HW6-Solutions

Problem 1

1- An analyst wanted to fit the regression model Yi = B0 + B1 Xi1 + B2 Xi2 + B3 Xi3 + Ei, i = 1,. . . ,n, by the method of least squares when it is known that B2 = 4. How can the analyst obtain the desired fit by using a multiple regression computer program? (20pts)

Solution:

Full model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

The reduced model below:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} Y_i = \beta_0 + \beta_1 X_{i1} + 4X_{i2} + \beta_3 X_{i3} Y_i - 4X_{i2} = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3}$$

Problem 2- Refer to the Commercial Properties data and problem in Assignment 5. (25 pts)

a) Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X4; with X1 given X4; with X2 , given X1 and X4; and with X3 , given X1, X2 and X4. (10pts)

b)Test whether X3 can be dropped from the regression model given that X1, X2 and X4 are retained. Use the F test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test? (5pts)

c)Test whether both X2 and X3 can be dropped from the regression model given that X1 and X4 are retained; use alpha =.01. State the alternatives, decision rule, and conclusion. What is the P-value of the test? (5pts)

d)Test whether, Beta1 = -.1 and, Beta2 =.4; Use alpha=.01. State the alternatives, full and reduced models, decision rule, and conclusion. (5pts)

a)

__Solution:

$$SSR(X_4) = 67.775$$
$$SSR(X_1|X_4) = 42.275$$
$$SSR(X_2|X_4, X_1) = 27.857$$
$$SSR(X_3|X_4, X_1, X_2) = 0.420$$
$$SSE(X_1, X_2, X_3, X_4) = 98.231$$

```
library(knitr)
Commercial.Properties <- read.csv("/cloud/project/Commercial Properties.csv")
f1<-lm(Y~X4+X1+X2+X3,data=Commercial.Properties)
anova(f1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X4         1 67.775  67.775 52.4369 3.073e-10 ***
## X1         1 42.275  42.275 32.7074 2.004e-07 ***
## X2         1 27.857  27.857 21.5531 1.412e-05 ***
## X3         1  0.420   0.420  0.3248    0.5704
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

Solution:

$$H_o : \beta_3 = 0 H_a : \beta_3 \neq 0$$

Pvalue of the test is 0.5704. Accept the null,

$$\beta_3$$

can be dropped from the model.

See below for the Rcode

```r
f1r<-lm(Y~X4+X1+X2,data=Commercial.Properties)
anova(f1r,f1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X4 + X1 + X2
## Model 2: Y ~ X4 + X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     77 98.650
## 2     76 98.231  1   0.41975 0.3248 0.5704
```

c)

Solution:

$$H_o : \beta_2 = \beta_3 = 0 H_a : Either\ \beta_2\ or\ \beta_3\ not\ equal\ to\ 0$$

Pvalue of the test is 6.682e-05 $<$ 0.05. Reject the null,

$$both\ \beta_2\ or\ \beta_3\ can\ Not\ be\ dropped\ from\ the\ model.$$

See below for the Rcode

```r
f1cr<-lm(Y~X4+X1,data=Commercial.Properties)
anova(f1cr,f1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X4 + X1
## Model 2: Y ~ X4 + X1 + X2 + X3
```

```
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     78 126.508
## 2     76  98.231  2    28.277 10.939 6.682e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d)

Solution:

$$H_o : \beta_1 = -0.1 \;, \beta_2 = 0.4 H_a : \textit{Not both equalities in } H_o \textit{ hold}$$

The reduced model is

$$Y_i + 0.1X_{i1} - 0.4X_{i2} = \beta_0 + \beta_3 X_3 + \beta_4 X_{i4} Y_i^* = \beta_0 + \beta_3 X_3 + \beta_4 X_{i4}$$

SSE.f=98.231 dF.f= 76 SSE.r=110.141 dF.r=78 F.test = ((110.141-98.231)/2)/(98.231/76)=4.607303

Pvalue of the test is 0.0129 which is greater than alpha=0.01. Accept the null.

