Homework 5

Anahi Rodriguez

Due 10/5/2022

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

The point value for each exercise is noted in the exercise title.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment. We'll also read in data from a csv file. To access the data, you'll want to download the dataset from Canvas, and place it in the same folder as this R Markdown document. You'll then be able to use the following code to load in the data.

```
library(ggplot2)
library(MASS)
```

Exercise 1: Cat Model [20 points]

For this exercise we will use the `cats` dataset from the `MASS` package. You should use `?cats` to learn about the background of this dataset.

part a

Fit the following simple linear regression model in `R`. Use heart weight as the response and body weight as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in an `R` object called `cat_model`. Run a summary of the model and report the following:

- The estimated value for β_0, along with an interpretation
- The estimated value for β_1, along with an interpretation
- The estimated value for σ (no interpretation needed)

- The estimated value for R^2, along with an interpretation

Make sure that all interpretations include units and are given in the context of the problem.

```
# Use this code chunk for your answer.
cat_model = lm(data = cats, Hwt ~ Bwt)
summary(cat_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: The estimated value of $\beta_0 = -0.3567$, telling us that we should expect the average heart weight in g to equal -0.3567 for a cat with a body weight of 0 kilograms.

The estimated value of $\beta_1 = 4.0341$, telling us that for every additional gram heavier a cats body weight is, we expect it's heart weight to increase by 4.0341 grams on average.

The estimated value for $\sigma = 1.452$

The estimated value for $R^2 = 0.6466$, telling us that 64.66% of the variation in the heart weight for a cat can be explained by its linear relationship with its body weight.

part b

Does the estimate for β_0 provide a meaningful value? Is it reliable? Explain.

Answer: The estimated value for β_0 is not meaningful because it is not realistic to see a cat with a body weight of 0 kilograms nor do we expect a negative heart weight. It is not reliable because a body weight of 0 kilograms is not within the range of body weight in our sample.

part c

Calculate the estimated standard error for the estimated slope. That is, calculate the **standard deviation** for the sampling distribution of the estimated slope.

Then, compare this value to that reported in the R output from Exercise 1a.

```
# Use this code chunk for your answer.
x_bar = mean(cats$Bwt)
sxx = sum((cats$Bwt - x_bar)^2)
sqrt((1.452^2)/sxx)
```

```
## [1] 0.2501972
```

```r
sqrt((1.452^2)/sxx) - 0.2503
```

```
## [1] -0.0001028456
```

Comparison: These values are very similar with the written out calculation having more precision

---

Exercise 2: Cat Inference [30 points]

For this exercise, we will continue analyzing the model fit in Exercise 1.

part a

Let's assume that for another animal, the relationship between body weight and heart weight is such that a 1 kg increase in body weight tends to result in a 3.25 g increase in heart weight, on average.

Might this same relationship be true for cats, or is there evidence that this relationship for cats is different?

Use a t test to test:

- $H_0 : \beta_1 = 3.25$
- $H_1 : \beta_1 \neq 3.25$

Report the following in your document outside the code block.

- The value of the updated test statistic
- Whether the p-value of the test will be less than 0.05, or greater than 0.05
- A statistical decision at $\alpha = 0.05$

```r
# Use this code chunk for your answer, if needed.
summary(cat_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
```

```r
t_stat = (4.0341 - 3.25)/0.2503
t_stat
```

```
## [1] 3.132641
```

Answer: the test statistic = 3.132641, and since this value is larger than 2, the p-value will be smaller than 0.05 indicating that we will reject the null in favor of a different value.

part b

Compute a 70% confidence interval for β_0. *(You may use the `confint` function for this assignment.)*

Then, based on this confidence interval, anticipate the decision of the following two hypothesis tests:

- H_0: $\beta_0 = 0$ vs.H_1: $\beta_0 \neq 0$
- H_0: $\beta_0 =$ -1 vs. H_1: $\beta_0 \neq$ -1

Explain your reasoning.

