Regression Analysis Tutoring9

Seung Bong Jung

Seoul National University

November 1, 2021

Maximum Likelihood Estimation

Suppose that we observe independent Z_i, $1 \le i \le n$, and assume that Z_i is generated from a distribution with pdf $f_i(\cdot, \boldsymbol{\theta})$ in a model $\{f_i(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.

- Likelihood of $\boldsymbol{\theta}$: We call $\prod_{i=1}^{n} f_i(z_i, \boldsymbol{\theta})$, as a function of $\boldsymbol{\theta}$, the likelihood of $\boldsymbol{\theta}$ given the observations $(z_1, \ldots, z_n)$. We call its logarithm, $\sum_{i=1}^{n} \log f_i(z_i, \boldsymbol{\theta})$, the log-likelihood of $\boldsymbol{\theta}$.
- Maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}}(z_1, \ldots, z_n) := \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \log f_i(z_i, \boldsymbol{\theta}).$$

- Computational difficulty: Typically, the maximization of a likelihood function is a nonlinear optimization problem which does not have an explicit solution.

Generalized Linear Models

Two assumptions of a generalized linear model are:

- The density function of Y given a set of predictors $x_1, \ldots, x_p$ belongs to an exponential family given by

$$\begin{aligned} \mathsf{pdf}_{Y|x_1,\ldots,x_p}(y) = \exp[&a(\phi)^{-1}(y\theta(x_1,\ldots,x_p) - b(\theta(x_1,\ldots,x_p))) \\ &+ c(y,\phi)], \end{aligned}$$

  where the functions a, b and c are fully specified, ϕ is termed as the dispersion parameter and θ is called the canonical parameter function;

- For a function g called link

$$g(\mathsf{E}(Y|x_1,\ldots,x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Dichotomous Response

Suppose that the response variable is dichotomous taking only the values 0 and 1. In this case, the mean function is $\mathsf{P}(Y = 1|x_1, \dots, x_p)$.

- Estimating the mean function based on the least squares estimator that minimizes

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

  is not suitable since the estimator does not have a correct range.

- A usual practice is to consider a link function, say g, such that (i) it is strictly increasing and (ii) its inverse maps $\mathbb{R}$ to the range $[0, 1]$, and then to assume that the mean function obeys the model

$$g(\mathsf{E}(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Logistic Regression Model

- Logistic regression model: Take the link

$$g(u) = \log(\frac{u}{1-u})$$

  i.e, assume

$$\mathsf{E}(Y|x_1,\ldots,x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1+\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}.$$

  The function g given above is called the logit link. It is the inverse of the logistic function.

- There are at least two motivations for modeling $\mathsf{E}(Y|x_1,\ldots,x_p)$ in this way.

Motivation I: Binary Choice model

- Binary choice model in economics: The observed response Y takes value 1 when a latent (unobserved) response $Y* = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p - \epsilon$ is greater than 0.
- In case ϵ has a logistic distribution with distribution function

$$\mathsf{P}(\epsilon \leq u) = e^u/(1 + e^u).$$

$$\begin{aligned}
\mathsf{E}(Y|x_1, \ldots, x_p) &= \mathsf{P}(Y^* \geq 0) \\
&= \mathsf{P}(\epsilon \leq \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \\
&= \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}
\end{aligned}$$

Motivation II: Convexity of Likelihood

- Think of estimating $\boldsymbol{\beta}$ by the maximum likelihood method. Given the observations $(\mathbf{x}_1, \ldots \mathbf{x}_p, \mathbf{Y})$, the log-likelihood of $\boldsymbol{\beta}$ equals

$$L(\boldsymbol{\beta}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{Y}) = \sum_{i=1}^{n} Y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i),$$

  where $p_i = \mathsf{E}(Y_i|x_{i1}, \ldots, x_{ip}) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$.

