الفصل الرابع

مقدمة في تحليل الإنحدار الخطي

بعد الانتهاء من دراسة هذا الفصل سيكون القارئ قادراً على ما يلي:

- تعريف الانحدار الخطي البسيط
- تعريف الانحدار الخطي المتعدد
- تطوير نموذج الانحدار الخطي البسيط والمتعدد بناء على بيانات معطاة
- تحديد صلاحية وجودة نموذج الانحدار الخطي البسيط والمتعدد
- إستخدام برنامج EXCEL في تطوير نموذج التقدير والقيام بالتقدير

مقدمة

في هذا الفصل سيتم مناقشة مفهوم الإنحدار واهمية نماذج الإنحدار الخطي ليس فقط في دراسة وتحديد العلاقات بين المتغيرات وإنما في تقدير أحد المتغيرات بمعلومية المتغيرات الأخرى. وقد تزايدت أهمية نماذج الإنحدار باعتبارها من اهم الأدوات في صناعة القرار. كما سيتطرق هذا الفصل الى كيفية بناء نماذج الإنحدار وتحديد صلاحيتها في دراسة الظواهر. ثم ستتم مناقشة كيفية تحديد جودة نموذج الإنحدار الذي تم بناؤه وتحديد مدى كفاءته في دراسة الظاهرة قيد الدراسة. كما يُعتبر تحليل الانحدار كذلك من أكثر واهم الطرق الإحصائية التي تستخدم في البحث العلمي.

أهمية نماذج الإنحدار The Importance of Regression Models

تُعتبر نماذج تحليل الانحدار من اهم الأدوات المستخدمة في صناعة القرار عند المدراء في العصر الحاضر. حيث تُستخدم في دراسة العلاقات والتأثيرات بين الكثير من العوامل والمتغيرات وتقدير أحد المتغيرات بناءً على معرفة متغيراتٍ اخرى. كما تقيس كيفية تأثير التغير في المتغير المستقل (المتغيرات المستقلة) على المتغير التابع.

فعلى سبيل المثال تُستخدم نماذج الإنحدار في دراسة العلاقة بين مستوى التعليم والدخل ، وسعر المنزل ومساحته ، وحجم المبيعات للشركة والمبالغ التي تنفقها الشركة على الإعلانات. كما تستخدم الشركة نماذج الإنحدارعندما تحاول تحديد الموقع الأفضل لإنشاء متجر جديد أو مكتب فرعي. كما ان الشركات غالبا ما تستخدم نماذج الإنحدار عند تقدير التكلفة.

كما وضحنا سابقاً، هناك غرضان لتحليل الانحدار. الأول هو فهم العلاقة بين المتغيرات مثل نفقات الإعلان والمبيعات. الغرض الثاني هو التنبؤ بقيمة متغير واحد بناء على قيمة الآخر. ولهذا السبب، يُعد الانحدار تقنية تنبؤ مهمة للغاية في هذا الفصل، سيتم تطوير نموذج الانحدار الخطي البسيط أولاً ، ثم سيتم استخدام نموذج انحدار متعدد أكثر تعقيداً لدمج المزيد من المتغيرات في النموذج. في أي نموذج انحدار، يسمى المتغير المراد التنبؤ به المتغير التابع أو متغير الاستجابة ونرمز له ب Y. ويقال إن قيمة هذا تعتمد على قيمة متغير مستقل، والذي يسمى أحيانا متغيرا تفسيرياً أو متغيرا تنبؤياً ونرمز له ب X.

بشكلٍ عام، هناك خطوتان أساسيتان في تحليل الإنحدار (البسيط او المتعدد) هما:

الخطوة الأولى: بناء وتطوير معادلة الإنحدار الخطي (تطوير نموذج التقدير)

الخطوة الثانية: القيام بعملية التقديربإستخدام النموذج الذي تم تطويره وقياس صلاحيته وجودته.

لماذا نستخدم تحليل الإنحدار؟ Why to use Regression Analysis?

كما أسلفنا سابقاً، نستخدم تحليل الإنحدار من أجل ما يلي:

1. تقدير قيمة المتغير التابع بناءً على قيمة المتغير المستقل.
2. تحديد التأثير الكبير الذي يحدثه المتغير المستقل على المتغير التابع.
3. تحديد قوة العلاقة بين المتغير المستقل والمتغير التابع.

الرسوم النقطية Scatter Diagram

تُعتبر الرسوم النقطية من أهم الأدوات التي تستخدم في دراسة العلاقة بين زوجين من البيانات، نستخدم الرسوم النقطية أو Scatter Diagram والذي يمثل كل متغير على محور من محاوره ويتم رسم أزواج البيانات لتحديد العلاقة بينها. إذا كانت المتغيرات مترابطة، فستقع النقاط على طول خط أو منحنى. كلما كان الارتباط أقوى، كلما كانت النقاط أقرب الى الخط او المنحنى الذي يحدد العلاقة بين المتغيرين.

للتحقيق في العلاقة بين المتغيرات، من المفيد إلقاء نظرة على الرسم النقطي لبيانات المتغيرات. عادةً ما يتم رسم المتغير المستقل على المحور الأفقي ويتم رسم المتغير التابع على المحور الرأسي.

مثال 1

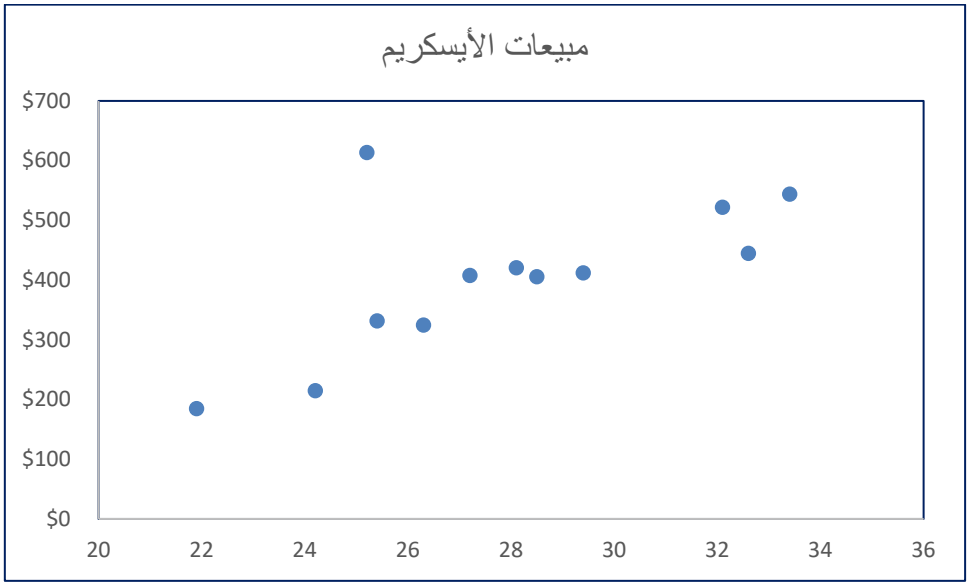
يتتبع متجر الآيس كريم المحلي كمية الآيس كريم التي يبيعونها في فصل الصيف وعلاقتها مع درجة الحرارة في كل يوم. وقد تم حصر المبيعات للأيس كريم لعدد 12 يوماً مع تحديد درجة الحرارة لكل يوم ، وكانت البيانات كالتالي:

مبيعات الأيسكريم	درجة الحرارة C°
$215	24.2
$325	26.3
$185	21.9
$332	25.4
$406	28.5
$522	32.1
$412	29.4
$614	25.2
$544	33.4
$421	28.1
$445	32.6
$408	27.2

والمطلوب رسم مخطط الرسوم النقطية التي تمثل العلاقة بين درجة حرارة الجو ومبيعات الآيسكريم في ذلك اليوم.

الحل

نمثل درجة الحرارة على محور X ومبيعات الأيسكريم على محور Y، فنحصل على الرسم النقطي المطلوب التالي



ونلاحظ من الرسم ان العلاقة بينهما طردية او موجبة حيث نلاحظ ارتفاع مبيعات الأيسكريم كلما إرتفعت درجة الحرارة في ذلك اليوم.

تحليل الإرتباط Correlation Analysis

معامل الارتباط هو مقياس لقوة العلاقة بين المتغيرات المستقلة والتابعة. هناك أنواع كثيرة من معاملات الارتباط بحسب طبيعة المتغيرات. ما سنركز عليه هنا هو معامل ارتباط بيرسون لقياس العلاقة بين متغيرين مستمرين. معامل ارتباط بيرسون، المعروف أيضًا باسم اختبار بيرسون الإحصائي، يقيس القوة بين المتغيرات المختلفة وعلاقاتها. يُعطى معامل إرتباط بيرسون بالعلاقة التالية

$$r = \frac{\sum(Y-\bar{Y})(X-\bar{X})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} = \frac{n\sum XY - \sum X\sum Y}{\sqrt{\left[n\sum X^2-(\sum X)^2\right]\left[n\sum Y^2-(\sum Y)^2\right]}}$$

