(1) تحديد مشاهدات قاصية في X (المتغير المستقل) - مصفوفة القبعة وقيم العزم (القيم النفوذية) :-

- حالات قاصية (حالات شاذة):

في تطبيقات تحليل الانحدار الخطي كثيرا ما تحتوي مجموعة من البيانات على بعض الحالات القاصية أوالمتطرفة أي أن هذه المشاهدات تكون منفصلة بوضوح عن بقية المشاهدات الأخرى .وقد تحتوي هذه الحالات القاصية على رواسب كبيرة ويكون لها في الغالب ،تأثيرات على دالة الانحدار التوفيقية ومن المهم دراسة الحالات القاصية بعناية وتقرير ما إذا كان ينبغي الاحتفاظ بها أو إلغاؤها ،وفي حالة الاحتفاظ بها ،تقرير ما إذا كان ينبغي تخفيض نفوذها في عملية التوفيق ، أو إعادة النظر في نموذج الانحدار.

وإحدى الخطى الأساسية في أي تحليل انحدار هي تحديد ما إذا كان نموذج الانحدار المدروس خاضعا لسطوة مشاهدة واحدة أو قلة من المشاهدات في مجموعة البيانات . وفي انحدار بمتغير مستقل واحد أو متغيرين يكون من السهل نسبياً التعرف على مشاهدات قاصية في قيم X (المتغير المستقل) أو في Y (المتغير التابع) بوسائل مثل الرسم الصندوقي " Box plot" ، رسوم الجذع و الورقة " stem-and-leaf plot" ، رسوم الانتشار " Scatter Plot" ،ورسوم البواقي " Residuals plot" . ودراسة ما إذا كان لها نفوذ له تأثيره على دالة الانحدار التوفيقية. إلا أنه عندما يشمل نموذج الانحدار الخطي على أكثر من متغيرين مستقلين ، يصبح التعرف على مشاهدات قاصية بالوسائل البيانية البسيطة أمرا صعبا ،ذلك لأن تفحص متغير بمفرده أو متغيرين لا يساعد تحديد القاصيات بالنسبة لنموذج انحدار متعدد المتغيرات.

وبعض القاصيات في متغير واحد قد لا تكون متطرفة في نموذج انحدار متعدد، وعلى العكس،قد لا نتمكن من اكتشاف قاصيات في عدد من المتغيرات عند تحليل يتطرق لمتغير واحد أو لمتغيرين منها.

لذلك نتعرف الآن على استخدام مصفوفة القبعة " Hat Matrix"التي قد تساعد في اكتشاف الحالات القاصية في نموذج الانحدار الخطي بأكثر من متغير مستقل.

- استخدام مصفوفة القبعة " H " للتعرف على مشاهدات قاصية في المتغير المستقل X :-

تعرف مصفوفة القبعة H بالعلاقة التالية :-

$$H = X\,(X'X)^{-1}\,X'$$

وحيث أن hii هو العنصر القطري لمصفوفة القبعة H .ويمكن الحصول عليه من العلاقة التالية :

$$h_{ii} = X_i'\,(X'X)^{-1}X_i$$

وحيث أن Xi تخص فقط المشاهدة i :-

$$X_i = \begin{bmatrix} 1 \\ X_{i,1} \\ . \\ . \\ . \\ X_{i,p-1} \end{bmatrix}$$

حيث أن X_i' هي الصف i من المصفوفة X المتعلق بالمشاهدة i .

وللعناصر القطرية hii في مصفوفة القبعة بعض الخواص، ونذكر على وجه الخصوص:

1 -قيمها تقع بين الصفر والواحد $0 \le hii \le 1$

2 -مجموعها يساوي p $\sum_{i=1}^{n} hii = p$

حيث p تساوي عدد المعالم في نموذج الانحدار الخطي.

- طريقة كشف القيم القاصية في المتغيرات المستقلة باستخدام قيم العزم (القيم الرافعة أوالنفوذية) :hii

تعتبر قيمة العزم (القيمة الرافعة أو النفوذية) hii كبيرة إذا تجاوزت ضعف متوسط قيم العزم. حيث أن متوسط قيم العزم يعطى بالعلاقة التالية:

$$\bar{h} = \frac{\sum_{i=1}^{n} hii}{n} = \frac{p}{n}$$

وأما ضعف متوسط قيم العزم فيعطى بالعلاقة :

$$2\,\bar{h} = \frac{2\,P}{n}$$

بالتالي فإن المشاهدات التي تكون قيم العزم لها أكبر من ضعف متوسط قيم العزم تعتبر قيم قاصية وفقا لهذه القاعدة، وهي مؤشر جيد لوجود مشاهدات قاصية. أي إذا كان :

$$hii > \frac{2P}{n}$$

والبينة الإضافية لمشاهدة قاصية هي وجود ثغرة بين قيم العزم hii لمعظم المشاهدات وقيمة (أو قيم) عزم كبيرة بصورة غير عادية.

