STA 5207: Homework 2

Due: Friday, January 26 by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output

Exercise 1 (Using `lm` for Estimation) [35 points]

For this exercise we will use the `faithful` dataset. This is a default dataset in `R` , so there is no need to load it (when using `R`). Otherwise, you can find the data in `faithful.csv` on Canvas. The dataset contains 272 measurements of the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park in Wyoming. You should use `?faithful` to learn about the dataset's background. The variables in the dataset are

- `eruptions`: Duration of the eruption in minutes.
- `waiting`: Waiting time before the next eruption in minutes.

Suppose we would like to predict the duration of an eruption of the Old Faithful geyser based on the waiting time before the eruption.

1. (5 points) Give the simple linear regression model for this data set and the assumptions about the error terms.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

2. (4 points) What is the interpretation of β_1 *in the context of this problem.*

   Mean change in eruptions per minute waiting.

3. (8 points) Give the estimated regression equation. What is the interpretation of $\hat{\beta}_1$ *in the context of this problem*?

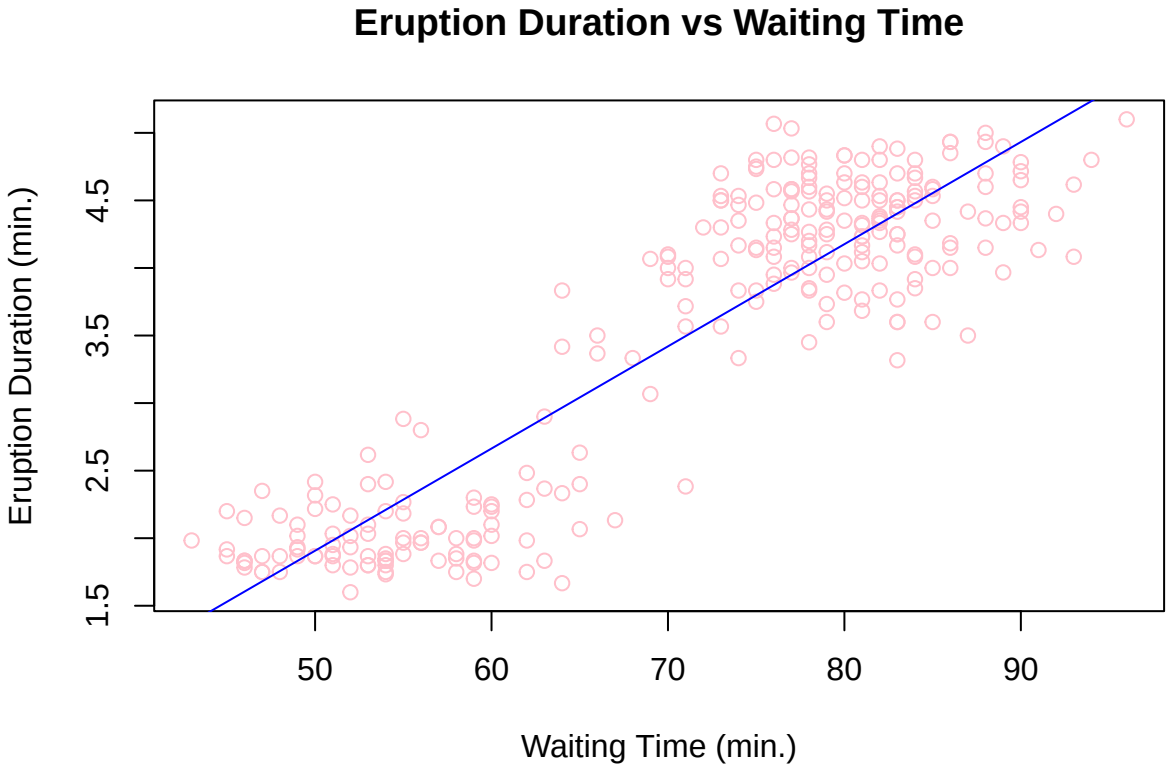
```
lmErupt = lm(eruptions ~ waiting, data = faithful)
```

   $eruptions = -1.87 + (.076)waiting + \epsilon$

   For every one minute of waiting time, the duration increases by .076 minutes.

