International Journal of Advanced Mathematics
Vol. 4, No. 1, April 2017, pp. 1-13
Available Online at http://acascipub.com/Journals.php

Research article

Optimally Rectifying Non-homogenous Poisson Probability Count Outcome Variation Forecasts in an Endemic County Level Syphilis Model

Grant Johnson, Brock Graham, Benjamin G. Jacob.

Department of Global Health, University of South Florida.
University of South Florida, 12901 Bruce B. Downs Boulevard, Tampa, FL 33612.

Corresponding Author: mmchild@health.usf.edu

Abstract:

Syphilis is an ongoing problem on the world stage, and thus the methods for analyzing its' spread are always being challenged. This investigation set out to assess the ongoing syphilis epidemic in Hillsborough County, Florida in hopes of determining critical correlates central to the increasing incidence. Instead, we discovered that the regression paradigms commonly utilized by researchers may be inherently flawed when applied to certain subsets of data. Utilizing logistic, Poisson and negative binomial regressions, we tested the model fit diagnostics against a 5-year array of incidence data, and found that in the case of syphilis, the models did not fit. Negative binomial regression with a heterogeneous mean did correct for overdispersion amongst the count variable findings, but further investigation is warranted into more certain and reliable methods of assessing syphilis outbreak data.

Keywords: Syphilis, Poisson, Negative binomial regression, overdispersion.

1.0 Introduction

Unfortunately for all of humankind, syphilisstill remains a problem. Known as "The Great Pretender", this disease is notorious for masking signs and symptoms within the body, causing many of the people that are infected with it to rest in false security while their body retains the infection. The symptomology and biology are not the only factors

of this bacterial infection that have led to frustration over the years however. For decades, researchers have been using a wide variety of frequentist, non-frequentist, geospatial and temporospatial methods to attempt to track, predict and defeat the outbreaks of this disease (Jacob et al, 2005, Griffith, 2005). Despite this multi-fronted attack, there is still not a consistent mathematical technique that can be relied upon for analyzing or predicting the mannerisms of syphilis infections.

One of the major goals of any epidemiologic endeavor focuses on a particular illness is to identify the major demographic covariates that can be used to determine what populations may be at the highest risk of becoming infected. This is no different in the history of syphilis. What does stand out however is the fact that other than MSM and HIV co-infection, very little demographic covariate information has been consistently and repetitively discovered for syphilis(CDC, 2006, Chen, 2002). This problem is the apparent statistical inconsistency for optimally quantitating demographic parameter estimators of the disease. The focus ofthis investigation is to attempt to rectify uncertainties in a regression model for optimally targeting demographic covariates associated with syphilis at the county level.

One of the most commonly employed statistical sampling methods in epidemiologic analysis is the frequentist method, known also as regression or linear analyses. This method of mathematical hypothesis testing employed to search for a causative correlation relationship for identifyingindependent variables and any number of dependent variables for any mathematical model, relies on a handful of assumptions being satisfied to ensure validity of the result (Pregibon, 1981). Shahmanesh et al. (2000) investigated the hypothesis that core populations were a reliable assumption for the spread of particular STIs in an urban setting, in this case Birmingham, England. This retrospective cross-sectional study employed a forward stepwise logistic regression model to assess correlation between patients with chlamydia and patients with gonorrhea based on the variables of ethnicity, age, sex and estimator for socioeconomic status using a super profile analysis, The findings of this article revealed a higher risk of infection for African Caribbean males under the age of 20 that lived in neighborhoods of similar sociodemographic indicators. Although finding of a sociodemographic profile for high-risk individuals suggest that testing an area for a targeted intervention on the part of the local health department could reduce the incidence of syphilis, the investigators of Shahmanesh et al. (2000) also found that when they tested the external validity of their findings, the results were not consistent in the neighboring counties.

Inconsistency in external validity is a common finding amongst researchers in not only syphilis outbreaks, but also in sexually transmitted infections such as chlamydia, gonorrhea and HIV. Johnson et al. (2016) evaluated frequentist models in STI analyses in South Africa against a microsimulation networking paradigm. They discovered that even after complex recalibration for various local prevalence and incidence trend statistics, the frequentist analysis consistently over-estimated levels of predictive prevalence. The investigative team did not test whether the misspecifications were due to violations of any stochastic, non-Gaussian assumptions of the frequency-based, explanative models, but found instead that the error only occurred when applied to STI's. Further, their findings suggested that the mathematical error may have been due to biological assumptions being violated which the mathematical algorithms were not able to account for.

Non-Gaussian linear and logistic regression models, as all other mathematical and statistical models, operate within a set series of predetermined assumptions. These assumptions must be followed in the design of the model, for any violations can skew the produced results, and invalidate any findings that the model produces. The assumptions for frequentist models are rather straightforward; total independence of the non-response, demographic covariates, no multicollinearity amongst independent variables, homoskedacity, and normally distributed uncertainties amongst the predictive variables. While a normal, or Gaussian, distribution is preferred amongst the observations, it is not a requirement for any model, as link functions and exponential family models can be utilized to overcome non-Gaussian distribution(Pregibon, D. (1981).

