QRM II Graded Assignment (5), Period 1 2025

Material by Sjoerd van Alten and Klervie Toczé

Assignment Group 8, Catherina Mikhail, Mayte Leegwater, Merel Vonk, Yasmina Andaou

05-09-2025

0 Introduction

This assignment is to be completed in groups of 3-4. Further, all students in your group need to be assigned to the same R tutorial group (Friday's tutorial). You can sign yourself up for a group on Canvas. Please do so **before the start of your first R tutorial on Friday September 5th.** You can use the Discussion Board in Canvas if you do not have a group yet or if your group is incomplete.

The assignment has 5 parts, and each part corresponds to the course material of that week (with the exclusion of week 6, for which there is no R programming material).

You are supposed to hand in these assignments on Canvas at the following dates:

- **Deadline 1** *Thursday September 25th, at 23:59pm*: you are supposed to hand in weeks 1, 2, and 3 of this assignment. This will determine 18% of your overall course grade
- **Deadline 2** *Thursday October 9th, at 23:59pm*: you are supposed to hand in weeks 4, and 5 of this assignment. This will determine 12% of your overall course grade

The R tutorials (each Friday) will consist of two halves. During the first half, you will discuss the tutorial exercises. These can be downloaded separately from Canvas. During the second half, you can work on this graded assignment within your own group. The purpose is that you find out how to work with R for doing statistical analyses by yourself. The tutorial exercises are meant to teach you basic commands to get you started, but to answer the problem sets in this assignment, you might need to research your own solutions, and use functions and commands not described in the tutorial exercises. Learning how to solve your own research problems is integral part of learning R. When you and your group get stuck on how to approach an exercise, the hierarchy in finding your way is as follows:

- use the concepts from the tutorial exercises;
- use the cheat sheets available on Canvas;
- use Google, YouTube, StackOverflow, or another website;
- ask the teacher.

The use of generative AI is **not** permitted and may result in a grade of 0. See the AI protocol in the course manual for details.

To answer the assignment, you can simply fill out this R markdown document. There are designated places which you can fill with R code. There are also designated spaces for you to answer each question. Often, the structure of an answer will be as follows. First, you type the R code in the designated box. This will show how you analyzed the data to get the answer to the question. Below the box for the R code, you will then summarize your answer to the question, i.e. what are the conclusions that you draw from the data analysis?

When handing in, you are supposed to submit this .Rmd file, and a knitted version of this document. You can knit this document to pdf, word, or html. Knitting to pdf requires you to have a .tex distribution installed on your computer. Knitting to Word requires you to have Word installed.

The exercises are designed such that you should be able to finish the majority of them during the tutorial each week. If you are not able to finish them fully during that time, you are expected to work on it in your own time using the computers on campus or your own device. It is best to meet as a group in-person when working together. If you want to work remotely, github is a good platform to guarantee smooth collaboration. Alternatively, you can email this .Rmd file back and forth to one another as a group, but this is not recommended as it is more cumbersome.

We encourage you to keep your code blocks, printing statements, and final answers, as short as possible. In any case, there is a page limit of 6 pages per week, which encompasses the total length of this document which consists of the questions, your coding lines, and your answers. When your answers to questions of the respective week exceed this page limit, they will not be graded, resulting in zero points.

Each week consists of 1, 2, or 3 subquestions. The total amount of points you can earn per week is 20 points.

1 Week 1

1. Find the dataset "movies2.tsv" on Canvas. Describe your data set: How many observations does it have. How many variables are there? How many subjects? What consists of a subject? [**4 points**]

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.5.1
```

```
movies2 = read_tsv("movies2.tsv")
```

```
## Rows: 606 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr   (8): keywords, original_language, title, genre, first_actor, first_act...
## dbl  (10): index, budget, popularity, revenue, runtime, vote_average, vote_c...
## date  (1): release_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Here I am computing the number of observations, variables and subjects
ncol(movies2)
```

```
## [1] 19
```

```
nrow(movies2)
```

```
## [1] 606
```

Your Answer:

This dataset has 606 observations. There are 19 variables and 606 subjects. A subject in a dataset is the unit of analysis that each row represents, in this case for example each row is gives all the information about 1 movie with variables such as the revenue, runtime, average voting and the other 16 variables.

2. Which of the following types of variables are present in your data set? (i) nominal; (ii) ordinal; (iii); interval; (iv) ratio. If present, name one example of such a variable present in your data set. **[4 points]**

