STAT 511 – Final Exam

Analysis of Elk feeding patterns at Yellowstone national park

Vignesh Ramchandran (vzr124@psu.edu)

1.) Problem Statement and Objectives

From 1984 to the present, Elk have been monitored and managed in the area surrounding Yellowstone national park (YNP). Elk use different parts of YNP at different times of the year, often spending more time in the higher elevations during the warm summer months, and spending more time at lower elevation during the cold winter months. Feeding programs exist in which, during cold winters, hay or other feed is left in places where Elk are likely to encounter it. Observers went to specific locations at different times of the year, and watched for Elk for 2 hours at dawn and dusk each day. Some observers stayed at sites for only one day, while others stayed at the site for up to 4 days. A dataset containing the counts of observed Elk (response) at specific locations is available along with requisite seasonal information to characterize the surrounding environment throughout the year.

There are 3 goals to this study:

1.) Identifying the relationship between the average numbers of Elk observed at a site and the various covariates.
2.) Determining the season when the predicted Elk counts are least accurate to devote more resources during that period and improve predictive accuracy.
3.) Compare 2 potential feeding sites for the Elks during winter to ensure feed is left at the site where more Elk are likely to visit.

2.) Analysis

2.1) Characteristics of interest for a site where an Elk has been sighted

There are 6 features at each site which are of keen interest in the study:

a) 'obs.day' – The number of days the observer stayed at the site.
b) 'season' – A categorical variable with 3 levels indicative of the season when Elk was sighted at the site namely Winter" (W), "Summer or Fall" (SF) and "Parturition" (P).
c) 'perc.forest' – Proportion of the observation site that is forested.
d) 'perc.open.water' – Proportion of the observation site that is open water (lakes/rivers).
e) 'perc.developed' – Proportion of the observation site that is developed (roads/buildings).

A Poisson regression model is best approach for the study since the response variable is count of the number of Elks sighted. All modeling and calculations are performed in R.

The model we are trying to fit to the data is of the form, $Y = Pois(\lambda)$ (where, Y is the count of the response and λ is the rate parameter).

The linear Poisson regression model is expressed as follows:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

For the average number of Elk observed at a site, the model between the count (response) and the covariates is:

$$log\{E(num.seen)\} = \beta_0 + \beta_1 obs.days + \beta_2 season + \beta_3 elev + \beta_4 perc.forest + \beta_5 perc.open.water + \beta_6 perc.developed$$

2.2) Relationship between Average number of Elk sighted and covariates

A Poisson regression model is fitted to the data using 'num.seen' as the count response and all 6 predictor variables as the response (Model 1). Table 1 shows the coefficients and p-values for the above model which enables us to estimate which predictors are significant.

Table 1: Coefficients and p-values for Model 1

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.19728	0.03188	6.189	6.06E-10
Season:SF	-0.10229	0.02335	-4.381	1.18E-05
Season:W	-0.18798	0.02784	-6.753	1.45E-11
elev	-0.15637	0.01702	-9.189	< 2e-16
Perc.forest	0.01681	0.03169	0.531	0.596
Perc.open.water	-0.87946	0.20772	-4.234	2.30E-05
Perc.developed	-1.08615	0.19165	-5.667	1.45E-08
Obs.days	-0.01622	0.01018	-1.594	0.111

Figure 1 shows the residuals, which were plotted against the fitted values with a clear trend observable between them, the residuals being separated on the basis of the response count. The QQ plot also did not adhere to the assumptions of normality. The partial residual plots were constructed to validate the presence of a linear relation between the response and the predictors. The partial residuals all showed a strong indication of a linear relationship.

The test for outliers was conducted using "outlierTest" from the "car" package in R. It revealed several outliers/influential points within the model, the influential observations were excluded from the original dataset before proceeding. The check for multicollinearity in the model through VIF values revealed that none of the covariates were correlated to one another.

