Four lecture notes

Part of syllabus for Master Course

ECON4160:

ECONOMETRICS – MODELLING AND SYSTEMS ESTIMATION

CONTENTS:

Erik Biørn

Department of Economics,
University of Oslo

Version of July 5, 2011

ECON 4160: ECONOMETRICS – MODELLING AND SYSTEMS ESTIMATION

Lecture note WH:

THE WU-HAUSMAN TEST FOR EXOGENEITY

Erik Biørn
Department of Economics

Version of July 5, 2011

1. Motivation

Sometimes we do not know or are in doubt whether a variable specified as a right-hand side variable in an econometric equation, is exogenous or endogenous. Consider the following equation, with i as subscript for the observation unit (individual, time period, etc.),

$$y_i = x_i\beta + z_i\gamma + u_i \tag{1}$$

We assume that exogeneity of the vector x_i is not subject to discussion. Exogeneity also of the scalar variable z_i relative to this equation then implies that both x_i and z_i are uncorrelated with the (zero mean) disturbance u_i. *We believe, however, that z_i may be endogenous* and therefore (in general) correlated with u_i. We want to test for this. We want to make a decision by means of a statistical test. Our null hypothesis is therefore

$$H_0: \ \mathrm{cov}(z_i, u_i) = \mathsf{E}(z_i u_i) = 0, \tag{2}$$

and the alternative hypothesis is

$$H_1: \ \mathrm{cov}(z_i, u_i) = \mathsf{E}(z_i u_i) \neq 0, \tag{3}$$

Assume that an *instrument* for z_i relative to (1) is available. It is the scalar variable w_i, i.e., w_i is correlated with z_i and uncorrelated with u_i:

$$\mathrm{cov}(w_i, u_i) = \mathsf{E}(w_i u_i) = 0, \tag{4}$$

Both z_i and w_i may be correlated with x_i. Assume that u_i is non-autocorrelated and has constant variance. We then know the following:

1. If z_i is exogenous and H_0 holds, then we know from Gauss-Markov's theorem that applying OLS on (1) gives the MVLUE (the Minimum Variance Linear Unbiased Estimators) of β and γ. These estimators, denoted as $(\widehat{\beta}_{OLS}, \widehat{\gamma}_{OLS})$, are therefore consistent.

2. If z_i is correlated with u_i and H_0 is violated, then $(\widehat{\beta}_{OLS}, \widehat{\gamma}_{OLS})$ are *both* inconsistent.

3. Estimating (1) by the two-stage least squares (2SLS), using w_i as an instrument for z_i gives consistent estimators of β and γ, denoted as $(\widehat{\beta}_{2SLS}, \widehat{\gamma}_{2SLS})$. Consistency is ensured regardless of whether H_0 or H_1 holds.

4. We then have two sets of estimators of β and γ: (i) $(\widehat{\beta}_{2SLS}, \widehat{\gamma}_{2SLS})$, is *consistent* both under H_0 and H_1, but *inefficient* under the former. (ii) $(\widehat{\beta}_{OLS}, \widehat{\gamma}_{OLS})$, is *consistent and efficient* under H_0, but *inconsistent* under H_1. Hence, $(\widehat{\beta}_{2SLS}, \widehat{\gamma}_{2SLS})$ is *more robust to inconsistency* than $(\widehat{\beta}_{OLS}, \widehat{\gamma}_{OLS})$. The price we have to pay when applying the former is, however, its loss of efficiency if H_0 is in fact true. Intuition then says that the "distance" between $(\widehat{\beta}_{2SLS}, \widehat{\gamma}_{2SLS})$ and $(\widehat{\beta}_{OLS}, \widehat{\gamma}_{OLS})$ should "on average" be "smaller" under H_0 than under H_1.

From the last statement in 4 we might proceed by estimating (1) by both OLS and 2SLS and investigate whether their discrepancy "seems large or small". The former suggests rejection and the latter non-rejection of exogeneity of z_i. Some people follow this strategy, but it can at most be an informal approach, and not strictly a test, since we do not know which distances should be denoted as "large" and which "small". A formal method is obviously needed.

Essentially, the Wu-Hausman test is a method for investigating whether the discrepancy is statistically large enough to call H_0 into doubt and accept H_1.

2. A simple implementation of the Wu-Hausman test

We formulate the assumed relationship between the instrument, the regressor vector x_i and the instrument w_i for z_i in (1) as follows:

$$z_i = x_i\delta + w_i\lambda + v_i, \tag{5}$$

and assume that

$$u_i = v_i\rho + \varepsilon_i, \tag{6}$$

where

$$\operatorname{cov}(\varepsilon_i, v_i) = \operatorname{cov}(\varepsilon_i, w_i) = \operatorname{cov}(x_i, v_i) = \operatorname{cov}(\varepsilon_i, x_i) = 0. \tag{7}$$

Remark 1: Equation (5) may be the *reduced form* for z_i in a multi-equation model to which (1) belongs, y_i is without doubt endogenous and z_i is potentially endogenous and determined jointly with y_i. Then (x_i, w_i) are the exogenous variables in the model. It is supposed that (1) is identified.

Remark 2: It is perfectly possible that z_i is a regressor variable affected by a *random measurement error*, where w_i is an instrument for the true (unobserved) value of z_i, and hence also for z_i itself.

From (4)–(7) it follows that:

$$\text{cov}(z_i, u_i) = \text{cov}(x_i\delta + w_i\lambda + v_i, v_i\,\rho + \varepsilon_i) = \rho\,\text{var}(v_i) \tag{8}$$

and therefore

$$\begin{aligned} H_0 &\implies \rho = 0, \\ H_1 &\implies \rho \neq 0, \end{aligned} \tag{9}$$

Inserting (6) into (1) gives

$$y_i = x_i\beta + z_i\gamma + v_i\rho + \varepsilon_i \tag{10}$$

Let the OLS estimates of (δ, λ) in (5) be $(\widehat{\delta}_{OLS}, \widehat{\lambda}_{OLS})$, and *compute the residuals*

$$\widehat{v}_i = z_i - x_i\widehat{\delta}_{OLS} - w_i\widehat{\lambda}_{OLS}. \tag{11}$$

Replace v_i with $\widehat{v}_i$ in (10), giving

$$y_i = x_i\beta + z_i\gamma + \widehat{v}_i\rho + \epsilon_i \tag{12}$$

Estimate the coefficients of (12), (β, δ, ρ), by OLS i.e., by regressing y_i on $(x_i, z_i, \widehat{v}_i)$. Test, by means of a t-test whether the OLS estimate of ρ is significantly different from zero or not.