See below for the Rcode

```
f1dr<-lm(Y+0.1*X1-0.4*X2~X3+X4,data=Commercial.Properties)
anova(f1dr)
```

```
## Analysis of Variance Table
##
## Response: Y + 0.1 * X1 - 0.4 * X2
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X3         1   9.205   9.205  6.5187   0.01263 *
## X4         1  31.872  31.872 22.5713 9.058e-06 ***
## Residuals 78 110.141   1.412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(f1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X4         1 67.775  67.775 52.4369 3.073e-10 ***
## X1         1 42.275  42.275 32.7074 2.004e-07 ***
## X2         1 27.857  27.857 21.5531 1.412e-05 ***
## X3         1  0.420   0.420  0.3248    0.5704
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
F.test=((110.141-98.231)/2)/(98.231/76)
F.test
```

```
## [1] 4.607303
```

```
1-pf(F.test,2,76)
```

```
## [1] 0.01292358
```

Problem 3

3- Refer to Brand preference data and problem in Assignment 5 (30 pts)

a) Transform the variables by means of the correlation transformation and fit the standardized regression model (10pts).

b) Interpret the standardized regression coefficient (5pts).

c) Transform the estimated standardized regression coefficients back to the ones for the fitted regression model in the original variables (5pts).

d) Calculate R2Y1, R2Y2, R212, R2Y1|2, R2Y2|1 and R2. Explain what each coefficient measures and interpret your results. (10pts)

a)

Solution:

Y=0.8923929X1 + 0.3945807X2 see below for the rcode

```
Brand.Preference <- read.csv("/cloud/project/Brand Preference.csv")
install.packages("QuantPsyc")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
library(QuantPsyc)
```

```
## Loading required package: boot

## Loading required package: MASS

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm
```

```
f3<-lm(Y~X1+X2,data=Brand.Preference)
summary(f3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = Brand.Preference)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## X1            4.4250     0.3011  14.695 1.78e-09 ***
## X2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

```
lm.beta(f3)
```

```
##        X1        X2
## 0.8923929 0.3945807
```

b)

Solution:

We see from the standardized regression coefficients that an increase of one standard deviation of X1 when X2 is fixed leads to a much larger increase in Y than does an increase of one standard deviation of X2 when X1 is fixed.

c)

Solution: **See below the code**

```
s<-sqrt(var(Brand.Preference))
sy<-s[1,1]
sx1<-s[2,2]
sx2<-s[3,3]
b1=(sy/sx1)*0.8923929
b2=(sy/sx2)*0.3945807
b0=mean(Brand.Preference$Y)-b1*mean(Brand.Preference$X1)-b2*mean(Brand.Preference$X2)
cbind(b0,b1,b2)
```

```
##         b0    b1    b2
## [1,] 37.65 4.425 4.375
```

d)

Solution:

$R^2_{Y1} = 0.7964 R^2_{Y2} = 0.1557 R^2_{12} = 0 R^2_{Y1|2} = 1566.45/1660.75 = 0.9432184 R^2_{Y2|1} = 306.25/400.55 = 0.7645737 R^2 = 0.9521$

see below for the R code

```
f3.1<-lm(Y~X1,data=Brand.Preference)
summary(f3.1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = Brand.Preference)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.475 -4.688 -0.100  4.638  7.525
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.775      4.395  11.554 1.52e-08 ***
## X1             4.425      0.598   7.399 3.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.349 on 14 degrees of freedom
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7818
## F-statistic: 54.75 on 1 and 14 DF,  p-value: 3.356e-06
```

```r
anova(f3.1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45  54.751 3.356e-06 ***
## Residuals 14  400.55   28.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f3.2<-lm(Y~X2,data=Brand.Preference)
summary(f3.2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = Brand.Preference)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.375  -7.312  -0.125   8.688  16.625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   68.625      8.610   7.970 1.43e-06 ***
## X2             4.375      2.723   1.607     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 14 degrees of freedom
## Multiple R-squared:  0.1557, Adjusted R-squared:  0.09539
## F-statistic: 2.582 on 1 and 14 DF,  p-value: 0.1304
```

```r
anova(f3.2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## X2         1  306.25  306.25  2.5817 0.1304
## Residuals 14 1660.75  118.62
```

