COMP 680
Statistics for Computing and Data Science
Week 12: Generalized Linear Models

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

Generalized Linear Models

- A large class of models!
  - where the error distribution is no longer restricted to be Gaussian

Generalized Linear Models

- A large class of models!
  - where the error distribution is no longer restricted to be Gaussian
- The conditional distribution of $Y|X$ follows exponential family
  - linear regression is one type of GLM
  - exponential family: Gaussian, Binomial, Poisson, Gamma, Beta ...

Generalized Linear Models

- A large class of models!
  - where the error distribution is no longer restricted to be Gaussian
- The conditional distribution of $Y|X$ follows exponential family
  - linear regression is one type of GLM
  - exponential family: Gaussian, Binomial, Poisson, Gamma, Beta ...
- Model the conditional mean of $Y|X$
  - $\mathbb{E}[Y|X] = X\beta$ for LM
  - $g(\mathbb{E}[Y|X]) = X\beta$ for GLM
  - different link functions g for specific conditional distributions

LM as GLM

A **generalized linear model** to model conditional mean $\mu = \mathbb{E}[Y|X]$:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

LM as GLM

A **generalized linear model** to model conditional mean $\mu = \mathbb{E}[Y|X]$:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

GLM

- a link function $g(\mu)$
  - describes how the transformed mean depends on the linear predictor
- a variance function $\phi V(\mu)$
  - how the variance of Y depends on the mean μ where ϕ is the dispersion parameter

LM as GLM

A **generalized linear model** to model conditional mean $\mu = \mathbb{E}[Y|X]$:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

GLM

- a link function $g(\mu)$
  - describes how the transformed mean depends on the linear predictor
- a variance function $\phi V(\mu)$
  - how the variance of Y depends on the mean μ where ϕ is the dispersion parameter

LM (a special case)

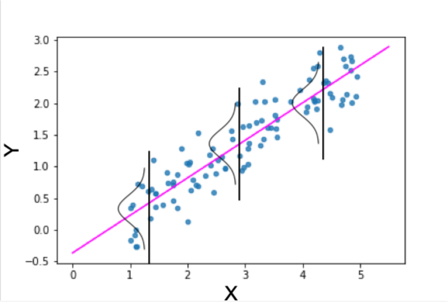
- $g(\mu) = \mu$
  - link function for LM is the identity function, i.e. no transformation needed
- $V(\mu) = 1$
  - for Gaussian distribution, variance does not depend on mean, and $\phi = \sigma^2$

Link Functions

Response Y	Conditional Distribution	Link Function
Continuous	Gaussian	Identity function
Continuous	Gamma	Negative inverse function
Continuous	Inverse Gaussian	Inverse squared function
Binary	Binomial	Logit function
Categorical	Multinomial	(generalized) Logit function
Ordinal	Multinomial	Logit or Probit function
Count	Poisson	Log function
Count	Negative Binomial	Log function

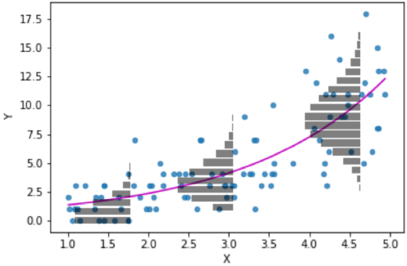
Illustration

- Linear regression



source

- Poisson regression



Outline

Binary Outcome

- Examples of binary responses
  - spam filter
  - fraud detection
  - cancer screening
  - ...

Binary Outcome

- Examples of binary responses
  - spam filter
  - fraud detection
  - cancer screening
  - ...
- In Machine Learning: Binary Classification

Binary Outcome

- Examples of binary responses
  - spam filter
  - fraud detection
  - cancer screening
  - ...
- In Machine Learning: Binary Classification
  - a type of supervised learning
  - Logistic Regression is one classifier
  - many others: KNN, NaiveBayes, DecisionTrees, SVM, ...

