Introduction to Generalized Linear Models

Jonathan Boss

June 27, 2022

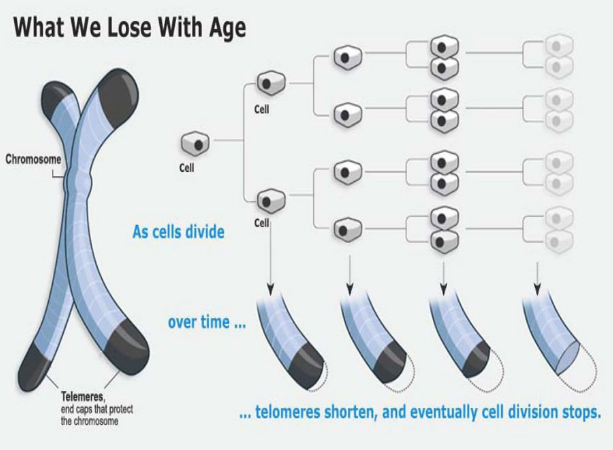
Big Data Summer Institute 2022
University of Michigan

Outline

Motivation

Regression Models

Regression is a statistical technique for modeling the relationship between explanatory variable(s) and response variable(s).



Regression allows us to model relationships adjusted for other factors!

Multivariable Linear Regression

Notation:

Y_i: Response for i-th observation

X_{ij}: j-th explanatory variable for i-th observation

Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2), \;\; i = 1, \ldots, n$$

Alternative Notation:

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \;\; \boldsymbol{x}_i^\top = (1, X_{i1}, \ldots, X_{ip}), \;\; \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$$

Multivariable Linear Regression Assumptions

Systematic Component:

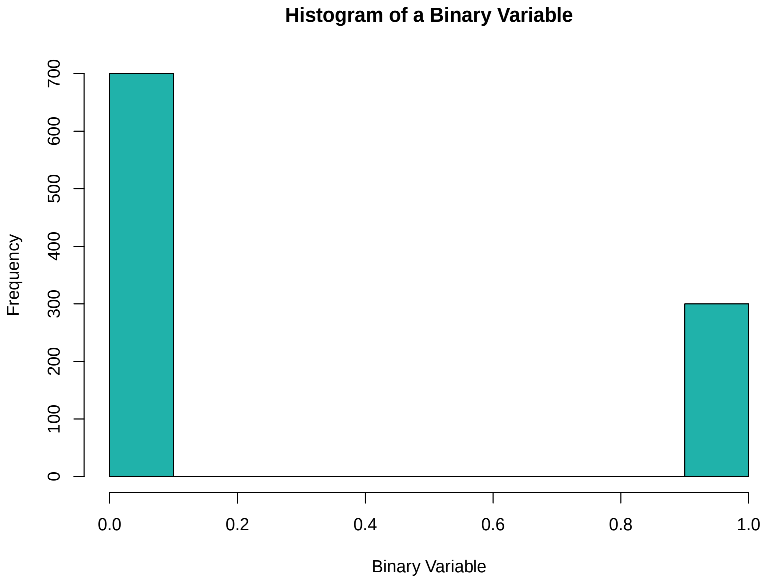
$$E[Y_i \mid \boldsymbol{x}_i] = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

We will sometimes use $E[Y_i]$ as shorthand for $E[Y_i \mid \boldsymbol{x}_i]$.

Random Component: At each level of the predictor, variation in the response is characterized as $N(0, \sigma^2)$

Independence Between Observations

What if Y_i is Binary?



What if Y_i is Binary?

If Y_i is binary, then

$$Y_i \sim \text{Bernoulli}(\pi_i), \;\; \pi_i = \pi(\boldsymbol{x}_i) = P(Y_i = 1 \mid \boldsymbol{x}_i)$$

Normality Assumption is violated!

Additionally,

$$E[Y_i] = \pi_i$$
$$V(Y_i) = \pi_i(1 - \pi_i) = E[Y_i]\{1 - E[Y_i]\}$$

Constant variance assumption is violated!

Predictions from the resulting linear regression model, $\widehat{Y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$, are not restricted to be between 0 and 1.

Idea: Model a function of $E[Y_i]$ rather than $E[Y_i]$ directly.