3. Conclusion

This leads to the following prescription for performing a Wu-Hausman test and estimating (1):

Rejection of $\rho = 0$ from OLS and t-test on (12) $\Longrightarrow$ rejection of H_0, i.e., rejection of exogeneity of z_i in (1). Stick to 2SLS estimation of (1).

Non-rejection of $\rho = 0$ from OLS and t-test on (12) $\Longrightarrow$ non-rejection of H_0, i.e., non-rejection of exogeneity of z_i in (1). Stick to OLS estimation of (1).

4. A more general exposition (Optional)

Consider the following, rather general, situation: We have two sets of estimators of a $(K \times 1)$ parameter vector $\boldsymbol{\beta}$ based on n observations: $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$. We want to test a hypothesis H_0 against a hypothesis H_1. A frequently arising situation may be that a regressor vector is uncorrelated with the equation's disturbance under H_0 and correlated with the disturbance under H_1. The latter may be due to endogenous regressors, random measurement errors in regressors, etc. The two estimators have the following properties in relation to the two hypotheses:

(i) $\widehat{\boldsymbol{\beta}}_0$, is *consistent and efficient* for $\boldsymbol{\beta}$ under H_0, but *inconsistent* under H_1.

(ii) $\widehat{\boldsymbol{\beta}}_1$ is *consistent* for $\boldsymbol{\beta}$ both under H_0 and H_1, but *inefficient* under H_0.

Hence, $\widehat{\boldsymbol{\beta}}_1$ is *more robust to inconsistency* than $\widehat{\boldsymbol{\beta}}_0$. The price we have to pay when applying $\widehat{\boldsymbol{\beta}}_1$ is, however, its *loss of efficiency* when H_0 is in fact true and the latter estimator should be preferred.

Intuition says that the 'distance' between the estimator vectors $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ should 'on average' be 'smaller' when H_0 is true than when H_1 is true. This intuition suggests that we might proceed by estimating $\boldsymbol{\beta}$ by both methods and investigate whether the discrepancy between the estimate vectors seems 'large' or 'small': $\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0$ 'large' suggests rejection, and $\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0$ small suggests non-rejection of H_0. Such a 'strategy', however can at most be an informal approach, not a statistical test, since we do not know which distances should be judged as 'large' and which be judged as 'small'. A formal method determining a critical region criterion is needed. At this point the Hausman specification test comes in.

Essentially, the Hausman test to be presented below is a method for investigating whether the discrepancy between $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_0$ is statistically large enough to call H_0 into doubt and accept H_1.

LEMMA, J.A. HAUSMAN: ECONOMETRICA (1978):

We have that:

$$
\begin{array}{lll}
(a) & \sqrt{n}(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \overset{d}{\rightarrow} \mathcal{N}(\mathbf{0}, \boldsymbol{V}_0) & \text{under } H_0 \\
(b) & \sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \overset{d}{\rightarrow} \mathcal{N}(\mathbf{0}, \boldsymbol{V}_1) & \text{under } H_0 \\
(c) & \boldsymbol{V}_1 - \boldsymbol{V}_0 & \text{is positive definite under } H_0
\end{array}
$$

Then:

$$
\begin{array}{lll}
(d) & \boldsymbol{S} = \mathsf{V}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0) = \mathsf{V}(\widehat{\boldsymbol{\beta}}_1) - \mathsf{V}(\widehat{\boldsymbol{\beta}}_0) & \text{is positive definite under } H_0, \\
(e) & Q = n(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0)'\widehat{\boldsymbol{S}}^{-1}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0) \overset{d}{\rightarrow} \chi^2(K) & \text{under } H_0,
\end{array}
$$

where $\widehat{\boldsymbol{S}}$ is a consistent estimator of $\boldsymbol{S} = \mathsf{V}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0)$, i.e., the variance-covariance matrix of the estimator difference vector $\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0$. Here Q is a quadratic form (and therefore a scalar) measuring the distance between the two estimator vectors. A crucial part of the lemma is that under H_0, Q is distributed as χ^2 with K degrees of freedom, K being number of parameters under test, and tends to be larger under H_1 than under H_0.

Test criterion: Reject H_0, at approximate significance level ε, when $Q > \chi^2_{1-\varepsilon}(K)$.

Further readings, proofs etc.

Davidson, R. and MacKinnon, J.G. (1993): *Estimation and Inference in Econometrics.* Oxford University Press, section 7.9.

Greene, W.H. (2008): *Econometric Analysis, Sixth edition.* Prentice-Hall, section 12.4.

Hausman, J.A. (1978): *Econometrica*, 46 (1978), 1251-1272.

Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, Section 6.2.1.

Wu, D.M. (1973): *Econometrica*, 41 (1973), 733-750.

ECON 4160: ECONOMETRICS – MODELLING AND SYSTEMS ESTIMATION

Lecture note OR:

TESTING FOR OVERIDENTIFYING RESTRICTIONS, ETC.

Erik Biørn,
Department of Economics

Version of July 5, 2011

Setting:

Consider the following equation, with i as subscript for the observation unit,

$$(1) \qquad y_i = z_i\gamma + u_i, \qquad i = 1, \dots, n.$$

Assume that we believe that the $1 \times k$ *vector* z_i *contains endogenous variables* and is therefore (in general) correlated with u_i. Some of the elements in z_i may, however, be exogenous. Assume that an *instrument vector (IV vector* for z_i relative to (1) is available. It is the $1 \times q$ vector w_i, with $q \geq k$. If w_i is a valid instrument, it should by *assumption be* correlated with z_i and uncorrelated with u_i. If z_i includes exogenous variables, say the subvector x_i (confer the notation in Lecture note WH, section 2), then w_i should also contain x_i as a subvector. The assumption that w_i and u_i are uncorrelated is called the orthogonality assumption.

