HUDM 5126 Linear Models and Regression Analysis Homework 7

Yifei Dong

10/14/2020

0. Data Preparation

```
setwd("/Users/yifei/Documents/Teachers College/Linear Models and Regression/Week 7/hw7")
getwd()
```

```
## [1] "/Users/yifei/Documents/Teachers College/Linear Models and Regression/Week 7/hw7"
```

```
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(ggplot2)
library(latex2exp)
```

1. Grade Point Average (Q 8.16)

Refer to **Grade point average** Problem 1.19. An assistant to the director of admissions conjectured that the predictive power of the model could be improved by adding information on whether the student had chosen a major field of concentration at the time the application was submitted. Assume that regression model (8.33) is appropriate, where X_1 is entrance test score and $X_2 = 1$ if student had indicated a major filed of concentration at the time of application and 0 if the major filed was undecided.

```
# Read Grade Point Average data first
data1 <- read.table(paste("http://users.stat.ufl.edu/",
                          "~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData",
                         "/Chapter%20%201%20Data%20Sets/CH01PR19.txt", sep=""),
                    header = FALSE)
data1 <- data1 %>%
  select("Y"=V1, "X1"=V2)
head(data1, 10) # check first 10 observations
```

```
##        Y X1
## 1  3.897 21
## 2  3.885 14
```

```
## 3  3.778 28
## 4  2.540 22
## 5  3.028 21
## 6  3.865 31
## 7  2.962 32
## 8  3.961 27
## 9  0.500 29
## 10 3.178 26
```

```r
# Read X2 dataset
data2 <- read.table(paste("http://users.stat.ufl.edu/",
                          "~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData",
                          "/Chapter%20%208%20Data%20Sets/CH08PR16.txt", sep=""),
                    header = FALSE)
data2 <- data2 %>%
  select("X2"=V1)
head(data2, 10)
```

```
##    X2
## 1   0
## 2   1
## 3   0
## 4   1
## 5   0
## 6   1
## 7   1
## 8   1
## 9   1
## 10  0
```

```r
# Combine two datasets
mydata <- cbind(data1, data2)
head(mydata, 10)
```

```
##        Y X1 X2
## 1  3.897 21  0
## 2  3.885 14  1
## 3  3.778 28  0
## 4  2.540 22  1
## 5  3.028 21  0
## 6  3.865 31  1
## 7  2.962 32  1
## 8  3.961 27  1
## 9  0.500 29  1
## 10 3.178 26  0
```

a. Explain how each regression coefficient in model (8.33) is interpreted here.

Recall the model 8.33:

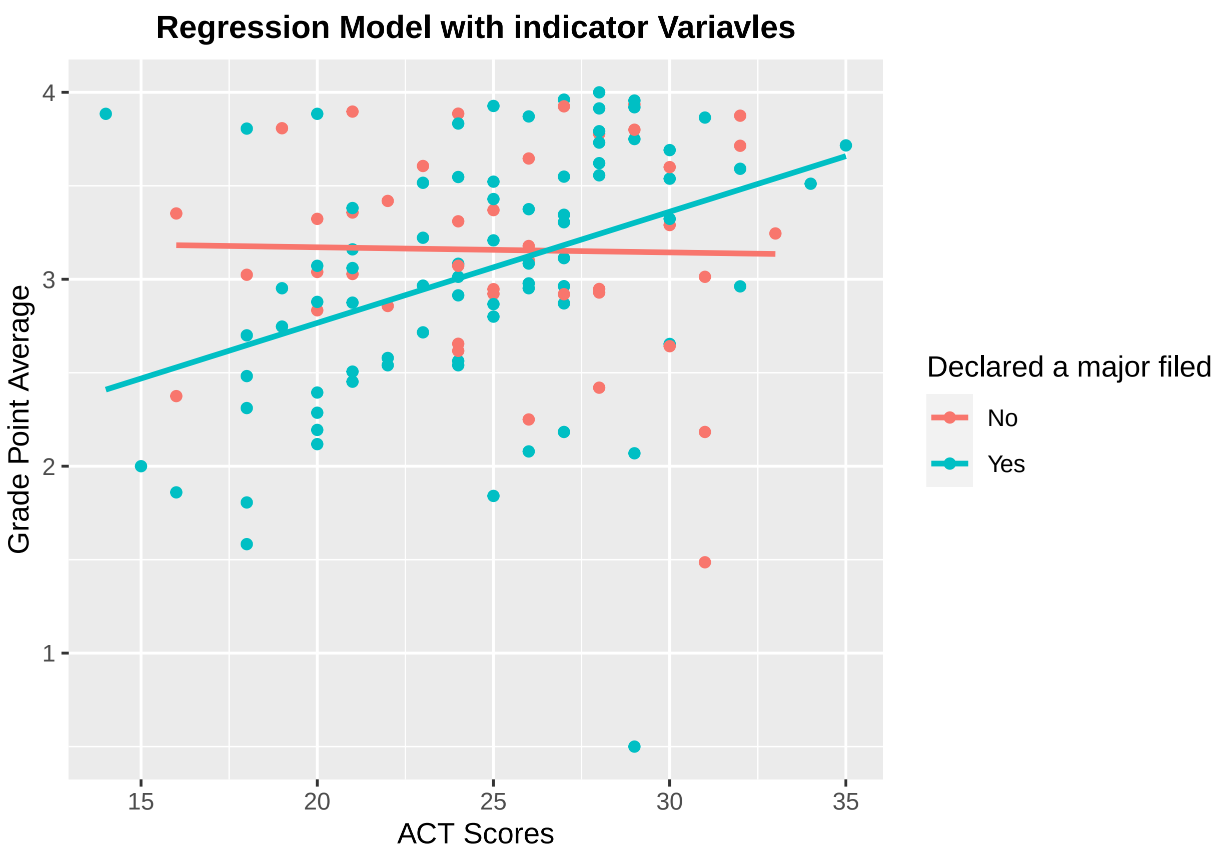
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{1}$$