Types of Outcome Data

Need a more general framework for non-normal outcome data:

- Continuous, non-normal response
  - Time-to-event data
- Binary response
  - Disease vs No Disease
- Nominal categorical response
  - Blood type, US state
- Ordinal categorical response
  - Likert scale data
- Count response
  - White blood cell count, number of insurance claims

Generalized Linear Models (GLMs)

Generalized Linear Models

Generalization here refers to the fact that we are:

- Removing the normality requirement
- Relaxing the constant variance assumption
- Allowing for a function of $E[Y_i]$ to be linear in the parameters

GLMs are based on the exponential family of distributions.

Exponential Family of Distribution

A distribution is in the exponential family of distributions if:

$$f(Y_i; \theta_i, \phi) = \exp\left\{ \frac{t(Y_i)\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

Notes:

- θ_i: parameter of interest, relates to the mean function $E[Y_i \mid \boldsymbol{x}_i]$
- ϕ: Dispersion parameter, relates to the variance
- $t(\cdot)$, $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are functions
- If $t(Y_i) = Y_i$, then the family is in canonical form and θ_i is called the canonical (natural) parameter.

Mean and Variance of Canonical Exponential Family

We can use maximum likelihood theory to show that:

$$E[Y_i] = \frac{d}{d\theta_i} b(\theta_i) = b'(\theta_i)$$

$$V(Y_i) = \frac{d^2}{d\theta_i^2} b(\theta_i) a(\phi) = b''(\theta_i) a(\phi)$$

Notice that $E[Y_i]$ depends only on the natural parameter, while $V(Y_i)$ depends on both the natural parameter and the dispersion parameter.

Example: Normal Response (with known σ^2)

Suppose that $Y_i \sim N(\mu_i, \sigma^2)$, as in linear regression. Then,

$$f(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 \right\}$$

Example: Normal Response (with known σ^2)

Suppose that $Y_i \sim N(\mu_i, \sigma^2)$, as in linear regression. Then,

$$
\begin{aligned}
f(Y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 - \log(2\pi\sigma^2) \right\}
\end{aligned}
$$

Example: Normal Response (with known σ^2)

Suppose that $Y_i \sim N(\mu_i, \sigma^2)$, as in linear regression. Then,

$$\begin{aligned} f(Y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 - \log(2\pi\sigma^2) \right\} \\ &= \exp\left\{ \frac{2Y_i\mu_i - Y_i^2 - \mu_i^2}{2\sigma^2} - \log(2\pi\sigma^2) \right\} \end{aligned}$$

Example: Normal Response (with known σ^2)

Suppose that $Y_i \sim N(\mu_i, \sigma^2)$, as in linear regression. Then,

$$
\begin{aligned}
f(Y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \mu_i)^2 - \log(2\pi\sigma^2) \right\} \\
&= \exp\left\{ \frac{2Y_i\mu_i - Y_i^2 - \mu_i^2}{2\sigma^2} - \log(2\pi\sigma^2) \right\} \\
&= \exp\left\{ \frac{Y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{Y_i^2}{2\sigma^2} - \log(2\pi\sigma^2) \right\}
\end{aligned}
$$

Example: Normal Response (with known σ^2)

Suppose that $Y_i \sim N(\mu_i, \sigma^2)$, as in linear regression. Then,

$$f(Y_i) = \exp\left\{ \frac{Y_i \mu_i - \mu_i^2/2}{\sigma^2} - \frac{Y_i^2}{2\sigma^2} - \log(2\pi\sigma^2) \right\}$$

The normal distribution is a member of the canonical exponential family:

$$\begin{aligned} t(Y_i) &= Y_i \\ \theta_i &= \mu_i \\ b(\theta_i) &= \mu_i^2/2 \\ a(\phi) &= \sigma^2 \\ c(Y_i, \phi) &= -\frac{Y_i^2}{2\sigma^2} - \log(2\pi\sigma^2) \end{aligned}$$

Mean and Variance: $E[Y_i] = b'(\theta_i) = \mu_i$ and $V(Y_i) = b''(\theta_i)a(\phi) = \sigma^2$

