LATE DIAGNOSIS OF HIV/AIDS BY PLACE OF RESIDENCE: A Geographic Analysis to Identify Potential Risk Factors for Late Diagnosis in the United States

by

MADELINE CANNON

A THESIS

Presented to the Department of Mathematics and Computer Science
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

Spring 2018

An Abstract of the Thesis of

Madeline Cannon for the degree of Bachelor of Science
in the Department of Mathematics and Computer Science to be taken Spring 2018

Title: Late Diagnosis of HIV/AIDS by Place of Residence

Approved: ____

Professor David Levin

Because of the long clinical latency stage of HIV/AIDS, many people with HIV are not diagnosed until they have already been living with the disease for several years. It is important to diagnose people with HIV as early as possible to improve their life expectancy and reduce their risk of transmitting the infection to others.

The purpose of my analysis was to determine the correlation between characteristics of an HIV-infected person's place of residence and the probability that they would not be diagnosed until they were within twelve months of developing Stage 3 AIDS. The two characteristics I studied were population density and prevalence of HIV/AIDS, and I controlled for poverty and race/ethnicity. My data represented 48,302 HIV/AIDS cases between 2010 and 2015 from four hundred counties/regions in thirteen states. The states and time period were selected based on data availability.

I found that population density and prevalence were both negatively correlated with late diagnosis, even when poverty and race/ethnicity were controlled. People living in rural areas with a lower prevalence of HIV/AIDS were more likely to be diagnosed late. This demonstrates the need for more research on rural, low-prevalence areas, which are the least studied parts of the United States in terms of HIV/AIDS research.

Acknowledgements

I would like to thank my Primary Advisor, Professor David Levin, for guiding me throughout this process. His willingness to meet with me so frequently and his attention to detail greatly improved the quality of my analysis.

I would like to thank the other members of my Thesis Committee, Professor Elaine Replogle and Professor Daniel Rosenberg, for agreeing to be on my committee and for their feedback. I particularly appreciate Professor Replogle's willingness to be my Second Reader on short notice.

I would like to thank Dr. Clare Evans for her guidance and feedback, even after she could no longer be my Second Reader. Her encouragement and background in epidemiology were immensely helpful throughout the thesis process.

I would like to thank the Departments of Health and Human Services in Minnesota, Nebraska, New Hampshire, Tennessee, and Texas for responding to my data request and taking the time to put together the data for my analysis.

Lastly, I would like thank my parents, my sisters, and my uncle for all their love and encouragement, and for travelling all this way to see my defense. Without their support throughout high school and college, I probably would not have majored in Mathematics and Computer Science or undertaken such a long and difficult thesis project.

This project has played an important role in shaping my interests and goals and I am very grateful to everyone who has helped me with it.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

A major obstacle in treating and preventing HIV/AIDS is its high rate of late diagnosis. Because of its long clinical latency stage, it can take over a decade after infection for a person with HIV to begin experiencing noticeable symptoms (CDC.gov). As a result, many people are not diagnosed until the disease has already progressed to AIDS. This is a serious issue for two reasons. One, studies have shown that the earlier a person with HIV begins treatment, the longer their life expectancy and the better their quality of life (Farnham et al., 2013). The other is that people who are aware of their serostatus are significantly less likely to transmit the disease to others (Farnham et al., 2013). This is likely because (1) people who are aware that they are HIV positive are substantially less likely to engage in high-risk behaviors such as unprotected sexual intercourse than those who are not aware of their status (Marks et al., 2005), and (2) treatment reduces an individual's likelihood of infecting others even if they do engage in high-risk behaviors (Giardi et al., 2007). There are still many factors that can prevent a person who has been diagnosed with HIV/AIDS from being treated, but diagnosis is a necessary prerequisite for treatment. Therefore, it is of critical importance that people with HIV are diagnosed as early as possible.

The purpose of this analysis was to study how a person's county of residence may affect their likelihood of being diagnosed late in the United States. There have been many studies of how poverty, race, sexuality, injection drug use, age, and other factors influence the probability of late diagnosis. However, there are very few studies concerning place of residence, and those that exist are limited to one state or region.

When looking at the places in the U.S. with the highest and lowest rates of late diagnosis, it seems that those with the highest rates tend to be rural areas with low prevalence of HIV/AIDS. In other words, a person with HIV living in a rural area in which HIV/AIDS is not very common is less likely to be diagnosed before the disease has progressed to Stage 3 than a person living in an urban area where HIV/AIDS is more common. Because prevalence of HIV/AIDS and population density are highly correlated (with most cases occurring in cities), it is difficult to tell if the highest rates of late diagnosis are occurring in rural areas or the areas with the lowest prevalence.

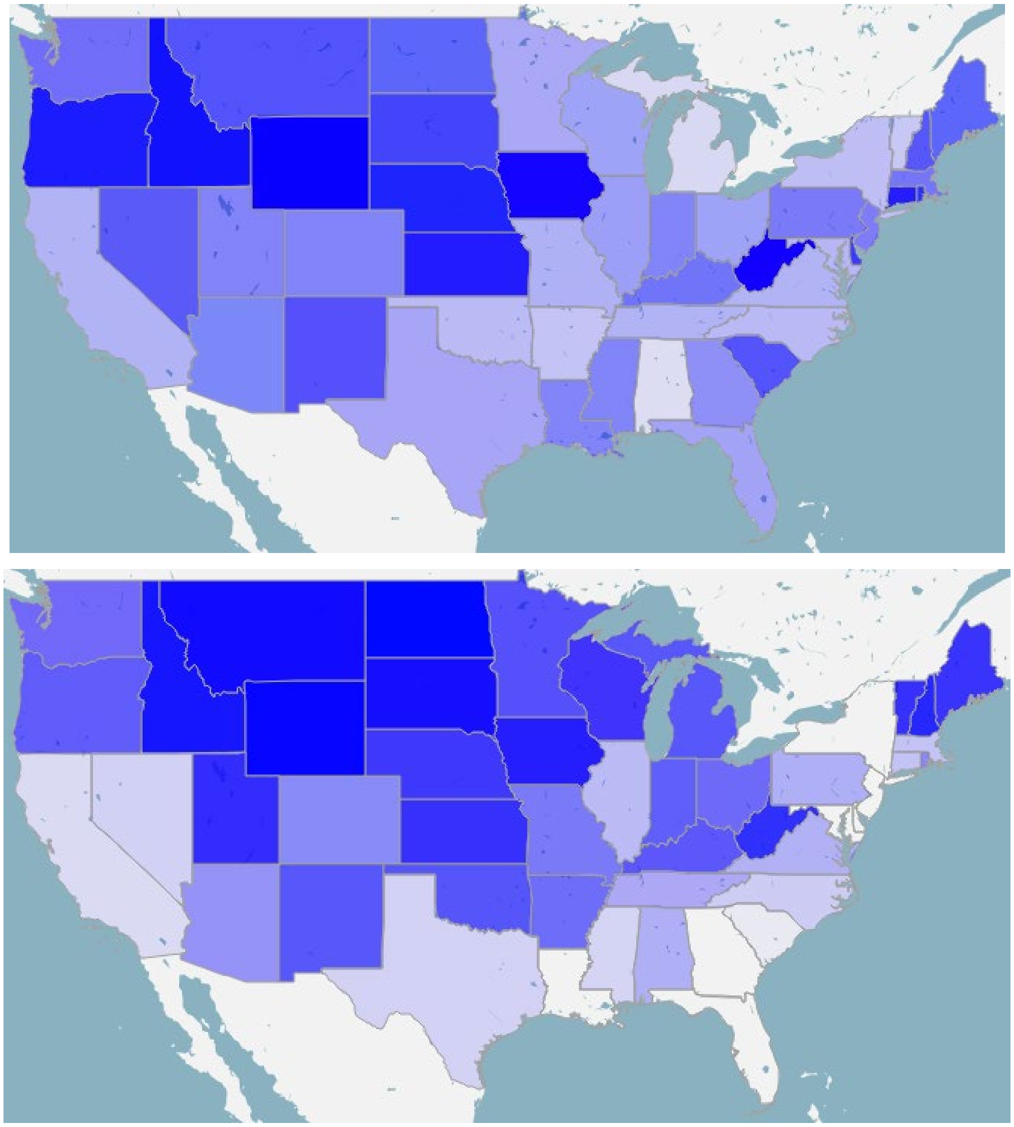


Figure 1: Late diagnosis and prevalence of HIV/AIDS in the contiguous United States

The first map shows late diagnosis in the contiguous United States between 2008 and 2014. In darker blue states, a higher percentage of HIV cases were not diagnosed until Stage 3. The second map shows low prevalence of HIV/AIDS in 2014. Darker blue states have lower prevalence and lighter blue states have higher prevalence. These maps were created using data from the CDC's 2011 and 2015 National HIV Surveillance System reports.

