Lab 1: Simple linear regression in R

Math 430, Winter 2017

0. Setup

The r markdwon template for this lab is available on the course webpage (right click and download the file).

We will use the `ggplot2` package for plotting during this lab. Make sure that you have installed it before running the command

```
library(ggplot2)
```

1. The data

How much is an extra square foot of space worth? In this lab, we investigate this question for homes in Saratoga County, New York. The data set contains the prices of 1,728 homes. To begin we must load the data set.

```
saratoga <- read.csv("https://github.com/math430-lu/data/raw/master/saratoga.csv")
```

1.1. Exploring the data

Before jumping into a regression analysis, it's a good idea to plot the variables involved to determine whether there is linear relationship between them.

Question 1. Create a scatterplot of price vs. living area and describe the relationship. (Look at the column names carefully before trying to create this plot!)

In addition to exploratory plots, we can quantify the strength and direction of the linear association between the price and living area using the correlation. To calculate the correlation in R we use the `cor` command.

Question 2. Calculate the correlation between price and living area.

2. Fitting a simple linear regression model in R

Having verified that a linear relationship is plausible between price and living area, we fit a regression model to further explore this relationship. In R, we use the `lm` function to fit a simple linear regression. The basic syntax follows the same pattern as before: `lm(y ~ x, data = data_frame)` where `y` is the response variable and `x` is the explanatory variable.

Question 3. Fit the regression model using living area to predict the price of a home.

Question 4. Overlay the fitted regression line on the scatterplot of price vs. living area.

Question 5. Report the fitted least squares regression equation.

Question 6. Interpret the estimate of the slope within the context of the problem.

Question 7. Interpret the estimate of the y-intercept within the context of the problem. Does this interpretation make sense in this context?

3. Inference for the slope

Up to this point we have simply described what our simple linear regression model reveals about the 1,728 houses contained in our sample; however, if we drew a new sample from the population it may give a different fitted least squares regression equation. In order to draw inferences about this linear relationship back to the population of all houses in Saratoga County, New York we must either construct confidence intervals for the slope and/or intercept (also called regression coefficients), or we must run hypothesis tests.

Question 8. Create a 95% confidence interval for the slope. Interpret this confidence interval within the context of the problem.

Question 9. Carry out a hypothesis testing whether the slope = 0. Report your p-value and a short conclusion.

Question 10. Do the results of your confidence interval and hypothesis test agree?