北京师范大学
香港浸会大学
联合国际学院
BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY
UNITED INTERNATIONAL COLLEGE

STAT2013 Regression Analysis

Semester 1, 2018-2019

Room: T3-401-R1

Instructor: Ye, Huajun Terry

TA：Jenna Otto

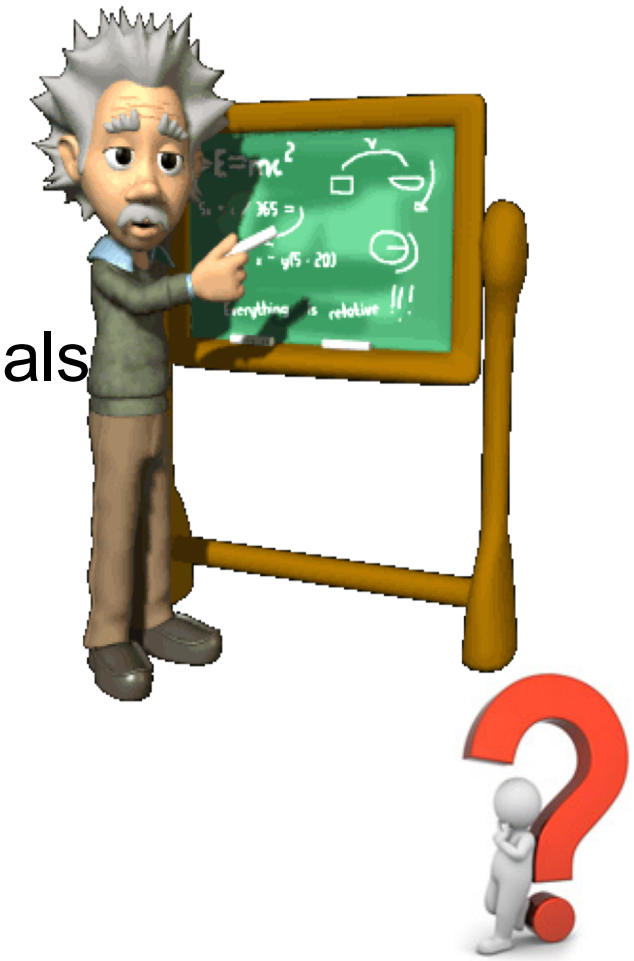
Email:hjye@uic.edu.hk

Jotto@uic.edu.hk (TA)

Chapter 5
Statistical Diagnostics

In this Chapter, we will know some important concepts.

- Introduction
- The Residuals
- Plotting of Residues
- Relation to Standardized PRESS Residuals
- Detection of Outliers
- Diagnostic Plots
- Influence of Cases



5.1 Introduction

So far we have discussed the regression models under a certain assumption (normality, constant variance etc). All of the methods that we have introduced are computed as if the model and the assumptions are correct, but in any practical problem assumptions are in doubt. Therefore, we need statistical diagnostics as follows:

- Is the model correct?
- Are there any outliers?
- Is the variance constant?
- Is the error normally distributed?

Please refer Chapters 5 and 6 of the textbook for the details.

Example 5.1.

Table 5.1 Four Hypothetical Data Sets (Anscombe data)

	Data Set Number					
	1-3	1	2	3	4	4
	Variable					
Case Number	X	Y	Y	Y	X	Y
1	10.0	8.04	9.14	7.46	8.0	6.58
2	8.0	6.95	8.14	6.77	8.0	5.76
3	13.0	7.58	8.74	12.74	8.0	7.71
4	9.0	8.81	8.77	7.11	8.0	8.84
5	11.0	8.33	9.26	7.81	8.0	8.47
6	14.0	9.96	8.10	8.84	8.0	7.04
7	6.0	7.24	6.13	6.08	8.0	5.25
8	4.0	4.26	3.10	5.39	19.0	12.50
9	12.0	10.84	9.13	8.15	8.0	5.56
10	7.0	4.82	7.26	6.42	8.0	7.91
11	5.0	5.68	4.74	5.73	8.0	6.89

Fit these 4 sets by

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Example 5.1. (*cont*)

Fit these 4 sets by

$$y = \beta_0 + \beta_1 x + \varepsilon$$

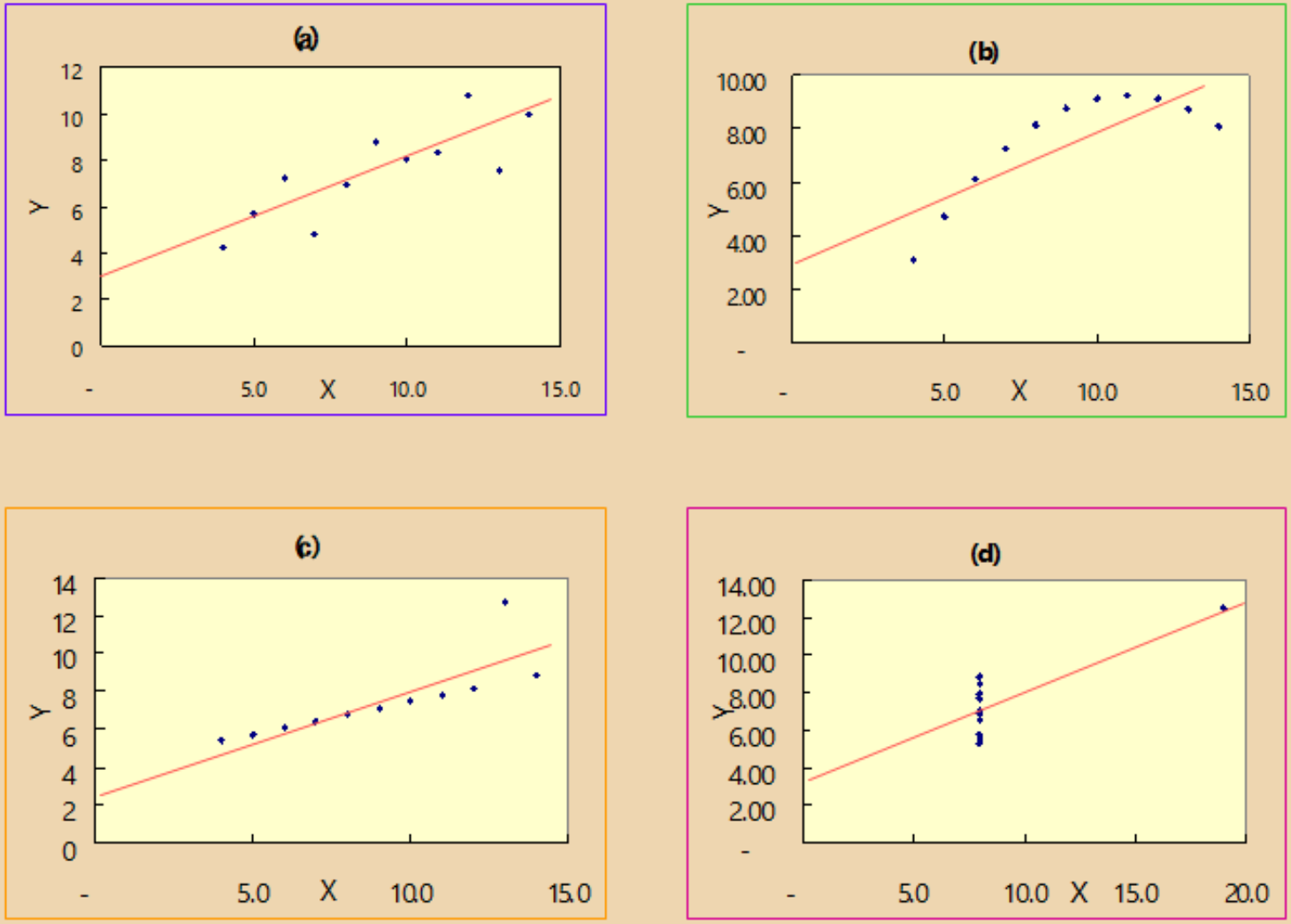
implies the regression equation:

$$\hat{y} = 3 + 0.5x$$

with $R^2 = 0.667$ and $\hat{\sigma}^2 = 13.75$.