ويمكن كتابتها كما يلي

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

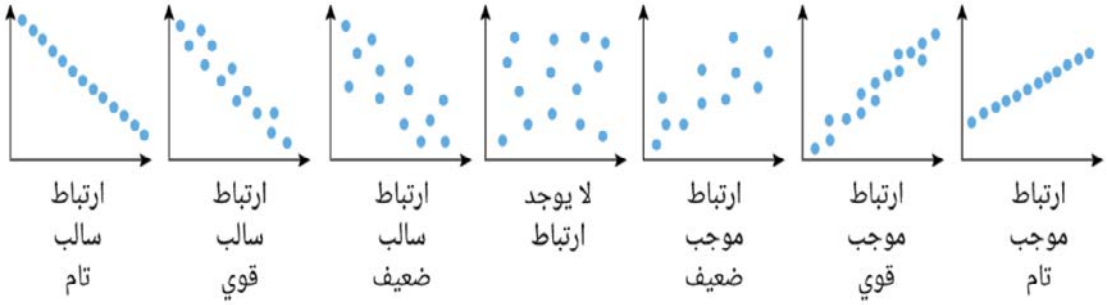
حيث

$$\sum(\mathrm{Y}-\overline{\mathrm{Y}})(\mathrm{X}-\overline{\mathrm{X}}) \qquad S_{XY}$$

$$\sum(\mathrm{X}-\overline{\mathrm{X}})^2 \qquad S_{XX}$$

$$\sum(\mathrm{Y}-\overline{\mathrm{Y}})^2 \qquad S_{YY}$$

تتراوح قيمة r بين -1.00 و +1.00 ، حيث أن قيمة ±1.00 تشير إلى وجود علاقة خطية تامة بين المتغيرات. و -1.00 تشير الى وجود علاقة عكسية تامة بين المتغيرات. كما تشير قيمة r القريبة من الصفر إلى وجود علاقة خطية ضعيفة أومعدومة بين المتغيرات. الشكل التالي يوضح الرسوم النقطية لبعض أنواع العلاقات بين المتغيرات.



الإنحدار الخطي البسيط Simple Linear Regression

في أي نموذج انحدار، هناك افتراض ضمني بوجود علاقة بين المتغيرات. هناك أيضا بعض الأخطاء العشوائية التي لا يمكن التنبؤ بها. يُعطى نموذج الانحدار الخطي البسيط بالصيغة التالية

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}$$

حيث

Y المتغير التابع أو متغير النتيجة أو الإستجابة او الظاهرة

X المتغير المستقل او المتغير التفسيري او السبب

β_0 ثابت النموذج وهو قيمة Y عندما تكون X صفراً

β_1 ميل خط الإنحدار الخطي (معدل تغير Y لوحدة التغير في X)

ε الخطأ العشوائي

ويُسمى هذا الإنحدار بالإنحدار الخطي البسيط لأن المتغير التابع Y يرتبط خطياً بمتغير واحد فقط هو المتغير المستقل X.

من أجل تطوير معادلة الإنحدار الخطي نحتاج أن نقوم بتقدير الثابت $\boldsymbol{\beta_0}$ وميل خط الإنحدار $\boldsymbol{\beta_1}$ بناءً على بيانات العينة، وتكون معادلة الإنحدار الخطي بناءً على بيانات العينة كما يلي

$$\hat{Y} = a_0 + b_1 X$$

حيث

$\hat{Y}$ القيمة التقديرية ل Y بناءً على نموذج التقدير

a_0 تقدير ل $\boldsymbol{\beta_0}$ بناءً على نتائج العينة

b_1 تقدير ل $\boldsymbol{\beta_1}$ بناءً على نتائج العينة

حيث أن الخطأ في التقدير هو الفرق بين القيمة الملاحظة والقيمة المقدرة بواسطة نموذج التقدير $\varepsilon = Y - \hat{Y}$ فسيتم إعتبار النموذج الأفضل هو النموذج الذي يقلل قيمة هذا الخطأ. وحيث ان الأخطاء بعضها موجبة وبعضها سالبة فسيتم إيجاد قيم الثابت والميل بناءً على ما يلي

$$Min \sum \varepsilon^2 = Min \sum (Y - \hat{Y})^2 = Min \sum (Y - a_0 - b_1 X)^2$$

حتى تحصل على القيمة الصغرى لمربعات الأخطاء(Ordinary Least Squares (OLS يجب أن يتم حساب قيمة الميل والثابت في معادلة الإنحدار بالصيغ التالية

$$\mathrm{b_1} = \frac{\sum(\mathrm{Y} - \overline{\mathrm{Y}})(\mathrm{X} - \overline{\mathrm{X}})}{\sum(\mathrm{X} - \overline{\mathrm{X}})^2} = \frac{\sum \mathrm{XY} - \mathrm{n}\overline{\mathrm{X}}\,\overline{\mathrm{Y}}}{\sum \mathrm{X}^2 - \mathrm{n}\overline{\mathrm{X}}^2}$$

$$\mathrm{a_0} = \overline{\mathrm{Y}} - \mathrm{b_1}\overline{\mathrm{X}}$$

كما أن إشارة الميل b_1 هي التي تحدد طبيعة العلاقة بين المتغير التابع والمتغير المستقل وكذا تحدد معدل التغير في الظاهرة *Y* لوحدة التغير في *X*.

مثال 2

لاحظت شركة السهم أن مبيعاتها لمنتج معين تعتمد على شمول وفعالية الحملات التسويقية التي تقوم بها، وقد قام فريق تطوير الاعمال بجمع البيانات المتعلقة بهذين المتغيرين لأجل دراسة العلاقة بينهما. البيانات التي تم تجميعها موضحة في الجدول التالي

المبيعات (وحدة $100,000)	تكلفة الإعلانات (وحدة $10,000)
3	6
4	8
6	9
4	5
2	4
5	10

والمطلوب

A. حدد المتغير المستقل والمتغير التابع مع توضيح السبب؟

B. ارسم مخطط الرسوم النقطية لتوضيح العلاقة بين المتغيرين

C. أوجد معادلة الإنحدار الخطي؟ ثم قم بوصف العلاقة بين تكلفة الإعلانات والمبيعات للشركة؟

D. قم بتقدير المبيعات المتوقعة إذا علمت ان تكلفة الحملة الإعلانية التي نفذتها الشركة هي 70,000 دولاراً؟

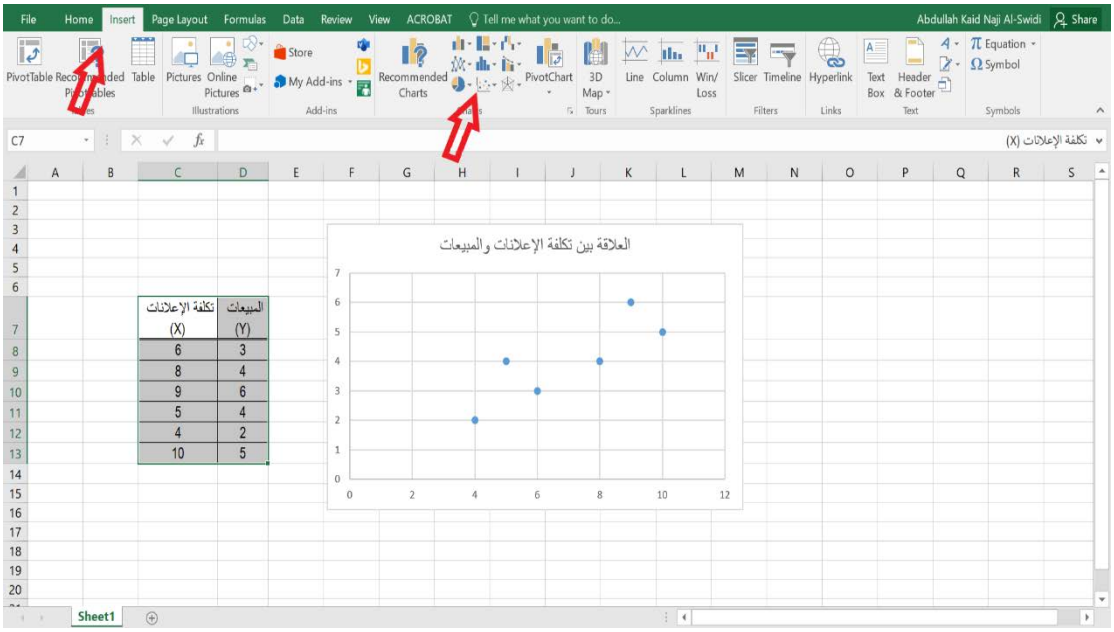
E. ما هو معدل التغير في المبيعات إذا زادت تكلفة الإعلانات ب 10,000 دولاراً؟

الحل

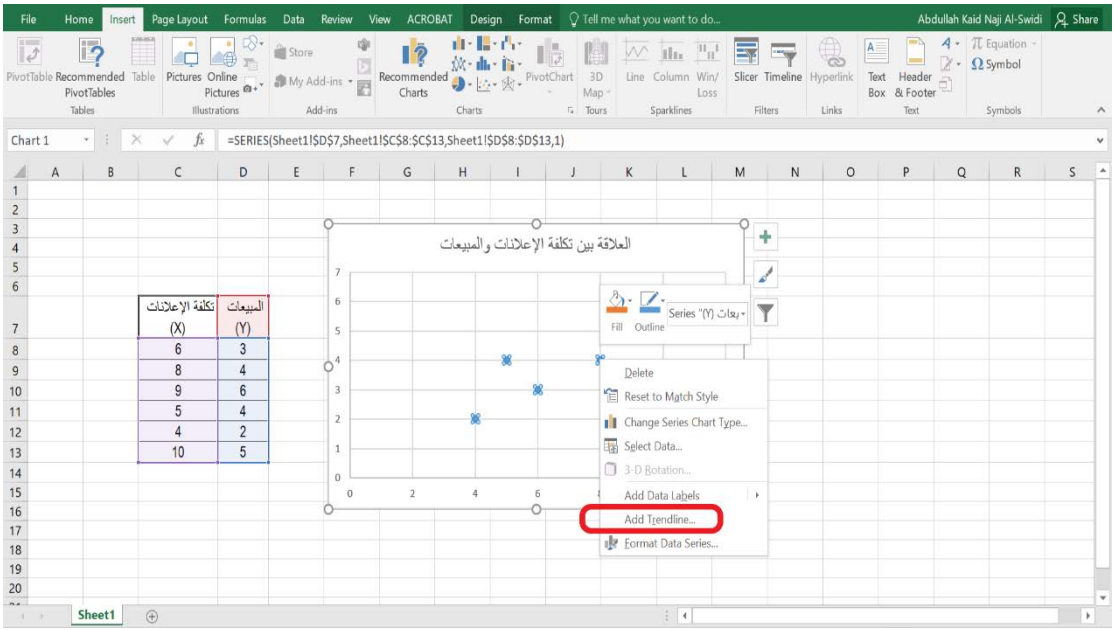
A. المتغير المستقل هو تكلفة الإعلانات حيث هو متغير السبب، والمتغير التابع هو المبيعات حيث يمثل متغير النتيجة.