Bernoulli and Binomial Distributions

- A Bernoulli random variable Y is a discrete random variable that can only take values 0 or 1:
  - a single parameter p: "success" probability
  - $\mathbb{P}(Y = 1) = p$: "success"
  - $\mathbb{P}(Y = 0) = 1 - p$: "failure"
  - $\mathbb{E}[Y] = p$, $\text{var}(Y) = p(1 - p)$

Bernoulli and Binomial Distributions

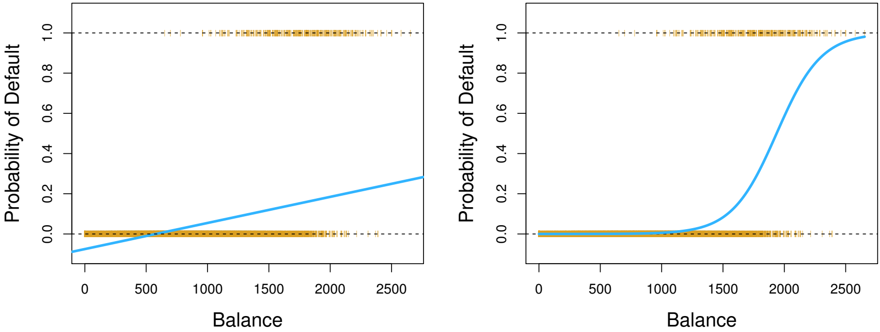
- A Bernoulli random variable Y is a discrete random variable that can only take values 0 or 1:
  - a single parameter p: "success" probability
  - $\mathbb{P}(Y = 1) = p$: "success"
  - $\mathbb{P}(Y = 0) = 1 - p$: "failure"
  - $\mathbb{E}[Y] = p$, $\text{var}(Y) = p(1 - p)$
- A Binomial random variable Y is the sum of n independent Bernoulli's:
  - two parameters: n and p
  - Y is a discrete random variable and can take values from 0 to n
  - $\mathbb{P}(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$: k "success" out of n trials
  - $\mathbb{E}[Y] = np$, $\text{var}(Y) = np(1 - p)$

Linear vs. Logistic Regression

Why not linear regression, model $\mathbb{E}[Y|X] = p_x$ with a straight line?

Linear vs. Logistic Regression

Why not linear regression, model $\mathbb{E}[Y|X] = p_x$ with a straight line?



source

Logit and Logistic

The logit function:

- $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$
- log odds ratio

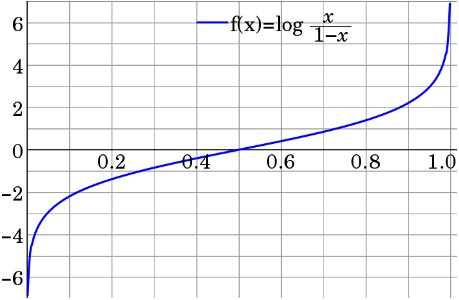
The logistic function:

- $f(x) = \frac{e^x}{1+e^x}$
- Sigmoid function

Logit and Logistic

The logit function:

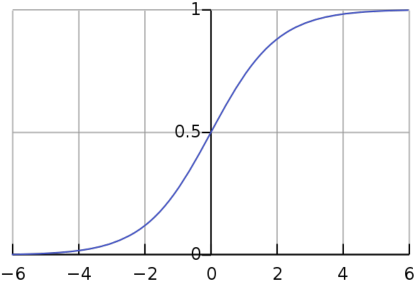
- $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$
- log odds ratio



source

The logistic function:

- $f(x) = \frac{e^x}{1+e^x}$
- Sigmoid function



Simple Logistic Model

$$\mathbb{P}(Y = 1|X) = p_x = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \Leftrightarrow \text{logit}(p_x) = \ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x$$

Simple Logistic Model

$$\mathbb{P}(Y = 1|X) = p_x = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \Leftrightarrow \text{logit}(p_x) = \ln\left(\frac{p_x}{1 - p_x}\right) = \beta_0 + \beta_1 x$$

- remember $\mu = \mathbb{E}[Y|X] = p_x$ for binary Y
- the link function $g(\mu)$ is the logit function
- log odds ratio is a linear function of X!

Model Fitting and Coefficients

- How to estimate model parameters: Maximum Likelihood Estimate
    - likelihood function?

Model Fitting and Coefficients

- How to estimate model parameters: Maximum Likelihood Estimate
  - likelihood function?
- Unlike linear models, does not have closed form solutions
  - numerical optimization - Newton's method

Model Fitting and Coefficients

- How to estimate model parameters: Maximum Likelihood Estimate
    - likelihood function?
- Unlike linear models, does not have closed form solutions
    - numerical optimization - Newton's method
- Point estimate for model coefficients: $\hat{\beta}_0$ and $\hat{\beta}_1$
    - can make inference
        - asymptotic normality of MLE

Model Fitting and Coefficients

- How to estimate model parameters: Maximum Likelihood Estimate
  - likelihood function?
- Unlike linear models, does not have closed form solutions
  - numerical optimization - Newton's method
- Point estimate for model coefficients: $\hat{\beta}_0$ and $\hat{\beta}_1$
  - can make inference
    - asymptotic normality of MLE
  - can make prediction
    - predict log odds $\rightarrow$ probability $\rightarrow$ label

