Introduction to Generalized Linear Models

Edps/Psych/Soc 589

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Fall 2018

Outline

- Introduction (motivation and history).
- "Review" ordinary linear regression.
- Components of a GLM.
  1. Random component.
  2. Structural component.
  3. Link function.
- Natural exponential family (technical).
- Normal Linear Regression re-visited.
- GLMs for binary data (introduction).
  Primary Example: High School & Beyond.
  1. Linear model for π.
  2. Cumulative Distribution functions (alternative links).
  3. Logistic regression.
  4. Probit models.

Outline (continued)

- GLMs for count data.
  1. Poisson regression for counts.
     Example: Number of deaths due to AIDs.
  2. Poisson regression for rates.
     Example: Number of violent incidents.
- Inference and model checking.
  1. Wald, Likelihood ratio, & Score test.
  2. Checking Poisson regression.
  3. Residuals.
  4. Confidence intervals for fitted values (means).
  5. Overdispersion.
- Fitting GLMS (a little technical).
  1. Newton-Raphson algorithm/Fisher scoring.
  2. Statistic inference & the Likelihood function.
  3. "Deviance".
- Summary

Introduction to Generalized Linear Modeling

Benefits of a model that fits well:

- The structural form of the model describes the patterns of interactions or associations in data.
- Inference for the model parameters provides a way to evaluate which explanatory variable(s) are related to the response variable(s) while (statistically) controlling for other variables.
- Estimated model parameters provide measures of the strength and (statistical) importance of effects.
- A model's predicted values "smooth" the data — they provide good estimates of the mean of the response variable.

Advantages of a Modeling Approach

Over Significance Testing

- Models can handle more complicated situations.

  For example, Breslow-Day is limited to $2 \times 2 \times K$ tables and does not provide estimates of common odds ratios for tables larger than 2×2.

  Loglinear models can be used to test for homogeneous association in $I \times J \times K$ (or higher–way) tables and provide estimates of common odds ratios.

- With models, the focus is on estimating parameters that describe relationships between/among variables.

A Little History

From Lindsey (who summary that from McCullagh & Nelder who got a lot from Stiegler)

- Multiple linear regression — normal distribution & identity link (Legendre, Guass: early 19th century).
- ANOVA — normal distribution & identity link (Fisher: 1920's – 1935).
- Likelihood function — a general approach to inference about any statistical model (Fisher, 1922).
- Dilution assays — a binomial distribution with complementary log-log link (Fisher, 1922).
- Exponential family — class of distributions with sufficient statistics for parameters (Fisher, 1934).
- Probit analysis — binomial distribution & probit link (Bliss, 1935).

A Little History (continued)

- Logit for proportions — binomial distribution & logit link (Berkson, 1944; Dyke & Patterson, 1952)
- Item analysis — Bernoulli distribution & logit link (Rasch, 1960).
- Log linear models for counts — Poisson distribution & log link (Birch, 1963).
- Regressions for survival data — exponential distribution & reciprocal or log link (Feigl & Zelen, 1965; Zippin & Armitage, 1966; Glasser, 1967).
- Inverse polynomials — Gamma distribution & reciprocal link (Nelder, 1966).
- Nelder & Wedderburn (1972): provided unification. They showed
  - All the previously mentioned models are special cases of a general model, "**Generalized Linear Models**"
  - The MLE for all these models could be obtained using same algorithm.
- All of the models listed have distributions in the "**Exponential Dispersion Family**"

Software Developments

- Computer software development in the 70's: "GLIM"
  Short for "**G**eneralized **L**inear **I**nteractive **M**odelling.
- Any statistician or researcher could fit a larger class of models (not restricted to normal).
- Growing recognition of the likelihood function as central to all statistical inference.
- Allowed experimental development of many new methods & uses for which it was never originally imagined.
- PROC GENMOD (GENeralized linear MODels) in SAS.
- glm package in R.

Limitations

- Linear function
- Responses must be independent
- There are ways around these by going to a slightly more general models and using more general software (e.g., SAS/NLMIXED, GLIMMIX, NLP, GAMs).
- R has specialized packages for some of the models that are not linear and/or dependent (e.g., packages lme4, logmult, gam).

Review of Ordinary Linear Regression

Linear (in the parameters) model for continuous/numerical response variable (Y) and continuous and/or discrete explanatory variables (X's).

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

where

$$e_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and independent.}$$

This linear model includes

- Multiple regression
- ANOVA
- ANCOVA

Simple Linear Regression

$$Y_i = \alpha + \beta x_i + e_i$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$ and independent.

We consider X as fixed, so $Y_i \sim \mathcal{N}(\mu_{(x_i)}, \sigma^2)$.

