

Logistic Regression

Wei Wang

Dec 7, 2010

Binary Response Data

- In a lot of applications, we are dealing with binary response data, rather than continuous response data.
- success/failure, won/lost, healthy/ill etc.
- We want a reasonable model that link predictors X and response y

Model Setup

- Response y takes value 1 and 0 with probabilities p and $1 - p$. This is the most basic Bernoulli model.
- A natural and simple way to go is to let

$$E y = p = X\beta$$

- Recall in simple linear regression, we have the similar property

$$E y = X\beta$$

Model Setup

- However, this doesn't work out well because $X\beta$ doesn't necessarily fall into $[0, 1]$ unit interval. (This is not a concern for simple linear regression.)
- If we still want to maintain **linearity** of our model, we need to come up with some transformation to scale $X\beta$ to $[0, 1]$ unit interval.
- It turns out that there are a lot of such transformations, the most widely used is inverse-logit function.

$$\begin{aligned}\text{logit}(p) &= \log \frac{p}{1-p} \\ \text{logit}^{-1}(a) &= \frac{\exp(a)}{1 + \exp(a)}\end{aligned}$$

Model Setup

- More formally, **logistic regression model** is given by

$$P(y = 1) = \text{logit}^{-1}(x^T \beta) = \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)}$$

$$P(y = 0) = 1 - P(y = 1) = \frac{1}{1 + \exp(x^T \beta)}$$

- Notice that unlike simple linear regression, logistic regression doesn't have a variance parameter. That is because Bernoulli distribution can be characterized solely by its mean p .

A Simple Example: Voting and Income

- Conservative parties generally receive more support among voters with higher income.
- We use a poll from 1992 Presidential Election.
- For 1197 respondents, $y_i = 1$ if he/she preferred Bush, $y_i = 0$ if he/she preferred Clinton. Third party preferred is excluded.
- Income level is dicretized into 5 categories.
- Using standard statistical software, the fitted model is given by

$$\text{logit}(P(y = 1)) = -1.40(0.19) + 0.33(0.06) * \text{income}$$

Voting and Income: Graphical Illustrations

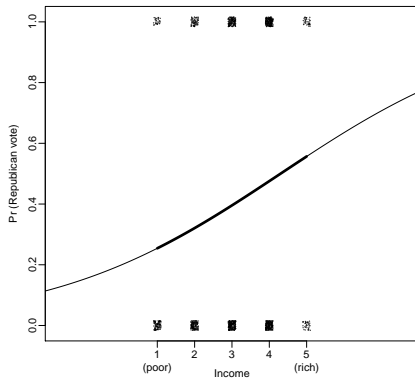


Figure: *Fitted model of voting-income logistic regression*

Interpretation of the Coefficients: Odds Ratio

- If two outcomes have the probabilities $(p, 1 - p)$, then $p/(1 - p)$ is called the **odds**.
- The ratio of two odds $\frac{p_1(1-p_2)}{(1-p_1)p_2}$ is called **odds ratio**.
- In the voting and income example, the slope of income is actually the odds ratio when income increases by **one** category.

$$\beta_{\text{income}} = \log \left(\frac{P(y = 1 | \text{income} = k + 1)}{P(y = 0 | \text{income} = k + 1)} \bigg/ \frac{P(y = 1 | \text{income} = k)}{P(y = 0 | \text{income} = k)} \right)$$

- Folks at biostat really like odds ratio. But for some applications, odds ratio seems obscure and difficult to understand.

Interpretation of the Coefficients: Probability Scale

We take a look at the standard logistic function curve.

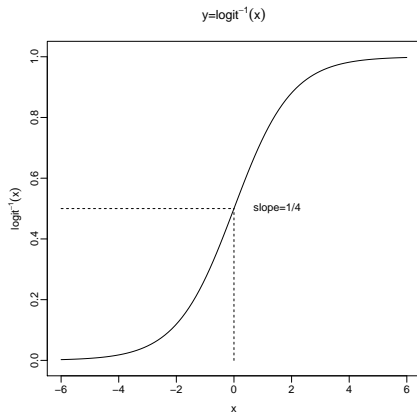


Figure: Curve of inverse logistic function.

Interpretation of the Coefficients: Probability Scale

- The logistic curve is steepest at its center, with slope $\beta/4$. As a rule of convenience, we can take $\beta/4$ as an upper bound of predictive difference corresponding to one unit difference in predictor. This is valid near the midpoint of the logistic curve.
- In the voting-income example, the fitted model suggests that a difference of 1 in income category corresponds to at most an $.33/4 \approx 8\%$ positive difference in the probability of supporting Bush.

