Dead Bird Surveillance as a Predictor of Equine Risk for West Nile Virus Infection

MATH-493 - Applied Biostatistics Final Project - GLM

Ernesto Bocini

Introduction

The efficacy of different surveillance methods for predicting and preventing human and veterinary illness from West Nile Virus (WNV) infection in the United States has been argument of debate since its introduction in 1999.This paper will discuss the effect of dead bird surveillance, which has been a focus of particular interest since bird mortality typically precedes human or equine WNV infection. This article will delve into the Poisson regression model of equine WNV (West Nile Virus) rate, examining how it varies based on the rate of WNV-positive dead birds. The analysis will also factor in population density, accounting for potential variations in the impact of population density on WNV rate. The Poisson regression model will demonstrate a strong match with the available data.

The study utilizes the variables described below:

- Equine cases (int): Count variable. Number of WNV-positive equine cases in the specific County.
- County (str): County area in South Carolina.
- Bird cases (int): Count data. Number of WNV-positive bird cases in the specific County.
- Farms (int): number of farms in the specific County.
- Area (int): area of the County in squared miles.
- Population (int): population of the specific County.
- Human density (float): human density of the county computed as Population/Area
- Positive bird rate (float): (PBR) # Bird Cases of West Nile / Human Population
- Positive Equine Rate (float): (PER) # of Equine Cases of West Nile / # Farms

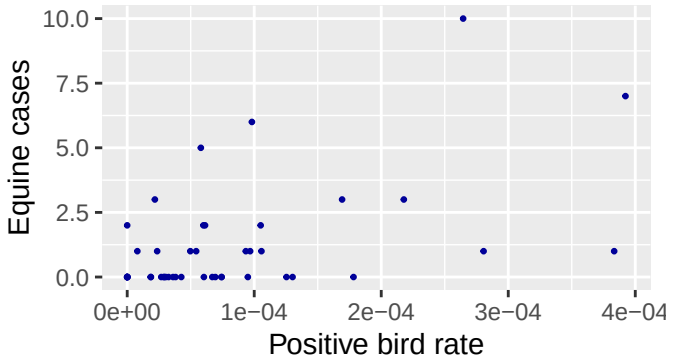
These data is coming from a combination of sources, among which the *South Carolina Department of Health and Environmental Control*, *(U.S. Census Bureau 2000*, and *United States Department of Agriculture's Census of Agriculture statistics.*

Exploratory Data Analysis

Before digging into model selection and implementation, it is crucial to conduct exploratory data analysis. It's important to conduct exploratory data analysis before implementing models because it helps us understand the data better, find patterns, and make informed decisions about how to build and process our models. Let's start by looking at some descriptive univariate statistics:

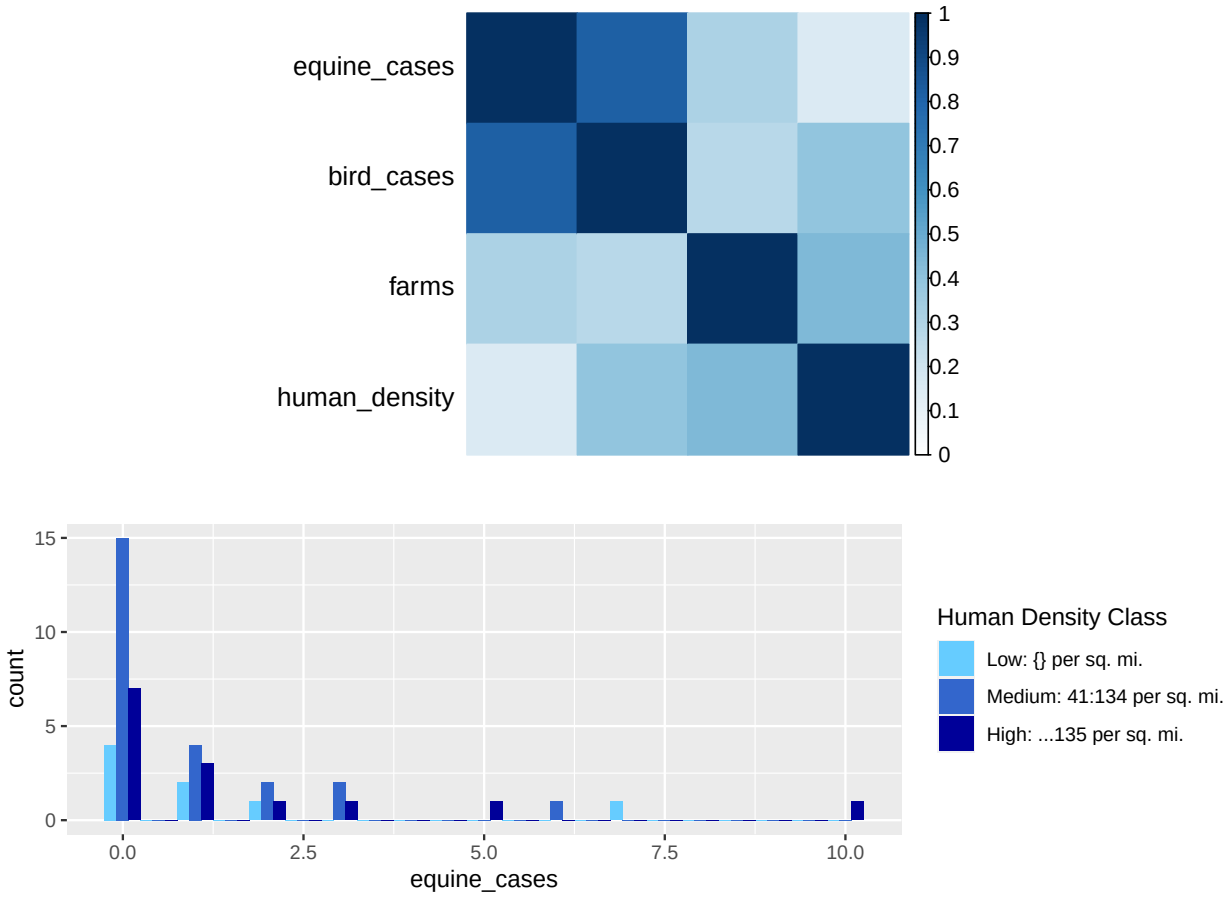
	equine_cases	bird_cases	population	farms	area
	Min. : 0.000	Min. : 0.000	Min. : 9960	Min. : 92.0	Min. : 360.0
	1st Qu.: 0.000	1st Qu.: 1.000	1st Qu.: 26831	1st Qu.: 254.2	1st Qu.: 498.2
	Median : 0.000	Median : 4.000	Median : 52598	Median : 348.0	Median : 616.0
	Mean : 1.174	Mean : 6.109	Mean : 87229	Mean : 437.8	Mean : 654.5
	3rd Qu.: 1.000	3rd Qu.: 6.750	3rd Qu.:118403	3rd Qu.: 590.5	3rd Qu.: 781.5
	Max. :10.000	Max. :52.000	Max. :379633	Max. :1271.0	Max. :1134.0

Each variable has 46 valid observations and looking at the table above, we can see the minimum, maximum, median, mean, and quartiles for each variable. The output shows that equine_cases and bird_cases have a positively skewed distribution because their mean values are greater than their median values. This is somehow expected because both the variables represent count data, which often follows a Poisson distribution, obviously positively skewed. Additionally, the following scatter plot of Equine cases versus Positive bird rate revealed a positive linear relationship, confirming that bird surveillance could potentially serve as a predictor of equine WNV infection.



To proceed with the exploration, the following correlation matrix shows a strong positive correlation between Equine cases and Bird cases (r=0.82), a weak-moderate correlation between equine cases and number of farms (r=0.31) and finally a weakly positive relation between Equine cases and Human Density (r=0.16). However, it also shows a moderate correlation between bird cases and human density (r=0.40). This is suggesting to include their interaction inside the final model. On top of this, please note that some of the regressors described in the introduciton were not taken into consideration for obvious reasons: Population and Area are both included in Human Density computation, PER was taking the dependent variable in his formula, and the categorical regressor County had only one sample for each County and

was giving no useful information. With this in mind, before entering the modelling section, a conditional histogram separated out by levels of density is plotted to have a clearer picture of the distribution of the dependent.