- Likelihood under the logistic model: Taking the logit link gives

$$L(\boldsymbol{\beta}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{Y}) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

$$- \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}),$$

  which is a strictly concave function of $\boldsymbol{\beta}$.

Fitting Logistic Regression Model

- Find $\hat{\beta}_j$,$0 \leq j \leq p$, that maximize $L(\boldsymbol{\beta})$ by some algorithm.
- Estimate $\mu_{x_1,\ldots,x_p} \equiv \mathsf{E}(Y|x_1,\ldots,x_p)$ by

$$\hat{\mu}_{x_1,\ldots,x_p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}$$

- The strict concavity of the likelihood function under the logistic model makes the maximization algorithm numerically stable.

Probit Model

- Basically, one may use other link functions, which lead to other regression models for the dichotomous response.
- Probit regression model: Take the link $g = \Phi^{-1}$ where $\Phi$is the CDF of $\mathsf{N}(0,1)$, i.e., assume

$$\mathsf{E}(Y|x_1,\ldots,x_p) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p).$$

  The function $g = \Phi^{-1}$ is called probit link.
- This can be also motivated by the binary choice model, now with $\epsilon \sim \mathsf{N}(0,1)$.
- Likelihood under the probit model:

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{x}_1,\ldots,\mathbf{x}_p,\mathbf{Y}) = &\sum_{i=1}^{n} Y_i \log\Big(\frac{\Phi(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 - \Phi(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}\Big) \\ &+ \sum_{i=1}^{n} \log(1 - \Phi(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})). \end{aligned}$$

Poisson Regression

- Suppose that the response Y represents the count of an event. Assume that the expected count depends on the values of predictors $x_1, \ldots, x_p$ and the count follows a Poisson distribution for each given set of predictor values.
- In the GLM framework, $\theta(x_1, \ldots, x_p) = \log \mathsf{E}(Y|x_1, \ldots, x_p)$, $b(u) = e^u$, $a(\phi) = 1$ and $c(y, \phi) = -\log(y!)$.
- Log-linear model: Taking the link $g(u) = \log(u)$, i.e., assuming $\log(\mathsf{E}(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ gives the log-likelihood

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{Y}) = &\sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \\ &- \sum_{i=1}^{n} e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}} - \sum_{i=1}^{n} \log(Y_i!). \end{aligned}$$

Newton-Raphson Algorithm

Suppose that we want to find $\hat{\boldsymbol{\beta}}$ such that $\mathbf{F}(\hat{\theta}) = \mathbf{0}$ in $\boldsymbol{\Theta}$ for some smooth nonlinear function $\mathbf{F} = (F_1, \ldots, F_k)^t$ that maps $\mathbb{R}^k$ to $\mathbb{R}^k$.

- Newton-Raphson algorithm:

$$\hat{\boldsymbol{\theta}}_{\text{new}} = \hat{\boldsymbol{\theta}}_{\text{old}} - \mathbf{F}'(\hat{\boldsymbol{\theta}}_{\text{old}})^{-1}\mathbf{F}(\hat{\boldsymbol{\theta}}),$$

  where $\mathbf{F}'(\mathbf{u})$ is $k \times k$ matrix whose (j,k)th entry equals $\partial F_j(\mathbf{u})/\partial u_k$.
- The algorithm is motivated from the first-order approximation

$$\mathbf{0} = \mathbf{F}(\boldsymbol{\theta}) \approx \mathbf{F}(\boldsymbol{\theta}_0) + \mathbf{F}'(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

  where $\boldsymbol{\theta} \approx \boldsymbol{\theta}_0$.

Iteratively Reweighted Least Squares

The Iteratively Reweighted Least Squares (IRLS) is a modified version of the Newton-Raphson algorithm for maximum likelihood estimation.

