Homework 3

BSTA 512/612

2024-02-15

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

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 from the last section "Regression with one categorical predictor" #2, 3, 5, 7.

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

Problems to turn in & more directions

Directions - important!!!

- Hypothesis tests
  - For every hypothesis test make sure to include the following:
    * Null & alternative hypotheses
    * Calculation of test statistic using the formula
    * Calculate the p-value directly using its probability distribution
    * Running the test using R
    * Conclusion in the context of the research problem. This includes referring to variables by what they actually are and not X_1, X_2, etc.

See additional instructions/ clarifications in green.

Tips

- You will be running *a lot* of different tests below. I highly recommend coming up with a naming convention that will easily help you keep track of what variables are being included in which models.
  - The names model1, model2, etc. will not be helpful.

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.

Problem 9.7 (a, b, e, f) from book

Note: (c) is in recommended problems below.

> Use the results from Problem 4 in Chapter 8, as well as the computer output given here, to answer the following questions about the data from that problem.

Note: Create all output you need using R instead of relying on the output given in the text.

Problem 4 from Chapter 8:

> A sociologist investigating the recent increase in the incidence of homicide throughout the United States studied the extent to which the homicide rate per 100,000 population (Y) is associated with the city's population size (X_1), the percentage of families with yearly income less than \$5,000 (X_2), and the rate of unemployment (X_3). Data are provided in the following table for a hypothetical sample of 20 cities.

(a)

> Conduct the overall regression F test for the model where Y is regressed on X_1, X_2, and X_3. Use alpha = 0.05. Interpret your result.

(b)

> Provide variables-added-in-order tests for the order X_2, X_1, and X_3.

This means that there are 3 tests: (1) test model with just X_2, (2) test adding X_1 to the model given that X_2 is already in the model, and (3) test adding X_3 to the model given that X_2, X_1 are already in the model. See the subsections below to divide up the work.

Test model with just X_2

Adding X_1 to the model given that X_2 is already in the model

Adding X_3 to the model given that X_2, X_1 are already in the model

(e)

> Provide variables-added-last tests for X_1, X_2, and X_3.

This means that there are 3 tests: (1) test adding X_1 to the model given that X_2, X_3 are already in the model, (2) test adding X_2 to the model given that X_1, X_3 are already in the model, and (3) test adding X_3 to the model given that X_1, X_2 are already in the model. See the subsections below to divide up the work.

- For adding X_2 and X_3 last to the model, you do not need to calculate the test statistic using the formula or the p-value directly using its probability distribution. You can instead run the appropriate tests in R.

Adding X_1 to the model given that X_2, X_3 are already in the model

Adding X_2 to the model given that X_1, X_3 are already in the model

Adding X_3 to the model given that X_1, X_2 are already in the model

(f)

> Provide the variables-added-last test for $X_4 = X_2X_3$ given that X_2 and X_3 are already in the model. Does X_4 significantly improve the prediction of Y given that X_2 and X_3 are already in the model?

Chapter 10 - not a book problem

Use the data from Chapter 9 Problem 5 to answer the questions below.

> An experiment was conducted regarding a quantitative analysis of factors found in high-density lipoprotein (HDL) in a sample of human blood serum. Three variables thought to be predictive of, or associated with, HDL measurement (Y) were the total cholesterol (X1) and total triglyceride (X2) concentrations in the sample, plus the presence or absence of a certain sticky component of the serum called sinking pre-beta or SPB (X3), coded as 0 if absent and 1 if present. The data obtained are shown in the following table.

(a)

Calculate the coefficient of determination $r^2_{Y|X1,X2,X3}$ and interpret this value in the context of the problem. Do the calculation using the formula and then check your answer with R. In particular, where in the R output do we find this value?

(b)

Calculate the partial coefficient of determination $r^2_{YX1|X2}$ and interpret this value in the context of the problem. Do the calculation using the formula and then check your answer with R.

(c)

Use $r^2_{YX1|X2}$ to calculate $r_{YX1|X2}$ and interpret this value in the context of the problem. Check your answer with R.

