Binary response models

or studying discrete choices between 2 alternatives

Lucas van der Velde

Warsaw School of Economics
FAME|GRAPE

October 2018

Which model are used to model probabilities?

Linear probability model (LPM)

Probit model

Logit model

Which model are used to model probabilities?

Linear probability model (LPM)

- $P(y_i = j) = G(x'\beta) = x'\beta$
- Simple to estimate and interpret
- Unrealistic predictions + heteroskedasticity
- Used when FE in regressors

Probit model

Logit model

Which model are used to model probabilities?

Linear probability model (LPM)

Probit model

- $P(y_i = j) = G(x'\beta) = \Phi(x'\beta)$
- Assumes a latent variable with normal distribution
- Most interpretations in terms of marginal probability effects

Logit model

- $P(y_i = j) = G(x'\beta) = \Lambda(x'\beta + u_i)$
- Assumes a latent variable with logistic distribution
- Can also be interpreted as linear model of Log-odds ratio

Interpreting the parameters

Linear probability model (LPM)

- $\partial P(y = j)/\partial x_k = \beta_k$
- β is the marginal effect

Interpreting the parameters

Linear probability model (LPM)

- $\partial P(y=j)/\partial x_k = \beta_k$
- β is the marginal effect

Non-linear models

- $\partial P(y=j)/\partial x_k = \beta_k * G'(x'\beta)$
- Marginal effect has same sign as β
- $|\text{Marginal effect}| < |\beta|$
- Marginal effect is largest when $x'\beta = 0$

Filling the $x'\beta$

We have three candidate solutions

1. Marginal effects at Mean
   - We evaluate x at E(x)
   - For binary / categorical data, it makes little sense
2. Marginal effects at specific values
   - Choose values of x that are interesting: e.g. women, aged 40-50, working...
   - Specify why were those values selected
3. Average Marginal Effects
   - Average MPE for all observations
   - Popular approach $\rightarrow$ not always sensitive

Other interpretations

Discrete changes

- When $\lim \Delta x_k \neq 0$
- Typical case: categorical variable $\Delta x_k = 1$
- $\Delta P(Y = j) = G(x'\beta + \Delta x_k \beta_k) - G(x'\beta)$
- Previous considerations for $x'\beta$ apply here as well.

Changes in odds-ratio

- Only for logit model
- $Odds = exp(x'\beta)$
- Relative change in odds $= exp(\Delta x'\beta) - 1$

Tests for parameter values

1 **Z-test for parameters**

- ML estimates are asymptotically normally distributed
- $Z = \frac{\hat{\theta} - \theta_0}{\sqrt{Asy.var(\theta)}}$
- $|Z| > 1.96$ We reject H0:$\theta = \theta_0$ at 5% level

2 **Wald test**

- Allows simultaneous tests of several restrictions
- It does not require any additional estimation
- General form: $(R\hat{\theta} - a) = 0$
- $W = (\mathbf{R}\hat{\theta} - \mathbf{a})`[\mathbf{R}Est.Asym.var(\hat{\theta})\mathbf{R}`](\mathbf{R}\hat{\theta} - \mathbf{a}) \sim \chi^2_q$

Tests for parameter values

3 **Likelihood ratio test**

- Akin to the F-test in linear regressions
- Requires the estimation of 2 models:
  - a Unrestricted
  - b Restricted or nested
- $LR = 2(ln\hat{L}(ur) - ln\hat{L}(r)) \sim \chi^2_q$

Model diagnostics

Homogeneity test

- Application of LR
- Requires the estimation of J+1 models:
  a. A model in J subsamples
  b. A pooled model
- $LR = 2(\sum_{j=1}^{J} ln\hat{L}(j) - ln\hat{L}(pooled)) \sim \chi^2_{(J-1)*K}$

Goodness of fit $\rightarrow$ Hosmer-Lemeshow

- Create 10 subsamples based on quantiles of $\hat{y} = x'\beta$
- $LM_{HL} = \sum_{d=1}^{10} \frac{n_d(\bar{y}_d - \bar{\bar{y}}_d)^2}{\bar{\bar{y}}_d(1-\bar{\bar{y}}_d)} \sim \chi^2_8$

  where d is the decile number, n_d is the number of observations in d, $\bar{y}_d$ is the empirical mean in d, $\bar{y}_d$ is the predicted mean in d

Model diagnostics: outliers

- **Using residuals**
  - Calculate standardized residuals (r_i^{std}) and check large values
  - Calculate influence $\rightarrow$ Pregiborns' Delta Beta statistic $= f(r_i^{std})$
  - If DBeta statistic is large (> 1) and a spike, re-examine the data

- **Using fitted values**
  - Find surprising predictions
  - Cases with the lowest (highest) $\hat{y}$ given that $y = 1$ ($y = 0$)
  - Do they have something in common?

