الارتباط والانحدار

- تعرضنا في الأجزاء السابقة إلى المقاييس الإحصائية التي تتعامل مع متغير واحد.
- مقاييس النزعة المركزية تصف لنا القيمة التي تقع في مركز مجموعة البيانات.
- مقاييس التشتت تقيس لنا درجة انتشار وتبعثر وتوزيع قيم البيانات.
- عندما يتعلق الأمر بمتغيرين أو أكثر ، فان الباحث يتطلع إلى قياس وتحديد درجة واتجاه العلاقة بين المتغيرات أو الظواهر قيد البحث .

- نقوم بعد ذلك باستخدام العلاقات الموجودة في التنبؤ بقيمة أحد المتغيرات بدلالة المتغير (المتغيرات) الأخرى . ومحاولة التعبير عن هذه العلاقات بدالة خطية ، لأن وجود علاقة أو ارتباط بين المتغيرات لا يعني بالضرورة إمكانية التعبير عن هذه العلاقة بشكل خطي .
- لو تناولنا الأبحاث الإحصائية لوجدنا أن هدفها الأساسي هو دراسة العلاقة بين المتغيرات . ويعبر عن هذه العلاقة بالأرقام أو القيم الكمية كما قد يعبر عنها بالوصف . فالإحصائي عندما يتحدث عن إحدى العلاقات الدالية بين متغيرين فانه يقصد بذلك أن المتغيرين يمكن ربطهما

- لنضرب مثلاً بالمكالمات الهاتفية ، فإذا عرفنا أن قيمة المكالمة الواحدة هي 20 فلساً ، وأن استخدام الهاتف كان لخمسة مرات ، فانه لمعرفة القيمة التي يجب دفعها هي 100 فلس . ويمكن التعبير عن ذلك بطريقة رياضية ، حيث نرمز لقيمة المكالمات بالرمز Y ولعدد مرات استعمال الهاتف بالرمز X . وبذلك تكون العلاقة الرياضية :

$$Y = 20\,X$$

- وتبقى هذه الصيغة صالحة مادامت قيمة المكالمة الواحدة ثابتة أي مساوية لعشرين فلساً . واذا ما زادت قيمة المكالمة إلى 25 فلساً فان المعادلة تصبح :

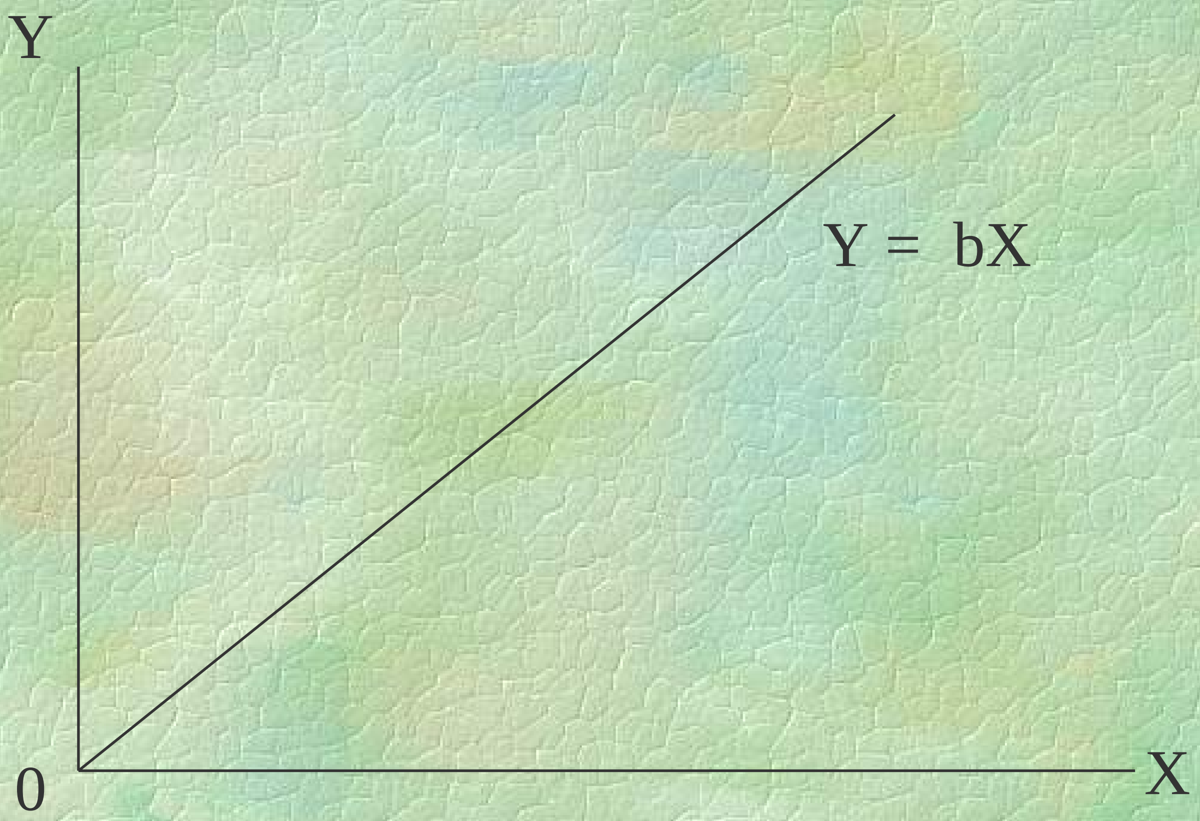
$$Y = 25\ X$$

- وعليه يمكن تعميم القاعدة على كل القيم التي تبلغها المكالمة الهاتفية (b) فتصبح الصياغة الرياضية في هذه الحالة كما في المعادلة :

$$Y = b\ X$$

- لزيادة الإيضاح يمكن التعبير عن هذه المعادلة بالرسم البياني ، حيث يخصص محور الإحداثيات الأفقية لقيم X ومحور الإحداثيات الرأسية لما يقابلها من قيم Y وبذلك نحصل على الشكل التالي (1) ، فتبقى b ثابتة في المعادلة يمثلها الخط المعروف بخط الانحدار ، وفيه نلاحظ أن لكل قيمة من قيم X تقابلها قيمة من قيم Y وان قيمة Y تساوي صفراً عندما تكون X مساوية للصفر . وينطلق هذا الخط الانحداري من أصل الإحداثيات الرأسية والأفقية أي من نقطة تقاطعهما وهي نقطة الصفر .

