Chapter 10

Diagnostic for Multiple Linear Regression

Instructor: Li C.X.

Outline

- Model Adequacy for a Predictor Variable
  - Added-Variable Plots
- Identifying outlying Y
  - Studentized Residuals
  - Studentized Deleted Residuals
- Identifying outlying X
  - Hat Matrix Leverage Values
- Identifying Influential Cases
  - DFFITS, Cook's Distance, DFBETAS
- Multicollinearity Diagnostic
  - Variance Inflation Factor

10.1 Model Adequacy for a Predictor

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Coefficient of Partial Determination

$$R^2_{Y1|2:(p-1)} = \frac{SSR\left(X_1 \mid X_2, \cdots, X_{p-1}\right)}{SSE\left(X_2, \cdots, X_{p-1}\right)}$$

- Define two variables:
  - residuals of predicting Y as function of $X_2, \cdots, X_{p-1}$

$$e_i(Y|X_2, \cdots, X_{p-1}) = Y_i - \widehat{Y}_i(X_2, \cdots, X_{p-1})$$

  - residuals of predicting X_1 as function of $X_2, \cdots, X_{p-1}$

$$e_i(X_1|X_2, \cdots, X_{p-1}) = X_{i1} - \widehat{X}_{i1}(X_2, \cdots, X_{p-1})$$

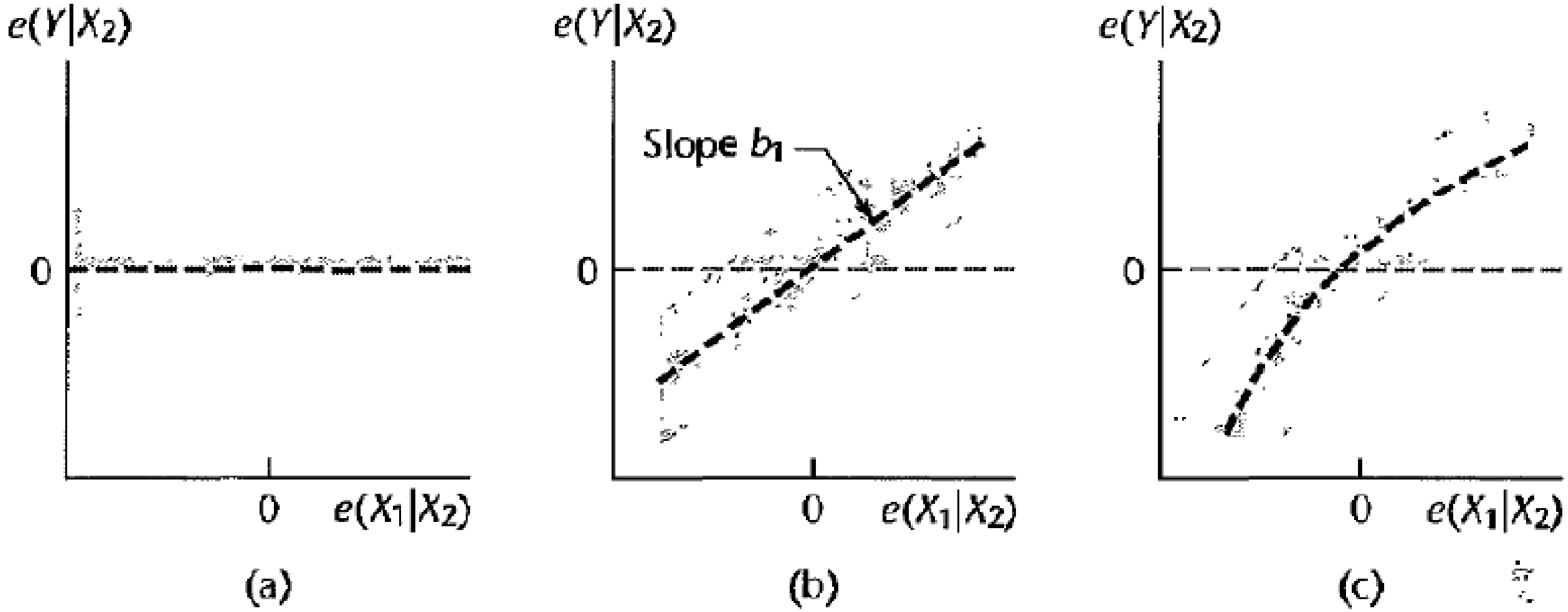
- $R^2_{Y1|2:(p-1)}$ equals to R^2 for regressing $e_i(Y|X_2, \cdots, X_{p-1})$ on $e_i(X_1|X_2, \cdots, X_{p-1})$
- Slope of the regression through the origin of $e_i(Y \mid X_2, \ldots, X_{p-1})$ on $e_i(X_1 \mid X_2, \ldots, X_{p-1})$ is the partial regression coefficient b_1

Added-variable plot

- Graphical way to determine partial relation between response and a given predictor, after controlling for other predictors
- Links with Coefficient of Partial Determination
- Algorithm (assume plot for X_1, given X_2):
  - Fit regression of Y on X_2, obtain residuals = $e_i(Y \mid X_2)$
  - Fit regression of X_1 on X_2, obtain residuals = $e_i(X_1 \mid X_2)$
  - Plot $e_i(Y \mid X_2)$ (vertical axis) versus $e_i(X_1 \mid X_2)$ (horizontal axis)
- The scatter plot of $e_i(Y \mid X_2)$ and $e_i(X_1 \mid X_2)$ provides a graphical representation of the strength of the relationship between Y and X_1, adjusted for X_2.
- The plot is called Added-Variable Plots or adjusted variable plots

Added-variable plot

- Slope of the regression through the origin of $e_i(Y \mid X_2)$ on $e_i(X_1 \mid X_2)$ is the partial regression coefficient for X_1