```
#Here I am looking at the head of the dataset to determine whether the variables are
↪  nominal, ordinal, interval or ratio
head(movies2)
```

```
## # A tibble: 6 x 19
##   index  budget keywords original_language title popularity release_date revenue
##   <dbl>   <dbl> <chr>    <chr>             <chr>      <dbl> <date>         <dbl>
## 1  3793   4  e6 <NA>     en                The ~      0.409 1999-07-16    0
## 2  3853   3.5e6 thrille~ en                2:13       1.27  2009-04-25    0
## 3  2476   0     poison ~ en                Whit~      9.42  2002-10-11    0
## 4   491   1.3e8 gladiat~ en                Pomp~     50.6   2014-02-18    1.18e8
## 5   540   7.5e7 rap mus~ en                Holl~     10.6   2003-06-09    5.11e7
## 6  3238   8  e6 califor~ en                Litt~     14.8   2006-07-26    1.01e8
## # i 11 more variables: runtime <dbl>, vote_average <dbl>, vote_count <dbl>,
## #   genre <chr>, release_year <dbl>, release_month <dbl>, release_day <dbl>,
## #   first_actor <chr>, first_actor_gender <chr>, director_first_name <chr>,
## #   director_gender <chr>
```

Your Answer:

The variable Keywords is nominal. The variable Vote Average is ordinal. The variable Budget is ratio. The variable release_year is interval.

3. A movie studio wants to know which types of movies give maximal profit. Perform the following steps to provide the movie studio with an analysis which corresponds to their request:

a. Create the variable profits as the revenue of a movie minus its budget. Report its mean, median, maximum, and minimum. **[2 points]**
b. Which movie has the highest profits in your data set and how much are these profits. Which movie has the lowest and how much are its profits? If multiple movies have the exact same highest or lowest profits, give only one example. **[2 points]**
c. Create a boxplot of the variable profits. Make sure it has an appropriate title, and appropriate titles and labels for the x- and y-axis. Give Q1, Q2, Q3, and Q4. What does this tell you about the nature of making money in the movies industry? **[2 points]**
d. Add a new variable to your data set the log of profits. When creating this variable, what happens to movies for which profits is zero or negative? What then happens when you calculate the mean of log of profits? **[2 points]**
e. For movies that have a profit of zero or less, replace log of profits with "NA". What is now the mean of log of profits? Create a boxplot for log of profits, again with an appropriate title, x- and y-axis labels. How does it compare to the boxplot you made under c.)? **[2 points]**
f. Create a scatterplot of with the runtime of movies on the x-axis and the average vote of movies on the y-axis. What do you conclude from the scatterplot? Are movies with a longer runtime considered worse or better by the audience, or does the audience not have a preference? Why do you think this is the case? **[2 points]**

For each step, you should provide first all the code you used to answer the question and then formulate an answer using full sentences.

Step a

```r
#Here I am making a new variable called profits
library(tidyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
movies2 = movies2 %>%
  mutate(profits = revenue -budget)
#Here I am computing the minimum, maximum and mean values of the variable profits
summary(movies2$profits)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -111007242   -1498570          0   55265374   59412351 1316249360
```

Your Answer:

The mean of profits is 55265374 dollars. The median of profits is 0. The minimum of profits is -111007242. The maximum of profits is 1316249360.

Step b

```r
#Here I am making a seperate dataset with 1 variable and 1 observation to see which movie
↪  has the highest profits and how much it is.
max_profit_movie = movies2[which.max(movies2$profits), "title"]
print(max_profit_movie)
```

```
## # A tibble: 1 x 1
##   title
##   <chr>
## 1 Furious 7
```

```r
max_profit = max(movies2$profits)

#Here I am doing the same thing as above but then for the lowest profits
min_profit_movie = movies2[which.min(movies2$profits), "title"]
print(min_profit_movie)
```