where X_{1i} stands for the ACT test score, and X_{2i} stands for whether student had chosen a major filed of concentration at the time the application was submitted.

If model 8.33 is applied, β_0 stands for the Y intercept for student who have not chosen the major field. We see that student's GPA is a linear function of ACT scores (X_1), with the same slope β_1 for both types of students. β_2 indicates how much higher(lower) the response function for student who had chosen a major field of concentration at the time the application was submitted than student who had not chosen a major field of concentration, for any given ACT scores. Thus, β_2 measures the differential effect of type of student. The Y intercept for students who had chosen a major field is $\beta_0 + \beta_2$. In general, β_2 shows how much higher(lower) the mean response line is for the student coded 1 than the line for the student coded 0, for any given level of X_1.

```r
attach(mydata)
# Let's visualize the model
g1 <- ggplot(mydata, aes(x = X1, y = Y, color = factor(X2)))+
  geom_point()+geom_smooth(method=lm, se=FALSE)+
  scale_x_continuous("ACT Scores")+scale_y_continuous("Grade Point Average")+
  ggtitle("Regression Model with indicator Variavles")+
  theme(plot.title = element_text(color = "black", size = 12, face = "bold", hjust = 0.5))+
  scale_colour_discrete(name="Declared a major filed", labels=c("No", "Yes"))
g1
```

```
## `geom_smooth()` using formula 'y ~ x'
```



b. Fit the regression model and state the estimated regression function.

```r
# Fit the regression model
reg1 <- lm(Y~X1+X2)
summary(reg1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70304 -0.35574  0.02541  0.45747  1.25037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19842    0.33886   6.488 2.18e-09 ***
## X1           0.03789    0.01285   2.949  0.00385 **
## X2          -0.09430    0.11997  -0.786  0.43341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6241 on 117 degrees of freedom
## Multiple R-squared:  0.07749,    Adjusted R-squared:  0.06172
## F-statistic: 4.914 on 2 and 117 DF,  p-value: 0.008928
```

So the estimated regression function is:

$$\widehat{Y_i} = 2.1984 + 0.0379X_{i1} - 0.0943X_{i2}$$

c. Test whether the X_2 variable can be dropped from the regression model; use $\alpha = .01$.State the alternatives, decision rule, and conclusion.

```r
reg2 <- lm(Y~X1)
anova(reg2, reg1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    118 45.818
## 2    117 45.577  1   0.24071 0.6179 0.4334
```

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

And the t-statistic is:

$$t^* = \frac{b_2}{s\{b_2\}}$$

```r
# Check partial F-statistics
qf(.99, df1=1, df2=117)
```

```
## [1] 6.856564
```

