Public Health Indicators Associated with Teen Birth Rate in Chicago

Francisca Moya Jimenez ‘22, Data Science Major Capstone

Background and Research Questions

Despite the overall decline in teen birth rate in the United States since 1991, Chicago has a teen birth rate higher than the national average of 39.1 births per 1,000 women ages 15 to 19 years old [1]. **This capstone project seeks to identify relevant public health indicators that are associated with teen pregnancy in the community areas in the City of Chicago**, which are dependable geographic units that divide the city into 77 areas from which the city government consistently collects data [2].

This study utilizes public health statistics and socioeconomic statistics collected from the 77 areas in the City of Chicago to fit a first-order model with particular interest in the following research questions:

(1) What variables are most relevant to predicting teen birth rate for community areas in Chicago?

(2) Are the identified predictors positively or negatively associated with teen birth rate given other predictors in the model?

Data

Description

The data was extracted from the Chicago Data Portal [3], which features data collected by the City of Chicago’s government. The final dataset merges several datasets available in the portal, which include:

- A list of active affordable housing developments supported by the City of Chicago
- A public health statistics dataset put together by the Chicago Department of Public Health, which contains a selection of 27 indicators of public health significance
- A COVID-19 community vulnerability index (CCVI) dataset, as calculated by the Department of Public Health
- A life expectancy dataset.

The resulting dataset has 37 variables, such as number of affordable housing buildings and diabetes-related deaths per 100,000 people for each area.

