

# Data modeling: CSCI E-106

Applied Linear Statistical Models  
Chapter 7 – Multiple Regression II

# Illustration for the extra sum of squares

**Example 1:** A study of the relation of amount of body fat: a sample of 20 healthy females: 25-34 years old

- $Y$  : body fat
- $X_1$ : triceps skinfold thickness
- $X_2$ : thigh circumference
- $X_3$ : midarm circumference

It would be very helpful if a regression model with some or all these predictor variables could provide reliable estimates of amount of body fat.

# Illustration for the extra sum of squares, cont'd

Table : Basic Data-Body Fat Example.

Subject	Triceps Skinfold Thickness	Thigh Circumference	Midarm Circumference	Body Fat
$i$	$X_{i1}$	$X_{i2}$	$X_{i3}$	$Y_i$
1	19.50	43.10	29.10	11.90
2	24.70	49.80	28.20	22.80
3	30.70	51.90	37.00	18.70
4	29.80	54.30	31.10	20.10
5	19.10	42.20	30.90	12.90
6	25.60	53.90	23.70	21.70
7	31.40	58.50	27.60	27.10
8	27.90	52.10	30.60	25.40
9	22.10	49.90	23.20	21.30
10	25.50	53.50	24.80	19.30
11	31.10	56.60	30.00	25.40
12	30.40	56.70	28.30	27.20
13	18.70	46.50	23.00	11.70
14	19.70	44.20	28.60	17.80
15	14.60	42.70	21.30	12.80
16	29.50	54.40	30.10	23.90
17	27.70	55.30	25.70	22.60
18	30.20	58.60	24.60	25.40
19	22.70	48.20	27.10	14.80
20	25.20	51.00	27.50	21.10

# Illustration for the extra sum of squares, cont'd

(a) Regression of  $Y$  on  $X_1$

(b) Regression of  $Y$  on  $X_2$

TABLE 7.2  
Regression  
Results for  
Several Fitted  
Models—Body  
Fat Example.

(a) Regression of $Y$ on $X_1$ $\hat{Y} = -1.496 + .8572X_1$			
Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	
Variable	Estimated Regression Coefficient	Estimated Standard Deviation	$t^*$
$X_1$	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66
(b) Regression of $Y$ on $X_2$ $\hat{Y} = -23.634 + .8565X_2$			
Source of Variation	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.30
Total	495.39	19	
Variable	Estimated Regression Coefficient	Estimated Standard Deviation	$t^*$
$X_2$	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

# Illustration for the extra sum of squares, cont'd

(c) Regression of  $Y$  on  $X_1$  and  $X_2$       (d) Regression of  $Y$  on  $X_1$ ,  $X_2$  and  $X_3$

TABLE 7.2  
(Continued).

(c) Regression of $Y$ on $X_1$ and $X_2$ $\hat{Y} = -19.174 + .2224X_1 + .6594X_2$			
Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	
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
(d) Regression of $Y$ on $X_1$ , $X_2$ , and $X_3$ $\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$			
Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	
Variable	Estimated Regression Coefficient	Estimated Standard Deviation	$t^*$
$X_1$	$b_1 = 4.334$	$s\{b_1\} = 3.016$	1.44
$X_2$	$b_2 = -2.857$	$s\{b_2\} = 2.582$	-1.11
$X_3$	$b_3 = -2.186$	$s\{b_3\} = 1.596$	-1.37

# Illustration for the extra sum of squares, cont'd

## Notations:

- Assume  $X_1$  is in the model
  - $SSR(X_1)$ : The regression sum of squares
  - $SSE(X_1)$ : The error sum of squares
  - measure the marginal effect of adding  $X_2$  (another variable) to the regression model when  $X_1$  is already in the model
    - $SSR(X_2|X_1)$ : The extra sum of squares gained by adding  $X_2$
- Assume  $X_1$  and  $X_2$  are in the model
  - $SSR(X_1, X_2)$ : The regression sum of squares
  - $SSE(X_1, X_2)$ : The error sum of squares
  - measure the marginal effect of adding  $X_3$  (another variable) to the regression model when  $X_1$  and  $X_2$  are already in the model
    - $SSR(X_3|X_1, X_2)$ : The extra sum of squares gained by adding  $X_3$

# Illustration for the extra sum of squares, cont'd

(a) Regression of Y on  $X_1$   
 $\hat{Y} = -1.496 + .8572X_1$

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	$t^*$
$X_1$	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

(c) Regression of Y on  $X_1$  and  $X_2$   
 $\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

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

## An extra sum of squares:

$$SSR(X_2 | X_1) = SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17$$

$$= SSR(X_1, X_2) - SSR(X_1) = 385.44 - 352.27 = 33.17$$

## $SSR(X_2 | X_1)$

- the marginal **increase** in the regression sum of squares (**SSR**)
- reflects the **additional** or **extra** reduction in the error sum of squares (**SSE**) associated with  $X_2$ , given that  $X_1$  is already included in the model

# Illustration for the extra sum of squares, cont'd

- The marginal reduction in the  $SSE$  = The marginal increase in  $SSR$
- $SSTO = SSR + SSE$ :
  - measure the variability of  $Y_i$  and does not depend on the regression model fitted
  - Any reduction in  $SSE$  implies an identical increase in  $SSR$