4. (4 points) Create a scatter plot of the data and add the fitted regression line. Make sure your plot is well labeled and is somewhat visually appealing.

```
waiting = faithful$waiting
plot(faithful$waiting, faithful$eruptions, xlab = "Waiting Time (min.)", ylab = "Eruption Duration
abline(lmErupt, col = "blue")
```



5. (5 points) Use your model to predict the duration of an eruption based on a waiting time of **80** minutes. Do you feel confident in this prediction? Briefly explain.

```
unseen <- data.frame(waiting = 80)
predict(lmErupt, newdata = unseen)
```

```
##       1
## 4.17622
```

Yes I am confident in this prediction because it is simply interpolating data that is already present.

6. (5 points) Use your model to predict the duration of an eruption based on a waiting time of **120** minutes. Do you feel confident in this prediction? Briefly explain.

```
unseen <- data.frame(waiting = 120)
predict(lmErupt, newdata = unseen)
```

```
##        1
## 7.201338
```

No, as it is extrapolating.

7. (2 points) Report the residual standard error (RSE) for the model.

```
summary(lmErupt)$sigma
```

```
## [1] 0.4965129
```

The RSE is around .50

8. (2 points) Give the value and interpretation of R^2 for the model.

```
summary(lmErupt)$r.squared
```

```
## [1] 0.8114608
```

The R^2 is around .81

Exercise 2 (Comparing Models With R^2) [30 points]

For this exercise, we will use the data stored in `goalies.csv`. It contains career data for all 716 players in the history of the National Hockey League to play goaltender through the 2014-2015 season. The variables in the dataset are:

- `Player`: NHL Player Name
- `First`: First year of NHL career
- `Last`: Last year of NHL carerr
- `GP`: Games Played
- `GS` : Games Started
- `W`: Wins
- `L`: Losses
- `TOL`: Ties/Overtime/Shootout Losses
- `GA`: Goals Against
- `SA`: Shots Against
- `SV`: Saves
- `SV_PCT`: Save Percentages
- `GAA`: Goals Against Average
- `SO`: Shutouts
- `MIN`: Minutes
- `G`: Goals (that the player recorded, not opponents)
- `A`: Assists (that the player recorded, not opponents)
- `PTS`: Points (that the player recorded, not opponents)
- `PIM`: Penalties in Minutes

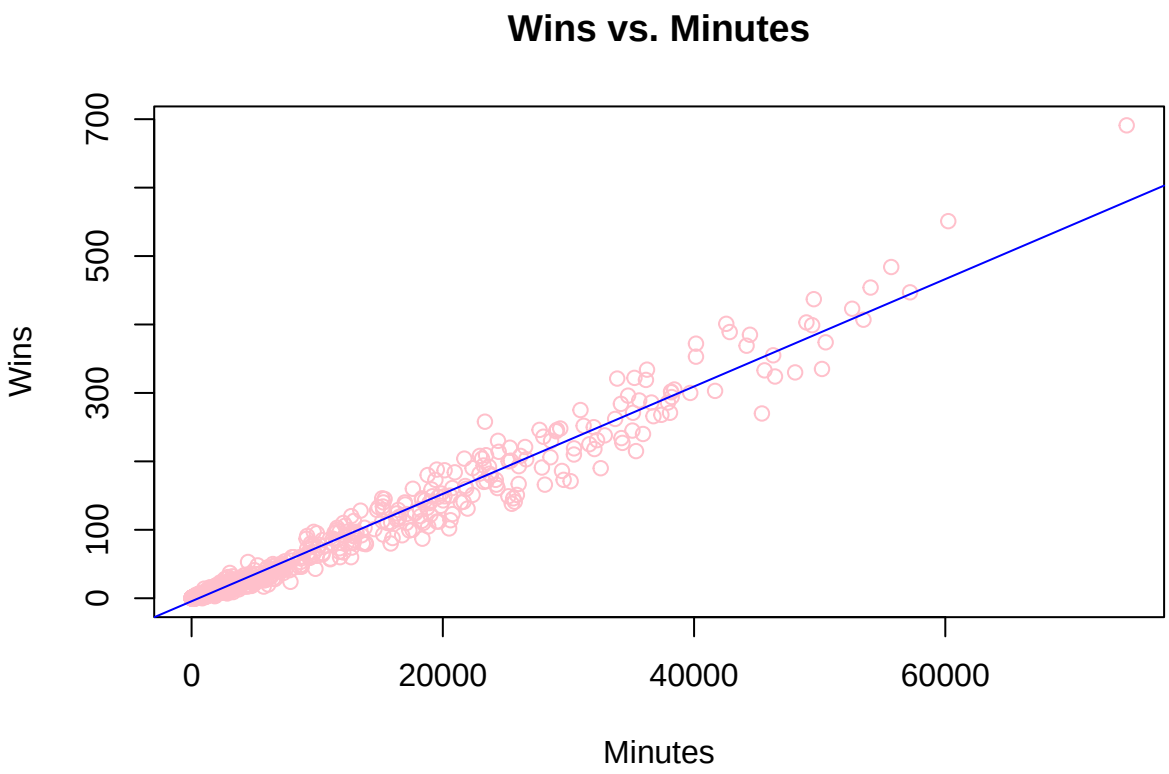
1. (6 points) Fit a model with "wins" as the response and "minutes" as the predictor. Report the value of R^2 for this model. Also provide a scatter plot of the fitted regression line.

```
wins = goalie$W
min = goalie$MIN

winMin = lm(wins ~ min, data = goalie)
summary(winMin)$r.squared
```

```
## [1] 0.9711568
```

```
plot(min, wins, xlab = "Minutes", ylab = "Wins", col = "pink", main = "Wins vs. Minutes")
abline(winMin, col = "blue")
```