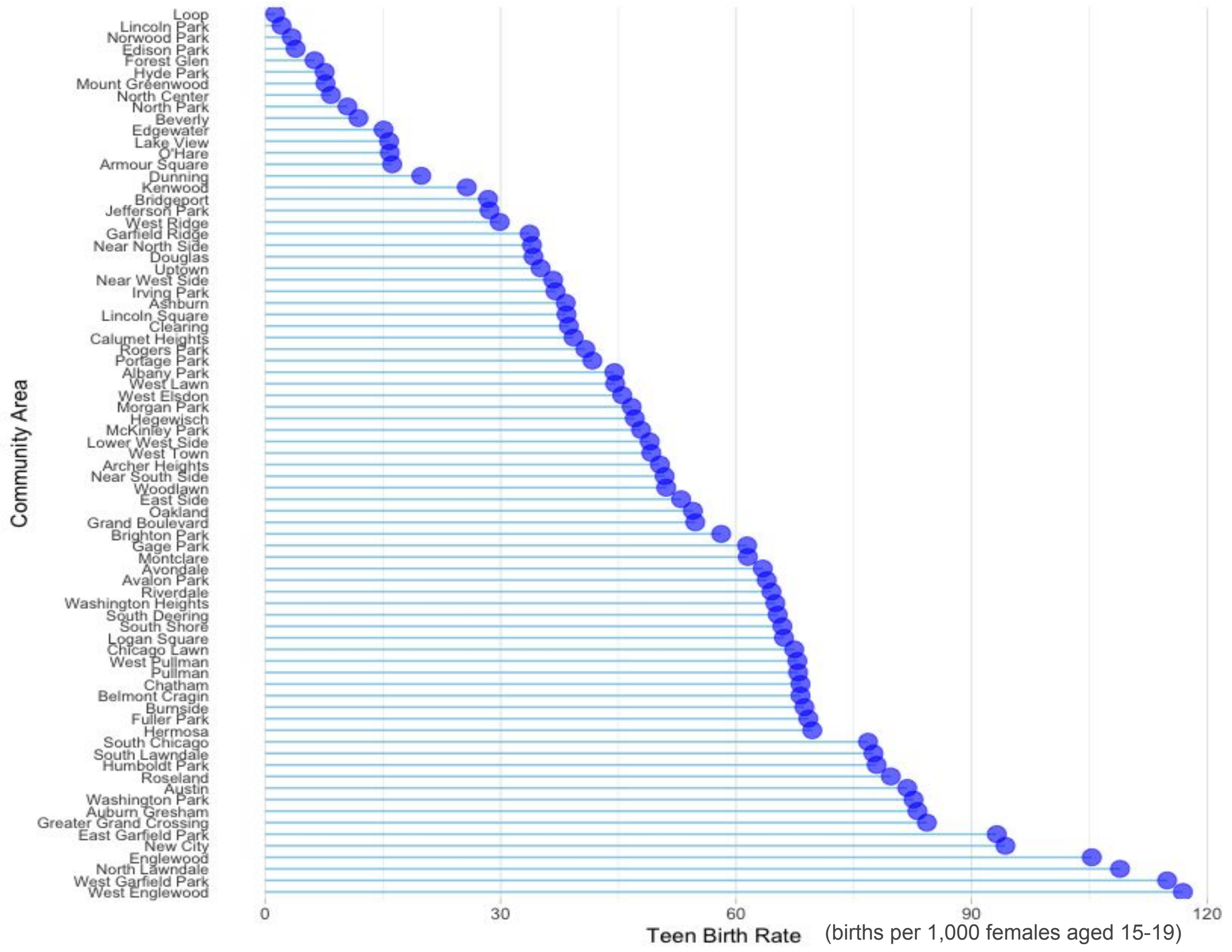

Figure I. Barplot of Teen Birth Rates by Community Area in Chicago.

Cleaning

No data cleaning was needed aside from merging the four datasets together using the community areas. The following variables had missing data:

- **Number of childhood blood lead level screening.** Data were insufficient to calculate the indicator for Riverdale [3]. Mean imputation was used since the distribution of the variable was fairly symmetric.
- **Childhood lead poisoning percentage.** Data were insufficient to calculate the indicator for Riverdale. Median imputation was used since the variable had a skewed distribution.
- **Rate of gonorrhea in males.** This variable had over 15% missingness, and it was removed from the set of predictors.
- **Rate of gonorrhea in females.** The variable was removed from the set of predictors as it had over 15% missingness.

Data Modeling: First-Order Model

A multiple linear regression was ran to identify the public health factors associated with teenage birth rate in Chicago. Stepwise elimination was performed based on a threshold of 10 for the variance inflation factor (VIF), and 7 variables were removed due to multicollinearity.

All-subset selection and automatic selection methods were used to perform variable selection. The BIC and Mallow’s Cp criteria yielded the same model (R-squared=0.89, R-squared adjusted = 0.88, p-value from the general F-test < $2.2*10^{-16}$). The model chosen under BIC/Mallow-CP had the lowest 5-fold cross-validation score with a value of 11.73 births per 1,000 teenagers aged 15 to 19. The best first-order model according to the cross-validation score has the following form:

$$\widehat{Teen\ Birth\ Rate} = 13.44 + 2.83 \cdot\ BirthRate - 0.77\ \cdot\ PrenatalCareInFirstTrimester + 0.39 \cdot\ BreastCancerInFemales + 0.19 \cdot\ DiabetesRelatedDeaths + 0.57 \cdot FirearmRelatedDeaths + 1.42 \cdot\ CrowdedHousing + 0.82 \cdot\ Unemployment$$

