Econometrics Block II

Juan Dolado

EUI

November 4, 2014

Outline

Revision: OLS

VOL. 96 NO. 4 ADDA AND CORNAGLIA: TAXES, CIGARETTE CONSUMPTION, AND SMOKING INTENSITY 1021

TABLE 4—DETERMINANTS OF SMOKING AS MEASURED BY LOG OF CIGARETTES, LOG OF COTININE CONCENTRATION AND LOG OF COTININE CONCENTRATION PER CIGARETTE SMOKED

	(1)		(2)		(3)		(4)	
	Log(Cig)[a]		Log(Cot)[a]		Log(Cot/Cig)[a]		Log(Cot/Cig)[b]	
Men	−0.05	(0.040)	−0.11	(0.060)	−0.06	(0.040)	0.11	(0.120)
Age	0.05**	(0.006)	0.05**	(0.008)	−0.01	(0.007)	0.00	(0.021)
Age squared (*100)	−0.1**	(0.001)	−0.04**	(0.010)	0.01	(0.007)	−0.00	(0.022)
Log income	−0.02	(0.026)	−0.05	(0.035)	−0.03	(0.027)	0.03	(0.020)
Education (years)	−0.01	(0.007)	−0.03**	(0.009)	−0.02**	(0.007)	−0.04	(0.061)
House size (number of bedrooms)	−0.04**	(0.009)	−0.09**	(0.010)	−0.05**	(0.009)	—	—
White	0.39**	(0.094)	0.36**	(0.130)	−0.03	(0.100)	0.16	(0.129)
African American	−0.05	(0.102)	0.51**	(0.140)	0.56**	(0.100)	0.64**	(0.140)
Family size	0.01	(0.010)	0.05**	(0.020)	0.04**	(0.010)	—	—
Attending church	−0.17**	(0.030)	−0.08**	(0.040)	0.09**	(0.030)	—	—
Living in urban area	−0.10**	(0.030)	−0.04	(0.041)	0.06*	(0.030)	—	—
Height (inches)	0.01*	(0.005)	0.01	(0.007)	−0.00	(0.006)	−0.01	(0.007)
Married	0.19**	(0.060)	0.10	(0.090)	−0.09	(0.070)	−0.02	(0.101)
Age started smoking	−0.02**	(0.003)	−0.03**	(0.004)	−0.00	(0.003)	−0.00	(0.010)
Filter							0.40	(0.372)
Nicotine yield							0.76**	(0.190)
Length of cigarette (cm)							0.06	(0.051)
Mentholated							0.09	(0.110)
Number of observations	3,424		3,424		3,424		590	

Notes: Robust standard errors in parenthesis. Regression also controls for year and region effects.
[a] Estimation done for years 1988–1994.
[b] Estimation done for 1999.
* Significant at the 10-percent level.
** Significant at the 5-percent level.

Interpretation of Marginal Effects

- **Linear model:** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \dfrac{\partial Y_i}{\partial X_i} = \beta_1$

  Interpretation: When X goes up by 1 unit, Y goes up by β_1 units.

- **Log-Log model** (constant elasticity model):

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i \qquad Y_i = e^{\beta_0} X_i^{\beta_1} e^{\varepsilon_i}$$

$$\frac{\partial Y_i}{\partial X_i} = e^{\beta_0} \beta_1 X_i^{\beta_1 - 1} e^{\varepsilon_i} \qquad \frac{\partial Y_i / Y_i}{\partial X_i / X_i} = \beta_1$$

  Interpretation: When X goes up by **1%**, Y goes up by β_1 **%**.

- **Log-lin model:**

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\begin{aligned} \frac{\partial Y_i}{\partial X_i} &= \beta_1 e^{\beta_0} e^{\beta_1 X_i} e^{\varepsilon_i} \\ \frac{\partial Y_i / Y_i}{\partial X_i} &= \beta_1 \end{aligned}$$

  Interpretation: When X goes up by 1 **unit**, Y goes up by $100\beta_1$ **%**.

Endogeneity and Simultaneity

- Consider the model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- In many problems studied in econometrics it is *not* possible to maintain restrictions requiring that the expected value of the latent variable in an equation is zero given the values of the right hand side variables in the equation:

$$E(\varepsilon|X) \neq 0$$

- This leads to a biased OLS estimate.
- There are many cases in which the OLS identification assumption does not hold:
  - simultaneous equations.
  - explanatory variables measured with error.
  - omitted variables correlated with explanatory variables.

Simultaneity

- **Definition:** Simultaneity arises when the causal relationship between Y and X runs both ways. In other words, the explanatory variable X is a function of the dependent variable Y, which in turn is a function of X.
- This arises in many economic examples:
  - Income and health.
  - Sales and advertizing.
  - Investment and productivity.
- What are we estimating when we run an OLS regression of Y on X? Is it the direct effect, the indirect effect or a mixture of both.

Implications of Simultaneity

- $$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + u_i & \text{(direct effect)} \\ \\ X_i = \alpha_0 + \alpha_1 Y_i + v_i & \text{(indirect effect)} \end{cases}$$
- Replacing the second equation in the first one, we get an equation expressing Y_i as a function of the parameters and the error terms u_i and v_i only. Substituting this into the second equation, we get X_i also as a function of the parameters and the error terms:
$$\begin{cases} Y_i = \dfrac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha 1 \beta_1} + \dfrac{\beta_1 v_i + u_i}{1 - \alpha_1 \beta_1} = B_0 + \tilde{u}_i \\ \\ X_i = \dfrac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \dfrac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1} = A_0 + \tilde{v}_i \end{cases}$$
- This is the **reduced form** of our model. In this rewritten model, Y_i is not a function of X_i and vice versa. However, Y_i and X_i are both a function of the two original error terms u_i and v_i.

Implications of Simultaneity

- Now that we have an expression for X_i, we can compute:

$$\begin{aligned} cov(X_i, u_i) &= cov\left(\frac{\alpha_0 + \alpha_1\beta_0}{1-\alpha_1\beta_1} + \frac{v_i + \alpha_1 u_i}{1-\alpha_1\beta_1}, u_i\right) \\ &= \frac{\alpha_1}{1-\alpha_1\beta_1} \; Var(u_i) \end{aligned}$$

which, in general is different from zero. Hence, with simultaneity, our assumption 1 is violated. **An OLS regression of Y_i on X_i will lead to a biased estimate of β_1**. Similarly, an OLS regression of X_i on Y_i will lead to a biased estimate of α_1.

What are we estimating?

- For the model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- The OLS estimate is:

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + \frac{cov(X_i, u_i)}{Var(X_i)} \\ &= \beta_1 + \frac{\alpha_1}{1 - \alpha_1 \beta_1} \frac{Var(u_i)}{Var(X_i)} \end{aligned}$$

- So
  - $E\hat{\beta}_1 \neq \beta_1$
  - $E\hat{\beta}_1 \neq \alpha_1$
  - $E\hat{\beta}_1 \neq$ an average of β_1 and α_1.

Identification

- Suppose a more general model:

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + u_i \\ X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2 Z_i + v_i \end{cases}$$

- We have two sorts of variables:
  - **Endogenous:** Y_i and X_i because they are determined within the system. They appear on the right and left hand side.
  - **Exogenous:** T_i and Z_i. They are determined outside of our model, and in particular are not caused by either X_i or Y_i. They appear only on the right-hand-side.

Example

- Consider a simple version of the Mincer model for returns to schooling with the following *structural equations*.

$$
\begin{aligned}
W &= \alpha_0 + \alpha_1 S + \alpha_2 Z + \varepsilon_1 \\
S &= \beta_0 + \beta_1 Z + \varepsilon_2
\end{aligned}
$$

  Here W is the log wage, S is years of schooling, Z is some characteristic of the individual, and ε_1 and ε_2 are unobservable latent random variables.

- We might expect those who receive unusually high levels of schooling given Z to also receive unusually high wages given Z and S, a situation that would arise if ε_1 and ε_2 were affected positively by ability, a characteristic not completely captured by variation in Z.

Example

- In this problem we might be prepared to impose the following restrictions.

$$\begin{aligned} E[\varepsilon_1|Z=z] &= 0 \\ E[\varepsilon_2|Z=z] &= 0 \end{aligned}$$

  but not

$$E[\varepsilon_1|S=s, Z=z]=0$$

  unless ε_1 was believed to be uncorrelated with ε_2.