```
## # A tibble: 1 x 1
##   title
##   <chr>
## 1 Mars Needs Moms
```

```r
min_profit = min(movies2$profits)
```

Your Answer:

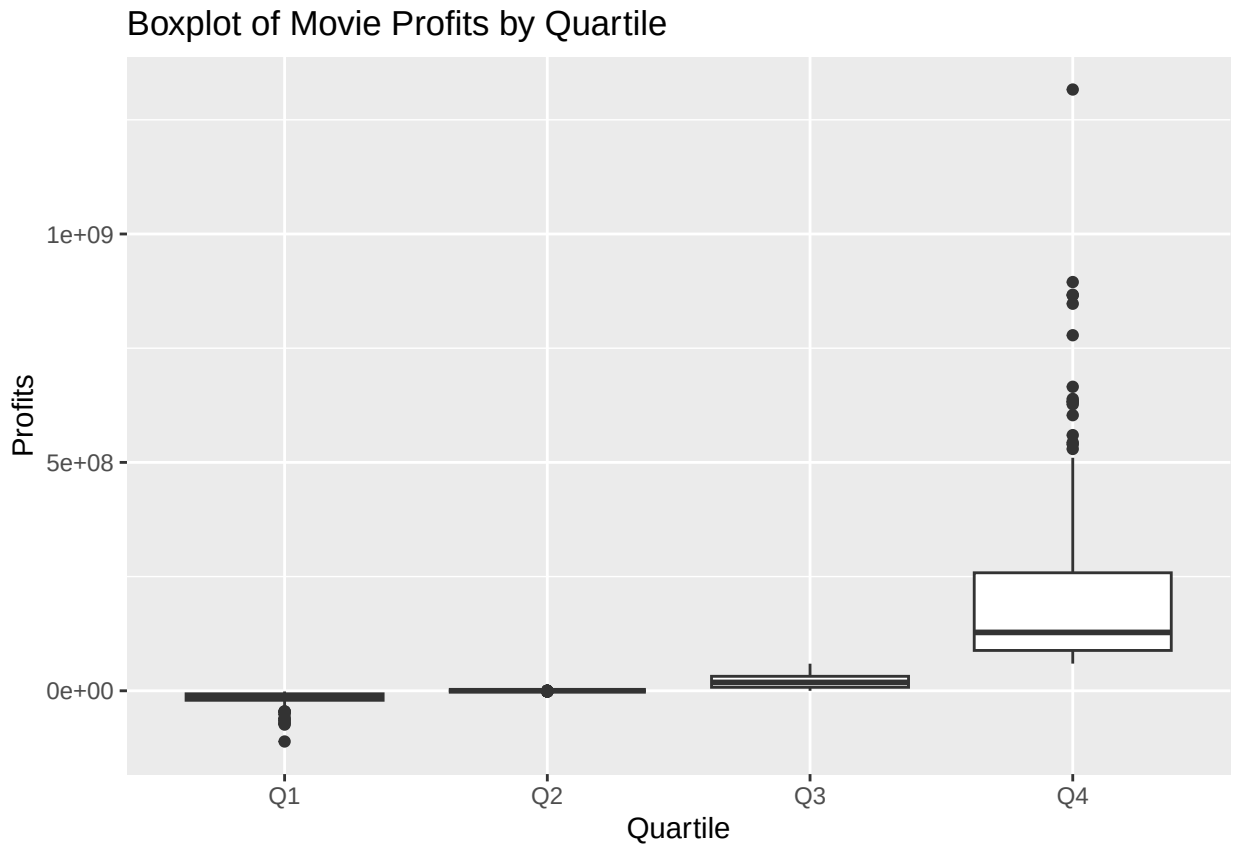
The movie with the highest profit of 1316249360 dollars is Furious 7. The movie with the lowest profit is Mars Needs Moms and has a negative profit of minus 111007242 dollars.

Step c

```r
#Here I am dividing the profits into quartiles as I have to plot a boxplot
movies2$quartile = cut(movies2$profits, breaks = quantile(movies2$profits, probs = c(0,
↪  0.25,0.5, 0.75, 1)), labels = c("Q1", "Q2", "Q3", "Q4"), include.lowest = T)

#Here I am plotting a boxplot for the profits with quartiles
library(ggplot2)

ggplot(movies2, aes(x = quartile, y = profits)) +
  geom_boxplot() +
  labs(title = "Boxplot of Movie Profits by Quartile", x = "Quartile", y = "Profits")
```



Your Answer:

The boxplot shows that most movies make very little profit, with many barely breaking even in Q2 or even losing money in Q1. Only a small share of films in the top quartile in Q4 generate significant profits, and a handful of blockbuster hits drive the industry's earnings. Overall, the movie industry is high-risk, high-reward, relying heavily on rare mega-hits to stay profitable.

Step d

```
#Here I make a new variable with the log transformed profits
movies2 = movies2 %>%
  mutate(log_profits = log(profits))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `log_profits = log(profits)`.
## Caused by warning in `log()`:
## ! NaNs produced
```