Body fat Example

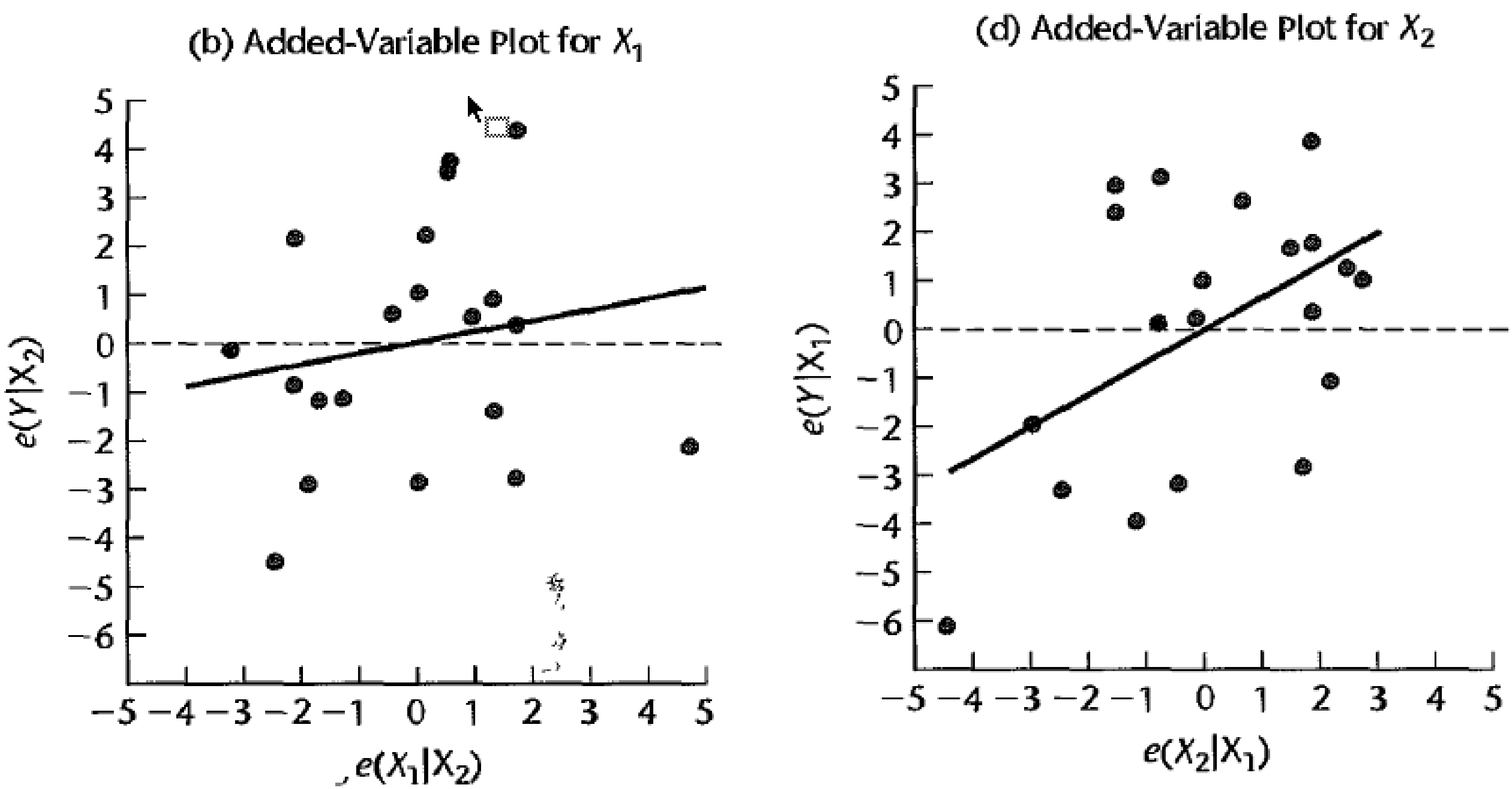
- Y: Body fat percentage
- Predictors

  1. triceps skin fold thickness(X1)
  2. thigh circumference (X2)
  3. midarm circumference (X3)

Subject i	Triceps Skinfold Thickness X_{i1}	Thigh Circumference X_{i2}	Midarm Circumference X_{i3}	Body Fat Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
...	...	...	...	...
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

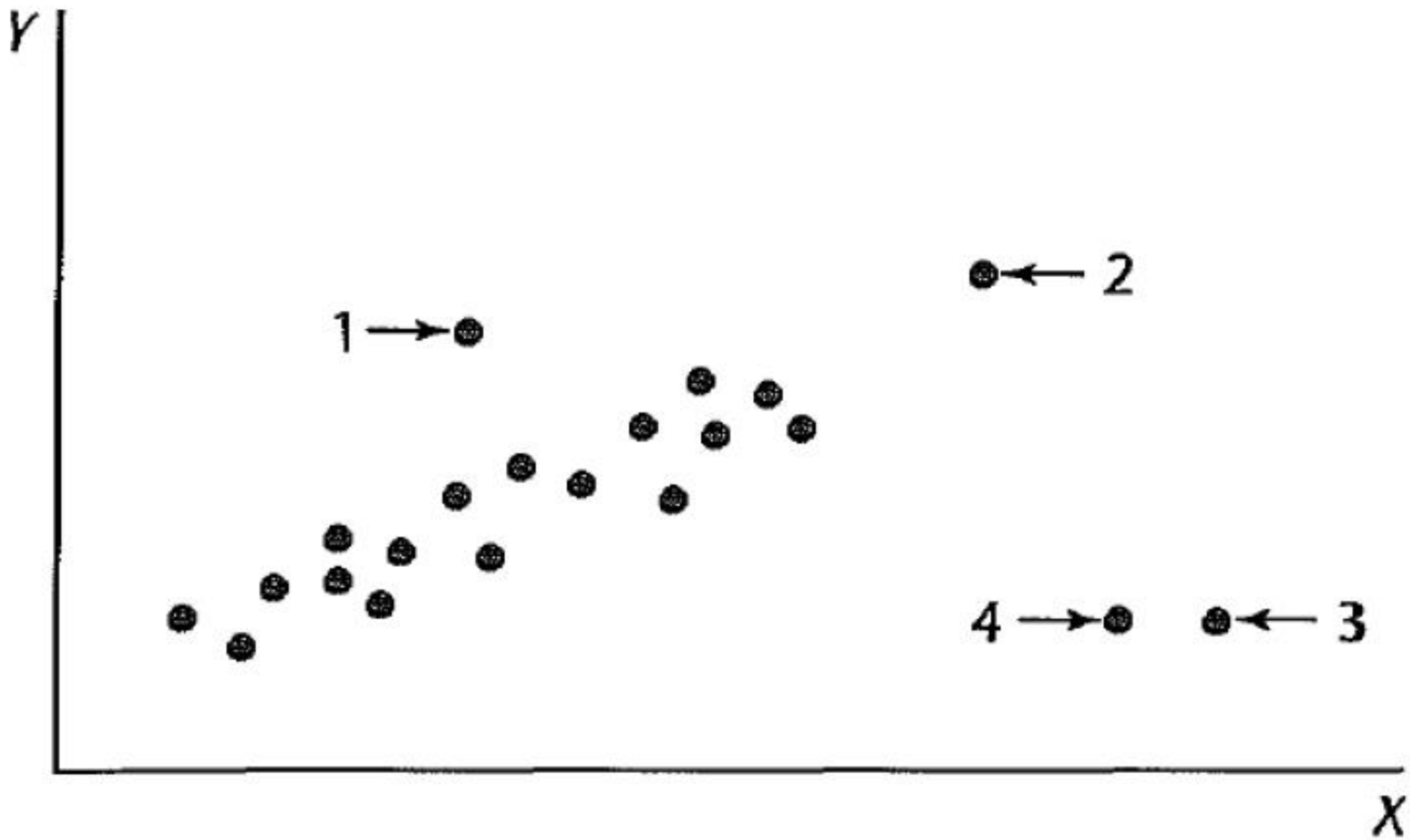
$$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26



