Homework 4

BSTA 512/612

Your name here!!!

2024-02-22

> **!** 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: 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.*

(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 $\widehat{\beta}_0, \widehat{\beta}_1$, 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.*

Question 1

Use the data from Chapter 12 Problem 3 to answer the questions below.

a)

How many dummy variable(s) do you need to create for the categorical variable Diet (protein-rich vs. protein-poor)? Create the dummy variable(s) with the reference cell coding approach{0,1}.

b)

At a level of significance alpha = .05, test whether if Age is significantly associated with Height. Would this association be modified depending on diet group (e.g., rich-protein or poor-protein)? In other words, is Diet an effect-modifier that changes the association between Height and Age? Justify your answer (e.g., perform a hypothesis test at a level of alpha =.1).

Note: recall that an effect modifier is an interaction.

c)

From the results obtained in part b, should we perform an assessment of a confounder for Diet? Justify your answer. Perform such an assessment if needed.

d)

Perform a regression analysis on the model obtained from the results obtained from parts a-c. Write down a general regression equation that is applicable to both groups—rich-protein vs.poor-protein. Write down regression lines for each specific groups—rich-protein or poor-protein.

Question 2

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)

Use $\alpha = 0.05$, test whether the (crude) association between Y and X1 could be established.

b)

Use $\alpha = 0.1$, test whether X3 is an effect modifier of the association between Y and X1.

Note: To identify effect modifiers, we perform a hypothesis test of interaction term, e.g.,X1*X3. That is: The full model includes X1, X3, X1*X3. the reduced model includes X1 and X3

c)

From the result obtained in part b, do we need to perform an assessment of a confounder for X3? Justify your answer. Perform such an assessment if needed.

d)

Perform an assessment of a confounder for X2 which potentially changes the association between Y and X1.

e)

From the results in parts a-d, what is your final association model?

Chapter 15: Polynomial Regression (questions NOT from book)

Use the data from Chapter 5 Problem 18 to answer the questions below.

a)

a. Obtain scatter plot: Y vs. X. Does linear trend support the relationship between Y and X?

b)

b. At the level alpha =.05, test whether the linear relationship could be established between Y and X.

c)

c. At the level alpha =.05, test whether the quadratic term (X^2) should be included in the model to improve the prediction in Y, given the linear term (X) is already in the model.

d)

d. From the result obtained from part c, should we test if the linear term (X) is necessary to be included in the model, given the quadratic term is already in the model? Explain your answer.

e)

e. From the results you obtain from parts c-d, should we further examine whether cubic term (X^3) or fourth polynomial degree (i.e., X^4) to improve the prediction in Y? Explain your answer and report the result of such a test if needed.

f)

f. From parts a – e, what is the final model that you have obtained? Interpret the R-square result from this model in the context of the study. Plot fitted value curve vs. X overlaid with scatter plot. Comments about the fitting model.