CS-E-106: Data Modeling

Assignment 6

Instructor: Hakan Gogtas
Submitted by: *Saurabh Kulkarni*

Due Date: *11/11/2019*

Question 1: An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_1 * X_{i3} + \epsilon_i$, i = 1, … , n, by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit by using a multiple regression computer program?

Solution:

Step 1: Create a new variable $Y^* = Y_i - \beta_2 * X_{i2} = Y_i - 4 * X_{i2}$ *Step 2:* Using Y^* as the new response variable, run the regression model: $Y^* = \beta_0 + \beta_1 * X_{i1} + \beta_3 * X_{i3} + \epsilon_i$ with least squares method. Using R functions: `model = lm(YStar~X1+X3, data)`. *Step 3:* Use the obtained coefficients as β_1 and β_3, assuming $\beta_2 = 4$

Question 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)

Solution:

```
properties_data = read.csv("Commercial Properties.csv")
lm_prop = lm(Y~X4+X1+X2+X3, data=properties_data)
summary(lm_prop)
```

```
##
## Call:
## lm(formula = Y ~ X4 + X1 + X2 + X3, data = properties_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

```
anova_F = anova(lm_prop)

anovaTable = data.frame(anova_F)
```

```r
totals = c(round(sum(anovaTable[,1])), round(sum(anovaTable[,2])), "", "", "")
anovaTable = rbind(anovaTable, totals)

#add names to the table
row.names(anovaTable) = c("SSR(X4)", "SSR(X1|X4)", "SSR(X2|(X4X1))","SSR(X3|(X4X1X2))", "SSE", "Total")
colnames(anovaTable) = c("DF", "Sum Sq.", "Mean Sq.", "F-Value", "Pr(>F)")

kable(anovaTable)
```

	DF	Sum Sq.	Mean Sq.	F-Value	Pr(>F)
SSR(X4)	1	67.7750979864736	67.7750979864736	52.4368960852129	3.07327030821117e-10
SSR(X1|X4)	1	42.2745683242813	42.2745683242813	32.7073986187373	2.00386962405898e-07
SSR(X2|(X4X1))	1	27.8574934834163	27.8574934834163	21.5530561280181	1.41220768697313e-05
SSR(X3|(X4X1X2))	1	0.419746262940206	0.419746262940206	0.324753365555366	0.570445705115829
SSE	76	98.2305939428886	1.29250781503801	NA	NA
Total	80	237			

(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)

Solution:

From the above ANOVA table, we can see that the P-value for SSR(X3|X4X1X2) is very high, which means that the extra regression sums of squares due to X3 is very low. Thus, X3 can be dropped. *See F-test below.*

```r
ssr = as.numeric(anovaTable["SSR(X3|(X4X1X2))","Sum Sq."])
sse = as.numeric(anovaTable["SSE","Sum Sq."])
df_diff = 1
df_E = as.numeric(anovaTable["SSE","DF"])

FStar = (ssr/df_diff) / (sse/df_E)
print(FStar)
```

```
## [1] 0.3247534
```

```r
print(paste("P-value:", 1-pf(FStar, df_diff, df_E)))
```

```
## [1] "P-value: 0.570445705115829"
```

```r
#alpha is given
alpha = 0.01

# df from Summary above in a
FTest = qf(1-alpha, df_diff, df_E)
print(FTest)
```

```
## [1] 6.980578
```

Hypotheses:

$H_0 : \beta_3 = 0$

$H_a : \beta_3 \neq 0$

Decision Rules:

If $F^* \leq 6.9805778$, conclude H_0

If $F^* > 6.9805778$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 0.3247534$, and $0.3247534 \leq 6.9805778$, we conclude H_0. Thus, X3 can be dropped from the model.

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

Solution:

```
ssr = sum(as.numeric(anovaTable[c("SSR(X3|(X4X1X2))","SSR(X2|(X4X1))"),"Sum Sq."]))
sse = as.numeric(anovaTable["SSE","Sum Sq."])

df_diff = 2
df_E = as.numeric(anovaTable["SSE","DF"])

