MATH 4780 (MSSC 5780) Homework 2

September 21 2022

Due Date: September 23, 2022 11:59 PM Henrique Medeiros Dos Reis

1 Homework Instruction and Requirement

- Homework 2 covers course materials of Week 1 to 4.
- Please submit your work in **one PDF** file including all parts 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, Quarto, 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

2.1 Simple Linear Regression

The following questions are based on the population and sample linear regression model defined in our course slides and textbook.

1. Find the least squares estimator b_0 and b_1 such that

$$(b_0, b_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Solution: The LSE of β_0 and β_1 that minimizes the error, can obtained by taking derivatives and setting them equal to 0.

$$\frac{\partial SS_{res}}{\partial \beta_0} = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\sum_{i=1}^{n}(y_i) - nb_0 - b_1\sum_{i=1}^{n}x_i = 0$$

Now, lets solve for b_0 given b_1

$$\sum_{i=1}^{n}(y_i) - nb_0 - b_1\sum_{i=1}^{n}x_i = 0 \Rightarrow -nb_0 = b_1\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}(y_i) =$$

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

Now, let's take the derivative with respect to β_1

$$\frac{\partial SS_{res}}{\partial \beta_1} = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$-2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i - b_0\sum_{i=1}^{n} x_i - b_1\sum_{i=1}^{n} x_i^2$$

Now, lets solve for b_1 given b_0

$$\sum_{i=1}^{n} x_i y_i = b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 \Rightarrow n\sum_{i=1}^{n} x_i y_i = nb_0\sum_{i=1}^{n} x_i + nb_1\sum_{i=1}^{n} x_i^2 = 0$$

Putting together the previous two equations we then have:

$$b_1(n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i^2) = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \Rightarrow$$

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

Therefore the estimators b_0 and b_1 are:

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

2. Show that $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ where $h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$.

Solution:

$$\hat{y}_i = b_0 + b_1 x_i \Rightarrow \hat{y}_i = \bar{y} - b_1\bar{x} + \frac{\sum_{j=1}^{n}(x_i - \bar{x})\sum_{j=1}^{n}(y_i - \bar{y})}{\sum_{j=1}^{n}(x_i - \bar{x})^2}$$

Since $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$.

$$\hat{y}_i = \bar{y} - b_1\bar{x} + \frac{\sum_{j=1}^{n}(x_i - \bar{x})y_i}{S_{xx}}$$

And $c_i = \frac{x_i - \bar{x}}{S_{xx}}$.

$$\hat{y}_i = \bar{y} - b_1\bar{x} + x_i \sum_{j=1}^{n} c_j y_j \Rightarrow$$

$$\hat{y}_i = \bar{y} - \bar{x}\sum_{j=1}^{n} c_j y_j + x_i \sum_{j=1}^{n} c_j y_j = \frac{1}{n}\sum_{j=1}^{n} y_j - \bar{x}\sum_{j=1}^{n} c_j y_j + x_i \sum_{j=1}^{n} c_j y_j$$

$$\hat{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_j - \bar{x}\sum_{j=1}^{n} c_j y_j + x_i \sum_{j=1}^{n} c_j y_j = \sum_{j=1} y_j(\frac{1}{n} - \bar{x}c_j + x_i c_j) \Rightarrow$$

$$\hat{y}_i = \sum_{j=1} y_j(\frac{1}{n} - \bar{x}\frac{x_j - \bar{x}}{S_{xx}} + x_i\frac{x_j - \bar{x}}{S_{xx}}) = \sum_{j=1} y_j(\frac{1}{n} + \frac{-\bar{x}x_j + \bar{x}^2 + x_i x_j - x_i\bar{x}}{S_{xx}})$$

$$\sum_{j=1} y_j(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}})$$

Therefore $\hat{y}_i = \sum_{j=1}^{n} y_j h_{ij}$

3. Remember that before training sample is collected, y_i are assumed random variables. Show that b_0 is an unbiased estimator for β_0, i.e., $E(b_0) = \beta_0$.