The goal of this analysis was to determine if there was a correlation between prevalence and late diagnosis or between population density and late diagnosis. My hypothesis was that there would be a strong correlation between prevalence and late diagnosis, but a weak correlation between population density and late diagnosis when prevalence was controlled. This is because relatively rural states with high prevalence (such as Alabama, Mississippi, and North Carolina) usually had low rates of late diagnosis. A possible explanation would be that people living in areas with low prevalence of HIV/AIDS have lower perceived risk—they do not consider themselves to be at risk of contracting HIV, so they are less likely to get tested for it. Another reason could be that low-prevalence states devote less funding to HIV/AIDS prevention and treatment programs. However, it could be that the urban/rural difference is the main predictor. People living in rural areas may have less access to HIV testing centers, resulting in a higher rate of late diagnosis.

To test these hypotheses, I performed a logistic regression analysis using data from four hundred counties and regions in thirteen states. The dependent variable was a binary outcome variable—one if the disease was already at Stage 3 at the time of diagnosis or if it progressed to Stage 3 within twelve months, and zero otherwise. The independent variables were population density of the person's county or region of residence, the prevalence of HIV in that county or region, the percentage of the population below the federal poverty line, the percentage of the population that was Black/African American, the percentage that was Hispanic/Latino, and the percentage that was another non-White race or ethnicity.

Glossary

Late diagnosis: An HIV case that has already progressed to AIDS at the time of diagnosis or progresses to AIDS within 12 months of diagnosis.

NHSS: National HIV Surveillance System. This is a CDC-funded system that collects data concerning HIV/AIDS from state and local health departments across the United States and publishes an annual report on the data, along with several intermittent supplemental reports.

PLWH: People Living With HIV

Stage 3 AIDS: The final stage of HIV, characterized by a CD4 count of less than two hundred cells per cubic millimeter of blood. This term is used interchangeably with "Stage 3" and "AIDS."

Chapter 2: Background

HIV/AIDS Basics

HIV (human immunodeficiency virus) is a virus that attacks the immune system, specifically the CD4 cells, which help the body fight off infections. It is transmitted through some bodily fluids, including blood, semen, vaginal fluid, and breast milk, and is most commonly spread by unprotected sexual contact and needle-sharing related to injection drug use. If left untreated, HIV will progress to AIDS (acquired immune deficiency syndrome), which is characterized by a CD4 count of less than two hundred cells per cubic millimeter of blood. People with AIDS have badly damaged immune systems and are vulnerable to opportunistic infections and infection-related cancers. It is the opportunistic illness and not the virus itself that causes death (AIDS.gov).

There is no cure for HIV, but HAART (highly active antiretroviral therapy) can greatly prolong the life of a person with HIV, prevent opportunistic illnesses, and reduce their risk of transmitting the disease to others if it is taken correctly and continuously. If the patient begins treatment before the disease is advanced, their life expectancy is close to that of someone without HIV. However, the more advanced the disease is as the beginning of treatment, the shorter their life expectancy (AIDS.gov).

Stages of HIV/AIDS

There are three stages of HIV infection: acute infection, clinical latency, and AIDS. The acute infection stage lasts two to four weeks, and is often characterized by severe flu-like symptoms. During this stage, the virus is rapidly reproducing in the body. The disease then progresses to the clinical latency stage. During this stage, the virus reproduces very slowly and the individual experiences few or no symptoms, though they can still transmit HIV to others. This stage lasts about ten years without treatment in the United States, though the length can vary widely depending on age, nutrition, stress, genetics, and other factors. The final stage is AIDS. At this stage, the immune system is badly damaged, leaving the person vulnerable to opportunistic illnesses. Without treatment, life expectancy for a person with AIDS is about three years after the latency period (AIDS.gov).

Brief History of HIV/AIDS in the United States

The first documented cases of HIV/AIDS in the United States were in 1981, when five young and previously healthy gay men in Los Angeles were diagnosed with a rare lung infection called *Pneumocystis carinii pneumonia (PCP)*. All five had other infections as well that were not usually seen in young and healthy individuals. At this time, the existence of AIDS was not yet known and medical professionals were unable to explain these infections. In the same year there were reports of a rare cancer called *Kaposi's sarcoma* in gay men in California and New York, and by the end of the year there had been 270 reported cases of severe immune deficiency among gay men and 121 deaths. This quickly turned into an epidemic, and by 1994, there had been 270,870 deaths in the United States, making it the leading cause of death for all Americans between the ages of 25 and 44 (AIDS.gov).

The epidemic finally began to slow in the mid-1990s. Concerted efforts on the part of activists, private organizations, and the U.S. government significantly reduced the number of new cases of HIV infection. The development of HAART (highly active antiretroviral therapy) in 1995 allowed people with HIV/AIDS to live almost as long as those without it (AIDS.gov).

However, in some subpopulations in the United States, HIV/AIDS continues to be a serious problem, and prevalence has actually increased in some groups (Castel et al, 2015). According to UNAIDS, HIV/AIDS has become a concentrated epidemic in the United States, meaning that less than 1% of the general population is infected and the majority of infections occur in high-risk subpopulations (Castel et al, 2015). These subpopulations include MSM (men who have sex with men), Blacks/African Americans, Hispanics/Latinos, and injection drug users. Prevalence is especially high among black MSM (or BMSM). Young BMSM alone (between the ages of 13 and 24) accounted for 58% of new infections among all MSM in 2010 (Castel et al, 2015).

Literature

Urban/Rural Differences

There have been many studies concerning late diagnosis of HIV/AIDS in the United States, but most of them focus on cities on the East and West coasts (Krawczyk et al., 2006). I have only been able to find two studies of late diagnosis in rural areas, and they are both limited to the South, where HIV/AIDS is more prevalent. This is an important gap in the literature, since many of the states with high rates of late diagnosis are relatively rural states outside of the South. In fact, the four states with the highest rates of late diagnosis between 2008 and 2014 were Wyoming, Iowa, West Virginia, and Idaho (CDC, 2011; CDC, 2015), which all fit this description.

One of the studies of late diagnosis in rural areas is Krawczyk et al. (2007), which provides a useful review of the epidemiological features of late diagnosis of HIV/AIDS in the South, including the effect of rural residence. They note that in 2001, the South accounted for two-thirds of rural AIDS cases throughout the United States, and 21% of AIDS cases within the South occurred in rural areas. They also note that in 2002, rural Southerners were significantly more likely to be uninsured than urban Southerners. However, this information is rather dated, as the epidemiology of HIV/AIDS in the United States has changed considerably since 2002.

The only study I have found that focuses on urban/rural differences in late diagnosis is by Trepka et al. (2014), who studied these differences in Florida. They found that late diagnosis was more common in rural areas than urban areas in Florida, with 35.8% of cases in rural areas being diagnosed late as opposed to 27.4% in urban areas. This difference persisted after controlling for age, sex, race/ethnicity, HIV transmission mode, country of birth, and diagnosis year. Interestingly, they found that predictors of late diagnosis were very different in rural and urban areas. In rural areas, men and people of older age were more likely to receive late diagnoses, while socioeconomic status and availability of healthcare resources were, surprisingly, not significant predictors. In urban areas, Hispanics/Latinos, Blacks/African Americans, people born outside the U.S., and people who were infected via heterosexual contact were at increased risk of late diagnosis, but these same patterns were not seen in the rural areas. This study highlights the importance of studying rural areas, as the predictors of late diagnosis are very different from the usual predictors found in the more-frequently studied urban areas. Therefore, intervention programs that are designed using research conducted in urban areas may not be very effective in rural areas.