As observable from Table 1, the predictors "obs.days" and 'perc.forest" are not significant at the 95% level of confidence and are hence discarded from the model. A check for all possible 2-way interactions between the remaining predictor variables was performed and realized that only the interaction between "elev" and "perc.developed" was significant at the 95% level of confidence and hence it was added to the model.

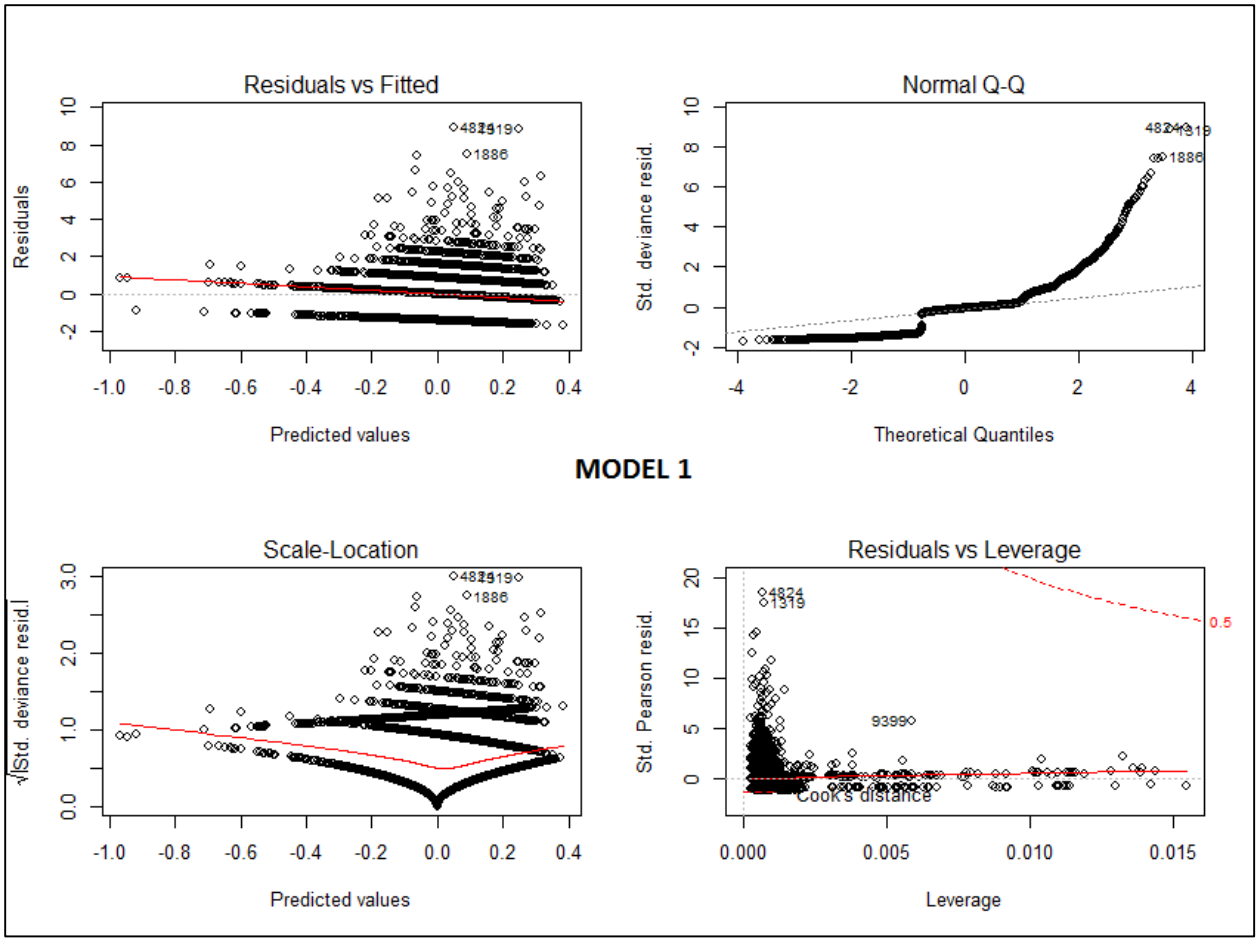

Figure 1: Residuals and QQ diagnostics for Model 1

The final model between the response and the predictors is (Model 2):

