Binary Response and Logistic Regression

Wei Wang

Dec 7, 2010

Outline

- Model Setup
- A simple example: Voting, Income and Gender
- Interpretation of coefficients
- Latent Variable Formulation: Probit and Robit model
- Estimation and Inference about Parameters
- Prediction
- Diagnostics

Binary Response Data

- In a lot of applications, we are dealing with binary response data, rather than continuous response data.
- sucess/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 natrual and simple way to go is to let

$$Ey = p = X\beta$$

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

$$Ey = X\beta$$

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.

$$\begin{aligned} \text{logit}(p) &= \log \frac{p}{1-p}, p \in (0, 1) \\ \text{logit}^{-1}(a) &= \frac{\exp(a)}{1+\exp(a)}, a \in \mathbb{R} \end{aligned}$$

Model Setup

- More formally, logistic regression model is given by

$$\begin{aligned} P(y = 1) &= \text{logit}^{-1}(x^T\beta) = \frac{\exp(x^T\beta)}{1 + \exp(x^T\beta)} \\ P(y = 0) &= 1 - P(y = 1) = \frac{1}{1 + \exp(x^T\beta)} \end{aligned}$$

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

A Simple Example: Voting and Income

- Repulicans generally receive more support among voters with higher income.
- We use a poll from 1992 Presidential Election.
- For 1197 respondents, $y_i = 1$ if the respondent preferred Bush, $y_i = 0$ if the respondent preferred Clinton. Third party preference is exculded.
- 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

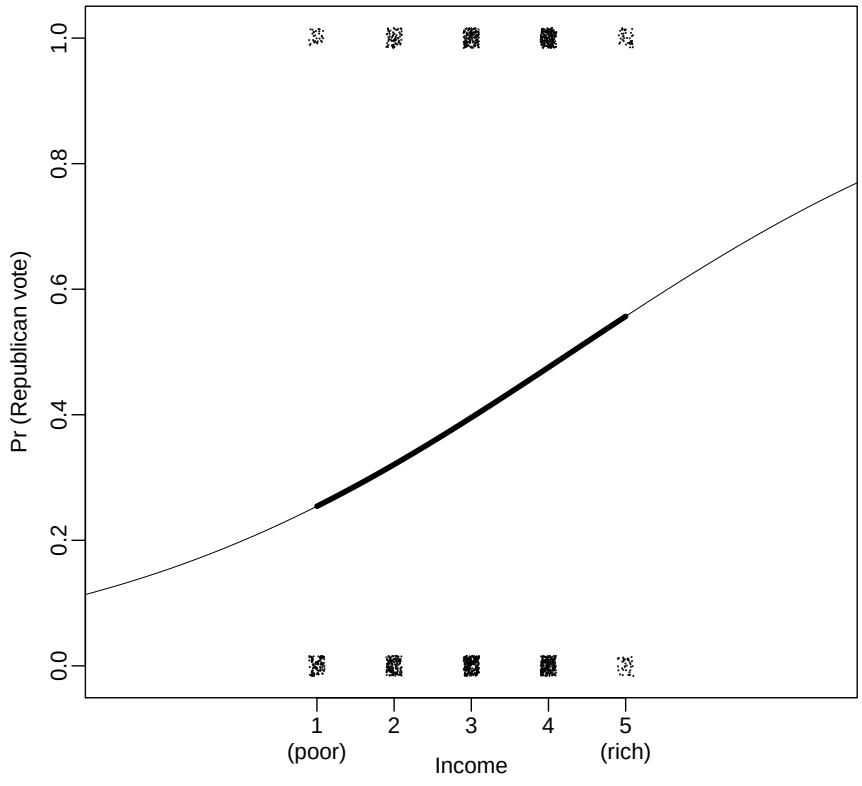


Figure: *Fitted model of voting-income logistic regression*

Voting and Income and Gender: Adding interaction

- Further, we throw in female indicator (0 for male and 1 for female) and income-female interaction into this model.
- The fitted model is given by

$$\begin{aligned}\text{logit}(\mathrm{P}(\mathrm{y}=1)) &= -1.89(0.33) + 0.49(0.10) * \text{income} \\ &\quad +0.78(0.40) * \text{female} \\ &\quad -0.28(0.12) * \text{income} \times \text{female}\end{aligned}$$

- It is widely believed that males tend to support Republican more than females. But the estimate for female here is 0.78. Is that a contradiction to popular belief? We cannot overlook interaction term.

