جامعة الملك سعود
قسم الإحصاء وبحوث العمليات

تحليل الإنحدار الخطي

LINEAR REGRESSION ANALYSIS

بإستخدام

EXCEL DATA ANALYSIS TOOL, MINITAB AND THE MATHEMATICAL MODELING LANGUAGE R

تأليف
د. عدنان ماجد عبد الرحمن بري
استاذ الإحصاء وبحوث العمليات المشارك

نماذج الإنحدار الخطي

تعاريف أساسية:

(1) **المصفوفة *Matrix*:**

المصفوفة **A** هي عبارة عن عناصر مرتبة في قائمة مستطيلة الشكل مثل

$$\mathbf{A} = \begin{pmatrix} 3 & 5 & 4 \\ 1 & 2 & 8 \end{pmatrix}$$

العنصر (i, j) للمصفوفة **A** يرمز له a_{ij} .

بعد *Dimension* أو حجم المصفوفة هو n (عدد الأسطر) بـ m (عدد الأعمدة). إذا كان $n = m$ فتسمى **A** مصفوفة مربعة. لكي نبين أبعاد المصفوفة نكتبها $\mathbf{A}_{n\times m}$.

(2) **المتجه *Vector*:**

المتجة هو مصفوفة تتكون من عمود واحد أو سطر واحد. يكتب متجه عمود على الشكل $\mathbf{a}_{n\times 1}$ و متجه سطر على الشكل $\mathbf{a}_{1\times n}$. سوف نفترض عند التكلم عن متجه انه متجه عمود مالم نذكر غير ذلك.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \qquad \mathbf{a}' = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}$$

بعض المتجهات الخاصة:

$$1\text{-}\ \mathbf{i}_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}, 1\ in\ the\ i\ th\ cell$$

$$2\text{-}\ \mathbf{j} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}, all\ cells\ 1$$

(3) منقول المصفوفة *Transpose*:

إذا كانت $\mathbf{A} = (a_{ij})$ مصفوفة ذات أبعاد $n \times m$ فإن منقول $\mathbf{A}$ ويرمزله بالرمز $\mathbf{A}'$ هو المصفوفة ذات الأبعاد $m \times n$ والتي عناصرها a_{ji}.

(4) المصفوفة المتناظرة *Symmetric*:

إذا كان $\mathbf{A}' = \mathbf{A}$ فإنه يقال أن المصفوفة $\mathbf{A}$ متناظرة.

خواص المنقول:

(أ) $(\mathbf{A}')' = \mathbf{A}$

(ب) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$

(ج) لأي مصفوفة $\mathbf{A}$ فإن $\mathbf{A}'\mathbf{A}$ و $\mathbf{AA}'$ تكون متناظرة.

ملاحظة: في جميع النتائج المطروحة يفترض أن جميع المصفوفات متوافقة *Conformable*. أي أن جميع العمليات المصفوفية (مثل الجمع والضرب) لها معنى. أي إذا كانت $\mathbf{A}_{n\times m}$ و $\mathbf{B}_{m\times p}$ فإن $\mathbf{AB}$ هي مصفوفة ذات أبعاد $n \times p$.

(5) **مصفوفة الوحدة *Identity Matrix*:**

مصفوفة الوحدة $\mathbf{I}$ ذات الأبعاد $n \times n$ تكتب

$$\mathbf{I} = \mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n\times n}$$

أي أن $a_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$.

(6) **مصفوفة الآحاد *Ones Matrix*:**

مصفوفة الآحاد $\mathbf{J}$ ذات الأبعاد $n \times n$ تكتب

$$\mathbf{J} = \mathbf{J}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}_{n\times n}$$

أي أن $a_{ij} = 1$ لجميع قيم i و j.

لاحظ أن

$$\mathbf{J} = \mathbf{j}\mathbf{j}'$$

حيث

$$\mathbf{j} = \mathbf{j}_n$$

هو متجه عمود ذا البعد $n \times 1$ وجميع عناصره الواحد الصحيح.

(7) **المصفوفة الصفرية *Null Matrix*:**

مصفوفة الأصفار **0** ذات الأبعاد $n \times n$ تكتب

$$\mathbf{0} = \mathbf{0}_n = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}_{n \times n}$$

أي أن $a_{ij} = 0$ لجميع قيم i و j.

(8) **المقلوب *Inverse*:**

إذا كانت **A** مصفوفة ذات بعد $n \times n$ وكان يوجد مصفوفة **C** بحيث

$$\mathbf{AC} = \mathbf{CA} = \mathbf{I}$$

فإن **A** مصفوفة غير شاذة *Nonsingular* والمصفوفة **C** تسمى مقلوب أو معكوس **A** ويرمز لها $\mathbf{A}^{-1}$. إذا كانت **A** غير شاذة فإن المقلوب $\mathbf{A}^{-1}$ يكون وحيد.

خواص المقلوب:

(أ) إذا كان كل من المصفوفات **A** و **B** غير شاذ فإن

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

(ب) إذا كانت **A** غير شاذة فإن

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$

(9) الإستقلال الخطي والرتبة *Linear Independence and Rank*:

المتجهات $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n$ ذات الأبعاد $n \times 1$ يقال انها غير مستقلة خطيا إذا وفقط إذا كان يوجد أعداد $c_1, c_2, ..., c_n$ بحيث

$$\sum_{i=1}^{m} c_i \mathbf{a}_i = 0$$

وعلى الأقل واحد من الأعداد c لايساوي الصفر أو

$$\sum_{i=1}^{m} c_i \mathbf{a}_i = 0 \Rightarrow c_1 = c_2 = \cdots = c_m = 0$$

فإن $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n$ تكون مستقلة خطيا.

ملاحظة:

إذا كانت $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n$ تكون أعمدة المصفوفة $\mathbf{A}$ ذات البعد $n \times m$ أي

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m)$$

فإن أعمدة المصفوفة $\mathbf{A}$ مستقلة خطيا إذا وفقط إذا :ان $\mathbf{Ac} = \mathbf{0} \Rightarrow \mathbf{c} = \mathbf{0}$ حيث $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_m)'$.

(10) رتبة المصفوفة $\mathbf{A}$ *Rank of Matrix*:

رتبة المصفوفة $\mathbf{A}$ تعرف كالتالي:

rank($\mathbf{A}$) = عدد أعمدة $\mathbf{A}$ المستقلة خطيا.

rank($\mathbf{A}$) = عدد أسطر $\mathbf{A}$ المستقلة خطيا.

نتيجة:

في أي مصفوفة عدد الأعمدة المستقلة خطيا يساوي عدد الأسطر المستقلة خطيا.

(11) لنفرض أن **A** مصفوفة ذات بعد $n \times p$ فإن

$$rank(\mathbf{A}) \leq min\{n,p\}$$

إذا كان

$$rank(\mathbf{A}) = min\{n,p\}$$

فإنه يقال أن المصفوفة **A** ذات مرتبة كاملة Full Rank .

إذا كان

$$rank(\mathbf{A}) = n$$

فإنا نقول أن المصفوفة ذات مرتبة عامود كاملة وإذا كانت

$$rank(\mathbf{A}) = p$$

فإنا نقول أن المصفوفة ذات مرتبة سطر كاملة.

خواص الرتب:

(أ) لأي مصفوفة **A** فإن

$$rank(\mathbf{A}) = rank\ (\ \mathbf{A}^{-1})$$

(ب)لأي مصفوفة **A** فإن

$$rank\ (\mathbf{A}'\mathbf{A}) = rank(\mathbf{A})$$

(ج) للمصفوفات المتوافقة فإن

$$rank(\mathbf{AB}) \leq\ rank(\mathbf{A})$$

و

$$rank(\mathbf{AB}) \leq\ rank(\mathbf{B})$$

(د) إذا كانت **B** غير شاذة فإن

$$rank(\mathbf{AB}) = rank(\mathbf{A})$$

(هـ) لأي مصفوفة **A** ذات بعد $n \times n$ فإن

$$rank(\mathbf{A}) = \mathrm{n} \Leftrightarrow \quad \mathbf{A}^{-1} \; exists \;\; (\text{موجودة})$$

(و) لأي مصفوفة $\mathbf{A}_{n\times n}$ ومتجه $\mathbf{b}_{n\times 1}$

$$rank(\mathbf{A},\mathbf{b}) \geq \; rank(\mathbf{A})$$

تعريف آخر للرتبة:

أي مصفوفة **A** ذات بعد $m \times n$ لها رتبة r إذا كانت أبعاد اكبر تحت مصفوفة غير شاذة من **A** هو $r \times r$.

(10) **أثارة المصفوفة *Trace*:**

أثارة المصفوفة المربعة $\mathbf{A} = \left(a_{ij}\right)_{n\times n}$ هو مجموع عناصر المحور الرئيسي أي

$$tr\left(\mathbf{A}\right) = \sum_{i=1}^{n} a_{ii}$$

(أ) إذا كانت $\mathbf{A}_{m\times n}$ و $\mathbf{B}_{n\times m}$ فإن $tr\left(\mathbf{AB}\right)=tr\left(\mathbf{BA}\right)$ (لاحظ أن $\mathbf{AB}$ و $\mathbf{BA}$ مربعين)

(ب) إذا كانت $\mathbf{A}_{n\times n}$ و $\mathbf{B}_{n\times n}$ فإن $tr\left(\mathbf{A}+\mathbf{B}\right)=tr\left(\mathbf{A}\right)+tr\left(\mathbf{B}\right)$

(ج) $tr\left(\lambda\mathbf{A}\right)=\lambda tr\left(\mathbf{A}\right)$

نماذج الإنحدار الخطي *Linear Regression Models* :

وهي النماذج التي تكون خطية في معالمها . الشكل العام لنموذج الإنحدار الخطي هو

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

حيث $\mathbf{y}$ متجه $n \times 1$ من الإستجابات *Responses* المشاهدة و $\mathbf{X}$ مصفوفة ذات بعد $n \times p$ من القيم الثابتة و $\boldsymbol{\beta}$ متجه $p \times 1$ من المعالم *Parameters* الثابتة غير المعلومة و $\boldsymbol{\varepsilon}$ متجة $n \times 1$ من (غير مشاهد *Unobserved*) الأخطاء العشوائية بمتوسط صفري.

يسمى النموذج خطي لأن متوسط متجه الإستجابة $\mathbf{y}$ خطي في المعالم المجهولة $\boldsymbol{\beta}$.

سوف يتركز إهتمامنا على تقدير هذه المعالم المجهولة وإختبار فرضيات لأي تشكيلة خطية من هذه المعالم.

أمثلة على النماذج الخطية:

مثال 1:

الإنحدار الخطي البسيط *Simple Linear Regression*:

لنعتبر النموذج الذي يكون فيه متغير الإستجابة y متصل خطيا بمتغير مستقل x بالعلاقة:

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

لقيم $i = 1,2,...,n$ حيث ε_i تعتبر دوما متغيرات عشوائية غير مترابطة بمتوسط صفر وتباين σ^2. بعض الأحيان تعتبر الأخطاء موزعة طبيعيا لكي نتمكن من إستخراج إستدلالات حول معالم الإنحدار β_0 و β_1. إذا افترضنا أن $x_1, x_2, ...x_n$ ثوابت محددة

مقاسة بدون أخطاء فعندئذ هذا النموذج يكون حالة خاصة من النموذج $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

حيث:

$$\mathbf{y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

ملاحظة:

إذا قيست x بأخطاء فإن نموذج الإنحدار الخطي لن يعتبر حالة خاصة من النموذج $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ لأن مصفوفة التصميم (*Design Matrix*) $\mathbf{X}$ عشوائية وليست ثابتة.

مثال 2:

الإنحدار الخطي المتعدد *Multiple Linear Regression*:

لنعتبر النموذج الذي يعتمد فيه متغير الإستجابة $\mathbf{y}$ خطيا على عدة متغيرات مستقلة $\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_k$ كما في العلاقة:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

لقيم $i = 1,2,...,n$ حيث ε_i تعتبر دوما متغيرات عشوائية غير مترابطة بمتوسط صفر وتباين σ^2. إذا كانت المتغيرات المستقلة ثوابت محددة أي مقاسة بدون أخطاء فعندئذ هذا النموذج يكون حالة خاصة من النموذج $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ حيث:

$$\mathbf{y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X}_{n\times(k+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta}_{(k+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

مثال:

في دراسة على تذوق جبنة الشدر حللت $n = 30$ عينة لمحتواها الكيميائي واخضعت لإختبارات تذوق وجمعت بيانات عن المتغيرات التالية:

y = نتيجة إختبار التذوق (*TASTE*).

x_1 = تركيز حامض الخليك (*ACETIC*).

x_2 = تركيز سولفيت الهيدروجين (*H2S*).

x_3 = تركيز الحامض اللبني (*LACTIC*).

ووضعت في الجدول التالي:

TASTE	ACETIC	H2S	LACTIC	TASTE	ACETIC	H2S	LACTIC
12.3	4.543	3.135	0.86	40.9	6.365	9.588	1.74
20.9	5.159	5.043	1.53	15.9	4.787	3.912	1.16
39.0	5.366	5.438	1.57	6.4	5.412	4.700	1.49
47.9	5.759	7.496	1.81	18.0	5.247	6.174	1.63
5.6	4.663	3.807	0.99	38.9	5.438	9.064	1.99
25.9	5.697	7.601	1.09	14.0	4.564	4.949	1.15
37.3	5.892	8.726	1.29	15.2	5.298	5.220	1.33
21.9	6.078	7.966	1.78	32.0	5.455	9.242	1.44
18.1	4.898	3.850	1.29	56.7	5.855	10.20	2.01
21.0	5.242	4.174	1.58	16.8	5.366	3.664	1.31
34.9	5.740	6.142	1.68	11.6	6.043	3.219	1.46
57.2	6.446	7.908	1.90	26.5	6.458	6.962	1.72
0.7	4.477	2.996	1.06	0.7	5.328	3.912	1.25
25.9	5.236	4.942	1.30	13.4	5.802	6.685	1.08
54.9	6.151	6.752	1.52	5.5	6.176	4.787	1.25

المتغيرات *ACETIC* و *H2S* ممثلة بالتدريج اللوغارثمي الطبيعي scale *Natural log*.
المتغير *LACTIC* لم يتم تحويله. لنفترض أن الباحثين وضعوا فرضية أن كل من الثلاثة المحتويات الكيميائية المتغيرة x_1 و x_2 و x_3 مهمة في وصف التذوق. وفي هذه الحالة فقد اعتمدوا مبدئيا نموذج الإنحدار التالي:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

لقيم $i = 1,2,...,30$. في الصيغة المصفوفية $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ يكتب

$$\mathbf{y}_{30\times1}=\begin{pmatrix}12.3\\20.9\\\vdots\\5.5\end{pmatrix},\quad \mathbf{X}_{30\times4}=\begin{pmatrix}1&4.543&3.135&0.86\\1&5.159&5.043&1.53\\1&5.366&5.438&1.57\\\vdots&\vdots&\vdots&\vdots\\1&6.176&4.787&1.25\end{pmatrix},\quad \boldsymbol{\beta}_{4\times1}=\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\end{pmatrix},$$

$\boldsymbol{\varepsilon}_{30\times1}=(\varepsilon_1,\varepsilon_2,...,\varepsilon_{30})'$.

وصف رياضي:

النموذج

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}+\varepsilon_i$$

بمعالم إنحدار $\beta_0,\beta_1,\beta_2,\cdots,\beta_k$ يسمى نموذج إنحدار خطي. كلمة " خطي " تشير إلى كيفية وجود كل حد من حدود المعادلة ولا يعنى شكل دالة الإنحدار

$$g(x_1,x_2,...,x_k)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}$$

وبدقة أكثر عندما نقول أن نموذج " نموذج خطي " فإننا نعني أن دالة الإنحدار الحقيقية g *خطية في المعالم*. رياضيا هذا يعني أن المشتقات k+1 الجزئية

$$\frac{\partial g(\beta_0,\beta_1,...,\beta_k)}{\partial\beta_i},\quad i=1,2,...,k$$

تكون كلها خالية من المعالم $\beta_0,\beta_1,\beta_2,\cdots,\beta_k$ فمثلا كل من نماذج الإنحدار الخطي التالية يمكن وضعها في الصيغة $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$:

$$y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 x_i^2}_{g(x)} + \varepsilon_i$$

$$y_i = \underbrace{\beta_0 + \beta_1 \log x_{i1} + \beta_2 \sqrt{\cos x_{i2}}}_{g(x_1, x_2)} + \varepsilon_i$$

من السهل تبيين أن كل نموذج من النماذج أعلاه خطي في المعالم. من جهة اخرى النموذج اللجستي:

$$y_i = \underbrace{\frac{\beta_0}{1 + \beta_1 e^{\beta_2 x_i}}}_{g(x)} + \varepsilon_i$$

غير خطي لأن

$$\frac{\partial}{\partial \beta_0}\left(\frac{\beta_0}{1 + \beta_1 e^{\beta_2 x_i}}\right) = \frac{1}{1 + \beta_1 e^{\beta_2 x_i}}$$

وهو غير خالي من معالم الإنحدار ولذلك لايمكن وضعه على الشكل $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

المتجهات والمصفوفات العشوائية *Random Vectors and Matrices*:

لنعتبر نموذج الإنحدار الخطي العام

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

حيث $\mathbf{y}$ متجه $n \times 1$ من الإستجابات المشاهدة و $\mathbf{X}$ مصفوفة ذات بعد $n \times p$ من القيم الثابتة و $\boldsymbol{\beta}$ متجه $p \times 1$ من المعالم الثابتة غير المعلومة و $\boldsymbol{\varepsilon}$ متجة $n \times 1$ من (غير مشاهد) الأخطاء العشوائية بمتوسط صفري. في هذا النموذج الكميات $\boldsymbol{\varepsilon}$ و $\mathbf{y}$ متجهات عشوائية (أي متجهات من المتغيرات العشوائية). متجه الخطأ $\boldsymbol{\varepsilon}$ يعتبر عشوائي حسب إفتراضنا. وحيث أن $\mathbf{y}$ يعتمد على $\boldsymbol{\varepsilon}$ فهو أيضا عشوائي. في معظم المشاكل التي تصادفنا مصفوفة التصميم $\mathbf{X}$ تعتبر ثابتة (غير عشوائية) ولكنه يمكن إعتبارها عشوائية في بعض المواقف (وسوف نناقش ذلك حينها). في كلا الحالتين سوف نستعرض هنا خواص المتجهات والمصفوفات العشوائية.

المتوسطات والتباينات:

مصطلح: لنفترض أن $y_1, y_2, ..., y_n$ متغيرات عشوائية. سوف نسمي

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

متجه عشوائي. دالة الكثافة الإحتمالية متعددة المتغيرات *Multivaiate pdf* لـ y يرمز لها $f_{\mathbf{Y}}(\mathbf{y})$. وهي تصف بشكل رياضي كيفية توزيع المتغيرات $y_1, y_2, ..., y_n$ مشتركة وهكذا تسمى أحيانا التوزيع المشترك *Joint Distribution*. إذا كانت $y_1, y_2, ..., y_n$ كل منها لها توزيع *iid* من $f_Y(y)$ فدالة التوزيع المشترك تعطى بالعلاقة

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i)$$

إذا كانت y_i مستقلة وتأتي من $f_{Y_i}(y_i)$ فإن دالة الكثافة المشتركة لـ y هي

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

لنفترض أن $E(y_i) = \mu_i$ و $V(y_i) = \sigma_i^2$ لقيم $i = 1,2,...,n$ و $Cov(y_i, y_j) = \sigma_{ij}$ حيث $i \neq j$. يعرف متوسط المتجه العشوائي $\mathbf{y}$ كالتالي:

$$\boldsymbol{\mu} = E(\mathbf{y}) = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

وتباين المتجه العشوائي هو المصفوفة $n \times n$

$$\boldsymbol{\Sigma} = Cov\left(\mathbf{y}\right) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

ملاحظة: تسمي المصفوفة $\boldsymbol{\Sigma}$ أيضا مصفوفة التباين - التغاير *Variance-Covariance Matrix* للمتغير العشوائي $\mathbf{y}$ لأنها تحوي التباينات $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ على المحور الرئيسي و $\binom{n}{2}$ من التغايرات $Cov\left(y_i, y_j\right)$ لقيم $i < j$ كعناصر فوق المحور وبما أن $Cov\left(y_i, y_j\right) = Cov\left(y_j, y_i\right)$ فإن مصفوفة التباين - التغاير $\boldsymbol{\Sigma}$ متناظرة.

مثال:

لنفترض أن $y_1, y_2, ..., y_n$ عينة *iid* بمتوسط $E\left(y_i\right) = \mu$ وتباين $V\left(y_i\right) = \sigma^2$ ولتكن $\mathbf{y} = \left(y_1, y_2, ..., y_n\right)'$. عندئذ

$$\boldsymbol{\mu} = E\left(\mathbf{y}\right) = \mu \mathbf{j}_n$$

و

$$\boldsymbol{\Sigma} = Cov\left(\mathbf{y}\right) = \sigma^2 \mathbf{I}_n$$

مثال:

نموذج جاوس - ماركوف *Gauss-Markov Model*: في نموذج الإنحدار الخطي العام $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ الأخطاء العشوائية $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ يفترض غالبا أنها متغيرات عشوائية مستقلة ولها توزيع متطابق *iid* بمتوسط صفري وتباين ثابت σ^2 ففي هذه الحالة

$$E\left(\boldsymbol{\varepsilon}\right)=\mathbf{0}_{n\times 1}$$

و

$$.Cov\left(\boldsymbol{\varepsilon}\right)=\sigma^{2}\mathbf{I}_{n}$$

مصطلح: لنفترض أن $z_{11},z_{12},...,z_{np}$ متغيرات عشوائية المصفوفة

$$\mathbf{Z}_{n\times p}=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix}$$

تسمى مصفوفة عشوائية. متوسط المصفوفة العشوائية $\mathbf{Z}$ يعرف كالتالي:

$$E\left(\mathbf{Z}\right)_{n\times p}=\begin{pmatrix} E\left(z_{11}\right) & E\left(z_{12}\right) & \cdots & E\left(z_{1p}\right) \\ E\left(z_{21}\right) & E\left(z_{22}\right) & \cdots & E\left(z_{2p}\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(z_{21}\right) & E\left(z_{n2}\right) & \cdots & E\left(z_{np}\right) \end{pmatrix}$$

نتيجة: لنفترض أن $\mathbf{y}$ متجه عشوائي بمتوسط $\boldsymbol{\mu}$ عندئذ

$$\boldsymbol{\Sigma}=Cov\left(\mathbf{y}\right)=E\left[\left(\mathbf{y}-\boldsymbol{\mu}\right)\left(\mathbf{y}-\boldsymbol{\mu}\right)'\right]=E\left(\mathbf{y}\mathbf{y}'\right)-\boldsymbol{\mu}\boldsymbol{\mu}'$$

يترك البرهان للطالب.

التغاير بين متجهين عشوائيين:

لنفترض أن $\mathbf{y}_{p\times 1}$ و $\mathbf{x}_{q\times 1}$ متجهات عشوائية بمتوسطات $\boldsymbol{\mu}_{\mathbf{Y}}$ و $\boldsymbol{\mu}_{\mathbf{X}}$ على التوالي. التغاير بين $\mathbf{y}$ و $\mathbf{x}$ هو المصفوفة $p\times q$ المعرفة كالتالي:

$$Cov\left(\mathbf{y},\mathbf{x}\right)=E\left[\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}\right)\left(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}\right)'\right]=\left(\sigma_{ij}\right)_{p\times q}$$

حيث

$$\sigma_{ij}=E\left\{\left[y_i-E\left(y_i\right)\right]\left[x_i-E\left(x_i\right)\right]\right\}=Cov\left(y_i,x_i\right)$$

حقائق:

$$Cov\left(\mathbf{y},\mathbf{y}\right)=\boldsymbol{\Sigma}=Cov\left(\mathbf{y}\right)$$

$$Cov\left(\mathbf{y},\mathbf{x}\right)=\left[Cov\left(\mathbf{x},\mathbf{y}\right)\right]'$$

مصطلح:

يقال أن المتجهات العشوائية $\mathbf{y}_{p\times 1}$ و $\mathbf{x}_{q\times 1}$ غير مترابطة *Uncorrelated* إذا كان

$$Cov\left(\mathbf{y},\mathbf{x}\right)=\mathbf{0}_{p\times q}$$

حقيقة:

إذا كان $Cov\left(\mathbf{y},\mathbf{x}\right)=\mathbf{0}_{p\times q}$ عندئذ

$$Cov\left(\mathbf{y},\mathbf{a}+\mathbf{B}\mathbf{x}\right)=\mathbf{0}$$

لجميع $\mathbf{a}$ و $\mathbf{B}$ غيرعشوائية ومتوافقة. (يترك برهانها كتمرين).

مصفوفات الترابط *Correlation Matrices*:

لنفترض أن $V\left(y_i\right)=\sigma_i^2$ لقيم $i=1,2,...,n$ و $Cov\left(y_i,y_j\right)=\sigma_{ij}$ حيث $i\neq j$.

مصفوفة الترابط لـ **y** هي

$$\mathbf{P}_{n\times n}=\left(\rho_{ij}\right)=\begin{pmatrix}1 & \rho_{12} & \cdots & \rho_{1n}\\ \rho_{21} & 1 & \cdots & \rho_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{n1} & \rho_{n2} & \cdots & 1\end{pmatrix}$$

حيث $\rho_{ij}=\dfrac{\sigma_{ij}}{\sigma_i\sigma_j}$.

ملاحظة: إذا عرفنا

$$\mathbf{D}_{n\times n}=diag\left(\sigma_1,\sigma_2,...,\sigma_n\right)\equiv\begin{pmatrix}\sigma_1 & 0 & \cdots & 0\\ 0 & \sigma_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_n\end{pmatrix}$$

فإن

$$\mathbf{P}=\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}$$

و

$$\mathbf{\Sigma}=\mathbf{DPD}$$

أي يمكن الحصول على كلا من مصفوفتي التغاير والترابط من بعضهما البعض.

التحويلات الخطية *Linear Transformations*:

لنفترض أن $y_1, y_2, ..., y_n$ متغيرات عشوائية وأن $a_1, a_2, ..., a_n$ ثوابت. عرف $\mathbf{a} = (a_1, a_2, ..., a_n)'$ و $\mathbf{y} = (y_1, y_2, ..., y_n)'$. الدالة

$$\mathbf{a}'\mathbf{y} = \sum_{i=1}^{n} a_i y_i$$

تسمى تركيب خطي *Linear Combination* للمتغيرات العشوائية $y_1, y_2, ..., y_n$.

متوسطات التراكيب الخطية *Means of Linear Combinations*:

إذا كان $\mathbf{a} = (a_1, a_2, ..., a_n)'$ متجه من الثوابت و $\mathbf{y} = (y_1, y_2, ..., y_n)'$ متجه عشوائي بمتوسط $\boldsymbol{\mu}$ عندئذ

$$E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\mu}$$

تمرين: برهن النتيجة السابقة.

حقائق: لنفترض أن $\mathbf{y} = (y_1, y_2, ..., y_n)'$ متجه عشوائي بمتوسط $\boldsymbol{\mu}$ ولتكن $\mathbf{X}$ مصفوفة عشوائية وليكن $\mathbf{A}$ و $\mathbf{B}$ و $\mathbf{a}$ و $\mathbf{b}$ مصفوفات ومتجهات متوافقة وغير عشوائية عندئذ:

$$E(\mathbf{A}\mathbf{y}) = \mathbf{A}\boldsymbol{\mu}$$
$$E(\mathbf{a}'\mathbf{X}\mathbf{b}) = \mathbf{a}'E(\mathbf{X})\mathbf{b}$$
$$E(\mathbf{A}\mathbf{X}\mathbf{B}) = \mathbf{A}E(\mathbf{X})\mathbf{B}$$

تباينات التراكيب الخطية *Variances of Linear Combinations*:

إذا كان $\mathbf{a}=(a_1,a_2,...,a_n)'$ متجه من الثوابت و $\mathbf{y}=(y_1,y_2,...,y_n)'$ متجه عشوائي بمتوسط $\boldsymbol{\mu}$ ومصفوفة تباين - تغاير $\boldsymbol{\Sigma}$ عندئذ:

$$V(\mathbf{a}'\mathbf{y})=\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$$

تمرين: برهن النتيجة السابقة.

نتيجة: لنفترض أن $\mathbf{y}=(y_1,y_2,...,y_n)'$ متجه عشوائي بمصفوفة تباين - تغاير $\boldsymbol{\Sigma}$ وليكن $\mathbf{a}$ و $\mathbf{b}$ متجهين متوافقين من الثوابت عندئذ

$$Cov(\mathbf{a}'\mathbf{y},\mathbf{b}'\mathbf{y})=\mathbf{a}'\boldsymbol{\Sigma}\mathbf{b}$$

تمرين: برهن النتيجة السابقة.

حقائق: : لنفترض أن $\mathbf{y}=(y_1,y_2,...,y_n)'$ متجه عشوائي بمصفوفة تباين - تغاير $\boldsymbol{\Sigma}$ وليكن $\mathbf{A}$ و $\mathbf{B}$ مصفوفات متوافقة وغير عشوائية عندئذ

$$Cov(\mathbf{A}\mathbf{y})=\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$
$$Cov(\mathbf{A}\mathbf{y},\mathbf{B}\mathbf{y})=\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}$$

نتيجة: إذا كان $\mathbf{y} = (y_1, y_2, ..., y_n)'$ متجه عشوائي بمتوسط $\boldsymbol{\mu}$ و $\mathbf{A}$ و $\mathbf{B}$ مصفوفات متوافقة وغير عشوائية عندئذ

$$E(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$
$$Cov(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$

خواص مصفوفات التباين ـ التغاير:

تعريف: المصفوفة المتناظرة $\mathbf{C}$ ذات أبعاد $n \times n$ يقال أنها موجبة المحدودية *Positive Definite* إذا حققت

$$\mathbf{a}'\mathbf{Ca} > 0$$

لجميع

$$\mathbf{a} \in R^n - \{0\}$$

ونقول أن $\mathbf{C}$ غير سالبة المحدودية *Non-negative Definite* إذا حققت

$$\mathbf{a}'\mathbf{Ca} \geq 0$$

لجميع

$$\mathbf{a} \in R^n$$

التعبير $\mathbf{a}'\mathbf{Ca}$ يسمى شكل رباعي *Quadratic Form*.

ملاحظة: المصفوفة غير سالبة المحدودية والتي ليست موجبة المحدودية يقال أنها موجبة نصف المحدودية *Positive Semidefinite*. أي أن **C** موجبة نصف المحدودية إذا كان

$$\mathbf{a}'\mathbf{C}\mathbf{a} \geq 0$$

لجميع

$$\mathbf{a} \in R^n$$

و

$$\mathbf{a}'\mathbf{C}\mathbf{a} = 0$$

لبعض

$$\mathbf{a} \neq \mathbf{0}$$

بأخذ مجموعتي المصفوفات معا أي موجبة المحدودية و موجبة نصف المحدودية نشكل مجموعة المصفوفات غير سالبة المحدودية *Non-negative Definite*.

نظرية: مصفوفة التباين - التغاير غيرسالبة المحدودية.

البرهان:

لنفترض أن المتجة العشوائي $\mathbf{y}_{n\times 1}$ له مصفوفة تباين - تغاير $\boldsymbol{\Sigma}$. نريد أن نبين أن

$$\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \geq 0$$

لجميع

$$\mathbf{a} \in R^n$$

لنعتبر

$$x = \mathbf{a}'\mathbf{y}$$

حيث **a** متجه من الثوابت ومتوافق. عندئذ فإن x متغير عشوائي عددي Scalar r.v. وبهذا فإن تباينه أكبر من أو يساوي الصفر أي

$$V\left(x\right)\geq 0$$

ولكن

$$V\left(x\right)=V\left(\mathbf{a}'\mathbf{y}\right)=\mathbf{a}'\mathbf{\Sigma a}$$

وحيث أن **a** إختياري فإن هذا يثبت النظرية.

حقيقة مهمة: إذا كانت **C** مصفوفة متناظرة ذات أبعاد $n\times n$ موجبة المحدودية فإنها لاتكون شاذة *Nonsingular*. أما إذا كانت موجبة نصف المحدودية فإنها شاذة *Singular*.

مجموع متجهات عشوائية:

ليكن **x** و **y** و **z** متجهات عشوائية ذات بعد $n\times 1$ ولنفترض أن $\mathbf{x}=\mathbf{y}+\mathbf{z}$ عندئذ

$$E\left(\mathbf{x}\right)=E\left(\mathbf{y}\right)+E\left(\mathbf{z}\right)$$

$$Cov\left(\mathbf{x}\right)=Cov\left(\mathbf{y}\right)+Cov\left(\mathbf{z}\right)+Cov\left(\mathbf{y},\mathbf{z}\right)+Cov\left(\mathbf{z},\mathbf{y}\right)$$

إذا كان **y** و **z** غير مترابطين فإن

$$Cov\left(\mathbf{x}\right)=Cov\left(\mathbf{y}\right)+Cov\left(\mathbf{z}\right)$$

التوزيع الطبيعي متعدد المتغيرات *Multivariate Normal Distribution*:

لنعتبر نموذج الإنحدار الخطي العام

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

حيث $\mathbf{y}$ متجه $n \times 1$ من الإستجابات المشاهدة و $\mathbf{X}$ مصفوفة ذات بعد $n \times p$ من القيم الثابتة و $\boldsymbol{\beta}$ متجه $p \times 1$ من المعالم الثابتة غير المعلومة و $\boldsymbol{\varepsilon}$ متجة $n \times 1$ من (غير مشاهد) الأخطاء العشوائية بمتوسط صفري. نموذج جاوس - ماركوف تحدد أيضا أن

$$E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}_{n\times 1}$$

و

$$Cov\left(\boldsymbol{\varepsilon}\right) = \sigma^2 \mathbf{I}_n$$

فرضية إضافية: بالإضافة للفرضية حول متوسط وتباين متجه الخطأ فإننا نفترض أيضا أن $\boldsymbol{\varepsilon}$ يتبع توزيع طبيعي متعدد المتغيرات.

لنفترض أن $z_1, z_2, ..., z_p$ هي *iid* بتوزيع طبيعي قياسي. دالة التوزيع المشترك للمتجه العشوائي $\mathbf{z} = \left(z_1, z_2, ..., z_p\right)'$ يعطي بالعلاقة

$$f_Z(\mathbf{z}) = \prod_{i=1}^{p} f_Z(z_i)$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\sum_i z_i^2/2} I(z_i \in R)$$
$$= (2\pi)^{-p/2} exp(-\mathbf{z}'\mathbf{z}/2) I(z_i \in R^p)$$

حيث $I(x \in S)$ دالة المؤشر وتعني

$$I(x \in S) = \begin{cases} 1, & x \in S \\ 0, & x \notin S \end{cases}$$

إذا كانت $\mathbf{z}$ لها دالة كثافة $f_Z(\mathbf{z})$ فإننا نقول أن $\mathbf{z}$ لها توزيع طبيعي قياسي متعدد المتغيرات. أي توزيع طبيعي متعدد المتغيرات بمتجه متوسط $\mathbf{0}_{p\times 1}$ ومصفوفة تغاير $\mathbf{I}_p$ ونكتب

$$.\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$$

في النموذج الخطي العام $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ يفترض عادة أن $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

تعريف:

الجزر التربيعي لمصفوفة متناظرة *Symmetric Square Root Matrix*:

إذا كانت $\mathbf{A}$ غير سالبة المحدودية فإنه بالإمكان إيجاد مصفوفة $\mathbf{A}^{1/2}$ والتي لها الخاصية:

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$$

$\mathbf{A}^{1/2}$ تسمي مصفوفة الجزر التربيعي المتناظرة.

خواص التوزيع الطبيعي متعدد المتغيرات:

لنفترض أن $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ وليكن $\mathbf{a}$ متجه $p \times 1$ و $\mathbf{b}$ متجه $k \times 1$ و مصفوفة $\mathbf{A}$ ببعد $k \times p$ عندئذ

$$z = \mathbf{a}'\mathbf{y} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$$
$$\mathbf{z} = \mathbf{A}\mathbf{y} \sim N_k(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$$
$$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b} \sim N_k(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$$

نتيجة: إذا كان $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ حيث $\boldsymbol{\Sigma}$موجبة المحدودية فإن

$$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I})$$

نتيجة: في النموذج الخطي العام $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ حيث $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. لاحظ أن $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ و $\boldsymbol{\Sigma} = Cov(\mathbf{y}) = \sigma^2\mathbf{I}$ وبما أن $\mathbf{y}$ هو تركيب خطي من $\boldsymbol{\varepsilon}$ فهو أيضا له توزيع طبيعي أي $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

نتيجة: وجدنا أن

$$\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

وأيضا

$$\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$$

و

$$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{M})\mathbf{y}$$

عندئذ

$$E(\hat{\mathbf{y}}) = E(\mathbf{My}) = \mathbf{M}E(\mathbf{y}) = \mathbf{MX\beta} = \mathbf{X\beta}$$

و

$$Cov(\hat{\mathbf{y}}) = Cov(\mathbf{My}) = \mathbf{M}Cov(\mathbf{y})\mathbf{M}' = \sigma^2\mathbf{MIM}' = \sigma^2\mathbf{M}$$

وبما أن $\hat{\mathbf{y}} = \mathbf{My}$ هي تركيب خطي من $\mathbf{y}$ فهي ايضا لها توزيع طبيعي أي

$$\hat{\mathbf{y}} \sim N_n(\mathbf{X\beta}, \sigma^2\mathbf{M})$$

تمرين: برهن أن $\hat{\mathbf{e}} \sim N_n\{\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{M})\}$.

توزيعات الأشكال التربيعية *Distributions of Quadratic Forms*:

لنفترض أن $\mathbf{y} = (y_1, y_2, ..., y_n)'$ متجه عشوائي وأن $\mathbf{A}$ مصفوفة مربعة متناظرة.

الكمية $\mathbf{y'Ay}$ تسمى شكل تربيعي.

لو كان متوسط $\mathbf{y}$ هو $\boldsymbol{\mu}$ ومصفوفة تغايره هي $\boldsymbol{\Sigma}$ فإن متوسط الشكل التربيعي $\mathbf{y'Ay}$ هو

$$E(\mathbf{y'Ay}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + tr(\mathbf{A}\boldsymbol{\Sigma})$$

وتباينه هو

$$V(\mathbf{y'Ay}) = 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} + 2tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma})$$

توزيع مربع كاي غير المركزي *Noncentral χ^2 Distribution* :

لنفترض أن $z \sim \chi^2(n)$ أي أن z له توزيع χ^2 (المركزي) بدرجات حرية $n > 0$.

دالة الكثافة الإحتمالية لـ z هي

$$f_Z(z|n) = \frac{1}{\Gamma\left(\frac{n}{2}\right)2^{n/2}} z^{\frac{n}{2}-1} e^{-z/2} I(z > 0)$$

ملاحظة: إذا كانت $z_1, z_2, ..., z_n$ لها توزيع $iid\ N(0,1)$ عندئذ

$$z_1^2 \sim \chi^2(1)$$

و

$$\mathbf{z}'\mathbf{z} = \sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$$

تعريف: المتغير العشوائي z يقال ان له توزيع مربع كاي غير مركزي بدرجات حرية $n > 0$ ومعلم غير مركزية $\lambda > 0$ إذا كانت له دالة كثافة

$$f_Z(z|n,\lambda) = \sum_{j=0}^{\infty} \frac{1}{\Gamma\left(\frac{n+2j}{2}\right) 2^{(n+2j)/2}} z^{\frac{n+2j}{2}-1} e^{-z/2} I(z>0)$$

ويكتب $z \sim \chi^2(n,\lambda)$. عندما تكون $\lambda = 0$ فإن توزيع مربع كاي غير المركزي يصبح مركزيا.

متوسط وتباين $z \sim \chi^2(n,\lambda)$ **:**

$$E(z) = n + 2\lambda$$

$$V(z) = 2n + 8\lambda$$

توزيع F **غير المركزي** ***Noncentral F Distribution*:**

المتغير العشوائي w يقال ان له توزيع F (المركزي) بدرجات حرية $n_1 > 0$ و $n_2 > 0$ إذا كانت دالة الكثافة له هي

$$f_W(w|n_1,n_2) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{n_1/2} w^{(n_1-2)/2}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left(1+\frac{n_1 w}{n_2}\right)^{(n_1+n_2)/2}} I(w>0)$$

ونكتب $w \sim F\left(n_1, n_2\right)$.

تعريف: المتغير العشوائي w يكون له توزيع F غير المركزي بدرجات حرية $n_1 > 0$ و $n_2 > 0$ ومعلم غير مركزية $\lambda > 0$ إذا كانت دالة الكثافة له هي

$$f_W\left(w \mid n_1, n_2, \lambda\right) = \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!} \frac{\Gamma\left(\frac{n_1 + 2j + n_2}{2}\right)\left(\frac{n_1 + 2j}{n_2}\right)^{(n_1+2j)/2} w^{(n_1+2j-2)/2}}{\Gamma\left(\frac{n_1 + 2j}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left(1 + \frac{n_1 w}{n_2}\right)^{(n_1+2j+n_2)/2}} I\left(w > 0\right)$$

ونكتب $w \sim F\left(n_1, n_2, \lambda\right)$ طبعا عندما $\lambda = 0$ فإن توزيع F غير المركزي يصبح مركزيا.

متوسط وتباين $w \sim F\left(n_1, n_2, \lambda\right)$:

$$E\left(w\right) = \frac{n_2}{n_2 - 2}\left(1 + \frac{2\lambda}{n_1}\right)$$

$$V\left(w\right) = \frac{2n_2^2}{n_1^2\left(n_2 - 2\right)}\left\{\frac{\left(n_1 + 2\lambda\right)^2}{\left(n_2 - 2\right)\left(n_2 - 4\right)} + \frac{n_1 + 4\lambda}{n_2 - 4}\right\}$$

المتوسط يكون موجودا فقط لقيم $n_2 > 2$ والتباين لقيم $n_2 > 4$.

تعريف: المصفوفة المثلية *Idempotent Matrix*:

يقال عن المصفوفة **A** انها مثلية إذا كانت تحقق

$$\mathbf{A}^2 = \mathbf{A}$$

ملاحظة: المصفوفة

$$\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

مصفوفة مثلية لأن

$$\begin{aligned}\mathbf{MM} &= \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\underbrace{\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{=\mathbf{I}}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I}\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{M}\end{aligned}$$

توزيع الشكل التربيعي $\mathbf{y}'\mathbf{A}\mathbf{y}$:

نظرية : لنفترض أن

$$\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{I})$$

إذا كانت $\mathbf{A}$ مصفوفة مثلية حيث

$$rank(\mathbf{A}) = s$$

عندئذ

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2(s, \lambda)$$

حيث

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

وعكسيا إذا كانت

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2(s, \lambda)$$

فإن $\mathbf{A}$ تكون مصفوفة مثلية برتبة s

و

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

البرهان:

سوف نبرهن الكفاية *Sufficiency* ($\Leftarrow$) الضرورة *Necessity* ($\Rightarrow$) برغم ان برهانه مشوق وممتع إلا انه قليل الأهمية. لنفترض أن

$$\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{I})$$

وأن $\mathbf{A}$ مثلية من الترتبة s.

يمكن وضع

$$\mathbf{A} = \mathbf{P}_1\mathbf{P}_1'$$

حيث $\mathbf{P}_1$ لها ابعاد $p \times s$ ولها الخاصية

$$\mathbf{P}_1'\mathbf{P}_1 = \mathbf{I}_s$$

(برهن هذا) وهكذا

$$\mathbf{y}'\mathbf{A}\mathbf{y} = \mathbf{y}\mathbf{P}_1\mathbf{P}_1'\mathbf{y} = \mathbf{x}'\mathbf{x}$$

حيث $\mathbf{x} = \mathbf{P}_1'\mathbf{y}$.

بما ان

$$\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{I})$$

و

$$\mathbf{x} = \mathbf{P}_1'\mathbf{y}$$

عبارة عن تركيب خطي من $\mathbf{y}$ فإن

$$\mathbf{x} \sim N_s\left(\mathbf{P}_1'\boldsymbol{\mu}, \mathbf{P}_1'\mathbf{I}\mathbf{P}_1\right) \sim N_s\left(\mathbf{P}_1'\boldsymbol{\mu}, \mathbf{I}_s\right)$$

كما ان

$$\mathbf{y}'\mathbf{A}\mathbf{y} = \mathbf{x}'\mathbf{x} \sim \chi^2\left\{s, \frac{1}{2}\left(\mathbf{P}_1'\boldsymbol{\mu}\right)' \mathbf{P}_1'\boldsymbol{\mu}\right\}$$

لاحظ ان

$$\lambda \equiv \frac{1}{2}\left(\mathbf{P}_1'\boldsymbol{\mu}\right)' \mathbf{P}_1'\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{P}_1\mathbf{P}_1'\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

وهو المطلوب.

نظرية : لنفترض أن

$$\mathbf{y} \sim N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

حيث

$$rank\left(\boldsymbol{\Sigma}\right) = p$$

إذا كانت $\mathbf{A\Sigma}$ مصفوفة مثلية حيث

$$rank\left(\mathbf{A}\boldsymbol{\Sigma}\right) = s$$

عندئذ

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2\left(s, \lambda\right)$$

حيث

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

وعكسيا إذا كانت

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2\left(s, \lambda\right)$$

فإن $\mathbf{A}$ تكون مصفوفة مثلية برتبة s و

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

البرهان: هذه النظرية تعميم للنظرية السابقة وبرهانها يتبع الخطوط العامة للبرهان السابق ويترك للطالب كتمرين. (تلميح: خذ $\mathbf{x} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$).

إستقلال الأشكال الرباعية:

نظرية: لنفترض أن

$$\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

إذا كان

$$\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$$

عندئذ

$$\mathbf{y}'\mathbf{A}\mathbf{y}$$

و

$$\mathbf{B}\mathbf{y}$$

مستقلين.

نظرية: لنفترض أن

$$\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

إذا كان

$$\mathbf{B\Sigma A = 0}$$

عندئذ

$$\mathbf{y'Ay}$$

و

$$\mathbf{y'By}$$

مستقلين.

تقدير معالم نموذج الإنحدار الخطي العام بطريقة المربعات الدنيا:

لنعتبر نموذج الإنحدار الخطي العام

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

حيث $\mathbf{y}$ متجه $n \times 1$ من الإستجابات *Responses* المشاهدة و $\mathbf{X}$ مصفوفة ذات بعد $n \times p$ من القيم الثابتة و $\boldsymbol{\beta}$ متجه $p \times 1$ من المعالم *Parameters* الثابتة غير المعلومة و $\boldsymbol{\varepsilon}$ متجة $n \times 1$ من (غير مشاهد *Unobserved*) الأخطاء العشوائية بمتوسط صفري.

يسمى النموذج غير خطي لأن متوسط متجه الإستجابة $\mathbf{y}$ خطي في المعالم المجهولة $\boldsymbol{\beta}$.

سوف يتركز إهتمامنا على تقدير هذه المعالم المجهولة وإختبار فرضيات لأي تشكيلة خطية من هذه المعالم.

مبدأ المربعات الدنيا يقول أوجد قيمة $\boldsymbol{\beta}$ التي تقلل *Minimizes* مجموع مربعات الخطأ

$$Q(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

أي

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

بإشتقاق $Q(\boldsymbol{\beta})$ بالنسبة للمعلم $\boldsymbol{\beta}$ ووضع المشتقة مساوية للصفر وحل المعادلات الطبيعية نجد

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \Rightarrow \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

وحيث أن $\boldsymbol{\beta}$ غير معروفة فإنها تقدر كالتالي:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

الآن إذا كانت $\mathbf{X}'\mathbf{X}$ غير شاذة فيكون

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

هو الحل الوحيد لمقدر المربعات الدنيا للمعلم $\boldsymbol{\beta}$.

القيم المطبقة والبواقي *Fitted values and Residuals*:

القيم المطبقة هي $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)'$ حيث

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

وهي قيم المتغير التابع أو الإستجابة التي تعطى من معادلة الإنحدار.
أيضا

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y}$$

حيث

$$\mathbf{M} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

البواقي هي القيم (المشاهدة - القيم المطبقة) أي $\mathbf{e} = \left(e_1, e_2, ..., e_n\right)'$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

أو

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{M}\mathbf{y} = \left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}$$

الإنحدار الخطي البسيط *Simple Linear Regression*:

الغرض من تحليل الإنحدار هو نمذجة العلاقة بين متغير إستجابة y مع واحد أو أكثر من المتغيرات المستقلة $x_1,x_2,...,x_p$ مثلا. أي اننا نريد إيجاد دالة g والتي تصف العلاقة بين y و $x_1,x_2,...,x_p$. لنعتبر النموذج الإحصائي التالي:

$$y = g\left(x_1,x_2,...,x_p\right)+\varepsilon$$

حيث $E\left(\varepsilon\right)=0$. كما هو واضح فإن النموذج يتكون من جزئين

(1) الجزء المحدد $y = g\left(x_1,x_2,...,x_p\right)$

(2) الجزء العشوائي ε. الخطأ العشوائي ε يدل على الحقيقة على انه لن يكون هناك علاقة كاملة وتامة بين y و $g\left(x_1,x_2,...,x_p\right)$.

في نموذج الإنحدار نفترض المتغيرات المستقلة $x_1,x_2,...,x_p$ ثابتة وقد تم قياسها بدون خطأ.

حالة خاصة: عندما $p=1$ و

$$g\left(x_1,x_2,...,x_p\right)=g\left(x_1\right)=\beta_0+\beta_1x_1$$

يسمى النموذج بنموذج الإنحدار الخطي البسيط.

تقدير المربعات الدنيا والقيم المطبقة والبواقي *Least-squares Estimation, Fitted Values, and Residuals*:

النموذج:

نموذج الإنحدار الخطي البسيط لعينة من n من الأفراد يكتب على الشكل:

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

لقيم $i = 1,2,...,n$. معالم الإنحدار هي β_0 و β_1 ويفترض انها ثابتة (غير عشوائية). الأخطاء ε_i يفترض عادة انها متغيرات عشوائية موزعة نفس التوزيع ومستقلة *iid* بمتوسط صفري وتباين ثابت σ^2.

الشكل المصفوفي:

سوف نكتب نموذج الإنحدار الخطي البسيط على الشكل $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ كالتالي:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$
$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$$

فرضية أن الأخطاء ε_i متغيرات عشوائية موزعة نفس التوزيع ومستقلة *iid* بمتوسط صفري وتباين ثابت σ^2 يعني أن

$$E\left(\boldsymbol{\varepsilon}\right)=\mathbf{0}$$

و

$$cov\left(\boldsymbol{\varepsilon}\right)=\sigma^2\mathbf{I}_n$$

تقدير المربعات الدنيا:

نريد تقدير $\boldsymbol{\beta}$ في نموذج الإنحدار الخطي البسيط. طالما تظل $\left(x_1,x_2,...,x_p\right)'$ ليست من أضعاف $\mathbf{j}_n$ وكذلك واحدة من $x_i\neq 0$ فإن $\text{rank}\left(\mathbf{X}\right)=2$ والمصفوفة $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ تكون موجودة ووحيدة وفي الحقيقة الحسابات التالية تبين ذلك:

$$\mathbf{X}'\mathbf{X}=\begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix},\quad \left(\mathbf{X}'\mathbf{X}\right)^{-1}=\begin{pmatrix} \frac{1}{n}+\frac{\bar{x}^2}{\sum_i\left(x_i-\bar{x}\right)^2} & -\frac{\bar{x}}{\sum_i\left(x_i-\bar{x}\right)^2} \\ -\frac{\bar{x}}{\sum_i\left(x_i-\bar{x}\right)^2} & \frac{1}{\sum_i\left(x_i-\bar{x}\right)^2} \end{pmatrix}$$

و

$$\mathbf{X}'\mathbf{y}=\begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

وهكذا فإن مقدر المربعات الدنيا الوحيد يعطى بالعلاقة:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \dfrac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} \end{pmatrix}$$

$$\mathbf{M} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \begin{pmatrix} \dfrac{1}{n} + \dfrac{\left(x_1 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} & \cdots & \dfrac{1}{n} + \dfrac{\left(x_1 - \bar{x}\right)\left(x_n - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} \\ \vdots & \ddots & \vdots \\ \dfrac{1}{n} + \dfrac{\left(x_1 - \bar{x}\right)\left(x_n - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} & \cdots & \dfrac{1}{n} + \dfrac{\left(x_n - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} \end{pmatrix}$$

والقيم المطبقة

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{M}\mathbf{y} = \begin{pmatrix} \dfrac{1}{n} + \dfrac{\left(x_1 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} & \cdots & \dfrac{1}{n} + \dfrac{\left(x_1 - \bar{x}\right)\left(x_n - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} \\ \vdots & \ddots & \vdots \\ \dfrac{1}{n} + \dfrac{\left(x_1 - \bar{x}\right)\left(x_n - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} & \cdots & \dfrac{1}{n} + \dfrac{\left(x_n - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \begin{pmatrix} \bar{y} + \hat{\beta}_1\left(x_1 - \bar{x}\right) \\ \bar{y} + \hat{\beta}_1\left(x_2 - \bar{x}\right) \\ \vdots \\ \bar{y} + \hat{\beta}_1\left(x_n - \bar{x}\right) \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

البواقي:

$$\hat{\mathbf{e}} = \left(\mathbf{I} - \mathbf{M}\right)\mathbf{y} = \mathbf{y} - \overline{\mathbf{y}} = \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix} = \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{pmatrix}$$

خواص مقدرات المربعات الدنيا:

النموذج $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ يسمى أحيانا نموذج جاوس - ماركوف *Gauss-Markov Model* حيث $\mathbf{X}$ مصفوفة ذات بعد $n \times p$ برتبة p (أي انها كاملة الرتبة) و

$$E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}$$

و

$$\mathrm{Cov}\left(\boldsymbol{\varepsilon}\right) = \sigma^2 \mathbf{I}_n$$

لقد وجدنا مقدر المربعات الدنيا للمعلم $\boldsymbol{\beta}$ يعطى بالعلاقة:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

متوسط $\boldsymbol{\beta}$ هو

$$\begin{aligned} \because E\left(\mathbf{y}\right) &= E\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) \\ &= E\left(\mathbf{X}\boldsymbol{\beta}\right) + E\left(\boldsymbol{\varepsilon}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} \\ &= \mathbf{X}\boldsymbol{\beta} \end{aligned}$$

$$E\left(\hat{\boldsymbol{\beta}}\right)=E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right]=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(\mathbf{y}\right)=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\beta}$$

أي أن $\hat{\boldsymbol{\beta}}$ مقدر غير محايز *Unbiased Estimator* ومصفوفة التباين- التغاير للمقدر $\hat{\boldsymbol{\beta}}$ هي

$$\begin{aligned}\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right)&=\mathrm{Cov}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right]\\&=\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathrm{Cov}\left(\mathbf{y}\right)\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]'\\&=\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\sigma^2\mathbf{I}_n\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]'\\&=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

لنموذج الإنحدار الخطي البسيط

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}=\begin{pmatrix}\dfrac{1}{n}+\dfrac{\bar{x}^2}{\sum_i\left(x_i-\bar{x}\right)^2} & -\dfrac{\bar{x}}{\sum_i\left(x_i-\bar{x}\right)^2}\\ -\dfrac{\bar{x}}{\sum_i\left(x_i-\bar{x}\right)^2} & \dfrac{1}{\sum_i\left(x_i-\bar{x}\right)^2}\end{pmatrix}$$

أي أن

$$\mathrm{V}\left(\hat{\beta}_0\right)=s_{00}\sigma^2=\frac{1}{n}+\frac{\bar{x}^2}{\sum_i\left(x_i-\bar{x}\right)^2}$$

$$\mathrm{V}\left(\hat{\beta}_1\right)=s_{11}\sigma^2=\frac{1}{\sum_i\left(x_i-\bar{x}\right)^2}$$

$$\mathrm{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\right)=s_{01}\sigma^2=\frac{-\bar{x}}{\sum_i\left(x_i-\bar{x}\right)^2}$$

تقدير تباين الخطأ:

تقدر σ^2 بالعلاقة

$$\hat{\sigma}^2 = s^2 \equiv \text{MSE} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}}{n-2}$$

وهو مقدر غير محايز.

تقدير مصفوفة التباين ـ التغاير:

تقدر $\text{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ بالعلاقة

$$\hat{\text{Cov}}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

عناصر هذه المصفوفة مفيدة في بناء فترات ثقة ولإختبار فرضيات حول β_0 و β_1.

الفرضية الطبيعية *Normality Assumption*:

في نموذج الإنحدار الخطي البسيط

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

لقيم $i = 1,2,...,n$ إذا افترض أن الأخطاء $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ متغيرات عشوائية لها توزيع متطابق ومستقل طبيعي $iid\ N\left(0,\sigma^2\right)$ (أي فرضية نموذج جاوس ـ ماركوف مضاف إليها الطبيعية). وبصورة مصفوفية

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0},\sigma^2\mathbf{I}\right)$$

فإن

- كلا من $\hat{\beta}_0$ و $\hat{\beta}_1$ موزعة طبيعيا.

- المتغير العشوائي $\frac{(n-1)MSE}{\sigma^2} \sim \chi^2_{n-2}$

- خطأ المتوسط الربع MSE مستقل عن كلا من $\hat{\beta}_0$ و $\hat{\beta}_1$.

مناقشة: إذا كانت

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

فإن

$$\mathbf{y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$$

وبما أن

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

تركيب خطي من $\mathbf{y}$ فهي ايضا موزعة طبيعيا. لاحظ أيضا أن

$$(n-2)MSE = SSE = \mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}$$

والذي يمكن إثبات أن

$$\sigma^{-2}\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y} \sim \chi^2_{n-r}$$

وحيث أن

$$r = \operatorname{rank}\left(\mathbf{X}\right) = 2$$

وأخيرا نلاحظ أن

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{B}\mathbf{y}$$

حيث

$$\mathbf{B} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

فإن $\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}$ و $\hat{\boldsymbol{\beta}}$ مستقلتين لأنه بوضع $\mathbf{A} = \left(\mathbf{I}-\mathbf{M}\right)$ في

$$\mathbf{B\Sigma A} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\left(\mathbf{I}-\mathbf{M}\right) = \mathbf{0}$$

مما يبين أن $\hat{\boldsymbol{\beta}}$ و MSE مستقلين.

الإستدلال الإحصائي في نموذج الإنحدار الخطي البسيط:

فترات الثقة وإختبار الفرضيات للمعالم β_0 و β_1 و σ^2:

المقدرات $\hat{\beta}_0$ و $\hat{\beta}_1$ هي مقدرات نقطة لمعالم المجتمع β_0 و β_1 على التوالي. سوف نناقش الآن كيفية الحصول على فترات ثقة وإختبار فرضيات حول معالم الإنحدار β_0 و β_1.

إستدلال حول β_1 **:**

من الفقرات السابقة وجدنا

$$\hat{\beta}_1 \sim N\left(\beta_1, s_{11}\sigma^2\right)$$

حيث

$$s_{11} = 1/\sum_i \left(x_i - \bar{x}\right)^2$$

وهكذا بوضعها في الشكل القياسي

$$z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s_{11}\sigma^2}} \sim N(0,1)$$

أيضا بما أن

$$\frac{(n-1)MSE}{\sigma^2} \sim \chi^2_{n-2}$$

و أن $\hat{\boldsymbol{\beta}}$ و MSE مستقلين فإن

$$t \equiv \frac{\hat{\beta}_1 - \beta_1}{\sqrt{s_{11}MSE}} = \frac{\left(\hat{\beta}_1 - \beta_1\right)\Big/\sqrt{s_{11}\sigma^2}}{\sqrt{\frac{(n-2)MSE}{\sigma^2}\Big/(n-2)}} \sim t_{n-2}$$

وهكذا فإن $100(1-\alpha)\%$ فترة ثقة للمعلم β_1 تعطى بالعلاقة

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{s_{11}MSE}$$

حيث $t_{n-2,\alpha/2}$ ترمز للربيع $\alpha/2$ الأعلى لتوزيع t بدرجات حرية n - 2. أيضا إذا أردنا أن نختبر عند مستوى α الفرضية

$$H_0: \beta_1 = \beta_{1,0}$$

ضد

$$H_1: \beta_1 \neq \beta_{1,0}$$

لقيمة معينة $\beta_{1,0}$ فإننا نستخدم

$$t \equiv \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s_{11} MSE}}$$

كإحصائة إختبار مع مجال رفض $R = \left\{ t : t > t_{n-2,\alpha/2} \text{ or } t < -t_{n-2,\alpha/2} \right\}$.

إستدلال حول β_0 :

بنفس الطريقة يمكن أن نبين أن

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{s_{00} MSE}} \sim t_{n-2}$$

حيث

$$s_{00} = \frac{1}{n} + \frac{\bar{x}^2}{\sum_i \left(x_i - \bar{x} \right)^2}$$

و $100(1-\alpha)\%$ فترة ثقة للمعلم β_0 تعطى بالعلاقة

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \sqrt{s_{00} MSE}$$

أيضا أن نختبر عند مستوى α الفرضية

$$H_0: \beta_0 = \beta_{0,0}$$

ضد

$$H_1: \beta_0 \neq \beta_{0,0}$$

لقيمة معينة $\beta_{0,0}$ فإننا نستخدم

$$t \equiv \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{s_{00} MSE}}$$

كإحصائة إختبار مع مجال رفض $R = \left\{ t : t > t_{n-2,\alpha/2} \text{ or } t < -t_{n-2,\alpha/2} \right\}$.

إستدلال حول σ^2:

بما أن

$$(n-1)MSE \,/\sigma^2 \sim \chi^2_{n-2}$$

فإنه يتبع أن

$$P\left\{ \chi^2_{n-2,1-\alpha/2} \leq \frac{(n-1)MSE}{\sigma^2} \leq \chi^2_{n-2,\alpha/2} \right\} = 1-\alpha$$

وتبعا لذلك فإن $100(1-\alpha)\%$ فترة ثقة للمعلم σ^2 تعطى بالعلاقة

$$\cdot \left(\frac{(n-1)MSE}{\chi^2_{n-2,\alpha/2}}, \frac{(n-1)MSE}{\chi^2_{n-2,1-\alpha/2}} \right)$$

تحليل التباين للإنحدار الخطي البسيط *The Analysis of Variance for Simple Linear Regression*:

يستخدم تحليل التباين لإختبار معنوية الإنحدار ويتم بتجزأة التغير الكلي المشاهد في **y**.

أولا نبدأ بتوضيح أن النموذج

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

يمكن أن يوضع على الشكل

$$y_i = \gamma_0 + \gamma_1 (x_i - \bar{x}) + \varepsilon_i$$

وهذا يسمى "إعادة تعلمة" النموذج *Reparameterisation* وكلا النموذجين يعطى جداول تحليل تباين متطابقة. الميزة في إستخدام

$$y_i = \gamma_0 + \gamma_1 (x_i - \bar{x}) + \varepsilon_i$$

أن أعمدة المصفوفة

$$\mathbf{X}_* = (\mathbf{j} \quad \mathbf{x}_0) = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_1 - \bar{x} \end{pmatrix}$$

متعامدة (لاحظ أن $rank(\mathbf{X}_*) = 2$).

ملاحظة: النموذج

$$y_i = \gamma_0 + \gamma_1 (x_i - \bar{x}) + \varepsilon_i$$

لقيم $i = 1,2,...,n$ يكتب بالشكل المصفوفي

$$\mathbf{y} = \mathbf{X}_*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

لننظر للمتطابقة

$$\begin{aligned}\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} &= \mathbf{y}'\mathbf{M}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{M})\mathbf{y} \\ &= \mathbf{y}'\left(n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{M})\mathbf{y}\end{aligned}$$

الآن

$$\mathbf{y}'\left(n^{-1}\mathbf{J}\right)\mathbf{y}$$

و

$$\mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}$$

أشكال رباعية مستقلة لأن

$$\begin{aligned}\left(n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}\left(\mathbf{M} - n^{-1}\mathbf{J}\right) &= n^{-1}\sigma^2\left(\mathbf{J}\mathbf{M} - \mathbf{J}\right) \\ &= n^{-1}\sigma^2\left(\mathbf{J} - \mathbf{J}\right) = \mathbf{0}\end{aligned}$$

كما ان من السهل إثبات أن

$$\mathbf{y}'\left(n^{-1}\mathbf{J}\right)\mathbf{y}$$

و

$$\mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}$$

كل منهم مستقل عن

$$\mathbf{y}'(\mathbf{I} - \mathbf{M})\mathbf{y}$$

وهكذا فإن مجموع المربعات غير المصحح $\mathbf{y}'\mathbf{y}$ *Uncorrected Sum* of *Squares*
وضع على شكل مجموع ثلاثة أشكال رباعية مستقلة عن بعضها البعض أي

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\left(n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}$$

والذي يكتب أيضا على الشكل

$$\mathbf{y}'\left(\mathbf{I} - n^{-1}\mathbf{J}\right)\mathbf{y} = \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}$$

أو بالشكل غير المصفوفي

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

مجاميع المربعات هذه تسمى كالتالي:

$$\mathbf{y}'\left(\mathbf{I} - n^{-1}\mathbf{J}\right)\mathbf{y}$$

مجموع المربعات الكلي المصحح *Corrected total Sum* of *Squares* ويرمز له *SST*.

$$\mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}$$

مجموع مربعات الإنحدار المصحح *Corrected Regression Sum of Squares* ويرمز له *SSR*.

$$\mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}$$

مجموع مربعات البواقي *Residuals Sum of Squares* ويرمز له *SSE*.

وفي نموذج الإنحدار الخطي البسيط بما أن 2 = rank($\mathbf{X}_*$) = rank(**M**) فيكون

$$rank\left(\mathbf{I}-n^{-1}\mathbf{J}\right)=n-1$$

$$rank\left(\mathbf{M}-n^{-1}\mathbf{J}\right)=2-1=1$$

$$rank\left(\mathbf{I}-\mathbf{M}\right)=n-2$$

وهكذا

$$rank\left(\mathbf{I}-n^{-1}\mathbf{J}\right)=rank\left(\mathbf{M}-n^{-1}\mathbf{J}\right)+rank\left(\mathbf{I}-\mathbf{M}\right)$$

وهذه هي كل درجات الحرية و درجات حرية الإنحدار و درجات حرية الخطأ على التوالي في جدول تحليل التباين للإنحدار الخطي البسيط التالي:

Source	*df*	*SS*	*MS*	*F*
Regression	1	*SSR*	*MSR*	*F = MSR/MSE*
Error	*n* - 2	*SSE*	*MSE*	
Total	*n* - 1	*SST*		

إختبار الفرضيات:

أهم إختبار في حالة الإنحدار الخطي البسيط هو

$$H_0:\quad \beta_1=0$$

ضد

$$H_1:\quad \beta_1\neq 0$$

وهذا تماما هو إفتراضنا تحت الفرضية البديلة أن

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon, i = 1,2,...,n$$

ويسمى النموذج الكامل وتحت الفرضية الصفرية أن

$$y_i = \beta_0 + \varepsilon, i = 1,2,...,n$$

ويسمى النموذج المخفض.

تحت الفرضية الصفرية H_0

$$\begin{aligned} E(SSR) &= E\left\{\mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}\right\} \\ &= \beta_0 \mathbf{j}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\beta_0 \mathbf{j} + \mathrm{tr}\left\{\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\sigma^2 \mathbf{I}\right\} \\ &= \sigma^2 rank\left(\mathbf{M} - n^{-1}\mathbf{J}\right) \\ &= \sigma^2 \end{aligned}$$

و

$$E(MSE) = \sigma^2$$

وهكذا تحت H_0 إحصائة الإختبار F تقدر شيئا قريب من 1.

إذا كانت H_0 غير صحيحة

$$
\begin{aligned}
E\left(SSR\right) &= E\left\{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}\right\} \\
&= \boldsymbol{\gamma}'\mathbf{X}_*'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{X}_*\boldsymbol{\gamma} + \mathrm{tr}\left\{\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}\right\} \\
&= \sigma^2 + \boldsymbol{\gamma}'\mathbf{X}_*'\mathbf{X}_*\boldsymbol{\gamma} \\
&= E\left(MSR\right)
\end{aligned}
$$

وبما أن

$$E\left(MSE\right) = \sigma^2$$

فإن F تقدر شيئا أكبر من 1. وهكذا فإن قيم F الكبيرة لا تتناسق مع النموذج المخفض في حال كون H_0 صحيحة.