Since $q \geq k$, the equation is identified, exactly identified if $q = k$ and overidentified if $q > k$ – in both cases provided that the orthogonality conditions – w_i and u_i being uncorrelated – hold. Under overidentification, $q-k$ overidentifying restrictions are imposed. Remember:

[1] Overidentification implies that we have more instruments than needed for estimability (consistent estimation).

[2] In a simultaneous equation model where the model's exogenous (predetermined) variables are candidates for being valid IVs, the system's structural form (SF) imposes restrictions on the coefficient of the system's reduced for (RF). For examples of such overidentifying restrictions for linear simultaneous systems, see Lecture Note E in "Six notes", section 10 [in particular (E-68)–(E-69)].

[3] Asserting that a set of IVs contribute to overidentification may be correct of wrong.

Assume that (1) is estimated by the Two-Stage Least Squares (2SLS), which is equivalent to using w_i as an instrument for z_i chosen in the optimal way. This gives consistent estimators of γ, denoted as $\widehat{\gamma}_{2SLS}$. If $q = k$, there are no overidentifying restrictions to test. If $q > k$, we may want to test the $q - k$ overidentifying restrictions.

Testing of overidentifying restrictions:

A simple test of these overidentifying restrictions (several others exist) can then be performed as follows [see Davidson & MacKinnon (2004) for elaboration]:

Step 1: Compute the residuals from the 2SLS estimation, assuming w_i – with $q > k$ – to be a valid IV vector for z_i:

$$\widetilde{u}_i = y_i - z_i \widehat{\gamma}_{2SLS} \qquad i = 1, \ldots, n. \tag{2}$$

Step 2: Regress $\widetilde{u}_i$ on the full IV vector, w_i.

Step 3: Compute R^2 from this regression. Denote it as $R^2_{\tilde{u}}$

Step 4: If the orthogonality condition is valid, then $nR^2_{\tilde{u}} \sim \chi^2(q-k)$ holds approximately, $\chi^2(q-k)$ denoting the χ^2-distribution with $q-k$ degrees of freedom. If $nR^2_{\tilde{u}} > \chi^2_{1-\alpha}(q-k)$, where $\chi^2_{1-\alpha}(q-k)$ is the $(1-\alpha)$-quantile of this distribution, the overidentifying restrictions are rejected at a significance level α.

Further readings, proofs, etc.:

Davidson & MacKinnon (1993): *Estimation and Inference in Econometrics.* Oxford University Press, section 7.8.

Davidson & MacKinnon (2004): *Econometric Theory and Methods.* Oxford University Press, section 8.6.

Testing of orthogonality conditions:

See Lecture note WH.

On the 'weak instrument' problem:

"Archimedes said, "Give me the place to stand, and a lever long enough, and I will move the Earth"...... Economists have their own powerful lever: the instrumental variable estimator. The instrumental variable estimator can avoid the bias that ordinary least squares suffers when an explanatory variable in a regression is correlated with the regression's disturbance term. But, like Archimedes lever, instrumental variable estimation requires both a valid instrument on which to stand and an instrument that isn't too short (or "too weak")." [M.P. Murray (2006, p.111)].

Further readings:

BOUND, JAEGER & BAKER (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *Journal of the American Statistical Association* 90, 443-450.

MURRAY (2006). Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives* 20, Number 4, 111-132.

NELSON & STARTZ (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63, S125-S140.

STAIGER & STOCK (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557-586.

STOCK, WRIGHT, & YOGO (2002): A Survey of weak instruments and weak identification in Generalized Method of Moments. *Journal of the American Statistical Association* 20, 518-529.

ECON 4160: ECONOMETRICS – MODELLING AND SYSTEMS ESTIMATION

Lecture note DL:

DISTRIBUTED LAGS – AN ELEMENTARY DESCRIPTION

Erik Biørn
Department of Economics

Version of July 5, 2011

1. A reinterpretation of the regression model

Consider a static regression equation for time series data:

$$(1) \qquad y_t = \alpha + \beta_0 x_t + \beta_1 z_t + u_t, \qquad t = 1, \dots, T.$$

We know that it is possible to let, for example, $z_t = x_t^2$, $z_t = \ln(x_t)$, etc. Econometrically, we then consider x_t and z_t as two different variables, even if they are functions of the same basic variable, x. Usually, x and z thus defined are correlated, but they will not be perfectly correlated. (Why?)

Let us now choose $z_t = x_{t-1}$ and *treat x_t and z_t econometrically as two different variables*. We then say that z_t is the value of x_t backward time-shifted, or *lagged*, one period. Then x_t and z_t will usually be correlated, but not perfectly, unless $x_t = a + bt$, in which case $x_t = x_{t-1} + b$. When $z_t = x_{t-1}$, (1) reads

$$(2) \qquad y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + u_t, \qquad t = 2, \dots, T,$$

where we, for simplicity, assume that u_t is a classical disturbance. We than say that (2) is a *distributed lag equation*. By this term we mean that the effect on y of a change in x is distributed over time. We have:

$$\begin{aligned}
\frac{\partial y_t}{\partial x_t} &= \beta_0 = \text{short-run, immediate effect,} \\
\frac{\partial y_t}{\partial x_{t-1}} &= \beta_1 = \text{effect realized after one period,} \\
\frac{\partial y_t}{\partial x_t} + \frac{\partial y_t}{\partial x_{t-1}} &= \beta_0 + \beta_1 = \text{long-run effect} \\
&= \text{total effect of a change lasting for at least two periods.}
\end{aligned}$$

Unsually, there are no problems in estimating such a model. We treat it formally as a regression equation with two right hand side (RHS) variables. We may then use classical regression, OLS (or GLS), as before. Note that when (y_t, x_t) are observed for $t = 1, \dots, T$, eq. (2) can only be estimated from observations $t = 2, \dots, T$ since one observation is "spent" in forming the lag.