Prevalence of HIV/AIDS

I have not been able to find any studies of the association between prevalence of HIV/AIDS in a person's place of residence and late diagnosis. This also seems to be an important gap in the literature, as there is a fairly clear inverse relationship between prevalence and late diagnosis. According to data from the CDC's 2011 and 2015 NHSS reports, eight of the ten states with the highest rates of late diagnosis between 2008 and 2014 had below-average prevalence (CDC, 2011; CDC, 2015).

Though there is a lack of studies of prevalence and late diagnosis, the association between perceived risk and late diagnosis is well documented (Girardi et al., 2007; Hall et al., 2013; Krawczyk et al., 2006; Mugavero et al., 2007; Mukolo et al., 2013; Trepka et al., 2014). People who do not believe they are at risk of contracting HIV are less likely to get tested before they begin experiencing serious symptoms, resulting in a higher rate of late diagnosis. It is possible that people who live in areas with low prevalence of HIV/AIDS would have lower perceived risk, resulting in a higher rate of late diagnosis.

Poverty

There is a well-documented association between poverty and late diagnosis of HIV/AIDS, an association that is found in many high-income and low-income countries around the world (Mukolo et al., 2012). Studies have found that people with low income are more likely to have been tested for HIV but less likely to receive an early diagnosis, reflecting the fact that uninsured and low-income people are more likely to experience suboptimal healthcare (Krawczyk et al., 2006).

Race/Ethnicity

Significant racial/ethnic differences have been found in studies of late diagnosis. CDC data from 1996 to 2001 showed that Hispanics/Latinos had the highest rates of late diagnosis (46.7%), Whites had the second highest (40.1%), and Blacks/African Americans had the lowest (39.4%). The same general pattern was found from 2001 to 2005, with Hispanics/Latinos at 57.7%, Whites at 54.1%, and Blacks/African Americans at 53.1% (Chen et al., 2011). (An interesting aspect of this data is that it shows that overall rates of late diagnosis increased over time, possibly due to lower perceived risk as prevalence of HIV/AIDS dropped.)

Numerous studies have found that Hispanics are significantly more likely than non-Hispanic Whites and non-Hispanic Blacks to receive late diagnoses, as Chen et al. (2011) found in their extensive review of the subject. Among Hispanics, males are more likely than females to be diagnosed late. Foreign-born Hispanics are also at increased risk of late diagnosis. In 2005, 55% of diagnoses among PLWH born in Mexico and 59% of diagnoses among PLWH born in Central America were late, compared with 40% of diagnoses among U.S.-born Hispanics (Chen et al., 2011). This is likely due to barriers in access to care and limited English proficiency (Chen et al., 2011).

Chapter 3: Methods

Data

My data is 48,302 HIV/AIDS cases that were diagnosed between 2010 and 2015, from four hundred counties and regions in thirteen states (Colorado, Kansas, Louisiana, Maryland, Minnesota, Nebraska, New Hampshire, Nevada, Tennessee, Texas, Virginia, and Washington). The states and time period were selected based on data availability.

The dependent variable was a binary variable—one if the case was diagnosed late, zero if it was not. The data from Colorado, Kansas, Louisiana, Maryland, Nevada, Virginia, and Washington came from reports published on each state health department's website. The data from Minnesota, Nebraska, New Hampshire, Tennessee, and Texas was prepared for me by each state health department in response to my data requests. Some states (Kansas, Louisiana, Nebraska, Nevada, New Mexico, Tennessee, and Virginia) had a policy against publishing or releasing county-level data because of confidentiality concerns. For these states I used larger regions made up of several counties. For example, Tennessee was divided into its fourteen regional health departments, as seen in Figure 2.

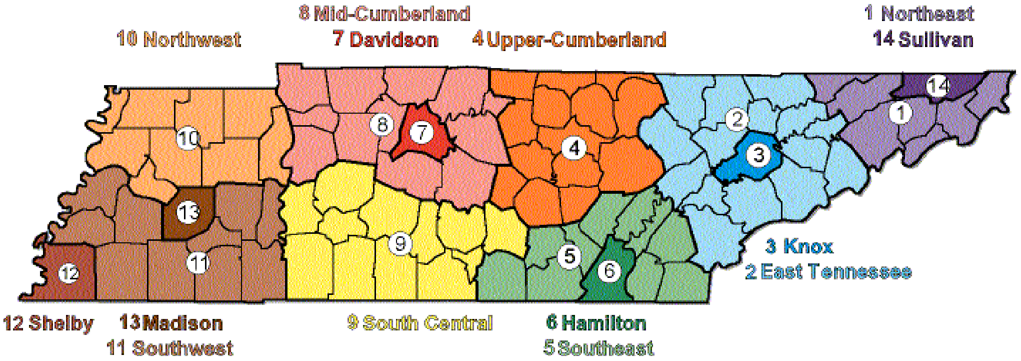


Figure 2: Map of Tennessee Health Department Regions

The Tennessee Health Department could not release county-level data to me for confidentiality reasons, so instead they provided data from each of the fourteen Health Department Regions. Six other states were also divided into regions in a similar way.

I studied six independent variables, using data from the U.S. Census website and from AIDSVu, which is an interactive online map of HIV/AIDS prevalence in the United States, using data from the CDC's national HIV surveillance database. It was created by Emory University's Rollins School of Public Health in partnership with Gilead Sciences, Inc. and the Center for AIDS Research at Emory University. The independent variables are described in Table 1.

Name	Label	Description	Source
Population density	density	Population per square mile, 2010	Census
Percent in poverty	poverty	Persons in poverty, percent, 2016 estimate	Census
Percent Black	black	Black or African American alone, percent, 2016 estimate	Census
Percent Hispanic	hispanic	Hispanic or Latino, percent, 2016 estimate	Census
Percent other non-white	other	Not white, black/African American, or Hispanic/Latino, percent, 2016 estimate	Census
Prevalence	prevalence	People living with diagnosed HIV per 100,000 people, 2014	AIDSVu

Table 1: Independent variables

The first column is the name by which I will refer to this variable. The second column is the one-word label I used in the analysis of the data. The third column is a description of the variable. The fourth column is the source of the data for this variable.

Originally I intended to use variables for each racial/ethnic group (non-Hispanic White, Black/African American, American Indian/Alaska Native, Asian, Native Hawaiian/Pacific Islander, and Hispanic/Latino). However, this resulted in a collinearity problem, as each variable depended on the others; if one percentage was higher, then the other percentages would be lower. Using a binary variable (White vs. non-White) would solve the collinearity problem, but it would mask differences among racial/ethnic groups. Studies have found that Blacks have a lower rate of late diagnosis than Whites and Hispanics have a higher rate than Whites (Chen et al., 2011), so these differences are important. Instead I took out the White variable and combined American Indian/Alaska Native, Asian, and Native Hawaiian/Pacific Islander into one category.

Because most counties are majority White, taking out this variable will eliminate most (though not all) of the collinearity problem, while including all of the non-White groups ensures that it is still indirectly represented (the “null model” in which they are all zero would be an all-White county). Though combining American Indian/Alaska Native, Asian, and Native Hawaiian/Pacific Islander into one category will mask the differences between these groups, they make up a small enough percentage of PLWH in the U.S. that I did not think it would significantly reduce the quality of my data.

Logistic Regression

To analyze the data, I performed a logistic regression analysis. A regression analysis is a type of statistical analysis that estimates the relationship between one or more independent variables and a dependent response variable. A simple linear regression model has one independent variable, X, and one dependent variable, Y. The model assumes that these variables are related to each other by the equation,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where ε is a random error with an expected value of zero. The coefficient β_1 shows the relationship between X and Y. For example, say Y was the number of new HIV/AIDS cases in a particular state and X was the number of dollars per capita spent on HIV/AIDS prevention in that state. If for each dollar increase in funding per capita, the number of HIV/AIDS cases dropped by ten, then β_1 would be -10. β_0 is the y-intercept, in this case the number of cases when there is zero funding for HIV/AIDS prevention.

