Research Methods in Political Science I

Logistic Regression (2)

Yuki Yanai

School of Law and Graduate School of Law

December 16, 2015

KOBE UNIVERSITY

Today's Menu

1. Logistic Regression by ML
   - Example Logistic Regression
   - Computation
2. Evaluating Logistic Regression Results
   - Evaluating the Fit

Question

Example 1: Explain the win-lose in SMDs by the previous wins

Does the previous wins affect the victory in SMDs? If it does, how much? (fake data)

- Response y the number of winning candidates by previous wins
- Predictor t (terms): non-negative integer

We'd like to fit a logistic curve to summarize the data

Checking Variables

Previous wins (t_i)	Candidates (n_i)	Winning Candidates (y_i)
0	3	1
1	2	1
2	1	0
3	2	1
4	3	2
5	3	2
6	0	0
7	1	1
Total	15	8

Logistic Regression

- We model this problem with logistic regression:

$$
\begin{aligned}
p_i &= \Pr(y_i|n_i,\theta_i) = \binom{n_i}{y_i}\theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \\
\theta_i &= \frac{\exp(\beta_1+\beta_2 t_i)}{1+\exp(\beta_1+\beta_2 t_i)} \\
Y_i &\sim \text{Bin}(n_i,\theta_i)
\end{aligned}
$$

- θ_i the success probability for a Bernoulli trial
- Y_i are independent
- parameters: β_1 and β_2

Specifying Likelihood Function

- Let $\binom{n_i}{y_i} = a_i$
- Likelihood function for the i-th observation:

$$
\begin{aligned}
L_i(\beta) &= p_i = a_i \theta_i^{t_i} (1-\theta_i)^{n_i - t_i} \\
&= a_i \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{n_i - y_i}
\end{aligned}
$$

- Since y_i are independent of each other, the likelihood function for the data is

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} L_i(\beta) \\
&= \prod_{i=1}^{n} a_i \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{n_i - y_i}
\end{aligned}
$$

Specifying Log-Likelihood Function

- Ignoring the constant term, the log-likelihood is

$$\begin{aligned}
\log L(\beta) &= \log \prod_{i=1}^{n} L_i(\beta) \\
&= \sum_{i=1}^{n} \log \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{n_i - y_i} \\
&= \sum_{i=1}^{n} \log \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i}
\end{aligned}$$

- perform further calculation with R

Question

Example 2: Explain the win-lose in SMDs by the electoral expenditure

Does the amount of electoral spending (million yen) affect the victory in SMDs? How much? (fake data)

- Response r: win = 1, lose = 0
- Predictor x (expenditure): non-negative real number (million yen)

We'd like to fit a logistic curve

Logistic Regression

- Model this problem with logistic regression

$$\begin{aligned} \theta_i &= \Pr(r_i = 1) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \\ r_i &\sim \text{Bern}(\theta_i) \end{aligned}$$

- θ_i: success probability of a Bernoulli trial
- r_i, $(i = 1, 2, \ldots, n)$ are mutually independent
- parameters: β_1 and β_2

Specifying Log-Likelihood Function

- Likelihood function for the i-th observation

$$
\begin{aligned}
L_i(\boldsymbol{\beta}) &= \Pr(r_i|\boldsymbol{\beta}, \mathbf{x}) \\
&= \theta_i^{r_i}(1-\theta_i)^{1-r_i} \\
&= \left(\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right)^{r_i}\left(\frac{1}{1+\exp(\beta_1+\beta_2 x_i)}\right)^{1-r_i}
\end{aligned}
$$

- If r_i are independent of each other, the likelihood is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta})$$

- $\boldsymbol{\beta} = [\beta_1, \beta_2]^T$
- $\mathbf{x} = [x_1, \ldots, x_n]^T$

Specifying the Log-Likelihood Function

- The log-likelihood is

$$
\begin{aligned}
\log L(\beta) &= \log \prod_{i=1}^{n} L_i(\beta) \\
&= \sum_{i=1}^{n} \log \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{r_i} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{1-r_i} \\
&= \sum_{i=1}^{n} \log \theta_i^{r_i} (1 - \theta_i)^{1-r_i}
\end{aligned}
$$

- We perform further calculation with R

How to find the Maximum

- Ideal: Differentiate the (log-)likelihood function and find the maximum (i.e. solve the score equation)
- Problem: Can't always solve the equation
- Reality: "Search" the maximum by numerical methods (computation)
  - Bisection method（二分法）
  - Gradient method（勾配法）
  - **Newton (Newton-Raphson) method**
  - etc.

Hit Ratio

- Prediction by logistic regression: the probability that each response equals 1
- What we'd like to know is if the response is 0 or 1
- Predict the response using the predicted probabilities
  1. Predict $y_i = 1$ if $\Pr(y_i = 1)$ exceeds (or falls below) a certain threshold (usually 0.5)
  2. Simulation
- Calculate the ratio of observations whose predicted value matches the observed value
- We use the ratio as an index of fit
- Baseline: $\max(\bar{y}, 1 - \bar{y})$

ROC Curves

- ROC (receiver operating characteristic, 受信者操作特性) Curve
- Plot true positive rate (TPR, sensitivity) versus false positive rate (FPR, $1-$ specificity)
- Predict the response is 1 for $\pi > c$ and 0 for $\pi \leq c$
- Draw a curve by changing the value of c from 1 to 0
- Random response (noise): ROC should be 45 degree line
- Accurate model: ROC curve should bend toward the upper left corner
- Good model has a large area under the curve (AUC)

AIC

- Akaike Information Criterion (AIC)

$$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2k$$

- k is the number of free parameters
- Better model has smaller AIC
  - Better as the maximum of the log-likelihood gets larger
  - Better with fewer parameters