Vrije Universiteit Amsterdam
Econometrics II (E_EOR2_TR2)

BSc EOR and EDS, Periods 4 and 5, 2025

Assignment 1 (Period 4): binary and multinomial data

Part A: Binary data

Suppose we have a latent variable y_i^* with

$$y_i^* \;=\; x_i'\beta + e_i \;=\; \beta_0 + \beta_1 x_{i1} + e_i,$$

where e_i is an error term with a certain continuous distribution (independent of x_i) with cumulative distribution function $G(.)$: $\Pr(e_i \leq a) = G(a)$.

$$y_i = I\{y_i^* > 0\} = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Do **not** assume that the distribution is symmetric around 0 (i.e., do **not** assume that $G(a) = 1 - G(-a)$).

(1a) Derive that we have probabilities $\Pr(y_i = 1|x_i) = 1 - G(-x_i'\beta)$ and $\Pr(y_i = 0|x_i) = G(-x_i'\beta)$.

(1b) Further assume that the n observations are independent. Derive the loglikelihood $\ln p(y_1, \ldots, y_n|x_1, \ldots, x_n)$.

As an alternative to the binary probit and logit models, one can make use of the so-called *Gompit model* (also called the *complementary log-log model* or *cloglog model*). In this model one assumes that the error terms e_i have a Gumbel distribution with $G(a) = \exp(-\exp(-a))$.[1] This is a distribution with positive skewness, where large positive values are more likely than large negative values. The Gompit model can be better than the probit/logit model, if we have an *asymmetric* situation where very low values of $x_i'\beta$ can still yield $y_i = 1$ (because e_i can have a huge positive value), whereas very high values of $x_i'\beta$ can not yield $y_i = 0$ (because e_i can not have a huge negative value). Roughly stated, there can be 'positive surprises', but not 'negative surprises'.

We can also switch how we define $y_i = 1$ and $y_i = 0$ (that is, define $\tilde{y}_i = 1 - y_i$, where $\tilde{y}_i = 0$ means 'yes' and $\tilde{y}_i = 1$ means 'no'), and estimate a Gompit model with dependent variable $\tilde{y}_i$. This can be useful if we have the opposite situation where 'negative surprises' are possible, but 'positive surprises' are not. This is equivalent with estimating a so-called *log-log model* for the original dependent variable y_i, where we have $G(a) = 1 - \exp(-\exp(a))$.[2]

[1] The Gumbel distribution is another name for the extreme value distribution and it is also sometimes named the Gompertz distribution, although the Gompertz distribution typically refers to a different distribution. In the Gompit model we have $\Pr(y_i = 1|x_i) = 1 - \exp(-\exp(x_i'\beta)))$.

[2] In the log-log model we have $\Pr(y_i = 1|x_i) = 1 - (1 - \exp(-\exp(-x_i'\beta))) = \exp(-\exp(-x_i'\beta))$, where the name *log-log* stems from the fact that $x_i'\beta = -\ln(-\ln(\Pr(y_i = 1|x_i)))$.

The probit and logit models are *symmetric* in the sense that in the logit model both 'positive surprises' and 'negative surprises' are possible (with huge positive or negative e_i), whereas in the probit model there are no 'positive surprises' or 'negative surprises' (i.e., there are no huge positive or negative e_i). Another symmetric model is the Gosset[3] model in which

$$G(a) = \int_{-\infty}^{a} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}} dz$$

is the CDF of the Student-t distribution with unknown degrees of freedom parameter ν (where $\nu > 0$ does not need to be an integer), where a small value of ν indicates that 'positive surprises' and 'negative surprises' are possible (with huge positive or negative e_i), whereas for $\nu \to \infty$ the Gosset model tends to the probit model, because for $\nu \to \infty$ the Student-t distribution tends to the normal distribution.

(1c) The file `BinaryDataXX.xlsx` (where `XX` is your group number) contains $n = 1500$ observations of binary dependent variable y_i and explanatory variable x_{i1}. This is a simulated dataset, but you can interpret this as a dataset of $n = 1500$ employees at a certain company where $y_i = 1$ if person i has a management job, $y_i = 0$ if person i does not have a management job, and x_{i1} is the number of years of education of person i.

Estimate the binary logit, probit, Gompit (cloglog), log-log and Gosset models using the method of maximum likelihood, where for each model the optimal value of β needs to be found and where for the Gosset model also the optimal value of ν (with $\nu > 0$) needs to be found. Compute the loglikelihood, Akaike Information Criterion (AIC), Schwarz Criterion (SC) and percentage correctly predicted (with threshold 0.5) for these 5 models. Which model has the best performance? Which model has the worst performance? Do these findings suggest the presence of 'positive surprises' and/or 'negative surprises'?
For your chosen best model, show the graph of the estimated probability $\widehat{\Pr}(y_i = 1|x_{i1})$, compute the partial effect at the average (PEA) and the average partial effect (APE), interpret the PEA and APE, and explain the difference between the PEA and APE.

