Generalized Linear Models Notes

Dustin Leatherman

April 17, 2020

Contents

1 Types of Regression

1.1 Logistic

A measure of the relationship between categorical data using a Binary distribution. (Bernoulli for Logistic, Gaussian for OLS Regression).

$E(y) = \frac{1}{1+e^{-(a+\beta_k x_k)}}$

- Deviance is used to measure lack-of-fit. (instead of Sum of Squares)
- Likelihood Ratio Test (LRT) also used

1.1.1 Examples

- Predicting whether a political candidate wins an election
- Predicting Admission into a Program

1.2 Poisson

Used for Count data. Y refers to the **number of occurrences** of an event and is assumed to have a Poisson Distribution:

$P(Y = k|x_1, x_2, ..., x_m) = \frac{e^{-\mu}\mu^k}{k!}$

Log Link: $y = exp(\beta_0 + \Sigma\beta_i x_i)$

The model is used for goodness-of-fit tests. Recall the following about the Poisson Distribution:

$$\begin{aligned} var(Y) &= \mu \\ E(Y) &= \mu \end{aligned} \tag{1}$$

1.2.1 Mean > Variance

- Use Logistic Regression to adjust standard errors
- Use Negative Binomial Regression

1.2.2 Over dispersion

This occurs when the observed variance is larger than the assumed (theoretical) variance.

1.2.3 Interpretation

Exponentiated coefficients are multiplicative analgous to odds ratios but called *incidence rate ratios.*

For example, if $e^b = 2$, then rate doubles for each unit change in x. If $e^b = 0.5$, then rate halves. If $e^b = 5$ and x decreases, then $\frac{1}{5} = 0.2$

1.2.4 Examples

- Number of people in line in front of you at the grocery store
- Number of awards earned by students at one High school

1.3 Negative Binomial

Used for Count Data. Describes probabilities of the occurrence of whole numbers greater than or equal to zero. Y is the number of times an event has occurred. One parameterization is:

$P(y) = P(Y|y) = \frac{\Gamma(y+\frac{1}{a})}{\Gamma(y+1)\Gamma(\frac{1}{a})}(\frac{1}{1+a\mu})^{\frac{1}{a}}$, $\mu > 0, a > 0$

a is the heterogeneity parameter.

The traditional model is: $log(\mu) = \beta_0 + \Sigma\beta_i x_i$

The response variable may be over or under-dispersed. The model models the log of expected count as a function of predictors.

1.3.1 Model Fit Statistics

- Log-likelihood
- deviance
- Pearson chi-square dispersion

- AIC
- BIC

1.3.2 Interpretation

For one unit of change in the predictor var, the difference in the logs of expected counts in the response is expected to change by β_i, given that all other predictors are held constant.

1.3.3 Examples

- Attendance behavior based on enrolled program and a standardized test score
- Number of hospital visits by seniors based on characteristics and types of health plans.

1.4 Gamma

Used for continuous, positive, right-skewed data where the variance is nearly constant.

Log Link: $\mu = exp(\beta_0 + \Sigma\beta_i x_i)$

1.4.1 Examples

- Study of damage done to cars in an insurance claim/

1.5 Zero-Inflated Poisson (ZIP)

Used to model count data with an excessive amount of zeros. There are two parts to this model.

Both Poisson Count and Logit Zero models should have good predictors. They are not required to have the *same* predictors.

This should be used for large sample sizes.

1.5.1 Poisson Count model

Generates counts, some of which may be zero.

$P(y_j = h_i) = (1 - \pi)\frac{\lambda^{h_i} e^{-\lambda}}{h_i!}$, $h_i \geq 1$

- y_j: any non-negative integer value

- λ_i: expected Poisson count for the ith individual
- π: probability of extra zeros

$E(Y) = (1-\pi)\lambda \; var(Y) = \lambda(1-\pi)(1-\lambda\pi)$

1.5.2 Logit "Zero" Model

Used for predicting excess zeros. This is a binary distribution that generates zeros.

$P(y_i = 0) = \pi + (-\pi)e^{-\lambda}$

Issues that can occur

- Perfect Prediction
- Separation or Partial Separation

1.6 Zero-Inflated Negative Binomial

Similar to ZIP. This is used for over-dispersed count response variables. The Count model in this case is **Negative Binomial** instead of Poisson.

