الفصل التاسع

تحليل الإنحدار الخطي البسيط

SIMPLE LINEAR REGRESSION ANALYSIS

محتويات الفصل :

(٩-١) نظرة عامة على محتويات الفصل.

(٩-٢) العلاقة بين متغيرين : نموذج الإنحدار الخطى البسيط .

(٩-٣) تقدير معالم نموذج الإنحدار الخطى البسيط .

(٩-٤) الإستنتاج الإحصائي لنموذج الإنحدار الخطى البسيط .

(٩-٥) درجة الإعتماد على التقديرات والتنبؤات .

(٩-٦) العوامل المؤثرة فى الأخطاء المعيارية للإنحدار .

(٩-٧) الإرتباط : قياس العلاقة الخطية بين Y , X .

(٩-٨) نموذج الإنحدار البسيط : مثال شامل .

(٩-٩) ملخص .

ملحق 9 : تعليمات الحاسب بإستخدام Minitab ; SAS

الفصل التاسع

تحليل الإنحدار الخطي البسيط

SIMPLE LINEAR REGRESSION ANALYSIS

(٩-١) نظرة عامة على محتويات الفصل Bridging To New Topics

سوف نتعرف - فى هذا الفصل - على طرق لدراسة العلاقة الإحصائية بين المتغيرات. ويعرف النموذج الإحصائي - كما هو معرف فى الفصل الأول - بأنه معادلة رياضية توضح كيف يرتبط متغير ما بمتغير أو بعدة متغيرات أخرى. وبناء النماذج الإحصائية تساعدنا فى تعريف العوامل المحددة لمعظم إختلاف نواتج العمليات . كما تساعدنا أيضاً على التنبؤ . فبفرض وجود نموذج إحصائى، يمكن التنبؤ بقيمة أحد المتغيرات عند معلومية قيم المتغير أو المتغيرات الأخرى .

بفرض أن مثمن عقارات يود تقدير القيمة السوقية لمنزل (القيمة السوقية هى القيمة المتوقعة لسعر بيع المنزل ، إذا ما تم بيعه) فيجب أن يكون لديه معلومات عن عينة مناظرة لأسعار بيع عدة منازل لها نفس خصائص هذا المنزل. والمشكلة الإحصائية هي: بالإعتماد على بيانات هذه العينة، كيف يمكن التنبؤ بسعر البيع لهذا المنزل بالتحديد .

بفرض وجود معلومات عن أسعار البيع لمجموعة من المنازل المناظرة فإن أفضل مقدر **best predicator** هو $\overline{X}$ (الوسط الحسابى لأسعار عينة المنازل المناظرة). ويكون متوسط العينة $\overline{X}$ أفضل مقدر إذ لم تتوافر معلومات أخرى غير سعر البيع. ومن خلال تقدير العلاقة بين سعر بيع المنزل وحجمه، يمكننا تحسين دقة التنبؤ بدرجة كبيرة. وتحليل الإنحدار: هو الطريقة التى تهتم ببناء العلاقة بين متغير (سعر البيع) ومتغير أو عدة متغيرات أخرى مثل (الحجم) .

وكما يوضح هذا المثال ، فإن هذا المدخل فى التنبؤ يبحث فى درجة إعتماد متغير ما على متغير آخر. ويعد تحليل الإنحدار طريقة لدراسة الارتباط **association** بين متغير ما (سعر البيع) ومتغير آخر أو أكثر (مثل الحجم) . فإذا كان الإهتمام بمتغير واحد فقط، سمى هذا الأسلوب «تحليل الإنحدار الخطى البسيط» **Simple Linear Regression Analysis** وهو موضوع هذا الفصل. بينما إذا كان الإهتمام بعدة متغيرات أخرى سمى هذا الأسلوب «تحليل الإنحدار المتعدد» **Multiple Linear Regression Analysis** وهو موضوع الفصل العاشر. ونموذج الإنحدار : هو معادلة رياضية تهتم بالتنبؤ بقيم متغير ما بالإعتماد على قيم متغير أو عدة متغيرات أخرى .

العمليات الحسابية في تحليل الانحدار تعتبر مجهدة بصورة واضحة عن تلك التي قابلناها في الفصول (٥-٨)، وفى الواقع فإن معظم تطبيقات الإنحدار الخطى المتعدد، يكون من الصعب استخراج نتائجها، بدون إستخدام البرامج الإحصائية مثل **Minitab** أو **SAS** . ومع ذلك فإننا سوف

نفترض أن الطالب يمكنه إستخدام الآلة الحاسبة لإيجاد تحليل جيد، قبل المحاولة على الحاسب. لذلك فإن معظم الأمثلة والتمارين لهذا الفصل سوف تحتوى على مجموعات صغيرة وبسيطة من البيانات. في الفصل العاشر سوف نعتمد كلية على الحاسب الآلي. حيث أن مفاهيم الانحدار المتعدد في معظمها هي امتداد للانحدار البسيط، يكون من المهم أن نتعلم جيداً هذه المفاهيم في هذا الفصل.

(٩-٢) العلاقة بين متغيرين : نموذج خط الإنحدار البسيط :

Relationships Between Two Variables: The Simple Linear Regression Model

بفرض أننا نرغب فى التنبؤ بسعر بيع أحد المنازل آخذاً فى الإعتبار حجم هذا المنزل. وبفرض أن Y تشير إلى سعر البيع ؛ X المساحة بالقدم المربع، سنشير إلى Y على أنه المتغير التابع أو متغير الاستجابة **response variable** وإلى X على أنه المتغير المفسر/المتنبئ به (المتغير الذى بنى عليه التنبؤ) **predictor variable**.

(٩-٢-١) علاقات الأرتباط مقابل علاقات السبب والنتيجة :

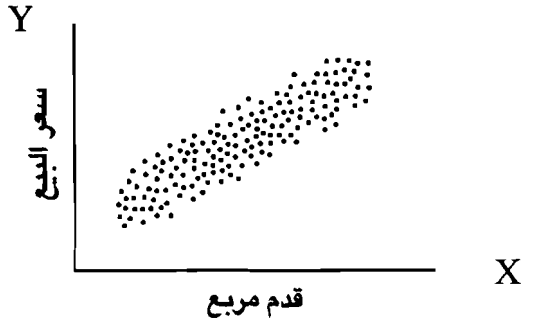
إذا كانت معلومية قيم X سوف تفيد فى التنبؤ بقيم Y، فإننا نقول انه يوجد إقتران أو إرتباط **assosciation** بين Y , X . ومن الأهمية فهم أن وجود الإرتباط بين متغيرين لا يعنى بالضرورة وجود علاقة السبب والنتيجة. والسببية تعنى أن التغير فى X يتسبب فى تغير مناظر فى Y بفرض ثبات جميع العوامل الأخرى المؤثرة على Y . وهذا ربما يكون حقيقيا للعلاقة بين حجم المنزل وسعر البيع، فزيادة حجم المنزل سوف تزيد قيمته إذا بقيت العوامل الأخرى ثابتة على ما هى عليه .

ويمكن أن يستخدم تحليل الإنحدار كنموذج إرتباط بين المتغيرات، ولكن لا يمكن أن نتأكد من أن هذه العلاقة سببية. والفكرة الأساسية هي تثبيت جميع العوامل الأخرى المؤثرة على المتغير التابع، مع تغيير المتغير المفسر.

شكل الإنتشار : Scatter Diagrams

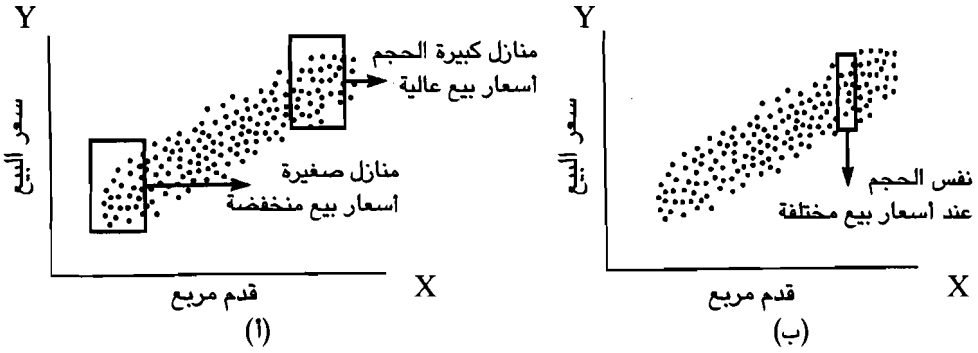
نرغب الآن في إنشاء نموذج يوضح العلاقة المصاحبة (إذا كان هناك علاقة) بين Y ، X. وسوف نبدأ بتحديد مجتمع الدراسة الذي نريد تقدير أو إستنتاج معلماته. إفترض أننا سوف نقوم بمشاهدة سعر البيع وكذلك الحجم لكل منزل في مجتمع الدراسة (في الحقيقة لانستطيع عمل هذه الدراسة نظراً لأن معظم هذه المنازل لم تباع حديثاً وبالتالي لن نستطيع معرفة سعر البيع). وبرسم البيانات المشاهدة، سوف نستطيع تكوين فكرة عن العلاقة بين Y ، X.

بفرض تسجيل السعر والحجم لكل منزل فى مجتمع المنازل المناظرة، ومن خلال الرسم البيانى للبيانات المسجلة، يمكن ملاحظة طبيعة العلاقة بين Y , X . وبفرض أن الشكل البيانى (شكل ٩-١) يوضح العلاقة بين سعر البيع Y والحجم X ، هذا الشكل يسمى بالشكل الإنتشارى. والشكل الانتشاري قدم من قبل في الفصل الثاني، كوسيلة لفحص طبيعة العلاقة بيانياً بين متغيرين.



شكل رقم (٩-١) : الشكل الانتشاري للمتغير Y مقابل X

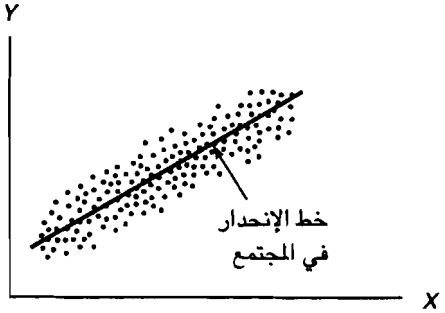
ولتوضيح شكل العلاقة التي قد تكون موجودة بين X , Y ننظر إلى (شكل ٩-٢). يلاحظ فى الجزء (أ) أن عدد كبير من المنازل (ذات قيم X الكبيرة) تميل إلى إرتفاع أسعار البيع (قيم كبيرة لقيمة Y) ، بينما يلاحظ فى الجزء (ب) أن العلاقة غير تامة لأن معظم المنازل التي لها نفس الحجم تكون عند أسعار مختلفة، ويحدث هذا لأن المنازل التى لها نفس الحجم تختلف عن بعضها البعض أخذا في الاعتبار عدة عوامل أخرى، مثل وجود مكيف، عدد الحمامات، وجود مدفأة. لذلك قد يبدو أن هناك ارتباط بين X، Y ولكنه بالتأكيد غير تام.



شكل رقم (٩-٢) : معنى العلاقة بين سعر البيع والحجم

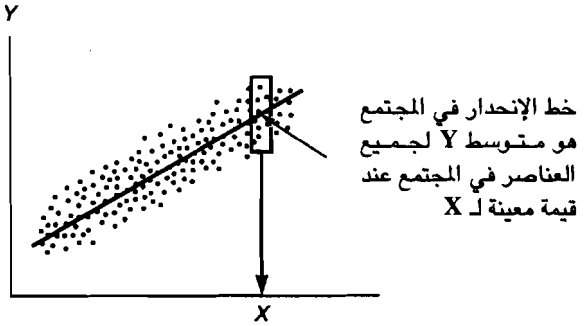
(٩-٢-٢) نموذج الإنحدار :

يبدأ تحليل الإنحدار بتعريف نموذج الإنحدار - بأنه معادلة رياضية تصف العلاقة بين X , Y فى المجتمع ، وعندما نهتم بمتغير تفسيرى واحد، فإن الشكل الإنتشارى يعتبر الخطوة المبدئية الأهم لتحديد نموذج الإنحدار المناسب. وفى إعتقادك من رؤية شكل (٩-١) ما هو شكل العلاقة بين X , Y ؟ إن هذه العلاقة تقترب من نموذج الخط المستقيم كما يوضحها شكل (٩-٣) التالى :



شكل (٩-٣): نموذج الانحدار المفترض للعلاقة بين X، Y

ويلاحظ من شكل (٩-٣) السابق أن معظم النقط لا تقع على الخط المستقيم، ويرجع ذلك لوجود متغيرات أخرى بخلاف المتغير X . لذلك فإن نموذج الإنحدار لا يمثل كل نقطة تماماً، ولكن قيم Y تميل إلى الإرتفاع مباشرة فى المجتمع بإرتفاع قيم X . وعلى سبيل المثال، المنازل الكبيرة يكون لها أسعار بيع أعلى. وبالتالي فإن نموذج الإنحدار يعكس القيمة المتوسطة للمتغير Y (أى متوسط سعر البيع) عند قيمة معينة للمتغير X (أي الحجم) فى المجتمع ، ويوضح ذلك شكل (٩-٤) التالى :



شكل (٩-٤)

نموذج الانحدار: خط يصور متوسط Y عند أي قيمة معينة لـ X

ومن المهم أن نكون قادرين على التعبير عن العلاقة الانحدارية بين متغيرين X , Y فى نموذج رياضى. وعلى سبيل المثال فإن الشكل الإنتشارى السابق يوضح أن نموذج الإنحدار المناسب فى المجتمع يمكن تمثيله بخط مستقيم. ويمكن كتابة معادلة الخط المستقيم على النحو التالى :

$$Y = \beta_0 + \beta_1 X \qquad \textbf{(9.1)}$$

حيث β_0 تشير إلى الجزء المقطوع من محور Y (قيمة Y عندما X=0)، β_1 تشير إلى ميل الخط المستقيم، وقد استخدمت الحروف اليونانية لتميز هذه الكميات لأنها تصف المجتمع، الجزء المقطوع β_0 ، الميل β_1 هى معلمات أو مؤشرات نموذج الإنحدار. وهذه القيم عادة غير معلومة ولكن يمكن تقديرها كما سنوضح فيما بعد .

ولتوضيح علاقة الخط المستقيم، دعنا نفترض معلومية معادلة الإنحدار فى المجتمع للمتغير Y (سعر البيع)، X (الحجم بالقدم المربع) هي :

$$Y = 60{,}000 + 30X$$

إذن الجزء المقطوع من محور Y هو ($\beta_0 = 60{,}000$) والميل ($\beta_1 = 30$) . ويمكن توضيح هذه المعادلة على النحو التالى :

1- المنزل الذى حجمه (X=3000) قدم مربع، يكون سعر بيعه فى المتوسط (Y = $150,000)

2- اذا كانت (X=0) ، فإن سعر البيع فى المتوسط يكون $60,000 . وهذا قد يمثل سعر البيع المتوسط لمنزل خالى (سيتم إيضاح ذلك فيما بعد) وهذا يحدث فقط إذا كانت البيانات تحتوى على قيمة (X = 0) .

3 - مقدار التغير فى Y عندما تتغير X من صفر إلى 3000 هو : (150,000 - 60,000 = 90,000)
وهذا يؤكد أن الميل هو (90,000 / 3000 = 30)

4- بما أن الميل يساوي 30 ، فإن متوسط سعر البيع يزيد بمقدار $30 مع إضافة قدم مربع واحد إلى مساحة المنزل .

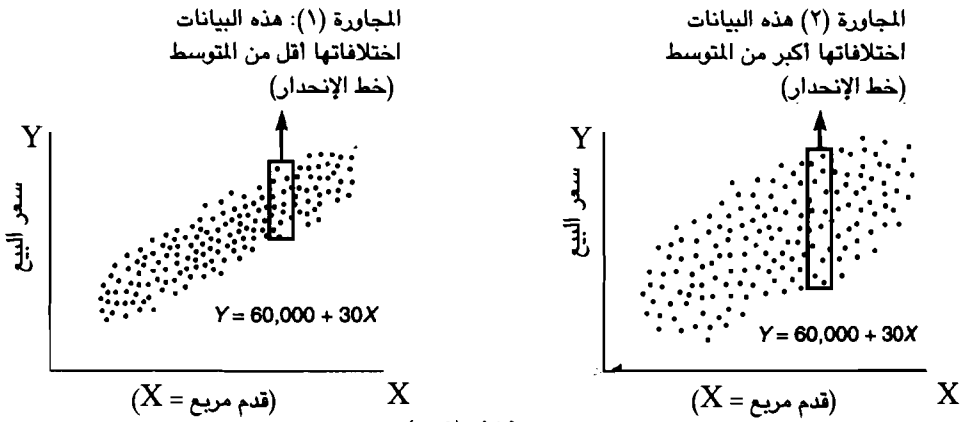
وبفرض أننا نرغب فى تقدير سعر البيع لمنزل مساحته 2800 قدم مربع. افترض أيضاً أن متوسط سعر البيع فى المجتمع لمنازل يمكن مقارنتها هو 120,000$ وأن متوسط مساحة المنزل2,000 قدم مربع. اذا كان حجم المنزل ليس محلا للاهتمام، فإن أفضل تقدير لسعر بيع أى منزل، هو متوسط سعر البيع فى المجتمع أى 120,000$. ولكن نموذج الإنحدار يستخدم المعلومات المتاحة عن مساحة المنزل فى التنبؤ بمتوسط سعر بيعه، وعلى هذا فإن نموذج الإنحدار يشير إلى أن متوسط سعر بيع أى منزل مساحته (X=2800) قدم هو :

$$(60,000 + 30\ (2800) = \$144000)$$

وهذا هو سعر البيع المتنبأ به للمنزل موضوع الاهتمام. وبدلاً من استخدام متوسط المجتمع كله كسعر تنبؤي، فإنه يمكننا الآن استخدام المتوسط فقط لتلك المنازل محل المقارنة والتي حجمها 2800 قدم مربع. ويلاحظ أن سعر البيع المتنبأ به يفوق المتوسط لسعر البيع 120000$، وذلك لأن المنزل الذي مساحته 2800 قدم مربع يكون أكبر من المتوسط 2000 قدم مربع لمجتمع المنازل المقارنة.

والحقيقة أننا لا نستطيع تحديد قيم β_0 , β_1 لأننا لا نستطيع ملاحظة المجتمع بالكامل. وإذا أمكننا سحب عينة ممثلة، فيمكن تقدير قيم β_0 , β_1 وعندئذ يمكن إستخدام تقدير لنموذج الإنحدار واستخدامه فى التنبؤ، وسوف نرى كيف يتم ذلك فيما بعد .

إن مجرد وجود نموذج انحدار لتوضيح العلاقة بين Y، X لايعني بالضرورة أن النموذج مناسب لتوفيق البيانات. وبسبب هذه الحقيقة، فإن التنبؤات الناتجة عن النموذج المستخدم يجب أن تكون قريبة نسبياً من القيم الفعلية المناظرة Y. وبعبارة أخرى انه يجب أن يكون توفيق البيانات بإستخدام نموذج الانحدار جيد. وللتوضيح، افترض أن لدينا بيانات تمثل المنازل في مجاورتين مختلفتين، افترض أيضا أنه قد تم استخدام معادلتى انحدار متماثلتين لهذه البيانات، كما يتضح ذلك من شكل (٩-٥). لاحظ انه، علي الرغم من استخدام معادلتي انحدار متماثلتين فإن البيانات من المجاورة الأولى يتم توفيقها بخط الانحدار بطريقة أفضل من البيانات من المجاورة الثانية: وبعبارة أخرى، فإن القيم الفردية Y تختلف قليلاً عن متوسط Y لكل قيم X. وكنتيجة لذلك فإن أخطاء التنبؤ سوف تكون أقل بالنسبة للمنازل من المجاورة الأولى عنها بالنسبة للمنازل من المجاورة الثانية. ولذلك نقول أن التوفيق أفضل في المجاورة الأولى لأن أخطاء التنبؤ تكون أقل.



شكل (٩-٥)

توضيح اختلاف البيانات حول خطى انحدار متماثلين

ونحتاج أن نكون قادرين على تقدير الدرجة التي تختلف بها البيانات عن خط الانحدار. فعند قيمة ما للمتغير X، فإن الاختلاف بين قيمة Y وخط الانحدار يسمى خطأ error، ويرمز لها بالرمز epsilon (ε). هذا الاختلاف هو الخطأ الذي نقع فيه إذا تم استخدام خط الانحدار للتنبؤ بالقيم الفردية للمتغير Y. وحتى يمكن الاحتساب لهذه الاختلافات عن خط الانحدار، فإن نموذج انحدار خطى كامل لتوضيح العلاقة بين المتغيرات Y، X لابد أن يحتوى على عنصر الخطأ كما يلي:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{الجزء المحدد: متوسط Y عند معلومية X}} + \underbrace{\varepsilon}_{\text{الجزء العشوائي: الانحراف غير المفسر من خط انحدار المجتمع}} \qquad (9.2)$$

تعرف المعادلة (9.2) باسم نموذج الإنحدار الخطى البسيط، ويوضح أن قيمة Y فى المجتمع تحدد من خلال :

(1) $(\beta_0 + \beta_1 X)$ المتوسط لجميع مفردات المجتمع والتي لها نفس قيم X .

(2) ε قيمة إختلاف المتغير Y عن خط إنحدار المجتمع .

والجزء الأول $(\beta_0 + \beta_1 X)$ يسمى بالجزء المحدد لأنه يتحدد بالكامل من خلال قيمة المتغير X . وهذا الجزء المحدد يطلق عليه خط إنحدار المجتمع **population regression line**. أما عنصر الخطأ ε ، فيطلق عليه الجزء العشوائي **random Component** لأن قيمته لأي قيمة فردية في المجتمع يفترض أن تختلف بطريقة غير متوقعة لجميع مفردات المجتمع والتي لها نفس قيم X . ولهذا السبب فإنه يشار إليه كخطأ عشوائى **Random error**. وعموماً فإن الاستنتاج الاحصائي بإستخدام اسلوب الإنحدار يعتمد أساساً على فرض عشوائية الخطأ .

بالإضافة إلى ذلك فإن الخطأ العشوائى لقيم X يقاس عن طريق تباين الخطأ **error variance** والذي يرمز له بالرمز σ_ε^2. وهذا أيضا هو نفس تباين أخطاء التنبؤ والتي تصاحب مفردات المجتمع. أي أن σ_ε^2 عبارة عن تباين Y لجميع مفردات المجتمع والتي تأخذ قيمة عامة X. ويلاحظ أنه مثل قيم β_0 , β_1 فإن σ_ε^2 عبارة عن معلمة غير معلومة والتي لابد من تقديرها بإستخدام بيانات عينة الدراسة. وكما يدلنا الشكل (٩-٥)، فإن قيمة σ_ε^2 تدل على مدى قرب توفيق بيانات العينة بإستخدام نموذج الانحدار. فإذا كانت العلاقة بين Y، X علاقة خطية حقيقية، فإن التوفيق باستخدام التعبير (9.2) لبيانات عينة الدراسة جيد حيث يكون حجم σ_ε^2 صغير نسبيا. ولكن إذا كان توفيق البيانات ليس على درجة عالية من الجودة فإن حجم σ_ε^2 سوف يكون كبير وسيظهر الخطأ في التعبير (9.2) بأنه عنصر كبير ذو دلالة.

ان الهدف الأساسي في تحليل الإنحدار هو أن نحدد معادلة انحدار لها معنى وتوفق بيانات عينة الدراسة بحيث يكون تباين الخطأ أصغر مايكون. والجدير بالذكر أن توفيق نموذج مفترض لبيانات عينة لايعنى ببساطة أن هذا النموذج كافى sufficient لهذا التوفيق، ولكن يجب: (١) اختبار نموذج مناسب (بمعنى أنه يمكننا رسم شكل الانتشار والتأكد من أن خط الإنحدار هو المناسب). (٢) تقييم نموذج الانحدار (٣) نأخذ في الاعتبار أن اضافة متغيرات مفسرة إضافية للنموذج من شأنه تخفيض تباين الخطأ (وسوف يكون هذا هو موضوع الفصل العاشر).

مثال (٩-١)

قامت شركة كهرباء محلية بإختيار عدة منازل متشابهة مأهولة بالسكان لتستخدمها فى بناء نموذج تجريبى لإستهلاك الطاقة (كيلو وات / يوم) كدالة فى درجات حرارة العالية خلال أشهر الشتاء فى منطقة جغرافية ما. وتم الحصول على البيانات التالية خلال 18 يوم .

درجة الحرارة	-1	1.5	3.5	-3	.5	2.5	4	5	-5
الطاقة المستخدمة	94	81	79	97	88	75	74	67	107
درجة الحرارة	-.5	9	9.5	7	3	-2	6	8	11
الطاقة المستخدمة	86	58	55	65	73	91	65	58	52

(أ) حدد المتغير التابع والمتغير المفسر فى هذا التطبيق ، بناءً على هدف شركة الكهرباء فى تقديم نموذج الإنحدار ؟

(ب) إرسم البيانات ، بالإعتماد على الشكل الإنتشارى ، هل يظهر وجود علاقة بين الطاقة المستخدمة يومياً ودرجات الحرارة المرتفعة؟ هل تعتقد أن هذه العلاقة خطية؟ إذا كانت كذلك ، وضح ذلك؟

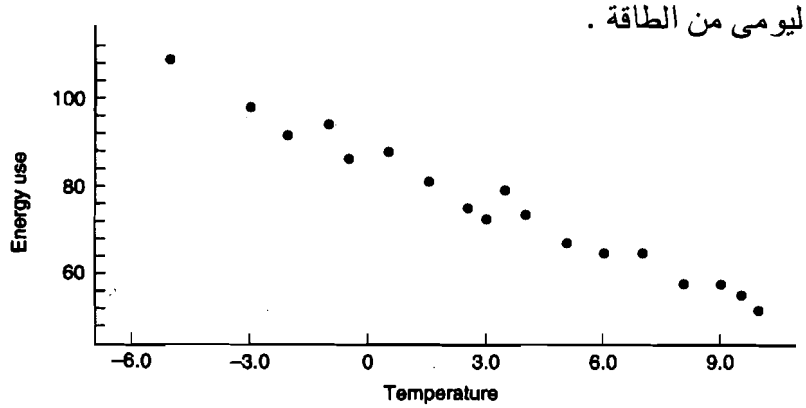
(ج) إذا كانت العلاقة التقريبية بين الطاقة ودرجات الحرارة المرتفعة تمثل خط مستقيم ، ما هى إشارة ميل هذا الخط؟ هل إجابتك تتفق مع معلوماتك عن العلاقة بين الطاقة المستخدمة ودرجات الحرارة المرتفعة ؟

الحـــل :

(أ) حيث أن شركة الكهرباء تريد التنبؤ بإستهلاك الطاقة، فإن إستهلاك الطاقة هو المتغير التابع. ويكون المتغير المفسر هو درجة الحرارة اليومية لأن شركة الكهرباء تشعر بأنه يفسر درجة الإختلاف فى الإستهلاك اليومى للطاقة .

(ب) الشكل الإنتشارى موضح فى شكل (٩-٦) . يظهر الشكل السابق التوقع الطبيعى للعلاقة بين الطاقة المستخدمة وإرتفاع درجات الحرارة. وبإرتفاع درجات الحرارة، تنخفض الطاقة المستخدمة. وفى الواقع فإنه يمكن ملاحظة - بالعين المجردة - أن النقط تقترب من الخط المستقيم، لذلك فإن أفضل توفيق لهذه البيانات هو الخط المستقيم، وتظهر العلاقة أيضاً صغر حجم الخطأ العشوائى .

(ج) وبما أن إستهلاك الطاقة يتناقص - فى شكل خطى - بزيادة درجات الحرارة فى أيام الشتاء فإننا نتوقع أن يكون ميل خط الإنحدار سالب. وهذا يتمشى مع فهمنا المسبق لتأثير درجة الحرارة على الإستهلاك اليومى من الطاقة .



شكل (٩-٦)
الشكل الإنتشارى للطاقة المستخدمة مقابل درجة الحرارة

مثال (٩-٢)

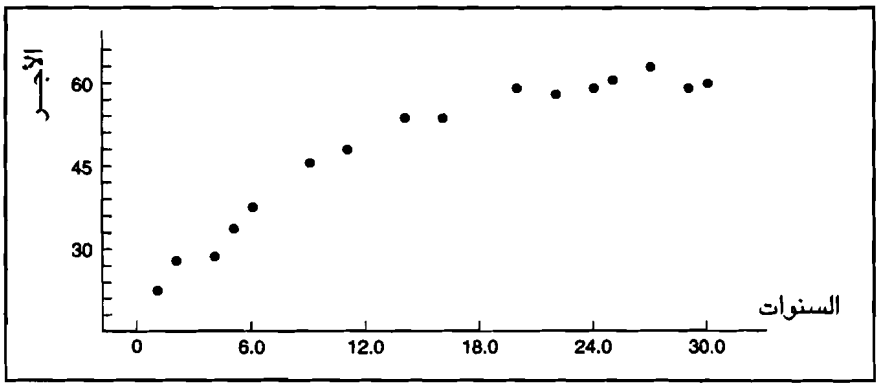
نفذت دراسة لفحص العلاقة بين عدد سنوات الخبرة والأجر السنوى لمجموعة من الأفراد فى حرفة معينة في منطقة جغرافية ما لعينة من 16 حرفى . (الأجر بألف دولار)

السنوات (X)	1	2	4	5	6	9	11	14
الأجـــر (Y)	23	27	29	34	46	46	48	54
السنوات (X)	16	20	22	24	25	27	29	30
الأجـــر (Y)	54	29	58	59	61	63	59	60

بالإعتماد على الشكل الإنتشارى ، هل يظهر علاقة بين X , Y ؟ وهل هذه العلاقة خطية؟ إشرح ؟

الحل

يوضح شكل (٩-٧) شكل الانتشار . وبناءً على ذلك الشكل ، فإنه لايوجد أى شك بأن هناك علاقة بين الأجر وسنوات الخبرة ، ولكن يوجد شك فى أن الخط المستقيم يعتبر أفضل تمثيل للبيانات . فعلى سبيل المثال ، لاحظ أنه كلما زادت سنوات الخبرة ، فإن الأجر يتزايد . ولكن معدل الزيادة في الأجر ينخفض (تكون الزيادة بطيئة) كلما زاد عدد سنوات الخبرة . وبالتالي فإن شكل الانتشار يوضح أن العلاقة تكون في شكل منحنى . أى أن استخدام الخط المستقيم لن يكون هو التمثيل المناسب في توفيق البيانات (وسوف نوضح كيفية التعامل مع المنحنى في الفصل ١٥).



شكل (٩-٧)
الشكل الانتشاري للأجر مقابل سنوات الخبرة

مثال (٩-٣)

رغب مدير الجامعة فى دراسة العلاقة بين متوسط درجات أو نقط التخرج (GPA) والعمر لعينة من 15 طالب فى مرحلة البكالوريوس فى كلية التجارة . وكانت البيانات كالتالى :

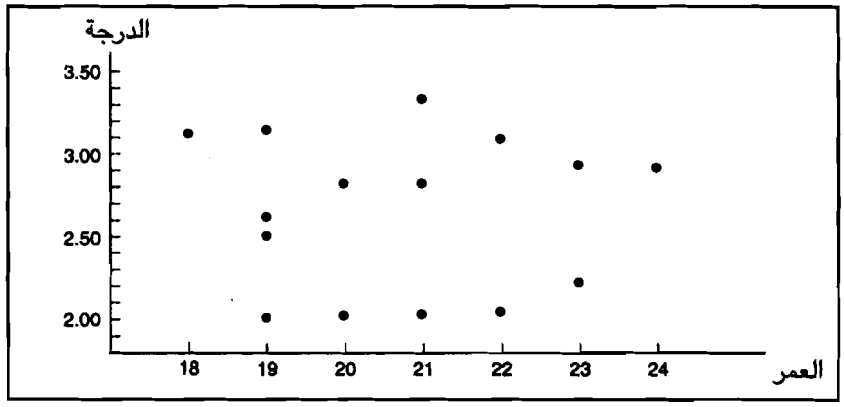
العمر (X)	20	21	19	19	20	21	18	22
متوسط الدرجات (Y)	2.13	3.38	2.07	2.63	2.85	2.16	3.15	3.12
العمر (X)	23	19	22	24	23	19	21	
متوسط الدرجات (Y)	2.97	2.52	2.17	2.95	2.25	3.17	2.85	

انشأ الشكل الإنتشارى، ثم صف العلاقة بين متوسط الدرجات والعمر اذا كان يبدو أن هناك علاقة ؟

الحل :

يوضح شكل (٩-٨) شكل الانتشار. ومن هذا الشكل، يتضح لنا انه لاتوجد علاقة واضحة بين متوسط التقدير (متوسط نقط الدرجات) والعمر.

أيضاً يلاحظ أن متوسط GPA للطلاب صغيرى السن يساوي تقريباً متوسط GPA للطلاب كبيرى السن. ويكون أفضل ما نستطيع عمله هنا هو أن نقوم برسم خط مستقيم بمجرد النظر بحيث يكون موازيا للمحور الأفقي (X-axis) ويتقاطع مع المحور الرأسي (Y-axis) عند 2.7 تقريباً. وحيث أن ميل أي خط مستقيم يوازي المحور الأفقي يساوي الصفر، فإن هذا يعنى انه لاتوجد علاقة خطية بين المتغيرات Y، X. وبالتالي فإن المعلومات عن سن الطلاب لايكون لها قيمة كبيرة عند التنبؤ بمتوسط نقطة التقدير. وأى خط مستقيم يتوازى مع المحور X فإن ميله يساوى الصفر، ويعنى ذلك أنه لا توجد علاقة خطية بين X , Y .



شكل (٩-٨)
الشكل الانتشاري لمتوسط الدرجات مقابل العمر

(٩-٢-٣) إستخدامات نماذج الإنحدار :

تعتبر نماذج الإنحدار أداة مفيدة جداً فى ادارة العمليات التجارية . بصفة عامة يمكن إستخدام نماذج الإنحدار فى نقطتين أساسيتين: أولاً فى إظهار شكل العلاقة، ثانياً فى التنبؤ والتقدير .

إن الهدف الأساسي لنموذج الانحدار هو تمثيل نظام معقد في شكل بسيط، ذو معنى حتى يمكنه تزويدنا بفهم أفضل بخصائص ذلك النظام. هذا الفهم يكون من بين اهتمامات المدير أو متخذ القرار، لتوضيح ذلك افترض أن خط الانحدار المقدر للمثال (٩-١) (الطاقة المستهلكة يومياً مقابل درجات الحرارة العالية). على الصورة التالية:

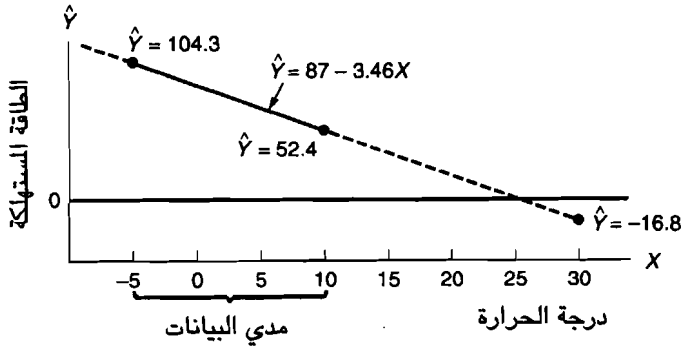
$$\hat{Y} = 87 - 3.46\,X$$

حيث $\hat{Y}$ «Y - hat» قيمة Y المقدرة عند قيمة معينة للمتغير X. وتوضح المعادلة السابقة أن الطاقة اللازمة تنخفض فى المتوسط بمقدار 3.46 كيلو وات تقريباً مع إرتفاع درجة الحرارة بمقدار درجة واحدة. وهذا هو تفسير تقدير الميل. ومن المهم أن تضع في ذهنك أن الميل هو الأساس في وصف العلاقة الخطية بين متغيرين، فإذا كان X، Y يرتبطا بعلاقة خطية، فإن الميل لايمكن أن يكون صفر. فإذا كان الميل مساوياً للصفر فإن العلاقة بين X، Y لا يمكن أن تكون خطية.

والقواعد الإحصائية للتعليق على نموذج الإنحدار يجب أن تكون خلال الفترة التى تقع داخلها قيم المتغير X، لأننا سنقع فى خطأ إذا كان التعليق على النموذج يتضمن قيم تقع خارج نطاق المتغير X، إلا إذا كنا نتحدث عن القواعد النظرية. وبالرجوع إلى نموذج الطاقة المستهلكة، فيلاحظ أن مدى درجات الحرارة بين (5-)، (10)، لذلك فإن النموذج سوف يستخدم فى التنبؤ بالطاقة المستهلكة خلال هذا المدى لدرجات الحرارة. ولتوضيح ذلك، نفترض أننا استخدمنا هذا النموذج فى التنبؤ بالطاقة المستهلكة خلال يوم صيفى عند درجة حرارة 30، فتنبؤ النموذج يكون:

$$(\ \hat{Y} = 87 - 3.46(30) = -16.8 \text{ kilowatts})$$

ومن الواضح أنها نتيجة غير منطقية، لأنه من الخبرة، يمكن القول أن الطاقة المستهلكة يومياً ستزيد مع إرتفاع درجات الحرارة صيفاً، بسبب إستخدام التكييف، وهذا ما يوضحه شكل (٩-٩).



شكل (٩-٩): العلاقة الخطية بين درجة الحرارة والطاقة المستهلكة

ما هو التفسير المناسب للجزء المقطوع من محور Y؟ توضح معادلة الإنحدار أن متوسط الطاقة المستهلكة يومياً 87 كيلو وات عندما تكون درجة الحرارة مساوية للصفر أى X = صفر، وهذا تفسير مناسب لتقدير قيمة β_0. فى بعض التطبيقات قد لايكون لهذا التفسير معنى، فعندما نضع X = صفر، يجب أن تكون هذه القيمة ضمن نطاق X، فى هذه الحالة يكون التفسير له معنى. بينما إذا لم يحتوى نطاق X على القيمة صفر، فإننا نتوقع الحصول على قيمة ليس لها معنى وقد تكون خيالية. وبالرجوع إلى مثال الطاقة المستهلكة، نجد نطاق X من (5-) إلى (10) من درجات، ولأن هذا المدى يحتوى على قيمة X = صفر. فإن الجزء المقطوع من محور Y سوف يكون له معنى.

وبإختصار فإن التعليق على نموذج الإنحدار المقدر، يكون من خلال مدى قيم X فى العينة، ومن خلال هذا المدى يكون مفتاح التعليق على الميل المقدر. وهو يشرح مقدار التغير فى Y المناظر للتغير فى X .

التقدير والتنبؤ :

بعد تكوين وجهة نظر عن العلاقة، وفهمها، والاعتقاد بأن نموذج الانحدار المقدر يمثل بدقة هذه العلاقة، فإننا غالباً نريد استخدام النموذج للتنبؤ بقيم متغير الاستجابة، فإذا نظرنا إلى مثال تقييم الأصول، فإن المسعر أو المثمن يستطيع تقدير القيمة السوقية للمنزل عن طريق التعويض بحجم المنزل بين مجموعة أخرى من الاعتبارات، في معادلة الانحدار المقدرة. وفي هذا السياق فإنه يوجد طريقتين مختلفتين يمكن استخدام نموذج الإنحدار فيهما:

(أ) التقدير : يستخدم نموذج الإنحدار المقدر فى تقدير القيمة المتوسطة للمتغير Y عند قيمة معلومة لـX . بفرض أن شركة للإستثمار تود المقارنة بين القيم الفعلية لأصول قابلة للمقارنة فى منطقتين، وبالتحديد هى ترغب فى مقارنة متوسط أسعار البيع فى المدينة (أ) والمدينة (ب). يمكن أن يستخدم تحليل الإنحدار لهذا الغرض. عينات من الأصول القابلة للمقارنة من كل مدينة يجب تحديدها. وباستخدام نموذج الانحدار، حيث أن Y تشير إلى سعر البيع (10 آلاف دولار) ، X القدم المربع وكانت نتائج النماذج كالآتى : (X بآلاف الأقدام المربعة)

المدينة (أ) : ($\hat{Y} = .8 + 6X$) ونطاق X من 2 إلى 3.2

المدينة (ب) : ($\hat{Y} = 1.1 + 5.2X$) ونطاق X من 2 إلى 3.2

لاحظ أن هذه النماذج مناسبة عندما يكون X من 2 إلى 3.2 ، فمثلاً المنازل التى لها مساحة قدرها 2200 قدم مربع أى ($X = 2.20$) فإن هذه النماذج تقدر متوسط أسعار البيع كالآتي :

فى المدينة (أ) : متوسط سعر البيع 140,000 دولار .

فى المدينة (ب) : متوسط سعر البيع 125,400 دولار .

(تم الحصول على هذه التقديرات بالتعويض عن X بالقيمة 2.20 فى كل معادلة) .

النقطة الهامة هنا أن شركة الاستثمار مهتمة بالقيمة المتوسطة Y عند قيمة معينة X، وليس لديها اهتمام خاص بالتنبؤ بسعر البيع لأى منزل مساحته 2200 قدم مربع.

(ب) التنبؤ : قد يستخدم نموذج انحدار مقدر للتنبؤ بقيمة Y لكل مفردة من مفردات المجتمع (مجتمع الدراسة)، بمعلومية قيمة معينة للمتغير X. وبالرجوع إلى مثال تقييم الأصول من وجهة نظر المسعر أو المثمن. وبصفة خاصة، إذا كان خط الانحدار المقدر للمدينة A هو: ($\hat{Y} = .8 + 6.0\,X$). لن يكون هذا النموذج كافى للمسعر ليقوم بتقدير قيمة متوسط سعر البيع: ويجب عليه أن يتنبأ بسعر البيع لكل مفردة في المجتمع. هذا التنبؤ يمكن الحصول عليه بنفس الطريقة التي نقدر بها قيمة المتوسط للمتغير Y، بمعلومية قيمة X. حيث نقوم ببساطة بالتعويض بقيمة X في معادلة الانحدار. ولهذا تكون قيمة سعر البيع المتوقع أو المتنبأ به لمنزل مساحته 2200 قدم مربع ($X = 2.2$) هي 140000 ($\hat{Y} = .8 + 6.0\,(2.2) = 14$). وربما تتساءل لماذا نفرق بين تقدير المتوسطات والتنبؤ بكل قيمة من قيم Y ، إذا كان لهما نفس النتائج لنفس قيمة X. ويرجع السبب

في ذلك للدقة، فإذا أردنا استخدام النموذج، فإنه يكون من الأهمية بمكان أن نعلم ما هو حجم الخطأ في حالة التقدير وكذلك في حالة التنبؤ. وكما سنرى في الجزء (٩-٥) فإن الخطأ يكون أكبر في حالة التنبؤ بقيمة Y بوصفها عنصر من مجتمع عنها في حالة تقدير متوسط قيم Y لكل عناصر المجتمع والذي يحتوى على قيمة معينة للمتغير X.

تمارين

(٩-١) فى دراسات الإنحدار، ما هو الفرق بين المتغير التابع والمتغير المفسر (المستقل) ؟

(٩-٢) فى دراسات الإنحدار، ماذا نريد أن نحدد بالنسبة لكل من المتغير التابع والمتغير المفسر (المستقل)؟

(٩-٣) إشرح ما هو المطلوب لبناء دليل علاقة السبب والنتيجة بين المتغيرين X , Y فى تحليل الإنحدار ؟

(٩-٤) إشرح الهدف من نموذج الإنحدار ؟

(٩-٥) ما هو الأسلوب الذى يتم إستخدامه مبدئياً لتحديد نموذج الإنحدار المناسب ؟

(٩-٦) إفترض أن نموذج الإنحدار المقدر لتحديد العلاقة بين أسعار البيع ومساحة المنزل فى جزء معين فى المدينة هو عبارة عن :

$$\hat{Y} = .5 + 6.8\,X$$

(أ) ماذا يعنى الرقم (.5) فى هذه العلاقة ؟

(ب) ماذا يعنى الرقم (6.8) فى هذه العلاقة ؟

(ج) هل تحديد هذه المعادلة يعنى بالضرورة أنها تعبر عن العلاقة بين سعر البيع ومساحة المنزل ؟ إشرح ذلك ؟

(د) إذا كانت إجابتك فى الجزء (ج) بالنفى، ما هى المعلومات أو التحليل الاضافي الضروري لإنشاء نموذج إنحدار مناسب أو غير مناسب لوصف العلاقة بين سعر البيع ومساحة المنزل ؟

(٩-٧) إفترض أن نموذج الإنحدار المقترح لتمثيل العلاقة بين الأجور وعدد سنوات الخبرة المذكورة فى المثال (٩-٢) هو :

$$\hat{Y} = 28.9 + 1.26\,X$$

(أ) ماذا يعنى الرقم (28.9) فى هذه العلاقة ؟

(ب) ماذا يعنى الرقم (1.26) فى هذه العلاقة ؟

(ج) هل تحديد هذه المعادلة يعنى بالضرورة أنها تعبر عن العلاقة بين الأجر وعدد سنوات الخبرة ؟ إشرح ذلك ؟

(د) إذا كانت إجابتك فى الجزء (ج) بالنفى ما هى المعلومات المطلوبة أو التحليل المطلوب لإنشاء نموذج إنحدار مناسب أو غير مناسب لتمييز وتوضيح العلاقة بين الأجر وسنوات الخبرة ؟

(٩-٨) عرف مكوني نموذج الإنحدار البسيط ($Y = \beta_0 + \beta_1 X + \varepsilon$) وإشرح معناهما؟

(٩-٩) أذكر وإشرح نوعى الإستخدام لنماذج الإنحدار ؟

(٩-١٠)محلل تأمينى يريد أن يحدد إلى أي مدى يرتبط دخل الأسرة ومقدار مبلغ تأمين الحياة على رب الأسرة. فإذا كانت درجة الأرتباط قوية، فإن الدخل يكون بمثابة مؤشر جيد على مقدار مبلغ التأمين الذى تطلبه شركة التأمين. وبناءاً على بيانات مجمعة من 18 أسرة تم الحصول على البيانات التالية (تم كتابة البيانات بآلاف الدولارات) :

الدخل	45	20	40	40	47	30	25	20	15
مبلغ التأمين	70	50	60	50	90	55	55	35	40
الدخل	35	40	55	50	60	15	30	35	45
مبلغ التأمين	65	75	105	110	120	30	40	65	80

(أ) ارسم شكل الإنتشار لهذه البيانات ، هل أمكنك تحديد العلاقة بين هذين المتغيرين كما يظهر من الرسم، وإذا كان الأمر كذلك، ما هو نوع تلك العلاقة المفترضة ؟

(ب) وإذا كانت تلك العلاقة خطية، هل تعتقد أن ميل ذلك الخط يكون موجب أم سالب؟ إشرح ذلك ؟

(٩-١١) يعلم طلبة الجامعة أنه كلما زاد متوسط نقط التخرج (GPA) كلما كان هناك فرصة أكبر للحصول على عمل أفضل عند التخرج. إفترض أن البيانات التالية توضح متوسط نقط التخرج لعدد 15 خريج من تلك الجامعة وبداية رواتبهم (بآلاف الدولارات) :

متوسط نقط التخرج (GPA)	2.95	2.30	3.40	3.60	3.20	2.85	3.10	2.85
بداية المرتب	23.5	25.0	26.1	27.4	26.2	20.0	23.0	23.8
متوسط نقط التخرج (GPA)	3.05	2.70	2.75	3.10	3.15	2.95	2.75	
بداية المرتب	20.7	19.4	20.5	22.2	24.0	22.2	21.8	

(أ) ارسم شكل الإنتشار لهذه البيانات ، هل أمكنك تحديد العلاقة بين هذين المتغيرين كما يظهر فى الرسم ، وإذا كان الأمر كذلك، ما هو نوع تلك العلاقة المفترضة؟

(ب) وإذا كانت تلك العلاقة خطية، هل تعتقد أن ميل ذلك الخط يكون موجب أم سالب؟ إشرح ذلك؟

(٩-١٢) ما يلى بيانات تعبر عن الطول X والوزن Y لعينة عشوائية مكونة من 10 إناث عاملين بإحدى الشركات الكبرى :

الطول (بالبوصة)	68	67	65	68	64
الوزن (بالرطل)	119	118	129	135	123
الطول (بالبوصة)	67	66	65	64	66
الوزن (بالرطل)	140	125	132	118	130

(أ) ارسم شكل الإنتشار لهذه البيانات ، هل أمكنك تحديد العلاقة بين هذين المتغيرين كما يظهر من الرسم، وإذا كان الأمر كذلك، ما هو نوع تلك العلاقة المفترضة ؟

(ب) وإذا كانت تلك العلاقة خطية، هل تعتقد أن ميل ذلك الخط يكون موجب أم سالب؟ إشرح ذلك ؟

(٩-١٣) كيف يتأثر إستهلاك الكحوليات بسعر بيعها ؟ توضح البيانات التالية الأسعار النسبية (بالسنت) للكحوليات لكل وحدة إستهلاك (باللتر) من الكحوليات وذلك فى الفترة من 1948 - 1967 في مدينة إونتاريو بكندا .

(أ) ارسم شكل الإنتشار لهذه البيانات، هل أمكنك تحديد العلاقة بين هذين المتغيرين كما يظهر من الرسم، وإذا كان الأمر كذلك، ما هو نوع تلك العلاقة المفترضة ؟

السنة	السعر النسبي (X)	وحدة الإستهلاك (Y)	السنة	السعر النسبي (X)	وحدة الإستهلاك (Y)
1948	5.7	7.09	1958	4.3	7.96
1949	5.8	7.18	1959	4.3	7.77
1950	5.5	7.23	1960	4.3	8.14
1951	5.2	7.23	1961	4.3	8.14
1952	5.1	7.32	1962	4.1	8.23
1953	5.5	7.64	1963	4.0	8.46
1954	5.6	7.73	1964	3.9	8.74
1955	4.7	7.55	1965	3.8	8.77
1956	4.5	7.91	1966	3.9	9.18
1957	4.4	7.86	1967	3.5	8.91

(ب) وإذا كانت تلك العلاقة خطية، هل تعتقد أن ميل ذلك الخط يكون موجب أم سالب؟ إشرح ذلك ؟

(٩-١٤) كيف يمكن التنبؤ بقيمة الضرائب التى يدفعها المواطن (دافع الضريبة) بمعرفة دخله الإجمالى؟ البيانات التالية توضح عينة عشوائية مكونة من 14 مواطن بحيث يتضح منها الدخل الإجمالى ونسبة الضرائب المدفوعة فى سنة معينة :

الدخل الإجمالي (بالآف الدولارات)	25.6	42.2	57.6	98.8	10.4	30.1	40.0
نسبة الضرائب المدفوعة	15.4	16.8	19.7	21.7	10.8	15.2	18.9
الدخل الإجمالي (بالآف الدولارات)	29.3	16.1	18.0	88.2	34.0	22.1	70.0
نسبة الضرائب المدفوعة	15.9	12.0	14.1	21.1	17.6	14.8	21.6

(أ) ارسم شكل الإنتشار لهذه البيانات ، هل أمكنك تحديد العلاقة بين هذين المتغيرين كما يظهر من الرسم، وإذا كان الأمر كذلك، ما هو نوع تلك العلاقة المفترضة ؟

(ب) وإذا كانت تلك العلاقة خطية، هل تعتقد أن ميل ذلك الخط يكون موجب أم سالب؟ إشرح ذلك؟

(٩-٣) تقدير معالم نموذج الإنحدار البسيط :

Estimating The Parameters of The Simple Linear Model

لتوفيق نموذج الإنحدار المفترض ليمثل بيانات العينة، فإن أول خطوة هى تقدير معالم النموذج β_1 , β_0 لنموذج الانحدار البسيط. وكنتيجة لتقدير β_1 , β_0 ، نكون في موقف يتطلب تقدير تباين الخطأ σ_ε^2. ولكننا سندرس أولاً طرق الحصول على بيانات العينة.

(٩-٣-١) الحصول على بيانات العينة :

يوجد ثلاث طرق للحصول على بيانات العينة :

1 – المعاينة العشوائية البسيطة **Simple Random Sampling**: أحياناً يتم إختيار العينة العشوائية البسيطة بحيث يكون كلا من X , Y متغيرات عشوائية. غير أنه من الأفضل عند تقدير خط إنحدار المجتمع إختيار قيم X بعناية ثم تحديد قيم Y المناظرة لها تلقائياً .

2 – المعاينة العشوائية لقيم X المختارة **Random Sampling for Selected X-Values**: يتم إختيار العينة العشوائية لقيم Y بناءاً على قيم X المحددة مسبقاً. فعلى سبيل المثال، افترض أننا مهتمين بأسعار البيع Y للمنازل ذات الأحجام التى تتراوح بين 1,500 إلى 2,500 قدم مربع. فإننا يمكن أن نختار الأحجام 1,500 و 2,000 و 2,500 ثم نختار عينة عشوائية من المنازل لكل فئة، مع إهمال المنازل ذات الأحجام التى تخرج عن هذا المدى. وإختيار مدى لقيم X مرغوب جداً ليس فقط لأنه يسمح لنا بمشاهدة قيم Y داخل المدى الذي نهتم به لـ X، ولكن لأنه أيضاً يوفر فرصة لزيادة درجة الإعتماد على الإستنتاجات المتعلقة بعملية التحليل، وهذا ما سوف نناقشه فى الجزء (٩-٦) .

3 – البيانات الملائمة **Convenience Data** : في بعض الأحيان لا يكون ممكناً إجراء معاينة عشوائية؛ هنا يمكن أن نحصل على البيانات التى حدثت فعلا، أي المتاحة. ففى مثال مثمن العقارات ، نلاحظ أن البيانات المتاحة ممثلة فى المنازل التى بيعت بالفعل. وأن هذه البيانات تحددت بواسطة ملاك المنازل وليس بواسطة الإختيار العشوائى .

ولعمل تحليل انحدار، فلابد أن نفترض أن «اختيار ملاك المنازل» يمثل المجتمع بالنسبة للعلاقة بين سعر البيع وحجم المنزل. وفي بعض الحالات، يكون من الخطأ عمل هذا الافتراض. فإذا إفترضنا أن معظم المنازل في البيانات الملائمة قديمة. وبالتالي فإن النموذج الذي تم التوصل إليه من هذه البيانات لن يعكس العلاقة الحقيقية بين سعر البيع وحجم المنازل لكل المنازل في المجاورة.

وفى جميع الحالات السابقة يفترض أن قيم X مقاسة بدون أخطاء، ويكون تحليل الإنحدار مناسباً للاوضاع الثلاث. وهناك نقطتين هامتين يجب أن نفهمهما: (1) بتحديدك المسبق لمدى X ، تكون قادراً على تحسين الإستنتاج (وهذا موضح فى الجزء (٩-٦))، (2) إذا رغبت فى إستخدام البيانات الملائمة، فيجب عليك أن تقرر أولاً ما إذا كانت البيانات تمثل البيئة التى ترغب فى صنع إستنتاجات عنها بكفاية أم لا . فإذا كانت الإجابة لا، فإن الإستنتاجات المبنية على تحليل الإنحدار يمكن أن تكون بها أخطاء بدرجة ملموسة .

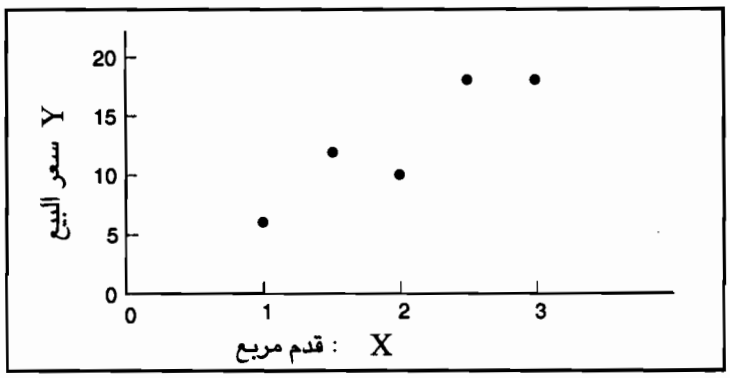
(٩-٣-٢) طريقة المربعات الصغرى :

بالعودة إلى المعادلة (9.2)، كيف يمكن تقدير β_0 , β_1 أيضاً σ_ε^2 من بيانات العينة؟ بفرض أننا حصلنا على البيانات التالية من عينة من خمس منازل وبفرض أن Y تشير إلى سعر البيع بالعشرة آلاف دولار بمعنى أن Y=10 تعنى أن سعر البيع 100,000 دولار، كما أن X تشير إلى المساحة لأقرب الف قدم مربع. البيانات هنا مقربة لرقم صحيح لتبسيط العرض.

X	Y
1	6
1.5	12
2	10
2.5	18
3	18

الوسط $\overline{Y} = 12.8$ ؛ $\overline{X} = 2$

هل تُظهر هذه البيانات وجود علاقة بين سعر البيع (Y) والحجم (X) فى المجتمع؟ عادة يستخدم الشكل الإنتشارى كأداة بسيطة لإظهار هذه العلاقة. ويلاحظ من الشكل الإنتشارى فى (٩-١٠)، أن سعر البيع يميل إلى الإرتفاع بزيادة الحجم ولكن هذه العلاقة غير تامة، فعندما إنخفضت Y من 12 إلى 10 ارتفعت X من 1.5 إلى 2 ، أيضاً ارتفعت قيمة X من 2.5 إلى 3 بينما ظلت قيمة Y ثابتة. وهذا يعنى وجود عدة عوامل أخرى تؤثر على سعر بيع المنزل بالإضافة إلى الحجم .

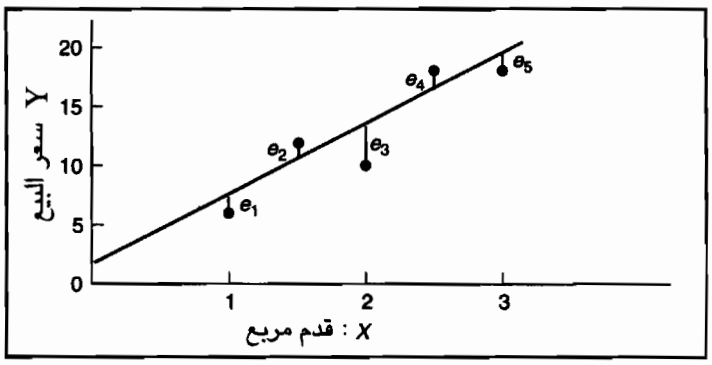


شكل (٩-١٠)

الشكل الانتشاري للبيانات السابقة

ولتقدير خط الإنحدار فى المجتمع برسم خط مستقيم يمر خلال بيانات العينة، فإن ذلك يتم بالعين المجردة وفق التقدير الشخصي. ولكن للحصول على أفضل خط، فإن اجراء معين يجب أن يستخدم. أولاً، يجب أن نحدد ماذا نقصد باللفظ «أحسن خط» «**the best line**» . يلاحظ من الشكل (٩-١٠) أننا لا يمكن أن نرسم خط مستقيم يمر بجميع النقاط الخمسة. والآن نرغب فى رسم خط واحد فقط، بحيث تكون المسافة الرأسية من النقط إلى الخط أقل ما يمكن. هذه المسافات الرأسية تمثل أخطاء تمثل أخطاء العينة التي يجب أن تحدث لو أن هذا الخط استخدم لتقدير متوسط Y عند كل قيمة لـ X في العينة. أخطاء العينة هذه تسمى بالبواقي **residuals** ويرمز لها بالرمز e . البواقي في هذا المثال موضحة فى الشكل (٩-١١) .

ولأنه توجد خمسة بواقى، فيجب تلخيصها بطريقة ما، وإحدى هذه الطرق هى إختيار الخط الذي يصغر مجموع البواقي، ولكن هذه الطريقة غير عملية، لأنه يوجد أكثر من خط مجموع البواقى لكل منهم يساوى الصفر. والبعض لايلائم البيانات مطلقا. على سبيل المثال، مجموع البواقى تساوى الصفر عند الخط $Y = \overline{Y}$.



شكل (٩-١١) البواقي في مثال تقييم الأصول

وحتى إذا كانت البواقى لهذا الخط كبيرة، فإن البواقى الموجبة تلاشى البواقى السالبة والمحصلة تكون صفر. وبدلاً من ذلك، فإن الطريقة المفضلة هي تصغير مجموع مربعات البواقى، والخط الذى يجعل مجموع مربعات البواقى $(\sum e_i^2)$ أقل ما يمكن يسمى بخط المربعات الصغرى أو خط الإنحدار المقدر **least squares line or the estimated regression line**. ومجموع مربعات البواقى لهذا الخط أقل من أى خط آخر. والأسلوب المستخدم فى إيجاد خط المربعات الصغرى يسمى، طريقة المربعات الصغرى **Method of Least Squares** .

ولتحديد خط المربعات الصغرى، يجب تقدير قيم β_0 , β_1 . هذه التقديرات تشير إلى الجزء المقطوع من محور Y، والميل، على الترتيب، وتستخدم معادلات المربعات الصغرى فى تحديدهما.

الأحصاءات b_0 ، b_1 تعرف على أنها تقديرات المربعات الصغرى للمعالم β_0 ، β_1 على التوالي، ويتم تحديدها باستخدام علم التفاضل. وتفاصيل ذلك هى خارج نطاق عرض هذا الكتاب.

$$b_1 = \frac{SP(XY)}{SS(X)} \qquad (9.3)$$

$$b_0 = \overline{Y} - b_1\overline{X} \qquad (9.4)$$

حيث :

$$SP(XY) = \sum(X_i - \overline{X})(Y_i - \overline{Y})$$

$$= \sum(X_iY_i - \frac{(\sum X_i)(\sum Y_i)}{n} \qquad (9.5)$$

و :

$$SS(X) = \sum(X_i - \overline{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} \qquad (9.6)$$

وفى معظم الأحيان، يستخدم الحاسب لتحديد تقديرات المربعات الصغرى. اذا استخدمت الآلة الحاسبة الشخصية، لاحظ أن الشق الثاني من المعادلات (9.5)، (9.6) تبسط العمليات الحسابية بشكل ملحوظ. وتكون القيم SP(XY)، SS(X) عبارة عن القيمة المبدئية والتي تستخدم لحساب المقدرات b_0 ، b_1. وسوف تظهر هذه القيم في الكثير من المعادلات التي سوف تظهر في هذا الفصل. والقيمة

SS(X)، والتي تعبر عن مجموع مربعات X، تظهر في البسط في التعبير الحسابي الذي نستخدمه لحساب تباين العينة للمتغير X. وهي ببساطة تقيس الأختلاف الكلي **total variation** للمتغير X. أما الكمية SP(XY) فهو عبارة عن كمية لم تتعرض لها حتى الآن وهى تعبر عن مجموع حاصل ضرب X، Y. وتعرف بأنها تغايرات «**Covariation**»- أي الدرجة التي تتأثر بها الاختلافات أو الانحرافات في قيم Y نتيجة الاختلاف أو الانحراف في قيم X. وبإستخدام المثال السابق، سوف نوضح حساب تقديرات المربعات الصغرى، الجزء الثابت والميل بإستخدام هذه الطريقة المتدرجة في الحساب.

X	Y	XY	X^2
1	6	6	1
1.5	12	18	2.25
2	10	20	4
2.5	18	45	6.25
3	18	54	9

$$\sum X_i = 10 \qquad \sum Y_i = 64 \qquad \sum X_i Y_i = 143 \qquad \sum X_i^2 = 22.5$$

$$\overline{Y} = \frac{64}{5} = 12.8 \qquad , \qquad \overline{X} = \frac{10}{5} = 2$$

مجموع حاصل الضرب:

$$SP(XY) = 143 - \frac{(10)(64)}{5} = 15$$

مجموع مربعات X

$$SS(X) = 22.5 - \frac{(10)}{5} = 2.5$$

بالتعويض بهذه القيم فى المعادلات (9.3) ، (9.4) يمكن إيجاد الجزء الثابت والميل (تقديرات المربعات الصغرى) .

$$b_1 = \frac{15}{2.5} = 6 \qquad , \qquad b_0 = 12.8 - 6(2) = .8$$

وعلى ذلك ، يكون خط الإنحدار المقدر (خط المربعات الصغرى) هو :

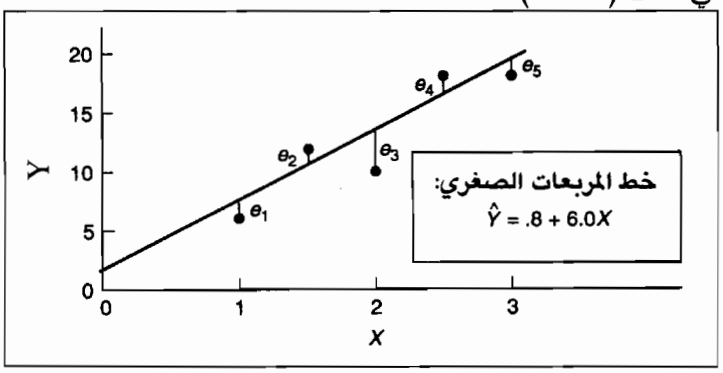
$$\hat{Y} = .8 + 6X$$

تذكر أن خط المربعات الصغرى، يجعل مجموع مربعات البواقى أقل ما يمكن. بفرض أننا نرغب فى تحديد مجموع مربعات البواقى فى المثال السابق. مبدئياً تذكر أن البواقى عند كل نقطة فى العينة هى الفرق بين القيمة الفعلية Y والقيمة المتنبأ بها لـ Y عن طريق خط المربعات الصغري.

