Math 3424
Chapter 5,6,7.3,8. Residual Analysis, Transformation, Influence Diagnostics & Multicollinearity

The residuals

To study the residuals, the basic model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \qquad \mathrm{Var}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$$

The fitted value $\hat{\boldsymbol{Y}}$ corresponding to the observed value $\boldsymbol{Y}$ are then given by

$$\begin{aligned} \hat{\boldsymbol{Y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \boldsymbol{H}\boldsymbol{Y} \end{aligned}$$

$\boldsymbol{H}$ is called the *hat matrix* because it transforms the vector of observed $\boldsymbol{Y}$ into the vector of fitted responses $\hat{\boldsymbol{Y}}$, usually read as y-hat. The vector of residuals $\hat{\boldsymbol{e}}$ is defined by

$$\hat{\boldsymbol{e}} = [\boldsymbol{I} - \boldsymbol{H}]\,\boldsymbol{Y}$$

Difference between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$

The errors $\boldsymbol{e}$ are unobservable random variables, assumed to have zero mean and uncorrelated elements, each with common variance σ^2. The mean and variance of $\hat{\boldsymbol{e}}$ are

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{e}}) &= \boldsymbol{0} \\ \mathrm{Var}(\hat{\boldsymbol{e}}) &= \sigma^2\,(\boldsymbol{I} - \boldsymbol{H}) \end{aligned}$$

In scalar form, the variance of the ith residual is

$$\mathrm{var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$$

Studentized residual

Since $\mathrm{var}(\hat{e}_i)$ will be small whenever h_{ii} is large, so cases with $\boldsymbol{x}_i$ near $\bar{\boldsymbol{x}}$ will have larger residuals, on the average, than cases far from $\bar{\boldsymbol{x}}$.

We consider two very closely related Studentizations that differ only by the choice of estimator for σ^2.

1. The first uses $\hat{\sigma}^2$ to estimate σ^2, giving the formula

   $$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

   The r_i are called *internally Studentized residuals* because the estimate of σ^2 uses all of the data including the ith case.

2. The second scaling uses an estimate of σ^2 obtained when the ith case is excluded from the regression, *externally Studentized residual*, i.e.

   $$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{-i}\sqrt{1 - h_{ii}}}\,.$$

Fig. 5.1 Ideal residual plot

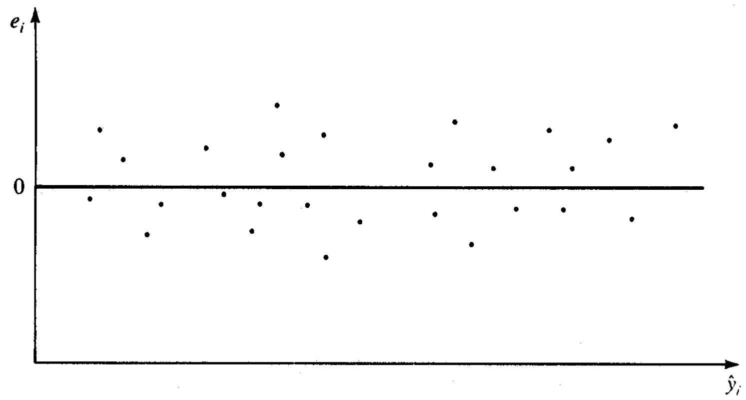
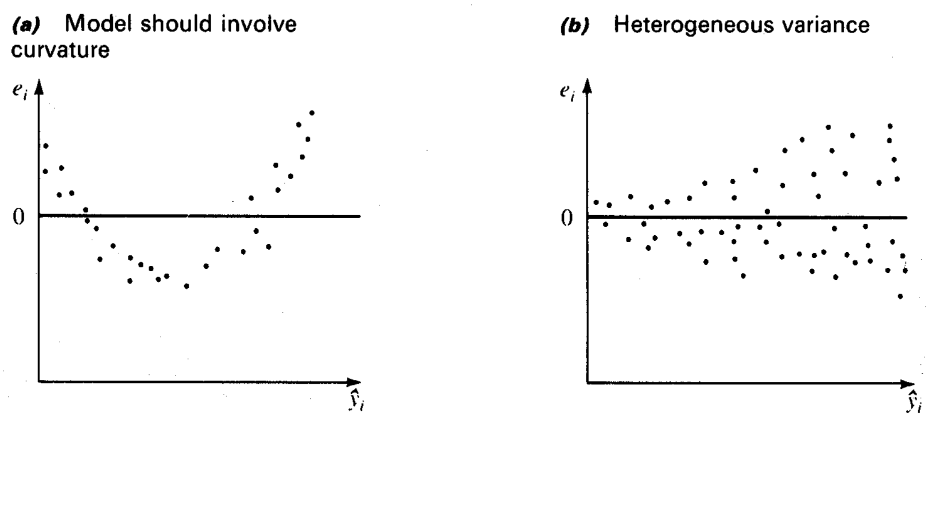


