L11: Qualitative Response Regressions

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Revision: November 25, 2013

Today we are going to learn...

1. **The logit model**

2. **The probit model**

3. **The Poisson regression model**

The logit model

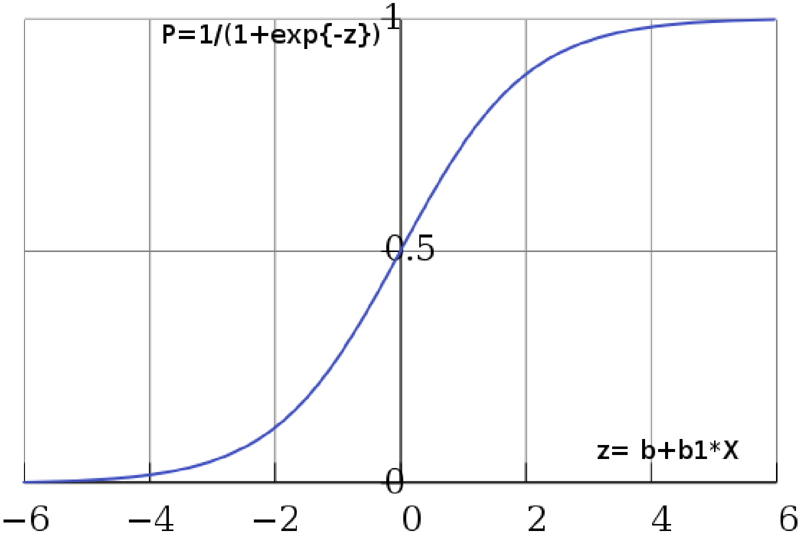
- We sometimes obtain such data
  - $$Y_i = \begin{cases} 1 & \text{The family own a house} \\ 0, & \text{The family does not own a hose} \end{cases}$$
  - X_i be the income of a family.
- We are interested in modeling such relationships between
  - $P_i = E(Y_i = 1|X_i)$ (the probability of a family owns a house) and
  - X_i (income)
- The **logit model** connects the two variables in this way

$$P_i = \frac{1}{1 + \exp(-(\beta_1 + \beta_2 X_i))}$$

- Alternatively we can write the model in this way

$$\log \frac{P_i}{1 - P_i} = \beta_1 + \beta_2 X_i$$

where $P_i/(1 - P_i)$ is called **odds ratio**: the ratio of probability of a family will own a house to the probability of not owing a house.


P=1/(1+exp{-z})
1
0.5
0
z= b+b1*X
−6
−4
−2
0
2
4
6

The logit model
↪ The interpretation

Let $I_i = \beta_1 + \beta_2 X_i$ and $L_i = \log\{P_i/(1 - P_i)\}$ be the **logit function**

- If P_i goes from 0 to 1 (I_i goes from $-\infty$ to ∞), the logit L_i goes from $-\infty$ to ∞.
- The logit L_i is now linear with X_i.
- If L is positive, something interesting happens (e.g. in favor of owing a house). If L is negative, the other way around.
- The slope β_2 measures the change in L for a unit change in X_i.

The logit model

↪ The estimation with grouped data

- Instead having individual base data, the data are grouped according to income level as follows (taking the income and owning a house example)
  - N_i: number of families at income X_i
  - n_i: number of families owning a house
- The estimation
  - Let $\hat{P}_i = n_i/N_i$
  - Obtain $\hat{L}_i = \log(\hat{P}_i/(1-\hat{P}_i))$
  - Liner regression $\hat{L}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- The underlying assumptions

$$u_i \sim N\left(0, \frac{1}{N_i P_i (1-P_i)}\right)$$

  which due to the fact that if P_i follows Bernoulli distribution, then the variance of $\hat{L}_i$ will be $1/\{N_i P_i (1-P_i)\}$.
- The estimated variance of the error term is

$$\hat{\sigma}^2 = \frac{1}{N_i \hat{P}_i (1-\hat{P}_i)}$$

- **Question**: how to apply GLS when heteroscedasticity presents?