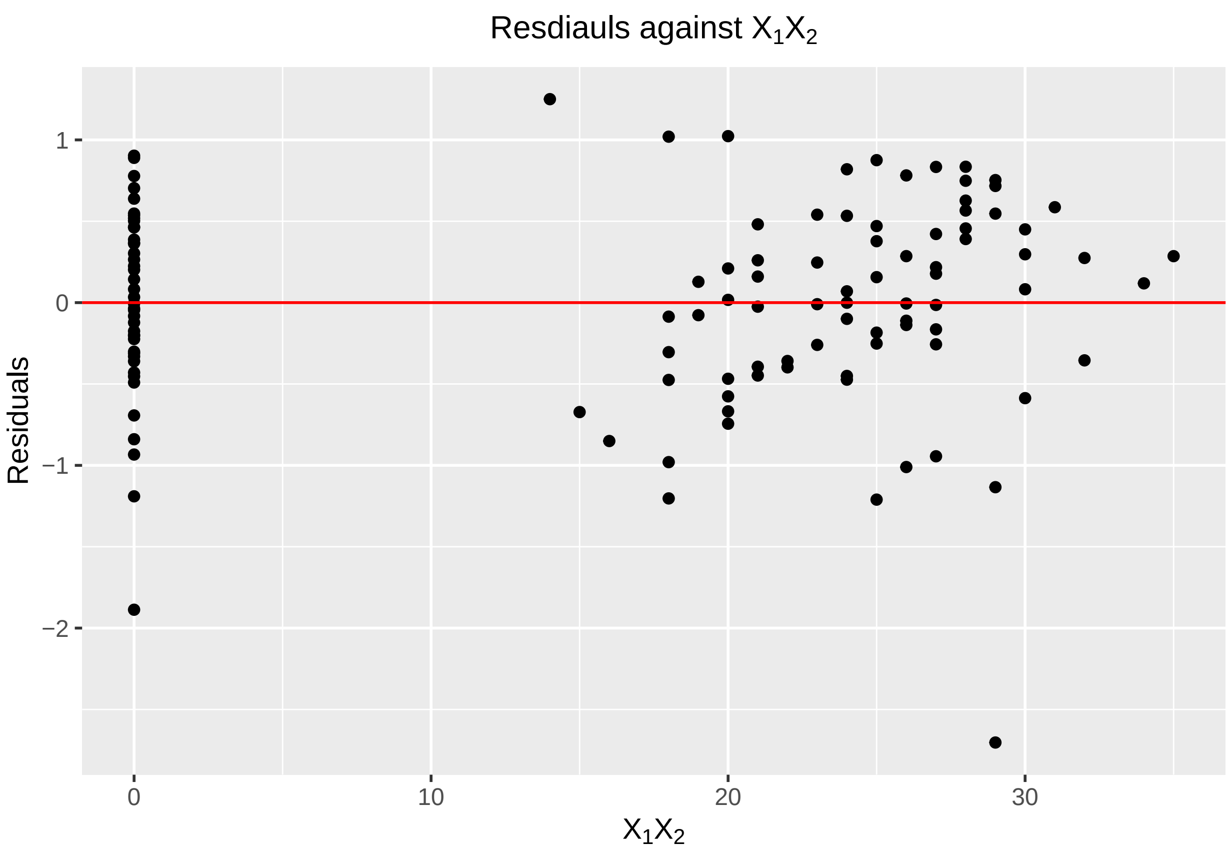
The decision rule is, if $|t^*| \leq t(1 - \alpha/2, n - p)$, conclude H_0, otherwise, conclude H_1. Since the p-value of the t test is $0.43341 > .01$, we cannot reject the null hypothesis. There we conclude that X_2 can be dropped from the regression model. We could use partial F-statistics as well, $F^* = 0.6179 < 6.8566$, therefore, we cannot reject null hypothesis.

d. Obtain the residuals for regression model (8.33) and plot them against X_1X_2. Is there any evidence in your plot that it would be helpful to include an interaction term in the model?

```r
# Obtain residuals and generate X1X2
e <- residuals(reg1)
X1X2 <- X1*X2
mydata <- cbind(mydata, e, X1X2)
head(mydata, 10)
```