In regression, the focus is on the mean or expected value of Y, i.e.,

$$\begin{aligned} E(Y_i) &= E(\alpha + \beta x_i + e_i) \\ &= \alpha + \beta x_i + E(e_i) \\ &= \alpha + \beta x_i \end{aligned}$$

GLMs go beyond Simple Linear Regression

Generalized Linear Models go beyond this in two major respects:

- The response variable(s) can have a distribution other than normal — any distribution within a class of distributions known as "exponential family of distributions".
- The relationship between the response (Y) and explanatory variables need not be simple ("identity"). For example, instead of

$$Y = \alpha + \beta x$$

we can allow for transformations of Y

$$g(Y) = \alpha + \beta x$$

- These are ideas we'll come back to after we go through an example ....

Counts of T_4 cells/mm in Blood Samples

From Lindsey (1997) from Altman (1991). The counts equal T_4 cells/mm in blood samples from 20 patients in remission from Hodgkin's disease & 20 other patients in remission from disseminated malignancies:

Hodgkin's		Non-Hodgkin's	
396	568	375	375
1212	171	752	208
554	1104	151	116
257	435	736	192
295	397	315	1252
288	1004	657	700
431	795	440	771
1621	1378	688	426
902	958	410	979
1283	2415	377	503

Is there a "difference" in cell counts between the two diseases?

What is Meant by "Difference"?

- Mean counts
- Variability
- Overall form of the distribution

Naive Approach: Assume a normal distribution and do a "t-test" (i.e., compute difference between means and divide by s.e. of difference).

More Sophisticated approach: Assume a Poisson distribution and compute difference between log of the means (i.e., ratio of means).

Summary of some possibilities and results

Model	AIC		"Likelihood ratio"	$\sqrt{\text{"Wald"}}$
	No difference	Difference	Difference in $-2\log(L)$	Estimate /s.e.
Normal	608.8	606.4	4.5	2.17
Normal log link	608.8	606.3	4.5	2.04
Gamma	591.2	587.9	5.2	2.27
Inverse Gaussian	589.9	588.1	3.8	1.87
Poisson	11652.0	10285.3	1368.96	36.52
Negative Binomial	591.1	587.9	5.3	2.37

- AIC = weighs goodness-of-fit & model complexity (smaller is better)
- Wald = $(\widehat{\text{parameter}}/\widehat{\text{standard error}})^2$.
- I get slightly different results than published and between SAS & R.
- Assumptions?

Considering Assumptions

T_4 cells/mm and Hodgkin's disease data continued

- Independence of observations
- Recall that with Poission

$$\mu = \sigma^2$$

- Sample statistics

disease	N	Mean	Variance
hodgkin	20	823.20	320792.27
non-hodgkin	20	521.15	85568.77

- Homogeneity assumption is suspect.

Components of Generalized Linear Models

There are 3 components of a generalized linear model (or GLM):

1. **Random Component** — identify the response variable (Y) and specify/assume a probability distribution for it.
2. **Systematic Component** — specify what the explanatory or predictor variables are (e.g., X_1, X_2, etc). These variable enter in a linear manner

$$\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

3. **Link** — Specify the relationship between the mean or expected value of the random component (i.e., $E(Y)$) and the systematic component.

Components of Simple linear regression

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- Random component: Y is the response variable and is normally distributed. . . generally we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- X is the explanatory variable is linear in the parameters. . .

$$\alpha + \beta x_i$$

- Identity link.

$$g(E(Y_i)) = E(Y_i) = \alpha + \beta x_i$$

Closer look at each of these components. . . .

Random Component

Let N = sample size and suppose that we have $Y_1, Y_2, \ldots, Y_N$ observations on our response variable and that the observations are all independent. Y's are discrete variables where Y is either

Dichotomous (binary) with a fixed numbers of trials. success/failure correct/incorrect agree/disagree academic/non-academic program	Counts (including cells of a contingency table): Number of people who die from AIDS during a given time period. Number of times a child tries to take a toy away from another child. Number of times patents generated by firms.
Binomial distribution.	Poisson distribution

Distributions for Discrete Variables

Thus the two distributions we will be primarily using are

- Binomial
- Poisson

With GLMs, you can use any distribution that belongs to the "exponential family of distributions". This is a wide class of distributions that have many of the "nice" properties of the Normal distribution. (we'll look at this in a bit more detail later).

Systematic Component

As in ordinary regression, we will be modelling means. The focus is on the expected value of our response variable

$$E(Y) = \mu$$

We want to investigate whether and how μ varies as a function of the levels of our predictor or explanatory variables, X's.

The systematic component of the model consists of a set of explanatory variables and some linear function of them.

$$\beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k.$$

This linear combination of our explanatory variables is referred to as a "linear predictor".

