Statistical Analysis of Life Expectancy and Community Health in the City of Chicago

Eduardo Martinez

Illinois Institute of Technology

Abstract

Health status can be broken down into three key components: physical health, mental health, and social health. Life expectancy is a key indicator used to assess a person's physical health. Gender and racial/ethnic health disparities exists nationally and locally among the 77 community areas (CAs) in the City of Chicago. Health policies are critical part of governments at both the national and local level. However, there is not a systematic and consistent way of tracking health data/records. It may be beneficial to rely on historic and recent community level health data when proposing health policies. One key source of such data is the American Community Survey (ACS). In this study, indicators measured in the ACS are used to predict life expectancy based on data for the CAs in Chicago, IL. While each model yielded very small values for MSE and RMSE in both the *training set* and the *test set*, no subset model significantly outperformed the other. As a result, it is unclear which indicators should be the focus of potential future health policies. Future studies should consider different data sets.

Introduction

Health is a key factor considered when evaluating the overall quality of a person's life. Healthy individuals typically have favorable qualities (e.g., longer life expectancy), while unhealthy individuals are often diagnosed with severe health conditions (e.g., diabetes and high cholesterol) [17]. The World Health Organization (WHO) identifies physical, mental, and social well-being as the main components of health [27]. In addition, the WHO has been considered adding spiritual health as a fourth component [20].

Generally, people from various countries perceive the environment as the key determinant of a person's health status [20]. Assessing mental and social health relies on personal accounts, which are subjective in nature. For example, a psychologist meets with patients consistently over long periods of time in order to gain an understanding of their mental and social well-being. By comparison, physical health can be objectively measured via quick and accurate methods. For example, blood test can be used to quantitatively measure the concentration of glucose and cholesterol in a patients blood; further, there are consistent scales used throughout the population to determine if a patient has high cholesterol or diabetes [18].

The WHO's main objective is to help each person in the population maximize their health status [27]. Similarly, improving public health is a priority among government agencies and officials, who often consider data when making critical health policies at a national level [13]. For instance, when Barack Obama was President of the United States, he passed the Affordable Care Act (ACA), which is still being used in 2022. While a large proportion of the U.S. population has supported the ACA, many Americans strongly oppose the program [13]. At the moment, there is not a nonpartisan solution for improving public health. Analyzing public health data at a deeper level may help uncover a

The United States is one of the largest countries in both total population and area. Attempts at solving public health issues may not be as effective at national level compared to a community level approach. For instance, the health status of Idaho City, a small town in Idaho with an estimated 520 total residents of which approximately 90.2% are White (Non-Hispanic) [7], is likely different than residents in Chicago, IL where there is an estimated 2.7 million residents of which approximately one-third are White (Non-Hispanic) [6]. National data and health statistics do not provide enough details on the status of community health. Furthermore, national averages mask local level disparities [10].

Instead, community level data should be considered. Health records can provide a crucial details about historic community health and provide insights on future health outcomes [25]. Historic and recent health records are already used be physicians when deciding how to best treat patients [25]. Local governments should consider systematic use of health data to help guide health policies.

Since physical, mental, and social well-being are the main components of health [27], it is difficult to pinpoint select indicators representative of a person's health status. Typically, average life expectancy is used as the main indicator of health; life expectancy data helps guide health policies and is one of the key health variables considered in research studies [15].

Geographic and racial/ethnic health disparities have existed within Chicago's community areas [19]. Recent studies indicate that these disparities persist [2, 11]. There is evidence of discrimination against racial minorities when they seek healthcare [2]. Racial discrimination is associated with poor health outcome [2]. Consequently, racial minorities do not receive adequate healthcare at a equal rate compared to non-minorities.

Premature mortality is defined as deaths before age 65 [12]. Annual data of Chicago's 77 community areas (CAs) from 2011 to 2015 showed a significant negative relationship between age-adjusted premature mortality rates and the diversity of a CA [12]; that is, age-adjusted premature mortality rates were

highest among the least diverse neighborhoods, where diversity was measured using the Index of Concentration at the Extremes [12].

Disparities in birth outcomes have existed for at least 20 years [26]. Moreover, there are indirect factors that may be associated with negative birth outcomes. From 2008 to 2013, there was yearly average of 61.5 (SD = 40.3) crime incidents per 1000 persons for Chicago's 77 CAs [16]. Increased crime rates in a CA was significantly related to increased rates of adverse pregnancy outcomes [16].

Gender disparities in pay have existed and continue to exist with no clear end in sight [3]. Consequently, women are disproportionately affected by poverty, which is associated with poor health [3]; these disparities become more pronounced among other gender minorities (e.g., transgenders). It is unclear if these gender disparities persist in regards to life expectancy.

In Chicago, Hispanics had the highest life expectancy in 2010, living an average of 84.6 years [10]. On the other hand, non-Hispanic Blacks (Blacks) had the lowest life expectancy, averaging about 71.7 years [10]. Examining indicators that such CAs perform well and poorly in can provide insight on how to best support such CAs.

Mortality rates are another indicator of health status. Disparities in mortality rates existed in 2004 [26]. More recently, an analysis of Non-Hispanic White (White) and non-Hispanic Black (Black) all-cause mortality rates in the United States (U.S.) uncovered disparities between these two races [1]. From 2016 to 2018, all-cause mortality rates were 24

Long-term changes in health disparities have not decreased nationally nor locally in Chicago. Data on multiple health indicators measuring mortality rates, birth outcomes, and disease incidence from 1990 to 2010 indicate that the general trend has remained constant; in fact, some indicators have widened disparities [11]. This means efforts to decrease health disparities have been ineffective despite the fact they are often areas of emphasis in health policies [13, 11, 3].

Statistical analysis of community health data can help identify select indicators that are most influential determinants of a person's health status and life expectancy.

Methods

Community-level data corresponding to each of the 77 community areas (CAs) in Chicago, IL was obtained and curated. The final dataset measured *average life expectancy*, which was the response variable of interest, and 27 other health-related indicators, which were the predictor variables considered (28 variables total). Regression analysis methods were applied to identify which of the 27 predictor variables are best at predicting *average life expectancy*.

If there are variables that are significantly better at predicting *average life expectancy*, then such indicators have a particularly strong relationship with life expectancy [5]. By contrast, marginal differences between a set of predictors suggests several predictors have strong relationships with life expectancy. Although identifying predictors with an especially strong relationship with life expectancy can provide useful insight, regression models do not imply causation.

Demographics

The population demographics for the entire City of Chicago and for the 77 CAs can be provided in Table 1. These measures are not equivalent because the City of Chicago data considers all households sampled in Chicago. In contrast, the statistics for the 77 CAs computes the average for each CA first [6]. Thus, CAs are equally weighted even if the number of households sampled in each CA differ. In other words, statistics for the 77 CAs represent "averages of averages".

Table 1: Demographics (Based on 5-year averages from 2015-2019)

		All 77 Community Areas (CAs)				
	Chicago	**Mean**	**SD**	**Min**	**Median**	**Max**
Total Population	2,709,534	34,240	22,182	1,997	28,332	93,941
Gender						
% Male	48.64	47.78	3.4	35.16	48.04	57.97
% Female	51.32	52.22	3.4	42.03	51.96	64.84
Race/Ethnicity						
% Non-Hispanic White	33.28	28.03	26.76	0.5	15.99	83.51
% Non-Hispanic Black	29.19	38.07	39.07	0.37	15.83	96.54
% Hispanic or Latino	28.79	26	26.91	0.09	14.11	89.18
% Asian or Pacific Islander	6.56	5.95	9.58	0	2.47	60.79
Age Group						
% Children (0-17 years)	20.9	21.87	6.08	5.11	22.1	41.23
% Young Adults (18-39 years)	37.33	33.79	8.78	21.32	31.41	64.02
% Middle-Aged Adults (40-64 years)	29.34	30.3	3.95	20.41	30.63	38.12
% Seniors (65 and older)	12.44	14.03	4.6	5.43	13.74	27.94

Sources: ACS Table B01001; Decennial Census Table P012

The Data

American Community Survey

Data for 26 of the 27 predictor variables was collected by the U.S. Census Bureau's annual American Community Survey (ACS), administered annual households sampled throughout the nation. (Annual version of the ACS can be found in [22]). Although different versions were administered each year, only minor changes if any were made. Response data for each household is averaged and organized by various geographic areas, including states, counties, tracts, and block groups, which are defined in the ACS handbook for All Data Users [24]. In terms of geographical area, CAs are smaller than counties but larger than Census tracts [23].

Community-level data was obtained from the Chicago Health Atlas, which is a data portal created by the Chicago Department of Public Health (CDPH) and the Population Health Analytics Metrics Evaluation (PHAME) Center at the University of Illinois Chicago (UIC) [6]. From the Chicago Health Atlas provided data for each CA and the City of Chicago overall was acquired. The Chicago Health Atlas was able to derive data for each CA because tracts are drawn such that they do not overlap with multiple CAs [23].

The specific ACS data used in this study consisted of 5-year annual average from 2015-2019 for each of the 77 CA in Chicago, IL [6]. 5-year averages provide better accuracy than single years because only small subsets of all households in each CA completed the survey each year [24]. In other words, the cumulative sample over the over the 60 month period is more representative of all households in each CA compared to a smaller samples surveyed each year.

Life Expectancy

Data for *average life expectancy*, was provided by the Illinois Department of Public Health (IDPH) via Death Certificates Data Files. Community-level data was also provided by the Chicago Health Atlas [6].

Although these values are also averaged over five years, the annual values are exact and not estimated like the ACS data [6]. The exact ages of each person's death is provided in Death Certificates. *average life expectancy*is calculated based on the average age of people who died in a given year.

Federally Qualified Health Centers

The only predictor variable considered not sourced by the ACS or organized in the Chicago Health Atlas was the number of federally qualified health centers (FQHC). The source of these quantities is the Health Resources & Services Administration (HRSA) Data Warehouse. The Heartland Alliance's Social IMPACT Research Center organized data for each CA in its data portal [9].

FQHC is the only variable that is measured in counts rather than averages. The exact number of FQHC is known (no estimation required), and the number did not change from 2015-2019 [9]. This means the 5-year average is same as each annual value from 2015-2019.

Mathematical Background

Multiple linear regression (MLR) using ordinary least squares (OLS) was used to fit models to the community-level data collected by the ACS. Details are explained throughout the paper, and complete details can be found in *Regression Analysis by Example* by Chatterjee & Hadi [5] and *The Elements of Statistical Learning* by Hastie et al. [8].

Let Y be a vector containing the true values of the response variable, *average life expectancy*, and let $X = (X_1 \mid X_2 \mid \ldots \mid X_p)_{n\times p}$ be a matrix containing the n rows of data for each of the p predictor variables in a regression model.

A general MLR equation is written

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \ , \qquad (1)$$

where $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_p$ are unknown constant coefficients such that β_0 is an intercept term and β_j for $j \in 1:p$ are separate coefficients multiplied to the associated predictor variable [5].

ε is a random variable normally distributed with mean 0 and variance σ^2. That is, $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2)$

Equation 1 can be expressed in matrix notation by defining $\mathbf{X} = (\mathbf{1} \mid X_1 \mid X_2 \mid \ldots \mid X_p)_{n\times b}$ [5], where $b = p + 1$ is the number of coefficients in a model:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

The true value of the coefficients $\boldsymbol{\beta}$ are unknown, but they can be estimated using observed data, x_i. The estimated coefficients are denoted by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$. Typically, a dataset is uniquely split into a *training set* and *test set*, where model coefficients are estimated using the *training set*, and the model's accuracy can be evaluated using the *training set* [5]. (Methods of evaluating a model's accuracy are described in Model Accuracy.)