Homoskedacity (i.e., common variance) is one of the primary assumptions that a modeler makes when selecting covariates for a study of any kind, the other being independence of the variables. It is therefore vital that a syphilis researcher does not automatically skip over consideration of this assumption whilst designing anendemic

model. Homoskedacity is the assumption that the variance of all of the selected independent covariates is equivalent. This equivalence is a necessity for optimally regressively quantitating reliable correlation values, however it is often close to impossible to find in syphilis sampled datasets since they frequently deal withsubjective human behavioral covariates. Socially dependent covariates may violate regression covariates(Pregibon, D. 1981).Therefore, it is often up to the syphilis researcher to utilize specific modeling techniques to either eliminate the outliers, or artificially generate a normalized variance through log-transformation methodologies of a negative binomial framework.A negative binomial regression with a non-homogenous gamma distributed mean can compensate for violations of homoskedascity (Haight 1967). Forward and reverse stepwise regressions are one such method, in which an algorithm will individually remove and/or replace single covariates from the model and report the alterations, if any, of the pseudo-R^2 correlation value. If there is no changes of this explanatory diagnostic value, then the demographic syphilis sampled covariates are homoscedastic, however any changes, especially large ones, can indicate heteroskedacity, and the model designer must then decide whether or not to remove the selected variables, or control for them in another manner.

In a non-spatial, syphilis-related, frequentist analysis, the tendency of demographic covariates may be found more often in one setting than another. These tendencies may be easily misjudged by the frequentist model as spurious correlations, which could skew the findings toward or away from the null hypothesis. A multivariate linear regression (with confounding factors for adjustment) may explore whether associations, of syphilis sampled demographic data has normal distribution of residuals and fulfils the other assumptions of regression analysis.

Multicollinearity (incidental lack of independence of the covariates, in regression space) is a phenomenon that is easily missed during model design, leading to alterations of pseudo-R^2 values.If two or more covariate numerical values are dependent upon one another misspecifications may be forecasted (Pregibon, D. (1981). This assumption is vital to test for correlation thus in an endemic model. Since assumed syphilis –related explanatory independent variables are in fact dependent on one another, when placed together in a model, they can artificially inflate the R^2 correlation value, thereby skewing the findings away from the null hypothesis.

Following this line of thought, and after an extensive literature analysis, we noticed that many researchers found mildly differing results when utilizing regression methods to analyze and predict syphilis outbreaks. We have hypothesized that the reason for this may indeed lie in a violation of assumptions being made in the design of these regressive models. While attempting to utilize frequentist analytic methods to determine critical correlates in a syphilis outbreak in Hillsborough County, Florida (see Graph 1.1) we noticed that the model fit diagnostics were far off the expected range. Experimentation with the data showed us that the issue was not in our methods or model design, but that there may be inherent error and violation of, or failure to properly account for, certain regression assumptions when attempting to apply frequentist methodologies to syphilis sampled demographic datasets. Thus, our research objectives in this comparative syphilis-related regression analysiswere; 1) to compare pseudo-R^2 values rendered from a dichotomous logistic bivariate model and a Poisson probability paradigm to determine non-robustness and 2) utilize diagnostic, residual, forecastsrendered from a negative binomial regression analysis employing a non-homogenous gamma-distributed mean to compensate for violations of assumptions (i.e., extreme outliers, over-dispersion). Although this model alludes to syphilis endemicity at the county level, we envision employing this model framework for other sexually transmitted diseases.

2.0 Methods:

All data was obtained from the Florida Department of Health in Hillsborough County, and all analysis was performed utilizing SAS Studio v9.04. 2116 cases were documented between the years 2010 and 2014, ranging across 75 zip codes. The data was analyzed and cleaned, repairing errors in recording, and restricting the zip codes to the 65 that make up Hillsborough County. These zip codes were also broken into four distinct geographic zones of the county; northwest (NW), northeast (NE), southwest (SW), and southeast (SE).

Forward and reverse stepwise regression analysis of potential demographic covariate candidates was performed utilizing the REG procedure. This method identified which of the variables held enough of an observable

linear or logistic relationship with the response variable to be considered for correlation analysis. The stepwise analyses also tested for and ruled out any risk of multicollinearity amongst the sampled variables.

