Section 4.4 Additional Sum of Squares Principle (Gas Consumption)

```
> dat=read.table("http://www.stat.ucla.edu/~hqxu/stat100C/data/gasconsumption.txt", h=T) #Table 1.4, p.
> dim(dat)  # 38*8
[1] 38  8
> dat[1,]  # view first case to get the variable names
   MPG   GPM   WT DIS NC  HP  ACC ET
1 16.9 5.917 4.36 350  8 155 14.9  1
> g = lm(GPM ~ WT + DIS + NC + HP + ACC + ET, dat)   # response GPM=100/MPG
> summary(g)  # Table 4.2, p. 107

Call:
lm(formula = GPM ~ WT + DIS + NC + HP + ACC + ET, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4996 -0.2547  0.0402  0.1956  0.6455

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.599357   0.663403  -3.918 0.000458 ***
WT           0.787768   0.451925   1.743 0.091222 .
DIS         -0.004890   0.002696  -1.814 0.079408 .
NC           0.444157   0.122683   3.620 0.001036 **
HP           0.023599   0.006742   3.500 0.001431 **
ACC          0.068814   0.044213   1.556 0.129757
ET          -0.959634   0.266785  -3.597 0.001104 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.313 on 31 degrees of freedom
Multiple R-Squared: 0.9386,Adjusted R-squared: 0.9267
F-statistic: 78.94 on 6 and 31 DF,  p-value: < 2.2e-16

> SSE.g = sum(resid(g)^2); SSE.g  # SSE=e'e
[1] 3.037351
> sum(SSE.g)/(38-6-1)  # estimate of sigma^2
[1] 0.09797905
>
> X=as.matrix( cbind(1, dat[,3:8]) )  # the model matrix X
> XX=t(X) %*% X  # a 7*7 matrix
> round( solve(XX), 4)   # round to 4 decimal places
          1      WT     DIS      NC      HP     ACC      ET
1    4.4918  0.6045  0.0019 -0.1734 -0.0210 -0.2361  0.1872
WT   0.6045  2.0845 -0.0092 -0.2052 -0.0239 -0.1081  0.7099
DIS  0.0019 -0.0092  0.0001 -0.0005  0.0001  0.0005 -0.0030
NC  -0.1734 -0.2052 -0.0005  0.1536  0.0009 -0.0011 -0.2001
HP  -0.0210 -0.0239  0.0001  0.0009  0.0005  0.0017 -0.0073
ACC -0.2361 -0.1081  0.0005 -0.0011  0.0017  0.0200 -0.0051
ET   0.1872  0.7099 -0.0030 -0.2001 -0.0073 -0.0051  0.7264
>
> # section 4.4, test whether beta4=beta5=beta6=0, the reduced model is
> gr = lm(GPM ~ WT + DIS + NC, dat)   #
> summary(gr)
```

```
Call:
lm(formula = GPM ~ WT + DIS + NC, data = dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.637375 -0.304545  0.003127  0.238986  0.644544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.639152   0.486450  -3.370 0.001886 **
WT           2.332838   0.288895   8.075 2.05e-09 ***
DIS         -0.010637   0.002697  -3.943 0.000381 ***
NC           0.218151   0.115982   1.881 0.068572 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.376 on 34 degrees of freedom
Multiple R-Squared: 0.9028,Adjusted R-squared: 0.8942
F-statistic: 105.3 on 3 and 34 DF,  p-value: < 2.2e-16

> SSE.gr = sum(resid(gr)^2); SSE.gr  # SSE(gr)
[1] 4.805601
>
> F0= (SSE.gr - SSE.g)/3 / (SSE.g/31); F0
[1] 6.015743
> 1- pf(F0, 3, 31)   # p-value is defined as the upper tail probability
[1] 0.002363145
>
> # test whether beta5=0, the reduced model is
> g5 = lm(GPM ~ WT + DIS + NC + HP  + ET, dat)   #
> SSE.g5 = sum(resid(g5)^2); SSE.g5  #
[1] 3.274703
> F0= (SSE.g5 - SSE.g)/1 / (SSE.g/31); F0
[1] 2.422476
> sqrt(F0)   # = the t-value of ACC in summary(g)
[1] 1.556431
> 1- pf(F0, 1, 31)   # same as the p-value of ACC in summary(g)
[1] 0.1297572
>
> # an easy way to do F test in R
> anova(gr, g)   # p. 118
Analysis of Variance Table

Model 1: GPM ~ WT + DIS + NC
Model 2: GPM ~ WT + DIS + NC + HP + ACC + ET
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     34 4.8056
2     31 3.0374  3    1.7683 6.0157 0.002363 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Section 4.4 Additional Sum of Squares Principle (Oral Contraceptive)

An experiment was conducted to determine the effects of 5 different oral contraceptive (OC) on high-density lipocholestrol (HDLC). 50 women were randomly divided into 5 groups; 10 women in each group. The variables are z =initial HDLC (before taking OCs), y =final HDLC (after 6 months) and OC type= 1,2,3,4 or 5.