الشكل (1)

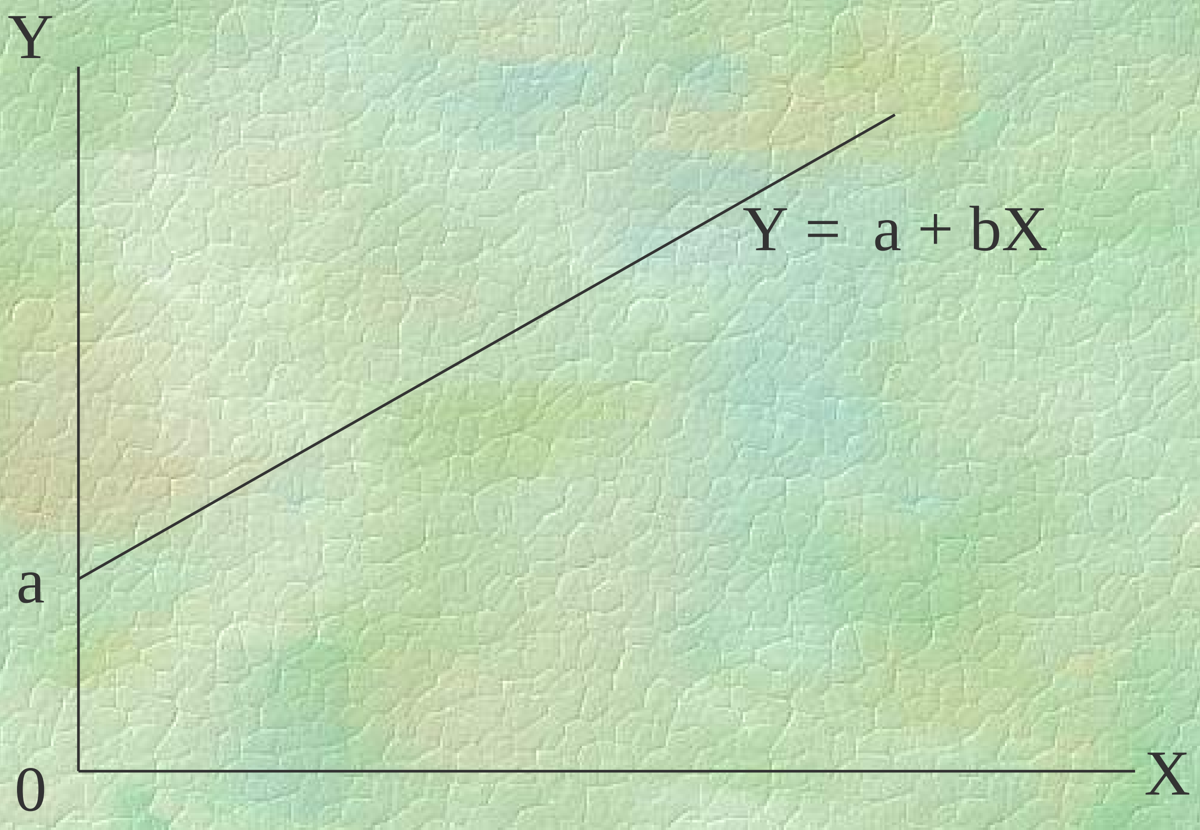


- ولكن صاحب الهاتف لا يدفع لشركة المواصلات قيمة المكالمات الهاتفية فقط ، بل يقوم بدفع قيمة الاشتراك وهي قيمة تضاف إلى قيمة المكالمات الهاتفية . وهذا الاشتراك يدفعه المشترك سواء استعمل هاتفه أم لم يستعمله . وعليه فإننا يجب أن نضيف إلى معادلتنا السابقة الرمز (a) للدلالة على قيمة الاشتراك الثابتة، فتصبح المعادلة التالية كما يلي :

$$Y = a + bX$$

- ويعبر عن هذه المعادلة بالرسم البياني التالي { الشكل (2)} فاصبح الخط الانحداري لا يمر بأصل الإحداثيات كما في الشكل السابق بل يقطع مستقيم الإحداثيات الرأسمالية في نقطة تساوي قيمة الاشتراك . أي أن نقطة (a) تساوي القيمة التي يدفعها المشترك عندما لا يستخدم هاتفه .

الشكل (2)



- ولكن العلاقة الدالية لا يشترط فيها دائماً أن تتبع خطاً مستقيماً بل قد تتبع خطاً منحنياً . أي أنـها لا تكون دائماً بسيطة بل قد تكون مركبة ، ولحساب قيمة العلاقة بين المتغيرين X و Y فان يجب معرفة قيمة المعاملين a و b ، وبمعرفة قيمة المتغير المستقل يستخرج قيمة المتغير التابع Y أو العكس .

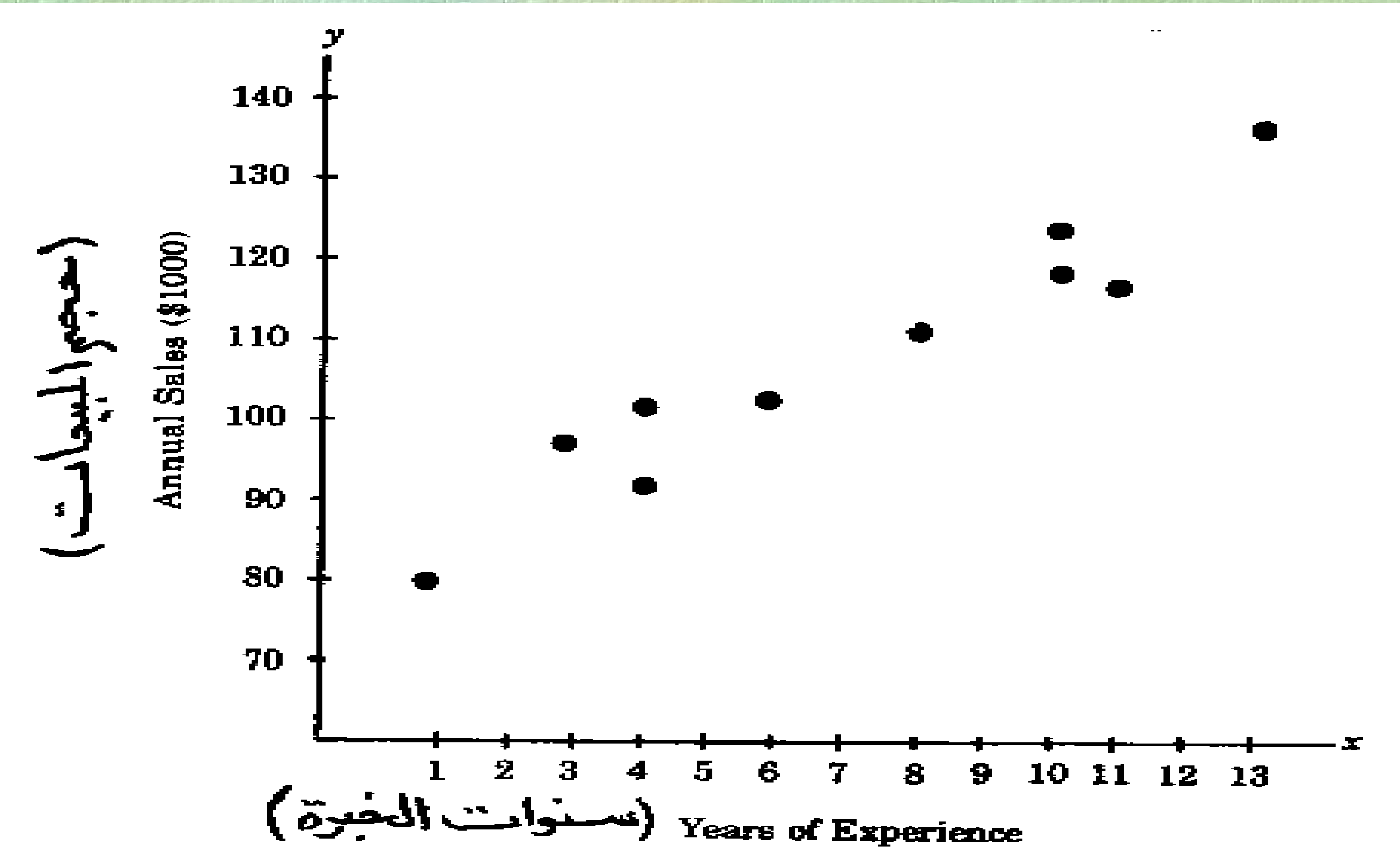
- ولكن ما يحصل في الواقع قد لا يكون بالشكل الذي افترضناه في مثالنا بخصوص الهاتف . فالعلاقة بين الظاهرتين قد لا تكون خطية بشكل كامل . ولبيان ذلك يفترض أن مشكلتنا هي في بيان أو إيجاد العلاقة بين حجم المبيعات وسنوات خبرة الباعة (مندوبي المبيعات) . فلو قام المدير بتسجيل بيانات تتضمن حجم المبيعات السنوية وسنوات الخبرة لمندوبي المبيعات . وكانت هذه البيانات كما في الجدول (1) :