To estimate β_0 and β_1, you would make n observations of the form (x_i, Y_i), where i = 1, …, n. Then you would use the least squares method, which finds the values of $\hat{\beta}_0$

and $\hat{\beta}_1$ that minimize the sum of the squared differences between the observed value Y_i and the predicted value $\hat{\beta}_0 + \hat{\beta}_1 x_i$ for each i. Put symbolically, it finds the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. This is done by taking the derivatives of this expression with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and setting them equal to zero, as in Figure 4.

$$\frac{\partial}{\partial \beta_0} \sum_i (Y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_i (Y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_i (Y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_i x_i (Y_i - \beta_0 - \beta_1 x_i) = 0$$

Figure 3: Finding estimates for β_0 and β_1 in a simple linear regression model

From the University of Hong Kong Department of Statistics and Actuarial Science.

Solving these equations results in $\hat{\beta}_0 = \bar{Y} - (S_{xy}/S_{xx})\bar{x}$ and $\hat{\beta}_1 = S_{xy}/S_{xx}$, where

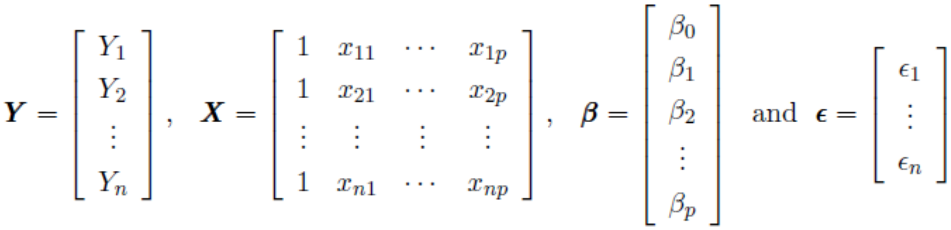
$$S_{xy} = \textstyle\sum_i x_i y_i - n\bar{x}\bar{Y} \text{ and } S_{xx} = \textstyle\sum_i x_i^2 - n\bar{x}^2.$$

Multiple linear regression is the same as simple linear regression except there is more than one explanatory independent variable, modeled by the equation,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

where p is the number of explanatory variables. Extending our previous example, say that Y is the number of HIV/AIDS cases in a state, X_1 is funding per capita, X_2 is the poverty rate, and X_3 is the percentage of the population that is white. Then β_1 will tell you the relationship between funding and HIV/AIDS incidence when poverty and percent white are controlled. In other words, each β_j shows the individual contribution of the jth variable to the number of new HIV/AIDS cases in the state.

To find the estimates of each β_j, make n observations of the form (x_{i1}, x_{i2}, …, x_{ip}, Y_i), where i = 1, …, n. Let



$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

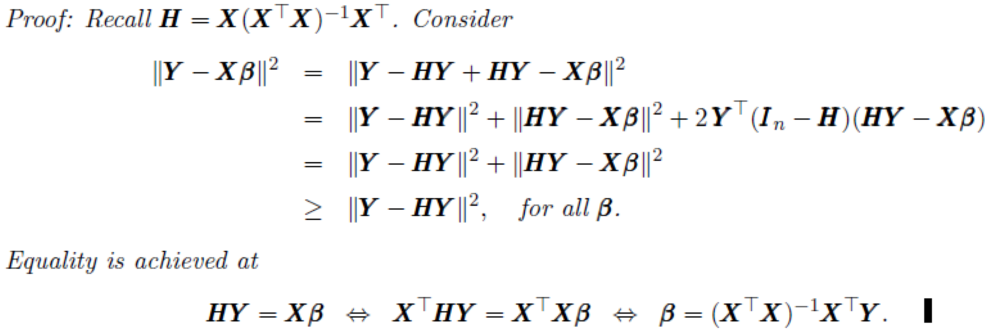
Figure 4: Variables in a Multiple Linear Regression model

From the University of Hong Kong Department of Statistics and Actuarial Science.

Using the least squares method, find $\boldsymbol{\beta}$ that minimizes $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$. The result is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

The proof is given in Figure 5.



Proof: Recall $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. *Consider*

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2 &= \|\boldsymbol{Y} - \boldsymbol{HY} + \boldsymbol{HY} - \boldsymbol{X\beta}\|^2 \\ &= \|\boldsymbol{Y} - \boldsymbol{HY}\|^2 + \|\boldsymbol{HY} - \boldsymbol{X\beta}\|^2 + 2\boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{HY} - \boldsymbol{X\beta}) \\ &= \|\boldsymbol{Y} - \boldsymbol{HY}\|^2 + \|\boldsymbol{HY} - \boldsymbol{X\beta}\|^2 \\ &\geq \|\boldsymbol{Y} - \boldsymbol{HY}\|^2, \quad \textit{for all } \boldsymbol{\beta}. \end{aligned}$$

Equality is achieved at

$$\boldsymbol{HY} = \boldsymbol{X\beta} \;\Leftrightarrow\; \boldsymbol{X}^\top\boldsymbol{HY} = \boldsymbol{X}^\top\boldsymbol{X\beta} \;\Leftrightarrow\; \boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}. \quad \blacksquare$$

Figure 5: Proof that the least squares estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$

From the University of Hong Kong Department of Statistics and Actuarial Science.

Logistic regression is a variation on multiple regression that is used when the observed dependent variables Y_i are binary and the predicted value Y is a probability. This is the case in my analysis—the observed variables are binary (Y_i = 1 if the person was diagnosed late, Y_i = 0 otherwise) and the predicted value is the probability of being diagnosed late in a particular county given the county's demographic variables. In the multiple linear regression model, we had the equation,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

that would predict an outcome Y given a set of independent variables $X_1, \ldots, X_p$. However, a probability must be between 0 and 1, so we must ensure that the predicted probability is in this range. The logistic function,

$$\sigma(t) = (1 + \mathrm{e}^{-t})^{-1},$$

takes any real number t and maps it onto the range [0, 1]. This allows us to take Y, a linear combination of $X_1, \ldots, X_p$, and plug it into the logistic function to get a probability. Therefore, we can predict the probability of late diagnosis like so:

$$\mathrm{P}(\text{late diagnosis} \mid X_1, \ldots, X_p) = (1 + \mathrm{e}^{-Y})^{-1}, \text{ where } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Performing a logistic regression analysis will give you the estimated coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_1$ of the equation,

$$\log(p/(1 - p)) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p,$$

where p is the predicted probability of late diagnosis. To find the relationship between each $\hat{\beta}_i$ and the probability of late diagnosis, find

$$\mathrm{e}^{\beta 0 + \beta i}/(1 + \mathrm{e}^{\beta 0 + \beta i}) - \mathrm{e}^{\beta 0}/(1 + \mathrm{e}^{\beta 0}).$$

This is the difference between the probability if $X_i = 1$ and all of the other X's are zero and the probability if all of the X's are zero. This is equivalent to the change in probability for each unit increase of X_i. For example, say we perform a logistic regression analysis and the coefficient of poverty is 0.006222 and the intercept is 0.6565. Then the change in probability for each unit increase of poverty is

$$\mathrm{e}^{0.6565 + 0.006222}/(1 + \mathrm{e}^{0.6565 + 0.006222}) - \mathrm{e}^{0.6565}/(1 + \mathrm{e}^{0.6565}) \approx 0.0014$$

This would mean that for each percentage point increase in poverty, the percentage of HIV cases diagnosed late increases by 0.14 percentage points.

A fundamental assumption of regression analyses is that the independent variables are independent from one another. If two or more variables are highly correlated, they will compete and cancel out each other’s effects. This results in high standard errors and changes in the magnitudes and possibly the signs of certain variables, making it difficult to assess their relative importance. A common method to check for multicollinearity is to find the Variance Inflation Factor of each variable. The VIF is the variance of the model with all of the variables divided by the variance of the model with one variable, which shows how much the variance of that particular variable is increased by collinearity. A typical rule of thumb is that VIF’s should not exceed ten. If this is the case, the model may need to be adjusted, either by collecting more data or eliminating some variables.

RStudio

All of my analysis was done using RStudio. I used the GLM function to perform the regression, the VIF function in the CAR package to find the Variance Inflation Factors, and the pairs and histogram functions to visualize the data.