Eq. (2) may be generalized to a *distributed lag equation with K lags*:

$$
\begin{aligned}
(3) \qquad y_t &= \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_K x_{t-K} + u_t \\
&= \alpha + \textstyle\sum_{i=0}^{K} \beta_i x_{t-i} + u_t, \qquad t = K+1, \ldots, T.
\end{aligned}
$$

Neither are there any problems, *in principle*, with estimating such a model. We treat it formally as a regression equation with K+1 RHS variables plus an intercept. Note that (3) can only be estimated from observations $t = K+1, \ldots, T$ since K observations are disposed of in forming the lags. In the model we may include other variables, $z_t, q_t, \ldots$, with similar lag-distributions. The lag distribution in (3) is characterized by the coefficient sequence $\beta_0, \beta_1, \ldots, \beta_K$, and we have

$$
\begin{aligned}
\frac{\partial y_t}{\partial x_t} &= \beta_0 = \text{short-run effect}, \\
\frac{\partial y_t}{\partial x_{t-i}} &= \beta_i = \text{effect realized after } i \text{ periods } (i = 1, \ldots, K), \\
\sum_{i=0}^{K} \frac{\partial y_t}{\partial x_{t-i}} &= \sum_{i=0}^{K} \beta_i = \text{long-run effect} \\
&= \text{total effect of a change lasting for at least } K+1 \text{ periods.}
\end{aligned}
$$

It follows that we also have

$$
\begin{aligned}
\frac{\partial y_{t+i}}{\partial x_t} &= \beta_i \quad (i = 1, \ldots, K), \\
\sum_{i=0}^{K} \frac{\partial y_{t+i}}{\partial x_t} &= \sum_{i=0}^{K} \beta_i.
\end{aligned}
$$

Eq. (3) exemplifies a *finite lag distribution*. The entire process is exhausted in the course of a finite number of periods, $K+1$. We loose K observations when we form the lags. Estimating (3) by OLS involves *estimating $K+2$ coefficients* $(\alpha, \beta_0, \beta_1, \ldots, \beta_K)$ *from $T-K$ observations*. The number of *degrees of freedom* we dispose of in the estimation is thus $T - K - K - 2 = T - 2(K+1)$. We see that *we spend two degrees of freedom for each additional lag we include in the equation.* This may give rise to multicollinearity problems and imprecise estimates. In order to have a positive number of degrees of freedom, we must have $T > 2(K+1)$.

2. Why are distributed lag equations interesting?

Distributed lag equations can be used to model:

- *Technical lags:* Ex.: Dynamic production processes.
- *Behavioural lags:* Delays from a signal is received by an agent until he/she responds.
- *Institutional lags:* Ex.: Delays from the time a sum of money is paid from the paying institution until it is received by or registered in the accounts of the receiving institution.

- *Dynamization of theories*: We have a static theory which we want to "dynamize" in order to take account of, estimate, or test for possible sluggishness in the responses predicted by the theory.

Examples:

1. Relationship between investment and scrapping of capital.
2. Delays in the shifting of cost changes into output prices.
3. Delays in the shifting of consumer price changes into changes in wage rates.
4. Delays in production to order. Lags from an order is delivered until production starts. Lags from production starts until it is finished.
5. Production processes in Ship-building, Building and Construction industries.

Note that the necessity of modeling lag distributions and the way we model such distributions is *more important the shorter the unit observation period in our data set is.* That is, the modeling related to a given problem is, cet. par., more important for quarterly data than for annual data, more important for monthly data than for quarterly data, more important for weekly data than for monthly data, and so on.

3. Imposing restrictions on distributed lag equations

Because of the potentially low number of degrees of freedom and the potential multicollinearity problem that arise when the coefficients in the lag distributions are unrestricted, we may want to impose restrictions on the lag coefficients. In this way we may "save" coefficients and at the same time ensure that the coefficient sequence $\beta_0, \beta_1, \ldots, \beta_K$ exhibits a more "smooth" pattern. We shall illustrate this idea by means of two examples.

Example 1: Polynomial lag distributions. Finite lag distribution

Assume that the $K+1$ coefficients in the lag distribution are *restricted to lie on a polynomial of degree P*, i.e.,

$$(4)\quad \beta_i = \gamma_0 + \gamma_1 i + \gamma_2 i^2 + \cdots + \gamma_P i^P = \gamma_0 + \sum_{p=1}^{P} \gamma_p i^p, \qquad i = 0, 1, \ldots, K;\; P < K,$$

where $\gamma_0, \gamma_1, \ldots, \gamma_K$ are $P+1$ unknown coefficients. It is important that P is smaller than (and usually considerably smaller than) K. Inserting (4) into (3), we obtain

$$\begin{aligned}(5)\, y_t &= \alpha + \sum_{i=0}^{K} \left(\gamma_0 + \sum_{p=1}^{P} \gamma_p i^p \right) x_{t-i} + u_t \\ &= \alpha + \gamma_0 \sum_{i=0}^{K} x_{t-i} + \gamma_1 \sum_{i=0}^{K} i x_{t-i} + + \cdots + \gamma_P \sum_{i=0}^{K} i^P x_{t-i} + u_t, \qquad t = K+1, \ldots, T.\end{aligned}$$

This is an equation of the form

$$(6) \qquad y_t = \alpha + \gamma_0 z_{0t} + \gamma_1 z_{1t} + \gamma_2 z_{2t} + \cdots + \gamma_P z_{Pt} + u_t, \quad t = K+1, \dots, T,$$

where the $P+1$ (observable) RHS variables are

$$z_{0t} = \textstyle\sum_{i=0}^{K} x_{t-i}, \quad z_{pt} = \sum_{i=0}^{K} i^p x_{t-i}, \qquad p = 1, \dots, P.$$

We see that z_{0t} is a non-weighted sum of the current and the lagged x values, and $z_{1t}, \dots, z_{Pt}$ are weighted sums with the weights set equal to the lag length raised to powers $1, \dots, P$, respectively.

