2012-03-06

The Effect of Smoking on Tuberculosis Incidence in Burdened Countries

Natalie Noel Ellison
Brigham Young University - Provo

The Effect of Smoking on Tuberculosis Incidence in Burdened Countries

Natalie Noel Ellison-Munson

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Bruce J. Collings, Chair
Dennis H. Tolley
Candace Berrett

Department of Statistics

Brigham Young University

April 2012

ABSTRACT

The Effect of Smoking on Tuberculosis Incidence in Burdened Countries

Natalie Noel Ellison-Munson
Department of Statistics, BYU
Master of Science

It is estimated that one third of the world's population is infected with tuberculosis. Though once thought a "dead" disease, tuberculosis is very much alive. The rise of drug resistant strains of tuberculosis, and TB-HIV coinfection have made tuberculosis an even greater worldwide threat. While HIV, poverty, and public health infrastructure are historically assumed to affect the burden of tuberculosis, recent research has been done to implicate smoking in this list.

This analysis involves combining data from multiple sources in order determine if smoking is a statistically significant factor in predicting the number of incident tuberculosis cases in a country. Quasi-Poisson generalized linear models and negative binomial regression will be used to analyze the effect of smoking, as well as the other factors, on tuberculosis incidence. This work will enhance tuberculosis control efforts by helping to identify new hypotheses that can be tested in future studies. One of the main hypotheses is whether or not smoking increases the number of tuberculosis cases above and beyond the effects of other factors that are known to influence tuberculosis incidence. These known factors include TB-HIV coinfection, poverty and public health infrastructure represented by treatment outcomes.

Keywords: Smoking, Tuberculosis, Poisson Regression, Negative Binomial Regression

ACKNOWLEDGMENTS

Upon finishing this thesis, I am overwhelmed with gratitude. There are so many people who have helped me on the road to receiving my Master's degree. I would first like to thank my mother. For anyone who is familiar with my statistics career, you would know that I would never have received this degree without her help. I am also grateful for her constant willingness to be an additional statistical consultant when I was in need of assistance. My father and siblings have always provided constant support and encouragement, which I am extremely grateful for. So many people having faith in you makes you feel as though your desires really can come true. Last but not least I would like to thank my husband, who has been a great support. Not only did he sleep in the hall of the Talmadge building while I stayed all night doing homework, but he still married me even through the stress of the past two years. Thanks to my mentors, especially Dr. Scott, Dr. Collings, Dr. Tolley and Dr. Berrett for their valuable input and direction, and to all of you who believed I could accomplish this. Because of you all, I did!

Last and most importantly, I thank my Heavenly Father for all of the above, and for the opportunity to attend this amazing university, to learn and grow academically and spiritually. And I am especially grateful for the Gospel of Jesus Christ in my life and the mission I served in Ukraine. Growing to love the people there and coming home with a positive TB test, helped me to realize the importance of controlling this disease for so many around the world. I hope this research helps others in their quest to stop the spread of TB in every country.

CONTENTS

CHAPTER 1

TUBERCULOSIS: A REEMERGING DISEASE

Tuberculosis (TB) is a contagious bacterial infection caused by *Mycobacterium tuberculosis*. TB usually affects the lungs (pulmonary TB), however, other parts of the body can also be affected (extrapulmonary TB). TB is an airborne disease; this means it can be spread by a cough or a sneeze from a person with active TB. There are two types of TB, latent and active. Latent TB infection describes a person who has TB bacteria in their body, but has yet to develop symptoms. Someone with active TB, commonly referred to as TB disease, is symptomatic and smear positive. Smear positive means that TB organisms are present in the patient's sputum. Symptoms of TB include fever, fatigue, weight loss and a persistent cough (New York State Department of Health 2000). Left untreated, those with active TB have the risk of infecting those they come in contact with.

1.1 Tuberculosis: A Reemerging Threat

While some people consider TB to be a "dead" disease that was eradicated years ago, it is, in fact, very much alive in the world today. It is estimated that more than two billion people are infected with TB bacilli (WHO 2010); this is approximately one third of the world's population. One in every 10 of these individuals with latent TB infection will develop active TB within their lifetime. Each person with active TB infects between 10 and 15 people every year (WHO 2010).

TB is a thriving disease. In 2009, there were 9.4 million incident (new) and 14 million prevalent (total) cases of TB worldwide. This statistic implies that there were more new cases of TB than existing cases. This shows TB to be a quickly spreading disease that when caught can be treated. In addition to being a prevalent disease, TB is also deadly—

approximately 1.7 million people died from TB in 2009 (WHO 2010a). This is approximately 4,700 deaths per day. In order to understand why TB has become a global concern, it is necessary to review the history of the disease.

TB has affected human health for thousands of years—500,000 year old fossils show evidence of TB infection (Avasthi 2007). Consumption, a common name for TB in the past, increased dramatically in Europe and North America during the seventeenth and eighteenth centuries. After this time, the disease began to decline. However, TB prevalence hit its peak in 1800 and remained high for the next 100 years. By the turn of the twentieth century, there were only a few professional and governmental organizations specifically engaged with TB (Murray 2004). The development of antibiotics in the mid-1900s led to near perfect control of tuberculin disease. With these drugs, treatment against TB made revolutionary progress. However, no new drugs have been developed to combat TB in the last 40 years (Migliori et al. 2010), which is one of the main reasons for the resurgence of TB in the late 1900s (Murray 2004).

Poor patient management, non-adherence to treatment, and poor national support of TB control has accelerated the current drug-resistant TB epidemic (Davies 2001). When compared with other viruses and most other bacterial pathogens *mycobacterium Tuberculosis* has slow replication and mutation rates. However, for over forty years, it had ample time to adjust to the current drugs that exist to fight it. The evolving of *mycobacterium Tuberculosis* resulted in drug-resistant strains of TB. There are two forms of drug resistant TB: multi-drug resistant (MDR-TB) and extremely drug-resistant tuberculosis (XDR-TB).

The strains of MDR-TB are resistant to at least the two most powerful first-line anti-TB drugs, isoniazid and rifampicin (Migliori et al. 2010). Among TB patients that were identified in 2009, it is estimated that 250,000 had MDR-TB. In 2008 there were estimated to be

440,000 new cases of MDR-TB worldwide. Worldwide 27 countries account for 86% of these cases; they are termed the 27 high MDR-TB burden countries (WHO 2010a). XDR-TB is MDR-TB that is also resistant to three or more of the six classes of second line drugs (Migliori et al. 2010). By July 2010, at least one case of XDR-TB had been reported by each of 58 countries and territories (WHO 2010a). In addition to drug resistance, TB-HIV coinfection has caused a resurgence in TB.

People with weakened immune systems are at risk of developing active TB (New York State Department of Health 2000). Those who are HIV-positive have weakened immune systems and are therefore more susceptible to TB. In fact, TB is the leading killer among people living with HIV (WHO 2010). Of the 9.4 million incidence cases in 2009, about 1.1 million (12%) were among HIV positive people. In that year, there were an estimated 0.4 million deaths among incidence TB cases that were HIV-positive. This makes up approximately 23.5% of all TB deaths in 2009 (WHO 2010a).

Tuberculosis is an evolving disease, with HIV coinfection and drug resistant strains keeping it alive. These factors further its spread and threaten the containment of this disease. Although TB incidence rates were increasing in the 1990s, overall incidence rates are currently on a decline. However, TB is evolving and posing a potentially imminent threat to public health (WHO 2010a).

1.2 The Need to Control TB

In the following chapter, factors that complicate the tuberculosis problem will be analyzed and statistics will be given. These statistics and results confirm that tuberculosis is a re-emerging worldwide problem. Complications due to smoking together with new drug resistant TB and HIV coinfection present obstacles in attempts to contain this disease. While the burden of TB is very slowly decreasing globally, it is not enough to reach all of the

epidemiological impact targets set for 2015. The World Health Organization (WHO) set a goal to reach TB elimination by 2050, however, with the current statistics, this is unfeasible. In order to reach these goals, specific considerations must be taken to combat TB-HIV coinfection and drug resistant TB. To insure long term TB control, precautions must be taken to limit the impact of TB risk factors, such as smoking (Lönnroth and Raviglione 2008).

CHAPTER 2

KNOWN INFLUENTIAL FACTORS OF TB

2.1 HIV: HIV-TB Coinfection

Human immunodeficiency virus (HIV) can cause people to develop acquired immunodeficiency syndrome (AIDS). A person is diagnosed with AIDS when their helper-T cells (adaptive immune system) levels drop below a critical value. This reduction in adaptive immunity leaves the infected individual with a weakened immune system, making them more susceptible to other diseases. HIV is a retrovirus that forms in many mutated strains, making it difficult to develop an effective vaccine (Bauman 2011). Due to this fact, the disease has continued to spread and has become a worldwide burden. For example, in the United States, approximately 1.1 million people were living with HIV infection at the end of 2006. It was also estimated that as many as 21% are unaware of their infection (CDC 2011).

The symptoms of Tuberculosis are enhanced in the presence of HIV. People who have latent TB infection (LTBI) are 20 to 30 times more likely to develop active disease if they also have an HIV infection. As stated earlier, active TB is contagious and symptomatic. The effects of TB infection on HIV patients is catastrophic: one in every four deaths among people with HIV is due to tuberculosis (CDC 2011). The Center for Disease Control (CDC) and WHO have incorporated HIV-TB coinfection into their treatment plans. The CDC has recommended that all people diagnosed with HIV should be tested for TB as soon as possible. The combination of the HIV epidemic and the spread of drug resistance are main barriers to dealing effectively with TB (Blöndal 2007).

2.2 Poverty

Poor social and economic circumstances have been associated with ill health by almost all health indicators. The effects of deprivation are specifically related to respiratory and infectious diseases, both of which describe tuberculosis. Historically TB has been associated with high levels of poverty. The decline in tuberculosis could potentially be attributed not only to more effective treatments, but also the reduced overcrowding, and improved nutrition and social conditions (Spence et al. 1993).

A study was performed in Liverpool, England, to evaluate whether or not the historic link between poverty and tuberculosis still exists. This was a retrospective study performed in 1993 that involved 344 residents of Liverpool that had tuberculosis between 1985 and 1990 inclusive. For that six year period, the number of all recorded forms of tuberculosis was correlated with four indices of poverty: council housing, free school meals, the Townsend overall deprivation index, and the Jarman index. The results showed the highest correlation between the rate of tuberculosis and the Jarman index, with $r = 0.73$ and a corresponding p-value less than 0.0001. The Jarman index includes an index of ethnic minorities in addition to some aspects of social deprivation (Spence et al. 1993). Ethnic minorities appear to still be correlated with TB spread. For example, recent outbreaks in the United States have occurred in port cities, such as New York City and San Francisco, that contain a lot of immigrants and ethnicities.

In addition to the Jarman and Townsend indices, this study showed statistically significant correlations with the simple indices of deprivation: council housing and children receiving free school meals. This indicates that the historical link between poverty and tuberculosis still exists. While the exact connection remains unclear and unproven, some connections can be inferred. Poverty usually results in fairly poor nutrition, which likely leads to a

compromised immune system. The immunocompromised individual is susceptible to any opportunistic pathogen, such as *mycobacterium Tuberculosis*.

Poverty also implies smaller living quarters for more people. Since tuberculosis is an airborne disease, overcrowded living conditions facilitate its spread (Spence et al. 1993). In another retrospective study, it was indicated that substandard housing conditions increased the probability (hazard ratio 1.07-3.11, p-value=0.03) that an individual will default from (fail to complete) tuberculosis treatment (Franke et al. 2008). Defaulting in treatment leads to additional spreading of tuberculosis and the development of drug resistant strains of tuberculosis.

2.3 Public Health Infrastructure: Treatment Outcomes

Poverty and public health infrastructure are connected factors. The more money a country has, the more money they are able to spend on health care, and vice versa. The main risk factor for developing *Mycobacterium tuberculosis* drug resistance is previous treatment against TB. When compared with new patients, previously treated tuberculosis cases were significantly more likely to have resistance to one, two, three, or four drugs. This study also showed that as the total time of prior anti-TB treatment increased, there was a linear increase in the likelihood of having MDR-TB (Migliori et al. 2010). Patients who default from treatment are also more likely to infect other individuals. In 2009, it was estimated that 3.3% of all new TB cases had MDR-TB (WHO 2010a). While previous treatment is the main risk factor for developing drug resistance, it is apparent that increasingly more people become initially infected with drug resistant strains of tuberculosis. Treatment of drug resistant strains of TB is also more complicated. Treatment of MDR-TB requires longer treatment time and expensive second-line drugs that have harsher side effects (Blöndal 2007).

In the 1990s, WHO developed a new strategy for tuberculosis containment: Directly Ob-

served Treatment Short-Course (DOTS). In this program, someone associated with the program watches the patients take their medication everyday, and keeps a record of daily intake. The program workers are stationed at a clinic or make house calls to watch patients take their medication. General tuberculosis treatment lasts 6-9 months and requires daily medication. Adhering to this daily regiment is necessary for treatment success. In short, DOTS makes sure that individuals adhere to treatment. In 2003, countries involved in the DOTS program achieved on average 82% treatment success. In 2004, approximately 83% of the world's population was living in countries or parts of countries covered by this strategy (Blöndal 2007). From the above facts, it is easy to see why successful treatment outcomes are so important. The better a country is able to deal with its existing cases of TB, the greater the decrease in the number of new TB cases and drug resistant cases.

2.4 Smoking and Tuberculosis

While gender, age, and ability to pay are associated with effective tuberculosis control, this project addresses one of the barriers to tuberculosis control governed by personal choice: smoking. If smoking is truly a risk factor for TB, it is a very prevalent factor. Tobacco use is the leading cause of preventable death. Currently it kills more than five million people every year (WHO 2009b). As regulations on smoking have become more strict in the United States, tobacco companies have had to turn to other markets. Therefore, the burden of tobacco smoking is highest in low- and middle-income countries where public health infrastructure is weak (WHO 2009b).

Tobacco consumption and TB have had a long history. In the United States in the early 1900s, individuals who chewed tobacco were encouraged to switch to smoking. It was believed that spitting chewing tobacco facilitated the transmission of TB. They believed that smoking would avoid this transmission and improve public health. However, in light of new research, this message may have inadvertently encouraged a more risky behavior. Studies

have shown that smoking is associated with increased risk of TB infection, increased risk of conversion from latent to active TB, and increased risk of mortality and symptoms due to TB (Hassmiller 2006).

In Hong Kong, China, a study was performed to assess the effects of smoking on cause of death. It was a case-control study which looked at the past smoking habits of all Chinese adults in Hong Kong who died in 1998. They compare 27,507 deceased cases to 13,054 live controls. The results showed that smoking was correlated to an increased risk in mortality from TB. In men between the ages of 35 and 69, the relative risk of death from TB of smokers compared to non-smokers was 2.54 (95% confidence interval: 1.24,5.22). Relative risk is the decrease risk of a given activity or treatment in relation to a control activity, divided by the control event rate. In other words, TB smokers are 2.54 times more likely to die from TB than non-smokers. This risk was smaller for men over the age of 70: 1.63 (95% confidence interval: 1.01, 2.64). This study also showed a dose response relationship—the more cigarettes smoked per day, the more likely you were to die from TB (Lam et al. 2001).

A study done in Estonia showed the association between development of active TB and smoking. It was a case-control study which compared 248 TB patients that were treated in a hospital in Tallinn between January 1999 and June 2000 to 248 people selected from the Population Registry. The study showed that the adjusted odds ratio of diagnosis with active pulmonary TB for current smokers compared to non-smokers was 4.62 (95% confidence interval: 2.44, 8.73), when controlling for place of birth, marital status, and education. For former smokers, this adjusted ratio was 2.3 (95% confidence interval: 1.3,4.2). There was also shown to be an increased risk of developing disease when exposed to second hand smoke. The odds ratio of those exposed to passive smoking compared to those unexposed was 2.31 (95% confidence interval: 1.3,4.2) (Tekkel et al. 2002).

The possible effects of smoking on characteristics of TB were explored by a study conducted in Cataluña, Spain. This was a cross-sectional study that compared 13,038 TB patients, smokers and non-smokers. Patients were those registered in the Tuberculosis Control Programme in Cataluña, Spain between 1 January 1996 and 31 December 2002. Regardless of age, TB patients who smoked were 80% more likely to require hospitalization. The mean duration of hospitalization for these patients was estimated to be 9.4 days longer than for non-smokers. Overall, 34.1% of smokers had cavitary lesions, or "holes" in the lungs, compared to 19.2% of non-smokers. Also, 56.2% of smoking TB patients were smear positive compared with 39.3% of non-smokers (Altet-Gõmez et al. 2005).

A study conducted in Malaysia further describes the association between smoking and TB. This study was conducted in the State of Penang in Malaysia among newly diagnosed TB patients. It was estimated that 40.27% of patients were smokers, and 13.9% were ex-smokers. This study was especially interesting because it also investigated tobacco related knowledge, attitudes and behaviors of TB patients who are smokers. These are variables in research that have not previously been explored. It showed that 78.5% of participants in the study reported that they were in support of the ongoing Malaysian government's campaigns against tobacco use. Fewer than half of the participants that smoked had ever attempted to quit smoking (41.2%). Also, only 47.5% of the study participants had knowledge about the body system on which cigarette smoking has the greatest negative effect. Of the smokers in the study, 53.8% said that smokeless tobacco is a safe, harmless product. Additionally, 60% felt that smokeless tobacco can increase athletic performance (Awaisu et al. 2010).

CHAPTER 3

STUDY DESIGN

3.1 Study Objectives

The spread and containment of TB is aggravated by the factors outlined in chapter two: HIV, poverty, public health infrastructure represented by treatment outcomes, and smoking. While these factors have been researched, the aim of this study is to combine all of these suspected risk factors in a generalized linear model to predict the number of TB incidence cases in a country. The main goal of this study is to examine the relationship of smoking and tuberculosis above and beyond known risk factors of the disease. Many studies have shown the effects of each of these factors individually, while few combine these factors and assesses potential interactions.

This analysis will investigate hypotheses regarding smoking as a factor that is correlated with the incidence of tuberculosis. The number of incident cases of tuberculosis is the number of identified, new cases of tuberculosis reported. This work will lead to future studies to enhance tuberculosis control efforts by investigating the correlation of smoking, TB-HIV coinfection, poverty, and treatment outcomes on a country's number of incident tuberculosis cases.

3.2 Data

For this analysis, data were collected from multiple countries. Smoking percents, TB-HIV confection proportions, gross national income (GNI), and treatment outcomes were all recorded for each country. While this information was not located in one data set, a combination of three data sets yielded all the necessary data for analysis: 2011 WHO Global

Tuberculosis Control report, WHO: 2009 Report on the Global Tobacco Epidemic, and GNI (Atlas method) from the World Bank published in April 2008.

All of the data were analyzed for the year 2006. While more recent treatment outcomes and overall country statistics exist, the most recent smoking percentages per country were recorded in 2006. The WHO report *Global Tuberculosis Control* from 2011 will be used to obtain the estimated number of incident cases of TB, TB-HIV cases, treatment outcomes, and population for each country.

Estimated number of incidence cases of TB was used as the response variable. Since this value can only be expressed in non-negative integer values, Poisson and negative-binomial regression were used. Using these methods also required an offset term, i.e. estimated population in each country. These models and offset term will be discussed later in Chapter 4.

For this analysis, treatment outcomes for smear positive as well as re-treatment cases were considered. A re-treatment case is a patient who has already been treated for TB once before. This patient either defaulted (failed to complete treatment) or the treatment was unsuccessful. In order to have treatment effects be comparable across countries, they were represented by proportions. For example, the number of new smear-positive cases that were cured was divided by the cohort size of new smear-positive cases of TB for that country. The treatment outcomes are represented as the proportion of successes/failures among new smear-positive pulmonary TB cases. The same was done to represent TB-HIV coinfection. The number of HIV-positive TB cases was divided by the number of incident cases of TB. Therefore, HIV is represented in terms of the proportion of incident TB cases that are HIV-positive. By using proportions, the effect of these outcomes will be easier to interpret and generalize to the entire population.

In order to address the questions concerning smoking and TB, overall smoking percents were taken from each country. These were gathered from the *WHO: Report on the Global Tobacco Epidemic* (2009). This report gave adult daily smoking prevalence: age-standardized prevalence rates for adult daily smokers of tobacco, weighted by sex in the year 2006.

In order to account for the significance of income in predicting incidence of TB, it was necessary to obtain information on income for each country. The World Development Indicators database by the World Bank published on 11 April 2008 recorded GNI for each country in the year 2006. While some countries did not have data available for 2006, the majority of countries used in the analysis did. For countries that did not have data from 2006, estimates from 2004 or 2005 GNI were used (World Bank 2011). The values of GNI were recorded in millions and showed the entire country's income. GNI per capita (in dollars) was obtained by multiplying the given value (GNI in millions) by one million and dividing by the country's population. GNI per capita was used in this analysis because it gives a more accurate representation of a country's wealth, because it has been adjusted for population.