Let's plot the data.

Figure 5.1 Four hypothetical data sets.



Example 5.1. (*cont*)

We might conclude that the linear regression model is equally appropriate for each of them. But they are wrong!

From the plots, we find:

(a) The model is appropriate;

(b) The model is not linear;

(c) There is an ***outlier***. If we delete it, the regression line becomes

$$\hat{y} = 4 + 0.346x$$

(d) The slope parameter estimate $\hat{\beta}_1$ is largely determined by the value of y_8. If x_8 were deleted, we could not even estimate β_1. We must distrust an aggregate analysis that is so heavily dependent upon a single case. The point (x_8, y_8) is called ***strongly influence point***.

5.2 The Residuals

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

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

The LSE of $\boldsymbol{\varepsilon}$ is

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

The $\boldsymbol{e}$ is called the vector of residuals.

Some discussion on residuals on a simply linear model can refer to the textbook, Section 2.12, pp 57-66.

Note that:

$$\mathrm{E}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\,\mathrm{E}(\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

$$\mathrm{Cov}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\,\mathrm{Cov}(\boldsymbol{y})\,(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2\,(\boldsymbol{I} - \boldsymbol{H})$$

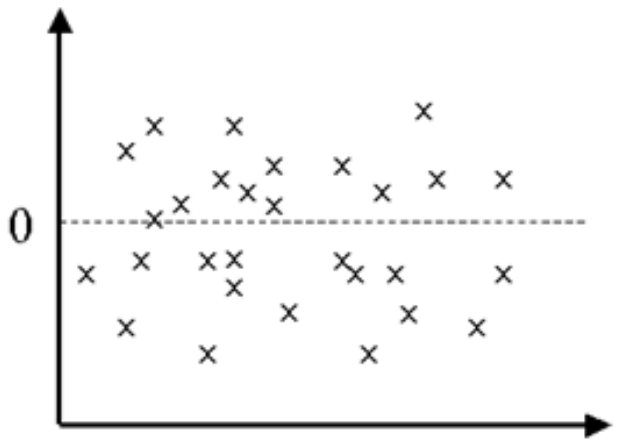
where $\boldsymbol{H} = (h_{ij})$

$$= \sigma^2 \begin{pmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ -h_{21} & 1-h_{22} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \cdots & 1-h_{nn} \end{pmatrix}$$

We can define the *studentized residuals*:

$$\boxed{r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}}$$

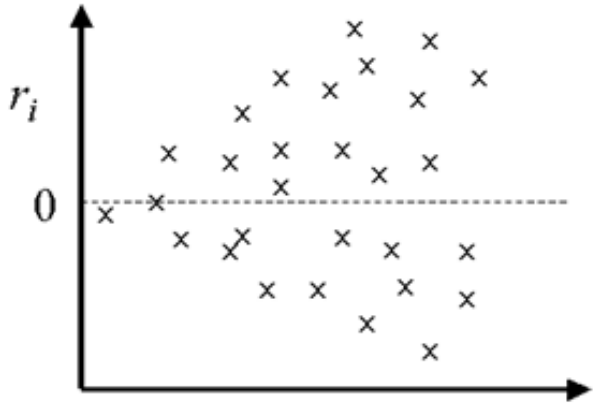
5.3 Plotting of Residues



(a) $\hat{y}_i$ or x_i

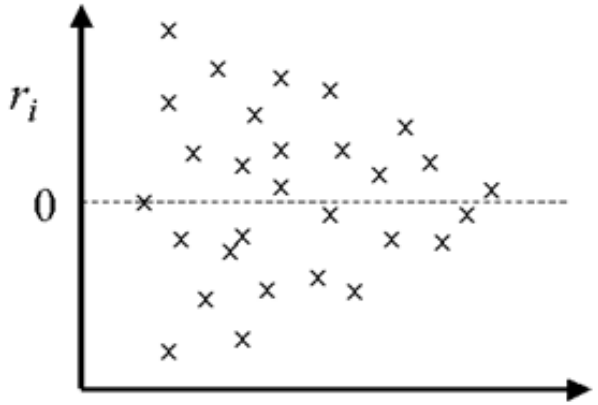
A null plot:

(normal one)



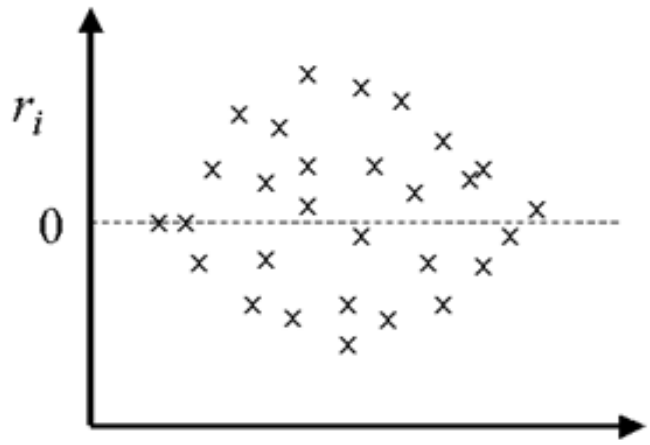
(b) $\hat{y}_i$ or x_i

Right-opening megaphone:
It suggests variance increasing with the quality plotted on the x-axis. This will often occur if an intrinsically positive response varies over a wide range, say from near zero into the thousands.



(c) $\hat{y}_i$ or x_i

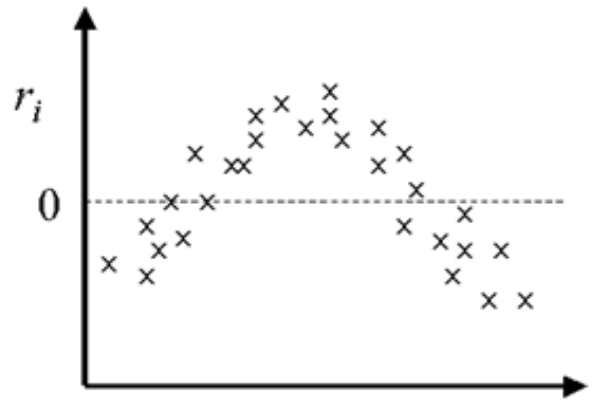
Left-opening megaphone:
It suggests variance decreasing with the quality plotted on the x-axis.