As expected, there are different count levels for different density classes. What is most important to take home from this plot, however, is that for ll the classes, the count of equine cases sis positively skewed and may suggest the implementation of a Poisson. this plot is suggesting that a Poisson may be a good proxy for our dependent variable in all the density classes. Another important observation is the presence of an evident outlier, which may effect the accuracy of the Poisson model and should be treated with caution in the following section.

Overall, the EDA supports the inclusion of Bird Counts, Human density, Farms, and the interaction term PBR*HD in the final Poisson regression model, as they demonstrated relevant relationships with Equine cases. However, since we are more interested in explaining the positive equine rate (PER) rather than the absolute count of equine cases, we can use PBR instead of Bird Counts and include the log number of farms as an offset variable. By doing so, we restrict the regression coefficient of the offset variable to be 1, thus allowing our model to represent rates rather than counts. Note that, we do this instead of directly modelling the PER, in order to still be ale to use Poisson likelihood functions, which would no longer be possible in case of using directlu the rate as ta response.

Statistical Analysis

Poisson Model

At this point, we are ready to perform our Poisson model analysis. The Poisson regression model is part of the large family of Generalized Linear Models (GLM) and has the following general form:

$$Y|X_1, X_2, X_3, ..., X_n \sim Pois(\lambda).$$
$$log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n.$$

This model is identical to the linear regression model, except for the addition of the log term on the left-hand side of the equation. As suggested by the EDA we fit the following model:

$$Y = equine\_cases;\ X_1 = PBR;\ X_2 = Human\_Density;\ X_3 = Farms.$$

$$Y|X_1, X_2, X_3 \sim Pois(\lambda)$$
$$log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + log(X_3)$$

Hence, Poisson regression analysis was used to model the number of WNV-positive equids per county as a function of *PBR* and *human_density*, with the log of the number of farms per county as an offset variable. An interaction term for PBR*human population density was included to capture the effect modification of the association by the level of urbanization. The results are show in the following table:

term	estimate	std.error	statistic	p.value
(Intercept)	-6.9824313	0.3664332	-19.0551275	0.0000000
PBR	5329.8421025	1595.6417764	3.3402498	0.0008370
human_density	0.0005903	0.0016728	0.3528668	0.7241883
PBR:human_density	26.0049514	12.3817783	2.1002598	0.0357060

The model output includes the estimated coefficients for each predictor variable and their standard errors, as well as their corresponding p-values testing whether the coefficients are significantly different from zero. Before commenting these results and performing a more in-depth model assessment, it's good habit to check the fundamental assumption behind the Poisson regression, i.e., the (conditional) mean $\mathsf{E}[y] = \mu$ assumed to be equal to the variance $\mathsf{VAR}[y] = \mu$. In order to do this, we can perform a dispersion test which assesses the hypothesis that this assumption holds (equidispersion) against the alternative that the variance is of the form:

$$H_1 : VAR[y] = (1 + \alpha) \cdot \mu = dispersion \cdot \mu\ .$$

Hence, the alternative corresponds to the specification of a Negative Binomial with linear variance, or quasi-Poisson model with dispersion parameter. Overdispersion corresponds to

$\alpha > 0$ and underdispersion to $\alpha < 0$. The coefficient α can be estimated by an auxiliary OLS regression and tested with the corresponding t (or z) statistic which is asymptotically standard normal under the null hypothesis. By performing this test [2], we obtain a significant result (p-value = 0.01) against the null hypothesis, hence favouring the alternative of true dispersion greater than 1, with dispersion around 1.5. This overdispersion becomes even smaller by removing the outlier spotted during the EDA, but still remains significant. However, since it's not too large of a dispersion, the next model we fit is a Quasi-Poisson, without the need to go for a Negative-Binomial.

