MULTIPLE LINEAR REGRESSION MODEL OF VISCERAL LEISHMANIASIS IN BIHAR, INDIA

by

DARREN SHEETS

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN MATHEMATICS

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2009

ACKNOWLEDGEMENTS

I would first like to thank my supervising professor Dr. Hristo Kojouharov for taking the time and energy to find an interesting project for my thesis, and then encouraging me throughout the project. I would also like to thank Dr. Anuj Mubayi for his encouragement and direction throughout the project from start to finish. I also thank the members of my committee, Dr. D.L. Hawkins and Dr. Danny Dyer, for their contributions both in and outside of the classroom.

I also thank the many teachers and professors that have inspired and motivated me over the years, in particular, Dr. Don Leake and Dr. Janna Cowen. If not for these two professors, I would not be where I am today.

Most importantly I would like to thank my parents, Donald and Benita Sheets, who always supported me through my wildest endeavors, as well as my sisters Tricia and Julie. Lastly, I thank my many friends for their support and encouragement. In particular, Jason Gyamerah for all his time spent discussing my project.

April 17, 2009

ABSTRACT

MULTIPLE LINEAR REGRESSION MODEL

OF VISCERAL LEISHMANIASIS

IN BIHAR, INDIA

Darren Sheets, M.S.

The University of Texas at Arlington, 2009

Supervising Professor: Hristo Kojouharov

Visceral Leishmaniasis (VL) is one of the world's worst parasitic killers, second only to Malaria, claiming nearly 500,000 lives each year. The disease attacks the spleen, liver, and bone marrow, and if left untreated is nearly always fatal. Whilst the disease is found all around the world, it is primarily prevalent in developing countries, in particular India. The most affected state in India is Bihar, where the disease is endemic. While other research has been conducted with emphasis on the effect of climate variables on the disease incidence rate, this analysis focuses on socio-economic variables such as literacy rate, housing structure, and working environment, to study their roles on the incidence rate. A Multiple linear regression model that includes these socio-economic factors as independent variables was initially developed and it explained 92% of the observed variance. The model was then reduced via stepwise regression and two models that explained 81% and 63% of the observed variance were used to help determine the most significant variables, such as housing and literacy rates. Modest comments are made on possible measures that could be taken to decrease the VL incidence rate, along with limitations of the model and suggestions for further research on this topic.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1 Description and History

For the last 20 years South Asia has experienced a resurgence of kala-azar, also known as Anthroponotic Visceral Leishmaniasis. Leishmaniasis is a family of vector-borne diseases caused by an infection from a single celled organism called a protozoan. The vector, an organism which does not cause the disease but instead spreads the infection from one host to another, in this case is the female sandfly [1]. The two most common forms of Leishmaniasis that are observed are cutaneous and visceral. Cutaneous Leishmaniasis (CL) is the most common form of Leishmaniasis. CL infects the skin, and is characterized by raised, red lesions that appear at the site of the bite. The lesion then ulcerates and is susceptible to infection from bacteria, often causing permanent scaring. Visceral Leishmaniasis (VL), while not the most common form, is the most severe form of Leishmaniasis. VL is the second-largest parasitic killer in the world after Malaria [2]. The parasite attacks internal organs such as the liver, spleen, and bone marrow. Untreated the death rate is 90%, compared to a death rate with treatment of 10% [8]. There is also growing concern of the number of cases of HIV infected individuals contracting VL [3].

Anthroponotic refers to the fact that humans are the reservoir for this form of VL. This means that while a human is infected with the disease a sandfly would be capable of receiving the disease from the infected host during a blood meal, and transmitting the disease to a different human during a different blood meal. There is also a Zoonotic form of VL in which animals are the reservoir, and therefore the disease is transmitted only by animal hosts and humans are dead end hosts.

VL exists in 88 countries on five continents; however countries hit especially hard are India, Bangladesh, Nepal, Sudan, and Brazil [8]. These five countries contain 90% of the

estimated 500,000 cases that occur annually worldwide. India, Nepal, and Bangladesh account for an estimated 300,000 cases annually and 60% of the global burden [4]. It is estimated that 200 million people worldwide are at risk of contracting the disease, with 62 countries already endemic for VL. Rural communities are often hit hardest, and are also frequently the poorest and unable to afford the appropriate treatments. The cost of therapies and medicine that are available ranges from 30 to 300 US dollars, with varying efficiency and side effects [5]. There are no vaccines available to prevent infection, thus protection against sandfly bites is regarded as the best prevention from contracting the disease.

1.2 Goal of Project

This paper will focus on the state of Bihar, India which experienced its last major outbreak of VL in the early 1990's affecting hundreds of thousands of people [4]. In particular 21 of the 38 districts in Bihar are of particular interest, as they represent an area of the state that typically floods with the onset on the rainy season, and hence experience elevated levels of the disease.

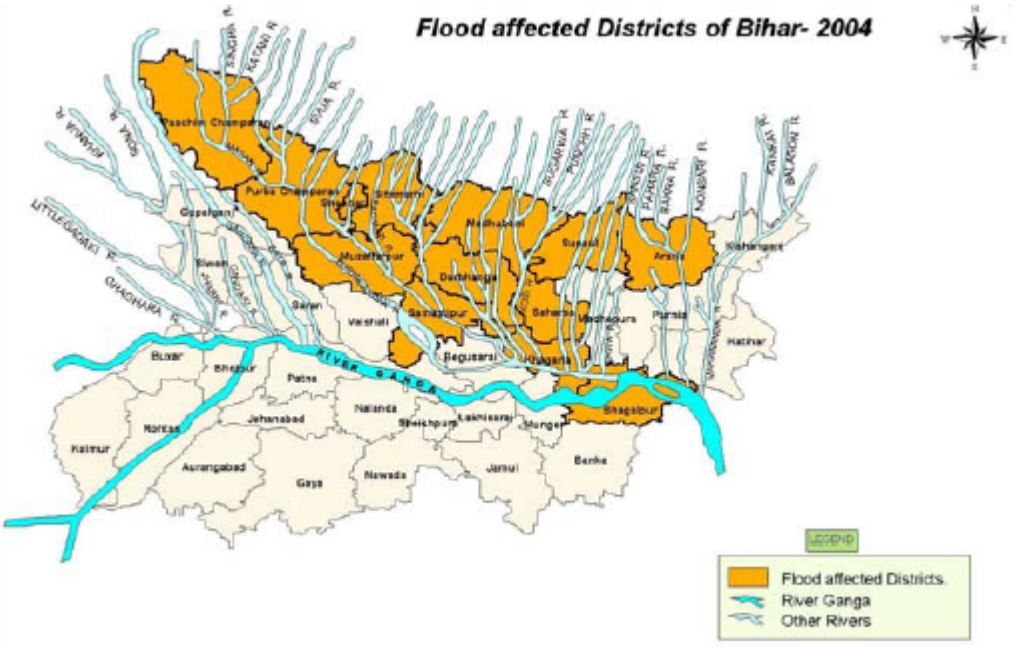


Figure 1.1: The gold colored districts represent regions often affected by floods, and are also heavily hit by VL.

Bihar lags behind the other states of India with respect to social and economic development, partially attributable to an ineffective central government [18]. However, the state government has made significant changes in improving educational facilities, infrastructure, and health care facilities [19]. Previous research on diseases related to VL, which is discussed in the following section, has shown how local climate variables such as rainfall and temperature can affect the incidence rate of the disease occurring. These variables can also be useful in attempting to predict when outbreaks are at a higher risk of occurring. This paper will focus primarily on how socio-economic variables such as literacy rates, construction of homes, type of employment, etc., are affecting the incidence rate of VL in Bihar, India. It is important to consider these variables because VL is a disease that is associated with third world countries. There are climate conditions in parts of the western world that are similar to those of India and other countries containing the disease. However, the disease is not observed in these parts of the western world, while some tropical and subtropical countries like India experience endemic levels of the disease. Potential reasons for differences in incidence rate between western and developing nations include effectiveness of government, poverty levels, income levels, and standard of living measures. With this in mind it's intuitive to begin to think of how the disease is related to these socio-economic variables.

CHAPTER 2

REVIEW OF PREVIOUS RESEARCH/STATISTICAL METHODS

2.1 Previous Research

2.1.1. American Cutaneous Leishmaniasis

Chaves et al. (2008) showed that the prevalence of American Cutaneous Leishmaniasis (ACL) is related to the region in which individuals work or live. The main result thus far is that infection is highest among individuals that live near forest edges. Workers that extract natural resources in forested areas are also at an elevated risk of ACL [6]. However, not considering the full multidimensionality of factors that exist in biological systems can lead to biased results. This is clear from Schmalhausen's law, which states that biological systems in extreme or unusual conditions with respect to one environmental variable are more vulnerable to small changes in any other environmental variable [7].

Therefore, Chaves et al. further considers a social marginalization index which is a measure of social-economic wellbeing. Factors that make up this index include; income, literacy rate, average distance to health centers, and level of education [6]. After comparing the index and ACL occurrence rates, statistical analysis is used to determine if there is a significant overlap of the two regions.

To determine where high levels of ACL are occurring, Chaves et al. used Kuldorff's Scan Statistic. This method assumes cases are generated by an inhomogeneous Poisson distribution. Then by moving a circular window through the study area, clusters containing an excess of cases can be detected [6].

The technique of Local Indicators of Spatial Autocorrelation (LISA) is used to determine patterns of clustering in the social marginality index. LISA compares a variable of interest in a given region with those in near-by regions. The degree of similarity is compared to that expected by random chance to determine where regions of low or high values occur [6].

When the two regions are compared it is discovered that disease incidence and the social marginalization index achieve their highest values in the same region [6]. Thus topics of further research should focus on socio-economic factors to determine their role in the spread of ACL. The specific methods used by Chaves et al. are beyond the scope of this paper, however, they are included to motivate the focus on socio-economic variables.

2.1.2. Dengue Fever

Several papers have been produced on the topic of Dengue Fever, a vector borne disease significant in countries such as south-east Asia, south Asia, and Latin America. In this case the vector is the mosquito, and the disease affects approximately 100 million people each year. Severe cases of Dengue Fever (Dengue Hemorrhagic Fever) result in a mortality rate of 15% for untreated cases and 5% for treated cases [12]. The susceptibility of the vector has been shown to be dependent on temperature, as well as increases in Dengue Fever related to higher rainfall [13]. All of these characteristics are closely related to VL, and therefore a thorough understanding of the techniques used in these papers is beneficial in understanding the models used in this paper.

This paper will focus on the use of multivariate linear regression analysis (discussed in later sections) to determine the relationship that exists between the incidence rate of VL and socio-economic variables for the state of Bihar, India. Regression analysis has been shown to be an important tool for the understanding of other vector borne diseases such as Dengue Fever. Joseph Keating investigated the relationship between cyclical Dengue Fever incidence and seasonal temperature fluctuation, and while his model was weakened due to positive autocorrelation and inconsistent prediction during peak months, a blueprint for future research and expansion was created [9]. His main goal was to show to what extent linear regression analysis could be used to model vector borne diseases. Autocorrelation is the correlation between error estimates (or residuals) at different points in time. If the monthly incidence rate of Dengue Fever is directly related to the next month's incidence rate, variance will also be directly

rated. A Durbin-Watson test can be administered to understand the extent of this effect. The test computes the following d value from the formula:

$$d = \frac{\sum_{t=2}^{T}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{T}\epsilon_t{}^2}$$

This value is then compared to upper and lower critical values. At significance level α, if $d<d_{lower}(\alpha)$ then there is statistical evidence at the α level that the error terms are positively auto correlated. If $d>d_{upper}(\alpha)$ there is statistical evidence at the alpha level that the error terms are not positively auto correlated. If d is between the critical levels, then the test is inconclusive. The Durbin-Watson test is not always relevant however. For instance, if the dependent variable is in a lagged form as an independent variable, or if the error is not normally distributed.

