Homework 12*

Advanced Methods for Data Analysis (36-402)

Due Friday May 3, 2019, at 3:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as *code*, and knitted or scanned/merged pdf or html as *writeup*). Make sure that everything you submit is readable.

Elephants in Kenya

A population of elephants in Kenya was studied for 8 years. The theory is that the success of mating increases as a function of age. The data consist of two variables, X = age (at the beginning of the study) and Y = number of successful matings over the duration of the study. What distribution is appropriate for these data? Analyze the Elephant data to predict Y based on X using a generalized linear model. To load the data, you can run the following commands

```
library(Sleuth3)
dat <- case2201
```

Part a

Produce 1D and 2D EDA plots with legible X and Y-axes, and a caption. Write a paragraph summarizing these plots. Describe the relationship between Y and X. Fit a Poisson regression model, call this Model 1. Why is this a sensible choice? Write down the fitted model, and add the fitted line on the 2D plot of the data. Describe the findings of your regression model in a few sentences. This should include a CI for the slope coefficient, and an interpretation of what that coefficient measures.

Part b

Perform a residual analysis (e.g. look at deviance residuals. See Demo_11.1_GLM and lecture notes!)

Part c

Provide the fitted rate of successful mating **per year** as a function of age, and provide the variance-covariance matrix of the regression coefficients (`vcov` command in `R`). Fit a new Poisson model, this time using `offset=8`. This will yield an yearly rate as opposed to a rate per 8 years.

Part d

Provide a point estimate with associated 95% confidence interval for the **yearly rate** of successful mating of animals that are 30 year old at the start of the study. Think carefully: are we asking you to compute a prediction interval or a confidence interval?

First, write down the parameter of interest, which should have the form $\theta = T + xU$, where T and U are unknown and need to be estimated. Then, proceed in two (or three) ways:

*Not for sharing, even after the class is over.

1. Use the `predict.glm` functions to get the prediction on the response scale, with its standard error, and set a normal confidence interval.
2. Start with the MLEs of the β's along with their variance-covariance matrix (part c). Then, obtain the MLE of the relevant linear combination of β's, along with an estimate of its standard error. Recall, $\text{SE}(c\hat{\beta}) = |c|\sqrt{\text{Var}(\hat{\beta})}$, for c constant. Then, set a 95% CI for that linear combination based on the normal approximation. Invert that CI back onto the desired scale (that is, apply the inverse link function to the CI). Does this match the previous confidence interval?
3. Bonus (+10 points, you should solve this easily given the hint for part g in HW 11). Consider the MLE of the parameter of interest to be a function of the MLEs of the β's. Linearize that function by using a multivariate first order Taylor expansion around the true β's. Then, use the variance-covariance matrix of the β's to obtain the variance of the linear expansion of the estimate of the parameter of interest. Argue that the estimate is approximately normal and thus set a normal-based confidence interval for the parameter of interest. Does this match the previous confidence intervals?

Part e

How old must animals be on average (at the start of the study) to have at least a 50% chance of one successful mating per year? That is, we are interest in the value of `Age` such that $\mathbb{P}(\text{matings in 1 year} = 0) \leq 0.5$. Recall that in our model, we have assumed that the number of matings per Year given Age is distributed Poisson(λ_2(Age)). Through the use of the log link function, we have modeled $\log \lambda_2(\text{Age}) = \beta_0 + \text{Age}\beta_1$. Now, we can proceed as follows:

1. First, if $U \sim \text{Poisson}(\lambda)$, what is $\mathbb{P}(U = 0)$ equal to?

   Hint: recall that $\mathbb{P}(U = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
2. Use the result above to argue that $\lambda_2(\text{Age}) \geq \log 2$ implies $\mathbb{P}(\text{matings in 1 year} = 0) \leq 0.5$.
3. Finally, re-arrange the inequality above to find that the parameter of interest is $\theta := \frac{\log(\log 2) - \beta_0}{\beta_1}$.
4. Propose an estimator for θ (e.g. it could be the plug-in estimator, which places $\hat{\cdot}$ on unknown quantities). Give the point-estimate.
5. (Bonus, + 10 points) Use the first-order Taylor expansion trick to construct a confidence interval for θ.

Part f

Now suppose that, instead of fitting a Poisson regression, you fit a linear regression using the log of the response as the outcome variable. Can you think of any problems with this approach? How would this approach differ from a Poisson model? There is no need to give lengthy answers, but your answers should hit the crux of the matter.