Quantitative Data Analysis

Master in Management - Academic Year 2023/2024

Ricardo Moura

ISEG

March 13, 2024

Binary models

Now let us consider that Y has only two values, $Y \in \{0, 1\}$.

Then, $E(Y|\mathbf{X}) = P(Y = 1|\mathbf{X}) = p(\mathbf{X})$ and then $0 < E(Y|\mathbf{X}) < 1$

If we write the estimated equation as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

remember that $\hat{y}$ is the predicted prob. of success and $\hat{\beta}_0$ is the predicted probability of success when all x_j are zero. *Ceteris paribus* an increase of 1 in x_j means an increase of β_j in the $p(\mathbf{X})$.

But this models have the problem that can have values < 0 or > 1.

Solution: $P(Y = 1|\mathbf{X}) = G(\mathbf{X}\beta)$ that is strictly between 0 and 1.

Binary Models

Most of the estimation is based in the maximum likelihood estimation

$$L(\beta|\mathbf{data}) = \prod_{i=1}^{n} G(\mathbf{x}_i'\beta)^{y_i} \left(1 - G(\mathbf{x}_i'\beta)\right)^{1-y_i}$$

where $G(\mathbf{x}_i'\beta) = P(y_i = 1|\mathbf{x}_i)$
And maximizing the logarithm of the likelihood function

$$l(\beta|\mathbf{data}) = ln(L(\beta|\mathbf{data})) = \sum_{i=1}^{n} \left(y_i ln[G(\mathbf{x}_i'\beta)] + (1 - y_i) ln[1 - G(\mathbf{x}_i'\beta)]\right)$$

Binary models

Model: $E(Y_i|\mathbf{X}_i) = G(\mathbf{X}_i'\beta)$; Most well known models

- Probit: Based on the normal cdf

$$G(\mathbf{x}_i'\beta) = \Phi(\mathbf{x}_i'\beta) = \int_{-\infty}^{\mathbf{X}\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{-(\mathbf{x}_i'\beta)^2}{2}} d\mathbf{x}\beta$$

- Logit: Based on the $logit(p) = log\left(\frac{p}{1-p}\right)$

$$G(\mathbf{x}_i'\beta) = \Lambda(\mathbf{x}_i'\beta) = \frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{x}_i'\beta}}$$

- Cloglog: Based in the complementary log-log function $cloglog(p) = log(-log(1-p))$

$$G(\mathbf{x}_i'\beta) = 1 - e^{-e^{\mathbf{x}_i'\beta}}$$

logit Y X_1 X_2
probit Y X_1 X_2
cloglog Y X_1 X_2

Binary Models - Partial effects

The partial effects are based in the partial derivative

$$g(z) = \frac{dG(z)}{dz}$$

For $\Delta x_j = 1$ we have $DeltaP(Y = 1|\mathbf{X}) = \beta_j g(\mathbf{x}_i'\boldsymbol{\beta})$

- Logit: $\beta_j \lambda(\mathbf{x}_i'\boldsymbol{\beta}) = \beta_j \Lambda(\mathbf{x}_i'\boldsymbol{\beta})\left[1 - \Lambda(\mathbf{x}_i'\boldsymbol{\beta})\right]$
- Probit: $\beta_j \phi(\mathbf{x}_i'\boldsymbol{\beta}) = \beta_j \frac{1}{\sqrt{2\pi}} e^{-\frac{-(\mathbf{x}_i'\boldsymbol{\beta})^2}{2}}$
- Cloglog: $g(\mathbf{x}_i'\boldsymbol{\beta}) = \left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right] e^{\mathbf{x}_i'\boldsymbol{\beta}} = e^{\mathbf{x}_i'\boldsymbol{\beta}} e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}}$

Binary Models - Partial effects

Calculation of partial effects:

- Average partial effect: calculate the effect for each individual and then average them
  `margins, dydx(varlist)`
- Partial effect evaluated at the mean: average each regressor and then replace in the partial effect
  `margins, dydx(varlist) atmeans`
- Replace specific values in $\mathbf{X}$
  `margins, dydx(varlist) at(...)`

Binary Models - Significance

Test for some regressors joint significance

- Models:
  - Unrestricted: $G(\beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g + \beta_{g+1} X_{g+1} + \cdots + \beta_p X_p)$
  - Restricted: $G^*(\beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g)$
- Hypothesis:

$$H_0 : \beta_{g+1} = \cdots = \beta_p = 0 \quad \textit{versus} \quad H_1 : \mathrm{No} H_0$$

  Assuming H_0 we select the restricted model
- likelihood ratio test:
  $LR = 2[L_{unrestricted}(\boldsymbol{\beta}|data) - L_{Restricted}(\boldsymbol{\beta}|data)] \sim \chi^2_{p-g}$
- Wald test (Robust for large samples):
  $W = \hat{\boldsymbol{\beta}}'_D [Var(\hat{\boldsymbol{\beta}}_D)]^{-1} \hat{\boldsymbol{\beta}}_D \sim \chi^2_{p-g}$ where $\hat{\boldsymbol{\beta}}_D = (\hat{\beta}_{g+1}, \ldots, \hat{\beta}_p)$
  In this course we may use wald test in stata after estimating the unrestricted model using
  `test` X_{g+1} `...`X_p

Binary Models - Significance

Test for global significance is a part of the test of some regressors joint singnificance where $H_0 : \beta_1 = \cdots = \beta_p = 0$

Test for individual significance

- The wald test resumes to be the square of $Z = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim N(0, 1)$

which are included in the software (`probit`, `logit`,...) output

Binary Models - RESET

1. Estimate the full model

$$P(Y = 1|x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_p X_p)$$

2. Obtain $(\mathbf{X}\hat{\beta})^2$, $(\mathbf{X}\hat{\beta})^3$, ...
   `predict XB, xb` (after estimating the model)
   `gen XB_2=XB^2` `gen XB_3=XB^3`
3. Estimate the auxiliary(artificial) model

$$P(Y = 1|x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_p X_p + \gamma_1(\mathbf{X}\hat{\beta})^2 + \gamma_2(\mathbf{X}\hat{\beta})^3 + \ldots)$$

4. Apply a LR/Wald Test for the joint significance of $(\mathbf{X}\hat{\beta})^2, (\mathbf{X}\hat{\beta})^3, \ldots,$ that is, test if $\gamma_1 = \gamma_2 = \cdots = 0$

Binary Models - Selection criteria

Among models that where not rejected by RESET test, you may use the correct classifications

	$y_i = 1$	$y_i = 0$	Total
$\hat{y}_i = 1$	n_{11} (TP)	False Positive (FP)	
$\hat{y}_i = 0$	False Negative (FN)	n_{00} (TN)	
Total	n_1	n_0	n

Table: Classification Table. Stata: `estat classification`

- $\hat{y}_i = \begin{cases} 1, & if \quad P(y_i \hat{=} 1|x_i) \geq 0.5 \\ 0, & if \quad P(y_i \hat{=} 1|x_i) < 0 \end{cases}$
- Accuracy: % of correct classifications $\frac{n_{11}+n_{00}}{n} \times 100\%$
- Recall or Sensitivity: % of 1's correctly classified: $\frac{n_{11}}{n_1} \times 100\%$
- Specificity: % of 0's correctly classified: $\frac{n_{00}}{n_0} \times 100\%$