أى أن $e_i = Y_i - \hat{Y}$. الحسابات اللازمة لتقدير البواقي كما يلي :

X	الفعلية Y	القيمة التقديرية $\hat{Y} = .8 + 6X$	البواقي $e = Y - \hat{Y}$	مربعات البواقي $e^2 = (Y - \hat{Y})^2$
1.0	6	6.8	-.8	.64
1.5	12	9.8	2.2	4.84
2.0	10	12.8	-2.8	7.84
2.0	18	15.8	2.2	4.84
3.0	18	18.8	-.8	.64
			$\sum(Y_i - \hat{Y}_i) = 0$	$\sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2 = 18.8$

ومجموع مربعات البواقى (إختصاراً SSE) الناتجة من خط المربعات الصغرى $\hat{Y} = .8 + 6X$ هى (SSE = 18.8) ومجموع مربعات البواقى ستكون أكبر لأى خط آخر لهذه البيانات. لاحظ أيضاً أن مجموع البواقى يساوى صفر، وهذا يعنى أن النقط أدني خط المربعات الصغرى تعادل النقط التي تقع أعلى خط المربعات الصغرى من حيث طول الأبعاد الرأسية. خط المربعات الصغرى والبواقي موضحة في شكل (٩-١٢).



شكل (٩-١٢) خط المربعات الصغرى والبواقي

(٩-٣-٣) تقدير تباين الخطأ σ_ε^2

لقد تعلمنا كيفية تقدير β_0 , β_1 ؛ الميل والجزء الثابت على الترتيب في خط إنحدار المجتمع. وتحتاج أيضاً إلى معرفة كيف يتم تقدير تباين الخطأ σ_ε^2 . والذي يصف درجة اقتراب توفيق خط الانحدار للبيانات، ولاداء ذلك نقدم المقدر S_e^2 للمعلمة σ_ε^2 بأسلوب مماثل عندما قدمنا تقدير تباين المجتمع في الفصل الثاني.

وبفرض أن W_1 , W_2 .... W_n على الترتيب هى مشاهدات عينة عشوائية. (استخدمنا الرمز W بدلا من X أو Y منعا للتداخل أو الارتباك). فإن صيغة التباين تكون :

$$S^2 = \frac{\sum(W_i - \overline{W})^2}{n-1}$$

حيث البسط هو مجموع مربعات الانحرافات بين قيم W ووسطها من العينة $\overline{W}$ ولأن المقام (n-1) فإن S^2 تصبح تقدير غير متحيز للمقدار σ^2 . ولأننا استخدمنا $\overline{W}$ كتقدير لمتوسط المجتمع غير المعلوم، فإنه تم طرح واحد من حجم العينة فى المقام لكى تصبح S^2 تقدير غير متحيز، واذا كان لدينا معلمتين غير معلومتين، عندئذ نطرح 2 من حجم العينة n فى المقام ، وهكذا .

الصيغة S_e^2 مثل الصيغة S^2 . تذكر أن خط الإنحدار فى المجتمع غير معلوم. ومعنى ذلك أن S_e^2 يقيس الإختلافات فى مشاهدات العينة (قيم Y) عن خط الإنحدار المقدر (قيم $\hat{Y}$) . وهذه الإختلافات تمثل البواقى $e_i = Y_i - \hat{Y}_i$ وطبقاً لذلك فإن بسط S_e^2 هو مجموع المربعات البواقى $SSE = \sum(Y_i - \hat{Y}_i)^2$. وماذا عن المقام ؟ لتحديد قيم $\hat{Y}$ فإننا قدرنا معلمتين β_0 , β_1 ، ولكى نجعل S_e^2 تقدير غير متحيز للمعلمة σ_ε^2 ، يجب أن نطرح 2 من حجم العينة n وعلى ذلك فإن المقام يصبح (n-2) ويمكن تعريف S_e^2 كالآتى:

$$S_e^2 = \frac{SSE}{n-2} \qquad \textbf{(9.7)}$$

where: $$SSE = \sum(Y_i - \hat{Y}_i)^2 \qquad \textbf{(9.8)}$$

هي مجموع مربعات البواقي (خطأ العينة). وبالتالي، يصبح تقدير خطأ الإنحراف المعيارى σ_ε على الصورة:

$$S_e = \sqrt{S_e^2} \qquad (9.9)$$

ومن الأهمية بمكان فهم ماذا يقيس S_e^2. S_e^2 تقيس إلى أى مدى يكون خط المربعات الصغرى ملائم لقيم Y فى العينة. فإذا كان الخط ملائماً تماماً، فإن كل البواقي يجب أن تساوى صفر وبالتالي تصبح S_e^2 تساوى صفر. كبر البواقي يعني كبر S_e^2 وإنخفاض درجة الملائمة، ويسمى الإحصاء S_e^2 بتباين البواقى أو متوسط مربعات الخطأ **residual variance or mean square for error (MSE)** وكذلك يسمى الإحصاء S_e الإنحراف المعيارى للبواقى أو الجذر التربيعى لمتوسط مربع الخطأ **residual standard deviation** أو **root mean square for error (RMSE)**.

ويمكن حساب S_e^2 ، S_e من المثال السابق كالتالى : بما أن SSE = 18.8

فإن :

$$S_e^2 = \frac{18.8}{5-2} = 6.2667$$

$$, \; S_e = \sqrt{6.2267} = 2.5033$$

(٩-٣-٤) معامل التحديد : تجزئة الإختلاف الكلى :

الهدف الأساسى من تحليل الإنحدار هو إيجاد نموذج يلائم البيانات ويتسق نظرياً، بقدر الامكان مع المعلومات غير الإحصائية. ولكى نتأكد من أن نموذج الخط المستقيم يلائم البيانات، يجب أن نناقش تباين البواقى S_e^2 أو MSE كمقياس لدرجة الملائمة. وبشكل موضوعى فإن أفضل توفيق يجعل S_e^2 أقل ما يمكن يأتى بإضافة متغيرات تفسيرية أخرى كما سنرى ذلك في الفصل العاشر. ولأن قيمة S_e^2 تعتمد على كيفية قياس قيم Y فى العينة، لذلك فإننا لا يمكن أن نعلق على قيمة S_e^2 إلا إذا أخذنا فى الإعتبار كيفية قياس قيم Y . لذلك فإننا سوف نتحدث عن مقياس آخر لا يأخذ فى الإعتبار كيفية قياس قيم Y فى العينة، ألا وهو معامل التحديد ويرمز له بالرمز r^2 .

ولتعريف r^2 ، يمكن أن نستفيد من أسلوب تحليل التباين الذي قدم في الفصل الثامن. ففى تحليل التباين، كانت الفكرة الأساسية هى تقسيم مجموع الإختلافات فى العينة(SST) إلى قسمين :

1- الإختلافات بسبب الفروق بين متوسطات المعالجات (SSTR) .

2 - الإختلافات بسبب العوامل العشوائية (SSE) .

حيث : SST = SSTR + SSE

هذا الأسلوب يمكن تطبيقه فى تحليل الإنحدار بنفس الصورة. لاحظ، أنه إذا كانت Y فى المجتمع تأخذ قيم ثابتة، فإنه يمكن التنبؤ بها بدقة تامة، لكن تحدث مشكلة فى التنبؤ عندما تختلف قيم Y داخل المجتمع .

بفرض أنه تم تحديد خط إنحدار المربعات الصغرى من بيانات عينة، السؤال الآن هو: لماذا تختلف قيم Y فى العينة ؟ يرجع ذلك لسببين محتملين :

1 - بسبب العلاقة الإنحدارية مع X . نحن نعلم أن المتغير Y مرتبط مع المتغير X، وقيم X فى العينة متغيرة. إذن مع كل تغير فى قيم X تتغير قيم Y ، فمثلاً أسعار المنازل الكبيرة تميل إلى الإرتفاع أكثر من المنازل الأقل مساحة، فى ظل ثبات باقى العوامل .

2- بسبب العوامل العشوائية، هناك عوامل أخرى تؤثر فى قيم Y أيضاً. ويفترض أنها تتغير عشوائياً، فمثلاً قد يباع منزلين لهما نفس المساحة بأسعار مختلفة، لإختلاف الموقع، وعدد المطابخ وعوامل أخرى .

الآن، كيف نقيس هذين النوعين من الاختلافات؟ في الفصل الثاني، قدمنا طريقة تحديد الاختلاف الكلي لأي متغير، ثم وضحت بعمق في الفصل الثامن، واستخدمت من قبل في هذا الفصل عند حساب الاختلافات الكلية في قيم X [الصيغة (9.6)]. وعلى ذلك، فإن الاختلاف الكلي في قيم Y في العينة يمكن تحديده بالصيغة (9.10):

$$SS(Y) = \sum(Y_i - \overline{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i^2)}{n} \qquad \textbf{(9.10)}$$

لاحظ أن هذا التعبير يناظر SST في تحليل التباين فى الفصل الثامن وسوف نستخدم الرمز SST للإشارة إلى هذا الإختلاف . إذن SS (Y) , SST متساويان ويشيران إلى نفس الكمية . والإختلاف الراجع إلى العوامل العشوائية هو ببساطة الإختلاف غير المفسر فى قيم Y من خط المربعات الصغرى، وهو مجموع مربعات البواقى ونشير إليه بالرمز SSE . والفرق بين الاختلاف الكلي SST والاختلاف غير المفسر SSE بنموذج الإنحدار، يسمى مجموع مربعات الإنحدار **regression sum of squares** ويرمز إليه بالرمز SSR ، وعليه : SSR = SST - SSE ، لذلك فإن مجموع مربعات الإنحدار تشير إلى الإختلافات فى قيم Y في العينة والتى يمكن تفسيرها فى ضوء اختلافات قيم X في العينة. ويمكن تقسيم الإختلافات الكلية مثلما ظهرت فى الفصل الثامن كالآتى :

$$SST = SSR + SSE$$

مجموع المربعات الكلى = مجموع مربعات الإنحدار + مجموع مربعات الخطأ

ويمكن حساب مجموع المربعات الكلى لقيم Y فى المثال السابق كالآتى :

$$SST = SS(Y) = (6^2 + 12^2 + \ldots\ldots\ldots 18^2) - \frac{(6+12+\ldots\ldots+18)^2}{5} = 108.8$$

وسبق أن حددنا مجموع مربعات البواقى ، (SSE = 18.8) ، إذن يمكن إيجاد مجموع مربعات الإنحدار .

$$SSR = SST - SSE = 108.8 - 18.8 = 90$$

ودرجة ملائمة معادلة الإنحدار المقدرة تتحدد بمقارنة مجموع مربعات الإنحدار SSR بمجموع المربعات الكلى SST . ومعامل التحديد (r^2) **The Coefficient of determination** هو نسبة مجموع مربعات الإنحدار إلى مجموع المربعات الكلى ويعطى بالعلاقة التالية :

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} \qquad \textbf{(9.11)}$$

ومن المثال السابق ، توفر لدينا: (SST = 108.8) , (SSE = 18.8) , (SSR = 90) بالتالي يكون معامل التحديد:

$$\therefore\ r^2 = \frac{90}{108.8} = \frac{108.8 - 18.8}{108.8} = .8272$$

معنى معامل التحديد (r^2) :

معامل التحديد ما هو إلا وصف إحصائي يوضح نسبة التغير الكلى فى قيم Y فى العينة والتى يمكن تفسيرها فى ضوء علاقة خط الإنحدار مع التغير فى X وهذا منطقيا لأنه كلما زادت الاختلافات بين قيم Y، كلما صعبت عملية التنبؤ. وعموماً فإنه إلى الحد الذي نستطيع شرح الأختلاف الأساسي أو المبدئي في قيم Y عن طريق معادلة الانحدار، نكون قادرين على التنبؤ بقيمة Y. ومعامل التحديد هو إحصاء محرر من وحدات القياس يأخذ أى قيمة داخل المدى (صفر ، 1). فإذا كان (r^2) قريب من الصفر، معنى ذلك أن معادلة خط الإنحدار المقدرة تفسر القليل من الاختلافات في قيم Y. وإذا كان (r^2) قريب من الواحد الصحيح دل ذلك على أن معادلة خط الإنحدار المقدرة يمكنها تفسير جزء كبير من إجمالى الإختلاف فى المتغير Y. وكلما زادت قيمة r^2، كلما كانت معادلة خط الانحدار المقدرة أكثر ملائمة وأكثر فاعلية في التنبؤ بـ Y. وحيث أن r^2 مقياس محرر، فغالبا ما يساء تفسيره. يجب أن تعلم أن r^2 لا تعنى قياس. ولا تقيس (r^2) مدى صحة نموذج الإنحدار المقدر، بمعنى قيمة r^2 لايمكن أن تشير إلى أن معادلة إنحدار Y على X في المجتمع، هي علاقة خط مستقيم تام، أي أنه يقيس فقط كيف أن الأختلافات الكلية في قيم Y في العينة، يمكن تفسيرها بمعادلة الانحدار المقدر (خط المربعات الصغري). وقد يكون من الشائع أن تكتشف أن معامل التحديد عند استخدام خط المربعات الصغرى كبير (إفترض أن $r^2 = 90$) ورغم هذا قد نجد أن نموذج آخر يكون أكثر صلاحية لتوفيق البيانات.

تفسير آخر لمعامل التحديد (r^2) :

فى الجزء (٩-١) ، وضحنا أن تحليل الإنحدار يقدم الوسائل التى من خلالها يمكن إدخال المعلومات الإضافية (المتغير X) إلى التحليل . ويمكن أيضاً رؤية معامل التحديد على أنه مؤشر إحصائي وصفى **descriptive statistic** يوضح كيفية أن إدخال المتغير X يساعد كثيراً فى التنبؤ بـ Y. هذه الفكرة مباشرة ومبنية على أساس التفكير التالى :

1 – نقيس أولاً خطأ التنبؤ الكلى **the total prediction error** الذى سوف يحدث ، إذا لم يتم أخذ X فى الإعتبار، فإن أفضل تنبؤ لقيم Y سوف يكون ببساطة متوسط العينة $\overline{Y}$. فى هذه الحالة نقيس خطأ التنبؤ الكلى على أنه مجموع مربعات الإنحرافات بين قيم Y الفعلية ، $\overline{Y}$ أى تكون : $\Sigma(Y_i - \overline{Y})^2$. لاحظ أن هذه الكمية هي SST

2 – نقيس بعد ذلك خطأ التنبؤ الكلى الذى يحدث عندما تستخدم X للتنبؤ (أى إستخدام خط إنحدار المربعات الصغرى) . وهذا يعني مجموع مربعات الإنحرافات بين قيم Y وتنبؤات الإنحدار لها أى $\Sigma(Y_i - \hat{Y}_i)^2$ وهذا يعطي مجموع مربعات البواقى، ويرمز لها بالرمز SSE .

3 – الإنخفاض فى خطأ التنبؤ الكلى كنتيجة لإستخدام X هو : (SSR = SST - SSE) ، أي الفرق بين خطأ التنبؤ الكلى بدون إستخدام X ، (SST) وخطأ التنبؤ الكلى عندما يتم إستخدام X، (SSE) .

4 – معامل التحديد ($r^2 = SSR / SST$) ، يمكن رؤيته على أنه ذلك الجزء من خطأ التنبؤ الكلى الذى يتم التخلص منه عن طريق إستخدام X .

وفى مثال تثمين العقارات ، وجدنا أن ($r^2 = .8272$) . هذا يعنى أن 82.72 % من الإختلافات فى قيم Y فى العينة، تم تفسيرها عن طريق العلاقة الخطية المقدرة مع X . وبمعنى آخر، يمكننا القول أن

إستخدام خط المربعات الصغرى للتنبؤ يؤدى إلى تخفيض 82.72 % من خطأ التنبؤ الكلى والذى كان سيحدث إذا تم إستخدام متوسط العينة فقط للتنبؤ .

مثال (٩-٤)

يرغب مدير مطعم عش البلبل Bird's Nest فى الحصول على نموذج يوضح إلى أى مدى يكون إيراد الفترة المسائية مرتبط بعدد «العملاء أو الزبائن Covers» أى عدد الأشخاص الذى يطلبون وجبة. البيانات التالية لست فترات مسائية حديثة (الإيرادات معبر عنها بمئات الدولارات) :

الفترة المسائية	عدد الوافدين (العملاء)	الإيراد
1	15	5
2	20	8
3	50	12
4	30	10
5	25	9
6	40	13

(a) حدد خط إنحدار المربعات الصغرى للتنبؤ بدخل الفترة المسائية بمعلومية عدد الوافدين، وفسر تقديرات الميل slope والجزء المقطوع intercept .

(b) إلى أى مدى لقيم X تكون تفسيراتك فى جزء (a) صحيحة؟ لماذا تكون محدودة في هذا المدى؟

(c) كون شكل الإنتشار وارسم عليه خط المربعات الصغرى. هل يبدو أن النموذج الخطى مناسب ؟

(d) قدر تباين الخطأ وإحسب معامل التحديد . إستخدم هذه المعلومات لوصف مدى ملائمة نموذج المربعات الصغرى .

الحل

(a) المتغير التابع هنا هو Y ويمثل دخل الفترة المسائية. المتغير المفسر هو X ، ويمثل عدد الوافدين للفترة المسائية. نستخدم Minitanb لتحديد الكميات التالية :

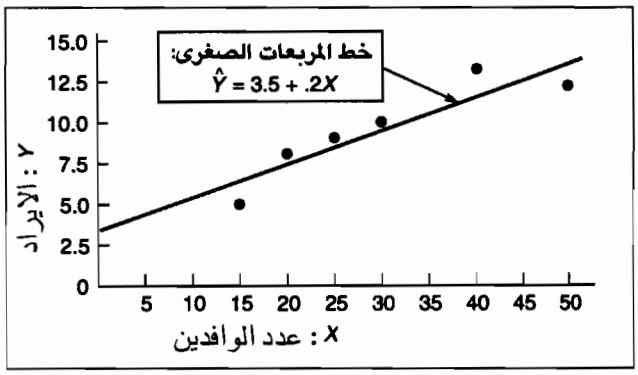
$$b_0 = 3.5 \; ; \; b_1 = .2 \; ; \; SST = 41.5 \; ; \; SSR = 34.0 \; ; \; SSE = 7.5 \; ; \; SS(X) = 850$$

حيث أن تقديرات المربعات الصغرى هى $(b_0 = 3.5)$, $(b_1 = .2)$ ، فإن خط المربعات الصغرى يكون:

$$\hat{Y} = 3.5 + .2X \qquad \text{for } X \text{ in } (15,50)$$

تقدير الميل هو $b_1 = .2$ ؛ هذا يعنى أن لكل وافد إضافى، متوسط الإيراد المقدر يزيد بمقدار 0.2 (أو 20$) . وحيث أن البيانات الممثلة لا تشتمل على معلومات عندما تكون (X=0) ، فإن تفسير تقدير الجزء المقطوع $(b_0 = 3.5)$ يكون غير ذى معنى. فى الواقع ، نعلم أنه لن يكون هناك أى إيراد من الوجبات إذا كانت $X = 0$ أى اذا كان لا يوجد أى عميل .

(b) كما أشرنا فى إجابة جزء (a) ، تفسير النموذج يكون صحيح إحصائياً فقط للأيام التى يكون فيها عدد الوافدين بين 15 , 50 . وحيث انه لا يوجد لدينا بيانات خارج هذا المدى؛ لذلك، لا تستطيع أن نستنتج إحصائياً أن نفس العلاقة الخطية السابقة تكون مناسبة عندما يكون عدد الوافدين أقل من 15 أو يزيد عن 50 .



شكل (٩-١٣)
شكل الإنتشار وخط المربعات الصغرى للإيراد Y مقابل عدد الوافدين X

(c) شكل الإنتشار وخط المربعات الصغرى معطى فى شكل (٩-١٣) ، بناء على هذا الشكل، التوفيق الخطى **linear fit** يبدو معقول. ومع ذلك يكون من الصعب قول هذا بثقة عندما يكون حجم العينة صغير جداً .

(d) حيث أن (SSE = 7.5) , (n=6) يكون تباين البواقى :

$$S_e^2 = \frac{SSR}{n-2} = \frac{7.5}{4} = 1.875$$

والإنحراف المعيارى للبواقى يكون :

$$S_e = \sqrt{1.875} = 1.3693$$

ويكون معامل التحديد للعينة عند معلومية: (SSR = 34.0) , (SST = 41.5) :

$$r^2 = \frac{34}{41.5} = .8193$$

التفسير المناسب لذلك هو ، أن 81.93% من الاختلافات اليومية في إيرادات العينة، تفسر عن طريق الإختلافات اليومية فى عدد الوافدين. بناء على ذلك ، يمكننا القول أن (1-.8193 = .1807) أو من 18.07% من الاختلافات اليومية فى عينة الإيرادات لا تفسر عن طريق الإختلاف اليومى فى الوافدين ولكن يمكن إعتبارها خطأ عشوائى. وإذا اعتقدت الإدارة أن التوفيق مناسب، يمكن زيادة حجم العينة، أو يمكن إدخال متغيرات تنبؤية أي تفسيرية **Predictor Variables** إضافية فى محاولة لحساب بعض الإختلاف غير المفسر هذا .

تمارين

(٩-١٥) إعتبر بيانات العينة التالية :

X	2	6	8	10	15
Y	50	35	30	44	20

(a) كون شكل الإنتشار .

(b) ارسم خط مستقيم على شكل الإنتشار السابق، بحيث يكون أفضل تمثيل للعلاقة الخطية بين X , Y من وجهة نظرك.

(c) عين البواقى الخمسة (الإنحرافات عن قيم Y فى العينة من الخط الذي تم رسمه في (كا) على شكل الإنتشار .

(d) حدد قيم البواقى الخمسة عن طريق قياسهم بالمسطرة .

(e) احسب مجموع مربعات البواقى من جزء (d) .

(٩-١٦) بإستخدام بيانات التمرين (٩-١٥) أجب على ما يلى :

(a) حدد خط المربعات الصغرى وفسر تقديرات الميل والجزء المقطوع .

(b) ارسم خط إنحدار المربعات الصغرى من جزء (a) على أن يوقع فى شكل الإنتشار السابق .

(c) عين البواقى الخمسة على شكل الإنتشار .

(d) حدد قيم البواقى الخمسة عن طريق طرح قيم Y التى تم التنبؤ بها من معادلة المربعات الصغرى من القيم الفعلية المناظرة .

(e) حدد مجموع مربعات البواقى من جزء (d) . قارن هذا المجموع بمجموع مربعات البواقى الذى حصلت عليه فى جزء (e) تمرين (٩-١٥) . أيهما أقل ؟ على أى شئ يدل هذا بخصوص درجة التوفيق النسبي relative fits للخطين ؟

(٩-١٧) بإستخدام بيانات التمرين (٩-١٦) :

(a) احسب تباين البواقى S_e^2 وفسر معناه .

(b) احسب SSR , SSE , SST . وضح أى جانب للبيانات يوصف عن طريق كل من هذه الكميات .

(c) احسب معامل التحديد r^2 وفسر معناه .

(٩-١٨) إفترض أنك حصلت على الحسابات التالية فى توفيق خط مستقيم لعينة من البيانات:

$$n = 15 \;;\quad \Sigma Y_i = 1{,}133 \;;\quad \Sigma X_i = 23 \;;\quad SP(XY) = -1{,}535.286 \;;$$
$$SS(X) = 863.7333 \;;\quad \Sigma Y_i^2 = 89.091 \;;\quad SSE = 782.8288$$

(a) حدد خط المربعات الصغرى وفسر نتيجة تقديرات الميل والجزء المقطوع لخط أنحدار المجتمع .

(b) احسب تباين الباقى S_e^2 وفسر معناه .

(c) احسب معامل التحديد r^2 وفسر معناه .

(٩-١٩) إفترض أن نتائج تحليل الإنحدار فى الحسابات التالية :

$$n = 31 \;;\quad \Sigma Y_i = 4{,}212 \;;\quad \Sigma X_i = 2{,}856 \;;\quad \Sigma X_i Y_i = 391{,}442.01 \;;$$
$$\Sigma Y_i^2 = 582{,}904.36 \;;\quad \Sigma X_i^2 = 264.770 \;;\quad SSE = 3{,}630.1392$$

أجب على نفس أسئلة تمرين (٩-١٨) .

(٩-٢٠) اعتبر بيانات العينة التالية :

X	1	2	3	4	5	6
Y	2	4	4	6	9	10

(a) كون شكل إنتشار لهذه البيانات . هل تبدو العلاقة الخطية مقبولة ؟

(b) افترض أن التوفيق الخطى مناسباً، حدد خط المربعات الصغرى وفسر ميله والجزء المقطوع ؟

(c) احسب تباين الباقى S_e^2 ومعامل التحديد r^2 وفسر معنى كل منهم ؟

(٩-٢١) إعتبر بيانات العينة التالية :

X	2	2	4
Y	4	6	11

(a) كون شكل إنتشار لهذه البيانات .

(b) مفترضاً أن العلاقة الخطية مناسبة، حدد خط المربعات الصغرى .

(c) ارسم خط المربعات الصغرى وكل من الخطوط الأخرى التالية على شكل الإنتشار من جزء (a) : $(\hat{Y} = 1 + 2X)$, $(\hat{Y} = 7)$. أى هذه الخطوط يعد الأفضل في تمثيل قيم عينة Y ؟ لماذا ؟

(d) لكل خط فى جزء (c) ، حدد البواقى .

(e) لكل خط فى جزء (c) ، احسب مجموع البواقى. بماذا تدل نتائجك بخصوص فائدة مجموع البواقى كمؤشر لدرجة توفيق الخط ؟

(f) لكل خط فى جزء (c) احسب مجموع مربعات البواقى. لأى خط يكون SSE الأصغر ؟

(٩-٢٢) اعتبر بيانات العينة التالية :

X	3	5	5	7	9	9
Y	6	2	1	-1	-4	-8

(a) كون شكل إنتشار لهذه البيانات . هل العلاقة الخطية تبدو مقبولة ؟ .

(b) إفترض أن التوفيق الخطى مناسب، حدد خط المربعات الصغرى وفسر ميله والجزء المقطوع .

(c) احسب تباين البواقى S_e^2 ومعامل التحديد r^2 وفسر معناهم .

(٩-٢٣) بالاشارة إلى تمرين (٩-١٢).

(a) إفترض أن العلاقة الخطية مناسبة، حدد خط المربعات الصغرى وفسر التقديرات الناتجة للميل والجزء المقطوع لخط إنحدار المجتمع .

(b) احسب تباين الباقى S_e^2 ومعامل التحديد r^2 وفسر نتائجك .

(٩-٢٤) بالاشارة إلى التمارين (٩-١٥) ، (٩-٢٠) ، (٩-٢٢) ، (٩-٢٣) :

(a) قارن تباينات البواقى لخط إنحدار المربعات الصغرى الذى حددته فى هذه التمارين. بناء على هذه المقارنة فقط ، هل يمكنك تحديد أى خط ، يعتبر أكثر ملائمة لبيانات العينة؟

(b) بناء على معامل التحديد الذى حسبته فى هذه التمارين ، هل يمكنك تحديد الخط الذى يوفق أفضل قيم العينة Y ؟ أى خط مربعات صغرى يقدم أسوأ توفيق ؟ وضح إجابتك .

(c) هل إجابتك لجزء (b) تتفق مع تفسيراتك لشكل الإنتشار لهذه التمرينات ؟ وضح .

(٩-٢٥) بالاشارة إلى تمرين(٩-١٤) . افترض أن النموذج الخطى مناسباً وأن خط المربعات الصغرى تم تحديده ليكون $\hat{Y} = 12.0 + .116X$.

(a) فسر تقديرات الميل والجزء المقطوع لهذه المشكلة .

(b) حدد البواقى واستخدمها لتحديد تباين البواقى .

(c) احسب معامل التحديد r^2 . فسر المعنى لنتيجتك .

(d) بناء على إجابتك لجزء (c) فقط ، هل يمكنك إستنتاج أن خط المربعات الصغرى مناسب (ملائم) للتقدير والتنبؤ؟ وضح .

(٩-٤) الإستنتاجات الإحصائية المتعلقة بنموذج الإنحدار الخطى البسيط :

Statistical Inferences For The Simple Linear Regression Model

ناقشنا فيما سبق تباين البواقى S_e^2 ومعامل التحديد r^2 كمقاييس لمدى توفيق معادلة الإنحدار المقدرة (خط المربعات الصغرى) لقيم Y للعينة الممثلة. وناقشنا أيضاً الإستخدامات المتنوعة لنموذج الإنحدار. ومع ذلك ، وقبل إستخدام هذا النموذج ، من المهم التأكد إذا كان النموذج ملائم للإستخدام المطلوب. وتقييم أداء النموذج يتضمن مناقشة القضايا التالية :

(1) هل بيانات العينة تدل على وجود فعلى لإرتباط خطى بين Y ، X فى المجتمع؟ من الممكن أن لا يكون هناك إرتباط على الإطلاق أو أن الإرتباط غير خطى. إذا كان ذلك ، فإن العلاقة المقترحة عن طريق خط المربعات الصغرى يمكن ببساطة أن تكون نتيجة لتغيرات المعاينة العشوائية **random sampling variability** . شكل الإنتشار يقدم فكرة أولية لهذا السؤال. ميل المجتمع β_1 هو المعلمة الرئيسية **key parmeter** هنا . إذا كانت β_1 = صفر ، لا يكون هناك إرتباط خطى بين Y ، X لذلك ، إذا كان دليل العينة لا يساند وجود علاقة خطية بين Y ، X فى المجتمع ، سوف يكون من الحماقة بناء قرارات جادة على خط المربعات الصغرى .

(2) ما مدى دقة التقديرات أو التنبؤات المبنية على خط المربعات الصغرى؟ إذا كنت ترغب فى إتخاذ قرارات بناء على هذه التقديرات أو التنبؤات . يكون من المهم فهم كيف يمكن أن تبتعد عن ذلك.

(3) هل الشروط ضرورية لتحليل الإنحدار المقدم فى هذا التطبيق؟ كما سترى الآن ، فإن استدلالات الإنحدار (لكى تحقق أول قضيتين) تكون مبنية على فروض معينة عن المجتمع. إذا كانت هذه الفروض غير صحيحة ، فإنه ربما تنتج بعض النتائج المضللة أو حتى السخيفة.

(4) هل من الممكن أن يمتد نموذج الإنحدار المفترض لكى يتضمن المتغيرات التنبؤية (التفسيرية) المحتملة الأخرى؟ على سبيل المثال ، فى مشكلة تثمين العقارات ، هل من الممكن أن تكون

متغيرات أخرى مثل عدد الحمامات bathrooms ، وجود مدفأة، أو عمر المنزل، يكون لها تأثير ذو معنى فى سعر البيع للمنزل؟ هذا بالطبع محتمل جداً ؛ وهذا الفرض سيتم تناوله بالتفصيل فى فصل (١٠) .

(٩-٤-١) فروض النموذج Model Assumptions

تكون الإجراءات المتضمنة فى إستنتاجات الإنحدار صحيحة فقط إذا وجدت شروط معينة للمجتمع. حيث أننا لا نستطيع ملاحظة المجتمع كله ، فإننا يجب أن نكون مستعدين لإفتراض وجود هذه الشروط . وفيما بعد نقدم طرقاً لفحص مصداقية هذه الشروط . وتكون إستنتاجات الإنحدار مبنية على أساس الفروض الأربعة التالية وهى تتعلق بالإرتباط بين Y ، X فى المجتمع . وهى تتركز على وجه الخصوص على النموذج الخطى البسيط المعطى بالمعادلة (9.2) :

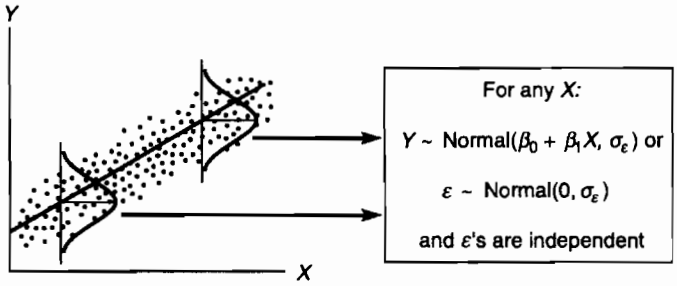
(1) النموذج الخطى البسيط يمثل بشكل صحيح الإرتباط بين متغير الإستجابة response والمتغير المفسر predictor وهذا يعنى أن لكل قيم X التى تقع داخل مدى بيانات العينة، قيمة متوسطة لمتغير الاستجابة Y تعطى عن طريق خط إنحدار المجتمع عند هذه القيمة للمتغير X . بتعبير آخر ، فإنه لأى قيمة X ، يكون هناك مجتمع لقيم Y حيث أن ($E(Y) = \beta_0 + \beta_1 X$)ومتوسط الخطأ العشوائي يساوى صفر.

(2) تباين الخطأ σ_ε^2 يكون ثابت **The error variance σ_ε^2 is constant** ، تباين ثابت للخطأ يعنى أن إختلاف قيم Y حول خط إنحدار المجتمع هو نفسه الإختلاف لكل قيم X داخل مدى بيانات العينة. وإذا كان إختلاف قيم Y حول معادلة الإنحدار تعتمد على قيم X ، فإن σ_ε^2 تكون غير ثابتة لكل قيم X . هذا يستلزم إستخدام إجراء بديل للإنحدار معروف على أنه المربعات الصغرى المرجحة **weighted least squares** وهى خارج نطاق عرض هذا الكتاب.

(3) الأخطاء العشوائية مستقلة **The random errors are independent**. هذا الفرض ينطوى على أن الخطأ المصاحب لأى قيمتين من قيم Y يكون مستقل. على سبيل المثال، بمعلومية قيمة واحدة لـ Y تكون أعلى أو أسفل خط إنحدار المجتمع لا يخبرنا بشئ عن ما إذا كانت قيمَة أخرى أعلى أو أسفل خط الإنحدار .

(4) الخطأ العشوائى يتبع توزيع طبيعي **The random errors are normally distributed**. هذا الفرض ينطوى على أن قيم Y توزع طبيعياً حول خط إنحدار المجتمع. ولذلك فإننا نفترض أن ε هو متغير عشوائي طبيعى بمتوسط 0 وإنحراف معيارى σ_ε ، لذلك لأى قيمة X، تكون Y متغير عشوائى طبيعى بمتوسط $(\beta_0 + \beta_1 X)$ وإنحراف معيارى σ_ε.

هذه الفروض يمكن تلخيصها بمصطلحات إحصائية عن طريق قول أن توزيع قيم Y ؛ بمعلومية X يتبع التوزيع الطبيعى بمتوسط قدره $(\beta_0 + \beta_1 X)$ وتباين قدره σ_ε^2 أي أن $Y \sim Normal\ (\beta_0 + \beta_1 X,\ \sigma_\varepsilon)$ ويكون التعبير المكافئ أن توزيع الخطأ يتبع التوزيع الطبيعى بمتوسط 0 وتباين σ_ε^2 أى أن $\varepsilon \sim Normal\ (0,\ \sigma_\varepsilon)$ ، حيث أن الإنحراف المعيارى للخطأ σ_ε يكون ثابتاً على المدى الكلى لقيم العينة X . هذه الفروض موضحة فى شكل (٩-١٤) .



شكل (٩-١٤) فروض نموذج الإنحدار

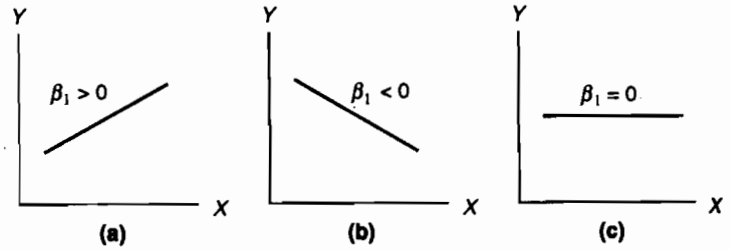
فحص فروض الإنحدار Checking the Regression Assumptions

إن الفروض الثلاثة الأولى السابقة أكثر أهمية من الفرض الرابع (الأخير) . وإستنتاجات الإنحدار تكون لدرجة معقولة غير حساسة لمخالفة فرض الاعتدالية (**normality**) . وعلى الجانب الآخر، لا يكون هناك تحليل إنحدار كامل بدون التأكد من صحة الفروض الثلاث الأولى. ولكى يتم هذا ، نفحص بمجرد النظر **visually** رسم البواقى المصاحبة لخط المربعات الصغرى. ويكون هدفنا هو إكتشاف نموذج بين البواقى قد يوحى بمخالفة ممكنة لفرض ما . حيث أن البواقى تمثل الأخطاء العشوائية، فيجب أن لا تظهر نموذج يمكن تمييزه .

قبل أن نبدأ مع إستنتاجات الإنحدار، هناك إعتبار واحد متبقى فى أى تحليل إنحدار والذى لا يمكن التأكيد عليه بشدة. وهو يتعلق بطبيعة بيانات العينة. البيانات المستخدمة فى تقدير معادلة الإنحدار يجب أن تكون ممثلة، بمعنى أن تكون بيانات مقطعية (**Cross - Section**) لبيئة ما نرغب فى دراستها وإذا لم تكن كذلك، فإن معادلة المربعات الصغرى لن تمثل البيئة المهتم بها، وهكذا يمكن أن تعطى نتائج مضللة للغاية .

(٩-٤-٢) المتوسط والخطأ المعيارى للتقديرات b_0 , b_1 :

تعتبر مقدرات المربعات الصغرى b_0 , b_1 أفضل الاحصاءات المتعلقة بتقدير ميل المجتمع β_1 والجزء المقطوع β_0، على الترتيب . وإذا كانت β_1 تساوى صفر، فإنه لا يوجد ارتباط خطى بين X , Y فى المجتمع . بتعبير آخر فإن العلاقة الخطية بين X , Y تتواجد للمجتمع فقط إذا كانت β_1 لا تساوى صفر . شكل (٩-١٥) يوضح خط إنحدار المجتمع $(E(Y) = \beta_0 + \beta_1 X)$ وجود إرتباط خطى في الأجزاء (a) , (b) بينما يوضح الجزء (c) عدم وجود علاقة خطية بين X , Y، والإستنتاجات التى تتعلق بالميل β_1 تكون أكثر أهمية إلى حد بعيد عن الإستنتاجات عن الجزء المقطوع β_0 .



شكل (٩-١٥) وجود أو عدم وجود علاقة خطية بين X , Y

ربما تتذكر من الفصل الخامس وحتى الفصل الثامن أن بداية الإستنتاج الإحصائي هى تحديد توزيع المعاينة لأفضل إحصاء **the best statistic** . حيث أن b_1 هى أفضل إحصاء لإثبات تواجد علاقة خطية، يجب أن نحدد توزيع المعاينة لها. وربما يكون هناك صعوبة فى البداية بإعتبار أن b_1 متغير عشوائى. هذا بالطبع ولأن b_1 إحصاء لعينة؛ فإن b_0 تكون كذلك، وتعتمد قيمهم الفعلية فى أى تطبيق على العينة الممثلة المحددة التى يتم اختيارها .

نحدد أولاً المتوسط والخطأ المعيارى لـ b_1؛ ثم نحدد توزيع المعاينة لها . لقد تم ايضاح أن المتوسط والخطأ المعيارى لـ b_1 يكون كما يلى :

$$E(b_1) = \beta_1 \qquad \textbf{(9.12)}$$

$$SE(b_1) = \sqrt{\frac{\sigma_\varepsilon^2}{SS(X)}} \qquad \textbf{(9.13)}$$

حيث أن SS (X) هى مجموع المربعات الكلى للمتغير X كما تم تعريفه فى معادلة (9.6). بإسترجاع أن خطأ التباين غير المعروف σ_ε^2 يجب أن يقدر عن طريق تباين البواقى S_e^2 . بذلك يكون الخطأ المعيارى المقدر لـ b_1 يكون :

$$SE(b_1) = \sqrt{\frac{S_e^2}{SS(X)}} \qquad \textbf{(9.14)}$$

وقد تم ايضاح أيضاً أن المتوسط والخطأ المعيارى المقدر لـ b_0 على الصورة التالية :

$$E(b_0) = \beta_0 \qquad \textbf{(9.15)}$$

$$SE(b_0) = \sqrt{\frac{S_e^2 \sum X_i^2}{nSS(X)}} \qquad \textbf{(9.16)}$$

على التوالى . لذلك فإن إحصاء المربعات الصغرى b_1 يكون مقدر غير متحيز لميل المجتمع β_1، وكذلك يكون الإحصاء b_0 للجزء المقطوع β_0.

ولتوضيح حساب $SE(b_1)$، $SE(b_0)$، بإسترجاع مثال (٩-٤) (مشكلة مطعم عش البلبل Bird's Nest). فى هذا المثال، حددنا أن: $(b_1 = .2)$, $(b_0=3.5)$ ، $(SS(X) = 850)$ ، $(S_e^2= 1.875)$ ، $(n = 6)$. بإستخدام قيم X ، يمكننا أيضاً تحديد أن $(\sum X_i^2= 6{,}250)$. بالتعويض فى المعادلات (14.9) ، (16.9) نجد أن الخطأ المعيارى المقدر لـb_1 يكون :

$$SE(b_1) = \sqrt{\frac{1.875}{850}} = .047$$

والخطأ المعيارى المقدر لـ b_0 يكون :

$$SE(b_0) = \sqrt{\frac{(1.875)(6{,}250)}{(6)(850)}} = 1.5158$$

والآن ، توقف دقيقة لتجيب عن سؤال طالما قمت بالإجابة عنه عدة مرات من قبل : بمعلومية أن $(b_1 = .2)$, $(SE(b_1) = .047)$ ، هل تجد أنه يكون مقبول الإدعاء بأن ميل المجتمع β_1 يساوى صفر؟ إذا إستخدمنا فكرة فترة الثقة، سوف نجد أن أكثر من أربع أخطاء معيارية أصغر من $(b_1 = .2)$ المقدرة ويوجد القيمة 0 من بينهم. حيث أن $(\beta_1 = 0)$ تبدو غير مقبولة، فإن هذا التحليل يدل على أن الإرتباط الخطى بين X , Y يوجد بالفعل. وسوف نعود لهذا النوع من التفكير بعد تحديد توزيع المعاينة للمقدر b_1 .

(٩-٤-٣) توزيع المعاينة للمقدر b_1 : The Sampling Distribution of b_1

إذا كانت الفروض الأساسية عن المجتمع صحيحة، فإنه يمكن توضيح أن توزيع المعاينة لإحصاء المربعات الصغرى b_1 هو توزيع طبيعى بمتوسط وخطأ معيارى كما هو معطى عن طريق معادلة (9.12) ، (9.13) على التوالى . الآن، لماذا يتوزع المقدر b_1 طبيعياً؟ من الممكن توضيح أن معادلة (9.3) للاحصاء b_1 هى توليفة خطية لمتغيرات عشوائية طبيعية (قيم العينة Y)* . نتيجة لذلك، b_1 نفسها تتوزع طبيعياً . (انظر الفصل (٥-٥-٢) لمراجعة توزيع التوليفة الخطية لمتغيرات عشوائية طبيعية)

حيث أن b_1 تتوزع طبيعياً، يمكننا بسهولة جعلها معيارية وتحويلها إلى الإحصاء Z إذا كنا نعلم تباين الخطأ σ_ε^2. ويمكننا بعد ذلك إستخدام التوزيع الطبيعى المعيارى لعمل إستدلات عن الميل β_1. لكن طالما أن تباين الخطأ σ_ε^2 غير معروف، يجب أن نبدل σ_ε^2 فى معادلة (9.13) بتباين البواقى S_e^2 . كما تتوقع فإن جعل b_1 معيارية فى هذه الطريقة يؤدى إلى الإحصاء T التالية :

$$T = \frac{b_1 - \beta_1}{SE(b_1)} \qquad \textbf{(9.17)}$$

التى لها توزيع T بدرجات حرية (n-2) ، لأن مقام المعادلة لتباين البواقى S_e^2 هو (n-2)، لذلك فإن الإستنتاج الإحصائى المتعلق بالميل β_1 يكون مبنى على أساس إحصاء المربعات الصغرى b_1 ، ويتضمن توزيع T بدرجات حرية (n-2) .

(٩-٤-٤) فترات الثقة وإختبار الفروض لـ β_1 :

Confidence Intervals and Hypothesis Testing For β_1:

إستخدام توزيع T لعمل إستنتاجات إحصائية يجب أن يكون شيء مألوفاً لديك من الآن. بالنسبة إلى الإستنتاجات عن الميل β_1 ، فإن الاجراءات تبقى كما هى في جوهرها. (يمكنك مراجعة الفصول (٥-٥-٤)، (٦-٤-٣) في هذا الشأن)

فترات الثقة لـβ_1 :

طالما أن أساس فهم الإرتباط بين X , Y هو الميل β_1 لخط إنحدار المجتمع، يكون من الضرورى أن نفهم الدقة التى تم بها تقدير β_1 . كما هو الحال دائماً فى التقدير، نصف دقة المقدر عن طريق حساب هامش خطأ المعاينة له وفترة الثقة الناتجة. وبناء على مناقشة فترات الثقة فى جزء(٦-٤-٣)، فأن فترة ثقة $100(1-\alpha)\%$ للميل β_1 تعطى عن طريق :

$$b_1 \pm t_{1-\alpha/2,\ n-2}\ SE(b_1) \qquad \textbf{(9.18)}$$

حيث :

$$\text{Margin of sampling error (هامش خطأ المعاينة)} = t_{1-\alpha/2,\ n-2}\ SE(b_1) \qquad \textbf{(9.18)}$$

($t_{1-\alpha/2,\ n-2}$)هى القيمة الجدولية المناسبة لتوزيع T (انظر جدول C فى الملحق فى نهاية هذا الكتاب) .

---

* تذكر أن قيم X يمكن اختيارها. لذلك تم إعتبارها على أنها ثابتة بدلاً من اعتبارها كميات عشوائية طالما أنه تم قياسها بدون خطأ .

للتوضيح ، دعنا نحدد ثقة 95% للميل β_1 فى مشكلة مطعم عش البلبل (مثال ٩-٤) . حددنا من قبل أن $(b_1 = .2)$, $(SE(b_1) = .047)$, $(n=6)$. لمستوى ثقة 95% ودرجات حرية $(n - 2 = 4)$ ، نجد أن القيمة الجدولية هى :($t_{.975,4} = 2.776$) . هكذا تكون فترة ثقة 95% للمعلمه β_1 هى :

$$.2 \pm (2.776)(.047) = .2 \pm .13$$

أو (.33 , .07) . كما فى الحالات السابقة ، يمكننا إعتبار مجموعة قيم مقبولة لـ β_1 ضمن الفترة (.33 , .07) وتكون القيم خارج هذه الفترة غير مقبولة. على وجه الخصوص ، لاحظ أن القيمة $(\beta_1 = 0)$ لا تكون واقعة فى تلك الفترة . وهذا يعنى أن الإدعاء بأنه لا يوجد إرتباط خطى بين X , Y فى المجتمع غير مقبول ومن الممكن أن يرفض .

على الرغم من أن فترة الثقة 95% لـ β_1 تدل على وجود إرتباط خطى بين X , Y ، الا ان هذه الفترة تدل على أن β_1 لم يتم تقديرها بدقة كبيرة ، على سبيل المثال ، الفترة (.33 , .07) لقيم β_1 تعنى أنه إذا زاد عدد الوافدين X بواحد فى اليوم فإن متوسط الإيراد Y ربما يزيد بمقدار قليل مثل 7$ إذا كانت $(\beta_1 = .07)$ أو بمقدار كبير 33$ إذا كانت $(\beta_1 = .33)$. هذا النقص فى دقة تقدير β_1 لايكون مفاجئاً إذا اعتبرنا أن العينة مكونة من بيانات لست فترات مسائية جديدة (n=6) . (تعمدنا إستخدام عينة صغيرة لتسهيل التمثيل ، لكن عدم ملائمة مثل هذه العينة الصغيرة لمعظم التطبيقات الفعلية سوف يظهر الآن) .

إختبار الفروض لـ β_1 :

فى مشكلة مطعم عش البلبل ، إفترض أننا نرغب إختبار فرض العدم **null hypothesis** بعدم وجود إرتباط خطى بين الإيراد (Y) وعدد الوافدين (X) مقابل الفرض البديل أنه يوجد إرتباط خطى . فيما يتعلق بالنموذج الخطى البسيط ، معادلة (9.2)، فإن هذه الفروض من الممكن أن تصاغ كالآتى:

$$H_0: \quad \beta_1 = 0 \qquad \textbf{(9.20)}$$
$$H_a: \quad \beta_1 \neq 0$$

كما فعلنا سابقاً، فترات الثقة يمكن إستخدامها لتحديد ما إذا كان دليل العينة يخالف إدعاء H_0 بعدم وجود إرتباط خطى بين X , Y . إذا كان الصفر لا يدخل ضمن الفترة بمستوى ثقة عالى ، فإن دليل العينة يخالف إدعاء عدم وجود إرتباط خطى بين X , Y ويساند وجود إرتباط خطى (الفرض البديل) . حيث أن فترة ثقة 95% وهى: (.33 , .07) لمشكلة المطعم لا تتضمن صفر ، فإن دليل العينة يخالف الفرض العدمى H_0 ويساند وجود علاقة خطية بين الإيراد وعدد العملاء . لاحظ أنه عن طريق النظر لشكل الإنتشار فى شكل (٩-١٣) ، سوف نصل لنفس الإستنتاج .

يمكن الوصول لنفس النتيجة عن طريق إستخدام مدخل القيمة P-value P. القيمة P تظهر إلى أى مدى يخالف دليل العينة الفرض بعدم وجود إرتباط خطى بين X , Y . كما سبق ، كلما صغرت القيمة P ، كلما ضعف دليل قبول H_0 ، وهكذا يصبح الفرض البديل أكثر قوة .

إفترض أن ادعاء الفرض العدمى صحيح (أي $\beta_1 = 0$) ، نجد أن قيمة الاحصاء T (معادلة {9.17}) لمشكلة مطعم عش البلبل Bird's Nest هى :

$$T = \frac{.2 - 0}{.047} = 4.26$$

بالنسبة للفرض البديل ذو الجانبين ، القيمة P هي إحتمال أن الاحصاء T مع أربع درجات حرية سوف يعطى قيمة سواء أكانت أكبر من 4.26 أو أقل من 4.26-. والكمبيوتر يشير إلى أن هذا الإحتمال هو {2 (.0065) = .013} . وإذا لم يكن لدينا مدخل إلى الكمبيوتر، يمكننا تقريب القيمة P عن طريق إستخدام جدول C في الملحق. نفحص الصف الخاص بأربعة درجات حرية ونستخرج القيم 3.747 , 4.604 التى تحصر (T = 4.26)، نجد أن المنطقة إلى اليمين من 4.604 هى (.005 = 1-.995)، والمنطقة إلى اليمين من 3.747 هى (.01 = 1-.099) هكذا، المنطقة إلى اليمين من T = 4.26 تكون بين .005 , .01. وحيث أن الفرض البديل ذو جانبين ، فإن القيمة P المطلوبة تكون بين (.01 = 2(.005)) , (.02 = 2(.01)) . القيمة P التى تكون أقل من .02 تعتبر غالباً صغيرة جداً لكى تساند إدعاء Ho بعدم وجود إرتباط خطى. لذلك، من الممكن أن نستنتج أن هناك فعلاً إرتباط خطى بين الإيراد وعدد الوافدين للمجتمع . (لكن ضع فى ذهنك ما يتعلق بعدم الدقة فى تقدير الميل β_1 يكون أساسه أن العينة الصغيرة) .

(٩-٤-٥) إستخدام أسلوب تحليل التباين فى الإنحدار الخطى البسيط :

An Analysis of Variance Approach in Simple Linear Regression

فى تعريف معامل التحديد r^2 فى الجزء (٩-٣-٤)، أخذنا فكرة عن تحليل التباين عن طريق تقسيم الإختلاف الكلى (SST) فى قيم العينة Y إلى مركبتين: مجموع مربعات الإنحدار (SSR)، الذى يحسب الإختلاف فى قيم Y بالعينة التى يمكن تفسيرها عن طريق الإختلاف فى قيم X بالعينة ، ومجموع مربعات الخطأ (SSE) ، الذى يحسب الإختلاف فى قيم Y والتي لايمكن تفسيرها عن طريق الإختلاف فى قيم X – أى ، الإختلاف الذى يرجع إلى أسباب عشوائية .

كبديل لإجراء T فى الجزء السابق ، يمكننا إستخدام تحليل التباين لإختبار الفرض العدمى .

$$H_0: \beta_1 = 0$$

مقابل الفرض البديل ذو الطرفين :

$$H_a: \beta_1 \neq 0$$

وإجراء تحليل التباين معادل لإجراء الاحصاء T إذا كان الفرض البديل ذو طرفين. هذا يعنى أن قيم P دائماً تكون متماثلة في الحالتين ولذلك فإن الإستنتاج يكون دائماً هو نفس القرار. ومع ذلك ، فإن إجراء تحليل التباين يقدم فكرة عن المصادر الممكنة للإختلاف والتى تكون بشكل خاص مفيدة إذا أخذ في الاعتبار نماذج معقدة (كما تم مناقشته فى فصل 10) . إضافة إلى ذلك، مخرجات معظم الحزم الإحصائية للكمبيوتر تشتمل على معلومات وثيقة الصلة لكل من المؤشر الإحصائي T وإجراء تحليل التباين. إذا درست كلا الإجرائيين، عملياً فإن كل هذه المخرجات سوف تكون مفيدة لك ، وسوف تكون مهيأ بشكل أفضل لدراسة تطور أكثر النماذج تعقيداً

وبإسترجاع تحليل التباين ، فإن كل مجموع مربعات مرتبط بعدد معين من درجات الحرية. كما فى فصل (8)، فمجموع المربعات الكلى SST له (n-1) درجات حرية. عدد درجات الحرية لمجموع مربعات الباقى SSE هو (n-2) ، حيث أن هناك معلمتان يتم تقديرهما (β_0 , β_1) فى معادلة SSE . ولأن درجات الحرية لمجموع المربعات الكلى تحسب عن طريق جمع درجات الحرية لمجموع مربعات الإنحدار ودرجات الحرية لمجموع مربعات البواقى .

(df for SST = df for SSR + df for SSE)

هذا يترك درجة حرية واحدة لمجموع SSR. ويكون صحيح دائماً أن درجات حرية SSR مماثلة لعدد الحدود فى نموذج الإنحدار المفترض الذى يتضمن متغيرات تفسيرية. وحيث يوجد واحد فقط لمثل هذا الحد للنموذج الخطى البسيط (هذا الحد هو $\beta_1 X$) تكون هناك درجة حرية واحدة لمجموع مربعات الإنحدار SSR .

والآن نتذكر ثانية من الفصل الثامن أن متوسط المربعات يعرف على أنه مجموع المربعات مقسوما على درجات الحرية لهذا المجموع. وبالتالى فإن متوسط مربعات الانحدار تكون :

$$MSR = \frac{SSR}{1} \qquad (9.21)$$

بينما متوسط مربعات الخطأ (تباين البواقى) يكون :

$$MSE = S_e^2 = \frac{SSE}{n-2} \qquad (9.22)$$

لإختبار الفرض العدمى أنه لا يوجد إرتباط خطى بين X , Y . نقارن متوسط المربعات للإنحدار بمتوسط المربعات للخطأ. هكذا، فإن المؤشر الإحصائى فى إجراء تحليل التباين يكون عبارة عن النسبة التالية :

$$F = \frac{MSR}{MSE} \qquad (9.23)$$

إذا كان الفرض العدمى بعدم وجود إرتباط خطى بين X , Y صحيحا، فإن توزيع المعاينة لهذه النسبة هو توزيع F بدرجات حرية 1 ، (n-2) .

سوف نستخدم هذه البديهيات لتحديد أى قيم للنسبة F سوف تساعدنا على إستنتاج أنه توجد علاقة خطية بين X , Y . وإذا كانت Y لها علاقة خطية بالمتغير X، فإن SSR والتي تمثل الاختلاف أو التغير في قيم Y والذي يفسر عن طريق الاختلاف في قيم X، يجب أن يكون كبير نسبيا. في المقابل، فإن الجزء غير المفسر للمتغير Y وهو SSE يجب أن يكون صغيراً نسبياً. وحيث أن SSR كبير، SSE صغير، فإن بيانات العينة توضح أن هناك علاقة خطية بين Y ، X. لذلك فإنه كلما كانت قيمة F كبيرة كلما كان هناك دليل قوى على وجود الارتباط الخطى بين X , Y . بعبارة أخرى فإنه كلما كبرت قيمة F ، كلما صغرت قيمة P ، وكلما قوى الدليل ضد الفرض العدمي بعدم وجود إرتباط خطى بين X , Y .

وللتوضيح ، دعنا نستخدم مثال مطعم عش البلبل Bird's Nest حيث أن :

$$SST = 41.5 \quad , \quad SSR = 34.0 \quad , \quad SSE = 7.5 \quad , \quad n = 6$$

وجدول تحليل التباين (ANOVA) للإختبار ($H_0: \beta_1 = 0$) مقابل ($H_a: \beta_1 \neq 0$) موضح فى جدول (٩-١).

جدول (٩-١)

جدول ANOVA لمشكلة مطعم عش البلبل Bird's Nest

المصدر	d f	مجموع المربعات	متوسط المربعات	قيمة F	قيمة P
عدد الوافدين	1	34.0	34.5	18.13	.0131
الخطأ	4	7.5	1.875		
إجمالي	5	41.5			

قيمة P هى (.0131) ولقد تم تحديدها فى الجدول بإستخدام الحاسب. ويمكننا تقريب قيمة P هذه عن طريق إستخدام جدول E فى الملحق كما يلى . بإستخدام درجة حرية واحدة (1) للبسط ، (4) للمقام ، نبحث عن القيمتين اللتين تحدان قيمة (F = 18.13) . هاتان القيمتان هما (12.22) ، (21.20). (لاحظ أن كل منهما تظهر فى صفحة مختلفة) . نلاحظ أن المنطقة على اليمين من 12.22 تكون (.025 = .975-1) ، والمنطقة على يمين 21.20 تكون (0.1 = .99-1) وهكذا فإن المنطقة على اليمين من (F = 18.13) تكون بين .025 , .01 لذلك فإن قيمة P تكون بين (.01) , (.025) .

كما سبق أن أوضحنا فإن خطوات إيجاد تحليل التباين و T تكون متعادلة. دعنا نتحقق من ذلك فى مثال مطعم عش البلبل. حيث أن قيمة T كانت (T = 4.26) . لاحظ أنه إذا تمكنا بتربيع قيمة (T = 4.26) ، نحصل على ($\{4.26\}^2 = 18.15$)، التى تعطي قيمة المؤشر الإحصائى F فى جدول تحليل التباين (مع وجود بعض التقريب) . لاحظ أيضاً أن قيم P هي نفس القيمة لاتتغير(.0131) . وفى الحقيقة، يمكن إثبات أن مربع المتغير العشوائى T بدرجات حرية γ هو المتغير العشوائى F بدرجات حرية 1 للبسط و γ للمقام ؛ أى أن:

$$T_v^2 = F_{1,v} \qquad \textbf{(9.24)}$$

لذلك ، فإن الأحصاء T (معادلة (9.17)) والأحصاء F الناتج عن تحليل التباين (معادلة 9.23) يكونا متكافئان فعلاً .

مثال (٩-٥)

فى مثال تثمين العقارات الذى تمت مناقشته فى البنود (٩-٣-٢)، (٩-٣-٤)، حددنا أن خط الإنحدار المقدر هو :

$$\hat{Y} = .8 + 6.0X, \quad SS(X) = 2.5, \quad SST = 108.8, \quad SSR = 90.0, \quad SSE = 18.8,$$
$$S_e^2 = 6.2667, \text{ and } n=5$$

(a) اعتماداً على فترة ثقة 95% للمعلمة β_1 ، هل ميل المجتمع β_1 تم تقديره بدقة معقولة؟ إشرح .

(b) ما المنتظر أن تكون عليه قيمة العينة لـ b_1 ، الميل المقدر ، إذا لم يكن هناك فعلاً إرتباط خطى بين سعر البيع (Y) وحجم المنزل (X) فى المجتمع ؟

(c) اعتماداً على إجراء تحليل التباين ، هل دليل العينة يخالف الفرض القائل بأنه لا يوجد ارتباط خطى بين X , Y .

الحل

(a) من معادلة (9.14) الخطأ المعيارى المقدر لـ b_1 هو :

$$SE(b_1) = \sqrt{\frac{6.2667}{2.5}} = 1.5832$$

عند درجات حرية (n-2=3) ومستوى ثقة 95% ، نجد أن قيمة T الجدولية Quantile value هى ($t_{.975,3} = 3.182$) . حيث أن ($b_1 = 6$) فإن فترة الثقة 95% لـβ_1 تكون

$$6 \pm (3.182)(1.5832) = 6 \pm 5.04$$

أو (11.04 , .96) . لاحظ أن القيمة $\beta_1 = 0$ لا تدخل ضمن هذه الفترة، وهذا يدل على وجود الإرتباط الخطى بين X , Y . فى نفس الوقت، وبرغم ذلك فإن هذه الفترة توضح أن الميل لم يتم

تقديره بدقة كبيرة. فعندما تزيد X بوحدة واحدة (1,000 قدم مربع) ، فإن الزيادة المنتظرة فى متوسط سعر البيع يمكن أن تكون قليلة بمقدار $9,600 (إذا كانت $\beta_1 = .96$) أو كبيرة بمقدار $110,400 (إذا كانت $\beta_1 = 11.04$) . وهذا المدى الواسع يدل على نقص أو ضعف درجة الدقة فى تقدير β_1. وبدون شك فإن حجم العينة الصغير (n=5) ساهم فى ذلك .

(b) قيمة إحصاء ميل المربعات الصغرى هى ($b_1 = 6$) . ويتم تحديد إحتمال ملاحظة مثل هذه النتيجة في العينة عند عدم وجود إرتباط خطى بين X , Y عن طريق حساب القيمة P . حيث تخبرنا التجارب أن الميل يجب أن يكون موجب إذا وجد إرتباط خطى بين سعر البيع وحجم المنزل، دعنا نختبر الفرض العدمي ($H_0 : \beta_1 = 0$) مقابل الفرض البديل ذو الطرف الواحد ($H_0: \beta_1 > 0$) . عند ($b_1 = 6$) ، ($SE(b_1) = 1.5832$)، فإن قيمة الإحصاء T (معادلة {9.17}) تكون :

$$T = \frac{6 - 0}{1.5832} = 3.79$$

القيمة P هى إحتمال أن الإحصاء T بدرجات حرية ($n-2 = 3$) تأخذ قيمة أكبر من 3.79 أي :

$$P\text{-value} = P(T_3 > 3.79)$$

والقيمة P الفعلية المحسوبة من الكمبيوتر هى (.0161) . من جدول C فى الملحق، نرى أن ($T = 3.79$) تقع بين القيم الجدولية ($\{t_{.975,3}\}$ 3.182) ، ($\{t_{.99,3}\}$ 4.541) ؛ لذلك، نقرب القيمة P لكى تكون بين (.01) ، (.025) . ولأن القيمة P صغيرة جداً، فإن دليل العينة يخالف فرض العدم ويساند الفرض البديل بوجود إرتباط خطى موجب بين X , Y فى المجتمع .

(c) تذكر أن ($SST = 108.8$) , ($SSR = 90.0$) , ($SSE = 18.8$) بإستخدام هذه الكميات ، يكون جدول تحليل التباين لإختبار ($H_0 : \beta_1 = 0$) مقابل الفرض البديل ذو الطرفين ($H_a : \beta_1 \neq 0$) ، القيمة P المعطاة فى جدول (٩-٢) تم إستنباطها بالكمبيوتر. وهى بالضبط ضعف القيمة P فى جزء(b) ، حيث أن القيمة P للاحصاء T فى جزء (b) تتعلق بفرض بديل ذو طرف واحد. وكما هو متوقع، تكون القيمة F للمقدار (14.36) هى مربع القيمة T فى جزء (b) ($3.79^2 = 14.36$) . وحيث أن القيمة P صغيرة، فإن دليل العينة الحالى يظهر ليخالف الفرض العدمي بعدم وجود إرتباط خطى بين سعر البيع وحجم المنزل، كما فى جزء (b) .

جدول (٩-٢)

جدول ANOVA لمشكلة تقييم الملكية

المصدر	d f	مجموع المربعات	متوسط المربعات	قيمة F	قيمة P
الحجم	1	90.0	90.0	14.36	.0322
الخطأ	3	18.8	6.2667		
إجمالي	4	108.8			

إستخدام الكمبيوتر :

الأمثلة التى إستخدمناها لتوضيح إجراءات الإنحدار كانت سهلة نسبياً فيما يتعلق بالعمليات الحسابية، لكن من الواضح أنه من المرغوب فيه إستخدام الكمبيوتر لتنفيذ تحليل التباين عندما يكون متاح لك ذلك . ونعرض مخرجات Minitab لمثال (٩-١) (الطاقة مقابل درجة الحرارة) ، (٩-٢) (الأجر مقابل سنوات الخبرة) .