Voting and Income and Gender: Graphical Illustrations

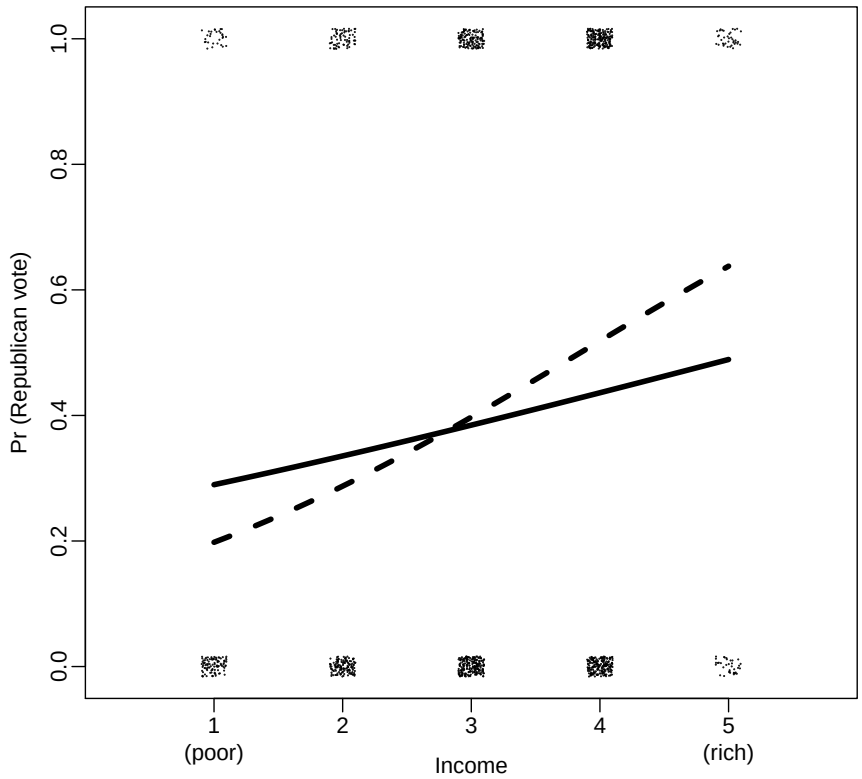


Figure: *Fitted model of income-voting curves for male (dashed line) and female (solid line)*

Interpretation of the Coefficients: Odds Ratio

- If two outcomes have the probabilites $(p, 1-p)$, 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.

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

- 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.