It was quickly discovered that Dengue cases reported increased about 12 weeks following peak temperature [9]. Such an effect could be due to a decrease in the extrinsic incubation period of the pathogen, resulting in an increase of the number of infectious vectors (in this case the mosquito) at a given time [10]. A null hypothesis of temperature lagged three months is not a significant factor was tested against the alternative hypothesis of temperature lagged three months is significant in Dengue monthly incidence. A first order regression model was created using the mean temperature data lagged three months, and the data lagged three months was shown to be significant and therefore retained in the model. A scatter diagram of temperature versus cases reported suggested curvilinearity may be present, and therefore stepwise regression was then used to create a second order model. This second order model was shown to be an improvement, producing a $\mathcal{R}^2$ statistic of 0.71 (versus a $\mathcal{R}^2$ statistic of 0.62 for the first order model). $\mathcal{R}^2$, or the coefficient of determination, is the proportion of variance in the data accounted for in the model. $\mathcal{R}^2$ is defined as:

$$\mathcal{R}^2 \stackrel{\text{def}}{=} 1 - \frac{SS_{error}}{SS_{total}}$$

Where SS_{error} is the sum of the square residuals, and SS_{total} is the total sum of squares. The sample information will ultimately be judged on its goodness of fit based on the

residuals, or the estimates of the error terms. These residuals are assumed to be a normal random variable with mean zero and variance σ^2. Residuals can be plotted in several ways to detect possible problems with the underlying assumptions. A popular test is the Durbin-Watson test, and when conducted for this data indicated a positive autocorrelation was present. The incidence of Dengue during a given month was highly correlated with the next month's incidence rate of Dengue. This is a violation of one of the inferences associated with the least squares method, and therefore jeopardizes the reliability of the model in the absence of considering the residuals. Overall, this model may be limited in the presence of residual autocorrelation and a small sample size. However, the original goal was achieved to present a regression model with the incidence rate of a vector borne disease as the dependent variable and one climatic variable (temperature) as an independent variable.

An outbreak of Dengue Fever occurred in 2002 in the state of Colima, Mexico. The authors, Gerardo Chowell and Fabio Sanchez, set out to show the correlation between Dengue incidence and several climate variables. Specifically, the correlation between incidence rate and precipitation, maximum and minimum temperature, mean temperature, and evaporation was considered by the authors. An average of these variables was obtained from meteorological offices, thus producing time-series data for 2001-2002. Lagged cross-correlation analysis was also considered due to the time it takes for mosquitoes to develop into adults, become infectious, infect a host, and the host led to clinical symptoms [12]. Questions to be answered are this; will all the climatological variables have lag effects, and if so will they be equal, and how significant are the climatological variables?

To answer these questions, first univariate regression analysis was used to determine the amount of variance contributable to each variable, and then multiple linear regression is used including all five climatological variables. Correlation coefficients were initially used to analyze the relationship between Dengue incidence rate and the individual climatological variables. Correlation coefficients indicate the direction and strength of a linear relationship between data. Calculated by the following:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(X-\mu_Y)]}{\sigma_X \sigma_Y}$$

The correlation coefficient must be between -1 and 1, where a positive number represents a positive slope of the regression line, and vice versa. It was found that minimum temperature had the largest correlation coefficient, .79, indicating individually that it has the strongest linear relationship with Dengue incidence rate. Mean temperature had the second strongest relationship, .74, followed by precipitation (.57), evaporation (.41), and max temperature (.29) [12]. It's of interest to lag certain climatological variables and observe if the correlation coefficient can be increased. Maximum temperature achieved its highest correlation at a lag of one month, while evaporation achieved its maximum at three months. The other three variables were found to be most correlated without a lag period [12]. Before the linear regression analysis, the square root transformation was applied to the Dengue Fever incidence rate. This was done to stabilize the variance and to linearize the relationship with climate variables. Performing univariate regression analysis without any lagging effects, minimum temperature explained the maximum amount of variance (75%). The second largest variance was explained by mean temperature (74%), followed by precipitation (34%), evaporation (30%), and maximum temperature (17%) [12]. A multiple linear regression model containing all the climatological variables explained 94% of the recorded variance with a p-value of less than 0.001. A reduced model containing only precipitation and evaporation as climatic variables explained 88% of the variance, and was significant at the 99% level. Another reduced model containing precipitation and maximum temperature explained 79% of the variance at the 99% confidence level. A full model with the appropriate lag adjustments for each variable reduced the overall observed variance from 94% to 86%. Chowell et al. found climatological variables were significantly related to Dengue incidence rate. It was also discovered that three of the five climate variables achieved their highest correlation with incidence rate without a lag period, while the remaining variables required lag periods (unequal) to achieve their maximum correlation. Similar results

have been observed with lag periods of 6-16 weeks for the same variables [14]. Further research indicates the importance of including several factors into a model, such as intensity of public health interventions, use of insecticides, and educating the public on disease spread [12].

2.1.3. Scorpion Stings in Colima, Mexico

Another interesting use of regression analysis involves predicting the incidence level of scorpion stings in Mexico with climatic variables. Scorpionism is a health issue in Africa, Middle East, India, Central Asia, and America. The highest fatality rate is nearly 1000 per year in Mexico [15]. Of the nearly 1500 scorpion species in the world, 25 are dangerous to humans, 8 of which are found in Mexico [16]. An increase in scorpion movements has been observed during the warmer months in Guanajuato, Mexico [16]. The impact of rainfall has been inconclusive, as studies have shown both increases and decreases in scorpion activity during the rainy season [17]. Using data from the Health Ministry and Mexican Institute of Public Health, it was discovered that nearly 70% of all cases occurred in the coastal municipalities where temperature surpasses 26 degrees Celsius during the summer. This time period is also associated with an increase in scorpion activity. For further analysis, the relationship between sting incidence and precipitation, minimum and maximum temperature, mean temperature and evaporation was considered. The data was for two full years (2000-2001), from eight meteorological offices located in 8 of 10 municipalities in Colima. Univariate analysis is first carried out on each individual climatological variable to determine the amount of variance attributable to that variable in the absence of the others. A multiple linear regression model using backward elimination is used, removing predictors at the 90% level or less, and leaving predictors significant at 95% of better. Their results showed approximately 3 stings per year per 1000 people in the cities of Colima and Villa de Alvarez, and 18-30 stings per year per 1000 people in the rest of the municipalities [17]. There are few stings or rain when the minimum temperature is less than 16 degrees Celsius [17]. Sting incidence increases monotonically with the minimum temperature [17]. When rainfall is above 30 mm per month, the scorpion sting incidence rate is independent of actual rainfall [17]. Evaporation was shown to be insignificant

with incidence rate [17]. In 8 out of the 10 municipalities the number of scorpion stings was proportional to the population. For these municipalities a simple linear regression model was used where the dependent variable was number of scorpion stings, and the independent variable was population size. This model explained 98.6% of the measured variance. For these same municipalities, correlation coefficients were also calculated to determine how related the climatological variables are. The highest correlation with sting incidence rate was with minimum temperature (r=.87), followed by mean temperature (.80), and precipitation (.72). Before carrying out regression analysis on the climate variables, the square root transformation was applied to precipitation to adjust for the non-linearity properties. In univariate analysis the maximum explained variance came from minimum temperature (75.41%), followed by mean temperature (63.25%), precipitation (60.58%), evaporation (11%), and maximum temperature (10.47%). A multiple linear regression model including all five climate variables (square root transformation for precipitation) explained 81.07% of the variance. Following the backward elimination procedure left the model with one predictor significant at the 95% level, minimum temperature, explaining 75.41% of variance with a p-value less than 0.0001. The same reduced model using minimum temperature as the only independent variable was applied to the other two municipalities. This model explained 14.77% of the observed variance, possibly this low due to limited climate information for this municipality (one station). These results showed a strong positive association between sting incidence rate and minimum temperature. There was also threshold results showing that sting incidence rates are independent of rainfall when rainfall is greater than 30 mm per month, and very few stings when less than 30 mm per month. The models derived here indicate that for every 1 degree Celsius increase in the minimum temperature there will be an increase of .2 stings per month per 1000 people in 8 of the 10 municipalities. This result can be used to help determine the proper allocation of anti-venom given changes in minimum temperature. Further research suggestions include considering the premise condition index (PCI) of households to determine if this plays a significant role, along with other socio-economic factors that could be at work.

Given this previous research, the topic of this paper will be focused primarily around the use of regression analysis to determine the relationship between the incidence of VL and several socio-economic variables. Climate variables were the main focus of the research described above and the results often lead to providing climate indicators of when possible outbreaks may occur. By using several socio-economic variables it is desired that the results would lead to a more pro-active list of variables that could reduce and control outbreaks, rather than solely indicate when outbreaks are at greater risk.

2.2 Statistical Methods

2.2.1. Simple Linear Regression

There are several statistical methods that will be used in this paper, and used in references of this paper, that make it appropriate to review the methods that will be used. The first concept is that of simple linear regression. Linear regression is a form of regression analysis in which the relationship between one dependent variable and one or more independent variables is modeled. The mathematical representation of this is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

Where the Y_i (the regressand) is the dependent variable, the X_{ip} (the regressors) are the independent variables, and ε_i is the error from the approximation of the model. The least-squares method of curve fitting that minimizes the residual (error estimate) sum of squares is used to estimate the parameters β_i for i=1,…,p. Letting n=1 for illustrative the above equation can be re-written as follows:

$$\epsilon = Y - (\beta_0 + \beta_1 X)$$

Now, since ε_i represents the vertical distance between the observed values and the model, it's desirable to minimize the sum of ε_i squared. This produces:

$$L \overset{\text{def}}{=} \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Taking partial derivatives with respect to each of the two parameters and setting equal to zero will minimize the residual sum of squares, thus:

$$\frac{\partial L}{\partial \beta_0} = 2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

and

$$\frac{\partial L}{\partial \beta_1} = 2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

Further algebra leads to the following two estimates for β_0 and β_1, denoted by $\widehat{\beta_1}$ and $\widehat{\beta_0}$

$$\widehat{\beta_1} = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n}$$

and

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

This can easily be generalized for n parameters.

2.2.2. Multiple Linear Regression

The simple linear regression model can also be generalized into n response variables, giving rise to the multiple regression model:

$$Y = X\beta + \varepsilon$$

Y is an nx1 vector containing the independent response variables, X is a matrix of the form:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

β is a (p+1) x 1 vector containing $\beta_o, \beta_1, \ldots, \beta_p$ and ε is a n x 1 column containing the residuals.

Given this generalized model, the specific model used for this paper will now be discussed. The vector Y which contains the independent response variables will consist of the actual disease incidence rate in a given district of Bihar, India. For example, the first entry in the vector Y would be the actual disease incidence in the district of Araria, India, which is 183.2. This represents 183.2 cases of the disease per 10,000 people. In this manner data for 20 other districts of Bihar are entered into the vector. The matrix X, aside from the first column of ones, will consist of 15 variables each as a column of X. For example the first variable, population density, is entered as the second column of X, and represents the population density of each district. Specifically, the population density of Araria, India is 751 population per kilometer squared, which would represent the $X(1, 1)$ entry. In the same manner as filling out the Y vector, the second column of X is filled out with population density data for each specific district. The matrix X is filled out in a similar manner will all the other variables chosen to be in the model. These variables are listed below:

1. Population density represented as population per kilometer squared.
2. Literacy rate for males is represented as a percentage of district male population that is literate.
3. Literacy rate for females is represented as a percentage of district female population that is literate.
4. The number of medical facilities per million people for each district.
5. Percentage of district population with no level of education.
6. Percentage of district population with graduate level of education.
7. Percentage of district population described as a main worker. These are workers employed in industry or factories in cities.
8. Percentage of district population described as a marginal worker. These are self-employed workers that sell goods in open markets in the cities.

