١٠ العلاقة بين متغيرين: الانحدار والارتباط البسيطان

Simple regression and correlation

١ - مقدمة

٢ - معادلة خط الانحدار

٣ - استخدام المصفوفات فى الانحدار

٤ - مصادر الاختلاف فى الانحدار الخطي

٥ - القيم المعدلة لأثر انحدار Y على X

٦ - الانحرافات المعيارية وحدود الثقة للتقديرات المختلفة

٧ - اختبار معنوية معامل الانحدار

٨ - المقارنة بين معاملى الانحدار

٩ - التوزيع ذو المتغيرين

١٠ - العلاقة بين معامل التحديد وخطأ التقدير

١١ - تقييم ملائمة نموذج التحليل

١٢ - الارتباط البسيط

١٣ - العلاقة بين معامل الانحدار ومعامل الارتباط

١٤ - اختبار معنوية معامل الارتباط وتقدير حدود الثقة له

١٥ - اختبار تساوى معاملى ارتباط

١٦ - اختبار تجانس عدد من معاملات الارتباط

١٧ - السبب والأثر فى تحليل الارتباط والانحدار

١٨ - تباين الدالة الخطية

١٩ - ارتباط الرتب

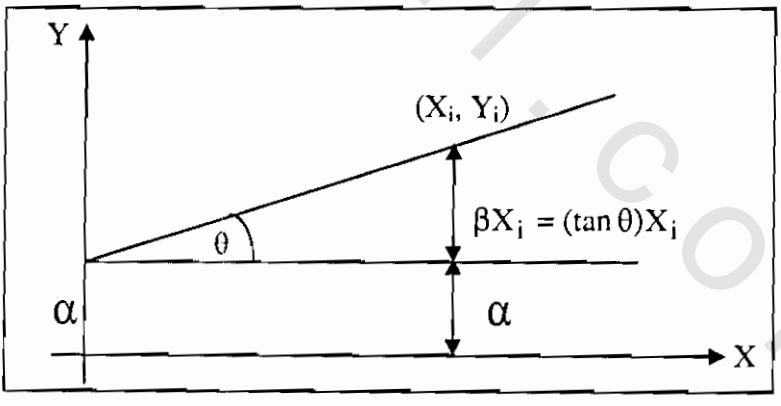
١٠-١ مقدمة

يطلق على الانحدار regression الارتداد أو الاعتماد ويستخدم لدراسة العلاقة بين متغيرين أحدهما Y والذى يعتمد فى قيمته على متغير آخر X. ويطلق على Y المتغير التابع أو المعتمد dependent variable بينما يطلق على X المتغير المستقل independent variable وقد يسمى أيضاً بالمتغير التفسيرى explanatory variable أى الذى يفسر التغيرات فى Y. ويوجد أمثلة كثيرة لذلك منها دراسة العلاقة بين الأسمدة X وكمية المحصول Y، دراسة العلاقة بين العمر عند أول ولادة X ومحصول اللبن Y، دراسة العلاقة بين الطول X والوزن Y، دراسة العلاقة بين محيط الصدر X والوزن Y. وعادة ما يكون المتغير المستقل هو المتغير الأسهل فى القياس عن المتغير التابع. ويعبر عن العلاقة بين المتغيرين رياضياً $Y = f(X)$، أى أن Y دالة X ويعبر عنها إحصائياً بالانحدار.

وقد تعبر المعادلة $Y = 3 + 5X$ على سبيل المثال عن العلاقة بين المتغيرين، وهى تمثل معادلة خط مستقيم صورته العامة:

$$Y = \alpha + \beta X \qquad (١٠-١)$$

حيث α هى الجزء المقطوع من محور الصادات intercept، أى قيمة Y عندما تكون قيمة X مساوية للصفر، β تمثل ميل slope الخط المستقيم، أى ظل tan الزاوية θ التى يصنعها المستقيم فى الاتجاه الموجب لمحور السينات. كما يطلق على β معامل الانحدار أو الاعتماد egression coefficient. وشكل ١٠-١ يمثل تلك العلاقة بيانياً.

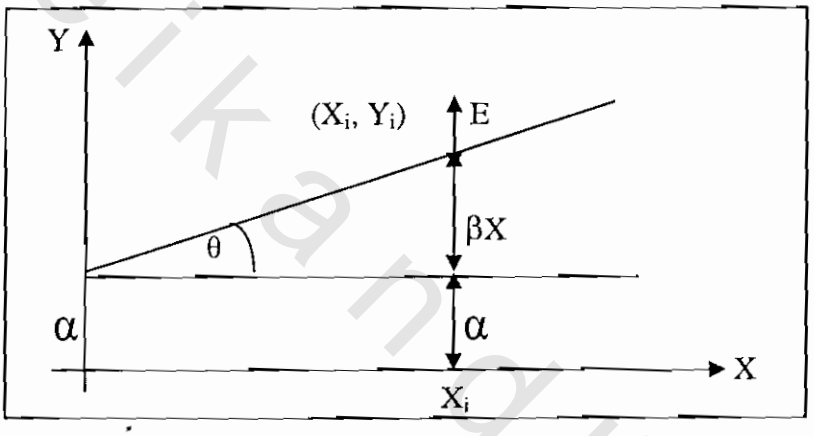


شكل ١٠-١ تمثيل العلاقة $Y = \alpha + \beta X$ بيانياً

ومن الشكل يلاحظ أن لكل قيمة من المتغير المستقل X قيمة تقابلها للمتغير التابع Y وتقع كل قيم Y على خط مستقيم واحد.

ولدراسة العلاقة بين متغيرين وليكن أحدهما وزن الحيوان بالكيلوجرام والآخر العمر بالأسبوع، فإن أول ما يجب عمله هو تمثيل تلك العلاقة بيانياً على أن يمثل المحور السينى المتغير المستقل، وهو العمر فى مثل هذه العلاقة، ويمثل المحور الصادى الوزن فى شكل يعرف بشكل الانتشار scatter diagram ومنه يتضح إذا كانت العلاقة خطية أو غير ذلك. وإذا كانت العلاقة خطية، أى أن النقط فى شكل الانتشار تتجمع حول خط مستقيم، حيث غالباً أن العلاقة لا تتحقق تماماً بمعنى أن بعض النقط تكون فوق الخط والبعض تحته، وقد تكون هناك عدة نقط على الخط. والمتغير التابع Y يمكن التعبير عنه بالعلاقة (٢-١٠) والشكل ٢-١٠ التاليين:

$$Y = \alpha + \beta X + \in \qquad (٢-١٠)$$



شكل ٢-١٠ تمثيل العلاقة $Y = \alpha + \beta X + \in$ بيانياً

وتمثل $\in$ الخطأ الخاص المصاحب للمتغير Y ويطلق عليه الخطأ العشوائى أو التجريبى وقد يرجع إلى:

١ - خطأ فى قياس المتغير Y errors of measurement

٢ - قد يكون هناك متغيرات أخرى غير X تؤثر على Y ولكن أهملت باعتبار أن X هو المتغير الأساسى محل الدراسة، ويعبر عنها بالمتغيرات المحذوفة omitted variables والتى تدخل ضمن مكونات الخطأ.

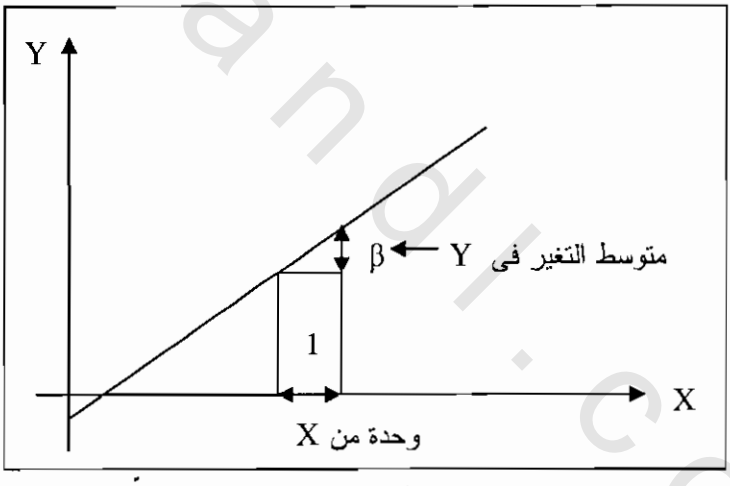
والعلاقة (٢-١٠) هنا تمثل علاقة حقيقية true relationship والهدف هو تقدير معالم هذه العلاقة وهما α و β. و هذه العلاقة مثل للبيانات البيولوجية biological data لوجود عنصر الخطأ والذى لا يمكن التحكم فيه.

١٠-٢ معادلة خط الانحدار

يعرف معامل انحدار المتغير التابع Y على المتغير المستقل X بأنه متوسط التغير فى المتغير التابع Y عندما يتغير المتغير المستقل X بمقدار الوحدة وذلك فى مدى معين للمتغير X، ويرمز له بالرمز β_{YX}.

فمثلاً عند دراسة العلاقة بين الوزن بالكيلوجرام والعمر باليوم فى عجول التسمين كان معامل اعتماد (انحدار) الوزن على العمر هو 0.8 كج/يوم، أى $\hat{\beta}_{YX} = 0.8$ كج/يوم. حيث $\hat{\beta}_{YX}$ هو تقدير غير متحيز لمعامل الانحدار فى العشيرة. وهذا يعنى أنه بزيادة العمر بمقدار يوم فإن الوزن يزيد بمقدار 0.8 كج.

أما إذا كان معامل الانحدار $\hat{\beta}_{YX} = -3$ وحدة من Y / وحدة من X، فإن هذا يعنى أنه بزيادة X بمقدار الوحدة فإن Y تنقص بمقدار 3 وحدات. ومعامل الانحدار لابد وأن يكون مميزاً وقيمته تتراوح بين $-\infty$ إلى $+\infty$ ويمثل شكل ١٠-٣ تعريف معامل الانحدار هندسياً.



شكل ١٠-٣ تمثيل معامل الانحدار هندسياً

وترجع أهمية دراسة الانحدار إلى:

١ - معرفة ما إذا كانت Y تعتمد على X وللحصول على مقياس لتلك العلاقة.

٢ - التنبؤ بقيم Y عند معرفة قيم X، وتفسير Y بواسطة X (وذلك فى مدى معين).

٣ - تحديد شكل منحنى الانحدار regression curve

٤ – معرفة الأخطاء الحقيقية الموجودة فى التجربة بعد التعديل (التصحيح) لأثر المتغير المستقل.

٥ – اختبار الفروض التى قد يضعها الباحثون حول العلاقة بين المسبب cause وتأثيره effect.

قد لا يفضل لسبب أو لآخر دراسة كل أفراد العشيرة وعلى ذلك يكتفى بدراسة العلاقة بين المتغيرين فى عينة أو جزء (عينة) من هذه العشيرة وذلك بغرض الوصول إلى تقدير لكل من α، β والتى يعبر عنها بنموذج التقدير estimated model التالى:

$$Y = a + bX + e \qquad (٣-١٠)$$

حيث b, a هما تقديران غير متحيزين لكل من α، β السابق الإشارة إليهما فى (١٠-٢)، e تقدير لعنصر الخطأ غير المعروف. ولكى تتم عملية التقدير يلزم مجموعة من المشاهدات للمتغير X وما يقابلها من المتغير Y.

وفى حالة الانحدار الخطى البسيط simple linear regression يمكن رسم الخط بنقطتين وفى هذه الحالة لا يمكن تقدير خطأ، أما إذا كان هناك مجموعة من النقط فيلزم توفيق خط يسمى خط الانحدار regression line يمثل هذه العلاقة ومعادلته هى:

$$\hat{Y} = a + bX \qquad (٤-١٠)$$

وتقدير كل من b ,a يكون بإحدى الوسيلتين التاليتين أو غيرهما:

١- حساب الفروق (الأخطاء) بغض النظر عن الإشارة مع محاولة جعل مجموعها أقل ما يمكن. أو

٢- حساب مجموع مربع الأخطاء $\sum e^2$ من النموذج (١٠-٤)، حيث $e_i = Y_i - a - bX_i$ وجعله أقل ما يمكن. وهذه الطريقة هى الأكثر شيوعاً واستخداماً وتعرف بطريقة المربعات الصغرى least squares method.

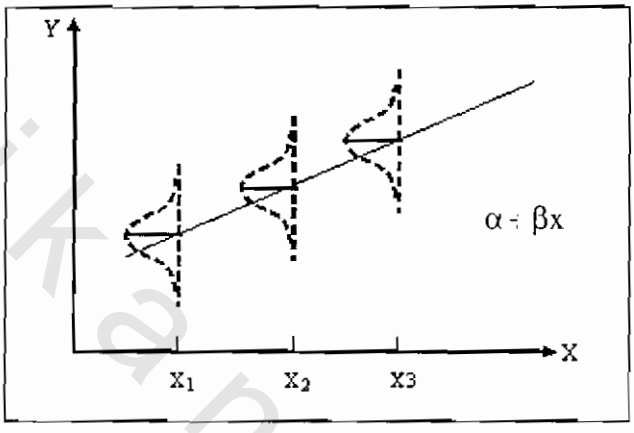
ولكى يتم ذلك توضع افتراضات لازمة لعملية التقدير وهى:

١- المتغير X ثابت fixed فى المعاينات المتكررة repeated sampling أى ليس له توزيع احتمالى.

٢- لا يوجد ارتباط بين كل من المتغير المستقل X ومكون الخطأ e.

٣- لكل قيمة للمتغير X_i يوجد توزيع للمتغير Y متوسطه يقع على خط الانحدار $\hat{\mu}_{y.x_i} = a + \hat{b}X_i$، وتختلف المتوسطات للتوزيعات ولكن لها نفس التباين $\sigma^2_{Y.X}$ وشكل ١٠-٤ يبين ذلك.

٤- الأخطاء e_i مستقلة وتتوزع طبيعياً بمتوسط يساوى الصفر وتباين $\sigma^2_{Y.X}$ وهذا الفرض مهم عند اختبار معنوية الانحدار.



شكل ١٠-٤ التوزيع الطبيعى لقيم Y حول خط الانحدار $\alpha + \beta x$ لبعض قيم X

ولتطبيق طريقة المربعات الصغرى، أى لجعل مجموع مربعات الأخطاء $\sum e^2$ أقل ما يمكن، يستخدم كل من التفاضل الجزئى وتفاضل دالة الدالة على النحو التالى:

بما أن: $Y_i = a + bX_i + e_i$

إذاً $e_i = Y_i - a - bX_i$

وبتربيع الطرفين: $e_i^2 = (Y - a - bX_i)^2$

وبالجمع لكل قيم e_i من 1 إلى n حيث n عدد أزواج المشاهدات:

$$\sum e_i^2 = \sum (Y - a - bX_i)^2$$

بالتفاضل الجزئى لكل من a ، b ومساواة الناتج بالصفر:

$$\frac{\partial \sum e^2}{\partial a} = 2\sum (Y_i - a - bX_i)(-1) = 0$$

$$\therefore \sum Y_i = na + b\sum X_i \qquad (٥-١٠)$$

$$\frac{\partial \sum e^2}{\partial b} = 2\sum (Y_i - a - bX_i)(-X_i) = 0$$

$$\therefore \sum X_i Y_i = a\sum X_i + b\sum X_i^2 \qquad (٦-١٠)$$

وتسمى المعادلتان (١٠-٥)، (١٠-٦) بالمعادلتين الاعتياديتين normal equations وبحل هاتين المعادلتين آنياً يمكن الحصول على:

$$a = \overline{Y} - b\overline{X} \qquad (٧-١٠)$$

$$b_{yx} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - (\sum X)^2/n} \qquad (٨-١٠)$$

ويمكن الوصول إلى صورة أخرى لمعامل الانحدار b إذا تم التعبير عن قيم المتغيرين كانحراف عن متوسطهما، أى استخدام y بدلاً من Y حيث $y = Y - \overline{Y}$ ، x بدلاً من X حيث $x = X - \overline{X}$. وبالتالى يمكن استنتاج أن:

$$b_{yx} = \frac{\sum xy}{\sum x^2} \qquad (٩-١٠)$$

وإذا قسم كل من بسط ومقام العلاقة (١٠-٩) على درجات الحرية n-1 فإن:

$$b_{yx} = \frac{\sum xy/(n-1)}{\sum x^2/(n-1)} = \frac{Cov(X,Y)}{V(X)} \qquad (١٠-١٠)$$

ويعرف $Cov(X,Y)$ بالتغاير covariance وهو عبارة عن الجزء من التباين المشترك joint variance بين قيم المتغيرين X، Y وقيمته قد تكون موجبة أو سالبة أو صفر.

ومعامل الانحدار b أو b_{yx} (معامل انحدار Y على X) يأخذ أى قيمة ويستمد إشارته من إشارة التغاير.

وتكون معادلة الخط المستقيم (معادلة الانحدار) عبارة عن:

$$\hat{Y} = a + bX \qquad (١٠-١١)$$

وبالتعويض عن a بقيمتها $a = \overline{Y} - b\overline{X}$

$$\hat{Y} = \overline{Y} - b\overline{X} + bX = \overline{Y} + b(X - \overline{X}) = \overline{Y} + bx \qquad (١٠-١٢)$$

أى أن:

$$\hat{y} = \hat{Y} - \overline{Y} = bx \qquad (١٠-١٣)$$

ومن خصائص خط الانحدار:

١ – مجموع الانحرافات عن خط الانحدار يساوى صفر.

٢ – مجموع مربع الانحرافات عنه أقل ما يمكن.

٣ – نقطة تقاطع المتوسطين $(\overline{X}, \overline{Y})$ تقع على هذا الخط.

مثال ١٠-١

فى دراسة للعلاقة بين الطول X بالسنتيمتر والوزن Y بالكيلوجرام فى عشيرة ما، كانت البيانات كالتالى:

الزوج	١	٢	٣	٤	٥	٦	٧	٨	٩	١٠
الطول	150	175	170	160	150	155	152	180	154	165
الوزن	50	77	65	63	61	62	55	75	55	68

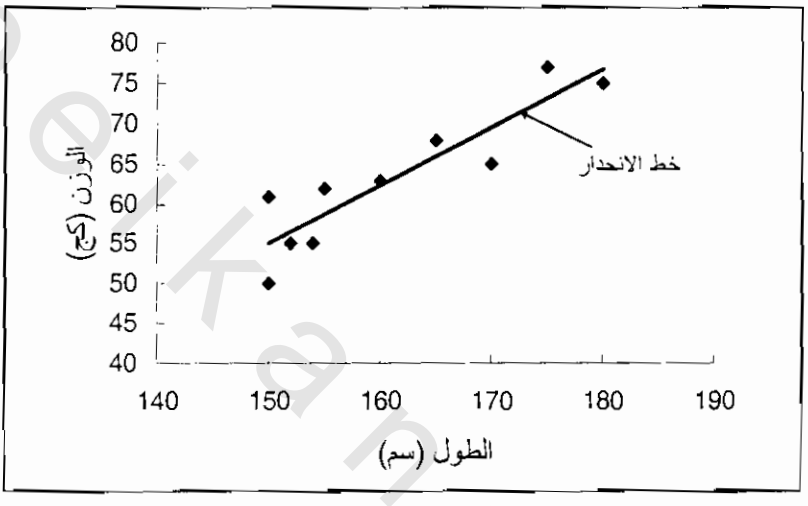
المطلوب:

١ – تمثيل تلك العلاقة بيانياً.

٢ – حساب معامل انحدار الوزن على الطول

٣ – كتابة معادلة خط الانحدار ورسم خط الانحدار رسماً دقيقاً.

التمثيل البيانى يمثله شكل ١٠-٥، ويلاحظ من دراسة شكل الانتشار أن العلاقة بين الطول والوزن علاقة أقرب ما تكون إلى الخطية (فى المدى المدروس). وعلى ذلك يمكن تقدير معادلة الخط المستقيم (معادلة الانحدار) $\hat{Y} = a + bX$ الذى يمثل هذه العلاقة بحيث يكون مجموع الانحرافات عن هذا الخط مساوية للصفر ومجموع مربع الانحرافات عن نفس الخط أقل ما يمكن، أى أن التقدير يتم بطريقة المربعات الصغرى.



