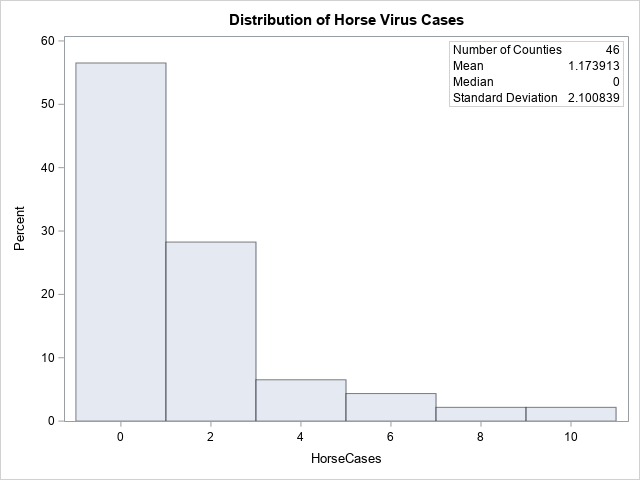
2018 Qualifying Exam

Question 3: Veterinary investigators are interested in determining whether the rate of horse virus can be predicted by examining the rate of bird virus and population density per county.

Part A:

Figure 1 shows a histogram of the counts of horse virus across all 46 counties of interest. It is important to note that 26 out of 46 counties (57%) report zero cases of horse virus. When all counties are included, the mean number of horse cases is 1.2. Of counties that report at least one count, the mean is 2.7.



Part B:

Here we use a Poisson model to evaluate whether horse virus rate is associated with bird virus rate. The outcome in the model is the log of the standardized horse virus rate, which is the number of horse virus counts in a given county divided by the number of farms in that county. The offset (log(N)) is the log of the number of farms in the county. The log-linear model allows us to model the expected value of the rate of horse virus given values of standardized bird virus rate and population density.

Model 1: log(mu/N) = beta0 + x1\*beta1 + x2\*beta2 + x1\*x2\*beta3

log(mu) = log(N) + beta0 + x1\*beta1 + x2\*beta2 + x1\*x2\*beta3

log(mu) = log of the standardized horse virus rate

log(N) = log of the number of farms

x1 = standardized bird virus rate

x2 = population density

beta0 = log of the expected standardized horse virus rate if the standardized bird virus rate and population density are zero

beta1 = increment to log of the standardized horse virus rate for a one unit increase in standardized bird virus rate

beta2 = increment to log of the standardized horse virus rate for one unit increase in population density

beta3 = increment to log of the standardized horse virus rate for a one unit increase in standardized bird virus rate by one unit increase in population density

In model 1 the expected rate of horse virus per farm count (mu/N) is the exponentiated value of matrix of variables **X** multiplied by their estimated regression coefficients, **B**. Table 1 shows the regression parameter estimates and associated standard errors.

| **Table 1: Analysis Of Maximum Likelihood Parameter Estimates, Poisson Model** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -6.9824 | 0.3664 | -7.7006 | -6.2642 | 363.10 | <.0001 |
| **BirdRate** | 1 | 5329.842 | 1595.642 | 2202.441 | 8457.242 | 11.16 | 0.0008 |
| **PopDensity** | 1 | 0.0006 | 0.0017 | -0.0027 | 0.0039 | 0.12 | 0.7242 |
| **BirdRate\*PopDensity** | 1 | 26.0050 | 12.3818 | 1.7371 | 50.2728 | 4.41 | 0.0357 |
| **Scale** | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 |  |  |