جدول(٩-٣)

مخرجات المثال (٩-١) بإستخدام Minitab

```
The regression equation is energy = 87.0 - 3.46 temp

Predictor          Coef        Stdev       t-ratio          p
Constant        86.9957       0.7572        114.89      0.000
temp            -3.4642       0.1395        -24.84      0.000

s=2.586   R-sq=97.5%   R-sq(adj)=97.3%

Analysis of Variance

SOURCE              DF          SS          MS          F          p
Regression           1      4123.5      4123.5     616.82      0.000
Error               16       107.0         6.7
Total               17      4230.5

      ROW        temp      energy        yhat    residual
        1        -1.0          94      90.460     3.54010
        2         1.5          81      81.799    -0.79943
        3         3.5          79      74.871     4.12894
        4        -3.0          97      97.388    -0.38828
        5         0.5          88      85.264     2.73638
        6         2.5          75      78.335    -3.33524
        7         4.0          74      73.139     0.86104
        8         5.0          67      69.675    -2.67477
        9        -5.0         107     104.317     2.68335
       10        -0.5          86      88.728    -2.72781
       11         9.0          58      55.818     2.18197
       12         9.5          55      54.086     0.91407
       13         7.0          65      62.746     2.25360
       14         3.0          73      76.603    -3.60315
       15        -2.0          91      93.924    -2.92409
       16         6.0          65      66.211    -1.21059
       17         8.0          58      59.282    -1.28222
       18        10.0          52      52.354    -0.35384
```

لبيانات عينة مثال (٩-١) ، فإن مخرجات الكمبيوتر لتحليل التباين، تكون متضمنة نسخة لقيم X , Y ، $\hat{Y}$ والبواقى معطاة فى جدول (٩-٣). لاحظ أن معادلة المربعات الصغرى Energy= 87-3.46 Temp. تعطى الإستخدام الأول للأسماء التى نستخدمها لتسمية المتغيرات X , Y (عنوان العمود يمكن أن يستخدم أيضاً). تذكر أن ذلك يتم تطبيقه على مدى درجات الحرارة

للعينة من (-5) إلى (10) درجات سليزية. ثم تعرف قيم إحصاءات المربعات الصغرى، ($b_0 = 86.9957, b_1 = -3.4642$) مع أخطائها المعيارية المقدرة ($SE(b_0) = .7572$, $SE(b_1) = .1395$)، قيمة الأحصاء T، وقيم P المناظرة. المخرجات تستمر بتقديم الإنحراف المعيارى للباقى ($S_e = 2.586$)، قيمة r^2 معبر عنها كنسبة مئوية (97.5%) ، وقيمة r^2 المعدلة (التى بيناها فى فصل ١٠) أخيراً، جدول ANOVA معطى لإختبار فرض العدم ($H_0 : \beta_1 = 0$) مقابل الفرض البديل ($H_a : \beta_1 \neq 0$) .

ومن مخرجات جدول (٩-٣) ، هناك شك صغير جداً في وجود علاقة خطية بين إستهلاك الطاقة ودرجة الحرارة في المجتمع (بيانات العينة تخالف بسهولة فرض العدم ($H_0 : \beta_1 = 0$) بإستخدام كل من الإحصاءة T أو F. بالإضافة إلى ذلك ، 97.5% من الإختلاف فى قيم عينة الطاقة تفسر عن طريق الإختلاف فى درجات الحرارة. هذه النتيجة، مع الحقيقة الهامة أن الميل β_1 تم تقديره بدقة كبيرة كما وضحنا فى بند (٩-٥) ، تقودنا إلى تصديق أن خط المربعات الصغرى: (الحرارة 3.46 - 87.0 = الطاقة) يكون ملائماً للتقدير والتنبؤ ضمن مدى درجات الحرارة المستخدم فى تحديد هذا الخط.

ولبيانات عينة مثال (٩-٢)، فإن مخرجات الكمبيوتر، تتضمن قائمة قيم $\hat{Y}$ والبواقى، مقدمة فى جدول (٩-٤). كما لاحظنا سابقاً (بناء على شكل الإنتشار) ، لا يوجد بالفعل أدنى شك أنه توجد علاقة بين الأجر وسنوات الخبرة. بالرغم من أن هذا التحليل يظهر العلاقة الخطية بين الأجر وسنوات الخبرة، فإنه يظل هناك سؤال عن ما إذا كان الخط المستقيم هو أفضل وصف لهذه العلاقة. نقول هذا بسبب أنه فى شكل الإنتشار (شكل ٩-٧) نرى أنه بعد 20 سنة أو أكثر ، لايجب استخدام خط المربعات الصغرى (سنة 1.26 + 28.9 = الأجر) للتقدير أو التنبؤ وبالتالي يجب أن نحصل على النموذج الذى له درجة إنحناء أفضل، ويصف بطريقة أفضل طبيعة الإرتباط (مثل هذه النماذج تمت مناقشتها فى فصل ١٠) .

جدول (٩-٤)

نتائج مثال (٩-٢) بإستخدام برنامج Minitab

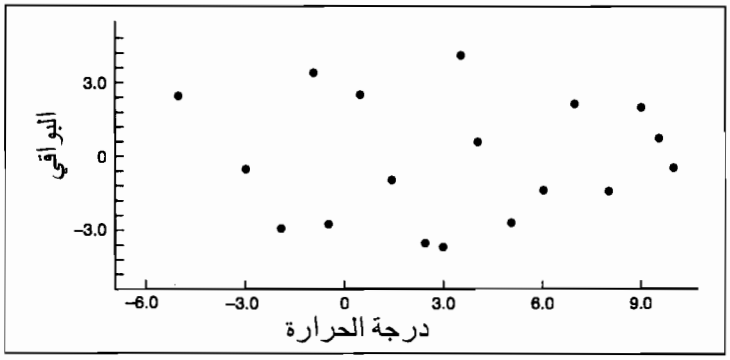
```
The regression equation is salary = 28.9 + 1.26 years

Predictor          Coef        Stdev      t-ratio          p
Constant         28.946        2.325        12.45      0.000
years            1.2607       0.1279         9.86      0.000

s = 5.017   R - sq = 87.4%   R - sq(adj) = 86.5%

Analysis of Variance

SOURCE              DF          SS          MS          F        p
Regression           1      2446.7      2446.7      97.22    0.000
Error               14       352.3        25.2
Total               15     2799.00

ROW              years      salary        yhat    residual
 1                   1          23     30.2064    -7.20641
 2                   2          27     31.4671    -4.46709
 3                   4          29     33.9885    -4.98847
 4                   5          34     35.2492    -1.24916
```

5	6	38	36.5098	1.49015
6	9	46	40.2919	5.70809
7	11	48	42.8133	5.18671
8	14	54	46.5953	7.40465
9	16	54	49.1167	4.88328
10	20	59	54.1595	4.84053
11	22	58	56.6809	1.31915
12	24	59	59.2022	-0.20222
13	25	61	60.4629	0.53709
14	27	63	62.9843	0.01571
15	29	59	65.5057	-6.50566
16	30	60	66.7663	-6.76635

(٩-٤-٦) مقدمة لتحليل البواقى An Introduction to the Analysis of Residuals

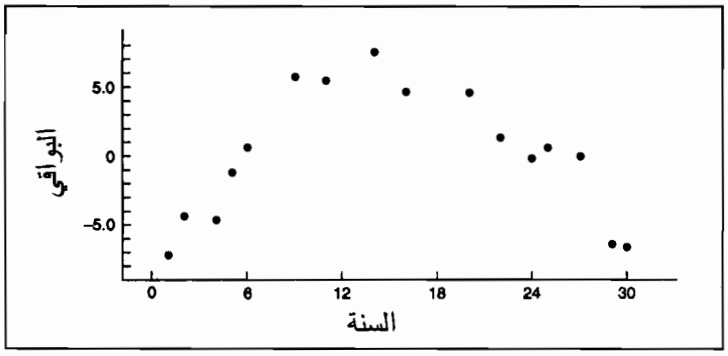
فى هذا الجزء ، نقدم تحليل البواقى بإستخدام مثال (٩-١)، (٩-٢) كما هو موضح. وتحليل البواقى هو أداة هامة فى محاولة لتحسين معادلة الإنحدار المقدرة. نقدم تحليل البواقى هنا ونوضح الموضوع بعمق أكبر فى فصل (١٠) .

ففى الجزء (٩-٤-١) ، ناقشنا الفروض الأساسية لإستنتاجات الإنحدار. الفرض الأول هو أن النموذج الخطى البسيط يمثل بشكل دقيق الإرتباط بين متغير الإستجابة response ومتغيرات التنبؤ predictor . إذا كان هذا الفرض صحيح، يكون شكل البواقى (على المحور الرأسى) مقابل قيم X المناظرة (على المحور الأفقى) يجب أن لا يظهر نموذج مميز. سبب ذلك أنه إذا كان النموذج صحيح، فإن البواقى تمثل أخطاء عشوائية بحتة؛ هكذا فإن، البواقى يجب ألا تظهر نموذج ما على الإطلاق عندما ترسم مقابل أى متغير. إذا ما تم ظهور نموذج ما، فإن البواقى ربما لا تمثل أخطاء عشوائية بحتة. تبعاً لذلك، هذا الفرض الأساسى ربما لا يكون صحيحاً فى بعض الحالات.



شكل (٩-١٦)

يوضح البواقى مقابل درجة الحرارة فى المثال (٩-١)



شكل (٩-١٧)

يوضح البواقى مقابل سنوات الخبرة فى المثال (٩-٢)

فى شكل (٩-١٦) تظهر البواقى لخط المربعات الصغرى لمعادلة (٩-١) (الطاقة اليومية المستخدمة مقابل درجات الحرارة اليومية) تم رسمها مقابل درجات الحرارة المناظرة. بالمثل، فى شكل (٩-١٧) ، البواقي لمثال (٩-٢) (الأجر مقابل سنوات الخبرة) تم رسمها مقابل سنوات الخبرة المناظرة. دعنا نبدأ مع شكل (٩-١٦). حيث يظهر نموذج غير مميز للبواقى بالنسبة لقيم درجات الحرارة. بتعبير آخر، يظهر عدم وجود إرتباط بين البواقي وقيم درجات الحرارة. لكن الآن تفحص البواقى فى شكل (٩-١٧) . يجب أن يكون واضح لك أن نفس الشئ لا يمكن قوله هنا. فهو يظهر نموذج يمكن تمييزه للبواقى بالنسبة لعدد سنوات الخبرة. لاحظ أن البواقى فى أدنى النهاية لمدى سنوات الخبرة تكون سالبة وفى منتصف المدى تكون موجبة وفى النهاية العليا تكون سالبة مرة أخرى. مثل هذا التغير لنموذج على شكل U يقترح وجود إنحناءة فى الإستجابة بالنسبة للزيادة في سنوات الخبرة. وبالتالي نحتاج إلى نموذج يدخل درجة الأنحناء فى العلاقة .

تمارين

(٩-٢٦) فيما يتعلق بالإستنتاج الإحصائى للنموذج الخطى البسيط، ما هى أهم معلمة؟ ولماذا تكون هذه المعلمة هامة ؟

(٩-٢٧) لماذا يكون فرض العينة المثلة **representative sample** للبيانات مهم ؟

(٩-٢٨) هل يمكن لفرض النموذج الخطى البسيط الذي يمثل الأرتباط بين X , Y أن يتم إثبات صحته بمعلومية معامل التحديد r^2 ؟ وضح ذلك .

(٩-٢٩) لبيانات عينة ما ، إفترض أن المربعات الصغرى المقدرة للميل هى $(b_1 = 6.5)$ والخطأ المعيارى المقدر لها هو $(SE(\beta_1) = 1.5)$.

(أ) بدون أى معلومات أخرى، ماذا يمكن استنتاجه بطريقة معقولة عن ميل المجتمع β_1 فى هذه الحالة؟ فسر إستنتاجك .

(ب) ماذا تعنى إجابتك لجزء (a) عن العلاقة الخطية بين X , Y ؟

(٩-٣٠) أجب عن نفس الأسئلة كما فى تمرين (٩-٢٩) إذا كانت $(b_1 = -2.4)$ ، $(SE(b_1) = 2.6)$.

(٩-٣١) أجب عن نفس الأسئلة كما فى تمرين (٩-٢٩) إذا كانت $(b_1 = 12.6)$ ، $(SE(b_1) = 7.0)$

(٩-٣٢) حدد فترة الثقة 95% للميل β_1 فى التمارين التالية وإستخدم كل فترة لتحديد ما إذا كان دليل العينة يخالف إفتراض عدم وجود إرتباط خطى بين X , Y :

(أ) تمرين (٩-١٥) .
(ب) تمرين (٩-٢٠) .
(ج) تمرين (٩-٢٢) .
(د) تمرين (٩-٢٣) .

(٩-٣٣) للتمارين التالية ؛ إلى أى مدى يساند دليل العينة الإعتقاد بأن هناك علاقة خطية موجبة بين X , Y ؟

(أ) تمرين (٩-٢٠) .
(ب) تمرين (٩-٢٣) .

(٩-٣٤) للتمارين التالية ؛ إلى أى مدى يساند دليل العينة الإعتقاد بأن هناك علاقة خطية سالبة بين X , Y ؟

(أ) تمرين (٩-١٥) .
(ب) تمرين (٩-٢٢) .

(٩-٣٥) للتمارين التالية ، إستخدم أسلوب تحليل التباين لإختبار الفرض العدمى بعدم وجود إرتباط خطى بين X , Y . هل يجب أن تكون نتائجك مختلفة عن تلك التى فى تمرين (٩-٢٣) ؟ فسر :

(أ) تمرين (٩-١٥) .
(ب) تمرين (٩-٢٠) .
(ج) تمرين (٩-٢٢) .
(د) تمرين (٩-٢٣) .

(٩-٣٦) بالرجوع إلى تمرين (٩-١٠) .

(أ) إفترض وجود علاقة خطية بين مبلغ تأمين الحياة والدخل، حدد خط المربعات الصغرى وفسر تقديرات الميل والجزء المقطوع من المحور الرأسى .

(ب) بناء على فترة ثقة 95% للميل β_1 ، هل β_1 تم تقديرها بدقة كبيرة؟ وضح .

(ج) بناء على إجراء تحليل التباين، هل دليل العينة يخالف الفرض بعدم وجود علاقة بين مبلغ تأمين الحياة والدخل ؟

(د) بإستخدام تحليل البواقى، هل اكتشفت أى إنتهاكات ملحوظة للفروض؟ وضح .

(٩-٣٧) ارجع إلى تمرين (٩-١١) أجب على أسئلة مماثلة كالتى فى تمرين (٩-٣٦) .

(٩-٣٨) ارجع إلى تمرين (٩-١٣) أجب على أسئلة مماثلة كالتى فى تمرين (٩-٣٦) .

(٩-٣٩) ارجع إلى تمرين (٩-١٤) ، (٩-٢٥) .

(أ) بناء على فترة ثقة 95% لقيمة β_1 ، هل β_1 تم تقديرها بدقة كبيرة؟ وضح .

(ب) إلى أى مدى يخالف دليل العينة الإدعاء بعدم وجود علاقة خطية بين نسبة الضرائب المسددة والنمو السنوى فى الدخل ؟

(ج) بإستخدام تحليل البواقى، هل اكتشفت أى نموذج ملحوظ؟ وضح .

(د) من خلال إجابتك للاجزاء (أ) إلى (ج) هل يمكنك إستنتاج أن خط المربعات الصغرى مناسب أو ملائم للتقدير والتنبؤ ؟ وضح .

(٩-٤٠) قام مدير إدارة الموظفين بعمل إختبار ذكاء لكل ممثلى المبيعات الجدد. إدارة المبيعات كان إهتمامها متعلق بمدى قدرة الإختبار على التنبؤ بحجم المبيعات النهائى. الآتى هو درجات الإختبار والمبيعات الأسبوعية (بآلاف الدولارات) لثمانية من ممثلى المبيعات :

درجات الأختبار	المبيعات
55	8
60	12
85	28
75	24
80	18
85	16
65	15
55	12

(أ) عين متغير الاستجابة والمتغير المفسر وبرر إجابتك .

(ب) كون شكل الإنتشار وحدد ما إذا كان الإرتباط الخطى واضحاً .

(ج) مفترضاً الإرتباط الخطى، حدد خط المربعات الصغرى وفسر تقديرات الميل والجزء المقطوع من المحور الرأسى .

(د) بناء على فترة الثقة 95% لـ β_1 ، هل يمكنك القول بأن β_1 تم تقديرها بدقة كبيرة؟ وضح .

(٩-٤١) إذا كانت Express Graphics شركة طباعة تطبع عبوات (صناديق) مختلفة بشكل كبير مثل علب للسجائر، صناديق مطهرات (منظفات) ، صناديق لمستحضرات التجميل، مستحضرات صيدلية، وللوجبات السريعة. الأعمال تختلف فى الحجم، الشكل، جودة الطابعة ، ونوع الورق المستخدم. وتقوم الشركة بعمل حوالى 200 عملية شهرياً. المدير مهتم بدرجة التوافق بين سعر البيع المستهدف (X) للعملية والمبلغ النهائي لفاتورة الحساب (Y) وفيما يلى بيانات عينة عشوائية لعدد 15 عملية :

السعر المستهدف (X)	المبلغ المدون بالفاتورة (Y)	السعر المستهدف (X)	المبلغ المدون بالفاتورة (Y)
$ 551	$ 328	$ 5,292	$ 6,417
469	543	83	85
1,882	2,577	2,336	2,178
545	404	123	127
13,596	15,090	4,285	4,349
929	292	76	115
633	1,045	125	381
		44	122

(أ) كون شكل الإنتشار . هل هناك علاقة بين مبلغ الفاتورة والسعر المستهدف ، وهل العلاقة فى شكل خط مستقيم ؟ وضح .

(ب) إفترض العلاقة الخطية، حدد خط المربعات الصغرى وفسر تقديرات الميل والجزء المقطوع من المحور الرأسى .

(ج) حدد فترة ثقة 95% للميل β_1 . هل تعتقد أن الميل المقدر فى جزء (ب) دقيق بشكل كاف ؟ وضح .

(د) ماذا يخبرك تحليل البواقى لهذه المشكلة ؟ وضح .

(٩-٤٢) قام مسئول فى شركة طباعة بتطوير سعر مستهدف لعملية محتملة والتى يمكن أن تنتقل إلى العميل. المتغير الأساسى الذى يؤثر فى تكلفة الإنتاج هو سرعة آلة الطباعة فى العمل. فى الماضى، سرعة الآلة كان يتم تقديرها على أساس الخبرة الشخصية فى اجتماعات تسعير أسبوعية . هذا التقدير، بجانب تقديرات أخرى يستخدم كمدخلات لبرنامج الكمبيوتر لتقدير تكلفة الإنتاج. فكر محلل السوق فى تحديد أى التقديرات هى الأفضل لسرعة آلة الطباعة يمكن الحصول عليها بإستخدام نموذج الإنحدار. وفى تحليل أولى، قام بجمع البيانات التالية لعينة من 15 عملية، سرعة آلة الطباعة (مئات الصور لكل ثانية) وصعوبة التدوين أو التسجيل registration فى طباعة الصندوق (تصنيف 1 = صعب ، 2 = عادى ، 3 = سهل ، تم تحديدها بالتقدير الشخصى) تم تسجيلها :

صعوبة التدوين	سرعة آلة الطباعة	صعوبة التدوين	سرعة آلة الطباعة
3	107	1	74
3	95	2	69
3	104	2	71
3	45	3	67
2	69	3	109
2	100	3	114
2	99	3	94
		3	120

(أ) حدد متغير الاستجابة والمتغير المفسر وبرر اختيارك .

(ب) أجب على الأجزاء (ب) إلى (د) فى تمرين (٩-٤٠) .

(ج) بماذا تقترح معادلتك المقدرة هل يوجد فرق بين سرعة آلة الطباعة للوظائف المصنفة1 (صعب) والوظائف 2 (عادية) ، فى المتوسط ؟

(٩-٤٣) يرغب مدير مؤسسة ما فى تحديد تأثير المدد الزمنية المختلفة على عدد الوحدات التى يتم تجميعها عن طريق مشغلى نظام تجميع قبل أن يأخذ المشغلين فترة راحة. وقد شك المدير فى أن فترات العمل الأطول قبل فترة الراحة تتجه لتخفيض الإنتاجية. وفى تجربة على فترات العمل تم تجربة فترات راحة 1، 2 ، 3 ، 4 ساعات لكل فترة، تم ملاحظة عدد الوحدات المجمعة لأربعة مشغلين . تم ملاحظة النتائج التالية :

الساعات قبل فترة الراحة	1	2	3	4
الوحدات المجمعة	25,29,23,31	55,65,63,54	73,75,74,71	90,88,91,87

(أ) أجب عن الأجزاء (أ) - (د) لتمرين(٩-٤٠) .

(ب) مستخدماً تحليل البواقى، هل تكتشف أى نموذج ملحوظ؟ وضح .

(ج) هل تعتقد أن عينة البيانات هذه تبرهن بوضوح أن إنتاجية نظام التجميع تعتمد فعلاً على المدة الزمنية قبل أن يأخذ المشغل فترة راحة؟ فسر إعتقادك .

(٩-٥) درجة الإعتماد على التقديرات والتنبؤات : The Reliability of Estimates and Predictions

بشكل عام إذا أخفق الإحصاء T أو تحليل التباين فى تقدير أن Y مرتبطة خطياً مع X ، في هذه الحالة يجب الا يستخدم خط المربعات الصغرى للتقدير أو التنبؤ. وحتى لوكانت هذه الأساليب تساند وتدعم وجود علاقة خطية، فمن الممكن ألا يكون خط المربعات الصغرى مقنع للتقدير أو التنبؤ .

ولما كان الهدف الأساسى من تحليل الإنحدار هو تقديم نموذج يمثل الوضع الفعلي بدرجة معقولة ويوفق البيانات الممثلة إلى الحد الذى يكون فيه تباين الأخطاء العشوائية صغيراً بدرجة كافية. مثل هذا النموذج يقدم درجة دقة مقبولة عندما يستخدم للتقدير أو التنبؤ. لذلك، فإن الهدف البديل يكون في تقديم نموذج يحقق درجة دقة مقنعة فى تقديراته وتنبؤاته. مع بقاء هذا الهدف فى الذهن، نركز إهتمامنا على الأخطاء المعيارية لمقدرات الإنحدار وحدود الخطأ المصاحبة لهم .

(٩-٥-١) درجة الإعتماد على b_1 فى تقييم العلاقة الخطية بين X , Y

The Reliability of b_1 in Assessing the Linear Relationship Between Y and X

كما أشرنا سابقاً ، فإن الميل β_1 لخط إنحدار المجتمع هو المحدد الرئسى للعلاقة الخطية بين X , Y . لذلك من المهم معرفة دقة b_1 ، والتى تعتبر أفضل مقدر للاستدلال حول β_1 . وبالرجوع لمثال (٩-٥) (التعامل مع مشكلة تقييم الملكية) ، اكتشفنا نقص واضح فى درجة دقة تقدير β_1 ، حد الخطأ في ($b_1 = 6.00$) كان 5.04 ؛ لذلك كان مدى القيم المقبولة للمقدار β_1 واسع بشكل غير مقبول. هذا النقص فى درجة الدقة ناتج مباشرة من أن الخطأ المعيارى لقيمة b_1 كبير. ونلاحظ هنا أنه كلما كبر الخطأ المعيارى، كلما كبر حد أو هامش خطأ المعاينة، وكلما كبر مدى قيم β_1 المقبولة .

وعموما فإنه لأى درجة دقة فى تقدير β_1 ، فإن الخطأ المعيارى لقيمة b_1 يجب أن يكون أقل كثيراً من حجم قيمة b_1 . بتعبير آخر فإن نسبة قيمة b_1 إلى خطأها المعيارى يجب أن تكون كبيرة تماماً فى الحجم. لكن لاحظ أن هذه النسبة هى ببساطة القيمة T $\{T = b_1 / SE(b_1)\}$ للفرض العدمى ($H_0: \beta_1 = 0$) . لذلك ، كلما كبر مقدار القيمة T ، تكون درجة الدقة أفضل فى تقدير الميل β_1 .

وجدول (٩-٥) يوضح قيم b_1 وأخطائهم المعيارية، وقيم T للأمثلة الأربعة المستخدمة فى هذا الفصل. لاحظ أن أفضل درجة دقة فى تقدير الميل b_1 تتحقق فى مثال الطاقة مقابل درجة الحرارة لأن القيمة T لها أكبر مقدار مطلق ($T = -24.84$) . وهذا الأمر لايدهشنا كثيراً إذا قارنا شكل الإنتشار لمثال الطاقة مقابل درجة الحرارة (شكل٩-٦) مع أشكال الأمثلة الأخرى ، أشكال (٩-٧)، (٩-١٠)، (٩-١٣) .

جدول (٩-٥)

مقارنة الأخطاء المعيارية والقيم T للأمثلة الأربعة

مثال	قيمة b_1	الخطا المعياري	قيمة T
تثمين العقارات أو تقييم الملكية	6.0	1.5832	3.79
مطعم عش البلبل Bird's Nest	.2	.0470	4.26
الأجر مقابل الخبرة	1.2607	.1279	9.86
الطاقة مقابل درجة الحرارة	-3.4641	.1395	-24.84

كما تعلم ، يمكن إستخدام فترات الثقة لوصف درجة دقة المقدرات. لأحصاء المربعات الصغرى b_1، فترة ثقة $\{100(1-\alpha)\%\}$ للمقدار β_1 تم تقديمها فى معادلة (9.18) وهامش خطأ المعاينة تم تقديمه فى معادلة (9.19) . لمثال الأجر مقابل الخبرة، تكون فترة ثقة 95% للمعلمة β_1 هى {القيمة الجدولية هى $(t_{.975,14} = 2.145)$ } .

$$1.2607 \pm (2.145)(.1279) = 1.2607 \pm .2743$$

أو (.99, 1.54) . هذه الفترة تعنى أنه عندما يزيد عدد سنوات الخبرة بسنة واحدة، متوسط الأجر ربما يزيد بكمية قليلة 990$ (إذا كانت $\beta_1 = .99$) أو يزيد بكمية كبيرة 1,540 $ (إذا كانت $\beta_1 = 1.54$) بثقة 95% . ويكون مدى هذه الفترة أقل بكثير من الموجود فى مشاكل تقييم الملكية ومطعم عش البلبل Bird's Nest .

فيما يتعلق بمثال الطاقة مقابل درجة الحرارة، تكون فترة ثقة 95% للمعلمة β_1 هى {القيمة الجدولية هى $(t_{.975,16} = 2.120)$} .

$$-3.4642 \pm (2.120)(.1395) = -3.4642 \pm .2957$$

أو (-3.76, -3.17) . هذه الفترة تعنى أنه عندما تزيد درجة الحرارة اليومية بدرجة واحدة مئوية، فإن متوسط الطاقة المستخدمة ربما ينخفض بمقدار كبير مثل3.76 كيلو وات (إذا كانت $\beta_1 = -3.76$) أو بمقدار صغير مثل 3.17 كيلو وات (إذا كانت $\beta_1 = -3.17$) بثقة 95% . صغر أو ضيق المدى لهذه الفترة بالمقارنة بالموجود فى الثلاث أمثلة الأخرى لا يترك شك أن ميل المجتمع β_1 لمثال الطاقة مقابل درجة الحرارة تم تقديره بأفضل درجة دقة .

(٩-٥-٢) تقدير متوسط Y ، بمعلومية X : Estimating the Mean of Y , Given X

كما سبق أن أشرنا فى الجزء (٩-٢-٣) فإن هناك استخدامين أوليين لنماذج الإنحدار : (1) لإكتساب فكرة عن عملية تحليل الانحدار عن طريق دراسة العلاقة بين متغيرات الانحدار. (2) للتنبؤ أو التقدير. والإستخدام الأخير يتكون من أحد حالتين: فى الحالة الاولى، نرغب فى التنبؤ بقيم Y الفردية بمعلومية أن X تساوى بعض القيم الخاصة، التى تشير إليها بالرمز x . فى الحالة الأخرى، نرغب فى تقدير متوسط قيمة Y ، بمعلومية أن X تساوى x . لتمييز أفضل بين هاتين الحالتين ، نستخدم المصطلح تنبؤ prediction ليشير إلى التنبؤ بقيم Y الفردية وتقدير estimation ليشير إلى التقدير لمتوسط قيمة Y .

هناك حاجة لتقدير الأساليب التى ذكرت سابقاً . ففى مثال تثمين العقارات ، إفترض أننا نريد تقدير متوسط سعر البيع لكل المنازل المتشابهة والتى مساحتها 2,200 قدم مربع . نعوض ببساطة عن (X = 2.2) فى معادلة الإنحدار ($\hat{Y} = .8 + (6.0)(2.2)$). هكذا حددنا متوسط سعر البيع المقدر لكى يكون $\hat{Y} = .8 + (6.0)(2.2) = 14.0$ أى يكون 140000$. من الضرورى إدراك درجة الدقة لتقدير ما قبل إستخدامه في إتخاذ قرارات هامة . دعنا نقوم بفحص الدقة التى يمكن توقعها للتقدير 140,000$ لمتوسط سعر البيع للمنازل المتشابهة التى مساحتها 2,200 قدم مربع . درجة الدقة لأى تقدير تتحدد عن طريق هامش خطأ المعاينة له أو ، بمعنى مساو ، فترة الثقة المصاحبة . فى كلا الأمرين ، تكون الخطوة الأولى هى تحديد الخطأ المعيارى للمقدر .

ويمكن أن يتم التوضيح جبرياً أن الخطأ المعيارى لقيمة $\hat{Y}$ ، بمعلومية أن (X = x) ، يعطى عن طريق :

$$SE(\hat{Y}) = \sqrt{\sigma_\varepsilon^2\left[\frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.25)}$$

وحيث أن تباين الخطأ σ_ε^2 غير معلوم ، يتم تقديره بإستخدام تباين البواقى S_e^2 وهكذا ، الخطأ المعيارى المقدر للمقدار $\hat{Y}$ ، بمعلومية أن X = x يكون :

$$SE(\hat{Y}) = \sqrt{S_e^2\left[\frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.26)}$$

الآن ، المقدر ($\hat{Y} = b_0 + b_1X$) يمكن توضيحه على أنه متغير عشوائى يتبع التوزيع الطبيعى* . وحيث أننا قدرنا σ_ε^2 بقيمة S_e^2 ، فإن الإستنتاجات عن متوسط المجتمع Y عندما تكون (X = x) يكون مبنى على أساس توزيع T بدرجات حرية n-2 . وبالتالى فإن فترة ثقة (100(1-α)%) لمتوسط قيم Y عندما تكون (X = x) يكون كما يلى :

$$\hat{Y} \pm t_{1-\alpha/2,\, n-2} \sqrt{S_e^2\left[\frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.27)}$$

حيث هامش أو حد الخطأ Margin error يساوى :

$$t_{1-\alpha/2,\, n-2} \sqrt{S_e^2\left[\frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.28)}$$

نعود الآن لمثال تقييم الملكية لتحديد الدقة المتوقعة لتقديرنا 140,000$ لمتوسط سعر البيع للمنازل المتشابهة مع (X = 2.2) (2,200 قدم مربع) . لقد وجدنا سابقاً أن ($\overline{X} = 2.0$) ، ($S_e^2 = 6.2667$) ، ($SS\{X\} = 2.5$) ، n = 5 . بالتالي الخطأ المعيارى المقدر لـ $\hat{Y}$ ، بإستخدام معادلة (9.26) هو :

$$SE(\hat{Y}) = \sqrt{(6.2667)\left[\frac{1}{5} + \frac{(2.2 - 2.0)^2}{2.5}\right]} = 1.1634$$

من المعادلة (9.27) ، فترة ثقة 95% لمتوسط سعر البيع لكل المنازل مع (X = 2.2) هو : (لاحظ أن $t_{.975,3} = 3.18$)

$$14.0 \pm (3.182)(1.1634) = 14.0 \pm 3.702$$

* السبب: رأينا فيما سبق أن b_1 تتوزع طبيعياً. بأسلوب مشابه، bo يمكن توضيح أنها تتوزع توزيعا طبيعياً . حيث ان X ثابتة، ($\hat{Y} = b_0 + b_1X$) هي توليفة خطية لمتغيرين عشوائيين طبيعيين، b_1X, b_0 وهكذا، $\hat{Y}$ تكون أيضاً موزعة توزيعاً طبيعيا، بمعلومية X.

أو (10.298 , 17.702) يمكننا القول أنه بدرجة ثقة 95% فإن متوسط سعر البيع للمنازل التى مساحتها 2,200 قدم مربع يمكن أن يكون منخفض بمقدار 102,980$ أو مرتفع بمقدار 177,020$. هل يشير هذا إلى درجة دقة جيدة ؟ يبدو هنا مدى واسع تماماً للقيم المقبولة لمتوسط سعر البيع. لذلك فإن المتوسط المقدر وهو 140,000$ لا يمكن إعتباره دقيق جداً .

دعنا ننظر الآن إلى مثال استهلاك الطاقة. إفترض أننا نرغب فى تقدير متوسط الإستهلاك عندما تكون درجة الحرارة اليومية العالية هى 3 درجات سليزية. وجدنا أن خط المربعات الصغرى لهذا المثال هو ($\hat{Y} = 87.0 - 3.46X$) ؛ وهكذا فإن متوسط الطاقة المقدرة المستخدمة لدرجة الحرارة العالية 3 درجات هى ($\hat{Y} = 87.0 - 346(3) = 76.62$) . لتحديد الخطأ المعيارى لهذا التقدير، نحسب $\overline{X}$, SS(X) بإستخدام قيم درجات الحرارة الموجودة فى مثال (٩-١). ينتج من هذا أن ($\overline{X} = 3.2222$) ، ($SS(X) = 343.6111$) . ولقد سبق أن علمنا من جدول (٩-٣) أن :

($S_e^2 = \frac{107}{16} = 6.6875$)، ($n = 18$) لهذا المثال. لذلك ، الخطأ المعيارى المقدر يكون :

$$SE(\hat{Y}) = \sqrt{(6.6875)\left[\frac{1}{18} + \frac{(3-3.2222)^2}{343.6111}\right]} = .6103$$

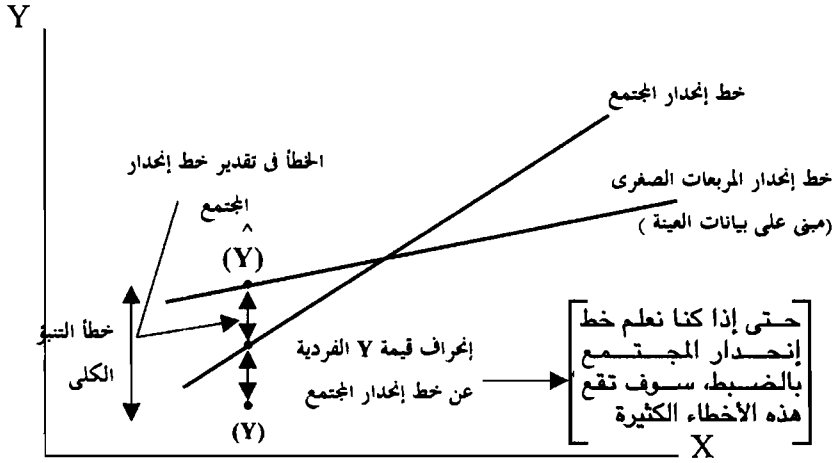
وبإستخدام فترة الثقة 95% لمتوسط استهلاك الطاقة عندما تكون درجة الحرارة اليومية العليا 3 درجات سليزية هى (حيث $t_{.975,16} = 2.120$) .

$$76.62 \pm (2.120)(.6103) = 76.62 \pm 1.29$$

أو (75.33 , 77.91) . وبمستوى ثقة 95% ، نستنتج أن متوسط إستهلاك الطاقة عندما تكون درجة الحرارة العليا اليومية هى 3 درجات سليزية، يمكن أن يكون منخفض إلى المقدار 75.33 أو مرتفع إلى المقدار 77.91 كيلو وات. هذا المدى يبدو صغير حقاً، وهكذا يدل على وجود دقة كبيرة فى التقدير 76.62 كيلو وات .

(٩-٥-٣) التنبؤ بقيم Y الفردية بمعلومية X : Predicting an Individual Y - Value , Given X

بمعلومية قيمة X ، فإن التنبؤات **predictions** بقيم Y الفردية تكون مماثلة لتقدير متوسط **estimates** قيمة Y المناظرة للقيمة X . كلاهما يتم تحديده عن طريق التعويض بقيمة X فى معادلة الإنحدار المقدرة ($\hat{Y} = b_0 + b_1X$) . ومع ذلك فإن أخطاء التنبؤ prediction لقيم Y الفردية يكون أكبر من التى تكون للمتوسط. السبب ليس صعب الفهم. فى كلا الحالتين، التنبؤ أو التقدير **prediction or estimate** يكون للقيمة $\hat{Y}$ لخط المربعات الصغرى المناظرة لقيمة X المعطاة. حيث أن خط المربعات الصغرى من المحتمل ألا يمثل خط إنحدار المجتمع تماماً، يكون نتيجة هذا الخطأ. هذا الخطأ هو نفسه عند التنبؤ بقيم Y الفردية وعند تقدير متوسط المجتمع. ولكن التنبؤ بقيم Y الفردية تكون معرضة لخطأ إضافى أيضاً، لأن قيمة Y الفردية من المحتمل ألا تقع على خط إنحدار المجتمع مباشرة. إذا كنا نعلم خط الانحدار بدقة تامة، فإنه يمكن أن نقدر متوسط Y لهذه القيمة X بدون خطأ، لكننا لن نكون قادرين على التنبؤ بقيمة Y الفردية بشكل تام. هذا هو سبب أن الأخطاء تكون أكبر عندما نتنبأ بقيم Y الفردية **individual** منها عندما نقدر متوسط **mean** المجتمع للمتغير Y بمعلومية قيمة معينة X . شكل (٩-١٨) يوضح هذه النقطة .



شكل (٩-١٨)

شكل يوضح عنصرى أخطاء التنبؤ (1) التقدير غير الصحيح لخط إنحدار المجتمع (2) إنحراف نقطة البيانات الفردية عن خط إنحدار المجتمع

دعنا نرى كيف أن مصدر الخطأ الاضافي تؤثر فى الخطأ المعيارى للتنبؤات لقيم Y الفردية. تعلمنا من جزء (٩-٥-٢) أن تباين الوسط المقدر للمتغير Y ، بمعلومية X ، هو :

$$Var(\hat{Y}) = \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]$$

نعلم أيضاً أن تباين قيم Y الفردية حول خط إنحدار المجتمع هو :

$$V.r\ (\varepsilon) = \sigma_\varepsilon^2$$

يمكن توضيح رياضياً أن تباين أخطاء التنبؤ لقيم Y الفردية هو مجموع هذين التباينين – أى أن :

$$\sigma_\varepsilon^2 + \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]$$

بأخذ σ_ε^2 عامل مشترك ، يمكننا تبسيط هذا إلى :

$$\sigma_\varepsilon^2\left[1 + \frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right] \qquad \textbf{(9.29)}$$

والخطأ المعيارى للتنبؤ بقيم Y الفردية ، عندما (X = x) يتم إيجاده بأخذ الجزر التربيعى لمعادلة (9.29) . ولكى نميزه عن الخطأ المعيارى لتقدير المتوسط للمتغير Y بمعلومية (X = x) [الذى يرمز له (SE{ $\hat{Y}$ })]، هذا الخطأ المعيارى يرمز له $SE_p(\hat{Y})$. الدليل p يدل على التنبؤ بقيم Y الفردية . لذلك فإن :

$$SE_p(\hat{Y}) = \sqrt{\sigma_\varepsilon^2\left[1 + \frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.30)}$$

ولقد سبق أن أوضحنا أن تباين الخطأ σ_ε^2 يتم تقديره عن طريق تباين البواقى S_e^2 ؛ لهذا فإن الخطأ المعيارى المقدر يكون :

$$SE_p(\hat{Y}) = \sqrt{S_e^2\left[1 + \frac{1}{n} + \frac{(x - \overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.31)}$$

بمقارنة الصيغ (9.26) ، (9.31) ، يجب أن ترى أن الخطأ المعيارى للتنبؤ بقيم Y الفردية أكبر من الخطأ المعيارى لتقدير متوسط قيمة Y ، بمعلومية X . هذا يوضح لنا أننا نتوقع درجة دقة أقل فى التنبؤ بقيم Y الفردية عنه فى تقدير متوسط Y ، بمعلومية أن (X = x) .

لقياس درجة دقة التنبؤات، نحسب فترة تقدير اعتماداً على $SE_p(\hat{Y})$. نسمى هذه الفترة فترة تنبؤ **prediction interval** لنميزها بوضوح عن فترة الثقة لمتوسط mean (Y) المعطى فى معادلة (9.27)، وتكون فترة التنبؤ بدرجة ثقة (100(1- α)%) لقيمة Y الفردية ، وبمعلومية (X = x) كالتالى :

$$\hat{Y} \pm t_{1-\alpha/2,\, n-2}\sqrt{S_e^2\left[1+\frac{1}{n}+\frac{(x-\overline{X})^2}{SS(X)}\right]} \qquad \textbf{(9.32)}$$

ففى مثال تقييم الملكية (تثمين العقارات) ، إفترض أننا نريد التنبؤ بسعر بيع منزل واحد حجمه 2,200 قدم مربع (X = 2.2) . التنبؤ يكون تماماً مثل تقدير متوسط سعر البيع عندما X = 2.2 أى يكون($\hat{Y}$= .8+(6)(2.2)=14.0) أو$140,000 . لكن الخطأ المعيارى لهذا التقدير يكون أكبر. تذكر أن ($\overline{X}$= 2.0)، (SS(X) = 2.5) , (S_e^2 = 6.2667) , (n=5) نحسب الخطأ المعياري المقدر لهذا التنبؤ كمايلي :

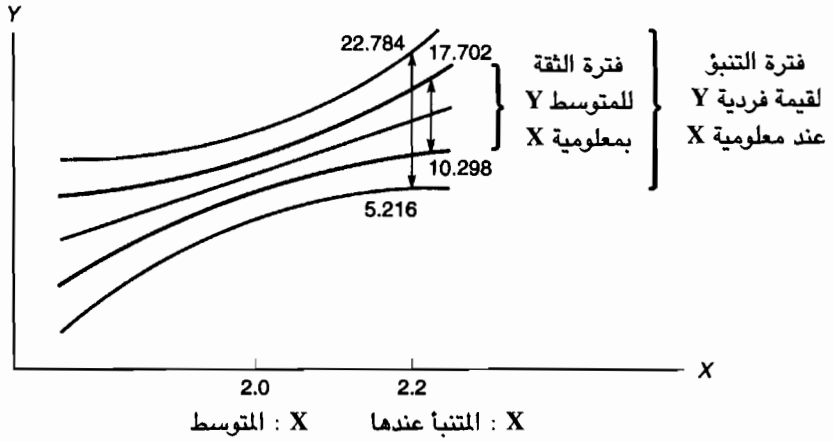
$$SE_p(\hat{Y}) = \sqrt{(6.2667)\left[1+\frac{1}{5}+\frac{(2.2-2.0)^2}{2.5}\right]} = 2.7605$$

وهو بشكل أساسى أكبر من الخطأ المعيارى المقدر من متوسط سعر البيع عندما (X = 2.2) ، {SE($\hat{Y}$) = 1.1634}. بإستخدام المعادلة (9.32) ، نجد أن فترة تقدير 95% لقيمة Y الفردية، بمعلومية (X = 2.2) هى :

$$14.0 \pm (3.182)\,(2.7605) = 14.0 \pm 8.784$$

أو (22.784 , 5.216) . للمنزل الفردى الذى مساحته 2,200 قدم مربع ، سعر البيع يمكن أن يكون منخفض إلى المقدار $52,160 أو مرتفع إلى المقدار $227,840 بدرجة ثقة 95%. من الواضح أن ، هذا المدى واسع جداً لكى يمكننا من الإعتماد بشكل جاد على التنبؤ $140,000 .

نستخدم التنبؤ الآن في مثال تقييم الملكية لتوضيح العلاقة بين فترات التقدير **prediction intervals** لقيم Y الفردية وفترات الثقة لمتوسط **mean** قيمة Y ، بمعلومية X ؛ انظر شكل (٩-١٩).



شكل (٩-١٩) مقارنة فترات الثقة لمتوسط Y ،
وفترات التقدير لقيمة Y الفردية ، بمعلومية X على مدى قيم X

«حزمة أو نطاق الثقة **confidence bands**» تمثل الفترات التى تناظر مدى قيم X. هناك فترتين عند X=2.2 حددت سابقا في هذا البند، تم ايضاحهما على الرسم.

لاحظ مايلي (1) فترة التنبؤ تكون أوسع دائماً من فترة الثقة المناظرة لها عند المتوسط، (2) الفترات تصبح أكثر إتساعاً كلما اعتبرنا قيم X أبعد عن المركز والسبب فى ذلك موضح فى جزء (٩-٦).

مثال (٩-٦)

(أ) بالرجوع لمشكلة مطعم عش البلبل Bird's Nest ، حددنا أن ($\hat{Y}$= 3.5 + .2X) ، (S_e^2 = 1.875) ، (SS(X) = 850) ، ($\overline{X}$ = 30) ، (n = 6) حدد فترة ثقة 95% لمتوسط الإيراد Y وكذلك فترة تنبؤ 95% لإيراد فترة مسائية فردية، بمعلومية أنه فى كلتا الحالتين (X = 30) من الزبائن. فسر نتائجك.

(ب) بالرجوع لمثال الأجر مقابل الخبرة، علمنا من جدول (٩-٤) أن ($\hat{Y}$ = 28.9 + 1.26X)، (S_e^2 = 352.3/14 = 25.1643) ، وحددنا سابقاً أن (SS(X) = 1,539.4375) ، ($\overline{X}$ = 15.3125) ، (n = 16). المطلوب تكوين 95% فترة ثقة لمتوسط الأجر Y وكذلك فترة تنبؤ 95% بأجر معين فردى، اذا علمت أن (X = 15) سنة خبرة.

الحل

(أ) التقدير لمتوسط الإيراد عندما (X = 30) هو ($\hat{Y}$ = 3.5 + .2(30) = 9.5) أو 950$. الخطأ المعيارى لهذا التقدير هو:

$$SE\ (\hat{Y}) = \sqrt{(1.875)\left[\frac{1}{6} + \frac{(30-30)^2}{850}\right]} = .5590$$

وبالتالى فإن فترة ثقة 95% لمتوسط الإيراد هى ($t_{.975,4}$ = 2.776)

$$9.5 \pm (2.776)\ (.5590) = 9.5 \pm 1.552$$

أو (7.95 , 11.05). نستطيع أن نستنتج بدرجة ثقة 95% أن متوسط الإيراد عندما (X = 30) متردد يمكن أن يكون الإيراد منخفض إلى المقدار 975$ أو مرتفع إلى المقدار 1,105$.

الإيراد المتنبأ به لفترة مسائية عند (X = 30) متردد هو نفسه متوسط الإيراد المقدر – أى أن ($\hat{Y}$ = 9.5) (أو 950$). لكن الخطأ المعيارى لهذا التنبؤ يكون أكبر بكثير.

$$SE_P(\hat{Y}) = \sqrt{(1.875)\left[1 + \frac{1}{6} + \frac{(30-30)^2}{850}\right]} = 1.4790$$

هكذا تكون فترة التنبؤ 95% بالإيراد لفترة مسائية واحدة مع 30 عميل هى:

$$9.5 \pm (2.776)\ (1.4790) = 9.5 \pm 4.11$$

أو (5.39 , 13.61). ومن ثم نستنتج بدرجة ثقة 95% أن إيراد الفترة المسائية الفردية يمكن أن يكون منخفض بمقدار 539$ أو مرتفع إلى المقدار 1,361$، بمعلومية أن (X = 30) عميل.

(ب) متوسط الأجر المقدر عندما (X = 15) سنة خبرة هو :

($\hat{Y}$= 28.9 + (1.26)(15) = 47.8) أو $47,800 . ويكون الخطأ المعيارى لهذا التقدير هو:

$$SE\ (\hat{Y}) = \sqrt{(25.1643)\left[\frac{1}{16} + \frac{(15-15.3125)^2}{1,539.4375}\right]} = 1.2547$$

وفترة الثقة 95% لمتوسط الأجر هى: ($t_{.975,14}$ = 2.145) .

$$47.8 \pm (2.145)(1.2547) = 47.8 \pm 2.69$$

أو (45.11 , 50.49) . ونستنتج بدرجة ثقة 95% وعندما تكون (X = 15) سنة من الخبرة، ان متوسط الأجر يمكن أن يكون منخفض بمقدار $45,110 أو مرتفع بمقدار $50,490 .

والتنبؤ بالأجر الفردى مع 15 سنة من الخبرة هو أيضاً ($\hat{Y}$ = 47.8) أو $47,800 . ويكون الخطأ المعيارى :

$$SE_P(\hat{Y}) = \sqrt{(25.1643)\left[1 + \frac{1}{16} + \frac{(15-15.3125)^2}{1,539.4375}\right]} = 5.1709$$

لذلك تكون فترة التنبؤ بدرجة ثقة 95% للأجر الفردى مع 15 سنة من الخبرة كما يلى :

$$47.8 \pm (2.145)(5.1709) = 47.8 \pm 11.09$$

أو (36.71 , 58.89) . ويمكننا إستنتاج أنه بدرجة ثقة 95% ، أن فرداً مع 15 سنة خبرة يمكن أن يكون الأجر منخفض إلى 36,710 أو مرتفع إلى $58,890 .

إستخدام الكمبيوتر :

يمكننا إستخدام الكمبيوتر لتحديد تقديرات وتنبؤات بالإضافة إلى فترات الثقة والتنبؤ، الأمر الفرعى PREDICT x فى برنامج Minitab، حيث x هى قيمة X المرغوبة، يلى الأمر REGRESS يعطينا قيمة $\hat{Y}$, SE ($\hat{Y}$) ، فترة ثقة 95% لمتوسط قيمة Y ، وفترة تنبؤ 95% لقيمة Y الفردية .

نستخدم الأمثلة الأربعة التى ناقشناها على مدى هذا الفصل لتوضيح إستخدام Minitab لهذا الغرض. لكل مشكلة، حددنا ثلاثة قيم للمتغير X : واحدة قريبة من مركز مدى X والاثنين الآخرين قريبين من الحدود (الأطراف) لتقييم الملكية ، (X = 1, 2, 3) ؛ (عش البلبل Bird's Nest) ، (X = 15, 30, 45) ؛ للأجر ، (X = 3, 15, 27) للطاقة (X= -4, 3, 9) . المخرجات تتكون من قيم $\hat{Y}$ ، SE ($\hat{Y}$)، فترات ثقة 95% ، فترات تنبؤ، وهى معطاة لكل مثال فى جداول(٩-٦)، (٩-٧)، (٩-٨)، (٩-٩) على الترتيب . ومن هذه النتائج، يمكن ملاحظة أن أفضل تقدير وتنبؤ من حيث الدقة عندما تكون قيمة X المرغوبة هى نفسها القيمة $\overline{X}$. كلما بعدت قيمة X المطلوبة عن $\overline{X}$ فإن مقدار الدقة فى التقدير والتنبؤ تنخفض .

جدول (٩-٦)

فترة الثقة وفترة التنبؤ لمثال تقييم الملكية عندما X = 3,2,1

Fit	Stdev.	95% C.I.	95% P.I.
6.80	1.94	(0.63,12.97)	(-3.28,16.88)
12.80	1.12	(9.24,16.36)	(4.07,21.53)
18.80	1.94	(12.63,24.97)	(8.72,28.88)

جدول (٩-٧)
فترة الثقة وفترة التنبؤ لمثال مطعم عش البلبل عندما X = 15 , 30 , 45

Fit	Stdev.	95% C.I.	95% P.I.
6.500	0.899	(4.002,8.998)	(1.950,11.050)
9.500	0.559	(7.947,11.053)	(5.392,13.608)
12.500	0.899	(10.002,14.998)	(7.950,17.050)

جدول (٩-٨)
فترة الثقة وفترة التنبؤ لمثال الأجر عندما X = 3 , 15 , 27

Fit	Stdev.	95% C.I.	95% P.I.
32.73	2.01	(28.41,37.05)	(21.13,44.32)
47.86	1.25	(45.16,50.55)	(36.76,58.95)
62.98	1.95	(58.80,67.17)	(51.44,74.53)

جدول (٩-٩)
فترة الثقة وفترة التنبؤ لمثال الطاقة عندما X = -4 , 3 , 9

Fit	Stdev.	95% C.I.	95% P.I.
100.852	1.177	(98.356,103.349)	(94.828,106.877)
76.603	0.610	(75.309,77.897)	(70.970,82.236)
55.818	1.010	(53.676,57.960)	(49.923,61.704)

(٩-٥-٤) ملخص الاستدلال حول نموذج الإنحدار الخطى البسيط :

إجراءات الإستنتاج الإحصائى للنموذج الخطى البسيط يمكن تلخيصها كما يلى :

ملخص إستنتاجات عن نموذج الإنحدار الخطى البسيط

- لإختبار ($H_0: \beta_1 = 0$) إستخدم :

(1) الأحصاء T : $T = \dfrac{b_1 - 0}{SE(b_1)}$

حيث:

$$SS(X) = \sum X_i^2 - \left[(\sum X_i)^2 / n\right] \quad , \quad SE(b_1) = \sqrt{S_e^2 / SS(X)}$$

أو

(2) تحليل التباين: الأحصاء F : $F = \dfrac{MSR}{MSE}$

- فترة ثقة $(100\{1-\alpha\}\%)$ لميل المجتمع β_1 هى:

$$\hat{Y} \pm t_{1-\alpha/2,\, n-2}\ SE(b_1)$$

حيث :

$$SE(b_1) = \sqrt{\frac{S_e^2}{SS(X)}}$$

- فترة ثقة $(100\{1-\alpha\}\%)$ لمتوسط Y ، بمعلومية $(X = x)$ ، هو :

$$\hat{Y} \pm t_{1-\alpha/2,\, n-2}\ SE\ (\hat{Y})$$

حيث :

$$SE\ (\hat{Y}) = \sqrt{S_e^2\left[\frac{1}{n} + \frac{(x-\overline{X})^2}{SS(X)}\right]}$$

- فترة تنبؤ $(100\{1-\alpha\}\%)$ لقيمة Y الفردية، بمعلومية $(X = x)$ ، هى :

$$\hat{Y} \pm t_{1-\alpha/2,\, n-2}\ SE_p(\hat{Y})$$

حيث :

$$SE_P(\hat{Y}) = \sqrt{S_e^2\left[1 + \frac{1}{n} + \frac{(x-\overline{X})^2}{SS(X)}\right]}$$

(٩-٦) العوامل التى تؤثر فى الأخطاء المعيارية للإنحدار : بعض إعتبارات التصميم

Factors that Affect Regression Standard Error : Some Design Considerations

إذا كانت لدينا الفرصة لرسم مجموعة من بيانات العينة، هل هناك أى شئ يمكننا عمله لتأكيد الدقة المقبولة لتقديرات الإنحدار المختلفة؟ للإجابة على هذا السؤال، نكون بحاجة لفحص العوامل التى تؤثر على الأخطاء المعيارية لمقدرات الإنحدار. سيكون هناك اهتمام خاص بالأخطاء المعيارية لكل من: لمقدر الميل b_1 ، قيمة المتوسط المقدر لـ Y ، بمعلومية $(X = x)$ ، القيمة المتنبأ بها لقيمة Y الفردية ، بمعلومية $(X = x)$. المعادلات لهذه الأخطاء المعيارية سنعيد تقديمها مرة أخرى :

$$SE(b_1) = \sqrt{\frac{\sigma_\varepsilon^2}{SS(X)}}$$

$$SE\ (\hat{Y}) = \sqrt{\sigma_\varepsilon^2\left[\frac{1}{n} + \frac{(x-\overline{X})^2}{SS(X)}\right]}$$

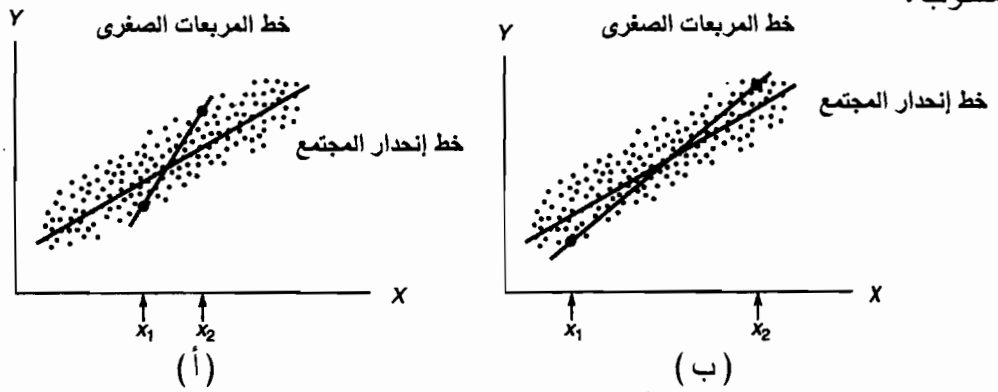
$$SE_P(\hat{Y}) = \sqrt{\sigma_\varepsilon^2\left[1 + \frac{1}{n} + \frac{(x-\overline{X})^2}{SS(X)}\right]}$$

دقق النظر في هذه المعادلات، هل الأخطاء المعيارية بها شئ مشترك؟ الإجابة، بالطبع نعم، الثلاث معادلات يعتمدون على تباين الخطأ σ_ε^2 . ويعتمدون أيضاً على SS(X) ، وهو الذي يقيس الإختلاف فى قيم X . أخيراً، الخطأين المعياريين الأخيرين يعتمدان بشكل مباشر على حجم العينة n والمقدار $(x-\overline{X})^2$، وهو يمثل مربع البعد بين القيمة المعينة x عن قيمة متوسط العينة $\overline{X}$.

دعنا نقوم بفحص هذه العوامل مع الأخذ فى الإعتبار مدى تأثيرها على دقة التقدير أو التنبؤ وبماذا توحى عند إختيار بيانات العينة .

1 . تباين الخطأ . تتذكر أن σ_ε^2 يقيس إلى أى مدى تتجمع فيه القيم Y حول خط إنحدار المجتمع . وهكذا ، كلما كان تباين الخطأ كبير ، كلما قلت درجة الإعتماد على التقديرات والتنبؤات . مع ذلك ، ضع في ذهنك أن تباين الخطأ المثالي لن يكون كبيراً، لأن الأخطاء العشوائية الحقيقية تكون صغيرة نسبياً. حيث أن تباين الخطأ σ_ε^2 تم تقديره عن طريق تباين البواقى S_e^2 ، فإنه دائماً ما يوصى بالتأكيد بأن قيمة S_e^2 لن تكون كبيرة بسبب إهمال باقى الحدود أو المتغيرات التفسيرية أو المتنبأ عندها **Predictor Variables** والتى يكون مطلوب إدخالها فى نموذج الإنحدار لكى يتم تحسينه .

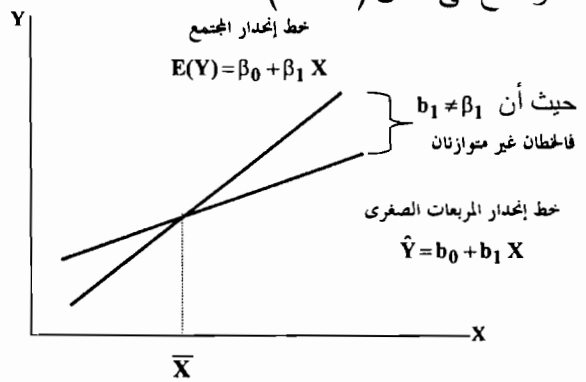
2 . الإختلاف فى قيم عينة المتغير المستقل أوالتفسيري **Predictor Variable** . كلما كان SS(X) كبيراً، كلما إتجهت التقديرات والتنبؤات لأن تكون أكثر دقة. بإسترجاع أن SS(X) يقيس الإختلاف الكلى لقيم X فى العينة. هكذا، كلما كبر مدى القيم فى بيانات العينة للمتغير المفسر X، كلما كان من المتوقع أن تكون تقديراتنا أكثر دقة. وإذا فكرنا لماذا يكون هذا صحيح؟ إفترض أننا نرغب فى تحديد خط إنحدار لعينة بها فقط (n=2) من المشاهدات. نقوم بعمل هذا بطريقتين مختلفتين . الطريقة الأولى، نختار قيمتين لـ X قريبتين جداً من بعضهما. الطريقة الثانية، نختار قيمتين لـ X بعيدتين عن بعضهما (الحالة الأخيرة تتضمن إختلافات Variability أكثر لعينة X). إفترض فى كل حالة أن أول قيمة مشاهدة Y تصادف أن تكون أصغر من المتوقع (بعبارة أخرى، أسفل خط إنحدار المجتمع)، والثانية تصادف أن تكون أكبر من المتوقع. هذا الوضع موضح فى شكل (٩-٢٠). لاحظ ماذا يحدث عندما ننشأ خطوط المربعات الصغرى. نقط البيانات المختارة تم تميزها عن طريق نقط كبيرة سوداء. فى شكل (٩-٢٠ أ)، حيث تكون قيمتى X قريبتين من بعضهما، نلاحظ أن خط المربعات الصغرى يكون أكثر إنحداراً من خط إنحدار المجتمع. فى شكل (٩-٢٠ب)على الجانب الآخر، عندما تكون قيمتي X بعيدتين جداً عن بعضهما، خط المربعات الصغرى يكون قريباً جداً من خط إنحدار المجتمع. هذا المثال البسيط يوضح لماذا تتحسن درجة دقة التقديرات والتنبؤات عندما يكون لإختلاف أكبر بين قيم X . هذه الفائدة الكبيرة تكون ممكنة فقط إذا كان لدينا مرونة فى إختيار قيم X والتي عندها نشاهد أو نسجل قيم عينة Y . وإذا أمكننا فعل هذا ، يجب علينا إختيار قيم X التي ينتج عنها أكبر اختلاف على المدى المطلوب .



شكل (٩-٢٠) تأثير تباعد قيم X على دقة خط المربعات الصغرى

3. حجم العينة . المقدار 1/n فى معادلة $SE(\hat{Y})$ ، $SE_p(\hat{Y})$ يدل على أن هذه الأخطاء المعيارية تنخفض كلما زادت قيمة n . ولا يجب أن تكون هذه النتيجة مفاجأة لنا، حيث أن الزيادة فى حجم العينة يجعل التقديرات اكثر موثوقية، كما هو متوقع.

4. اقتراب قيمة X المطلوبة من $\overline{X}$. إن المقدار $(X-\overline{X})^2$ فى معادلة $SE(\hat{Y})$ ، $SE_p(\hat{Y})$ يدل على أنه كلما ابتعدت قيمة x المطلوبة عن $\overline{X}$ كثيراً، كلما زاد الخطأ المعياري. وهكذا فإن قيم العينة X يجب أن يتم اختيارها إلى الحد الذى تكون فيه قيمة X المطلوبة للتقدير أو التنبؤ قريبة من متوسط قيم العينة X . هل تأثير هذا العامل يكون مفهوماً لك بديهياً؟ أعتبر التوضيح التالي: إحصاء المربعات الصغرى b_1 والذى سبق تعريفه بأنه المقدر الوحيد لميل المجتمع β_1، لذلك فإن قيمة b_1 والتى تعتمد فى حسابها على بيانات عينة معينة، سوف يكون من المؤكد وجود خطأ بمقدار ما بها. حيث أن ميل خط المربعات الصغرى سوف يختلف نوعاً ما عن ميل خط إنحدار المجتمع، فإن الخطان لن يكونا متوازيان. هذا موضح فى شكل (٩-٢١).



شكل (٩-٢١)

تباعد خط المربعات الصغرى عن خط إنحدار المجتمع

عادة ما يتقاطع هذان الخطان بالقرب من مركز المدى لقيم (X) ، أى قريباً من $(X=\overline{X})$. كلما بعدت قيمة x المعينة عن $\overline{X}$، كلما إزداد ابتعاد (إنحراف) الخطان. الفرق بين الخطان يمثل خطأ التقدير الذى يحدث لأى قيمة معينة X . لذلك فإن خطأ التقدير يتجه للتزايد كلما إعتبرنا قيم X بعيدة عن $\overline{X}$.

تمارين

(٩-٤٤) إرجع إلى تمرين (٩-١٥) :

(أ) إستخدم خط المربعات الصغرى فى تحديد الوسط الحسابى لـ Y عندما (X = 7) .

(ب) بناء على فترة ثقة (95%) لمتوسط Y ، هل هذا المتوسط تم تقديره بدقة كبيرة؟ وضح .

(ج) إستخدم خط المربعات الصغرى للتنبؤ بقيمة Y عندما (X = 7). هل تختلف إجابتك عن التى كانت فى جزء (أ) ؟ وضح .

(د) بناء على فترة تنبؤ (95%) لقيمة Y الفردية عندما (X = 7) ، هل هذه القيمة قدرت بدقة معقولة؟ وضح .

(هـ) قارن إجابتك للفترات فى أجزاء (ب) ، (د). أى فترة تكون أوسع ولماذا تكون هذه الفترة أوسع ؟

(٩-٤٥) إرجع إلى تمرين (٩-٢٠) ثم أجب عن كل أجزاء تمرين(٩-٤٤) عندما (X = 4)

(٩-٤٦) إرجع إلى تمرين (٩-٢٢) ثم أجب عن كل أجزاء تمرين (٩-٤٤) عندما (X = 8)

(٩-٤٧) إرجع إلى تمرين (٩-١٢) ثم أجب عن كل أجزاء تمرين (٩-٤٤) عندما (X = 67)

(٩-٤٨) إرجع إلى تمرين (٩-١٠) ، وتمرين (٩-٣٦). وهي تتضمن العلاقة بين مبلغ التأمين على الحياة (Y) والدخل العائلى (X) :

(أ) قدر المبلغ المتوسط للتأمين عندما تكون دخول العائلة هى X حيث X تأخذ 60, 35, 15 (بآلاف الدولارات) .

(ب) لكل قيم X فى جزء (أ)، حدد فترات ثقة (95%) لمتوسط مبلغ تأمين الحياة. لأى قيمة لـ X تكون دقة التقدير أفضل ولماذا ؟

(ج) عند قيم X فى جزء (أ) ، حدد فترات تنبؤ (95%) لمبلغ تأمين الحياة الفردى ، وأجب على نفس السؤال كما فى جزء (ب).

(د) مع إعتبار إجابتك على هذا التمرين ، إضافة إلى تمارين (٩-١٠) ، (٩-٣٦)، هل تعتقد الآن أن خط المربعات الصغرى بالفعل ملائم ومناسب للتقدير والتنبؤ؟ فسر إجابتك .

