Regression 7

Multiple Linear Regression II

1 Introduction

Let us consider the following two models:

Model R: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

Model F: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

We get "Model F" by adding one more predictor (X_2) to the "Model R." We call Model F *full model* and Model R *reduced Model*. When we wish to test whether the term $\beta_2 X_2$ can be dropped from the full model, we can do the test:

$$H_0 : \beta_2 = 0 \quad \text{(Model R)} \quad \text{versus} \quad H_1 : \beta_2 \neq 0 \quad \text{(Model F)}.$$

We will show that this test involves the differences between SSE of the reduced model and that of the full model.

Usually the smaller SSE is desirable (equivalently the larger SSR is desirable). Hence, we are of particular interest in the reduction of SSE after adding predictor(s) to the given regression model (Model R). We use the following notation:

$$\begin{aligned}
\mathrm{SSE}(X_1) &= \mathrm{SSE} \text{ when } X_1 \text{ only is in the model} \\
\mathrm{SSE}(X_1, X_2) &= \mathrm{SSE} \text{ when both } X_1 \text{ and } X_2 \text{ are in the model} \\
&= \mathrm{SSE}(X_2, X_1)
\end{aligned}$$

$$\begin{aligned}
\mathrm{SSR}(X_1) &= \mathrm{SSR} \text{ when } X_1 \text{ only is in the model} \\
\mathrm{SSR}(X_1, X_2) &= \mathrm{SSR} \text{ when both } X_1 \text{ and } X_2 \text{ are in the model} \\
&= \mathrm{SSR}(X_2, X_1)
\end{aligned}$$

$$\begin{aligned}
\mathrm{SSR}(X_2|X_1) &= \mathrm{SSR}(X_1, X_2) - \mathrm{SSR}(X_1) \\
&= \text{Increase in SSR when } X_2 \text{ is added to a model} \\
&\quad\ \text{involving only } X_1 \text{ and the intercept.} \\
&= \mathrm{SSE}(X_1) - \mathrm{SSE}(X_1, X_2) \\
&= \text{Reduction of SSE when } X_2 \text{ is added to a model} \\
&\quad\ \text{involving only } X_1 \text{ and the intercept.} \\
&= \text{Extra sum of squares} \\
&\neq \mathrm{SSR}(X_1|X_2) \text{ in general}
\end{aligned}$$

Remark 7.1.

1. As we add more predictors, SSE never increases.
2. SSTo $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ does not depend on the regression model fitted. However, we can think of SSTo as SSE when the null model (intercept β_0 only) is used.
3. If SSE never increases as more predictors are added, why we do not include all the possible predictors.
   (a) Parsimony principle:
   Given two models that perform almost equally well in terms of prediction, one should choose the model that is more parsimonious (simple).

(b) Prediction principle:

The model should give predictions that are as accurate as possible, not just for current observation, but for future observations as well.

$\triangle$

2 ANOVA Results

Let us consider the following two models:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k_1} X_{k_1} + \epsilon \qquad \text{(Reduced)}$$

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k_1} X_{k_1} + \beta_{k_1+1} X_{k_1+1} + \cdots + \beta_{k_1+k_2} X_{k_1+k_2} + \epsilon \qquad \text{(Full)}$$

These models can be expressed in matrix notation. We need to partition the $\mathbf{X}$ matrix as

$$\underset{n\times p}{\mathbf{X}} = [\underset{n\times 1}{\mathbf{1}}, \underset{n\times k_1}{\mathbf{X}_A}, \underset{n\times k_2}{\mathbf{X}_B}],$$

where $p = 1 + k_1 + k_2$, $\mathbf{X}_A = [X_1, \ldots, X_{k_1}]$, and $\mathbf{X}_B = [X_{k_1+1}, \ldots, X_{k_1+k_2}]$.

Similarly, let us partition $\boldsymbol{\beta}$ as

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix},$$

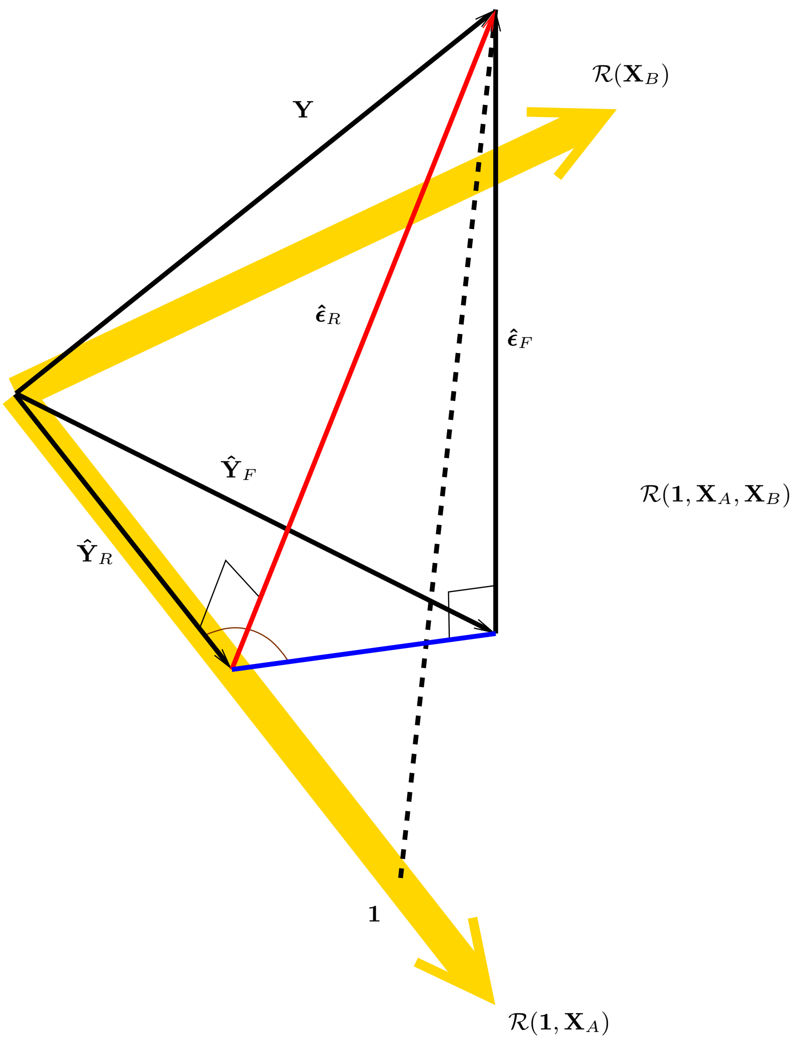
where $\underset{k_1\times 1}{\boldsymbol{\beta}_A} = [\beta_1, \ldots, \beta_{k_1}]'$, and $\underset{k_2\times 1}{\boldsymbol{\beta}_B} = [\beta_{k_1+1}, \ldots, \beta_{k_1+k_2}]'$. The partitioned vectors $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$ are the vectors of parameters corresponding to $\mathbf{X}_A$ and $\mathbf{X}_B$ respectively.

1. Reduced model.

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}_A \boldsymbol{\beta}_A + \boldsymbol{\epsilon}$$

$$\hat{\mathbf{Y}}_R = [\mathbf{1}, \mathbf{X}_A]\Big([\mathbf{1}, \mathbf{X}_A]'[\mathbf{1}, \mathbf{X}_A]\Big)^{-1}[\mathbf{1}, \mathbf{X}_A]'\mathbf{Y}.$$

Projection of $\mathbf{Y}$ onto the reduced and full space.



