Endogenous switching or sample selection models for count data

July 9, 2025

Introduction

In two seminal papers, Heckman (1976) and Heckman (1979) considered the case where two variable are jointly determined: a binomial variable and an outcome continuous variable. For example, the continuous variable can be the wage and the binomial variable the labor force participation. In this case, the wage is observed only for the sub-sample of the individuals who works. If the unobserved determinants of labor force participation are correlated with the unobserved determinants of wage, estimating the wage equation only on the subset of individuals who work will result in an inconsistent estimator. This case is called the **sample selection** model.

An other example is the case where the binomial variable is a a dummy for private vs public sector. In this case the wage is observed for the whole sample (for the individuals in the public and in the private sector). But, once again, if the unobserved determinants of the chosen sector are correlated with those of the wage, estimating the wage equation only will leads to inconsistent estimators. This case is called the **endogenous switching** model.

Two consistent methods of estimation can be used in this context:

- the first one is a two-step method where, in a first step, a probit model for the binomial variable is estimated and, in a second step, the outcome equation is estimated by OLS with a supplementary covariate which is a function of the linear predictor of the probit,[1]
- the second one is the maximum likelihood estimation of the system of the two equations, assuming that the error terms are jointly normally distributed.

---

[1]More precisely the inverse Mills ratio.

Sample selection and endogenous switching for count data

Let y be a count response (for the sake of simplicity a Poisson variable) and d a binomial variable. The value of d is given by the sign of $\alpha^\top z + \nu$, where ν is a standard normal deviate, z a vector of covariates and α the associated vector of unknown parameters. The distribution of y_n is Poisson with parameter λ_n, given by $\ln \lambda_n = \beta^\top x_n + \epsilon_n$ where x is a second set of covariates (which can overlap with z), β is the corresponding set of unknown parameters and ϵ is a random normal deviate with 0 mean and a standard deviation equal to σ. ϵ and ν being potentially correlated, their joint distribution has to be considered:

$$
\begin{aligned}
f(\epsilon, \nu; \sigma, \rho) &= \frac{1}{2\pi\sqrt{1-\rho^2}\sigma} e^{-\frac{1}{2}\frac{\left(\frac{\epsilon}{\sigma}\right)^2 + \nu^2 - 2\rho\left(\frac{\epsilon}{\sigma}\right)\nu}{1-\rho^2}} \\
&= \frac{1}{\sqrt{2\pi}}\sigma e^{-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{(\nu-\rho\epsilon/\sigma)}{\sqrt{1-\rho^2}}\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
f(\epsilon, \nu; \sigma, \rho) &= \frac{1}{2\pi\sqrt{1-\rho^2}\sigma} e^{-\frac{1}{2}\frac{\left(\frac{\epsilon}{\sigma}\right)^2 + \nu^2 - 2\rho\left(\frac{\epsilon}{\sigma}\right)\nu}{1-\rho^2}} \\
&= \frac{1}{\sqrt{2\pi}}\sigma e^{-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{(\nu-\rho\epsilon/\sigma)}{\sqrt{1-\rho^2}}\right)^2}
\end{aligned}
$$

The second expression classically gives the joint distribution as the product of the marginal distribution of ϵ and the conditional distribution of ν. For $d = 1$, the unconditional distribution of y is obtained by integrating out $g(y_n \mid x_n, \epsilon_n, \nu_n, d_n = 1)$ with respect with the two random deviates:

$$
P(y_n \mid x_n, d_n = 1) = \int_{-\infty}^{+\infty} \int_{-\alpha^\top z_n}^{+\infty} g(y_n \mid x_n, \epsilon_n, d_n = 1) f(\epsilon, \nu) d\epsilon d\nu
$$

$$
\begin{aligned}
P(y_n \mid x_n, d_n = 1) &= \int_{-\infty}^{+\infty} g(y_n \mid x_n, \epsilon_n, d_n = 1) \left(\int_{-\alpha^\top z_n}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{(\nu-\rho\epsilon/\sigma)}{\sqrt{1-\rho^2}}\right)^2} d\nu \right) \\
&\times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2} d\epsilon
\end{aligned}
$$