$$\log\{\mathbf{E}(\mathbf{num.seen})\} = \beta_0 + \beta_1 \mathbf{seasonSF} + \beta_2 \mathbf{seasonW} + \beta_3 \mathbf{elev} + \beta_4 \mathbf{perc.open.water} + \beta_5 \mathbf{perc.developed} + \beta_6 \mathbf{elev} * \mathbf{perc.developed}$$

Figure 2 and Table 2 describe the residuals, Q-Q plots, coefficients and p-values for the final model. All the terms are significant at the 95% confidence level as seen in Table 2. It can also be observed that the residuals and QQ plot retain the form similar to what was observed for Model 1 but the number of levels in the residuals vs fitted values has decreased due to the elimination of outliers.

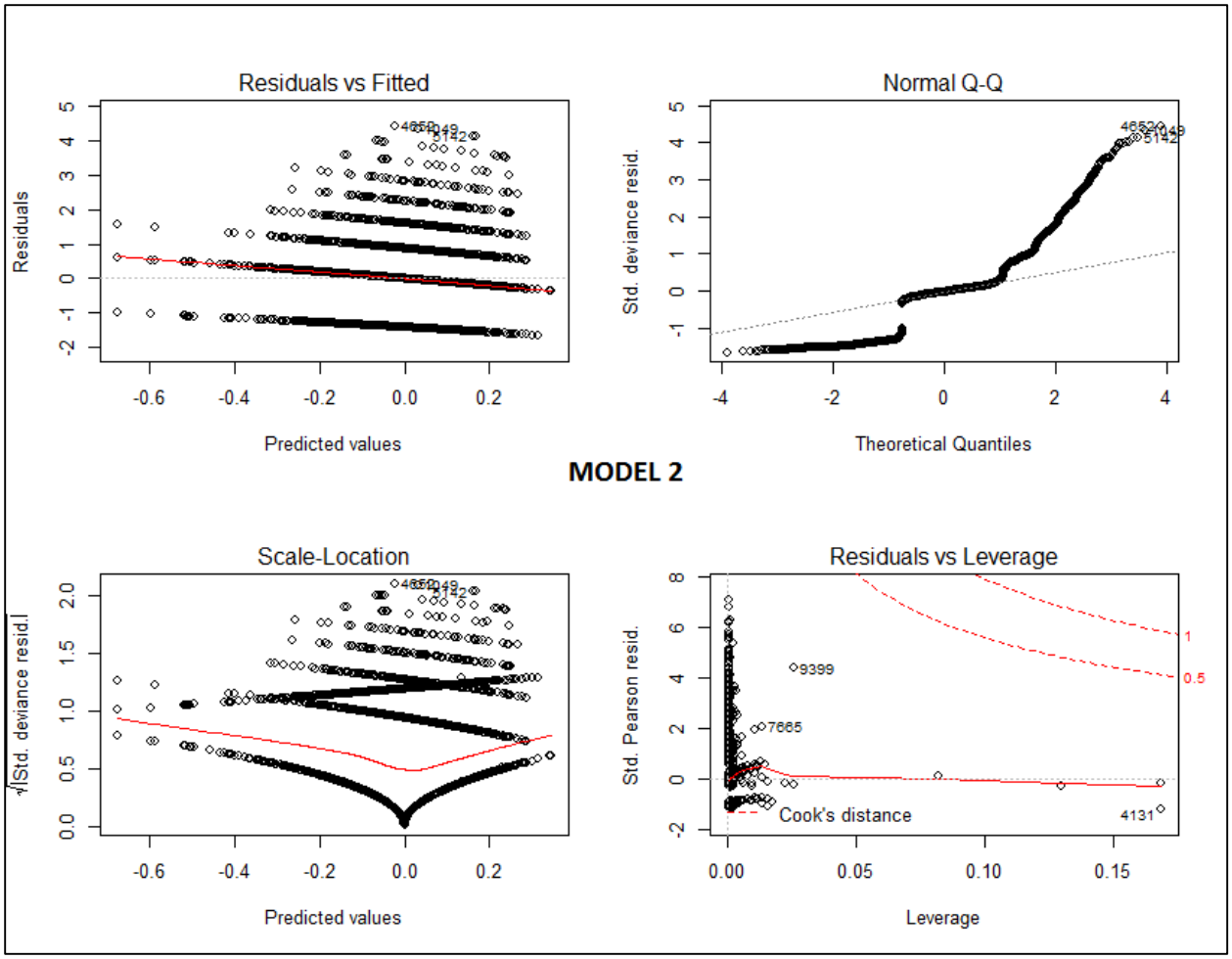

Figure 2: Residuals and QQ diagnostics for Model 2

Table 2: Coefficients and p-values for Model 2

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.11939	0.01687	7.078	1.46E-12
Season:SF	-0.07964	0.02369	-3.362	0.000774
Season:W	-0.16779	0.02826	-5.938	2.89E-09
Elev	-0.15375	0.01641	-9.371	< 2e-16
Perc.open.water	-0.81967	0.20693	-3.961	7.46E-05
Perc.developed	-1.73186	0.34364	-5.04	4.66E-07
Elev*Perc.developed	-1.56635	0.56623	-2.766	0.00567

To ascertain that the distribution of the residuals is consistent, we used coefficients from Model 2 to simulate a Poisson distribution and compare the residuals in both cases for similarity. Figure 3 is the simulated residuals obtained using Model 2. It can be seen that they are similar qualitatively and hence, the model used is acceptable.