The R^2 is around .97

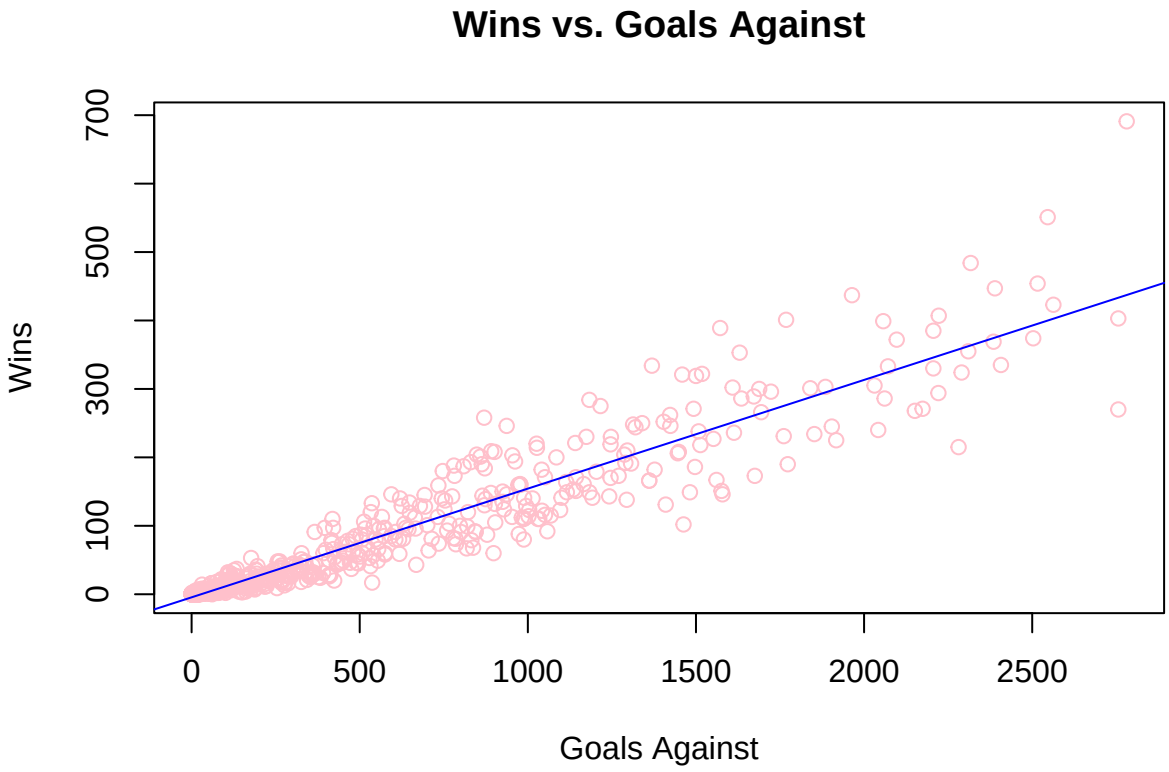
2. (6 points) Fit a model with "wins" as the response and "goals against" as the predictor. Report the value of R^2 for this model. Also provide a scatter plot of the fitted regression line.

```
wins = goalie$W
GA = goalie$GA

winGA = lm(wins ~ GA, data = goalie)
summary(winGA)$r.squared
```

```
## [1] 0.9007736
```

```
plot(GA, wins, xlab = "Goals Against", ylab = "Wins", col = "pink", main = "Wins vs. Goals Against"
abline(winGA, col = "blue")
```



