STA303H5S - Winter 2014: Data Analysis II

LECTURE 7: Generalized Linear Models (GLM)

Ramya Thinniyam

January 30th, 2014

A Motivating Example: Donner Party Case Study

- In April 1846, a group of 87 pioneers set out for California by wagon train
- Some pioneers got stuck in Sierra Nevada mountains in November due to difficult conditions (harsh weather, unsuitable travel equipment, splits within the group, etc).
- Only some survived
- They were rescued in April 1847

Data: For Adults (15 years of age or older):

Age
Gender
Whether the subject survived or not

Questions of Interest:

1. Were women or men more likely to survive?
2. Were younger pioneers more likely to survive than older ones?

Interested in modelling the odds of survival based on gender and age.

Q: Can we use the usual linear regression / ANOVA to model this? Why or why not?

A:

Linear Regression / ANOVA vs. GLM

Linear Regression / ANOVA:

- Response - quantitative
- Predictors - categorical/quantitative
- Model expected value of response linearly: $E(Y_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$
- Linear relationship between response and predictors
- Errors are normally distributed so response is assumed to be normal
- Gauss-Markov conditions are assumed
- Parameters estimated by Ordinary Least Squares / Maximum Likelihood

Linear Regression / ANOVA vs. GLM

Generalized Linear Models (GLM):

- More flexible framework for modelling responses with different distributions
- Response - quantitative or qualitative
- Predictors - categorical/quantitative
- Model a transformation of a parameter of the response linearly: For ex, $g(E(Y_i)) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$
- Linear relationship between transformation of a parameter of the response and predictors
- Response does not have to be normally distributed: can be Bernoulli, Binomial, Poisson, Gamma, Normal, etc.
- Response does not need to have constant variance
- Parameters estimated by Iteratively Reweighted Least Squares / Maximum Likelihood

Components of a GLM:

1. Random Component: specifies the response and a probability distribution for the response from the Exponential Family.
2. Systematic Component: specifies the linear combination of the predictors. A linear predictor, $\eta = X\beta$.
3. Link Function, g: specifies how the random and systematic components are related. A function g such that $g(E(Y)) = \eta$. The key idea in GLM is the link function which links the mean of the response to the linear predictor.

Response: Y , Predictors / Explanatory variables: $X_1, \ldots, X_p$

Model: $g(E(Y)) = X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

What is the Exponential Family of Distributions?

A distribution from the Exponential Family has the form:

$$f_Y(y|\theta, \tau) = h(y, \tau) \exp\left(\frac{b(\theta)T(y) - A(\theta)}{d(\tau)}\right).$$

- τ: dispersion parameter; usually known and it is related to the variance.
- θ: related to the mean.
- If $b(\theta)$ is the identity function then the distribution is said to be in the “canonical” (natural) form. (Any distribution can be converted to the canonical form by transforming θ.)
- If $T(y)$ is the identity function and τ is known, then θ is called the “canonical parameter” (natural parameter) and the following hold:
  - $\mu = E(Y) = A'(\theta)$
  - $Var(Y) = A''(\theta)d(\tau)$

Examples: Normal, Binomial, Poisson, Exponential, Gamma, Geometric, etc.

Link Functions

The link function is a monotone function, g, that specifies how the mean of the response is related to the explanatory variables in the linear predictor: $g(E(Y)) = X\beta$ or $g(\mu) = X\beta$.

Notes:

- Link function is a transformation of the mean of the response and not of the data
- Model predicts a transformation of the parameter: support of distribution is not necessarily the same type of data as the parameter being predicted
- Usually choose canonical links
- In some cases, domain of the canonical link is not same as the domain of the mean: be careful when doing Maximum Likelihood Estimation or use non-canonical link

Choices for Link Functions:

1. Identity Link: $g(\mu) = \mu$
   Model: $E(Y) = X\beta$: usual linear regression / ANOVA
   Distribution: $Y|X \sim$ Normal $\rightarrow$ STA302 linear regression

2. Log Link: $g(\mu) = \log(\mu)$
   Model: $\log(E(Y)) = X\beta$. $E(Y)$ must be positive.
   "Log-linear" model useful for count data
   Distribution: $Y|X \sim$ Poisson

3. Logit Link: $g(\mu) = \log(\frac{\mu}{1-\mu})$
   Model: $\log\left(\frac{E(Y)}{1-E(Y)}\right) = X\beta$. $0 \leq E(Y) \leq 1$.
   "Logistic" model useful for binary/Binomial data
   Distribution: $Y|X \sim$ Binomial

Logistic Regression

Suppose response is a success or a failure.
$Y|X \sim$ Bernoulli(π), where $\pi = P$(success).
Then, $E(Y|X) = \pi$.

Logit Link: $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \to$ log odds in favour of a success

Model: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$
called "Logistic Regression" Model

Invert: $\pi = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} = \frac{e^{\eta}}{1+e^{\eta}}$
called the "logistic function"

Logistic Function

Q: What does the Logistic function looks like?

A:

Logistic Regression Model

- $E(Y_i|X_{i1},\ldots,X_{ip}) = \pi_i$
- $Var(Y_i|X_{i1},\ldots,X_{ip}) = \pi_i(1-\pi_i)$
- Model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$
- This model does not predict if the response is 0 or 1. It does predict the log odds of the response being 1 (i.e. log odds of a success)
- log odds $\in (-\infty, \infty)$
- As π increases, odds of success and log odds of success increase