MATH 4780 (MSSC 5780) Homework 3

Due Date: October 7, 2022 11:59 PM Henrique Medeiros Dos Reis

1 Homework Instruction and Requirement

- Homework 3 covers course materials of Week 1 to 6.
- Please submit your work in **one PDF** file including all parts (Section 2 and 3) to **D2L > Assessments > Dropbox**. *Multiple files or a file that is not in pdf format are not allowed.*
- In your homework, please number and answer questions **in order**.
- Your answers may be handwritten on the Mathematical Derivation and Reasoning part. However, you need to scan your paper and make it a PDF file.
- Your entire work on Statistical Computing and Data Analysis should be completed by any word processing software (Microsoft Word, Google Docs, (R)Markdown, LaTex, etc) and your preferred programming language. Your document should be a PDF file.
- Questions starting with **(MSSC)** are for MSSC 5780 students. MATH 4780 students could possibly earn extra points from them.
- It is your responsibility to let me understand what you try to show. If you type your answers, make sure there are no typos. I grade your work based on *what you show, not what you want to show.* If you choose to handwrite your answers, write them neatly. If I can't read your sloppy handwriting, your answer is judged as wrong.

2 Mathematical Derivation and Reasoning

The simple linear regression (SLR) and multiple linear regression (MLR) models and notations are the same as defined in our course slides and textbook.

In simple linear regression,

1. Show that $r^2 = \frac{SS_R}{SS_T} = R^2$, that is, the square of the sample correlation coefficient between y and x is equal to the coefficient of determination.

   Solution:

   $\sum_{i=1}^{n}(y_i - \overline{y})^2 = SS_T$, $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = SS_R$, And

$$r^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} \right)^2$$

Then:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

Since

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

We can write:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{y})^2 = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \Rightarrow$$

$$R^2 = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^n (y_i - \overline{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2}$$

Therefore:

$$R^2 = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^n (y_i - \overline{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2} = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2}} \right)^2 = r^2$$

2. Show that $\text{Cov}(b_0, b_1) = \frac{-\bar{x}\sigma^2}{S_{xx}}$. [Hint: Use the fact that $\text{Cov}(aX + bY, W) = a\text{Cov}(X, W) + b\text{Cov}(Y, W)$ and $\text{Cov}(\bar{y}, b_1) = 0$]. This tells us the least-squares intercept and slope are negatively correlated when the average of x is positive.

Solution:

Since $b_0 = \bar{y} - b_1\bar{x}$, then:

$$\text{Cov}(\bar{y} - b_1\bar{x}, b_1) = \text{Cov}(\bar{y}, b_1) - \bar{x}\text{Cov}(b_1, b_1)$$

Using the hint $\text{Cov}(\bar{y}, b_1) = 0$, we can simplify:

$$\text{Cov}(b_0, b_1) = -\bar{x}\text{Cov}(b_1, b_1) \Rightarrow$$

Then since $\text{Cov}(X, X) = E(x^2) - E(x)^2 = Var(x)$ and $b_1 = \frac{S_{xy}}{S_{xx}}$

$$-\bar{x}Var(b_1) = -\bar{x}Var(\frac{S_{xy}}{S_{xx}}) = -\bar{x}\frac{1}{S_{xx}^2}\sum_{i=1}^n (x_i - \bar{x})Var(y_i) \Rightarrow$$

Finally, using the facts $\frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{S_{xx}}$ and $Var(y_i) = \sigma^2$

$$\frac{-\bar{x}\sigma^2}{S_{xx}} = \text{Cov}(b_0, b_1)$$

3. **(MSSC)** Show that for the test $H_0 : \beta_1 = 0$; $H_1 : \beta_1 \neq 0$, the square of t test statistic is equal to the F test statistic, i.e., $t_{test}^2 = F_{test}$.

Solution:

$$F_{test} = \frac{MS_{reg}}{MS_{res}};\ t_{test} = \frac{b_1}{se(b_1)} = \frac{b_1}{\sqrt{\frac{MS_{res}}{S_{xx}}}}$$

$$t^2 = \left(\frac{b_1}{\sqrt{\frac{MS_{res}}{S_{xx}}}}\right)^2 = \frac{b_1^2}{\frac{MS_{res}}{S_{xx}}} = \frac{b_1^2 S_{xx}}{MS_{res}} \Rightarrow$$

$$\frac{\sum_{i=1}^n (b_1 x - b_1\bar{x})^2}{MS_{res}} = \frac{\sum_{i=1}^n (\bar{y} + b_1 x - b_1\bar{x} - \bar{y})^2}{MS_{res}} = \frac{\sum_{i=1}^n (b_0 + b_1 x_i - \bar{y})^2}{MS_{res}} \Rightarrow$$

$$\frac{\sum_{i=1}^n (\hat{y} - \bar{y})^2}{MS_{res}} = \frac{MS_{reg}}{MS_{res}} = F_{test}$$

4. **(Optional)** Show that $\text{Cov}(\bar{y}, b_1) = 0$. This tells us that the sample mean of response is uncorrelated with the least-squares slope.

