Econometrics

Binary Choice and Selection Models

Jose Angel Garcia Sanchez

Université Paris 1 Panthéon-Sorbonne
jagarsanc@gmail.com

November 2025

Outline

1. Maximum Likelihood

2. Models of Discrete Choice

3. Sample Selection Models

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Presentation

- An ordinary linear model can be estimated via OLS
- But there exists another method: *Maximum Likelihood*
- We write the *likelihood* of the sample, that we maximize with respect to the parameters we'd like to estimate
- We will then get the optimal parameters
- In the OLS method, we minimized the sum of squared residuals: it is again an optimization process

The model (simple case)

- Let's consider an ordinary linear model $y_t = a + b.x_t + u_t$
- Assume that the error term u_t follows a normal distribution $N(0, \sigma^2)$
- We consider a, b, and x as fixed
- Thus y_t follows a normal distribution as well: $N(a + b.x_t, \sigma^2)$
- The "likelihood" of the sample is the probability that the data in the sample indeed occurred

Estimation

- The distribution of y depends on parameters a and b that are unknown
- If we think that the sample was correctly sampled, then parameters a and b are such that the probability of having sampled this particular data is maximum
- In other words, these parameters should maximize the likelihood of the sample
- This is the ML method: Maximum Likelihood

Computation

- We assume that individuals in the sample are independent
- The likelihood of the sample is thus the product of the likelihood of each individual
- For this reason, we usually take its log to get a sum (easier to handle)
- We thus deal with the "log-likelihood"
- Parameters a and b are obtained by maximizing the log-likelihood, which is the same as maximizing the likelihood since the log function is strictly increasing (bijective)
- In practice: usual maximization techniques, or by iterative methods
- With ML, same results as OLS, with $\hat{\sigma}$ downward biased ($\hat{\sigma}_{ML} = SSR/N$)

Application

- $\forall t,\ y_t \hookrightarrow N(a + b.x_t, \sigma^2)$ so its probability density is:
- $f(y_t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_t - a - b.x_t)^2\right)$
- So the likelihood L of the sample is: $L = \prod f(y_t)$
- The log-likelihood is $LL = log(L) = \sum log(f(y_t))$
- $LL = -\frac{T}{2} ln(2\pi) - T ln\sigma - \frac{1}{2\sigma^2}\sum(y_t - a - b.x_t)^2$
- In the end, we get: $Max_{(a,b)}(LL) \iff Min_{(a,b)}(\sum(y_t - a - b.x_t)^2)$
- In this special case, we get to the same optimization as OLS: OLS and ML estimators are identical for a and b in this case
- Similarly, we find $\hat{\sigma}^2_{ML} = \frac{SSR}{N}$, which is biased downwards
- Indeed, we found earlier that an unbiased estimator was $\hat{\sigma}^2_{OLS} = \frac{SSR}{N-k}$
- However, although biased, it is consistent since in a large sample, N and $N - k$ are almost the same

Maximum Likelihood General Properties

- ML estimators can be biased
- But they are Consistent
- Asymptotically normally distributed
- Asymptotically efficient (smallest variance among all consistent asymptotically normal estimators)
- Caution: the likelihood must be well specified

Tests

No OLS tests any more, since we are not any more in the OLS framework. Assume we want to test a linear restriction on parameter θ: $R\theta = q$. Possibilities, which are asymptotically equivalent are:

- Wald test: estimate parameter θ by ML, then check whether $R\hat{\theta} - q$ is close to zero, using its asymptotic variance-covariance matrix (this procedure is close to tests we are used to)
- Likelihood ratio test: estimate model with and without constraint, and check if the difference between the 2 LL is significantly different from 0
- Lagrange multiplier test: estimate the model with the restriction then check if first order condition from the unrestricted model is significantly violated in that case

Binary Choice

- Example: for individual i, $y_i = 1$ if unemployed, $y_i = 0$ if employed
- We want to explain y by age, education, etc:
- OLS fits a line, but here there is no real scatterplot since y can take only 2 values, 0 and 1
- Plus, many other problems, the main one being that predictions will be out of range
- So preferred method: ML

Discrete Choice Models and OLS (1)

- Example: for individual i, $y_i = 1$ if unemployed, $y_i = 0$ if employed
- We want to explain y by age and education:
- $y_i = \alpha + \beta age_i + \gamma education_i + \varepsilon_i = x_i b + \varepsilon_i$
- So $E(y_i) = x_i b$
- And we know that $E(y_i) = 1 * P(y_i = 1) + 0 * P(y_i = 0)$, so $P(y_i = 1) = x_i b$
- $x_i b$ should lie between 0 and 1, which imposes strong restrictions on x and $b \Rightarrow$ inconvenient