FIGURE 5.2 **Residual plots indicating violation of assumptions:**



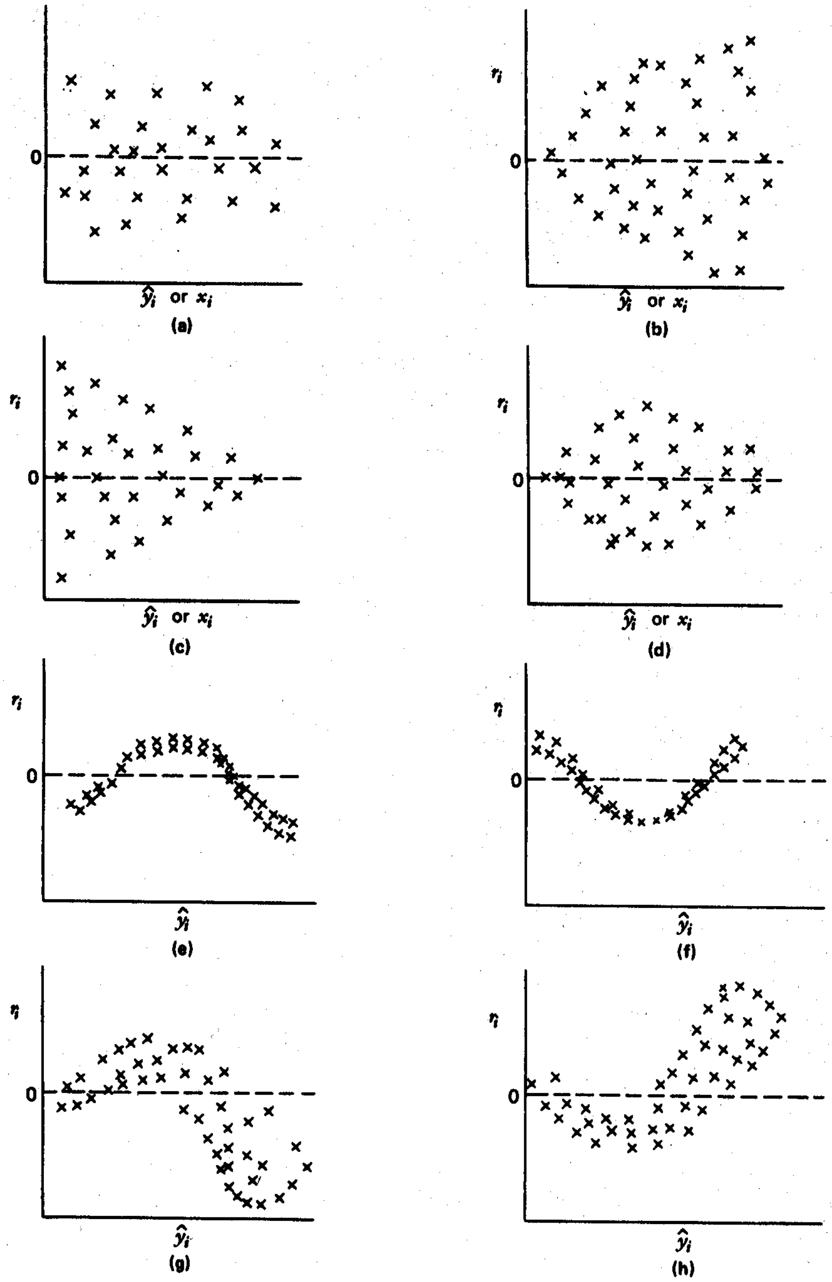


Figure 6.3 Residual plots: (*a*) null plot; (*b*) right-opening megaphone; (*c*) left-opening megaphone; (*d*) double outward bow; (*e*) nonlinearity; (*f*) nonlinearity; (*g*) nonlinearity and nonconstant variance; (*h*) nonlinearity and nonconstant variance.

Normality assumption

A normal probability plot is a quantile-quantile plot of the data. The empirical quantiles are plotted against the quantiles of a standard normal distribution. The vertical coordinate is the ordered data value, $x_{(i)}$, and the horizontal coordinate is

$$\Phi^{-1}((i-3/8)/(n+1/4))$$

where Φ^{-1} is the inverse of the standard normal distribution function and n is the number of nonmissing data values.

Transformation

Transformation on y

1. Some Suggestions

Table 6.1 Common variance stabilizers

Transformation	Situation	Comments
$\sqrt{Y}$	$\mathrm{var}(e_i) \propto E(Y_i)$	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
$\log Y$	$\mathrm{var}(e_i) \propto [E(Y_i)]^2$	This transformation is very common; it is a good candidate if the range of Y is very broad, say from 1 to several thousand; all Y_i must be strictly positive
$\log(Y+1)$	As above	Used if $Y_i=0$ for some cases
$1/Y$	$\mathrm{var}(e_i) \propto [E(Y_i)]^4$	Appropriate when responses are "bunched" near zero, but, in markedly decreasing numbers, large responses 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
$\sin^{-1}(\sqrt{Y})$	$\mathrm{var}(e_i) \propto E(Y_i)(1-E(Y_i))$	For binomial proportions ($0 \leqslant Y_i \leqslant 1$)

2. Box-Cox Transformation

Tukey (1957) introduced a family of power transformations such that

$$y_i^{(\lambda)} = \begin{cases} y_i^{\lambda} & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}$$

