Introduction to Generalized Linear Models in R

Paul Stey

January 17, 2018

Table of Contents

Reading Data in to R

<EXAMPLES_IN_R>

Four Books

- "*Data Analysis using Regression and Multilevel/Hierarchical Models*", Gelman & Hill
- "*Applied Logistic Regression*", Hosmer *et al.*
- "*An Introduction to Generalized Linear Models*", Dobson & Barnett
- "*Categorical Data Analysis*", Agresti

Terminology

"General Linear Model" $\neq$ "Generali**zed** Linear Model"

1. "General linear model" refers to models with a continuous outcome variable, and assumption of normality
   - ANOVA (and friends)
   - Linear regression
2. Term "Generali**zed** Linear Model" is usually used to refer to a family of models for categorical and/or non-normally distributed outcome variables

Terminology (*cont.*)

"Covariate" = "Predictor"

"Binomial logistic regression" = "logit regression" or "logit model"

Terminology (*cont.*)

Regression vs. Classification

Recap of Linear Models

Linear Regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

1. Outcome variable (y) is continuous
2. Can have one or many predictor variables
3. Predictors can be continuous or categorical
4. Examples:
   - Estimating effect square footage on home price
   - Effect of age and weight on blood pressure

Assumptions of Linear Regression

1. $E(y_i) = \mu_i = \beta_0 + \beta_1 x_i$
    - Or equivalently $E(\varepsilon_i) = 0$
    - The means of $E(y_i)$ are on a straight line
2. $var(y_i) = \sigma^2$
    - Or equivalently $var(\varepsilon_i) = \sigma^2$
    - Known as *homoscedasticity*
3. $cov(y_i, y_j) = 0$
    - Or equivalently $cov(\varepsilon_i, \varepsilon_j) = 0$
    - Errors are uncorrelated
4. y_i is normally distributed
    - Needed when using maximum likelihood estimation (MLE), but not ordinary least squares (OLS)

Limitations of Linear Regression

1. Relationship might not be linear
2. Often doesn't make sense for y to increase to infinity as x goes to infinity (e.g., probability of dying)

Interactions between Predictors

In some cases we are curious whether two predictors interact. We can estimate this effect easily in R, and it allows us to test whether a given predictor behaves differently depending on the value of another predictor. This often called a “moderation effect”.

Interactions between Predictors

For instance, suppose we believe that bodyweight and age both predict blood pressure.

But we might also believe that bodyweight become *an especially strong* predictor in older individuals. We can test the interaction explicitly.

Age * Weight Interaction

For example:

```
fm1 <- lm(bp ~ age + weight + age*weight, dat)
```

The above model has 4 regression coefficients—one for intercept, age, `weight`, and `age*weight`).

The coefficient associated with the `age*weight` term would tell us whether or not a significant interaction exists.

Why Logistic Regression?

Linear regression assumes a continuous outcome variable

If the outcome variable is *not* continuous, we need a different approach.

In the case of a binary outcome variable, we model $\Pr(y_i = 1)$

Binomial Logistic Regression

Logistic Regression

1. Used when outcome variable takes one of two values (e.g., 0 or 1, "lived" or "died")
2. Similar structure as linear regression
   - Estimate effects of predictors on outcome
   - Can have one or many predictors
3. Can answer similar kinds of questions as linear regression, for example:
   - "*What is the effect of the predictor, x, on the outcome y?*"

Logistic Regression vs. Linear Regression

Differences from linear regression:

1. Assumes outcome is bounded by 0 and 1, that is $0 \leq E(y_i) = \pi_i \leq 1$
2. Variance of y is *not* constant (i.e., not the same for all y_i)
3. Similarly, variance of ε is not constant
4. Computational differences (i.e., closed-form vs numerical methods)

Components of Generalized Linear Models

Recall the form of the linear model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

which can also be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}\boldsymbol{\beta}$ is the systemic component and $\boldsymbol{\varepsilon}$ is the random component.

Components of Generalized Linear Models (*cont.*)

Form of GLM:

$$g(\mu) = \mathbf{X}\boldsymbol{\beta}$$

Generalized linear models have 3 components:

1. Systemic component
   - Same as linear regression (e.g., $\mathbf{X}\boldsymbol{\beta}$)
2. Response distribution assumption
   - Random component of the model
   - Specifies the probabilistic mechanism by which responses were generated
3. Link function
   - This is $g(\cdot)$ in equation above

The Link Function

Link function is a characteristic feature of generalized linear models

A link function:

1. Connects the systemic component to response (i.e., "links" them)
   - Allows us to map a linear function with range $(-\infty, \infty)$ to some new range; e.g., $(0, 1)$
2. Differs according to the species of GLM in question (and even within)
3. Similar to "activation functions" in artificial neural networks

Binomial Logistic Regression

Logistic regression with a single predictor:

$$\pi(x_1) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

$$= \frac{1}{1 + \exp(-\eta)}$$

where $\eta = \beta_0 + \beta_1 x_1$

Binomial Logistic Regression

$$\pi(x_1) = \frac{\exp\left(\beta_0 + \beta_1 x_1\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1\right)}$$

Note that the $\beta_0 + \beta_1 x_1$ in the above equation is the same as we saw in linear regression. This is called the "linear predictor" in logistic regression