OLS is the main method used to estimate $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ are the values that minimize

$$S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{y} - \boldsymbol{X\beta})^T(\mathbf{y} - \boldsymbol{X\beta}) \qquad (2)$$

If $\mathbf{X}$ has full rank and it is invertible, then a unique solution is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y. \qquad (3)$$

The resulting estimated OLS regression model is expressed using the following equation:

$$\hat{y} = X\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p \ , \qquad (4)$$

where $\hat{\beta} \sim N(\boldsymbol{\beta}, (X^T X)^{-1}\sigma^2)$. Since the true value for σ^2 is usually unknown, an unbiased estimator of σ^2 is used instead:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ , \qquad (5)$$

where $\hat{y}_i$ denoted values fitted by a model and y_i are the true values of a predicted variable [8]. Note, as an unbiased estimator, $E(\hat{\sigma}^2) = \sigma^2$ [8].

Let e_i denote a model's residuals/errors such that $e_i = y_i - \hat{y}_i$. Then, Equation 5 can be written as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(e_i)^2}{n-p-1} = \frac{\overbrace{\boldsymbol{e}^T\boldsymbol{e}}^{\text{Matrix Form}}}{n-p-1} \qquad (6)$$

† *Full variable names in* Table 2

where I_n is an identity matrix, $e_i \sim N(0, \sigma^2(I_n - H)$, and H is referred to as the "hat" or "projection" matrix [5, 8]:

$$H = X(X^T X)^{-1} X^T \tag{7}$$

By substituting Equation 3,

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy. \tag{8}$$

Variance Inflation Factor (VIF)

VIF is a measure used to detect collinearity among a set of p predictors [5]. Collinearity arises when two or more predictors are highly correlated with each other. The VIF of the j^{th} predictor can be calculated using the following formula:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \ , \ j = 1, \ldots, p \ , \tag{9}$$

where $R_j^2 = 1 - \frac{\text{SSE}}{\text{SST}}$ corresponds to the the regression model such that X_j is the response variable predicted using all other variables.

If $\text{VIF}_j > 10$, then collinearity will likely be a problem in a model [5]. Suppose a pair of predictors in a regression model exceed the cutoff value, then the variance of one predictor may be accounted for by the other. As a result, it may be beneficial to select a smaller subset model to avoid collinearity issues.

Outlier Detection

Standardized residuals, point leverage, and Cook's distance were the primary statistics used in identifying and analyze potential outliers. All details described in this section are from *Regression Analysis by Example* by Chatterjee & Hadi [5]. Each CA corresponds to a row in the dataset. If an outlier is detected, it means the associated CA yields extremely different values compared to the other CAs.

Standardized Residuals (r_i):

$$r_i = \frac{e_i}{s(\mathbf{e})} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \ , \tag{10}$$

The exact value for the residual standard error (RSE), σ is unknown, so the unbiased estimator $\hat{sigma}$ is used instead. By dividing by $\hat{sigma}$, the resulting random variable is assumed to have a standard normal distribution:

$$e_i \sim N(0, \sigma^2(I_n - h_i)) \implies r_i \sim N(0, 1) \tag{11}$$

Based on the empirical rule of a normal distribution, 95% of values fall within 2σ and 99.7% fall within 3σ [21]. Given $r_i \sim N(0, 1)$, $2\sigma = 2$ and $3\sigma = 3$. The *training set* contains data for $n_1 = 45$ CAs; this means around 2.25 CAs should have $|r_i| > 2$ and 0.135 CAs should have $|r_i| > 3$. The value $0.135 \approx 0$ suggests CAs with $|r_i| > 3$ are likely outliers. Moreover, if there are more than 5 CAs with $|r_i| > 2$ (approximately twice the expected amount of 2.25), then OLS assumptions are violated. That is, the standardized residuals are not sampled from a standard normal distribution.

Point Leverage (h_i):

A point's leverage is given by the associated diagonal element of H defined in equation (7) and they typically denoted h_i instead of H_{ii}.

Let $\bar{x}_1$ be the mean of an arbitrary predictor variable used to predict *average life expectancy*. For each data point in the vector x_1 defined as $x_1 = (x_{1,1} \quad x_{2,1} \quad \ldots \quad x_{n,1})^T$, leverage h_i measures the distance between $x_{1,i}$ and $\bar{x}_1$ for each $i \in 1 : n$.

Given values of a predictor variable x_i (vector with n rows), its leverage can be calculated as follows:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{12}$$

If h_i very small, then the observed response y_i has small influence value of $\hat{y}_i$, while large values for h_i are more influential. Influential data points will yield much different regression results when they are removed or added to a model. Chatterjee & Hadi *recommended* the following cutoff value for determining if a point has high leverage:

$$h_i > \frac{2b}{n} = \frac{2(p+1)}{n} \tag{13}$$

Cook's Distance (D_i):

$$D_i = \frac{r_i^2}{p+1} \times \frac{h_i}{1 - h_i} \ , i = 1, \ldots, n \tag{14}$$

† *Full variable names in* Table 2

Similarly to leverage, Cook's distance is used to identify highly influential data points. Large values for Cook's distance indicate that a point is highly influential. Specifically, removing it from a regression model will result in a much different model when refitting the remaining variables. Generally, a point is considered highly influential is greater than the following cutoff value: Chatterjee & Hadi *recommended* the cutoff value defined below:

$$D_i > F(p = 0.5,\ df_1 = p+1,\ df_2 = n-p-1)\ , \quad (15)$$

where F denotes the F-distribution with $p+1$ and $n-p-1$ degrees of freedom.

It is important to emphasize, the cutoff values for Cook's distance and high leverage data points are recommendations rather than strict cutoff values [5]; different sources might recommend different cutoff values. Making decision solely based on the recommended values can be erroneous [5, 8].

In addition, Cook's distance, leverage values, and standardized residuals of each data point (CA) should be compared to all other points (CA). It is particularly useful to compare values visually (e.g., graphs, charts, tables) because outliers should stand out. A specific data point "stands out" when few (if any) points have values nearby; conversely, a data point does not stands out if there are several other points nearby [5].

If a data point exceeds the cutoff values and visually stands out, then it likely is an outlier [5]. However, if a data point exceeds the cutoff values but does not visually stand out, then it may not be an outlier, especially when there are many other points with much larger values. Before deciding to remove an outlier, it is crucial to analyze how much a regression model changes when the potential outlier is removed, since influential data points should yield significant changes in the resulting regression model [5].

Model Selection

The details described in this section are based on Chapter 11 of *Regression Analysis by Example* by Chatterjee & Hadi [5] and Chapter 7 of *Elements of Statistical Learning* by Hastie et al. [8].

To reduce overfitting and potential issues caused due to collinearity in the data, best subset models were selected using the following model selection statistics and information criteria:

1. Adjusted Multiple Correlation Coefficient (R^2_{adj}),
2. Akaike Information Criterion (AIC),
3. Bias Corrected AIC (AIC_c),
4. Bayes Information Criterion (BIC), and
5. Mallow's Statistic (C_p).

In each equation below, the sum of squared errors, $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$. The terms "residuals" and "errors" are used interchangeably but refer to the same values, e_i. The abbreviation SSE is used to differentiate it from the sum of squares for regression (SSR), where $\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$. SSE and SSR are related to the sum of squares in total (SST) such that $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{SSR} + \text{SSE}$. Recall, the number of coefficients, including the intercept, in a model is denoted b such that $b = p+1$.

Note, R^2_{adj} is optimized when its value is maximized. All other criteria are optimized when they are minimized [5, 8].

Adjusted Multiple Correlation Coefficient (R^2_{adj}):

$$R^2_{adj} = 1 - \frac{\text{SSE}/(n-b)}{\text{SST}/(n-1)} = 1 - \frac{\text{SSE}\cdot(n-1)}{\text{SST}\cdot(n-b)} \quad (16)$$

The unadjusted $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$. Values for both R^2_{adj} and R^2 range between 0 and 1 such that the model with the largest value fits the data most adequately [5]. However, R^2 does not account fo the number of predictors, p, in a model. In fact, R^2 increases as p increases. Consequently, R^2 will likely result a model overfitting the data. On the other hand, each predictor variable added causes R^2_{adj} to decreases, assuming all other terms remain constant. Thus, R^2_{adj} penalizes models for each added term.

Based on the equation for R^2_{adj}, it appears that the rate at which the penalty changes is quiet small even as $p \to \infty$. Consider the isolated penalty component $\frac{1}{n-p-1} = \frac{1}{44-p}$. Then, for $p = 0, 5, 10, 15, 20$ the associated values are $0.023, 0.026, 0.029, 0.034, 0.042$. A model with 20 predictors will severely overfit the data, and it is hypothesized that R^2_{adj} does

† *Full variable names in* Table 2

not effectively prevent overfitting compared to the remaining information criteria.

Akaike Information Criterion (AIC):

$$\text{AIC} = n \log(\text{SSE}/n) + 2b \tag{17}$$

The penalty term is $2b = 2p + 2$. AIC encounters the same potential issues as in R^2_{adj}; the penalty component increase by 2 for each added predictor even as $p \to \infty$. Consequently, it is not expected that AIC will prevent overfitting significantly better than R^2_{adj}. For $p = 0, 5, 10, 15, 20$ the penalty values are $2, 4, 6, 8, 10$, respectively.

Bias Corrected AIC (AIC_c):

$$\text{AIC}_c = \text{AIC} + \frac{2(b+2)(b+3)}{n-b-3} \tag{18}$$

Given 45 CAs in the *training set*, the additional penalty term can be expressed as $\frac{2p^2+14p+24}{41-p}$. Notice, the numerator increases quadratically while the denominator decrease linearly. This means that the penalty rapidly gets more severe as p increases.

In fact, for $n = 45$ and $p = 0, 5, 10, 15, 20$ the *added* penalty values are $0.6, 4, 12, 26, 53$, respectively, appearing to increase exponentially.

Bayes Information Criterion (BIC):

$$\text{BIC} = n \ln(\text{SSE}/n) + b \ln(n) \tag{19}$$

The only difference between AIC and BIC is the penalty term. Since $\ln(n = 45) \approx 3.8$, so each added predictor is penalized to a greater extend compared to AIC, but the rate of change is constant. Note, the penalty term for BIC is larger than the AIC penalty when $n > 8$ [5].

Mallows Statistic (C_p):

$$C_p = \frac{\text{SSE}_{RM}}{\hat{\sigma}^2_{FM}} + (2b - n) \ , \tag{20}$$

where SSE_{RM} is the SSE of reduced (subset) model, while $\hat{\sigma}^2_{FM}$ is the estimated residual standard error for the full mode; equivalently, $\hat{\sigma}^2_{FM}$ is the means squared error (MSE) of the full model.x Thus, $\hat{\sigma}^2_{FM}$ is a constant. The the penalty term, $(2b - n)$, is unlike the ones in the previous criteria because the entire equation takes into account bias and variance. Accordingly, C_p is minimized for values C_p close to p.