for $y_i > 0$. However, this family has been modified by Box and Cox (1964) to take account of the discontinuity at $\lambda = 0$, such that

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

where λ can be estimated from the data. The aims of the Box-Cox transformations is to ensure that the usual assumptions for linear model are more likely to hold after the transformation. That is,

$$\boldsymbol{Y}^{*} \sim N\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}\right)$$

One main convenience in the Box-Cox transformation is that statistical inference on the transformation parameter λ is available via the maximum likelihood (ML) approach.

In relation to the original observations the likelihood function, $l\left(\lambda, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}\right)$, is equal to

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{||\boldsymbol{Y}^{*} - \boldsymbol{X\beta}||^2}{2\sigma^2}\right\} J(\lambda, \boldsymbol{Y})$$

where

$$\begin{aligned} J(\lambda, \boldsymbol{Y}) &= \textstyle\prod_{i=1}^{n} y_i^{\lambda-1} \\ &= GM(y)^{n(\lambda-1)} \end{aligned}$$

and $GM(y)$ is the geometric mean.

Substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ into the likelihood equation, we obtain

$$L(\lambda) = -\frac{n}{2} \log \tilde{\sigma}^2(\lambda) + \log J$$

where $n\tilde{\sigma}^2 = \boldsymbol{Y}^{*T}\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\}\boldsymbol{Y}^{*}$. Write $\boldsymbol{Z}^{\lambda} = \boldsymbol{Y}^{*}/J^{1/n}$, the profile log likelihood for λ can be written as

$$L(\lambda) = -\frac{n}{2} \log[RSS_{\lambda}(\boldsymbol{Z})]$$

where $\boldsymbol{Z}^{\lambda}$ be an $n \times 1$ vector with ith element z_i^{λ} defined by

$$
z_i^{\lambda} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda[GM(y)]^{\lambda-1}} & \lambda \neq 0 \\ \\ GM(y)\log(y_i) & \lambda = 0 \end{cases}
$$