(d) $\hat{y}_i$ or x_i

Double outward bow:

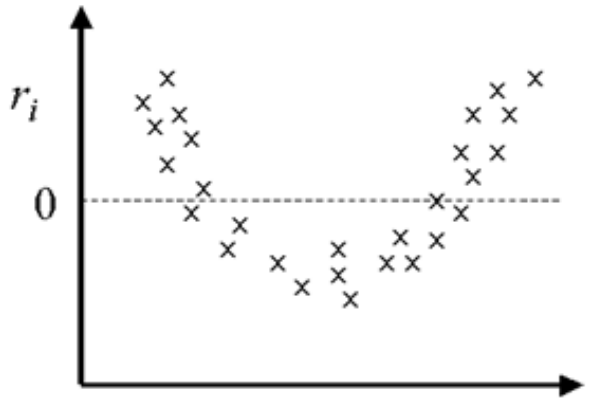
It can occur if the response is constrained to lie between a minimum and a maximum value, for example, a percentage between 0 and 100. Large and small percentages are often less variable than the percentages near 50%.



(e) $\hat{y}_i$

Non-linearity:

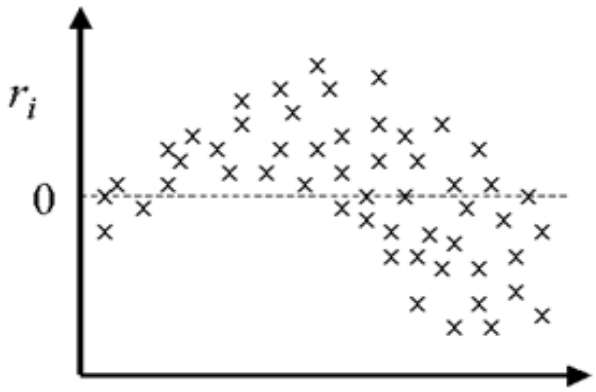
This will often call for transformation of the data, either the response or the predictors, or use of nonlinear models.



(f) $\hat{y}_i$

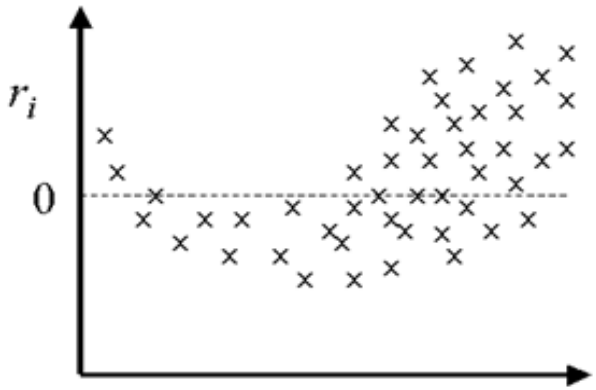
Non-linearity

This will often call for transformation of the data, either the response or the predictors, or use of nonlinear models



(g) $\hat{y}_i$

Non-linearity and
Non-constant variance:



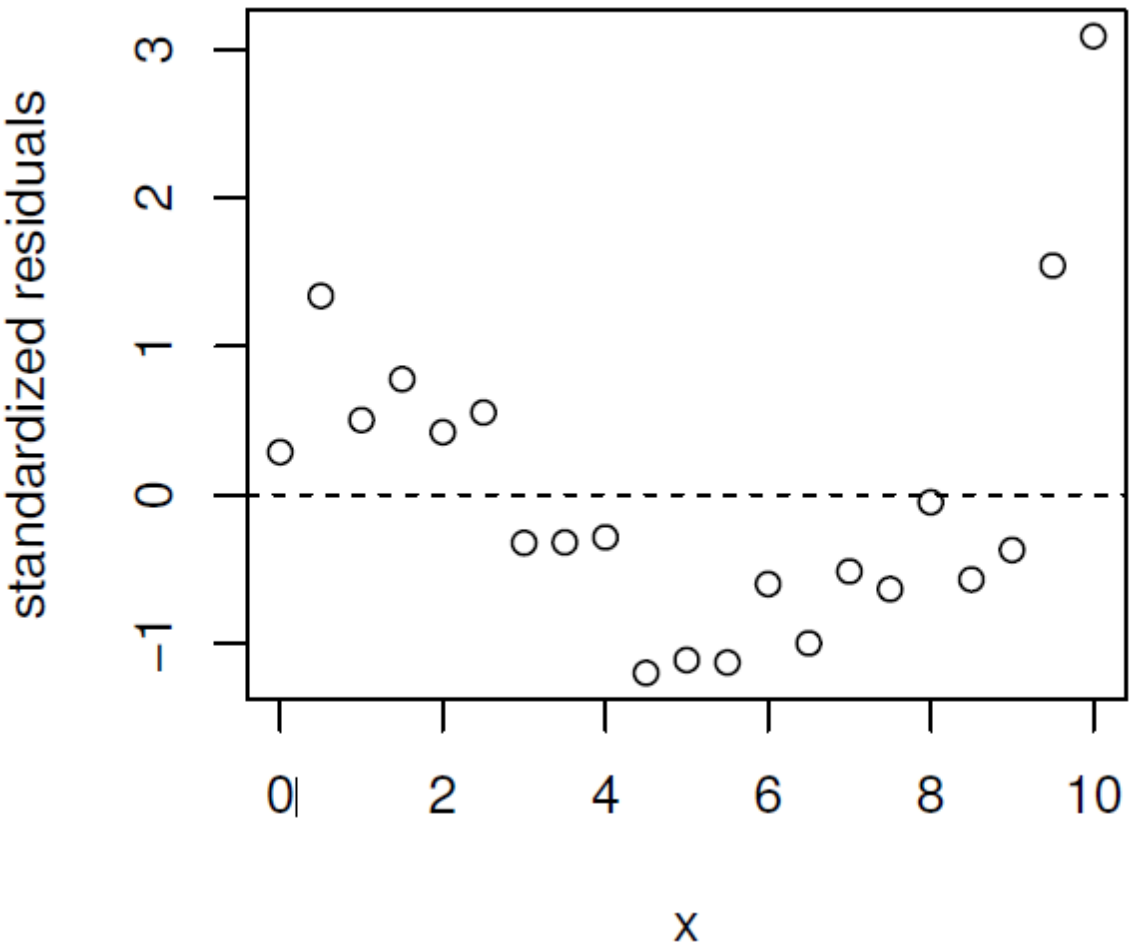
(h) $\hat{y}_i$

Non-linearity and
Non-constant variance

Linearizable Regression Functions with Corresponding Transformations.

Curve	Function	Transformation	Linear Model
Power	$Y = \alpha X^{\beta}$	$Y' = \ln(Y),\ X' = \ln(X)$	$Y' = \ln(\alpha) + \beta X'$
Exponential	$Y = \alpha e^{\beta X}$	$Y' = \ln(Y)$	$Y' = \ln(\alpha) + \beta X$
Inverse Exponential	$Y = \alpha e^{\beta / X}$	$Y' = \ln(Y),\ X' = \frac{1}{X}$	$Y' = \ln(\alpha) + \beta X'$
	$Y = \alpha + \beta \ln(X)$	$X' = \ln(X)$	$Y = \alpha + \beta X'$
Hyperbola	$Y = \frac{X}{\alpha X + \beta}$	$Y' = \frac{1}{Y},\ X' = \frac{1}{X}$	$Y' = \alpha + \beta X'$
Logistic	$Y = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}$	$Y' = \frac{Y}{1 - Y}$	$Y' = \alpha + \beta X$

Figure 5.2: Residual plot of exponentially-related data. A pattern exists, suggesting the data are not linearly related.



As with the second simulated example, sometimes finding a solution for one problem (i.e. nonlinearity) will result in a violation of another assumption (i.e. non-constant variance of errors). This is usually due to the fact that transformations are done while ignoring the original errors.

