Chapter 3

Multiple Linear Regression

Chapter 3d

$SS_R(\boldsymbol{\beta})$ and $SS_R(\boldsymbol{\beta_2}|\boldsymbol{\beta_1})$ notations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon, \quad p = k + 1$$

$$y = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-r-1} x_{p-r-1})$$ p-r regression coefficients

$$+ (\beta_{p-r} x_{p-r} + ... + \beta_k x_k) + \epsilon$$

r regression coefficients

$$\underset{n\times 1}{\boldsymbol{y}} = \overset{n\times p}{\boldsymbol{X}}\overset{p\times 1 \text{ (where } p=k+1)}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}} = (\underset{n\times(p-r)}{\boldsymbol{X_1}}, \overset{n\times r}{\boldsymbol{X_2}}) \begin{bmatrix} \overset{(p-r)\times 1}{\boldsymbol{\beta_1}} \\ \underset{r\times 1}{\boldsymbol{\beta_2}} \end{bmatrix} + \boldsymbol{\epsilon} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{\epsilon}$$

$\boldsymbol{\beta_1}$ contains the first $p - r$ regression coefficients.

$\boldsymbol{\beta_2}$ contains the last r regression coefficients.

$$\boldsymbol{X_1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-r-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-r-1} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-r-1} \end{bmatrix}_{n\times(p-r)} \qquad \boldsymbol{X_2} = \begin{bmatrix} x_{1,p-r} & x_{1,p-r+1} & \dots & x_{1,k} \\ x_{2,p-r} & x_{2,p-r+1} & \dots & x_{2,k} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{n,p-r} & x_{n,p-r+1} & \dots & x_{n,k} \end{bmatrix}_{n\times r}$$

$$\boldsymbol{\beta_1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_{p-r-1} \end{bmatrix}_{(p-r)\times 1} \qquad \boldsymbol{\beta_2} = \begin{bmatrix} \beta_{p-r} \\ \beta_{p-r+1} \\ . \\ . \\ . \\ \beta_k \end{bmatrix}_{r\times 1}$$

multiple linear regression
model with K parameters

1. $$SS_R(\boldsymbol{\beta}) = SS_R(\beta_0, \beta_1, ..., \beta_k) \equiv \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

$$SS_R(\boldsymbol{\beta_1}) = SS_R(\beta_0, \beta_1, ..., \beta_{p-r-1}) \equiv \hat{\boldsymbol{\beta}}_{\boldsymbol{1}}'\boldsymbol{X}_{\boldsymbol{1}}'\boldsymbol{y}$$

$$SS_R(\boldsymbol{\beta_2}) = SS_R(\beta_{p-r}, \beta_{p-r+1}, ..., \beta_k) \equiv \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}'\boldsymbol{X}_{\boldsymbol{2}}'\boldsymbol{y}$$

vectors/matrices

2. $$SS_R(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}) \equiv SS_R(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) - SS_R(\boldsymbol{\beta_1})$$

r p-r

$$SS_R(\beta_{p-r}, ..., \beta_k|\beta_0, ..., \beta_{p-r-1})$$

$$\equiv \underbrace{SS_R(\beta_0, \beta_1, ..., \beta_k)}_{\text{based on all regressor variables}} - \underbrace{SS_R(\beta_0, \beta_1, ..., \beta_{p-r-1})}_{\text{first } p-r}$$

$SS_R(\beta_1)$ β_0 $\beta_1 \ldots \beta_{p-r-1}$

$SS_R(\beta_1|\beta_2)$ β_0 $\beta_1 \ldots \beta_{p-r-1}$ $\beta_{p-r} \ldots \beta_K$

Interpretation of $SS_R(\boldsymbol{\beta})$ and $SS_R(\boldsymbol{\beta_2}|\boldsymbol{\beta_1})$

$SS_R(\boldsymbol{\beta})$ denotes the regression sum of squares due to $\boldsymbol{\beta}$.