\ Suppose $\mathbf{A}_{n\times n}$ is a symmetric idempotent matrix.

5. **(MSSC)** Show that the eigenvalues of $\mathbf{A}$ are either zero or one, and therefore $\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A})$.

Solution:

Since A is symetric and idempotent, then $A = A^-1$, and $A = A^2$

Then $Tr(A) = \sum_{i=1}^n \lambda_i$, where all the eigenvalues are either 0 or one, which is the same as the number of nonzero eigenvalues.Since it is a sum of 1s. And by the definition \$rank(A) = \$ the number of nonzero λ_i. Therefore $\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A})$

In multiple linear regression, let the hat matrix $\mathbf{H}_{n\times n} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

6. Show that $\mathbf{H}$ and $(\mathbf{I} - \mathbf{H})$ are symmetric and idempotent.

Solution:

Let's start with H $H = X(X'X)^{-1}X'$ and $H' = (X(X'X)^{-1}X')'$.

$$X(X'X)^{-1}X' = (X(X'X)^{-1}X')' = X'((X'X)^{-1})'(X')' \Rightarrow$$

$$X'((X'X)')^{-1}X = X(X'X)^{-1}X'$$

$H = X(X'X)^{-1}X'$ and $H^2 = (X(X'X)^{-1}X')(X(X'X)^{-1}X')$

$$X(X'X)^{-1}X'X(X'X)^{-1}X',\ \text{since } X'X(X'X)^{-1} = I \Rightarrow$$

$$X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X'$$

Then $I - H$, since we just proved H is symetric and idenpotent:

$$I - H = (I - H)' = I' - H' = I - H$$

And:

$$(I - H)^2 = I^2 - 2IH + H^2 = I - 2H + H = I - H$$

7. Show that $\mathrm{tr}(\mathbf{H}) = p$.

Solution:

$$Tr(H) = Tr(X(X'X)^{-1}X')$$

Since $Tr(AB) = Tr(BA)$, and H is $p \times p$

$$Tr((X'X)^{-1}X'X) = Tr(I) = \sum_{i=1}^{p} I_{ii} = p$$

3 Statistical Computing and Data Analysis

Please perform a data analysis using `R` or your preferred language. **Any results should be generated by computer outputs, and your work should be done entirely by a computer. Handwriting is not allowed. Relevant code should be attached.**

We use the same data set `mpg.csv` for data analysis.

```
mpgData <- read.csv("mpg.csv")
head(mpgData, 10)
```

```
##        y    x1  x2  x3   x4   x5 x6 x7    x8   x9  x10 x11
## 1  18.90 350.0 165 260 8.00 2.56  4  3 200.3 69.9 3910   1
## 2  17.00 350.0 170 275 8.50 2.56  4  3 199.6 72.9 3860   1
## 3  20.00 250.0 105 185 8.25 2.73  1  3 196.7 72.2 3510   1
## 4  18.25 351.0 143 255 8.00 3.00  2  3 199.9 74.0 3890   1
## 5  20.07 225.0  95 170 8.40 2.76  1  3 194.1 71.8 3365   0
## 6  11.20 440.0 215 330 8.20 2.88  4  3 184.5 69.0 4215   1
## 7  22.12 231.0 110 175 8.00 2.56  2  3 179.3 65.4 3020   1
## 8  21.47 262.0 110 200 8.50 2.56  2  3 179.3 65.4 3180   1
## 9  34.70  89.7  70  81 8.20 3.90  2  4 155.7 64.0 1905   0
## 10 30.40  96.9  75  83 9.00 4.30  2  5 165.2 65.0 2320   0
```

1. Fit a MLR model $y = \beta_0 + \beta_1 x_1 + \beta_6 x_6 + \epsilon$ relating gasoline mileage y (miles per gallon) to engine displacement x_1 and the number of carburetor barrels x_6. Interpret the regression coefficients β_1 and β_6.