Example: Poisson Response

Suppose that $Y_i \sim \text{Poisson}(\lambda_i)$, where $Y_i \in \{0\} \cup \mathbb{Z}^+$

$$f(Y_i) = \frac{e^{-\lambda_i}\lambda_i^{Y_i}}{Y_i!}$$

Example: Poisson Response

Suppose that $Y_i \sim \text{Poisson}(\lambda_i)$, where $Y_i \in \{0\} \cup \mathbb{Z}^+$

$$
\begin{aligned}
f(Y_i) &= \frac{e^{-\lambda_i}\lambda_i^{Y_i}}{Y_i!} \\
&= \exp\left\{ Y_i \log(\lambda_i) - \lambda_i - \log(Y_i!) \right\}
\end{aligned}
$$

Example: Poisson Response

Suppose that $Y_i \sim \text{Poisson}(\lambda_i)$, where $Y_i \in \{0\} \cup \mathbb{Z}^+$

$$\begin{aligned} f(Y_i) &= \frac{e^{-\lambda_i}\lambda_i^{Y_i}}{Y_i!} \\ &= \exp\left\{ Y_i \log(\lambda_i) - \lambda_i - \log(Y_i!) \right\} \end{aligned}$$

The Poisson distribution is a member of the canonical exponential family:

$$\begin{aligned} t(Y_i) &= Y_i \\ \theta_i &= \log(\lambda_i) \\ b(\theta_i) &= \lambda_i = e^{\theta_i} \\ a(\phi) &= 1 \\ c(Y_i, \phi) &= -\log(Y_i!) \end{aligned}$$

Mean and Variance: $E[Y_i] = b'(\theta_i) = \lambda_i$ and $V(Y_i) = b''(\theta_i)a(\phi) = \lambda_i$

Exercise: Binary Response

Suppose that $Y_i \sim \text{Bernoulli}(\pi_i)$

$$f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \quad Y_i \in \{0,1\}$$

Questions:

- Is the Bernoulli distribution a member of the canonical exponential family? If yes, what is $E[Y_i]$ and $V(Y_i)$?

Canonical Exponential Family:

$$f(Y_i;\theta_i,\phi) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i,\phi)\right\}$$

$$E[Y_i] = b'(\theta_i), \quad V(Y_i) = b''(\theta_i)a(\phi)$$

Solution: Binary Response

Suppose that $Y_i \sim \text{Bernoulli}(\pi_i)$

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$$

Solution: Binary Response

Suppose that $Y_i \sim \text{Bernoulli}(\pi_i)$

$$\begin{aligned} f(Y_i) &= \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \exp\left\{ Y_i \log(\pi_i) + (1-Y_i)\log(1-\pi_i) \right\} \end{aligned}$$

Solution: Binary Response

Suppose that $Y_i \sim \text{Bernoulli}(\pi_i)$

$$\begin{aligned} f(Y_i) &= \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \exp\left\{ Y_i \log(\pi_i) + (1-Y_i)\log(1-\pi_i) \right\} \\ &= \exp\left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right\} \end{aligned}$$

Solution: Binary Response

Suppose that $Y_i \sim \text{Bernoulli}(\pi_i)$. Then,

$$f(Y_i) = \exp\left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right\}$$

Bernoulli distribution is a member of the canonical exponential family:

$$\begin{aligned} t(Y_i) &= Y_i \\ \theta_i &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \implies \pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \\ b(\theta_i) &= -\log(1-\pi_i) = \log(1+e^{\theta_i}) \\ a(\phi) &= 1 \\ c(Y_i, \phi) &= 0 \end{aligned}$$

$E[Y_i] = b'(\theta_i) = e^{\theta_i}/(1+e^{\theta_i})$, $\quad V(Y_i) = b''(\theta_i)a(\phi) = e^{\theta_i}/(1+e^{\theta_i})^2$

Generalization Checklist

Generalization here refers to the fact that we are:

- Removing the normality requirement ✓
- Relaxing the constant variance assumption ✓
- Allowing for a function of $E[Y_i]$ to be linear in the parameters ?