(٩-٤٩) إرجع إلى تمرين (٩-١١) ، (٩-٣٧). وهو يتضمن العلاقة بين بداية الأجر (Y) ومتوسط نقط التقدير (X) . أجب عن أسئلة مماثلة لتلك الموجودة فى تمرين (٩-٤٨) عندما تكون متوسطات نقط التقدير X عبارة عن 2.70 , 3 , 3.30 = X.

(٩-٥٠) بالرجوع إلى تمرين (٩-١٣)، (٩-٣٨) والذى يتضمن العلاقة بين الكمية المستهلكة من الكحول (Y) والسعر النسبى للكحول (X) . أجب عن أسئلة مماثلة لتلك الموجودة فى تمرين (٩-٤٨)عندما تكون الأسعار النسبية للمتغير X هى 3.5 , 4.6 , 5.8 سنت .

(٩-٥١) بالرجوع إلى تمرين (٩-١٤) ، (٩-٢٥) ،(٩-٣٩) والذى يتضمن العلاقة بين النسبة المئوية للضرائب المسددة (Y) وإجمالى الدخل (X) . أجب عن أسئلة مماثلة لتلك الموجودة فى تمرين (٩-٤٨)عندما يكون إجمالى الدخل X مساوياً 11 ، 42 ، 98 (بآلاف الدولارات) .

(٩-٥٢) بالرجوع إلى تمرين (٩-٤١) ، الذى يتضمن العلاقة بين مبلغ الفاتورة (Y) ، سعر البيع المحدد (X) لمؤسسة Express Graphic :

(أ) قدر متوسط مبلغ الفاتورة للأعمال التى سعرها المستهدف يساوى $5,000 .

(ب) حدد هامش خطأ المعاينة لتقديرك فى جزء (أ) بناء على مستوى ثقة (95%) .

(ج) بماذا توحى إجاباتك عن الجزء (أ)، (ب) للادارة بخصوص مصداقية سعر البيع المستهدف؟ فسر إجابتك .

(د) إفترض أنه تم سؤالك لتحديد هامش خطأ المعاينة للتنبؤ بالمبلغ الفعلى بالفاتورة لعمل فردى الذى كان سعره المستهدف $5,000 . بدون حساب أى شئ، هل هامش خطأ المعاينة هذا سوف يكون أقل من ، مساوى ، أو أكبر من المحسوب فى جزء (ب) ، ولماذا؟

(٩-٥٣) إرجع إلى تمرين (٩-٤٢) ، الذى يتضمن العلاقة بين سرعة الطباعة (Y) وصعوبات التسجيل (X) لشركة الطباعة Express Graphics :

(أ) تنبأ بسرعة الطبع للنشاط أو العملية المصنفة 1 (صعب) .

(ب) حدد فترة تنبؤ (95%) لسرعة الطباعة عندما $X = 1$.

(ج) إذا كنت المدير ، كيف ستصف فائدة التنبؤ فى جزء (أ) بناء على إجابتك لجزء (ب) ؟

(٩-٥٤) بالرجوع إلى تمرين (٩-١٢)، الذى يوضح العلاقة بين الطول (X) والوزن (Y) للموظفين الإناث فى شركة كبيرة. وبإفتراض أنه توجد مرونة لإختيار عدد الموظفين الإناث الذين يكون طولهم فى المدى من 62 إلى 70 بوصة. إذا علم أن العلاقة خطية بشكل قاطع، لأى أطوال يجب أن تختار لمشاهدة الأوزان؟ ولماذا يجب أن يتم إختيار هذه الأطوال؟

(٩-٥٥) بالرجوع إلى تمرين (٩-١٠)، وفيه يرغب محلل التأمين تحديد العلاقة بين الدخل العائلى (X) ومبلغ تأمين الحياة (Y) . إفترض أن المحلل يرغب فى تحديد هذه العلاقة للدخول العائلية فى المدى من $20,000 إلى $100,000 . وبإفتراض أن المحلل لديه مرونة فى إختيار الدخول العائلية (قيم X) لملاحظة مبالغ تأمين الحياة :

(أ) إذا علم أن العلاقة خطية بشكل قاطع، لأى دخول عائلية يجب أن يلاحظ المحلل مبالغ تأمين الحياة ولماذا ؟

(ب) إفترض أن هذه العلاقة تأخذ شكل منحنى . هل سوف تكون إجابتك نفسها كما في جزء (أ)؟ وضح .

(٩-٧) الإرتباط : قياس الإرتباط الخطى بين X , Y :

Correlation: Measuring The Linear Association Between Y and X

فى الأجزاء السابقة ، إفترضنا الإرتباط الخطى بين المتغيرات X , Y . بتحديد خط المربعات الصغرى بناء على العينة التى يمثلها، كنا قادرين على صياغة نموذج لتوضيح طبيعة هذه العلاقة الخطية. فى الواقع، كنا قادرين على تحديد ما اذا كان الإرتباط الخطى بين X , Y يمكن إعتباره مقبول عن طريق إختبار الفرض العدمى ($H_0 : \beta_1 = 0$) أم لا . إضافة لذلك، عرفنا معامل التحديد r^2، وهو مقياس نسبى لمدى جودة توفيق الخط المستقيم فى قيم العينة Y . قريب من ذلك يوجد مقياس إحصائى وصفى معروف بمعامل الإرتباط مرتبط بتلك الأفكار. فى الواقع، قيمة معامل الإرتباط هى الجذر التربيعى لمعامل التحديد. معامل الإرتباط، يرمز له بالرمز r ، يقيس درجة الإرتباط الخطى بين متغيرين X , Y بناء على عينة من المشاهدات. مثل r^2 فإن r هو مقياس محرر **Free Scale**؛ هكذا ، يكون تفسيره مستقل عن الوحدات التى تقيسهما قيم X , Y . وسوف نقدم الصيغة التى

تحدد معامل الإرتباط ، ثم نركز بصفة خاصة على تفسيره .

تحديد معامل الإرتباط :

صيغة تحديد معامل الإرتباط وطريقة اشتقاقها تكون خارج نطاق هذا الكتاب وهي كالتالى :

$$r = \frac{SP(XY)}{\sqrt{SS(X)SS(Y)}} \qquad \textbf{(9.33)}$$

للتوضيح، دعنا نحسب قيمة r لمثال تقييم الملكية. وجدنا أن (SP(XY) = 15.0) ، (SS(X) = 2.5) ، (SS(Y) = 108.8) ، وهكذا يكون معامل الإرتباط :

$$r = \frac{15.0}{\sqrt{(2.5)(108.8)}} = .9095$$

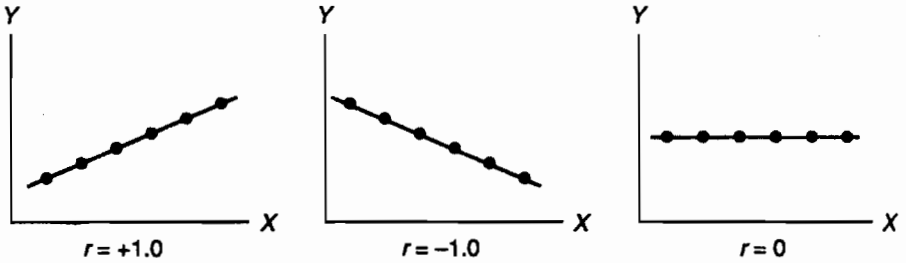
تفسير معامل الإرتباط :

الآن ، ماذا تعنى هذه النتيجة ؟ أولاً r تكون دائماً واقعة بين1- , 1+ . هذا يكون صحيح لأى بيانات ، بغض النظر عن الوحدات الأصلية. لغرض المناقشة، إعتبر القيم $r = +1$, $r = -1$, $r = 0$

(1) (r = +1) تدل على العلاقة الخطية التامة بين X , Y مع وجود ميل موجب . أى أن هناك علاقة موجبة تامة بين المتغيرين X , Y .

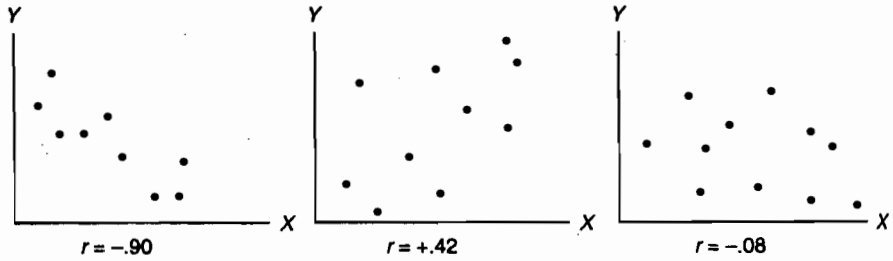
(2) (r = -1) تدل على العلاقة الخطية التامة بين X , Y لكن مع ميل سالب. أى أن هناك علاقة سالبة تامة بين المتغيرين X , Y .

(3) (r = 0) تدل على عدم وجود علاقة خطية بين X , Y. كما نرى، هذا يعنى أن الميل لخط إنحدار المجتمع يكون 0 ، هذه الحالات الثلاث موضحة فى شكل (٩-٢٢) .



شكل (٩-٢٢)
طبيعة العلاقة الخطية عندما r تساوى 1+ , 1- , 0

الحالات الموضحة فى الشكل نادراً ما تحدث فى الحياة العملية. وهكذا فإن r تكون دائماً قيمة ما غير صفرية بين 1- , 1+ . قيمة r تعتمد على كل من قيمة الميل b_1 وإختلاف بيانات العينة حول خط المربعات الصغرى. بشكل عام ، r تتناسب طردياً مع الميل b_1 ، وتتناسب عكسياً مع الإنحراف المعيارى للبواقى S_e. أمثلة نموذجية موضحة فى شكل (٩-٢٣) .



شكل(٩-٢٣)

أمثلة لمعاملات الإرتباط لبيانات ثلاث عينات

من شكل (٩-٢٣) يمكننا إستنتاج أنه كلما اقتربت قيمة r من 1+ أو 1- على السواء، كلما قوى الإرتباط الخطى بين X , Y . على الجانب الآخر، كلما قربت قيمة r من 0 ، كلما كان الإرتباط الخطى بين X , Y أكثر ضعفاً .

المقارنة بين r^2 , r , b_1 فى وصف الإرتباط الخطى :

تذكر أن إحصاء المربعات الصغرى b_1 تقدم معلومات تفصيلية عن العلاقة الخطية، لأنها تقدر التغير المحدد فى Y المناظر للتغير فى X . معامل التحديد أيضاً يقدم معلومات عن العلاقة الخطية لأنه مقياس حر لمدى جودة توفيق الخط المستقيم لقيم Y . حيث أن معامل الإرتباط r يقيس درجة الإرتباط الخطى بين X , Y ، فلن يكون مفاجأة أن نجد أن هذه الكميات الثلاث تكون مرتبطة عن قرب .

كما ذكرنا ، معامل التحديد هو مربع معامل الإرتباط ، لمثال تقييم الملكية، وجدنا (r = .9095) ، لذلك يكون معامل التحديد ($r^2 = .9095^2 = .8272$) ، كما علمناها منذ قليل، اذا كان أحد هذه الكميات يمكن تحديده من الكمية الأخرى، لماذا يكون كل منها شائع الإستخدام؟ الإجابة بسيطة : أن تفسيراتهم مختلفة نوعاً ما ، حيث أن :

$$r = SP(XY) / \sqrt{SS(Y)SS(X)} \quad , \quad b_1 = SP(XY) / SS(X)$$

ويمكن توضيح أن إحصاء المربعات الصغرى b_1 ومعامل الإرتباط r مرتبطين رياضياً عن طريق :

$$b_1 = r\sqrt{\frac{SS(Y)}{SS(X)}} \qquad \textbf{(9.34)}$$

لاحظ أنه إذا كانت (r = 0) فإن ($b_1 = 0$) والعكس بالعكس. أكثر من ذلك، فإن إشارة b_1 هى نفسها إشارة r دائماً – بعبارة أخرى، إذا كانت قيم r^2 ، b_1 معلومة فإن r وإشارتها تكون معلومة، حيث أن $\mathbf{r = \sqrt{r^2}}$ وإشارة r هى نفسها إشارة b_1 .

مثال (٩-٧)

بالنسبة لمثال الطاقة مقابل درجة الحرارة، حدد وفسر الإرتباط بإستخدام مخرجات الكمبيوتر المعطاة فى جدول (٩-٣).

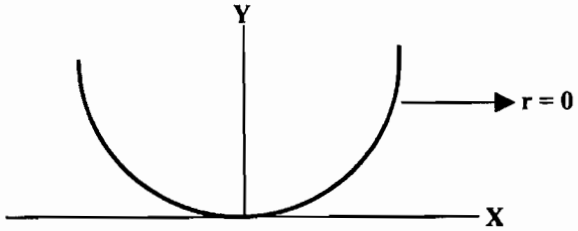
الحل

من مخرجات الكمبيوتر ، رأينا أن ($b_1 = -3.464$) ، ($r^2 = .975$) . لذلك ، قيمة r تساوى $\sqrt{.975}$،

حيث أن قيمة b_1 سالبة فإن ($r = -\sqrt{.975} = -.98746$). وحيث أن قيمة معامل الإرتباط (-.987) قريبة من (-1) فإن هذا يدل على الإرتباط الخطى القوى بين الطاقة ودرجة الحرارة وحيث أن r سالبة، فإن العلاقة تكون عكسية. أى إذا ارتفعت درجة الحرارة فى الشتاء فإن مقدار الطاقة المستخدمة فى التدفئة تقل .

معامل الإرتباط عندما تكون العلاقة بين Y ، X غير خطية :

وأخيراً، فإنه من الأهمية بمكان أن نؤكد مرة أخرى على أن r تقيس فقط درجة الإرتباط الخطى بين X, Y . ومن الممكن أن تكون X , Y مرتبطين بطريقة غير خطية. للتوضيح، إعتبر الرسم فى شكل (٩-٢٤)، حيث تكون Y مرتبطة بشكل تام مع X ، لكن بشكل غير خطى على الإطلاق. فى مثل هذه الحالة فإن معامل الإرتباط يكون 0 . لهذا السبب عندما كنا نفسر ماذا تقيس r^2 فى جزء (٩-٣-٤) ، ذكرنا أن r^2 لا يمكن أن تقيس صحة الإنحدار المفترض .



شكل (٩-٢٤)

الإرتباط التام غير الخطى بين X , Y عندها يكون معامل الإرتباط يساوى الصفر

تمارين

(٩-٥٦) وضح المقصود بمعاملات الإرتباط الافتراضية التالية :

(أ) (r = -1) (ب) (r = 0) (ج) (r = +1)

(٩-٥٧) عند إنشاء نموذج الإنحدار الخطى ، وجد أن

(SS(Y) = 129) ، (SS(X) = 9.5) ، (SP(XY) = 21.4) . حدد معامل الإرتباط .

(٩-٥٨) عند إنشاء نموذج إنحدار خطى، وجد أن (b_1= -2.58)، (r^2 = .8238) ، حدد معامل الإرتباط .

(٩-٥٩) عند إنشاء نموذج إنحدار خطى ، وجد أن (SS(Y) = 98) ، (SS(X) = 38.6) ، (r = .6525) حدد تقدير المربعات الصغرى للميل .

(٩-٦٠) إذا عوضنا عن قيم X : (4 , 3 , 2 , 1 , 0 , -1 , -2 , -3 , -4) فى العلاقة الرياضية ($Y = X^2$)، نحصل على قيم Y المناظرة : (16 , 9 , 4 , 1 , 0 , 1 , 4 , 9 , 16) .

(أ) ارسم علاقة X , Y .

(ب) حدد معامل الإرتباط بإستخدام قيم X , Y . هل أنت مندهش من إجابتك ؟ وضح .

(٩-٦١) إستخدم المعلومات المتاحة لك، وحدد معامل الإرتباط للتمارين التالية :

(أ) تمرين (٩-١٥) . (ب) تمرين (٩-٤٠). (ج) تمرين (٩-٤١) .

(٩-٦٢) أظهرت إحدى الدراسات أن معامل الإرتباط بين مكاسب ممثلى المبيعات (Y) وعدد المكالمات التى يقومون بها لمحاولة البيع (X_1) هو 44. وأوضحت نفس الدراسة أن معامل الإرتباط بين المكاسب (Y) وعدد الساعات المنقضية فى المهام الإدارية (X_2) هو 55.- . أى عامل ، X_1 أو X_2 يظهر إرتباط خطى أكثر قوة بالمكاسب ؟ برر إجابتك .

(٩-٦٣) يقوم مكتب القبول بأحد الجامعات بدراسة مؤشرات الطلاب بالمرحلة الثانوية وأدائهم فى الكليات ، حيث يقاس ذلك عن طريق متوسط تقديراتهم فإذا توافرت بيانات عن الإنتهاء من مرحلة الثانوية وكذا مستوى الأداء بالكلية لعينة مكونة من 498 طالب حديث. وكان معامل الإرتباط بين تقدير الكلية (GPA) وتقدير المدرسة الثانوية (GPA) ووجد أنه يساوى 42. ووجد أن معامل الإرتباط بين تقدير الكلية (GPA) وترتيب المدرسة الثانوية يساوى 36 .

(أ) بناءً على العلاقة الخطية ، أيهما يبدو المؤشر الأقوى لأداء الكلية ؟ وضح .

(ب) فى التنبؤ بأداء الكلية، إلى أى مدى يساعد هذا فى معرفة تقديرات طلبة المدارس الثانوية GPAs (في مقابل عدم وجود معلومات عن المدرسة الثانوية على الإطلاق)؟

(ج) أى جزء للاختلافات Variability بين تقديرات طلبة الكليات (GPAs) تم تفسيره عن طريق الإختلافات بين تقديرات مدارسهم الثانوية (GPAs) ؟

(٩-٨) الإنحدار الخطى البسيط : مثال شامل : A Comprehensive Example

هذا المثال مبنى على تطبيق فعلى فى شركة Xerox . البعد الوحيد عن الواقعية هو التغيير فى البيانات . البيانات الأصلية مسجلة، وتم تبسيط المثال عن طريق تقديم عينة صغيرة لبيانات مصطنعة. ظروف العمل ، الإصدارات، والقواعد الإحصائية المتضمنة مطابقة للحالة الفعلية. الحسابات تم تنفيذها بإستخدام الكمبيوتر .

فى شركة Xerox . تم إجراء بحث موسع لدراسة السوق. تم أخذ عينة من 500 مؤسسة من قائمة 900,000 مؤسسة صغيرة فى الولايات المتحدة، تم الحصول عليها من خلال شركة Dun & Bradstreet . المجتمع محل الاهتمام يتكون من كل المؤسسات الصغيرة فى الولايات المتحدة. إطار إختيار العينة يتكون من قائمة Dun & Bradstreet . «المؤسسة الصغيرة» تم تعريفها على أنها المؤسسة التى تضم 50 موظف أو أقل . بعض المعلومات الوصفية تم تقديمها عن طريق Dun & Bradstreet لكل مؤسسة فى المجتمع . المتغير المستقل (المفسر) الأساسي للدراسة كان عدد الموظفين فى كل مؤسسة.

تم زيارة كل مؤسسة فى العينة ، وتم الحصول على معلومات تفصيلية بخصوص طريقة النسخ أو التصوير. البند الأساسى لمعلومات هذه الدراسة كان متوسط عدد النسخ التى تقوم بها المؤسسة بتصويرها فى اليوم . هذا المتغير يسمى «حجم النسخ» . الإدارة فى Xerox تعتقد أن حجم أو عدد النسخ التي يتم تصويرها للمؤسسات يتجه للزيادة كلما تزايد عدد الموظفين – موظفين أكثر، يعني تصوير نسخ أكثر فى المتوسط. ويعتقدون أيضاً أن طرق التصوير (النسخ) فى المؤسسات تعتمد نوعاً ما على نوع المؤسسة – أى ، الصناعة التى تنتمى إليها ، يتم تعيين هذا عن طريق كود التصنيف الصناعى المعيارى (SIC) للمؤسسة، أيضاً تم تقديمه عن طريق Dun & Bradstreet . الأمثلة

للمؤسسات هى «الكليات والجامعات» ، «البنوك» ، «المصانع» فى هذا المثال، نختبر البيانات التى تخص المؤسسات المصرفية .

ومن البنوك التي تم إختيارها في هذه الدراسة، كان الاهتمام بعدد من 8 إلى 50 من العاملين بهذه البنوك للتأكيد على درجة الدقة المقبولة فى التنبؤات والتقديرات، أراد فريق الإستقصاء أن ينشأ إختلاف كبير بين قيم X (عدد الموظفين) ضمن المدى المهتم به. بشكل أولى، أراد الفريق أخذ عينة لعدد نسخ التصوير (قيم Y) لخمس بنوك بعدد موظفين (8-10) لكل بنك وخمس بنوك بعدد موظفين (48-50) لكل بنك (القيم المتطرفة فى مدى X) . لكن أحد الأشخاص فى فريق الإستقصاء ذكر أن قسم التخطيط يرغب فى تقدير حجم الأوراق التي تم تصويرها (نسخها) للبنوك بعدد موظفين 30 . إضافة لذلك ، تم إدراك أنه إذا كان هناك شكل منحنى فى العلاقة بين حجم النسخ وعدد الموظفين ، فلا يمكن إكتشافها بإستخدام القيم المتطرفة فقط فى مدى X . لذلك، تم إختيار 5 بنوك فى المدى المتوسط (28-30) موظف لكل بنك. إضافة إلى أطراف المدى X. كنتيجة لذلك، تم الحصول على بيانات العينة التالية :

عدد الموظفين (X)	8	9	9	10	10	28	29	29
حجم النسخ اليومى(Y)	8	13	17	15	23	56	46	60
عدد الموظفين (X)	30	30	48	48	49	50	50	
حجم النسخ اليومى(Y)	52	65	86	90	95	99	88	

فإذا رغبت إدارتان في شركة Xerox فى إستخدام هذه الدراسة :

(1) قسم التخطيط يرغب فى تقدر حجم الأوراق المصورة (النسخ) للبنوك التي بها عدد 30 موظف (30 كان رقم من أرقام benchmark العديدة فى عمليتهم التخطيطية) . وقد استخدموا معلومات Dun & Bradstreet حيث يوجد 800 مؤسسة بنكية بها 30 موظف تقريباً، لذلك كانوا بحاجة إلى تقدير متوسط حجم النسخ لهذه المؤسسات .

(2) إدارة المبيعات ترغب فى عمل نموذج للتنبؤ بحجم النسخ للبنوك الفردية (المفردة) ، بمعلومية عدد الموظفين. ممثلى المبيعات يمكنهم إستخدام هذا التنبؤ فى إعداد عروضهم لعملائهم فيما يتعلق بالحجم الملائم للناسخين للعميل . الغرض من تحليل الإنحدار فى هذا المثال هو تقديم نموذج يشبع حاجات كل من قسم التخطيط وإدارة المبيعات .

أسئلة مناسبة لهذه الدراسة :

١- كيف يظهر حجم النسخ (Y) ليكون معتمد على عدد الموظفين (X) ؟ هنا يجب علينا تقدير علاقة X , Y ببيانات العينة .

٢- هل علاقة Y تجاه X فى العينة تقدم دليل قوى على الإرتباط بين X , Y فى المجتمع؟ أو بمعنى اخر، هل العلاقة بينهما ترجع إلى المعاينة العشوائية التي تمت بين مجتمع لايوجد فيه إرتباط بين X، Y ؟

٣- إفترض أن العينة تدل بشكل مقنع على أن حجم النسخ مرتبط بعدد الموظفين. ما مدى دقة نموذج إنحدار المربعات الصغرى فى تصوير العلاقة Y تجاه X ؟ هل هذا المستوى من الدقة يجعل معادلة الإنحدار المقدرة صالحة للإستخدام عن طريق كل من التخطيط للمبيعات أو التخطيط السوقى؟ هل هناك أى إخلال ملحوظ لإفتراضات الإنحدار الضرورية ؟

٤- إفترض أن قسم التخطيط أو دراسة السوق كان يستخدم معادلة إنحدار المربعات الصغرى لتقدير متوسط حجم النسخ للمؤسسات فى المجتمع بعدد 30 موظف .

(أ) ما هو تقديرهم .

(ب) ما هى الدقة التى يتوقع أن يكون عليها هذا التقدير ؟.

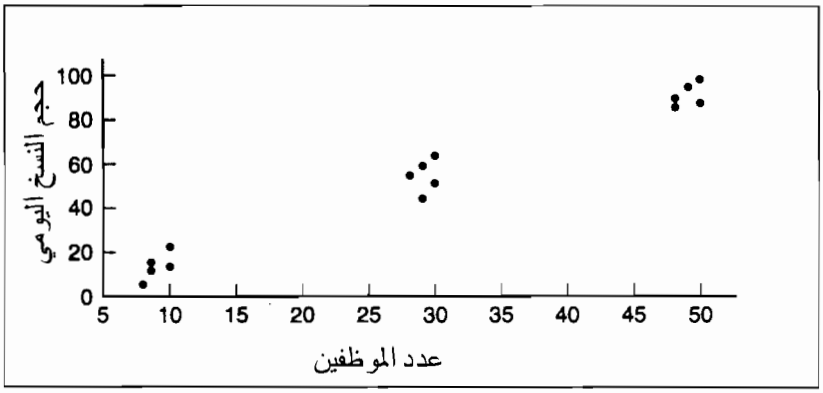
٥- إفترض أن إدارة المبيعات استخدمت معادلة إنحدار المربعات الصغرى فى التنبؤ بحجم النسخ لمؤسسة بنكية جديدة (ليست فى العينة) بعدد 30 موظف :

(أ) ما هو العدد الذى يمكنهم التنبؤ به ؟

(ب) ما مدى الدقة التى يتوقعونها لتنبؤهم ؟

إجابة الأسئلة ١ ، ٢ :

شكل الإنتشار يقدم دليل أولى واضح للعلاقة Y تجاه X . شكل الإنتشار (٩-٢٥) يظهر وجود إرتباط خطى قوى بين حجم النسخ وعدد الموظفين .



شكل(٩-٢٥)

الشكل الإنتشارى لحجم النسخ مقابل لعدد الموظفين

إضافة لذلك ، لا يوجد شكل إنحنائي يمكن تمييزه فى هذه العلاقة، لذلك فإن توفيق الخط المستقيم يكون كافياً تماماً. ولقد تم الحصول على خط المربعات الصغرى والمعلومات الهامة الأخرى عن طريق البرنامج الإحصائى Minitab . ومخرجات الكمبيوتر معطاة فى جدول (٩-١٠).

بناء على هذا المثال، فإن تقدير متوسط حجم النسخ Y بمعلومية عدد الموظفين X ، وعن طريق خط المربعات الصغرى .

$$\hat{Y} = -1.82 + 1.92X$$

وحيث أن مدى X لا يتضمن 0 ، فإن تقدير الجزء المقطوع ليس له معنى حقيقى فى هذه المشكلة. لكن تقدير الميل ($b_1 = 1.92$) له بالفعل معنى معين. فهو يعنى أن موظف واحد إضافى يؤدى إلى وجود زيادة إضافية مقدارها 1.92 نسخ كل يوم ، فى المتوسط .

إجابة السؤال ٣ :

بناء على مخرجات الكمبيوتر ، يجب أن يكون واضحاً لك من خلال قيم T أو F ($F = 503,02$ أو $T = 22.43$) أن الفرض العدمى بعدم وجود إرتباط خطي بين Y ، X ($H_0 : b_1 = 0$) يتعارض بشدة مع بيانات العينة (قيمة P تكون فعلياً 0). الأكثر أهمية أن ميل خط إنحدار المجتمع تم تقديره بدقة كبيرة. الدليل الواضح على هذا، هى حقيقة أن الخطأ المعيارى للاحصاء b_1 صغير جداً = ($SE(b_1)$ .08574) بالنسبة لميل المربعات الصغرى $b_1 = 1.92295$. بعبارة أخرى، قيمة T ($T = 22.43$) كبيرة تماماً. كنتيجة لذلك، فترة الثقة للميل β_1 يتوقع أن تكون ضيقة. على سبيل المثال فترة الثقة (95%) للمؤشر β_1 هى : ($t_{.975,13} = 2.160$):

$$1.92295 \pm (2.160)(.08574) = 1.92295 \pm .185$$

أو (2.108 , 1.738) بدرجة ثقة (95%) ، هذه الفترة تعنى أن موظف إضافى يمكن أن يساهم بمقدار صغير من النسخ مثل 1.738 لكل يوم أو مقدار كبير إضافى من النسخ مثل 2.108 لكل يوم ، فى المتوسط، وضيق هذه الفترة يوضح أن الميل β_1 تم تقديره بدرجة دقة كبيرة .

لاحظ أيضاً أن قيمة معامل التحديد قريبة جداً من واحد ($r^2 = .975$) هذا يعنى أن 97.5% من الإختلاف الكلى فى عينة حجم النسخ تم تفسيره عن طريق خط المربعات الصغرى ($\hat{Y} = -1.82 + 1.92X$) .

أخيراً، عندما رسمنا البواقى فى شكل (٩-٢٦) لاحظنا عدم وجود نمط يمكن تميزه. هذا يعنى أننا لم نكتشف أى اختلال ملحوظ للفروض. لذلك فإن توفيق الخط المستقيم يظهر أنه كاف تماماً ويجب إستخدامه للتقدير والتنبؤ من خلال المدى محل الاهتمام في الدراسة .

جدول (٩-١٠)

نتائج مخرجات المثال الشامل بإستخدام Minitab

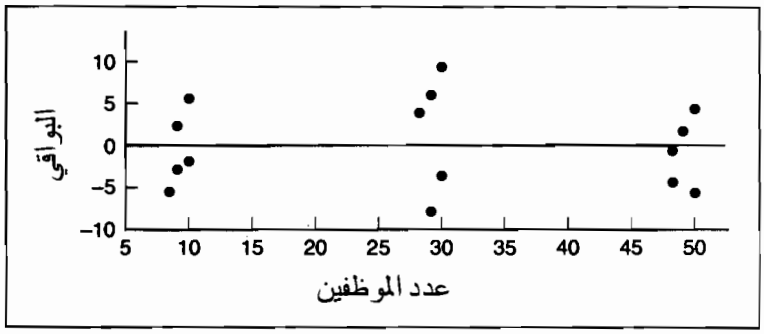
```
The regression equation is volume = -1.82 + 1.92 number

Predictor        Coef          Stdev        t-ratio       p
Constant       -1.822          2.861         -0.64      0.535
number        1.92295        0.08574         22.43      0.000

s=5.402    R-sq=97.5%    R-sq(adj)=97.3%

Analysis of Variance

SOURCE          DF          SS          MS          F       p
Regression       1       14679       14679     503.02   0.000
Error           13         379          29
Total           14       15058
```

Fit	STdev.Fit	95% C.I.	95% P.I.
55.87	1.40	(52.85, 58.88)	(43.81, 67.92)

ROW	number	volume	yhat	residual
1	8	8	13.5617	-5.56165
2	9	13	15.4846	-2.48460
3	9	17	15.4846	1.51540
4	10	15	17.4076	-2.40755
5	10	23	17.4076	5.59245
6	28	56	52.0207	3.97934
7	29	46	53.9436	-7.94361
8	29	60	53.9436	6.05639
9	30	52	55.8666	-3.86656
10	30	65	55.8666	9.13344
11	48	86	90.4797	-4.47966
12	48	90	90.4797	-0.47966
13	49	95	92.4026	2.59739
14	50	99	94.3256	4.67444
15	50	88	94.3256	-6.32556

إجابات الأسئلة ٤ ، ٥

من جدول (٩-١٠) لاحظنا أنه إذا إستخدمت إدارة التخطيط السوقى النموذج عندما (X = 30)، فإن متوسط حجم النسخ المقدر يكون 55.87 = ($\hat{Y}$) ، فترة ثقة 95% تكون (52,85 , 58.88). لذلك فإنه عندما تكون (X = 30) موظف، فإن متوسط عدد النسخ يمكن أن يكون صغير بمقدار 52.85 أو كبيراً بمقدار 58.88 بدرجة ثقة (95%) .



شكل (٩-٢٦)

الشكل الإنتشارى للبواقى لحجم النسخ مقابل عدد الموظفين

إذ إستخدمت إدارة المبيعات النموذج عندما (X = 30) ، حجم النسخ المتنبأ به للمؤسسة البنكية بعدد 30 موظف هو أيضاً 55.87 = ($\hat{Y}$) . لكن ، كما نتوقع فترة تنبؤ (95%) تكون (43.81 , 67.92) وتكون أوسع من فترة الثقة المناظرة لمتوسط حجم النسخ، بمعلومية (X = 30) موظف. فترة التنبؤ (43.81 , 67.92) تعنى أن بنك جديد بعدد (X = 30) موظف يمكنه إضافة عدد نسخ قليلة بمقدار 43.81 لكل يوم أو كبيرة بمقدار 67.92 بدرجة 95% . كل من فترات الثقة والتنبؤ يجب أن يكون مفيداً تماماً لتخطيط السوق والمبيعات، على التوالى ، لأنها تكون فترات ضيقة نسبياً .

(٩-٩) ملخص : Summary

فى هذا الفصل ، عرضنا الأشكال المختلفة لتحليل الإنحدار الخطى البسيط لبناء الإرتباط بين المتغير التابع **response variable** والمتغير المفسر **predictor variables** المحتمل. وافترضنا أن العلاقة بين هذه المتغيرات هى الخط المستقيم واستخدمنا تحليل الإنحدار لتقدير وتقييم حدود هذه العلاقة الخطية المفترضة .

الغرض الأساسى لنموذج الإنحدار هو تمثيل النظم المعقدة ببساطة وبشكل سهل وتقديم فهم أفضل لخصائص النظام . إضافة لذلك ، نموذج الإنحدار يستخدم للتقدير والتنبؤ بقيم المتغير التابع. بالتالي، يكون من المهم التأكيد على ما إذا كان نموذج الإنحدار يؤدى بشكل كاف للإستخدام المطلوب أم لا. فى هذا السياق ، تظهر ثلاث أسئلة هامة : (1) هل بيانات العينة تدل على وجود إرتباط خطى بين المتغيرين؟ (2) ما مدى دقة التقديرات و / أو التنبؤات ؟ (3) هل يمكن تحسين نموذج الإنحدار المفترض عن طريق إعتبار متغيرات مفسرة أخرى محتملة؟ هذه الأسئلة يمكن الاجابة عليها عن طريق مزيج من أساليب الرسم البيانى والإستنتاجات الإحصائية على المعالم الهامة للنموذج المفترض .

المراجـــــع : References

1. N. Draper and H. Smith . *Applied Regression Analysis*, 2nd ed. New York: Wiley, 1981.
2. W. Mendenhall and T. Sincich. *A Second Course in Business Statistics: Regression Analysis*, 4th ed. San Francisco: Dellen, 1993 .
3. R. B. Miller and D. W. Wichern. *Intermediate Business Statistic: Analysis of Variance, Regression, and Time Series.* New York: Holt, Rinehart & Winston, 1977 .
4. J. Neter W. Wasserman, and M. Kutner. *Applied Linear Statistical Models,* 2nd ed. Homewood, Il: Richard D. Irwin, 1985 .
5. M. Younger. *A Handbook for Linear Regression*, 2nd ed. Boston : Duxbury Press, 1985 .

تمارين إضافية :

بالنسبة للتمارين الإضافية التالية ، يكون الهدف النهائى هو تحديد إذا كان خط المربعات الصغرى له معنى وكاف للتقدير والتنبؤ أم لا عن طريق تحديد قوة العلاقة الخطية بين المتغيرات المفسرة والتابعة. ويجب أن يشتمل التحليل على الآتى : شكل إنتشار ، تحديد وتفسير خط المربعات الصغرى ، تطبيق وتفسير الإحصاء F , T ، تحديد وتفسير فترة ثقة (95%) لميل المجتمع β_1 ، تحديد وتفسير r^2 ثم تحليل للبواقى .

(٩-٦٤) يرغب محلل إحصائي فى مستشفى ما أن يختبر الدرجة التى يتم بها تحديد فترة الإقامة بالمستشفى للعناية الطبية لمريض ما ، يتم تحديدها عن طريق عمر المريض. ولدينا عينة من 42 مريض كانت كما يلى :

العمر	63	66	67	68	68	69	69	69
مدة الإقامة (بالأيام)	3	16	6	9	3	4	8	19
العمر	70	70	72	73	74	83	84	85
مدة الإقامة (بالأيام)	9	6	7	10	7	16	21	8
العمر	88	66	70	72	77	77	78	78
مدة الإقامة (بالأيام)	10	9	8	10	17	18	12	9

(٩-٦٥) يرغب محلل لنظم المدارس الحكومية بأحد مدن الولايات المتحدة أن يختبر كيف ستعكس درجات أحد إختبارات SAT لمتوسط التقديرات (GPA) لخريجى مدرسة المدينة الثانوية. وبأخذ بيانات عينة من 30 من الخريجين من هذه المدارس كانت كما يلى :

GPA	4.9	4.7	4.0	3.7	4.3	3.5
SAT	1235	1105	1020	1000	1190	1010
GPA	3.4	3.8	3.2	5.0	4.4	4.5
SAT	1125	1020	975	1390	1050	1205
GPA	4.8	4.6	3.6	3.7	4.5	4.6
SAT	1300	1100	970	1110	1290	1250
GPA	3.8	3.7	3.5	4.0	4.5	4.1
SAT	1010	1310	1000	950	1275	950
GPA	3.4	3.0	2.8	2.9	3.4	3.4
SAT	990	895	890	920	1270	1210

(٩-٦٦) يهتم مدير الأفراد فى بنك ما بتقييم التعليم الذى يحدث فى القسم الأكاديمى لبرنامج التدريب الإدارى. التعليم تم تقييمه لعينة من 28 متدرب بناء على درجاتهم فى إختبارات تتم قبل وبعد المشاركة فى برنامج التدريب. الإهتمام كان منصباً على التنبؤ بالدرجة بعد البرنامج بناء على الدرجة المناظرة قبل البرنامج ويتم الحصول على بيانات العينة التالية :

المتدرب	1	2	3	4	5	6
قبل	35	39	38	41	45	51
بعد	44	66	51	63	62	66
المتدرب	7	8	9	10	11	12
قبل	36	41	41	43	38	43
بعد	57	63	60	63	60	58
المتدرب	13	14	15	16	17	18
قبل	31	33	46	33	44	33
بعد	55	52	75	50	55	54
المتدرب	19	20	21	22	23	24
قبل	44	42	38	47	41	38
بعد	58	63	58	65	63	67

المتدرب	25	26	27	28		
قبل	41	42	34	34		
بعد	64	57	62	52		

(٩-٦٧) إستجابة للضغط لتقليل الإنبعاثات من المركبات المتطايرة التى يسببها إستخدام حبر الطباعة لمصنع ما على أساس مواد مذيبة، قام مدير مؤسسة بالتبديل إلى الحبر المُصنع على أساس الماء. كنتيجة للتغير، أصبح كل من الإنتاجية ومقدار النفايات موضع إهتمام. السؤال التالى إذا كان مقدار النفايات لكل 1,000 ياردة مرتبطة بالإنتاجية التى تقاس بعدد الياردات المطبوعة لكل وردية عمل. فإذا كانت بيانات الإنتاجية والنفايات لعينة مكونة من 23 من المطبوعات بإستخدام الحبر المصنع على أساس الماء كما يلى :

الياردات لكل وردية عمل	20,000	21,882	20,800	23,273	44,387
النفايات لكل 1,000 ياردة	35.76	31.15	45.09	14.31	14.65
الياردات لكل وردية عمل	37,576	48,000	20,500	21,714	18,759
النفايات لكل 1,000 ياردة	15.31	24.60	45.63	29.81	21.38
الياردات لكل وردية عمل	18,000	18,065	19,429	9,143	24,000
النفايات لكل 1,000 ياردة	27.71	24.40	50.62	63.54	51.11
الياردات لكل وردية عمل	38,316	21,429	38,261	33,333	33,500
النفايات لكل 1,000 ياردة	12.31	19.05	10.40	22.44	44.36
الياردات لكل وردية عمل	40,000	23,273	17,524		
النفايات لكل 1,000 ياردة	17.31	35.94	41.04		

(٩-٦٨) يرغب محلل إحصائي فى دراسة تغيب الموظفين الطويل عن العمل فى كل من مؤسستى Louisville , Richmond . لعينة من 30 موظف فى كل مؤسسة تم تسجيل عدد الغائبين على مدى فترة 3 سنوات وتم تسجيل سنوات الخدمة. يعتقد أن عدد الغائبين يتأثر بسنوات الخدمة. حلل كل مؤسسة على حدة ثم قارن نتائجك. إذا أتيحت لنا عينة من بيانات كل مؤسسة كالتالى :

Richmond:

الغائبين	18	14	24	5	7	0	8
سنوات الخدمة	9	9	21	15	15	15	17
الغائبين	13	2	0	5	11	10	1
سنوات الخدمة	21	18	16	15	9	14	8
الغائبين	7	2	21	18	2	12	9
سنوات الخدمة	10	15	9	12	15	16	14
الغائبين	5	9	13	16	11	9	23
سنوات الخدمة	14	14	8	8	15	13	10
الغائبين	5	13					
سنوات الخدمة	8	14					

Louisville:

الغائبين	24	0	18	28	30	51	48
سنوات الخدمة	9	31	17	19	19	16	24
الغائبين	3	14	19	50	9	13	9
سنوات الخدمة	30	17	19	7	13	7	7
الغائبين	4	50	49	15	11	15	9
سنوات الخدمة	20	16	9	12	11	18	7
الغائبين	13	5	6	33	64	12	8
سنوات الخدمة	24	18	12	7	7	16	14
الغائبين	0	3					
سنوات الخدمة	15	17					

(٩-٦٩) يرغب محلل استثمار أن يختبر ما إذا كانت الأسعار المنخفضة لمدة 52 أسبوع تؤثر على الأسعار المرتفعة لمدة 52 أسبوع أم لا للمركز الرئيسى لمؤسسات فى فرجينيا. بيانات العينة لعدد 15 من شركات فرجينيا كانت كما يلى :

Sock (المخازن)	52 - week High (52 اسبوع مرتفع)	52 - week Low (52 اسبوع منخفض)
Best Products	$ 16.38	$ 9.00
CFB	36.50	24.25
Circuit City	33.38	11.75
CSX	37.50	25.63
Dominion Resources	25.19	17.38
Du Pont	90.88	59.50
Ethyl	22.75	13.31
James River	35.00	22.00
MCI	13.13	6.13
Robins	15.63	7.75
Philip Morris	78.00	39.25
Jefferson Bank	39.50	29.00
United Virginia Bank	35.38	22.63
Basset Furniture	50.25	33.50
Sovran	44.25	29.38

(٩-٧٠) يهتم أستاذ جامعى بتحديد إمكانية إستخدام درجات الطالب فى إختبار ما فى التنبؤ بدرجاته فى الإختبار التالى. البيانات التالية تمثل الدرجات فى أول اختبارين فى نظم المعلومات مصنفة لـ 25 طالب فى السنة الثانية، 25 طالب فى السنة الثالثة : حلل كل سنة جامعية على حدة وقارن بين النتائج التى حصلت عليها .

طلبة السنة الثانية Sophomores				طلبة السنة الثانية Sophomores			
اختبار 1	اختبار 2	اختبار 1	اختبار 2	اختبار 1	اختبار 2	اختبار 1	اختبار 2
62	75	80	93	60	53	76	53
68	70	82	82	62	52	76	60
68	82	82	70	62	68	76	68
70	63	84	67	64	63	78	58
70	72	84	77	64	73	78	68
72	75	84	75	66	62	78	62
72	78	86	82	68	58	80	68
76	65	88	82	70	53	80	75
76	78	90	78	70	52	84	80
76	65	92	87	72	58	86	88
78	67	94	78	72	60	86	80
78	53	94	85	74	72	88	70
78	85			76	73		

حالة دراسية (٩-١): تحليل الخصائص البيئية للمياه فى نومنى كريك : NOMINI CREEK

قام مجموعة من المواطنين بالخدمة كمراقبى ومحللي جودة المياه لمنبع نهر Potomac فى جهد مشترك مع Alliance . الغرض من عملهم هو تقديم قاعدة من البيانات بخصوص جودة المياه يمكن إستخدامها فى المستقبل لتحديد كيف أن استخدام الأرض يؤثر في جودة المياه.

الهدف من دراسة هذه الحالة هو توضيح وعرض نموذج للعلاقة بين درجة حرارة المياه ومتغيرين بيئيين هامين ، مستوى الأكسجين المنحل Dissolved Oxygen (DO) ، عمق وضوح الماء Secchi depth (SD) كل من (DO) ، (SD) يعتبران خصائص لجودة المياه فيما يتعلق بالأوضاع البيئية . درجة الحرارة هى العامل المشتبه فيه الذى يجب إستخدامه فى تحليلات مستقبلية لتأثيرات إستخدام الأرض. البيانات تم تقديمها عن طريق أحد المواطنين المحلين وتتكون من 151 مشاهدة أسبوعية لدرجة حرارة المياه و(DO) ، (SD) لمنطقة Nomini Creek (الذى يغذى نهر Potomac) على مدى 3سنوات. درجة حرارة الماء تم قياسها بالدرجات السيليزية بالترمومتر. الأكسجين المنحل يعين بجزء لكل مليون وتم تحديده عن طريق إجراء تجارب تفصيلية كل أسبوع. عمق وضوح الماء تم قياسه بالمتر بإستخدام أسطوانة القياس، التى تكون مستديرة ومقسمة إلى دوائر ملونة تبادلياً أسود وأبيض .

فى دراسة هذه الحالة ، عليك إختبار إختلاف مستوى الأكسجين المنحل وعمق وضوح المياه على مر الزمن، والعلاقة بين مستوى الأكسجين المنحل ودرجة حرارة المياه وبين عمق وضوح الماء ودرجة حرارة الماء . إستخدم التفكير الإحصائى وطرق ومفاهيم هذا الفصل، بالإضافة إلى الفصول السابقة. يجب أن يشتمل تقريرك على مناقشة كاملة لطبيعة هذه العلاقات ، تقرير عن درجة الثقة المناسبة لكل نموذج، مدى الإختلاف للأكسجين المنحل والعمق لتوقعه فى المستقبل، إذا لم يكن هناك تأثيرات عكسية لإستخدام الأرض، وأى تحفظات لك عن ملاءمة الطرق التى استخدمتها.

البيانات محفوظة في القرص المرن المرفق في ملف يسمى CASE 0901 . دليل الأعمدة هي الأسبوع = C1 ؛ مستوى الأكسجين (D0) = C2 ؛ العمق (SD) = C3 ؛ درجة حرارة الماء = C4 .

ملحق ٩ : Appendix - 9

تعليمات الحاسب الآلى عند إستخدام البرامج الإحصائية SAS , Minitab

سوف يستخدم المثال (٩-١) كنموذج لتوضيح تعليمات SAS , Minitab والتى تستخدم للحصول على نتائج حل تلك المشكلة .

(١) إستخدام البرنامج الإحصائى Minitab :

يمكن إستخدام الأوامر SET , NAME لإدخال البيانات كما سبق أن ذكرنا - ثم يتم إستخدام الأمر PLOT لعمل الرسم البيانى للبيانات. أما الأمر REGRESS فيستخدم للحصول على (بين أشياء كثيرة أخرى) معادلة المربعات الصغرى والأخطاء المعيارية وكذلك جدول تحليل التباين ANOVA ويتم إستخدام الأمر الفرعى PREDICT لتقدير والتنبؤ بقيم Y .

والتعليمات التالية تولد الشكل (٩-٦) والجدول (٩-٣) وكذلك الجدول (٩-٩) . لاحظ أننا فى الأمر NAME قمنا بوضع عنوان العمود C4 العنوان YHAT (لتوضيح القيمة المقدرة للمتغير Y) ، العمود C5 بالعنوان RESIDUAL (البواقي). في الأمر REGRESS فإننا نعرف العمود الخاص بقيم Y للعينة (C2) وعدد المتغيرات المفسرة (1)، حيث أن قيم المتغير المفسر تظهر فى العمود الأول (C1) والعمود الذى يتم فيه تخزين العينة المتنبأ بها للمتغير التابع Y (C4) . ونعرف عمود آخر فى جملة REGRESS (C3) حيث يقوم Minitab بتخزين معلومات أخرى إضافية عن البواقى والتى لم يتم تغطيتها فى هذا الكتاب ، الأمر الفرعي RESIDUAL (C5) يقوم بتخزين البواقي فى العمود المذكور. وأخيراً فإن الأوامر الفرعية الثلاثة للتنبؤ PREDICT حيث أن الجملة الأخيرة PREDICT تم إنهائها بنقطة. هذه الأوامر الفرعية الثلاثة تقوم بإعطائنا المخرجات أو النتائج والتى توجد فى جدول (٩-٩).

```
MTB> name c1 = 'temp' c2 = 'energy'   c4 = 'ymat'  c5 = 'residual'
MTB> 1 1.5 3. -3 0.5 2.5  4  5  -5 -0.5  9  9.5  7  3 -2  6  8  10
DATA> end
MTB> set c2
DATA>   94 81 79 97 88 75 74 67 107 86 58 55 65 74 91 65 58 52
DATA> end
MTB> plot c2 c1
MTB> regress y c2 1 c1 c3 c4;
SUBC> residual c5;
SUBc> predict -4;
SUBC> predict 3;
SUBC> predict 9.
MTB> print c1 c2 c4 c5
```

ولعمل الشكل البيانى للبواقى كما فى شكل (٩-١٦) ، فإننا نستخدم الأمر plot كالآتى - حيث أن C5 عبارة عن العمود الذى يحتوى على البواقى ، C1 عبارة عن العمود الذى يحتوى على قيم X .

```
MTB  >   Plot c5 c1
```

للحصول فقط على المعلومات الموجودة بجدول (٩-٣) (معادلة المربعات الصغرى ، جدول تحليل التباين ، . . . إلخ) فإننا نستخدم الأمر REGRESS بدون أى أوامر فرعية أخرى (أى لا تضع

علامة ؛ فى نهاية تلك الجملة) .

(٢) البرنامج الإحصائى SAS :

التعليمات التالية تولد أشكالاً وجداول مكافئة للشكل (٩-٦)، (٩-١٦)، الجداول (٩-٣)، (٩-٩). وهذه المخرجات أو النتائج الخاصة ببرنامج SAS تم وصفها بعد التعليمات التالية. وكما سبق أن قدمنا قبل ذلك يتم إستخدام الجملة INPUT لوضع عناوين لعينة للمتغيرات X ، Y وتم بإستخدام TEMP ، ENERGY على التوالى .

لاحظ أن الثلاثة سطور الأخيرة من البيانات توضح قيم X التى ترغب فى إيجاد التنبؤ لقيمة Y المقابلة وقيم X تكون متبوعة بنقط. وهذه هى الطريقة المستخدمة فى البرنامج الإحصائى SAS أما جملة ENERGY*TEMP PLOT , PROC PLOT فإنه تقوم بإنتاج بيانات مكافئة لتلك الموجودة بجدول (٩-٣) ، (٩-٩) أما الجمل الأربعة الأخيرة فإنها يمكننا من الحصول على شكل بيانى للبواقى مثل ذلك الشكل الموضح فى شكل (٩-١٦). لاحظ أنه فى جملة MODEL قمنا بتعريف المتغير التابع ENERGY وكذلك المتغير المفسر TEMP وتم وضع جميع الخيارات R , P فإن البرنامج الإحصائي SAS يمدنا بقيم Y المقدرة ($\hat{Y}$) والبواقى بفترة ثقة (95%) لمتوسط قيم Y ، وفترات تنبؤ بمقدار (95%) بقيم Y الفردية ، على التوالى .

أما المعلومات المكافئة لمخرجات أو نتائج Minitab والموجودة بجدول (٩-٩) فإنها توجد فى الصفوف الثلاثة الأخيرة فى الأعمدة من 2 إلى 6 من نتائج SAS بعد جدول تحليل التباين ANOVA وتقديرات المؤشر. أما الأعمدة الأربعة الأخيرة فى هذه النتائج فإنها تزودنا بمعلومات إضافية والتى نغطيها فى هذا الكتاب. ولإيجاد جدول تحليل التباين فقط ومقدرات المؤشر فإننا نستخدم الجملة PROG REG والجملة MODGL بدون عمل أى خيارات إضافية .

```
DATA;
INPUT TEMP ENERGY;
CARDS;
-1 94
1.5 81
3.5 79
-3 97
0.5 88
2.5 75
4 74
5 67
-5 107
-0.5 86
9 58
9.5 55
7 65
3 73
-2 91
6 65
8 58
10 52
-4 .
3 .
9 .
PROC PLOT;
PLOT ENERGY * TEMP;
PROC REG;
MODEL ENERGY = TEMP/ P R CLM CLI;
OUTPUT OUT = A
RESIDUAL = RESID;
PROC PLOT DATA = A;
PLOT RESID * TEMP;
```

Model: MODEL1
Dependent Variable: ENERGY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	4123.53791	4123.53791	616.822	0.0001
Error	16	106.96209	6.68513		
C Total	17	4230.50000			

Root MSE	2.58556	R-square	0.9747
Dep Mean	75.83333	Adj R-sq	0.9731
C.V.	3.40953		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > ?T?
INTERCEP	1	86.995715	0.75722995	114.887	0.0001
TEMP	1	-3.464188	0.13948302	-24.836	0.0001

Obs	Dep Var ENERGY	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
1	94.0000	90.4599	0.847	88.6633	92.2565	84.6919	96.2280	3.5401	2.443	1.449	? ?** ?	0.126
2	81.0000	81.7994	0.655	80.4108	83.1881	76.1451	87.4537	-0.7994	2.501	-0.320	? ? ?	0.004
3	79.0000	74.8711	0.611	73.5765	76.1656	69.2391	80.5030	4.1289	2.512	1.643	? ?*** ?	0.080
4	97.0000	97.3883	1.060	95.1402	99.6364	91.4640	103.3	-0.3883	2.358	-0.165	? ? ?	0.003
5	88.0000	85.2636	0.718	83.7415	86.7858	79.5751	90.9522	2.7364	2.484	1.102	? ?** ?	0.051
6	75.0000	78.3352	0.618	77.0258	79.6447	72.6999	83.9706	-3.3352	2.511	-1.328	? **? ?	0.053
7	74.0000	73.1390	0.619	71.8267	74.4512	67.5030	78.7750	0.8610	2.510	0.343	? ? ?	0.004
8	67.0000	69.6748	0.658	68.2800	71.0695	64.0190	75.3306	-2.6748	2.500	-1.070	? **? ?	0.040
9	107.0	104.3	1.299	101.6	107.1	98.1829	110.5	2.6833	2.236	1.200	? ?** ?	0.243
10	86.0000	88.7278	0.801	87.0306	90.4250	81.9899	94.4657	-2.7278	2.458	-1.110	? **? ?	0.065
11	58.0000	55.8180	1.010	53.6761	57.9599	49.9333	61.7028	2.1820	2.380	0.917	? ?* ?	0.076
12	55.0000	54.0859	1.067	51.8243	56.3475	48.1566	60.0153	0.9141	2.355	0.388	? ? ?	0.015
13	65.0000	62.7464	0.806	61.0385	64.4543	57.0054	68.4874	2.2536	2.457	0.917	? ?* ?	0.045
14	73.0000	76.6032	0.610	75.3096	77.8967	70.9715	82.2349	-3.6032	2.513	-1.434	? **? ?	0.061
15	91.0000	93.9241	0.950	91.9108	95.9374	88.0849	99.7633	-2.9241	2.405	-1.216	? **? ?	0.115
16	65.0000	66.2106	0.722	64.6797	67.7415	60.5197	71.9015	-1.2106	2.483	-0.488	? ? ?	0.010
17	58.0000	59.2822	0.903	57.3678	61.1966	53.4764	65.0880	-1.2822	2.423	-0.529	? *? ?	0.019
18	52.0000	52.3538	1.125	49.9694	54.7383	46.3765	58.3311	-0.3538	2.328	-0.152	? ? ?	0.003
19	.	100.9	1.177	98.3566	103.3	94.8298	106.9	.	.	.		.
20	.	76.6032	0.610	75.3096	77.8967	70.9715	82.2349	.	.	.		.
21	.	55.8180	1.010	53.6761	57.9599	49.9333	61.7028	.	.	.		.

Sum of Residuals 0
Sum of Squared Residuals 106.9621
Predicted Resid SS (Press) 132.2064

الفصل العاشر

الإنحـــدار الخطـــي المتعـــدد

MULTIPLE LINEAR REGRESSION

محتويات الفصل :

(١٠-١) نظرة عامة على محتويات الفصل .

(١٠-٢) نموذج الإنحدار الخطى المتعدد .

(١٠-٣) تقدير مؤشرات نموذج الإنحدار الخطى المتعدد .

(١٠-٤) درجة جودة النموذج: الإستنتاج الإحصائي للإنحدار الخطى المتعدد .

(١٠-٥) إدخال المعلومات الوصفية فى معادلة الإنحدار الخطى المتعدد: المتغيرات الوهمية .

(١٠-٦) المنحنى الخطى لنماذج الإنحدار .

(١٠-٧) اكتشاف النقص في النموذج وتجنب العوائق : تحليل البواقى والإرتباط الخطى

(١٠-٨) معيار لإختيار أفضل مجموعة من المتغيرات التفسيرية .

(١٠-٩) الإنحدار الخطى المتعدد : مثال شامل .

(١٠-١٠) ملخص .

ملحق ١٠ : تعليمات الحاسب الآلى لإستخدام برامج SAS وMinitab .

الفصل العاشر

الإنحـــدار الخطـــي المتعـــدد

MULTIPLE LINEAR REGRESSION

(١٠-١) نظرة عامة على محتويات الفصل : Bridging To New Topics

فى هذا الفصل ، نوسع فى مفاهيم الإنحدار الخطى البسيط. حيث أننا سنتعامل مع :

١- أكثر من متغير مفسر فى نموذج الإنحدار .

٢- إستخدام معلومات وصفية فى نموذج الإنحدار .

٣- إستخدام تحليل الإنحدار لنموذج العلاقات غير الخطية .

٤- تجنب المشاكل الشائعة فى تطبيق تحليل الإنحدار .

ويظل الهدف الأساسى وهو تحديد النموذج الأمثل الذى يجعل توفيق البيانات الممثلة ذو معنى، حتى نصل إلى أصغر خطأ عشوائى ممكن . وحيث أن التقدير والتنبؤ يظلان الأسباب الرئيسية لإستخدام نماذج الإنحدار، فإن النموذج الملائم يكون النموذج الذى يزودنا بالدقة الكافية عندما يستخدم للتقدير أو التنبؤ. ويسمى نموذج الإنحدار الذى يحتوى على أكثر من متغير مفسر بنموذج الإنحدار الخطى المتعدد. والإنحدار الخطى المتعدد ومفاهيمه يعتبر إمتداداً للإنحدار الخطى البسيط. ومع ذلك تكون التعبيرات الحسابية أكثر تعقيداً وتشتمل على جبر المصفوفات والتى تكون خارج نطاق هذا الكتاب. والإنحدار الخطى المتعدد وحساباته تكون مجهدة جداً بإنجازها باليد. بالتالى، فإن معظم الحسابات فى هذا الفصل ستكون معتمدة على إستخدام الحاسب الآلى. وكنتيجة لذلك يتحول مجهودك إلى تفسير مخرجات الحاسب الآلى. مع ذلك نؤخر تفسيرات الحاسب الآلى حتى يتم كشف بعض النقاط الأساسية الضرورية لك لفهم ذلك .

تحليل الإنحدار يمكن أن يكون أداة قوية لتحديد أسباب الإختلاف لعملية المخرجات والأكثر عمومية لفهم أفضل لصنع وإتخاذ القرار. وهو بصفة خاصة، مفيد عندما لا تستطيع التحكم فى مستويات المتغيرات الأساسية لكى ندير التجربة المصممة. لهذا نطبق عادة تحليل الإنحدار عندما تستخدم البيانات الملائمة. ومن الأهمية أن نتحكم فى إستخدامه لفهم قيوده وإدراك نتائج سوء إستخدامه. بالإضافة لفهم أسس تحليل الإنحدار المقدمة فى الفصل التاسع. فإن موضوعات هذا الفصل تتمثل فى التعلم الصحيح لإجراء نموذج إنحدار محسن وتعلم أسلوب تحديد أخطاء تطبيقات الإنحدار .

(١٠-٢) نموذج الإنحدار الخطى المتعدد : The Multiple Linear Regression Model

وسوف نقدم بناء نموذج يوضح العلاقة بين متغير تابع Y وأكثر من متغير مفسر. وبفرض أن K تمثل عدد المتغيرات المفسرة، فإن نموذج الإنحدار الخطى المتعدد يمكن التعبير عنه كإمتداد لنموذج الإنحدار الخطى البسيط كما يلى حيث :

X_1 , X_2 , ... , X_k تكون عبارة عن K من المتغيرات المفسرة

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_K X_K}_{\text{متوسط Y عند وجود قيم معينة للمتغيرات } X_1 ، X_2 ، \ldots\ldots ، K_K \text{ (العنصر المحدد)}} + \underbrace{\varepsilon}_{\text{الخطأ العشوائي والإنحراف غير المتوقع للمقدار Y عن نموذج الإنحدار للمجتمع (العنصر العشوائي)}} \qquad (10.1)$$

وكما فى الإنحدار الخطى البسيط، يحتوى نموذج الإنحدار الخطى المتعدد على مكونين: مكون العنصر المحدد ($\beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_K X_K$) ، ومكون العنصر العشوائي (ε). يعرف المكون ($\beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_K X_K$) بنموذج انحدار المجتمع **population regression model**، ويفترض بأنه متوسط Y بشرط معرفة قيم محددة للمتغيرات X_1 , X_2 , ... , X_K. أما (ε) فهو يمثل الخطأ العشوائى المرتبط بأى عنصر أو متغير فى مجتمع الدراسة ، وهو يمثل الإختلاف غير المفسر بين قيم Y عند مجموعة قيم X_1 , X_2 , ... , Xk . كما فى الفصل التاسع، فإن الخطأ العشوائى يمكن قياسه بواسطة تباين الخطأ σ_ε^2 وهو مثل المعالم β_0 ، β_1 ، .... ، β_K ، تكون غير معروفة (مجهولة) والتى يجب تقديرها اعتماداً على عينة الدراسة. وتظهر الأهمية القصوى للمقدار σ_ε^2 بقياسه دقة توفيق المنحنى. وبالطبع فإن التقدير الأقل لقيمة σ_ε^2 يعنى توفيق أفضل للبيانات .

نتذكر من الفصل التاسع ، أن شكل الإنتشار يزودنا بالمعانى الأساسية لنوعية العلاقة المحددة التى توجد بين المتغير التابع Y والمتغير المفسر X . ولسوء الحظ عندما يكون هناك أكثر من متغير مفسر. فإن الشكل البياني لا يساعدنا في رسم قيم Y مقابل عدة متغيرات لـ X، حتى يمكن معرفة وتحديد نوعية العلاقة بينهما. ومع ذلك، هناك إجراءات أخرى سوف تساعدنا على ذلك وسوف نستعرضها فى الاجزاء التالية :

ويكون من الأهمية فهم أنه عندما نصف نموذج إنحدار المجتمع المعطى فى المعادلة (10.1) كعلاقة خطية فهذا يعنى «خطية تتعلق بالمعالم» (β_1 ,β_2 , ... β_K) هذه الجملة تعنى أن كل المعالم تظهر بأس واحد. ليس هناك معالم تكون هى نفسها أس، أو مضروبة فى، أو مقسومة على معلمة أخرى. فعلى سبيل المثال يكون النموذج :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \varepsilon \qquad (10.2)$$

نموذج خطى فيما يتعلق بالمعالم β_0 , β_1 , β_2, β_3, β_4 على الرغم من أن العلاقة بين المتغير التابع والمتغيرات المفسرة X_1 , X_2 غير خطية . لهذا عندما نقول نموذج إنحدار خطى متعدد، نعنى بالخطية هنا معالم النموذج وليس للمتغيرات المفسرة للنموذج.

ولتوضيح نموذج الإنحدار المتعدد، نعتبر المثال التالى للمتغير التابع Y (المبيعات اليومية من الأيس كريم) ، والمتغيرات المفسرة X_1 = السعر للوحدة، X_2 = درجة الحرارة اليومية . فى هذه العلاقة نتوقع زيادة المبيعات إذا إنخفض السعر أو اذا إرتفعت درجة الحرارة. لهذا نتوقع أن تكون العلاقة السالبة مع السعر والعلاقة الموجبة مع درجة الحرارة.

تفسير معادلة إنحدار المجتمع : Interpreting the population Regression Equation

إفترض أن نموذج إنحدار المجتمع لمثال الأيس كريم السابق :

$$Y = 0 - 1.6\,X_1 + 4X_2$$

بالطبع فى التطبيقات الفعلية. هذا النموذج لا يكون معروف، إنما يقدر بإستخدام بيانات عينة ممثلة للمجتمع، لكن دعنا نفسر قيم المعالم كما حددناها :

١- ($\beta_0 = 0$) تمثل متوسط المبيعات اليومية إذا كان X_1، X_2 مساوى للصفر. ومن المستبعد أن يكون السعر مساوى للصفر (الايس كريم يعطي مجانا) أو أن درجة الحرارة تكون مساوية للصفر، هنا يقول النموذج أن متوسط المبيعات اليومى صفر .