2. Full model.

   The regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}_A \boldsymbol{\beta}_A + \mathbf{X}_B \boldsymbol{\beta}_B + \boldsymbol{\epsilon}$$

$$\hat{\mathbf{Y}}_F = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

From the Pythagorean triangle, it follows that

$$\|\mathbf{Y} - \hat{\mathbf{Y}}_R\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}_F\|^2 + \|\hat{\mathbf{Y}}_F - \hat{\mathbf{Y}}_R\|^2.$$

Hence, we have

$$\begin{aligned}\|\hat{\mathbf{Y}}_F - \hat{\mathbf{Y}}_R\|^2 &= \|\mathbf{Y} - \hat{\mathbf{Y}}_R\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}_F\|^2 \\ &= \mathrm{SSE}(\mathbf{X}_A) - \mathrm{SSE}(\mathbf{X}_A, \mathbf{X}_B) \\ &= \mathrm{SSR}(\mathbf{X}_A, \mathbf{X}_B) - \mathrm{SSR}(\mathbf{X}_A).\end{aligned}$$

We will denote $\|\hat{\mathbf{Y}}_F - \hat{\mathbf{Y}}_R\|^2$ as $\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A)$, that is,

$$\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A) = \mathrm{SSR}(\mathbf{X}_A, \mathbf{X}_B) - \mathrm{SSR}(\mathbf{X}_A) = \mathrm{SSE}(\mathbf{X}_A) - \mathrm{SSE}(\mathbf{X}_A, \mathbf{X}_B).$$

This is the <u>increase</u> in SSR when $\mathbf{X}_B$ are added to a model involving only $\mathbf{X}_A$ and the intercept, *or*, the <u>decrease</u> in SSE when $\mathbf{X}_B$ are added to a model involving only $\mathbf{X}_A$ and the intercept.

Example 7.1. Decomposition of SSTo and SSE.

1. Let $\mathbf{X}_A = X_1$ and $\mathbf{X}_B = X_2$.

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

2. Let $\mathbf{X}_A = X_2$ and $\mathbf{X}_B = X_1$.

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

3. Let $\mathbf{X}_A = [X_1, X_2]$ and $\mathbf{X}_B = X_3$.

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

4. Let $\mathbf{X}_A = [X_1, X_2, \ldots, X_{k-1}]$ and $\mathbf{X}_B = X_k$.

$$\begin{aligned}\mathrm{SSR}(X_k|X_1, \ldots, X_{k-1}) &= \mathrm{SSR}(X_1, \ldots, X_k) - \mathrm{SSR}(X_1, \ldots, X_{k-1}) \\ &= \mathrm{SSE}(X_1, \ldots, X_{k-1}) - \mathrm{SSE}(X_1, \ldots, X_k).\end{aligned}$$

Using

$$\mathrm{SSR}(X_1,\ldots,X_k)=\mathrm{SSR}(X_1,\ldots,X_{k-1})+\mathrm{SSR}(X_k|X_1,\ldots,X_{k-1}),$$

we have

$$\begin{aligned}
\mathrm{SSTo} &= \mathrm{SSR}(X_1,\ldots,X_k)+\mathrm{SSE}(X_1,\ldots,X_k)\\
&= \mathrm{SSR}(X_1,\ldots,X_{k-1})+\mathrm{SSR}(X_k|X_1,\ldots,X_{k-1})+\mathrm{SSE}(X_1,\ldots,X_k)\\
&= \mathrm{SSR}(X_1)+\mathrm{SSR}(X_2|X_1)+\mathrm{SSR}(X_3|X_1,X_2)+\cdots\\
&\quad +\mathrm{SSR}(X_k|X_1,\ldots,X_{k-1})+\mathrm{SSE}(X_1,\ldots,X_k).
\end{aligned}$$

ANOVA decomposition of $\mathrm{SSR}(X_1,\ldots,X_k)$: Sequential F-test.

Source	SS	df
Regression	$\mathrm{SSR}(X_1,\ldots,X_k)$	k
1. X_1	$\mathrm{SSR}(X_1)$	1
2. $X_2\|X_1$	$\mathrm{SSR}(X_2\|X_1)$	1
3. $X_3\|X_1,X_2$	$\mathrm{SSR}(X_3\|X_1,X_2)$	1
$\vdots$	$\vdots$	$\vdots$
k. $X_k\|X_1,\ldots,X_{k-1}$	$\mathrm{SSR}(X_k\|X_1,\ldots,X_{k-1})$	1
Error	$\mathrm{SSE}(X_1,\ldots,X_k)$	$n-(k+1)$
Total	SSTo	$n-1$

‖

Example 7.2. Body Fat Example in Table 7.1 on Page 257 of Kutner et al. (2005).

Minitab

Read Data

```
MTB > read c1 c2 c3 c11 ;
SUBC>     file "S:\LM\CH07TA01.txt" .

MTB > name c1  'X1'
MTB > name c2  'X2'
MTB > name c3  'X3'
MTB > name c11 'Y'
```

Model: $Y=\beta_0+\beta_1X_1+\epsilon$

```
MTB > regr c11 1 c1

Regression Analysis: Y versus X1

The regression equation is
Y = - 1.50 + 0.857 X1

Predictor    Coef  SE Coef      T      P
Constant   -1.496    3.319  -0.45  0.658
X1         0.8572   0.1288   6.66  0.000

S = 2.81977   R-Sq = 71.1%   R-Sq(adj) = 69.5%

```

```
Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  352.27  352.27  44.30  0.000
Residual Error  18  143.12    7.95
Total           19  495.39

Unusual Observations
Obs    X1       Y     Fit  SE Fit  Residual  St Resid
  3  30.7  18.700  24.820   0.938    -6.120    -2.30R
R denotes an observation with a large standardized residual.
```

Model: $Y = \beta_0 + \beta_2 X_2 + \epsilon$

```
MTB > regr c11 1 c2

Regression Analysis: Y versus X2

The regression equation is
Y = - 23.6 + 0.857 X2

Predictor     Coef  SE Coef      T      P
Constant   -23.634    5.657  -4.18  0.001
X2          0.8565   0.1100   7.79  0.000

S = 2.51024   R-Sq = 77.1%   R-Sq(adj) = 75.8%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  381.97  381.97  60.62  0.000
Residual Error  18  113.42    6.30
Total           19  495.39
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (X_1 first)

```
MTB > regr c11 2 c1 c2

Regression Analysis: Y versus X1, X2

The regression equation is
Y = - 19.2 + 0.222 X1 + 0.659 X2

Predictor     Coef  SE Coef      T      P
Constant   -19.174    8.361  -2.29  0.035
X1          0.2224   0.3034   0.73  0.474
X2          0.6594   0.2912   2.26  0.037

S = 2.54317   R-Sq = 77.8%   R-Sq(adj) = 75.2%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  385.44  192.72  29.80  0.000
Residual Error  17  109.95    6.47
Total           19  495.39

Source  DF  Seq SS
X1       1  352.27
X2       1   33.17
```

Model: $Y = \beta_0 + \beta_2 X_2 + \beta_1 X_1 + \epsilon$ (X_2 first)

```
MTB > regr c11 2 c2 c1

Regression Analysis: Y versus X2, X1

The regression equation is
Y = - 19.2 + 0.659 X2 + 0.222 X1

Predictor     Coef  SE Coef      T      P
Constant   -19.174    8.361  -2.29  0.035
X2          0.6594   0.2912   2.26  0.037
```

```
X1           0.2224   0.3034   0.73  0.474

S = 2.54317   R-Sq = 77.8%   R-Sq(adj) = 75.2%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  385.44  192.72  29.80  0.000
Residual Error  17  109.95    6.47
Total           19  495.39

Source  DF  Seq SS
X2       1  381.97
X1       1    3.47
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

```
MTB > regr c11 3 c1 c2 c3

Regression Analysis: Y versus X1, X2, X3

The regression equation is
Y = 117 + 4.33 X1 - 2.86 X2 - 2.19 X3

Predictor    Coef  SE Coef      T      P
Constant   117.08    99.78   1.17  0.258
X1          4.334    3.016   1.44  0.170
X2         -2.857    2.582  -1.11  0.285
X3         -2.186    1.595  -1.37  0.190

S = 2.47998   R-Sq = 80.1%   R-Sq(adj) = 76.4%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       3  396.98  132.33  21.52  0.000
Residual Error  16   98.40    6.15
Total           19  495.39

Source  DF  Seq SS
X1       1  352.27
X2       1   33.17
X3       1   11.55
```

R

Ⓡ Read Data

```
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH07TA01.txt")
>
> x1 = mydata[,1]
> x2 = mydata[,2]
> x3 = mydata[,3]
> y  = mydata[,4]
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

```
> LM1 = lm ( y ~ x1 )
> summary(LM1)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1195 -2.1904  0.6735  1.9383  3.8523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -1.4961     3.3192  -0.451    0.658
x1            0.8572     0.1288   6.656 3.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-Squared: 0.7111, Adjusted R-squared: 0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06

