CSSS 510: Lab 3

Logistic Regression

2017-10-13

0. Agenda

1. Deriving a likelihood function for the logistic regression model
2. Fitting a logit model using `optim()` and `glm()`
3. Simulating predicted values and confidence intervals
4. Simulating first differences
5. Assessing model fit
   - Likelihood ratio test
   - Akaike Information Criterion
   - Bayesian Information Criterion
   - Average vs Predicted Plots
   - ROC plots
   - Residual vs Leverage Plots

1. Deriving a likelihood function for the logistic regression model

Recall from lecture the logit model:

$$\begin{aligned} y_i &\sim \text{Bern}(y_i|\pi_i) \\ \pi_i &= \text{logit}^{-1}(\boldsymbol{x}_i\boldsymbol{\beta}) \\ \pi_i &= \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{x}_i\boldsymbol{\beta})} \end{aligned}$$

In the simple case, this stems from the latent variable model:

$$y^* = \beta_0 + \beta_1 x + \epsilon$$

where the relationship between latent variable y^* and the explanatory variable x is modeled using simple linear regression, and the binary outcome y is a function of the sign of y^*:

$$y = \begin{cases} 1, & \text{if } y^* > 0 \\ 0, & \text{if } y^* \leq 0 \end{cases} \tag{1}$$

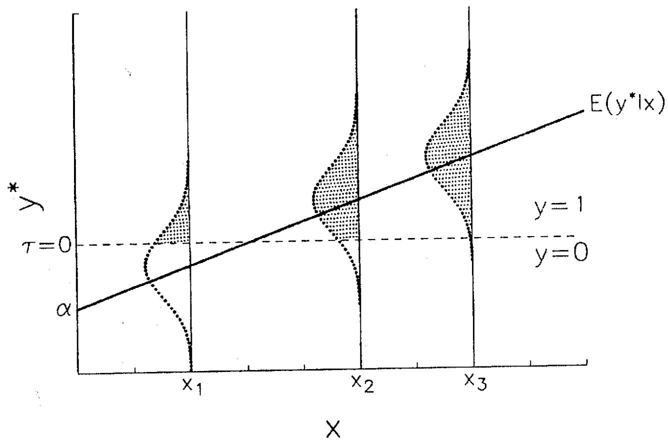

Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

The logistic regression model is obtained if we assume the errors of this latent variable model follow a standard logistic distribution. Recall that the pdf and cdf of the standard logisitic distribution are as follows:

$$f(t) = \frac{\exp(t)}{(1+\exp(t))^2}$$
$$F(t) = \frac{\exp(t)}{1+\exp(t)}$$

We therefore have the following:

$$\begin{aligned}
\Pr(y=1|x) &= \Pr(y^* > 0|x) \\
&= \Pr(\beta_0 + \beta_1 x + \epsilon > 0|x) \\
&= \Pr(\epsilon > -(\beta_0 + \beta_1 x)) \\
&= \Pr(\epsilon < \beta_0 + \beta_1 x) \\
&= F(\beta_0^L + \beta_1^L x)
\end{aligned}$$

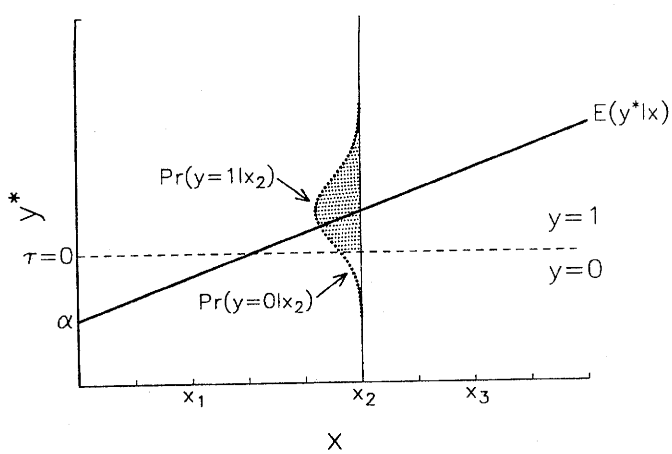

Figure 3.4. Probability of Observed Values in the Binary Response Model

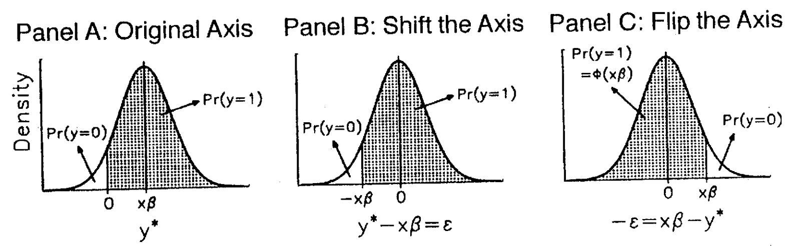

Figure 3.5. Computing $\Pr(y = 1 \mid x)$ in the Binary Response Model

Since we assume the errors follow a standard logistic distribution, we have

$$\Pr(y = 1|x) = F(\beta_0^L + \beta_1^L x) = \frac{\exp(\beta_0^L + \beta_1^L x)}{1 + \exp(\beta_0^L + \beta_1^L x)}$$