شكل ١٠-٥ انحدار الوزن على الطول

وحيث إن الوزن قد يعتمد على الطول فإن الوزن Y هو المتغير التابع والطول X هو المتغير المستقل. ومن المعطيات:

$$n = 10$$

$$\overline{X} = 161.1 \qquad \sum X = 1611 \qquad \sum X^2 = 260595$$

$$\overline{Y} = 63.1 \qquad \sum Y = 631 \qquad \sum XY = 102415$$

وبتطبيق المعادلة (١٠-٨) فإن:

$$b_{yx} = \frac{102415 - \frac{(1611)(631)}{10}}{260595 - \frac{(1611)^2}{10}} = \frac{760.9}{1062.9} = 0.716 \text{ kg/cm}$$

وحيث إن: $a = \overline{Y} - b\overline{X}$ فإن $a = 63.1 - (161.1)(0.716) = -52.228 \text{ kg}$

ولرسم خط الانحدار (شكل ١٠-٥) تؤخذ نقطة تقاطع متوسطى المتغيرين والجزء المقطوع من المحور الصادى a ويتم التوصيل بين النقطتين للحصول على خط الانحدار الذى معادلته ستكون كما يلى:

$$\hat{Y} = -52.228 + 0.716X$$

وتسمى معادلة الخط المستقيم هذه معادلة الانحدار ويطلق عليها أيضاً معادلة التنبؤ prediction equation والتى منها يمكن حساب أى قيمة متوقعة لـــ Y إذا علمت قيمة X، فمثلاً القيمة المتوقعة للوزن إذا كان الطول 185 سم هى:

$$\hat{Y} = -52.228 + (0.716)(185) = 80.232 \text{ kg}$$

والقيمة المتوقعة للوزن إذا كان الطول 172 سم هى:

$$\hat{Y} = -52.228 + (0.716)(172) = 70.924 \text{ kg}$$

ويمكن أن تقدر القيمة المتوقعة من (١٠-١٢) كما يلى:

$$\hat{Y} = 63.1 + 0.716(x) = 63.1 + (0.716)(172 - 161.1)$$
$$= 63.1 + (0.716)(10.9) = 70.904$$

والفرق الناتج بين 70.924 ، 70.904 راجع لخطأ التقريب rounding error.

مثال ١٠-٢

يمكن حل مثال ١٠-١ باستخدام PROC REG فى برنامج SAS، وتستخدم هذه الطريقة عند الرغبة فى الحصول على تقدير كل من a (الجزء المقطوع من المحور الصادى intercept) و b (معامل الانحدار slope) بالإضافة إلى تحليل التباين والذى سوف يتم تناوله فى الأجزاء التالية.

```
DATA WEIGHT;
INPUT HEIGHT WEIGHT @@;
CARDS;
150 50 175 77 170 65 160 63 150 61
155 62 152 55 180 75 154 55 165 68
PROC REG;
MODEL WEIGHT = HEIGHT;
RUN;
```

لاحظ أنه يمكن إضافة بعض الخيارات إلى الـ model منها على سبيل المثال:

Model weight = height / XPX I;

وهذا يؤدى إلى الحصول على مصفوفة $X'X$ ومعكوسها، أى $(X'X)^{-1}$

نتيجة التحليل:

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	544.70676	544.70676	34.53	0.0004
Error	8	126.19324	15.77416		
Corrected Total	9	670.90000			

Root MSE	3.97167	R-Square	0.8119
Dependent Mean	63.10000	Adj R-Sq	0.7884
Coeff Var	6.29425		

الجزء المقطوع من المحور الصادى (a)

معامل الانحدار b

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-52.22693	19.66572	-2.66	0.0290
height	1	0.71587	0.12182	5.88	0.0004

لاحظ وجود بعض النتائج الأخرى والتى سوف يتم تناولها فى الأجزاء التالية.

١٠-٣ استخدام المصفوفات فى الانحدار

استخدام المصفوفات له كثير من الميزات من أهمها أن التعامل الرياضى فيها يكون أكثر اختصاراً وأوضح رؤية، أى أنه بمجرد كتابة المشكلة وحلها عن طريق المصفوفات فإنه يمكن تطبيق الحل على أى مشكلة بغض النظر عن عدد الحدود فى معادلة الانحدار. والخطوات التالية تبين كيفية حل مثال ١٠-١ باستخدام المصفوفات.

سبق عرض المعادلة (١٠-٣) وهى $Y = a + bX + e$ ، هذه المعادلة يمكن إعادة كتابتها باستخدام المصفوفات كالتالى:

$$Y = X\beta + \varepsilon \qquad (١٤-١٠)$$

وباستخدام بيانات المثال ١٠-١ يمكن تكوين المصفوفات التالية:

$$Y = \begin{bmatrix} 50 \\ 77 \\ \vdots \\ 55 \\ 68 \end{bmatrix}, X = \begin{bmatrix} 1 & 150 \\ 1 & 175 \\ \vdots & \vdots \\ 1 & 154 \\ 1 & 165 \end{bmatrix}, \beta = \begin{bmatrix} a \\ b \end{bmatrix}, \text{and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix}$$

حيث:

Y : متجهة vector حجمه 10 x 1 يحتوى على قيم المتغير التابع

X : مصفوفة حجمها 2 x 10 تحتوى على العوامل المستقلة

β : متجهة حجمه 2 x 1 يحتوى على المعالم المراد تقديرها وهى a، b

ε : متجهة حجمه 10 x 1 يمثل الخطأ

وبتطبيق المعادلة $Y = X\beta + \varepsilon$ (١٤-١٠) يمكن الحصول على:

$$Y = \begin{bmatrix} 50 \\ 77 \\ \vdots \\ 55 \\ 68 \end{bmatrix} = \begin{bmatrix} a + 150\,b \\ a + 175\,b \\ \vdots \\ a + 154\,b \\ a + 165\,b \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix}$$

$$(10 \times 1) \quad (10 \times 1) \quad (10 \times 1)$$

المعادلتان الاعتياديتان normal equations (١٠-٥)، (١٠-٦) يمكن التعبير عنهما بالمصفوفات كالتالى:

$$X'X\hat{\beta} = X'Y \qquad (١٠-١٥)$$

وبالتعويض:

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 150 & 175 & \cdots & 154 & 165 \end{bmatrix} \begin{bmatrix} 1 & 150 \\ 1 & 175 \\ \vdots & \vdots \\ 1 & 154 \\ 1 & 165 \end{bmatrix} = \begin{bmatrix} 10 & 1611 \\ 1611 & 260595 \end{bmatrix}$$

$$(2 \times 10) \quad (10 \times 2) \quad (2 \times 2)$$

والصورة العامة لهذا الجزء:

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$(2 \times n) \qquad (n \times 2) \qquad (2 \times 2)$$

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 150 & 175 & \cdots & 154 & 165 \end{bmatrix} \begin{bmatrix} 50 \\ 77 \\ \vdots \\ 55 \\ 68 \end{bmatrix} = \begin{bmatrix} 631 \\ 102415 \end{bmatrix}$$

$$(2 \times 10) \qquad (10 \times 1) \qquad (2 \times 1)$$

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$(2 \times n) \qquad (n \times 1) \qquad (2 \times 1)$$

وبالتالى تكون الصورة العامة للمعادلة (١٠-١٥) كالتالى:

$$\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \qquad (١٦-١٠)$$

وبالتعويض

$$\begin{bmatrix} 10 & 1611 \\ 1611 & 2600595 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 631 \\ 102415 \end{bmatrix}$$

حل المعادلة (١٠-١٦) سوف يعطى تقديراً لكل من a، b بطريقة المربعات لصغرى. ولعمل ذلك لابد من الحصول على معكوس inverse المصفوفة التى فى لجانب الأيسر لهذه المعادلة أى إيجاد $(X'X)^{-1}$ وبالتالى يكون:

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (١٠-١٧)$$

بمعنى أن:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \qquad (١٠-١٨)$$

ويمكن الحصول على هذا المعكوس كالتالى:

$$(X'X)^{-1} = \begin{bmatrix} \dfrac{\sum X_i^2}{n\sum (X_i - \overline{X})^2} & \dfrac{-\overline{X}}{\sum (X_i - \overline{X})^2} \\ \dfrac{-\overline{X}}{\sum (X_i - \overline{X})^2} & \dfrac{1}{\sum (X_i - \overline{X})^2} \end{bmatrix} \qquad (١٠-١٩)$$

وبأخذ عامل مشترك للجانب الأيمن تصبح المعادلة (١٠-١٩) كالتالى:

$$(X'X)^{-1} = \frac{1}{n\sum (X_i - \overline{X})^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} \qquad (١٠-٢٠)$$

وبالتعويض

$$(X'X)^{-1} = \begin{bmatrix} 24.517358 & -0.151566 \\ -0.151566 & 0.0009408 \end{bmatrix}$$

وبتطبيق المعادلة (١٠-١٨) يكون تقدير كل من a، b كالتالى:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 24.517358 & -0.151566 \\ -0.151566 & 0.0009408 \end{bmatrix} \begin{bmatrix} 631 \\ 102415 \end{bmatrix} = \begin{bmatrix} -52.22693 \\ 0.7158717 \end{bmatrix}$$

وهى نفس النتائج المتحصل عليها سابقاً مع بعض الاختلافات نتيجة للتقريب. ومن ذلك تتضح إمكانية وسهولة تطبيق طريق المصفوفات على أى أمثلة أخرى باستخدام نفس المعادلتين (١٠-١٨) و (١٠-٢٠).

مثال ١٠-٣

حل مثال ١٠-١ باستخدام طريقة **IML** فى برنامج SAS، وتستخدم هذه الطريقة عند الرغبة فى استخدام المصفوفات فى الحل

```
PROC IML;
X={1 150, 1 175, 1 170, 1 160, 1 150,
    1 155, 1 152, 1 180, 1 154, 1 165};
 XP = X`;
XPX = XP*X;
   Y= {50, 77, 65, 63, 61, 62, 55, 75, 55, 68};
YP = Y`;
XPY = XP*Y;
XPXINV=INV(XPX);
SOL = XPXINV*XPY;
PRINT XP YP XPX XPY XPXINV SOL;
QUIT;
```

لاحظ:

طريقة كتابة المصفوفات حيث تكتب جميع عناصر المصفوفة داخل قوسين من النوع{ }.

تنتهى عناصر كل صف بفصله عادية.

استخدام علامة ` للحصول على المقلوب transpose (X`) واستخدام INV للحصول على المعكوس inverse.

لابد أن ينتهى البرنامج بكلمة quit.

نتائج التحليل:

XP

	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10
ROW1	1	1	1	1	1	1	1	1	1	1
ROW2	150	175	170	160	150	155	152	180	154	165

YP

	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10
ROW1	50	77	65	63	61	62	55	75	55	68

XPX

	COL1	COL2
ROW1	10	1611
ROW2	1611	260595

XPY

	COL1
ROW1	631
ROW2	102415

XPXINV

	COL1	COL2
ROW1	24.517358	-0.151566
ROW2	-0.151566	0.0009408

SOL

	COL1
ROW1	-52.517358
ROW2	0.7158717

صندوق ١٠-١

- عندما يكون هناك متغيران أحدهما وليكن Y يعتمد على متغير آخر وليكن X فيعرف الأول بالتابع والثانى بالمستقل.
- ولكل من المتغيرين تباين ولكن بينهما أيضاً تباين مشترك يطلق عليه التغاير covariance.
- يمكن تقنين العلاقة بين هذين المتغيرين باستخدام معادلة الخط المستقيم باعتبار X على المحور السينى و Y على المحور الصادى وحيث إن العلاقة بين المتغيرين عادة ما تكون غير تامة أى أن Y تحتوى على جزء غير راجع إلى X وهو ما يطلق عليه الخطأ وبالتالى فإن معادلة الخط المستقيم يضاف إليها مكون يعبر عن هذا الخطأ وتكون معادلة الانحدار هى:

$$Y = a + bX + e$$

حيث a ، b ، ثابتان، e الخطأ فى المشاهدة. كما هو الحال فى معادلة الخط المستقيم فإن a تمثل الجزء المقطوع من المحور الصادى بينما b هى ظل زاوية تقاطع الخط المستقيم مع المحور السينى.

- يقدر معامل انحدار Y على X من

$$b_{YX} = \frac{\sum XY - (\sum X \sum Y)/n}{\sum X^2 - (\sum X)^2/n}$$

١٠-٤ مصادر الاختلاف فى الانحدار الخطى

بالرجوع إلى البيانات التى فى مثال ١٠-١ وجد أنه عندما كانت قيمة X للفرد الأول 150 سم فإن قيمة Y له 50 كج، بينما قيمة X للفرد الثانى 175 سم وقيمة Y له 77 كج. ومعنى ذلك أن الوزن أثقل للطول الأكبر، كما يلاحظ أنه بينما كانت قيمة X للفرد الخامس هى أيضاً 150 سم كان الوزن المقابل لها 61 كج. أى أن هناك أيضاً فرقاً فى الوزن حتى لو تساوت الأطوال. ويطلق على هذا الفرق غير المعروف مسبباته خطأ error أو من خط الانحدار from regression. وبالتالى فإن مصادر الاختلاف فى الانحدار الخطى جزء منه راجع إلى اعتماد الوزن على الطول والجزء الآخر خاص بكل فرد. ويمكن بيان ذلك رياضياً كما يلى:

فى الانحدار البسيط الذى يمثل العلاقة بين المتغيرين Y,X، يمكن التعبير عن أى مشاهدة بالنموذج المبين فى (١٠-٢) أى:

$$Y = \alpha + \beta X + \in$$

حيث تمثل $\alpha + \beta X$ متوسط المتغير Y والتى تقابل قيمة محددة للمتغير X ويرمز له μ_{yx} وتمثل $\in$ مكون الخطأ error component، أى أن قيمة Y تمثل مجموع متغيرين أحدهما المتوسط والآخر الخطأ العشوائى. وفى النموذج التقديرى estimated model فإن:

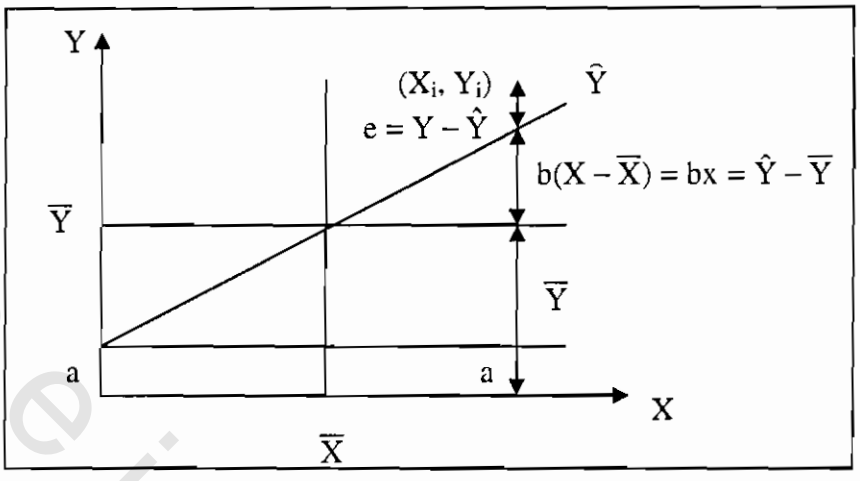
$$Y = a + bX + e$$
$$Y = \overline{Y} + b(X - \overline{X}) + e$$
$$Y = \overline{Y} + (\hat{Y} - \overline{Y}) + (Y - \hat{Y})$$

ى أن:

$$(Y - \overline{Y}) = (\hat{Y} - \overline{Y}) + (Y - \hat{Y}) \qquad (١٠-٢١)$$

يمكن توضيح ذلك بالشكل ١٠-٦.

ويتضح من الشكل ١٠-٦ إن انحراف القيمة عن متوسطها عبارة عن مجموع جدين. الأول يمثل انحراف القيمة المتوقعة عن المتوسط، والتى تمثل الجزء الراجع إلى اعتماد Y على X ويسمى due to regression أى $(\hat{Y} - \overline{Y}) = b(X - \overline{X}) = bx$. أما الحد الثانى فهو يمثل انحراف القيمة Y عن القيمة المتوقعة $\hat{Y}$ والواقعة على خط الانحدار، وهذا الحد يمثل الخطأ العشوائى ويسمى from regression، أى $e = d_{Y.X} = Y - \hat{Y} = Y - a - bX = y - bx$.



شكل ١٠-٦ مصادر الاختلاف فى Y

ويلاحظ أن مجموع انحرافات قيم $\hat{Y}$ عن المتوسط يساوى صفر، أى $\sum(\hat{Y} - \overline{Y}) = 0$، ومجموع انحرافات Y عن خط الاعتماد أيضاً يساوى صفر أى $\sum(Y - \hat{Y}) = 0$.

وبالتالى من المعادلة (١٠-٢١) يمكن إثبات أن:

$$\sum(Y - \overline{Y})^2 = \sum(\hat{Y} - \overline{Y})^2 + \sum(Y - \hat{Y})^2 \qquad (٢١-١٠)$$

أى أن:

$$TSS = RSS + ESS$$

حيث:

TSS عبارة عن مجموع المربعات الكلى total sum of squares،

RSS عبارة عن مجموع المربعات الراجع لانحدار Y على X،
sum of squares due to regression

ESS عبارة عن مجموع المربعات عن خط الانحدار،
sum of squares from regression

هذه القيم يمكن الحصول عليها كالتالى:

$$RSS = \sum(\hat{Y} - \bar{Y})^2 = b^2\sum x^2 = b\sum xy = \frac{(\sum xy)^2}{\sum x^2}$$

$$ESS = \sum y^2 - RSS = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$$

والجدول ١٠-١ يبين مصادر الاختلاف فى Y لبيانات مثال ١٠-١

جدول ١٠-١ مصادر الاختلاف فى قيم Y وتقسيم التباين إلى مصادره المختلفة

X	Y	$\hat{Y}$	$Y-\bar{Y}$	$\hat{Y}-\bar{Y}$	$e = Y-\hat{Y}$	$(\hat{Y}-\bar{Y})^2$	$(Y-\bar{Y})^2$
150	50	55.2	-13.1	-7.9	-5.2	62.41	27.04
175	77	73.1	13.9	10.0	3.9	100.00	15.21
170	65	69.4	1.9	6.3	-4.4	39.69	19.36
160	63	62.3	-0.1	-0.8	0.7	0.64	0.49
150	61	55.3	-2.1	-7.9	5.8	62.41	33.64
155	62	58.7	-1.1	-4.4	3.3	19.36	10.89
152	55	56.6	-8.1	-6.5	-1.6	42.45	2.56
180	75	76.6	11.9	13.5	-1.6	182.25	2.56
154	55	58.0	-8.1	-5.1	-3.0	26.01	9.00
165	68	65.9	4.9	2.8	2.1	7.84	4.41
1611	**631**	**631**	**0.0**	**0.0**	**0.0**	**542.86**	**125.16**

مثال ١٠-٤

قسم مجموع المربعات الكلى فى Y إلى مكوناته مستخدماً بيانات المثال ١٠-١.

مجموع المربعات الكلى:

$$TSS = \sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} = 40487 - \frac{(631)^2}{10} = 670.9$$

مجموع المربعات الراجع إلى الانحدار:

$$RSS = \sum(\hat{Y} - \overline{Y})^2 = b^2 \sum x^2 = b \sum xy$$

$$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n} = 260595 - \frac{(1611)^2}{10} = 1062.9$$

$$RSS = (0.716)^2 (1062.9) = 544.9$$

مجموع مربعات الخطأ (من خط الانحدار):

$$ESS = \sum(Y - \hat{Y})^2 = TSS - RSS = 670.9 - 544.7 = 126.2$$

وقيم كل من RSS و ESS هى نفس القيم المتحصل عليها فى جدول ١٠-١ والاختلافات بينها راجع إلى التقريب.

يمكن الحصول على نفس النتائج باستخدام المصفوفات كالتالى:

$$TSS = Y'Y - n\overline{Y}^2 \qquad (٢٢-١٠)$$

$$RSS = b'X'Y - n\overline{Y}^2 \qquad (٢٣-١٠)$$

$$ESS = TSS - RSS = Y'Y - b'X'Y \qquad (٢٤-١٠)$$

وباستخدام بيانات مثال ١٠-١
مجموع المربعات الكلى:

$$TSS = Y'Y - n\overline{Y}^2 = \begin{bmatrix} 50 & 77 & \cdots & 55 & 68 \end{bmatrix} \begin{bmatrix} 50 \\ 77 \\ \vdots \\ 55 \\ 68 \end{bmatrix} - \left[(10)(63.1)^2\right] = 670.9$$

و هذا المجموع له 9 درجات حرية أى (n-1).

مجموع المربعات الراجع إلى الانحدار:

$$RSS = \begin{bmatrix} -52.22693 & 0.71587 \end{bmatrix} \begin{bmatrix} 631 \\ 102415 \end{bmatrix} - \left[(10)(63.1)^2\right] = 544.534$$

وهذا المجموع له درجة حرية واحدة.

مجموع مربعات الخطأ (من خط الانحدار):

$$ESS = TSS - RSS = 670.9 - 544.534 = 126.357$$

و هذا المجموع له 8 درجات حرية أى (n-2).