جدول (1)

مندوبي المبيعات	1	2	3	4	5	6	7	8	9	10
سنوات الخبرة	1	3	4	4	6	8	10	10	11	13
حجم المبيعات (ألف دينار)	80	97	92	102	103	111	119	123	117	136

- إن الخطوة الأولى للبحث عن علاقة هي رسم بياني للبيانات أعلاه على الشكل أدناه (3) حيث مثلت سنوات الخبرة على الإحداثي الأفقي والمبيعات السنوية على الإحداثي الرأسي. ونكون بذلك قد حصلنا على رسم انتشاري . وقد أعطي هذا الاسم نظراً لانتشار وتبعثر النقاط على الشكل أو الرسم . وهذا الرسم الانتشاري المبين أدناه يساعد على تكوين استنتاج مبدئي حول إمكانية وجود علاقة بين المتغيرات.

الشكل (3)


y
140
130
120
110
100
90
80
70
(حجم المبيعات) Annual Sales ($1000)
x
1
2
3
4
5
6
7
8
9
10
11
12
13
(سنوات الخبرة) Years of Experience

- عموماً لجأ الإحصائيون إلى تصنيف المتغيرات إلى متغيرات مستقلة ومتغيرات تابعة .
- يستخدم هذا التصنيف من أجل تحديد المتغير الذي يفسر (المستقل) والمتغير المفسَّر (التابع).
- في مثالنا هذا فانه يشار إلى سنوات الخبرة كمتغير مستقل ، ويستخدم للتنبؤ بحجم المبيعات أو المتغير التابع .
- جرت العادة في الرسم الانتشاري أن يكون المتغير المستقل على الإحداثي الأفقي والمتغير التابع على الإحداثي الرأسي .

- يشير الرسم إلى أن هناك فرصة طيبة لوجود علاقة بين المتغيرات . حيث أن سنوات الخبرة المنخفضة يترافق مع انخفاض حجم المبيعات ، وارتفاع سنوات الخبرة يترافق في الغالب مع ارتفاع حجم المبيعات السنوية.
- تبدو هذه العلاقة بين المتغيرين خطاً مستقيماً تقريباً أو معادلة خطية .
- نبين في ما يلي كيفية تطوير مثل هذه العلاقة الخطية

خط الانحدار البسيط:

- نحاول هنا حصر أنفسنا بمهمة إيجاد خط مستقيم يمثل الرسم الانتشاري للبيانات أفضل تمثيل. أي أننا سنوفق معادلة على الشكل التالي :

$$\hat{y} = a + bx$$

حيث:

$\hat{y}$ = القيمة المقدرة للمتغير التابع .

x = قيمة المتغير المستقل .

a = تقاطع المحور y (أي أنها قيمة y عندما تكون x= 0)

b = تغير المتغير التابع y نتيجة لتغير المتغير المستقل (ميل خط الانحدار).

- في تحليل الانحدار تكون المعادلة المستخدمة لوصف البيانات هي معادلة الانحدار المقدرة.
- نركز في هذا الجزء على معادلات خط الانحدار التي تأخذ شكل خط مستقيم.
- يشار إلى معادلة خط الانحدار بأنها خط الانحدار المقدر . ويشار عموماً إلى علاقة الخط المستقيم المعنية بمتغيرين أحدهما مستقل والآخر تابع على أنها انحدار خطي بسيط .
- لتوضيح فكرة خط الانحدار البسيط ، نأخذ المثال المقدم في جدول (1) بخصوص حجم المبيعات السنوية وسنوات الخبرة لعشرة مندوبي مبيعات . فباستخدام الرسم الانتشاري المبين في شكل (3)

- سنحاول في البداية وضع خط يعتقد أنه الأقرب لإحداثيات البيانات المرسومة بحيث نتمكن من تحديد بداية الخط (a). إضافة إلى ذلك، b_1 فإننا باستعمال أي نقطتين ، يمكننا تحديد الميل (   ) . وإذا ما كلفنا شخصين (باحثين) ليقوما بهذه المهمة، فإنهما سيخرجا بمعادلتي خط انحدار مختلفين، ولتكونا كما يلي:

$$\hat{y} = 78 + 4.2x$$

الباحث الأول : $\hat{y} = 81 + 3.9x$

الباحث الثاني :