The R^2 is around .90

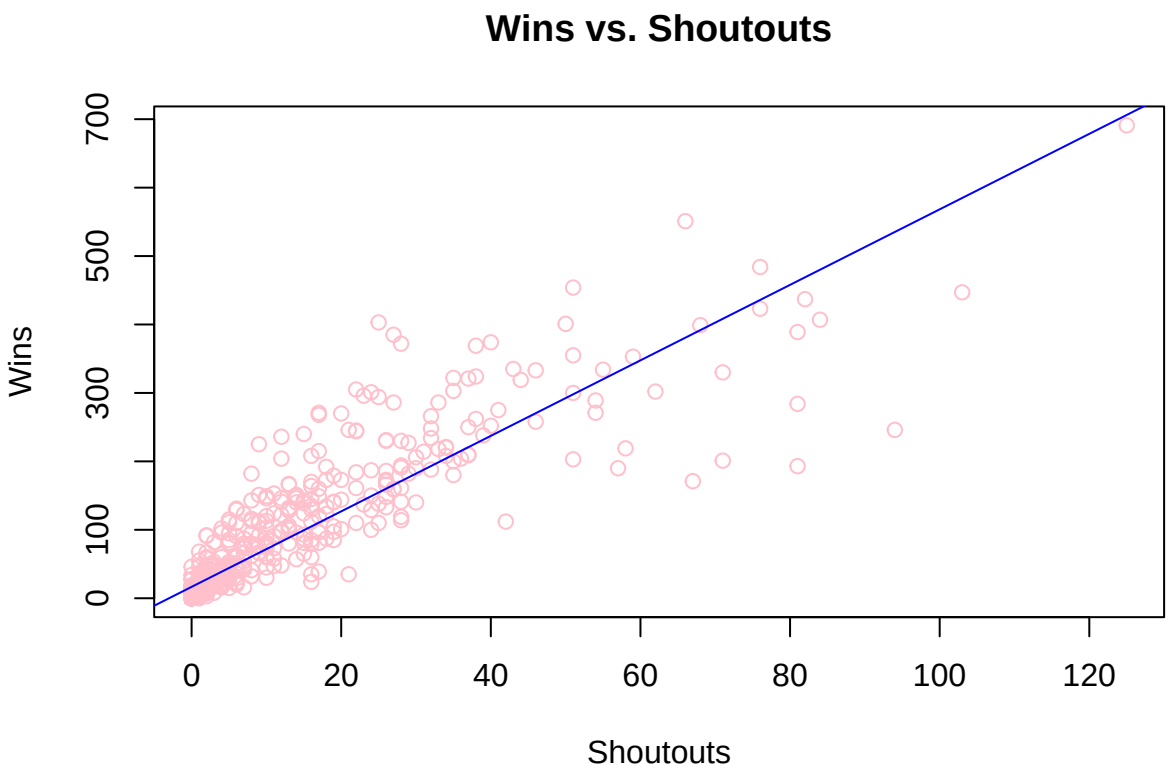
3. (6 points) Fit a model with "wins" as the response and "shutouts" as the predictor. Report the value of R^2 for this model. Also provide a scatter plot of the fitted regression line.

```
wins = goalie$W
SO = goalie$SO

winSO = lm(wins ~ SO, data = goalie)
summary(winSO)$r.squared
```

```
## [1] 0.7934997
```

```
plot(SO, wins, xlab = "Shoutouts", ylab = "Wins", col = "pink", main = "Wins vs. Shoutouts")
abline(winSO, col = "blue")
```



The R^2 is around .80

4. (12 points) If we use R^2 to measure each model's goodness-of-fit, which of the three predictors fits the data better? Briefly explain.

   Minutes would fit the data better as the R^2 is higher indicating that the model fits the data better.

Exercise 3 (Using `lm` for Inference) [35 points]

For this exercise, we will use the `cats` dataset from the `MASS` package. To load the dataset in `R`, run `data(cats, package='MASS')`. You can also find that data in `cats.csv` on Canvas. The dataset contains the following variables:

- `Sex`: The gender of the cat.
- `Bwt`: The body weight of the cat in kilograms.
- `Hwt`: The weight of the cat's heart in grams.

