Homework 2

BSTA 512/612

2024-02-01

❗ Important

THIS PAGE IS UNDER CONSTRUCTION!! It's likely that I will be making changes to this assignment at this time!

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.
  - Questions **not requiring a sentence are**

        - Ch 7 # 1, 2, 5
        - Ch 6 # 5, 6
        - Ch 14 # 2, 12, 14

> 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.

```r
library(haven)
here()
```

```
[1] "/Users/wakim/Library/CloudStorage/OneDrive-OregonHealth&ScienceUniversity/Teaching/Clas
```

```
#fatal_dep = read_sas("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. The “units” for fatalism and depression are scores.

Part c

Obtain the interquartile range (IQR) for fatalism (IQR is Q3-Q1). Make a new variable FatalismIQR, equal to Fatalism centered at its median, and divided by the IQR. This is one way of standardizing a variable, and can be used for variables that have symmetric or skewed distribution.

Re-run the regression from Part b using this new variable. Interpret the new slope and intercept. Which of the following are the same as in 1b: intercept, slope?

Part d

F-test???

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

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

(1)

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

(2)

Using the values in the `Df` and `Sum Sq` columns, show how the remaining values in the table are calculated.

(3)

Calculate the SSY and its degrees of freedom. Use these values to calculate the standard deviation of the outcome variable.

(4)

What are the hypotheses being tested in the ANOVA table? Make sure to include a description of the parameter being tested in the context of the research question.

(5)

Using just the values in the ANOVA table, find the value of the t-distribution test statistic for testing the slope that one would find in the regression output of the linear model. What are the degrees of freedom for that t-distribution?

(6)

Show the regression output for the linear model. Using just the regression output, calculate the SSR and SSE values in the ANOVA table.

Question 1 (chapter 6)

Use the data from Chapter 5 Question 9 to answer the following questions. Use the log-transformed values as given in the dataset.

Note: the question numbers below do not refer to questions from the textbook. Complete the problems below instead of the ones in the book.

(1)

Create a scatterplot of the dependent and independent variables, and in words describe the their relationship. Is it reasonable to use a linear regression to model the relationship?

(2)

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

(3)

Test whether the two variables are significantly correlated. Do this using the formula and then check your work with R's test for correlations. Make sure to include the hypotheses and a conclusion.

(4)

Calculate the confidence interval for ρ using the formula and verify that it matches the confidence interval in R's test output. Include an interpretation of the confidence interval and also explain why the confidence interval is consistent with the p-value.

(5)

Calculate the coefficient of determination using the ANOVA table output, and confirm that it matches the value in the R output (what R output shows this and what is it labeled as?).

(6)

What is another way to calculate the coefficient of determination? Do the calculation and verify that you have the same answer.

(7)

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

Note: the question numbers below do not refer to questions from the textbook. Complete the problems below instead of the ones in the book.