تحليل الانحدار
Regression Analysis

مقدمة

تحليل الانحدار هو أداة إحصائية تقوم ببناء نموذج إحصائي وذلك لتقدير العلاقة بين متغير كمي واحد وهو المتغير التابع ومتغير كمي آخر أو عدة متغيرات كمية وهي المتغيرات المستقلة، بحيث ينتج معادلة إحصائية توضح العلاقة بين المتغيرات. ويمكن استخدام هذه المعادلة في معرفة نوع العلاقة بين المتغيرات وتقدير المتغير التابع باستخدام المتغيرات الأخرى. وعندما تكون العلاقة في النموذج الإحصائي بين متغير تابع واحد ومتغير مستقل واحد، فإن هذا النموذج هو أبسط نماذج الانحدار ويسمى النموذج الخطي أو البسيط Simple Linear Regression، وعندما تكون عدد المتغيرات المستقلة أكثر من متغير كمي واحد فإن النموذج يسمى نموذج الانحدار المتعدد Multiple Regression.

نموذج الانحدار البسيط Simple Linear regression

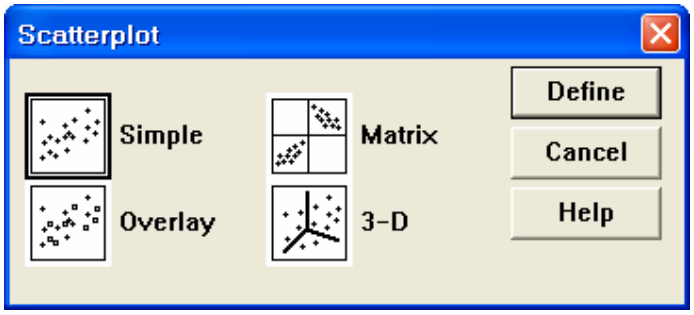
نموذج الانحدار البسيط هو نموذج إحصائي يقوم بتقدير العلاقة التي تربط بين متغير كمي واحد وهو المتغير التابع مع متغير كمي آخر وهو المتغير المستقل. وينتج من هذا النموذج معادلة إحصائية خطية يمكن استخدامها لتفسير العلاقة بين المتغيرين أو تقدير قيمة المتغير التابع عند معرفة قيمة المتغير المستقل. ويمكن صياغة العلاقة الإحصائية بالنموذج التالي:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

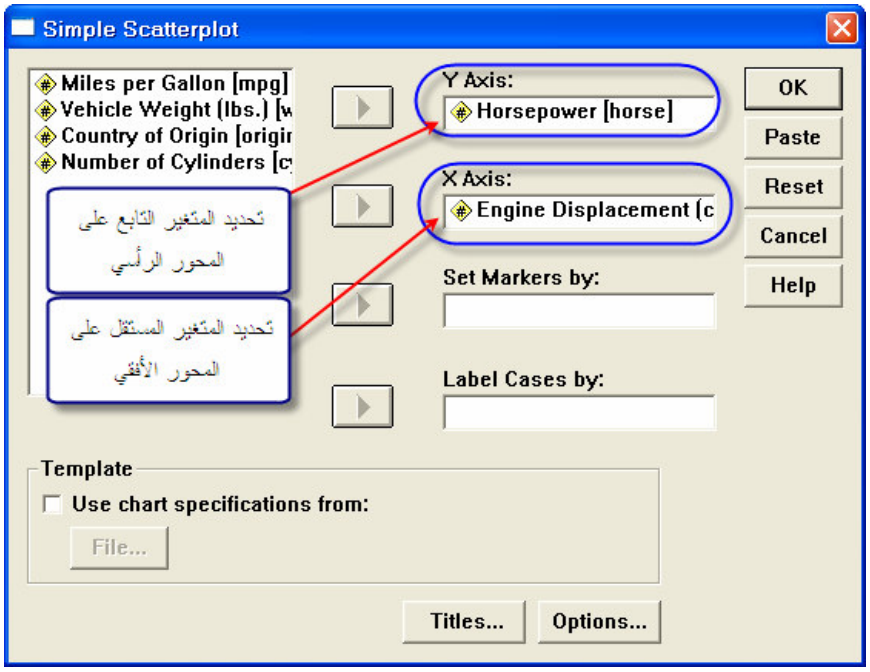
حيث y_i هو المتغير التابع و x_i هو المتغير المستقل و ε_i هو الخطأ العشوائي و β_0 هي قيمة ثابتة تعبر عن قيمة y عندما تكون قيمة x تساوي الصفر و β_1 تعبر عن ميل الخط المستقيم الذي يوضح العلاقة.