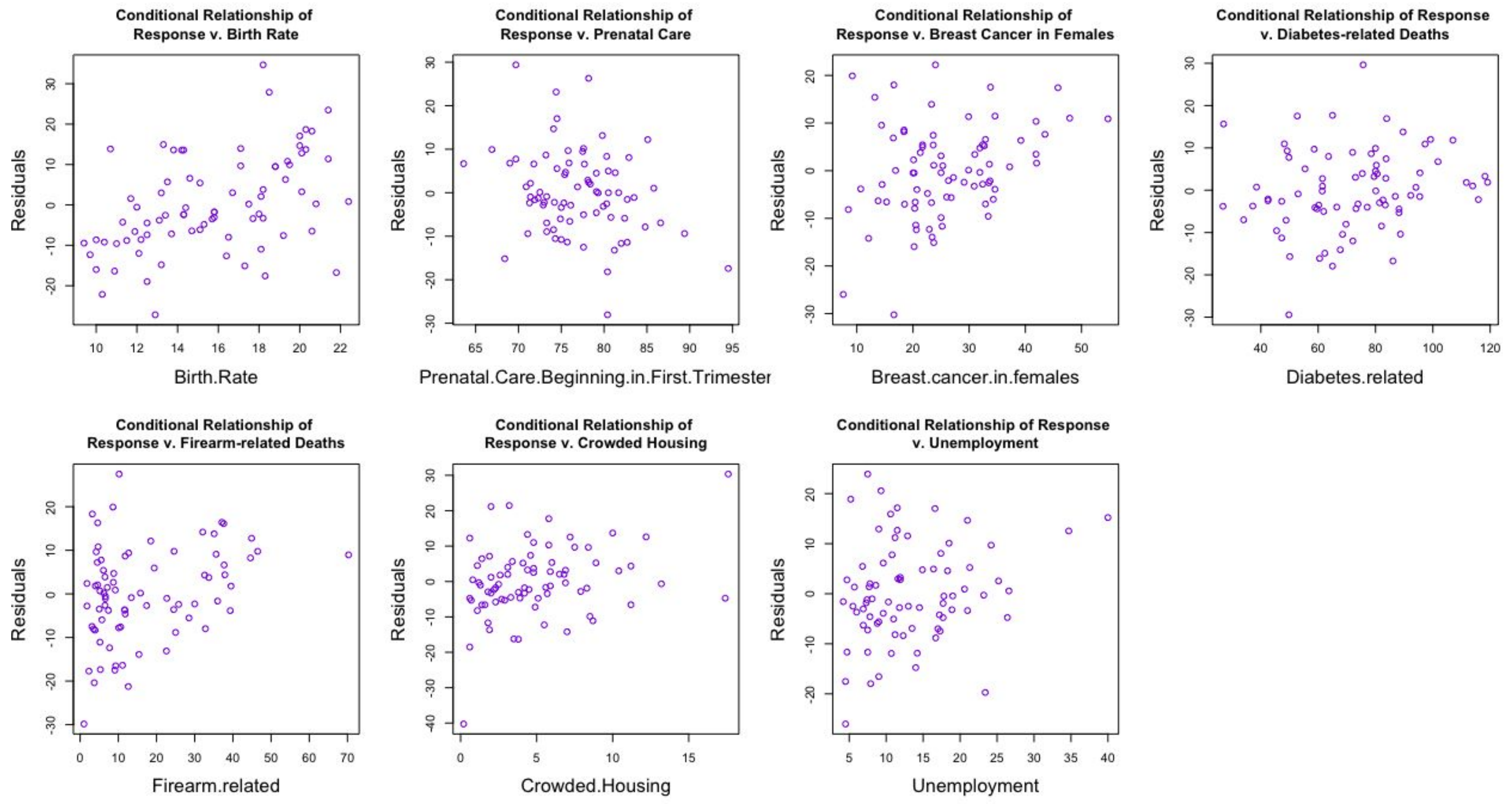

Figure II. Relationship between Teen Birth Rate and Each Predictor in the First-order Model Conditional on All Other Predictors. The scatterplots show the residuals of the model predicting teen birth rate on all other predictors but the predictor on the x-axis.

Data Modeling: Interaction Model

Interaction terms were added to the first-order model using a stepwise process with the first-order model as a lower boundary and the model with all interactions as the upper boundary. The interaction model minimizes the AIC criteria and has the lowest cross-validation score (10.34 births per 1,000 teenagers aged 15 to 19). The model has an R-squared of 0.90, R-squared adjusted of 0.89, p-value from the general F-test < $2.2*10^{-16}$. It includes all of the variables in the first order model as well as 2 interaction terms and it has the following form:

$$\widehat{Teen\ Birth\ Rate} = 167.75 - 2.45 \cdot BirthRate - 2.56 \cdot PrenatalCareInFirstTrimester - 0.20 \cdot BreastCancerInFemales - 1.98 \cdot DiabetesRelatedDeaths + 0.15 \cdot FirearmRelatedDeaths + 1.48 \cdot CrowdedHousing + 0.80 \cdot Unemployment + 0.02 \cdot BreastCancerInFemales \cdot FirearmRelatedDeaths + 0.03 \cdot PrenatalCareInFirstTrimester \cdot DiabetesRelatedDeaths$$

Plots of interactions are shown below. Interactions that are not part of the final model are omitted.

