Generalized Linear Models: Advanced Topic

BDSI 2019

Holly Hartman

6/19/2019

Review of Previous Models:

- Linear regression:

Review of Previous Models:

- Linear regression: Appropriate when Y is continuous and the errors are normally distributed
- Logistic regression:

Review of Previous Models:

- Linear regression: Appropriate when Y is continuous and the errors are normally distributed
- Logistic regression: Appropriate when Y is Bernoulli.

Other types of outcomes:

- Count data - How many times does a patient visit the hospital for a particular condition?

Other types of outcomes:

- Count data - How many times does a patient visit the hospital for a particular condition?
- Rates - How many time does a patient experience heartburn per month?

Other types of outcomes:

- Count data - How many times does a patient visit the hospital for a particular condition?
- Rates - How many time does a patient experience heartburn per month?
- Multinomial - What influences choice of mode of commuting to work (car, bike, walking, public transportation, etc)?

Other types of outcomes:

- Count data - How many times does a patient visit the hospital for a particular condition?
- Rates - How many time does a patient experience heartburn per month?
- Multinomial - What influences choice of mode of commuting to work (car, bike, walking, public transportation, etc)?
- Ordered categorical - Can we predicting BMI category (underweight, normal weight, overweight, obese) based on genetics?

Other types of outcomes:

- Count data - How many times does a patient visit the hospital for a particular condition?
- Rates - How many time does a patient experience heartburn per month?
- Multinomial - What influences choice of mode of commuting to work (car, bike, walking, public transportation, etc)?
- Ordered categorical - Can we predicting BMI category (underweight, normal weight, overweight, obese) based on genetics?
- Time-to-Event (censored) - Does the treatment for a particular disease improve 5-year survival?

Other types of outcomes:

- Count data - How many times does a patient visit the hospital for a particular condition?
- Rates - How many time does a patient experience heartburn per month?
- Multinomial - What influences choice of mode of commuting to work (car, bike, walking, public transportation, etc)?
- Ordered categorical - Can we predicting BMI category (underweight, normal weight, overweight, obese) based on genetics?
- Time-to-Event (censored) - Does the treatment for a particular disease improve 5-year survival?

Overview of Generalized Linear Models

Flexible method to fit outcomes from different distributions. The family of distributions that is valid for GLM is called the exponential family.

The expected value of Y, conditional on covariates X, is modeled using a **function** of a linear model.

Linear regression and logistic regression are both examples of generalized linear models.

GLM components

1. Exponential family of distribution, the random component

Y_i assumed to follow canonical exponential family
$f(Y_i|\theta, \phi) = \exp\left(\frac{Y_i\theta - b(\theta)}{a(\phi)} + c(Y, \phi)\right)$

2. Linear predictor, the systematic component

$\eta_i \equiv \boldsymbol{x}_i^T \beta$

3. Link function, g

Connect $\boldsymbol{x}_i$ and $\boldsymbol{\mu}_i$ such that $E(Y_i|\boldsymbol{x}) = \mu = g^{-1}(\eta_i)$.

GLM components

1. **Exponential family of distribution, the random component**

Y_i assumed to follow canonical exponential family
$f(Y_i|\theta,\phi) = \exp\left(\frac{Y_i\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right)$

2. Linear predictor, the systematic component

$\eta_i \equiv \boldsymbol{x}_i^T\beta$

3. Link function, g

Connect $\boldsymbol{x}_i$ and $\boldsymbol{\mu}_i$ such that $E(Y_i|\boldsymbol{x}) = \mu = g^{-1}(\eta_i)$.

Distributions in exponential family

Exponential family is distributions that can be written in the form
$f(Y|\theta, \phi) = \exp\left(\frac{t(Y)\theta - b(\theta)}{a(\phi)} + c(Y, \phi)\right)$

- θ is the canonical parameter, typically unknown (location, mean)
- ϕ is the dispersion parameter, typically known (scale, variance)

To show a distribution is an exponential family, then it just needs to be rearranged to match the above formula. Depending on the number of parameters, θ and ϕ can be vectors.