ويعتبر متوسط مجموع مربعات الخطأ، $(ESS)/(n-2)$، تقدير غير متحيز للتباين σ^2 ويرمز له بالرمز $S^2_{y.x}$.

لاحظ أن نفس النتائج تم الحصول عليها عند استخدام اختيار PROC REG فى برنامج SAS فى فصل ١٠-٤ مع ملاحظة وجود خطأ التقريب.

١٠-٥ القيم المعدلة لأثر انحدار Y على X (Adjusted Y)

كما ذكر سابقاً $Y = \overline{Y} + bx + e$

وعلى ذلك فقيمة Y المعدلة، ويرمز لها Y_A، أى بعد إزالة الجزء الراجع للانحدار bx من قيمة Y هى:

$$\text{Adjusted } Y = Y_A = \overline{Y} + e = Y - bx \qquad (١٠-٢٥)$$

وتفيد القيم المعدلة فى المقارنة بين الأفراد أو بين المتوسطات المعدلة، أى بعد إزالة أثر المتغير المستقل. فمثلاً إذا كان معامل انحدار الوزن على العمر هو 0.8 كج لكل يوم وكان هناك حيوانان الأول عمره 140 يوم ووزنه 200 كج والثانى عمره 168 يوم ووزنه 220 كج، فأيهما أثقل وزناً؟ وللإجابة على ذلك يجب أولا إزالة أثر الاختلاف فى العمر للحيوانين. فإذا كان متوسط العمر هو 105 يوماً ومتوسط الوزن 190 كج فإن:

$$Y_{A_1} = Y_1 - bx_1 = 200 - 0.8(140 - 105) = 172 \text{ kg}$$

$$Y_{A_2} = Y_2 - bx_2 = 220 - 0.8(168 - 105) = 169.6 \text{ kg}$$

ومعنى ذلك أنه لو أن الحيوانين كانا عند نفس العمر (105 يوم مثلاً) لكان التوقع أن يزن الحيوانان 172 و 169.6 كج، على التوالى. وعلى ذلك فالحيوان الأول أكثر وزناً من الحيوان الثانى بعد إزالة العمر وبالتالى يكون هو الأثقل وزناً.

وتجدر الإشارة هنا أنه بحساب معامل الانحدار فإنه يمكن استخراج كل العلاقة بين المتغيرين Y,X وما يترك فى Y بعد هذه العلاقة وهو e فإن العلاقة بينه وبين كل من المتغيرات الأخرى تساوى صفراً أى أن: $Cov(e,X) = Cov(e,\hat{Y}) = 0$.

ويمكن التأكد من ذلك بحساب كل من $\sum e\hat{Y}$ و $\sum eX$ من جدول ١٠--١ والتى ستكون مساوية للصفر أيضاً.

١٠-٦ الانحرافات المعيارية وحدود الثقة للتقديرات المختلفة

Standard deviation and confidence limits of estimates

١٠-٦-١ الانحرافات المعيارية

سبق اعتبار أن Y تقدير غير متحيز لـــ $\mu_{y.x}$ وكذلك $\hat{Y},b,a$ تقديرات غير متحيزة لكل من $\mu_{Y.X},\beta,\alpha$ على التوالى.

تباين $\overline{Y}$ هو: $\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$ وذلك من المعادلة (٣-١٣) بالباب الثالث.

تباين b:

$$\sigma_b^2 = V\left(\frac{\sum xy}{\sum x^2}\right)$$

$$= \frac{1}{(\sum x^2)^2}\left[V(x_1y_1 + x_2y_2 + \ldots + x_ny_n)\right]$$

$$= \frac{1}{(\sum x^2)^2}\left[x_1^2V(y_1) + x_2^2V(y_2) + \cdots + x_n^2V(y_n)\right]$$

وحيث إن: $V(y_1) = V(y_2) = \cdots = V(y_n) = \sigma^2$ ، فإن:

$$\sigma_b^2 = \frac{1}{(\sum x^2)^2}\left[x_1^2\sigma^2 + x_2^2\sigma^2 + \cdots + x_n^2\sigma^2\right]$$

$$= \frac{1}{(\sum x^2)^2}\sigma^2\sum x^2 = \frac{\sigma^2}{\sum x^2} \qquad (٢٦-١٠)$$

تباين الجزء المقطوع من محور الصادات α:

يمكن بنفس المفهوم إثبات أن:

$$\sigma_a^2 = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum x^2}\right) \qquad (٢٧-١٠)$$

وتباين متوسط العشيرة الذى يقابل قيمة محددة لـــ X أى $\mu_{y.x}$ هو:

$$\sigma_{\mu_{y.x}}^2 = \sigma^2\left(\frac{1}{n} + \frac{x_i^2}{\Sigma x_i^2}\right) \qquad (٢٨-١٠)$$

وتباين القيمة المتوقعة التى تقابل قيمة محددة لـــ X هى:

$$\sigma_{\hat{Y}}^2 = \sigma^2\left(1 + \frac{1}{n} + \frac{x_i^2}{\Sigma x_i^2}\right) \qquad (٢٩-١٠)$$

حيث القيمة المتوقعة $\hat{Y}$ هى تقدير لنقطة جديدة لـــ Y التى تقابل قيمة محددة لـــ X. وبأخذ الجذر التربيعى للمعادلات السابقة يمكن الحصول على الانحراف المعيارى للتقديرات المختلفة.

فمثلاً الانحراف المعيارى لمعامل الانحدار SE_b أو S_b هو

$$S_b = \frac{S_{y.x}}{\sqrt{\sum x^2}} \qquad (٣٠-١٠)$$

حيث $S_{y.x}^2$ هى تقدير لـــ σ^2 علماً بأن $S_{y.x}^2 = ESS/(n-2)$ كما ذكر من قبل. و ESS هى مجموع مربعات الخطأ و n عدد أزواج المتغيرات (المشاهدات).

وتعبر $S_{y.x}$ عن الخطأ القياسى للتقدير أو الانحراف المعيارى لـ Y باعتبار أن X ثابتة.

ومما سبق يمكن كتابة مصفوفة التباين والتغاير variance-covariance matrix للتقديرات المختلفة كالتالى:

$$V(\hat{\beta}) = \begin{bmatrix} V(a) & Cov(a,b) \\ Cov(a,b) & V(b) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sigma^2 \sum X_i^2}{n\sum (X_i - \overline{X})^2} & -\frac{\overline{X}\sigma^2}{\sum (X_i - \overline{X})^2} \\ -\frac{\overline{X}\sigma^2}{\sum (X_i - \overline{X})^2} & \frac{\sigma^2}{\sum (X_i - \overline{X})^2} \end{bmatrix} \qquad (٣١-١٠)$$

وبأخذ عامل مشترك σ^2 فإن المعادلة (١٠-٣١) تصبح:

$$V(\hat{\beta}) = (X'X)^{-1}\sigma^2 \qquad (٣٢-١٠)$$

وللحصول على تقدير لمتوسط العشيرة (باستخدام جبر المصفوفات) عند قيمة معينة ولتكن X_o ضع $X'_o = [1 \quad X_o]$، وبالتالى يمكن تقدير متوسط العشيرة عند قيمة X_o كما يلى:

$$\hat{\mu}_{y.x} = [1 \quad X_o]\begin{bmatrix} a \\ b \end{bmatrix} = X'_o b = b'X_o$$

وحيث إن تباين $\hat{\mu}_{y.x}$ عبارة عن

$$V(\hat{\mu}_{y.x}) = V(a) + 2X_o cov(a,b) + X_o^2 V(b)$$

فإن

$$V(\hat{\mu}_{y.x}) = [1 \quad X_o]\begin{bmatrix} V(a) & cov(a,b) \\ cov(a,b) & V(b) \end{bmatrix}\begin{bmatrix} 1 \\ X_o \end{bmatrix}$$

$$= X'_o (X'X)^{-1}\sigma^2 X_o$$

$$= X'_o (X'X)^{-1} X_o \sigma^2$$

وهذه مطابقة تماماً للمعادلة (١٠-٢٩).

عندما تكون $X_o = 156$ (فى مثال ١٠-١) فإن

$$\hat{\mu}_{y.x} = [1 \quad 156]\begin{bmatrix} -52.22693 \\ 0.71587 \end{bmatrix} = 59.45$$

$$V(\hat{\mu}_{y.x}) = X'_o(X'X)^{-1}X_o\sigma^2_{y.x}$$

وحيث إن $\sigma^2_{y.x} = 15.77416$ تعتبر تقديراً غير متحيز لـ $\sigma^2_{y.x}$ فإن

$$V(\hat{\mu}_{y.x}) = [1 \quad 156]\begin{bmatrix} 24.517358 & -0.151566 \\ -0.151566 & 0.0009408 \end{bmatrix}\begin{bmatrix} 1 \\ 156 \end{bmatrix}(15.77416)$$
$$= 1.956841$$

وبالتالى

$$S_{\hat{\mu}_{Y.X}} = \sqrt{1.956841} = 1.398871$$

تقدير تباين القيمة المتوقعة $\hat{Y}$ عند قيمة معينة للمتغير X:

عندما يكون الغرض هو التنبؤ prediction بقيمة المتغير التابع Y عند قيمة معينة للمتغير X ولتكن X_o (وليس التقدير estimation كما هو الحال عند تقدير متوسط العشيرة عند نفس النقطة X) فإن القيمة المتوقعة تكون هى نفسها مساوية لمتوسط العشيرة عند تلك النقطة ولكن بتباين أكبر هو:

$$V(\hat{Y}_{X=X_o}) = [1 + X'_o(X'X)^{-1}X_o]\sigma^2_{y.x}$$

وتستخدم $S^2_{Y.X}$ فى حالة عدم معرفة $\sigma^2_{Y.X}$

١٠-٦-٢ حدود الثقة لكل من a، b، $\hat{\mu}_{Y.X}$، $\hat{Y}$

بنفس المفهوم الذى استخدم لإيجاد حدود الثقة لكل من المتوسط والتباين فى الباب السادس فصل ٦-٢ يمكن إيجاد حدود الثقة كما يلى:

حدا الثقة لـــ a هما:

$$(\overline{Y} - b\overline{X}) \pm t_{\alpha} S_{y.x} \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum x^2}} \qquad (٣٣-١٠)$$

حدا الثقة لـــ b هما:

$$b \pm t_{\alpha} \frac{S_{y.x}}{\sqrt{\sum x^2}} \qquad (٣٤-١٠)$$

حدا الثقة لـــ $\hat{\mu}_{Y.X}$: أى حدا الثقة لمتوسط العشيرة الذى يقابل قيمة معينة لـــ X (حيث $\mu_{Y.X} = \overline{Y} + bx$) هما:

$$\hat{Y} \pm t_{\alpha} S_{y.x} \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum x^2}} \qquad (٣٥-١٠)$$

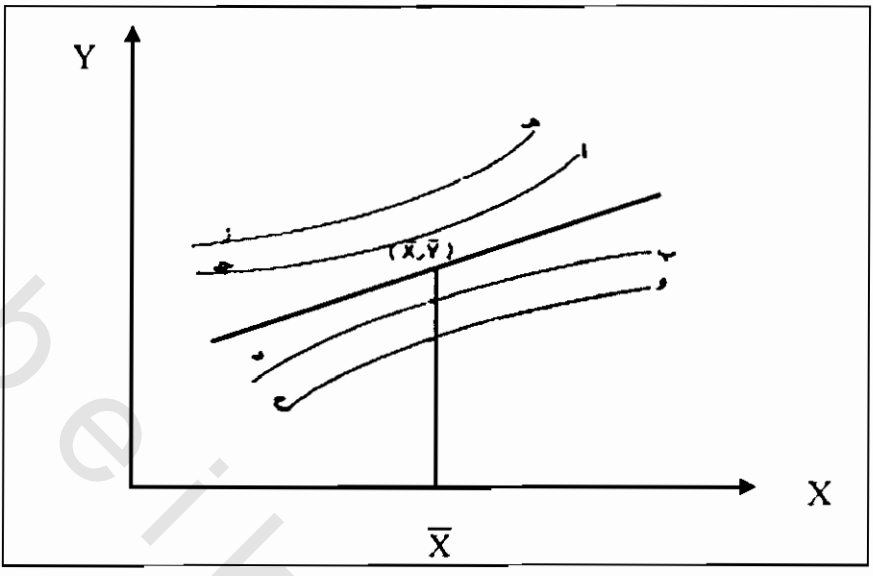
حيث $\hat{Y} = \overline{Y} + bx$.

وهذه تمثل حدى الثقة لخط الانحدار. وهذه القيم تشكل خطين وتسمى المساحة المحصورة بينهما بحزام الثقة confidence belt عند مستوى احتمالى α (مستوى المعنوية).

أما حدا الثقة للقيمة المتوقعة $\hat{Y}$ فهما:

$$(\overline{Y} - bx) \pm t_{\alpha} S_{y.x} \sqrt{1 + \frac{1}{n} + \frac{\overline{X}^2}{\sum x^2}} \qquad (٣٦-١٠)$$

وشكل ١٠-٧ يبين حزام الثقة حول $\hat{\mu}_{Y.X}$ وحزام الثقة حول $\hat{Y}$.



شكل ١٠-٧ حزام الثقة حول $\hat{\mu}_{Y.X}$ وهو أ ب ج د وحزام الثقة حول $\hat{Y}$ وهو هـ و ز ح.

وفى المعادلات السابقة t_α هى قيمة t الجدولية بدرجات حرية $n-2$ ومستوى معنوية α.

مثال ١٠-٥

احسب حدى الثقة لمعامل اعتماد Y على X باستخدام بيانات المثال ١٠-١ علماً بأن مستوى المعنوية 5%.

حدا الثقة هما:

$$b \pm t_\alpha \frac{S_{y.x}}{\sqrt{\sum x^2}}$$

حيث

$$S_{Y.X} = \sqrt{\frac{\sum y^2 - [(\sum xy)^2/(\sum x^2)]}{(n-2)}} \quad ، \quad \sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n} \quad ،$$

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

وبالتعويض فإن:

$$\sum x^2 = 260595 - \frac{(1611)^2}{10} = 1062.9$$

$$\sum y^2 = 40487 - \frac{(631)^2}{10} = 670.9$$

$$\sum xy = 102415 - \frac{(1611)(631)}{10} = 760.9$$

$$S_{Y.X} = \sqrt{\frac{670.9 - [(760.9)^2/(1062.9)]}{(10-2)}} = \sqrt{\frac{670.9 - 544.7}{8}} = 3.97$$

$$S_b = \frac{S_{y.x}}{\sqrt{\sum x^2}} = \frac{3.97}{\sqrt{1062.9}} = 0.122$$

وبالتالى فإن حدى الثقة لمعامل الاعتماد (الانحدار) هما:

$$0.72 \pm (2.306)(0.122) = 0.72 \pm 0.28$$

أى أن الحد الأدنى للثقة هو 0.44 والحد الأعلى للثقة هو 1.00

مثال ١٠-٦

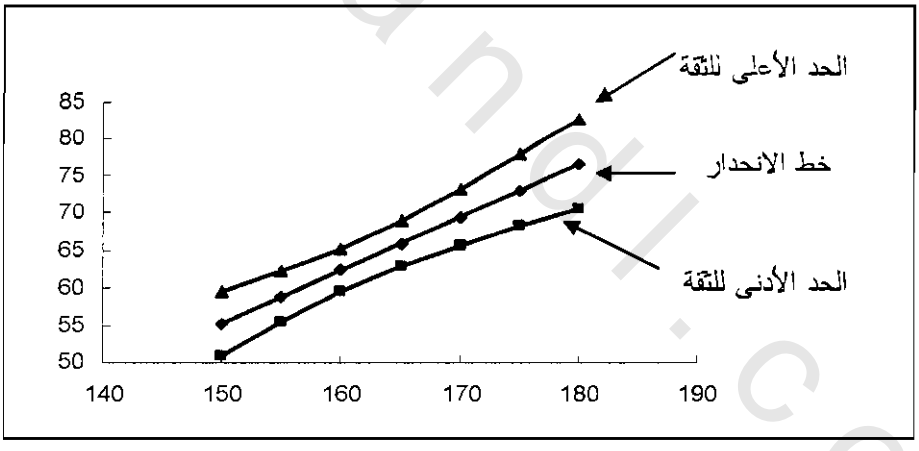
احسب حدود الثقة لخط الانحدار بمستوى معنوية 5% باستخدام بيانات المثال ١٠-١، وماذا تعنى ثم بين ذلك بيانياً.

لحساب حدود الثقة لخط الانحدار يلزم أولاً حساب حدود الثقة لعدد من $\mu_{y.x}$ ثم تمثل هذه النقط بيانياً ويتم التوصيل فيما بينها فينتج خطين أحدهم يمثل الحد الأعلى لخط الانحدار والآخر يمثل الحد الأدنى لخط الانحدار.

حيث إن قيمة t الجدولية بمستوى معنوية 5% ودرجات حرية 8 هى 2.306 وقيمة $S_{y.x} = 3.97$ وباستخدام المعادلة (١٠-٣٥) فإنه يمكن الحصول على النقاط التالية:

حدود الثقة		$\hat{Y} \pm t_{\alpha} S_{y.x} \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum x^2}}$	$\hat{Y}$	X
الحد الأعلى	الحد الأدنى			
59.40	50.90	55.15 ± 4.25	55.15	150
62.09	55.37	58.73 ± 3.36	58.73	155
65.22	59.40	62.31 ± 2.91	62.31	160
68.99	62.79	65.89 ± 3.10	65.89	165
73.29	65.65	69.47 ± 3.82	69.47	170
77.91	68.19	73.05 ± 4.86	73.05	175
82.68	70.58	76.63 ± 6.05	76.63	180

ويمكن تمثيل حدود الثقة لخط الانحدار لهذه البيانات بيانياً بالشكل التالي:



وحزام الثقة هذا يعنى أن احتمال أن يحوى الحزام خط الاعتماد الحقيقى $(\alpha+\beta X)$ هو 0.95.

١٠-٧ اختبار معنوية معامل الانحدار

يمكن اختبار الفرض القائل بأن معامل انحدار العشيرة β يساوى الصفر، أى لا توجد علاقة بين المتغيرينX, Y باستخدام اختبار t كما يلى:

فرض العدم: $H_o: \beta = 0$. وتحت صحة فرض العدم فإن $\beta = 0$ وبالتالى:

$$t = \frac{b - \beta}{S_b} = \frac{b\sqrt{\sum x^2}}{S_{y.x}}$$

وتقارن هذه القيمة بقيمة t الجدولية بدرجات حرية $n-2$ ومستوى معنوية وليكن α.

أما إذا كان فرض العدم هو $Ho: \beta = c$ حيث c هى قيمة معينة لمعامل انحدار العشيرة، فيكون الفرض البديل أن $H_1: \beta \neq c$ وتكون

$$t = \frac{b - c}{S_b}$$

مثال ١٠-٧

اختبر معنوية معامل انحدار الوزن على الطول لبيانات مثال ١٠-١ باستخدام اختبار t.

فرض العدم: $H_o: \beta = 0$ ومن مثال ١٠-٥ $S_b = S_{y.x}/\sqrt{\sum x^2} = 0.12177$ إذاً

$$t = \frac{b - \beta}{S_b} = \frac{0.71587}{0.12177} = 5.88$$

ومن جدول t تظهر القيم $t_{(8,0.05)} = 2.306$ ، $t_{(8,0.01)} = 3.350$

وبالتالى يرفض فرض العدم وبالتالى فإن معامل الانحدار يختلف عن الصفر معنوياً، أى أن العلاقة بين Y، X معنوية بدرجة ثقة 99%.

يمكن اختبار معنوية معامل الانحدار باستخدام تحليل التباين عن طريق تكوين جدول تحليل التباين ANOVA table. فكما سبق وذكر أن مجموع المربعات الكلى فى Y يقسم إلى مكونين أحدهما راجع إلى الانحدار due to regression والآخر عن

خط الانحدار deviations from regression. وبالتالى فإن جدول ١٠-٢ لتحليل لتباين يمكن تكوينه كالتالى:

جدول ١٠- ٢ تحليل التباين باستخدام تحليل الانحدار

SOV مصدر التباين	df	SS مجموع المربعات	MS متوسط المربعات	F
Due to regression راجع للانحدار	1	$RSS = b'X'Y$		$RSS/S^2_{y.x}$
From regression من الانحدار	n-2	$ESS = \sum y^2_{y.x}$ $= Y'Y - b'X'Y$	$S^2_{y.x}$	
C. Total **الكلى المصحح**	**n-1**	$\sum y^2 = Y'Y - n\overline{Y}^2$		

وتقارن قيمة F المحسوبة بقيمة F الجدولية بدرجات حرية $1، n-2$ ومستوى معنوية وليكن α.