# Illustration for the extra sum of squares, cont'd

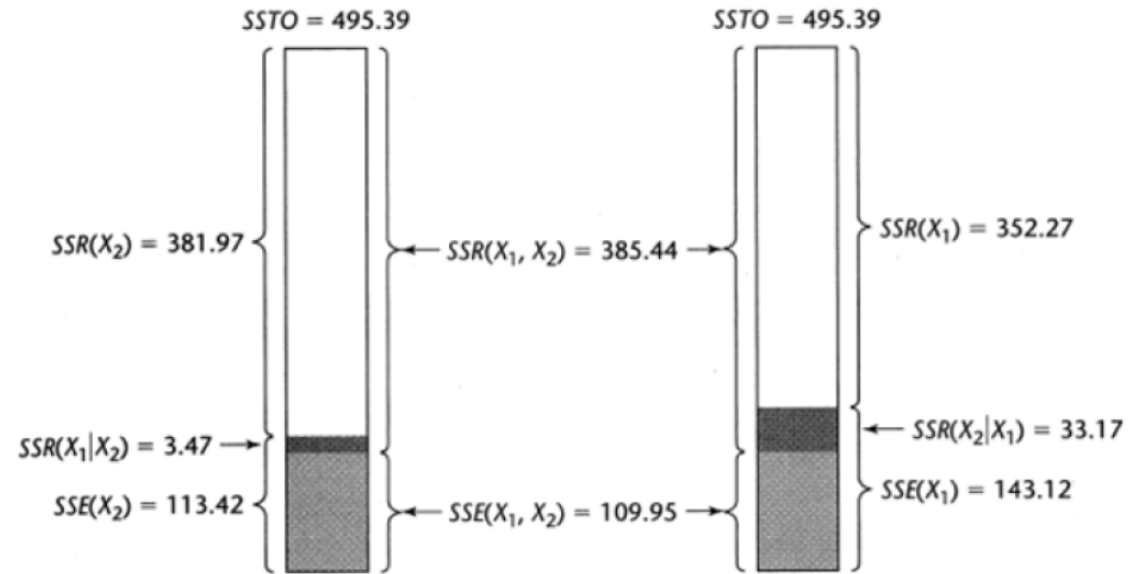


Figure : Schematic Representation of Extra Sums of Squares-Body Fat Example.

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17 \\ &= SSR(X_1, X_2) - SSR(X_1) = 385.44 - 352.27 = 33.17 \end{aligned}$$

# Illustration for the extra sum of squares, cont'd

(c) Regression of  $Y$  on  $X_1$  and  $X_2$   
 $\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

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

(d) Regression of  $Y$  on  $X_1$ ,  $X_2$ , and  $X_3$   
 $\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	$t^*$
$X_1$	$b_1 = 4.334$	$s\{b_1\} = 3.016$	1.44
$X_2$	$b_2 = -2.857$	$s\{b_2\} = 2.582$	-1.11
$X_3$	$b_3 = -2.186$	$s\{b_3\} = 1.596$	-1.37

An extra sum of squares: adding  $X_3$

$$\begin{aligned}
 SSR(X_3 | X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = 109.95 - 98.41 = 11.54 \\
 &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) = 396.98 - 385.44 = 11.54
 \end{aligned}$$

An extra sum of squares: adding  $X_2, X_3$

$$\begin{aligned}
 SSR(X_2, X_3 | X_1) &= SSE(X_1) - SSE(X_1, X_2, X_3) = 143.12 - 98.41 = 44.71 \\
 &= SSR(X_1, X_2, X_3) - SSR(X_1) = 396.98 - 352.27 = 44.71
 \end{aligned}$$

# Illustration for the extra sum of squares, cont'd

## Extra Sums of Squares

- An extra sum of squares measures the marginal decrease in the error sum of squares when one or several predictor variables are added to the regression model, given that other variables are already in the model.
- Equivalently, one can view the extra sum of squares as measuring the marginal increase in the regression sum of squares
- Extra:  $SSE \downarrow$ ;  $SSR \uparrow$

# Definitions

Extra Sums of Squares for two variables:

If  $X_1$  is the extra variable:

$$\begin{aligned} SSR(X_1 | X_2) &= SSE(X_2) - SSE(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_2) \end{aligned}$$

If  $X_2$  is the extra variable:

$$\begin{aligned} SSR(X_2 | X_1) &= SSE(X_1) - SSE(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_1) \end{aligned}$$

# Definitions, cont'd

Extra Sums of Squares for three variables:

If  $X_3$  is the extra variable:

$$\begin{aligned} SSR(X_3 | X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\ &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \end{aligned}$$

If  $X_2, X_3$  are the extra variables:

$$\begin{aligned} SSR(X_3, X_2 | X_1) &= SSE(X_1) - SSE(X_1, X_2, X_3) \\ &= SSR(X_1, X_2, X_3) - SSR(X_1) \end{aligned}$$

Extensions for more variables are straightforward and easily follow as above.

# Decomposition of SSR into Extra Sums of Squares

Consider the case of two X variables. We begin with the SSTO identity (2.50) for variable  $X_1$ :

$$SSTO = SSE(X_1) + SSR(X_1)$$

From slide 12:

$$SSR(X_2/X_1) = SSE(X_1) - SSE(X_1, X_2) \Rightarrow SSE(X_1) = SSR(X_2/X_1) + SSE(X_1, X_2)$$

Then,

$$\begin{aligned} SSTO &= SSR(X_1) + SSE(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_1) + SSR(X_1) + SSE(X_1, X_2) \\ &= SSR(X_1, X_2) + SSE(X_1, X_2) \end{aligned}$$

Also From slide 12;  $SSR(X_2, X_1) = SSR(X_1) + SSR(X_2/X_1)$

# Decomposition of SSR into Extra Sums of Squares, cont'd

Decomposition  $SSR(X_2, X_1) = SSR(X_1) + SSR(X_2/X_1)$

- $SSR(X_1)$  : measuring the **contribution** by including  $X_1$  alone in the model
- $SSR(X_2/X_1)$ : measuring the **addition contribution** when  $X_2$  is included, given that  $X_1$  is already in the model
- The order of the X variables is **arbitrary**