```
# Use this code chunk for your answer.
confint(cat_model, parm = '(Intercept)', level = 0.7,)
```

```
##                  15 %      85 %
## (Intercept) -1.076791 0.3634664
```

Answer: We would fail to reject the null for both hypothesis tests

part c

The inference procedures from parts a and b being valid relies on some assumptions being met. First, write out the four assumptions.

Answer: Normality of errors, constant variance of y at each x, linear relationship between x and y, and independent true errors

part d

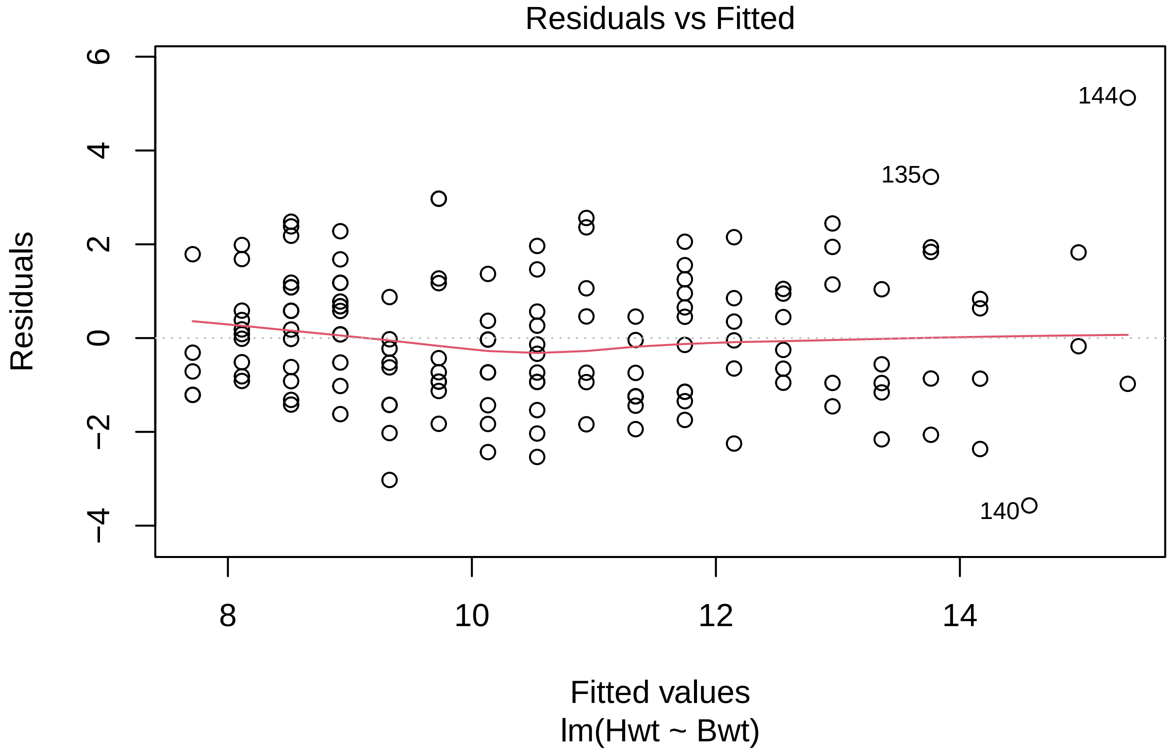
Which of these four assumptions cannot be checked with a plot?

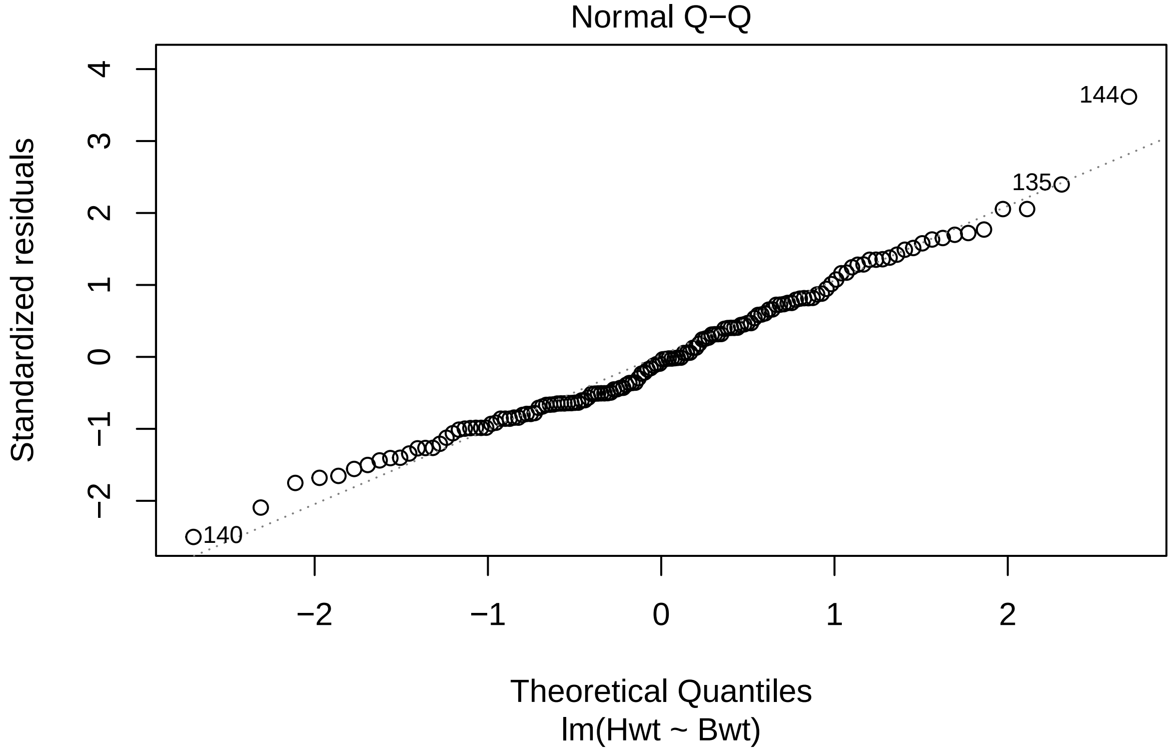
Answer: Independent true errors

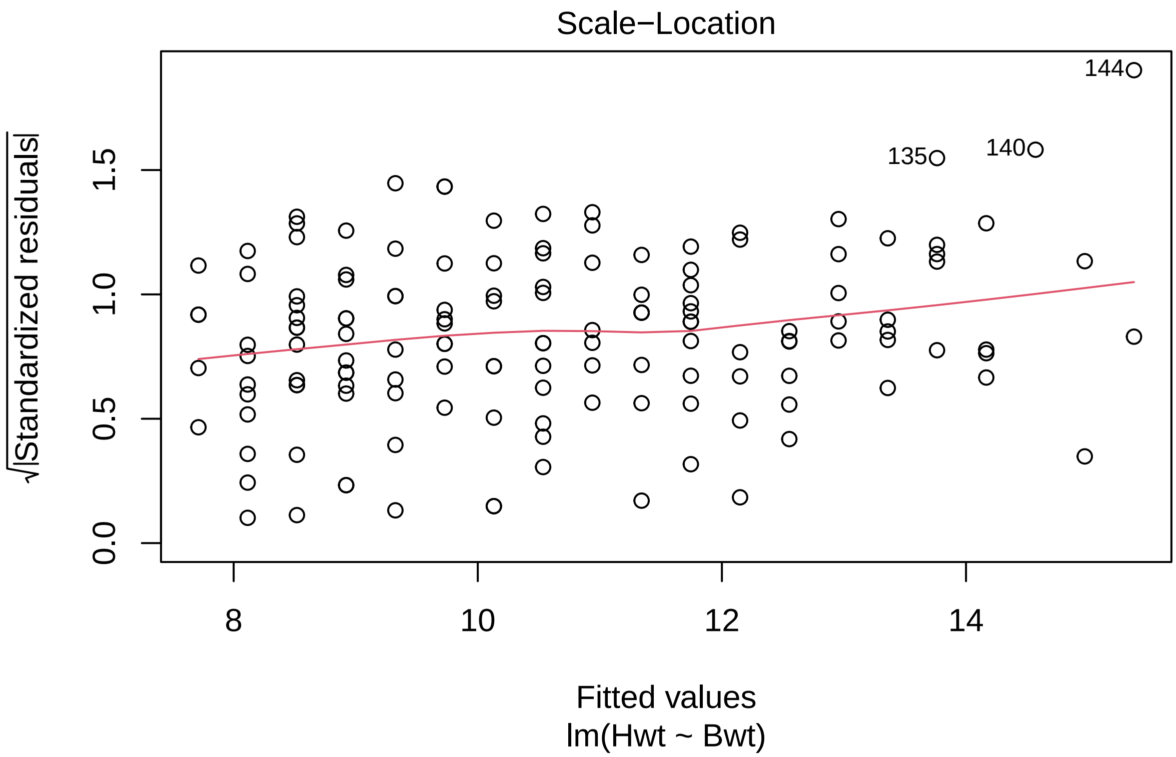
part e

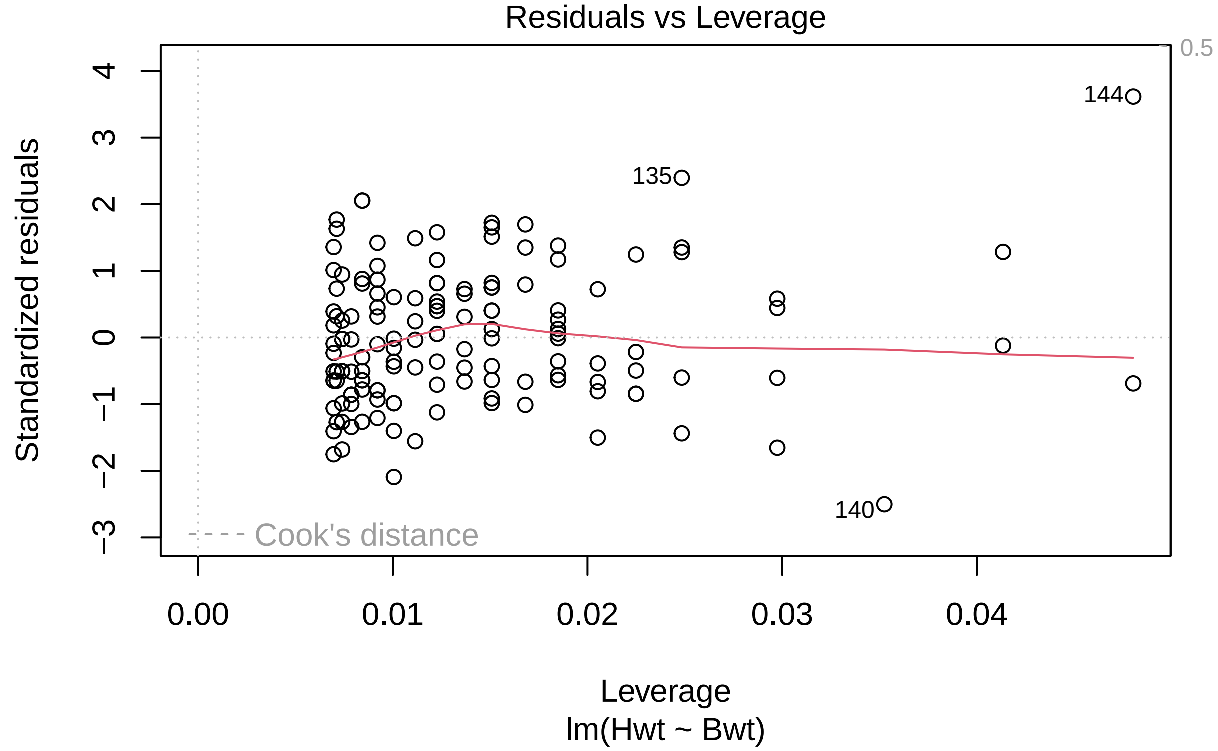
Create the two plots (it's ok if you end up with 4 plots) that we can use to check our assumptions for the cats data. Interpret these two plots. Be sure to specify if the assumptions seem reasonable and describe what you see that supports your conclusions.