10.2 Outlying Y Observations

- Outlying cases



Studentized Residuals

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}, \quad \underset{n\times 1}{\boldsymbol{\varepsilon}} \sim N\left(\underset{n\times 1}{\mathbf{0}}, \sigma^2 \underset{n\times n}{\mathbf{I}}\right)$$

Model Errors (unobserved): $\varepsilon_i = Y_i - \left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right)$

Residuals (observed)

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(b_0 + b_1 X_{i1} + \ldots + b_{p-1} X_{i,p-1}\right)$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \quad \text{where } \underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{E}\left\{e\right\} = \mathbf{E}\left\{\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}\right\} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{E}\left\{\mathbf{Y}\right\} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\boldsymbol{\sigma}^2\left\{\mathbf{e}\right\} = \left(\mathbf{I} - \mathbf{H}\right)\boldsymbol{\sigma}^2\mathbf{I}\left(\mathbf{I} - \mathbf{H}\right)' = \sigma^2\left(\mathbf{I} - \mathbf{H}\right)$$

$$\underset{n\times 1}{\mathbf{e}} \sim N\left(\underset{n\times 1}{\mathbf{0}}, \sigma^2\left(\underset{n\times n}{\mathbf{I}} - \underset{n\times n}{\mathbf{H}}\right)\right)$$

Studentized residuals

$$\underset{n\times 1}{\mathbf{e}} \sim N\left(\underset{n\times 1}{\mathbf{0}}, \sigma^2\left(\underset{n\times n}{\mathbf{I}} - \underset{n\times n}{\mathbf{H}}\right)\right)$$

$$E\{e_i\} = 0; \quad \text{Let } h_{ij} = (i, j)^{th} \text{ element of } \underset{n\times n}{\mathbf{H}} = \mathbf{X(X'X)^{-1}X'}$$

$$\sigma^2\{e_i\} = \sigma^2(1-h_{ii}), \qquad \sigma\{e_i, e_j\} = -h_{ij}\sigma^2 \quad \forall i \neq j$$

$$s^2\{e_i\} = MSE(1-h_{ii}), \qquad s\{e_i, e_j\} = -h_{ij}MSE \quad \forall i \neq j$$

- Studentized residual

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \quad h_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j,$$

$$\text{where } \underset{p\times 1}{\mathbf{x}_i} = \begin{bmatrix} 1 & X_{i1} & \cdots & X_{i,p-1} \end{bmatrix}'$$

- Semi-studentized residual

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

Studentized Deleted Residuals

Deleted Residual---$(X_i,\ Y_i)$ was not used to fit the model

$$d_i = Y_i - \hat{Y}_{i(i)},$$

$\hat{Y}_{i(i)}$: fitted value when regression is fit on the other n - 1 cases

$$\mathbf{b}_{(i)} = \left(\mathbf{X}'_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})}\right)^{\mathbf{-1}} \mathbf{X}'_{(\mathbf{i})}\mathbf{Y}_{(i)} \sim N\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}'_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})}\right)^{\mathbf{-1}}\right)$$

$$\hat{Y}_{i(i)} = b_{0(i)} + b_{1(i)}X_{i1} + \ldots + b_{p-1(i)}X_{i,p-1} = \mathbf{x}_i'\mathbf{b}_{(i)}$$

where $\mathbf{x}_i' = \begin{bmatrix} 1 & X_{i1} & \cdots & X_{i,p-1} \end{bmatrix}$

$$\underset{(n-1)\times p}{\mathbf{X}_{(\mathbf{i})}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{i-1,1} & X_{i-1,2} & \cdots & X_{i-1,p-1} \\ 1 & X_{i+1,1} & X_{i+1,1} & & X_{i+1,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

- Variance of deleted residuals

$$\operatorname{var}\{d_i\} = \operatorname{var}(Y_i) + \operatorname{var}(\hat{Y}_{i(i)}) = \sigma^2 + \operatorname{var}\left\{\mathbf{x}_i'\mathbf{b}_{(i)}\right\}$$

$$= \sigma^2 + \mathbf{x}_i' \operatorname{var}\left\{\mathbf{b}_{(i)}\right\}\mathbf{x}_i = \sigma^2\left[1 + \mathbf{x}_i'\left(\mathbf{X}'_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})}\right)^{-\mathbf{1}}\mathbf{x}_i\right]$$

$$s^2\{d_i\} = MSE_{(i)}\left[1 + \mathbf{x}_i'\left(\mathbf{X}'_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})}\right)^{-1}\mathbf{x}_i\right]$$

Studentized Deleted Residuals

- Studentized deleted residual $$t_i = \frac{d_i}{s\{d_i\}}$$
- Can calculate d_i and t_i in a single model fit $\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$,

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}},$$

$$\text{var}(d_i) = \frac{\text{var}(e_i)}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}}, \qquad s^2\{d_i\} = \frac{MSE_{(i)}}{1-h_{ii}},$$

$$SSE_{(i)} = SSE - \frac{e_i^2}{1-h_{ii}}$$

$$(n-p-1)MSE_{(i)} = (n-p)MSE - \frac{e_i^2}{1-h_{ii}} \Rightarrow MSE_{(i)} =$$

- **PRESS (PRE**diction **S**um of **S**quares) in chapter 9

$$PRESS_p = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{i(i)}\right)^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

Studentized Deleted Residuals

- Studentized deleted residual

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{(1-h_{ii})\sqrt{MSE_{(i)}/(1-h_{ii})}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

- Comparison with studentized residual

$$\boxed{t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}} \quad \text{vs} \quad \boxed{\frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}}$$

$$SSE = (n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

$$\Rightarrow \quad t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} = \frac{e_i\sqrt{n-p-1}}{\sqrt{SSE(1-h_{ii})-e_i^2}}$$

Studentized Deleted Residuals

- If there are no outlying observations,

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}\left(1-h_{ii}\right)}} = \frac{e_i\sqrt{n-p-1}}{\sqrt{SSE\left(1-h_{ii}\right)-e_i^2}} \sim t\left(n-p-1\right)$$

- Adjust for n outlier tests using Bonferroni

$$\text{Outlier if } \left|t_i\right| \geq t\left(1-\left(\frac{\alpha}{2n}\right), n-p-1\right)$$

