Problem Set 7

Regression

[YOUR NAME]

Due Date: 2024-03-01

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...` . Accept defaults and save this file as `[LAST NAME]_ps7.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps7.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `mv.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/mv.Rds).

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 8 total points, plus two extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compiled the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Also note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _____

Question 0

Require `tidyverse` and load the `mv.Rds` (https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/5_Regression/data/mv.Rds?raw=true') data to an object called `movies` .

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
movies <- read_rds('https://github.com/jbisbee1/DS1000_S2024/blob/main/data/mv.Rds?raw=true')
```

Question 1 [2 points]

In this problem set, we will answer the following research question: “Do movies that have a higher Bechdel Score (`bechdel_score`) make more money (`gross`)?” First, write out a **theory** that answers this question and transform it into a **hypothesis**. To learn more about the Bechdel Score, please read this Wikipedia entry (https://en.wikipedia.org/wiki/Bechdel_test).

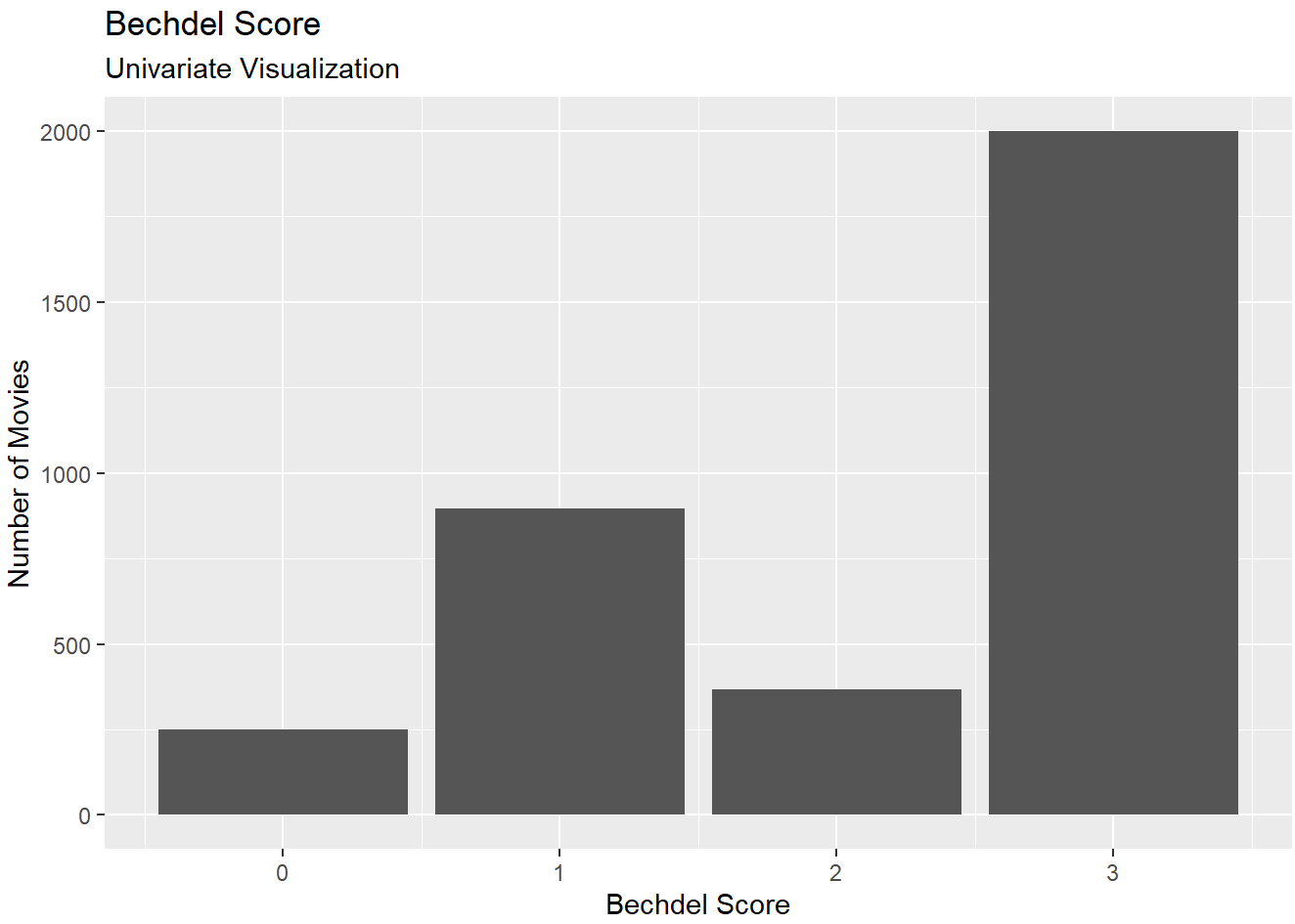
> I theorize that movies which have deeper female characters are appealing to women and men, but particularly women who have been underrepresented in cinema. Women comprise half of the population and pay money to watch movies. Therefore, I hypothesize that the relationship between the Bechdel Score (X) and the movie’s gross (Y) should be positive.

Question 2 [1 point]

Based on your theory, which variable is the X variable (i.e., the independent variable or the predictor)? Which variable is the Y variable (i.e., the dependent variable or the outcome)? Use **univariate** visualization to create two plots, one for each variable. Do you need to apply a log-transformation to either of these variables? Why? Then create a multivariate visualization of these two variables, making sure to put the independent variable on the x-axis and the dependent variable on the y-axis. Make sure to log the data if you determined this was necessary in the previous question! Does the visualization support your hypothesis?

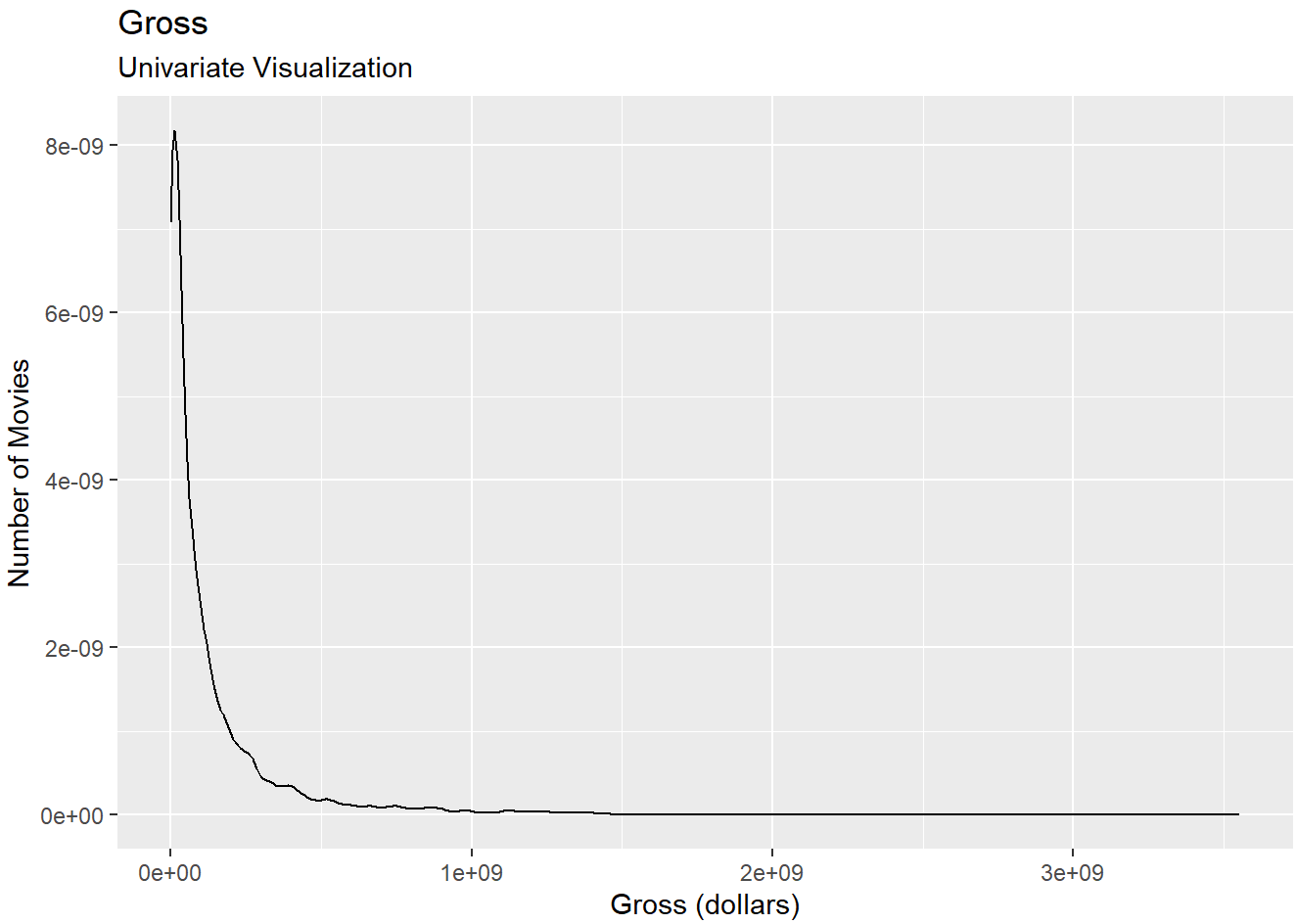
```
# Predictor: Bechdel
movies %>%
  ggplot(aes(x = bechdel_score)) +
  geom_bar() +
  labs(title = 'Bechdel Score',
       subtitle = 'Univariate Visualization',
       x = 'Bechdel Score',
       y = 'Number of Movies')
```