2 Exponential Family of Distributions

A distribution belongs to the exponential family if it can be written in the following form:

$$\begin{aligned} f(y:\theta) =& s(y)t(\theta)e^{a(y)b(\theta)} \\ =& exp[a(y)b(\theta) + c(\theta) + d(y)] \end{aligned} \tag{2}$$

Distribution	Natural Parameter	c	d
Poisson	$\log\theta$	$-\theta$	$-logy!$
Normal	$\frac{\mu}{\sigma^2}$	$\frac{-\mu^2}{2\sigma^2} - 0.5log(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$
Binomial	$log(\frac{\pi}{1-\pi})$	$nlog(1-\pi)$	$log(nchoosey)$

2.1 Properties

$$\begin{aligned} E(a(Y)) =& -c'(\theta)/b'(\theta) \\ var(a(Y)) =& \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \end{aligned} \tag{3}$$

2.1.1 Score Statistic

$$\begin{aligned} l(\theta;y) =& a(y)b(\theta) + c(\theta) + d(y) \\ U(\theta;y) =& \frac{dl(\theta;y)}{d\theta} = a(y)b'(\theta) + c'(\theta) \\ E(U) =& b'(\theta)E[a(Y)] + c'(\theta) \\ =& b'(\theta)\frac{-c'(\theta)}{b'(\theta)} + c'(\theta) = 0 \\ var(U) =& [b'(\theta)^2]var[a(Y)] \\ =& b''\frac{(\theta)c'(\theta)}{b'(\theta)} - c''(\theta) \\ var(U) =& E(U^2) = -E(U') \end{aligned} \tag{4}$$

- U: A random variable called the **Score Statistic**
- var(U): Information Matrix

3 Outliers & Influential Obs

3.1 Explanatory Variable Pattern (EVP)

Sometimes, converting bernoulli random variables to binomial is helpful for running goodness-of-fit measures and residuals.

This format has one row for each unique set of explanatory variables. Suppose there are 6 observations with a bernoulli RV value of 1 and an Age of 30. This would be converted to a single row with (age=30, n=6, fail=5, y=1)

When fitting models in this form, use mod.fit <- glm(y/n ~ B1, data = ...)

3.2 Pearson Residual

Pearson Residual: $e_j = \frac{\text{observed - predicted}}{\sqrt{\hat{Var}(Observed)}} = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)}}$ with a binomial version of the data, there's a possibility the sample size is large enough for normal approximation to work. With continuous variables, this is not the case and this residual should be interpreted with caution.

Outliers: ± 2.576 though the effect on the model should be examined.

Standardized Pearson Residual: $e_j = \frac{e_j}{\sqrt{1-h_j}} = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)(1-h_j)}}$

where h_j is the jth diagonal of the hat matrix.

3.3 Pearson Statistic

$\chi^2 = \sum_{j=1}^{J} e_j^2$ J: num of explanatory variable patterns.

Can be approximated by $\chi^2_{J-(k+1)}$ distribution where k + 1 is the num of parameters estimating.

$$\begin{aligned} H_0 &: logit(\pi)\alpha + \beta_1 x_1 + ... + \beta_k x_k \text{ k + 1 parameters} \\ H_A &: \text{Saturated Model (J parameters)} \end{aligned} \tag{5}$$

The "saturated" model contains an estimate per explanatory variable pattern.

3.3.1 Influential Obs

$\chi^2 \approx e_j^2$ (squared standardized residual) is used to calculate the influence. The same statistic is used for outliers.

$e_j^2 > \chi^2_{0.95,1} = 3.84$ or $e_j^2 > \chi^2_{0.99,1} = 6.63$ may indicate an outlier or influential EVP.

A measure similar to Cook's distance can also be used.

$\Delta\hat{\beta}_j = \frac{e_j^2 h_j}{(1-h_j)}$

Large values indicate an explanatory variable pattern may be influential. Its effect on the $\hat{\beta}$'s can be seen by temporarily removing the variable from the dataset and refitting the model.

3.4 Diagnostic Plots

- When there is one explanatory var, plotting e_j, e_j^2 and/or $\Delta\hat{\beta}_j$ is helpful. Doesnt work if the explanatory variable is binary.
- e_j, e_j^2 and/or $\Delta\hat{\beta}_j$ vs the observation number
- e_j^2 and $\Delta\hat{\beta}$ vs the estimated probabilities **or** proportion to n_j.
- e_j^2 vs estimated probabilities with the plotting point proportional to $\Delta\hat{\beta}$ can help combine different influence measures.