Pseudo-R^2 measures

Non-linear models do not have a perfect equivalence of R^2

Alternatives based on log-likelihoods

- Deviance $= ln\hat{L}(Saturated) - ln\hat{L}(Current\ model)$.
  (Large values = worse models)
- McFadden's $R^2 = 1 - \frac{ln\hat{L}(current)}{ln\hat{L}(intercept)}$
- Efron's $R^2 = 1 - \frac{\sum_i = 1^N (y_i - \hat{\pi}_i)^2}{\sum_i = 1^N (y_i - \bar{y})^2}$

Classification table

Simple idea: count correct predictions by the model

Define $\hat{y} = I_1(\hat{\pi} > 0.5)$

	$y = 0$	$y = 1$
$\hat{y} = 0$	$n_{0,0}$	$n_{0,1}$
$\hat{y} = 1$	$n_{1,0}$	$n_{1,1}$

Classification table

Simple idea: count correct predictions by the model

Define $\hat{y} = I_1(\hat{\pi} > 0.5)$

	$y = 0$	$y = 1$
$\hat{y} = 0$	$n_{0,0}$	$n_{0,1}$
$\hat{y} = 1$	$n_{1,0}$	$n_{1,1}$

1. **% Right predictions** $\Rightarrow (n_{0,0} + n_{1,1})/N$
2. **Sensitivity** $\Rightarrow (n_{1,1})/(n_{0,1} + n_{1,1})$
3. **Specificity** $\Rightarrow (n_{0,0})/(n_{1,0} + n_{0,0})$

Count R^2

R^2 measures based on the classification table

- Count $R^2 = (n_{0,0} + n_{1,1})/N$

Count R^2

R^2 measures based on the classification table

- Count $R^2 = (n_{0,0} + n_{1,1})/N$
- Problem: we can always get 50%+ correct predictions by setting all y to 1 or 0

Count R^2

R^2 measures based on the classification table

- Count $R^2 = (n_{0,0} + n_{1,1})/N$
- Problem: we can always get 50%+ correct predictions by setting all y to 1 or 0
- Adjusted Count $R^2 = (n_{0,0} + n_{1,1} - n_{max})/(N - n_{max})$

  where $n_max = max(n_0, n_1)$
  - Adjusted count is ≤ 1
  - Adjusted count is > 0 only if + correct predictions than by choosing a single number

ROC curve

- Receiver operating characteristic curve
- A generalized version of the classification table
- Idea: A threshold $= 0.5$ is arbitrary. Instead calculate sensitivity & specificity for all $p \in (0, 1)$
- Then, plot sensitivity and (1-specificity) and calculate the area under the curve
- $\uparrow$ area $\rightarrow$ better models
- Rule of thumb: if area $< .7$ model is not too good.

Exercise: determinants of employment

Using the database ISSP_2014_PL, we will explore the determinants of taking up employment among populaiton in the working age (25-55 years old). We would like to run the model

$$P(work|x) = F(\text{gender, age, education level, married, city, kids})$$

- Estimate using a non-linear probability model
- Use Wald and LR to test significance of # kids parameters. Drop from regression if not significant.
- Test the quality of the model using the classification table and the ROC curve and the Hosmer-Lemeshow test
- Estimate marginal effects for age at the means and the average marginal effect for age
- Interpret the coefficients as changes in the odds-ratio

Homework (I)

We would like to include a dummy variable indicating whether the partner works.

- Add the variable to the model and test for significance using LR-test.
- Estimate discrete changes in probability for those with and without a working partner
- Derive the classification table. Does it offer better predictions than a model without information on the partner?
- Stata does not have a command for the homogeneity test. Try to create a series of commands that allow to run the test.

Homework (II): advanced tasks

Back to the linear probability model (commands not written)

- Run a linear regression and verify whether a) predicted probabilities are on the (0,1) interval; and b) non-normal distribution of residuals.
- Derive the classification table for a cutoff $= 0.5$.
- Run the Hosmer-Lemeshow test.

Summary

- **Probit and logit models offer sensible alternatives to LPM**
  - Address predictions outside of range + $\rightarrow$ flexibility
  - Interpretation of parameters is more cumbersome.
- **Marginal effects and discrete changes**
  - At the mean or at specific values
  - Average marginal effect $\rightarrow$ More popular.
- **Relevance of a model**
  - Several measures based on Log-likelihoods...
  - yet, measures based on correct predictions are more intuitive

For more information:
Winkelmann and Boes (c.2), Greene and Hensher (c.2), Long and Freese (c.4)

Thank you for your attention

Lucas van der Velde

Contact: lvandervelde@grape.org.pl
lvelde@sgh.waw.edu.pl