9. Percentage of district population described as a non-worker. Non-workers are generally farmers or related to agriculture in some form.
10. Percentage of district households classified as permanent. These are homes constructed of brick.
11. Percentage of district households classified as semi-temporary. These are homes constructed of hard mud.
12. Percentage of district households classified as temporary. These are homes constructed of bamboo and hay.
13. Percentage of villages within a district that have access to bus services.
14. Total rainfall in millimeters per district.
15. Number of rainy days per year in each district.

The model can now be written as follows:

$$\begin{aligned} Y = \beta_0 &+ \beta_1(population\ density) + \beta_2(male\ literacy\ rate) + \beta_3(female\ literacy\ rate) \\ &+ \beta_4(medical\ facilities) + \beta_5(no\ level\ of\ education) \\ &+ \beta_6(graduate\ level\ of\ education) + \beta_7(main\ worker\ type) \\ &+ \beta_8(marginal\ worker\ type) + \beta_9(non-worker\ type) \\ &+ \beta_{10}(permanent\ house\ type) + \beta_{11}(semi-temporary\ house\ type) \\ &+ \beta_{12}(temporary\ house\ type) + \beta_{13}(bus\ services) + \beta_{14}(annual\ rainfall) \\ &+ \beta_{15}(rainy\ days) + \varepsilon \end{aligned}$$

The majority of the data (literacy rates, education level, worker types, house types, number of medical facilities and bus services) came from the 2001 census carried out by India. The remaining variables (population density and rainfall data) were collected from a statistical survey carried out in India in 1999. The incidence rate of the disease is an estimation of the actual number of cases, given the number of cases that are reported. The reason the estimation is used instead of the number of cases reported is because underreporting of the disease is

thought to be widespread. For instance, due to the physical scars the disease can induce on infected individuals they may be too shameful or embarrassed to seek medical attention, particularly women. VL is also commonly misdiagnosed in favor of other diseases, therefore decreasing the number of cases reported. The number of cases reported comes from government ran hospitals, and does not reflect the number of cases reported in privately owned health care facilities. The particular methods that calculate this estimation based on reported cases is beyond the scope of this paper, however, reference to Mubayi's et al. paper in the references should be read to gain a greater knowledge of this method [21].

2.2.3. Stepwise Regression Using Forward Selection

While regression analysis will be used to determine the final validity of the models, the forward-selection stepwise regression procedure will be used to select the most significant variables for the model, thereby reducing the number of variables. This is desirable given the large number of variables in the model to begin with, and will reduce the model to its most significant variables so that more specific conclusions can be drawn. The procedure begins by selecting the "best" one variable model as determined by the variable with the most significant F-test statistic. The F-test compares the ratio of the mean sum of squares for the model and the mean square error to that of a critical F value determined by the models degrees of freedom. Having a small probability of the calculated F statistic being greater than the critical F value is desirable, as this indicates a small probability of committing the type I error of rejecting the null hypothesis when it is in fact true. The process continues adding variables one at a time based upon having the most significant F statistic and being more significant than a predetermined level. If an added variable becomes less significant than the predetermined level in a following round, it is removed and the process reevaluated without that variable. The process ends when either every variable has been added, or there are no variables more significant than the predetermined level.

2.2.4. Box-Cox Transformation

The linear regression model used in this paper begins by assuming only first order terms will be included in the model. However, some of the variables could have quadratic (or other powers), exponential, or logarithmic relationships with respect to the incidence rate. Not capturing the best possible relationship between the data could significantly reduce the power of the model and any results that would be concluded. It is desired to inspect the data for any non-linear relationships that exist, and when discovered, to replace the first order data with the appropriate transformation that represents a better fit. Scatter plots of the variables against the incidence rate can be created, and the data visually inspected to search for any non-linear relationships that may exist. However, visually inspecting the data introduces a large amount of subjectivity as to which relationships exist, or visual inspection may not reveal obvious existing relationships. The Box-Cox method automatically suggests the appropriate transformation to use from the family of power transformations. The family of power transformations can be represented as follows, where Y' represents the new data to be used after the transformation has been made:

$$Y' = Y^{\lambda}$$

Commonly used members of the power transformation family include:

$$\lambda = 2 \qquad Y' = Y^2$$

$$\lambda = .5 \qquad Y' = \sqrt{Y}$$

$$\lambda = 0 \qquad Y' = log_e Y \quad (by\ definition)$$

$$\lambda = -.5 \qquad Y' = \frac{1}{\sqrt{Y}}$$

$$\lambda = -1 \qquad Y' = \frac{1}{Y}$$

Where λ represents the specific power transformation to be used. Notice this family of power transformations contains the square root transformation, log transformation and reciprocal

transformation for λ equal to .5, 0, and -1 respectively. The regression model can now be written where the independent variable is a member of the family of power transformations:

$$Y_i^{\lambda} = X\beta + \varepsilon$$

The Box-Cox method calculates the maximum likelihood estimate (MLE) of the parameter λ; the estimate is denoted by $\hat{\lambda}$ [20]. Many software packages can provide the MLE of λ, however there are alternative methods if this is not possible. One such procedure is to conduct a numerical search of likely values, such as λ = -2, λ = -1.75,…, λ = 1.75, λ = 2 [20]. Then for each value of λ, Y_i^{λ} values are standardized so the error term is not dependent on the λ [20]:

$$W_i = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right) & \lambda \neq 0 \\ K_2(log_e Y_i) & \lambda = 0 \end{cases}$$

where:

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Once W_i has been obtained for a given λ, regression against X can be ran and the SS_{error} obtained. When SS_{error} is minimized $\hat{\lambda}$ is the MLE of λ. It should also be noted that the Box-Cox method is used primarily to provide a guide to the selection of λ, and that overly precise results are not typically used. For instance, a λ of .6 would suggest the square root transformation since .6 is relatively close to .5 which represents the square root transformation [20].

2.3 Assumptions of the Model

The linear regression model makes several assumptions concerning the variables and error terms that will need to be examined before and after the final model is determined. These

assumptions provide parameter estimates that will be unbiased, efficient, and consistent and are listed as follows:

1. The sample is representative of the population. The sample in this case comes from 21 districts of the state of Bihar. Therefore, any results obtained will be applicable to Bihar, India, or regions of similar climate and socio-economic backgrounds.
2. The error is a random variable with mean of zero. Once analysis is ran, residual plots will be analyzed to verify this. A residual plot that appears to be equally spread with no definitive shape would be expected to have a mean of zero.
3. The independent variables are error free. If this is not met, there are other methods that can correct this problem. However, this assumption will be kept for this model.
4. The predictor variables are linearly independent. This is a difficult assumption to satisfy given the real world data used for the model. Obviously some variables may be closely related, such as literacy rate and level of education. However, in order to obtain a model some assumptions must be relaxed, and it should be observed this may be a limitation of the model and its results.
5. The errors are uncorrelated.
6. The variance of the errors is constant across all observations.

CHAPTER 3

RESULTS

3.1 Regression Results

The regression model was evaluated using MatLab and SAS software packages. The full model containing all 16 variables described in the previous section was initially run without applying transformations to any variables. This was done so that comparisons could be made, and thus the effectiveness of the transformations could be analyzed. This full model with no transformations produces a $\mathcal{R}^2$ statistic of .62 and a probability of .84 of the model being insignificant.

Scatter diagrams of each independent variable and the disease incidence rate (see Appendix A) were created and visually inspected before running the Box Cox transformation. With the small sample size of 21 it is visually difficult to see any signs of non-linearity in the data. Hence, the Box Cox transformation was used to detect any possible non-linearity trends. The results of the Box Cox transformation procedure are as follows:

1. An λ equal to 1.66 for population density suggests squaring this variable.
2. An λ equal to -.86 for female literacy rate suggests taking the reciprocal.
3. An λ equal to -.77 for no level of education suggests taking the reciprocal.
4. An λ equal to 1.75 for non-workers suggests squaring this variable.
5. An λ equal to -.08 for semi-permanent housing suggests taking the natural log.
6. An λ equal to .38 for temporary housing suggests taking the natural log.
7. An λ equal to -.09 for bus services suggests taking the natural log.
8. An λ equal to 1.48 for rain days suggests squaring the variable.

After making the appropriate transformation for the variables listed above, the full model with 16 variables using the transformed data was run again using regression analysis. This

transformed model produces a $\mathcal{R}^2$ statistic of .92 and a probability of .07 of the model being insignificant. The residual plots for this data can be found in Appendix B. SAS also produces individual parameter estimates for each β_i and tests the following hypothesis:

Null hypothesis: H_0: β_i=0

Alternative hypothesis H_1: $\beta_i \neq 0$

Choosing the variables significant at the level α=.05, the variables male literacy rate, number of medical facilities, marginal workers, permanent house type, temporary house type, and rainfall are selected to represent an additional model. This reduced model of six significant variables was run in SAS and produces a $\mathcal{R}^2$ statistic of .45 and a probability of .15 of the model being insignificant.

3.2 Stepwise Regression Results

Next, the full model using the transformed variables will be analyzed using forward selection stepwise regression. Two models were formed from this method by using two different criterions for selecting significant variables. The first model was stopped once no variable met a significance level of .4, while the second model was stopped once no variable met a significance level of .3. The reason two criterions were selected was to provide flexibility in the tradeoffs between the accuracy of the model and the number of variables in the model. The stopping criterion of .4 will allow more variables into the model thus making it more accurate. The stopping criterion of .3 will reduce the number of variables, and hence give a more precise picture of significant variables, at the cost of accuracy for the model. The first model selected the following variables in the following order:

1. Number of rainy days ($\mathcal{R}^2$ = .28, p-value .01)
2. Semi-temporary house type (.44, .01)
3. Permanent house type (.50, .01)
4. Literacy rate of males (.54, .01)
5. Marginal Worker (.59, .01)

6. Temporary house type (.63, .02)
7. Non-Worker (.65, .03)
8. Literacy rate of females (.69, .03)
9. Total yearly rainfall (.73, .03)
10. Graduate level of education (.79, .03)
11. Number of medical facilities per million people (.81, .04)

The second model selected the following variables in the following order:

1. Number of rainy days ($\mathcal{R}^2 =$.28, p-value .01)
2. Semi-temporary house type (.44, .01)
3. Permanent house type (.50, .01)
4. Literacy rate of males (.54, .01)
5. Marginal Worker (.59, .01)
6. Temporary house type (.63, .02)

Pearson's correlation coefficients were also calculated to determine what, if any, correlation exists between the variables and incidence rate. The greatest correlation coefficient (r=.53) was with the number of rainy days, followed by semi-temporary house type (r=-.49), and literacy rate of males (r=-.48).

CHAPTER 4

DISCUSSION AND FURTHER RESEARCH

The full model with no transformations was initially shown to be a poor choice for a model producing a relatively low $\mathcal{R}^2$ value of .62 and a high probability of the model being insignificant, .84. The linear regression model used assumes that the variables have a linear relationship with the dependent variable, and if this is not the case the model can be significantly compromised, as observed with the model not accounting for transformations. The model with Box Cox transformations was shown to be a significant improvement over the first model, producing a high $\mathcal{R}^2$ value, .92, and a relatively low probability of the model being insignificant, .07. However, this model still contains 16 variables and it's difficult to draw any conclusions for what specific variables are the most significant. Intuitively it was thought that by selecting the most significant variables from this model at the .05 level, that the model could be reduced so that more specific variable analysis could be completed. However, this was shown to be an incorrect method for this model, as the model containing the six variables significant at the .05 level was shown to produce a very low $\mathcal{R}^2$ value of .45 with a relatively high probability of the model being insignificant, .15. Forward selection stepwise regression was effective in eliminating variables, however, a relatively large amount of variables had to be retained in each of the two stepwise selected models to maintain statistically significant $\mathcal{R}^2$ values of .81 and .63. Individual correlation coefficients between the incidence rate and all the independent variables were computed, but few were of significance and only three were mentioned because of their relatively high value amongst the others.

From all this it is desired to discuss the most significant variables from all the various models, and make comments on how these variables could be used to have an effect on the disease incidence rate.