```
#Here I compute the mean but we were not sure to remove the NAs or not so in the first
↪  one we did not remove it and in the second we did to see if we would get different
↪  answers
mean(movies2$log_profits)
```

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

```
mean(movies2$log_profits, na.rm = T)
```

```
## [1] -Inf
```

Your Answer:

When you do not remove the NAs, while computing the mean of the log transformed profits you get an NaN, so a not a number error. However, if you do remove the NAs while computing the mean of the log transformed profits you get a minus Inf.

Step e

```
#Here I put NA instead of the profits that are negative or equal to 0, then I used the
↪  same code as before to calculate the logs and then I calculated the mean of the logs
↪  of the profits again.
movies2$profits = ifelse(movies2$profits <= 0 , NA, movies2$profits)
movies2 = movies2 %>%
  mutate(log_profits = log(profits))
mean(movies2$log_profits, na.rm = T)
```

```
## [1] 17.62456
```

```
#As I have to make a boxplot, I first divided the logs into quartiles
movies2$quartiles = cut(movies2$log_profits, breaks = quantile(movies2$log_profits, probs
↪  = c(0, 0.25,0.5, 0.75, 1), na.rm = T), labels = c("Q1", "Q2", "Q3", "Q4"),
↪  include.lowest = T)

# Here I am plotting the boxplot with the log transformed profits with quartiles
ggplot(movies2, aes(x = quartiles, y = log_profits)) +
  geom_boxplot() +
  labs(title = "Boxplot of Movies Log Transformed Profits", x = "Quartiles", y = "Log of
  ↪  Profits")
```

```
## Warning: Removed 310 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

Your Answer:

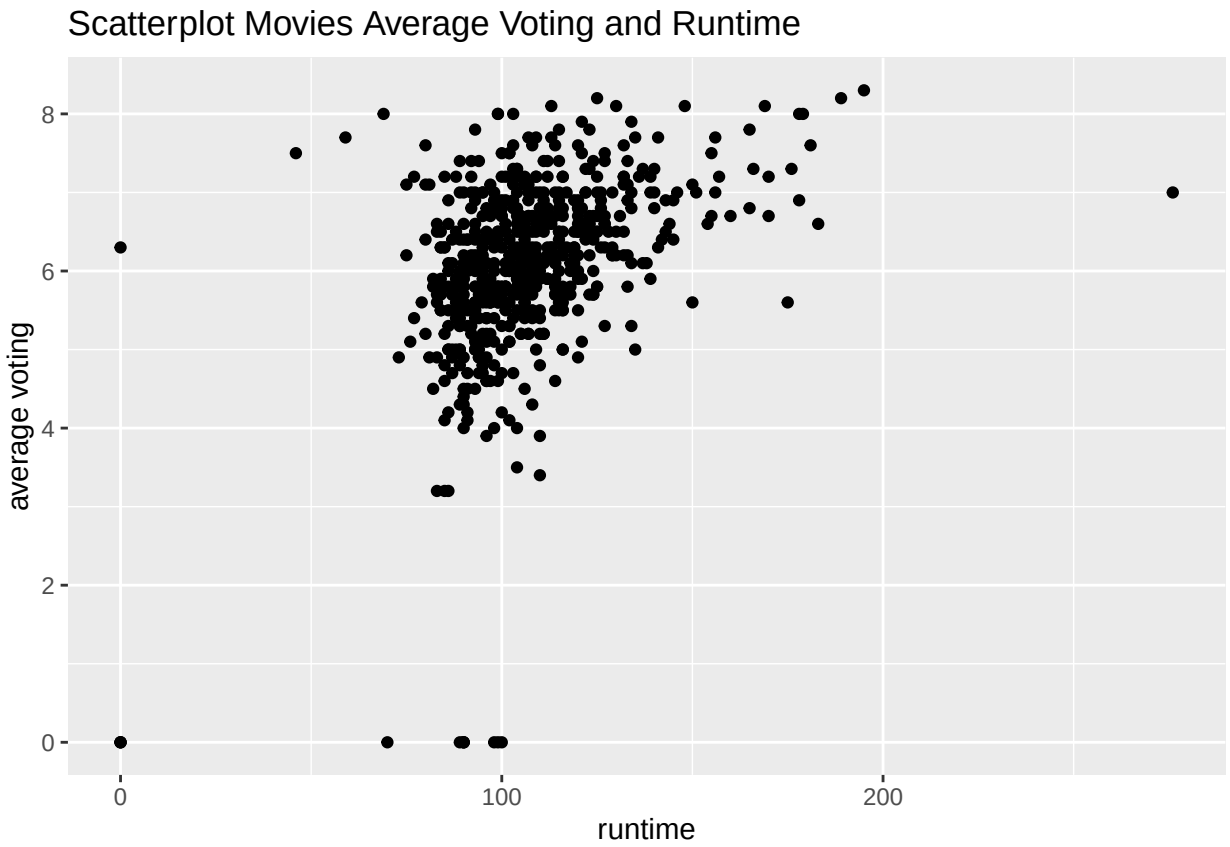
The mean of the log of the profits is now 17.62456.

