Johnson et al., J Remote Sens GIS 2018, 7:2
DOI: 10.4172/2469-4134.1000243

Journal of Remote Sensing & GIS

Research Article Open Access

Employment of a Negative Binomial Regression Probabilistic Paradigm with a Non-Homogenous Gamma-Distributed Mean to Compensate for Over-Poissonian Variation in a County Level Syphilis Model

Grant Johnson, Brock Graham, Samuel Alao and Benjamin G Jacob*

Department of Global Health, College of Public Health, University of South Florida, Tampa, FL, USA

***Corresponding author:** Benjamin G Jacob, Department of of Global Health, College of Public Health, University of South Florida, Tampa, FL, USA, Tel: +813-974-9784; E-mail: bjacob1@health.usf.edu

Rec date: December 14, 2016; **Acc date:** April 30, 2018; **Pub date:** May 03, 2018

Abstract

Syphilis is an ongoing problem on the world stage, and thus the methods for analyzing it's spread are always being tested, challenged, and continuing to grow. This investigation set out to assess the ongoing syphilis epidemic in Hillsborough County, FL in hopes of determining critical correlates central to the increasing incidence. Instead, we discovered that the frequentist paradigms commonly utilized by researchers may be inherently flawed when applied to certain subsets of data. Utilizing logistic, Poissonian, 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, and perhaps other sexually transmitted diseases as well, the models did not fit. Negative binomial regression did correct for overdispersion amongst the Poissonian findings, but further investigation is warranted into more certain and reliable methods of assessing STI outbreak data

Keywords: Regression; Homoskedacity; Autocorrelation; Dispersion

Introduction

Unfortunately for all of the humankind, syphilis remains a problem. Known as "The Great Pretender", this disease is known for masking its 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 [1]. 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 focused 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 [2,3].

This problem, the apparent statistical inconsistency of demographic effectivity of the disease, was the focusing question that sparked this investigation. One of the most commonly employed statistical sampling methods in the epidemiologic analysis is the frequentist method, known also as regression or linear analyses. This method of mathematical hypothesis testing is used to search for a causative correlation relationship between an investigator-identified independent variable and any number of dependent variables [4] and, as with any other mathematical model, relies on a handful of assumptions being satisfied to ensure validity of the result [1].

Shahmanesh et al. [5] 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 used a forward stepwise logistic regression model to assess the correlation between patients with chlamydia and patients with gonorrhea based on the variables of ethnicity, age, sex and super profile analysis, an estimator for socioeconomic status. The findings of this article discussed a higher risk of infection for African-Caribbean males under the age of 20 that lived in neighborhoods of similar sociodemographic indicators [5]. This finding of a sociodemographic profile for high-risk individuals could be used in that testing area for a targeted intervention on the part of the local health department to attempt to reduce the incidence of disease. Unfortunately, the investigators 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. [6] 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 test whether the misspecifications were due to the stochastic, non-Gaussian assumptions of frequency-based explanative models, but found that the propagational error only occurred when applied to STI's [6]. They discussed that this finding suggested that the mathematical error may have been due to biological assumptions being violated that the mathematical algorithms were not or could not 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 and will 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 covariates, no multicollinearity amongst independent variables, homoskedacity, and no autocorrelation amongst the 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 distributions.

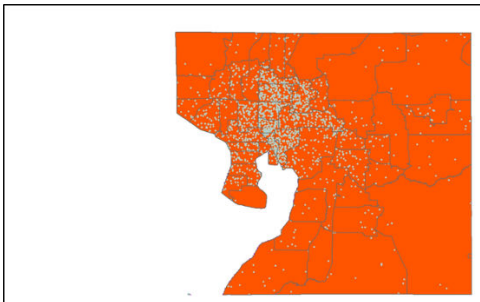
Homoskedacity is one of the primary assumptions that a modeler makes when selecting covariates for a study of any kind, the other being independence of the variables. It is, therefore, vital that one does not automatically skip over consideration of this whilst designing a design model. Homoskedacity is the assumption that the variance of all of the selected independent covariates is equivalent. This equivalence is a necessity for clean, reliable correlation values, however, it is often close to impossible to find in nature, especially when dealing with human populations and behavior. Therefore, it is often up to the designer to utilize specific modeling techniques to either eliminate the outliers or artificially generate a normalized variance, as in through the log-transformation methodologies of the negative binomial framework. 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 change of this value, then the covariates are homoskedastic, however, any changes, especially large ones, indicate heteroskedacity, and the model designer must then decide whether or not to remove the selected variables, or control for them in another manner.