FStar = (ssr/df_diff) / (sse/df_E)
print(FStar)
```

```
## [1] 10.9389
```

```
print(paste("P-value:", 1-pf(FStar, df_diff, df_E)))
```

```
## [1] "P-value: 6.68213642763815e-05"
```

```
#alpha is given
alpha = 0.01

# df from Summary above in a
FTest = qf(1-alpha, df_diff, df_E)
print(FTest)
```

```
## [1] 4.89584
```

Hypotheses:

$H_0 : \beta_2 = \beta_3 = 0$

H_a : Not both βs equal to zero

Decision Rules:

If $F^* \leq 4.8958399$, conclude H_0

If $F^* > 4.8958399$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 10.9389047$, and $10.9389047 > 4.8958399$, we conclude H_1. Not both βs equal to zero.

(d) Test whether, 1 = -.1 and, 2 =.4; Use =.01. State the alternatives, full and reduced models, decision rule, and conclusion. (5pts)

Solution:

```
Y = properties_data$Y+0.1*properties_data$X1-0.4*properties_data$X2
lm_prop_R = lm(Y~properties_data$X3+properties_data$X4)
summary(lm_prop_R)
```

```
##
## Call:
## lm(formula = Y ~ properties_data$X3 + properties_data$X4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8267 -0.6642 -0.0671  0.5533  3.5096
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.094e+01  2.446e-01  44.737  < 2e-16 ***
## properties_data$X3 2.142e+00  9.906e-01   2.162   0.0337 *
## properties_data$X4 5.804e-06  1.222e-06   4.751 9.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.188 on 78 degrees of freedom
## Multiple R-squared:  0.2716, Adjusted R-squared:  0.253
## F-statistic: 14.55 on 2 and 78 DF,  p-value: 4.28e-06
```

```r
anova_R = anova(lm_prop_R)
anova_R
```

```
## Analysis of Variance Table
##
## Response: Y
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## properties_data$X3  1   9.205   9.205  6.5187   0.01263 *
## properties_data$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
```

```r
anova(lm_prop_R, lm_prop)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ properties_data$X3 + properties_data$X4
## Model 2: Y ~ X4 + X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     78 110.141
## 2     76  98.231  2    11.911 4.6076 0.01292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSE_R = anova_R["Residuals", "Sum Sq"]
SSE_F = anova_F["Residuals", "Sum Sq"]
df_R = anova_R["Residuals", "Df"]
df_F = anova_F["Residuals", "Df"]

FStar = ((SSE_R-SSE_F)/(df_R-df_F))/(SSE_F/df_F)
print(FStar)
```

```
## [1] 4.60764
```

```r
#alpha is given
alpha = 0.01
```

```r
# df from Summary above in a
FTest = qf(1-alpha, (df_R-df_F), df_F)
print(FTest)
```

```
## [1] 4.89584
```

Hypotheses:

$H_0 : \beta_1 = -0.1$ & $\beta_2 = 0.4$

H_a : Not both $\beta_1 = -0.1$ & $\beta_2 = 0.4$ hold

Decision Rules:

If $F^* \leq 4.8958399$, conclude H_0

If $F^* > 4.8958399$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 4.6076402$, and $4.6076402 \approx 4.8958399$, at the boundary of decision rule. We may wish to conduct further analyses whether $\beta_1 = -0.1$ & $\beta_2 = 0.4$ can hold true at the same time.

Question 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).

Solution:

```r
standardize_corr = function(df){
  cols = colnames(df)
  df_new = df
  n = nrow(df)
  for(c in cols){
    mu = mean(df[, c])
    s = sqrt(var(df[, c]))
    df_new[, c] = (df[, c]-mu)/(s*sqrt(n-1))
  }
  df_new
}
```

```r
brand_data = read.csv("Brand Preference.csv")
brand_data_new = standardize_corr(brand_data)
#spot check
summary(brand_data_new)
```