- Considering just the first (W) equation,

$$E[W|S=s, Z=z]=\alpha_0+\alpha_1 s+\alpha_2 z+E[\varepsilon_1|S=s, Z=z]$$

- A variable like S, appearing in a structural form equation and correlated with the latent variable in the equation, is called an *endogenous* variable.

Reduced Form Equations

- Substitute for S in the wage equation:

$$\begin{aligned} W &= (\alpha_0 + \alpha_1 \beta_0) + (\alpha_1 \beta_1 + \alpha_2) Z + \varepsilon_1 + \alpha_1 \varepsilon_2 \\ S &= \beta_0 + \beta_1 Z + \varepsilon_2 \end{aligned}$$

- Equations like this, in which each equation involves exactly one endogenous variable are called *reduced form* equations.
- The restrictions $E[\varepsilon_1 | Z = z] = 0$ and $E[\varepsilon_2 | Z = z] = 0$ imply that

$$\begin{aligned} E[W|Z &= z] = (\alpha_0 + \alpha_1 \beta_0) + (\alpha_1 \beta_1 + \alpha_2) z \\ E[S|Z &= z] = \beta_0 + \beta_1 z \end{aligned}$$

- Given enough (at least 2) distinct values of z and knowledge of the left hand side quantities we can solve for $(\alpha_0 + \alpha_1 \beta_0)$, $(\alpha_1 \beta_1 + \alpha_2)$, β_0 and β_1. So, the values of these *functions* of parameters of the structural equations *can* be identified.

Reduced Form Equations

- In practice we do not know the left hand side quantities but with enough data we can estimate the data generating values of $(\alpha_0 + \alpha_1\beta_0)$, $(\alpha_1\beta_1 + \alpha_2)$, β_0 and β_1, for example by OLS applied first to (W, Z) data and then to (S, Z) data.
- The values of β_0 and β_1 *are* identified but the values of α_0, α_1 and α_2 are *not*, for without further restrictions their values cannot be deduced from knowledge of $(\alpha_0 + \alpha_1\beta_0)$, $(\alpha_1\beta_1 + \alpha_2)$, β_0.

Identification using an Exclusion Restriction

- One restriction we might be prepared to add to the model is the restriction $\alpha_2 = 0$. Whether or not that is a reasonable restriction to maintain depends on the nature of the variable Z.
- If Z were a measure of some characteristic of the environment of the person at the time that schooling decisions were made (for example the parents' income, or some measure of an event that perturbed the schooling choice) then we might be prepared to maintain the restriction that, given schooling achieved (S), Z does not affect W, i.e. that $\alpha_2 = 0$.
- This restriction may be sufficient to identify the remaining parameters. If the restriction is true then the coefficients on Z become $\alpha_1\beta_1$.
- We have already seen that (the value of) the coefficient β_1 is identified. If β_1 is not itself zero (that is Z does indeed affect years of schooling) then α_1 is identified as the ratio of the coefficients on Z in the regressions of W and S on Z. With α_1 identified and β_0 already identified, identification of α_0 follows directly.

Indirect Least Squares Estimation

- Estimation could proceed under the restriction $\alpha_2 = 0$ by calculating OLS (or GLS) estimates of the "reduced form" equations:

$$\begin{aligned} W &= \pi_{01} + \pi_{11} Z + U_1 \\ S &= \pi_{02} + \pi_{12} Z + U_2 \end{aligned}$$

where

$$\begin{aligned} \pi_{01} &= \alpha_0 + \alpha_1 \beta_0 & \pi_{11} &= \alpha_1 \beta_1 \\ \pi_{02} &= \beta_0 & \pi_{12} &= \beta_1 \\ U_1 &= \varepsilon_1 + \alpha_1 \varepsilon_2 & U_2 &= \varepsilon_2 \end{aligned}$$

and

$$E[U_1 | Z = z] = 0 \quad E[U_2 | Z = z] = 0$$

Indirect Least Squares Estimation

solving the equations:

$$\begin{aligned} \hat{\pi}_{01} &= \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\beta}_0 & \qquad \hat{\pi}_{11} &= \hat{\alpha}_1 \hat{\beta}_1 \\ \hat{\pi}_{02} &= \hat{\beta}_0 & \qquad \hat{\pi}_{12} &= \hat{\beta}_1 \end{aligned}$$

given values of the $\hat{\pi}$'s for values of the $\hat{\alpha}$'s and $\hat{\beta}$'s, as follows.

$$\begin{aligned} \hat{\alpha}_0 &= \hat{\pi}_{01} - \hat{\pi}_{02} \left(\hat{\pi}_{11} / \hat{\pi}_{12} \right) & \qquad \hat{\alpha}_1 &= \hat{\pi}_{11} / \hat{\pi}_{12} \\ \hat{\beta}_0 &= \hat{\pi}_{02} & \qquad \hat{\beta}_1 &= \hat{\pi}_{12} \end{aligned}$$

- Estimators obtained in this way, by solving the equations relating structural form parameters to reduced form parameters with OLS estimates replacing the reduced form parameters, are known as *Indirect Least Squares* estimators. They were first proposed by Jan Tinbergen in 1930.

Over Identification

- Suppose that there are *two* covariates, Z_1 and Z_2 whose impact on the structural equations we are prepared to restrict so that *both* affect schooling choice but *neither* affect the wage given the amount of schooling achieved:

$$\begin{aligned} W &= \alpha_0 + \alpha_1 S + \varepsilon_1 \\ S &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon_2 \end{aligned}$$

- the reduced form equations are as follows

$$\begin{aligned} W &= \pi_{01} + \pi_{11} Z_1 + \pi_{21} Z_2 + U_1 \\ S &= \pi_{02} + \pi_{12} Z_1 + \pi_{22} Z_2 + U_2 \end{aligned}$$

where

$$\begin{aligned} \pi_{01} &= \alpha_0 + \alpha_1 \beta_0 & \pi_{11} &= \alpha_1 \beta_1 & \pi_{21} &= \alpha_1 \beta_2 \\ \pi_{02} &= \beta_0 & \pi_{12} &= \beta_1 & \pi_{22} &= \beta_2 \end{aligned}$$

and

$$U_1 = \varepsilon_1 + \alpha_1 \varepsilon_2 \qquad U_2 = \varepsilon_2.$$

Over Identification

- The values of the reduced form equations' coefficients are identified under restrictions .
- Note, there are *two* ways in which the coefficient α_1 can be identified, as follows
$$\alpha_1 = \alpha_1^{Z_1} = \frac{\pi_{11}}{\pi_{12}} \qquad \alpha_1 = \alpha_1^{Z_2} = \frac{\pi_{21}}{\pi_{22}}$$
- In this situation we say that the value of the parameter α_1 is *over identified*.
- We will usually find that $\hat{\alpha}_1^{Z_1} \neq \hat{\alpha}_1^{Z_2}$ even though these are both estimates of the value of the same structural form parameter.
- If the discrepancy was found to be very large then we might doubt whether the restrictions of the model are correct. This suggests that tests of over identifying restrictions can detect misspecification of the econometric model.
- If the discrepancy is not large then there is scope for combining the estimates to produce a single estimate that is more efficient than either taken alone.

Instrumental Variables

- Consider the linear model for an outcome Y given k covariates X

$$Y = X\beta + \varepsilon$$

- Suppose that the restriction $E[\varepsilon|X = x] = 0$ cannot be maintained but that there exist m variables Z for which the restriction $E[\varepsilon|Z = z] = 0$ can be maintained. It implies:

$$E[Y - X\beta|Z = z] = 0$$

and thus that

$$E[Z'(Y - X\beta)|Z = z] = 0$$

which implies that, unconditionally

$$E[Z'(Y - X\beta)] = 0.$$

and thus

$$E[Z'Y] = E[Z'X]\beta.$$

Instrumental Variables

- First suppose $m = k$, and that $E[Z'X]$ has rank k. Then β can be expressed in terms of moments of Y, X and Z as follows

$$\beta = E[Z'X]^{-1}E[Z'Y].$$

and β is (just) identifiable. This leads directly to an analogue type estimator:

$$\hat{\beta} = (Z'X)^{-1}(Z'Y)$$

In the context of the just identified returns to schooling model this is the Indirect Least Squares estimator.