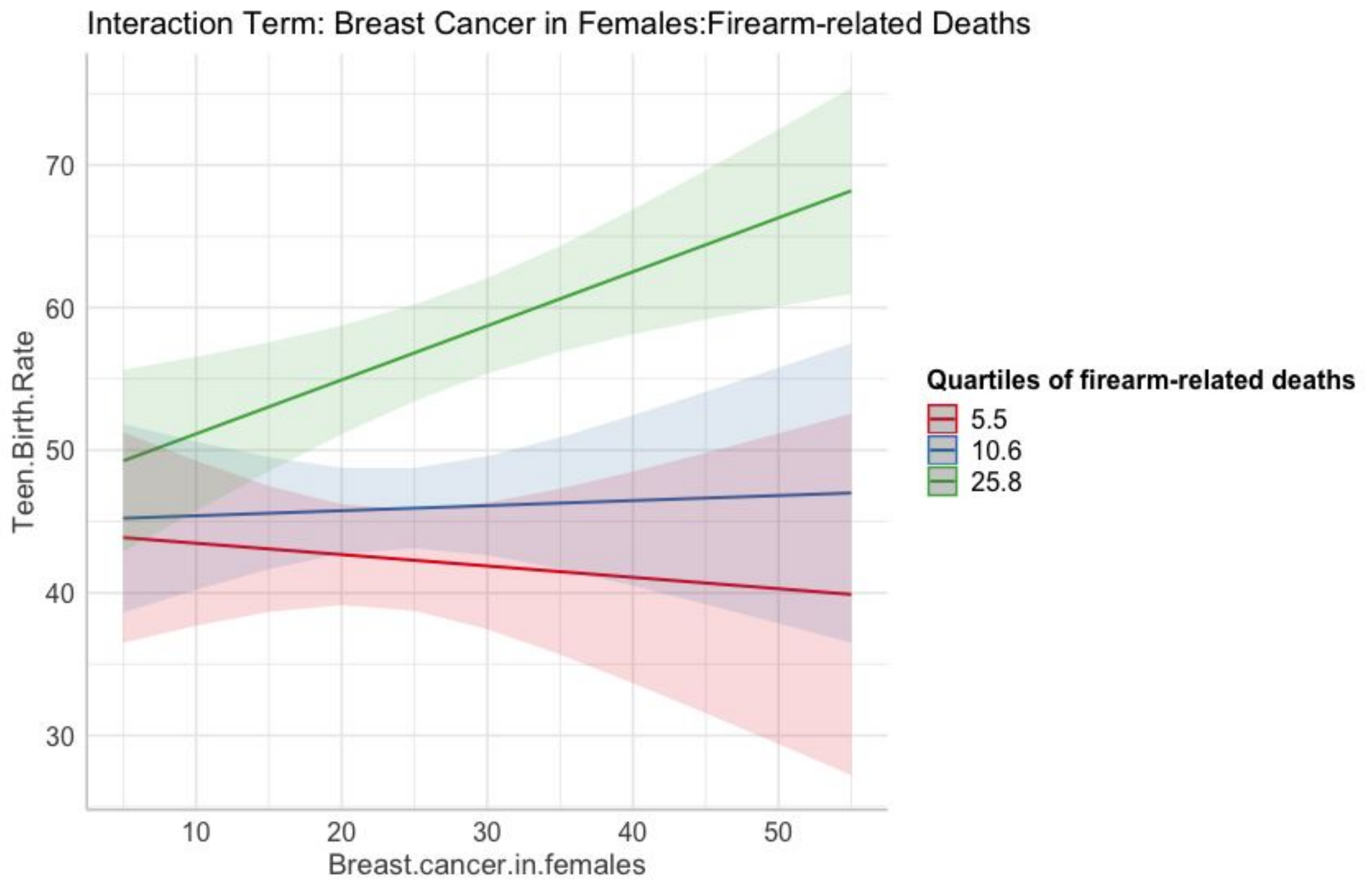

Figure III. Interaction Term between Breast Cancer in Females and Firearm-Related Deaths in the Best Interaction Model. Axes limits were calculated based on the quartiles of firearm-related deaths.