Discrete Choice Models and OLS (2)

- Furthermore, ε_i can take only 2 values: $-x_i b$ with probability $1 - x_i b$, or $1 - x_i b$ with probability $x_i b$
- So that: $V(\varepsilon_i) = x_i b(1 - x_i b)$
- There is heteroskedasticity

Discrete Choice Models and OLS (3)

- Heteroskedasticity could be dealt with using White's standard errors (see lecture on FGLS)
- Non-normality could be dealt with using a large sample size (Central Limit Theorem)
- But the main problem is that we need $E(y_i)$ to belong to [0;1]
- For these various reasons, OLS is inappropriate

Solution

- We do as if there was an unobserved continuous variable y^* such that $\forall i, y_i^* \geq 0 \Rightarrow y_i = 1$ and $y_i^* < 0 \Rightarrow y_i = 0$
- This new variable y^* is continuous: we can thus refer to our usual model, with some adaptations

The Model

- Assume that $y^* = X\beta + u$
- Let F be the probability distribution function (pdf) of u, with a symmetric distribution
- $P(y_i = 1) = P(y^* > 0) = P(X\beta + u > 0) = P(u > -X\beta) = P(u < X\beta) = F(X\beta)$
- Usual functions for F are the normal (*Probit model*) or the logistic (*Logit model*)

Distributions

- Normal: $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
- Logistic: $F(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$
- These 2 are very close, and give similar results
- Expectation: 0 for both distributions
- Variance: 1 for the standard normal, $\pi^2/3$ for the logistic
- Characteristics of the *Logit*: extreme events have higher probability (heavier tails); parameters are interpreted more easily

Log-likelihood (1)

Likelihood function for the entire sample:

$$L = \prod P[y_i = 1|x_i, \beta]^{y_i} P[y_i = 0|x_i, \beta]^{1-y_i}$$

So that when $y = 0$, only the right-hand-side part is there, and when $y = 1$, only the left-hand-side part is there.
And we have $P[y_i = 1|x_i, \beta] = P[y_i^* > 0|x_i, \beta] = F(x_i\beta)$, so

$$LL = \sum y_i log(F(x_i\beta)) + \sum (1 - y_i) log(1 - F(x_i\beta))$$

Maximizing this log-likelihood amounts to differentiating this expression with respect to parameter β and setting it to zero (it can be shown that it is globally concave).

Log-likelihood (2)

$$LL = \sum y_i log(F(x_i\beta)) + \sum (1-y_i) log(1-F(x_i\beta))$$

The first order condition is the following (derivative wrt β is 0): $\frac{dLL}{d\beta} = 0$, which in fact is a column vector of derivatives (derivative of LL wrt β_1, β_2, etc)

$$\frac{dLL}{d\beta} = \sum \left[\frac{y_i - F(x_i\beta)}{F(x_i\beta)(1-F(x_i\beta))} f(x_i\beta) \right] x_i' = 0$$

For the logistic distribution, $F(x) = \frac{e^x}{1+e^x}$ and $f(x) = \frac{e^x}{(1+e^x)^2}$ so:

$$\frac{dLL}{d\beta} = \sum \left(y_i - \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} \right) x_i' = 0$$

The optimal value for β is found with optimization techniques (iterative)

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Remark: Scaling

- Recall that $P(y_i = 1) = P(y^* > 0) = P(X\beta + u > 0) = P(u > -X\beta) = P(u < X\beta) = F(X\beta)$: we then find the optimal estimate for β by maximizing the corresponding likelihood
- But $P(u < X\beta) = F(X\beta)$ only if u is assumed to be a standardized distribution (of variance 1 for the Probit)
- u could in fact have *any* variance σ, in that case one has to write: $P(u < X\beta) = P(u/\sigma < X\beta/\sigma) = F(X\beta/\sigma)$
- Which means that in fact, the estimate we compute is really β/σ, and there is no way to identify the two separately
- So the convention is to set $\sigma = 1$ for the Probit model, and $\pi/\sqrt{3}$ for the Logit model
- Since this is a multiplicative scaling, this doesn't change a thing as for the interpretation of parameters (signs and odds-ratios stay the same), nor for predictions
- The only consequence is that parameters for the Probit and Logit will be scaled differently: $\hat{b}_{logit} \simeq (\pi/\sqrt{3})\hat{b}_{probit}$