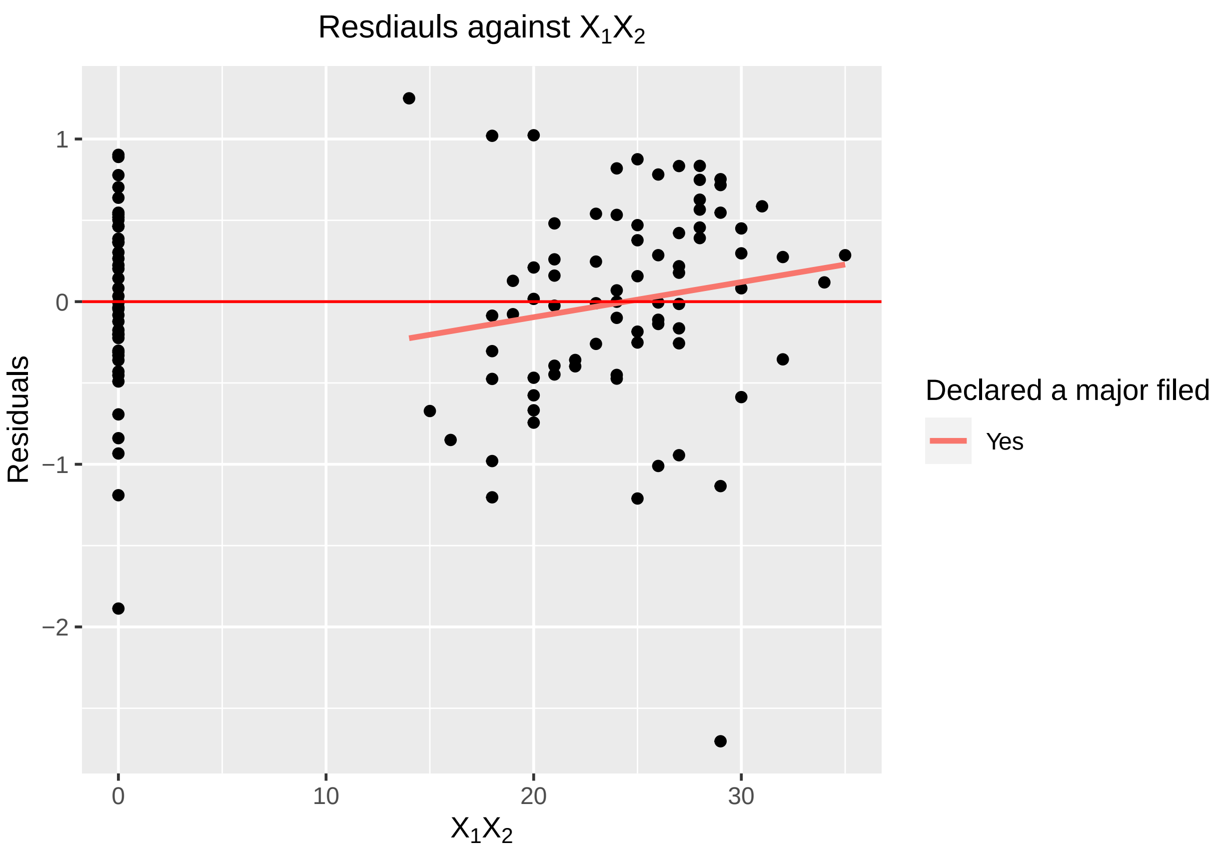
```
##        Y X1 X2            e X1X2
## 1  3.897 21  0  0.902807465    0
## 2  3.885 14  1  1.250369139   14
## 3  3.778 28  0  0.518549716    0
## 4  2.540 22  1 -0.397782574   22
## 5  3.028 21  0  0.033807465    0
## 6  3.865 31  1  0.586171749   31
## 7  2.962 32  1 -0.354722215   32
## 8  3.961 27  1  0.833747605   27
## 9  0.500 29  1 -2.703040323   29
## 10 3.178 26  0 -0.005662355    0
```

```r
# Residuals against X1X2
g2 <- ggplot(mydata, aes(x = X1X2, y = e))+
  geom_point(color="black")+xlab(TeX("X_{1}X_{2}"))+
  ylab("Residuals")+ggtitle(TeX("Resdiauls against X_{1}X_{2}"))+
  theme(plot.title = element_text(color = "black", face = "bold", size = 12, hjust = 0.5))+
  geom_hline(yintercept = 0, color="red", linetype=1)
g2
```