$$SSR(X_2, X_1) = SSR(X_2) + SSR(X_1/X_2)$$

# Decomposition of SSR into Extra Sums of Squares, cont'd

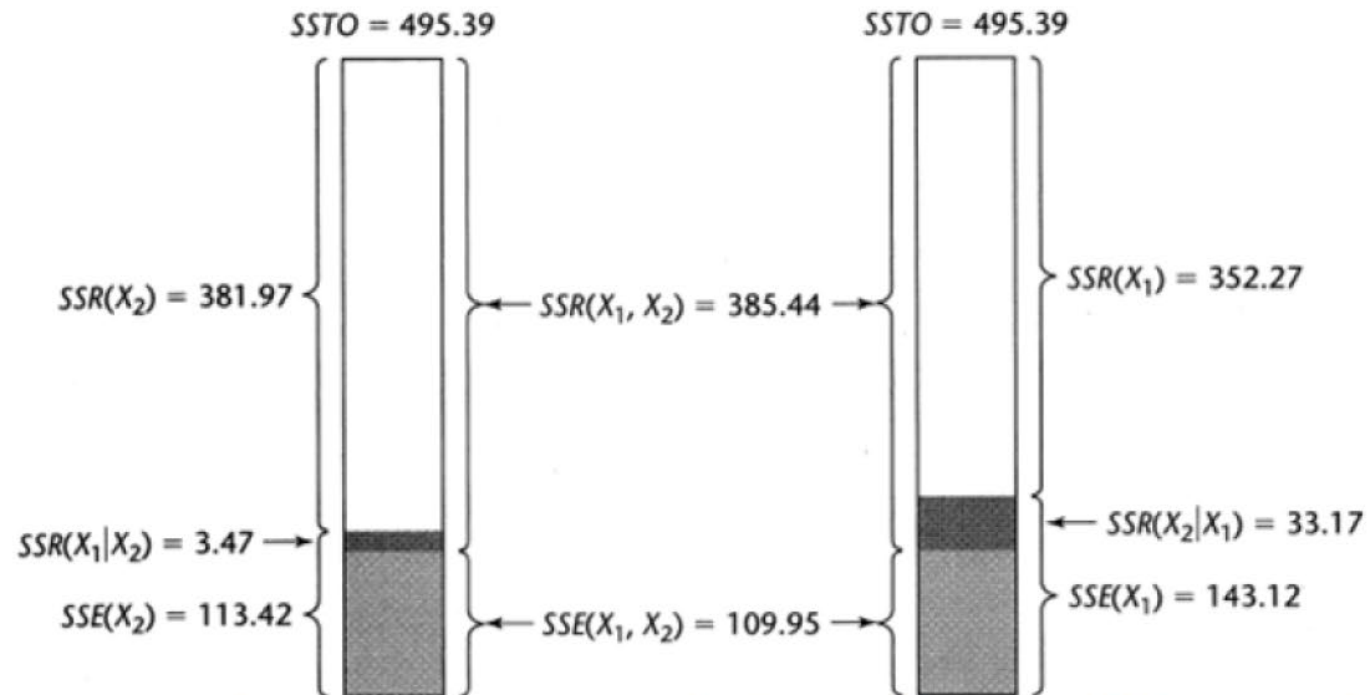


Figure : Schematic Representation of Extra Sums of Squares-Body Fat Example.



# Decomposition of SSR into Extra Sums of Squares, cont'd

When the regression model contains **three  $X$  variables**  
 **$(X_1, X_2, X_3)$** :

$$\begin{aligned} SSR(X_1, X_2, X_3) &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \\ &= SSR(X_2) + SSR(X_3|X_2) + SSR(X_1|X_2, X_3) \\ &= SSR(X_3) + SSR(X_1|X_3) + SSR(X_2|X_1, X_3) \\ &= SSR(X_1) + SSR(X_2, X_3|X_1) \end{aligned}$$

The number of possible decompositions becomes vast as **the number of  $X$  variables in the regression model increases**.

# ANOVA Table Containing Decomposition of SSR

Example of ANOVA Table With Decomposition Three X Variables.

Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
$X_1$	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 X_1$	$SSR(X_2 X_1)$	1	$MSR(X_2 X_1)$
$X_3 X_1, X_2$	$SSR(X_3 X_1, X_2)$	1	$MSR(X_3 X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n - 4$	$MSE(X_1, X_2, X_3)$
Total	$SSTO$	$n - 1$	

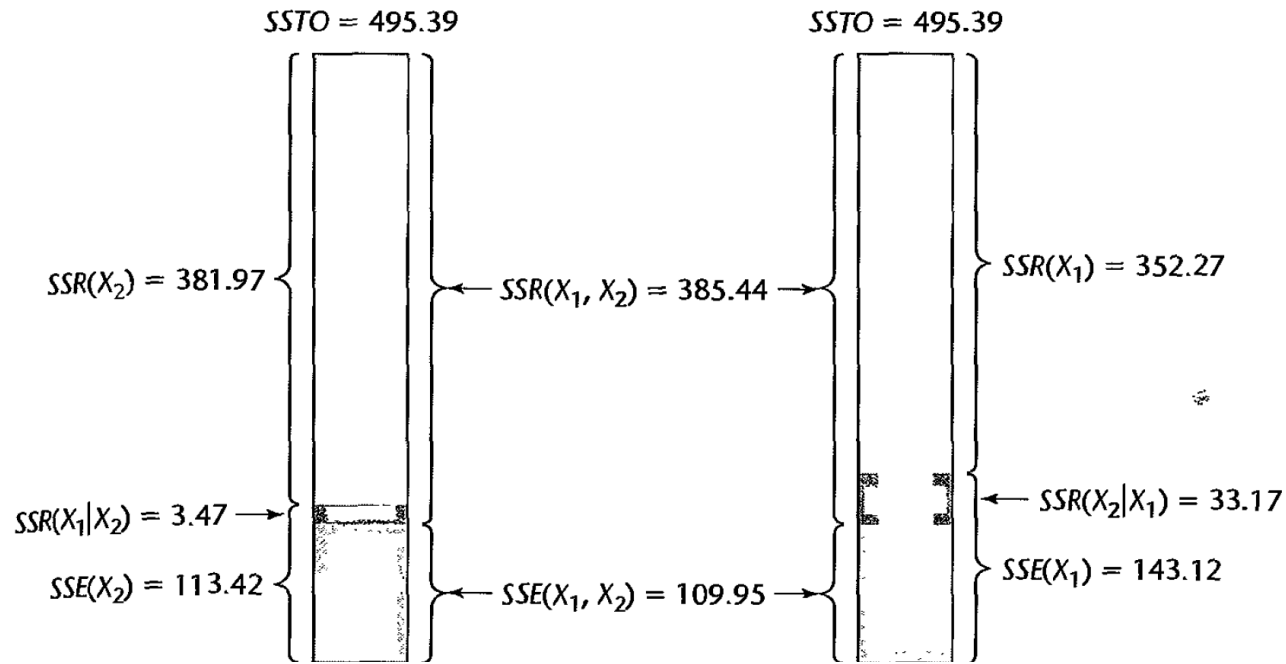
# ANOVA Table Containing Decomposition of SSR, cont'd

