3 Dec

Transformation on X

- Naive method $\quad \dfrac{\max(x_i)}{\min(x_i)} > 10 \Rightarrow \log_e(x)$

- Partial Residual ~~model~~ plot

x_l - non-linear relationship with y

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{l-1} x_{i,l-1} + \boxed{f_l(x_{i,l})} + \beta_{l+1} x_{i,l+1} + \cdots + \beta_p x_{i,p}$$

least squares estimates for $\beta_0, \beta_1, \ldots, \boxed{\beta_l}, \ldots, \beta_p$

$$\Rightarrow y_i \approx \hat\beta_0 + \hat\beta_1 x_{i1} + \cdots + \hat\beta_{l-1} x_{i,l-1} + \boxed{f_l(x_{i,l})} + \hat\beta_{l+1} x_{i,l+1} + \cdots + \hat\beta_p x_{i,p}$$

$$\Rightarrow \hat e_i = y_i - \hat y_i$$

$$= y_i - (\hat\beta_0 + \hat\beta_1 x_{i1} + \cdots + \hat\beta_{l-1} x_{i,l-1} + \hat\beta_l x_{i,l} + \hat\beta_{l+1} x_{i,l+1} + \cdots + \hat\beta_p x_{ip})$$

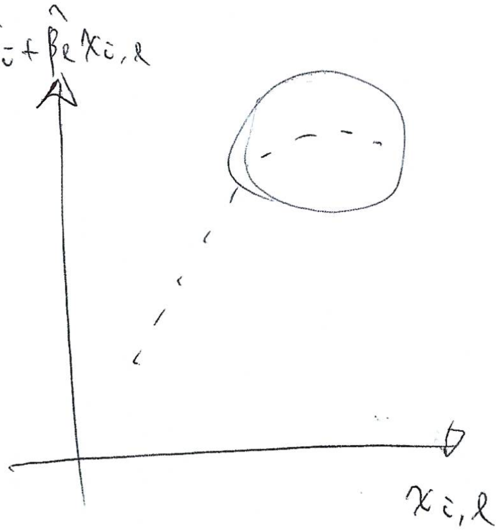
$$\Rightarrow \hat e_i = f_l(x_{i,l}) - \hat\beta_l x_{i,l}$$

$$\Rightarrow f_l(x_{i,l}) = \underbrace{\hat e_i + \hat\beta_l x_{i,l}}_{l^{th}\text{ partial residual}}$$

Partial residual plot

$\hat e_i + \hat\beta_l x_{i,l}$ vs $x_{i,l}$

Subjective !



- Box - Tidwell transformation
  - power transformation
- ~~Interpolati~~ Interpolation spline / smoothing spline

~~Gen~~ generalized additive model

$\Rightarrow$ non-linear part of x_l

e.g.	y - area	m^2	$\sqrt{y}$	$\log(y)$	y
	x - perimeter	m	x	$\log(x)$	x^2

<u>Outlier</u>

<u>Influential Point</u>

$\Big\}\ \Leftarrow$ unusual obs.

Outlier — affect intercept & test H_0 & p-value

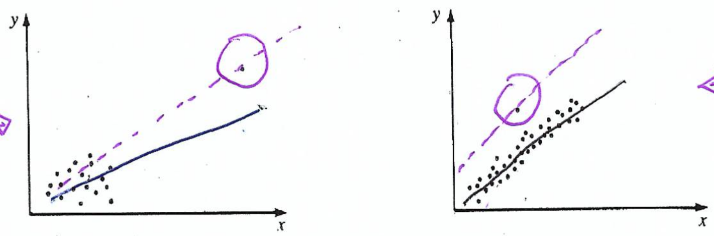
Influential point — affect reg. coeff.

Fig. 6.1

(a) **Single influential observation remote from center**

(b) **Single observation with error in *y*-direction**

Diff. reg. coeff. ⇒ influential point



← Different intercept ⇒ outlier

Fig. 6.2 Large HAT diagonal but not influential observation

far away from $\bar{x}$ ⇒ large h_{ii}

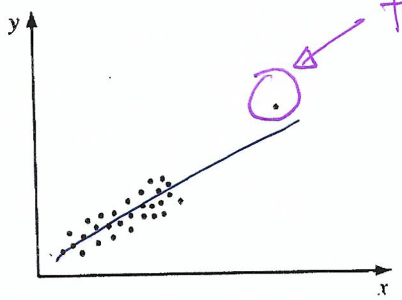
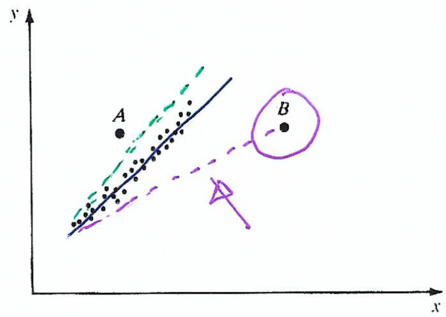


