

Lecture 5 Exercises

Isabel Fulcher

8/13/2018

Install packages

```
library(tidyverse) #ggplot2, dplyr, etc.  
library(reshape2) #need this for melt()  
library(knitr) #need this for kable  
library(MASS) #contains dataset
```

Exercise I

Load the `birthwt` data. This data contains 189 observations, 9 predictors, and an outcome, birthweight, available both as a continuous measure and a binary indicator for low birth weight.

```
data(birthwt)  
head(birthwt)
```

1. Plot a scatterplot of birthweight (`bwt`) and mother's weight (`lwt`).
2. Use OLS to fit the regression of birthweight on mother's weight.
3. Extract the following: estimated coefficients, standard errors, variance-covariance matrix, and confidence intervals.
4. Plot the regression line and interpret the intercept and slope
5. Does the interpretation of the intercept make sense? How might we change this?
6. Now, we want to fit a model that includes race, mother's age, and smoking status in the model. Race takes on value 1 for white, 2 for black, and 3 for other. Mother's age is continuous. Smoking status is binary. Write out the regression function we may be interested in.
7. Use OLS to calculate the coefficient estimates in this model.
8. Interpret all the coefficient estimates.
9. Print the results in Rmarkdown using `kable()`.

Group Exercises

From the course website, load the North Carolina infant mortality dataset. This contains information on all 225,152 births in North Carolina from 2003-2004.

Group 1

The goal of this exercise is to emulate “sampling” from this North Carolina birth population and see how variability in our estimates will change with sample size.

1. You are interested in how maternal age affects birthweight. Write the form of the linear regression model.
2. Take a sample of size $n = 100$ from this population, fit the linear regression from part 1 and extract the coefficient estimate for gestational age, i.e. $\hat{\beta}_1$.
3. Repeat part 2 $b = 500$ times and plot the estimated coefficients in a histogram.
4. For the following sample sizes, $N = \{25, 50, 100, 500, 1000, 5000, 10000\}$, repeat questions 2-3. Save in a dataframe so you can plot your results.
5. Find a creative way to plot your results, and include some reference to the population β_1 . Interpret these results.
6. If you had instead extracted the standard error estimate for maternal age, what would you expect to happen?
7. Confirm your intuition by repeating this procedure again for the standard error of the beta coefficient estimate at various sample sizes.

Group 2

The purpose of this exercise is to practice working with regression models with interaction terms.

1. Take a random sample of size 10,000 from the NC dataset to work with for this problem. Make sure everyone in your group uses the same seed, so that you draw the same sample.
2. For this problem, you will be working with the following model where Y is birth weight, X_1 is weight gain during pregnancy and X_2 is smoking. What does β_3 represent? Why might this be of interest?

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

3. Create a scatter plot of maternal weight gain and birth weight. Color observations according to smoking status.
4. Use the expression $\hat{\beta}$ given in the slides to find the estimates of the coefficients. Note: for this question, you will need to create the “design” matrix, \mathbf{X} .
5. Fit this regression using the `lm()` function. How does this compare with the results from part 4?
6. Interpret the coefficients for weight gain and smoker. Be as precise as possible.
7. Plot the regression line for smokers and non-smokers on part 3. Hint: use `stat_function()` in `ggplot` and define your own function.
8. Do you see large differences in the slopes of these lines? Which p-value in the regression output formally tests this? Does this align with your expectations?