لكي نحدد كيفية " كبر" F لكي نرفض H_0 (عند مستوى معنوية α مثلا) فإنه من الضروري فرض الطبيعية للأخطاء أي

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0},\sigma^2\mathbf{I}\right)$$

الإحصائة F:

في الإنحدار الخطي البسيط الإحصائة F هي نسبة MSR و MSE أي

$$F = \frac{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/\left(2-1\right)}{\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}/\left(n-2\right)}$$

توزيع F عندما تكون H_0 صحيحة :

سبق أن بينا أن

$$\mathbf{y}'\underbrace{\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)}_{\mathbf{A}}\mathbf{y}\sim\chi^2\left(1,\lambda=0\right)$$

وحيث أن

$$\mathbf{A\Sigma}=\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}=\mathbf{M}-n^{-1}\mathbf{J}$$

مصفوفة مثلية برتبة 1 و

$$\lambda=\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}=\frac{1}{2}\beta_0\mathbf{j}'\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\beta_0\mathbf{j}=0$$

كما انه يمكن إثبات أن

$$\sigma^{-2}\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}\sim\chi^2\left(n-2\right)$$

أيضا نعلم أن

$$\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}$$

و

$$\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}$$

أشكال رباعية مستقلة لأن

$$\left(\mathbf{I}-\mathbf{M}\right)\sigma^2\mathbf{I}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)=0$$

وهكذا فإن الإحصائة

$$F=\frac{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/(2-1)}{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-2)}=\frac{\sigma^{-2}\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/(2-1)}{\sigma^{-2}\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-2)}\sim F(1,n-2)$$

أي توزيع F المركزي بدرجات حرية 1 و n - 2 وترفض H_0 عند مستوى معنوية α عندما $F>F_{\alpha}(1,n-2)$.

توزيع F عندما تكون H_0 غيرصحيحة :

تحت عدم صحة الفرضية الصفرية H_0 فإن

$$\mathbf{y}'\underbrace{\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)}_{\mathbf{A}}\mathbf{y}\sim\chi^2\left(1,\lambda=\boldsymbol{\gamma}'\mathbf{X}_*'\mathbf{X}_*\boldsymbol{\gamma}/2\sigma^2\right)$$

وبما أن

$$\mathbf{A\Sigma}=\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}=\mathbf{M}-n^{-1}\mathbf{J}$$

مصفوفة مثلية برتبة 1 و

$$\lambda=\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}=\frac{1}{2}\boldsymbol{\gamma}'\mathbf{X}_*'\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{X}_*\boldsymbol{\gamma}=\boldsymbol{\gamma}'\mathbf{X}_*'\mathbf{X}_*\boldsymbol{\gamma}/2\sigma^2$$

وهكذا عندما تكون H_0 غير صحيحة فإن إحصائة الإختبار

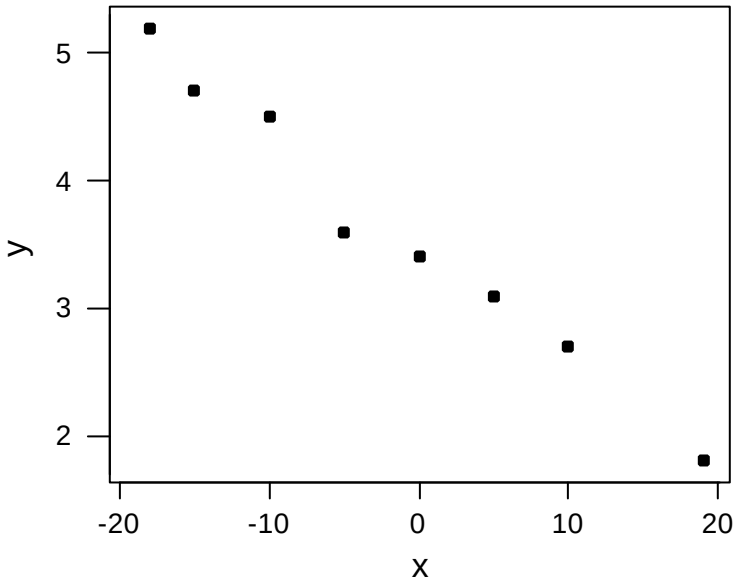
$$F=\frac{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/(2-1)}{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-2)}=\frac{\sigma^{-2}\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/(2-1)}{\sigma^{-2}\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-2)}\sim F(1,n-2,\lambda)$$

وبما أن F غير المركزية يزداد عشوائيا في معلمه عدم التمركز λ فإننا نتوقع أن تكون F كبيرة عندما تكون H_0 غير صحيحة.

مثال:

البيانات التالية عن إستهلاك الأكسجين للطيور (y) مقاسة عند مختلف درجات الحرارة (x).

x (*degrees Celcius*)	-18	-15	-10	-5	0	5	10	19
y (*ml/g/hr*)	5.2	4.7	4.5	3.6	3.4	3.1	2.7	1.8



من الرسم نرى أن النموذج $y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \varepsilon_i$ يبدو مناسبا. في الشكل المصفوفي لدينا $\mathbf{y} = \mathbf{X}_*\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ حيث $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)'$ و

$$\mathbf{X}_*' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -19.75 & -16.75 & -11.75 & -6.75 & -1.75 & 3.25 & 8.25 & 17.25 \end{pmatrix}$$

تحليل التباين *ANOVA* هو

Source	*df*	*SS*	*MS*	*F*
Regression	1	8.745	8.745	308.927
Error	6	0.170	0.028	
Total	7	8.915		

التحليل:

نريد أن نختبر الفرضية

$$H_0: \quad \gamma_1 = 0$$

ضد

$$H_1: \quad \gamma_1 \neq 0$$

ويجب أن نرفض H_0 لأن $F = 308.93$ وهي أكبر بكثير من $F_{0.05}(1,6) = 5.99$.

معامل التحديد *The Coefficient of Determination*:

بما أن $SST = SSR + SSE$ فإنه يتبع أن جزء التغير الكامل في البيانات والمفسر بواسطة النموذج هو

$$R^2 = \frac{SSR}{SST}$$

الإحصائة R^2 تسمى معامل التحديد. واضح أن

$$0 \le R^2 \le 1$$

كلما تكون R^2 اكبر كلما يكون الجزء المحدد من النموذج $\beta_0 + \beta_1 x$ يفسر التغير في البيانات. وهكذا قيمة R^2 "قريبة" من 1 تؤخذ كدليل على ان نموذج الإنحدار عمل جيدا في وصف التغير غي البيانات. أي لوكانت $R^2 = 0.97$ فهذا يعنى أن 97% من التغير في y تم تفسيره بواسطة x. لاحظ أن الأخذ بقيمة R^2 يكون صحيحا فقط تحت فرضية صحة النموذج.

ملاحظات:

1) إذا كانت R^2 صغيرة فإن هذا قد يعني أنه يوجد الكثير من التغير العشوائي في البيانات بحيث أن خط الإنحدار برغم انه مناسب إلى انه لم يفسر إلا هذه النسبة الصغيرة من التغير في البيانات.

2) قد تكون R^2 قريبة جدا من 1 ولكن نموذج الخط المستقيم قد لايكون النموذج الأفضل لوصف البيانات. قد تكون R^2 جدا مرتفعة ولكنها قد تكون غير مهمة لأنها تفترض أن نموذج الخط المستقيم صحيح بينما يكون هناك نموذج أفضل لوصف البيانات.

إختبار نقص التطبيق *Lack of Fit Test*:

هذا الإختبار يجرى عندما يكون هناك تكرارات للإستجابة y عند قيم لمتغير مستقل x أي لنفترض لقيم معينة x_j توجد عدة إستجابات y_{ij} حيث $i = 1,2,...,n_j$ و $j = 1,2,...,m$. لإجراء إختبار نقص التطبيق نحسب:

1) الجزء الأول: نوجد مجموع مربعات إنحرافات الإستجابة عن متوسطها لكل التكرارات أي $\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2$ ثم نجمع مجاميع المربعات هذة على كل الحالات أي

$$SSP = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2$$

يسمى SSP **مجموع مربعات خطأ نقي *Pure Error SS***. لاحظ انه في حالة وجود إستجابة واحدة عند قيمة لمتغير مستقل x_j فإن $y_{1j} = \bar{y}_j$ وبالتالي $y_{1j} - \bar{y}_j = 0$ وبالتالي x_j لايشارك في SSP. درجات الحرية التابعة لـ SSP هي $n - m$ لأنه يوجد $n_j - 1$ درجات حرية لكل x_j فيكون مجموع درجات الحرية هو

$\sum_{j=1}^{m}\left(n_j - 1\right) = \sum_{j=1}^{m} n_j - m = n - m$ ونلاحظ ايضا أن x_j التي ليست لها تكرارات في الإستجابة أي $n_j = 1$ وبالتالي $n_j - 1 = 0$ لاتشارك أيضا في درجات الحرية. متوسط مربع الخطأ النقي MSE_P يعطى بالعلاقة

$$MSE_P = \frac{SSP}{n - m}$$

السبب في تسمية "خطأ نقي" هو أن MSE_P مقدر غير حيازي لتباين الخطأ σ^2 بغض النظر عن طبيعة دالة الإنحدار فـ MSE_P يقيس التغير في توزيع y بدون الإعتماد على أي فرضية حول طبيعة علاقة الإنحدار. ولهذا فإنه مقياس نقي لتباين الخطأ.

2) الجزء الثاني: نحسب $SSL = SSE - SSP$ حيث SSL ترمز لمجموع مربعات نقص التطبيق *Lack of Fit SS* ويمكن أن نبين أنه

$$SSL = \sum_{j=1}^{m} n_j \left(\bar{y}_j - \hat{y}_j \right)^2$$

حيث $\hat{y}_j$ ترمز للقيم المطبقة عندما $x = x_j$. درجات الحرية التابعة SSL هي $m-2$ لأنه يوجد m قيمة لـ x و 2 درجة حرية إستخدمت لتقدير المعالم لحساب القيم المطبقة $\hat{y}_j$ ولهذا متوسط مربع نقص التطبيق MSL يعطى بالعلاقة

$$MSL = \frac{SSL}{m-2}$$

إختبار *F*:

الإحصائة لإختبار نقص التطبيق هي

$$F^* = \frac{MSL}{MSP}$$

هذه الإحصائة تتبع توزيع $F\left(m-2, n-m\right)$ ونقرر عدم ملائمة التطبيق إذا كانت

$$F^* > F\left(m-2, n-m\right)$$

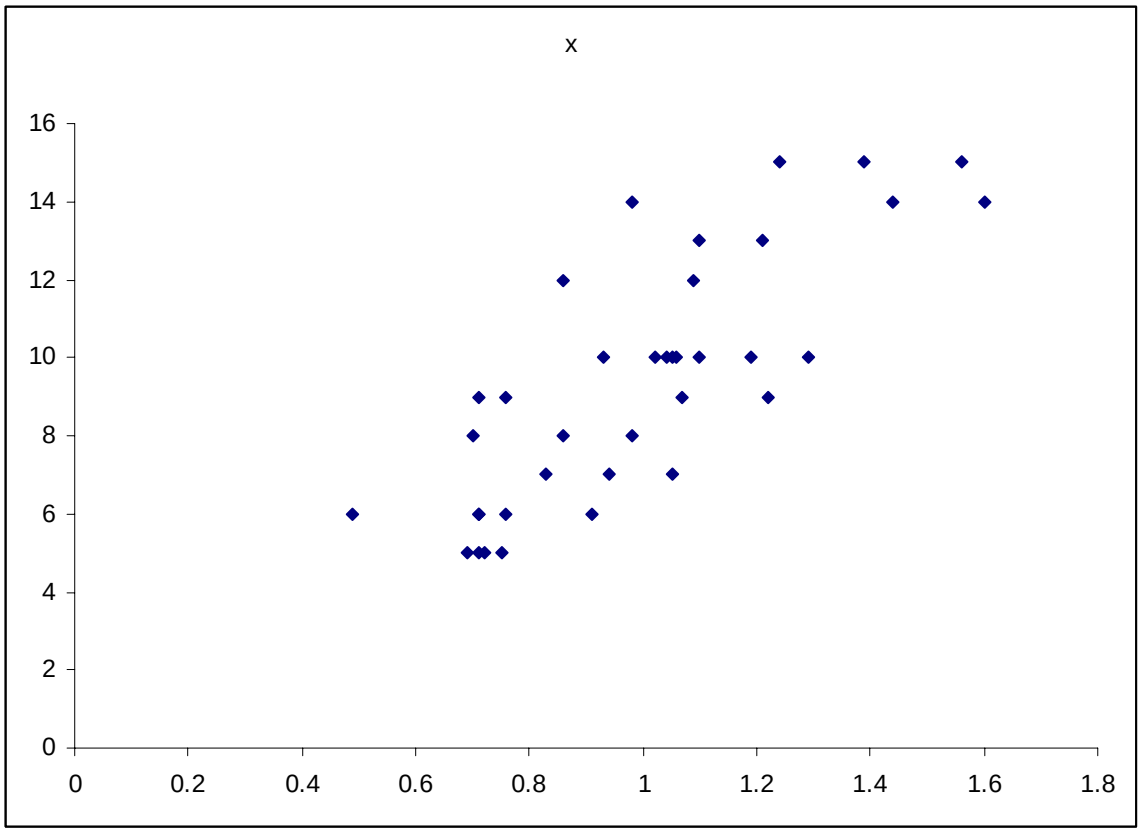
مثال:

البيانات التالي تعطي y كمتغير إستجابة للمتغير المستقل x:

y	x
0.69	5
1.39	15
1.09	12
0.76	9
0.86	12
1.04	10
1.10	13
1.02	10
1.10	10
1.07	9
0.94	7
0.98	14
0.71	5
0.75	5
1.05	10
1.44	14
0.76	6
0.72	5
1.29	10
0.71	6
0.86	8
0.93	10
0.91	6
0.70	8
1.56	15
1.06	10
0.71	9
1.19	10
0.83	7
0.49	6
1.24	15
1.22	9
0.98	8
0.71	6
1.60	14
1.21	13
1.05	7

بإستخدام Excel Anaysis ToolPac:

نرسم المشاهدات



نلاحظ تكرارات الإستجابة عند قيم مختلفة.

SUMMARY OUTPUT

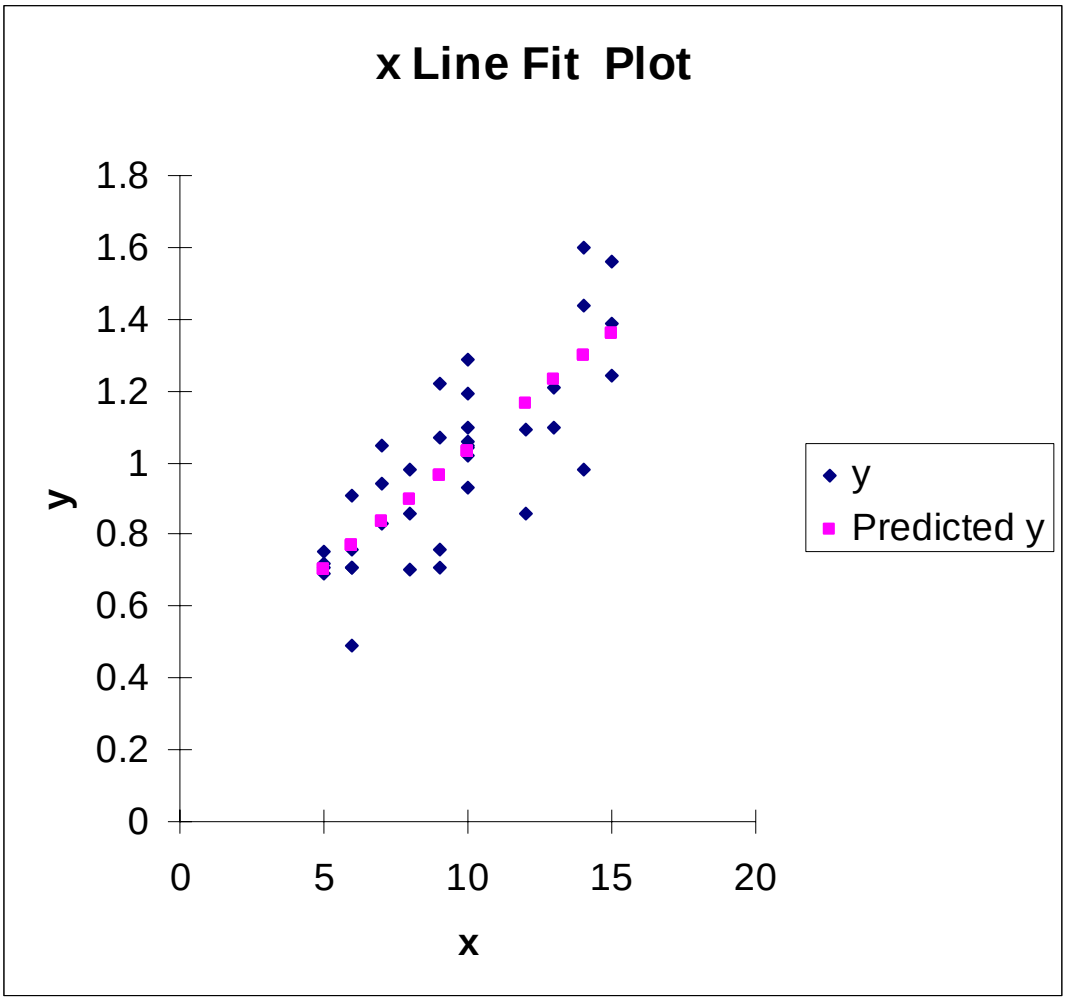
Regression Statistics	
Multiple R	0.79915365
R Square	0.63863746
Adjusted R Square	0.62831985
Standard Error	0.15921456
Observations	37

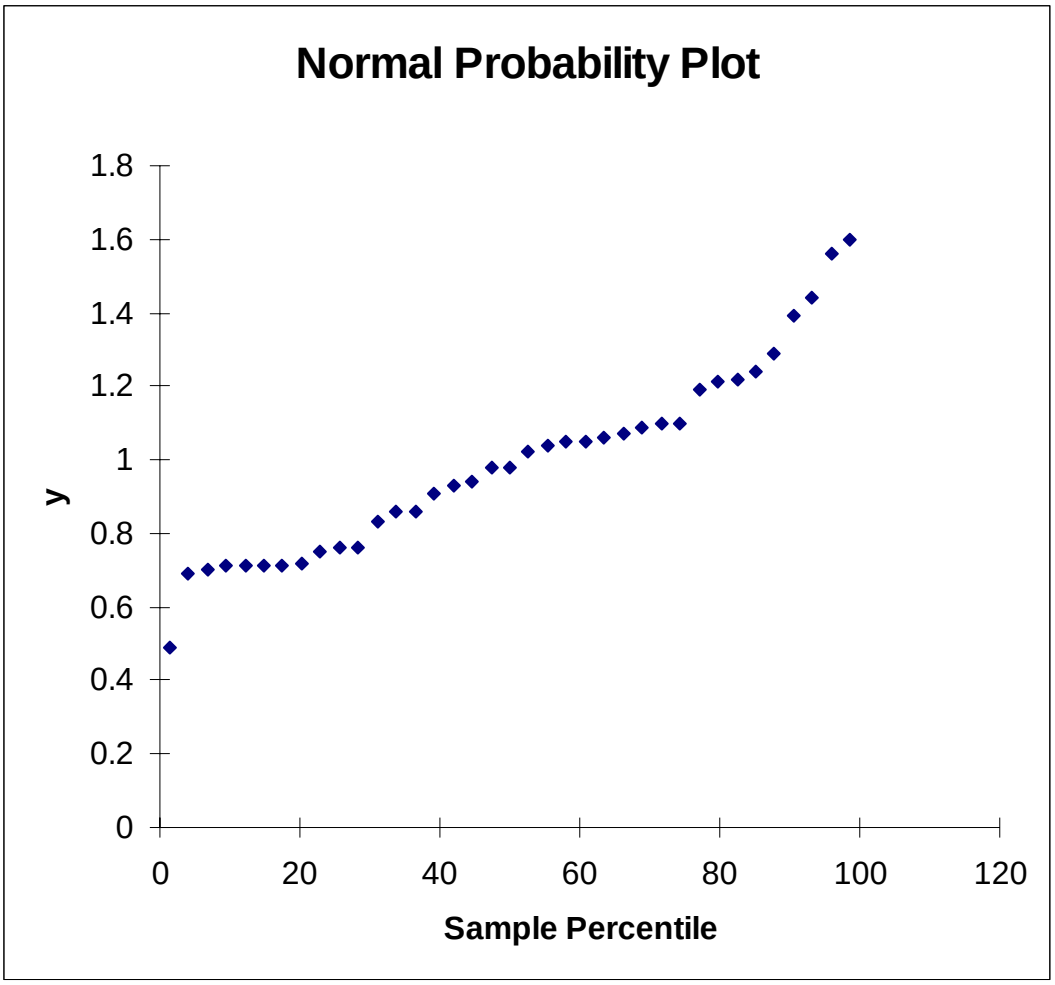
ANOVA

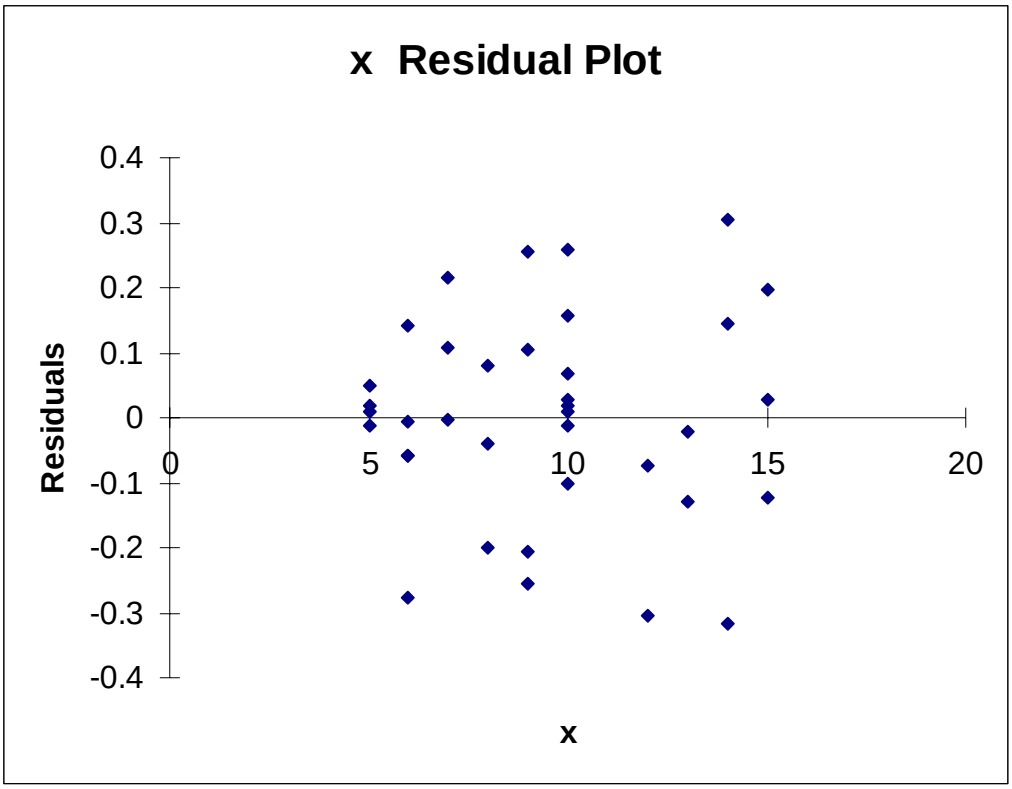
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.56805636	1.56805636	61.85802905	3.03091E-09

Residual	35	0.88722 4721	0.0253 49278
Total	36	2.45528 1081	

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.3707 62048	0.08326 3886	4.4528 5543	8.2707 4E-05	0.20172 7374	0.5397 96723	0.2017 27374	0.5397 96723
x	0.0660 97139	0.00840 3972	7.8649 87543	3.0309 1E-09	0.04903 6168	0.0831 5811	0.0490 36168	0.0831 5811

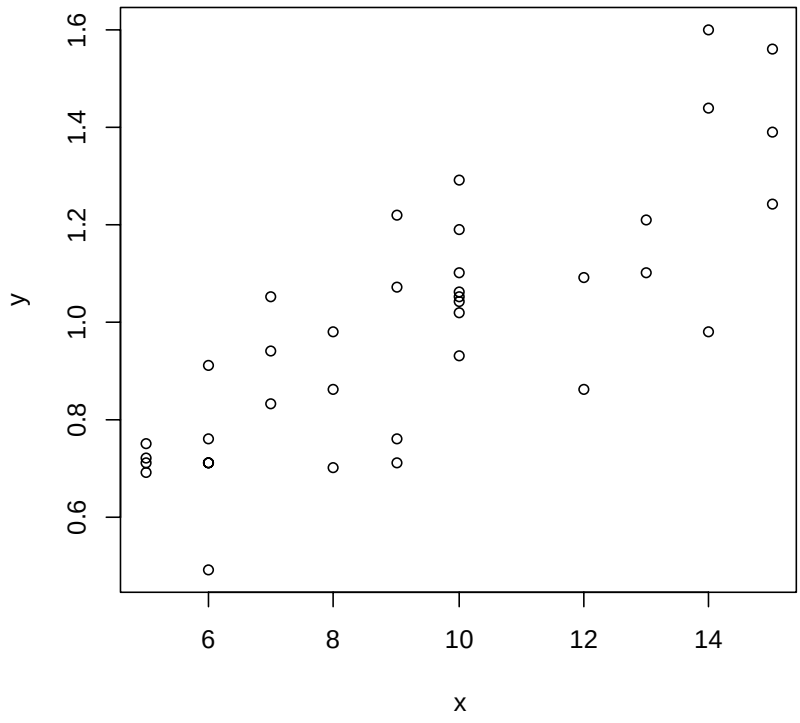



Normal Probability Plot
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
y
0
20
40
60
80
100
120
Sample Percentile


x Residual Plot
Residuals
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0
5
10
15
20
x

بإستخدام R:

```
> x =
c(5,15,12,9,12,10,13,10,10,9,7,14,5,5,10,14,6,5,10,
6,8,10,6,8,15,10,9,10,7,6,15,9,8,6,14,13,7)
> y =
c(0.69,1.39,1.09,0.76,0.86,1.04,1.10,1.02,1.10,1.07
,0.94,0.98,0.71,0.75,1.05,1.44,0.76,0.72,1.29,0.71,
0.86,0.93,0.91,0.70,1.56,1.06,0.71,1.19,0.83,0.49,1
.24,1.22,0.98,0.71,1.60,1.21,1.05)
> plot(x,y)
>
```



نستخدم lm

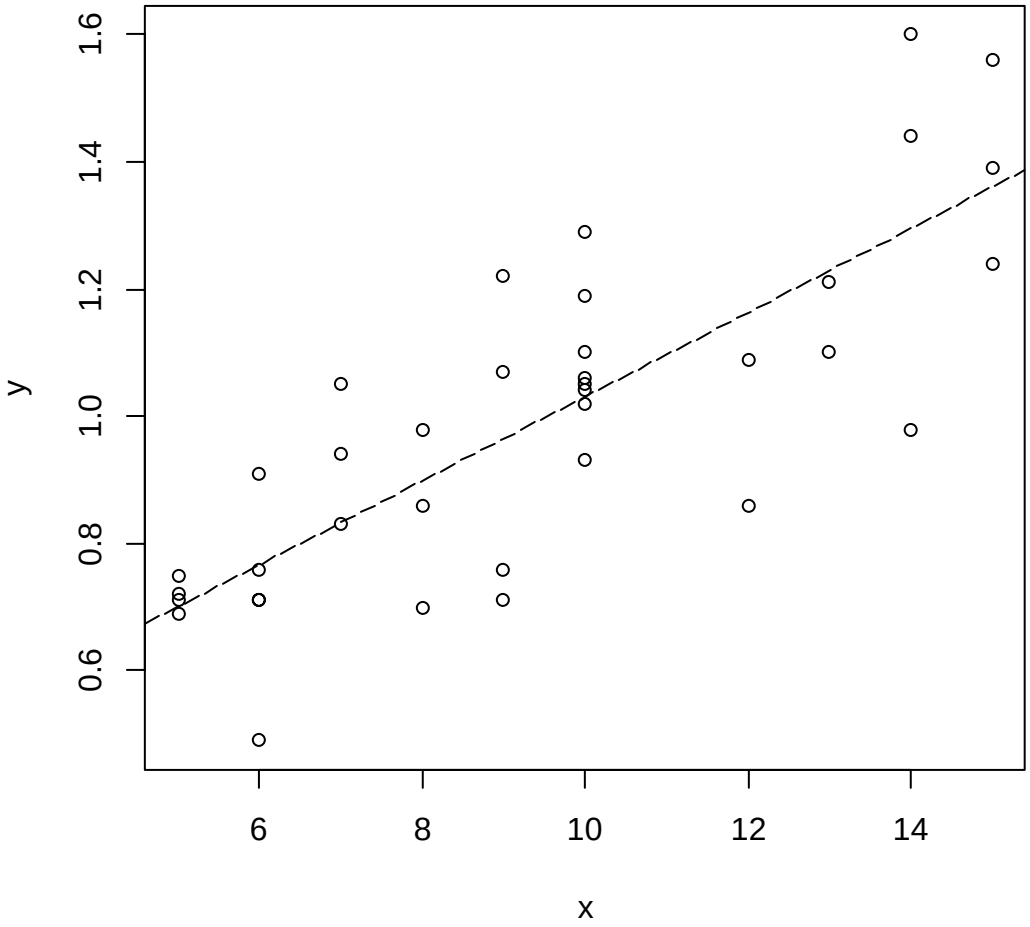
```
> yxfit = lm(y ~ x)
```

```
> summary(yxfit)

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.316122 -0.073928  0.008267  0.104364  0.303878

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.370762   0.083264   4.453 8.27e-05 ***
x           0.066097   0.008404   7.865 3.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.1592 on 35 degrees of freedom
Multiple R-squared: 0.6386,     Adjusted R-squared: 0.6283
F-statistic: 61.86 on 1 and 35 DF,  p-value: 3.031e-09

> anova(yxfit)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1.56806 1.56806  61.858 3.031e-09 ***
Residuals 35 0.88722 0.02535
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
>
```

ويرسم خط الإنحدار مع المشاهدات

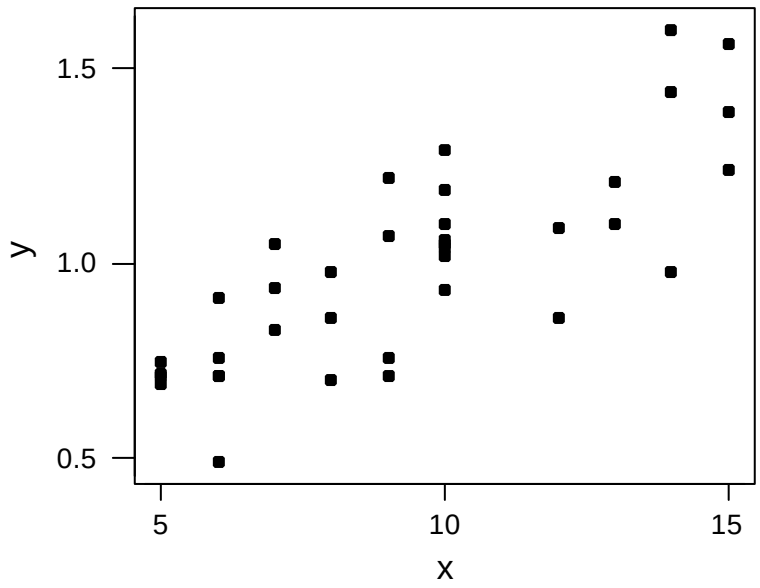
```
> plot(x,y)
> abline(yxfit$coef, lty = 5)
>
```



ويمكن الحصول على نتائج وتشخيصات كثيرة تترك كتمرين (انظر تحليل الإنحدار بإستخدام R في نهاية الكتاب)

بإستخدام Minitab:

ونرسم العلاقة بينهم في رسم إنتشار



Regression Analysis: y versus x

The regression equation is
y = 0.371 + 0.0661 x

Predictor	Coef	SE Coef	T	P
Constant	0.37076	0.08326	4.45	0.000
x	0.066097	0.008404	7.86	0.000

S = 0.1592 R-Sq = 63.9% R-Sq(adj) = 62.8%
PRESS = 0.995706 R-Sq(pred) = 59.45%

Analysis of Variance

Source	DF	SS	MS	F	P

Regression	1	1.5681	1.5681	61.86	0.000
Residual Error	35	0.8872	0.0253		
Lack of Fit	8	0.1740	0.0217	0.82	0.589
Pure Error	27	0.7132	0.0264		
Total	36	2.4553			

No evidence of lack of fit (P > 0.1)

الإنحدار المتعدد *Multiple Linear Regression*:

التقدير:

لقد أفترضنا في نموذج الإنحدار أن مصوفة التصميم $\mathbf{X}$ ذات رتبة كاملة ولذا يوجد حل وحيد لمقدر مربعات الدنيا لمعالم النموذج يعطى بالعلاقة

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

لقد بينا سابقا أن

$$E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$$
$$Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

والمقدر غير الحيازي *Unbiased* لـ σ^2 هو كمابينا سابقا

$$MSE = \frac{\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}}{rank\left(\mathbf{I}-\mathbf{M}\right)} = \frac{\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}}{n-p} = \frac{\hat{e}'\hat{e}}{n-p}$$

وبهذا فإن مقدر لـ $Cov\left(\hat{\boldsymbol{\beta}}\right)$ يعطى

$$\hat{Cov}\left(\hat{\boldsymbol{\beta}}\right) = MSE\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

نظرية جاوس ـ ماركوف

المقدرات الخطية:

مقدر خطي غير حيازي لـ $\boldsymbol{\beta}$ هو أي دالة خطية على الشكل

$$t(\mathbf{y}) = \mathbf{c} + \mathbf{A}\mathbf{y}$$

والذي يحقق

$$E\{t(\mathbf{y})\} = E(\mathbf{c} + \mathbf{A}\mathbf{y}) = \boldsymbol{\beta}$$

لجميع $\boldsymbol{\beta} \in R^p$ لاحظ أن

$$\boldsymbol{\beta} = E(\mathbf{c} + \mathbf{A}\mathbf{y}) = \mathbf{c} + \mathbf{A}E(\mathbf{y}) = \mathbf{c} + \mathbf{A}\mathbf{X}\boldsymbol{\beta}$$

وحيث أن هذه النتيجة صحيحة لجميع $\boldsymbol{\beta} \in R^p$ فهي صحيحة عندما

$$\mathbf{c} = 0$$

و

$$\mathbf{A}\mathbf{X} = \mathbf{I}$$

وهكذا فإننا نعتبر مقدرات من الشكل

$$\mathbf{A}\mathbf{y}$$

حيث

$$\mathbf{A}\mathbf{X} = \mathbf{I}$$

طبعا إذا كانت

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$$

فعندئذ

$$\mathbf{Ay} = \left(\mathbf{X'X}\right)^{-1}\mathbf{Xy} = \hat{\boldsymbol{\beta}} \text{ و } \mathbf{AX} = \mathbf{I}$$

نظرية:(نظرية جاوس - ماركوف)

إذا كانت

$$E\left(\mathbf{y}\right) = \mathbf{X}\boldsymbol{\beta}$$

و كانت $\mathbf{X}$ ذات رتبة كاملة وكان

$$Cov\left(\boldsymbol{\varepsilon}\right) = \sigma^2\mathbf{I}_n$$

عندئذ

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}$$

هو أفضل مقدر خطي غير حيازي (*BLUE*) *Best Linear Unbiased Estimator* للمعالم $\boldsymbol{\beta}$.

البرهان:

لنفترض أن $\mathbf{Ay}$ مقدر آخر غير حيازي للمعلم $\boldsymbol{\beta}$ والذي له

$$\mathbf{AX} = \mathbf{I}$$

يكفي أن نبين أن

$$Cov\left(\mathbf{Ay}\right) - Cov\left(\hat{\boldsymbol{\beta}}\right)$$

مصفوفة غير سالبة المحدودية. لدينا

$$Cov\left(\mathbf{Ay}\right)=\mathbf{A}Cov\left(\mathbf{y}\right)\mathbf{A}'=\sigma^2\left\{\mathbf{AA}'-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right\}$$

يكفي أن نبين أن

$$\mathbf{AA}'-\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

مصفوفة غير سالبة المحدودية. لنكتب $\mathbf{AA}'$ كالتالي

$$\mathbf{AA}'=\left\{\mathbf{A}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right\}\left\{\mathbf{A}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right\}'$$

نفكك هذا بإستخدام الحدود

$$\mathbf{AA}'-\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

و

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

فينتج (تحقق من ذلك)

$$\mathbf{AA}'=\left\{\mathbf{A}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right\}\left\{\mathbf{A}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right\}'+\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

بحيث أن

$$\mathbf{AA}'-\left(\mathbf{X}'\mathbf{X}\right)^{-1}=\left\{\mathbf{A}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right\}\left\{\mathbf{A}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right\}'$$

وهذا واضح انه مصفوفة غير سالبة محددة.

نتيجة: لنفترض أن

$$E\left(\mathbf{y}\right)=\mathbf{X}\boldsymbol{\beta}$$

و كانت $\mathbf{X}$ ذات رتبة كاملة وكان

$$cov\left(\boldsymbol{\varepsilon}\right)=\sigma^{2}\mathbf{I}_{n}$$

ليكن $\boldsymbol{\alpha}'\boldsymbol{\beta}$ تركيب خطي من $\beta_0,\beta_1,...,\beta_k$. عندئذ $\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}}$ هو مقدر المربعات الدنيا لـ $\boldsymbol{\alpha}'\boldsymbol{\beta}$ و $\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}}$ هو *BLUE*.

تطبيق:

في نموذج الإنحدار الخطي البسيط نعتبر التركيب الخطي

$$E\left(y\left|x_0\right.\right)=\beta_0+\beta_1x_0$$

وهو متوسط y عندما $x=x_0$ الـ *BLUE* لـ $E\left(y\left|x_0\right.\right)$ هو $\hat{\beta}_0+\hat{\beta}_1x_0$ حيث $\hat{\beta}_0$ و $\hat{\beta}_1$ هي مقدرات المربعات الدنيا للمعالم β_0 و β_1 على التوالي.

إعادة معلمة النموذج:

كما في حالة الإنحدار الخطي البسيط نكتب الإنحدار المتعدد بالشكل المعاد فيه معلمة النموذج

$$\begin{aligned}y_i&=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...,\beta_kx_{ik}+\varepsilon_i\\&=\alpha+\beta_1\left(x_{i1}-\overline{x}_1\right)+\beta_2\left(x_{i2}-\overline{x}_2\right)+...+\beta_k\left(x_{ik}-\overline{x}_k\right)+\varepsilon_i\end{aligned}$$

لقيم $i = 1,2,...,n$ حيث

$$\alpha = \beta_1 \bar{x}_1 + \beta_1 \bar{x}_2 + ... + \beta_k \bar{x}_k$$

وفي الشكل المصفوفي تكتب

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \begin{pmatrix} \mathbf{j} & \mathbf{X}_1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\varepsilon} \end{aligned}$$

حيث $\mathbf{j}$ متجه $n \times 1$ من الوحدة و $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, ..., \beta_k)'$ و

$$\mathbf{X}_1 = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

والنموذج المعاد معلمته

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_* \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ &= \begin{pmatrix} \mathbf{j} & \mathbf{X}_c \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\varepsilon} \end{aligned}$$

حيث $\mathbf{j}$ متجه $n \times 1$ من الوحدة و $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, ..., \beta_k)'$ و

$$\mathbf{X}_c = \left(\mathbf{I} - n^{-1}\mathbf{J}\right)\mathbf{X}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix}$$

المصفوفة $\left(\mathbf{I}-n^{-1}\mathbf{J}\right)$ تسمي مصفوفة التمركز *Centering Matrix* النموذج العادي (غير الممركز) والنموذج المعاد معلمته (الممركز) متكافئة تماما.

التقدير: نستخدم المعادلات الطبيعية *Normal Equations* لإيجاد مقدرات المربعات الدنيا للمعلم $\boldsymbol{\gamma}$. لاحظ أن

$$\mathbf{X}_*'\mathbf{X}_* = \begin{pmatrix}\mathbf{j}'\\ \mathbf{X}_c'\end{pmatrix}\begin{pmatrix}\mathbf{j} & \mathbf{X}_c\end{pmatrix} = \begin{pmatrix}\mathbf{j}'\mathbf{j} & \mathbf{j}'\mathbf{X}_c\\ \mathbf{X}_c'\mathbf{j} & \mathbf{X}_c'\mathbf{X}_c\end{pmatrix} = \begin{pmatrix}n & \mathbf{0}'\\ \mathbf{0} & \mathbf{X}_c'\mathbf{X}_c\end{pmatrix}$$

أيضا

$$\mathbf{X}_*'\mathbf{y} = \begin{pmatrix}\mathbf{j}'\\ \mathbf{X}_c'\end{pmatrix}\mathbf{y} = \begin{pmatrix}n\bar{y}\\ \mathbf{X}_c'\mathbf{y}\end{pmatrix}$$

وهكذا مقدر المربعات الدنيا الوحيد للمعلم $\boldsymbol{\gamma}$

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{X}_*'\mathbf{X}_*\right)^{-1}\mathbf{X}_*'\mathbf{y} = \begin{pmatrix}n^{-1} & \mathbf{0}'\\ \mathbf{0} & \left(\mathbf{X}_c'\mathbf{X}_c\right)^{-1}\end{pmatrix}\begin{pmatrix}n\bar{y}\\ \mathbf{X}_c'\mathbf{y}\end{pmatrix} = \begin{pmatrix}\bar{y}\\ \left(\mathbf{X}_c'\mathbf{X}_c\right)^{-1}\mathbf{X}_c'\mathbf{y}\end{pmatrix}$$

متجه القيم المطبقة

$$\hat{\mathbf{y}} = \mathbf{M}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}_*\hat{\boldsymbol{\gamma}} = \left(\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\right)'$$

حيث

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ..., \hat{\beta}_k x_{ik}\\ &= \hat{\alpha} + \hat{\beta}_1\left(x_{i1} - \bar{x}_1\right) + \hat{\beta}_2\left(x_{i2} - \bar{x}_2\right) + ... + \hat{\beta}_k\left(x_{ik} - \bar{x}_k\right)\end{aligned}$$

لقيم $i = 1,2,...,n$. متجة البواقي

$$\hat{\mathbf{e}} = \left(\hat{e}_1, \hat{e}_2, ..., \hat{e}_n\right)'$$

حيث

$$\hat{e}_i = y_i - \hat{y}_i$$

مجموع المربعات غير المصحح *Uncorrected Sum of Squares* $\mathbf{y}'\mathbf{y}$ يوضع على شكل مجموع ثلاثة أشكال رباعية مستقلة عن بعضها البعض أي

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\left(n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}$$

والذي يكتب أيضا على الشكل

$$\mathbf{y}'\left(\mathbf{I} - n^{-1}\mathbf{J}\right)\mathbf{y} = \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y} + \mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}$$

أو بالشكل غير المصفوفي

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

مجاميع المربعات هذه تسمى كالتالي:

$$\mathbf{y}'\left(\mathbf{I}-n^{-1}\mathbf{J}\right)\mathbf{y}$$

مجموع المربعات الكلي المصحح *Corrected total Sum* of *Squares* ويرمز له *SST*.

$$\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}$$

مجموع مربعات الإنحدار المصحح *Corrected Regression Sum of Squares* ويرمز له *SSR*.

$$\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}$$

مجموع مربعات البواقي *Residuals Sum of Squares* ويرمز له *SSE*.

وفي نموذج الإنحدار الخطي المتعدد أن

$$rank(\mathbf{M}) = rank(\mathbf{X}_*) = p = k+1$$

فيكون

$$rank\left(\mathbf{I}-n^{-1}\mathbf{J}\right) = n-1$$

$$rank\left(\mathbf{M}-n^{-1}\mathbf{J}\right) = (k+1)-1 = k$$

$$rank\left(\mathbf{I}-\mathbf{M}\right) = n-k-1$$

وهكذا

$$rank\left(\mathbf{I}-n^{-1}\mathbf{J}\right) = rank\left(\mathbf{M}-n^{-1}\mathbf{J}\right) + rank\left(\mathbf{I}-\mathbf{M}\right)$$

وهذه هي كل درجات الحرية و درجات حرية الإنحدار و درجات حرية الخطأ على التوالي في جدول تحليل التباين للإنحدار الخطي المتعدد التالي:

Source	*df*	*SS*	*MS*	*F*
Regression	k	*SSR*	*MSR*	$F = MSR/MSE$
Error	$n-k-1$	*SSE*	*MSE*	
Total	$n-1$	*SST*		

تقدير الأرجحية العظمى *Maximum Likelihood Estimation*:

لنعتبر النموذج العام

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

حيث

$$\boldsymbol{\varepsilon} \sim N_p\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

(فرضية جاوس - ماركوف مع الطبيعية) وفي هذه الحالة فإننا نعرف أن

$$\mathbf{y} \sim N_n\left(\mathbf{X\beta}, \sigma^2\mathbf{I}\right)$$

وسوف نفترض أن مصفوفة التصميم $\mathbf{X}$ ذات رتبة كاملة. على الخصوص نفترض أن $\mathbf{X}$ ذات بعد $n \times p$ حيث $p = k+1$ وأن $\text{rank}(\mathbf{X}) = p$ سوف نرمز لفضاء المعالم كالتالي

$$\boldsymbol{\Theta} = \left\{ \left(\boldsymbol{\beta}, \sigma^2 \right) : \left(\boldsymbol{\beta}, \sigma^2 \right) \in R^p \times R^+ \right\}$$

طريقة اخرى لتقدير $\boldsymbol{\beta}$ و σ^2 هي طريقة تقدير الأرجحية العظمى فتحت فرضية النموذج الطبيعي للمتغير $\mathbf{y}$ فإن دالة الأرجحية للمعالم $\boldsymbol{\beta}$ و σ^2 لـ $\mathbf{y} \in R^p$ تعطى بالعلاقة

$$L\left(\boldsymbol{\beta}, \sigma^2 \middle| \mathbf{y} \right) = (2\pi)^{-n/2} \left[\det\left(\sigma^2 \mathbf{I} \right) \right]^{-1/2} \exp\left\{ -\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right) \middle/ 2\sigma^2 \right\}$$

ولوغارثم دالة الأرجحية هو

$$\log L\left(\boldsymbol{\beta}, \sigma^2 \middle| \mathbf{y} \right) = \frac{-n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right) \Big/ 2\sigma^2$$

لكل قيمة لـ σ^2 الـ $\log L\left(\boldsymbol{\beta}, \sigma^2 \middle| \mathbf{y} \right)$ تأخذ قيمة عظمى بإختيار قيمة $\boldsymbol{\beta}$ التي تعظم

$$\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)$$

أي أن مقدر المربعات الدنيا

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{y}$$

هي أيضا مقدرات الأرجحية العظمى (MLE) تحت فرضية النموذج الطبيعي. بتعويض $\mathbf{X}\hat{\boldsymbol{\beta}}$ في لوغارثم دالة الأرجحية والإشتقاق بالنسبة لـ σ^2 نجد

$$\frac{\partial \log L\left(\boldsymbol{\beta},\sigma^2|\mathbf{y}\right)}{\partial \sigma^2}=-\frac{n}{2\sigma^2}+\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)\Big/2\sigma^4$$

وبوضع هذه العلاقة مساوية للصفر والحل لـ σ^2 نجد

$$\hat{\sigma}^2=\frac{\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n}=\frac{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}}{n}$$

من السهل أن نبين أن قيمة $\hat{\sigma}^2$ هذه تعظم $\log L\left(\boldsymbol{\beta},\sigma^2|\mathbf{y}\right)$ (برهن هذا!).

مقدر الأرجحية العظمى (MLE) لـ σ^2 محايز ونادرا مايستخدم في التطبيقات ويستخدم بدلا عنه متوسط مربع الخطأ

$$MSE=\frac{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}}{rank\left(\mathbf{I}-\mathbf{M}\right)}=\frac{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}}{n-p}$$

نتائج عن التوزيعات:

تحت الفرضية الطبيعية للنموذج وبأخذ $p=k+1$ نجد

$$\hat{\boldsymbol{\beta}}\sim N_p\left[\boldsymbol{\beta},\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]$$

$$n\hat{\sigma}^2/\sigma^2\sim\chi^2\left(n-p\right)$$

أو

$$\left(n-p\right)\mathrm{MSE}/\sigma^2\sim\chi^2\left(n-p\right)$$

و $\hat{\boldsymbol{\beta}}$ مستقلة عن كل من $\hat{\sigma}^2$ و MSE.

إستدلالات حول الإنحدار المتعدد:

لننظر لنموذج الإنحدار المتعدد

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

لقيم $i = 1,2,...,n$ حيث

$$E\left(\varepsilon_i\right) = 0$$

و

$$V\left(\varepsilon_i\right) = \sigma^2$$

و

$$Cov\left(\varepsilon_i, \varepsilon_j\right) = 0$$

لقيم $i \neq j$ (هذه فرضيات جاوس - ماركوف المعتادة). في الشكل المصفوفي نكتب النموذج على الشكل

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

حيث

$$E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}$$

و

$$Cov\left(\boldsymbol{\varepsilon}\right) = \sigma^2 \mathbf{I}$$

كما يفترض أن $\mathbf{X}$ ذات رتبة كاملة. النموذج المعاد معلمته

$$\begin{aligned}\mathbf{y} &= \mathbf{X}_{*}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ &= \begin{pmatrix}\mathbf{j} & \mathbf{X}_c\end{pmatrix}\begin{pmatrix}\alpha \\ \boldsymbol{\beta}_1\end{pmatrix} + \boldsymbol{\varepsilon}\end{aligned}$$

حيث $\mathbf{j}$ متجه $n \times 1$ من الوحدة و $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, ..., \beta_k)'$ و

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix}$$

جدول تحليل التباين للإنحدار الخطي المتعدد:

Source	*df*	*SS*	*MS*	*F*
Regression	k	*SSR*	*MSR*	*F = MSR/MSE*
Error	$n-k-1$	*SSE*	*MSE*	
Total	$n-1$	*SST*		

و مجاميع المربعات:

$$\mathbf{y}'\left(\mathbf{I}-n^{-1}\mathbf{J}\right)\mathbf{y}$$

مجموع المربعات الكلي المصحح *Corrected total Sum* of *Squares* ويرمز له SST.

$$\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}$$

مجموع مربعات الإنحدار المصحح *Corrected Regression Sum of Squares* ويرمز له *SSR*.

$$\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}$$

مجموع مربعات البواقي *Residuals Sum of Squares* ويرمز له *SSE*.

إختبار إجمالي:

يستخدم إختبار F إجمالي لإختبار معنوية الإنحدار فالإختبار اعلاه يختبر

$$H_0: \boldsymbol{\beta}_1 = \mathbf{0}$$

ضد

$$H_1: \boldsymbol{\beta}_1 \neq \mathbf{0}$$

ولتوضيح هذا لنفترض أن الفرضية $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ صحيحة أي أننا نفترض النموذج المختزل

$$\mathbf{y} = \alpha\mathbf{j} + \boldsymbol{\varepsilon}$$

ويكون

$$
\begin{aligned}
E\left(SSR\right) &= E\left\{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}\right\} \\
&= \alpha\mathbf{j}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\alpha\mathbf{j} + \mathrm{tr}\left\{\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}\right\} \\
&= \sigma^2 rank\left(\mathbf{M}-n^{-1}\mathbf{J}\right) = k\,\sigma^2
\end{aligned}
$$

لاحظ أن

$$\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\alpha\mathbf{j} = \mathbf{0}$$

وواضح أن

$$E\left(MSR\right) = \sigma^2$$

تحت الفرضية الصفرية وهكذا تحت صحة الفرضية الصفرية فإن F تقدر شيئا قريب من 1 (k?) وبما أن

$$E\left(MSE\right) = \sigma^2$$

دائما مهما كانت الفرضية الصفرية. من جهة اخرى عندما $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ غير صحيحة عندئذ

$$
\begin{aligned}
E\left(SSR\right) &= E\left\{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}\right\} \\
&= \boldsymbol{\gamma}'\mathbf{X}_*'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{X}_*\boldsymbol{\gamma} + \mathrm{tr}\left\{\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}\right\} \\
&= \begin{pmatrix}\alpha & \boldsymbol{\beta}_1'\end{pmatrix}\begin{pmatrix}\mathbf{j}' \\ \mathbf{X}_c'\end{pmatrix}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\begin{pmatrix}\mathbf{j} & \mathbf{X}_c\end{pmatrix}\begin{pmatrix}\alpha \\ \boldsymbol{\beta}_1\end{pmatrix} + k\,\sigma^2 \\
&= \begin{pmatrix}\alpha & \boldsymbol{\beta}_1'\end{pmatrix}\begin{pmatrix}\mathbf{j}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{j} & \mathbf{j}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{X}_c \\ \mathbf{X}_c'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{j} & \mathbf{X}_c'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{X}_c\end{pmatrix}\begin{pmatrix}\alpha \\ \boldsymbol{\beta}_1\end{pmatrix} + k\,\sigma^2 \\
&= \begin{pmatrix}\alpha & \boldsymbol{\beta}_1'\end{pmatrix}\begin{pmatrix}\mathbf{0} & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}_c'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{X}_c\end{pmatrix}\begin{pmatrix}\alpha \\ \boldsymbol{\beta}_1\end{pmatrix} + k\,\sigma^2 \\
&= k\,\sigma^2 + \boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1
\end{aligned}
$$

عندما تكون الفرضية الصفرية غير صحيحة

$$E\left(MSR\right)=\sigma^2+\boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1/k>\sigma^2$$

والتي تدل على أن القيم الكبيرة لـ F تكون ضد الفرضية الصفرية.

توزيع F تحت الفرضية الطبيعية:

بالإضافة لفرضية جاوس - ماركوف إذا أضفنا فرض الطبيعية للأخطاء أي $\boldsymbol{\varepsilon}\sim N_n\left(\mathbf{0},\sigma^2\mathbf{I}\right)$. ففي الإنحدار الخطي المتعدد الإحصائة F هي نسبة MSR و MSE أي

$$F=\frac{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/k}{\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}/\left(n-k-1\right)}$$

توزيع F عندما تكون H_0 صحيحة :

سبق أن بينا أن

$$\mathbf{y}'\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}\sim\chi^2\left(1,\lambda=0\right)$$

وحيث أن

$$\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}=\mathbf{M}-n^{-1}\mathbf{J}$$

مصفوفة مثلية برتبة k و

$$\lambda=\frac{1}{2}\alpha\mathbf{j}'\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\alpha\mathbf{j}=0$$

كما انه يمكن إثبات أن

$$\sigma^{-2}\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y} \sim \chi^2(n-k-1)$$

أيضا نعلم أن

$$\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}$$

و

$$\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}$$

أشكال رباعية مستقلة لأن

$$(\mathbf{I}-\mathbf{M})\sigma^2\mathbf{I}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)=0$$

وهكذا فإن الإحصائة

$$F=\frac{\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/k}{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-k-1)}=\frac{\sigma^{-2}\mathbf{y}'\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}/k}{\sigma^{-2}\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-k-1)}\sim F(k,n-k-1)$$

أي توزيع F المركزي بدرجات حرية k و $n-k-1$ وترفض H_0 عند مستوى معنوية α عندما $F>F_\alpha(k,n-k-1)$.

توزيع F عندما تكون H_0 غيرصحيحة :

تحت عدم صحة الفرضية الصفرية H_0 فإن

$$\mathbf{y}'\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\mathbf{y}\sim\chi^2\left(k,\lambda=\boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1/2\sigma^2\right)$$

وبما أن

$$\sigma^{-2}\left(\mathbf{M}-n^{-1}\mathbf{J}\right)\sigma^2\mathbf{I}=\mathbf{M}-n^{-1}\mathbf{J}$$

مصفوفة مثلية برتبة k و

$$\lambda = \frac{1}{2}\boldsymbol{\gamma}'\mathbf{X}_*'\sigma^{-2}\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{X}_*\boldsymbol{\gamma} = \boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1 / 2\sigma^2$$

وهكذا عندما تكون H_0 غير صحيحة فإن إحصائة الإختبار

$$F = \frac{\mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}/k}{\mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}/\left(n - k - 1\right)} = \frac{\sigma^{-2}\mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}/k}{\sigma^{-2}\mathbf{y}'\left(\mathbf{I} - \mathbf{M}\right)\mathbf{y}/\left(n - k - 1\right)} \sim F\left(k, n - k - 1, \lambda\right)$$

وبما أن F غير المركزية تزداد عشوائيا في معلمه عدم التمركز λ فإننا نتوقع أن تكون F كبيرة عندما تكون H_0 غير صحيحة.

مثال:

في مثال تذوق الجبن لننظر لإختبار

$$H_0: \mathrm{y_i} = \beta_0 + \varepsilon_i$$

النموذج المختزل ضد النموذج الكامل

$$H_0: \mathrm{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

جدول تحليل التباين *ANOVA* هو

Source	*df*	*SS*	*MS*	*F*	*Pr > F*
Regression	3	4994.51	1664.84	16.22	< 0.0001
Error	26	2668.38	102.63		
Total	29	7662.89			

نلاحظ أن إحصائية F المستخدمة لإختبار H_0 ضد H_1 كبيرة جدا بحيث أنها لاتتناسب مع النموذج المختصر. وهذا يبين أن واحدا من $\mathbf{x}_1$ أو $\mathbf{x}_2$ أو $\mathbf{x}_3$ مهم في وصف التذوق. لاحظ قيمة ألـ p (*p-value*) وهي 0.0001 وهو إحتمال كون H_0 صحيحة.

إختبار النماذج المختزلة ضد النماذج الكاملة: معالجة عامة

سوف نناقش بشكل عام إختبار النموذج المختصر ضد النموذج الكامل. F الكليه التي ناقشناها سابقا هي حالة خاصة من المعالجة العامة. لإعطاء حافذ نعتبر المثال التالي

مثال:

في تجربة تذوق الجبن لنفترض أن الباحث يعتقد أن $\mathbf{x}_1$ و $\mathbf{x}_2$ ليست مهمه في وصف $\mathbf{y}$ ففي هذه الحالة ربما نختبر

$H_0: \quad \beta_1 = \beta_2 = 0$

ضد

$H_1: \quad \text{not } H_0$

بوضع هذا الإختبار على شكل النموذج المختصر والنموذج الكامل

$$H_0: \quad y_i = \gamma_0 + \gamma_1 x_{i3} + \varepsilon_i$$

للنموذج المختصر ضد

$$H_1: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

للنموذج الكامل. هذا يعني أن الباحث يرغب أن يعرف فيما إذا كانت البيانات تنطبق على النموذج المختصر بشكل جيد كما هي تنطبق على النموذج الكامل. لاحظ أنه إذا كان النموذج المختصر جيد فإن النموذج الكامل يكون جيد أيضا. المهم هنا فيما إذا كان النموذج المختصر جيد.

سوف نبدأ بالنموذج الكامل ونفترض انه صحيح

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

والنموذج المختصر

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

في كلا النموذجين نفترض $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ و $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ (فرضية جاوس - ماركوف) لبيانات تذوق الجبن يكتب $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ الشكل:

$$\mathbf{y}_{30\times1}=\begin{pmatrix}12.3\\20.9\\\vdots\\5.5\end{pmatrix},\quad \mathbf{X}_{30\times4}=\begin{pmatrix}1&4.543&3.135&0.86\\1&5.159&5.043&1.53\\1&5.366&5.438&1.57\\\vdots&\vdots&\vdots&\vdots\\1&6.176&4.787&1.25\end{pmatrix},\quad \boldsymbol{\beta}_{4\times1}=\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\end{pmatrix},$$

$\boldsymbol{\varepsilon}_{30\times1}=(\varepsilon_1,\varepsilon_2,...,\varepsilon_{30})'$.

والنموذج المختصر $y_i=\gamma_0+\gamma_1 x_{i3}+\varepsilon_i$ على الشكل:

$$\mathbf{y}_{30\times1}=\begin{pmatrix}12.3\\20.9\\\vdots\\5.5\end{pmatrix},\quad \mathbf{X}_{30\times4}=\begin{pmatrix}1&0.86\\1&1.53\\1&1.57\\1&1.81\\\vdots&\vdots\\1&1.25\end{pmatrix},\quad \boldsymbol{\gamma}_{2\times1}=\begin{pmatrix}\gamma_0\\\gamma_1\end{pmatrix}$$

$\boldsymbol{\varepsilon}_{30\times1}=(\varepsilon_1,\varepsilon_2,...,\varepsilon_{30})'$.

تحت فرضية النموذج الكامل نكتب

$$\mathbf{M}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

وتحت فرضية النموذج المختصر نكتب

$$\mathbf{M}_0=\mathbf{X}_0\left(\mathbf{X}_0'\mathbf{X}_0\right)^{-1}\mathbf{X}_0'$$

لنعتبر الكمية

$$\frac{\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}}{rank\left(\mathbf{M}-\mathbf{M}_0\right)}=\frac{\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}}{p-p_0}$$

حيث $rank\left(\mathbf{M}\right)=p$ و $rank\left(\mathbf{M}_0\right)=p_0$ الآن

$$E\left\{\frac{\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}}{p-p_0}\right\}=\left(p-p_0\right)^{-1}\left[\boldsymbol{\gamma}'\mathbf{X}_0'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{X}_0\boldsymbol{\gamma}+\mathrm{tr}\left\{\left(\mathbf{M}-\mathbf{M}_0\right)\sigma^2\mathbf{I}\right\}\right]$$
$$=\left(p-p_0\right)^{-1}\left\{0+\sigma^2\left(p-p_0\right)\right\}=\sigma^2$$

أي عندما يكون النموذج المختصر صحيح فإن $\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}/p-p_0$ هي مقدر لـ σ^2. عندما يكون النموذج الكامل صحيح فإن

$$E\left\{\frac{\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}}{p-p_0}\right\}=\left(p-p_0\right)^{-1}\left[\boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{X}\boldsymbol{\beta}+\mathrm{tr}\left\{\left(\mathbf{M}-\mathbf{M}_0\right)\sigma^2\mathbf{I}\right\}\right]$$
$$=\left(p-p_0\right)^{-1}\left\{\boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{X}\boldsymbol{\beta}+\sigma^2\left(p-p_0\right)\right\}$$
$$=\sigma^2+\boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{X}\boldsymbol{\beta}/\left(p-p_0\right)$$

وبالتالي فإن $\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}/p-p_0$ تقدر كمية أكبر من σ^2.

إذا كانت $\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}/p-p_0$ ليست أكبر بكثير من σ^2 فهذا يعني أن النموذج المختصر صحيح وأن قيم $\mathbf{y}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{y}/p-p_0$ أكبر بكثير من σ^2 تدل على أن النموذج المختصر غير صحيح لأنها تبين أن $\boldsymbol{\beta}'\mathbf{X}'\left(\mathbf{M}-\mathbf{M}_0\right)\mathbf{X}\boldsymbol{\beta}>0$. في التطبيق العملي σ^2 لاتكون معلومة ولهذا يستخدم

$$MSE=\frac{\mathbf{y}'\left(\mathbf{I}-\mathbf{M}\right)\mathbf{y}}{n-p}$$

كمقدر غير حيازي لـ σ^2 تحتف فرضية النموذج الكامل. المناقشة السابقة تؤدي لإستخدام الإحصائة

$$F = \frac{\mathbf{y}'(\mathbf{M}-\mathbf{M}_0)\mathbf{y}/(p-p_0)}{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-p)}$$

بإضافة الفرضية أن $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0},\sigma^2\mathbf{I})$ وتحت صحة النموذج المختصر

$$\mathbf{y}'\underbrace{\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)}_{\mathbf{A}}\mathbf{y} \sim \chi^2(p-p_0,\lambda=0)$$

لأن $\mathbf{A\Sigma} = \sigma^{-2}(\mathbf{M}-\mathbf{M}_0)\sigma^2\mathbf{I} = \mathbf{M}-\mathbf{M}_0$ مصفوفة مثلية برتبه $p-p_0$ و

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\gamma}'\mathbf{X}_0'\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)\mathbf{X}_0\boldsymbol{\gamma} = 0$$

ونعلم أن

$$\sigma^{-2}\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y} \sim \chi^2(n-p)$$

وايضا نعلم ان

$$\mathbf{y}'(\mathbf{M}-\mathbf{M}_0)\mathbf{y}$$

و

$$\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}$$

مستقلتين لأن

$$(\mathbf{I}-\mathbf{M})\sigma^2\mathbf{I}(\mathbf{M}-\mathbf{M}_0) = \mathbf{0}$$

فالإحصائة

$$F=\frac{\mathbf{y}'(\mathbf{M}-\mathbf{M}_0)\mathbf{y}/(p-p_0)}{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-p)}=\frac{\mathbf{y}'\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)\mathbf{y}/(p-p_0)}{\mathbf{y}'\sigma^{-2}(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-p)}\sim F(p-p_0,n-p)$$

لها توزيع F المركزي بدرجات حرية $p-p_0$ (للبسط) و $n-p$ (للمقام) .

لإجراء إختبار مستوى α فإننا نرفض الفرضية الصفرية إذا كان

$$F>F_\alpha(p-p_0,n-p)$$

من جهة اخرى إذا كانت الفرضية الصفرية غير صحيحة عندئذ

$$\mathbf{y}'\underbrace{\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)}_{\mathbf{A}}\mathbf{y}\sim\chi^2(p-p_0,\lambda)$$

وبما أن $\mathbf{A\Sigma}=\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)\sigma^2\mathbf{I}=\mathbf{M}-\mathbf{M}_0$ مصفوفة مثلية برتبه $p-p_0$ و

$$\lambda=\frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}=\frac{1}{2}\boldsymbol{\beta}'\mathbf{X}'\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)\mathbf{X}\boldsymbol{\beta}$$

وفي هذه الحالة الإحصائة

$$F=\frac{\mathbf{y}'(\mathbf{M}-\mathbf{M}_0)\mathbf{y}/(p-p_0)}{\mathbf{y}'(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-p)}=\frac{\mathbf{y}'\sigma^{-2}(\mathbf{M}-\mathbf{M}_0)\mathbf{y}/(p-p_0)}{\mathbf{y}'\sigma^{-2}(\mathbf{I}-\mathbf{M})\mathbf{y}/(n-p)}$$
$$\sim F\left(p-p_0,n-p,\lambda=\frac{1}{2}\boldsymbol{\beta}'\mathbf{X}\sigma^{-2}(\mathbf{M}-\mathbf{M})\mathbf{X}\boldsymbol{\beta}\right)$$

وحيث أن توزيع F غير المركزي يزداد عشوائيا في معلم عدم التمركز λ فإن القيم الكبيرة لـ F تدل على عدم صحة النموذج المختصر.

مجموع مربعات الإنحدار *Regression Sums of Squares*:

مجموع مربعات الإنحدار المعدل (*Corrected*) للنموذج الكامل

$$SSR_F = \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}$$

وللنموذج المختصر

$$SSR_R = \mathbf{y}'\left(\mathbf{M}_0 - n^{-1}\mathbf{J}\right)\mathbf{y}$$

وحيث أن مجموع مربعات الإنحدار SSR لايتناقص أبدا بإضافة متغيرات مستقلة جديدة فإن

$$SSR_F = \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y} \geq \mathbf{y}'\left(\mathbf{M}_0 - n^{-1}\mathbf{J}\right)\mathbf{y} = SSR_R$$

وبالتالي:

1- إذا كان $SSR_F = \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}$ و $SSR_R = \mathbf{y}'\left(\mathbf{M}_0 - n^{-1}\mathbf{J}\right)\mathbf{y}$ "قريبان" فإن المتغيرات المستقلة الإضافية لاتضيف شيئا يذكر وبالتالي فإن النموذج المختصر يبلي حسنا في وصف البيانات مثل النموذج الكامل.