3.5 Model Selection Process

What explanatory variables should be in the model? Should interactions or quadratic terms be included?

1. Find all possible one variable logistic regression models. LRT is preferred to test model parameters with Logistic Regression due to the χ^2 approximation for the LRT statistic.

2. Put all variables found in 1 in a logistic regression model. Perform backwards elimination.

3. Determine if the quadratic or interaction terms are needed in the model. The best way is to add them and see if they are significant.

4. Convert data to explanatory variable pattern

5. Examine how well the model fits the data. Make any changes necessary.

   Calculate Pearson residuals, standardized residuals, Pearson Statistic, $\Delta\hat{\beta}$'s and LRT. Construct diagnostic plots and determine if changes need to be made.

6. Use the model.

4 Linear Mixed Models (LMM)

Mixed Modeling is a model with both *random* and *fixed* effects.

- It is often found in Physical, Biological, Medicine, and Social Sciences.
- Can be used in Longitudinal Studies
- Can be used for clusters where there is staggered entry, dropout, missing data, and mistime visits.

Example Study

Suppose a study involving different treatments (A, B, Control) to groups of rats were conducted.

4.1 Fixed Effect

An effect which is generally being tested for. Characteristics such as Gender, Age, and Blood Type would be considered fixed effects if they are used as independent variables.

$Y_i = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 BloodType + \epsilon_i$

Helps explain the variance of Y at each level of the data.

4.2 Random Effect

An effect which is being considered a sample from a much larger distribution. Values for these are treated as *random samples*.

They are variables specific to the data sample.

- Allow us to account for correlation among observations within the same level-2 or higher units.
  - correlations among observations within the same school
- Allow us to partition the total variance of Y into levels that correspond with the multilevel structure of the data.
  - How much of the variation in student math scores can be attributed to student-level variability (level 1) vs school-level variability (level 2)?

$Y_i = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 BloodType + b_{0j} + \epsilon_{ij}$

b_{0j}: Cluster-specific random deviations ϵ_{ij}: Subject-within-cluster-specific errors

This LMM is referred to as the **Variance Components** model because it partitions the total variation in the outcome into between-cluster variation and within-cluster variation.

- variance of the random intercepts is the between-cluster variation. Also referred to as the Level 2 variance.
- Variance of the residuals is the within-cluster variation, known as the Level 1 variance.

Introduction of a *Random Effect* creates a variance component.

$Var(ij) = \sigma_p^2 + \sigma_r^2$

Random effects usually include a random intercept for each level of clustering to account for possible correlation within clusters, and to make inference to the larger population of clusters.

4.3 Multi-Level Data

Level 1 is the smallest grain of data where the outcome variable of interest is measured.

Levels 2+ capture higher level information.

- cluster-levels for Clustered Data

- Subject-level for Longitudinal Data
- Subject and cluster levels for clustered-longitudinal data

Questions to Drive Analysis

- Is the data "clustered", "longitudinal", or "clustered-longitudinal"?
- How many levels are there? 2, 3, or more?
- What defines each level?
- What is the outcome of interest and is it measured at Level 1?
- What other variables are of interest at each level?

  Using single-level (OLS, GLM) analysis leads to:

  - Unit of Analysis problem
    - School or child?
  - Aggregation Bias
    - School SES or child SES?
  - Incorrectly estimated precision or standard errors
    - incorrect p-values and thus conclusions

4.3.1 Clustered Data

An outcome is measured once for each subject, and subjects "belong to" clusters, such as families, schools, or neighborhoods. These outcomes are likely to be correlated with other members of the "cluster".

Each "Level" represents a factor that can be thought of as a random sample from a larger population.

- students in a two-level clustered dataset can be thought of as a random sample of students within each school.

4.3.2 Repeated Measures

Multiple observations for one treatment level **or** the same subject.

4.3.3 Examples

Level 3	Level 2	Level 1
	School	[Student 1, Student 2, Student 3]
	Child	[MeasurementTime 1, MeasurementTime 2]
	Rat	[Treatment A (Region 1), Treatment A (Region 2)]
School	Classroom	[Student 1, Student 2, ...]
School	Student	[Score Grade 1, Score Grade 2, ...]