It means that if we fit the model

$$
\boldsymbol{Z}^{\lambda} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}
$$

and compute the residual sum of squares, say $RSS_{\lambda}(\boldsymbol{Z})$, is for each value of λ. The maximum likelihood of λ can be chosen to minimize $RSS_{\lambda}(\boldsymbol{Z})$.

The Box & Cox method is applicable only if the response is strictly positive. If zero or negative values occur, the usual method is to add a constant to the response before applying the method; unfortunately, very little information is available in the data to help choose the added constant.

<u>Transformation on x</u>

We could simply use the natural logarithm transformation of X_j if the ratio of the largest observed value of of X_j to the smallest observed of of X_j is greater than about 10.

Identify any unusual observations

1. Outlier

   Cases that do not follow the same model as the rest of the data are called *outliers*.

   Suppose that the ith case is a candidate for an outlier. We assume that the model for all other cases is

$$y_j = \boldsymbol{x}_j^T \boldsymbol{\beta} + e_j \qquad j \neq i$$

   but for case i, the model is

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \delta + e_i$$

   The ith response y_i has expected value different from $\boldsymbol{x}_j^T \boldsymbol{\beta}$ by the amount δ. Therefore, we can test the ith case to be an outlier if we have a test of $\delta = 0$.

   The test statistic is

$$\begin{aligned} t_i &= \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{-i}\sqrt{1 + \boldsymbol{x_i}^T (\boldsymbol{X_{(i)}}^T \boldsymbol{X_{(i)}})^{-1} \boldsymbol{x_i}}} \\ &= \frac{y_i - \boldsymbol{x_i}^T \left(\hat{\boldsymbol{\beta}} - \dfrac{(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x_i} \hat{e}_i}{1 - h_{ii}} \right)}{\hat{\sigma}_{-i} \sqrt{1 + \boldsymbol{x_i}^T \left((\boldsymbol{X}^T \boldsymbol{X})^{-1} + \dfrac{(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x_i} \boldsymbol{x_i}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1}}{1 - h_{ii}} \right) \boldsymbol{x_i}}} \\ &= \frac{\hat{e}_i}{\hat{\sigma}_{-i}\sqrt{1 - h_{ii}}} \end{aligned}$$

   The test statistic for testing outlier is, in fact, *externally Studentized residual* and has a t distribution with $n - (p + 1) - 1$ degrees of freedom.

   The technique we use to find critical values is based on the *Bonferroni inequality*. We choose the critical value to be the $(\alpha/n) \times 100\%$.

   OR,

   (a) The i^{th} observation is NOT an outlier if $|t_i| < 2$
   (b) The i^{th} observation is possibly an outlier if $2 \leq |t_i| < 3$
   (c) The i^{th} observation is an outlier if $|t_i| > 3$

   If a set of data has more than one outlier, the cases may mask each other, making finding outliers difficult.

2. High-leverage point

It is an observation which is far away from the centroid (sample mean) of all data. We make use of the fact that

$$\sum_{i=1}^{n} h_{ii} = p'$$

where h_{ii} is the diagonal entry of HAT matrix $\boldsymbol{X}\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\ \boldsymbol{X}^T$ and p' is the number of parameters in the model. As a result, the average h_{ii}, namely p'/n, provides a norm.

Remarks

(a) An observation is a high-leverage point if it has a hat-diagonal h_{ii} greater than $2p'/n$.

(b) The hat-matrix only depends on the design matrix and not on the response variables y_i. For simple linear regression,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

(c) If the observation y_i corresponding to a leverage point lies close to the general trend in the data, the point is called a good leverage point, and there is no reason to do anything about the data point.

However, if y_i differs from the main trend, in particular, if y_i corresponds to an outlier, the point is called a bad leverage point, and should be removed from the data set.