- تطبيق (1-1) : -

في تجربة لعميل من قسم الميكانيكية سحبت عينة مكونة من 18 محرك من محركات توربينات الغاز حيث تم

قياس الجهد الناتج عن هذه المحركات في توليفات مختلفة من السرعة X1، و كذلك قياس تمديد الاستشعار X2، وكانت البيانات كالتالي :

التمديد =X2 (بوصة)	السرعة =X1 (بوصة/ثانية)	الجهد = Y فولت	العينة
0	6336	1.95	1
0	7099	2.5	2
0	8026	2.93	3
0	6230	1.69	4
0	5369	1.23	5
0	8343	3.13	6
0.006	6522	1.55	7
0.006	7310	1.94	8
0.006	7974	2.18	9
0.006	8501	2.7	10
0.012	6646	1.32	11
0.012	7384	1.6	12
0.012	8000	1.89	13
0.012	8545	2.15	14
0.018	6755	1.09	15
0.018	7362	1.26	16
0.018	7934	1.57	17
0.018	8554	1.92	18

المصدر :مركز الاستشارات الهندسية في جامعة فرجينيا ، معهد البوليتكنيك.

- التحليل الإحصائي :-

باستخدام برنامج الحزم الإحصائية Minitab سوف نقوم باستخراج قيم الرافعة لمصفوفة القبعة واكتشاف المشاهدات القاصية في السرعة X1 (بوصه/ثانية) والتمديد X2 (بوصه).

$$H_0: B_i = 0 \quad vs \quad H_1: B_i \neq 0 \quad i = 0,1,2$$

```
MTB > copy c2-c4 m2
MTB > Tran m2 m1
MTB > print m1

Data Display

 Matrix M1

   1     1     1     1     1     1     1.00     1.00     1.00     1.00     1.00
6336  7099  8026  6230  5369  8343  6522.00  7310.00  7974.00  8501.00  6646.00
   0     0     0     0     0     0     0.01     0.01     0.01     0.01     0.01

   1.00     1.00     1.00     1.00     1.00     1.00     1.00
7384.00  8000.00  8545.00  6755.00  7362.00  7934.00  8554.00
   0.01     0.01     0.01     0.02     0.02     0.02     0.02

MTB > Multiply M1 M2 m3.
MTB > Invert M3 m4.
MTB > Multiply M2 M4 m5.
MTB > Multiply M5 m1 m6.
MTB > Diagonal M6 c7.
MTB > prin c7
```

Data Display

```
C7
0.16813    0.129856   0.203451   0.180514   0.344897   0.258850
0.106380   0.060196   0.095191   0.171084   0.136976   0.076233
0.089473   0.149695   0.255876   0.186366   0.172565   0.214262

MTB > Name c4 "e(Y/X1,X2)" c5 "hii"
MTB > Regress 'Y' 2 'X1' 'X2';
SUBC>   Residuals 'e(Y/X1,X2)';
SUBC>   Hi 'hii';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis: Y versus X1; X2