B. يتم رسم العلاقة بين تكلفة الإعلانات والمبيعات بإستخدام تطبيق Excel كما يلي

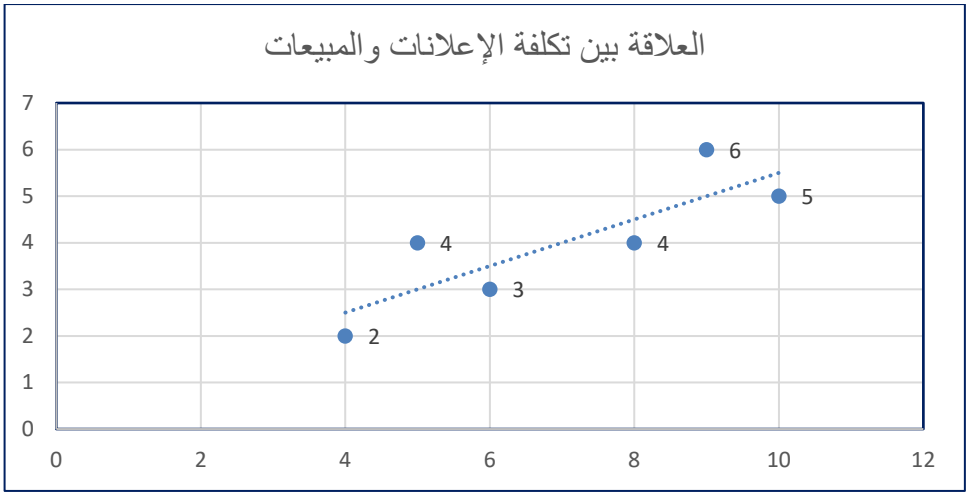
حيث يتم إختيار قائمة insert ثم إختيار الرسوم النقطية



كما يمكن تمثيل البيانات بخط الإنحدار من خلال الضغط بالزر الأيمن على نقط البيانات في الرسم النقطي وإختيار Add trendline



فنحصل على خط التباين كما في الرسم التالي



ويتضح من الرسم ان العلاقة طردية أو موجبة بمعنى أن المبيعات تزداد بزيادة الإستثمار في الإعلانات كما سيتم توضيحها فيما بعد.

C. يتم إشتقاق معادلة الانحدار الخطي كما يلي

نقوم بإيجاد قيمة الثوابت التالية من اجل تحديد معدلة الانحدار $\bar{Y} = a_0 + b_1 X$

$$b_1 = \frac{\sum(Y-\bar{Y})(X-\bar{X})}{\sum(X-\bar{X})^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$a_0 = \bar{Y} - b_1\bar{X}$$

تكلفة الإعلانات (X)	المبيعات (Y)	XY	X^2
6	3	18	36
8	4	32	64
9	6	54	81
5	4	20	25
4	2	8	16
10	5	50	100
$\sum$ المجموع	42	24	182	322

وعليه فإن $\bar{X} = \frac{42}{6} = 7$ ، $\bar{Y} = \frac{24}{6} = 4$

$$b_1 = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{182 - 6*4*7}{322 - 6*49} = 0.5$$

$$a_0 = 4 - 0.5 * 7 = 0.5$$

وحيث ان إشارة الميل موجبة، فالعلاقة بين المتغيرين طردية او موجبة. ويتم كتابة معادلة الانحدار الخطي كما يلي

$$\hat{Y} = 0.5 + 0.5\,X$$

D. بحسب المطلوب وهو تقدير المبيعات إذا كانت تكلفة الإعلانات 70,000 دولاراً(X=7).

$$\hat{Y} = 0.5 + 0.5 * 7 = 4$$

والذي يعني أن المبيعات المقدرة باستخدام نموذج الإنحدار الخطي ستكون 400,000 دولاراً.

E. لإيجاد معدل التغير في المبيعات إذا زادت تكلفة الإعلانات ب 10,000 دولاراً (أي بمقدار 1). حيث انه من المعلوم ان معدل تغير Y لوحدة تغير X هو الميل والذي يساوي (b1=0.5) وحيث ان العلاقة بين المتغيرين طردية، فيمكن القول أنه إذا زادت الشركة تكلفة الإعلانات بمقدار الوحدة (10,000دولاراً) فإن الزيادة في المبيعات ستكون بمقدار نصف الوحدة (50,000 دولاراً). ويمكن إثبات ذلك كما يلي

$$if\ X = 2 \rightarrow \hat{Y} = 0.5 + 0.5 * 2 = 1.5$$

$$if\ X = 3 \rightarrow \hat{Y} = 0.5 + 0.5 * 3 = 2$$

والفرق بينهما هو 0.5 كما يشير إلى ذلك ميل خط الإنحدار.

تحديد معامل التحديد

Coefficient of Determination

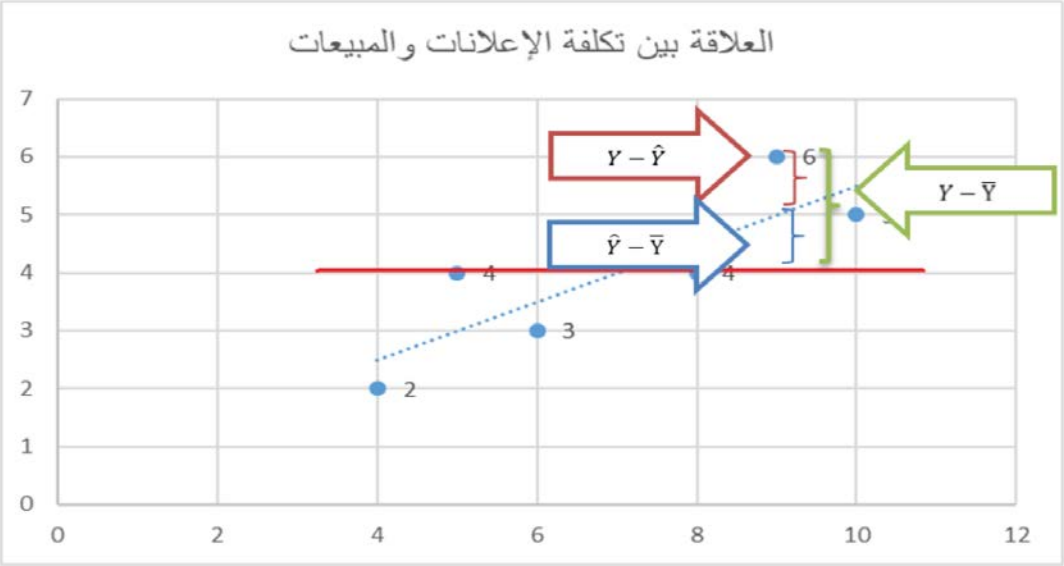
بالنظر في المثال السابق ومن اجل تحليل الخطأ نجد ما يلي

$$SST = \sum \varepsilon^2 = \sum (Y - \bar{Y})^2 = \sum \{(Y - \hat{Y}) + (\hat{Y} - \bar{Y})\}^2$$

$$SST = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})(\hat{Y} - \bar{Y})$$

وحيث ان $\sum (Y - \hat{Y})(\hat{Y} - \bar{Y}) = 0$ فيكون

$$SST = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2 = SSE + SSR$$



وبالتطبيق على المثال السابق نجد أن

تكلفة الإعلانات (X)	المبيعات (Y)	$\hat{Y}$	$(Y-\bar{Y})^2$	$(Y-\hat{Y})^2$	$(\hat{Y}-\bar{Y})^2$
6	3	3.5	1.00	0.25	0.25
8	4	4.5	0.00	0.25	0.25
9	6	5	4.00	1.00	1.00
5	4	3	0.00	1.00	1.00
4	2	2.5	4.00	0.25	2.25
10	5	5.5	1.00	0.25	2.25
$\bar{X}=7$	$\bar{Y}=4$		10.00	3.00	7.00
			SST	SSE	SSR

حيث أن $SST = SSE + SSR = 3 + 7 = 10$

SST هو التغير الكلي في Y

SSE هو حجم التغير في Y الذي لا يستطيع النموذج تفسيره

SSR هو حجم التغير في Y الذي يمكن تفسيره بواسطة النموذج

ويتم تعريف معامل التحديد R^2 بأنه مقدار التغير في Y الذي يستطيع النموذج تفسيره. أو هو نسبة التباين الذي يمكن تفسيره بواسطة النموذج من التباين الكلي. ويتم حسابه بالصيغة التالية:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

وبالتالي فإن معامل التحديد تتراوح قيمته بين الصفر والواحد الصحيح.