For example, take the power function

$$Y = \alpha X^{\beta} + \epsilon.$$

We assume the errors have mean 0, thus,

$$E[Y] = \alpha X^{\beta}. \qquad Y = \alpha X^{\beta} + \epsilon = \alpha X^{\beta}\left(1 + \frac{\epsilon}{E[Y]}\right).$$

$$
\begin{aligned}
Y' = \ln(Y) &= \ln(\alpha X^{\beta}) + \ln\left(1 + \frac{\epsilon}{E[Y]}\right) \\
&= \ln(\alpha) + \beta \ln(X) + \ln\left(1 + \frac{\epsilon}{E[Y]}\right) \\
&= \ln(\alpha) + \beta X' + \epsilon^{*}.
\end{aligned}
$$

$$
\begin{aligned}
E\left[1 + \frac{\epsilon}{E[Y]}\right] &= 1 + \frac{E[\epsilon]}{E[Y]} = 1 \quad (b/c\ E[\epsilon] = 0) \\
Var\left[1 + \frac{\epsilon}{E[Y]}\right] &= \frac{Var[\epsilon]}{E[Y]^2} = \frac{1}{(\alpha X^{\beta})^2}\sigma^2.
\end{aligned}
$$

So, the variance of the new residuals is a function of the data. As the X values increase, the variance will probably decrease.

Non-constant residual variance violates the assumptions of the linear regression model. When non-constant variance is diagnosed, but exact variances are unknown we could contemplate two remedies.

(a) The weighted least squares could be used.

$$Q = \sum_{i=1}^{n} (y_i - \beta x_i)^2 w_i \rightarrow \min$$

When non-constant variance is diagnosed, but exact variances are unknown

We could contemplate two remedies.

(b) Using a variance stabilizing transformation.

Transformation	Situation	Comments
$\sqrt{Y}$	$\mathrm{var}(e_i) \propto \mathrm{E}(Y_i)$	The theoretical basis is for counts from the Poisson distribution
$\sqrt{Y} + \sqrt{Y+1}$	As above	For use when some Y_i's are zero or very small; this is called the Freeman-Tukey (1950) transformation
$\log Y$	$\mathrm{var}(e_i) \propto [\mathrm{E}(Y_i)]^2$	This transformation is very common; it is a good candidate if the range of Y is very broad, say from to several thousand; all Y_i must be strictly positive
$\log(Y+1)$	As above	Used if $Y_i = 0$ for some cases
$1/Y$	$\mathrm{var}(e_i) \propto [\mathrm{E}(Y_i)]^4$	Appropriate when response are "bunched" near zero, but, in markedly decreasing numbers, large response do occur; e.g., if the response is a latency or response time for a treatment or a drug, some subjects may respond quickly while a few take much longer; the reciprocal transformation changes the scale of time per response to the rate of response, response per unit time; all Y_i must be positive
$1/(Y+1)$	As above	Used if $Y_i = 0$ for some cases
$\sin^{-1}(\sqrt{Y})$	$\mathrm{var}(e_i) \propto \mathrm{E}(Y_i)(1-\mathrm{E}(Y_i))$	For binomial proportions ($0 \le Y_i \le 1$)

- Using a variance stabilizing transformation.
- Another idea is to make a diagnostic plot of $log_{10} | e_i |$ against $log_{10} (x_i)$. If the points in this plot are approximately linear with slope b , it can be shown that a power transformation, $y \text{-->} y^{(1-b)}$, where y^0 is interpreted as $log\ (y)$, will make the residual variance more nearly equal. A famous power transformation is proposed by Cox-Box (1964)

Box-Cox transformation

- The Box-Cox transformation is defined as

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

- The transformation requires that $y_i > 0$, i=1, 2, 3,……, n .

For a studentized residual plot, the points should be in $\pm t_{0.005,n-p}$ as the upper / lower limits.

By the technique in the quality control, we can conclude the model/data is not normal if any one or more of the criteria is met:

(a) One or more points outside of the control limits.

(b) A run of at least 8 points, where the type of run could be either a run up or down, a run above or below the center line.

(c) An unusual or non-random pattern in the plot.

Example 5.3. Population Data

The following data give the population (in millions) of the U.S. from the year 1790 through the year 1970.

Year	y (Population)	Year	y (Population)
1790	3.929	1890	62.947
1800	5.308	1900	75.994
1810	7.239	1910	91.972
1820	9.638	1920	105.710
1830	12.866	1930	122.775
1840	17.069	1940	131.669
1850	23.191	1950	151.325
1860	31.443	1960	179.323
1870	39.818	1970	203.211
1880	50.155		

Consider a simple linear regression fit to the 19 data points with a model of the type $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, 2, \ldots, 19)$

Example 5.3. Population Data

The least squares fit produces the regression

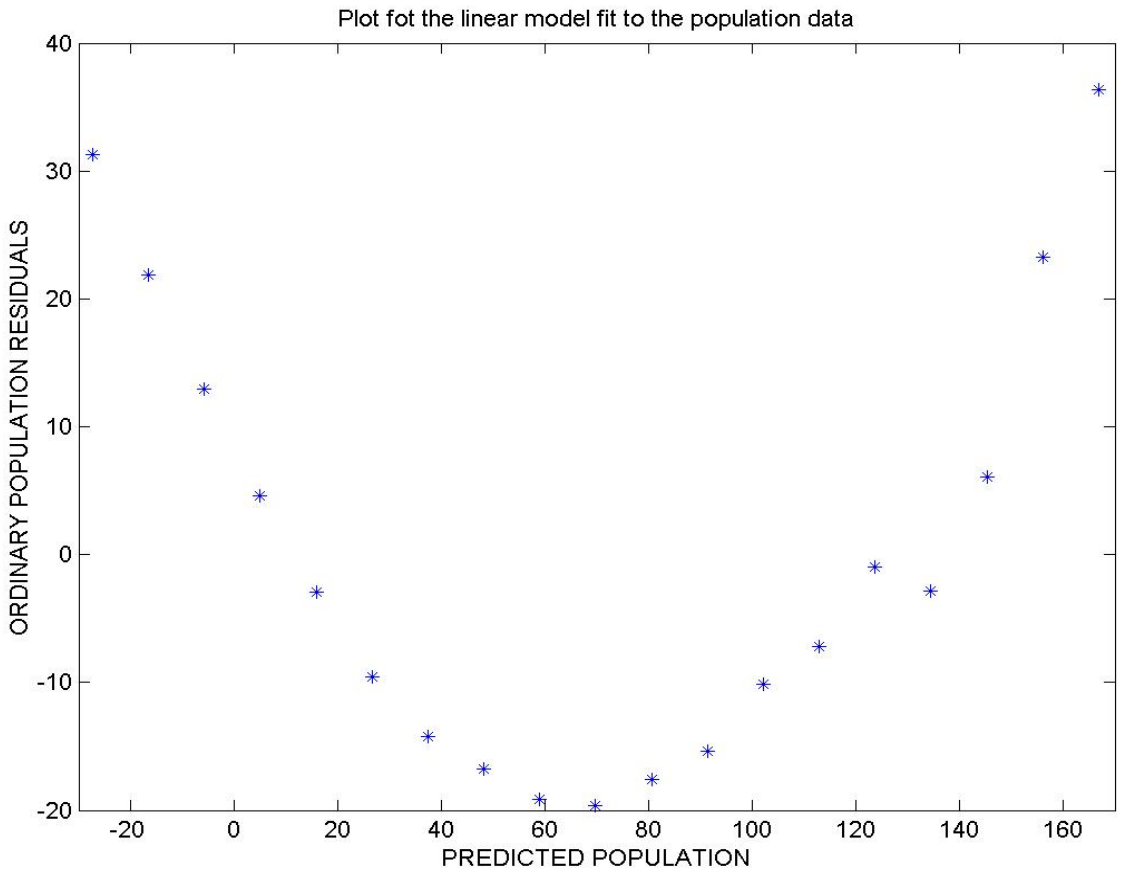
$$\hat{y} = -1958.366 + 1.078795x$$

with $R^2 = 0.9223$ and $s = 18.1274$. The listing below shows the fitted population $\hat{y}$ and residuals.