وحيث إن t هى $(b\sqrt{\sum x^2})/(S_{y.x})$ فإن قيمة t^2 هى $(b^2\sum x^2)/(S^2_{y.x})$ وهى نفس قيمة F أى أن $F = t^2$ فى حالة الانحدار البسيط.

مثال ١٠-٨

كون جدول تحليل التباين واختبر معنوية معامل الانحدار باستخدام بيانات المثال ١٠-٤.

لاحظ أن جدول تحليل التباين المطلوب قد تم الحصول عليه باستخدام برنامج SAS عند حل مثال ١٠-٢، والجدول كان كالتالى:

SOV	df	SS	MS	F
Due to regression	1	544.7	5.447	34.52**
From regression	8	126.2	15.775	
C. Total	**9**	**670.9**		

قيم F الجدولية $F_{(1,8,0.05)} = 5.32$ ، $F_{(1,8,0.01)} = 11.26$

قيمة F المحسوبة أكبر من قيمة F الجدولية عند مستوى معنوية 5%، 1% .

إذاً العلاقة معنوية بدرجة ثقة 99% بين المتغيرين X و Y.

وحيث إن قيمة t المستخرجة فى المثال ١٠-٧ هى 5.88 ومربعها هو 34.57 ، وهى نفس قيمة F (مع مراعاة الاختلاف البسيط بين مربع قيمة t وقيمة F نتيجة للتقريب) وكلتا الطريقتين متطابقتان تماماً.

صندوق ١٠-٢

- معامل الانحدار هو متوسط التغير فى المتغير التابع Y لكل وحدة تغير فى المتغير المستقل X. وهو قيمة مميزة (أى كج، سم، أردب، وحدة سماد ... الخ) وتتراوح قيمته بين $+\infty$ و $-\infty$.
- بينما تتيح دراسة الانحدار بحث العلاقة بين متغيرين يفترض أن أحدهما مستقل وثابت (أى بدون توزيع احتمالى) والآخر تابع وله توزيع احتمالى (طبيعى مثلا) فإنه أيضا يمكن من:

- تقسيم التباين الكلى Σy^2 فى Y إلى جزء راجع إلى التغير فى X وهو $b\Sigma xy$ وجزء آخر راجع إلى عوامل غير X وهو $\Sigma y^2 - b\Sigma xy$.
- استنباط معادلة الانحدار $\hat{Y} = \alpha + bX$ وبذلك يمكن التنبؤ بأى قيمة للمتغير التابع عند أى قيمة للمتغير المستقل فى حدود مدى الدراسة.
- اختبار المعنوية للعلاقة b_{YX} وكذلك استنباط حدود الثقة عند مستوى احتمالى معين.

١٠-٨ المقارنة بين معاملى الانحدار

يستخدم اختبار t لمقارنة معاملى الانحدار فى عينتين لمعرفة ما إذا كان هذان المعاملان هما تقدير لنفس معامل انحدار العشيرة β أم لا بنفس المفهوم المستخدم فى مقارنة المجاميع group comparison، كما هو موضح فى الباب الثامن ويكون ذلك كما يلى:

$$H_o : \beta_1 - \beta_2 = 0 \quad \text{بمعنى} \quad H_o : \beta_1 = \beta_2$$

وتحت صحة فرض العدم فإن الكمية $\beta_1 - \beta_2 = 0$ تتبع توزيع t بدرجات حرية $[(n_1 - 2) + (n_2 - 2)]$ وبالتالى فإن

$$t = \frac{(b_1 - b_2) - (\beta_1 - \beta_2)}{\sqrt{S^2_{b_1 - b_2}}} = \frac{b_1 - b_2}{S_{b_1 - b_2}} \qquad (١٠-٣٧)$$

حيث

$S^2_{b_1 - b_2} = S^2_p / \sum x_1^2 + S^2_p / \sum x_2^2$ ،

$\sum x_1^2$ مجموع مربع الانحرافات عن المتوسط بالنسبة للمتغير X_1 فى العينة الأولى،

$\sum x_2^2$ مجموع مربع الانحرافات عن المتوسط بالنسبة للمتغير X_2 فى العينة الثانية،

S^2_p تمثل التباين المشترك pooled variance والذى يعبر عنه كما يلى

$$S^2_p = \frac{(\sum y_1^2 - b_1^2 \sum x_1^2) + (\sum y_2^2 - b_2^2 \sum x_2^2)}{(n_1 - 2) + (n_2 - 2)}$$

أى مجموع المربعات عن خط الاعتماد فى العينة الأولى + مجموع المربعات عن خط الاعتماد فى العينة الثانية مقسوماً على مجموع درجات الحرية للخطأ فى العينتين.

كما يستخدم أيضاً تحليل التباين لإجراء هذا الاختبار ويمكن الرجوع فى ذلك إلى Steel and Torrie (1980).

مثال ١٠-٩

قياسان Y,X أخذا على أفراد عينتين وكانت البيانات التالية:

العينة الأولى:

$n_1 = 10$ ، $b_1 = 1.26$ وحدة من Y لكل وحدة من X

مجموع المربعات عن خط الاعتماد = 2.3 ، $\sum x_1^2 = 17.5$

العينة الثانية:

$n_2 = 10$ ، $b_2 = 1.35$ وحدة من Y لكل وحدة من X

مجموع المربعات عن خط الاعتماد = 4 ، $\sum x_2^2 = 19$

اختبر الفرض القائل بأن معاملى الانحدار هما نفس التقدير لمعامل انحدار العشيرة.

الحل:

$$S_{y.x}^2 = \frac{2.3+4}{10+10-4} = 0.39 \quad ، \quad H_o : \beta_1 - \beta_2 = 0$$

$$S_{b_1-b_2}^2 = 0.39\left(\frac{1}{17.75}+\frac{1}{19}\right) = 0.04 \quad \text{وبالتالى} \quad S_{b_1-b_2} = \sqrt{0.04} = 0.2$$

$$\therefore t = \frac{|1.35-1.26|}{0.2} = 0.045$$

قيمة t الجدولية $t_{(16,0.05)} = 2.12$ وهى أكبر من t المحسوبة (0.45) وبالتالى لا يرفض فرض العدم وعلى هذا فإن b_2, b_1 هما تقديران لنفس معامل انحدار العشيرة β وتحسب β العامة للمثال لأنه ليس هناك سبب لوجود معاملى انحدار ولكن معامل انحدار واحد.

١٠-٩ التوزيع ذو المتغيرين

فى كثير من الحالات يكون لكل من المتغيرين Y,X توزيع احتمالى بمعنى أن أزواج المتغيرين كل منهما مسحوب عشوائياً. فمثلاً كمية اللبن التى تعطيها البقرة فى الموسم والعمر عند أول ولادة لكل بقرة إذا سحبت 20 بقرة عشوائيا وتم تسجيل العمر

وكمية اللبن لكل بقرة. أو عند دراسة العلاقة بين عدد لطع دودة ورق القطن وكمية المحصول فى عدد من القطع اختيرت عشوائياً. فى مثل هذه الحالات تكون العينات مسحوبة من عشيرة ذات متغيرين، وفى هذه الحالة لا توضع قيود أو شروط على أى من المتغيرين، كأن يكون أحدهما ثابتاً fixed مثلاً. ويمكن حساب معاملين للانحدار أحدهما معامل انحدار المتغير الأول مثلاً على المتغير الثانى، والآخر معامل انحدار المتغير الثانى على المتغير الأول. ولو أنه فى كثير من الأحيان يكون الاهتمام مقصوراً على حساب معامل انحدار واحد له معنى معين. وفى العينات المسحوبة من عشيرة ذات متغيرين فإنه يمكن تلخيص بيانات العينة المسحوبة وذلك بحساب متوسط وتباين كل من المتغيرين على حدة وكذلك حساب التغاير بينهما ويمكن أن يعبر عن التغاير بمعامل التحديد coefficient of determination أو الجذر التربيعى له والذى يعرف بمعامل الارتباط البسيط كما سيتضح فيما بعد.

مثال ١٠-١٠

فى عينة من 8 أزواج كانت البيانات التالية:

رقم المشاهدة	١	٢	٣	٤	٥	٦	٧	٨
Y_1	6	4	0	10	14	2	12	8
Y_2	9	3	2	10	15	8	11	12

احسب معامل انحدار Y_1 على Y_2 وكذلك معامل انحدار Y_2 على Y_1 مع تقسيم التباين فى كل من المتغيرين إلى مكوناته.

الحل:

$$\sum Y_1 = 56 \qquad \overline{Y}_1 = 7 \qquad \sum Y_1^2 = 560 \qquad \sum y_1^2 = 168$$

$$\sum Y_2 = 70 \qquad \overline{Y}_2 = 8.75 \qquad \sum Y_2^2 = 748 \qquad \sum y_2^2 = 135.5$$

$$\sum Y_1 Y_2 = 620 \qquad \sum y_1 y_2 = 130$$

معامل انحدار Y_1 على Y_2 0.959 وحدة من Y_1 لكل وحدة من Y_2 ناتج من

$$b_{Y_1Y_2} = \frac{\sum y_1 y_2}{\sum y_2^2} = \frac{130}{135.5} = 0.959$$

معامل انحدار Y_2 على Y_1 0.774 وحدة من Y_2 لكل وحدة من Y_1 ناتج من

$$b_{Y_2Y_1} = \frac{\sum y_1 y_2}{\sum y_1^2} = \frac{130}{168} = 0.774$$

تباين المتغير الأول: $S_{Y_1}^2 = \frac{168}{7} = 24$ والانحراف المعيارى 4.9

تباين المتغير الثانى: $S_{Y_2}^2 = \frac{135.5}{7} = 19.36$ والانحراف المعيارى 4.4

التغاير بين المتغيرين: $\text{cov}(Y_1, Y_2) = \frac{130}{7} = 18.57$

إذا اعتبر أن Y_2 هو المتغير التابع فإن مجموع المربعات الراجع لانحدار Y_2 على Y_1 (RSS):

$$RSS = \frac{(\sum y_1 y_2)^2}{\sum y_1^2} = \frac{(130)^2}{168} = 100.6$$

مجموع المربعات الكلى للمتغير Y_2: $TSS = 135.5$

مجموع المربعات عن خط انحدار Y_2 على Y_1:

$$ESS = 135.5 - 100.6 = 34.9$$

وعلى ذلك فإن:

$$\frac{RSS}{TSS} = \frac{(\sum y_1 y_2)^2 / \sum y_1^2}{\sum y_2^2} = \frac{(130)^2/168}{135.5} = 0.74$$

أما إذا اعتبر أن Y_1 هو المتغير التابع فإن مجموع المربعات الراجع لانحدار Y_2 على Y_1:

$$RSS = \frac{(\sum y_1 y_2)^2}{\sum y_2^2} = \frac{(130)^2}{135.5} = 124.7$$

مجموع المربعات عن خط انحدار Y_1 على Y_2:

$$ESS = 168 - 124.7 = 43.3$$

وعلى ذلك فإن:

$$\frac{RSS}{TSS} = \frac{(\sum y_1 y_2)^2 / \sum y_1^2}{\sum y_1^2} = \frac{(130)^2/135.5}{168} = 0.74$$

أما حاصل ضرب معاملى الانحدار

$$(b_{Y_1Y_2})(b_{Y_2Y_1}) = (0.959)(0.774) = 0.74$$

وبالتالى يوجد معادلتان للانحدار يمثلان خطى الانحدار يمر كل منهما بنقطة تقاطع متوسطى المتغيرين ولا يشترط أن ينطبق الخطين كل منهما على الآخر، إذ لا يتحقق ذلك إلا إذا كان مجموع مربع الانحرافات عن خط الانحدار مساوياً للصفر أى أن جميع النقط واقعة على خط الانحدار بدون أى أخطاء كما يلاحظ أن معامل انحدار Y_2 على Y_1 ليس هو مقلوب معامل انحدار Y_1 على Y_2.

معادلة خط انحدار Y_2 على Y_1 هى:

$$\hat{Y}_2 = a_1 + b_{Y_2Y_1} Y_1$$

حيث:

$$a_1 = \overline{Y}_2 - b_{Y_2Y_1} \overline{Y}_1 \quad \text{وبالتعويض} \quad a_1 = 8.75 - (0.774)(7) = 3.33$$

$$\hat{Y}_2 = 3.33 + 0.774 Y_1 \quad \text{وعلى ذلك فإن معادلة خط الانحدار}$$

أما معادلة خط انحدار Y_1 على Y_2 فهى:

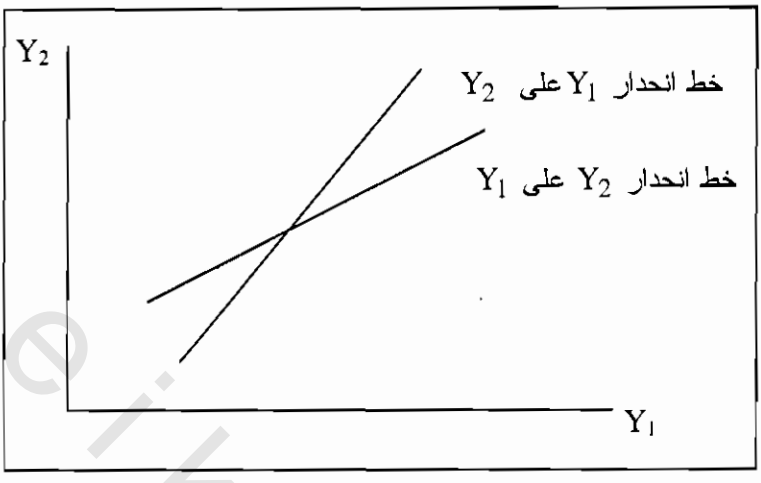
$$\hat{Y}_1 = a_2 + b_{Y_1Y_2} Y_2$$

حيث:

$$a_2 = \overline{Y}_1 - b_{Y_1Y_2} \overline{Y}_2 \quad \text{وبالتعويض} \quad a_2 = 7 - (0.959)(8.75) = -1.39$$

$$\hat{Y}_2 = -1.39 + 0.9594 Y_2 \quad \text{وعلى ذلك فإن معادلة خط الانحدار}$$

ويبين الشكل التالى خطى الانحدار.



ويلاحظ أيضاً أن النسبة بين مجموع المربعات الراجعة للانحدار إلى مجموع المربعات الكلى عندما كان Y_2 هو المتغير التابع هى نفسها عندما كان Y_1 هو المتغير التابع، أى أن هذه النسبة والتى تساوى 0.74 هى نفسها بغض النظر عن كون أى المتغيرين هو التابع وهذه النسبة تساوى حاصل ضرب معاملى الانحدار أى أن:

$$(b_{Y_1Y_2})(b_{Y_2Y_1}) = \frac{(\Sigma y_1y_2)^2/\Sigma y_1^2}{\Sigma y_2^2} = \frac{(\Sigma y_1y_2)^2/\Sigma y_2^2}{\Sigma y_1^2}$$

ويرمز لهذه الكمية بالرمز r^2 ويطلق عليها معامل التحديد coefficient of determination أى أن:

$$r^2 = (b_{Y_1Y_2})(b_{Y_2Y_1}) = \frac{(\Sigma y_1y_2)^2}{(\Sigma y_1^2)(\Sigma y_2^2)} \qquad (٣٨-١٠)$$

والجذر التربيعى لمعامل التحديد (r) يطلق عليه معامل الارتباط (التلازم) البسيط. وكثيراً ما يعرف معامل الارتباط بأنه معامل اعتماد قياسى standard regression، أى أنه إذا قيست المشاهدات فى كل من المتغيرين بوحدات قياسية (أى مقسومة على انحرافها المعيارى) كان معامل اعتماد Y على X فى هذه الحالة يساوى معامل اعتماد X على Y يساوى الارتباط بينهما، فإذا رمز لمعامل الاعتماد القياسى هذا بالرمز b°_{YX} فإن:

$$b^{\circ}_{YX} = \frac{\sum \frac{x}{\sigma_x} \sum \frac{y}{\sigma_y}}{\sum \frac{x^2}{\sigma_x^2}} = \frac{\sum xy}{\sum x^2} . \frac{\sigma_x^2}{\sigma_x \sigma_y}$$

$$b^{\circ}_{YX} = b_{yx} \frac{\sigma_x}{\sigma_y} = b_{xy} \frac{\sigma_y}{\sigma_x} = r \qquad (٣٩-١٠)$$

وسوف يتم فيما بعد شرح معامل الارتباط بالتفصيل.

١٠-١٠ العلاقة بين معامل التحديد r^2 وخطأ التقدير $S_{y.x}$

سبق بيان أن مجموع مربع الانحرافات عن خط الانحدار وكان رمزه ESS أو $d^2_{y.x}$ هو

$$ESS = \sum d^2_{y.x} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = \sum y^2 \left(1 - \frac{(\sum xy)^2}{\sum x^2 \sum y^2}\right)$$

وحيث إن:

$$(b_{Y_1Y_2})(b_{Y_2Y_1}) = \frac{(\sum y_1 y_2)^2}{(\sum y_1^2)(\sum y_2^2)} = r^2$$

$$ESS = \sum y^2 (1 - r^2) \qquad (٤٠-١٠)$$

وحيث إن ESS هى مجموع مربعات فإن قيمتها تساوى صفر على الأقل ولكى يتحقق ذلك يلزم أن تكون قيمة r^2 ما بين الصفر والواحد الصحيح أى أن:

$$\left(0 \le r^2 \le 1\right) \qquad (٤١-١٠)$$

وهذه إحدى خصائص معامل التحديد.

وعندما تكون n أى حجم العينة كبيراً فإن $n-1$ تساوى تقريباً $n-2$ وعلى ذلك فإن العلاقة (١٠-٤٠) تصبح:

$$\frac{ESS}{n-2} \cong \frac{\sum y^2}{n-1}(1-r^2)$$

أى أن:

$$S^2_{y.x} \cong S^2_y(1-r^2) \qquad (١٠-٤٢)$$

وبأخذ الجذر التربيعى للطرفين:

$$S_{y.x} \cong S_y\sqrt{(1-r^2)} \qquad (١٠-٤٣)$$

حيث $S_{y.x}$ تمثل خطأ التقدير أما S_y فتمثل الانحراف المعيارى للمتغير Y.

وأيضاً:

$$\frac{S^2_{y.x}}{S^2_y} \cong 1-r^2 \qquad (١٠-٤٤)$$

$$r^2 \cong 1-\frac{S^2_{y.x}}{S^2_y} \cong \frac{S^2_y - S^2_{y.x}}{S^2_y} \qquad (١٠-٤٥)$$

وعندما تكون $r^2 = 1$، وهو الحد الأقصى لقيمة معامل التحديد، فإن: $S^2_{y.x}$ لابد وأن تساوى الصفر. وعلى ذلك فإن قيمة r^2 تساوى صفر فى حالة عدم وجود علاقة بين المتغيرين أى عندما يكون التباين هو نفسه التباين عن خط الانحدار وكما أن r^2 تساوى الواحد الصحيح عندما تكون جميع النقط واقعة على خط الانحدار.

وإذا وصف معامل التحديد r^2 كما سبق بأنه النسبة من التباين فى Y والتى ترجع إلى اعتماد Y على X فإن المقدار $(1-r^2)$ يمثل النسبة من التباين فى أحد المتغيرين الخالية من تأثير المتغير الثانى. وعلى سبيل المثال إذا كانت $r^2 = 0.74$ (كما فى مثال ١٠-١٠) فمعنى ذلك أن 74% من الاختلاف أو التباين فى أحد المتغيرين ترجع إلى الاختلافات فى قيم المتغير الآخر وأن $1-0.74 = 0.26$ من التباين راجعة إلى الخطأ العشوائى أو التباين الغير منسوب إلى تباين قيم المتغير الثانى.

ويمكن اختبار مدى معنوية معامل التحديد باختبار فرض العدم $H_o: \rho = 0$ ضد الفرض البديل $H_o: \rho \neq 0$ حيث ρ عبارة عن معامل ارتباط العشيرة. ولاختبار ذلك الفرض تستخدم t بدرجات حرية $n-2$ كالتالى:

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (١٠-٤٦)$$

وعلى الرغم من أن معامل التحديد هو المقياس الأكثر استخداماً، إلا أن هذا لا يعطى دليلا كافياً على أن r^2 فعلا تفسر قيمة الاختلافات فى المتغير التابع والراجعة إلى المتغير المستقل. ولذلك لابد من تقييم مدى ملائمة evaluating the fit النموذج المستخدم فى التحليل.