Rain was a significant part of the model, in particular the number of rainy days. Annual rainfall was one of the six variables that individually were significant to the full model, while the number of rainy days was the first variable to be included in the stepwise forward selection procedure and also had the largest correlation coefficient. This is not too surprising as previous research has shown rain in particular forms plays a significant role in incidence rate of vector borne diseases similar to VL. This suggests the number of rainy days, aside from strictly total amounts of rain, should be considered as a possible indicator of where VL outbreaks may be at greater risk. An increase in the number of rainy days allows for an increase in the amount of time and places for sand flies to breed and their young to reach a mature age. Districts that experience an elevated number of rainy days could put in place increased preventative measures to help decrease the number of places sandflies live and breed. Possible examples include increased use of lime in households to decrease the number of crevasses where sand flies lay eggs, or increased destruction of non-essential standing waters that increases sandfly populations.

House type in general was also a significant portion of the model, as all three house types were included in the six variable stepwise reduced model with a significant $\mathcal{R}^2$ value of .82. Two of the house types, permanent and temporary, individually had significant values at the .05 level for the full model, and semi-temporary house type had the second largest correlation coefficient. This seems logical because many people spend the majority of their time at home, and while there remain relatively inactive giving sand flies an increased chance of making a bite and transmitting the disease.

Literacy rate of males was also a significant variable, since it was included in the six variable stepwise reduced model, had the third largest correlation coefficient, and also a significant value at the .05 level for the full model. This implies increasing the number of literate males could help decrease the disease incidence rate. This seems appropriate because as the number of literates increases there will be more people aware of how the disease is spread,

what the symptoms of the disease are, and measures that can be taken to reduce risk of contracting the disease. Moreover, male members are usually the head of a household and their direction can impact the life style of the family.

Non-workers and marginal workers are included in the stepwise regression model, and marginal workers were individually significant in the full model. This indicates worker type may also be significant to the disease incidence rate. This is logical since marginal and non-workers spend more time where the vector is present than main workers, where they are more likely to contract the disease. Possible measures that could be taken include providing disease information to workers in these specific classes, or performing vector reduction procedures such as insecticide spraying of work areas before workers begin a particular project.

Some additional notes, literacy rate of females was included in the 11 variable stepwise reduced model, and since it is most likely just as easy to improve the literacy rate of females and males together, rather than males individually, it is best to think of improving the overall literacy rate of a population as a means of possibly reducing disease incidence rate.

While the number of medical facilities was individually significant in the full model, it appeared as the last variable added in the stepwise regression with 11 variables, and also had a relatively low correlation coefficient of -.14. This suggests the number of medical facilities may not have a very strong relationship with the disease incidence rate relative to other variables considered in this study, and given the relatively large cost of increasing the number of medical facilities; this may not be the best use of government funds to decrease VL incidence rate.

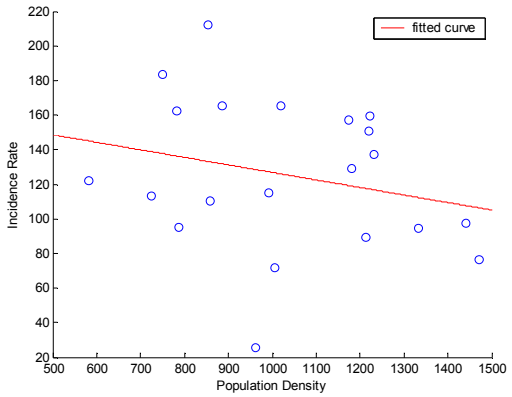
This paper began with the hypothesis that the disease may be strongly related to socio-economic variables, based on the fact that the disease is prevalent in third world countries but nearly non-exist in developed countries, despite the fact of similar climate characteristics. The regression model essentially showed that a significant model could be produced, and socio-economic variables such as literacy rate, house type, and worker type could be significantly

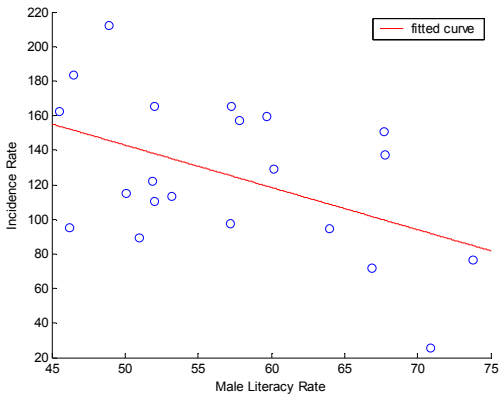
related to disease incidence rate. Government funds spent towards the goals of increasing literacy rates of the population so as to understand the disease, combined with improving housing conditions and working condition for non-main workers, could decrease the disease incidence rate.

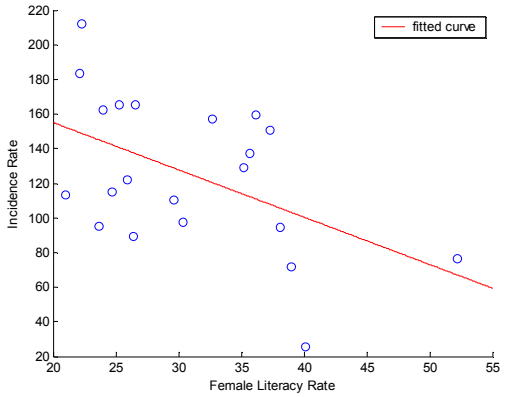
The limitations of the model should be clearly understood so as to not overestimate the importance of the results. The main limitation is associated with the small sample size of 21 observations, the districts of Bihar, India. With a small sample it may be inappropriate to use the central limit theorem to assume the parameter estimates are normally distributed. It's not possible to overcome the limitation of a small sample is this case, because there are no additional districts in Bihar to consider given the question at hand. Broader areas where the disease exits, such as Bangladesh, Nepal, Sudan, and Brazil could be included to increase the sample size, however this would be answering a different question than the one posed in this paper. Another limitation is that spatial data was used instead of time series data. For instance, rainfall can vary from year to year; global economic situations can affect what type of work people conduct and also affect government funds for services such as education, medical facilities, and bus services. A study of Bihar with time-series spatial data when compared to this study could be beneficial in a more complete understanding of the disease situation. Under reporting of the disease incidence is also a limitation, requiring the use of an estimation of this parameter. Further research suggestions should focus on improving these limitations.

