Project 3: Solar Equity

Day 2

Task 2: Polynomial Regression

Many students may have already performed **polynomial regression** but may not realize it. A first order polynomial is a linear fit. Given a set of data points, a linear regression will try to fit a line through the data such the error between the data points and the line is minimized. A common way to evaluate how good the fit is (how closely the data points are to the line) is by using the **mean squared error (MSE)**. An introduction and example of the MSE can be found here.

Task 2.1: Practice with polynomial regression

Write a script that does the following:

- Generate a polynomial function of your choice, with any order (so generate X and Y vectors, where Y = f(X);
- Plot the data. Use markers rather than lines.
- Modify the function so that the data shown on your plot is noisy. You may choose how much noise to add but you should be able to see it on the plot. Hint: use the `rand` function to generate a vector of noise.

Now experiment with the function `MSE_polyFit.` What does the code do? How might you manually select the input `poly_order`?

Question 2.1: Try running your code with different amounts of noise. How does the Mean Squared Error (MSE) change? Be specific with how you generated the noise and MSE values.

Question 2.2: How does MSE change when you change poly_order?

Question 2.3: Suppose you are analyzing real data. Do you think it is a good idea to select the order to minimize MSE? What could be the problems with this approach?

Task 2.2: Applying to real data

Now let's apply polynomial regressions to our actual data. Look at the trend between solar cell adoption (`normalized_existing_installation`) and median income (`Med_HHD_Inc_ACS_09_13`) for each majority race/ethnicity. First try to fit polynomial fits of orders ranging from 1 to 5 for the Black majority table by changing the function input `poly_order.` Which polynomial order had the lowest MSE? Which do you think is the right one to use here (there is no right or wrong answer) Be sure to comment on the effects of outliers.

Now generate plots for each majority race/ethnicity and “No majority.” In total you should have five plots, each showing individual data points for each tract with polynomial fit overlaid. Be sure to include titles and label axes.

TO SUBMIT:

(a) Your modified code. Be sure to generate plots for Task 2.2 and includes comments explaining any new code that you added.

(b) A text, Word or PDF file with answers to each of the questions here.