Solution:

$$\beta_0 = E[b_0] = E[\bar{y} - b_1\bar{x}] = E[\bar{y}] - E[b_1\bar{x}] \Rightarrow$$

$$\beta_0 = E[\sum_{i=1}^{n}\frac{y_i}{n}] - E[b_1\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} E[y_i] - \bar{x}E[b_1]$$

Since $E[b_1] = \beta_1$ and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Then:

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} E[\beta_0 + \beta_1 x_i + \epsilon_i] - \bar{x}\beta_1 = \frac{1}{n}\sum_{i=1}^{n}(E[\beta_0] + E[\beta_1 x_i] + E[\epsilon_i]) - \bar{x}\beta_1 \Rightarrow$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 = \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \frac{1}{n}\sum_{i=1}^{n}\beta_1 x_i - \bar{x}\beta_1 \Rightarrow$$

$$\beta_0 = \frac{1}{n}n\beta_0 + \frac{\sum_{i=1}^{n}\beta_1 x_i}{n} - \bar{x}\beta_1 = \beta_0 + \beta_1\frac{\sum_{i=1}^{n} x_i}{n} - \bar{x}\beta_1 = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x}$$

Therefore $E[b_0] = \beta_0$

4. Show that $\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, i.e., $SS_T = SS_R + SS_{res}$.

Solution:

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

Now let $a = (y_i - \hat{y}_i)$ and $b = (\hat{y}_i - \overline{y})$ then $(a+b)^2 = a^2 + 2ab + b^2 \Rightarrow (y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + (\hat{y}_i - \overline{y})^2$. Then:

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

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

For this expression to be true, we need to prove that $2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$

$$2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0 = \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \Rightarrow$$

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))(\beta_0 + \beta_1 x_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \hat{y} + \beta_1\bar{x} - \beta_1 x_i)(\hat{y} - \beta_1\bar{x} + \beta_1 x_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y} - \beta_1(x_i - \bar{x}))(\beta_1(x_i - \bar{x}))$$

$$\sum_{i=1}^{n}\beta_1(x_i - \bar{x})(y_i - \bar{y}) - \beta_1^2(x_i - \bar{x}) = \beta_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \beta_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Where:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{S_{xx}}$$

Then:

$$\frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{S_{xx}} - [\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}]^2 S_{xx} = 0 \Rightarrow$$

$$\frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{S_{xx}} - \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{S_{xx}} = 0$$

Therefore

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

5. **(MSSC)** Show that $SS_{res} = SS_T - b_1 S_{xy}$, i.e., $SS_R = b_1 S_{xy}$. This proof tells us that SS_{res} is the variation with all variation explained by the association of x and y removed.
Solution:

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,\ SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2, S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Then we can show:

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

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

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

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

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

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=0=SS_{res}$$

And $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}e_i=0$.

6. **(MSSC)** Show that $E(MS_{res})=\sigma^2$ and $E(MS_R)=\sigma^2+\beta_1^2S_{xx}$. [*Hint:* Use the fact that a χ_k^2 random variable has the mean value k.] This proof tells us that MS_R is also an estimator for σ^2. Although it is usually biased, it is unbiased when $\beta_1=0$. Would $MS_R\geq MS_{res}$ always hold? Why or why not? Please explain.

Solution:

part 1

$$E(MS_{res})=\sigma^2\Rightarrow\sigma^2=\frac{SS_{Res}}{n-2}$$

$$E(MS_{res})=E[\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}]=\frac{1}{n-2}E[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2]\Rightarrow$$

$$\frac{1}{n-2}E[SS_{res}]=\frac{1}{n-2}((n-2)\sigma^2)=\sigma^2$$

part 2

$$E(MS_R)=E[\frac{SS_r}{1}]=E[b_1^2S_{xx}]=S_{xx}E[b_1^2]\Rightarrow$$

$$S_{xx}[Var(b_1)+E[b_1]^2]=S_{xx}[\frac{\sigma^2}{S_{xx}}+E[b_1]^2]=S_{xx}[\frac{\sigma^2}{S_{xx}}+\beta_1^2]\Rightarrow$$

$$\sigma^2+\beta_1^2S_{xx}=E[MS_R]$$

The property $MS_R\geq MS_{res}$ will always hold since $\sigma^2\geq\sigma^2+\beta_1^2S_{xx}$, where the smallest value we can have for β_1^2 is 0, which in that case $\sigma^2=\sigma^2+0^2S_{xx}$. Otherwise the value of $\sigma^2+\beta_1^2S_{xx}$ will always be greater than just σ^2