3.3 Preliminary Data Analysis

As expected, there is a lot of variation in incidence cases of TB per country. For example, the United States has very few cases of TB. This results in few reported treatment outcomes. The original WHO report on TB had statistics for 213 countries. Only 85 countries had complete information for all of the variables of interest. These missing values will be discussed later in Section 5.2.

Colinearity

In this section, a preliminary analysis of the data was performed. In the WHO TB report, there were five different treatment outcomes for both smear-positive (sp.) and re-treatment

(ret.) cases. These outcomes are as follows: cured, completed, failed, defaulted and died. The enumerated list below shows the definitions for these different treatment outcomes.

1. Cured: Initially smear-positive patient who was smear-negative in the last month of treatment and on at least one previous occasion
2. Failed: Smear-positive patient who remained smear-positive at month 5 or later during treatment
3. Completed: Patient who completed treatment but did not meet the criteria for cured or failed
4. Died: Patient who died from any cause during treatment
5. Defaulted: Patient whose treatment was interrupted for 2 consecutive months or more

It is expected that each country has consistent treatment outcomes whether or not they are treating smear-positive or re-treatement cases. It was also expected that treatment outcomes within a given country are correlated, for this reason only a few of these factors were included in the model. Table 3.1 shows correlations of interest among treatment outcomes.

Table 3.1: Correlations of treatment outcomes.

	sp.cured	sp.cmplt	sp.died	sp.fail	ret.cured	ret.cmplt	ret.died	ret.fail
sp.cured	1.00	-0.77	-0.28	0.02	0.73	-0.52	0.01	0.09
sp.cmplt	-0.77	1.00	0.02	-0.24	-0.54	0.68	-0.14	-0.29
sp.died	-0.28	0.02	1.00	0.21	-0.22	0.11	0.46	0.12
sp.fail	0.02	-0.24	0.21	1.00	-0.24	-0.04	0.28	0.89
ret.cured	0.73	-0.54	-0.22	-0.24	1.00	-0.70	-0.13	-0.13
ret.cmplt	-0.52	0.68	0.11	-0.04	-0.70	1.00	-0.19	-0.13
ret.died	0.01	-0.14	0.46	0.28	-0.13	-0.19	1.00	0.15
ret.fail	0.09	-0.29	0.12	0.89	-0.13	-0.13	0.15	1.00

Any two factors that have a correlation of 0.5 or greater would have caused problems with colinearity in a model. Therefore, only one of each of these pairs was included in the model.

Since there is more interest in active new cases of TB, when given the choice smear-positive outcomes were included instead of the re-treatment cases. By eliminating factors that have high colinearity, the factors of smear-positive cured, died, failed, and re-treatment died were left in the model. Table 3.2 below shows 6 observations of interest from the combined data sets. India and China have the highest TB incidence, according to this data set. In contrast, Jamaica and Mauritius have the lowest TB incidence. Russia and Ukraine have been included in table 3.2 because of their high smoking prevalence and high rate of drug-resistant tuberculosis.

Table 3.2: Response and explanatory variables that were used for analysis.

Country	Population	TB Inc.	TB-HIV	Smoke (%)	GNI per capita
1. India	1,157,038,539	2,400,000	0.0542	15	785.75
2. China	1,314,581,402	1,200,000	0.0142	31	1,993.75
3. Jamaica	2,696,334	180	0.2778	13	3,524.79
4. Mauritius	1,266,684	290	0.0448	14	5377.82
5. Russia	143,510,059	150,000	0.0226	44	5,730.107
6. Ukraine	46,591,797	47,000	0.0915	39	1,947.75

	sp.cured	sp.cmplt	sp.died	sp.fail	ret.cured	ret.cmplt	ret.died	ret.fail
1.	0.84	0.02	0.05	0.02	0.46	0.26	0.07	0.04
2.	0.93	0.02	0.02	0.01	0.85	0.05	0.02	0.02
3.	0.08	0.33	0.18	0	0	0.8	0	0
4.	0.46	0.46	0.03	0	0.57	0.423	0	0
5.	0.56	0.03	0.13	0.15	0.20	0.27	0.11	0.22
6.	0.54	0.05	0.12	0.12	0.27	0.16	0.14	0.20

Preliminary Plots

In order to illustrate the relationship between factors, scatter plots of the factors against the log of incidence cases of TB were plotted. Figure 3.1 shows the known risk factors of TB, GNI per capita and TB-HIV coinfection proportion against the log of the number of

TB incidence cases. Figure 3.2, shows the affect of death for any reason among TB patients before completion of treatment (smear-positive and re-treatment). Figure 3.3 shows how smear-positive cure and failure rates affect the log of TB incidence for each country. Lastly, Figure 3.4 explores the effect that smoking percent has on incidence. As seen from these figures, TB has a strong relationship with GNI per capita and TB-HIV coinfection. The proportion of re-treatment and smear positive cases that die before completing treatment appears to affect the number of incident cases of TB. It appears that the number of incident TB cases decreases as more people die before completing treatment. Unfortunately, the proportion cured and failed for smear-positive cases appear to have little correlation with incident cases of TB. This is also the case for smoking. Smoking by itself appears to be uninformative for predicting the incidence number of TB cases.

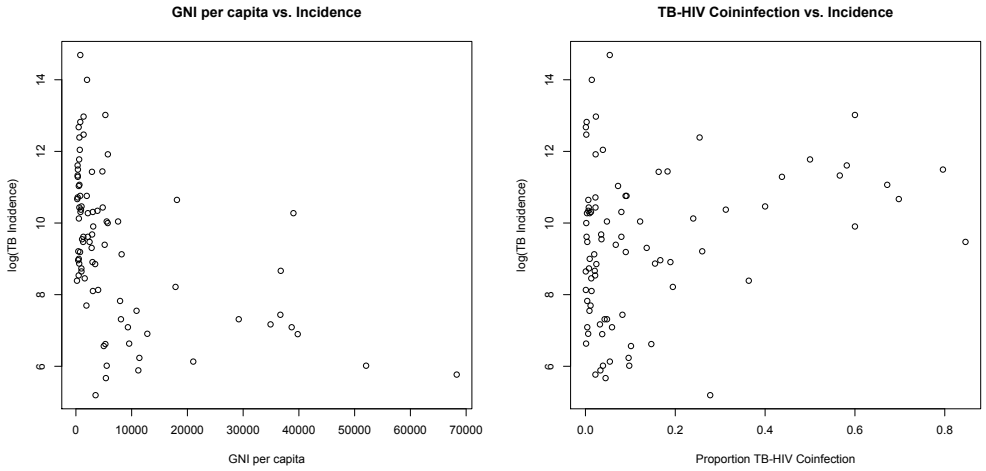

Figure 3.1: Left panel: log of TB incidence plotted against GNI per capita for each country. Right panel: log of TB incidence plotted against proportion of TB-HIV confection among incidence cases for each country.

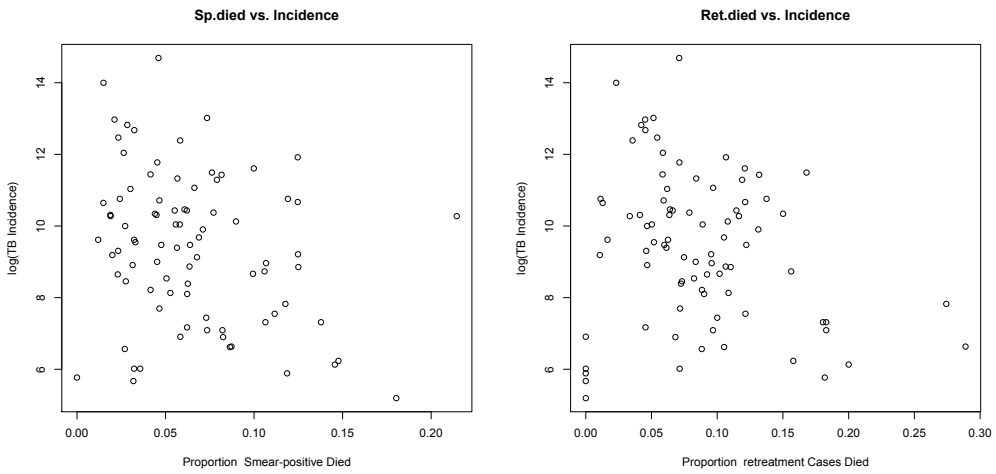

Figure 3.2: Left panel: log of TB incidence plotted against the proportion of smear-positive cases that died before completion of treatment for each country. Right panel: log of TB incidence plotted against proportion of re-treatment cases that died before completion of treatment for each country.

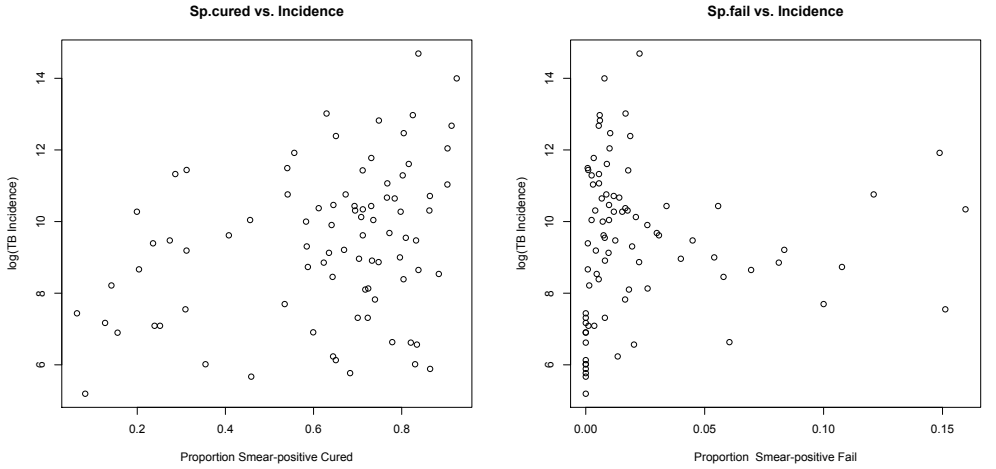

Figure 3.3: Left panel: log of TB incidence plotted against the proportion of smear-positive cases cured for each country. Right panel: log of TB incidence plotted against proportion of smear-positive cases failed for each country.

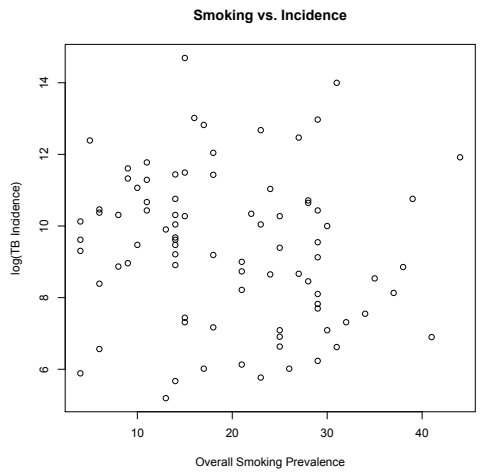

Figure 3.4: The log of TB incidence for each country plotted against smoking prevalence.

Summary Statistics

The summary statistics of the data illustrate the spread and distribution of the data. Table 3.3 shows summary statistics for the known risk factors of TB and smoking percentages. Table 3.4 shows the summary statistics of the treatment outcomes previously outlined.

Table 3.3: Summary statistics for number of incident cases of TB, proportion HIV+ among incident TB cases (TB-HIV), GNI per capita, and smoking prevalence.

	TB Incidence	TB-HIV	GNI per capita	Smoking (%)
Min	180	0.0007	191.1	4
1st Qu	2,500	0.0123	764.2	13
Median	13,000	0.0448	2879.7	18
Mean	90,151	0.1384	7772.8	19.78
3rd Qu	45,000	0.1630	7564.7	28
Max	2,400,000	0.8461	68308.5	44

These summary statistics illustrate that there is a large range in the number of incident cases of TB between countries. This indicates overdispersion (excess variation) for the Poisson

Table 3.4: Summary statistics for treatment outcomes.

	sp.cured	sp.cmplt	sp.died	sp.fail	ret.cured	ret.cmplt	ret.died	ret.fail
Min	0.06	0	0	0	0	0	0	0
1st Qu	0.54	0.04	0.03	0.003	0.23	0.07	0.05	0.01
Median	0.70	0.08	0.06	0.01	0.46	0.17	0.08	0.03
Mean	0.63	0.13	0.07	0.02	0.44	0.22	0.09	0.04
3rd Qu	0.80	0.14	0.08	0.02	0.63	0.31	0.11	0.06
Max	0.93	0.73	0.21	0.16	1	0.80	0.29	0.22

model. Quasi-Poisson and negative binomial are two modeling approaches that were used to address overdispersion. Theses models are discussed in further detail in Chapter 4.

3.4 Benefits

This analysis provided interesting results that could be used as future hypotheses. These results could provide a starting point for others interested in TB control. The results from future analysis and studies that have been based on these hypotheses may potentially lay the groundwork for public health programs to reduce the incidence of tuberculosis.

CHAPTER 4

MODELS

4.1 Overview

Modeling count response data has been a popular topic in the subject of ecology and medicine (Hoef and Boveng (2007), Lindén and Mäntyniemi (2011), Helgason et al. (2004)). A count response consists of any discrete response of counts (0,1,2 ...) (Hilbe 2007, p. 8). Often, data on organisms come in the form of counts, and it is desired to relate these counts to environmental conditions (Hoef and Boveng 2007, p. 2766). In the context of this project, counts of organisms are similar to incident counts of tuberculosis. In this project, it is desired to relate the conditions of the "environment," or country, to the count of new tuberculosis cases. Linear regression has been commonly used to model the relationship, however, it is not the most appropriate modeling method for count data. Therefore, there has been increasing interest in regression models that use Poisson or negative binomial distributions. The method of using linear models on non-normally distributed data is referred to as *generalized linear modeling* (GLM). In the next sections, Poisson and negative binomial regression will be discussed. In addition, the problem of "overdispersion" will be defined and addressed.

4.2 Poisson Regression

The standard or basic method used to model count response data is Poisson regression (Hilbe 2007, p. 1). A variety of other count models are based upon this basic count model. The Poisson distribution has the following characteristics:

$$f_Y(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} \qquad y = 0, 1, 2, 3, \ldots; \qquad \mu > 0. \tag{4.1}$$

The random variable Y in the above equation is the count response, and the parameter μ is the mean. Often, this parameter is referred to as the rate or intensity parameter. The Poisson distribution does not have a scale (variance) parameter, instead the scale is assumed to equal the location parameter (μ) (Hilbe 2007, p. 8).

A canonical link allows data which follow a Poisson distribution to be modeled by regression. While there are multiple link functions, the most common link used for Poisson regression is the natural log, simply referred to as the "log link:"

$$ln(\mu) = \eta = \beta_0 + \beta_1 x_1 \dots \beta_k x_k. \tag{4.2}$$

Or in matrix form this can be written:

$$ln(\mu) = \eta = X\beta. \tag{4.3}$$

After the Poisson regression model has been fit, it is important to interpret the coefficients. In this model, the coefficients show the effect of each factor on the log of the random variable. The coefficients also show the effect each factor has on the log of the count response, for each unit increase of the factor. For interpretation, the coefficients in the model may also be untransformed by taking the exponential of $X\beta$, as shown below:

$$\mu = e^{\eta} = e^{X\beta} \tag{4.4}$$

$$\mu = e^{\eta} = exp(\beta_0 + \beta_1 x_1 \dots \beta_k x_k) \tag{4.5}$$

$$\mu = e^{\eta} = e^{\beta_0} e^{\beta_1 x_1} e^{\cdots} e^{\beta_k x_k}. \tag{4.6}$$

Poisson models are generally used in either interpretation of exponentiated estimated slopes (Equation 4.6) or to summarize predicted counts based on a set of explanatory variables. The exponentiated estimated slopes indicate the expected change or difference in the incidence rate ratio of the outcome based on the changes in one or more explanatory predictors (Hilbe 2007, p. 43).

For evaluating the effect of smoking on tuberculosis incidence, the additive model is used below for illustrative purposes. Below is the equation of the additive model for tuberculosis incidence that includes all of the variables of interest. Note that x_1 = GNI per capita, x_2= TB-HIV Coinfection, x_3 = Sp. cured, x_4 =Sp. fail, x_5 = Sp. died, x_6 = Ret. died:

$$ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6. \tag{4.7}$$

In this equation, μ is the expected number of tuberculosis incidence cases in a country. However, when modeling TB incidence, it was necessary that the population of a given country be taken into account. Countries with larger populations will naturally have more tuberculosis incidence. In order to account for population, an offset term was used. An offset is a constant term that is added in each regression. For example, instead of modeling just the number off incidence cases of tuberculosis, the number of cases relative to population (pop) will be modeled:

$$\ln\left(\frac{\mu}{\text{pop}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$\ln(\mu) - \ln(\text{pop}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \ln(\text{pop}). \tag{4.8}$$

As seen in the algebra of the equations above, the modeling of incidence cases relative to population is the equivalent of adding a constant term (offset) to the regression model: ln(pop). The population of each country was known and was added to the model in each regression as the natural log of population.

4.3 Overdispersion

As stated earlier, the Poisson has the unique characteristic that the distribution's mean is equal to its variance, this relationship is termed *equidispersion*. Therefore, modeling using the Poisson distribution assumes the equality of the data's mean and variance–a property that is rarely found in real data. Generally the data tend to be *Poisson overdispersed*, the

variance of the distribution is greater than the mean. More commonly, this situation is termed *overdispersion* or the data is referred to as *overdispersed*. The dispersion parameter is often given the symbol ϕ. When $\phi = 1$ the condition of equidispersion is met. In contrast, $\phi > 1$ indicates overdispersion and $\phi < 1$ indicates underdispersion, a far less common problem in actual data.

Violations of equidispersion indicate potential correlation in the data, which can affect the standard errors of parameter estimates. There are other models that deal with data that violate the inherent assumptions of the Poisson model. Two of the most common model alternatives are discussed in Section 4.5 and 4.6

4.4 Goodness-of-fit Statistics

Fitting a model may be regarded as a way of representing a set of observed values, y, with a set of fitted values, $\hat{\mu}$, that were derived from the model. Generally, the observed y's will not equal the $\hat{\mu}$'s, which will result in a discrepancy. While a small discrepancy is acceptable, a large one is not. Measures of discrepancy are referred to as "goodness of fit." There are many ways to evaluate goodness of fit; however, in this section only the two most popular goodness of fit statistics are discussed.

One of the most common goodness of fit indicators is the Pearson X^2 fit statistic. It is defined as the following:

$$\text{Pearson } X^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}, \tag{4.9}$$

where $V(\mu_i)$ is the variance function for the distribution concerned. The Pearson X^2 statistic is the sum of all model Pearson residuals (Hilbe 2007, p. 73). A scaled Pearson X^2 statistic is the Pearson X^2 statistic divided by the overdispersion parameter (ϕ) for a given model. The Pearson X^2 statistic catches the excess variability. A Pearson $X^2 > 1$ indicates overdispersion, $X^2 < 1$ indicates underdispersion, and $X^2 \approx 1$ indicates that the condition

of equidispersion is met for Poisson regression. A scaled Pearson X^2 statistic indicates good or poor model fit and whether or not the correct distribution assumptions are met (Poisson, etc).

In order to explain the the next fit statistic, it is necessary to review some parameterizations of the generalized linear model. For generalized linear models, the response variable must follow the form of an exponential family. Each component in Y has a distribution in the exponential family, following the form:

$$f(y_i; \theta_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)\right). \tag{4.10}$$

Both $a(\phi)$ and $b(\theta)$ will be used later in this section. The function $a(\phi)$ is often in the form

$$a(\phi) = \phi/w. \tag{4.11}$$

In this form, ϕ is referred to as the *dispersion parameter* and w is a known *prior weight* that varies from observation to observation (McCullagh and Nelder 1989, p. 28-29). If no weight is specified, $w_i = 1$ for all observations. No weights were specified for this analysis, meaning $w_i = 1$ for every observation (SAS Institute Inc 2008, p. 1402).

The next fit statistic is the logarithm of a ratio of likelihoods, called *deviance*. Given n observations, a model with up to n parameters can be fit. The simplest, *null*, model has only one parameter which is a common expected value for all observations. In this model, all of the variation between observations is attributed to randomness. At the other extreme is the *full model*. This model contains n parameters, one for each observation, and therefore the derived fitted values match the data exactly.

