Lab 8 - GLM Midterm Review

In this lab we're going to go over a review of doing GLM analysis in `R` as a preparation for the midterm.

First off, the general steps you want to follow for your statistical analysis are as follows:

1. State research questions and the statistical methods (e.g. logistic regression, poisson regression) you intend to use to analyze your data for informing your research.
2. Choose the model you intend to use to answer your research question, and verify that that modelling assumptions hold for inference. (It is also good to state any potential worries you may have about any modelling assumptions)
3. Run hypothesis testing relevant to your initial research questions. State conclusions in the context of the problem.
4. Make final conclusions, reiterating the model you chose, any potential worries for using this data for inference, the results you inferred about the population, and any other relevant information corresponding to your research question.

Now we'll go over relevant things to know about completing these steps.

Writing out your model

When writing out your model, make sure it's in the context of your data (e.g. stating what each variable corresponds to in your data set)! Below we write the models for the general case.

Binomial/Logistic Regression Case

The model we fit is

$Y_i|X_i \sim Binomial(n_i, p_i)$ where the probability of success p_i has the following relationship with the predictors: $p_i = logit(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})$, where $logit$ is the logistic function from class.

The β parameters, the counts n_i, and the predictors X_i are all constant variables. n_i is the largest count for each observation.

Y_i is a random variable with the binomial distribution stated as form above.

We assume Y_i are independent.

Poisson Regression Case

The model we fit is

$Y_i|X_i \sim Poisson(\lambda_i)$ where the rate or the mean of Y_i has the following relationship with the predictors: $\lambda_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})$.

The β parameters, and the predictors X_i are all constant variables.

Y_i is a random variable with the Poisson distribution stated as form above.

We assume Y_i are independent.

Choosing between the Binomial and Poisson case

When you get a data set and you need to decide if a binomial or poisson regression is more appropriate, consider the following:

1. Is your data a count of some event occurring? (If not, neither binomial or poisson is appropriate)
2. Is there an obvious upper bound on the number of events that can occur? (If not, a Poisson regression is probably more appropriate)
3. Is the maximum number of events fixed, but the events are rare events (e.g. the probability of an event is less than 5%?)? (If yes, then either a Poisson regression of Binomial regression can be used, but a Poisson regression doesn't need to specify the n_i parameter)
4. Is the maximum number of events fixed, and the event occurs relatively frequently? (If yes, a binomial regression is probably more appropriate).

Model diagnostics

To do model diagnostics, comment on independence as a reasonable assumption, look at cook's distance to see if outliers are driving the conclusions of the inference, and run a deviance goodness-of-fit test.

Note that sometimes you can plot out your data to see if the logistic or exponential rate assumptions are reasonable, although when your response is only 0 and 1, this is not possible (see HW4 Q2 solutions).

Interpretting Coefficients

Note that interpreting coefficients is different for each model we fit, and the differences are quite subtle.

Binomial/Logistic Case

β_1 is the increase in the log odds for every unit increase in X_1 *when holding all other predictors constant.*

$\exp(\beta_1)$ is the *factor increase* in the *odds* for every unit increase in X_1 when holding all other predictors.

(A factor increase, means it is the amount we multiply the odds by. So the odds will go from $ODDS$ to $ODDS \times \exp(\beta_1)$)

Note that any calculated confidence interval that is calculated should also keep these interpretations in mind.

Poisson Case

β_1 is the increase in the *log rate (or log mean)* for every unit increase in X_1 *when holding all other predictors constant.*

$\exp(\beta_1)$ is the *factor increase* in the rate (or mean) for every unit increase in X_1 when holding all other predictors.

Again: any calculated confidence interval that is calculated should also keep these interpretations in mind.

Testing 1 parameter

$H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.

(Note we can have one-sided tests, also the null mean doesn't need to be 0)

For both model types, use the estimate and standard error from the summary output of the GLM to test 1 parameter. Use a normal distribution for calculating the pvalues and critical values for the confidence interval.

Note that the normal distribution is only appropriate when not accounting for over-dispersion.

Testing multiple parameters

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$ vs H_a : at least one of the slopes are not equal.

If answering your research question corresponds to testing multiple parameters, be **sure** to use the `anova` function with the `test = 'LRT'` to do a likelihood ratio test.

E.g. if you have a categorical variable with >2 levels and are interested in a testing if your data is significantly explained by this predictor, then you need to use a likelihood ratio test!! Very important.

Accounting for Overdispersion

If dependence is a potential issue in your dataset, one can also directly test for overdispersion (see lecture slide 115). If this test is significant, one can account for overdispersion in their model (although if the pvalue for this test is <0.001, this may signify other assumptions in your model do not hold).

Account for overdispersion by using the `quasibinomial` or `quasipoisson` model. Note that the summary table now uses the t-distribution with the residual deviance degrees of freedom (see summary table for residual deviance degrees of freedom).

Also, for the overdispersed models, with the `anova` function use `test = 'F'` to account for overdispersion when looking at multiple parameters.

Offsets in Poisson models

Offsets can be used in any glm model, although they tend to be most intuitable in Poisson models.

If your mean rate should always be proportional to a variable OFF, then when specifying the glm, you should set `offset = log(OFF)`.

See lecture slides 118 and 119 for examples.

AIC and BIC for model selection

The `bestglm` function can be useful for model selection!

Lastly!

Make sure to read the questions in full! On the last prelim, the quesiton asked to look at polynomial terms, although many people didn't include that in their solutions!

If you have any questions make sure you ask before the midterm is out on Wednedsay morning! Good luck!