ويفترض النموذج أن y متغير عشوائي يتبع التوزيع الطبيعي بمتوسط يرتبط بقيمة x وبتباين ثابت على اختلاف قيم x. وبذلك فإنه وتبعاً للفرضية السابقة فإن $\varepsilon_i \sim N\left(0,\sigma^2\right)$.
ويمكن تقدير معالم النموذج باستخدام طريقة المربعات الصغرى.
وسيتم فيما يلي استخدام بيانات السيارات Cars.sav لتقدير نماذج إحصائية مختلفة والقيام باختبار الفرضيات حول النماذج. لنفرض أن لدينا الرغبة في تقدير العلاقة بين قوة السيارة بالحصان كمتغير تابع وسعة الاسطوانات كمتغير مستقل، لذلك سيتم استخدام نموذج الانحدار البسيط لتقدير العلاقة بين المتغيرين ثم اختبار معالم النموذج. إلا أن الخطوة الأولى قبل تقدير

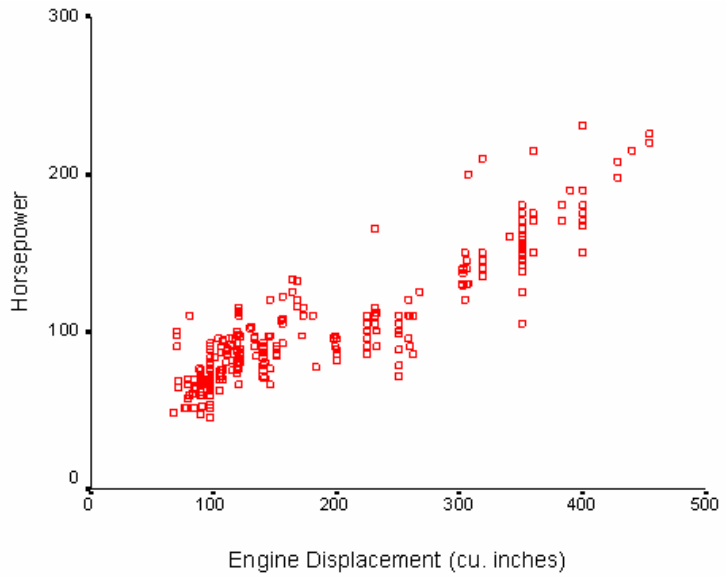
النموذج هو فحص العلاقة بين المتغيرين بيانياً للتأكد من كون العلاقة خطية أو غير خطيـــة. وللقيام بذلك يمكن رسم انتشار البيانات باختيار الأمر Scatter من قائمـــة Graphs ليظهـــر مربع الحوار التالي:



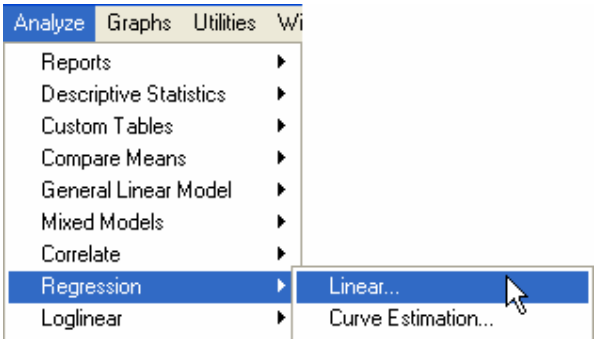
وباختيار نوع الأمر Simple ثم النقر على Define يظهر مربع الحوار التالي:



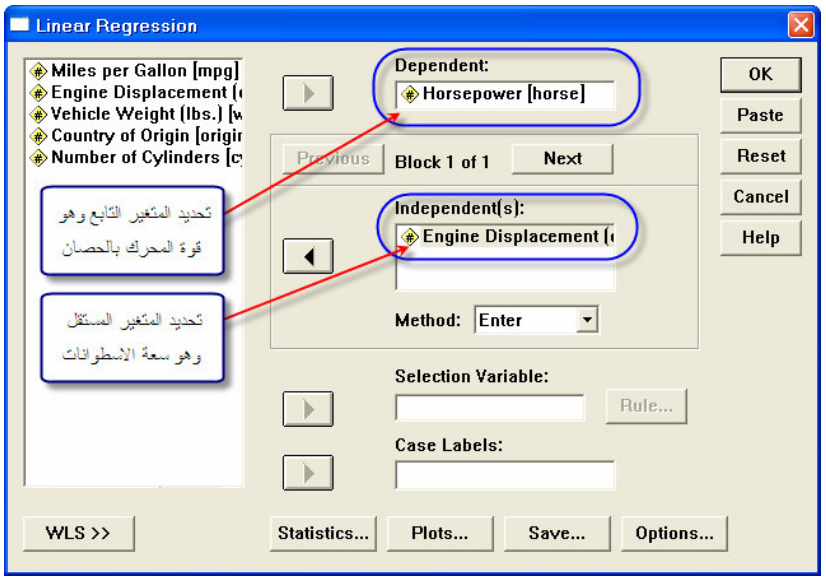
وبالنقر على OK يظهر الرسم التالي على شاشة عارض النتائج.



ويلاحظ من الرسم أنه يمكن تمثيل العلاقة بخط مستقيم وهي علاقة إيجابية، لذلك فإن نموذج الانحدار الخطي سيكون مناسب لهذه العلاقة. ولتقدير النموذج الخطي نستخدم الأمر



وبذلك يظهر مربع الحوار التالي



وبالنقر على OK تظهر النتائج التالية.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	461270.9	1	461270.874	1628.778	.000[a]
	Residual	110165.0	389	283.201		
	Total	571435.9	390			

a. Predictors: (Constant), Engine Displacement (cu. inches)

b. Dependent Variable: Horsepower

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.434	1.796		22.519	.000
	Engine Displacement (cu. inches)	.329	.008	.898	40.358	.000

a. Dependent Variable: Horsepower

وتحتوي النتائج على جدول ANOVA والذي يوضح مقدار ما يفسر النموذج الخطي من البيانات، وتشير القيمة الكبيرة لـ F أن نموذج الانحدار الخطي يفسر جزء كبير من البيانات وإن الاختلافات العشوائية قليلة. ويختبر جدول ANOVA معنوية النموذج باستخدام توزيع F، حيث يختبر الفرضية التالية

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

وبناء على الجدول السابق فإن فرضية العدم مرفوضة. وفي الحقيقة فإن اختبار F يختبر معنوية العلاقة بين المتغيرين ولا يختبر معنوية المعلمة β_1 كمعلمة للنموذج. وتبرز أهمية اختبار F عندما يوجد في النموذج أكثر من متغير مستقل واحد، فإن اختبار F في هذه الحالة يختبر معنوية جميع معالم النموذج دفعة واحدة.

$$H_0 : \beta_1 = 0; \beta_2 = 0; \ldots; \beta_k = 0$$
$$H_a : \beta_1 \neq 0; \beta_2 \neq 0; \ldots; \beta_k \neq 0$$

