Week 3 - Homework

STAT 420, Summer 2017, Dalpiaz

Directions

- Be sure to remove this section if you use this `.Rmd` file as a template.
- You may leave the questions in your final document.

Exercise 1 (Using `lm` for Inference)

For this exercise we will use the `cats` dataset from the `MASS` package. You should use `?cats` to learn about the background of this dataset.

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

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

Store the results in a variable called `cat_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 heart weight for body weights of 2.5 and 3.0 kilograms Which of the two intervals is wider? Why?

(e) Use a 95% prediction interval to predict the heart weight for body weights of 2.5 and 4.0 kilograms.

(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 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 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 $\alpha = 0.05$
- 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 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 $\alpha = 0.05$
- 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`)

Write a function named `get_p_val_beta_1` that performs the test

$$H_0 : \beta_1 = \beta_{10} \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_{10}$$

for the linear model

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

The function should take two inputs:

- A model object that is the result of fitting the SLR model with `lm()`
- A hypothesized value of β_1, β_{10}, with a default value of 0

The function should return a named vector with elements:

- `t`, which stores the value of the test statistic for performing the test
- `p_val`, which stores the p-value for performing the test

(a) After writing the function, run these three lines of code:

```
get_p_val_beta_1(cat_model, beta_1 = 4.2)
get_p_val_beta_1(cholesterol_model)
get_p_val_beta_1(hdl_model)
```

(b) Return to the goalies dataset from the previous homework, which is stored in `goalies.csv`. Fit a simple linear regression model with `W` as the response and `MIN` as the predictor. Store the results in a variable called `goalies_model_min`. After doing so, run these three lines of code:

```r
get_p_val_beta_1(goalies_model_min)
get_p_val_beta_1(goalies_model_min, beta_1 = coef(goalies_model_min)[2])
get_p_val_beta_1(goalies_model_min, beta_1 = 0.008)
```

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:

- $\beta_0 = 3$
- $\beta_1 = 0.75$
- $\sigma^2 = 25$

We will use samples of size $n = 42$.

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

```r
birthday = 18760613
set.seed(birthday)
n = 42
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. Do **not** use the `confint()` function for this entire exercise.

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

```
birthday = 18760613
set.seed(birthday)
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 $\alpha = 0.01$?