(d)

Explain how the interpretations of $r^2_{YX1|X2}$ and $r_{YX1|X2}$. In particular, what information do each of these values tell us that the other does not?

(e)

Calculate the partial coefficient of determination $r^2_{YX1|X2X3}$ and interpret this value in the context of the problem. Do the calculation using the formula and then check your answer with R.

(f)

Use $r^2_{YX1|X2X3}$ to calculate $r_{YX1|X2X3}$ and interpret this value in the context of the problem. Check your answer with R.

(g)

Use your answers to parts (b, c, e, f), to discuss the change in the first-order partial correlation to the second-order partial correlation.

Regression with one categorical predictor (Prequel to Ch 11 & 12)

Penguins: Flipper length vs. species

For this problem we will be using the `penguins` dataset from the `palmerpenguins` R package.

Description from help file:

> Includes measurements for penguin species, island in Palmer Archipelago, size (flipper length, body mass, bill dimensions), and sex.

More info about the data are at https://allisonhorst.github.io/palmerpenguins/.

```
# first install the palmerpenguins package
# install.packages("palmerpenguins")
library(palmerpenguins)
data(penguins)

# run the command below to learn more about the variables in the penguins dataset
```

```
# ?penguins
```

(1) Outcome averages stratified by category levels

Calculate the average flipper lengths stratified by each of the penguin species.

(2) Visualize the "regression"

Make a scatterplot of flipper lengths by species, and include diamond-shape points for the averages of the flipper lengths for each of the species.

(3) Regression equations

Before running the regression in R, we are going to find the regression equation"manually."

Write out the regression equation using LaTeX math markup (see class notes) that models the flipper length by penguin species. Do not yet insert values for the regression coefficients, i.e. us the generic coefficients $\widehat{\beta}_0, \widehat{\beta}_1$, etc. Use Adelie as the reference level.

(4) Interpret coefficients

How do we interpret each of the regression coefficients for this model? *Write out a separate interpretation for each of the coefficients.*

(5) Regression coefficients "manually"

"Manually" calculate the values for each of the coefficients, and update the regression model with the values inserted.
You must show your work for this. Do not run the linear model in this step to get the values.

(6) Regression table with `lm()` function

Run the linear regression of flipper lengths vs. species in R, and display the regression table output. Which species did R choose as the reference level, and how did you determine this?

(7) Mean calculation using regression output

Calculate the mean flipper length of penguins in the Chinstrap and Gentoo species using *only* the results from the regression table. *You must show your work.*

Recommended extra problems do not turn in (will be not graded)

Below are some problems I *highly* recommend working on. They will not be graded and you do not need to turn them in.

Problem 9.7 (c)

Do this problem after completing 9.7 (b) above

> Provide variables-added-in-order tests for the order X_3, X_1, and X_2.

This means that there are 3 tests: (1) test model with just X_3, (2) test adding X_1 to the model given that X_3 is already in the model, and (3) test adding X_2 to the model given that X_3, X_1 are already in the model. See the subsections below to divide up the work.

Test model with just X_3

Adding X_1 to the model given that X_3 is already in the model

Adding X_2 to the model given that X_3, X_1 are already in the model

Chapter 10 (c) using the formula

Do this after completing Chapter 10 (c) above.

Calculate $r_{YX1|X2}$ using the formula and check that your answer matches that of Chapter 10 (c) above.

Regression with one categorical predictor (Prequel to Ch 11 & 12): Change the reference level to Gentoo

After completing exercises (1) - (7) in the section *Regression with one categorical predictor (Prequel to Ch 11 & 12)*, do the problems below.

(8)

Write out the regression equation using LaTeX math markup (see class notes) that models the flipper length by penguin species. Do not yet insert values for the regression coefficients, i.e. us the generic coefficients $\hat{\beta}_0, \hat{\beta}_1$, etc. Use Gentoo as the reference level.

(9)

How do we interpret each of the regression coefficients for this model? *Write out a separate interpretation for each of the coefficients.*

(10)

"Manually" calculate the values for each of the coefficients, and update the regression model with the values inserted. *You must show your work for this. Do not run the linear model in this step to get the values.*