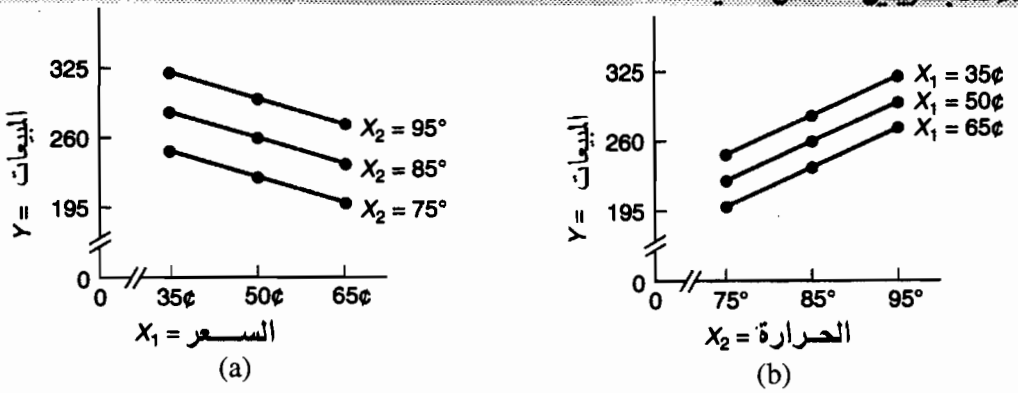
٢- المعامل الخاص بالسعر ($\beta_1 = -1.6$) يعنى أن المتوسط اليومى للمبيعات يقل بواسطة 1.6 أوقية لكل سنت يرفع به السعر، وذلك مع ثبات درجة الحرارة. وكتوضيح إعتبر أنه فى يوم معين ($X_1 = 50$) , ($X_2 = 85°$). القيمة المتوقعة للمبيعات {$Y = 0-1.6(50) + 4(85)= 260$} أوقية. لكن إذا زادت X_1 إلى 60 ، وظلت X_2 85° ، تكون القيمة المتوقعة 244 أوقية. أى أن القيمة المتوقعة للمبيعات إنخفضت بمقدار (16 = 244 - 260) أوقية أو 1.6 أوقية لكل سنت يزيد به السعر.

٣- بالمثل معلمة درجة الحرارة ($\beta_2 = 4$) تشير إلى أن متوسط المبيعات اليومية تزيد بمقدار 4 أوقية لكل زيادة درجة واحدة فى درجة الحرارة اليومية مع ثبات السعر .

وكتوضيح إضافى للنموذج المفسر فإن الجدول التالى يعطى قيم متوسط المبيعات الخاصة بتسع توليفات سعرية حرارية :

	PRICE X_1		
Temperature X_2	65	50	35
75	196	220	244
85	236	260	484
95	276	300	324

يوضح شكل(١٠-١ أ) العلاقة بين متوسط المبيعات اليومية مع السعر عندما تكون درجة الحرارة ثابتة. ويوضح شكل (١٠-١ ب) العلاقة بين متوسط المبيعات مع درجة الحرارة عندما يكون السعر ثابت.



شكل رقم (١٠): متوسط المبيعات مقابل ثلاث مستويات للسعر (أ)، ومقابل ثلاث مستويات للحرارة (ب)

ومن الشكل السابق نلاحظ الآتى :

١- العلاقة الخطية بين متوسط Y , X_1 لها نفس ميل الإنحدار لأى قيمة X_1 طالما X_2 ثابتة. ونفس الجملة صحيحة للعلاقة الخطية بين Y , X_2 لأى قيمة X_2 طالما X_1 ثابتة. لهذا فإن الخطوط تكون متوازية .

٢-عند درجة حرارة معينة X_2، الفرق بين أى خطين يشتملا على السعر X_1 يكون ثابتا. فمثلاً، عند $X_1 = 35$، $X_1 = 50$، نجد أن المسافة بين الخطوط تكون 24 وحدة دائماً. وهكذا فإن متوسط Y يتناقص بمقدار 24 وحدة، كلما زادت X_1 بمقدار 15 سنت، بشرط أن تظل X_2 ثابتة. هذا التناقص في Y هو 1.6 أوقية لكل زيادة واحد سنت في السعر.

٣-عند قيمة معينة للسعر X_1، الفرق بين أى خطين يشتملا على درجة الحرارة X_2 يكون ثابتا. فمثلاً، عند $X_2=85°$ ، $X_2=95°$ نجد أن المسافة بين الخطوط تكون ثابتة 40 وحدة. وهكذا فإن متوسط Y يتزايد بمقدار 40 وحدة، كلما زادت درجة الحرارة 10°، بشرط أن تظل X_1 ثابتة. هذه الزيادة في متوسط Y وهي 4 أوقيات لكل زيادة درجة واحدة في الحرارة.

(١٠-٣) تقدير معالم نموذج الإنحدار الخطى المتعدد :

Estimating the Parameters of the Multiple Linear Regression Model :

لتقدير نموذج إنحدار المجتمع ، تستخدم بيانات عينة لتقدير معالم النموذج β_0 ، β_1 ، . . . . ، β_n وكذلك تباين الخطأ σ_ε^2. وطرق الحصول على بيانات عينة هى نفسها كما فى جزء (٩-٣) . من هذه الطرق ، تسجيل قيم Y عند القيم التي سبق تحديدها للمتغيرات التفسيرية $X_1, X_2, \ldots, X_K$. أو استخدام البيانات الملائمة. ومن المهم أن تكون هناك عناية فائقة في إختيار قيم المتغيرات المستقلة، ولكن في كثير من الحالات لايكون أمامنا فرصة لاختيار البيانات. يجب أن تضع في ذهنك أن بيانات العينة التي تستخدم لتقدير معالم النموذج، يجب أن تكون ممثلة ومعبرة للبيئة التي نرغب في دراستها.

(١٠-٣-١) طريقة المربعات الصغرى The Method of least Squares

كما فى الفصل التاسع ، تستخدم طريقة المربعات الصغرى لتحديد أفضل الإحصاءات لتقدير المعالم β_0 ، β_1 ، . . . . ، β_k وتستخدم كتقديرات القيم التى تكون مجموع مربعات البواقى أصغر

لبيانات العينة. وتقديرات المربعات الصغرى للمعالم β_0 ، β_1 ، . . . . ، β_k يرمز لها بالرموز b_0 , ... , b_K على التوالى. وكما فى الإنحدار البسيط فإننا نحدد قيم b_0 ..... , b_K التى تجعل مجموع مربعات الخطأ أقل ما يمكن. بمعلومية هذه القيم تكون معادلة إنحدار المربعات الصغرى كما يلى :

$$\hat{Y} = b_0 + b_1X_1 + \ldots\ldots\ldots\ldots + b_KX_K \qquad \textbf{(10.3)}$$

وصيغ المربعات الصغرى لـ b_0 ، ..... ، b_K لن تعرض هنا لأنها تشمل جبر المصفوفات. بالتالى نعتمد كلياً على إستخدام الحاسب الآلى فى تحديد تقديرات المربعات الصغرى لمعالم النموذج. ففى نموذج الآيس كريم. إفترض أننا حصلنا على البيانات الممثلة لعينة 10 أيام:

Y (daily sales)	374	386	471	429	391	475	428	412	405	341
X_1 (Price)	35	35	35	50	50	50	50	65	65	65
X_2 (high temperature)	74	82	94	93	82	96	91	93	88	78

وكما سنرى فيما بعد عندما نفسر نتائج الحاسب الآلى لهذا المثال. فإن تقدير المربعات الصغرى هي: $(b_0 = 25.8777)$, $(b_1 = -1.3418)$, $(b_2 = 5.1953)$ للنموذج $Y= \beta_0 + \beta_1 X_1+ \beta_2X_2 + \varepsilon$ ، ولهذا فإن معادلة الإنحدار لأصغر المربعات تكون :

$$\hat{Y} = 25.8777 - 1.3418X_1 + 5.1953X_2$$

(١٠-٣-٢) تقدير تباين الخطأ σ_ε^2 Estimating the Error. Variance σ_ε^2

إن إجراء تقدير تباين الخطأ σ_ε^2 يتشابه مع ما قدم في الفصل التاسع. وهذا يعنى أن المقدر S_e^2 يجب أن يعتمد على مدى انحراف القيم الفعلية Y عن القيم المتنبأ بها ($\hat{Y}$) والمحددة من معادلة المربعات الصغرى. هذه الانحرافات هي البواقي. لهذا فإن بسط التباين S_e^2 يظل مجموع مربعات البواقى، كما كان فى الإنحدار الخطى البسيط والمقام للتباين S_e^2 في الأنحدار الخطى البسيط (n-2). والآن ماذا نتوقع أن يكون مقام S_e^2 فى الإنحدار الخطى المتعدد؟ نذكر أنه فى تحديد قيم $\hat{Y}$ ، يجب تقدير (K + 1) معلمة أو مؤشر (β_0 , β_1 , .. β_K) للنموذج ($Y= \beta_0 + \beta_1 X_1+ \beta_2X_2 +\ldots+\ldots+ \beta_KX_k + \varepsilon$) لذا ولكى نجعل تباين البواقى S_e^2 مقدر غير متحيز للبواقى σ_ε^2 ، يجب طرح (K+1) وحدة من n فى مقام S_e^2 . طبقا لذلك ، يعرف تباين البواقى S_e^2 كما يلى :

$$S_e^2 = \frac{SSE}{n-(k+1)} \qquad \textbf{(10.4)}$$

حيث أن :

$$SSE = \Sigma(Y_i - \hat{Y}_i)^2 \qquad \textbf{(10.5)}$$

وهو يعبر عن مجموع مربعات البواقى . وكما سبق يمكن تقدير σ_ε بواسطة الإنحراف المعيارى للبواقى .

$$S_e = \sqrt{S_e^2} \qquad \textbf{(10.6)}$$

ويظل تباين البواقى S_e^2 مقياساً لكيفية توفيق معادلة المربعات الصغرى لقيم Y من العينة. إذا كان التوفيق تام، كل البواقى تساوى الصفر، وبالتالى S_e^2 تساوى الصفر. عندما يشتمل النموذج على العديد من المتغيرات المفسرة (بعضاً منها قد لا يكون مساعد فى تفسير الإختلاف فى قيم Y) ، يكون تباين البواقى S_e^2 ذو أهمية كبيرة كمعيار لجودة التوفيق وسيكتشف ذلك لاحقاً. مرة أخرى، يعرف S_e^2 بمتوسط مربعات الخطأ (MSE) ، أما S_e يشير إلى جذر متوسط مربعات الخطأ (RMSE) .

على الرغم من إستخدام الحاسب الآلى بشكل كبير. إلا أنه يمكننا حساب تباين البواقى S_e^2 بالحسابات اليدوية إذا أعطينا معادلة المربعات الصغرى المنسجمة مع بيانات العينة. للتوضيح، بالرجوع لمعادلة المربعات الصغرى لمثال الأيس كريم وهي : ($\hat{Y} = 25.8777 - 1.3418X_1 + 5.1953X_2$) حيث X_1 تمثل السعر، X_2 درجة الحرارة. فى اليوم الأول من العينة، السعر ($X_1 = 35$ cents) ودرجة الحرارة ($X_2 = 74°$ F) . بهذه القيم المفسرة فإن $\hat{Y} = 25.8777 - 1.3418(35) + 5.1953(74) = 363.368$ لذلك ، البواقى ($Y_1 - \hat{Y}_1 = 374 - 363.368 = 10.632$) . فى اليوم الثانى عندما ($X_1 = 35$) , ($X_2 = 82$) ، فإن المبيعات المتنبأ بها ($\hat{Y} = 404.930$) . وتكون البواقى ($Y_2 - \hat{Y}_2 = 386 - 404.930 = -18.93$) . وبإستمرار هذا الأسلوب نجد أن قيم$\hat{Y}$ وبواقيها لمثال للأيس كريم كما يلى :

Price (X_1)	Temperature (X_2)	Sales (Y)	Predicted ($\hat{Y}$)	Residual $e = y - \hat{Y}$
35	74	374	363.368	10.631
35	82	386	404.93	-18.93
35	94	472	467.274	4.726
50	93	429	44.952	-12.852
50	82	391	384.804	6.196
50	96	475	457.534	17.462
50	91	428	431.562	-3.562
65	93	412	421.826	-9.826
65	88	405	395.849	9.151
65	78	341	343.897	-2.897

وكنتيجة لهذا ، فإن مجموع مربعات البواقى :

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = (10.632)^2 + (-18.93)^2 + \ldots + (-2.897)^2 = 1,206.158$$

ولهذا ، فإن تباين البواقى يكون :

$$S_e^2 = \frac{1,206.158}{10-3} = 172.308$$

والإنحراف المعيارى للبواقى يكون :

$$S_e = \sqrt{172.308} = 13.127$$

(١٠-٣-٣) معامل التحديد The Coefficient of Determination

معامل التحديد فى الإنحدار الخطى المتعدد له نفس التفسير كما فى الإنحدار الخطى البسيط. فهو يمثل نسبة إجمالى الإختلاف فى قيم العينة Y التى تفسر بواسطة المتغيرات التفسيرية فى معادلة إنحدار المربعات الصغرى .

إفترض أن لدينا عينة والتى تحدد لها معادلة المربعات الصغرى. نسأل نفس السؤال البسيط كما فى الفصل التاسع: لماذا تختلف قيم Y فى العينة ؟ مرة أخرى، هناك إجابتان محتملتان :

١-تختلف قيم Y فى العينة ، لأن المتغير التابع يكون مرتبطاً بالمتغيرت المفسرة X_1 , ... , X_k . فكلما تغيرت قيم المتغيرات المفسرة، تميل قيم Y للتغير. فى مثال الأيس كريم، إذا تغير السعر (و / أو) درجة الحرارة فإن مستوى المبيعات اليومية تميل للتغير .

٢- تختلف قيم Y فى العينة ، بسبب عوامل أخرى غير المتغيرات التفسيرية فى النموذج. على سبيل المثال، إذا لم يتغير السعر ودرجة الحرارة لعدة أيام، فإن مستوى المبيعات اليومى سوف يختلف لأسباب أخرى. والأسباب الإضافية للإختلاف تفترض أنها تؤثر في المتغير التابع في نمط عشوائي.

وكما فى الفصل التاسع ، فإن الإختلاف الكلى فى العينة لقيم Y يقاس بواسطة SST ، مجموع المربعات الكلى [انظر للصيغة (9.10)]. بالإضافة إلى الإختلاف غير المفسر فى قيم Y والذي يقاس أيضاً بواسطة SSE ، مجموع مربعات البواقى. الفرق بين SST , SSE يعطى مجموع مربعات الإنحدار SSR والتى تقيس الإختلاف فى Y، والتى تكون راجعة إلى التغيرات بين قيم المتغيرات التفسيرية فى النموذج ، وبهذا فإن:

$$SST = SSR + SSE \qquad \textbf{(10.7)}$$

ومعامل التحديد يكون عبارة عن النسبة بين مجموع مربعات الإنحدار إلى مجموع المربعات الكلى.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \qquad \textbf{(10.8)}$$

نلاحظ أن حرف R الكبير يستخدم عادة ليشير إلى معامل التحديد فى الإنحدار الخطى المتعدد، بينما r الصغيرة تشير إلى معامل التحديد فى الإنحدار الخطى البسيط .

فى مثال الأيس كريم حددنا (SSE = 1,206.158) ولحساب مجموع المربعات الكلى فإن :

$$SST = (374)^2 + (386)^2 + ... + (341)^2 - \frac{(374 + 386 + ... + 341)^2}{10}$$

$$= 15,760.1$$

إذن :

$$SSR = 15,760,1 - 1,206.158 = 14,553.942$$

$$R^2 = \frac{14,553.942}{15,760.1} = .9235$$

وهذا يعنى أن 92.35% من إجمالى الإختلاف فى قيم Y يرجع إلي الأختلافات في قيم العينة: قيم X_1 (السعر)، X_2 (درجة الحرارة) .

الاستخدام المعيب لـ R^2 : Misapplication of R^2

فى سياق الحديث عن تحليل الإنحدار الخطى المتعدد عادة ما يساء فهم R^2 أو يساء استخدامه. ومن الأهمية أن تعلم أن R^2 لا يمكن أن تقل عندما تضاف متغيرات مفسرة إلى نموذج الإنحدار، حتى ولو كانت هذه المتغيرات لا تساهم بمعلومات إضافية للتنبؤ بقيمة Y . وهذا صحيح لأن الإختلاف غير المفسر فى العينة لقيم Y ، كما قيست بواسطة SSE تتناقص بوضوح عندما يكون هناك حد اضافي، قد أضيف لنموذج الإنحدار بينما يظل مجموع المربعات الكلى SST ثابتا بصرف النظر عن عدد المكونات فى النموذج (لأن SST تكون محددة كلياً بواسطة قيم X) . لهذا، فإن مجموع مربعات الإنحدار SSR (الإختلاف المفسر) يجب أن يزيد على الأقل عندما تضاف حدود جديدة إلى النموذج.

فإذا استخدمت R^2 ، لتحديد ما إذا كان يجب إضافة عناصر جديدة للنموذج أم لا. إذن السؤال لا يكون ما اذا كانت هناك زيادة في قيمة R^2 عند إضافة متغيرات جديدة ولكن بكم تزيد R^2 . R^2 الكبيرة لا تعنى بالضرورة نموذج أفضل. فى الحقيقة R^2 الكبيرة بدرجة كافية يمكن تحقيقها ببساطة باضافة متغيرات تفسيرية، البعض منها ربما يساهم في تفسير القليل من التغيرات في قيم Y بالعينة. فى مثال الأيس كريم، نجد أن إضافة درجة الحرارة للنموذج المحتوى على السعر فقط يعتبر مفيد إذا كانت الإضافة تزيد مجموع مربعات الإنحدار و بالتالى تزيد R^2 . بعض المحللين يخطئون بضم عدد كبير من المتغيرات المفسرة فى النموذج كأساس للحصول على قيمة عالية لقيمة R^2 .

وهناك إستخدام خاطئ آخر شائع، وفيه يفترض أن النموذج يكون جيداً إذا كانت R^2 عالية، وغير جيد (أي سئ) إذا كانت R^2 منخفضة. فالنموذج الذى به ($R^2 = 0.6$) ربما يكون جيد إذا كان الهدف إنشاء علاقة موجودة بين قيمY والمتغيرات المفسرة فى النموذج. وعلى العكس فإن نموذج به ($R^2 = 0.9$) قد يكون نموذج غير جيد إذا كان ($R^2 = .99$) يمكن تحقيقها بأشتمال النموذج على متغيرات مفسرة لها معنى أكثر وضوحاً .

معامل التحديد المعدل The Adjusted Coefficient of Determination

كما سبق أن ذكرنا فإن R^2 تزيد بزيادة العناصر المضافة لنموذج الإنحدار. فبإضافة عناصر كافية للنموذج يمكن أن تقترب R^2 من الواحد الصحيح. بهذا فإن R^2 التى قيمتها تساوى 95. تكون مؤثرة لنموذج به أربعة عناصر عن نموذج به 30 عنصر. لهذا السبب فإن هناك علاقة بديلة لقياس جودة التوفيق قد أقترحت لتأخذ عناصر النموذج فى الحسبان. وهذا المقياس الوصفى لجودة توفيق معادلة المربعات الصغرى يسمى معامل التحديد المعدل ويعرف كما يلي:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SST}} \qquad \textbf{(10.9)}$$

حيث P عدد عناصر النموذج شاملة الجزء المقطوع من المحور الرأسى. بمعنى آخر P عدد المعالم β فى النموذج. لاحظ أن تعبير R_a^2 تختلف عن R^2 فقط بالكسر $\left(\frac{n-1}{n-p}\right)$. حيث أن هذا الكسر يجب أن يزيد عن الواحد، فإن قيمة R_a^2 دائماً أقل من R^2 . بالإضافة لذلك، من الممكن لقيمة R_a^2 أن تتناقص عندما تضاف متغيرات مفسرة غير مناسبة لنموذج الإنحدار. لهذا فإنه فى تحليل الإنحدار الخطى المتعدد يفضل معامل التحديد المعدل R_a^2 على R^2 للمقارنة بين نماذج الإنحدار المتنافسة. ومن المهم أن نعلم أن R_a^2 تمثل مؤشر إحصائى وصفى فقط. وسوف نعرض إجراء استنتاجى فى الجزء (١٠-٥) لنقرر ما إذا كان مساهمة العناصر المضافة فى نموذج الإنحدار تكون كافية لتبرير بقائها فى النموذج.

ولتوضيح حساب R_a^2 وبالعودة إلى مثال الأيس كريم حيث حجم العينة ($n = 10$) , ($p = 3$) عناصر فى النموذج ومجموع مربعات البواقى ($SSE = 1,206.158$) ومجموع المربعات الكلى ($SST = 15,760.1$) فإن :

$$R_a^2 = 1 - \left(\frac{10-1}{10-3}\right)\frac{1,206.158}{15,760.1} = .9016$$

مثال (١٠-١)

إفترض أن نموذج الإنحدار لمتغير تابع Y فى مقابل متغيرين مفسرين X_1 , X_2 ، وجد أن معامل التحديد ($R^2 = .621$) ، ومعامل التحديد المعدل ($R_a^2 = .585$). وعند إضافة المتغير X_3 لمعادلة الإنحدار Y فى مقابل X_1 , X_2 , X_3 ، وجد أن ($R^2 = .629$) ، ($R_a^2 = .575$) هل زيادة R^2 تشير إلى أن ضم X_3 أفضل للنموذج ؟

الحل

الإجابة وبصوت عالى (لا) ، لأن قيمة أى متغير (أو أى حد)، حتى لو كان غير مناسبا، سوف يزيد من R^2 بمقدار صغير. والزيادة الصغيرة هنا من 621. إلى 629. عندما أضيف X_3 للنموذج. من المؤكد أن هذه الاضافة ليست مؤثرة. هذه النتيجة مدعمة بوضوح من تناقص قيمة R_a^2 من 585. إلى 575. والتى تعنى أن ضم X_3 ليس مضموناً من الناحية الإحصائية.

تمـــارين

(١٠-١) إفترض أن النموذج ($Y = 1,5 + 1,2X_1 + .15X_2$) هو المستخدم فى وكالة تأجير سيارات لتقدير تكلفة الصيانة السنوية Y (بالألف دولار) كدالة فى عدد السيارات المؤجرة X_1 ومتوسط عدد الأميال لكل سيارة X_2 (بالألف ميل) .

أ - اشرح معنى قيمة كل معلمة .

ب- حدد التكلفة السنوية للصيانة لكل التوليفات التالية لقيم X_1 , X_2 :

$(600 , 400 , 200 = X_1)$ $\qquad$ $(40 , 30 , 20 = X_2)$

(ملاحظة : سيوجد لدينا تسع توليفات)

ج- إستخدم نتائج (ب) لرسم العلاقة بين Y , X_1 مع ثبات X_2 ثم العلاقة بين Y , X_2 مع ثبات X_1.

(١٠-٢) إفترض النموذج الذي يعبر عن المتغير التابع Y كدالة فى المتغيرات التفسيرية X_1، X_2 كالتالى : $(Y = 15 + 6X_1 - 2X_2 - 1.5X_2^2)$

أ - ارسم العلاقة بين Y , X_1 عندما تكون $(X_2 = 2)$

ب- ارسم العلاقة بين Y , X_2 عندما تكون $(X_1 = 1)$

(١٠-٣) حدد أى واحد من النماذج الإنحدارية التالية تعتبر خطية بالنسبة لمعالمها. اشرح إجابتك.

أ - $Y = \beta_0 + \beta_1 X + \beta_2 X_2^2 + \varepsilon$

ب - $Y = \beta_0 + \beta_1 \log X_1 + \beta_2 X_2 + \varepsilon$

ج - $Y = \beta_0 e^{\beta_1 X} + \varepsilon$

د - $Y = \beta_0 + \beta_1 X_1^{\beta_2} + \beta_3 X_2 + \varepsilon$

(١٠-٤) البيانات العينة التالية تكون لها معادلة المربعات الصغرى كما يلى :

$$\hat{Y} = 5.34 + .6065X_1 + .0942X_2$$

Y	75	58	50	100	60
X_1	100	66	88	150	75
X_2	22	75	44	33	100

أ - بالنسبة لهذه البيانات . هل تقدير bo = 5.34 له معنى حقيقي؟ أعطى سبباً لإجابتك .

ب- قدر متوسط Y عندما $(X_1 = 90)$, $(X_2 = 65)$

ج- حدد قيمة SSE ثم حدد تباين البواقى S_e^2 .

د - حدد معامل R^2 وأشرح معناه لهذا التمرين .

هـ - حدد معامل التحديد المعدل R_a^2 وأعطى سبباً لماذا قيمة R_a^2 المعدلة أصغر بكثير من R^2 لهذا التمرين .

(١٠-٥) بالإشارة إلى تمرين رقم (١٠-٤) وفق الخط المستقيم لبيانات العينة بإستخدام X_1 فقط .

أ - حدد SSE وتباين البواقى للخط المستقيم وقارن نتائجك مع (ج) فى تمرين رقم(١٠-٤)، واشرح النتائج التي توصلت إليها.

ب- حدد R^2 , R_a^2 وقارن النتائج مع (د) فى تمرين (١٠-٤) واشرح ما تنتج عنه تلك المقارنة.

ج - من وجهة نظر إجابتك للأجزاء (أ) ، (ب) لهذا التمرين ما هو الإستنتاج المعقول الذى يمكن أن تستنتجه بخصوص X_2 .

(١٠-٦) إفترض أن نموذج الإنحدار $(Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon)$ تم توفيقه لعينة من (n=10) مشاهدات وأعطى النتائج التالية: $R_2 = .927$, $SSE = 121.2$. فيما بعد أضيف X_3 إلى مجموعة المتغيرات المفسرة ليصبح النموذج على الصورة $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$ وأعطى النتائج التالية:

$$R^2 = .930 , SSE = 116.2$$

أ - فى النموذج الأول، $(R^2 = .927)$ تعتبر قيمة عالية. هل هذا يشير إلى أن النموذج جيد التنبؤ بالمتغير Y ؟ علق على إجابتك .

ب- عند إضافة X_3 للنموذج فإن SSE انخفضت وارتفعت R^2 . هل هذه النتائج تشير إلى أن النموذج الثانى محسن عن النموذج الأول ؟ إشرح لماذا نعم أو لماذا لا .

ج- حدد تباين البواقى S_e^2 ومعامل التحديد المعدل لكل نموذج (ملاحظة : إستخدم R^2 , SSE لإيجاد SST) . ماذا تستنتج من هذه المعلومات عن المزايا النسبية لكلا النموذجين) .

(١٠-٧) تم تحديد نموذج إنحدار ليعبر عن المتغير Y فى مقابل 4 متغيرات تفسيرات لعينة مكونة من 22 مشاهدة وكانت SSE = 1225 ، SST = 14570 لهذا النموذج .

أ – حدد R^2 واشرح معناه .

ب– معتمداً على قيمة R^2 فى (أ)، هل تعتقد أن النموذج جيد للتنبؤ (Y)؟ وضح إجابتك .

ج – حدد قيمة R_a^2 .

(١٠–٨) إفترض أن نموذج المربعات الصغرى على الصورة التالية :

$$\hat{Y} = 22 + 5.2X_1 - .32X_1 + 2.7X_3 - 9.6X_4$$

تم استخدامه لتوفيق عينة مكونة من (n = 28) مشاهدة ، وكانت ($R^2 = .65$) ، (SSE = 15.22) لهذا النموذج .

أ – هل SSE , R^2 تفيد في أن هذا النموذج جيد التنبؤ بـالمتغير Y ؟ وضح إجابتك .

ب– حدد قيم SSR , SST , R_a^2 واشرح لماذا R_a^2 أصغر من R^2 .

(١٠–٩) إفترض اننا نستخدم نموذج المربعات الصغرى التالى :

$$\hat{Y} = 3.06 - .22X_1 + 8.44X_2 - 2.44X_3$$

لعينة (n = 100) مشاهدة، وكانت: $(\Sigma(Y_i - \overline{Y})^2 = 855.1)$, $(\Sigma(Y_i - \hat{Y}_i)^2 = 144.3)$ لهذا النموذج .

أ – حدد قيم SSE , SST , R^2

ب – حدد R_a^2 . هل يمكنك القول بأن النموذج جيد التنبؤ بالمتغير Y ؟ اشرح إجابتك .

(١٠–١٠) اذا كانت معادلة المربعات الصغرى للبيانات التالية هي :

$$\hat{Y} = 120.86 + 23.492X_1 - 2.2596X_2$$

Y	158	165	145	172	179	152	154
X_1	2.2	2.1	1.9	2.4	2.8	2.3	2.6
X_2	5	3	8	6	2	10	12

أ– بخصوص هذه البيانات ، هل تقدير ($b_0 = 120.86$)له معنوية حقيقية ؟ اشرح إجابتك .

ب– قدر متوسط Y عندما $(X_1 = 2.5)$, $(X_2 = 4)$

ج – حدد SSE , S_e^2

د – حدد R^2 , R_a^2 . هل يمكنك القول أن معادلة النموذج جيدة للتنبؤ بالمتغير Y ؟ اشرح إجابتك .

(١٠–١١) بالإشارة إلى تمرين(١٠–١٠)، وفق خطان مستقيمان، أحدهما بإستخدام X_1 فقط والآخر بإستخدام X_2 فقط .

أ – حدد SSE، S_e^2 لكلا الخطين وقارن بين نتائجك ونتائج الجزء (ج) فى تمرين(١٠–١٠) إشرح إجابتك .

ب– حدد R^2 , R_a^2 للخطين وقارن نتائجك بنتائج الجزء (د) فى تمرين (١٠–١٠) اشرح النتائج التي تصل إليها .

ج- ارسم البواقى لخط الإنحدار الذي به X_1 ، فى مقابل القيم المناظرة لمتغير X_2 . افعل نفس الشيء للبواقى الخاصة بخط الإنحدار X_2 مقابل قيم X_1 المناظرة؟ اشرح إجابتك .

د - من خلال إجابتك فى الأجزاء (أ) إلى (ج) لهذا التمرين. ما الإستنتاج المعقول الذي توصلت إليه لكلا المتغيرين X_1 , X_2 ؟

(١٠-٤) كيف يكون النموذج جيداً ؟ الإستنتاج الإحصائى للإنحدار الخطى المتعدد

How good is The Model? Statistical Inference for Multiple Linear Regression

يستخدم الإنحدار المتعدد بنفس الطريقة التى يستخدم بها الإنحدار الخطى البسيط، بهدف الوصول إلى علاقة مفهومة بين المتغيرات العملية، لتقدير متوسط Y أو التنبؤ بقيم Y بمعلومية مجموعة من قيم المتغيرات المفسرة فى النموذج. للتوضيح نعود إلى مثال الأيس كريم كما ناقشناه من قبل حيث تكون معادلة المربعات الصغرى : ($\hat{Y} = 25.8777 - 1.3418X_1 + 5.1953\ X_2$) وتساعدنا المعادلة على فهم حساسية المبيعات Y للتغيرات فى السعر X_1 ، ودرجة الحرارة X_2 . ويزودنا النموذج أيضاً بتقدير لمتوسط المبيعات أو التنبؤ بمبيعات يوم عند سعر ودرجة حرارة معينة. لكن كما كان الحال فى الفصل التاسع، لايمكن إستخدام نموذج المربعات الصغرى إلا إذا تم تقييمه بكل دقة .

والقضايا التي تتعلق بتقييم نموذج المربعات الصغرى، هى نفس القضاياالتي وضحت في الجزء (٩-٤)، ونعيدها مرة أخرى وهى:

١- هل بيانات العينة تشير بقناعة إلى العلاقة الموجودة بين Y المتغيرات المفسرة كما حددت فى نموذج الإنحدار ؟

٢- ما دقة التقديرات أو التنبؤات بمعادلة المربعات الصغرى ؟ وما هى إعتبارات الدقة التي تتضمن تقدير المعالم β's وتقديرات وتنبؤات المتغير التابع Y ؟

٣- هل هناك أى خلل يمكن توضيحه بالنسب للفروض الأساسية للإستنتاج الإحصائى؟ فكما فى الفصل التاسع يلعب تحليل البواقى دوراً كبيراً فى هذا الخصوص .

عموماً، سيعاد حل هذه القضايا بإستنتاج إحصائى ملائم لمعادلة المربعات الصغرى. مشتملاً فترات الثقة وإختبارات الفروض للمعالم β's . وهذه الإستنتاجات تعتبر إمتداداً لتلك الموجودة فى الفصل التاسع والخاصة بنموذج الإنحدار البسيط. فمثلا تعتمد طرق الإستنتاج الإحصائي على إفتراضات موجودة فى جزء (٩-٤-١) وسوف نعيدها هنا مرة أخرى.

ملخص لفروض الإستنتاج فى الإنحدار الخطى المتعدد :

Summary of Assumptions for Inferences in Multiple linear Regression

١-نموذج الانحدار المحدد له صيغة صحيحة. فنموذج الإنحدار ($Y = \beta_0 + \beta_1X_1 + ... + \beta_KX_K + \varepsilon$) يمثل بصورة صحيحة شكل العلاقة بين المتغير التابع ومجموعة المتغيرات المفسرة. عند قيم معينة X_1 , X_2 , ... , X_k للمتغيرات المفسرة يكون { $E(Y) = \beta_0 + \beta_1X_1 + ... + \beta_KX_K$ } ويكون متوسط الخطأ العشوائي مساويا للصفر. معنى ذلك أنه عندما تكون تقديرات المربعات الصغرى محددة فإن معادلة المربعات الصغرى ($\hat{Y} = b_0 + b_1X_1 + ... + b_KX_K$) تقدر متوسط قيمة Y لمجموعة من قيم المتغيرات المفسرة .

٢- تباين الخطأ مقدار ثابت . تباين الخطأ σ_ε^2 مقدار ثابت لكل قيم المتغيرات المفسرة . لهذا فإن مدى إختلافات قيم Y من نموذج الإنحدار تكون واحدة بغض النظر عن قيم المتغيرات المفسرة .

٣- الأخطاء العشوائية تكون مستقلة وتتبع التوزيع الطبيعي. الأخطاء العشوائية المرتبطة بقيم Y تكون مستقلة إحصائياً عن بعضها البعض. وتتوزع طبقاً للتوزيع الطبيعي .

(١٠-٤-١) الإستنتاجات الإحصائية للنموذج الكامل : أسلوب تحليل التباين
Statistical Inferences on the Overall Model: An Analysis of Variance Approach

من المحتمل أن بعض أو كل المتغيرات التفسيرية في معادلة المربعات الصغرى لا تكون مفيدة في تفسير الإختلاف في قيم Y . ومن الأغراض الأساسية لتقييم النموذج هو تحديد أي من هذه المتغيرات المفسرة، أن وجد، يجب أن يتواجد في معادلة الإنحدار. لكن يجب أن نبحث أولاً ما إذا كانت العلاقة بين Y وأي من المتغيرات المفسرة المحددة موجودة أم لا . لهذا نستخدم تحليل التباين .

ففي مثال الأيس كريم ربما نسأل ما إذا كانت توجد علاقة واضحة بين المبيعات Y وأي من المتغيرات التفسيرية: السعر X_1 ، ودرجة الحرارة اليومية X_2 .افترض أننا مؤقتا أفترضنا أنه لاتوجد علاقة بين لا وكل من X_1، X_2. هذا يعني أن المعاملات β_1 ، β_2 في نموذج إنحدار المجتمع تساوي صفر. سنأخذ هذا على أنه الفرض العدمي. إذا كانت العلاقة غير واضحة، إذن على الأقل فإن واحدة من المعالم β_1 ، β_2 لا تساوي الصفر، ويكون هذا هو شكل الفرض البديل .

في نموذج الإنحدار : $(Y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + \varepsilon)$ نجد أن إجراء تحليل التباين يتشابه مع ماجاء في الفصل التاسع، ويكون ملائماً لإختبار الفرض العدمي :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$$

في مقابل الفرض البديل: على الأقل واحد من المعالم β_1 ، β_2 ، . . . ، β_k لايساوي الصفر: H_a

تذكر أن نموذج إنحدار المجتمع يتكون من جزئين (جزء محدد وجزء عبارة عن خطأ عشوائي) بنفس الأسلوب كما في الفصل التاسع نجد أنه في تحليل التباين يمكن تجزئة مجموع المربعات الكلي إلي جزئين واللذان يمثلان جزئي نموذج الإنحدار. وهكذا يتم تجزئة الإختلاف الكلي لقيم Y إلي جزئين : (1) تغير أو إختلاف يعزى إلى تغير في قيم المتغيرات المفسرة في النموذج. (2) تغير أو إختلاف يعزى إلى الخطأ العشوائي. وبلغة مجاميع المربعات (SST = SSR + SSE) . هذه الكميات تم تفسيرها في جزء (١٠-٣-٣) ووضحت بإستخدام مثال الأيس كريم .

ومن المهم فهم أن SSR , SSE لهما علاقة تكاملية. تظل SST ثابتة لكل بيانات العينة المعطاة بصرف النظر عن أي متغيرات تفسيرية موجودة في النموذج، لأنها ببساطة تعتمد فقط على قيم Y. لكن SSR , SSE مجموعهما يجب أن يساوي SST . لذا إذا زاد أحدهما ينقص الآخر. أفضل معادلة مربعات صغرى تلائم بيانات العينة، تلك التي لها أصغر قيمة لـ SSE وأكبر قيمة لـ SSR. في الحالات المتطرفة، فإن معادلة المربعات الصغرى التي تلائم بيانات العينة بدقة تامة يكون فيها: SST = SSR & SSE = 0.

الأحصاء الأساسي في منهج تحليل التباين هو النسبة بين مجموع مربعات الانحدار والخطأ، أي: MSR/MSE. وكما في الفصل التاسع، فإن مجموع المربعات طور بالقسمة على درجات الحرية الملائمة والمناظرة له. نتذكر من الجزء (٩-٤-٥) أن عدد درجات الحرية لـ SSR

هو عدد الحدود التي تشتمل على المتغيرات التفسيرية في نموذج الانحدار. في النموذج: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + \varepsilon$ ، يوجد K من مثل هذه الحدود. وبالتالى فهناك K من درجات الحرية مقترنة بـ SSR. في الجزء (١٠-٣-٢) بينا أنه طالما أن هناك (K+1) من المعالم β يجب تقديرها، فإن المقام عند تحديد متوسط مربعات الخطأ (تباين البواقي S^2_e) يكون : n-(K+1). وعلى ذلك، يوجد n-(K+1) من درجات الحرية تكون مرتبطة بـ SSE. بهذا نجد أن متوسط مربعات الانحدار والخطأ يكونا على الصورة التالية:

$$MSR = \frac{SSR}{K} \qquad (10.10)$$

$$MSE = \frac{SSE}{n-(K+1)} \qquad (10.11)$$

الأحصاء F هو نسبة متوسطى المربعين:

$$F = \frac{MSR}{MSE} \qquad (10.12)$$

وإذا كان الفرض العدمى وهو عدم وجود علاقة بين Y والمتغيرات التفسيرية X_1، X_2 صحيحاً، فإن توزيع المعاينة لهذه النسبة هو توزيع F بـ K درجة حرية البسط، n-(K+1) درجة حرية المقام. وكما سبق القول من قبل، فإنه كلما كبرت قيمة F صغرت قيمة P ويكون هناك دليل قوى ضد الفرض العدمى القائل بإنه ليس هناك إرتباط بين Y والمتغيرات المفسرة .

وللتوضيح ، نرجع إلى مثال الأيس كريم ، حيث أن (n = 10) حجم العينة، (k=2) حدين يشملا متغيرات مفسرة ومجموع البواقى SSE تساوى 1,206.158 ومجموع مربعات الإنحدار (SSR = 14,553.942) ومجموع المربعات الكلى تساوى 15,760.10 ويكون الفرضين العدمي والبديل:

$H_0 : \beta_1 = \beta_2 = 0$

H_a : على الأقل واحد من المعالم β_1 أو β_2 لا يساوى الصفر

ويعطى جدول تحليل التباين (١٠-١) قيمة F المحسوبة 42.232 وقيمة P تساوى 0001. وحيث أن قيمة P تساوى الصفر تقريباً. فإن دليل العينة يتعارض بقوة مع الفرض العدمى القائل بعدم وجود علاقة بين Y والمتغيرات المفسرة (X_2 , X_1) . لهذا يمكننا أن نصل لدرجة ثقة كبيرة أنه توجد علاقة بين المبيعات وواحد على الأقل من المتغيرات المفسرة، السعر X_1 ، ودرجة الحرارة X_2 .

جدول (١٠-١) : جدول تحليل التباين الكلي لمثال الأيس كريم

Source	df	SS	MS	F-value	P-value
Due to regression	2	14,553.942	7,276.971	4.0232	.0001
Due to error	7	1,206.158	172.308		
Total	9	15,760.100			

(١٠-٤-٢) تقييم المساهمة الفردية لمتغير تفسيري: الأحصاء T :

Evaluating the Contribution of an Individual Predictor Variable: The T Statistic

إفترض أن إجراء تحليل التباين للنموذج الكلي لم يكتشف بوضوح عن العلاقة بين Y وواحد على الأقل من المتغيرات المفسرة كما رأينا فى مثال الأيس كريم. تكون الخطوة المنطقية التالية هى تحديد أى المتغيرات المفسرة يظهر ليساهم فى تفسير الإختلاف فى قيم Y ، وأى منها ليس له هذا التأثير ويمكن تحقيق ذلك باجراء اختبارات فردية لكل معامل β الذى يتضمنه المتغير المفسر .

وفيما يتعلق بنموذج الإنحدار :($Y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + \varepsilon$)، يكون الفرض العدمى لإختبار المساهمة الفردية للمتغير المفسر X_i هو $(H_o: \beta_i = 0 \text{ For } i = 1, ..., k)$ ويقر هذا الفرض العدمى بانه على الرغم من تغير المتغير المفسر X_i، فإن قيمة متوسط Y تظل ثابتة طالما ظلت قيمة المتغيرات الأخرى المفسرة فى النموذج بدون تغيير. ويوجد تفسير آخر فى غاية الأهمية لهذا الفرض العدمى وهو أن إضافة المتغير X_i إلى النموذج الذى يحتوى بالفعل على متغيرات مفسرة أخرى، لا يحسن من عملية التنبؤ بقيم Y : وهذا معناه، إنه إذا كانت $(H_o: \beta_i = 0)$ صحيحة، فإن : X_i لاتمدنا بمعلومات مفيدة لتقدير قيم Y ، أى لاتمدنا بمعلومات أكثر من تلك المعلومات التى تم الحصول عليها بواسطة المتغيرات الأخرى المفسرة. وبالتالى فإن $(\beta_i = 0)$ عبارة عن المساهمة الهامشية للمتغير المفسر X_i فى التنبؤ بقيمة Y، فى وجود كل جميع المتغيرات المفسرة الأخرى للنموذج، هذه المساهمة تساوى صفر. وربما يكون الفرض البديل من طرفين (ذيلين) two-sided أو في جانب واحد (ذيل واحد) one-sided. ويعتمد ذلك على العلاقة الخاصة بين X_i , Y . ونلاحظ أن معظم برامج الحاسبات تفترض وجود الفرض البديل ذو الجانبين two-sided. ولهذا السبب سوف نتبنى ذلك الفرض في هذا الفصل.

لا تندهش إذا علمت أن الإحصاء T هو المؤشر الملائم لاستنتاج المساهمة الهامشية للمتغير المفسر فى حضور كل المتغيرات المفسرة الأخرى فى النموذج ومعرفة قيمة مقدر b_i للمتغيرات والخطأ المعيارى للمقدر b_i ، يمكن إختبار الفرض العدمى$(H_o: \beta_i = 0)$ بواسطة قيمة T المحسوبة .

$$T_i = \frac{b_i - 0}{SE(b_i)} \qquad \textbf{(10.13)}$$

أو تحديد نتيجة P أو إنشاء فترة ثقة للمقدار β_i . وكالعادة ، فإن قيمة P الصغيرة تتعارض مع الفرض العدمى وتقترح أن X_i يساهم مساهمة هامشية فى التنبؤ بقيمة Y . وعكس ذلك إذا كانت فترة الثقة للمقدار β_i لا تحتوى على الصفر. لا يكون الفرض العدمى على نحو يوهم بأنه مقبول وبالتالى يكون هناك دليل على أن X_i له مساهمة هامشية للتنبؤ بقيمة Y . وفى كلتا الحالتين فإن توزيع المعاينة للإحصاء T هو توزيع T بدرجة حرية $\{n - (K+1)\}$ (وهي نفسها درجة حرية SSE) لهذا فإن فترة الثقة $\{100(1-\alpha)\%\}$للمؤشر β_i تكون :

$$b_i \pm t_{1-\alpha/2,\, n-(k+1)} SE(b_i) \qquad \textbf{(10.14)}$$

حيث: $t_{1-\alpha/2,[n-(k+1)]}$ قيمة جدولية من توزيع T (انظر جدول C فى الملحق).

والتعبيرات التى تحدد قيمة $SE(b_i)$ أي الخطأ المعياري لتقديرات المربعات الصغرى وتشمل جبر المصفوفات والتى لن نقوم بذكرها هنا. وعموماً فإن مخرجات الحاسب الآلى تمدنا دائماً بالمقدرات والأخطاء المعيارية دون مجهود .

استخدام الحاسب الآلي: Using the Computer

كما رأينا في الفصل التاسع، فإن كل من Minitab، SAS (والعديد من البرامج الاحصائية الجاهزة الأخرى) تزودنا بالمعلومات المناسبة والمطلوبة في عمليات تحليل الانحدار. وسوف نستخدم بيانات مثال الأيس كريم السابق لتوضيح مخرجات الحاسب عند إستخدام Minitab وكذلك SAS (انظر جدول ١٠-٢) للبرنامج SAS فإننا إستخدمنا الأمر PROC GLM بدلاً من PROC REG لأن ذلك يزودنا بمعلومات مفيدة في إختيار المتغيرات المفسرة والتي يتضمنها النموذج، كما سوف يتم مناقشتها في الجزء التالي Next Section.

لاحظ أن مخرجات Minitab (بخلاف الجزء الأخير منها) عبارة عن إمتداد طبيعي من المخرجات الخاصة بالانحدار الخطي البسيط والتي تم مناقشتها في الفصل التاسع. وفي مخرجات SAS وجدنا أن مقدرات المربعات الصغرى للمعلمات β_0، β_1، β_2 تحت العمود المعنون بكلمة Estimate في الجزء الأخير من المخرجات. وكما لاحظنا فإن ($b_0 = 25,8777$)، ($b_1 = -1.3418$)، ($b_2 = 5,1953$): وبالتالي فإن تقدير المربعات الصغري هو ($\hat{Y} = 25.877 - 1.3418\ X_1 + 5.1953X_2$).

ومن هذا الجزء من مخرجات SAS فإننا نجد أيضاً الأخطاء المعيارية لمقدرات المربعات الصغرى في العمود الذي عنوانه Std Error of Estimate وبالتالي فإن ($SE(b_0)= 51.3260$)، ($SE(b_1)= .3620$)، ($SE(b_2)= .5839$) ويكون تفسير الجزء الأخير من مخرجات Minitab وكذلك الجزء الأوسط من مخرجات SAS (والذي يتبع كلا منهما لجدول تحليل التباين ANOVA) في الجزء التالي Next Section، ولتقييم كيفية توفيق معادلة المربعات الصغرى بقيم Y، فإنه تم عمل هذه الملاحظات بالاعتماد على مخرجات البرامج الاحصائية Minitab أو SAS.

(١) تحليل النموذج بشكل عام: بالنسبة للنموذج بشكل عام فإن الفرض العدمي ($H_o: \beta_1 = \beta_2 = 0$) يتعارض بشكل كبير من بيانات المثال حيث أن قيمة P-value والتي تناظر قيمة (F والتي قيمتها 42.23) عبارة عن قيمة صغيرة جداً (.0001). وبالتالي فإننا نعلم انه ربما يكون السعر أو درجة الحرارة أو كلاهما تساعد في شرح الاختلاف في قيم Y في العينة.

(٢) تحليل المساهمة الهامشية الفردية للمتغيرات المفسرة: والآن نستطيع تقييم المساهمات الفردية للسعر في وجود مساهمة درجة الحرارة، أو المساهمة الفردية لدرجة الحرارة في وجود مساهمة السعر. وسوف نلاحظ أن الفروض العدمية الفردية ($H_o: \beta_1 = 0$)، ($H_o: \beta_2 = 0$) تتعارض مع بيانات العينة لأن قيمة P-value المناظرة مقدار صغير للغاية. فمثلاً بالنسبة للسعر نجد أن قيمة T هي ($T_1 = -1.3418/.3620=-3.71$) حيث تكون قيمة P-value هى (.0076). أما بالنسبة لدرجة الحرارة فإن قيمة T هى ($T_2 = 5.1953/0=.5839=8.90$) وتكون قيمة P-value هى (.0001). وفي

جدول (١٠-٢)

مخرجات الانحدار بالبرامج Minitab -SAS لبيانات مثال الأيس كريم

General Linear Models Procedure

Dependent Variable: SALES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14553.94187442	7276.97093721	42.23	0.0001
Error	7	1206.15812558	172.30830365		
Corrected Total	9	15760.10000000			

R-Square	C.V.	Root MSE	SALES Mean
0.923468	3.191497	13.12662575	411.30000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PRICE	1	912.66666667	912.66666667	5.30	0.0549
TEMP	1	13641.27520776	13641.27520776	79.17	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PRICE	1	2367.16968325	2367.16968325	13.74	0.0076
TEMP	1	13641.27520776	13641.27520776	79.17	0.0001

تابع : جدول (١٠-٢)

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	25.87773161	0.50	0.6296	51.32600406
PRICE	-1.34175131	-3.71	0.0076	0.36200155
TEMP	5.19529086	8.90	0.0001	0.58389598

Minitab

```
The regression equation is
sales = 25.9 - 1.34 price + 5.20 temp

Predictor        Coef      Stdev     t-ratio      p
Constant        25.88      51.33        0.50    0.630
price         -1.3418     0.3620       -3.71    0.008
temp           5.1953     0.5839        8.90    0.000

s = 13.13   R-sq = 92.3%   R-sq(adj) = 90.2%

Analysis of Variance

SOURCE        DF        SS        MS        F        p
Regression     2   14553.9    7277.0    42.23    0.000
Error          7    1206.2     172.3
Total          9   15760.1

SOURCE        DF    SEQ SS
price          1     912.7
temp           1   13641.3
```

مخرجات SAS ظهرت قيمة T في العمود الذي له العنوان الذي (T for H_0 : Prmeter = 0) والصفوف (PRICE nd TEMP.). وتوجد قيم P المناظرة في العمود التالي إلى اليمين. ولذلك فإننا نخلص بأن السعر يمدنا بمعلومات متزايدة مفيدة غير تلك المعلومات التي تساهم بها درجة الحرارة. كما أن درجة الحرارة تمدنا بمعلومات متزايدة مفيدة غير تلك المعلومات التي يساهم السعر. وبالتالي فإن معادلة المربعات الصغرى والتي تحتوى كل من السعر ودرجة الحرارة تكون حقا أفضل في التنبؤ بقيمة Y من أى نموذج يحتوى على واحد فقط من تلك المتغيرات.

(٣) تحليل دقة معاملات الانحدار المقدرة: إن دقة المعلومات β_1 ، β_2 المقدرة تعتبر جيدة. حيث أن الأخطاء المعيارية لمقدرات المربعات الصغرى تعتبر أصغر من قيمة المقدرات. وفي الحقيقة فإنه بمقارنة حجم قيم T-values (T_1 = -3.71)، (T_2 = 8.90) فإننا نستطيع أن نخلص إلى أن β_2 يتم تقديرها بدقة أكثر نسبياً عن β_1. وبإستخدام المعادلة رقم (10.14) نستطيع تكوين أو إنشاء فترات ثقة لكل من β_1 ، β_2 ، فعلى سبيل المثال نجد أن فترة ثقة 95% للمعلمة β_1 : ($t_{.975,7}$ = 2.365) هي:

$$- 1.3418 \pm (2.365)\ (.3620) = - 1.3418 \pm .8561$$

أو هى (.4857 - ، 2.1979-). وهذا يعنى أنه، إذا ظلت الحرارة ثابتة، فإن متوسط المبيعات ربما ينخفض أو يتناقص بما يساوى حوالى 2.2 أوقية على الأكثر أو بما يساوى 49. أوقية على الأقل

لكل تخفيض بمقدار ١ سنت فى السعر. كما أن فترة الثقة 95% للمعلمة β_2 هي

$$5,1953 \pm (2.365)\ (.5839) = 5.1953 \pm 1.3809$$

أو هى (6.562 ، 3.8144). وهذا يعني أنه إذا ظل السعر ثابتا، فإن متوسط المبيعات ربما يرتفع أو يتزايد بما يساوى 6.58 أوقية تقريباً على الأكثر أو بما يساوى 3.81 أوقية على الأقل عند ارتفاع درجة الحرارة اليومية بدرجة واحدة.

(٤) تحليل معامل التحديد: نجد أن قيم كل من R^2 (.9235) وكذلك R^2 المعدل Adjusted R^2 (.9016) كبيراً بدرجة معقولة. وهذا معناه أن معظم الاختلافات في قيم العينة Y تم شرحه عن طريق معادلة المربعات الصغرى والتي تحتوى على السعر ودرجات الحرارة. وهذا ليس معناه أنه يمكننا تحسين هذا النموذج عن طريق إضافة متغيرات مفسرة جديدة.

ولعله من المعقول أن نكون متفائلين بأن معادلة المربعات الصغرى ($\hat{Y} = 25.8666 - 1.3418X_1 + 5.1953X_2$) تكون مناسبة للتقدير والتنبؤ في حدود مدى الاسعار ودرجات الحرارية التي توجد في بيانات العينة. ويتبقى فقط إجراء تحليل البواقي لإختبار صحة فروض النموذج.

(١٠-٤-٣) إختبارات إضافية عن المساهمات الفردية للمتغيرات المفسرة : مبدأ مجموع المربعات الإضافية : Extra Sum of Squares Principle

فى الجزء السابق، تعلمت كيف يستخدم الأحصاء T فى فحص المساهمة الهامشية لمتغير مفسر معين إذا كانت معادلة المربعات الصغرى تشمل كل المتغيرات المفسرة الأخرى. وفى هذا الجزء نقدم مفهوم مرن خاص يميز مجموع المربعات الإضافى للإسهام الهامشى للمتغير المفسر ويسمى «مبدأ مجموع المربعات الأضافي **Extra Sum of Squares Principle**». يساعدنا هذا الإجراء على فهم أفضل لفكرة الإسهام الهامشى ومشكلة الازدواج الخطى بين المتغيرات المستقلة **Collinearity** والتى سوف تناقش فى الجزء (١٠-٧). ولحسن الحظ فإن المعلومات الدائمة المتعلقة بمبدأ مجموع المربعات الإضافى توجد فى العديد من حزم الحاسب الآلى ومنها Minitab , SAS (والجزء من مخرجات SAS ، Minitab والتى لم تناقش فى مشكلة الأيس كريم تتعلق بهذه الجزئية).

ونلاحظ أنه فى بعض تطبيقات الإنحدار، لا يزودنا المتغير المفسر بمعلومات إضافية فى النموذج الذى يحتوى على متغيرات مفسرة أخرى، حتى إذا كان هذا المتغير المفسر مفيد فى التنبؤ بالمتغير التابع عندما يكون هو المتغير المفسر الوحيد فى النموذج. فى بعض التطبيقات الأخرى، قد يزودنا المتغير المفسر بمعلومات إضافية تفيد فى النموذج الذى يحتوى على متغيرات مفسرة أخرى، حتى إن لم يكن كذلك عندما يؤخذ فى الإعتبار وحده. معرفةمثل تلك المشكلة الدقيقة يمكن الحصول عليها وفصلها بإستخدام مبدأ مجموع المربعات الإضافى. ويعتمد هذا المبدأ الأساسى على طبيعة مجموع المربعات الكلية وطبيعة مجموع مربعات الإنحدار وطبيعة مجموع مربعات البواقى. والتى تستحق إعادة ذكرها هنا :

١- فى حالة معرفة قيم Y لعينة معينة فإن : مجموع المربعات الكلى SST لا تكون متأثرة عندما تضاف عناصر تشمل متغيرات مفسرة جديدة فى نموذج الإنحدار.

٢- مجموع مربعات الأخطاء SSE يقل ولو بمقدار قليل بزيادة عدد العناصر المضافة إلى النموذج .

٣- مجموع مربعات الإنحدار SSR تزيد على الأقل بمقدار قليل بزيادة العناصر المضافة للنموذج، والزيادة فى مجموع مربعات الإنحدار تتوافق مع إنخفاض مجموع مربعات الأخطاء .

مما تقدم ، فإن الأسلوب المنطقى فى الإنحدار الخطى المتعدد هو إضافة عناصر إلى النموذج فقط إذا كان إضافة تلك العناصر يقلل SSE بقدر كبير. وبالتالى يزيد SSR بقدر كبير .

وسوف نوضح هذا المبدأ بمثال الأيس كريم حيث أن :

$\{MSE(X_1, X_2)=172.308\}$ ، $\{SSR(X_1, X_2)=14,553.942\}$ ، $\{SSE(X_1, X_2)=1,206.158\}$ ، $(SST=15,760.1)$

ولمساعدة القارئ في إدراك المتغيرات المستقلة فى معادلة الإنحدار فقد تم تحديدها لكل من مجموع مربعات الإنحدار والخطأ ومتوسط مجموع مربعات الخطأ. لهذا فإننا نعرف SSR , SSR" (X_1, X_2) للنموذج الذي يحتوى على متغيرات مفسرة (X_2, X_1) ". ولتحديد مجموع المربعات الإضافية نشير إلى الجزء المتوسط من مخرجات SAS فى جدول (١٠-٢) ولقد تم تقديمه هنا للتأكد. ونلاحظ أن هناك عمودان جدد تم تعريفهما كما يلي: Type I SS وكذلك Type III SS. ومثلما يوجد عمودان خاصان بمجموع المربعات، فإن هناك طرق عديدة مفيدة لتطبيق مبدأ مجموع المربعات الإضافي.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PRICE	1	912.66666667	912.66666667	5.30	0.0549
TEMP	1	13641.27520776	13641.27520776	79.17	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PRICE	1	2367.16968325	2367.16968325	13.74	0.0076
TEMP	1	13641.27520776	13641.27520776	79.17	0.0001

- الأحصاء الهامشى F المكافئ للأحصاء T : النوع الثالث SS III

النوع الثالث III لمجموع المربعات يقيس التزايد فى SSR والذى يكون نتيجة إضافة متغير مفسر للنموذج الذى يتضمن كل المتغيرات المفسرة الأخرى. على سبيل المثال الأثر المتزايد للمتغير الإضافى X_3 إلى نموذج الإنحدار الذى يشمل بالفعل X_2, X_1 . نقارن مجموع مربعات الخطأ للنموذجين: واحد يتضمن الثلاث متغيرات المفسرة والآخر يشمل X_2, X_1 فقط ونفترض أن النتائج كما يلى :

Model	SST	SSE	SSR
$\hat{Y} = b_0 + b_1X_1 + b_2X_2$	100	40	60
$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$	100	25	75
Extra sum of squares		15	15

لاحظ أن إضافة X_3 قلل مجموع مربعات الخطأ من 40 إلى 25 . ويمثل هذا «مجموع المربعات الإضافى» الأثر المضاف لإضافة X_3 إلى النموذج ويعرف مجموع المربعات الإضافية هنا $SSR(X_1, X_2, X_3)$ والتى تعنى زيادة فى SSR نتيجة لإضافة X_3 إلى النموذج الذى يحتوى سابقاً على X_2, X_1 .

وبتطبيق هذا المفهوم على مثال الأيس كريم نجد أن العمود المحدد كنوع ثالث SS III فى مخرجات SAS تزودنا بالقيم التالية :

$\{SSR\ (X_2|\ X_1) = 13,64.275\}$, $\{SSR\ (X_1\ |\ X_2) = 2,367.17\}$

وهنا نجد أن $SSR\ (X_1\ |\ X_2) = 2,367.170$ تمثل الإنخفاض فى مجموع مربعات الخطأ الذى يمكن أن يعزى إلى السعر X_1 عندما يضاف إلى النموذج الذى يحتوى فعلاً على درجة الحرارة X_2. ويتم شرح مجموع المربعات الإضافى فى الجدول التالى :

Model	SST	SSE	SSR
$\hat{Y} = b_0 + b_2X_2$	15,760.1	3,573.328	12,186.772
$\hat{Y} = b_0 + b_2X_2 + b_1X_1$	15,760.1	1,206.158	14,553.942
Extra sum of squares		2,367.170	$2,367.170 \Rightarrow SSR(X_1\|X_2)$

بالمثل فإن $\{SSR = (X_2|\ X_1) = 13,641.275\}$ تمثل إنخفاض في مجموع مربعات الخطأ الذى يعزى إلى إضافة درجة الحرارة إلى النموذج الذى يحتوى بالفعل على السعر. ويكون مجموع المربعات الإضافى كما يلى :

Model	SST	SSE	SSR
$\hat{Y} = b_0 + b_1X_1$	15,760.1	14,847.433	912.667
$\hat{Y} = b_0 + b_1X_1 + b_2X_2$	15,760.1	1,206.158	14,553.942
Extra sum of squares		13,641.275	$13,641.275 \Rightarrow SSR(X_2\|X_1)$

وكلما زاد الإنخفاض فى مجموع مربعات الخطأ، كلما زادت فائدة إضافة متغير تفسيرى إلى النموذج الذى يحتوى بالفعل على متغيرات تفسيرية أخرى .

هل أنت مندهش، أن درجة الحرارة أكثر فائدة نسبياً فى التنبؤ بقيمة Y عن السعر؟، لايجب أن تندهش فنحن بالفعل توصلنا لمثل ذلك معتمدين على الأحصاء T . فربما يكون هناك بعض الإتصال بين النوع الثالث III لمجموع المربعات والأحصاء T ، وسنلاحظ كما فى حالة الأحصاء T أن المساهمة الإضافية لكل متغير مفسر فى وجود كل المتغيرات المفسرة الأخرى يمكن تحديدها بواسطة الأحصاء F . أولاً ، يجب تحديد متوسط المربعات المناظر لمجموع المربعات الإضافية للكميات $\{SSR = (X_1|\ X_2)\}$, $\{SSR\ (X_2|\ X_1)\}$ حيث أن كل مجموع مربعات يمثل أثر إضافة عنصر إلى النموذج وحيث توجد درجة حرية واحدة فقط مرتبطة به. لهذا فإن: $[MSR\ (X_1|\ X_2) = SSR\ (X_1|\ X_2)\ /1]$ ، $[MSR(X_2|\ X_1) = SSR\ (X_2|\ X_1)\ /1]$ ، بالتالى، فإن متوسط المربعات ومجموع المربعات يكونا متطابقان. ولإختبار الأثر الإضافى عند اضافة X_1 إلى النموذج الذى يحتوى بالفعل على المتغير X_2 فإننا نحدد نسبة $\{SSR\ (X_1|\ X_2)\}$ إلى $\{MSE\ (X_1, X_2)\}$. وبالمثل فإنه لإختبار الأثر الإضافى المتزايد لإضافة X_2 إلى النموذج الذى يحتوى على المتغير X_1 نحدد نسبة $\{SSR\ (X_2|\ X_1)\}$ إلى $\{MSE\ (X_2, X_1)\}$ هذه النسب هى إحصاءات F، وتسمى باحصاءات F الهامشية أو الجزئية لأختبار الفروض $H_o: \beta_1=0$ ، $H_o: \beta_2=0$.

بتطبيق ذلك على مثال الأيس كريم نجد أن قيمة F الجزئية للإختبار $(H_o: \beta_1=0)$ فى وجود X_2 هو:

$$F_1 = \frac{SSR(X_1\,|\,X_2)/1}{MSE(X_1,X_2)/1} = \frac{2,367.170}{172.308} = 13.74$$

بينما تكون قيمة F الجزئية لإختبار $(H_0: \beta_2=0)$ فى وجود X_1 :

$$F_2 = \frac{SSR(X_2|X_1)/1}{MSE(X_2,X_1)} = \frac{13,641.275}{172.308} = 79.17$$

هذه القيم للاحصاء F بالإضافة إلى قيم P $(P_r > F)$ توجد فى مخرجات البرنامج الإحصائى SAS المجاورة لعمود النوع الثالث لمجموع المربعات (Type III SS) لنفس الفروض. نعود إلى الإحصاء T حيث $(T_1 = -3.71)$, $(T_2 = 8.9)$. والآن ماذا تعتقد عندما نقوم بتربيع قيم T هذه؟ تحصل بالطبع على قيم F المناظرة، لماذا؟ لأنه كما ذكرنا فى الفصل التاسع فأن مربع المتغير العشوائي T بدرجات حرية V يكون مساوياً لقيمة المتغير العشوائى F بدرجة حرية واحدة فى البسط ودرجات حرية V فى المقام. لهذا فإن قيم P (.0001 , .0076) تكون هى نفسها كما تم الحصول عليها مع قيم T ، لأن الإحصاء F الهامشي من النوع III من مجموع المربعات، تكون مساوية للإحصاء T المناظر.

- الإحصاءات الهامشية F لآخر متغير مفسر تم إدخاله للنموذج : النوع SS I

يمكن الإستفادة من تطبيق مبدأ مجموع المربعات الإضافية بطرق أخرى. حيث أنه إذا إفترضنا أن لدينا أربعة متغيرات مفسرة فى نموذج إنحدار : X_1 , X_2 , X_3 , X_4 ويستخدم أسلوب إختبار أثر إضافة متغير واحد فى كل مرة، وهذا معناه أننا نستخدم أربعة نماذج : النموذج الأول سوف نعتبر أن به متغير مفسر واحد X_1 ثم نوجد النموذج للإنحدار الثاني مع المتغيرات X_1 , X_2 . ثم نوجد النموذج الثالث مع X_1 , X_2 , X_3 . ثم نوجد النموذج الرابع مع X_1 , X_2 , X_3 , X_4 . وفى كل حالة نلاحظ كيف أن إضافة متغير تفسيري يخفض فى مجموع المربعات للخطأ. ويسمى هذا التغير المتتالى فى مجموع مربعات الخطأ بمجموعات المربعات من النوع الأول [Type I] وتستخدم الملاحظات التالية :

$SSR(X_1)$ = مجموع مربعات الإنحدار عندما يكون X_1 المتغير المفسر الوحيد .