Probit Model and Latent Variable Formulation

- Another popular model for binary response data is Probit Model.
- In Probit Model, we choose a different mapping from \mathbb{R} to $[0, 1]$ unit interval: the CDF of standard normal.
- If ξ is a standard normal random variable and Φ is its CDF,

$$p = \text{probit}(x^T \beta) = P(\xi + x^T \beta > 0) = \Phi(-x^T \beta)$$

Probit Model and Latent Variable Formulation

- We can view probit model in a **Latent Variable Formulation**. Each discrete outcome y_i is associated with a continuous, unobserved z_i

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases}$$
$$z_i = \mathbf{x}_i^T \boldsymbol{\beta} + \xi_i$$

ξ_i are i.i.d. standard normal random variables.

- In fact, we can incorporate logistic regression model into this formulation if we let ξ_i follows **logistic distribution**, which is defined by

$$P(\xi_i < x) = \text{logit}^{-1}(x)$$

Probit Model and Latent Variable Formulation

- A lot of statistical problems are easier to understand and tackle from a Latent Variable Formulation point of view.
- In practice, logistic regression and probit regression are not very different. In fact, logistic distribution is very close to normal distribution with mean 0 and standard deviation 1.6.
- So to use logistic or probit regression is mostly just a matter of taste.

Estimation

- Maximum likelihood approach is used to estimate the parameters of logistic regression

$$\hat{\mathbf{b}} = \arg \max \log L(\beta)$$

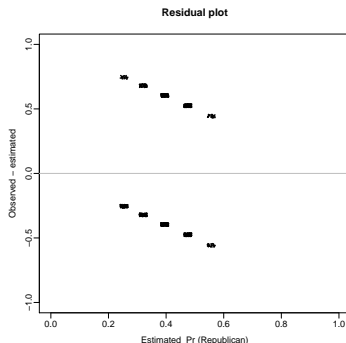
- The covariance matrix of the estimators is given by the Hessian matrix of loglikelihood function.

$$\mathbf{s}^2\{\mathbf{b}\} = -\left(\frac{\partial^2 \log L(\beta)}{\partial \beta^2}\right)^{-1}$$

- Iterative Weighted Least Squares** scheme is used to estimate the parameters. In some literature, it is also called **Fisher's Scoring Method**.
- These estimations are routinely provided by standard softwares.

Model Checking: Residuals

- Based on the same rational of diagnostics of simple linear model, we may want to look at the **Fitted values v.s. Residuals** plot.
- Because of the discreteness of the response, the plot doesn't show us much useful information.



Model Checking: Residuals

- Instead, we can look at the binned residual plot.

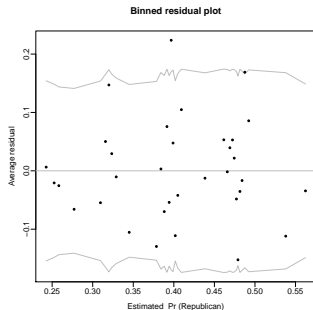


Figure: Binned residual plot for voting-income example

- We can see that most of the points lie inside the 95% interval and no dramatic pattern appears.

Model Checking: Error Rate

- Another thing that we might want to look at is Error Rate.
- If we do deterministic prediction, i.e. guessing $y_i = 1$ when fitted value $\text{logit}^{-1}(x_i^T \hat{\beta}) > .5$ and $y_i = 0$ if otherwise, then we define the proportion of wrong prediction as **Error Rate**.
- Error Rate need to be at least less than .5(the error rate when we just randomly guess).

Model Checking: Deviance

- **Deviance** is defined as -2 times the logarithm of the likelihood function. The better the model fit, the lower the deviance is.
- If a predictor that is random noise is added to the model, we expect deviance to decrease on average by 1.
- If a predictor that is informative is added to the model, we expect deviance to decrease on average by more than 1. The larger the increase, the more relevant the predictor is for our model.
- Deviance is a fundamental concept in model comparison and selection.

Model Checking: Deviance

- For the voting-income example, the null model (only has intercept) has a deviance of 1591.
- Added income as a predictor, the model has a deviance of 1557.
- This shows that income is informative for predicting vote outcome.

Software Implementations

- R: `glm`
- Matlab: `glmfit`