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*.

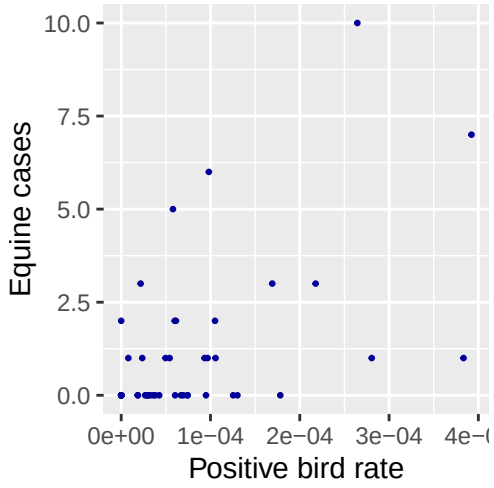
Exploratory Data Analysis

Before digging into model selection and implementation, it is crucial to conduct exploratory data analysis 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:

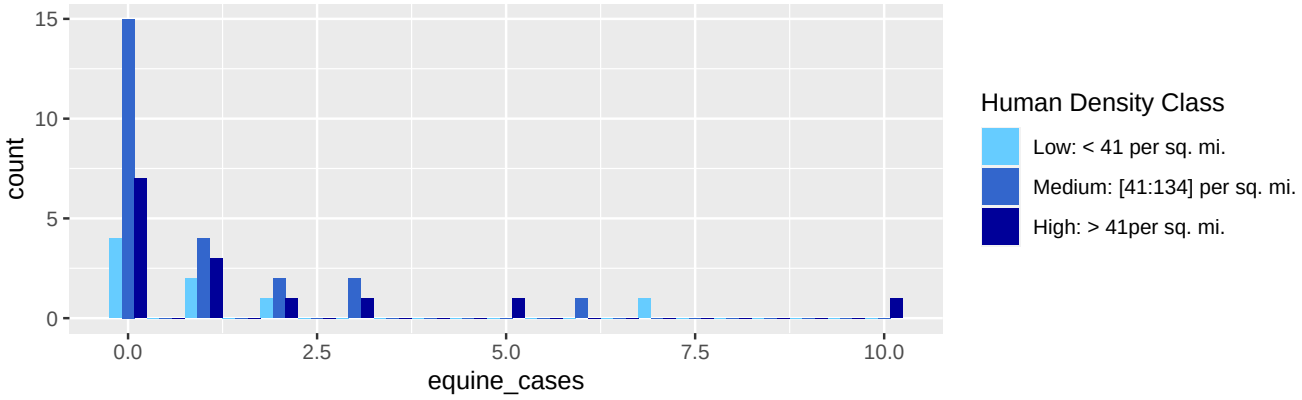
Variable	Min	Q1	Median	Q3	Mean	Max
equine_cases	0	0.00	0	1.00	1.17	10
bird_cases	0	1.00	4	6.75	6.11	52
population	9960	26830.75	52598	118403.25	87229.37	379633
farms	92	254.25	348	590.50	437.83	1271
area	360	498.25	616	781.50	654.54	1134

Each variable has 46 valid observations. The output of the table 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.

Correlation Matrix	equine_cases	bird_cases	farms	human_density
equine_cases	1	0.82	0.31	0.16
bird_cases	0.82	1	0.28	0.4
farms	0.31	0.28	1	0.45
human_density	0.16	0.4	0.45	1



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 all the classes, the count of equine cases is 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 directly the rate as ta response.

Statistical Analysis

Poisson Model

The Poisson regression model is part of the large family of Generalized Linear Models (GLM) and it models the relationship between the response variable and the predictor variables using the logarithm of the expected response, known as the log-link function. The model assumes that the logarithm of the expected response is a linear combination of the predictor variables. This linear relationship is modeled using regression coefficients.

Mathematically, the Poisson regression model can be represented as:

$$Y|X_1, X_2, ..., X_n \sim Pois(\mu).$$

$$\mu = exp(\eta),\ \eta = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n.$$

where μ is the mean (and variance) of the Poisson distribution. In this analysis, I initially included all the regressors described in the introduction, as well as additional variables suggested by the exploration analysis. I also incorporated the interaction term between "PBR" and "human_density" to capture their combined effect and included the offset term obtained from the logarithm of the number of farms. Then I performed backward selection while checking for multicollinearity in the regressors. The backward selection is based on the Akaike Information Criterion (AIC), a popular method to compare different models which penalizes the model complexity by considering the goodness of fit and the number of parameters. Mathematically, the AIC can be computed as:

$$AIC = -2 * ln(L) + 2 * n$$

Where n is the number of parameters, and $L = L(\hat{\theta})$ is the maximum value of the likelihood function of the model. The AIC penalizes complex models and aims to find the best balance between goodness of fit and model complexity. Backward selection helps select the most important predictors while considering the AIC. Checking for multicollinearity ensures independent contributions of variables, improving reliability and interpretability of the model.