Model Accuracy

The most common measures used to evaluate a model's accuracy are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

$$\text{MSE} = \frac{\text{SSE}}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{21}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{22}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{23}$$

Smaller values for these statistics indicate a model is more accurate. However, these statistics tend to get larger as the predicted variable gets larger. For larger values,

MSE and RMSE terms are related to the each of the model selection criteria such that the best subset model of size b is the only that minimizes MSE and RMSE. This is not surprising given Ordinary Least Squares (OLS) regression was applied.

Final Sample

Initial analysis resulted in two CAs (Burnside and Fuller Park) being dropped from the final sample as they were considered extreme outliers. Note, complete details on this process quickly explained below will be given for the 75 final CAs considered.

The outlier detection procedures outlined in Outlier Detection were completed after selecting the best subset models based on the criteria in Model Selection. The standardized residuals for Burnside and Fuller Park were greater than two standard deviation in absolute value for every model considered; Burnside's point leverage was high in every model considered, while Fuller Park's Cook's distance indicated that the CA is a highly influential observation. Moreover, Burnside and Fuller Park had an average population of 1,997 and 2,397 residents, respectively, which are the two smallest CAs overall. Further, they often had some of the largest standard errors based on ACS estimates [6, 24].

† *Full variable names in* Table 2

Burnside and Fuller park are believed to be outliers, so they were excluded from this analysis.

The final sample consisted of 75 CAs randomly split into a *training set* and *test set* of size $n_1 = 45$ and $n_2 = 30$, respectfully.

Table 2: Statistics for All Regression Variables (Based on 5-Year Averages from 2015-2019)

			Final 75 Community Areas (CAs)				
Code	**Variable**	**Chicago**	**Mean**	**SD**	**Min**	**Median**	**Max**
LE	**Average Life Expectancy**	**77.29**	**76.98**	**4.37**	**68.02**	**79**	**82.94**
FQHC	Number of Federally Qualified Health Centers	188	2.49	3.09	0	1	12
MHI	Median Household Income	61,784	62,874	28,831	16,501	56,529	134,527
PCI	Per Capita Income	39,356	35,131	20,178	12,844	28,703	102,621
% of Residents							
PE	Preschool Enrollment	57.89	58.03	15.44	19.02	57.28	89.22
HSGR	High School Graduation Rate	85.12	84.04	9.54	56.33	85.64	97.95
CGR	College Graduation Rate	39.48	32.91	21.77	5.99	26.15	83.77
UER	Unemployment Rate	8.06	10.21	6.65	0.71	8.7	30.88
RB	Rent Burdened	45.97	47.67	9.32	24.84	49.43	64.78
SRB	Severely Rent Burdened	23.91	25.6	8.02	11.09	25.7	40.87
SPH	Single Parent Households	7.39	9.32	6.76	0.96	6.83	35.91
CH	Crowded Housing	3.62	3.63	2.21	0.23	3.03	10.41
VH	Vacant Housing	12.16	11.73	6.14	4.12	10.28	34.86
UIR	Uninsured Rate	9.65	8.99	4.25	2.25	8.12	22.52
PR	Poverty Rate	18.39	19.57	11.03	3.51	17.67	53.86
FS	Food Stamps (SNAP)	18.26	21.66	14.35	1.46	18.24	61.36
HPNRFS	Households in Poverty Not Receiving FS	48.05	53.51	18.59	21.65	52.72	95.34
PAI	Public Assistance Income (Cash Welfare)	3.15	3.61	1.95	0.59	3.68	9.05
AD	Ambulatory Difficulty	6.06	6.82	3.14	1.88	5.91	14.48
CD	Cognitive Difficulty	3.78	4.09	1.82	1.27	3.64	8.47
HD	Hearing Difficulty	2.15	2.41	0.85	1.04	2.42	4.91
ILD	Independent Living Difficulty	4.18	4.71	2.01	1.31	4.44	9.71
SCD	Self-Care Difficulty	2.33	2.62	1.29	0.5	2.43	5.93
VD	Vision Difficulty	2.39	2.55	1.23	0.48	2.32	6.25
FB	Foreign Born	20.64	18.54	13.51	0.8	17.42	49.31
LEP	Limited English Proficiency	7.64	7	7.34	0	3.99	31.33
Values Range from 0 to 0.83							
EDI	Economic Diversity Index	0.79	0.7	0.13	0.33	0.75	0.81
Composite Score Out of 100							
HI	Hardship Index	62.17	60.05	29.65	1.61	65.16	98.58

Sources: ACS; IDPH Death Certificates Data Files; HRSA Data Warehouse
Complete definitions for each variable can be found on https://chicagohealthatlas.org/indicators

The mean values of the predictor variables for the 75 CAs can be interpreted as "averages of averages" because they are based on averages for the households that completed the ACS for each unique CA. Consequently, averages for each CA are equally weighted even if an unequal number of households completed the ACS [6]. That is, descriptive statistics for the 75 CA are based on 75 data points.

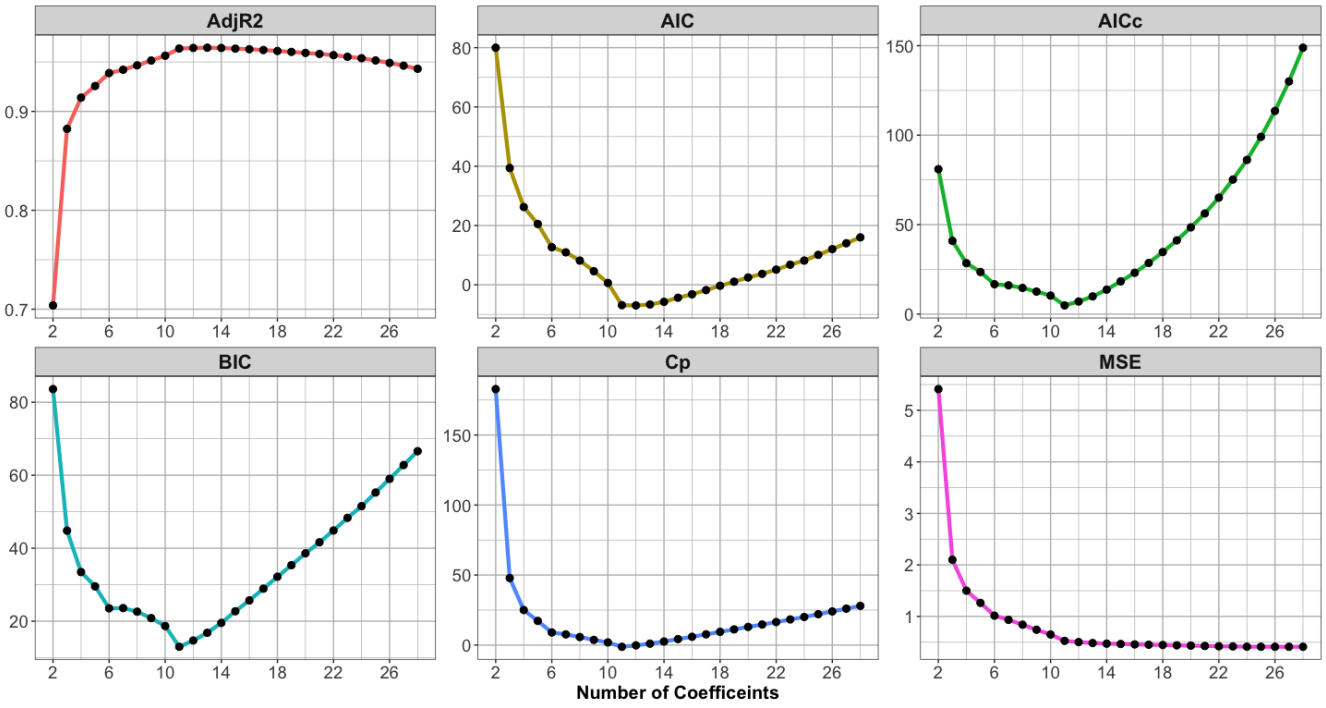
In comparison, the values of the predictor variables for Chicago, IL are averages of all households sampled

in the entire geographical area of City of Chicago [24]. The ACS was given to thousands of households in Chicago, IL between 2015-2019 [24]. Suppose N total households completed the ACS from 2015-2019. Then, the averages for Chicago, IL are based on N data points such that $N >> 75$.

The only exception is number of FQHC since it represents counts and the values are known rather than estimated. The the values for FQHC in Table 2 for the 75 CAs represents "averages of total counts". For Chicago, the value for FQHC represents the overall total or the sum of all FQHCs in all 77 CAs.

Note, median household income in Chicago, IL has remained roughly constant between 1969 to 2014 when adjusted for inflation [14]. In terms of inflation-adjusted 2021 dollars, the 2007-2011 median household income was approximately $57,000 USD; from 2011-2015, median household income dropped slightly to $55,000 USD; then, it increased for 2015-2019, reaching a median of nearly $63,000 as shown in Table 2

Figure 1: Information Criteria and MSE - Best Subset Models



Best Subset Models

Figure 1 shows how each model selection criteria and MSE change as the best subset model includes additional coefficients beginning with $b = 2$ (an intercept and one predictor variable). Recall, the best subset of each size b yielded the smallest MSE (and therefore RMSE).

R^2_{adj} was maximized by a model with 13 coefficients ($R^2_{adj} = 0.964$). However, the model with 12 coefficients is simpler and its R^2_{adj} is approximately equal ($R^2_{adj} = 0.964$). Therefore, the 12 coefficient model was selected as the best subset model for R^2_{adj}. Similarly, AIC was minimized by the same subset model with 12 coefficients (AIC $= -6.97$). Let *Model 2* (M_2) denote the subset model selected by R^2_{adj} and AIC. Table 4 provides M_1's regression summary.

AIC_c, BIC, and C_p were minimized by the subset model with 11 coefficients ($AIC_c = 4.85, \text{BIC} = 12.99, C_p = -1.24$). Let *Model 1* ($M_1$) denote the best subset model identically selected by AIC_c, BIC, and C_p. Table 5 provides M_2's regression summary.

It is hypothesized that models M_1 and M_2 overfit the data, especially since the *training set* only contains 45 rows of data. Two *Alternative Models*,

† *Full variable names in* Table 2

H_1 and H_2, are hypothesized to outperform M_1 and M_2 on the *test set*. H_1 was selected because it is the smallest model ($b = 7$) such that MSE < 1, and H_2 was selected to test if a simple two predictor model with an intercept ($b = 3$) will be sufficient; also, the increments in which MSE increases become relatively small for models larger than H_2 as shown in Figure 1. The regression summaries for H_1 and H_2 are in Table 6 and Table 7, respectively.

There were 13 variables not selected by any subset model. Let OV denote the model containing an intercept and these 13 other variables. Since OV's predictor were not selected in previous subset models, it is hypothesized that most predictors will not be statistically significant. In addition, it is hypothesized that OV will overfit the *training set* compared to the other subset models. This means OV may perform better or approximately the same as the other subset models on the *training set*, but OV's accuracy in the *test set*will be much worse [5]. The regression summary for OV is in Table 8.

The Full Model (FM) contains all 27 predictor variables and an intercept ($b = 28$). It is hypothesized that FM will drastically overfit the data. Specifically, FM will outperform all subset models on the *training set*, but its accuracy on the *test set*will be statistically worse than most if not all subset models. The FM's regression summary is in Table 3.

