DATA 310 Homework 6

Marc Trussler

2/19/2021

Problem 1

We are going to work again the ACS County Data to investigate the relationship between median household income and the percent of children living in poverty in counties. Load in the "ACSCountyData.Rdata" dataframe.

(a) First, to make things more readable, recode the `median.income` variable to be expressed in thousands of dollars.

(b) Plot the relationship between median income on the x axis and percent child poverty on the y axis and describe what you see.

(c) Run a bi-variate linear regression on this relationship, discuss what the coefficients (including the intercept) mean, and visualize the result on top of the scatterplot you produced above.

Just to make things easier for the next step, you may want to use code similar to this to plot the result:

```r
p <- ggplot(acs, aes(x=median.income, y=percent.child.poverty)) + geom_point() +
  ylim(0,100) +
  labs(x="Median Income (Thousands)", y = "Percent Childen in Poverty") +
  geom_smooth(method = lm, formula = y ~ poly(x, 1), se = FALSE)
```

(d) Looking at this relationship visually, why doesn't this regression satisfy Gauss-Markov Assumption 2 (functional form)? Use the `poly()` function to add the square of `median.income` to your model and determine whether this improves model fit, making reference to both the visual change in the regression line and to the R^2 of each model.

(e) In this new regression with a second order polynomial term, what is the the effect of an additional $1000 in median income when median income is at $30k? What is the the effect of an additional $1000 in median income when median income is at $100k? Does this make theoretical sense?

(f) A possible confounding variable to this relationship is the unemployment rate, which may affect both the median income of a county and the percent of children living in poverty. Use the `cor()` function to investigate the relationships between median income, unemployment, and child poverty. Based on the patten of correlations, what is likely to happen to the coefficient on `median.income` if you add unemployment rate to the first regression model (the one without the polynomial terms)?

(g) Run this regression with unemployment rate and median income (no polynomial terms), and determine the degree to which the coefficient on `median.income` changes. Interpret the other coefficients in the model as well, being sure to adjust your language to the fact that there are now multiple indpeendent variables.

(h) Another possible confounding variable is the census region people are living in. For example, living in the south could be associated with both lower average incomes and more child poverty. Create an indicator variable for the 4 census regions (or change the variable into a factor variable) and then re-estimate the regression with median income and unemployment to take into account what region each county is in. Interpret the coefficients from this regression.

(i) It's possible that the effect of median income is different conditional on whether a county is urban or not. Create an indicator variable for whether a county is urban (population density greater or equal to 1000) or not. Interact this variable with median income in the regression with unemployment rate and census region indicators. Interpret the coefficients on median income, the urban indicator, and the interaction term.