

# Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 10 – Building the Regression Model II: Diagnostic

# Overview

- a number of refined diagnostics for checking the adequacy of a regression model
  - detecting improper functional form for a predictor variable
  - outliers
  - influential observations
  - multicollinearity

# Added Variable Plots

## Previous

- Chap. 3,6: check whether a **curvature effect** for that variable is required in the model
- **the residual plots vs. the predictor variables**: determine whether it would be helpful to add one or more of these variables to the model

## Limitation:

- not properly show the nature of **the marginal effect of a predictor variable**, **given the other predictor variables in the model**

# Added Variable Plots, cont'd

- *partial regression plots* or *adjusted variable plots*; provide **graphic information** about the **marginal importance** of a predictor variable  $X_k$ , given the other predictor variables already in the model
- In an added-variable plot, **both  $Y$  and  $X_k$**  under consideration are **regressed against the other predictor variables** and **residuals are obtained for each**.
- the plot of these residuals:
  - the **marginal importance** of this variable in **reducing the residual variability**
  - provide information about **the nature of the marginal regression relation** for  $X_k$  under consideration for possible inclusion in the regression model

# Added Variable Plots, cont'd

- Illustration: the regression effect for  $X_1$ , given that  $X_2$  is already in the model

regress  $Y$  on  $X_2$

$$\hat{Y}_i(X_2) = b_0 + b_2 X_{i2}$$
$$e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$

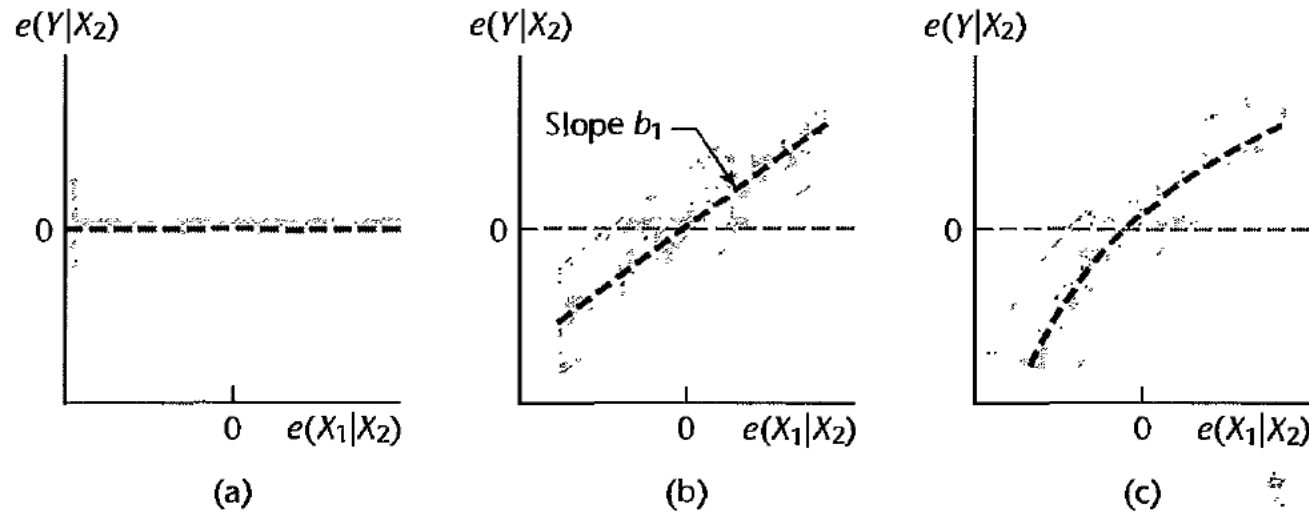
regress  $X_1$  on  $X_2$

$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2}$$
$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

⇒ added variable for predictor variable  $X_1$ :

Plot  $e(Y|X_2)$  vs  $e(X_1|X_2)$

# Added Variable Plots, cont'd



- Figure (a) shows a horizontal band, indicating that  $X_1$  contains no additional information useful for predicting  $Y$  beyond that contained in  $X_2$ , so that it is not helpful to add  $X_1$  to the regression model.
- Figure (b) shows a linear band with a nonzero slope. This plot indicates that a linear term in  $X_1$  may be a helpful addition to the regression model already containing  $X_2$ .
- Figure (c) shows a curvilinear band, indicating that the addition of  $X_1$  to the regression model may be helpful and suggesting the possible nature of the curvature effect by the pattern shown.