Special Case: Wald Estimator

- A special (and simple case) consists of a binary instrument. $Z \in \{0,1\}$. In this case, the IV estimator simplifies to:

$$\hat{\beta}_{Wald} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(X|Z=1) - E(X|Z=0)}$$

- This estimator is just a function of conditional means that are easy to compute with simple statistical tools. For instance if Y are earnings, X is college attendance and Z an indicator of living close to a college, the numerator is just the difference in earning of those close or far away to a college, and the denominator is the difference in college attendance for those close of far away from the college.

Wald Estimator

- Note that $\bar{Z} = N_1/N$.
- Proof:

$$
\begin{aligned}
Z'Y &= \frac{1}{N}\sum_i (Z_i - \bar{Z})(Y_i - \bar{Y}) \\
&= \frac{-\bar{Z}}{N}\sum_{I_0}(Y_i - \bar{Y}) + \frac{1-\bar{Z}}{N}\sum_{I_1}(Y_i - \bar{Y}) \\
&= \frac{-N_1}{N}\frac{N_0}{N}\bar{Y}_0 + \frac{N_1 N_0}{N^2}\bar{Y} + \frac{N_0}{N}\frac{N_1}{N}\bar{Y}_1 - \frac{N_0 N_1}{N^2}\bar{Y} \\
&= \frac{N_0 N_1}{N^2}(\bar{Y}_1 - \bar{Y}_0)
\end{aligned}
$$

Similarly, $Z'X = \frac{N_0 N_1}{N^2}(\bar{X}_1 - \bar{X}_0)$

Example: Return to Schooling

- What is the effect of additional years of schooling on wages?
- One can regress wages on schooling, but schooling may be endogenous.
- Angrist and Krueger (1991) "Does Compulsory School Attendance Affect Schooling and Earnings?" QJE, introduce an IV method, based on particular features of the US school system:
  - Children entering school have to be 6 by January 1rst.
  - Children have to remain in school until their sixteenth birthday.
  - Children born earlier in the year enter school later.

Example: Return to Schooling

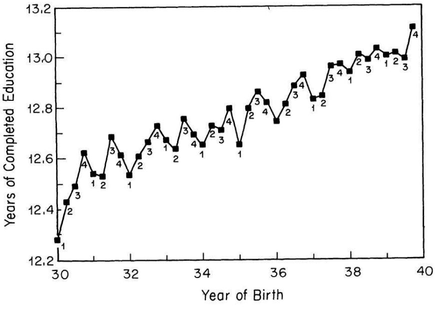


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

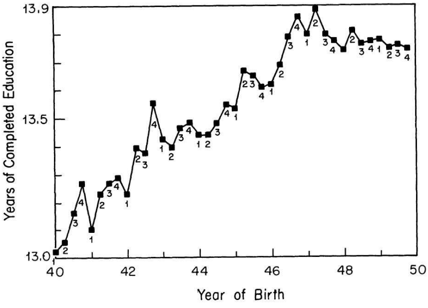


FIGURE II
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Example: Return to Schooling

TABLE I
THE EFFECT OF QUARTER OF BIRTH ON VARIOUS EDUCATIONAL OUTCOME VARIABLES

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]
Completed master's degree	1930–1939	0.09	−0.001	0.002	−0.001	1.7
			(0.001)	(0.001)	(0.001)	[0.1599]
	1940–1949	0.11	0.000	0.004	0.001	3.9
			(0.001)	(0.001)	(0.001)	[0.0091]
Completed doctoral degree	1930–1939	0.03	0.002	0.003	0.000	2.9
			(0.001)	(0.001)	(0.001)	[0.0332]
	1940–1949	0.04	−0.002	0.001	−0.001	4.3
			(0.001)	(0.001)	(0.001)	[0.0050]

Example: Return to Schooling

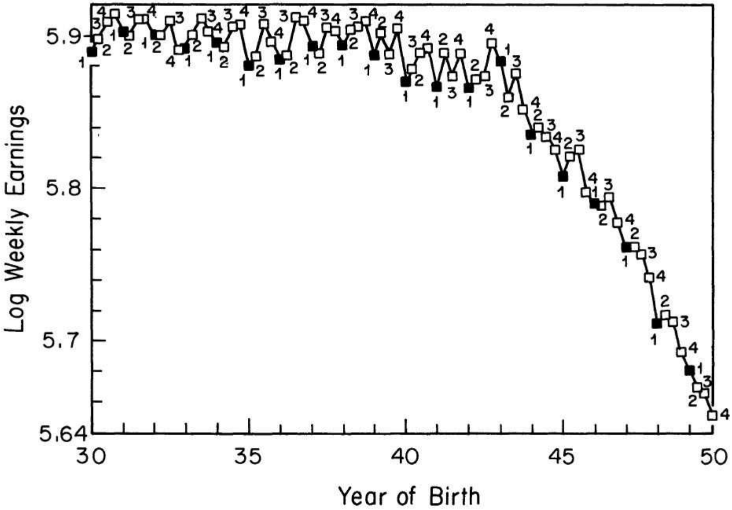


FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

Example: Return to Schooling: Wald Estimator

Panel B: Wald Estimates for 1980 Census—Men Born 1930–1939

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.8916	5.9027	−0.01110
			(0.00274)
Education	12.6881	12.7969	−0.1088
			(0.0132)
Wald est. of return to education			0.1020
			(0.0239)
OLS return to education			0.0709
			(0.0003)

a. The sample size is 247,199 in Panel A, and 327,509 in Panel B. Each sample consists of males born in the United States who had positive earnings in the year preceding the survey. The 1980 Census sample is drawn from the 5 percent sample, and the 1970 Census sample is from the State, County, and Neighborhoods 1 percent samples.

b. The OLS return to education was estimated from a bivariate regression of log weekly earnings on years of education.

Generalised Method of Moments estimation

- Suppose that $m > k$. We will not find a solution since we have $m > k$ equations in k unknowns.
- Define a family of estimators, $\hat{\beta}_W$ as

$$\hat{\beta}_W = \arg\min_{\beta} \left(Z'Y - Z'X\beta\right)' W \left(Z'Y - Z'X\beta\right)$$

  where W is a $m \times m$ full rank, positive definite symmetric matrix.
- This M-estimator is an example of what is known as the *Generalised Method of Moments* (GMM) estimator.
- Different choices of W lead to different estimators unless $m = k$.
- The choice among these is commonly made by considering their accuracy. We consider the limiting distribution of the GMM estimator for alternative choices of W and choose W to minimise the variance of the limiting distribution of $n^{1/2}(\hat{\beta}_W - \beta_0)$.
- In standard cases this means choosing W to be proportional to a consistent estimator of the inverse of the variance of the limiting distribution of $n^{1/2}\left(Z'Y - Z'X\beta\right)$.

Generalised Instrumental Variables Estimation

- Write $\hat{\beta}_W$ explicitly in terms of sample moments:

$$\hat{\beta}_W = \arg\min_{\beta} \left(\frac{Z_n' y_n - Z_n' X_n \beta}{n^{1/2}} \right)' W \left(\frac{Z_n' y_n - Z_n' X_n \beta}{n^{1/2}} \right)$$

- Consider what the (asymptotically) efficient choice of W is by examining the variance of $n^{-1/2}(Z_n' y_n - Z_n' X_n \beta)$.
- We have, since $y_n = X_n \beta + \varepsilon_n$,

$$n^{-1/2}(Z_n' y_n) - n^{-1/2}(Z_n' X_n)\beta = n^{-1/2}(Z_n' \varepsilon_n)$$

and *if* we suppose that $Var(\varepsilon_n | Z_n) = \sigma^2 I_n$,

$$Var\left(n^{-1/2}(Z_n' \varepsilon_n) | Z_n \right) = \sigma^2 (n^{-1} Z_n' Z_n).$$

This suggests choosing $W = (n^{-1} Z_n' Z_n)^{-1}$ leading to the following minimisation problem:

$$\hat{\beta}_n = \arg\min_{\beta} \left(Z_n' y_n - Z_n' X_n \beta \right)' \left(Z_n' Z_n \right)^{-1} \left(Z_n' y_n - Z_n' X_n \beta \right)$$

Generalised Instrumental Variables Estimation

- The first order conditions for this problem, satisfied by $\hat{\beta}_n$ are:

$$2\hat{\beta}_n'(X_n'Z_n)(Z_n'Z_n)^{-1}(Z_n'X_n) - 2(X_n'Z_n)(Z_n'Z_n)^{-1}(Z_n'y_n) = 0$$

  leading to the following estimator.

$$\hat{\beta} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}Z'y$$

  This is known as the *generalised instrumental variable estimator* (GIVE). Also know as 2SLS.