- Each extra sum of squares involving
  - a single extra  $X$  variable has associated with it one degree of freedom
  - two extra  $X$  variables have two degrees of freedom
- Mean squares:

$$MSR(X_2|X_1) = \frac{SSR(X_2|X_1)}{1}$$

$$MSR(X_2, X_3|X_1) = \frac{SSR(X_2, X_3|X_1)}{2}$$

# ANOVA Table Containing Decomposition of SSR, cont'd



Source of Variation	SS	df	MS
Regression	396.98	3	132.33
$X_1$	352.27	1	352.27
$X_2 X_1$	33.17	1	33.17
$X_3 X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

Extra sums of squares are of interest because they occur in a variety of tests about regression coefficients where the question of concern is whether certain  $X$  variables can be dropped from the regression model.

# Test whether a Single $\beta_k = 0$

- Test whether  $\beta_k X_k$  can be dropped from a multiple regression model

$$H_0 : \beta_k = 0$$

$$H_a : \beta_k \neq 0$$

- Test statistics in (6.51b):  $t^* = \frac{b_k}{s\{b_k\}}$
- The general linear test approach (Sec. 2.8): Full model vs. Reduced model

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

# Test whether a Single $\beta_k = 0$ , cont'd

- The general linear test approach (Sec. 2.8) involves an extra sum of squares:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad \text{Full Model}$$

- To test the alternatives:

$$H_0: \beta_3 = 0 \text{ vs. } H_a: \beta_3 \neq 0$$

- When  $H_0$  holds:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{Reduced Model}$$

- The test whether or not  $\beta_3 = 0$  is a marginal test, given  $X_1, X_2$  are already in the model

# Test whether a Single $\beta_k = 0$ , cont'd

- Steps:

1.  $SSE(F) = SSE(X_1, X_2, X_3)$ ,  $df_F = n - 4$

2.  $SSE(R) = SSE(X_1, X_2)$ ,  $df_R = n - 3$

3. The general linear test statistic (2.70):

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \\ &= \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n - 3) - (n - 4)} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} \\ &= \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} \\ &= \frac{MSR(X_3|X_1, X_2)}{MSE(X_1, X_2, X_3)} \end{aligned}$$

# Test whether a Single $\beta_k = 0$ , cont'd

**TABLE 7.4**  
ANOVA Table  
with  
Decomposition  
of *SSR*—Body  
Fat Example  
with Three  
Predictor  
Variables.

Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>
Regression	396.98	3	132.33
$X_1$	352.27	1	352.27
$X_2 X_1$	33.17	1	33.17
$X_3 X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

Body Fat Example: Testing,  $H_0: \beta_3 = 0$  vs.  $H_a: \beta_3 \neq 0$

$$F^* = \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} = \frac{11.54}{1} \div \frac{98.41}{16} = 1.88$$

$$F^* = 1.88 \leq 8.53 = F(0.99; 1, 16) \Rightarrow \text{conclude } H_0 \text{ ( } \alpha = 0.01 \text{)}$$

- $X_3$  can be dropped from the regression model that already contains  $X_1, X_2$



# Test whether a Single $\beta_k = 0$ , cont'd

- R Codes For Extra Sum Of Squares With The Body Fat Example:

```
ex <- read.table("CH07TA01.txt",header=F)
n<-length(ex$V1)
frm1 <- lm(V4~V1+V2+V3,data=ex)
frm2 <- lm(V4~V1+V2,data=ex)
SSE1 <-deviance(frm1)
SSE2 <-deviance(frm2)
F<-((SSE2-SSE1)/1)/(SSE1/(n-4))
```

# Test whether Several $\beta_k = 0$

- Test whether  $\beta_2 X_2$  and  $\beta_3 X_3$  can be dropped from the full model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad \text{Full Model}$$

- Alternative

$$H_0: \beta_2 = \beta_3 = 0$$

$$H_a: \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal } 0$$

- When  $H_0$  holds:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \quad \text{Reduced Model}$$

# Test whether Several $\beta_k = 0$

- Test statistics
  1.  $SSE(F) = SSE(X_1, X_2, X_3)$ ,  $df_F = n - 4$
  2.  $SSE(R) = SSE(X_1)$ ,  $df_R = n - 2$
  3. The general linear test statistic (2.70):

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \\ &= \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(n - 2) - (n - 4)} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} \\ &= \frac{SSR(X_2, X_3 | X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{n - 4} \\ &= \frac{MSR(X_2, X_3 | X_1)}{MSE(X_1, X_2, X_3)} \end{aligned}$$

# Test whether a Single $\beta_k = 0$ , cont'd

**TABLE 7.4**  
ANOVA Table  
with  
Decomposition  
of *SSR*—Body  
Fat Example  
with Three  
Predictor  
Variables.

Source of Variation	<i>SS</i>	<i>df</i>	<i>MS</i>
Regression	396.98	3	132.33
$X_1$	352.27	1	352.27
$X_2 X_1$	33.17	1	33.17
$X_3 X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

Can both  $X_2$  and  $X_3$  be dropped from the full model?

$$F^* = \frac{SSR(X_2, X_3|X_1)}{2} \div MSE(X_1, X_2, X_3) = \frac{33.17 + 11.54}{2} \div 6.15 = 3.63$$

$F^* = 3.63 \sim 3.63 = F(0.99; 2, 16) \Rightarrow$  at the boundary of the decision rule

We may wish to make further analyses before deciding whether  $X_2$  and  $X_3$  should be dropped from the regression model that already contains  $X_1$ .