In real situations, the full model represents an over-fit model and the null model is too simple. However, the full model gives a baseline for measuring the discrepancy of intermediate models that have p parameters ($p < n$). The discrepancy of a fit is proportional to

two times the difference between the maximum log likelihood achievable and that achieved by the model being fit. In the following equation for discrepancy, $\hat{\theta} = \theta(\hat{\mu})$ and $\tilde{\theta} = \theta(y)$ are the estimates of the canonical parameters under the two models. Assuming $a_i(\phi) = \phi/w$ the discrepancy can be written:

$$\sum 2w_i y_i(\tilde{\theta_i} - \hat{\theta_i}) - b(\tilde{\theta_i}) + b(\hat{\theta_i})/\phi = D(y; \hat{\mu})/\phi. \tag{4.12}$$

D(y; $\hat{\mu}$) is the *deviance* of the current model. Lower positive values of the deviance statistic indicate a better fitted model (Hilbe 2007, p. 41). The scaled deviance is defined as follows:

$$D^*(y; \hat{\mu}) = D(y; \hat{\mu})/\phi. \tag{4.13}$$

Consistent with the scaled Pearson X^2 statistic, scaled deviance around one indicates good model fit. The above equation for deviance is the general form for any exponential family class. More specifically, the deviance for Poisson distributed data is:

$$2\sum_i \left(y * \ln\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right) \tag{4.14}$$

Both the Pearson X^2 and the deviance have exact X^2 null distributions. The deviance statistic is useful in the fact that it is additive for nested sets of models if maximum likelihood estimates are used. The same is not true of the Pearson X^2 statistic (McCullagh and Nelder 1989, p. 34). Since the deviance is additive for nested models, it is possible to perform a test to compare full and reduced nested models. This is accomplished by taking the difference in deviance (D) for the two models and then comparing that difference to a X^2 with degrees of freedom equal to the difference in degrees of freedom (df) for the two models:

$$D_{full} - D_{reduced} \sim X^2(df_{reduced} - df_{full}). \tag{4.15}$$

The null hypothesis of this test is that the reduced model is as good at describing the data as the full model.

While the deviance has superior additive properties, the Pearson X^2 is direct in its interpretation. Simulation studies have indicated that Pearson dispersion better captures the excess variability and adjusts standard errors in a way that reflects what the standard errors would be if the excess variability were not present in the data (Hilbe 2007, p. 73-74). The scaled Pearson X^2 and deviance statistics are often used to assess the excess variability in Poisson models.

4.5 Quasi-Poisson

When the scaled Pearson X^2 is greater than one the model does not have adequate fit. A possible reason for the lack of fit is that the equivariance assumption is violated. The overdipsersion can be corrected by introducing a dispersion parameter (ϕ) into the relationship between the variance and the mean:

$$Var(Y_i) = \phi\mu. \tag{4.16}$$

This method is based on the quasi-likelihood approach, and thereby earned the name of quasi-Poisson (QP). This method permits estimation of parameters and inference testing without full knowledge of the probability distribution of the data. When $\phi = 1$, the quasi-Poisson is the generalized Poisson, meaning that the equivariance condition is met. However, as stated earlier, when $\phi > 1$ the data is overdispersed. It is important to note that the quasi-Poisson does not fit a new likelihood, it simply gives a correction term for testing the parameter estimates under the Poisson model. If overdispersion is modest, this method produces appropriate inference. It has been suggested that the dispersion parameter ϕ be estimated as a ratio of the deviance or the Pearson X^2 to its associated degrees of freedom. As stated earlier, simulations have shown that the Pearson X^2 test statistic captures the variance of a distribution better. Therefore, for this analysis, the dispersion parameter was estimated using the scaled Pearson X^2 test statistic (Pedan 2001, p. 2).

While scaling the variance is a viable option, most count models require more sophisticated adjustments. A more sophisticated alternative to the quasi-Poisson is using the negative binomial distribution (Hilbe 2007, p. 10). The negative binomial is frequently used to model overdispersed count data, and will be discussed in the next section.

4.6 Negative Binomial Regression

Negative binomial regression (NB) can be considered another form of nonlinear regression, or part of the family of generalized linear models. The negative binomial model can be derived using the Poisson distribution, as a Poisson-gamma mixture model. It can be derived from the Poisson distribution when the mean parameter for the Poisson is not identical for all members of the population, but itself follows a gamma distribution. Considering its connections to contagion data, the explanation of this derivation follows. The contagion or mixture concept of the negative binomial originated with Eggenberger and Polya in 1923 (Hilbe 2007, p. 16).

It is assumed that the data, given some contagion distribution, follow a Poisson distribution. The contagion distribution itself follows a gamma distribution. Written out in distribution terms:

$$Y_i|\lambda_i \sim Poi(\lambda_i) \tag{4.17}$$

$$\lambda_i \sim Gamma(\alpha_i, \beta_i). \tag{4.18}$$

In other words, the distribution of Y is mixed with an underlying distribution that represents the contagious nature of the disease. In order to derive the distribution it is necessary to outline the parameterization that will be used for both the Poisson and the gamma distributions. Below are the probability functions for both distributions (Poisson and gamma

respectively):

$$f(y|\lambda) = \frac{e^{-y}\lambda^{y}}{y!} \tag{4.19}$$

$$f(\lambda|\alpha\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}e^{\frac{-\lambda}{\beta}}. \tag{4.20}$$

The joint distribution can then be written:

$$f(y,\lambda) = f(y|\lambda)f(\lambda) \tag{4.21}$$

$$= \frac{\lambda^{y+\alpha-1}e^{-\lambda-\frac{\lambda}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)y!}. \tag{4.22}$$

By integrating over λ, the distribution of Y was found. After some algebraic work, which can be found in Appendix A, the distribution of Y follows a negative binomial. The resultant mass function of Y and its distribution are shown below:

$$f(y) = \binom{y+\alpha-1}{y}\left(\frac{1}{\beta+1}\right)^{\alpha}\left(1-\frac{1}{\beta+1}\right)^{y} \tag{4.23}$$

$$\sim NegBin\left(r=\alpha, p=\frac{1}{\beta+1}\right). \tag{4.24}$$

Using the estimates of r and p for the negative binomial, estimates for the parameters of the gamma distribution can also be obtained. The algebra used to solve for these estimates can also be found in Appendix A. Using r and p from above, the distribution of λ_i is:

$$\lambda_i \sim Gamma\left(r, \frac{1-p_i}{p_i}\right). \tag{4.25}$$

It is important to note that the computing software used in this analysis, SAS, uses a slightly different parameterization than written above. For SAS, the parameters for the negative binomial are κ and μ. Using algebra, these parameters can be written in the form of the most common negative binomial parameterization, using r and p. The algebraic work can be seen in Appendix B. The parameters r and p are written below in terms of SAS parameters for the negative binomial:

$$r = \frac{1}{\kappa} \tag{4.26}$$

$$p = \frac{1}{1+\kappa\mu}. \tag{4.27}$$

The negative binomial distribution naturally accounts for overdispersion, because its variance is always greater than its mean. For this reason, the negative binomial has more flexibility in modeling the relationship between the expected value and the variance of Y than the restrictive Poisson model. The variance for Y using the negative binomial (SAS's parameterization) is:

$$Var(y) = \mu + \kappa\mu^2, \tag{4.28}$$

where κ is a distribution parameter that is equal to $\frac{1}{r}$ (4.26). It can be seen that as κ gets small, the negative binomial model approaches a Poisson model.

Using SAS's parameterization the canonical link takes the form $\ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right)$ (Hilbe 2007, p. 83). As with Poisson regression, this link allows for linear regression of the variables. The negative binomial will also use the ln(pop) as an offset term.

$$\ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \ln(\text{pop}) \tag{4.29}$$

CHAPTER 5

RESULTS

5.1 Overview

Preliminary models and analysis showed that the data for tuberculosis incidence is overdispersed (Section 4.3). As outlined in Chapter 4, quasi-Poisson and negative binomial regression were used to model tuberculosis incidence; this accounts for the excess variation within the data. Since treatment outcomes could potentially interact with one another, it is important to allow for this in the model. The other health/poverty (TB-HIV coinfection, Smoking and GNI per capita) indicators may also have potential interactions. While there are insufficient degrees of freedom to model all possible interactions of all variables of interest, many interactions are still estimable. For this reason, the full model will include all factors individually, as well as interactions of treatment outcomes and interactions of the other three factors. The full model is written out below; μ is the expected value of TB incidence:

$$\begin{aligned}
ln(\mu) = {} & \beta_0 + \beta_1 x_{\text{TB-HIV}} + \beta_2 x_{\text{GNI per capita}} + \beta_3 x_{\text{Smoking}} + \beta_4 (x_{\text{TB-HIV}} \times x_{\text{GNI per capita}}) \\
& + \beta_5 (x_{\text{GNI per capita}} \times x_{\text{Smoking}}) + \beta_6 (x_{\text{TB-HIV}} \times x_{\text{Smoking}}) \\
& + \beta_7 (x_{\text{TB-HIV}} \times x_{\text{GNI per capita}} \times x_{\text{Smoking}}) \\
& + \beta_8 x_{\text{SP cured}} + \beta_9 x_{\text{SP died}} + \beta_{10} x_{\text{SP fail}} + \beta_{11} x_{\text{R died}} \\
& + \beta_{12} (x_{\text{SP cured}} \times x_{\text{SP died}}) + \beta_{13} (x_{\text{SP cured}} \times x_{\text{SP fail}}) + \beta_{14} (x_{\text{SP cured}} \times x_{\text{R died}}) \\
& + \beta_{15} (x_{\text{SP died}} \times x_{\text{SP fail}}) + \beta_{16} (x_{\text{SP died}} \times x_{\text{R died}}) + \beta_{17} (x_{\text{SP fail}} \times x_{\text{R died}}) \\
& + \beta_{18} (x_{\text{SP cured}} \times x_{\text{SP died}} \times x_{\text{SP fail}}) + \beta_{19} (x_{\text{SP cured}} \times x_{\text{SP died}} \times x_{\text{R died}}) \\
& + \beta_{20} (x_{\text{SP died}} \times x_{\text{SP fail}} \times x_{\text{R died}}) \\
& + \beta_{21} (x_{\text{SP cured}} \times x_{\text{SP fail}} \times x_{\text{SP died}} \times x_{\text{R died}}) + ln(x_{\text{Population}}). \qquad (5.1)
\end{aligned}$$

Backwards selection was used to select an appropriate model for both the quasi-Poisson and negative binomial models. From the full model (5.1), coefficients with a p-value of 0.05 or greater were dropped from the model. Terms were dropped one-by-one, starting with the term with the highest p-value. This process continued until all terms remaining in the model were significant at the 0.05 level. If a significant interaction included a particular term, that term's main effect was also included in the model, whether or not the main effect was significant.

The full model is also useful for performing full vs. reduced model tests as outlined in Section 4.4 (4.15). The full model, mentioned above, will be compared via deviance to the model being investigated (4.15). In Section 5.2, the problem of missing data is addressed. The quasi-Poisson generalized linear model and negative-binomial regression model are discussed in Sections 5.3 and 5.4, respectively. Section 5.5 compares and contrasts the models fit in the previous two sections. Quasi-Poisson and negative binomial models using only smoking as a predictor is considered in Section 5.6. Lastly, Section 5.7 compares all the models used to describe TB incidence that are discussed in this chapter.

5.2 Missing Data

As mentioned in the preliminary data analysis in Chapter 3, there were initially 213 data points, however, all but 85 points had missing values in at least one of the variables of interest (response or explanatory). A t-test was performed to compare the mean number of TB incidence cases for the observations that had missing values, and the 85 observations that were retained. Assuming that the variance of TB incidence between the two groups is not equal, $t = -1.7721$ and has an associated p-value of 0.079. This indicates that the difference in average TB Incidence for the two groups (missing and retained) is insignificant. However, if it is assumed that the variances are equal, and therefore using the pooled sample variance when calculating the test statistic, $t = -2.0108$ with an associated p-value of 0.046. This

borderline p-value indicates that the mean of TB Incidence for the missing data is different from the 85 retained observations. Summary statistics for TB incidence for both groups are listed in Table 5.1. Among the missing observations are four countries that have no recorded TB Incidence: Aruba, Monaco, San Marino, and the U.S. Virgin Islands. Figure 5.1 shows the difference between omitted (due to missing values) and retained data points.

Table 5.1: Summary statistics of TB incidence for observations that contain missing values and observations that do not (complete).

TB Incidence Statistics	Missing Data	Complete
Minimum	0	180
1st Quartile	53	2,275
Median	570	12,500
Mean	20670	75210
3rd Quartile	9,000	42,750
Maximum	1,200,000	2,400,000

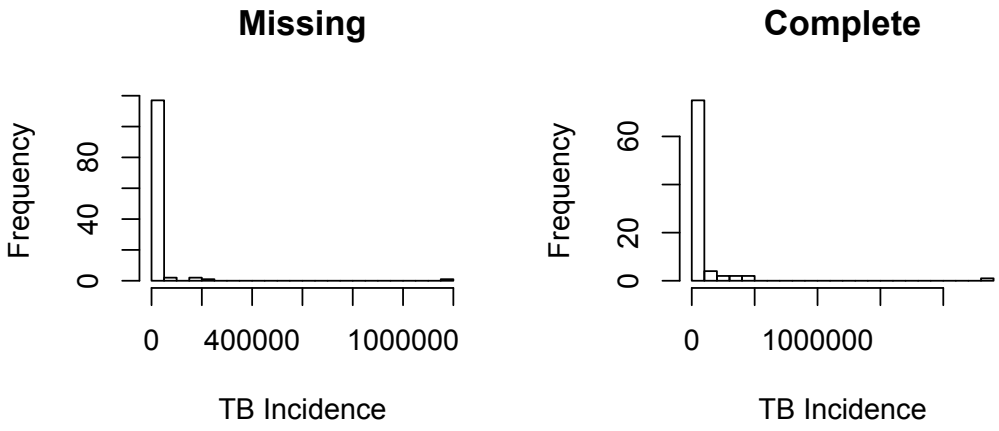

Figure 5.1: Distribution of data that contains missing values (left) and retained data (right).

Table 5.1 and Figure 5.1 show that observations with larger values of TB Incidence are being retained. However, observations with smaller values of TB incidence are more likely to have missing data and be omitted. This is expected. Countries that are highly burdened with tuberculosis also have WHO programs integrated into their health care. The WHO pays for collection and analysis of data concerning contagious diseases/epidemics like tuberculosis. This is one of the reasons why countries with higher TB incidence have observations lacking in missing values.

Approximately 32.2% of missing observations are only missing data for one variable; 15.7% are missing information for two variables. While 57.5% of the incomplete observations are missing values for less than or equal to three variables, 23.6% are missing data for 12 variables or more. There are nine countries missing data for 15 and 16 variables, 12 of which are used in calculating treatment outcomes. These nine countries are Aruba, Bermud, British Virgin Islands, Monaco, Montserrat, Netherlands Antilles, San Marino, Turks and Caicos Islands, and the US Virgin Islands. Note that the majority of the these nine countries are islands, and that they also have lower or no recorded levels of TB incidence.

Iceland, another country with missing values, only has estimated TB incidence of 13. However, Iceland is only missing data for the variables that are used to calculate re-treatment outcomes. With estimated TB incidence of 6,100, France is missing values for all the variables that are used to calculate treatment outcomes—smear-positive and re-treatment. These results suggest that islands/territories have the most incomplete observations and countries with relatively low TB incidence are missing values related to treatment outcomes. Countries that have negligible amounts of TB will be less likely to record treatment outcomes (smear-positive or re-treatment) because there are few observed TB cases, as in France and Iceland.

Of the observations that are missing data, the majority are missing values for smoking prevalence. About 10.7% of all missing values are smoking prevalence; 9.1% of all missing values are TB-HIV coinfection. Re-treatment outcomes also have more missing values than smear-positive treatment outcomes. Approximately 45.2% of all missing values are in variables connected to re-treatment and 29.4% are in smear-positive related fields.

India and China have the largest TB incidence recorded in the WHO database. Both of these countries have values for every observation of interest. This analysis is interested in modeling countries which have a tuberculosis burden. For the purposes of this analysis, it is not necessary to generalize results over islands/territories or countries that have small TB incidence and do not record treatment outcomes. For these reasons, it is reasonable to use the remaining 85 countries that have a TB problem.

5.3 Quasi-Poisson

The quasi-Poisson model was fit using the scaled Pearson X^2 statistic (4.9) as an estimate of ϕ as recommended in Section 4.5. An offset term of the natural log of population (4.8) was also included in the model. The resulting model, fit using backwards elimination from the full model (5.1), is :

$$\begin{aligned}\ln(\mu) = {} & \beta_0 + \beta_1 x_{\text{TB-HIV}} + \beta_2 x_{\text{GNI per capita}} + \beta_3 (x_{\text{TB-HIV}} \times x_{\text{GNI per capita}}) \\ & + \beta_4 x_{\text{SP cured}} + \beta_5 x_{\text{SP died}} + \beta_6 x_{\text{R died}} \\ & + \beta_7 (x_{\text{SP cured}} \times x_{\text{SP died}}) + \beta_8 (x_{\text{SP cured}} \times x_{\text{R died}}) + \beta_9 (x_{\text{SP died}} \times x_{\text{R died}}) \\ & + \beta_{10} (x_{\text{SP cured}} \times x_{\text{SP died}} \times x_{\text{R died}}) + \ln(x_{\text{Population}}). \end{aligned} \tag{5.2}$$

The fit statistics for this model are shown in Table 5.2. The parameter estimates from the fitted model are shown in Table 5.3. In Table 5.3 the scale parameter is equal to $\sqrt{\phi}$ (SAS Institute Inc 2008). Therefore, the dispersion parameter (ϕ) is estimated to be 18,210.56. The model fit through backwards selection used p-values when testing for the significance of

covariates. Parameter estimates from the quasi-Poisson model are equal to the parameter estimates from the unadjusted Poisson model if the given explanatory variables are the same. However, p-values for the quasi-Poisson are accurate because the corresponding standard errors have been adjusted for overdispersion.

Table 5.2: Model fit statistics for the quasi-Poisson model.

Criterion	DF	Value	Value/DF
Deviance	74	949032.2784	12824.7605
Scaled Deviance	74	52.1144	0.7042
Pearson Chi-Square	74	1347580.4594	18210.5467
Scaled Pearson χ^2	74	74	1
Log Likelihood		5058.5556	
AIC		950009.9603	
BIC		950036.8294	

Table 5.3: Parameter estimates for the quasi-Poisson model.

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	-3.4394	1.6578	0.0380
TB-HIV Coinfection	1.9799	0.4427	<.0001
GNI per capita	-1.29×10^{-4}	3.71×10^{-5}	0.0005
TB-HIV Coinfection*GNI per capita	5.64×10^{-4}	1.28×10^{-4}	<.0001
SP cured	-4.2762	1.8951	0.0240
SP died	-85.0734	41.6320	0.0410
Ret died	-64.1399	20.0538	0.0014
SP cured*SP died	112.3332	55.8977	0.0445
SP cured*Ret died	96.3281	24.0807	<.0001
SP died*Ret died	1117.544	344.1760	0.0012
SP cured*SP died*R died	-1629.01	438.1294	0.0002
Scale	134.9465		

The terms TB-HIV coinfection, GNI per capita, TB-HIV coinfection*GNI per capita, SP cured, SP died, Ret died, SP cured*SP died, SP cured*Ret died, SP died*Ret died, and SP cured*SP died*R died are included in the model as being significant predictors of TB incidence. A one percent increase in TB-HIV coinfection is expected to increase the log of

TB incidence by about 0.02. As GNI per capita increases the expected log of tuberculosis incidence goes down. Death among smear-positive and re-treamtment cases, as well as the proportion of smear-positive cases that are cured, appears to decrease the expected log of TB incidence for a country.

It is interesting to note from Table 5.3 above, that TB-HIV coinfection is not only a significant factor, but also *positively* correlated with TB incidence. These findings correlate with WHO conclusions. HIV is known to worsen symptoms of the disease in TB patients; TB-HIV coinfected cases do not survive long. As stated in the WHO report, there was a decline in TB prevalence in the 1990s, but a rise in TB incidence. This means that the number of new cases each year was increasing, but the number of already existent cases was decreasing. According to the WHO this contrast "can be explained by a decrease in the average duration of disease ..., combined with a comparatively short duration of disease among HIV-positive cases" (WHO 2009a, p. 12).

From the negative coefficients of treatment outcomes, it can be seen that incidence decreases when the number of active TB cases decreases. People with TB-HIV coinfection will not survive long, thereby decreasing the number of active TB cases. For this reason, it was thought that the coefficient for TB-HIV coinfection would be negative. The positive coefficient indicates that although TB disease is short-lived in coinfected patients, TB-HIV coinfection is responsible for some of the increase in TB incidence. The positive coefficient indicates that the greater the proportion of coinfected cases, the greater the TB incidence. From this, it appears that even though TB-HIV co-infected patients die quickly, they are still having a significant effect on the spread of tuberculosis; causing more incident cases within a country.

Initially it was seen in Section 3.3 that treatment outcomes are significantly correlated.

In Section 2.3, it was established that public health infrastructure is important to TB control. It is understandable that the treatment outcomes of SP cured, SP died, Ret died are significant factors, resulting in the decrease of TB incidence when proportions of these variables are high. However, it is surprising to see significant two and three way interactions of treatment outcomes. Since collinearity in treatment outcomes has already been addressed (Section 3.3), it would be assumed that the remaining treatment outcomes would be significant, but independent of each other. However, from the significant interactions it can be seen that the treatment outcomes are related. The results of this model suggest that improving all aspects of TB treatment may be necessary for decreasing TB incidence.

