Assignment 2

For all of the below, include a link to your code.

1. Write your own original code that produces a dataset that conforms to the classic univariate regression model. Your data set should have 999 observations and a Normal error term. The slope of the coefficient on your regressor should be positive. Now include a single outlier, such that when you fit a regression to your 1000 data points, the slope of your regression line is negative. Your answer to this question should consist of:
   (a) Your original data-generating equation

   (b) Regression results for the original 999 (copy/paste the "summary" output)

   (c) Regression results with the outlier included (copy/paste "summary" output)

   (d) A properly-labeled data visualization that shows the regression line based on the original 999 points, and another differentiated regression line (on the same axes) based on 1000 points.

   (e) No more than 3 sentences that would serve as a caption for your figure if it were to be included in an econometrics textbook to illustrate the dangers of extrapolation.

2. *NOTE: FOR THIS PROBLEM (AND THIS PROBLEM ONLY), USE ONLY THE CONTROL GROUP. DO NOT USE ANY UNITS FOR WHICH TREATMENT == 1.* Using the Lalonde data set and a linear model that predicts re78 as a linear additive function of age, educ, re74, re75, educ*re74, educ*re75, age*re74, age*re75, age*age, and re74*re75, estimate:
   - the 95% interval of expected values for re78, for every unit (i.e., each age 17-55, spanning the age range in the data set), using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You should not incorporate simulated sigmas, and you should hold educ, re74, and re75 at their medians. Even include ages that are not covered by the data (e.g., 47, 49, etc.).

   - the 95% interval of expected values for re78, for every unit, using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You should not incorporate simulated sigmas, and you should hold educ, re74, and re75 at their 75% quantiles.

   - the 95% prediction interval for re78, for every unit (i.e., each age, spanning the age range in the data set), using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You will need to incorporate simulated sigmas, and you should hold educ, re74, and re75 at their medians.

   - the 95% prediction interval for re78, for every unit, using simulation (i.e., 10000 simulated predictions for every row from 10000 sets of coefficients). You will

need to incorporate simulated sigmas, and you should hold educ, re74, and re75 at their 75% quantiles.

Your answer to this question should consist of the following:

(a) A table with the relevant point estimates (e.g., the bounds of the prediction intervals of y for the different ages, and the medians of the other predictors)

(b) 1 figure for the 2 interval analyses with expected values, and 1 figure for the 2 interval analyses with predicted values. The "scatterplots" don't have to show the original data--all I am interested in are the prediction intervals for each age. Each of these figures should show how the intervals change over time (i.e., over the range of ages in the data set). Be sure to label your plot's features (axis, title, etc.).

E.g.: https://gist.github.com/diamonaj/75fef6eb48639c2c36f73c58d54bac2f

3. Obtain the PlantGrowth dataset in R.

Specify a regression model in which the dependent variable is *weight* and the independent variable is an indicator of treatment1 (set the value = 1) or control (set the value = 0). This means you will discard observations associated with treatment2.

Then, bootstrap the 95% confidence intervals for the value of the coefficient for treatment. Then, obtain the analytical confidence interval for the coefficient value using the standard error that pops out of a regression (or equivalently, in R, you can use the *confint* function). Compare the two confidence intervals--one obtained via simulation, the other via the formula.

NOTE: Make sure that you don't use a 'canned' bootstrap function -- please code the bootstrap routine manually.

Your answer to this question should consist of the following:

(a) A table with the relevant results (bounds on the 2 confidence intervals).

(b) 1 histogram (properly labeled) showing your bootstrap-sample results. How you do this one is up to you.

(c) No more than 3 sentences summarizing the results and drawing any conclusions you find relevant and interesting.

4. Write your own function (5 lines max) that takes Ys and predicted Ys as inputs, and outputs R^2. Copy/paste an example using the *PlantGrowth* data (from #3 above) that shows it working.

5. Obtain the *nsw.dta* dataset from http://users.nber.org/~rdehejia/data/nswdata2.html.

Read the description of this data set provided on the page. If you proceed with this work in R (recommended) use the *foreign* library to open it (so you can use *read.dta*).

Use this *nsw.dta* dataset to estimate the probability of being assigned to the treatment group (vs. the control group) for every observation in the data set. Your logistic regression model should be a linear additive function of all predictors available to you -- no interaction terms needed. NOTE: re78 is not a predictor because it postdates the treatment. (In other words, it's an outcome.)

Your answer to this question should consist of the following:

(a) Two properly labeled histograms: one in red (showing the distribution of the treatment group's estimated probabilities) and one in blue (showing the distribution of the control group's estimated probabilities). Extra credit for a legend in the plot.

(b) No more than 3 sentences summarizing the differences between the two distributions of estimated probabilities, and whether/not your results are surprising and/or intuitive.