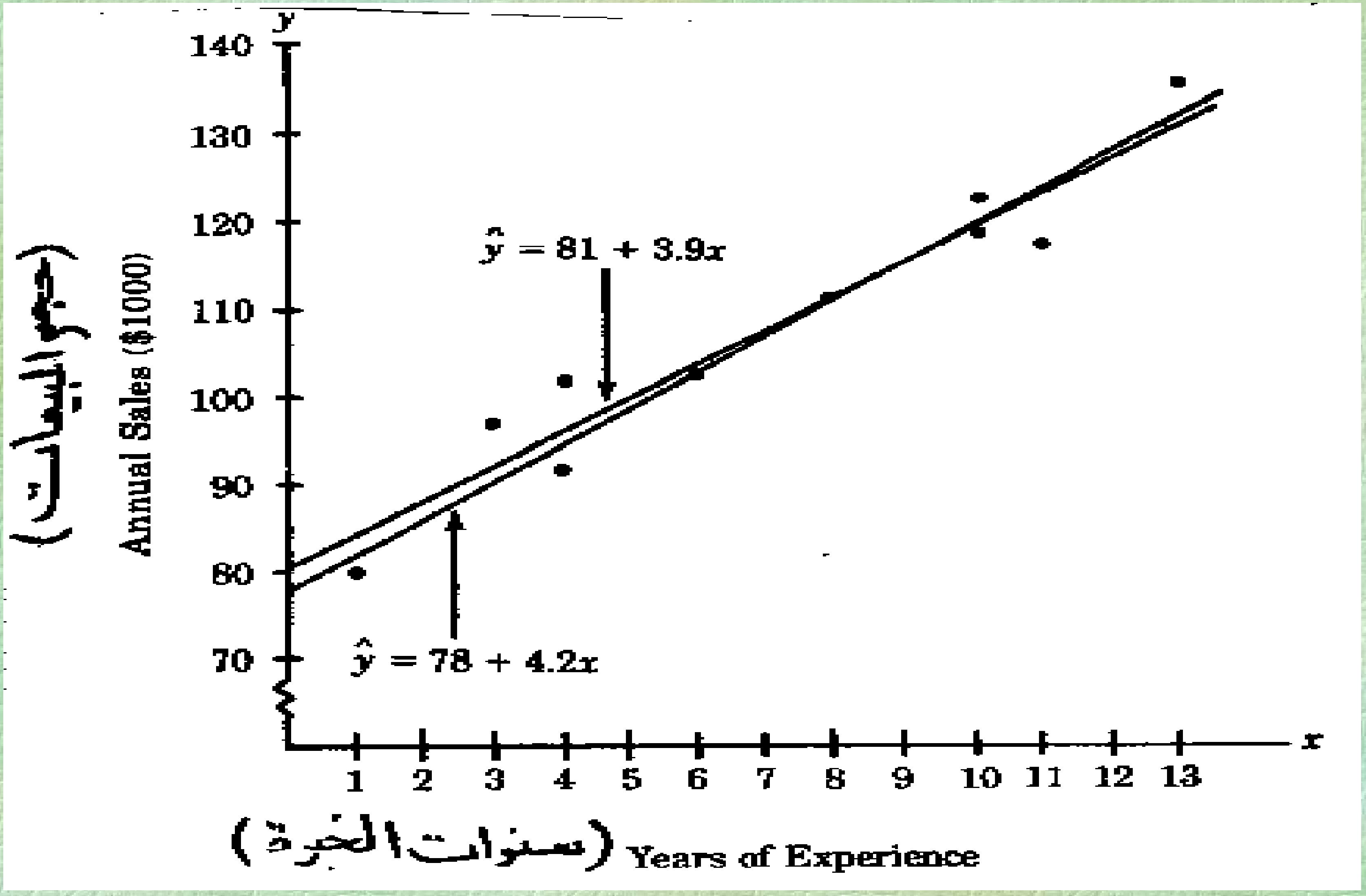
حيث: $\hat{y}$

= القيمة المقدرة للمبيعات السنوية (بآلاف الدنانير) .

x = سنوات الخبرة في مجال المبيعات .

- يبين الشكل (4) مدى ملاءمة هذين الخطين المستقيمين للبيانات الموضحة في الشكل (3) . ولكن رغم أن الخطين مشابهين إلى حد كبير للخط الذي يمكن أن نرسمه ، إلا أننا لا نمتلك معياراً لاختيار أفضل خط. إن اختلاف نظرة الأشخاص لنفس البيانات وخروجهم بعلاقات رياضية مختلفة يمكن أن يؤثر على صانعي القرار . فلو وضع شخص آخر خطاً مختلفاً ، ربما يتساءل صانع القرار ، هل يمكن أن يكون هناك قرار آخر؟ وعليه فإننا بحاجة إلى اتفاق على معيار أو مقياس معين لاختيار خط الانحدار

الشكل (4)



y
140
130
120
110
100
90
80
70
Annual Sales ($1000)
(حجم المبيعات)
$\hat{y} = 81 + 3.9x$
$\hat{y} = 78 + 4.2x$
x
1 2 3 4 5 6 7 8 9 10 11 12 13
Years of Experience
(سنوات الخبرة)

طريقة المربعات الصغرى:

- هذه الطريقة تعطي أكفأ تقدير لمعادلة الانحدار الخطي ، بحيث تجعل من مجموع مربعات الاختلافات بين القيم المشاهدة للمتغير التابع (Y) والقيم المقدرة لهذا المتغير ($\hat{Y}$) عند نهايتها الصغرى . ولنلق نظرة متفحصة على مدلول مجموع مربعات الاختلافات بين القيم المشاهدة والقيم المقدرة للمتغير التابع ، ومعناها الحقيقي .

- ومن أجل مزيد من التوضيح سنسلط الضوء على جزء من الرسم الانتشاري فقط لأول مشاهدتين وقيم المقدرة لهما على خط الانحدار $\hat{y} = 78 + 4.2x$.
- وبالنظر إلى الشكل التالي (5)، يمكن ملاحظة الفرق بين القيمة المشاهدة من المبيعات السنوية لمندوب المبيعات الأول (y_1) والقيمة المقدرة لهذه المبيعات ($\hat{y}_1$) هو كما يلي : $(y_1 - \hat{y}_1) = 80 - 82.2 = -2.2$

- وبالتالي فان مربع الفرق سيكون :

$$(y_1 - \hat{y}_1)^2 = (-2.2)^2 = 4.84$$

- وسيكون الفرق للمشاهدة الثانية: $(y_2 - \hat{y}_2) = 6.4$

- ومربع الفرق $(y_2 - \hat{y}_2)^2 = 40.96$

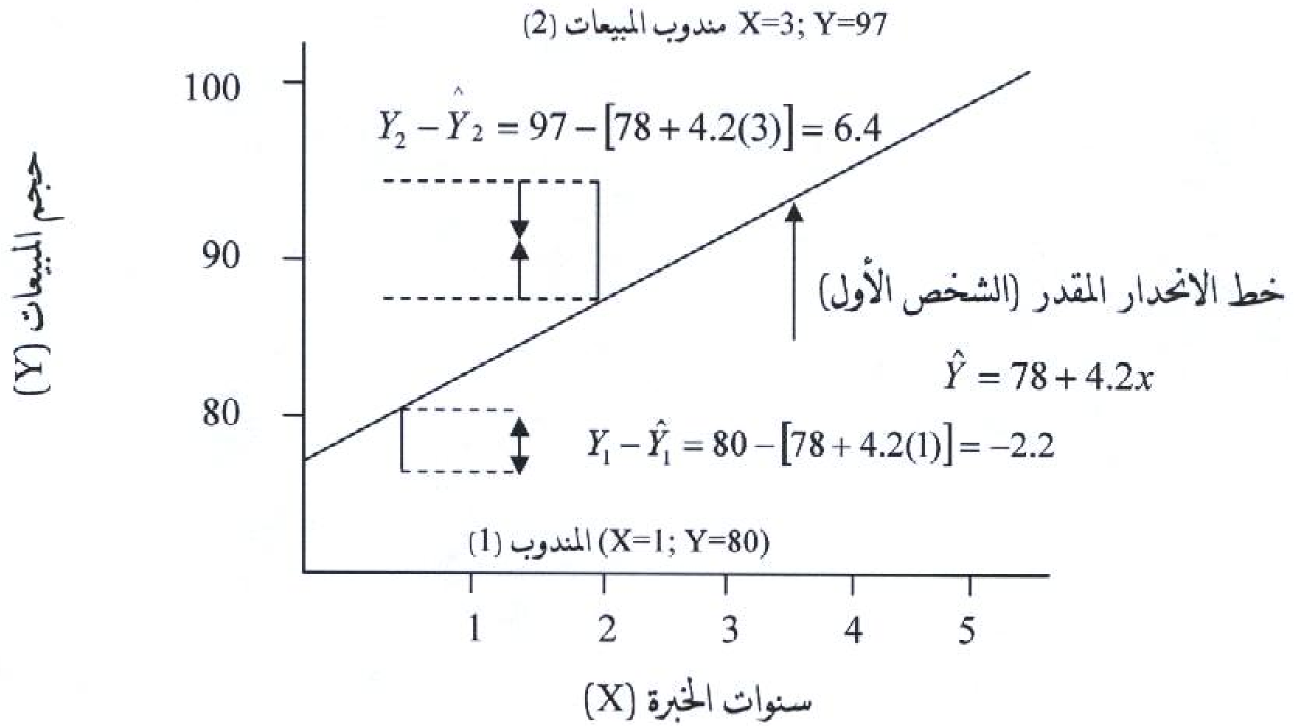
ويبين الجدول التالي (2) بقية الحسابات للمشاهدات الثمانية الأخرى .