Linear Predictor

This restriction to a linear predictor is not all that restrictive.

For example,

- $x_3 = x_1 x_2$ — an "interaction".
- $x_1 \Rightarrow x_1^2$ — a "curvilinear" relationship.
- $x_2 \Rightarrow \log(x_2)$ — a "curvilinear" relationship.

$$\beta_o + \beta_1 x_1^2 + \beta_2 \log(x_2) + \beta_3 x_1^2 \log(x_2)$$

This part of the model is very much like what you know with respect to ordinary linear regression.

The Link Function

"Left hand" side of an equation/model — the random component,	"Right hand" side of the equation— the systematic component; that is,
$E(Y) = \mu$	$\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

We now need to "link" the two sides.

How is $\mu = E(Y)$ related to $\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$?
We do this using a "Link Function" $\Longrightarrow g(\mu)$

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

More about the Link Function

- Important things about $g(.)$:
  - This function $g(.)$ is "monotone" — as the systematic part gets larger, μ gets larger (or smaller).
  - The relationship between $E(Y)$ and the systematic part can be non-linear.
- Some common links are
  - Identity (ordinary regression, ANOVA, ANCOVA):

$$E(Y) = \alpha + \beta x$$

  - Log link which is often used when Y is nonnegative (i.e., $0 \leq Y$)

$$\log(E(Y)) = \log(\mu) = \alpha + \beta x$$

    This yields a "loglinear" model.
  - Logit link, which is often used when $0 \leq \mu \leq 1$ (when response is dictohomous/binary and we're interested in a probability).

$$\log(\mu/(1-\mu)) = \alpha + \beta x$$

General Model Formula for a GLM

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

The links ones given on previous slide and below are special ones (depending on the assumed distribution):

Distribution	"Natural Parameter"	"Canonical Link"
Normal	μ	Identity
Poisson	$\log(\mu)$	log
Binomial	$\log(\mu/(1-\mu))$	logit

Natural Exponential Family of Distributions

"One-Parameter Exponential Distribution"

Probability density or mass functions belonging to the natural exponential family have the general form

$$f(y_i;\theta_i) = \exp\{a(y_i)b(\theta_i) - c(\theta_i) + d(y_i)\}$$

where

y_i is an observation ($i = 1, \ldots, N$).
θ_i is the parameter of the distribution for i and $b(\theta_i)$ is the location parameter (i.e., the mean; other parameters such are variance are often considered "nuisance" parameters).
$a(.)$, $b(.)$, $c(.)$, and $d(.)$ are all functions.

When $a(y_i) = y_i$, then the density/mass is in "canonical form", and we have

$$f(y_i;\theta_i) = \exp\{y_i b(\theta_i) - c(\theta_i) + d(y_i)\}$$

When in canonical form, the "natural parameter" is $b(\theta_i)$.

According to Webester's Dictionary

Canonical means

- conforming to a general rule
- reduced to the simplest or clearest scheme possible
- the simplest form of a matrix (specifically the form of a square matrix that has zero off-diagonals).

Now for some examples. . .

The Poisson Distribution

$$f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

where

$y = 0, 1, 2, \ldots$

$\theta = \mu$ (the parameter of the distribution).

Now to put this in canonical form:

$$\begin{aligned} f(y;\mu) &= \exp\left(\log\left(\frac{\mu^y e^{-\mu}}{y!}\right)\right) \\ &= \exp\left(y\log(\mu) + e^{-\mu} - \log(y!)\right) \\ &= \exp\left(yb(\mu) - c(\mu) + d(y)\right) \end{aligned}$$

- $a(y) = y$
- $b(\mu) = \log(\mu)$, the natural parameter.
- $c(\mu) = \exp(-\mu)$.
- $d(y) = -\log(y!)$.

The canonical link for the Poisson distribution: $\log(\bullet)$.

The Binomial Distribution

$$f(y;\pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y}$$

where $y = 0, 1, \ldots, n$.
$n =$ number of trials.
$\pi =$ probability of a success .
π is the parameter of interest and n is assumed to be known.
We now re-express the distribution as

$$\begin{aligned} f(y;\pi) &= \exp\left(\log\left[\left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y}\right]\right) \\ &= \exp\left(y\log(\pi) + (n-y)\log(1-\pi) + \log\left(\left(\begin{array}{c} n \\ y \end{array} \right)\right)\right) \\ &= \exp(y\log(\pi/(1-\pi)) + n\log(1-\pi) + \log\left(\begin{array}{c} n \\ y \end{array} \right)) \\ &= \exp(yb(\pi) - c(\pi) + d(y)) \end{aligned}$$

Canonical Form of the Binomial Distribution

$$\begin{aligned} f(y;\pi) &= \exp(y\log(\pi/(1-\pi)) + n\log(1-\pi) + \log\left(\begin{array}{c} n \\ y \end{array}\right)) \\ &= \exp(yb(\pi) - c(\pi) + d(y)) \end{aligned}$$

where
$a(y) = y$
$b(\pi) = \log(\pi/(1-\pi))$, the natural parameter

$c(\pi) = -n\log(1-\pi)$
$d(y) = \log\left(\left(\begin{array}{c} n \\ y \end{array}\right)\right)$.