Fig. 6.3 **Point *B* is clearly influential**



Is ~~i~~ an obs. an outlier ? ~~jth obs~~ H_0: ith obs. is not an outlier

$$y_j = \beta_0 + \beta_1 x_{j1} + e_j \qquad j \neq i \qquad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \delta \end{pmatrix}$$

$$y_i = \beta_0 + \delta + \beta_1 x_{i1} + e_i$$

$$\underset{\sim}{X} = \begin{pmatrix} 1 & x_{11} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{i-1,1} & 0 \\ \boxed{1} & \boxed{x_i} & \boxed{1} \\ 1 & x_{i+1,1} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & 0 \end{pmatrix} \leftarrow \text{ith row}$$

H_0 = ith obs. is not an outlier

$\Rightarrow H_0 : \delta = 0$

Using results in Chapter 1

$\Rightarrow \hat{\delta}$, s.e. of $\hat{\delta}$ $\longleftarrow$ $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{Y}$

$\Rightarrow t_i = \hat{\delta} / \text{s.e. of } \hat{\delta}$ $\quad Var(\hat{\underset{\sim}{\beta}}) = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \sigma^2$

$$\Rightarrow t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}} \quad \text{(externally studentized residual)}$$

$\sim t$ (# of obs. $-$ # of unknown para)

$\uparrow$ $\qquad$ n $\qquad$ $(1 + p + 1)$ $\leftarrow \delta$

$$\frac{\text{Res S.S.}}{\text{d.f}} \sim \chi^2$$

$\Rightarrow$ Can't reject H_0 $\Rightarrow$ ith obs is not an outlier

Reject H_0 $\Rightarrow$ ith obs is an outlier

(~~det~~ delete this obs)

$t_1, \dots, t_n$ $-$ multiple tests of n obs.

$\Downarrow$

n tests

$H_0 : \delta_1 = 0, \quad \dots, \quad H_0 : \delta_n = 0$

Bonferroni correction

$$\alpha \longrightarrow \frac{\alpha}{\underbrace{\#\text{ of multiple tests}}_{= n}}$$

Reject H_0 if

$$|t_i| \geq t_{\frac{\alpha}{2n}} (n - (p+2))$$

- If all $|t_i| < 3 \Rightarrow$ no outlier

- If $\max(t_i) >$ critical value, ~~dete~~ delete the corresponding obs.

~~the~~ & then check the 2nd largest of $|t_i|$

If $\max(t_i) <$ Critical value $\Rightarrow$ no outlier

<u>Influential point</u>

	Name	Expression	Cutoff point
	Student Residual	$r_i = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$	$\lvert r_i\rvert > 2$
outlier $\rightarrow$	Rstudent	$t_i = \dfrac{\hat{e}_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}$	$\lvert t_i\rvert > t_{\alpha/(2n)}$
Influential point $\rightarrow$	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	$D_i = \left(\dfrac{t_i^2}{p'}\right)\left(\dfrac{h_{ii}}{1-h_{ii}}\right)$	$D_i >> 1$ — large h_{ii}, large $\lvert t_i\rvert$
	DFFITS	$(\text{DFFITS})_i = \dfrac{\hat{y}_i - \hat{y}_{i,-i}}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} = \underbrace{(\text{Rstudent})_i}_{t_i}\left(\dfrac{h_{ii}}{1-h_{ii}}\right)^{1/2}$	$> 2\sqrt{p'/n}$
	$(\text{DFBETAS})_{j,i}$, $j=1,\dots,p$	$(\text{DFBETAS})_{j,i} = \dfrac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{\hat{\sigma}_{-i}\sqrt{c_{jj}}} = \dfrac{r_{j,i}}{\sqrt{\boldsymbol{r}'_j\boldsymbol{r}_j}}\dfrac{\underbrace{(\text{Rstudent})_i}_{t_i}}{\sqrt{1-h_{ii}}}$	$> 2/\sqrt{n}$
	Cov Ratio	$(\text{COVRATIO})_i = \dfrac{(\hat{\sigma}_{-i})^{2p'}}{\hat{\sigma}^{2p'}}\left(\dfrac{1}{1-h_{ii}}\right)$	$> 1 + 3p'/n$ or $< 1 - 3p'/n$