```
The regression equation is
Y = - 1.64 + 0.000556 X1 - 67.4 X2
```

Predictor	Coef	SE Coef	T	P
Constant	-1.6413	0.2470	-6.64	0.000
X1	0.00055571	0.00003448	16.11	0.000
X2	-67.396	4.481	-15.04	0.000

S = 0.124548 R-Sq = 96.1% R-Sq(adj) = 95.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5.6942	2.8471	183.54	0.000
Residual Error	15	0.2327	0.0155		
Total	17	5.9269			

Source	DF	Seq SS
X1	1	2.1849
X2	1	3.5094

حيث أن قيمة المتوسط لقيم العزم تساوي:

$$\bar{h} = \frac{3}{18} = 0.166667$$

وقيمة ضعف متوسط قيم العزم تساوي :

$$\frac{2*p}{n} = \frac{2*3}{18} = 0.333333$$

- جدول (1-1) قيم العزم والمشاهدات القاصية في المتغيرات المستقلة المؤثرة على الجهد.

I	Y	X1	X2	e(Y/X1,X2)	hii	big (hii)
1	1.95	6336	0	0.070315	0.168134	
2	2.5	7099	0	0.196309	0.129856	
3	2.93	8026	0	0.111166	0.203451	
4	1.69	6230	0	-0.13078	0.180514	
5	1.23	5369	0	-0.112315	0.344897	0.344897
6	3.13	8343	0	0.135006	0.25885	
7	1.55	6522	0.006	-0.028672	0.10638	
8	1.94	7310	0.006	-0.076571	0.060196	
9	2.18	7974	0.006	-0.205562	0.095191	
10	2.7	8501	0.006	0.02158	0.171084	
11	1.32	6646	0.012	0.076795	0.136976	
12	1.6	7384	0.012	-0.053318	0.076233	
13	1.89	8000	0.012	-0.105635	0.089473	
14	2.15	8545	0.012	-0.148496	0.149695	
15	1.09	6755	0.018	0.190598	0.255876	
16	1.26	7362	0.018	0.023283	0.186366	
17	1.57	7934	0.018	0.015417	0.172565	
18	1.92	8554	0.018	0.020878	0.214262	

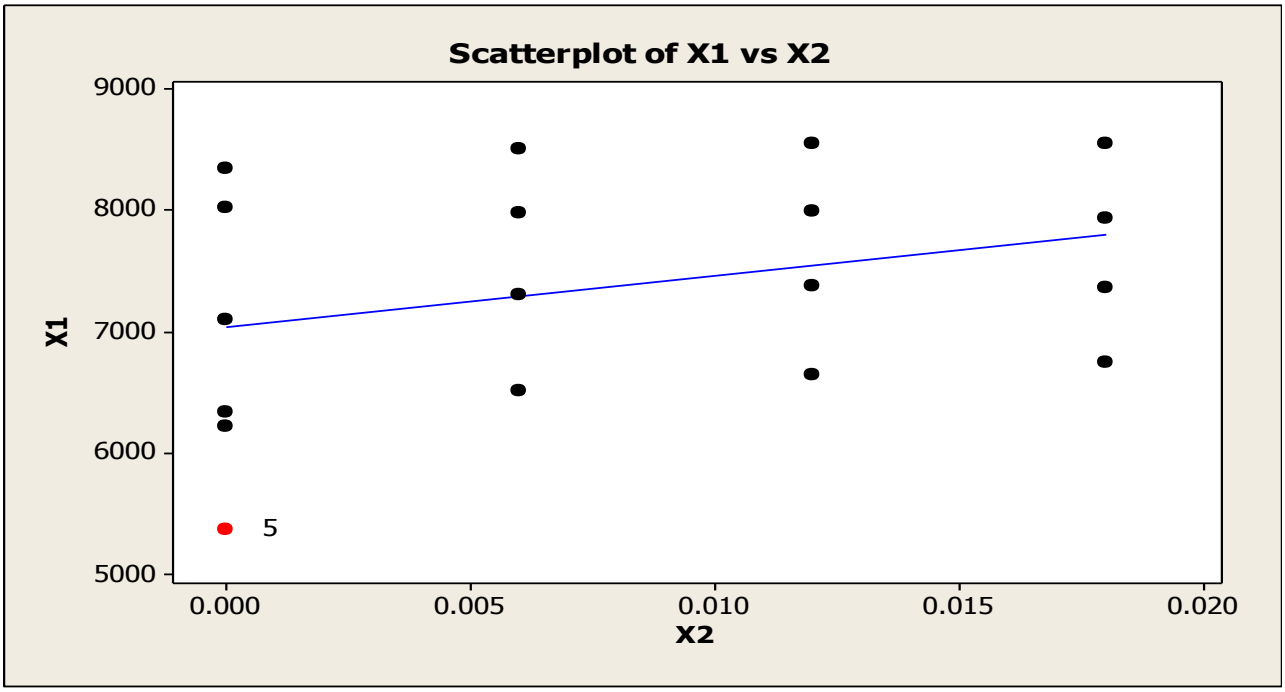
- الاستنتاج :

نلاحظ من الجدول (1-1) ، أن هناك نقطة شاذة واحدة في المتغيرات المستقلة أي أنها أكبر من ضعف متوسط قيم العزم حيث تشير إلى أنها مشاهدة قاصية في بيانات السرعة أو التمديد و قد تؤدي إلى تأثير في تحليل الانحدار الخطي وهي المشاهدة رقم 5 حيث أن قيمة العزم لها تساوي:

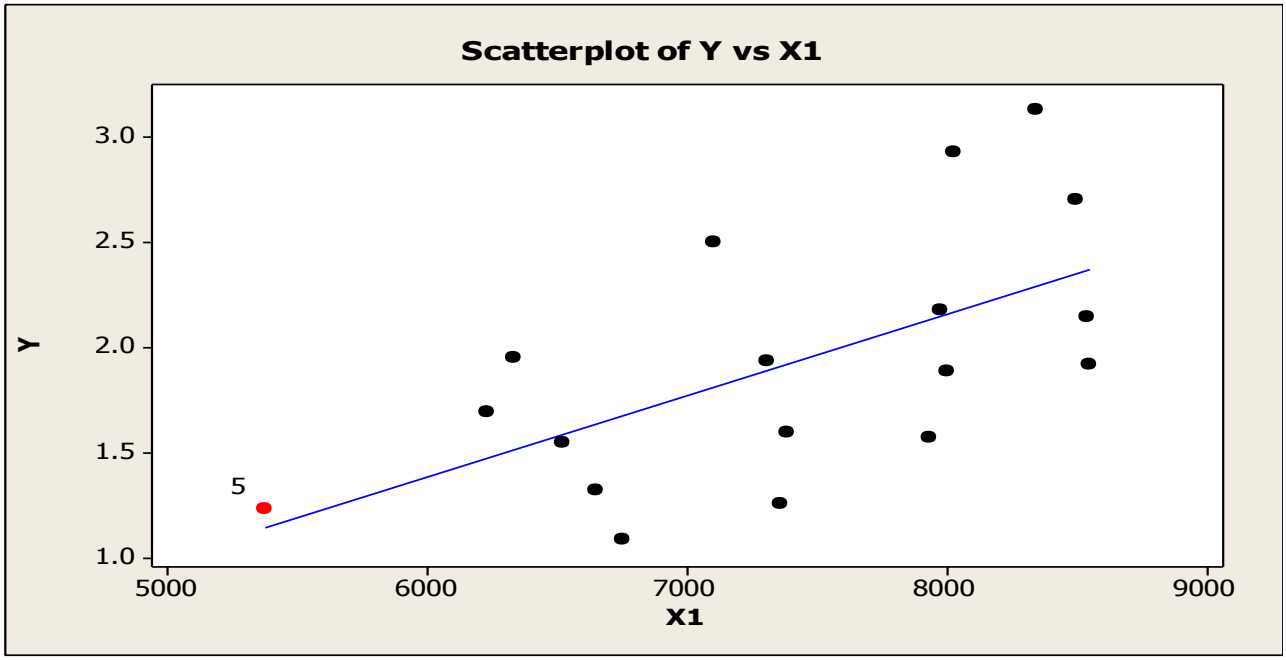
$$h_{5,5} = 0.344897$$

وقد قمنا برسم شكل (1-1) لانتشار المتغيرات على بعضها وتحديد المشاهدة القاصية باللون الأحمر أما المشاهدات الأخرى بدوائر باللون الأسود ونلاحظ أن المشاهدة القاصية غير منسجمة مع بقية النقاط الأخرى.

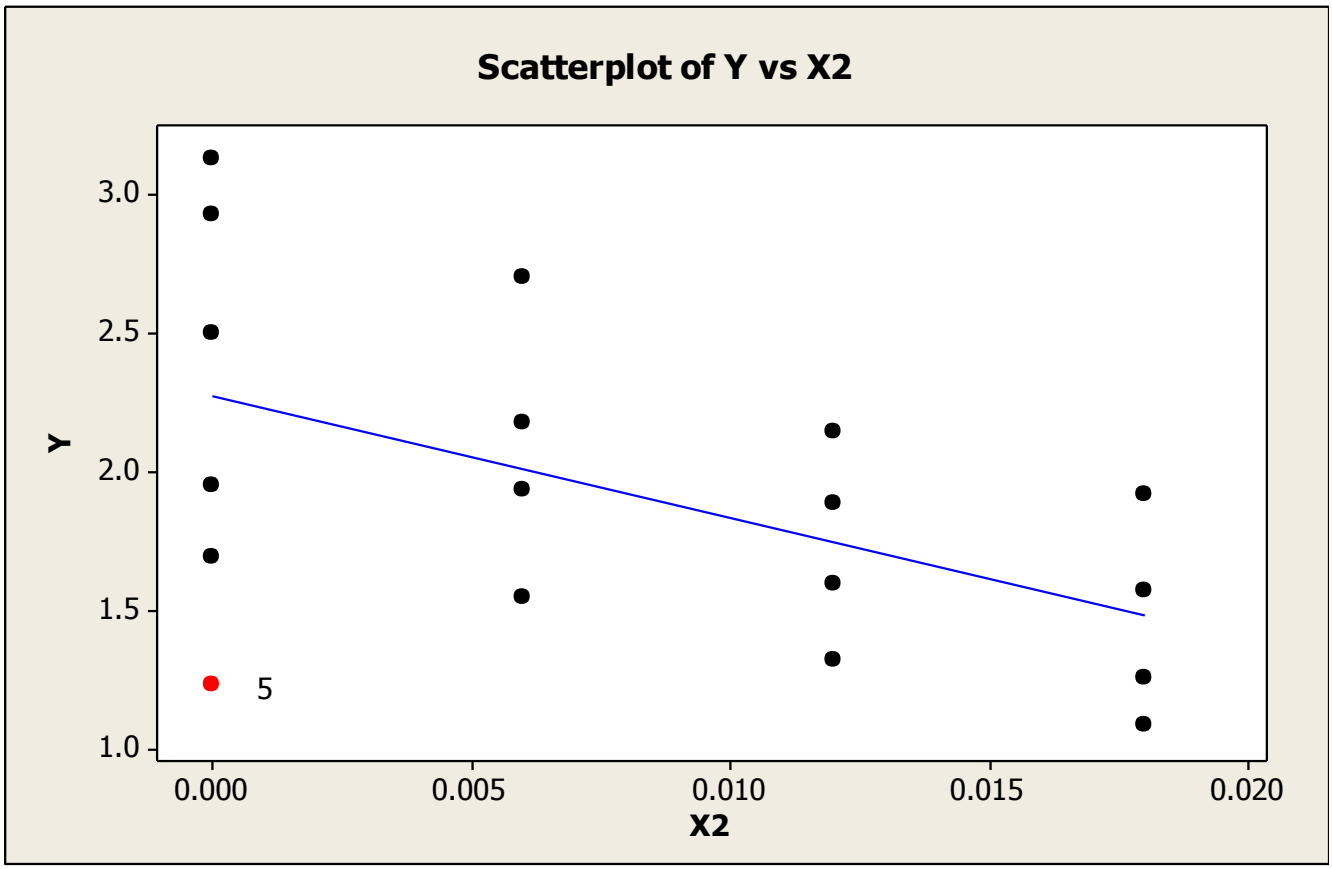
- شكل (1-1) : رسم انتشار السرعة (بوصه/ثانية) X1 على التمديد (بوصه) X2.



