Stat 311 Homework 3

This assignment will the `UBSprices` data set that is part of the alr4 package. You must first install the alr4 package. The template for this assignment includes the addition of `library(alr4)` to the set-up chunk. To access the data in R, use `data(UBSprices)`. The data will be in the object `UBSprices`. Use `?UBSprices` to bring up the help file that provides a description of the variables in the data set. [NOTE: The year 2003 was before the major recession that hit much of the world around 2006, and the year 2009 may reflect changes in prices due to the recession]. This assignment focuses on simple linear regression.

1. Some basic univariate EDA before thinking about regression
   a) What are the units for the rice variables? Why did the author of the data set choose these units?
   b) Use the ggplot2 package to make histograms and boxplots of `rice2003` and `rice2009`. Use `ggarrange` from the `ggpubr` package (install the package first if you did not install it when playing with the demo code in Quantitative.Rmd) all four graphs in a single figure. I recommend using 5 bins for each of the histograms. Summarize the distributions of rice 2003 and rice 2009.
   c) Create 7-number summary statistics for `rice2003` and `rice2009`. Briefly explain how the summary statistics are consistent with what you found in the graphs in part (a). Are there any missing values for either variable?
2. Making and interpreting scatterplots.
   a) Make a scatterplot of `rice2009` (y) on `rice2003` (x) and the line $y = x$ to the scatterplot. [Hint: to add the line, use `geom_abline(slope=1, intercept = 0)` if plotting with `ggplot2`, or use `abline(0,1)` if using `plot`].
   b) Describe the joint relationship between the two variables plotted in part(a). What is the main difference between points that fall above the $y = x$ line and points that fall below this line?
   c) Which city had the largest increase in rice price and what is that increase? Which city had the largest decrease in rice price and what is that decrease? [We have included some code to help with these questions. Make sure you understand what the code is doing]
3. Fit a linear regression for `rice2009` on `rice2003`. Create a second scatterplot that shows the regression line. Write out the regression equation. Interpret the estimated slope parameter for the regression line in the context of the problem. [Round both parameter estimates to one decimal place]
4. What is the coefficient of determination for the regression of `rice2009` on `rice2003`? Provide an interpretation of this value in the context of the problem.
5. What is the residual standard error for the regression of `rice2009` on `rice2003`? Provide an interpretation of this value in the context of the problem.
6. What is the expected value of the price of rice in 2009 when the price of rice in 2003 is 25 minutes of labor? You can hardcode the numbers using the rounded numbers you wrote out in problem 3, or you can use `lm.out$coef[1]` to pull of the y-intercept and `lm.out$coef[2]` to get the slope, where `lm.out` is the name you used for the saved linear model object. The answers will be similar within rounding error]
7. Chicago shows a 2003 rice price of 7 minutes of labor. What is the regression residual for Chicago based on the fitted regression line? Same rules for calculation as given in problem 6.
8. The data point for Mumbai appears to have high leverage. Explain what is meant by this.

Stat 311 Homework 3

9. Is the data point for Mumbai (observation 36) influential? Set up a new regression to help you answer this question. [Hint: To remove an observation from a data frame, use `NewDat <- OldDat[-row,]`]. Explain whether you would consider Mumbai influential.
10. Create a residual plot for the original regression of `rice2009` on `rice2003`. Do you see any patterns in this plot that indicates a violation of any of our regression assumptions? Explain.
11. Make a histogram and normal probability plot of the residuals for the original regression, putting both plots in one figure. Do you think the residuals are approximately normally distributed? Explain.
12. Do you think that `rice2003` is a useful variable for estimating the price of rice in 2009 using a linear model? Explain.