Link Function

Generalized Linear Model:

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad \mu_i = E[Y_i]$$

Details:

- $g(\cdot)$ is called the link function, connects μ_i and $\boldsymbol{x}_i$
- $g(\cdot)$ is required to be monotone and differentiable
- $g(\cdot)$ is called the canonical link if $\eta_i = \theta_i$, where $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
- Linearity assumption now applies to $g(\mu_i)$, $g(\mu_i) \in (-\infty, \infty)$
- Still assume that $Y_1, \ldots, Y_n$ are independent

Canonical Link Examples

Normal Response:

$$\theta_i = \mu_i, \quad \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \implies \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad E[Y_i] = \mu_i$$

Bernoulli Response:

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right), \quad \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \implies \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad E[Y_i] = \pi_i$$

Poisson Response:

$$\theta_i = \log(\lambda_i), \quad \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \implies \log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad E[Y_i] = \lambda_i$$

Canonical Link or Non-Canonical Link?

Canonical links mostly lead to mathematical/algorithmic simplifications, but are not intrinsically better to use than non-canonical links.

The link function is often chosen based on (not an exhaustive list):

- Type of response variable
- The desired interpretability of parameters in your model
- Model fit
- Whether the model specification makes conceptual sense

My recommendation is to default to the canonical link, and only use non-canonical links if there is an explicit rationale.

GLM Specification (Canonical Exponential Family)

- **Random Component:** Assume that $Y_1, \ldots, Y_n$ come from a distribution within the exponential family of distributions:

$$f(Y_i; \theta_i, \phi) = \exp\left\{ \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

- **Systematic Component (Linear Predictor):** $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
- **Link Function:** $\eta_i = g(\mu_i) \implies \mu_i = g^{-1}(\eta_i)$

Specific Types of GLMs

GLMs for Continuous Responses

Linear Regression Model:

- Assumes a normally distributed response
- Generally good for symmetric responses
- Response takes values in $(-\infty, \infty)$

Gamma Regression Model:

- Assumes a gamma distributed response
- Less common, but applicable for right-skewed responses
- Response takes values in $(0, \infty)$

Note: Alternatively, we can log-transform a right-skewed, positive response variable and use the linear regression framework.

Bernoulli Responses

Logistic Regression (Canonical Link):

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \pi_i = P(Y_i = 1 \mid \boldsymbol{x}_i)$$

Use as the default link function for binary responses.

Probit Regression:

$$\Phi^{-1}(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad \Phi(\cdot) \text{ is the standard normal CDF}$$

Use when you can think of your binary response as being obtained by thresholding a normally distributed latent variable.

Complementary log-log (cloglog) Regression:

$$\log\{-\log(1-\pi_i)\} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

Use when you can think of your binary response as quantifying whether a count response is nonzero, with the count being Poisson distributed.

http://bayesium.com/which-link-function-logit-probit-or-cloglog/

Multinomial Responses

Generalized Logit Model (Nominal):

$$\log\left(\frac{\pi_{ij}}{\pi_{i0}}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \quad j = 1, \ldots, J$$

$$\pi_{ij} = P(Y_i = j \mid \boldsymbol{x}_i) = \frac{\{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)\}}{1 + \sum_{k=1}^J \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}, \quad \pi_{i0} = 1 - \sum_{k=1}^J \pi_{ik}$$

Can also use this model for ordinal data.

Cumulative Logit Model (Ordinal):

$$\log\left(\frac{P(Y_i \le j)}{P(Y_i > j)}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \quad j = 0, \ldots, J-1$$

Count Responses

Poisson Regression (Likelihood):

$$\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad E[Y_i] = V(Y_i) = \lambda_i$$

λ_i controls the rate at which events happen.

Poisson Regression (Quasi-Likelihood):

$$\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

$$a(\phi) = \phi \text{ instead of } a(\phi) = 1 \implies E[Y_i] = \lambda_i, \quad Var[Y_i] = \phi\lambda_i$$

- Used to correct for overdispersion ($V(Y_i) > E[Y_i]$)
- Estimation of $\boldsymbol{\beta}$ is unchanged from regular Poisson regression
- Standard errors corresponding to $\widehat{\boldsymbol{\beta}}$ are generally larger when outcome is truly overdispersed

Offset:

$$\log(\lambda_i) = \log(T_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad T_i = \text{time over which counts were obtained}$$

Count Response Example in R

Data Example: Seizure Counts for Epileptic Individuals

Study Details (Thall and Vail, 1990):

- $n = 59$ participants with epilepsy.
- Randomized to Progabide ($n_t = 31$) or placebo ($n_p = 28$).
- Number of seizures were recorded during an 8-week baseline period.
- Seizure counts were recorded for 4 successive 2-week periods.

