Exponentional conditional mean models with endogeneity

June 8, 2025

The linear conditional mean model

The linear model $y_n = \beta^\top x_n + \epsilon_n$, with $\mathrm{E}(\epsilon \mid x) = 0$ implies a linear conditional mean function: $\mathrm{E}(y \mid x) = \beta^\top x$. It implies the following moment conditions:

$$\mathrm{E}\left(y - \beta^\top x\right) = 0$$

which can be estimated on a given sample by:

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - \beta^\top)x_n = X^\top(y_n - \beta^\top)/N$$

Solving this vector of K empirical moments for β leads to the OLS estimator.

If some of the covariates are endogenous, a consistent estimator can be obtained if there is a set of $L \geq K$ series exogenous Z, some of them being potentially elements of X. Then, the L moment conditions can be estimated by:

$$\bar{m} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \beta^\top)z_n = Z^\top\epsilon/N$$

The variance of $\sqrt{N}\bar{m}$ is $\Omega = N\mathrm{E}\left(\bar{m}\bar{m}^\top\right) = \frac{1}{N}\mathrm{E}\left(Z^\top\epsilon\epsilon^\top Z\right)$, which reduce to $\sigma_\epsilon^2 Z^\top Z$ if the errors are iid and, more generaly, can be consistently estimated by $\frac{1}{N}\sum_{n=1}^{N}\hat{\epsilon}_n^2 z_n z_n^\top$ where $\hat{\epsilon}$ are the residuals of a consistent estimation.

The IV estimator minimise the quadratic form of the moments with the inverse of its variance assuming that the errors are iid:

$$\epsilon^\top Z \left(\sigma^2 Z^\top Z\right)^{-1} Z^\top \epsilon = \epsilon^\top P_Z \epsilon / \sigma^2$$

which is the IV estimator. As $P_Z X$ is the projection of the column of X on the subspace generated by the columns of Z, this estimator can be performed by first regressing every covariates on the set of instruments and then regressing the response on these fitted values (2SLS).

The IV estimator is consistent but inefficient if the errors are not iid. In this case, a more efficient estimator can be obtained by minimizing $\bar{m}^\top \hat{\Omega}^{-1} \bar{m}$ with:

$$\hat{\Omega} = \frac{1}{N} \sum_{n=1}^{N} \hat{\epsilon}_n^2 z_n z_n^\top$$

where $\hat{\epsilon}$ can be the residuals of the IV estimator. This is the GMM estimator.

The exponential linear conditional mean model

The linear model is often inappropriate if the conditional distribution of y is asymetric. In this case, a common solution is to use $\ln y$ instead of y as the response.

$$\ln y_n = \beta^\top x_n + \epsilon$$

This is of course possible only if $y_n > 0 \forall n$. An alternative is to use an exponential linear conditional mean model, with additive:

$$y_n = e^{\beta^\top x_n} + \epsilon_n$$

or with multiplicative errors:

$$y_n = e^{\beta^\top x_n} \nu_n$$

If all the covariates are exogenous, $\mathrm{E}\left(y - e^{\beta^\top x} \mid x\right) = 0$ which corresponds to the following empirical moments:

$$X^\top \left(y - e^{\beta^\top x}\right) / N = 0$$

This define a non-linear system of K equations with K unknown parameters (β) that is in particular used when fitting a Poisson model with a log link for count data. It can also be used with any non-negative response.

If some of the covariates are endogenous, as previously an IV estimator can be defined. For additive errors, the empirical moments are:

$$\frac{1}{N}\sum_{n=1}^{N}(y_n - e^{\beta^\top x_n})z_n = Z^\top(y_n - e^{\beta^\top x_n})/N$$

and the IV estimator minimize:

$$\epsilon^\top Z\left(\sigma^2 Z^\top Z\right)^{-1} Z^\top \epsilon = \epsilon^\top P_Z \epsilon/\sigma^2$$

Denoting $\hat{\epsilon}$ the residuals of this regression, the same $\hat{\omega}$ matrix can be constructed and used in a second step to get the more efficient GMM estimator.

With additive errors, the only difference with the linear case is that the minimization process results in a set of non linear equations, so that some numerical methods should be used.

With mlultiplicative errors, we have: $\nu_n = y_n/e^{\beta^\top x_n}$, with $\mathrm{E}(nu_n \mid x_n) = 1$ if all the covariates are exogenous. Defining $\tau_n = \nu_n - 1$, the moment conditions are then:

$$\mathrm{E}\left((y/e^{\beta^\top x} - 1)x_n\right) = 0$$

If some covariates are endogenous, this should be replaced by:

$$\mathrm{E}\left((y/e^{\beta^\top x} - 1)z_n\right) = 0$$

which leads to the following empirical moments:

$$\frac{1}{N}\sum_{n=1}^{N}(y_n/e^{\beta^\top x_n} - 1)z_n = Z^\top(y/e^{X^\top\beta} - 1)/N = Z^\top\tau_n$$

Minimizing the quadratic form of these empirical moments with $(Z^\top Z)^{-1}$ or $\left(\sum_{n=1}^{N}\hat{\tau}_n^2 z_n z_n^\top\right)^{-1}$ leads respectively to the IV and the GMM estimators.