- شكل (1-2) :رسم انتشار الجهد (فولت) Y على السرعة (بوصه/ثانية) X1.

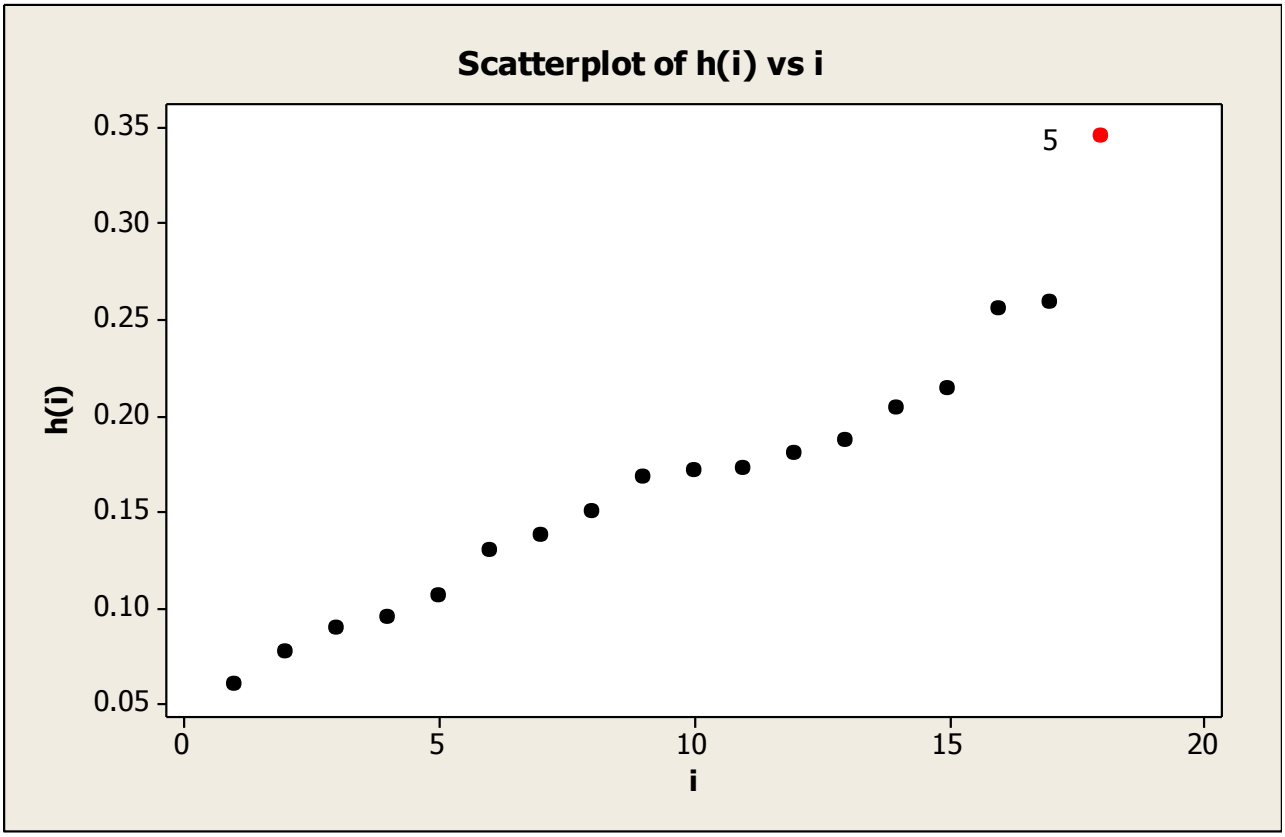


- شكل (3-1) رسم انتشار الجهد (فولت) Y على التمديد (بوصه) X2.



نلاحظ من رسومات الانتشار السابقة للمتغيرات أننا لا نستطيع بسهولة اكتشاف جميع القيم القاصية في النموذج و يرجع ذلك لأن النموذج يعتمد على أكثر من متغير . لذلك فإن طريقة اكتشاف المشاهدات القاصية بواسطة قيم العزم و الطرق الأخرى أفضل من طريقة مشاهدة القيم من خلال الرسم.

- شكل (4-1): رسم انتشار (الترتيب التصاعدي لقيم العزم) h(i) مقابل أرقام المشاهدات i.



نلاحظ أن قيمة العزم للمشاهدة رقم 5 تساوي 0.344897 وتفصلها فجوة تساوي 0.17823 عن متوسط قيم العزم.

- تطبيق (2-1) : -

في بيانات لمشروع تنفيذي اختيرت 33 شركة لمعرفة تكلفة أو تعويضات هذا المشروع (بآلاف الدولارات) ، وبعض المتغيرات المؤثرة عليه مثل المبيعات (بملايين الدولارات) ، والأرباح (بملايين الدولارات)، وتكلفة العمالة وكانت البيانات كما يلي:-

تكلفة العمالة X3	الأرباح X2	المبيعات X1	التكلفة "التعويضات " Y	الشركة
48,000	128.1	4600.6	450	1
55,900	783.9	9,255.40	387	2
13,783	136	1,526.20	368	3
27,765	179	1,683.20	277	4
3,4000	231.5	2,752.80	676	5
2,6500	329.5	2,205.80	454	6
30,800	381.8	2,384.60	507	7
41,000	237.9	2,746.00	496	8
25,900	222.3	1,434.00	487	9
8,600	63.7	470.6	383	10
21,075	149.5	1,508.00	311	11
6,874	30	464.4	271	12
39,000	577.3	9,329.30	524	13
34,300	250.7	2,377.50	498	14
19,405	82.6	1,174.30	343	15
3,586	61.5	409.3	354	16
3,905	90.8	724.7	324	17
4,139	63.3	578.9	225	18
6,255	42.8	966.8	254	19
10,605	48.5	591	208	20
65,392	310.6	4,933.10	518	21
89,400	491.6	7,613.20	406	22
55,200	228	3,457.40	332	23
7,800	54.6	545.3	340	24
337,119	3011.3	22,862.80	698	25
52,000	203	2,361.00	306	26
50,500	201	2,614.10	613	27
18,625	121.3	1,013.20	302	28
97,937	194.6	4,560.30	540	29
12,300	63.4	855.7	293	30
71,800	352.1	4,211.60	528	31
87,700	655.2	5,440.40	456	32
14,600	97.5	1,229.90	417	33

- التحليل الإحصائي :-

باستخدام برنامج الحزم الإحصائية Minitab سوف نقوم باستخراج قيم العزم لمصفوفة القبعة واكتشاف المشاهدات القاصية في المتغيرات المستقلة (المبيعات ، الأرباح ، العمالة).

$$H_0: B_i = 0 \qquad vs \qquad H_1: B_i \neq 0 \qquad i = 0,1,2,3$$