```
# Use this code chunk for your answer.
plot(cat_model)
```


Residuals vs Fitted
Residuals
6
4
2
0
−2
−4
144
135
140
8
10
12
14
Fitted values
lm(Hwt ~ Bwt)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
144
135
140
−2
−1
0
1
2
Theoretical Quantiles
lm(Hwt ~ Bwt)





```
shapiro.test(resid(cat_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(cat_model)
## W = 0.9845, p-value = 0.1046
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(cat_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  cat_model
## BP = 9.5294, df = 1, p-value = 0.002022
```

Answer: I believe that the fitted values vs residuals plot shows that x and y have a reasonably linear relationship as the values are generally centered around 0. Using the QQ-plot to check for normality of errors shows that this assumption is met. Additionally, the shapiro-wilk test supports this conclusion as we fail to reject the null of normally distributed errors. I would say that the assumption of equal variance is not met as the fitted vs. residual plot graphs seem to have a few observations outside of the variance of 2 which many of the other observations fall within. Additionally, the results of the Breush-Pagan test tell us that we reject the null of homoskedasticity.

Exercise 3: Cat Model Predictions [20 points]

Continue analyzing the cats dataset using the model generated in Exercise 1.

part a

Use a 99% confidence interval to estimate the mean heart weight for body weights of 2.1 kilograms as well as for 2.8 kilograms.

Which of the two intervals is wider? How might you use the variance formula for $\hat{y}(x)$ to know beforehand which would be wider?