If u_t is a classical disturbance, we can estimate $\alpha, \gamma_0, \gamma_1, \dots, \gamma_P$ by applying OLS on (6). The number of degrees of freedom is then $T - K - P - 2$, i.e. we gain $K - P$ degrees of freedom as compared with free estimation of the β's. Let the estimates be $\widehat{\alpha}, \widehat{\gamma}_0, \widehat{\gamma}_1, \dots, \widehat{\gamma}_P$. We can then estimate the original lag coefficients by inserting these estimates into (4), giving

$$(7) \;\; \widehat{\beta}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 i + \widehat{\gamma}_2 i^2 + \cdots + \widehat{\gamma}_P i^P = \widehat{\gamma}_0 + \textstyle\sum_{p=1}^{P} \widehat{\gamma}_p i^p, \quad i = 0, 1, \dots, K; \;\; P < K.$$

Exercise: Show that these estimators are unbiased and consistent. Hint: Use Gauss-Markov's and Slutsky's theorems. Will these estimators be Gauss-Markov estimators (MVLUE)? How would you proceed to test whether a third degree polynomial ($P = 3$) gives a significantly better fit than a second degree polynomial ($P = 2$)?

Let us consider the special case with *linear lag distribution* and zero restriction and the far endpoint of the distribution. Let $P = 1$ and

$$\gamma_1 = -\frac{\gamma_0}{K+1},$$

and let γ_0 be a free parameter. Inserting these restrictions into (4), we get

$$\beta_i = \gamma_0 \left(1 - \frac{i}{K+1}\right), \qquad i = 0, 1, \dots, K.$$

Then (5) becomes

$$y_t = \alpha + \gamma_0 \sum_{i=0}^{K} \left(1 - \frac{i}{K+1}\right) x_{t-i} + u_t, \qquad t = K+1, \dots, T.$$

This lag distribution has only two unknown coefficients, so that we can estimate α and γ_0 by regressing y_t on $\sum_{i=0}^{K}(1 - i/(K+1))x_{t-i}$ from the $T - K$ observations available. Finally, we estimate β_i by $\widehat{\beta}_i = \widehat{\gamma}_0(1 - i/(K+1))$

Exercise: Find, by using the formula for the sum of an arithmetic succession, an expression for the long-run effect in this linear lag distribution model. How would you estimate it?

Example 2: Geometric lag distribution. Infinite lag distribution

We next consider the equation

$$(8) \qquad y_t = \alpha + \beta x_t + \lambda y_{t-1} + \varepsilon_t, \qquad |\lambda| < 1, \quad t = 2, \dots, T,$$

in which we have included the value of the LHS variable lagged one period as an additional regressor to x_t, let $\boldsymbol{x} = (x_1, \dots, x_T)$, and assume that

$$(9) \qquad E(\varepsilon_t|\boldsymbol{x}) = 0, \qquad E(\varepsilon_t\varepsilon_s|\boldsymbol{x}) = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s. \end{cases}$$

We call (8) an *autoregressive equation of the first order in* y_t *with an exogenous variable* x_t. What kind of lag response will this kind of model involve?

Let us in (8) insert backwards for $y_{t-1}, y_{t-2}, \dots$, giving

$$\begin{aligned} y_t &= \alpha + \beta x_t + \lambda(\alpha + \beta x_{t-1} + \lambda y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \alpha(1+\lambda) + \beta(x_t + \lambda x_{t-1}) + \lambda^2((\alpha + \beta x_{t-2} + \lambda y_{t-3} + \varepsilon_{t-2}) + \varepsilon_t + \lambda\varepsilon_{t-1} \\ &\vdots \\ &= \alpha(1 + \lambda + \lambda^2 + \cdots) + \beta(x_t + \lambda x_{t-1} + \lambda^2 x_{t-1} + \cdots) + \varepsilon_t + \lambda\varepsilon_{t-1} + \lambda^2\varepsilon_{t-2} + \cdots \ , \end{aligned}$$

since $|\lambda| < 1$. Hence, using the summation formula for a convergent infinite geometric succession, we have at the limit

$$(10) \qquad y_t = \frac{\alpha}{1-\lambda} + \beta\sum_{i=0}^{\infty}\lambda^i x_{t-i} + \sum_{i=0}^{\infty}\lambda^i\varepsilon_{t-i}.$$

Comparing this equation with (3), we see that y is related to x via a lag distribution with an infinitely large number of terms ($K \to \infty$), with the lag coefficients given by

$$(11) \qquad \beta_i = \beta\lambda^i, \qquad i = 0, 1, \dots \ ,$$

and with a disturbance

$$(12) \qquad u_t = \sum_{i=0}^{\infty}\lambda^i\varepsilon_{t-i}.$$

Eq. (11) implies:

$$(13) \qquad \beta_0 = \beta, \quad \beta_1 = \beta\lambda, \quad \beta_2 = \beta\lambda^2, \quad \beta_3 = \beta\lambda^3, \ \dots \ .$$

The short-run effect is thus $\beta_0 = \beta$. The long-run effect is

$$\sum_{i=0}^{\infty} \beta_i = \beta \sum_{i=1}^{\infty} \lambda^i = \frac{\beta}{1-\lambda}, \tag{14}$$

when exploiting the assumption $|\lambda| < 1$ and the formula for the sum of an infinite convergent geometric succession.

We denote a lag distribution with an infinite number of terms an *infinite lag distribution.* Since the lag coefficients in (10) decline as a convergent infinite geometric succession in the lag number. We denote the lag distribution in (9) a *geometric lag distribution.* Geometric lag distributions play an important role in dynamic econometrics. An advantage with it is that we do not need to be concerned with specifying the maximal lag K, which may often be difficult. For practical purposes, we consider $K \to \infty$ as an approximation.