The first boxplot of the profits before being log-transformed shows that movie profits are extremely skewed, with most films making very little and a few blockbusters dominating the scale. This highlights the "winner-takes-most" nature of the industry but makes it hard to see differences among the majority of movies. The second boxplot, so the log-transformed profits, reduces the effect of extreme outliers, spreading the data more evenly across quartiles. This makes the relative differences between groups clearer and shows a smoother upward trend in profits from Q1 to Q4. Together, they reveal both the extreme inequality of raw profits and the more balanced progression when viewed on a log scale.

Step f

```
#Here I plot a scatterplot with runtime on x-axis and on the y-axis the vote average

ggplot(movies2, aes(x = runtime, y = vote_average)) +
  geom_point() +
  labs(title = "Scatterplot Movies Average Voting and Runtime", x = "runtime", y =
  ↪  "average voting")
```



Your Answer:

The scatterplot shows a weak positive relationship between runtime and ratings: longer movies tend to score slightly higher, though the effect is small. Most films cluster around 90–120 minutes with average ratings, while very long films (150+ minutes) often maintain strong scores around 7–8. This may reflect that only higher-quality projects get extended runtimes, supported by bigger budgets and stronger storytelling. Short films under 60 minutes generally perform poorly. Overall, runtime has some influence, but we think that audience satisfaction depends far more on quality than length.

2 Week 2

1 Is your dataset movies2.tsv the full population, or is it a sample of a larger population? If the latter, how would you describe the full population? [**4 points**]

Your Answer:

Write your formulated response here.

2

a. For which actor in your data set do you observe the most movies? [**2 points**]
b. What is the average revenue of the movie in which this actor plays and does the revenue lie above or below the revenue of an average movie according to your data set? [**2 points**]

c. How trustworthy do you consider your conclusion to answer 2b? Use the term "law of large numbers" in your explanation. [**2 points**]

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

3 For this question, you will assume that your data set is the full population.

a. Recode profits such that it is expressed in millions. What is the variance of the variable profits (in millions) in your data set? **[2 points]**
b. Create a new data set, called movies_sample. Make sure that it is a random sample of your data set of 25 movies. What is the variance of profits in this random sample? How does it compare to the variance of profits in 2a? **[2 points]**
c. In a for loop, create 100 different samples of 25 movies, as in b, and estimate the variance within each sample. Save the variance of each sample in a vector called sample_vars. So the first position of the vector would have the variance of the first sample, the second position the variance of the second sample, etc. Print the start of this vector. **[2 points]**
d. Summarize and make a histogram of sample_vars. What is the mean, standard deviation and shape of its distribution? **[2 points]**
e. In your opinion, is a sample of 25 movies sufficient to get a reliable estimate of the population variance of profits, using the sample variance? Explain? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step d

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step e

Your Answer:

Write your formulated response here.

Your answer here

3 Week 3

For the next part of the assignment, assume that the movies in your data frame are a random sample of a larger population of movies.

1

a. Create a new data set that only includes movies that are of the genre "Thriller". For these thriller movies, give a 99 percent confidence interval for the variable *runtime*. Interpret the result. **[2 points]**
b. Now, assume that the variance of *runtime* amongst thriller movies in your data is exactly the same as the variance of *runtime* in the population. Under this assumption, give a 99 percent confidence interval for the variable *runtime* among thriller movies. Interpret the result. Is you confidence interval wider or less wide than the one you found under question 1a? Why is that the case? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

2

a. Using an appropriate five-step procedure, set up a test for the null hypothesis that the variance of runtime equals 500. Clearly state your null hypothesis, alternative hypothesis your test statistic, your critical value, and your conclusion. [**2 points**]
b. For the validity of your test in 2a, what assumption about the distribution of revenue needs to hold? Make an appropriate plot to test this assumption. What do you conclude? [**2 points**]