```
## Warning: Removed 4162 rows containing non-finite values (`stat_count()`).
```



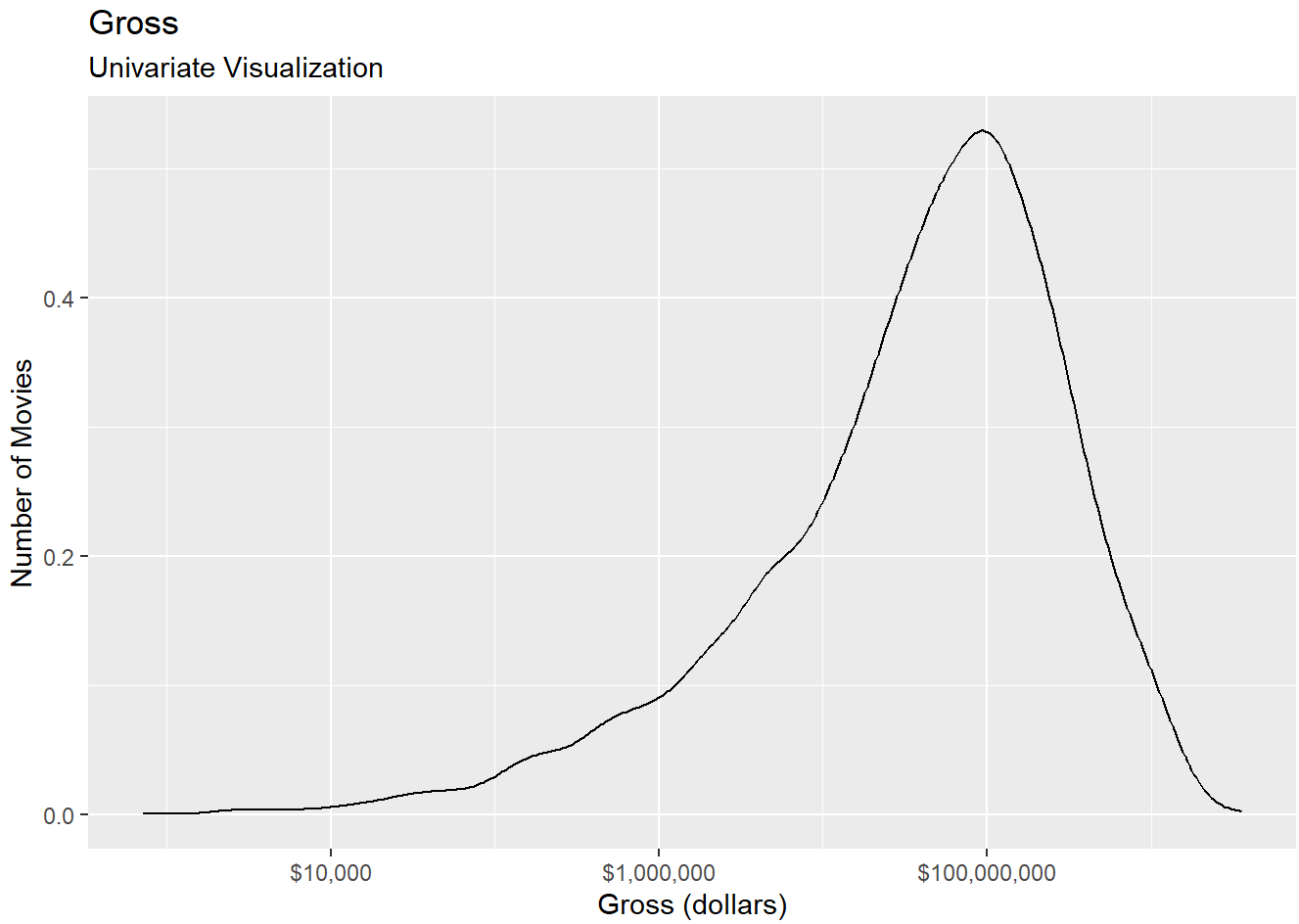