In summary,

1. Hypothesis 1: H_1 will be more accurate than M_1 and M_2 on the *test set*.
2. Hypothesis 2: H_2 will be least accurate on the *training set*, but on the *test set* its accuracy will be comparable to other models; it may even be more accurate than some models.
3. Hypothesis 3: Most predictors in the OV model will not yield statistically significant p-values.
4. Hypothesis 4: OV will perform the worse on the *test set* compared to all other subset models.
5. Hypothesis 5: FM will be most accurate model on the *training set* but its accuracy will sharply drop such that it will become the least accurate model on the *test set*.

Results

Model Coefficients and VIF

Coefficient tables display the estimated coefficients ($\hat{\boldsymbol{\beta}}$), their standard errors (Std. Error), and p-values (p-value) with significance codes (Signif.) *if* a predictor statistically significant at a significance level $alpha = 0.05$. Significance codes are defined below:

$$\text{Signif.} = \begin{cases} * & \text{, if p-value} < 0.05 \\ ** & \text{, if p-value} < 0.01 \\ *** & \text{, if p-value} < 0.001 \\ & \text{, if p.value} \geq 0.05 \text{ (Not Signif.)} \end{cases}$$

Additionally, the VIF values for each predictor in a model are plotted. Recall, the cutoff value for high VIF is 10 9. For each variable with VIF > 10, there is a likely at least one set of two or more predictors that demonstrate collinearity. This cutoff value is denoted by a red dashed line in each plot.

Note, all models happen to have the same estimated value for their intercept, $\hat{\beta}_0$.

The Full Model ($p = 7$)

The Full Model (FM) included all 27 predictors and an intercept. Only three predictors were statistically significant: (1) EDI = Economic Diversity Index; (2) PCI = Per Capita Income; & (3) SCD = Self-Care Difficulty. The lack of significant variables might be a consequence of collinearity between multiple sets of predictors.

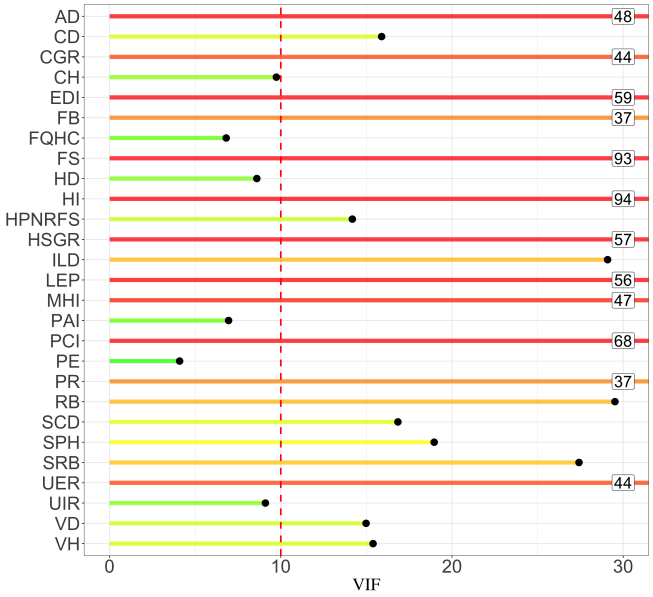
In Figure 2, only 6 of the 27 predictor variables had VIF below the cutoff value. Since many VIF have very large values, some values are labeled instead of expanding the axis limits. Despite being statistically significant, the predictors EDI, PCI, and SCD had VIFs exceeding the cutoff value.

† *Full variable names in* Table 2

Table 3: Full Model Coefficients

	$\hat{\beta}$	Std. Error	p.value	Signif.
(Intercept)	76.93	0.16	0.000	***
AD	0.81	1.09	0.469	
CD	-0.62	0.63	0.333	
CGR	2.05	1.04	0.066	
CH	-0.71	0.49	0.168	
EDI	-2.71	1.20	0.038	*
FB	1.01	0.96	0.307	
FQHC	-0.16	0.41	0.699	
FS	-1.34	1.51	0.387	
HD	0.21	0.46	0.650	
HI	0.70	1.52	0.650	
HPNRFS	-0.50	0.59	0.412	
HSGR	0.24	1.18	0.840	
ILD	1.47	0.85	0.101	
LEP	2.20	1.18	0.079	
MHI	0.62	1.08	0.571	
PAI	0.04	0.41	0.920	
PCI	-2.74	1.30	0.050	*
PE	-0.34	0.32	0.301	
PR	-1.80	0.95	0.076	
RB	0.87	0.85	0.323	
SCD	-1.71	0.65	0.017	*
SPH	0.49	0.68	0.483	
SRB	-1.01	0.82	0.238	
UER	-0.04	1.04	0.969	
UIR	0.06	0.47	0.908	
VD	-0.48	0.61	0.439	
VH	0.47	0.62	0.460	

Figure 2: VIF of Predictors in the Full Model



Model 1 $(p = 10)$

Model 1 (M_1) is based on the best subset model with 10 predictors selected by AIC_c, BIC, and C_p.

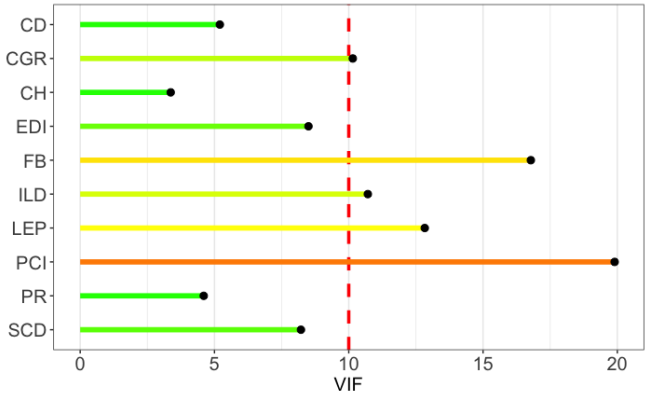
Table 4: Model 1 Coefficients

	$\hat{\beta}$	Std. Error	p.value	Signif.
(Intercept)	76.93	0.12	0.000	***
CD	-1.04	0.29	0.001	***
CGR	1.30	0.40	0.003	**
CH	-0.65	0.23	0.008	**
EDI	-2.26	0.37	0.000	***
FB	1.90	0.52	0.001	***
ILD	1.83	0.41	0.000	***
LEP	1.45	0.45	0.003	**
PCI	-1.96	0.56	0.001	**
PR	-1.66	0.27	0.000	***
SCD	-1.40	0.36	0.000	***

All predictors in Model 1 are statistically significant with p.value < 0.001 $(**)$.

The VIF for PCI (Per Capita Income) and FB (Foreign Born) are large and stand out compared to other predictors, especially PCI. FB and PCI likely introduce a set of collinear predictors to M_1. Consequently, issues caused by collinearity may arise in the *test set*.

Figure 3: VIF of Predictors in Model 1



† *Full variable names in* Table 2

Model 2 ($p = 11$)

Model 2 (M_2) is based on the best subset model selected by AIC and R^2_{adj}.

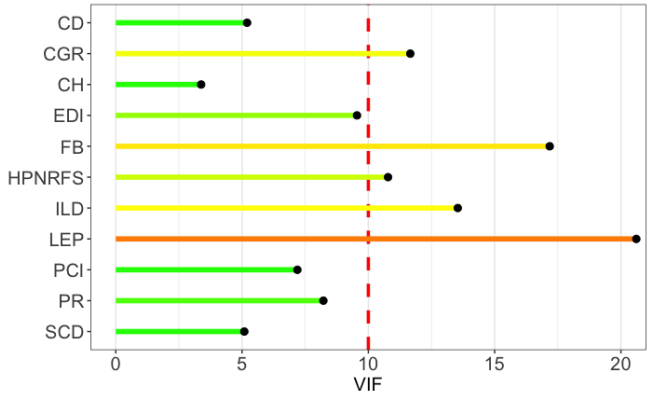
Table 5: M_2 Coefficients

	$\hat{\beta}$	Std. Error	p.value	Signif.
(Intercept)	76.93	0.12	0.000	***
CD	-1.05	0.28	0.001	***
CGR	1.50	0.43	0.001	**
CH	-0.63	0.23	0.009	**
EDI	-2.42	0.39	0.000	***
FB	1.80	0.52	0.001	**
ILD	1.87	0.41	0.000	***
LEP	1.58	0.46	0.002	**
PCI	-2.09	0.57	0.001	***
PR	-1.91	0.33	0.000	***
SCD	-1.40	0.36	0.000	***
HPNRFS	-0.35	0.28	0.220	

Note, all predictor variables in M_1 are also in M_2. The unique predictor added in M_2 was HPNRFS (Households in Poverty Not Receiving Food Stamps). HPNRFS was the only predictor not statistically significant.

Compared to M_1, the VIF for PCI in M_2 dropped below the cutoff line while the new predictor HPNRFS is slightly above the cutoff value. This suggests PCI and HPNRFS may be a set of collinear predictors.

Figure 4: VIF of Predictors in Model 2



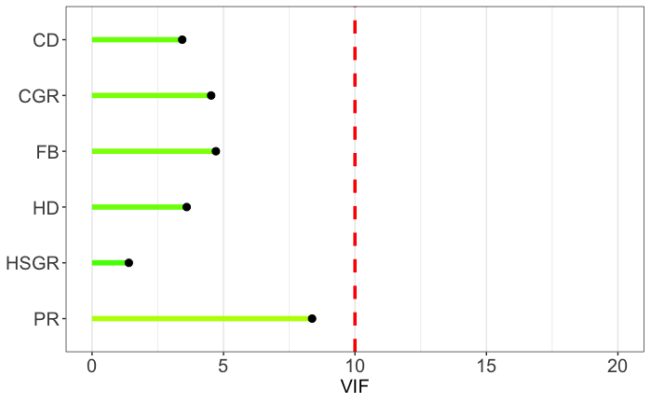
Model H_1 ($p = 6$)

H_1 is the smallest subset model such that MSE < 1. All predictors in the model are statistically significant and no VIF values exceed the cutoff line. Therefore, the H_1 does not raise concerns *yet*.

Table 6: H_1 Coefficients

	$\hat{\beta}$	Std. Error	p.value	Signif.
(Intercept)	76.93	0.16	0.000	***
CD	-1.56	0.29	0.000	***
CGR	2.40	0.34	0.000	***
FB	0.97	0.34	0.008	**
PR	-1.49	0.30	0.000	***
HD	0.80	0.19	0.000	***
HSGR	-1.82	0.46	0.000	***

Figure 5: VIF of Predictors in H_1



Model H_2 ($p = 2$)

H_2 is used test if a simple model is sufficient. The VIF values are not plotted because there are only two variables, so they have the same VIF. Specifically, VIF $= 1.2$. Therefore, H_2 does not raise concerns *yet* since both of the predictors are statistically significant with very small VIF values.

Table 7: H_2 Coefficients

	$\hat{\beta}$	Std. Error	p.value	Sig.
(Intercept)	76.93	0.22	0	***
FB	2.02	0.25	0	***
FS	-2.86	0.25	0	***

† *Full variable names in* Table 2

Model OV $(p = 13)$

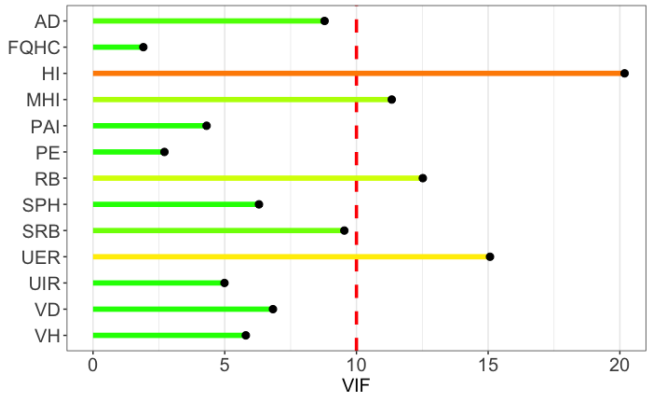
OV only consists of all *other variables* not selected in any other subset model (M_1, M_2, H_1, H_2).