> anova(LM1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ⓡ Model: $Y = \beta_0 + \beta_2 X_2 + \epsilon$

```
> LM2 = lm ( y ~ x2 )
> summary(LM2)

Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4949 -1.5671  0.1241  1.3362  4.4084

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
x2            0.8565     0.1100   7.786   3.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-Squared: 0.771,  Adjusted R-squared: 0.7583
F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07

> anova(LM2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x2         1 381.97  381.97  60.617 3.6e-07 ***
Residuals 18 113.42    6.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ⓡ Model:$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (X_1 first)

```
> LM3 = lm ( y ~ x1 + x2 )
> summary(LM3)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9469 -1.8807  0.1678  1.3367  4.0147


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
x1            0.2224     0.3034   0.733   0.4737
x2            0.6594     0.2912   2.265   0.0369 *
---
```

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

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-Squared: 0.7781, Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06


> anova(LM3)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 352.27  352.27 54.4661 1.075e-06 ***
x2         1  33.17   33.17  5.1284    0.0369 *
Residuals 17 109.95    6.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ⓡ Model: $Y = \beta_0 + \beta_2 X_2 + \beta_1 X_1 + \epsilon$ (X_2 first)

```
> LM4 = lm ( y ~ x2 + x1 )
> summary(LM4)

Call:
lm(formula = y ~ x2 + x1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9469 -1.8807  0.1678  1.3367  4.0147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
x2            0.6594     0.2912   2.265   0.0369 *
x1            0.2224     0.3034   0.733   0.4737
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-Squared: 0.7781, Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

> anova(LM4)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x2         1 381.97  381.97  59.057 6.281e-07 ***
x1         1   3.47    3.47   0.537    0.4737
Residuals 17 109.95    6.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

```
> LM5 = lm ( y ~ x1 + x2 + x3 )
> summary(LM5)

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(LM5)
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
```

3 Tests concerning regression parameters

Theorem 7.1 (Fundamental Theorem of ANOVA).

Suppose that the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *with* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ *is true and we partition* $\mathbf{X}$ *and* $\boldsymbol{\beta}$ *as*

$$\underset{n\times p}{\mathbf{X}} = \left[\ \underset{n\times 1}{\mathbf{1}},\ \underset{n\times k_1}{\mathbf{X}_A},\ \underset{n\times k_2}{\mathbf{X}_B}\ \right] \quad \text{and} \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \underset{k_1\times 1}{\boldsymbol{\beta}_A} \\ \underset{k_2\times 1}{\boldsymbol{\beta}_B} \end{bmatrix}.$$

Then we have the following results:

(a) $\dfrac{\mathrm{SSE}(\mathbf{X}_A, \mathbf{X}_B)}{\sigma^2} \sim \chi^2_{n-p}$

(b) $\dfrac{\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A)}{\sigma^2} \sim \chi^2_{k_2}$ *under* $H_0 : \boldsymbol{\beta}_B = \mathbf{0}$.

Proof. See Cochran (1934) or Appendix VI of Scheffé (1959). □

Notice that $\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A)$ is independent of $\mathrm{SSE}(\mathbf{X}_A, \mathbf{X}_B)$. Thus, we have

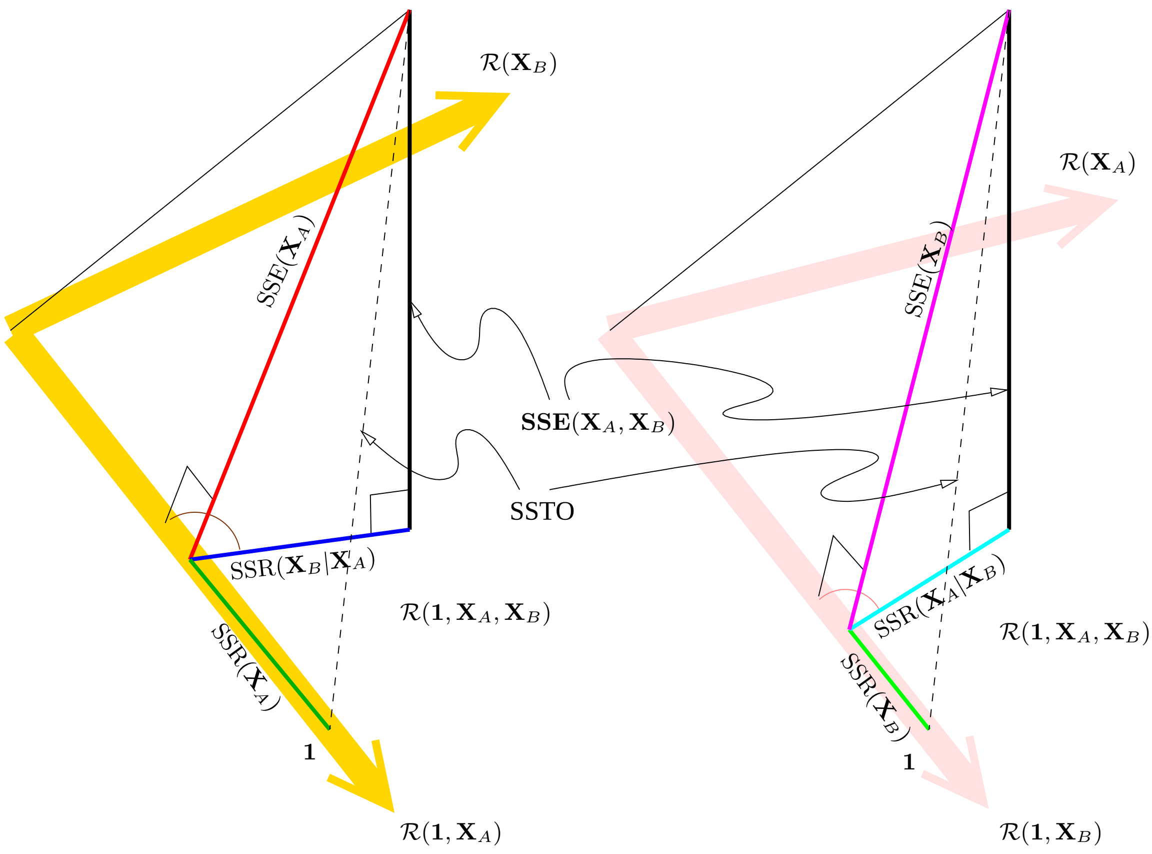
$$F = \frac{\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A)/k_2}{\mathrm{SSE}(\mathbf{X}_A, \mathbf{X}_B)/(n-p)} \sim F(k_2, n-p).$$

$$F = \frac{\dfrac{\Delta \text{SSR}}{\Delta \text{df}}}{\dfrac{\text{SSE (full)}}{\text{df (full)}}} = \frac{\dfrac{\Delta \text{SSE}}{\Delta \text{df}}}{\dfrac{\text{SSE (full)}}{\text{df (full)}}} = \frac{\dfrac{\text{SSE (reduced)} - \text{SSE (full)}}{\text{df (reduced)} - \text{df (full)}}}{\dfrac{\text{SSE (full)}}{\text{df (full)}}}$$

ANOVA decomposition

$\mathbf{X}_A$ first			$\mathbf{X}_B$ first		
Source	SS	df			
A1. $\mathbf{X}_A$	$\text{SSR}(\mathbf{X}_A)$	k_1	B1. $\mathbf{X}_B$	$\text{SSR}(\mathbf{X}_B)$	k_2
A2. $\mathbf{X}_B \mid \mathbf{X}_A$	$\text{SSR}(\mathbf{X}_B \mid \mathbf{X}_A)$	k_2	B2. $\mathbf{X}_A \mid \mathbf{X}_B$	$\text{SSR}(\mathbf{X}_A \mid \mathbf{X}_B)$	k_1
A3. Error	$\text{SSE}(\mathbf{X}_A, \mathbf{X}_B)$	$n-p$	B3. Error	$\text{SSE}(\mathbf{X}_A, \mathbf{X}_B)$	$n-p$
Total	SSTo	$n-1$	Total	SSTo	$n-1$