وبالتطبيق على المثال السابق نجد أن معامل التحديد

$$R^2 = \frac{SSR}{SST} = \frac{7}{10} = 0.7$$

وهذا يعني أن المتغير المستقل (تكلفة الإعلانات) تستطيع تفسير 70% من التباين في المبيعات.

تحديد جودة نموذج التقدير Model Quality Identification

يتم تحديد جودة نموذج التقدير من خلال ثلاثة مؤشرات

أولاً: قيمة معامل التحديد R^2 Coefficient of Determination

والتي تحدد قدرة نموذج التقدير على شرح التغير في الظاهرة Y. فكلما إقتربت قيمة معامل التحديد من الواحد كلما كان النموذج أقدر على شرح الظاهرة بشكلٍ أفضل.

ثانياً: صلاحية نموذج التقدير Model Validity

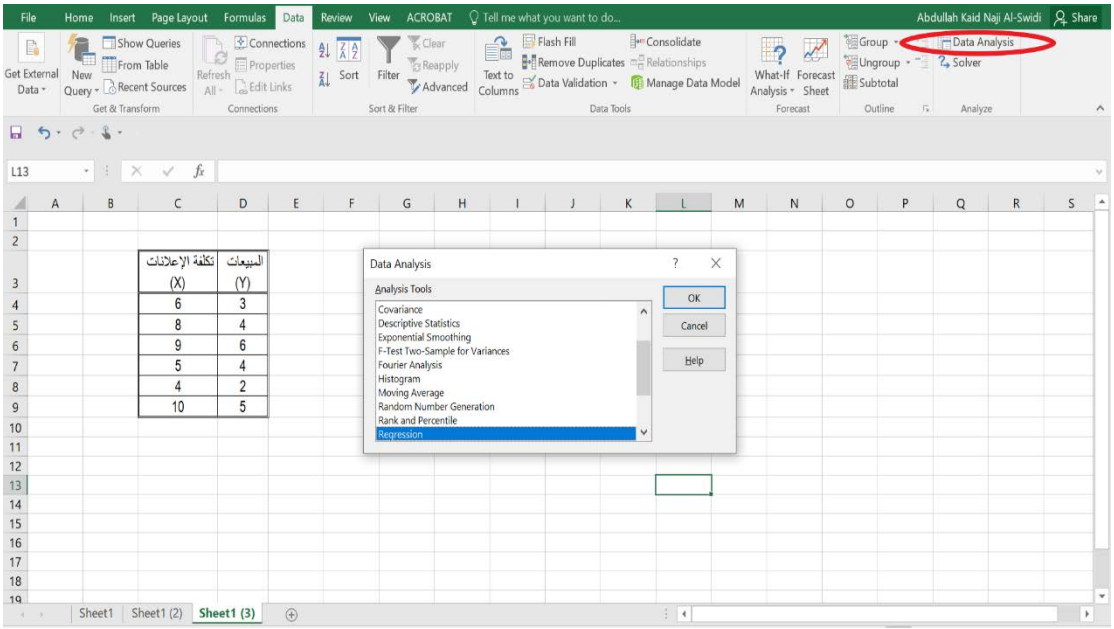
والتي توضح فعالية المتغير(ات) المستقلة في شرح المتغير Y

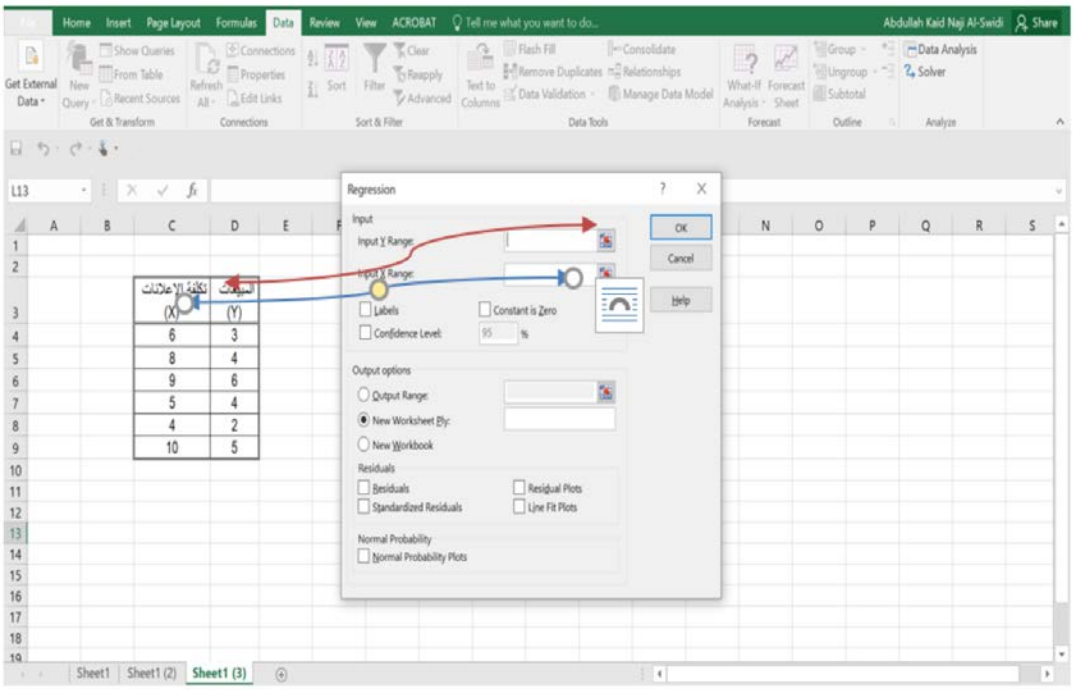
ثالثاً: الدلالة الإحصائية للمتغير(ات) المستقلة Statistical Significance

والتي توضح مدى دلالة كل متغير في نموذج التقدير والقدرة النسبية لكل متغير في شرح التغير في الظاهرة Y.

من أجل توضيح النقاط السابقة سنقوم بتحليل الإنحدار بإستخدام تطبيق EXCEL كما يلي

من خلال قائمة Data في تطبيق EXCEL نحصل على قائمة Data Analysis ونختار منها Regression





يتم تحديد بيانات المتغيرين والضغط على Labels ليقوم التطبيق بقراء أسماء المتغيرات في الصف الأول ثم يتم الضغط على OK لنحصل على النتائج التالية:

SUMMARY OUTPUT

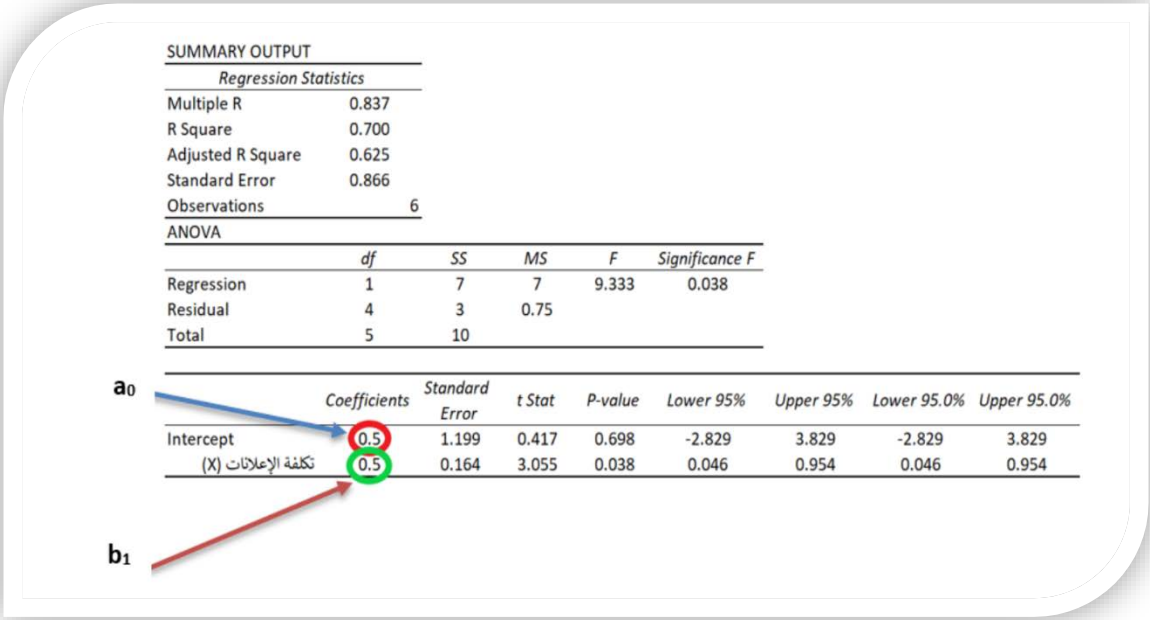
Regression Statistics	
Multiple R	0.837
R Square	0.700
Adjusted R Square	0.625
Standard Error	0.866
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7	7	9.333	0.038
Residual	4	3	0.75		
Total	5	10			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.5	1.199	0.417	0.698	-2.829	3.829	-2.829	3.829
تكلفة الإعلانات (X)	0.5	0.164	3.055	0.038	0.046	0.954	0.046	0.954

نقوم اولاً بكتابة معادلة الانحدار الخطي



SUMMARY OUTPUT

Regression Statistics

Multiple R	0.837
R Square	0.700
Adjusted R Square	0.625
Standard Error	0.866
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7	7	9.333	0.038
Residual	4	3	0.75		
Total	5	10			