After this selection process, I am happy to see that the resulting model correspond to the one suggested in the exploration, i.e.:

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

$$Model\ 1 : Y|X_1, X_2, X_3 \sim Pois(\mu).$$
$$\mu = exp(\eta),\ \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + log(X_3).$$

Hence, we 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. The results are show in the following table:

term	estimate	std.error	statistic	p.value
(Intercept)	-6.98	0.37	-19.06	0.00
PBR	5329.84	1595.64	3.34	0.00
human_density	0.00	0.00	0.35	0.72
PBR:human_density	26.00	12.38	2.10	0.04

The model output includes estimated coefficients, standard errors, and p-values for each predictor variable. The significance of each coefficient is assessed to determine if it significantly deviates from zero. "Human_Density" is retained in the model despite not being individually significant because it contributes to a significant interaction term.

Before interpreting the results and conducting a thorough model assessment, it is important to check the fundamental assumption of Poisson regression: $\mathsf{E}[y] = \mu$ is assumed to be equal to $\mathsf{VAR}[y] = \mu$. A dispersion test supports an alternative hypothesis of $\mathsf{VAR}[y] = (1 + \alpha) \cdot \mu$ with a significant p-value (0.01), indicating overdispersion with an estimated parameter of around 1.5. Although not extreme, a Quasi-Poisson model is preferred over a Negative Binomial model, which would be more suitable for large overdispersion.

Quasi-Poisson Model

It is important to note that over-dispersion can have a significant impact on the model's results, as it can lead to incorrect conclusions about the relationships between the variables. For this reason, I performed again the backward select + VIF check, using the same starting point as in the previous Poisson model fitting. At the end of the selection, the way this model is specified is the same as before, with exception that now the variance is assumed to be $\mathsf{VAR}[y] = (1+\alpha)\cdot\mu$:

$$\textit{Model 2}: Y|X_1, X_2, X_3 \sim Quasi - Pois(\mu, \alpha).$$

$$\mu = exp(\eta),\ \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + log(X_3).$$

where $\alpha = 1.5$ is the overdispersion and was estimated previously. As expected, since the overdispersion is not so large, modelling the equine cases under this new set of assumptions lead to the following similar results:

term	estimate	std.error	statistic	p.value
(Intercept)	-6.98	0.48	-14.55	0.00
PBR	5329.84	2090.29	2.55	0.01
human_density	0.00	0.00	0.27	0.79
PBR:human_density	26.00	16.22	1.60	0.12

The intercept term has an estimated coefficient of -6.98, representing the expected log rate of equine cases when all predictors are zero. "PBR" has a coefficient of 5329.84, indicating its association with an increase in the expected log rate of equine cases. "Human_density" has a positive but insignificant coefficient, suggesting weak evidence of its association with the expected log rate of equine cases. However, the significant interaction term "PBR*human_density" (coefficient: 26.00) indicates that the effect of "PBR" on the expected log rate of equine cases depends on the level of "human_density".

Model 2, considering overdispersion, will be used as the final model for its theoretical correctness and similar results. Model assessment will be based on this model.

Model assessment

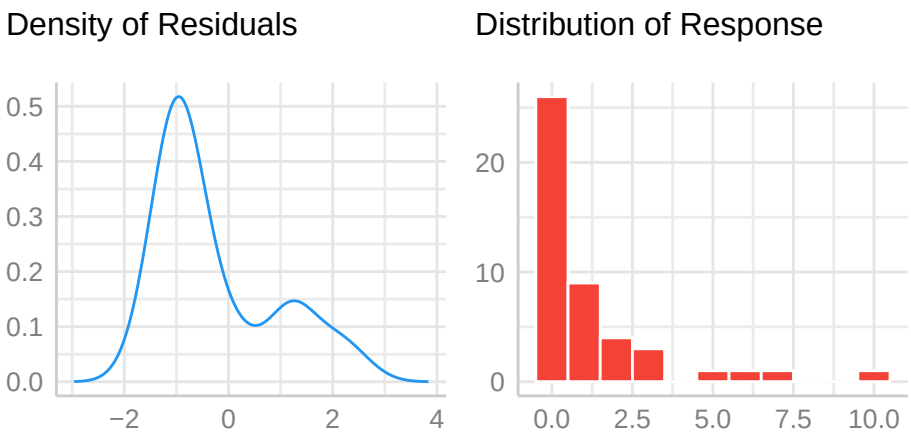
A final model assessment is now carried out to make sure that the assumptions of the *Model 2* are verified. We proceed assumption per assumption

Model assumptions:

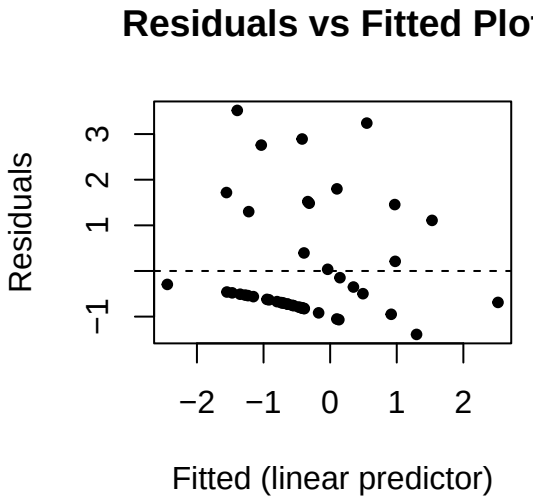
1) Equine_cases follows a Quasi-Poisson distribution:

$Y|X_1, X_2, X_3 \sim Quasi - Bin(\mu, \alpha)$ where $\hat{\mu} = exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + log(X_3))$ and α is the .over-dispersion parameter. In order to check this assumption I firstly proceed by graphical assessment. The density plot and histogram of the response indicate a good fit with the Quasi-Poisson distribution.In order to be more sure about this, I evaluated the fit of the model by grouping the data based on population density in 5 equally sized groups, and then compared the observed and expected numbers of equids within each group using a Pearson chi-square statistic (4 degrees of freedom). This grouping allowed me to compare the observed and expected numbers of equids within each specific population density range separately, providing a more focused analysis.

Using the chi-squared distribution with the d.o.f., we find the cumulative probability of obtaining a test statistic as extreme as or more extreme than the observed test statistic. Finally, we subtract the cumulative probability from 1 to obtain the p-value and a significant result would suggest poor model fit. Since the resulting p-value is $0.22 > 0.05$, we conclude the model is well fitted.



2) Linear relation between log count and linear predictor: Once again, this assumption can be easily evaluated using graphical representation, particularly the Residuals vs Fitted Plot: The diagnostic plot of the residuals reveals a random scattering around the zero line, indicating that there is no apparent pattern. This observation suggests that the assumption of a linear relationship between the log count and the linear predictor is satisfied, as there are no systematic deviations from linearity.



3) The observations are independent: The independence assumption of the observations cannot be determined without additional information about the data collection process. Hence, we assume independence for the analysis.

Conclusions

Throughout our analysis, we aimed to develop a model that accurately predicts the rate of equine West Nile Virus (WNV) cases based on various predictors. Here is a recap of the steps we took to arrive at the final model:

1. Data Exploration: I performed exploratory data analysis to understand the distributions and relationships between variables. This involved visualizations, summary statistics, and identifying any outliers or unusual patterns. I identified PBR, Human_Density and Farms to be particularly important.
2. Model Selection: I started by fitting a Poisson regression model with equine cases as the response variable and all the provided predictors. However, we observed overdispersion, suggesting that the Poisson model might not be appropriate. Hence, I considered alternative models to address the overdispersion issue, quasi-Poisson regression. I then used model selection criteria, such as AIC, to backward compare the models and select the best-fitting one.
3. Final Model: After conducting the backward selection, we arrived at the final model:

$$Y|X_1 = PBR, X_2 = HD, X_3 = FARMS \sim Quasi - Pois(\mu, \alpha)$$

$\hat{\mu} = exp(\hat{\eta})$, $\hat{\eta} = -6.98 + 5329.84 \times P\hat{B}R + 0.001 \times \hat{HD} + 26.00 \times P\hat{B}R \times hatHD + log(F\hat{A}RMS)$

4. Model Assessment: I evaluated the assumptions of the final model, such as linearity and independence of residuals. I also conducted goodness-of-fit tests to ensure the model adequately represents the data.

To conclude, this study suggests that monitoring dead birds is a useful tool for predicting West Nile Virus (WNV) infection in both animals and humans. We adjusted for the effect of human population density on the association between dead bird counts and WNV cases in horses. Despite this adjustment, we still found a strong connection between WNV-related bird deaths and the risk of WNV infection in horses. These findings have important implications for public health as birds, which transmit WNV, can affect both equids and humans. Although there may be some biases and limitations in our study, the strong association supports the idea that WNV-related bird deaths can predict veterinary and human WNV infection. Monitoring dead birds may be more effective than mosquito surveillance for WNV prediction in South Carolina, where this study took place. Further research is needed to assess its value in predicting human WNV infection and to compare it with other surveillance methods.

References

[1] R.S. Roberts and I.M. Foppa (2006). "Prediction of Equine Risk of West Nile Virus Infection Based on Dead Bird Surveillance," Vector-Bourne and Zoonotic Diseases, Vol. 6, #1, pp. 1-6

[2] Kleiber, C., & Zeileis, A. (2021). AER: Applied Econometrics with R. Version 1.2-9.1. URL: https://CRAN.R-project.org/package=AER