

# Homework 2

BSTA 512/612

Your name here - update this!!!!

2024-02-01

! Important

**This page is no longer under construction. You may start this homework!  
(1/25/24)**

## Directions

- [Download the .qmd file here.](#)
- You will need to download the datasets. Use [this link to download](#) the homework datasets needed in this assignment. If you do not want to make changes to the paths set in this document, then make sure the files are stored in a folder named “data” that is housed in the same location as this homework .qmd file.
- Please upload your homework to [Sakai](#). **Upload both your .qmd code file and the rendered .html file**
- For each question, make sure to include all code and resulting output in the html file to support your answers.
- Show the work of your calculations using R code within a code chunk. Make sure that both your code and output are visible in the rendered html file. This is the default setting.
- If you are computing something by hand, you may take a picture of your work and insert the image in this file. You may also use LaTeX to write it inline.
- Write all answers in complete sentences as if communicating the results to a collaborator. This means including a sentence summarizing results in the context of the research study.

### 💡 Tip

It is a good idea to try rendering your document from time to time as you go along! Note that rendering automatically saves your qmd file and rendering frequently helps you catch your errors more quickly.

## Question 1

This homework assignment is based on data collected as part of an observational study of patients who suffered from stroke.

Dataset: The main goal was to study various psychological factors: optimism, fatalism, depression, spirituality, and their relationship with stroke severity and other health outcomes among the study participants. Data were collected using questionnaires during a baseline interview and also medical chart review. More information about this study can be found in the article [Fatalism, optimism, spirituality, depressive symptoms and stroke outcome: a population based analysis](#).

The dataset that you will work with is called `completedata.sas7bdat`. It is SIMILAR but does not exactly match the data in the article. It contains information on complete cases (i.e. excludes participants who had missing data on one or more variables of interest) who suffered a stroke. The two variables we are interested in are:

- Covariate: **Fatalism** (larger values indicate that the individual feels less control of their life)
  - Scores range from 8 to 40
- Outcome: **Depression** (larger values imply increased depression)
  - Scores range from 0 to 27

For our homework purposes we will assume they are continuous.

```
fatal_dep = read_sas(here("./homework/data/completedata.sas7bdat"))
```

### Part a

Plot the data, with title and axis labels, for Depression (y-axis) vs. Fatalism (x-axis). Comment on what you see.

### Part b

Fit a linear regression model to estimate the association between the predictor Fatalism and the outcome Depression.

Interpret the slope and intercept. Does the intercept make sense?

#### **i** Note

Make sure to include the confidence interval. Whenever asked to interpret coefficients, you must include confidence intervals. Also, the “units” for fatalism and depression are scores.

### Part c

In your dataset, make a new variable FatalismC, equal to Fatalism centered at its median (C is for centered).

$$\text{FatalismC} = \text{Fatalism} - \text{median of Fatalism}$$

This is one way of centering a variable, and can be used when the intercept estimate does not make sense. (Hint: the `mutate()` function will work well here!)

Plot the data, with title and axis labels, for Depression (y-axis) vs. FatalismC (x-axis).

### Part d

Re-run the regression from Part b using this new variable for FatalismC. Interpret the new slope and intercept. Which of the following are the same as Part b: intercept, slope?

#### **i** Note

Make sure to include the confidence interval. Whenever asked to interpret coefficients, you must include confidence intervals. Also, the “units” for the centered fatalism is still the score.

### Part e

From the above interpretations, what would be the equivalent conclusion from a hypothesis test for the association between Depression and Fatalism?

**i** Note

You do not need to go through the whole process for the hypothesis test. You only need to state whether it is rejected or not and site the confidence interval as evidence.

## Question 2

This question and data are adapted from [this textbook](#).

In an experiment designed to describe the dose–response curve for vitamin K, individual rats were depleted of their vitamin K reserves and then fed dried liver for 4 days at different dosage levels. The response of each rat was measured as the concentration of a clotting agent needed to clot a sample of its blood in 3 minutes. The results of the experiment on 12 rats are given in the following table; values are expressed in common logarithms for both dose and response.