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

3. There is an argument going on in the movie studio. *Bob* claims that they should make higher-quality movies, as this will bring in more profits. *Chantal* disagrees. She tells Bob that mediocre movies bring in the most profits. You are asked to advise on who is right.

a. Create a new variable called vote_average_rounded. Make sure this variable is the same as vote_average, but without any decimals (i.e., a 6.3 becomes a 6, a 8.7 an 8, etc.). Display a histogram of vote_average_rounded. [**2 points**]
b. Create a scatter plot with vote_average_rounded on the x axis and the mean of profits within each category of vote_average_rounded on the y-axis. Make sure it has an appropriate title, and appropriate titles and labels for the x- and y-axis. At which rating of movies are profits the highest? [**3 points**]
c. Recreate the scatter plot with year on the x axis and mean_profits on the y-axis, but now add bars around each point, indicating the 95% confidence interval. [**3 points**]
d. Write an advice to settle the argument between Bob and Chantal. [**4 points**]

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step d

Your Answer:

Write your formulated response here.

4 Week 4

1. There is another argument going on in the movie studio. *Bob* claims that production budgets are getting out of hand, and that the studio should focus on making cheaper movies. *Chantal* disagrees. She tells Bob that "Every dollar we spend on movie production is more than offset by the increase in movie profits''.

   a. Set up a regression model to test Chantal's claim, and estimate it. That is, estimate:

   $$\text{Profits}_i = \beta_0 + \beta_1 \text{Budget}_i + \varepsilon_i.$$

   Print a summary of your estimated model. **[2 points]**
   b. What is the estimated value of β_1 and how do you interpet it? **[2 points]**
   c. Test for the null hypothesis that $\beta_1 \geq 0$. Report the p-value and state your conclusion. **[2 points]**
   d. Next, estimate the model

   $$\text{Log Profits}_i = \beta_0 + \beta_1 \text{Log Budget}_i + \varepsilon_i.$$

   When creating the variables Log Profits and Log Budget, make sure that movies with a Revenue or Budget of zero are assigned the value "NA". Print a summary of your estimated model **[2 points]**
   e. What is the estimated value of β_1 and how do you interpet it? **[2 points]**
   f. Which model has better fit? The level-level model or the log-log model? Explain. **[2 points]**
   g. Who do you think is correct? Bob or Chantal? What would you advise the movie studio to do? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step d

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step e

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step f

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step g

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

2

a. Make a plot with a 95% confidence interval with the mean log of budget on the y-axis, and whether the first actor of the movie is male or female on the x-axis. What do you conclude? **[2 points]**
b. Estimate the following simple OLS model: $log(budget)_i = \beta_0 + \beta_1(FirstActorMale)_i + \varepsilon_i$. Is the estimated coefficient for β_1 significantly different from zero? How do you interpret its estimate, and how does this relate to your conclusion in 2a? **[2 points]**
c. Now, have a close look at your data frame. Can you find any instances of male first actors who are wrongly labeled as being female, or vice versa? What would such mislabelling mean for the coefficient you estimated under 2b? **[2 points]**

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

5 Week 5

a. Create a plot of the mean profits by month of release. Do you see any indication that month of release matters to the profits of the movie? [**2 points**]

b. Estimate an OLS model which has as dependent variable the log of profits of a movie, and as independent variable the log of budget, a dummy for whether the movie was released in english or not, and a linear term for the month of release. Show a summary of the resulting model and interpret each coefficient. [**4 points**]

c. Test for the hypothesis that the coefficient that belongs to month of release is zero. [**2 points**]

d. Based on your plot in a.) do you consider the choice that month of release enters the model linearly under b.) reasonable? Estimate a specification that allows for a more flexible curve. In this new specification, test for the null hypothesis that month of release does not impact profits. This might require testing multiple terms at once. [**4 points**]

e. One executive at the studio wants to time the release of the movie to a specific month of the year such that they can maximize revenue. Based on your model under d.), What would you advise the movie studio regarding the timing of the release of the movie? [**2 points**]

The movie studio that you work at is releasing a new movie in 2026. It will be an English-spoken Thriller movie with a budget of 40,000,0000.

f. Estimate a model that is able to predict the revenue of this movie. Give its predicted revenue and include a 99% prediction interval. [**6 points**]

step a

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step b

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step c

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step d

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step e

```
#WRITE YOUR CODE HERE
```

Your Answer:

Write your formulated response here.

step f

```
#WRITE YOUR CODE HERE
```