# Comments

- Testing whether a single  $\beta_k$  equals zero:
  - 1 the  $t^*$  test statistic
  - 2 the  $F^*$  general linear test statistic
- Testing whether several  $\beta_k$  equal zero:
  - 1 the  $F^*$  general linear test statistic
- General linear test statistic can be expressed in term of the coefficients of multiple determination  $R^2$

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \\ &= \frac{R_F^2 - R_R^2}{df_R - df_F} \div \frac{1 - R_F^2}{df_F} \end{aligned}$$

# Comments, cont'd

Can both  $X_2$  and  $X_3$  be dropped from the full model?

$$F^* = \frac{0.80135 - 0.71110}{(20-2) - (20-4)} \div \frac{1-0.80135}{16} = 3.63$$

$$F^* = \frac{R_{Y|123}^2 - R_{Y|1}^2}{(n-2) - (n-4)} \div \frac{1-R_{Y|123}^2}{n-4} = 3.63$$

Test Statistics:

$$F^* = \frac{R_F^2 - R_R^2}{df_R - df_F} \div \frac{1 - R_F^2}{df_F}$$

is not appropriate when the full and reduced regression models do not contain  $\beta_0$

# Summary of Tests Concerning Regression Coefficients

- Test whether **all**  $\beta_k = 0$

$$\text{overall } F \text{ test: } F^* = \frac{MSR}{MSE} \sim F(p-1, n-p)$$

- Test whether **a single**  $\beta_k = 0$

$$\text{partial } F \text{ test: } F^* = \frac{MSR(X_k | X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1})}{MSE}$$

$$\sim F(1, n-p)$$

$$\Leftrightarrow t^* = \frac{b_k}{s\{b_k\}}$$

# Summary of Tests Concerning Regression Coefficients, cont'd

- Test whether **some**  $\beta_k = 0$

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

$$\text{partial } F \text{ test: } F^* = \frac{MSR(X_q, \dots, X_{p-1} | X_1, \dots, X_{q-1})}{MSE} \\ \sim F(p - q, n - p)$$



# Summary of Tests Concerning Regression Coefficients, cont'd

- When tests about regression coefficients are desired that do not involve testing whether one or several  $\beta_k$  equal zero, extra sums of squares cannot be used and the general linear test approach requires separate fittings of the full and reduced models.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad \text{Full Model}$$

- Wish to test:  $H_o: \beta_1 = \beta_2$  vs.  $H_a: \beta_1 \neq \beta_2$

$$Y_i = \beta_0 + \beta_c(X_{i1} + X_{i2}) + \beta_3 X_{i3} + \varepsilon_i \quad \text{Reduced Model}$$

- Wish to test:  $H_o: \beta_1 = 3, \beta_3 = 5$  vs.  $H_a: \text{not both equalities in } H_o \text{ holds}$
- Under  $H_o$ ,  $\beta_1 X_1$  and  $\beta_3 X_3$  are known constants

$$Y_i - 3X_{i1} - 5X_{i3} = \beta_0 + \beta_2 X_{i2} + \varepsilon_i \quad \text{Reduced Model}$$

# Coefficients of Partial Determination

- $R^2$ : measures the **proportionate reduction in the variation of Y** achieved by the introduction of the entire set of X considered in the model
- **Coefficient of partial determination**: measures the **marginal contribution on one X variable** when all others are already included in the model

# Coefficients of Partial Determination, cont'd

Illustration: two predictor variables

$$\text{Model } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

- $SSE(X_2)$ : measures the variation in  $Y$  when  $X_2$  is included in the model
- $SSE(X_1, X_2)$ : measures the variation in  $Y$  when  $X_1, X_2$  are included in the model
- $R^2_{Y1|2}$ : the coefficient of partial determination between  $Y$  and  $X_{i1}$ , given that  $X_2$  is in the model

$$R^2_{Y1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

# Coefficients of Partial Determination, cont'd

General Case: coefficients of partial determination to three or more  $X$  variables in the model

$$R^2_{Y1|23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)}$$

$$R^2_{Y2|13} = \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)}$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)}$$

$$R^2_{Y4|123} = \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$

# Coefficients of Partial Determination, cont'd

- Body Fat Example:

$$R_{Y2|1}^2 = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{143.12} = 0.232$$

$$R_{Y3|12}^2 = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{11.54}{109.95} = 0.105$$

$$R_{Y1|2}^2 = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = 0.031$$

# Comments

- The coefficients of partial determination can take on values between 0 and 1.
- Other interpretation: with a coefficient of simple determination
  - Residuals: regress  $Y$  on  $X_2 \Rightarrow e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$
  - Residuals: regress  $X_1$  on  $X_2 \Rightarrow e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$
  - $R^2$  between  $e_i(Y|X_2)$  and  $e_i(X_1|X_2)$  will be the same as  $R_{Y1|2}^2$
- **added variable plots** or **partial regression plots** (Chapter 10): the strength of the relationship between  $Y$  and  $X_1$  adjusted for  $X_2$

$$e_i(Y|X_2) \text{ vs. } e_i(X_1|X_2)$$

# Comments, cont'd

- Body Fat Example:

$$R_{Y_1|2}^2 = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = 0.031$$

```
ex<-read.table("CH07TA01.txt",header=F)
res1<-lm(V4~V2,data=ex)$residuals
res2<-lm(V1~V2,data=ex)$residuals
fitres<-summary(lm(res1~res2))
fitres$ r.squared
[1] 0.03061875
```