Example: Showing the normal distribution is an exponential family

Assume unknown mean, known variance (can be shown other ways, but this is for simplicity)

$$f(Y) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(y-\mu)^2/2\sigma^2)$$

Example: Showing the normal distribution is an exponential family

Assume unknown mean, known variance (can be shown other ways, but this is for simplicity)

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

Example: Showing the normal distribution is an exponential family

Assume unknown mean, known variance (can be shown other ways, but this is for simplicity)

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

Example: Showing the normal distribution is an exponential family

$$f(Y)\exp\left(\frac{\mu y-\mu^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi)-\log(\sigma)\right)$$

$t(Y)=y$

$b(\theta)=\mu^2/2$

$a(\phi)=\sigma^2$

$c(Y,\phi)=\frac{y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi)-\log(\sigma)$

Example: Showing the Poisson distribution is an exponential family

$$p(Y) = \frac{e^{-\lambda}\lambda^{Y}}{Y!}$$

Example: Showing the Poisson distribution is an exponential family

$$
\begin{aligned}
p(Y) &= \frac{e^{-\lambda}\lambda^{Y}}{Y!} \\
&= \frac{\exp\left(Ylog(\lambda) - \lambda\right)}{Y!}
\end{aligned}
$$

Example: Showing the Poisson distribution is an exponential family

$$
\begin{aligned}
p(Y) &= \frac{e^{-\lambda}\lambda^{Y}}{Y!} \\
&= \frac{\exp\left(Ylog(\lambda) - \lambda\right)}{Y!} \\
&= \exp\left(Ylog(\lambda) - \lambda - log(Y!)\right)
\end{aligned}
$$

Example: Showing the Poisson distribution is an exponential family

$$P(Y) = \exp\left(Ylog(\lambda) - \lambda - log(Y!)\right)$$

$t(Y) = Y$

$\theta = log(\lambda)$

$b(\theta) = \lambda = e^{\theta}$

$a(\phi) = 1$

$c(Y, \phi) = -log(Y!)$

Exponential family: Means and variances

$$E(Y_i) \equiv \mu = b'(\theta)$$

$$V(Y_i) = b''(\theta)a(\phi)$$

Exponential family: Means and variances - Example

From before, for the normal distribution, $\theta = \mu$, $b(\theta) = \mu^2/2$, and $a(\phi) = \sigma^2$

$E(Y_i) = b'(\theta) = \mu$

$V(Y_i) = b''(\theta)a(\phi) = \sigma^2$

Exponential family: Means and variances - Example

From before, for the Poisson distribution, $\theta = log(\lambda)$, $b(\theta) = e^{\theta}$, and $a(\phi) = 1$

$E(Y_i) = b'(\theta) = e^{\theta} = \lambda$

$V(Y_i) = b''(\theta)a(\phi) = e^{\theta} = \lambda$

Exponential family: Distributions

Many common distributions are exponential families

- Normal
- Exponential
- Poisson
- Bernoulli
- Beta
- Chi-squared

There are more that are exponential families, with some constraints

- Binomial (fixed # trials)
- Multinomial (fixed # trials)
- Negative binomial (fixed # failures)

GLM components

1. Exponential family of distribution, the random component

Y_i assumed to follow canonical exponential family
$f(Y_i|\theta, \phi) = \exp\left(\frac{Y_i\theta - b(\theta)}{a(\phi)} + c(Y, \phi)\right)$

2. Linear predictor, the systematic component

$\eta_i \equiv \boldsymbol{x}_i^T \beta$

3. **Link function,** g

Connect $\boldsymbol{x}_i$ and $\boldsymbol{\mu}_i$ such that $E(Y_i|\boldsymbol{x}) = \mu = g^{-1}(\eta_i)$.

Link functions

The link function describes how the mean response $E(Y_i) = \mu_i$ is related to the covariates.

The link function is such that $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i$.

Recall that $E(g(Y)) \neq g(E(Y))$. We will be working with $g(E(Y))$, not $E(g(Y))$.

g must be monotone (non-decreasing OR non-increasing)

g must be differentiable

Examples of valid link functions

- Identity $\eta_i = \mu_i$
- Logit $\eta_i = log\left(\frac{\mu_i}{1-\mu_i}\right)$
- Log $\eta_i = log(\mu_i)$
- Probit $\eta_i = \Phi^{-1}(\mu_i)$

Link Functions

Using different link functions results in different interpretations of covariate estimates

Identity link - additive effect of covariates

Log link - multiplicative effect of covariates (odds)

Logit link - multiplicative effect of covariates (odds ratio)

The canonical link

The canonical link function is such that $\eta = g(\mu) = \theta$ where θ is the canonical parameter of the exponential family distribution. This has nice properties and tends to be the default, but it is not necessary to use.