Figure 5.2 shows the residuals and Cook's distances for the fitted values given by the quasi-Poisson model and shows that there are some problems with model fit. The maximum Pearson residual was 493.0487, which was observed for Jamaica, the country with the smallest observed TB incidence. This, as well as other evidence to be discussed later (Section 5.5), shows that the quasi-Poisson doesn't model countries with low TB incidence well. Negative binomial regression will be fit next because of its built-in accommodation for larger variance (4.28).

5.4 Negative Binomial

Like the quasi-Poisson, an offset term of the natural log of population was used in the negative binomial model. Again, backwards selection was used from the full model (5.1) in order to obtain an equally sufficient, more parsimonious model. The resulting model is:

$$\begin{aligned}\ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right) &= \beta_0 + \beta_1 x_{\text{TB-HIV}} + \beta_2 x_{\text{GNI per capita}} \\ &\quad + \beta_3 x_{\text{SP cured}} + \beta_4 x_{\text{SP fail}} + \beta_5 x_{\text{SP died}} + \beta_6 x_{\text{R died}} \\ &\quad + \beta_7(x_{\text{SP cured}} \times x_{\text{SP died}}) + \beta_8(x_{\text{SP died}} \times x_{\text{R died}}) \\ &\quad + \beta_9(x_{\text{SP cured}} \times x_{\text{SP died}} \times x_{\text{R died}}) + \ln(x_{\text{Population}}). \qquad (5.3)\end{aligned}$$

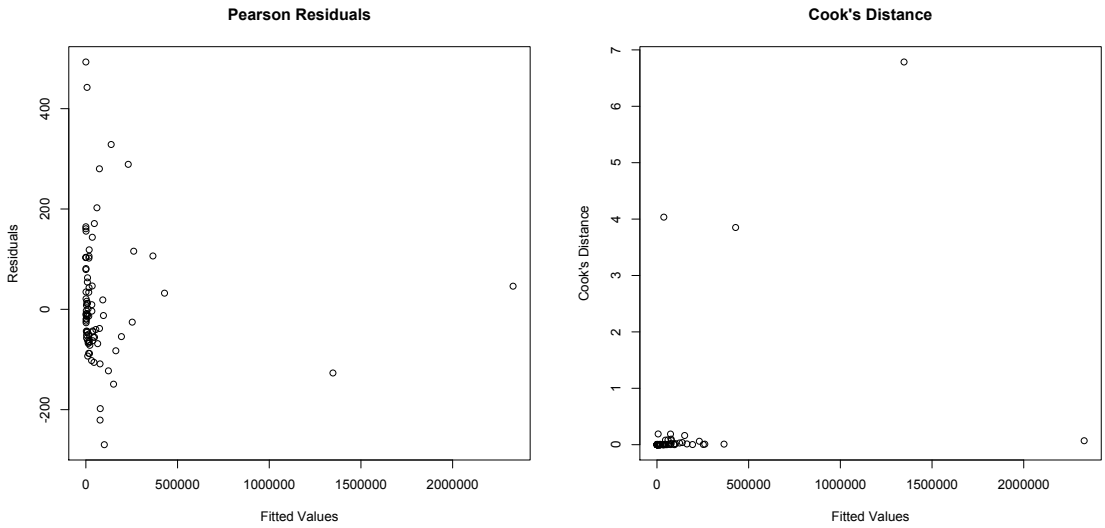

Figure 5.2: Left panel: Pearson residuals for the quasi-Poisson model. Right panel: Cook's distances for quasi-Poisson model.

Goodness of fit statistics for this model are in Table 5.4. As can be seen, the scaled deviance (4.13) and scaled Pearson X^2 (4.9) statistics are close to one, indicating a good model fit.

The parameter estimates are in Table 5.5 below. The deviance (4.15) of the full model fit

Table 5.4: Model fit statistics for negative binomial regression model.

Criterion	DF	Value	Value/DF
Deviance	74	90.4530	1.2060
Scaled Deviance	74	90.4530	1.2060
Pearson Chi-Square	74	73.1012	0.9747
Scaled Pearson χ^2	74	73.1012	0.9747
Log Likelihood		92593171.342	
AIC		1793.6016	
BIC		1820.4707	

using negative binomial (NB) regression was 90.4530 on 75 degrees of freedom. The comparison of deviance from that of the full (5.1) and reduced (5.3) model resulted in a p-value of 0.99, indicating that the reduced model is adequate, and is as informative as the full model. The significant covariates in this model are: TB-HIV coinfection, GNI per capita, SP cured, SP died, Ret died, SP fail, SP cured*Ret died, SP died*Ret died, and SP cured*SP died*R

died. Consistent with the quasi-Poisson model, it's expected that a decrease is seen in TB incidence as the proportion of smear-positive and retreatment cases that die increases. As TB-HIV coinfection increases, TB incidence also appears to increase. An increase in the proportion of smear-positive cases that fail treatment, seems to increase TB incidence.

Table 5.5: Parameter estimates for negative binomial regression model.

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	-6.2570	0.4796	<.0001
TB-HIV Coinfection	3.8222	0.3896	<.0001
GNI per capita	-4.03 $\times 10^{-5}$	7.23 $\times 10^{-6}$	<.0001
SP cured	0.3045	0.6574	0.6433
SP died	-23.5218	3.4926	<.0001
Ret died	-27.1544	7.5249	0.0003
SP fail	6.5643	.3329	0.0049
SP cured*Ret died	31.3349	11.4959	0.0064
SP died*Ret died	337.7698	64.7909	<.0001
SP cured*SP died*R died	-341.135	90.6177	0.0002
Dispersion	0.3920	0.0567	

This model is fairly consistent with the quasi-Poisson. Both models show the interaction of treatment outcomes to be highly significant. Again, this could indicate that it is necessary for countries to address all aspects of TB treatment if they desire to control TB incidence. Unlike the quasi-Poisson model, the negative binomial model suggests that the proportion of smear-positive cases that fail treatment is a significant factor in predicting TB incidence. Recall that SP fail describes smear-positive patients who remained smear-positive at month 5 or later during treatment (Section 3.1). In short, fail describes a patient who received treatment for active tuberculosis and was not "cured." The main effect of smear-positive cured is insignificant, but it is significant in the treatment interactions. For this model, treatment of smear-positive cases that are unsuccessful is more significant than the proportion of smear-positive cases that are cured. The interaction of TB-HIV coinfection and GNI per capita is also no longer significant.

Figure 5.3 shows the residuals and Cook's distances for the fitted values given by the negative binomial regression. The Cook's distances and residuals appear to be small and have constant variance. Comparing these diagnostics with the quasi-Poisson diagnostics, it appears that negative binomial regression is, overall, better suited to address overdispersion when modeling tuberculosis incidence. However, it is important to note that in the quasi-Poisson only standard errors were adjusted for overdispersion, not parameter estimates. This is a possible reason why the Pearson X^2 and Cook's distances for this model are large compared to the negative binomial model. The largest Pearson residual is 2.346 and occurred when modeling Cambodia. Cambodia has a relatively large TB incidence: 62,000. This suggests that the negative binomial is possibly not a good fit countries with a larger TB burden. The strengths and short comings of the negative binomial model and quasi-Poisson model is another interesting result that is discussed in the next section.

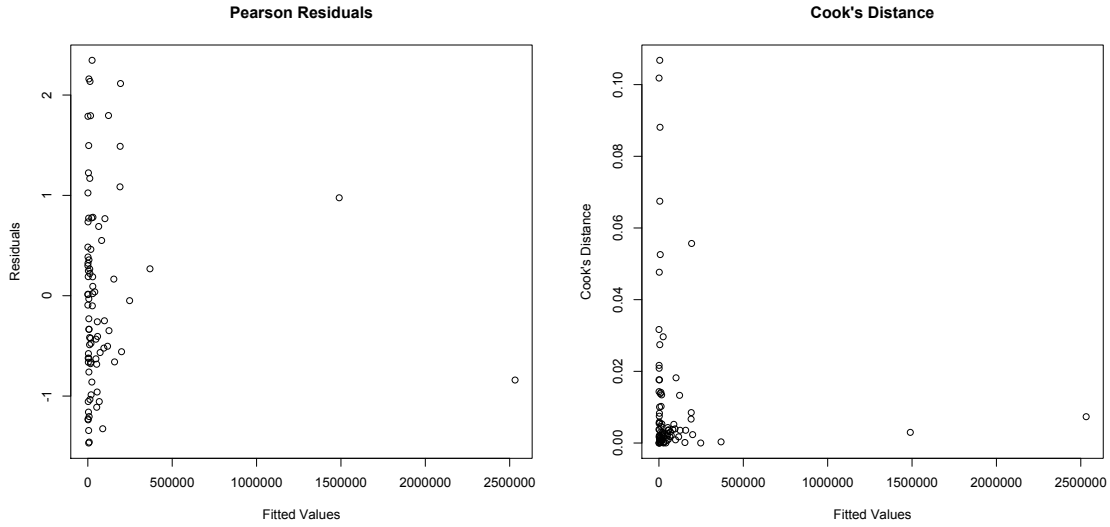

Figure 5.3: Left panel: Pearson residuals for the negative binomial regression model. Right panel: Cook's Distances for negative binomial regression model.

5.5 Comparison: Quasi-Poisson and Negative Binomial

Choosing the Best Model for TB Incidence

In this section the quasi-Poisson and negative binomial regression methods and results are compared. Figure 5.5 contains a plot of the fitted values and 95% confidence intervals for the quasi-Poisson and negative binomial models on separate graphs. The confidence intervals, fitted values, and actual TB incidence values in Figure 5.5 are all represented in the natural log of TB incidence. The confidence intervals and fitted values are easier to see and distinguish on the log scale. However, Figure 5.5 contains the actual values of TB incidence, fitted values, and confidence limits for both models. As can be seen there is significant interval overlap for both models. The majority of quasi-Poisson confidence intervals are wider than the negative binomial confidence intervals. For countries with TB incidence less than or equal to 7,800, 80% of the quasi-Poisson confidence intervals are larger than the negative binomial confidence intervals. For countries with higher TB incidence, greater than 7,800, only 68% of the quasi-Poisson confidence intervals are larger than the negative binomial confidence intervals. This indicates that negative binomial estimates become more uncertain as TB incidence increases.

As stated earlier, the different fits of the quasi-Poisson and negative binomial models yield a very interesting result. Recall that in the quasi-Poisson model the variance was a linear function of the mean (4.16). The variance of the quasi-Poisson indicates a constant overdispersion in the data. For negative binomial regression, the variance is quadratic in the mean. In the negative binomial case, the overdispersion (variance in excess of μ) is the multiplicative factor of $1 + \kappa\mu$, which depends on μ. Since the negative binomial model fits better overall than the quasi-Poisson, the data tend to have an overdispersion that increases with the mean (Hoef and Boveng 2007). This means that the excess variation in TB incidence increases as the number of incidence cases increases.

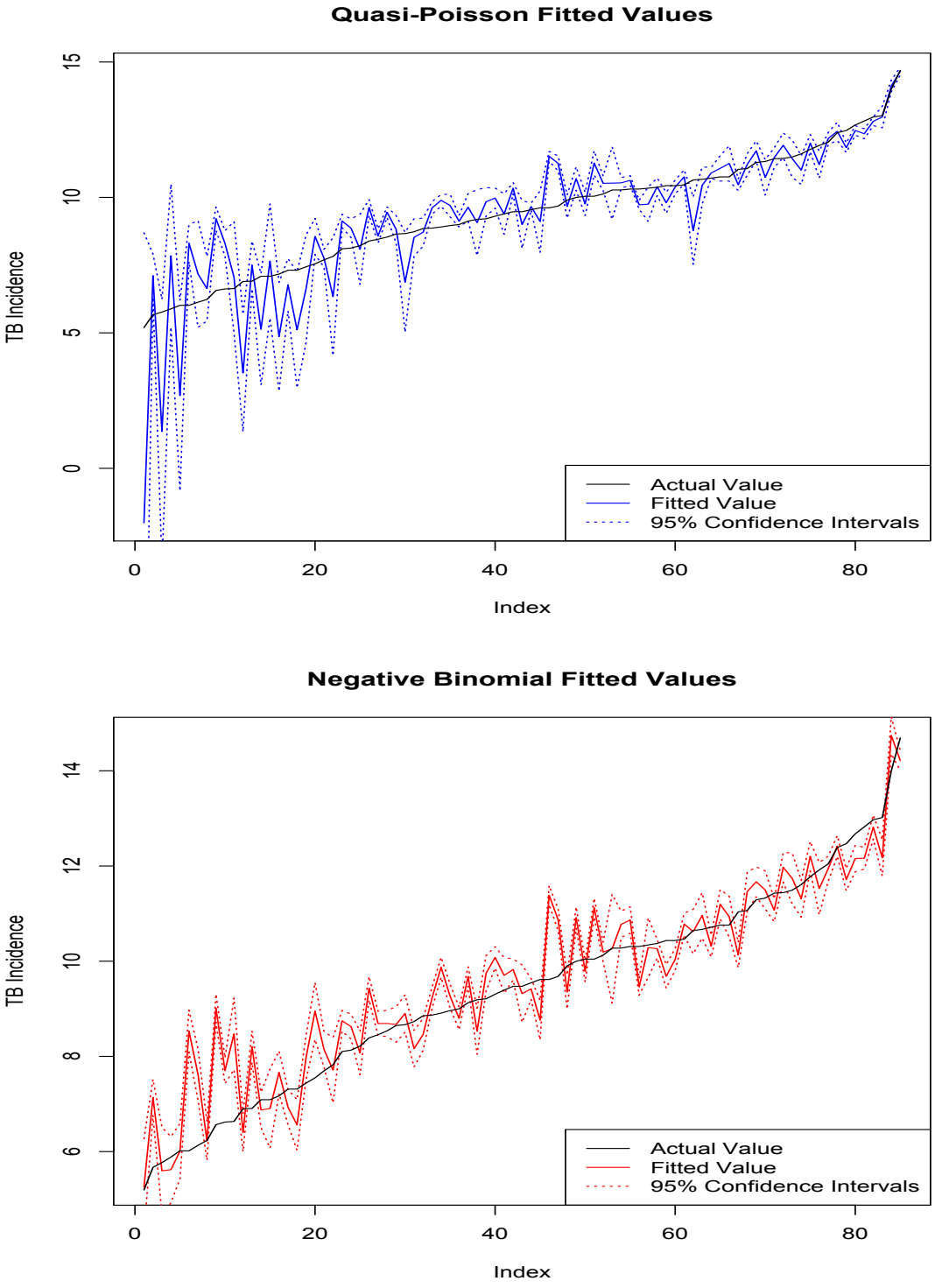

Figure 5.4: Top: Fitted values and 95% confidence intervals for the quasi-Poisson model. Bottom: Fitted values and 95% confidence intervals for the negative binomial model (all in the natural log of TB incidence).

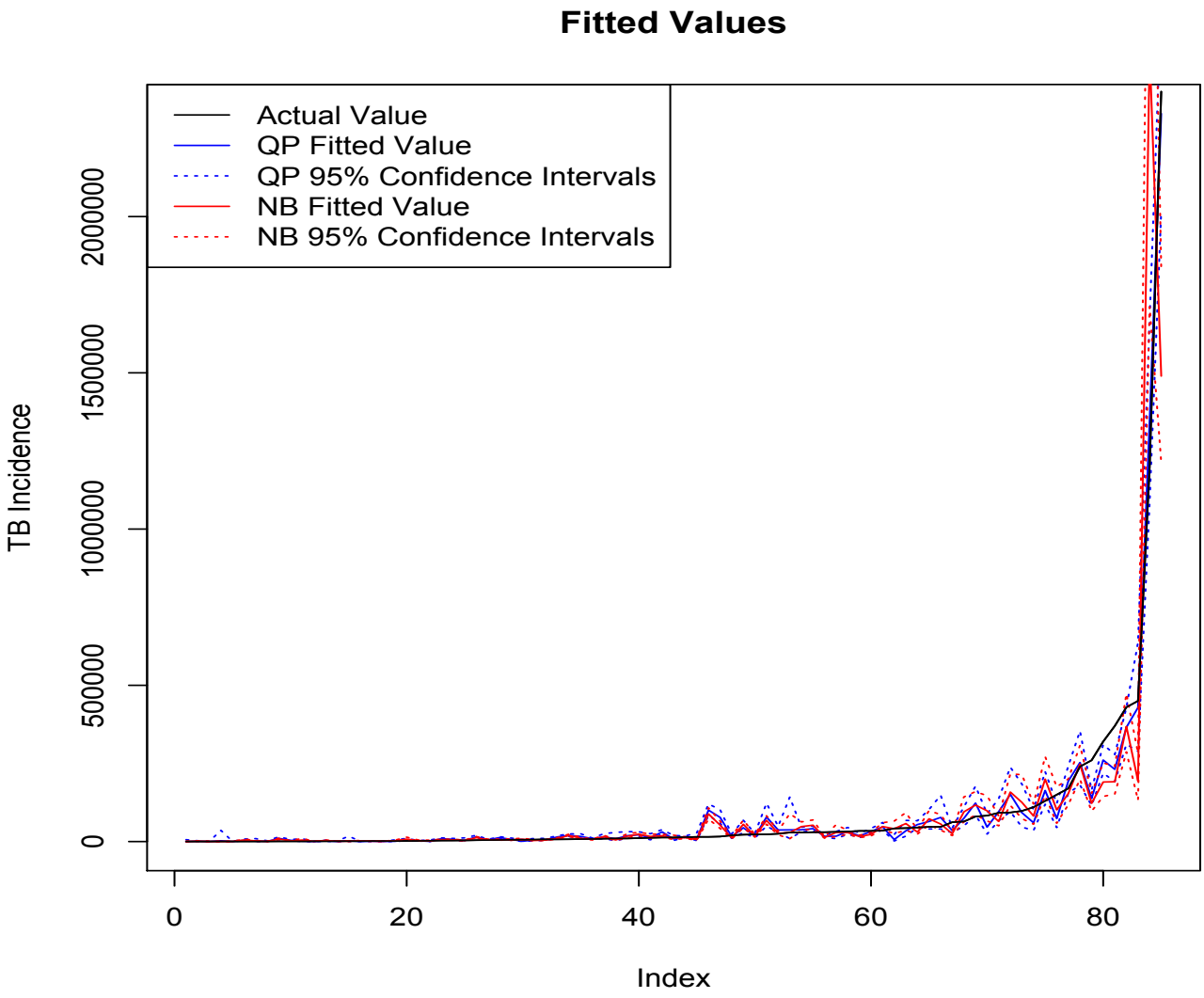

Figure 5.5: Fitted values and 95% confidence intervals for the quasi-Poisson model and the negative binomial regression model (actual values of TB incidence).

This property is implemented in the iterative process used to fit the quasi-Poisson and negative binomial models. These regression models were fit using iteratively reweighed least squares (IRLS). IRLS assigns different weights to observations as it performs an iterative fitting process. When fitting the quasi-Poisson, IRLS assigns weights (W) that are directly proportional to the mean:

$$W = diag\left(\frac{\mu_1}{\phi} \ldots \frac{\mu_n}{\phi}\right). \tag{5.4}$$

When fitting the negative binomial regression model, IRLS assigns weights that have a concave relationship to the mean. This means that observations with average values of the response variable receive more weight than extreme observations. The weights used for IRLS negative binomial regression are shown in the equation below:

$$W = diag\left(\frac{\mu_1}{1+\kappa\mu_1} \ldots \frac{\mu_n}{1+\kappa\mu_n}\right). \tag{5.5}$$

Because different weights are used when fitting quasi-Poisson models and negative binomial regression models, observations in the different models receive different weights. It is important to understand that the negative binomial regression gives countries with low and average TB incidence more weight, relative to quasi-Poisson. However, the quasi-Poisson model was driven by countries with a larger number of TB incidence cases (Hoef and Boveng 2007, p. 2772). The difference in model fits could be due to a ceiling effect in tuberculosis incidence. In other words, as TB incidence increases beyond a certain point the relationship of the variance and expected value of TB incidence becomes simplified. This could indicate that after a certain point, TB incidence becomes less complex, less variable, and better modeled by the quasi-Poisson. However, below this point, a more complex relationship of the mean and variance is needed; better modeled by negative binomial regression. It could also be that negative binomial regression better models countries with smaller TB incidence because the iterative fitting process gives smaller observations more weight. Whether or not this result is due to a ceiling effect or just because smaller observations are given more weight,

negative binomial regression better models countries with smaller TB incidence. Likewise, the quasi-Poisson model is better at modeling countries with higher TB incidence. Table 5.6 shows the percentage of time a given model's fitted values are closer to actual value of TB incidence than the other proposed model.

Table 5.6: The percentage of fitted values that were more accurate than the other proposed model.