# Coefficients of Partial Correlation

- Coefficient of partial correlation: (Chapter 9)

$$r_{Y2|1} = \sqrt{R_{Y2|1}^2}$$

- the same sign with the regression coefficient
- Expressed in terms of simple or other partial correlation coefficients:

$$R_{Y2|1}^2 = [r_{Y2|1}]^2 = \frac{(r_{Y2} - r_{12}r_{Y1})^2}{(1 - r_{12}^2)(1 - r_{Y1}^2)}$$

$$R_{Y2|13}^2 = [r_{Y2|13}]^2 = \frac{(r_{Y2|3} - r_{12|3}r_{Y1|3})^2}{(1 - r_{12|3}^2)(1 - r_{Y1|3}^2)}$$

$r_{Y1}$ : correlation of  $Y$  and  $X_1$

$r_{12}$ : correlation of  $X_1$  and  $X_2$



# Standardized Multiple Regression Model

- Roundoff errors tend to enter normal equations calculations primarily when the **inverse of  $X'X$**  is taken.
  - determinant that is close to zero: some variables are highly intercorrelated
  - the element of  **$X'X$**  substantially different: the entries in  **$X'X$**  cover a wide range magnitudes

Roundoff errors  $\Rightarrow$  *standartized regression*

- Transformation: correlation transformation
  - Transformed variables fall between -1 and 1
  - becomes much less subject to roundoff errors

# Lack of Comparability in Regression Coefficients

- differences in the units:

$$\hat{Y}_i = 200 + 20000X_1 + 0.2X_2$$

- Y: dollars;  $X_1$ : thousand dollars;  $X_2$ :cents
- Is  $X_1$  the only important predicted variable ?

# Correlation Transformation

- help with controlling roundoff errors
- expressing the regression coefficients in the same units
- Y is a normal random variable  $\Rightarrow Z = \frac{Y - \mu}{\sigma}$
- **Standardizing:** involving **centering** and **scaling** the variable

# Correlation Transformation, cont'd

- The usual standardizations of the variables:

$$\frac{Y_i - \bar{Y}}{s_Y}; \quad s_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n - 1}}$$
$$\frac{X_{ik} - \bar{X}_k}{s_k}; \quad s_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n - 1}} \quad (k = 1, \dots, p - 1)$$

- The correlation transformation:

$$Y_i^* = \frac{1}{\sqrt{n - 1}} \left( \frac{Y_i - \bar{Y}}{s_Y} \right)$$
$$X_{ik}^* = \frac{1}{\sqrt{n - 1}} \left( \frac{X_{ik} - \bar{X}_k}{s_k} \right) \quad (k = 1, \dots, p - 1)$$

# Standardized Regression Model

A standardized regression model:

$$Y_i^* = \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

- no need for intercept

$$\beta_k = \left( \frac{s_Y}{s_k} \right) \beta_k^* \quad (k = 1, \dots, p-1)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \cdots - \beta_{p-1} \bar{X}_{p-1}$$

# X'X Matrix for Transformed Variables

- $\mathbf{r}_{XX}$ : correlation matrix of the  $X$  variables

$$\mathbf{r}_{XX}_{(p-1) \times (p-1)} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix}$$

- $\mathbf{r}_{YX}$ : correlation between  $Y$  and each of  $X$  variables:

$$\mathbf{r}_{YX}_{(p-1) \times 1} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix}$$

# X'X Matrix for Transformed Variables, cont'd

- The transformed variables: (no column of 1 in  $\mathbf{X}$ )

$$\mathbf{X}_{n \times (p-1)} = \begin{bmatrix} X_{11}^* & \cdots & X_{1,p-1}^* \\ X_{21}^* & \cdots & X_{2,p-1}^* \\ \vdots & & \vdots \\ X_{n1}^* & \cdots & X_{n,p-1}^* \end{bmatrix}$$
$$\Rightarrow \mathbf{X}'\mathbf{X}_{(p-1) \times (p-1)} = \mathbf{r}_{XX}$$

- All of the elements of  $\mathbf{X}'\mathbf{X}$  are between -1 and 1
- $\sum (X_{i1}^*)^2 = 1$
- $\sum X_{i1}^* X_{i2}^* = \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{[\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2]^2}$

# Estimated Standard Regression Coefficients

- the least squares estimator:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- The least squares normal equations and estimators of the regression coefficients of the standardized regression model:

$$\mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX} \Rightarrow \mathbf{b} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$$

$$\underset{(p-1) \times 1}{\mathbf{b}} = \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{p-1}^* \end{bmatrix}$$

- $b_1^*, \dots, b_{p-1}^*$ : standardized regression coefficients



# Estimated Standard Regression Coefficients, cont'd

- The standardized parameters vs. the original parameters

- $b_k = \left(\frac{s_Y}{s_k}\right) b_k^* \quad k = 1, \dots, p - 1$

- $b_o = \bar{Y} - b_1\bar{X}_1 - \dots - b_{p-1}\bar{X}_{p-1}$

- Illustration for  $p - 1 = 2$ :

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

$$\mathbf{r}_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix}$$

$$\mathbf{r}_{XX}^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

$$\mathbf{b} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} r_{Y1} - r_{12}r_{Y2} \\ r_{Y2} - r_{12}r_{Y1} \end{bmatrix}$$

$$b_1^* = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2}$$

$$b_2^* = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2}$$

# Estimated Standard Regression Coefficients, cont'd

- Dwane Studios Example:

(a) Original Data			
Case $i$	Sales $Y_i$	Target Population $X_{i1}$	Per Capita Disposable Income $X_{i2}$
1	174.4	68.5	16.7
2	164.4	45.2	16.8
...	...	...	...
20	224.1	82.7	19.1
21	166.5	52.3	16.0
	$\bar{Y} = 181.90$	$\bar{X}_1 = 62.019$	$\bar{X}_2 = 17.143$
	$s_Y = 36.191$	$s_1 = 18.620$	$s_2 = .97035$

# Estimated Standard Regression Coefficients, cont'd

- Dwane Studios Example:

(b) Transformed Data			
$i$	$Y_i^*$	$X_{i1}^*$	$X_{i2}^*$
1	-.04637	.07783	-.10205
2	-.10815	-.20198	-.07901
...	...	...	...
20	.26070	.24835	.45100
21	-.09518	-.11671	-.26336