Variance of 2SLS Estimator

- By noting $P_Z = Z(Z'Z)^{-1}Z'$, we can rewrite the estimator as:

$$\hat{\beta}_{2SLS} = (X'P_ZX)^{-1}X'P_ZY$$

- The 2SLS estimator can be shown to be asymptotically normal distributed with estimated asymptotic variance

$$V(\hat{\beta}_{2SLS}) = N(X'P_ZX)^{-1}[X'Z(Z'Z)^{-1}\hat{\Sigma}(Z'Z)^{-1}Z'X](X'P_ZX)^{-1}$$

with $\hat{\Sigma} = N^{-1}\sum_i \hat{\epsilon}_i^2 z_i z_i'$

- Under homoskedasticity,

$$V(\hat{\beta}_{2SLS}) = \sigma^2(X'P_ZX)^{-1}$$

GIVE and Two Stage OLS

- Suppose there is a model for X,

$$X = Z\Phi + V$$

where $E[V|Z] = 0$. The OLS estimator of Φ is

$$\hat{\Phi}_n = \left(Z_n' Z_n\right)^{-1} Z_n' X_n$$

and the "predicted value" of X for a given Z is

$$\hat{X}_n = Z_n \left(Z_n' Z_n\right)^{-1} Z_n' X_n.$$

Note that

$$\hat{X}_n' \hat{X}_n = X_n' Z_n (Z_n' Z_n)^{-1} Z_n' X_n$$

and

$$\hat{X}_n' y_n = X_n' Z_n (Z_n' Z_n)^{-1} Z_n' y_n.$$

So the Generalised Instrumental Variables Estimator can be written as

$$\hat{\beta}_n = \left(\hat{X}_n' \hat{X}_n\right)^{-1} \hat{X}_n' y_n.$$

that is, as the OLS estimator of the coefficients of a linear relationship between y_n and the *predicted values* of X_n got from OLS estimation of a linear relationship between X_n and the instrumental variables Z_n.

Example: Return to Schooling: TSLS

To improve efficiency of the estimates and control for age-related trends in earnings, we estimated the following TSLS model:

(1) $$E_i = X_i\pi + \Sigma_c Y_{ic}\delta_c + \Sigma_c \Sigma_j Y_{ic} Q_{ij} \theta_{jc} + \epsilon_i$$

(2) $$\ln W_i = X_i\beta + \Sigma_c Y_{ic}\xi_c + \rho E_i + \mu_i,$$

where E_i is the education of the ith individual, X_i is a vector of covariates, Q_{ij} is a dummy variable indicating whether the individual was born in quarter j ($j = 1,2,3$), and Y_{ic} is a dummy variable indicating whether the individual was born in year c ($c = 1, \ldots, 10$), and W_i is the weekly wage. The coefficient ρ is the return to education. If the residual in the wage equation, μ, is correlated with years of education due to, say, omitted variables, OLS estimates of the return to education will be biased.

Example: Return to Schooling: OLS and TSLS

TABLE V
OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1930–1939: 1980 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0711	0.0891	0.0711	0.0760	0.0632	0.0806	0.0632	0.0600
	(0.0003)	(0.0161)	(0.0003)	(0.0290)	(0.0003)	(0.0164)	(0.0003)	(0.0299)
Race (1 = black)	—	—	—	—	−0.2575	−0.2302	−0.2575	−0.2626
					(0.0040)	(0.0261)	(0.0040)	(0.0458)
SMSA (1 = center city)	—	—	—	—	0.1763	0.1581	0.1763	0.1797
					(0.0029)	(0.0174)	(0.0029)	(0.0305)
Married (1 = married)	—	—	—	—	0.2479	0.2440	0.2479	0.2486
					(0.0032)	(0.0049)	(0.0032)	(0.0073)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	−0.0772	−0.0801	—	—	−0.0760	−0.0741
			(0.0621)	(0.0645)			(0.0604)	(0.0626)
Age-squared	—	—	0.0008	0.0008	—	—	0.0008	0.0007
			(0.0007)	(0.0007)			(0.0007)	(0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]	—	22.5 [29]	—	19.6 [27]

a. Standard errors are in parentheses. Sample size is 329,509. Instruments are a full set of quarter-of-birth times year-of-birth interactions. The sample consists of males born in the United States. The sample is drawn from the 5 percent sample of the 1980 Census. The dependent variable is the log of weekly earnings. Age and age-squared are measured in quarters of years. Each equation also includes an intercept.

Example: Does Trade cause Growth?

- Jeffrey A. Frankel and David Romer (1999) "Does Trade cause Growth?", *AER*.
- Regressing growth rates on trade may lead to biased estimates. Issues with omitted variables.
- Idea: instrument trade with geographical information: distance to trade partners, whether country is landlocked, whether they share a border.
- They construct a prediction of trade, based on these variables.

Example: Does Trade cause Growth?

First stage equation:

$$\ln(\tau_{ij}/GDP_i) = \alpha X_{ij} + v_i$$

TABLE 1—THE BILATERAL TRADE EQUATION

	Variable	Interaction
Constant	−6.38	5.10
	(0.42)	(1.78)
Ln distance	−0.85	0.15
	(0.04)	(0.30)
Ln population	−0.24	−0.29
(country *i*)	(0.03)	(0.18)
Ln area	−0.12	−0.06
(country *i*)	(0.02)	(0.15)
Ln population	0.61	−0.14
(country *j*)	(0.03)	(0.18)
Ln area	−0.19	−0.07
(country *j*)	(0.02)	(0.15)
Landlocked	−0.36	0.33
	(0.08)	(0.33)
Sample size	3220	
R^2	0.36	
SE of regression	1.64	

Notes: The dependent variable is $\ln(\tau_{ij}/\mathrm{GDP}_i)$. The first column reports the coefficient on the variable listed, and the second column reports the coefficient on the variable's interaction with the common-border dummy. Standard er-

Example: Does Trade cause Growth?

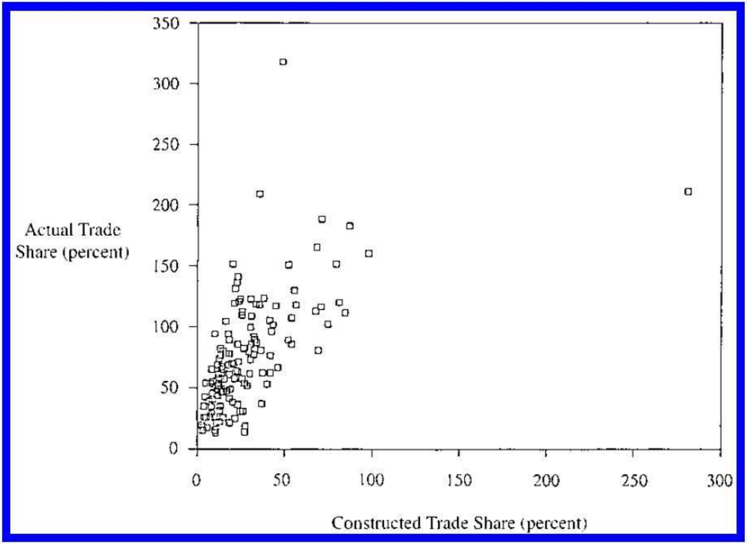


FIGURE 1. ACTUAL VERSUS CONSTRUCTED TRADE SHARE

Example: Does Trade cause Growth?

- The equation of interest is:

$$\ln Y_i = \beta_0 + \beta_T T_i + X_i \beta_X + u_i$$

where Y_i is income per capita in country i and T_i is the trade share.