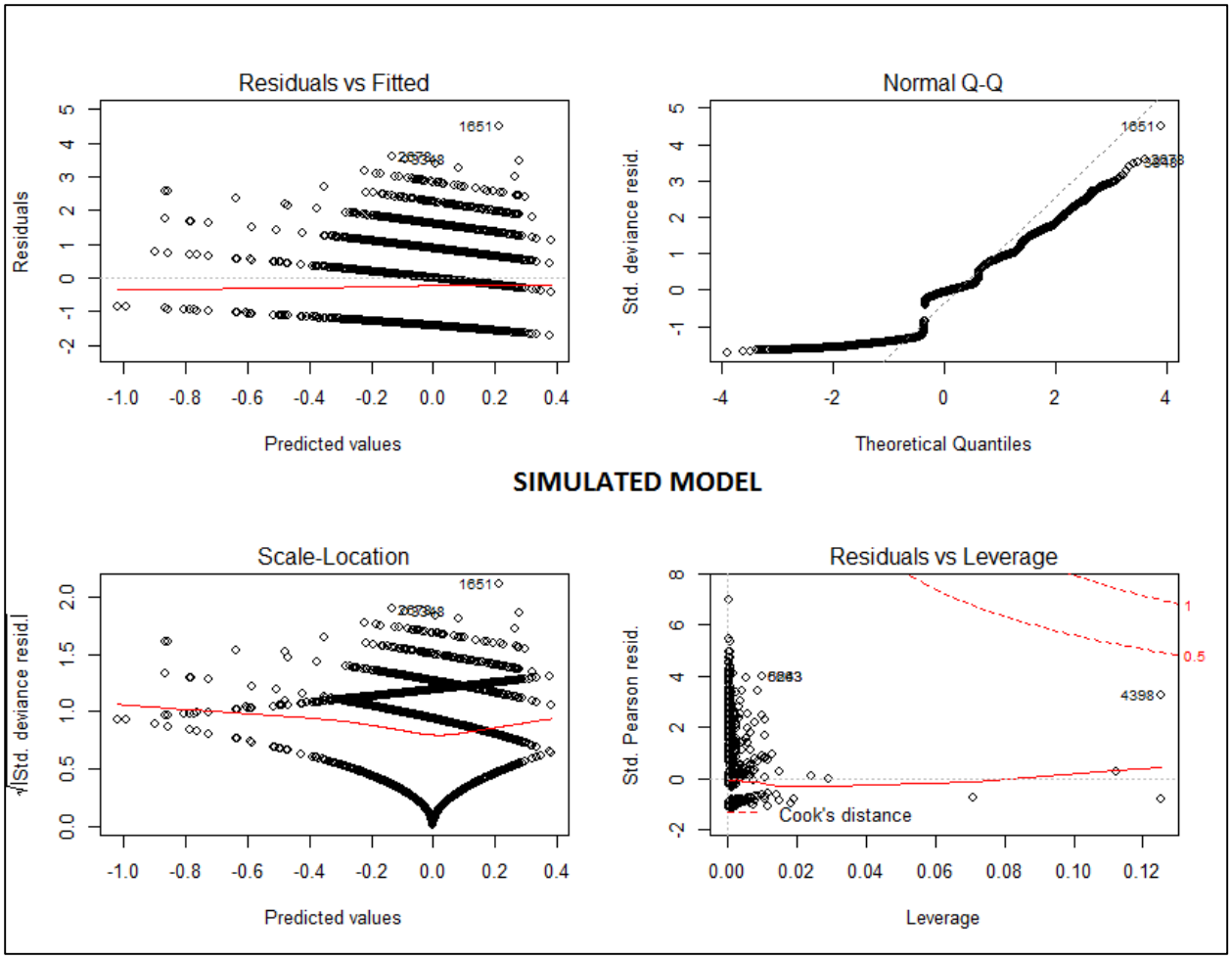

Figure 3: Residuals and QQ diagnostics simulated using Model 2 to validate residuals

The final model between the average number of Elk observed at a site and the predictors is:

$$\log(E(num.seen)) = 0.1194 - 0.079(seasonSF) - 0.167(seasonW) - 0.154(Elev) - 0.82(Perc.open.water) - 1.732(Perc.developed) - 1.566(Elev * Perc.developed)$$

The average number of Elk observed at each site is obtained is taking exponent of the linear equation above. To understand the relationship between "E(num.seen)" (average number of Elk seen) and the predictors Figure 4 shows the plots between them to provide clarity.