Autocorrelation is a concept most commonly described in spatial analyses. It is the habit of certain covariates to reduce or increase euclidean distances from one another in a natural setting. In a non-spatial frequentist analysis, this is seen as a tendency of covariates to be found more often in one setting than another. These clustering tendencies can and will be easily misjudged by the frequentist model as spurious correlations and can very quickly skew the findings toward or away from the null hypothesis.

Multicollinearity, on the other hand, is a phenomenon that is easily missed during model design, as it is the incidental lack of independence of the covariates, leading to alterations of pseudo-R^2 values based on two or more covariates numerical values being dependent upon one another. This assumption is vital to test for in frequentist tests for correlation. Since the previously assumed 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 frequentist regressional methods to analyze and predict syphilis outbreaks. We have hypothesized that the reason for this may indeed lie in a violation of biological assumptions being made in the designs of many regressive models.

While attempting to utilize frequentistic analytic methods to determine critical correlates in a syphilis outbreak in Hillsborough County, Florida (see Graph 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 biological assumptions when attempting to apply frequentist methodologies to a sexually transmitted disease data background. Thus, the research objectives in this comparative regression analysis are 1) to compare pseudo-R^2 values rendered from a dichotomous logistic bivariate model and a Poisson probability paradigm to determine robustness and 2) utilize residual forecast from a negative binomial regression analysis employing a non-homogenous gamma-distributed mean to compensate for violations of assumptions (extreme outliers, over-dispersion). Although this model alludes to syphilis endemicity at the county level, we envision using this model for other sexually transmitted diseases.



Graph 1: Distribution of Hillsborough County syphilis cases, 2010-2014.

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 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 variables.

A bivariate logistic regression was first used to examine the possibility of a correlation between covariates. A dichotomous variable (DV) was generated as the response variable by separating the cases into the duration of infection categories, split at the one-year mark. This mark was determined by the disease stage variable; all cases that were considered primary, secondary or early latent 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. [7] to analyze covariates for a Bayesian analysis of a syphilis outbreak. According to the results of the stepwise analysis (results not shown) the age, sex, race, and quadrant variables were set against the response variable for analysis.

A Poissonian regression analysis was also used 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 a number of cases falling into each of the potential variablilistic 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 analyses 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 Poissonian model which would result in an underdispersion of the model data, a negative binomial regression was also performed on the aforementioned data [8].

Results

As the focus of this analysis is on the appropriate application of frequency analyses in reference to STI data, we will be assessing the fit diagnostics for our results, rather than the outputs of the covariate analyses to determie validity of findings. The logistic regression returned an R^2 value of 0.1041, as can be seen in Figure 1.

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

Figure 1: Logistic 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	

Figure 2: Poisson regression fit diagnostics.

The AIC and other model fit diagnostics are all in the high two thousand. The Poisson and negative binomial analyses are both best analyzed for model fit by measuring deviance over degrees of freedom, or V/DF, to assess dispersion. The V/DF in the Poisson paradigm is 3.8081, as can be seen in Figures 2 and 3 shows the fit diagnostics of the same data, ran within the negative binomial regression instead, this time with a V/DF of 1.0247.

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	

Figure 3: Negative binomial fit diagnostics.