(d) A high leverage point will affect the variance of the LS estimate of regression coefficients.

3. Influential observation

It is the observation that causes the LS estimates to be substantially different from what they would be if it is removed from the data.

(a) Cook's distance

Cook's distance is one of the commonly used influence measures in data analysis. It is a measure of change in the LS estimates, $\hat{\boldsymbol{\beta}}$, when the i^{th} observation is deleted. Specifically, the Cook's Distance is defined by the distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{-i}$ after "standardization", i.e.,

$$
\begin{aligned}
D_i &= \frac{(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})}{p' \hat{\sigma}^2} \\
&= \frac{(\hat{\boldsymbol{Y}}_{-i} - \hat{\boldsymbol{Y}})^T (\hat{\boldsymbol{Y}}_{-i} - \hat{\boldsymbol{Y}})}{p' \hat{\sigma}^2} \\
&= \left(\frac{\hat{e}_i^2}{(1-h_{ii})^2} \right) \left(\frac{h_{ii}}{p' \hat{\sigma}^2} \right) \\
&= \left(\frac{r_i^2}{p'} \right) \left(\frac{h_{ii}}{1-h_{ii}} \right)
\end{aligned}
$$

where r_i is the i^{th} studentized residual. As before, D_i becomes large with either a poor fit (large r_i) at the i^{th} point or high leverage (h_{ii} close to 1.0), or both. A simple operation guideline of $D_i > 1$ has been suggested. Others have indicated that $D_i > 4/n$.

A large value of D_i implies that the i^{th} observation exerts undue influence on the set of coefficients. To determine which specific coefficients are affected, one must direct attention to the $(\mathrm{DFBETAS})_{j,i}$.

(b) Influence on the fitted value (DFFITS)

The "DF" prefix means the difference between the result with $\boldsymbol{x}_i$ and without $\boldsymbol{x}_i$.

$$
\begin{aligned}
(\mathrm{DFFITS})_i &= \frac{\hat{y}_i - \hat{y}_{i,-i}}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} \\
&= \left[\frac{\hat{e}_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}} \right] \left[\frac{h_{ii}}{1-h_{ii}} \right]^{1/2} \\
&= (R - \mathrm{student})_i \left[\frac{h_{ii}}{1-h_{ii}} \right]^{1/2}
\end{aligned}
$$

If the data point is an outlier (larger R-student in magnitude) or is a high leverage point (h_{ii} close to 1.0), DFFITS will tend to be large. The diagnostic is produced by the impact of leverage and errors in the y-direction. A general cutoff is 2 and a size-adjusted cutoff is $2\sqrt{p'/n}$ (Belsley, Kuh & Welsch (1980)).

(c) Influence on the regression coefficients (DFBETAS)

$$
(\mathrm{DFBETAS})_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{\hat{\sigma}_{-i}\sqrt{c_{jj}}}
$$

where c_{jj} is the jth diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Let the $(k+1) \times n$ matrix

$$
\boldsymbol{R} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T
$$

with the (q,s) element denoted by $r_{q,s}$. Then

$$(\text{DFBETAS})_{j,i} \;=\; \frac{r_{j,i}}{\sqrt{\boldsymbol{r}'_j\boldsymbol{r}_j}}\,\frac{1}{\sqrt{1-h_{ii}}}\,(R-\text{student})_i$$

Again, the diagnostic represents the combination of leverage measures and the impact of errors in the y-direction. The value of $r_{j,i}/\sqrt{\boldsymbol{r}'_j\boldsymbol{r}_j}$ is a normalized measure impact of errors in the y-direction. A general cutoff is 2 and a size-adjusted cutoff is $2/\sqrt{n}$ (Belsley, Kuh & Welsch (1980)).
A large value (in magnitude) of ($\text{DFBETAS}_{j,i}$) indicates that the ith observation has a sizable impact on the jth regression coefficient. The sign of $(\text{DFBETAS})_{j,i}$ may also be meaningful. A wrong sign of a coefficient may be a result of one erroneous observation or perhaps a model fallacy in the region of the observation.