$SSR(X_2|X_1)$ = مجموع المربعات الإضافية عندما يضاف X_2 إلى النموذج مقارنة بمجموع مربعات الإنحدار عندما يحتوى النموذج على X1 فقط .

$SSR(X_3|X_1, X_2)$ = مجموع المربعات الإضافية عندما يضاف X_3 إلى النموذج مقارنة بمجموع مربعات الإنحدار للنموذج الذى يحتوى على X_1 , X_2 .

$SSR(X_4|X_1, X_2, X_3)$ = مجموع المربعات الاضافية عندما يضاف X_4 إلى النموذج مقارنة بمجموع مربعات الإنحدار للنموذج الذي يحتوى على X_1 , X_2, X_3.

على سبيل المثال، إفترض أن مجموع مربعات الخطأ والإنحدار كما يلى للأربعة نماذج :

Model	SST	SSE	SSR	Type I SS
$\hat{Y} = b_0 + b_1X_1$	100	50	50	50
$\hat{Y} = b_0 + b_1X_1 + b_2X_2$	100	40	60	10
$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$	100	25	75	15
$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$	100	5	95	20
Total				95

والنوع الأول (I) لمجموع المربعات هو عبارة عن زيادة متتالية فى مجموع مربعات الإنحدار كمتغيرات مضافة إلى النموذج واحداً بعد الآخر. لاحظ أن مجموع المربعات الكلى للنوع I فى الجدول يكون 95 هو نفسه مجموع مربعات الإنحدار للنموذج النهائي الذى يشمل X_1 , X_2 X_3 , X_4. وهذا صحيح بصفة عامة. مجموع المربعات من النوع I الذي يمثل تجزئة لمجموع مربعات الإنحدار المتضمن كل المتغيرات المفسرة لأى نموذج له أربعة متغيرات مفسرة.

$$SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) + SSR(X_4|X_1, X_2, X_3) \quad (10.15)$$

وكل مجموع مربعات من النوع I يكون له درجة حرية واحدة لأنها تمثل مساهمة عنصر واحد فقط يشمل متغير مفسر. ويساعدنا الحاسب الآلى فى تحديد مجموع المربعات من النوع I. وعموماً، ليس من الضرورى توفيق أربعة نماذج. وبدلاً من ذلك فإن البرنامج الإحصائى SAS يزودنا بكل مجموع المربعات من النوع I عندما نوفق نموذج كامل يحتوى على كل المتغيرات المفسرة وهذا موضح فى العمود (Type I SS).

فعلى سبيل المثال، إذا كان هناك متغيران مفسران فى مثال الأيس كريم، فمن الممكن تجزئة مجموع مربعات الإنحدار (عندما يكون كلا المتغيرين السعر X_1 ودرجة الحرارة X_2 فى النموذج) إلى مكونات كل واحد منهم عبارة عن النوع الأول من مجموع مربعات الإنحدار بدرجة حرية واحدة. فإذا إفترضنا أن ترتيب الدخول للمتغيرات المفسرة إلى النموذج تكون السعر أولاً ثم درجة الحرارة. فتجزئة مجموع مربعات الإنحدار إلى مكونات يعزى إلى ترتيب دخول كل متغير مفسر ويكون :

$$SSR(\text{Price}, \text{Temperature}) = SSR(\text{Price}) + SSR(\text{Temperature} \mid \text{Price}) \quad (10.16)$$

مجموع مربعات الإنحدار للنموذج الذى يحتوى على السعر ودرجة الحرارة المعطى فى جدول (١٠-٢) يساوى (14,553.942) وقيمة المكونات $SSR(X_1)$ ، $SSR(X_2|X_1)$ موجودة فى مخرجات البرنامج الإحصائى SAS فى عمود (Type I SS). ($SSR(X_1) = 912.667$) تمثل مجموع مربعات الإنحدار عندما يكون السعر فقط فى النموذج وتمثل ($SSR(X_2|X_1) = 13,641.775$) إنخفاض مجموع مربعات الخطأ الذى يحدث عندما يضاف X_2 إلى النموذج السابق المحتوى على X_1. للتأكيد على هذه التجزئة نلاحظ أن :

$$SSR(X_1, X_2) = 14,553.942 = 912,667 + 13,641.275)$$

نريد أن نكون قادرين على تحديد المساهمات الإضافية المتزايدة للمتغيرات المفسرة كلما أضيفت بالتتابع لمعادلة الإنحدار. ولأى من المتغيرات التفسيرية، هل نتوقع أن تكون مساهمته أكثر من إضافة متغيرات غير مرتبطة فى النموذج؟ لتنفيذ هذا التقييم نقوم بتكوين الإحصاء F لكل متغير مفسر بالاعتماد على قيمة مجموع مربعاتها من النوع I. ويكون بسط الإحصاء F مجموع المربعات للنوع I مقسوماً على درجة حريتها، والمقام يكون متوسط مربع الخطأ لكل النموذج $MSE(X_1|X_2)$. لهذا فإن قيمة F الناتجة عن مساهمة X_1 فقط هى :

$$F_1 = \frac{SSR(X_1)/1}{MSE(X_1, X_2)} = \frac{912.667}{172.308} = 5.30$$

وقيمة F للمساهمة الإضافية للمتغير X_2 عندما تضاف إلى النموذج المحتوى على X_1 يكون :

$$F_2 = \frac{SSR(X_2|X_1)/1}{MSE(X_1,X_2)} = \frac{13,641.275}{172.308} = 79.17$$

ونجد أن هذه القيم للاحصاء F وما يقابلها من قيم P موجودة فى الجانب الأيمن لعمود Type I SS. لاحظ أن قيمة P لمساهمة السعر صغيرة لكن لا يمكن إهمالها (.0549) . وهذا يدل على أنه إذا كان السعر هو المتغير المفسر الوحيد فى النموذج فإن مساهمته فى إختلاف Y غير مؤكد. ومن المهم المقارنة بين قيم P وما يقابلها من مجموع مربعات السعر للنوع III ، والفرق قليل ويقدر بمقدار (.0076). وتشير قيمة P السابقة أننا متأكدين تماماً بأن السعر فى وجود درجة الحرارة يساعد فى تفسير الإختلاف للمبيعات اليومية. ما سبب هذا الإختلاف أو الفرق؟ يرجع هذا الفرق إلى العلاقة بين المتغيرات المفسرة. وهذا يعنى أن السعر ودرجة الحرارة مرتبطين بشكل ما. عندما يكون هناك علاقة بين المتغيرات التفسيرية، فإنه يظهر تعارض معتمداً على قيم P للنوع I والنوع III لمجموع المربعات. وعندما يكون التعارض كبير (بمعنى قيم P مختلفة تماماً)، فإن ذلك يدل على وجود ما يعرف بالإرتباط الخطى بين المتغيرات المفسرة. والازدواج الخطى Collinearity بين المتغيرات مشكلة كبيرة فى تطبيقات الإنحدار المتعدد. وسوف نقوم بتعريفه وفحصه فى الجزء (١٠-٧).

والبرنامج الاحصائي Minitab وغيرها من الحزم الإحصائية تزودنا بمعلومات عن إحصاء F الجزئية - بالخصوص مخرجات Minitab - تتضمن ما يكافئ (Type I SS) لكن لا تزودنا بقيم F الجزئية أو قيم P ويمكنك أن تقوم بعمل ذلك بنفسك بإستخدام المعلومات المعطاة. وفى الجزء الأخير من مخرجات Minitab فى جدول (١٠-٢) لمثال الأيس كريم ، نلاحظ أنه تحت العمود SEQ SS تعطي مكونات $SSR(X_1, X_2)$. هذه الأرقام هي نفسها كما فى مخرجات SAS بعد تجنيب فروق التقريب : $(SSR(X_1) = 912.7)$ ، $(SSR(X_2|X_1) = 13,641.3)$ ويجب علينا أن نشير إلى أن ترتيب دخول المتغيرات التفسيرية إلى النموذج يتم تعريفه بواسطة المستخدم لأمر الإنحدار فى البرنامج الإحصائى. فعلى سبيل المثال فى مثال الأيس كريم إذا حددت درجة الحرارة أولاً X_2 ثم تبعها X_1 فإن المكونات $(SSR(X_1, X_2) = 14,553.942)$ تكون :

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2) \qquad \textbf{(10.17)}$$

وفي هذا السياق ، فإن $SSR(X_2)$ هو مجموع مربعات الإنحدار عندما يكون X_2 فقط فى النموذج (درجة الحرارة)، وتكون $SSR(X_1|X_2)$ هى مجموع المربعات الإضافية التي أزيلت من SSE عندما تضاف X_1 السعر إلى النموذج. وعند إستبدال ترتيب الدخول إلى النموذج من المتغيرات التفسيرية من الممكن التعرف على عدة علاقات لمجموع مربعات الإنحدار كما وضحت فى التعبيرات (10.16)، (10.17).

مثال (١٠-٢)

بالإشارة إلى مثال الأيس كريم ، وبالإعتماد على الإحصاء F الجزئية ، حدد ما إذا كانت درجة الحرارة تساعد فى شرح الإختلاف بين قيم Y عندما تكون درجة الحرارة فقط هي المتغير المفسر فى النموذج .

الحل

جدول(١٠-٢) يظهر جزء من مخرجات البرنامج الإحصائي SAS التى تشمل كلا من السعر ودرجة الحرارة. ونحن نعلم أن مجموع مربعات الإنحدار للنموذج بأكمله $SSR(X_1,X_2)$ تساوى 14,553.942 وأن متوسط مجموع مربعات الخطأ $MSE(X_1 , X_2)$ تساوى 172.308 ومجموع المربعات الإضافية عندما تضاف X_1 إلى النموذج السابق المتضمن فقط X_2 يكون $SSR(X_1| X_2)= 2,367$، (مجموع مربعات النوع الثالث) ونريد تحديد $SSR(X_2)$ ، مجموع المربعات الإضافى الذي يستبعد من SSE عندما تضاف درجة الحرارة X_2 للنموذج ($Y = \beta_0 + \varepsilon$)والتى لا تتضمن متغيرات مفسرة. من التعبير (10.17) نعلم أن :

$$SSR (X_1 , X_2) = SSR(X_2) + SSR(X_1 | X_2)$$

بحلها بالنسبة لقيمة $SSR(X_2)$ وإحلال الكميات المعلومة نجد أن :

$$SSR(X_2) = SSR (X_1 , X_2) - SSR (X_1 | X_2)$$

$$= 14,553.942 - 2,367.17 = 12,186.772$$

حيث أن متوسط مربعات الخطأ للنموذج بأكمله هى: $[(MSE(X_1 , X_2) = 172.308]$ ، تكون قيمة F كما يلي:

$$F = \frac{12,186.772 / 1}{177.308} = 70.73$$

وقيمة P المقابلة تكون (.0001) مساوية للصفر. بالتالى درجة الحرارة غالباً تساعد فى تفسير وشرح الإختلاف فى Y عندما تكون هى المتغير المفسر الوحيد فى النموذج .

مثال (١٠-٣)

إفترض أن X_3 , X_2 , X_1 ثلاث متغيرات مفسرة في نموذج للتنبؤ بالمتغير التابع Y . معتمداً على عينة مناسبة، إفترض بعض الحزم الإحصائية المستخدمة (مثل SAS) لتحديد معادلة المربعات الصغرى ($\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$) .

(أ) إذا كان ترتيب دخول المتغيرات X_1 ثم X_2 ثم X_3، حدد ملاحظة مناسبة للثلاث مكونات الخاصة بقيمة $SSR (X_1 , X_2 , X_3)$.

(ب) أجب نفس السؤال كما فى (أ) إذا كان ترتيب الدخول X_2 ثم X_1 ثم X_3 .

(ج) أجب نفس السؤال كما فى (أ) إذا كان ترتيب الدخول X_3 ثم X_2 ثم X_1 .

(د) بصرف النظر عن طريقة الدخول، حدد الملاحظة المناسبة لمجموع المربعات التى تكون مدرجة فى العمود (Type III SS) اذا تم إستخدام الأمر PROC GLM إذا كانت الحزمة الإحصائية SAS هى المستخدمة .

الحل

الإجابة الأجزاء الثلاث الأولى هى مجموع المربعات الهامشية المرتبطة بترتيب دخول المتغيرات وهى:

(أ) $SSR = (X_1 , X_2 , X_3) = SSR(X_1) + SSR(X_2| X_1) + SSR (X_3| X_1, X_2)$

حيث $SSR(X_1)$ تمثل الانخفاض فى مجموع مربعات الخطأ عندما X_1 تكون هي أول متغير يدخل النموذج.

$SSR(X_2|X_1)$ وتمثل الإنخفاض الاضافي فى مجموع مربعات الخطأ SSE عندما تضاف X_2 للنموذج المحتوى فعلاً على X_1 .

$SSR(X_3|X_1, X_2)$ وتمثل الإنخفاض الاضافي فى مجموع مربعات الخطأ عندما تضاف X_3 إلى النموذج المحتوى فعلاً على X_1 , X_2 .

(ب) $SSR(X_1, X_2, X_3) = SSR(X_2) + SSR(X_1|X_2) + SSR(X_3|X_2, X_1)$

(ج) $SSR = (X_1, X_2, X_3) = SSR(X_3) + SSR(X_2|X_3) + SSR(X_1|X_3, X_2)$

مكونات الأجزاء (ب) ، (ج) لها نفس المعنى كما في الجزء (أ).

(د) المكونات التى يجب إدراجها فى عمود (Type III SS) هى التى تكون مقياس للأثر المضاف لكل متغير مفسر إذا أدخل فى النموذج ويكون ذلك بعد أن يشتمل النموذج على كل المتغيرات التفسيرية الأخرى. وتكون المكونات :

$$SSR(X_1|X_2, X_3),\ SSR(X_2|X_1, X_3) \text{ and } SSR(X_3|X_1, X_2)$$

مثال (١٠-٤)

تم التعاقد مع شركة تسويق لتقدير مقدار إنفاق الأسرة على الغذاء بالإعتماد على دخل وحجم الأسرة، وتشمل البيانات التالية الإنفاق الشهرى على الطعام Y (بآلاف الدولارات) فى مقابل الدخل الشهرى X_1 (بآلاف الدولارات) حجم العائلة X_2 وذلك لعدد 15 عائلة مختارة عشوائياً من منطقة جغرافية معينة :

Y	X_1	X_2	Y	X_1	X_2
.43	2.1	3	.29	1	5
31	1.1	4	1.29	8.9	3
.32	.9	5	.35	2.4	2
.46	1.6	4	.35	1.2	4
1.25	6.2	4	.78	4.7	3
.44	2.3	3	.43	3.5	2
.52	1.8	6	47.	2.9	3
			.38	1.4	4

حدد ما إذا كان حجم ودخل العائلة يساهم فى شرح عمليات الإختلاف بين أرقام إنفاق الأسرة فى العينة الموضحة فى الجدول .

الحل :

لنموذج الإنحدار ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$) تم توضيح نتائج البرنامج SAS فى جدول (١٠-٣) ومنها يمكن إستنتاج ما يلى :

١- معادلة المربعات الصغرى :

$$(\hat{Y} = -.1605 + .1487X_1 + .0769X_2)$$

جدول (١٠-٣)

مخرجات SAS لمثال (١٠-٤)

General Linear Models Procedure

Dependent Variable: FOOD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.35954215	0.67977108	113.14	0.0001
Error	12	0.07209785	0.00600815		
Corrected Total	14	1.43164000			

R-Square	C.V.	Root MSE	FOOD Mean
0.949640	14.40749	0.07751228	0.53800000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
INCOME	1	1.27162442	1.27162442	211.65	0.0001
SIZE	1	0.08791773	0.08791773	14.63	0.0024

Source	DF	Type III SS	Mean Square	F Value	Pr > F
INCOME	1	1.33664408	1.33664408	222.47	0.0001
SIZE	1	0.08791773	0.08791773	14.63	0.0024

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-.1604580427	-1.78	0.1012	0.09038910
INCOME	0.1487270228	14.92	0.0001	0.00997132
SIZE	0.0769151943	3.83	0.0024	0.02010687

نلاحظ أن علامات التقديرات ($b_1 = .1487$) , ($b_2 = .0769$) موجبة، وينبغي أن تكون كذلك، العائلة الأكبر دخلاً هي الأكثر إنفاقاً في إستهلاك الطعام وذلك لحجم معين للعائلة، (العائلة الأكبر حجماً هي الأكثر إنفاقاً في إستهلاك الطعام وذلك حسب دخل معين) .

٢- لتقييم النموذج ككل فإن الفرض العدمي (H_0: $\beta_1 = \beta_2 = 0$) يتعارض مع ما يحدث للعينة. ويتضح ذلك بواسطة تحليل التباين (P-value = .0001) . ومعنى هذا أن دخل العائلة أو حجمها أو كلاهما يساهم في شرح الإختلاف في الإنفاق على الطعام .

٣- لتقييم المساهمات الفردية للمتغيرات المفسرة فإن الفرض العدمي (H_0: $\beta_1 = 0$) من السهل أن يتم تعارضه أي رفضه بالاعتماد على T أو الإحصاء F الهامشية وتكون (P-value = .0001) . لهذا فإن مساهمة دخل الأسرة مع وجود حجم الأسرة مفيداً بالفعل. من ناحية أخرى، فإن فرض (H_0: $\beta_2 = 0$) أيضاً يتعارض حيث أن (P-value = .0024) . لهذا فإن حجم العائلة X_2 يظهر ليساهم بصفة متزايدة للتنبؤ بقيمة Y أكثر من المساهمة الناشئة بواسطة دخل الأسرة X_1 .

(١٠-٤-٤) إستخدام نموذج المربعات الصغرى في التقدير والتنبؤ :

Using the least Squares Model for Estimation and Prediction

كما في حالة الإنحدار الخطي البسيط في الفصل التاسع، طالما أصبح لدينا رؤية واضحة عن العلاقة بين المتغير التابع ومجموعة المتغيرات المفسرة، فإننا نريد إستخدام معادلة المربعات الصغرى للتقدير والتنبؤ. وعلى وجه الخصوص نريد تقدير متوسط قيمة Y أو التنبؤ بالقيم الفردية للمتغير Y، بإستخدام قيم مجموعة المتغيرات المفسرة في معادلة المربعات الصغرى .

وللتوضيح ، سوف نعود لمثال (١٠-٤). إفترض أن شركة تسويق تريد تقدير المتوسط الشهرى المنفق على الطعام بواسطة ($X_1 = 1\{\$1000\}$) ، ($X_2 = 3$) أفراد العائلة، وبالتعويض فى معادلة المربعات الصغرى:

($\hat{Y} = -.1605 + .1487X_1 + 0.769X_2$) تحصل على ($\hat{Y} = -.1605 + 1487(1) + .0769(3) = .219$)

لهذا فإن متوسط الإنفاق المقدر يكون 219$. ويكون الإنفاق المتنبأ به للعائلة الواحدة عند تكون ($X_1 = 1$) , ($X_2 = 3$) أيضاً يساوى 219$. لكن في ظل مجموعة من قيم X ، فإن دقة التقدير لمتوسط قيم Y تكون أفضل من تقدير قيم Y الفردية المتنبأ بها. نعلم أن الدقة محددة بواسطة الخطأ المعيارى المرتبط بالتقدير أو التنبؤ. ومما يؤسف له فإن التعبيرات الرياضية للأخطاء المعيارية تشمل جبر المصفوفات وهي ليست موضوع هذا الكتاب. ومع ذلك، فإن كل الحزم الإحصائية ومنها SAS , Minitab مزودة بفترات ثقة لمتوسط قيم Y . وفترات تنبؤ بمفردات قيم Y المعطاة لمجموعة المتغيرات المفسرة فى معادلة المربعات الصغرى . وفيما يلى فترة ثقة وفترة تنبؤ (95%) للمثال (١٠-٤) بمعلومية قيم X_1 , X_2 ولقد تم تنفيذ الحسابات بإستخدام الحاسب الآلى :

X_1	X_2	$\hat{Y}$	Confidence Interval	Prediction Interval
1.0	3	.2190	(.1473,.2908)	(.0355,.4026)
3.0	3	.5165	(.4647,5682)	(.3398,.6931)
6.0	3	.9626	(.8848,1.0405)	(.7767,1.1486)
1.0	4	.2929	(.2392,.3526)	(.1177,.4741)
3.0	4	.5934	(.5467,.6401)	(.4181,.7687)
6.0	4	1.0396	(.9517,1.1274)	(.8492,1.2300)

دعنا نفسر المعنى لأحد هذه الأسر. للأسرة المكونة من 4 أفراد بدخل شهرى , 3000$ قدرنا متوسط الإنفاق على الطعام فى الشهر بمقدار (593.4$) وفترة ثقة 95% (.6401 , .5467) تعنى أن متوسط الإنفاق الشهرى يمكن أن يكون بحد أدنى 546.7$ وحد أعلى 640.1$. ولنفس الأسرة نتنبأ باستهلاك شهرى للطعام (593.40$) حيث أن {($\hat{Y} = .5934$)} . لكن فترة التنبؤ (.7687 , .4181) 95% تعنى أن الإنفاق الفعلى يكون الحد الأدنى له 418.1$ والحد الأعلى 5768.70$.

تمارين

(١٠-١٢) عند توفيق النموذج ($Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon$) لعينة مكونة من 15 مشاهدة. حددت الكميات التالية :

$SST = 200$, $SSR(X_1, X_2) = 140$, $SSR(X_1) = 90$ and $SSR(X_1|X_2) = 80$

أ - هل اكتشفت علاقة بين قيم X_2 , X_1 , Y ؟

(ملاحظة : اختبر الفرض العدمى $H_0 : \beta_1 = \beta_2 = 0$) .

ب- هل مساهمة X_1 يساعد فى شرح الإختلاف فى قيم Y عندما يكون X_1 المتغير المفسر الوحيد؟

ج- هل مساهمة X_2 الزائدة مفيدة فى شرح الإختلاف فى قيم Y فى وجود X_2 ؟

د - هل مساهمة X_2 الزائدة مفيدة فى وجود X_1 ؟

هـ - هل مساهمة X_2 مفيدة عندما يكون X_2 هو المتغير المفسر الوحيد فى النموذج ؟

و - بالإعتماد على إجابتك فى الأجزاء السابقة للتمرين، ما استنتاجك حول المتغيرات المفسرة X_2 , X_1 ؟

(١٠-١٣) بالإشارة إلى تمرين (١٠-١٢) حدد قيمة R_a^2 للنموذج المحتوى على X_2 , X_1 واشرح معنى ذلك ؟

(١٠-١٤)عند توفيق النموذج ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$) لعينة (n= 23) مشاهدة ، حصلنا على الكميات التالية :

$SST = 500 , SSE (X_1, X_2) = 200 , SSR (X_2) = 200$ and $SSR(X_2| X_1) =275$

أجب عن جميع الأجزاء الواردة فى تمرين (١٠-١٢) ؟

(١٠-١٥) عند توفيق النموذج ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$) لعينة من 14 مشاهدة. حددت الكميات التالية :

$SST = 800 , SSE (X_1, X_2, X_3) = 100 , SSR (X_1) = 200$ and $SSR(X_2| X_1) = 470$

أ - هل اكتشفت العلاقة بين قيم X_3 , X_2 , X_1 , Y ؟

ب- ما قيمة R^2 المعدلة للمعادلة المتضمنة كل المتغيرات المفسرة الثلاث؟ هل تشير هذه القيمة لجودة النموذج فى التنبؤ ؟ اشرح .

ج - هل مساهمة X_1 مفيد فى شرح الإختلاف فى قيم Y عندما يكون هو المتغير المفسر الوحيد فى النموذج ؟

د - هل المساهمة الإضافية لـ X_2 مفيد فى وجود X_1 ؟

هـ - هل المساهمة الإضافية لـ X_3 مفيد فى وجود X_2 , X_1 ؟

و - ما قيمة T من المساهمة الأضافية للمتغير X_3 فى وجود X_2 , X_1 ؟

(١٠-١٦) بالإشارة إلى تمرين (١٠-١٥) افترض أن $SSR (X_2) = 350$

أ - هل مساهمة X_2 مفيد عندما يكون X_2 المتغير المفسر الوحيد ؟

ب - هل مساهمة X_1 الإضافية مفيد فى وجود X_2 ؟

(١٠-١٧) قامت جامعة ما بدراسة لتحديد ما إذا كان هناك علاقة بين بداية المرتبات Y بالألف دولار ومتوسط نقط التخرج X_1 (أي تقدير التخرج بمتوسط النقاط) وعمر الخريجين X_2 لطلاب فى كلية التجارة وحصلت على البيانات التالية :

بداية المرتب	متوسط النقط	العمر
25.5	2.95	22
27	3.2	23
28.1	3.4	23
29.4	3.6	23
28.2	3.2	27
22	3.85	22
25	3.1	25
25.8	3.85	28
22.7	3.05	23
21.4	2.7	22
22.5	2.75	28
24.2	3.1	22
26	3.15	26
24.2	2.95	23
23.8	2.75	26

حدد ما إذا كان متوسط النقط والعمر يساهموا على التوالى فى شرح الإختلاف فى بداية المرتب للعينة.

(١٠-١٨) افترض معادلة المربعات الصغرى ($\hat{Y} = 50 + .04X_1 + 25X_2 + 8X_3$) الممثلة لعينة مكونة من 20 مشاهدة وتمثل Y سعر مكيف هواء كدالة فى ثلاثة متغيرات (X_1) ، (X_2) ، (X_3)، افترض أيضاً أن $(SE(b_1) = 005)$, $(SST = 1,200)$, $(SSE\ (X_1 , X_2 , X_3) = 200)$
$(SE\ (b_2) = 2.5)$, $(SE\ (b_3) = 6.5)$

أ - هل إشارات معاملات X_1 , X_2 , X_3 صحيحة ؟ اشرح .

ب- هل استطعت تحديد العلاقة بين السعر و X_1، X_2 ، X_3 ؟ دعم إجابتك .

ج - حدد فترة ثقة 95% للمعالم β_1 , β_2 , β_3 ، معتمداً على هذه الفترات، ماذا نستنتج عن المساهمة الإضافية لكل متغير مفسر فى وجود المتغيرات الأخرى؟ اشرح .

(١٠-١٩) بفرض أن معادلة المربعات الصغرى ($\hat{Y} = 2.8 + 6.5X_1 - 3.2X_2$) قد تم تحديدها من 15 قراءة سنوية للمناطق الخاصة لتخزين الكمبيوتر. حيث Y تمثل عدد الحاسبات الشخصية المباعة فى المنطقة كدالة في كل من تكلفة ترويج كل منطقة X_1 وعدد الحاسبات المخزنة فى كل منطقة X_2. افترض أيضاً أن :

$$SE(b_1) = .62 , SE(b_2) = .4 , SSR = (X_1, X_2) = 750 , SST= 800 ,$$

أ – هل إشارات معاملات X_1 , X_2 صحيحة ؟ اشرح .

ب– هل اكتشفت العلاقة بين المبيعات وX_1، X_2 ؟ وضح إجابتك .

ج – حدد فترات ثقة 95% للمعالم β_1 , β_2 . معتمداً على هذه الفترات، ما استنتاجك حول المساهمة الإضافية لكل متغير مفسر فى وجود المتغير الآخر ؟ اشرح .

(١٠–٥) ادخال المعلومات الوصفية فى الإنحدار الخطى المتعدد : المتغيرات الوهمية

Incorporating Qualitative Information in Multiple Linear Regression:Dummy Variables

في هذا الفصل تم دراسة وتحليل نماذج الإنحدار الخطى المتعدد. والآن نستطيع تناول بعض الطرق كامتداد مفيد لنماذج الإنحدار. ففى هذا الجزء، نعرض طريقة للاستفادة من المعلومات الوصفية فى نموذج الإنحدار .

فى جميع المشاكل التى تم مناقشتها حتى الآن، كانت المتغيرات المفسرة كالسعر ودرجة الحرارة والدخل وحجم العائلة متغيرات كمية تأخذ قيماً. وعادة توجد بعض المتغيرات الوصفية كالحالة الإجتماعية والجنس وكلها عوامل هامة نحتاج إلى إدخالها فى نموذج الإنحدار، ولكن المتغيرات الوصفية ليس لها مقياس معروف جيداً. على سبيل المثال الحالة الإجتماعية تعزى إلى شخص متزوج أو غير ذلك. أيضاً التخصيص الدقيق لطالب التجارة، أما محاسبة أو إدارة الأعمال أو مالية. .الخ. كيف إذن تظهر المعلومات الوصفية فى معادلة الإنحدار؟ تستخدم المتغيرات الوهمية كمتغير إصطناعى Dummy Variable والذى يخصص دائماً قيم للواحد أو الصفر. القيمة (1) تشير إلى وجود الصفة، (0) يشير إلى غياب الصفة .

وللتوضيح ، سوف نعود إلى مثال الأيس كريم ، والغرض من تحليل الإنحدار هو تحسين وزيادة البصيرة خلال العمليات . بالتالى الوصول إلى أفضل القرارات التى يمكن إتخاذها. لهذا نتسائل ما هي العوامل الأخرى غير السعر ودرجة الحرارة التي يجب حسابها والتي تؤثر على الإختلاف فى مبيعات الأيس كريم اليومية. أحد الاحتمالات أن يكون الطلب عليها أكبر فى الأجازات عنها فى أيام الأسبوع الأخرى. إذا كان هذا صحيحاً فإننا نختار تخفيض السعر فى أيام الأسبوع، ونرفع السعر في عطلات نهاية الاسبوع. نوع اليوم (عطلة / يوم عادى) يعتبر معلومة وصفية يمكن دمجها فى نموذج الإنحدار بإنشاء متغير وهمى . ويمكن تعريف المتغير X_3 كما يلى :

$$X_3 = \begin{cases} 1 & \text{أيام العطلات} \\ 0 & \text{أيام العمل} \end{cases}$$

وبيانات العينة لمثال الأيس كريم مع توضيح نوع اليوم كما يلى :

اليوم Day	Y(sales) المبيعات	X_1 (Price) السعر	X_2(temperature) درجة الحرارة	X_3 (type of day نوع اليوم)
1	374	35	74	1
2	386	35	82	0
3	472	35	94	1
4	429	50	93	0
5	391	50	82	1
6	475	50	96	1
7	428	50	91	1
8	412	65	93	0
9	405	65	88	1
10	341	65	78	0

ويكون نموذج الإنحدار : $(Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon)$

لاحظ أنه إذا كان نوع اليوم غير هام ، إذن $(\beta_3 = 0)$ ويكون الفرض العدمي الإضافي المراد اختباره هو : $(H_0: \beta_3=0)$ في وجود السعر ودرجة الحرارة. من مخرجات البرنامج الإحصائي SAS يظهر النموذج في جدول (١٠-٤). والإستنتاجات التالية تظهر من المخرجات كما يلي :

جدول (١٠-٤)

مخرجات SAS لمثال الأيس كريم شاملاً نوع اليوم كمتغير وهمي

General Linear Models Procedure

Dependent Variable: SALES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15456.98030414	5152.32676805	101.99	0.0001
Error	6	303.11969586	50.51994931		
Corrected Total	9	15760.10000000			

R-Square	C.V.	Root MSE	SALES Mean
0.980767	1.728115	7.10773869	411.30000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PRICE	1	912.66666667	912.66666667	18.07	0.0054
TEMP	1	13641.27520776	13641.27520776	270.02	0.0001
DAY	1	903.03842972	903.03842972	17.87	0.0055

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PRICE	1	1470.50695790	1470.50695790	29.11	0.0017
TEMP	1	12660.27582653	12660.27582653	250.60	0.0001
DAY	1	903.03842972	903.03842972	17.87	0.0055

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	15.30937021	0.55	0.6030	27.90393324
PRICE	-1.10118194	-5.40	0.0017	0.20410657
TEMP	5.03906542	15.83	0.0001	0.31831702
DAY	20.24521411	4.23	0.0055	4.78851344

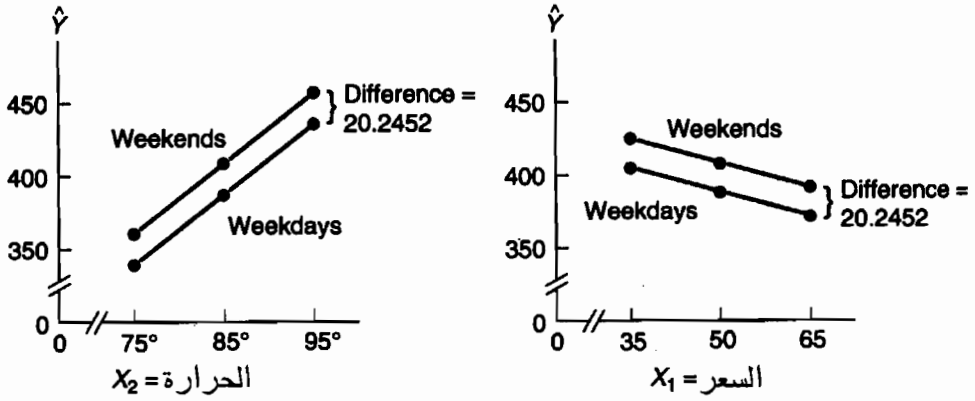
١- معادلة المربعات الصغرى هي :

$$\{\ \hat{Y} = 15.3094 - 1.1012X_1 + 5.0391X_2 + 20.2452X_3\}$$

ويمثل تقدير المربعات الصغرى $b_3 = 20.2452$ الفرق بين متوسط المبيعات اليومي عندما $(X_3=1)$ (يوم عطلة) ومتوسط المبيعات عندما $(X_3 = 0)$ (يوم عمل)، إذا كان السعر ودرجة الحرارة ثابتين . ويمكن أن ترى ذلك بواسطة التعبير في معادلة المربعات الصغرى في كلا الحالتين كما يلي :

$$\text{If } X_3 = 1 :\ \hat{Y} = 15.3094 - 1.1012X_1 + 5.0391X_2 + 20.2452$$
$$\text{If } X_3 = 0 :\ \hat{Y} = 15.3094 - 1.1012X_1 + 5.0391X_2 + 0$$

لاحظ أن المعادلتين متطابقتان فيما عدا أن الثابت $(b_3 = 20.2452)$ يضاف في حالة تقدير المبيعات في أيام العطلات . وطبيعة كلا المعادلتين موضحتان في شكل (١٠-٢) لقيمة $\hat{Y}$ في مقابل X_2 على فرض أن $(X_1 = 50)$ وكذلك $\hat{Y}$ مقابل X_1 على فرض أن $(X_2 = 85)$.



شكل (١٠-٢)

(أ) $\hat{Y}$ مقابل X_2 عندما $(X_1 = 50)$

(ب) تمثيل $\hat{Y}$ مقابل X_1 عندما $(X_2 = 85)$ لكل من أيام العطلات وأيام العمل

٢- يتعارض الفرض العدمى ($H_0: \beta_3=0$) بقوة مع بيانات العينة، حيث أن قيمة P والتي تعتمد أيضاً على قيمة الاحصاء T أو F الجزئية صغير جداً (.0055) . وهذا يعنى أن معرفة نوع اليوم يساعد فى التنبؤ بالمبيعات إذا كان السعر ودرجة الحرارة معروفين ويقدر متوسط مبيعات في الإجازات بمقدار 20.2452 فى اليوم أكثر من مبيعات يوم العمل العادي.

٣-نموذج المربعات الصغرى الذي يتضمن نوع اليوم يعتبر أفضل فى التقدير والتنبؤ عن النموذج الذي لا يتضمن نوع اليوم. وهذا يعتبر صحيحاً لأن تباين البواقى S_e^2 يكون أصغر في هذه الحالة. (50.52 فى مقابل 172.308) . بالإضافة إلى ذلك، فإن الأخطاء المعيارية لتقديرات المربعات الصغرى لمعاملات السعر ودرجة الحرارة تكون أصغر من قبل. حيث أصبحت قيمة T الآن: (-5.4)، (15.836) الآن فى مقابل قيمة T (-3.7)، (8.96) بالنسبة لكل من للسعر ودرجة الحرارة.

تخصيص قيم الصفر (0) والواحد الصحيح (1) للتعبير عن المتغيرات الوهمية :

لادخال المتغير الوصفى «نوع اليوم» فى مثال الأيس كريم إلى معادلة الانحدار، إفترض أننا خصصنا القيم 0 , 1 للمتغير X_3 بطريقة عكسية كما يلى :

$$X_3 = \begin{cases} 0 & \text{أيام العطلات} \\ 1 & \text{أيام العمل} \end{cases}$$

هل يصنع ذلك إختلافاً؟ الإجابة لا . فمعاملات الإنحدار ستكون هي نفسها ما عدا b_3 ستتغير إشاراتها فقط من الموجب إلى السالب. والتفسير أن متوسط المبيعات فى أيام العمل يكون (20.2452) أقل من أيام العطلات. بالطبع هذا هو نفس التفسير السابق. لهذا فتخصيص قيم 1 , 0 للمتغير الوهمى يكون إختيارياً. ولكن نحتاج فقط أن نتذكر أن معامل المتغير الوهمى فى معادلة الإنحدار دائماً يمثل الفرق بين حالة (1) ، حالة (0) .

ماذا إذا كان هناك أكثر من حالتين للمتغير الوصفى ؟

ماذا يحدث لو أن المتغير الوصفي له أكثر من حالتين؟ عامل آخر يؤثر على مبيعات الأيس كريم هو درجة صفاء السماء أي هل الجو مشمس أم به غيوم أو ممطر. لهذا يجب أن نأخذ فى الإعتبار

متغير وصفى آخر (حالة السماء) فى مثال الأيس كريم بثلاث حالات ممكنة وهي مشمس، مغيم، ممطر. الثلاث حالات للسماء يمكن تضمينها في النموذج بتعريف متغيرين وهميين على النحو التالي:

$$X_4 = \begin{cases} 1 & \text{جو مشمس} \\ 0 & \text{غير ذلك} \end{cases}$$

$$X_5 = \begin{cases} 1 & \text{به غيوم} \\ 0 & \text{غير ذلك} \end{cases}$$

هذان المتغيران الوهميان يمكن أن يحددا الثلاث حالات للسماء كما يلي :

	X_4	X_5
يوم مشمس	1	0
يوم به غيوم	0	1
يوم ممطر	0	0

وبصفة عامة ، فإن عدد المتغيرات الوهمية المطلوبة تكون أقل من عدد الحالات الممكنة للمتغير الوصفى بواحد . لاحظ أن واحد من هذه الحالات يكون الحالة الأساسية وهى الحالة التى تشير إلى الحالة التى تكون فيها كل المتغيرات الوهمية تساوى صفر. ولهذا فإن الحالة الأساسية فى حالة الأيس كريم هى الممطرة .

إذا تم إدخال المتغير الوصفى (حالة السماء) إلى معادلة الانحدار فى مثال الأيس كريم، فإننا يمكن أن نفسر معاملات المربعات الصغرى للمتغيرات X_5 , X_4 بواسطة فحص معادلة المربعات الصغرى لكل حالة كما يلى :

إذا كان $(X_4 = 1)$, $(X_5 = 0)$ (مشمس) : $\hat{Y} = b_0+b_1X_1+b_2X_2+b_3X_3+b_4$

إذا كان $(X_4 = 0)$, $(X_5 = 1)$ (معتم) : $\hat{Y} = b_0+b_1X_1+b_2X_2+b_3X_3+b_5$

إذا كان $(X_4 = 0)$, $(X_5 = 0)$ (ممطر) : $\hat{Y} = b_0+b_1X_1+b_2X_2+b_3X_3$

وكنتيجة لذلك فان التقدير (b_4) يمثل الفرق فى متوسط المبيعات بين اليوم المشمس $(X_4 = 1)$ والحالة الأساسية وهى اليوم الممطر. وتقدر (b_5) بأنها الفرق فى متوسط المبيعات بين اليوم الذى به غيوم $(X_5 = 1)$ واليوم الممطر. فى النهاية التقدير $(b_4 - b_5)$ يمثل الفرق فى متوسط المبيعات بين اليوم المشمس واليوم الذى به غيوم . وعلى العموم، فإن معامل أي متغير وهمى يقدر الفرق بين متوسط المتغير التابع (Y) عندما يساوى المتغير الوهمى (1) والمتوسط التابع للحالة الأساسية .

مثال (١٠-٥)

يرغب أحد خبراء إدارة الأفراد في تحديد العوامل التى تفسر المرتبات الأولية التى يحصل عليها خريجى المدارس التجارية، ويعتقد أن تقدير الخريج (GPA) وكذلك نوع التخصص أو كلاهما يؤثران على ذلك. وفيما يلى عينة مكونة من 15 خريج من المدارس التجارية حيث يمثل (Y) بداية مربوط المرتب (بآلاف الدولارات) ، (X_1) تعبر عن التقدير الخاص بالخريج (GPA) .

Y	X_1	التخصص Major	Y	X_1	التخصص Major
21.5	2.95	Management إدارة	24.7	3.05	Accounting محاسبة
23.0	3.20	Management إدارة	23.4	2.70	Accounting محاسبة
24.1	3.40	Management إدارة	20.5	2.75	Finance تمويل
25.4	3.60	Management إدارة	22.2	3.10	Finance تمويل
24.2	3.20	Management إدارة	24.0	3.15	Finance تمويل
24.0	2.85	Accounting محاسبة	22.2	2.95	Finance تمويل
27.0	3.10	Accounting محاسبة	21.8	2.75	Finance تمويل
27.8	2.85	Accounting محاسبة			

والمطلوب : إيجاد النموذج الملائم لهذه البيانات ، ثم تقييمه وتفسيره .

الحل :

حيث أن المتغير الوصفى (التخصص) يحتوى على ثلاث حالات، فإننا نحتاج إلى تعريف متغيرين وهميين كما يلى:

$$X_2 = \begin{cases} 1 & \text{إدارة أعمال} \\ 0 & \text{خلاف ذلك} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{محاسبة} \\ 0 & \text{خلاف ذلك} \end{cases}$$

لهذا فإن المتغيرين الوهميين X_3, X_2 يوضحا الثلاث تخصصات، حيث يكون تخصص التمويل هو الحالة الأساسية :

التخصص	X_2	X_3
إدارة أعمال	1	0
محاسبة	0	1
تمويل	0	0

تم توضيح مخرجات البرنامج الإحصائى Minitab فى جدول (١٠-٥) للنموذج ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$) من هذه المخرجات يمكن إستنتاج الآتى :

جدول (١٠-٥)

مخرجات ميني تاب لمثال(١٠-٥)

```
The regression equation is
salary = 6.00 + 5.49 gpa - 0.312 dummya + 3.40 dummyb

Predictor        Coef        Stdev       t-ratio        p
constant        5.997        4.942         1.21    0.250
gpa             5.491        1.672         3.28    0.007
dummya        -0.3120       0.9212        -0.34    0.741
dummyb         3.4047       0.7395         4.60    0.000

s = 1.167   R-sq = 73.2%   R-sq(adj) = 65.9%
```

تابع : جدول (١٠-٥) مخرجات ميني تاب لمثال(١٠-٥)

```
Analysis of Variance

SOURCE        DF        SS        MS        F        p
Regression     3    40.974    13.658    10.04    0.002
Error         11    14.970     1.361
Total         14    55.944

SOURCE        DF    SEQ SS
gpa            1     5.934
dummya         1     6.193
dummyb         1    28.847
```

١-معادلة المربعات الصغرى :

$$Y = 5.997 + 5.491X_1 - .312X_2 + 3.405X_3$$

والتقدير ($b_2 = -.312$) يعنى أنه فى المتوسط، يحصل تخصص إدارة الأعمال على أقل من المتخصص في التمويل . بالمثل تقدير ($b_3 = 3.405$) يعنى أنه فى المتوسط فإن المحاسب يحصل على ($3,405) أكثر من المتخصص فى التمويل .

٢-R^2 المعدلة (65.9%) وإختبار الفرض العدمي (H_o: $\beta_1 = \beta_2 = \beta_4 = 0$) وكانت قيمة P للاحصاء ($F = 10.04$) يساوى (.002) لهذا فإن GPA (و/أو) كل من المتغيرين الوهميين يظهر لهما على الأقل نفس الأثر على بداية المرتب.

٣-الفرض (H_o: $\beta_1 = 0$) ، (H_o: $\beta_3 = 0$) للمساهمة الفردية التي تعتمد على الأحصاء T تتعارضان مع بيانات العينة حيث أن قيمة P تساوى .007 , .000 على التوالى. وهذا يعنى أن التمييز بين المحاسب ($X_3 = 1$) والتمويل ($X_3 = 0$) مفيداً . لكن الفرض (H_o: $\beta_2 = 0$) لا يتعارض مع بيانات العينة حيث أن قيمة P تساوى .741. ويظهر بالتالى أن التمييز بين إدارة الأعمال ($X_2 = 1$)، والتخصص التمويل ($X_2 = 0$) لا يكون مفيداً. بالتالى يمكن إسقاط X_2 من النموذج ونعاود توفيقه فقط مع المتغير X_1 (GPA) والمتغير الوهمى X_3، حيث ($X_3 = 1$) تشير إلى تخصص المحاسبة، ($X_3 = 0$) تشير إلى إدارة الأعمال أو التمويل.

ويوضح جدول (١٠-٦) مخرجات النموذج ($Y = \beta_0 + \beta_1X_1 + \beta_3X_3 + \varepsilon$) بواسطة البرنامج الإحصائي Minitab . من هذه المخرجات نلاحظ أن المقدرة التنبؤية لمعادلة المربعات الصغرى ($\hat{Y} = 6.894 + 5.152X_1 + 3.495X_3$) أفضل من قبل. وتزداد قيمة R المعدلة من 65.9% إلى 68.5% كما أن قيمة تباين البواقى S_e^2 تقل من 1.361 إلى 1.206 ، الأخطاء المعيارية لقيم b_1 , b_2 تكون أصغر من ذى قبل .

جدول (١٠-٦)

مخرجات مينى تاب المحسنة لمثال (١٠-٥)

```
The regression equation is
salary = 6.89 + 5.15 gpa + 3.49 dummyc

Predictor        Coef      Stdev    t-ratio       p
Constant        6.894      4.016       1.72   0.112
gpa             5.152      1.288       4.00   0.002
dummyc         3.4946     0.6643       5.26   0.000

s = 1.123   R-sq = 73.0%   R-sq(adj) = 68.5%
```

تابع : جدول (١٠-٦)

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	40.818	20.409	16.19	0.000
Error	12	15.126	1.260		
Total	14	55.944			

SOURCE	DF	SEQ SS
gpa	1	5.934
dummyc	1	34.884

تمارين

(١٠-٢٠) يريد مثمن عقارات أن ينشأ نموذج إنحدار ويريد المثمن أن يتضمن النموذج المتغيرات المفسرة الآتية : الأماكن المراد تدفئتها (بالقدم المربع) عدد الحمامات، الشكل المعمارى (معاصر - على طراز المستعمرات البريطانية - تقليدى) وما إذا كان فى المنزل مكان تدفئة أم لا ، المتغير التابع يكون القيمة المثمنة. أنشئ النموذج الذى يحاول المثمن العقارى أن يوفقه لتمثيل عينة من البيانات .

(١٠-٢١) شركة (R. L. Williams) تؤجر الحاسبات الصغيرة وأيضاً تحافظ على خدمتها خلال مدة التعاقد. يريد مدير الخدمات تكوين نموذج للعلاقة بين سنوات الخدمة الممثلة فى الخبرة ومتوسط الوقت (فى الساعة) الذى تستغرقه للإصلاح . تتطلب بعض الحاسبات الصغيرة التقليدية وقت أكثر من غيرها. تستخدم الشركة ثلاث نماذج A , B , C معتمدة على عينة حددت معادلة المربعات الصغرى التالية لمتوسط وقت الإصلاح .

$$\hat{Y} = 20 - .2X_1 + 1.1X_2 - .5X_3$$

حيث X_1 عدد سنوات الخبرة فى الخدمة، X_2 , X_3 متغيرات وهمية معرفة كالتالى :

$$X_2 = \begin{cases} 1 & \text{متوسط وقت التصليح للنموذج A} \\ 0 & \text{غير ذلك} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{متوسط وقت التصليح للنموذج B} \\ 0 & \text{غير ذلك} \end{cases}$$

أ - فسر قيم المعاملات $(b_2 = 1.1)$, $(b_3 = -.05)$

ب- عبر عن معادلة المربعات الصغرى لثلاث معادلات كل واحدة تمثل العلاقة بين متوسط وقت التصليح لكل نموذج من نماذج الحاسب الصغير. فسر الإختلاف بين المعادلات الثلاثة .

ج - فسر معامل المربعات الصغرى $(b_1 = -.02)$.

(١٠-٢٢) اعتبر معادلة المربعات الصغرى $(\hat{Y} = 60 + 2X_1 - 5X_2 + 22X_3 - 3.5X_4)$ حيث: X_4 عبارة عن متغير كمى أما X_1 , X_2 , X_3 فهى عبارة عن متغيرات وهمية تمثل متغير وصفي بالشروط التالية:

$X_1 = \begin{cases} 1 & \text{إذا تحقق الشرط (1)} \\ 0 & \text{غير ذلك} \end{cases}$ $X_2 = \begin{cases} 1 & \text{إذا تحقق الشرط (2)} \\ 0 & \text{غير ذلك} \end{cases}$ $X_3 = \begin{cases} 1 & \text{إذا تحقق الشرط (3)} \\ 0 & \text{غير ذلك} \end{cases}$

فإذا كانت (X_4=10) إستخدام النموذج في التنبؤ بالمتغير Y عندما :

أ – المتغير الوصفى تحقق الشرط 1 .

ب– المتغير الوصفى تحقق الشرط 2 .

ج – المتغير الوصفى تحقق الشرط 3 .

د – المتغير الوصفى تحقق الشرط 4 .

هـ – هل الإختلافات التى حصلت عليها فى الأجزاء من (أ) حتى (د) ترجع فقط إلى المعلومات التي نحصل عليها من معاملات المربعات الصغرى ($b_1 = 2$)، ($b_2 = -5$)، ($b_3 = 22$) دون التعويض في معادلة المربعات الصغري؟ إشرح ذلك.

(١٠–٢٣) يرغب مدير احدى شركات التأمين، في تحديد العلاقة بين Y (عدد أيام العمل المفقودة بواطة ضحايا الحوادث)، (X_1 = العمر) ، (X_2 = النوع). فإذا تم إختيار 25 تقرير عن خسائر، وتم استخدام نموذج الانحدار التالى:

$$\hat{Y} = 21.4 - .075X_1 - 1.5X_2$$

ولهذه المعادلة وجد أن ($SE(b_2) = .99$, $SE(b_1) = .11$, $SSE = 3.81$, $SST = 4.75$)

أ – هل اكتشفت العلاقة بين المتغير التابع Y والمتغيرات المفسرة X_1 , X_2 ؟ وضح إجابتك .

ب– هل المساهمة الهامشية للعمر مفيدة فى وجود النوع ؟

ج – هل المساهمة الهامشية للنوع مفيدة فى وجود العمر ؟

د – ما استنتاجاتك للأجزاء أ ، ب ، حول إمكانية جعل قسط التأمين يعتمد على الدخل بدلا من الاعتماد على العمر ونوع المؤمن عليه عندما يفقد وقت العمل بسبب الحادث.

(١٠–٢٤) مدير مكتب نظم معلومات يهتم بكفاءة الحاسب الآلى أثناء وقت العمل اليومى أو ضغط العمل من (10 صباحاً إلى 3 مساءً) فى مقابل الأوقات خفيفة العمل(من 8-10 صباحاً ، 3-5 مساءً) وتقاس الكفاءة بواسطة إجمالى الوقت الذي ينتهى الحاسب الآلى من أداء وظيفته. جمعت البيانات من عينة مكونة من 36 وظيفة: 18 أثناء فترات العمل التي بها ضغط، 18 أثناء الفترات الخفيفة. وحيث أن حجم الوظيفة معقد يمكن توقع أثرها على الوقت الكلى. الوقت الفعلى للحاسب الآلى (الوقت الفعلى هو الوقت الذى يستغرقه الحاسب فى إنهاء الوظيفة مطروحاً منه أوقات انتظار وظائف الأخرى) بالتالى يمكن أن يضم النموذج المتغيرات التالية:

Y : الوقت الكلي

$X_1 = \begin{cases} 1 & \text{فترة ضغط} \\ 0 & \text{فترة خفيفة} \end{cases}$

X_2 = وقت وحدة التشغيل للحاسب

وكانت النتائج كالاتي:

$$\hat{Y} = 14.31 + 4.5X_1 + .33X_2$$

وكانت: $(R^2 = .77\ ,\ SST = 3225\ ,\ SS(b_1) = .09\ ,\ SE(b_2) = .11)$

أ – حدد SSR , SSE

ب– هل يمكن تحديد العلاقة بين X_2 , X_1 , Y كمجموعة ؟ وضح إجابتك .

ج – هل المساهمات الهامشية للمتغيرات X_2 , X_1 مفيدة ؟ اشرح استنتاجك حول كل متغير .

د – هل إشارات b_2 , b_1 متسقة مع العلاقات المتوقعة بين الوقت: (والوقت المضغوط/ الخفيف)، ووقت تشغيل الحاسب ؟

هـ – ما هى معادلة المربعات الصغرى لفترة العمل المضغوط ؟ لفترة العمل الخفيف ؟

(١٠–٢٥) شركة خدمات تتخذ نموذج الإنحدار التالى في تقييم المواقع الخاصة بتقديم الخدمة وتعتمد فى دراستها على 46 محطة خدمة . (محطات خدمة بترولية)

$$\hat{Y} = -17.4 + 20X_1 + .0033X_2 + 3.1X_3 + 2.2X_4$$

حيث :

Y = متوسط مبيعات البنزين فى اليوم (بالألف دولار) .

X_1 = سعر الجالون .

X_2 = حجم المرور (متوسط عدد السيارات التى تمر على الموقع فى اليوم) .

X_3 = طريقة الوصول: (1) إذا كان هناك إتجاهان ، (0) إذا كان هناك اتجاه واحد.

X_4 = brand image (1 "brand", 0 "off")

فإذا حصلت على المعلومات التالية لهذا النموذج

$$SST = 180,\ SSE = 35,\ SE(b_1) = 14,\ SE(B_2) = .0005,\ SE(b_3) = .60 \text{ and } SE(b_4) = 1.8$$

أ – حدد R_a^2 معتمداً على هذه القيمة هل يمكنك القول أن معادلة التنبؤ جيدة ؟ اشرح .

ب– كمجموعة ، هل المتغيرات المفسرة مرتبطة كلياً مع متوسط المبيعات اليومية؟ دعم إجابتك.

ج – فيما يتعلق بالمساهمات الهامشية لكل متغير مفسر، أى واحد منهم يظهر أكثر أهمية ؟ الأقل أهمية .

د – معتمداً على إجابة الجزء ج، أى المتغيرات يمكنك حذفه ولماذا ؟

هـ – ما هى معادلة المربعات الصغرى لمحطة الخدمة التى تبيع (name brand) بنزين والتي تقع على طريق ذو إتجاهين؟

(١٠–٢٦) فى مدينة ما، تم اختيار 5 منازل مباعة حديثا عشوائيا من كل منطقة من ثلاث مناطق مختلفة (A، B، C) في تلك المدينة. ولقد تم مقارنة سعر البيع Y بالقيمة المقدرة X_1 والمحددة من مكتب مثمن عقارى. وكانت بيانات العينة كما يلى (حيث أن سعر البيع وقيمة الأصول مقدرة بآلاف الدولارات):

المنطقة	X_1	Y
A	53.1	62.5
A	62	56.8
A	67.8	62.6
A	73.4	61.2
A	79.6	68.6
B	83.9	95.2
B	88.4	103.4
B	92.3	103.3
B	97.8	136.8
B	100.8	134.3
C	116.5	142.8
C	121.8	145.6
C	126.2	152.5
C	132.6	147.4
C	140.5	167.8

اختار نموذج مناسب لهذه البيانات، وقيم النموذج المختار.

(١٠-٦) المنحنى الخطى لنموذج الإنحدار : Curvilinear Regression Models

فى العديد من التطبيقات ، قد لا تكون العلاقة خطية بين المتغير التابع Y وواحد أو أكثر من المتغيرات المفسرة . ولتوضيح ذلك نرجع إلى مثال (٩-٢) (المرتب مقابل الخبرة) . وبالنظر إلى شكل (٩-٧)، (٩-١٧) نلاحظ أن كلاهما يحتوى على عنصر مربع (X^2) فى نموذج الإنحدار. ويمكننا دمج العنصر التربيعى ضمن المتغيرات المفسرة فى نموذج الإنحدار كما يلي :

$$(Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon)$$

حيث : $\beta_2 X^2$ العنصر المربع. ويظل بذلك نموذج خطى متعدد حيث أن الخطية تتعلق بالمعالم β_0 , β_1 , β_2. سوف نعالج X (سنوات الخبرة) ، X^2 (سنوات الخبرة المربعة) كمتغيرين تفسيريين. لاحظ أنه إذا كان العنصر التربيعى مفيداً ، فإن β_2 لا تساوى الصفر . ومن نتائج برنامج SAS لهذا المثال المستخدم والموضح بجدول (١٠-٧) يمكن إستنتاج الآتى من هذه المخرجات :

جدول (١٠-٧)
مخرجات SAS المرتب مقابل الخبرة

General Linear Models Procedure

Dependent Variable: SALARY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2766.59844714	1383.29922357	555.00	0.0001
Error	13	32.40155286	2.49242714		
Corrected Total	15	2799.00000000			

R-Square	C.V.	Root MSE	SALARY Mean
0.988424	3.272005	1.57874227	48.25000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
YEARS	1	2446.67975316	2446.67975316	981.65	0.0001
YEARSQRD	1	319.91869398	319.91869398	128.36	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
YEARS	1	821.06024971	821.06024971	329.42	0.0001
YEARSQRD	1	319.91869398	319.91869398	128.36	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	19.98007592	18.54	0.0001	1.07779724
YEARS	3.21518871	18.15	0.0001	0.17714553
YEARSQRD	-0.06339113	-11.33	0.0001	0.00559526

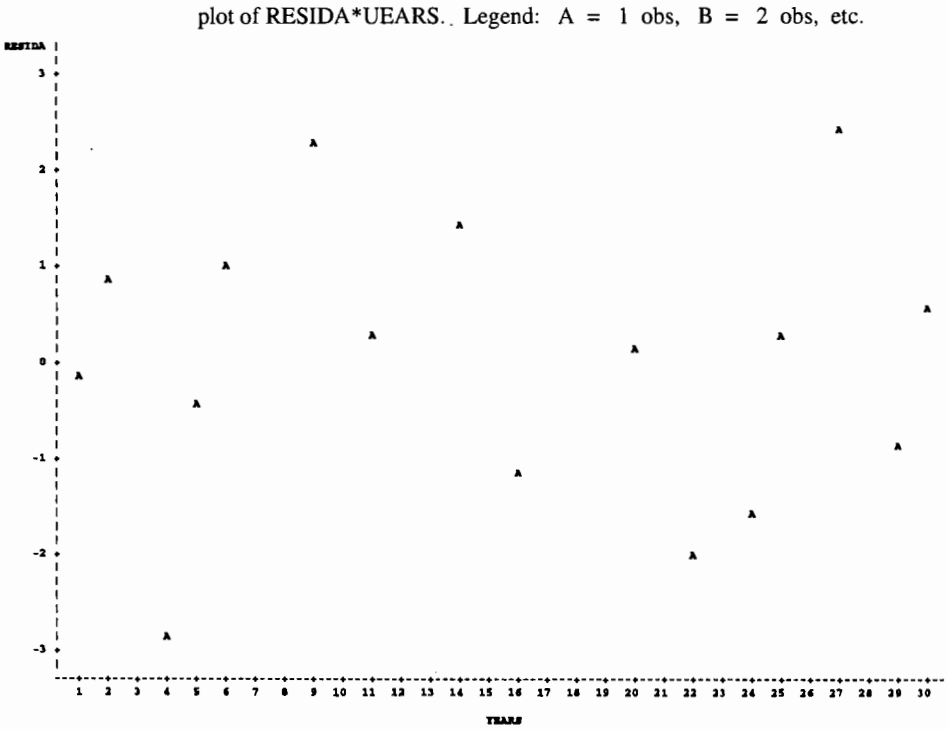
١- معادلة المربعات الصغرى هى :

$$(\hat{Y} = 19.9801 + 3.2152X - 0634X^2)$$

لاحظ أن إشارة معامل الحد التربيعى سالبة (b_2 = - .0634). وهذا يعنى أن منحنى المربعات الصغرى يتسطح كلما زادت قيمة X . وهذا يتسق مع شكل الإنتشار (٩-٧) .

٢-إن وجود الحد التربيعى ضمن معادلة الانحدار مفيداً، حيث أن الفرض (H_0: β_2=0) يتناقض مع بيانات العينة (وتكون قيمة P = .0001 سواء المصاحبة للاحصاء T أو F الجزئية).

٣- تكون معادلة المربعات الصغرى المتضمنة الحد التربيعى أفضل فى التقدير والتنبؤ عن النموذج الخطى. ويكون تباين البواقى S_e^2 أقل من قيمتها من قبل. وتتزايد قيمة R^2 من 874. إلى 988. . وتقدير المعالم β_1 لعنصر $\beta_1 X$ يكون أدق من ذى قبل (وقيمة T تكون الآن 18.5 فى مقابل قيمة T من قبل وهو 9.86) بالإضافة إلى ذلك، فإن الشكل البيانى للبواقى للتوفيق التربيعى مقابل عدد سنوات الخبرة والموضح فى شكل (١٠-٣)يظهر نمط لايمكن تحديده. ومما لا شك فيه فإن معادلة المربعات الصغرى التربيعية تكون أفضل فى التقدير والتنبؤ عن معادلة المربعات الصغرى الخطية وذلك داخل مدى بيانات العينة المعطاة.



شكل (١٠-٣)

شكل البواقى للتوفيق التربيعى للدخل مقابل عدد سنوات الخبرة

تمارين

(١٠-٢٧) افترض أن معادلة المربعات الصغرى ($\hat{Y} = 110 + 4.5X - .25X^2$) ممثلة للعلاقة بين المبيعات (Y) والسعر (X) افترض أن :

$SE(b_2) = .06$, $SE(b_1) = .5$, $n = 15$, $SSR(X, X^2) = 300$, $SST = 400$.

أ - إذا كان المدى المستخدم لقيم X فى تحديد هذه المعادلة من 5 إلى 20، بين واشرح المشكلة إذا أردت التنبؤ بالمبيعات عندما (X = 40) .

ب- حدد قيم R^2 , R_a^2 لهذه المعادلة .

ج - هل تنصح بضم العنصر المربع إلى النموذج ؟ وضح اجابتك .

(١٠-٢٨) استخدم البيانات التالية لتحديد معادلة المربعات الصغرى :

X	10	15	22	28	40	45	55	60
Y	20.1	23	24.7	27.3	39.7	31.4	33	34.4

أ - مثل البيانات بيانياً معتمداً على شكل الإنتشار، ما هوالنموذج الملائم لتمثيل العلاقة بين X، Y ؟

ب- وفق النموذج طبقاً لإجابتك فى (أ) وقيم نتائج معادلة المربعات الصغرى؟

(١٠-٢٩) اذا اردنا تحديد النموذج الملائم لتمثيل العلاقة بين المبيعات الشهرية (Y) فى إقليم ما، وعدد المبيعات (X) المخصصة فى الإقليم. وفيما يلى رقم المبيعات لعدد 12 إقليم:

عدد الوحدات المخصصة (X)	2	2	2	3	3	4	4	5	6	7	9	9
المبيعات (Y)	41	52	48	56	66	77	72	85	80	90	92	85

أ – مثل البيانات بيانيا. ما هو النموذج الملائم لتمثيل العلاقة بين X , Y .

ب– وفق النموذج طبقاً لإجابتك فى (أ) وقيم نتائج معادلة المربعات الصغرى؟

(١٠–٣٠) إذا علمت أن الطلب على سلعة ما يتغير بسبب التغير فى سعر الوحدة، البيانات التالية تحتوى على (Y) الطلب على المنتج كدالة فى (X) مدى السعر العادل .

بالدولار X	8.8	9.7	9.9	10.3	11	12.5	13.2	14.8	15.8	17.4	18.2
بالوحدة Y	360	305	230	242	180	172	121	83	122	91	105

أجب على الجزئين (أ)، (ب) المذكورين في التمرين (١٠–٢٩) باستخدام بيانات هذا التمرين.

(١٠–٧) اكتشاف النقص في النموذج وتجنب العوائق : تحليل البواقى والأزدواج الخطى:

Detecting Model Deficiencies and Avoiding Pitfalls: Residual Analysis and Collinearity

T، F يعتبرا من الأحصاءات الهامة في اكتشاف المساهمة المفيدة للمتغيرات الفردية، ولكنهما غير كافيان لتوضيح ما إذا كان نموذج المربعات الصغرى دقيق ومضمون. وليس هناك تحليل إنحدار كامل ما لم يتم فحص الإنتقاضات الممكنة فى النموذج أو المشاكل التى توجد به، حتى لو كانت النتائج التى تعتمد على المؤشران F , T الهامشية والتى قد تظهر جودة النموذج. دائما توجد مشكلتان واضحتان وهما التناقضات المحتملة عن فروض النموذج والأزدواج الخطى Callinearity والذى قد يحدث عندما يكون هناك إرتباط قوى بين بعض المتغيرات المفسرة. وسوف نتعرض لهاتان المشكلتان فى هذا الجزء .