TABLE 3—TRADE AND INCOME

	(1)	(2)	(3)	(4)
Estimation	OLS	IV	OLS	IV
Constant	7.40	4.96	6.95	1.62
	(0.66)	(2.20)	(1.12)	(3.85)
Trade share	0.85	1.97	0.82	2.96
	(0.25)	(0.99)	(0.32)	(1.49)
Ln population	0.12	0.19	0.21	0.35
	(0.06)	(0.09)	(0.10)	(0.15)
Ln area	−0.01	0.09	−0.05	0.20
	(0.06)	(0.10)	(0.08)	(0.19)
Sample size	150	150	98	98
R^2	0.09	0.09	0.11	0.09
SE of regression	1.00	1.06	1.04	1.27
First-stage F on excluded instrument		13.13		8.45

Notes: The dependent variable is log income per person in 1985. The 150-country sample includes all countries for

Examples: Measurement Errors

- Suppose we are measuring the impact of income, X, on consumption, Y. The true model is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\beta_0 = 0, \ \ \beta_1 = 1$$

- Suppose we have two measures of income, both with measurement errors.
  - $\check{X}_{1i} = X_i + v_{1i}, \ \ s.d.(v_{1i}) = 0.2 * \bar{Y}$
  - $\check{X}_{2i} = X_i + v_{2i}, \ \ s.d.(v_{2i}) = 0.4 * \bar{Y}$

  If we use $\check{X}_2$ to instrument $\check{X}_1$, we get:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (\check{X}_{2i} - \bar{\check{X}}_2)(Y_i - \bar{Y})}{\sum_{i=1}^{N} (\check{X}_{2i} - \bar{\check{X}}_2)(\check{X}_{1i} - \bar{\check{X}}_1)}$$

Examples: Measurement Errors

- Results:

Method	Estimate of β_1
OLS regressing Y on $\check{X}_1$	0.88
OLS regressing Y on $\check{X}_2$	0.68
IV, using $\check{X}_2$ as instrument	0.99

IV with Potentially Heterogeneous Outcomes

- Up to now, we have assumed that the effect of X on Y is common to all individuals.
- We are now considering the case with heterogenous outcomes:

$$Y_i = \alpha + \beta_i X_i + u_i$$

- Suppose that we have an instrument Z_i. What is the IV methodology uncovering?
- To fix ideas, we could be interested in the effect of college ($X_i = 1$) on earnings (Y_i), where Z_i is an indicator variable equal to 1 if the individual is eligible for a voucher to pay for part of the tuition fees.

The Setup

- Suppose we are interested in an outcome Y_i, as a function of a variable X_i. Suppose to make things simple that X_i takes only two values, 0 or 1 (a dummy variable, termed also the treatment). This variable is potentially endogenous.
- Suppose that we have an instrument Z_i, which also takes two values, 0 or 1. Define $X_{0i} = X_i | Z_i = 0$, we can write:

$$X_i = X_{0i} + (X_{1i} - X_{0i})Z_i$$

  which implies:

$$X_i = \pi_0 + \pi_{1i} Z_i + \xi_i$$

  with $\pi_{1i} = (X_{1i} - X_{0i})$ and $\pi_0 = E(X_{0i})$.
- Denote by $Y_i(x, z)$ the outcome of individual i when $X_i = x$ and $Z_i = z$.

Assumption 1: Independence

- The instrument Z_i is independent of the potential outcomes and of the explanatory variable X_i. We write it:

$$\{Y_i(X_{1i}, 1), Y_i(X_{0i}, 0), X_{1i}, X_{0i}\} \perp Z_i$$

- Note that this assumption is enough to interpret causally the reduced form, the regression of Y_i on Z_i.
- In our example, the voucher should be given randomly, and not specifically to those who would go to college any way, nor to those with higher productivity in the labor market.

Assumption 2: Exclusion restriction

- Y_i is a function of only X_i. In other words, Z_i does not affect Y_i directly, but only through its effect on X_i
- If this is verified, we can write more simply: $Y_i = Y_i(X_i)$ as the value of Z_i is irrelevant once we know X_i.

$$\begin{aligned} Y_i &= Y_i(0, Z_i) + [Y_i(1, Z_i) - Y_i(0, Z_i)]X_i \\ &= Y_{i0} + [Y_i(1) - Y_i(0)]X_i \\ &= \alpha + \beta_i X_i + u_i \end{aligned}$$

- In our example, earnings depend only on going to college or not. The fact that one received a voucher has to be irrelevant. Hence the instrument cannot be a prize based on merit, which could be observable by the employer and give some further signal about quality, over and above college education.

Assumption 3: First stage

- We assume that $E(X_{1i} - X_{0i}) \neq 0$, which means that the instrument (Z_i) actually influence the explanatory variable X_i.
- In our example, the voucher has to induce students to go to college. Otherwise, there is no variation in the first stage and the IV would break down.

Assumption 4: Monotonicity

- $X_{1i} - X_{0i} \geq 0$ or $X_{1i} - X_{0i} \leq 0$, $\forall i$. This means that the instrument affects every one in the same way.
- In our example, it means that the voucher would only induce more people to go to college. What we are ruling out is that someone who would have gone to college without a voucher, would decide not to if he/she was given a voucher. This population is sometimes called the *defiers*.
- This is an assumption that is not needed in the homogenous case.

LATE Theorem

- Suppose the following properties hold true:
  - independence: $\{Y_i(X_{1i}, 1), Y_i(X_{0i}, 0), X_{1i}, X_{0i}\} \perp Z_i$,
  - exclusion: $Y_i(x, 0) = Y_i(x, 1) \forall d$,
  - the existence of a first stage: $E(X_{1i} - X_{0i}) \neq 0$,
  - monotonicity: $X_{1i} - X_{0i} \geq 0$ or $X_{1i} - X_{0i} \leq 0$, $\forall i$,
- LATE theorem:

$$
\begin{aligned}
LATE &= \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(X_i|Z_i = 1) - E(X_i|Z_i = 0)} \\
&= E(Y_{i1} - Y_{i0}|X_{1i} > X_{0i}) \\
&= E(\beta_i|\pi_{1i} > 0)
\end{aligned}
$$

- What IV identifies is not necessarily $E(\beta_i)$.

Proof

- From the independence assumption:
  $E[Y_i|Z_i = 1] = E[Y_i(0) + [Y_i(1) - Y_i(0)]X_{1i}]$ and
  $E[Y_i|Z_i = 0] = E[Y_i(0) + (Y_i(1) - Y_i(0))X_{0i}]$ so that the numerator is:

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[Y_i(1) - Y_i(0))(X_i(1) - X_i(0)]$$

- The monotonicity assumption states that $(X_i(1) - X_i(0))$ is either one or zero.

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[Y_i(1) - Y_i(0)|X_{1i} > X_{0i}]P(X_i(1) > X_i(0))$$

- A similar arguments leads to:

$$E(X_i|Z_i = 1) - E(X_i|Z_i = 0) = E[X_i(1) - X_i(0)] = P(X_i(1) > X_i(0))$$

Interpretation of LATE

- Note that the IV estimates depends on the IV we choose. Different IV will give us different results. This is because the population for which $X_{1i} > X_{0i}$ depends on the instrument at hand. Take for instance the example of a school voucher. If the voucher is equal to €100, many students would be indifferent, and only the ones who are most successful at learning will enrol in college. In this case, the LATE estimate will be large. As the voucher is increased, say to €10,000, many more students with lower ability will enrol, so that the LATE estimate will be lower.
- The IV uncover the causal effect for the population under study. This is called internal validity. However, this estimate may have little to say about the effect of schooling in general $(E(\beta_i))$ or for a different experiment (external validity).

Section 3

Regression Discontinuity

Regression Discontinuity (RD)

- The basic idea is to exploit a local discontinuity in the explanatory variable. It requires the existence of a discontinuity at $X = x$.
- RD estimates the local average treatment effect (LATE) of the treatment at a given point.
- Under suitable assumptions, RD is like a randomized experiment at that cutpoint.
- Requires minimal assumptions to get a causal estimate, but it requires to have a sharp discontinuity in the data, at a point which is economically interesting.