وفي حالة رفض فرضية العدم ننتقل إلى اختبار معالم النموذج بحث يتم اختبار معنوية كل معلم من معالم النموذج بصورة منفصلة عن الآخر، أما إذا لم يتم رفض فرضية العدم، فليس هناك حاجة لاختبار معالم النموذج مما يشير إلى أن النموذج المستخدم غير مناسب. كذلك فإنه يمكن حساب معامل التحديد للنموذج والذي يساوي

$$R^2 = \frac{SSR}{SST} = \frac{461270.9}{571435.9} = 0.807$$

وبذلك فإن النموذج يفسر 80.7% من الاختلافات في قيم المتغير التابع، في حين 19.3 من الاختلافات ناتجة من عوامل عشوائية. ولمعرفة مدى تشتت الخطأ العشوائي حول خط الانحدار، يستخدم متوسط مجموع مربعات الفروق للبواقي Residuals والذي يساوي:

$$MSE = \frac{110165}{389} = 283.20$$

وتشير قيم MSE الصغيرة إلى تركز البيانات حول خط الانحدار، إلا أنه يجب أن لا نهمل تأثير قيم المتغير التابع الكبيرة على تباين الخطأ العشوائي.

ويحتوي الجدول الثاني على تقديرات لمعالم النموذج مع اختبارات T لمعنوية معالم النموذج، حيث كانت قيمة $(\beta_0 = 40.434)$ وقيمة $(\beta_1 = 0.329)$، وبناء على ذلك فإن النموذج المقدر هو:

$$\hat{y} = 40.434 + 0.329x$$

ويختبر الجدول الثاني معنوية معالم النموذج بصورة منفصلة عن بعضها البعض، حيث يختبر الفرضيتين التاليتين:

الفرضية الأولى

$$H_0 : \beta_0 = 0$$
$$H_a : \beta_0 \neq 0$$

الفرضية الثانية

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

بناء على نتائج الجدول فإنه يمكن رفض فرضية العدم في الفرضية الأولى لصالح الفرضية البديلة حيث أن القيمة المحسوبة لاختبار T تساوي:

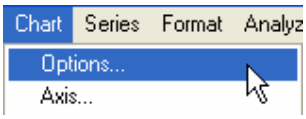
$$t^* = \frac{b_0}{s(b_0)} = 22.52$$

وقيمة P-Value تساوي الصفر. كذلك فإنه يمكن رفض فرضية العدم في الفرضية الثانية لصالح الفرضية البديلة حيث أن القيمة المحسوبة لاختبار T تساوي:

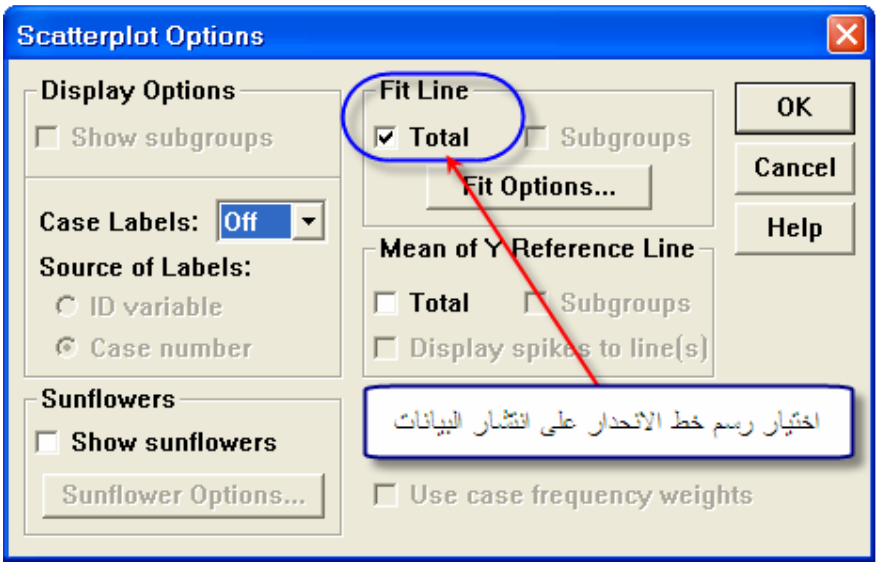
$$t^* = \frac{b_1}{s(b_1)} = 40.36$$

وقيمة P-Value تساوي الصفر. إلا أنه تجدر الإشارة هنا إلى أن اختبار الفرضية الأولى لا يعني شيء إذا لم نتمكن من رفض الفرضية الثانية. والسبب في ذلك في أن النموذج لن يكون مناسب لتقدير العلاقة في المقام الأول.

