Lab 6 Activity

In this lab activity you will use the `insurance.csv` dataset. First, click here to download the dataset. These data contain the information on yearly health insurance charges for 1,338 individuals. Here is a description of the variables in the dataset:

variable	**description**
age	Age of primary beneficiary
sex	Insurance contractor gender, female, male
bmi	Body mass index
children	Number of children covered by health insurance / Number of dependents
smoker	Whether the beneficiary is a smoker
region	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest
charges	Individual medical costs billed yearly by health insurance in $

1. Load the data and run a regression where `charges` is the DV (y) and `age` is the IV (x). Before looking at the results, whuld you expect the slope of this regression to be positive or negative? why? Print your results and write out the regression equation given the estimated intercept and the regression slope.

2. Interpret the intercept. What is the meaning of the intercept in this case? Is it meaningful given the scale of the data? Justify your answer.

3. Interpret the slope. What is the meaning of the slope in this case? According to the regression equation, what would be the predicted yearly health insurance charge for someone who is 40 years old?

4. Create a scatterplot of `charges` on the y-axis and `age` on the x-axis with a regression line. Additionally, also create a scatterplot of the residuals of `charges` on the y-axis and `age` on the x-axis, along with a line at $y = 0$. (**Note:** you will notice that the two plots are *very* similar. in fact, the residual plot is the regression plot but tilted such that the regression line is flat)

- What do you make of the pattern of the residuals? Do you think that the regression line adequately described the relation between `charges` and `age`? Are the some points that are not well predicted by the regression line?

- You should notice some unusual patterns in the residuals plots (i.e., separate clusters of points). Look back a the table with the description of each variable; can you identify any variable that may explain the unusual pattern in the residuals and the relatevely poor performance of our regression? (this question is a bit more conceptual, so we can discuss in class)