Year	$\hat{y}$ (Fitted Population)	$e_i = y_i - \hat{y}_i$	Year	(Fitted Population)	$e_i = y_i - \hat{y}_i$
1790	-27.324	31.253	1890	80.555	-17.608
1800	-16.536	21.844	1900	91.343	-15.349
1810	-5.748	12.987	1910	102.131	-10.159
1820	5.040	4.598	1920	112.919	-7.209
1830	15.828	-2.962	1930	123.707	-0.932
1840	26.616	-9.547	1940	134.495	-2.826
1850	37.404	-14.213	1950	145.283	6.042
1860	48.192	-16.749	1960	156.071	23.252
1870	58.980	-19.162	1970	166.859	36.352
1880	69.767	-19.612			

Example 5.3. Population Data

The following plot of the residuals against the fitted population. It is clear from the plot that there are grave doubts the model specification.



Example 5.3. Population Data

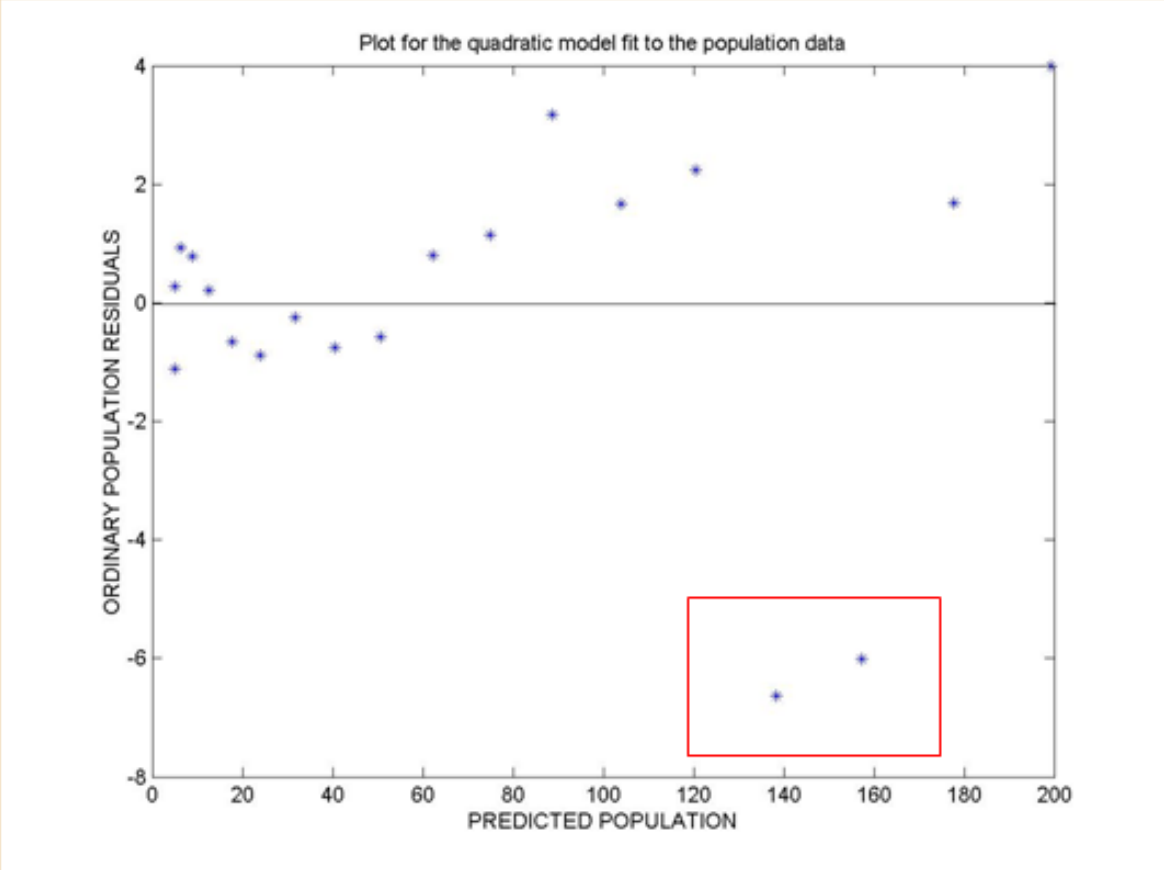
One would expect that the introduction of a quadratic term would produce a residual plot that would be closer to the "ideal" appearance. A quadratic regression model fit to the data results in a fitted model given by

$$\hat{y} = 20450.434 - 22.7806x + 0.0063456x^2$$

with $R^2 = 0.9983$ and $s = 2.781$.

Example 5.3. Population Data

The following plot shows the residuals against the fitted population for the quadratic model.



Example 5.3. Population Data

✓ Two particular residuals appear considerably below the others. These two residuals represent years 1940 and 1950. We could conclude that the picture still depicts a possible model misspecification.

✓ There is an obvious explanation for the phenomenon seen here. The population during the period 1940 through 1950 was influenced by the war years, suggesting a need for a model shift, or a categorical variable.

Example 5.3. Population Data

This was attempted by assigning a 1 to the categorical variable for the years 1940 to 1950 and zero to the other years.

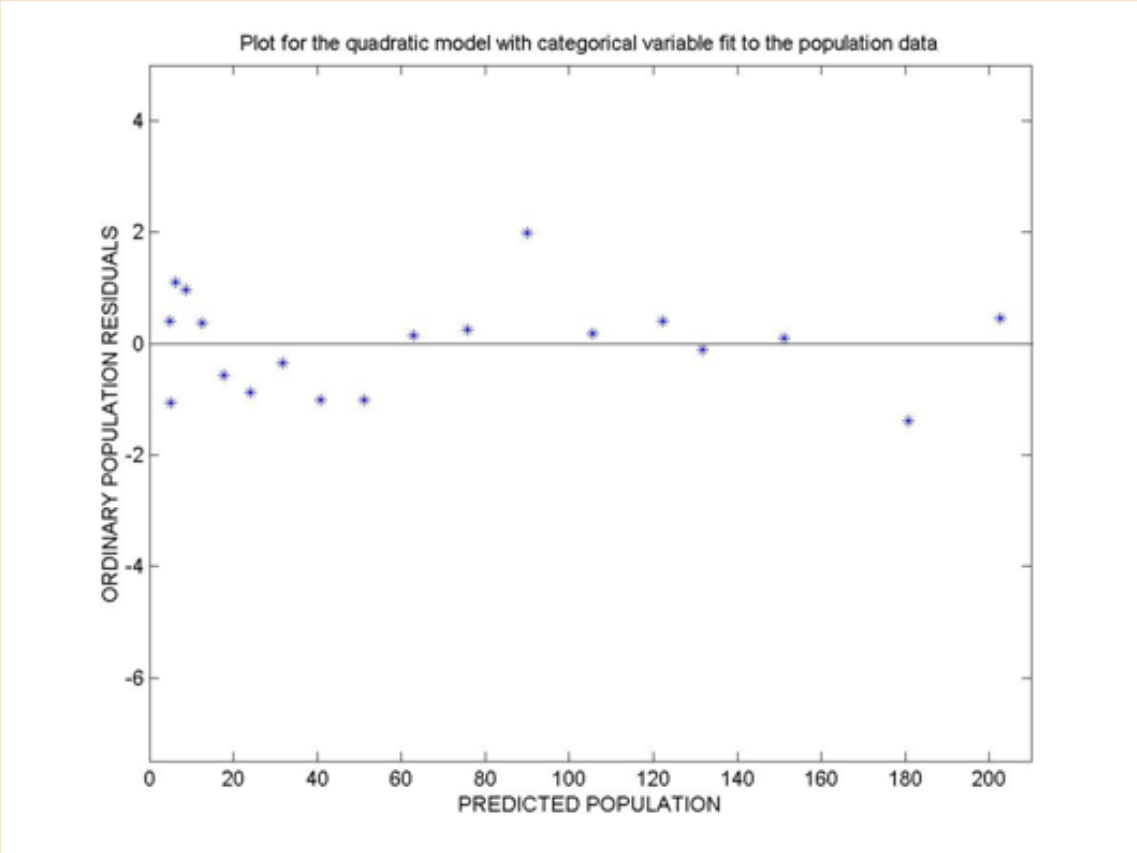
The resulting regression equation is given by