Discussion

In literature, Poissonian probabilistic paradigms have revealed robust pseudo-R^2 values based on count-variables of time-series parametrizable syphilis explanators when compared with dichotomous, binomialized frequency models. Logistic regression commonly employs log-transformed binary dependent variables (example, 0=infected, 1=non-infected). Since Poissonian models employ actual non-log-transformed independent 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 residual outputs would not be the equivalence to the mean. Hence, the pseudo-R^2 values rendered from the probabilistic poissonian paradigm would be over dispersed due to unquantitated outliers (example, extreme observations). Fortunately, a negative binomial regression with a non-homogenous gamma-distributed mean can compensate for over-Poissonian variation.

As can be seen above, 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 represents a highly non-linear relationship between the elicited variables and the model variance as designed within the boundaries set by the model design [4]. While this does not indicate a poor model fit in and of itself, it does highlight underlying issues within the dataset of heteroskedacity, and non-gaussian distribution of the mean. This combination of issues within the logistic framework will still allow the model to run but may skew the impact of the correlations the regression generates.

The Poissonian regression also showed some very 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. Variations of the V/DF

score greater than or less than 1.0 represent an over- or under-dispersion, respectively, of the model data. In the case of the Poisson model, the V/DF score is 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 concluded 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 discussed was the idea of frequentist modeling not being able to account for some biological assumptions, as was mentioned in Johnson et al. [6]. 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 the likelihood of infection.

This line of questioning the mathematical ability to adequately account for interpersonal factors specific to sexual activity lends us to also believe that these findings 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 include the availability and thoroughness of the recorded incidence data. While the data recovered from the Department of Health was very thorough in its 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 [2,3,5,6,9].

Conclusion

Even though the negative binomial regression was able to compensate for the outliers in the Poissonian county-level syphilis model, the residual outputs cannot reveal geolocation data (example, clustering tendencies such as negative autocorrelation.) In order to implement control strategies for county-level syphilis, it is vital to generate forecast maps of geolocations where hyperendemic transmission occurs. An Eigenfunction decomposition algorithm can cartiographically delineate georeferenced explanatory predictors. Spatially weighted algorithms can prioritize varying and constant intra-cluster covariates associated with syphilis county-level prevalence.

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 sexually transmitted 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. While the remarkable effectiveness of the negative binomial regression as a correction tool for Poissonian error was impressive, it is our consideration that given the findings discussed previously, it would be unwise for the prudent researcher to rely on any conclusions discovered based on these methods.

References

1. Jacob BG, Krapp F, Ponce M, Zhang N, Caliskan S, et al. (2013) A Bayesian Poisson specification with a conditionally autoregressive prior and a residual 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: 122-143.
2. Centers for Disease Control and Prevention (CDC) (2006) Primary and secondary syphilis--United States, 2003-2004. 55: 269-273.
3. Chen SY, Gibson S, Katz MH, Klausner JD, Dilley JW, et al. (2002) Continuing increases in sexual risk behavior and sexually transmitted diseases among men who have sex with men: San Francisco, Calif, 1999-2001. American Journal of Public Health 92: 1387-1388.
4. Pregibon D (1981) Logistic Regression Diagnostics. Annals of Statistics 9: 705-724.
5. Shahmanesh M, Gayed S, Ashcroft M, Smith R, Roopnarainsingh R, et al. (2000) Geomapping of chlamydia and gonorrhoea in Birmingham. Sexually Transmitted Infections 76: 268-272.
6. Johnson LF, Geffen N (2016) A Comparison of Two Mathematical Modeling Frameworks for Evaluating Sexually Transmitted Infection Epidemiology. Sexually Transmitted Diseases 43: 139-146.
7. Gesink Law DC, Bernstein KT, Serre ML, Schumacher CM, Leone PA, et al. (2006) Modeling a Syphilis Outbreak Through Space and Time Using the Bayesian Maximum Entropy Approach. Annals of Epidemiology 16: 797-804.
8. de Araujo CL, Shimizu HE, de Sousa AIA, Hamann EM (2012) The incidence of congenital syphilis in Brazil and its relationship with the family health strategy. Revista de Saude Publica 46: 479-486.
9. 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: 29-39.