Solution:

```
mpg_engDisp_carBar_lm <- lm(y~x1+x6, data = mpgData)
mpg_engDisp_carBar_lm
```

```
##
## Call:
## lm(formula = y ~ x1 + x6, data = mpgData)
##
## Coefficients:
## (Intercept)           x1           x6
##    32.88455     -0.05315      0.95922
```

Our linear model then is $y = 32.88 - 0.05x_1 + 0.96x_6$ which means that: If all else is held constant, and we increase the engine displacement by 1 unit, the mpg is going to decrease, on average, 0.05. And if all else is held constant, and we increase the number of carburetor barrels by one, the mpg is going to increase, on average, 0.96.

2. Write down the H_0 and H_1 for testing significance of regression, and construct the ANOVA table to test the significance. Explain your decision rule and conclusion.

Solution:

We will test $H_0 : \beta_1 = \beta_6 = 0$ against $H_1 : \beta_j \neq 0$ for at least one j. Since we have a multiple linear regression model, we want to test it against a null model, such that we can test the whole model not just one specific variable.

```
null_model <- lm(y ~ 1, data = mpgData)
anova(null_model, mpg_engDisp_carBar_lm)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x1 + x6
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     31 1237.54
## 2     29  263.23  2    974.31 53.669 1.79e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova test, we can see that our $p-value$ is small enough for us to reject H_0 in favor of H_1.

3. Obtain R^2 and R^2_{Adj} for this MLR model. Compare these to the R^2 and R^2_{Adj} for the SLR model relating mileage to engine displacement.

Solution:

```
mpg_disp_lm <- lm(formula = y~x1, data = mpgData)
sum_mlr<-summary(mpg_engDisp_carBar_lm)
sum_lr<-summary(mpg_disp_lm)
cat("R^2 for MLR model: ", sum_mlr$r.squared, ", ",
    "R_{adj}^2 for MLR model: ", sum_mlr$adj.r.squared,
    "\n","R^2 for LR model: ", sum_lr$r.squared, ", ",
    "R_{adj}^2 for LR model:  ", sum_lr$adj.r.squared, "\n")
```

```
## R^2 for MLR model:  0.7872928 ,  R_{adj}^2 for MLR model:  0.7726233
##  R^2 for LR model:  0.7722712 ,  R_{adj}^2 for LR model:   0.7646803
```

Obviously, R^2 for the MLR model is higher, but not by a lot. And for R^2_{adj} the difference is even smaller.

4. Find a 95% confidence interval (CI) for β_1. Interpret your results.

   Solution:

```
(ci<- confint(mpg_engDisp_carBar_lm))
```

```
##                   2.5 %      97.5 %
## (Intercept) 29.74428901 36.02481266
## x1          -0.06569892 -0.04059641
## x6          -0.41164739  2.33009349
```

The 95 confidence interval for β_1 is $(-0.0656, -0.0406)$. Which means that there is a 95 chance that the true value for β_1 is inside that interval.

5. With $\alpha = 0.05$, do the marginal test $H_0 : \beta_6 = 0$. Interpret your results.

   Solution:

   In order to test for $H_0 : \beta_6 = 0$, we need to look at the p-value of the coefficient for β_6

```
sum_mlr$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 32.88455083 1.535407938 21.417468 2.546135e-19
## x1          -0.05314767 0.006136843 -8.660425 1.549965e-09
## x6           0.95922305 0.670277025  1.431084 1.630948e-01
```

We can see that the p-value is 0.1631, which means that we fail to reject the H_0 with $\alpha = 0.05$, since $0.1631 \geq 0.05$.

6. Find a 95% CI on the mean gasoline mileage when $x_1 = 275$ in^3 and $x_6 = 2$ barrels.

   Solution:

```
predict(mpg_engDisp_carBar_lm,
        newdata = data.frame(x1 = 275, x6 = 2),
        interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 20.18739 18.87221 21.50257
```

The 95% CI on the mean gasoline mileage when $x_1 = 275$ in^3 and $x_6 = 2$ barrels is $(18.872, 21.503)$

7. Find a 95% prediction interval (PI) for a new observation on gasoline mileage when $x_1 = 275$ in^3 and $x_6 = 2$ barrels.

Solution:

```
predict(mpg_engDisp_carBar_lm,
        newdata = data.frame(x1 = 275, x6 = 2),
        interval = "predict", level = 0.95)
```