١٠-١١ تقييم ملائمة نموذج التحليل Evaluating the fit

حتى الآن تم شرح النتائج الأساسية والتى تستخدم للوصول إلى استنتاجات فيما يتعلق بمفهوم النموذج الخطى البسيط. وهذه النتائج تكون صالحة وذات معنى حتى الآن فيما يتعلق فقط بالخطأ التجريبى (أو المتبقى residual) فى النموذج المستخدم. وكثيراً ما يجب تحليل مكونات هذا المتبقى من خلال عمل بعض الرسومات البيانية والتحليلية. وكما سبق فإن كبر قيمة معامل التحديد ومعنوية اختبار t له لا يؤكدان أن النموذج المستخدم يعتبر هو الأكثر ملائمة لتحليل البيانات محل الدراسة. والمثال التالى يوضح ذلك. ولابد من الأخذ فى الاعتبار أن التقييم الدقيق للمتبقى له أهمية فى التأكيد على المحافظة على الافتراضات الخاصة باستخدام نظرية المربعات الصغرى.

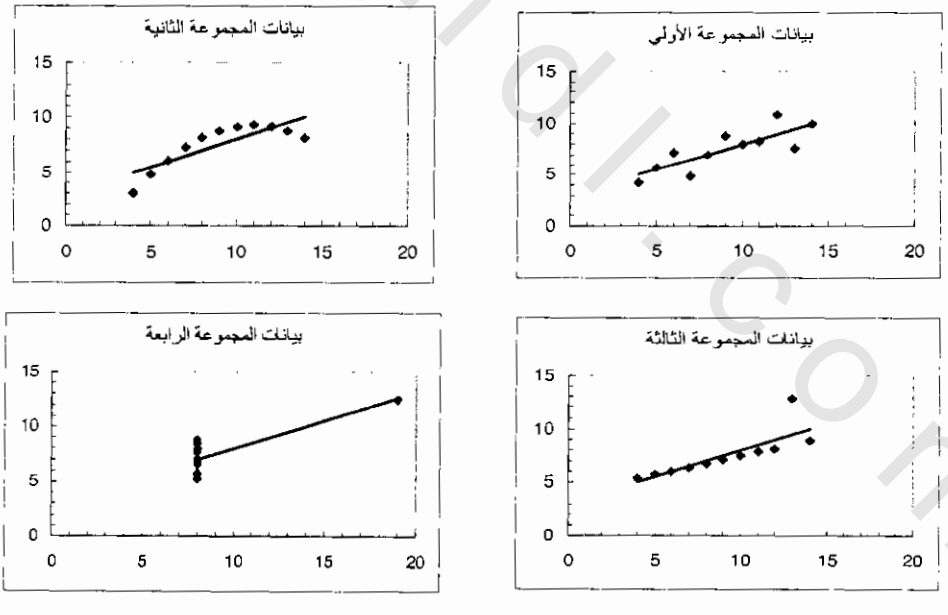
مثال ١٠-١١

يوضح جدول ١٠-٣ أربعة مجموعات من البيانات لمتغيرين X، Y وهذه المجموعات لها نفس الإحصاءات. والشكل ١٠-٨ يوضح الشكل الانتشارى لهذه البيانات مع خط الانحدار لكل مجموعة (المصدر: Chatterjee and Price, 1991).

يظهر تحليل بيانات هذا المثال باستخدام معادلة الانحدار الخطى البسيط أن معادلة الانحدار للمجموعات الأربع من البيانات متطابقة وهى $\hat{Y} = 3 + 0.5X$ ومعامل التحديد قيمته $r^2 = 0.67$ للأربعة مجموعات من البيانات. ولكن الشكل الانتشارى١٠-٨ وخط الانحدار لمجموعات البيانات كل على حده توضح أن التحليلات التى اعتمدت على المتوسط ومعامل الانحدار ومعامل التحديد لم تنجح جميعها فى الكشف عن الاختلافات بين أنماط المجموعات الأربع من البيانات وبالتالى يكون التحليل غير صحيح.

جدول ١٠-٣ أربعة مجموعات من البيانات لمتغيرين X، Y لها نفس المتوسط

مجموعة ١		مجموعة ٢		مجموعة ٣		مجموعة ٤		
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4	
10	8.04	10	9.14	10	7.46	8	6.58	
8	6.95	8	8.14	8	6.77	8	5.76	
13	7.58	13	8.74	13	12.74	8	7.71	
9	8.81	9	8.77	9	7.11	8	8.84	
11	8.33	11	9.26	11	7.81	8	8.47	
14	9.96	14	8.10	14	8.84	8	7.04	
6	7.24	6	6.13	6	6.08	8	5.25	
4	4.26	4	3.10	4	5.39	19	12.50	
12	10.84	12	9.13	12	8.15	8	5.56	
7	4.82	7	7.26	7	6.42	8	7.91	
5	5.68	5	4.74	5	5.73	8	6.89	
9	**7.5**	**9**	**7.5**	**9**	**7.5**	**9**	**7.5**	**المتوسط**



شكل ١٠-٨ شكل انتشارى مع خط الانحدار للأربعة مجموعات من بيانات مثال ١٠-١١.

لاحظ أن المخالفات الصغيرة small violations والخاصة بالتحليل بطريقة المربعات الصغرى لا تؤثر بدرجة كبيرة فى استنتاجات التحليل بينما مخالفات نموذج التحليل المستخدم تؤدى إلى تغيير جذرى فى استنتاجات التحليل.

يعتبر تحليل المتبقى analysis of residual طريقة سهلة للكشف عن مدى ملائمة نموذج تحليل الانحدار. ويوجد طريقتان لتحليل المتبقى: الطريقة الأولى تعتمد على الرسم البيانى حيث يمثل المحور الصادى القيم القياسية للمتبقى والمحور السينى عبارة عن القيم المتوقعة $\hat{Y}$ أو قيم المتغير المستقل X_i، أما الطريقة الثانية فهى تعتمد على تقسيم المتبقى إلى جزئين، الجزء الأول يسمى الخطأ النقى pure error والجزء الثانى يطلق عليه عدم الكفاية lack of fit.

١٠-١١-١ تحليل المتبقى بطريقة الرسم البيانى

من المعادلتين (١٠-٣) و (١٠-٤) وجدول تحليل التباين ١٠-٢ يمكن حساب قيمة الانحراف (المتبقى) عن خط الاعتماد لكل قيمة من المتغير المعتمد Y_i كالتالى:
$e_i = Y_i - \hat{Y}_i$ وهذه القيم يمكن تحويلها إلى قيم قياسية باستخدام $e_{is} = e_i / S_{Y.X}$.

بصفة عامة عندما يتم اختيار النموذج الصحيح فإن القيم القياسية للمتبقى تتراوح بين 2 إلى 2- وتتوزع عشوائياً حول الصفر. والرسم البيانى لابد أن لا يظهر نمط واضح للاختلافات فى هذه القيم.

مثال ١٠-١٢

استخدم بيانات المثال ١٠-١١ لرسم القيم المتبقية القياسية e_{is} على المحور الصادى وكل من القيم المتوقعة $\hat{Y}_i$ وقيم المتغير المستقل X_i للمجموعات الأربع من البيانات. $S_{y.x} = 1.24$ ، $a = 3$ ، $b_{y.x} = 0.5$ للمجموعات الأربع من البيانات.

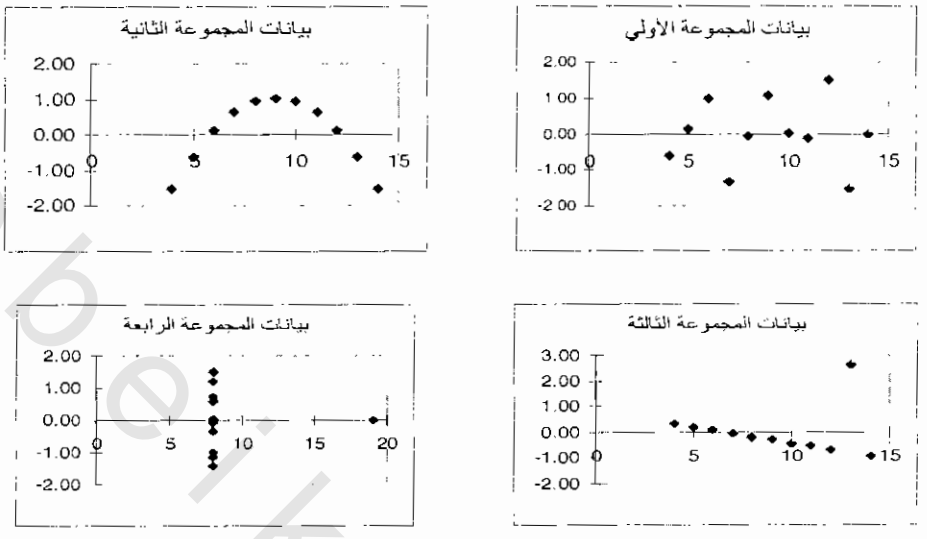
يوضح جدول ١٠-٤ قيم المتغير المستقل و القيم المتوقعة والقيم المتبقية القياسية لمجموعات البيانات الأربع المذكورة فى مثال ١٠-١١، والشكلين ١٠-٩ و ١٠-١٠ يوضحان القيم المتبقية القياسية على المحور الصادى وقيم المتغير المستقل والقيم المتوقعة على المحور السينى للمجموعات الأربع، على الترتيب.

يتضح من الشكلين ١٠-٩ و ١٠-١٠ أن بيانات المجموعة الأولى فقط هى التى يمكن تحليلها بالنموذج $Y = a + bX + e$ وكان توزيع القيم المتبقية القياسية لهذه المجموعة يبدو عشوائياً حول الصفر وتراوحت قيمه بين ± 2 ولم يظهر توزيع هذه القيم أى نمط معين بتغير قيم العامل المستقل أو القيم المتوقعة. بينما الأشكال البيانية الخاصة بمجموعات البيانات الثانية والثالثة والرابعة أظهرت جميعها نمطاً معيناً غير

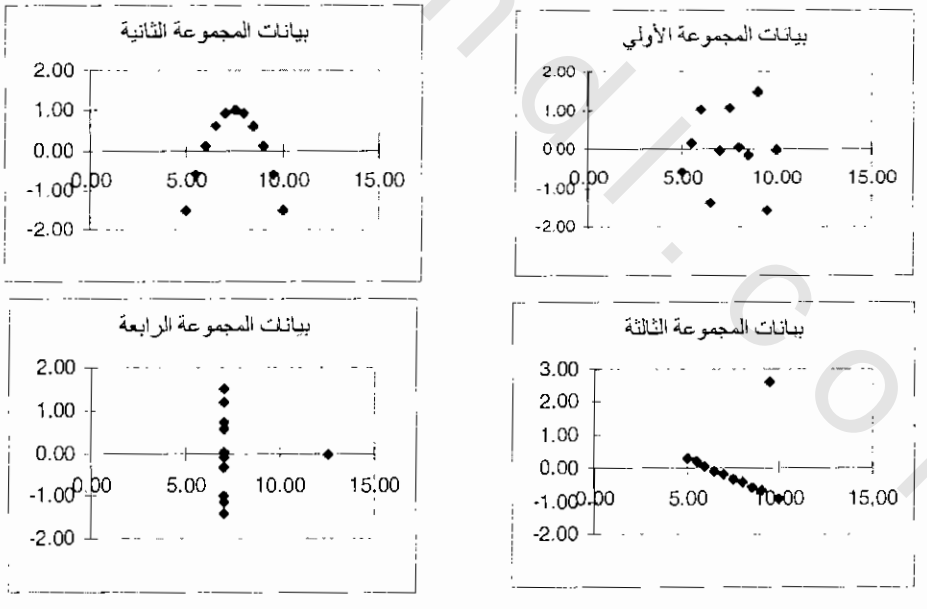
عشوائى لتوزيع القيم المتبقية القياسية بتغير سواء العامل التابع أو القيم المتوقعة مما يدل على عدم صلاحية النموذج $Y = a + bX + e$ فى تحليل هذه المجموعات من البيانات ولابد من البحث عن نموذج آخر غير النموذج الخطى البسيط. ومعنى هذا أن هناك ربما عوامل أخرى تؤثر على المتغير التابع و/أو العلاقة بين X و Y ليست بالخطية ولكنها قد تكون أكث تعقيداً.

جدول ١٠-٤ قيم المتغير المستقل و القيم المتوقعة والقيم المتبقية القياسية للمجموعات الأربع من البيانات المذكورة فى مثال ١٠-١١

مجموعة ١			مجموعة ٢			مجموعة ٣			مجموعة ٤		
X_1	$\hat{Y}_1$	e_{1s}	X_2	$\hat{Y}_2$	e_{2s}	X_3	$\hat{Y}_3$	e_{3s}	X_4	$\hat{Y}_4$	e_{4s}
10	8.0	0.03	10	8.0	0.92	10	8.0	-0.44	8	7.0	-0.34
8	7.0	-0.04	8	7.0	0.92	8	7.0	-0.19	8	7.0	-1.00
13	9.5	-1.55	13	9.5	-0.61	13	9.5	2.62	8	7.0	0.57
9	7.5	1.06	9	7.5	1.03	9	7.5	-0.32	8	7.0	1.49
11	8.5	-0.14	11	8.5	0.61	11	8.5	-0.56	8	7.0	1.19
14	10.0	-0.03	14	10.0	-1.54	14	10.0	-0.94	8	7.0	0.03
6	6.0	1.00	6	6.0	0.11	6	6.0	0.06	8	7.0	-1.41
4	5.0	-0.60	4	5.0	-1.54	4	5.0	0.32	19	12.5	0.00
12	9.0	1.49	12	9.0	0.11	12	9.0	-0.69	8	7.0	-1.16
7	6.5	-1.36	7	6.5	0.61	7	6.5	-0.06	8	7.0	0.74
5	5.5	0.15	5	5.5	-.061	5	5.5	0.19	8	7.0	-0.09



شكل ١٠-٩ القيم المتبقية القياسية على المحور الصادى وقيم المتغير المستقل على المحور السينى للأربعة مجموعات من البيانات.



شكل ١٠-١٠ القيم المتبقية القياسية على المحور الصادى والقيم المتوقعة على المحور السينى للمجموعات الأربع من البيانات.

١٠-١١-٢ طريقة الخطأ النقى وعدم الكفاية Lack of fit and pure error

كما سبق فإن $e_i = Y_i - \hat{Y}_i$ تمثل المتبقى عند X_i وهذه تمثل القيمة التى تختلف بها القيمة الحقيقية Y_i عن القيمة المتوقعة $\hat{Y}_i$. وأيضا سبق إيضاح أن $\sum e_i = 0$. تشتمل قيمة المتبقى هذه على جميع المعلومات المتاحة والتى فشل النموذج المستخدم فى التحليل فى توظيفها لتفسير التباين فى المتغير المعتمد Y. هذا المتبقى يمكن تقسيمه إلى جزئين الأول يعرف بالخطأ النقى pure error، كما سبق، ويرمز له بالرمز S_e^2 والجزء الثانى يعرف بعدم الكفاية lack of fit ويرمز له بالرمز MS_L وهذا الجزء يمثل عدم ملائمة النموذج المستخدم فى تحليل البيانات محل الدراسة. يجب ملاحظة أنه يلزم وجود مشاهدات متكررة لكل من X، Y حتى يمكن تحليل المتبقى. ولحساب ذلك افترض وجود بيانات مكررة يمكن تعريفها كالتالى:

$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ عبارة عن عدد n_1 من المشاهدات المكررة عند X_1

$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ عبارة عن عدد n_2 من المشاهدات المكررة عند X_2

Y_{ju} عبارة عن المشاهدة u عند X_j حيث $u = 1, 2, \cdots, n_j$

$Y_{m1}, Y_{m2}, \ldots, Y_{mn_m}$ عبارة عن عدد n_m من المشاهدات المكررة عند X_m

بمعنى أن $n = \sum_{j=1}^{m}\sum_{u=1}^{n_j} 1 = \sum_{j=1}^{m} n_j$ مشاهدة.

مساهمة n_1 من المشاهدات عند X_1 فى مجموع مربعات الخطأ النقى عبارة عن مجموع المربعات الداخلى لقيم Y_{1u} حول متوسطهم $\overline{Y}_1$ بمعنى أن:

$$\sum_{u=1}^{n_1}(Y_{1u} - \overline{Y}_1)^2 = \sum_{u=1}^{n_1} Y_{1u}^2 - n_1\overline{Y}_1^2 = \sum_{u=1}^{n_1} Y_{1u}^2 - \left(\sum_{u=1}^{n_1} Y_{1u}\right)^2 / n_1 \qquad (١٠-٤٧)$$

تجميع pooling مجموع المربعات الداخلية الراجعة إلى كل المكررات يؤدى إلى الحصول على مجموع المربعات الراجع إلى الخطأ النقى

$$\sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju} - \overline{Y}_j)^2 \qquad (١٠-٤٨)$$

ودرجات الحرية عبارة عن:

$$n_e = \sum_{j=1}^{m}(n_j - 1) = \sum_{j=1}^{m} n_j - m \qquad (٤٩-١٠)$$

وبما أن مجموع مربعات الخطأ النقى هو جزء من مجموع المربعات للمتبقى فإنه يمكن التعبير عن المتبقى للمشاهدة u عند X_j كالتالى:

$$Y_{ju} - \hat{Y}_j = (Y_{ju} - \overline{Y}_j) - (\hat{Y}_j - \overline{Y}_j)$$

لاحظ أن جميع القيم المكررة عن أى X_j سوف يكون لهما نفس القيمة المتوقعة $\hat{Y}_j$. بتربيع جانبى المعادلة (١٠-٤٩) والجمع على كل من u و j يمكن الحصول على:

$$\sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju} - \hat{Y}_j)^2 = \sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju} - \overline{Y}_j)^2 + \sum_{j=1}^{m}\sum_{u=1}^{n_j}(\hat{Y}_{ju} - \overline{Y}_j)^2 \qquad (٥٠-١٠)$$

لاحظ أن الجانب الأيسر من هذه المعادلة عبارة عن مجموع مربعات المتبقى، أما الجانب الأيمن فيحتوى على مجموع مربعات الخطأ الحقيقى وهو

$$\sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju} - \overline{Y}_j)^2$$

أما الجزء الثانى فهو مجموع مربعات عدم الكفاية .

مثال ١٠-١٣

الجدول التالى يوضح عدد 24 مشاهدة بعضها مكرر. ومعادلة الانحدار الخطى البسيط لها $\hat{Y} = 1.436 + 0.338X$. يوضح جدول ١٠-٥ تحليل التباين لهذه البيانات، (المصدر: Draper and Smith, 1981).

مشاهدة	X	Y	مشاهدة	X	Y	مشاهدة	X	Y	مشاهدة	X	Y
١	1.3	2.3	٧	3.3	1.8	١٣	4.7	5.4	١٩	5.3	2.1
٢	1.3	1.8	٨	3.7	3.7	١٤	4.7	3.2	٢٠	5.7	3.4
٣	2.0	2.8	٩	3.7	1.7	١٥	4.7	1.9	٢١	6.0	3.2
٤	2.0	1.5	١٠	4.0	2.8	١٦	5.0	1.8	٢٢	6.0	3.0
٥	2.7	2.2	١١	4.0	2.8	١٧	5.3	3.5	٢٣	6.3	3.0
٦	3.3	3.8	١٢	4.0	2.2	١٨	5.3	2.8	٢٤	6.7	5.9

جدول ١٠-٥ تحليل بيانات مثال ١٠-١٣

SOV	df	SS	MS	F
Due to regression	1	6.326	6.926	6.569
From regression	22	21.192	$0.963 = S^2$	

الخطوة التالية هى حساب مجموع مربعات الخطأ النقى وبالطرح من مجموع مربعات الخطأ الراجع للانحدار يمكن الحصول على مجموع مربعات عدم الكفاية.

١ - مجموع مربعات الخطأ النقى لمكررات القيم Y عند $X = 1.3$ عبارة عن $(0.5)(2.3-1.8)^2 = 0.125$ وهذه القيمة لها درجة حرية واحدة.

٢ - مجموع مربعات الخطأ النقى لمكررات القيم Y عند $X = 4.7$ عبارة عن

$$(5.4)^2 + (3.2)^2 + (1.9)^2 - (3)[(5.4+3.2+1.9)/3]^2$$

$$= 43.01 - (10.5)^2/3 = 6.26$$

وهذه القيمة لها 2 درجة حرية.