(1d) For the symmetric models (probit, logit and Gosset) we have $G(a) = 1 - G(-a)$, so that

$$\begin{aligned} \Pr(y_i = 1|x_i) &= 1 - G(-x_i'\beta) = G(x_i'\beta) \\ \Pr(y_i = 0|x_i) &= G(-x_i'\beta) = 1 - G(x_i'\beta) \end{aligned}$$

Define $\tilde{y}_i = 1 - y_i$, where $\tilde{y}_i = 0$ means 'yes' and $\tilde{y}_i = 1$ means 'no', so that we have

$$\begin{aligned} \Pr(\tilde{y}_i = 1|x_i) &= \Pr(y_i = 0|x_i) = 1 - G(-x_i' \cdot (-\beta)) \\ \Pr(\tilde{y}_i = 0|x_i) &= \Pr(y_i = 1|x_i) = G(-x_i' \cdot (-\beta)) \end{aligned}$$

which is described by the same model (probit, logit or Gosset) with coefficients $-\beta$ instead of β. Therefore the loglikelihood, AIC and SC are the same in the symmetric models with y_i and $\tilde{y}_i$ (and also the percentage correctly predicted, unless one or more of the estimated probabilities are exactly equal to the threshold, but that typically does not occur), so that we do not need to consider symmetric models for dependent variable $\tilde{y}_i$.
Estimate a logit model for dependent variable $\tilde{y}_i = 1 - y_i$ (using the method of maximum likelihood) and check that indeed the estimated coefficients are (approximately) $(-1)\times$ the estimated coefficients in the model for dependent variable y_i, and that the loglikelihood, AIC, SC and percentage correctly predicted are (approximately) the same.

[3] 'Student' was the pseudonym of William Sealy Gosset who wrote one of the first articles about the Student-t distribution.

Part B: Multinomial data

The file `MultinomialDataXX.xlsx` (where `XX` is your group number) contains $n = 1500$ observations of dependent variable y_i (with $y_i = 0$, $y_i = 1$ or $y_i = 2$) and explanatory variable x_{i1}. This is a simulated dataset, but you can interpret this as a dataset of $n = 1500$ employees at a different company with two departments A and B, where $y_i = 0$ if person i has an administrative job at department A, $y_i = 1$ if person i has an administrative job at department B, $y_i = 2$ if person i has a management job, and x_{i1} is the number of years of education of person i.

(1e) Estimate an ordered logit model for this dataset, using the method of maximum likelihood. For what value of x_{i1} do you have $\widehat{\Pr}(y_i = 2|x_{i1}) = \frac{1}{2}$ and $\widehat{\Pr}(y_i = 0 \text{ or } 1|x_{i1}) = \frac{1}{2}$? Show the graphs of the estimated probabilities $\widehat{\Pr}(y_i = 0|x_{i1})$, $\widehat{\Pr}(y_i = 1|x_{i1})$ and $\widehat{\Pr}(y_i = 2|x_{i1})$. Explain why the estimates of the parameters match with the graphs of these estimated probabilities.

(1f) For the ordered logit model we have $G(a) = 1 - G(-a)$, so that

$$
\begin{aligned}
\Pr(y_i = 0|x_i) &= G(\tau_1 - x_i'\beta) = 1 - G(-\tau_1 + x_i'\beta) \\
\Pr(y_i = 1|x_i) &= G(\tau_2 - x_i'\beta) - G(\tau_1 - x_i'\beta) \\
&= [1 - G(-\tau_2 + x_i'\beta)] - [1 - G(-\tau_1 + x_i'\beta)] \\
&= G(-\tau_1 + x_i'\beta) - G(-\tau_2 + x_i'\beta) \\
\Pr(y_i = 2|x_i) &= 1 - G(\tau_2 - x_i'\beta) \\
&= 1 - [1 - G(-\tau_2 + x_i'\beta)] \\
&= G(-\tau_2 + x_i'\beta)
\end{aligned}
$$