$SS_R(\boldsymbol{\beta_2}|\boldsymbol{\beta_1})$ denotes the regression sum of squares due to $\boldsymbol{\beta_2}$ given that $\boldsymbol{\beta_1}$ is already in the model.

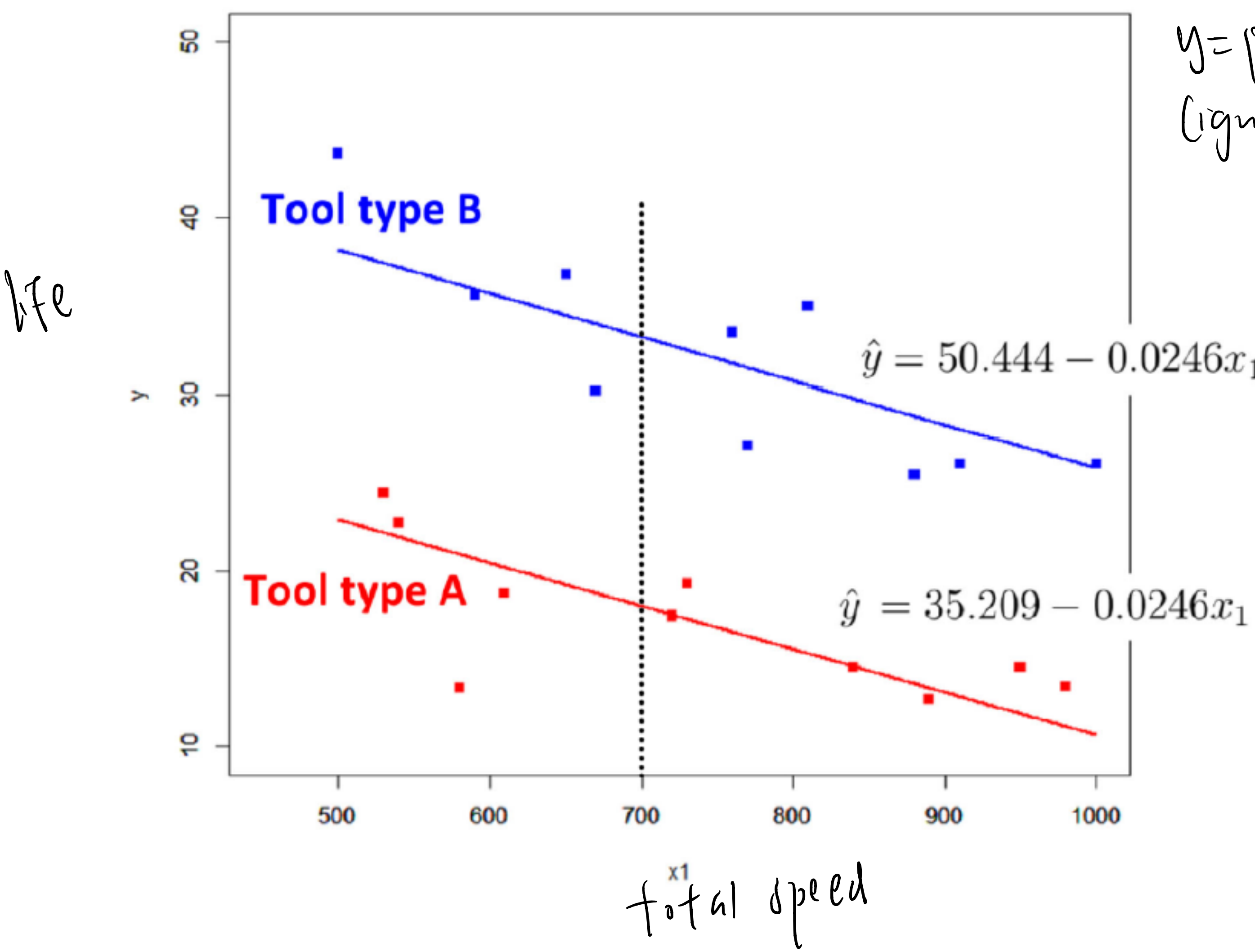
tool speed

tool type

For example, suppose we are fitting the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ where $y =$ is the life of a cutting tool, $x_1 =$ tool speed, $x_2 = 0$ if tool is of type A, $x_2 = 1$ if tool is of type B. If we fix the tool speed $x_1 = 700$ rpm and compare the two types of tools, any difference in life will be due to the tool types and not the speed. If we test $SS_R(\beta_2|\beta_0, \beta_1)$ we are comparing the life of the two types of tools after accounting for speed.