Table 8: *OV* Model Coefficients

	$\hat{\beta}$	Std. Error	p.value	Signif.
Intercept	76.93	0.28	0.000	***
AD	-1.94	0.83	0.026	*
FQHC	-0.57	0.39	0.154	
HI	2.03	1.26	0.119	
MHI	-0.10	0.95	0.918	
PAI	-0.58	0.58	0.326	
PE	0.41	0.46	0.379	
RB	0.38	0.99	0.706	
SPH	-2.17	0.71	0.004	**
SRB	-1.25	0.87	0.160	
UER	-1.56	1.09	0.162	
UIR	-0.53	0.63	0.407	
VD	0.70	0.73	0.350	
VH	0.35	0.68	0.610	

Only 2 out of 13 predictors were statistically significant, reinforcing Hypothesis 3. However, only 4 of 13 predictors had VIF values above the cutoff line, which seems contradicting given the lack of significant predictors. M_2 has 11 predictors and M_1 has 10, yet both M_2 and M_1 had 5 predictors with VIF values above the cutoff line. This suggests that *OV*'s model accuracy could refute Hypothesis 4.

Figure 6: VIF of Other Predictors in Model *OV*



Model Diagnostics

For each model, four diagnostic plots based on the *training set* are provided in each of the following figures.

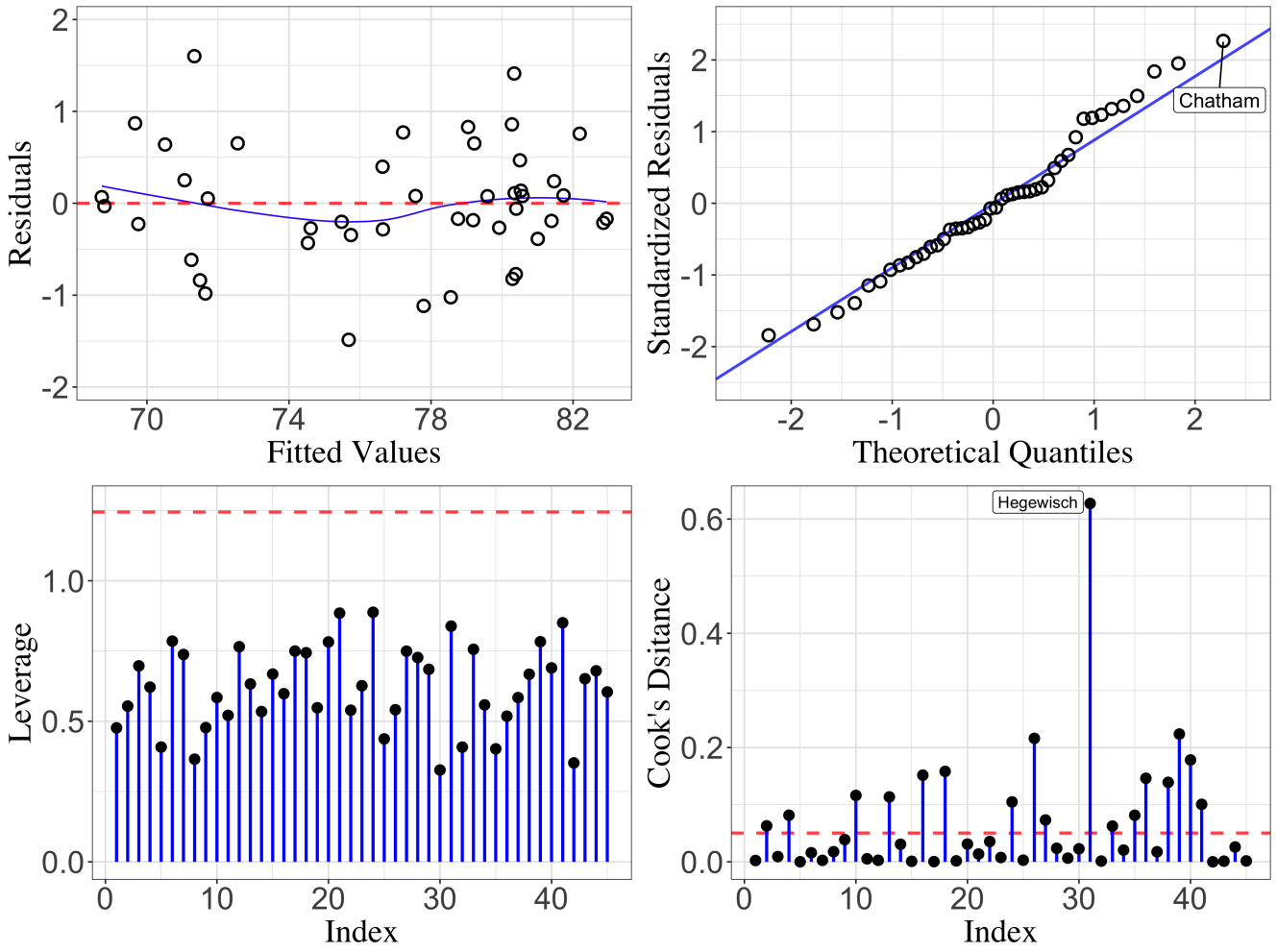
1. Fitted values versus residuals (top left)
2. Normal Q-Q plot (top right)
3. Point leverage for each CA (bottom left)
4. Cook's distance for each CA (bottom right)

The top two plots (1. and 2.) are used to assess regression assumptions about the residuals. Specifically, the residuals should be independently distributed with equal variance (6), and the standardized residuals should follow a standard normal distribution (11).

The bottom two plots (3. and 4.) are used to identify highly influential observations with cutoff values denoted by the red dashed line (12, 14).

† *Full variable names in* Table 2

Figure 7: *FM* Diagnostic Plots (Training Set)

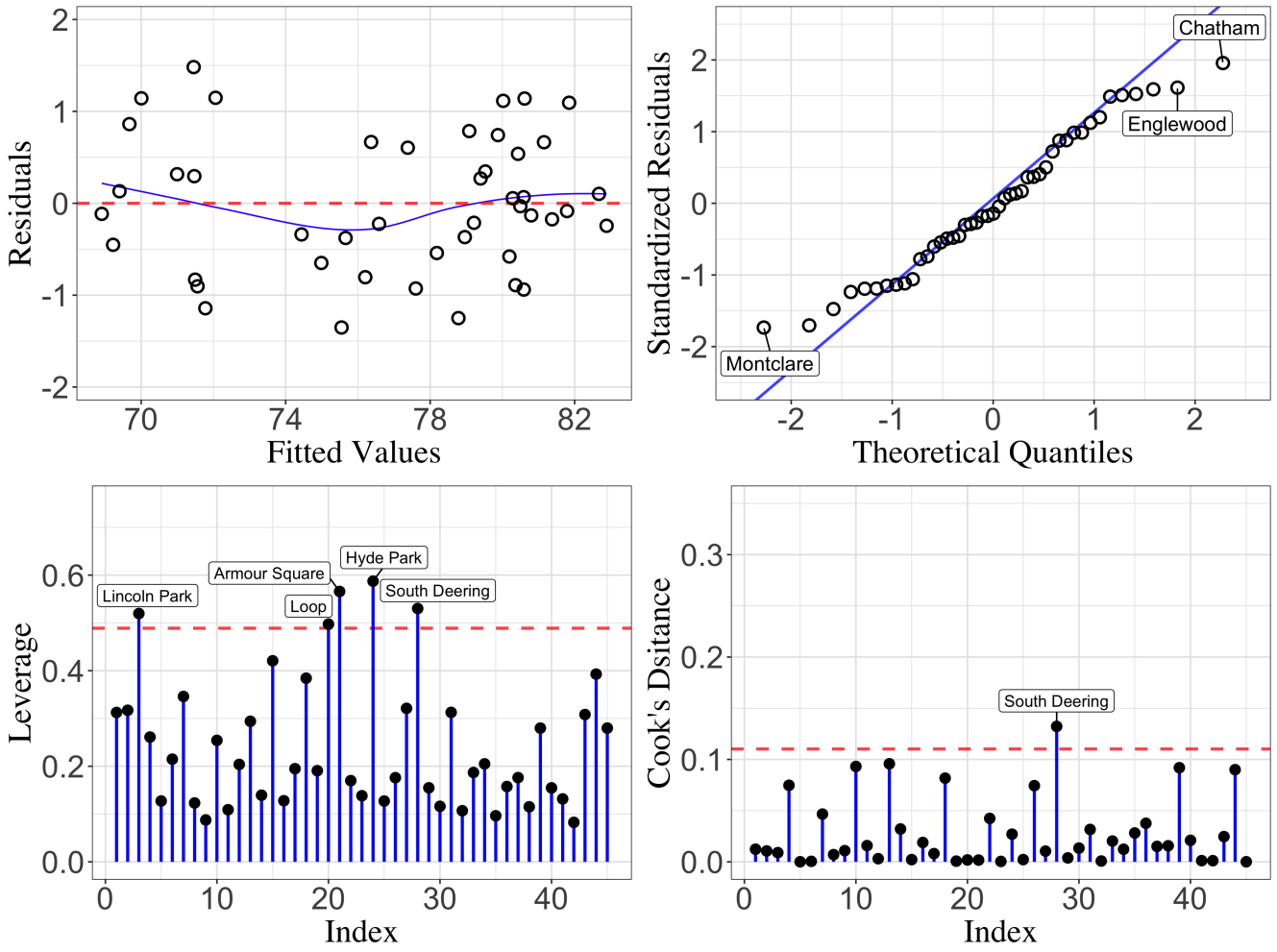


No assumptions about the residuals appear to be violated. There is one variable with standardized residuals slightly greater than 2, but this is expected based on the empirical rule regarding standard normal distribution (10).

Hegewisch appears to be highly influential point. Although no point leverage values exceeded the associated cutoff value, several points exceeding the cutoff value for Cook's distance. However, Hegewisch was the only CA labeled because it especially stands out.