$$\hat{y} = 20982.754 - 2336638\,x + 0.006507\,x^2 - 8.7415\,z$$

where z is 0 or 1 depending on the year. The R^2 and root residual mean square are given by $R^2 = 0.9998$ and $s = 0.9374$, an appreciable improvement over the previous model.

Example 5.3. Population Data

The following plot shows the residuals against the fitted population.



5.4 Relation to Standardized PRESS Residuals

The ith *PRESS* residual is

$$e_{i,-i} = \frac{e_i}{1-h_{ii}}$$

and the variance of $e_{i,-i}$ is

$$\begin{aligned}\mathrm{Var}(e_{i,-i}) &= \frac{1}{(1-h_{ii})^2}\left[\sigma^2(1-h_{ii})\right] \\ &= \frac{\sigma^2}{1-h_{ii}}.\end{aligned}$$

Thus we can standardize the *PRESS* residual and obtain

$$\begin{aligned}\frac{e_{i,-i}}{\sigma_{i,-i}} &= \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} \\ &= \frac{e_i}{\sigma\sqrt{1-h_{ii}}}\end{aligned}$$

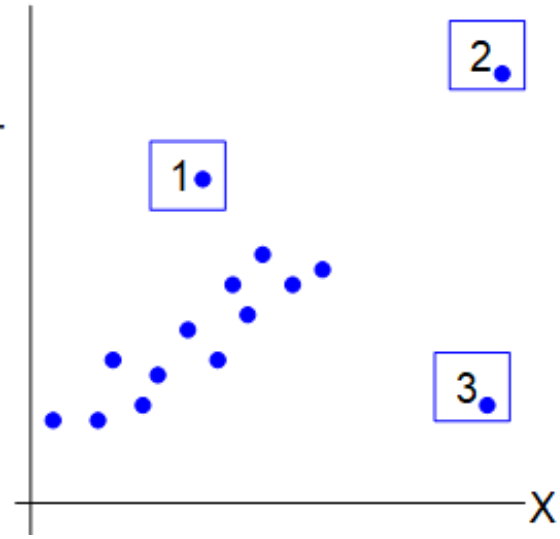
5.5 Detection of Outliers

Outliers and influential points are both unusual observations.

An outlier usually lies far from the fitted regression line and produces a large residual.

An influential point is often separated from the other observations in the x-direction. It does not need to produce a large residual. But removing it would markedly change the regression line.

1. Outliers are not necessarily influential.
2. Observations separated in the x-direction are not necessarily influential.
3. Influential observations are not necessarily outliers.



STATISTICAL DIAG

5.5 Detection of Outliers

One important assumption made in regression analysis is that the model used is appropriate for all of the data. Cases that do not follow the same model as the rest of the data are called **outliers**, and are important function of case analysis is to identify such cases.

The following are certainly possibilities:

(i) Mean Shift Outlier:

There is a breakdown in the model at the ith point, producing a "location shift", i.e. $E(\varepsilon_i) = \Delta_i \neq 0$.

(ii) Variance Shift Outlier:

There is a breakdown in the model at the ith point, and $Var(\varepsilon_i)$ exceeds the error variance at the other data locations.

In this section we consider mean shift outlier only.

A mean shift model:

$$y_j = x'_j\beta + \varepsilon_j, \quad j \neq i$$

$$y_i = x'_i\beta + \Delta_i + \varepsilon_i$$

An Outlier Test:

Suppose the ith case is suspected to be an outlier.

(i) Consider the model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \Delta_i \mathbf{u} + \boldsymbol{\varepsilon} \qquad u_j = 0 \text{ for } j \neq i, \ \& \ u_i = 1 \text{ otherwise}$$

The estimated coefficient for $\mathbf{u}$ is the estimate of the mean shift Δ_i. The t-statistic for testing $H_0: \Delta_i = 0$ VS $H_1: \Delta_i \neq 0$ is the appropriate test statistic.

(ii) Consider the *PRESS* residual

$$e_{i,-i} = y_i - \hat{y}_{i,-i}$$

$$\mathrm{E}(e_{i,-i}) = \Delta_i$$

We need to standardize the *PRESS* residual to determine if the deviation of $e_{i,-i}$ from 0 is more than mere chance.

$$e_{i,-i} / \sigma_{i,-i} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

Here $\sigma_{i,-i}$ should be estimated by

$$s_{-i}^2 = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}$$

Put this estimate to replace $\sigma_{i,-i}$, we have a test statistic, called *R*-student, given by

$$t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}}$$

If the hypothesis is true, $t \sim t_{n-p-1}$. The *R*-student statistic is also appropriate in testing

$$H_0 : \sigma_1 = 0$$

$$H_1 : \sigma_1 \neq 0$$

where $\mathrm{Var}(\varepsilon_i) = \sigma^2 + \sigma_1^2$.

How to handle outliers

1. Check original source of data. Remove the outlier if it appears to be an error.
2. Provide two data analysis. (with and without outliers)
3. Use indicator variable. (See population example)
4. Use weighted least squares estimation. Put less weight on outliers.

5.6 Diagnostic Plots

a) Residual against estimated response plots :

$$e_j \quad \text{against} \quad \hat{y}_j$$

b) Partial regression plots :

$$\mathbf{e}_y \mid \boldsymbol{x}_{-j} \quad \text{against} \quad \varepsilon_{x_j} \mid \boldsymbol{x}_{-j}$$

c) Component - plus - resident plots :

$$\mathbf{e}_{y|X} + \boldsymbol{x}_j b_j \quad \text{against} \quad \boldsymbol{x}_j$$

d) Normal residual plots

5.6 Diagnostic Plots

a) Residual against estimated response plots :

Plot $e_{y|X}$ against $\hat{y}_j$

If the model is correct, we expect to see a plot with random pattern such that the variance of $e_{y|X}$ at different values of $\hat{y}_j$'s are about constant.