Comparing the life of two types of tools after accounting for tool speed

$y = \beta_0 + \beta_2 x_2 + \varepsilon$
(ignoring tool speed)



What is $SS_R(\beta_0)$?

$$y_i = \beta_0 + \epsilon_i, \quad i = 1, 2, ..., n$$

$$\boldsymbol{y} = \boldsymbol{X}\beta_0 + \boldsymbol{\epsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$$\hat{\beta}_0 = (\underbrace{\boldsymbol{X'X}}_{n})^{-1}\boldsymbol{X'y} = n^{-1}\sum_{i=1}^{n} y_i = \bar{y}$$

$$SS_R(\beta_0) = \hat{\beta}_0\boldsymbol{X'y} = \bar{y}\sum_{i=1}^{n} y_i = n\bar{y}^2$$

$SS_R(\beta_0)$ denotes the regression sum of squares due to β_0.

Fitting the model $y = \beta_0 + \epsilon$ and testing $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$

1. Note that $E(y) = \beta_0$, therefore testing $H_0 : \beta_0 = 0$ is the same as testing whether the random sample $y_1, y_2, ..., y_n$ taken from a normal population has mean β_0. This can be done using the t-test of a population mean assuming that the population variance is unknown (a test procedure you learned in ST1131 Introduction to Statistics):
$t = \dfrac{\bar{y} - 0}{s_y/\sqrt{n}}$ where $s_y = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Reject H_0 if $|t| > t_{\alpha/2,n-1}$.

2. According to the decomposition of variance

$$\boldsymbol{y}'\boldsymbol{y} = SS_{Res} + \hat{\beta}_0 \boldsymbol{X}'\boldsymbol{y}$$

$SS_R(\beta_0) \neq SS_R$

$$\boldsymbol{y}'\boldsymbol{y} = SS_{Res} + SS_R(\beta_0)$$

$$\boldsymbol{y}'\boldsymbol{y} = SS_{Res} + n\bar{y}^2$$

$$F = \frac{SS_R(\beta_0)/1}{SS_{Res}/(n-1)}$$

Reject H_0 if $F > F_{\alpha,1,n-1}$.

3. $$F = \frac{SS_R(\beta_0)/1}{SS_{Res}/(n-1)}$$

$$= \frac{n\bar{y}^2}{[\boldsymbol{y'y} - n\bar{y}^2]/(n-1)}$$

$$= \frac{n\bar{y}^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \left[\frac{(\bar{y} - 0)}{s_y/\sqrt{n}}\right]^2$$

$$= t^2$$

$$y'y = \sum y_i^2$$

$$\sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

The F test is equivalent to the t test for testing $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$.

Example - Cholesterol age data

```
# test beta0
summary(lm(cholesterol~1))
anova(lm(cholesterol~1))
# alternative approach
bary <- mean(cholesterol)
stdy <- sqrt(var(cholesterol))
t <- (bary - 0)/(stdy/sqrt(length(cholesterol)))
t

# plot cholesterol versus age
plot(age,cholesterol,pch=16,col="blue")

#obtain beta0
COEF <- coef(lm(cholesterol~1))
names(COEF)
names(COEF) <- NULL
beta0 <- COEF[1]
beta0
lines(c(10,70),c(beta0,beta0))
```

```
> # test beta0
> summary(lm(cholesterol~1))

Call:
lm(formula = cholesterol ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.03636 -0.83636  0.06364  0.86364  1.46364 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.1364     0.1863   16.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.07 on 32 degrees of freedom
```