Quasi-Poisson Model

The way this model is specified is the same as before, with exception that now the variance is assumed to be the following:

$$VAR[Y] = \alpha \times \lambda \ \ , \lambda = 1.5 \ \ (overdispersion)$$

Modelling the equine cases under this new set of assumptions lead to the following (similar) results:

term	estimate	std.error	statistic	p.value
(Intercept)	-6.9824313	0.4800277	-14.5458929	0.0000000
PBR	5329.8421025	2090.2915830	2.5498080	0.0145166
human_density	0.0005903	0.0021914	0.2693639	0.7889687
PBR:human_density	26.0049514	16.2201362	1.6032511	0.1163741

The coefficient estimate for the intercept term is -6.98, indicating the expected log rate of equine cases when all predictor variables are zero. The estimated coefficient for "PBR" is 5329.84, suggesting that an increase in "PBR" is associated with an increase in the expected log rate of equine cases. The estimated coefficient for "human_density" is positive but not significant, suggesting that there is weak evidence for an association between "human_density" and the expected log rate of equine cases. However, the estimated coefficient for the interaction term "PBR*human_density" is 26.00 and significant at the 0.05 level, indicating that the effect of "PBR" on the expected log rate of equine cases depends on the level of "human_density".

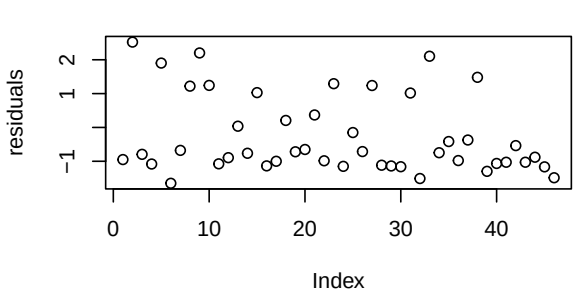
In order to get insights on how well the model is fitting the data, a Pearson-Chi-Squared goodness of fit test test was performed. However, since the amount of data at hand is very small, it is convenient to make some grouping of the data and perform the goodness of fit test on the groups instead, as it increases the effective sample size, addresses sparse data, stabilizes assumptions, and reduces noise. The test confirms that the model is not fitting the data well with moderate evidence.

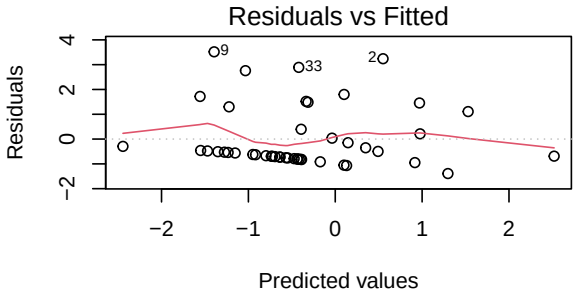
Model assessment

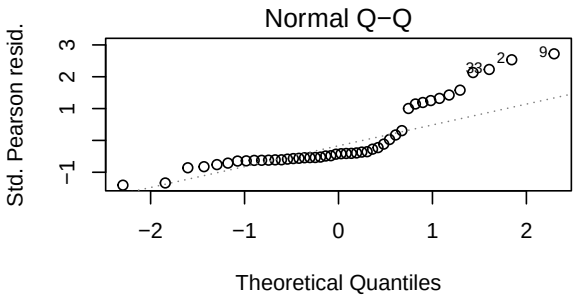
To perform model assessment for the GLM-Quasi-Poisson model with equine_cases as the dependent variable and PBR, human_density, and PBR*human_density as the independent variables, we first need to state the model assumptions.

Model assumptions:

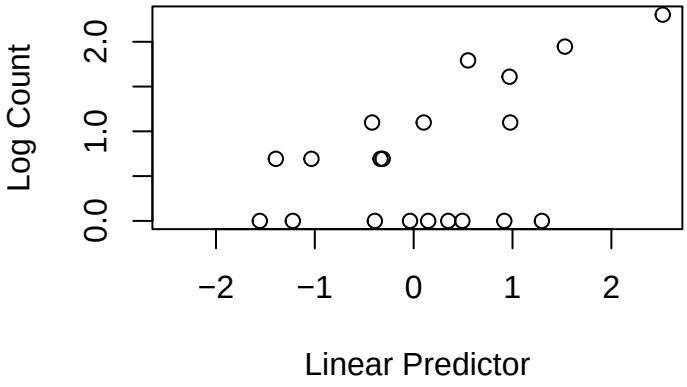
- Count outcome: equine_cases follows a Quasi-Poisson distribution.







- Independent observations: The observations in the dataset are independent of each other.
- Linear relation between log count and linear predictor: There is a linear relationship between the log * count of equine_cases and the linear predictor PBR + human_density + PBR*human_density.



Conclusions

add them

References