```
# Use this code chunk for your answer.
predict(cat_model, newdata = data.frame('Bwt' = c(2.1, 2.8)), interval = 'confidence', level = 0.99)
```

```
##         fit       lwr       upr
## 1  8.114869  7.599225  8.630513
## 2 10.938713 10.618796 11.258630
```

```
mean(cats$Bwt)
```

```
## [1] 2.723611
```

```
sd(cats$Bwt)
```

```
## [1] 0.4853066
```

Answer: The interval for 2.1 is wider because 2.1 is farther form the average value for body weight compared to 2.8. The variance formula is calculated using the difference between individual observations of x and the mean value of x.

part b

Use a 99% prediction interval to predict the heart weights for body weights of 4.2 kilograms. Interpret this interval in context. Do you trust this prediction interval? Explain why or why not.

```
# Use this code chunk for your answer.
predict(cat_model, newdata = data.frame('Bwt' = c(4.2)), interval = 'prediction', level = 0.99)
```

```
##       fit      lwr      upr
## 1 16.5864 12.66088 20.51192
```

```
summary(cats$Bwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.000   2.300   2.700   2.724   3.025   3.900
```

Answer: I would not trust this prediction interval due to extrapolation as 4.2 kg is outside of the range of body weights of cats in our sample

part c

A confidence interval for the mean heart weights is reported as: (13.3427, 14.1827). Determine the cat body weight that this confidence interval corresponds to.

```
# Use this code chunk, if needed, to perform any calculations.
bounds_avg = (13.3427 + 14.1827)/2
x_star = (bounds_avg - as.numeric(coef(cat_model)[1]))/as.numeric(coef(cat_model)[2])
x_star
```

```
## [1] 3.500035
```

Answer: This corresponds to a cat body weight of 3.500035 kgs.

part d

Suppose that I want to generate a confidence interval for the mean heart weights that is narrower than that described in part c. Describe two ways that I can create a narrower confidence interval.

Answer: Use values for body weight closer to the average body weight or decrease the level of confidence you are calculating (for example, calculate 70% CI instead of 90% CI)

Exercise 4: Simulations [25 points]

Consider the model

$$Y = 4 + 0x + \varepsilon$$

with

$$\varepsilon \sim N(\mu = 0, \sigma^2 = 36)$$

Before answering the following parts, set a seed value equal to **your** birthday, as was done in class (this time in the format yyyymmdd).

```
birthday = 20010216
set.seed(birthday)
```

part a

Repeat the process of simulating $n = 125$ observations from the above model 2500 times. Use the x created in the code chunk below for all iterations.

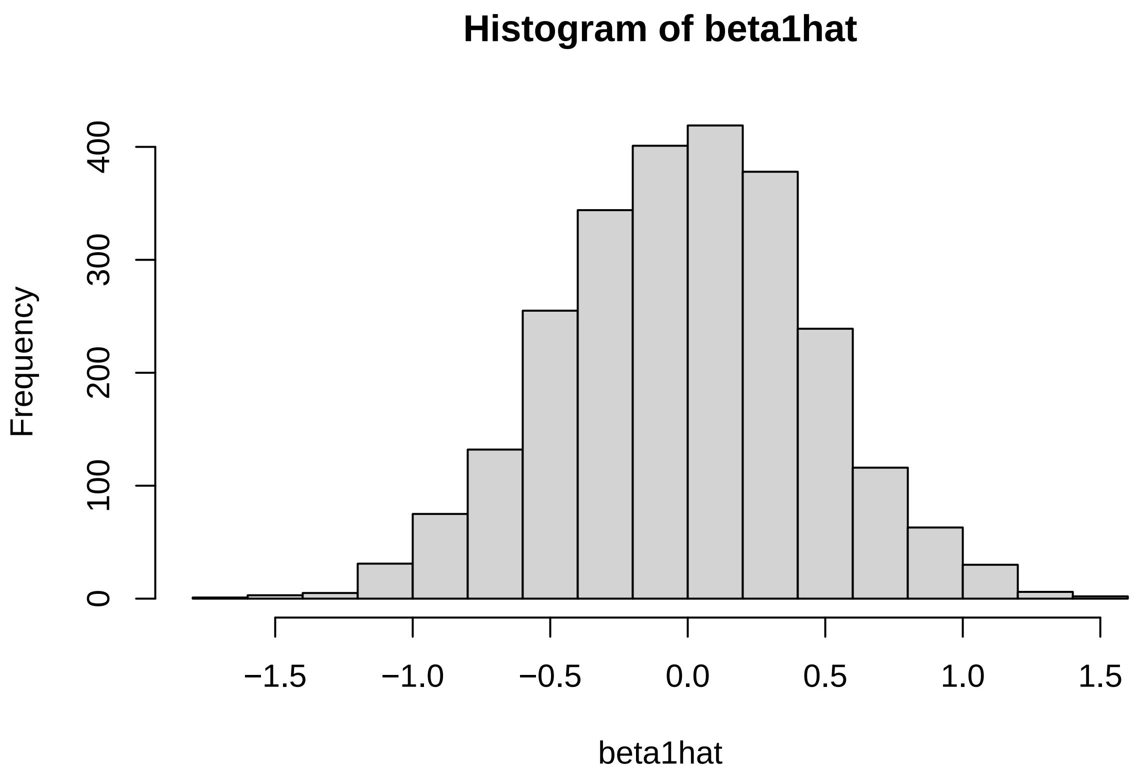
```
n = 125
x = runif(n, -2, 2)
reps = 2500
beta0 = 4
beta1 = 0
sigma = 6