(c) Fitted Standardized Model	
$\hat{Y}^* = .7484X_1^* + .2511X_2^*$	

$$\hat{Y}^* = 0.7484X_1^* + 0.2511X_2^*$$

$$\hat{Y} = -68.860 + 1.455X_1 + 9.365X_2$$

# Estimated Standard Regression Coefficients, cont'd

## Ex p277

```
library(QuantPsyc)
```

```
ex7.5<-read.table("CH07TA05.txt")
```

```
fit<-lm(V1~V2+V3,data=ex7.5)
```

```
fit
```

Call:

```
lm(formula = V1 ~ V2 + V3, data = ex7.5)
```

Coefficients:

(Intercept)	V2	V3
-68.857	1.455	9.366

```
lm.beta(fit)
```

V2	V3
0.7483670	0.2511039

# Estimated Standard Regression Coefficients, cont'd

$$\hat{Y}^* = 0.7484X_1^* + 0.2511X_2^*$$

- Does  $X_1$  have much greater impact on sales than  $X_2$ ? ( $\because b_1^* > b_2^*$ )
- One must be **cautious about interpreting any regression coefficient** whether standardized or not.
  - caution if the predictor variables
  - $r_{12} = 0.781$  in the Dwaine Studios data

# Estimated Standard Regression Coefficients, cont'd

To shift from the standardized regression coefficients  $b_1^*$  and  $b_2^*$  back to the regression coefficients for the model with the original variables:

$$b_1 = \left( \frac{s_Y}{s_1} \right) b_1^* = \frac{36.191}{18.620} \times 0.7484 = 1.4546$$

$$b_2 = \left( \frac{s_Y}{s_2} \right) b_2^* = \frac{36.191}{0.97035} \times 0.2511 = 9.3652$$

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 = 181.90 - 1.45 \times 62.02 - 9.36 \times 17.14 = -68.86$$

$$\hat{Y} = -68.86 + 1.455X_1 + 9.365X_2$$

# Multicollinearity and Its Effects

## Questions:

- What is the relative importance of the effects of the different predictor variables?
- What is the magnitude of the effect of a given predictor variable on the response variable?
- Can any predictor variable be dropped from the model because it has little or no effect on the response variable?
- Should any predictor variables not yet included in the model be considered for possible inclusion?
- intercorrelation or multicollinearity: the predictor variables are correlated among themselves

# Uncorrelated Predicted Variables

Example:  $Y$  - crew productivity;  $X_1$ -the effect of work crew size;  $X_2$ -level of bonus pay

- $r_{12}^2 = 0 \Rightarrow$  the predictor variables are uncorrelated
- $SSR(X_1|X_2) = 231.125 = SSR(X_1)$
- $SSR(X_2|X_1) = 171.125 = SSR(X_2)$

---

Case $i$	Crew Size $X_{i1}$	Bonus Pay (dollars) $X_{i2}$	Crew Productivity $Y_i$
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60



# Uncorrelated Predicted Variables, cont'd

Example:  $Y$  - crew productivity;  $X_1$ -the effect of work crew size;  $X_2$ -level of bonus pay

**TABLE 7.7**  
Regression  
Results when  
Predictor  
Variables Are  
Uncorrelated—  
Work Crew  
Productivity  
Example.

(a) Regression of $Y$ on $X_1$ and $X_2$			
$\hat{Y} = .375 + 5.375X_1 + 9.250X_2$			
Source of Variation	SS	df	MS
Regression	402.250	2	201.125
Error	17.625	5	3.525
Total	419.875	7	

(b) Regression of $Y$ on $X_1$			
$\hat{Y} = 23.500 + 5.375X_1$			
Source of Variation	SS	df	MS
Regression	231.125	1	231.125
Error	188.750	6	31.458
Total	419.875	7	

(c) Regression of $Y$ on $X_2$			
$\hat{Y} = 27.250 + 9.250X_2$			
Source of Variation	SS	df	MS
Regression	171.125	1	171.125
Error	248.750	6	41.458
Total	419.875	7	

# Uncorrelated Predicted Variables, cont'd

- When two or more predictor variables are **uncorrelated**, the **marginal contribution of one predictor variable in reducing the error sum of squares** the other predictor variables are in the model is exactly the **same as** when **this predictor variable is in the model alone**.

$$b_1 = \frac{\frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} - \left[ \frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2} \right]^{1/2} r_{Y2}r_{12}}{1 - r_{12}^2} \Rightarrow b_1 = \frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} \quad \text{when } r_{12} = 0$$

# Multicollinearity and Its Effects

Predictor variables are perfectly correlated:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Figure : Example of Perfectly Correlated Predictor Variables.

**TABLE 7.8**  
Example of  
Perfectly  
Correlated  
Predictor  
Variables.

Case <i>i</i>	$X_{i1}$	$X_{i2}$	$Y_i$	Fitted Values for Regression Function	
				(7.58)	(7.59)
1	2	6	23	23	23
2	8	9	83	83	83
3	6	8	63	63	63
4	10	10	103	103	103