جدول (2): حساب مجموع مربعات الاختلافات لخط الانحدار المقدر

$$\hat{y} = 78 + 4.2x$$

مندوب المبيعات	سنوات الخبرة (X_i)	القيم المشاهدة لحجم المبيعات (Y_i)	القيم المقدرة للمبيعات $(\hat{y}_i = 78 + 4.2x_i)$	الفروق بين القيم المشاهدة والمقدرة $(y_i - \hat{y}_i)$	مربعات الفروق $(y_i - \hat{y}_i)^2$
1	1	80	78+4.2(1)=82.2	-2.2	4.84
2	3	97	78+4.2(3)=90.6	6.4	40.96
3	4	92	78+4.2(4)=94.8	-2.8	7.84
4	4	102	78+4.2(4)=94.8	7.2	51.84
5	6	103	78+4.2(6)=103.2	-0.2	0.04
6	8	111	78+4.2(8)=111.6	-0.6	0.36
7	10	119	78+4.2(10)=120.0	-1.0	1.00
8	10	123	78+4.2(10)=120.0	3.0	9.00
9	11	117	78+4.2(11)=124.2	-7.2	51.84
10	13	136	78+4.2(13)=132.6	3.4	11.56
المجموع					**179.28**

الشكل (5)



(2) مندوب المبيعات X=3; Y=97
100
90
80
$Y_2 - \hat{Y}_2 = 97 - [78 + 4.2(3)] = 6.4$
خط الانحدار المقدر (الشخص الأول)
$\hat{Y} = 78 + 4.2x$
$Y_1 - \hat{Y}_1 = 80 - [78 + 4.2(1)] = -2.2$
(1) المندوب (X=1; Y=80)
1
2
3
4
5
سنوات الخبرة (X)
حجم المبيعات (Y)

- من جدول (2) نلاحظ أن مجموع مربعات الاختلافات = 179.28. وقد قمنا أيضاً باحتساب مجموع مربعات الاختلافات للخط المقدر الآخر: $\hat{y} = 81 + 3.9x$ والذي أعد من قبل الباحث الثاني في مثالنا أعلاه. وقد كانت النتيجة 172.32. وحيث أن مجموع المربعات لهذا الخط المقدر هي أصغر أو أقل من الخط الأول فإننا نعتبر هذا الخط أكفأ لتقدير خط الانحدار من الخط الأول . ولكن طريقة المربعات الصغرى لا تعتبر هذين الخطين هما الخطين الوحيدين. فهي التي توجد في الحقيقة خط الانحدار المقدر الذي يعطي أصغر مجموع لمربعات الفروق (الاختلافات) لكل الخطوط المختارة .

تقدير خط الانحدار بطريقة المربعات الصغرى:

- عند استعمال طريقة المربعات الصغرى لتحديد خط الانحدار المقدر $\hat{y} = a + bx$ ، فان الإحصائيين يرون بأن أفضل قيم لكل من a و b يمكن إيجادها باستعمال المعادلات التالية :

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \quad ....(1)$$

$$a = \bar{y} - b_1 \bar{x} \quad ....(2)$$

حيث أن :

x_i = قيم المتغير المستقل لـ (i) من المشاهدات .

y_i = قيم المتغير التابع لـ (i) من المشاهدات .

$\bar{x}$ = قيمة متوسط المتغير المستقل .

$\bar{y}$ = قيمة متوسط المتغير التابع .

n = عدد المشاهدات .

- ويبين الجدول التالي (3) بعض الحسابات الضرورية لاحتساب خط الانحدار المقدر .

جدول (3): بعض الحسابات الضرورية لتقدير معاملات خط الانحدار

x_i^2	$x_i y_i$	y_i	سنوات الخبرة x_i	مندوبي المبيعات (i)
1	80	80	1	1
9	291	97	3	2
16	368	92	4	3
16	408	102	4	4
36	618	103	6	5
64	888	111	8	6
100	1190	119	10	7
100	1230	123	10	8
121	1287	117	11	9
169	1768	136	13	10
632	8,128	1,080	70	المجموع

- وباستعمال الأرقام والقيم الموجودة في الجدول (3) وباستخدام الصيغ الرياضية 1 و 2 الآنفة الذكر ، يمكننا احتساب الميل (b) ونقطة تقاطع خط الانحدار مع (y) وهي a كما يلي :

(1) احتساب $\bar{x}$ و $\bar{y}$:

$$\bar{x} = \frac{\sum x_i}{\eta} = \frac{70}{10} = 7$$

$$\bar{y} = \frac{\sum y_i}{\eta} = \frac{1080}{10} = 108$$

(2) احتساب الميل (b) :

$$b = \frac{n\sum x_i \sum y_i - \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{10(8,128) - 70(1080)}{10(632) - (70)^2}$$

$$= \frac{5680}{1420} = 4$$

(3) احتساب تقاطع (a) :

$$a = \bar{y} - b_1 \bar{x}$$

$$= 108 - 4(7)$$

$$= 80$$

وعليه تكون معادلة خط الانحدار المقدرة كما يلي :

$$\hat{y} = 80 + 4x$$

- بملاحظة أن الميل b موجب ، هذا يعني أنه كلما زادت سنوات الخبرة (x) لدى العاملين في المبيعات كلما زاد حجم المبيعات (y) . ويتضح من المسألة الموجودة بين أيدينا أن هناك علاقة إيجابية بين x و y . ولكن في حالات أخرى قد يكون الميل سالباً ليشير إلى أنه كلما زادت x انخفضت y ، وفي هذه الحالة تكون العلاقة بين x و y سالبة أو عكسية .

- ولو قمنا باحتساب مجموع مربعات الفروق (الاختلافات) بين القيم المشاهدة والقيم المقـــدرة استنتـــاداً إلى خط الانحدار المقدر $\hat{y} = 80 + 4x$ بطريقة المربعات الصغرى $\sum (y - \hat{y})^2 = 170$ ، فإننا نجد أن وهي قيمة أقل من مجموع المربعات للخطين الذين نوقشا سابقاً . ومن المهم الإشارة هنا إلى أن طريقة المربعات الصغرى تضمن عدم وجود أي خط آخر يمكن أن يؤمن مجموع مربعات أصغر من في مثالنا الحالي . لذلك نعتبر أن طريقة المربعات الصغرى هي التي تضمن أكفأ تقدير لخط الانحدار .