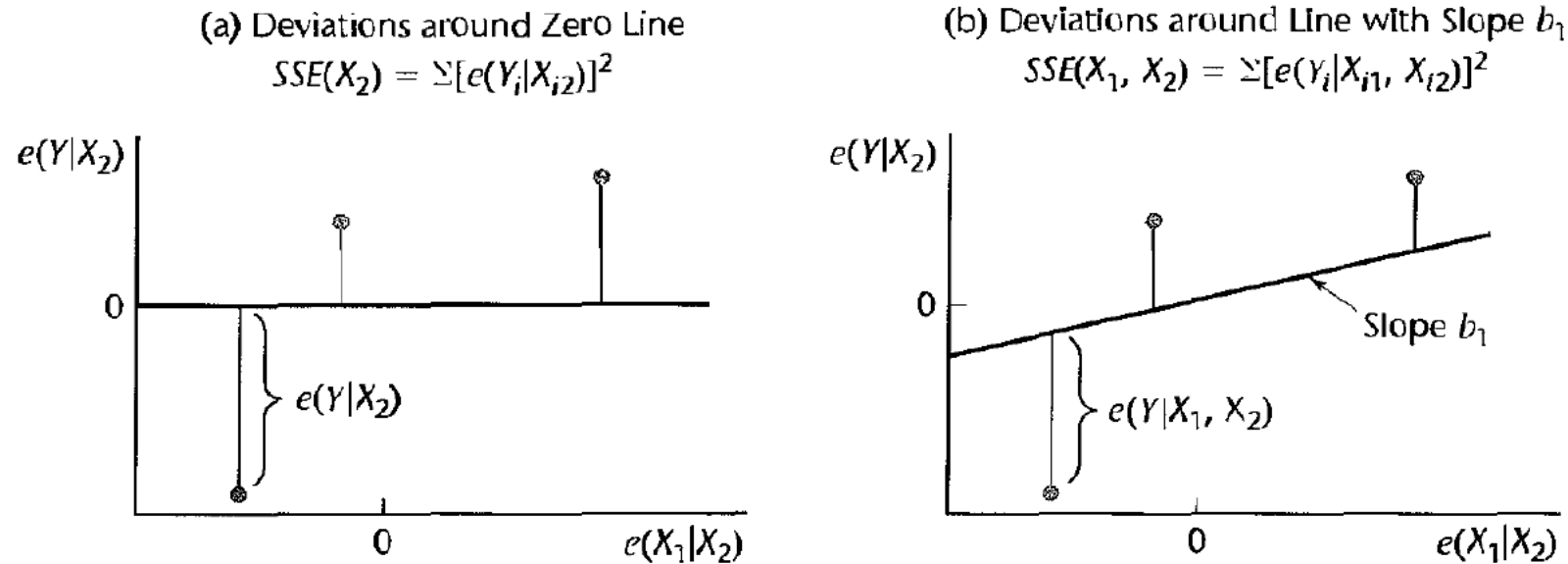
# Added Variable Plots, cont'd

## Added-variable plots:

- providing information about the possible nature of the **marginal relationship** for a predictor variable, given the other variables already in the regression model
- the **strength** of the relationship
- useful for **uncovering outlying** data points that may have a **strong influence in estimating the relationship of the predictor variable  $X_k$  to the response variable**

# Added Variable Plots, cont'd

**FIGURE 10.2 Illustration of Deviations in an Added-Variable Plot.**



- $SSE(X_2)$
- $SSE(X_1, X_2)$
- Difference  $(SSE(X_2) - SSE(X_1, X_2))$ :  $SSR(X_1 | X_2)$ ; provides information about the marginal strength of the linear relation of  $X_1$  to the response variable, given that  $X_2$  is in the model



# Added Variable Plots, cont'd

Example:

Manager $i$	Average Annual Income (thousand dollars) $X_{i1}$	Risk Aversion Score $X_{i2}$	Amount of Life Insurance Carried (thousand dollars) $Y_i$
1	45.01	6	91
2	57.20	4	162
3	26.85	5	11
...	...	...	...
16	46.13	4	91
17	30.37	3	14
18	39.06	5	63

- $r_{12}=0.254$
- $\hat{Y} = -205.72 + 6.288X_1 + 4.738X_2$
- Residual plot: a linear relation for  $X_1$  is not appropriate in the model already containing  $X_2$

# R Code

```
attach(Dataset_10TA01)
fit<-lm(Y~X1+X2)
par(mfrow=c(1,3))
plot(X1,resid(fit),pch=16)
abline(0,0,lty=2,col="gray")
```