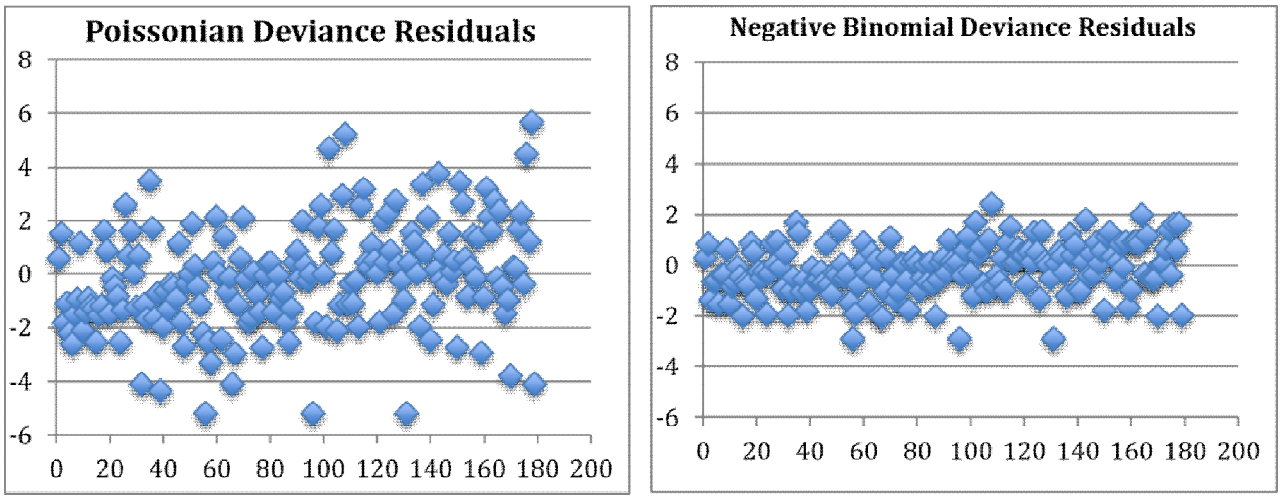
A bivariate, logistic regression was first employed to examine the possibility of correlation existing between the sampled covariates. A dichotomous variable (DV) was generated as the response variable by separating the cases into duration of infection categories, split at the one-year mark. This mark was determined by the disease stage variable; all cases were classified primary, secondary or early latent which were assumed to be infections lasting less than one year, whereas all cases labeled late latent were assumed to have lasted greater than one year. This method was also employed by Gesink et al (2006) to analyze covariates for a Bayesian analysis of a syphilis outbreak.

In the stepwise analysis (results not shown) the age, sex, race, and quadrant variables were set against the response variable for analysis. A Poisson regression probability analysis was conducted to analyze covariates, utilizing the GENMOD procedure in SAS 9.04 with a logistic link. The response variable (N) was generated as a count variable representing number of cases falling into each of the potential variable categories. The independent variables analyzed were quadrant, race (Caucasian vs. African American), ethnicity (Hispanic vs. non-Hispanic) and sex (male vs. female). To generate the response variable "N", our data was first sorted by quadrant, year, race, ethnicity and sex by means of the SORT procedure. The resulting data was then analyzed via the MEANS procedure with respect to the sorting priority, and the data output generated 179 unique individual count variables to correspond to the analytic criteria. In an attempt to account for the known assumption violations of the count variable model and for identifying underdispersion of the model data, a negative binomial regression was also performed on the aforementioned data.

3.0 Results:

As the focus of this analysis was on the appropriate application of frequency analyses in reference to endemic, syphilis sampled, demographic data, we assessed the fit diagnostics for our results, rather than the outputs of the covariate analyses to determine validity of findings. The logistic regression returned an R^2 value of 0.1041, as can be seen in Figure 3.1.

Figure 3.1 Regression model fit diagnostics for syphilis sampled, demographic data for Hillsborough County



The Poisson and negative binomial regression analyses were analyzed for model fit by measuring deviance over degrees of freedom, or V/DF, to assess dispersion as in Haight, (1967). The V/DF in the Poisson paradigm was 3.8081, which revealed the fit diagnostics of the syphilis data, ran within the negative binomial distribution(see Appendix A).

Discussion

In literature, Poissonian probabilistic paradigms have revealed robust pseudo-R^2 values based on countvariables of optimally parameterizable, time-series, syphilis explanatory when compared with dichotomous, binomialized frequency models. Logistic regression commonly employs log-transformed, binary, dependent variables (example, 0 = infected, 1= non-infected). Since endemic, Poissonian syphilismodels employ actual non-log-transformed independent demographic variables, the regressions are more robust. However unfortunately, in Poissonian syphilis-oriented, forecast, vulnerability, linear analysis, over-dispersion would be common, since the variance in the explanative residual outputs may not be equivalent to the mean. Hence, the pseudo-R^2 values rendered from a probabilistic Poissonian paradigm would be over dispersed due to unquantitated outliers (e.g., extremeoutliers). Fortunately, a negative binomial regression with a non-homogenous gamma-distributed mean can compensate for over-Poissonian variation in a endemic, county-level, syphilis model.

As seen from our models the model fit diagnostics from both the logistic and Poissonian regressions do not fall within the normally expected paradigm. The R^2 value of 0.1041 represented a highly non-linear relationship between the elicited variables and the model variance as designed within the boundaries set by the model design. While this output did not indicate a poor model fit in and of itself, it did highlight underlying issues within the sampled Hillsborough County, syphilis dataset of heteroskedacity, based on a non-Gaussian distribution of the mean. This combination of issues within the logistic framework may still allow anendemic syphilismodel to run, but the regression may skew the impact of the correlations generated.

