Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Model Description and Estimation

Georgia Tech

About This Lesson

Georgia Tech

Logistic Regression Model

Data: $\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Model: We model the *probability of success given the predictor(s)*

$$\mathrm{p} = \mathrm{p}(X_1, \cdots, X_p) = \Pr(Y = 1 \mid X_1, \cdots, X_p)$$

by linking p to the predicting variables through the ***logit*** *link function* g:

$$\mathrm{g}(\mathrm{p}) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

OR

$$\boxed{\mathrm{p}(X_1, \cdots, X_p) = \frac{\mathrm{e}^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + \mathrm{e}^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}}$$

Georgia Tech

Model Interpretation

- The probability of success given one predicting variable $X = x$ is

  $$\mathrm{p} = \mathrm{p}(x) = \Pr(Y = 1 \mid x)$$

- The logit function $\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 x$ is the ***log odds*** function.

- The exponential of the logit function $\frac{\mathrm{p}(X)}{1-\mathrm{p}(X)} = \mathrm{e}^{\beta_0 + \beta_1 x}$ is the ***odds*** of $Y = 1$ at $X = x$

- The odds at $X = a$ versus $X = b$ is equal to the ***odds ratio***:

  $$\frac{\mathrm{e}^{\beta_0 + \beta_1 a}}{\mathrm{e}^{\beta_0 + \beta_1 b}} = \mathrm{e}^{\beta_1 (a-b)}$$

Georgia Tech

Model Interpretation

If we calculate the odds ratio of the odds at $X = b + 1$ versus $X = b$, we have

$$\frac{e^{\beta_0+\beta_1(b+1)}}{e^{\beta_0+\beta_1 b}} = e^{\beta_1}$$

- The regression coefficient β_1 can be interpreted as the log of the odds ratio for an increase of one unit in the predicting variable.
- If X a dummy variable of a categorical factor, interpret as the log of odds ratio of one category versus baseline.
- Interpret β with respect to the odds of success, not directly with respect to the response variable.

Model Estimation

Model the probability of success given predictor(s):

$$\text{Logit}\left(\Pr(Y = 1 \mid X_1, \cdots, X_p)\right) = \text{Logit}\left(\text{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Parameters: $\beta_0, \beta_1, \cdots, \beta_p$

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \mathcal{L}(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i=1}^{n} \text{p}(\boldsymbol{X}_{i,1}, \boldsymbol{X}_{i,2}, \cdots, \boldsymbol{X}_{i,p})^{Y_i} \left(1 - \text{p}(\boldsymbol{X}_{i,1}, \boldsymbol{X}_{i,2}, \cdots, \boldsymbol{X}_{i,p})\right)^{1-Y_i}$$

or

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \ell(\beta_0, \beta_1, \cdots, \beta_p) = \max_{\beta_0,\beta_1,\cdots,\beta_p} \log(\mathcal{L}(\beta_0, \beta_1, \cdots, \beta_p))$$

$$= \max_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} \left(\boldsymbol{Y}_i \log\left(\frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}\right) + (1 - \boldsymbol{Y}_i) \log\left(\frac{1}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}\right)\right)$$

Model Estimation (cont'd)

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} \left(\boldsymbol{Y}_i \log\left(\frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \right) + (1-\boldsymbol{Y}_i)\log\left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \right) \right)$$

- Maximizing the (log-)likelihood function with respect to $\beta_0,\beta_1,\cdots,\beta_p$ in closed form expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters.
- Use numerical algorithm to estimate $\beta_0,\beta_1,\cdots,\beta_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$

Upshot: The estimated parameters and their standard errors are approximate estimates. Do not attempt to do it yourself! Use statistical software to derive the estimated regression coefficients.

Georgia Tech

Summary

Georgia Tech