BTRY 6020: Module 6
Generalized Linear Models

Spring 2025

Logistics

- Continue Module 6 on Generalized Linear Models
- Assessment deadline extended to Friday 18 April
- Lab is a little ahead, so
- Final project instructions posted on Friday
- R Markdown file, similar to module assessments

Recap

Generalized Linear Model

Logistic regression is a specific example of a **Generalized Linear Model** (GLM)

In general GLM's have 3 pieces

- **Distribution (family)**: What is the distribution of the dependent variable Y?
- **Link function**: The conditional mean of

$$E(Y_i \mid \mathbf{X_i}) = g^{-1}(a_i)$$

  is a function of some input; equivalent to saying:

$$g(E(Y_i \mid \mathbf{X_i})) = a_i$$

- **Linear model of covariates**: The "input" a_i is a linear function of some covariates so that

$$a_i = b_0 + \sum_k b_k x_{ik}$$

Generalized Linear Model

Putting everything together, we have

$$g(E(Y_i \mid \mathbf{X_i})) = b_0 + \sum_k b_k x_{ik}$$

- Bernoulli or Binomial Data: $E(Y_i \mid \mathbf{X_i}) = \theta = P(Success)$,

$$\log\left(\frac{\theta}{1-\theta}\right) = b_0 + \sum_k b_k x_{ik}$$

- Poisson Data: $E(Y_i \mid \mathbf{X_i}) = \theta$,

$$\log(\theta) = b_0 + \sum_k b_k x_{ik}$$

Discussion

What are examples from your field where a glm might be useful?

- Binary outcome?
- Count valued outcome?

Maximum likelihood estimation

Selecting model parameters

- In linear regression, we selected $\hat{\mathbf{b}}$ to minimize the RSS

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - (\hat{b}_0 + \sum_k \hat{b}_k x_{i,k}) \right)^2$$

- In GLM's we don't minimize the sum of squares, but we instead maximize a likelihood function
- Likelihood function can be for various tasks used similar to how we use RSS

Likelihood functions

Given a distribution, the density/mass function tells us the probability of a particular outcome

Suppose we observe, a Binomial random variable, Y, with m trials and the probability of each trial is θ

- Probability of outcome y given some parameter theta

$$P(Y = y) = \frac{m!}{y!(m-y)!}\theta^y(1-\theta)^{m-y}$$

- If $m = 5$ and $\theta = .7$ and $y = 2$ then,

$$P(Y = 2) = \frac{5!}{2!(5-2)!}.7^2(1-.7)^{5-2} = .1323$$

- If $m = 5$ and $\theta = .7$ and $y = 5$ then,

$$P(Y = 5) = \frac{5!}{5!(5-5)!}.7^5(1-.7)^{5-5} = .1681$$

Likelihood functions

- Density/Mass functions are different for different distributions
- Given the distribution parameters, they tell us the probability that a specified outcome will occur
- We can also go the other direction: given a model and observed outcomes, what is the value of the parameters which make the outcome most likely
- **Likelihood function**: fix the data, and consider the probability a function of the parameter(s)

Likelihood functions

Given a distribution, the likelihood tells us the probability of a particular outcome

- Given an outcome y, for some θ the likelihood is

$$l(\theta; y) = \frac{m!}{y!(m-y)!}\theta^y(1-\theta)^{m-y}$$

- If $m = 5$ and $y = 2$, if $\theta = .7$ then,

$$l(\theta = .7; y = 2) = \frac{5!}{2!(5-2)!}.7^2(1-.7)^{5-2} = .1323$$

- If $m = 5$ and $y = 2$, if $\theta = .4$ then,

$$l(\theta = .4; y = 2) = \frac{5!}{2!(5-2)!}.5^2(1-.5)^{5-2} = .3456$$

- If $m = 5$ and $y = 2$, if $\theta = .2$ then,

$$l(\theta = .2; y = 2) = \frac{5!}{2!(5-2)!}.2^2(1-.2)^{5-2} = .2048$$

Likelihood functions

When the parameters of a distribution are unknown, then one way to estimate them is to select the parameter under which the outcome has the highest likelihood. This is called the **maximum likelihood estimate** (MLE).