Model	Overall	TB Incidence $\leq$ 7,800	TB Incidence $>$ 7,800
quasi-Poisson	49.4%	34.3%	60%
negative binomial	50.6%	65.7%	40%

Over all of the observations, the negative binomial fitted values are closer to the true values of TB incidence. However, the results show that the quasi-Poisson fits better for observations that have a higher TB incidence ($>$ 7,800). The negative binomial model fits best when modeling countries with lower TB incidence ($\leq$ 7,800). The cutoff between "low" and "high" TB incidence (7,800) was selected because this value caused a natural break in the accuracy of the negative binomial model. The quasi-Poisson model began to outperform the negative binomial when TB incidence was greater than 7,800. This result was seen earlier in the diagnostics for each model. Recall that the largest Pearson residual from the quasi-Poisson model was for the country with the smallest TB incidence (Jamaica); the largest Pearson residual for negative binomial regression model was a country with relatively large TB incidence (Cambodia). In this data set, 35 countries have TB incidence less than or equal to 7,800. The remaining 50 countries are classified as having high TB incidence ($>$ 7,800).

A few countries were highlighted in order to illuminate this difference. India and China have the largest and second largest TB incidence in the world, respectively. Table 5.7 shows TB incidence and fitted values from both models for India and China. As indicated by

the the raw residual (error), the quasi-Poisson provided more accurate predicted values for India and China. This is consistent with the majority of higher burdened TB countries (TB incidence $>$ 11,000). The opposite is seen when looking at the two countries with the lowest TB Incidence.

Table 5.7: Actual and fitted values for the two countries with highest TB incidence for both models.

	Country	TB Incidence	Fitted	Error
QP	India	2,400,000	2,329,584	70,416
NB			1,489,667	910,333
QP	China	1,200,000	1,347,267	-147,267
NB			2,531,901	-1,331,901

Jamaica and Norway are the lowest and third lowest in TB incidence, respectively. Table 5.8 gives the actual and fitted values for these two countries using the negative binomial and quasi-Poisson models. As expected, the model fit using negative binomial regression more adequately describes countries with lower TB incidence. This is consistent with the majority of lower burdened TB countries (TB incidence $\leq$ 7,800).

Table 5.8: Actual and fitted values for the countries with the lowest and third lowest values of TB incidence for both models.

	Country	TB Incidence	Fitted	Error
QP	Jamaica	180	0.13	179.87
NB			191	-11
QP	Norway	320	3.87	316.13
NB			269.23	50.77

From these results, whether due to weights or a ceiling effect, it is apparent that the quasi-Poisson model (Section 5.3) more adequately describes countries with higher levels of TB incidence. Likewise, the negative binomial model (Section 5.4) gives better predictions for countries with low TB incidence. The coefficients (Table 5.3 and Table 5.5) from these

different models can now be compared to examine the different factors that influence TB incidence in high and low burdened countries.

Interpreting the Results of the Best Fit Models

In the negative binomial model, it appears that countries with low TB incidence are possibly more effected by the amount of smear-positive cases that fail treatment than are cured (Table 5.5). When a treatment is unsuccessful, that individual is more likely to develop drug resistance (Section 1.1). Because of this association, the significance of smear-positive cases that fail treatment would indicate that, in low burdened countries, the problem at hand is the development of drug resistant tuberculosis. SP fail is a more significant predictor in low burden countries than the amount of SP cases that are cured, indicating that controlling the sheer number of TB cases is not as important for low burdened countries. A possible reason for this is because countries with low TB incidence have already kept the number of TB cases down. However, the new threat could be the development of drug resistant TB, which these countries are not as prepared to fight. For the United States, obviously a low burdened country, this is the aspect of TB that is most concerning. The CDC has put in significant effort to fight normal strains of TB, but has not developed significant strategies to combat the development of drug resistance. However, countries with low TB incidence might be able to minimize drug resistance by making treatments (smear-positive and re-treatment) more successful possibly by stimulating the WHO DOTS program (Section 2.3).

In higher burdened countries (quasi-Poisson model), the proportion of smear-positive cases that are cured is a statistically significant predictor of TB incidence, while the proportion of smear-positive cases that fail treatment is statistically insignificant (Table 5.3). This result could indicate that countries with high TB incidence have a problem with contagious cases spreading additional TB disease. For these countries, the main concern could be controlling the sheer numbers of contagious TB cases, as indicated by the significance of SP cured in

predicting TB incidence.

Even though high burdened countries might need to focus on treatment success and low burden countries may need to focus on effectively treating active TB cases, both may also need to employ good overall TB treatment. Both high and low TB burdened countries have statistically significant treatment outcome interactions, indicating that potentially all facets of TB treatment must be monitored. The significant interactions show that even though each country may have aspects of treatment they need to focus on, it might also be necessary to improve all aspects of TB care if they desire to decrease incidence.

The larger coefficient on GNI per capita for the quasi-Poisson could indicate that an increase/decrease in wealth has more of an effect on TB incidence in high burdened counties than in low burdened countries. This could be possible because high burdened countries tend to be poorer. With very poor countries, a slight increase in wealth can, in turn, increase funding to public health care and ultimately TB control. A wealthy country that experiences a decrease in wealth can cut other luxuries while preserving the country's health care. A loss of funds in poor countries can result in the cutting of health programs because there are fewer luxuries to cut.

5.6 One Predictor: Smoking

The comparison of countries with high and low TB incidence via the quasi-Poisson and negative binomial models resulted in interesting results. The results and conclusions are merely inferences that require additional studies and analysis. However, the initial question of how smoking affects TB incidence has yet to be answered. Even though smoking prevalence was not a significant factor for this type of data for these two models, does not mean it is not an important predictor of TB incidence. This could be due to confounding between other significant predictors and smoking. For this purpose, a negative binomial regression

and quasi-Poisson model with the afore mentioned offset was fit with only one predictor: smoking. Both models were fit in order to accurately describe the effect of smoking on tuberculosis incidence in both low and high burdened countries. These models are written out below. The quasi-Poisson model is as follows:

$$\ln(\mu) = \beta_0 + \beta_1 x_{\text{Smoking}} + \ln(\text{pop}). \tag{5.6}$$

The negative binomial regression model is as follows:

$$\ln\left(\frac{1}{1+\kappa\mu}\right) = \beta_0 + \beta_1 x_{\text{Smoking}} + \ln(\text{pop}). \tag{5.7}$$

The results showed that without the other predictors smoking is significant in predicting tuberculosis incidence, however, it is negatively correlated with TB incidence in both models.

The goodness of fit statistics for the quasi-Poisson model are in Table 5.9. The scaled deviance and scaled Pearson X^2 statistics are close to one, indicating good model fit. The parameter estimates for the effect of smoking and the dispersion parameter can be seen in Table 5.10. Based on this model the log of TB incidence is estimated to drop 0.0288 for every one percent increase in smoking prevalence. This is not intuitive, however, this result is discussed after the negative binomial regression has been analyzed.

Table 5.9: Model fit statistics for quasi-Poisson model with smoking as the only predictor.

Criterion	DF	Value	Value/DF
Deviance	83	3566234.1687	42966.6767
Scaled Deviance	83	70.2261	0.8461
Pearson Chi-Square	83	4214921.0426	50782.1812
Scaled Pearson χ^2	83	83	1.0
Log Likelihood		1788.2348	
AIC		3567193.8516	
BIC		3567198.7369	

The goodness of fit statistics for the negative binomial model are in Table 5.11. The scaled deviance and scaled Pearson X^2 statistics are close to one, indicating a good model fit. The

Table 5.10: Parameter estimates for quasi-Poisson with smoking as the only predictor.

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	-5.9299	0.2015	<.0001
Smoking	-0.0288	0.0094	0.0021
Scale	225.349		

parameter estimate for the effect of smoking and the dispersion parameter are in Table 5.12. This model predicts that TB incidence decreases as smoking prevalence increases.

Table 5.11: Model fit statistics for negative binomial model with smoking as the only predictor.

Criterion	DF	Value	Value/DF
Deviance	83	99.7968	1.2024
Scaled Deviance	83	99.7968	1.2024
Pearson Chi-Square	83	97.7718	1.1780
Scaled Pearson χ^2	83	97.7718	1.1780
Log Likelihood		92593115.986	
AIC		1888.3129	
BIC		1895.6409	

Table 5.12: Parameter estimates for Negative binomial with smoking as the only predictor.

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	-5.6208	0.2663	<.0001
Smoking	-0.0467	0.0121	0.0001
Dispersion	1.1470	0.1530	

Generally, it is assumed that whatever health problem you have, smoking makes them worse. Because of this, the negative correlation of smoking and TB incidence is not intuitive; however, further analysis indicated that smoking prevalence is confounded with other significant factors. First, it will be interesting to discuss some countries and their relative smoking prevalence and TB incidence. India, the country with the highest TB incidence, only has a smoking prevalence of 15%. Jamaica, the country with the lowest TB incidence, has a

smoking prevalence of 13%; while this percent is smaller than India's, it is only 2% less. However, Norway, which only has 320 incident TB cases, has a smoking prevalence of 23%. From these examples, there is a lot of variation in smoking prevalence that is not directly related to TB incidence. However, smoking appears to be correlated with known predictors of TB incidence in unexpected ways. Table 5.13 shows correlation (r) of smoking with TB incidence and other variables of interest.

Table 5.13: Correlation (r) of smoking with other factors and TB incidence.

	Smoking Prevalence
TB Incidence	-0.0024
GNI per capita	0.2379
SP Cured	-0.0916
SP Died	0.1114
SP Failed	0.2983
Ret Died	0.2238

While these correlations may be surprising initially, they are understandable. In Section 2.4, previous studies showed increase risk of death from TB disease among smokers. For this reason, it is not surprising that smoking is positively correlated with proportion of retreatment and smear-positive cases that died. Previous studies correlated smoking with worsened TB disease; making the disease harder to treat. This reasoning could explain the negative correlation of smoking with proportion of smear-positive cases cured and positive correlation with smear-positive cases that were not successfully treated. The most baffling result is the 0.2379 correlation between smoking and GNI per capita. Smoking is generally considered a plaque of the poor. However, with this correlation, it appears that smoking may not solely be a vice of the poor. A basic linear model to predict GNI per capita using smoking prevalence was fit. This linear model estimates that for every one percent increase in smoking prevalence, there is a \$316.4 increase in GNI per capita (p-value=0.0283). In conclusion, the negative coefficient on smoking seems to be counterintuitive at first glance, but it could be

due to the positive correlation between smoking prevalence and other significant predictors of TB incidence (GNI per capita, SP and Ret died), which are in turn negatively correlated with TB incidence. Under these conditions, the negative smoking coefficient is reasonable and not unexpected.

Even though these are simple models, the fit is better than expected. However these models still have larger residuals and Cook's distances than the models with multiple predictors. Figure 5.6 shows the residuals and Cook's distances for fitted values from the quasi-Poisson model with one factor. Figure 5.7 shows the residuals and Cook's distances for the fitted values given by the negative binomial regression with smoking as the only predictor. In the next section the predicted values of these models will be compared with the values predicted from the models fit in Section 5.3 and 5.4.

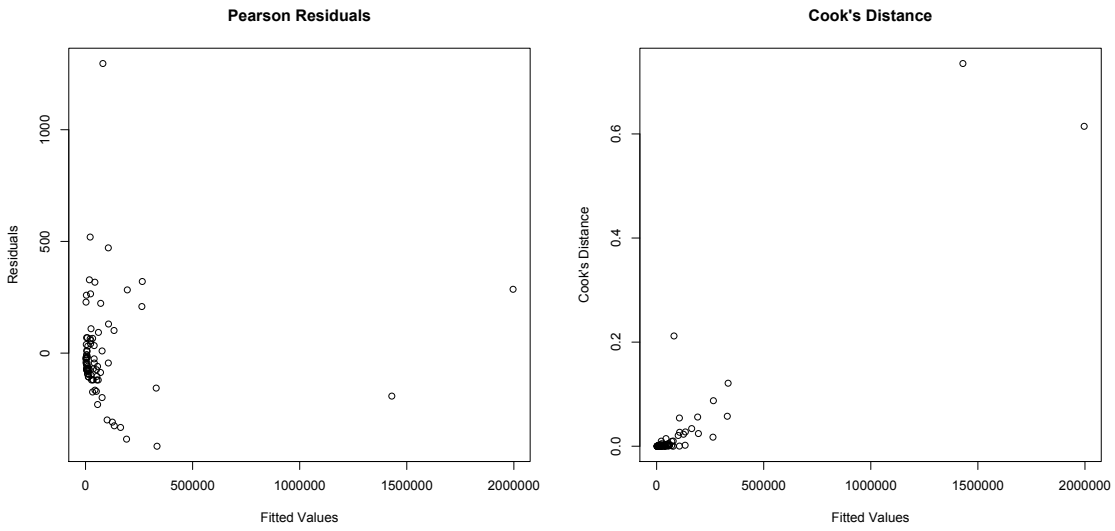

Figure 5.6: Left panel: Pearson residuals for the quasi-Poisson model with one predictor. Right panel: Cook's distances for quasi-Poisson regression model with one predictor.

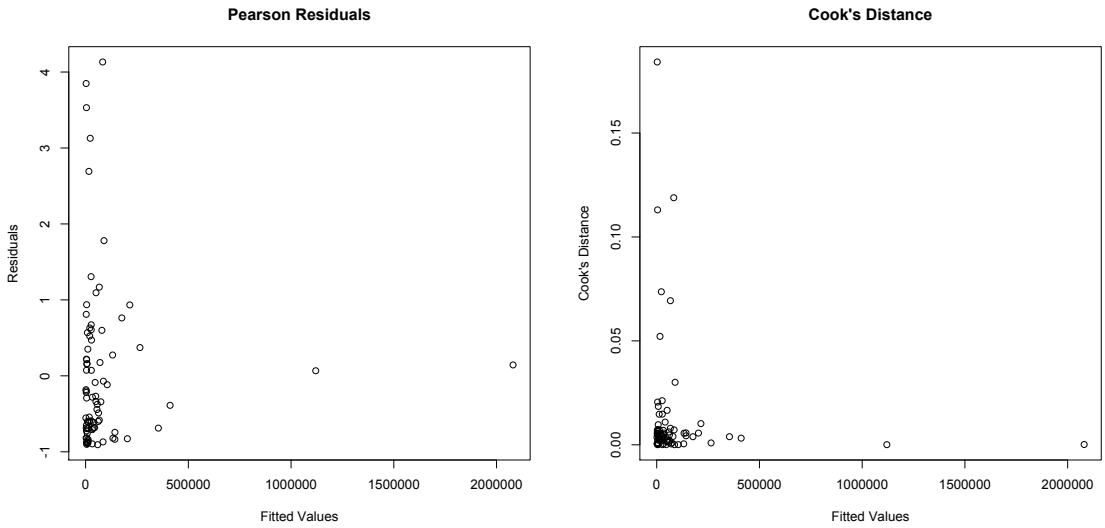

Figure 5.7: Left panel: Pearson residuals for the negative binomial regression model with one predictor. Right panel: Cook's distances for the negative binomial regression model with one predictor.

5.7 Overall Comparisons

In the above sections, four different models were used to estimate TB incidence. Some models confirmed hypotheses, while others assisted in connecting factors that affect tuberculosis incidence. The models with a single predictor (smoking prevalence) did not fit the data as well as models with multiple predictors, but these simple models helped to investigate the effect of smoking on tuberculosis incidence. In this section, model predictions will be compared for the four models previously fit in Sections 5.3, 5.4, and 5.6.

Figure 5.8 shows the log of the predicted values for TB incidence of all four models against the log of the actual values of tuberculosis incidence. Since there is a large spread in tuberculosis incidence, these values have been plotted against an index, corresponding to relative number of tuberculosis incidence. All data sets have been ranked in order of increasing tuberculosis incidence. For example, index one is the country with the smallest TB incidence.

The last plotted value, index 85, is India, the country with the largest TB incidence. Plotting the values by an index of increasing TB incidence makes the model fits more visible because the spread of TB incidence between countries is so large. This plot gives a visual representation of model performance and in which range each model is preferred. The models fit with smoking as the only predictor seem to be similar for both quasi-Poisson and negative binomial. However, these models have much larger discrepancies from the actual values of TB incidence than the models with multiple predictors.

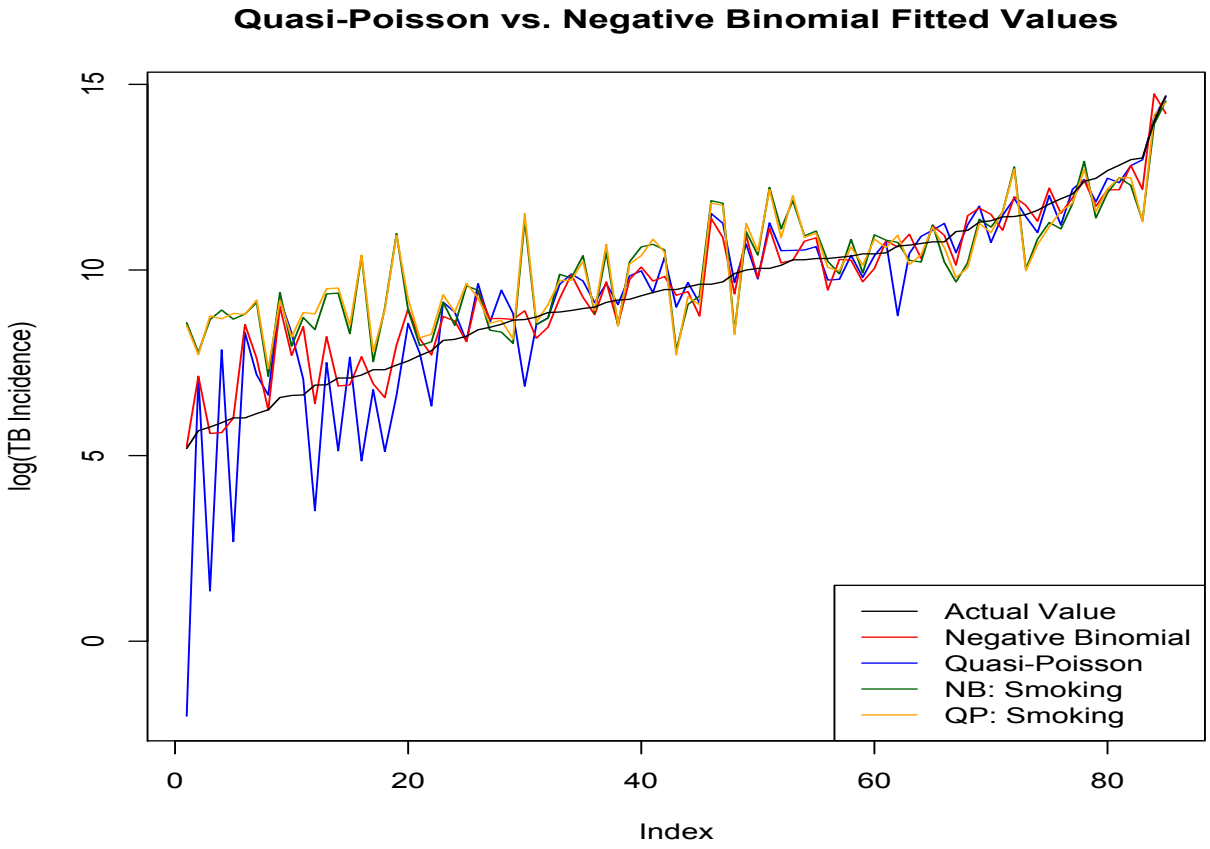

Figure 5.8: Fitted values against actual values for tuberculosis incidence on the log scale.

As can be seen from the graph, negative binomial regression with multiple predictors fits countries with smaller TB incidence better. Quasi-Poisson with multiple factors poorly fits

countries with small TB incidence, but excels at modeling countries with high TB incidence. While this result is hard to see when all model fits are plotted, it can be seen from individual plots of low and high TB incidence with only the fitted lines of the quasi-Poisson and negative binomial regression models with multiple predictors. Figure 5.9 shows low and high TB incidence countries with fitted values for only quasi-Poisson and negative binomial models that contain multiple predictors. From this figure, it is readily seen that for low TB incidence, the negative binomial fitted values (red line) are closer to the actual values of TB incidence. For high levels of TB incidence, the quasi-Poisson fitted values (blue line) are closer to the actual value of TB incidence. The graphical representations in this section are consistent with the other diagnostic plots, statistics and conclusions discussed in this chapter. In the next and final section of this chapter, the exposure rate of TB is investigated by modeling the underlying gamma distributions from the negative binomial regression model with multiple predictors.

