CS-E-106: Data Modeling

Assignment 3

Instructor: Hakan Gogtas
Submitted by: *Saurabh Kulkarni*

Due Date: *10/07/2019*

Solution 1:

```
reg_loop <- function(df, x_cols, y_str) {
  r2_list = c()
  lm_fits = list({})
  for(i in 1:length(x_cols)){
    x_str = x_cols[i]
    formula = as.formula(paste(y_str,"~", x_str))
    lm = lm(formula, data=df)
    r2_list[[i]] = summary(lm)$r.squared
  }
  r2_df = data.frame(cbind(x_cols, r2_list))
  ordered_df = r2_df[order(r2_list, decreasing = TRUE),]
  print(paste("Variable with maximum R-squared:",x_cols[which(r2_list == max(r2_list))]))
  print(paste("R-squared value:",r2_list[which(r2_list == max(r2_list))]))
  return(ordered_df)
}
```

```
cdi = read.csv("cdi.csv")
colnames(cdi)
```

```
##  [1] "Identification.number"
##  [2] "County"
##  [3] "State"
##  [4] "Land.area"
##  [5] "Total.population"
##  [6] "Percent.of.population.aged.18.34"
##  [7] "Percent.of.population.65.or.older"
##  [8] "Number.of.active.physicians"
##  [9] "Number.of.hospital.beds"
## [10] "Total.serious.crimes"
## [11] "Percent.high.school.graduates"
## [12] "Percent.bachelor.s.degrees"
## [13] "Percent.below.poverty.level"
## [14] "Percent.unemployment"
## [15] "Per.capita.income"
## [16] "Total.personal.income"
## [17] "Geographic.region"
```

```
exc = c("Identification.number", "Number.of.active.physicians")
x_cols = setdiff(colnames(cdi), exc)
y_str = "Number.of.active.physicians"
r2_df = reg_loop(df=cdi, x_cols = x_cols, y_str=y_str)
```

```
## [1] "Variable with maximum R-squared: County"
## [1] "R-squared value: 0.921236638865801"
```

Thus, **county accounts for maximum variability in the number of active physicians**. The remainder variables and their respective R^2 is given below in descending order of importance.

```r
r2_df
```

```
##                                x_cols              r2_list
## 1                              County    0.921236638865801
## 7            Number.of.hospital.beds    0.903382565497334
## 14             Total.personal.income    0.898913655463206
## 4                   Total.population    0.884067412249688
## 8              Total.serious.crimes    0.673153752663095
## 13                 Per.capita.income   0.0999411008221881
## 2                              State    0.063458287522148
## 10        Percent.bachelor.s.degrees   0.0560578858594275
## 5   Percent.of.population.aged.18.34   0.0143279081163583
## 3                          Land.area  0.00609565213240784
## 11       Percent.below.poverty.level  0.00411345912581384
## 12              Percent.unemployment  0.00255187801271758
## 15                 Geographic.region 0.000607428795977754
## 9      Percent.high.school.graduates  1.8046222736129e-05
## 6  Percent.of.population.65.or.older 9.78832264836169e-06
```

Solution 2:

```r
confint_regions <- function(df) {
  regions = levels(factor(df$Geographic.region))
  lm_fits = list({})
  for(i in regions){
    formula = as.formula(paste("Number.of.active.physicians","~", "Percent.bachelor.s.degrees"))
    lm = lm(formula, data=df[df$Geographic.region==i,])
    lm_fits[[i]] = lm
    print(paste("Region:",i))
    print(confint(lm, level=0.9)[2,])
    print(summary(lm)$coefficients)
    cat("\n")
  }

}
```

```r
confint_regions(df=cdi)
```

```
## [1] "Region: 1"
##       5 %     95 %
## 37.52342 99.62067
##                              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                -399.33942  429.03669 -0.9307815 0.3541860656
## Percent.bachelor.s.degrees   68.57205   18.70308  3.6663503 0.0003944817
##
## [1] "Region: 2"
##       5 %     95 %
## 15.34141 88.23268
##                              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)                -163.67212  463.91810 -0.3528039 0.72493602
## Percent.bachelor.s.degrees   51.78704   21.96372  2.3578445 0.02021586
##
## [1] "Region: 3"
```

```
##       5 %      95 %
## 21.38184 58.19190
##                              Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)                 -16.12031  251.04144 -0.06421374 0.9488855299
## Percent.bachelor.s.degrees  39.78687   11.12036  3.57784135 0.0004669194
##
## [1] "Region: 4"
##         5 %       95 %
##  -2.927066 149.839187
##                               Estimate Std. Error    t value  Pr(>|t|)
## (Intercept)                 -265.54556 1063.69869 -0.2496436 0.8035455
## Percent.bachelor.s.degrees   73.45606   45.86403  1.6016049 0.1134469
```

Thus, the slopes for the regression lines for differen regions vary from one another.

Solution 3:

(a)

```r
gpa = read.csv("GPA.csv")
lm_gpa = lm(GPA~ACT,  data=gpa)
summary(lm_gpa)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## ACT          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

```r
anova(lm_gpa)
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value   Pr(>F)
## ACT         1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b)

$MSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

MSR measures the effect of the regression line in explaining the total variation in Y_i.