The Poissonian syphilis regression model forecast also revealed poor model fit diagnostics. The general assumption for a good model design amongst Poisson regressions is that the model deviance divided by the degrees of freedom should be equal to or close to 1.0(Pregibon, D. 1981).Variations of the V/DF score greater than or less than 1.0 represent an over- or under-dispersion, respectively, of the model data (Haight, 1967). In the case of the Poissonian syphilis model for the Hillsborough county study site, the V/DF score was 3.8081, representing a large over-dispersion of the data. When the negative binomial regression was applied to the identical dataset, the V/DF reduced to 1.0247; a dramatic movement towards a nearly ideal model design.

We conclude that the surprising success of this method is due to the ability of the negative binomial regression to alter the landscape of the base model data in such a manner that the non-homogenous gamma distributed mean and heteroskedastic nature become less apparent. By reducing these, and effectively pushing the variance of the model closer to the mean, the model becomes a stronger fit due to the gross outliers no longer skewing the normality of the data distribution as much.

Another conclusion that we derived from this research was the idea that frequentist syphilis endemic modeling is not able to account for some biological assumptions, as was mentioned in Johnson et al (2016). Based on the data provided, and the principles of frequentist assumptions, we believe that frequentist modeling may be incapable of differentiating the risk profiles amongst syphilis incidence data between casual contacts, such as family and coworkers, and sexual contacts, such as significant others, sex workers, etc. As some diseases, such as TB do not need such precise differentiation, this would not matter, but in the case of STI analysis the model must be able to focus on the risk applied only to those that the infected are actively having unprotected sexual contact with. If, as this theory suggests, a model is unable to differentiate the risks between casual and sexual contacts, then it will be likely to over- or under-estimate odds ratios associated with likelihood of syphilisinfection.

Questioning the mathematical ability to adequately account for interpersonal demographic factors specific to sexual activity lends us to also believe that these finding are most likely applicable in the broader sense of all sexually transmitted infections. Further investigation will be needed to confirm this hypothesis.

Some limitations that we faced in the pursuit of our goals in this study include the availability and thoroughness of the recorded incidence data. While the data recovered from theHillsborough Department of Health was very thorough in recording, we determined that the inclusion of two particular variables would have greatly increased the ability to determine accurate analysis results. These two variables are number of sex partners, and sexual orientation. We came to this conclusion based on the heavy use of both variables in many other statistical analyses amongst all manners of sexually transmitted infections (Shahmanesh et al, 2000; CDC, 2006; Chen et al 2002; Johnson et al 2016; Prabhakararao et al, 2014).

4.0 Conclusion:

Even though the negative binomial regression was able to compensate for the outliers in ta Poissonian, county-level, syphilis, probability model, the residual outputs cannot reveal location data (e.g., clustering tendencies

such as negative autocorrelation.) In order to implement control strategies for county-level syphilis model, it is vital to generate forecast maps of geographic locations where hyperendemic transmission occurs. An eigenfunction decomposition algorithm may cartographically delineate georeferenced syphilis explanatory predictors. Spatially weighted algorithms can prioritize varying and constant intra-cluster covariates associated with syphilis county-level prevalence (Griffith 2003).

Given the above observations, we find it reasonable to conclude that while frequentist methodologies are incredibly reliable in many, if not most other applications of disease analysis, they are not the best option for analysis of covariance as applied to syphilis infections. In the future, researchers may prefer to use other, more sophisticated methods for analyzing outbreaks of syphilis, such as geospatial analysis or Bayesian analysis, which have been used to great effect in similar areas of interest (e.g., Jacob et al. 2013). While the remarkable effectiveness of the negative binomial regression as a correction tool for correcting Poissonian error was impressive, it is our consideration that given the findings discussed previously, it would be unwise for a prudent researcher to rely on any conclusions discovered based on these methods.

References:

[1] Centers for Disease Control and Prevention (CDC). Primary and secondary syphilis--United States, 2003-2004. MMWR Morb Mortal Wkly Rep 2006; 55: 269-73. Retrieved June 11, 2016 from: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5510a1.htm

[2] Chen, S. Y., Gibson, S., Katz, M. H., Klausner, J. D., Dilley, J. W., Schwarcz, S. K., … McFarland, W. (2002). Continuing increases in sexual risk behavior and sexually transmitted diseases among men who have sex with men: San Francisco, Calif, 1999-2001 [3]. *American Journal of Public Health*, *92*(9), 1387–1388. http://doi.org/http://dx.doi.org/10.2105/AJPH.92.9.1387-a

[3] de Araujo, C. L., Shimizu, H. E., de Sousa, A. I. A., & Hamann, E. M. (2012). Incidence of congenital syphilis in Brazil and its relationship with the family health strategy. *Revista de Saude Publica*, *46*(2), 479–486.