(١٠–٧–١) تحليل البواقى : The Analysis of Residuals

إن أفضل طريقة لفحص الانحرافات عن فروض النموذج، يكون عن طريق تحليل البواقى. حيث أن تحليل البواقى يحدد المشاكل غير العادية فى بيانات العينة ويقترح طرق لتحسين النموذج وعلى الخصوص سوف نفحص ثلاث مشاكل شائعة وهي : (١) العلاقة بين المتغير التابع وواحد أو أكثر من المتغيرات المفسرة والتى قد لا تكون خطية. (٢) قد لا يكون تباين الخطأ σ_ε^2 ثابت . (٣) قد لا يشمل النموذج واحد أو أكثر من المتغيرات الهامة . أيضاً سيتم النظر إلى مشكلة القيم المتطرفة outliers أى للمفردات التى تكون مشاهداتها مختلفة وبعيدة عن بيانات العينة .

ويعنى تحليل البواقى تحليل الرسم البيانى للبواقى. فإذا كانت معادلة المربعات الصغرى جيدة ولا يوجد بها تناقضات فإن الشكل البيانى للبواقى فى مقابل كل متغير مفسر أو قيم $\hat{Y}$ يجب أن لا يظهر عنه نموذج معين. وبعبارة أخرى يجب أن لا توجد علاقة بين البواقى والمتغيرات المفسرة أو بين البواقى وقيم Y المقدرة. لكن إذا كان هناك شكل أو نمط يمكن توضيحه، فإنه يمكن أن يكون نقص أو عيب فى معادلة المربعات الصغرى .

ولإكتشاف المشاكل أو العيوب الثلاث الشائعة نجرى الآتى :

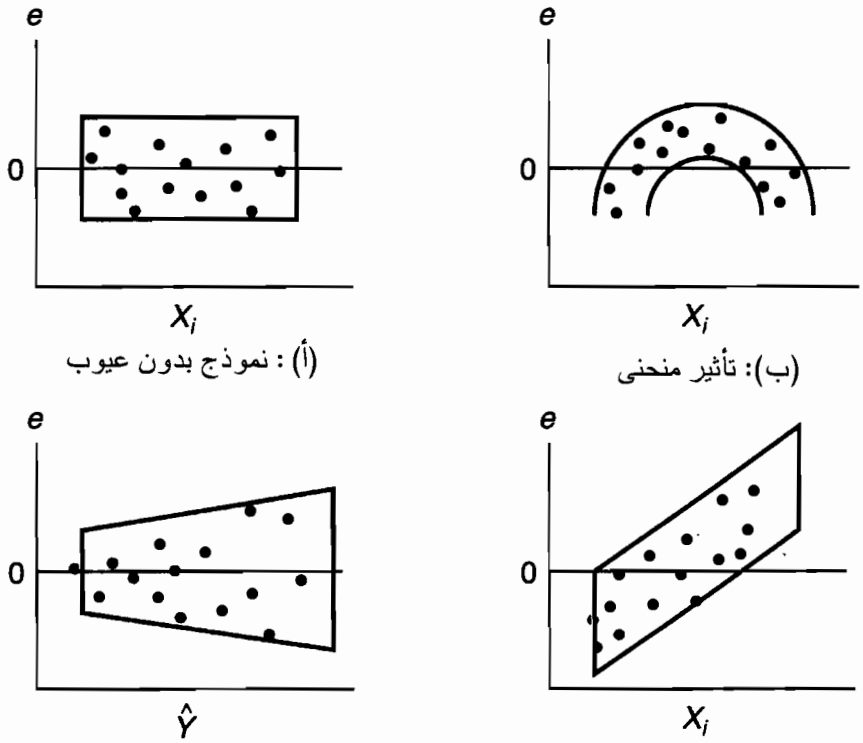
١– إكتشاف الإنحناء (التقوس) لتحديد ما إذا كانت العلاقة لها إنحناء يتعلق ببعض المتغيرات المفسرة. ويتم رسم شكل بيانى للبواقى فى مقابل كل متغير مفسر فى معادلة المربعات الصغرى .

٢-إكتشاف عدم ثبات تباين الخطأ. لتحديد هل تباين الخطأ ثابت، فإننا نقوم برسم الشكل البيانى بين البواقى وقيم$\hat{Y}$.

٣-إكتشاف عدم وجود متغير تفسيري هام. لتحديد ما إذا كان من الممكن ادخال متغير تفسيري هام في نموذج الانحدار (لم يكن موجوداً من قبل)، نقوم برسم البواقي مقابل قيم ذلك المتغير.

إذا كانت معادلة المربعات الصغرى خالية من أي عيوب، فإن البواقي تميل إلى أن تقع في شكل حزمة أفقية تتمركز حول الصفر، مع عدم وجود ميل لأن تكون موجبة أو سالبة بأنتظام. عموماً فان أي إختلاف عن هذا السلوك قد يفسر بوجود بعض النقائص أو العيوب في النموذج.

شكل(١٠-٤) يصور أشكال البواقى عندما تكون : (أ) معادلة المربعات الصغرى لا تحتوى على أى تناقضات. (ب) الأثر التربيعى للمتغير المفسر والذي يجب تضمينه فى النموذج. (ج) تباين الخطأ لا يكون ثابت (في مثل هذه الحالة، يمكن استخدام طرقة المربعات الصغري المرجحة كعلاج لهذه المشكلة). (د) حذف متغير مفسر يظهر إرتباط خطى قوى مع البواقى ويجب أن يضم فى نموذج الإنحدار .



شكل (١٠-٤)
أشكال البواقى

رصد البواقى تساعدنا أيضاً فى إكتشاف المشاهدات المتطرفة والتى تعرف باسم outliers . البواقي المرتبطة بهذه المشاهدات المتطرفة outliers عادة ما تكون كبيرة جداً في الحجم مقارنة بالبواقى الأخرى. ويمكن أن تخلق القيم المتطرفة مشكلة لأن لها أثر غير متكافئ على قيم معاملات المربعات الصغرى المقدرة فى النموذج. وعندما نحدد مشاهدة بأنها outlier ، ويجب إستبعادها من بيانات العينة المستخدمة لتحسين معادلة الإنحدار المقدرة. إذا كانت كل بيانات عينة ممثلة بشكل صحيح، فإن

إزالة أى مشاهدة من العينة تولد أثر صغير على معادلة المربعات الصغرى. ومع ذلك فنحن نحذر، من أنه حتى اذا كانت المشاهدة شبه قيمة متطرفة outliers ، فلايوجد سبب لاستبعادها إلا إذا وجد دليل قوى يدل على أن المشاهدة لاتمت بصلة لموضوع الدراسة. أمثلة للقيم المتطرفة والتي تشمل أخطاء التسجيل والبيانات لحوادث غير العادية مثل الأضرابات، الكوارث الطبيعية، الأعطال غير المتوقعة للآلات.

مثال (١٠-٦)

تريد شركة صناعية أن تتنبأ بتكلفة الوحدة المصنعة Y لواحدة من منتجاتها كدالة فى معدل الإنتاج المتغير X_1 والمواد الخام وتكلفة العمالة X_2 ولقد جمعت البيانات لمدة 20 شهر. إستخدم هذه البيانات فى تحديد أفضل معادلة مربعات صغرى للتنبؤ بتكلفة الوحدة .

Y	X_1	X_2	Y	X_1	X_2
13.59	87	80	15.93	102	116
15.71	78	95	16.45	82	117
15.97	81	106	19.02	74	127
20.21	65	115	18.16	85	133
24.64	51	128	18.57	86	135
21.25	62	128	17.01	90	136
18.94	70	115	18.03	93	140
14.85	91	92	19.22	81	142
15.18	94	93	21.12	72	148
16.30	100	111	23.32	60	150

الحـــل :

فى البداية نفترض النموذج الخطى :

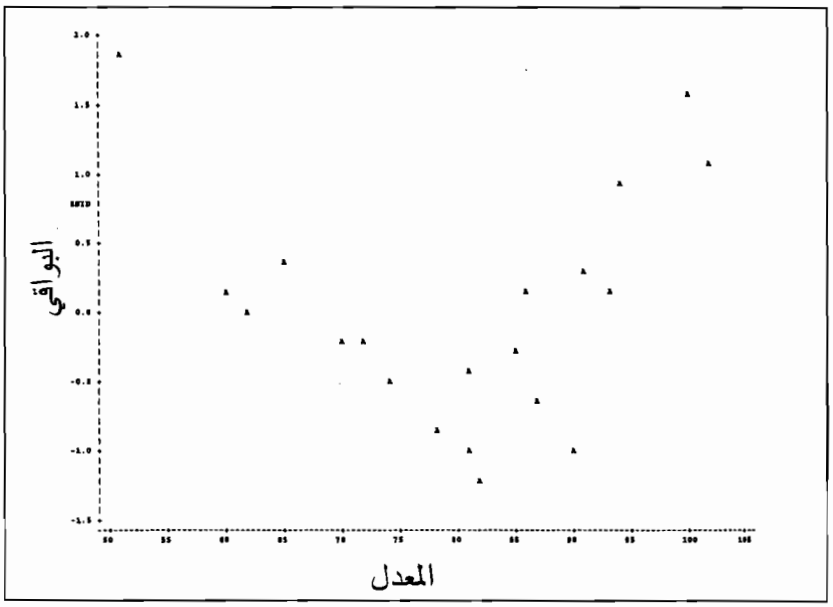
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

نتائج أو مخرجات برنامج SAS لهذا النموذج معطى فى جدول (١٠-٨). من هذه المخرجات نلاحظ أن (b_1 = -.1377) . (b_2 = .0742) أى قيمة سالبة لمعامل X_1 وقيمة موجبة لمعامل X_2. الخطأ المعيارى لهذه التقديرات صغير قياساً بقيم المعاملات نفسها. لهذا فإن، β_1, β_2 قدرت بدقة معقولة وقيمة R^2 تساوى 914. وهي مرتفعة نسبياً، والأثر المضاف لكل متغير مفسر فى حضور المتغيرات الأخرى واضح تماماً، حيث أن قيم P لإختبار T , F الهامشية تكون صغيرة جداً (كلاهما 001.) وتكون النتيجة اننا نحصل على معادلة تنبؤ جيدة، هل أنت موافق؟ لكن انتظر ماذا عن شكل البواقى؟ وفى شكل (١٠-٥) يظهر شكل البواقى. والجزء (أ) يبين البواقى مقابل X_1 (المعدل)، والجزء (ب) يظهر البواقى فى مقابل X_2 (العمالة). وعلى الرغم من عدم ظهور نمط أو نموذج متعلق بقيمة X_2، فإن نموذج تربيعي محدد يمكن اكتشافه مع المتغير X_1 (المعدل). وبهذا يفضل أن ندخل عنصر تربيعى للمتغير X_1 فى معادلة الإنحدار.

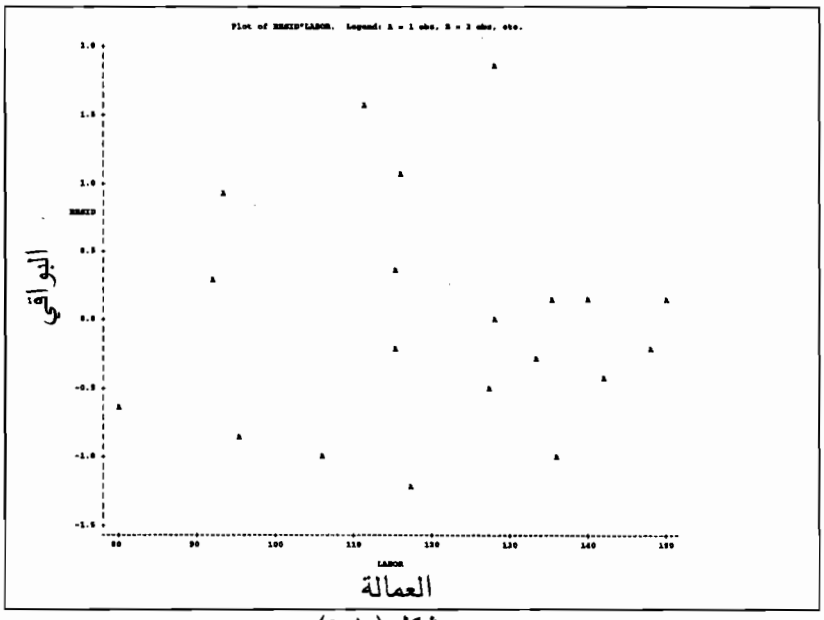
وإذا أردنا الآن محاولة تحسين النموذج عن طريق اضافة العنصر التربيعي للمقدار X_1 ، بالتالي يكون النموذج.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \varepsilon$$

ونحصل على نتائج البرنامج الإحصائى SAS المعطاة فى جدول (١٠-٩). بناء على هذه النتائج الجديدة هل تستطيع أن ترى أن ادخال العنصر التربيعى فى النموذج قد ساعد على تحسين التوفيق في معادلة المربعات الصغرى؟ الفرض العدمى (H_0: β_3=0) يكون من السهل إنكاره حيث أن (P-value = .0001) ، وتكون المعالم β_1 , β_2 المقدرة أفضل من ذى قبل من حيث الدقة، لأن قيم T تكون أكبر وتباين S_e^2 البواقي يكون أقل من ذي قبل. (1911. لهذا النموذج في مقابل 7996. للنموذج السابق) وقيمة R^2 تزداد بالتالى من 914. إلى 981. ، وفى النهاية يتم تمثيل البواقى الجديدة فى شكل (١٠-٦) والذى لا يظهر نموذج أو شكل واضح لتلك البواقى .



شكل (١٠-٥)
(أ) البواقى مقابل المعدل



شكل (١٠-٥)
(ب) البواقى مقابل العمالة

جدول (١٠-٨)
مخرجات SAS لمثال (١٠-٦)

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	12226767.88076240	2037794.64679375	16.16	0.0001
Error	17	2144200.74423751	126129.45554338		
Corrected Total	23	14370968.62500000			

R-Square	C.V.	Root MSE	Y Mean
0.850796	16.41824	355.14709001	2163.12500000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	3180503.76444430	3180503.76444430	25.22	0.0001
X2	1	1355656.00475436	1355656.00475436	10.75	0.0044
X3	1	727758.15719231	727758.15719231	5.77	0.0280
X4	1	4814280.29319091	4814280.29319091	38.17	0.0001
X5	1	2126706.26423271	2126706.26423271	16.86	0.0007
X6	1	21863.39694791	21863.39694791	0.17	0.6824

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	1247739.29877925	1247739.29877925	9.89	0.0059
X2	1	1476207.38278585	1476207.38278585	11.70	0.0033
X3	1	566435.02679629	566435.02679629	4.49	0.0491
X4	1	506795.34905793	506795.34905793	4.02	0.0612
X5	1	1758155.66446358	1758155.66446358	13.94	0.0017
X6	1	21863.39694791	21863.39694791	0.17	0.6824

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	7268.399177	3.09	0.0067	2353.287660
X1	0.537548	3.15	0.0059	0.170909
X2	-31.919953	-3.42	0.0033	9.330326
X3	-324.098428	-2.12	0.0491	152.936144
X4	-95.344409	-2.00	0.0612	47.564965
X5	-599.216956	-3.73	0.0017	160.495808
X6	-148.079139	-0.42	0.6824	355.666795

جدول (١٠-٩)
مخرجات SAS المعدلة لمثال (١٠-٦)

General Linear Models Procedure

Dependent Variable: COST

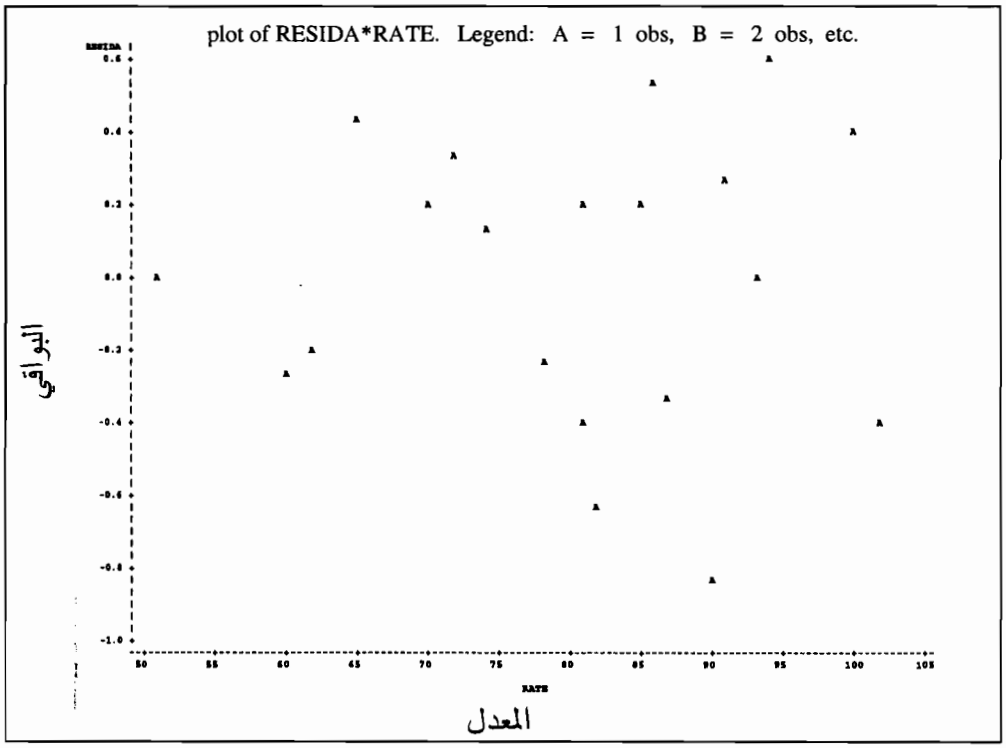
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	154.92332501	51.64110834	270.29	0.0001
Error	16	3.05692999	0.19105812		
Corrected Total	19	157.98025500			

R-Square	C.V.	Root MSE	COST Mean
0.980650	2.405161	0.43710196	18.17350000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
RATE	1	107.72559542	107.72559542	563.84	0.0001
LABOR	1	36.66174860	36.66174860	191.89	0.0001
RATESQRD	1	10.53598099	10.53598099	55.15	0.0001

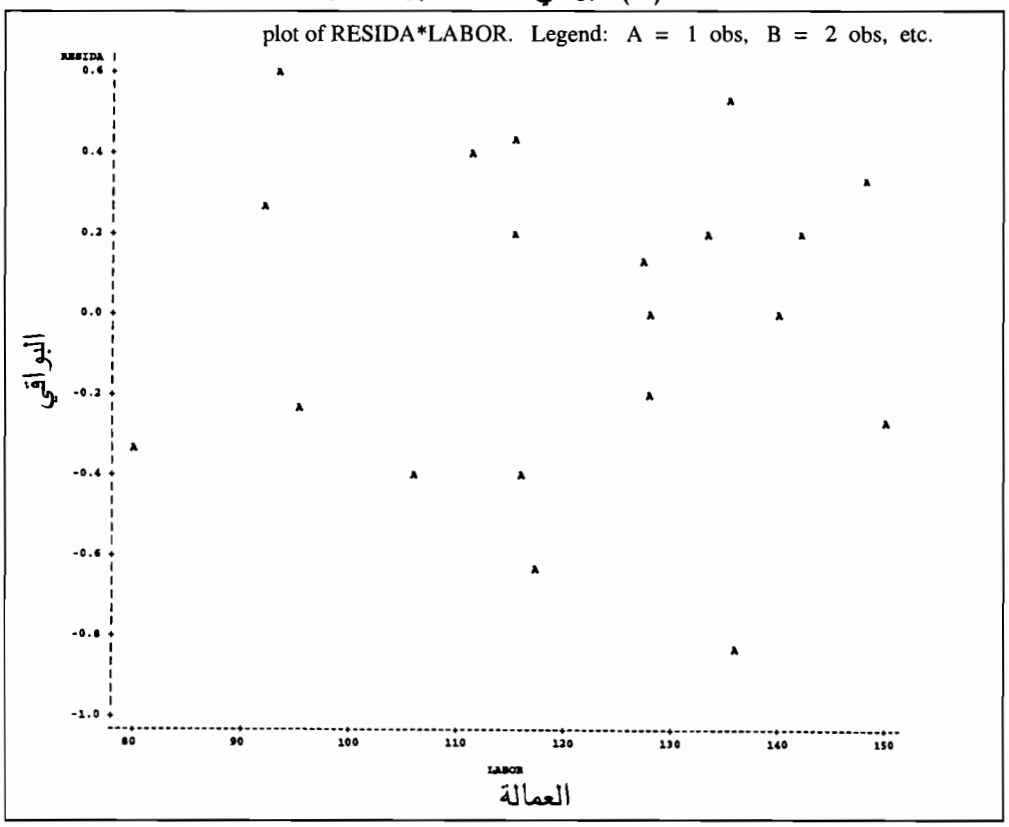
Source	DF	Type III SS	Mean Square	F Value	Pr > F
RATE	1	16.15567401	16.15567401	84.56	0.0001
LABOR	1	35.76285262	35.76285262	187.18	0.0001
RATESQRD	1	10.53598099	10.53598099	55.15	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	41.55145864	13.64	0.0001	3.04686620
RATE	-0.70027157	-9.20	0.0001	0.07615295
LABOR	0.07334821	13.68	0.0001	0.00536113
RATESQRD	0.00362371	7.43	0.0001	0.00048798



شكل (١٠-٦)

(أ) البواقي معدلة مقابل المعدل



شكل (١٠-٦)

(ب) البواقي معدلة مقابل العمالة

مثال(١٠-٧)

بالإشارة لمثال (١٠-٤) . إستخدم دخل العائلة (X_1) كمتغير مفسر وحيد. إستخدم النموذج ($Y = \beta_0 + \beta_1 X_1 + \varepsilon$) لتوفيق بيانات العينة. ثم ارسم شكلا لتوضيح البواقى الناتجة عن معادلة المربعات الصغرى فى مقابل قيم (X_2) حجم العائلة. ماذا ترى؟ إشرح معنى ذلك .

الحل

بإستخدام نتائج البرنامج الإحصائى Minitab والموضحة فى جدول (١٠-١٠) يمكن تحديد الآتى :

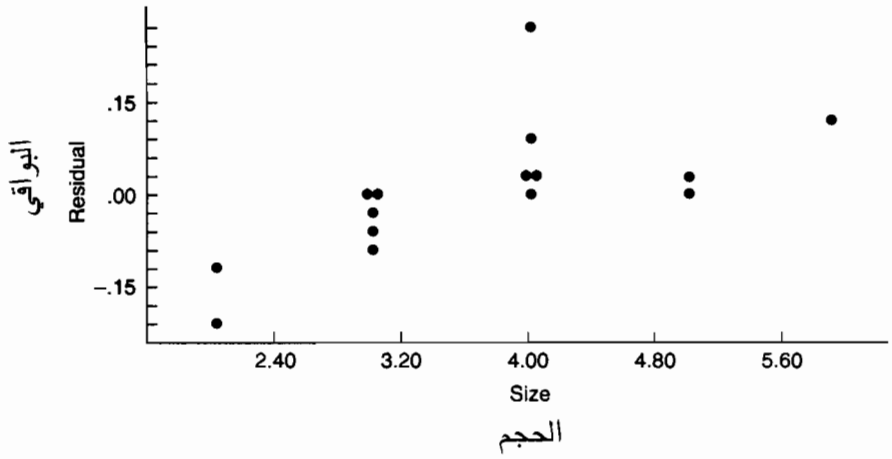
١-خط معادلة المربعات الصغرى ($\hat{Y} = .1619 + .1343\ X_1$) وتقدير المربعات الصغرى للميل ($b_1 = .1343$) موجب كما هو متوقع .

٢-الفرض العدمى ($H_0: \beta_1 = 0$) يتناقض بوضوح مع دليل العينة (P-value = .000) لهذا تظهر بقوة علاقة خطية بين الإنفاق على الطعام ودخل العائلة .

وشكل الإنتشار للبواقى الناشئة عن معادلة المربعات الصغرى $\hat{Y} = .1619 + .1343X_1$ المرسومة فى مقابل قيم X_2 (حجم العائلة)، وشكل (١٠-٧) يشير إلى أن هناك اتجاه خطى لأعلى. وهذا يوضح أن حجم العائلة كمتغير مفسر سوف يحسن معادلة المربعات الصغرى بصورة واضحة، ونحن نعلم أن هذا صحيح من نتائج المثال (١٠-٤) .

جدول (١٠-١٠)
مخرجات ميني تاب لمثال (١٠-٧)

```
The regression equation is
food = 0.162 + 0.134 income

Predictor        Coef        Stdev      t-ratio        p
Constant      0.16190      0.04680         3.46    0.004
income        0.13432      0.01322        10.16    0.000

s = 0.1109   R-sq = 88.8%   R-sq(adj) = 88.0%

Analysis of Variance

SOURCE        DF        SS        MS        F        p
Regression     1    1.2716    1.2716   103.31    0.000
Error         13    0.1600    0.0123
Total         14    1.4316

Unusual Observations

Obs.  income    food     Fit  Stdev. Fit  Residual  St. Resid

  5    6.20  1.2500  0.9947      0.0533    0.2553     2.62R
  9    8.90  1.2900  1.3574      0.0856   -0.0674    -0.95 X

R denotes an obs. with a large st. resid.
X denotes an obs whose X value gives it large influence.
```



شكل (١٠-٧) : الشكل البيانى للبواقى مقابل لحجم العائلة فى مثال (١٠-٧)

(١٠-٧-٢) مشكلة الأزدواج الخطى : The Problem of Collinearity

المشكلة المتكررة فى الإنحدار الخطى المتعدد هي مشكلة إرتباط المتغيرات التفسيرية. إذا كان هذا الإرتباط قليل فإن التأثير يكون صغير، ولكن إذا كان هناك ارتباط قوى بين متغيرين أو أكثر من المتغيرات التفسيرية فإن النتائج تكون وخيمة. هذا يعنى أن مثل هذه المتغيرات تقدم معلومات زائدة عن الحاجة وبالتالى نتائج الإنحدار يمكن أن تكون غامضة جداً. خاصة المتعلقة بقيم تقديرات المربعات الصغرى.

وينشأ عن الإرتباط القوى بين متغيرين (أو أكثر) من المتغيرات التفسيرية، حالة تسمى الأزدواج الخطى Collinearity، وتعرف أيضاً (بالأزدواج الخطى المتعدد) . وترجع مشكلة الأزدواج الخطى إلى نقص البيانات . وهذا ثمن ندفعه عندما لا نستطيع استخدام تصميم التجارب للحصول على بيانات ونضطر إلى الإعتماد على بيانات ملائمة بديلة .

وسبق أن ذكرنا أن معادلة المربعات الصغرى تساعدنا على تقدير المتوسط للمتغير التابع أو تساعد في التنبؤ بالمتغيرات التابعة الفردية بدقة مناسبة. ولا تمنع الإرتباط الخطى جودة التوفيق ولا من التقدير أو التنبؤ داخل مدى قيم المتغيرات التفسيرية . الأزدواج الخطى يؤثر على تقديرات المربعات الصغرى لأنها تميل لتخفيض الدقة المتعلقة بالآثار الإضافية للأزدواج الخطى للمتغيرات المفسرة. وبعبارة أخرى عندما يكون هناك متغيران مرتبطين خطياً فإن معاملات المربعات الصغرى لا تقيس أثر كل منهما على المتغير التابع. وإلى حد ما فإنها تعكس الأثر التابع الذى يكون معرض لإرتباط قد يحدث للمتغيرات المفسرة فى معادلة المربعات الصغرى .

للتوضيح ، إفترض أننا ندخل المتغير التفسيرى (الرطوبة) "humidity" فى مثال الأيس كريم بالإضافة إلى السعر ودرجة الحرارة (لقد اسقطنا نوع اليوم المتغير الوهمي للخطة حتى يتضح لنا بجلاء تأثير الأزدواج الخطي). ومن الممكن أن الرطوبة ودرجة الحرارة مرتبطان بدرجة كبيرة، لهذا فإن كلاهما يشير ببساطة إلى مستوى الراحة لليوم المعطى (حار ورطب في يوم وبارد جاف فى يوم آخر) . وفيما يلى البيانات التى نتعامل معها :

Day	Daily sales	Price	Temperature	Relative Humidity
1	374	35	74	50
2	386	35	82	72
3	472	35	94	92
4	429	50	93	88
5	391	50	82	70
6	475	50	96	94
7	428	50	91	85
8	412	65	93	89
9	405	65	88	80
10	341	65	78	60

لاحظ أنه فى الأيام التى تكون درجة الحرارة منخفضة فيها، تكون الرطوبة أيضاً منخفضة. وعندما تكون درجة الحرارة عالية تكون هى أيضاً كذلك. في هذه البيانات لا توجد أيام فيها درجة الحرارة عالية وتكون الرطوبة منخفضة أو العكس. هذا النوع من الحالات يكون أساس الازدواج الخطى، بمعنى يكون هناك إرتباط قوى بين المتغيرين المفسرين. ونموذج الإنحدار الذى يحتوى درجة الحرارة والرطوبة يكون مرجحاً للتعرض لمشكلتين بسبب الازدواج الخطى :

١- إذا كان المتغيران المفسران عاليان في الإرتباط فإنهما لا يعطيا بمعلومات أساسية مضافة غير تلك البيانات التى نحصل عليها بواسطة المتغيرات الأخرى. لهذا فإن الأثر الفردي لا يتضح لكل متغير على الرغم من أن كل متغير بنفسه يكون مفيداً فى شرح الإختلاف بين قيم Y . وقيم P للاحصاء F الهامشية لآخر متغير يدخل فى النموذج (Type I SS in SAS)، ربما يتناقض مع قيم P للاحصاء F الهامشية للأثر الإضافى للمتغير المفسر فى وجود كل المتغيرات التفسيرية الأخرى فى النموذج (الأحصاء T أو Type III SS in SAS). هذا السلوك يمكن توضيحه فى مخرجات SAS فى جدول (١٠-١١) لمثال الأيس كريم للسعر ودرجة الحرارة والرطوبة كمتغيرات مفسرة. لاحظ أن قيم P للأحصاء F الهامشية أو T للأثر الإضافى للحرارة أو للرطوبة في وجود المتغيرين الأخرين تكون 0627. , 2893. على التوالى وهذا يدل على أنه لا يمكن إعتبار أى من التأثيرات مفيدة فى شرح الإختلاف فى قيم Y للعينة. لكن عندما ننظر إلى للأحصاء F الهامشية لدرجة الحرارة فى وجود السعر فقط (Type I SS) نجد أن قيمة P هى 0001. وهذا يدل على أن درجة الحرارة فى وجود السعر فقط تكون مفيدة تماماً فى شرح الإختلاف فى قيم Y . هذا النوع من التناقض يكون سببه راجع إلى الإرتباط بين درجة الحرارة والرطوبة .

جدول (١٠-١١)

مخرجات SAS لمثال الأيس الكريم للمتغيرات الحرارة، السعر، الرطوبة النسبية

General Linear Models Procedure

Dependent Variable: SALES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14775.53863450	4925.17954483	30.01	0.0005
Error	6	984.56136550	164.09356092		
Corrected Total	9	15760.10000000			

R-Square	C.V.	Root MSE	SALES Mean
0.937528	3.114491	12.80990089	411.30000000

تابع : جدول (١٠-١١)

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PRICE	1	912.66666667	912.66666667	5.56	0.0564
TEMP	1	13641.27520776	13641.27520776	83.13	0.0001
HUMID	1	221.59676008	221.59676008	1.35	0.2893

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PRICE	1	2547.48181406	2547.48181406	15.52	0.0076
TEMP	1	854.12480200	854.12480200	5.21	0.0627
HUMID	1	221.59676008	221.59676008	1.35	0.2893

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-217.6070310	-1.01	0.3514	215.4286267
PRICE	-1.4123454	-3.94	0.0076	0.3584521
TEMP	10.5048795	2.28	0.0627	4.6044332
HUMID	-2.7621885	-1.16	0.2893	2.3769354

٢- معاملات المربعات الصغرى للمتغيرات المرتبطة تكون عالية التقلب ولها أخطاء معيارية كبيرة. هذا يعنى أنه حتى في حالة حدوث تغيرات طفيفة فى بيانات العينة، فإنها يمكن أن تسبب تغيرات كبيرة فى أحد المعاملات. بالتالى فإن معاملات المربعات الصغرى للمتغيرات المرتبطة خطياً لا يعتمد عليها في تفسير علاقات النموذج. على الجانب الآخر فإنها تميل إلى أن تعوض بعضها البعض. إذا كانت المعاملات كبيرة نسبياً للعينة المعطاة فإن معاملات المتغيرات المرتبطة يمكن أن تعوض وتصبح صغيرة. كنتيجة لذلك فإن، كل الآثار للمتغيرات المرتبطة (كمجموعة) فى التقدير أو التنبؤ تقريباً تكون مستقرة ومتعادلة. وللتوضيح بإستخدام مثال الأيس كريم، فإن معامل المربعات الصغرى لدرجة الحرارة يكون $\{b_2 = 5.1953\}$ بخطأ معياري (.5839) عندما لا يدخل عامل الرطوبة فى النموذج. لكن يصبح المعامل ($b_2 = 10.5049$) والخطأ المعيارى 4.6044 عندما تكون الرطوبة مضافة للنموذج [انظر جدول (١٠-١١)]. ويكون معامل المربعات الصغرى للرطوبة سالب (-2.7622). ومن الشائع أن معامل المتغير المرتبط خطياً تكون له إشارة خاطئة كما فى هذه الحالة. لذلك، فإن معادلة المربعات الصغرى تكون غير جيدة. وتكون الرسالة أن الإرتباط أو الأزدواج الخطى يسبب معادلات نموذج غير واقعية ولا تكون مفيدة لغرض التفسير حتى لو أن التنبؤات ظلت ثابتة.

ولقد تم إشتقاق طرق إحصائية معقدة لتتبع وجود عمليات الأزدواج الخطى، لعل أبسطها هو ملاحظة تناقض النتائج بين نوعى الأحصاء F الهامشية أو ملاحظة تقلب المعاملات، والذى يساعد فى هذا الخصوص. والحل الأفضل هو تجنب الإرتباط أو الأزدواج الخطى ككل. ويمكن تحقيق ذلك بواسطة إستخدام تجارب مصممة جيدا للحصول على بيانات عينة، ويكون ذلك بأختيار قيم المتغيرات التفسيرية التى تؤكد غياب الإرتباط الخطى. ففى مثال الأيس كريم، يعنى ذلك أن نلاحظ المبيعات للأيام بدرجة حرارة عالية ورطوبة منخفضة أو درجة حرارة منخفضة ورطوبة عالية. بالإضافة للأيام التى يكون فيها كلاهما عالى، أو كلاهما منخفض. لسوء الحظ فإنه عندما تكون البيانات الملائمة هى فقط المصدر الوحيد للمعلومات، فإن هذا الأسلوب يكون غير واضح أو غير مقبول عادة.

كيف إذن نعالج الموقف عندما لا يكتشف الإرتباط أو الأزدواج الخطى؟ يمكن إتباع هذه الطرق المباشرة:

١- إن أمكن، أضف لبيانات العينة قيم المتغيرات المرتبطة التى تميل إلى تقليل شدة أو حدة الإرتباط. فى مثال الأيس، كريم هذا يعنى إضافة المبيعات المشاهدة للأيام ذات درجة الحرارة العالية والرطوبة منخفضة أو العكس.

٢-إحذف واحد أو أكثر من المتغيرات المرتبطة. فى مثال الأيس كريم مثلا، نحذف الرطوبة ويضم السعر ودرجة الحرارة ونوع اليوم، سنحصل على معادلة إنحدار أفضل للتقدير والتنبؤ .

٣- نشكل متغير مفسر جديد والذى يخدم كمؤشر أو دليل للمتغيرات المرتبطة. فى مثال الأيس كريم، يمكن إنشاء متغير جديد عبارة عن متوسط المتغيرات: درجة الحرارة والرطوبة. وبإستخدام المتغير الجديد فى النموذج بدلاً من درجة الحرارة والرطوبة، نحذف الإرتباط أو الأزدواج الخطى بينما يحتفظ النموذج بالمعلومات لكلا المتغيرين درجة الحرارة والرطوبة.

مثال (١٠-٨)

البيانات التالية تمثل متوسط درجة حرارة الجو Y فى يناير لعدد 24 محطة قياس الطقس فى فرجينيا، حيث أن كل محطة تحدد بخط العرض X_1 ، وخط الطول X_2 ، وإرتفاع سطح البحر X_3 على التوالى. بإستخدام X_1 , X_2, X_3 أوجد النموذج الخطى المناسب لهذه البيانات ثم وفق البيانات بهذا النموذج :

Y(temperature)	X_1(latitude)	X_2(longitude)	X_3(elevation)
37.9	37.35	79.52	975
28.7	38.52	78.43	3.535
38.3	37.08	77.95	440
37.3	37.53	79.68	870
31.5	37.08	81.33	3.300
35.0	37.38	80.08	1.890
36.0	38.03	78.52	870
37.4	36.83	79.37	700
40.4	37.28	75.97	11
35.8	37.77	78.15	300
35.3	38.47	78.00	420
33.2	38.45	78.93	1.400
41.3	36.90	76.20	25
34.7	38.45	77.67	300
38.0	37.33	78.38	450
34.2	36.93	80.30	2.600
35.4	38.30	77.47	100
35.7	37.37	80.00	1.524
39.7	36.68	76.78	80
40.5	37.30	77.30	40
31.6	38.00	79.83	2.238
40.0	37.08	76.35	10
36.1	37.78	79.43	1.060
34.1	39.12	77.72	500

الحل

جدول(١٠-١٢) يوضح مخرجات النموذج ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$) وعلى الرغم من أن التوفيق يظهر جيداً (متوسط مربعات الخطأ قليلة و R^2 كبيرة)، إلا أن هناك تعارض بين نوعى الأحصاء F الهامشية. على سبيل المثال. الأثر الإضافى لخط الطول فى وجود خط العرض فقط (Type I SS) يكون مفيداً جداً فى شرح الإختلاف فى درجة الحرارة (P-value = .0001) لكن نفس الأثر فى وجود كل من خط العرض وإرتفاع مستوى سطح البحر

غيـر مفـيد (P-value = .2511) وهذا يدل على أن خط الطول وإرتفاع سطح البحر ربما يكونا مرتبطان. وهذا متوقع إذا أخذنا جغرافية ولاية فرجينيا في الاعتبار. فكلما انتقلنا من الشرق للغرب يتزايد خط الطول ونلاحظ ميل للتزايد فى الارتفاع عن سطح البحر. أي أن هناك إرتباط بين خط الطول والإرتفاع عن سطح البحر.

وحيث أننا لا نستطيع زيادة بيانات العينة بسبب الظروف الجغرافية، دعنا نخرج إما خط الطول أو الإرتفاع عن سطح البحر من نموذج الإنحدار. ويوضح جدول (١٠-١٣) مخرجات SAS لخط العرض وخط الطول فقط. وجدول (١٠-١٤) يوضح مخرجات SAS لخط العرض والإرتفاع عن سطح البحر. ومن الواضح أنه حتى الآن فإن معادلة الإنحدار التى تحتوى فقط على خط العرض وإرتفاع سطح البحر تكون أفضل بكثير عن التى تحتوى على خط العرض والطول. على سبيل المثال، فإن تباين البواقى يكون (.7645) بالمقارنة بالمقدار 2.8939 ومعامل التحديد R^2 تكون .9285 فى مقابل .7293. والأهم من كل هذا فإن دقة تقدير المعالم لخط العرض وإرتفاع سطح البحر يكون أفضل من تقدير معالم خط العرض وخط الطول (قيمة T تساوى 8.91- ، 13.13- فى مقابل 5.31- ، 5.49-) على التوالى :

جدول (١٠-١٢)

مخرجات SAS لمثال (١٠-٨)

خط الطول، خط العرض، الارتفاع عن سطح البحر: متغيرات تفسيرية

Dependent Variable: TEMP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	209.50414409	69.83471470	93.08	0.0001
Error	20	15.00543924	0.75027196		
Corrected Total	23	224.50958333			

R-Square	C.V.	Root MSE	TEMP Mean
0.933163	2.394699	0.86618241	36.17083333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LAT	1	76.65032249	76.65032249	102.16	0.0001
LONG	1	87.08679739	87.08679739	116.07	0.0001
ELEV	1	45.76702422	45.76702422	61.00	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LAT	1	61.76107570	61.76107570	82.32	0.0001
LONG	1	1.04809648	1.04809648	1.40	0.2511
ELEV	1	45.76702422	45.76702422	61.00	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	151.7116876	7.75	0.0001	19.57344898
LAT	-2.5354467	-9.07	0.0001	0.27945148
LONG	-0.2306516	-1.18	0.2511	0.19514853
ELEV	-0.0020749	-7.81	0.0001	0.00026566

جدول (١٠-١٣)

مخرجات SAS لمثال (١٠-٨)

خط الطول، خط العرض: متغيرات تفسيرية

General Linear Models Procedure

Dependent Variable: TEMP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	163.73711988	81.86855994	28.29	0.0001
Error	21	60.77246346	2.89392683		
Corrected Total	23	224.50958333			

R-Square	C.V.	Root MSE	TEMP Mean
0.729310	4.703111	1.70115456	36.17083333

تابع : جدول (١٠-١٣)

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LAT	1	76.65032249	76.65032249	26.49	0.0001
LONG	1	87.08679739	87.08679739	30.09	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LAT	1	81.73903603	81.73903603	28.25	0.0001
LONG	1	87.08679739	87.08679739	30.09	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	252.8595453	8.77	0.0001	28.82433419
LAT	-2.8802143	-5.31	0.0001	0.54194325
LONG	-1.3803345	-5.49	0.0001	0.25162394

جدول (١٠-١٤)

مخرجات SAS لمثال (١٠-٨)

خط العرض والارتفاع عن سطح البحر : متغيرات تفسيرية

General Linear Models Procedure

Dependent Variable: TEMP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	208.45604761	104.22802381	136.34	0.0001
Error	21	16.05353572	0.76445408		
Corrected Total	23	224.50958333			

R-Square	C.V.	Root MSE	TEMP Mean
0.928495	2.417226	0.87433065	36.17083333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LAT	1	76.65032249	76.65032249	100.27	0.0001
ELEV	1	131.80572512	131.80572512	172.42	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LAT	1	60.71812652	60.71812652	79.43	0.0001
ELEV	1	131.80572512	131.80572512	172.42	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	132.1137062	12.58	0.0001	10.49873166
LAT	-2.4894338	-8.91	0.0001	0.27932969
ELEV	-0.0023118	-13.13	0.0001	0.00017606

تمارين

(١٠-٣١) إستخدم بيانات مثال الأيس كريم (والمذكور في الجزء (١٠-٣)) لتوفيق النموذج الخطى مستخدماً السعر (X_1) كمتغير مفسر. ارسم البواقى فى مقابل قيم درجة الحرارة. بناءً على النتائج التي توصلت إليها، هل هي نتائج غير متوقعة، إشرح.

(١٠-٣٢) بالإشارة إلى تمرين (٩-١٤)، (٩-٣٩) فى الفصل التاسع. هذه التمارين تشمل العلاقة بين نسبة الضرائب المدفوعة (Y) وإجمالى الدخل السنوى (X). معتمداً على إجابتك فى الجزء (جـ) من التمرين (٩-٣٩)، أضف إلى النموذج المفترض الحد الذي تعتقد أنه يجب إضافته، ووفق النموذج الجديد للبيانات، وقيم نتائج معادلة المربعات الصغرى.

(١٠-٣٣) شركة بناء كبيرة تريد دراسة العلاقة بين حجم العروض X (بالمليون دولار) وتكلفة أعداد عروض الشركة Y (بالألف دولار) للعروض الاثنتى عشر التالية والتي تمثل عينة :

Y	X	Y	X
20	3.37	20.5	2.43
28.2	3.75	16.1	1.61
47	10.89	67.6	11.4
15	1.5	40.4	6.4
25	4.76	29.9	6
52.5	8.4	33.1	7.31

أ – أنشئ شكل الإنتشار ووفق النموذج الملائم لهذه البيانات .

ب– قيم نتائج معادلة المربعات الصغرى وارسم البواقى فى مقابل قيم X .

ج– بناءً على النتائج التي يظهرها شكل البواقى . هل تُستخدم هذه المعادلة فى التقدير والتنبؤ؟ دعم إجابتك .

(١٠–٣٤) يوجد شك بأن الغياب بسبب مرض المديرين يمكن تقليله بواسطة إخضاعهم لبرنامج ممارسة التمارين الرياضة أو تقليل إستهلاك الكافيين . فى تجربة تشمل 20 مدير من الذين يشربوا القهوة ولا يمارسوا التمارين بإنتظام ، قام 10 مديرين بتقليل إستهلاك القهوة إلى أقل من فنجان واحد فى اليوم وقاموا بالاشتراك في ممارسة التمارين الرياضية ، 10 مديرين مازالوا على عاداتهم . وتم تحصيل أيام الغياب لجميع المديرين لأكثر من سنة . وكانت معادلة الانحدار المناسبة لتوفيق البيانات هي :

$$\hat{Y} = 2.1 - .66X_1 - 15.4X_2$$

حيث Y = عدد مرات الغياب ،

X_1 = متوسط إستهلاك القهوة فى اليوم (بالأوقية)

X_2 = 1 إذا كان المدير يمارس الرياضة ، وتساوى صفر إذا كان غير ذلك .

ويبدو أن النموذج تم توفيقه جيداً ، حيث: (R_a^2= .85) ، كما كانت (P-value=.001) للاختبار F للفرض العدمي ($H_0 : \beta_1 = \beta_2 = 0$) . ومع ذلك يبدو أن إشارة b_1 خطأ والقيمة السالبة لـ b_1 تفيد بأن تناقص إستهلاك القهوة يزيد الغياب – بالإضافة لذلك (b_2 = -15.4) تدل على تقليل 15.4 يوم كنتيجة لممارسة الرياضة (التخفيض بهذه القيمة الكبيرة يبدو وغير واقعي) . اشرح هذه النتائج .

(١٠–٣٥) البيانات التالية تحتوى على درجة الحرارة Y (كيفية الشعور بالدفء) ، X_1 درجة حرارة الجو والرطوبة النسبية X_2 .

Y	66	72	77	67	73	78	68	74	79
X_1:°F	70	75	80	70	75	80	70	75	80
X_2 :%	20	20	20	30	30	30	40	40	40

أ – احسب معامل الإرتباط بين X_1 , X_2 . بناءً على هذه النتيجة ما مدى العلاقة الخطية بين X_1 , X_2 ؟

ب- وفق النموذج($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$) لبيانات العينة وقيم المعادلة . هل اكتشفت الازدواج الخطى بين X_1 , X_2 ؟.

ج - وفق الخط المستقيم لبيانات العينة بإستخدام X_1 فقط ، ثم بإستخدام X_2 فقط ، ثم قارن بين معاملات المربعات الصغرى والتى حددتها فى هذا الجزء وبين التى حددتها فى (ب) ؟

د - بناءً على نتائج الجزء (أ) . هل اندهشت لما اكتشفته فى الأجزاء (ب) ، (ج) ؟ اشرح إجابتك .

(١٠-٣٦) اجريت تجربة لتحديد العلاقة بين متوسط درجات طلاب كلية ما ، (1) عدد ساعات الدراسة فى الأسبوع ، (2) عدد ساعات مشاهدة التليفزيون فى الأسبوع . وتم ملاحظة هذه المتغيرات خلال فصل دراسي لعينة مكونة من 50 طالبا . هل يمكنك توضيح الصعوبات التي يمكن أن تظهر لنتائج معادلة المربعات الصغرى؟ ما اقتراحاتك؟

(١٠-٣٧) اذا فرض وجود متغيرين تفسيريين فى معادلة إنحدار وكانت درجة الإرتباط بينها عالية، بأى طريقة يؤثر ذلك بالضرر على معادلة المربعات الصغرى؟

(١٠-٣٨) تحت أى ظرف يجب إزالة واحدة من المشاهدات من مجموعة بيانات العينة المستخدمة فى تحديد معادلة المربعات الصغرى؟

(١٠-٣٩) إذا كان المطلوب توفيق النموذج : ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$) للبيانات التالية :

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
17	.297	.310	.290	17	.099	.092	.074
17	.360	.390	.369	73	.420	.452	.425
35	.075	.058	.047	17	.189	.178	.153
69	.114	.100	.081	35	.369	.391	.364
69	.229	.213	.198	69	.142	.124	.105
173	.315	.304	.267	35	.094	.087	.072
173	.477	.518	.496	35	.171	.161	.145
17	.072	.063	.047	52	.378	.420	.380

أ - هل اكتشفت العلاقة بين Y والثلاث متغيرات المفسرة كمجموعة ؟ دعم إجابتك.

ب- هل اكتشفت أى ازدواج إرتباط خطى بين المتغيرات الثلاث المفسرة؟ دعم إجابتك .

ج- افترض أنك حذفت X_3 من النموذج. أعد توفيق البيانات وحدد ما إذا كان حذف X_3 حسن الموقف أم لا؟

(١٠-٨) معيار لإختبار أفضل مجموعة من المتغيرات التفسيرية :

Criteria for Selecting The Best Set of Predictor Variables

وكما سبق أن إقترحنا ، فإن مشكلة هامة تظهر فى تحليل الإنحدار وهى تحديد أي المتغيرات المفسرة فى القائمة الأصلية يجب أن تضم إلى نموذج الإنحدار. ونحن نعتقد أنه لكل محلل أو باحث وجهة نظر لشرح القائمة الرئيسية للمتغيرات التفسيرية التى يعتقد أنها تكون ذو أهمية لشرح

الإختلاف فى المتغير التابع. وما نحتاجه هو طريقة لتحديد المتغيرات التفسيرية من القائمة الأساسية لهذه المتغيرات التفسيرية والتى تظهر أفضل مجموعة لشرح معظم الإختلاف فى قيم Y . وكلمة أفضل (best) هنا تعنى أن نتائج معادلة المربعات الصغرى تزودنا بالدقة الكافية للتقدير والتنبؤ داخل نطاق المتغيرات التفسيرية وبدون أى إمكانية لظهور تناقضات .

عند تحديد أفضل معادلة مربعات صغرى، فإن هناك معيارين شائعين من أهم المعايير المفيدة وهما تباين البواقى (وهذا بمعنى آخر مكافئ لقيمة R^2 المعدلة)، والأخطاء المعيارية لمعاملات المربعات الصغرى. وربما تكون لاحظت أن هذان المعياران لهما دور كبير فى إرشادنا عند التحليل فى هذا الفصل .

١- تباين البواقى S_e^2 : وتباين البواقى هو نفسه متوسط مربعات الخطأ (MSE). حيث أن (MSE) هو مجموع مربعات البواقى مقسوما على درجات الحرية للمجموع SSE . وتأخذ MSE في الحسبان عدد الحدود الموجودة فى النموذج من خلال درجات الحرية. بينما لا تزيد قيمة SSE إذا تم السماح لعدد إضافي من المتغيرات التفسيرية بالدخول للنموذج، S_e^2 يمكن أن تزيد إذا كان الإنخفاض فى SSE صغير جداً ولا يحتمل فقدان درجات حرية إضافية. على سبيل المثال انظر للجداول (١٠-٥)، (١٠-٦) والتى تكون معادلة المربعات الصغرى تحتوى على متغيرين تفسيريين، انظر جدول (١٠-٦) وفيها يكون تباين البواقي (1.26)، أصغر من نظيره في النموذج بالثلاث متغيرات (1.361) جدول (١٠-٥). مع معيار تباين البواقى نحدد مجموعة المتغيرات التفسيرية التى تقلل إما S_e^2 أو تقللها إلى النقطة التى لا تستطيع إدخال متغيرات تفسيرية أخرى للنموذج لأنها ستكون غير نافعة.

فى الجزء (١٠-٣-٣) نتذكر أن قيمة R^2 المعدلة تأخذ في اعتبارها عدد الحدود فى النموذج. لهذا السبب يعتبر هذا المعيار مكافئ لتباين البواقى. وإستخدام قيمة R^2 المعدلة تحدد لنا مجموعة المتغيرات التفسيرية التى تعظم R_a^2 أو تقريباً تعظمها للنقطة التى تكون إضافة متغيرات تفسيرية أخرى بعدها غير مفيدة .

٢- الأخطاء المعيارية لمعاملات المربعات الصغرى: تعتبر دقة المعلمات أو المؤشرات المقدرة لنموذج الإنحدار المفترض واحدة من أهم الإعتبارات عن تحديد أفضل المتغيرات التفسيرية. وكلما صغرت الأخطاء المعيارية لمعاملات المربعات الصغرى، كلما كانت الدقة أفضل، وكلما كان استخدام المربعات الصغرى أفضل للتقدير والتنبؤ. وهذا يعنى أنه كلما صغرت قيمة الأخطاء المعيارية بالنسبة لمقدرات المربعات الصغرى المناظرة، كلما كبرت قيمة T . وعندما تكون قيم T كبيرة تكون قيم P المقابلة صغيرة. وكنتيجة لهذا فإن الأثر الإضافى لكل متغير مفسر فى وجود المتغيرات المفسرة الأخرى فى المجموعة الأفضل يكون مفيداً تماماً فى شرح الإختلاف فى قيم Y .

إستخدام هذه المعايير، يمكننا من إيجاد أفضل مجموعة متغيرات تفسيرية وذلك بتحديد وتقييم كل معادلات المربعات الصغرى الخطية المشتملة على القائمة الأساسية للمتغيرات التفسيرية. وإذا كان هناك متغيران تفسيريان فى القائمة الأساسية، فان هذا يعنى أن إجمالى الثلاث معادلات : معادلتان تحتوى كل منهما على متغير واحد فقط، ومعادلة تحتوى على المتغيران معاً. إذا كان هناك ثلاث متغيرات تفسيرية، فإنه يجب أن يكون هناك سبعة معادلات مربعات صغرى: ثلاث معادلات تحتوى كل منهم على متغير واحد فقط وثلاثة تحتوى على متغيرين ومعادلة أخيرة تحتوى على الثلاث متغيرات معاً. بصفة عامة، إذا احتوت القائمة الأساسية على K متغير مفسر، فهناك $(2^k - 1)$

من المعادلات الخطية للمربعات الصغرى الممكنة. كل واحدة منها تحتوى على الأقل على متغير واحد مفسر .

أساليب إختيار المتغير Variable Selection Techniques

عندما تكون K كبيرة ($K \geq 5$) . فالتحديد والتقييم لكل معادلات الإنحدار الخطية ربما لا تكون عملية. ولمثل هذه الحالة، فإن أساليب إختيار المتغير المستخدمة يمكن أن تزودنا بمعلومات مفيدة بدون تقييم كل المعادلات الممكنة. وعلى العموم فإن هذه الأساليب لا تعتبر أساليب متساوية مع أساليب التقييم لكل المعادلات الممكنة معا والتي تستخدم المعايير السابقة. وأشهر أسلوب لاختيار المتغيرات هو الإنحدار المتدرج **stepwise regression** لتحديد أفضل مجموعة متغيرة مفسرة. وهناك نوعان أساسيان لهذا الأسلوب: الإختيار الأمامى **forward selection** والحذف الخلفى **backward eliminatin** .

أسلوب الإختيار الأمامى : (في حالة الانحدار المتدرج) Forward Selection

يبدأ أسلوب الإختيار الأمامى بمعادلة لا تحتوى على متغيرات مفسرة ($\hat{Y} = \overline{Y}$). المتغير المفسر الأول الداخل إلى النموذج هو الذى ينتج عنه أكبر تخفيض فى مجموع مربعات الأخطاء. وإذا اعتمد على قيمة P فإن هذا المتغير يكون مفيداً فى شرح الإختلاف فى قيم Y، وبالتالي يبقى في النموذج ويتم البحث عن متغير ثان. المتغير الثانى الذى يتم إدخاله للنموذج هو الذى ينتج عنه أكبر تخفيض فى مجموع مربعات الأخطاء فى وجود المتغير الأول. إذا كان الأثر الإضافى للمتغير الثانى مفيداً حقاً وذلك عن طريق معرفة قيمة P ، فإن المتغير الثانى يبقى بالنموذج ونبحث عن متغير مفسر ثالث. وتستمر العملية بهذا الأسلوب حتى يكون الأثر المضاف للمتغير المفسر الأخير المدخل للنموذج غير مفيد .

وأسلوب الإختيار الأمامى تم تعديله بحيث أن إمكانية إلغاء متغير أخذت في الاعتبار كل مرحلة. هذا التعديل ينتج ما هو معروف فى الحزم الإحصائية بأسلوب الإنحدار المتدرج (**stepwise regression**) . مع هذه الطريقة فإن المتغير المفسر والذي تم إدخاله في مرحلة مبكرة، يمكن حذفه في مرحلة لاحقة. ويكون القرار معتمداً على مدى التخفيض في مجموع مربعات الأخطاء، ويكون معتمدا أيضاً على مزيج خاص من المتغيرات فى نموذج الإنحدار .

أسلوب الحذف الخلفى : (في الانحدار المتدرج) Backward Elimination

عملية الحذف الخلفى تبدأ بنموذج الإنحدار الذى يحتوى على كل المتغيرات التفسيرية فى القائمة الأساسية، ثم يتم حذف المتغيرات الأقل أهمية متغير بعد الآخر، وتحدد هذه الأهمية بمدى مساهمتها فى تخفيض مجموع مربعات الخطأ (أى نحذف المتغيرات الأقل تأثير فى تخفيض مجموع مربعات الأخطاء). على سبيل المثال، المتغير المحذوف الأول يكون المتغير الذى ينتج عنه إنخفاض صغير فى مجموع مربعات الأخطاء فى وجود المتغيرات الأخرى. وتنتهى العملية عندما يكون الأثر الإضافى لكل المتغيرات الباقية مفيداً اعتماداً على قيم P-value المناسبة .

وتزودنا العديد من الحزم الإحصائية مثل Minitab , SAS بهذه الأساليب لإختيار المتغيرات (سواء الاختيار الأمامي، الاختيار الامامي المعدل أو الحذف الخلفي). ويجب أن نلاحظ أن أى إجراء منهم لا يجب إعتباره كبديل لتقييم النموذج. وعموماً فإن العديد من أوجه التقييم والتى تتضمن تحليل البواقى وتناقضات أو عيوب النموذج يظل مسئولية المستخدم وليس على برنامج الحاسب .

مثال (١٠-٩)

بالإشارة إلى مثال الأيس كريم إفترض أن قائمة المتغيرات التفسيرية الأساسية تحتوى على السعر، ودرجة الحرارة ونوعية اليوم والرطوبة النسبية. إستخدم أسلوب الاختيار الأمامي المعدل والحذف الخلفى لتحديد أفضل مجموعة للمتغيرات التفسيرية.

الحل

يوضح جدول (١٠-١٥) وجدول (١٠-١٦) نتائج أو مخرجات البرنامج الإحصائي SAS لعمليات التعديل الأمامية والحذف الخلفى. لاحظ أن كلا الإجرائين يصلا إلى نفس الإستنتاج. وأفضل مجموعة متغيرات مفسرة للعينة المعطاة هى السعر ودرجة الحرارة ونوعية اليوم، لهذا فإنه كما سبق القول فى الجزء (١٠-٥) أن معادلة المربعات الصغرى : ($\hat{Y}= 15.3094 - 1.1012X_1 + 5.039 X_2 + 20.2452X_3$) تمكننا من التقدير والتنبؤ بدقة كافية.

ومن الجداول (١٠-١٥)، (١٠-١٦) لاحظ العمود المعنون (Type II SS) ، المقادير فى هذا العمود هى مجموع المربعات الناشئة عن مبدأ مجموع المربعات الإضافية. كل كمية تمثل المقدار الذى يمكن أن يزداد بها مجموع مربعات الخطأ، إذا تم حذف المتغير المفسر (رأس الصف) من نموذج الإنحدار. هذا يعنى أنه كلما إرتفعت القيمة فى هذا العمود، كلما صغرت قيمة P ، كلما زادت الأهمية للأثر الإضافى للمتغير المفسر المقابل .

بالإضافة إلى ذلك، شاهد القيمة المعروفة على أنها C (P) . هذه القيمة لإحصاء يسمى الإحصاء C_p، (C_p Statistic). والإحصاء C_p هو معيار آخر لتحديد مدى جودة معادلة المربعات الصغرى فيما يتعلق بالتقدير والتنبؤ. وعلى الرغم أن المناقشة المستفيضة لهذا الإحصاء C_p هي خارج نطاق هذا الكتاب، لكن من الممكن القول بأن معادلة المربعات الصغرى والتي لها قيمة C_p قريبة من عدد المعاملات فى النموذج، شاملة الجزء المتطوع، تكون مرغوبة في التقدير والتنبؤ. من جدول (١٠-١٥)، (١٠-١٦) لاحظ أن ($C_p = 4.19$) لأفضل معادلة مربعات صغرى. وتقرب هذه القيمة من الرقم 4 وهو عدد حدود النموذج بالإضافة إلى الجزء المقطوع من المحور الرأسى .

جدول (١٠-١٥)

مخرجات SAS لمثال الأيس كريم باستخدام أسلوب الاختيار الأمامي المعدل

Stepwise Procedure for Dependent variable Sales

Step 1 Variable TEMP Entered R-square = 0.77326744 C(p) = 66.93862745

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	12186.77219117	12186.77219117	27.28	0.0008
Error	8	3573.32780883	446.66597610		
Total	9	15760.10000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-10.80516477	81.08638826	7.93139807	0.02	0.8973
TEMP	4.84621314	0.92778982	12186.77219117	27.28	0.0008

Bounds on condition number: 1, 1

Step 2 Variable PRICE Entered R-square = 0.92346761 C(p) = 20.62005248

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	14553.94187442	7276.97093721	42.23	0.0001
Error	7	1206.15812558	172.30830365		
Total	9	15760.10000000			

تابع : جدول (١٠-١٥)

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	25.87773161	51.32600406	43.80096790	0.25	0.6296
PRICE	-1.34175131	0.36200155	2367.16968325	13.74	0.0076
TEMP	5.19529086	0.58389598	13641.27520776	79.17	0.0001

Bounds on condition number: 1.026712, 4.106846

Step 3 Variable DAY Entered R-square = 0.98076664 C(p) = 4.18726737

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	15456.98030414	5152.32676805	101.99	0.0001
Error	6	303.11969586	50.51994931		
Total	9	15760.10000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	15.30937021	27.90393324	15.20711073	0.30	0.6030
PRICE	-1.10118194	0.20410657	1470.50695790	29.11	0.0017
TEMP	5.03906542	0.31831702	12660.27582653	250.60	0.0001
DAY	20.24521411	4.78851344	903.03842972	17.87	0.0055

Bounds on condition number: 1.11323, 9.729814

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable SALES

Step	Variable Entered	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	TEMP		1	0.7733	0.7733	66.9386	27.2839	0.0008
2	PRICE		2	0.1502	0.9235	20.6201	13.7380	0.0076
3	DAY		3	0.0573	0.9808	4.1873	17.8749	0.0055

جدول (١٠-١٦)

مخرجات SAS لمثال الأيس كريم باستخدام أسلوب الحذف الخلفي

Backward Elimination procedure for dependent variable SALES

Step 0 All Variables Entered R-square = 0.98445731 C(p) = 5.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	15515.14558190	3878.78639547	79.17	0.0001
Error	5	244.95441810	48.99088362		
Total	9	15760.10000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-112.82158874	120.76038653	42.76128091	0.87	0.3931
PRICE	-1.15426971	0.20681485	1526.04177981	31.15	0.0025
TEMP	7.85847018	2.60643466	445.34605432	9.09	0.0296
DAY	18.92077590	4.86963138	739.60694740	15.10	0.0116
HUMID	-1.46141210	1.34121509	58.16527776	1.19	0.3256

Bounds on condition number: 71.95518, 581.2342

Step 1 Variable HUMID Removed R-square = 0.98076664 C(p) = 4.18726737

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	15456.98030414	5152.32676805	101.99	0.0001
Error	6	303.11969586	50.51994931		
Total	9	15760.10000000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	15.30937021	27.90393324	15.20711073	0.30	0.6030
PRICE	-1.10118194	0.20410657	1470.50695790	29.11	0.0017
TEMP	5.03906542	0.31831702	12660.27582653	250.60	0.0001
DAY	20.24521411	4.78851344	903.03842972	17.87	0.0055

Bounds on condition number: 1.11323, 9.729814

تابع : جدول (١٠-١٦)

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable SALES

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	HUMID	3	0.0037	0.9808	4.1873	1.1873	0.3256

تمارين

(١٠-٤٠) ناقش أهم معيارين لتحديد أى المتغيرت التفسيرية فى مجموعة يجب أن تتضمنها معادلة الإنحدار.

(١٠-٤١) ناقش ما إذا كانت معادلة المربعات الصغرى الناتجة بإستخدام إسلوب إختيار المتغيرات تعتبر أفضل معادلة إنحدار تستخدم فى التقدير أوالتنبؤ ؟

(١٠-٤٢) بالإشارة إلى تمرين (١٠-٣٩) إستخدم طريقة الحذف الخلفى والأمامى المعدل لإختيار المتغيرات لتحديد أفضل مجموعة متغيرات مفسرة تستخدم فى معادلة الإنحدار. فسر ما تجده؟

(١٠-٤٣) بالإشارة إلى مثال (١٠-٨) فى الجزء (١٠-٧) . بين ما إذا كانت طريقة الاختيار الأمامى المعدل والحذف الخلفى لإختيار المتغيرات المفسرة، تؤدي إلى أن أفضل معادلة انحدار تحتوى فقط على latitude وكذلك elevation، (خط العرض والارتفاع عن سطح البحر).