Interpreting Parameter Estimates

Interpretation of logistic regression parameter estimates:

1. Slightly different than linear regression
2. Recall our model is $\Pr(y_i = 1) = \text{logit}^{-1}(\mathbf{X}\boldsymbol{\beta})$
3. Regression parameters estimates are on logit scale (log odds),
   - It's common to exponentiate $\widehat{\boldsymbol{\beta}}$
   - Value of $\exp(\beta_j)$ is the odds ratio of 1-unit increase on x_j

Logistic Regression Examples

<EXAMPLES_IN_R>

Model Evaluation

- Recall that R^2 in linear regression gives us a nice method of evaluating models (i.e., proportion of variance explained).
- However, in logistic regression, there is no direct analogue to R^2 (but there are some similar measure)
- Thus, we tend to rely on the information-based criteria discussed previously (e.g., AIC, BIC)
  - These also have the advantage of penalizing unnecessary model complexity

Choosing a Link Function

Several link function options for modeling binomial data:

1. Logit link (most common, by far)
2. CDF of normal distribution (probit regression)
3. CDF of t-distribution ("robit" model; robust binomial regression)
   - Degrees of freedom parameter allows for flexibility in accommodating outliers

Poisson Regression

1. Form of Poisson model for single predictor

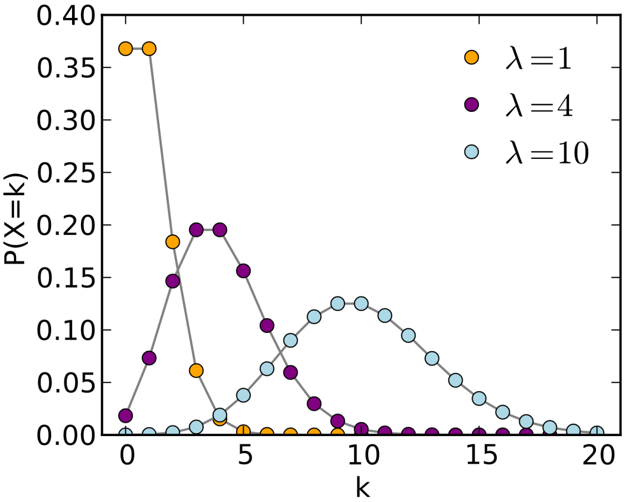
$$\log(\mu) = \beta_0 + \beta_1 x_1$$

2. Link function is $\log(\cdot)$
3. We use Poisson regression when we model count data
    - Number of offspring an individual has
    - Number bacterial colonies in Petri dish

4. As we saw with logistic regression, we *could* use a linear model instead,[1] but our parameter estimates would be biased, and our model inaccurate

[1] Don't do this.

Poisson Distribution



Poisson Regression

- As with linear and logistic regression, we can use Poisson regression to estimate effects of predictors on some outcome
- We can also use fitted Poisson regression models to predict future values of some outcome variable given known values for the covariates
- Frequently used for modeling rare events

Assumptions of Poisson Regression

1. Log-transformed outcomes are linearly related to predictors
2. Observations are independent
3. Distributional assumption: $y_i|x_i \sim \text{Poisson}(\lambda_i)$

Assumptions of Poisson Regression (*cont.*)

- Note that the assumption $y_i|x_i \sim \text{Poisson}(\lambda_i)$ has some important implications.
- The Poisson distribution has a single parameter, λ, which is both its mean and variance.
- It is frequently the case we will have data where the variance greatly exceeds the mean. When this happens, it is wise to consider similar alternatives to the Poisson model

Alternatives to Poisson Models

1. Quasi-Poisson regression
2. Zero-inflated Poisson regression
3. Negative Binomial regression

Evaluation of Poisson Regression Models

- As with logistic regression, there is no direct counterpart to the R^2 in linear regression
- Poisson regression models can be compared using AIC and BIC as we saw with linear and logistic regression

Interpreting Poisson Regression Parameters

- We can exponentiate Poisson regression parameter estimates, and then treat them multiplicative effects

Poisson Regression Examples

<EXAMPLES_IN_R>

Survival Analysis

- Sometimes called event history analysis
- Strictly speaking, survival analysis is not in the family of generalized linear models
- Survival models have some similarities with logistic and Poisson regression
- Key idea is survival analysis is to model the time until and event occurs

Cox Proportional Hazards Models

- Perhaps the most common method of modeling time-to-event data is the Cox proportional hazards (PH) model
- The Cox PH model has the form

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

$$= \lambda_0(t)\exp(X_i \cdot \beta)$$

where $\lambda_0(t)$ is the baseline hazard function

Not Discussed GLMs

- Multinomial logistic regression can accommodate problems in which we have more than 2 discrete categories in our outcome variable. Multinomial models also use the logit link function, and have a similar structure as binomial logistic regression
- Ordered logistic regression can be used when the outcome variable has more than 2 categories, and they have some logical ordering (e.g., *"poor"*, *"fair"*, *"good"*)
- Penalized regression methods (e.g., ridge regression, lasso) can be applied to logistic regression and Poisson regression, as well as Cox PH models (see glmnet package in R)
- Hierarchical / Mixed-Effects Models

References

1. "*Data Analysis using Regression and Multilevel/Hierarchical Models*", Gelman & Hill
2. "*Applied Logistic Regression*", Hosmer *et al.*