† *Full variable names in* Table 2

Figure 8: M_1 Diagnostic Plots (Training Set)

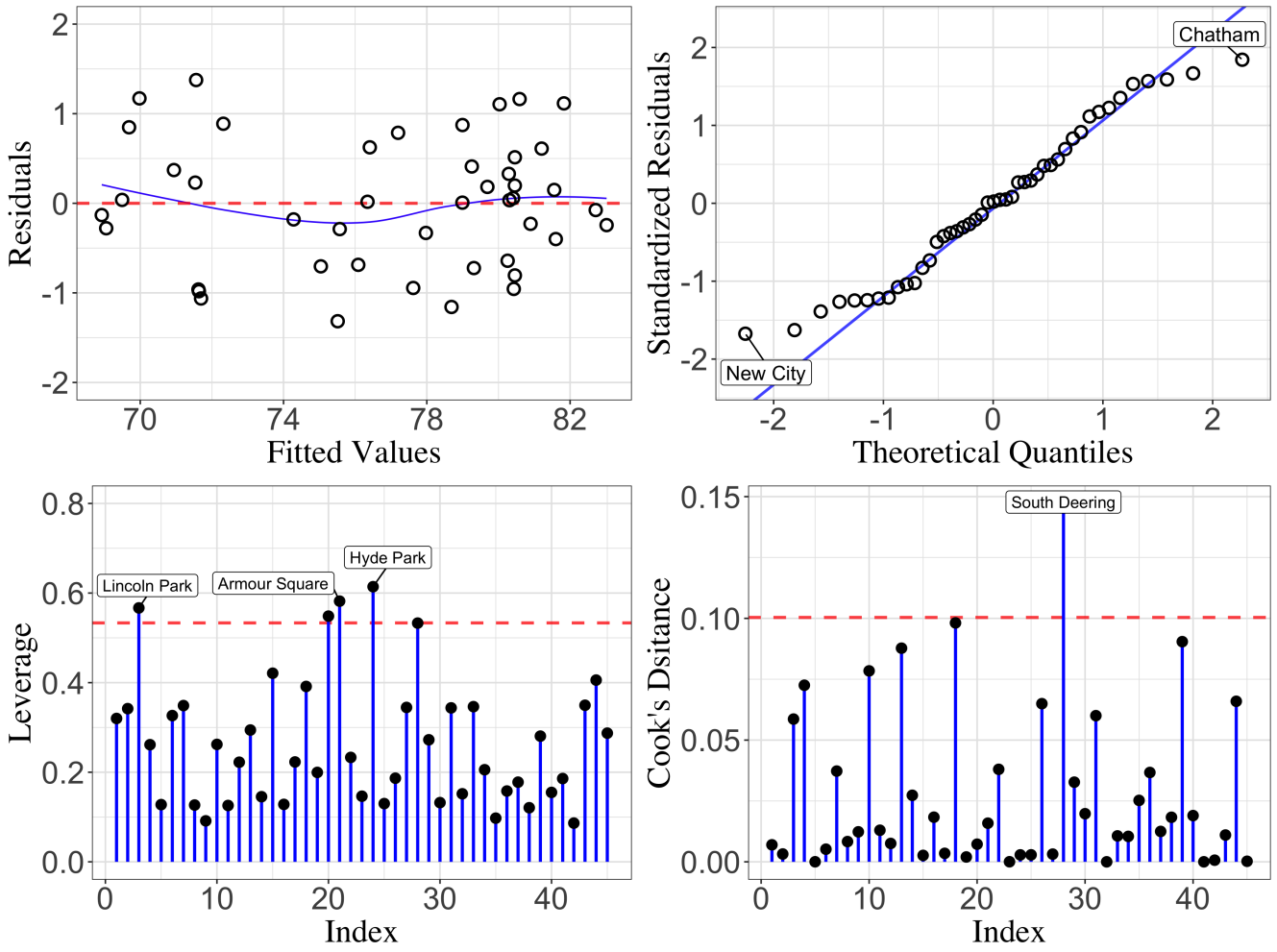


No assumptions about the residuals are clearly violated. The Normal Q-Q plot is showing signs of potentially tailing off based on the three CAs labeled. Nonetheless, this is still not cause for concern since 2-3 unusual points are expected based on the empirical rule.

Although several CAs have leverage above the cutoff value and South Deering has a Cook's distance above the cutoff, no point(s) stand out. For instance, the points above the cutoff line are justly slightly above, but there are several points just below the cutoff value (i.e., nearby values).

† *Full variable names in* Table 2

Figure 9: M_2 Diagnostic Plots (Training Set)

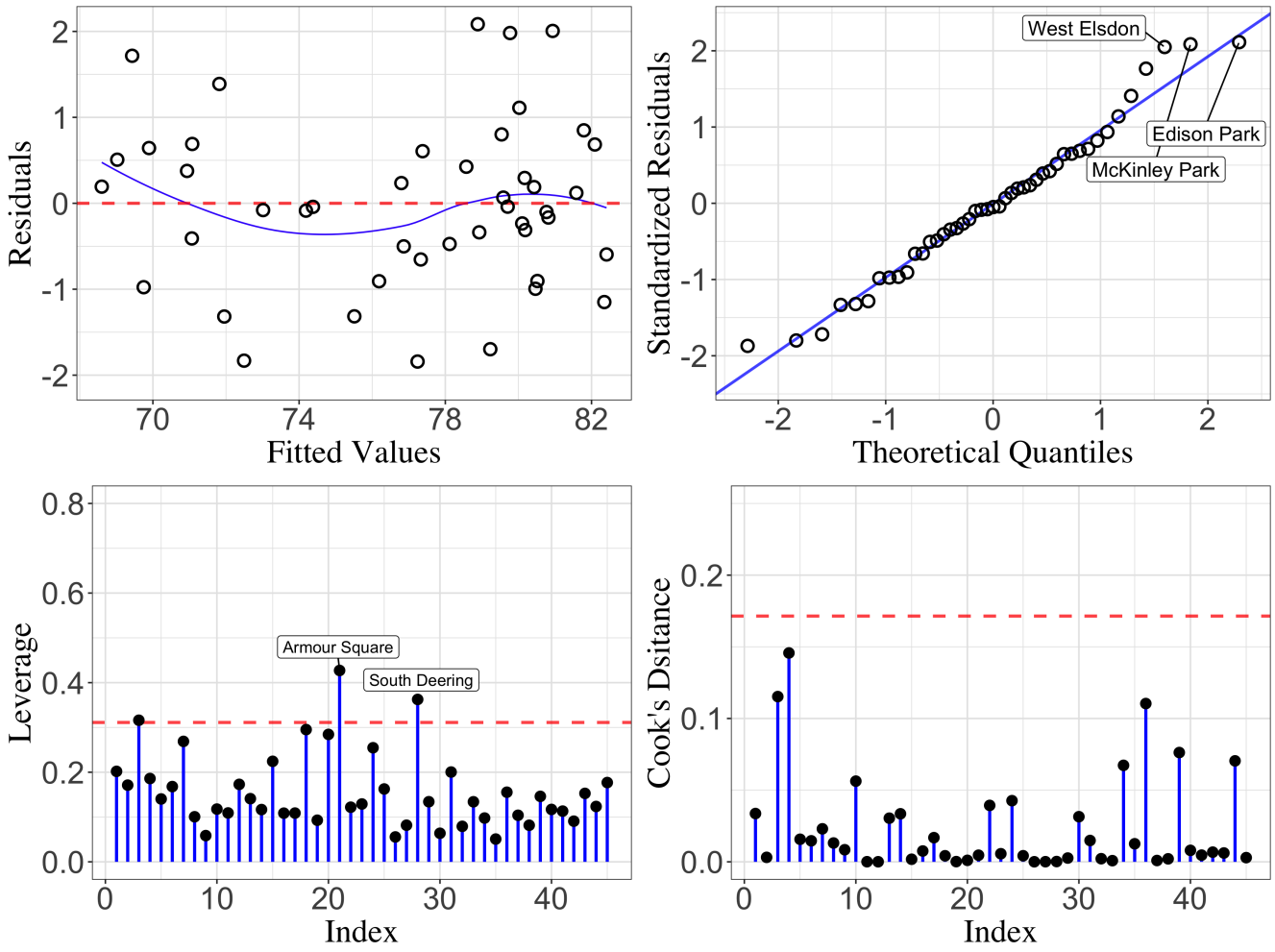


No residual assumptions are violated. The Normal Q-Q shows similar signa of tailing off like M_1, but this only starts among the two outer points.

South Deering appears to be the only CA that stands out based on its Cook's distance.

† *Full variable names in* Table 2

Figure 10: H_1 Diagnostic Plots (Training Set)

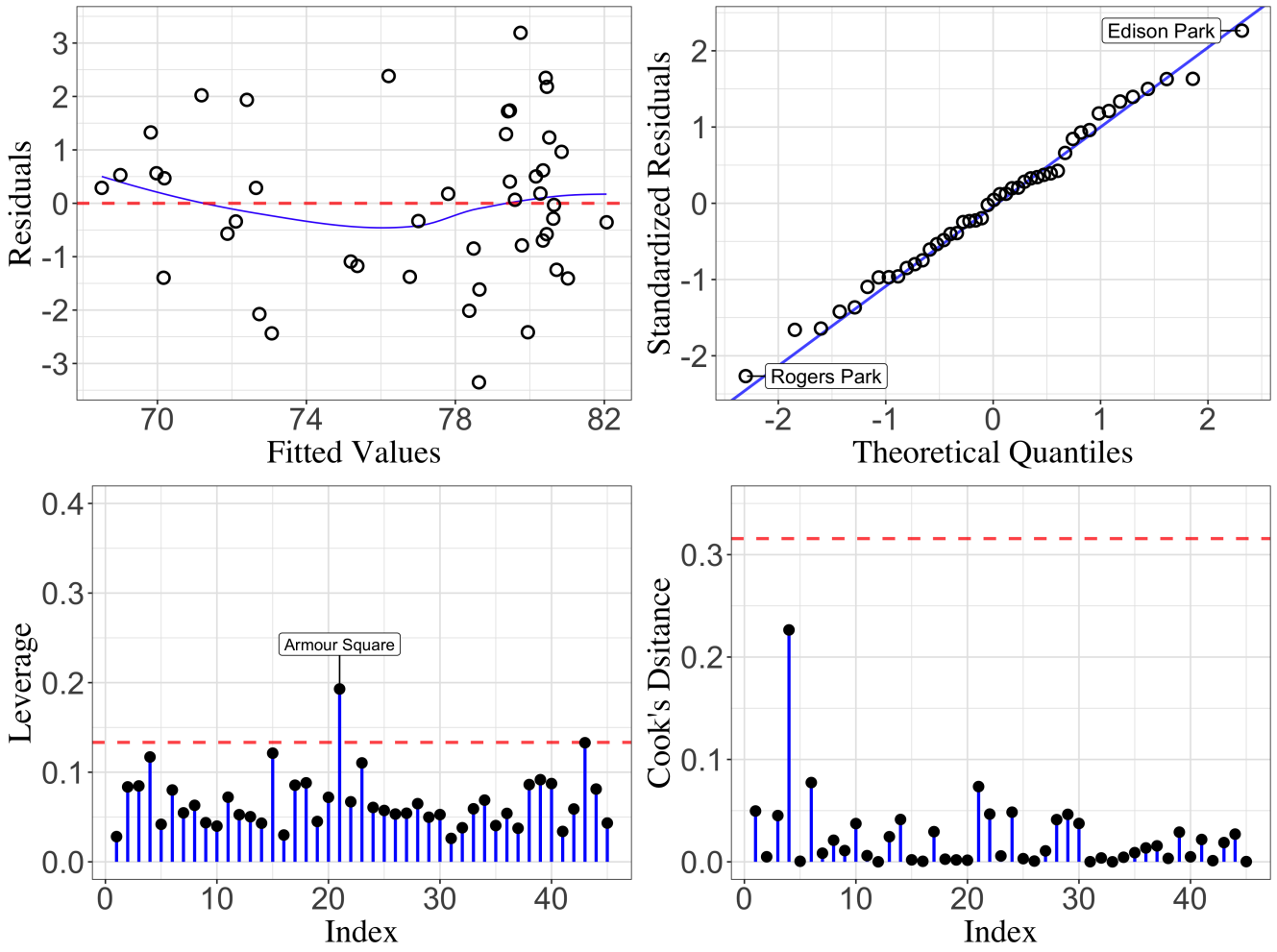


No assumptions about the residuals appear to be violated. No point stands out as being especially influential except for potentially Armour Square, but it does not particularly stand out despite exceeding the cutoff line.

