Econ 103: Final Project

Your Name

9/05/2021

Final Project Instructions

For this final project we will use the following dataset whose observations are various counties in the U.S. The outcome variable of interest is the "mobility" variable, which is a measure of the causal effect of growing up in a county on a person's likelihood to attain an income in the top 25% of the income distribution. This measure is taken from the Raj Chetty et. al paper, "Where is the Land of Opporunity? The Goegraphy of Intergenerational Mobility in the United States" (AER, 2014).

The other variables in this dataset are various demographic characteristics of these counties taken from the 2015 CPS (Current Population Survey).

Non-Numeric Variables:

- cz_name: County Zone Name
- stateabbrv: State Abbreviation
- name: County Names

Numeric Variables:

- population: The population of the county
- density: The poulation density of the county (number of people per square mile)
- male: The percentage of the county's population that is male
- white: The percentage of the county's population that reports as white
- black: The percentage of the county's population that reports as black
- amind: The percentage of the county's population that reports as American Indian
- asian: The percentage of the county's population that reports as asian
- highschool: The percentage of the county's population that has graduated highschool
- college: The percentage of the county's population that has graduated college
- dropout: The percentage of the county's population that has droppped out of college
- labforce: The percentage of the county's population that is participating in the labor force
- manufacturing: The percentage of the county's population that works in manufacturing
- income: The average income of the county
- gini: The county's gini coeffecient (a measure of income inequality within the county, higher = more unequal)
- rentpct: The average rent as a percentage of income in the county
- poverty: The percentage of the county's population who are living below the federal poverty line

We would like to explore the (associative!) relationship between these demographic variables and county mobility rates.

Single Linear Regression

Plotting

For this section, we will investigate the relationship between average rent as a percentage of income and county mobility rates. Let's start investigating this relationship by making a scatter plot of the two variables, with rent as a percentage of income on the y-axis and mobility rates on the x-axis. Be sure to give your plot x-labels, y-labels, and a title. Below the code chunk, discuss the relationship you see between rent as a percentage of income and county mobility rates. In your discussion comment on

- Whether there looks like there is a positive or negative association between these two variables
- Whether the relationship looks linear or non-linear

```
# Put your plot here and discuss below
```

Exploratory Data Analysis

Using the functions "mean", "median" and "var", report the sample mean, sample median, and sample variance for both mobility and average rent as percentage of income. Put all discussion below the code chunk.

```
n = nrow(data)

# Code for sample mean, median, and variance of rentpct

# Code for sample mean, median, and variance of mobility

# Discuss below:
```

Linear Model

Next, let's formally investigate this relationship between average rent as a percentage of income and county mobility rates by running a linear regression of mobility against rentpct. Below this code chunk answer the following:

- Interpret the estimated slope and coeffecient in context.
- Can we reject the null hypothesis the rentpct has a positive association with mobility rates against the alternative hypothesis that it has a negative association at level $\alpha = 0.01$?
- What is the Rˆ2 of this regression? How can we interpret this in context?
- What is the predicted value of mobility for a county whose average rent as a percentage of income is 27?

```
# Code for linear regression
```

Visualizing our regression line

Now let's visualize this. Add the regression line to your scatter plot from before. Discuss below this code chunk whether or not you think homoskedasticity is violated.

```
# Scatterplot
```

Interpretation

Discuss below whether we can interpret these results as evidence of a causal relationship between rent as a percentage of income and county mobility rates. List two possible confounding factors that we are not including in this regression, one within the dataset and one from outside the dataset. Is excluding these biasing the coeffecient on rentpct upwards or downwards. Discuss why you think so, using both formulas from class and intuition to back up your statments.

Answer:

Multiple Linear Regression

Model Selection I

Pick 3 variables that you think are important for explaining county mobility rates (could include rentpct). Discuss below why you think each variable would be an important explanatory covariate.

Answer:

Exploratory Data Analysis

Answer the following questions:

- For each explanatory variable, after controlling for the other two explanatory variables, do you expect the relationship between the explanatory variable and county mobility rates to be positive or negative?
- Plot the scatter plot between each of your explanatory variables and the outcome variable of interest of mobility? Based on these scatter plots does the relationship between each explanatory variable and mobility look to be increasing or decreasing? Does the relationship look linear or non-linear?

For each plot make sure you include an x-label, a y-label, as well as a title. Put all discussion below this code chunk.

```
# Inclde code for plotting below.
```

Multicollinearity

Using the function cor(x,y), calculate the correlation coeffecient between each of the chosen explanatory variables. Do we look to have a multicollinearity problem? How would a high correlation coeffecient affect the variance of our estimates?

```
# Ifneeded, get help by running the below line
?cor

# Calculate correlation coeffecients
```

Regression Implementation

Regress county mobility rates against your 3 variables. Interpret each of your slope coeffecients in context as well as the Rˆ2 from this regression.

Pick one hypothesized relationship from the last selection. Can you reject the null hypothesis that this prediction is incorrect and conclude in favor of your hypothesized relationship? Use level $\alpha = 0.01$. Report both the test statistic t* and the p-value.

Put all discussion below the code chunk.

```
# Linear Regression Code



# Hypothesis Testing Code
```

Plotting

Pick one of your explanatory variables to visualize. Make a scatter plot of this variable and mobility. On top of this scatter plot, fix the other two explanatory variables at their mean values and use your fitted multiple regression model to plot the estimated relationship between your chosen explanatory variable and mobility. Based on this plot, does homoskedasticity look to violated?

Remember to add an x-label, y-label, and plot title. Put all discussion below the code chunk.

```
# Code for plotting below
```

Hypothesis Testing

Use the model above to test the null hypothesis that the sum of all the slope coeffecients is equal to one against the alternative that it is not equal to one. Do we reject the null hypothesis at level $\alpha = 0.05$?

Hint: the vcov(regression) command may be useful here. Put all discussion below the code chunk.

```
# Help for the 'vcov' command
?vcov
# Code for hypothesis testing
```

Model Selection II

Consider adding three more terms to this model. These terms could be new variables from the dataset, transformations of your existing terms, or interaction terms. For each additional term, provide a short justification for why you are considering adding this specific term. The justification could be based on the scatter plots above or some economic theory justification (diminishing marginal returns, etc.)

Re-run your regression with all these transformed terms (as well as the original terms). Is your adjusted Rˆ2 higher or lower than before? Run an F-test to determine whether adding these new terms significantly increased the explanatory power of the model. State your null hypothesis and alternative hypothesis in terms of the full model parameters. Use level $\alpha = 0.1$ and interpret the conclusions of this test in context.

Put your conclusions below this code chunk.

```
# Code for new model with transformations

# Code for F-Test
```

Conclude

In a few sentences, comment on what you have learned about the associative relationship between mobility rates and the explanatory variables considered above. Which variables seem to be the most important? Are there any variables that you thought would be significant that weren't? Was your model able to explain more or less of the variation in mobility rates than you expected?

Answer:

Finish Up

Knit this document to pdf and submit the pdf, as well as your .rmd file, to CCLE (remember to change your name at the top under "author"). Congrats on completing Econ 103!