ANOVA decomposition



SSTO	$\text{SSR}(\mathbf{X}_A)$	$\text{SSR}(\mathbf{X}_B \mid \mathbf{X}_A)$	$\mathbf{SSE}(\mathbf{X}_A, \mathbf{X}_B)$
SSTO	$\text{SSR}(\mathbf{X}_B)$	$\text{SSR}(\mathbf{X}_A \mid \mathbf{X}_B)$	$\mathbf{SSE}(\mathbf{X}_B, \mathbf{X}_A)$

3.1 Overall F test

This tests the significance of all predictors at once.

$$H_0 : \boldsymbol{\beta^*} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta^*} \neq \mathbf{0},$$

where $\boldsymbol{\beta^*} = [\beta_1, \beta_2, \ldots, \beta_k]'$ and $k = p - 1$. That is

$$\begin{aligned} &H_0 : Y = \beta_0 + \epsilon && \text{(no predictors model)} \\ &H_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon && \text{(full model)} \end{aligned}$$

The ANOVA decomposition is

$$\begin{aligned} \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2 &= \|\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \\ \text{SSTo} \quad &= \quad \text{SSR} \quad + \quad \text{SSE} \end{aligned}$$

ANOVA decomposition

Source	SS	df	F
Regression$(X_1, \ldots, X_k)$	SSR	k	$F = \frac{\text{SSR}/k}{\text{SSE}/(n-p)}$
Error	SSE	$n-p$	
Total	SSTo	$n-1$	

Decision rule: $F \sim F(k, n-p)$ under H_0.

If $F \leq F(1-\alpha; k, n-p)$, *conclude* H_0.

If $F > F(1-\alpha; k, n-p)$, *conclude* H_1.

If H_0 is rejected, we know that at least one of the parameters $\beta_1, \ldots, \beta_k$ is non-zero. But we don't know which one(s).

3.2 Partial F test

This tests the significance of a group of additional predictors, say

$$\mathbf{X}_B = [X_{k_1+1}, \ldots, X_{k_1+k_2}] \qquad \text{(last } k_2 \text{ predictors)}$$

given that the rest

$$\mathbf{X}_A = [X_1, X_2, \ldots, X_{k_1}] \qquad \text{(first } k_1 \text{ predictors)}$$

are already in the model. We want to test

$$\boxed{H_0 : \boldsymbol{\beta}_B = \mathbf{0} \ \ \text{versus} \ \ H_1 : \boldsymbol{\beta}_B \neq \mathbf{0}.}$$

That is

$$\begin{aligned} H_0 &: \mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}_A \boldsymbol{\beta}_A + \boldsymbol{\epsilon} && \text{(reduced)} \\ H_1 &: \mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}_A \boldsymbol{\beta}_A + \underline{\mathbf{X}_B \boldsymbol{\beta}_B} + \boldsymbol{\epsilon} && \text{(full model)}, \end{aligned}$$

equivalently

$$\boxed{\begin{aligned} H_0 &: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k_1} X_{k_1} + \epsilon \\ H_1 &: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k_1} X_{k_1} + \beta_{k_1+1} X_{k_1+1} + \cdots + \beta_{k_1+k_2} X_{k_1+k_2} + \epsilon \end{aligned}}$$

The ANOVA decomposition is

$$\begin{aligned} \|\mathbf{Y} - \hat{\mathbf{Y}}_R\|^2 &= \|\hat{\mathbf{Y}}_F - \hat{\mathbf{Y}}_R\|^2 \ + \|\mathbf{Y} - \hat{\mathbf{Y}}_F\|^2 \\ \text{SSE}(\mathbf{X}_A) \ &= \text{SSR}(\mathbf{X}_B|\mathbf{X}_A) + \text{SSE}(\mathbf{X}_A, \mathbf{X}_B) \end{aligned}$$

$\text{SSR}(\mathbf{X}_B|\mathbf{X}_A)$ is the drop in SSE when $\mathbf{X}_B = [X_{k_1+1}, \ldots, X_{k_1+k_2}]$ are added to the model with $\mathbf{X}_A = [X_1, \ldots, X_{k_1}]$ already in.

ANOVA decomposition ($\mathbf{X}_A$ first)

Source	SS	df	F
A1. $\mathbf{X}_A$	$\text{SSR}(\mathbf{X}_A)$	k_1	
A2. $\mathbf{X}_B|\mathbf{X}_A$	$\text{SSR}(\mathbf{X}_B|\mathbf{X}_A)$	k_2	$F = \frac{\text{SSR}(\mathbf{X}_B|\mathbf{X}_A)/k_2}{\text{SSE}(\mathbf{X}_A,\mathbf{X}_B)/(n-p)}$
A3. Error	$\text{SSE}(\mathbf{X}_A, \mathbf{X}_B)$	$n-p$	
Total	SSTo	$n-1$	

Decision rule: $F \sim F(k_2, n-p)$ under H_0.

If $F \leq F(1-\alpha; k_2, n-p)$, *conclude* H_0.

If $F > F(1-\alpha; k_2, n-p)$, *conclude* H_1.

Remark 7.2.

1. $\mathrm{SSR}(\mathbf{X}_A)$ and $\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A)$ add up to $\mathrm{SSR}(\mathbf{X}_A, \mathbf{X}_B)$, the SSR that appears in the overall F test. Thus, we have decomposed SSR for the full model into two pieces — a piece due to the effect of adding $\mathbf{X}_B$ after $\mathbf{X}_A$ are already in.

2. $\mathrm{SSR}(\mathbf{X}_B|\mathbf{X}_A)$ and $\mathrm{SSE}(\mathbf{X}_A, \mathbf{X}_B)$ add up to $\mathrm{SSE}(\mathbf{X}_A)$, the SSE from the reduced model that contains only $\mathbf{X}_A$.

3. In the special case that $\mathbf{X}_A = [X_1, \ldots, X_{k-1}]$ and $\mathbf{X}_B = [X_k]$ where $k = p - 1$, we are testing the effect of the last variable X_k

$$\begin{aligned} H_0 &: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \epsilon \\ H_1 &: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \underline{\beta_k X_k} + \epsilon. \end{aligned}$$

   That is, we are testing

$$H_0 : \beta_k = 0 \quad \text{versus} \quad H_1 : \beta_k \neq 0.$$

   In this special case,

$$F = \frac{\mathrm{SSR}(X_k|X_1, \ldots, X_{k-1})/1}{\mathrm{SSE}(X_1, \ldots, X_k)/(n-p)} = \frac{\mathrm{SSR}(X_k|X_1, \ldots, X_{k-1})}{\mathrm{MSE}} = T^2 = \left[\frac{\hat\beta_k}{\mathrm{SE}(\hat\beta_k)}\right]^2,$$

   where $\mathrm{SE}(\hat\beta_k) = \sqrt{\mathrm{MSE}\,[(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}$ and $\mathrm{MSE} = \mathrm{SSE}(X_1, \ldots, X_k)/(n-p)$. Notice that if $T \sim t(df)$, then $T^2 \sim F(1, df)$. From this, we have the following result:

$$\mathrm{SSR}(X_k|X_1, \ldots, X_{k-1}) = \mathrm{MSE}\left[\frac{\hat\beta_k}{\mathrm{SE}(\hat\beta_k)}\right]^2.$$

   Thus, the T-statistics from the table of coefficients give the significance of each variable given that all other variables are already in the model.

△

3.3 Sequential F tests

Sequential F tests pertain to the effects of adding each variable in sequence. Let us consider the following models:

$$
\begin{aligned}
Y &= \beta_0 + \epsilon \\
Y &= \beta_0 + \beta_1 X_1 + \epsilon \\
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\
&\vdots \\
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon
\end{aligned}
$$

1. Effect of X_1 given that no variables are in the model
2. Effect of X_2 given that X_1 is in the model
3. Effect of X_3 given that X_1 and X_2 are in the model
4. Effect of X_4 given that $X_1, \ldots, X_3$ are in the model
5. and so on $\cdots$.