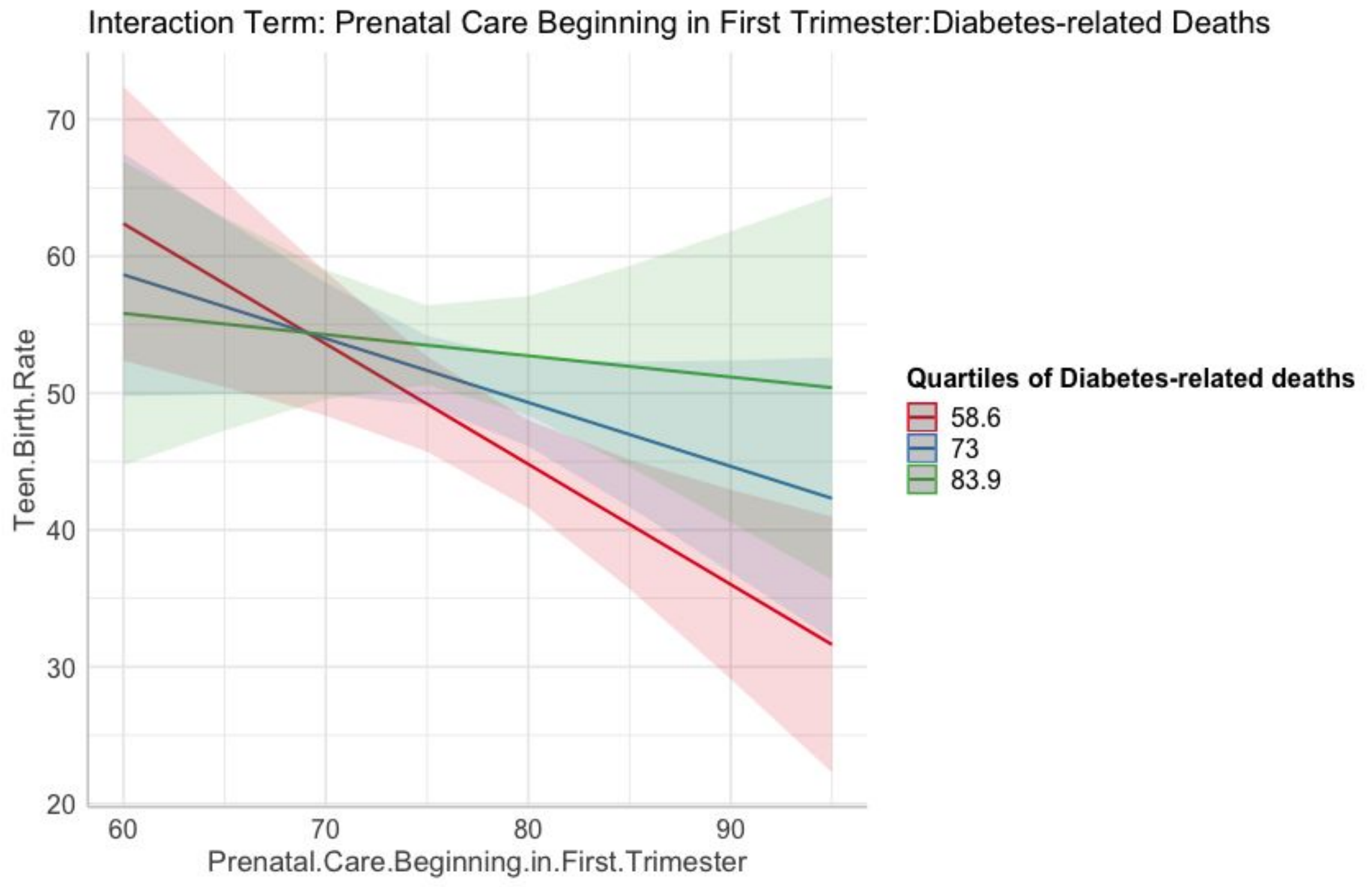

Figure III. Interaction Term between Prenatal Care Beginning in First Trimester and Diabetes-Related Deaths. Axes limits were calculated based on the quartiles of diabetes-related deaths.

Results

- The first-order model shows a positive relationship between teen birth rate and the predictors birth rate, breast cancer in females, diabetes-related deaths, firearm-related deaths, crowded housing and unemployment. Prenatal care beginning in first trimester is negatively associated with teen birth rate. At higher levels of firearm-related deaths, teen birth rate and breast cancer in females seem to have a strong positive linear association.
- The City of Chicago could use this information to develop social programs that address socioeconomic and public health issues such as crowded housing and unemployment, which are affecting communities at high risk of teenage pregnancy.
- Despite the differences in income between areas with the highest teen birth rate ($10,559 for the West Englewood) and lowest birth rate ($67,699 for the Loop) [4], income was not chosen as a predictor in the variable selection process. This suggests that socioeconomic inequality in Chicago can be better summarized by looking at the public health conditions of these areas.