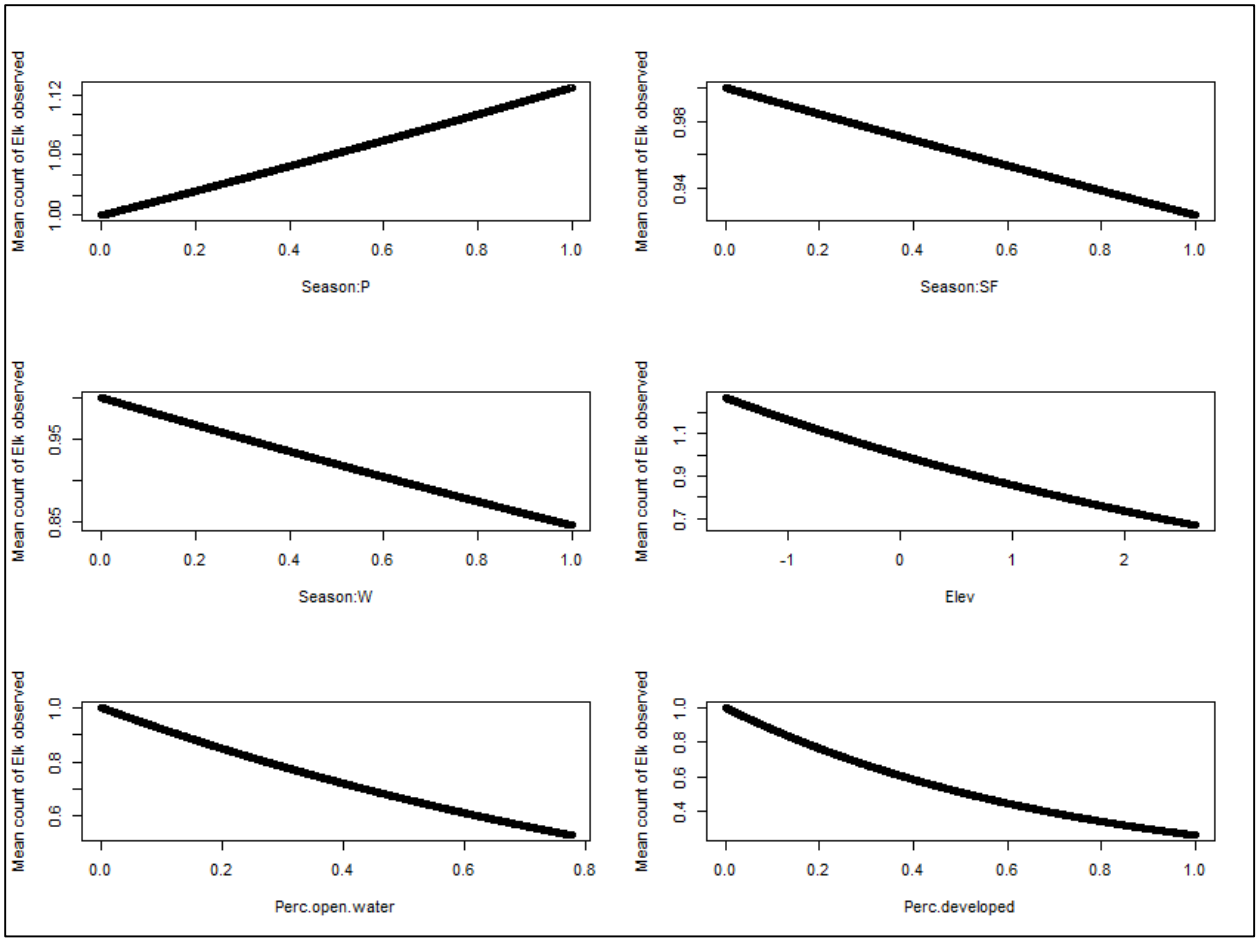

Figure 4: Relationship between Average number of Elk observed and the predictors

It can be observed from the above figure that, the only possible relationship is between the average count observed and the intercept (seasonP), which indicates that the number of Elk that can be sighted is higher by about ~12% during the parturition period between spring and summer. This would indicate that Elk are averse to winters (decrease by ~8%) and the hot summers/fall seasons (decrease by ~17%).

The response is negatively related to all the other predictors. The relation between the numbers sighted and the other predictors is slightly intuitive because with increase in elevation difficult in breathing for the Elk may decrease the numbers (~15%) that can be seen at high elevations. The increase in perc.open.water would mean lesser ground covered in grass to feed on for them and increase in "perc.developed" would mean more humans in the vicinity that would make them uncomfortable.

2.3) Determination of the season when Elk sighting prediction is the lowest

Determining which season from the given 3 has the least accuracy of prediction implies the season during which the mean squared prediction error (MSPE) is the highest. To determine the MSPE for the data on the basis of the seasons we separate/partition the data into 3 parts, each part corresponding to one of the seasons ("W", "P" and "SF"). Since, the seasons are constant for each of the 3 partitions resp. we can exclude the "season" variable from the model and proceed to estimate the best linear model between the average numbers of Elk sighted and the predictors.

Since we are employing the best model approach using the "step" function, all the possible 2 way interactions between the predictors were included for a more robust model. The selection of the best model was done through AIC based variable elimination.