We can write (10) as

$$y_t = \frac{\alpha}{1-\lambda} + \beta \sum_{i=0}^{\infty} \lambda^i x_{t-i} + u_t. \tag{15}$$

Since (12) implies

$$u_{t-1} = \sum_{i=0}^{\infty} \lambda^i \varepsilon_{t-i-1} = \sum_{j=1}^{\infty} \lambda^{j-1} \varepsilon_{t-j}$$

and

$$u_t = \lambda \sum_{i=1}^{\infty} \lambda^{i-1} \varepsilon_{t-i} + \varepsilon_t$$

it follows that

$$u_t = \lambda u_{t-1} + \varepsilon_t. \tag{16}$$

This shows that u_t defined by (12) is an autoregressive process of the first order, an AR(1)-process. We can the state the following conclusion: *A first order autoregressive equation in y_t with an exogenous variable x_t and with the autoregressive parameter $|\lambda| < 1$ is equivalent to expressing y_t as an infinite, geometric lag distribution in x_t with an AR(1) disturbance with autoregressive parameter λ.*

How could we estimate this kind of model? First, application of OLS on (10) will not work, because it has an infinite number of RHS variables. Now, we know that

$$\operatorname{cov}(x_t, u_t) = 0,$$

since x_t is exogenous. Moreover, since lagging (10) one period yields

$$y_{t-1} = \frac{\alpha}{1-\lambda} + \beta \sum_{i=0}^{\infty} \lambda^i x_{t-i-1} + \sum_{i=0}^{\infty} \lambda^i \varepsilon_{t-i-1},$$

we have that

$$\mathrm{cov}(y_{t-1}, \varepsilon_t) = \mathrm{cov}\left(\frac{\alpha}{1-\lambda} + \beta \sum_{i=0}^{\infty} \lambda^i x_{t-i-1} + \sum_{i=0}^{\infty} \lambda^i \varepsilon_{t-i-1}, \varepsilon_t\right) = 0,$$

because, in view of (9), ε_t is uncorrelated with all past ε's. Application of OLS on (8) is therefore consistent, since its disturbance is uncorrelated with both of its RHS variables. After having obtained the estimates $\widehat{\alpha}$, $\widehat{\beta}$, and $\widehat{\lambda}$, we can estimate the intercept of (10) by $\widehat{\alpha}/(1-\widehat{\lambda})$ and the coefficients in (10), i.e. the lag responses, using (13), by means of

$$\widehat{\beta}_0 = \widehat{\beta}, \quad \widehat{\beta}_1 = \widehat{\beta}\widehat{\lambda}, \quad \widehat{\beta}_2 = \widehat{\beta}\widehat{\lambda}^2, \quad \widehat{\beta}_3 = \widehat{\beta}\widehat{\lambda}^3, \dots .$$

The long-run effect can be estimated as

$$\sum_{i=0}^{\infty} \widehat{\beta}_i = \frac{\widehat{\beta}}{1-\widehat{\lambda}}.$$

Exercise: Show that $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_3$ as well as the corresponding estimator of the long-run coefficient are consistent.

ECON 4160: ECONOMETRICS – MODELLING AND SYSTEMS ESTIMATION

Lecture note DC:

ANALYSING DISCRETE CHOICE – AN ELEMENTARY EXPOSITION

Erik Biørn
Department of Economics

Version of July 5, 2011

1. Background

Several economic variables are observed as the results of individuals' choices between a *limited number* of alternatives. In this note, we shall assume that only two alternatives are available, e.g.: purchase/not purchase a car, apply for/not apply for a job, obtain/not obtain a loan, travel to work by own car/public transport. These are examples of genuine *qualitative choices*. Since there are two alternatives, we call it a binomial (or binary) choice. We represent the outcome of the choice by a binary variable. We are, from more basic courses in econometrics, familiar with using linear regression analysis in connection with models with binary *right-hand-side* (RHS) variables (dummy regressor variables). *The models we now consider, have binary left-hand-side (LHS) variables.*

Let the two possible choices be denoted as *'positive response'* and *'negative response'*, respectively, assume that n individuals are observed, and let

$$(1)\qquad y_i = \begin{cases} 1 & \text{if individual } i \text{ responds positively,} \\ 0 & \text{if individual } i \text{ responds negatively,} \end{cases} \qquad i = 1, \dots, n.$$

Moreover, we assume that the individuals are observed independently of each other. How should we model the determination of y_i? As potential explanatory (exogenous) variables we have the vector $x_i = (1, x_{1i}, x_{2i}, \dots, x_{Ki})$, some of which are continuous and some may be binary variables.

2. Why is a linear regression model inconvenient?

Let us first attempt to model the determination of y_i by means of a standard linear regression model:

$$(2)\qquad y_i = x_i\beta + u_i, \qquad i = 1, \dots, n,$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_K)'$. Which consequences will this have?

First, the LHS variable is, by nature, different from the RHS variables. By (2) we attempt to put "something discrete" equal to "something continuous".

Second, the choice of values for y_i, i.e., 0 and 1, is arbitrary. We might equally well have used (1,2), (5,10), (2.71,3.14), etc. This would, however, have changed the β's, which means that the β's get no clear interpretation.

Third, let us imagine that we draw a scatter of points exhibiting the n y and x values, the former being either zero or one, the latter varying continuously. It does not seem very meaningful, or attractive, to draw a straight line, or a plane, through this scatter of points in order to minimize a squared distance, as we do in classical regression analysis.

Fourth, according to (1) and (2), the disturbance u_i can, for each x_i, only take one of two values:

$$u_i = \begin{cases} 1 - x_i\beta & \text{if individual } i \text{ responds positively,} \\ -x_i\beta & \text{if individual } i \text{ responds negatively.} \end{cases}$$

Let P_i denote the probability that individual i responds positively, i.e., $P(y_i = 1) = P(u_i = 1 - x_i\beta)$. It is commonly called the *response probability* and $1 - P_i$ is called the non-response probability. The last statement is then equivalent to

$$(3) \qquad u_i = \begin{cases} 1 - x_i\beta & \text{with probability } P_i = P(y_i = 1), \\ -x_i\beta & \text{with probability } 1 - P_i = P(y_i = 0). \end{cases}$$

For this reason, it is, for instance, *impossible that u_i can follow a normal distribution*, even as an approximation.

Fifth, let us require that the expectation of u_i, conditional on the exogenous variables, is zero, as in standard regression analysis. Using the definition of an expectation in a discrete probability distribution, this implies

$$E(u_i|x_i) = (1 - x_i\beta)P_i + (-x_i\beta)(1 - P_i) = P_i - x_i\beta = 0.$$

Hence

$$(4) \qquad P_i = x_i\beta, \qquad i = 1, \dots, n,$$

so that (2) is equivalent to

$$(5) \qquad y_i = P_i + u_i, \qquad i = 1, \dots, n,$$

i.e., the disturbance has the interpretation as the difference between the binary response variable and the response probability. The response probability is usually a continuously varying entity. The variance of the disturbance is, in view of (3) and (4), when we use the definition of a variance in a discrete probability distribution,

$$(6) \qquad \text{var}(u_i|x_i) = (1 - x_i\beta)^2 P_i + (-x_i\beta)^2(1 - P_i) = (1 - x_i\beta)x_i\beta.$$

We note that this disturbance variance is a function of both x_i and β. This means, on the one hand, that the disturbance is heteroskedastic, on the other hand that its variance depends on the slope coefficients of (2).