ويمكن تلخيص حسابات مجموع مربعات الخطأ النقى فى الجدول التالى:

df	$\sum_{u=1}^{n_j}(Y_{ju} - \overline{Y}_j)^2$	مستوى X
1	0.125	1.3
1	0.845	2.0
1	2.000	3.3
1	2.000	3.7
2	0.240	4.0
2	6.260	4.7
2	0.980	5.3
1	0.020	6.0
11	**12.470**	**المجموع**

وبالتالى يمكن إعادة كتابة جدول تحليل التباين ١٠-٥ فى جدول آخر ١٠-٦ والذى يظهر مجموع مربعات عدم الكفاية ومجموع مربعات الخطأ النقى بدرجات حرية = درجات الحرية من خط الاعتماد – درجات حرية عدم الكفاية، أى 11 = 11 – 22

جدول ١٠-٦ تحليل التباين لبيانات مثال ١٠-١٣ مع إظهار قيمة مجموع مربعات كل من عدم الكفاية والخطأ النقى

SOV	df	SS	MS	F
Due to regression	1	6.326	6.926	6.569
From regression	22	21.192	$0.963 = S^2$	
Lack of fit	11	8.722	$0.793 = MS_L$	0.699
Pure error	11	12.470	$1.134 = S_e^2$	
C. Total	**23**			

لاحظ أن F المحسوبة لعدم الكفاية حسبت من $F = 0.793/1.134 = 0.699$ وهى غير معنوية بمستوى $\alpha = 0.05$ وبالتالى فإن النموذج المستخدم يعتبر كافياً لتفسير التباين فى قيم Y.

مثال ١٠-١٤

يمكن استخدام برنامج SAS لإجراء اختبار عدم الكفاية lack of fit لبيانات مثال ١٠-١٣

```
DATA LACK;
INPUT X Y @@;
CARDS;
1.3 2.3 1.3 1.8 2 2.8 2 1.5 2.7 2.2 3.3 3.8
3.3 1.8 3.7 3.7 3.7 1.7 4 2.8 4 2.8 4 2.2
4.7 5.4 4.7 3.2 4.7 1.9 5 1.8 5.3 3.5 5.3 2.8
5.3 2.1 5.7 3.4 6 3.2 6 3 6.3 3 6.7 5.9
PROC SORT;
BY Y;
PROC RSREG;
MODEL Y = X / COVAR = 1 LACKFIT;
RUN;
```

لاحظ:

لابد من أن يتم ترتيب البيانات على المتغير التابع ولذلك استخدم اختيار PROC SORT باستخدام Y كمتغير للترتيب.

استخدم اختيار PROC RSREG (quadratic response surface) بدلا من PROC REG.

استخدم اختيار COVR = 1 مع النموذج للإشارة إلى أن هناك متغيراً مستقلاً واحداً فقط.

استخدم اختيار lackfit مع النموذج للحصول على مجموع المربعات الراجع لكل من الخطأ النقى وعدم الكفاية.

النتائج:

The RSREG Procedure

Response Surface for Variable Y

Response Mean	2.858333
Root MSE	0.981503
R-Square	0.2298
Coefficient of Variation	34.3383

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Covariates	1	6.324667	0.2298	6.57	0.0178
Linear	0	0	0.0000	.	.
Quadratic	0	0	0.0000	.	.
Crossproduct	0	0	0.0000	.	.
Total Model	1	6.324667	0.2298	6.57	0.0178

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	11	8.723666	0.793061	0.70	0.7183
Pure Error	11	12.470000	1.133636		
Total Error	22	21.193666	0.963348		

صندوق ١٠-٣

- تسمى نسبة التباين الراجع للاعتماد إلى التباين الكلى $(b\sum xy/\sum y^2)$ بمعامل التحديد r^2.
- لا يجب أخذ ارتفاع r^2 ومعنويتها دليلا على أن النموذج المفترض قد استخلص كل المعلومات فى Y، ففى كثير من هذه الحالات تكون مازالت معلومات فى المتبقى لم يستخلصها النموذج.
- فى الحالات التى يكون فيها أكثر من قيمة للمتغير المعتمد Y لنفس قيمة المتغير المستقل X، فإنه يمكن تقسيم مجموع مربعات المتبقى إلى جزئين:

  جزء راجع لكفاية النموذج الإحصائى المستخدم والآخر خطأ نقى، وبهذا يمكن أخذ فكرة عن مدى ملائمة النموذج الإحصائى وعن القدر الذى استخلص به المعلومات فى Y.

١٠-١٢ الارتباط البسيط Simple correlation

(معامل الارتباط لبيرسون Pearson correlation coefficient)

يقيس معامل الارتباط شدة العلاقة بين متغيرين أو هو مقياس لدرجة تغير متغيرين معاً وهو تقدير غير متحيز لمعامل ارتباط المتغيرين فى العشيرة ويحسب من المعادلة التالية:

$$r_{Y_1Y_2} = \frac{\sum(Y_1-\overline{Y}_1)(Y_2-\overline{Y}_2)}{\sqrt{\sum(Y_1-\overline{Y}_1)^2}\sqrt{(Y_2-\overline{Y})^2}} = \frac{\sum y_1y_2}{\sqrt{\sum y_1^2}\sqrt{\sum y_2^2}} \qquad (١٠-٥١)$$

وبقسمة البسط والمقام على درجات الحرية $n-1$ فإن:

$$r_{Y_1Y_2} = \frac{(\sum y_1y_2)/(n-1)}{\sqrt{(\sum y_1^2)/(n-1)}\sqrt{(\sum y_2^2)/(n-1)}}$$

$$= \frac{\text{cov}(Y_1, Y_2)}{(S_{Y_1})(S_{Y_2})} \qquad (٥٢-١٠)$$

حيث S_{Y_1} هى الانحراف المعيارى للمتغير Y_1، S_{Y_2} هى الانحراف المعيارى للمتغير Y_2.

وفى مثال ١٠-١٠ فإن معامل الارتباط $r_{Y_1Y_2} = \frac{130}{\sqrt{168}\sqrt{135.5}} = 0.86$

ومعامل الارتباط هو الجذر التربيعى لحاصل ضرب معاملى الانحدار، أى أنه عبارة عن المتوسط الهندسى لهما أى:

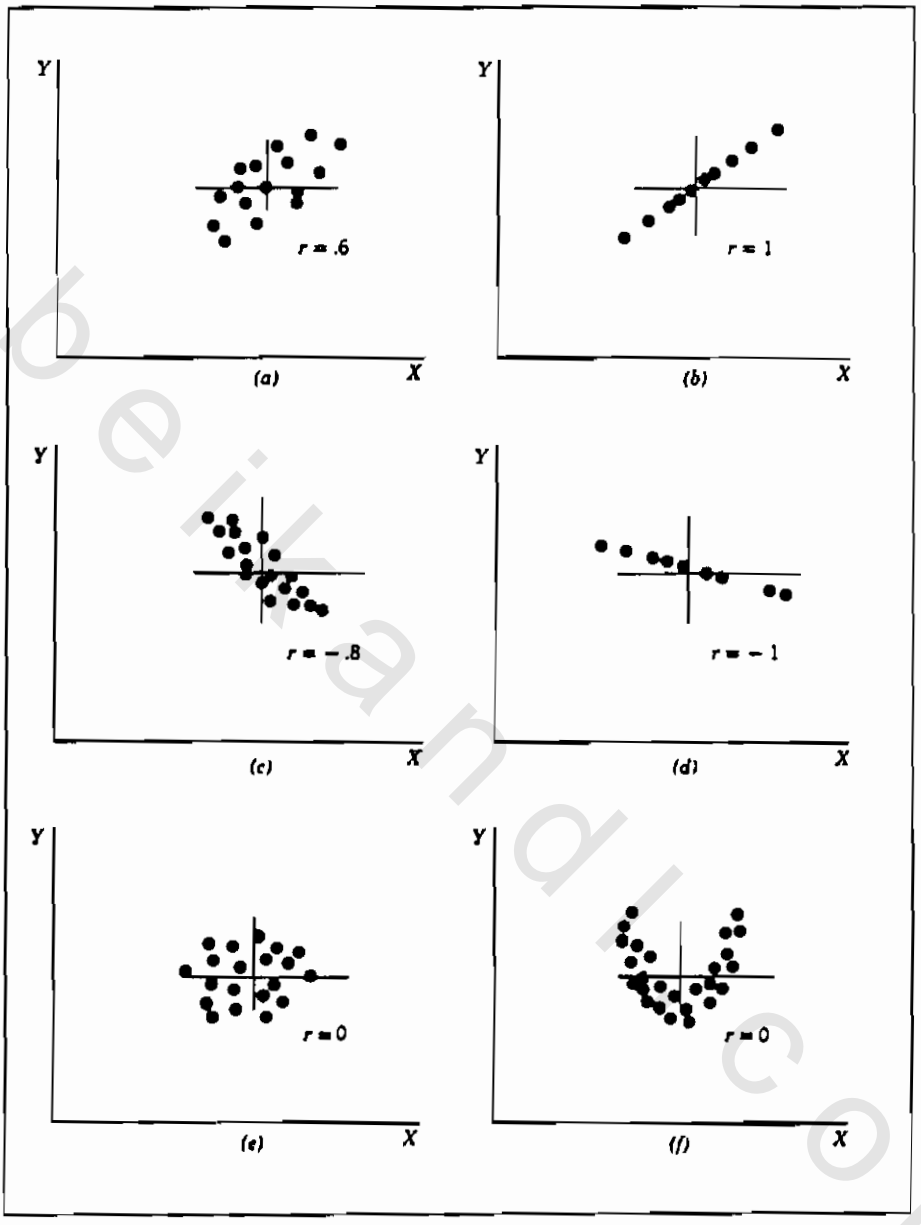
$$r = \sqrt{(b_{Y_1Y_2})(b_{Y_2Y_1})} \qquad (٥٣-١٠)$$

وفى مثال ١٠-١٠ فإن معامل الارتباط $r = \sqrt{(0.959)(0.774)} = 0.86$

وإذا كان معامل التحديد، وهو مربع معامل الارتباط، تنحصر قيمته بين الصفر والواحد الصحيح فإن معامل الارتباط تنحصر قيمته بين 1+ ، 1- أى أن:

$$(-1 \le r \le +1) \qquad (٥٤-١٠)$$

ويكون معامل الارتباط موجباً عندما تكون القيم الكبرى للمتغير الأول تقابلها القيم الكبرى للمتغير الثانى، أى أن تغير المتغيرين يكون فى نفس الاتجاه. ويكون معامل الارتباط سالباً عندما تكون القيم الكبرى لأحد المتغيرين تقابلها القيم الصغرى للمتغير الآخر. ويمكن من الأشكال الانتشارية (١٠-١١ a، b، c، d، e، f) الاستدلال على الارتباط. الشكل (b) يمثل علاقة خطية كاملة موجبة $r = 1$، والشكل (d) يمثل علاقة خطية كاملة سالبة $r = -1$. الشكل (a) يمثل علاقة غير كاملة موجبة $r = 0.6$ والشكل (c) يمثل علاقة غير كاملة سالبة $r = -0.8$. عند مقارنة الشكلين (e) و (f) يلاحظ أن الارتباط قيمته صفر فى الحالتين ولكن الشكل (f) يظهر بوضوح وجود علاقة قوية بين X و Y رغم أن $r = 0$ ومعنى ذلك أن معامل الارتباط فى هذه الحالة لا يعنى عدم وجود علاقة بين المتغيرين ولكنه يعنى عدم وجود علاقة خطية.



شكل ١٠-١١ حالات مختلفة تمثل العلاقة بين متغيرين X و Y

وعلى عكس معامل الاعتماد الذى هو قيمة مميزة فإن معامل الارتباط كمية مجردة أو مطلقة، أى مستقلة عن وحدات القياس. كما يمكن حسابه بوحدات مختزلة

لكل من المتغيرين ولا يلزم لها التعديل بعد ذلك. والملاحظ من دراسات البيانات البيولوجية أن معامل الارتباط نادراً ما يكون أعلى من 0.9.

مثال ١٠-١٥

احسب معامل انحدار Y_2 على Y_1 وكذا معامل انحدار Y_1 على Y_2 ومعامل الارتباط بينهما للبيانات التالية، ثم مثل العلاقة بين المتغيرين بيانياً.

2	0	8	4	1	$:Y_1$
3	1	9	5	2	$:Y_2$

من البيانات:

$$n=5 \qquad \sum Y_1=15 \qquad \sum Y_1^2=85$$

$$\sum Y_1Y_2=100 \qquad \sum Y_2=20 \qquad \sum Y_2^2=120$$

$$b_{Y_1Y_2}=\frac{\sum Y_1Y_2-\frac{\sum Y_1\sum Y_2}{n}}{\sum Y_1^2-\frac{(\sum Y_1)^2}{n}}=\frac{100-\frac{(15)(20)}{5}}{120-\frac{(20)^2}{5}}=1$$

$$b_{Y_2Y_1}=\frac{\sum Y_1Y_2-\frac{\sum Y_1\sum Y_2}{n}}{\sum Y_2^2-\frac{(\sum Y_2)^2}{n}}=\frac{100-\frac{(15)(20)}{5}}{85-\frac{(15)^2}{5}}=1$$

معامل انحدار Y_1 على Y_2 عبارة عن واحد صحيح، أى وحدة من Y_1 لكل وحدة من Y_2، وكذلك معامل انحدار Y_2 على Y_1.

معادلتى خطى الانحدار هما:

$$\hat{Y}_1=-1+Y_2$$

$$\hat{Y}_2=1+Y_1$$

ومعامل الارتباط هو:

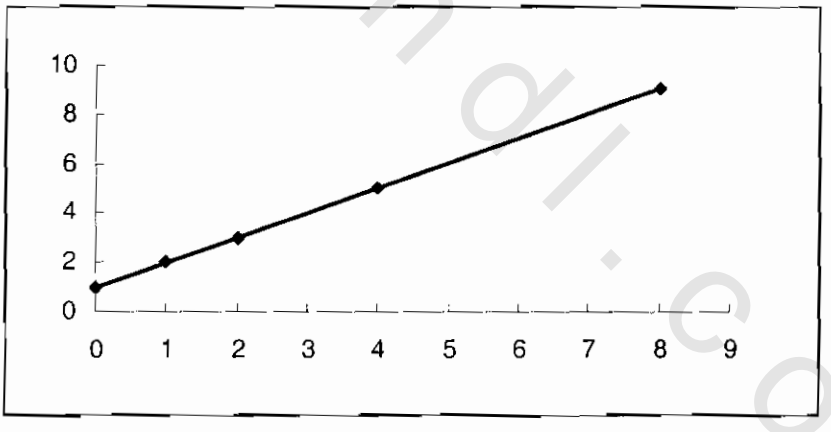
$$r_{Y_2Y_1} = \frac{\sum Y_1Y_2 - \frac{\sum Y_1 \sum Y_2}{n}}{\sqrt{\sum Y_1^2 - \frac{(\sum Y_1)^2}{n}}\sqrt{\sum Y_2^2 - \frac{(\sum Y_2)^2}{n}}}$$

$$= \frac{100 - \frac{(15)(20)}{5}}{\sqrt{85 - \frac{(15)^2}{5}}\sqrt{120 - \frac{(20)^2}{5}}} = 1$$

وكما سبق يمكن استخدام معاملى الانحدار فى حساب معامل الارتباط كالتالى:

$$r = \sqrt{(b_{Y_1Y_2})(b_{Y_2Y_1})} = \sqrt{(1)(1)} = 1$$

والشكل ١٠-١٢ يمثل العلاقة بين المتغيرين بيانياً.



شكل ١٠-١٢ التمثيل البيانى لخطى الانحدار فى مثال ١٠-١٥

لاحظ أن جميع النقط تقع على خط واحد، خط انحدار Y_2 على Y_1 هو نفسه خط الانحدار Y_1 على Y_2 وهذا لا يحدث إلا عندما يكون الارتباط تاماً بين المتغيرين وهى الحالة التى فيها تكون القيم المتوقعة predicted values هى نفسها القيم الفعلية actual values حيث $S^2_{Y_1Y_2} = 0$.

مثال ١٠-١٦

استخدام برنامج SAS فى حساب معامل انحدار Y_2 على Y_1 وكذا معامل انحدار Y_1 على Y_2 ومعامل الارتباط بينهما لبيانات مثال ١٠-١٥.

```
DATA RC;
INPUT Y1 Y2 @@;
CARDS;
1 2 4 5 8 9 0 1 2 3
PROC REG;
MODEL Y1 = Y2;
MODEL Y2 = Y1;
PROC CORR;
RUN;
```

لاحظ:

كل معامل انحدار يراد حسابه يوضع فى نموذج model منفصل .

استخدام PROC CORR لحساب معامل الارتباط.

النتائج

The REG Procedure
Model: MODEL1

Dependent Variable: Y1
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40.00000	40.00000	Infty	<.0001
Error	3	0	0		
Corrected Total	4	40.00000			

Root MSE	0	R-Square	1.0000	
Dependent Mean	3.00000	Adj R-Sq	1.0000	Coeff Var 0

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.00000	0	-Infty	<.0001
Y2	1	1.00000	0	Infty	<.0001

Dependent Variable: Y2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40.00000	40.00000	Infty	<.0001
Error	3	0	0		
Corrected Total	4	40.00000			

Root MSE	0	R-Square	1.0000
Dependent Mean	4.00000	Adj R-Sq	1.0000
Coeff Var	0		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.00000	0	Infty	<.0001
Y1	1	1.00000	0	Infty	<.0001

The CORR Procedure

2 Variables: Y1 Y2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y1	5	3.00	3.16228	15.00	0	8.00
Y2	5	4.00	3.16228	20.00	1.00	9.00

Pearson Correlation Coefficients, N = 5
Prob > |r| under H0: Rho=0

	Y1	Y2
Y1	1.00000	1.00000
Y2	1.00000	1.00000

١٠-١٣ العلاقة بين معامل الانحدار ومعامل الارتباط

١- حيث إن معامل انحدار Y_2 على Y_1 هو

$$b_{Y_2Y_1} = \frac{\text{cov}(Y_1, Y_2)}{S^2_{Y_1}}$$

فإنه بضرب كل من البسط والمقام فى S_{Y_2}

$$b_{Y_2Y_1} = \left(\frac{\text{cov}(Y_1, Y_2)}{S^2_{Y_1}}\right)\left(\frac{S_{Y_2}}{S_{Y_2}}\right) = \left(\frac{\text{cov}(Y_1, Y_2)}{(S_{Y_1})(S_{Y_2})}\right)\left(\frac{S_{Y_2}}{S_{Y_1}}\right)$$

$$\hat{b}_{Y_2Y_1} = (r)\left(\frac{S_{Y_2}}{S_{Y_1}}\right) \qquad (٥٥-١٠)$$

وحيث إن $S_{y_1} = \sqrt{(\sum y_1^2)/(n-1)}$ ، $S_{y_2} = \sqrt{(\sum y_2^2)/(n-1)}$

$$\hat{b}_{Y_2Y_1} = r\sqrt{(\sum y_2^2)/(\sum y_1^2)} \qquad (٥٦-١٠)$$

أى أن:

$$r = b_{Y_2Y_1}\sqrt{(\sum y_1^2)/(\sum y_2^2)} \qquad (٥٧-١٠)$$

وبالمثل فإن:

$$r = b_{Y_1Y_2}\sqrt{(\sum y_2^2)/(\sum y_1^2)} \qquad (٥٨-١٠)$$

٢- حيث إن معادلة خط انحدار Y_2 على Y_1 هى $\hat{Y}_2 = \overline{Y}_2 + b(Y_1 - \overline{Y}_1)$ وبالتعبير عن Y_1 ، Y_2 بوحداتهم القياسية standard units أى:

$$Y_2^\circ = \frac{Y_2 - \overline{Y}_2}{S_{Y_2}} \quad ، \quad Y_1^\circ = \frac{Y_1 - \overline{Y}_1}{S_{Y_1}}$$

وبالتعويض فى معادلة خط الانحدار عن Y_1 ، Y_2 بوحداتهم القياسية فإن

$$\hat{Y}_2^{\circ} S_{y_2} = b_{Y_2 Y_1} Y_1^{\circ} S_{y_1}$$

$$\hat{Y}_2^{\circ} = b_{Y_2 Y_1} \frac{S_{Y_1}}{S_{Y_2}} Y_1^{\circ}$$

$$\hat{Y}_2^{\circ} = r Y_1^{\circ} \qquad (٥٩-١٠)$$

حيث $\hat{Y}_2^{\circ}$ تمثل القيمة المتوقعة معبراً عنها بوحدات قياسية. وباستخدام الوحدات القياسية يصبح معامل الارتباط r هو نفسه معامل الانحدار والاختلافات بينهما تتلاشى.

٣- سبق بيان أن معامل الارتباط هو المتوسط الهندسي لمعاملي الانحدار، أي:

$$r = \pm\sqrt{(b_{Y_1 Y_2})(b_{Y_2 Y_1})}$$

وخطي الانحدار يتطابقان عندما تكون $r = \pm 1$، ويكونان قريبين جداً من بعضهما عندما يكون معامل الارتباط قريب من ± 1.