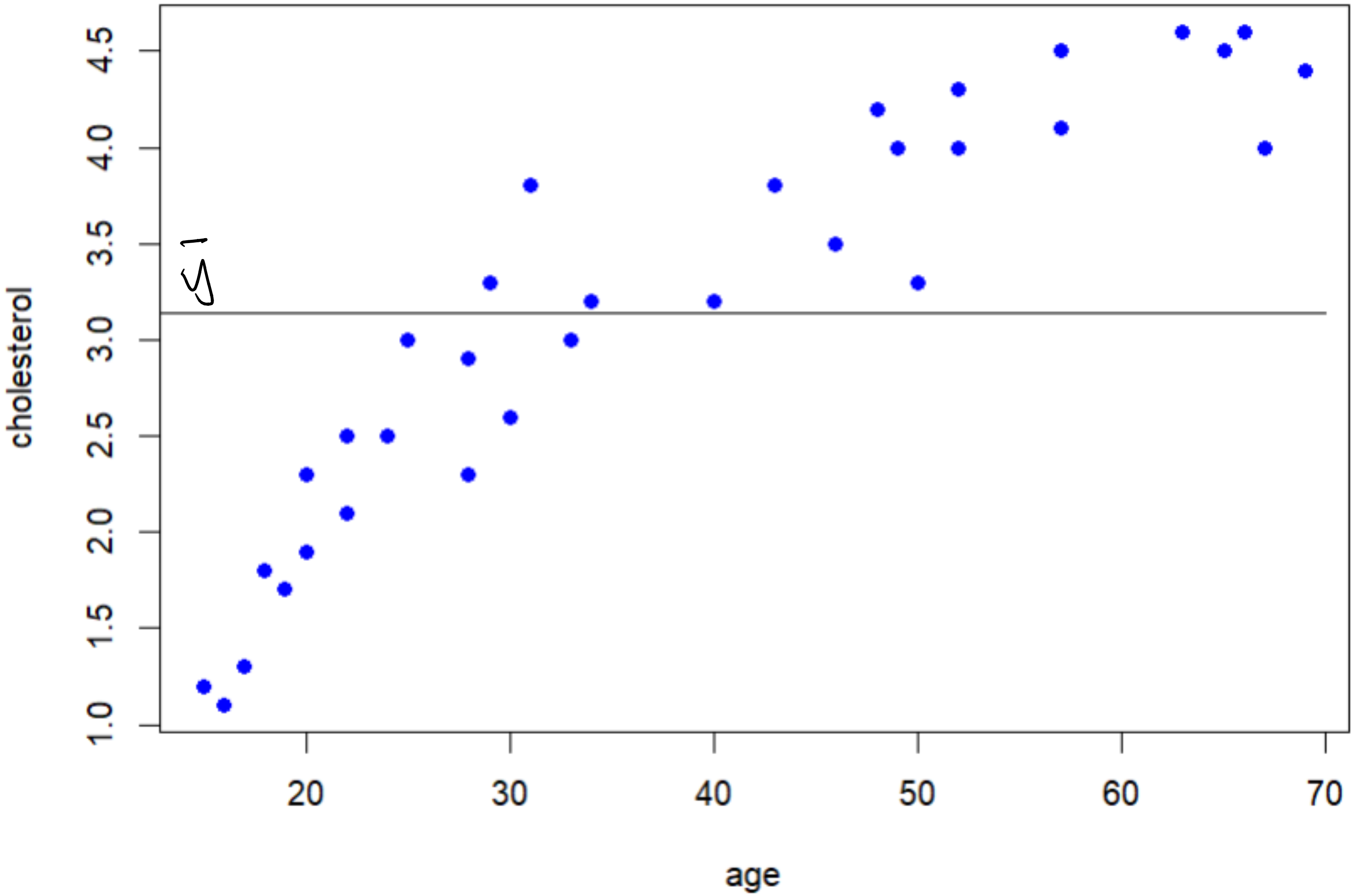
$\hat{\beta}_0$

p value for testing if $H_0: \beta_0 = 0$

```
> anova(lm(cholesterol~1))
Analysis of Variance Table

Response: cholesterol
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 32 36.636  1.1449
```

```
> # alternative approach
> bary <- mean(cholesterol)
> stdy <- sqrt(var(cholesterol))
> t <- (bary - 0)/(stdy/sqrt(length(cholesterol)))
> t
[1] 16.83845
```



cholesterol
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
20
30
40
50
60
70
age