2- إذا كان $SSR_F = \mathbf{y}'\left(\mathbf{M} - n^{-1}\mathbf{J}\right)\mathbf{y}$ و $SSR_R = \mathbf{y}'\left(\mathbf{M}_0 - n^{-1}\mathbf{J}\right)\mathbf{y}$ غير "قريبان" فإن المتغيرات المستقلة الإضافية تضيف إضافة مهمة وبالتالي فإن النموذج المختصر لا يبلي حسنا في وصف البيانات مثل النموذج الكامل.

وبهذا فإن الكمية

$$SSR_F - SSR_R = \mathbf{y}'\left(\mathbf{M} - \mathbf{M}_0\right)\mathbf{y}$$

تستخدم لإختبار (أنظر F أعلاه) أي من النموذجين نختار فإذا كان الفرق كبير فهذا يعني ان النموذج المختصر لايعطي وصفا للبيانات في جودة النموذج الكامل .

ملاحظة: لاحظ أن $SSR_F - SSR_R = SSE_R - SSE_F$ وهذا يعني أن الفرق في مجموع مربعات الإنحدار يساوي الفرق في مجموع مربعات الخطأ (في القيمة المطلقة).

جدول تحليل الإنحدار للمثال:

ANOVA للنموذج المختصر:

Source	*df*	*SS*	*MS*	*F*	*Pr > F*
Model	1	3800.398	3800.398	27.55	< 0.0001
Error	28	3862.489	137.946		
Corrected Total	29	7662.887			

ANOVA للنموذج الكامل:

Source	*df*	*SS*	*MS*	*F*	*Pr > F*
Model	3	4994.509	1664.836	16.22	< 0.0001
Error	26	2668.378	102.629		
Corrected Total	29	7662.887			

وهكذا بأخذ

$$\mathbf{y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{y} = SSR_F - SSR_R = 4994.509 - 3800.398 = 1194.111$$

و $p - p_0 = 2$ و $n - p = 26$ نجد

$$F = \frac{\mathbf{y}'(\mathbf{M} - \mathbf{M}_0)\mathbf{y}/(p - p_0)}{\mathbf{y}'(\mathbf{I} - \mathbf{M})\mathbf{y}/(n - p)} = \frac{1194.111/2}{102.630} = 5.82$$

وبما أن $F = 5.82 > F_{2,26,0.05} = 3.369$ فإننا نرفض H_0 فإننا نستنتج أن النموذج المختزل ليس في مثل جودة النموذج الكامل في وصف البيانات.وأن المتغيرين *ACETIC* (x_1) و *H2S* (x_2) يضيفان معلومات مهمه للنموذج.

مجاميع المربعات التتابعية والجزئية *Sequential and Partial Sums of Squares*:

مجاميع المربعات التتابعية تقوم بتركيم مجاميع المربعات التابعة للإنحدار وقيمتها تعتمد على الترتيب الذي تضاف به المتغيرات المستقلة x's للنموذج. لننظر إلى تفكيك *SSR* في الجدول التالي:

Sequential ANOVA	*df*	*SS*
Regression on x_1 (*after* β_0)	1	$R(\beta_1 \mid \beta_0)$
Regression on x_2 (*after* β_0 *and* x_1)	1	$R(\beta_2 \mid \beta_0, \beta_1)$
Regression on x_3 (*after* β_0 *and* x_1 *and* x_2)	1	$R(\beta_3 \mid \beta_0, \beta_1, \beta_2)$
⋮		
Regression on x_k (*after* β_0 *and* $x_1, x_2, ..., x_k$)	1	$R(\beta_k \mid \beta_0, \beta_1, ..., \beta_{k-1})$

حقيقة:

مجاميع المربعات التتابعية مجموعها هو *SSR* المصحح في جدول *ANOVA* الكلي أي:

$$SSR = R(\beta_1 \mid \beta_0) + R(\beta_2 \mid \beta_0, \beta_1) + R(\beta_3 \mid \beta_0, \beta_1, \beta_2) + \cdots + R(\beta_k \mid \beta_0, \beta_1, ..., \beta_{k-1})$$

ملاحظة:

مجاميع المربعات التتابعية تساعد في تقييم نفعية إضافة متغيرات مستقلة x إلى النموذج في خطوات تدريجية. حيث توجد حالات تكون فيها هذه الطريقة من التفكير مفيدة جدا فمثلا لننظر إلى النموذج التكعيبي

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

ولنفترض أننا مهتمين فيما إذا كانت الحدود التربيعية والتكعيبية مهمه أي فيما إذا كانت $\beta_2 = \beta_3 = 0$ نستطيع الإجابة على هذا بفحص كل من

$$R\left(\beta_2 \middle| \beta_0, \beta_1\right)$$

و

$$R\left(\beta_3 \middle| \beta_0, \beta_1, \beta_2\right)$$

فمثلا إذا كان كليهما كبير فهذا يعني أن x^2 و x^3 يجب أن يظلا في النموذج. أما إذا كانت $R\left(\beta_2 \middle| \beta_0, \beta_1\right)$ كبيرة ولكن $R\left(\beta_3 \middle| \beta_0, \beta_1, \beta_2\right)$ ليست كذلك فهذا يوجب إستخدام نموذج تربيعي فقط. يوجد خلاف في كيفية التوجه عندما تكون $R\left(\beta_3 \middle| \beta_0, \beta_1, \beta_2\right)$ كبيرة و $R\left(\beta_2 \middle| \beta_0, \beta_1\right)$ صغيرة.

عدم التفرد:

ترتيب تحليل التباين التتابعي غير وحيد. إذا تغير الترتيب الذي تضاف به المتغيرات إلى النموذج فإن مجاميع المربعات التتابعية سوف يتغير أيضا (ولكن مجموعهم سيصل SSR). التالي مستخرجات من *Minitab* توضح ذلك

النموذج الكامل:

$$TASTE = \beta_0 + \beta_1\ ACETIC + \beta_2\ H2S + \beta_3\ LACTIC + \varepsilon$$

مجاميع المربعات التتابعية هي:

Source	DF	Seq SS
ACETIC	1	2314.1
H2S	1	2147.1
LACTIC	1	533.3

وهكذا فإن

$$R\left(\beta_1 \middle| \beta_0\right) = 2314.1$$

$$R\left(\beta_2 \middle| \beta_0, \beta_1\right) = 2147.1$$

$$R\left(\beta_3 \middle| \beta_0, \beta_1, \beta_2\right) = 533.3$$

ويلاحظ أن مجموعها SSR = 4994.5 وذلك من جدول تحليل التباين

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	4994.5	1664.8	16.22	0.000
Residual Error	26	2668.4	102.6		
Total	29	7662.9			

إختبار إحصائية F يختبر تتابعيا فيما إذا كانت إضافة متغير تابع مهمه للنموذج أم لا.

فمثلا F =2314.1/102.6 = 22.55 تختبر الفرضية:

H_0 : $TASTE = \gamma_0 + \varepsilon$

H_1 : $TASTE = \beta_0 + \beta_1\ ACETIC + \varepsilon$

وواضح أن *ACETIC* يجب أن تضاف إلى النموذج الذي يحوي التقاطع فقط أي

TASTE = - 61.5 + 15.6 ACETIC

ثم $F = 2147.1/102.6 = 20.93$ تختبر الفرضية

H_0 : TASTE = γ_0 + γ_1 ACETIC + ε

H_1 : TASTE = β_0 + β_1 ACETIC + β_2 H2S + ε

وهذا يؤكد إضافة *H2S* إلى النموذجح الذي يحوي *ACETIC* والتقاطع أي

TASTE = - 26.9 + 3.80 ACETIC + 5.15 H2S

أخيرا $F = 533.3/ 102.6 = 5.2$ تختبر الفرضية

H_0 : TASTE = γ_0 + γ_1 ACETIC + γ_2 H2S + ε

H_1 : TASTE = β_0 + β_1 ACETIC + β_2 H2S + β_3 LACTIC + ε

وهكذا فإن *LACTIC* يجب أن يضاف إلى النموذج الذي يحوي *H2S* و *ACETIC* والتقاطع أي النموذج

TASTE = - 28.9 + 0.33 ACETIC + 3.91 H2S + 19.7 LACTIC

مجاميع المربعات التتابعية لترتيب مختلف:

لنفترض إدخال المتغيرات المستقلة في نموذج الإنحدار بالترتيب التالي:

$$\text{TASTE} = \beta_0 + \beta_1\ \text{H2S} + \beta_2\ \text{LACTIC} + \beta_3\ \text{ACETIC} + \varepsilon$$

وهذا يعطي مجاميع المربعات التتابعية التالية:

Source	DF	Seq SS
H2S	1	4376.8
LACTIC	1	617.1
ACETIC	1	0.6

وهذا يعطي:

$$R\left(\beta_1 \middle| \beta_0\right) = 4376.8$$

$$R\left(\beta_2 \middle| \beta_0, \beta_1\right) = 617.1$$

$$R\left(\beta_3 \middle| \beta_0, \beta_1, \beta_2\right) = 0.6$$

ويلاحظ أن مجموعها لايزال $SSR = 4994.5$ وذلك من جدول تحليل التباين

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	4994.5	1664.8	16.22	0.000
Residual Error	26	2668.4	102.6		
Total	29	7662.9			

من المستغرب والمهم ملاحظته انه حسب هذا الترتيب فإن *ACETIC* يجب ألا يضاف إلى نموذج يحوي مسبقا *H2S* و *LACTIC* !!! تذكر انه في الترتيب السابق اضيف *ACETIC* إلى نموذج يحوي فقط التقاطع β_0.

وفي التلخيص:

الإحصائة F التي تعتمد على مجاميع المربعات التتابعية تستخدم لإختبار فيما إذا كانت x_1 يجب ان تضاف إلى نموذج يحوي سابقا x_1 و x_2 و ... و x_{l-1} و β_0 لقيم $l = 1,2,..,k$ أي ان F_1 يختبر

$$H_0: \quad y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_{l-1} x_{i(l-1)} + \varepsilon_i$$
$$H_1: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{l-1} x_{i(l-1)} + \beta_l x_{il} + \varepsilon_i$$

ونحسبه من نسبة $R\left(\beta_l \middle| \beta_0, \beta_1, \ldots, \beta_{l-1}\right)$ إلى *MSE* من جدول تحليل التباين الكلي أي

$$F_l = \frac{R\left(\beta_l \middle| \beta_0, \beta_1, \ldots, \beta_{l-1}\right)}{MSE}$$

عندما تكون H_0 صحيحة فإن $F_l \sim F_{1,n-k-1}$ والقيم الكبيرة لـ F_l تكون ضد H_0.

مجاميع المربعات الجزئية *Partial Sums of Squares*:

وتساعد على تقييم إضافة متغير مستقل x إلى نموذج يحوي مسبقا كل المتغيرات المستقلة 1 - k و β_0 أي لقيم $l = 1,2,..,k$

$$Partial\ SS\ for\ x_l = R\left(\beta_l \middle| \beta_0, \beta_1, \ldots, \beta_{l-1}, \beta_{l+1}, \ldots, \beta_k\right)$$

بخلاف مجاميع المربعات التتابعية فإن مجاميع المربعات الجزئية لاتجمع لتساوي *SSR*.

إختبارات *F* الجزئية:

مجاميع المربعات الجزئية تمكننا من دراسة تأثير وضع x_l معينة كآخر متغير في النموذج أي أنها تمكننا من إختبار لقيم $l = 1,2,..,k$

$$H_0: \quad y_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_{l-1} x_{i(l-1)} + \gamma_{l+1} x_{i(l+1)} + \cdots + \gamma_k x_{ik} + \varepsilon_i$$
$$H_1: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (the\ \ full\ \ \ mod\, el)$$

الإحصائة F المستخدمة لإختبار H_0 ضد H_1 تعطى بالعلاقة

$$F_l = \frac{R\left(\beta_l \middle| \beta_0, \beta_1, \ldots, \beta_{l-1}, \beta_{l+1}, \ldots, \beta_k\right)}{\text{MSE}}$$

عندما تكون H_0 صحيحة عندئذ $F_l \sim F_{1,n-k-1}$ والقيم الكبيرة لـ F_l تكون ضد H_0.

مثال:

في بيانات التذوق لكي نحصل على مجاميع المربعات الجزئية بواسطة *Minitab* نطبق النموذج بإستخدام جميع المتغيرات ونجعل المتغير المراد حساب مجموع مربعة الجزئي آخر متغير في المعادلة ويكون مجموع مربعه الجزئي هو مجموع مربعه التتابعي كما في المستخرج التالي:

```
Analysis of Variance
Source          DF        SS          MS       F       P
Regression       3       4994.5     1664.8   16.22    0.000
Residual Error 26       2668.4      102.6 = MSE
Total          29       7662.9

TASTE = - 28.9 + 3.91 H2S + 19.7 LACTIC + 0.33 ACETIC
Source       DF      Seq SS
H2S           1      4376.8
LACTIC        1       617.1
ACETIC        1         0.6 = Partial SS for ACETIC

TASTE = - 28.9 + 0.33 ACETIC + 19.7 LACTIC + 3.91 H2S
Source       DF      Seq SS
ACETIC        1      2314.1
LACTIC        1      1672.7
H2S           1      1007.7 = Partial SS for H2S

TASTE = - 28.9 + 0.33 ACETIC + 3.91 H2S + 19.7 LACTIC
Source       DF      Seq SS
ACETIC        1      2314.1
H2S           1      2147.1
LACTIC        1       533.3 = Partial SS for LACTIC
```

ومنها نحصل على الجدول التالي:

Source	*df*	*PSS*	*MS*	*F Value*	*Pr > F*
ACETIC	1	0.6	0.6	0.6/102.6 = 0.006	0.942
H2S	1	1007.7	1007.7	1007.7/102.6 = 9.82	0.0042
LACTIC	1	533.3	533.3	533.3/102.6 = 5.2	0.0311

ومنها نستطيع إختبار تأثير وضع أي متغير كآخر متغير في النموذج.

إستنتاجات عن المعالم المفردة في الإنحدار المتعدد *Inference for Individual Parameters in ML*:

في نموذج الإنحدار الخطي المتعدد

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

لقيم $i = 1,2,...,n$ حيث $\varepsilon_i \sim iid\ N\left(0,\sigma^2\right)$ قد يساعد تكوين فترة ثقة أو إختبار فرضية لمعالم واحد β_j في إعطاء فكرة عن أهمية المتغير المستقل x_j في النموذج الكامل.

فترات ثقة *Confidence Intervals*:

بما أن $\hat{\beta}_j \sim N\left(\beta_j, s_{jj}\sigma^2\right)$ حيث $s_{jj} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}$ العنصر j على قطر المصفوفة $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ لقيم $j = 1,2,...,k$ فإن

$$z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s_{jj}\sigma^2}} \sim N\left(0,1\right)$$

و

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s_{jj}\,\mathrm{MSE}}} \sim t_{n-k-1}$$

فترات الثقة وإختبار الفرضيات تعتمد على هذه القيمة المحورية و هكذا $100(1-\alpha)\%$ فترة ثقة للمعلم β_j يعطى بالعلاقة

$$\hat{\beta}_j \pm t_{n-k-1,\alpha/2}\sqrt{s_{jj}\,\mathrm{MSE}}$$

إختبارات فرضيات *Hypothesis Tests*:

لكي نختبر

$$H_0: \quad \beta_j = \beta_{j,0}$$

ضد أحد البدائل

$$H_1: \quad \begin{cases} \beta_j \neq \beta_{j,0} \\ \beta_j > \beta_{j,0} \\ \beta_j < \beta_{j,0} \end{cases}$$

نستخدم الإحصائة

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s_{jj}\,\mathrm{MSE}}} \sim t_{n-k-1}$$

ومنطقة الرفض تقع في الذيل أو الذيول المناسبة لتوزيع t_{n-k-1}.

ملاحظة: لكي نقيم فيما إذا كان x_j مفيد في وصف y مع وجود كل المتغيرات المستقلة الأخرى في النموذج نستطيع إختبار $H_0: \beta_j = 0$ ضد $H_1: \beta_j \neq 0$ ومن المهم أن نلاحظ أن إختبارات من هذا النوع هي إختبارات **<u>شرطية</u>** أي تشترط وجود المتغيرات المستقلة الأخرى في النموذج.

ملاحظة: الإختبارات السابقة يمكن إجرائها بشكل افضل عن طريق النموذج الكامل والنموذج المختصر كما اوردنا سابقا. فلإختبار $H_0 : \beta_j = 0$ يكون الإختبار

$$H_0 : \quad y_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_{l-1} x_{i(l-1)} + \gamma_{l+1} x_{i(l+1)} + \cdots + \gamma_k x_{ik} + \varepsilon_i$$

ووجدنا أن مجاميع التربيع الجزئية تعطي إختبارا من هذا النوع عن طريق الإحصائة

$$F_j = \frac{R\left(\beta_l \middle| \beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right)}{MSE}$$

والتي هي موزعة تحت الفرضية الصفرية $F_j \sim F_{1,n-k-1}$.

علاقة جبرية:

بمساواة F_j و

$$t^2 = \left(\frac{\hat{\beta}_j - 0}{\sqrt{s_{jj} MSE}} \right)^2$$

نجد

$$Partial\ SS\ for\ x_j = R\left(\beta_j \middle| \beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right) = \frac{\hat{\beta}_j^2}{s_{jj}}$$

حيث $s_{jj} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}$ لقيم $j = 1,2,...,k$.

مثال:

في مثال تذوق الجبن لنعتبر النموذج الكامل

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

لتقييم أهمية تركيزسلفات الهيدرون H2S وتأثيرة على التذوق نستطيع إختبار $H_0 : \beta_2 = 0$ ضد $H_1 : \beta_2 \neq 0$ وذلك بأخذ $\hat{\beta}_2 = 3.9$ و $s_{22} = 0.015$ و $MSE = 102.63$ نجد

$$t = \frac{\hat{\beta}_2 - 0}{\sqrt{s_{22} MSE}} = \frac{3.9}{\sqrt{0.015 \times 102.63}} = 3.13$$

ونجد أنها أكبر من $t_{26,0.025} = 2.056$ وهكذا عند $\alpha = 0.05$ مستوى معنوية فإننا نجد دليل معنوي على أن تركيز سلفات الهيدروجين بعد التعديل لتأثيرات تركيزات كل من *ACETIC* و *LACTIC* مهم في وصف التذوق.

ملاحظة: وجدنا في مثال سابق أن

$$F = \frac{R\left(\beta_2 \middle| \beta_0, \beta_1, \beta_3\right)}{MSE} = \frac{1007.69}{102.63} = 9.82$$

لاحظ أن $t^2 = (3.13)^2 \simeq F$ لاحظ أيضا أن

$$\frac{\hat{\beta}_2^2}{s_{22}} = \frac{(3.912)^2}{0.0152} \simeq 1007.69 = R\left(\beta_2 \middle| \beta_0, \beta_1, \beta_3\right) = Partial \;\; SS \;\; for \;\; x_2$$

فترات ثقة لـ $E\left(y|x\right)$ وفترات تنبؤ لقيمة مستقبلية لـ y في الإنحدار المتعدد:

في نموذج الإنحدار الخطي المتعدد

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

لقيم $i = 1,2,...,n$ حيث $\varepsilon_i \sim iid\ N\left(0,\sigma^2\right)$ أو تكافئيا $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ حيث $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0},\sigma^2\mathbf{I}\right)$ هدفنا الحصول على $100(1-\alpha)$ في المئة فترة ثقة لـ

$$E\left(y|x_0\right) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}$$

أي متوسط الإستجابة y عندما $\mathbf{x}' = \mathbf{x}_0' \equiv \left(x_{01}, x_{02}, ..., x_{0k}\right)$.

مقدر المربعات الدنيا: لنعرف $\mathbf{a}' = \left(1, \mathbf{x}_0'\right) = \left(1, x_{01}, x_{02}, ..., x_{0k}\right)$ لاحظ أن

$$E\left(y|x_0\right) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k} = \mathbf{a}'\boldsymbol{\beta} \equiv \theta$$

لتقدير $\mathbf{a}'\boldsymbol{\beta}$ نستخدم $\hat{\theta} = \mathbf{a}'\hat{\boldsymbol{\beta}}$ حيث $\hat{\boldsymbol{\beta}}$ مقدر المربعات الدنيا للمعلم $\boldsymbol{\beta}$.

تحت فرضيات النموذج من السهل أن نبين أن

$$\hat{\theta} = \mathbf{a}'\hat{\boldsymbol{\beta}} \sim N\left\{\theta, \sigma^2 \mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}\right\}$$

(برهن ذلك) وأن

$$t = \frac{\hat{\theta} - \theta}{\sqrt{MSE\ \mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}}} \sim t_{n-k-1}$$

وهكذا نستطيع إستخدام t كقيمة محورية لإشتقاق فترة ثقة ولإختبار فرضيات حول $\theta = \mathbf{a}'\boldsymbol{\beta}$. فترة $100(1-\alpha)\%$ ثقة لـ $\theta = \mathbf{a}'\boldsymbol{\beta} = E\left(y \mid x_0\right)$ تعطى بالعلاقة

$$\hat{\theta} \pm t_{n-k-1,\alpha/2}\sqrt{MSE\ \mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}}$$

وإختبار مستوى α من الشكل $H_0 : \theta = \theta_0$ ضد بديل بجهة أو جهتين يستخدم

$$t = \frac{\hat{\theta} - \theta_0}{\sqrt{MSE\ \mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}}}$$

كإحصائة إختبار. منطقة الرفض تكون في الذيل المناسب لتوزيع t_{n-k-1}.

مثال:

لبيانات تذوق الجبن نريد إيجاد 95% فترة ثقة لـ $E\left(y \mid x_0\right) = \beta_0 + 5\beta_1 + 6\beta_2 + \beta_k$

هنا $\mathbf{x}_0 = (5,6,1)$ وهكذا $\mathbf{a}' = (1,5,6,1)$ فيكون

$$\hat{E}\left(y\,|x_0\right)=\hat{\theta}=\mathbf{a}'\hat{\boldsymbol{\beta}}=(1,5,6,1)\begin{pmatrix}-28.88\\0.33\\3.91\\19.67\end{pmatrix}\simeq 15.9$$

و

$$\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}=(1,5,6,1)\begin{pmatrix}3.79&-0.76&0.09&-0.07\\-0.76&0.19&-0.02&-0.13\\0.09&-0.02&0.02&-0.05\\-0.07&-0.13&-0.05&0.73\end{pmatrix}\begin{pmatrix}1\\5\\6\\1\end{pmatrix}\simeq\sim 0.18$$

وهكذا فترة ثقة 95% للمقدار $E\left(y\,|x_0\right)=\beta_0+5\beta_1+6\beta_2+\beta_k$ تعطى بواسطة

$$15.9\pm 2.056\sqrt{102.63\times 0.18}$$

أو $(7.3,24.5)$ وهكذا عندما تكون $x_{01}=5$ و $x_{02}=6$ و $x_{03}=1$ فإننا واثقين بـ 95% أن متوسط معدل التذوق $E\left(y\,|x_0\right)$ يكون بين 7.3 و 24.5 .

فترات تنبؤ *Prediction Intervals*:

للكمية $\hat{\theta}=\mathbf{a}'\hat{\boldsymbol{\beta}}$ من السهل بيان أن $100(1-\alpha)\%$ فترة تنبؤ لقيمة y عندما $\mathbf{x}'=\mathbf{x}_0'\equiv\left(x_{01},x_{02},...,x_{0k}\right)$ تعطى بالعلاقة

$$\hat{\theta}\pm t_{n-k-1,\alpha/2}\sqrt{MSE\left\{1+\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}\right\}}$$

هنا $\mathbf{a}_0 = (1, \mathbf{x}_0) = (1, x_{01}, x_{02}, ..., x_{0k})$. لاحظ الواحد الصحيح المضاف تحت الجزر.

مثال:

للتذوق نريد فترة تنبؤ 95% للإستجابة y عندما $\mathbf{x}_0 = (5,6,1)$ أي $\mathbf{a}' = (1,5,6,1)$ وتعطى بالعلاقة

$$15.9 \pm 2.056\sqrt{102.63 \times (1 + 0.18)}$$

أو $(-6.62, 38.43)$ و وهكذا عندما تكون $x_{01} = 5$ و $x_{02} = 6$ و $x_{03} = 1$ فإننا واثقين بـ 95% أن معدل التذوق y يكون بين 6.62– و 38.43.

ملاحظة: القيمة السالبة غير منطقية!

تحليل البواقي وتشخيص النموذج *Residual Analysis and Model Diagnostics* :

في نموذج الإنحدار الخطي المتعدد

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

لقيم $i = 1,2,...,n$ حيث $\varepsilon_i \sim iid\ N(0, \sigma^2)$ أو في الشكل المصفوفي $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ حيث $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

يلاحظ أننا وضعنا بعض الإدعائآت عن تركيب الأخطاء. لذلك من المهم التأكد من هذه الإدعائات مثل التوزيع الطبيعي وثبات التباين ألخ عن طريق تحليل البواقي.

مصفوفة التقدير أو مصفوفة الهات *Hat Matrix*:

سبق أن عرفنا البواقي على أنها القيم المشاهدة ناقص القيم المطبقة أي $e_i = y_i - \hat{y}_i$ حيث y_i القيمة المشاهدة رقم i و $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$ القيمة المطبقة لها أي القيمة المحسوبة من النموذج. وبالشكل المصفوفي $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ حيث $\mathbf{y}$ متجة القيم المشاهدة و $\hat{\mathbf{y}}$ متجه القيم المطبقة والذي يكتب على الشكل

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\,\mathbf{y}$$

المصفوفة

$$\mathbf{M} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

تسمى مصفوفة التقدير أو مصفوفة هات *Hat Matrix* ويرمز لها أحينا $\mathbf{H}$.

حقائق:

المصفوفة $\mathbf{M}$ لها عناصر على المحور الرئيسي

$$h_{ii} = \mathbf{x}_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i$$

حيث $\mathbf{x}_i'$ الصف i في مصفوفة التصميم $\mathbf{X}$. القيمة h_{ii} تسمى **<u>الرفع أو الرفعية</u>** ***<u>Leverage</u>*** للحالة i حيث $i = 1,2,...,n$. يمكن إثبات أن

$$E(\mathbf{e}) = \mathbf{0}$$

و

$$cov(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{M}) = \begin{pmatrix} \sigma^2(1-h_{11}) & -\sigma^2 h_{12} & \cdots & -\sigma^2 h_{1n} \\ -\sigma^2 h_{21} & \sigma^2(1-h_{22}) & \cdots & -\sigma^2 h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\sigma^2 h_{n1} & -\sigma^2 h_{n2} & \cdots & \sigma^2(1-h_{nn}) \end{pmatrix}$$

وهكذا فإن التباين للباقي رقم i هو $V(e_i) = \sigma^2(1-h_{ii})$ لقيم $i = 1,2,...,n$.

تحت فرضيات نموذجنا يكون $e_i \sim N\{0, \sigma^2(1-h_{ii})\}$.

بخلاف أخطاء جاوس - ماركوف والتي يفترض أن لها تباين ثابت الأخطاء الناتجة عن طريق التقدير بواسطة المربعات الدنيا ليس لها تباين ثابت كما أن هذه الأخطاء مترابطه.

تشخيص التباينات غير الثابتة و القصور في النموذج *Diagnosing Nonconstant Variance and other Model Inadequacies* :

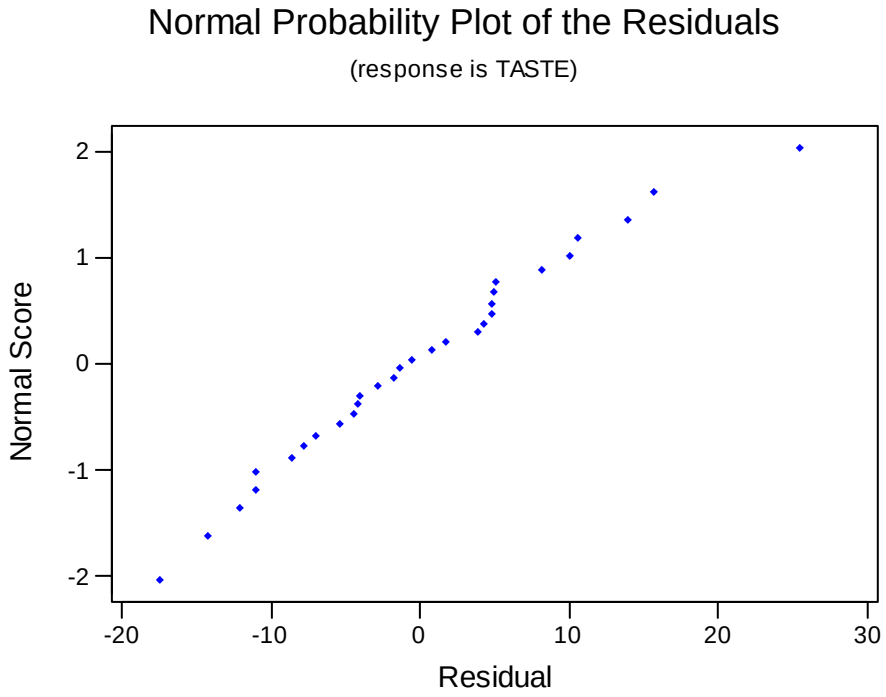
لتشخيص عدم ثبات التباين و عدم التحديد *Misspecification* في النموذج نستخدم وسيلة بصرية مثل رسم البواقي الناتجة ضد القيم المطبقة أي رسم e_i ضد $\hat{y}_i$. إذا كان النموذج صحيح فإن $cov(\mathbf{e}, \hat{\mathbf{y}}) = \mathbf{0}$ أي $cov(e_i, \hat{y}_i) = 0$ لجميع قيم $i = 1,2,...,n$ وهذا يعني أن البواقي والقيم المطبقة غير مترابطة ولهذا فإن رسومات البواقي ضد القيم المطبقة والتي تعطي أنماط غير عشوائية تعطي إشارة لعدم بوجود مشاكل في فرضيات النموذج.

مثال:

سوف نحلل البواقي لبيانات التذوق الناتجة من تطبيق النموذج الكامل

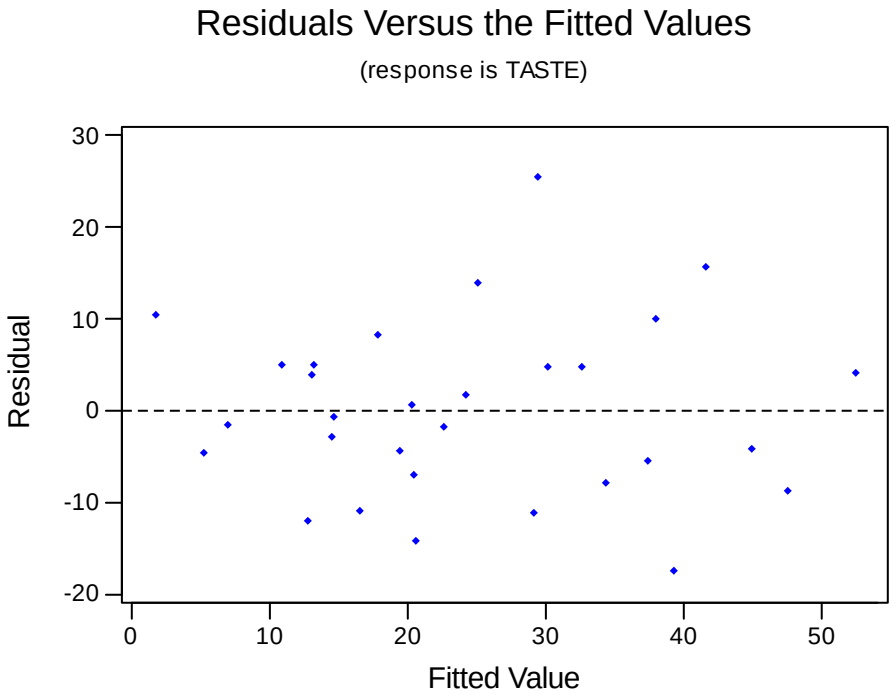
```
TASTE = - 28.9 + 0.33 ACETIC + 3.91 H2S + 19.7 LACTIC
```

الشكل التالي يعطي رسومات الإحتمالات الطبيعية Normal Probability Plots أو تسمى أحيانا qq - plot



ويبدو منها ان فرضية التوزيع الطبيعي لايشكل مجال للقلق.

الرسم التالي يعطي رسومات البواقي



وكما نلاحظ فإنه لايوجد أي إقتراح لعدم تحديد أو عدم ثبات في التباين.

كشف الخوارج وتحليل النفوذ *Outlier Detection and Influence Analysis*:

كشف الخوارج (القيم الشاذة): نستخدم البواقي للمساعدة في تقييم فيما إذا كانت حالة خاصة من الحالات المشاهدة هي خارجة *Outlier*. تحت فرضيات النموذج البواقي $e_i \sim N\left\{0,\sigma^2\left(1-h_{ii}\right)\right\}$ حيث h_{ii} هو العنصر i على المحور الرئيسي في المصفوفة **M** أو الرفعية للنقطة i . لكي نتغلب على مشكلة عدم تساوي التباينات نحسب البواقي المتلمذة *Studentized Residuals* (تسمي متلمذة وذلك على إسم مكتشف توزيع t والذي كان ينشر أبحاثة تحت توقيع "تلميذ") لكي نحصل على تباينات متساوية.

1. **بواقي متلمذة داخلية *Internally Studentized Residuals*** :

$$r_i = \frac{e_i}{\sqrt{s^2\left(1-h_{ii}\right)}}$$

حيث $s^2 = MSE$ تحسب من جميع قيم البيانات. من السهل إثبات أن $E\left(r_i\right)=0$ و $V\left(r_i\right)\simeq 1$. قيم $\left|r_i\right|$ اكبر من 3 تستدعي التحقق منها.

2. **بواقي متلمذة خارجية *Externally Studentized Residuals***:

$$t_i = \frac{e_i}{\sqrt{s_{-i}^2\left(1-h_{ii}\right)}}$$

حيث $s_{-i}^2 = MSE$ تحسب من كل البيانات **<u>ماعدى</u>** الحالة i. يمكن إثبات أن

$$s_{-i}^2 = \frac{\left(n-k\right)s^2 - e_i^2/\left(1-h_{ii}\right)}{n-k-1}$$

تحت فرضيات نموذج الإنحدار المتعدد تكون $t_i \sim t_{n-k-1}$ وهكذا عند مستوى معنوية α يمكننا تصنيف المشاهدة i كخارجة إذا كانت $\left|t_i\right| \geq t_{n-k-1,\alpha/2n}$.

مثال:

في مثال تذوق الجبن لدينا $n = 30$ حالة أو مشاهدة وتصنف المشاهدة على انها خارجة عند مستوى معنوية $\alpha = 0.05$ إذا كان $|t_i| \geq t_{26,0.000833} = 3.51$ ولكن من المخرجات من *Minitab* نجد

```
TRES1
 1.14909  -0.17502   1.45150   1.03122  -0.14710
 0.19153   0.51757  -1.87803   0.50140   0.07755
 0.48512   1.73956  -0.46995   0.81793   3.01551
-0.42633   0.51873  -1.47907  -1.14634  -0.98723
-0.06280  -0.42027  -0.58929   0.46860   0.38716
-0.31871  -0.83076  -1.24793  -0.76771  -1.23299

 Maximum of TRES1 = 3.0155
```

وحيث أن 3.0155 أقل من 3.51 نستطيع أن نقول انه لايوجد دليل على وجود حالة خارجة بإستخدام هذا المعيار.

تعريف: في تحليل الإنحدار يقال عن حالة أنها **نافذة أو مؤثرة *Influential*** إذا تسبب إزالتها من النموذج في تغير كبير في التحليل (مثل تغير كبير في تقدير معاملات الإنحدار أو في جدول تحليل التباين أو في قيمة R^2 الخ). القيمة النافذة أو المؤثرة لايحتاج أن تكون قيمة خارجة (أو العكس). ولكن في معظم الأحيان فإن المشاهدات الخارجة تكون نافذة.

مسافة كوك *Cook's Distance*: لقياس نفوذ الحالة i اقترح العالم كوك (1997) الإحصائة التالية

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}\right)' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}\right)}{(k+1)\ \mathrm{MSE}} = \frac{r_i^2}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

حيث $\hat{\boldsymbol{\beta}}_{-i}$ مقدر المربعات الدنيا للمعالم $\boldsymbol{\beta}$ بعد إزالة الحالة i من النموذج و MSE يحسب من كامل النموذج شاملا جميع الحالات. القيم الكبيرة للإحصاءة D_i تتبع لقيم نافذة أو مؤثرة. وهناك قاعدة غير رسمية تقول أن نصنف مشاهدة كمشاهدة نافذة إذا كانت D_i لها $D_i \geq 4/n$. هناك أنواع أخرى من الإحصائآت لتشخيص النفوذ مثل *DFITS* و *DFBETAS* و مقياس هادي للنفوذ *Hadi's Influence Measure* الخ ولكن مسافة كوك هي الأكثر إستخدام.

تعددية خطية مشتركة *Multicollinearty*:

وتختصر خطية مشتركة *Collinearity* وتحدث عند وجود علاقة خطية أو قريبة من الخطية بين إثنان أو اكثر من المتغيرات المستقلة $x_1, x_2, ..., x_k$ في النموذج.

تعريف: مجموعة من متغيرات الإنحدار $x_1, x_2, ..., x_k$ يقال أنها خطية مشتركة تماما *Exactly Collinear* إذا وجدت ثوابت $c_0, c_1, ..., c_k$ (ليست كلها صفر) بحيث

$$\sum_{j=1}^{k} c_j x_j = c_0$$

مثال:

لننظر البيانات التالية:

x_1	x_2	x_3
1	8	7
2	5	3
6	10	4

هذه المتغيرات خطية مشتركة تماما لأن $\sum_{j=1}^{3} c_j x_j = c_0$ بأخذ $c_1 = -1$ و $c_2 = 1$ و $c_3 = -1$ و $c_0 = 0$ أي أن $x_3 = x_2 - x_1$.

فإذا كانت مجموعة من المتغيرات $x_1, x_2, ..., x_k$ خطية مشتركة تماما فإن هذا يعني أنه يوجد واحدة x_j والتي يمكن كتابتها كتركيب خطي من x's الأخرى أي

$$c_j x_j = c_0 - c_1 x_1 - \cdots - c_{j-1} x_{j-1} - c_{j+1} x_{j+1} - \ldots - c_k x_k$$

$$\Rightarrow x_j = \frac{c_0}{c_j} - \frac{c_1}{c_j} x_1 - \cdots - \frac{c_{j-1}}{c_j} x_{j-1} - \frac{c_{j+1}}{c_j} x_{j+1} - \ldots - \frac{c_k}{c_j} x_k$$

وهكذا في هذه الحالة فإن x_j لاتضيف أي معلومات جديدة للإنحدار ليست موجودة مسبقا في المتغيرات الأخرى. في هذه الحالة النادرة مصفوفة التصميم $\mathbf{X}$ لايكون لها رتبة كاملة والمصفوفة $\mathbf{X'X}$ لايوجد لها مقلوب وبالتالي لانستطيع تقدير $\boldsymbol{\beta}$ بشكل وحيد.

ملاحظة: عند وجود خطية مشتركة تقريبا (والتي توجد كثيرا تطبيقيا) أي يوجد ثوابت $c_0, c_1, ..., c_k$ (ليست كلها صفر) بحيث

$$\sum_{j=1}^{k} c_j x_j \simeq c_0$$

وعندها يقال أن متغيرات الإنحدار خطية مشتركة تقريبيا. إذا كانت متغيرات الإنحدار خطية مشتركة تقريبا ولكنها غير خطية مشتركة بشكل كامل فإن $(\mathbf{X}'\mathbf{X})^{-1}$ لاتكون موجودة ولكننا نستطيع حساب $\hat{\boldsymbol{\beta}}$ بشكل وحيد ومع هذا فإن $V\left(\hat{\beta}_j\right)$ سيكون كبير جدا والذي سيجعل مقدر $\hat{\beta}_j$ أقل دقة. نعلم أن $V\left(\hat{\beta}_j\right)=s_{jj}\sigma^2$ حيث $s_{jj}=(\mathbf{X}'\mathbf{X})^{-1}_{jj}$. فإذا كان لدينا خطية مشتركة تقريبا فإن الحدود s_{jj} تصبح كبيرة وهذا بدوره يضخم $V\left(\hat{\beta}_j\right)$ ويمكن تبيان أن

$$V\left(\hat{\beta}_j\right)=\sigma^2\left(\frac{1}{1-R_j^2}\right)\left\{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{+j}\right)^2\right\}^{-1}$$

حيث R_j^2 معامل التحديد الناتج من إنحدار x_j على المتغيرات الأخرى. كلما تكون درجة الخطية المشتركة بين x_j و بقية المتغيرات كبيرة كلما كانت R_j^2 كبيرة. وعندما R_j^2 تقترب من واحد تكبر $V\left(\hat{\beta}_j\right)$ بدون حدود. وهكذا فإن تعددية الخطية المشتركة يمكن أن تؤثر وبشكل كبير في تضخيم تباين مقدرات المربعات الدنيا. وهذا بدوره سوف يكون له تأثير على نوعية التطبيق بالمربعات الدنيا ومن ثم سوف يؤثر على دقة فترات الثقة وعلى إختبار الفرضيات الخ.

طرق قياس الخطية المشتركة:

1. أسهل طريقة لتشخيص الخطية المشتركة هي حساب ترابط العينة الثنائي بين المتغيرات. أي حساب $r_{jj'}$ لجميع $j\neq j'$ و $j=1,2,...,k$. ولكن هذه الطريقة غير مضمونة لكشف الخطية المشتركة لأن بعض الخطية المشتركة قد يتضمن ثلاثة أو أكثر

من المتغيرات وأي مجموعة جزئية منها تتكون من متغيرين قد لاتكشف الخطية المشتركة.

2. حساب **عامل تضخم التباين (*Variance Inflation Factor* (*VIF*)):**

ويعرف بالعلاقة

$$VIF_j = \frac{1}{1 - R_j^2}$$

لقيم $j = 1,2,...,k$. قيم VIF_j التي تكون أكبر من 10 مؤشر لخطية مشتركة قوية. هذا المقياس قد لايعمل بشكل جيد لأن R_j^2 حساسة بالنسبة للقيم الخارجة.

مثال:

عامل تضخم التباين للمتغيرات *ACETIC* و *H2S* و *LACTIC* من مخرجات *Minitab*

```
The regression equation is
TASTE = - 28.9 + 0.33 ACETIC + 3.91 H2S + 19.7 LACTIC

Predictor      Coef    SE Coef       T      P     VIF
Constant     -28.88    19.74     -1.46  0.155
ACETIC         0.328    4.460      0.07  0.942   1.8
H2S            3.912    1.248      3.13  0.004   2.0
LACTIC        19.670    8.629      2.28  0.031   1.9
```

هي على التوالي (من العامود الأخير تحت *VIF*) 1.8 و 2.0 و 1.9 وهي ليست كبيرة بشكل يثير الشك في وجود خطية مشتركة.

معيار إختيار أفضل نموذج *Criteria for Choice of Best Model*:

هناك معايير عدة لإختيار أفضل نموذج من بين عدة نماذج مقترحة لوصف بيانات معطاة في حالة الإنحدار الخطي البسيط أو المتعدد وهي:

1. معامل التحديد ويعطي النسبة من التغير الكلي في البيانات والذي تم تفسيره بواسطة النموذج ويعطى بالعلاقة:

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

ومن البديهي القيم الكبيرة لـ R^2 توحي أن النموذج قد أسترد كثير من التغير في البيانات عن طريق الإنحدار.

2. متوسط مربع الخطأ MSE وهو المقدر غير الحيازي للتباين الخطأ σ^2 وطبعا يختار النموذج الذي يعطي أقل متوسط مربع خطأ.

3. معيار المعلومات الذتي AIC ويعطى بالعلاقة

$$AIC = ln\left(\hat{\sigma}^2\right) + \frac{n + 2k}{n}$$

حيث n عدد المشاهدات و k عدد المتغيرات المستقلة ويؤخذ النموذج الذي يعطي أقل قيمة للمعيار.

a3. معيار المعلومات الذاتي غير الحيازي $AICc$ ويعطى بالعلاقة

$$AICc = ln\left(\hat{\sigma}^2\right) + \frac{n + k}{n - k - 2}$$

ويؤخذ النموذج الذي يعطي أقل قيمة للمعيار.

4. معيار المعلومات لشفارتز *SIC* ويعطى بالعلاقة

$$SIC = ln\left(\hat{\sigma}^2\right) + \frac{k\ ln(n)}{n}$$

ويؤخذ النموذج الذي يعطي أقل قيمة للمعيار.

ملاحظة:

R^2 لايمكن أن تتناقص أبدا (وغالبا ما تزداد) بإضافة متغير مستقل إضافي حتى لو كان هذا المتغيرالإضافي لايضيف أي تفسير جديد. أو حتى لو كان هذا المتغير الإضافي ليس له مغزى أو منطق!

R^2 المعدلة *Adjusted*:

السبب في أن R^2 تصبح كبيرة كلما اضيف متغير مستقل للنموذج بعكس *MSE* هو ان هذا المعيار لايوجد عليه عقاب أو تغريم *Penality* بنقصان درجات الحري $n-k-1$ بينما *SSR* يزداد. ولهذا السبب إستخدام R^2 فقط كإحصائة للتفضيل بين النماذج المقترحة خطر جدا! لهذا فإن شكل آخر مطور أو معدل هو *Adjusted* R^2 والذي يأخذ في الإعتبار هذه المشكلة ويعطى بالعلاقة

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

وبالتالي زيادة k تسبب نقص في $n-k-1$ وهكذا R_a^2 تنقص. وبهذا المتغيرات الإضافية لن تزيد بالضرورة قيمة R_a^2.

ملاحظة:

المعايير R_a^2 و MSE ليست بالضرورة ذات منحى تنبئي أي انهم لايعطو إشارة أو تقييم على مقدرة النموذج في أغراض التنبؤ. هناك معيارين لهذا الغرض هما $PRESS$ و مالو C_s.

معيار *PRESS*:

لتكن $\hat{y}_{i,-i}$ تمثل القيمة المطبقة غي الإنحدار بعد إستقصاء أو إبعاد الحالة i أي

$$\hat{y}_{i,-i} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{-i}$$

حيث $\mathbf{x}_i'$ هو السطر i للمصفوفة $\mathbf{X}$ و $\hat{\boldsymbol{\beta}}_{-i}$ مقدر المربعات الدنيا للمعلم $\boldsymbol{\beta}$ بعد إقصاء الحالة i. الكمية $e_{i,-i} = y_i - \hat{y}_{i,-i}$ تسمى الباقي i الـ $PRESS$ وبما اننا نبعد كل مشاهدة واحدة في كل مرة فإن بواقي "برس" هي أخطاء تنبؤ حقيقية لأن $\hat{y}_{i,-i}$ مستقلة عن y_i. طبعا نرغب في كون هذه البواقي صغيرة ولهذا فإن الإحصائة

$$PRESS = \sum_{i=1}^{n} e_{i,-i}^2$$

كل ما تكون صغيرة تؤخذ كمقياس للنموذج الأفضل.

ملاحظة:

قد يبدو إستخدام مقياس برس يحتاج إلى حساب n نموذج إنحدار لحسابه ولكن توجد طريقة سهلة جدا لحسابه من النموذج الكامل كالتالي:

$$PRESS = \sum_{i=1}^{n} e_{i,-i}^{2} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

حيث e_i البواقي المحسوبة من النموذج الذي يحوي كل الحالات و h_{ii} عناصر المحور الرئيسي للمصفوفة **M**.

مالو C_s ***Mallows'***:

المبدأ الأساسي لهذه الإحصائة هو معاقبة الباحث (وضع جزاء) على زيادة التطبيق *Overfitting* (وهو عملية وضع متغيرات مستقلة كثيرة غير ضرورية في النموذج تخالف مبدأ الشح *Parsimony Principle* (أنظر كتاب بناء النماذج للمؤلف)) وتحت التطبيق *Underfitting* (وهو عدم أخذ المتغيرات المستقلة المهمة في النموذج).

مثال:

لنفترض لدينا 6 متغيرات مستقلة $x_1, x_2, ..., x_6$ والمتغير التابع y يتبع في الحقيقة النموذج

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon$$

طبعا في الحقيقة التطبيقية نحن لانعرف النموذج الصحيح ولهذا فإننا نطبق النموذج

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$$

وفي هذه الحالة فإننا قمنا بتحت تطبيق وأما إذا طبقنا النموذج

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

ففي هذه الحالة فإننا قمنا بفوق تطبيق.

وهكذا:

1. إذا أجرينا تحت تطبيق فإن معالم الإنحدار المقدرة و *MSE* ستكون حيازية وبالذات فإن *MSE* سيقدر بزيادة *Overestimate* القيمة الحقيقية σ^2 لأننا تجاهلنا متغيرات مهمة.
2. إذا أجرينا فوق تطبيق فإن معالم الإنحدار و *MSE* ستكون غير محايزة ولكن هناك مخاطرة تضخم التباينات لمعالم الإنحدار نظرا للخطية المشتركة.

مالو C_s تدمج معاقبة كلا من الإنحياز و تضخم التباينات في إحصائة واحدة. ليكن MSE_s متوسط مربع الخطأ لنموذج مقترح بـ $s \le m$ من المتغيرات المتاحة. عندئذ فإن هذا النموذج المقترح له إحصائية C_s تعطى بالتالي:

$$C_s = (s+1) + \frac{(MSE_s - MSE_m)(n-s-1)}{MSE_m}$$

$$C_s = (s+1) + (n-s-1)\left(\frac{MSE_s}{MSE_m} - 1\right)$$

ماهي القيم الجيدة للإحصائة C_s؟ إذا كان النموذج المقترح بـ s من المتغيرات في الحقيقة صحيح فإن كلا من MSE_m و MSE_s تقدر نفس الكمية أي σ^2 وفي هذة الحالة يكون

$$E\left(C_s\right)=\left(s+1\right)+\underbrace{E\left\{\frac{\left(MSE_s-MSE_m\right)\left(n-s-1\right)}{MSE_m}\right\}}_{\simeq 0}\simeq s+1$$

وهكذا فإن قيم $C_s \simeq s+1$ تكون مفضلة. النماذج التي يكون فيها C_s اكبر بكثير من $s+1$ ربما لاتشمل التركيب الصحيح من المتغيرات المستقلة اي يحصل تحت تطبيق. طبعا النموذج الكامل والذي فيه $s=m$ من المتغيرات المستقلة يكون فيه $C_s=s+1$.

مثال:

سوف نقوم بحساب قيم $PRESS$ و R^2 و R_a^2 و MSE و C_p لبيانات التذوق والتي سوف نطبق عليها 7 نماذج ممكنة كالتالي:

Number in Model	PRESS	R-Square	Adjusted R-Square	MSE	C(p)	Variables in Model
1	3687.95	0.5712	0.5559	117.359	6.018	h2s
1	4375.64	0.4959	0.4779	137.946	11.635	lactic
1	6111.25	0.3020	0.2771	191.027	26.116	acetic
2	3135.38	0.6517	0.6259	98.849	2.005	h2s lactic
2	3877.49	0.5822	0.5512	118.579	7.195	acetic h2s
2	4535.47	0.5203	0.4847	136.150	11.818	acetic lactic
3	3402.17	0.6518	0.6116	102.630	4.0000	acetic h2s lactic

والمخرجات من *Minitab*

Best Subsets Regression: TASTE versus ACETIC, H2S, LACTIC

Response is TASTE

Vars	R-Sq	R-Sq(adj)	C-p	S	ACETIC	H2S	LACTIC
1	57.1	55.6	6.0	10.833		X	
1	49.6	47.8	11.6	11.745			X
1	30.2	27.7	26.1	13.821	X		
2	65.2	62.6	2.0	9.9423		X	X
2	58.2	55.1	7.2	10.889	X	X	
2	52.0	48.5	11.8	11.668	X		X
3	65.2	61.2	4.0	10.131	X	X	X

نلاحظ أن افضل نموذج لوصف بيانات التذوق هو الذي يحوي *H2S* و *LACTIC*.

ترابط الأخطاء وإختبار دوربن - واتسون Correlation of Errors and Durbin - Watson Test:

في معالجتنا للإنحدار الخطي بشكل عام كنا نفترض عدم ترابط أو إستقلالية الأخطاء. في كثير من التطبيقات المالية والإقتصادية تكون الأخطاء مترابطة وأحينا تشكل الأخطاء متسلسلة زمنية أي

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t, \quad t = 1,2,...,n$$

حيث المعلم $|\rho|<1$ و $v_t \sim iid\ N\left(0,\sigma^2\right)$.

لإختبار الترابط (أو الترابط الذاتي) أي

$$H_0:\ \rho=0$$

ضد

$$H_a:\ \rho\neq 0$$

نستخدم الإحصائة

$$D=\frac{\sum_{t=2}^{n}\left(e_t-e_{t-1}\right)^2}{\sum_{t=1}^{n}e_t^2}$$

حيث $e_t=y_t-\hat{y}_t$ لقيم $t=1,2,...,n$ البواقي.

طريقة الإختبار الدقيقة غير متوفرة ولكن دوربن و واتسون أوجدا حد أدنى d_L و حد أعلى d_U بحيث أن قيم D التي تقع خارج هذه الحدين تؤدي لقرار محدد فإذا كانت $D>d_U$ فلا نرفض H_0 أما إذا كانت $D<d_L$ نرفض H_0 و إذا كانت $d_L\le D\le d_U$ فإن الإختبار غير حاسم.

طرق إختيار المتغيرات التتابعي *Sequential Variable Selection Procedures*:

الإختيار التتابعي للمتغيرات: طرق إختيار أفضل مجموعة جزئية تعتمد على تقييم كل المجموعات الجزئية من النموذج الكامل والتعرف على أفضل نماذج مختزلة حسب معايير معينة. حساب جميع النماذج الممكنة هي أنسب طريقة لإختيار المتغيرات ولكن الحسابات المترتبة على ذلك مهيلة وقد لاتكون عملية فمثلا إذا كان لدينا 8 متغيرات مستقلة $x_1, x_2, ..., x_8$ فإنه يوجد $2^8 = 256$ نموذج لكي تؤخذ في الإعتبار!

الإختيار الأمامي Forward Selection:

1. إبدأ بنموذج يحوي الحد الثابت (القطع) فقط.
2. إعتبر كل النماذج التي تحوي متغير واحد وإختار النموذج الذي له أكبر t وذلك بمقارنتها مع قيمة معينة مسبقا t_c تسمى القطع *Cutoff Value*.
3. إعتبر كل النماذج التي تحوي متغيرين وذلك بإضافة متغير مستقل جديد للنموذج الذي يحوي متغير واحد اختير حسب الخطوة السابقة وإختار النموذج بمتغيرين الذي فيه المتغير الثاني له أكبر t وذلك طالما $t \geq t_c$.
4. نستمر بهذه الطريقة (حيث الخطوة التالية إعتبار نماذج من ثلاثة متغيرات) وتتوقف العملية عندما لايوجد متغير للإضافة له $t \geq t_c$.

بإستخدام هذا الخوارزم فإن عدد النماذج التي يمكن حسابها تكون على الأكثر $k(k+1)/2$ وهي أقل بكثير من 2^k إذا كانت k كبيرة. قيمة القطع التي تؤخذ عادة هي $t_c = 2$ لأنها تعطي قاعدة شاملة لإختبار له حجم تقريبي $\alpha = 0.05$ لأهمية متغير الإنحدار. وبمجرد إختيار متغير بهذه الطريقة فيجب أن يظل في النموذج.

الإختيار الخلفي *Backward Selection*:

1. إبدأ بالنموذج الكامل الذي يحوي كل المتغيرات k المستقلة.

2. لنعتبر إزالة متغير واحد وذلك بإقصاء المتغير الذي له أصغر t وذلك بمقارنتها مع قيمة معينة مسبقا t_c.

3. ثم أعيد تطبيق نموذج يحوي $k-1$ متغيرمستقل المتبقية ومن ثم نحذف المتغير الذي له أصغر t بحيث $t < t_c$.

4. ونستمر على هذه الطريقة حتى لايتبقى متغير يمكن حذفه له $t < t_c$.

بإستخدام هذا الخوارزم عدد النماذج التي يمكن تطبيقها على الأكثر $k-1$ نموذج وهي أقل بكثير من 2^k إذا كانت k كبيرة. مرة اخرى قيمة نقطة القطع $t_c = 2$ لنفس الأسباب المذكورة أعلاه.

الإختيار الخطوي *Stepwise Selection*:

1. نفس الثلاثة خطوات في الإختيار الأمامي.

2. الآن لننظر في حذف متغير له $t < t_c$ حتى لوكان أول متغير أُضيف . وهذا الذي يجعل الإختيار الخطوي مختلف عن الإختيار الأمامي حيث يمكن حذف متغير من النموذج.

3. بالنموذج الناتج في الخطوة السابقة والذي قد يحوي متغير أو إثنين نضيف متغير له أكبر $t \geq t_c$.

4. بالنموذج الناتج في الخطوة السابقة والذي قد يحوي متغيرين أو ثلاثة نحذف أي متغير له أصغر $t < t_c$.

5. ونستمر حتى تنتهي جميع المتغيرات المقترحة.

تمرين: قم بإجراء الطرق الثلاثة السابقة على بيانات التذوق وأوجد أفضل نموذج يصف البيانات.

مثال على الرفعية Example of Leverage:

سبق أن عرفنا الرفع أو الرفعية لنقطه (x_i, y_i) بـ $h_{ii} = \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i$ حيث $\mathbf{x}_i'$ الصف i في مصفوفة التصميم $\mathbf{X}$. حيث $i = 1,2,...,n$

ويمكن إثبات أنها تساوي $h_{ii} = (1/n) + \left\{ (x_i - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}$

تكتب $\hat{\beta}_1$ ميل خط الإنحدار كالتالي: $\hat{\beta}_1 = \sum_{i=1}^{n} p_i \left\{ (y_i - \bar{y}) / (x_i - \bar{x}) \right\}$ حيث $p_i = h_{ii} - (1/n)$ لاحظ أن $\sum_{i=1}^{n} p_i = 1$ (ملاحظة: $p_i = h_{ii} - (1/n)$ يطلق عليها أيضا الرفعية) أحد الخطوط التي يمكن أن يكون لها ميل $(y_i - \bar{y}) / (x_i - \bar{x})$ تمر خلال (x_i, y_i) و $(\bar{x}, \bar{y})$ أي أن ميل خط الإنحدار هو متوسط موزون لميل خطوط تمر بين كل نقطة والنقطة المتوسطة $(\bar{x}, \bar{y})$.

1- الأوزان في التركيب الموزون هي الرفعيات.

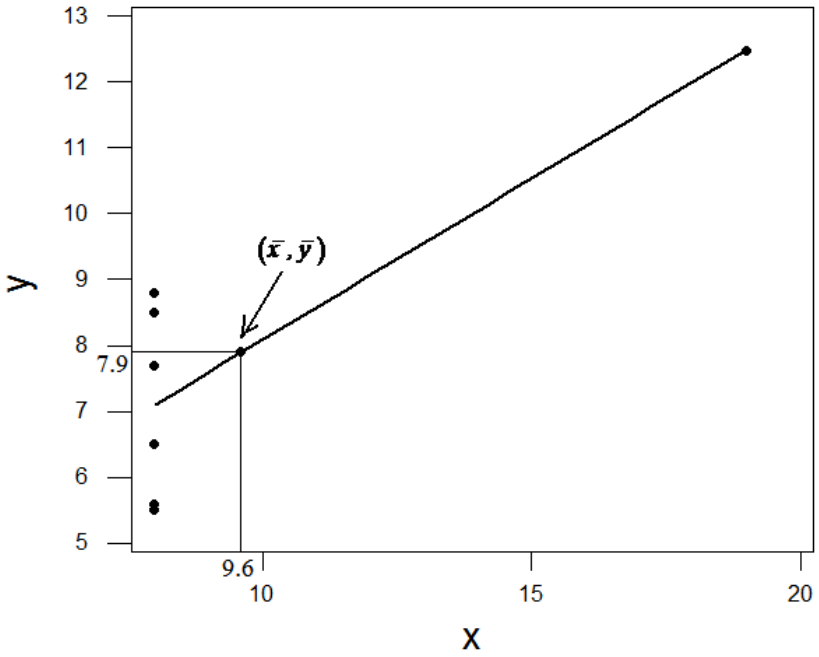
2- إذا كان لنقطة رفعية كبيرة فإن ميل خط الإنحدار يقترب أكثر من الخط الممتد من تلك النقطة والنقطة المتوسطة. مما يعني ان في الإنحدار أن النقط التي لها رفعية كبيرة هي من النقاط المهمة. والنقاط التي رفعيتها صغيرة لايحسب حسابها في الإنحدار إذ يمكن تحريكها او إزاحتها من البيانات ومعادلة خط الإنحدار بدون ترك تأثير يذكر.

مثال: البيانات التالية وضعت خصيصا لتوضيح الرفعية

y	6.5	5.8	7.7	8.8	8.5	12.5	5.6
x	8	8	8	8	8	19	8

خط الإنحدار هو

$y = 3.26 + 0.49\,x$



لاحظ أن خط الإنحدار قريب جدا من الخط الذي يمر بالنقطة (19,12.5) والنقطة (9.6,7.9) (النقطة المتوسطة $(\bar{x},\bar{y})$).

x_i	8	8	8	8	8	19	8	Sum
$(x_i - \bar{x})^2$	2.5	2.5	2.5	2.5	2.5	88.9	2.5	103.9
p_i	0.024	0.024	0.024	0.024	0.024	0.856	0.024	1.000

$\bar{x} = 9.571$

إذا حركنا واحدة من النقاط حيث $x = 8$ والتي لها رافعية صغيرة فإن خط الإنحدار لن يتغير إلا بمقدار ضئيل. فمثلا إذا الغينا النقطة الولى (8,6.5) وأعدنا تطبيق خط الإنحدار للنقاط 6 الاخرى سنجد أن خط الإنحدار يصبح $y = 3.48 + 0.47\,x$ وإذا ازحنا النقطة

(8,6.5) إلى (12,6.5) فيصبح خط الإنحدار $y = 3.58 + 0.43\ x$ ولكن إذا أزحنا (19,12.5) إلى (12,12.5) فيصبح خط الإنحدار $y = -\ 3.55 + 1.34\ x$.

تمرين: يترك للطالب التحقق من ذلك من البيانات ورسم الأشكال المناسبة وتوضيح الفرق بينهما.

مثال على المشاهدات النافذة أو المؤثرة Influential Observations:

طريقة أخرى لقياس أهمية مشاهدة عند حساب خط الإنحدار هي إجراء التالي:

1- حساب قيمتها المطبقة.

2- تزاح المشاهدة ويطبق خط إنحدار بدونها وتحسب قيمتها المطبقة من الخط الجديد.

3- ننظر إلى الفرق التربيعي بين القيمتين المطبقتين ممعيرة بالتباين في القيمة المطبقة للنموذج الكامل.

إذا كانت القيمة المشاهدة مهمة أو نافذة فسنرى قيمة كبيرة في الفرق الممعير.

هذه العملية هي لحساب الإحصائة التالية والتي سبق أن عرفناها بمسافة كوك

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}\right)' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}\right)}{(k+1)\ \mathrm{MSE}} = \frac{r_i^2}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

حيث $\hat{\boldsymbol{\beta}}_{-i}$ مقدر المربعات الدنيا للمعالم $\boldsymbol{\beta}$ بعد إزالة الحالة i من النموذج و MSE يحسب من كامل النموذج شاملا جميع الحالات. القيم الكبيرة للإحصاءة D_i تتبع لقيم نافذة أو مؤثرة. وهناك قاعدة غير رسمية تقول أن نصنف مشاهدة كمشاهدة نافذة إذا كانت D_i لها $D_i \geq 4/n$.

يمكن أيضا إثبات أن مسافة كوك والتي تستخدم لقياس النفوذ لنقطة (x_i, y_i) تعطى للإنحدار الخطي البسيط بالعلاقة:

$$D_i = \frac{\left(\hat{y}_i - \hat{y}_{(i)}\right)^2}{2MSE\left(h_{ii} - \frac{1}{n}\right)}$$

حيث $\hat{y}_{(i)}$ القيمة المطبقة لـ y_i عند إزاحة النقطة (x_i, y_i).

مثال: البيانات التالية وضعت خصيصا لتوضيح القيم النافذة ومسافة كوك

y	6.5	5.8	7.7	8.8	8.5	12.5	5.6
x	12	8	8	8	8	19	8

خط الإنحدار هو

$y = 3.58 + 0.43\,x$

عند $x = 19$ فإن $y_6 = 11.7$ و $MSE = 2.98$ و $p_6 = 0.86$.

نزيح هذه النقطة فيصبح خط الإنحدار

$y = 8.84 - 0.2\,x$

عند $x = 19$ فإن $y_{(6)} = 5.14$ ومسافة كوك لها 7.23 وبتكرار هذا لجميع النقاط

x_i	12	8	8	8	8	19	8
y_i	6.5	5.8	7.7	8.8	8.5	12.5	5.6
D_i	<u>2.05</u>	0.08	0.03	0.17	1.12	<u>7.23</u>	0.10

النقاط ذات الرفعية العالية ليست بالضرورة لها مسافة كوك معنوية.

النقاط النافذة مهمة جدا لأنها تؤثر بشكل كبير على خط الإنحدار. ويجب أن تفحص بعناية ويتحقق عنها فقد تكون نقطة خارجة ولأن قيمتها تؤثر كثيرا على الإنحدار فيجب أن نكون متأكدين أن قيمتها صحيحة.

وجود قيمة أو قيم نافذة قد يعني:

1- النموذج صحيح ولكن يوجد خطأ في قياس النقطة النافذة.

2- قيمة النقطة النافذة صحيحة ولكن النموذج غير صحيح حيث ان لم يستطيع نمذجة النقطة بشكل جيد.

فإذا كان النموذج غير صحيح فقد يكون هذا بسبب:

1- العلاقة بين x و y ليست خطية في فترة من قيم x التي تشمل هذه النقطة النافذة.

2- عدم تساوي تباينات الخطأ *Heteroscedasticity* .

3- يوجد متغير مستقل آخر لـ y والذي يأخذ قيمة مختلفة للقيمة النافذة.