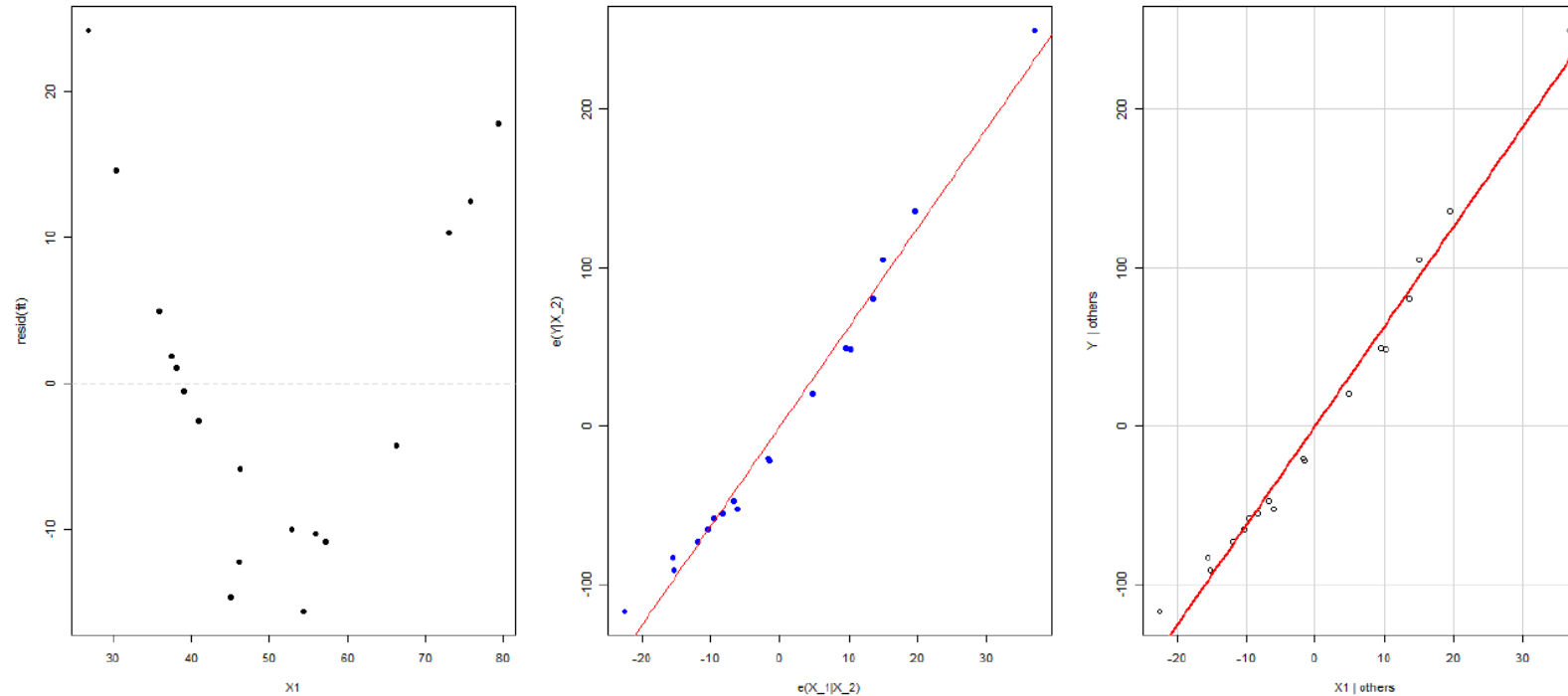
## Method 1:

```
plot(resid(lm(Y~X2)) ~ resid(lm(X1~X2)),col="blue",pch=16,xlab="e(X_1|X_2)", ylab="e(Y|X_2)")
abline(lm(resid(lm(Y~X2))~resid(lm(X1~X2)))),col="red")
```

## Method 2: using avPlot()

```
library(car)
avPlot( model=lm( Y~X1+X2 ), variable=X1 )
```

# Added Variable Plots, cont'd



- $\hat{Y}(X_2) = 50.70 + 15.54X_2$ ;  $\hat{X}_1(X_2) = 40.779 + 1.718X_2$
- Plots: through (0,0);  $b_1 = 6.2880$ ;
- suggest the **curvilinear relation** between  $Y$  and  $X_1|X_2$  is **strongly positive**; a slight concave upward shape
- $R^2_{Y1|2} = 0.984$

# Added Variable Plots, cont'd

## Comments:

- An added-variable plot only suggests the nature of the functional relation in which a predictor variable should be added to the regression model but does not provide an analytic expression of the relation.
- Added-variable plots need to be used with caution for identifying the nature of the marginal effect of a predictor variable.
  - may not show the proper form of the marginal effect of a predictor variable if the functional relations for some or all of the predictor variables already in the regression model are misspecified
  - the relations of the predictor variable to the response variable are complex
  - high multicollinearity among the predictor variables

# Added Variable Plots, cont'd

- Any fitted multiple regression function can be obtained from a sequence of fitted partial regressions. Ex: having  $e(Y|X_2)$ ,  $e(X_1|X_2)$

$$\Rightarrow e(Y|X_2) = 6.2880[e(X_1|X_2)]$$

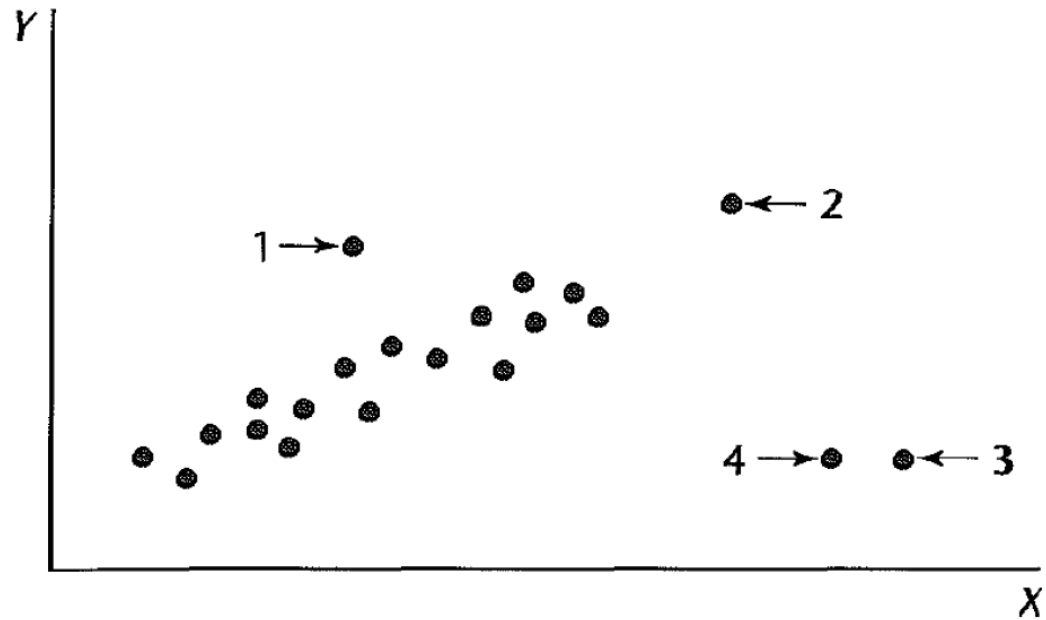
$$\Rightarrow [\hat{Y} - \hat{Y}(X_2)] = 6.2880[X_1 - \hat{X}_1(X_2)]$$

$$\Rightarrow \hat{Y} = -205.72 + 6.2880X_1 + 4.737X_2$$

# Identifying Outlying $Y$ Observations

Outlying or Extreme:

- the observations for these cases are well separated from the remainder of the data
- large residuals; have dramatic effects



# Identifying Outlying $Y$ Observations, cont'd

- Outlying:  $Y$  value,  $X$  values or both
- Not all outlying cases have a strong influence on the fitted regression function.
- A basic step: determine if the regression model under consideration is heavily influenced by one or a few cases in the data set

# Identifying Outlying $Y$ Observations, cont'd

Two refined measures for identifying cases with outlying  $Y$  observations:

- Residuals, Semistudentized Residuals:

$$e_i = Y_i - \hat{Y}_i; \quad e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- Hat matrix:  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \Rightarrow \hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\Rightarrow \sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$$

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}), \quad h_{ii}: \text{the } i\text{th elements on the diagonal of } \mathbf{H}$$

$$\sigma\{e_i, e_j\} = -h_{ij}\sigma^2, \quad i \neq j$$

$$\Rightarrow s^2\{e_i\} = MSE(1 - h_{ii}), \quad s\{e_i, e_j\} = -h_{ij}MSE$$



# Identifying Outlying $Y$ Observations, cont'd

- $h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$ ,  $\mathbf{X}_i = [1 \ X_{i,1} \ \cdots \ X_{i,p-1}]'$

Small data set:  $n = 4$

(a) Data and Basic Results							
$i$	(1) $X_{i1}$	(2) $X_{i2}$	(3) $Y_i$	(4) $\hat{Y}_i$	(5) $e_i$	(6) $h_{ii}$	(7) $s^2\{e_i\}$
1	14	25	301	282.2	18.8	.3877	352.0
2	19	32	327	332.3	-5.3	.9513	28.0
3	12	22	246	260.0	-14.0	.6614	194.6
4	11	15	187	186.5	.5	.9996	.2

(b) H				(c) $s^2\{e\}$					
$\begin{bmatrix}$	.3877	.1727	.4553	-.0157	$\begin{bmatrix}$	352.0	-99.3	-261.8	9.0
$\begin{bmatrix}$	.1727	.9513	-.1284	.0044	$\begin{bmatrix}$	-99.3	28.0	73.8	-2.5
$\begin{bmatrix}$	.4553	-.1284	.6614	.0117	$\begin{bmatrix}$	-261.8	73.8	194.6	-6.7
$\begin{bmatrix}$	-.0157	.0044	.0117	.9996	$\begin{bmatrix}$	9.0	-2.5	-6.7	.2

- $\hat{Y} = 80.93 - 5.84X_1 + 11.32X_2$
- $MSE = 574.9$
- $s^2\{e_1\} = 574.9(1 - 0.3877) = 352.0$

# Identifying Outlying $Y$ Observations, cont'd

- Deleted Residuals: The difference between  $Y_i$  and  $\hat{Y}_{i(i)}$ : (*PRESS prediction error*)

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

- $h_{ii}$ : the larger will be the deleted residuals as compared to the ordinary residual
- the estimated variance of  $d_i$ :

$$s^2\{d_i\} = MSE_{(i)}(1 + \mathbf{X}'_i(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}_i) = \frac{MSE_{(i)}}{1 - h_{ii}}$$
$$\Rightarrow \frac{d_i}{s\{d_i\}} \sim t((n - 1) - p)$$

# Identifying Outlying $Y$ Observations, cont'd

- Studentized Deleted Residuals:

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

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

$$\Rightarrow t_i = e_i \left[ \frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right]^{1/2}$$

- $h_{ii}$ : the larger will be the deleted residuals as compared to the ordinary residual

# Identifying Outlying $Y$ Observations, cont'd

- the estimated variance of  $d_i$ :

$$s^2\{d_i\} = MSE_{(i)}(1 + \mathbf{x}'_i(\mathbf{x}'_{(i)}\mathbf{x}_{(i)})^{-1}\mathbf{x}_i) = \frac{MSE_{(i)}}{1 - h_{ii}}$$

$$\Rightarrow \frac{d_i}{s\{d_i\}} \sim t((n - 1) - p)$$

- **Test for Outliers:** whose studentized deleted residuals are **large in absolute value**
  - If the regression model is appropriate, so that no case is outlying. Each  $t_i \sim t(n - p - 1)$ .
  - $|t_i|$ : the appropriate Bonferroni critical value:  
 $t(1 - \alpha/2n; n - p - 1)$