(d) *COVRATIO*
The covariance ratio (COVRATIO) is defined as

$$COVRATIO \;=\; \frac{|(\boldsymbol{X}^T_{-i}\boldsymbol{X}_{-i})^{-1}\hat{\sigma}_{-i}|}{|(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\sigma}|}$$

where $\boldsymbol{X}_{-i}$ represent the design matrix without the i^{th} observation. A value of this ratio close to 1 would indicate lack of influence of the i^{th} observation. The observation is worth investigation if it is $> 1+3p'/n$ or $< 1-3p'/n$ (Belsley, Kuh & Welsch (1980)).

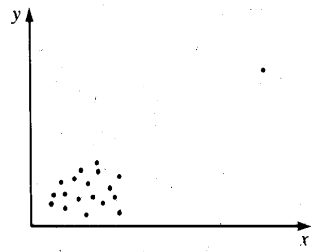
Summary

To investigate the influence of a case more closely, the analyst should delete it and recompute the analysis to see exactly what aspects of it have changed.

Name in SAS output	Expression	Cutoff point
Student Residual	$r_i = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$	$\lvert r_i\rvert > 2$
Rstudent	$t_i = \dfrac{\hat{e}_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}$	$\lvert t_i\rvert > t_{\alpha/(2n)}$
Hat Diag H	$h_{ii} = \boldsymbol{x_i}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_i}$	$h_{ii} > 2p'/n$
Cook's D	$D_i = \left(\dfrac{t_i^2}{p'}\right)\left(\dfrac{h_{ii}}{1-h_{ii}}\right)$	$D_i >> 1$
Dffits	$(\text{DFFITS})_i = \dfrac{\hat{y}_i - \hat{y}_{i,-i}}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} = (\text{Rstudent})_i\left(\dfrac{h_{ii}}{1-h_{ii}}\right)^{1/2}$	$> 2\sqrt{p'/n}$
$(\text{DFBETAS})_{j,i}$	$(\text{DFBETAS})_{j,i} = \dfrac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{\hat{\sigma}_{-i}\sqrt{c_{jj}}} = \dfrac{r_{j,i}}{\sqrt{\boldsymbol{r}'_j\boldsymbol{r}_j}}\dfrac{(R-\text{student})_i}{\sqrt{1-h_{ii}}}$	$> 2/\sqrt{n}$
Cov Ratio	$(\text{COVRATIO})_i = \dfrac{(\hat{\sigma}_{-i})^{2p'}}{\hat{\sigma}^{2p'}}\left(\dfrac{1}{1-h_{ii}}\right)$	$> 1 + 3p'/n$ or $< 1 - 3p'/n$

Fig. 6.1

(a) Single influential observation remote from center



(b) Single observation with error in y-direction

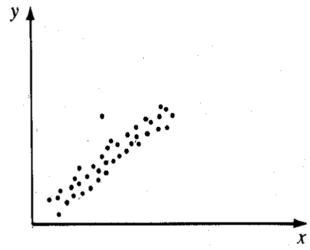


Fig. 6.2 Large HAT diagonal but not influential observation

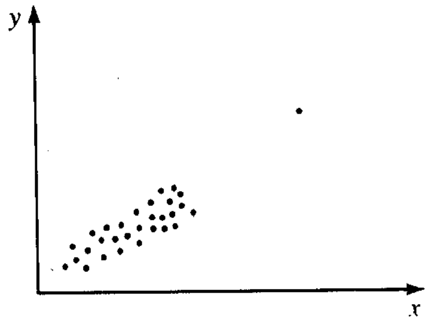
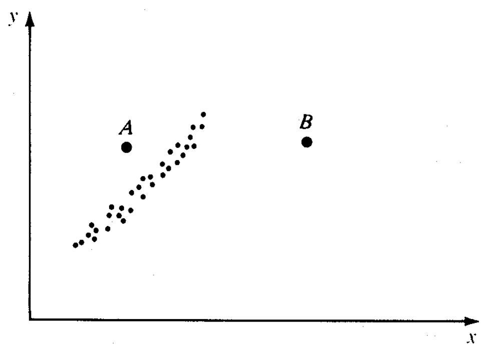