الجزء العملي على نماذج الإنحدار الخطي:

مثال (1):

إنحدار خطي بسيط:

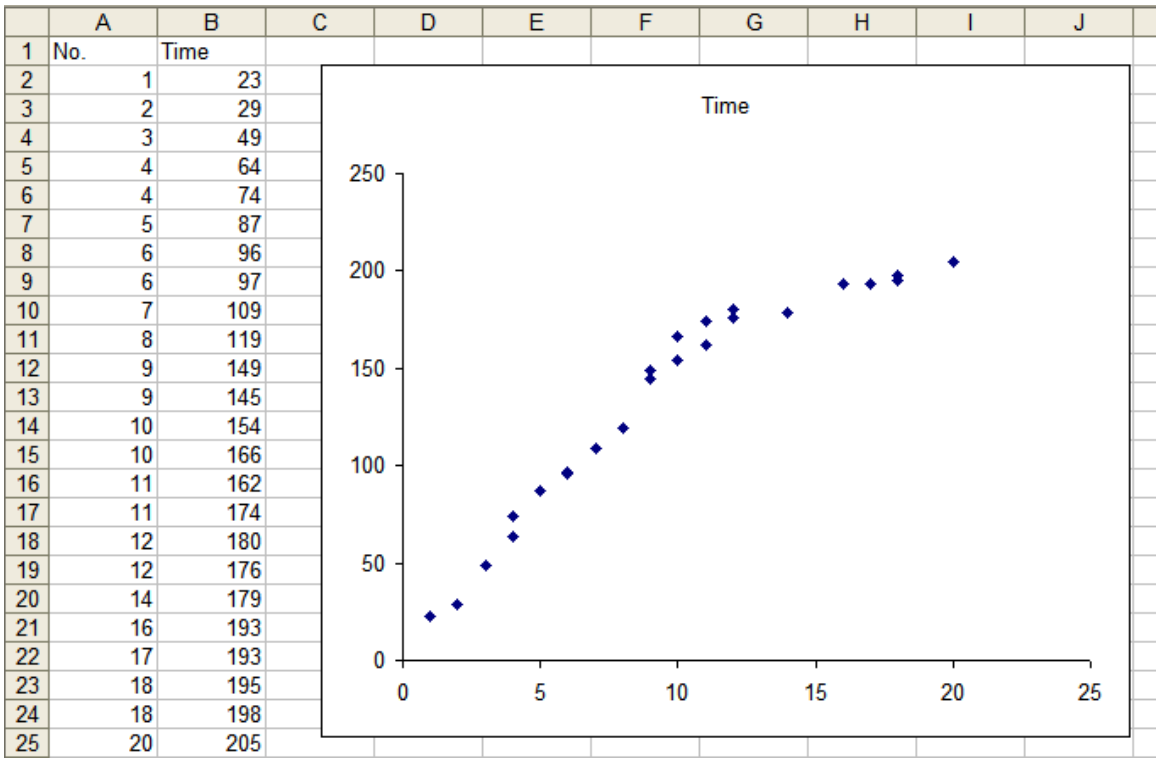
Data Display

البيانات:

Row	No.	Time
1	1	23
2	2	29
3	3	49
4	4	64
5	4	74
6	5	87
7	6	96
8	6	97
9	7	109
10	8	119
11	9	149
12	9	145
13	10	154
14	10	166
15	11	162
16	11	174
17	12	180
18	12	176
19	14	179
20	16	193
21	17	193
22	18	195
23	18	198
24	20	205

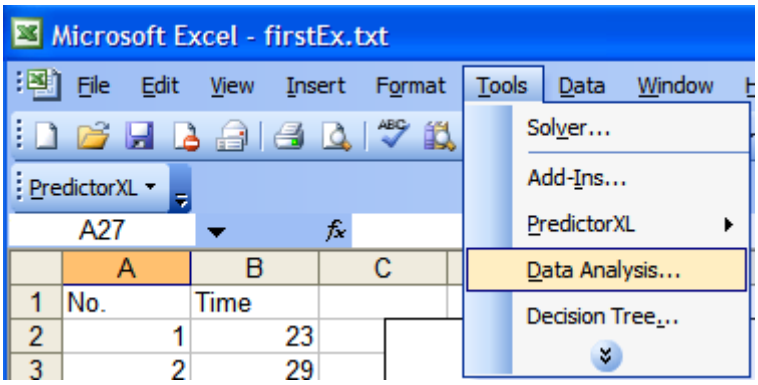
بإستخدام Excel:

ندخل البيانات ونرسمها

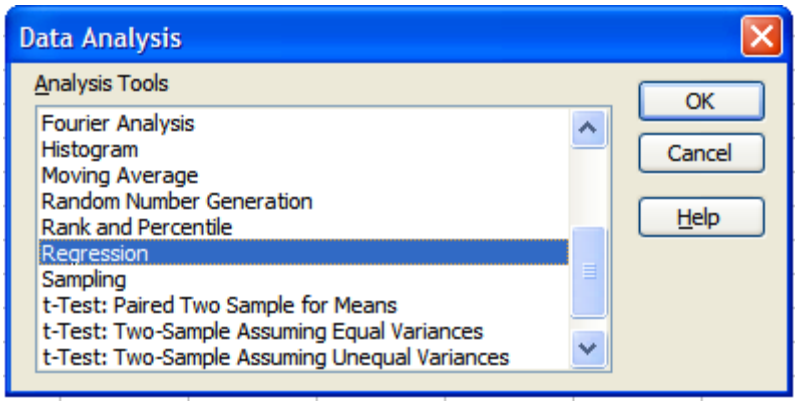


	A	B
1	No.	Time
2	1	23
3	2	29
4	3	49
5	4	64
6	4	74
7	5	87
8	6	96
9	6	97
10	7	109
11	8	119
12	9	149
13	9	145
14	10	154
15	10	166
16	11	162
17	11	174
18	12	180
19	12	176
20	14	179
21	16	193
22	17	193
23	18	195
24	18	198
25	20	205

من القائمة الرئيسية نختار Tools ومن قائمة الإسقاط الناتجة نختار Data Analysis

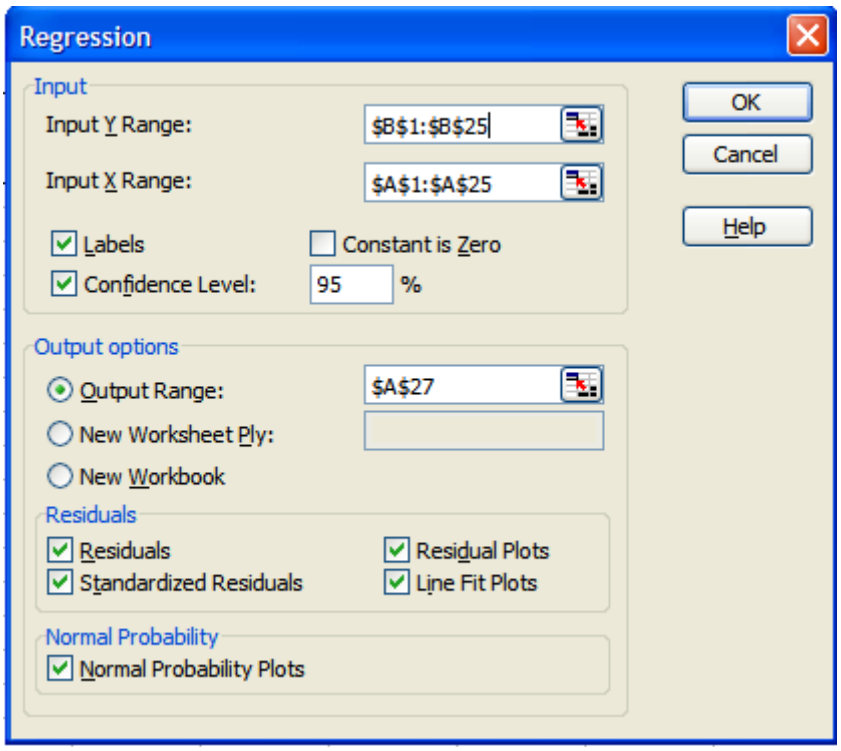


فتظهر النافذة



نختار Regression

فتظهر النافذة



ندخل المطلوب كما هو مبين

فينتج

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.946885334
R Square	0.896591836

Adjusted R Square	0.891891465
Standard Error	18.75343075
Observations	24

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	67084.79437	67084.79437	190.7491	2.55586E-12
Residual	22	7737.20563	351.691165		
Total	23	74822			

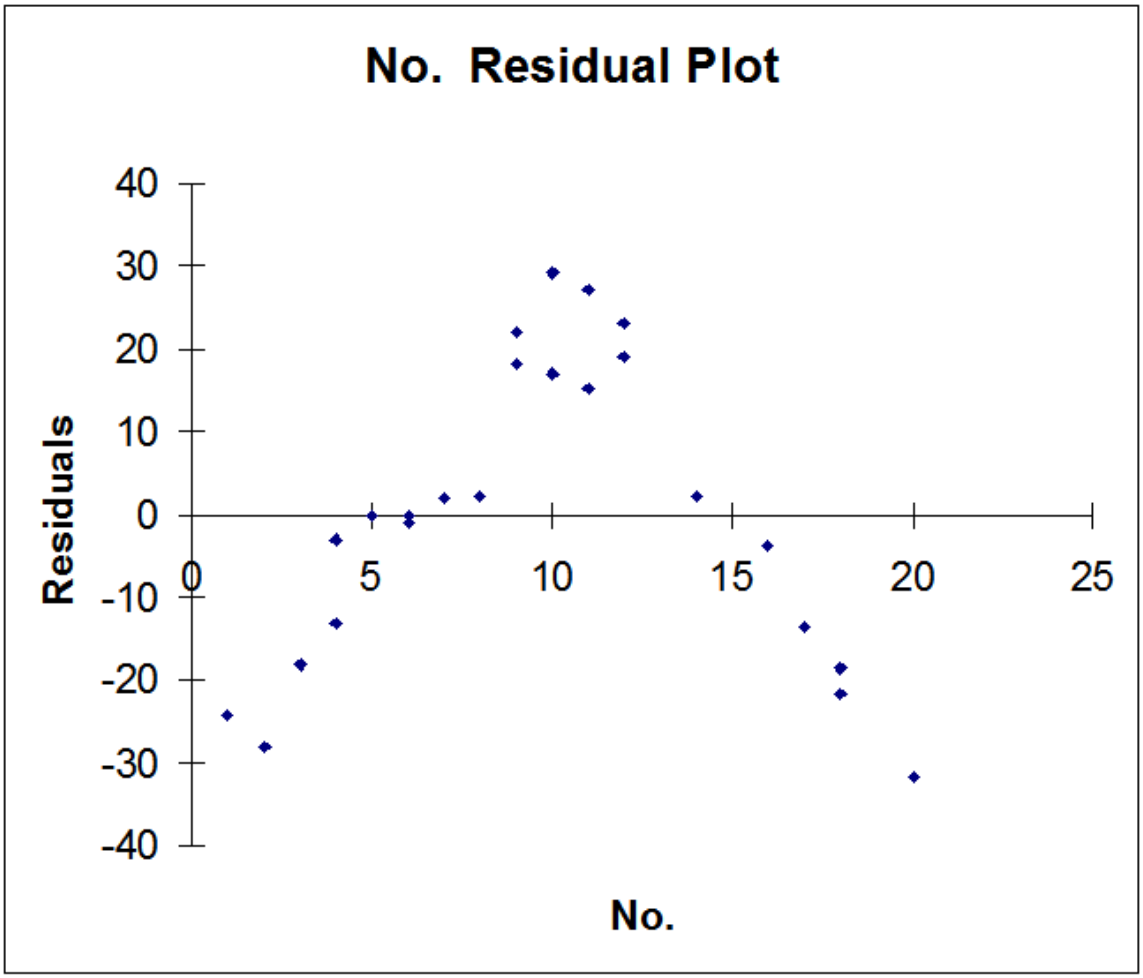
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	37.21272918	7.985251513	4.660182483	0.000120321	20.65233121	53.77312716	20.65233121	53.77312716
No.	9.969504291	0.721842166	13.81119691	2.55586E-12	8.472495271	11.46651331	8.472495271	11.46651331

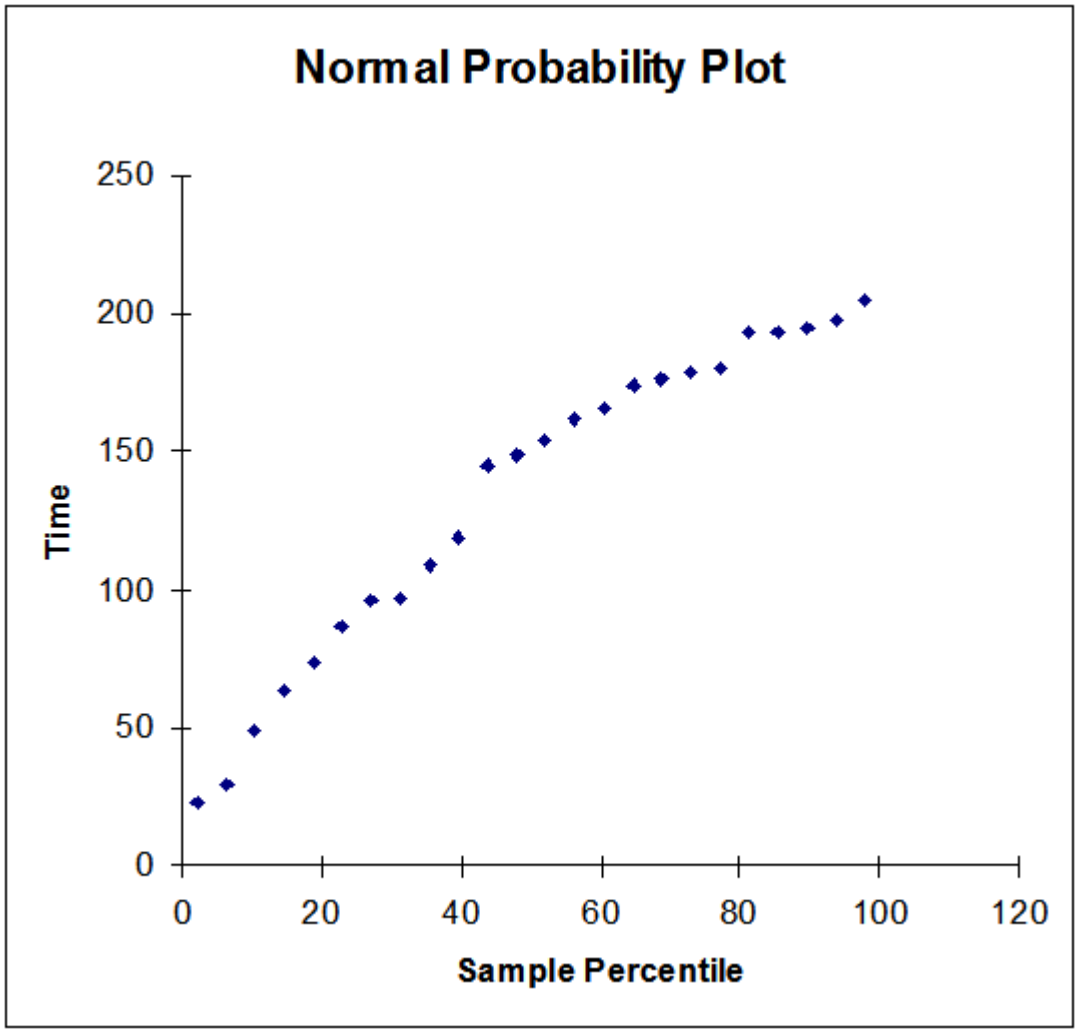
RESIDUAL OUTPUT

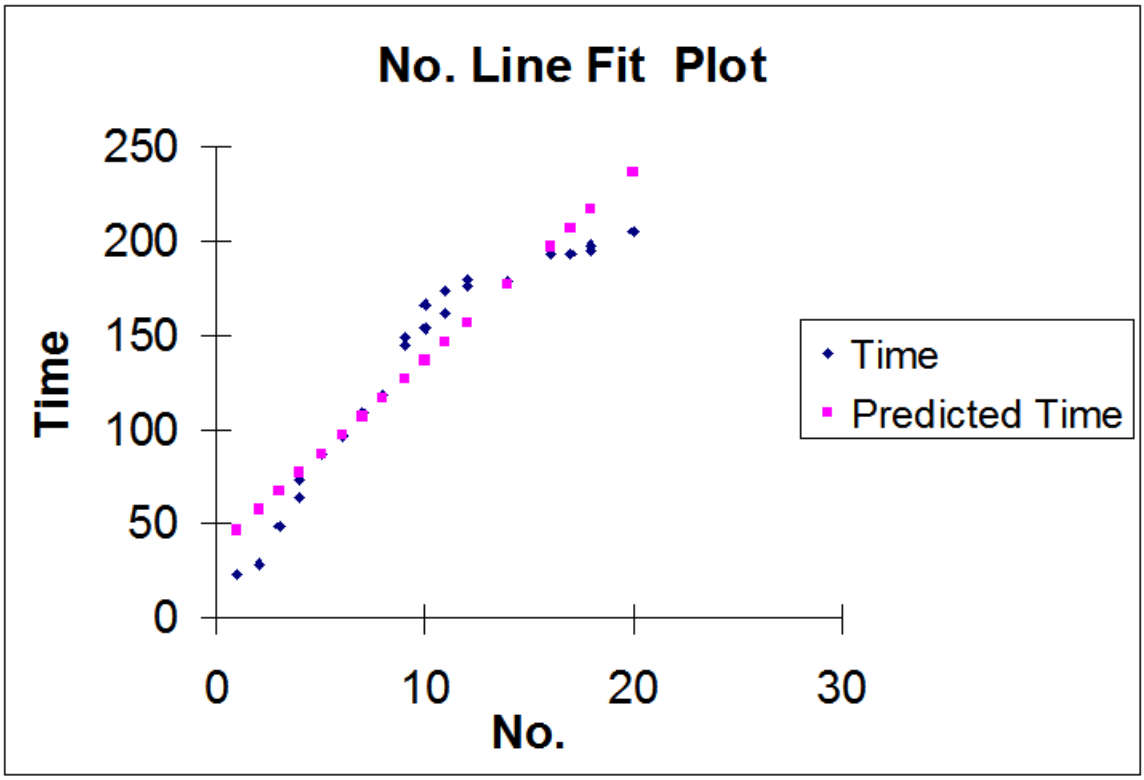
Observation	*Predicted Time*	*Residuals*	*Standard Residuals*
1	47.18223347	-24.18223347	-1.318463946
2	57.15173776	-28.15173776	-1.534889294
3	67.12124205	-18.12124205	-0.988006518
4	77.09074634	-13.09074634	-0.713733787
5	77.09074634	-3.090746342	-0.168513699
6	87.06025063	-0.060250633	-0.003284986
7	97.02975492	-1.029754923	-0.056144307
8	97.02975492	-0.029754923	-0.001622298
9	106.9992592	2.000740786	0.109084407
10	116.9687635	2.031236496	0.110747094
11	126.9382678	22.06173221	1.202849957
12	126.9382678	18.06173221	0.984761922
13	136.9077721	17.09222792	0.931902601
14	136.9077721	29.09222792	1.586166706
15	146.8772764	15.12272362	0.82452127
16	146.8772764	27.12272362	1.478785376
17	156.8467807	23.15321933	1.262360028
18	156.8467807	19.15321933	1.044271993
19	176.7857892	2.214210754	0.120723218
20	196.7247978	-3.724797827	-0.20308346
21	206.6943021	-13.69430212	-0.74664086
22	216.6638064	-21.66380641	-1.181154243
23	216.6638064	-18.66380641	-1.017588217
24	236.602815	-31.60281499	-1.723048957

PROBABILITY
OUTPUT

Percentile	*Time*
2.083333333	23
6.25	29
10.41666667	49
14.58333333	64
18.75	74
22.91666667	87
27.08333333	96
31.25	97
35.41666667	109
39.58333333	119
43.75	145
47.91666667	149
52.08333333	154
56.25	162
60.41666667	166
64.58333333	174
68.75	176
72.91666667	179
77.08333333	180
81.25	193
85.41666667	193
89.58333333	195
93.75	198
97.91666667	205


No. Residual Plot
Residuals
40
30
20
10
0
-10
-20
-30
-40
0
5
10
15
20
25
No.


Normal Probability Plot
250
200
150
100
50
0
Time
0
20
40
60
80
100
120
Sample Percentile

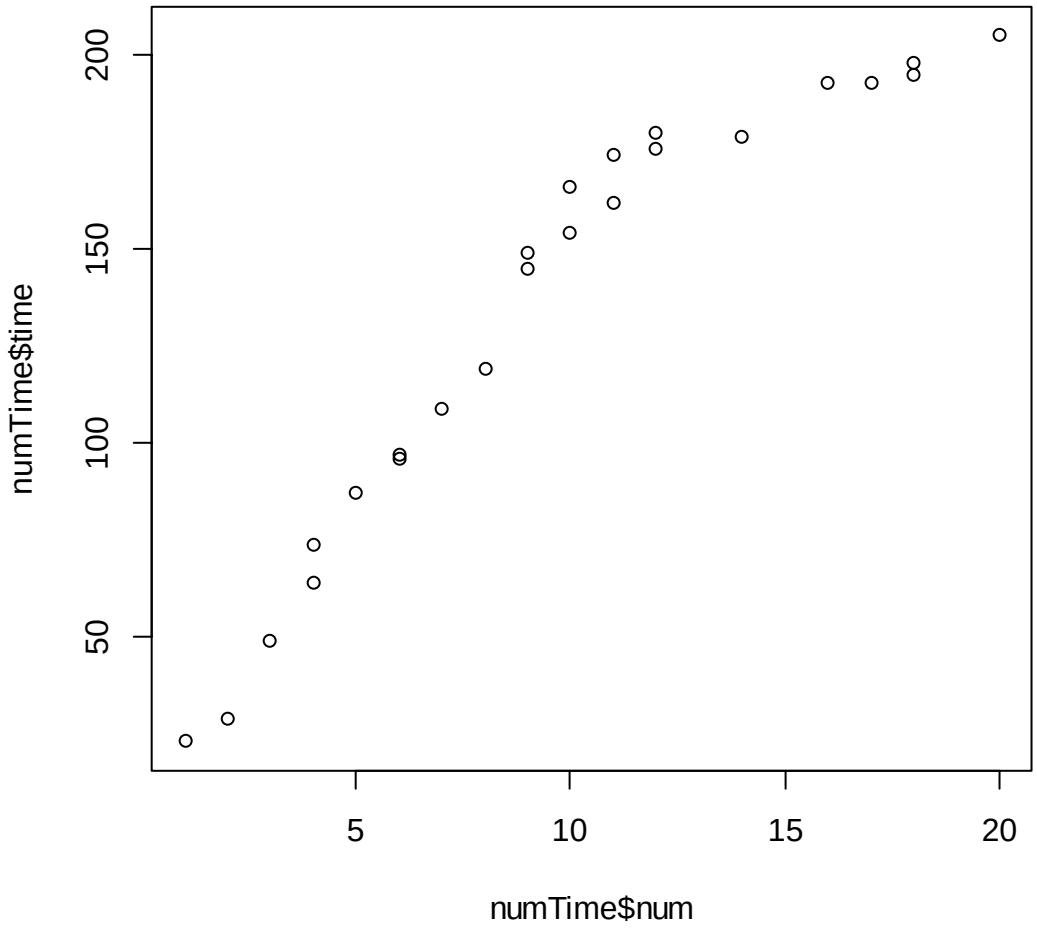

No. Line Fit Plot
250
200
150
100
50
0
Time
0
10
20
30
No.
Time
Predicted Time

بإستخدام R:

تدخل البيانات وترسم

```
> numTime = read.table(file.choose(), header =
TRUE)
> numTime
   num time
1    1   23
2    2   29
3    3   49
4    4   64
5    4   74
6    5   87
7    6   96
8    6   97
9    7  109
10   8  119
11   9  149
12   9  145
13  10  154
14  10  166
15  11  162
16  11  174
17  12  180
18  12  176
19  14  179
20  16  193
21  17  193
```

```
22  18  195
23  18  198
24  20  205
> plot(numTime$num,numTime$time)
>
```

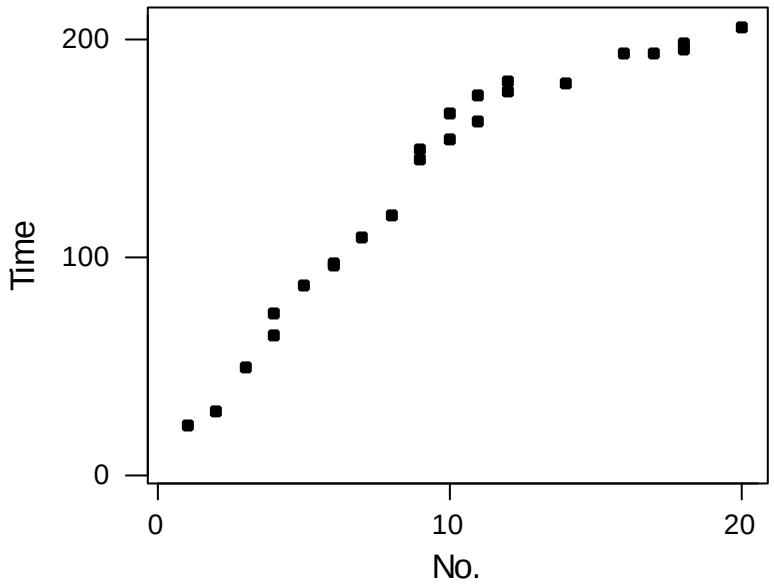



```
> fitnum = lm( numTime$time ~ numTime$num)
> summary(fitnum)

Call:
```

```
lm(formula = numTime$time ~ numTime$num)

Residuals:
    Min      1Q  Median      3Q     Max
-31.603 -14.801  -0.045  17.335  29.092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2127     7.9853    4.66  0.00012
***
numTime$num   9.9695     0.7218   13.81 2.56e-12
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.75 on 22 degrees of
freedom
Multiple R-squared: 0.8966,     Adjusted R-squared:
0.8919
F-statistic: 190.7 on 1 and 22 DF,  p-value:
2.556e-12

>
```

بإستخدام Minitab:

ترسم البيانت:



الأوامر التالية تطبق إنحدار خطي بسيط للمتغير التابع *Time* على المتغير المستقل *No.*

```
MTB > Name c3 = 'RESI1' c4 = 'SRES1' c5 = 'TRES1' c6 = 'HI1' c7 = 'COOK1' &
CONT>      c8 = 'DFIT1' c9 = 'COEF1' c10 = 'FITS1' K1 = 'MSE1' &
CONT>      m1 = 'XPXI1' m2 = 'RMAT1'
MTB > Regress 'Time' 1 'No.';
SUBC>   Residuals 'RESI1';
SUBC>   SResiduals 'SRES1';
SUBC>   Tresiduals 'TRES1';
SUBC>   Hi 'HI1';
SUBC>   Cookd 'COOK1';
SUBC>   DFits 'DFIT1';
SUBC>   Coefficients 'COEF1';
SUBC>   Fits 'FITS1';
SUBC>   MSE 'MSE1';
SUBC>   XPXInverse 'XPXI1';
```

```
SUBC>   RMatrix 'RMAT1';
SUBC>  GHistogram;
SUBC>  GNormalplot;
SUBC>  GFits;
SUBC>  GOrder;
SUBC>  RType 1;
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

نتائج التطبيق:

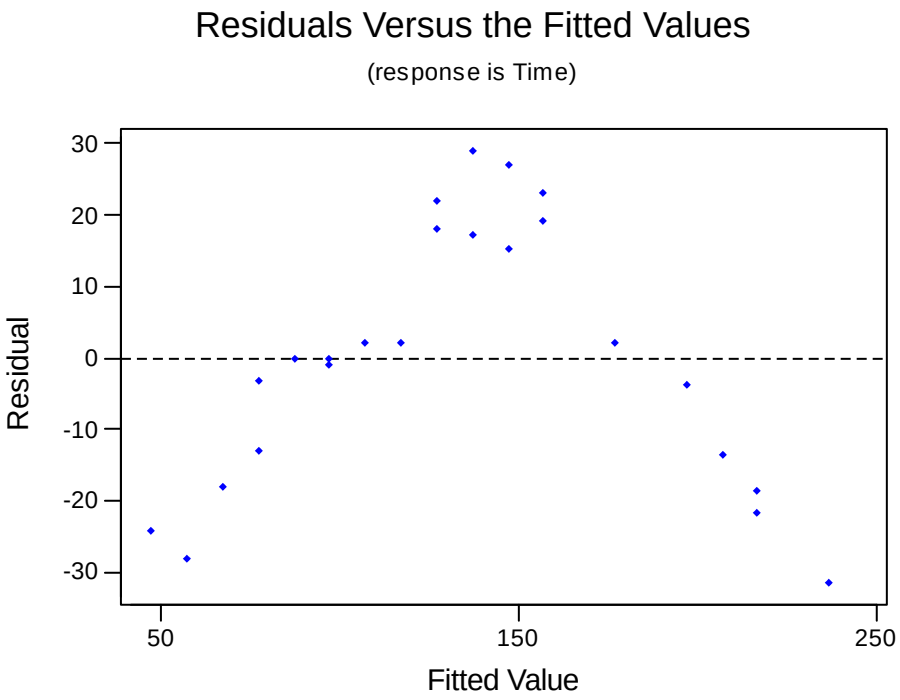
Regression Analysis: Time versus No.

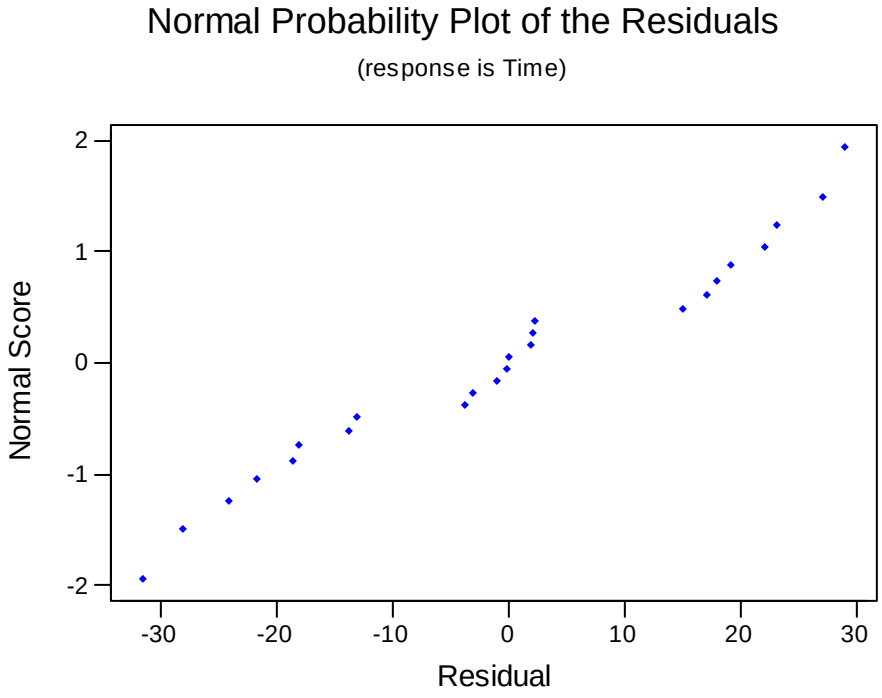
```
The regression equation is
Time = 37.2 + 9.97 No.

Predictor        Coef     SE Coef           T        P
Constant       37.213       7.985       4.66    0.000
No.            9.9695      0.7218      13.81    0.000

S = 18.75              R-Sq = 89.7%         R-Sq(adj) = 89.2%
PRESS = 9621.32        R-Sq(pred) = 87.14%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1       67085       67085    190.75    0.000
Residual Error    22        7737         352
  Lack of Fit     15        7522         501     16.33    0.001
  Pure Error       7         215          31
Total             23       74822

10 rows with no replicates

Obs        No.       Time         Fit      SE Fit    Residual    St Resid
  1        1.0      23.00       47.18        7.36      -24.18       -1.40
```

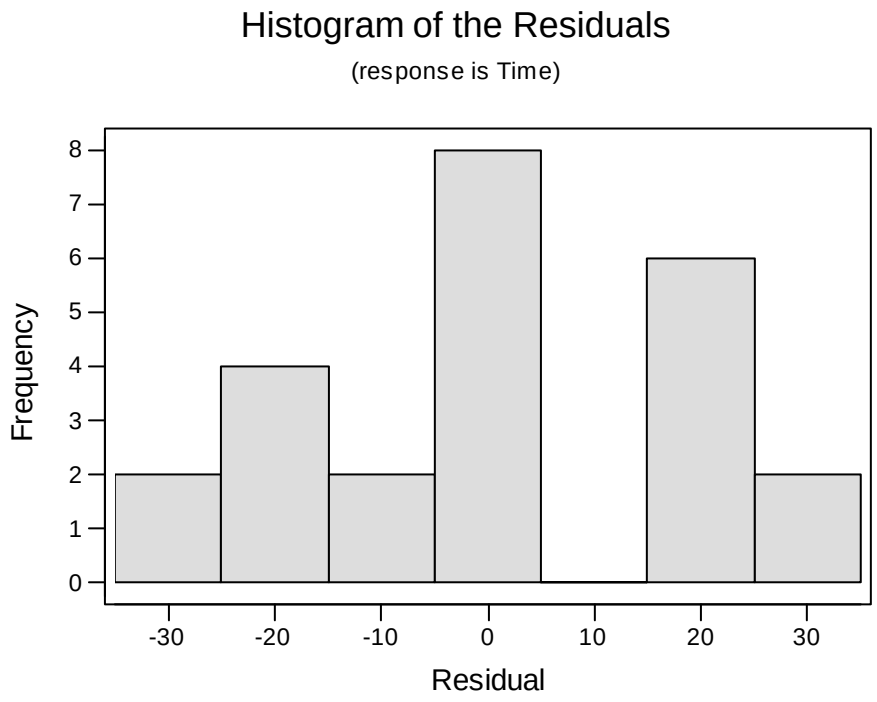
2	2.0	29.00	57.15	6.75	-28.15	-1.61
3	3.0	49.00	67.12	6.17	-18.12	-1.02
4	4.0	64.00	77.09	5.62	-13.09	-0.73
5	4.0	74.00	77.09	5.62	-3.09	-0.17
6	5.0	87.00	87.06	5.12	-0.06	-0.00
7	6.0	96.00	97.03	4.67	-1.03	-0.06
8	6.0	97.00	97.03	4.67	-0.03	-0.00
9	7.0	109.00	107.00	4.30	2.00	0.11
10	8.0	119.00	116.97	4.02	2.03	0.11
11	9.0	149.00	126.94	3.86	22.06	1.20
12	9.0	145.00	126.94	3.86	18.06	0.98
13	10.0	154.00	136.91	3.83	17.09	0.93
14	10.0	166.00	136.91	3.83	29.09	1.58
15	11.0	162.00	146.88	3.94	15.12	0.82
16	11.0	174.00	146.88	3.94	27.12	1.48
17	12.0	180.00	156.85	4.17	23.15	1.27
18	12.0	176.00	156.85	4.17	19.15	1.05
19	14.0	179.00	176.79	4.92	2.21	0.12
20	16.0	193.00	196.72	5.94	-3.72	-0.21
21	17.0	193.00	206.69	6.51	-13.69	-0.78
22	18.0	195.00	216.66	7.10	-21.66	-1.25
23	18.0	198.00	216.66	7.10	-18.66	-1.08
24	20.0	205.00	236.60	8.36	-31.60	-1.88

Durbin-Watson statistic = 0.24

Lack of fit test
Possible curvature in variable No. (P-Value = 0.000)
Possible lack of fit at outer X-values (P-Value = 0.000)
Overall lack of fit test is significant at P = 0.000


Residuals Versus the Fitted Values
(response is Time)
Residual
Fitted Value
30
20
10
0
-10
-20
-30
50
150
250


Normal Probability Plot of the Residuals
(response is Time)
Normal Score
Residual
2
1
0
-1
-2
-30
-20
-10
0
10
20
30



Row	RESI1	SRES1	TRES1	HI1	COOK1	DFIT1	COEF1
1	-24.1822	-1.40196	-1.43535	0.154022	0.178923	-0.612446	37.2127
2	-28.1517	-1.60913	-1.67370	0.129699	0.192938	-0.646120	9.9695
3	-18.1212	-1.02331	-1.02446	0.108340	0.063617	-0.357101	
4	-13.0907	-0.73173	-0.72377	0.089944	0.026459	-0.227536	
5	-3.0907	-0.17276	-0.16890	0.089944	0.001475	-0.053100	
6	-0.0603	-0.00334	-0.00326	0.074511	0.000000	-0.000926	
7	-1.0298	-0.05670	-0.05540	0.062041	0.000106	-0.014247	
8	-0.0298	-0.00164	-0.00160	0.062041	0.000000	-0.000412	
9	2.0007	0.10960	0.10711	0.052534	0.000333	0.025222	
10	2.0312	0.11089	0.10837	0.045990	0.000296	0.023795	
11	22.0617	1.20218	1.21513	0.042410	0.032003	0.255721	
12	18.0617	0.98421	0.98348	0.042410	0.021450	0.206971	
13	17.0922	0.93108	0.92815	0.041793	0.018905	0.193837	
14	29.0922	1.58477	1.64508	0.041793	0.054770	0.343564	
15	15.1227	0.82481	0.81860	0.044139	0.015707	0.175907	
16	27.1227	1.47930	1.52301	0.044139	0.050525	0.327277	
17	23.1532	1.26632	1.28492	0.049447	0.041708	0.293062	
18	19.1532	1.04755	1.04998	0.049447	0.028542	0.239478	
19	2.2142	0.12236	0.11959	0.068955	0.000554	0.032546	

```
 20   -3.7248  -0.20940  -0.20479   0.100315   0.002445  -0.068383
 21  -13.6943  -0.77862  -0.77142   0.120440   0.041507  -0.285459
 22  -21.6638  -1.24824  -1.26516   0.143527   0.130553  -0.517913
 23  -18.6638  -1.07538  -1.07941   0.143527   0.096898  -0.441873
 24  -31.6028  -1.88243  -2.00795   0.198593   0.439052  -0.999559


FITS1
    47.182    57.152    67.121    77.091    77.091    87.060    97.030
    97.030   106.999   116.969   126.938   126.938   136.908   136.908
   146.877   146.877   156.847   156.847   176.786   196.725   206.694
   216.664   216.664   236.603
```

مثال (2)

بيانات تقييم برنامج أخبار تلفزيون

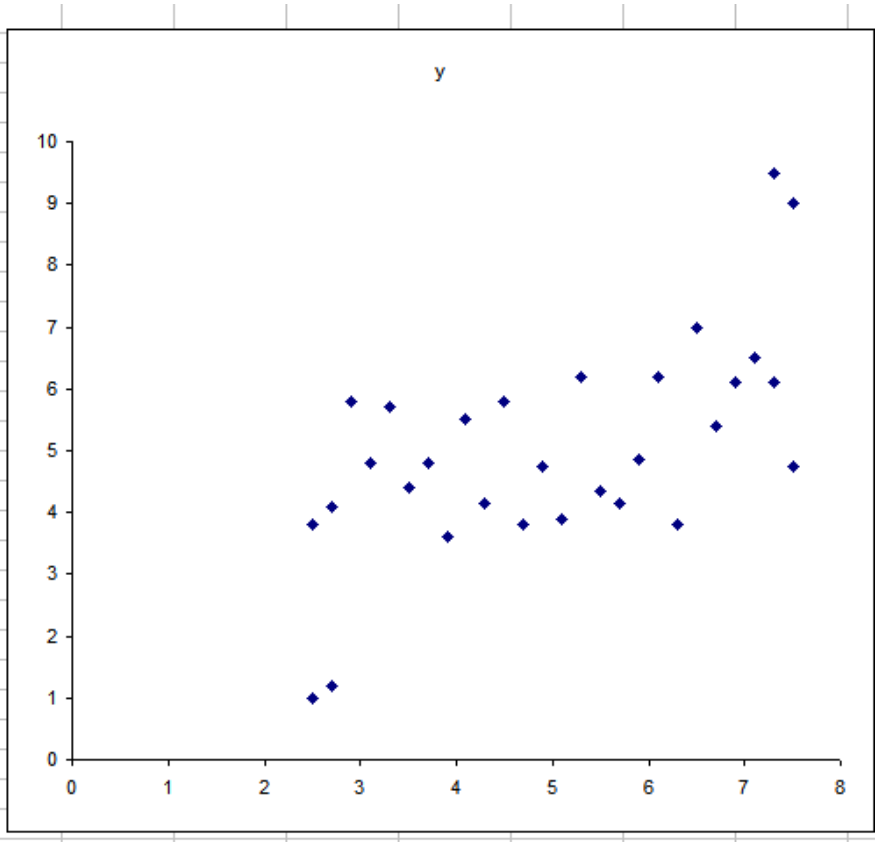
يريد مدير أحد شركات التلفزيون معرفة تأثير البرنامج المذاع مباشرة قبل الأخبار على بقاء المشاهد لمتابعة الأخبار. لتحقيق ذلك جمعت بيانات عن كثافة مشاهدة البرنامج السابق للأخبار x وكثافة مشاهدة الأخبار y كالتالي:

x	y	x	y
2.50	3.80	5.50	4.35
2.70	4.10	5.70	4.15
2.90	5.80	5.90	4.85
3.10	4.80	6.10	6.20
3.30	5.70	6.30	3.80
3.50	4.40	6.50	7.00
3.70	4.80	6.70	5.40
3.90	3.60	6.90	6.10
4.10	5.50	7.10	6.50
4.30	4.15	7.30	6.10
4.50	5.80	7.50	4.75
4.70	3.80	2.50	1.00
4.90	4.75	2.70	1.20
5.10	3.90	7.30	9.50
5.30	6.20	7.50	9.00

نريد تحليل هذه البيانات بإستخدام الإنحدار الخطي البسيط وإستخراج كل النتائج الممكنة.

بإستخدام Excel:

x	y
2.5	3.8
2.7	4.1
2.9	5.8
3.1	4.8
3.3	5.7
3.5	4.4
3.7	4.8
3.9	3.6
4.1	5.5
4.3	4.15
4.5	5.8
4.7	3.8
4.9	4.75
5.1	3.9
5.3	6.2
5.5	4.35
5.7	4.15
5.9	4.85
6.1	6.2
6.3	3.8
6.5	7
6.7	5.4
6.9	6.1
7.1	6.5
7.3	6.1
7.5	4.75
2.5	1
2.7	1.2
7.3	9.5
7.5	9



SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.961916468
R Square	0.925283292
Adjusted R Square	0.890800534
Standard Error	1.480878046
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	787.5780062	787.5780062	359.132733	1.65897E-17
Residual	29	63.59699379	2.192999786		
Total	30	851.175			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
x	0.97318	0.051353172	18.95079768	7.0355E-18	0.86815455	1.07821607	0.86815455	1.07821607

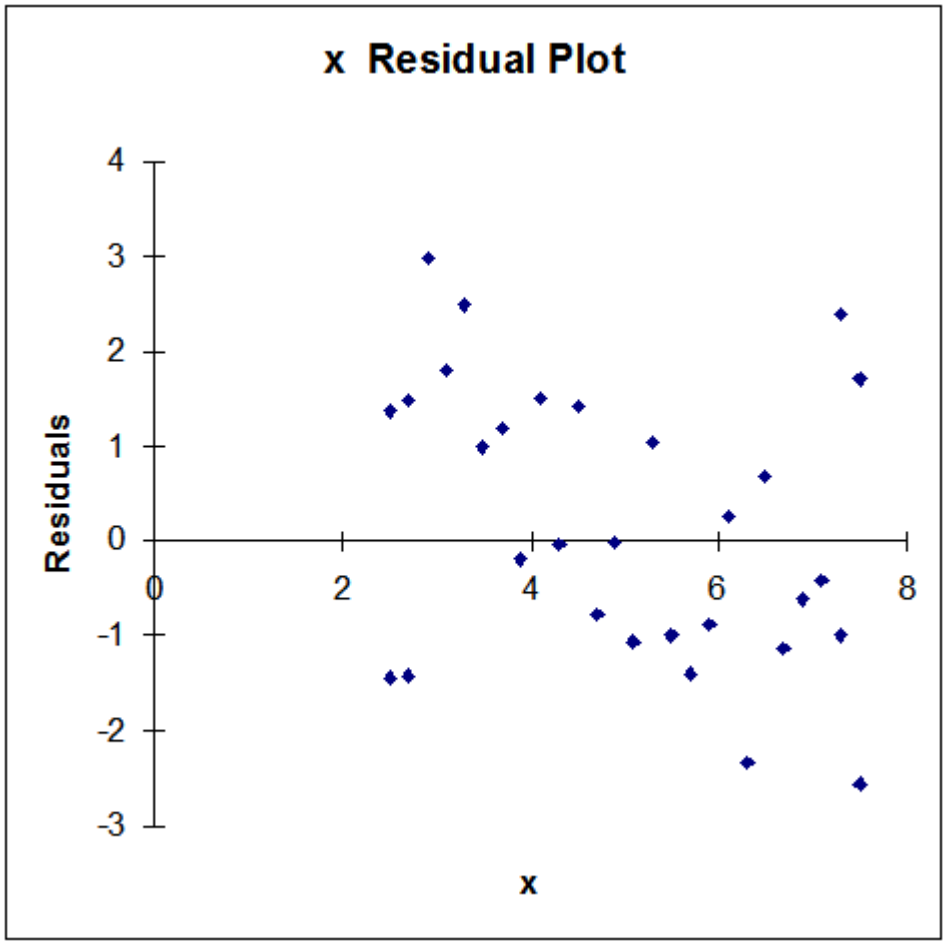
RESIDUAL OUTPUT

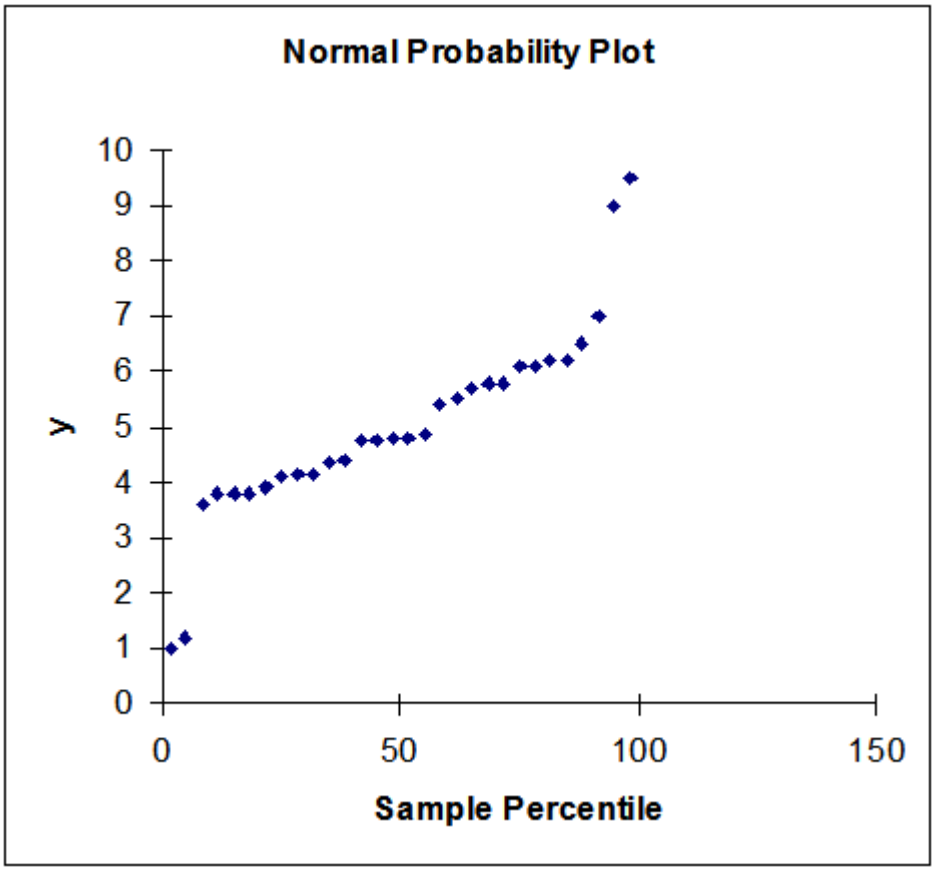
Observation	*Predicted y*	*Residuals*	*Standard Residuals*
1	2.43296	1.367041054	0.938909841
2	2.6276	1.472404339	1.011275352
3	2.82223	2.977767623	2.045187536
4	3.01687	1.783130907	1.224688279
5	3.21151	2.488494192	1.709145222
6	3.40614	0.993857476	0.68260025
7	3.60078	1.19922076	0.823647666
8	3.79542	-0.195415955	-0.134215401
9	3.99005	1.509947329	1.037060593
10	4.18469	-0.034689386	-0.023825332
11	4.37933	1.420673898	0.9757459
12	4.57396	-0.773962818	-0.531572409
13	4.7686	-0.018599533	-0.012774514
14	4.96324	-1.063236249	-0.730250913
15	5.15787	1.042127035	0.715752702
16	5.35251	-1.00250968	-0.688542748
17	5.54715	-1.397146396	-0.959586763
18	5.74178	-0.891783112	-0.612493631
19	5.93642	0.263580173	0.181031884
20	6.13106	-2.331056543	-1.601014045
21	6.32569	0.674306741	0.463126717
22	6.52033	-1.120329974	-0.769463971
23	6.71497	-0.61496669	-0.422370839

24	6.9096	0.409603406	-0.281323423
25	7.10424	1.004240121	-0.689731248
26	7.29888	2.548876837	-1.750617173
27	2.43296	1.432958946	-0.984183505
28	2.6276	1.427595661	-0.980499899
29	7.10424	2.395759879	1.645453529
30	7.29888	1.701123163	1.168363798

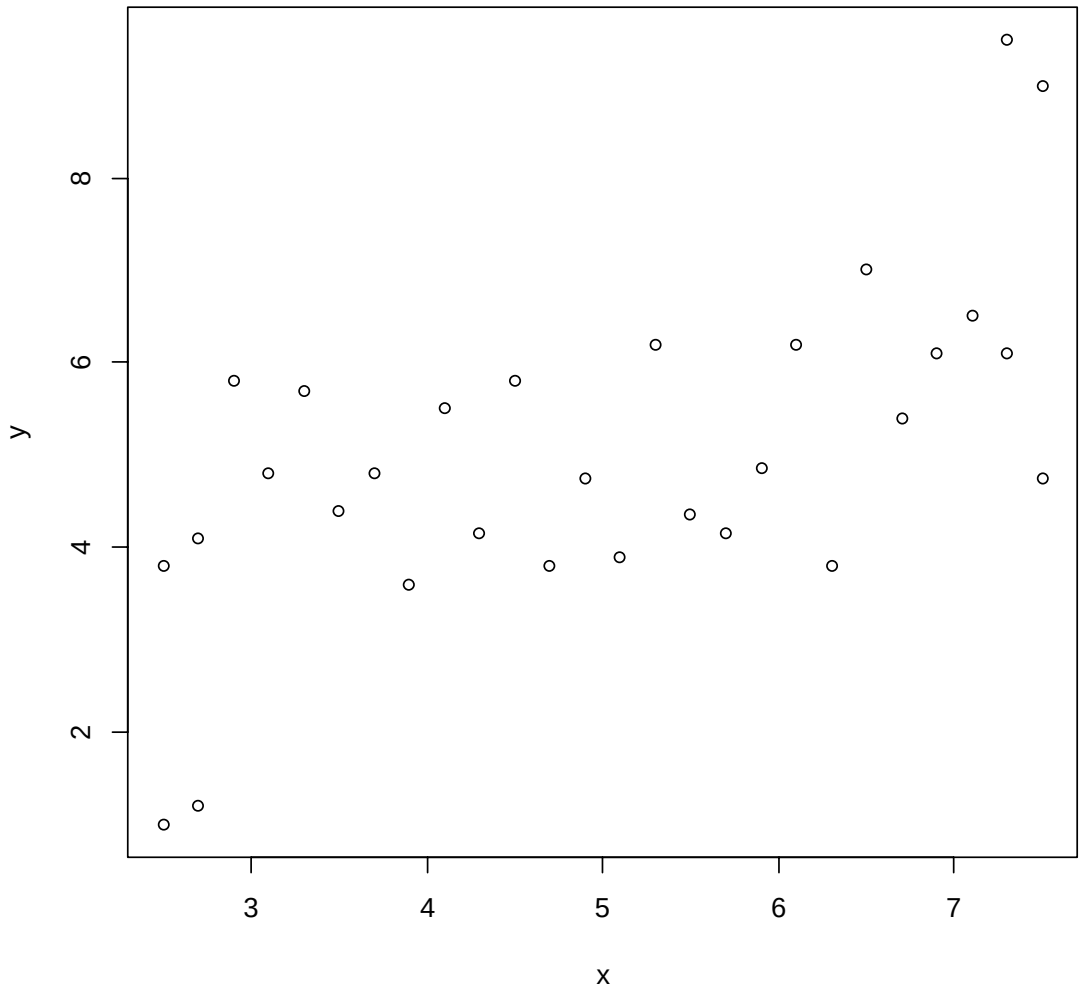
PROBABILITY OUTPUT

Percentile	*y*
1.666666667	1
5	1.2
8.333333333	3.6
11.66666667	3.8
15	3.8
18.33333333	3.8
21.66666667	3.9
25	4.1
28.33333333	4.15
31.66666667	4.15
35	4.35
38.33333333	4.4
41.66666667	4.75
45	4.75
48.33333333	4.8
51.66666667	4.8
55	4.85
58.33333333	5.4
61.66666667	5.5
65	5.7
68.33333333	5.8
71.66666667	5.8
75	6.1
78.33333333	6.1
81.66666667	6.2
85	6.2
88.33333333	6.5
91.66666667	7
95	9
98.33333333	9.5


x Residual Plot
Residuals
x
4
3
2
1
0
-1
-2
-3
0
2
4
6
8


Normal Probability Plot
y
10
9
8
7
6
5
4
3
2
1
0
0
50
100
150
Sample Percentile

بإستخدام R:

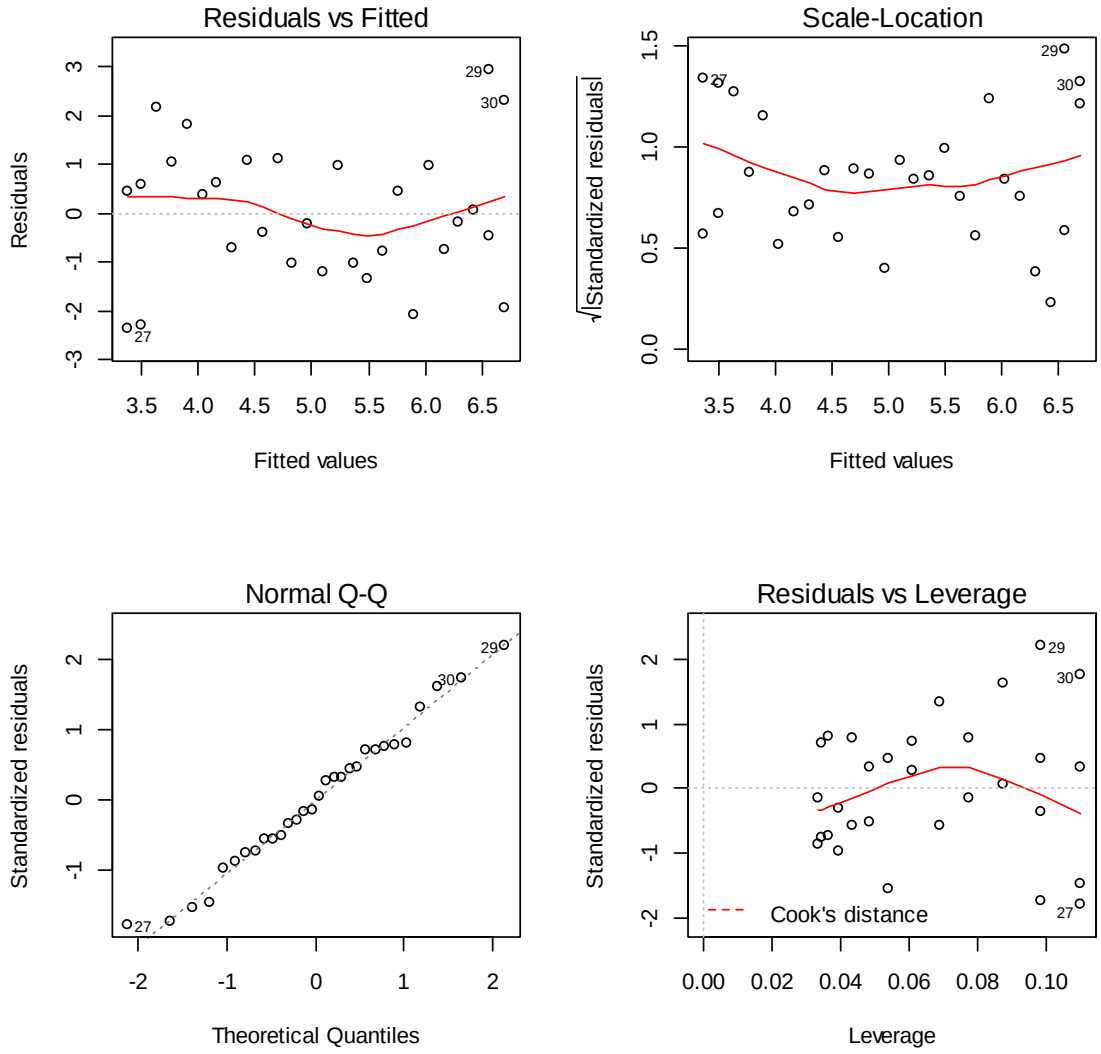



```
> tv = lm( y ~ x)
> summary(tv)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.36994 -0.95755 -0.06405  0.96824  2.93634

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   1.7065     0.8172   2.088 0.045977 *
x             0.6654     0.1552   4.287 0.000194 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 1.402 on 28 degrees of freedom
Multiple R-squared: 0.3963,     Adjusted R-squared: 0.3747
F-statistic: 18.38 on 1 and 28 DF,  p-value: 0.0001939

> anova(tv)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 36.116  36.116  18.378 0.0001939 ***
Residuals 28 55.026   1.965
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
> library()
> confint(tv)
                  2.5 %    97.5 %
(Intercept) 0.03267164 3.3804035
x           0.34743079 0.9832875
>
> vcov(tv)
             (Intercept)          x
(Intercept)   0.6677421 -0.1204470
x            -0.1204470  0.0240894
>
> influence(tv)
$hat
         1          2          3          4          5
6          7          8          9         10
0.10994525 0.09817766 0.08739070 0.07758438 0.06875868
0.06091362 0.05404920 0.04816540 0.04326224 0.03933971
        11         12         13         14         15
16         17         18         19         20
0.03639781 0.03443654 0.03345591 0.03345591 0.03443654
0.03639781 0.03933971 0.04326224 0.04816540 0.05404920
        21         22         23         24         25
26         27         28         29         30
0.06091362 0.06875868 0.07758438 0.08739070 0.09817766
0.10994525 0.10994525 0.09817766 0.09817766 0.10994525
```

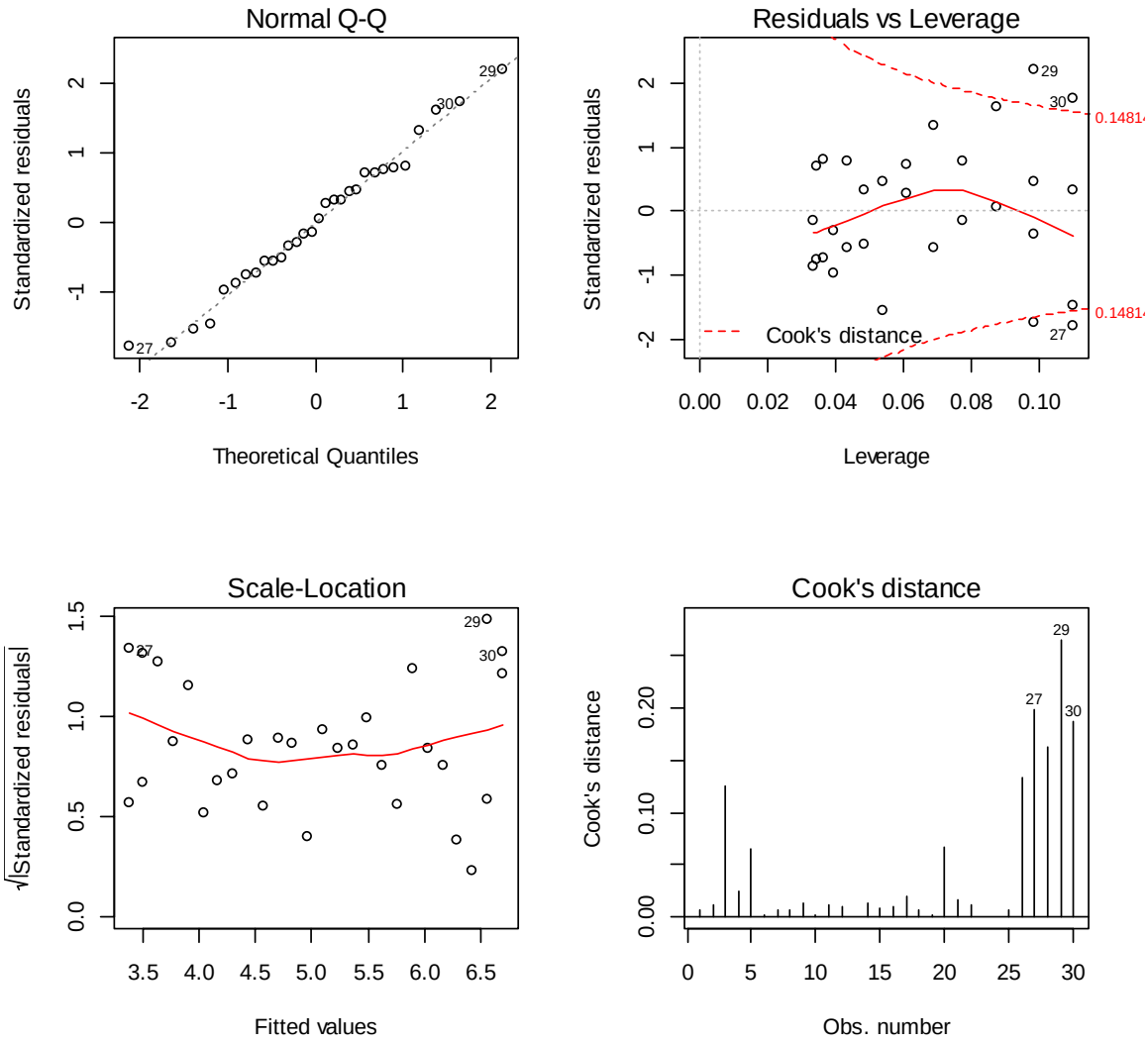
```
$coefficients
    (Intercept)             x
1   0.090142346 -0.0148072101
2   0.115383467 -0.0186634611
3   0.384222102 -0.0610368461
4   0.167391053 -0.0260278521
5   0.265495393 -0.0402289651
6   0.048649179 -0.0071407563
7   0.075459247 -0.0106403596
8  -0.074247333  0.0099365803
9   0.098552993 -0.0122861336
10 -0.033138434  0.0037298055
11  0.072990788 -0.0069923478
12 -0.055371282  0.0039369690
13 -0.008851455  0.0002749468
14 -0.033771557 -0.0015216983
15  0.014969623  0.0036830674
16 -0.002834800 -0.0064623095
17  0.013438516 -0.0120499080
18  0.017844308 -0.0090190250
19 -0.015569150  0.0061589781
20  0.102797787 -0.0353475022
21 -0.060444419  0.0189652689
22  0.058167126 -0.0171060053
23  0.017797694 -0.0049870640
24 -0.007254133  0.0019578887
25  0.055337797 -0.0144951331
26  0.262224961 -0.0670263514
27 -0.496742960  0.0815973586
28 -0.445112564  0.0719976720
29 -0.350452555  0.0917972298
30 -0.310250617  0.0793020117

$sigma
       1        2        3        4        5        6
7        8        9       10       11       12
1.424887 1.422449 1.359398 1.412562 1.381831 1.425747
1.422103 1.420864 1.412109 1.425228 1.411222 1.413156
      13       14       15       16       17       18
19       20       21       22       23       24
1.426954 1.408131 1.414965 1.413620 1.402794 1.419266
1.425007 1.365874 1.414565 1.419421 1.427036 1.427517
      25       26       27       28       29       30
1.424489 1.371240 1.343235 1.349139 1.297651 1.348052

$wt.res
```

```
          1           2           3           4           5
6           7           8           9
 0.43006456  0.59699273  2.16392090  1.03084906  1.79777723
0.36470540  0.63163357 -0.70143826  1.06548991
         10          11          12          13          14
15          16          17          18
-0.41758192  1.09934624 -1.03372559 -0.21679742 -1.19986925
0.96705892 -1.01601291 -1.34908474 -0.78215657
         19          20          21          22          23
24          25          26          27
 0.43477159 -2.09830024  0.96862793 -0.76444390 -0.19751573
0.06941244 -0.46365939 -1.94673122 -2.36993544
         28          29          30
-2.30300727  2.93634061  2.30326878

>
> layout(matrix(c(1,2,3,4),2,2))
> plot(tv)
>
```



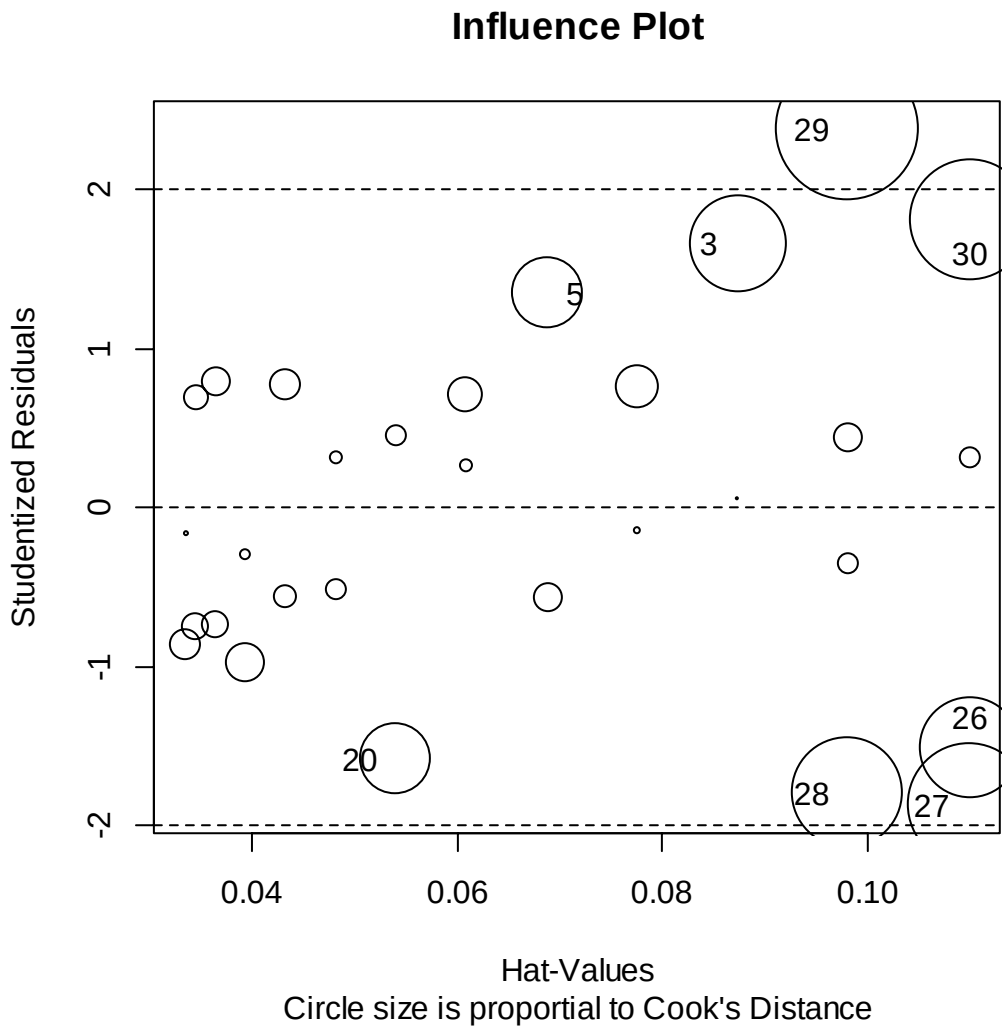

```
> library(car)
Loading required package: MASS
Loading required package: nnet
> outlierTest(tv)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
29 2.382803           0.024482      0.73447
>
> qqPlot(tv, main = "QQ PLOT")
> leveragePlots(tv)
> avPlots(tv)
```

```
> cutoff=4/((length(y)-length(tv$coefficients)-1))
> plot(tv, cook.levels = cutoff)
> plot(tv, which = 4, cook.levels = cutoff)
> abline(0,0)
>
```