According to the residuals against X_1X_2 plot, I find there is a systematic pattern when $X_2 = 1$. There is a positive relationship between X_1X_2 and residuals when $X_2 = 1$ (See graph below, which uses the simple linear model to fit the relationship between X_1X_2 and residuals when $X_2 = 1$, other methods may even better).

```
g3 <- ggplot(mydata, aes(x = X1X2, y = e, color=factor(X2)))+
  geom_point(color="black")+xlab(TeX("X_{1}X_{2}"))+
  ylab("Residuals")+ggtitle(TeX("Resdiauls against X_{1}X_{2}"))+
  theme(plot.title = element_text(color = "black", face = "bold", size = 12, hjust = 0.5))+
  stat_smooth(method = "lm", formula=y~x, se=FALSE)+
  scale_color_discrete(name="Declared a major filed", labels=c("Yes", "No"))+
  geom_hline(yintercept = 0, color="red", linetype=1)
g3
```



Continued: Grade Point Average (Q 8.20)

a. Fit regression model (8.49) and state the estimated regression model.

Recall the model 8.49:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \tag{2}$$

```r
reg3 <- lm(Y~X1*X2)
summary(reg3)
```

```
##
## Call:
## lm(formula = Y ~ X1 * X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80187 -0.31392  0.04451  0.44337  1.47544
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.226318   0.549428   5.872 4.18e-08 ***
## X1          -0.002757   0.021405  -0.129   0.8977
## X2          -1.649577   0.672197  -2.454   0.0156 *
## X1:X2        0.062245   0.026487   2.350   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6124 on 116 degrees of freedom
## Multiple R-squared:  0.1194, Adjusted R-squared:  0.09664
## F-statistic: 5.244 on 3 and 116 DF,  p-value: 0.001982
```

So the estimated regression function is:

$$\widehat{Y}_i = 3.2263 - 0.0028X_{i1} - 1.6496X_{i2} + 0.0622X_{i1}X_{i2}$$

b. Test whether the interaction term can be dropped from the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. If the interaction term cannot be dropped from the model, describe the nature of the interaction effect.

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

And the t-statistic is:

$$t^* = \frac{b_3}{s\{b_3\}}$$

```r
anova(reg1, reg3)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 * X2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    117 45.577
## 2    116 43.506  1    2.0713 5.5226 0.02046 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The decision rule is, if $|t^*| \leq t(1-\alpha/2, n-p)$, conclude H_0, otherwise, conclude H_1. Since the p-value of t-test is 0.0205, which is smaller than .05. Therefore, we reject the null hypothesis H_0 and conclude that the interaction term cannot be dropped from the model.

Alternatively, we can use partial F test.

$$H_0 : \beta_q = \beta_{q+1} = ... = \beta_{p-1} = 0$$

$$H_1 : \text{not all of the } \beta_k \text{ in } H_0 \text{ equal } 0$$

In this case, we only test β_3, therefore, $q = 3$. The test is a single regression coefficient equal zero.

$SSR(X_1X_2|X_1, X_2) = 2.0713$, $SSE(X1, X2, X_1X_2) = 43.506$. The equation for test statistic is:

$$F^* = \frac{SSR(X_1X_2|X_1, X_2)}{p-q} \div \frac{SSE(X1, X2, X_1X_2)}{n-p} \tag{3}$$

The decision rule is, if $F^* \leq F(p-q, n-p)$, conclude H_0, otherwise, conclude H_1.

```r
# check F-statistics
qf(.95, df1=1, df2=116)
```

```
## [1] 3.922879
```

therefore, $F^* = \frac{2.0713}{4-3} \div \frac{43.506}{120-4} = \frac{2.0713*116}{43.506} = 5.5227$, which matches the result we calculated from using anova built in command. Since F-Statistics is larger than 3.9229, we conclude H_1.

The Y intercept is $\beta_0 = 3.2263$ and the slope is $\beta_1 = -.0028$ for the response function for student who haven't chosen major field. The response function for student who had chosen a major field of concentration has Y intercept $\beta_0 + \beta_2 = 3.2263 - 1.6496 = 1.5767$, and slope $\beta_1 + \beta_3 = -0.00276 + 0.06225 = 0.0595$. We see that β_2 here indicates how much smaller is the Y intercept of the response function for student coded 1 than that for the student coded 0. Similarly, β_3 indicates how much greater is the slope of the response function for the student coded 1 than student coded 0 (the difference in slope between two groups).