# Identifying Outlying $Y$ Observations, cont'd

## Body fat example

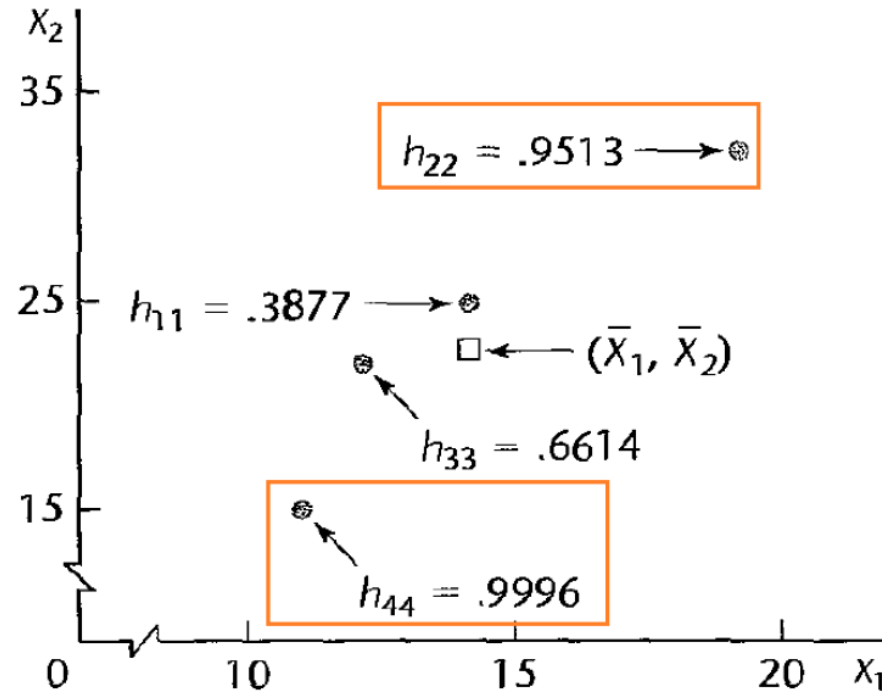
$i$	(1) $e_i$	(2) $h_{ii}$	(3) $t_i$
1	-1.683	.201	-.730
2	3.643	.059	1.534
3	-3.176	.372	<u>-1.656</u>
4	-3.158	.111	-1.348
5	.000	.248	.000
6	-.361	.129	-.148
7	.716	.156	.298
8	4.015	.096	<u>1.760</u>
9	2.655	.115	1.117
10	-2.475	.110	-1.034
11	.336	.120	.137
12	2.226	.109	.923
13	-3.947	.178	<u>-1.825</u>
14	3.447	.148	1.524
15	.571	.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

- $|t| < 3.252 = t(1 - \alpha / 2n; n - p - 1)$
- The Bonferroni procedure provides a **Conservative test** for the presence of an outlier.

# Identifying Outlying $X$ Observations

- Using  $\mathbf{H}$  for identifying outlying  $X$
- $0 \leq h_{ii} \leq 1 \quad \sum_{i=1}^n h_{ii} = p$
- $h_{ii}$ : called *leverage*; measure the distance between  $X_i$  and  $\bar{X}$ 
  - large  $h_{ii} \Rightarrow X_i$  distant from the center of all  $X$ s

**FIGURE 10.6**  
Illustration of  
Leverage  
Values as  
Distance  
Measures—  
Table 10.2  
Example.



# Identifying Outlying $X$ Observations, cont'd

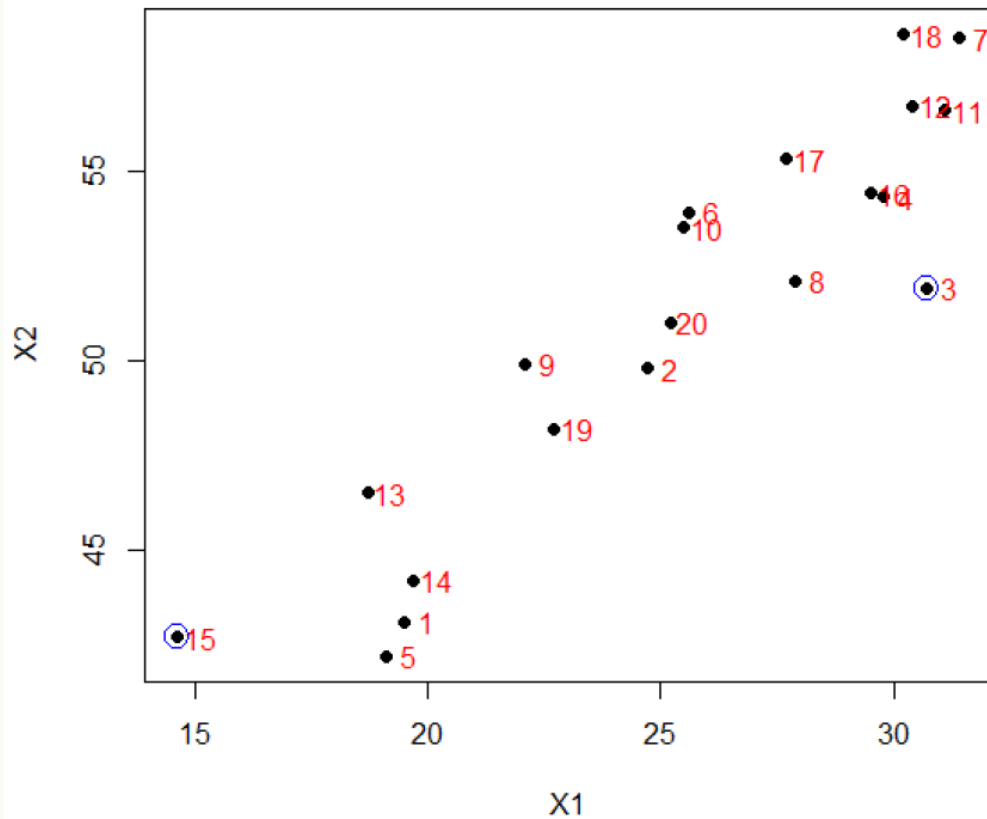
- If  $X_i$  is **outlying**  $\Rightarrow$  has a **large leverage**  $h_{ii}$
- $\hat{Y}_i$ : a linear combination of  $Y$  ( $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ )
- $h_{ii}$ : the weight of  $Y_i \Rightarrow$  the larger is  $h_{ii}$ , the more important is  $Y_i$  in determining  $\hat{Y}_i$
- The larger is  $h_{ii}$ , the smaller is  $\sigma^2\{e_i\}$ .
- $h_{ii} = 1 \Rightarrow \sigma^2\{e_i\} = 0$
- Rule:

$$h_{ii} > 2\bar{h} = 2\frac{\sum h_{ii}}{n} = 2\frac{p}{n} \quad \left(\frac{2p}{n} \leq 1\right)$$

$$\begin{cases} \text{very high leverage: } h_{ii} > 0.5 \\ \text{moderate leverage: } h_{ii} : 0.2 \sim 0.5 \end{cases}$$

# Identifying Outlying $X$ Observations, cont'd

## Body fat example



- $2p/n = 0.30$
- $h_{33} = 0.372;$   
 $h_{15,15} = 0.333$



# Identifying Outlying $X$ Observations, cont'd

```
attach(Dataset_07TA01)
fit<-lm(Y~X1+X2+X3)
n<-length(Y); p = 3
plot(X2~X1, pch=16 )
text(X1+0.5, X2,
labels=as.character(1:length(X1)),col="red")
hii<-hatvalues(fit)
index<-hii>2*p/n
points( X1[index], X2[index], cex=2.0, col="blue")
```

# DFFITS measure

- Influence on **single** fitted value: **DFFITS**-measure the influence that case  $i$  has on  $\hat{Y}_i$


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

- Rule:

$$\begin{cases} |DFFITS| > 1 & \text{for small to medium data sets} \\ |DFFITS| > 2\sqrt{p/n} & \text{for large data sets} \end{cases}$$

- If  $X_i$  is an outlier and has a high  $h_{ii}$ ,  $(DFFITS)_i$  will tend to be large absolutely.

# Cook's Distance

- Influence on **all** fitted value: *Cook's distance*-consider the influence of the  $i$ th case on all  $n$  fitted values 

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

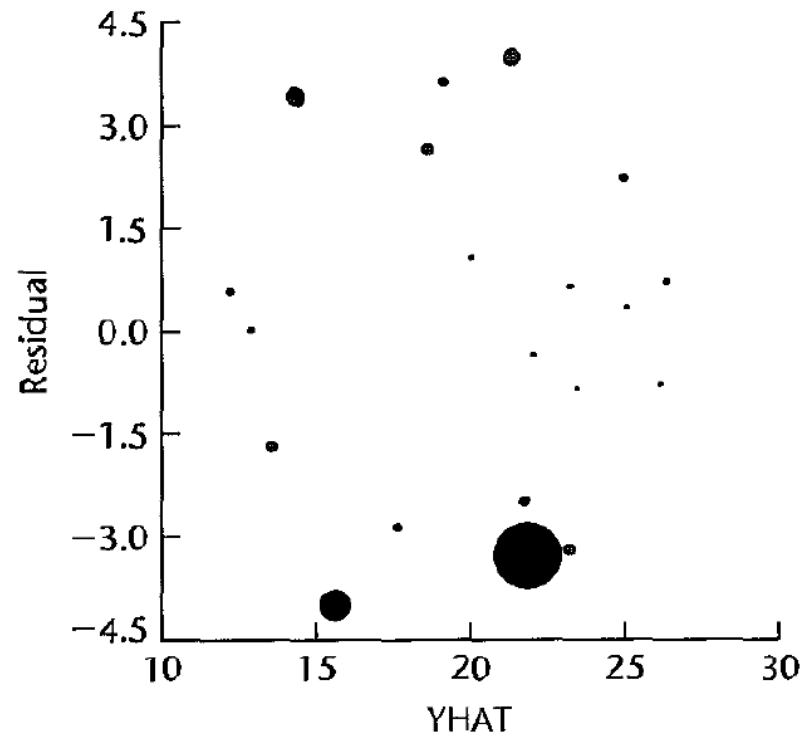
- 1 the size of  $e_i$
- 2 the leverage  $h_{ii}$
- 3  $e_i \uparrow$  or  $h_{ii} \uparrow \Rightarrow D_i \uparrow$