Remark: Predicted and Actual Frequency

- $\hat{p}_i = \hat{P}(y_i = 1) = F(x_i\hat{\beta})$
- So $\hat{p}_i = 1/(1 + \exp(-x_i\hat{\beta})) = \exp(x_i\hat{\beta})/(1 + \exp(x_i\hat{\beta}))$
- $\hat{\beta}$ is the solution of the first order condition, so that we get:

$$\sum \left(y_i - \frac{\exp(x_i\hat{\beta})}{1 + \exp(x_i\hat{\beta})} \right) x_i' = 0$$

$$\sum (y_i - \hat{p}_i) x_i' = 0$$

$$\sum \hat{p}_i x_i' = \sum y_i x_i'$$

If there is a constant term in x, then $\sum \hat{p}_i = \sum y_i$: the predicted frequency is exactly equal to the actual frequency. This holds as well for the Probit model (near equality).

Interpreting Parameters (1)

Let x_i be a row vector of observations (corresponding to individual i) and β be a colum vector of parameters.

- $P(y_i = 1) = F(x_i\beta)$
- Parameters enter non-linearly in the expression
- If we differentiate $F(x_i\beta)$ with respect to variable x_k, calling f the derivative of F, we get:

$$\frac{dF(x_i\beta)}{dx_{i,k}} = f(x_i\beta)\beta_k$$

With f either the standard normal density ϕ (*probit*) or logistic density: $\exp(x)/(1+\exp(x))^2$ (*logit*)
The effect of a change in $x_{i,k}$ depends on the values of x_i, so it is different for each person.

Interpreting Parameters (2)

$$\frac{dF(x_i\beta)}{dx_{i,k}} = f(x_i\beta)\beta_k$$

- The *sign* of the effect of a change in $x_{i,k}$ can however be determined
- It is the sign of β_k

Marginal Effects

- We still miss a way to compute the effect on X on y
- We can compute the *marginal effects* of a variable on y
- Consider the average person in the sample and compute her prediction $\hat{y}_0 = F(\bar{X}\hat{\beta})$
- Then increase variable x_1 by one unit, and compute again the prediction, call it $\hat{y}_1$
- The marginal effect of variable x_1 is just $\hat{y}_1 - \hat{y}_0$
- It will tell by how much the probability of getting outcome $[y = 1]$ is increased by increasing x_1 by one unit: e.g. increasing education by one year, switching from male to female (if gender is a dummy), etc
- Always mention that this was computed for the average person
- These can also be computed for specific individuals (e.g. the average 40-year-old white collar worker)

Logit vs. Probit

- Since the two distributions are quite similar, predictions will be similar as well
- However, values of parameters differ: this is because we use different formulas (recall the different scaling)
- We usually get $\hat{b}_{logit} \simeq 1.6\,\hat{b}_{probit}$
- We can find as well: $\hat{b}_{logit} \simeq (\pi/\sqrt{3})\hat{b}_{probit}$
- Again, this does not mean that the impact of explanatory variables is higher with the Logit model

Logit: Odds-ratios

- Parameters β do not have a linear impact: we can compute marginal effects, but the Logit allows more information
- Recall that $P(y_i = 1) = P(y^* > 0) = F(X_i\beta) = \frac{1}{1+e^{-X_i\beta}}$
- I.e. $\forall i, P(Y_i = 1) = p_i$ and $P(Y_i = 0) = 1 - p_i$
- Let's define c_i as $c_i = \frac{p_i}{1-p_i}$
- There are c_i more chances that $[Y_i = 1]$ happens than $[Y_i = 0]$ happens
- We have $p_i = \frac{1}{1+e^{-X_i\beta}} = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$
- And $1 - p_i = \frac{1}{1+e^{X_i\beta}}$
- So $c_i = e^{X_i\beta}$

Odds-ratios (2)

- We thus get $c_i = e^{X_i \beta}$
- But $X_i \beta = \beta_0 + \sum x_{i,j} \beta_j$
- So $c_i = e^{\beta_0 + \sum x_{i,j} \beta_j} = e^{\beta_0} \prod e^{x_{i,j} \beta_j}$
- If variable $x_{i,k}$ increases by 1 unit, c_i is multiplied by e^{β_k}
- If variable x_k increases by 1 unit, it multiplies the *chances* that $[y_i = 1]$ happens by the value of e^{β_k}
- We call *odds-ratio* associated to variable x_k the value e^{β_k}

Goodness of Fit (1)

- Pseudo-R^2: $R^2 = \frac{LL_{fit} - LL_0}{LL_{max} - LL_0}$
- LL_{fit}: Log-likelihood of the model
- LL_0: Log-likelihood of a model with only a constant, where the estimated probability is estimated by the proportion of ones in the sample
- LL_{max}: Maximum Log-likelihood attainable. In our case, 0 because a model that predicts perfectly the observed values has likelihood 1, and $log(1) = 0$.