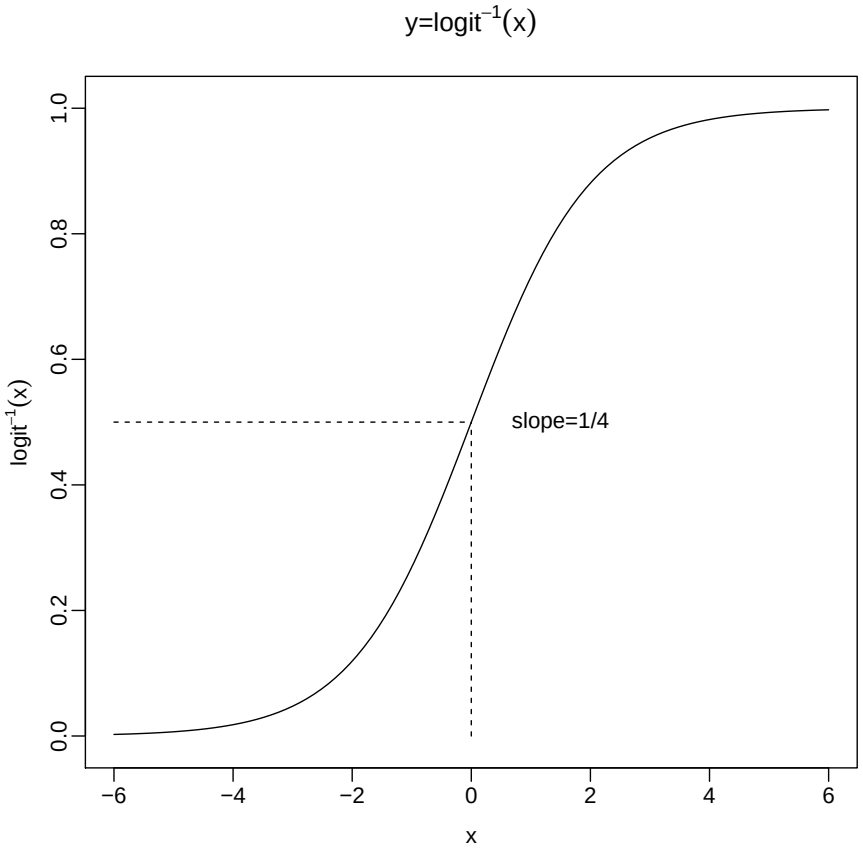


Figure: Curve of inverse logistic function.

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 dicrete outcome y_i is associated with a continuous, unobserved z_i

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\
z_i &= x_i^T \beta + \xi_i
\end{aligned}
$$

  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.

Robust Regression: Robit Model

- When a regression model can have occasionally very large errors, it is appropriate to use a Student-t distribution for the errors.

- The Robit (robustified) Model is given by

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\
z_i &= x_i^T \beta + \xi_i \\
\xi_i &\sim t_\nu\left(0, \frac{\nu - 2}{\nu}\right)
\end{aligned}
$$

with degree of freedom ν is estimated from the data and the t distribution is scaled to have standard deviation 1.

Estimation

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

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

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

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

- Iterative Weighted Least Squares scheme is used to estimate the parameters. In some literature, it is also called Fisher's Scoring Method.

- These estimates are routinely provided by standard softwares.

Inference about parameters: Wald Test

- The Wald Test is used for testing and confidence interval building for a single parameter.
- It is given by

$$\frac{b_k - \beta_k}{s\{b_k\}} \rightarrow N(0,1)$$

- We can do hypothesis testing and build confidence interval based on this.

Inference about parameters: Likelihood Ratio Test

- Likelihood Ratio Test is used when we want to determine whether a subset of the X predictor can be dropped. It is derived from general Likelihood Ratio Test framework.
- If the predictors are like this

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}$$

- And we want to test

$$\begin{aligned} H_0 : &\quad \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0 \\ H_a : &\quad \text{not all the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned}$$

- Let $L(F)$ represent the likelihood of the full model and $L(R)$ represent the likelihood of the reduced model (H_0).

Inference about parameters: Likelihood Ratio Test

- Then we find the maximum likelihood of both the full and the reduced models, $\max_\beta L(F)$ and $\max_\beta L(R)$.

- The Likelihood Ratio Test statistic is given by

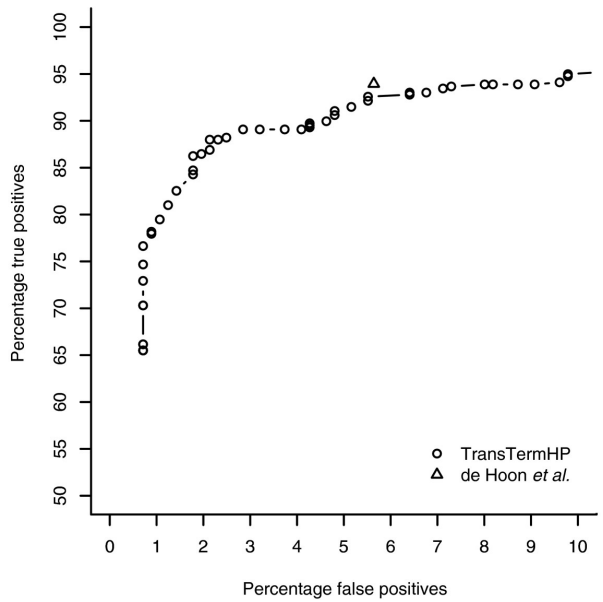
$$-2 \log \frac{\max_\beta L(F)}{\max_\beta L(R)} \rightarrow \chi^2_{p-q}$$

Prediction and ROC curve

- Naive prediction is to choose 0.5 as the cutoff point to predict, i.e. predicting $y_i = 1$ when fitted value $\text{logit}^{-1}(x_i^T \hat{\beta}) > .5$ and $y_i = 0$ if otherwise.
- Error Rate should be at least less than .5 (the error rate when we just randomly guess).
- But to fully express the predicting power of the model, we might want to find the best cutoff points. ROC curve is for displaying this information.