The logit model

↬ Example with grouped data

- N_i: number of families at income X_i
- n_i: number of families owning a house

TABLE 15.5 Data to Estimate the Logit Model of Home Ownership

X (thousands of dollars) (1)	N_i (2)	n_i (3)	$\hat{P}_i$ (4) = (3) ÷ (2)	$1-\hat{P}_i$ (5)	$\frac{\hat{P}_i}{1-\hat{P}_i}$ (6)	$\hat{L}_i = \ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right)$ (7)	$N_i\hat{P}_i(1-\hat{P}_i) = w_i$ (8)	$\sqrt{w_i} = \sqrt{N_i\hat{P}_i(1-\hat{P}_i)}$ (9) = $\sqrt{(8)}$	$\hat{L}_i^* = \hat{L}_i\sqrt{w_i}$ (10) = (7)(9)	$\hat{X}_i^* = \hat{X}_i\sqrt{w_i}$ (11) = (1)(9)
6	40	8	0.20	0.80	0.25	−1.3863	6.40	2.5298	−3.5071	15.1788
8	50	12	0.24	0.76	0.32	−1.1526	9.12	3.0199	−3.4807	24.1592
10	60	18	0.30	0.70	0.43	−0.8472	12.60	3.5496	−3.0072	35.4960
13	80	28	0.35	0.65	0.54	−0.6190	18.20	4.2661	−2.6407	55.4593
15	100	45	0.45	0.55	0.82	−0.2007	24.75	4.9749	−0.9985	74.6235
20	70	36	0.51	0.49	1.04	0.0570	17.49	4.1816	0.1673	83.6506
25	65	39	0.60	0.40	1.50	0.4054	15.60	3.9497	1.6012	98.7425
30	50	33	0.66	0.34	1.94	0.6633	11.20	3.3496	2.2218	100.4880
35	40	30	0.75	0.25	3.0	1.0986	7.50	2.7386	3.0086	95.8405
40	25	20	0.80	0.20	4.0	1.3863	4.00	2.000	2.7726	80.0000

Take home exercise: implement this in R

The logit model

↬ Estimate logit model with ungrouped (individual) data

- **The idea**: using maximum likelihood method with binomial distribution.
- One owns a house ($Y = 1$) or do not own a house ($Y = 0$) can be represented with **Bernoulli distribution**

$$\Pr(y; p) = p^y (1-p)^{1-y} \quad \text{for } y \in \{0, 1\}.$$

- The likelihood function is as follows

$$l(\beta) = \sum_{n=1}^{N} \{y_i \log P_i + (1 - y_i) \log(1 - P_i)\}$$

where

$$P_i = \frac{1}{1 + \exp(-(\beta_1 + \beta_2 X_i))}$$

- Note that the sum of n Bernoulli samples will be **binomial** distributed.
- To obtain $\hat{\beta}$, use Newton-Raphson algorithm

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta}$$

Example 15.5

The probit model
↬ Again the owning housing example

- Let $I_i = \beta_1 + \beta_2 X_i$, where X_i is the income for *i*th family
- We would like to find the **critical value** (threshold) of I_i^* so that if $I_I > I_i^*$, the family will own a house ($Y_i = 1$), and $I_I \leqslant I_i^*$, the family will not own a house ($Y_i = 0$).
- Under the normality assumption, we have

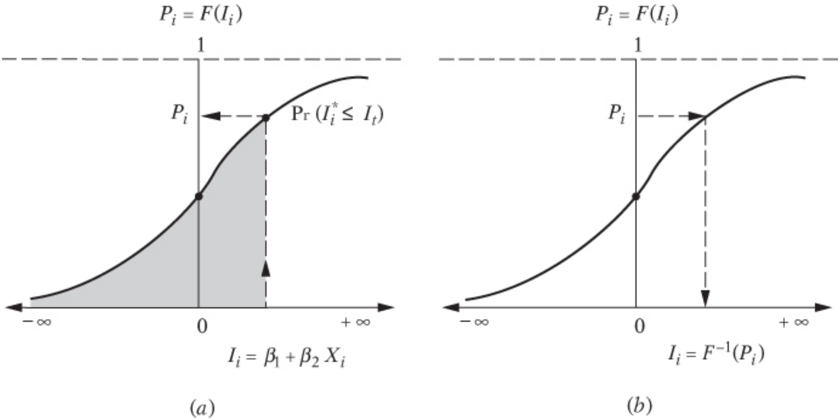
$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} \exp\left\{\frac{-z^2}{2}\right\} dz$$

- And our previous example will be expressed in this way

$$P_i = P(Y = 1|X) = P(I_i^* \leqslant I_i) = P(Z \leqslant \beta_1 + \beta_2 X_i) = F(\beta_1 + \beta_2 X_i)$$

The probit model

↪ The illustrative figure



The probit model
↪ Estimation with grouped data

- First obtain $\hat{P}_i$ as in logit model for grouped data
- Transform the CDF into PDF

$$\hat{I}_i = F^{-1}(\hat{P}_i)$$

where $F()$ is the CDF function of normal distribution and $F^{-1}()$ is the inverse of the CDF, i.e. the PDF

- Now you make regression

$$\hat{I}_i = \beta_1 + \beta_2 X_i$$

- **Question:** How do we do the prediction?