```
# Outcome: gross
movies %>%
  ggplot(aes(x = gross)) +
  geom_density() +
  labs(title = 'Gross',
       subtitle = 'Univariate Visualization',
       x = 'Gross (dollars)',
       y = 'Number of Movies')
```

```
## Warning: Removed 3668 rows containing non-finite values (`stat_density()`).
```



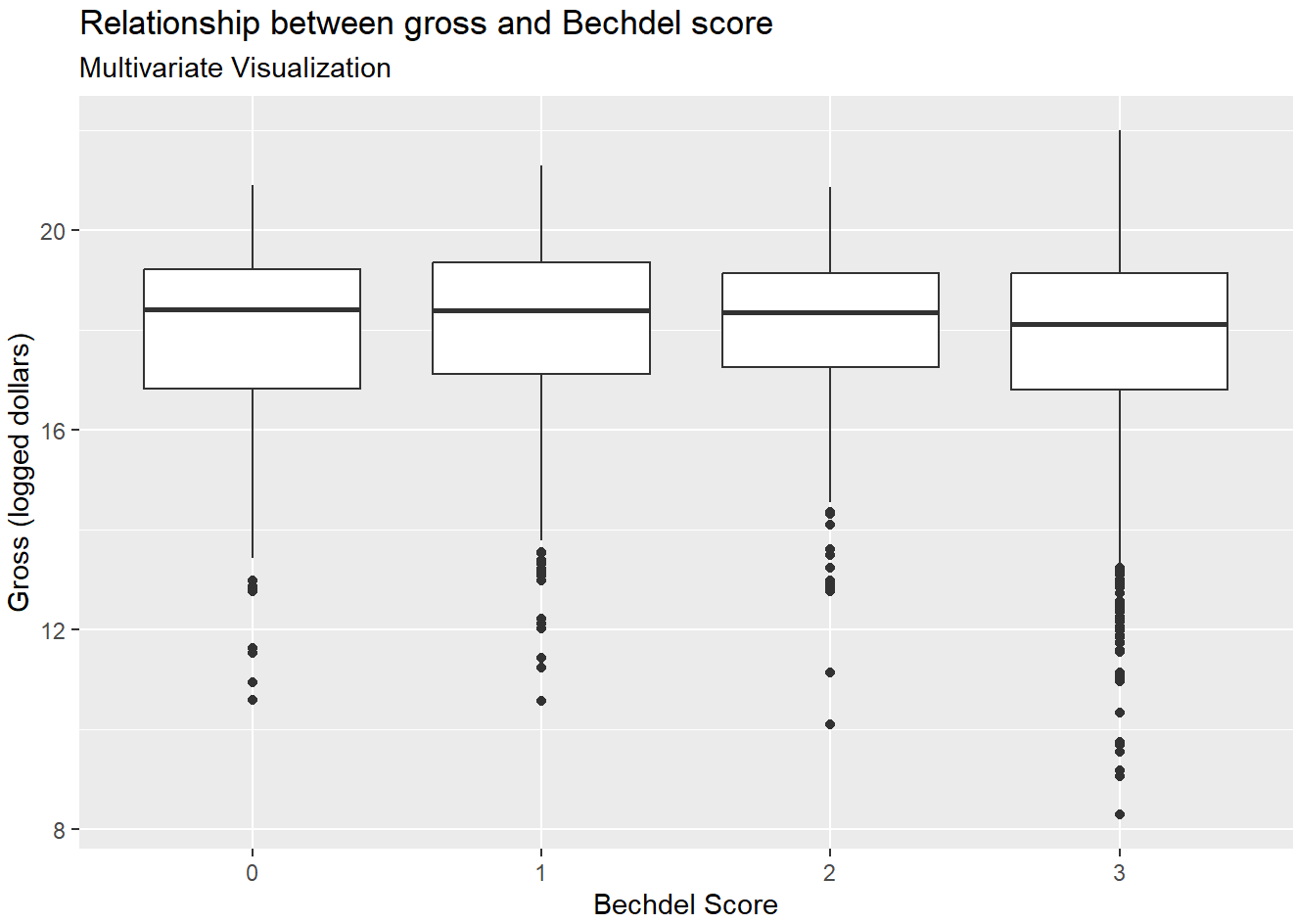