Thus far, H_1 appears to be the strongest model.

† *Full variable names in* Table 2

Figure 11: H_2 Diagnostic Plots (Training Set)

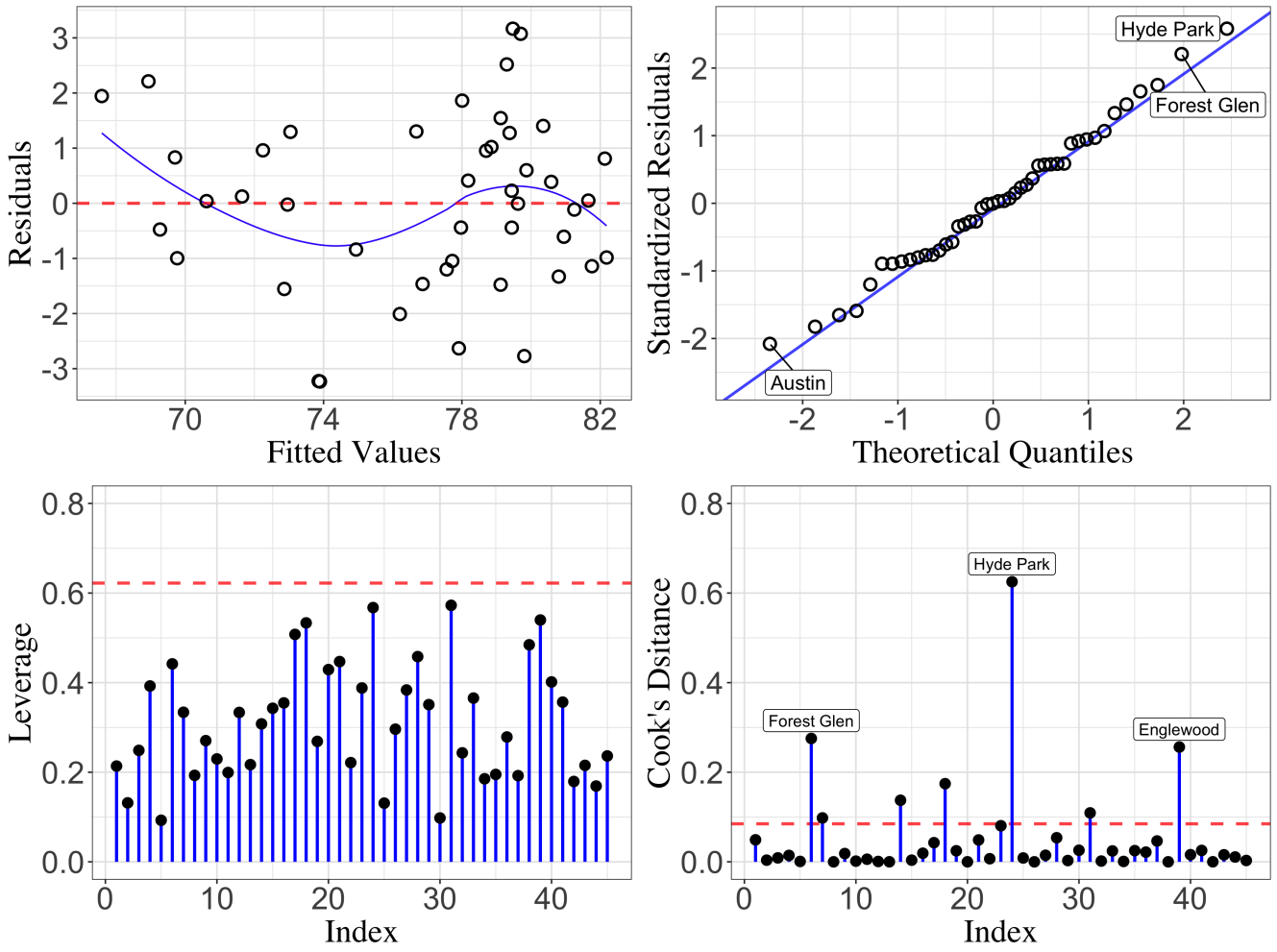


The fitted versus residual plot for H_2 has shape somewhat resembling a megaphone. In other words, residuals appear to be getting larger from left two right. For instance, $0 < |e_i| < 1$ for $\hat{y} < 70$; then, $0 < |e_i| < 2.5$ for $70 \leq \hat{y} \leq 78$; at the end, $0 < |e_i| < 3.5$ for $\hat{y} > 78$. The assumptions of independently distributed residuals is potentially violated in H_2. In contrast, the assumption that residuals are normally ditributed is strongly reinforced by the Normal Q-Q.

Armour Square's point leverage and Edison Park's Cook's distance stand out, making these CAs potential influential points. Note, Edison Park ($D_4 \approx 0.23$ was not labelled because it fell below the cutoff value, but its Cook's distance is more than 2 times greater the next closes value. Hence, it stands out more than Armour Square's leverage.

† *Full variable names in* Table 2

Figure 12: *OV* Diagnostic Plots (Training Set)



Note, the blue line is fitted using loess method smoothed to capture underlying trends. Accordingly, their appears to be a sinusoidal pattern in the residual vs fits plot. Also, residuals seem to have an increasing trend, although it is not super clear. The Normal Q-Q plot looks very consistent.

No points stand out in the leverage plot, but Hyde Park exceeds the cutoff line and very clearly stands out in the Cook's distance plot. Forest Glen and Englewood also appear to be influential, although their Cook's distance is less than half of Hyde Park's Cook's distance.

† *Full variable names in* Table 2

Model Performance Results

Figure 13: Model Coverage on the Test Set (95% Prediction Interval)

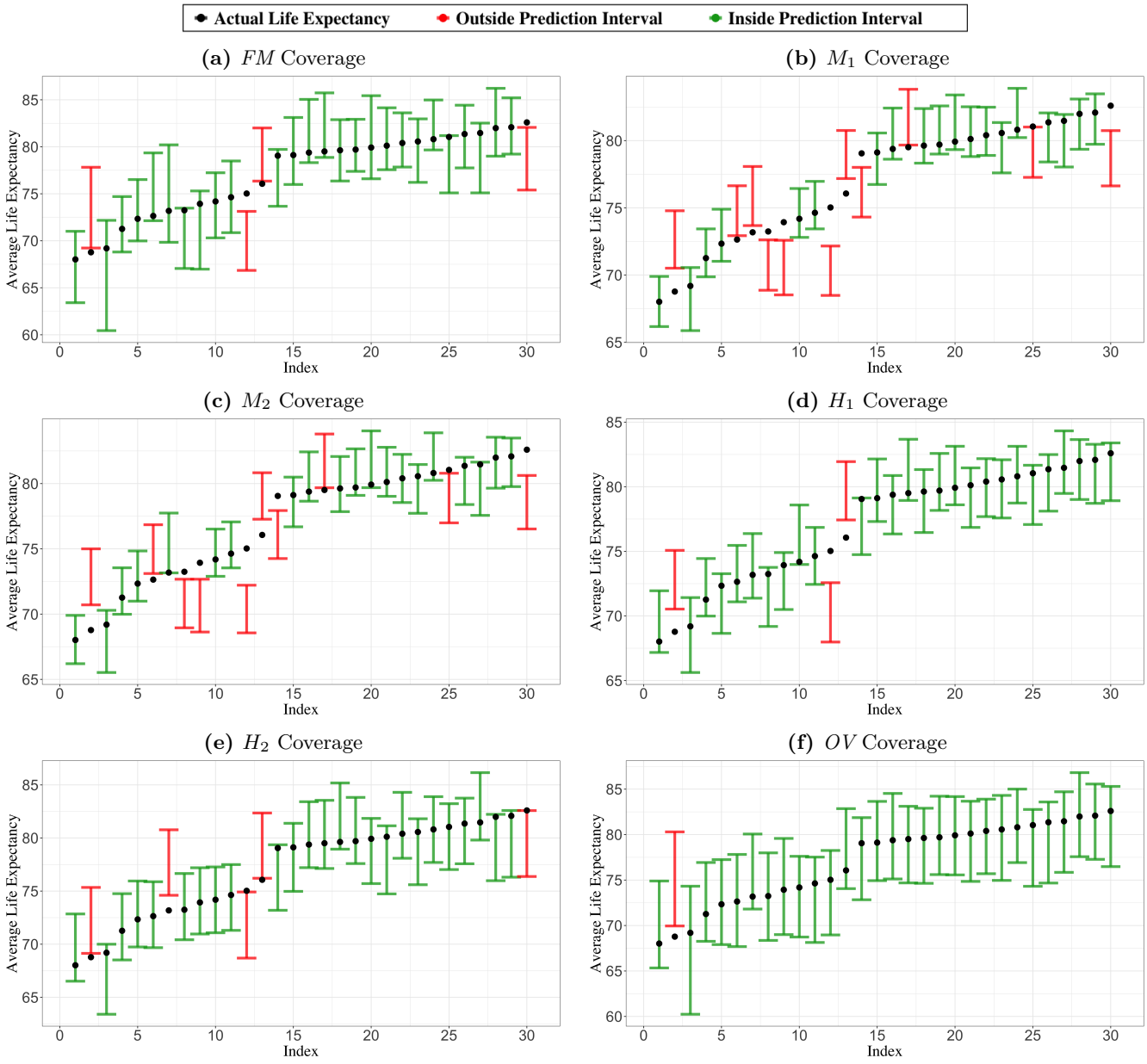


Table 9: Model Error Statistics (Training Set)

	p	MSE	RMSE	MAE
FM	27	0.411	0.641	0.493
M_1	10	0.526	0.725	0.603
M_2	11	0.502	0.709	0.582
H_1	6	0.934	0.967	0.754
H_2	2	2.100	1.449	1.174
OV	13	2.395	1.548	1.245

Table 10: Model Error Statistics (Test Set)

	p	MSE	RMSE	MAE
FM	27	5.085	2.255	1.753
M_1	10	4.066	2.016	1.601
M_2	11	4.235	2.058	1.633
H_1	6	2.918	1.708	1.322
H_2	2	3.948	1.987	1.565
OV	13	3.075	1.754	1.231

† *Full variable names in* Table 2

Discussion

In Table 9 and Table 10, key performance statistics compared between the *training set* and *test set*. MAE may be the most intuitive statistics for interpreting model performance since the magnitude of errors are not modified. Consider, the MAE for the Full Model (Table 9) was 0.493; in other words, the fitted values deviated from the true *average life expectancy* by about 0.493 years on average.

In Figure 13, the actual *average life expectancy* (black points) are compared to each model's prediction interval (error bars) in order to see if the interval covers the true outcome on the *test set*. If a model prediction interval covers the true value, the error bar is green; otherwise, the error bar is red. Given the error bars represent a 95% confidence band and the *test set* includes 30 CAs, the prediction interval should cover the actual value approximately 28.5 times.

H_1 slightly outperformed both M_1 and M_2 on the *test set* with H_1 yielding an MAE 0.279 and 0.311 smaller than M_1 and M_2, respectively. The prediction interval for H_1 included the actual *average life expectancy* 27 out of 30 times (Figure 13d), so H_1 had a coverage rate of of 90%. This rate is much better than the coverage rate of 63.3% yielded by both M_1 and M_2. These results support Hypothesis 1. In addition, H_1 yielded the lowest MSE and RMSE overall on the *test set*.