١٠-١٤ اختبار معنوية معامل الارتباط وتقدير حدود الثقة له

إذا سحب عدد كبير من العينات العشوائية حجم كل منها n من عشيرة تتبع التوزيع الطبيعي لمتغيرين معامل الارتباط بينهما $\rho = 0$ وحسب معامل الارتباط لكل عينة، فإن توزيع معاملات ارتباط العينات يقترب من التوزيع الطبيعي ويزداد اقتراباً منه بزيادة حجم العينة. وبالتالي فإن التوزيع العيني لمعاملات الارتباط يكون متماثلاً أيضاً حتى عندما يكون توزيع أحد المتغيرين غير طبيعي كأن يكون قيماً ثابتة fixed مثلاً يختارها الباحث على أن يكون توزيع المتغير الآخر طبيعياً. أما إذا كان معامل ارتباط العشيرة $\rho \neq 0$ فإن التوزيع العيني لمعاملات الارتباط لا يتبع التوزيع الطبيعي وإنما يكون التوزيع ملتويا skewed مهما كان حجم العينات. ويمثل شكل ١٠-١٣ التوزيع العيني لمعاملات الارتباط المسحوبة من عشيرة معامل الارتباط يساوى صفراً وكذلك التوزيع العيني لمعاملات ارتباط مسحوبة من عشيرة معامل الارتباط بها لا يساوى الصفر.

صندوق ١٠-٤

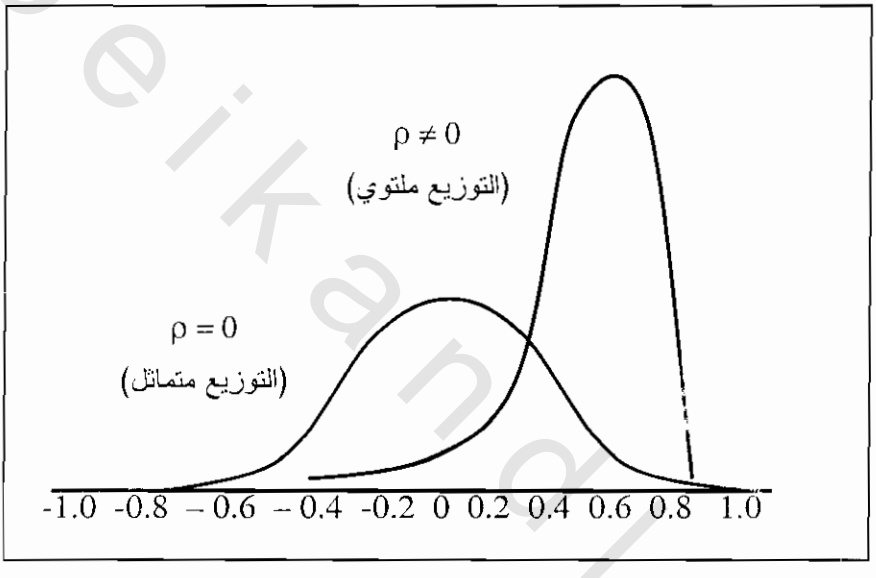
- معامل الارتباط (r) بين متغيرين هو مقياس لشدة العلاقة بينهما. تتراوح قيمته بين 1+، 1- وكلما بعدت قيمته عن الصفر ازدادت شدة العلاقة بين المتغيرين.
- معامل الارتباط كمية مطلقة لا تميز.
- العلاقة الرياضية بين معاملى الاعتماد والارتباط وثيقة جدا، ويمكن تعريف الأخير بأنه الجذر التربيعى لحاصل ضرب معامل اعتماد X على Y ومعامل اعتماد Y على X . ويعرف معامل الارتباط كذلك بأنه معامل الانحدار القياسى أو المعيارى standard (أى بين المتغيرين بعد تعييرهما $(Y-\overline{Y})/\sigma_Y$) لأحد المتغيرين على الآخر

$$r = b_{Y_1Y_2}/\sqrt{\sum y_2^2/\sum y_1^2}$$
$$= b_{Y_2Y_1}/\sqrt{\sum y_1^2/\sum y_2^2}$$
$$= \frac{\sum y_1 y_2}{\sqrt{(\sum y_1^2)(\sum y_2^2)}}$$

- يسمى مربع معامل الارتباط (r^2) بمعامل التحديد أى النسبة من تباين المتغير المرتبط بالمتغير الآخر أو الممكن تفسيرها من قبل المتغير الآخر.
- تختبر معنوية r فى حالة فرض العدم ($\rho = 0$) عن طريق اختبار t أو مباشرة عن طريق جداول خاصة.
- فى حالة ما إذا كان فرض العدم ($\rho \neq 0$) فإن هذه الحالة تحتاج إلى تحوير خاص.

١٠-١٤-١ اختبار معنوية معامل الارتباط عندما يكون معامل الارتباط فى العشيرة يساوى صفراً $(\rho = 0)$

يستخدم فى مثل هذه الحالة، والتى يكون فيها توزيع r متماثلا حول المنتصف (كما فى شكل ١٠-١٣) طريقتان لاختبار معنوية العلاقة بين متغيرين. الأولى اختبار معنوية الانحدار باستخدام اختبار t ودرجات حرية $n-2$ ومستوى معنوية وليكن α. والثانية اختبار معنوية معامل الارتباط باستخدام الجداول الخاصة بذلك. وكل من الاختبارين يعطى نفس النتيجة ويمكن استخدام أى منهما.



شكل ١٠-١٣ التوزيع العينى لمعاملات الارتباط

أولا: استخدام اختبار t:

$$H_o : \rho = 0$$

$$t = \frac{b}{S_b}$$

بدرجات حرية $n-2$ ومن العلاقة (١٠-٥٦)

$$b = r\frac{S_{Y_2}}{S_{Y_1}}$$

$$\hat{t} = r\frac{S_{Y_2}}{S_{Y_1}}/S_b = r\frac{S_{Y_2}}{S_{Y_1}}/\sqrt{\left[\frac{\sum y_2^2 - r^2\sum y_2^2}{(n-2)\sum y_1^2}\right]}$$

$$= \frac{rS_{Y_2}}{S_{Y_1}}\sqrt{\frac{(n-2)\sum y_1^2}{\sum y_2^2(1-r^2)}}$$

وحيث إن:

$$\frac{S_{y_2}}{S_{y_1}} = \sqrt{\frac{\sum y_2^2}{\sum y_1^2}}$$

$$\hat{t} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (١٠-٦٠)$$

حيث يعبر عن $\frac{(1-r^2)}{(n-2)}$ بتباين معامل الارتباط.

مثال ١٠-١٧

فى مثال ١٠-١٠ اختبر معنوية معامل الارتباط إذا كان معامل ارتباط العشيرة يساوى صفراً.

حيث $n = 8$، $r = 0.86$، درجات الحرية $n-2 = 8-2 = 6$ وحيث إن:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(0.86)(\sqrt{6})}{\sqrt{1-(0.86)^2}} = 8.09$$

وبما إن $t_{(6,0.05)} = 2.447$، $t_{(6,0.01)} = 3.707$ وهذه القيم أعلى من قيمة t المحسوبة (8.09) وبالتالى يرفض فرض العدم وهذا يعنى أن العلاقة بين المتغيرين معنوية بدرجة ثقة 99%.

ثانيا: وضعت جداول بمستوى معنوية 5% وأيضاً بمستوى معنوية 1% لاختبار معنوية معامل الارتباط مباشرة وذلك بدرجات حرية $n-2$ (جدول ٧ ملحق أ)، وتتوقف نتيجة الاختبار على حجم العينة وعلى قيمة r فمثلاً:

العينة الأولى: $r_1 = 0.3$ $n_1 = 10$ درجات الحرية = 8

العينة الثانية: $r_2 = 0.3$ $n_1 = 100$ درجات الحرية = 98

العينة الثالثة: $r_3 = 0.7$ $n_3 = 10$ درجات الحرية = 8

على ذلك عند درجات حرية 8 كان معامل الارتباط والذى قيمته 0.3 غير معنوى بينما معامل الارتباط وقيمته 0.7 فهو معنوى بمستوى معنوية 5% عند نفس درجات الحرية. بينما معامل الارتباط (0.3) والذى كان غير معنوى عند درجات حرية 8 يكون معنوياً بمستوى معنوية 1% عند درجات حرية 98، ويجب ملاحظة أنه بزيادة درجات الحرية فإن الحد الأدنى لمعامل الارتباط المعنوى يقل.

والعلاقة المعنوية بين المتغيرين لا تعنى بالضرورة أن أحد المتغيرين هو سبب التغير فى المتغير الآخر (أى علاقة سببية) فقد يكون السبب راجع إلى مصادر أخرى. وإذا كانت $r = 0.3^{**}$ وبالرغم من أن معامل الارتباط معنوى جداً – أى رفض فرض العدم القائل بأنه ليس هناك علاقة بين المتغيرين وقبول الفرض البديل بوجود علاقة بينهما – إلا أن هذا يعنى أن $(0.3)^2 = 0.09$ من تباين Y_2 يرجع للعلاقة بين المتغيرين Y_1، Y_2 وأن $(1 - 0.09) = 0.91$ من التباين خالى من تأثير تلك العلاقة ويرجع إلى مصادر أخرى ذكرت فيما سبق عند الحديث عن عنصر الخطأ فى فصل (١٠-١). وتبين الاختبارات الإحصائية أن هناك علاقة خطية بميل لا يساوى الصفر ومعنوى. ولكن هذا لا يعنى أن معظم الاختلافات فى Y_2 هى نتيجة انحدارها على Y_1 بل قد يكون هناك مصادر أخرى للتباين.

وجدير بالملاحظة أنه إذا كان معامل انحدار b_{YX} معنوياً عند مستوى معنوية معين كان أيضاً كل من b_{XY}، r معنوياً عند نفس المستوى.

١٠-١٤-٢ اختبار معنوية معامل الارتباط عندما يكون معامل ارتباط العشيرة لا يساوى صفراً $(\rho \neq 0)$

لإجراء مثل هذا الاختبار فإنه يفترض أن يكون توزيع كل من المتغيرين طبيعياً، أى يشترط أن تكون أزواج المتغيرات فى العينة تمثل عينة عشوائية مأخوذة من عشيرة ذات متغيرين وهى السابق الإشارة إليها. وإذا كان ρ لا يساوى صفراً فإن التوزيع العينى لمعاملات الارتباط يكون ملتوياً ولا يجوز استخدام t والتى تستخدم فقط عندما يكون التوزيع متماثلاً أى عندما تكون $\rho = 0$. ولقد وجد Fisher أنه بتحويل r إلى قيمة أخرى هى Z تصبح الأخيرة ذات توزيع طبيعى. وتحسب Z كما يلى:

$$Z_r = \frac{1}{2}[\log_e(1+r) - \log_e(1-r)] \qquad (٦١-١٠)$$

حيث $\log_e$ هو اللوغاريتم الطبيعى أى ln ، وتتوزع Z طبيعياً بمتوسط Z_ρ حيث $Z\rho = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho}$ وانحراف معيارى $\sigma_Z = \frac{1}{\sqrt{n-3}}$.

ووضعت جداول لتحويل r إلى ما يقابلها من قيم Z (جدول ٨ ملحق أ) وأخرى للتعبير عن قيم Z بما يقابلها من قيم Z (جدول ٩ ملحق أ) ويمكن تغير أى منهما للأخرى بدرجة كافية من الدقة.

وعلى ذلك فالكمية $\frac{Z_r - Z_\rho}{\sqrt{n-3}}$ تتوزع طبيعياً أيضاً. وتستخدم جداول التوزيع الطبيعى أو جدول t عند درجات حرية ∞ ويفضل ألا يقل حجم العينة عن 50 حتى يكون معامل الارتباط أكثر وثوقاً ...

مثال ١٠-١٨

أخذت عينة عشوائية من عشيرة ذات متغيرين وكان معامل الارتباط 0.5، اختبر ما إذا كانت هذه العينة مسحوبة من عشيرة معامل الارتباط بها 0.6 علماً بأن حجم العينة كان 50 فرداً بدرجة ثقة 95% .

$H_0 : \rho = 0.6$

إذا كانت $r = 0.5$ ، فإن قيمة Z المقابلة لها هى $Z = 0.549$

وعند قيمة $\rho = 0.6$ قيمة Z المقابلة لها هى $Z\rho = 0.693$

وحيث إن $\sigma_Z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{50-3}} = 0.146$

وبالتالى

$$\frac{|Z_r - Z_\rho|}{\sigma_Z} = \frac{|0.549 - 0.693|}{0.146} = 0.99$$

وهذه القيمة (0.99) تقارن بقيمة $t_{(\infty, 0.05)} = 1.96$. وعلى ذلك لا يرفض فرض العدم بأن معامل ارتباط العينة 0.5 لا يختلف عن معامل ارتباط العشيرة 0.6 أى أن هذه العينة مسحوبة من عشيرة معامل الارتباط بها 0.6.

١٠-١٤-٣ تقدير حدود الثقة لمعامل ارتباط العشيرة باستخدام بيانات العينة

حيث إن التوزيع العينى لمعاملات الارتباط يصبح ملتوياً إذا كان معامل الارتباط فى العشيرة لا يساوى الصفر وحيث إن Z تتوزع طبيعياً بغض النظر عن حجم العينة بمتوسط Z_ρ وتباين $\frac{1}{n-3}$ فإن حدود الثقة لـــ Z_ρ هى:

$$Pr(Z_r - Z_{\alpha/2}\sigma_Z \le Z_\rho \le Z_r + Z_{\alpha/2}\sigma_Z) = 1-\alpha \qquad (١٠-٦٢)$$

يمكن أن يعبر عن قيمة $Z_{\alpha/2}$ بأنها قيمة t عند مستوى معنوية α ودرجات حرية ∞.

وباستخدام (جدول ٩ ملحق أ) يمكن إيجاد قيم r المقابلة والتى تمثل حدود الثقة لمعامل ارتباط العشيرة.

مثال ١٠-١٩

ما هى حدود الثقة لمعامل ارتباط العشيرة إذا كان معامل ارتباط العينة 0.425 وحجم العينة 35 وذلك بدرجة ثقة 99% ؟

بما أن $r = 0.425$ وباستخدام (جدول٩ ملحق أ) فإن $Z_r = 0.454$

$$t_{(\infty,0.01)} = 2.576 \text{ ، } \sigma_Z = \frac{1}{\sqrt{35-3}} = \frac{1}{32} = 0.177$$

∴ حدى الثقة لـــ Z_ρ هما:

$$0.454 \pm (2.576)(0.177) = 0.454 \pm 0.455$$

أى هما 0.001- يمثل الحد الأدنى و 0.909+ يمثل الحد الأعلى. وبتحويل Z إلى قيم r المقابلة يكون الحد الأدنى 0.001- والحد الأعلى 0.721 بدرجة ثقة 99% .

لاحظ أن حدى الثقة ليسا على نفس المسافة على كلا الجانبين لمعامل ارتباط العينة حيث التوزيع غير متماثل كما أنه إذا زاد حجم العينة فإن فترة الثقة تنكمش والعكس صحيح.

١٠-١٥ اختبار تساوى معاملى ارتباط لكل من عينتين مأخوذتين من نفس العشيرة معامل الارتباط بها لا يساوى صفراً

أى اختبار فرض العدم بأن الفرق بين معاملى الارتباط يساوى الصفر. ولكى يتم ذلك تحول كل من قيمتى معامل الارتباط إلى قيم Z المقابلة ثم تختبر معنوية الفرق بين قيمتى Z وما ينطبق على Z فهو ينطبق على r.

أى تحول r_1 إلى Z_1 حيث تباين Z_1 هو $\sigma^2_{Z_1} = \frac{1}{n_1 - 3}$

و تحول r_2 إلى Z_2 وتباين Z_2 هو $\sigma^2_{Z_2} = \frac{1}{n_2 - 3}$

وعلى ذلك فإن $\sigma^2_{Z_1 - Z_2} = \sigma^2_{Z_1} + \sigma^2_{Z_2}$ والانحراف المعيارى للفرق بين Z_1 و Z_2 هو

$$S_{Z_1 - Z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \qquad (١٠-٦٣)$$

وحيث إن $\frac{Z_1 - Z_2}{S_{Z_1 - Z_2}}$ تتبع التوزيع الطبيعى فيمكن مقارنتها بقيمة t عند مستوى معنوية α ودرجات حرية ∞.

مثال ١٠-٢٠

فى تجربة كانت البيانات كالتالى:

العينة الأولى: $n_1 = 15$ ، $r_1 = 0.32$

العينة الثانية: $n_2 = 35$ ، $r_2 = 0.65$

اختبر ما إذا كان معاملا الارتباط للعينتين مسحوبين من نفس العشيرة التى معامل الارتباط بها لا يساوى صفراً.

بما أن $r_1 = 0.32$ فإن $Z_1 = 0.332$ ،

$r_2 = 0.65$ فإن $Z_1 = 0.775$

والتباين لكل من Z_1 و Z_2 عبارة عن $\sigma^2_{Z_1} = \frac{1}{12}$ ، $\sigma^2_{Z_2} = \frac{1}{32}$

والانحراف المعيارى للفرق بينهما من العلاقة (١٠-٦٣)

أى أن:

$$S_{Z_1-Z_2} = \sqrt{\frac{1}{12}+\frac{1}{32}} = 0.338$$

وبالتالى فإن

$$t = \frac{|0.332-0.775|}{0.338} = 1.31$$

وتقارن قيمة t المحسوبة بالقيمة $t_{(\infty,0.05)} = 1.96$ وبالتالى فإن Z_1 و Z_2 من نفس العشيرة وبالتالى فإن معاملى الارتباط هما لعينتين من نفس العشيرة.

١٠-١٦ اختبار تجانس عدد من معاملات الارتباط وتجميعهم فى قيمة واحدة كتقدير أكثر صلاحية لمعامل ارتباط العشيرة

إذا كانت العينات مأخوذة من نفس العشيرة تحول قيم معاملات الارتباط للعينات لمختلفة إلى قيم Z المناظرة لها ثم يختبر الفرض بأن معاملات ارتباط العينات هى تقدير لنفس معامل ارتباط العشيرة باستخدام المعادلة التالية:

$$\sum W_i (Z_i - \overline{Z}_W)^2 \qquad (١٠-٦٤)$$

والتى تساوى

$$\sum W_i Z_i^2 - (\sum W_i Z_i)^2 / \sum W_i \qquad (١٠-٦٥)$$

والتى تتوزع حسب توزيع مربع كاى χ^2 بدرجات حرية = (عدد معاملات ارتباط العينات - 1) حيث

$$W_i = 1/\sigma^2_{Z_i} = (n_i - 3) \qquad (١٠-٦٦)$$

أى أن W_i هى مقلوب تباين Z_i وتستخدم كنوع من عوامل الوزن weight حيث يعطى وزن أكبر للعينات كبيرة الحجم ووزن أقل للعينات الأقل حجما. وإذا كانت قيمة مربع كاى غير معنوية فهذا يعنى أن قيم Z متجانسة وبالتالى فإن قيم r للعينات تكون مأخوذة من نفس العشيرة ويحسب المتوسط $\overline{Z}_W$ حيث إنه يساوى مجموع قيم Z المرجحة بمقلوب تبايناتها على مجموع الأوزان أى أن:

$$\overline{Z}_W = \frac{\sum W_i Z_i}{\sum W_i} \qquad (٦٧-١٠)$$

ثم تستخرج قيمة r المقابلة لقيمة $\overline{Z}_W$ وتكون كتقدير لمعامل ارتباط العشيرة وتباين هذا المتوسط هو:

$$\sigma^2_{\overline{Z}_W} = \frac{1}{\sum (n_i - 3)} \qquad (٦٨-١٠)$$

وبذا يكون انحرافه المعيارى

$$\sigma_{\overline{Z}_W} = \sqrt{\frac{1}{\sum (n_i - 3)}} \qquad (٦٩-١٠)$$

ولقد وجد Fisher أن قيم Z تكون متحيزة biased فكل منها أكبر بهذه الكمية:

$$\frac{\rho}{2(n_i - 1)} \qquad (٧٠-١٠)$$

ويظهر أثر هذا التحيز واضحاً عند الرغبة فى الحصول على تقدير لمعامل الارتباط فى العشيرة من عدد كبير من معاملات الارتباط للعينات المختلفة، حيث تتجمع هذه الكمية الصغيرة لكل ويظهر تأثيرها؛ ولذا يجب التصحيح لهذا التحيز بالعلاقة التالية:

$$\text{corected } Z_i = \hat{Z}_i = Z_i - \left(\frac{\rho}{2(n_i - 1)} \right) \qquad (٧١-١٠)$$

وإذا لم يكن معامل الارتباط فى العشيرة معروفاً، وهذا ما يحدث غالباً، تستخدم r المقابلة $\overline{Z}_W$ بدلاً من ρ ثم يحسب متوسط قيم Z المصححة والموزونة كما يلى:

$$\overline{Z}^{\circ}_W = \frac{\sum W_i Z^{\circ}_i}{\sum W_i} \qquad (٧٢-١٠)$$

وتكون قيمة r المقابلة لقيمة $\overline{Z}^{\circ}_W$ هى معامل الارتباط فى العشيرة.