```
movies %>%
  ggplot(aes(x = gross)) +
  geom_density() +
  scale_x_log10(label = scales::dollar) +
  labs(title = 'Gross',
       subtitle = 'Univariate Visualization',
       x = 'Gross (dollars)',
       y = 'Number of Movies')
```

```
## Warning: Removed 3668 rows containing non-finite values (`stat_density()`).
```




```
movies %>%
  drop_na(bechdel_score) %>%
  mutate(log_gross = log(gross)) %>%
  ggplot(aes(x = factor(bechdel_score),y = log_gross)) +
  geom_boxplot() +
  labs(title = 'Relationship between gross and Bechdel score',
       subtitle = 'Multivariate Visualization',
       x = 'Bechdel Score',
       y = 'Gross (logged dollars)')
```

```
## Warning: Removed 1114 rows containing non-finite values (`stat_boxplot()`).
```



According to my theory, the Bechdel score is the predictor / independent / X variable and the movie's gross is the outcome / dependent / Y variable. Univariate visualization revealed that the gross variable is highly skewed, meaning that I should use a log transformation. The multivariate visualization does not support my hypothesis. Movies that score a 1 on the Bechdel score make slightly more money than those which score a 0, but those that score the highest (a 3) make less money than all other categories.

Question 3 [2 points]

Now estimate the regression using the `lm()` function. Describe the output of the model in English, talking about the intercept, the slope, and the statistical significance.

```
movies_analysis <- movies %>%
  mutate(log_gross = log(gross)) %>%
  drop_na(log_gross,bechdel_score)

model_gross_bechdel_score <- lm(formula = log_gross ~ bechdel_score,
                        data = movies_analysis)

summary(model_gross_bechdel_score)
```

```
##
## Call:
## lm(formula = log_gross ~ bechdel_score, data = movies_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4747 -0.9196  0.3796  1.3447  4.2295
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.04380    0.09552 188.903   <2e-16 ***
## bechdel_score -0.09411    0.03927  -2.397   0.0166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.981 on 2395 degrees of freedom
## Multiple R-squared:  0.002393,   Adjusted R-squared:  0.001976
## F-statistic: 5.744 on 1 and 2395 DF,  p-value: 0.01662
```