Understanding SS_R and $SS_R(\boldsymbol{\beta})$

$$\boldsymbol{y}'\boldsymbol{y} = SS_{Res} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

$\sum (y_i - \bar{y})^2$

$$\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2 = SS_{Res} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

$$SS_T = SS_{Res} + SS_R$$

$$SS_R = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - n\bar{y}^2$$

$$= SS_R(\boldsymbol{\beta}) - SS_R(\beta_0)$$

$$= SS_R(\beta_0, \beta_1, ..., \beta_k) - SS_R(\beta_0)$$

$$= SS_R(\beta_1, ..., \beta_k | \beta_0)$$

SS_R is the regression sum of squares due to $\beta_1, ..., \beta_k$ given that β_0 is already in the model. Therefore we use SS_R to test $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.

To test $H_0 : \beta_0 = \beta_1 = \beta_2 = ... = \beta_k = 0$, we would use $SS_R(\boldsymbol{\beta})$ where $SS_R(\boldsymbol{\beta}) = SS_R(\beta_0, \beta_1, ..., \beta_k) \equiv \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$. Note that this hypothesis is usually not tested because in general β_0 is not zero.

Fitting multiple linear regression model using the R function `lm`

1. The R function `lm` can be used to fit a multiple linear regression model. For example, we can use `lm(y ~ x1+x2+x3)` to fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$.

2. The extractor `anova` can be used to obtain an ANOVA table. For example, `anova(lm(y ~ x1+x2+x3))` will produce Table 1. It can be verified easily that the regression sum of squares for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, $SS_R(\beta_1, \beta_2, \beta_3|\beta_0)$ can be obtained by adding $SS_R(\beta_1|\beta_0)$, $SS_R(\beta_2|\beta_0, \beta_1)$ and $SS_R(\beta_3|\beta_0, \beta_1, \beta_2)$.

3. The extractor `summary` can be used to obtain estimates of regression coefficients etc. For example, `summary(lm(y ~ x1+x2+x3))` will produce Table 2. The p-value for testing $H_0 : \beta_3 = 0$ versus $H_0 : \beta_3 \neq 0$ from Table 2 is exactly the same as the p-value calculated using $SS_R(\beta_3|\beta_0, \beta_1, \beta_2)$ in Table 1.

$k = 3$

$p = k + 1 = 4$

$n - p = n - 4$

$\sum (y_i - \bar{y})^2$

$SS_R(\beta_0 \beta_1) - SS_R(\beta_0)$

$\sum y_i^2 \leftarrow n$

Table 1: Analysis of variance table

Source of variation	DF	Sum of Squares	Mean Square	F
$x1$	1	$SS_R(\beta_1\|\beta_0)$	$MS_R(\beta_1\|\beta_0)$	$MS_R(\beta_1\|\beta_0)/MS_{Res}$
$x2$	1	$SS_R(\beta_2\|\beta_0, \beta_1)$	$MS_R(\beta_2\|\beta_0, \beta_1)$	$MS_R(\beta_2\|\beta_0, \beta_1)/MS_{Res}$
$x3$	1	$SS_R(\beta_3\|\beta_0, \beta_1, \beta_2)$	$MS_R(\beta_3\|\beta_0, \beta_1, \beta_2)$	$MS_R(\beta_3\|\beta_0, \beta_1, \beta_2)/MS_{Res}$
Residual	$n - 4$	SS_{Res}	MS_{Res}	
Total	$n - 1$	SS_T		