```
> influencePlot(tv, id.method = "identify", main =
"Influence Plot", sub = "Circle size is proportial
to Cook's Distance")
warning: nearest point already identified
warning: nearest point already identified
warning: no point within 0.25 inches
     StudRes         Hat     CookD
```

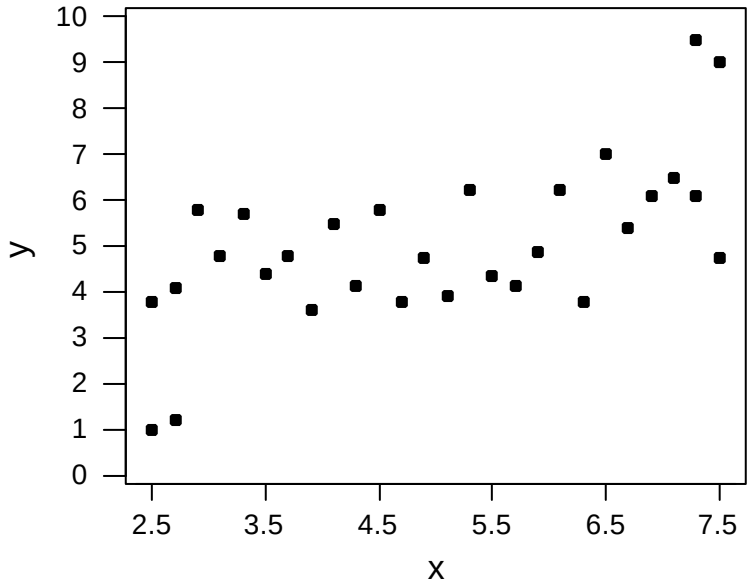
```
3   1.666296 0.08739070 0.3535650
5   1.348186 0.06875868 0.2553392
20 -1.579511 0.05404920 0.2601201
26 -1.504819 0.10994525 0.3658118
28 -1.797539 0.09817766 0.4036090
29  2.382803 0.09817766 0.5146025
30  1.811046 0.10994525 0.4328090
>
```



تمرين: يترك إستخراج النتائج والتشخيصات للطالب وتفسير النتائج.

بإستخدام Minitab:

أولا: نرسم البيانات في رسم إنتشار



يبدو أن هناك علاقة خطية بين كثافتي المشاهدة للبرنامجين ونوجد هذه العلاقة كالتالي:

```
MTB > Regress 'y' 1 'x';
SUBC>  GHistogram;
SUBC>  GNormalplot;
SUBC>  GFits;
SUBC>  GOrder;
SUBC>  RType 1;
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: y versus x

The regression equation is
y = 1.71 + 0.665 x

Predictor	Coef	SE Coef	T	P
Constant	1.7065	0.8172	2.09	0.046
x	0.6654	0.1552	4.29	0.000

S = 1.402 R-Sq = 39.6% R-Sq(adj) = 37.5%
PRESS = 65.5095 R-Sq(pred) = 28.12%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36.116	36.116	18.38	0.000
Residual Error	28	55.026	1.965		
Lack of Fit	24	32.090	1.337	0.23	0.991
Pure Error	4	22.936	5.734		
Total	29	91.142			

22 rows with no replicates

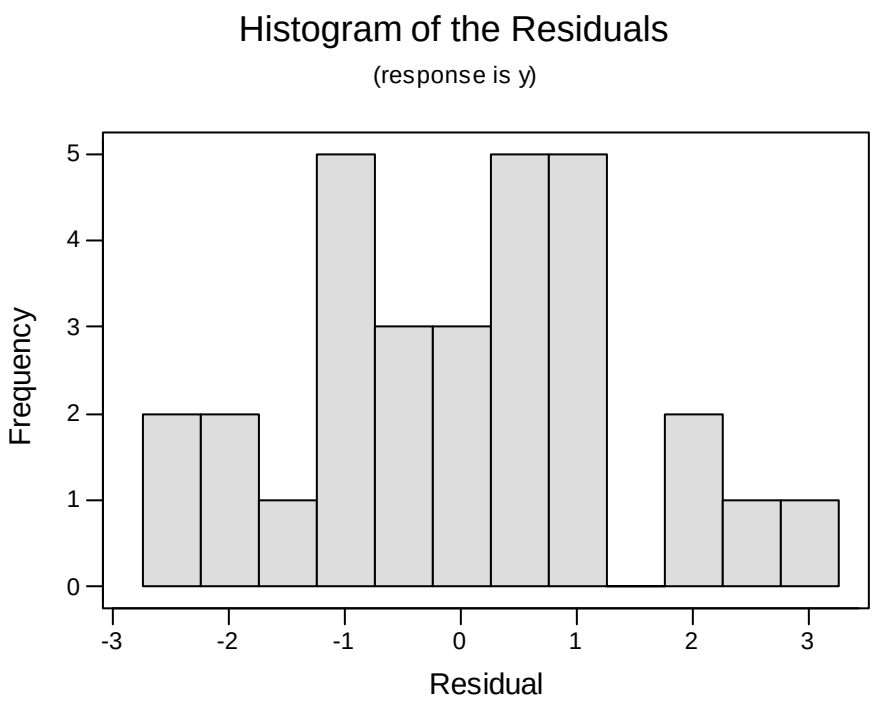
Obs	x	y	Fit	SE Fit	Residual	St Resid
1	2.50	3.800	3.370	0.465	0.430	0.33
2	2.70	4.100	3.503	0.439	0.597	0.45
3	2.90	5.800	3.636	0.414	2.164	1.62
4	3.10	4.800	3.769	0.390	1.031	0.77
5	3.30	5.700	3.902	0.368	1.798	1.33
6	3.50	4.400	4.035	0.346	0.365	0.27
7	3.70	4.800	4.168	0.326	0.632	0.46
8	3.90	3.600	4.301	0.308	-0.701	-0.51
9	4.10	5.500	4.435	0.292	1.065	0.78
10	4.30	4.150	4.568	0.278	-0.418	-0.30
11	4.50	5.800	4.701	0.267	1.099	0.80
12	4.70	3.800	4.834	0.260	-1.034	-0.75
13	4.90	4.750	4.967	0.256	-0.217	-0.16
14	5.10	3.900	5.100	0.256	-1.200	-0.87
15	5.30	6.200	5.233	0.260	0.967	0.70
16	5.50	4.350	5.366	0.267	-1.016	-0.74
17	5.70	4.150	5.499	0.278	-1.349	-0.98
18	5.90	4.850	5.632	0.292	-0.782	-0.57

19	6.10	6.200	5.765	0.308	0.435	0.32
20	6.30	3.800	5.898	0.326	-2.098	-1.54
21	6.50	7.000	6.031	0.346	0.969	0.71
22	6.70	5.400	6.164	0.368	-0.764	-0.57
23	6.90	6.100	6.298	0.390	-0.198	-0.15
24	7.10	6.500	6.431	0.414	0.069	0.05
25	7.30	6.100	6.564	0.439	-0.464	-0.35
26	7.50	4.750	6.697	0.465	-1.947	-1.47
27	2.50	1.000	3.370	0.465	-2.370	-1.79
28	2.70	1.200	3.503	0.439	-2.303	-1.73
29	7.30	9.500	6.564	0.439	2.936	2.21R
30	7.50	9.000	6.697	0.465	2.303	1.74

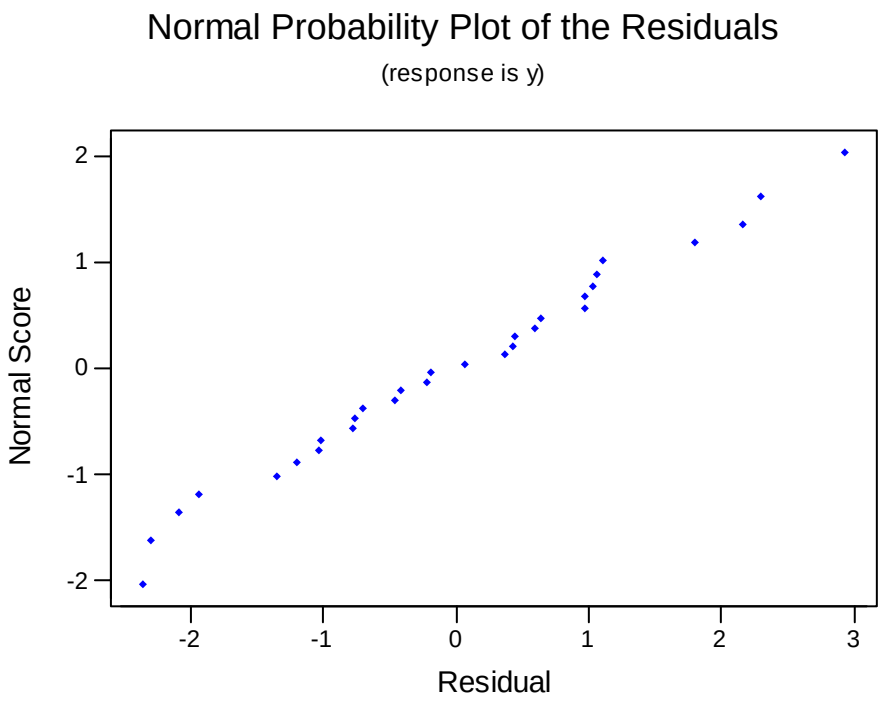
R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.50
Possible lack of fit at outer X-values (P-Value = 0.021)
Overall lack of fit test is significant at P = 0.021

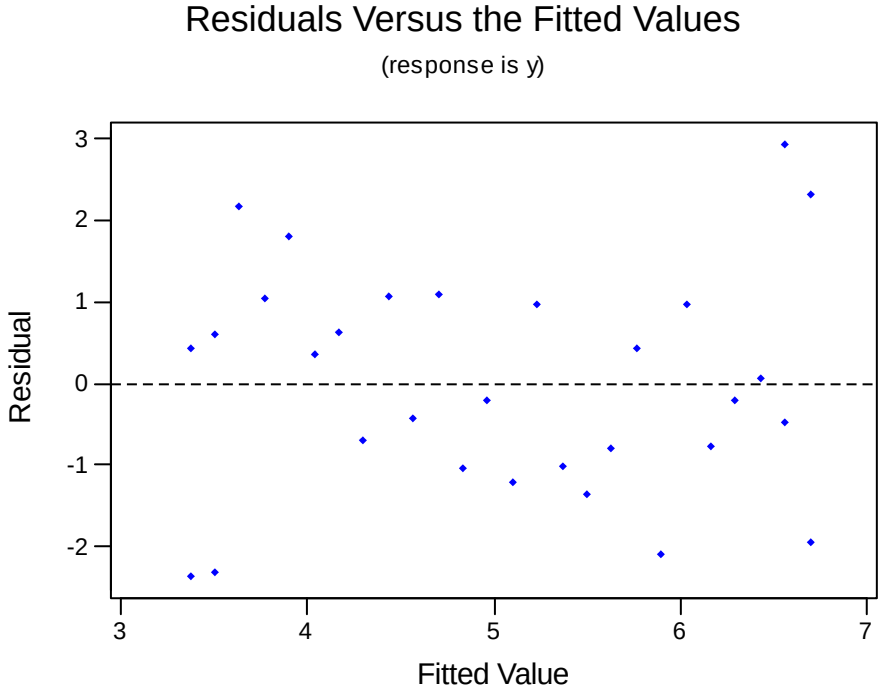
Residual Histogram for y



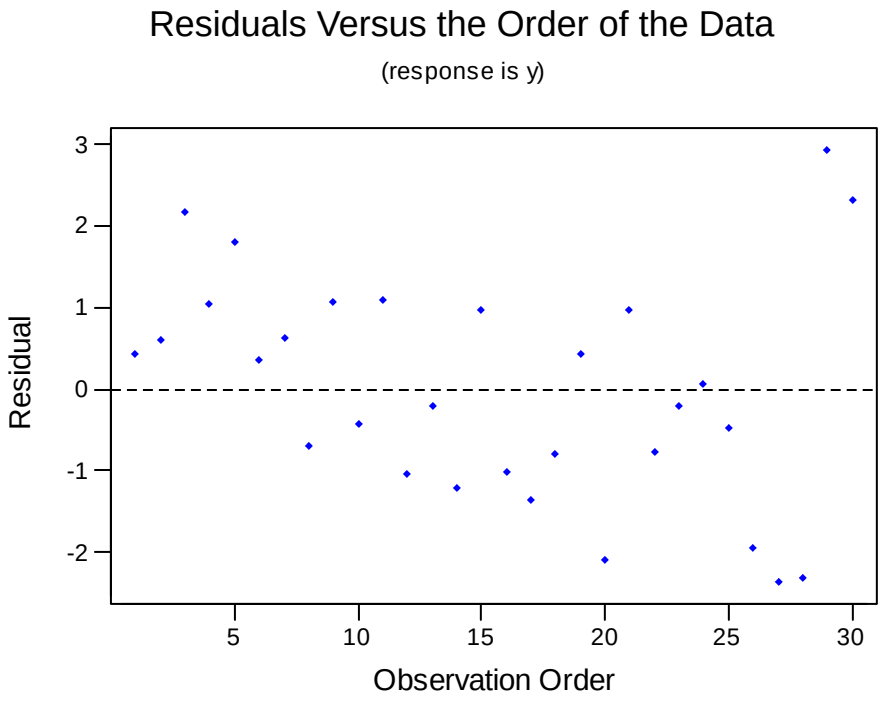
Normplot of Residuals for y



Residuals vs Fits for y



Residuals vs Order for y



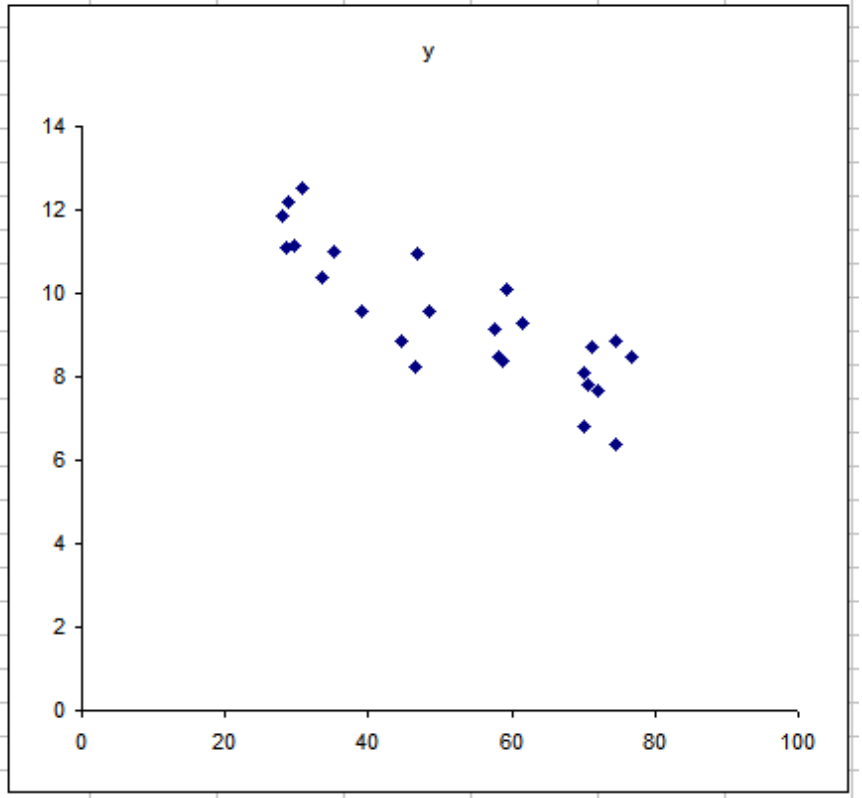
مثال (3)

البيانات التالية لإستخدام البخار في أحد المصانع y ومتوسط الضغط الجوي x .

x	y
35.3	10.98
29.7	11.13
30.8	12.51
58.8	8.40
61.4	9.27
71.3	8.73
74.4	6.36
76.7	8.50
70.7	7.82
57.5	9.14
46.4	8.24
28.9	12.19
28.1	11.88
39.1	9.57
46.8	10.94
48.5	9.58
59.3	10.09
70.0	8.11
70.0	6.83
74.5	8.88
72.1	7.68
58.1	8.47
44.6	8.86
33.4	10.36
28.6	11.08

بإستخدام Excel:

x	y
35.3	10.98
29.7	11.13
30.8	12.51
58.8	8.4
61.4	9.27
71.3	8.73
74.4	6.36
76.7	8.5
70.7	7.82
57.5	9.14
46.4	8.24
28.9	12.19
28.1	11.88
39.1	9.57
46.8	10.94
48.5	9.58
59.3	10.09
70	8.11
70	6.83
74.5	8.88
72.1	7.68
58.1	8.47
44.6	8.86
33.4	10.36
28.6	11.08



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.84524406
R Square	0.714437521
Adjusted R Square	0.702021761
Standard Error	0.890124516
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	45.5924	45.5924	57.54279	1.05E-07
Residual	23	18.2234	0.792322		
Total	24	63.8158			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*

Interc ept	13.62298 927	0.581463	23.42 88	1.5E-17	12.4201 4	14.8258 4	12.42014	14.82584
x	-0.079828 693	0.010524	-7.585 7	1.05E-07	-0.1016	-0.05806	-0.1016	-0.05806

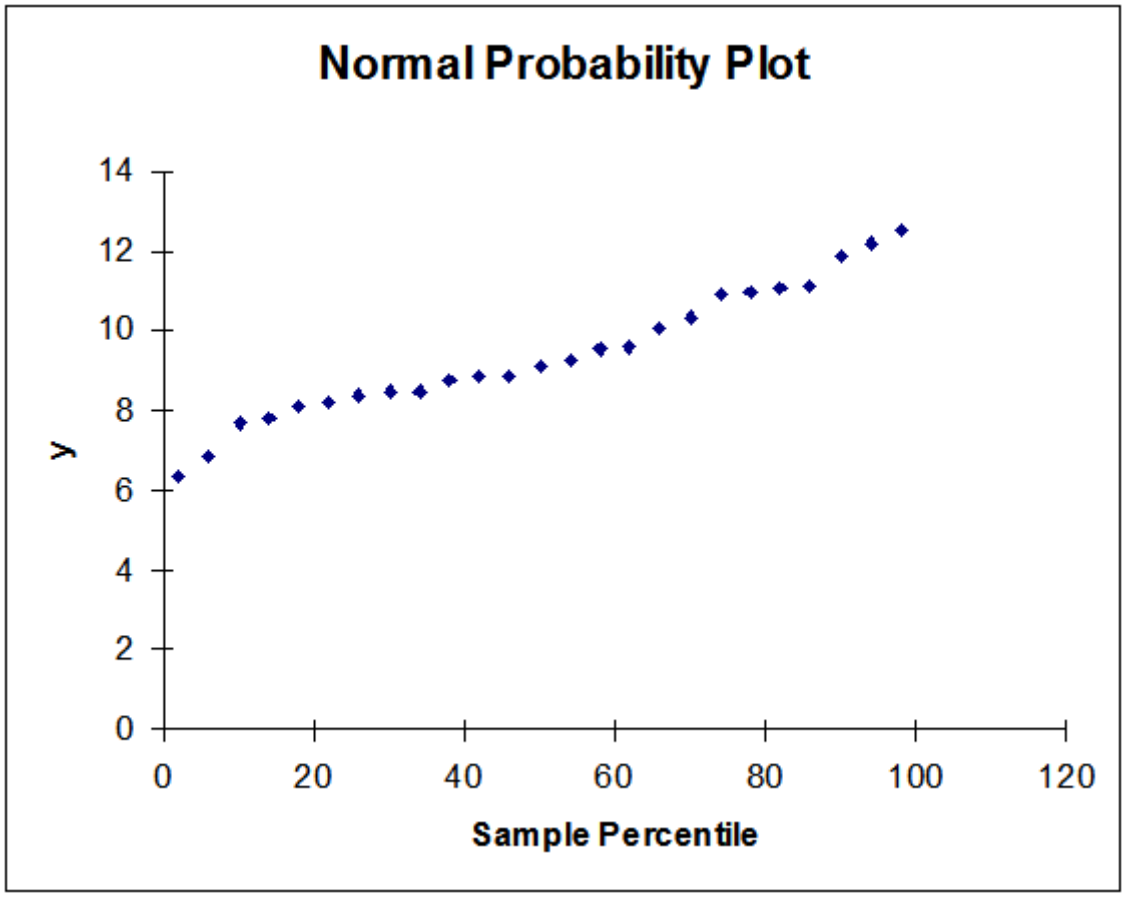
RESIDUAL OUTPUT

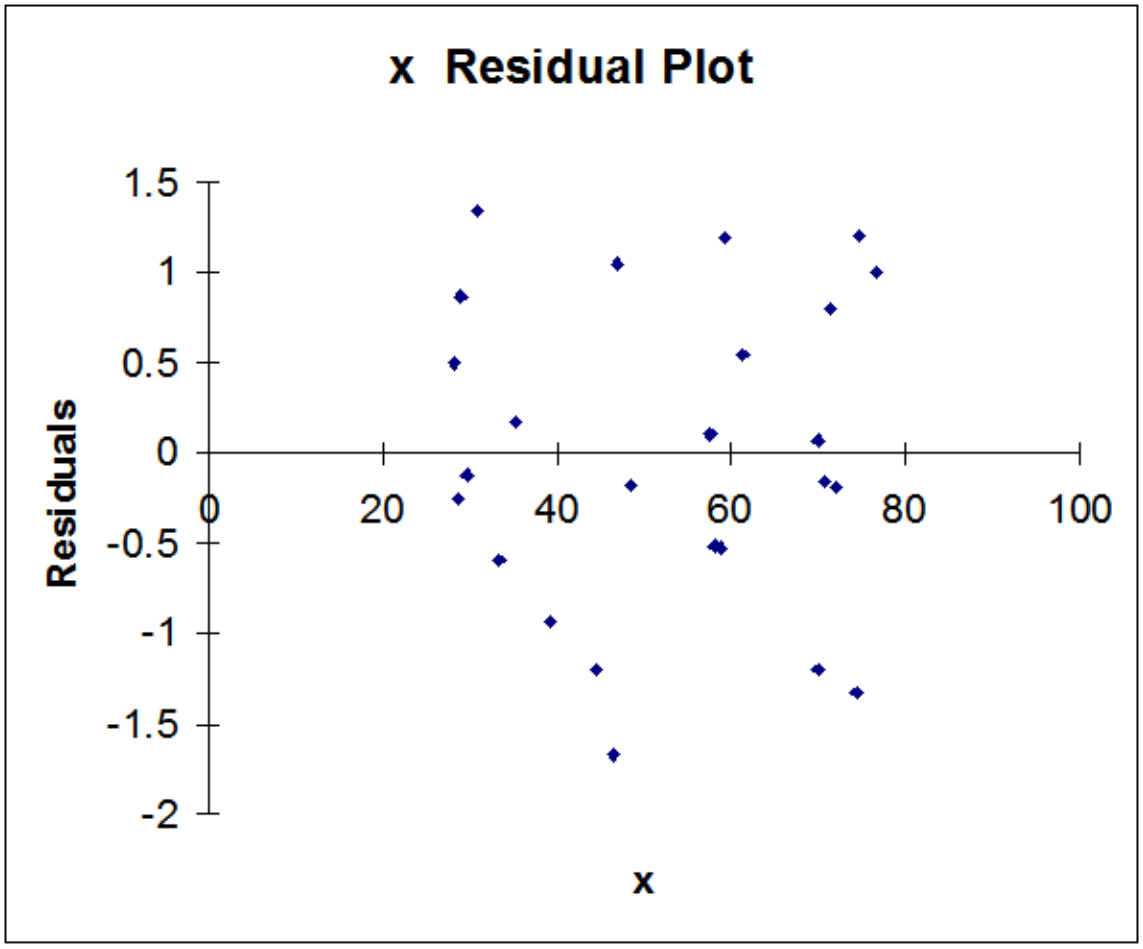
Observation	*Predicted y*	*Residuals*	*Standard Residuals*
1	10.80503639	0.174964	0.200788419
2	11.25207708	-0.12208	-0.140095783
3	11.16426551	1.345734	1.54436632
4	8.929062101	-0.52906	-0.607152228
5	8.721507499	0.548493	0.629450575
6	7.931203435	0.798797	0.916699783
7	7.683734486	-1.32373	-1.519119097
8	7.500128491	0.999872	1.147453602
9	7.979100651	-0.1591	-0.182584076
10	9.032839403	0.107161	0.122977615
11	9.918937899	-1.67894	-1.92675091
12	11.31594003	0.87406	1.003072145
13	11.37980299	0.500197	0.574026623
14	10.50168736	-0.93169	-1.0692054
15	9.887006421	1.052994	1.208416546
16	9.751297643	-0.1713	-0.196581356
17	8.889147755	1.200852	1.378099308
18	8.034980736	0.075019	0.086092186
19	8.034980736	-1.20498	-1.382837169
20	7.675751616	1.204248	1.38199672
21	7.86734048	-0.18734	-0.214992134
22	8.984942187	-0.51494	-0.590948199
23	10.06262955	-1.20263	-1.380138941
24	10.95671091	-0.59671	-0.684786074
25	11.33988864	-0.25989	-0.298248478

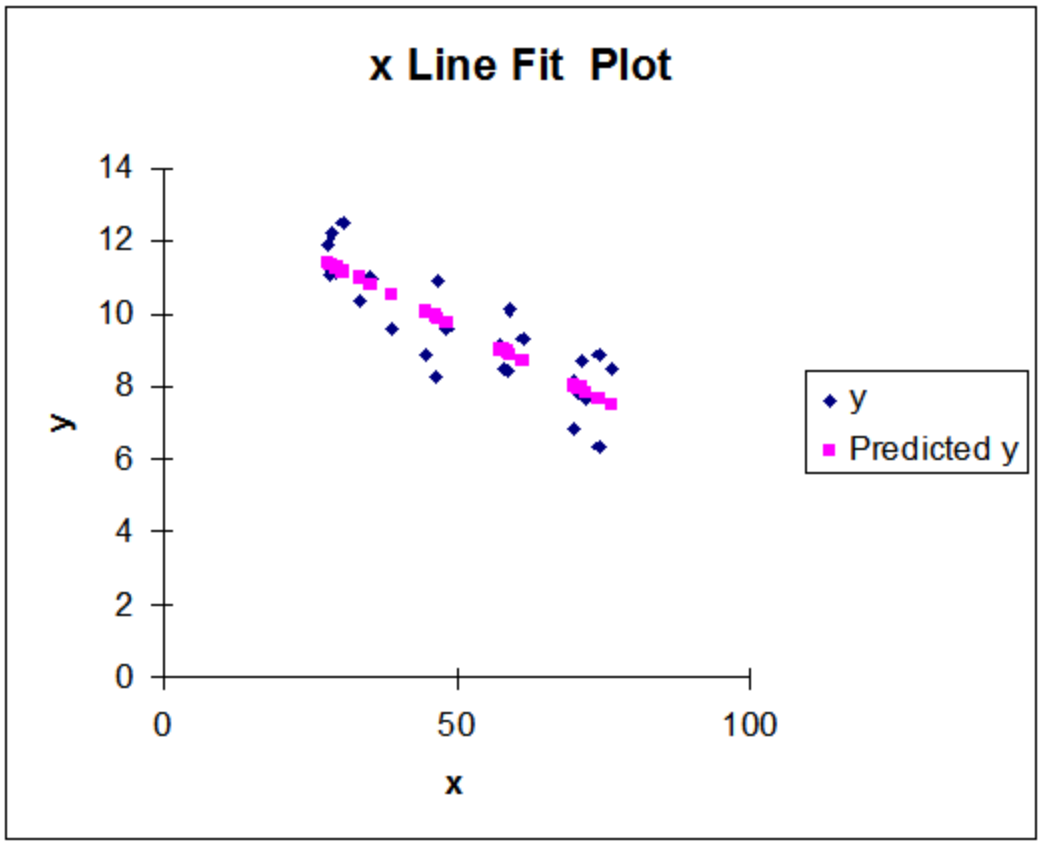
PROBABILITY OUTPUT

Percentile	*y*
2	6.36
6	6.83
10	7.68
14	7.82
18	8.11
22	8.24
26	8.4
30	8.47

34	8.5
38	8.73
42	8.86
46	8.88
50	9.14
54	9.27
58	9.57
62	9.58
66	10.09
70	10.36
74	10.94
78	10.98
82	11.08
86	11.13
90	11.88
94	12.19
98	12.51

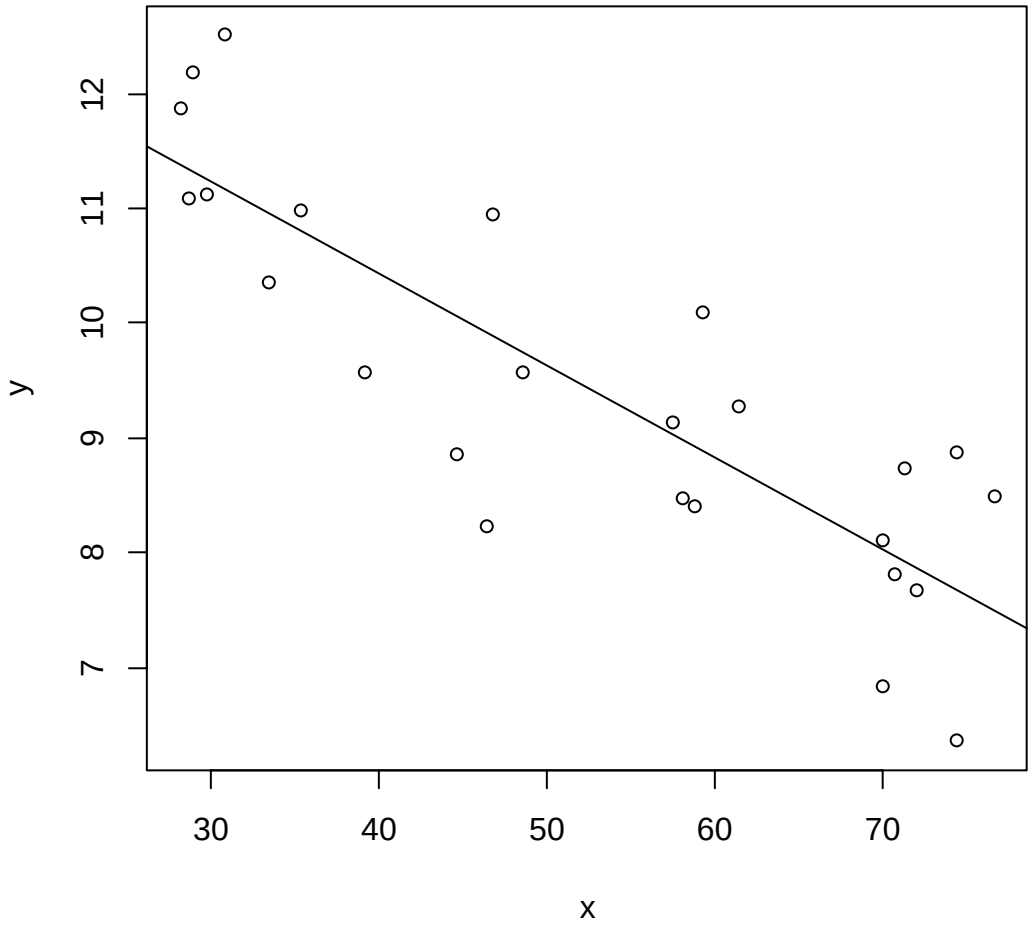



x Residual Plot
Residuals
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
20
40
60
80
100
x


x Line Fit Plot
14
12
10
8
6
4
2
0
y
0
50
100
x
y
Predicted y

بإستخدام R:

```
> rm(list=ls())
> sf = read.table(file.choose(), header = TRUE)
> x = sf$x
> y = sf$y
> fitsf = lm(y ~ x)
> plot(x,y)
> abline(fitsf$coef, lty=7)
>
```

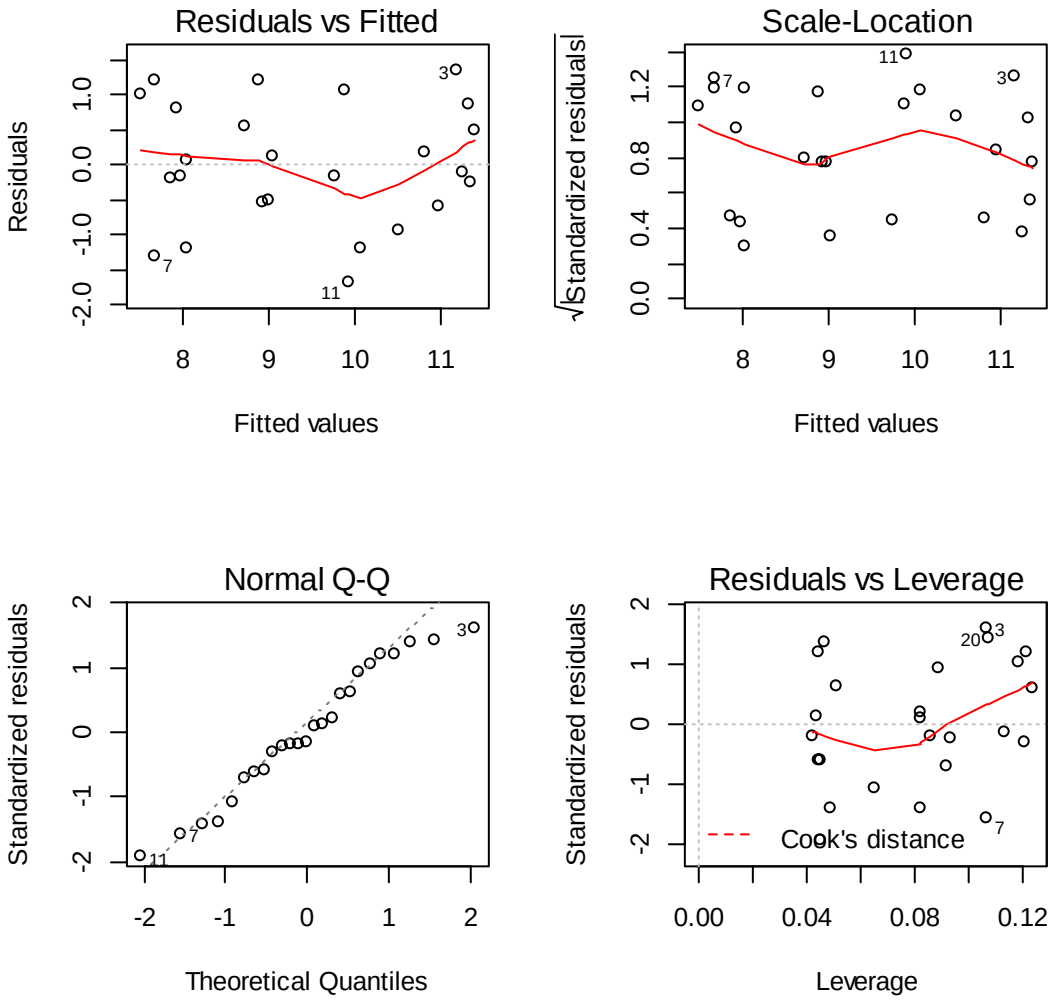



```
> summary(fitsf)

Call:
lm(formula = y ~ x)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-1.6789 -0.5291 -0.1221  0.7988  1.3457

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.62299    0.58146  23.429  < 2e-16 ***
x           -0.07983    0.01052  -7.586 1.05e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.8901 on 23 degrees of freedom
Multiple R-squared: 0.7144,     Adjusted R-squared: 0.702
F-statistic: 57.54 on 1 and 23 DF,  p-value: 1.055e-07

> confint(fitsf)
                2.5 %      97.5 %
(Intercept) 12.4201404 14.82583815
x           -0.1015984 -0.05805901
> anova(fitsf)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 45.592  45.592  57.543 1.055e-07 ***
Residuals 23 18.223   0.792
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
> vcov(fitsf)
             (Intercept)             x
(Intercept)  0.338099795 -0.0058252268
x           -0.005825227  0.0001107458
> influence(fitsf)
$hat
         1          2          3          4          5
6          7          8          9
0.08183288 0.11329874 0.10642607 0.04537290 0.05082408
0.08887748 0.10642607 0.12118198 0.08579127
        10         11         12         13         14
15         16         17         18
0.04335597 0.04537290 0.11850951 0.12389918 0.06547376
0.04470199 0.04234960 0.04627444 0.08231790
        19         20         21         22         23
24         25
0.08231790 0.10703688 0.09314896 0.04422816 0.04894552
0.09152619 0.12050967
```

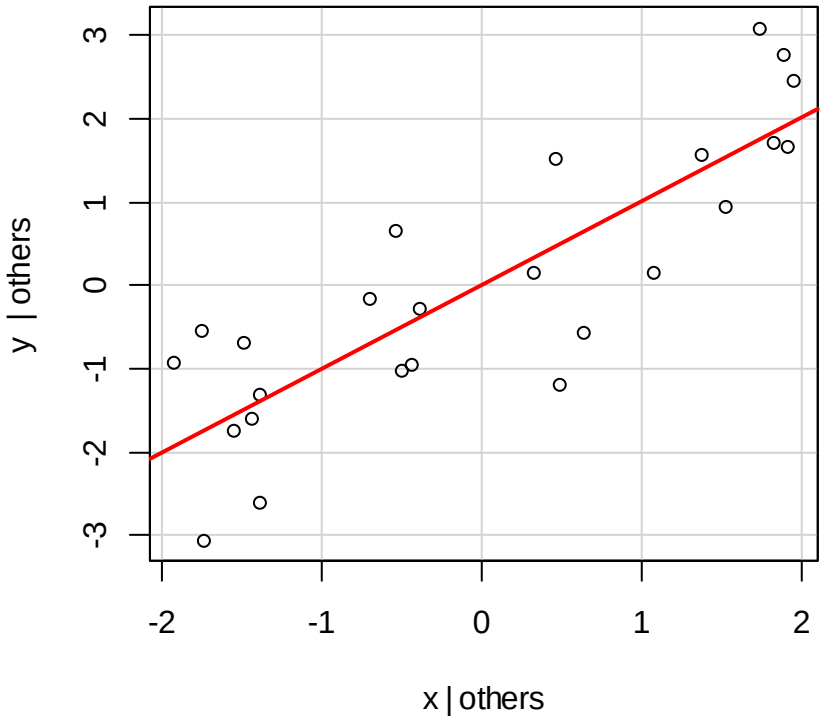
```
$coefficients
     (Intercept)             x
1   0.0318595521 -4.607843e-04
2  -0.0286864966  4.406744e-04
3   0.3016179953 -4.588925e-03
4   0.0030941497 -4.802752e-04
5  -0.0142722999  7.107751e-04
6  -0.0854660866  2.291535e-03
7   0.1781756907 -4.513905e-03
8  -0.1560823084  3.832550e-03
9   0.0161975832 -4.402818e-04
10  0.0004452368  7.671961e-05
11 -0.1505179977  1.524116e-03
12  0.2124387570 -3.284714e-03
13  0.1256779718 -1.955144e-03
14 -0.1388303089  1.881213e-03
15  0.0910937469 -8.935943e-04
16 -0.0125467852  1.025071e-04
17 -0.0116582409  1.179143e-03
18 -0.0071878728  1.988178e-04
19  0.1154536559 -3.193469e-03
20 -0.1631951470  4.128119e-03
21  0.0213536956 -5.630615e-04
22  0.0002351958 -4.141832e-04
23 -0.1249560257  1.413976e-03
24 -0.1189911260  1.762700e-03
25 -0.0639608968  9.912725e-04

$sigma
        1         2         3         4         5         6
7         8         9        10
0.9092969 0.9097100 0.8580290 0.9027782 0.9021803 0.8924705
0.8597681 0.8812645 0.9094381 0.9098300
       11        12        13        14        15        16
17        18        19        20
0.8331372 0.8882236 0.9029703 0.8866315 0.8806691 0.9093644
0.8715551 0.9099767 0.8697224 0.8686288
       21        22        23        24        25
0.9091629 0.9031753 0.8713273 0.9002894 0.9082101

$wt.res
          1           2           3           4           5
6           7           8
 0.17496361 -0.12207708  1.34573449 -0.52906210  0.54849250
0.79879656 -1.32373449  0.99987151
```

```
          9          10          11          12          13
14          15          16
-0.15910065  0.10716060 -1.67893790  0.87405997  0.50019701
-0.93168736  1.05299358 -0.17129764
         17          18          19          20          21
22          23          24
 1.20085225  0.07501926 -1.20498074  1.20424838 -0.18734048
-0.51494219 -1.20262955 -0.59671091
         25
-0.25988864

> layout(matrix(c(1,2,3,4),2,2))
> plot(fitsf)
>
```

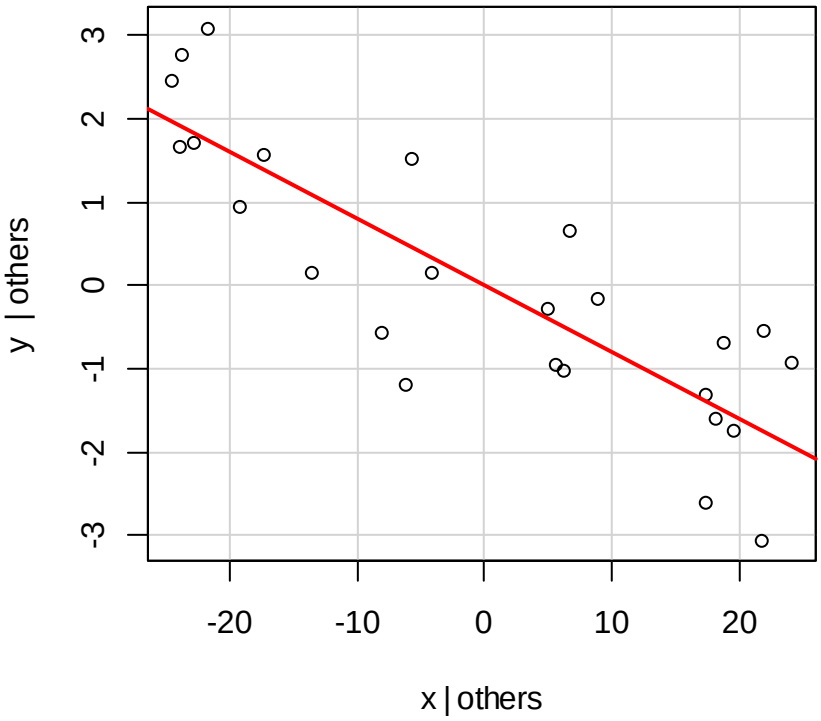



```
> library(car)
```

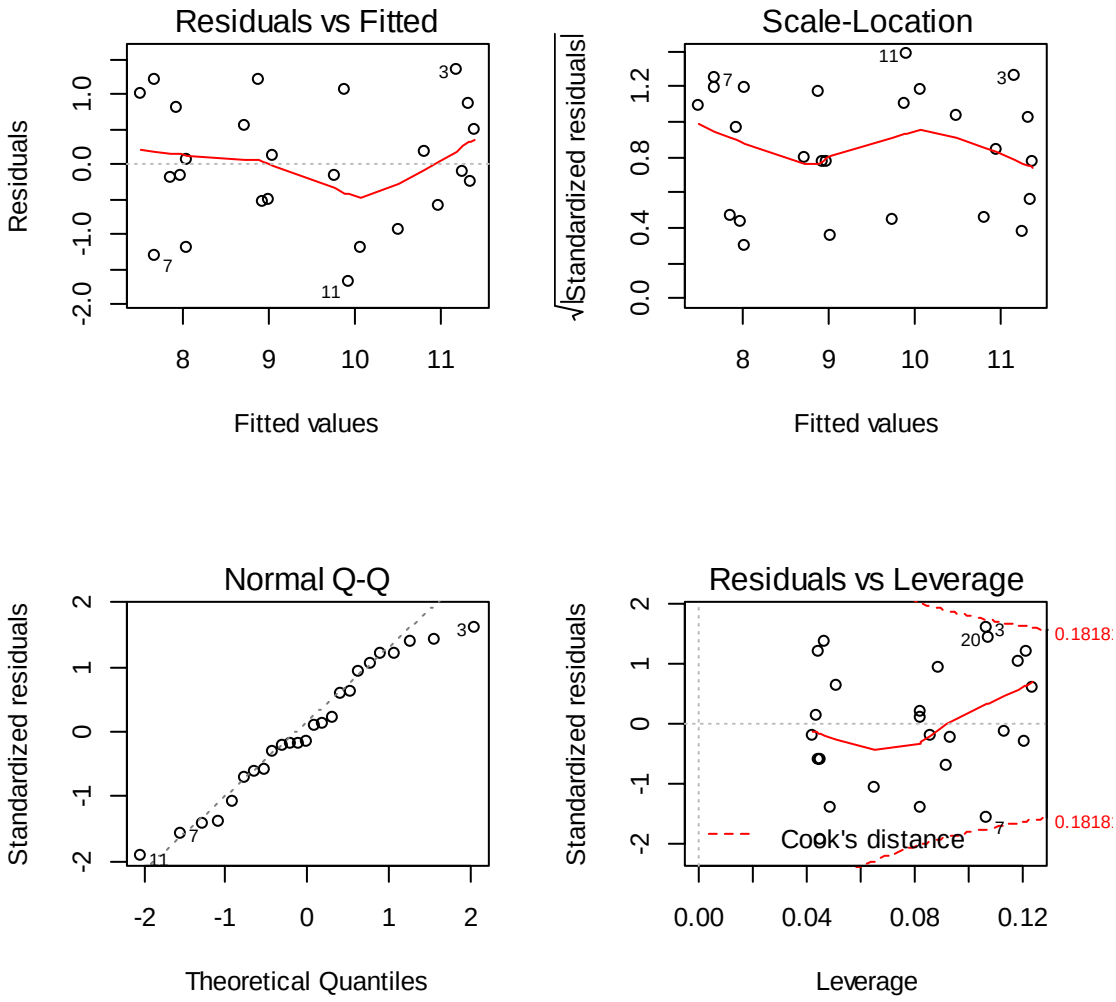
```
Loading required package: MASS
Loading required package: nnet
> outlierTest(fitsf)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
11 -2.062534           0.051159           NA
>
> library(car)
Loading required package: MASS
Loading required package: nnet
> outlierTest(fitsf)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
11 -2.062534           0.051159           NA
> layout(matrix(1), widths=lcm(12), heights=lcm(12))
> leveragePlots(fitsf)
>
```



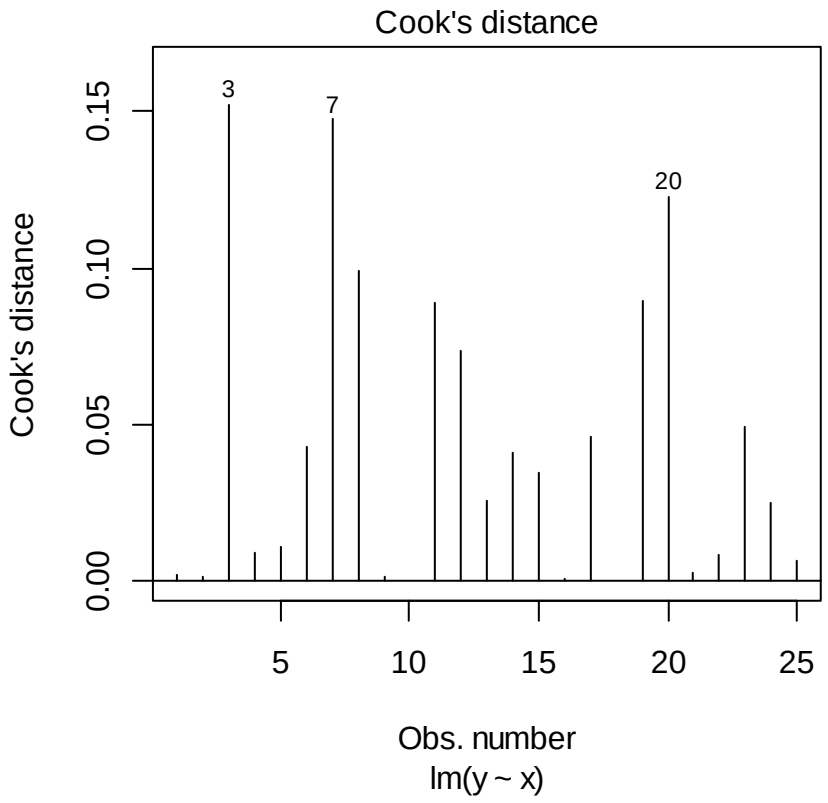

```
> avPlots(fitsf)
>
```




```
> cutoff=4/((length(y)-length(fitsf$coefficients)-1))
> layout(matrix(c(1,2,3,4),2,2))
> plot(fitsf, cook.levels = cutoff)
>
```

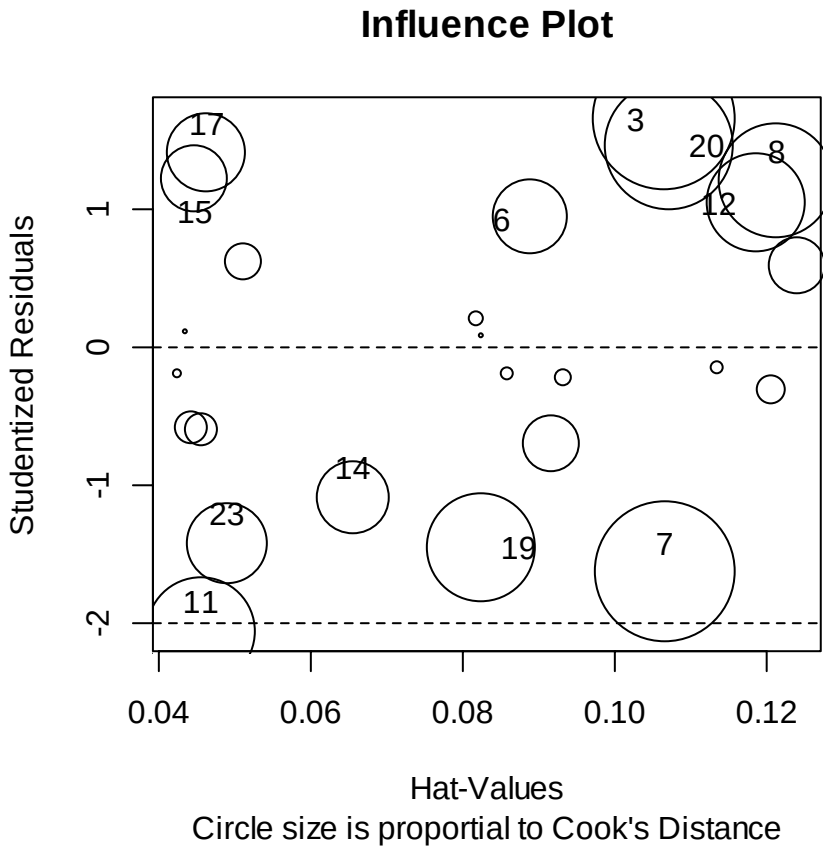



```
> layout(matrix(1), widths=lcm(12), heights=lcm(12))
> plot(fitsf, which = 4, cook.levels = cutoff)
> abline(0,0)
>
```




```
> influencePlot(fitsf, id.method = "identify", main =
"Influence Plot", sub = "Circle size is proportial to
Cook's Distance")
warning: no point within 0.25 inches
warning: no point within 0.25 inches
warning: nearest point already identified
warning: nearest point already identified
warning: no point within 0.25 inches
      StudRes        Hat     CookD
3   1.6591749 0.10642607 0.3902897
6   0.9376784 0.08887748 0.2076297
7  -1.6287495 0.10642607 0.3839093
8   1.2102871 0.12118198 0.3146288
11 -2.0625345 0.04537290 0.2976003
12  1.0481178 0.11850951 0.2711655
14 -1.0870043 0.06547376 0.2026500
```

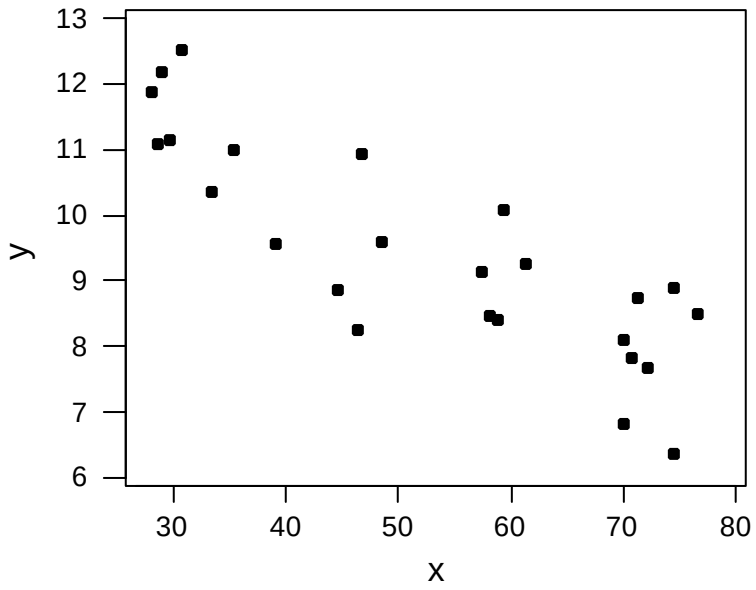
```
15  1.2233297 0.04470199 0.1851333
17  1.4108570 0.04627444 0.2151647
19 -1.4462832 0.08231790 0.2992742
20  1.4671182 0.10703688 0.3504964
23 -1.4152979 0.04894552 0.2222374
>
```



تمرين: يترك إستخراج النتائج والتشخيصات للطالب وتفسير النتائج.

بإستخدام Minitab:

نرسم العلاقة بين المتغيرين في رسم إنتشار



تبدو هناك علاقة خطية عكسية نوجدها كالتالي:

```
MTB > Regress 'y' 1 'x';
SUBC>  GHistogram;
SUBC>  GNormalplot;
SUBC>  GFits;
SUBC>  GOrder;
SUBC>  RType 1;
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: y versus x

The regression equation is

y = 13.6 - 0.0798 x

Predictor	Coef	SE Coef	T	P
Constant	13.6230	0.5815	23.43	0.000
x	-0.07983	0.01052	-7.59	0.000

S = 0.8901 R-Sq = 71.4% R-Sq(adj) = 70.2%

PRESS = 21.4938 R-Sq(pred) = 66.32%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	45.592	45.592	57.54	0.000
Residual Error	23	18.223	0.792		
Lack of Fit	22	17.404	0.791	0.97	0.680
Pure Error	1	0.819	0.819		
Total	24	63.816			

23 rows with no replicates

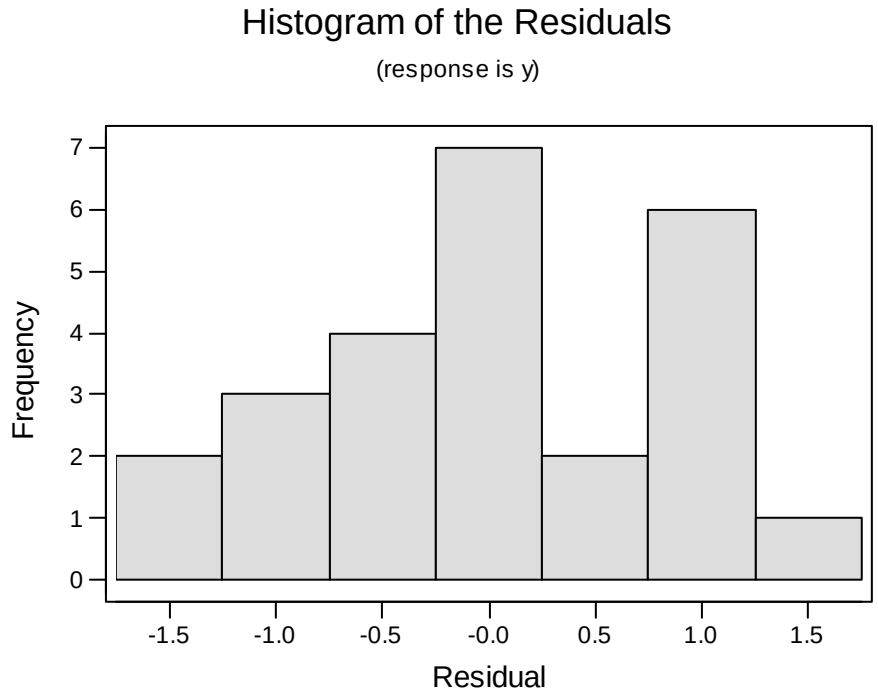
Obs	x	y	Fit	SE Fit	Residual	St Resid
1	35.3	10.980	10.805	0.255	0.175	0.21
2	29.7	11.130	11.252	0.300	-0.122	-0.15
3	30.8	12.510	11.164	0.290	1.346	1.60
4	58.8	8.400	8.929	0.190	-0.529	-0.61
5	61.4	9.270	8.722	0.201	0.548	0.63
6	71.3	8.730	7.931	0.265	0.799	0.94
7	74.4	6.360	7.684	0.290	-1.324	-1.57
8	76.7	8.500	7.500	0.310	1.000	1.20
9	70.7	7.820	7.979	0.261	-0.159	-0.19
10	57.5	9.140	9.033	0.185	0.107	0.12
11	46.4	8.240	9.919	0.190	-1.679	-1.93
12	28.9	12.190	11.316	0.306	0.874	1.05
13	28.1	11.880	11.380	0.313	0.500	0.60
14	39.1	9.570	10.502	0.228	-0.932	-1.08
15	46.8	10.940	9.887	0.188	1.053	1.21
16	48.5	9.580	9.751	0.183	-0.171	-0.20
17	59.3	10.090	8.889	0.191	1.201	1.38
18	70.0	8.110	8.035	0.255	0.075	0.09
19	70.0	6.830	8.035	0.255	-1.205	-1.41

20	74.5	8.880	7.676	0.291	1.204	1.43
21	72.1	7.680	7.867	0.272	-0.187	-0.22
22	58.1	8.470	8.985	0.187	-0.515	-0.59
23	44.6	8.860	10.063	0.197	-1.203	-1.39
24	33.4	10.360	10.957	0.269	-0.597	-0.70
25	28.6	11.080	11.340	0.309	-0.260	-0.31

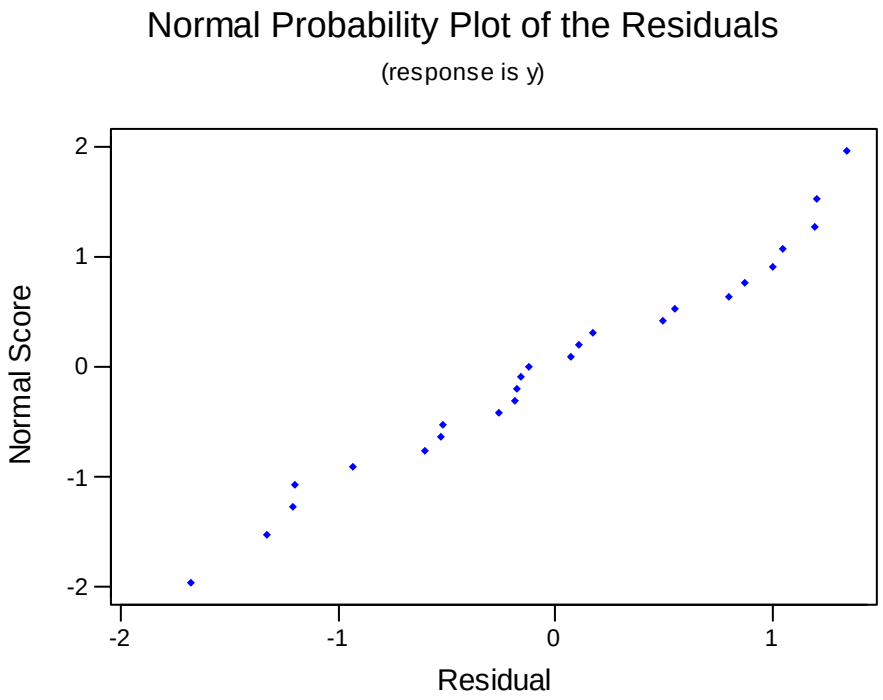
Durbin-Watson statistic = 2.70

No evidence of lack of fit (P > 0.1)

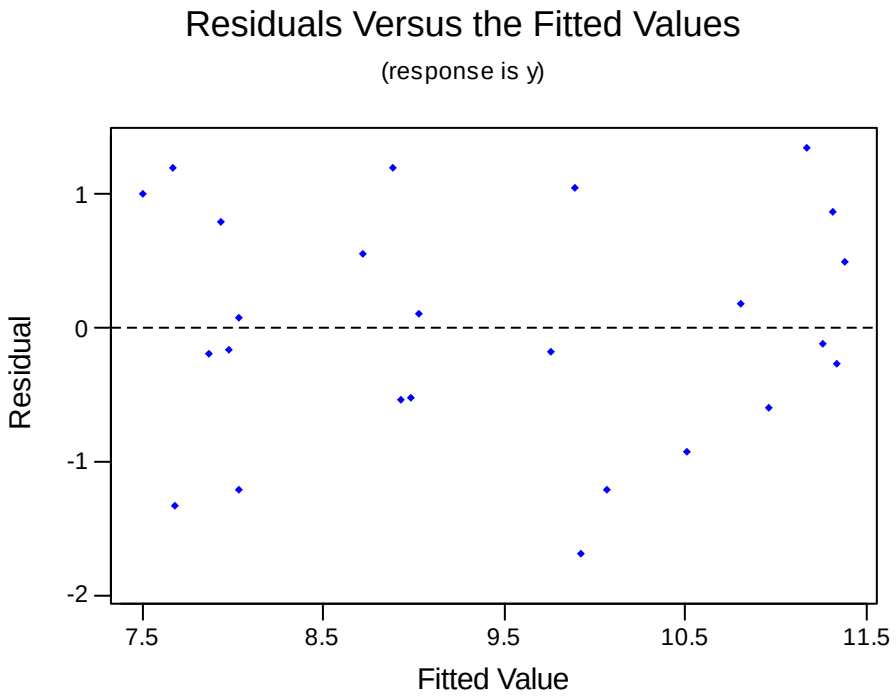
Residual Histogram for y



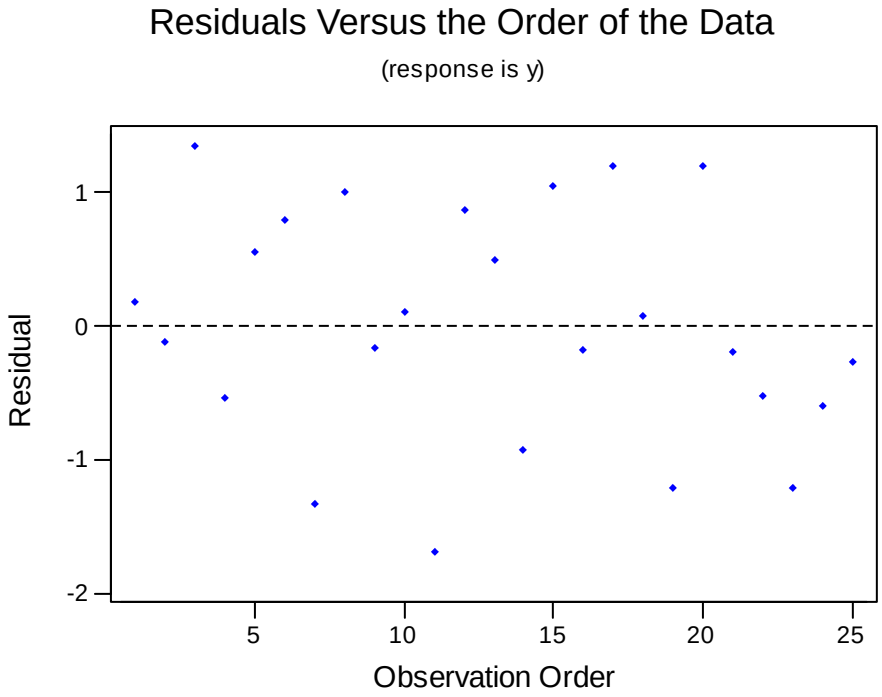
Normplot of Residuals for y



Residuals vs Fits for y



Residuals vs Order for y



مثال (4)

بيانات مصنع يعمل بالبخار

Obs	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	10.98	5.20	0.61	7.4	31	20	22	35.3	54.8	4
2	11.13	5.12	0.64	8.0	29	20	25	29.7	64.0	5
3	12.51	6.19	0.78	7.4	31	23	17	30.8	54.8	4
4	8.40	3.89	0.49	7.5	30	20	22	58.8	56.3	4
5	9.27	6.28	0.84	5.5	31	21	0	61.4	30.3	5
6	8.73	5.76	0.74	8.9	30	22	0	71.3	79.2	4
7	6.36	3.45	0.42	4.1	31	11	0	74.4	16.8	2
8	8.50	6.57	0.87	4.1	31	23	0	76.7	16.8	5
9	7.82	5.69	0.75	4.1	30	21	0	70.7	16.8	4
10	9.14	6.14	0.76	4.5	31	20	0	57.5	20.3	5
11	8.24	4.84	0.65	10.3	30	20	11	46.4	106.1	4
12	12.19	4.88	0.62	6.9	31	21	12	28.9	47.6	4
13	11.88	6.03	0.79	6.6	31	21	25	28.1	43.6	5
14	9.57	4.55	0.60	7.3	28	19	18	39.1	53.3	5
15	10.94	5.71	0.70	8.1	31	23	5	46.8	65.6	4
16	9.58	5.67	0.74	8.4	30	20	7	48.5	70.6	4
17	10.09	6.72	0.85	6.1	31	22	0	59.3	37.2	6
18	8.11	4.95	0.67	4.9	30	22	0	70.0	24.0	4
19	6.83	4.62	0.45	4.6	31	11	0	70.0	21.2	3
20	8.88	6.60	0.95	3.7	31	23	0	74.5	13.7	4
21	7.68	5.01	0.64	4.7	30	20	0	72.1	22.1	4
22	8.47	5.68	0.75	5.3	31	21	1	58.1	28.1	6
23	8.86	5.28	0.70	6.2	30	20	14	44.6	38.4	4
24	10.36	5.36	0.67	6.8	31	20	22	33.4	46.2	4
25	11.08	5.87	0.70	7.5	31	22	28	28.6	56.3	5

الحل بواسطة Excel:

SUMMARY
OUTPUT

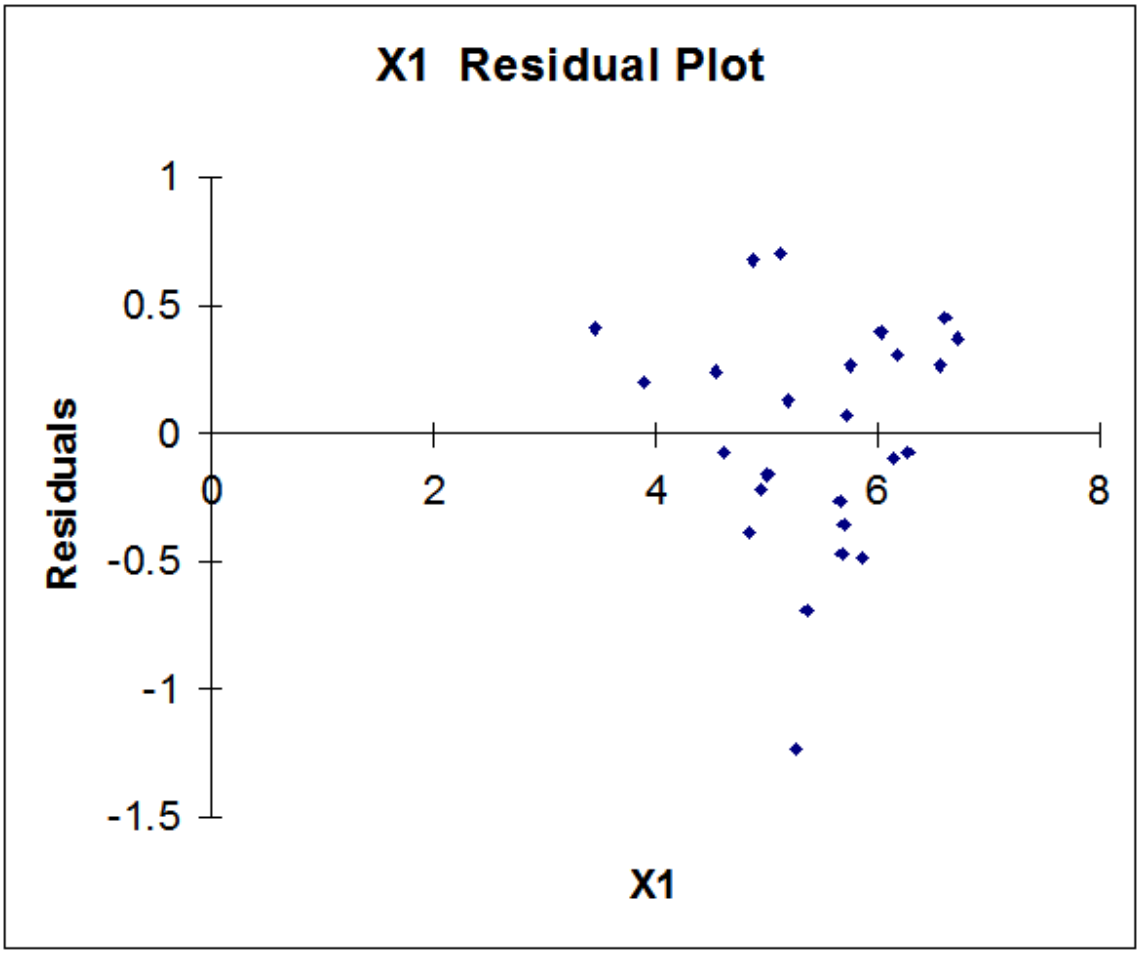
Regression Statistics	
Multiple R	0.961093086
R Square	0.923699919
Adjusted R Square	0.87791987

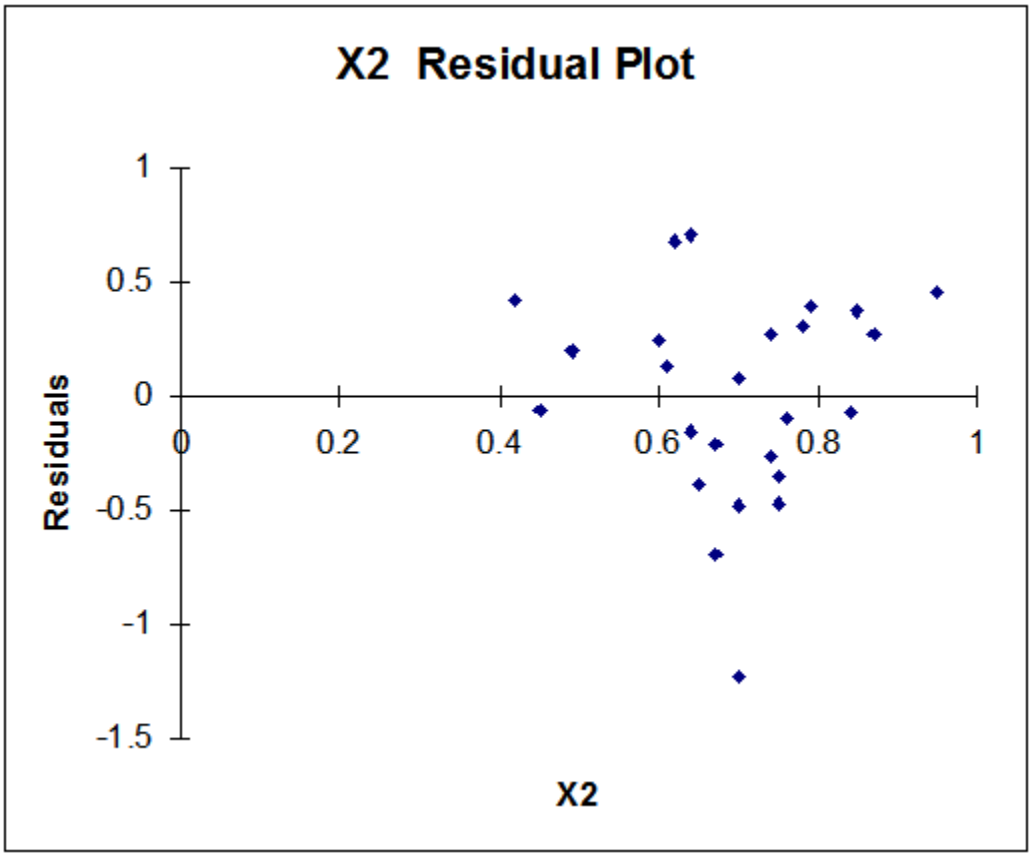
Standard Error	0.569745599
Observations	25

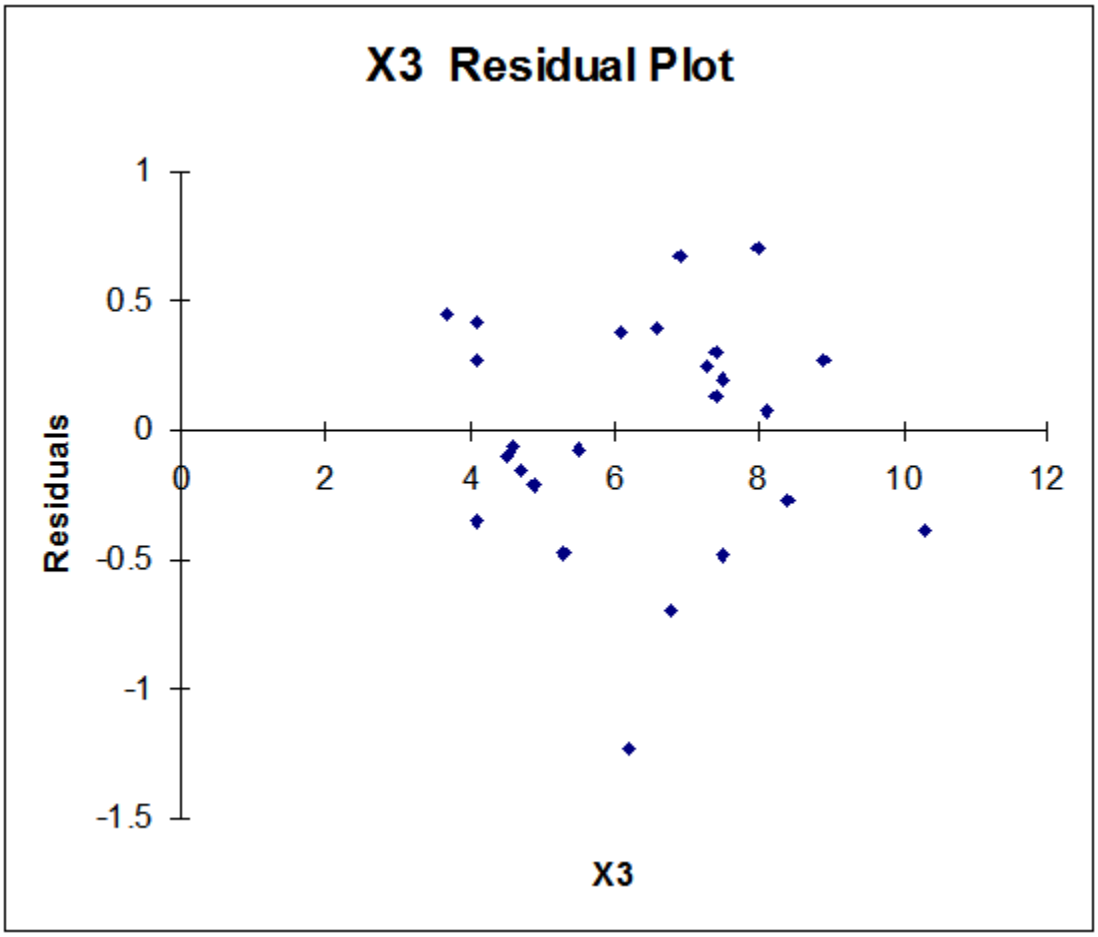
ANOVA

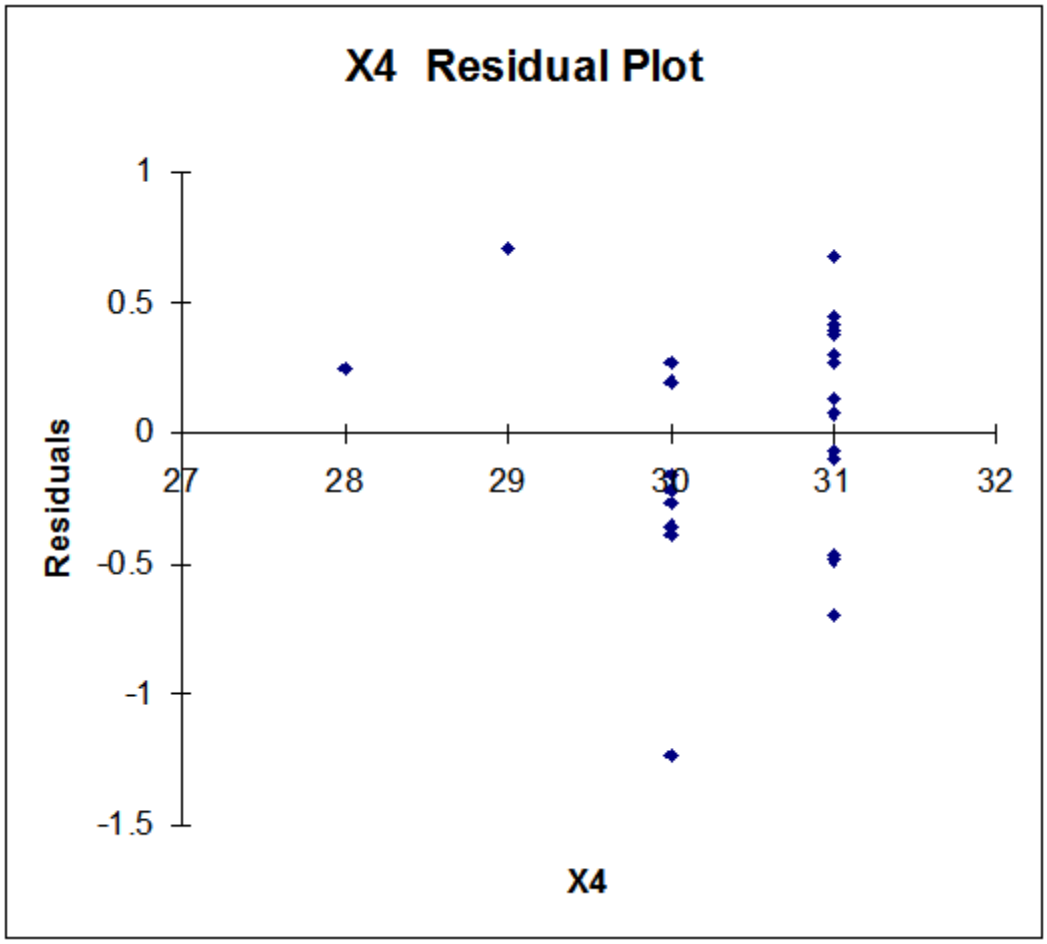
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	58.94665	6.549628	20.17691	7.9699E-07
Residual	15	4.869151	0.32461		
Total	24	63.8158			

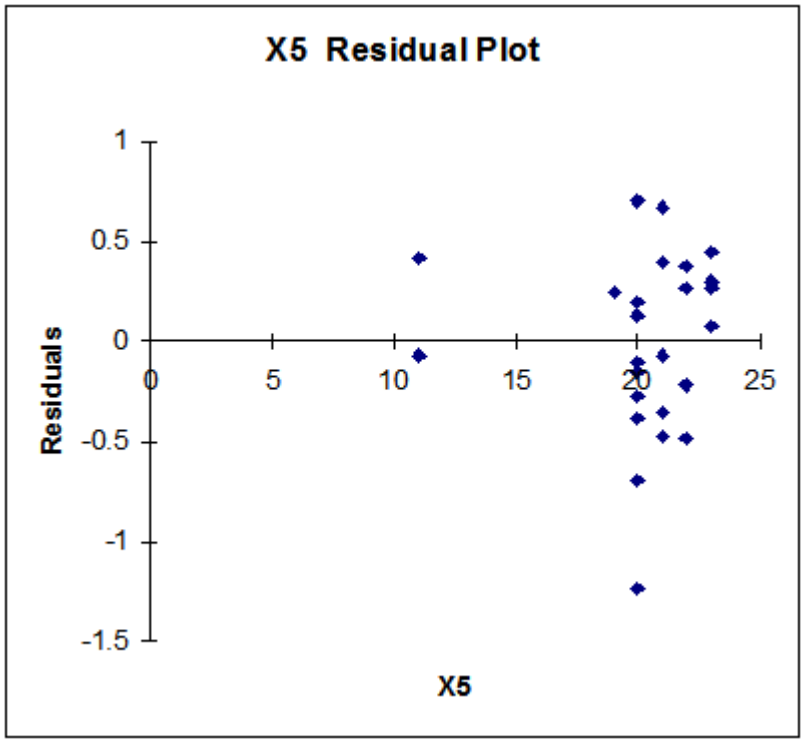
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.894213294	6.996358848	0.270742773	0.790278926	-13.01817252	16.80659911	-13.01817252	16.80659911
X1	0.705406145	0.564900703	1.248725912	0.230903451	-0.498651196	1.909463486	-0.498651196	1.909463486
X2	-1.893719353	4.146291477	-0.456726056	0.654412006	-10.73133304	6.943891691	-10.73133304	6.943891691
X3	1.134220124	0.746089998	1.520218911	0.149252575	-0.456033055	2.724473303	-0.456033055	2.724473303
X4	-0.118762995	0.204614466	-0.580423257	0.570246821	-0.317362414	0.554888404	-0.317362414	0.554888404
X5	-0.179345234	0.080948622	-2.215544008	0.042611222	-0.006807332	0.351883136	-0.006807332	0.351883136
X6	-0.018178638	0.024508026	-0.741742231	0.469698884	-0.070416258	0.034058982	-0.070416258	0.034058982
X7	-0.077417474	0.016591896	-4.665981085	0.000304499	-0.112782264	-0.042052684	-0.112782264	-0.042052684
X8	-0.085846561	0.051998921	-1.650929664	0.119528831	-0.196679636	0.024986515	-0.196679636	0.024986515
X9	0.345005329	0.210695918	1.637456158	0.122337337	-0.794093046	0.104082388	-0.794093046	0.104082388

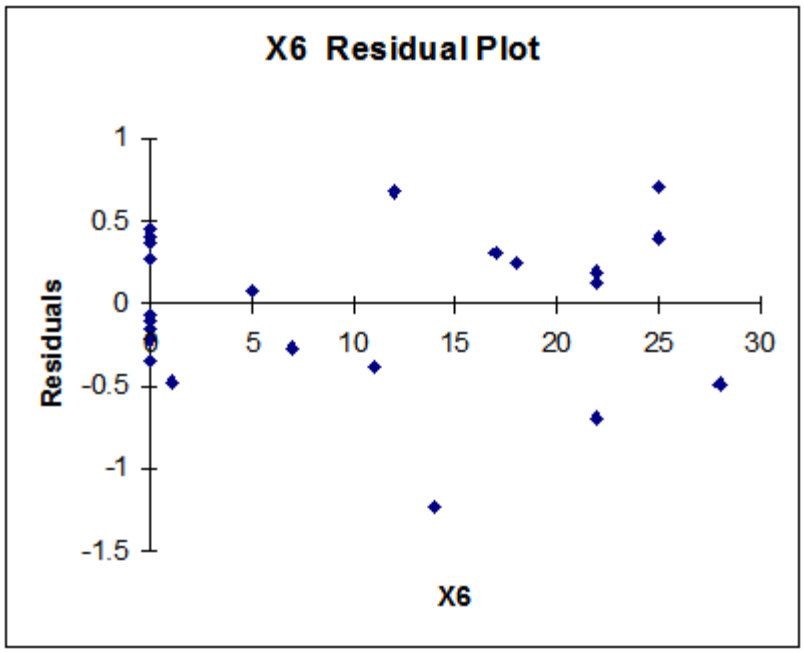

X1 Residual Plot
1
0.5
0
-0.5
-1
-1.5
Residuals
0
2
4
6
8
X1

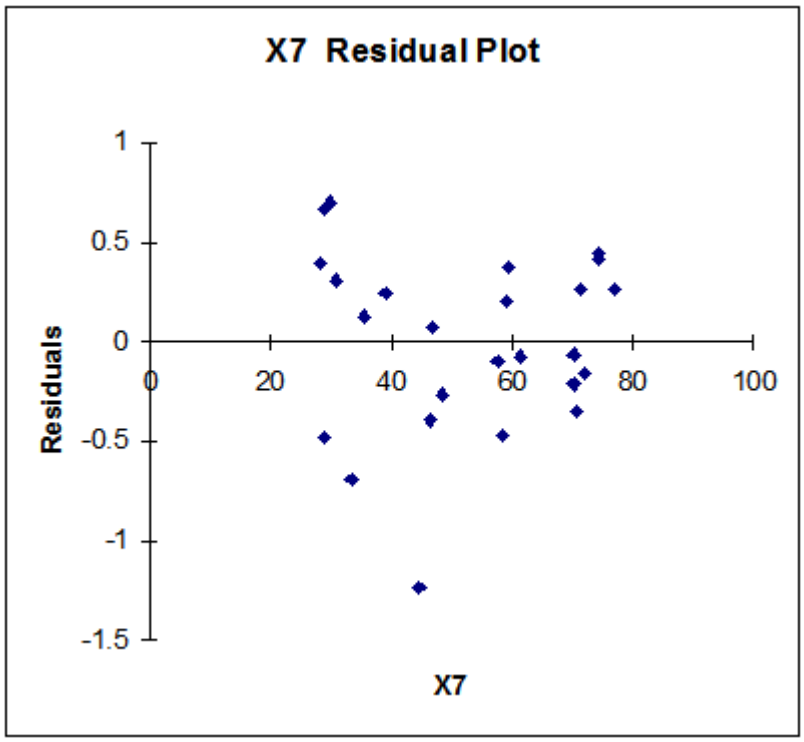

X2 Residual Plot
1
0.5
0
-0.5
-1
-1.5
0
0.2
0.4
0.6
0.8
1
Residuals
X2

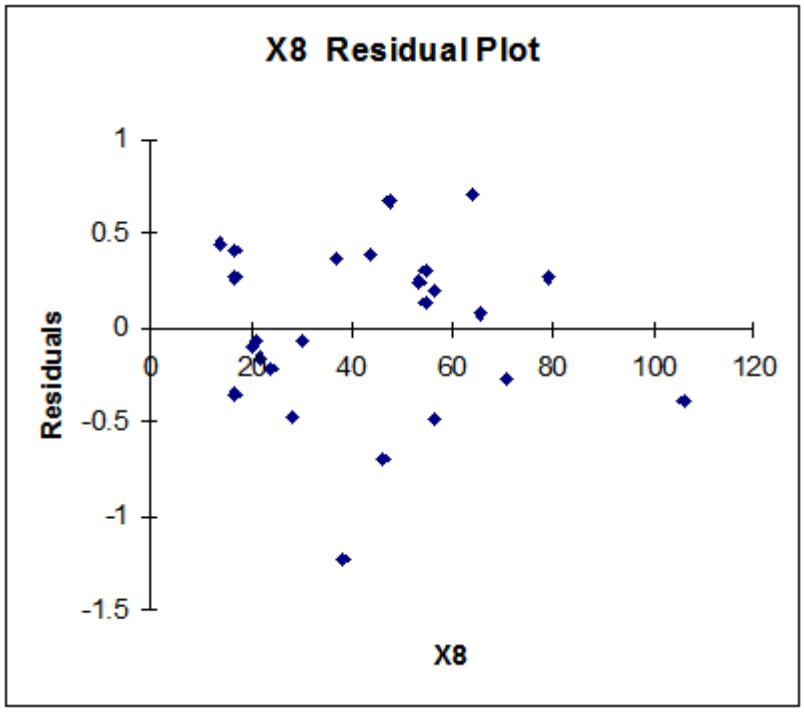

X3 Residual Plot
1
0.5
0
-0.5
-1
-1.5
Residuals
0
2
4
6
8
10
12
X3

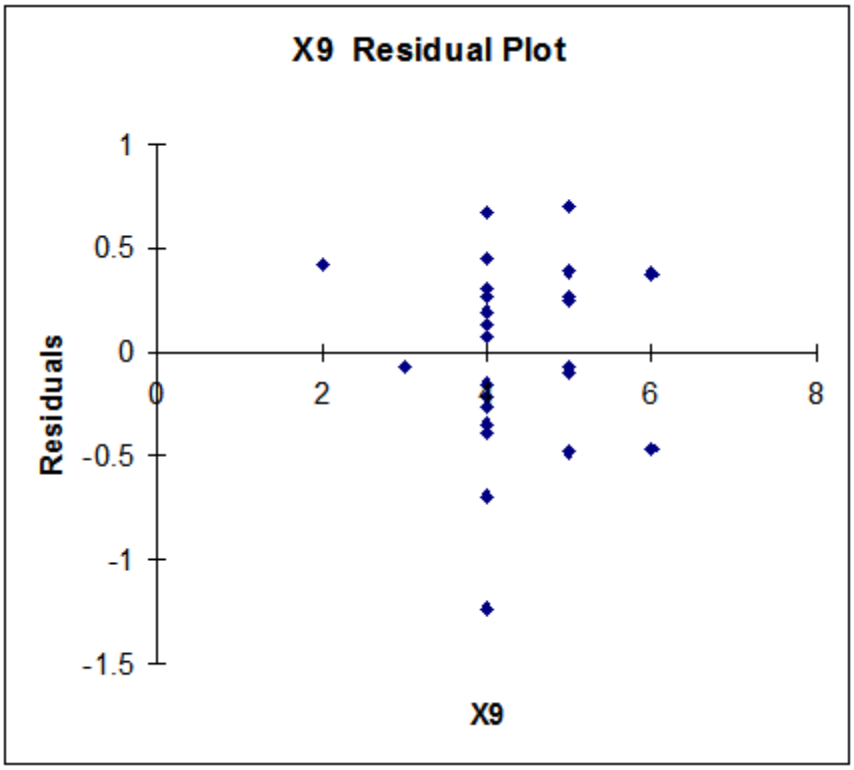

X4 Residual Plot
Residuals
X4
1
0.5
0
-0.5
-1
-1.5
27
28
29
30
31
32

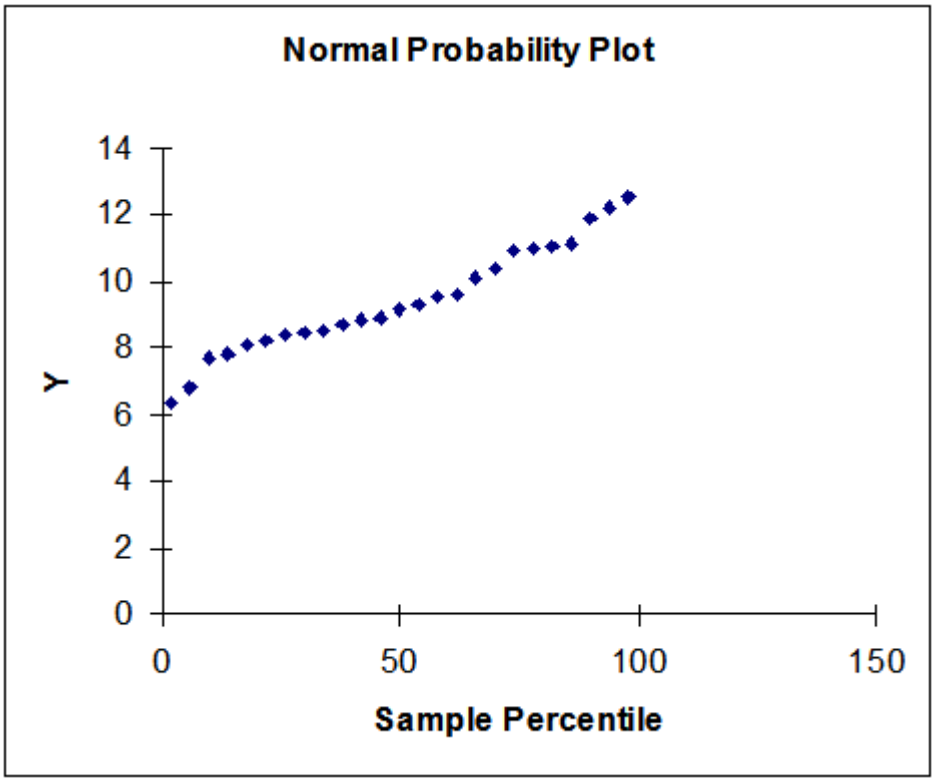

X5 Residual Plot
Residuals
X5
1
0.5
0
-0.5
-1
-1.5
0
5
10
15
20
25


X6 Residual Plot
Residuals
1
0.5
0
-0.5
-1
-1.5
5
10
15
20
25
30
X6


X7 Residual Plot
Residuals
1
0.5
0
-0.5
-1
-1.5
0
20
40
60
80
100
X7


X8 Residual Plot
Residuals
X8
1
0.5
0
-0.5
-1
-1.5
0
20
40
60
80
100
120


X9 Residual Plot
Residuals
X9
1
0.5
0
-0.5
-1
-1.5
0
2
6
8


Normal Probability Plot
14
12
10
8
6
4
2
0
Y
0
50
100
150
Sample Percentile

الحل بواسطة R:

```
> rm(list=ls())
> stf =read.table(file.choose(), header = TRUE)
> stf
   Obs     Y   X1   X2   X3 X4 X5 X6   X7    X8 X9
1    1 10.98 5.20 0.61  7.4 31 20 22 35.3  54.8  4
2    2 11.13 5.12 0.64  8.0 29 20 25 29.7  64.0  5
3    3 12.51 6.19 0.78  7.4 31 23 17 30.8  54.8  4
4    4  8.40 3.89 0.49  7.5 30 20 22 58.8  56.3  4
5    5  9.27 6.28 0.84  5.5 31 21  0 61.4  30.3  5
6    6  8.73 5.76 0.74  8.9 30 22  0 71.3  79.2  4
7    7  6.36 3.45 0.42  4.1 31 11  0 74.4  16.8  2
8    8  8.50 6.57 0.87  4.1 31 23  0 76.7  16.8  5
9    9  7.82 5.69 0.75  4.1 30 21  0 70.7  16.8  4
10  10  9.14 6.14 0.76  4.5 31 20  0 57.5  20.3  5
11  11  8.24 4.84 0.65 10.3 30 20 11 46.4 106.1  4
12  12 12.19 4.88 0.62  6.9 31 21 12 28.9  47.6  4
13  13 11.88 6.03 0.79  6.6 31 21 25 28.1  43.6  5
14  14  9.57 4.55 0.60  7.3 28 19 18 39.1  53.3  5
15  15 10.94 5.71 0.70  8.1 31 23  5 46.8  65.6  4
16  16  9.58 5.67 0.74  8.4 30 20  7 48.5  70.6  4
17  17 10.09 6.72 0.85  6.1 31 22  0 59.3  37.2  6
18  18  8.11 4.95 0.67  4.9 30 22  0 70.0  24.0  4
19  19  6.83 4.62 0.45  4.6 31 11  0 70.0  21.2  3
20  20  8.88 6.60 0.95  3.7 31 23  0 74.5  13.7  4
21  21  7.68 5.01 0.64  4.7 30 20  0 72.1  22.1  4
22  22  8.47 5.68 0.75  5.3 31 21  1 58.1  28.1  6
23  23  8.86 5.28 0.70  6.2 30 20 14 44.6  38.4  4
24  24 10.36 5.36 0.67  6.8 31 20 22 33.4  46.2  4
25  25 11.08 5.87 0.70  7.5 31 22 28 28.6  56.3  5
> fitst = lm( Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
X9, data = stf)
> summary(fitst)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
X9,
    data = stf)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23126 -0.26697  0.07533  0.30451  0.70427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.89421    6.99636   0.271 0.790279
```

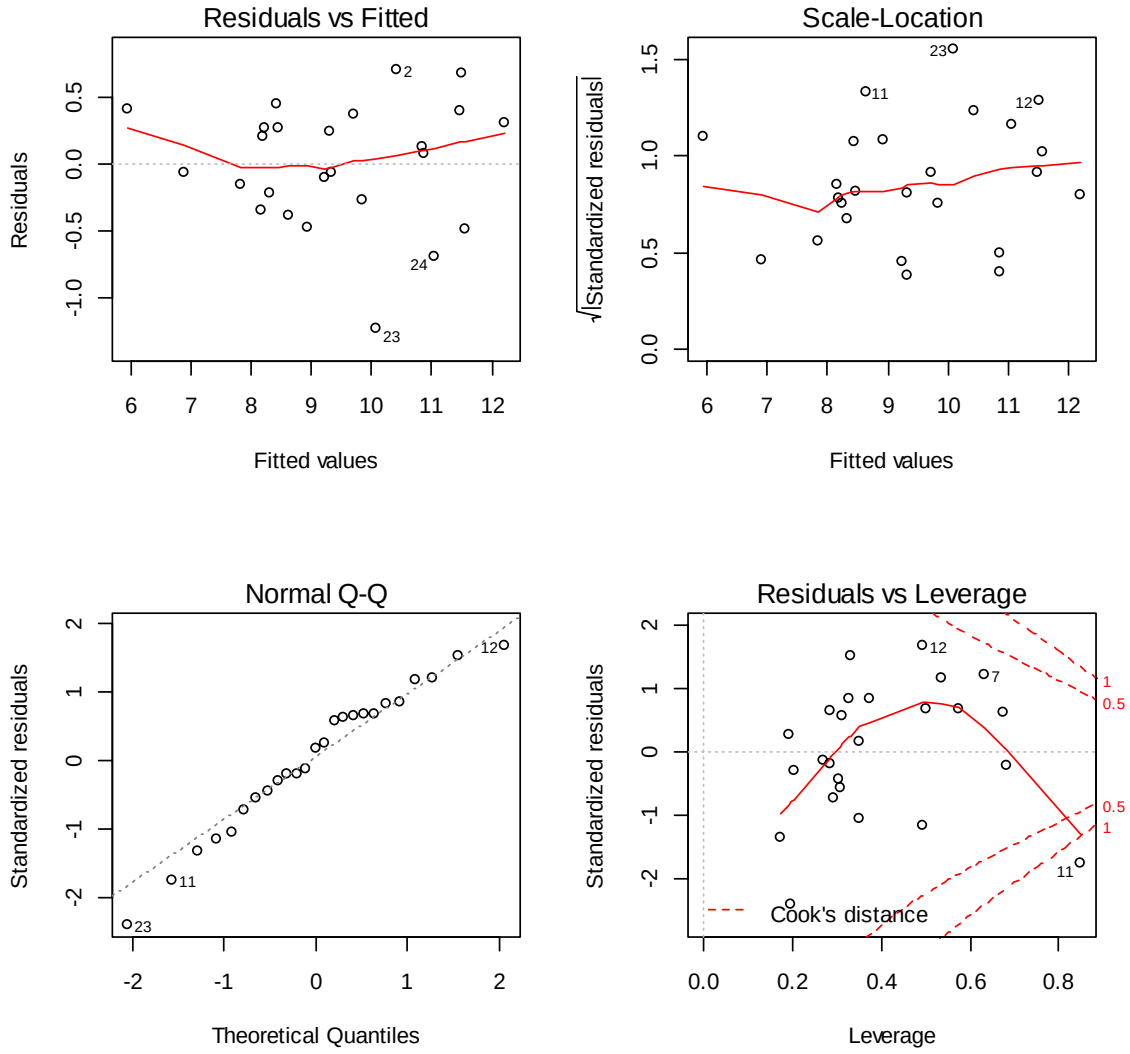
```
X1           0.70541    0.56490   1.249 0.230903
X2          -1.89372    4.14629  -0.457 0.654412
X3           1.13422    0.74609   1.520 0.149253
X4           0.11876    0.20461   0.580 0.570247
X5           0.17935    0.08095   2.216 0.042611 *
X6          -0.01818    0.02451  -0.742 0.469699
X7          -0.07742    0.01659  -4.666 0.000304 ***
X8          -0.08585    0.05200  -1.651 0.119529
X9          -0.34501    0.21070  -1.637 0.122337
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.5697 on 15 degrees of freedom
Multiple R-squared: 0.9237,     Adjusted R-squared: 0.8779
F-statistic: 20.18 on 9 and 15 DF,  p-value: 7.97e-07

> confint(fitst)
                    2.5 %      97.5 %
(Intercept) -13.018172593 16.80659918
X1           -0.498651201  1.90946349
X2          -10.731330438  6.94389173
X3           -0.456033062  2.72447331
X4           -0.317362416  0.55488841
X5            0.006807331  0.35188314
X6           -0.070416258  0.03405898
X7           -0.112782264 -0.04205268
X8           -0.196679637  0.02498652
X9           -0.794093048  0.10408239
> anova(fitst)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1  9.3698  9.3698 28.8647 7.760e-05 ***
X2         1  1.8295  1.8295  5.6360 0.0313762 *
X3         1 16.9620 16.9620 52.2535 2.932e-06 ***
X4         1  0.7259  0.7259  2.2363 0.1555476
X5         1  5.5695  5.5695 17.1574 0.0008688 ***
X6         1 13.2026 13.2026 40.6723 1.242e-05 ***
X7         1  9.8937  9.8937 30.4787 5.873e-05 ***
X8         1  0.5233  0.5233  1.6120 0.2235554
X9         1  0.8704  0.8704  2.6813 0.1223373
Residuals 15  4.8692  0.3246
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
```

```
> vcov(fitst)
             (Intercept)           X1            X2
X3            X4            X5
(Intercept)  48.94903713  1.673886033 -10.167034461 -
2.2429638221 -1.2904945886  0.0776986442
X1            1.67388603  0.319112804  -2.018603483 -
0.1120892239 -0.0582188944  0.0135525752
X2          -10.16703446 -2.018603483  17.191733015
1.3288420150  0.2698644583 -0.2088262840
X3           -2.24296382 -0.112089224   1.328842015
0.5566502847  0.0125570129 -0.0214247909
X4           -1.29049459 -0.058218894   0.269864458
0.0125570129  0.0418670797 -0.0011063551
X5            0.07769864  0.013552575  -0.208826284 -
0.0214247909 -0.0011063551  0.0065526794
X6           -0.04413755  0.001060047   0.008041426
0.0001915862  0.0005676597 -0.0003717627
X7           -0.05646543  0.000188617   0.007288507
0.0050532717  0.0006243732 -0.0002600118
X8            0.13848178  0.007277608  -0.086462577 -
0.0384576770 -0.0004240573  0.0012996046
X9           -0.36821699 -0.030060881   0.011290039 -
0.0365417280  0.0154060223 -0.0019175442
                        X6            X7            X8
X9
(Intercept) -4.413755e-02 -0.0564654319  1.384818e-01 -
3.682170e-01
X1           1.060047e-03  0.0001886170  7.277608e-03 -
3.006088e-02
X2           8.041426e-03  0.0072885067 -8.646258e-02
1.129004e-02
X3           1.915862e-04  0.0050532717 -3.845768e-02 -
3.654173e-02
X4           5.676597e-04  0.0006243732 -4.240573e-04
1.540602e-02
X5          -3.717627e-04 -0.0002600118  1.299605e-03 -
1.917544e-03
X6           6.006433e-04  0.0003025874  3.772940e-06
4.649118e-05
X7           3.025874e-04  0.0002752910 -3.134071e-04
1.906615e-04
X8           3.772940e-06 -0.0003134071  2.703888e-03
2.841869e-03
X9           4.649118e-05  0.0001906615  2.841869e-03
4.439277e-02
> influence(fitst)
$hat
```

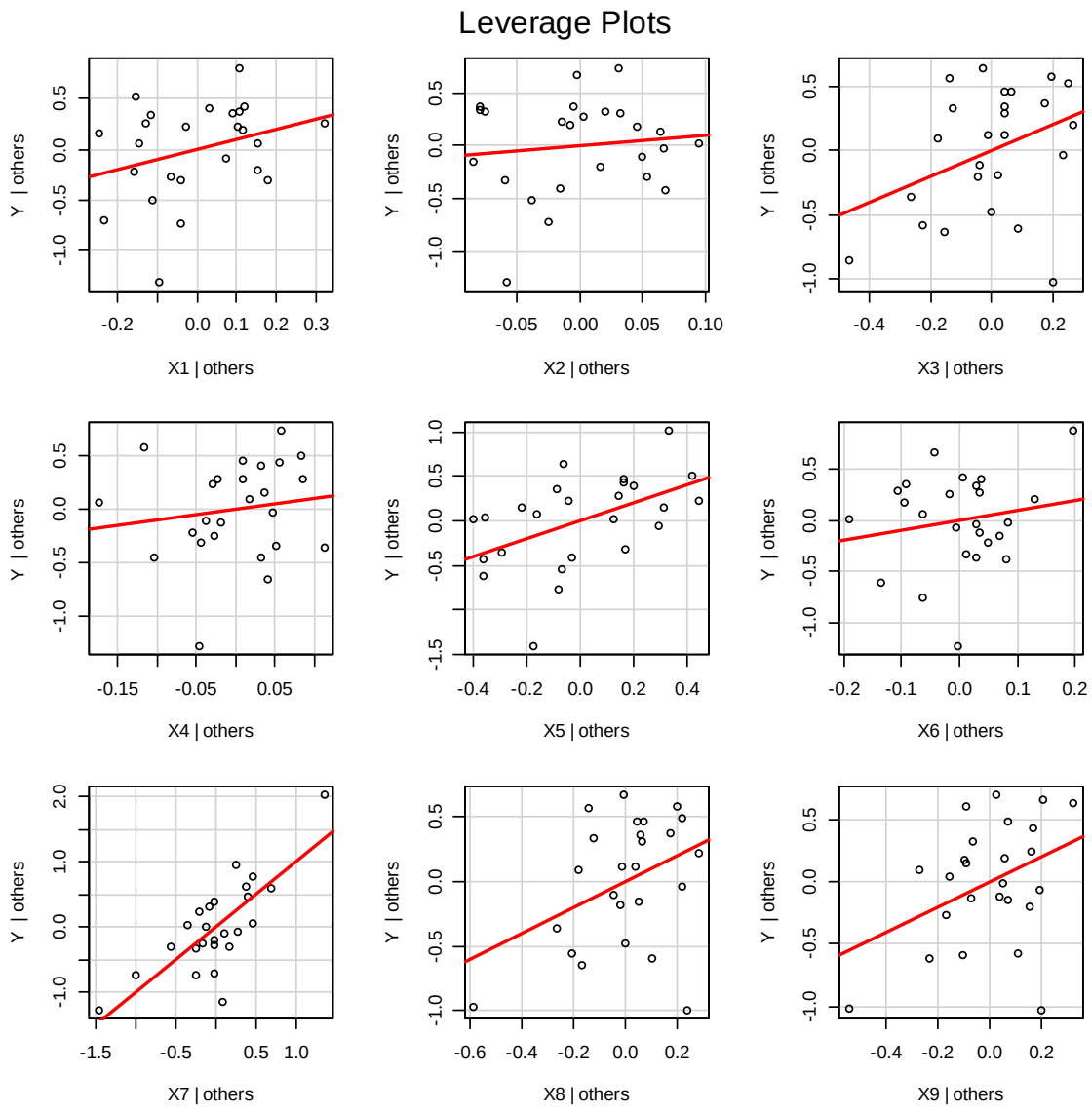
```
        1         2         3         4         5         6
7         8         9
0.1913205 0.3312540 0.2856863 0.6738105 0.2706685 0.5030608
0.6337615 0.3141305 0.2926615
       10        11        12        13        14        15
16        17        18
0.2836230 0.8499781 0.4935469 0.3292774 0.5737078 0.3494442
0.3078923 0.3730104 0.3053461
       19        20        21        22        23        24
25
0.6817352 0.5345388 0.2027363 0.4955507 0.1980320 0.1751443
0.3500826

$coefficients
   (Intercept)          X1          X2           X3
X4           X5            X6
1  -0.21800312  0.02309612 -0.20170881 -0.003078935
0.006337641  0.002536393  0.0010493734
2   4.67054029  0.16024224 -0.90496918 -0.216294175 -
0.132241793 -0.007849371  0.0047781929
3   0.25000355  0.07163349 -0.24717145  0.027132021 -
0.010182219  0.007611083 -0.0012023572
4  -2.75949908 -0.12404022  0.13696609  0.159805659
0.056766979  0.013685499  0.0119272088
5   0.26626985  0.02144728 -0.22760488 -0.039162627 -
0.004964470  0.003931231  0.0002823905
6  -0.72265679  0.06992953 -0.04494625  0.205090728 -
0.016980574 -0.002949093  0.0008940449
7  -1.75858516 -0.39111529  2.50730422  0.073605141
0.102336994 -0.051147083 -0.0004904631
8  -0.19044307  0.05764782 -0.35807653 -0.102860036
0.004200718  0.007120669  0.0037903570
9  -2.18950791 -0.10640395  0.75367933  0.168461624
0.055803969 -0.016145153  0.0006494244
10 -0.29030122 -0.02963009  0.25938111  0.054950371
0.002687165 -0.001879493  0.0011918765
11 -2.34370200  0.41093054 -1.13624032  1.838085286 -
0.145529182  0.009531227  0.0078228559
12 -0.32331508 -0.28603477  0.11168035 -0.058565130
0.083378730  0.049525769 -0.0268993117
13 -1.18922963 -0.10402138  1.23531676  0.056731597
0.035024828 -0.023733585  0.0064468363
14  3.48453612 -0.02146809  0.23000291  0.037004134 -
0.108679431 -0.010583304 -0.0020919918
15 -0.08846768  0.01901102 -0.20622454  0.007498724
0.002137001  0.005470621 -0.0015193366
```

```
16 -0.20321695  0.02073820 -0.64374349 -0.135533684
0.018526236  0.015716556  0.0019491122
17 -1.52071827  0.02511348  0.07711466  0.177248953
0.021073715 -0.014935317 -0.0022549661
18 -0.30811002  0.02748952  0.13967582 -0.009526192
0.009020142 -0.015336667  0.0021380538
19 -0.34410654 -0.09694565  0.57465226  0.013198401
0.009056595  0.007203387 -0.0001119971
20  0.11436858 -0.15456629  2.15894791 -0.180993544
0.008500937 -0.009918716  0.0090300660
21 -0.39169888 -0.01988675  0.27624733  0.013022469
0.011767491 -0.006967666  0.0006707893
22  3.08831617  0.30404771 -1.01934061 -0.001352524 -
0.114536595  0.007627804  0.0078127568
23 -1.47348222  0.19968596 -2.50590703 -0.472766194
0.076204151  0.030551349 -0.0005018897
24  1.36785824  0.04800463 -0.59264906 -0.110808292 -
0.038401500  0.007788040 -0.0054000136
25  0.81249472 -0.18561614  1.42052370  0.173437998 -
0.025789735 -0.013669552 -0.0102481177
               X7            X8           X9
1   2.313311e-04  0.0001968401 -0.003791305
2  -2.810250e-03  0.0145484900  0.038461481
3  -2.143843e-03 -0.0023646335 -0.055741182
4   9.808092e-03 -0.0102845240  0.037393540
5  -2.628407e-04  0.0026354652 -0.002541349
6   5.970640e-03 -0.0114034933 -0.036805557
7   2.025926e-04 -0.0050938017 -0.034190738
8   2.412844e-03  0.0068835257  0.014513211
9   1.493949e-03 -0.0100612163  0.030477398
10  1.049548e-03 -0.0035598789 -0.009074007
11  1.297432e-02 -0.1471536748 -0.236019207
12 -1.984782e-02  0.0009776713 -0.015699233
13  4.790840e-04 -0.0040144922  0.014355521
14 -2.375183e-03 -0.0036489780  0.020657746
15 -4.922690e-04 -0.0004120171 -0.007602931
16 -9.719926e-05  0.0082260268  0.030478505
17  2.326190e-03 -0.0113806432  0.064684487
18  3.462968e-04  0.0015877746  0.008903900
19 -4.151817e-04 -0.0008977983  0.004550437
20  2.244123e-03  0.0115712076 -0.080070603
21 -6.733625e-05 -0.0003982006  0.003288684
22  1.772988e-03 -0.0003373666 -0.202942595
23  1.250138e-03  0.0355286490  0.122551294
24 -2.113277e-04  0.0083259275  0.038385835
25 -2.130798e-03 -0.0120402193 -0.030340215
```

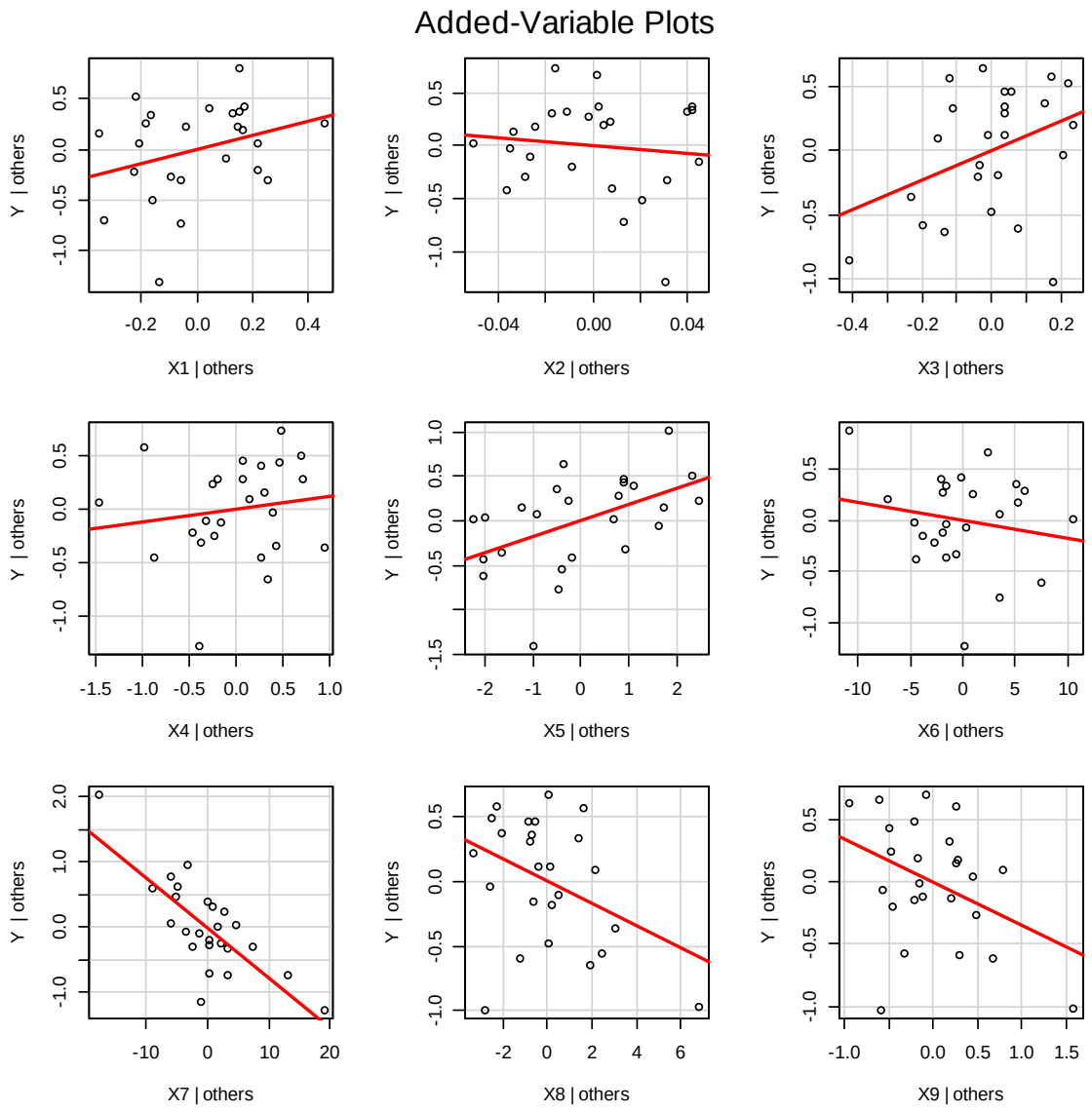
```
$sigma
        1         2         3         4         5         6
7         8         9
0.5885100 0.5429731 0.5818284 0.5823820 0.5893364 0.5807982
0.5605651 0.5834106 0.5790701
       10        11        12        13        14        15
16        17        18
0.5889050 0.5250316 0.5324251 0.5756793 0.5811829 0.5892142
0.5834730 0.5761104 0.5857448
       19        20        21        22        23        24
25
0.5888442 0.5628226 0.5878454 0.5624277 0.4612705 0.5530523
0.5676903

$wt.res
          1           2           3           4           5
6           7           8
 0.12823682  0.70426655  0.30450950  0.19849401 -0.06994273
0.26989035  0.41483765  0.26707756
          9          10          11          12          13
14          15          16
-0.35146807 -0.09951181 -0.38924605  0.67531536  0.39230402
0.24457719  0.07533160 -0.26697093
         17          18          19          20          21
22          23          24
 0.37350835 -0.21378339 -0.06869271  0.44965286 -0.15791662
-0.47144589 -1.23126397 -0.69584655
         25
-0.48191312

> layout(matrix(c(1,2,3,4),2,2))
> plot(fitst)
>
```




```
> library(car)
Loading required package: MASS
Loading required package: nnet
> outlierTest(fitst)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
23 -2.980691          0.0099241       0.2481
> layout(matrix(1), widths=lcm(12),
heights=lcm(12))
> leveragePlots(fitst)
>
```

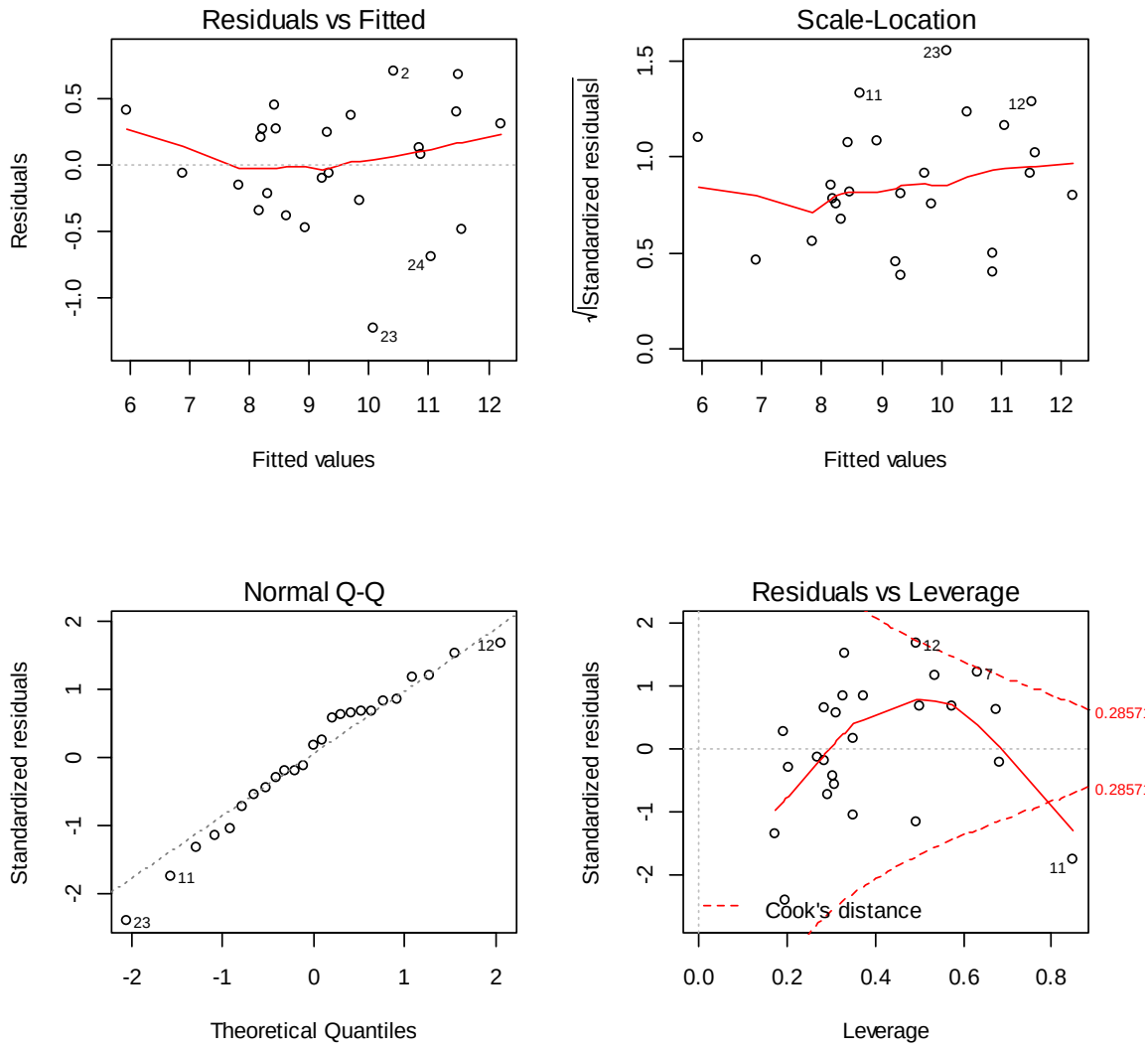



```
> avPlots(fitst)
>
```

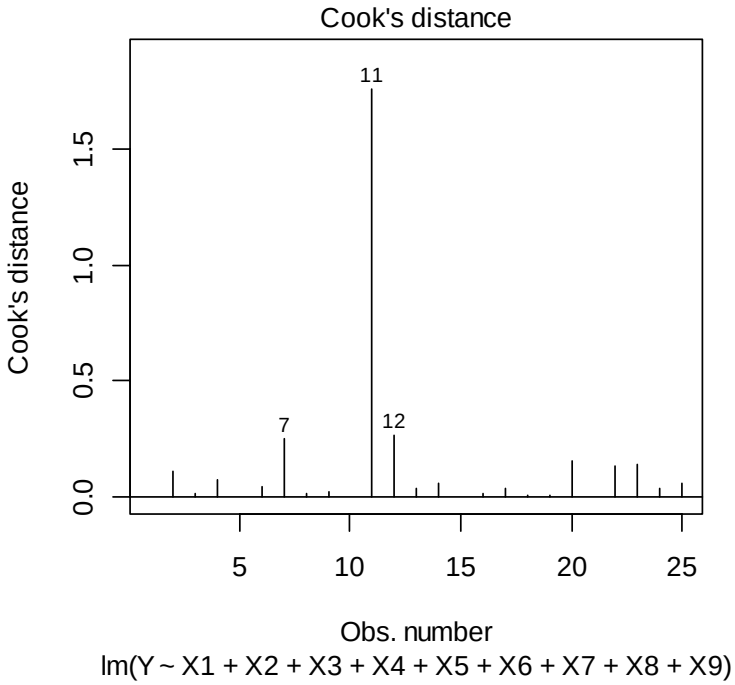



```
> layout(matrix(1), widths=lcm(12),
heights=lcm(12))
> qqPlot(fitst, main = "QQ PLOT")
>
```


```
> cutoff=4/((length(stf$Y)-
length(fitst$coefficients)-1))
> layout(matrix(c(1,2,3,4),2,2))
> plot(fitst, cook.levels = cutoff)
>
```

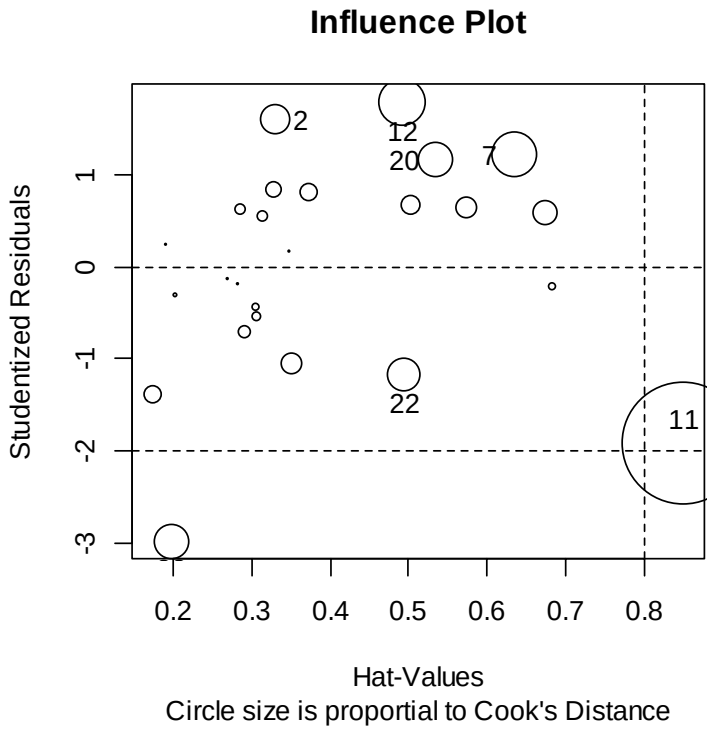



```
> layout(matrix(1), widths=lcm(12),
heights=lcm(12))
> plot(fitst, which = 4, cook.levels = cutoff)
> abline(0,0)
>
```

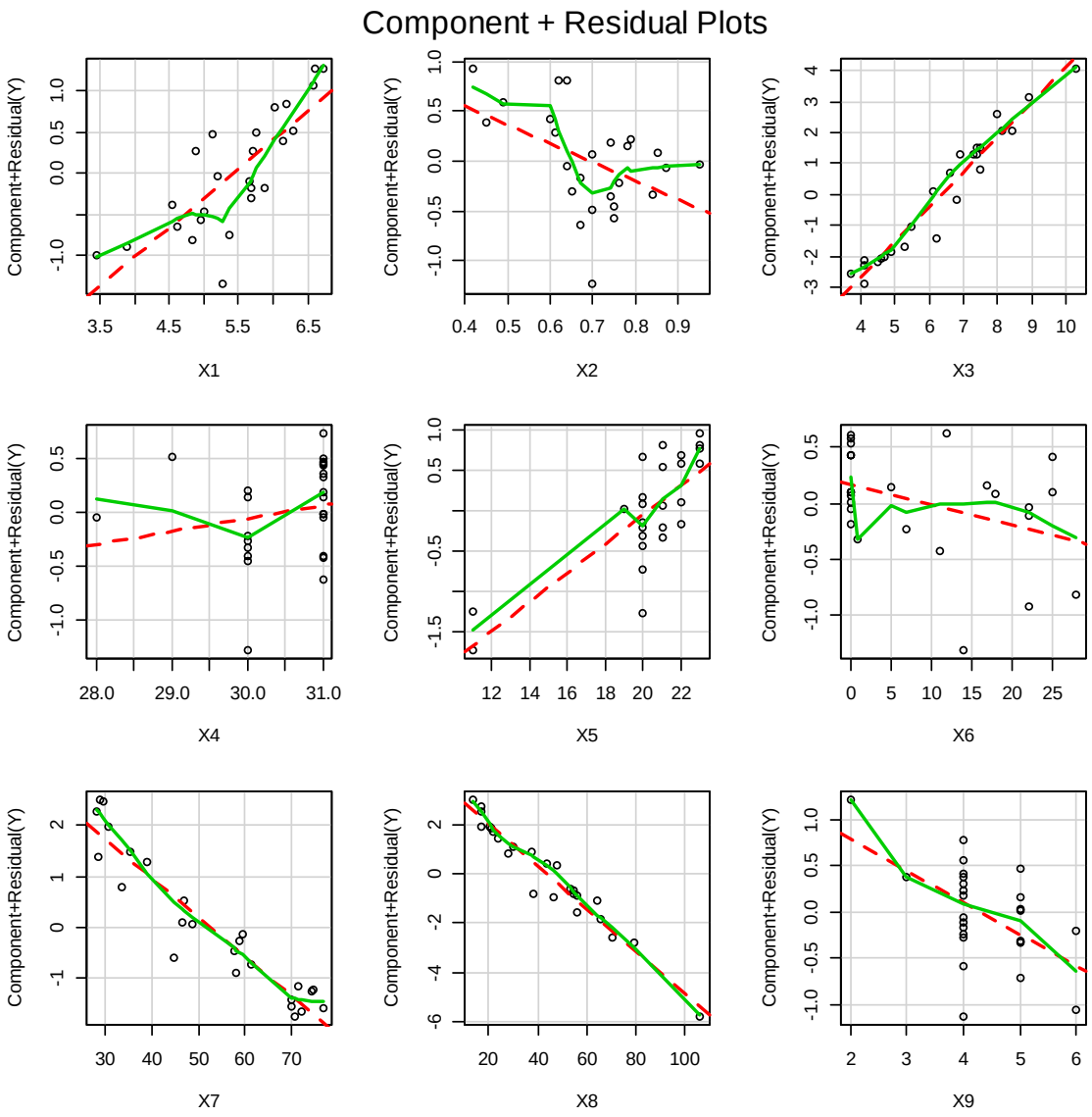



```
> influencePlot(fitst, id.method = "identify", main
= "Influence Plot", sub = "Circle size is
proportial to Cook's Distance")
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
     StudRes       Hat     CookD
2   1.586091 0.3312540 0.3364151
7   1.222841 0.6337615 0.5004908
11 -1.914086 0.8499781 1.3276771
12  1.782290 0.4935469 0.5199369
20  1.171019 0.5345388 0.3920148
22 -1.180202 0.4955507 0.3651551
```

```
23 -2.980691 0.1980320 0.3792109
>
```




```
> vif(fitst)
       X1        X2        X3        X4        X5        X6        X7        X8
15.746595 20.137114 126.625618   1.836626   4.411920   4.695013   6.067426 107.590891
       X9
 2.385046
>
> crPlots(fitst)
>
```




```
> durbinWatsonTest(fitst)
 lag Autocorrelation D-W Statistic p-value
   1       0.4542126      1.040501   0.012
 Alternative hypothesis: rho != 0
>
> library(gvlma)
> gvmodel <- gvlma(fitst)
> summary(gvmodel)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
X9,
    data = stf)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.23126 -0.26697  0.07533  0.30451  0.70427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.89421    6.99636   0.271 0.790279
X1           0.70541    0.56490   1.249 0.230903
X2          -1.89372    4.14629  -0.457 0.654412
X3           1.13422    0.74609   1.520 0.149253
X4           0.11876    0.20461   0.580 0.570247
X5           0.17935    0.08095   2.216 0.042611 *
X6          -0.01818    0.02451  -0.742 0.469699
X7          -0.07742    0.01659  -4.666 0.000304 ***
X8          -0.08585    0.05200  -1.651 0.119529
X9          -0.34501    0.21070  -1.637 0.122337
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.5697 on 15 degrees of freedom
Multiple R-squared: 0.9237,     Adjusted R-squared: 0.8779
F-statistic: 20.18 on 9 and 15 DF,  p-value: 7.97e-07


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = fitst)

                     Value p-value
Decision
Global Stat        10.5230 0.03248 Assumptions NOT
satisfied!
Skewness            2.1206 0.14533    Assumptions
acceptable.
Kurtosis            0.2134 0.64411    Assumptions
acceptable.
Link Function       4.9650 0.02587 Assumptions NOT
satisfied!
Heteroscedasticity  3.2240 0.07256    Assumptions
acceptable.
>
> fit1 <- lm( Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
X9, data = stf)
```

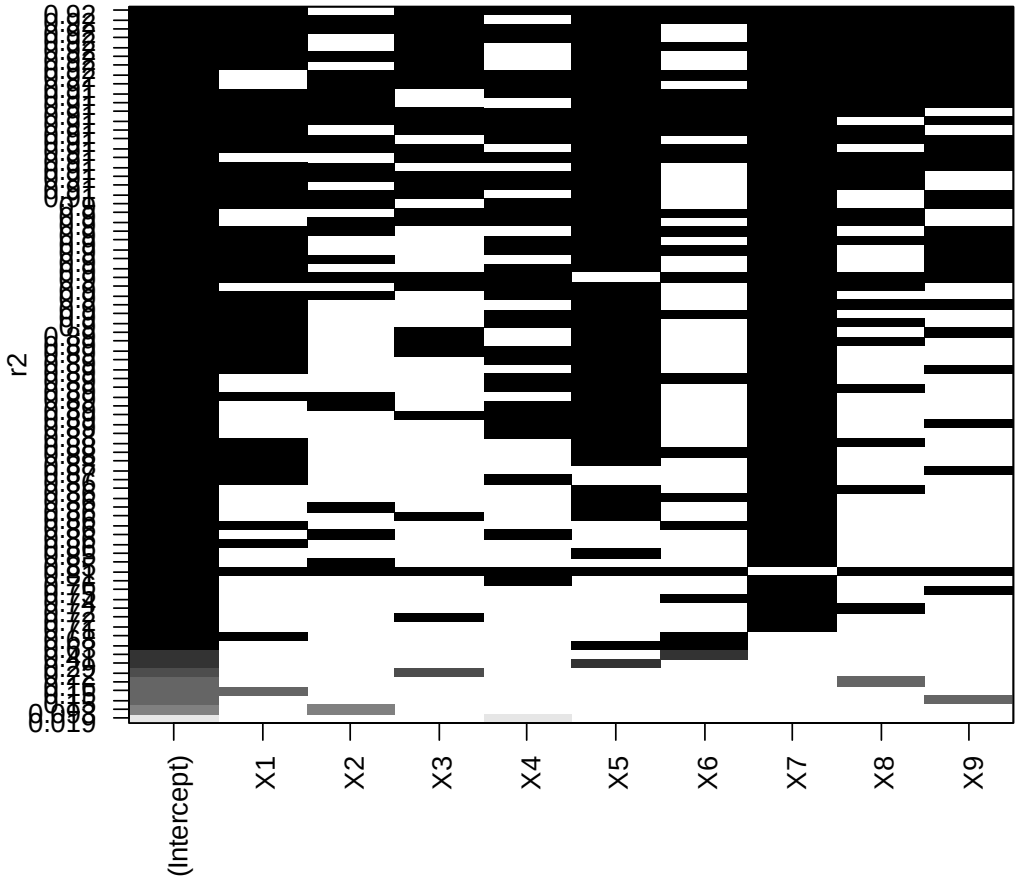
```
> fit2 <- lm( Y ~ X1 + X2 + X3 + X4, data = stf)
> anova(fit1, fit2)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     15  4.869
2     20 34.929 -5   -30.059 18.52 6.103e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
>
> # Stepwise Regression
> library(MASS)
> fitst <- lm( Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
X9, data = stf)
> step <- stepAIC(fitst, direction="both")
Start:  AIC=-20.9
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9

       Df Sum of Sq     RSS      AIC
- X2    1    0.0677  4.9369 -22.5536
- X4    1    0.1094  4.9785 -22.3436
- X6    1    0.1786  5.0477 -21.9984
<none>               4.8692 -20.8989
- X1    1    0.5062  5.3753 -20.4264
- X3    1    0.7502  5.6193 -19.3165
- X9    1    0.8704  5.7395 -18.7875
- X8    1    0.8847  5.7539 -18.7250
- X5    1    1.5934  6.4625 -15.8213
- X7    1    7.0672 11.9364  -0.4822

Step:  AIC=-22.55
Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8 + X9

       Df Sum of Sq     RSS      AIC
- X6    1    0.1626  5.0995 -23.7434
- X4    1    0.1902  5.1271 -23.6086
<none>               4.9369 -22.5536
+ X2    1    0.0677  4.8692 -20.8989
- X9    1    0.8642  5.8011 -20.5207
- X1    1    0.9226  5.8595 -20.2702
- X3    1    1.1727  6.1096 -19.2255
- X8    1    1.3012  6.2381 -18.7051
- X5    1    1.9757  6.9125 -16.1385
- X7    1    7.0000 11.9368  -2.4812
```

```
Step:  AIC=-23.74
Y ~ X1 + X3 + X4 + X5 + X7 + X8 + X9

       Df Sum of Sq     RSS      AIC
- X4    1    0.2263  5.3258 -24.6577
<none>               5.0995 -23.7434
+ X6    1    0.1626  4.9369 -22.5536
+ X2    1    0.0518  5.0477 -21.9984
- X9    1    0.8583  5.9578 -21.8543
- X3    1    1.1508  6.2503 -20.6562
- X1    1    1.2612  6.3607 -20.2185
- X8    1    1.2683  6.3678 -20.1905
- X5    1    1.8377  6.9372 -18.0496
- X7    1   12.2597 17.3592   4.8811

Step:  AIC=-24.66
Y ~ X1 + X3 + X5 + X7 + X8 + X9

       Df Sum of Sq     RSS      AIC
<none>               5.3258 -24.6577
+ X4    1    0.2263  5.0995 -23.7434
+ X6    1    0.1988  5.1271 -23.6086
+ X2    1    0.1329  5.1929 -23.2896
- X3    1    1.2189  6.5447 -21.5054
- X8    1    1.3888  6.7146 -20.8648
- X9    1    1.4168  6.7426 -20.7606
- X5    1    1.6566  6.9824 -19.8870
- X1    1    2.6204  7.9463 -16.6543
- X7    1   12.9509 18.2768   4.1688
> step$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9

Final Model:
Y ~ X1 + X3 + X5 + X7 + X8 + X9


  Step Df   Deviance Resid. Df Resid. Dev       AIC
1                           15   4.869151 -20.89891
2 - X2  1 0.06771323        16   4.936864 -22.55364
3 - X6  1 0.16263227        17   5.099496 -23.74335
4 - X4  1 0.22633618        18   5.325832 -24.65767
> # All Subsets Regression
```

```
> library(leaps)
> leaps<-regsubsets(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 +
X8 + X9, data = stf, nbest=10)
> # view results
> summary(leaps)
Subset selection object
Call: regsubsets.formula(Y ~ X1 + X2 + X3 + X4 + X5 + X6 +
X7 + X8 +
    X9, data = stf, nbest = 10)
9 Variables  (and intercept)
   Forced in Forced out
X1     FALSE      FALSE
X2     FALSE      FALSE
X3     FALSE      FALSE
X4     FALSE      FALSE
X5     FALSE      FALSE
X6     FALSE      FALSE
X7     FALSE      FALSE
X8     FALSE      FALSE
X9     FALSE      FALSE
10 subsets of each size up to 8
Selection Algorithm: exhaustive
          X1  X2  X3  X4  X5  X6  X7  X8  X9
1  ( 1 )  " " " " " " " " " " " " "*" " " " "
1  ( 2 )  " " " " " " " " " " " " "*" " " " "
1  ( 3 )  " " " " " " " " " " "*" " " " " " "
1  ( 4 )  " " " " "*" " " " " " " " " " " " "
1  ( 5 )  " " " " " " " " " " " " " " "*" " "
1  ( 6 )  "*" " " " " " " " " " " " " " " " "
1  ( 7 )  " " " " " " " " " " " " " " " " "*"
1  ( 8 )  " " "*" " " " " " " " " " " " " " "
1  ( 9 )  " " " " " " " "*" " " " " " " " " "
2  ( 1 )  "*" " " " " " " " " " " " "*" " " "
2  ( 2 )  " " " " " " " " " "*" " " "*" " " "
2  ( 3 )  " " "*" " " " " " " " " " "*" " " "
2  ( 4 )  " " " " " " " "*" " " " " "*" " " "
2  ( 5 )  " " " " " " " " " " " " "*" " " "*"
2  ( 6 )  " " " " " " " " " " "*" "*" " " " "
2  ( 7 )  " " " " " " " " " " " " "*" "*" " "
2  ( 8 )  " " " " "*" " " " " " " "*" " " " "
2  ( 9 )  "*" " " " " " " " " " "*" " " " " "
2  ( 10 ) " " " " " " " " " "*" "*" " " " " "
3  ( 1 )  " " " " " " " "*" "*" " " "*" " " "
3  ( 2 )  "*" " " " " " " " " "*" " " "*" " " "
3  ( 3 )  "*" " " " " " " " " " " " "*" " " "*"
3  ( 4 )  "*" " " " " " "*" " " " " "*" " " "
3  ( 5 )  " " " " " " " " " "*" " " "*" "*" " "
```

```
3  ( 6 )  " " " " " " " " " "*" "*" "*" " " " " "
3  ( 7 )  " " "*" " " " " " "*" " " "*" " " " " "
3  ( 8 )  " " " " "*" " " " "*" " " "*" " " " " "
3  ( 9 )  "*" " " " " " " " " " "*" "*" " " " " "
3  ( 10 ) " " "*" " " " "*" " " " " "*" " " " " "
4  ( 1 )  "*" " " " " "*" "*" " " "*" " " " " "
4  ( 2 )  "*" " " " " " " " "*" " " "*" " " " "*"
4  ( 3 )  " " " " " " " "*" "*" "*" "*" " " " " "
4  ( 4 )  " " " " " " " "*" "*" " " "*" "*" " " "
4  ( 5 )  "*" "*" " " " " " "*" " " " "*" " " " " "
4  ( 6 )  " " "*" " " " "*" "*" " " "*" " " " " "
4  ( 7 )  " " " " "*" "*" "*" " " "*" " " " " " "
4  ( 8 )  " " " " " " "*" "*" " " "*" " " " "*"
4  ( 9 )  "*" " " " " " " " "*" " " "*" "*" " " "
4  ( 10 ) "*" " " " " " " " "*" "*" "*" " " " " "
5  ( 1 )  "*" "*" " " " " " "*" " " "*" " " " "*"
5  ( 2 )  "*" " " " " " "*" "*" " " "*" " " " "*"
5  ( 3 )  " " " " "*" "*" "*" " " "*" "*" " " "
5  ( 4 )  "*" "*" " " " "*" "*" " " "*" " " " " "
5  ( 5 )  "*" " " " " " " " "*" " " "*" "*" "*"
5  ( 6 )  "*" " " " " " "*" "*" "*" "*" " " " " "
5  ( 7 )  "*" " " " " " "*" "*" " " "*" "*" " " "
5  ( 8 )  "*" " " "*" " " " "*" " " "*" " " " "*"
5  ( 9 )  "*" " " "*" " " " "*" " " "*" "*" " " "
5  ( 10 ) "*" " " "*" "*" "*" " " "*" " " " " "
6  ( 1 )  "*" " " "*" " " "*" " " "*" "*" "*" "
6  ( 2 )  "*" "*" " " " " " "*" " " "*" "*" "*" "
6  ( 3 )  "*" " " "*" "*" "*" " " "*" "*" " " "
6  ( 4 )  "*" "*" "*" " " "*" " " "*" " " " "*"
6  ( 5 )  "*" "*" " " "*" "*" " " "*" " " " "*"
6  ( 6 )  " " " " "*" "*" "*" "*" "*" "*" " " "
6  ( 7 )  " " "*" "*" "*" "*" " " "*" "*" " " "
6  ( 8 )  "*" "*" " " " " " "*" "*" "*" " " "*"
6  ( 9 )  "*" " " " " " "*" "*" " " "*" "*" "*"
6  ( 10 ) "*" " " " " " "*" "*" "*" "*" " " "*"
7  ( 1 )  "*" " " "*" "*" "*" " " "*" "*" "*" "
7  ( 2 )  "*" " " "*" " " "*" "*" "*" "*" "*" "
7  ( 3 )  "*" "*" "*" " " "*" " " "*" "*" "*" "
7  ( 4 )  " " "*" "*" "*" "*" " " "*" "*" "*" "
7  ( 5 )  "*" "*" " " " " " "*" "*" "*" "*" "*" "
7  ( 6 )  "*" " " "*" "*" "*" "*" "*" "*" " " "
7  ( 7 )  "*" "*" " " "*" "*" " " "*" "*" "*" "
7  ( 8 )  "*" "*" "*" " " "*" "*" "*" " " "*" "
7  ( 9 )  " " " " "*" "*" "*" "*" "*" "*" "*" "
7  ( 10 ) "*" "*" "*" "*" "*" " " "*" "*" " " "
8  ( 1 )  "*" " " "*" "*" "*" "*" "*" "*" "*" "
8  ( 2 )  "*" "*" "*" " " "*" "*" "*" "*" "*" "
```

```
8  ( 3 )  "*" "*" "*" "*" "*" " " "*" "*" "*"
8  ( 4 )  " " "*" "*" "*" "*" "*" "*" "*" "*"
8  ( 5 )  "*" "*" " " "*" "*" "*" "*" "*" "*"
8  ( 6 )  "*" "*" "*" "*" "*" "*" "*" "*" " "
8  ( 7 )  "*" "*" "*" "*" "*" "*" "*" " " "*"
8  ( 8 )  "*" "*" "*" "*" " " "*" "*" "*" "*"
8  ( 9 )  "*" "*" "*" "*" "*" "*" " " "*" "*"
>
> # plot a table of models showing variables in each model.
> # models are ordered by the selection statistic.
> plot(leaps,scale="r2")
>
```



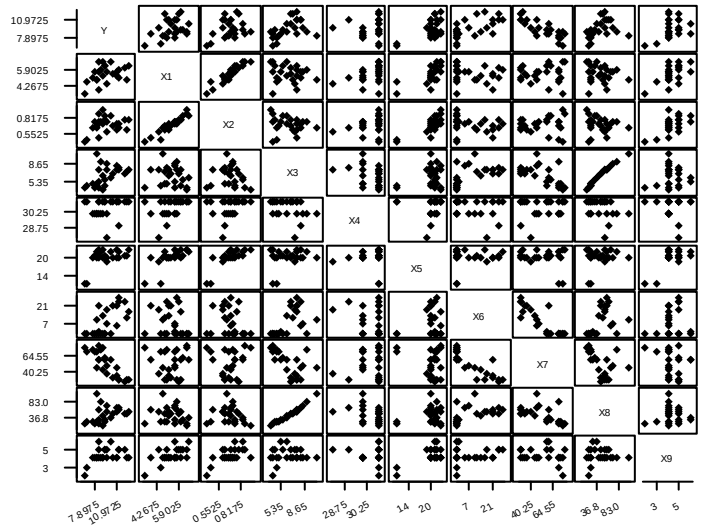
الحل بواسطة Minitab:

نرسم البيانات في رسم مصفوفي لمعرفة العلاقات بين المتغيرات ثنائيا:

```
MTB > MatrixPlot 'Y' 'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'X9';
SUBC>   Symbol;
SUBC>   ScFrame.
```

MatrixPlot 'Y' 'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'X9';

```
MTB >
```



يلاحظ وجود علاقات خطية بين $X1$ و $X2$ وكذلك بين $X3$ و $X8$ وبين $X6$ و $X7$ وقد تسبب هذة تعددية خطية مشتركة *Multicollinearity*.