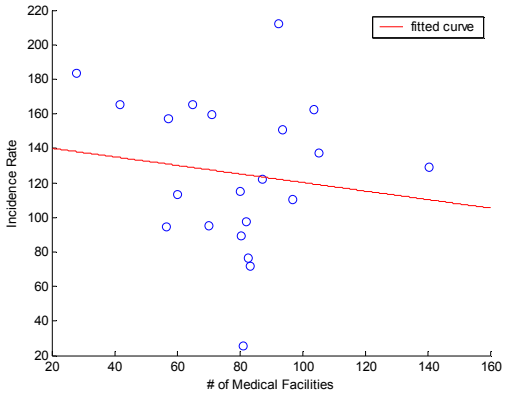
APPENDIX A

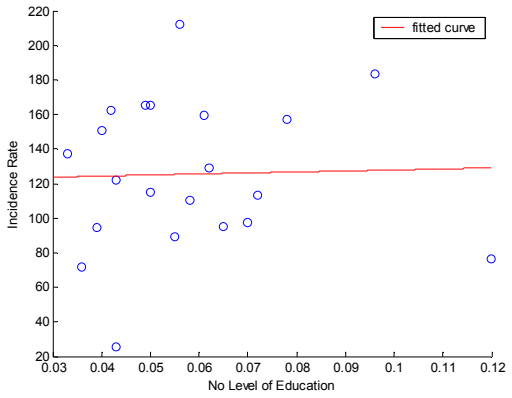
SCATTER DIAGRAMS OF DATA

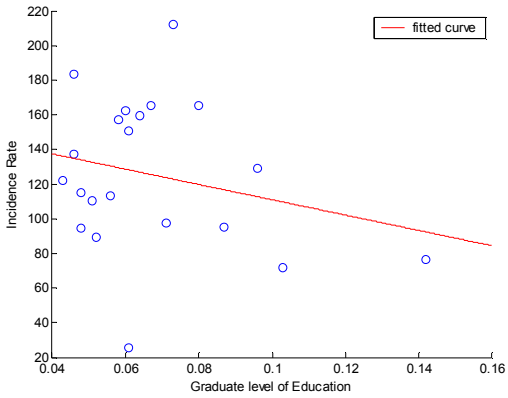

fitted curve
Incidence Rate
Population Density

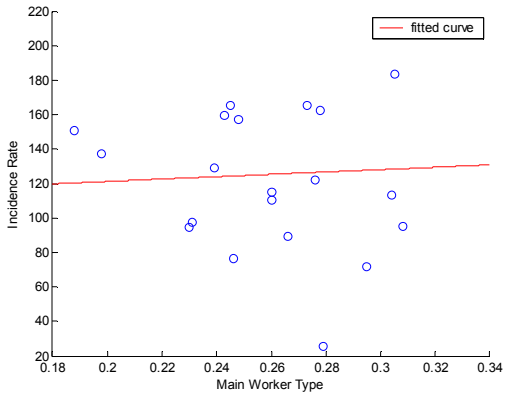

fitted curve
Incidence Rate
Male Literacy Rate

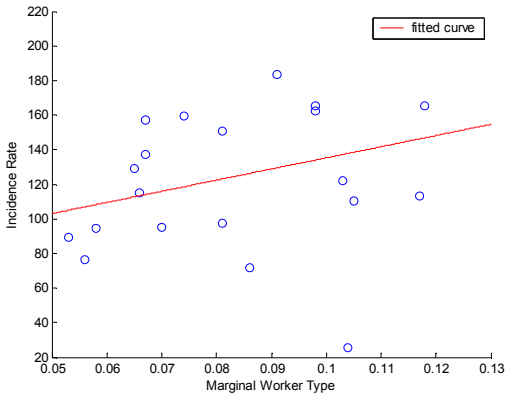

fitted curve
Incidence Rate
Female Literacy Rate

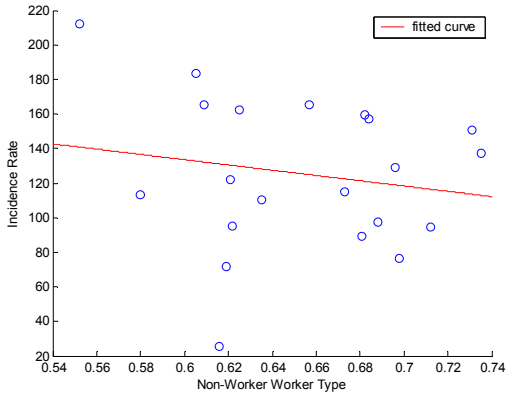

fitted curve
Incidence Rate
# of Medical Facilities

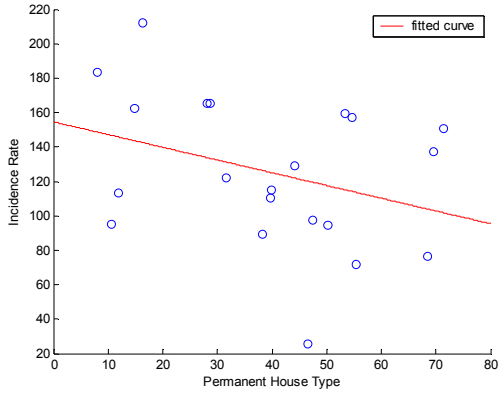

fitted curve
Incidence Rate
No Level of Education

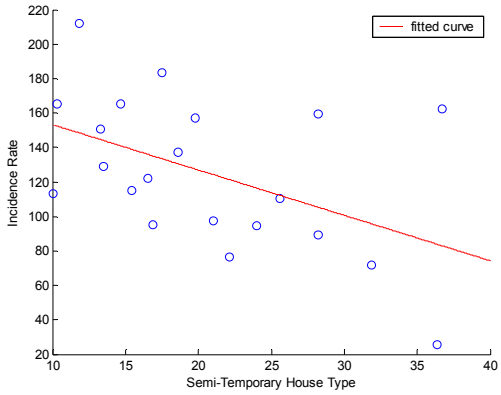

fitted curve
Incidence Rate
Graduate level of Education

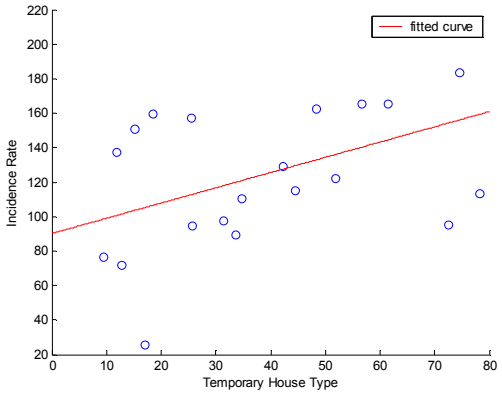

fitted curve
Incidence Rate
Main Worker Type

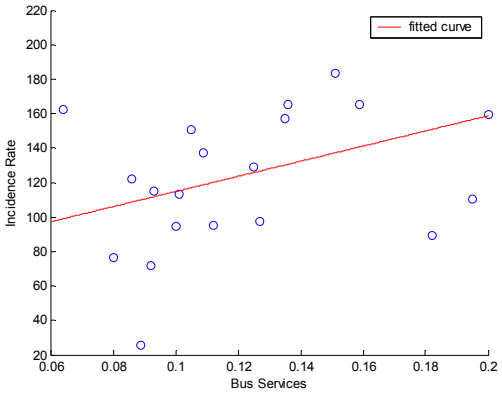

fitted curve
Incidence Rate
Marginal Worker Type

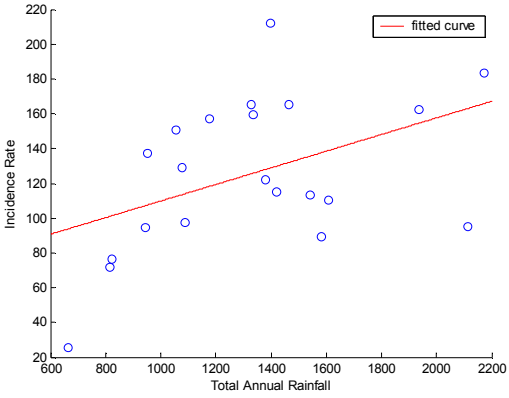

fitted curve
Incidence Rate
Non-Worker Worker Type

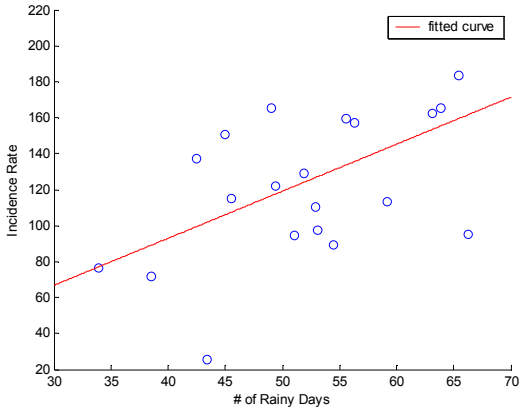

fitted curve
Incidence Rate
Permanent House Type


fitted curve
Incidence Rate
Semi-Temporary House Type


fitted curve
Incidence Rate
Temporary House Type

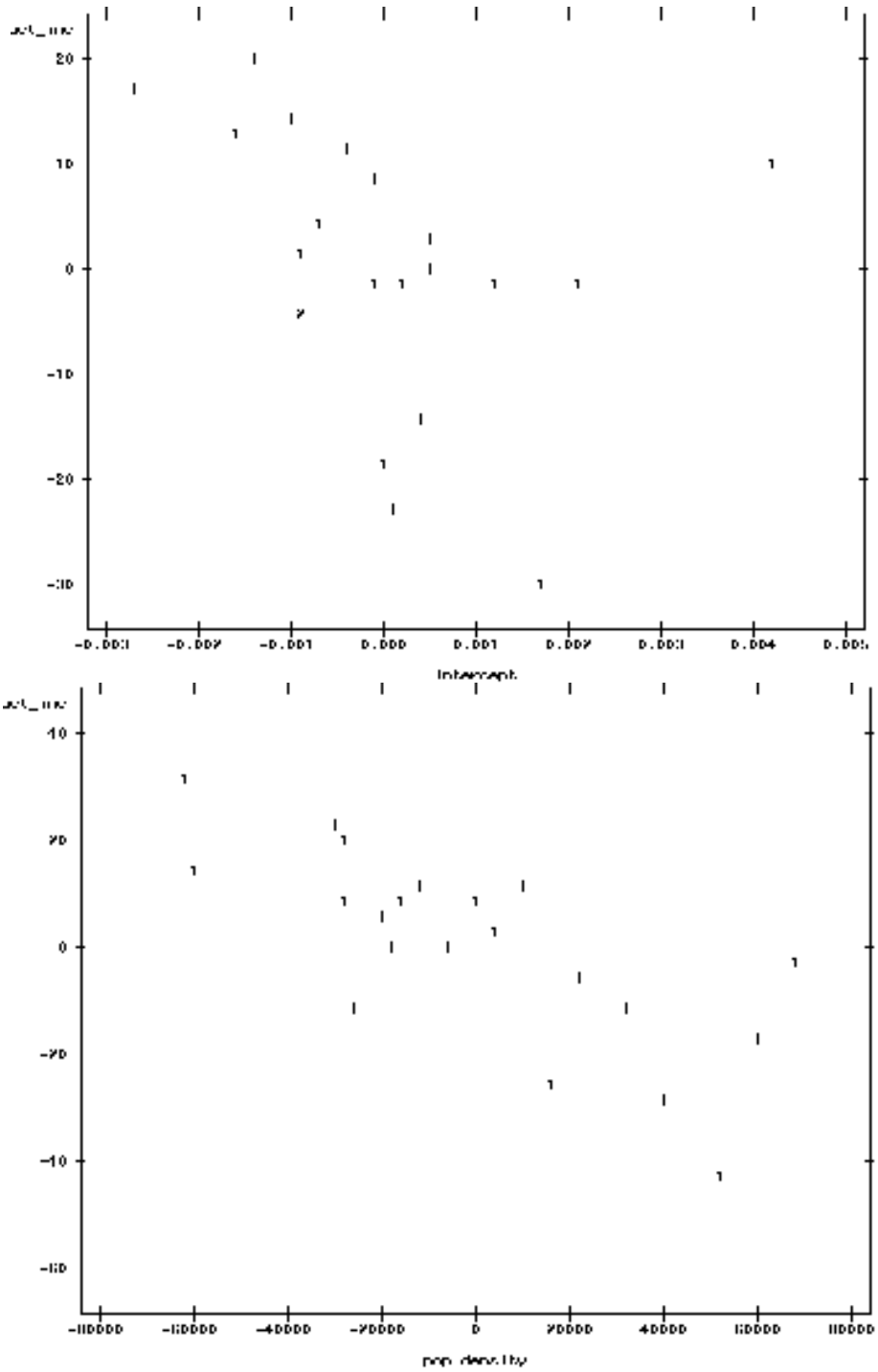

fitted curve
Incidence Rate
Bus Services
220
200
180
160
140
120
100
80
60
40
20
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2

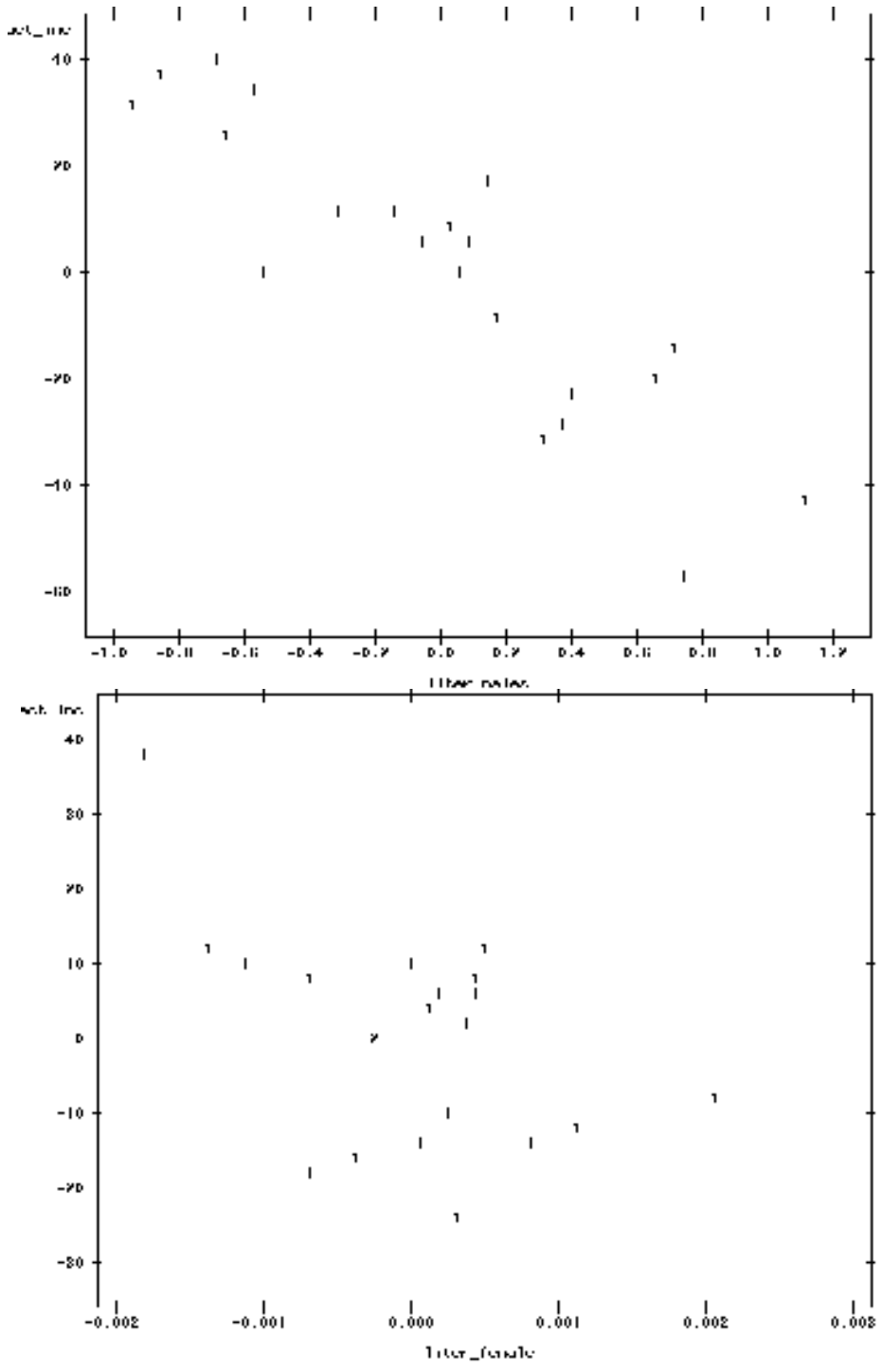

fitted curve
Incidence Rate
Total Annual Rainfall
220
200
180
160
140
120
100
80
60
40
20
600
800
1000
1200
1400
1600
1800
2000
2200

