

Homework 1

BSTA 512/612

Your name here - update this!!!!

2024-01-25

 Important

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

Directions

- You will need to download use [this link to download](#) the HW1 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 your HW1 .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.
- Write all answers in complete sentences as if communicating the results to a collaborator.

 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

The following questions were adapted from [this textbook](#). Questions 1-4 are optional if you would like a short review of 511 and coding practice.

Question 1

Please use R code to determine the following answers. (*adapted from problem 3.3 in Applied Regression Analysis and Other Multivariable Methods*)

i Type `?pnorm` in the console to get some information on a potentially helpful function.

Part a

From a normal distribution with mean 4 and standard deviation 6, what is $P(X > 2)$?

Part b

From a normal distribution with mean 4 and standard deviation 6, for what value (in place of ??) would $P(X > ??) = 0.1$?

Question 2

Suppose that the height (H) of assigned-male-at-birth (AMAB) patients registered at a clinic has the normal distribution with mean 70 inches and variance 4. (*adapted from problem 3.11 in Applied Regression Analysis and Other Multivariable Methods*)

Part a

For a random sample of patients of size $n = 25$, the expression $P(\bar{H} < 65)$, in which \bar{H} denotes the sample mean height, is equivalent to saying $P(Z < ?)$

i Z is a standard normal random variable.

Part b

Using the `pnorm` function, show that the probability expressions in Part a are equal.

Part c

Find an interval (a, b) such that $P(a < \bar{H} < b) = 0.80$ for the same random sample in Part a.

Question 3

Test the null hypothesis that the true population average height is the same for two independent groups from one hospital versus the alternative hypothesis that these two population averages are different, using the following data:

- Group 1: [69.25, 72.80, 68.73, 72.01, 70.36, 71.49, 72.73]
- Group 2: [67.54, 68.51, 71.84, 70.59, 71.52, 71.50]

You may assume that the populations from which the data come are each normally distributed, with equal population variances. What conclusion should be drawn, with $\alpha = 0.05$?

! Please attempt this problem using R. Take a look at the information for the `t.test` function. You will need to set `x`, `y`, `alternative`, and `var.equal=T`. You can use the below groups coded in R.

```
grp1 = c(69.25, 72.80, 68.73, 72.01, 70.36, 71.49, 72.73)
grp2 = c(67.54, 68.51, 71.84, 70.59, 71.52, 71.50)
```

Question 4

The choice of an alternative hypothesis (H_A or H_1) should depend primarily on (choose all that apply)

- a. the data obtained from the study.
- b. what the investigator is interested in determining.
- c. the critical region.
- d. the significance level.
- e. the power of the test.

Question 5

Visit this site on dplyr. <https://dplyr.tidyverse.org/reference/index.html#groups>

For one of the functions that we have not discussed in class, please use it on the `dds.discr` dataset.

Question 6

The accompanying table gives the dry weights (Y) of 11 chick embryos ranging in age from 6 to 16 days (X). Also given in the table are the values of the common logarithms of the weights (Z).

- Load the dataset using the `readxl` package.
 - This `readxl` package was installed as a part of the tidyverse, however it does not get loaded when you load the tidyverse package and thus you need to do that separately.
 - Use the command `read_excel()`, as shown below

```
library(readxl)
# you might need to update the location of the data file
# you can choose whatever name you like for the tibble when loading it into R's workspace
ch05q01 <- read_excel("./data/CH05Q01.xls")
```

Part a

Create the scatterplots in R using ggplot the above dataset. Observe the following two scatter diagrams. Describe the relationships between age (X) and dry weight (Y) and between age and log10 dry weight (Z).

Part b

State the simple linear regression models for these two regressions: Y regressed on X and Z regressed on X.

i This is asking for the regression models BEFORE you find the values of the coefficients.

Part c

Determine the least-squares estimates of each of the regression lines in part (b).

i Now get the regression coefficients using R and plug them into the regression models from (b). You can get the coefficients from the R output - you don't have to use the formulas.

Part d

Create a line on your plots. Which of the two regression lines has the better fit? Based on your answers to parts (a)–(c), is it more appropriate to run a linear regression of Y on X or of Z on X? Explain.

Part e

For the regression that you chose as being more appropriate in part (d), find 95% confidence intervals for the true slope and intercept. Interpret each interval with regard to the null hypothesis that the true value is 0.

i You can get the CI's from the R output - you don't have to use the formulas.

Part f

For the regression that you chose as being more appropriate in part (d), add 95% confidence and prediction bands. Using your sketch, find and interpret an approximate 95% confidence interval for the mean response of an 8-day-old chick.

Question 7

Below are the mathematical steps for determining the coefficient estimates using OLS for simple linear regression. For each blank explanation, please describe in words how we got from the previous mathematical step to the current one.

Step 1: Set up SSE

Mathematical Step	Explanation
1. $SSE = \sum_{i=1}^n \hat{\epsilon}_i^2$	Start with the sum of square errors (SSE)

Mathematical Step	Explanation
2. $SSE = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	
3. $SSE = \sum_{i=1}^n (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2$	
3. $SSE = \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$	

Step 2: Minimize SSE with respect to the coefficient estimates

Question 8

Quick True/False question: Is ordinary least squares the only way to find the best fit line for linear regression?