10.3 Outlying X-Cases

- Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'} = (h_{ij})_{n\times n}$
- In simple linear regression with $p-1=1$,

$$\underset{n\times 2}{\mathbf{X}} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \qquad \mathbf{X'X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}, \quad (\mathbf{X'X})^{-1} = \frac{1}{SS_{XX}} \begin{bmatrix} \frac{SS_{XX}}{n} + \overline{X}^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}$$

Let $\mathbf{x}_i' = [1 \quad X_i]$, *then*

$$h_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j = [1 \;\; X_i](\mathbf{X}'\mathbf{X})^{-1}[1 \;\; X_j]' = \frac{1}{n} + \frac{(X_i - \overline{X})(X_j - \overline{X})}{SS_{XX}}$$

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SS_{XX}} \Rightarrow \boxed{h_{ii} > 0 \text{ and } \sum_{i=1}^n h_{ii} = 2, \text{ if } n > 1}$$

Hat Matrix

In general

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}, \quad h_{ij} = \mathbf{x}_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_j, \quad \mathbf{x}_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

- Properties: $\sum_{i=1}^{n} h_{ii} = trace\left(\mathbf{H}\right) = trace\left(\mathbf{I}_p\right) = p$

$$\underset{n\times n}{\mathbf{H}}\ \underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{X}} \Rightarrow \sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$$

$$\underset{n\times n}{\mathbf{H}} = \underset{n\times n}{\mathbf{H}}\ \underset{n\times n}{\mathbf{H}} \Rightarrow h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h_{ij}^2 \geq 0$$

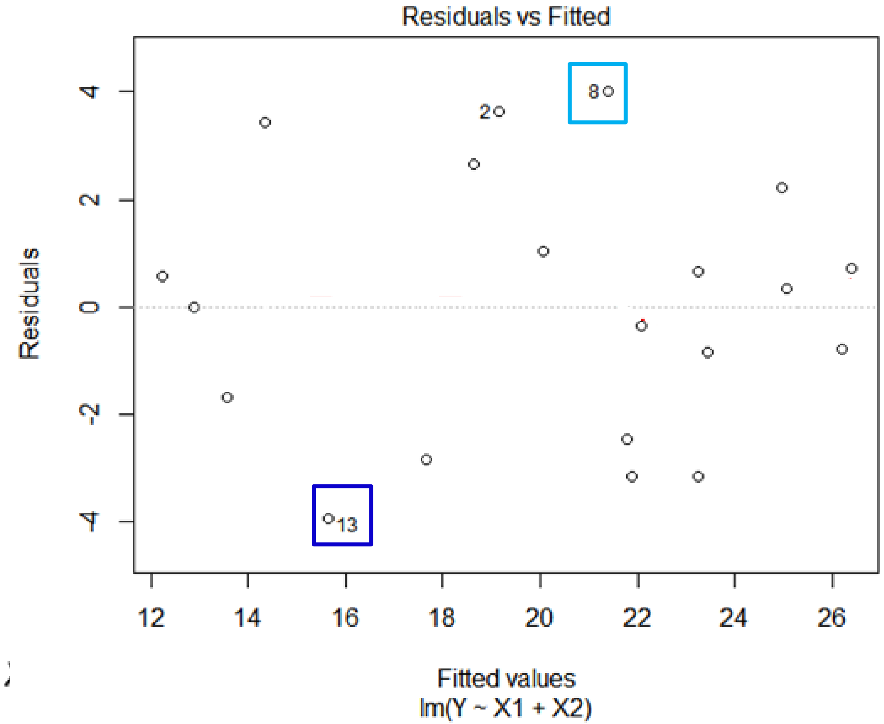
$$(\mathbf{I}-\mathbf{H})^2 = \mathbf{I}-\mathbf{H} \Rightarrow 1-h_{ii} = \sum_{j=1}^{n} (I_{ij}-h_{ij})^2 \geq 0$$

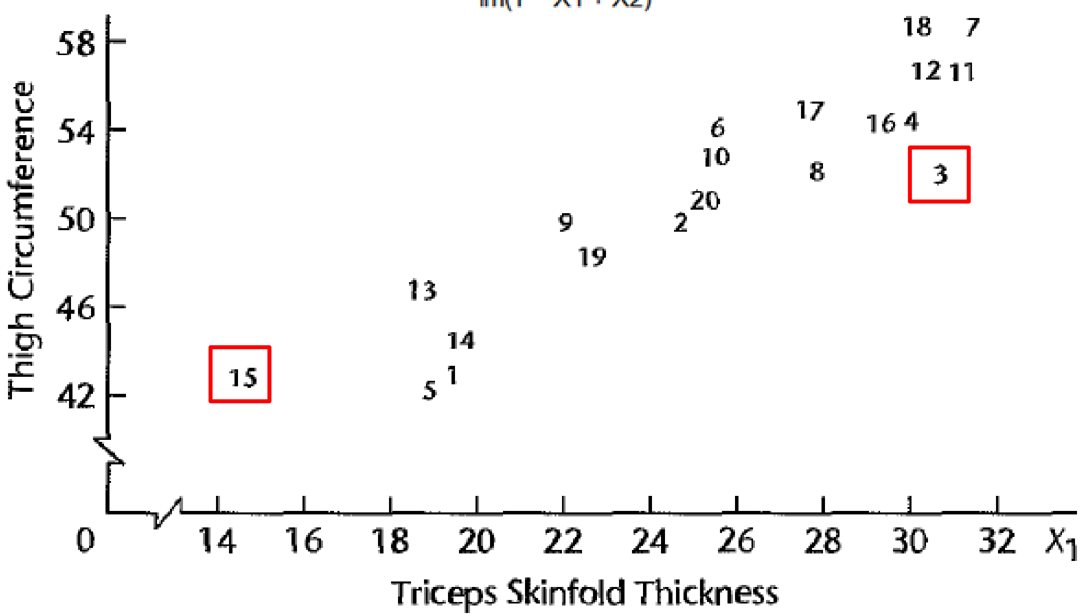
Hat Matrix Leverage Values

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X'X}\right)^{-1}\mathbf{X'}, \quad h_{ii} = \mathbf{x}_i'\left(\mathbf{X'X}\right)^{-1}\mathbf{x}_i,$$

$$0 \le h_{ii} \le 1, \quad \sum_{i=1}^{n} h_{ii} = trace\left(\mathbf{H}\right) = p$$