5.6 Diagnostic Plots

Simple linear regression:

- Data: y and x.
- A scatter of y vs x shows how y is related to x

Multiple regression

- Data: y and at least 2 independent variables
- A scatter plot of y vs x_i is not so informative.
- The plot may not truly highlight the role of the individual variable since the interdepedenency among other x_i's has not taken into account.

To determine the appropriateness of the specified relationship between y and X_j,

i) initially, one may try to plot $e_{y|X}$ vs. X_j.

That is, plot residuals vs. j^{th} independent variable.

- If there is a random scattering of points, the relationship between y and X_j is specified correctly in the model.

- If there is a relationship between the points of $e_{y|X}$ vs. X_k (for example, a quadratic relationship), this suggests X_k is not specified correctly in the model.

ii) A better approach is to use partial residual plots.

b) Partial residual plots:

- In multiple regression the correlations among predictors often makes it difficult to see the *unique* effects of one predictor, controlling for the others.
- ***Partial regression residual plots*** are designed to show the relationship between y and each X_k, after the effects of all other predictors have been removed.

To see how y and x_j are related. (after adjusting for the presence of other x's), plot $e_{y|X_{-j}}$ against $e_{x_j|X_{-j}}$

Variation in the plot shows how strong is the term $\hat{\beta}_j x_j$ in the regression.

eg. If y and x_j are related in a linear manner, then the plot should show a linear trend with slope $\hat{\beta}_j$ and intercept 0.

b) Partial residual plots:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{X} = [\boldsymbol{x}_j : \boldsymbol{X}_{-j}]$$

The vector of residuals for $\boldsymbol{y}$ against $\boldsymbol{X}_{-j}$ is

$$\mathbf{e}_{\boldsymbol{y}|\boldsymbol{X}_{-j}} = \boldsymbol{y} - \mathrm{x}_{-j}(\boldsymbol{X}'_{-j}\,\boldsymbol{X}_{-j})^{-1}\ \mathrm{x}'_{-j}\,\boldsymbol{y}$$

The vector of residuals for $\boldsymbol{x}_j$ against $\boldsymbol{X}_{-j}$ is

$$\mathbf{e}_{\boldsymbol{x}_j|\boldsymbol{X}_{-j}} = \mathbf{x}_j - \boldsymbol{X}_{-j}(\boldsymbol{X}'_{-j}\,\boldsymbol{X}_{-j})^{-1}\,\boldsymbol{X}'_{-j}\,\boldsymbol{x}_j$$

Example 5.6 Partial Regression Plot

X1	X2	X3	Y
1.00	13.0	1.50	0.330
1.40	19.3	3.00	0.366
1.80	25.0	1.00	0.294
2.20	10.0	2.50	0.476
2.60	16.0	0.50	0.209
3.00	22.0	2.00	0.451
3.40	28.0	3.50	0.482

$$\hat{y} = 0.2024 + 0.0372\,x_1 - 0.0034\,x_2 + 0.0769\,x_3$$

Use x_1, x_2 and remove x_3, get the following regression model:

$$\boxed{\hat{y} = 0.2804 + 0.0532\,x_1 - 0.0013\,x_2}$$

Example 5.6 (cont):

$$\boldsymbol{e}_{y|x_{-3}} = \boldsymbol{y} - \boldsymbol{H}_{-3}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H}_{-3})\boldsymbol{y}$$

where $\boldsymbol{H}_{-3} = \boldsymbol{X}_{-3}(\boldsymbol{X}'_{-3}\boldsymbol{X}_{-3})^{-1}\boldsymbol{X}'_{-3}$

The regression of x_3 against the variables in x_1 & x_2:

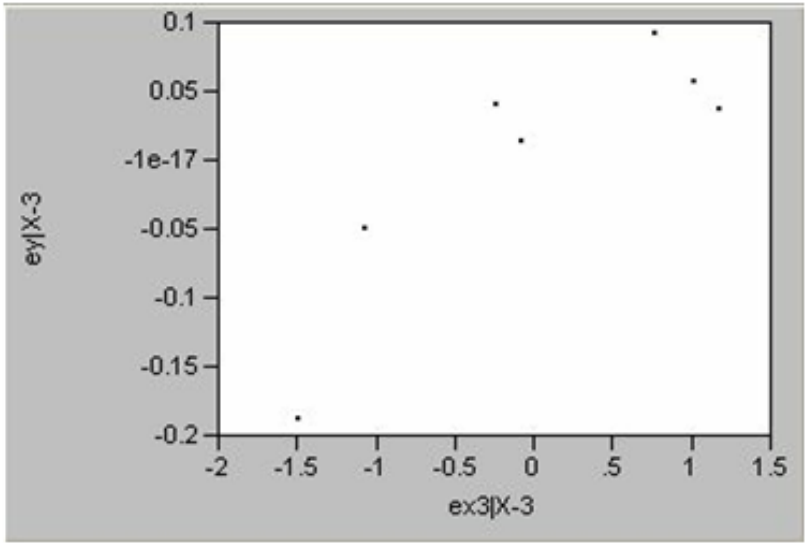
$$\boxed{\hat{x}_3 = 1.0139 + 0.2083x_1 + 0.027x_2}$$

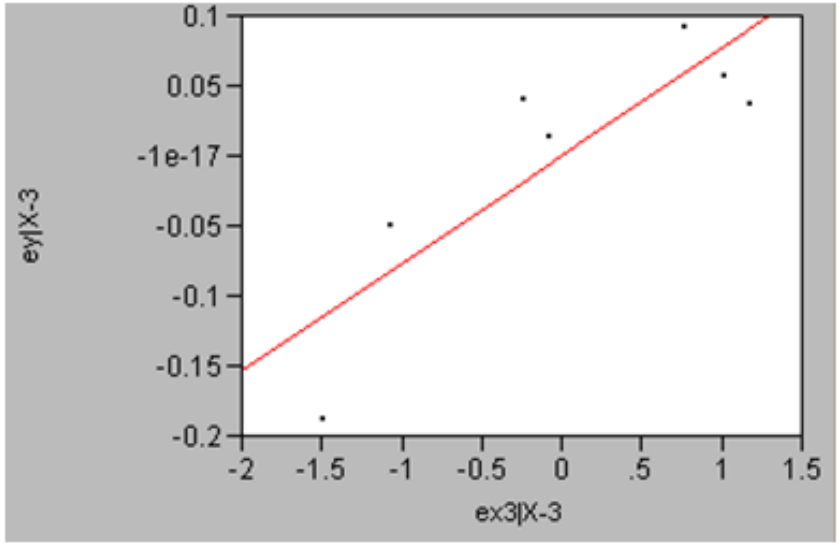
With the corresponding residual:

$$\boldsymbol{e}_{x_3|\boldsymbol{X}_{-3}} = \boldsymbol{x}_3 - \boldsymbol{X}_{-3}(\boldsymbol{X}'_{-3}\boldsymbol{X}_{-3})\boldsymbol{X}'_{-3}\boldsymbol{x}_3$$

X1	X2	X3	Y	$e_{y\|X_{-3}}$	$e_{x_3\|X_{-3}}$
1.00	13.0	1.50	0.330	0.0134	-0.0833
1.40	19.3	3.00	0.366	0.0360	1.1667
1.80	25.0	1.00	0.294	-0.0494	-1.0833
2.20	10.0	2.50	0.476	0.0916	0.7500
2.60	16.0	0.50	0.209	-0.1888	-1.5000
3.00	22.0	2.00	0.451	0.0398	-0.2500
3.40	28.0	3.50	0.482	0.0574	1.0000

