STAT 420: Homework 04

Fall 2020, D. Unger

Due: Tuesday, October 6 by 11:30 PM CT

Contents

Directions

Students are encouraged to work together on homework. However, sharing, copying or providing any part of a homework solution or code is an infraction of the University's rules on Academic Integrity. Any violation will be punished as severely as possible.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

- Your assignment must be submitted through the submission link on **Compass 2g.** You are required to attach one `.zip` file, named `hw03_yourNetID.zip`, which contains:
  - Your RMarkdown file which should be saved as `hw04_yourNetID.Rmd`. For example `hw04_dunger.Rmd`.
  - The result of knitting your RMarkdown file as `hw04_yourNetID.html`. For example `hw04_dunger.html`.
  - Any raw data supplied by me for the assignment. For example `nutrition.csv`.
- Your resulting `.html` file will be considered a "report" which is the material that will determine the majority of your grade. Be sure to visibly include all `R` code and output that is relevant to answering the exercises. (You do not need to include irrelevant code you tried that resulted in error or did not answer the question correctly.)
- You are granted an unlimited number of submissions, but only the last submission *before* the deadline will be viewed and graded.
- If you use this `.Rmd` file as a template, be sure to remove the directions section. Consider removing `eval = FALSE` from any code chunks provided in the template, if you would like to run that code as part of your assignment.

- Your `.Rmd` file should be written such that, if it is placed in a folder with any data your are asked to import, it will knit properly without modification.
- Unless otherwise stated, you may use `R` for each of the exercises.
- Be sure to read each exercise carefully!
- Include your Name and NetID in the final document, not only in your filenames.

Assignment

Exercise 1 (Using `lm` for Inference)

For this exercise we will again use the `faithful` dataset. Remember, this is a default dataset in `R`, so there is no need to load it. You should use `?faithful` to refresh your memory about the background of this dataset about the duration and waiting times of eruptions of the Old Faithful geyser in Yellowstone National Park.

(a) Fit the following simple linear regression model in `R`. Use the eruption duration as the response and waiting time as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in a variable called `faithful_model`. Use a t test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

(b) Calculate a 99% confidence interval for β_1. Give an interpretation of the interval in the context of the problem.

(c) Calculate a 90% confidence interval for β_0. Give an interpretation of the interval in the context of the problem.

(d) Use a 95% confidence interval to estimate the mean eruption duration for waiting times of 75 and 80 minutes. Which of the two intervals is wider? Why?

(e) Use a 95% prediction interval to predict the eruption duration for waiting times of 75 and 100 minutes.

(f) Create a scatterplot of the data. Add the regression line, 95% confidence bands, and 95% prediction bands.

Exercise 2 (Using `lm` for Inference)

For this exercise we will again use the `diabetes` dataset, which can be found in the `faraway` package.

(a) Fit the following simple linear regression model in `R`. Use the total cholesterol as the response and weight as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in a variable called `cholesterol_model`. Use an F test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The ANOVA table (You may use `anova()` and omit the row for Total.)
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

(b) Fit the following simple linear regression model in `R`. Use HDL as the response and weight as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in a variable called `hdl_model`. Use an F test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The ANOVA table (You may use `anova()` and omit the row for Total.)
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

Exercise 3 (Inference "without" `lm`)

For this exercise we will once again use the data stored in `goalies.csv`. It contains career data for all 716 players in the history of the National Hockey League to play goaltender through the 2014-2015 season. The two variables we are interested in are:

- `W` - Wins
- `MIN` - Minutes

Fit a SLR model with `W` as the response and `MIN` as the predictor. Test $H_0 : \beta_1 = 0.008$ vs $H_1 : \beta_1 < 0.008$ at $\alpha = 0.01$. Report the following:

- $\hat{\beta}_1$
- $SE[\hat{\beta}_1]$
- The value of the t test statistic
- The degrees of freedom
- The p-value of the test
- A statistical decision at $\alpha = 0.01$

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

You should use `lm()` to fit the model and obtain the estimate and standard error. But then you should directly calculate the remaining values. Hint: be careful with the degrees of freedom. Think about how many observations are being used.

Exercise 4 (Simulating Sampling Distributions)

For this exercise we will simulate data from the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where $\epsilon_i \sim N(0, \sigma^2)$. Also, the parameters are known to be:

- $\beta_0 = 4$
- $\beta_1 = 0.5$
- $\sigma^2 = 25$

We will use samples of size $n = 50$.

(a) Simulate this model 1500 times. Each time use `lm()` to fit a SLR model, then store the value of $\hat{\beta}_0$ and $\hat{\beta}_1$. Set a seed using **your** UIN before performing the simulation. Note, we are simulating the x values once, and then they remain fixed for the remainder of the exercise.

```r
uin = 123456789
set.seed(uin)
n = 50
x = seq(0, 20, length = n)
```

(b) For the *known* values of x, what is the expected value of $\hat{\beta}_1$?

(c) For the known values of x, what is the standard deviation of $\hat{\beta}_1$?

(d) What is the mean of your simulated values of $\hat{\beta}_1$? Does this make sense given your answer in **(b)**?

(e) What is the standard deviation of your simulated values of $\hat{\beta}_1$? Does this make sense given your answer in **(c)**?

(f) For the known values of x, what is the expected value of $\hat{\beta}_0$?

(g) For the known values of x, what is the standard deviation of $\hat{\beta}_0$?

(h) What is the mean of your simulated values of $\hat{\beta}_0$? Does this make sense given your answer in **(f)**?

(i) What is the standard deviation of your simulated values of $\hat{\beta}_0$? Does this make sense given your answer in **(g)**?

(j) Plot a histogram of your simulated values for $\hat{\beta}_1$. Add the normal curve for the true sampling distribution of $\hat{\beta}_1$.

(k) Plot a histogram of your simulated values for $\hat{\beta}_0$. Add the normal curve for the true sampling distribution of $\hat{\beta}_0$.

Exercise 5 (Simulating Confidence Intervals)

For this exercise we will simulate data from the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where $\epsilon_i \sim N(0, \sigma^2)$. Also, the parameters are known to be:

- $\beta_0 = 1$

- $\beta_1 = 3$
- $\sigma^2 = 16$

We will use samples of size $n = 20$.

Our goal here is to use simulation to verify that the confidence intervals really do have their stated confidence level.

(a) Simulate this model 2000 times. Each time use `lm()` to fit a SLR model, then store the value of $\hat{\beta}_0$ and s_e. Set a seed using **your** UIN before performing the simulation. Note, we are simulating the x values once, and then they remain fixed for the remainder of the exercise.

```r
uin = 123456789
set.seed(uin)
n = 20
x = seq(-5, 5, length = n)
```

(b) For each of the $\hat{\beta}_0$ that you simulated calculate a 90% confidence interval. Store the lower limits in a vector `lower_90` and the upper limits in a vector `upper_90`. Some hints:

- You will need to use `qt()` to calculate the critical value, which will be the same for each interval.
- Remember that `x` is fixed, so S_{xx} will be the same for each interval.
- You could, but do not need to write a `for` loop. Remember vectorized operations.

(c) What proportion of these intervals contain the true value of β_0?

(d) Based on these intervals, what proportion of the simulations would reject the test $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ at $\alpha = 0.10$?

(e) For each of the $\hat{\beta}_0$ that you simulated calculate a 99% confidence interval. Store the lower limits in a vector `lower_99` and the upper limits in a vector `upper_99`.

(f) What proportion of these intervals contain the true value of β_0?

(g) Based on these intervals, what proportion of the simulations would reject the test $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ at $\alpha = 0.01$?