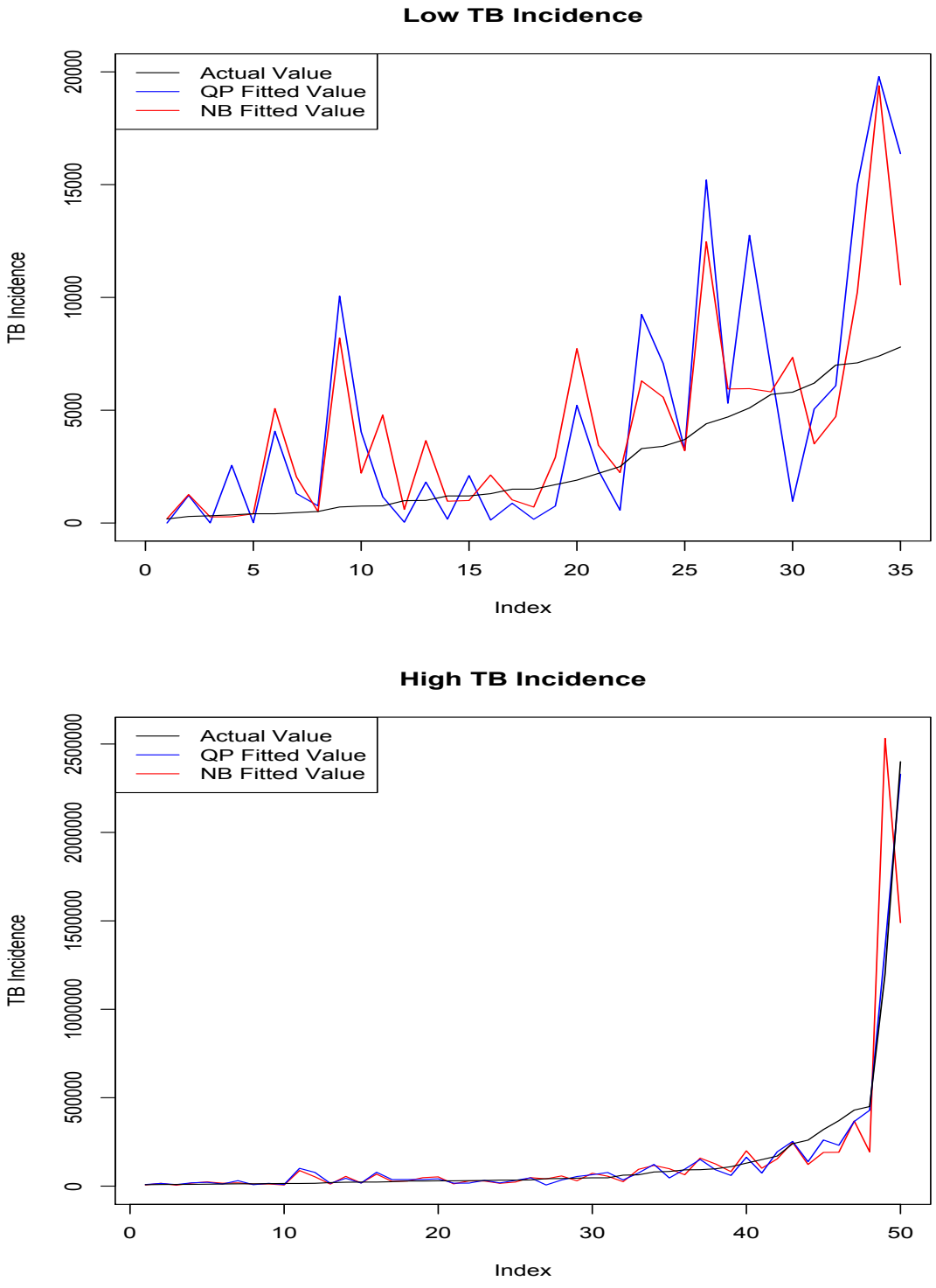

Figure 5.9: Top: Fitted values for the quasi-Poisson model and the negative binomial regression for countries with low TB incidence. Bottom: Fitted values for quasi-Poisson model and the negative binomial regression for countries with high TB incidence.

5.8 Underlying Gamma Distribution

As stated in Section 4.6, the negative binomial can be thought of as a combination of Poisson-gamma mixture models. In this section, the resultant gamma distribution from the negative binomial is explored. The underlying gamma distribution has an interesting interpretation when modeling contagious diseases. In this case, the underlying gamma distribution can be thought to express the contagious nature of TB, or the rate of exposure to TB. The parameters of the gamma distribution are shown in terms of r and p in Equation 4.25. It is assumed that r is constant for every country, however, the value of p varies from country to country. For this reason, there is an underlying gamma distribution for every country in the data set. A comparison of underlying gamma distributions for wealthy vs. poor countries is considered. Then, the underlying gamma distributions of countries with high and low TB incidence is compared.

The gamma distributions were generated using the specified r and p for each country. The x-values used to generate the gamma distribution ranged from 0 to 100,000. The plots only extend to 8,000 on the x-axis because the gamma distributions appear the same after this point. Figure 5.10 shows all 85 underlying gamma distributions. From this figure, it appears that countries with high TB incidence have gamma distributions that have a larger spread. Countries with low TB incidence appear to have underlying gamma distributions that have less variation. The same can be seen when looking at the underlying gamma distributions grouped by country wealth.

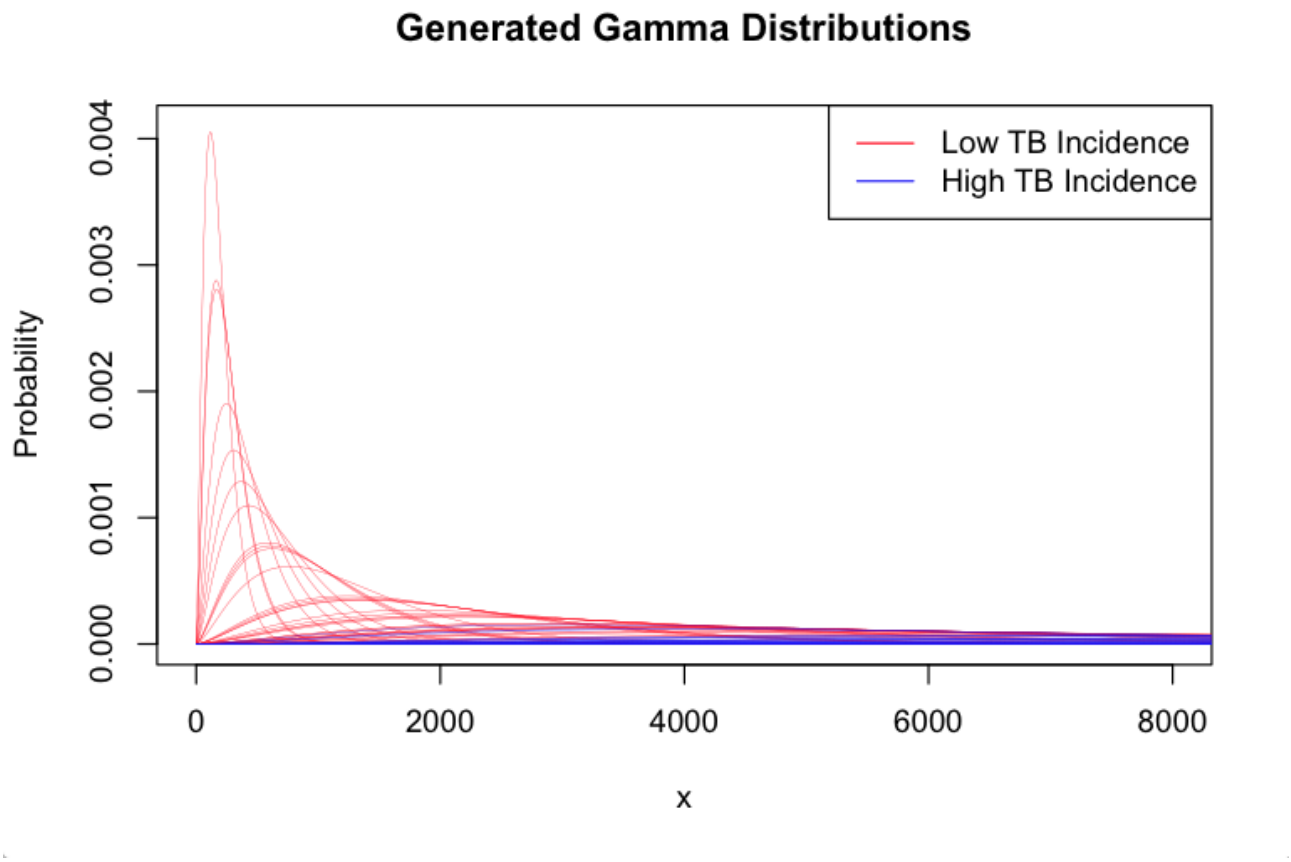

Figure 5.10: All 85 generated gamma distributions from the negative binomial regression, for both high and low TB incidence countries grouped by TB incidence.

Figure 5.11 again shows all the underlying gamma distributions for the 85 countries that were analyzed, only this time they are grouped differently. The 42 countries with the lowest GNI per capita of the 85 countries were classified as "poor." These countries had a GNI per capita less than $2,825. Any country with GNI per capita greater than or equal $2,825, i considered "wealthy." In reality, this value of GNI per capita splits the countries in half; the lower half is classified as poor and the upper half is classified as wealthy. Richer countries tend to have less variation in their underlying gamma distributions. Poorer countries appear to have more variation in their gamma distributions. In summary of Figures 5.10 and 5.11,

it appears that countries with higher TB incidence and lower income have greater spread in their underlying gamma distributions. It also appears that countries with low TB incidence and wealth have less spread in their underlying gamma distributions.

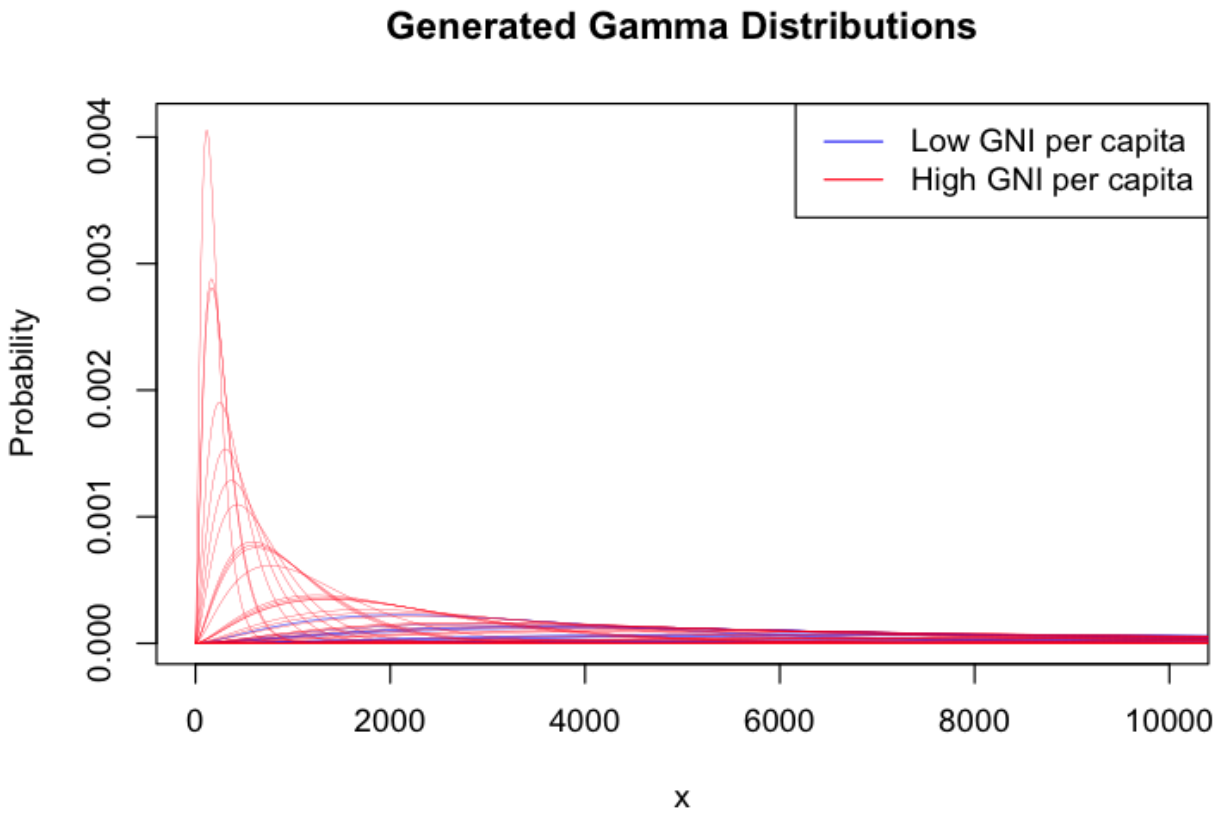

Figure 5.11: All 85 generated gamma distributions from the negative binomial regression, for both poor and rich countries.

For this analysis, the underlying gamma distribution can be thought of as the exposure to tuberculosis for individuals within a given country. For countries with high TB burden, there are extremes in exposure. These graphs imply that there are people who are constantly exposed, and those who will never encounter the disease. This is similar for poor countries. This could be because there is an uneven distribution of wealth in poor coun-

tries; some people are very rich and others are very poor. Due to this discrepancy, the poor will almost always be exposed because of their living situations while the rich may never enter a neighborhood that is infested with TB. This large variation in exposure could explain the spread of the underlying gamma distribution for poor and high burdened countries.

Inhabitants of countries with low TB incidence appear to have less variation in exposure. This could be because countries with low TB incidence do not have many regions that are infested with TB disease. Most TB cases in low burdened countries are few and far between. For this reason, most people are never or rarely exposed to TB. This could be a possible reason of why the variation in exposure is so small for low burdened countries. The same could be true for richer countries. In richer countries, the standard of living is higher. Most people can afford to live in areas that are clean and spacious, qualities which are not conducive to TB spread. It could be that these people will rarely, if ever, be exposed to tuberculosis. This reason could possibly explain the lack of spread in the underlying gamma distributions of wealthy and low burdened countries.

CHAPTER 6

CONCLUSION

In this project, different variables and their connection to tuberculosis incidence was investigated. Chapter 1 discussed the fact that tuberculosis is a re-emerging epidemic and laid a foundation for the importance of watching and studying this disease. A literary review in Chapter 2 indicated that poverty, public health infrastructure, HIV and smoking are all things that have been historically linked to tuberculosis. In Chapter 3, a preliminary analysis of the data revealed collinearity of treatment outcomes. The quasi-Poisson and negative binomial regression models as well as the violation of equidspersion were discussed in Chapter 4. In Chapter 5, the fitted models were evaluated and compared.

In regards to the models, both quasi-Poisson and negative binomial regression were useful in predicting TB incidence. Using IRLS, the negative binomial regression model gives relatively less weight to large/extreme observations when compared to the quasi-Poisson model. Therefore, smaller observations have a greater effect on the estimates for the negative binomial model. For this reason, negative binomial regression was a better model for countries with lower TB incidence. The quasi-Poisson model, on the other hand, better modeled countries with higher TB incidence. The models that included smoking as the only predictor provided interesting results, but were not good predictors of TB incidence per country.

It was indicated that there were statistically significant interactions of treatment outcomes for countries with high or low TB incidence. These statistically significant interactions suggest that it might be necessary to improve all aspects of TB care to reduce TB incidence.

For countries with high TB incidence, the main concern might be the sheer number of active TB cases. The proportion of smear-positive cases that die was a significant factor in predicting TB incidence in these countries. Although increased death rate does appear to hinder the spread of the disease, it is not the preferred method of control. This indicates that a potential emphasis in these countries could be to cure both new and re-treatment cases. It could be that because of the high number of TB cases, these countries are usually very diligent in overseeing the treatment of new cases. For countries with low TB incidence, it could be that their challenge is reducing drug resistance. Unsuccessful treatment (failures) give rise to drug resistant strains of tuberculosis. Treatment failure in smear-positive cases was a significant factor in predicting TB incidence in these countries that have an overall lower TB burden. Drug-resistant strains escalate the threat of TB everywhere. If drug resistance is truly the concern, these countries may need to oversee the treatment of their patients more diligently. However, it could be that because of the lower incidence in these countries, they are inclined to be more relaxed.

In models with smoking as the only predictor, smoking was shown to be a significant factor in both the quasi-Poisson and negative binomial regression models (estimated coefficient of -0.0288 with p-value=0.0021, estimated coefficient of -0.0467and p-value=.0001, respectively). Although the coefficients are quite small, they are both negative and both significant. These results seemed counter intuitive until it was indicated that, in this type of data, smoking was confounded in a surprising way with other factors: it was positively correlated with significant factors that are negatively correlated with TB incidence. For example, smoking prevalence was positively correlated with smear-positive and re-treatment deaths, which were negatively correlated with TB incidence. It was also negatively correlated with smear positive and re-treatment cases that were cured, which are negatively correlated with TB incidence. Because of these relationships, it can be seen how, with this type of data, smok-

ing could be negatively correlated with TB incidence. One of the confounding factors was surprising, however. Smoking is generally considered to be habit of the poor and should be negatively correlated with income, but, in this data, it was positively correlated with GNI per capita. This suggests that as smoking increases, GNI per capita is expected to increase, and as GNI per capita (and smoking) increase, TB incidence decreases. The linear model results showed that for every one percent increase in smoking prevalence, GNI per capita is expected to increase by \$316.40 (p-value=0.0283). Because of all these underlying correlations, smoking prevalence cannot be accurately separated from all the other factors, but these confounding factors make it easier to understand the unexpected negative correlation between smoking and TB incidence.

The underlying gamma distribution from the Poisson-gamma mixture model (negative binomial regression) showed the exposure rate of tuberculosis for different countries. Underlying gamma distributions showed more variation in exposure for poor and high TB burdened countries. Wealthier countries and countries with low TB incidence seemed to have less spread in their underlying gamma distributions. These results could indicate that in poor and high burdened countries, people are either constantly exposed, never exposed, or somewhere in between. For wealthy countries with low TB incidence, there could be less spread in exposure because most people are rarely, or never exposed to TB.

While all these results are interesting, it is important to remember that this analysis was performed to create future hypotheses to test. This analysis was a retrospective study performed at only one point in time, in the year 2006. These results are not conclusive and cannot be directly applied to TB treatment. However, these interesting results provide a starting point for public health facilities and people who wish to further study TB containment. Hopefully, these hypotheses will be used in future research to help stop the spread of tuberculosis.

BIBLIOGRAPHY

Altet-Gõmez, M., Alcaide, J., Godoy, P., Romero, M., and Hernandez, d. R. I. (2005), "Clinical and epidemiological aspects of smoking and tuberculosis: a study of 13,038 cases." *The International Journal of Tuberculosis and Lung Disease*, 9, 430–436.

Avasthi, A. (2007), "Oldest Human TB Case Found in 500,000-Year-Old Fossil," *National Geographic*.

Awaisu, A., Haniki, M., Mohamed, N., Aziz, N. A., Sulaiman, S. A. S., Noordin, N. M., Muttalif, A. R., and Mahayiddin, A. A. (2010), "Tobacco use prevalence, knowledge, and attitudes among newly diagnosed tuberculosis patients in Penang State and Wilayah Persekutuan Kuala Lumpur, Malaysia," *PubMed*.

Bauman, R. W. (2011), *Microbiology with Diseases by Taxonomy*, Benjamin Cummings.

Blöndal, K. (2007), "Barriers to reaching the targets for tuberculosis control: multidrug-resistant tuberculosis," *Bulletin of the World Health Organization*, 387–394.

CDC (2011), "HIV and TB," Center for Disease Control.

Davies, P. (2001), "Drug-resistant tuberculosis," *J R Soc Med*, 94, 261–263.

Franke, M. F., Appleton, S. C., Bayona, J., Arteaga, F., Palacious, E., Llaro, K., Shin, S. S., Becerra, M. C., Murray, M. B., and Mitnick, C. D. (2008), "Risk Factors and Mortality Associated with Default from Multidrug-Resistant Tuberculosis Treatment," *CID*, 46, 1844–1851.

Hassmiller, K. (2006), "The Association between Smoking and Tuberculosis," *Salud Publica de Mexico*, 48, Supplement 1.

Helgason, T., Tómasson, H., and Zoega, T. (2004), "Antidepressants and public health in Iceland," *The British Journal of Psychiatry*, 184, 157–162.

Hilbe, J. M. (2007), *Negative Binomial Regression*, Cambridge University Press.

Hoef, J. M. V., and Boveng, P. L. (2007), "Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count data?" *Ecology*, 88, 2766–2772.

Lam, T., Ho, S., Hedley, A., Mak, K., and Petro, R. (2001), "Mortality and smoking in Hong Kong: case-control study of all adult deaths in 1998," *BMJ*, 323.

Lindén, A., and Mäntyniemi, S. (2011), "Using the negative binomial distribution to model overdispersion in ecological data," *Ecology*, 92, 1414–1421.

Lönnroth, K., and Raviglione, M. (2008), "Global Epidemiology of Tuberculosis: Prospects for Control," *Seminars in Respiratory and Critical Care Medicine*, 29, 481–491.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models Second Edition*, Chapman & Hall.

Migliori, G. B., Centis, R., Lange, C., Richardson, M. D., and Sotgiu, G. (2010), "Emerging Epidemic of Drug-Resistant Tuberculosis in Europe, Russia, China, South America and Asia: Current Status and Global Perspectives," *Current Opinion in Pulmonary Medicine*, 16, 171–179.

Murray, J. F. (2004), "A Century of Tuberculosis," *American Journal of Respiratory and Critical Care Medicine*, 169, 1181–1186.

New York State Department of Health (2000), "Tuberculosis (TB)," Department of Health.

Pedan, A. (2001), "Analysis of Count Data Using the SAS System," *SUGI*, 247–26.

SAS Institute Inc (2008), *SAS/STAT®9.2 Users Guide*, Cary, NC: SAS Institute Inc.