```
MTB > Regress 'Y' 9 'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'X9';
SUBC>  GHistogram;
SUBC>  GNormalplot;
SUBC>  GFits;
```

```
SUBC>  RType 1;
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: Y versus X1, X2, X3, X4, X5, X6, X7, X8, X9

```
The regression equation is
Y = 1.89 + 0.705 X1 - 1.89 X2 + 1.13 X3 + 0.119 X4 + 0.179 X5 - 0.0182 X6
           - 0.0774 X7 - 0.0858 X8 - 0.345 X9

Predictor        Coef     SE Coef          T        P       VIF
Constant        1.894       6.996       0.27    0.790
X1             0.7054      0.5649       1.25    0.231      15.7
X2             -1.894       4.146      -0.46    0.654      20.1
X3             1.1342      0.7461       1.52    0.149     126.6
X4             0.1188      0.2046       0.58    0.570       1.8
X5            0.17935     0.08095       2.22    0.043       4.4
X6           -0.01818     0.02451      -0.74    0.470       4.7
X7           -0.07742     0.01659      -4.67    0.000       6.1
X8           -0.08585     0.05200      -1.65    0.120     107.6
X9            -0.3450      0.2107      -1.64    0.122       2.4

S = 0.5697             R-Sq = 92.4%         R-Sq(adj) = 87.8%
PRESS = 18.9953        R-Sq(pred) = 70.23%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         9     58.9466      6.5496     20.18    0.000
Residual Error    15      4.8692      0.3246
Total             24     63.8158

No replicates. Cannot do pure error test.

Source       DF      Seq SS
```

X1	1	9.3698
X2	1	1.8295
X3	1	16.9620
X4	1	0.7259
X5	1	5.5695
X6	1	13.2026
X7	1	9.8937
X8	1	0.5233
X9	1	0.8704

Obs	X1	Y	Fit	SE Fit	Residual	St Resid
1	5.20	10.980	10.852	0.249	0.128	0.25
2	5.12	11.130	10.426	0.328	0.704	1.51
3	6.19	12.510	12.205	0.305	0.305	0.63
4	3.89	8.400	8.202	0.468	0.198	0.61
5	6.28	9.270	9.340	0.296	-0.070	-0.14
6	5.76	8.730	8.460	0.404	0.270	0.67
7	3.45	6.360	5.945	0.454	0.415	1.20
8	6.57	8.500	8.233	0.319	0.267	0.57
9	5.69	7.820	8.171	0.308	-0.351	-0.73
10	6.14	9.140	9.240	0.303	-0.100	-0.21
11	4.84	8.240	8.629	0.525	-0.389	-1.76
12	4.88	12.190	11.515	0.400	0.675	1.67
13	6.03	11.880	11.488	0.327	0.392	0.84
14	4.55	9.570	9.325	0.432	0.245	0.66
15	5.71	10.940	10.865	0.337	0.075	0.16
16	5.67	9.580	9.847	0.316	-0.267	-0.56
17	6.72	10.090	9.716	0.348	0.374	0.83
18	4.95	8.110	8.324	0.315	-0.214	-0.45
19	4.62	6.830	6.899	0.470	-0.069	-0.21
20	6.60	8.880	8.430	0.417	0.450	1.16
21	5.01	7.680	7.838	0.257	-0.158	-0.31
22	5.68	8.470	8.941	0.401	-0.471	-1.17
23	5.28	8.860	10.091	0.254	-1.231	-2.41R
24	5.36	10.360	11.056	0.238	-0.696	-1.34
25	5.87	11.080	11.562	0.337	-0.482	-1.05

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.04

No evidence of lack of fit (P > 0.1)

```
MTB > Stepwise 'Y' 'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'X9';
SUBC>   AEnter 0.15;
SUBC>   ARemove 0.15;
SUBC>   Constant;
SUBC>   Press.
```

Stepwise Regression: Y versus X1, X2, X3, X4, X5, X6, X7, X8, X9

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is    Y     on  9 predictors, with N =   25


    Step          1        2        3        4        5
Constant   13.62299  9.47422  8.56626  0.09878 -2.96806

X7          -0.0798  -0.0798  -0.0758  -0.0756  -0.0739
T-Value       -7.59   -10.59   -10.16   -10.50   -10.30
P-Value       0.000    0.000    0.000    0.000    0.000

X1                     0.76     0.49     0.30
T-Value                4.78     2.30     1.26
P-Value               0.000    0.031    0.221

X5                            0.108    0.142    0.199
T-Value                        1.85     2.36     4.86
P-Value                       0.079    0.029    0.000

X4                                      0.29     0.40
T-Value                                 1.61     2.56
P-Value                                0.123    0.018

S             0.890    0.637    0.605    0.583    0.591
R-Sq          71.44    86.00    87.96    89.34    88.49
R-Sq(adj)     70.20    84.73    86.24    87.21    86.85
C-p            35.1      8.5      6.7      6.0      5.6
PRESS       21.4938  11.0951  10.3443  9.69846  9.93612
R-Sq(pred)    66.32    82.61    83.79    84.80    84.43
```

```
 More? (Yes, No, Subcommand, or Help)
SUBC> Yes

 No variables entered or removed

 More? (Yes, No, Subcommand, or Help)
SUBC> no


MTB > BReg 'Y' 'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7' 'X8' 'X9' ;
SUBC>   NVars 1 9;
SUBC>   Best 5;
SUBC>   Constant.
```

Best Subsets Regression: Y versus X1, X2, X3, X4, X5, X6, X7, X8, X9

Response is Y

Vars	R-Sq	R-Sq(adj)	C-p	S	X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9
1	71.4	70.2	35.1	0.89012							X		
1	41.0	38.5	94.9	1.2790						X			
1	28.7	25.6	119.1	1.4061					X				
1	22.5	19.1	131.4	1.4664				X					
1	15.6	11.9	145.0	1.5306								X	
2	86.0	84.7	8.5	0.63716	X						X		
2	84.9	83.5	10.7	0.66157					X		X		
2	84.7	83.3	11.1	0.66691		X					X		
2	75.6	73.3	29.1	0.84207				X			X		
2	74.9	72.7	30.3	0.85248							X		X
3	88.5	86.8	5.6	0.59136				X	X		X		
3	88.0	86.2	6.7	0.60493	X				X		X		
3	86.5	84.6	9.5	0.64038	X						X		X
3	86.4	84.4	9.8	0.64339	X			X			X		
3	86.4	84.4	9.8	0.64397					X		X	X	
4	89.3	87.2	6.0	0.58318	X			X	X		X		
4	89.1	87.0	6.4	0.58877	X				X		X		X
4	89.0	86.8	6.6	0.59265				X	X	X	X		
4	88.8	86.6	6.9	0.59665				X	X		X	X	
4	88.8	86.5	7.0	0.59817	X	X			X		X		

```
   5   90.0         87.3         6.7   0.58088   X X     X   X   X
   5   89.9         87.2         6.9   0.58304   X     X X   X   X
   5   89.8         87.1         7.1   0.58665       X X X   X X
   5   89.7         87.0         7.2   0.58686   X X   X X   X
   5   89.7         87.0         7.2   0.58691   X       X   X X X
   6   91.7         88.9         5.4   0.54395   X   X   X   X X X
   6   90.8         87.7         7.2   0.57264   X X     X   X X X
   6   90.7         87.6         7.4   0.57532   X   X X X   X X
   6   90.5         87.4         7.6   0.57965   X X X   X   X   X
   6   90.4         87.2         7.9   0.58425   X X   X X   X   X
   7   92.0         88.7         6.7   0.54770   X   X X X   X X X
   7   92.0         88.7         6.8   0.54917   X   X   X X X X X
   7   91.9         88.5         7.0   0.55269   X X X   X   X X X
   7   91.2         87.6         8.3   0.57416     X X X X   X X X
   7   91.1         87.4         8.5   0.57839   X X     X X X X X
   8   92.3         88.4         8.2   0.55548   X   X X X X X X X
   8   92.2         88.3         8.3   0.55781   X X X   X X X X X
   8   92.1         88.1         8.6   0.56168   X X X X X   X X X
   8   91.6         87.4         9.6   0.57962     X X X X X X X X
   8   91.2         86.8        10.3   0.59263   X X   X X X X X X
   9   92.4         87.8        10.0   0.56975   X X X X X X X X X


MTB >
```

مثال (5)

بيانات النادي الصحي:

No.	Y	X1	X2	X3	X4
1	481	217	67	270	91
2	292	141	52	190	66
3	338	152	58	203	68
4	357	153	56	183	70
5	396	180	66	170	77
6	429	193	71	178	82
7	345	162	65	160	74
8	469	180	80	170	84
9	425	205	77	188	83
10	358	168	74	170	79
11	393	232	65	220	72
12	346	146	68	158	68
13	279	173	51	243	56
14	311	155	64	198	59
15	401	212	66	220	77
16	267	138	70	180	62
17	404	147	54	150	75
18	442	197	76	228	88
19	368	165	59	188	70
20	295	125	58	160	66
21	391	161	52	190	69
22	264	132	62	163	59
23	487	257	64	313	96
24	481	236	72	225	84
25	374	149	57	173	68
26	309	161	57	173	65
27	367	198	59	220	62
28	469	245	70	218	69
29	252	141	63	193	60
30	338	177	53	183	75

بإستخدام Excel:

SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.922790566
R Square	0.85154243
Adjusted R Square	0.827789218
Standard Error	28.82427852
Observations	30

ANOVA

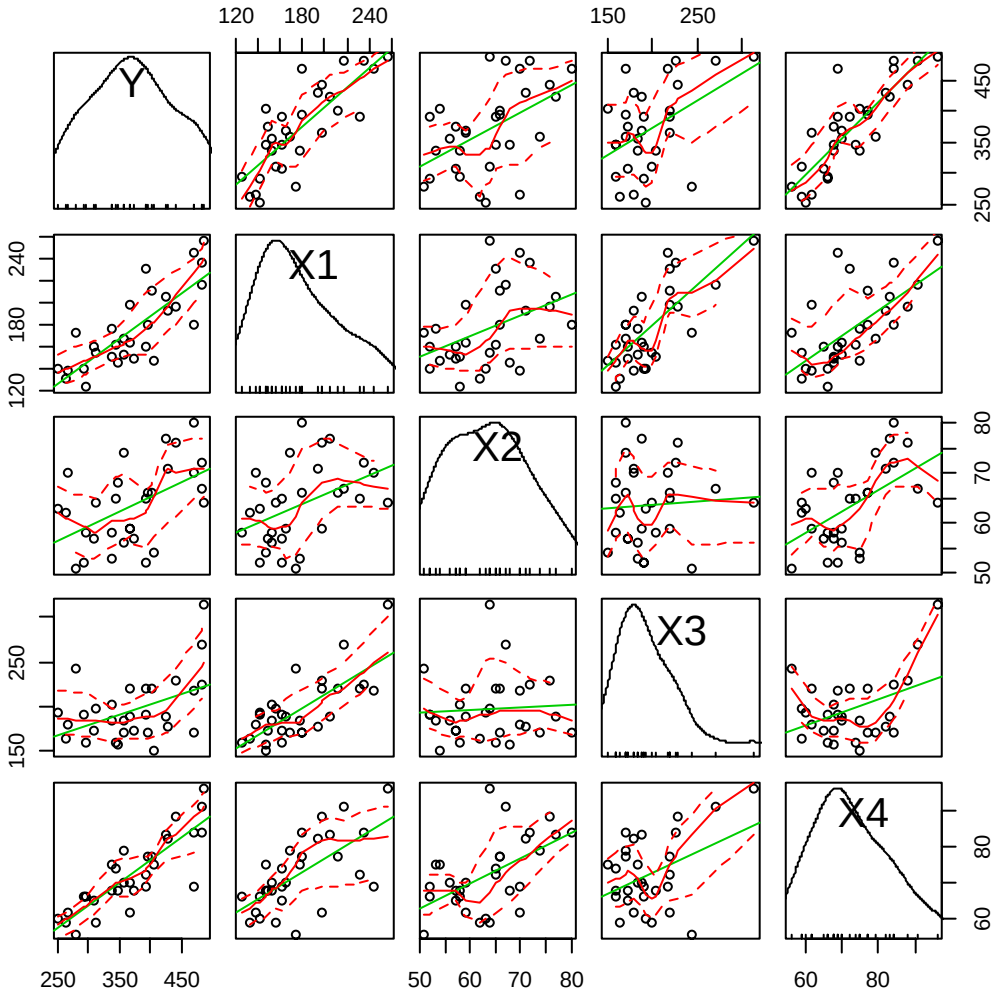
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	119140.8909	29785.22271	35.84957 08	5.14212E-10
Residual	25	20770.97581	830.8390324		
Total	29	139911.8667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-8.863	55.520	-0.160	0.874	-123.208	105.483	-123.208	105.483
X1	1.243	0.284	4.370	0.000	0.657	1.829	0.657	1.829
X2	-0.502	0.867	-0.579	0.568	-2.288	1.284	-2.288	1.284
X3	-0.476	0.241	-1.977	0.059	-0.971	0.020	-0.971	0.020
X4	3.938	0.752	5.235	0.000	2.389	5.487	2.389	5.487

بإستخدام R:

```
> rm(list=ls())
> # sports club data
> sclub = read.table(file.choose(), header = T)
> sclub
   No.   Y  X1 X2  X3 X4
1    1 481 217 67 270 91
2    2 292 141 52 190 66
3    3 338 152 58 203 68
4    4 357 153 56 183 70
5    5 396 180 66 170 77
6    6 429 193 71 178 82
7    7 345 162 65 160 74
8    8 469 180 80 170 84
9    9 425 205 77 188 83
10  10 358 168 74 170 79
11  11 393 232 65 220 72
12  12 346 146 68 158 68
13  13 279 173 51 243 56
14  14 311 155 64 198 59
15  15 401 212 66 220 77
16  16 267 138 70 180 62
17  17 404 147 54 150 75
18  18 442 197 76 228 88
19  19 368 165 59 188 70
20  20 295 125 58 160 66
21  21 391 161 52 190 69
22  22 264 132 62 163 59
23  23 487 257 64 313 96
24  24 481 236 72 225 84
25  25 374 149 57 173 68
26  26 309 161 57 173 65
27  27 367 198 59 220 62
28  28 469 245 70 218 69
29  29 252 141 63 193 60
30  30 338 177 53 183 75
> fit = lm( Y ~ X1 + X2 + X3 + X4, data = sclub)
> summary(fit)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = sclub)

Residuals:
    Min      1Q  Median      3Q     Max
-54.843 -19.570  -5.422  18.847  44.499
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.8626    55.5199  -0.160 0.874456
X1            1.2429     0.2844   4.370 0.000191 ***
X2           -0.5017     0.8671  -0.579 0.568001
X3           -0.4756     0.2406  -1.977 0.059174 .
X4            3.9379     0.7522   5.235 2.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 28.82 on 25 degrees of freedom
Multiple R-squared: 0.8515,     Adjusted R-squared: 0.8278
F-statistic: 35.85 on 4 and 25 DF,  p-value: 5.142e-10

> plot(cor(sclub))
> library(car)
Loading required package: MASS
Loading required package: nnet
> scatterplotMatrix(~ Y + X1 + X2 + X3 + X4, data= sclub)
>
```


Y
X1
X2
X3
X4
120
180
240
150
250
450
350
250
240
180
120
80
70
60
50
250
150
80
60
250
350
450
50
60
70
80
60
80

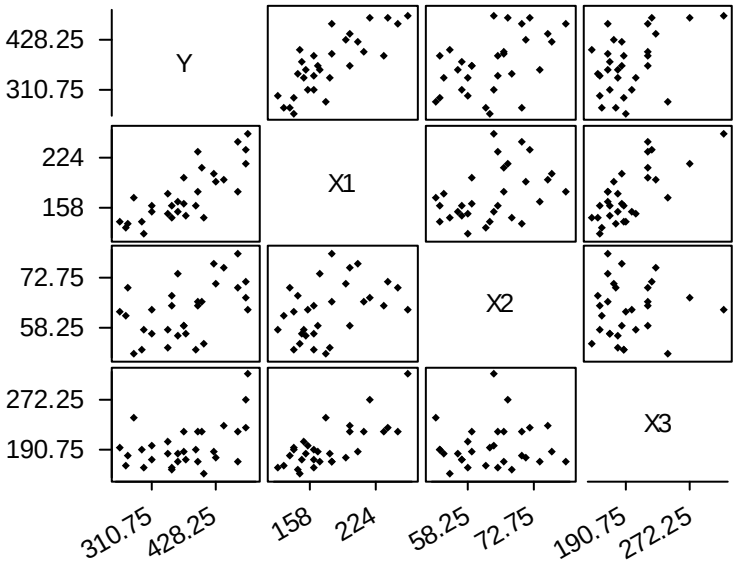
بإستخدام Minitab:

نفحص العلاقات الثنائية بين المتغيرات بالرسم المصفوفي

```
MTB > MatrixPlot 'Y' 'X1' 'X2' 'X3';
SUBC>   Symbol;
SUBC>   ScFrame.
```

MatrixPlot 'Y' 'X1' 'X2' 'X3';

```
MTB >
```



يلاحظ خلو البيانات من التعددية الخطية المشتركة.

```
MTB > Regress 'Y' 3 'X1' 'X2' 'X3';
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
```

```
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: Y versus X1, X2, X3

```
The regression equation is
Y = 66.5 + 1.76 X1 + 1.19 X2 - 0.423 X3

Predictor        Coef     SE Coef          T        P       VIF
Constant        66.52       76.12       0.87    0.390
X1             1.7648      0.3781       4.67    0.000       3.1
X2              1.191       1.142       1.04    0.307       1.4
X3            -0.4232      0.3413      -1.24    0.226       2.6

S = 40.92              R-Sq = 68.9%         R-Sq(adj) = 65.3%
PRESS = 61648.5        R-Sq(pred) = 55.94%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3       96372       32124     19.18    0.000
Residual Error    26       43540        1675
Total             29      139912

No replicates. Cannot do pure error test.

Source       DF      Seq SS
X1            1       89117
X2            1        4680
X3            1        2575

Obs         X1          Y         Fit      SE Fit    Residual    St Resid
  1        217     481.00      415.02       18.81       65.98        1.82
  2        141     292.00      296.88       14.07       -4.88       -0.13
  3        152     338.00      317.94       12.10       20.06        0.51
  4        153     357.00      325.79       10.63       31.21        0.79
  5        180     396.00      390.85       11.74        5.15        0.13
  6        193     429.00      416.36       12.81       12.64        0.33
  7        162     345.00      362.12       11.30      -17.12       -0.44
  8        180     469.00      407.52       18.67       61.48        1.69
```

```
 9        205     425.00      440.45       15.67      -15.45       -0.41
10        168     358.00      379.19       14.20      -21.19       -0.55
11        232     393.00      460.27       16.39      -67.27       -1.79
12        146     346.00      338.30       12.73        7.70        0.20
13        173     279.00      329.74       18.73      -50.74       -1.39
14        155     311.00      332.49       11.72      -21.49       -0.55
15        212     401.00      426.16       10.75      -25.16       -0.64
16        138     267.00      317.25       17.77      -50.25       -1.36
17        147     404.00      326.78       16.51       77.22        2.06R
18        197     442.00      408.21       17.57       33.79        0.91
19        165     368.00      348.42        8.76       19.58        0.49
20        125     295.00      288.49       13.38        6.51        0.17
21        161     391.00      332.18       13.61       58.82        1.52
22        132     264.00      304.33       12.52      -40.33       -1.04
23        257     487.00      463.84       26.06       23.16        0.73 X
24        236     481.00      473.55       15.53        7.45        0.20
25        149     374.00      324.15       10.71       49.85        1.26
26        161     309.00      345.33       11.55      -36.33       -0.93
27        198     367.00      393.12       11.14      -26.12       -0.66
28        245     469.00      490.01       19.12      -21.01       -0.58
29        141     252.00      308.71       14.45      -56.71       -1.48
30        177     338.00      364.57       16.29      -26.57       -0.71

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

Durbin-Watson statistic = 1.82

Lack of fit test
Possible interactions with variable X1 (P-Value = 0.001)
Possible lack of fit at outer X-values       (P-Value = 0.042)
Overall lack of fit test is significant at P = 0.001

MTB > Stepwise 'Y' 'X1' 'X2' 'X3';
SUBC>   AEnter 0.15;
SUBC>   ARemove 0.15;
SUBC>   Constant;
SUBC>   Press.
```

Stepwise Regression: Y versus X1, X2, X3

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

 Response is    Y     on  3 predictors, with N =   30


    Step          1        2
Constant      94.61   131.50

X1             1.56     1.98
T-Value        7.01     6.27
P-Value       0.000    0.000

X3                     -0.56
T-Value                -1.80
P-Value                0.083

S              42.6     41.0
R-Sq          63.69    67.58
R-Sq(adj)     62.40    65.18
C-p             4.3      3.1
PRESS       57119.7  57808.9
R-Sq(pred)    59.17    58.68
 More? (Yes, No, Subcommand, or Help)
SUBC> yes

 No variables entered or removed

 More? (Yes, No, Subcommand, or Help)
SUBC> no


MTB > BReg 'Y' 'X1' 'X2' 'X3' ;
SUBC>   NVars 1 3;
SUBC>   Best 3;
SUBC>   Constant.
```

Best Subsets Regression: Y versus X1, X2, X3

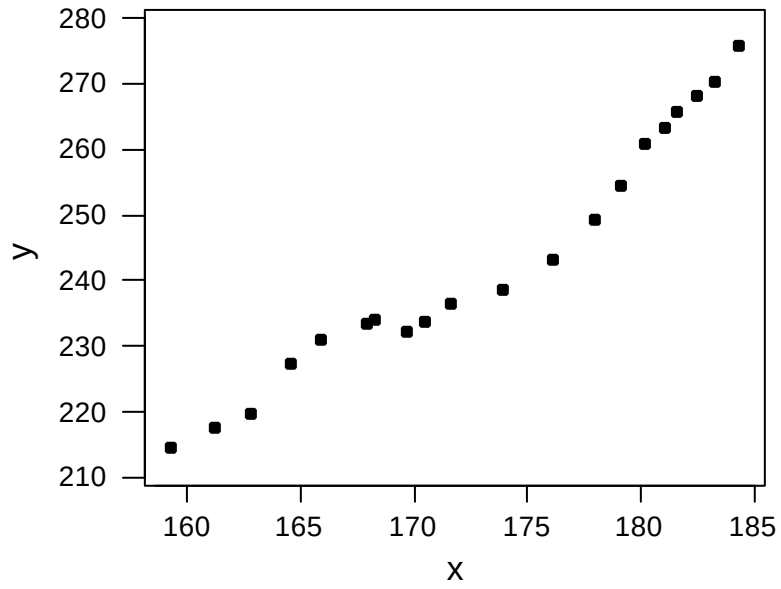
```
Response is Y

                                                  X X X
```

Vars	R-Sq	R-Sq(adj)	C-p	S	1	2	3
1	63.7	62.4	4.3	42.592	X		
1	25.1	22.4	36.6	61.172		X	
1	20.4	17.5	40.5	63.070			X
2	67.6	65.2	3.1	40.988	X		X
2	67.0	64.6	3.5	41.328	X	X	
2	42.8	38.6	23.8	54.439		X	X
3	68.9	65.3	4.0	40.922	X	X	X

مثال (6)

Regression with Serially Correlated Errors: Expenditure Data

T	Y	X
1	214.6	159.3
2	217.7	161.2
3	219.6	162.8
4	227.2	164.6
5	230.9	165.9
6	233.3	167.9
7	234.1	168.3
8	232.3	169.7
9	233.7	170.5
10	236.5	171.6
11	238.7	173.9
12	243.2	176.1
13	249.4	178.0
14	254.3	179.1
15	260.9	180.2
16	263.3	181.1
17	265.6	181.6
18	268.2	182.5
19	270.4	183.3
20	275.6	184.3



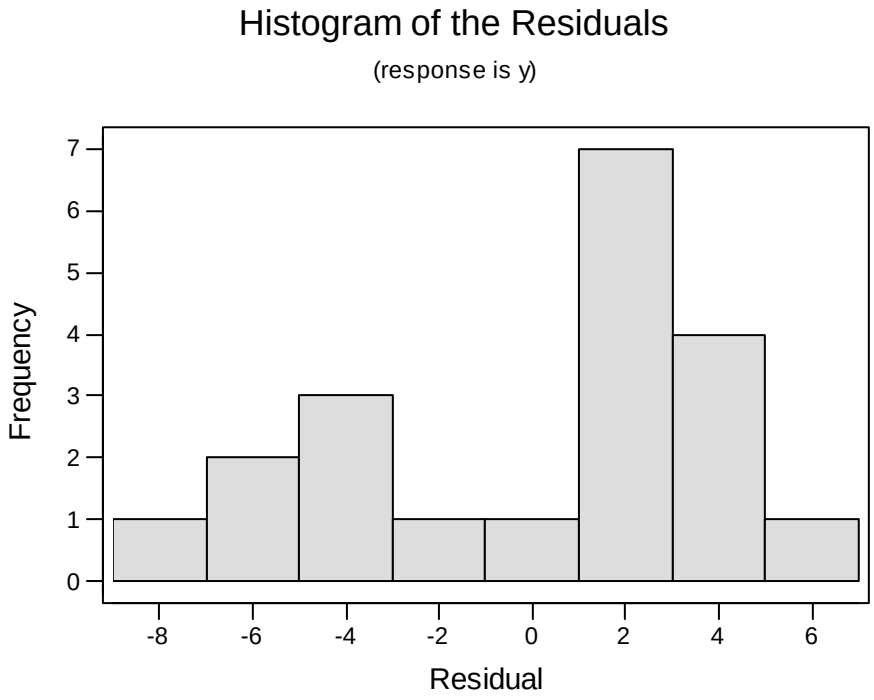

```
MTB > Regress 'y' 1 'x';
SUBC>  GHistogram;
SUBC>  GNormalplot;
SUBC>  GFits;
SUBC>  GOrder;
SUBC>  RType 1;
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: y versus x

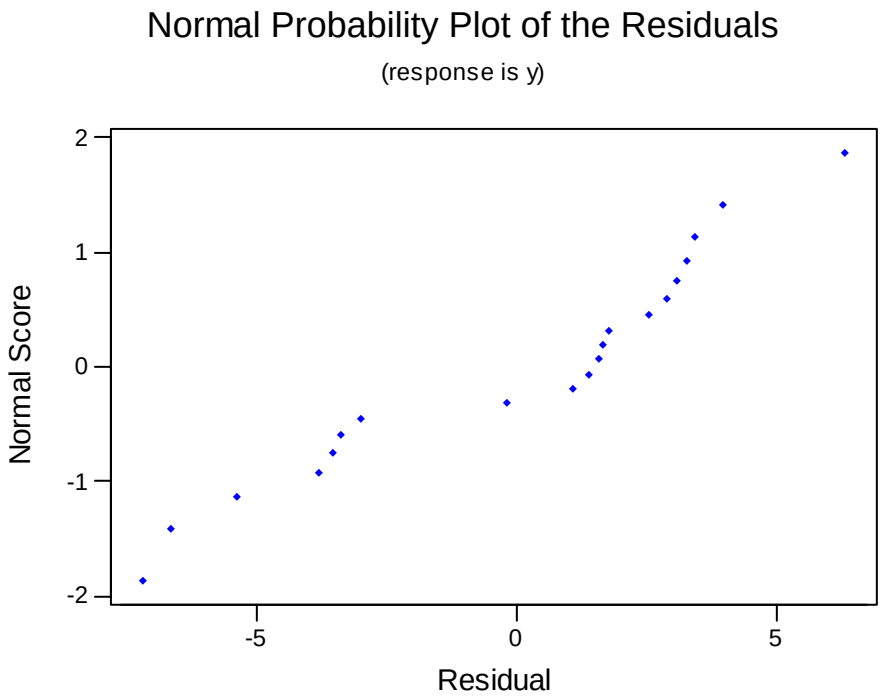
```
The regression equation is
y = - 155 + 2.30 x

Predictor        Coef     SE Coef          T        P
Constant      -154.95       19.88      -7.79    0.000
```

```
x              2.3018      0.1148      20.06    0.000

S = 3.987              R-Sq = 95.7%         R-Sq(adj) = 95.5%
PRESS = 346.619        R-Sq(pred) = 94.81%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      6395.2      6395.2    402.34    0.000
Residual Error    18       286.1        15.9
Total             19      6681.3

No replicates. Cannot do pure error test.

Obs          x          y         Fit      SE Fit    Residual    St Resid
  1        159    214.600     211.722       1.817       2.878        0.81
  2        161    217.700     216.095       1.630       1.605        0.44
  3        163    219.600     219.778       1.480      -0.178       -0.05
  4        165    227.200     223.921       1.321       3.279        0.87
  5        166    230.900     226.914       1.215       3.986        1.05
  6        168    233.300     231.517       1.072       1.783        0.46
  7        168    234.100     232.438       1.048       1.662        0.43
  8        170    232.300     235.660       0.973      -3.360       -0.87
  9        171    233.700     237.502       0.940      -3.802       -0.98
 10        172    236.500     240.034       0.908      -3.534       -0.91
 11        174    238.700     245.328       0.896      -6.628       -1.71
 12        176    243.200     250.392       0.956      -7.192       -1.86
 13        178    249.400     254.765       1.054      -5.365       -1.40
 14        179    254.300     257.297       1.127      -2.997       -0.78
 15        180    260.900     259.829       1.208       1.071        0.28
 16        181    263.300     261.901       1.280       1.399        0.37
 17        182    265.600     263.052       1.322       2.548        0.68
 18        183    268.200     265.123       1.400       3.077        0.82
 19        183    270.400     266.965       1.472       3.435        0.93
 20        184    275.600     269.267       1.565       6.333        1.73


Durbin-Watson statistic = 0.33

Lack of fit test
Possible curvature in variable x (P-Value = 0.000)
Overall lack of fit test is significant at P = 0.000
```

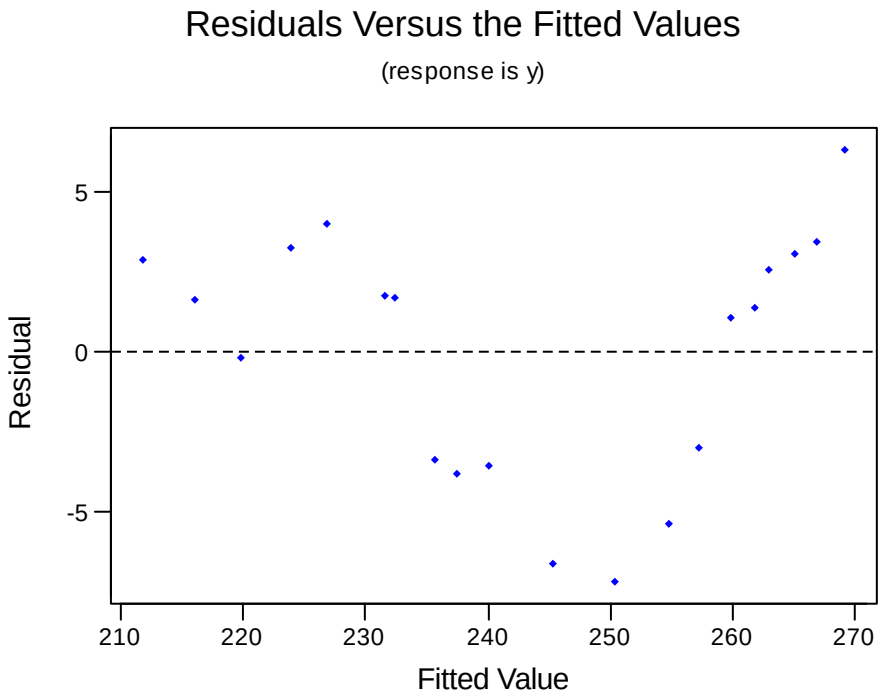
Residual Histogram for y



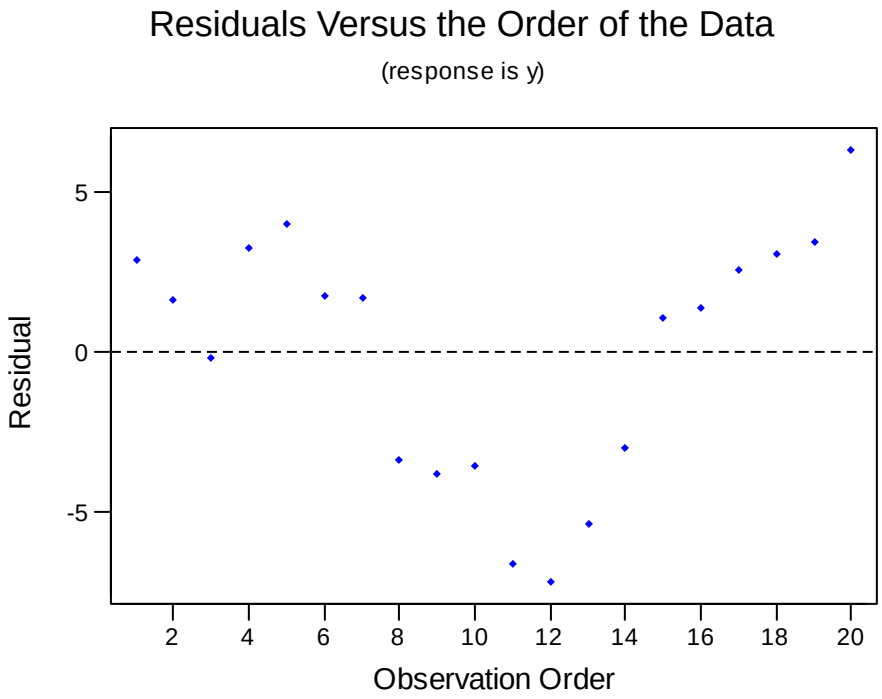
Normplot of Residuals for y

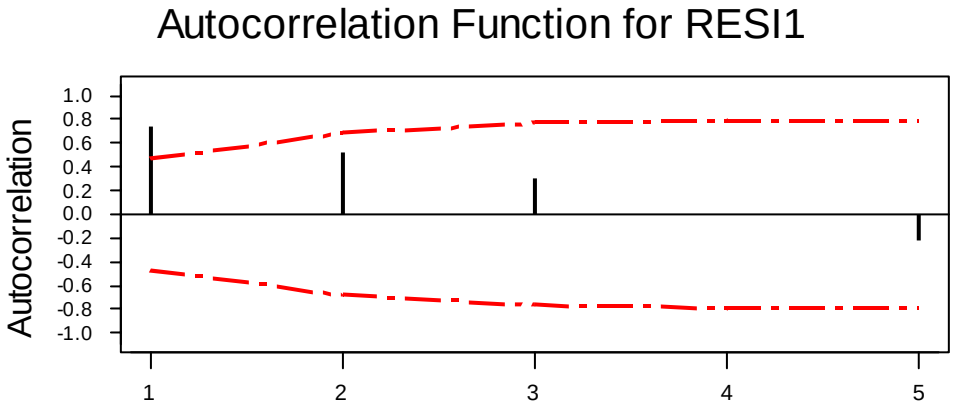


Residuals vs Fits for y



Residuals vs Order for y





Lag	Corr	T	LBQ
1	0.75	3.37	13.11
2	0.52	1.60	19.76
3	0.30	0.81	22.03
4	-0.01	-0.02	22.03
5	-0.22	-0.59	23.49

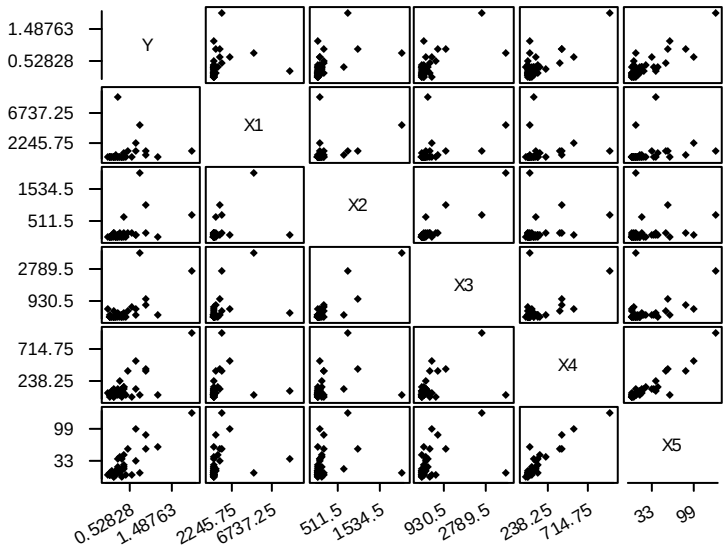
مثال (7)

Influence Diagnostics: Solid Waste Data

Obs	X1	X2	X3	X4	X5	Y
1	102	69	133	125	36	0.3574
2	1120	723	2616	953	132	1.9673
3	139	138	46	35	6	0.1862
4	221	637	153	115	16	0.3816
5	12	0	1	9	1	0.1512
6	1	50	3	25	2	0.1449
7	1046	127	313	392	56	0.4711
8	2032	44	409	540	98	0.6512
9	895	54	168	117	32	0.6624
10	0	0	2	0	1	0.3457
11	25	2	24	78	15	0.3355
12	97	12	91	135	24	0.3982
13	1	0	15	46	11	0.2044
14	4	1	18	23	8	0.2969
15	42	4	78	41	61	1.1515
16	87	162	599	11	3	0.5609
17	2	0	26	24	6	0.1104
18	2	9	29	11	2	0.0863
19	48	18	101	25	4	0.1952
20	131	126	387	6	0	0.1688
21	4	0	103	49	9	0.0786
22	1	4	46	16	2	0.0955
23	0	0	468	56	2	0.0486
24	7	0	52	37	5	0.0867
25	5	1	6	95	11	0.1403
26	174	113	285	69	18	0.3786
27	0	0	6	35	4	0.0761
28	233	153	682	404	85	0.8927
29	155	56	94	75	17	0.3621
30	120	74	55	120	8	0.1758
31	8983	37	236	77	38	0.2699
32	59	54	138	55	11	0.2762
33	72	112	169	228	39	0.3240

```
 34    571     78     25    162     43  0.3737
 35    853   1002   1017    418     57  0.9114
 36      5      0     17     14     13  0.2594
 37     11     34      3     20      4  0.4284
 38    258      1     33     48     13  0.1905
 39     69     14    126    108     20  0.2341
 40   4790   2046   3719     31      7  0.7759
```

الرسم المصفوفي



يلاحظ علاقة خطية بين $X4$ و $X5$.

```
MTB > Regress 'Y' 5 'X1' 'X2' 'X3' 'X4' 'X5';
SUBC>   Residuals 'RESI1';
SUBC>   SResiduals 'SRES1';
SUBC>   Tresiduals 'TRES1';
SUBC>   Hi 'HI1';
SUBC>   Cookd 'COOK1';
SUBC>   DFits 'DFIT1';
SUBC>   Coefficients 'COEF1';
```

```
SUBC>   Fits 'FITS1';
SUBC>   MSE 'MSE1';
SUBC>   XPXInverse 'XPXI1';
SUBC>   RMatrix 'RMAT1';
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: Y versus X1, X2, X3, X4, X5

```
The regression equation is
Y = 0.122 -0.000053 X1 +0.000045 X2 +0.000249 X3 -0.000867 X4 + 0.0134 X5

Predictor        Coef     SE Coef          T        P       VIF
Constant      0.12176     0.03162       3.85    0.000
X1        -0.00005284  0.00001786      -2.96    0.006       1.4
X2          0.0000454   0.0001534       0.30    0.769       5.6
X3         0.00024949  0.00008831       2.83    0.008       6.8
X4         -0.0008666   0.0003761      -2.30    0.027       8.3
X5           0.013385    0.002279       5.87    0.000       7.7

S = 0.1502             R-Sq = 85.0%         R-Sq(adj) = 82.7%
PRESS = 6.97367        R-Sq(pred) = 0.00%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     4.32851     0.86570     38.39    0.000
Residual Error    34     0.76677     0.02255
Total             39     5.09528

No replicates. Cannot do pure error test.

Source       DF      Seq SS
X1            1     0.16258
X2            1     1.02780
```

X3	1	1.16960
X4	1	1.19050
X5	1	0.77804

Obs	X1	Y	Fit	SE Fit	Residual	St Resid
1	102	0.3574	0.5262	0.0386	-0.1688	-1.16
2	1120	1.9673	1.6891	0.1281	0.2782	3.55RX
3	139	0.1862	0.1821	0.0315	0.0041	0.03
4	221	0.3816	0.2917	0.0901	0.0899	0.75
5	12	0.1512	0.1270	0.0313	0.0242	0.17
6	1	0.1449	0.1298	0.0315	0.0151	0.10
7	1046	0.4711	0.5602	0.0534	-0.0891	-0.63
8	2032	0.6512	0.9622	0.0723	-0.3110	-2.36R
9	895	0.6624	0.4458	0.0312	0.2166	1.47
10	0	0.3457	0.1356	0.0309	0.2101	1.43
11	25	0.3355	0.2597	0.0264	0.0758	0.51
12	97	0.3982	0.3441	0.0262	0.0541	0.37
13	1	0.2044	0.2328	0.0272	-0.0284	-0.19
14	4	0.2969	0.2132	0.0283	0.0837	0.57
15	42	1.1515	0.9202	0.1173	0.2313	2.47RX
16	87	0.5609	0.3046	0.0447	0.2563	1.79
17	2	0.1104	0.1877	0.0289	-0.0773	-0.52
18	2	0.0863	0.1465	0.0306	-0.0602	-0.41
19	48	0.1952	0.1771	0.0306	0.0181	0.12
20	131	0.1688	0.2119	0.0366	-0.0431	-0.30
21	4	0.0786	0.2253	0.0292	-0.1467	-1.00
22	1	0.0955	0.1463	0.0313	-0.0508	-0.35
23	0	0.0486	0.2168	0.0541	-0.1682	-1.20
24	7	0.0867	0.1692	0.0305	-0.0825	-0.56
25	5	0.1403	0.1880	0.0322	-0.0477	-0.32
26	174	0.3786	0.3699	0.0276	0.0087	0.06
27	0	0.0761	0.1465	0.0309	-0.0704	-0.48
28	233	0.8927	1.0742	0.0671	-0.1815	-1.35
29	155	0.3621	0.3021	0.0254	0.0600	0.41
30	120	0.1758	0.1356	0.0430	0.0402	0.28
31	8983	0.2699	0.1496	0.1439	0.1203	2.80RX
32	59	0.2762	0.2551	0.0267	0.0211	0.14
33	72	0.3240	0.4896	0.0328	-0.1656	-1.13
34	571	0.3737	0.5365	0.0430	-0.1628	-1.13
35	853	0.9114	0.7766	0.1024	0.1348	1.23 X
36	5	0.2594	0.2876	0.0319	-0.0282	-0.19
37	11	0.4284	0.1597	0.0293	0.2687	1.82

38	258	0.1905	0.2488	0.0266	-0.0583	-0.39
39	69	0.2341	0.3243	0.0261	-0.0902	-0.61
40	4790	0.7759	0.9562	0.1416	-0.1803	-3.60RX

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

Durbin-Watson statistic = 1.70

Lack of fit test
Possible interactions with variable X2 (P-Value = 0.004)
Possible interactions with variable X3 (P-Value = 0.026)
Possible curvature in variable X4 (P-Value = 0.045)
Possible interactions with variable X4 (P-Value = 0.016)
Possible interactions with variable X5 (P-Value = 0.007)
Possible lack of fit at outer X-values (P-Value = 0.008)
Overall lack of fit test is significant at P = 0.004

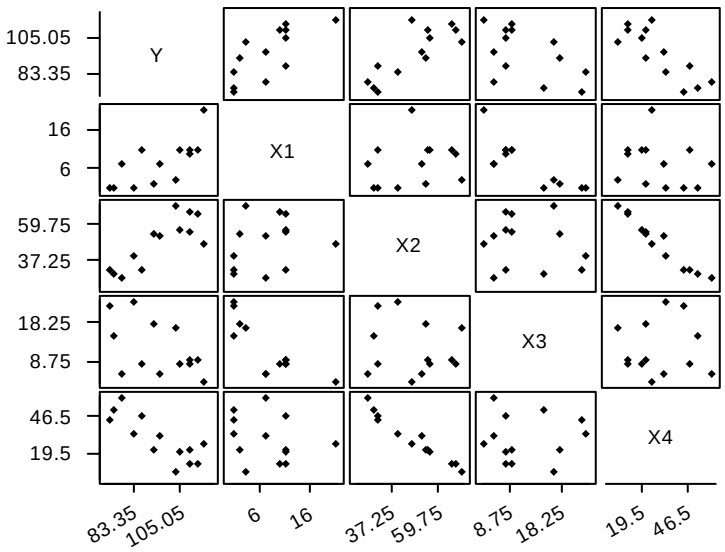
RESI1	SRES1	TRES1	HI1	COOK1	DFIT1
-0.168833	-1.16334	-1.16962	0.066073	0.0160	-0.3111
0.278244	3.55058	4.40978	0.727690	5.6148	7.2087
0.004059	0.02765	0.02724	0.044031	0.0000	0.0058
0.089915	0.74843	0.74350	0.360015	0.0525	0.5576
0.024236	0.16502	0.16264	0.043572	0.0002	0.0347
0.015066	0.10261	0.10111	0.044040	0.0001	0.0217
-0.089117	-0.63492	-0.62925	0.126436	0.0097	-0.2394
-0.311024	-2.36320	-2.54672	0.231927	0.2811	-1.3994
0.216626	1.47466	1.50162	0.043127	0.0163	0.3188
0.210053	1.42935	1.45248	0.042369	0.0151	0.3055
0.075796	0.51272	0.50709	0.030973	0.0014	0.0907
0.054059	0.36558	0.36088	0.030424	0.0007	0.0639
-0.028426	-0.19247	-0.18972	0.032742	0.0002	-0.0349
0.083662	0.56723	0.56149	0.035398	0.0020	0.1076
0.231344	2.46704	2.68240	0.610079	1.5871	3.3553
0.256313	1.78772	1.85034	0.088504	0.0517	0.5766
-0.077257	-0.52422	-0.51855	0.036922	0.0018	-0.1015
-0.060238	-0.40974	-0.40467	0.041630	0.0012	-0.0843
0.018082	0.12298	0.12119	0.041427	0.0001	0.0252
-0.043113	-0.29603	-0.29202	0.059467	0.0009	-0.0734
-0.146652	-0.99560	-0.99546	0.037891	0.0065	-0.1976

-0.050772	-0.34569	-0.34117	0.043495	0.0009	-0.0728
-0.168163	-1.20054	-1.20865	0.129993	0.0359	-0.4672
-0.082528	-0.56124	-0.55550	0.041213	0.0023	-0.1152
-0.047652	-0.32486	-0.32055	0.045961	0.0008	-0.0704
0.008657	0.05865	0.05778	0.033694	0.0000	0.0108
-0.070369	-0.47883	-0.47333	0.042326	0.0017	-0.0995
-0.181491	-1.35077	-1.36797	0.199508	0.0758	-0.6829
0.059979	0.40522	0.40019	0.028567	0.0008	0.0686
0.040206	0.27943	0.27561	0.081994	0.0012	0.0824
0.120324	2.80131	3.14674	0.918193	14.6796	10.5422
0.021099	0.14277	0.14070	0.031583	0.0001	0.0254
-0.165648	-1.13032	-1.13510	0.047681	0.0107	-0.2540
-0.162848	-1.13188	-1.13673	0.082143	0.0191	-0.3401
0.134763	1.22733	1.23686	0.465400	0.2186	1.1540
-0.028216	-0.19228	-0.18953	0.045175	0.0003	-0.0412
0.268718	1.82437	1.89235	0.037992	0.0219	0.3761
-0.058320	-0.39461	-0.38965	0.031448	0.0008	-0.0702
-0.090200	-0.60992	-0.60420	0.030209	0.0019	-0.1066
-0.180335	-3.59933	-4.50719	0.888691	17.2390	-12.7355

مثال (8)

Variable Selection in Multiple Regression: Hald Data

X1	X2	X3	X4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4



يبدو أن هناك علاقة خطية بين $X2$ و $X4$.

```
MTB > BReg 'Y' 'X1' 'X2' 'X3' 'X4' ;
SUBC>   NVars 1 4;
SUBC>   Best 4;
SUBC>   Constant.
```

Best Subsets Regression: Y versus X1, X2, X3, X4

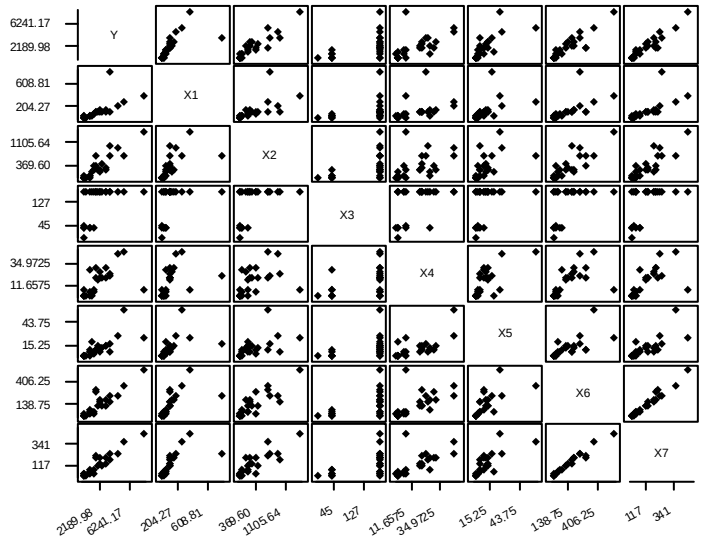
Response is Y

Vars	R-Sq	R-Sq(adj)	C-p	S	X 1	X 2	X 3	X 4
1	67.5	64.5	138.7	8.9639				X
1	66.6	63.6	142.5	9.0771		X		
1	53.4	49.2	202.5	10.727	X			
1	28.6	22.1	315.2	13.278			X	
2	97.9	97.4	2.7	2.4063	X	X		
2	97.2	96.7	5.5	2.7343	X			X
2	93.5	92.2	22.4	4.1921			X	X
2	84.7	81.6	62.4	6.4455		X	X	
3	98.2	97.6	3.0	2.3087	X	X		X
3	98.2	97.6	3.0	2.3121	X	X	X	
3	98.1	97.5	3.5	2.3766	X		X	X
3	97.3	96.4	7.3	2.8638		X	X	X
4	98.2	97.4	5.0	2.4460	X	X	X	X

مثال (9)

Multicollinearity in Linear Regression: Manhours Data

Y	X1	X2	X3	X4	X5	X6	X7
180.23	2.00	4.00	4.0	1.26	1	6	6
182.61	3.00	1.58	40.0	1.25	1	5	5
199.92	5.30	1.67	42.5	7.79	3	25	25
284.55	7.00	2.37	168.0	1.00	1	7	8
267.38	16.50	8.25	168.0	1.12	2	19	19
164.38	16.60	23.78	40.0	1.00	1	13	13
999.09	25.89	3.00	40.0	0.00	3	36	36
931.84	31.92	40.80	168.0	5.52	6	47	47
944.21	39.63	50.86	40.0	27.37	10	77	77
1103.24	44.42	159.75	168.0	0.60	18	48	48
1387.82	54.48	207.08	40.0	7.77	6	66	66
1489.50	56.63	373.42	168.0	6.03	4	36	37
1845.89	95.00	368.00	168.0	30.26	9	292	196
1891.70	96.67	206.67	168.0	17.86	14	120	120
1880.84	96.83	677.33	168.0	20.31	10	302	210
2268.06	97.33	255.08	168.0	19.00	6	165	130
3036.63	102.33	288.83	168.0	21.01	14	131	131
2628.32	110.24	410.00	168.0	20.05	12	115	115
3559.92	113.88	981.00	168.0	24.48	6	166	179
2227.76	134.32	145.82	168.0	25.99	12	192	192
3115.29	149.58	233.83	168.0	31.07	14	185	202
4804.24	188.74	937.00	168.0	45.44	26	237	237
5539.98	274.92	695.25	168.0	46.63	58	363	363
8266.77	384.50	1473.66	168.0	7.36	24	540	453
3534.49	811.08	714.33	168.0	22.76	17	242	242



هناك علاقة خطية بين $X7$ وكلا من $X1$ و $X6$.

```
MTB > Regress 'Y' 7 'X1' 'X2' 'X3' 'X4' 'X5' 'X6' 'X7';
SUBC>   Residuals 'RESI1';
SUBC>   SResiduals 'SRES1';
SUBC>   Tresiduals 'TRES1';
SUBC>   Hi 'HI1';
SUBC>   Cookd 'COOK1';
SUBC>   DFits 'DFIT1';
SUBC>   Coefficients 'COEF1';
SUBC>   Fits 'FITS1';
SUBC>   MSE 'MSE1';
SUBC>   XPXInverse 'XPXI1';
SUBC>   RMatrix 'RMAT1';
SUBC>   Constant;
SUBC>   VIF;
SUBC>   DW;
SUBC>   Press;
SUBC>   Pure;
SUBC>   XLOF;
SUBC>   Brief 3.
```

Regression Analysis: Y versus X1, X2, X3, X4, X5, X6, X7

The regression equation is
Y = 148 - 1.29 X1 + 1.81 X2 + 0.59 X3 - 21.5 X4 + 5.6 X5 - 14.5 X6 + 29.4 X7

Predictor	Coef	SE Coef	T	P	VIF
Constant	148.2	221.6	0.67	0.513	
X1	-1.2874	0.8057	-1.60	0.129	2.2
X2	1.8096	0.5152	3.51	0.003	4.5
X3	0.590	1.800	0.33	0.747	1.4
X4	-21.48	10.22	-2.10	0.051	2.4
X5	5.62	14.76	0.38	0.708	3.7
X6	-14.515	4.226	-3.43	0.003	37.2
X7	29.360	6.370	4.61	0.000	63.7

S = 455.5 R-Sq = 96.1% R-Sq(adj) = 94.5%
PRESS = 212840622 R-Sq(pred) = 0.00%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	87382503	12483215	60.17	0.000
Residual Error	17	3526698	207453		
Total	24	90909201			

No replicates. Cannot do pure error test.

Source	DF	Seq SS
X1	1	37281241
X2	1	38025044
X3	1	418515
X4	1	1739294
X5	1	4362717
X6	1	1149032
X7	1	4406662

Obs	X1	Y	Fit	SE Fit	Residual	St Resid
1	2	180.2	222.9	215.1	-42.6	-0.11
2	3	182.6	223.8	171.3	-41.2	-0.10

3	5	199.9	390.2	163.8	-190.2	-0.45
4	7	284.6	360.1	192.5	-75.6	-0.18
5	17	267.4	510.3	189.8	-243.0	-0.59
6	17	164.4	370.6	171.3	-206.2	-0.49
7	26	999.1	695.2	183.9	303.9	0.73
8	32	931.8	893.0	171.0	38.8	0.09
9	40	944.2	824.2	234.9	120.0	0.31
10	44	1103.2	1280.2	278.1	-176.9	-0.49
11	54	1387.8	1323.0	172.1	64.8	0.15
12	57	1489.5	1307.0	210.9	182.5	0.45
13	95	1845.9	1707.9	336.3	138.0	0.45
14	97	1891.7	1973.4	132.9	-81.7	-0.19
15	97	1880.8	2750.6	275.7	-869.7	-2.40R
16	97	2268.1	1631.2	161.8	636.9	1.50
17	102	3036.6	2210.5	124.2	826.2	1.89
18	110	2628.3	2191.4	137.6	436.9	1.01
19	114	3559.9	4229.9	340.4	-670.0	-2.21R
20	134	2227.8	2697.8	277.0	-470.1	-1.30
21	150	3115.3	3134.8	289.5	-19.5	-0.06
22	189	4804.2	4388.4	303.9	415.8	1.23
23	275	5540.0	5864.8	403.8	-324.8	-1.54
24	385	8266.8	7858.2	425.3	408.6	2.51R
25	811	3534.5	3695.1	452.8	-160.6	-3.28RX

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

Durbin-Watson statistic = 1.92
Possible lack of fit at outer X-values (P-Value = 0.053)
Overall lack of fit test is significant at P = 0.053

RESI1	SRES1	TRES1	HI1	COOK1	DFIT1
-42.642	-0.10621	-0.10307	0.223020	0.000	-0.0552
-41.219	-0.09767	-0.09478	0.141423	0.000	-0.0385
-190.247	-0.44766	-0.43687	0.129383	0.004	-0.1684
-75.551	-0.18302	-0.17773	0.178565	0.001	-0.0829
-242.960	-0.58682	-0.57515	0.173692	0.009	-0.2637
-206.249	-0.48870	-0.47748	0.141438	0.005	-0.1938
303.856	0.72922	0.71878	0.163058	0.013	0.3173
38.814	0.09195	0.08922	0.140991	0.000	0.0361
120.005	0.30752	0.29917	0.265919	0.004	0.1801

-176.916	-0.49041	-0.47917	0.372677	0.018	-0.3693
64.772	0.15360	0.14912	0.142839	0.000	0.0609
182.505	0.45210	0.44126	0.214468	0.007	0.2306
137.980	0.44926	0.43846	0.545310	0.030	0.4802
-81.728	-0.18760	-0.18218	0.085100	0.000	-0.0556
-869.744	-2.39894	-2.86152	0.366384	0.416	-2.1760
636.880	1.49591	1.55736	0.126260	0.040	0.5920
826.176	1.88533	2.05664	0.074343	0.036	0.5828
436.923	1.00631	1.00671	0.091288	0.013	0.3191
-670.013	-2.21380	-2.54579	0.558461	0.775	-2.8631
-470.080	-1.30018	-1.32917	0.369886	0.124	-1.0184
-19.486	-0.05541	-0.05376	0.403966	0.000	-0.0443
415.820	1.22568	1.24539	0.445203	0.151	1.1156
-324.818	-1.54199	-1.61300	0.786106	1.092	-3.0923
408.565	2.50518	3.05998	0.871789	5.334	7.9792
-160.643	-3.27951	-5.24939	0.988434	114.892	-48.5277

إستعراض مفصل لإستخدام R في تحليل الإنحدار:

مقدمة عن R:

أنواع البيانات Data Types:

1) **متجهات Vectors:**

```
a  = c(1,2,5.3,6,-2,4)
```

متجه عددي

```
b = c( "one","two","three")
```

متجه نصي

```
c = c(TRUE,TRUE,FALSE,TRUE,FALSE)
```

متجه منطقي

للإشارة لعنصر نستخدم التحت تنصيص subscripts كالتالي:

```
a[c(2,4)]
```

وهذا يشير للعنصر الثاني والعنصر الرابع للمتجه a.

2) **مصفوفات Matrices:**

كل أعمدة المصفوفة يجب أن تكون من نفس النوع mode إما عددية أو نصية أو منطقية الخ ولها نفس الطول الشكل العام لتعريف مصفوفة هو:

```
mymat = matrix( vector, nrow = r, ncol = c, byrow = FALSE,
dimnames = list( char_vector_rownames,
char_vector_colnames))
```

الخيار `byrow = TRUE` يبين ان المصفوفة يجب ان تقرأ سطرا بسطر و `byrow = FALSE` يبين ان المصفوفة يجب ان تقرأ عمودا عمود (وهو القيمة الإفتراضية). `dimnames` إختيارية لإعطاء أسماء للأعمدة و الأسطر.

أمثلة:

```
y = matrix(1:20, nrow =5, ncol = 4)
```

تولد مصفوفة 5X4 .