- Given an outcome y, for some θ the likelihood is $l(\theta = .7; y) = \frac{m!}{y!(m-y)!}\theta^y(1-\theta)^{m-y}$
- If $m = 5$ and $y = 2$, if $\theta = .7$ then,

$$l(\theta = .7; y) = \frac{5!}{2!(5-2)!}.7^2(1-.7)^{5-2} = .1323$$

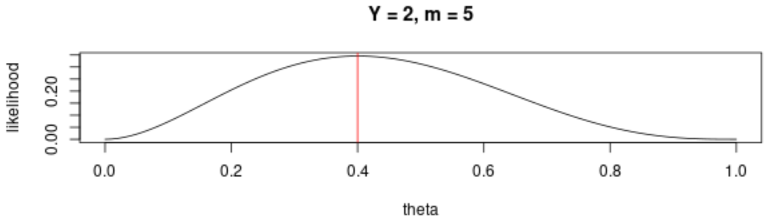
- If $m = 5$ and $y = 2$, if $\theta = .4$ then,

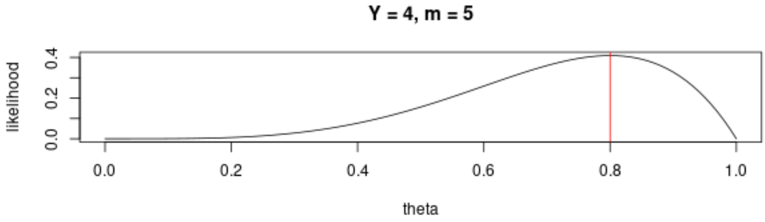
$$l(\theta = .4; y) = \frac{5!}{2!(5-2)!}.5^2(1-.5)^{5-2} = .3456$$

- If $m = 5$ and $y = 2$, if $\theta = .2$ then,

$$l(\theta = .2; y) = \frac{5!}{2!(5-2)!}.2^2(1-.2)^{5-2} = .2048$$

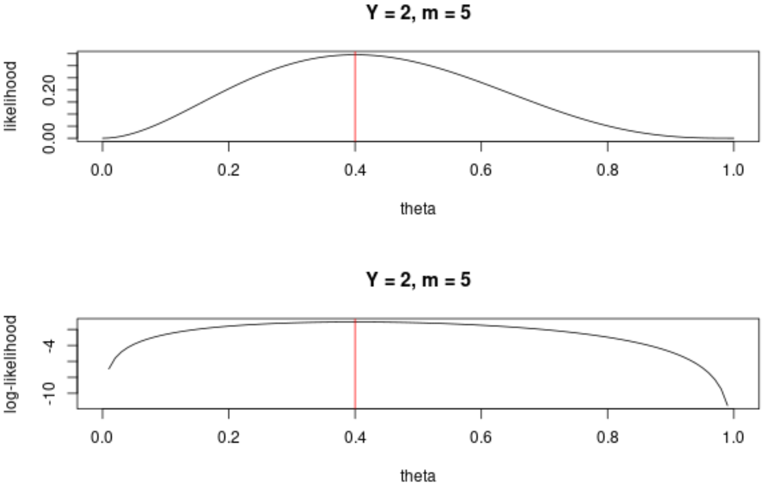
Likelihood functions





Log-Likelihood functions

Choosing the parameter which maximizes the log of the likelihood is equivalent to maximizing the likelihood, and often easier mathematically



Log-Likelihood functions

Choosing the parameter which maximizes the log of the likelihood is equivalent to maximizing the likelihood, and often easier mathematically

For binomial data, when $\theta = E(Y_i \mid \mathbf{X_i})$, the log-likelihood is

$$\ell(\theta; \mathbf{y}) = \sum_{i}^{n} \left[\log\left(\frac{m!}{y_i!(m-y_i)!} \right) + y_i \log(\theta(\mathbf{X_i})) + (m - y_i) \log(1 - \theta(\mathbf{X_i})) \right]$$

For Poisson data, when $\theta = E(Y_i \mid \mathbf{X_i})$, the log-likelihood is

$$\ell(\theta; \mathbf{y}) = -n\theta + \sum_{i=1}^{n} [y_i \log(\theta(\mathbf{X_i})) - \log(y_i!)]$$

Log-Likelihood functions

Choosing the parameter which maximizes the log of the likelihood is equivalent to maximizing the likelihood, and often easier mathematically

For Gaussian data, when $\theta = E(Y_i \mid \mathbf{X_i})$ and $\mathrm{var}(Y_i) = \sigma^2$, the log-likelihood is

$$\ell(\mathbf{y}; \theta) = -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \theta(\mathbf{X_i}))^2}{\sigma^2} - \log(\sigma\sqrt{2\pi})$$