$SS_R(\beta_1 | \beta_0, \beta_2, \beta_k)$

Table 2: Estimates of regression coefficients etc

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$
$x1$			$\pm\sqrt{MS_R(\beta_1\|\beta_0,\beta_2,\beta_3)/MS_{Res}}$	
$x2$			$\pm\sqrt{MS_R(\beta_2\|\beta_0,\beta_1,\beta_3)/MS_{Res}}$	
$x3$			$\pm\sqrt{MS_R(\beta_3\|\beta_0,\beta_1,\beta_2)/MS_{Res}}$	

The pr2103 data

The data set displayed in Figure 5 was collected from a class of 83 students.

The following variables were measured from each student.

y = systolic blood pressure in mmHg

x_1 = diastolic blood pressure in mmHg

x_2 = number of heart beats per minute

x_3 = weight in kg

x_4 = height in m

x_5 = age in years

x_6 = exam score

x_7 = f for female and m for male

x_8 = religion: c for Christianity, b for Buddhism, i for Islam and o for others

x_9 = blood type: a, b, ab, o

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9
1	125	73	83	52	165	23	66	f	c	a
2	136	81	95	80	157	22	73	f	b	ab
3	132	79	83	52	160	20	61	f	c	a
4	117	69	70	48	160	21	42	f	c	a
5	128	91	69	43	158	20	66	f	o	b
6	105	62	71	53	168	21	58	f	b	o
7	127	72	87	48	166	20	45	f	b	b
8	96	64	58	60	175	23	73	m	b	ab
9	103	71	62	50	160	21	54	f	o	o
10	99	60	58	55	160	20	69	f	b	o
11	112	81	63	40	148	21	72	f	o	o
12	114	73	77	41	155	20	47	f	b	b
13	99	75	65	47	165	21	64	f	b	o
.										
.										
.										
75	117	75	86	45	161	20	75	f	b	a
76	131	96	68	55	157	20	62	f	b	b
77	117	79	74	54	150	20	56	f	o	b
78	115	78	63	53	154	20	59	f	i	a
79	111	81	84	40	154	20	95	f	b	o
80	115	76	78	46	161	22	52	f	c	o
81	103	73	70	42	142	20	64	f	c	o
82	116	81	62	60	171	22	61	m	b	b
83	100	61	100	52	163	20	94	f	i	b

```r
#ch3_pr2103.R
library(MASS)
rm(list = ls())
dat <- read.csv("D:\\nus_teaching\\st3131\\data\\pr2103.csv",
                      header = T, sep=",")
dat
names(dat)
attach(dat)

summary(lm(y~x1+x2+x3))
anova(lm(y~x1+x2+x3))
anova(lm(y~1),lm(y~x1+x2+x3))
```

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

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.1586  -7.3682  -0.5432   5.8787  29.8728 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.6946    11.7860   1.077  0.28472    
x1            0.6383     0.1193   5.352 8.30e-07 ***
x2            0.3684     0.1092   3.375  0.00115 ** 
x3            0.6186     0.1362   4.542 1.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.62 on 79 degrees of freedom
Multiple R-squared:  0.5686,	Adjusted R-squared:  0.5522 
F-statistic: 34.71 on 3 and 79 DF,  p-value: 2.056e-14

> anova(lm(y~x1+x2+x3))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x1         1 8413.4  8413.4  74.560 5.028e-13 ***
x2         1 1009.6  1009.6   8.947  0.003706 ** 
x3         1 2327.8  2327.8  20.629 1.973e-05 ***
Residuals 79 8914.3   112.8                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(lm(y~x1+x2+x3))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x1         1 8413.4  8413.4  74.560 5.028e-13 ***
x2         1 1009.6  1009.6   8.947  0.003706 ** 
x3         1 2327.8  2327.8  20.629 1.973e-05 ***
Residuals 79 8914.3   112.8                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(lm(y~1),lm(y~x1+x2+x3))
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
1     82 20665.1                                  
2     79  8914.3  3     11751 34.712 2.056e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

add

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

The End