▪ ومع الاعتقاد بأن معادلة خط الانحدار المقدرة بطريقة المربعات الصغرى تصف العلاقة بين x و y بشكل دقيق ، فانه يبدو أنه من المعقول استخدام هذه المعادلة الرياضية في تقدير قيم y إذا أعطينا قيم x . وفي مثالنا الحالي المتعلق بمسألة حجم المبيعات السنوية وسنوات الخبرة ، فانه يمكننا استخدام معادلة خط الانحدار المقدرة بطريقة المربعات الصغرى $\hat{y} = 80 + 4x$ في تقدير حجم المبيعات السنوية المتوقعة لثلاثة متقدمين لوظيفة مندوب مبيعات . ويبين الجدول (4) نتائج عملية التقدير .

جدول (4) : استخدام معادلة خط الانحدار في تقدير حجم المبيعات السنوية لثلاثة متقدمين

المتقدم للوظيفة	سنوات الخبرة	حجم المبيعات السنوية المقدرة (بالألف دولار)
رشوان محمد	5	$80 + 4(5) = 100$
علي مرتّجي	2	$80 + 4(2) = 88$
يحيى الربيعي	0	$80 + 4(0) = 80$

- من الجدول السابق نرى أن تحليل الانحدار السابق يعطي دعماً لقرار أن السيد رشوان محمد هو الشخص الأفضل كمندوب مبيعات ، حيث يمكن تقدير مبيعاته خلال السنة الأولى من التحاقه بالشركة بحوالي 100,000 دولار ، ويفوق بذلك المبيعات المقدره لعلي مرتجى المتقدم الثاني للوظيفة بحوالي 12,000 دولار. في هذا المثال لمسنا أحد أهم استخدامات تحليل الانحدار : بتزويد صيغة رياضية تدعم بالمعلومات صانعي القرار في المجال التجاري والاقتصادي .

- إن معادلة خط الانحدار المقدرة أعلاه تسمى معادلة خط انحدار y على x حيث أننا من خلالها يمكن تقدير أو التنبؤ بقيمة y عند معرفة قيمة x . وإذا ما كان المتغير التابع هو x والمتغير المستقل هو y فان المعادلة المقدرة ستكون معادلة انحدار x على y وبالتالي فإننا سنتنبأ من خلالها بقيم x بعد معرفة قيم y .

قياس جودة التقدير:

- والخطوة التالية في دراسة الانحدار هي قياس جودة التقدير في نموذج خط الانحدار. ولكن قبل ذلك لا بد من وضع مجموعة من الفرضيات (في قيم البيانات) تمكننا من إنجاز تحليلات إحصائية إضافية .
- يفترض النموذج أن القيم الموضوعة للمتغير المستقل x يمكن أن تقابلها قيم للمتغير التابع y . وبالتالي فان تغير y ناتج عن (1) التغير في x (2) التغير الباقي e وهو تغير عشوائي غير مفسر من قبل النموذج.

- سوف نستعرض هنا في هذا الجزء طريقة إحصائية لقياس أو وصف جودة خط الانحدار المقدر . وقد كنا قد استعرضنا طريقة المربعات الصغرى كتقنية من أجل تقليل مجموع مربعات الفروق (الاختلافات) بين القيم المشاهدة للمتغير التابع (y_i) والقيم المقدرة له ($\hat{y}_i$). ويلاحظ أن الفروق تمثل في الحقيقة خطأ استعمال $\hat{y}_i$ كقيم مقدرة لـ y_i . وعليه فان مجموع المربعات الناتجة يمكن أن يشار إليها على أنها مجموع مربعات الأخطاء ، وسوف نرمز إليها بـ SSE .

$$SSE=\sum(y_i-\hat{y}_i)^2$$

حيث أن :

- وبالعودة إلى مثالنا السابق المتعلق بحجم المبيعات وسنوات الخبرة في البيع فقد كانت $\sum(y_i - \hat{y}_1)^2 = 170$ وعليه فان SSE = 170 هي عبارة عن مقدار الخطأ الناتج عن استخدام خط الانحدار المقدر ($\hat{y} = 80 + 4x$) لتقدير قيم y .

- ولكننا لو سئلنا عن تقدير حجم المبيعات السنوية لمندوب المبيعات دون اللجوء إلى أو دون معرفة سنوات الخبرة لكل مندوب وبدون اللجوء إلى معادلة خط الانحدار ، فان متوسط مبيعات العينة Y وهي $\bar{Y} = 108$ ، سيكون هو التقدير الأمثل بالنسبة لنا . والآن انظر إلى الأخطاء أو الفروق التي حصلنا عليها لو استعملنا المتوسط لكل مندوبي المبيعات (انظر الجدول 5). إن مجموع مربعات التشتت للمشاهدات الحقيقية حول المتوسط (108) هو 2,442 . وهذه القيمة تمثل مجموع المربعات الكلي للفروق قبل تحليل الانحدار ، ويشار إليها عموماً على أنها المجموع الكلي لمربعات

$$SST = \sum (Y_i - \bar{Y})^2 = 2442$$

الاختلاف حول المتوسط . وسوف نرمز لها بـ SST . وفي

جدول (5): مجموع مربعات الاختلاف حول المتوسط
SST (108)

مندوب المبيعات	سنوات الخبرة (x_i)	(y_i)	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	1	80	-28	784
2	3	97	-11	121
3	4	92	-16	256
4	4	102	-6	36
5	6	103	-5	25
6	8	111	3	9
7	10	119	11	121
8	10	123	15	225
9	11	117	9	81
10	13	136	28	784
$\sum$				2242

- فلو كان مقدار الاختلاف هو 2442 (SST) قبل تحليل الانحدار و (170) SSE بعد تحليل الانحدار ، فيمكن الاستنتاج أن الفرق (2272) هو مجموع مربعات الاختلاف (التباين) المفسرة بواسطة معادلة خط الانحدار . وعموماً يدعى مجموع المربعات هذا بمجموع مربعات الانحدار ويرمز له بــ SSR وهو الاختلاف المفسر . ورغم أن الطريقة أعلاه هي الطريقة المستخدمة عادة في احتساب الاختلاف المفسر (SSR)، إلا أنه يمكن تبيان طريقة مباشرة لاحتساب SSR باستخدام الصيغة التالية:

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

- واحتساب (SSR) باستخدام الصيغة أعلاه مبين في جدول (6) ، وقد تم الحصول على قيمة (2272=SSR) وهي قيمة مطابقة لما تم الحصول عليه أعلاه . والعلاقة بين SSE ، SST و SSR تشكل أساساً لواحدة من أهم النظريات في الإحصاء التطبيقي . وتقول هذا النظرية ، بشكل عام : أن مجموع مربعات المشاهدات حول متوسطها (SST) يمكن تجزئتها إلى جزئين : SSE و SSR بحيث :

$$SST = SSE + SSR$$

جدول (6) : الاحتساب المباشر لمجموع مربعات الانحدار (SSR)

