Week3 GLM Part 2

Continue Binomial (Logistic) Regression

$logit(\pi) = \log \frac{\pi}{1-\pi}$ (log odds)

$\pi = \exp(logit(\pi)) = \exp(\log \frac{\pi}{1-\pi}) = \frac{\pi}{1-\pi}$ (odds)

Log odds 之间是差的关系，加减，一个比另一个多多少。

odds 之间是比例关系，一个是另一个的多少倍，百分之多少。

For example, β_7 represents `ruralUrban`, $x_7 = 1$ if rural, 0 if urban.

Rural: log odds$= \beta_0 + \beta_1 x_1 + \cdots + \beta_7 \cdot 1$

Rural: log odds$= \beta_0 + \beta_1 x_1 + \cdots + \beta_7 \cdot 0$

$[\log \text{odds}]_{rural} - [\log \text{odds}]_{urban} = \beta_7 = 0.43$， 差的关系

$\text{odds}_{rural} = \exp(\beta_0 + \cdots + \beta_7)$ Since $x_7 = 1$

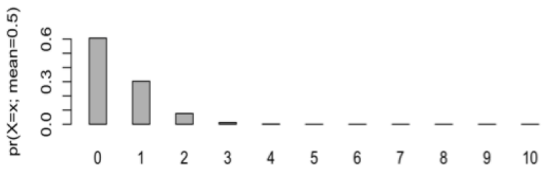
$\text{odds}_{rural} = \exp(\beta_0 + \cdots + 0)$ Since $x_7 = 0$

$\text{odds ratio} = \frac{\text{odds}_{rural}}{\text{odds}_{urban}} = \exp(\beta_7)$, 比例关系

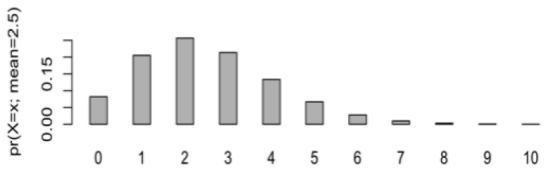
odds ratio: 保持其他条件(x_i) 不变，只改变一个x, 对应的β所能带来的两种情况的ratio。

Poisson Regression

$$X_i \sim \text{Poisson}(\lambda_i)$$
$$pr(X_i = x) = \lambda_i^x \exp(-\lambda_i)/x!$$
$$X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$$



$\lambda = 0.5$



$\lambda = 2.5$

Poisson Distribution's mean and variance should be the same.

	Poisson distribution	Normal distribution
Parameter	λ	μ, σ^2
Use case	Usually used for count or rate data	Continuous data most appropriate
Shape	Skew (depending on the value of λ	Symmetric around the mean ("bell-curve")
Parameter note	Variance = mean (same parameter)	Variance and mean different parameters

Components of Poisson GLM

The components of a GLM for a count response are:

1. **Random Component**: $Y_i \sim \text{Poisson}(\lambda)$ and $E(Y) = \text{var}(Y) = \mu$.
2. **Systematic component**: Your model matrix and parameters, $\boldsymbol{X\beta}$
3. **Link**:

- We could use an **identity Link** which would give us:

$$\mu = \boldsymbol{X\beta}$$

  But, just as for binomial data, with an identity link the model can yield $\mu < 0$ where we only want $\mu \geq 0$.

Or we could use:

- The **log link** (most common and the canonical* link)

$$\log(\mu) = \boldsymbol{X\beta}$$

  *Canonical link functions have nice theoretical properties that make them often the most popular link function to use. We won't go deeper than that in this course.

Offsets

Offsets allow us adjust for the time period under consideration. That is, the exposure period.

An offset term can be thought of as the log of the time period under study when we are doing Poisson regression with a log link.

E.g.,

$$\log(\frac{children}{month}) = \boldsymbol{X\beta}$$

$$\log(children) = \boldsymbol{X\beta} + \log(month)$$

When using an offset like this we often say we are fitting a *rate model*.

You will need the `offset()` function in your model formula.