Primary Research Question:

- Is Progabide use associated with fewer numbers of seizures in epileptic individuals during the final two week period of follow-up?

Data Example: Seizure Counts for Epileptic Individuals

```r
#Read in data and load necessary libraries
library(MASS)
library(dplyr)
library(ggplot2)
library(grid)
library(gridExtra)
data("epil") #Type ?epil to see dataset details
epil.follow.up.4 <- epil %>% filter(period == 4)
```

Variables in Dataset:

- y: seizure count for the corresponding two week period
- trt: treatment, either placebo or Progabide
- base: seizure count in the 8-week baseline period
- age: individual's age in years
- V4: binary (0, 1) indicator variable for the 4th period
- subject: subject identifier, 1 to 59
- period: indicator of the two-week time period (1, 2, 3 or 4)
- lbase: log-counts for the baseline period, centered to have mean zero
- lage: log-age, centered to have mean zero

Poisson Regression Model with Canonical Link

We will use a Poisson regression model, since we have a count response.

Random Component:

$$Y_i \sim Poisson(\lambda_i)$$

Systematic Component and Link Function:

$$\log(\lambda_i) = \beta_0 + \beta_t trt_i + \beta_a age_i + \beta_b base_i$$

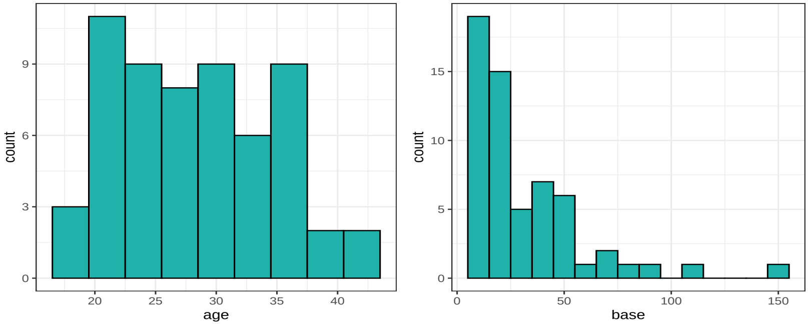
Note that we do not need to be concerned with an offset term, because the follow-up time is the exact same for all individuals!

Descriptive Statistics

```
#Data Exploration
h1 <- ggplot(data = epil.follow.up.4, aes(x = age)) +
  geom_histogram(binwidth = 3, fill = "lightseagreen", color = "black") + theme_bw()

h2 <- ggplot(data = epil.follow.up.4, aes(x = base)) +
  geom_histogram(binwidth = 10, fill = "lightseagreen", color = "black") + theme_bw()

grid.arrange(h1, h2, nrow = 1, ncol = 2)
```



Need to log-transform the baseline number of seizures!

Descriptive Statistics

```
#Crude Overdispersion Check
c(mean(epil.follow.up.4$y), var(epil.follow.up.4$y))
```

Note that the empirical variance (93.1) $\gg$ empirical mean (7.3)!

This suggests that we will need to account for overdispersion.

Updated Regression Model

Random Component:

$$f(Y_i; \lambda_i, \phi) = \exp\left\{ \frac{Y_i \log(\lambda_i) - \lambda_i}{\phi} - \log(Y_i!) \right\}$$

Note that we have added overdispersion parameter ϕ.

Systematic Component and Link Function:

$$\log(\lambda_i) = \beta_0 + \beta_t trt_i + \beta_a age_i + \beta_b lbase_i$$

Note that we are now adjusting for *lbase* instead of *base*.