By symmetry of the normal distribution, the term in bracket is:

$$\Phi\left(\frac{\alpha^\top z_n + \rho/\sigma\epsilon}{\sqrt{1-\rho^2}}\right)$$

which is the probability that $d=1$ for a given value of ϵ. The density of y given that $d=1$ is then:

$$P(y_n \mid x_n, d_n = 1) = \int_{-\infty}^{+\infty} \frac{e^{-\exp(\beta^\top x_n + \epsilon_n)} e^{y_n(\beta^\top x_n + \epsilon_n)}}{y_n!} \Phi\left(\frac{\alpha^\top z_n + \rho/\sigma\epsilon}{\sqrt{1-\rho^2}}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2} d\epsilon \tag{1}$$

By symmetry, it is easily shown that $P(y_n \mid x_n, d_n = 0)$ is similar except that $\Phi\left((\alpha^\top z_n + \rho/\sigma\epsilon)/\sqrt{1-\rho^2}\right)$ is replaced by $1-\Phi\left((\alpha^\top z_n + \rho/\sigma\epsilon)/\sqrt{1-\rho^2}\right)$ or $\Phi\left(-(\alpha^\top z_n + \rho/\sigma\epsilon)/\sqrt{1-\rho^2}\right)$, so that a general formulation of the distribution of y is, denoting $q = 2d - 1$:

$$P(y_n \mid x_n) = \int_{-\infty}^{+\infty} \frac{e^{-\exp(\beta^\top x_n + \epsilon_n)} e^{y_n(\beta^\top x_n + \epsilon_n)}}{y_n!} \Phi\left(q_n \frac{\alpha^\top z_n + \rho/\sigma\epsilon}{\sqrt{1-\rho^2}}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2} d\epsilon \tag{2}$$

There is no closed form for this integral but, using the change of variable $\eta = \epsilon/\sqrt{2}/\sigma$, we get:

$$P(y_n \mid x_n) = \int_{-\infty}^{+\infty} \frac{e^{-\exp(\beta^\top x_n + \sqrt{2}\sigma\eta)} e^{y_n(\beta^\top x_n + \sqrt{2}\sigma\eta)}}{y_n!} \Phi\left(q_n \frac{\alpha^\top z_n + \sqrt{2}\rho\eta}{\sqrt{1-\rho^2}}\right) \frac{1}{\sqrt{\pi}} e^{-\eta^2} d\eta$$

which can be approximated using Gauss-Hermite quadrature. Denting η_r the nodes and ω_r the weights:

$$P(y_n \mid x_n) \approx \sum_{r=1}^{R} \omega_r \frac{e^{-\exp(\beta^\top x_n + \sqrt{2}\sigma\eta_r)} e^{y_n(\beta^\top x_n + \sqrt{2}\sigma\eta_r)}}{y_n!} \Phi\left(q_n \frac{\alpha^\top z_n + \sqrt{2}\rho\eta_r}{\sqrt{1-\rho^2}}\right) \frac{1}{\sqrt{\pi}} e^{-\eta_r^2}$$

For the exogenous switching model, the contribution of one observation to the likelihood is given by Equation 2. For the sample selection model, the contribution of one observation to the likelihood is given by Equation 1 if $d_n = 1$. If $d_n = 0$, y is unobserved and the contribution of such observations to the likelihood is the probability that $z_n = 0$, which is:

$$P(d_n = 0 \mid x_n) = \int_{-\infty}^{+\infty} \Phi\left(q_n \frac{\alpha^\top z_n + \rho/\sigma\epsilon}{\sqrt{1-\rho^2}}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2} d\epsilon$$

The **ML** estimator is computing intensive as the integral has no closed form. One alternative is to use non-linear least squares, by first computing the expectation of y. Terza (1998) showed that, in the endogenous switching case:

$$\text{E}(y_n \mid x_n) = \exp\left(\beta^\top x_n + \ln \frac{\Phi\left(q_n(\alpha^\top z_n + \theta)\right)}{\Phi\left(q_n(\alpha^\top z_n)\right)}\right)$$

For the sample selection case, we have:

$$\text{E}(y_n \mid x_n, z_n = 1) = \exp\left(\beta^\top x_n + \ln \frac{\Phi(\alpha^\top z_n + \theta)}{\Phi(\alpha^\top z_n)}\right)$$