Choosing a different link function may result in estimates that are not feasible (negative probabilities, for example).

Distribution	Canonical Link
Normal	Identity
Binomial	Logit
Poisson	Log

Canonical link - Examples

Normal:

$E(Y) = \mu$, also $\mu = \theta$. $\theta = E(Y)$ and so the identity link function is the canonical link.

Poisson:

From before, $E(Y) = \lambda$ and $e^{\theta} = \lambda$. Then we see that $\theta = \log(\lambda)$. It follows that $\theta = \log(E(Y))$ and so the log link is the canonical link.

Example using count data in R

This data is from a study examining counts of seizures in people with epilepsy. The number of seizures were measured during an 8 week baseline period. Then counts were recorded for 4 successive 2-week periods.

Example using count data in R

```r
library(MASS)
?epil
```

y - the count for the 2-week period.

trt - treatment, “placebo” or “progabide”.

base - the counts in the baseline 8-week period.

age - subject’s age, in years.

V4 - $0/1$ indicator variable of period 4.

subject - subject number, 1 to 59.

period - period, 1 to 4.

lbase - log-counts for the baseline period, centred to have zero mean.

lage - log-ages, centered to have zero mean.

Data exploration

```
epil2<-epil[epil$period == 4, ]
summary(epil2)
```

```
##        y                  trt          base             age       
##  Min.   : 0.000   placebo  :28   Min.   :  6.00   Min.   :18.00  
##  1st Qu.: 3.000   progabide:31   1st Qu.: 12.00   1st Qu.:23.00  
##  Median : 4.000                  Median : 22.00   Median :28.00  
##  Mean   : 7.305                  Mean   : 31.22   Mean   :28.34  
##  3rd Qu.: 8.000                  3rd Qu.: 41.00   3rd Qu.:32.00  
##  Max.   :63.000                  Max.   :151.00   Max.   :42.00  
##        V4       subject        period      lbase         
##  Min.   :1   Min.   : 1.0   Min.   :4   Min.   :-1.36249  
##  1st Qu.:1   1st Qu.:15.5   1st Qu.:4   1st Qu.:-0.66934  
##  Median :1   Median :30.0   Median :4   Median :-0.06321  
##  Mean   :1   Mean   :30.0   Mean   :4   Mean   : 0.00000  
##  3rd Qu.:1   3rd Qu.:44.5   3rd Qu.:4   3rd Qu.: 0.55932  
##  Max.   :1   Max.   :59.0   Max.   :4   Max.   : 1.86303  
##       lage          
##  Min.   :-0.42941  
##  1st Qu.:-0.18429  
##  Median : 0.01242  
```

Modeling

Recall:

$$E(Y) \equiv \mu$$

We are going to use a log link since this is count data which is modeled using the Poisson distribution

$$\log(\mu_i) = X_i^T \beta$$

We will use the model:

$$\log(\mu_i) = \beta_0 + \beta_1 Trt_i + \beta_2 Age_i + \beta_3 Base_i$$

Modeling

```
summary(glm(y ~ trt + age + base, family = poisson,
            data = epil2))
```

```
##
## Call:
## glm(formula = y ~ trt + age + base, family = poisson, data = epil2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1636  -1.0246  -0.1443   0.4865   3.8993
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.775574   0.284598   2.725  0.00643 **
## trtprogabide -0.270482   0.101868  -2.655  0.00793 **
## age           0.014044   0.008580   1.637  0.10169
## base          0.022057   0.001088  20.267  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

Interpretation of a Poisson Model - categorical covariate

Treatment effect estimate: -0.2705

"Progabide changes $\log(\mu)$ by -0.2705 more than the placebo." This is hard to interpret.

Interpretation of a Poisson Model - categorical covariate

Treatment effect estimate: -0.2705

"Progabide changes $\log(\mu)$ by -0.2705 more than the placebo." This is hard to interpret.

"Progabide lowers the log of the expected number of seizures by 0.2705."

Interpretation of a Poisson Model - categorical covariate

Treatment effect estimate: -0.2705

"Progabide changes $\log(\mu)$ by -0.2705 more than the placebo." This is hard to interpret.

"Progabide lowers the log of the expected number of seizures by 0.2705."

Since this is a log model then we can use $\exp(\beta)$ and this is an estimate of the rate ratio.

"A person using Progabide is expected to have $\exp(-0.2705) \times 100 = 76.3\%$ of the number of seizures as a person using the placebo."

Interpretation of a Poisson Model - categorical covariate

Treatment effect estimate: -0.2705

"Progabide changes $\log(\mu)$ by -0.2705 more than the placebo." This is hard to interpret.

"Progabide lowers the log of the expected number of seizures by 0.2705."

Since this is a log model then we can use $\exp(\beta)$ and this is an estimate of the rate ratio.

"A person using Progabide is expected to have $\exp(-0.2705) \times 100 = 76.3\%$ of the number of seizures as a person using the placebo."

"A person using Progabide is expected to have $(1 - \exp(-0.2705)) \times 100 = 23.7\%$ fewer seizures than a person using the placebo."

Interpretation of a Poisson Model - continuous covariate

Base # seizures estimate: 0.022057

Interpretation of a Poisson Model - continuous covariate

Base # seizures estimate: 0.022057

"Increasing the base number of seizures by 1 increases $\log(\mu)$ by 0.022057"

"Increasing the base number of seizures by 1 is expected to result in $(\exp(-0.2705) - 1) \times 100 = 2.2\%$ more seizures."

Interpretation of a Poisson Model - scaling parameters

Is 1 more seizure at baseline a good unit of measurement?

```
summary(epil2$base)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   12.00   22.00   31.22   41.00  151.00
```

```
sd(epil2$base)
```

```
## [1] 26.87716
```

Interpretation of a Poisson Model - scaling parameters

Is 1 more seizure at baseline a good unit of measurement?

```
summary(epil2$base)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   12.00   22.00   31.22   41.00  151.00
```

```
sd(epil2$base)
```

```
## [1] 26.87716
```

"Increasing the base number of seizures by 10 is expected to result in $(\exp(10 \times -0.2705) - 1) \times 100 = 24.7\%$ more seizures."

"Increasing the base number of seizures by 20 is expected to result in $(\exp(20 \times -0.2705) - 1) \times 100 = 55.4\%$ more seizures."

Estimating counts from parameters

Expected number of seizures at the conclusion of the trial for a 28 year old person, with 22 seizures during the baseline period on the placebo is:

$$\begin{aligned}\exp(\beta_0 + \beta_2 \times 28 + \beta_3 \times 22) &= \exp(0.7756 + 0.0140 \times 28 + 0.0220 \times 22) \\ &= \exp(1.652) \\ &= 5.22\end{aligned}$$

The intercept would be the expected number of counts for a person on the placebo of age 0 with 0 seizures at baseline - somewhat nonsensical! We could center our covariates to fix this.

Example of rate data in R

We are going to be using data on non-melanoma skin cancer cases.

Two cities: Minneapolis and Dallas

Eight age ranges: 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 84+

Source: https://rpubs.com/kaz_yos/poisson

Example of rate data in R - creating data set

```
## Create a dataset manually
nonmel <- read.table(header = TRUE,
                     text = "
   cases city u1 u2 u3 u4 u5 u6 u7      n
1      1    0  1  0  0  0  0  0  0 172675
2     16    0  0  1  0  0  0  0  0 123065
3     30    0  0  0  1  0  0  0  0  96216
4     71    0  0  0  0  1  0  0  0  92051
5    102    0  0  0  0  0  1  0  0  72159
6    130    0  0  0  0  0  0  1  0  54722
7    133    0  0  0  0  0  0  0  1  32185
8     40    0  0  0  0  0  0  0  0   8328
9      4    1  1  0  0  0  0  0  0 181343
10    38    1  0  1  0  0  0  0  0 146207
11   119    1  0  0  1  0  0  0  0 121374
12   221    1  0  0  0  1  0  0  0 111353
13   259    1  0  0  0  0  1  0  0  83004
14   310    1  0  0  0  0  0  1  0  55932
15   226    1  0  0  0  0  0  0  1  29007
16    65    1  0  0  0  0  0  0  0   7583
")
```

Example of rate data in R - creating data set