```
MTB > Regress 'Y' 3 'X1' 'X2' 'X3';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis: Y versus X1; X2; X3

```
The regression equation is
Y = 344 + 0.0141 X1 - 0.127 X2 + 0.00135 X3


Predictor        Coef    SE Coef      T      P
Constant       344.45      25.22  13.66  0.000
X1            0.01413    0.01418   1.00  0.327
X2            -0.1270     0.1309  -0.97  0.340
X3          0.0013506  0.0009734   1.39  0.176


S = 103.860   R-Sq = 37.2%   R-Sq(adj) = 30.7%


Analysis of Variance

Source          DF      SS     MS     F      P
Regression       3  185131  61710  5.72  0.003
Residual Error  29  312821  10787
Total           32  497952


Source  DF  Seq SS
X1       1  163582
X2       1     784
X3       1   20766

Unusual Observations

Obs     X1      Y    Fit  SE Fit  Residual  St Resid
  5   2753  676.0  399.9    18.3     276.1     2.70R
 13   9329  524.0  455.6    75.2      68.4     0.95 X
 25  22863  698.0  740.3    99.9     -42.3    -1.49 X
 29   4560  540.0  516.5    65.2      23.5     0.29 X
```

نلاحظ أن أحد المتغيرات المستقلة غير معنوي ولذلك نقوم بحذف X2 (الأرباح) الذي يقابل أكبر قيمة من قيم p-value وبعد حذف هذا المتغير نقوم بتوفيق دالة الانحدار Y على المتغيرين المستقلين الآخرين X1,X3 ، فحصلنا على:

```
MTB > Regress 'Y' 2 'X1' 'X3';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis: Y versus X1; X3