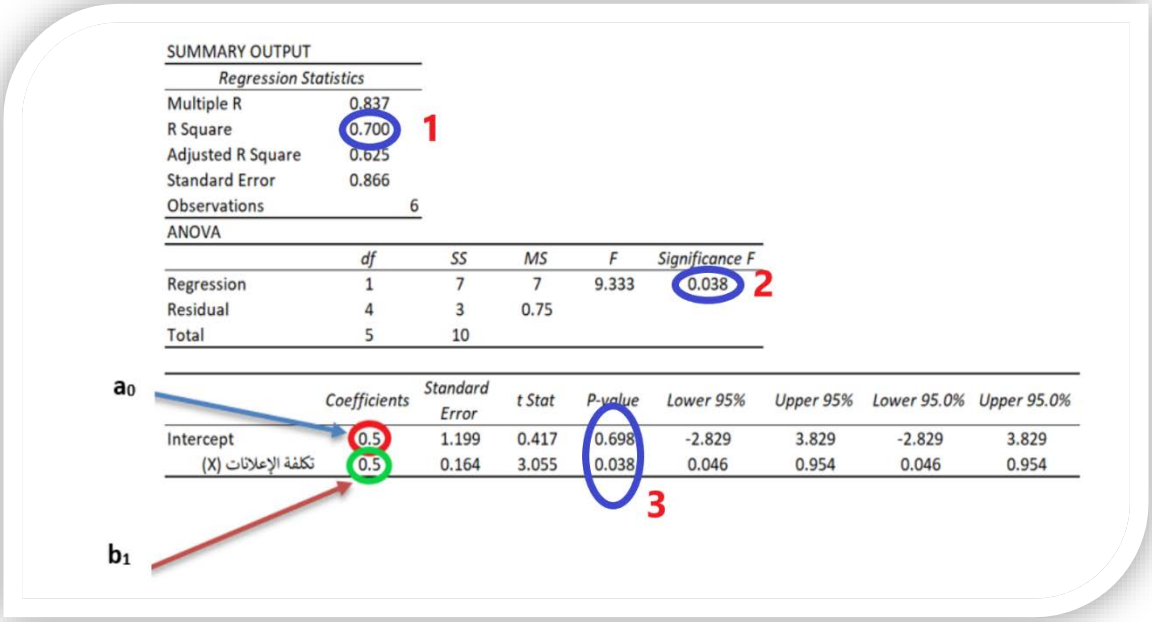
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.5	1.199	0.417	0.698	-2.829	3.829	-2.829	3.829
تكلفة الإعلانات (X)	0.5	0.164	3.055	0.038	0.046	0.954	0.046	0.954

a_0

b_1

$$Y = 0.5 + 0.5X$$

ثم نقوم بدراسة جودة النموذج من خلال معامل التحديد، وصلاحية النموذج، ودلالة المتغيرات.



SUMMARY OUTPUT

Regression Statistics

Multiple R	0.837
R Square	0.700
Adjusted R Square	0.625
Standard Error	0.866
Observations	6

1

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7	7	9.333	0.038
Residual	4	3	0.75		
Total	5	10			

2

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.5	1.199	0.417	0.698	-2.829	3.829	-2.829	3.829
تكلفة الإعلانات (X)	0.5	0.164	3.055	0.038	0.046	0.954	0.046	0.954

3

a_0

b_1

أولاً: قيمة معامل التحديد R^2 Coefficient of Determination

من الشكل أعلاه يتضح أن معامل التحديد هو $\boldsymbol{R^2 = 70\%}$ وهو ما يعني ان متغير (تكلفة الإعلانات) يستطيع أن يشرح 70% من التغير في المبيعات وهي نسبة جيدة تعكس جودة النموذج.

ثانياً: صلاحية نموذج التقدير Model Validity

تقاس صلاحية النموذج من خلال رفض فرضية ان قيمة الميل في معادلة الإنحدار صفرية كما يلي

$$H_0: \beta_1 = 0 \; vs \; H_1: \beta_1 \neq 0$$

إذا إخترنا مستوى دلالة $\alpha = 0.05$ ، فبإستخدام النتائج من تطبيق EXCEL نجد ان الإحتمالية p=0.038 في جدول ANOVA وبالتالي نقوم بمقارنة قيمة p مع α فإذا كانت

A. $p > \alpha$ لا نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج غير صالح للتقدير

B. $p < \alpha$ نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج صالح للتقدير

ومن الملاحظ في المثال الذي نناقشه أن $p = 0.038 < \alpha = 0.05$ وعليه فإن النموذج الذي يحدد العلاقة بين تكلفة الإعلانات والمبيعات له قوة تقديرية وهو صالح للتقدير.

ثالثاً: الدلالة الإحصائية للمتغير(ات) المستقلة Statistical Significance

لتحليل الدلالة الإحصائية للمتغير المستقل نختبر الفرض التالي (سيتطابق مع الفرض المتعلق بصلاحية النموذج في حالة الانحدار الخطي البسيط)

$$H_0: \beta_1 = 0 \; vs \; H_1: \beta_1 \neq 0$$

إذا إخترنا مستوى دلالة $\alpha = 0.05$ ، فبإستخدام النتائج من تطبيق EXCEL نجد ان الإحتمالية p=0.038 في الجدول الثالث عند الرقم 3 وبالتالي نقوم بمقارنة قيمة p مع α فإذا كانت

A. $p > \alpha$ لا نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج غير صالح للتقدير

B. $p < \alpha$ نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج صالح للتقدير

ومن الملاحظ في المثال الذي نناقشه أن $p = 0.038 < \alpha = 0.05$ وعليه فإن المتغير المستقل (تكلفة الإعلانات) هو متغير ذو دلالة إحصائية عند مستوى دلالة $\alpha = 0.05$.

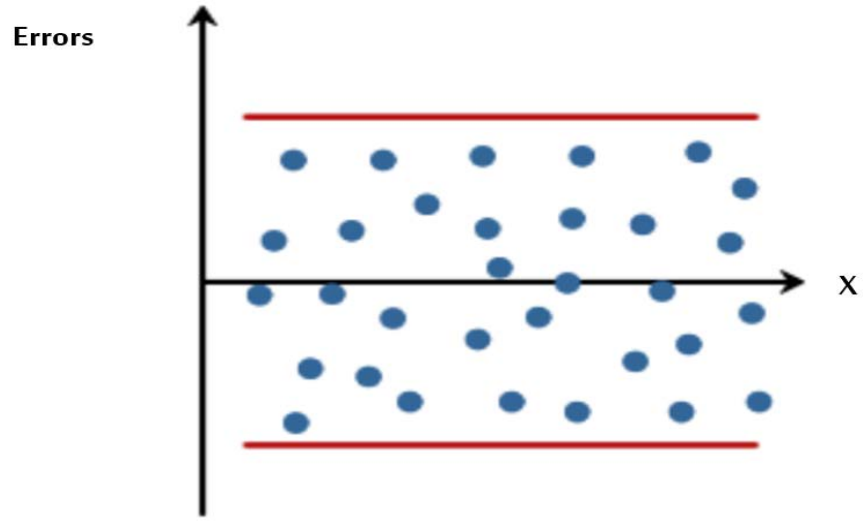
فرضيات الانحدار الخطي Assumptions of the Regression Model

هناك بعض الفرضيات لإجراء تحليل الانحدار، حيث يمكن إجراء اختبارات إحصائية لتحديد ما إذا كان النموذج مفيدا أم لا.

هذه الفرضيات هي:

1. الأخطاء مستقلة عن بعضها.
2. الأخطاء موزعة بحسب التوزيع الطبيعي.
3. متوسط الأخطاء صفراً.
4. الأخطاء لها تباين ثابت بغض النظر عن قيمة X.

من الممكن التحقق من البيانات لمعرفة ما إذا كانت هذه الافتراضات مستوفاة من عدمه. ويمكن التحقق من كل تلك الفرضيات من خلال رسم الأخطاء مع قيم المتغير المستقل فإذا ظهر توزيع الأخطاء عشوائياً بغض النظر عن قيم X وكان هناك ثبات في التباين، فنقول ان الفرضيات قد تحققت.



الإنحدار الخطي المتعدد Multiple Linear Regression

الانحدار المتعدد، المعروف أيضا باسم الانحدار الخطي المتعدد (MLR) ، هو تقنية إحصائية تستخدم متغيرين مستقلين أو أكثر للتنبؤ بنتيجة المتغير التابع او متغير الاستجابة. ويُعتبر تمديد للانحدار الخطي البسيط. يمكن بواسطة الانحدار الخطي البسيط تفسير العلاقة بين متغيرات مستقلة متعددة ومتغير تابع واحد. تعمل هذه المتغيرات المستقلة كمتغيرات تنبؤية أو تفسيرية. كما تُساعد هذه التقنية على تفسير الظواهر المختلفة من خلال بناء نماذج تحوي العديد من المتغيرات ذات الصلة. وتُستخدم هذه التقنية في مجموعة كبيرة من المجالات والدراسات والتخصصات، بما في ذلك العلوم الطبيعية والمعملية والبحوث الطبية والصيدلانية والاقتصاد القياسي والاستدلال المالي وغيرها.