$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$

MSE measures the mean variation of Y_i around the regresion line. Its the average of all the squared distances by which the regression line missed the actual Y_i.

$E[MSE] = \sigma^2$

$E[MSR] = \sigma^2 + \beta_1 \sum_{i=1}^{n} (X_i - \bar{X})^2$

Thus, MSE and MSR will estimate same quantity when $\beta_1 = 0$ i.e. $Y_i = \bar{Y}$

(c)

Null Hypothesis: $H_0 : \beta_1 = 0$; *Alternate Hypothesis:* $H_1 : \beta_1 \neq 0$

Decision Rule:

$F^* = \frac{MSR}{MSE}$

- If $F* \leq F(1-\alpha; 1, n-2)$, conclude H_0;
- If $F* \geq F(1-\alpha; 1, n-2)$, conclude H_1

```
MSR = 3.5878
MSE = 0.3883
F = MSR/MSE
print(F)
```

```
## [1] 9.239763
```

```
help(pf)
pf(q=0.01, 1, 118)
```

```
## [1] 0.07948598
```

Result:

Thus, since $F^* > F(1-\alpha; 1, n-2)$, we conclude that $H_1 : \beta_1 \neq 0$ holds.

(d)

The absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model is SST - SSE = SSR = 3.588 (from the ANOVA table above).

The relative measure is given by $\frac{SSR}{SST} = \frac{3.588}{3.588+45.818} = 0.0726$. This measure is also known as the R^2 or the coefficient of determination.

```
R_sq = 3.588/(3.588+45.818)
R_sq
```

```
## [1] 0.07262276
```

(e)

```
r = sqrt(R_sq)
r
```

```
## [1] 0.2694861
```

Looking at the summary of the regression model for GPA dataset, β_1 is positive. Hence, $r = +0.27$.

(f)

Operationally, R^2 has more clear interpretation.

- R^2 is the proportion of total variation in Y explained by X. Thus, it is a relative measure of improvement that was made by the introduction of X in the regression model. This can be used in a more direct way compared to r which measures linear association between X and Y.
- R^2 is on a scale of 0 to 1 (1 indicating the highest correlation), whereas, r ranges from -1 to 1 (the extremes indicating highest correlation). Meaning both r=-1 and r=+1 can mean the same level of association between X and Y. Also, the objective of the coefficients of correlation/determination is to measure the overall effectiveness of the model rather than looking at which direction the regression line is going. This is better accomplished by R^2.

Solution 4:

(a)

```
crime = read.csv("Crime Rate.csv")
cor.test(crime$Y,crime$X,method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  crime$Y and crime$X
## t = -4.1029, df = 82, p-value = 9.571e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5761223 -0.2175580
## sample estimates:
##        cor
## -0.4127033
```

(b)

```
lm_crime = lm(Y~X, data=crime)
summary(lm_crime)
```

```
##
## Call:
## lm(formula = Y ~ X, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## X            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```
anova(lm_crime)
```

```
## Analysis of Variance Table
##
## Response: Y
```

```
##           Df    Sum Sq  Mean Sq F value    Pr(>F)    
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis: $H_0 : \beta_1 = 0$; *Alternate Hypothesis:* $H_1 : \beta_1 \neq 0$

Decision Rule:

$F^* = \frac{MSR}{MSE}$

- If $F* \leq F(1-\alpha; 1, n-2)$, conclude H_0;
- If $F* \geq F(1-\alpha; 1, n-2)$, conclude H_1

```r
MSR = 93462942
MSE = 5552112
F = MSR/MSE
print(F)
```

```
## [1] 16.83376
```

```r
pf(q=0.01, 1, 82)
```

```
## [1] 0.07941159
```

Result:

Thus, since $F^* > F(1-\alpha; 1, n-2)$, we conclude that $H_1 : \beta_1 \neq 0$ holds.

(c)

```r
cor.test(crime$Y,crime$X, method="spearman")
```

```
## Warning in cor.test.default(crime$Y, crime$X, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
## 
##  Spearman's rank correlation rho
## 
## data:  crime$Y and crime$X
## S = 140839, p-value = 5.359e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho 
## -0.4259324
```

(d)

Null Hypothesis: There is no association between X and Y; *Alternate Hypothesis:* There is an association between X and Y

Decision Rule:

$t^* = \frac{r_s\sqrt{n-2}}{1-r_s^2}$

- If $|t^*| \leq t(1-\alpha/2; n-2)$, conclude H_0;
- If $|t^*| \geq t(1-\alpha/2; n-2)$, conclude H_1

```r
r_s = -0.4259324
n = nrow(crime)
```

```r
t = (r_s*sqrt(n-2)/(1-r_s^2))
t
```

```
## [1] -4.711787
```

```r
pt(0.005, 82)
```

```
## [1] 0.5019886
```

Result:

Thus, since $|t^*| \geq t(1-\alpha/2; n-2)$, we conclude that H_1 that there is an association between X and Y.