[4] Garnett, G. P., Aral, S. O., Hoyle, D. V., Cates, W. J., & Anderson, R. M. (1997). The Natural History of Syphilis: Implications for the Transmission Dynamics and Control of Infection. *Sexually Transmitted Diseases*, *24*(4), 185–200. Retrieved from http://journals.lww.com/stdjournal/pages/articleviewer.aspx?year=1997&issue=04000&article=00002&type=abstract

[5] Gesink Law, D. C., Bernstein, K. T., Serre, M. L., Schumacher, C. M., Leone, P. A., Zenilman, J. M., … Rompalo, A. M. (2006). Modeling a Syphilis Outbreak Through Space and Time Using the Bayesian Maximum Entropy Approach. *Annals of Epidemiology*, *16*(11), 797–804. http://doi.org/10.1016/j.annepidem.2006.05.003

[6] Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. http://doi.org/10.1037/1082-989X.3.4.424

[7] Jacob, B. G., Krapp, F., Ponce, M., Zhang, N., Caliskan, S., Griffith, D. A., … Novak, R. J. (2013). A Bayesian Poisson specification with a conditionally autoregressive prior and a r esidual Moran ' s coefficient minimization criterion for quantitating leptokurtic distributions in regression-based multi-drug resistant tuberculosis treatment protocols. *Journal of Public Health and Epidemiology*, *5*(March), 122–143. http://doi.org/10.5897/JPHE12.076

[8] Johnson, L. F., & Geffen, N. (2016). A Comparison of Two Mathematical Modeling Frameworks for Evaluating Sexually Transmitted Infection Epidemiology. *Sexually Transmitted Diseases*, *43*(3), 139–146. http://doi.org/10.1097/OLQ.0000000000000412

[9] Law, D. C. G., Serre, M. L., Christakos, G., Leone, P. A., & Miller, W. C. (2004). Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. *Sexually Transmitted Infections*, *80*(4), 294–9. http://doi.org/10.1136/sti.2003.006700

[10] Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., … Murray, C. J. L. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*(9859), 2095–2128. http://doi.org/10.1016/S01406736(12)61728-0

[11] Michaud, J. M., Johnson, S. M., & Ellen, J. (2004). Comparison of sex partner meeting venues and residences of syphilis cases in Baltimore. *Sexually Transmitted Diseases*, *31*(4), 239–242. http://doi.org/10.1097/01.OLQ.0000118541.44827.92

[12] Prabhakararao, G. (2014). Mathematical Modeling Of Syphilis Disease A Case Study With Reference To Anantapur District- Andhrapradesh- India. *International Journal of Engineering Research and Applications*, *4*(10), 29–39.

[13] Pregibon, D. (1981) Logistic Regression Diagnostics, Annals of Statistics, Vol. 9, 705-724.

[14] Shahmanesh, M., Gayed, S., Ashcroft, M., Smith, R., Roopnarainsingh, R., Dunn, J., & Ross, J. (2000). Geomapping of chlamydia and gonorrhoea in Birmingham. *Sexually Transmitted Infections*, *76*(April 1996), 268–272. http://doi.org/10.1136/sti.76.4.268

[15] Singh, A. E., & Romanowski, B. (1999). Syphilis: Review with emphasis on clinical, epidemiologic, and some biologic features. *Clinical Microbiology Reviews*, *12*(2), 187–209.

[16] Sirotin, N., Strathdee, S. A., Lozada, R., Abramovitz, D., Semple, S. J., Bucardo, J., & Patterson, T. L. (2010). Effects of government registration on unprotected sex amongst female sex workers in Tijuana; Mexico. *International Journal of Drug Policy*, *21*(6), 466–470. http://doi.org/10.1016/j.drugpo.2010.08.002

[17] Zhang, X., Zhang, T., Pei, J., Liu, Y., Li, X., & Medrano-Gracia, P. (2016). Time Series Modelling of Syphilis Incidence in China from 2005 to 2012. *PloS One*, *11*(2), e0149401. http://doi.org/10.1371/journal.pone.0149401

International Journal of Advanced Mathematics
Vol. 4, No. 1, April 2017, pp. 1-13
Available Online at http://acascipub.com/Journals.php

Graph 1.1: Distribution of Hillsborough County syphilis cases, 2010-2014between 2010 and 2014.

This graphic depicts the distribution of incident cases of syphilis in Hillsborough County, Florida
Graphic was generated in GIS v10.3, and all points are randomly generated, representing a single case as dictated by zip code locational data.

Appendix A.
Logistic regression fit diagnostics

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	2313.774	2133.008
SC	2319.294	2199.245
-2 Log L	2311.774	2109.008

R-Square	0.1041	**Max-rescaled R-Square**	0.1457

This figure displays the R^2 and AIC fit diagnostics of the logistic regression analysis.

Poisson regression fit diagnostics

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	171	651.1834	3.8081
Scaled Deviance	171	651.1834	3.8081
Pearson Chi-Square	171	662.3209	3.8732
Scaled Pearson X2	171	662.3209	3.8732
Log Likelihood		3362.0338	
Full Log Likelihood		-635.9867	
AIC (smaller is better)		1287.9733	
AICC (smaller is better)		1288.8204	
BIC (smaller is better)		1313.4724	

This figure displays the V/DF and other fit diagnostics of the Poissonian regression analysis.