Sixth, we know that any probability should belong to the interval (0,1), but there is no guarantee that the RHS of (4) should be within these two bounds. This is a serious limitation of the linear model (2) – (4).

We can therefore conclude that there are considerable problems involved in modeling the determination of the binary variable y_i by the linear regression equation (2).

3. A better solution: Modeling the response probability

We have seen that there is no guarantee that $P_i = x_i\beta$ belongs to the interval (0,1). A more attractive solution than (2) is to model the mechanism determining the individual response by choosing, *for the response probability*, a (non-linear) functional form such that it will always belong to the interval (0,1). We therefore let

$$P_i = F(x_i\beta) \tag{7}$$

and *choose* F *such that its domain is* $(-\infty, +\infty)$ *and its range is (0,1).* Moreover, we require that F is monotonically increasing in its argument, which means that

$$F(-\infty) = 0, \qquad F(+\infty) = 1, \qquad F'(x_i\beta) \geq 0.$$

Two choices of such an F function have become popular: The first is

$$P_i = P(y_i = 1) = F(x_i\beta) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = \frac{1}{1 + e^{-x_i\beta}}, \tag{8}$$

which is the cumulative distribution function (CDF) of the *logistic* distribution. The second is

$$P_i = P(y_i = 1) = F(x_i\beta) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du, \tag{9}$$

which is the CDF of the *standardized normal distribution*, i.e., the N(0,1) distribution. Both these distributions are symmetric. The response mechanism described by (8) is called the **Logit model**. The response mechanism described by (9) is called the **Probit model**. Their *non-response* probabilities are, respectively

$$1 - P_i = P(y_i = 0) = 1 - F(x_i\beta) = \frac{1}{1 + e^{x_i\beta}} = \frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}}$$

and

$$1 - P_i = P(y_i = 0) = 1 - F(x_i\beta) = \int_{x_i\beta}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

4. A closer look at the logit model's properties

Let us take a closer look at the logit model, its interpretation and estimation procedures. An advantage with this model is that it, unlike the Probit model, expresses the response probability in closed form, i.e., by an explicit algebraic expression, not by an integral.

From (8) it follows that the ratio between the response and the non-response probabilities in the Logit model, often denoted as the *odds-ratio*, is

$$\frac{P_i}{1-P_i} = \frac{P(y_i = 1)}{P(y_i = 0)} = e^{x_i\beta}.$$

This ratio is monotonically increasing in P_i from zero to infinity as P_i increases from zero to one. Taking its logarithm we get the *log-odds-ratio*, i.e., the logarithm of the ratio between the response and the non-response probabilities,

$$\ln\left(\frac{P_i}{1-P_i}\right) = \ln\left(\frac{P(y_i = 1)}{P(y_i = 0)}\right) = x_i\beta. \tag{10}$$

This shows that *the Logit-model is parametrized in such a way that the log-odds ratio is linear in the explanatory variables (possibly after a known transformation).* The coefficient β_k represents the effect on the log-odds ratio of a one unit increase in x_{ki}. The log-odds ratio has the general property of being monotonically increasing in P_i from minus infinity to plus infinity as P_i increases from zero to one. This is reassuring as, in principle, $x_i\beta$ may take any value.

Since (8) implies

$$\ln(P_i) = x_i\beta - \ln(1 + e^{x_i\beta}), \qquad \ln(1 - P_i) = -\ln(1 + e^{x_i\beta}),$$

it is not difficult to show that

$$\frac{\partial \ln(P_i)}{\partial(x_i\beta)} = 1 - P_i, \qquad \frac{\partial \ln(1-P_i)}{\partial(x_i\beta)} = -P_i, \tag{11}$$

from which it follows that the effect on the log of the response probabilities of a change in the k'th explanatory variable is

$$\frac{\partial \ln(P_i)}{\partial x_{ki}} = (1 - P_i)\beta_k, \qquad \frac{\partial \ln(1-P_i)}{\partial x_{ki}} = -P_i\beta_k, \tag{12}$$

or equivalenly,

$$\frac{\partial P_i}{\partial x_{ki}} = P_i(1 - P_i)\beta_k, \qquad \frac{\partial (1-P_i)}{\partial x_{ki}} = -P_i(1-P_i)\beta_k. \tag{13}$$

Obviously, the latter derivatives add to zero, as they should.

5. A random utility-based interpretation of the model

We next give a utility-based interpretation of the qualitative response model. In doing this, we introduce the concept of a *random utility.* The choice of the individuals is not determined deterministically, but as the outcome of a stochastic process.

Assume that individual i's (perceived) utility of taking a decision leading to the response we observe is given by

$$y_i^* = c_0 + x_i\beta - \epsilon_i, \tag{14}$$

where y_i^* is the random utility of individual i, x_i is a vector of observable variables determining the utility, c_0 is a "threshold value" for the utility, *common to all individuals*, and ϵ_i is a disturbance with zero mean, conditional on x_i, *not to be confused with the disturbance u_i in (2).* The utility y_i^*, unlike y_i, is a latent (unobserved) variable, which is assumed to vary continuously. Two individuals with the same x_i vector have the same (expected) y_i^*. Finally, ϵ_i is a stochastic variable, with CDF conditionally on x_i equal to $F(\epsilon_i)$. We assume that ϵ_i has a symmetric distribution, so that it is immaterial whether we assign a positive or negative sign to it in the utility equation (14). The latter turns out to be the most convenient.