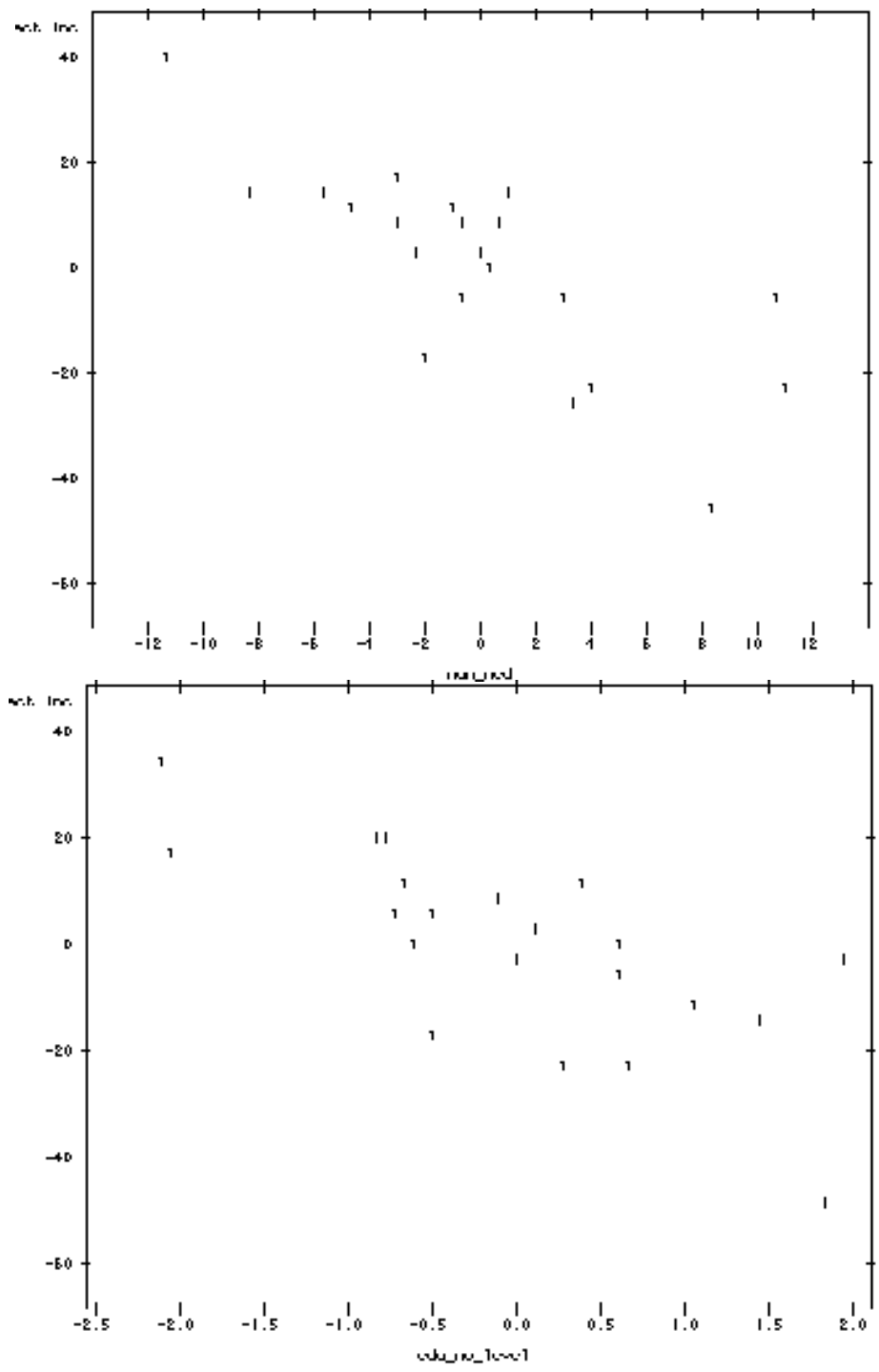

fitted curve
Incidence Rate
# of Rainy Days
220
200
180
160
140
120
100
80
60
40
20
30
35
40
45
50
55
60
65
70