The sequence of sequential F tests depends on the order in which the variables are added. We can decompose the ANOVA table as follows:

ANOVA decomposition of $\mathrm{SSR}(X_1, \ldots, X_k)$

Source		SS	df
Regression		$\mathrm{SSR}(X_1, \ldots, X_k)$	k
L_1	X_1	$\mathrm{SSR}(X_1)$	1
L_2	$X_2 \mid X_1$	$\mathrm{SSR}(X_2 \mid X_1)$	1
L_3	$X_3 \mid X_1, X_2$	$\mathrm{SSR}(X_3 \mid X_1, X_2)$	1
	$\vdots$	$\vdots$	$\vdots$
L_k	$X_k \mid X_1, \ldots, X_{k-1}$	$\mathrm{SSR}(X_k \mid X_1, \ldots, X_{k-1})$	1
Error		$\mathrm{SSE}(X_1, \ldots, X_k)$	$n-(k+1)$
Total		SSTo	$n-1$

Note that

$$\text{SSR}(X_1, \ldots, X_j) = L_1 + \cdots + L_j$$

$$\text{SSE}(X_1, \ldots, X_j) = L_{j+1} + \cdots + L_k + \text{SSE}(X_1, \ldots, X_k).$$

A sequential F test is like a partial F test. For example, the sequential F test for X_3 is testing

$$H_0 : \beta_3 = 0 \quad \text{versus} \quad H_1 : \beta_3 \neq 0,$$

that is

$$\boxed{\begin{aligned} &H_0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ &H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \end{aligned}}$$

where both H_0 and H_1 assume $\beta_4 = \beta_5 = \cdots = \beta_k = 0$, *i.e.*, H_1 is thought of as the full model. If $\beta_4 = \beta_5 = \cdots = \beta_k = 0$ is true, then we can collapse $L_4, L_5, \ldots, L_k$ of the ANOVA table into error to get $\text{SSE}(X_1, X_2, X_3)$. Then we have the following F test statistic

$$F = \frac{\text{SSR}(X_3|X_1, X_2)/1}{\text{SSE}(X_1, X_2, X_3)/(n-4)} \sim F(1, n-4) \quad \text{under } H_0.$$

The above F test is essentially the same as the t-test with

$$T = \frac{\hat{\beta}_3}{\text{SE}(\hat{\beta}_3)}.$$

Note that $T^2 \sim F(1, n-4)$ under H_0.

We can also do sequential F tests for groups of variables. For example,

$$\boxed{\begin{aligned} &H_0 : Y = \beta_0 + \beta_1 X_1 + \epsilon \\ &H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \end{aligned}}$$

$$F = \frac{\text{SSR}(X_2, X_3|X_1)/2}{\text{SSE}(X_1, X_2, X_3)/(n-4)} \sim F(2, n-4) \quad \text{under } H_0.$$

Note that $\text{SSR}(X_2, X_3|X_1) = \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1, X_2)$.

4 Coefficients of partial determination

The *coefficient of partial determination* is defined as the relative marginal reduction in the variation in Y associated with $\mathbf{X}_B$ when $\mathbf{X}_A$ is already in the model. Let us consider the following model:

$$H_0 : \mathbf{Y} = \beta_0\mathbf{1} + \mathbf{X}_A\boldsymbol{\beta}_A + \boldsymbol{\epsilon},$$

$$H_1 : \mathbf{Y} = \beta_0\mathbf{1} + \mathbf{X}_A\boldsymbol{\beta}_A + \mathbf{X}_B\boldsymbol{\beta}_B + \boldsymbol{\epsilon}.$$

The $\text{SSE}(\mathbf{X}_A)$ measures the variation in Y when the reduced model ($\mathbf{Y} = \beta_0\mathbf{1} + \mathbf{X}_A\boldsymbol{\beta}_A + \boldsymbol{\epsilon}$) is used and the $\text{SSE}(\mathbf{X}_A, \mathbf{X}_B)$ measures the variation in Y when the full model ($\mathbf{Y} = \beta_0\mathbf{1} + \mathbf{X}_A\boldsymbol{\beta}_A + \mathbf{X}_B + \boldsymbol{\epsilon}$) is used. The coefficient of partial determination between Y and $\mathbf{X}_B$ with $\mathbf{X}_A$ already in the model is defined as follows:

$$R^2(Y, \mathbf{X}_B|\mathbf{X}_A) = \frac{\text{SSE}(\mathbf{X}_A) - \text{SSE}(\mathbf{X}_A, \mathbf{X}_B)}{\text{SSE}(\mathbf{X}_A)} = \frac{\text{SSR}(\mathbf{X}_B|\mathbf{X}_A)}{\text{SSE}(\mathbf{X}_A)}.$$

$$\boxed{R^2(Y, \mathbf{X}_B|\mathbf{X}_A) = \frac{\Delta\text{SSE}}{\text{SSE(reduced)}}.}$$

This $R^2(Y, \mathbf{X}_B|\mathbf{X}_A)$ is interpreted as the proportion of the variation in Y explained by $\mathbf{X}_B$ after $\mathbf{X}_A$ is included in the model.

When *only one more* predictor X_k is added to the full model (*i.e.*, $\mathbf{X}_B = [X_k]$), the square root of a coefficient of partial determination is called a *coefficient of partial correlation*. The sign is the same as the regression coefficient in the regression.

$$r(Y, X_k|\mathbf{X}_A) = \text{sign}(\hat{\beta}_k)\sqrt{R^2(Y, X_k|\mathbf{X}_A)} = \text{sign}(\hat{\beta}_k)\sqrt{\frac{\text{SSR}(X_k|\mathbf{X}_A)}{\text{SSE}(\mathbf{X}_A)}}.$$

Example 7.3. Body Fat Example in Table 7.1 on Page 257 of Kutner et al. (2005).

`Minitab`

Read Data

```
MTB > read c1 c2 c3 c11 ;
SUBC>     file "S:\LM\CH07TA01.txt" .

MTB > name c1  'X1'
MTB > name c2  'X2'
MTB > name c3  'X3'
MTB > name c11 'Y'
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (X_1 first)

```
MTB > regr c11 2 c1 c2

Regression Analysis: Y versus X1, X2

```

```
The regression equation is
Y = - 19.2 + 0.222 X1 + 0.659 X2

Predictor      Coef   SE Coef      T      P
Constant    -19.174     8.361  -2.29  0.035
X1           0.2224    0.3034   0.73  0.474
X2           0.6594    0.2912   2.26  0.037

S = 2.54317   R-Sq = 77.8%   R-Sq(adj) = 75.2%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  385.44  192.72  29.80  0.000
Residual Error  17  109.95    6.47
Total           19  495.39

Source  DF  Seq SS
X1       1  352.27
X2       1   33.17
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

```
MTB > regr c11 3 c1 c2 c3

Regression Analysis: Y versus X1, X2, X3

The regression equation is
Y = 117 + 4.33 X1 - 2.86 X2 - 2.19 X3

Predictor     Coef  SE Coef      T      P
Constant    117.08    99.78   1.17  0.258
X1           4.334    3.016   1.44  0.170
X2          -2.857    2.582  -1.11  0.285
X3          -2.186    1.595  -1.37  0.190

S = 2.47998   R-Sq = 80.1%   R-Sq(adj) = 76.4%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       3  396.98  132.33  21.52  0.000
Residual Error  16   98.40    6.15
Total           19  495.39

Source  DF  Seq SS
X1       1  352.27
X2       1   33.17
X3       1   11.55
```

R

Ⓡ Read Data

```
> ## TABLE 7.1: Body Fat Example (pg. 257)
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH07TA01.txt")
> x1 = mydata[,1]
> x2 = mydata[,2]
> x3 = mydata[,3]
> y  = mydata[,4]
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (X_1 first)

```
> LM3 = lm ( y ~ x1 + x2 )
> anova(LM3)
Analysis of Variance Table

Response: y
```

```
              Df  Sum Sq  Mean Sq  F value     Pr(>F)
x1             1  352.27   352.27  54.4661  1.075e-06 ***
x2             1   33.17    33.17   5.1284     0.0369 *
Residuals     17  109.95     6.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

```
> LM5 = lm ( y ~ x1 + x2 + x3 )
> anova(LM5)
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
```

From the Minitab or R results, we can get $\text{SSR}(X_2|X_1) = 33.17$ and $\text{SSE}(X_1) = 33.17 + 109.95$. It follows that

$$
\begin{aligned}
R^2(Y, X_2|X_1) &= \frac{\text{SSE}(X_1) - \text{SSE}(X_1, X_2)}{\text{SSE}(X_1)} \\
&= \frac{\text{SSR}(X_2|X_1)}{\text{SSE}(X_1)} = \frac{33.17}{33.17 + 109.95} = 0.2318,
\end{aligned}
$$

and $r(Y, X_2|X_1) = \text{sign}(\hat{\beta}_2)\sqrt{R^2(Y, X_2|X_1)} = 0.481$.