Negative binomial fit diagnostics

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	171	175.2194	1.0247
Scaled Deviance	171	175.2194	1.0247
Pearson Chi-Square	171	159.1518	0.9307
Scaled Pearson X2	171	159.1518	0.9307
Log Likelihood		3495.0598	
Full Log Likelihood		-502.9606	
AIC (smaller is better)		1023.9213	
AICC (smaller is better)		1024.9864	
BIC (smaller is better)		1052.6078	

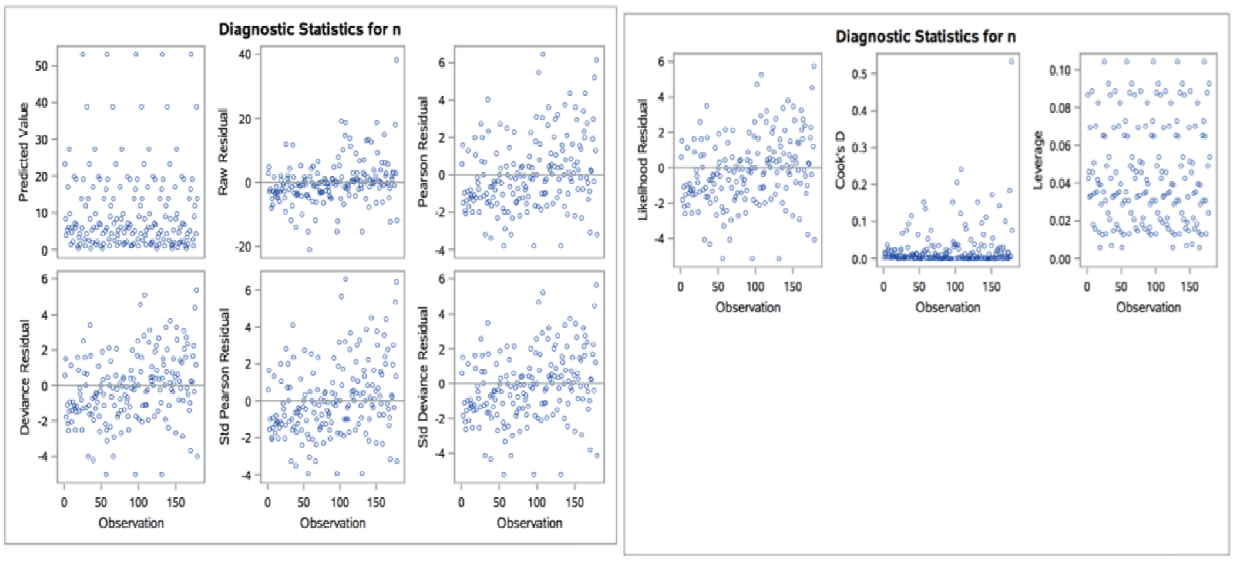
This figure displays the V/DF and other fit diagnostics of the Negative binomial regression analysis.

International Journal of Advanced Mathematics
Vol. 4, No. 1, April 2017, pp. 1-13
Available Online at http://acascipub.com/Journals.php