Not Accounting for Overdispersion

```r
#Regular Poisson Regression
poisson.reg.full <- glm(y ~ factor(trt) + age + lbase, family = "poisson", data = epil.follow.up.4)
summary(poisson.reg.full)
```

```
##
## Call:
## glm(formula = y ~ factor(trt) + age + lbase, family = "poisson",
##     data = epil.follow.up.4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5962  -1.1318   0.1552   0.8062   3.6635
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.37181    0.26731   5.132 2.87e-07 ***
## factor(trt)progabide -0.15726    0.10144  -1.550    0.121
## age                   0.01100    0.00823   1.337    0.181
## lbase                 1.17365    0.06819  17.211  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 476.25  on 58  degrees of freedom
## Residual deviance: 145.98  on 55  degrees of freedom
## AIC: 341.74
##
## Number of Fisher Scoring iterations: 5
```

Accounting for Overdispersion

```r
#With Correction for Overdispersion
poisson.reg.full <- glm(y ~ factor(trt) + age + lbase, family = "quasipoisson", data = epil.follow.up.4)
summary(poisson.reg.full)
```

```
##
## Call:
## glm(formula = y ~ factor(trt) + age + lbase, family = "quasipoisson",
##     data = epil.follow.up.4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5962  -1.1318   0.1552   0.8062   3.6635
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.37181    0.42241   3.248  0.00199 **
## factor(trt)progabide -0.15726    0.16030  -0.981  0.33087
## age                   0.01100    0.01301   0.846  0.40116
## lbase                 1.17365    0.10776  10.892  2.4e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.497075)
##
##     Null deviance: 476.25  on 58  degrees of freedom
## Residual deviance: 145.98  on 55  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Interpretation of Progabide Coefficient

Treatment Effect Estimate:

$$\widehat{\beta}_t = -0.157$$

Mathematical Meaning of Treatment Effect:

$$\log(E[Y_i \mid trt_i = 1]) - \log(E[Y_i \mid trt_i = 0]) = \beta_t$$

Interpretation: Progabide lowers the log of the expected number of seizures by 0.157 when compared with the placebo, adjusted for age and the number of baseline seizures.

Not a very intuitive interpretation!

Interpretation of Progabide Coefficient

Rate Ratio:

$$e^{\widehat{\beta}_t} = 0.854$$

Mathematical Meaning of Rate Ratio:

$$\frac{E[Y_i \mid trt_i = 1]}{E[Y_i \mid trt_i = 0]} = e^{\beta_t}$$

Interpretations:

(i) A person using Progabide is expected to have 85.4% of the number of seizures as they would using the placebo, adjusted for age and the number of baseline seizures.

(ii) A person using Progabide is expected to have 14.6% fewer seizures than they would using the placebo, adjusted for age and the number of baseline seizures.

Predictions

General Formula for GLM Predictions:

$$\widehat{Y}_i = g^{-1}(\boldsymbol{x}_i^\top \widehat{\beta})$$

The predicted value for the first participant is:

$$Y_1 = 3, \quad \widehat{Y}_1 = \exp(\widehat{\beta}_0 + \widehat{\beta}_t \times 0 + \widehat{\beta}_a \times 31 + \widehat{\beta_b} \times -0.7563538) = 2.28$$

```r
#Predicted number of seizures in the final two-week follow-up period value for the first participant
pred.obs <- epil.follow.up.4[1,]
eta.1.hat <- predict.glm(poisson.reg.full, newdata = pred.obs)
Y.1.hat <- exp(eta.1.hat)
Y.1 <- pred.obs$y
```

Inference (Confidence Intervals)

```
#Get 95% Confidence Interval for Treatment
ci95.beta <- confint(poisson.reg.full)
ci95.beta.t <- ci95.beta[row.names(ci95.beta) == "factor(trt)progabide",]
ci95.rr <- exp(ci95.beta.t)
ci95.rr
```

Interpretation: The probability that the true rate ratio is between 0.62 and 1.17 is 0.95.

Many other inferential techniques you can employ with GLMs!

Summary

- GLMs are useful for modeling many different types of responses
- Requires Specification of:
  - A random component from the exponential family
  - Systematic component
  - Link function
- Many of the concepts that apply to multivariable linear regression continue to apply when using GLMs.

Contact Information

E-mail: bossjona@umich.edu

References

P. F. Thall and S. C. Vail. Some covariance models for longitudinal count data with over-dispersion. *Biometrics*, 46(3):657–671, 1990.