- h_{ii} known as the leverage of ith case. It is a measure of distance between the X_i value and the mean of the X values.
- Cases with X-levels close to the "center" of the sampled X-levels will have small leverages. Cases with "extreme" levels have large leverages
- Large leverage values: $h_{ii} > 2p/n$





$$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

i	(1) e_i	(2) h_{ii}	(3) t_i
1	−1.683	.201	−.730
2	3.643	.059	1.534
3	−3.176	$\boxed{.372}$	−1.656
4	−3.158	.111	−1.348
5	.000	.248	.000
6	−.361	.129	−.148
7	.716	.156	.298
8	$\boxed{4.015}$	.096	$\boxed{1.760}$
9	2.655	.115	1.117
10	−2.475	.110	−1.034
11	.336	.120	.137
12	2.226	.109	.923
13	$\boxed{-3.947}$	.178	$\boxed{-1.825}$
14	3.447	.148	1.524
15	.571	$\boxed{.333}$	.267
16	.642	.095	.258
17	−.851	.106	.344
18	−.783	.197	.335
19	−2.857	.067	−1.176
20	1.040	.050	.409

$2p/n=0.3$; $t(0.975;16)>2$

Hat Matrix Leverage Values

$$\hat{\mathbf{Y}} = \mathbf{HY} \quad \Rightarrow \quad \hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{i1} Y_1 + h_{i2} Y_2 + \cdots + h_{in} Y_n$$

Note that $\sum_{j=1}^{n} h_{ij} = 1$

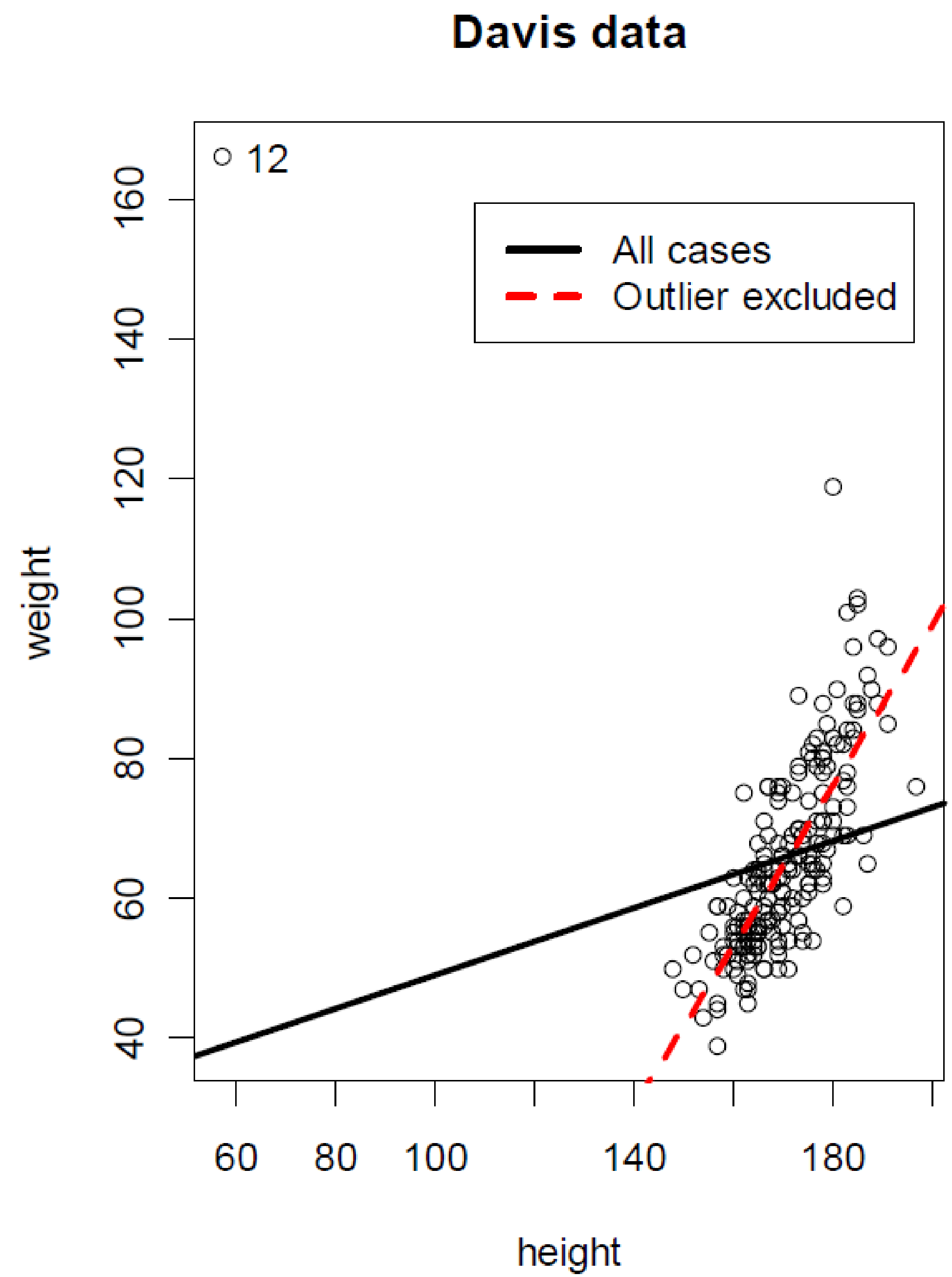
- Thus h_{ii} is a measure of how much Y_i is contributing to the prediction $\hat{Y}_i$
- Cases with large leverages have the potential to "pull" the regression equation toward their observed Y-values.

Leverage values for new observations: $h_{\text{new,new}} = \mathbf{X}_{\text{new}}' \left(\mathbf{X}'\mathbf{X}\right)^{\mathbf{-1}} \mathbf{X}_{\text{new}}$

New cases with leverage values larger than those in original dataset are extrapolations

10.4 Identifying Influential Cases

- These data are the Davis data in the car package
- It is clear that observation 12 is ***influential***
- The model including obs. observation 12 does a poor job of representing the trend
- in the data; The model excluding observation12 does much better



Influence: Davis Data

Model including all cases

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.2662	14.9504	1.69	0.0926
height	0.2384	0.0877	2.72	0.0072