Model Interpretation

Linear Regression:

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{\beta}_0$: average value of Y when $X = 0$
- $\hat{\beta}_1$: average change in Y when X increase by 1 unit
- $\hat{\beta}_1 > 0$ and significant: X is positively correlated with Y

Model Interpretation

Linear Regression:

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{\beta}_0$: average value of Y when $X = 0$
- $\hat{\beta}_1$: average change in Y when X increase by 1 unit
- $\hat{\beta}_1 > 0$ and significant: X is positively correlated with Y

Logistic Regression

- $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{\beta}_0$: log odds of $Y = 1$ when $X = 0$
- $\hat{\beta}_1$: change in log odds of $Y = 1$ when X increase by 1 unit
- $\hat{\beta}_1 > 0$ and significant: X is positively correlated with log odds of $Y = 1$

Multiple Covariates

- simple linear regression $\rightarrow$ multiple regression
  - interpretation of single slope means controlling others

Multiple Covariates

- simple linear regression $\rightarrow$ multiple regression
  - interpretation of single slope means controlling others
- simple logistic regression $\rightarrow$ multiple logistic regression
  - A set of p covariates $X_1, X_2, \cdots X_p$

$$\mathbb{P}(Y = 1|X) = p_x = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \Leftrightarrow$$

$$\text{logit}(p_x) = \ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - interpretation of single slope means controlling others (**holding others constant**), exactly the same!

Prediction

- From predicted probability $\hat{p} = \mathbb{P}(\hat{Y} = 1)$ to predicted outcome $\hat{Y}$
  - a natural threshold 0.5
  - $\hat{p} = \mathbb{P}(\hat{Y} = 1) \geq 0.5 \Rightarrow \hat{Y} = 1$
  - $\hat{p} = \mathbb{P}(\hat{Y} = 1) < 0.5 \Rightarrow \hat{Y} = 0$

Prediction

- From predicted probability $\hat{p} = \mathbb{P}(\hat{Y} = 1)$ to predicted outcome $\hat{Y}$
  - a natural threshold 0.5
  - $\hat{p} = \mathbb{P}(\hat{Y} = 1) \geq 0.5 \Rightarrow \hat{Y} = 1$
  - $\hat{p} = \mathbb{P}(\hat{Y} = 1) < 0.5 \Rightarrow \hat{Y} = 0$
- Pearson residue

$$\frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

Prediction

- From predicted probability $\hat{p} = \mathbb{P}(\hat{Y} = 1)$ to predicted outcome $\hat{Y}$
  - a natural threshold 0.5
  - $\hat{p} = \mathbb{P}(\hat{Y} = 1) \geq 0.5 \Rightarrow \hat{Y} = 1$
  - $\hat{p} = \mathbb{P}(\hat{Y} = 1) < 0.5 \Rightarrow \hat{Y} = 0$
- Pearson residue

$$\frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

- Can vary threshold to minimize error measure
  - sensitivity: true positive rate
  - specificity: true negative rate

Outline

Categorical Outcome

- More than 2 outcomes
  - multi-class classification

Categorical Outcome

- More than 2 outcomes
  - multi-class classification
- Multinomial regression
  - extension on logistic regression
  - multi-class logistic regression

Multinomial Distribution

- A Multinomial random variable (vector) $Y = (Y_1, \cdots Y_K)$
  - K parameters: n and probabilities $(p_1, p_2, \cdots p_K)$
  - $\sum_{k=1}^{K} p_k = 1$
  - Y represents number of k different outcomes

Multinomial Distribution

- A Multinomial random variable (vector) $Y = (Y_1, \cdots Y_K)$
  - K parameters: n and probabilities $(p_1, p_2, \cdots p_K)$
  - $\sum_{k=1}^{K} p_k = 1$
  - Y represents number of k different outcomes
- A Binomial is a special case of Multinomial where $K = 2$
  - $(p_1 = p, p_2 = 1 - p)$

Generalized Logit

- You have total of K categories, pick a category to be baseline

$$\mathbb{P}(Y = j|X) = p_x^j = \frac{e^{\beta_0^j + \beta_1^j X_1 + \cdots + \beta_p^j X_p}}{1 + \sum_{k=1}^{K-1} e^{\beta_0^k + \beta_1^k X_1 + \cdots + \beta_p^k X_p}}, \quad j = 1, \cdots K - 1$$