ويُعطى نموذج الانحدار المتعدد بالصيغة التالية

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

حيث

Y المتغير التابع او متغير الإستجابة

X_i المتغير المستقل رقم i

β_0 ثابت النموذج، قيمة Y عندما تكون X=0

β_i معامل المتغير X_i

n عدد المتغيرات المستقلة في النموذج

ε الخطأ العشوائي

ويمكن تقدير معالم النموذج السابق β_0 و β_i كما في النموذج التالي عن طريق تقنية مربعات الأخطاء الصغرى OLS :

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

حيث

$\hat{Y}$ القيمة المقدرة ل Y

b_0 ثابت النموذج بناءً على بيانات العينة

b_i معامل المتغير X_i

مثال 3

أرادت وزارة التربية والتعليم في إحدى الدول دراسة بعض العوامل التي تؤثر على متوسط أداء الطلاب في الشهادة الثانوية. وقد حدد الخبراء أن متوسط رواتب المدرسين وسنوات الخبرة في التدريس في نفس المدرسة تُعتبر من العوامل الهامة في تحديد أداء الطلاب. وقد قام مختصون بجمع البيانات المتعلقة بمتوسط أداء الطلاب، ومتوسط رواتب المدرسين ومتوسط سنوات الخدمة للمدرسين في كل مدرسة من عدد 12 مدرسة تم إخضاعها للدراسة. وقد تم وضع البيانات في الجدول التالي

المدرسة	متوسط معدلات الطلاب	متوسط رواتب المدرسين	متوسط سنوات خدمة المدرس في المدرسة
1	81	$34,300	9.3
2	78	28700	10.1
3	76	26500	7.6
4	77	36200	8.2
5	84	35900	8.8
6	86	32500	12.7
7	79	31800	8.4
8	91	38200	11.5
9	68	27100	8.3
10	73	31500	7.3
11	90	37600	12.3
12	85	40400	14.2

والمطلوب

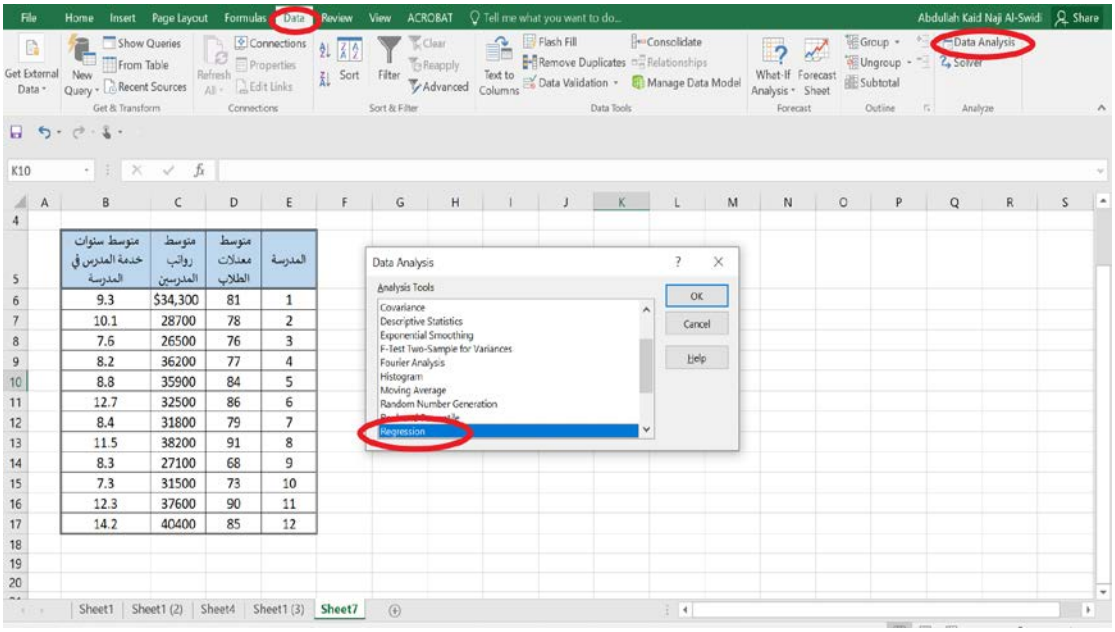
A. بإستخدام تطبيق EXCEL ، قم إيجاد معادلة الانحدار الخطي

B. هل يُعتبر هذين العاملين جيدين في تحديد متوسط درجات الطلاب؟ وضح

C. إذا قامت الوزارة بزيادة متوسط رواتب المدرسين في إحدى المدارس ب 1000 دولار، فما هو متوسط الزيادة في متوسط درجات الطلاب في تلك المدرسة.

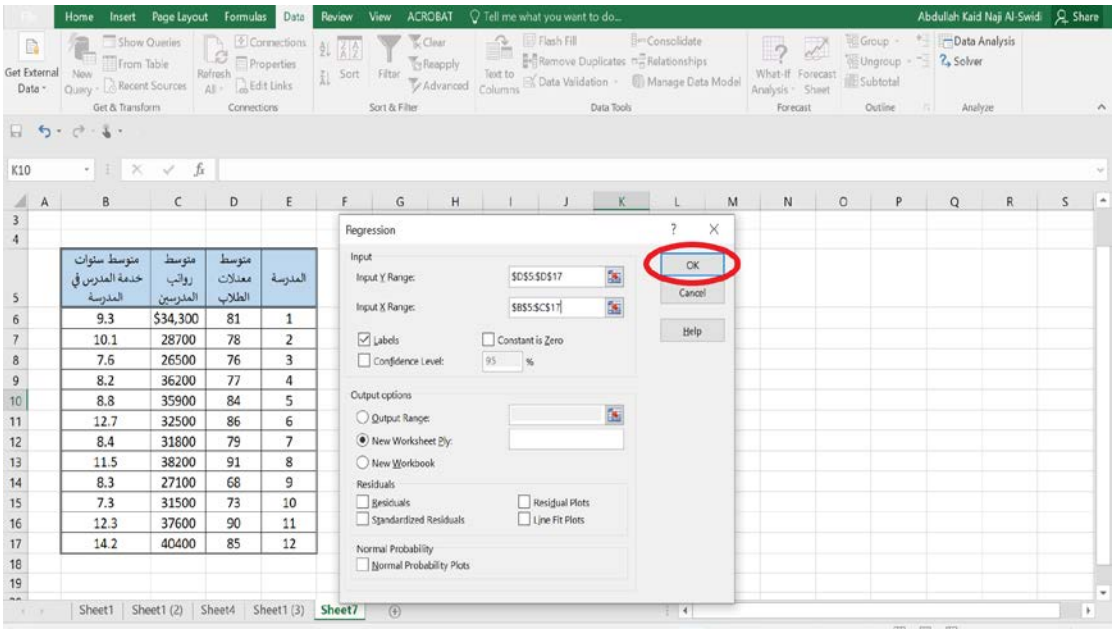
الحل

A. بعد تجهيز البيانات في تطبيق EXCEL نقوم بإختيار قائمة Data ثم نختار Data Analysis ونختار منها Regression كما في الشكل التالي

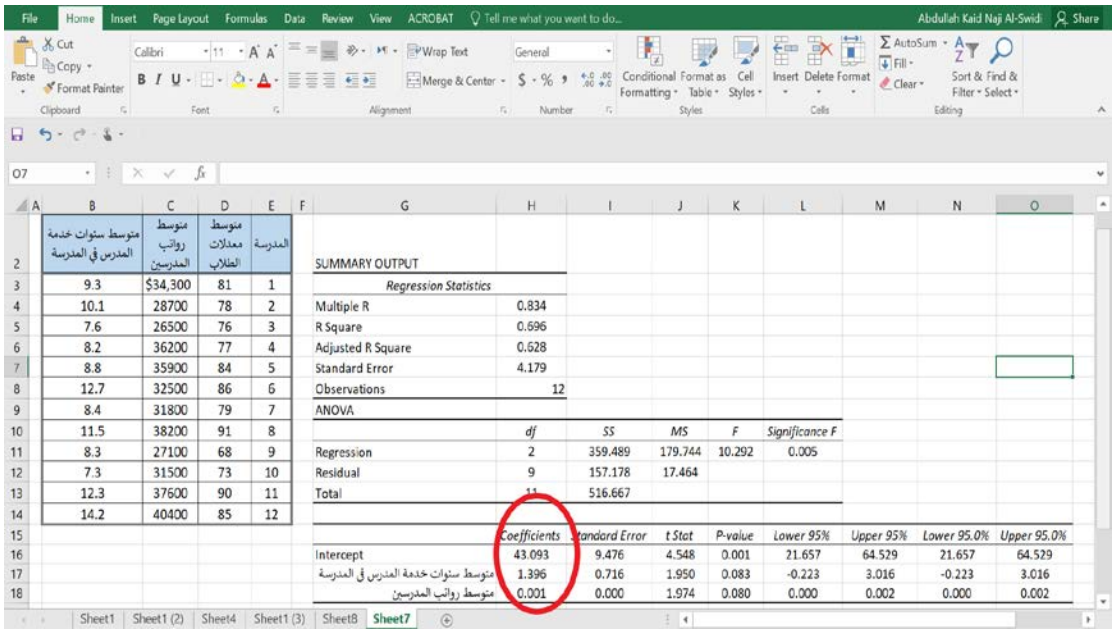


المدرسة	متوسط معدلات الطلاب	متوسط رواتب المدرسين	متوسط سنوات خدمة المدرس في المدرسة
1	81	$34,300	9.3
2	78	28700	10.1
3	76	26500	7.6
4	77	36200	8.2
5	84	35900	8.8
6	86	32500	12.7
7	79	31800	8.4
8	91	38200	11.5
9	68	27100	8.3
10	73	31500	7.3
11	90	37600	12.3
12	85	40400	14.2

حيث ان متوسط معدلات الطلاب هو المتغير التابع الذي يتأثر ب متوسط رواتب المدرسين وسنوات خدمتهم في المدرسة (المتغيرات المستقلة)، نقوم بتحديد موقع بيانات المتغير التابع والمتغيرات المستقلة ونضغط على OK كما في الشكل التالي