Residual standard error: 14.86

Multiple R-Squared: 0.0359

Model excluding observation #12

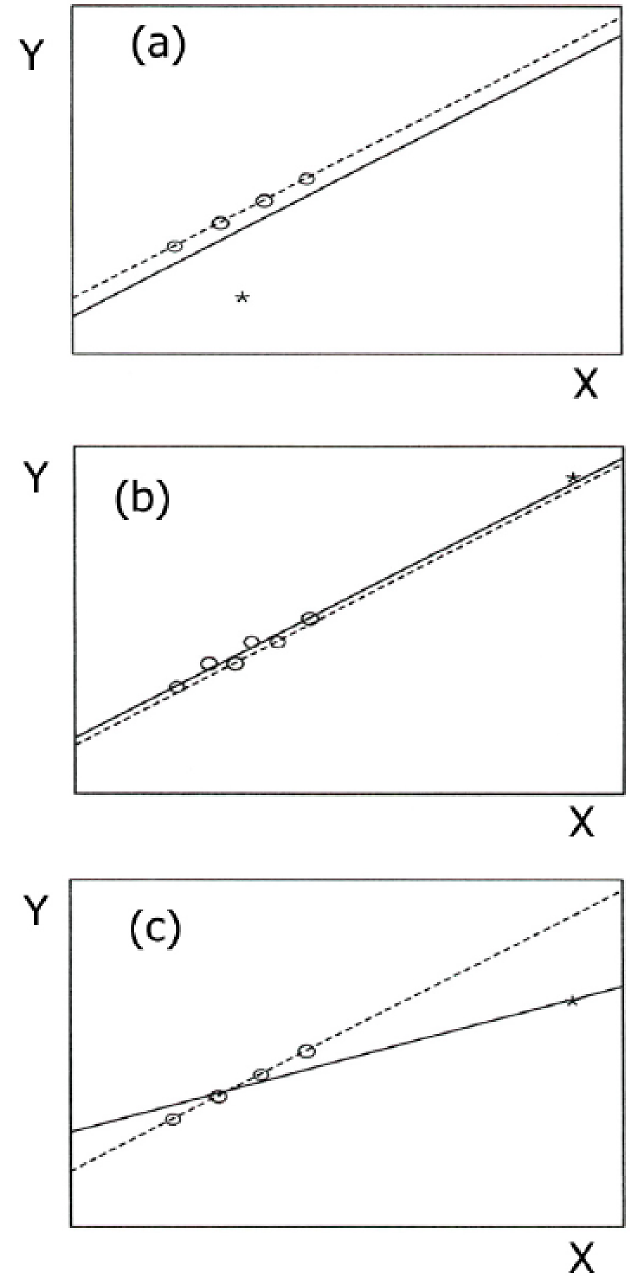
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−130.7470	11.5627	−11.31	0.0000
height	1.1492	0.0677	16.98	0.0000

Residual standard error: 8.523

Multiple R-Squared: 0.594

Types of Unusual Observations

- Figure (a): Unusual Y value has little influence.
- Figure (b): High leverage has no influence
- Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value) results in strong influence. When this case is deleted both the slope and intercept change dramatically.



Adapted from Figure 11.1 (Fox, 1997)

Identifying Influential Cases

- After identifying cases as outlying, we would like to ascertain that these cases are influential, i.e., whether its exclusion causes major changes in the fitted regression function. Three measures:
- Influence on Single Fitted Value (Difference between the fitted values, DFFITS).
  - Influential if |DFFITS| larger than 1 for small to medium data sets, or larger than $2\sqrt{p/n}$ for large data sets.

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} = t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

Estimated sd of $\hat{Y}_i$

$$\text{where } t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} = \frac{e_i\sqrt{n-p-1}}{\sqrt{SSE(1-h_{ii})-e_i^2}}$$

Identifying Influential Cases

- Influence on all fitted values (Cook's Distance)

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE} = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)'\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)}{pMSE} = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{\left(1-h_{ii}\right)^2}\right] = \frac{h_{ii}}{p\left(1-h_{ii}\right)}\tilde{e}_i^2$$

where $\tilde{e}_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$ is studentized residual;

Problem cases are $D_i > F\left(0.50; p, n-p\right)$

- Influence on the Regression Coefficients (DFBETAS).

$$\left(DFBETAS\right)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}$$

Estimated sd of b_k

where c_{kk} is the k-th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1} = (c_{ij})_{p\times p}$

Problem cases are >1 for small to medium sized datasets, $> 2/\sqrt{n}$ for larger ones.

Body fat Example

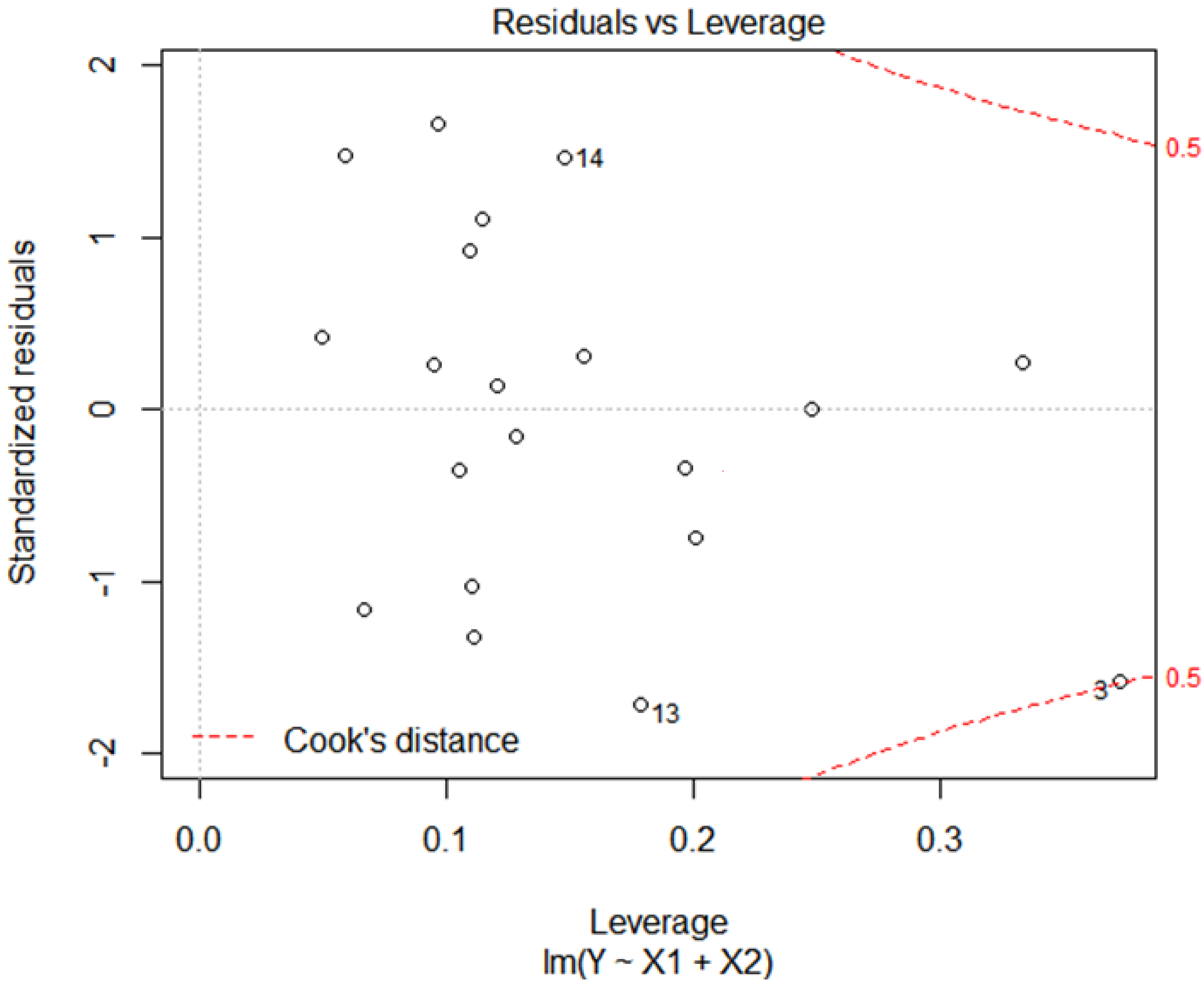
$$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