Greene (2001) noted that a first order Taylor series of $\ln \frac{\Phi(\alpha^\top z_n + \theta)}{\Phi(\alpha^\top z_n)}$ around $\theta = 0$ gives: $\theta\phi(\alpha^\top z_n)/\Phi(\alpha^\top z_n)$, which is the inverse mills ratio that is used in the linear case in order to correct the inconsistency due to sample selection. As α can be consistently estimated by a probit model, the **NLS** estimator is obtained by minimizing with respect to β and θ the sum of squares of the following residuals:

$$y_n - e^{\beta^\top x_n + \ln \frac{\Phi(\hat{\alpha}^\top z_n + \theta)}{\Phi(\hat{\alpha}^\top z_n)}}$$

As it is customary for two-step estimators, the covariance matrix of the estimators should take into account the fact that α has been estimated in the first step. Moreover, only $\theta = \rho\sigma$ is estimated. To retrieve an estimator of σ, Terza (1998) proposed to insert into the log-likelihood function the estimated values of α, β and θ and then to maximize it with respect to σ^2.[2]

The `escount` function

```
library(micsr)
```

The `escount` function estimates the endogenous switching and the sample selection model for count data. The first one is obtained by setting the `model` argument to `'es'` (the default) and the second one to `'ss'`. The estimation method is selected using the `method` argument, which can be either `'twostep'` for the two-step non-linear least squares model (the default) or `'ML'` for maximum likelihood. The model is described by an extended formula (using the **Formula** package) of the form:

```
y + d ~ x + y + z + d | x + y + w
```

---

[2]These once again requires to use Gauss-Hermite quadrature, but the problem is considerably simpler as the likelihood is maximized with respect with only one parameter.

which indicates that the two responses are `y` (the count) and `d` (the binomial variable), that the covariates are `x`, `y` and `z` for the count equation and `x`, `y` and `w` for the switching/selection equation. When there are two large sets of covariates that overlap, it is possible to define the second set of covariates by updating the first one:

```
y + d ~ x + y + z + d | . - d - z + w
```

`R` is an integer that indicate the number of points used for the Gauss-Hermite quadrature method. Relevant only for the **ML** method, the `hessian` argument is a boolean: if `TRUE`, the covariance matrix of the coefficients is estimated using the numerical hessian, which is computed using the `hessian` function of the **numDeriv** package, otherwise, the outer product of the gradient is used. `escount` returns an object of class `escount` which inherits from `lm`.

Trips demand

Terza (1998) analyzed the number of trips taken by 577 individuals in the United States in 1978 the day before they were interviewed. The `trips` data set is included in the **micsr** package. A major determinant of trips demand is the availability of a car in the household. Terza (1998) advocates that the unobserved determinants of the decision of having a car may be correlated with those of the trip demand equation. In this case the estimation of the Poisson model will lead to inconsistent estimators. The covariates are the share of trips for work or school (`workschl`), the number of individuals in the household (`size`), the distance to the central business district (`dist`) a factor (`smsa`) with two levels for small and large urban area, the number of full-time worker in household (`fulltime`), the distance from home to nearest transit node, household income divided by the median income of the census tract (`realinc`), a dummy if the survey period is Saturday or Sunday (`weeekend`) and a dummy for owning at least a car (`car`). Although the coefficients are identified, as in the classic Heckman's model by the non-linearity of the correction term, Terza (1998) use a different set of covariates for the binomial and the count parts of the model. Namely, the `weekend` covariate is removed and `adults` is added in the binomial part of the model.

We first compute the two-step **NLS** estimator. The `model` and `method` arguments needn't to be set are the default values are `es` (endogenous switching) and `twosteps` (two-step **NLS**).

```
trips_2s <- escount(trips + car ~ workschl + size + dist + smsa +
                      fulltime + distnod + realinc + weekend +
                      car | . - car - weekend + adults,
                  data = trips)
names(trips_2s)
```

```
[1] "coefficients"  "sigma"          "rho"          "vcov"
[5] "residuals"     "fitted.values" "model"        "terms"
```

5

```
 [9] "value"          "npar"          "df.residual"   "xlevels"
[13] "na.action"      "call"          "first"         "est_method"
```