نقوم بكتابة معادلة الانحدار الخطي بناءً على نتائج التحليل كما يلي



متوسط سنوات خدمة المدرس في المدرسة	متوسط رواتب المدرسين	متوسط معدلات الطلاب	المدرسة
9.3	$34,300	81	1
10.1	28700	78	2
7.6	26500	76	3
8.2	36200	77	4
8.8	35900	84	5
12.7	32500	86	6
8.4	31800	79	7
11.5	38200	91	8
8.3	27100	68	9
7.3	31500	73	10
12.3	37600	90	11
14.2	40400	85	12

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.834
R Square	0.696
Adjusted R Square	0.628
Standard Error	4.179
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	359.489	179.744	10.292	0.005
Residual	9	157.178	17.464		
Total	11	516.667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	43.093	9.476	4.548	0.001	21.657	64.529	21.657	64.529
متوسط سنوات خدمة المدرس في المدرسة	1.396	0.716	1.950	0.083	-0.223	3.016	-0.223	3.016
متوسط رواتب المدرسين	0.001	0.000	1.974	0.080	0.000	0.002	0.000	0.002

$$\widehat{Y} = \mathbf{43.093} + \mathbf{1.396} * \text{سنوات الخدمة} + \mathbf{0.001} * \text{متوسط الرواتب}$$

B. أولاً: نختبر صلاحية النموذج

تقاس صلاحية النموذج من خلال رفض فرضية ان قيمة الميول في معادلة الإنحدار صفرية كما يلي

$$H_0: \beta_1 = \beta_2 = 0 \; vs \; H_1: \beta_i \neq 0, i = 1 \; or \; 2$$

إذا إخترنا مستوى دلالة $\alpha = 0.05$ ، فبإستخدام النتائج من تطبيق EXCEL نجد ان الإحتمالية p=0.005 في جدول ANOVA وبالتالي نقوم بمقارنة قيمة p مع α فإذا كانت

A. $p > \alpha$ لا نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج غير صالح للتقدير

B. $p < \alpha$ نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج صالح للتقدير

ومن الملاحظ في المثال الذي نناقشه أن $p = 0.005 < \alpha = 0.05$ وعليه فإن النموذج الذي يحدد العلاقة بين رواتب المدرسين وسنوات خدمتهم له قوة تقديرية ويُعتبر صالحاً للتقدير.

ثانياً: نختبر الدلالة الإحصائية للمتغيرات

لتحليل الدلالة الإحصائية للمتغير المستقل المتعلق بمتوسط سنوات خدمة المدرسين نختبر الفرض التالي

$$H_0: \beta_1 = 0 \; vs \; H_1: \beta_1 \neq 0$$

إذا إخترنا مستوى دلالة $\alpha = 0.1$ ، فبإستخدام النتائج من تطبيق EXCEL نجد ان الإحتمالية p=0.083 في الشكل أدناه وبالتالي نقوم بمقارنة قيمة p مع α فإذا كانت

A. $p > \alpha$ لا نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج غير صالح للتقدير

B. $p < \alpha$ نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج صالح للتقدير

ومن الملاحظ في المثال الذي نناقشه أن $p = 0.083 < \alpha = 0.1$ وعليه فإن المتغير المستقل (متوسط سنوات الخدمة) هو متغير ذو دلالة إحصائية عند مستوى دلالة $\alpha = 0.1$.

وبالمثل، لتحليل الدلالة الإحصائية للمتغير المستقل المتعلق بمتوسط رواتب المدرسين نختبر الفرض التالي

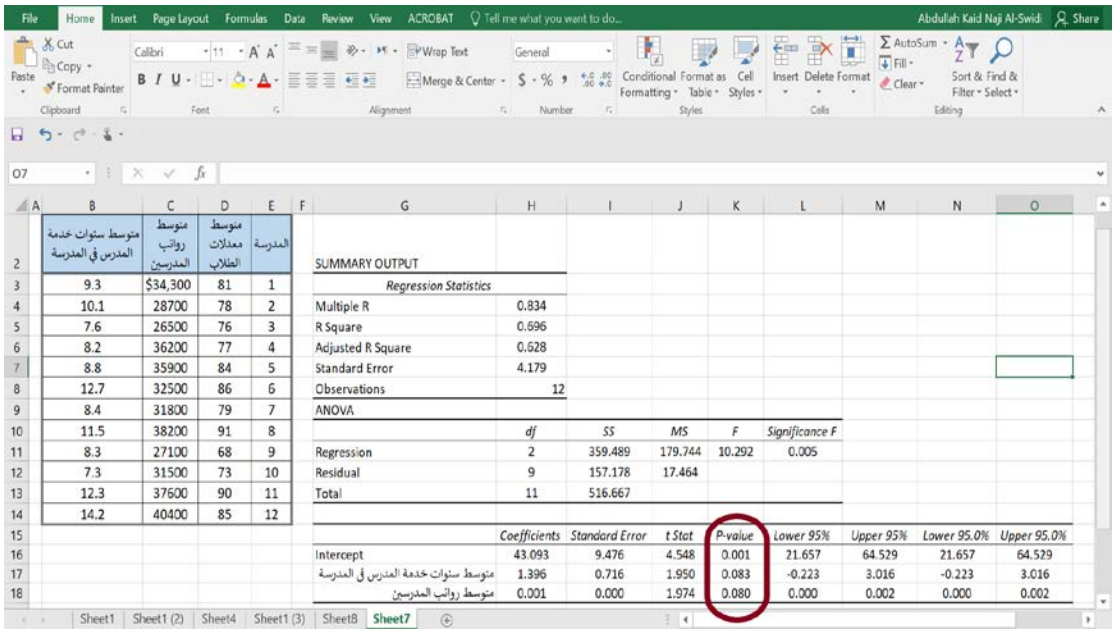
$$H_0: \beta_2 = 0 \; vs \; H_1: \beta_2 \neq 0$$

إذا إخترنا مستوى دلالة $\alpha = 0.1$ ، فبإستخدام النتائج من تطبيق EXCEL نجد ان الإحتمالية p=0.080 في الشكل أدناه وبالتالي نقوم بمقارنة قيمة p مع α فإذا كانت

A. $p > \alpha$ لا نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج غير صالح للتقدير

B. $p < \alpha$ نرفض الفرضية الصفرية H_0 ويعني ذلك ان النموذج صالح للتقدير

ومن الملاحظ في المثال الذي نناقشه أن $p = 0.080 < \alpha = 0.1$ وعليه فإن المتغير المستقل (متوسط رواتب المدرسين) هو متغير ذو دلالة إحصائية عند مستوى دلالة $\alpha = 0.1$.



متوسط سنوات خدمة المدرس في المدرسة	متوسط رواتب المدرسين	متوسط معدلات الطلاب	المدرسة
9.3	$34,300	81	1
10.1	28700	78	2
7.6	26500	76	3
8.2	36200	77	4
8.8	35900	84	5
12.7	32500	86	6
8.4	31800	79	7
11.5	38200	91	8
8.3	27100	68	9
7.3	31500	73	10
12.3	37600	90	11
14.2	40400	85	12

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.834
R Square	0.696
Adjusted R Square	0.628
Standard Error	4.179
Observations	12

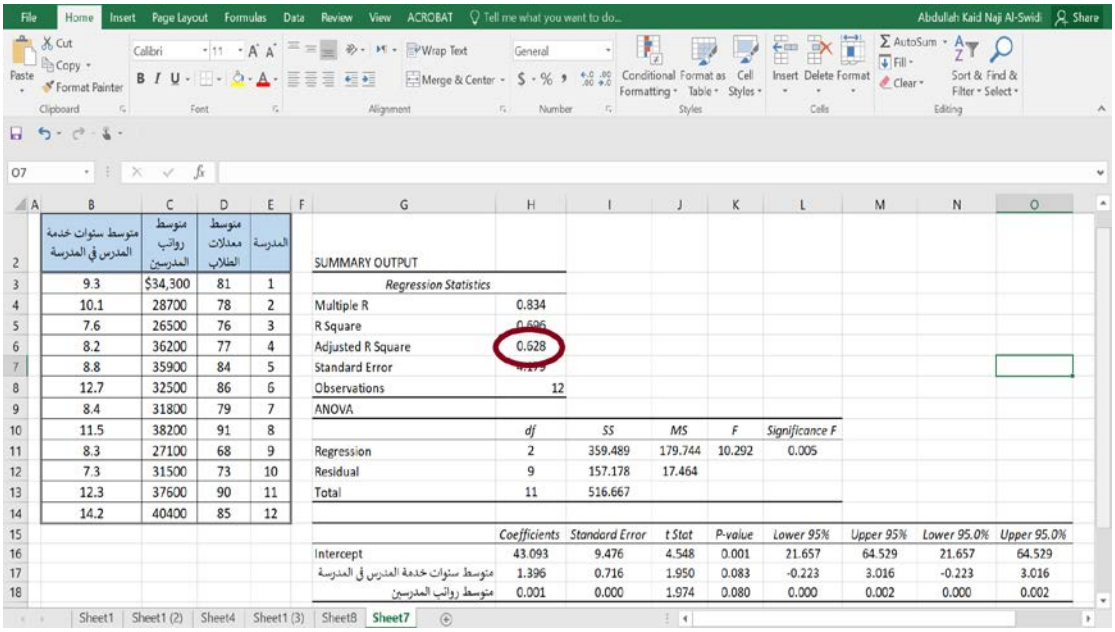
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	359.489	179.744	10.292	0.005
Residual	9	157.178	17.464		
Total	11	516.667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	43.093	9.476	4.548	0.001	21.657	64.529	21.657	64.529
متوسط سنوات خدمة المدرس في المدرسة	1.396	0.716	1.950	0.083	-0.223	3.016	-0.223	3.016
متوسط رواتب المدرسين	0.001	0.000	1.974	0.080	0.000	0.002	0.000	0.002

ثالثاً: قيمة معامل التحديد R^2

حيث ان النموذج يشرح 62.8% من التباين في المتغير التابع ، فيُعتبر النموذج مناسب الى حدٍ ما في شرح الظاهرة.