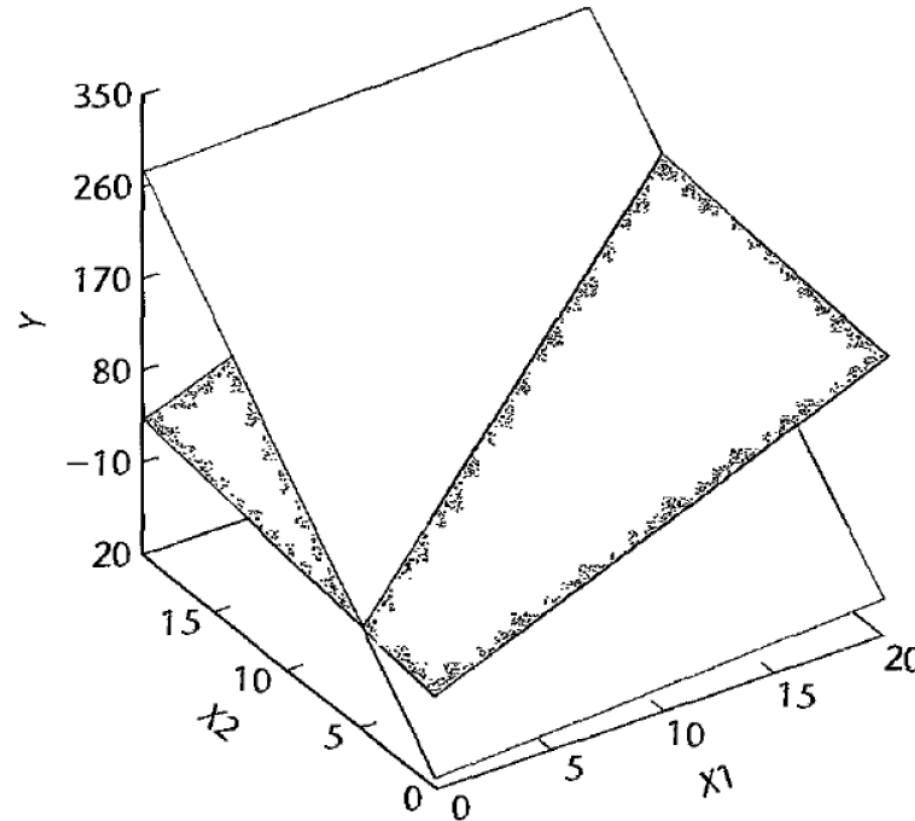
Response Functions:

$$\hat{Y} = -87 + X_1 + 18X_2 \quad (7.58)$$
$$\hat{Y} = -7 + 9X_1 + 2X_2 \quad (7.59)$$

# Multicollinearity and Its Effects, cont'd

Figure : Two Response Planes That Intersect when  $X_2 = 5 + 0.5X_1$ .

**FIGURE 7.2**  
**Two Response**  
**Planes That**  
**Intersect when**  
 $X_2 = 5 + .5X_1$ .



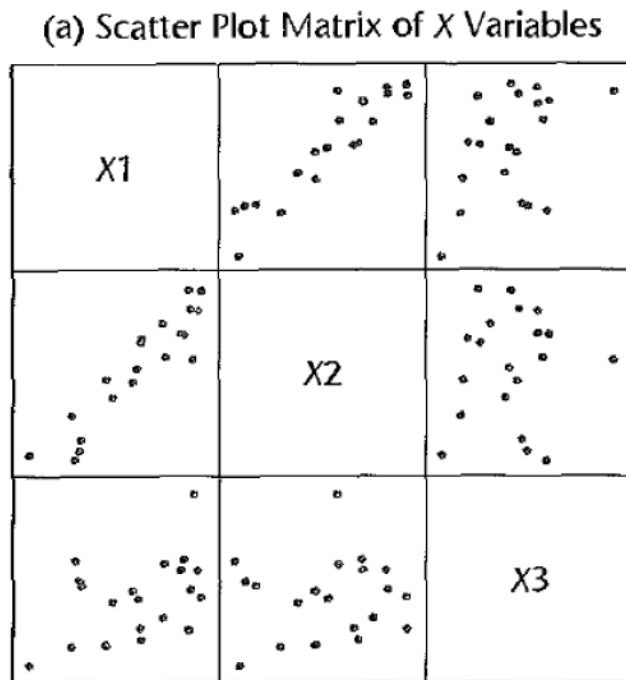
# Multicollinearity and Its Effects, cont'd

- When  $X_1$  and  $X_2$  are perfectly correlated, many different response functions will lead to the same perfectly fitted values for the observations.
- The perfect relation between  $X_1$  and  $X_2$  did not inhibit the ability to obtain a good fit to the data.
- Since many different response functions provide the same good fit, we cannot interpret any one set of regression coefficients as reflecting the effects of the different predictor variables.

# Multicollinearity and Its Effects, cont'd

**Figure :** Scatter Plot Matrix and Correlation Matrix of the Predictor Variables-Body Fat Example.

**FIGURE 7.3**  
Scatter Plot  
Matrix and  
Correlation  
Matrix of the  
Predictor  
Variables—  
Body Fat  
Example.



(b) Correlation Matrix of X Variables

$$r_{XX} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

# Multicollinearity and Its Effects, cont'd

Effects on Regression Coefficients:  $X_1$ , triceps skinfold thickness, varies markedly depending on which other variables are included in the model:

Variables in Model	$b_1$	$b_2$
$X_1$	.8572	—
$X_2$	—	.8565
$X_1, X_2$	.2224	.6594
$X_1, X_2, X_3$	4.334	−2.857

- The story is the same for the regression coefficient for  $X_2$  : the regression coefficient  $b_2$  changes sign when  $X_3$  is added to the model that includes  $X_1$  and  $X_2$

# Multicollinearity and Its Effects, cont'd

Effects on  $s\{b_k\}$

Variables in Model	$s\{b_1\}$	$s\{b_2\}$
$X_1$	.1288	—
$X_2$	—	.1100
$X_1, X_2$	.3034	.2912
$X_1, X_2, X_3$	3.016	2.582

- The **high degree of multicollinearity** among the predictor variables is responsible for the **inflated** variability of the estimated regression coefficients.



# Multicollinearity and Its Effects, cont'd

Effects on fitted values and predictions

<b>Variables in Model</b>	<b><i>MSE</i></b>
$X_1$	7.95
$X_1, X_2$	6.47
$X_1, X_2, X_3$	6.15

- Estimated means and Predicted values are not affected

# Multicollinearity and Its Effects, cont'd

## Effects on the test statistics

- It is possible that when individual t tests are performed, neither  $\beta_1$  or  $\beta_2$  is significant.
- However, when the F test is performed for both  $\beta_1$  and  $\beta_2$ , the results may still be significant.
- Need for more powerful diagnostics for multicollinearity