Chapter 4: Results

First I looked at a scatterplot of each independent variable graphed against the proportion of cases that were diagnosed late, as shown in Figure 6. These graphs show the relationship between each variable and late diagnosis when none of the other variables are controlled. Population density, percent Black, percent other non-White, and prevalence were all negatively correlated with late diagnosis, while percent in poverty and percent Hispanic were positively correlated.

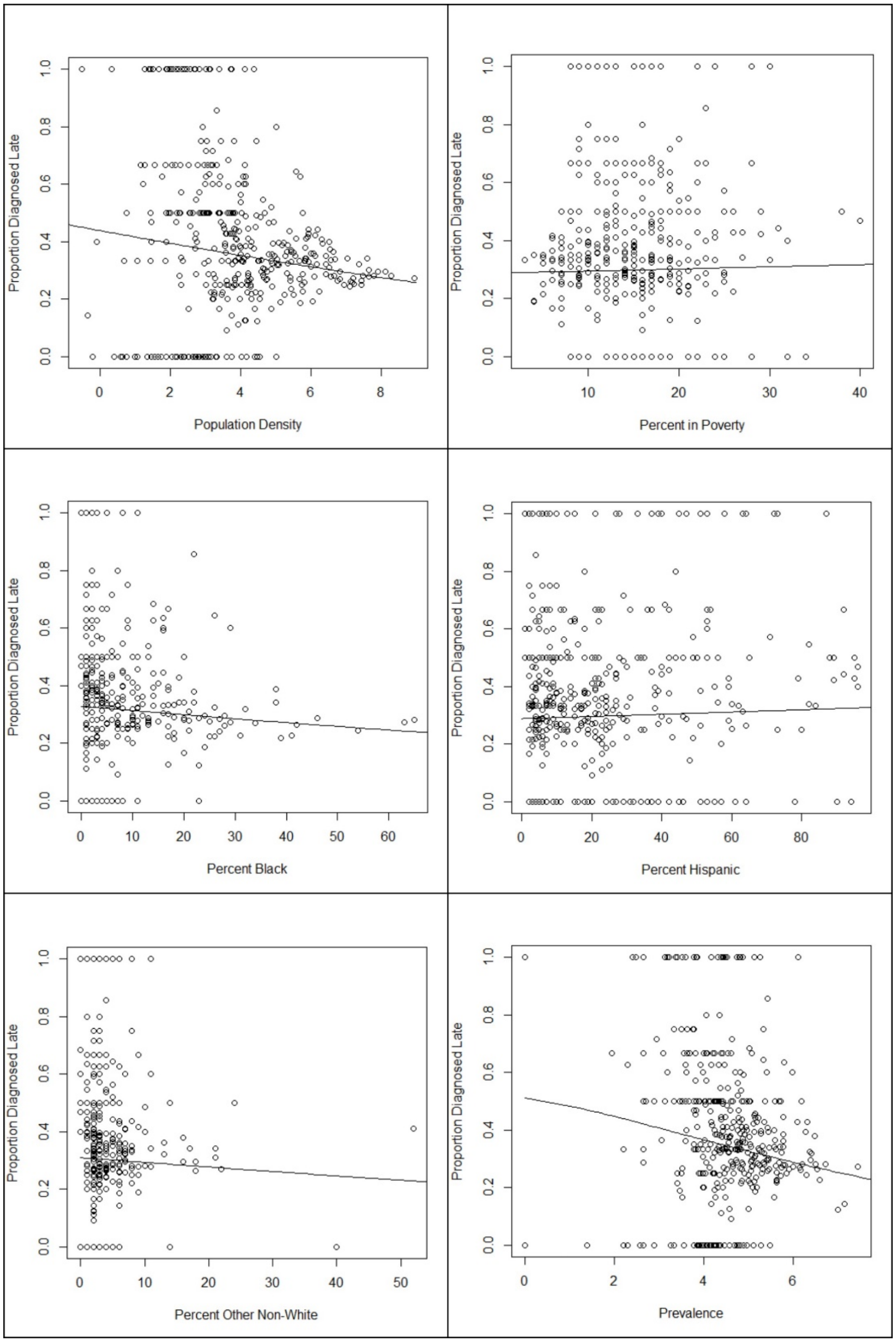


Figure 6: Scatterplots of each independent variable graphed against the percentage of cases diagnosed late

Note: I took the log of population density and prevalence to make them more normally distributed.

Next I ran the logistic regression model without any adjustments to the variables. The p-values are shown in Table 2.

Variable	**P-value**
Population density	0.334
Percent in poverty	0.020
Percent Black	Less than 0.001
Percent Hispanic	0.021
Percent other non-White	Less than 0.001
Prevalence	0.073

Table 2: p-values in initial analysis

The results of the initial analysis were that percent Black and percent other non-White were highly significant in explaining variation in late diagnosis, percent in poverty and percent Hispanic were significant, prevalence was slightly significant, and density was not significant.

These results showed that percent Black and percent other non-White were the most significant variables in explaining variation in late diagnosis. Percent in poverty and percent Hispanic were also significant, though not to the same extent. Prevalence had a p-value of 0.073, putting it just above the conventional 0.05 threshold, so this analysis would not demonstrate the significance of it in explaining variation in late diagnosis. Population density had a very high p-value so it would not be considered significant. Poverty was positively correlated with late diagnosis, meaning people in counties/regions with high poverty rates were more likely to be diagnosed late. Percent Black, percent Hispanic, percent other non-White, and prevalence were all negatively correlated with late diagnosis.

I checked the Variance Inflation Factor values for each variable to make sure there was not too much collinearity. The results are shown in Table 3.

Variable	**VIF**
Population density	5.187
Poverty	2.006
Black	2.711
Hispanic	1.852
Other	1.550
Prevalence	7.547

Table 3: Variance Inflation Factors in initial analysis

None of the Variance Inflation Factors exceed ten, so multicollinearity problems can probably be safely ignored.

A typical rule of thumb is that no Variance Inflation Factor should exceed ten. The highest VIF here was the value for prevalence, which was 7.55, so I did not think it was necessary to remove any variables. The correlations between each of the independent variables are shown in Table 4.

	Percent in poverty	Percent Black	Percent Hispanic	Percent other non-White	Prevalence
Population density	-0.096	0.425	-0.075	0.252	0.572
Percent in poverty		0.174	0.584	-0.213	0.204
Percent Black			-0.185	0.042	0.626
Percent Hispanic				-0.252	-0.008
Percent other non-White					0.039

Table 4: Correlations between independent variables

Correlation coefficients are between -1 and 1, where 1 means there is a perfect positive relationship, -1 means there is a perfect negative relationship, and 0 means there is no relationship.

The strongest correlations were between percent Black and prevalence (0.626), percent Hispanic and percent in poverty (0.584), and population density and prevalence

(0.572). All of these were positively related. There was a weak positive correlation between population density and percent Black (0.425). Because percent Black and population density are both positively correlated with prevalence, it may be that some of each of their correlations is explained by the other variable. The rest of the variables were very weakly related to each other. None of the correlations between the race/ethnicity variables had a magnitude greater than 0.252, which shows that taking out the percent White variable likely reduced most of the collinearity issues associated with using percentages. The strongest correlation between the race/ethnicity variables was percent Hispanic vs. percent other non-White (-0.252), which is likely because there were several counties in Texas with extremely high Hispanic populations (up to 96%). In these counties the high percentage of Hispanic residents would significantly reduce the percentages of all the other races/ethnicities, which is probably why percent Hispanic was negatively associated with both percent Black and percent other non-White.

The distributions of and relationships between each of the variables is shown in Figure 7.