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

$p' \times n$

$r_{qs} = (q,s)^{th}$ element in R

e.g. ~~obs.~~ k^{th} obs — with large value of h_{ii}

large value of $|t_i|$

— it may be an influential point

$\hat{\beta}_j$, $\hat{\beta}_{j,(-k)}$ changes $j=1,\dots,p$

If there is significant ~~diff~~. between $\hat{\beta}_j$ & $\hat{\beta}_{j(-k)}$, $j=1,\dots,p$

$\Rightarrow$ kth obs is an influential point

Significant changes

$\hat{\beta}_j$		$\hat{\beta}_{j(-k)}$
sign +ve	$\Rightarrow$	sign. -ve
sign +ve/-ve	$\Rightarrow$	insignificant
insignificant	$\Rightarrow$	sign +ve/-ve

~~Multi multi~~

Multicollinearity (among $\underset{\sim}{X}$)

linear regression $\quad Var(\hat{\underset{\sim}{\beta}}) = \boxed{(\underset{\sim}{X}^T \underset{\sim}{X})^{-1}} \sigma^2$

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)

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

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

← z-scores of X_1 & X_2

sample mean $= 0$

sample variance $= 1$

$r_{12} = \sum_{i=1}^{n} (X_{i1}^* - \bar{X}_1^*)(X_{i2}^* - \bar{X}_2^*)$

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

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

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

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

$H_0: \beta_1 = 0$

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

$t_1 = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{\hat{\sigma}}$

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

(first case)

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

$H_0: \beta_1 = 0$

$$t_2 = \frac{\hat{\beta}_1 - 0}{\text{s.e. of } \hat{\beta}_1} = \frac{\hat{\beta}_1}{\sqrt{63.94\hat{\sigma}^2}} = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{63.94}}$$

$\gamma_{12} = 0.99215$ — 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 $\sum_{j=1}^{p} c_j x_j^* \approx 0$ (2nd case)

Consider a regression with two perdictors:

$t_1 > t_2$ for the same $\hat{\beta}_1$

$\Rightarrow$ more diff. to reject Ho for the 2nd case because r_{12} for 2nd case $= 0.99215$

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

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

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

$\Rightarrow$ X_1 & X_2 strongly linear related

$\Updownarrow$

problem of multicollinearity

- delete the indep. varable
- Ridge regression

When $p > 2$, $Var(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{S_{x_1x_1}} \right) \boxed{\left(\frac{1}{1-R_j^2} \right)}$ — VIF

R_j^2 = coeff. of determination of the x_j regression on all other indep. variables x_k, $k \neq j$

$1 - \frac{\text{Res S.S.}}{\text{total S.S.}} = \frac{\text{Reg S.S.}}{\text{total S.S.}}$

Variance Inflation Factor

$R_j^2 \to 1 \Rightarrow Var(\hat{\beta}_1) \to$ large

$\Rightarrow t$ for testing $\beta_1 = 0 \to$ small

- VIF > 10 $\Rightarrow$ problem of multicollinearity $\to$ Can't reject H_0

- condition index

$\underset{\sim}{X}^T \underset{\sim}{X} \Rightarrow$ eigenvalues

$\sqrt{\frac{\max(\text{eigenvalue})}{\text{each eigenvalue}}} > 100 \Rightarrow$ problem of multicollinearity

Condition index

A condition index of 30 to 100 indicates moderate to strong collinearity

e.g.

	VIF
X1	$\boxed{9597.571}$
X2	$\boxed{17.94}$
X3	$\boxed{8993.086}$
X4	23.29386
X5	4.27984

$\leftarrow$ largest > 10 $\Rightarrow$ delete X_1

$\Rightarrow$

X2	7.9258
X3	$\boxed{23.9268}$
X4	12.7060
X5	3.36087

[illegible]

- Descriptive stat.

  ~~quant~~ which one is y ? $y <$ quantitative / binary ..

  quantitative variable / categorical variable

  $\Rightarrow$ clean the data

- Full model
- Residual analysis – residual plot / Q-Q plot

  ⓐ transformation on y (Box-Cox transformation) $\Rightarrow \lambda$

  " " x (naive, partial residual plot, Box-tidewell transformation, spline & etc)

  ⓑ outlier / influential point

  ⓒ multicollinearity – before you make transformation on x

  – quantitative variable only

- Model selection
- Check residual again

  Add the originally deleted outlier / influential point into the best model

- hypothesis testing ?

  prediction ?