Goodness of Fit (2)

- We can evaluate the goodness of fit of the model as well by comparing correct and incorrect predictions
- Predicted outcomes: since $0 \leq \hat{p}_i \leq 1$, we need to choose a cut-off point c that will determine if the prediction will be 0 or 1 (usually: $c = 1/2$)
- ROC: Receiver Operating Characteristics curve: plots the fraction of $y = 1$ values correctly classified (true positives) against the fraction of $y = 0$ incorrectly classified (false positives) as c varies
- *Sensitivity* (sensibilité): true positive rate in predictions
- *Specificity* (spécificité): true negative rate in predictions
- ROC: *Sensibility* wrt *(1-Specificity)*

The ROC Curve

- Comes from signal detection theory, developed during WW2
- Accuracy of the classification rule is measured with the area under the ROC curve
- The area represents the ability of the procedure to correctly classify individuals, and is comprised between 0.5 and 1
- Worst classification rule would yield a diagonal ROC curve
- Trade-off between sensitivity and specificity: the more bowed the curve is, the better the model is
- The closer the ROC curve to the upper-left corner, the higher the predictive power
- The optimal cut-off point is determined by the analyst depending on the issue studied

Could Heteroskedasticity be an Issue?

- In Probit and Logit models, homoskedasticity of u is assumed in the latent variable framework (normalization: $\sigma = 1$)
- Recall that in the Probit model, only β/σ can be identified
- Heteroskedasticity means that σ is a function of some variables, say $\sigma^2 = \exp(Z\gamma)$
- Remark 1: it is best not to have a constant in variables Z, because then testing $\gamma = 0$ amounts to testing $\sigma^2 = \exp(0) = 1$ which means homoskedasticity
- Remark 2: variables Z should be exogenous variables not already in the model (see Cameron & Trivedi)
- In this case, we model $P(y_i = 1) = P(y^* > 0) = F(X_i\beta/\sigma_i)$ and plug-in $\sigma_i = \exp(Z_i\gamma)$ in the likelihood
- The *hetprobit* command provides this estimation, along with a test of whether there was indeed heteroskedasticity (test of $\gamma = 0$)

Could Endogeneity be an Issue?

- Just like OLS, Probit and Logit estimates are inconsistent if one or several explanatory variables are endogenous
- The usual latent variable model is written: $y^* = X\beta + x\gamma + u$, with x endogenous
- In that case, given we find enough instruments Z, the endogenous variable should be instrumented: $x = Zb + Xc + v$
- To estimate this now 2-equation model in the maximum likelihood framework, we only need to postulate that (u, v) follow a bivariate normal disribution and plug this into the likelihood
- Testing endogeneity of x amounts to testing whether u and v are uncorrelated
- All this requires both normality and homoskedasticity of u and v, so Probit is required (no Logit here)

About Selection

- Ex: consider the interpretation of average scores over time of an achievement test
- A decline over time may be due to real deterioration in student knowledge, or it may just reflect a selection effect, i.e. more students have been taking the test over time and the new test takers are the relatively weaker students
- Selection can arise from self-selection (individuals choose to participate or not in a particular activity) or sample selection (those who participate in the activity are oversampled)
- These problems are treated alike in Econometrics through *sample selection models*

The Heckit Model (or Generalized Tobit)

- Say a variable y_2^* is partially observed (y_2^* is the latent variable)
- Say we have another latent variable y_1^*, such that y_2^* is observed only if $y_1^* > 0$
- Ex: health expenditures or wages
- y_1 is participation to the labor market: $y_1 = 0$ (does not participate) or $y_1 = 1$ (participates)
- y_2 is observed wage, that cannot be observed if the person does not even belong to the labor market: in that case, she could perhaps get a non zero wage, but her $y_1 = 0$
- We'll see here why the Probit model can be useful for estimation of the 1st step equation
- Note: in the *standard* Tobit model, y_2 is observed only if it is above (or below) a certain threshold

The Model

Latent variables:

- Participation equation: $y_1 = 1$ if $y_1^* > 0$, $y_1 = 0$ otherwise
- Outcome equation: $y_2 = y_2^*$ if $y_1 = 1$, y_2 is missing otherwise

$$\begin{pmatrix} u_1 & u_2 \end{pmatrix} \hookrightarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}\right).$$

The Logit model could not be used here for the 1st equation since the error term is normal: we have to consider the Probit model.