```
> dat=read.table("http://www.stat.ucla.edu/~hqxu/stat100C/data/contraceptive.txt", h=T)  #  Table 1.3, p
> dim(dat)  # 10*10
[1] 10 10
> dat[1,]  # view first case to get the variable names
  Gp1Y Gp1X Gp2Y Gp2X Gp3Y Gp3X Gp4Y Gp4X Gp5Y Gp5X
1   43   49   58   56  100  102   50   57   41   37
> # we need rearrange the data
> y=c(dat[,1], dat[,3], dat[,5], dat[,7], dat[,9])  # stack 5 columns as a vector
> z=c(dat[,2], dat[,4], dat[,6], dat[,8], dat[,10])  # stack 5 columns as a vector
> type=c(rep(1, 10), rep(2,10), rep(3, 10), rep(4, 10), rep(5, 10))  # OC type
> # create 5 dummy variables
> x1=ifelse(type==1, 1, 0)
> x2=ifelse(type==2, 1, 0)
> x3=ifelse(type==3, 1, 0)
> x4=ifelse(type==4, 1, 0)
> x5=ifelse(type==5, 1, 0)
> # fit the full model (4.47) without intercept
> gf=lm(y~z+x1+x2+x3+x4+x5 -1)  # -1: a model without intercept
> summary(gf)
Call:
lm(formula = y ~ z + x1 + x2 + x3 + x4 + x5 - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-19.079  -4.768  -1.078   4.817  19.016

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
z    0.7124     0.0978   7.284 4.39e-09 ***
x1  12.0519     6.0817   1.982  0.05379 .
x2  12.7612     6.5252   1.956  0.05687 .
x3   8.3214     6.6711   1.247  0.21886
x4  12.8709     6.3073   2.041  0.04732 *
x5  17.7616     5.8665   3.028  0.00411 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.545 on 44 degrees of freedom
Multiple R-Squared: 0.9841,Adjusted R-squared: 0.9819
F-statistic: 454.3 on 6 and 44 DF,  p-value: < 2.2e-16

> # fit the reduced model (4.48) with  intercept
> gr=lm(y~z )  #
> summary(gr)
Call:
lm(formula = y ~ z)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-19.165  -5.179  -1.016   5.245  17.244

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.47718    5.88575   2.800  0.00735 **
z            0.64980    0.09716   6.688 2.21e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.816 on 48 degrees of freedom
Multiple R-Squared: 0.4824,Adjusted R-squared: 0.4716
F-statistic: 44.73 on 1 and 48 DF,  p-value: 2.214e-08

> # test whether beta1=beta2=...=beta5
> anova(gr, gf)  # p. 119
Analysis of Variance Table

Model 1: y ~ z
Model 2: y ~ z + x1 + x2 + x3 + x4 + x5 - 1
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     48 2931.96
2     44 2504.99  4    426.96 1.8749 0.1318
> # fit the reduced model (4.49) without intercept
> gr2=lm(y~z -1)  # -1: without intercept
> summary(gr2)
Call:
lm(formula = y ~ z - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-18.1835  -5.8115  -0.1437   6.6054  22.2387

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
z  0.91696    0.01948   47.08   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.343 on 49 degrees of freedom
Multiple R-Squared: 0.9784,Adjusted R-squared: 0.9779
F-statistic:  2216 on 1 and 49 DF,  p-value: < 2.2e-16

> # test whether beta1=beta2=...=beta5=0
> anova(gr2, gf)  # p. 120
Analysis of Variance Table

Model 1: y ~ z - 1
Model 2: y ~ z + x1 + x2 + x3 + x4 + x5 - 1
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     49 3410.7
2     44 2505.0  5     905.7 3.1816 0.01545 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```