The impact of Business Types on Poverty Rate

Luke Borowy, Tae Kosmo, Jair Vargas. Group C2

Purpose: We want to understand various factors that influence the amount of poverty in an area. In particular, we want to study how the poverty rate of a county is influenced by the type of businesses that are located there. We expect that some types of business will tend to increase the wealth of an area, hence lowering the poverty rate. For example, health care businesses are likely more concentrated in wealthy counties, leading to a negative association with poverty rate. In contrast, there may be an increase in poverty associated with vice related businesses. We want to determine which specific business types have the greatest impact on poverty rate in an area. Additionally, we anticipate that there will be significant interaction and multicollinearity effects between our variables because of the ways that local businesses affect each other and the local economy.

Data: Our data is from the United States Environmental Protection agency. It can be located at their Environmental Dataset Gateway, at https://edg.epa.gov/EPADataCommons/public/ORD/NHEERL/EQI/ . We are specifically using the dataset of EQI data from July 2013. This dataset contains 227 variables categorized into 5 domains. These are Air, Water, Land, Built, and Socioeconomic. These variables are intended to describe the environmental health of an area, as well as the human-made effects. The EPA gathered information from several other sources in order to assemble this complete report. There is a description of the dataset itself available at https://edg.epa.gov/EPADataCommons/public/ORD/NHEERL/EQI/EQI%20Technical%20Report_Final.pdf

Population: The observational units in our dataset are all 3141 individual counties from every state in America. This data is complete, and covers every case that we are interested in. In order to conduct a better analysis, we will first take a random sample of the counties, to obtain 300 counties. This should address some independence issues arising from the geographical proximity of counties. Therefore we will be trying to make inferences about the larger population (all counties) by analyzing our random sample.

Response variable: The response variable that we want to study is the percent of people with income below the poverty level, or `pct_pers_lt_pov`. This variable is measured in percent. We expect that the percentage to lie between 3 and 30 percent for most counties, with the possibility of outliers in a few especially rich or poor areas.

Explanatory variables: There are 227 different variables available in this dataset. However, we are focusing only on 9 variables in the Built domain in order to explain the poverty rate. These variables are

- Vice related businesses: `rate_al_pn_gm_env`
    - Casinos, alcohol, etc
- Entertainment related businesses: `rate_ent_env`
- Education related businesses: `rate_ed_env`
- Negative food related businesses: `rate_food_env_neg`
    - Sell fast food, food trucks, etc
- Positive food related businesses: `rate_food_env_pos`
    - Sell healthier foods, like grocery stores, sit-down restaurants, etc
- Healthcare related businesses: rate_hc_env
- Recreation related businesses: `rate_rec_env`
- Transportation related businesses: `rate_trans_env`
- Civic related businesses: `rate_civic_env`
    - Government, non-profits, etc

Each one of these variables is measured by the number of businesses divided by the county population. We expect that we will apply log transformations to make these variables more useful for linear regression.