7. **(Optional)** Obtain the maximum likelihood estimator for β_0, β_1 and σ^2.

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 your computer**. **Handwriting is not allowed**. **Attach your code at the end of your homework PDF file**.

The data set `mpg.csv` presents data on the gasoline mileage performance of 32 different automobiles. (Table B.3 in the textbook)

To import the data set into your R session, use `read.csv()` like

```
data_name_you_like <- read.csv("mpg.csv")
```

Once you load the data set, type its name on the R console. The data should be a data frame with 32 rows and 12 columns that looks like

```
##       y  x1  x2  x3   x4   x5 x6 x7    x8   x9  x10 x11
## 1 18.90 350 165 260 8.00 2.56  4  3 200.3 69.9 3910   1
## 2 17.00 350 170 275 8.50 2.56  4  3 199.6 72.9 3860   1
## 3 20.00 250 105 185 8.25 2.73  1  3 196.7 72.2 3510   1
## 4 18.25 351 143 255 8.00 3.00  2  3 199.9 74.0 3890   1
## 5 20.07 225  95 170 8.40 2.76  1  3 194.1 71.8 3365   0
## 6 11.20 440 215 330 8.20 2.88  4  3 184.5 69.0 4215   1
```

If this is what you get, you are good to start!

1. Fit a simple linear regression model relating gasoline mileage y (miles per gallon) to engine displacement x_1 (cubic inches). Explain your coefficients. Any potential concern?

   Solution:

```
reg_fit_mpg_disp <- lm(formula = y~x1, data = mpg)
reg_fit_mpg_disp
```

```
##
## Call:
## lm(formula = y ~ x1, data = mpg)
##
## Coefficients:
## (Intercept)           x1
##    33.72268     -0.04736
```

There is a negative linear relationship between mpg and engine displacement. Therefore, vehicles that have a higher engine displacement will do less mileage per gallon.

2. Provide the 95% CI for β_0, β_1 and σ^2.

   Solution:

```
#confidence interval for \beta_0 and \beta_1
confint(reg_fit_mpg_disp, level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 30.77383383 36.67151954
## x1          -0.05694883 -0.03777032
```

```r
#confidence interval for \sigma^2
alpha <- 0.95
SS_res <-(sum(reg_fit_mpg_disp$residuals^2))
lower_bd <- SS_res / qchisq(alpha / 2, df = reg_fit_mpg_disp$df, lower.tail = FALSE)
upper_bd <- SS_res / qchisq(alpha / 2, df = reg_fit_mpg_disp$df, lower.tail = TRUE)


cat("Lower bound: ",lower_bd,'\n')
```

```
## Lower bound:  9.451467
```

```r
cat("Upper bound: ",upper_bd)
```

```
## Upper bound:  9.765482
```

3. With $\alpha = 0.05$, test if β_1 is significantly different from 0. Provide procedure and steps, for example, H_0 and H_1, the test statistic or p-value, and decision rule.