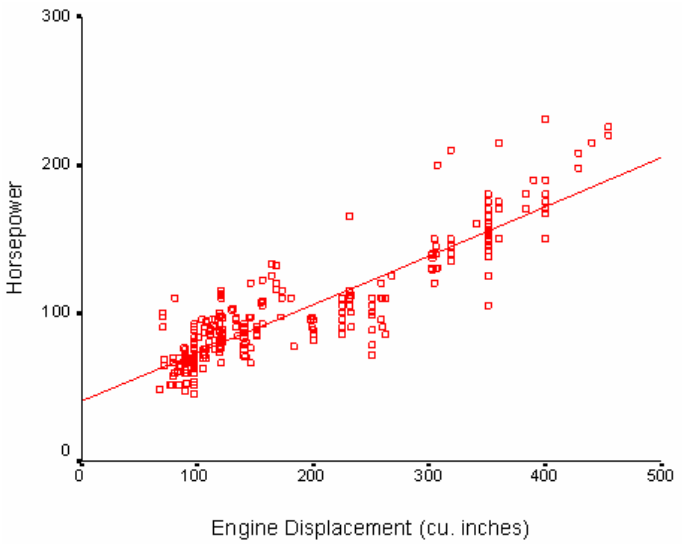
ويمكن عرض خط الانحدار بيانياً مع رسم الانتشار السابق وذلك بالنقر المزدوج على رسم الانتشار السابق لتنشيط محرر الرسومات ثم اختيار الأمر



ليظهر مربع الحوار التالي:

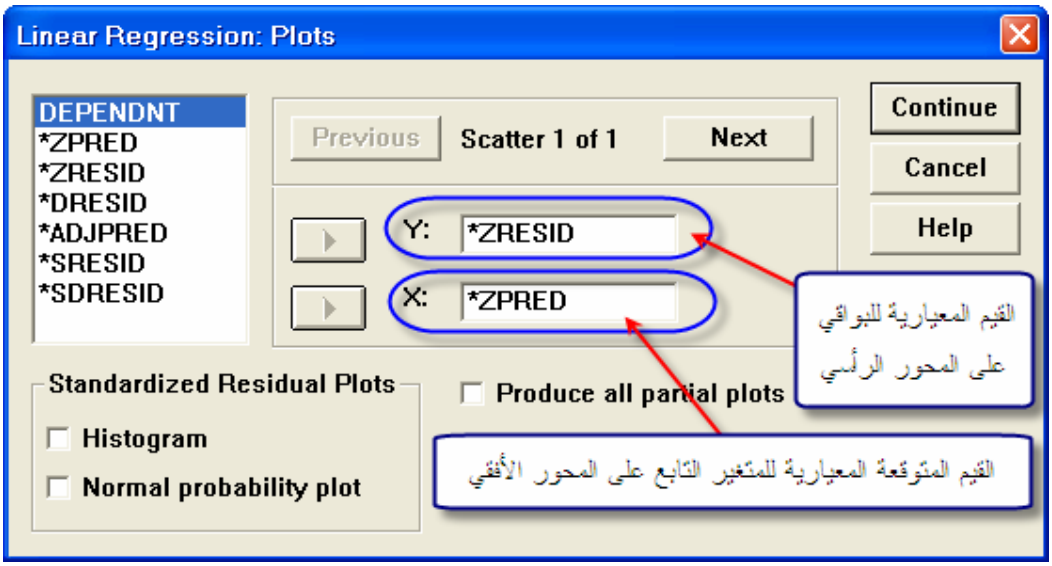


وبذلك يظهر خط الانحدار على رسم الانتشار

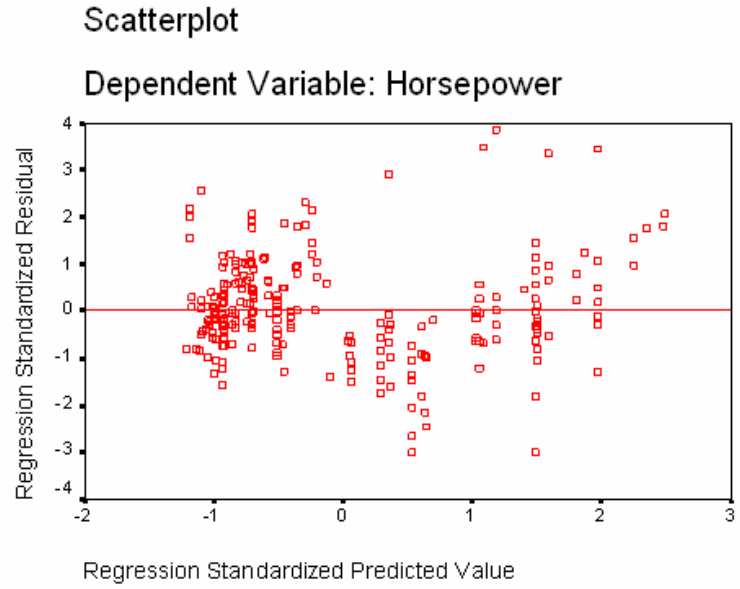


ومن الخطوات الضرورية في تحليل الانحدار هو معرفة انتشار الخطأ العشوائي المعياري مع قيم المتغير التابع المعيارية، ويتم استخدام القيم المعيارية لتلافي تأثير المقياس للمتغير التابع.

ويتم ذلك بالنقر على زر Plots في مربع حوار Linear Regression ليظهر مربع الحوار التالي:



وبالنقر على Continue ثم OK تظهر النتائج التالية:



ويلاحظ أن البواقي المعيارية تنتشر وبشكل جيد حول خط الصفر، مما يدل على أن النموذج جيد، ويمكن كذلك من مربع الحوار السابق اختبار توزيع البواقي بيانياً باستخدام Histogram و Normal probability plot والتأكد بأنها تتبع التوزيع الطبيعي.

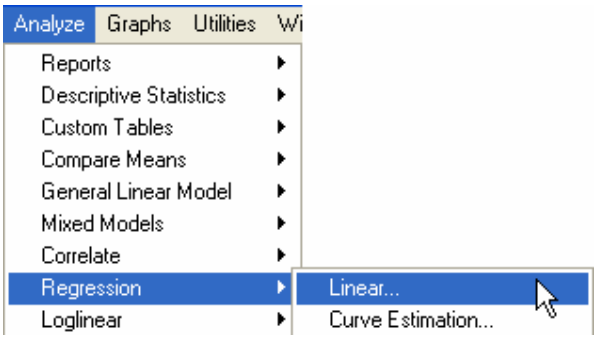
تحليل الانحدار المتعدد Multiple Linear Regression

يعتبر نموذج تحليل الانحدار المتعدد من أكثر أدوات التحليل الإحصائي استخداماً، ويهتم نموذج الانحدار المتعدد بتقدير العلاقة بين متغير كمي وهو المتغير التابع وعدة متغيرات كمية أخرى وهي المتغيرات المستقلة. وبافتراض وجود متغير تابع ومتغيرين مستقلين، فإنه يمكن صياغة النموذج على النحو التالي:

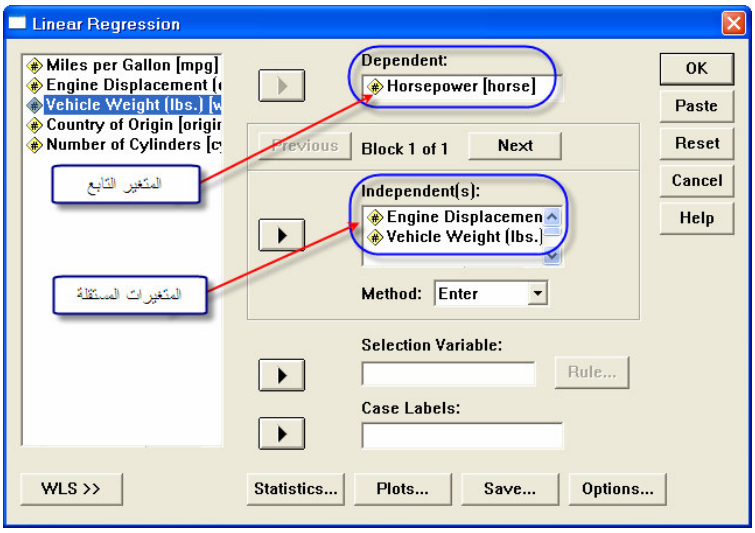
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