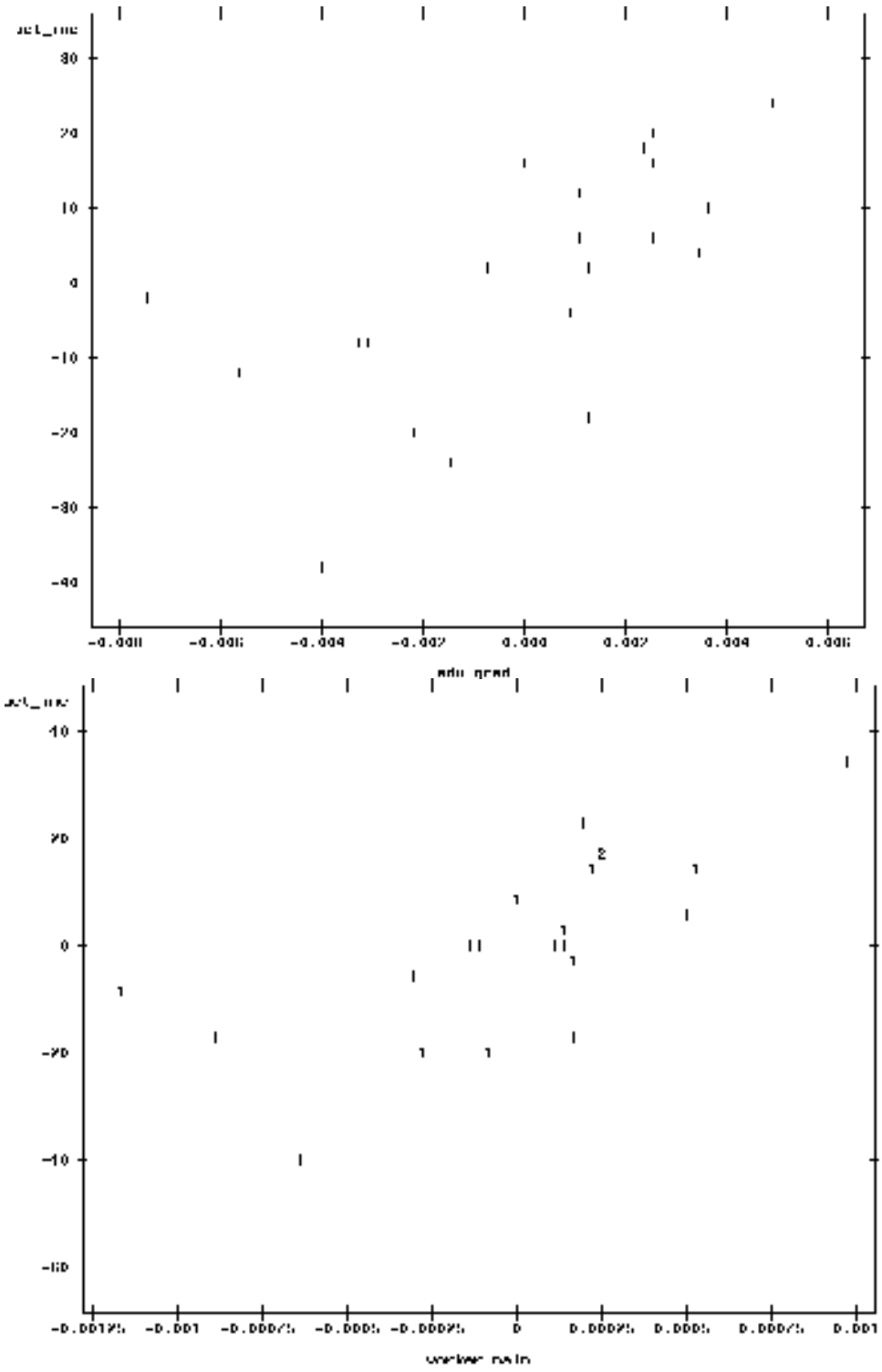
APPENDIX B

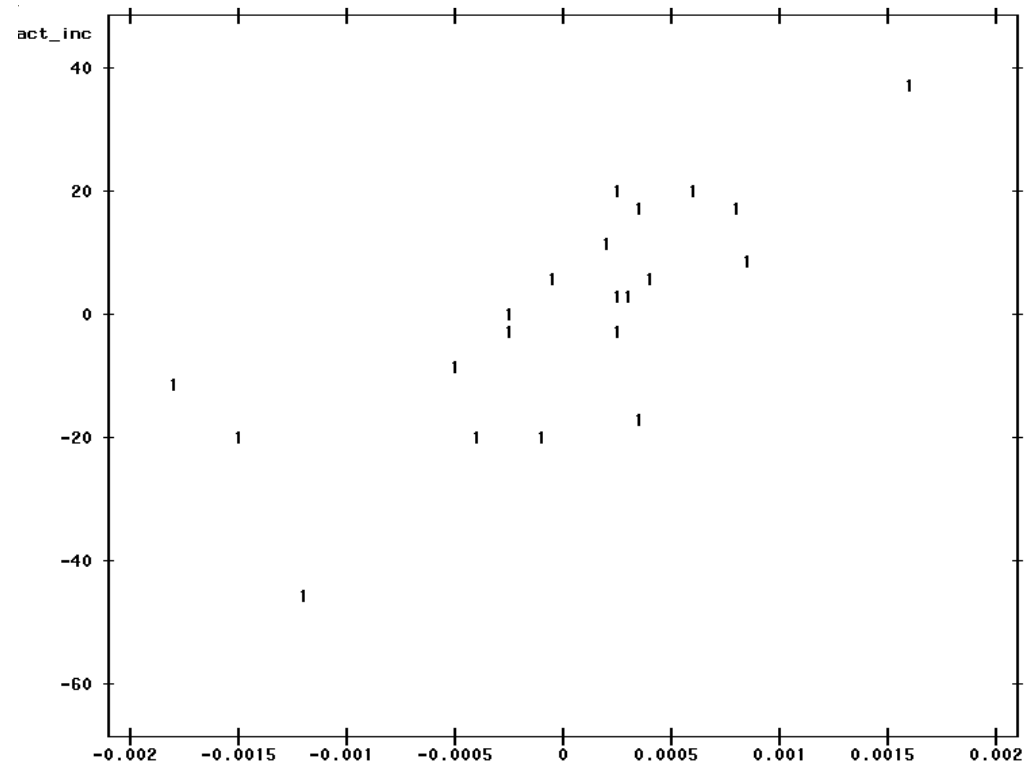
RESIDUAL PLOTS

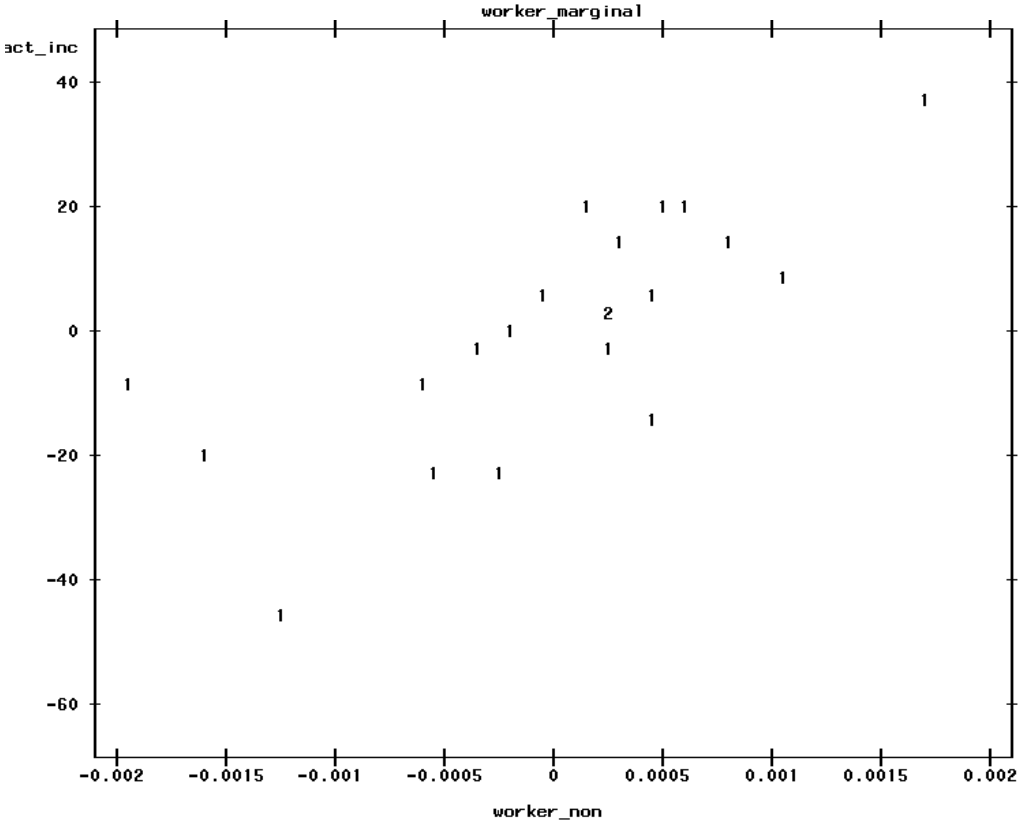


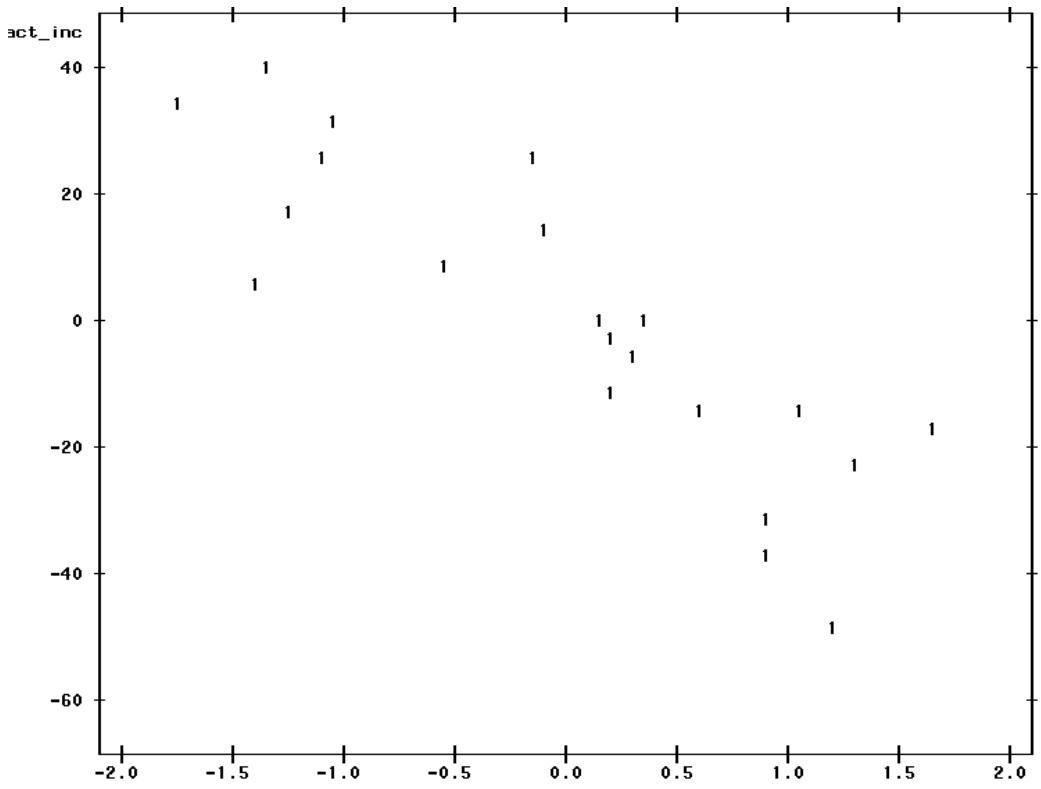


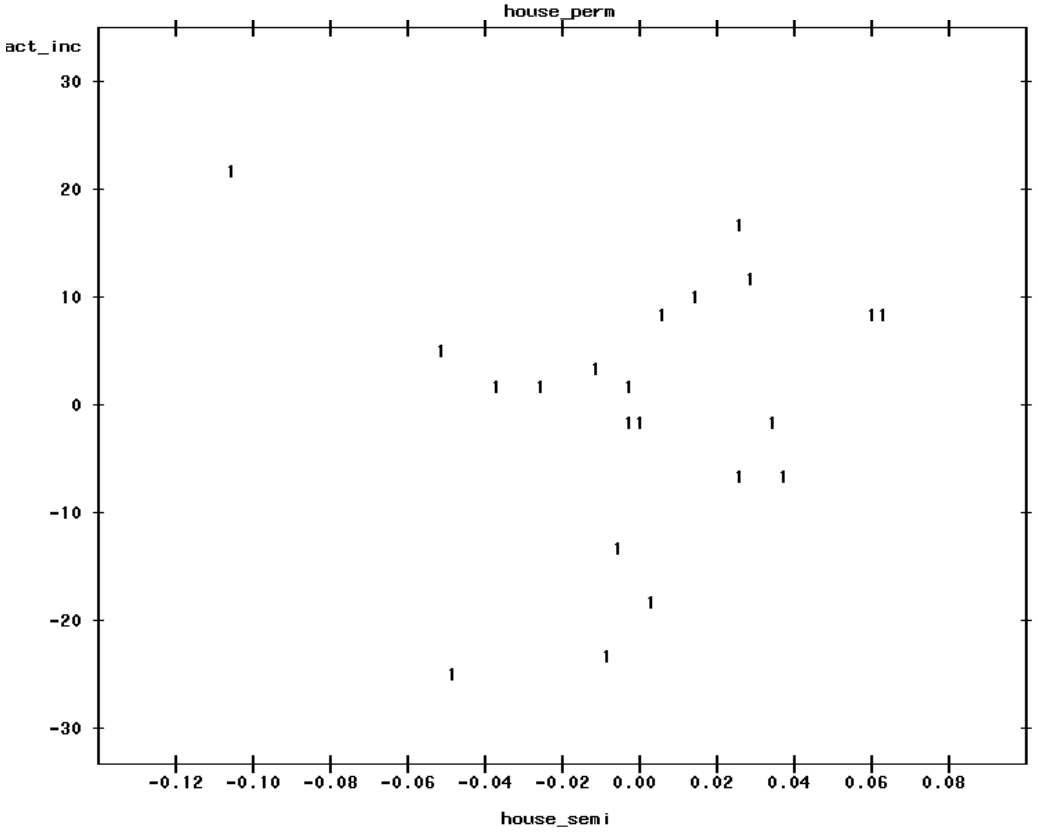

sch. inc.
40
20
0
-20
-40
-60
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
sch. inc.
40
20
0
-20
-40
-60
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

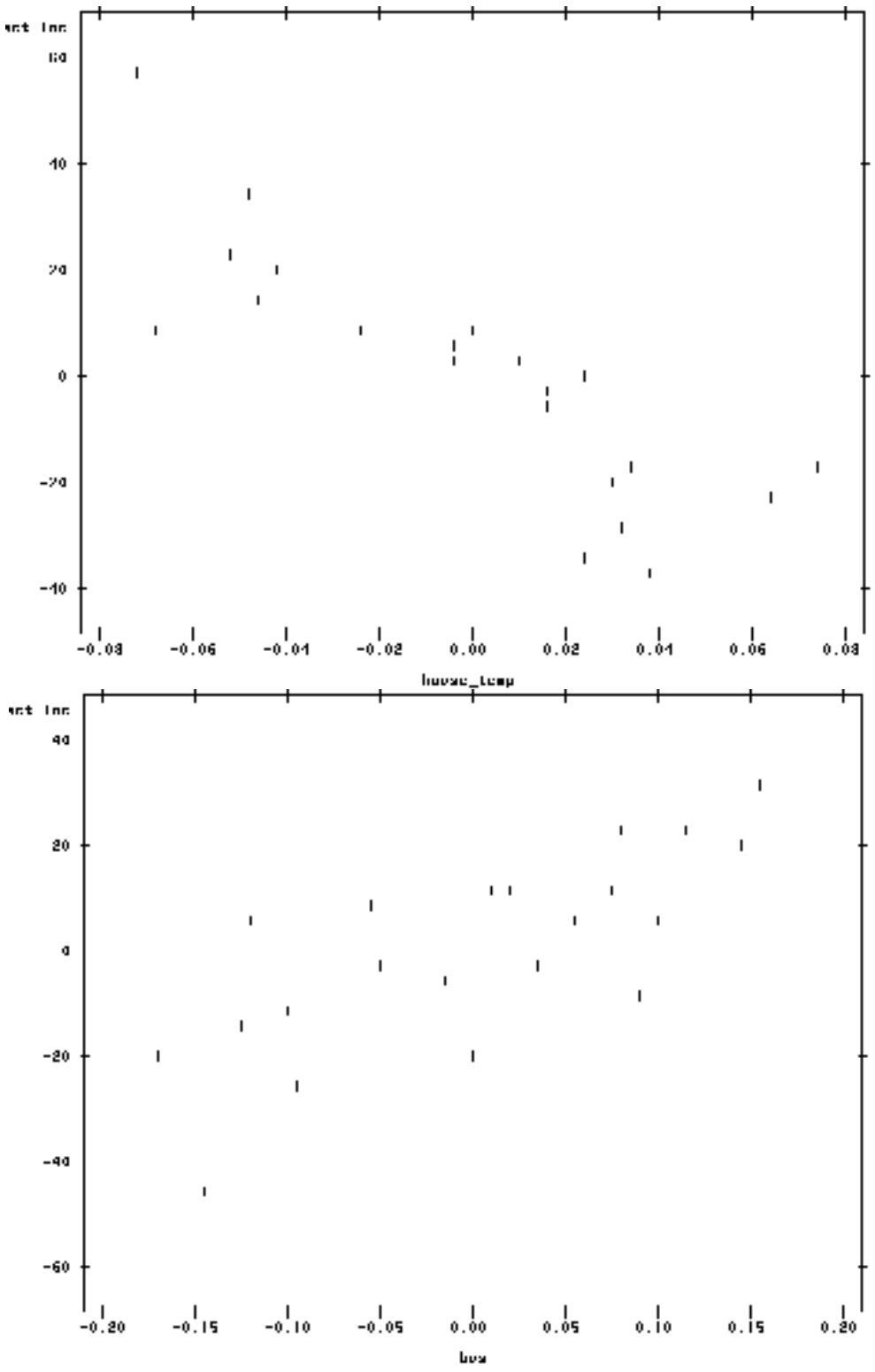


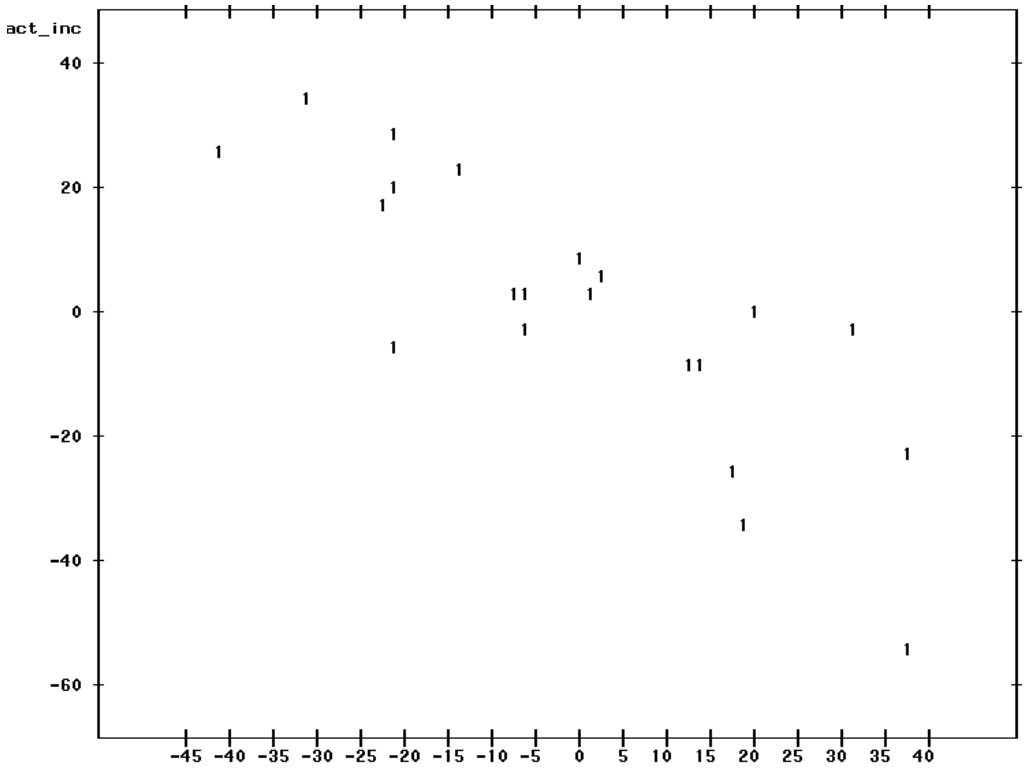

act_inc
40
20
0
-20
-40
-60
-0.002
-0.0015
-0.001
-0.0005
0
0.0005
0.001
0.0015
0.002
worker_marginal