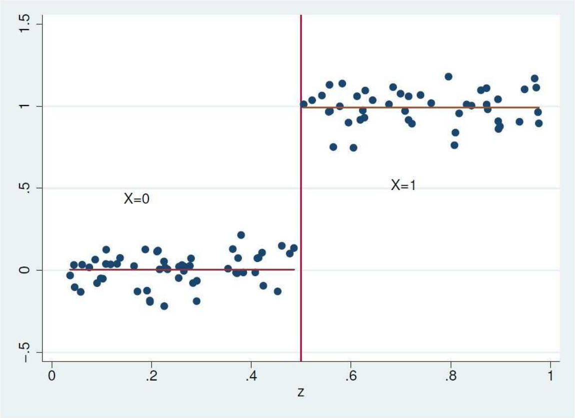
Formal Description

- Denote by Y_i the outcome variable, and by X_i the explanatory one, which can take only two values $\{0, 1\}$. We call Y_{0i} and Y_{1i} the outcome for individual i under the two possible regimes.
- Note that in the data we only observe one of these outcomes, so that we cannot compute $E(Y_{1i} - Y_{0i})$ the average treatment effect.
- Suppose that there exists a continuous variable Z_i such as:

$$X_i = 1\{Z_i \geq c\}$$

- We assume that $E(Y_{1i}|Z)$ and $E(Y_{0i}|Z)$ are continuous in Z at the cutoff point c.

Graphical example



Defining the estimator

- We define the causal effect of X on Y as:

$$\tau_{RD} = E[Y_{1i} - Y_{0i} | Z_i = c]$$

- This cannot be computed with data, but we can compute something close to it:

$$\begin{aligned}
\hat{\tau}_{RD} &= \lim_{\varepsilon \to 0^+} E[Y_i | Z_i = c + \varepsilon] - \lim_{\varepsilon \to 0^-} E[Y_i | Z_i = c + \varepsilon] \\
&= \lim_{\varepsilon \to 0^+} E[Y_{1i} | Z_i = c + \varepsilon] - \lim_{\varepsilon \to 0^-} E[Y_{0i} | Z_i = c + \varepsilon]
\end{aligned}$$

- Some extrapolation is required because by design there are no individuals observed at the exact threshold. In other words, there is no overlap of individuals with similar Zs.

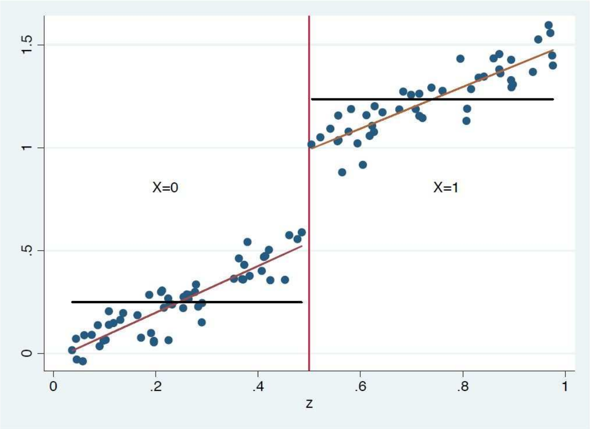
Practical considerations

- In practice, the extrapolation may impact on the results. With few data points on each side of the discontinuity, the estimation will rely on points too far off. The model may include a trend, which will confound the results:

$$Y_i = \alpha_0 + \alpha_1 Z_i + \beta X_i + u_i$$

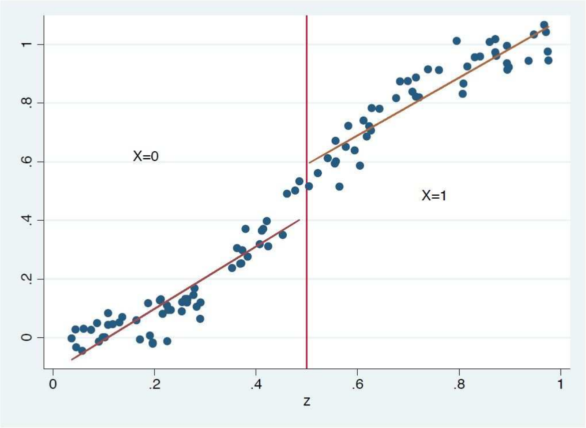
  In this case, we need to perform two separate regressions, for X_i below and above the cutoff, of Y_i on Z_i. This can be generalized to a case where the effect of Z_i is non linear but (sufficiently) smooth. However, if the non-linearity is at the discontinuity, one may get spurious effects.
- To evaluate the plausibility of uncovering a causal estimate, one have to show that other explanatory variables do not jump around the discontinuity. We need to ensure that they are smooth.

Graphical example 2



Note: The DGP is: $y_i = \alpha_0 + \alpha_1 Z_i + \beta X_i + u_i$

Graphical example 3



Note: The DGP is in fact: $y_i = \Phi(\alpha(z_i - \gamma)) + \beta X_i + u_i$ with $\beta = 0$

Example

- Card, Chetty and Weber (2007) *QJE* "CASH-ON-HAND AND COMPETING MODELS OF INTERTEMPORAL BEHAVIOR: NEW EVIDENCE FROM THE LABOR MARKET"
- Investigate the effect of severance payment on subsequent unemployment duration. In a permanent income model, the payment should have almost no effect on behavior.
- Individuals in work for more than 3 years are entitled to about €2,300 when laid-off.

Duration of unemployment

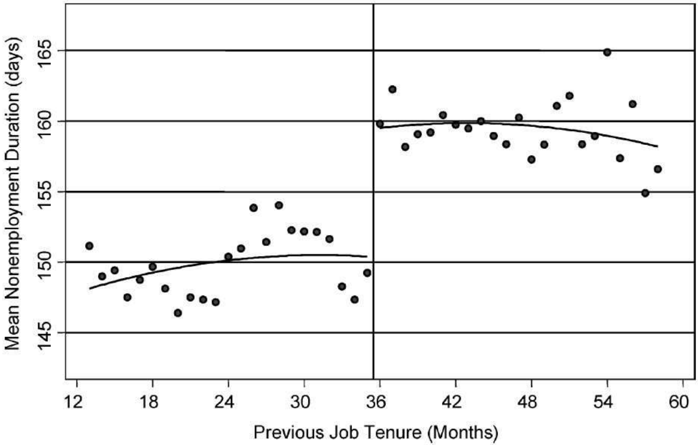


FIGURE V
Effect of Severance Pay on Nonemployment Durations

Card, Chetty and Weber (2007)

Prior Number of Layoffs

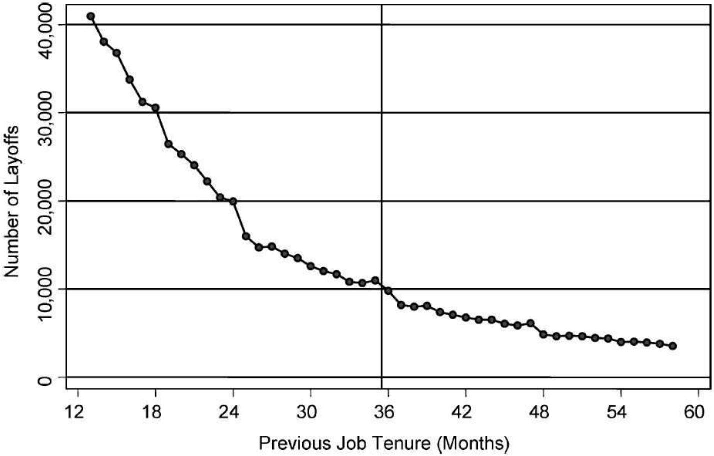


FIGURE II
Frequency of Layoffs by Job Tenure

Card, Chetty and Weber (2007)