The Likelihood

$$L = \prod [P(y^*_{1,i} \leq 0)]^{1-y_{1,i}} [f(y_{2,i}|y^*_{1,i} > 0) * P(y^*_{1,i} > 0)]^{y_{1,i}}$$

- First term: discrete contribution when $y^*_{1,i} \leq 0$
- Second term: continuous contribution when $y^*_{1,i} > 0$

Conditional Means

$$E(y_2|x, y_1^* > 0) = E(x_2b_2 + u_2|x_1b_1 + u_1 > 0) = x_2b_2 + E(u_2|u_1 > -x_1b_1)$$

In the normal case, it can be shown that:

$$E(y_2|x, y_1^* > 0) = x_2b_2 + \sigma_{1,2}\lambda(x_1b_1)$$

With λ the *inverse Mill's ratio*: $\lambda(z) = \frac{\phi(z)}{\Phi(z)}$

- If u_1 and u_2 are independent, then $E(u_2|u_1 > -x_1b_1) = 0$ and OLS are consistent in the second equation
- Otherwise, they are not because we have to take into account the link between the 2 error terms: the sample selection bias $\sigma_{1,2}\lambda(x_1b_1)$
- Error terms represent unobservable heterogeneity: it is likely that individual heterogeneity has an impact both on the decision to participate in the health care system and on the subsequent health care expenditures
- The sign of $corr(u_1, u_2)$ is the sign of $\sigma_{1,2}$

The Heckman Two-Step Estimator

We have:

$$E(y_2|x, y_1^* > 0) = x_2 b_2 + \sigma_{1,2}\lambda(x_1 b_1)$$

The Heckman two-step estimator suggests to first estimate λ, then to estimate the model replacing λ by its estimate $\hat{\lambda}$

1. Probit regression of y_1 on x_1 (indeed, $P(y_1^* > 0) = \Phi(x_1 b_1)$)
2. Compute $\hat{\lambda} = \frac{\phi(x_1 \hat{b}_1)}{\Phi(x_1 \hat{b}_1)}$
3. Estimate the following regression: $E(y_2|x, y_1^* > 0) = x_2 b_2 + \sigma_{1,2}\hat{\lambda}$
4. The sign of the coefficient corresponding to $\hat{\lambda}$ will give the sign of the correlation between the 2 error terms
5. And if this coefficient is not significant, the 2 equations can be considered to be independent

Remarks

- Two-step Heckman estimates are consistent
- This method is easy and fast to implement
- However, there is a efficiency loss (not appropriate in small samples)
- A more widely used method is Maximum Likelihood on the full model
- Notice that the use of a Logit would not be appropriate here

Identification Considerations

- Theoretically, the exact same regressors can appear in both equations because the second set of equations appear in a non linear way: no strict multicollinearity
- However, in that case, the model is close to unidentified because the inverse Mill's ratio is almost linear
- This leads to a great instability of parameters (see slides on near-perfect multicollinearity)
- Thus, at least one exclusion restriction is usually required: we need that at least one variable in the first equation (selection) be absent from the second equation (conditional outcome)

Remarks

- If we find out that there is in fact no correlation between the 2 equations (check ρ), they can be estimated independently through a Two-Part model (see next slide)
- If there is no correlation: selection is said to be made on observables, because it is fully explained by regressors
- If there is correlation: selection is said to be made on unobservables, because it is partly explained by error terms, that comprise all the unobservable information
- The Two-Part model can be used as well if the correlation of the inverse Mill's ratio with regressors is too strong (near-to-perfect multicollinearity)

The Two-Part Model

This model intentionally does not take into account potential correlation between the error terms of the 2 equations

1. Participation equation: Probit or Logit that gives the probability to participate
2. Conditional outcome equation: anything we like (linear model, count data etc) that gives the level of activity, given that the individual participates (warning: need distribution truncated at 0, such as ztp or ztnb); so more flexible than the Generalized Tobit framework

And the prediction of the model is the product of the two. Notice that we lose assessment of selection effect but gain in flexibility (no need to stick to Normal models)