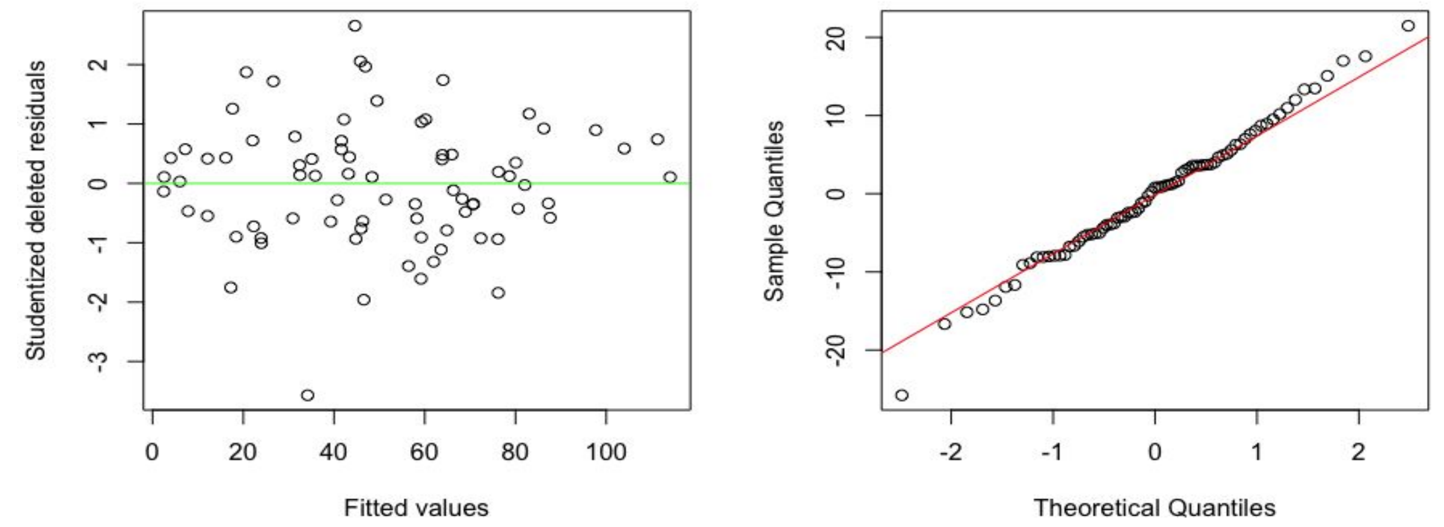

Figure IV. Visualizations to evaluate the linear regression assumptions for the first-order model. The studentized deleted residual versus fitted value plot on the left and the Q-Q plot on the right do not show violations to the equal variance assumption or major departures from normality.

Data Ethics and Limitations

According to the CDC, ethnic and racial minorities are at most risk for teenage pregnancies [1]. However, variables such as race or ethnicity are not included in the model to avoid perpetuating biases about racial and ethnic minorities regarding teen pregnancy. This project seeks to identify a combination of factors that encapsulate some of the social determinants of health that often affect racial and ethnic minorities.

Some limitations of this project include:

- The data used to fit the model includes information that partly goes back to 2010, which might mean that the data is no longer current.
- The geographic proximity of the community areas may mean that the independence assumption may not be perfectly met.

References

[1] “About Teen Pregnancy.” *Centers for Disease Control and Prevention,* Centers for Disease Control and Prevention, 15 Nov. 2021, https://www.cdc.gov/teenpregnancy/about/index.htm
[2] Hrobowski, Gianni, and Noelle Sanzeri. “Teen Birth Rates Hit Historic Lows in Chicago.” *Teen Pregnancy Rates in Chicago from 2010-2020*, The Red Line Project, 8 May 2020, http://redlineproject.org/teen_births.php.
[3] “Chicago's Neighborhoods.” *The University of Chicago–Chicago Studies*, The University of Chicago, 31 Aug. 2021, https://chicagostudies.uchicago.edu/neighborhoods
[4] “Codebook: Selected Public Health Indicators by Chicago Community Area.” City of Chicago Open Data Portal, https://data.cityofchicago.org/api/assets/2107948F-357D-4ED7-ACC2-2E9266BBFFA2