Define $\tilde{y}_i = 2 - y_i$, where we have the opposite ordering with $\tilde{y}_i = 0$ meaning 'high outcome', $\tilde{y}_i = 1$ meaning 'middle outcome' and $\tilde{y}_i = 2$ meaning 'low outcome', so that we have

$$
\begin{aligned}
\Pr(\tilde{y}_i = 0|x_i) &= \Pr(y_i = 2|x_i) = G(-\tau_2 - x_i' \cdot (-\beta)) \\
\Pr(\tilde{y}_i = 1|x_i) &= \Pr(y_i = 1|x_i) = G(-\tau_1 - x_i' \cdot (-\beta)) - G(-\tau_2 - x_i' \cdot (-\beta)) \\
\Pr(\tilde{y}_i = 2|x_i) &= \Pr(y_i = 0|x_i) = 1 - G(-\tau_1 - x_i' \cdot (-\beta))
\end{aligned}
$$

which is an ordered logit model with coefficients $-\beta$ (instead of β) and thresholds $-\tau_2$ and $-\tau_1$ (instead of τ_1 and τ_2). Therefore the loglikelihood, AIC and SC are the same in the ordered logit models with y_i and $\tilde{y}_i$ (and also the percentage correctly predicted, unless there are observations that have two or three equal estimated probabilities, but that typically does not occur), so that we do not need to consider the ordered logit model for dependent variable $\tilde{y}_i$.
Estimate an ordered logit model for dependent variable $\tilde{y}_i = 2 - y_i$ (using the method of maximum likelihood) and check that indeed the estimated coefficients are (approximately) $(-1)\times$ the estimated coefficients in the model for dependent variable y_i, and that the loglikelihood is (approximately) the same.

(1g) Estimate a multinomial logit model for this dataset, using the method of maximum likelihood. Show the graphs of the estimated probabilities $\widehat{\Pr}(y_i = 0|x_{i1})$, $\widehat{\Pr}(y_i = 1|x_{i1})$ and $\widehat{\Pr}(y_i = 2|x_{i1})$. Explain why the estimates of the parameters match with the graphs of these estimated probabilities.

Consider the alternative dependent variables y_i' and y_i'' (where the ordering of the three job categories is different, each time with a different 'middle category'):

$$\tilde{y}_i' = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i = 2 \\ 2 & \text{if } y_i = 1 \end{cases} \qquad \text{and} \qquad y_i'' = \begin{cases} 0 & \text{if } y_i = 1 \\ 1 & \text{if } y_i = 0 \\ 2 & \text{if } y_i = 2 \end{cases}$$

(1h) Consider the following four models:

- the ordered logit model with dependent variable y_i
- the ordered logit model with dependent variable y_i'
- the ordered logit model with dependent variable y_i''
- the multinomial logit model with dependent variable y_i.

Based on the AIC, which model is the best and which model is the worst of these four models? Explain why these models are the best and worst models in this case.

(1i) In the multinomial logit model we have odds ratios

$$\begin{aligned} \frac{\Pr(y_i = 1|x_i)}{\Pr(y_i = 0|x_i)} &= \exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1}) \\ \frac{\Pr(y_i = 2|x_i)}{\Pr(y_i = 0|x_i)} &= \exp(\beta_0^{(2)} + \beta_1^{(2)} x_{i1}) \end{aligned}$$

For the alternative dependent variable y_i'' this implies

$$\begin{aligned} \frac{\Pr(y_i'' = 1|x_i)}{\Pr(y_i'' = 0|x_i)} &= \frac{\Pr(y_i = 0|x_i)}{\Pr(y_i = 1|x_i)} = \exp(-\beta_0^{(1)} - \beta_1^{(1)} x_{i1}) \\ \frac{\Pr(y_i'' = 2|x_i)}{\Pr(y_i'' = 0|x_i)} &= \frac{\Pr(y_i = 2|x_i)}{\Pr(y_i = 1|x_i)} = \exp((\beta_0^{(2)} - \beta_0^{(1)}) + (\beta_0^{(2)} - \beta_0^{(1)}) x_{i1}) \end{aligned}$$

which is a multinomial logit model with coefficients $-\beta_0^{(1)}$, $-\beta_1^{(1)}$, $\beta_0^{(2)} - \beta_0^{(1)}$ and $\beta_0^{(2)} - \beta_0^{(1)}$. Therefore the loglikelihood, AIC and SC are the same in the multinomial logit models with y_i and y_i'' (and also the percentage correctly predicted, unless there are observations that have two or three equal estimated probabilities, but that typically does not occur), so that we do not need to consider the multinomial logit model for dependent variable y_i'' (or any other ordering that is different from y_i).
Estimate a multinomial logit model for dependent variable y_i'' (using the method of maximum likelihood) and check that indeed the estimated coefficients have (approximately) the aforementioned relationship with the estimated coefficients in the multinomial logit model for dependent variable y_i, and that the loglikelihood is (approximately) the same.