The canonical link is the logit— the log of the odds

Exponential Dispersion Family

Generalization of the one-parameter exponential family: includes a constant scale parameter ϕ.
The canonical form of the exponential dispersion family:

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i b(\theta_i) - c(\theta_i)}{r_i(\phi)} + d(y_i,\phi)\right]$$

where $r_i(\phi)$ is a function of the dispersion parameter.
Notes:

- For Poisson and Binomial $r_i(\phi) = 1$.
- If ϕ is known and $r_i(\phi) = r(\phi)$, then back to one-parameter exponential family.

With this generalization...

Normal Distribution

$$f(y;\mu;\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-1}{2\sigma^2}(y-\mu)^2\right)$$

θ (parameter of the distribution) is μ, the mean.
The variance σ^2 is considered a "nuisance" parameter.
Putting $f(y;\mu)$ into it's canonical form.

$$\begin{aligned}
f(y;\mu;\sigma^2) &= \exp\left(\log\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\right)\right)\exp\left(\frac{-1}{2\sigma^2}(y-\mu)^2\right) \\
&= \exp\left(\log\left((2\pi\sigma^2)^{-1/2}\right)\right)\exp\left(\frac{-1}{2\sigma^2}(y-\mu)^2\right) \\
&= \exp\left[\frac{y\mu-\mu^2/2}{2\sigma^2} - 1/2(\log(2\pi\sigma^2)) - \frac{y^2}{2\sigma^2}\right] \\
&= \exp\left[\frac{yb(\mu)-c(\mu)}{r(\sigma^2)+d(y_i,\sigma^2)}\right]
\end{aligned}$$

The Canonical Form of the Normal Distribution

$$
\begin{aligned}
f(y;\mu;\sigma^2) &= \exp\left[\frac{y\mu - \mu^2/2}{2\sigma^2} - 1/2(\log(2\pi\sigma^2)) - \frac{y^2}{2\sigma^2}\right] \\
&= \exp\left[\frac{yb(\mu) - c(\mu)}{r(\sigma^2)} + d(y_i, \sigma^2)\right]
\end{aligned}
$$

where

$$
\begin{aligned}
a(\mu) &= y \\
b(\mu) &= \mu \\
c(\mu) &= \mu^2/2. \\
d(y;\phi) &= -1/2(\log(2\pi\sigma^2)) - y^2/(2\sigma^2). \\
r(\sigma^2) &= \sigma^2
\end{aligned}
$$

So,

$b(\mu) = \mu$ is the "natural parameter".

The canonical link is the identity.

The Normal GLM: Ordinary linear regression

- Generalized linear models go beyond ordinary linear regression in two ways
  1. The random component can be something other than Normal.
  2. We can model a function of the mean.
- GLM have a definite advantage over the "traditional" way of analyzing non-normal responses (Y). The traditional way to handle non-normal responses:
  1. Transform your data so that responses are approximately Normal with constant variance.
  2. Use least squares.
- Transforming to normality with constant variance very rarely works. . .

Problem with the Traditional Approach

- A transformation that produces constant variance may not yield normally distributed response.
  Counts that have a Poisson distribution where $E(Y) = \mu$ and $\text{Var}(Y) = \mu$.
  Binomial distributed responses where $E(Y) = n\pi$ and $\text{Var}(Y) = n\pi(1 - \pi)$.
- Linear models often fit discrete data very badly — they can yield predicted values of μ that are outside the range of possible values for Y.
  - Consider counts that have a Poisson distribution where $Y \geq 0$.
  - Consider Binomial distributed responses where $0 \leq \pi \leq 1$.
  - Linear models can yield negative predictions.

Advantage of GLM over Traditional Regression

- You don't have to transform Y to normality.
  The choice of link is separate from choice of random component.
  If the link produces additive effects, then don't need constant variance. (I'll show an example of this next week).
- The models are fit using maximum likelihood. Thus optimal properties of estimators.

Next we'll now talk about GLMS for

1. Dichotomous (binary) data — linear, logit, probit and logistic regression models. (introduce them now and go into much more detail later).
2. Poission regression for count data — these are very similar to regression that you are familiar with, but with a twist.