```
exp(18.04380)
```

```
## [1] 68599788
```

```
(exp(-0.09411)-1)*100
```

```
## [1] -8.981736
```

The model indicates that movies which scored a zero had an average gross of 18.04 logged dollars, or $6.85 million. Each additional point of Bechdel score is associated with an decrease of 0.094 logged dollars, which is better expressed as an almost 9 percentage point decline. This relationship is statistically significant, with a p-value of 0.017, meaning we are about 98.3% confident in our conclusion.

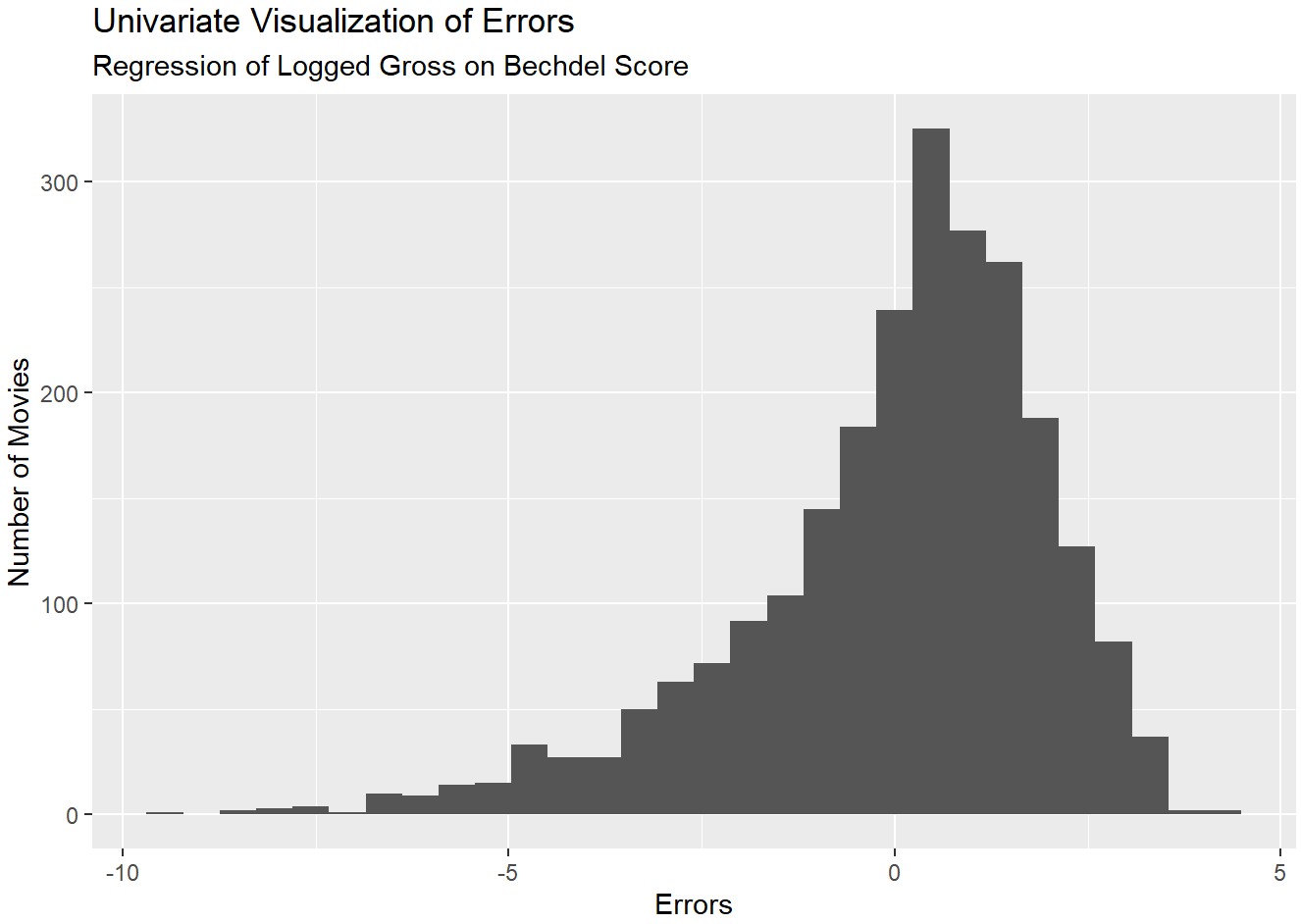
Question 4 [2 points]