```
##        Y                   X1                 X2
##  Min.   :-0.46786   Min.   :-0.3354   Min.   :-0.25
##  1st Qu.:-0.20293   1st Qu.:-0.1677   1st Qu.:-0.25
##  Median : 0.02818   Median : 0.0000   Median : 0.00
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00
##  3rd Qu.: 0.18602   3rd Qu.: 0.1677   3rd Qu.: 0.25
##  Max.   : 0.41149   Max.   : 0.3354   Max.   : 0.25
```

```r
lm_brand_new = lm(Y~., data=brand_data_new)
summary(lm_brand_new)
```

```
##
## Call:
## lm(formula = Y ~ ., data = brand_data_new)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.099209 -0.039740  0.000564  0.035794  0.094699
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.238e-17  1.518e-02   0.000        1
## X1           8.924e-01  6.073e-02  14.695 1.78e-09 ***
## X2           3.946e-01  6.073e-02   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06073 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
```

β_0 is pretty much 0. Thus,

Regression Function: $Y = 0.892 * X_1 + 0.3946 * X_2$

```
lm_brand = lm(Y~., data=brand_data)
```

Double-check Using QuantPsyc library:

```
library(QuantPsyc)
```

```
## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma

## The following object is masked from 'package:alr3':
##
##     wool

## The following object is masked from 'package:car':
##
##     logit

##
## Attaching package: 'QuantPsyc'

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

```
lm.beta(lm_brand)
```

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

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

Solution:

Interpretation:

- We can see that the β_0 is almost equal to 0, which is expected since Y is now centered at 0 (based on definition and summary of the new data in part (a)).
- We also see that β_1 and β_2 are less than one and their standard deviations are also a lot less than the model without standardization.
- However, the R^2 is the same which means that the is equally powerful in terms of predictive capability.

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

Solution:

```
df = brand_data

b1_star = as.numeric(lm.beta(lm_brand)["X1"])
b1 = (sqrt(var(df[, "Y"]))/sqrt(var(df[, "X1"])))*b1_star
print(b1)
```

```
## [1] 4.425
```

```
b2_star = as.numeric(lm.beta(lm_brand)["X2"])
b2 = (sqrt(var(df[, "Y"]))/sqrt(var(df[, "X2"])))*b2_star
print(b2)
```

```
## [1] 4.375
```

```
b0 = mean(df$Y)-b1*mean(df$X1)-b2*mean(df$X2)
print(b0)
```

```
## [1] 37.65
```

Double-check

```
lm_brand$coefficients
```

```
## (Intercept)          X1          X2
##      37.650       4.425       4.375
```

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

Solution:

```
df = brand_data

R2_Y1 = anova(lm(Y~X1, data=df))[1,2]/sum(anova(lm(Y~X1, data=df))[1:2,2])

R2_Y2 = anova(lm(Y~X2, data=df))[1,2]/sum(anova(lm(Y~X2, data=df))[1:2,2])

R2_12 = cor(df$X1, df$X2)^2

R2_Y1_2 = anova(lm(Y~X2+X1, data=df))[2,2]/sum(anova(lm(Y~X2+X1, data=df))[2:3,2])

R2_Y2_1 = anova(lm(Y~X1+X2, data=df))[2,2]/sum(anova(lm(Y~X1+X2, data=df))[2:3,2])

R2 = sum(anova(lm(Y~X1+X2, data=df))[1:2,2])/sum(anova(lm(Y~X1+X2, data=df))[1:3,2])
```

$R^2_{Y1} = 0.796365$

$R^2_{Y2} = 0.155694$

$R^2_{12} = 0$

$R^2_{Y1|2} = 0.9432184$

$R^2_{Y2|1} = 0.7645737$

$R^2 = 0.952059$

Question 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)