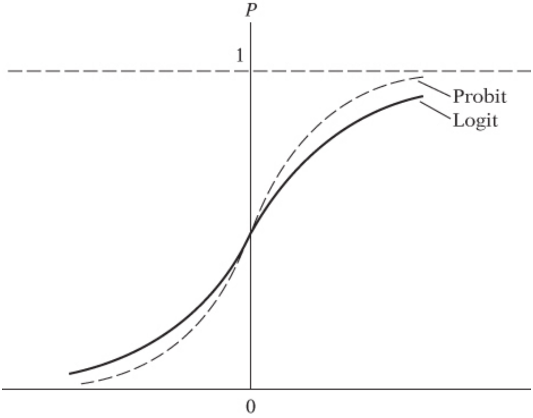
The probit model
↪ Estimation with grouped data

- **Exercise 15.6**
- **Take home question:**
  - For the grouped data, how would implement the probit model if the assumption follows **student -t** distribution?

The probit model
↪ Estimation with ungrouped data

Example 15.5 continued

Comparison between the probit and logit curve



Poisson distribution
↪ Count type data

- Let's think about this type of data
  - How many days do you take for vacation?
  - How often do you go to the gym per week?
  - How often are you absent of the class per semester?
- The characteristics
  - The variables are nonnegative.
  - The variables are discrete.
  - Some are rare/infrequent counts.
- What kind of distribution captures such phenomena?

Poisson distribution

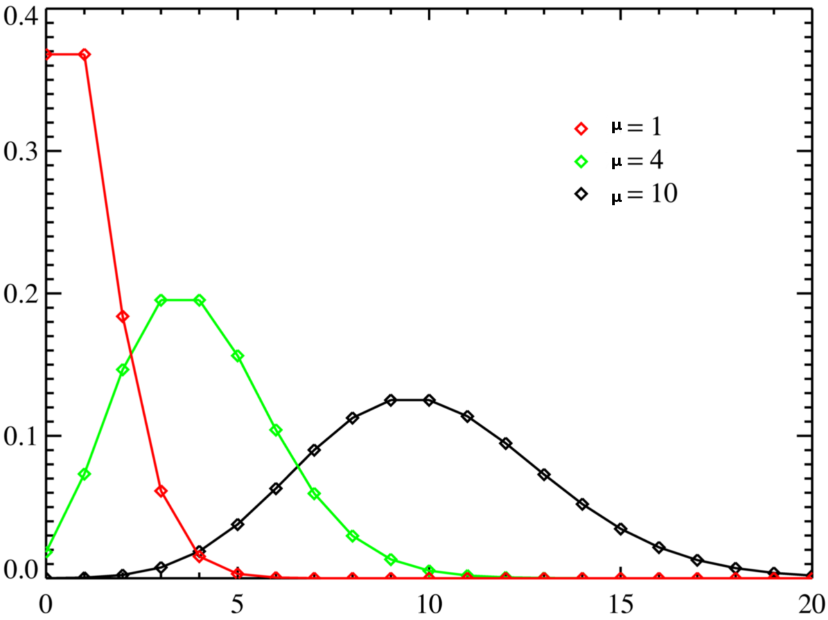
- The probability mass function

$$p(Y) = \frac{\mu^Y e^{-\mu}}{Y!}$$

  for $y = 0, 1, 2, ...$ where y is the occurrence of particular event and $Y! = Y \times (Y-1) \times (Y-2) \times ... \times 1$ is the factorial.

- The Poisson distribution has the same mean and variance

$$E(Y) = Var(Y) = \mu$$



0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
μ = 1
μ = 4
μ = 10

The Poisson model

- We model the mean value (positive) of Y_i with covariates X_1, X_2,...,X_k

$$\mu_i = E(Y_i) = exp(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

or alternatively we write the model as

$$p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \text{ where } \mu_i = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k$$

- The interpretation of the model
  - How frequently the event happens to ith observations on average?

$$\mu_i = E(Y_i) = exp(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

  - What is the probability the event happens exactly y_i times to ith observation?

$$p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

  - What is the probability the event happens at most y_i times to ith observation?

$$\sum_{l=0}^{y_i} \frac{\mu_i^{l} e^{-\mu_i}}{l!}$$

  - What is the probability the event happens at lest y_i times to ith observation?

$$1 - \sum_{l=0}^{y_i} \frac{\mu_i^{l} e^{-\mu_i}}{l!}$$

The Poisson model

- Estimate the Poisson model with maximum likelihood method
  - The likelihood
$$p(y_1, y_2, ...y_n) = \prod_{i=1}^{n} p(y_i)$$
  - The log likelihood
$$\begin{aligned}\log p(y_1, y_2, ...y_n) &= \sum_{i=1}^{n} \log p(y_i) = \sum_{i=1}^{n} [y_i \log(u_i) - u_i - \log(y_i!)] \\ &= \sum_{i=1}^{n} [y_i(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k) \\ &\qquad - \exp(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k) - \log(y_i!)]\end{aligned}$$
  - Then maximize $\log p(y_1, y_2, ...y_n)$ with respect to $\beta_1, \beta_2, ..., \beta_k$

The Poisson model

Example 15.8