Spence, D., Hotchkiss, J., Williams, C., and Davies, P. (1993), "Tuberculosis and Poverty," *BMJ*, 307, 759–761.

Tekkel, M., Rahu, M., Loit, H.-M., and Barburin, A. (2002), "Risk factors for pulmonary tuberculosis in Estonia," *The International Journal of Tuberculosis and Lung Disease*, 6, 878–894.

WHO (2009a), "Global Tuberculosis Control: WHO report 2009," World Health Organization.

—— (2009b), "WHO Report On The Global Tobacco Epidemic, 2009," World Health Organization.

WHO (2010), "10 Facts About Tuberculosis," World Health Organization.

WHO (2010a), "Global Tuberculosis Control: WHO report 2010," World Health Organization.

—— (2010b), "World Health Statistics," World Health Organization.

World Bank (2011), "Total GNI 2006, Atlas method," World Development Indicators database.

APPENDICES

APPENDIX A

NEGATIVE BINOMIAL DERIVATION FROM POISSON-GAMMA MIXTURE MODEL

The distribution of the data (y_i) follows a poisson distribution of another factor λ_i that follows a gamma distribution:

$$Y_i|\lambda_i \sim Poi(\lambda_i),$$
$$\lambda_i \sim Gamma(\alpha_i, \beta_i).$$

The joint distribution can be written:

$$\begin{aligned} f(y,\lambda) &= f(y|\lambda)f(\lambda) \\ &= \frac{\lambda^{y+\alpha-1}e^{-\lambda-\frac{\lambda}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)y!}. \end{aligned}$$

By integrating over λ, the distribution of y can be found. The arithmetic for this solution is shown below.

$$\begin{aligned} f(y) &= \int_0^{\text{inf}} f(y,\lambda)d\lambda \\ &= \frac{\lambda^{y+\alpha-1}e^{-\lambda-\frac{\lambda}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)y!} \\ &= \frac{\lambda^{y+\alpha-1}e^{\lambda}(\frac{\beta+1}{\beta})}{\Gamma(\alpha)y!} \\ &= \frac{\lambda^{y+\alpha-1}e^{\lambda}(\frac{\beta+1}{\beta})}{\Gamma(\alpha)y!} \end{aligned}$$

The integral was manipulated to have a gamma kernel. After the integral of the kernel integrated to one, the equation left is as follows:

$$\begin{aligned} f(y) &= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{1}{\beta+1}\right)^{\alpha} \left(\frac{\beta}{\beta+1}\right)^{y} \\ &= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{1}{\beta+1}\right)^{\alpha} \left(1-\frac{1}{\beta+1}\right)^{y} \\ &= \binom{y+\alpha-1}{y} \left(\frac{1}{\beta+1}\right)^{\alpha} \left(\frac{\beta}{\beta+1}\right)^{y} \\ &\sim NegBin\left(r=\alpha, p_i=\frac{1}{\beta_i+1}\right). \end{aligned}$$

The parameter estimates from the negative binomial distribution, can be used to solve for the parameter estimates of the distribution of λ_i. This distribution follows a gamma distribution with parameters:

$$\begin{aligned} \alpha &= r \\ \beta &= \frac{1-p_i}{p_i}. \end{aligned}$$

APPENDIX B

PARAMETERIZATION OF NEGATIVE BINOMIAL FOR SAS

The probability density function of the negative binomial distribution in SAS is:

$$\begin{aligned} f(y) &= \frac{\Gamma(y+\frac{1}{\kappa})}{\Gamma(y+1)\Gamma(\frac{1}{\kappa})} \times \frac{(\kappa\mu)^y}{(1+\kappa\mu)^{y+\frac{1}{\kappa}}} \\ &= \binom{y+\frac{1}{\kappa}-1}{y} \left(\frac{1}{1+\kappa\mu}\right)^{\frac{1}{\kappa}} \left(1-\frac{1}{1+\kappa\mu}\right)^y. \end{aligned}$$

The parameterization in SAS has parameters κ and μ. This is not the normal parameterization for the negative binomial. Usually the parameters are r and p and the probability density function is written as follows:

$$f(x) = P(X = x|r,p) = \binom{r+x-1}{x} p^r (1-p)^r. \tag{B. 1}$$

By comparing the SAS and general mass function for the negative binomial distribution, r and p can be written in terms of κ and μ. For the commonly used negative binomial parameterization,

$$\begin{aligned} r &= \frac{1}{\kappa} \\ p &= \frac{1}{1+\kappa\mu}. \end{aligned}$$

APPENDIX C

R CODE

```
#####Read in Data Sets#####

####General Stats and TB burden####

TB_burden <- read.csv("/Users/Noel/Documents/Research/TB Data/

TB_burden_countries_2012-01-19-1.csv",header=TRUE)


####Treatment outcomes####

TB_outcomes <- read.csv("/Users/Noel/Documents/Research/TB Data/

TB_outcomes_2012-01-19.csv",header=TRUE)


####Smoking Percents####

TB_smoke <- read.csv("/Users/Noel/Documents/Research/TB Data/TB_Smoke.csv",

header=TRUE)


####GNI####

TB_GNI <- read.csv("/Users/Noel/Documents/Research/TB Data/TB_GNI.csv",

header=TRUE)


###Select only year 2006###

TB_burden_2006 <- TB_burden[which(TB_burden$year==2006),]

TB_outcomes_2006 <- TB_outcomes[which(TB_outcomes$year==2006),]


####Create Data set that only contains variables of interest####
```

```
TB <- cbind(TB_burden_2006[,1], TB_burden_2006$e_pop_num,
TB_burden_2006$e_inc_num, TB_burden_2006$e_inc_tbhiv_num/TB_burden_2006$e_inc_num,
TB_smoke[ ,2], TB_GNI[,2], TB_outcomes_2006[ ,8:13], TB_outcomes_2006[ ,21:26])

colnames(TB) <- c("Country","Population", "TB_Incidence",
"TBHIV_Incidence", "Smoking", "GNI", "SP_cohort", "SP_cured", "SP_Cmplt",
 "SP_Died", "SP_Fail", "SP_Def", "R_cohort", "R_cured", "R_Cmplt", "R_Died",
 "R_Fail", "R_Def")

##Test for difference in means between ommitted and not##
##Which observations have missing values and are being omitted##
omitted <- 1
for(i in 1:213){
omitted <- rbind(omitted, any(is.na(TB[i,])))
}

omitted <- omitted[-1,]
omitted <- as.vector(omitted)

##Account for the ones that have no cohorts (SP or RET)##
omitted.data <- c(which(omitted==1), 42)

##Observations that are retained##
kept <- which(omitted==0)
kept <- kept[-17]
kept <- as.vector(kept)
```

```
##Omitted data set and retained (kept) data##
omitted.part <- TB[omitted.data,]
kept <- TB[kept,]

##Histogram of retained and omitted data##
par(mfrow=c(1,2))
hist(omitted.part$TB_Incidence, breaks=20, main="Missing",
xlab="TB Incidence")
hist(kept$TB_Incidence, breaks=20, main="Complete",
xlab="TB Incidence")

summary(omitted.part$TB_Incidence)
summary(kept$TB_Incidence)
library(xtable)
xtable(as.matrix(summary(omitted.part$TB_Incidence)))
xtable(as.matrix(summary(kept$TB_Incidence)))

#Box plots of retained and omitted data#
boxplot(kept$TB_Incidence)
boxplot(omitted.part$TB_Incidence)

##4 observations missing TB_Incidence##
omitted.part[which(is.na(omitted.part$TB_Incidence)=="TRUE"),1]

##What variables are the observations missing##
##miss=variables (column number) that are missing values##
miss <- 0
```

```
for(i in 1:127){
miss <- c(miss, which(is.na(omitted.part[i,])=="TRUE"))
}
miss <- miss[-1]

##How often are the different variables missing##
freq <- tapply(miss,miss,length)
as.numeric(names(freq)[which.max(freq)])

##How many missing values are there per observation##
##num.miss = number of missing values per observation##
num.miss <- 0
for(i in 1:127){
num.miss <- c(num.miss, length(which(is.na(omitted.part[i,])=="TRUE")))
}
num.miss <- num.miss[-1]

##Frequency of number of missing values##
freq <- tapply(num.miss,num.miss,length)
as.numeric(names(freq)[which.max(freq)])

##overall number of missing values##
bot <- sum(num.miss)

##Observations with greater than 14 missing values##
ov.14 <- which(num.miss>14)
omitted.part[ov.14,1]
```

```
##France and Iceland##
omitted.part[which(omitted.part$Country=="Iceland"),]
omitted.part[which(omitted.part$Country=="France"),]

##The omitted data ordered as a function of decreasing TB incidence##
order(omitted.part$TB_Incidence)

##Test for difference in means##
##Not assuming equal variance of TB incidence for omitted and retained##
t_test <- t.test(omitted.part$TB_Incidence, kept$TB_Incidence,
var.equal=FALSE)

##Assuming equal variance of TB incidence for omitted and retained##
t_test <- t.test(omitted.part$TB_Incidence, kept$TB_Incidence,
var.equal=TRUE)

##Mean and variance of omitted observations##
mean.o <- mean(omitted.part$TB_Incidence)
var.o <- var(omitted.part$TB_Incidence)

##Mean and variance of retained observations##
mean.k <- mean(kept$TB_Incidence)
var.k <- var(kept$TB_Incidence)

###Remove Missing Values and output data set, 87 observations###
TB.reduced <- na.omit(TB)
```

```
##Take out observations that have no observed treatment outcomes##
which(TB.reduced$SP_cohort==0)
which(TB.reduced$R_cohort==0)

##Italy (SP and R) and Kuwait ##
TB.reduced[c(37, 42), ]

##Remove Italy and Kuwait from data set##
TB <- TB.reduced[-c(37,42), ]

##Put all variables of interest into the model: Treatment outcomes,
GNI per capita, and log of population##
###Create percents for treatment outcomes###
SP_cured_p <- TB$SP_cured/TB$SP_cohort
SP_cmplt_p <- TB$SP_Cmplt/TB$SP_cohort
SP_died_p <- TB$SP_Died/TB$SP_cohort
SP_fail_p <- TB$SP_Fail/TB$SP_cohort
SP_def_p <- TB$SP_Df/TB$SP_cohort

R_cured_p <- TB$R_cured/TB$R_cohort
R_cmplt_p <- TB$R_Cmplt/TB$R_cohort
R_died_p <- TB$R_Died/TB$R_cohort
R_fail_p <- TB$R_Fail/TB$R_cohort
R_def_p <- TB$R_Df/TB$R_cohort

##Merge outcome percents and the rest of the data set##
```

```
outcome.percents <- cbind(SP_cured_p, SP_cmplt_p, SP_died_p, SP_fail_p,
SP_def_p, R_cured_p, R_cmplt_p, R_died_p, R_fail_p, R_def_p)

TB <- cbind(TB, outcome.percents)
TB <- as.data.frame(TB)

## Log of Population##
LP <- log(TB$Population)

##GNI converted from millions ##
GNInm <- TB$GNI*1000000

##GNI per capita##
GNIpercap <- GNInm/TB$Population
TB.final <- cbind(TB, LP, GNIpercap)

##Output data set to be used for analysis in SAS##
write.table(TB.final,"/Users/Noel/Documents/Research/TB Data/TBall.csv", sep=",",
quote=FALSE)

##Fit quasi-Poisson in R for comparison##
mod<-glm(TB_Incidence~TBHIV_Incidence+GNIpercap
+TBHIV_Incidence*GNIpercap
+SP_cured_p
+SP_died_p
+R_died_p
+SP_cured_p:SP_died_p
```

```
+SP_cured_p:R_died_p
+SP_died_p:R_died_p
+SP_cured_p:SP_died_p:R_died_p+ offset(LP),family=quasipoisson(link='log'),
data=TB.final)
CI.mod <- predict(mod,type="response",se.fit=T)

###Create summary statistics###
summary(TB)

##Plots of TB incidence and GNIpercap, both for incidence and incidence rate##
par(mfrow=c(1,2))
plot(TB.final$GNIpercap, TB.final$TB_Incidence)
plot(TB.final$GNIpercap, TB.final$TB_Incidence/TB.final$Population)

plot(TB.final$GNIpercap, log(TB.final$TB_Incidence))
plot(TB.final$GNIpercap, log(TB.final$TB_Incidence/TB.final$Population))

plot(TB.final$GNIpercap, 1/(TB.final$TB_Incidence))
plot(TB.final$GNIpercap, 1/(TB.final$TB_Incidence/TB.final$Population))

##Plots of GNIpercap with TB incidence with labels (log and not)##
par(mfrow=c(1,2))
plot(TB.final$GNIpercap, TB.final$TB_Incidence, xlab="GNI per capita",
ylab="TB Incidence")
plot(TB.final$GNIpercap, log(TB.final$TB_Incidence), xlab="GNI per capita",
ylab="log(TB Incidence)")
```

```
##Plot GNI per capita against smoking prevalence##
par(mfrow=c(1,1))
plot(TB.final$GNIpercap, TB.final$Smoking, xlab="GNI per capita",
ylab="Smoking Percent")

##Look at correlation of treatment outcomes##
TB.collinearity <- TB[ ,-c(1:6, 7:18)]
cor(TB.collinearity)

library(xtable)
xtable(cor(TB.collinearity))

##Preliminary Plots##
attach(TB.final)
plot(TB.final$GNIpercap, log(TB.final$TB_Incidence),
ylab="log(TB Incidence)", main="GNI per capita vs. Incidence",
xlab="GNI per capita")
plot(TBHIV_Incidence, log(TB_Incidence), ylab="log(TB Incidence)",
xlab="Proportion TB-HIV Coinfection", main="TB-HIV Coininfection vs. Incidence")

plot(SP_cured_p, log(TB_Incidence), ylab="log(TB Incidence)",
xlab="Proportion Smear-positive Cured", main="Sp.cured vs. Incidence")
plot(SP_fail_p, log(TB_Incidence), ylab="log(TB Incidence)",
xlab="Proportion  Smear-positive Fail", main="Sp.fail vs. Incidence")

plot(SP_died_p, log(TB_Incidence), ylab="log(TB Incidence)",
xlab="Proportion  Smear-positive Died", main="Sp.died vs. Incidence")
```

```
plot(R_died_p, log(TB_Incidence), ylab="log(TB Incidence)",
xlab="Proportion  retreatment Cases Died", main="Ret.died vs. Incidence")

plot(Smoking, log(TB_Incidence), ylab="log(TB Incidence)",
xlab="Overall Smoking Prevalence", main="Smoking vs. Incidence")

##Break GNI per capita into four groups and analyze the effect on TB incidence##
TB.low <- TB.final[which(TB.final$GNIpercap <= 764.2),]
TB.med1 <- TB.final[which(TB.final$GNIpercap > 764.2 &
TB.final$GNIpercap <= 2880),]
TB.med2 <- TB.final[which(TB.final$GNIpercap > 2880 &
TB.final$GNIpercap <= 7565),]
TB.high <- TB.final[which(TB.final$GNIpercap > 7565),]

par(mfrow=c(1,2))

##TB Incidence vs. GNI per capita##
plot(TB.low$GNIpercap, TB.low$TB_Incidence, xlab="Low GNI per capita",
ylab="TB Incidence")
plot(TB.low$GNIpercap, log(TB.low$TB_Incidence), xlab="Low GNI per capita",
ylab="log(TB Incidence)")

plot(TB.med1$GNIpercap, TB.med1$TB_Incidence, xlab="Med1 GNI per capita",
ylab="TB Incidence")
plot(TB.med1$GNIpercap, log(TB.med1$TB_Incidence), xlab="Med1 GNI per capita",
ylab="log(TB Incidence)")
```

```
plot(TB.med2$GNIpercap, TB.med2$TB_Incidence, xlab="Med2 GNI per capita",
ylab="TB Incidence")
plot(TB.med2$GNIpercap, log(TB.med2$TB_Incidence), xlab="Med2 GNI per capita",
ylab="log(TB Incidence)")

plot(TB.high$GNIpercap, TB.high$TB_Incidence, xlab="High GNI per capita",
ylab="TB Incidence")
plot(TB.high$GNIpercap, log(TB.high$TB_Incidence), xlab="High GNI per capita",
ylab="log(TB Incidence)")

##GNI per capita vs. Smoking##
par(mfrow=c(2,2))
plot(TB.low$GNIpercap, TB.low$Smoking, xlab="Low GNI per capita",
ylab="Smoking", main="Low GNIpercap vs. Smoking")

plot(TB.med1$GNIpercap, TB.med1$Smoking, xlab="Med1 GNI per capita",
 ylab="Smoking", main="Med1 GNIpercap vs. Smoking")

plot(TB.med2$GNIpercap, TB.med2$Smoking, xlab="Med2 GNI per capita",
ylab="Smoking", main="Med2 GNIpercap vs. Smoking")

plot(TB.high$GNIpercap, TB.high$Smoking, xlab="High GNI per capita",
ylab="Smoking", main="High GNIpercap vs. Smoking")

##Output data sets of low, med1, med2, and high GNI per capita##
write.table(TB.low, "D:/TBlow.csv", sep=",", quote=FALSE)
write.table(TB.med1, "D:/TBmed1.csv", sep=",", quote=FALSE)
```

```
write.table(TB.med2, "D:/TBmed2.csv", sep=",", quote=FALSE)
write.table(TB.high, "D:/TBhigh.csv", sep=",", quote=FALSE)

##Testing full vs. reduced model from SAS##
#Quasi-Poisson#
pchisq(949032.2784-719784.5504, 74-64)

#Negative binomial#
1-pchisq(90.7877-89.7877, 75-64)

##Read in fit Quasi-Poisson##
poi <- read.csv("/Users/Noel/Documents/Research/TB Data/POIfit2.csv",
header=TRUE)
poi <- poi[order(poi$TB_Incidence, decreasing=FALSE), ]
poi[which(poi$res==max(poi$res)),]
poi[which(poi$cd==max(poi$cd)),]

##Read in fit NB##
nb <- read.csv("/Users/Noel/Documents/Research/TB Data/NBfit2.csv",
header=TRUE)
nb <- nb[order(nb$TB_Incidence, decreasing=FALSE), ]
nb[which(nb$res==max(nb$res)),]
nb[which(nb$cd==max(nb$cd)),]

nb[which(nb$fitted==max(nb$fitted)),]
nb[which(nb$TB_Incidence==max(nb$TB_Incidence)),]
```

```
##Fit with smoking only##
smk <- read.csv("/Users/Noel/Documents/Research/TB Data/smokenb.csv",
header=TRUE)
smk <- smk[order(smk$TB_Incidence, decreasing=FALSE), ]
smk[which(smk$res==max(smk$res)),]
smk[which(smk$cd==max(smk$cd)),]

smkp <- read.csv("/Users/Noel/Documents/Research/TB Data/smokepoi.csv",
header=TRUE)
smkp <- smkp[order(smkp$TB_Incidence, decreasing=FALSE), ]
smkp[which(smkp$res==max(smkp$res)),]
smkp[which(smkp$cd==max(smkp$cd)),]

##Plot residuals##
plot(poi$fitted, poi$res, main="Pearson Residuals", ylab="Residuals",
xlab="Fitted Values")
plot(nb$fitted, nb$res, main="Pearson Residuals", ylab="Residuals",
xlab="Fitted Values")
plot(smk$fitted, smk$res, main="Pearson Residuals", ylab="Residuals",
 xlab="Fitted Values")
plot(smkp$fitted, smkp$res, main="Pearson Residuals", ylab="Residuals",
xlab="Fitted Values")

##Plot Cooks Distance##
plot(poi$fitted, poi$cd, main="Cook's Distance", ylab="Cook's Distance",
 xlab="Fitted Values")
plot(nb$fitted, nb$cd, main="Cook's Distance", ylab="Cook's Distance",
```

```
xlab="Fitted Values")
plot(smk$fitted, smk$cd, main="Cook's Distance", ylab="Cook's Distance",
xlab="Fitted Values")
plot(smkp$fitted, smkp$cd, main="Cook's Distance", ylab="Cook's Distance",
 xlab="Fitted Values")

plot(poi$lev, main="Leverage", ylab="Leverage")
plot(nb$lev, main="Leverage", ylab="Leverage")
plot(smk$lev, main="Leverage", ylab="Leverage")

##Look at large TB countries vs. smaller##
xtable(rbind(cbind(nb[85,c(1,3,17)],(nb[85,3] - nb[85,17])),
cbind(poi[85,c(1,3,17)], (nb[85,17] - poi[85,17]))))
xtable(rbind(nb[84,c(1,3,17)],poi[84,c(1,3,17)]))

rbind(nb[1,c(1,3,17)],poi[1,c(1,3,17)])
rbind(nb[2,c(1,3,17)],poi[2,c(1,3,17)])
rbind(nb[3,c(1,3,17)],poi[3,c(1,3,17)])

rbind(nb[24,c(1,3,17)],poi[24,c(1,3,17)])
rbind(nb[25,c(1,3,17)],poi[25,c(1,3,17)])

rbind(nb[84,c(1,3,17)],poi[84,c(1,3,17)])
rbind(nb[85,c(1,3,17)],poi[85,c(1,3,17)])

##Compare the QP and NB fitted values for each country##
comp <- NA
```

```
result <- NA

for(i in 1:85){
int <-rbind(nb[i,c(1,3,17)],poi[i,c(1,3,17)])
comp <- rbind(comp, int)
}

##Calculate the number of countries who have a smaller raw
residual by fitting NB (when compared to QP)##
dnb <- NULL
dpoi <- NULL
result <- NULL

for(i in 1:85){
dnb <- abs(nb[i,3] - nb[i,17])
dpoi <- abs(nb[i,3] - poi[i,17])
result[i] <- dnb < dpoi
}
freq <- tapply(result,result,length)

##The precent of time NB is better than QP for TB <= 7800##
result.1 <- result[1:35]
freq <- tapply(result.1,result.1,length)

##The precent of time NB is better than QP for TB > 7800##
result.2 <- result[36:85]
freq <- tapply(result.2,result.2,length)
```

```
##Which country is the cut-off##
nb[which(nb$TB_Incidence==7800), ]

##Double check that 35 countries are in the "low TB incidence" category##
dim(nb[which(nb$TB_Incidence<=7800), ])

##From SAS CI limits with log##
par(mfrow=c(1,1))
plot(log(poi$fitted), col="blue", type="l", ylab="TB Incidence",
main="Quasi-Poisson Fitted Values")
points(log(poi$LCI), col="blue", type="l", lty=3)
points(log(poi$UCI), col="blue", type="l", lty=3)
points(log(nb$TB_Incidence), type="l")
legend("bottomright", c("Actual Value","Fitted Value","95% Confidence Intervals"),
 lty=c(1,1,3), col=c("black","blue","blue"))

plot(log(nb$fitted), col="red", type="l", ylab="TB Incidence",
main="Negative Binomial Fitted Values")
points(log(nb$LCI), col="red", type="l", lty=3)
points(log(nb$UCI), col="red", type="l", lty=3)
points(log(nb$TB_Incidence), type="l")
legend("bottomright", c("Actual Value","Fitted Value","95% Confidence Intervals"),
lty=c(1,1,3), col=c("black","red","red"))

##Both on Same Plot##
plot(log(poi$fitted), col="blue", type="l", ylab="log(TB Incidence)",
```

```
main="Fitted Values")
points(log(poi$LCI), col="blue", type="l", lty=3)
points(log(poi$UCI), col="blue", type="l", lty=3)
points(log(nb$TB_Incidence), type="l")

points(log(nb$fitted), col="red", type="l")
points(log(nb$LCI), col="red", type="l", lty=3)
points(log(nb$UCI), col="red", type="l", lty=3)
points(log(nb$TB_Incidence), type="l")

legend("bottomright", c("Actual Value","QP Fitted Value",
"QP 95% Confidence Intervals","NB Fitted Value","NB 95% Confidence Intervals"),
lty=c(1,1,3,1,3), col=c("black","blue","blue","red","red"))

##From SAS CI limits without log##
plot(poi$fitted, col="blue", type="l", ylab="TB Incidence", main="Fitted Values")
points(poi$LCI, col="blue", type="l", lty=3)
points(poi$UCI, col="blue", type="l", lty=3)
points(nb$TB_Incidence, type="l")
#legend("topleft", c("Actual Value","Fitted Value","95% Confidence Intervals"),
 lty=c(1,1,3), col=c("black","blue","blue"))#

points(nb$fitted, col="red", type="l")
points(nb$LCI, col="red", type="l", lty=3)
points(nb$UCI, col="red", type="l", lty=3)
points(nb$TB_Incidence, type="l")
```

```
legend("topleft", c("Actual Value","QP Fitted Value","QP
95% Confidence Intervals","NB Fitted Value","NB 95% Confidence Intervals"),
lty=c(1,1,3,1,3), col=c("black","blue","blue","red","red"))

##Both together without the log##
plot(poi$fitted, col="blue", type="l", ylab="TB Incidence",
main="Quasi-Poisson Fitted Values")
points(poi$LCI, col="blue", type="l", lty=3)
points(poi$UCI, col="blue", type="l", lty=3)
points(nb$TB_Incidence, type="l")
points(nb$fitted, col="red", type="l")
points(nb$LCI, col="red", type="l", lty=3)
points(nb$UCI, col="red", type="l", lty=3)
points(nb$TB_Incidence, type="l")
legend("topleft", c("Actual Value","Fitted Value","95% Confidence Intervals"),
lty=c(1,1,3), col=c("black","blue","blue"))

last.nb <- nb[36:85, ]
last.poi <- poi[36:85, ]
last.cil.nb <- nb[36:85, 19]
last.ciu.nb <- nb[36:85, 20]
last.cil.p <- poi[36:85, 19]
last.ciu.p <- poi[36:85, 20]

first.nb <- nb[1:35, ]
first.poi <- poi[1:35, ]
first.cil.nb <- nb[1:35, 19]
```

```
first.ciu.nb <- nb[1:35, 20]
first.cil.p <- poi[1:35, 19]
first.ciu.p <- poi[1:35, 20]


par(mfrow=c(1,1))
plot(log(last.poi$fitted), col="blue", type="l", ylab="log(TB Incidence)",
main="Quasi-Poisson vs. Negative Binomial Fitted Values")
points(log(last.nb$fitted), col="red", type="l")
points(log(last.nb$TB_Incidence), type="l")

##Graph of fitted values for high TB incidence (high.pdf)##
plot(1, type="n", axes=F, ylim=c(-1000,2550000), xlim=c(1,50),
ylab="TB Incidence", main="High TB Incidence", xlab="Index")
points(last.nb$fitted, col="red", type="l",ylab="TB Incidence",
main="High TB Incidence", xlab="Index")
points(last.poi$fitted, col="blue", type="l")
points(last.nb$TB_Incidence, type="l")

legend("topleft", c("Actual Value","QP Fitted Value", "NB Fitted Value"),
lty=c(1,1,1), col=c("black","blue","red"))

##If interested in having the confidence intervals##
points(last.cil.nb,type="l",col="red", lty=3)
points(last.ciu.nb,type="l",col="red", lty=3)
points(last.cil.p,type="l",col="blue", lty=3)
points(last.ciu.p,type="l",col="blue", lty=3)
```

```
legend("topleft", c("Actual Value","QP Fitted Value","QP 95
% Confidence Intervals", "NB Fitted Value","NB 95% Confidence Intervals"),
lty=c(1,1,3,1,3), col=c("black","blue","blue","red","red"))

##Graph of fitted values for low TB incidence (low.pdf)##
plot(1, type="n", axes=F, ylim=c(0,20000), xlim=c(0,35), ylab="TB Incidence",
main="Low TB Incidence", xlab="Index")
points(first.poi$fitted, col="blue", type="l")
points(first.nb$fitted, col="red", type="l")
points(first.nb$TB_Incidence, type="l")
legend("topleft", c("Actual Value","QP Fitted Value", "NB Fitted Value"),
lty=c(1,1,1), col=c("black","blue","red"))

##If interested in confidence intervals#
points(first.cil.nb,type="l",col="red", lty=3)
points(first.ciu.nb,type="l",col="red", lty=3)
points(first.ciu.p,type="l", col="blue", lty=3)
points(first.cil.p,type="l",col="blue", lty=3)

##NB confidence intervals are wider##
ci <- (nb[,20]-nb[,19]>poi[,20]-poi[,19])
which((nb[,20]-nb[,19]>poi[,20]-poi[,19]))
freq <- tapply(ci,ci,length)

##Percent of nb ci's that are larger for low TB incidence##
ci.1 <- ci[1:35]
freq <- tapply(ci.1,ci.1,length)
```

```
##Percent of nb ci's that are larger for high TB incidence##
ci.2 <- ci[36:85]
freq <- tapply(ci.2,ci.2,length)

##Plot of fitted values for ALL regression methods##
plot(log(poi$fitted), col="blue", type="l", ylab="log(TB Incidence)",
main="Quasi-Poisson vs. Negative Binomial Fitted Values")
points(log(nb$fitted), col="red", type="l")
points(log(smk$fitted), col="dark green", type="l")
points(log(smkp$fitted), col="orange", type="l")
points(log(nb$TB_Incidence), type="l")
legend("bottomright", c("Actual Value","Negative Binomial","Quasi-Poisson",
"NB: Smoking", "QP: Smoking"), lty=c(1,1,1,1,1), col=c("black","red","blue",
 "dark green", "orange"))


##Look at the Gamma distribution, in order of increasing TB incidence##
k <- 0.3920
m <- nb$fitted

r <- 1/k
p <- 1/(1+k*m)

##Put into Gamma distributions##
a <- r
b <- (1-p)/p
```

```
##Plot all generated Gammas##
plot(1, type="n", axes=F, ylim=c(0,.0041), xlim=c(0,8000), ylab="Probability",
xlab="x", main="Generated Gamma Distributions")

x <- seq(0, 100000, 0.1)
for(i in 1:35){
  points(x, dgamma(x, shape=a,scale=b[i]), col="red", type="l", lwd=.25)
}
for(i in 36:85){
  points(x, dgamma(x, shape=a,scale=b[i]), col="blue", type="l", lwd=.25)
}
legend("topright", c("Low TB Incidence", "High TB Incidence"), lty=c(1,1), c
ol=c("red","blue"))


##Look at the Gamma distribution, in order of increasing GNI per capita##
nb.gni <- nb[order(nb$GNIpercap, decreasing=FALSE), ]
k <- 0.3920
m <- nb.gni$fitted

r <- 1/k
p <- 1/(1+k*m)

##Put into Gamma distributions##
a <- r
b <- (1-p)/p
```

```
##Plot all generated Gammas##
plot(1, type="n", axes=F, ylim=c(0,.0041), xlim=c(0,10000), ylab="Probability",
 xlab="x", main="Generated Gamma Distributions")

x <- seq(0,100000, 0.1)
for(i in 1:42){
  points(x, dgamma(x, shape=a,scale=b[i]), col="blue", type="l", lwd=.25)
}
for(i in 43:85){
  points(x, dgamma(x, shape=a,scale=b[i]), col="red", type="l", lwd=.25)
}
legend("topright", c("Low GNI per capita", "High GNI per capita"), lty=c(1,1),
 col=c("blue","red"))

##Do gamma of 2 high and 2 low incidence countries##
k <- 0.3920
m <- nb$fitted

r <- 1/k
p <- 1/(1+k*m)

##Put into Gamma distributions##
a <- r
b1 <- (1-p[1])/p[1]
b2 <- (1-p[2])/p[2]
```

```
b3 <- (1-p[85])/p[85]
b4 <- (1-p[84])/p[84]

x <- seq(0, 10000, .10)
plot(1, type="n", axes=F, ylim=c(0,.004), xlim=c(0,10000), ylab="Probability",
xlab="x", main="Gamma Distribution: TB Incidence")
points(x, dgamma(x, shape=a,scale=b1), type="l", lwd=1,lty=1,
main="Gamma Distribution: TB Incidence", col="blue", ylab="Probability")
points(x, dgamma(x, shape=a,scale=b2), type="l", lty=1, lwd=1, col="dark blue",
 ylab="Probability")
points(x, dgamma(x, shape=a,scale=b3), type="l", lwd=1, col="red",
ylab="Probability")
points(x, dgamma(x, shape=a,scale=b4), type="l", lty=1, lwd=1, col="dark red",
main="Gamma Distribution", ylab="Probability")
legend("topright", title="Low TB Incidence ", c("Jamaica", "Mauritius"),
lty=c(1,1),
col=c("blue","dark blue"))
legend(x=7060,y=0.00319, title="High TB Incidence ", c("China      ",
"India      "),
lty=c(1,1), col=c("red","dark red"))

##Based on GNI per capita##
nb.gni <- nb[order(nb$GNIpercap, decreasing=FALSE), ]
nb.gni[c(1,2,84,85),]
k <- 0.3920
m <- nb.gni$fitted
```

```
r <- 1/k
p <- 1/(1+k*m)

##Put into Gamma distributions##
a <- r
b1 <- (1-p[1])/p[1]
b2 <- (1-p[2])/p[2]

b3 <- (1-p[85])/p[85]
b4 <- (1-p[84])/p[84]

x <- seq(0, 100000, 0.1)
plot(1, type="n", axes=F, ylim=c(0,.003), xlim=c(0,10000),
ylab="Probability", xlab="x", main="Gamma Distribution: GNI per capita")
points(x, dgamma(x, shape=a,scale=b3), type="l", lwd=1,
col="blue", main="Gamma Distribution: GNI per capita", ylab="Probability")
points(x, dgamma(x, shape=a,scale=b4), type="l", lty=1, lwd=1,
col="dark blue", main="Gamma Distribution", ylab="Probability")
points(x, dgamma(x, shape=a,scale=b1), type="l", lwd=1,lty=1,
col="red", main="Gamma Distribution", ylab="Probability")
points(x, dgamma(x, shape=a,scale=b2), type="l",  lty=1, lwd=1,
col="dark red", main="Gamma Distribution", ylab="Probability")
legend("topright", title="Low GNI per capita ", c("Eritrea  ", "Malawi  "),
 lty=c(1,1),
col=c("red","dark red"))
legend(x=6830,y=0.00239, title="High GNI per capita",
 c("Denmark", "Norway"), lty=c(1,1), col=c("blue","dark blue"))
```

```
legend("topright", c("Low GNI per capita","Eritrea", "Malawi",
"High GNI per capita",
"Denmark", "Norway"), lty=c(0,1,3,0,1,3), col=c("red","red","red",
"blue", "blue","blue"), ncol=2)

Interpret into gamma parameters
a <- r
b <- (1-avg.p)/avg.p
b.all <- (1-p)/p

m.gam <- a*b
var.gam <- a*b^2

m.gam.all <- a*b.all
var.gam.all <- a*b.all^2

plot(m.gam.all, ylab="Variance", main="Gamma Distributions Variance")
plot(var.gam.all, ylab="Mean", main="Gamma Distributions Expected Value")

x <- seq(0, 10000, 0.1)
plot(x, dgamma(x, shape=a,scale=b), type="l", lwd=2, main="Gamma Distribution",
ylab="Probability")
text(7900, 0.00022, expression(f(Lambda) == frac(1, b^a * Gamma(a)) *
      Lambda^(a - 1) * exp(-frac(Lambda, b))))
```

```
##Plot of all generated gamma distributions##
plot(1, type="n", axes=F, ylim=c(0,.0041), xlim=c(0,3000), ylab="Probability",
xlab="x", main="Generated Gamma Distributions")

x <- seq(0, 3000, 0.1)
for(i in 1:85){
  points(x, dgamma(x, shape=a,scale=b.all[i]), type="l", lwd=.25)
}

TBcor <- cbind(cor(TB$Smoking, TB$GNI), cor(TB$Smoking, TB$SP_cured_p), )
colnames(TBcor) <- c("GNI", "SP_cured_p")


###Smoking Correlations###
cor(TB$Smoking, TB$GNI)
cor(TB$Smoking, TB$SP_cured_p)
cor(TB$Smoking, TB$SP_cmplt_p)
cor(TB$Smoking, TB$SP_fail_p)
cor(TB$Smoking, TB$SP_died_p)
cor(TB$Smoking, TB$R_cured_p)
cor(TB$Smoking, TB$R_cmplt_p)
cor(TB$Smoking, TB$R_fail_p)
cor(TB$Smoking, TB$R_died_p)


###Interesting Countries###
```

```
Jamaica <- nb[1,]

Denmark <- nb[5,]

China <- nb[which(nb$Country=="China"),]

India <- nb[which(nb$Country=="India"),]

Russia <- nb[which(nb$Country=="Russian Federation"),]

Ukraine <- nb[which(nb$Country=="Ukraine"),]


plot(nb$fitted, nb$res)

plot(nb$TB_Incidence-nb$fitted)


##Smoking prevalence to predict GNI per capita##

 mod <- lm(smkp$GNIpercap~smkp$Smoking)

 summary(mod)
```