وإذا كان حجم العينات لا يختلف كثيراً فقد يهمل الوزن ويحسب المتوسط حيث يقسم مجموع قيم Z المقابلة لقيم r على عددها ثم تستخرج قيمة r المقابلة.

مثال ١٠-٢١

البيانات التالية مأخوذة من عشيرة ما:

العينة	معامل الارتباط	حجم العينة
١	0.62	30
٢	0.59	5
٣	0.65	12
٤	0.60	17

هل هذه العينات مأخوذة من عشيرة معامل الارتباط فيها وليكن ρ، وما هو التقدير الأكثر صلاحية لهذا المعامل؟

العينة	معامل الارتباط r_i	حجم العينة	Z_i	الوزن $W_i = n_i - 3$	W_iZ_i	$W_iZ_i^2$
1	0.62	30	0.725	27	19.575	14.192
2	0.59	5	0.678	2	1.356	0.919
3	0.65	12	0.775	9	6.975	5.406
4	0.60	17	0.693	14	9.702	6.723
المجموع		**64**		**52**	**37.608**	**27.240**

ولاختبار فرض العدم بأن معاملات الارتباط فى العينات المختلفة لها نفس معامل الارتباط فى العشيرة فإن:

$$\chi^2 = \Sigma(n_i - 3)Z_i^2 - \frac{[\Sigma(n_i - 3)Z_i]^2}{\Sigma(n_i - 3)}$$

حيث إن

وبالتعويض $\chi^2 = 27.240 - \frac{(37.608)^2}{52} = 0.04$ بدرجات حرية $(3 = 1 - 4)$.

والقيمة الجدولية $\chi^2_{(3,0.05)} = 7.81$ ، وعلى ذلك فإن χ^2 المحسوبة غير معنوية أى أن العينات التى قدر فيها معاملات الارتباط مأخوذة من نفس العشيرة. ويكون التقدير الأكثر صلاحية لمعامل الارتباط فى العشيرة هو القيمة المقابلة $\overline{Z}_W$.

وحيث إن $\bar{Z}_W = \frac{\sum W_i Z_i}{\sum W_i} = \frac{37.608}{52} = 0.72$ فإن قيمة r التى تقابلها هى 0.62.

ولتصحيح التحيز لتقدير معامل الارتباط فى العشيرة يستخدم التقدير 0.62 بدلاً من ρ فى الكمية $\frac{\rho}{(2)(n_i - 1)}$ حيث إن معامل الارتباط فى العشيرة غير معروف كالتالى:

العينة	حجم العينة	الوزن $W_i = n_i - 3$	r_i	Z_i	$\frac{\rho}{2(n_i - 1)}$	Z المصححة Z°	$W_i Z_i^\circ$ المصححة والموزونة
١	30	27	0.62	0.725	0.011	0.714	19.278
٢	5	2	0.59	0.678	0.078	0.600	1.200
٣	12	9	0.65	0.775	0.028	0.747	6.723
٤	17	14	0.60	0.693	0.019	0.674	9.436
المجموع	**64**	**52**					**36.637**

وبالتالى فإن $\bar{Z}^\circ_W = 36.637/52 = 0.704$ وقيمة r المقابلة هى 0.6 والتى تعتبر تقديراً غير متحيزاً لمعامل الارتباط فى العشيرة.

١٠-١٧ السبب والأثر فى تحليل الارتباط والانحدار

Cause and effect in regression and correlation analysis

عند تفسير نتائج تحليل الانحدار والارتباط لابد أن يكون من الواضح تماماً أنه ليس بالضرورة أن نتائج التحليل تدل على السبب والأثر cause and effect، فعلى سبيل المثال فى إحدى الدراسات بلغ معامل الارتباط بين دخل الأسرة واستهلاك اللحوم فى فترة زمنية معينة 0.6، هذا ليس بالضرورة دليلا على أن زيادة دخل الأسرة يؤدى إلى زيادة استهلاك اللحوم أو زيادة استهلاك اللحوم يؤدى إلى زيادة دخل الأسرة ولكن هذا المعامل يدل على أن المتغيرين يتغيران سوياً فى نفس الاتجاه وذلك قد يكون بسبب وجود متغير ثالث، وليكن ارتفاع مستوى المعيشة بصفة عامة، والذى يؤثر فى كل من الدخل واستهلاك اللحوم. وفى حالة ما إذا أمكن تثبيت هذا العامل

الثالث والتعامل معه بطريقة إحصائية معينة فإنه فى هذه الحالة فقط فإن معامل الارتباط قد يكون ذا مدلول معين. وهذا هو الهدف من دراسة الارتباط المتعدد أو الجزئى والذى سوف يتم تناوله فيما بعد. وبالمثل عند دراسة الانحدار البسيط فإن الحصول على معامل انحدار معنوى لا يدل أيضاً على السبب والأثر إلا إذا أخذت عوامل أخرى فى الاعتبار.

وعلى الرغم من أن نتائج دراسة الانحدار والارتباط البسيط ليست دليلا على السبب والأثر فإن تلك النتائج ربما تعطى بعض الاقتراحات التى قد تساعد فى دراسة السبب والأثر فمثلاً وجد أن التدخين له علاقة ارتباط عالية بالإصابة بسرطان الرئة مما أوحى إلى ضرورة دراسة السبب والأثر بين المتغيرين لاستيضاح العلاقة البيوكيميائية بينهما.

١٠-١٨ تباين الدالة الخطية Variance of a linear function

إذا كانت $L_1 = Y_1 + Y_2$ فإن تباين L_1 عبارة مجموع التباينين لكل من Y_1 و Y_2.

أى أن:

$$\sigma^2_{L_1} = \sigma^2_{y_1} + \sigma^2_{y_2} \qquad (٧٣-١٠)$$

وذلك إذا كان Y_1 و Y_2 غير مرتبطين uncorrelated. وأيضاً فإن تباين الفرق بين هذين المتغيرين يساوى مجموع التباينين. أى أنه إذا كانت $L_2 = Y_1 - Y_2$ فإن:

$$\sigma^2_{L_2} = \sigma^2_{y_1} + \sigma^2_{y_2} \qquad (٧٤-١٠)$$

أى أن تباين الفرق يساوى تباين المجموع فى حالة المتغيرات غير المرتبطة. ويمكن إثبات ذلك كما يلى:

حيث إن التباين لأى متغير وليكن L هو:

$$V(L) = \frac{\sum (L - \mu_L)^2}{N}$$

وحيث إن $L = Y_1 + Y_2$ فإن $\mu_L = \mu_{Y_1} + \mu_{Y_2}$ وبالتالى:

$$V(L) = \frac{1}{N}\sum[Y_1 + Y_2 - (\mu_{Y_1} + \mu_{Y_2}]^2$$

$$= \frac{1}{N}\sum[(Y_1 - \mu_{Y_1}) + (Y_2 - \mu_{Y_2})]^2$$

$$= \frac{1}{N}[\sum(Y_1 - \mu_{Y_1})^2 + \sum(Y_2 - \mu_{Y_2})^2 + 2\sum(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})$$

$$= \frac{\sum(Y_1 - \mu_{Y_1})^2}{N} + \frac{\sum(Y_2 - \mu_{Y2})^2}{N} + 2\frac{\sum(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})}{N}$$

وبالتالى فإن

$$\sigma_L^2 = \sigma_{y_1}^2 + \sigma_{y_2}^2 + 2\mathrm{cov}(y_1, y_2) \qquad (٧٥-١٠)$$

وإذا كان المتغيران غير مرتبطين فإن الحد الأخير (التغاير covariance) يساوى صفراً.

وبالمثل يمكن إثبات أن تباين الفرق يساوى مجموع التباينين إذا كان المتغيران غير مرتبطين.

أما إذا كان المتغيران مرتبطين فإن تباين المجموع كما فى (١٠-٧٥) وحيث إن

$$\rho = \frac{\mathrm{cov}(y_1, y_2)}{\sigma_{y_1}\sigma_{y_2}}$$

أى أن $\mathrm{cov}(y_1, y_2) = \rho\sigma_{y_1}\sigma_{y_2}$، وعلى ذلك تصبح المعادلة (١٠-٧٥) على الصورة التالية:

$$\sigma_{L_1}^2 = \sigma_{y_1}^2 + \sigma_{y_2}^2 + 2\rho\sigma_{y_1}\sigma_{y_2} \qquad (٧٦-١٠)$$

وعلى ذلك فإن الارتباط الموجب يزيد من تباين المجموع عن مجموع التباين بالمقدار $2\rho\sigma_{y_1}\sigma_{y_2}$ والارتباط السالب يخفضه بنفس الكمية، أى

$$\sigma_{L_1}^2 = \sigma_{y_1}^2 + \sigma_{y_2}^2 - 2\rho\sigma_{y_1}\sigma_{y_2} \qquad (٧٧-١٠)$$

لاحظ هنا أن الارتباط الموجب ينقص تباين الفرق عن مجموع التباينين بينما الارتباط السالب يزيده عن مجموع التباينين. وفى تجارب الأزواج paired expernments فى تحليل t فإن الهدف الأساسى هو إيجاد الارتباط الموجب بين فردى الزوج وبالتالى ينقص تباين الفرق عن مجموع التباينين بالكمية $2\rho\sigma_{y_1}\sigma_{y_2}$

وبنفس المفهوم يمكن إثبات أن:

$$\sigma^2_{(a_1Y_1+a_2Y_2)} = a_1^2\sigma^2_{y_1} + a_2^2\sigma^2_{y_2} + 2\rho a_1 a_2 \sigma_{y_1}\sigma_{y_2} \qquad (٧٨-١٠)$$

حيث إن a_1 ، a_2 ثابتان.

وللتعميم : إذا كانت L دالة خطية بالصورة $L = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$ فإن تباين L هو:

$$\begin{aligned}\sigma^2_L &= a_1^2\sigma^2_{y_1} + a_2^2\sigma^2_{y_2} + \ldots + 2a_1a_2\rho_{y_1y_2}\sigma_{y_1}\sigma_{y_2} \\ &\quad + \ldots + 2a_{(n-1)}a_n\rho_{y_{(n-1)}\cdot y_n}\sigma_{y_{(n-1)}}\sigma_{y_n} \qquad (٧٩-١٠) \\ &= \sum a_i^2\sigma^2_{y_i} + \sum_i\sum_{j\neq i} a_ia_j\rho_{y_iy_j}\sigma_{y_i}\sigma_{y_j}\end{aligned}$$

وإذا كانت المتغيرات غير مرتبطة فإن الحد الثانى من المعادلة فى الطرف الأيمن يساوى صفرا وتصبح المعادلة:

$$\sigma^2_L = \sum a_i^2\sigma^2_{y_i} \qquad (٨٠-١٠)$$

وإذا كان هناك دالتان خطيتان فى نفس المتغيرات: ولتكن الدالة الأولى $L_3 = a_1Y_1 + a_2Y_2$ والدالة الثانية $L_4 = b_1Y_1 + b_2Y_2$ ، فإن التغاير بينهما:

$$\begin{aligned}\mathrm{cov}(L_3, L_4) &= a_1b_1\sigma^2_{y_1} + a_2b_2\sigma^2_{y_2} \\ &\quad + (a_1b_2 + a_2b_1)\mathrm{cov}(Y_1, Y_2) \qquad (٨١-١٠)\end{aligned}$$

وللتعميم إذا كانت:

$$L_1 = a_1Y_1 + \ldots + a_nY_n \quad ، \quad L_2 = b_1Y_1 + \ldots + b_nY_n$$

فإن:

$$\text{cov}(L_1, L_2) = \sum a_i b_i \sigma_i^2 + \sum_i \sum_{j>i} (a_i b_j + a_j b_i)\text{cov}(Y_i, Y_j) \qquad (٨٢-١٠)$$

١٠-١٩ ارتباط الرتب Rank correlation

فى حالة المتغيرات التى لا تتبع توزيع معين أو التى تتبع توزيعاً غير معروف المعالم (الثوابت) فإن حساب r كتقدير لمعامل الارتباط فى العشيرة يكون غير صالح وخاصة عندما يكون التوزيع ذو المتغيرين بعيداً عن التوزيع الطبيعى ولذلك فإن إجراء تحوير للمتغيرين أملاً فى أن يكون توزيعهما المشترك وثيق الشبه للتوزيع الطبيعى ذى المتغيرين يجعل تقدير ρ ممكنا فى بعض الأحيان وغير ممكن فى البعض الآخر. ولكن هل المتغيران مرتبطان وهل يتغيران فى نفس الاتجاه أم فى اتجاهين متضادين؟ وقد تم فيما سبق إيضاح أنه فى حالة اختبار فرض العدم بأن معامل الارتباط فى العشيرة يساوى صفراً فإنه يمكن استخدام r على أن يتوزع أحد المتغيرين طبيعياً، أما عندما يكون توزيع المتغيرين غير طبيعى فإن أحسن طريقة للإجابة على السؤال السابق هو ترتيب كل من المتغيرين (تنازلياً وتصاعدياً) ثم يختبر التوافق بين الترتيبين. وفى حالة البيانات غير المرتبة فإن أول خطوة هى ترتيب كل من المتغيرين كل منهما على حده ويستخدم معامل ارتباط الرتب r_s والذى وضعه Spearman حيث:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \qquad (٨٣-١٠)$$

حيث $d_i = Y_{1_i} - Y_{2_i}$ ، Y_{1_i} ، Y_{2_i} هما رتبة الفرد i فى كل من المتغير الأول والثانى على التوالى. فإذا كانت Σd_i^2 تساوى صفراً فإن $r_s = \pm 1$.

ومعامل ارتباط الرتب تنحصر قيمته بين 1− فى حالة عدم التوافق المطلق discordance و 1+ فى حالة التوافق التام complete concordance. ولاختبار معنوية معامل ارتباط الرتب للعينات التى حجمها عشرة أزواج أو أقل يستخدم الحد الأدنى لقيمة r_s المعنوية على مستوى 5% ومستوى 1% كما وصفه Kendall (1970) فى جدول ١٠-٧، حيث إن توزيع r وتوزيع r_s متساويان فى حالة اختبار فرض العدم $r = 0$.

ويلاحظ أنه عندما يكون حجم العينة 4 مثلاً فإن الحد الأدنى لمعنوية r_s يكون أكبر من الواحد الصحيح عند مستوى 5% وأيضاً 1% وبالتالى فإن معنوية معامل

ارتباط الرتب فى مثل هذه الحالة لن تتحقق حيث إن الحد الأعلى له يساوى الواحد الصحيح.

ويستخدم معامل ارتباط الرتب فى الحالات التى يتعذر فيها قياس المتغيرات بطريقة كمية وإنما تعطى رتباً أو درجات، ولقد استنبط Kendall مقياساً آخر لدرجة التوافق بين المتغيرين لن يتم تناوله هنا فى الوقت الحالى.

جدول ١٠-٧ معنوية معامل ارتباط الرتب للعينات التى حجمها عشرة أزواج أو أقل

حجم العينة	الحد الأدنى لمعنوية r_s	
	5%	1%
٤ أو أقل	-	-
٥	1.000	-
٦	0.886	1.000
٧	0.786	0.929
٨	0.738	0.857
٩	0.683	0.817
١٠	0.648	0.781
١١ أو أكثر	يستخدم جدول ٧ ملحق أ بدرجات حرية $n-2$	

مثال ١٠-٢٢

حساب معامل ارتباط الرتب واختبار معنويته.

رقم البقرة	الترتيب حسب كمية اللبن	الترتيب حسب المطابقة للنموذج	الفرق d_i	d_i^2
١	3	2	1	1
٢	4	3	1	1
٣	1	1	0	0
٤	2	5	-3	9
٥	5	4	1	1
المجموع			**0**	**12**

بتطبيق المعادلة (١٠-٨٣) فإن

$$r_s = 1 - \frac{(6)(12)}{(5)(25-1)} = 0.4$$

ومن القيمة الجدولية (حجم العينة = 5) يتضح أن معامل ارتباط الرتب غير معنوى أى ليس هناك علاقة بين الترتيبين.

تمارين الباب العاشر

١٠-١ في عينة ما وجدت القيم التالية للعلاقة بين الوزن Y_2 بالكيلوجرام والارتفاع Y_1 بالسنتيمتر:

$$n = 12 \qquad \sum Y_1 = 228 \qquad \sum Y_2 = 450$$

$$\sum y_1 y_2 = -542.5 \qquad \sum y_1^2 = 961 \qquad \sum y_2^2 = 1225$$

المطلوب حساب:

١ - معامل انحدار الوزن على الارتفاع ومعامل انحدار الارتفاع على الوزن.

٢ - معامل الارتباط.

٣ - تحليل التباين في الوزن إلى مكوناته.

٤ - ايجاد قيمة $\hat{Y}_2$ عندما تكون Y_1 تساوى 30, 19, 22 سنتيمتر.

٥ - تمثيل العلاقة بين المتغيرين بيانياً.

١٠-٢ في التمرين السابق اختبر الفرض بأن معامل انحدار الوزن على الارتفاع يساوى صفراً بمستوى معنوية 1% .

١٠-٣ في التمرين ١٠-١ اختبر الفروض الآتية:

١ - $H_0 : \rho = 0$

٢ - $H_0 : \rho = -3$

١٠-٤ في تجربة تسمين كان يتم وزن 20 حمل أسبوعياً فإذا كان رمز الوزن Y والأسابيع X وتوافرت البيانات التالية:

$$\sum Y = 400 \qquad \sum Y^2 = 9600 \qquad \sum XY = 5160$$

$$\sum X = 240 \qquad \sum X^2 = 3024$$

المطلوب حساب:

١ - معامل الانحدار b_{yx} ومعامل الارتباط بين المتغيرين.

٢- معامل التحديد.

٣- اختبر فرض العدم $H_0:\beta=0$ ، $H_0:\rho=0$. وضح جدول تحليل التباين.

٤- اكتب معادلة التنبؤ وارسم خط الانحدار رسماً دقيقاً.

١٠-٥ ما هي حدود الثقة لكل من الانحدار والجزء المقطوع من محور الصادات في تمرين ١٠-٤ وما هي حدود الثقة لخط الانحدار؟ بين ذلك بيانياً.

١٠-٦ إذا توافرت البيانات التالية:

العينة الأولى : معامل الارتباط = 0.325 حجم العينة = 25

العينة الثانية : معامل الارتباط = 0.34 حجم العينة = 15

العينة الثالثة : معامل الارتباط = 0.31 حجم العينة = 22

هل هذه العينات مسحوبة من عشيرة واحدة وما هو تقديرك لمعامل الارتباط؟ احسب الاختبارات الإحصائية اللازمة.

١٠-٧ إذا كانت البيانات التالية مسحوبة من عشيرة طبيعية ذات متغيرين

2.3	2.1	2.1	1.8	1.7	1.1	1.0	0.1	Y_1
2.3	2.8	2.8	1.8	3.0	1.5	2.7	0.7	Y_2
6.1	4.9	4.4	4.4	3.4	2.9	2.7	2.6	Y_1
6.6	4.0	3.8	2.7	3.2	2.6	4.4	2.7	Y_2

وكان معامل ارتباط العشيرة يساوى 0.7 ومعامل اعتماد العشيرة Y_2 على Y_1 يساوى 1

المطلوب حساب:

١- معامل الإرتباط بين المتغيرين ومعامل اعتماد Y_2 على Y_1 وارسم خط الانحدار رسماً دقيقاً.

٢- معامل التحديد.

٣- اختر أقل أربع قيم للمتغير Y_1 وقيم Y_2 المقابلة لها وأيضاً أكبر أربع قيم للمتغير Y_1 وقيم Y_2 المقابلة لتكوين عينة من 8 أزواج لهذه العينة. احسب المطلوب فى ١- وقارن النتائج المتحصل عليها.

٤- قسم كل من قيم Y_2 إلى مكوناتها ثم احسب $\Sigma e_1 Y_1$، $\Sigma e_1 Y_2$ وقارن بين $\Sigma e_1 Y_1$، Σe_1^2. اشرح معنى هذه النتائج.

١٠-٨ بين كيف أن تباين معامل الارتباط $= \frac{1-r^2}{n-2}$

حيث r معامل الارتباط و n حجم العينة (عدد أزواج المشاهدات).

١٠-٩ قام اثنان من المحكمين بترتيب 10 بقرات حسب حالتها الجسمية كما يلى:

رقم البقرة	المحكم الأول	المحكم الثانى
١	8	10
٢	9	8
٣	7	9
٤	5	3
٥	1	4
٦	4	2
٧	2	1
٨	3	5
٩	6	6
١٠	10	7

احسب معامل ارتباط الرتب واختبر معنويته.