F(30.6; 3,17)

	(1)	(2)	(3)			DFBETAS		
i	e_i	h_{ii}	t_i	$(DFFITS)_i$	D_i	b_0	b_1	b_2
1	−1.683	.201	−.730	−.366	.046	−.305	−.132	.232
2	3.643	.059	1.534	.384	.046	.173	.115	−.143
3	−3.176	.372	−1.656	−1.273	.490	−.847	1.183	1.067
4	−3.158	.111	−1.348	−.476	.072	−.102	−.294	.196
5	.000	.248	.000	.000	.000	.000	.000	.000
6	−.361	.129	−.148	−.057	.001	.040	.040	−.044
7	.716	.156	.298	.128	.006	−.078	−.016	.054
8	4.015	.096	1.760	.575	.098	.261	.391	−.333
9	2.655	.115	1.117	.402	.053	−.151	−.295	.247
10	−2.475	.110	−1.034	−.364	.044	.238	.245	−.269
11	.336	.120	.137	.051	.001	−.009	.017	−.003
12	2.226	.109	.923	.323	.035	−.131	.023	.070
13	−3.947	.178	−1.825	−.851	.212	.119	.592	−.390
14	3.447	.148	1.524	.636	.125	.452	.113	−.298
15	.571	.333	.267	.189	.013	−.003	−.125	.069
16	.642	.095	.258	.084	.002	.009	.043	−.025
17	−.851	.106	−.344	−.118	.005	.080	.055	−.076
18	−.783	.197	−.335	−.166	.010	.132	.075	−.116
19	−2.857	.067	−1.176	−.315	.032	−.130	−.004	.064
20	1.040	.050	.409	.094	.003	.010	.002	−.003

Body fat Example

$$D_i = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)'\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)}{pMSE} = \frac{h_{ii}}{p\left(1-h_{ii}\right)}\tilde{e}_i^2$$



10.5 Multicollinearity - Variance Inflation Factors

- Problems when predictor variables are correlated among themselves
  - Standard Errors of Regression Coefficients increase when predictors are highly correlated
  - Individual Regression Coefficients are not significant, although the overall model is
  - Point Estimates of Regression Coefficients are wrong sign (+/–)

Original Units for $X_1,\ldots,X_{p-1},Y$: $\quad \boldsymbol{\sigma}^2\{\mathbf{b}\} = \sigma^2(\mathbf{X'X})^{-1}$

Considering the standardized regression model, we have

Correlation Transformed Values: $X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik}-\overline{X}_k}{s_k}\right) \qquad Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i-\overline{Y}}{s_Y}\right)$

$$(X^*)'X^* = r_{XX} \qquad \boldsymbol{\sigma}^2\{\mathbf{b}^*\} = (\sigma^*)^2\mathbf{r}_{\mathrm{XX}}^{-1}$$

Variance Inflation Factor

$$\boldsymbol{\sigma}^2\left\{\mathbf{b}^*\right\}=\left(\sigma^*\right)^2\mathbf{r}_{\mathrm{xx}}^{-1} \qquad \sigma^2\left\{b_k^*\right\}=\left(\sigma^*\right)^2\left(VIF\right)_k$$

where $(VIF)_k$ is the k-th diagonal element of $\mathbf{r}_{\mathrm{xx}}^{-1}$

- It's called variance inflation factor (VIF) for b_k^*
- It can be shown

$$\left(VIF\right)_k=\frac{1}{1-R_k^2}$$

where R_k^2 is the coefficient of determination when X_{k} is regressed on the $p-2$ other X variables.

- For only two predictors, $\left(VIF\right)_1=\left(VIF\right)_2=\dfrac{1}{1-r_{12}^2}$
- Think about what happens when $R_k = 0$ and when R_k is close to ±1.

$$R_k^2=0\Rightarrow\left(VIF\right)_k=1, \qquad R_k^2=1\Rightarrow\left(VIF\right)_k=\infty,$$

$$0<R_k^2<1\Rightarrow\left(VIF\right)_k>1$$

Variance Inflation Factor

- VIF value measure how large is the variance relative to what the variance would be if the predictor variables were uncorrelated.

$$\text{if } \max\left((VIF)_1, \ldots, (VIF)_{p-1}\right) > 10$$

$$\text{or } \left(\overline{VIF}\right) = \frac{1}{p-1}\sum_{k=1}^{p-1}(VIF)_k \text{ is much larger than 1}$$

indicates there is serious multicollinearity problem.

- Body fat example: regression with three predictors X1~X3

Variable	b_k^*	$(VIF)_k$
X_1	4.2637	708.84
X_2	-2.9287	564.34
X_3	-1.5614	104.61

Maximum $(VIF)_k = 708.84$ $\quad$ $(\overline{VIF}) = 459.26$

Appendix:

Theory for adding variables

- Suppose we have the model

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon$$

and want to add the r predictors $\mathbf{X}_2$.

Then the model containing $\mathbf{X}_1$ and $\mathbf{X}_2$ can be written

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon.$$

- If the new predictors $\mathbf{X}_2$ are orthogonal to the old ones $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$ and

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1} & 0 \\ 0 & (\mathbf{X}_2^T\mathbf{X}_2)^{-1} \end{pmatrix}.$$

The least squares estimates are

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1} & 0 \\ 0 & (\mathbf{X}_2^T\mathbf{X}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_2^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y} \\ (\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{y} \end{pmatrix}$$

- When the new predictors are not orthogonal to the old ones, $\mathbf{X}_1^T\mathbf{X}_2 \neq \mathbf{0}$, the situation is more complicated.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon \\ &= \mathbf{X}_1\beta_1 + (\mathbf{H}_1 + \mathbf{I} - \mathbf{H}_1)\mathbf{X}_2\beta_2 + \epsilon \\ &= \mathbf{X}_1\theta + (\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2\beta_2 + \epsilon, \end{aligned}$$

where

$$\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$$

$$\theta = \beta_1 + (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2\beta_2$$

