Homework 4.

Problem #1 (finishing leftovers from previous HW)

For the *fl_crime.csv* data:

3. Proceed to fit a $crime \sim education$ linear regression, and

    d. Report and interpret the R^2 value.

4. Hard-code the calculation of R^2 value from scratch (explicitly applying formula from the top of slide #38), via only using the $y = fl\_crime\$crime$ and $y.hat = predict(lm.obj)$ as the quantities to work with. Double-check it with R^2 from part 3.

5. Calculate the predicted crime rate for education level of

    - 70,
    - 35.

    Comment on whether we can trust either of the predicted values, and why (name the issue).

Problem #2 (finishing leftovers from previous HW)

Broadband.csv contains data on each country's GDP (measured in billions USD) and the number of broadband subscribers. Proceed to

2. Fit a linear regression for the # of broadband subscribers onto GDP, and

    d. Report and interpret the R^2 value.

Problem #3

For the *fl_crime.csv* data, proceed to:

1. Subdivide *urbanization* variable into three groups: ≤ 33, $(34, 66]$, $(66, 100]$.

2. For each urbanization group:

    a. Provide a plot of crime rate (response) against education (explanatory). Fit simple linear regression of crime rate (response) onto education (explanatory), and add the fitted line to the plot.
    b. Calculate correlation between education and crime rate.

3. Compare the results from 2(a) with the line fitted for the whole data set: did the direction of the **overall** *crime-education* linear relationship change after conditioning on *urbanization* (at least for some *urbanization* levels)? Compare the **overall** correlation between *crime* and *education*, with the ones calculated in 2(b): did the directions change?

4. If you witnessed any changes as a result of breaking the data down by groups according to *urbanization*, what is the name for that phenomena? What do we call the *urbanization* variable with respect to the *crime-education* relationship then?

5. For the highest urbanization level, proceed to write down the fitted linear regression equation, interpret the intercept (does it make sense to interpret it?), the slope and the R^2.

6. For any of the urbanization levels, does there appear to be an influential observation? If yes - what is it? In either case, proceed to outline the ways to deal with influential observations introduced in class.

Problem #4

Broadband.csv contains data on each country's GDP (measured in billions USD) and the number of broadband subscribers.

1. Plot the broadband subscribers against GDP, along with a fitted regression line. Provide the correlation between the two variables. Does there appear to be a regression outlier? Which country is it? Proceed to dispose of it and:
   - Re-calculate the correlation. Compare it with the correlation you got prior to removing the outlier. What do you witness?
   - Re-fit the linear regression: Would you trust the new slope value more than before the outlier was removed? Why?

2. Provide the two fitted regression lines (least squares regression with & without the outlier) on the same plot. Make sure to apply different coloring to those lines. Comment on the differences between the lines.

Problem #5

3.91, 3.92, 4.2, 4.3