- The MLE is often given as the solution of the likelihood equation

$$\mathbf{F}(\boldsymbol{\theta}) := \frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbf{0}, \text{ where } L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(\mathbf{z}_i, \boldsymbol{\theta}).$$

- Write $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^t$, where $\mu_i \equiv \mu_i(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}(Z_i)$. Note that

$$\mathbf{F}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\frac{\partial L}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \boldsymbol{\theta}}\right) = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}^t}\right)^t \left(\frac{\partial L}{\partial \boldsymbol{\mu}}\right),$$

$$\mathbf{F}(\boldsymbol{\theta})' = \sum_{i,j=1}^{n} \left(\frac{\partial^2 L}{\partial \mu_i \mu_j}\right)\left(\frac{\partial \mu_i}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \mu_j}{\partial \boldsymbol{\theta}}^t\right) + \sum_{i=1}^{n} \left(\frac{\partial L}{\partial \mu_i}\right)\left(\frac{\partial^2 \mu_i}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^t}\right)$$

$$= \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}^t}\right)^t \left(\frac{\partial^2 L}{\partial \boldsymbol{\mu}\boldsymbol{\mu}^t}\right)\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}^t}\right) + \sum_{i=1}^{n} \left(\frac{\partial L}{\partial \mu_i}\right)\left(\frac{\partial^2 \mu_i}{\partial \boldsymbol{\theta}\boldsymbol{\theta}^t}\right)$$

Iterative Reweighted Least Squares

- Define

$$\mathbf{X}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}^t}, \mathbf{W}(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}}(\frac{\partial^2 L}{\partial \boldsymbol{\mu}\boldsymbol{\mu}^t}), \mathbf{Y}(\boldsymbol{\theta}) = \frac{\partial L}{\partial \boldsymbol{\mu}}.$$

Note that $\mathsf{E}_{\boldsymbol{\theta}}(\partial L/\partial \mu_i) = 0$ under some regularity conditions. We get

$$-\mathbf{F}'(\boldsymbol{\theta}) \approx \mathbf{X}(\boldsymbol{\theta})^t \mathbf{W}(\boldsymbol{\theta})\mathbf{X}(\boldsymbol{\theta}).$$

- Stuffing these ingredients into the Newton-Raphson algorithm gives the normal equation of a weighted least squares regression,

$$\mathbf{X}_{\text{old}}^t \mathbf{W}_{\text{old}} \mathbf{X}_{\text{old}} \hat{\boldsymbol{\theta}}_{\text{new}} = \mathbf{X}_{\text{old}}^t \mathbf{W}_{\text{old}} (\mathbf{W}_{\text{old}}^{-1} \mathbf{Y}_{\text{old}} + \mathbf{X}_{\text{old}} \hat{\boldsymbol{\theta}}_{\text{old}}),$$

where $\mathbf{X}_{\text{old}} = \mathbf{X}(\hat{\boldsymbol{\theta}}_{\text{old}})$, $\mathbf{W}_{\text{old}} = \mathbf{W}(\hat{\boldsymbol{\theta}}_{\text{old}})$ and $\mathbf{Y}_{\text{old}} = \mathbf{Y}(\hat{\boldsymbol{\theta}}_{\text{old}})$.

IRLS for Logistic Regression

- In this case, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$

$$\mu_i(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}.$$

- It follows that $\mathbf{X}(\boldsymbol{\beta})$ is a $n \times (p+1)$ matrix given by

$$\mathbf{X}(\boldsymbol{\beta}) = \mathbf{V}(\boldsymbol{\beta})\mathbf{X},$$

where $\mathbf{V} = \text{diag}(p_i(1-p_i))$ and $\mathbf{X}$ is the original design matrix. Also,

$$\mathbf{W}(\boldsymbol{\beta}) = \mathbf{V}(\boldsymbol{\beta})^{-1}, \mathbf{Y}(\boldsymbol{\beta}) = \left(\frac{Y_i - p_i(\boldsymbol{\beta})}{p_i(\boldsymbol{\beta})(1 - p_i(\boldsymbol{\beta}))}\right).$$