The matrices $\mathbf{X}_1$ and $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2$ are orthogonal, so estimates of $\boldsymbol{\theta}$ and $\beta_{\mathbf{2}}$ can be obtained separately, as above:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{y}$$

$$\hat{\beta}_2 = [\mathbf{X}_2^T (\mathbf{I} - \mathbf{H}_1) \mathbf{X}_2]^{-1} \mathbf{X}_2^T (\mathbf{I} - \mathbf{H}_1) \mathbf{y}.$$

we see that $\hat{\beta}_2$ is the result of regressing one set of residuals, $(\mathbf{I} - \mathbf{H}_1)\mathbf{y}$ on another $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2$. And

$$\hat{\beta}_1 = \hat{\boldsymbol{\theta}} - (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2 \hat{\beta}_2 = [\mathbf{X}_1^T \mathbf{X}_1]^{-1} \mathbf{X}_1^T (\mathbf{y} - \mathbf{X}_2 \hat{\beta}_2).$$

Derivations for case deletion

- A trick is to delete the ith case by adding its indicator, $\mathbf{u}_i$ as a new predictor in the model!
- Let $u_{ij} = 1$ for $j = i$ and $u_{ij} = 0$ for $j \neq i$.

$$Y_j = \beta_0 + \beta_1 X_{j1} + \ldots + \beta_{p-1} X_{j,p-1} + \gamma u_{ij} + \varepsilon_j, \quad j = 1, 2, \cdots, n$$

- Consider the expanded model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}_i\gamma + \epsilon.$$

- The sum of squares function is

$$SSE(\beta, \gamma) = \sum_{j \neq i}^{n} (y_j - \mathbf{x}_j^T \beta)^2 + (y_i - \mathbf{x}_i^T \beta - \gamma)^2.$$

- The deleted estimate for β minimizes the first term, and is based on all cases but the ith.
- Denote these case deleted estimates $\hat{\beta}_{(i)}$.
- The estimate for γ makes the second term zero, and is

$$\hat{\gamma} = y_i - \mathbf{x_i}^T \hat{\beta}_{(i)}$$

the *deleted residual*, $e_{(i)}$, formed using the prediction of $E[y_i]$ without case i.

- Using the theory developed for adding variables

$$\hat{\gamma} = [\mathbf{u}_i^T(\mathbf{I} - \mathbf{H})\mathbf{u}_i]^{-1}\mathbf{u}_i^T(\mathbf{I} - \mathbf{H})\mathbf{y} = \frac{e_i}{1 - h_{ii}} = e_{(i)}.$$

$$\hat{\beta}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{u}_i\frac{e_i}{1 - h_{ii}}) = \hat{\beta} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{e_i}{1 - h_{ii}}.$$

- When γ=0, the model is reduced to

$$Y_j = \beta_0 + \beta_1 X_{j1} + \ldots + \beta_{p-1} X_{j,p-1} + \varepsilon_j, \quad j = 1, 2, \cdots, n$$

Then $\quad SSR(u_i \mid \mathbf{X}) = SSE(\text{Reduced}) - SSE(\text{Full}) = SSE - SSE_{(i)}$

- The partial determination coefficient of

$$R^2_{u_i|\mathbf{X}} = \frac{\mathbf{u}_i^T(\mathbf{I}-\mathbf{H})\mathbf{Y}}{\sqrt{\mathbf{u}_i^T(\mathbf{I}-\mathbf{H})\mathbf{u_i} \cdot \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}}} = \frac{\mathbf{u}_i^T\mathbf{e}}{\sqrt{(1-h_{ii})SSE}} = \frac{e_i}{\sqrt{(1-h_{ii})SSE}}$$

where $R_{u_i|\mathbf{X}}$ is also the Pearson correlation two set of residuals,

$(\mathbf{I}-\mathbf{H})\mathbf{u_i}$ and $(\mathbf{I}-\mathbf{H})\mathbf{Y}$.

$$R_{u_i|\mathbf{X}} = \frac{SSR(u_i \mid \mathbf{X})}{SSE(\mathbf{X})} = \frac{SSE - SSE_{(i)}}{SSE}$$

- Therefore

$$SSE_{(i)} = SSE - SSR(u_i \mid \mathbf{X}) = SSE - R^2_{u_i|\mathbf{X}} SSE = SSE - \frac{e_i^2}{1-h_{ii}}$$

R code

```
#####Body fat example
dat = read.table('CH07TA01.txt')
X1 = dat[,1]; X2 = dat[,2]; X3 = dat[,3];Y = dat[,4]
fit = lm(Y~X1+X2); fit
n = nrow(dat)
###Added Variable Plot
par(mfrow=c(1,2))
fit2 = lm(Y~X2); fit12 = lm(X1~X2)
fit1 = lm(Y~X1); fit21 = lm(X2~X1)
plot(fit12$resi,fit2$resi, main='Added Variable Plot for X1')
abline(lm(fit2$resi~fit12$resi)); lm(fit2$resi~fit12$resi)
plot(fit21$resi,fit1$resi, main='Added Variable Plot for X2')
abline(lm(fit1$resi~fit21$resi)); lm(fit1$resi~fit21$resi)
```

```r
###Examine outlying Y observations
par(mfrow=c(1,1)); plot(fit)
p = 3
elist = fit$resi;  SSE = sum(elist^2)
X = cbind(1,X1,X2)
hlist = diag(X%*%solve(t(X)%*%X)%*%t(X))
tlist = elist*((n-p-1)/(SSE*(1-hlist)-elist^2))^(1/2)
cbind(elist,hlist,tlist)
max(abs(tlist));  qt(0.9975,n-p-1)
###Identifying outlying X observations
2*p/n
hlist  ##Case 3 and 15 larger than 2p/n
```

```
###Identifying influential cases
MSE = SSE/(n-p)
DFFITS = tlist * (hlist/(1-hlist))^0.5
Dlist = elist^2 /p/MSE*hlist/((1-hlist)^2)
clist = diag(solve( t(X)%*%X))
b = fit$coef;  DFBETAS = matrix(0,n,p)
for (i in 1:n){
    fiti = lm(Y[-i]~X1[-i]+X2[-i])
    bi = fiti$coef
    MSEi = sum(fiti$resi^2)/(n-1-p)
    DFBETAS[i,] = (b-bi)/sqrt(MSEi*clist)  }
cbind(Dlist,DFBETAS)
###VIF
Xmat = cbind(X1,X2,X3);VIF3 = diag(solve(cor(Xmat)));VIF3
Xmat = cbind(X1,X2);VIF2 = diag(solve(cor(Xmat)));VIF2
```

Homework

- P415

  10.9 (a)~(c)，(g)

  10.13