Sargan test

When the number of external instruments is greater that the number of endogenous variables, the empirical moments can't be simultanoulsy set to 0 and a quadratic form of the empirical moments is minimized. The value of the objective function at convergence time the size of the sample is, under the null hypothesis that all the instruments are exogenous a chi square with a number of degrees of freedom equal to the difference between the number of instruments and the number of covariates.

Cigarette smoking behaviour

Mullahy (1997) estimate a demand function for cigarettes which depends on the stock of smoking habits. This variable is quite similar to a lagged dependent variable and is likely to be endogenous as the unobservable determinants of current smoking behaviour should be correlated with the unobservable determinants of past smoking behaviour. The data set contains observations of 6160 males in 1979 and 1980 from the smoking supplement to the 1979 National Health Interview Survey. The response `cigarettes` is the number of cigarettes smoked daily. The covariates are the habit "stock" `habit`, the current state-level average per-pack price of cigarettes `price`, a dummy indicating whether there is in the state of residence a restriction on smokong in restaurants `restaurant`, the age `age` and the number of years of schooling `educ` and their squares, the number of family members `famsize`, and a dummy `race` which indicates whether the individual is white or not. The external instruments are cubic terms in `age` and `educ` and their interaction, the one-year lagged price of a pack of cigarettes `lagprice` and the number of years the state's restaurant smoking restrictions had been in place.

The data set is called `cigmales` and is lazy loaded while attaching the **micsr** package.

The starting point is a basic count model, ie a Poisson model with a log link:

```
library(micsr)
cigmales <- cigmales |>
    transform(age2 = age ^ 2, educ2 = educ ^ 2,
           age3 = age ^ 3, educ3 = educ ^ 3,
           educage = educ * age)
pois_cig <- glm(cigarettes ~ habit + price + restaurant + income + age +
                    age2 + educ + educ2 + famsize + race, data = cigmales,
                family = quasipoisson)
```

The IV and the GMM estimators are provided by the `expreg` function. Its main argument is a two-part formula, where the first part indicates the covariates and the second part the instruments. The instrument set can be constructed from the covariate set by indicating

which series should be omited (the endogenous variables) and which series should be added (the external instruments).

```
iv_cig <- expreg(cigarettes ~ habit + price + restaurant + income + age + age2 +
                        educ + educ2 + famsize + race | . - habit + age3 + educ3 +
                        educage + lagprice + reslgth, data = cigmales,
                    method = "iv")
gmm_cig <- update(iv_cig, method = "gmm")
```

The method argument unables to estimate either the instrumental variables or the general method of moments estimator. The Sargan test gives:

```
sargan(iv_cig) |> gaze()
```

```
chisq = 32.018, df: 4, pval = 0.000
```

```
sargan(gmm_cig) |> gaze()
```

```
chisq = 7.469, df: 4, pval = 0.113
```

Birth weight

The second data set used by Mullahy (1997) consists on 1388 observations on birthweight from the Child Health Supplement to the 1988 National Health Interview Survey. The response is birthweight `birthwt` in pounds and the covariates are the number of cigarettes smoked daily during the pregnancy `cigarettes`, the birth order `parity`, a dummy for white women `race` and child's sex `sex`. Smoking behaviour during the pregnancy is suspected to be correlated with some other unobserved "bad habits" that may be have a negative effect on birthweight. Therefore, performing a pseudo-Poisson regression should result in a upward bias in the estimation of the effect of smoking on birthweight. The external instruments are the number of years of education of the father `edfather` and the mother `edmother`, the family income `faminc` and the per-pack state excise tax on cigarettes.

```
ml_bwt <- glm(birthwt ~ cigarettes + parity + race + sex, data = birthwt,
              family = quasipoisson)
iv_bwt <- expreg(birthwt ~ cigarettes + parity + race + sex |
                    . - cigarettes + edmother + edfather + faminc +
                    cigtax, data = birthwt, method = "iv")
gmm_bwt <- update(iv_bwt, method = "gmm")
```

```
sargan(gmm_bwt)
```

```
    Sargan Test

data:  birthwt
chisq = 3.8743, df = 3, p-value = 0.2754
alternative hypothesis: the moment conditions are not valid
```

Mullahy, John. 1997. "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior." *The Review of Economics and Statistics* 79 (4): 586–93.