- Thus, we get the updating equation

$$\mathbf{X}^t \mathbf{V}_{\text{old}} \mathbf{X} \hat{\boldsymbol{\theta}}_{\text{new}} = \mathbf{X}^t \mathbf{V}_{\text{old}} (\mathbf{Y}_{\text{old}} + \mathbf{X} \hat{\boldsymbol{\theta}}_{\text{old}}).$$

IRLS for Poisson Regression

- In this case, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^t$

$$\mu_i(\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}).$$

- It follows that $\mathbf{X}(\boldsymbol{\beta})$ is a $n \times (p+1)$ matrix given by

$$\mathbf{X}(\boldsymbol{\beta}) = \mathbf{V}(\boldsymbol{\beta})\mathbf{X},$$

where $\mathbf{V} = \text{diag}(\mu_i)$ and $\mathbf{X}$ is the original design matrix. Also,

$$\mathbf{W}(\boldsymbol{\beta}) = \mathbf{V}(\boldsymbol{\beta})^{-1}, \mathbf{Y}(\boldsymbol{\beta}) = (\frac{Y_i - \mu_i(\boldsymbol{\beta})}{\mu_i(\boldsymbol{\beta})}).$$

- Thus, we get the updating equation

$$\mathbf{X}^t \mathbf{V}_{\text{old}} \mathbf{X} \hat{\boldsymbol{\theta}}_{\text{new}} = \mathbf{X}^t \mathbf{V}_{\text{old}} (\mathbf{Y}_{\text{old}} + \mathbf{X} \hat{\boldsymbol{\theta}}_{\text{old}}).$$

Least Squares Estimation of Nonlinear Regression Models

- Assume that

$$Y = f(x_1, \ldots, x_p, \boldsymbol{\theta}) + \epsilon, \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k,$$

  for some known nonlinear function f of $\boldsymbol{\theta}$ and that $\mathsf{E}(\epsilon) = 0$. Given a set of independent observations $(x_{i1}, \ldots, x_{ip}, Y_i)$ we maximize

$$L(\boldsymbol{\theta}) \equiv -\frac{1}{2}\sum_{i=1}^{n}(Y_i - f(x_{i1}, \ldots, x_{ip}, \boldsymbol{\theta}))^2 \text{ over } \boldsymbol{\Theta}.$$

- Putting this into the framework of IRLS, for example, we get

$$\mathbf{X}(\boldsymbol{\theta}) = (\frac{\partial f(x_{i1}, \ldots, x_{ip}, \boldsymbol{\theta})}{\partial \theta_j}), \mathbf{W}(\boldsymbol{\theta}) \equiv I,$$

$$\mathbf{Y}(\boldsymbol{\theta}) = (Y_i - f(x_{i1}, \ldots, x_{ip}, \boldsymbol{\theta})).$$

Least Squares Estimation of Nonlinear Regression Models

- This means that each updating step for obtaining $\hat{\boldsymbol{\theta}}_{\text{new}}$ is simply to do an ordinary least squares regression with the pseudo responses

$$Y_i - f(x_{i1}, \ldots, x_{ip}, \hat{\boldsymbol{\theta}}_{\text{old}}) + \sum_{j=1}^{p} \hat{\theta}_{\text{j,old}} \frac{\partial f(x_{i1}, \ldots, x_{ip}, \hat{\boldsymbol{\theta}}_{\text{old}})}{\partial \hat{\theta}_{\text{j,old}}}$$

and the pseudo predictors

$$\frac{\partial f(x_{i1}, \ldots, x_{ip}, \hat{\boldsymbol{\theta}}_{\text{old}})}{\partial \hat{\theta}_{\text{1,old}}}, \ldots, \frac{\partial f(x_{i1}, \ldots, x_{ip}, \hat{\boldsymbol{\theta}}_{\text{old}})}{\partial \hat{\theta}_{\text{p,old}}}$$