and E(ϵ)=0 and Var(ϵ) $= \frac{\pi^2}{3}$.

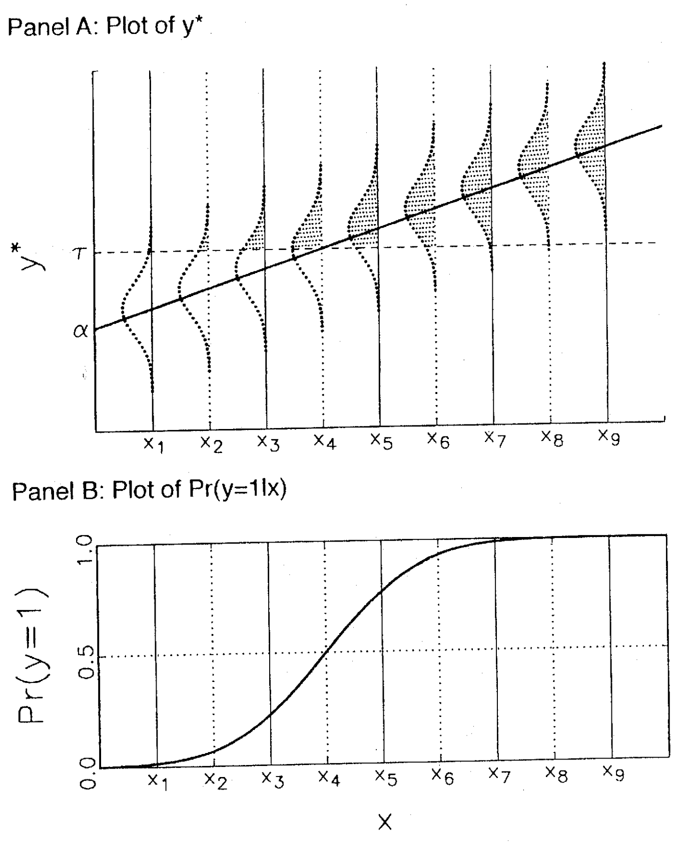

Figure 3.6. Plot of y^* and $\Pr(y = 1 \mid \mathbf{x})$ in the Binary Response Model

The logit function is the inverse of the logistic function:

$$\text{logit}(p) = \log\frac{p}{1-p}$$

or

$$\text{logit}^{-1}(p) = \frac{\exp(x)}{1+\exp(x)}$$

We therefore have the following

$$\Pr(y=1|x) = \text{logit}^{-1}(\beta_1^L + \beta_1^L x)$$

or

$$\text{logit}(\Pr(y=1|x)) = \beta_1^L + \beta_1^L x$$

or

$$\log\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = \beta_0^L + \beta_1^L x.$$

Recall from lecture that a Bernoulli distribution has the following pdf:

$$\Pr(y_i = 1|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

And the likelihood function can be derived from the joint probability:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\pi}|\boldsymbol{y}) &\propto \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{y}) &\propto \prod_{i=1}^{n} \left(\frac{1}{1+\exp(-\boldsymbol{x}_i\boldsymbol{\beta})}\right)^{y_i} \left(1 - \frac{1}{1+\exp(-\boldsymbol{x}_i\boldsymbol{\beta})}\right)^{1-y_i} \\
\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{y}) &\propto \prod_{i=1}^{n} (1+\exp(-\boldsymbol{x}_i\boldsymbol{\beta}))^{-y_i}(1+\exp(-\boldsymbol{x}_i\boldsymbol{\beta}))^{-(1-y_i)} \\
\log\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{y}) &\propto \sum_{i=1}^{n} -y_i\log(1+\exp(-\boldsymbol{x}_i\boldsymbol{\beta})) - (1-y_i)\log(1+\exp(\boldsymbol{x}_i\boldsymbol{\beta}))
\end{aligned}$$

2. Fitting a logit model using **optim**() and **glm**()

3. Simulating predicted values and confidence intervals

4. Simulating first differences

5. Assessing model fit