Example 5.6 (cont):





The regression of $e_{y|\boldsymbol{X}_{-3}}$ against $e_{x_3|\boldsymbol{X}_{-3}}$:

$$e_{y|\boldsymbol{X}_{-3}} = -0.000 + \boxed{0.0769} e_{x_3|\boldsymbol{X}_{-3}}$$

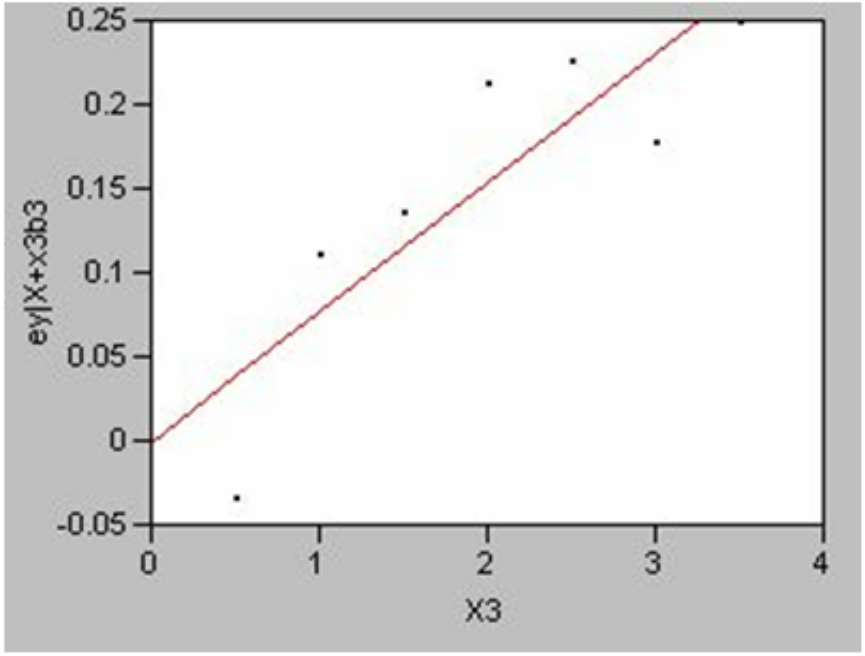
Note that the slope 0.0769 in this model is the same as in the model:

$$\hat{y} = 0.2024 + 0.0372\, x_1 - 0.0034\, x_2 + \boxed{0.0769}\, x_3$$

Example 5.6 (cont):

Plot $e_{y|X} + x_3 b_3$ against x_3

$e_{y\|X} + x_3 b_3$	x_3
0.1346	1.5
0.17714	3
0.10964	1
0.22576	2.5
-0.03572	0.5
0.2118	2
0.24832	3.5



c) Component-plus-residual plots:

Plot $e_{y|X} + x_j b_j$ against x_j

To see how y and x_j are related.

In some cases this can result in a more effective detection on non-linearity in the regressor x_j.

d) Normal residual plots

- Standard normal distribution $Z \sim N(0,1)$
- Normal distribution $X \sim (\mu, \sigma^2)$

$$\left.\right\} Z = \frac{X - \mu}{\sigma}$$

- $P[Z \leq \Phi^{-1}(w)] = w$
- Regression: Fit model; compute and rank the residuals
- Let $y_i - \hat{y}_i$ be the ith smallest residuals.

$$\text{Normal quantile-quantile plot: Plot } \frac{(y_i - \hat{y}_i)}{s} \text{ vs } \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

$$\text{Normal probability plot: Plot } \frac{i}{n} \text{ vs } \Phi\left(\frac{i - 0.375}{n + 0.25}\right)$$

- Ideal plot: Straight line; slope 1; goes through (0,0).

d) Normal quantile-quantile plot

Plot i^{th} ordered residuals vs $\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$

- If the residuals are normally distributed, we expect the plot looks linear with slope σ and intercept μ.
- Very little can be learned in cases where the sample size is small.

5.7 Influence of cases

Another aspect of case analysis is the attempt to understand the influence or importance of each case in fitting models. The general idea is to study changes in a specific part of the analysis when the data are slightly perturbed.

Cases whose removal causes major changes in the analysis are called ***influential***.

5.7 Influence of cases

- An outlier can be produced by model shift or heterogeneous variance.

- A large R student value, $t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}}$, is an indication of an unusually large error has occurred on the ith observation. But it does not access the observation's influence.

- Hat matrix diagonal h_{ii}: a measure of how extreme an individual observation is in the "space" of the explanatory variables. Observations with large h_{ii} are observations that are extreme in the x's.

5.7 Influence of cases

- For a regression model with *p* parameters: Any observation with $h_{ii}>2p/n$ has potential for exerting strong influence on the results. This does not apply for data set with $2p/n>1$.

- Pay attention on observation (s) that have $h_{ii}>2n/p$ or $|t_i|>2$.

- Having identifying the unusual observations, we attempt to understand the influence or importance of each of them in our fitted models.

- The general idea is to study changes in a specific part of the analysis when the data are slightly perturbed.

- Cases whose removal causes major changes in the analysis are called ***influential***.

A. Influence on the fitted value

What influence observation i has on the predicted value or fitted value $\hat{y}_i$. A diagnostic is given by:

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}}$$

The value of $(DFFITS)_i$ for the ith point represents the number of estimated standard errors that the fitted value $\hat{y}_i$ changes if the ith point is removed from the data set.

$$(DFFITS)_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}}\left[\frac{h_{ii}}{1-h_{ii}}\right]^{1/2}$$

$$= (R-student)_i\left[\frac{h_{ii}}{1-h_{ii}}\right]^{1/2}$$

DFFITS

- DF prefix means the difference between the result with x_i and without x_i.
- DFFITS is essentially the *R*-student value, magnified or shrunk according to $[h_{ii}/(1-h_{ii})]^{1/2}$.
- If an observation is an outlier (large R-student) or a high leverage point (h_{ii} close to 1.0), DFFITS will tend to be large.
- If H_{ii} is close to 0, the effect of *R*-student will certainly be moderated.
- See p.258 of the textbook.

B. Influence on the Regression Coefficients

For each regression coefficient, the influence diagnostics provide a statistic, which gives the number of standard errors that the coefficient changes if the ith observation were set aside.

$$(DFBETAS)_{j,i} = \frac{b_j - b_{j,-i}}{s_{-i}\sqrt{c_{jj}}}$$

A large value (in magnitude) of $(DFBETAS)_{j,i}$ indicates that the ith observation has a sizable impact on the jth regression coefficient. The sign of $(DFBETAS)_{j,i}$ may also be meaningful (see page 259).

C. Cook distance

$$D_i = \frac{\left(\boldsymbol{\beta}_{(i)} - \boldsymbol{\beta}\right)'(X'X)\left(\boldsymbol{\beta}_{(i)} - \boldsymbol{\beta}\right)}{ps^2}$$

which can measures influence by comparing $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_{(i)}$, which $\hat{\boldsymbol{\beta}}_{(i)}$ is the estimate of $\boldsymbol{\beta}$ computed without case i. It has several desirable properties:

(a) Contours of Constant D_i are ellipsoids.

(b) Contours can be thought of as defining the distance from $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_{(i)}$.

(c) D_i is invariant under the linear transformation of $\boldsymbol{X}$.

(d) Let $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}_{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)}$. Then $D_i = \dfrac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{ps^2}$