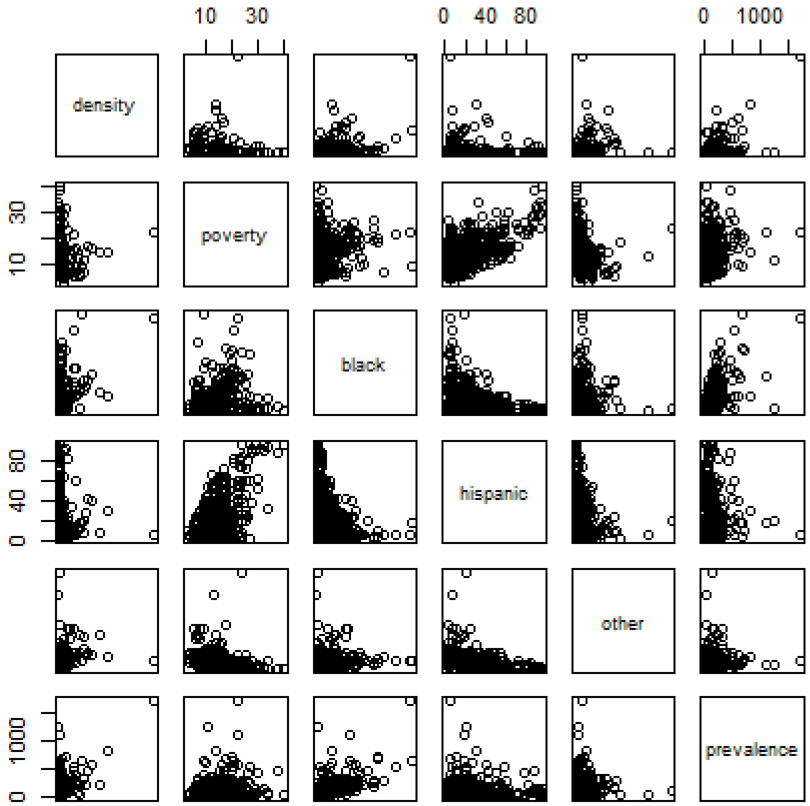


Figure 7: Relationships between variables

Each square shows a scatterplot of the data with the variable in same column on the x-axis and the variable in the same row on the y-axis. For example, the second square in the first row shows a scatterplot of poverty vs. density, with poverty on the x-axis and density on the y-axis.

All of the scatterplots with population density had most of the points concentrated on one side of the graph, so I looked at a histogram of the data.

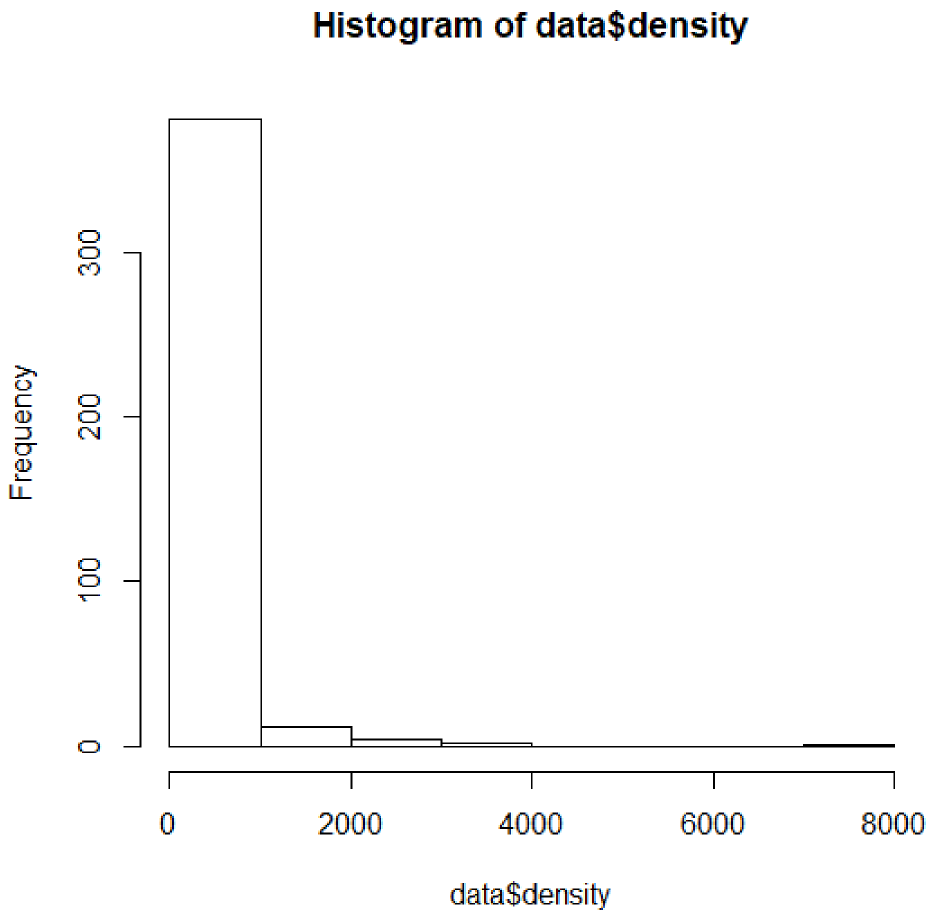


Figure 8: Histogram of population density

Almost all of the population densities are less than 1,000, with only a few outliers.

The histogram showed that almost all of the counties/regions in my data set had population densities under 1,000, with just a few outliers (the counties that contained cities such as Baltimore, Denver, Saint Paul, Dallas, and Houston). An assumption of regression analysis is that the variables are somewhat normally distributed, but this data was heavily skewed. To get a distribution closer to a normal distribution, I took the log of population density.

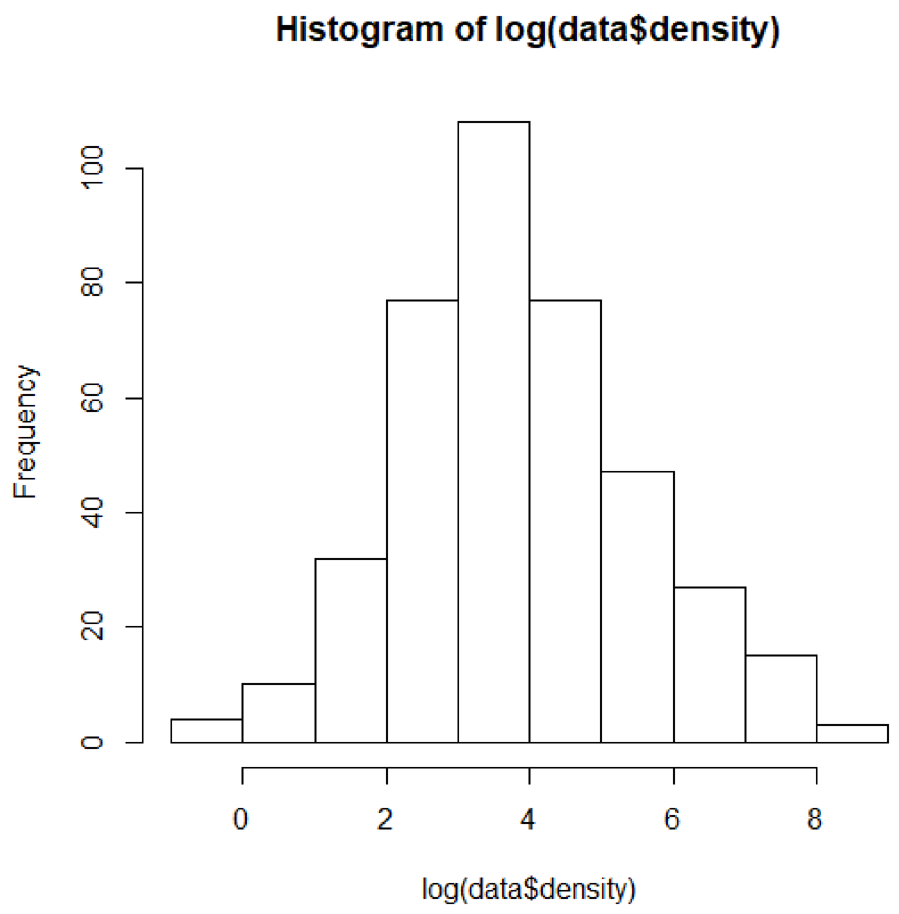


Figure 9: Histogram of the log of population density

The logs of the population densities are much closer to being normally distributed.

This resulted in an approximately normal distribution, so I replaced population density with the log of population density in my model. Prevalence was distributed similarly to population density, so I took the log of this variable as well.