To read more about the dataset type `?cats` in RStudio.

1. (7 points) Fit the following simple regression model with the cat's heart weight as the response and its body weight as the predictor:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

   Use a t-test to test the significance of the regression. Report the following:

   - The null and alternative hypothesis
   - The value of the test statistic
   - The p-value of the test
   - A statistical decision at $\alpha = 0.05$

- A conclusion in the context of the problem.

```
model <- lm(Hwt ~ Bwt, data = cats)
summary(model)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## Bwt           4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Answer:

Null Hypothesis H_0: $\beta_1 = 0$ (There is no linear relationship between body weight and heart weight)

Alternative Hypothesis H_1: $\beta_1 \neq 0$ (There is a significant linear relationship)

Test Statistic: The t-value for the coefficient β_1 is 16.119.

p-value: The p-value associated with the t-test is <2e-16, which is extremely small.

Based on the t-test, there is significant evidence to suggest that there is a relationship between the body weight and heart weight. The coefficient for Bwt is significantly different from zero, indicating that body weight has a significant impact on heart weight.

2. (8 points) Use an F-test to test the significance of the regression. Report the following **and** discuss how they compare to the answers from the t-test in part 1:

   - The null and alternative hypothesis
   - The value of the test statistic
   - The p-value of the test
   - A statistical decision at $\alpha = 0.05$
   - A conclusion in the context of the problem.

     Answer:

   - Null Hypothesis H_0: H_0 states that all regression coefficients, except the intercept, are equal to zero.

     H_0:$\beta_i = 0$

   - Alternative Hypothesis (H_1): H_1 suggests that at least one regression coefficient is not equal to zero.

     H_1:At least one β_i is not equal to 0

   - Test Statistic: The F-statistic is 259.8 with degrees of freedom 1 and 142.

- p-value: The p-value associated with the F-test is <2.2e-16, which is extremely small.
- Statistical Decision: At a significance level (α) of 0.05, since the p-value is much less than 0.05, we reject the null hypothesis.
- Conclusion: The F-test provides strong evidence against the null hypothesis, indicating that at least one of the regression coefficients is not equal to zero. This aligns with the results of the t-test for the individual coefficient, suggesting a significant relationship between body weight and heart weight in the context of the problem.

3. (5 points) Give the 99% confidence interval for β_1. Give an interpretation of the interval in the context of the problem.

```
confint(model, level = 0.99)
```

```
##                 0.5 %   99.5 %
## (Intercept) -2.164125 1.450800
## Bwt          3.380656 4.687469
```

With 99% confidence, we can say that the true effect of body weight on heart weight is estimated to be between 3.38 and 4.69.

4. (5 points) Give the 90% confidence interval for β_0. Give an interpretation of the interval in the context of the problem.

```
confint(model, level = 0.90)
```

```
##                   5 %      95 %
## (Intercept) -1.502834 0.7895096
## Bwt          3.619716 4.4484094
```

With 90% confidence, we can say that the true average heart weight when the body weight is zero is estimated to be between -1.50 and 0.79.

5. (5 points) Report a 95% confidence interval to estimate the mean heart weight for body weights of 2.5 and 3.0 kilograms. Which of the intervals is wider? Why?

```
new_data <- data.frame(Bwt = c(2.5, 3.0))
predict(model, newdata = new_data, interval = "confidence", level = 0.95)
```

```
##         fit       lwr       upr
## 1  9.728494  9.464902  9.992087
## 2 11.745526 11.469954 12.021097
```

The wider confidence interval for the body weight of 3.0 kilograms means we can not be as sure when estimating mean heart rate at 3.0 kilograms.

6. (5 points) Report a 95% prediction interval to predict the heart weight for body weights of 2.5 and 4.0 kilograms.

```
new_data <- data.frame(Bwt = c(2.5, 4.0))
predict(model, newdata = new_data, interval = "prediction", level = 0.95)
```

```
##         fit       lwr      upr
## 1  9.728494  6.845352 12.61164
## 2 15.779588 12.830180 18.72900
```

- For a body weight of 2.5 kilograms, the model predicts a heart weight between approximately 6.85 and 12.61 grams with 95% confidence.
- For a body weight of 4.0 kilograms, the model predicts a heart weight between approximately 12.83 and 18.73 grams with 95% confidence.