What we, as econometricians, observe is *whether or not the latent random utility is larger than or smaller than its "threshold value".* The former leads to a positive response, the latter to a negative response. This is the way the individuals preferences for the commodity or the decision is revealed to us. Hence, (14) implies

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* \geq c_0 \iff \epsilon_i \leq x_i\beta, \\ 0 & \text{if} \quad y_i^* < c_0 \iff \epsilon_i > x_i\beta. \end{cases} \tag{15}$$

This equation formally expresses the binary variable y_i as a *step function* of the latent, continuous variable y_i^*. From (15) we find that the reponse probability can be expressed as follows:

$$P_i = P(y_i = 1) = P(y_i^* \geq c_0) = P(\epsilon_i \leq x_i\beta) = F(x_i\beta), \tag{16}$$

since ϵ_i has CDF given by F.

The above argument gives a rationalization of assumption (7).

1. *If ϵ_i follows the logistic distribution, i.e., if $F(\epsilon_i) = e^{\epsilon_i}/(1+e^{\epsilon_i})$, then we get the Logit model.*

2. *If ϵ_i follows the normal distribution, i.e., if $F(\epsilon_i)$ is the CDF of the N(0,1) distribution, then we get the Probit model.*

Some people (economists, psychologists?) like this utility based interpretation and find that it has intuitive appeal and improves our understanding. Other people (statisticians?) do not like it and think that we can dispense with it. Adhering to the latter view, we may only interpret the model as a way of *parameterizing or endogenizing the response probability.*

Remark: Since the variance in the logistic distribution can be shown to be equal to $\pi^2/3$ and the variance of the standardized normal distribution, by construction, is one, the elements of the coefficient vector β will not have the same order of magnitude. Although the general shape of the two distributions, the bell-shape, is approximately the same (but the logistic has somewhat "thicker tails"), we have to multiply the coefficients of the Probit model by $\pi/\sqrt{3} \approx 1.8$ to make them comparable with those from the Logit model: $\widehat{\beta}_{logit} \approx 1.8\widehat{\beta}_{probit}$.

6. Maximum Likelihood estimation of the Logit model

Assume that we have observations on $(y_i, x_i) = (y_i, 1, x_{1i}, \dots, x_{Ki})$ for individuals $i = 1, \dots, n$. We assume that $\epsilon_1, \dots, \epsilon_n$ are stochastically independent. Then $(y_1|x_1), \dots, (y_n|x_n)$ will also be stochastically independent. Consider now

$$L_i = P_i^{y_i}(1-P_i)^{1-y_i} = \begin{cases} P_i & \text{for} \quad y_i = 1, \\ 1-P_i & \text{for} \quad y_i = 0. \end{cases} \tag{17}$$

We define, in general, the *Likelihood function* as the joint probability (or probability density function) of the endogenous variables conditional on the exogenous variables. In the present case, since the n individual observations are independent, it is the point probability of $(y_1, \dots, y_n)$ conditional on $(x_1, \dots, x_n)$, which is, in view of (17),

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i} = \prod_{\{i:y_i=1\}} P_i \prod_{\{i:y_i=0\}} (1-P_i), \tag{18}$$

where P_i is given by (8) and where $\prod_{\{i:y_i=1\}}$ and $\prod_{\{i:y_i=0\}}$ denotes the product taken across all i such that $y_i = 1$ and such that $y_i = 0$, respectively.

Maximum Likelihood (ML) method is a general method which, as its name indicates, chooses as the estimators of the unknown parameters of the model the values which maximize the likelihood function. Or more loosely stated, the method finds the parameter values which "maximize the probability of the observed outcome". Maximum Likelihood is a very well-established estimation method in econometrics and statistics. In the present case, the ML problem is therefore to maximize L, given by (18) with respect to $\beta_0, \beta_1, \dots, \beta_K$. Since the logarithm function is monotonically increasing, maximizing L is equivalent to maximizing $\ln(L)$, i.e.,

$$\ln(L) = \sum_{i=1}^{n} \ln(L_i) = \sum_{i=1}^{n} [y_i \ln(P_i) + (1-y_i)\ln(1-P_i)], \tag{19}$$

which is a simpler mathematical problem. Since (11) implies

$$\text{(20)} \qquad \frac{\partial \ln(P_i)}{\partial \beta_k} = (1-P_i)x_{ki}, \qquad \frac{\partial \ln(1-P_i)}{\partial \beta_k} = -P_i x_{ki},$$

it follows, after a little algebra, that

$$\text{(21)} \qquad \frac{\partial \ln(L)}{\partial \beta_k} = \sum_{i=1}^{n}(y_i - P_i)x_{ki} = \sum_{i=1}^{n}\left(y_i - \frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)x_{ki},$$

after inserting from (8). The first-order conditions for the ML problem, $(\partial \ln(L))/(\partial \beta_k) = 0, k = 0, 1, \ldots, K$, defining the ML estimators of $\beta_0, \beta_1, \ldots, \beta_K$, are thus

$$\text{(22)} \qquad \sum_{i=1}^{n} y_i x_{ki} = \sum_{i=1}^{n}\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)x_{ki}, \qquad k = 0, 1, \ldots, K.$$

This is a non-linear equation system which can be solved numerically, and the solution is not particularly complicated when implemented on a computer. The ML estimators cannot, however, be expressed in closed form.

From the estimates, we can estimate the effects of changes in the exogenous variables on the response and non-response probabilities for a value of the x_i vector we might choose (e.g. its sample mean), by inserting the estimated β_k's in (13). This may give rise to interesting interpretations of an estimated discrete choice model. We may, for instance, be able to give statements like the following: "A one per cent increase in the price ratio between public and private transport will reduce the probability that an individual (with average characteristics) will use public transport for his/her next trip by p per cent".

Remark: In the particular case where the response probability is the same for all individuals and equal to P for all i, and hence independent of x_i, then we have a *classical binomial situation* with, independent observations, two possible outcomes and constant probability for each of them. Then (21) simplifies to

$$\frac{\partial \ln(L)}{\partial \beta_k} = \sum_{i=1}^{n}(y_i - P),$$

so that the first order conditions become

$$\sum_{i=1}^{n} y_i = n\widehat{P} \iff \widehat{P} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y}{n},$$

where $y = \sum_{i=1}^{n} y_i$ is the number of individuals responding positively. The latter is the familiar estimator for the response probability in a binomial distribution.