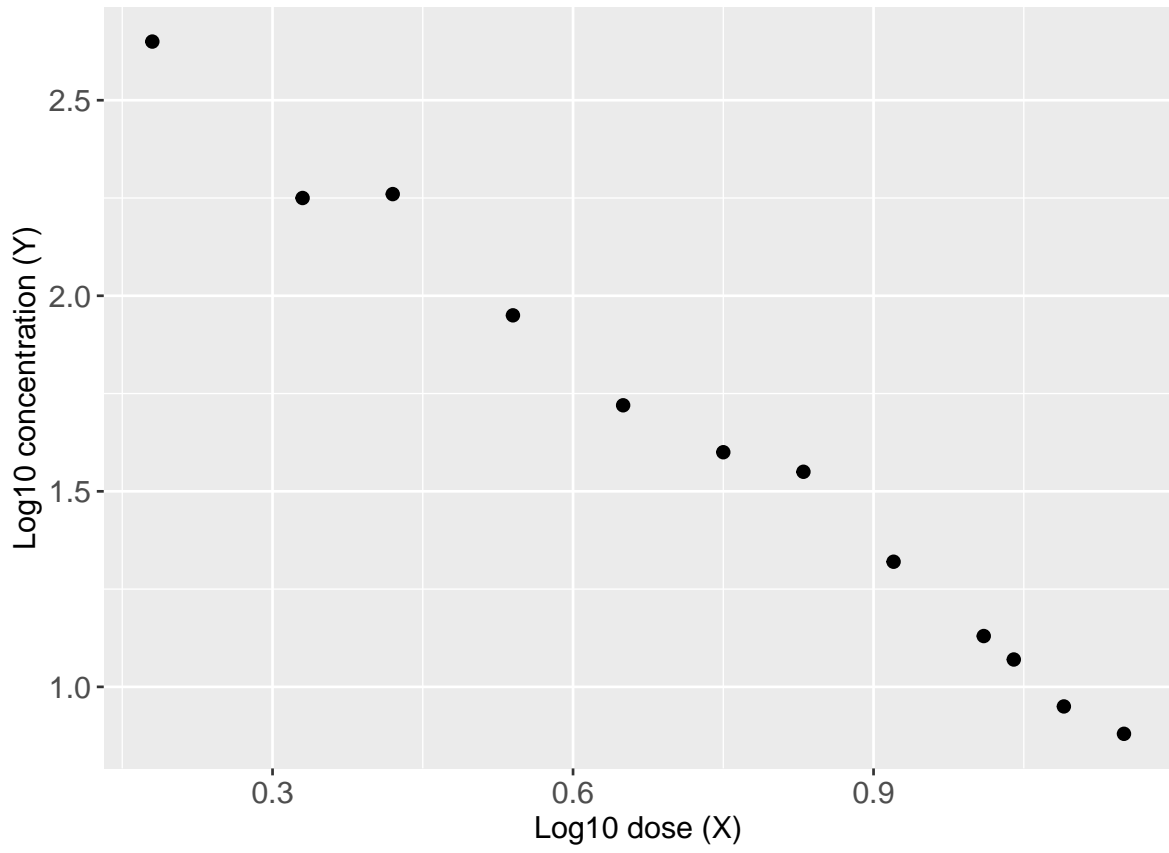
```
clot = read_excel(here("./homework/data/CH05Q09.xls"))
clot %>% gt() %>%
  cols_label(RAT = md("**Rat**"),
             LOGCONC = md("**Log10 Concentration (Y)**"),
             LOGDOSE = md("**Log10 Dose (X)**"))
```

| Rat | Log10 Concentration (Y) | Log10 Dose (X) |
|-----|-------------------------|----------------|
| 1   | 2.65                    | 0.18           |
| 2   | 2.25                    | 0.33           |
| 3   | 2.26                    | 0.42           |
| 4   | 1.95                    | 0.54           |
| 5   | 1.72                    | 0.65           |
| 6   | 1.60                    | 0.75           |
| 7   | 1.55                    | 0.83           |
| 8   | 1.32                    | 0.92           |
| 9   | 1.13                    | 1.01           |
| 10  | 1.07                    | 1.04           |
| 11  | 0.95                    | 1.09           |
| 12  | 0.88                    | 1.15           |

Use the log-transformed values as given in the dataset.

Use the following scatterplot to build your answers off of:

Concentration of clotting agent needed vs. dosage levels of liver



### Part a

Fit a linear regression model to the data and add the regression line to the plot.

### Part b

Use R to create the ANOVA table for the regression described in the exercise.

### Part c

Using the F-test, determine whether there is an association between the log<sub>10</sub> concentration and log<sub>10</sub> dose.

#### Note

Make sure to include all needed steps for an F-test. Calculating the F test statistic (step 5) is not needed if you use the ANOVA table. Make sure your conclusion connects back the research context.

### Part d

Rewrite your hypothesis test in Part c to show the null and alternative models that we are testing. Did we reject the smaller (reduced) model?

#### Note

You do not need to go through the hypothesis test process again. A quick statement on rejecting or not is okay.

#### Tip

If you prefer to write out the models by hand, remember that you can take a picture of your work and insert it into this document. HW0 can be a good reference for how we've done this before.

## Question 3

We will continue to work with the study and dataset from Question 2 above.

**Part a**

Find the correlation coefficient between the two variables. Is the value consistent with your description of the relationship in Question 2? Why or why not?

**Part b**

Calculate the coefficient of determination using linear regression summary output. Can we also calculate the coefficient of determination from the ANOVA in Question 2?

**Part c**

Give an interpretation of the coefficient of determination in the context of the study.