مندوب المبيعات	سنوات الخبرة x_i	حجم المبيعات y_i	حجم المبيعات المقدرة $(\hat{y} = 80 + 4x_i)$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	1	80	84	-24	576
2	3	97	92	-16	256
3	4	92	96	-12	144
4	4	102	96	-12	144
5	6	103	104	-4	16
6	8	111	112	4	16
7	10	119	120	12	144
8	10	123	120	12	144
9	11	117	124	16	256
10	13	136	132	24	576
$\sum$					2272

- والآن ، لنرى كيف يمكن استعمال العلاقة المبينة أعلاه في تطوير مقياس لجودة التوفيق لمعادلة خط الانحدار المقدرة .
- سيكون لدينا توفيق كامل لخط الانحدار المقدر لو كانت كل المشاهدات تقع على خط مستقيم . وفي هذه الحالة فان خط الانحدار المقدر بطريقة المربعات الصغرى سيمر في جميع النقاط . ولذلك ستكون SSE =0. وسيكون SSR = SST ، وبالتالي SST/SSR = 1. ومن جانب آخر فان توفيقاً ضعيفاً للبيانات المشاهدة ينتج عن SSE كبيرة . وحيث أن SSR + SSE = SST ، فان أكبر SSE (وبالتالي أضعف جودة توفيق) سيحدث عندما تكون SSE = SST ، وفي هذه الحالة فان SSR = صفر ، وعندها لن يكون لخط الانحدار المقدر أي دور في المساعدة بتخمين قيم y . وهكذا فان أسوأ توفيق ممكن لخط الانحدار يكون عندما تكون SSR = صفر وعندها تكون نسبة SST/SSR = صفر .

- وإذا ما استخدمنا نسبة SST/SSR لتقييم مدى جودة التقدير لخط الانحدار ، سيكون لدينا مقياس يمكن أن يأخذ قيماً بين صفر و 1 ، وكلما اقتربت القيمة من 1 يدل على جودة أفضل في التقدير لخط الانحدار. والكسر الناتج عن SST/SSR يسمى معامل التحديد ويرمز له بالرمز R^2 . وهو عبارة عن النسبة بين الاختلافات المفسرة والاختلاف الكلي .

معامل التحديد (R^2) = $\frac{SSR}{SST}$

وقيمة معامل التحديد في مثالنا السابق $R^2 = \frac{SSR}{SST} = \frac{2272}{2442} = 0.93$

- ومن أجل توضيح أكثر لــــ (R^2) ، يمكن التفكير بـــ SST كمقياس لمدى جودة $\hat{y}$ كمخمن لحجم المبيعات السنوية . وبعد توفيق خط الانحدار المقدر ، نحتسب SSE كمقياس الجودة المخمن لحجم المبيعات السنوية. وعليه فإن SSR - الفرق بين SST و SSE - يقيس حقيقة R^2 ذلك الجزء من SST المفسر من خلال توفيق خط الانحدار المقدر . وعليه فإننا نفكر بـــ كما في الصيغة التالية :

$$R^2 = \frac{\text{مجموع مربعات الاختلاف المفسرة بخط الانحدار}}{\text{مجموع مربعات الاختلاف قبل خط الانحدار}}$$

- وعندما يعبر عنها كنسبة مئوية R^2 ، فان كنسبة مجموع المربعات المفسرة بخط الانحدار المقدر. وفي مثالنا الماضي نستنتج أن معادلة خط الانحدار تفسر ما نسبة 93% من مجموع R^2 مربعات الكلي . ونكون في أشد السعادة إذا حصلنا على مثل هذه القيمة لـــ ، حيث أن الباحثين في الدراسات الاقتصادية والتجارية يشعرون بنتائج جيدة إذا حصلوا على R^2 مساوية لـ 0.60 وأكثر .

الارتباط:

- هناك بعض الحالات التي يكون فيها صانع القرار غير معني بالمعادلة التي تربط بين متغيرين أو بتخمين أو التنبؤ بالمتغير التابع إذا عرفنا قيمة المتغير المستقل . فهناك طريقة أو أسلوب إحصائي في هذه الحالات يستخدم للقياس ومعرفة اتجاه العلاقة بين متغيرين أو أكثر، وهذا الأسلوب الارتباط . ومن أهم مقاييس الارتباط ما يشار بمعامل الارتباط (R). ويكون هذا المعامل محصوراً بين +1 و -1 . ويشير +1 إلى أن العلاقة إيجابية (طردية) وكاملة بين المتغيرين ، وعندها ستكون جميع النقاط الموجودة في الرسم الانتشاري واقعة على خط مستقيم بميل إيجابي . إما القيمة -1 فتشير إلى أن العلاقة بين X و Y كاملة وبشكل سلبي أو عكسي وتقع جميع النقاط على خط مستقيم

- إذن قد يكون الارتباط بين المتغيرين طردياً (موجباً) بمعنى أن تزايد قيمة أحد المتغيرين يصاحبها تزايد في قيم لمتغير الآخر . وقد يكون الارتباط عكسياً (سالباً) إذا كانت القيم الصغيرة لأحد المتغيرين تصاحبها قيم كبيرة للمتغير الآخر . وعليه فان معامل الارتباط يعكس خاصتين هامتين هما : اتجاه العلاقة (طردي أو عكسي) ، ودرجة هذه العلاقة وقوتها . حيث تحدد الإشارات الجبرية اتجاه الارتباط بينما تعكس القيمة المطلقة درجة الارتباط .
- فيما يلي أهم الصيغ المستخدمة لاستنباط معامل الارتباط :

أولاً: تحديد معامل الارتباط من خلال تحليل الانحدار:

- فمن خلال مناقشتنا السابقة للانحدار الخطـــي كنـــاقد افترضنا أن معـــادلة الانـحدار الخطي $\hat{y} = a + bx$ بطريقة المربعات الصغرى هي . وفي r^2 هذه الحالة فان معامل الارتباط يمكن احتسابه من معامل التحديد ( ) كما يلي :

$$r = \mp\sqrt{r^2}$$

$$= \mp\sqrt{\qquad\qquad}$$

معامل التحديد

- وتتحدد إشارة معامل الارتباط من خلال إشارة الميل (b) في معادلة خط الانحدار . وفي مثالنا السابق فان معامل الارتباط هو :

$$r = \mp\sqrt{0.93}$$
$$= +0.96$$

- وحيث أن الميل (b = 4) كان موجباً فإن معامل الارتباط هو إيجابي أيضاً.