The result is a list with usual items, except `sigma` and `rho` which are the estimates of σ and ρ obtained from the estimation of θ. The print of the summary method returns the usual table of coefficients, the value of the objective function (the sum of squares residuals) and the estimated values of σ and ρ:

```
summary(trips_2s)
```

```
Two-steps
               Estimate  Std. Error z-value Pr(>|z|)
(Intercept)  -1.4455570  27.2358377 -0.0531   0.9577
workschl     -0.5543755  12.0877257 -0.0459   0.9634
size          0.1487720 113.9834105  0.0013   0.9990
dist         -0.0089621 328.3588966  0.0000   1.0000
smsa         -0.0088507  15.4127875 -0.0006   0.9995
fulltime      0.1059545  45.3698679  0.0023   0.9981
distnod       0.0043303 590.6098360  0.0000   1.0000
realinc       0.0066861 113.0591962  0.0001   1.0000
weekend      -0.1650841   9.9565303 -0.0166   0.9868
car           2.7957446  26.9980427  0.1036   0.9175
theta        -0.5662482  10.6388889 -0.0532   0.9576

NA: NULL
```

```
trips_pois <- glm(trips ~ workschl + size + dist + smsa + fulltime +
                      distnod + realinc + weekend + car,
                  data = trips, family = poisson)
trips_ml <- update(trips_2s, method = "ml")
```

The coefficient of `car` is equal to 1.413 for the Poisson model, which implies that the number of trips increases by $e^{1.413} - 1 = 311\%$ for individuals who belongs to households that own at least one car. The coefficient is much higher in the selection models (2.796 for the 2-step estimator and 2.160 for the **ML** estimator), which implies an increase of 767% for the **ML** estimator. The Poisson estimator is therefore downward biased, which indicates a negative correlation between the unobserved part of the trips demand equation and the propensity to own a car equation.

Physician advice and alcohol consumption

Kenkel and Terza (2001) investigate the effect of physician's advice on alcohol consumption. The outcome variable `drinks` is the number of drinks in the past 2 weeks and the selection variable `advice` is a dummy based on the respondents' answer to the question "Have you ever been told by a physician to drink less". The unobserved part of the equation indicating the propensity to receive an advice form the physician can obviously be correlated with the one of the alcohol consumption equation. The data set `drinks` is part of the **micsr** package. The covariates are monthly income in thousands of dollars (`income`), `age` (a factor with six 10 years categories of age), education in years (`educ`), `race` (a factor with levels `white`, `black` and `other`), the marital status (`marital`, a factor with levels `single`, `married`, `widow`, `separated`), the employment status (a factor `empstatus` with levels `other`, `emp` and `unemp`) and the region (`region`, a factor with levels `west`, `northeast`, `midwest` and `south`). For the binomial part of the model, the same covariates are used (except of course `advice`) and 10 supplementary covariates indicating the insurance coverage and the health status are added.

```
kt_pois <- glm(drinks ~ advice + income + age + educ + race + marital +
                   empstatus + region, data = drinks, family = poisson)
kt_ml <- escount(drinks + advice ~ advice + income + age + educ + race +
                    marital + empstatus + region | income + age + educ +
                    race + marital + empstatus + region + medicare +
                    medicaid + champus + hlthins + regmed + dri + limits +
                    diabete + hearthcond + stroke,
                data = drinks, method = "ml")
kt_2s <- update(kt_ml, method = "twostep")
```

The coefficient of `advice` in the alcohol demand equation is positive in the Poisson model, which would imply a positive effect of physical advice on alcohol consumption. The estimation of the endogenous switching model shows that this positive coefficient is due to the positive correlation between the error terms of the two equations (the unobserved propensities to drink and to receive an advice from a physician are positively correlated).

References

Greene, William H. 2001. "Fiml Estimation of Sample Selection Models for Count Data." In *Economic Theory, Dynamics and Markets: Essays in Honor of Ryuzo Sato*, edited by Takashi Negishi, Rama V. Ramachandran, and Kazuo Mino, 73–91. Boston, MA: Springer US.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement, Volume 5, Number 4*, 475–92.

———. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61.

Kenkel, Donald S., and Joseph V. Terza. 2001. "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect." *Journal of Applied Econometrics* 16 (2): 165–84.

Terza, Joseph V. 1998. "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects." *Journal of Econometrics* 84 (1): 129–54.