```
The regression equation is
Y = 354 + 0.0059 X1 + 0.000845 X3


Predictor       Coef    SE Coef      T      P
Constant      354.47      22.99  15.42  0.000
X1           0.00594    0.01139   0.52  0.606
X3         0.0008450  0.0008214   1.03  0.312


S = 103.759   R-Sq = 35.1%   R-Sq(adj) = 30.8%


Analysis of Variance

Source          DF      SS     MS     F      P
Regression       2  174974  87487  8.13  0.002
Residual Error  30  322978  10766
Total           32  497952


Source  DF  Seq SS
X1       1  163582
X3       1   11392

Unusual Observations


Obs     X1      Y    Fit  SE Fit  Residual  St Resid
  2   9255  387.0  456.7    60.9     -69.7     -0.83 X
  5   2753  676.0  399.6    18.3     276.4      2.71R
 13   9329  524.0  442.9    74.0      81.1      1.12 X
 25  22863  698.0  775.2    93.1     -77.2     -1.69 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

نلاحظ أن أحد المتغيرات المستقلة غير معنوي ولذلك نقوم بحذف X1 (المبيعات) الذي يقابل أكبر قيمة من قيم p-value وبعد حذف هذا المتغير نقوم بتوفيق دالة الانحدار Y على المتغير المستقل المتبقي X3 ، فحصلنا على:

```
MTB > Name c6 "e(Y/x3)" c7 "hii"
MTB > Regress 'Y' 1 'X3';
SUBC>   Hi 'hii';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis: Y versus X3