Generalized Logit

- You have total of K categories, pick a category to be baseline

$$\mathbb{P}(Y=j|X) = p_x^j = \frac{e^{\beta_0^j+\beta_1^j X_1+\cdots+\beta_p^j X_p}}{1+\sum_{k=1}^{K-1} e^{\beta_0^k+\beta_1^k X_1+\cdots+\beta_p^k X_p}}, \quad j=1,\cdots K-1$$

  - $K-1$ set of $p+1$ parameters to estimate
  - Softmax function

Generalized Logit

- You have total of K categories, pick a category to be baseline

$$\mathbb{P}(Y=j|X) = p_x^j = \frac{e^{\beta_0^j+\beta_1^j X_1+\cdots+\beta_p^j X_p}}{1+\sum_{k=1}^{K-1} e^{\beta_0^k+\beta_1^k X_1+\cdots+\beta_p^k X_p}}, \quad j=1,\cdots K-1$$

  - $K-1$ set of $p+1$ parameters to estimate
  - Softmax function

  - which one to be your baseline DOES NOT matter
  - interpretation of coefficients: odds of category j vs. baseline

- Binary Logistic is a special case of $K=2$ and $Y=0$ as baseline

Ordinal Outcome

- Ignore ordinal information and use Multinomial Regression
    - less efficient, $(K-1)(p+1)$ parameters

Ordinal Outcome

- Ignore ordinal information and use Multinomial Regression
  - less efficient, $(K-1)(p+1)$ parameters
- Use ordinal information
  - more efficient, $K+p-1$ parameters

Ordinal Outcome

- Ignore ordinal information and use Multinomial Regression
  - less efficient, $(K-1)(p+1)$ parameters
- Use ordinal information
  - more efficient, $K+p-1$ parameters
- Intuition
  - unlike multinomial case, now the categories are ordered
  - the difference between category $k-1$ and k should be comparable to the difference between category k and $k+1$

Ordinal Logistic Regression

- The Proportional-Odds Cumulative Logit Model

$$\log\left(\frac{\mathbb{P}(Y \leq k|X)}{1-\mathbb{P}(Y \leq k|X)}\right) = \beta_0^k + \beta_1 X_1 + \cdots \beta_p X_p, \quad k = 1, 2, \cdots K-1$$

Ordinal Logistic Regression

- The Proportional-Odds Cumulative Logit Model

$$\log\left(\frac{\mathbb{P}(Y \leq k|X)}{1-\mathbb{P}(Y \leq k|X)}\right) = \beta_0^k + \beta_1 X_1 + \cdots \beta_p X_p, \quad k = 1, 2, \cdots K-1$$

  - estimate $K-1$ intercept and p slopes

Ordinal Logistic Regression

- The Proportional-Odds Cumulative Logit Model

$$\log\left(\frac{\mathbb{P}(Y \leq k|X)}{1-\mathbb{P}(Y \leq k|X)}\right) = \beta_0^k + \beta_1 X_1 + \cdots \beta_p X_p, \quad k = 1, 2, \cdots K-1$$

  - estimate $K-1$ intercept and p slopes
- Interpretation of coefficients:
  - β_0^k is the log odds of falling into or below category k when all $X = 0$
  - β_j is increase in log odds of falling into or below any category associated with a one unit increase in X_j while controlling all other X

Outline

Poisson Distribution

- A Poisson random variable X is discrete and can take any integer values with single parameter λ:

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson Distribution

- A Poisson random variable X is discrete and can take any integer values with single parameter λ:

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- $\mathbb{E}[Y] = \mathrm{Var}(Y) = \lambda$

Poisson Regression

- Link function is natural log

$$\log(\mathbb{E}[Y|X]) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Poisson Regression

- Link function is natural log

$$\log(\mathbb{E}[Y|X]) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Interpretation of slopes:
  - one unit increase of X_j impacts average value of Y by multiplication of e^{β_j} while holding all other X constant

Poisson GLM for Rates

- Response Y has a Poisson distribution, and t is index of the time (or space).
    - count of incidents during specific period of time

Poisson GLM for Rates

- Response Y has a Poisson distribution, and t is index of the time (or space).
  - count of incidents during specific period of time
- In this case, model $\mathbb{E}[Y]/t$ instead:

$$\log(\mathbb{E}[Y]/t) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Poisson GLM for Rates