حيث y_i هو المتغير التابع و $(x_{1i}; x_{2i})$ هي المتغيرات المستقلة و ε_i هو الخطأ العشوائي و β_0 هي قيمة ثابتة تعبر عن قيمة y عندما تكون قيم $(x_1; x_2)$ تساوي الصفر و $(\beta_1; \beta_2)$ تعبر عن معاملات الانحدار للمتغيرات المستقلة. وتبقى فرضيات نموذج الانحدار الخطي سارية على نموذج الانحدار المتعدد بالإضافة إلى فرضية عدم وجود ارتباط بين المتغيرات المستقلة.

وسيتم استخدام بيانات السيارات لتكوين علاقة بين قوة السيارة بالحصان كمتغير تابع والمتغيرين المستقلين وهما سعة الاسطوانات ووزن السيارة. ولتقدير النموذج يستخدم الأمر



ليظهر مربع الحوار التالي والذي يمكن من خلاله تحديد النموذج الخطي وتحديد الإحصائيات والرسومات المطلوبة.



وبتحديد النموذج الخطي والنقر على OK تظهر النتائج التالية.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	463820.8	2	231910.394	836.139	.000[a]
	Residual	107615.1	388	277.359		
	Total	571435.9	390			

a. Predictors: (Constant), Vehicle Weight (lbs.), Engine Displacement (cu. inches)

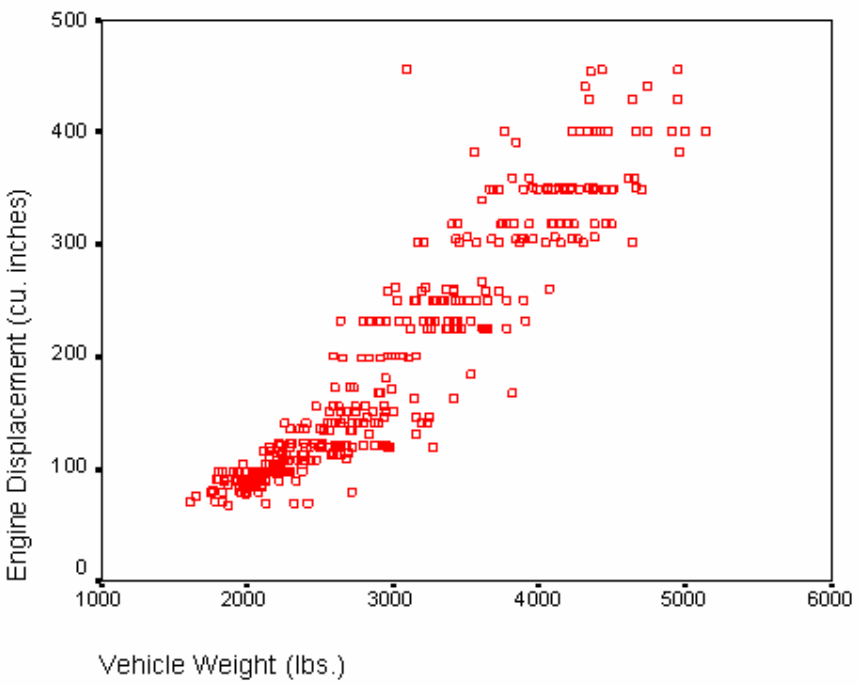
b. Dependent Variable: Horsepower

Coefficients[a]