المدرسة	متوسط معدلات الطلاب	متوسط رواتب المدرسين	متوسط سنوات خدمة المدرس في المدرسة
1	81	$34,300	9.3
2	78	28700	10.1
3	76	26500	7.6
4	77	36200	8.2
5	84	35900	8.8
6	86	32500	12.7
7	79	31800	8.4
8	91	38200	11.5
9	68	27100	8.3
10	73	31500	7.3
11	90	37600	12.3
12	85	40400	14.2

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.834
R Square	0.696
Adjusted R Square	0.628
Standard Error	4.279
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	359.489	179.744	10.292	0.005
Residual	9	157.178	17.464		
Total	11	516.667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	43.093	9.476	4.548	0.001	21.657	64.529	21.657	64.529
متوسط سنوات خدمة المدرس في المدرسة	1.396	0.716	1.950	0.083	-0.223	3.016	-0.223	3.016
متوسط رواتب المدرسين	0.001	0.000	1.974	0.080	0.000	0.002	0.000	0.002

C. لو قامت الوزارة برفع متوسط الرواتب في مدرسةٍ ما بمعدل 1000 دولار، فإنه بحسب معادلة الانحدار الخطي، ستكون الزيادة في متوسط الدرجات بمقدار $1000 * 0.001 =$ 1، أي بمقدار درجة واحدة في المتوسط.

الخاتمة

قمنا في هذا الفصل مناقشة مفهوم الانحدار الخطي البسيط والمتعدد مع تقديم مثلة عنهما. كما تمت مناقشة كيفية تحديد معامل التحديد وكذا خطوات تحديد جودة نموذج التقدير للمتغير التابع بمعلومية المتغير او المتغيرات المستقلة. كما قمنا بشرح كيفية الإستفادة من تطبيق EXCEL في تنفيذ تحليل الانحدار الخطي وكيفية قراءة مخرجاته والاستفادة منها في تطوير معادلة الانحدار الخطي وتحديد كيفية نموذج التقدير.

أسئلة وتمارين

أولاً : الأسئلة النظرية

1. يختلف نموذج الانحدار الخطي المتعدد عن نموذج الانحدار الخطي البسيط في أن نموذج الانحدار المتعدد يحتوي على أكثر من
   A. متغير مستقل.
   B. متغير تابع.
   C. الجزء المقطوع من محور Y (a).
   D. خطأ.

---

2. يتم اختبار صلاحية النموذج باستخدام قيمة *F*. فيكون النموذج صالحاً إذا
   A. كانت قيمة *F* منخفضة.
   B. كانت *F* significance منخفضة.
   C. قيمة معامل التحديد R^2 منخفضة.
   D. الميل b أقل من الجزء المقطوع من محور Y (a).

---

3. تُسمى النسبة المئوية للتباين في المتغير التابع الذي يتم شرحه بواسطة المتغيرات المستقلة ب
   A. معامل الارتباط r.
   B. متوسط مربعات الأخطاء MSE .
   C. معامل التحديد R^2.
   D. الميل b

---

4. يتم تعريف أفضل خط انحدار على أنه
   A. الخط الذي يحتوي على الحد الأدنى من مجموع مربعات الأخطاء.
   B. الحد الأقصى لمجموع مربعات الأخطاء.
   C. متوسط مربعات الأخطاء.
   D. لا شيء مما سبق.

---

5. تقاس نسبة التباين في المتغير التابع الذي يتم تفسيره بمعادلة الانحدار بواسطة
   A. معامل الارتباط.
   B. MSE

C. معامل التحديد.

D. الميل.

---

6. يقيس معامل الارتباط

A. درجة أو قوة العلاقة الخطية.

B. التباين في Y الذي تفسره معادلة الانحدار.

C. كل من A و بB.

D. لا شيء مما سبق.

---

7. عند استخدام المتغيرات الوهمية في معادلة الانحدار لنمذجة متغير نوعي أو فئوي ، يجب أن يساوي عدد المتغيرات الوهمية

A. عدد الفئات.

B. أكثر من عدد الفئات بواحد.

C. أقل من عدد الفئات بواحد.

D. عدد المتغيرات المستقلة الأخرى في النموذج.

---

8. أحد الافتراضات في تحليل الانحدار هو أن

A. الأخطاء لها متوسط 1.

B. الأخطاء لها متوسط 0.

C. البيانات (Y) لها متوسط 1.

D. البيانات (Y) لها متوسط 0.

---

9. في مخطط الرسوم النقطية ، يتم رسم المتغير المستقل على

A. المحور الرأسي

B. خط الاتجاه.

C. المحور الأفقي.

D. نقطة أصل المحور X

ثانياً المسائل العملية

السؤال 1

في دراسة العلاقة بين كمية الأمطار التي نزلت في 6 مناطق وبين إرتفاع شجرة الزيتون بعمر معين تم تجميع البيانات كما يلي:

كمية الأمطار	إرتفاع شجرة الزيتون
5	3
10	3
15	4
20	6
25	6
30	7

والمطلوب

A. أوجد معادلة الإنحدار الخطي
B. قم بتقييم صلاحية النموذج وجودته
C. قم بكتابة معادلة الإنحدار الخطي
D. كم تقدر كمية الأمطار التي ستهطل في منطقةٍ ما إذا كان إرتفاع شجرة الزيتون 5.5

السؤال 2

في دراسة العلاقة بين عدد الساعات الذي قضاها الطلبة في دراسة مادة الطرق الكمية في صناعة القرار وبين درجات الإختبار الأول كانت النتاج لعدد عشرة طلاب كما يلي:

درجة الإختبار	عدد ساعات الدراسة
84	17
21	1
91	22
60	12
44	7
37	4
73	14
65	10
58	9
31	4

والمطلوب

A. أوجد معادلة الإنحدار الخطي
B. قم بتقييم صلاحية النموذج وجودته
C. ما هو معدل الإرتفاع في درجة الطالب لكل ساعة دراسة إضافية
D. كم تقدر عدد الساعات التي درسها الطالب إذا علمت أنه قضى في الدراسة للإختبار 6 ساعات

السؤال 3

تقوم شركة الرواد للعقارات بتطوير نموذج يمكن من خلاله تحديد سعر المنزل من خلال تحديد بعض المتغيرات مثل مساحته بالقدم وعدد غرف النوم عمره بالسنوات وحالته. وبناءً على دراسة عينة من المنازل فقد تم وضع البيانات في الجدول التالي

الحالة	العمر بالسنوات	عدد الغرف	المساحة بالقدم	سعر البيع
جيدة	30	2	1,926	95,000
ممتازة	40	3	2,069	119,000
ممتازة	30	4	1,720	124,800
جيدة	15	3	1,396	135,000
جيدة	32	3	1,706	142,800
ممتازة	38	3	1,847	145,000
جيدة	27	2	1,950	159,000
ممتازة	30	2	2,323	165,000
جيدة	26	2	2,285	182,000
جيدة	35	4	3,752	183,000
جيدة	18	2	2,300	200,000
جيدة	17	2	2,525	211,000
ممتازة	40	2	3,800	215,000
ممتازة	12	3	1,740	219,000

والمطلوب

A. أوجد معادلة الإنحدار الخطي
B. قم بتقييم صلاحية النموذج وجودته
C. ما هو معدل الإرتفاع في سعر المنزل لكل سنة من عمره
D. كم تقدر سعر المنزل إذا علمت أن عمره 25 سنة ومساحته 1800 قدم
E. ما هو العامل الأهم في تحديد سعر المنزل المساحة أو عدد غرف النوم أو العمر؟

السؤال الرابع

توضح البيانات في الجدول التالي الرواتب الإبتدائية للطلاب الذي تخرجوا من جامعة قطر وتم توظيفهم حديثاً في سوق العمل بعد وقت قصير من التخرج. الجدول التالي يوضح الراتب الإبتدائي ومتوسط معدل الخريج (GPA) والتخصص (إدارة الاعمال أو غيرها).

الراتب	المعدل	التخصص
29,500	3.1	غير متعلق بإدارة أعمال
46,000	3.5	إدارة اعمال
39,800	3.8	إدارة اعمال
36,500	2.9	غير متعلق بإدارة أعمال
42,000	3.4	إدارة اعمال
31,500	2.1	غير متعلق بإدارة أعمال
36,200	2.5	إدارة اعمال

والمطلوب

A. باستخدام برنامج الإكسل أو POM for Windows، قم بتطوير النموذج الذي يمكن بواسطته التنبؤ بالمرتبات الإبتدائية للخريجين بناء على المعدل والتخصص.

B. استخدام هذا النموذج للتنبؤ بالمرتب الأساسي لخريج تخصصه إدارة اعمال ومعدله التراكمي 3.0.

C. ماذا يمكن القول بخصوص المقارنة بين خريجي إدارة الاعمال وغيرهم من حيث الراتب؟

D. هل تعتقد أن هذا النموذج صالح للتنبؤ ؟ قم بتبرير إجابتك باستخدام المعلومات المتوفرة في مخرجات برنامج التحليل .

E. ما هو العامل الأكثر أهمية في شرح الراتب المقدر للخريج؟