Similarly, we have

$$
\begin{aligned}
R^2(Y, X_3|X_1, X_2) &= \frac{\text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3)}{\text{SSE}(X_1, X_2)} \\
&= \frac{\text{SSR}(X_3|X_1, X_2)}{\text{SSE}(X_1, X_2)} = \frac{11.55}{11.55 + 98.40} = 0.1050477,
\end{aligned}
$$

and $r(Y, X_3|X_1, X_2) = \text{sign}(\hat{\beta}_3)\sqrt{R^2(Y, X_3|X_1, X_2)} = -\sqrt{0.1050477} = -0.3241$.

Remark 7.3.

1. The coefficient of partial determination between Y and X_k, $R^2(Y, X_k)$, can be thought of as the coefficient of partial determination between Y and X_k with the null model in the reduced model. Thus, we have

$$
R^2(Y, X_k|\text{Null}) = \frac{\text{SSE(Null)} - \text{SSE}(X_k)}{\text{SSE(Null)}} = \frac{\text{SSR}(X_k)}{\text{SSTo}}
$$

since SSE(Null) = SSTo. The coefficient of partial correlation between Y and X_k is given as follows

$$r(Y, X_k|\text{Null}) = \text{sign}(\hat{\beta}_k)\sqrt{\frac{\text{SSR}(X_k)}{\text{SSTo}}}.$$

2. The coefficient of multiple determination R^2 can be considered as the coefficient of partial determination between Y and $\mathbf{X}$ (all the predictors) with the null model in the reduced model. That is,

$$R^2 = R^2(Y, \mathbf{X}|\text{Null}) = \frac{\text{SSE(Null)} - \text{SSE}(\mathbf{X})}{\text{SSE(Null)}} = \frac{\text{SSTo} - \text{SSE}(\mathbf{X})}{\text{SSTo}} = \frac{\text{SSR}(\mathbf{X})}{\text{SSTo}}.$$

△

Another way to compute a partial correlation coefficient is using the residuals. We can find $r(Y, X_k|\mathbf{X}_A)$ in the following manner:

(i) Regress Y on $\mathbf{X}_A$ and save residuals

(ii) Regress X_k on $\mathbf{X}_A$ and save residuals

(iii) Calculate the ordinary correlation between the residuals from (i) and the residuals from (ii).

`Minitab`

Partial correlation between Y and X_2 with X_1 given

```
regr c11 1 c1;
resid c21.

regr c2 1 c1;
resid c22.

 MTB > corr c21 c22.

 Correlation of C21 and C22 = 0.481
```

Partial correlation between Y and X_3 with X_1 and X_2 given

```
regr c11 2 c1 c2;
resid c21.

regr c3  2 c1 c2;
resid c22.

 MTB > corr c21 c22.

 Correlation of C21 and C22 = -0.324
```

R

Ⓡ Partial correlation between Y and X_2 with X_1 given

```
> LM1 = lm ( y ~ x1 )
> c21 = resid(LM1)
> tmp = lm ( x2 ~ x1 )
> c22 = resid(tmp)
> cor(c21, c22)
[1] 0.4814109
```

Ⓡ Partial correlation between Y and X_3 with X_1 and X_2 given

```
> LM3 = lm ( y ~ x1 + x2 )
> c21 = resid(LM3)
> tmp = lm ( x3 ~ x1 + x2 )
> c22 = resid(tmp)
> cor(c21, c22)
[1] -0.3240520
```

$\|$

5 Standardized multiple regression model

As will be stated in the following section, strong multicollinearity can result in round-off errors in calculating $(\mathbf{X}'\mathbf{X})^{-1}$. Such errors can also occur when the predictors have substantially different magnitudes because they cause the entries in $\mathbf{X}'\mathbf{X}$ to cover a wide range of values.

To control round-off errors, we can transform the variables in the multiple linear regression model by standardizing the variables. Consider a normal random variable X with μ and σ. Then the standardized normal random variable

$$Z = \frac{Y - \mu}{\sigma}$$

is normally distributed with $N(0,1)$. By analogy with this, we can transform the variables

$$Y_i^* = \frac{1}{\sqrt{n-1}} \Big(\frac{Y_i - \bar{y}}{s_y} \Big) \tag{7.1}$$

and

$$X_{ik}^* = \frac{1}{\sqrt{n-1}} \Big(\frac{X_{ik} - \bar{X}_k}{s_k} \Big), \qquad (k = 1, \ldots, p-1)$$

where

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{y})^2 \quad \text{with} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

and

$$s_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ik}-\bar{X}_k)^2 \quad \text{with} \quad \bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ik}.$$

The above transform is called the *correlation transformation*, which makes all the entries in the $\mathbf{X}'\mathbf{X}$ matrix after this transformation fall on $[-1, 1]$.

The *standardized regression model* is defined by the correlation transformation and is as follows:

$$Y_i^* = \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \epsilon_i^*. \tag{7.2}$$

Notice that this standardized regression model does not include an intercept parameter β_0^*.

It is easily seen that the original parameters are related to

$$\beta_k = \left(\frac{s_y}{s_k}\right)\beta_k^* \qquad (k = 1, 2, \ldots, p-1) \tag{7.3}$$

$$\beta_0 = \bar{y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2 - \cdots - \beta_{p-1}\bar{X}_{p-1}.$$

Using the standardized variables, we form the matrices

$$\underset{n\times 1}{\overset{*}{\mathbf{Y}}} = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{bmatrix} \quad \text{and} \quad \underset{n\times(p-1)}{\overset{*}{\mathbf{X}}} = \begin{bmatrix} X_{11}^* & X_{12}^* & \cdots & X_{1,p-1}^* \\ X_{21}^* & X_{22}^* & \cdots & X_{2,p-1}^* \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}^* & X_{n2}^* & \cdots & X_{n,p-1}^* \end{bmatrix}. \tag{7.4}$$

Note that $\mathbf{r}_{XX} = \overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{X}}$ is the correlation matrix of the predictors and $\mathbf{r}_{YX} = \overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{Y}}$ is the vector of the coefficients of simple correlation between Y and each of the predictors X_k. Especially $\mathbf{r}_{XX}$ is given by

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

where

$$r_{jk} = \sum_{i=1}^{n} X_{ij}^{*} X_{ik}^{*} = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\sqrt{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}\sqrt{\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)^2}}.$$

Example 7.4. Dwaine Studios Example on Page 276. (Original data are in Figure 6.5b on Page 237).