- Rule:  $D_i \sim F(p, n - p)$

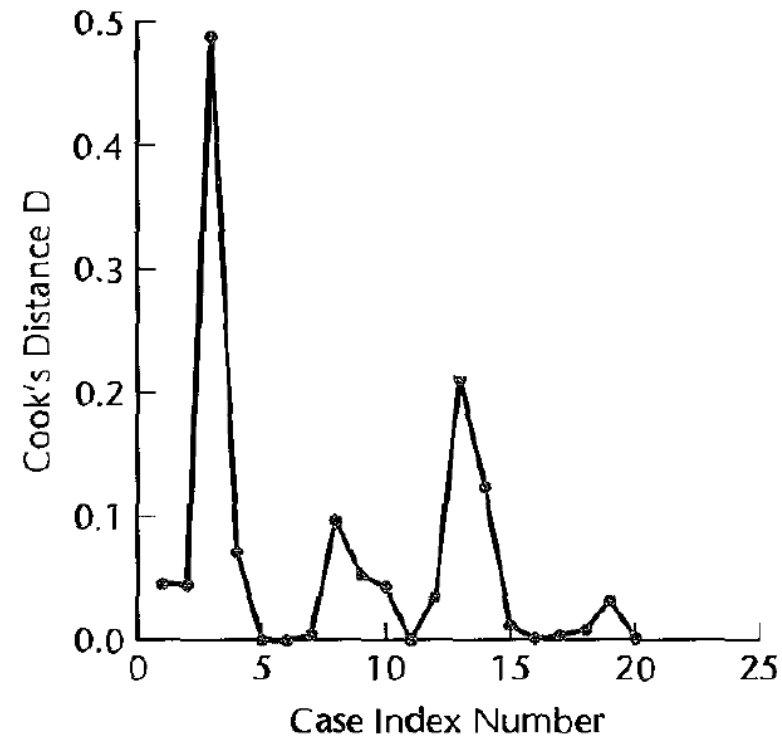
$$\begin{cases} \text{little influence : } P(F(p, n - p) \leq D_i) > 0.1 \text{ or } 0.2 \\ \text{major influence : } P(F(p, n - p) \leq D_i) > 0.5 \end{cases}$$

# Cook's Distance, cont'd

(a) Proportional Influence Plot



(b) Index Influence Plot



# DFBETAS

- Influence on regression coefficients: *DFBETAS*—the difference between  $b_k$  and  $b_{k(i)}$

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}, \quad k = 0, 1, \dots, p - 1$$

where  $c_{kk}$ : the  $k$ th diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$

- Rule:

$$\begin{cases} |DFBETAS| > 1 & \text{for small to medium data sets} \\ |DFBETAS| > 2\sqrt{n} & \text{for large data sets} \end{cases}$$

# DFBETAS, cont'd

```
# Body fat example (Table 10.4)
```

```
> influence.measures(fit)
```

```
Influence measures of
```

```
lm(formula = Y ~ X1 + X2) :
```

	dfb.1_	dfb.X1	dfb.X2	dffit	cov.r	cook.d	hat	inf
1	-3.05e-01	-1.31e-01	2.32e-01	-3.66e-01	1.361	4.60e-02	0.2010	
2	1.73e-01	1.15e-01	-1.43e-01	3.84e-01	0.844	4.55e-02	0.0589	
3	-8.47e-01	-1.18e+00	1.07e+00	-1.27e+00	1.189	4.90e-01	0.3719	*
4	-1.02e-01	-2.94e-01	1.96e-01	-4.76e-01	0.977	7.22e-02	0.1109	
5	-6.37e-05	-3.05e-05	5.02e-05	-7.29e-05	1.595	1.88e-09	0.2480	*
6	3.97e-02	4.01e-02	-4.43e-02	-5.67e-02	1.371	1.14e-03	0.1286	
7	-7.75e-02	-1.56e-02	5.43e-02	1.28e-01	1.397	5.76e-03	0.1555	
8	2.61e-01	3.91e-01	-3.32e-01	5.75e-01	0.780	9.79e-02	0.0963	
9	-1.51e-01	-2.95e-01	2.47e-01	4.02e-01	1.081	5.31e-02	0.1146	
10	2.38e-01	2.45e-01	-2.69e-01	-3.64e-01	1.110	4.40e-02	0.1102	
11	-9.02e-03	1.71e-02	-2.48e-03	5.05e-02	1.359	9.04e-04	0.1203	
12	-1.30e-01	2.25e-02	7.00e-02	3.23e-01	1.152	3.52e-02	0.1093	
13	1.19e-01	5.92e-01	-3.89e-01	-8.51e-01	0.827	2.12e-01	0.1784	
14	4.52e-01	1.13e-01	-2.98e-01	6.36e-01	0.937	1.25e-01	0.1480	
15	-3.00e-03	-1.25e-01	6.88e-02	1.89e-01	1.775	1.26e-02	0.3332	*
16	9.31e-03	4.31e-02	-2.51e-02	8.38e-02	1.309	2.47e-03	0.0953	
17	7.95e-02	5.50e-02	-7.61e-02	-1.18e-01	1.312	4.93e-03	0.1056	
18	1.32e-01	7.53e-02	-1.16e-01	-1.66e-01	1.462	9.64e-03	0.1968	
19	-1.30e-01	-4.07e-03	6.44e-02	-3.15e-01	1.002	3.24e-02	0.0670	
20	1.02e-02	2.29e-03	-3.31e-03	9.40e-02	1.224	3.10e-03	0.0501	

# DFBETAS, cont'd

Some final comments:

- Analysis of outlying and influential cases: a necessary component of good regression analysis
  - neither automatic nor foolproof
  - require good judgment by the analyst
- Methods described: ineffective
- **Extensions of the single-case diagnostic** procedures: computational requirements