- Response Y has a Poisson distribution, and t is index of the time (or space).
    - count of incidents during specific period of time
- In this case, model $\mathbb{E}[Y]/t$ instead:

$$\log(\mathbb{E}[Y]/t) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- the term $\log(t)$ is referred to as an offset, an adjustment term
- the mean count is proportional to t

Overdispersion

- In practice, the apparent variance of data often exceeds the mean, reflecting overdispersion in the model parameters.
  - cause covariates to be “more significant”
  - narrower confidence intervals than warranted by the data

Overdispersion

- In practice, the apparent variance of data often exceeds the mean, reflecting overdispersion in the model parameters.
  - cause covariates to be "more significant"
  - narrower confidence intervals than warranted by the data
- Possible solutions
  - quasi-Poisson or negative binomial generalized linear model
  - more parameter and no longer force mean = variance

Outline

Exponential Family

- A single-parameter exponential family PDF (or PMF) can be expressed in the form:

$$f_X(x \mid \theta) = h(x) \exp[\eta(\theta) \cdot T(x) + A(\theta)]$$

Exponential Family

- A single-parameter exponential family PDF (or PMF) can be expressed in the form:

$$f_X(x \mid \theta) = h(x) \exp[\eta(\theta) \cdot T(x) + A(\theta)]$$

- $\eta(\theta)$ is called the natural parameterization and is the link function!
  - each exponential distribution corresponds to a GLM
  - logistic and Poisson are most common

Exponential Family

- General:

$$f_X(x \mid \theta) = h(x) \exp[\eta(\theta) \cdot T(x) + A(\theta)]$$

Exponential Family

- General:

$$f_X(x \mid \theta) = h(x)\ \exp[\eta(\theta) \cdot T(x) + A(\theta)]$$

- Binomial:

$$\begin{aligned}\mathbb{P}(X = x \mid p) &= \binom{n}{k} p^x (1-p)^{n-x} \\ &= \binom{n}{k}\ \exp\left[x \ln\left(\frac{p}{1-p}\right) + n\ln(1-p)\right]\end{aligned}$$

Exponential Family

- General:

$$f_X(x \mid \theta) = h(x) \exp[\eta(\theta) \cdot T(x) + A(\theta)]$$

- Binomial:

$$\begin{aligned}\mathbb{P}(X = x \mid p) &= \binom{n}{k} p^x (1-p)^{n-x} \\ &= \binom{n}{k} \exp\left[x \ln\left(\frac{p}{1-p}\right) + n \ln(1-p)\right]\end{aligned}$$

- Poisson

$$\mathbb{P}(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{1}{x!} \exp[x \ln \lambda - \lambda]$$

Model Fit and Comparison

- What is the goodness-of-fit measure in linear models?

Model Fit and Comparison

- What is the goodness-of-fit measure in linear models?
  - residual sum of squares (RSS)
  - R^2 and adjusted R^2

Model Fit and Comparison

- What is the goodness-of-fit measure in linear models?
  - residual sum of squares (RSS)
  - R^2 and adjusted R^2
- The notion of deviances: $\approx$ - 2log likelihood, smaller $\rightarrow$ better fit.

Model Fit and Comparison

- What is the goodness-of-fit measure in linear models?
  - residual sum of squares (RSS)
  - R^2 and adjusted R^2
- The notion of deviances: $\approx$ - 2log likelihood, smaller $\rightarrow$ better fit.
- Software output usually includes:
  - the null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean).
  - the residual deviance can be compared to the null deviance:

Model Fit and Comparison

- What is the goodness-of-fit measure in linear models?
  - residual sum of squares (RSS)
  - R^2 and adjusted R^2
- The notion of deviances: $\approx$ - 2log likelihood, smaller $\rightarrow$ better fit.
- Software output usually includes:
  - the null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean).
  - the residual deviance can be compared to the null deviance:
- Akaike Information Criterion (AIC):
  - based on deviance, but penalize on model size
  - more useful in model comparison

Model Fit and Comparison

- Can compare AIC and choose the one with smaller AIC

Model Fit and Comparison

- Can compare AIC and choose the one with smaller AIC
- Can perform hypothesis test
  - Hosmer-Lemeshow Goodness of Fit for binary response
  - in R package "ResourceSelection"
  - significant results means poor model fit

Model Fit and Comparison

- Can compare AIC and choose the one with smaller AIC
- Can perform hypothesis test
  - Hosmer-Lemeshow Goodness of Fit for binary response
  - in R package "ResourceSelection"
  - significant results means poor model fit
- In general, GLM are estimated by MLE
  - Likelihood ratio test (χ^2 test)
  - significant results means larger model provide significant improvement

Outline