   Solution:

```r
(sum_reg_fit<-summary(reg_fit_mpg_disp))
```

```
##
## Call:
## lm(formula = y ~ x1, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7923 -1.9752  0.0044  1.7677  6.8171
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.722677   1.443903   23.36  < 2e-16 ***
## x1          -0.047360   0.004695  -10.09 3.74e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.065 on 30 degrees of freedom
## Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
## F-statistic: 101.7 on 1 and 30 DF,  p-value: 3.743e-11
```

As we can see, the p-value for β_1 is 3.74^{-11}, which is 0. Therefore we need to reject $H_0 : \beta_1 = 0$, since there is enough evidence at $\alpha = 0.05$ that $\beta_1 \neq 0$

4. Construct the ANOVA table and test for significance of regression.

Solution:

```
anova(reg_fit_mpg_disp)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 955.72  955.72  101.74 3.743e-11 ***
## Residuals 30 281.82    9.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. What percent of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement?

Solution:

```
sum_reg_fit$r.squared
```

```
## [1] 0.7722712
```

We can see that the multiple R^2 is equal to 0.7723, which means that 77.23% of the variability in mpg is explained by engine displacement

6. Find a 95% CI on the mean gasoline mileage if the engine displacement is 275 in^3 engine.

Solution:

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

```
##        fit      lwr      upr
## 1 20.69879 19.58807 21.80952
```

7. Suppose that we wish to predict the gasoline mileage obtained from a car with a 275 in^3 engine. Give a point estimate of mileage. Find a 95% prediction interval (PI) on the mileage. Compare the PI with the CI in 6. Explain the difference between them. Which one is wider, and why?

Solution:

```
#the prediction for that specific point
predict(reg_fit_mpg_disp, newdata = data.frame(x1= 275))
```

```
##        1
## 20.69879
```

```r
#the prediction interval for that estimation
predict(reg_fit_mpg_disp, newdata = data.frame(x1= 275), interval = "predict", level = 0.95)
```

```
##        fit      lwr      upr
## 1 20.69879 14.34147 27.05611
```

The prediction interval is wider, by a lot, since it includes the uncertainty about b_0, b_1, as well as the y_0 and ϵ, which are accounted for with the confidence interval.

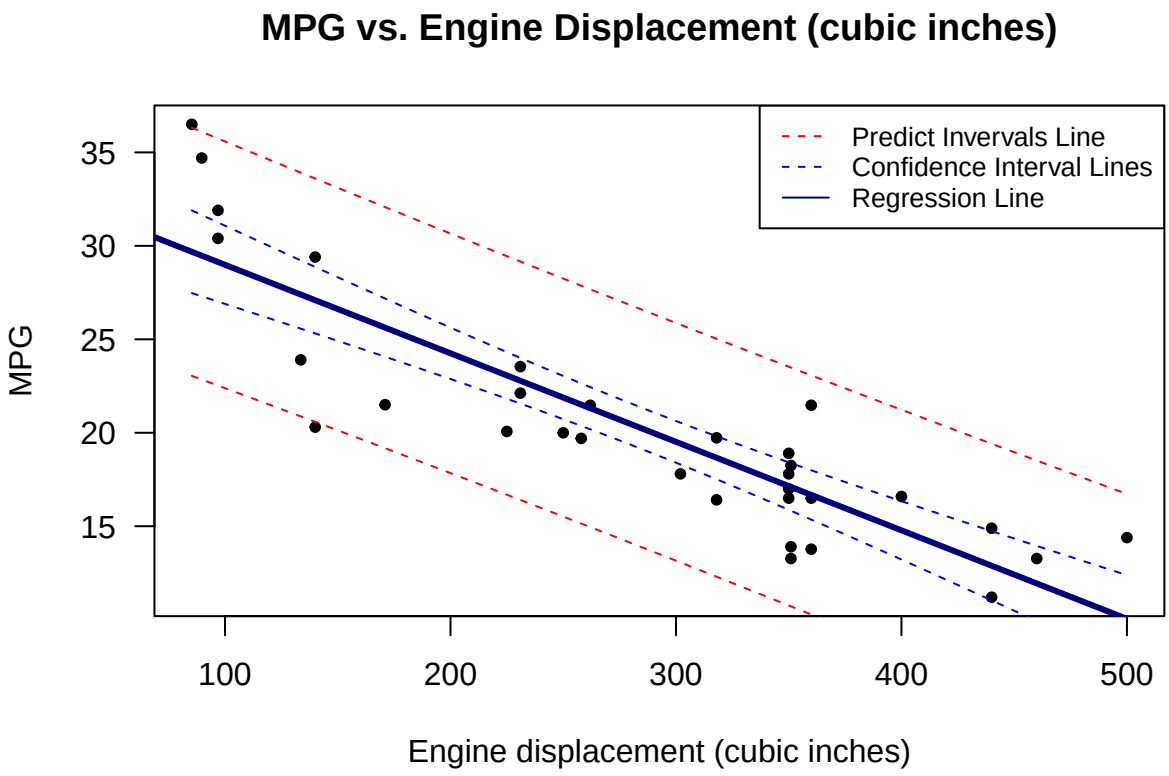
8. Plot data $\{(x_{1i}, y_i)\}_{i=1}^{32}$, the fitted regression line, CI for μ and PI for y in one figure. Add appropriate labels of axes, title, and legend. [*Hint:* Create a sequence of values of x, and obtain CI and PI for each value of x. Use `legend()` to add legends to a plot.]

Solution:

```r
plot(mpg$x1, mpg$y, pch=20, las =1 ,cex = 1,
     xlab = "Engine displacement (cubic inches)", ylab = "MPG",
     main = "MPG vs. Engine Displacement (cubic inches)")
abline(reg_fit_mpg_disp,col="navy", lwd=3)

#confidence interval
conf_x1<- seq(min(mpg$x1),max(mpg$x1), by=0.5)
conf_int<- predict(reg_fit_mpg_disp, newdata=data.frame(x1=conf_x1), interval="confidence",
                         level = 0.95)
lines(conf_x1, conf_int[,2], col="blue", lty=2)
lines(conf_x1, conf_int[,3], col="blue", lty=2)
#predict Interval
pred_x1<- seq(min(mpg$x1),max(mpg$x1), by=0.5)
conf_int<- predict(reg_fit_mpg_disp, newdata=data.frame(x1=pred_x1), interval="predict", level
lines(conf_x1, conf_int[,2], col="red", lty=2)
lines(conf_x1, conf_int[,3], col="red", lty=2)

#add legends
legend("topright", legend=c("Predict Invervals Line", "Confidence Interval Lines", "Regression
       col=c("red", "blue", "navy"), lty=c(2,2,1), cex=0.8)
```



9. Use the data and your fitted result to verify that

- (a) $\sum_{i=1}^{32}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$
- (b) $\sum_{i=1}^{32} y_i = \sum_{i=1}^{32} \hat{y}_i$
- (c) The LS regression line passes through the centroid $(\overline{x}, \overline{y})$
- (d) $\sum_{i=1}^{32} x_i e_i = 0$ (may not be exactly but numerically 0)
- (e) $\sum_{i=1}^{32} \hat{y}_i e_i = 0$ (may not be exactly but numerically 0)

Solution

(a)

```
sum(mpg$y - reg_fit_mpg_disp$fitted.values)
```

```
## [1] 3.552714e-15
```

which is numerically 0

(b)

```
sum(mpg$y)
```

```
## [1] 647.14
```

```
sum(reg_fit_mpg_disp$fitted.values)
```

```
## [1] 647.14
```

which is the same.

(c)

```
mean(mpg$x1)
```

```
## [1] 285.0437
```

```
mean(mpg$y)
```

```
## [1] 20.22312
```

```
#so we plug into the equation that we got for that line
# y =   33.72268 - 0.04736\beta_1
y =  33.72268 - 0.04736*mean(mpg$x1)
y
```

```
## [1] 20.22301
```

which shows that when we use $\bar{x}$ as the input for our fitted line, we get the same output as $\bar{x}$

(d) since $\sum_{i-1}^{32} \epsilon_i = \sum_{i-1}^{32} (y_i - \hat{y}_i)$

```
sum(mpg$x1 * (mpg$y - reg_fit_mpg_disp$fitted.values))
```

```
## [1] 6.181722e-13
```

which is numerically 0

(e)

```
sum(reg_fit_mpg_disp$fitted.values * (mpg$y - reg_fit_mpg_disp$fitted.values))
```

```
## [1] 1.101341e-13
```

which is numerically 0