```
The regression equation is
Y = 357 + 0.00124 X3


Predictor       Coef    SE Coef      T      P
Constant      356.98      22.22  16.07  0.000
X3         0.0012418  0.0003070   4.05  0.000


S = 102.534   R-Sq = 34.5%   R-Sq(adj) = 32.4%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  172041  172041  16.36  0.000
Residual Error  31  325911   10513
Total           32  497952


Unusual Observations

Obs      X3      Y    Fit  SE Fit  Residual  St Resid
  5   34000  676.0  399.2    18.1     276.8     2.74R
 25  337119  698.0  775.6    92.0     -77.6    -1.72 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

حيث أن قيمة المتوسط لقيم العزم تساوي:

$$\bar{h} = \frac{2}{33} = 0.0606061$$

وقيمة ضعف متوسط قيم العزم تساوي :

$$\frac{2 * p}{n} = \frac{2 * 2}{33} = 0.121212$$

- جدول (2-1)) قيم العزم والمشاهدات القاصية في المتغيرات المستقلة على التكلفة.

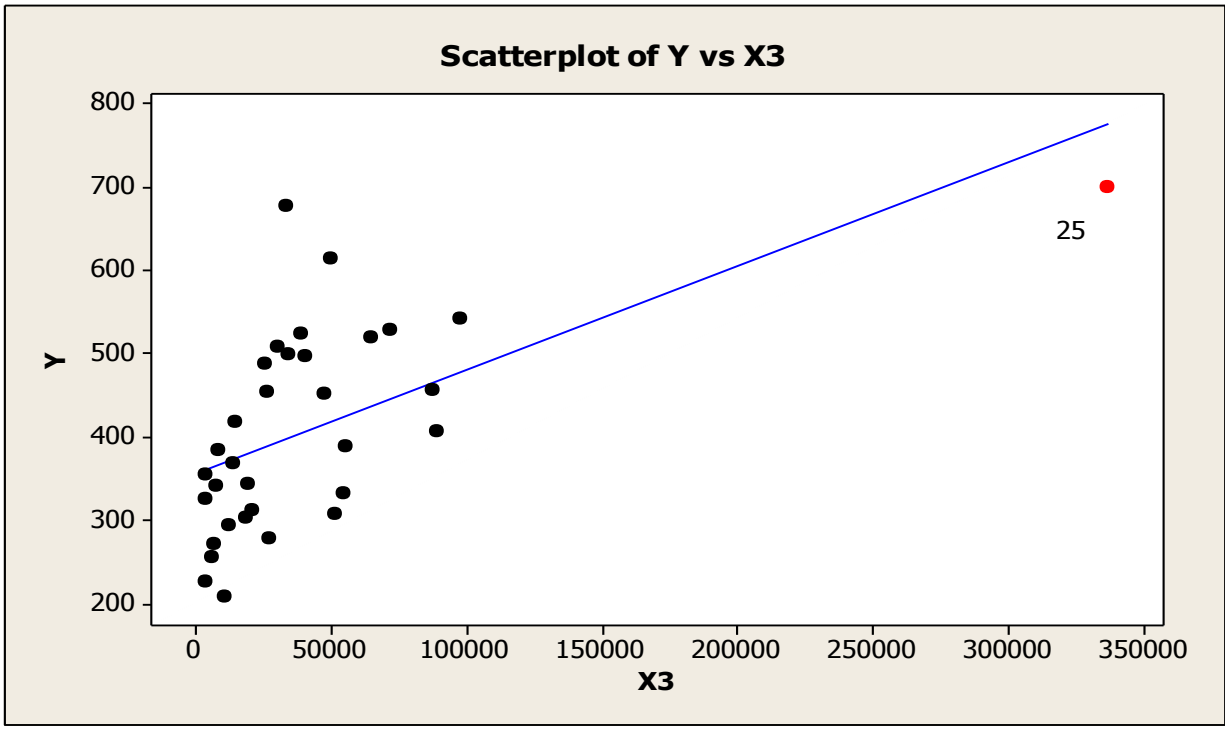
الشركة	التكلفة "التعويضات " Y	العمالة X3	e(Y/X)	hii	big hii
1	450	48,000	33.41	0.03052	
2	387	55,900	-39.4	0.031775	
3	368	13,783	-6.098	0.037999	
4	277	27,765	-114.461	0.032407	
5	676	34,000	276.796	0.031043	
6	454	26,500	64.11	0.032768	
7	507	30,800	111.77	0.031656	
8	496	41,000	88.103	0.030342	
9	487	25,900	97.855	0.03295	
10	383	8,600	15.338	0.040962	
11	311	21075	-72.154	0.034645	
12	271	6,874	-94.518	0.042056	
13	524	39,000	118.587	0.030453	
14	498	34,300	98.423	0.030995	
15	343	19,405	-38.08	0.035329	
16	354	3,586	-7.435	0.044287	
17	324	3,905	-37.831	0.044062	
18	225	4,139	-137.122	0.043899	
19	254	6,255	-110.75	0.042461	
20	208	10,605	-162.152	0.039759	
21	518	65,392	79.812	0.034764	
22	406	89,400	-62.002	0.049532	
23	332	55,200	-93.531	0.031619	
24	340	7,800	-26.668	0.041463	
25	698	337,119	-77.628	0.80529	0.80529
26	306	52,000	-115.557	0.031016	

