Homework 3

BSTA 512/612

Your name here!!!

2024-02-09

> **!** Important
>
> **THIS PAGE IS UNDER CONSTRUCTION!!**

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**
  - **Please rename you homework as `Lastname_Firstinitial_HW0.qmd`. This will help organize the homeworks when the TAs grade them.**
  - **Please also add the following line under `subtitle: "BSTA 512/612"`: `author: First-name Last-name` with your first and last name so it is attached to the viewable document.**
- 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.

Questions

Question 1

A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is "high" however, a physician must have a clear picture of the distribution of normal rates. To this end, Italian researchers measured the respiratory rates (in breaths/minute) of 618 children between the ages of 15 days and 3 years (measured in months).

The data and problem framing came from the `Sleuth3` package. Please make sure to run the following code to load the data. You can directly access the dataset `ex0824` from the package. I have included a new assignment of the data to `q1_data` if you would like to use that.

```
if(!require(Sleuth3)) { install.packages("Sleuth3"); library(Sleuth3) }
q1_data = ex0824
```

Part a

Create a scatterplot of the dependent and independent variables with both the best-fit line and a smoothed curve through the points. Describe the relationship between the dependent and independent variables, and also comment on whether you think it is reasonable to use a linear regression to model the relationship.

Part b

Write out the population regression model for the simple linear regression model. Please leave the variables untransformed for now.

Part c

Fit the regression model, display the regression table, and write out the fitted regression line.

Part d

Assess the normality of the model's fitted residuals by creating a histogram, density plot, and boxplot of the residuals to visually inspect the distribution of the residuals, and describe any deviations from normality.

Part e

Assess the normality of the model's fitted residuals by creating QQ plot of the residuals.

Bonus work, but not required: Compare the QQ plot to 4 such plots simulated from normal data, and discuss why or why not the residuals could have come from a normal distribution.

Part f

Test the normality of the model's fitted residuals and comment on whether the test's conclusion is consistent with your visual inspection or not. Make sure to include the hypotheses, needed R code, and a conclusion to the test based on the p-value (as shown in these slides).

Part g

Create a residual plot using ggplot and the residuals. Discuss what this means in the context of our model assumptions.

Part h

Determine whether there are any observations with high leverage. Please use the cutoff, $h_i > 6/n$. If there are observations with high leverage, print the observations and state how many high leverage points there are.

Part i

Print the 10 observations with highest Cook's distance. If there are observations with high Cook's distance ($d_i > 1$), state how many observations have high Cook's distance.

Part j

Create a histogram for rate. Describe its distribution shapes.

Part k

Using the above histogram, and Tukey's ladder of transformations, discuss the range of transformations that will be appropriate for **Rate**. Explain your reasoning.

Then use `gladder()` to decide on two possible transformations. Explain your reasoning.

Note: questions below will ask about model fit with the transformations. For now, just explain why you chose the two that you did.

Part l

Add the two rate transformations you chose above to the dataset. You do not need to print any output, just make sure the code is visible.

Part m

Create scatterplots using two transformed rates and age. Discuss if either transformation potentially improves the model fit. Explain why or why not. Note: including lines will help!

Part n

Using one of the transoformed outcomes, fit the regression model, display the regression table, and write out the fitted regression line.

Part o

Assess the normality of the model's fitted residuals by creating QQ plot of the residuals. Does the transformation improve the QQ plot?

Part p

Create a residual plot using ggplot and the residuals. Discuss what this means in the context of our model assumptions. Does the transformation improve our model assumptions?

Part q

Between the model with the untransformed outcome and the transformed outcome, which would you recommend using for analysis? (Hint: there are pros and cons to both models)

Question 2

This question uses the same dataset as HW 2, question 1.

This question 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`. The two variables we are interested in are:

- Covariate 1: `Fatalism` (larger values indicate that the individual feels less control of their life)
  - Potential scores range from 8 to 40
- Covariate 2: `Optimism` (larger values indicate that the individual feels higher levels of optimism)
  - Potential scores range from 6 to 24
- Covariate 3: `Spirituality` (larger values indicate that the individual has more belief in a higher power)
  - Potential scores range from 2 to 8
- Outcome: `Depression` (larger values imply increased depression)
  - Potential scores range from 0 to 27

For our homework purposes we will assume each variable is continuous.

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

Part a

Fit the regression model with all the covariates (Fatalism, Optimism, Spirituality), display the regression table, and write out the fitted regression line.

Part b

Interpret each coefficient (β_0, β_1, β_2, β_3).

Does the intercept make sense for the range of values that each covariate can take? Explain.

Part c

Recall in Homework 2, we ran a simple linear regression model for Depression vs. Fatalism with the following interpretation for the coefficient: For every 1 point higher fatalism score, there is an expected difference of 0.25 points higher depression score (95%CI: 0.17, 0.32).

Does the addition of Optimism and Spirituality change our coefficient estimate for Fatalism? (No need for an official hypothesis test here. I just want us to note some differences.)

Part c

From the fitted regression model, calculate the regression line when Optimism score is 10 and Spirituality score is 6.

Part d

Does at least one of the covariates contribute significantly to the prediction of Depression? (Note: this is an overall test. Please follow the hypothesis test steps. To complete step 4-6, simply output your ANOVA table.)

Part e

Does the addition of Spirituality add significantly to the prediction of Depression achieved by Fatalism and Optimism?

Part f

Does the addition of Spirituality and Optimism add significantly to the prediction of Depression achieved by Fatalism?