```r
## Create age.range variable and city variable
nonmel <- within(nonmel, {
    age.range <- rep(c("15_24",
                       "25_34",
                       "35_44",
                       "45_54",
                       "55_64",
                       "65_74",
                       "75_84",
                       "85+"), 2)
    age.range <- factor(age.range)
    age.range <- relevel(age.range, ref = "85+")

    city <- factor(city, 0:1, c("Minneapolis", "Dallas"))
})

## rop unnecessary columns
nonmel <- nonmel[c("cases","n","city","age.range")]
```

Example of rate data in R - creating data set

```
nonmel
```

```
##     cases      n        city age.range
## 1       1 172675 Minneapolis     15_24
## 2      16 123065 Minneapolis     25_34
## 3      30  96216 Minneapolis     35_44
## 4      71  92051 Minneapolis     45_54
## 5     102  72159 Minneapolis     55_64
## 6     130  54722 Minneapolis     65_74
## 7     133  32185 Minneapolis     75_84
## 8      40   8328 Minneapolis       85+
## 9       4 181343      Dallas     15_24
## 10     38 146207      Dallas     25_34
## 11    119 121374      Dallas     35_44
## 12    221 111353      Dallas     45_54
## 13    259  83004      Dallas     55_64
## 14    310  55932      Dallas     65_74
## 15    226  29007      Dallas     75_84
## 16     65   7583      Dallas       85+
```

Rate data - offsets

Problem: Number of cases will heavily depend on the population of the cities and the number of people in each age range.

To account for this, we will use a Poisson model (since this is count data) with an offset.

Offsets

Typical Poisson model:

$$\log(\mu) = \boldsymbol{x}^T \boldsymbol{\beta}$$
$$\mu = \exp(\boldsymbol{x}^T \boldsymbol{\beta})$$

Let the count response Y have an index t, then the sample rate is Y/t. Then the expected value of the rate is μ/t where μ is the expected count.

Poisson model with offset:

$$\log(\mu/t) = \boldsymbol{x}^T \boldsymbol{\beta}$$
$$\log(\mu) = \log(t) + \boldsymbol{x}^T \boldsymbol{\beta}$$
$$\mu = t \exp(\boldsymbol{x}^t \boldsymbol{\beta})$$

Poisson rate model

```
summary(glm(cases ~ city + age.range, offset = log(n), family = poisson,
            data = nonmel))
```

```
##
## Call:
## glm(formula = cases ~ city + age.range, family = poisson, data = nonmel,
##     offset = log(n))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.50598  -0.48566   0.01639   0.36926   1.24763
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.4834     0.1037 -52.890  < 2e-16 ***
## cityDallas       0.8039     0.0522  15.399  < 2e-16 ***
## age.range15_24  -6.1742     0.4577 -13.488  < 2e-16 ***
## age.range25_34  -3.5440     0.1675 -21.160  < 2e-16 ***
## age.range35_44  -2.3268     0.1275 -18.254  < 2e-16 ***
## age.range45_54  -1.5790     0.1138 -13.871  < 2e-16 ***
## age.range55_64  -1.0869     0.1109  -9.800  < 2e-16 ***
## age.range65_74  -0.5288     0.1086  -4.868 1.13e-06 ***
## age.range75_84  -0.1157     0.1109  -1.042    0.297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the coefficients

Basically, the same as a Poisson model.

"The risk of non-melanoma skin cancer is $\exp(0.8039) = 2.23$ times is higher in Dallas than Minneapolis."

Estimate rates given covariates

The main difference is that estimates using the coefficients are estimates of the rate, not counts.

The rate is in the same units as the offset. Since our offset is people, then the rate is per person.

"The rate of non-melanoma skin cancer among people age 65-74 in Dallas is $\exp(-5.4834 + 0.8039 - 0.5288) = 0.0054$ per person."

Estimate rates given covariates

The main difference is that estimates using the coefficients are estimates of the rate, not counts.

The rate is in the same units as the offset. Since our offset is people, then the rate is per person.

"The rate of non-melanoma skin cancer among people age 65-74 in Dallas is $\exp(-5.4834 + 0.8039 - 0.5288) = 0.0054$ per person."

Adjust this by multiplying the estimated effect by a different unit.

"The rate of non-melanoma skin cancer among people age 65-74 in Dallas is $\exp(-5.4834 + 0.8039 - 0.5288) \times 1000 = 5.4$ per 1000 people."