ثانياً: معامل ارتباط بيرسون:

- وهو معامل يمكن اللجوء إليه عندما لا يكون صانع القرار معنياً بالعلاقة الدالية بين المتغيرين x و y ، بل يكون معنياً فقط بالعلاقة بين المتغيرين من عدمها. فيمكن احتساب معامل الارتباط دون اللجوء إلى إنجاز تحليل خط الانحدار. وفي هذه الحالة فان الصيغة المستخدمة هي :

$$R = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- وفي مثالنا السابق حول حجم المبيعات السنوية فان الحسابات الضرورية لاستخدام الصيغة أعلاه موضحة في الجدول (7)

جدول (7) : الحسابات الضرورية لاستخراج معامل الارتباط

y_i^2	x_i^2	$x_i y_i$	حجم المبيعات y_i	سنوات الخبرة x_i	مندوب المبيعات
6400	1	80	80	1	1
9409	9	291	97	3	2
8464	16	368	92	4	3
10404	16	408	102	4	4
10609	36	618	103	6	5
12321	64	888	111	8	6
14161	100	1190	119	10	7
15129	100	1230	123	10	8
13689	121	1287	117	11	9
18496	169	1768	136	13	10
119.082	632	8,128	1080	70	المجموع

- وبتطبيق الصيغة الرياضية أعلاه فان :

$$R = \frac{10(8,128) - 70(1,080)}{\sqrt{10(632) - (70)^2}\sqrt{10(119,082) - (1,080)^2}}$$

$$= \frac{81,280 - 75,600}{\sqrt{1420}\sqrt{24,420}} = 0.96$$

- ويلاحظ أن قيمة معامل الارتباط التي حصلنا عليها هي نفس القيمة التي احتسبت كجذر تربيعي لمعامل التحديد.

- وهناك خصائص مميزة يتمتع بها معامل بيرسون ، تسهل من عملية احتسابه وهي :

1. أنه لا يتأثر بإضافة أو طرح أي ثابت على/من جميع قيم أي من المتغيرين .
2. كذلك فانه لا يتأثر بضرب جميع قيم أي من المتغيرين في أية كمية ثابتة .

- ولتوضيح أهمية هاتين الخاصيتين سنحاول طرح مثال مبسط يمكن من خلاله استغلال وتسهيل عملية الاحتساب للصيغة المذكورة لارتباط بيرسون .

جدول (8) : الحسابات الضرورية لاستخراج معامل الارتباط

الشهر	نفقات للدعاية x_i	حجم المبيعات بالألف y_i	$x_i y_i$	y_i^2	x_i^2
1	2	60	120	3600	4
2	5	100	500	10000	25
3	4	70	280	4900	16
4	6	90	540	8100	36
5	3	80	240	6400	9
المجموع	20	400	1680	33000	20

- وبتطبيق الصيغة المذكورة لمعامل بيرسون :

$$R = \frac{5(1680) - 20(400)}{\sqrt{5(90) - (20)^2}\,\sqrt{5(33300) - (400)^2}}$$

$$= \frac{8,400 - 8,000}{\sqrt{50\ X\ 5,000}} = \frac{400}{500}$$

$$= 0.8$$

- هذا يعني أن العلاقة بين نفقات الدعاية وحجم المبيعات طردية ، فكلما زادت نفقات الدعاية كلما زادت المبيعات .
- وإذا ما استخدمنا أحد الخاصيتين المذكورتين بقسمة (Y) على 10 فإن النتائج ستكون كالتالي :

جدول (9): الحسابات الضرورية لاحتساب معامل الارتباط

بعد قسمة أحد المتغيرين (y) على 10

الشهر	X_i	Y_i	$Y_i/10$	X_i^2	$(Y_i/10)^2$	$X_i(Y_i/10)$
1	2	60	6	4	36	12
2	5	100	10	25	100	50
3	4	70	7	16	49	28
4	6	90	9	36	81	54
5	3	80	8	9	64	24
المجموع	20	400	40	90	330	168

- وبتطبيق المعادلة المذكورة أعلاه حصلنا على معامل الارتباط = + 0.8 وهي نفس النتيجة التي حصلنا عليها عند استخدام العمود (Y) الأصلي .

ثالثاً: معامل ارتباط سبيرمان (ارتباط الرتب) Rank Correlation:

- يقدم معامل بيرسون مقياساً ممتازاً للارتباط ولكنه يعتبر أفضل ما يكون في ظروف معينة ويفقد ميزته تحت ظروف أخرى حيث لا يصلح للاستخدام إذا كان المتغير لا يمكن قياسه مثل تقديرات الطلاب حيث يمكن أن تعرف أن (أ) أفضل من (ب) دون معرفة بكم تماماً، فعندما تكون المعلومات في صورة كيفية أن تظهر أو تصف وضعاً معيناً كالحالة الزواجية أو التعليمية لأو حالة الطقس مثلاً، وفي هذه الحالات يتم اللجوء إلى بديل أكثر كفاءة يعتمد على ترتيب قيم كل متغير بينها وبين نفسها ثم تطرح الرتب من بعضها لتعرف بـ (d = فروق الرتب) ويحسب معامل ارتباط الرتب حسب المعادلة التالية :

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

- بتطبيق هذه المعادلة على المثال السابق (العلاقة بين الإنفاق على الدعاية وبين المبيعات) .

D^2	الفرق (d)	ترتيب قيم Y_i	حجم المبيعات Y_i	ترتيب قيم X_i	نفقات الدعاية X_i	الشهر
صفر	صفر	5	60	5	2	1
1	1	1	100	2	5	2
1	1	4	70	3	4	3
1	1	2	90	1	6	4
1	1	3	80	4	3	5
4						المجموع

$$
\begin{aligned}
r &= 1 - \frac{6 \times 4}{125 \quad - 5} \\
&= 1 - \frac{24}{120} \\
&= 1 - 0.2 \\
&= 0.8
\end{aligned}
$$

- وعادة ما يعطي بيرسون وسبيرمان قيماً متقاربة . ولو أن التطابق في هذا المثال لا يعدو أن يكون مجرد مصادفة . ويمكن التحقق بسهولة من أن معامل سبيرمان يتمتع كما يتمتع بيرسون بالخواص التي تم ذكرها لهذا المعامل وهي أنه لا يتأثر بطرح أو إضافة أو حتى قسمة أي من المتغيرين على أي ثابت.
- مثال : عند احتساب العلاقة بين عدد الأطباء ومعدل الوفيات في دولة الكويت حسب طريقة سبيرمان كانت النتائج كما هي مبينة أدناه :

السنوات	عدد الأطباء X_i	معدل الوفيات لكل ألف من السكان الكويتيين (Y_i)	ترتيب X_i	ترتيب Y_i	الفرق (d)	D^2
1973	990	6.7	16	1	15	225
1974	1087	6.6	15	2	13	169
1975	1178	6.1	14	4	10	100
1976	1341	5.6	13	5	8	64
1977	1567	6.2	12	3	9	81
1978	1703	5.4	11	6.5	4.5	20.25
1979	1957	5.4	10	6.5	3.5	12.25
1989	2341	5.1	9	8	1	1
1981	2580	4.5	8	10	2	4
1982	2734	4.6	7	9	2	4
1983	2834	4.2	6	11	5	25
1984	2983	3.9	5	13	8	64
1985	3095	4.0	4	12	8	64
1986	3129	3.5	2	14	12	144
1987	3128	3.4	3	15.5	12.5	156.25
1988	3256	3.4	1	15.5	14.5	210.25
المجموع						1344

$$r = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 1344}{(16)^3 - 16} = 1 - \frac{8064}{4096 - 16}$$

$$= 1 - \frac{8064}{4080} = -0.9764$$

- وتدل هذه النتيجة على أن العلاقة قوية وعكسية بين عدد الأطباء ومعدل الوفيات ، فكلما زاد عدد الأطباء كلما انخفضت نسبة الوفيات.