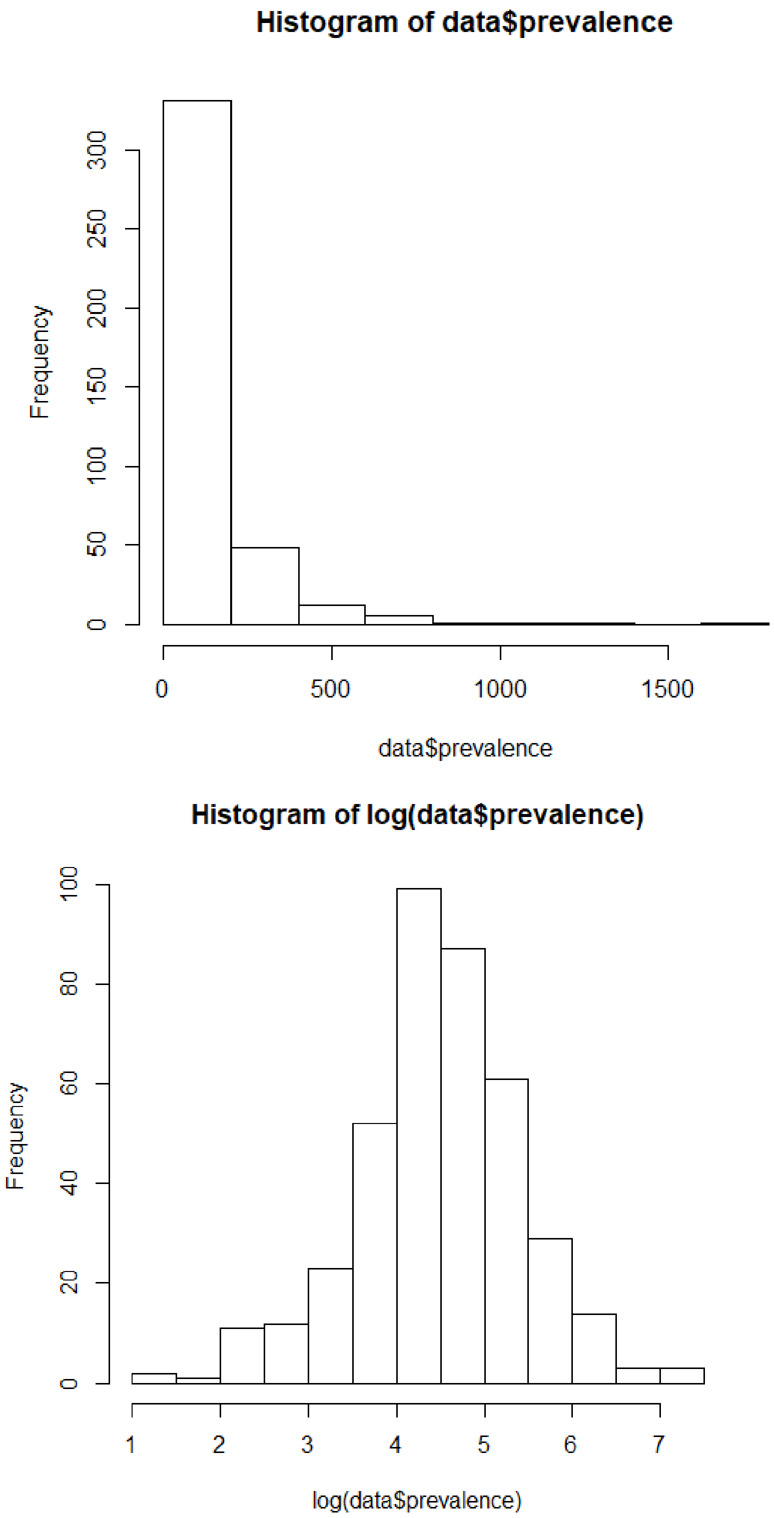


Figure 10: Histograms of prevalence and log of prevalence

Log(prevalence) is much closer to a normal distribution than prevalence.

The new p-values are shown in Table 5.

Variable	**P-value**	**Change in probability**
Log(population density)	Less than 0.001	-0.0112
Percent in poverty	0.077	0.0012
Percent black	0.277	-0.0003
Percent Hispanic	0.019	0.0004
Percent other non-white	0.885	-0.0001
Log(prevalence)	Less than 0.001	-0.0292

Table 5: p-values and changes in probability in final analysis

The new results showed that log(population density) and log(prevalence) were highly significant, percent Hispanic was significant, percent in poverty was slightly significant, and percent Black and percent other non-White were not significant.

The change in probability shows how much the probability of late diagnosis changes with each unit change in the variable. For example, for each one-unit increase in the log of population density, the probability of late diagnosis decreases by 0.0112 percentage points.

The log of population density and the log of prevalence were now the most significant variables in the model. Percent Hispanic was also significant, though not to the same extent, and poverty had a p-value of approximately 0.08, putting it slightly over 0.05. Percent Black and percent other non-White were no longer significant. Poverty and percent Hispanic were positively correlated with late diagnosis, while population density, percent Black, percent other non-White, and prevalence were negatively correlated.

The “change in probability” in Table 5 shows how much the probability of late diagnosis changes with each unit change of the variable in question. Each time the log of the population density increases by one, the probability of late diagnosis decreases by approximately one percentage point. This means that for each ten-fold increase in the population density (say from 100 to 1,000), the probability of late diagnosis decreases by one percentage point (say from 30% to 29%). For each percentage point increase in the poverty rate (say from 15% to 16%), the probability of late diagnosis increases by about 0.1 percentage points (say from 30% to 30.1%). For each percentage point increase in the percent of the population who is Black/African American, the probability of late diagnosis decreases by about 0.03 percentage points. For each percentage point increase in the percent of the population who is Hispanic/Latino, the probability of late diagnosis increases by about 0.04 percentage points. For each percentage point increase in the percent of the population who is not White, Black, or Hispanic, the probability of late diagnosis decreases by about 0.01 percentage points. And for each ten-fold increase in prevalence, the probability of late diagnosis decreases by about 3 percentage points.

Chapter 5: Conclusion

My analysis demonstrated that there is a strong negative correlation between population density and late diagnosis and between prevalence and late diagnosis. People with HIV in counties that are sparsely populated and with low rates of HIV/AIDS are more likely to be diagnosed late. This trend holds even when poverty and race/ethnicity are controlled. The relationship between population density and late diagnosis reflects the finding by Trepka et al. (2014) that people living in rural areas of Florida were more likely to be diagnosed late than people living in urban areas, and shows that the same is likely true for the thirteen states in my study.

The relationships between poverty, percent Black, and percent Hispanic are in line with the existing literature on late diagnosis. Poverty is positively associated with late diagnosis (Mukolo et al. 2013), which would explain the positive relationship in my analysis between the percentage of people below the poverty line and the percentage of HIV/AIDS cases diagnosed late. My results showed a negative relationship between percent Black and late diagnosis and a positive relationship between percent Hispanic and late diagnosis, which reflects the finding that Blacks/African Americans have the lowest rate of late diagnosis of any racial/ethnic group and Hispanics/Latinos have the highest (Chen et al. 2011).

The most important implication of my analysis is that there needs to be more research on late diagnosis in rural areas of the United States with low prevalence of HIV/AIDS. My results show that these places tend to have the highest rates of late diagnosis, but they are also the least studied parts of the country. Not only are these areas under-studied, but there has also been very little research on the correlation between population density and late diagnosis and between prevalence and late diagnosis.

An essential first step would be identifying the reasons why these places have higher rates of late diagnosis. My results demonstrate correlation but not causation; they show that people in rural, low-prevalence areas are more likely to be diagnosed late, but they do not show why this is. My guess would be that lower perceived risk is a major factor. Many studies have demonstrated a negative relationship between perceived risk and late diagnosis. This is likely a reason why subgroups with higher prevalence of HIV/AIDS (such as Blacks/African Americans, MSM, and injection drug users) have the lowest rates of late diagnosis. It would make sense that people living in areas where HIV/AIDS is highly prevalent would have higher perceived risk; they would be more likely to know people with HIV/AIDS, which would increase their own awareness of the risk. It is also possible that state health departments devote more resources to raising awareness of HIV/AIDS in high-prevalence areas, which might prompt people to get tested who would otherwise not have considered it. Another possibility is that people in rural areas may have less access to HIV testing facilities.