Minitab

Read Data

```
MTB >READ c1 c2 c3;
SUBC>   file "S:\LM\CH06FI05.txt" .
Entering data from file: S:\LM\CH06FI05.TXT
21 rows read.

MTB > name c11 'Y'
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
MTB > regr c3 2 c1 c2 .

Regression Analysis: C3 versus C1, C2

The regression equation is
C3 = - 68.9 + 1.45 C1 + 9.37 C2

Predictor     Coef  SE Coef      T      P
Constant    -68.86    60.02  -1.15  0.266
C1          1.4546   0.2118   6.87  0.000
C2           9.366    4.064   2.30  0.033
.......................
```

Using STCOEF.MAC

```
MTB > S:\LM\STCOEF c3 c1 c2 .
Executing from file: S:\LM\STCOEF.MAC

Standardized Regression Coefficients for C3

Row  Predictors    StdCoef
  1  C1           0.748367
  2  C2           0.251104
```

R

Ⓡ Read Data

```
> # Original data from Figure 6.5b on Page 237
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH06FI05.txt")
>
> x1 = mydata[,1]
> x2 = mydata[,2]
> y  = mydata[,3]
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
> ## Ordinary regression
> LM = lm ( y ~ x1 + x2 )
> coef(LM)
(Intercept)          x1          x2
 -68.857073    1.454560    9.365500
>
> ## Standardized regression
> n = length(y)
> x1star = (x1-mean(x1)) / sd(x1)
> x2star = (x2-mean(x2)) / sd(x2)
>  ystar = (y-mean(y)) / sd(y)
>
> LMstar = lm ( ystar ~ 0 + x1star + x2star )
> LMstar

Call:
lm(formula = ystar ~ 0 + x1star + x2star)

Coefficients:
x1star  x2star
0.7484  0.2511

> LMstar2 = lm ( ystar ~ x1star + x2star )
> LMstar2

Call:
lm(formula = ystar ~ x1star + x2star)

Coefficients:
(Intercept)       x1star       x2star
 -4.645e-17    7.484e-01    2.511e-01
```

Ⓡ Using coef.sd() R function

```
> ## Using coef.sd.R() function
> source("https://raw.githubusercontent.com/AppliedStat/LM/master/coef-sd.R")
> LM = lm ( y ~ x1 + x2 )
> coef.sd(LM)
       x1        x2
0.7483670 0.2511039
```

‖

6 Multicollinearity

Collinearity means that two predictors X_1 and X_2 are highly correlated. When there are more than two correlated predictors (say, $X_1, \ldots, X_k$), this condition is called multicollinearity and means that at least one X_j can be predicted with substantial accuracy from the others. That is, the regression of X_j on all the other predictors will have a high R^2 (coefficient of multiple determination).

In the most extreme form of collinearity or multicollinearity, one of the columns of $\mathbf{X}$ is a perfect linear combination of the others. Then $\mathbf{X}'\mathbf{X}$ has deficient rank (*i.e.*, is singular) and $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist, so $\hat{\boldsymbol{\beta}}$ can't be computed in the usual way.

Heuristically, suppose that we have two predictors X_1 and X_2 and they are highly collinear. Then X_2 is approximately linear in X_1, and vice versa, that is,

$$X_2 \approx a + bX_1, \qquad a, b \text{ constants.}$$

Then $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ becomes

$$\begin{aligned} Y &\approx \beta_0 + \beta_1 X_1 + \beta_2(a + bX_1) + \epsilon \\ &= (\beta_0 + a\beta_2) + (\beta_1 + b\beta_2)X_1 + \epsilon \end{aligned}$$

Two parameters,

$$\beta_0^* = \beta_0 + a\beta_2 \ \text{ and } \ \beta_1^* = \beta_1 + b\beta_2$$

are well estimated, but the original parameters β_0, β_1 and β_2 are not. For any $\hat{\beta}_0^* = \beta_0 + a\beta_2$ and $\hat{\beta}_1^* = \beta_1 + b\beta_2$, there are an infinite number of sets of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ that will satisfy.

The problem does not lie in that $\hat{\boldsymbol{\beta}}$ does not exist, but that an infinite number of $\hat{\boldsymbol{\beta}}$'s exist, all of which lead to the same (or nearly the same) fitted values $\hat{Y}$.

6.1 Effects of multicollinearity

Multicollinearity can make a variety of effects on the multiple linear regression such as (i) SSR and SSE, (ii) regression parameters, (iii) t-test statistic, and (iv) inflation of variance.

Effects on SSR and SSE

When the two predictors, say, X_1 and X_2 are highly correlated, then the SSR$(X_1|X_2)$ is very small compared to SSR(X_1). This is because X_2 contains much of the same information as X_1. So the the marginal contribution of X_1 in reducing the error sum of squares is comparatively small when X_2 is already in the regression model. Similarly SSR$(X_2|X_1)$ is very small compared to SSR(X_2). Note that if X_1 and X_2 are uncorrelated, then SSR$(X_2|X_1) =$ SSR(X_2) and SSR$(X_1|X_2) =$ SSR(X_1).

Multicollinearity also affects the coefficients of partial determination through its effects on SSE. When Y and X_1 are highly correlated, $R^2(Y, X_1|X_2) = \text{SSR}(X_1|X_2)/\text{SSE}(X_2)$ is very small compared to $R^2(Y, X_1) = \text{SSR}(X_1)/\text{SSTo}$.

Effects on regression parameters

If multicollinearity exists, the estimates of the regression parameters depend on the particular predictors. If a predictor is added to a regression model and this added predictor in the model is highly correlated to the other predictor(s) already included in the model, the estimates of the regression parameters can change dramatically.

Effects on t-test statistics and associated p-values

When multicollinearity exists, two or more correlated predictors contribute redundant information. This often causes the t-test statistics by relating a response variable to correlated predictor(s) to be smaller than the t-test statistics that would be obtained with correlated predictor(s) if separate regression analyses were run. That is, multicollinearity can cause some of the correlated predictors to appear to be less significant than they actually are.

Example 7.5. Example and Table 7.6 on Page 279 and Table 7.7 on Page 280 of Kutner et al. (2005).

Ⓡ R functions to construct Table 7.7.

```
> x1 = c(4,4,4,4, 6,6,6,6)
> x2 = c(2,2,3,3, 2,2,3,3)
> x3 = c(6,6,7,7, 8,9,9,9)
> y  = c(42,39,48,51, 49,53,61,60)
>
> #
> cor( cbind(x1,x2,x3) )
          x1        x2        x3
x1 1.0000000 0.0000000 0.9233805
x2 0.0000000 1.0000000 0.3077935
x3 0.9233805 0.3077935 1.0000000
>
> LMa = lm( y~x1 )
> LMb = lm( y~x2 )
> LMc = lm( y~x1 + x2 )
> LMd = lm( y~x2 + x1 )
> LMe = lm( y~x1 + x3 )
> LMf = lm( y~x3 + x1 )
```

‖

Predictors		X_1	X_2	X_1, X_2	X_1, X_3	X_1, X_2, X_3
	$\hat{\beta}_1$	5.375		5.375	-4.750	2.000
β_1	SE	1.983		0.6638	2.824	2.540
	t stat.	2.711		8.097	-1.682	0.787
	p-value	0.0351		0.00047	0.1534	0.4750
	$\hat{\beta}_2$		9.250	9.250		7.000
β_2	SE		4.553	1.3276		2.049
	t stat.		2.032	6.968		3.416
	p-value		0.0885	0.00094		0.0269
	$\hat{\beta}_3$				9.000	3.000
β_3	SE				2.318	2.191
	t stat.				3.883	1.369
	p-value				0.0116	0.2427

Source	SS	Source	SS	Source	SS	Source	SS
X_1	231.1	X_2	171.1	X_1	231.1	X_3	346.3
$X_2 \vert X_1$	171.1	$X_1 \vert X_2$	231.1	$X_3 \vert X_1$	141.8	$X_1 \vert X_3$	26.6
$X_3 \vert X_1, X_2$	5.6	$X_3 \vert X_1, X_2$	5.6	$X_2 \vert X_1, X_3$	35.0	$X_2 \vert X_1, X_3$	35.0

6.2 Variance inflation factor

Multicollinearity can cause the t-test statistic

$$T = \frac{\hat{\beta}_j}{\sqrt{\text{MSE} \cdot d_{jj}}},$$

where d_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to the parameter β_j. Especially when nearly-perfect dependence exists, the variances of the elements of $\hat{\boldsymbol{\beta}}$ are large. This can hinder our ability to assess and test the regression parameters.

Theorem 7.2 (VIF). *It can be shown that*

$$d_{jj} = \frac{1}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 \cdot (1 - R_j^2)}, \tag{7.5}$$

where R_j^2 is the multiple coefficient of determination calculated by using the regression model

$$X_{ij} = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_{j-1} X_{i,j-1} + \gamma_{j+1} X_{i,j+1} + \cdots + \gamma_{p-1} X_{i,p-1} + \eta_i. \tag{7.6}$$

This model expresses the predictor X_j as a function of the remaining predictors, $X_{i1}, \ldots, X_{i,j-1}, X_{i,j+1}, \ldots, X_{i,p-1}$. Because R_j^2 is the proportion of the total variation in the variable X_j explained by the regression model with the remaining predictors, it follows that R_j^2 is a measure of the multicollinearity between X_j and the remaining predictors, $X_{i1}, \ldots, X_{i,j-1}, X_{i,j+1}, \ldots, X_{i,p-1}$. The greater the multicollinearity is, the closer to one is R_j^2.