Fig. 6.3 **Point B is clearly influential**



Multicollinearity

e.g. $n = 8$

x_1	10	10	10	10	15	15	15	15
x_2	10	10	15	15	10	10	15	15

$\gamma_{12} = 0$ — linear independent (simple correlation coeff. between x_1 and x_2)

$$X_{i1}{}^* = \frac{X_{i1} - \bar{X}_1}{S_1}$$
$$X_{i2}{}^* = \frac{X_{i2} - \bar{X}_2}{S_2}$$

where $S_1^2 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2$ and $S_2^2 = \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2$

$$X^* = \begin{pmatrix} x_{11}^* & x_{12}^* \\ \vdots & \vdots \\ x_{n1}^* & x_{n2}^* \end{pmatrix}$$

$$\begin{aligned} \sum_{i=1}^n x_{i1}^{*2} &= \sum_{i=1}^n \left(\frac{x_{i1} - \bar{x}_1}{S_1}\right)^2 \\ &= \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}{S_1{}^2} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n x_{i1}^* x_{i2}^* &= \sum_{i=1}^n \left(\frac{x_{i1} - \bar{x}_1}{S_1}\right)\left(\frac{x_{i2} - \bar{x}_2}{S_2}\right) \\ &= \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{S_1 S_2} \end{aligned}$$

$$X^{*T} X^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad (X^{*T} X^*)^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \qquad \text{Var}(\hat{\beta}_0) = \sigma^2 \qquad \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$$

e.g. $n = 8$

x_1	10	11	11.9	12.7	13.3	14.2	14.7	15.0
x_2	10	11.4	12.2	12.5	13.2	13.9	14.4	15.0

$$\gamma_{12} = 0.99215 \text{ — linear dependent}$$

$$X^{*T}X^* = \begin{pmatrix} 1 & 0.99215 \\ 0.99215 & 1 \end{pmatrix} \qquad (X^{*T}X^*)^{-1} = \begin{pmatrix} 63.94 & -63.44 \\ -63.44 & 63.94 \end{pmatrix}$$

$$\text{Var}(\hat{\beta}_1) = 63.94\sigma^2 \qquad \text{Var}(\hat{\beta}_0) = 63.94\sigma^2$$

Multicollinearity occurs when there are near linear dependenceies among the x_j^* the column of X^*. That is, there is a set of constants (not all zero) for which $\sum_{j=1}^{p} c_j x_j^* \approx 0$

Consider a regression with two perdictors:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \\ &= {\beta_0}^* + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + e_i \end{aligned}$$

$$\underset{\sim}{X} = \begin{pmatrix} 1 & x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 \end{pmatrix}, \qquad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0^* \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} n & 0 & 0 \\ 0 & \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 & \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\ 0 & \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) & \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 \end{pmatrix} \qquad (X^T X)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{pmatrix}$$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \sigma^2 \frac{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2 - \left[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right]^2} \\
&= \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 - \frac{\left[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right]^2}{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} \\
&= \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 \left[1 - \frac{\left[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\right]^2}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}\right]} \\
&= \sigma^2 \left(\frac{1}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}\right)\left(\frac{1}{1-r_{12}^2}\right)
\end{aligned}
$$

Also,

$$
\mathrm{Var}(\hat{\beta}_2) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}\right)\left(\frac{1}{1-r_{12}^2}\right)
$$

When $p > 2$,

$$
\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\right)\left(\frac{1}{1-R_j^2}\right)
$$

R_j^2 : coefficient of multiple determination of the regression produced by regression X_j on the other predictors $(X_k, k \neq j)$

The variance inflation factor (VIF) for the ith regression coefficient is defined as $1/(1-R_j^2)$. The higher the multiple correlation in this artificial regression, the lower the precision in the estimate of the coefficient β_i.