```
cells = c(1,26,24,68)
rnames = c("R1", "R2")
cnames = c("C1", "C2")
mymat = matrix(cells, nrow =2, ncol =2,
byrow = TRUE, dimnames = list(rnames,
cnames))
```

للإشارة لعناصر مصفوفة نستخدم التحت تنصيص كالتالي:

```
y[,4]
```

ويعني العمود الرابع للمصفوفة.

```
y[3,]
```

الصف الثالث للمصفوفة.

```
y[2:4, 1:3]
```

الصفوف 2 و 3 و 4 وألأعمدة 1 و 2 و 3 .

المنظمات Arrays:

المنظمات مثل المصفوفات ولكن يمكن أن يكون لها أكثر من بعدين.

أطر البيانات Data Frames:

إطار البيانات أكثر عمومية من المصفوفة بحيث الأعمدة المختلفة يمكن أن يكون لها نوع مختلف.

مثال:

```
d = c(1, 2, 3, 4)
e = c("red", "blue", "red", NA)
f = c( TRUE, TRUE, FALSE, TRUE)
```

```
mydat = data.frame(d, e, f)
names(mydat) = c("ID", "Color", "Passed")
```

يوجد عدة طرق للإشارة لعنصر في إطار البيانات مثل:

```
mydat[3:5]
```

يشير للأعمدة 3 و 4 و 5

```
mydat[c("ID", "Color")]
```

يشير للأعمدة المسماه ID و Color

```
mydat$Color
```

يشير للمتغير (العمود) Color

القوائم Lists:

مجموعة مرتبة من الأشياء (المركبات). والقوائم تمكننا من تجميع مختلف الأشياء (التي قد تكون ليس لها علاقة ببعضها) تحت إسم واحد.

مثال:

قائمة بأربعة مركبات (نص و متجه عددي ومصفوفة وعدد)

```
w = list(name = ahmad, mynum = a, mymat = y,
age = 15)
```

مثال لقائمة مكونة من قائمتين.

```
v = c(list1, list2)
```

للإشارة لعنصر من قائمة نستخدم [[]] كالتالي:

```
mylist[[2]]
```

وتشير للمركبة الثانية من القائمة.

```
mylist[["mynumbers"]]
```

تشير للمركبة المسماة myname في القائمة.

العوامل Factors:

لجعل متغير إسمي nominal (أو رمزي) نستخدم العامل. العامل يخزن القيم الإسمية في متجه من الأرقام الصحيحة في المجال 1,…,k و يخزن القيم الأصلية في متجة داخلي مؤشر لهذه القيم.

مثال:

سوف نكون متغير يحوي نوع 20 ذكرا و 30 انثى

```
gender = c(rep("male",20), rep("female",
30))
gender = factor(gender)
summary(gender)
```

والعامل المرتب يستخدم لتمثيل متغير مرتب ordinal.

```
rating = c("large", "midium", "small")
rating = ordered(rating)
```

Rسوف يعامل العوامل كمتغيرات رمزية والعوامل الرتبية كمتغيرات رتبية في التحليل الإحصائي.

بعض الدوال المفيدة:

`length(object)`	عدد عناصر object
`str(object)`	تركيب العنصر
`class(object)`	نوع او فئة العنصر
`names(object)`	اسماء مركبات العنصر
`c(object, object, …)`	ضم أو جمع العناصر في متجه
`cbind(object, object, …)`	ضم أو جمع العناصر في عمود
`rbind(object, object, …)`	ضم أو جمع العناصر في سطر

يطبع الشيئ	`object`
يسرد المتغيرات الأشياء الحالية	`ls()`
يمسح الشيئ	`rm(object)`
يحرر وينسخ ويحفظ object ك newobj	`newobj = edit(object)`
يحرر الشيئ	`fix(object)`

إدخال و إستيراد بيانات Importing Data:

1- من ملف نصي محدد بفاصلة Comma Delimited Text File:

السطر الأول يحوي أسماء المتغيرات ويستخدم الفاصلة أو الفراغ كفاصل بين القيم

```
mydata = read.table("C://data/mydata.csv", header =
TRUE, sep = ",", row.names = "id")
```

أستخدمنا لأسماء الأسطر المتغير id.

ملاحظة:

يمكن إختيار الملف بالملاحة خلال الأدلة وإختيار الملف المناسب كالتالي:

```
mydata = read.table(file.choose(), header = TRUE,
sep = ",", row.names = "id")
```

تظهر نافذة تحوي محتويات الدليل الحالي أو دليل العمل والذي يحدد بـ

```
getwd()
setwd(dir)
```

ومن الأفضل جعل dir الدليل الذي يحوي البيانات

2- من Excel على شكل ملف csv:

السطر الأول يحوي أسماء المتغيرات ويستخدم الفاصلة أو الفراغ كفاصل بين القيم

```
mydata = read.csv("C://data/mydata.csv", header =
TRUE, sep = ",", row.names = "id")
```

أستخدمنا لأسماء الأسطر المتغير id.

ملاحظة:

لإستيراد عدد ضخم من البيانات على شكل مصفوفة ذات أبعاد مثلا 200X2000 نستخدم

```
mydata = matrix(scan("C://data/matrix.dat", n =
200*2000), 200, 2000, byrow = TRUE)
```

أو البدائل الأبطء

```
mydata =
as.matrix(read.table("C://data/matrix.dat"))
```

أو

```
mydata =
as.matrix(read.table("C://data/matrix.dat", header
= FALSE, nrows = 200)
```

لقرائة أنواع مختلفة من الملفات مثل ملفات SAS أو SPSS أو MINITAB ألخ

تحمل الحزمة foreign من CRAN (أرشيف R) فمثلا

للقرائة من MINITAB

```
library(foreign)
mydata = read.mtp( "C://Program Files/Minitab
15/English/Sample Data/Student1/Exam.mtp")
```

لاحظ أن نوع الملف mtp أي Minitab Portable Worksheet .

للقرائة من SAS:

يجب أن يكون الملف على شكل SAS XPORT format

```
mydata = read.xport( "C://Program Files/SAS/Sample
Data/Exam.xport")
```

ملاحظة:

نستخدم الإجراء

```
PROC EXPORT DATA = data-set OUTFILE = 'filename';
```

في SAS .

للقرائة من SPSS:

```
mydata = read.spss( "C://Program Files/SPSS/Sample Data/Exam.sav")
```

تصدير البيانات Exporting Data:

1- إلى ملف نصي محدد بفاصلة Comma Delimited Text File:

```
write.table(mydata,"C://data/mydata.txt", sep = "\t")
```

لاحظ أن \t تعني الضغط على enter .

2- إلى Excel على شكل ملف csv:

```
write.csv(mydata, "C://data/mydata.csv")
```

3- إلى SAS:

```
library(foreign)
write.foreign(mydata, "C://data/mydata.txt",
"C://data/mydata.sas", package = "SAS")
```

4- إلى SPSS:

```
library(foreign)
write.foreign(mydata, "C://data/mydata.txt",
"C://data/mydata.sps", package = "SPSS")
```

5- إلى **MINITAB:**

```
library(foreign)
write.foreign(mydata, "C://data/mydata.txt",
"C://data/mydata.mtp", package = "MINITAB")
```

تحليل الإنحدار الخطي البسيط والمتعدد Simple and Multiple Linear Regression:

يعطي R دعم شامل لتحليل الإنحدار وسوف نستعرض بعضها:

لتطبيق نموذج لمتغير تابع أو متغير إستجابه y على متغيرات مستقلة x1 و x2 و x3

مثلا نستخدم lm():

```
fit = lm(y ~ x1 + x2 + x3, data = mydata)
summary(fit)
```

summary() يطبع النتائج.

مثال لإنحدار خطي بسيط:

```
> x <- c(0.45, 0.08, -1.08, 0.92, 1.65, 0.53, 0.52, -2.15,
+ -2.20, -0.32, -1.87, -0.16, -0.19, -0.98, -0.20, 0.67,
+ 0.08, 0.38, 0.76, -0.78)
> y <- c(1.26, 0.58, -1.00, 1.07, 1.28, -0.33, 0.68, -2.22,
+ -1.82, -1.17, -1.54, 0.35, -0.23, -1.53, 0.16, 0.91,
+  0.22, 0.44, 0.98, -0.98)
> plot(x,y)
> fit = lm(y ~ x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9028 -0.1878  0.1102  0.2540  0.7741
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04574    0.10211   0.448     0.66
x            0.97810    0.09870   9.910 1.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.4485 on 18 degrees of freedom
Multiple R-squared: 0.8451,     Adjusted R-squared: 0.8365
F-statistic:  98.2 on 1 and 18 DF,  p-value: 1.027e-08

>
```

ونحصل على تفاصيل أكثر كالتالي:

```
> coefficients(fit)
(Intercept)           x
 0.04574141  0.97810494
> confint(fit)
                2.5 %    97.5 %
(Intercept) -0.1687853 0.2602682
x            0.7707388 1.1854711
> fitted(fit)
         1          2          3          4          5
6          7          8          9         10
 0.4858886  0.1239898 -1.0106119  0.9455980  1.6596146
0.5641370  0.5543560 -2.0571842 -2.1060895 -0.2672522
        11         12         13         14         15
16         17         18         19         20
-1.7833148 -0.1107554 -0.1400985 -0.9128014 -0.1498796
0.7010717  0.1239898  0.4174213  0.7891012 -0.7171804
> residuals(fit)
```

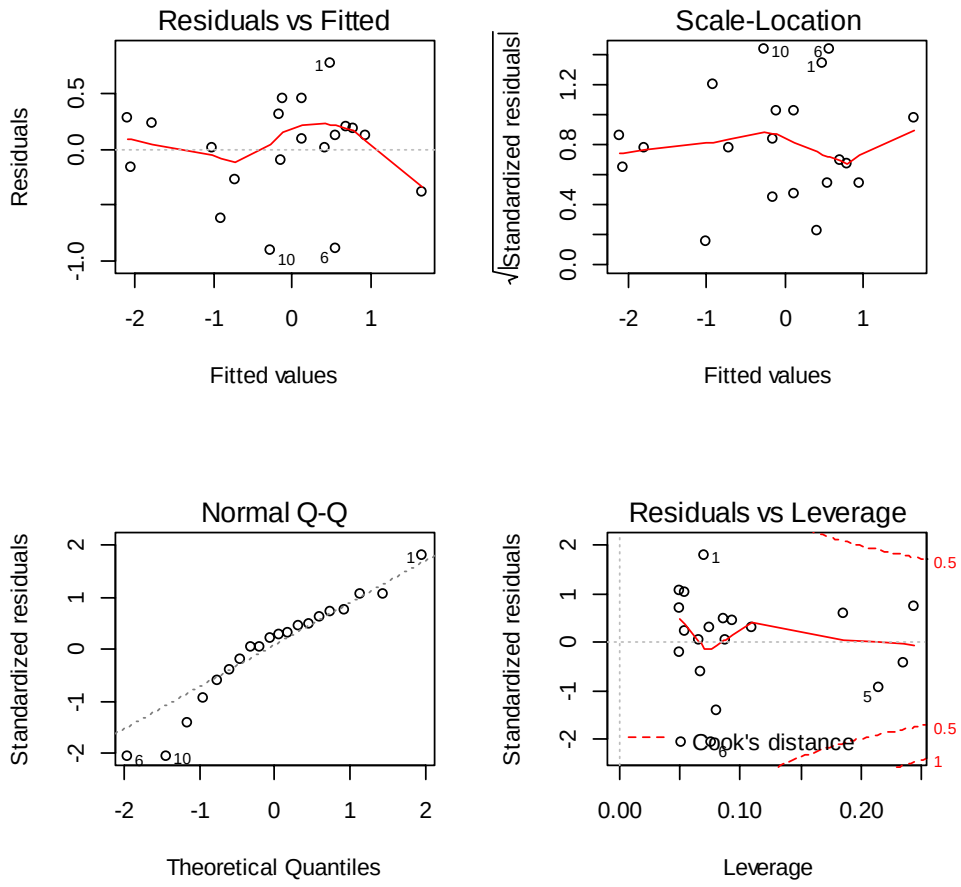
```
          1           2           3           4           5
6           7           8           9
 0.77411137  0.45601019  0.01061193  0.12440204 -0.37961457
-0.89413703  0.12564402 -0.16281579  0.28608946
         10          11          12          13          14
15          16          17          18
-0.90274783  0.24331483  0.46075538 -0.08990147 -0.61719857
0.30987958  0.20892828  0.09601019  0.02257871
         19          20
 0.19089883 -0.26281956
> anova(fit)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 19.7542 19.7542  98.201 1.027e-08 ***
Residuals 18  3.6209  0.2012
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
> vcov(fit)
            (Intercept)           x
(Intercept) 0.010426608 0.001894852
x           0.001894852 0.009742172
> influence(fit)
$hat
         1          2          3          4          5
6          7          8          9         10
0.07011673 0.05364919 0.08797421 0.11015500 0.21476650
0.07542075 0.07472385 0.23519414 0.24478564 0.05076278
        11         12         13         14         15
16         17         18         19         20
```

```
0.18595665 0.05005764 0.05000098 0.07988161 0.05000146
0.08619442 0.05364919 0.06598423 0.09412284 0.06660219

$coefficients
     (Intercept)             x
1   0.0466780442  2.598421e-02
2   0.0253390207  6.405844e-03
3   0.0004847252 -4.989847e-04
4   0.0084577527  7.545792e-03
5  -0.0325715908 -4.318514e-02
6  -0.0549535069 -3.393204e-02
7   0.0077034526  4.698771e-03
8  -0.0067229175  2.016112e-02
9   0.0117846944 -3.679306e-02
10 -0.0464269641  5.780255e-03
11  0.0102274979 -2.425365e-02
12  0.0244093791  8.104083e-04
13 -0.0047356726 -2.062376e-05
14 -0.0285759208  2.551755e-02
15  0.0162925775 -8.688490e-05
16  0.0132935968  9.572382e-03
17  0.0053349778  1.348712e-03
18  0.0013395076  6.725840e-04
19  0.0124313871  9.741403e-03
20 -0.0125257268  7.984163e-03

$sigma
1         2         3         4         5         6
7         8         9        10        11
0.4184332 0.4472904 0.4615051 0.4604033 0.4496652 0.4026533
0.4604244 0.4592987 0.4545538 0.4031031 0.4568547
```

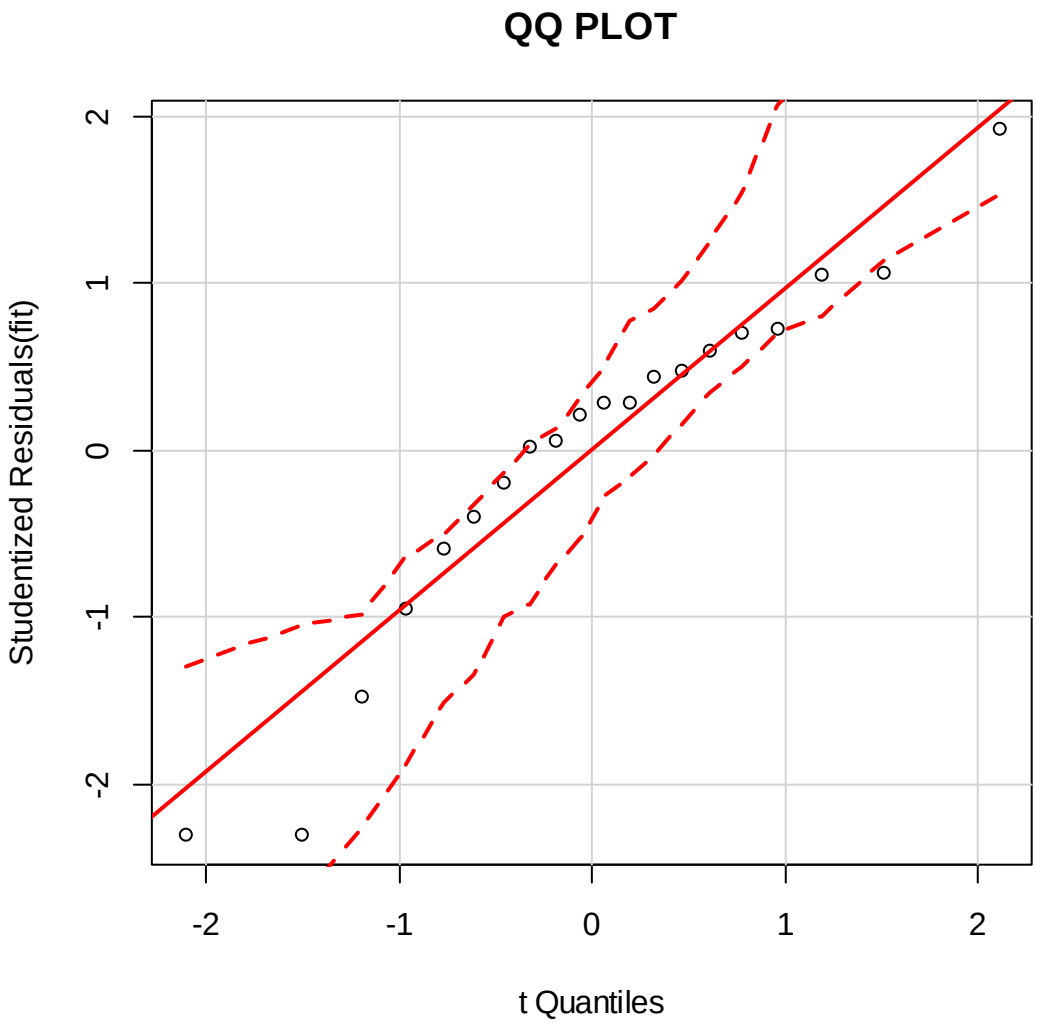
```
12        13        14        15        16        17
18        19        20
0.4470438 0.4609704 0.4343281 0.4550256 0.4584586 0.4608918
0.4614782 0.4589420 0.4567725

$wt.res
1           2           3           4           5
6           7           8           9
 0.77411137  0.45601019  0.01061193  0.12440204 -0.37961457
-0.89413703  0.12564402 -0.16281579  0.28608946
10          11          12          13          14
15          16          17          18
-0.90274783  0.24331483  0.46075538 -0.08990147 -0.61719857
0.30987958  0.20892828  0.09601019  0.02257871
19          20
 0.19089883 -0.26281956

>
```

والرسومات التشخيصية Diagnostic Plots:

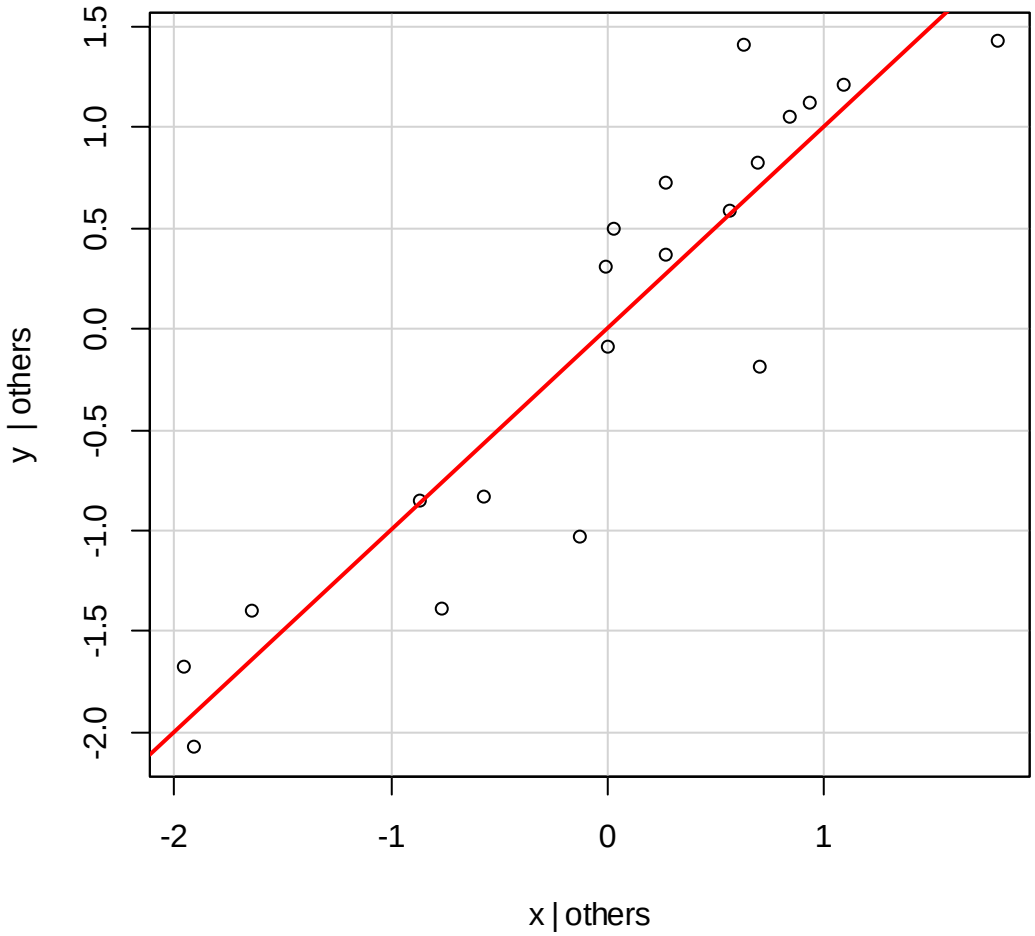
```
> layout(matrix(c(1,2,3,4),2,2))
> plot(fit)
>
```




```
> library(car)
> outlierTest(fit)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
6 -2.309409           0.033747      0.67494
> qqPlot(fit, main = "QQ PLOT")
>
```

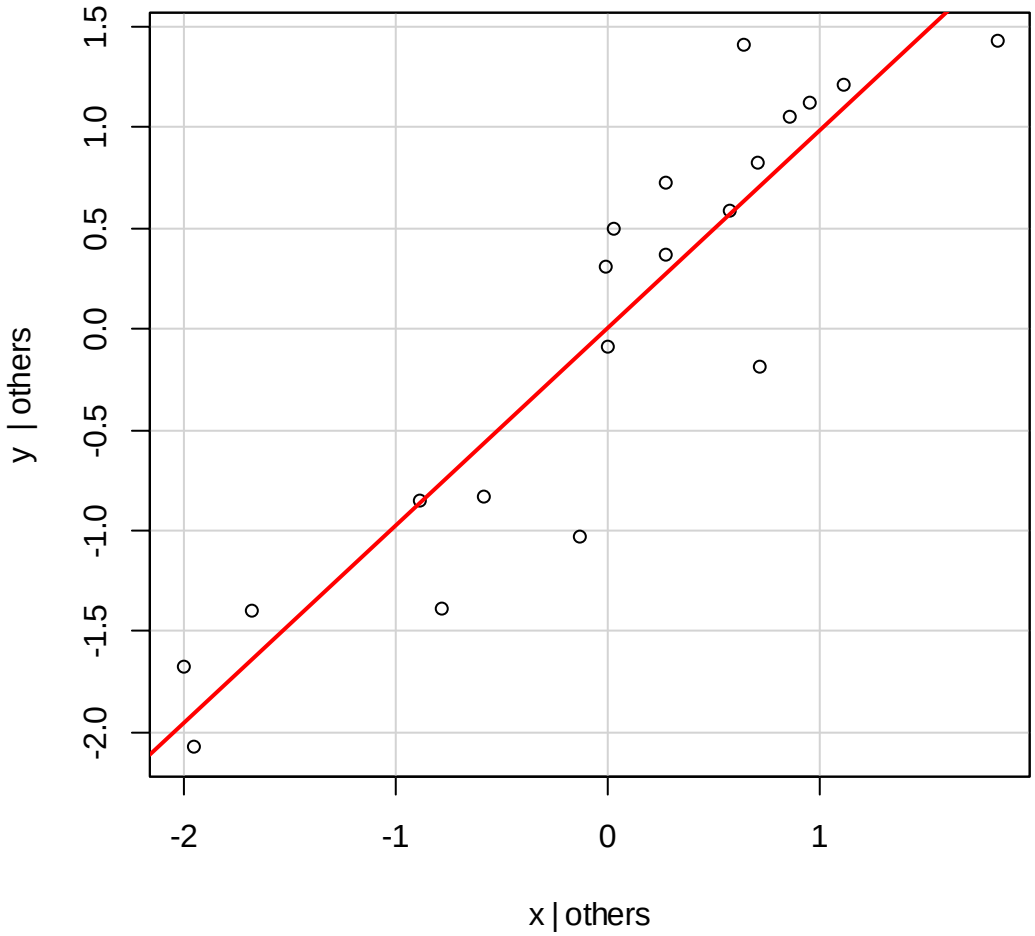



```
> qqPlot(fit, main = "QQ PLOT")
> leveragePlots(fit)
>
```



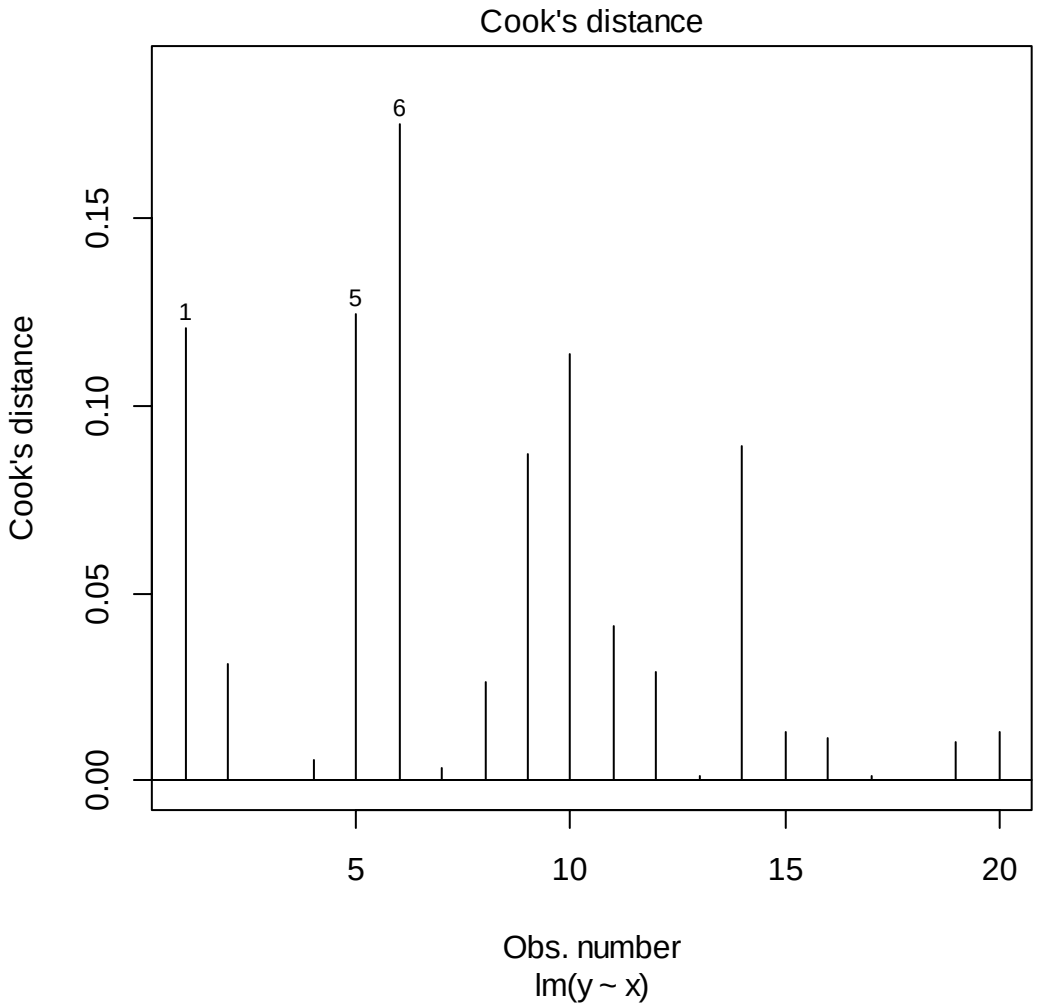
المشاهدات المؤثرة Influential Observations:

```
> # added variable plot
> avPlots(fit)
>
```



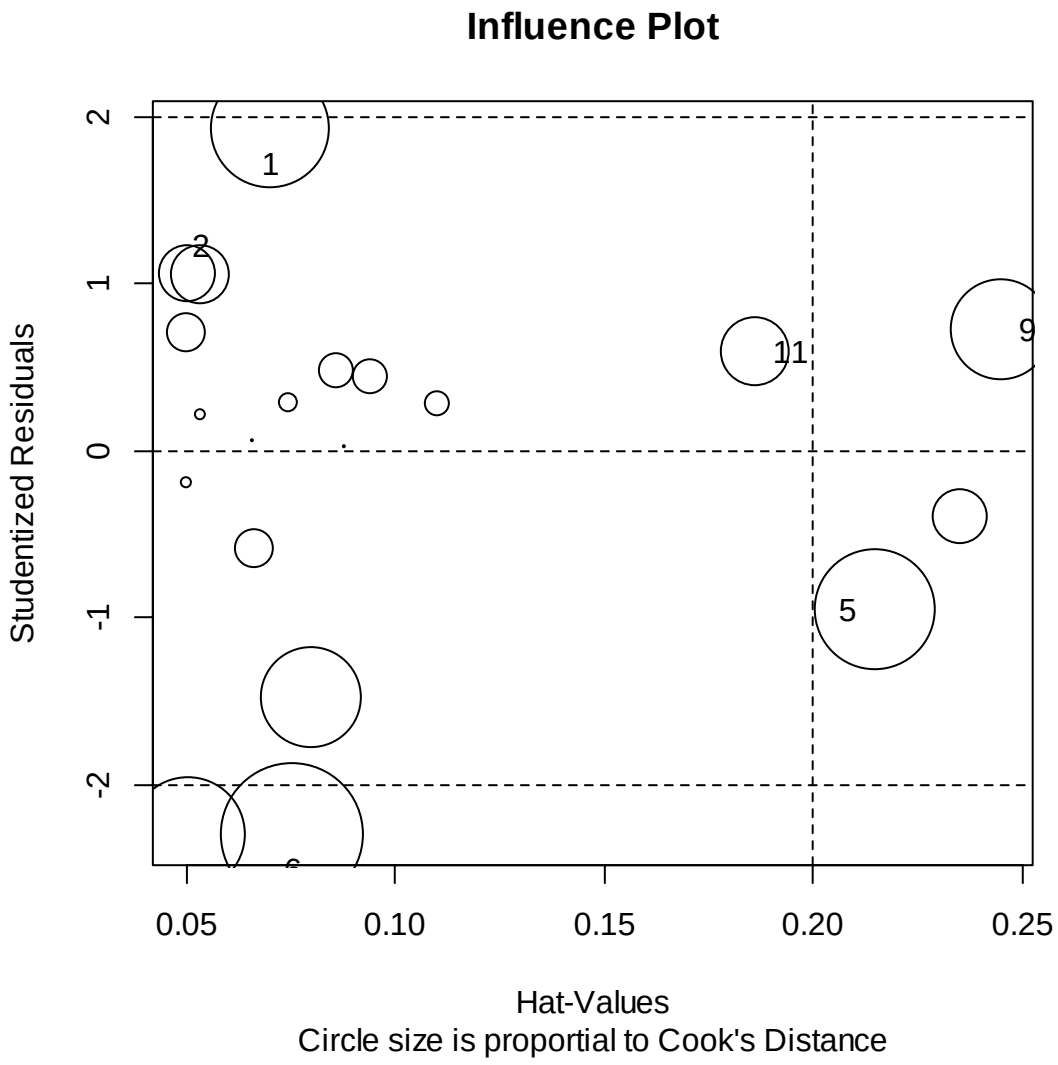

```
> # Cook's D plot
> # identify D values > 4/(n-k-1)
> cutoff=4/((length(y)-length(fit$coefficients)-1))
> plot(fit, cook.levels = cutoff)
Waiting to confirm page change...
Waiting to confirm page change...
Waiting to confirm page change...
Waiting to confirm page change...
> plot(fit, which = 4, cook.levels = cutoff)
> abline(0,0)
```

```
>
```




```
> # Influence Plot
> influencePlot(fit, id.method = "identify", main =
"Influence Plot", sub = "Circle size is proportial to
Cook's Distance")
warning: no point within 0.25 inches
      StudRes        Hat     CookD
1   1.9185059 0.07011673 0.3475352
```

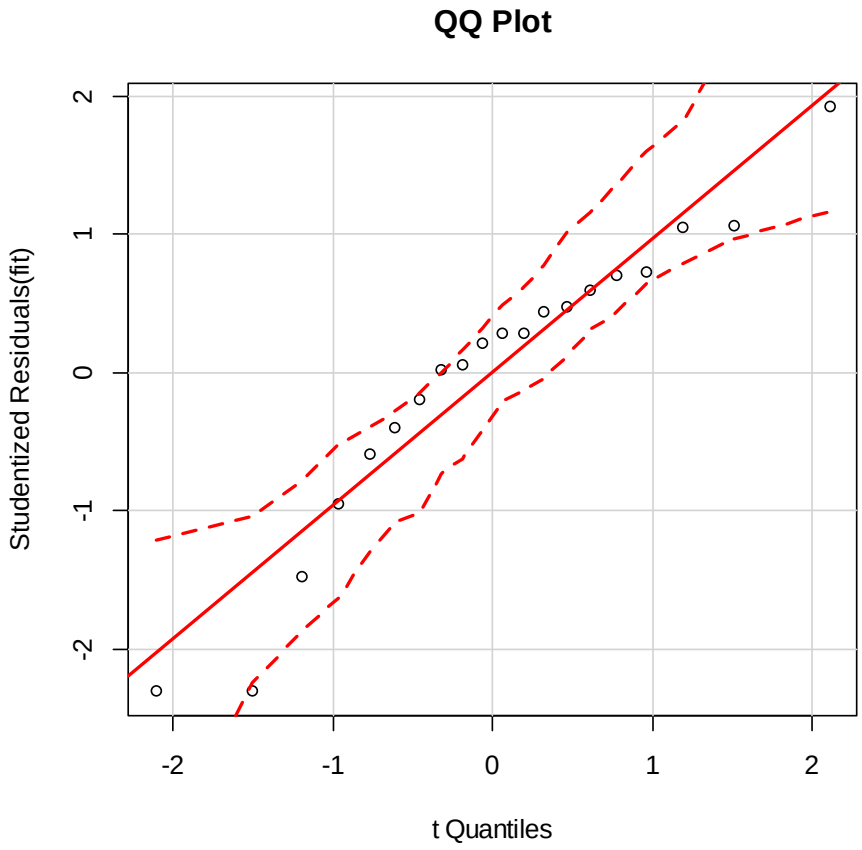
```
2   1.0479943 0.05364919 0.1759612
5  -0.9526956 0.21476650 0.3532159
6  -2.3094086 0.07542075 0.4187148
9   0.7242383 0.24478564 0.2954864
11  0.5902918 0.18595665 0.2032075
>
```



ملاحظة: هذا الرسم تفاعلي يمكن من إختيار المشاهدات المؤثرة عن طريق إختيارها بالفأرة ثم عند الإنتهاء من الإختيار يضغط بالفأرة اليمين ويختار stop.

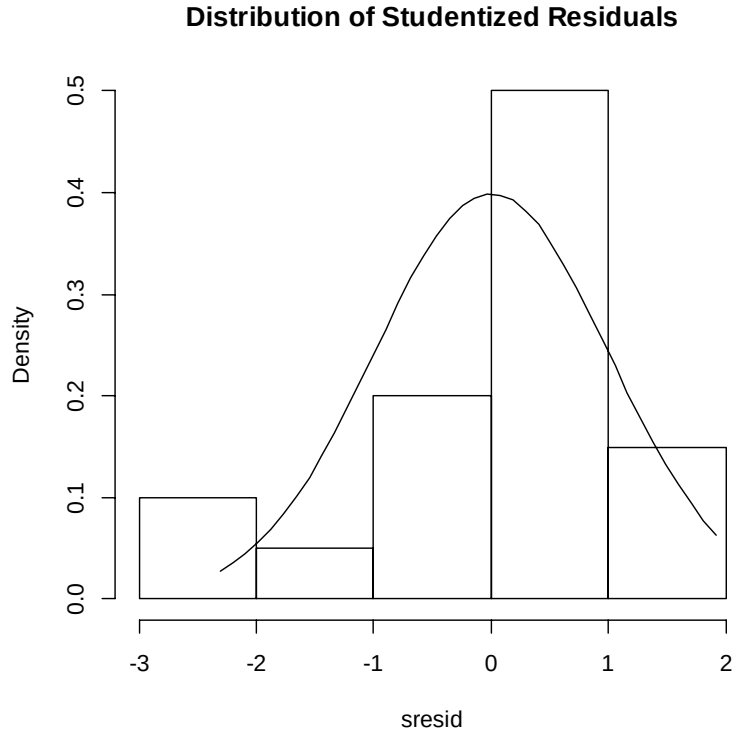
إختبار طبيعية البواقي:

```
> # Normality of Residuals
> # qq plot for studentized resid
> qqPlot(fit, main="QQ Plot")
```



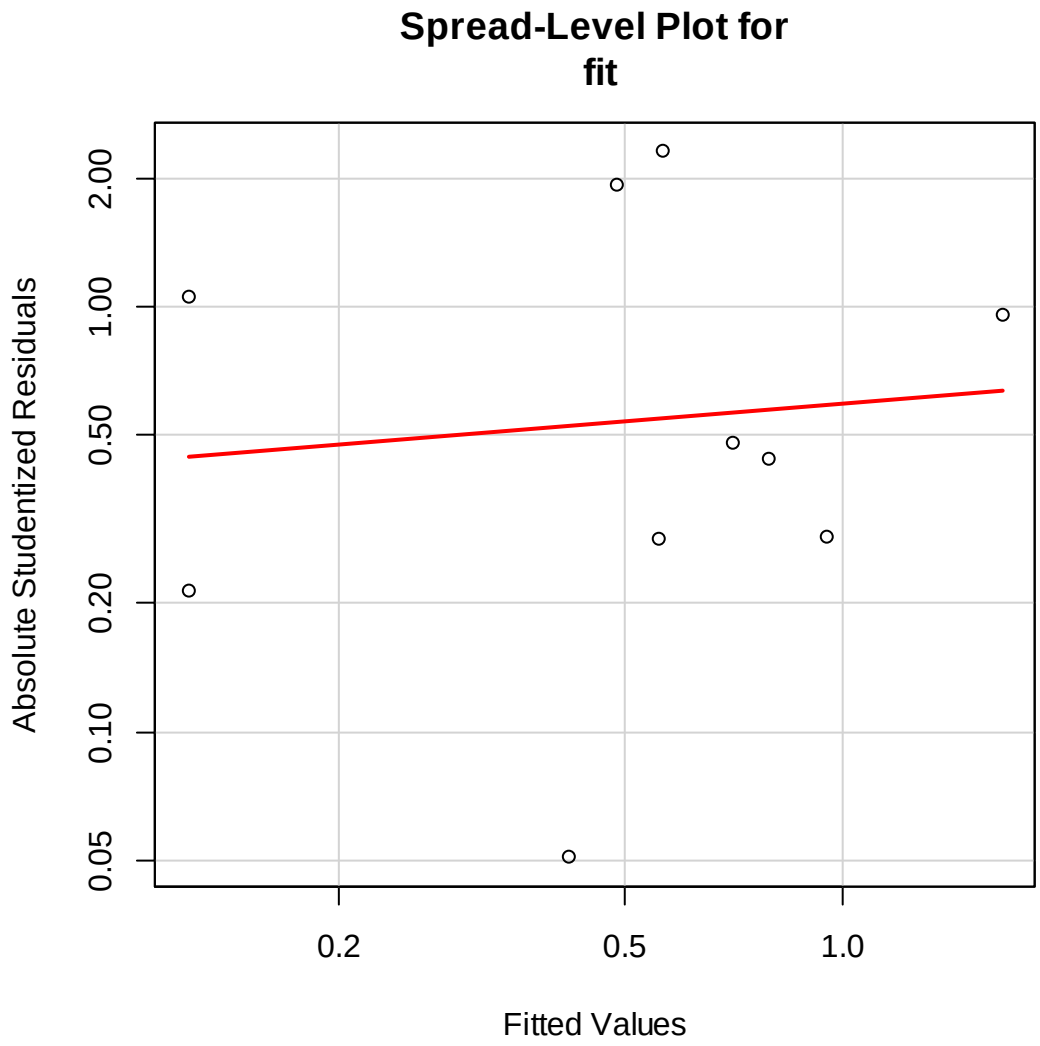
توزيع البواقي المتلمذة Studentized Residuals Distribution:

```
> # distribution of studentized residuals
> library(MASS)
> sresid <- studres(fit)
> hist(sresid, freq=FALSE, main="Distribution of
Studentized Residuals")
> xfit<-seq(min(sresid),max(sresid),length=40)
> yfit<-dnorm(xfit)
> lines(xfit, yfit)
>
```



إختبار ثبات تباين الخطأ:

```
> # Evaluate homoscedasticity
> # non-constant error variance test
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.3043143    Df = 1     p = 0.5811903
> # plot studentized residuals vs. fitted values
> spreadLevelPlot(fit)

Suggested power transformation:  0.8624161
Warning message:
In spreadLevelPlot.lm(fit) : 10 negative fitted
values removed
>
```

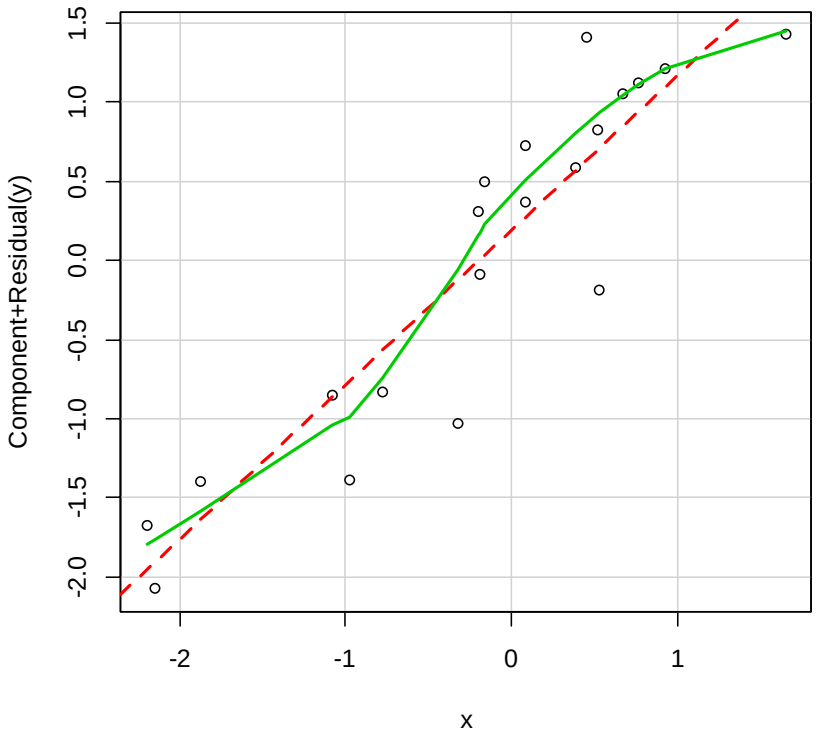


إختبار التعددية الخطية Multi-colinearity: (للإنحدار المتعدد)

```
> # Evaluate Collinearity
> vif(fit) # variance inflation factors
Error in vif.lm(fit) : model contains fewer than 2 terms
> sqrt(vif(fit)) > 2 # problem?
Error in vif.lm(fit) : model contains fewer than 2 terms
```

إختبار عدم الخطية Nonlinearity:

```
> # Evaluate Nonlinearity
> # component + residual plot
> crPlots(fit)
> # Ceres plots
> ceresPlots(fit)
Error in ceresPlot.lm(model, term, main = "", ...) :
  There are no covariates.
>
```



إختبار عدم إستقلال الأخطاء Non-independence of Errors:

```
> # Test for Autocorrelated Errors
> durbinWatsonTest(fit)
 lag Autocorrelation D-W Statistic p-value
   1    -0.008260341      1.831947   0.654
 Alternative hypothesis: rho != 0
>
```

مكتبة أو حزمة gvlma تقوم بتشخيص عام كالتالي:

```
> # Global test of model assumptions
> library(gvlma)
> gvmodel <- gvlma(fit)
> summary(gvmodel)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9028 -0.1878  0.1102  0.2540  0.7741

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04574    0.10211   0.448     0.66
x            0.97810    0.09870   9.910 1.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4485 on 18 degrees of freedom
Multiple R-squared: 0.8451,     Adjusted R-squared: 0.8365
F-statistic:  98.2 on 1 and 18 DF,  p-value: 1.027e-08
```

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = fit)

                     Value p-value                Decision
Global Stat        3.4018269  0.4930 Assumptions acceptable.
Skewness           1.4569383  0.2274 Assumptions acceptable.
Kurtosis           0.0008766  0.9764 Assumptions acceptable.
Link Function      0.0022666  0.9620 Assumptions acceptable.
Heteroscedasticity 1.9417454  0.1635 Assumptions acceptable.
>
```

إضافات للإنحدار المتعدد:

1- مقارنة النماذج

```
> # compare models
> fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)
> fit2 <- lm(y ~ x1 + x2)
> anova(fit1, fit2)
>
```

2- إختيار المتغيرات المستقلة (الإنحدار المتدرج):

```
> # Stepwise Regression
> library(MASS)
> fit <- lm(y~x1+x2+x3,data=mydata)
> step <- stepAIC(fit, direction="both")
> step$anova # display results
```

```
>
```

3- الإنحدار الجزئي

```
> # All Subsets Regression
> library(leaps)
> attach(mydata)
> leaps<-regsubsets(y~x1+x2+x3+x4,data=mydata,nbest=10)
> # view results
> summary(leaps)
> # plot a table of models showing variables in each model.
> # models are ordered by the selection statistic.
> plot(leaps,scale="r2")
> # plot statistic by subset size
> library(car)
> subsets(leaps, statistic="rsq")
```

الأهمية النسبية للمتغيرات المستقلة:

```
> # Calculate Relative Importance for Each Predictor
> library(relaimpo)
>
calc.relimp(fit,type=c("lmg","last","first","pratt"),rela=T
RUE)

> # Bootstrap Measures of Relative Importance (1000
samples)
> boot <- boot.relimp(fit, b = 1000, type = c("lmg","last",
"first", "pratt"), rank = TRUE,
  diff = TRUE, rela = TRUE)
> booteval.relimp(boot) # print result
> plot(booteval.relimp(boot,sort=TRUE)) # plot result
```

مثال بإستخدام R:

سوف نطبق نموذج الإنحدار الخطي البسيط:

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

على البيانات

$$\mathbf{Y} = \begin{pmatrix} 136 \\ 144 \\ 145 \\ 169 \\ 176 \\ 195 \\ 211 \\ 224 \\ 231 \\ 256 \\ 281 \\ 312 \\ 347 \\ 377 \\ 423 \\ 477 \\ 553 \\ 613 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 91 \\ 1 & 105 \\ 1 & 109 \\ 1 & 130 \\ 1 & 146 \\ 1 & 155 \\ 1 & 160 \\ 1 & 180 \\ 1 & 200 \\ 1 & 215 \\ 1 & 240 \\ 1 & 275 \\ 1 & 320 \\ 1 & 360 \\ 1 & 410 \\ 1 & 460 \\ 1 & 510 \\ 1 & 575 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{18} \end{pmatrix}$$

تقدر $\boldsymbol{\beta}$ من العلاقة (أنظر الورقة النظرية)

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

أدخل التالي في R

```
X = matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
91,105,109,130,146,155,160,180,200,215,240,275,320,360,410,
460,510,575),byrow=F,18,2)

Y = matrix(c(136,144,145,169,176,195,211,224,231,256,281,
312,347,377,423,477,553,613),byrow=F,18,1)

XX = solve(t(X)%*%X)
XY = t(X)%*%Y
beta = XX%*%XY
options(digits=4)
beta
        [,1]
[1,] 42.8931
[2,]  0.9692
```

أي

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 42.8931 \\ 0.9692 \end{pmatrix}$$

والنموذج المقدر

$$\hat{y}_i = 42.8931 + 0.9692 x_i \ , i = 1,2,...,18$$

تقدير الخطأ المعياري للبواقي:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i}{n-k}}$$

حيث $n = 18$ (حجم العينة) و $k = 2$ (عدد المعالم المقدرة) و

$\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1,2,...,18$

هي البواقي أو مقدرات للخطأ.

نحسب $\hat{y}_i, i = 1,2,...,18$ من

```
Ye = beta[1]*X[,1]+beta[2]*X[,2]
erro = Y-Ye
```

مجموع مربعات الأخطاء $\sum_{i=1}^{18} \hat{\varepsilon}_i^2$ يحسب من

```
sum(erro^2)
[1] 1523
```

ويكون الخطأ المعياري للبواقي

```
n = length(Y)
k = 1
sigma2 = sum(erro^2)/(n-k)
sigma = sqrt(sigma2)
sigma
[1] 9.465
```

إذا الخطأ المعياري للبواقي $\hat{\sigma} = 9.465$ بدرجات حرية 16

الخطأ المعياري لمقدرات المعالم:

الاخطاء المعيارية هي عناصر القطر لمصفوفة التباين - التغاير

$$V\left(\hat{\boldsymbol{\beta}}\right) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

```
var.beta = sigma2*XX
var.beta
         [,1]       [,2]
[1,] 20.78621 -0.0613161
[2,] -0.06132  0.0002378
```

قطر المصفوفة

```
diag(var.beta)
2.079e+01 2.378e-04
```

والأخطاء المعيارية للمعالم

```
sqrt(diag(var.beta))
[1] 4.55919 0.01542
```

ونضع النموذج على الشكل

$$\hat{y}_i = \underset{(4.559)}{42.8931} + \underset{(0.0154)}{0.9692}\, x_i \ , i = 1,2,...,18$$

معامل التحديد:

معامل التحديد R^2 غير المعدل يعطى بالعلاقة

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

ويحسب

```
R2 <- 1-sum(erro^2)/sum((Y-mean(Y))^2)
R2
[1] 0.9957
```

لحساب معامل التحديد المعدل R_a^2

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

```
R22 <- 1-(1-R2)*(18-1)/(18-2)
R22
[1] 0.9954
```

إختبار فرضيات:

سوف نختبر هل هناك علاقة بين المتغيرات المستقلة والمتغير التابع كالآتي:

$$H_0 : \beta_1 = 0$$

وبوضعها في شكل مصفوفة

$$H_0 : \mathbf{RB} = \mathbf{r}$$

حيث:

$$\mathbf{R} = [0 \quad 1], \mathbf{B} = [\beta_1], \mathbf{r} = [0]$$

ونحسب النسبة F

$$F = \frac{\left(\mathbf{R}\hat{\beta} - \mathbf{r}\right)' \left|\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}\right|^{-1} \left(\mathbf{R}\hat{\beta} - \mathbf{r}\right)/q}{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k}}$$

حيث q هو عدد عدد القيود على الفرضية الصفرية ويساوي هنا 1

```
n=18
k=2
R = matrix(c(0,1),1,2,byrow=T)
r = matrix(c(0),1,byrow=T)
q = 1
(t(R%*%beta-r)%*%solve(R%*%XX%*%t(R))%*%(R%*%beta-
r)/q)/(sum(erro^2)/(n-k))
     [,1]
[1,] 3717
>
qf(0.95,1,16)
[1] 4.494
```

وحيث أن $3717 \in [4.494, \infty)$ فإننا نرفض الفرضية الصفرية.

إستخدام lm:

```
summary(lm(Y~X[,2]))

Call:
lm(formula = Y ~ X[, 2])

Residuals:
   Min     1Q Median     3Q    Max
-17.25  -5.95   1.00   5.36  15.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.8931     4.6995    9.13  9.6e-08 ***
X[, 2]        0.9692     0.0159   60.97  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 9.76 on 16 degrees of freedom
Multiple R-squared: 0.996,      Adjusted R-squared: 0.995
F-statistic: 3.72e+03 on 1 and 16 DF,  p-value: <2e-16

>
```

R FUNCTIONS FOR REGRESSION ANALYSIS

Linear model

Anova: Anova Tables for Linear and Generalized Linear Models (car)

anova: Compute an analysis of variance table for one or more linear model fits (stats)

coef: is a generic function which extracts model coefficients from objects returned by modeling functions. coefficients is an alias for it (stats)

coeftest: Testing Estimated Coefficients (lmtest)

confint: Computes confidence intervals for one or more parameters in a fitted model. Base has a method for objects inheriting from class "lm" (stats)

deviance:Returns the deviance of a fitted model object (stats)

effects: Returns (orthogonal) effects from a fitted model, usually a linear model. This is a generic function, but currently only has a methods for objects inheriting from classes "lm" and "glm" (stats)

fitted: is a generic function which extracts fitted values from objects returned by modeling functions fitted.values is an alias for it (stats)

formula: provide a way of extracting formulae which have been included in other objects (stats)

linear.hypothesis: Test Linear Hypothesis (car)

lm: is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (stats)

model.matrix: creates a design matrix (stats)

predict: Predicted values based on linear model object (stats)

residuals: is a generic function which extracts model residuals from objects returned by modeling functions (stats)

summary.lm: summary method for class "lm" (stats)

vcov: Returns the variance-covariance matrix of the main parameters of a fitted model object (stats)

Model – Variables selection

add1: Compute all the single terms in the scope argument that can be added to or dropped from the model, fit those models and compute a table of the changes in fit (stats)

AIC: Generic function calculating the Akaike information criterion for one or several fitted model objects for which a log-likelihood value can be obtained, according to the formula -2*log-likelihood + k*npar, where npar represents the number of parameters in the fitted model, and k = 2 for the usual AIC, or k = log(n) (n the number of observations) for the so-called BIC or SBC (Schwarz's Bayesian criterion) (stats)

Cpplot: Cp plot (faraway)

drop1: Compute all the single terms in the scope argument that can be added to or dropped from the model, fit those models and compute a table of the changes in fit (stats)

extractAIC: Computes the (generalized) Akaike An Information Criterion for a fitted parametric model (stats)

leaps: Subset selection by `leaps and bounds' (leaps)

maxadjr: Maximum Adjusted R-squared (faraway)

offset: An offset is a term to be added to a linear predictor, such as in a generalised linear model, with known coefficient 1 rather than an estimated coefficient (stats)

step: Select a formula-based model by AIC (stats)

update.formula: is used to update model formulae. This typically involves adding or dropping terms, but updates can be more general (stats)

Diagnostics

cookd: Cook's Distances for Linear and Generalized Linear Models (car)

cooks.distance: Cook's distance (stats)

covratio: covariance ratio (stats)

dfbeta: DBETA (stats)

dfbetas: DBETAS (stats)

dffits: DFFTITS (stats)

hat: diagonal elements of the hat matrix (stats)

hatvalues: diagonal elements of the hat matrix (stats)

influence.measures: This suite of functions can be used to compute some of the regression (leave-one-out deletion) diagnostics for linear and generalized linear models (stats)

lm.influence: This function provides the basic quantities which are used in forming a wide variety of diagnostics for checking the quality of regression fits (stats)

ls.diag: Computes basic statistics, including standard errors, t- and p-values for the regression coefficients (stats)

outlier.test: Bonferroni Outlier Test (car)

rstandard: standardized residuals (stats)

rstudent: studentized residuals (stats)

vif: Variance Inflation Factor (car)

Graphics

ceres.plots: Ceres Plots (car)

cr.plots: Component+Residual (Partial Residual) Plots (car)

influence.plot: Regression Influence Plot (car)

leverage.plots: Regression Leverage Plots (car)

panel.car: Panel Function Coplots (car)

plot.lm: Four plots (selectable by which) are currently provided: a plot of residuals against fitted values, a Scale-Location plot of sqrt{| residuals |} against fitted values, a Normal Q-Q plot, and a plot of Cook's distances versus row

labels (stats)

prplot: Partial Residual Plot (faraway)

qq.plot: Quantile-Comparison Plots (car)

qqline: adds a line to a normal quantile-quantile plot which passes through the first and third quartiles (stats)

qqnorm: is a generic function the default method of which produces a normal QQ plot of the values in y (stats)

reg.line: Plot Regression Line (car)

scatterplot.matrix: Scatterplot Matrices (car)

scatterplot: Scatterplots with Boxplots (car)

spread.level.plot: Spread-Level Plots (car)

Tests

ad.test: Anderson-Darling test for normality (nortest)

bartlett.test: Performs Bartlett's test of the null that the variances in each of the groups (samples) are the same (stats)

bgtest: Breusch-Godfrey Test (lmtest)

bptest: Breusch-Pagan Test (lmtest)

cvm.test: Cramer-von Mises test for normality (nortest)

durbin.watson: Durbin-Watson Test for Autocorrelated Errors (car)

dwtest: Durbin-Watson Test (lmtest)

levene.test: Levene's Test (car)

lillie.test: Lilliefors (Kolmogorov-Smirnov) test for normality (nortest)

ncv.test: Score Test for Non-Constant Error Variance (car)

pearson.test: Pearson chi-square test for normality (nortest)

sf.test: Shapiro-Francia test for normality (nortest)

shapiro.test: Performs the Shapiro-Wilk test of normality (stats)

Variables transformations

box.cox: Box-Cox Family of Transformations (car)

boxcox: Box-Cox Transformations for Linear Models (MASS)

box.cox.powers: Multivariate Unconditional Box-Cox Transformations (car)

box.tidwell: Box-Tidwell Transformations (car)

box.cox.var: Constructed Variable for Box-Cox Transformation (car)

Ridge regression

lm.ridge: Ridge Regression (MASS)

Segmented regression

segmented: Segmented relationships in regression models (segmented)

slope.segmented: Summary for slopes of segmented relationships (segmented)

Generalized Least Squares (GLS)

ACF.gls: Autocorrelation Function for gls Residuals (nlme)

anova.gls: Compare Likelihoods of Fitted Objects (nlme)

gls: Fit Linear Model Using Generalized Least Squares (nlme)

intervals.gls: Confidence Intervals on gls Parameters (nlme)

lm.gls: fit Linear Models by Generalized Least Squares (MASS)

plot.gls: Plot a gls Object (nlme)

predict.gls: Predictions from a gls Object (nlme)

qqnorm.gls: Normal Plot of Residuals from a gls Object (nlme)

residuals.gls: Extract gls Residuals (nlme)

summary.gls: Summarize a gls Object (nlme)

Generalized Linear Models (GLM)

family: Family objects provide a convenient way to specify the details of the models used by functions such as glm (stats)

glm.nb: fit a Negative Binomial Generalized Linear Model (MASS)

glm: is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution (stats)

polr: Proportional Odds Logistic Regression (MASS)

Non linear Least Squares (NLS)

nlm: This function carries out a minimization of the function f using a Newton-type algorithm (stats)

nls: Determine the nonlinear least-squares estimates of the nonlinear model parameters and return a class nls object (stats)

nlscontrol: Allow the user to set some characteristics of the nls nonlinear least squares algorithm (stats)

nlsModel: This is the constructor for nlsModel objects, which are function closures for several functions in a list. The closure includes a nonlinear model formula, data values for the formula, as well as parameters and their values (stats)

Generalized Non linear Least Squares (GNLS)

coef.gnls: Extract gnls Coefficients (nlme)

gnls: Fit Nonlinear Model Using Generalized Least Squares (nlme)

predict.gnls: Predictions from a gnls Object (nlme)

Loess regression

loess: Fit a polynomial surface determined by one or more numerical predictors, using local fitting (stats)

loess.control: Set control parameters for loess fits (stats)

predict.loess: Predictions from a loess fit, optionally with standard errors (stats)

scatter.smooth: Plot and add a smooth curve computed by loess to a scatter plot (stats)

Splines regression

bs: B-Spline Basis for Polynomial Splines (splines)

ns: Generate a Basis Matrix for Natural Cubic Splines (splines)

periodicSpline: Create a Periodic Interpolation Spline (splines)

polySpline: Piecewise Polynomial Spline Representation (splines)

predict.bSpline: Evaluate a Spline at New Values of x (splines)

predict.bs: Evaluate a Spline Basis (splines)

splineDesign: Design Matrix for B-splines (splines)

splineKnots: Knot Vector from a Spline (splines)

splineOrder: Determine the Order of a Spline (splines)

Robust regression

lqs: Resistant Regression (MASS)

rlm: Robust Fitting of Linear Models (MASS)

Structural equation models

sem: General Structural Equation Models (sem)

tsls: Two-Stage Least Squares (sem)

Simultaneous Equation Estimation

systemfit: Fits a set of linear structural equations using Ordinary Least Squares (OLS), Weighted Least Squares (WLS), Seemingly Unrelated Regression (SUR), Two-Stage Least Squares (2SLS), Weighted Two-Stage Least Squares (W2SLS) or Three-Stage Least Squares (3SLS) (systemfit)

Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)

biplot.mvr: Biplots of PLSR and PCR Models (pls)

coefplot: Plot Regression Coefficients of PLSR and PCR models (pls)

crossval: Cross-validation of PLSR and PCR models (pls)

cvsegments: Generate segments for cross-validation (pls)

kernelpls.fit: Kernel PLS (Dayal and MacGregor) (pls)

msc: Multiplicative Scatter Correction (pls)

mvr: Partial Least Squares and Principal Components Regression (pls)

mvrCv: Cross-validation (pls)

oscorespls.fit: Orthogonal scores PLSR (pls)

predplot: Prediction Plots (pls)

scoreplot: Plots of Scores and Loadings (pls)

scores: Extract Scores and Loadings from PLSR and PCR Models (pls)

svdpc.fit: Principal Components Regression (pls)

validationplot: Validation Plots (pls)

Quantile regression

anova.rq: Anova function for quantile regression fits (quantreg)

boot.rq: Bootstrapping Quantile Regression(quantreg)

lprq: locally polynomial quantile regression (quantreg)

nlrq: Function to compute nonlinear quantile regression estimates (quantreg)

qss: Additive Nonparametric Terms for rqss Fitting (quantreg)

ranks: Quantile Regression Ranks (quantreg)

rq: Quantile Regression (quantreg)

rqss: Additive Quantile Regression Smoothing (quantreg)

rrs.test: Quantile Regression Rankscore Test (quantreg)

standardize: Function to standardize the quantile regression process (quantreg)

Linear and nonlinear mixed effects models

ACF: Autocorrelation Function (nlme)

ACF.lme: Autocorrelation Function for lme Residuals (nlme)

anova.lme: compare Likelihoods of Fitted Objects (nlme)

fitted.lme: Extract lme Fitted Values (nlme)

fixed.effects: Extract Fixed Effects (nlme)

intervals: Confidence Intervals on Coefficients (nlme)

intervals.lme: Confidence Intervals on lme Parameters (nlme)

lme: Linear Mixed-Effects Models (nlme)

nlme: Nonlinear Mixed-Effects Models (nlme)

predict.lme: Predictions from an lme Object (nlme)

predict.nlme: Predictions from an nlme Obj (nlme)

qqnorm.lme: Normal Plot of Residuals or Random Effects from an lme object (nlme)

random.effects: Extract Random Effects (nlme)

ranef.lme: Extract lme Random Effects (nlme)

residuals.lme: Extract lme Residuals (nlme)

simulate.lme: simulate lme models (nlme)

summary.lme: Summarize an lme Object (nlme)

glmmPQL: fit Generalized Linear Mixed Models via PQL (MASS)

Generalized Additive Model (GAM)

anova.gam: compare the fits of a number of gam models (gam)

gam.control: control parameters for fitting gam models (gam)

gam: Fit a generalized additive model (gam)

na.gam.replace: a missing value method that is helpful with gams (gam)

plot.gam: an interactive plotting function for gams (gam)

predict.gam: make predictions from a gam object (gam)

preplot.gam: extracts the components from a gam in a plot-ready form (gam)

step.gam: stepwise model search with gam (gam)

summary.gam: summary method for gam (gam)

Survival analysis

anova.survreg: ANOVA tables for survreg objects (survival)

clogit: Conditional logistic regression (survival)

cox.zph: Test the proportional hazards assumption of a Cox regression (survival)

coxph: Proportional Hazards Regression (survival)

coxph.detail: Details of a cox model fit (survival)

coxph.rvar: Robust variance for a Cox model (survival)

ridge: ridge regression (survival)

survdiff: Test Survival Curve Differences (survival)

survexp: Compute Expected Survival (survival)

survfit: Compute a survival Curve for Censored Data (survival)

survreg: Regression for a parametric survival model (survival)

Classification and Regression Trees

cv.tree: Cross-validation for Choosing tree Complexity (tree)

deviance.tree: Extract Deviance from a tree Object (tree)

labels.rpart: Create Split Labels For an rpart Object (rpart)

meanvar.rpart: Mean-Variance Plot for an rpart Object (rpart)

misclass.tree: Misclassifications by a Classification tree (tree)

na.rpart: Handles Missing Values in an rpart Object (rpart)

partition.tree: Plot the Partitions of a simple Tree Model (tree)

path.rpart: Follow Paths to Selected Nodes of an rpart Object (rpart)

plotcp: Plot a Complexity Parameter Table for an rpart Fit (rpart)

printcp: Displays CP table for Fitted rpart Object (rpart)

prune.misclass: Cost-complexity Pruning of Tree by error rate (tree)

prune.rpart: Cost-complexity Pruning of an rpart Object (rpart)

prune.tree: Cost-complexity Pruning of tree Object (tree)

rpart: Recursive Partitioning and Regression Trees (rpart)

rpconvert: Update an rpart object (rpart)

rsq.rpart: Plots the Approximate R-Square for the Different Splits (rpart)

snip.rpart: Snip Subtrees of an rpart Object (rpart)

solder: Soldering of Components on Printed-Circuit Boards (rpart)

text.tree: Annotate a Tree Plot (tree)

tile.tree: Add Class Barplots to a Classification Tree Plot (tree)

tree.control: Select Parameters for Tree (tree)

tree.screens: Split Screen for Plotting Trees (tree)

tree: Fit a Classification or Regression Tree (tree)

Beta regression

betareg: Fitting beta regression models (betareg)

plot.betareg: Plot Diagnostics for a betareg Object (betareg)

predict.betareg: Predicted values from beta regression model (betareg)

residuals.betareg: Residuals function for beta regression models (betareg)

summary.betareg: Summary method for Beta Regression (betareg)

Packages used:

car: Companion to Applied Regression

stasts: R statistical functions

lmtest: Testing Linear Regression Models

faraway: Functions and datasets for books by Julian Faraway

leaps: regression subset selection

nortest: Tests for Normality

MASS: Support Functions and Datasets for Venables and Ripley's MASS

segmented: Segmented relationships in regression models with breakpoints/changepoints estimation

nlme: Linear and Nonlinear Mixed Effects Models

splines: Interpolating Splines

sem: Structural Equation Models

systemfit: Estimating Systems of Simultaneous Equations

pls: Partial Least Squares and Principal Component regression

quantreg: Quantile Regression

gam: Generalized Additive Models

survival: Survival analysis, including penalised likelihood

tree: Classification and regression trees

rpart: Recursive Partitioning

betareg: Beta Regression