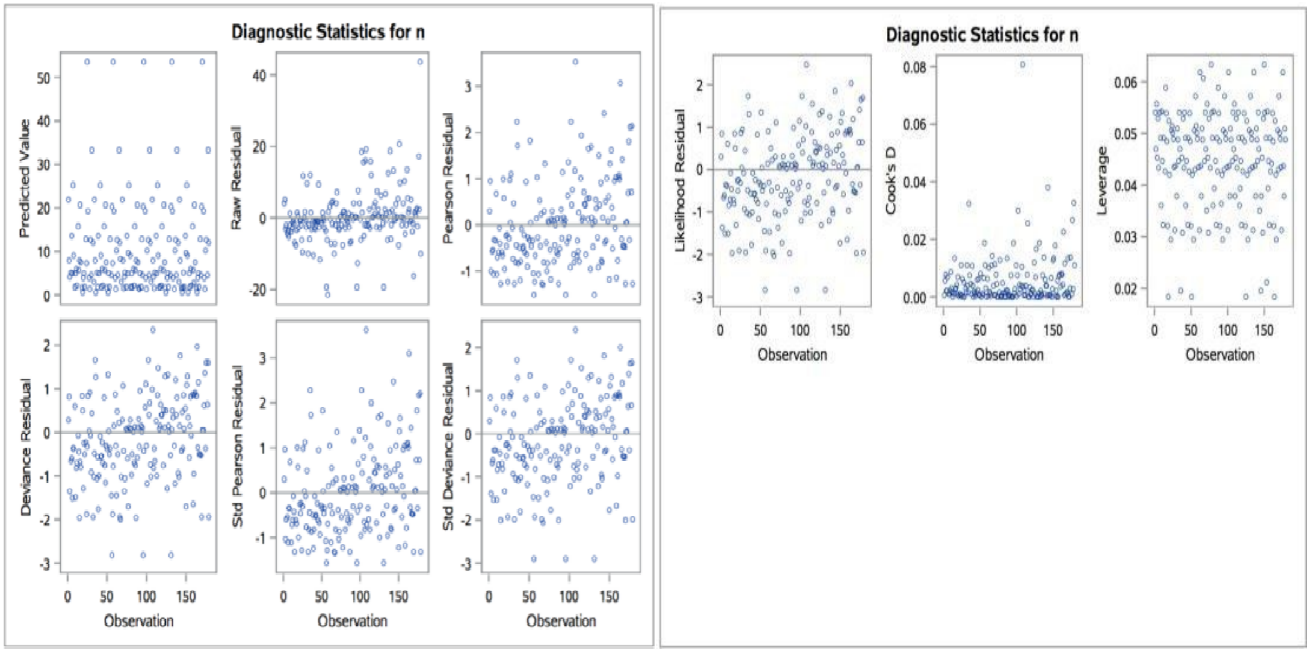
Please take note of scales of Y-Axes

Diagnostic Residual Plots

Poisson



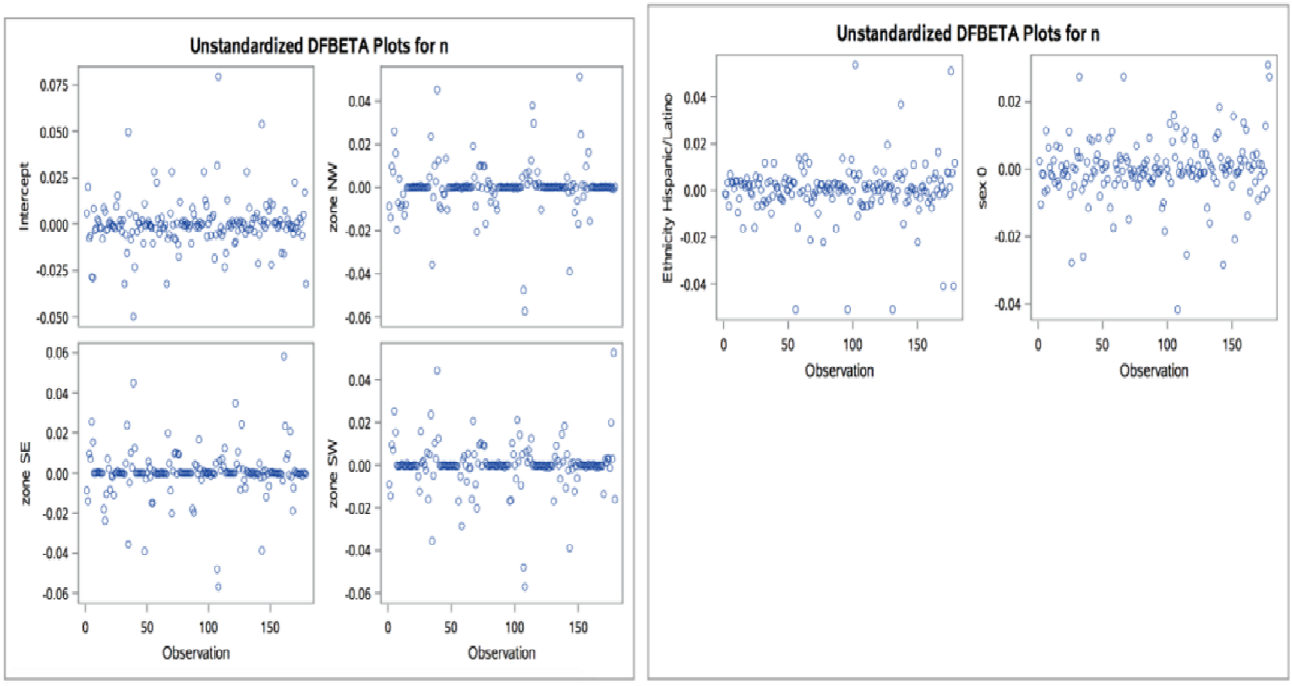
Negative Binomial



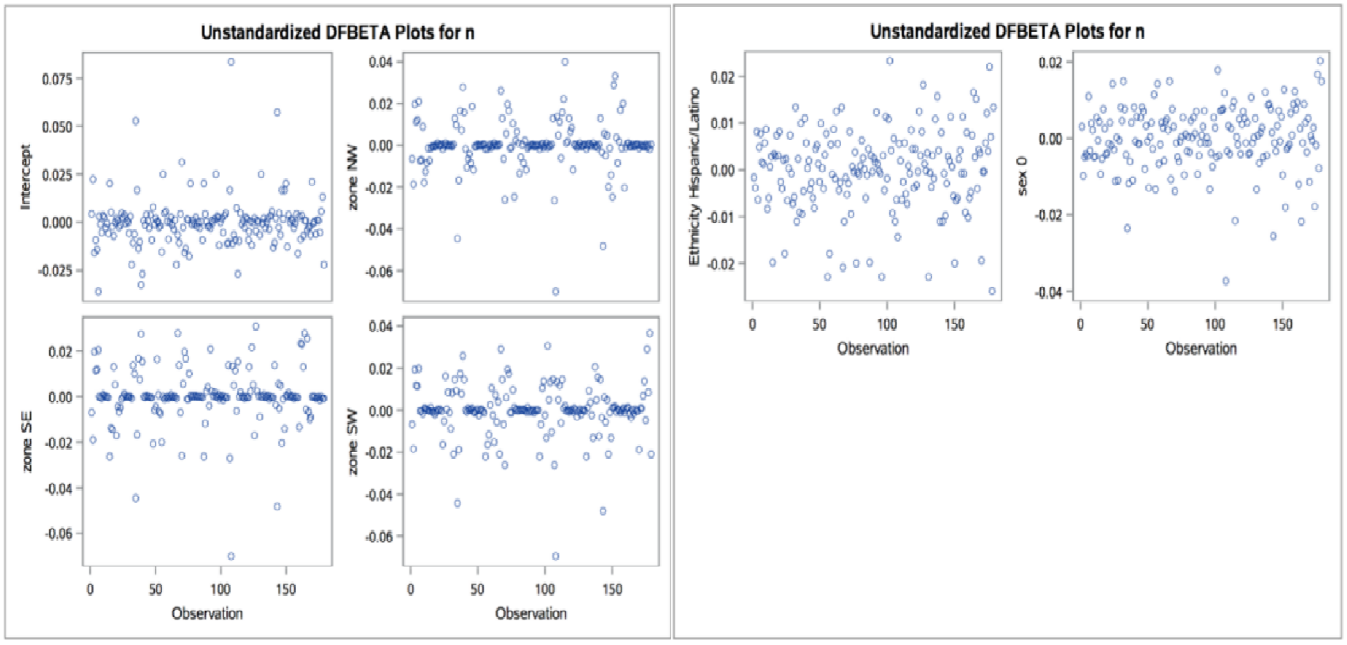
International Journal of Advanced Mathematics
Vol. 4, No. 1, April 2017, pp. 1-13
Available Online at http://acascipub.com/Journals.php