The deviance for this model is 62.01. If we divide this by the degrees of freedom, 42, we get 1.48. This value, which is far from 1, implies that overdispersion exists in this model. One of the drawbacks to modeling the horse virus rate using the Poisson distribution is that we must use the strong assumption that the mean of the distribution is equal to the variance. We see from the model’s regression diagnostics and the descriptive statistics that overdispersion exists, meaning the variance of the outcome is larger than the mean. Though the parameter estimates and their standard errors may not be reliable, we can still interpret the results of this preliminary model. We interpret the regression coefficient for the main effect of standardized bird virus rate to be the increment to the log of expected standardized horse virus rate for a one unit increase in standardized bird virus rate. If there is one additional bird per person per square mile with the virus we would expect 8.6 (log(5330)) additional horses per farm to have the virus in a given county. The population density has a nearly negligible effect on horse virus rate. The regression term for the interaction of standardized bird virus rate and population density (this is equivalent to non-standardized bird virus counts since standardized bird virus rate = bird virus counts/population density) indicates we would expect an increase of 3.3 (log(26.0)) cases of horse virus per farm for every one additional bird (not standardized for population density in the county) with the virus.

The model includes both the main effect of bird rate, and its interaction with population density (roughly translating to the standardized and non standardized bird counts) so we will perform a test of the null hypothesis that the regression parameters in the model for both of these terms are simultaneously equal to zero. The Wald chi square p value for this test with 2 degrees of freedom is 53.1 and has an associated p value of < 0.0001. Therefore we reject the null hypothesis of no association of bird count or rate with horse virus rate.

Part C:

Here we use a negative binomial model, which can account for overdispersion. The negative binomial distribution incorporates a dispersion parameter, so we do not need to assume the mean of the outcome distribution is equal to the variance. Table 2 shows the estimated regression parameters and associated standard errors for the negative binomial model.

| **Table 2: Analysis Of Maximum Likelihood Parameter Estimates, Neg Bin Model** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -6.9152 | 0.4230 | -7.7442 | -6.0863 | 267.32 | <.0001 |
| **BirdRate** | 1 | 4516.534 | 2474.606 | -333.604 | 9366.672 | 3.33 | 0.0680 |
| **PopDensity** | 1 | -0.0007 | 0.0024 | -0.0054 | 0.0040 | 0.08 | 0.7741 |
| **BirdRate\*PopDensity** | 1 | 44.6441 | 28.5155 | -11.2453 | 100.5335 | 2.45 | 0.1174 |
| **Dispersion** | 1 | 0.4182 | 0.3662 | 0.0752 | 2.3265 |  |  |

Table 2 shows the same parameter estimates and their standard errors for the negative binomial model. The dispersion parameter is 0.42. The strength of the association between horse rate and bird rate appears to have been attenuated but performing a similar test of the null hypothesis for both main effect and interaction term, the Wald chi square test statistic with 2 degrees of freedom is 20.4, with an associated p value of <0.0001. Using this model we interpret the regression parameters as follows. If there is one additional bird per person per square mile with the virus we would expect 8.4 (log(4512)) additional horses per farm to have the virus in a given county. The population density has a nearly negligible effect on horse virus rate. We would expect an increase of 3.8 (log(44.6)) cases of horse virus per farm for every one additional bird (not standardized for population density in the county) with the virus.

I am not sure if parts D and E are correct but here is the SAS resource:

https://support.sas.com/rnd/app/stat/examples/GENMODZIP/roots.htm

Part D:

As we saw from the descriptive statistics, the counts of horse virus tend towards zero. There are 26 counties with zero counts, and 20 counties with non-zero counts. We hypothesize that the counts of horse virus across counties follow a hybrid distribution. First, we assume an underlying binomial distribution, where a county either experiences horse virus cases or it does not. Using the sample of the data collected, assuming this is a representative sample, we estimate that 57% of all counties will not experience cases of horse virus. Of the counties that do experience horse virus, we can assume a Poisson distribution of those counts.