```
##        fit     lwr      upr
## 1 20.18739 13.8867 26.48808
```

The 95% PI on the mean gasoline mileage when $x_1 = 275$ in^3 and $x_6 = 2$ barrels is $(13.887, 26.488)$

8. In Homework 2 you were asked to compute 95% CI on mean gasoline mileage and PI on a car's gasoline mileage when the engine displacement $x_1 = 275$ in^3. Compare the length of these intervals to the length of the CI and PI from the question 6 and 7 above. Does adding x_6 to the model help in terms of prediction or uncertainty reduction?

Solution:

From Homework 2, we got CI to be (19.58807, 21.80952) and PI to be (14.34147 27.05611). Which when comparing to the results from questions 6 and 7, we can see that the CI interval from question 6 is just a little smaller than the CI from Homework 2, and the PI intervals are about the same size. Therefore adding x_6 does not help much the model in terms of prediction, and helps a little bit in uncertainty reduction.

9. Perform matrix operations to compute $(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$ and verify it is $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Solution:

```
X <- as.matrix(cbind(1, mpgData[, c(2, 7)]))
y <- as.matrix(mpgData$y)
b <- solve(t(X) %*% X) %*% t(X) %*% y

t(y-X%*%b)%*%(y-X%*%b)
```

```
##          [,1]
## [1,] 263.2345
```

```
sum((mpgData$y - mpg_engDisp_carBar_lm$fitted.values)^2)
```

```
## [1] 263.2345
```

We can see that they are the same.

10. Generate the predictor effect plot for x_1 and x_6. Explain the plot by discussing the effect of each predictor on y.

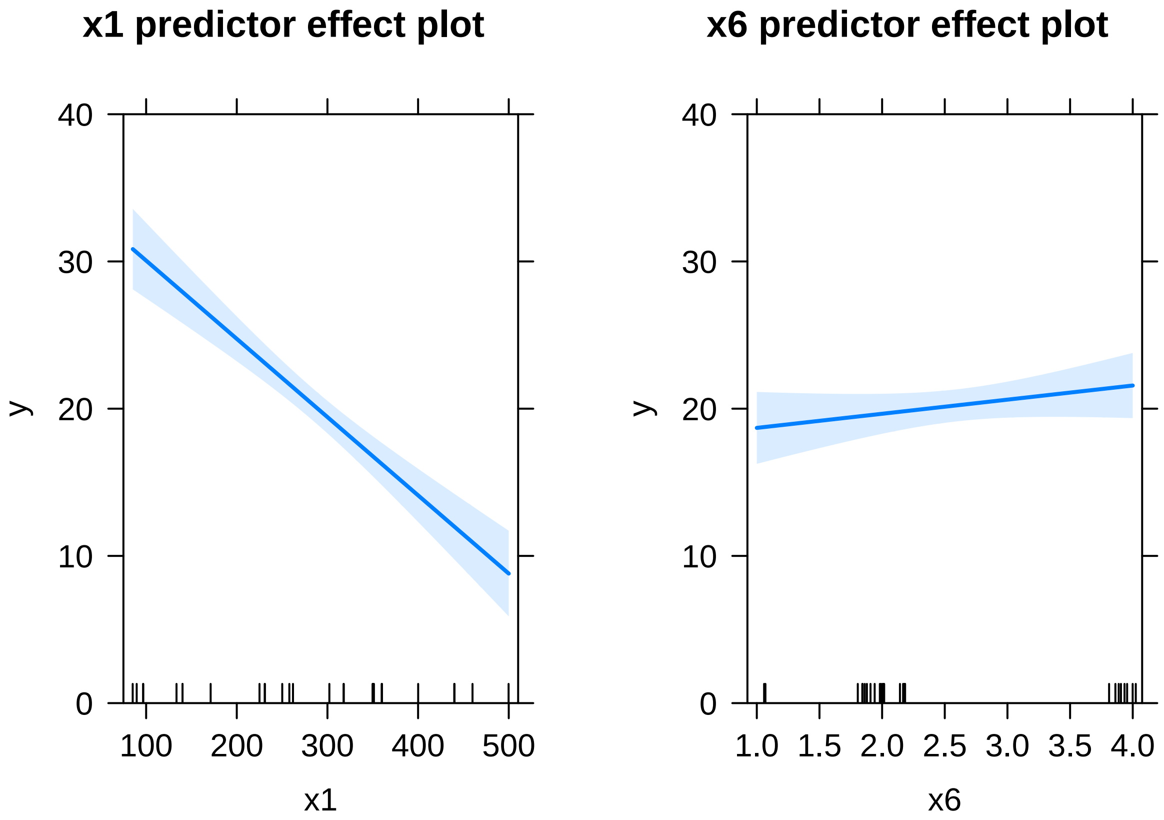
Solution:

```
library(effects)
```

```
## Loading required package: carData
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
plot(effects::predictorEffects(mod = mpg_engDisp_carBar_lm), axes=list(y=list(lim=c(0,40))))
```



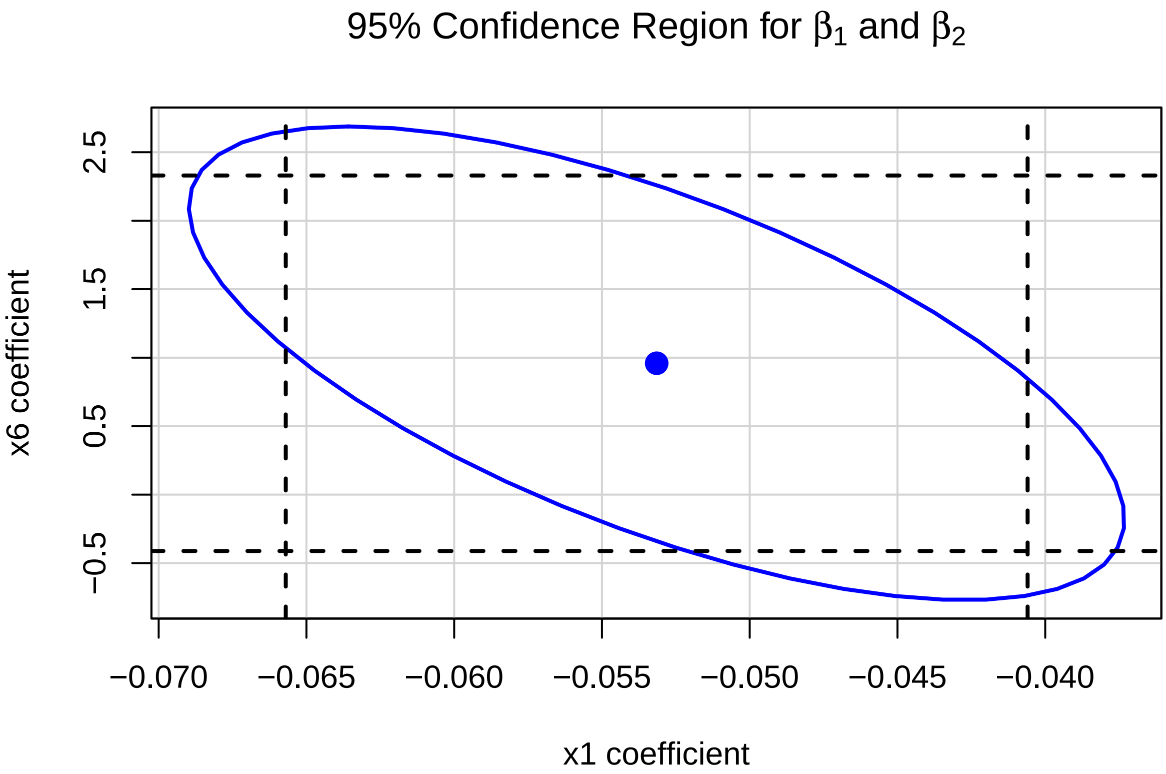
We can see that in the graph for X_1 we have a negative slope that is not close to 0. On the other hand x_6 the slope is very close to 0, so the prediction “power” of x_6 is really small.

11. Construct the 95% confidence region for the coefficients (β_1, β_6). Interpret the region.

Solution:

```
car::confidenceEllipse(
    mpg_engDisp_carBar_lm,
    levels = 0.95,
    which.coef = c("x1", "x6"),
    main = expression(
        paste("95% Confidence Region for ",
              beta[1], " and ",  beta[2])
        )
    )

abline(v = ci[2, ], lty = 2, lwd = 2)
abline(h = ci[3, ], lty = 2, lwd = 2)
```



So there is a region that is not included inside the blue elliptically-shaped region, the joint of both x_1 and x_6. Which happens because of the negative correlation between those two variables. And also, we may have values that are not included in the dashed lines, and included in the elliptically-shaped region, that are excluded from the marginal interval.