In class, we fit a Poisson regression model to the Chicago Crime dataset that incorporated information such as the year the crimes were reported and the community area; however, a potentially large confounding variable is the income level of the community area in which the crime was committed. In this assignment, I add yearly median household income from the American Community Survey (ACS) and then compare my model for the number of thefts that occur in Chicago to the corresponding model we fit in class. Due to the number of unique addresses in the Chicago Crime dataset and since the addresses must be manually downloaded from the Census, I decided to focus on the most prevalent crime reported, ‘theft’.

Recall, the Chicago Crime dataset contains address-level information for reported crimes from 2001 to 2018. It contains information such as the date and address of the crime, the crime type and description, and whether there was an arrest. Before continuing with any analysis, I keep reports with the primary type as “Theft” and date between 2012 and 2016.[[1]](#footnote-0)

I add the median household income information by determining which Census Tract each crime was reported in and then merging in the ACS data. Because the Chicago Crime dataset does not contain the exact addresses for which the crimes were reported, I approximately match the addresses by replacing the ‘XX’s in the addresses listed with ‘00’, attempting to find a match in the (batch address) Census Geocoder, and repeating the process with ‘01’ if no match was returned.[[2]](#footnote-1) I then merge in the ACS data by Census Tract. Finally, I aggregate the data to be the total number of thefts and the median household income for each community area and year. My final dataset, excluding missing values, consisted of 384 observations for the years 2012 to 2016. One important potential limitation of my model is communities that did not report a crime during a year are not included.

I then fit a Poisson regression model. Recall, the model in class incorporates four splines for the years from 2003 to 2018, community area as a categorical variable, and ten splines for the months in each year. My model, however, incorporates four splines for the years from 2012 to 2016, community area as a categorical variable, and six splines for income level. I used four splines for the year variable because human behavior does not reset at each year and this allows for direct comparison with the class model. I chose six splines for income level because neighborhoods are often colloquially classified as low-income, middle-class, and high-income areas, and six splines allows for a similar, but slightly more granular delineation. Additionally, one would not expect large changes in other unobserved factors (such as average education level) when the difference between the median incomes of two community areas is fairly low, e.g. 10k, but would expect large differences when the income gap between communities is fairly large. Splines account for this non-linear relationship.





Above are the diagnostic plots for the class model (Figures 1 and 2) and for my model (Figures 3 and 4). Figures 1 and 3 graph of the expected number of thefts by community area per day and per year, respectively. As expected, the number of reported thefts per year by community area decreases with trends similar to the per day trends. Figures 2 and 4 graph the estimated mean against the estimated variance of the number of thefts based on 8 groupings of the fitted values. My mean and variance are still linearly related, satisfying the quasi-Poisson regression condition; however, my data is much more over-dispersed than the data used in class. This may be due to the fact that the median household income by year has nearly the same number of unique values as the number of community areas.

Overall, it is difficult to discern whether adding the median household income improves the model or not. The over-dispersion of the model is a large concern; however, including the median income variable helps distinguish community areas outside of being a categorical variable.

Sources:

American Community Survey. <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_S1903&prodType=table>

Census Geocoder: <https://geocoding.geo.census.gov/geocoder/geographies/addressbatch?form>

1. These years were the only consecutive years that were available from the ACS. [↑](#footnote-ref-0)
2. Note, the Census Geocoder requires a specific file format in order to properly match the addresses. Additional cleaning of the addresses was performed to match the requirements. [↑](#footnote-ref-1)