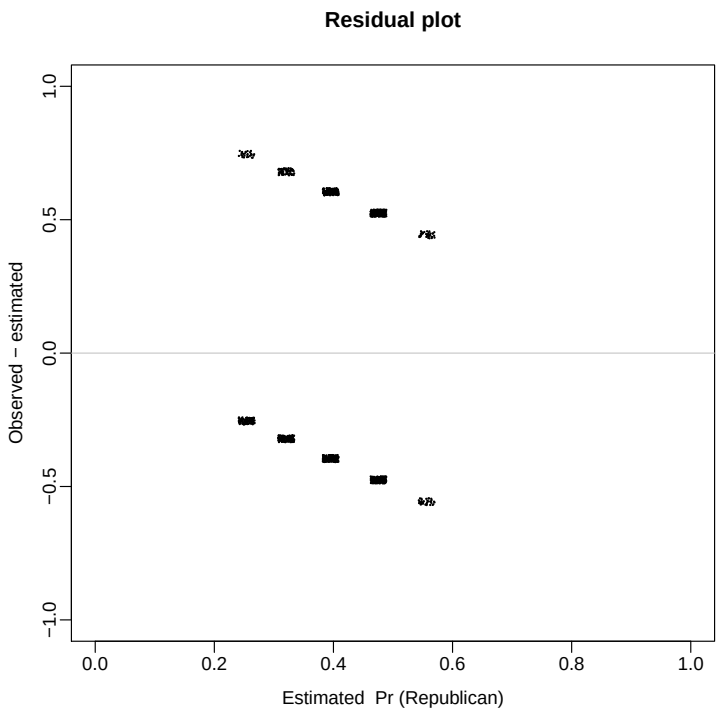
Predition and ROC curve

- ROC (receiver operating characteristic) curve comes from Signal Processing literature. It plots False Positive rate $(\mathrm{P}(\hat{\mathrm{Y}} = 1|\mathrm{Y} = 0))$ against True Positive rate $(\mathrm{P}(\hat{\mathrm{Y}} = 1|\mathrm{Y} = 1))$ for different cutoff point choices.
- The ROC curve should be above the diagonal line (which means random guessing), and the further away it is, the better the predicting power is.



Model Diagnostics: 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 Diagnostics: Residuals

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

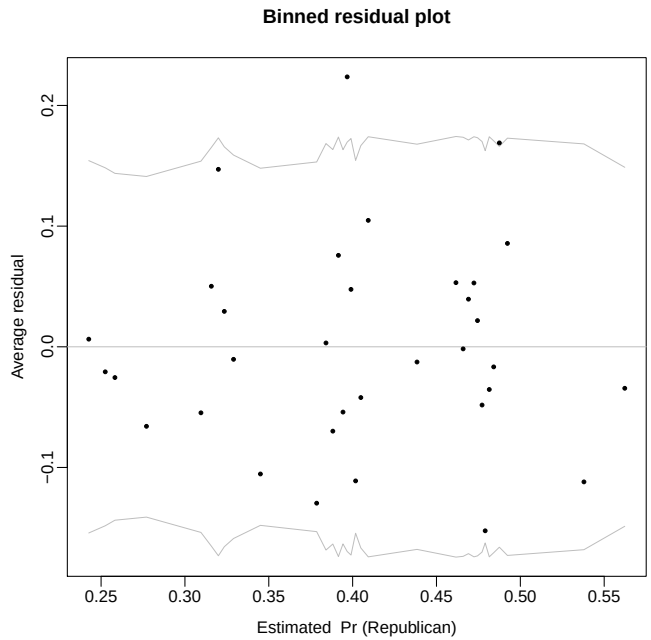


Figure: Binned residual plot for voting-income example

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

Model Diagnostics: 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 fundmental concept in model comparison and selection.

Model Diagnostics: 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(y ~ x1 + x2, family = binomial('logit'))`
  `glm(y ~ x1 + x2, family = binomial('probit'))`
  `glm(y ~ x1 + x2, family = binomial('cauchit'))`

- Matlab: `glmfit`