Prior Number of Jobs

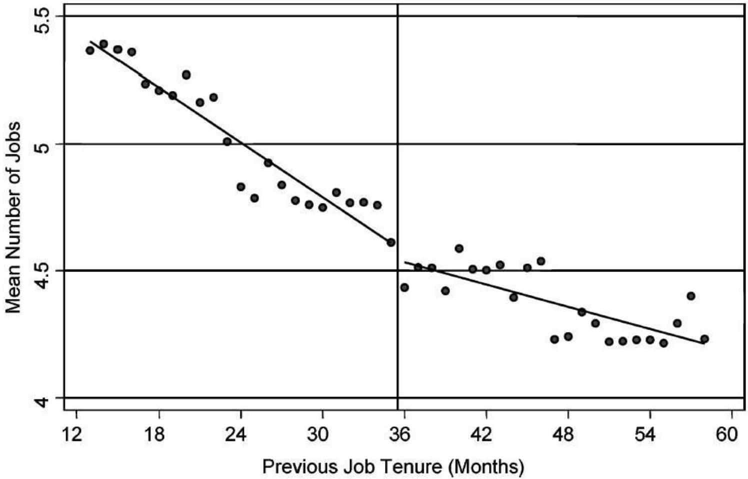


FIGURE IIIa
Number of Jobs Held by Job Tenure

Card, Chetty and Weber (2007)

Prior Annual Wage

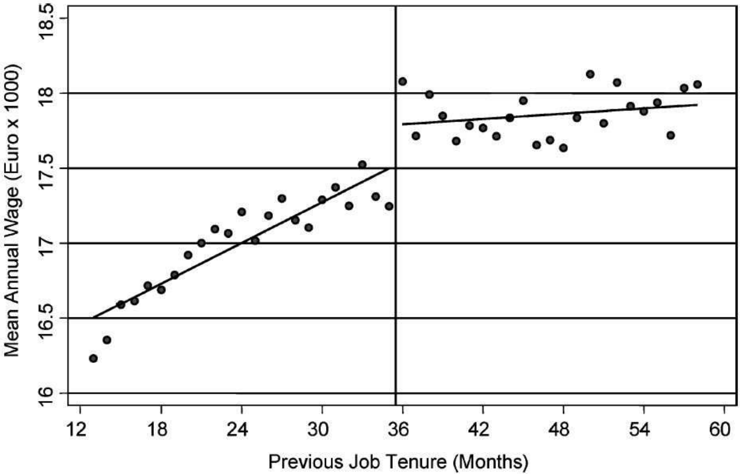


FIGURE IIIb
Wage by Tenure

Card, Chetty and Weber (2007)

Motivation

- Most of econometrics work is to assess the average effect.
- Sometimes, we are interested in a more general characterization of a policy. For example, what is the effect of a training program for unemployed individuals on the duration of unemployment.
  - The mean effect may be small.
  - The program may lengthen short duration spells and shorten very long spells.
- To investigate whether heterogenous effects exist, we need to extend the regression framework to investigate the effect at various points of the distribution.

Definition

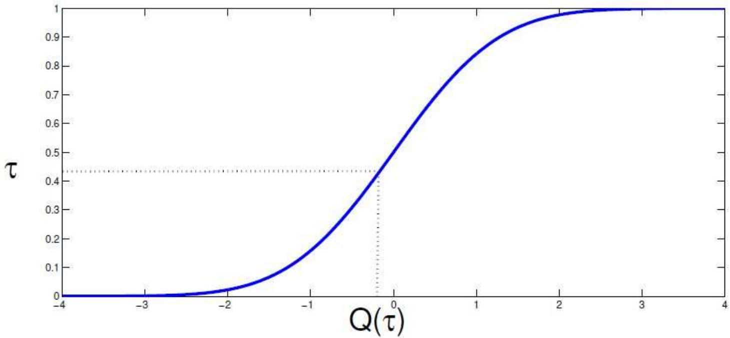
- Let Y be a real valued random variable:
- We can characterize it by its distribution function (or cdf):

$$F(y) = Prob(Y \leq y)$$

- or by its τth quantile of Y: for any $0 < \tau < 1$

$$Q(\tau) = \inf\{y : F(y) \geq \tau\}$$

Definition

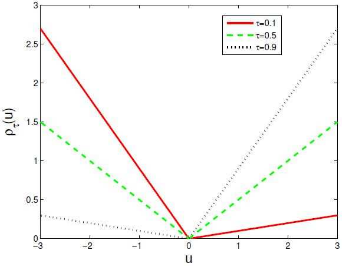


- For instance, $Q(1/2)$ is the median.

Quantiles and the Check Function

- Define the check function:

$$\rho_\tau(u) = u.(\tau - I_{u<0})$$



- The quantiles are the solution to a simple optimization problem

$$E\rho_\tau(Y - \hat{y}) = (\tau - 1)\int_{-\infty}^{\hat{y}} (y - \hat{y})dF(y) + \tau \int_{\hat{y}}^{\infty} (y - \hat{y})dF(y)$$

- Minimizing on $\hat{y}$ we get the FOC:

$$0 = (1 - \tau)\int_{-\infty}^{\hat{y}} dF(y) - \tau \int_{\hat{y}}^{\infty} dF(y) = F(\hat{y}) - \tau$$

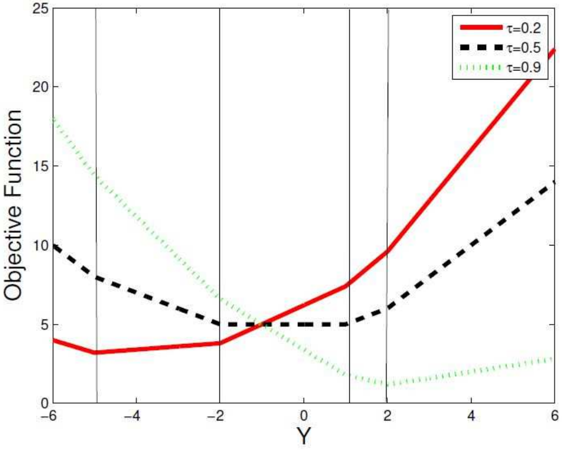
Empirical Estimator

- Suppose we have N data $\{Y_i\}$. The empirical objective function which defines the quantile is

$$\hat{\xi}_\tau = \arg\min_{\xi \in IR} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

- This function is piecewise linear and non-differentiable.
- The optimum may be flat: quantile not necessarily point-identified (in case you have ties).
- When there is point-identification, the solution is one of the observation.

Shape of the Objective Function



- Data $Y = \{-5, -2, 1, 2\}$.

Quantile Regression

- Suppose that we now have observable characteristics X.
- Instead of finding a unique value ξ_τ for each group of individual with similar X, we may want to explore how this value varies with different explanatory variables.

$$\xi_\tau = X\beta_\tau$$

- The objective function is then

$$R(\beta_\tau) = \sum_{i=1}^{N} \rho_\tau(Y_i - X_i\beta_\tau)$$

- It is differentiable except at the points where $Y_i = X_i\beta_\tau$.

Link with Standard Models: Homoskedasticity

- Consider the model with homoskedasticity:

$$Y_i = X_i\beta + u_i \qquad u_i \sim \mathcal{N}(0, \sigma^2)$$

- We have that

$$P(u < \sigma\Phi^{-1}(\tau)|X) = \tau$$

$$P(Y_i < X\beta + \sigma\Phi^{-1}(\tau)|X) = \tau$$

so

$$Q_\tau(Y|X) = X\beta + \sigma\Phi^{-1}(\tau)$$

In this model, the effect of X is the same for all quantiles, only the intercept differs with τ.

Link with Standard Models: Heteroskedasticity

- Consider the model with heteroskedasticity:

$$Y_i = X_i\beta + u_i \qquad u_i \sim \mathcal{N}(0, \sigma^2(X)) \qquad \sigma^2(X) = (X\lambda)^2$$

- We have that

$$P(u < X\lambda\Phi^{-1}(\tau)|X) = \tau$$

$$P(Y_i < X\beta + X\lambda\Phi^{-1}(\tau)|X) = \tau$$

so

$$Q_\tau(Y|X) = X\beta + X\lambda\Phi^{-1}(\tau) = X(\beta + \lambda\Phi^{-1}(\tau))$$

In this model, the effect of X differs across quantiles.

Properties of Quantiles

- Quantiles are equivariant to monotone transformations. That is,

$$Q_{h(Y)|X}(\tau|X) = h\left(Q_{Y|X}(\tau|X)\right)$$

  for any monotone function $h()$. For example, the conditional median of log earnings is the log of the conditional median of earnings. This follows from the fact that:

$$Prob(T < t|X) = Prob(h(T) < h(t)|X)$$

  Note that this is not true for means: $E(h(T)|X) \neq h(E(T|X))$.
- Quantiles are robust to outliers on Y.
- Median regression estimators can be more efficient than mean regression estimators when the error term is non-normal.
- Quantile regression allows one to detect heteroskedasticity.

