HUDM 5126 Linear Models and Regression Analysis Homework 6

Yifei Dong

10/7/2020

0. Data Preparation

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(rsq)
```

1. Commerical Properties

Refer to **Commercial properties** Problems 6.18

a). Obtain the analysis of variance tables that decomposes the regression sum of squares into extra sum of squares associated with X_4; with X_1, given X_4; with X_2,given X_1 and X_4; and with X_3, given X_1, X_2 and X_4.

```
# Load data
mydata <- read.table(paste("http://users.stat.ufl.edu",
                           "/~rrandles/sta4210/Rclassnotes",
                           "/data/textdatasets/KutnerData",
                           "/Chapter%20%206%20Data%20Sets/CH06PR18.txt", sep=""))

# Rename variales us deplyr package
mydata <- mydata %>%
  rename("Y"=V1, "X1"=V2, "X2"=V3, "X3"=V4, "X4"=V5)
head(mydata, 10) # Check the first 10 rows of dataset
```

```
##        Y X1    X2   X3     X4
## 1   13.5  1  5.02 0.14 123000
## 2   12.0 14  8.19 0.27 104079
## 3   10.5 16  3.00 0.00  39998
## 4   15.0  4 10.70 0.05  57112
## 5   14.0 11  8.97 0.07  60000
## 6   10.5 15  9.45 0.24 101385
```

```
## 7  14.0  2  8.00 0.19  31300
## 8  16.5  1  6.62 0.60 248172
## 9  17.5  1  6.20 0.00 215000
## 10 16.5  8 11.78 0.03 251015
```

```r
attach(mydata)
reg1234 <- lm(Y~X1+X2+X3+X4)
anova(reg1234)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 14.819  14.819 11.4649  0.001125 **
## X2         1 72.802  72.802 56.3262 9.699e-11 ***
## X3         1  8.381   8.381  6.4846  0.012904 *
## X4         1 42.325  42.325 32.7464 1.976e-07 ***
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_1, X_2, X_3, X_4) = 98.231$

```r
reg4 <-lm(Y~X4)
anova(reg4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  31.723 2.628e-07 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_4) = 67.775$

```r
reg41 <- lm(Y~X4+X1)
anova(reg41)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  41.788 8.076e-09 ***
## X1         1  42.275  42.275  26.065 2.275e-06 ***
## Residuals 78 126.508   1.622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_1|X_4) = 42.275$

```r
reg142 <- lm(Y~X1+X4+X2)
anova(reg142)
```

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

```
##             Df Sum Sq Mean Sq F value    Pr(>F)    
## X1          1 14.819  14.819  11.566  0.001067 ** 
## X4          1 95.231  95.231  74.331 6.439e-13 ***
## X2          1 27.857  27.857  21.744 1.287e-05 ***
## Residuals 77 98.650   1.281                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_2|X_1, X_4) = 27.857$

```r
reg1243 <-lm(Y~X1+X2+X4+X3)
anova(reg1243)
```

```
## Analysis of Variance Table
## 
## Response: Y
##             Df Sum Sq Mean Sq F value    Pr(>F)    
## X1          1 14.819  14.819 11.4649  0.001125 ** 
## X2          1 72.802  72.802 56.3262 9.699e-11 ***
## X4          1 50.287  50.287 38.9062 2.306e-08 ***
## X3          1  0.420   0.420  0.3248  0.570446    
## Residuals 76 98.231   1.293                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_3|X_1, X_2, X_4) = 0.420$

b). Test whether X_3 can be dropped from the regression model given that X_1, X_2 and X_4 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?

```r
anova(reg1243)
```


```
## Analysis of Variance Table
## 
## Response: Y
##             Df Sum Sq Mean Sq F value    Pr(>F)    
## X1          1 14.819  14.819 11.4649  0.001125 ** 
## X2          1 72.802  72.802 56.3262 9.699e-11 ***
## X4          1 50.287  50.287 38.9062 2.306e-08 ***
## X3          1  0.420   0.420  0.3248  0.570446    
## Residuals 76 98.231   1.293                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


ANSWER: $H_0 : \beta_3 = 0$ and $H_a : \beta_3 \neq 0$. Partial F-statistic is $F^* = 0.3248$ and corresponding P-value=0.5704, which is greater than 0.01. Therefore, we cannot reject the null hypothesis and conclude that X_3 is not significant and can be dropped from the regression model given that X_1, X_2 and X_4 are retained.

2. Continue: Commerical Properties

Refer to **Commerical properties** Problem 6.18 and 7.7. Calculate $R^2_{Y4}, R^2_{Y1}, R^2_{Y1|4}, R^2_{14}, R^2_{Y2|14}, R^2_{Y3|124}$ and R^2. Explain what each coefficient measures and interpret your results. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_4?