APPENDIX D

SAS CODE

```
/*Import csv file*/
proc print data=TB; run;
proc print data=TBall; run;
proc print data=TBlow; run;
proc print data=TBmed1; run;
proc print data=TBmed2; run;
proc print data=TBhigh; run;

 /*Fit using only Smoking*/
proc genmod data=TB plots=ALL;
model TB_Incidence = Smoking
/dist=nb
link=log offset=LP;
output out=fitsmoke pred=fitted cookd=cd leverage=lev RESCHI=res;
run;

proc genmod data=TB plots=ALL;
model TB_Incidence = Smoking
/dist=poisson
link=log scale=pearson offset=LP;
output out=fitpoi pred=fitted cookd=cd leverage=lev RESCHI=res;
run;
```

```
/*Final model 0.06: All TB data*/
proc genmod data=TB;
model TB_Incidence =
TBHIV_Incidence
Smoking
GNIpercap
TBHIV_Incidence*GNIpercap
Smoking*GNIpercap
TBHIV_Incidence*GNIpercap*Smoking
SP_cured_p
SP_died_p
R_died_p
SP_fail_p
SP_cured_p*SP_died_p
SP_cured_p*R_died_p
SP_cured_p*SP_fail_p
SP_died_p*R_died_p
SP_died_p*SP_fail_p
R_died_p*SP_fail_p
SP_cured_p*SP_died_p*R_died_p
SP_cured_p*SP_died_p*SP_fail_p
SP_died_p*R_died_p*SP_fail_p
SP_cured_p*SP_died_p*R_died_p*SP_fail_p/dist=negbin
offset=LP link=log;
run;
```

```
/*Final NB model*/
ods graphics on;
ods output parameterestimates=pe;
proc genmod data=TB plots=ALL;
model TB_Incidence = TBHIV_Incidence
GNIpercap
SP_cured_p
SP_died_p
R_died_p
SP_fail_p
SP_cured_p*R_died_p
SP_died_p*R_died_p
SP_cured_p*SP_died_p*R_died_p
/dist=negbin
offset=LP link=log;
run;
ods graphics off;

/*Get estimates for decimals*/
proc print data=pe; run;
 proc print data=pe label noobs;
      format estimate 12.10
             StdErr 12.10;
      title "Parameter Estimates";
      run;

/*Smoking only*/
```

```
proc genmod data=TB plots=ALL;
model TB_Incidence = Smoking
/dist=negbin
offset=LP link=log;
output out=fitnb pred=fitted cookd=cd leverage=lev RESCHI=res ;
run;

/*Fit final model with Quasi-Poisson*/
proc genmod data=TB;
model TB_Incidence =
TBHIV_Incidence
Smoking
GNIpercap
TBHIV_Incidence*GNIpercap
Smoking*GNIpercap
TBHIV_Incidence*GNIpercap*Smoking
SP_cured_p
SP_died_p
R_died_p
SP_fail_p
SP_cured_p*SP_died_p
SP_cured_p*R_died_p
SP_cured_p*SP_fail_p
SP_died_p*R_died_p
SP_died_p*SP_fail_p
R_died_p*SP_fail_p
SP_cured_p*SP_died_p*R_died_p
```

```
SP_cured_p*SP_died_p*SP_fail_p
SP_died_p*R_died_p*SP_fail_p
SP_cured_p*SP_died_p*R_died_p*SP_fail_p/dist=poisson
link=log scale=pearson offset=LP;
run;



/*Final quasi-Poisson model*/
ods graphics on;
ods output parameterestimates=pe;
proc genmod data=TB plots=ALL;
model TB_Incidence =
TBHIV_Incidence
GNIpercap
TBHIV_Incidence*GNIpercap
SP_cured_p
SP_died_p
R_died_p
SP_cured_p*SP_died_p
SP_cured_p*R_died_p
SP_died_p*R_died_p
SP_cured_p*SP_died_p*R_died_p
/dist=poisson
link=log scale=pearson offset=LP;
output out=fitpoi pred=fitted cookd=cd leverage=lev RESCHI=res;
run;
ods graphics off;
```

```
 proc print data=pe label noobs;
      format estimate 12.10
             StdErr 12.10;
      title "Parameter Estimates";
      run;

/*Estimate parameter Mu of Poisson*/
proc genmod data=TB;
model TB_Incidence = /dist=poisson;
ods output parameterestimates=pe;
run;
data pe;
set;
if _n_=1;
lamda = exp(estimate);
lower = exp(lowerwaldcl);
upper = exp(upperwaldcl);
run;
proc print;
var lamda lower upper;
run;


/*Estimate parameters of NB*/
proc genmod data=TB;
model TB_Incidence = /dist=negbin;
```

```
ods output parameterestimates=pe pred=p;
output out=diag2 pred=p;
run;
proc print data=pe; run;

proc print data=diag2; run;
proc transpose data=pe out=tpe;
var estimate;
id parameter;
run;
data tpe;
set;
if _n_=1;
nb_k = 1/dispersion;
nb_p = 1/(1+exp(intercept)*dispersion);
nb_mean = nb_k*(1-nb_p)/nb_p;
nb_var = nb_k*(1-nb_p)/nb_p**2;
run;
proc print;
run;

/*Plot histogram of the data*/
proc univariate data=TB noprint;
histogram TB_Incidence;
run;
```