Module 10: Generalized linear regression

Siyue Yang

06/22/2022

Outline

In this module, we will review generalized linear regression.

Logistic regression

- Each response is binary: $y_i = 1, 0$
- Explanatory variables x_i^T as usual
- Model

$$\operatorname{pr}\left(y_i = 1 \mid x_i\right) = p_i(\beta) = \frac{\exp\left(x_i^{\top} \beta\right)}{1 + \exp\left(x_i^{\top} \beta\right)}$$

Compared to linear regression

- Logistic regression
  - Regression

$$\mathbb{E}(y_i) = p_i = \frac{\exp\left(x_i^{\mathrm{T}}\beta\right)}{1 + \exp\left(x_i^{\mathrm{T}}\beta\right)}$$

  - Probability distribution

$$y_i \sim \text{Bernoulli}(p_i)$$

Compared to linear regression

- Logistic regression
  - Regression

$$\mathbb{E}(y_i) = p_i = \frac{\exp\left(x_i^{\mathrm{T}}\beta\right)}{1 + \exp\left(x_i^{\mathrm{T}}\beta\right)}$$

  - Probability distribution

$$y_i \sim \text{Bernoulli}(p_i)$$

- Linear regression
  - Regression

$$\mathbb{E}(y_i) = \mu_i = x_i^{\mathrm{T}}\beta$$

  - Probability distribution

$$y_i \sim \text{Normal}\left(\mu_i, \sigma^2\right)$$

Generalized linear models (GLMs)

- Generalized Linear Models extend the classical set-up to allow for a wider range of distributions
- GLMs have three pieces
  1. random component: $y_i \sim$ some distribution with $E[y_i|\mathbf{x}_i] = \mu_i$
  2. systematic component: $\mathbf{x}_i^T \beta$
  3. The link function that links the random and systematic components $g(u_i) = \mathbf{x}_i^T \beta$
- Distributions of y_i comes from exponential family.

Exponential family

The random variable Y belongs to the exponential family of distributions if its support does not depend upon any unknown parameters and its density or probability mass function takes the form

$$p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

GLMs: theory

- $p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$

GLMs: theory

- $p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$
- $\mathrm{E}\left(y_i \mid x_i\right) = b'\left(\theta_i\right) = \mu_i$ defines μ_i as a function of θ_i

GLMs: theory

- $p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$
- $\mathrm{E}\left(y_i \mid x_i\right) = b'\left(\theta_i\right) = \mu_i$ defines μ_i as a function of θ_i
- $g\left(\mu_i\right) = x_i^\top \beta = \eta_i$ links the n observations together via covariates

GLMs: theory

- $p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$
- $\mathrm{E}\left(y_i \mid x_i\right) = b'\left(\theta_i\right) = \mu_i$ defines μ_i as a function of θ_i
- $g\left(\mu_i\right) = x_i^\top \beta = \eta_i$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_i is the linear predictor

GLMs: theory

- $p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$
- $\mathrm{E}(y_i \mid x_i) = b'(\theta_i) = \mu_i$ defines μ_i as a function of θ_i
- $g(\mu_i) = x_i^\top \beta = \eta_i$ links the n observations together via covariates
- $g(\cdot)$ is the link function; η_i is the linear predictor
- $\mathrm{Var}(y_i \mid x_i) = \phi_i b''(\theta_i) = \phi_i V(\mu_i)$, $V(\cdot)$ is the variance function

GLMs in R

"glm" has several options for family:

$$\begin{aligned}&\text{binomial (link = "logit")}\\&\text{gaussian(link = "identity")}\\&\text{Gamma(link = "inverse")}\\&\text{inverse.gaussian(link} = \text{"}1/\mathrm{mu}^\wedge 2\text{")}\\&\text{poisson(link = "log")}\\&\text{quasi (link = "identity", variance = "constant")}\\&\text{quasibinomial (link = "logit")}\\&\text{quasipoisson(link = "log")}\end{aligned}$$

- Each of these is a member of the class of generalized linear models
- Generalized: distribution of response is not assumed to be normal
- Linear: some transformation of $E(y_i)$ is of the form $x_i^\top \beta$

Poisson regression

- the Poisson distribution is a useful starting point for data that counts events

$$f(y_i \mid x_i) = \frac{1}{y!}\mu_i^{y_i} e^{-\mu_i}, y_i = 0, 1, \ldots$$

$$f(y_i \mid x_i) = \exp\{y_i \log \mu_i - \mu_i - \log(y_i!)\}$$

- canonical parameter

$$\theta_i = \log(\mu_i)$$

- linear model:

$$\log(\mu_i) = x_i^\top \beta$$

- equivalently

$$E(y_i) = \mu_i = \exp\left(x_i^\top \beta\right)$$

Likelihood-based estimation and inference

- Maximum likelihood estimation, similar to linear regression but has to be estimated iteratively (using Newton Raphson / Method of Scoring)
- Inference based on the limiting distribution for MLE

$$\hat{\beta} \sim N\left(\beta, I(\hat{\beta})^{-1}\right)$$

  Standard errors are the square roots of the inverse of the information matrix.

Exercise

More math derivation exercises of inference of GLMs are in this week's exercises.

Thanks for spending 3 weeks with us!