Model 3 is a zero-inflated Poisson model. It does not account for overdispersion, but does account for the excess number of zero counts in the distribution of horse virus. Table 3 shows the parameter estimates and associated standard errors for model 3.

| **Table 3: Analysis Of Maximum Likelihood Parameter Estimates, ZIP** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -6.3740 | 0.4828 | -7.3202 | -5.4278 | 174.32 | <.0001 |
| **BirdRate** | 1 | 3845.254 | 1854.999 | 209.5225 | 7480.986 | 4.30 | 0.0382 |
| **PopDensity** | 1 | 0.0001 | 0.0019 | -0.0036 | 0.0038 | 0.00 | 0.9566 |
| **BirdRate\*PopDensity** | 1 | 21.4189 | 13.2298 | -4.5110 | 47.3489 | 2.62 | 0.1054 |
| **Scale** | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 |  |  |

The strength of the association between horse rate and bird rate appears to have been attenuated in this zero-inflated Poisson model, but performing a test of the null hypothesis for both main effect and interaction term, the Wald chi square test statistic with 2 degrees of freedom is 20.4, with an associated p value of <0.0001. Using this model we interpret the regression parameters as follows. If there is one additional bird per person per square mile with the virus we would expect 8.3 (log(3845)) additional horses per farm to have the virus in a given county. The population density has a nearly negligible effect on horse virus rate. We would expect an increase of 3.3 (log(21.4)) cases of horse virus per farm for every one additional bird (not standardized for population density in the county) with the virus.

Part E:

Model 4 accounts for both overdispersion and zero-inflation. Table 4 shows the parameter estimates and associated standard errors for model 4. The dispersion parameter in this model is 0.12, lower than Model 2, indicating that much of the overdispersion was due to the data being zero-inflated. Yet, using a zero-inflated model does not fully account for the overdispersion in this data.

| **Table 4: Analysis Of Maximum Likelihood Parameter Estimates, ZINB** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -6.3787 | 0.5150 | -7.3880 | -5.3694 | 153.43 | <.0001 |
| **BirdRate** | 1 | 3604.133 | 2173.857 | -656.548 | 7864.813 | 2.75 | 0.0973 |
| **PopDensity** | 1 | -0.0002 | 0.0022 | -0.0045 | 0.0041 | 0.01 | 0.9225 |
| **BirdRate\*PopDensity** | 1 | 26.6289 | 20.0907 | -12.7481 | 66.0059 | 1.76 | 0.1850 |
| **Dispersion** | 1 | 0.1295 | 0.2364 | 0.0036 | 4.6322 |  |  |

The strength of the association between horse rate and bird rate appears to have been attenuated in this zero-inflated Negative Binomial model compared to the Poisson, but performing a test of the null hypothesis for both main effect and interaction term, the Wald chi square test statistic with 2 degrees of freedom is 12.3, with an associated p value of 0.022, which is still statistically significant at a p<0.05 level. Using this model we interpret the regression parameters as follows. If there is one additional bird per person per square mile with the virus we would expect 8.2 (log(3845)) additional horses per farm to have the virus in a given county. The population density has a nearly negligible effect on horse virus rate. We would expect an increase of 3.3 (log(26.6)) cases of horse virus per farm for every one additional bird (not standardized for population density in the county) with the virus.

Part F:

I prefer model 4, which accounts for both overdispersion, caused by, and independent of, zero-inflation. In all of the models we find a statistically significant association between rate of bird virus and rate of horse virus, though accounting for overdispersion attenuates the strength of this association. The parameter estimates are not wildly different between each model, and the interpretations are similar in terms of the expected rate of the virus in horses when the rate of infected birds in the county is known. For every additional infected bird per person per square mile, we would expect anywhere from 3.3 to 3.8 horses per farm to contract the virus. The main difference between these models is the standard errors associated with the point estimates. Using the estimate and standard errors from the zero-inflated negative binomial model, our confidence intervals for the point estimate would be quite large and include values less than zero. Using the zero-inflated Poisson model we would have confidence intervals that did not span zero, but this inference may not be valid due to additional over-dispersion not accounted for by zero-inflation.

Part G:

[Not written, but general idea is in part F.]