Access to Food in Chicago:
a Hierarchical Perspective

Daniel Berry

Abstract—Easy access to healthy, nutrient rich foods is a current public health crisis in the United States. Areas that have limited access to healthy foods and easy access to unhealthy foods are called "food deserts". In the United States today, X million people are living in a food desert. Access to food can be measured at the city-block level and can have substantial variation within a single neighborhood. However, much publicly available city wide health data is available at the neighborhood level. This gives the data a natural hierarchical structure as city blocks are nested within neighborhoods. In this paper we model presence/absence of deserts using a combination of block-level and neighborhood-level data from the city of Chicago. We apply construct a hierarchical logistic regression model and demonstrate its superiority.

I. INTRODUCTION

Access to readily available and inexpensive healthy food options is important for personal health. Unfortunately, for some in society today, access is not equitable. Some towns and neighborhoods have excellent healthy food options. Others do not. In some locations it is difficult to impossible for many residents to find healthy food options. In these areas, known as food deserts, there is readily available processed, calorie dense but nutrient poor foods.

A person's access to food is determined by the city block in which they live. Any increase in precision beyond the city block level is meaningless. However, rarely is data reported at the city block level. Sampling variability at that level would make drawing conclusions difficult. For that and other reasons, often city data is reported at the neighborhood, zip code, or telephone area code level. This gives the data a natural hierarchical structure where city blocks are nested within neighborhoods.

A. Food Deserts

There is no single agreed upon definition of a food desert. Some common metrics are based on distances to nearest supermarket with cutoffs at 1 mile for urban areas and 10 miles for rural areas. 1 mile may seem like a short distance to travel, but in urban areas with low car ownership rates where residents rely on public transportation, a 1 mile journey can take a substantial amount of time. For this work we used the definition from of defining a city block as being in a food desert if the city block is more than 1 mile from a supermarket. Supermarket in this context is a grocery store that is larger than 10000 square feet that is not primarily a liquor store. Distance is defined as the great circle distance from the center of mass of the city block to the center of mass of the grocery store.

II. METHODS

In the following section we describe the procedures used. We begin by describing the data used and including data sources. Then we move on to describing the types of models used.

A. Data Gathering and Manipulation

All data was sourced from the generous Open Data Portal operated by the city of Chicago. The portal can be found at data.cityofchicago.org. We utilized the following data files:

1) Block level data:

a) Crimes 2001 - present: This file contains a record for crimes in Chicago since 2001 with information about the type of crime as well as its location. The location is pseudo-anonymized to be random but within the same city block. For each city block we counted the total number of crimes committed within 1 mile in 2009. Our hypothesis *a priori* was that food deserts were often located in high-crime areas.

b) 311 Service Requests: Vacant Buildings: This file contains a record for every 311 service call about an abandoned/unoccupied/unlawfully occupied building. For each city block we counted the number of calls for vacant buildings where the building was located within 1 mile of the city block. Our hypothesis was that food deserts were located in areas with higher levels of vacant buildings.

c) CTA Ridership: Avg. Weekly Boardings during October 2010: This file contains average weekly boardings for the month of October 2010 for every bus stop in Chicago. We counted the total boardings for all stops within a mile of each city block. Our hypothesis was that residents of food deserts would tend to have a higher reliance on public transportation than residents in other areas.

d) Census Block Population : This file contains the population of each city block.

2) Neighborhood level data:

a) Public Health Statistics: selected public health indicators by Chicago community area: This file contains a record for every neighborhood in Chicago with selected public health information. Examples include teenage births per 100,000 residents and Gonorrhea prevelence (cases per 100,000). Our hypothesis was that food deserts were likely to be underserved in a more general public health sense than just lacking access to food.

b) Census Data: Selected socioeconomic indicators : This file contains a record for every neighborhood in Chicago with information about the socioeconomic status of that neighborhood. Examples include percent of residents below the povery level and percent without a high school diploma.

c) Race by Community Area : This file contains a record for every neighborhood in Chicago with the number of residents of each race who reside in that neighborhood.

B. Generalized Linear Models

In ordinary linear regression (or ordinary least squares, OLS) we assume that our observations y are some linear combination of the covariates, x_i's that we have plus random error. In matrix notation this can be expressed as:

$$y = X\beta + \epsilon$$

where y is the vector of responses, X is the (model) matrix of data, β is the vector of regression coefficients, and $\epsilon \sim N(0, \sigma^2 I)$ is the random error.

Generalized linear models extends this to non-normally distributed responses through the use of a link function g:

$$Y = g^{-1}(X\beta)$$

. For 0,1 or binomially distributed responses, there are a variety of link functions available. One of the most common is the logit link: $g(x) = \log(\frac{x}{1-x})$. For simplicity, we chose the logit link although there are certainly other options available (probit, t, etc.).

C. Hierarchical Models

Hierarchical models are another extension to the family of linear models which allow for

D. Models Fit

1) Complete Pooling: To begin we have the simplest model: ordinary regression using only the block-level variables. This model pools together every neighborhood as if the neighborhood distinctions don't matter.

$$y_{ij} = \text{logit}^{-1}\left(\alpha + X_B\beta_B + \epsilon_{ij}\right)$$

Where $\epsilon_{ij} \sim N(0, \sigma^2)$

2) No Pooling: The next model has a different but nonrandom intercept for each neighborhood, a fixed effect for that neighborhood. This would correspond to our belief that the neighborhoods are each different from the others.

$$y_i = \text{logit}^{-1}\left(\alpha + X_B\beta_B + \gamma_j + \epsilon_{ij}\right)$$

Where $\epsilon_i \sim N(0, \sigma^2)$

3) Partial pooling: The next model has a random intercept for each neighborhood which corresponds to partially pooling the data together. For every neighborhood we use some of the information in other neighborhoods to estimate its intercept. That is, the intercepts in the previous model are shrunk toward the common mean.

$$y_i = \text{logit}^{-1}\left(\alpha_{j[i]} + X_B\beta_B + \epsilon_i\right)$$

Where $\epsilon_i \sim N(0, \sigma^2)$ and $\alpha_j \sim N(0, \sigma_\alpha^2)$

4) Hierarchical: The final and most complicated model that was fit was a hierarchical model including the neighborhood level predictors in estimating the random intercept for each neighborhood.

$$y_i = \text{logit}^{-1}\left(\alpha_{j[i]} + X_B\beta_B + \epsilon_i\right)$$

Where $\epsilon_i \sim N(0, \sigma^2)$ and $\alpha_j \sim N(X_N\beta_N, \sigma_\alpha^2)$

E. Model Comparison

Models were compared using AIC and cross validated Breir Score (Mean Square Error in the case of 2-class logistic regression)

III. RESULTS

A. Model Parameters

1) No Pooling:
2) Complete Pooling:
3) Partial Pooling:
4) Hierarchical:

IV. CONCLUSIONS