```r
summary(reg4)
```

```
##
## Call:
## lm(formula = Y ~ X4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1390 -0.7930  0.2890  0.9653  3.4415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.378e+01  2.903e-01  47.482  < 2e-16 ***
## X4          8.437e-06  1.498e-06   5.632 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 79 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2775
## F-statistic: 31.72 on 1 and 79 DF,  p-value: 2.628e-07
```

```r
anova(reg4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  31.723 2.628e-07 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER: $R^2_{Y4} = \frac{67.775}{67.775+168.782} = 0.2865$. This means that 28.65% of variation in Y can be explained by X_4.

```r
reg1 <-lm(Y~X1)
summary(reg1)
```

```
##
## Call:
## lm(formula = Y ~ X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1759 -0.9545  0.1705  0.9157  4.4444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.64918    0.28978  54.003   <2e-16 ***
```

```
## X1          -0.06489    0.02824  -2.298   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.675 on 79 degrees of freedom
## Multiple R-squared:  0.06264,    Adjusted R-squared:  0.05078
## F-statistic: 5.279 on 1 and 79 DF,  p-value: 0.02422
```

```r
anova(reg1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## X1         1  14.819 14.8185  5.2795 0.02422 *
## Residuals 79 221.739  2.8068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER: $R^2_{Y1} = \frac{14.819}{14.819+221.739} = 0.0626$. This means that 6.26% of variation in Y can be explained by X_1.

$R^2_{Y1|4} = \frac{SSE(X_4)-SSE(X_1,X_4)}{SSE(X_4)} = \frac{SSR(X_1|X_4)}{SSE(X_4)}$

```r
rsq.partial(reg41,reg4)
```

```
## $adjustment
## [1] FALSE
##
## $variables.full
## [1] "X4" "X1"
##
## $variables.reduced
## [1] "X4"
##
## $partial.rsq
## [1] 0.2504679
```

ANSWER: $R^2_{Y1|4} = 0.2505$. This measure is the coefficient of partial determination between Y and X_1, given that X_4 is in the model. Thus, this measures the proportionate reduction in the variation in Y remaining after X_4 is included in the model that is gained by also including X_1 in the model, which is 25.05%.

```r
summary(reg41)
```

```
##
## Call:
## lm(formula = Y ~ X4 + X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2032 -0.4593  0.0641  0.7730  2.5083
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.436e+01  2.771e-01  51.831  < 2e-16 ***
## X4           1.045e-05  1.363e-06   7.663 4.23e-11 ***
## X1          -1.145e-01  2.242e-02  -5.105 2.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.274 on 78 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.4515
## F-statistic: 33.93 on 2 and 78 DF,  p-value: 2.506e-11
```

```
anova(reg41)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  41.788 8.076e-09 ***
## X1         1  42.275  42.275  26.065 2.275e-06 ***
## Residuals 78 126.508   1.622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANSWER: $R^2_{14} = 0.4652$ This measures the variation of Y that can be determined by X_1 and X_4 together, which is 46.52%. To calculate it manually, this is equivalent to $(67.775+42.275)/(67.775+42.275+126.508) = 0.4652$

$R^2_{Y2|14} = \frac{SSE(X_1,X_4)-SSE(X_1,X_4,X_2)}{SSE(X_1,X_4)} = \frac{SSR(X_2|X_1,X_4)}{SSE(X_1,X_4)}$

```
reg14 <-lm(Y~X1+X4)
rsq.partial(reg142,reg14)
```

```
## $adjustment
## [1] FALSE
##
## $variables.full
## [1] "X1" "X4" "X2"
##
## $variables.reduced
## [1] "X1" "X4"
##
## $partial.rsq
## [1] 0.2202037
```

ANSWER: $R^2_{Y2|14} = 0.2202$ This measures is the coefficient of partial determination between Y and X_2, given that X_1 and X_4 is in the model. Thus, this measures the proportionate reduction in the variation in Y remaining after X_1 and X_4 is included in the model that is gained by also including X_2 in the model, which is 22.02%.

$R^2_{Y3|124} = \frac{SSE(X_1,X_2,X_4)-SSE(X_1,X_2,X_4,X_3))}{SSE(X_1,X_2,X_4)} = \frac{SSR(X3|X_1,X_2,X_4)}{SSE(X_1,X_2,X_4)}$

```r
reg124 <-lm(Y~X1+X2+X4)
rsq.partial(reg1243,reg124)
```

```
## $adjustment
## [1] FALSE
##
## $variables.full
## [1] "X1" "X2" "X4" "X3"
##
## $variables.reduced
## [1] "X1" "X2" "X4"
##
## $partial.rsq
## [1] 0.004254889
```

ANSWER: $R^2_{Y3|124} = 0.0043$. This measure is the coefficient of partial determination between Y and X_3, given that X_1, X_2 and X_4 is in the model. Thus, this measures the proportionate reduction in the variation in Y remaining after X_1, X_2 and X_4 is included in the model that is gained by also including X_3 in the model, which is only 0.425%.

```r
summary(reg1234)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4)
##
## 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 ***
## 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
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## 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
```

$R^2 = 0.5847$ This measures the variation of Y that can be explained by X_1, X_2, X_3 and X_4 together, which is 58.47%.

How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_4?

ANSWER: The degree of marginal linear association between Y and X_1, when adjusted for X_4 is $R^2_{Y1|4} =$ 25.05%.