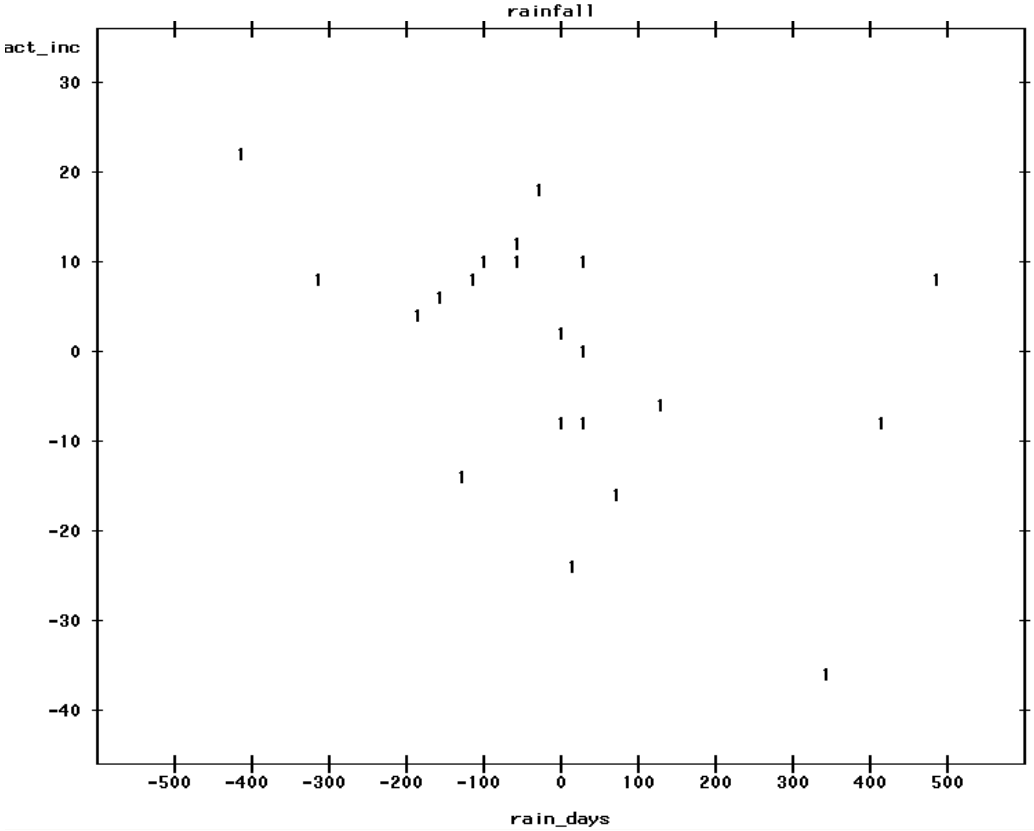

act_inc
40
20
0
-20
-40
-60
-0.002
-0.0015
-0.001
-0.0005
0
0.0005
0.001
0.0015
0.002
worker_non

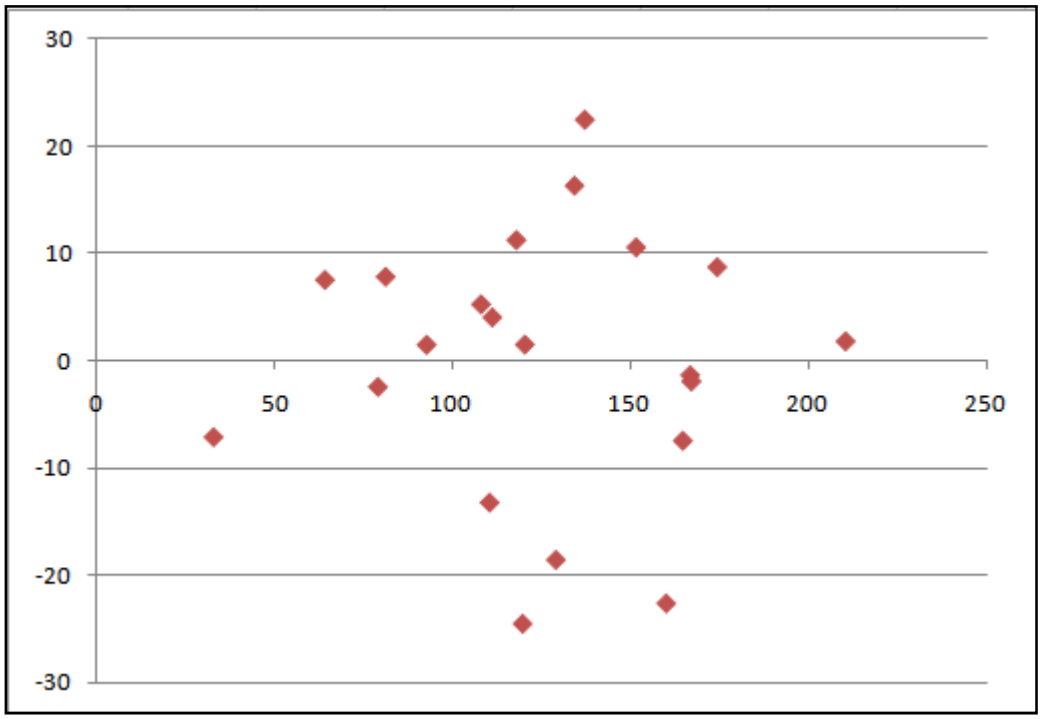

act_inc
40
20
0
-20
-40
-60
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
house_perm


act_inc
30
20
10
0
-10
-20
-30
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
house_semi


60
40
20
0
-20
-40
-0.08
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
40
20
0
-20
-40
-60
-0.20
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20


act_inc
40
20
0
-20
-40
-60
-45 -40 -35 -30 -25 -20 -15 -10 -5 0 5 10 15 20 25 30 35 40
rainfall


act_inc
30
20
10
0
-10
-20
-30
-40
-500 -400 -300 -200 -100 0 100 200 300 400 500
rain_days



Plot of the predicted values $\hat{y}$ (x-axis) against the residuals $\hat{\epsilon}$ (y-axis). It is ideal for the residuals to form a horizontal band, indicating equal variances and no dependence on $\hat{y}$.

REFERENCES

1. Myler P; Fasel N (editors). (2008). Leishmania: After The Genome. Caister Academic Press. ISBN 978-1-904455-28-8
2. A Small Charity Takes the Reins in Fighting a Neglected Disease, New York Times, July 31, 2006
3. R. Russo, F. Laguna, R. Lopez-Velez, F. J. Medrano, E. Rosenthal, B. Cacopardo, L Nigro. (2003) VL in those infected with HIV: clinical aspects and other opportunistic infections. Annals of Tropical Medicine & Parasitology, Vol. 97, Supplement No. 1, S99-S105.
4. Bern C, Hightower AW, Chowdhury R, Ali M, Amann J, Wagatsuma Y, et al. (May 2005) Risk factors for kala-azar in Bangladesh. Emerg Infect Dis. Available from http://www.cdc.gov/ncidod/EID/vol11no05/04-0718.htm
5. Institute for OneWorld Health, http://www.oneworldhealth.org/img/pdfdownloads/Leishmaniasis%20Fact%20Sheet.pdf
6. Luis Chaves, Justin Cohen, Mercedes Pascual, Mark Wilson (2008) Social exclusion modifies climate and deforestation impacts on a vector-borne disease. PLoS
7. Richard Lewontin & Richard Levins. "Schmalhausen's Law." Capitalism, Nature, Socialism, 11(4) (2000): 103–108
8. http://medical-dictionary.thefreedictionary.com/Leishmaniasis
9. Joseph Keating (2000) An investigation into the cyclical incidence of Dengue Fever. Social Science & Medicine 53 (2001) 1587-1597.
10. Patz, J.A., Epstein, P., Burke, T., & Balbus, J. (1996) Global climate change and emerging infectious disease. Journal of American Medical Association, 275(3), 2217-2223.

11. Richard Johnson & Dean Wichern (2007) Applied Multivariate Statistical Analysis 6E ISBN: 0-13-187715-1
12. Gerardo Chowell & Fabio Sanchez (2006) Climate-Based Descriptive Models of Dengue Fever: The 2002 Epidemic in Colima, Mexico. Journal of Environmental Health, Vol. 68 Number 10.
13. Mourya, D.T., Yadav, P., & Mishra, A.C. (2004) Effect of temperature stress on immature stages and susceptibility of *Aedes Aegypti* mosquitoes to chikungunya virus. American Journal Tropical Medicine and Hygiene, 70(4), 346-350.
14. Depradine, C.A. & Lovell, E. (2004) Climatological variables and the incidence of Dengue Fever in Barbados. International Journal of Environmental Health Research, 14, 429-441.
15. Russell FE. (2001) Toxic effects of terrestrial animal venoms and poisons. Toxicology-The basic science of poisons. 6E New York: McGraw-Hill pp 945-694
16. Dehesa-Davila M, Possani LD. (1994) Scorpionism and serotherapy in Mexico. Toxicon 32(9) 1015-1018
17. G. Chowell, J.M. Hyman, P. Diaz Duenas, & N.W. Hengartner. (2005) Predicting scorpion sting incidence in an endemic region using climatological variables. International Journal of Environmental Health Research 15(6): 425-435.
18. "National Human Development Report" (PDF). Planning Commission of the Union Government. 2001. http://planningcommission.nic.in/reports/genrep/nhdrep/nhdch1.pdf. Retrieved on 2008-08-10.
19. Goswami, Urmi A (2008-06-17). "Biharis get work at home, bashers realise their worth". The Economic Times. http://economictimes.indiatimes.com/News/PoliticsNation/Biharis_get_work_at_home_bashers_realise_their_worth/articleshow/3135697.cms. Retrieved on 2008-06-17.

20. Kutner, Nachtsheim, Neter, Li (2005) Applied Linear Statistical Models 5E

21. Mubayi, Chowell, Castillo-Chavez, Kribs-Zaleta, Siddiqui, Kumar, Cas (2008) Transmission Dynamics and Underreporting of Kala-Azar in the Indian State of Bihar.

BIOGRAPHICAL INFORMATION

Darren L. Sheets was born in Northfield, MN in 1983. He earned his B.S. degree in Mathematics from the University of Wisconsin – River Falls in 2005 and his M.S. degree in Mathematics (Statics Major) from the University of Texas at Arlington in 2009. He has worked in the insurance industry as a Corporate Financial Systems Analyst, as well as performing teaching assistantship duties in the Information Systems and Operations Management department at the University of Texas at Arlington. He will begin working towards his M.S. degree in Economics in the summer of 2009, with the intent of pursuing a PhD in Economics upon successful completion.