Lemma 7.3 (The inverse of the partitioned matrix). *Let $\mathbf{B}$ be an $n \times n$ matrix partitioned by*

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix},$$

where $\mathbf{B}_{ij}$ has size $m_i \times m_j$ for $i, j = 1, 2$, and where $m_1 + m_2 = m$. Then we have

$$\mathbf{B}^{-1} = \begin{bmatrix} [\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21}]^{-1} & -\mathbf{B}_{11}^{-1}\mathbf{B}_{12}[\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}]^{-1} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21}[\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21}]^{-1} & [\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}]^{-1} \end{bmatrix}.$$

Proof. See Graybill (1976). □

Proof of Theorem 7.2. For convenience, we let

$$\mathbf{X} = [\ \mathbf{X_1} \ \ \mathbf{X_2}\ ],$$

where

$$\mathbf{X_1} = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix} \quad \text{and} \quad \mathbf{X_2} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,j-1} & X_{1,j+1} & \dots & X_{1,p-1} \\ 1 & X_{21} & \dots & X_{2,j-1} & X_{2,j+1} & \dots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{n,j-1} & X_{n,j+1} & \dots & X_{n,p-1} \end{bmatrix}.$$

Then we have

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X_1'X_1} & \mathbf{X_1'X_2} \\ \mathbf{X_2'X_1} & \mathbf{X_2'X_2} \end{bmatrix}.$$

Applying Lemma 7.3 to the above matrix, we have

$$\begin{aligned} d_{jj} &= \Big[\mathbf{X_1'X_1} - \mathbf{X_1'X_2}(\mathbf{X_2'X_2})^{-1}\mathbf{X_2'X_1}\Big]^{-1} \\ &= \Big[\mathbf{X_1'}\Big\{\mathbf{I} - \mathbf{X_2}(\mathbf{X_2'X_2})^{-1}\mathbf{X_2}\Big\}\mathbf{X_1}\Big]^{-1} \\ &= \Big[\mathbf{X_1'}\big(\mathbf{I}-\mathbf{H}\big)\mathbf{X_1}\Big]^{-1} \\ &= \frac{1}{\mathbf{X_1'}\big(\mathbf{I}-\mathbf{H}\big)\mathbf{X_1}}, \end{aligned}$$

where $\mathbf{H} = \mathbf{X_2}(\mathbf{X_2'X_2})^{-1}\mathbf{X_2}$ is the hat (projection) matrix onto $\mathcal{R}(\mathbf{X_2})$. The term $\mathbf{X_1'}\big(\mathbf{I} - \mathbf{H}\big)\mathbf{X_1}$ is the SSE when we regress $\mathbf{X_1} = [X_{ij}]$ on $\mathbf{X_2} = [1, X_{i1}, \dots, X_{i,j-1}, X_{i,j+1}, X_{i,p-1}]$. Thus, we have

$$d_{jj} = \frac{1}{\mathrm{SSE}_j} = \frac{1}{\mathrm{SSTo}_j \cdot \dfrac{\mathrm{SSTo}_j - \mathrm{SSR}_j}{\mathrm{SSTo}_j}} = \frac{1}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2(1-R_j^2)},$$

where SSE_j, SSTo_j and SSR_j are SSE, SSTo and SSR under the regression model in (7.6), respectively. □

The VIF is defined as

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}, \qquad j = 1, 2, \dots, p-1$$

where R_j^2 is the coefficient of multiple determination when X_j is regressed on the $p-2$ other predictors in the model.

Theorem 7.4. *The* VIF_j *is the* j*-th diagonal element of the following matrix*

$$(\overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{X}})^{-1}\,\overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{X}}\,(\overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{X}})^{-1},$$

where $\overset{*}{\mathbf{X}}$ *is defined in (7.4).*

6.3 Diagnosing multicollinearity

(a) Scatter plot matrix of the predictors

(b) Correlation matrix of the predictors

(c) VIF (Variance Inflation Factor). If $\max(\mathrm{VIF}_1, \ldots, \mathrm{VIF}_{p-1}) > 10$, it indicates that multicollinearity may be unduly influencing the least squares estimates.

Example 7.6. VIF: Body Fat Example in Table 7.1 on Page 257 and Table 10.5 on Page 409 of Kutner et al. (2005).

Minitab

Read Data

```
MTB > read c1 c2 c3 c11 ;
SUBC>     file "S:\LM\CH07TA01.txt" .
Entering data from file: S:\LM\CH07TA01.TXT
20 rows read.

MTB > name c1  'X1'
MTB > name c2  'X2'
MTB > name c3  'X3'
MTB > name c11 'Y'
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

```
MTB > regr c11 3 c1 c2 c3 ;
SUBC> vif.

Regression Analysis: Y versus X1, X2, X3

The regression equation is
Y = 117 + 4.33 X1 - 2.86 X2 - 2.19 X3

Predictor     Coef  SE Coef      T      P      VIF
Constant    117.08    99.78   1.17  0.258
X1           4.334    3.016   1.44  0.170  708.843
X2          -2.857    2.582  -1.11  0.285  564.343
X3          -2.186    1.595  -1.37  0.190  104.606
```

R

Ⓡ Read Data

```
> mydata =
   read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH07TA01.txt")
>
> x1 = mydata[,1]
> x2 = mydata[,2]
> x3 = mydata[,3]
> y  = mydata[,4]
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (X_1 first)

```
> source("https://raw.githubusercontent.com/AppliedStat/LM/master/VIF.R")
>
> LM = lm (y ~ x1 + x2 + x3)
>
> vif(LM)
      x1       x2       x3
708.8429 564.3434 104.6060
```

‖

6.4 What to do?

(a) If the goal is to estimate β_0, β_1 and β_2, we can not do much. When we remove a multicollinear predictor (say, X_2) from the model, this changes the definition of β_1.

(b) If the goal is to predict Y, then we are OK, provided that we are predicting only over the region of X-space that contains the observed data (interpolation). Within the observed range, the full model and the two reduced models will all give similar fits. But, extrapolation is very unstable.

(c) Ridge regression.

In Chapter 6, we have studied how to estimate the regression parameters which are obtained by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}).$$

When strong multicollinearity exists, the calculation of $(\mathbf{X}'\mathbf{X})^{-1}$ is quite unstable or infeasible. The idea is to stabilize this calculation by adding positive values on the diagonal components of $\mathbf{X}'\mathbf{X}$. The basic idea is to use $(\mathbf{X}'\mathbf{X}+\mathbf{D})^{-1}$ instead of $(\mathbf{X}'\mathbf{X})^{-1}$ where $\mathbf{D}$ is a diagonal matrix. A better idea for simplification is to use a standardized regression model.

The ridge estimators of the parameters β_1^*, β_{p-1}^* under the standardized regression model are given by

$$\hat{\boldsymbol{\beta}}^{*R} = (\overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{X}} + c\mathbf{I})^{-1}(\overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{Y}}),$$

where $c \geq 0$ is a biasing constant. Again, note that $\mathbf{r}_{XX} = \overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{X}}$ is the correlation matrix of the predictors and $\mathbf{r}_{YX} = \overset{*}{\mathbf{X}}{}'\overset{*}{\mathbf{Y}}$ is the vector of the coefficients of simple correlation between Y and each of the predictors X_k.

The ridge estimators of parameters $\beta_0, \beta_1, \beta_{p-1}$ under the original regression model are

$$\beta_k^R = \frac{s_y}{s_k}\beta_k^{*R} \quad \text{and} \quad \beta_0 = \bar{Y} - \beta_1^R\bar{X}_1 - \cdots - \beta_{p-1}^R\bar{X}_{p-1}.$$

One disadvantage of this ridge regression is that the choice of c is somewhat subjective. Note that R has a `lm.ridge( )` function in the `MASS` library.

References

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system with applications to the analysis of variance. *Proceedings of the Cambridge Philosophical Society*, 30:178–191.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill, New York, 5th edition.

Scheffé, H. (1959). *The Analysis of Variance*. John Wiley & Sons, New York.