Now calculate the model's prediction errors and create both a univariate and multivariate visualization of them. Based on these analyses, would you say that your model does a good job predicting how much money a movie makes? **Make sure to reference both the univariate and multivariate visualization of the errors!** Use the prediction errors to calculate the RMSE in the full data. Then calculate the RMSE using 100-fold cross validation with an 80-20 split and take the average of the 100 estimates.

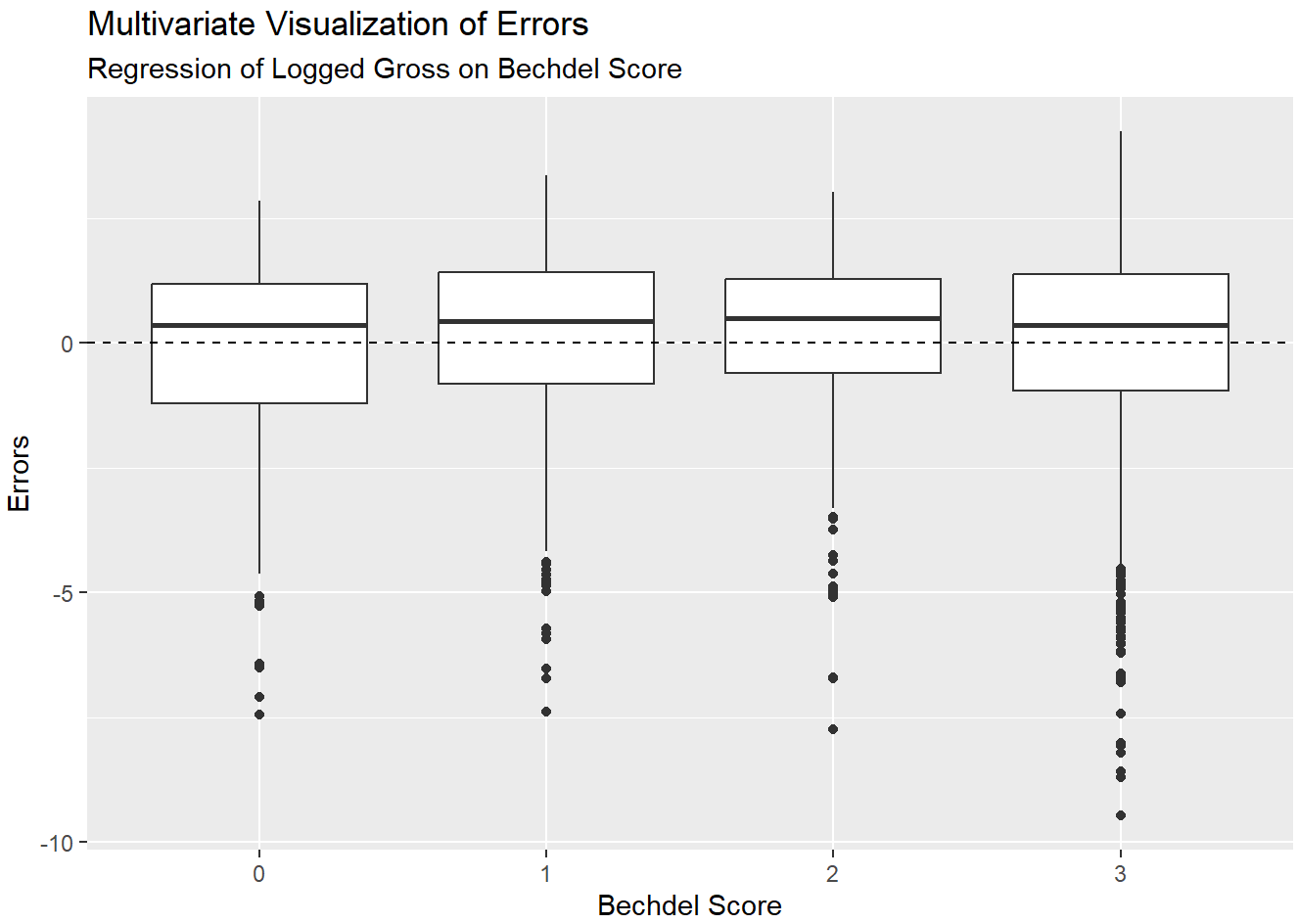
```
movies_analysis <- movies_analysis %>%
  mutate(preds = predict(model_gross_bechdel_score)) %>%
  mutate(errors = log_gross - preds)