الشركة	التكلفة "التعويضات " Y	العمالة X3	e(Y/X)	hii	big hii
27	613	50,500	193.305	0.030796	
28	302	18,625	-78.111	0.035666	
29	540	97,937	61.396	0.057274	
30	293	12,300	-79.256	0.038798	
31	528	71,800	81.854	0.037695	
32	456	87,700	-9.891	0.048147	
33	417	14,600	41.887	0.037576	

- الاستنتاج :-

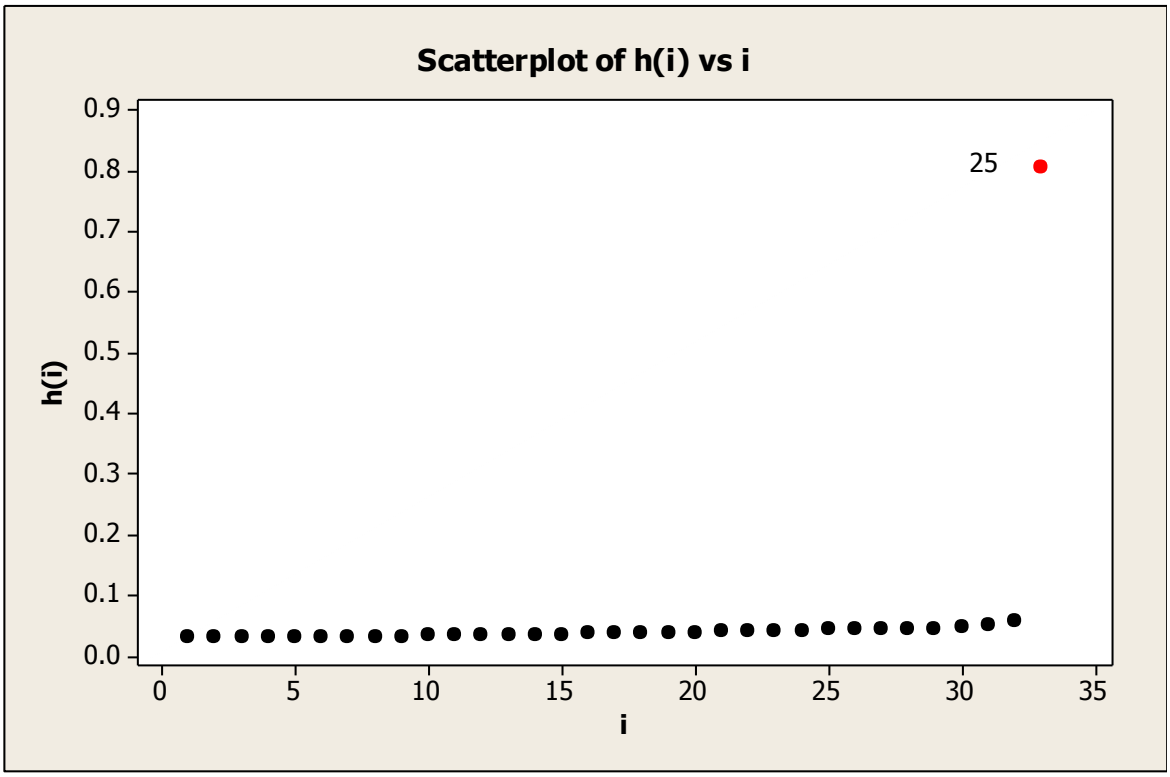
نلاحظ كما في الجدول (2-1) ، وجود نقطة واحدة قاصية (المشاهدة رقم 25) وهي تفوق ضعف متوسط قيم العزم ، مع ملاحظة أنها قيمة كبيرة جدا وذلك يدل على أنها قاصية بشكل كبير وتفصلها ثغرة كبيرة عن البيانات والمشاهدات الأخرى.

- شكل (5-1) رسم انتشار التكلفة Y على العمالة X3 .



نلاحظ من الرسم السابق وجود نقطة شاذة ولكننا لا نستطيع اكتشاف كل النقاط الشاذة من الرسم لذلك فإن طريقة اكتشاف القيم القاصية بواسطة قيم العزم والطرق الأخرى أفضل من طريقة الانتشار.

- شكل (6-1) رسم انتشار(الترتيب التصاعدي لقيم العزم) h(i) مقابل أرقام المشاهدات i .



نلاحظ أن قيمة العزم للمشاهدة رقم 25 تساوي $h_{25,25} = 0.80529$ تفصلها فجوة تساوي 0.7446839 عن قيمة متوسط العزم.