Solution:

```r
inc_cols = c("Number.of.active.physicians", "Total.population", "Total.personal.income",
             "Land.area", "Percent.of.population.65.or.older", "Number.of.hospital.beds",
             "Total.serious.crimes")
cdi_data = read.csv("CDI.csv")[, inc_cols]
colnames(cdi_data)
```

```
## [1] "Number.of.active.physicians"       "Total.population"
## [3] "Total.personal.income"             "Land.area"
## [5] "Percent.of.population.65.or.older" "Number.of.hospital.beds"
## [7] "Total.serious.crimes"
```

```r
cdi_data =
cdi_data %>%
  rename(
    Y = Number.of.active.physicians,
    X1 = Total.population,
    X2 = Total.personal.income,
    X3 = Land.area,
    X4 = Percent.of.population.65.or.older,
    X5 = Number.of.hospital.beds,
    X6 = Total.serious.crimes
  )

colnames(cdi_data)
```

```
## [1] "Y"  "X1" "X2" "X3" "X4" "X5" "X6"
```

```r
lm_cdi = lm(Y~X1+X2, data=cdi_data)
summary(lm_cdi)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = cdi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1849.1  -198.3   -71.4    39.7  3755.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.444e+01  3.283e+01  -1.963   0.0503 .
```

```
## X1           5.310e-04  2.775e-04   1.914   0.0563 .
## X2           1.072e-01  1.297e-02   8.269 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 568 on 437 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8993
## F-statistic:  1961 on 2 and 437 DF,  p-value: < 2.2e-16
```

```r
df = cdi_data

print(anova(lm(Y~X1+X2+X3,df)))
```

```
## 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
(r2_X3 = anova(lm(Y~X1+X2+X3,df))[3,2]/sum(anova(lm(Y~X1+X2+X3,df))[3:4,2]))
```

```
## [1] 0.02882495
```

```r
print(anova(lm(Y~X1+X2+X4,df)))
```

```
## 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
(r2_X4 = anova(lm(Y~X1+X2+X4,df))[3,2]/sum(anova(lm(Y~X1+X2+X4,df))[3:4,2]))
```

```
## [1] 0.003842367
```

```r
print(anova(lm(Y~X1+X2+X5,df)))
```

```
## 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
```

```
(r2_X5 = anova(lm(Y~X1+X2+X5,df))[3,2]/sum(anova(lm(Y~X1+X2+X5,df))[3:4,2]))
```

```
## [1] 0.5538182
```

```
print(anova(lm(Y~X1+X2+X6,df)))
```

```
## 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
```

```
(r2_X6 = anova(lm(Y~X1+X2+X6,df))[3,2]/sum(anova(lm(Y~X1+X2+X6,df))[3:4,2]))
```

```
## [1] 0.007323408
```

$R^2_{3|12} = 0.028825$

$R^2_{4|12} = 0.0038424$

$R^2_{5|12} = 0.5538182$

$R^2_{6|12} = 0.0073234$

(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)

Solution:

X_5 is the best predictor we can add to the model as it has the maximum coefficient of partial determination. Yes, the extra sum of squares associated with this variable is larger compared to other variables also, which makes sense since SST will remain constant.

(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 =.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)

Solution:

```
anova_x5 = anova(lm(Y~X1+X2+X5, data=cdi_data))
anova_x5
```

```
## 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
anova(lm_cdi, lm(Y~X1+X2+X5, data=cdi_data))
```

```
## 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
```

```r
ssr = as.numeric(anova_x5["X5","Sum Sq"])
sse = as.numeric(anova_x5["Residuals","Sum Sq"])
df_diff = 1
df_E = as.numeric(anova_x5["Residuals","Df"])

FStar = (ssr/df_diff) / (sse/df_E)
print(FStar)
```

```
## [1] 541.1801
```

```r
print(paste("P-value:", 1-pf(FStar, df_diff, df_E)))
```

```
## [1] "P-value: 0"
```

```r
#alpha is given
alpha = 0.01

# df from Summary above in a
FTest = qf(1-alpha, df_diff, df_E)
print(FTest)
```

```
## [1] 6.693358
```

Hypotheses:

$H_0 : \beta_5 = 0$

$H_a : \beta_5 \neq 0$

Decision Rules:

If $F^* \leq 6.6933576$, conclude H_0

If $F^* > 6.6933576$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 541.1800993$, and $541.1800993 > 6.6933576$, we conclude H_a. Thus, X5 adds valuable information to the model and is helpful in predicting the response, given X1 and X2 are already present.