beta0hat = vector(length = reps)
beta1hat = vector(length = reps)

for(i in 1:reps){
  epsilon = rnorm(n, 0, sigma)
  y = beta0 + beta1 * x + epsilon
  mymodel = lm(y ~ x)
  beta0hat[i] = coef(mymodel)[1] ## change this line of code so 00 is now the estimated intercept
  beta1hat[i] = coef(mymodel)[2] ## change this line of code so 01 is now the estimated slope
}
ests = data.frame(cbind(beta0hat, beta1hat))
```

The general structure has been provided for you. The last two lines of the for loop have placeholder values of 00 and 01. Adjust these values that are saved at the end of the for loop to be the appropriate values for beta0hat and beta1hat.

part b

Plot a histogram of `beta1hat` values based on your simulation. Describe this histogram, including comments on the shape, center, spread, and outliers.

```
# Use this code chunk for your answer.
hist(beta1hat)
```



Answer: This histogram is close to a normal distribution with a mean at 0 and the majority of observations within the -1 to 1 range. Additionally, there seem to exist a few outliers at both the positive and negative extremes of 2

part c

Import the `skeptic.csv` data (found on Canvas) and fit a SLR model. (Check the variable names in the skeptic dataset, as the names should indicate which to use as your y variable and which for your x variable).

Print the estimated coefficient for β_1.

```
# Use this code chunk for your answer.
setwd("~/Desktop/data")
df = read.csv("skeptic.csv")
lm1 = lm(y ~ x, data = df)
coef(lm1)[2]
```

```
##         x
## 0.6895595
```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4987  -4.1277  -0.1347   3.6636  15.3081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7269     0.5430   8.705 1.73e-14 ***
## x             0.6896     0.4939   1.396    0.165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.053 on 123 degrees of freedom
## Multiple R-squared:  0.0156, Adjusted R-squared:  0.007597
## F-statistic: 1.949 on 1 and 123 DF,  p-value: 0.1652
```

part d

Your goal for this part of the question is to determine if you think the skeptic data could reasonably have been simulated from the model described at the beginning of this exercise. I won't tell you exactly how to do that, but I will leave the following hint:

If the skeptic data really was simulated from this model, then how unusual is the $\hat{\beta}_1$ we found from the skeptic data? How can our simulated $\hat{\beta}_1$ values help us quantify this answer?

```r
# Use this code chunk for your answer.
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4987  -4.1277  -0.1347   3.6636  15.3081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7269     0.5430   8.705 1.73e-14 ***
## x             0.6896     0.4939   1.396    0.165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.053 on 123 degrees of freedom
## Multiple R-squared:  0.0156, Adjusted R-squared:  0.007597
## F-statistic: 1.949 on 1 and 123 DF,  p-value: 0.1652
```

```r
test_stat2 = (0.6896 - 0)/0.4939
test_stat2
```

```
## [1] 1.396234
```

part e

Based on your investigation in **part d**, do you think the skeptic data could have been simulated from the model provided in **part a**?

Answer: I believe that the skeptic data could have been simulated from the model provided in part a because based on a hypothesis test where $H_0 = 0$ and $H_1 \neq 0$, we fail to reject the null using the estimated $\beta_1 = 0.6896$

Exercise 5: Formatting [5 points]

The last five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- selected **page 1 (with your name)** and this page for this exercise (Exercise 5)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file