Code

```
fiji_model = glm(
  formula = children ~ offset(logYears) + ageMarried + literacy,
  family=poisson(link=log),
  data=fijiSub)
```

```
summary(fiji_model)
```

```
##
## Call:
## glm(formula = children ~ offset(logYears) + ageMarried + literacy,
##     family = poisson(link = log), data = fijiSub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3119  -0.5992   0.1261   0.6917   6.7558
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.211564   0.013581 -89.208  < 2e-16
## ageMarried0to15   -0.118943   0.020731  -5.737 9.61e-09
## ageMarried18to20   0.033068   0.020308   1.628    0.103
## ageMarried20to22   0.007782   0.023933   0.325    0.745
## ageMarried22to25  -0.004174   0.029553  -0.141    0.888
## ageMarried25to30   0.050164   0.047446   1.057    0.290
## ageMarried30toInf  0.122959   0.097554   1.260    0.208
## literacyno        -0.004526   0.016978  -0.267    0.790
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5746.5  on 4907  degrees of freedom
## Residual deviance: 5687.1  on 4900  degrees of freedom
## AIC: 19354
##
```

"Dispersion parameter for poisson family taken to be 1" means, the mean and variance for the poisson model are approximately equal $\rightarrow$ good model.

Gamma GLM

Random Component: $Y_i \sim Gamma(\phi, \nu), E(Y) = \phi\nu, var(Y) = \phi^2\nu$

Systematic Component: Model matrix and parameters. $X\beta$

Link:

log link: $\log(\mu) = X\beta$

inverse link: $\frac{1}{\mu} = \mu^{-1} = X\beta$

Diagnostics of GLMs

1. Assessing model fit is diicult for binary and count data!
2. Residuals don't have nice properties
3. Histograms can be useful, as can exploratory plots
4. A goodness of fit test is a test of how much data you have

Notes on GLMs with continuous data

- It's not easy to test which distribution is best, since they are not nested.
- GLMs are **not** often used with continuous data
  - they're almost always Binomial or Poisson
  - with the notable exception of the Weibull for event times
- 'standard practice' is to transform continuous data to normality (logs, Box-Cox)

Summary

Distribution of response, Y_i	Support of distribution	Use case	Link function name	Link function $X\beta = h(\mu)$	Mean function
Normal	$(-\infty, \infty)$ (reals)	Linear response*	Identity	$\boldsymbol{X\beta} = \mu$	$\mu = \boldsymbol{X\beta}$
Binomial	{0, 1} (integers, restricted)	Count of yes/no (TRUE/FALSE, 0/1) responses out of fixed number of trials	Logit	$\boldsymbol{X\beta} = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(\boldsymbol{X\beta})}{1+\exp(\boldsymbol{X\beta})}$
Poisson	0, 1, 2, 3, ... (integers)	Count of occurrences in fixed time or space	Log	$\boldsymbol{X\beta} = \ln(\mu)$	$\mu = \exp(\boldsymbol{X\beta})$
Gamma	$(0, \infty)$ (positive reals)	Wait times	Inverse log	$\boldsymbol{X\beta} = \mu^{-1}$ $X\beta = \log(\mu)$	$\mu = (\boldsymbol{X\beta})^{-1}$

Constraints on the vector of regression coefficients β in restricted model

In a likelihood ratio test, two models are required to perform the test. One should be nested within the other.

What are the constraints on the vector of regression coefficients, β?

- The constraints are on the variables in the complicated model but not in the simple model.
- For example, for a binary x, such as literacy, the constraint will be setting the $\beta = 0$ for literacy in the complicated model. The purpose is to ignore this x.

Check Assumptions

Linear Model

- Residual Plot: Constant Variance
- QQ-Plot: Normality
- Independence: explain with words

GLM

Check histogram to see if data follow a specific distribution.

判断一个variable是否有必要的方法

Every variable corresponds to a β, the β has an estimate after fitting a model, and the confidence interval can be obtained. If the confidence interval contain 0, then the variable may not be needed (没有作用).

e^{β} is the odds ratio, $e^{0} = 1$. Thus, if the odds ratio's confidence interval contain 1, the variable may not be needed.

These two methods are equivalent.