Future research on this topic might involve interviewing people with diagnosed HIV/AIDS about when they decided to get tested and what prompted them to do so. Comparing the answers of people in urban, high-prevalence areas and rural, low-prevalence areas could shed light on the reasons for the disparity. A limitation of my data is that it does not show any information about the individuals themselves, only the demographics of the counties in which they live. Interviews would resolve this issue.

A broader implication of my analysis is the importance of not letting low-prevalence areas slip through the cracks in public health research. It makes sense that the majority of studies of a particular disease should take place in areas where the disease is more prevalent, because they will represent a larger number of people. However, there still should be some attention given to low-prevalence areas, as the needs of these places may be dramatically different. The vast majority of studies of late diagnosis of HIV/AIDS occur in high-prevalence cities, but late diagnosis occurs at a higher rate in low-prevalence rural areas. As a result, the needs of people with HIV/AIDS in these places are being overlooked. My hope is that this study will prompt further research of late diagnosis in rural, low-prevalence parts of the United States so that people with HIV/AIDS in these places can begin treatment sooner and enjoy a longer life expectancy and better quality of life.

Bibliography

AIDS.gov. U.S. Department of Health & Human Services, www.aids.gov. Accessed 29 Apr. 2017.

AIDSVu.org. Emory University Rollins School of Public Health. Accessed 10 April 2018.

"About HIV/AIDS." *CDC*, March 16, 2018, https://www.cdc.gov/hiv/basics/whatishiv.html. Accessed 22 April, 2018.

Castel, Amanda D., Manya Magnus, and Alan E. Greenberg. "Update on the Epidemiology and Prevention of HIV/AIDS in the USA." *Current epidemiology reports* 2.2 (2015): 110-119.

Centers for Disease Control and Prevention (CDC). "Monitoring Selected National HIV Prevention and Care Objectives by Using HIV Surveillance Data—United States and 6 Dependent Areas—2011." *HIV Surveillance Supplemental Reports* 18.5: 1-47 (2012).

Centers for Disease Control and Prevention (CDC). "Monitoring Selected National HIV Prevention and Care Objectives by Using HIV Surveillance Data: United States and 6 Dependent Areas, 2014." *HIV Surveillance Supplemental Reports* 21.4: 1-87 (2015).

Colorado Department of Public Health and Environment (CDPHE). "HIV Surveillance Semiannual Report, 4th Quarter 2015." *Biannual HIV Surveillance Reports* (2015).

Chen, Nadine E., Joel E. Gallant, and Kathleen R. Page. "A systematic review of HIV/AIDS survival and delayed diagnosis among Hispanics in the United States." *Journal of Immigrant and Minority Health* 14.1 (2012): 65-81.

Farnham, Paul G., PhD, Chaitra Gopalappa, PhD, Stephanie L. Sansom, PhD, Angela B. Hutchinson, PhD, John T. Brooks, MD, Paul J. Weidle, PharmD, Vincent C. Marconi, MD, and David Rimland, MD. "Updates of lifetime costs of care and quality-of-life estimates for HIV-infected persons in the United States: late versus early diagnosis and entry into care." *JAIDS Journal of Acquired Immune Deficiency Syndromes* 64.2 (2013): 183-189.

Girardi, Enrico, MD, Caroline A. Sabin, PhD, and Antonella d'Arminio Monforte, MD. "Late diagnosis of HIV infection: epidemiological features, consequences and strategies to encourage earlier testing." *JAIDS Journal of Acquired Immune Deficiency Syndromes* 46 (2007): S3-S8.

Hall, H. Irene, Jessica Halverson, David P. Wilson, Barbara Suligoi, Mercedes Diez, Stéphane Le Vu, Tian Tang, Ann McDonald, Laura Camoni, Caroline Semaille, and Chris Archibald. "Late diagnosis and entry to care after diagnosis of human immunodeficiency virus infection: a country comparison." *PloS one* 8.11 (2013): e77763.

New Mexico's Indicator-Based Information System (NM-IBIS). "Health Indicator Report of HIV Infection – New Stage 3 Infections." *New Mexico Department of Health,* 22 January 2016, https://ibis.health.state.nm.us/indicator/view/HIVAIDSNewStage3.12Mos.Rate.Region.html. Accessed 10 April 2018.

Kansas Department of Health and Environment. "Integrated Epidemiological Profile: An Analysis of the HIV Epidemic in Kansas from 2010 – 2014." *Kansas HIV Epidemiological Profiles* (2015).

Krawczyk, Christopher S., et al. "Delayed access to HIV diagnosis and care: Special concerns for the Southern United States." *AIDS care* 18.S1 (2006): 35-44.

Louisiana Office of Public Health. "2014 STD/HIV Program Report." *Louisiana STD/HIV Annual Reports* (2015).

Marks, Gary, Nicole Crepaz, PhD, J. Walton Senterfitt, PhD, and Robert S. Janssen, MD. "Meta-analysis of high-risk sexual behavior in persons aware and unaware they are infected with HIV in the United States: implications for HIV prevention programs." *JAIDS Journal of Acquired Immune Deficiency Syndromes* 39.4 (2005): 446-453.

Maryland Department of Health and Mental Hygiene. "2012 Maryland Annual HIV Epidemiological Profile." *Maryland Annual Epidemiological Profiles* (2013).

Maryland Department of Health and Mental Hygiene. "2013 Maryland Annual HIV Epidemiological Profile." *Maryland Annual Epidemiological Profiles* (2014).

Maryland Department of Health and Mental Hygiene. "2014 Maryland Annual HIV Epidemiological Profile." *Maryland Annual Epidemiological Profiles* (2015).

Maryland Department of Health and Mental Hygiene. "2015 Maryland Annual HIV Epidemiological Profile." *Maryland Annual Epidemiological Profiles* (2016).

Minnesota Department of Health. Late HIV/AIDS diagnoses in Minnesota by county between 2010 and 2015. Unpublished data. (2017).

Mugavero, Michael J., MD, Chelsea Castellano, BS, David Edelman, MD, MHS, Charles Hicks, MD. "Late diagnosis of HIV infection: the role of age and sex." *The American journal of medicine* 120.4 (2007): 370-373.

Mukolo, Abraham, Raquel Villegas, Muktar Aliyu, and Kenneth A. Wallston. "Predictors of late presentation for HIV diagnosis: a literature review and suggested way forward." *AIDS and Behavior* 17.1 (2013): 5-30.

Nebraska Department of Health and Human Services. Late HIV/AIDS diagnoses in Nebraska by health department between 2010 and 2015. Unpublished data. (2017).

Nevada Division of Public and Behavioral Health. "HIV Epidemiological Profile: 2014 Update." *HIV Epidemiological Profiles* (2015).

Tennessee Department of Health. Late HIV/AIDS diagnoses in Tennessee by public health region between 2010 and 2015. Unpublished data. (2017).

Texas Department of State Health Services. Late HIV/AIDS diagnoses in Texas by county between 2010 and 2015. Unpublished data. (2017).

Trepka, Mary Jo, MD, MSPH, Kristopher P. Fennie, PhD, MPH, Diana M. Sheehan, MPH, Khaleeq Lutfi, MPH, Lorene Maddox, MPH, and Spencer Lieb, MPH. "Late HIV diagnosis: differences by rural/urban residence, Florida, 2007–2011." *Aids patient care and STDs* 28.4 (2014): 188-197.

United States Census Bureau. "QuickFacts." *Census*, https://www.census.gov/quickfacts. Accessed 10 April 2018.

University of Hong Kong Department of Statistics and Actuarial Science. "STAT2301/3600 Linear Statistical Analysis." Unpublished course outline. (2017).

Virginia Department of Health. "HIV Disease in the Central Region." *Regional Statistics* (2015).

Virginia Department of Health. "HIV Disease in the Eastern Region." *Regional Statistics* (2015).

Virginia Department of Health. "HIV Disease in the Northern Region." *Regional Statistics* (2015).

Virginia Department of Health. "HIV Disease in the Northwest Region." *Regional Statistics* (2015).

Virginia Department of Health. "HIV Disease in the Southwest Region." *Regional Statistics* (2015).

Washington State Department of Health. "HIV Surveillance Semiannual Report, 2016 Edition." *HIV Surveillance Semiannual Reports* (2017).