```r
f3.3<-lm(Y~X2+X1,data=Brand.Preference)
anova(f3.3)
```

```
## Analysis of Variance Table
```

```
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
f3.4<-lm(Y~X1+X2,data=Brand.Preference)
anova(f3.4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(f3.4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = Brand.Preference)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## X1            4.4250     0.3011  14.695 1.78e-09 ***
## X2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

```
cor(Brand.Preference)
```

```
##            Y        X1        X2
## Y  1.0000000 0.8923929 0.3945807
## X1 0.8923929 1.0000000 0.0000000
## X2 0.3945807 0.0000000 1.0000000
```

Problem 4

4-Refer to the CDI data set. For predicting the number of active physicians (Y) in a county, it has been decided to include total population (X1) and total personal income (X2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. Assume that a first-order multiple regression model is appropriate. (25 pts)

a) For each of the following variables, calculate the coefficient of partial determination given that X1 and X2 are included in the model: land area (X3), percent of population 65 or older (X4), number of hospital beds (X5), and total serious crimes (X6). (15pts)

b) On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables? (5pts)

c) Using the F* test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when X1 and X2 are included in the model; use alpha=.01. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other three potential predictor variables be as large as the one here? (5pts)

a)

Solution:

$$R^2_{Y3|12} = 4063370/140967081 = 0.02882496$$

$$R^2_{Y4|12} = 541647/140967081 = 0.03842365$$

$$R^2_{Y5|12} = 78070132/140967081 = 0.5538182$$

$$R^2_{Y6|12} = 1032359/140967081 = 0.007323405$$

see below for the Rcode

```
CDI <- read.csv("/cloud/project/CDI.csv")
Y=CDI$Number.of.active.physicians
X1=CDI$Total.population
X2=CDI$Total.personal.income
X3=CDI$Land.area
X4=CDI$Percent.of.population.65.or.older
X5=CDI$Number.of.hospital.beds
X6=CDI$Total.serious.crimes
f4.12<-lm(Y~X1+X2)
anova(f4.12)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## X1          1 1243181164 1243181164 3853.88 < 2.2e-16 ***
## X2          1   22058054   22058054   68.38 1.638e-15 ***
## Residuals 437  140967081     322579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f4.3<-lm(Y~X1+X2+X3)
anova(f4.3)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df     Sum Sq    Mean Sq  F value    Pr(>F)
## X1           1 1243181164 1243181164 3959.184 < 2.2e-16 ***
## X2           1   22058054   22058054   70.249 7.271e-16 ***
## X3           1    4063370    4063370   12.941 0.0003583 ***
## Residuals 436  136903711     313999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f4.4<-lm(Y~X1+X2+X4)
anova(f4.4)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df     Sum Sq    Mean Sq   F value    Pr(>F)
## X1           1 1243181164 1243181164 3859.8919 < 2.2e-16 ***
## X2           1   22058054   22058054   68.4870 1.571e-15 ***
## X4           1     541647     541647    1.6817    0.1954
## Residuals 436  140425434     322077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f4.5<-lm(Y~X1+X2+X5)
anova(f4.5)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## X1           1 1243181164 1243181164 8617.70 < 2.2e-16 ***
## X2           1   22058054   22058054  152.91 < 2.2e-16 ***
## X5           1   78070132   78070132  541.18 < 2.2e-16 ***
## Residuals 436   62896949     144259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f4.6<-lm(Y~X1+X2+X6)
anova(f4.6)
```

```
## Analysis of Variance Table
##
## Response: Y
##             Df     Sum Sq    Mean Sq   F value    Pr(>F)
## X1           1 1243181164 1243181164 3873.4274 < 2.2e-16 ***
## X2           1   22058054   22058054   68.7271 1.414e-15 ***
## X6           1    1032359    1032359    3.2166   0.07359 .
## Residuals 436  139934722     320951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

Solution: **X5 is th best variable as it has the highest coefficient of partial determination.**

c)

Solution:

$$H_o : \beta_5 = 0 H_a : \beta_5 \neq 0$$

Pvalue of the test is 2.2e-16. Reject the null,

$$\beta_5$$

is significant and it should be added to the model.

See below for the Rcode

```
f4.125<-lm(Y~X1+X2+X5)
anova(f4.12,f4.125)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X5
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1    437 140967081
## 2    436  62896949  1  78070132 541.18 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```