Using the best model obtained from the "step" function we can determine the predicted mean (μ') using the partitioned data as the input, the predicted mean can be used to generate a Poisson distribution of the count (response) for the number of Elks sighted in the area.

The MSPE is then calculated as the average of the squared difference between the observed and the simulated Elk counts. The "best" model used is the same for all the 3 partitions of the dataset for a more consistent result.

The best model obtained for this case is given by Model 3:

$$\begin{aligned}\log(E(num.seen) &= 0.032 - 0.278(Elev) - 0.07(perc.forest) - 0.378(perc.open.water) \\ &- 0.78(Perc.developed) + 0.426(Elev * Perc.forest) \\ &- 1.54(Perc.forest * Perc.open.water) - 6.93(perc.forest \\ &* perc.developed)\end{aligned}$$

Using the above model we estimated the predicted means for the 3 partitions respectively and proceeded to simulate the Poisson distributions for the respective data before determining the MSPE. Table 3 shows the comparison of the MSPE for the 3 seasons.

Table 3: Comparison of MSPE across the 3 seasons

Season	MSPE
Season:P	2.034696
Season:SF	1.687698
Season:W	1.587537

As seen from Table 3, we can conclude that the "Parturition" season between late spring and summer has the highest MSPE amongst the 3 seasons. Hence, more resources can be assigned and dedicated towards getting more accurate observations during that particular season and improve the predictive accuracy.

2.4) Comparison of sites for dropping off feed for the Elks

During the winter months due to most of the grass dying out from the cold, there is provision to drop off hay for the Elks to feed on during the winter months. There 2 sites that have been earmarked as potential sites where the hay can be dropped off.

The sites have the following characteristics:

Site 1: (Elev = -0.954, perc.forest = 0, perc.open.water = 0, perc.developed = 0)

Site 2: (Elev = 0.209, perc.forest = 0.222, perc.open.water = 0.01, perc.developed = 0)

Using the best model solution (Model 3) from the previous section we can determine the predicted mean for both sites.

The predicted means for sites 1 and 2 are:

Site 1: Predicted mean = 1.35

Site 2: Predicted mean = 0.97

Using the predicted means for both the sites, we can simulate Poisson distributions for the count of the number of Elks sighted in the observed area. For a comprehensive attempt at simulating the distribution we simulate 10000 observations at both sites using their respective predicted means and then evaluate the mean count of the number of Elks observed in the given sites.

The mean number of Elks observed at Site 1 are: 1.35

The mean number of Elks observed at Site 2 are: 0.99

Based on the above values, we can conclude that Site 1 is definitely the better choice for dropping off the hay as the mean number of Elks visiting Site 1 are definitely higher than Site 2 and Site 1 is more easily accessible than Site 2.

3.) Conclusions

The analysis of the Elk feeding patterns at Yellowstone national park revealed the dependency of the number of the Elks visiting a site on its environment and climate. The Elks prefer to visit a site when the climate is relatively cool ("Parturition" season) and also more Elks can be spotted at lower elevations than higher. The number of Elks in an area also increases with a decrease in the urban infrastructure in that area, an increase in the grazing area available that is not covered by open water bodies are also likely locations Elks may be spotted in.

The "Parturition" season was also deemed as the season where the predictive accuracy of the number of Elks observed was least and hence additional resources should be dedicated towards improving the accuracy of the predictive potential during the season the Elks prefer to feed.

Lastly, the sites compared to be designated as feeding sites where hay shall be deposited for the Elks to feed on was assigned to the site where the elevation was lower and there is lots of open space and area for the Elks to roam around and graze in (as evidenced by the percentage of forest cover, open water bodies and urban development all being 0). An additional benefit to the area was its ease access during the deposition phase.

References

1.) "Regression: Models, Methods and Applications", Fahrmeir, L., Kneib, T., Lang, S., Marx, B.