Unstandartized DF-Beta Residuals

Poisson

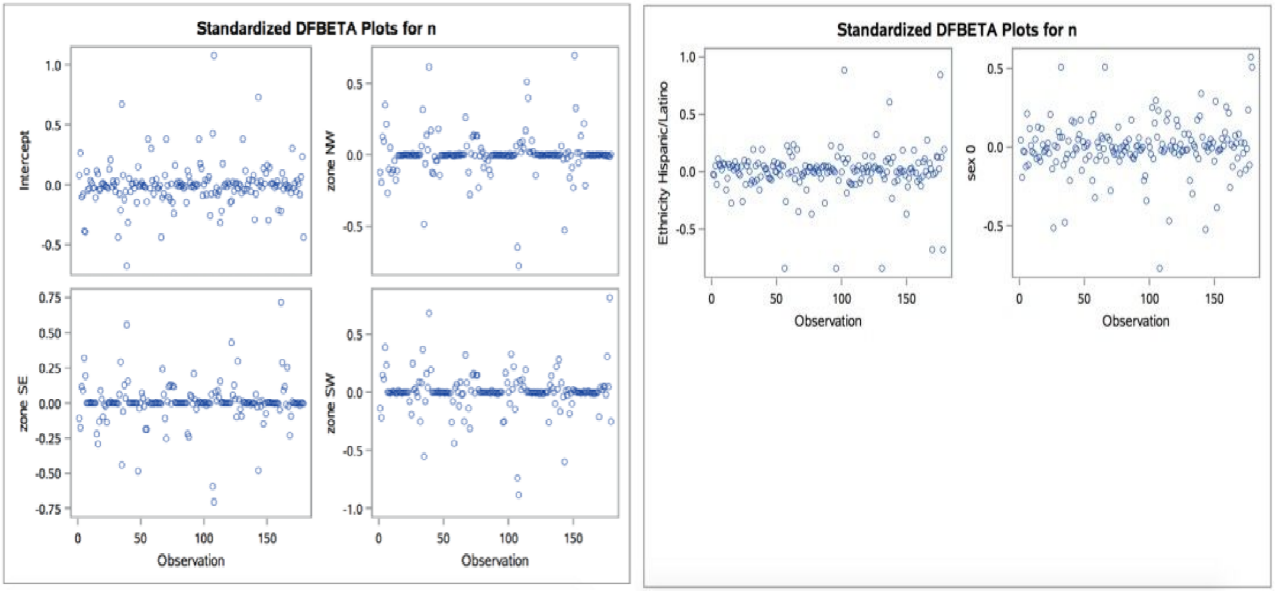


Negative Binomial



Standardized DF-Beta Residuals

Poisson



International Journal of Advanced Mathematics
Vol. 4, No. 1, April 2017, pp. 1-13
Available Online at http://acascipub.com/Journals.php

Negative Binomial