The *FM*'s error statistics were smallest in the *training set* but the largest in the *test set* in comparison to all other models. These results suggests *FM* overfit the *training set*. For instance, its MAE on the *test set* was more than 3 times greater its MAE on the *training set*. The error statistics for the *FM* reinforce Hypothesis 5. However, its coverage rate of 86.7% was still better than M_1 and M_2, contradicting Hypothesis 5. Although the *FM*'s performance was one of the poorest, it is unclear which model was the worst overall.

Out of all subsets, M_1 and M_2 yielded the largest error statistics and had the smallest coverage rate, especially M_2. While these means Hypothesis 2 is confirmed, it demonstrates alarming evidence against the statistical model selection criteria described in Model Selection.

Perhaps the most concerning evidence is revealed by the *OV* model's performance. Recall, *OV* consisted of the other variables not selected by any other subset (i.e., no statistical methods applied). Despite only have 2 statistically significant predictors out of 13 total, *OV* did not overfit the *training set*. On the contrary, the MAE for *OV* decreased on the *test set*, while its MSE and RMSE increased by the smallest proportion. Arguably, *OV* outperformed all other models since it yielded the lowest MAE highest coverage rate (96.7%) on the *test set* (H_1 is the other candidate for best model). Evidently, Hypothesis 4 was not supported.

Limitations

While there was no clear subset of predictors that performed significantly better than an alternative set of predictors in this study, this does not mean that there does not exist such a set.

There are other surveys used to collect community level health data that were not considered in this study. Initially, the goal of this study was to analyze two other different datasets, which are all found on the Chicago Health Atlas [6]. One of these dataset contained community-level data collected by the Healthy Chicago Community Survey (HCS) [6]. However, the dataset contained many missing values that limited any analysis. For instance, Armour Square, which is the only neighborhood with a predominantly Asian population, contained mostly missing values. This suggests that residents in this CA did not participate in the survey. Other CAs contained missing values as well. In addition, the most recent time period in which data was available was 2013 to 2017. Given the limitations of the HCS dataset, it was not analyzed.

Second, the Illinois Department of Public Health, Death Certificate Data Files were used to compile a dataset measuring mortality rates for the 77 CAs.

† *Full variable names in* Table 2

This data was very accurate with little to no margins of error because the cause of death is almost always known for each person. However, mortality rates were age-adjusted. In this study, the goal was to predict life expectancy, which is an age-related measure. Without a concrete understanding of the relationship between life expectancy and age-adjusted variables, generalization about the CAs in Chicago could be misleading.

ACS data split by gender, race/ethnicity, and/or age group for each of Chicago's 77 CAs was not analyzed. Various studies found evidence of gender and racial/ethnic health disparities among related indicators [1, 2, 11, 10, 12, 26, 3]. However, such data was not found in the Chicago Health Atlas. If such data is available, analysis of the ACS data by demographic splits (Table 1) could be insightful.

The data used in this study has its limitation because it represents sums of averages estimates. The ACS measures average estimates for each Census Tract over 5-year period [24]. Since Census Tracts are smaller geographical areas than CAs, the Chicago Health Atlas summed various averages to obtain averages for each CA.

By the central limit theorem [4], as sample sizes become increasingly larger, they become approximately normally distributed with sample mean, $\bar{X}$, equal to the population mean, μ. Consequently, the underlying variance of each CA is lost when taking the "averages of averages". Therefore, the standard deviation provided in Table 1 correspond to the variance of the averages rather than the variance of CAs. It would be erroneous to make generalizations about the City of Chicago using "averages of averages", especially when city-wide data is provided [21].

The ACS provides standard errors for their estimates at the Census tract level [24]. Standard error for each indicator and CA can be found on the Chicago Health Atlas [6]. If the sample size is small, then the level of uncertainty increases when estimating the true mean of all households in each CA [21]; this is referred to sampling error [24]. Hence, small differences are likely not statistically significant. Consequently, small differences between two values are likely insignificant [24].

The statistical methods used for model selection included Adjusted Multiple Correlation Coefficient (R^2_{adj}), Akaike Information Criterion (AIC), Bias Corrected AIC (AIC_c), Bayes Information Criterion (BIC), and Mallow's Statistic (C_p). There are so many more methods of selecting models and fitting effective regression models. For instance, Principle Component Analysis (PCA), Ridge Regression, and Weighted Least Squares are some methods used to account for collinearity and scale regression models based on the influence of each predictor. Such advanced methods may yield different results, especially considering the large variance in the response data collected by the ACS (Table 2).

By splitting the data into a *training set* and *test set* of 45 and 30 CAs, respectively, potential findings are limited and could be misleading. Ideally, the procedure outlined in this study should be repeated through bootstrap methods both in terms of the sample splits and their sizes. Since bootstrap methods were not used, findings can only be generalized to the sample considered. For instance, while the model selection criteria did not select the best subset model in this sample, it is unknown if these findings hold true when an alternative random sample is considered.

Conclusion

The goal of this project was to gain insight on health issues by analyzing community-level data. Although regression analysis was used to predict future outcomes, the primary purpose of applying regression analysis was to gain a deeper understanding about the relationship between *average life expectancy* and health-related indicators listed in Table 2.

Generally, proposed strategies for improving public health are partisan across the nation [13]. Studying data trends using statistical methods is a subjective approach of uncovering insight about public health. Accordingly, significant findings could provide a basis for a non-partisan solutions to improve public health, which is the primary objective of the World Health Organization [27].

† *Full variable names in* Table 2

However, based on the data used and statistical methods applied, there were mixed findings. Regression analysis did not uncover which set of indicators had significantly stronger relationship with *average life expectancy* compared to an alternative model. Every model was relatively successful at predicting *average life expectancy*. The worst model on the *test set* was the Full Model (*FM*), yet its MSE was 5.085, RMSE was 2.255, and its MAE was 1.753 (Table 10). Considering the actual values for *average life expectancy* had a mean of 76.98 and a standard deviation of 4.37 (Table 2), the *FM* was not a bad predictor. Because every model did relatively well, it was difficult to pinpoint which predictors were the best.

Overall, the models H_1 (Table 4) and *OV* (Table 8) were *slightly* better than all other models. Nonetheless, these models were not significantly better than other subset models, so it is unclear which if any indicators should be the focus of potential health policies seeking to increase *average life expectancy*.

† *Full variable names in* Table 2

References

[1] Benjamins, M. R., Silva, A., Saiyed, N. S., and De Maio, F. G. (2021). Comparison of all-cause mortality rates and inequities between Black and White populations across the 30 most populous US cities. *JAMA Network Open*, 4(1):1–14.

[2] Benjamins, M. R. and Whitman, S. (2014). Relationships between discrimination in health care and health care outcomes among four race/ethnic groups. *Journal of Behavioral Medicine*, 37:402–413.

[3] Buitrago, K. (2019). The gender disadvantage: Why inequity persists. Technical report, Heartland Alliance.

[4] Casella, G. and Berger, R. L. (2002). *Statistical learning*. Cengage: Belmont, CA, 2nd edition.

[5] Chatterjee, S. and Hadi, A. S. (2012). *Regression analysis by example*. John Wiley and Sons, Inc.: Hoboken, NJ, 5th edition.

[6] Chicago Health Atlas (2020). [Data Portal]. Data downloaded Mar. 2022, URL: https://chicagohealthatlas.org/.

[7] Data USA (2020). Idaho city, id. [Website]. Accessed May 2022, URL: https://datausa.io/profile/geo/idaho-city-id.

[8] Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The elements of statistical learning: Data mining, interence, and prediction*. Springer: New York, NY, 2nd edition.

[9] Heartland Alliance: Social IMPACT Research Center (2022). [Data Portal]. Data downloaded Feb. 2022, URL: https://www.heartlandalliance.org/heartland-alliance/research-and-policy/data-reports/chicago-data-dashboards/.

[10] Hunt, B. R., Tran, G., and Whitman, S. (2015). Life expectancy varies in local communities in Chicago: racial and spatial disparities and correlates. *Journal of Racial and Ethnic Health Disparities*, 2:425–433.

[11] Hunt, B. R. and Whitman, S. (2015). Black:White health disparities in the United States and Chicago: 1990–2010. *Journal of Racial and Ethnic Health Disparities*, 2:93–100.

[12] Lange-Maia, B. S., De Maio, F., Avery, E. F., Lynch, E. B., Laflamme, E. M., Ansell, D. A., and Shah, R. C. (2018). Association of community-level inequities and premature mortality: Chicago, 2011–2015. *Journal of Epidemiology and Community Health*, 72(12):1–5.

[13] Lantz, P. M. and Rosenbaum, S. E. (2020). The potential and realized impact of the Affordable Care Act on health equity. *Journal of Health Politics, Policy and Law*, 45(5):831–846.

[14] Lewis, J. and Paral, R. (2016). Two decades of household income trends in chicago community areas. Technical report, Rob Paral and Associates.

[15] Luya, M., Giulioa, P. D., Legoa, V. D., Lazarevič̆a, P., and Sauerberga, M. (2019). Life expectancy: Frequently used, but hardly understood. *Gerontology*, 66(1):95–103.

[16] Mayne, S. L., Pool, L. R., Grobman, W. A., and Kershaw, K. N. (2018). Associations of neighbourhood crime with adverse pregnancy outcomes among women in Chicago: Analysis of electronic health records from 2009 to 2013. *Journal of Epidemiology and Community Health*, 72(3):230–236.

[17] McCartney, G., Popham, F., McMaster, R., and Cumbers, A. (2019). Defining health and health inequalities. *Public Health*, 172:22–30.

[18] National Heart, Lung, and Blood Institute. (Mar. 2022). Blood Tests [Online]. Accessed Apr. 2022, https://www.nhlbi.nih.gov/health/blood-tests.

[19] Shah, A. M., Whitman, S., and Silva, A. (2006). Variations in the health conditions of 6 Chicago community areas: A case for local-level data. *American Journal of Public Health*, 96(8):1485–1491.

[20] Svalastog, A. L., Donev, D., Kristoffersen, N. J., and Gajović, S. (2017). Concepts and definitions of health and health-related values in the knowledge landscapes of the digital societya˙ *Croatian Medical Journal*, 58(6):431–435.

[21] Triola, M. F. (2012). *Elementary statistics*. Pearson: Boston, MA, 12th edition.

[22] United States Census Bureau (2020). Sample ACS & PRCS Forms and Instructions [Online]. Accessed Feb. 2022, URL: https://www.census.gov/programs-surveys/acs/about/forms-and-instructions.2019.html.

[23] United States Census Bureau (2022). Glossary. [Website]. Accessed Apr. 2022, URL: https://www.census.gov/programs-surveys/geography/about/glossary.html#.

[24] U.S. Census Bureau. (2020). *Understanding and using American Community Survey data: What all data users need to know*. U.S. Government Publishing Office, Washington D.C.

[25] Van Brunt, D. (2017). Community health records: Establishing a systematic approach to improving social and physical determinants of health. *American Journal of Public Health*, 107(3):407–412.

[26] Whitman, S., Silva, A., Shah, A. M., and Ansell, D. (2004). Diversity and disparity: GIS and small-area analysis in six Chicago neighborhoods. *Journal of Medical Systems*, 28(4):397–411.

[27] World Health Organization (2020). *Constitution of the World Health Organization*, 49th edition.