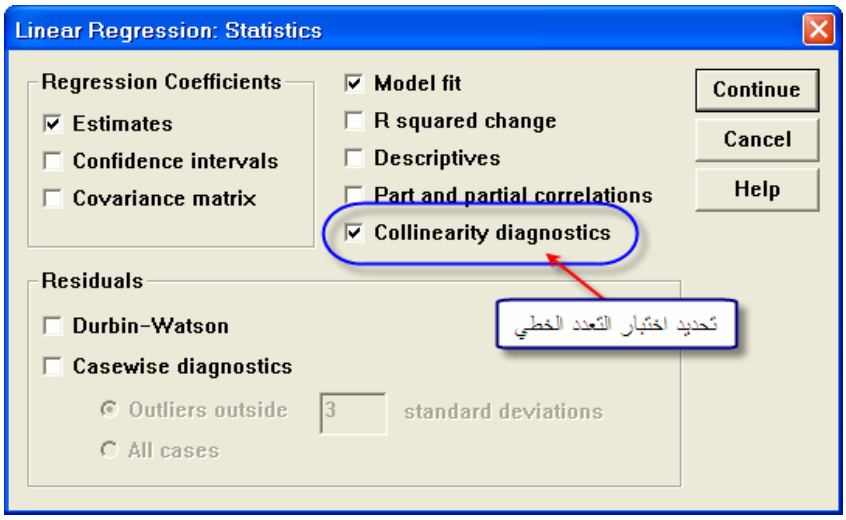
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	27.689	4.563		6.068	.000
	Engine Displacement (cu. inches)	.265	.023	.724	11.751	.000
	Vehicle Weight (lbs.)	8.454E-03	.003	.187	3.032	.003

a. Dependent Variable: Horsepower

ويختبر الجدول الأول معنوية النموذج في حين أن الجدول الثاني يختبر معنوية معالم النموذج، ويمكن استنتاج أن النموذج الخطي جيد وأن المعالم جميعها تختلف عن الصفر. ويمكن ملاحظة أن إضافة متغير جديد للنموذج الخطي البسيط لم يكن له تأثير ايجابي كبير على مجموع مربعات الفروق وبالتالي لم يفيد المتغير المستقل الجديد في تفسير الاختلاف في المتغير التابع. وقد يكون السبب الأساسي لهذا هو وجود ارتباط قوي بين المتغيرات المستقلة وهو ما يسمى بالتعدد الخطي أو Multicollinearity. ويمكن التأكد من وجود ارتباط بين المتغيرات المستقلة بتمثيلها بيانياً.



ويوضح الرسم البياني الارتباط القوي بين المتغيرات المستقلة مما يشير إلى أنه يمكن الاستغناء عن أحد المتغيرين والاكتفاء بمتغير واحد. ويمكن قياس مشكلة التعدد الخطي بالنقر على زر Statistics من مربع حوار Linear Regression ثم تحديد اختبار التعدد الخطي.



وبالنقر على Continue ثم OK، تظهر نتائج اختبار التعدد الخطي ضمن نتائج تحليل الانحدار.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Engine Displacement (cu. inches)	Vehicle Weight (lbs.)
1	1	2.874	1.000	.00	.00	.00
	2	.120	4.903	.15	.12	.00
	3	6.244E-03	21.454	.85	.88	1.00

a. Dependent Variable: Horsepower

ويمثل الجدول السابق حسابات مرتبطة بمصفوفة المتغيرات، حيث يرتبط الصف الأول بالقيمة الثابتة ويرتبط الصف الثاني بالمتغير المستقل الأول ويرتبط الصف الثالث بالمتغير المستقل الثاين. وتعتبر مشكل التعدد الخطي مؤثرة إذا كانت قيمة دليل الحالة Condition Index للمتغير كبيرة، فإذا زادت القيمة عن 15 فهذا مؤشر على وجود مشكلة التعدد الخطي. وبالنظر إلى الجدول السابق فإن قيمة دليل الحالة للمتغير الثاني تساوي 21.454، وبالنظر إلى المتغيرات التي يساهم بها المتغير الثاني، نجد أن المتغير المستقل الثاني وهو وزن السيارة يساهم بنسبة 88% من تباين المتغير المستقل الأول وهو سعة الاسطوانات.