# Univariate
movies_analysis %>%
  ggplot(aes(x = errors)) +
  geom_histogram() +
  labs(title = 'Univariate Visualization of Errors',
       subtitle = 'Regression of Logged Gross on Bechdel Score',
       x = 'Errors',
       y = 'Number of Movies')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# Multivariate
movies_analysis %>%
  ggplot(aes(x = factor(bechdel_score),y = errors)) +
  geom_boxplot() +
  geom_hline(yintercept = 0,linetype = 'dashed') +
  labs(title = 'Multivariate Visualization of Errors',
       subtitle = 'Regression of Logged Gross on Bechdel Score',
       x = 'Bechdel Score',
       y = 'Errors')
```




```
# RMSE
# RMSE Full Data
movies_analysis %>%
  mutate(se = errors^2) %>%
  summarise(mse = mean(se)) %>%
  mutate(rmse = sqrt(mse))
```

```
## # A tibble: 1 × 2
##     mse  rmse
##   <dbl> <dbl>
## 1  3.92  1.98
```

```
# RMSE 100-fold CV
set.seed(123)
cvRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:100) { # Loop 100 times
  inds <- sample(1:nrow(movies_analysis),size = round(.8*nrow(movies_analysis)),replace = F)

  train <- movies_analysis %>% slice(inds)
  test <- movies_analysis %>% slice(-inds)

  m <- lm(formula = log_gross ~ bechdel_score,data = train)

  test$preds <- predict(m,newdata = test)

  e <- test$log_gross - test$preds
  se <- e^2
  mse <- mean(se)
  rmse <- sqrt(mse)
  cvRes <- c(cvRes,rmse)
}

mean(cvRes)
```

```
## [1] 1.962773
```

```
mean(cvRes < 1.98)
```

```
## [1] 0.62
```

I would conclude that the model is poor based on the univariate and multivariate visualization of the errors. The univariate visualization of the errors is not symmetric around zero, meaning the model overpredicts the logged gross more than it underpredicts the logged gross. A good model should have symmetrically distributed errors around zero. The multivariate visualization errors is also poor, indicating that the model underpredicts logged gross across all Bechdel score values. The average RMSE calculated from 100 cross validation steps is slightly smaller than the RMSE calculated from the full data (1.96 in the cross validated analysis versus 1.98 in the full data). However, this is not a statistically significant difference, since the cross-validated RMSE is smaller than the full data RMSE only 62% of the time.

Extra Credit [2 Points]

Taking a step back, do you trust these results? What concerns might you have about the model? Can you propose a “control” to add that would speak to your concerns? Run the regression with this control and re-interpret the results.

I don’t trust these results, because they might be spurious. In particular, I worry that movies that score better on a Bechdel Test receive less funding. This assumption is based on my knowledge that most producers in Hollywood are men. They might therefore be less interested themselves in movies that score higher on the Bechdel test, and thus provide less funding for these types of movies. Since we know that a movie’s gross is a function of it’s budget, it might not be that having higher Bechdel scores CAUSE movies to make less money, but rather that higher Bechdel scores cause movies to get less funding, creating the spurious conclusion that the representation of women in film causes lower earnings.

To test my assumption, I propose adding the movie’s budget as a control to the regression. By adding this variable as a control, we can compare two movies that have the same budget, but one scores higher on the Bechdel test than the other. Based on this analysis (summarized below), I find that scoring higher on the Bechdel test is correlated with higher gross, reversing the conclusions drawn from above. Specifically, I find that a one unit increase in the Bechdel score is associated with a 0.099 increase in the logged gross, or roughly a 10 percentage point increase in the amount of money the movie makes. This conclusion is highly statistically significant (p-value < .01).

```
movies_analysis <- movies %>%
  mutate(log_gross = log(gross),
         log_budget = log(budget)) %>%
  drop_na(log_gross,log_budget,bechdel_score)

model_gross_bechdel_score_budget <- lm(log_gross ~ bechdel_score + log_budget,
                                       movies_analysis)

summary(model_gross_bechdel_score_budget)
```

```
##
## Call:
## lm(formula = log_gross ~ bechdel_score + log_budget, data = movies_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6328 -0.5309  0.1307  0.6739  7.9406
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.00049    0.35855   5.579 2.73e-08 ***
## bechdel_score  0.09924    0.02505   3.962 7.67e-05 ***
## log_budget     0.92154    0.01990  46.300  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.165 on 2055 degrees of freedom
## Multiple R-squared:  0.5113, Adjusted R-squared:  0.5108
## F-statistic:  1075 on 2 and 2055 DF,  p-value: < 2.2e-16
```