(١٠-٩) الإنحدار الخطى المتعدد : مثال شامل :
Multiple Linear Regression: A Comprehensive Example

من القضايا التى فحصت فى هذا الفصل، هناك عدد من الخطوات الضرورية التى يجب إتباعها عند تحسين نموذج الإنحدار الخطى المتعدد. هذه الخطوات هى :

خطوات تحسين نموذج الإنحدار

١-تحديد قائمة المتغيرات المفسرة الأساسية الواجب إعتبارها لتضمينها النموذج .

٢- الحصول على بيانات عينة وتقرير ما إذا كانت البيانات ممثلة للبيئة محل الدراسة .

٣-مبدئياً على الأقل، نفرض أن هناك علاقة بين المتغير التابع والمتغيرات المفسرة والممثلة بواسطة نموذج الإنحدار الخطى المعطي فى المعادلة (10.1) .

٤- توفيق هذا النموذج لبيانات العينة وتقييم معاملات المربعات الصغرى، وهل إشاراتها تتوافق مع العلاقات التى لها معنى؟

٥-تقييم معادلة المربعات الصغرى، ويجب أن يتضمن التقييم الحد الأدنى من تناقضات النموذج (تحليل البواقى) والمشاكل (الإرتباط أو الأزدواج الخطى) وتحديد أفضل مجموعة من المتغيرات المفسرة. وبالتالى تحسين النموذج كنتيجة من نتائج التقييم .

لتوضيح هذه الخطوات نعتبر المثال التالى : مدير شركة مرافق عامة يريد تقديم نموذج للعوامل التى تؤثر على إستخدام الكهرباء فى المنازل السكنية أثناء موسم إستخدام التدفئة (من نوفمبر إلى أبريل لمنطقة جغرافية معينة) .

قائمة المتغيرات المفسرة الممكنة :

يرى المدير أن كميات الكهرباء المستخدمة لهذه المنازل تعتمد على : (1) حجم المساحة التى يتم تدفئتها ، (2) كيف يتم عزل حوائط المنازل فى تلك المنطقة ، (3) نوعية نظام التدفئة فى المنازل ، (4) برودة الطقس ، (5) متوسط الدرجة التى تظهر على مقياس الحرارة (ترموستات) thermostut . ولقد قام المدير بتعريف المتغيرات التالية : Y = عدد الكيلو وات ساعة / شهر ، X_1 = مربع المساحة التى يتم تدفئتها ، X_2 = قيمة R التى توضح قوة مواد العزل ، X_3 = 1 إذا كان المنزل يتوافر فيه نوافذ معزولة ، 0 إذا لم تكن كذلك ، X_4 = متوسط درجة الحرارة ، X_5 = 1 إذا تم إستخدام آلة لتوليد درجة الحرارة ، 0 إستخدام قوة التيار الكهربائى ، X_6 = متوسط عدد الساعات المشمسة فى اليوم . ويستطيع المدير الحصول على هذه البيانات جميعها ما عدا درجات الحرارة التى توجد على الترموستات .

الحصول على بيانات العينة :

يختار المدير عينة ممثلة من 25 فاتورة عميل شهرية مأخوذة من عدة مواسم تدفئة حديثة . يحصل المدير على بيانات المتغيرات من X_1 حتى X_6 من الفاتورة المختارة . ومن سجلات الشركة وبواسطة إجراء إستطلاع للمنازل المختارة ومن معلومات خدمات الطقس القومية تكون البيانات للعينة كما يلى:

Y	X_1	X_2	X_3	X_4	X_5	X_6
2.405	1.400	0	0	40	0	11.0
1.064	1.650	11	1	41	1	11.3
2.203	1.680	19	0	41	0	10.9
2.535	1.820	14	0	36	1	10.5
1.801	1.750	0	0	43	1	11.7
1.068	1.900	30	1	38	1	11.0
2.972	1.880	11	0	38	1	10.8
1.545	1.600	0	1	40	1	10.9
2.141	2.000	25	0	42	0	10.8
1.670	1.850	11	0	43	0	11.2
1.236	2.050	19	1	48	0	11.7
1.912	2.080	14	0	47	0	11.5
1.825	2.140	25	1	39	0	10.7
1.988	2.150	0	0	50	0	12.0
788	2.200	22	0	45	1	11.4
400	2.310	11	0	33	0	9.7
2.072	2.420	19	1	45	0	11.6
2.644	2.480	11	0	38	0	10.4
2.786	2.130	19	1	34	0	9.8
2.704	2.500	24	0	34	1	9.7
3.073	2.300	19	0	33	0	10.2
2.263	2.750	19	1	36	1	9.9
4.075	3.000	11	0	32	0	9.7
1.665	3.100	24	0	41	1	11.1
3.480	3.400	11	1	38	0	10.5

Y = عدد الكيلو وات ساعة/شهر

X_1 = مربع المساحة المراد تدفئتها

X_2 = قيمة R التي توضح قوة مواد العزل

X_3 = وجود/أو عدم وجود نوافذ العزل

X_4 = متوسط درجة الحرارة

X_5 = استخدام آلة توليد/قوة تيار كهربائي

X_6 = متوسط الساعات المشمسة في اليوم

إستخدام نموذج الإنحدار الخطى المتعدد :

مبدئياً يفترض المدير النموذج التالى :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

يمثل العلاقة السليمة بين المتغير التابع Y والمتغيرات المفسرة من $X_1 \leftarrow X_6$

توفيق النموذج لبيانات العينة وتقييم معاملات المربعات الصغرى :

يوضح جدول (١٠-١٧) نتائج البرنامج SAS ومن هذه المخرجات أو النتائج تكون معادلة المربعات الصغرى :

$$\hat{Y} = 2268.7 + .6372X_1 - 21.898X_2 - 192.889X_3 - 111.5655X_4 - 437.0679X_5 + 318.6621X_6$$

وإشارات معاملات المربعات الصغرى للمتغيرات X_1 , X_2 , X_3 , X_4 , X_5 تتفق مع توقعات المدير. بمعنى أن المدير يتوقع علاقة موجبة بين X_1 , Y وسالبة بين Y , X_2, X_3 , X_4 , X_5 . والإشارة الموجبة لمعامل X_6 غير متوقعة. وهذا يدل على وجود نتيجة غير حقيقية أو غير منطقية وهى زيادة إستهلاك الكهرباء المستعملة فى التدفئة إذا زاد متوسط عدد الساعات المشمسة مع بقاء باقى المتغيرات الأخرى ثابتة .

جدول (١٠-١٧)

مخرجات SAS المبدئية للمثال الشامل

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	7241298.64550659	1206883.10758443	2.15	0.0976
Error	18	10113935.35449340	561885.29747186		
Corrected Total	24	17355234.00000000			

R-Square	C.V.	Root MSE	Y Mean
0.417240	35.82099	749.59008630	2092.60000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	2843552.22280672	2843552.22280672	5.06	0.0372
X2	1	907324.24266437	907324.24266437	1.61	0.2200
X3	1	328823.31140619	328823.31140619	0.59	0.4542
X4	1	2209824.99016773	2209824.99016773	3.93	0.0628
X5	1	847688.07049603	847688.07049603	1.51	0.2352
X6	1	104085.80796555	104085.80796555	0.19	0.6720

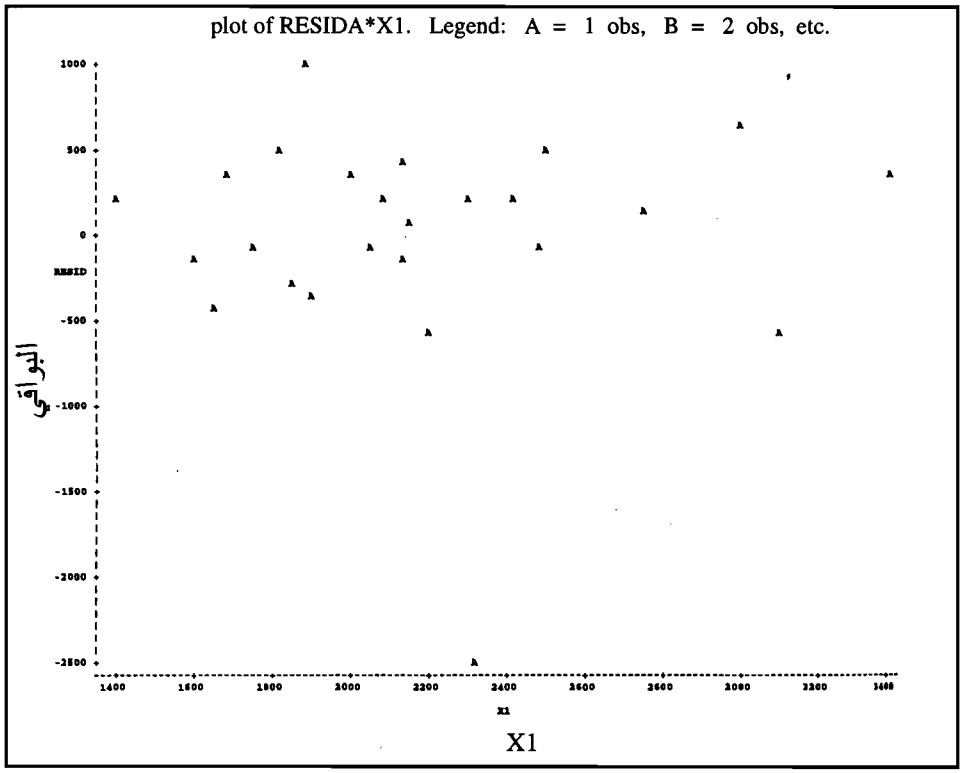
Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	1762790.89204180	1762790.89204180	3.14	0.0935
X2	1	707676.33222952	707676.33222952	1.26	0.2765
X3	1	203002.51041771	203002.51041771	0.36	0.5553
X4	1	695187.69761528	695187.69761528	1.24	0.2806
X5	1	950735.59502893	950735.59502893	1.69	0.2097
X6	1	104085.80796555	104085.80796555	0.19	0.6720

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	2268.695657	0.47	0.6412	4786.265029
X1	0.637212	1.77	0.0935	0.359756
X2	-21.897991	-1.12	0.2765	19.512404
X3	-192.889342	-0.60	0.5553	320.908604
X4	-111.565499	-1.11	0.2806	100.300410
X5	-437.067909	-1.30	0.2097	336.002785
X6	318.662114	0.43	0.6720	740.386528

تقييم معادلة المربعات الصغرى

من المعلومات الموضحة فى جدول (١٠-١٧) . تكون معادلة المربعات الصغرى غير مؤثرة. قيمة P للفرض العدمى (H_0: $\beta_0 = \beta_1 \ldots = \beta_6 = 0$) هى 0976. والتى تشير إلى أن الدليل مقابل الفرض العدمي هذا غير ملائم أو غير مقنع. بالطبع هذا يعنى أنه لا يوجد أى من المتغيرات المفسرة الستة يساعد فى شرح الإختلاف فى قيم Y . ونصل إلى نفس الإستنتاج بواسطة فحص قيم P المناظرة للإحصاء T أو الإحصاء F الهامشى (Type III SS) . فى الحقيقة المتغير المفسر الوحيد التى يبدو مساعداً وهو X_1 عندما يكون بمفرده فى النموذج (P-value = .0372) .

ولإكتشاف ما هو الخطأ، يرسم المدير البواقى فى مقابل كل متغير مفسر ويكتشف أن بيانات العينة تحتوى على مشاهدات غير مألوفة. والمشاهدات غير المعتادة أو غير المألوفة هذه تكون واضحة عند فحص الشكل البيانى للبواقى مع X_1 (مربع مساحة المكان المراد تدفئته) ، وتتضح في شكل (١٠-٨)، ونجد ذلك فى أسفل الشكل. وتظهر بيانات العينة أن منزلاً واحداً يستخدم فقط 40 كيلووات فى الساعة من الكهرباء. وجد المدير أن العائلة بعيدة عن المنزل لمدة الشهر بأكمله. ولأن المدير يعتقد أن هذه المشاهدات لا تمثل بيئة الدراسة فيستبعدها المدير ويتم توفيق النموذج على أن (n = 24) مشاهمة. وتكون مخرجات SAS الجديدة كما هو موضح(١٠-١٨).



شكل (١٠-٨)
البواقي (مبدائياً) مقابل X_1 في المثال الشامل

جدول (١٠-١٨)
مخرجات SAS المنقحة للمثال الشامل
General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	12226767.88076240	2037794.64679375	16.16	0.0001
Error	17	2144200.74423751	126129.45554338		
Corrected Total	23	14370968.62500000			

R-Square	C.V.	Root MSE	Y Mean
0.850796	16.41824	355.14709001	2163.12500000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	3180503.76444430	3180503.76444430	25.22	0.0001
X2	1	1355656.00475436	1355656.00475436	10.75	0.0044
X3	1	727758.15719231	727758.15719231	5.77	0.0280
X4	1	4814280.29319091	4814280.29319091	38.17	0.0001
X5	1	2126706.26423271	2126706.26423271	16.86	0.0007
X6	1	21863.39694791	21863.39694791	0.17	0.6824

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	1247739.29877925	1247739.29877925	9.89	0.0059
X2	1	1476207.38278585	1476207.38278585	11.70	0.0033
X3	1	566435.02679629	566435.02679629	4.49	0.0491
X4	1	506795.34905793	506795.34905793	4.02	0.0612
X5	1	1758155.66446358	1758155.66446358	13.94	0.0017
X6	1	21863.39694791	21863.39694791	0.17	0.6824

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	7268.399177	3.09	0.0067	2353.287660
X1	0.537548	3.15	0.0059	0.170909
X2	-31.919953	-3.42	0.0033	9.330326
X3	-324.098428	-2.12	0.0491	152.936144
X4	-95.344409	-2.00	0.0612	47.564965
X5	-599.216956	-3.73	0.0017	160.495808
X6	-148.079139	-0.42	0.6824	355.666795

والآن ، يبدو أن معادلة المربعات الصغرى مشجعة ومبشرة. فالإشارات لمعاملات الانحدار كما هي متوقعة. وإختبار الفرض ($H_0: \beta_1 = \beta_2 = \ldots \beta_6 = 0$) يتضح تناقضه (P-value = .0001) . وتكون الآثار المضافة للمتغيرات X_1 , X_2 , X_5 في وجود كل المتغيرات التفسيرية الأخرى مفيدة في شرح الإختلاف في قيمة Y (وتكون قيمة P لكل من T، F الهامشية هي .0059 , .0033 , .0017 على التوالي) ، الأثر الإضافي للمتغير X_6 في وجود كل المتغيرات يمكن إهماله، (P-value = .6824). والمتغيران X_3 , X_4 يكونا في المنطقة الرمادية وتكون قيمة P-value عبارة عن .0491، .06126 على التوالي. وربما يوجد هناك بعض الأزدواج الخطي يشمل X_4 حيث أن التأثير الإضافي للمتغير X_4 في وجود X_1 , X_2 , X_3 يكون مفيداً للغاية حيث أن (P-value = .0001). ولكن تأثير X_4 في وجود كل المتغيرات التفسيرية يكون غير مقنع (P-value=.0612). ويعتقد المدير أن إختيار المتغير ربما يساعد في حل الموقف الخاص بالمتغيرات X_3 , X_4 , X_6 فيقرر المدير إستخدام الحذف الخلفي ويوضح جدول (١٠-١٩) مخرجات البرنامج SAS . من معلومات هذا الجدول لاحظ أن التحسين قد حدث. والمتغير X_6 تم حذفه من النموذج. الأخطاء المعيارية لمعاملات المربعات الصغرى أصغر من ذي قبل، تباين البواقي انخفض (S_e^2= 120,336.9) عن القيمة السابقة: S_e^2= 126129.46 . الأثر الإضافي للمتغير X_4 في وجود كل المتغيرات الأخرى مفيد تماماً. ويظهر بعض الإرتباط الخطي بين X_4 , X_6، لهذا فإن ظهور X_6 في النموذج يزيل الأثر الإضافي للمتغير X_4 . والآن بما أنه تم حذف X_6 ، فقد وضحت أهمية X_4 . إن الأزدواج الخطي بين X_4 ، X_6 يمكن أيضاً أن يشاهد بواسطة عملية الإختيار الأمامي المعدلة المعطاة في جدول (١٠-٢٠) . لاحظ أنه بإستخدام هذا الأسلوب، فإن

X_6 هو أول متغير يدخل فى معادلة الإنحدار، لكنه يحذف فوراً عند دخول X_4 للمعادلة. على الرغم من عدم ظهور الرسم البيانى للبواقى مع المتغيرات التفسيرية، من X_1 إلى X_5 لأنها لاتكشف عن أية عيوب وتكون معادلة الإنحدار النهائية كما يلى :

$$\hat{Y}= 6356.174 + .5604X_1 - 31.2077X_2 - 327.503X_3 - 113.8952X_4 - 621.4582X_5$$

جدول (١٠-١٩)

مخرجات SAS للمثال الشامل بطريقة الحذف الخلفي

Backward Elimination Procedure for Dependent Variable Y

Step 0 All Variables Entered R-square = 0.85079637 C(p) = 7.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	12226767.880762	2037794.6467937	16.16	0.0001
Error	17	2144200.7442375	126129.45554338		
Total	23	14370968.625000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7268.39917674	2353.28766000	1203217.1876926	9.54	0.0067
X1	0.53754815	0.17090852	1247739.2987793	9.89	0.0059
X2	-31.91995251	9.33032584	1476207.3827858	11.70	0.0033
X3	-324.09842825	152.93614357	566435.02679629	4.49	0.0491
X4	-95.34440947	47.56496521	506795.34905793	4.02	0.0612
X5	-599.21695561	160.49580751	1758155.6644636	13.94	0.0017
X6	-148.07913902	355.66679485	21863.39694791	0.17	0.6824

Bounds on condition number: 10.06899, 145.1055

Step 1 Variable X6 Removed R-square = 0.84927501 C(p) = 5.17334093

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	12204904.483815	2440980.8967629	20.28	0.0001
Error	18	2166064.1411854	120336.89673252		
Total	23	14370968.625000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	6356.17398265	838.70144171	6911552.0204922	57.44	0.0001
X1	0.56037727	0.15811300	1511557.4703703	12.56	0.0023
X2	-31.20767520	8.95904711	1460151.7793736	12.13	0.0027
X3	-327.50300255	149.16934379	580056.39758100	4.82	0.0415
X4	-113.89524228	16.26040224	5904014.3421837	49.06	0.0001
X5	-621.45823581	147.82830473	2126706.2642327	17.67	0.0005

Bounds on condition number: 1.190054, 27.9493

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable Y

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X6	5	0.0015	0.8493	5.1733	0.1733	0.6824

جدول (١٠-٢٠)

مخرجات SAS : الانحدار المتدرج للمثال الشامل

Stepwise Procedure for Dependent Variable Y

Step 1 Variable X6 Entered R-square = 0.46500207 C(p) = 40.95672402

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	6682530.2130470	6682530.2130470	19.12	0.0002
Error	22	7688438.4119530	349474.47327059		
Total	23	14370968.625000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	11011.74505084	2027.13955118	10312416.862195	29.51	0.0001
X6	-815.85432662	186.57346835	6682530.2130470	19.12	0.0002

تابع : جدول (١٠-٢٠)

```
Bounds on condition number:            1,            1
-----------------------------------------------------------------------------------------------------------

Step 2   Variable X5 Entered        R-square = 0.59936476   C(p) = 27.64767530

                                        DF         Sum of Squares      Mean Square          F   Prob>F

                     Regression          2        8613452.1921840  4306726.0960920      15.71   0.0001
                     Error              21        5757516.4328160  274167.44918171
                     Total              23       14370968.625000

                                     Parameter         Standard          Type II
                     Variable         Estimate            Error   Sum of Squares          F   Prob>F

                     INTERCEP   11349.95309408   1800.01070306  10900686.507193      39.76   0.0001
                     X5          -575.46238274    216.84155965  1930921.9791370       7.04   0.0149
                     X6          -824.92989024    165.28866608  6829105.4957554      24.91   0.0001

Bounds on condition number:     1.000428,     4.001713
-----------------------------------------------------------------------------------------------------------

Step 3   Variable X2 Entered        R-square = 0.71543302   C(p) = 16.42306147

                                        DF         Sum of Squares      Mean Square          F   Prob>F

                     Regression          3       10281465.534196  3427155.1780655      16.76   0.0001
                     Error              20        4089503.0908036  204475.15454018
                     Total              23       14370968.625000

                                     Parameter         Standard          Type II
                     Variable         Estimate            Error   Sum of Squares          F   Prob>F

                     INTERCEP   13092.09145440   1669.87654601  12568685.985735      61.47   0.0001
                     X2            -32.36554326     11.33191775  1668013.3420124       8.16   0.0098
                     X5           -546.27893444    187.54279075  1734877.0571193       8.48   0.0086
                     X6           -942.16496764    148.52760338  8227731.3376228      40.24   0.0001

Bounds on condition number:     1.086352,     9.518733
-----------------------------------------------------------------------------------------------------------

Step 4   Variable X1 Entered        R-square = 0.77915256   C(p) = 11.16296945

                                        DF         Sum of Squares      Mean Square          F   Prob>F

                     Regression          4       11197176.987919  2799294.2469798      16.76   0.0001
                     Error              19        3173791.6370808  167041.66510951
                     Total              23       14370968.625000

                                     Parameter         Standard          Type II
                     Variable         Estimate            Error   Sum of Squares          F   Prob>F

                     INTERCEP   10907.96262573   1774.31772477  6313196.5562753      37.79   0.0001
                     X1             0.44928090      0.19188942  915711.45372280       5.48   0.0303
                     X2            -37.61269324     10.48456915  2149772.4741806      12.87   0.0020
                     X5           -490.70026816    171.16294115  1372896.8865486       8.22   0.0099
                     X6           -825.85362012    143.14195988  5560284.9134963      33.29   0.0001

Bounds on condition number:     1.262722,     18.62257
-----------------------------------------------------------------------------------------------------------

Step 5   Variable X3 Entered        R-square = 0.81553115   C(p) =  9.01805706

                                        DF         Sum of Squares      Mean Square          F   Prob>F

                     Regression          5       11719972.531705  2343994.5063409      15.92   0.0001
                     Error              18        2650996.0932954  147277.56073864
                     Total              23       14370968.625000

                                     Parameter         Standard          Type II
                     Variable         Estimate            Error   Sum of Squares          F   Prob>F

                     INTERCEP   10831.35915199   1666.54315151  6221131.3280929      42.24   0.0001
                     X1             0.46372804      0.18034323  973786.13561346       6.61   0.0192
                     X2            -34.31277902      9.99937887  1734212.5380407      11.78   0.0030
                     X3           -311.08282670    165.11190545  522795.54378534       3.55   0.0758
                     X5           -478.88165385    160.84078029  1305571.4326521       8.86   0.0081
                     X6           -815.92623151    134.51053808  5419082.3900737      36.80   0.0001

Bounds on condition number:     1.265009,     28.69316
-----------------------------------------------------------------------------------------------------------
```

تابع : جدول (١٠-٢٠)

Step 6 Variable X4 Entered R-square = 0.85079637 C(p) = 7.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	12226767.880762	2037794.6467937	16.16	0.0001
Error	17	2144200.7442375	126129.45554338		
Total	23	14370968.625000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7268.39917674	2353.28766000	1203217.1876926	9.54	0.0067
X1	0.53754815	0.17090852	1247739.2987792	9.89	0.0059
X2	-31.91995251	9.33032584	1476207.3827858	11.70	0.0033
X3	-324.09842825	152.93614357	566435.02679629	4.49	0.0491
X4	-95.34440947	47.56496521	506795.34905792	4.02	0.0612
X5	-599.21695561	160.49580751	1758155.6644636	13.94	0.0017
X6	-148.07913902	355.66679485	21863.39694791	0.17	0.6824

Bounds on condition number: 10.06899, 145.1055

Step 7 Variable X6 Removed R-square = 0.84927501 C(p) = 5.17334093

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	12204904.483815	2440980.8967629	20.28	0.0001
Error	18	2166064.1411854	120336.89673252		
Total	23	14370968.625000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	6356.17398265	838.70144171	6911552.0204922	57.44	0.0001
X1	0.56037727	0.15811300	1511557.4703703	12.56	0.0023
X2	-31.20767520	8.95904711	1460151.7793736	12.13	0.0027
X3	-327.50300255	149.16934379	580056.39758100	4.82	0.0415
X4	-113.89524228	16.26040224	5904014.3421837	49.06	0.0001
X5	-621.45823581	147.82830473	2126706.2642327	17.67	0.0005

Bounds on condition number: 1.190054, 27.9493

All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Entered	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X6		1	0.4650	0.4650	40.9567	19.1217	0.0002
2	X5		2	0.1344	0.5994	27.6477	7.0429	0.0149
3	X2		3	0.1161	0.7154	16.4231	8.1575	0.0098
4	X1		4	0.0637	0.7792	11.1630	5.4819	0.0303
5	X3		5	0.0364	0.8155	9.0181	3.5497	0.0758
6	X4		6	0.0353	0.8508	7.0000	4.0181	0.0612
7		X6	5	0.0015	0.8493	5.1733	0.1733	0.6824

- الملخص (١٠-١٠) Summary

فى هذا الفصل تم التوسع فى الطرق الإحصائية والتى سبق ذكرها فى الفصل التاسع بإستخدام أكثر من متغير واحد فى نموذج الإنحدار .

فى نموذج الإنحدار الخطى المتعدد، يمكن أن يتضمن النموذج متغيرات تفسيرية تكون وصفية أو تكون فى شكل علاقة غير خطية. ويكون مفتاح الأسئلة والتى تحتاج إلى إجابة هى نفس الأسئلة المطلوب الإجابة عنها فى الفصل التاسع. مع ذلك، فإن وجود العديد من المتغيرات التفسيرية فى النموذج يمكن أن يعقد من الإستنتاجات الإحصائية إلى حد ما، وهذا إلى حد كبير يسبب الإرتباط بين المتغيرات التفسيرية. والإرتباط بين متغيرين أو مجموعة من المتغيرات يسمى الإرتباط الخطى بين المتغيرات التفسيرية. والتحليل الإنحدارى المتعدد، يتعرض لإنتقاضات كبيرة للنموذج مثل الازدواج الخطى وإنشاء معيار لاختيار أفضل مجموعة متغيرات مفسرة لإستخدامها فى نموذج الإنحدار .

المراجع References

1. N. Draper and H. Smith. *Applied Regression Analysis,* 2nd ed. NeW York: Wiley, 1981.
2. W. Mendenhall and T. Sincich. *A Second Course in Business Statistics: Regression analysis*, 4th ed. San Francisco: Dellen, 1993.
3. R. B. Miller and D. W. Wichern. *Intermediate Business Statistics: Analysis of Variance, Regression, and Time Series.* New York: Holt, Rinehart, and Winstion, 1977.
4. J. Neter, W. Wasserman, and M. Kutner. *Applied Linear Statistical Models,* 2nd ed. Homewood, IL: Richard D. Irwin, 1985.
5. M. Younger. *A Handbook for Linear Regression,* nd ed. Boston: Duxbury Press, 1985.

تمارين إضافية

(١٠-٤٤) مدير أفراد يريد تكوين نموذج للمتغير (Y) والذي يعبر عن الرضاء عن وظيفة المدير وتأخذ الأرقام (من 1 : 10) كدالة فى , (X_1) العمر (X_2) معدل الأداء (من 1 إلى 5) ، (X_3) تمثل المرتب (بالألف دولار فى الشهر) ، (X_4) الوقت الذى تستغرقه الوظيفة (بالسنوات) . وكانت معادلة الإنحدار المعدلة كما يلى :

$$\hat{Y} = 3 + .05X_1 + 1.1X_2 - 6X_3 + .11X_4$$

أ - ما هى درجة الرضاء المتنبأ بها بواسطة النموذج لموظف عمره 40 سنة والذى معدلة 4 فى تقييم الأداء ويكسب 3950 دولار فى الشهر وخبرته الحالية 3 سنوات ؟

ب- هل المعاملات المقدرة صحيحة ؟ حدد لكل واحد من هذه المعاملات ما إذا كانت إشارته ملائمة أم لا . إذا بدا أن أحد هذه الاشارات خاطئة، هل يمكنك اقتراح السبب ؟

(١٠-٤٥)مدير برنامج دراسات عليا بأحد كليات التجارة يريد الحصول على أفضل تنبؤ للأداء فى برنامج MBA للكلية. إستخدم عينة مكونة من 25 طالب فى برنامج MBA وكون نموذج الإنحدار المتعلق بـ(GPA = Y) للطالب (تحول إلى رقم 4 كحد أقصى) ، ($GPA = X_1$)

لطالب جامعى (يحول إلى رقم حده الأقصى 4) ، X_2 = درجة الطالب فى GMAT. معادلة المربعات الصغرى هي :

$$\hat{Y} = -.25 + .5X_1 + .0085X_2 - .0000081X_2^2$$

أ – العينة المكونة من 25 طالب لهم مدى GPAs من 2.8 إلى 3.85 ودرجات GMAT مداها من 475 إلى 650. الآن يوجد لدى المدير متقدمين من الطلبة الجامعيين يتعدى GPA لـ 3.71 ودرجة GMAT لـ750 . هل نستخدم النموذج فى التنبؤ بقيمة GPA , MBA لهذا المتقدم؟ اشرح مدعماً إجابتك .

ب– حاول أن تستخدم النموذج فى التنبؤ بأداء المتقدمين فى الجزء (أ) وبالمتقدمين الآخريين الذين يكونوا طلبة GPA بتقدير 3.6 ودرجة الـ GMAT 650 . بماذا تخبرك هذه النتائج عن العلاقة بين Y , X_2 فى النموذج ؟

فى التمارين التالية سوف يكون هدفك النهائى هو تحديد ما إذا كان هناك علاقة بين المتغير التابع والمتغيرات المستقلة. إذا كانت هناك علاقة يكون المطلوب منك هو تقييم نموذج الانحدار بالكامل بهدف الحصول على أفضل معادلة إنحدار لإستخدامها فى التنبؤ والتقدير .

(١٠-٤٦) قام فريق من المحللين في مستشفى بتحليل درجة توقف طول فترة اقامة المريض بالمستشفى على عمر المريض، أيضاً فإنه لوحظ أن الرجل يقيم في المستشفى لمدة أطول من المرأة.

Women		Men	
Age	Length of stay (days)	Age	Length of stay (days)
63	3	66	9
66	16	70	8
67	6	72	10
68	9	77	17
68	3	77	18
69	4	78	12
69	8	78	9
69	19		
70	9		
70	6		
72	7		
73	10		
74	7		
83	16		
84	21		
85	8		
88	10		

(١٠-٤٧) شركة XRX تدير برنامج أبحاث مستمر لتحسين جودة الطباعة لطابعتها ومن المهم تحديد خصائص مساهمة الطباعة على إحساس العملاء بجودة الطباعة. وفى مسح أو استقصاء، قدر العملاء جودة الطباعة لعينة من المطبوعات. كل مطبوع حلل فى مختبر وكانت المعدلات من 0 إلى 10 لـ grceyness (X_1) الخلفية، X_2 = عدد البقع في الخلفية ، X_3 = حدة التصوير، X_4 = تحبير التصوير. فإذا كانت بيانات العينة كما يلي حدد أفضل معادلة إنحدار .

Y	X_1	X_2	X_3	X_4	Y	X_1	X_2	X_3	X_4
6	10	8	5	4	6	6	4	7	5
2	8	1	2	5	4	8	4	3	5
6	5	9	5	4	4	5	5	3	3
6	4	9	4	5	6	3	7	5	6
8	0	6	9	6	3	7	0	3	5
7	10	5	5	8	5	1	2	6	5
5	5	4	4	6	7	4	6	8	5
5	7	2	6	6	4	3	4	5	4
8	5	7	9	6	4	5	1	7	2
5	1	6	4	4	4	1	2	2	6

(١٠-٤٨) فى تمرين (٩-٦٨) طلب منك تحليل غياب الموظفين فى كل من مصنعين Richmond, Lousville . افترض أنه بالإضافة إلى سنوات الخدمة تم أخذ نوع الجنس للموظفين في الاعتبار. حدد للبيانات التالية معادلة الإنحدار الملائمة .

Richmond			Louisville		
Absences	Year of Service	Gender	Absences	Year of Service	Gender
18	9	M	24	9	M
14	9	M	0	31	M
24	21	M	18	17	M
5	15	F	28	19	M
7	15	F	30	19	F
0	15	M	51	16	M
8	17	F	48	24	M
13	21	M	3	30	F
2	18	F	14	17	M
0	16	M	19	19	F
5	15	M	50	7	M
11	9	M	9	13	M
10	14	F	13	7	M
1	8	M	9	7	M
7	10	F	4	20	M
2	15	M	50	16	F
21	9	M	49	9	M
18	12	M	15	12	M
2	15	M	11	11	M
12	16	F	15	18	F
9	14	F	9	7	F
5	14	M	13	24	M
9	14	F	5	18	F
13	8	M	6	12	M
16	8	M	33	7	F
11	15	F	64	7	M
9	13	M	12	16	F
23	10	F	8	14	M
5	8	M	0	15	M
13	14	M	3	17	M

(١٠-٤٩) قام محلل فى شركة تليفون بدراسة العلاقة بين زيادة خط التليفون السنوى فى منطقة ما والحالة السنوية للتوظيف فى مصانع البناء. وتكون الزيادة ممثلة في التغير فى عدد التليفونات فى المكان من سنة إلى أخرى. ويعتقد المحلل بأن الرقم القياسي للأسعار (CPI) ربما يرشد إلى الزيادة السنوية. فاذا تم تسجيل الزيادة السنوية للعشرة سنوات الأخيرة مع مؤشر الحالة السنوية الكلية للتوظف فى مصنع البناء ومؤشر الرقم القياسي للأسعار. حدد أفضل معادلة إنحدار .

CPI	Construction Employment	Area Gain	CPI	Construction Employment	Area Gain
195.4	130.2	446	298.4	113.9	688
217.4	138.4	591	.311.1	132.8	667
264.8	128.3	569	322.2	152.0	757
272.4	116.3	490	328.4	168.1	899
289.1	103.8	262	340.4	173.6	741

(١٠-٥٠) فى العديد من الشركات . تعتبر مشكلة تحديد أهم عوامل التنبؤ بكفاءة العمل للمستخدمين الحاليين عملية مستمرة، والإجراء الدائم يكون بعمل إختبارات ملائمة، وإتخاذ قرار التعيين الملائم يعتمد على درجات الاختبار. والسؤال المبدئي هو معرفة أي الاختبارات يساعد في التنبؤ بأداء الأفراد. افترض أن مكتب الأفراد فى إحدى الشركات الكبيرة أجرى أربعة إختبارات لوظائف معينة. هذه الإختبارات أديت لـ 20 فرد تم تعينهم بواسطة الشركة وبعد سنتين تم وضع درجات لكل واحد من الموظفين بحسب كفاءته فى الوظيفة. ودرجات كفاءة الوظيفة ودرجات الإختبارات الأربعة مسجلة تم تسجيلها كما يلى : حدد معادلة الإنحدار الملائمة .

Employee	Score	Test 1	Test 2	Test 3	Test 4
1	94	122	121	96	89
2	71	108	115	98	78
3	82	120	115	95	90
4	76	118	117	93	95
5	111	113	102	109	109
6	64	112	96	90	88
7	109	109	129	102	108
8	104	112	119	106	105
9	80	115	101	95	88
10	73	111	95	95	84
11	127	119	118	107	110
12	88	112	110	100	87
13	99	120	89	105	97
14	80	117	108	99	100
15	99	109	125	108	95
16	116	116	122	116	102
17	110	104	83	100	102
18	96	110	101	103	103
19	126	117	120	113	108
20	58	120	77	80	74

(١٠-٥١) كما في مثال (١٠-٨) تمثل البيانات التالية متوسط درجات الحرارة فى يناير لـ 26 محطة رصد فى Virginia . كل محطة رصد حددت بناء على خط العرض ومستوى سطح البحر. وبتجميع الـ 26 محطة مع الـ 24 فى المثال (١٠-٨) . حدد نموذج الإنحدار الملائم.

Temperature	Latitude	Longitude	Elevation
39.3	36.58	79.38	410
36.1	38.03	78.00	420
33.9	38.67	78.38	1,200
36.6	37.33	79.20	916
37.1	36.70	79.88	760
28.6	38.42	79.58	2,910

29.3	39.07	77.88	1,720
37.4	37.70	78.30	300
40.5	36.90	76.20	22
38.9	37.58	75.82	300
34.4	36.75	83.03	1,510
35.3	38.50	77.32	12
37.5	37.50	77.33	164
36.4	37.32	79.97	1,149
35.0	36.88	81.77	1,375
34.0	38.15	79.03	1,385
33.3	38.65	78.72	1,000
38.6	37.65	76.57	25
37.5	37.75	77.05	50
36.2	37.85	75.48	9
32.1	38.95	77.45	291
35.6	38.85	77.03	10
39.3	37.30	76.70	70
33.7	37.20	78.17	760
34.4	38.88	78.52	887
34.4	36.93	81.08	2,450

(١٠-٥٢) يدرس أحد طلاب الدراسات العليا أسعار كتب إدارة الأعمال فى الفصل الدراسي لربيع 1992، ولقد حدد الطالب المتغيرات التفسيرية التالية وهي عدد الصفحات ، نوع غطاء الكتاب (غلاف مقوى - غلاف خفيف) وتخصص الادارة (اقتصاد - محاسبة - إدارة، نظم المعلومات الادارية وغيرها) وكانت البيانات المجمعة كما يلى :

Price	Pages	Cover	major	Price	Pages	Cover	major
$20.65	486	S	Acc	**$51.65**	668	H	Mgt
18.95	522	S	Acc	50.00	440	H	Mgt
52.50	826	H	Acc	48.95	888	H	Mgt
55.65	810	H	Acc	50.65	690	H	Mgt
64.95	1,336	H	Acc	45.95	1,011	H	Mgt
56.25	857	H	Acc	30.00	507	H	Mgt
16.95	417	S	Econ	45.00	814	H	Mgt
13.95	207	S	Econ	19.95	182	S	MIS
35.15	460	S	Econ	29.95	731	S	MIS
48.75	826	H	Econ	44.95	866	S	MIS
42.95	828	H	Econ	56.25	826	H	MIS
48.75	644	H	Econ	44.95	797	H	MIS
50.90	986	H	Other	48.75	308	H	MIS
18.30	558	S	Other	37.95	585	H	MIS
46.90	1,398	H	Other	27.20	340	S	Other
25.95	518	S	Mgt	48.75	792	H	Other
50.00	1,070	H	Mgt	45.00	829	H	Other
48.50	586	H	Mgt	50.00	1,100	H	Other
47.95	732	H	Mgt	54.95	1,181	H	Other
47.85	679	H	Mgt	15.00	354	S	Other

(١٠-٥٣) فى دراسة حديثة تم دراسة تأثير العوامل المؤثرة على عدد الشكاوى عن العلاج طويل المدى لكبيرى السن في ولاية فريجينيا. ومن العوامل المحتملة كيفية التعامل مع الشكاوى وحلها. ففى الوقت الحالى يتم التعامل مع الشكاوى إما عن طريق برامج محلية أو بواسطة الولاية. كما أن هناك العديد من العوامل الإضافية والتى يمكن أن تؤثر على عدد الشكاوى وهى عدد الأسرة المتاحة فى المدى الطويل، سهولة وإمكانية الرعاية ومكان تقديم الرعاية «الخدمة» (ريف - حضر - . . .) والبيانات التالية قائمة على الشكاوى المستقصاة في خلال 1990 .

Area	Complaints	Number of beds	Location	Program
1	36	412	Rural	Local
2	22	280	Rural	Local
3	211	989	Rural	Local
4	5	650	Rural	State
5	77	1,789	Urban	Local
6	1	1,259	Rural	State
7	15	820	Rural	State
8	176	3,388	Mixed	Local
9	13	582	Rural	State
10	64	800	Urban	Local
11	28	648	Rural	State
12	3	1,364	Rural	State
13	3	494	Rural	State
14	0	475	Rural	State
15	273	3,117	Mixed	Local
16	14	698	Urban	State
17	8	801	Rural	State
18	4	810	Mixed	State
19	17	3,292	Mixed	State
20	5	356	Mixed	State

(١٠-٥٤) إهتمت احدى الدراسات الحديثة بتقدير تكاليف التصنيع الشهرية لشركة صناعية ، وتم اعتبار أربعة كميات لمتغيرات مفسرة :

١- قيمة الإنتاج المصنع والذى يمثل تجميع تكلفة كل الوظائف فى الشهر (X_1) .

٢-صافى المبيعات الشهرية (X_2) .

٣-تكلفة جدول الرواتب (X_3) .

٤- نسبة أجر الموظفين فى الساعة والتى تطبق مباشرة على وظيفة العميل (X_4) وقد جمعت البيانات التالية معتمدة على 18 شهر .

Month	Actual Manufacturing Expenses	X_1	X_2	X_3	X_4
1	765,715	736,932	1,070,857	431,313	70.9%
2	866,646	868,731	1,166,572	488,672	76.0
3	795,762	768,923	1,084,584	466,110	75.4
4	880,175	802,044	1,152,015	471,759	73.7
5	840,308	768,753	1,102,914	466,955	74.5

6	813,588	738,084	1,209,486	442,674	75.5
7	215,828	830,697	1,093,130	518,724	72.7
8	844,200	772,550	1,178,449	473,927	72.6
9	939,497	865,235	1,246,947	540,264	73.3
10	857,316	778,727	1,026,848	478,457	72.5
11	867,235	835,969	1,283,029	489,242	74.4
12	871,289	814,294	1,421,233	475,405	75.8
13	875,872	862,820	1,006,106	487,189	75.4
14	945,866	**1,005,529**	1,258,522	544,505	75.5
15	875,960	887.690,	1,177,637	489,028	77.0
16	1,014,107	**1,031,243**	1,252,445	578,567	75.3
17	939,557	978,892	1,227,701	525,516	76.9
18	850,136	816,093	1,221,124	478,830	75.6

(١٠-٥٥) هذا التمرين يعتبر تعميم لتمرين (٧-٦٤) حيث تم دراسة فائدة حضور المحاضرة قبل أخذهم الإمتحان. كان هناك اعتقاد بوجود عامل آخر لأداء الطلبة فى الإمتحان. بجانب المحاضرة وهو كفاءة الطالب. وقد استُخدم متوسط درجات الكلية (GPAs) كمعيار لهذه الكفاءة. فى النهاية كان تصنيف مستوى الطلبة (أولى، ثانية، . . . الخ). يمكن أن يشرح بعض الإختلافات المشاهدة فى درجات الإمتحان وكانت البيانات كما يلى :

Grade:	75	88	75	95	95	75	75	75	85	65	75	95
Attend?	1	1	0	1	0	0	0	1	0	0	0	0
GPA:	3.02	3.2	2.2	2.9	4.0	2.4	2.0	2.5	1.5	2.5	2.29	3.0
Class:	1	1	1	1	1	2	1	2	2	2	2	2

Grade:	85	88	72	82	82	82	85	92	98	78	85	78
Attend?	1	1	0	0	0	0	0	0	1	1	1	0
GPA:	2.75	2.2	2.1	3.9	2.3	2.0	2.27	2.75	3.23	3.16	2.15	2.5
Class:	2	2	2	2	2	2	2	2	2	2	2	2

Grade:	95	95	85	75	72	78	65	75	72	95	98	88
Attend?	1	1	0	0	0	0	0	0	0	1	1	1
GPA:	3.62	3.56	2.0	2.4	2.72	2.8	2.6	2.4	1.93	3.0	3.5	1.5
Class:	2	2	2	2	2	2	2	2	2	2	2	2

Grade:	75	82	85	78	88	92	65	78	88	82	88	78
Attend?	1	0	0	1	1	1	0	0	1	1	0	0
GPA:	2.5	3.2	2.1	2.8	3.2	2.1	2.2	2.75	3.0	3.0	2.0	3.4
Class:	2	2	2	2	2	2	2	2	2	2	2	2

Grade:	75	75	85	95	92	75	95	75	88	85	82	95
Attend?	0	0	0	1	0	0	1	0	1	1	0	0
GPA:	2.9	2.75	2.01	2.82	2.7	3.4	3.7	2.8	2.8	2.7	1.9	2.35
Class:	2	2	2	2	2	2	2	2	2	2	2	2

Grade:	85	95	65	82	75	98	65	98	75	85	85	78
Attend?	0	0	0	0	0	0	0	1	0	0	0	1
GPA:	3.3	3.4	2.7	2.1	1.96	2.6	2.2	3.65	2.6	2.7	2.9	2.4
Class:	2	3	3	3	3	3	3	3	3	3	3	3

Grade:	85	85	75	88	95	92	95	92	88	78	72	75
Attend?	1	0	1	0	1	0	1	1	0	0	0	1
GPA:	2.2	3.2	2.3	2.75	2.25	2.4	3.2	3.2	3.4	3.0	3.02	2.9
Class:	3	3	3	3	3	3	3	3	3	3	4	4

Grade:	92	72	78	92	92	82
Attend?	0	0	0	1	1	1
GPA:	2.7	2.8	2.8	3.7	3.45	3.1
Class:	4	4	4	4	4	4

Note: For Attend? : 1 = attended the study session; = did not attend.
For Class: 1 = freshman; 2= sophomore; 3=Junior; 4=senior

(١٠-٥٦) هذا التمرين يعتبر تعميم لتمرين (٢-٨٦)، (٨-٤٨) والذى يتعامل مع أسلوب العمل لزملاء العمل تحت قانون شركة Noknvp Bavers ويعتقد أن عدد الـ billable فى الساعة ترتبط مع عدد سنوات الخبرة لزملاء العمل وقسمهم الإدارى . وكلما زادت الخبرة كلما كان هنا الكثير من billable hours والبيانات التالية توضح billable ساعة لأكثر من 9 أشهر لكل 43 زميل حسب القسم الإدارى وعدد سنوات الخبرة .

Hre.:	802	1,287	1,225	1,178	1.275	767	1,424	1,328	1,223	790	1,399	1,434
Yrs.:	3	10	5	9	7	4	9	7	2	4	4	7
Dept.:	1	1	1	2	1	1	3	2	1	1	4	4

Hre.:	1,050	796	1,308	1,464	1,389	1,316	1,325	1,494	1,096	1,482	1,493	1,452
Yrs.:	3	5	5	8	5	6	6	7	3	15	2	7
Dept.:	5	6	6	6	7	4	8	1	1	3	7	3

Hre.:	1,060	1,407	1,067	934	901	1,400	1,320	1,321	1,256	858	1,346	885
Yrs.:	3	12	5	5	4	6	4	10	4	2	6	4
Dept.:	6	6	8	3	1	1	7	1	3	1	8	1

Hre.:	1,084	1,065	1.211	1,379	1,340	1,098	1,407
Yrs.:	5	4	4	10	8	7	5
Dept.:	5	5	1	3	6	5	1

Dept . Code :
1 = Business / commercial Litigation
2 = Labor relations
3 = Real estate
4 = Banking / finance
5 = Administrative
6 = Corporate
7 = Insurance / product liability
8 = Trusts / estates

حال دراسية (١٠-١) : فاعلية أحد الأجهزة الطبية :

قامت إحدي شركات إنتاج الأجهزة الطبية بتطوير شاشة Monitor أحد الأجهزة – قادر على تزويد الجراح بمعلومات جديدة عن حالة المريض أثناء إجراء العملية الجراحية. وتعتقد الشركة وكذلك بعض الأطباء اعتقاداً كبيراً بأن الجهاز يحسن من حالة المريض. فعلى سبيل المثال فانهم يعتقدون بأن الجهاز وإستخدامه قد يؤدي إلي التقليل من عمليات النزيف، ينتج عن إستخدامه قليل من الدماء التي تلوث وتقلل من التعقيدات التي يتعرض لها المريض من اجراء العملية، وهذا يؤدى إلى اختصار وقت إجراء العملية الجراحية. فاذا حصل أحد الدارسين بالمصنع (المحللين) على بعض بيانات عن العمليات التي تم إجرائها (بيانات كافية)، بعضها تم استخدام الجهاز في العمليات، وبعضها لم يتم استخدامه. وكان أمامه هدفين هما:

١ – تحديد بعض البيانات والتصريحات الايجابية والتي تقوم بها الشركة عن الجهاز المنتج.

٢ – تحديد تلك القضايا التي يجب أخذها في الاعتبار في ربع العام التالي عندما يبدأوا في عمل خطط دراسة خاصة بالمنتج في المستقبل (Prospective).

ولقد طلب منك مساعدة المحللين بعمل إستشارة لهم، المحلل يصر على أن هذه البيانات تتوقف على أحداث الماضى (retrospective) بمعنى أنه لم يتم جمعها كنتيجة عمل تجربة مصممة إحصائياً (بإستخدام العشوائية أو التعشية، والقطاعات، وهكذا . . .) ولكنها جمعت من الأداء الفعلي للعمليات في فترتين زمنيتين مختلفتين في الماضي.

١– بيانات من الفترة 1، والتي تعبر عن النتائج للعمليات الجراحية قبل استخدام الجهاز (أو قبل وجود هذا الجهاز).

٢ – بيانات من الفترة 2، والتي تعبر عن النتائج للعمليات الجراحية بعد استخدام الجهاز (أي بعد استخدام المستشفى للجهاز).

وتكون مساعدتك عن طريق فحص وتحليل البيانات آخذا في إعتبارك الأهداف:

١ – تحديد وتعريف التصريحات الايجابية الصحيحة والتي تقوم بها الشركة عن الجهاز المنتج.

٢ – فهم المواقف التي يساعد فيها هذا الجهاز في العمليات والتي لا يساعد فيها (في حالة وجود الجهاز).

يجب عليك استخدام الأساليب المناسبة والملائمة التي درستها في هذا الكتاب من الفصل الأول حتى الفصل العاشر.

وتوجد البيانات الخاصة بهذه الحالة على القرص المرن Disk في الملف المعنون Case 1001. وعموماً فإن المتغيرات التالية هي متغيرات البيانات – أيضاً الأعمدة المصاحبة وكذلك تعليقات مختصرة عن البيانات يمكن تلخيصها فيما يلي:

متغيرات المريض Patient Variables

C1 : نوع المريض (ذكر = 0، انثى = 1)

C2 : حجم المريض (مساحة مسطح الجسم بالبوصة المربعة)

C3 : سن المريض (بالسنوات)

متغيرات اسلوب العمل Operative Procedure Variables

C4 : بطاقة الجراح (ويعنون ذلك بالأرقام 1 ، 4 ، 5)

C5 : نوع الأسلوب المستخدم (حالات عادية = 0، حالات طوارئ = 1)

C6 : استخدام الجهاز (لايستخدم = 0، يستخدم = 1)

بعض المقاييس الداخلية للعمليات

C7 : الوقت الاجمالي لإجراء العملية (بالدقيقة)

C8 : تعداد كرات الدم (قبل اجراء العملية)

متغيرات ما بعد العملية ٢٤ ساعة بعد إجراء العملية

C9 : تعداد كرات الدم

متغيرات خاصة بالعملية (أثناء إجراء العملية)

C10 : الدم المسحوب أو المبذول من الصدر (بالوحدات)

C11 : نقل الدم المطلوب (حالة عدم النقل = 0، حالة النقل = 1)

تعليقات على المتغيرات

* C1 ، C2: نوع المريض وحجم الجسم. يعتقد الأطباء المعالجون بأن الإناث ذو الحجم الأصغر يكونون أكثر عرضة للنزيف خلال العملية.
* C3 عمر المرضى : المرض الأكبر سنا يكونون أكثر عرضة لصعوبات أكثر أثناء أداء العملية.
* C4 الاطباء : قد يختلف الأطباء في اسلوبهم، ودرجة معرفتهم وايلافهم للجهاز ومجتمع المرضى.
* C5 نوع الأسلوب المستخدم: حالات الطوارئ عادة ما تكون حالات غير جيدة عن استخدام الجهاز.
* C6 الجهاز : يفترض المصنع أن استخدام الجهاز يحسن من حالة المريض.
* C7 طول فترة إجراء العملية. حالات الطوارئ والحالات التي ينتج عنها مضاعفات تأخذ وقت أطول في اجراء العملية.
* C8 : عدد كرات اليوم قبل إجراء العملية. وهذا يعطي الأساس في المقارنة والتغير في تعداد كرات الدم والذي يرجع إلى النزيف خلال إجراء العملية. بالاضافة إلى ذلك إذا كانت عدد كرات الدم تبدأ منخفضة بصورة كبيرة، فإن هذا يعنى أن المريض ليس في حالة جيدة في بداية اجراء العملية.
* C9 : مستوى الكرات في الدم بعد ٢٤ ساعة من إجراء العملية. يعتقد بعض الأطباء أن انخفاض كرات الدم بشكل كبير غير مرغوب فيه.
* C10 : سحب أو بذل الدم. كلما زادت كمية الدماء النازفة من صدر المريض أثناء اجراء العملية أو بعدها، كلما كانت حالة المريض سيئة.

* C11 عمليات نقل الدم. وخصوصا نوع من نقل الدم والذي يسمى هذه الأيام HIV ولاتتم عملية نقل الدم إلا إذا كانت ضرورية.

تعليق عن البيانات :

* أحياناً، تكون قيم البيانات غير الموجودة missing.
* جميع البيانات تم ترتيبها بترتيب زمنى يتفق مع وقت إجراء العملية.

حالة دراسية (١٠-٢) تحليل مزايا الفريق المحلي: في مسابقة الدوري المحلي كرة القدم الأمريكية

هناك اعتقاد كبير أن الفريق المحلي له مزايا في جميع الأحداث الرياضية. وهذه الحالة الدراسية تعتبر فرصة لك في تحليل ما إذا كانت هذه المزايا موجودة في الدورى المحلي لكرة القدم الأمريكية National Football League (NFL)، وإذا كان الأمر كذلك، فإن الأهمية ترجع إلى الاختلاف في نوع المظهر الخارجي لملاعب الفريق المحلي، أيضاً تمتد الأهمية إلى القوة النسبية للفرق المنافسة.

والبيانات المتاحة تمثل حوالى 224 مباراة من مباريات (NFL) في موسم ١٩٩٢، وتم أخذ هذه البيانات من سجلات الرياضة، وتوجد على القرص المرن المرفق والذي يرافق هذا الكتاب تحت ملف اسمه Casel 1002. ويحتوى الملف 224 سطر من البيانات، سطر لكل مباراة تمت. و المتغيرات (الأعمدة) هي:

C1 : مؤشر عن ترتيب المباراة المحلية (1 إلى 18).

C2 : اسم الفريق المحلي .

C3 : اسم الفريق الزائر.

C4 : المخرجات ويعبر عنها (0 = حالة فوز الفريق الزائر، 1 = في حالة فوز الفريق المحلي).

C5 : النقاط للفريق المحلي.

C6 : النقاط للفريق الزائر.

C7 : هامش فوز الفريق المحلي أو المنزلي (أو خسارته) أي (C5 - C6).

C8 : مظهر ومساحة ملعب الفريق المحلي أو المنزلي.

C9: مظهر ومساحة ملعب الفريق الزائر.

C10: نسبة الفوز للفريق المنزلي أو المحلي في نهاية موسم الدوري.

C11: نسبة الفوز للفريق الزائر في نهاية موسم الدوري.

والمطلوب منك أن تتبع وتحلل البيانات باستخدام التفكير الاحصائي والاساليب التي درستها في الفصول من الأول حتى العاشر.

يجب أن تطور وتحدد نموذج يوضح العلاقة بين مزايا ملعب الفريق المحلي (إن وجدت) ومايلي: (١) ومساحة ومظهر ملعبى للفريقين، (٢) القوة النسبية للفريق المنافس في تقريرك لابد من ذكر نتائجك وان تبررها آخذ في الاعتبار وجود وكذلك حجم ملاعب الفريقين، التأثير على مزايا الفريق

المحلي با للعب على النجيل الصناعي (الأرض الصناعية) مع زيادة العيوب عندما يلعب علي أرض بها نجيل طبيعي أو حشائش طبيعية، وكذلك العلاقة بين مزايا الفريق المحلي والقوة النسبية للفرق المنافسة كما تقاس بالنسبة المئوية للفوز لهم في نهاية موسم الدوري.

وبأيجاد النموذج المناسب، سوف تستخدم الانحدار المتعدد. يجب أن تفكر في متغير الاستجابة. فلابد أن يكون فترة أو نسبة لتأكيد الفروض الضرورية، ولذلك فإنك لا نستطيع استخدام C4 (الخسارة أو المكسب) كمتغير استجابة.

ملحق ١٠ : Appendix -10

تعليمات الحاسب لإستخدام Minitab , SAS

سوف نستخدم مثال الأيس كريم (انظر جداول ١٠-٢، ١٠-٤، ١٠-١١، ١٠-١٥، ١٠-١٦) ومثال (١٠-١٦) (انظر جداول ١٠-٦، ١٠-٩) وأشكال (١٠-٥، ١٠-٦) لتوضيح تعليمات (إرشادات) الـ Minitab , SAS . لإيجاد مخرجات الحاسب لهذه التمارين .

مثال الأيس كريم ICE Cream parlor example

SAS:

تولد التعليمات الآتية المخرجات المعطاة فى جدول(١٠-٢) . لاحظ أن لدينا نوع اليوم والرطوبة كمتغيرات مفسرة فى جملة INPUT وبياناتها. حيث ستستخدم هذه المتغيرات التفسيرية بشكل مختصر.

```
DATA:
INPUT SALES PRICE TEMP DAY HUMID;
CARDS'
374   35   74   1   50
386   35   85   0   72
472   35   94   1   92
429   50   93   0   88
391   50   85   1   70
475   50   96   1   94
428   50   91   1   85
412   65   93   0   89
405   65   88   1   80
341   65   78   0   60
PROC GLM;
MODEL SALES = PRICE TEMP;
```

للحصول على المخرجات المعطاة فى جدول (١٠-٤) نعدل فقط جملة MODE كما يلى :

```
MODEL SALES = PRICE TEMP DAY;
```

للحصول على المخرجات المعطاة في جدول (١٠-١١) . والتى تتضمن الرطوبة كمتغير مفسر لكن لا تحتوى على نوع اليوم . نعدل أيضاً جملة MODEL كما يلى :

```
MODEL SALES = PRICE TEMP HUMID;
```

للحصول على المخرجات المعطاة فى جدول (١٠-١٥) للإنحدار STEPWISE سنستخدم PROC STEPWISE مع جملة MODEL كما يلى :

```
PROC STEPWISE;
MODEL SALES = PRICE TEMP DAY HUMID;
```

فى النهاية للحصول على مخرجات الجدول (١٠-١٦) لحذف الخلفى نستخدم PROC STEPWISE لكن نعدل فى جملة MODEL كالتالى :

```
MODEL SALES = PRICE TEMP DAY HUMID / B;
```

Minitab:

تستخدم أوامر Name ، READ لإدخال البيانات ويتضمن لنوعية اليوم والرطوبة كما يلى :

```
MTB >  name c1 = 'sales' c2 = 'price' c3 = ' temp' c4 = 'day' c5 = 'humid'
BTB >  read c1 - c5
DATA >  347  35  74  1  50
DATA >  386  35  82  0  72
DATA >  472  35  94  1  92
DATA >  429  50  93  0  88
DATA >  391  50  82  1  70
DATA >  475  50  96  1  94
DATA >  428  50  91  1  85
DATA >  412  65  93  0  89
DATA >  405  65  88  1  80
DATA >  341  65  78  0  60
DATA >  end
```

للحصول على مخرجات Minitab فى جدول (١٠-٢) نستخدم الأمر REGRESS ونشير إلى عدد المتغيرات التفسيرية التى نهتم بها والأعمدة التى تحتوى على بياناتهات كالتالى :

```
MTB > regress  Y c1  2  c2   c3
```

إذا أردنا طباعة قيم Y والبواقى نحدد العمود فى جملة REGRESS للقيم Y ونستخدم الأمر الفرعى RESIDUAL مع العمود المعمم للبواقى .

للحصول على معادلات الـMinitab للجداول (١٠-٤) ، (١٠-١١) سوف نعدل ببساطة أمر REGRESS كما يلى :

```
MTB > regress  Y  c1  3  c2  c3  c4
MTB > regress  Y  c1  3  c2  c3  c5
```

للحصول على معادلات الـ Minitab وإنحدار stepwise للـ SAS المعطى فى جدول (١٠-١٥) سنستخدم أمر stepwiese مع الأمر الفرعى F Enter ، Fremove حيث أن كلا الأمران الفرعيان يكونوا مساويان للقيمة 4 . هذه التعليمات تنسجم مع مخرجات الـ Minitab كما يلى :

من المخرجات نلاحظ أن المعلومات فى أخر عمود تحتوى على معادلات المربعات الصغرى للمتغيرات المفسرة الموجودة فى معادلة الإنحدار المنسجمة مع قيم T . أخر صفين فى العمود يكون

إنحراف البواقى المعيارية وقيمة R2 لأفضل معادلة إنحدار .

```
MTB > stepwise  Y  c1  c2  - c5 ;
SUBC > fenter = 4 ;
SUBC > fremove = 4.

STEPWISE REGRESSION OF sals ON 4
PREDICTORS, WITH N = 10

    STEP          1       2       3
CONSTANT     -10.81   25.88   15.31

temp           4.85    5.20    5.04
T-RATIO        5.22    8.90   15.83

price                 -1.34   -1.10
T-RATIO               -3.71   -5.40

day                   20.2
T-RATIO               4.23

S             21.1    13.1    7.11
R -SQ        77.33   92.35   98.08
```

للحصول على معادلات Minitab للحذف الخلفى المعطى فى جدول (١٠-١٦) نستخدم مرة أخرى الأمر STEPWISE المنسجم مع الأمر FENTER ، FREMOVE , ENTER . والآن جملة الأمر الفرعى FENTER تكون مجموعة من 10000 والأمر الفرعى ENTER لتحديد كل المتغيرات المفسرة التى نريد إدخالها. هذه الإرشادات تنسجم مع مخرجات 1 (MINITAb) كما يلى وكما من قبل المعلومات فى العمود الأخير يحتوى على النتائج المتعلقة بأفضل معادلة إنحدار :

```
MTB > stepwise   Y   c1   c2 - c5 ;
SUBC > fenter = 100000 ;
SUBC > fremove = 4 .
SUBC > enter c2 - c5 .

SATEPWISE REGRESSION OF sales
ON 4 PREDICTORS , WITH N = 10

    STEP          1       2
CONSTANT     -112.85   15.31

price        -1.15    -1.10
T-RATIO      -5.58    -5.40

temp          7.86     5.04
T-RATIO       3.02    15.83

day          18.9     20.2
T-RATIO      3.89     4.23
```

```
humid          -1.5
T-RATIO       -1.09

S              7.00      7.11
R -SQ         98.45     98.08
```

مثال (١٠-٦)

الإرشادات المتبعة للحصول على النتائج فى جدول (١٠-٨) ورسومات (١٠-٥) أ ، ب (a or b):

```
DATA;
INPUT COST RATE LABOR;
CARDS;
13.59    87      80
15.71    78      95
15.97    81      106
20.21    65      115
24.64    51      128
21.25    62      128
18.94    70      115
14.85    91      92
15.18    94      93
16.30    100     111
15.93    102     116
16.45    82      117
19.02    74      127
18.16    85      133
18.57    86      135
17.01    90      136
18.03    93      140
19.22    81      142
21.12    72      148
23.32    60      150
PROC GLM;
MODEL COST = RATE LABOR;
OUTPUT OUT = A
RESIDUAL = RESID;
PROC PLIT DATA = A;
PLOT RESID * RATE;
PLOT RESID * LABOR;
```

لتفسير العنصر المربع المشتمل على المعدل. نشكل العمود المحتوى على المعدلات المربعة بواسطة أمر الإدراج .

```
RATESQRD = RATE * RATE;
```

بين الجمل INPUT , CARDS ثم نتبع الإرشادات للحصول على نتائج جدول (١٠-٩) والأشال فى (١٠-٦).

```
DATA;
INPUT COST RATE LABOR;
RATESQRD = RATE * RATE;
CARDS;
13.59    87     80
15.71    78     95
15.97    81     106
20.21    65     115
24.64    51     128
21.25    62     128
18.94    70     115
14.85    91     92
15.18    94     93
16.30    100    111
15.93    102    116
16.45    82     117
19.02    74     127
18.16    85     133
18.57    86     135
17.01    90     136
18.03    93     140
19.22    81     142
21.12    72     148
23.32    60     150
PROC GLM;
MODEL COST = RATE LABOR ;
OUTPUT OUT = A
RESIDUAL = RESID;
PROC PLIT DATA = A;
PLOT RESID * RATE;
PLOT RESID * LABOR;
MINITAB
```

التعليمات المتبعة للحصول على معادلات الـ Minitab للجداول (١٠-٨)، (١٠-٩) والأشكال (١٠-٥)، (١٠-٦).

```
MTB > name c1 = `cost` c2 = `rate` c3 = `labor` c4 = `ratesqrd`
MTB > read c1 - c3
DATA > 13.59    87     80
DATA > 15.71    78     95
DATA > 15.97    81     106
DATA > 20.21    65     115
DATA > 24.64    51     128
DATA > 21.25    62     128
DATA > 18.94    70     115
DATA > 14.85    91     92
DATA > 15.18    94     93
DATA > 16.30    100    111
DATA > 15.93    102    116
```

```
DATA > 16.45    82      117
DATA > 19.02    74      127
DATA > 18.16    85      133
DATA > 18.57    86      135
DATA > 17.01    90      136
DATA > 18.03    93      140
DATA > 19.22    81      142
DATA > 21.12    72      148
DATA > 23.32    60      150
DATA > end
MTB >  let c4 = c2 * c2
MTB > regress Y c1  2 c2 c3;
SUBC > residual c5
MTB > plot c5  c2
MTB > plot c5  cE
MTB > regress Y c1 E c2 c3 c4;
SUBC > residual c6.
MTB > plot c6  c2.
MTB > lpot c6 cE.
```