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`.

```
# INSERT CODE HERE
```

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).

> Write answer here

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?

```
# Univariate visualization of predictor / X variable
movies %>%
  ggplot() + # Put the predictor on the x-axis
  geom_...() + # Choose the appropriate geom...()
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '')
```

```
## Error in movies %>% ggplot(): could not find function "%>%"
```

```
# Univariate visualization of outcome / Y variable
movies %>%
  ggplot() + # Put the outcome on the x-axis
  geom_...() + # Choose the appropriate geom...()
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '')
```

```
## Error in movies %>% ggplot(): could not find function "%>%"
```

```
# Multivariate visualization
movies %>%
  drop_na() %>% # Dropping missing observations from the variables
  mutate() %>% # Do you need to mutate anything?
  ggplot() + # Put the predictor on the x-axis and the outcome on the y-axis
  geom_...() + # Choose the appropriate geom...()
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '')
```

```
## Error in movies %>% drop_na() %>% mutate() %>% ggplot(): could not find function "%
>%"
```

Write answer here

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.

```
# Data wrangling
movies_analysis <- movies %>%
  mutate() %>% # Do you need to mutate anything?
  drop_na() # Dropping missing observations from the variables
```

```
## Error in movies %>% mutate() %>% drop_na(): could not find function "%>%"
```

```
model_gross_bechdel_score <- lm(formula = , # Write the regression formula here
                                data = ) # Put the data here
```

```
## Error in terms.formula(formula, data = data): argument is not a valid model
```

```
summary(model_gross_bechdel_score) # Interpret the regression output. Check here for hel
p: https://learneconomicsonline.com/blog/archives/1095
```

```
## Error in summary(model_gross_bechdel_score): object 'model_gross_bechdel_score' not f
ound
```

Write answer here

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() %>% # Create new variable of predicted values from the model
  mutate() # Calculate the errors
```

```
## Error in movies_analysis %>% mutate() %>% mutate(): could not find function "%>%"
```

```
# Univariate visualization of the errors
movies_analysis %>%
  ggplot() + # Put errors on the x-axis
  geom_...() + # Choose the apppropriate geom...()
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '')
```

```
## Error in movies_analysis %>% ggplot(): could not find function "%>%"
```

```
# Multivariate
movies_analysis %>%
  ggplot() + # Put errors on the y-axis and the predictor on the x-axis
  geom_...() + # Choose the apppropriate geom...()
  geom_hline() + # Add a horizontal dashed line at zero
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '')
```

```
## Error in movies_analysis %>% ggplot(): could not find function "%>%"
```

```
# RMSE Calculations
# RMSE Full Data
movies_analysis %>%
  mutate() %>% # Calculate the squared errors (SE)
  summarise() %>% # Calculate the mean of the squared errors (MSE)
  mutate() # Calculate the square root of the mean of the squared errors (RMSE)
```

```
## Error in movies_analysis %>% mutate() %>% summarise() %>% mutate(): could not find fu
nction "%>%"
```

```
# RMSE 100-fold CV
set.seed(123) # Set seed for consistency!
cvRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:100) { # Loop 100 times
  inds <- sample() # Create a list of random row numbers from 80% of the data, without r
eplacement

  train <- movies_analysis %>% slice() # Create the training dataset based on these rows
  test <- movies_analysis %>% slice() # Create the test dataset based on all other rows

  m <- lm(formula = ,
          data = ) # Calculate your regression on the training data

  test$preds <- predict(,
                        newdata = ) # Apply your model to the test data

  e <-  # Calculate the errors
  se <-  # Calculate the squared errors
  mse <- mean() # Calcualte the mean of the squared errors
  rmse <- sqrt()# Calculate the RMSE
  cvRes <- c() # Save the RMSE to a list of numbers
}
```

```
## Error in sample(): argument "x" is missing, with no default
```

```
mean(cvRes)
```

```
## Warning in mean.default(cvRes): argument is not numeric or logical: returning NA
```

```
## [1] NA
```

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

```
## [1] NaN
```

Write answer here

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?

Write answer here

```
# INSERT CODE HERE
```