When we re-write only in terms of $\theta(\mathbf{X_i})$, we get

$$\begin{aligned}\ell(\mathbf{y}; \theta) &= -\frac{n}{2} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{1}{n}\sum_i (y_i - \theta(\mathbf{X_i}))^2\right) \\ &\propto -\frac{n}{2}\log\left(\frac{1}{n}\sum_i (y_i - \theta(\mathbf{X_i}))^2\right)\end{aligned}$$

Log-Likelihood functions

Linear regression is a GLM assuming a Gaussian distribution and link function is just $g(a_i) = a_i$ (i.e. no transformation)

- Maximizing the log-likelihood

$$-\frac{n}{2}\log\left(\frac{1}{n}\sum_i (y_i - \theta(\mathbf{X_i}))^2\right)$$

  is equivalent to minimizing

$$\sum_i (y_i - \theta(\mathbf{X_i}))^2$$

- When we use the link function $g(a_i) = a_i$

$$g(E(Y_i \mid \mathbf{X_i})) = E(Y_i \mid \mathbf{X_i}) = b_0 + \sum_k b_k x_{i,k}$$

  so the procedure selects $\hat{\mathbf{b}}$ by minimizing

$$\sum_i (y_i - \theta(\mathbf{X_i}))^2 = \underbrace{\sum_i (y_i - (\hat{b}_0 + \sum_k \hat{b}_k x_{i,k}))^2}_{\text{RSS}}$$

Likelihood functions

Picking the parameters which have the maximum likelihood, often coincides with common sense procedures

- For Binomial, Poisson, and Gaussian data the maximum likelihood estimate (not for GLMs) is just the sample mean
- For GLMs, the procedure is a bit more complicated, but conceptually similar
- The likelihood is a function of the conditional mean, $\theta(\mathbf{X_i})$, which a function of $\hat{\mathbf{b}}$, so pick $\hat{\mathbf{b}}$ which maximizes the likelihood

GLMs and likelihood functions

For GLMs can use the likelihood function in very similar way in which we used the RSS for linear regression

- Fitting a GLM when assuming Gaussian data with $g(a_i) = a_i$ is equivalent to linear regression (see slides at end for more details)
- A way to measure how well the model fits our data
- A way test whether covariates are "statistically significant"
- A way to select which covariates model to include

Hypothesis Testing: multiple coefficients

Suppose we want to test $H_0 = b_1 = b_2 = \ldots = 0$
For linear regression, we used the F-test where:

$$F = \frac{[RSS(Null) - RSS(Alt)]/(p_{\text{alt}} - p_{\text{null}})}{RSS(Alt)/(n - p_{\text{alt}} - 1)}$$

is compared to an F distribution to compute p-values.

For GLM's, we use

$$\chi = 2(l_1 - l_0)$$

where l_1 is the log-likelihood of the alternative model and l_0 is the log-likelihood of the null model. To compute a p-value, the test statistic χ is compared to a χ^2 distribution.

Hypothesis Testing: single coefficient

Suppose we want to test $H_0 : b_k = \beta$

- For linear regression, we used a t-test
- Can do a similar test for GLM's using the `summary` function to form

$$z = \frac{\hat{b}_k - \beta}{\widehat{se(\hat{b})}}$$

  and compare to a $N(0, 1)$ (instead of T)
- For linear regression, using an F-test with 1 variable always gave exactly the same result
- For GLMs, using a χ^2 test with 1 variable will be similar as the z-test, but not quite the same when n is small
- Generally, using χ^2 test will be better when n is small

Model Selection

For linear regression:

$$\text{AIC} = -\frac{n}{2}\log(RSS/n) - (\text{number of parameters})$$

$$\text{BIC} = -\frac{n}{2}\log(RSS/n) - \frac{\log(n)}{2}(\text{number of parameters})$$

For GLM's,

$$\text{AIC} = -\ell(\hat{\theta}; y) - (\text{number of parameters})$$

$$\text{BIC} = -\ell(\hat{\theta}; y) - \frac{\log(n)}{2}(\text{number of parameters})$$

Summary

- GLMs are a very flexible class of models which have three parts
  - **Distribution (family)**: distribution of the dependent variable (conditional on covariates)
  - **Link function**: some specified function which takes an input and maps to the conditional mean of the dependent variable
  - **Linear model of covariates**: The value $a_i = b_0 + \sum_k b_k x_{ik}$
- GLMs are fit using a maximum likelihood principle
- Likelihood gives a way to measure "goodness of fit" for a particular model
- Using the likelihood allows us to do hypothesis testing