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

- Please upload your homework to Sakai. **Upload both your .Rmd code file and the knitted .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 knitted html file.
- Write all answers in complete sentences as if communicating the results to a collaborator.
  - Points (usually 0.5-1) will be deducted for not 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 (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.

Chapter 14

Use the data from Chapter 5 Question 8 to answer the following questions.

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 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. Are their any outliers in the data? If so, describe the points and why you think they are outliers.

(2)

Write out the regression equation for the simple linear regression model.

(3)

Assess the normality of the model's (ordinary) 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.

(4)

Assess the normality of the model's (ordinary) residuals by creating a normal probability plot of the residuals. Compare the normality probability plot to 8 such plots simulated from normal data, and discuss why or why not the residuals could have come from a normal distribution.

(5)

Test the normality of the model's (ordinary) residuals and comment on whether the test's conclusion is consistent with your visual inspection or not. Make sure to include the hypotheses and a conclusion to the test based on the p-value.

(6)

Create a residual plot using ggplot and the standardized residuals and discuss what this shows us in terms of whether the model assumptions have been met or not.

(7)

Determine whether there are any observations with high leverage. If there are observations with high leverage, identify their coordinates and describe how they relate to the other observations. Why would these points have high leverage compared to the other observations? Do you think removing the points would change the linear model much? (you do not need to remove the points and rerun the model, just comment on whether you think they are influential)

(8)

Determine whether there are any observations with high Cook's distance. If there are observations with high Cook's distance, identify their coordinates and describe how they relate to the other observations. Why would these points have high Cook's distance compared to the other observations? Do you think removing the points would change the linear model much? (you do not need to remove the points and rerun the model, just comment on whether you think they are influential)

(9)

Create histograms and density plots of the dependent and independent variables and describe their distribution shapes.

(10)

Use Tukey's ladder of transformations to choose two possible transformations for the dependent variable. Explain why you chose them. Note: questions below will ask about model fit with the transformations. For now, just explain why you chose the ones that you did.

(11)

Use Tukey's ladder of transformations to choose two possible transformations for the independent variable. Explain why you chose them. Note: questions below will ask about model fit with the transformations. For now, just explain why you chose the ones that you did.

(12)

Add the 4 transformations you chose above (2 for the dependent variable and 2 for the independent variable) to the dataset.

(13)

Create scatterplots using the transformed variables and discuss whether any of the transformations improve the model fit and why (or why not). Include plots with just the x or y variables transformed, and at least one plot with both the x and y variables transformed.

(14)

Run the various transformed models and save the output to use for the diagnostic questions below.

(15)

Compare the normal QQ plots of the different models and discuss whether any of the transformations improve the model fit and why (or why not).

(16)

Compare the residual plots of the different models and discuss whether any of the transformations improve the model fit and why (or why not).

(17)

Compare the leverage & Cook's distance of the different models and discuss whether any of the transformations improve the model fit and why (or why not).

(18)

Compare the R^2 values and F-test p-values of the different models and discuss whether any of the transformations improve the model fit and why (or why not).

(19)

Which of the models would you recommend using for analyses? Discuss why you chose the model and why you did not choose the other models.