# Variance Inflation Factor-VIF

Some problems: multicollinearity

- Adding or deleting  $X$ : change the regression coefficients
- Extra sum of squares: depending upon which other  $X_k$  variables are already included in the model
- $X_k$  highly correlated with each other  $\Rightarrow s\{b_k\} \uparrow$
- the estimated regression coefficients individually may not be statistically significant



# Variance Inflation Factor-VIF, cont'd

informal diagnostics for multicollinearity:

1. Large changes in  $b_k$  when  $X_k$  is added or deleted, or when an observation is altered or deleted
2. Nonsignificant results in individual tests on the regression coefficients for important predictor variables.
3. Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical consideration or prior experience.
4. Large  $r_{xx}$
5. Wide confidence interval for  $\beta_k$

# Variance Inflation Factor-VIF, cont'd

Important limitations:

- do not provide **quantitative measurements**
- may not identify the nature of the multicollinearity
- sometimes the observed behavior may **occur without multicollinearity** being present

# Variance Inflation Factor-VIF, cont'd

## Variance Inflation Factor (VIF)

- a formal method: detecting multicollinearity; widely accepted
- measure how much the variances of  $b_k$ s are inflated as compared to when the predictor variables are not linearly related.
- Illustration:
- Variance-covariance matrix of  $\mathbf{b}$ :  $\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

# Variance Inflation Factor-VIF, cont'd

Using the **standardized regression model**:

Variance-covariance matrix of  $\mathbf{b}^*$ :  $\sigma^2\{\mathbf{b}^*\} = (\sigma^*)^2 \mathbf{r}_{xx}^{-1}$

$(\sigma^*)^2$  = the error term variance for the transformed model

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

$$\Rightarrow \sigma^2\{b_k^*\} = (\sigma^*)^2 (VIF)_k = \frac{(\sigma^*)^2}{1 - R_k^2}$$

VIF for  $b_k^*$ :  $(VIF)_k = (1 - R_k^2)^{-1}$ ,  $k = 1, 2, \dots, p - 1$

$R_k^2$ :  $X_k$  is regressed on the  $p - 2$  other  $X_{k'}$ 's

- $R_k^2 = 0 \Rightarrow (VIF)_k = 1$ :  $X_k$  is **not linearly** related to  $X_{k'}$ 's
- $R_k^2 \neq 0 \Rightarrow (VIF)_k > 1$ : indicate **inflated variance** for  $b_k^*$  as a result of the **intercorrelations among the  $X$  variables**

# Variance Inflation Factor-VIF, cont'd

Perfect linear association with  $X_k \Rightarrow R_k^2 = 1 \Rightarrow VIF_k$  and  $\sigma^2\{b_k^*\}$  are unbounded

**Rule:** largest  $VIF$  value among all  $X$ s  $\Rightarrow$  as an indicator of the severity of multicollinearity

$$\max\{VIF_1, \dots, VIF_{p-1}\} > 10$$

# Variance Inflation Factor-VIF, cont'd

- If  $\overline{VIF} > 1 \Rightarrow$  serious multicollinearity problems

$$\because E \left\{ \sum_{k=1}^{p-1} (b_k^* - \beta_k^*)^2 \right\} = (\sigma^*)^2 \sum_{k=1}^p (VIF)_k$$

$\Rightarrow$  large  $\overline{VIF} \Rightarrow$  larger differences between  $b_k^*$  and  $\beta_k^*$

- If no linearly  $R_k^2 \equiv 0 \Rightarrow (VIF)_k = 1$

$$\Rightarrow E \left\{ \sum_{k=1}^{p-1} (b_k^* - \beta_k^*)^2 \right\} = (\sigma^*)^2 (p - 1)$$

$$\Rightarrow \overline{VIF} = \frac{(\sigma^*)^2 \sum_{k=1}^p (VIF)_k}{(\sigma^*)(p - 1)} = \frac{\sum_{k=1}^p (VIF)_k}{(p - 1)}$$

# Variance Inflation Factor-VIF, cont'd

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

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

Figure : VIF-Body Fat Example with three  $X$ s

- $VIF_3 = 105$
- $r_{13}^2 = 0.458^2 = 0.209764$ ,  $r_{23}^2 = 0.085^2$ : not large
- $X_3$  :  $R_3^2 = 0.990$ ; strong related to  $X_1, X_2$

# Variance Inflation Factor-VIF, cont'd

## Body Fat Example:

```
> summary(fit)
```

Call:

```
lm(formula = Y ~ X1 + X2 + X3)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.7263	-1.6111	0.3923	1.4656	4.1277

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	117.085	99.782	1.173	0.258
X1	4.334	3.016	1.437	0.170
X2	-2.857	2.582	-1.106	0.285
X3	-2.186	1.595	-1.370	0.190

Residual standard error: 2.48 on 16 degrees of freedom

Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641

F-statistic: 21.52 on 3 and 16 DF, p-value: 7.343e-06

```
> anova(fit)
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	352.27	352.27	57.2768	1.131e-06 ***
X2	1	33.17	33.17	5.3931	0.03373 *
X3	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> vif(fit)
```

X1	X2	X3
708.8429	564.3434	104.6060



# Variance Inflation Factor-VIF, cont'd

## Comments

- Some program: using  $1/VIF_k = 1 - R_k^2 < 0.01$  (0.001, 0.001)
- Limitation: cannot distinguish between several simultaneous multicollinearities
- Other methods: more complex than  $VIF$