Asymptotic Distribution

- For a given quantile τ:

$$Y_i = X_i\beta_\tau + u_{i,\tau} \qquad Quant_\tau(u_{i,\tau}|X_i) = 0$$

- Define the quantile residual as: $\hat{u}_i = Y_i - X_i\hat{\beta}_\tau$. Under weak conditions, $\sqrt{N}(\hat{\beta}_\tau - \beta_\tau)$ is asymptotically normal with variance $A^{-1}BA^{-1}$.

$$A = E[f_u(0|X_i)X_i'X_i]$$

and

$$B = \tau(1-\tau)E(X_i'X_i)$$

- A consistent estimator of A is:

$$\hat{A} = (2Nh_N)^{-1}\sum_{i=1}^{N} I\left(|\hat{u}_{i,\tau}| \leq h_N\right)X_i'X_i$$

where $\{h_N > 0\}$ is a nonrandom sequence shrinking to zero as $N \to \infty$. For instance $h_N = aN^{-1/3}$, $a > 0$.

Asymptotic Distribution

- If u_i and X_i are independent,

$$Avar(\sqrt{N}(\hat{\beta}_\tau - \beta_\tau) = \frac{\tau(1-\tau)}{[f_u(0)]^2} E(X_i' X_i)^{-1}$$

and $Avar(\hat{\beta}_\tau)$ is estimated as:

$$\widehat{Avar(\hat{\beta}_\tau)} = \frac{\tau(1-\tau)}{[f_u(0)]^2} \left(N^{-1} \sum_{i=1}^{N} X_i' X_i \right)^{-1}$$

with

$$\hat{f}_u(0) = (2Nh_N)^{-1} \sum_{i=1}^{N} I[|\hat{u}_{i,\tau}| \leq h_N]$$

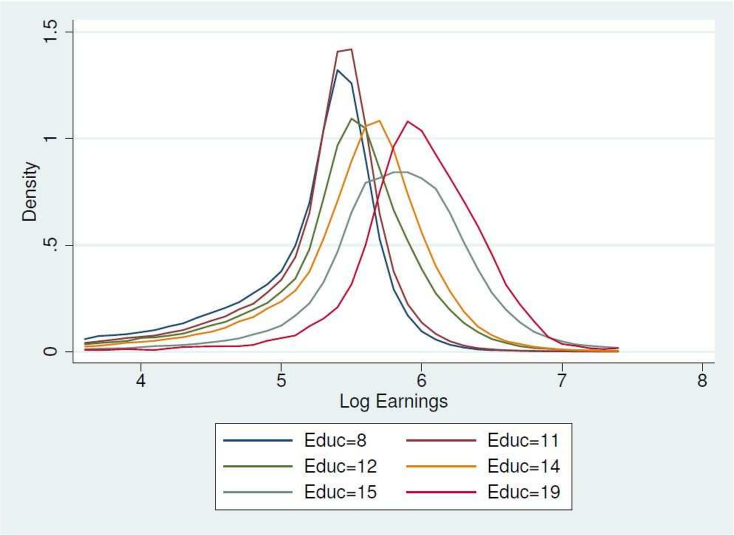
Example: Return to Education

- Data from the Swedish Inland Revenue, merged with Education Register.
- Data on 573,247 men. We record their mean earnings in 4 years and the number of years of education.
- How does schooling affect earnings? Its mean? Its distribution?

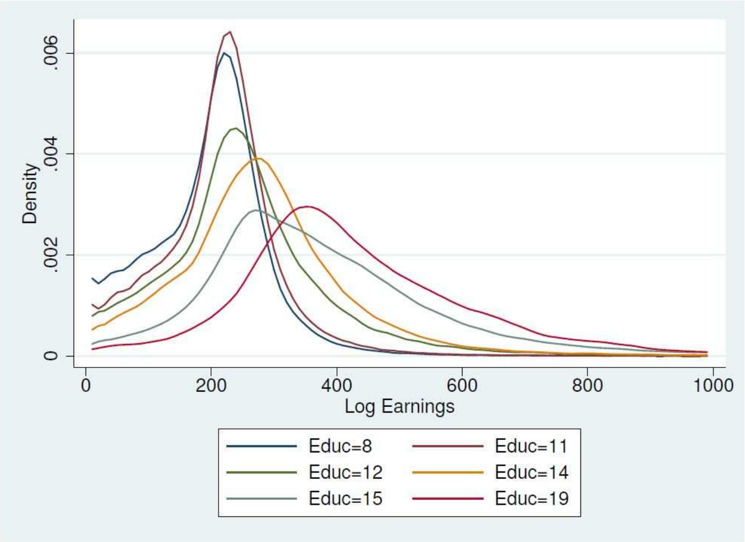
Return to Education

	Summary of logink		
Years of schooling	Mean	Std. Dev.	Freq.
8	5.0393588	.94620853	115182
11	5.182954	.79510023	219535
12	5.3377052	.84006766	66120
14	5.4758962	.74572105	83056
15.5	5.7930574	.69850008	67731
19	5.9640555	.57545454	5513
Total	5.2972029	.84945648	557137

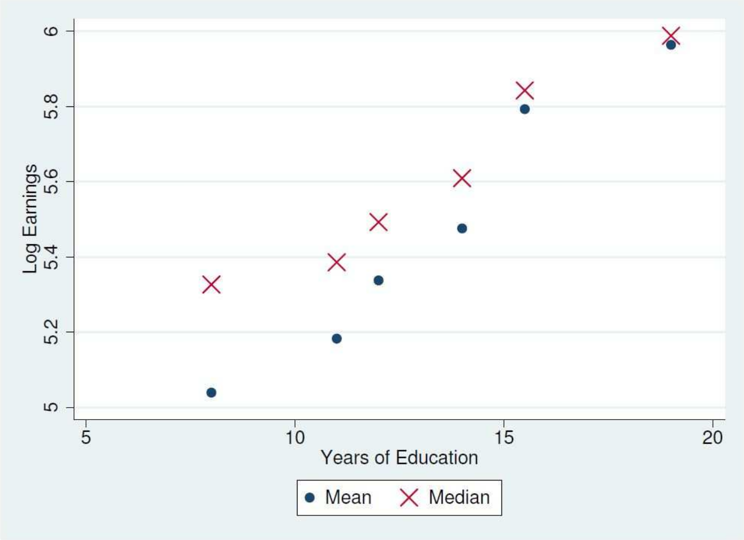
Return to Education



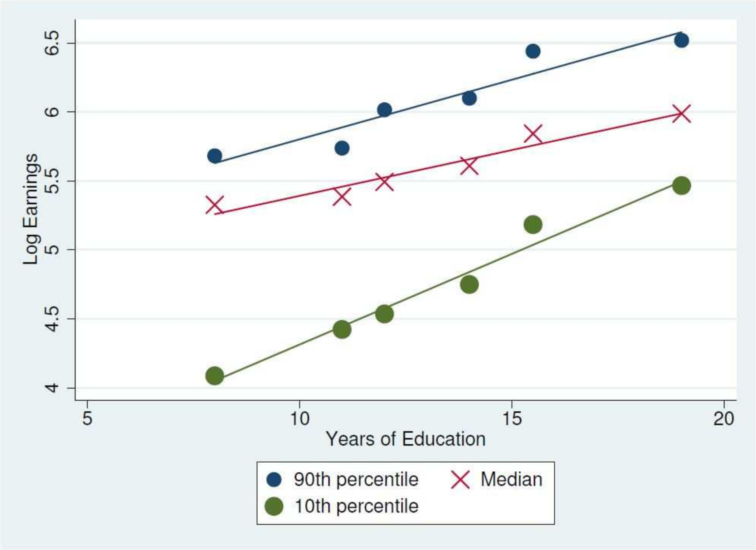
Return to Education



Return to Education



Return to Education



Return to Education

	OLS	$10th$	Median	$90th$
Years of Educ	.093 (.0004)	.14 (.001)	.064 (.00023)	.089 (.00041)
Constant	4.21 (.005)	2.94 (.014)	4.73 (.0028)	4.89 (.0048)

- Education has a complex effect on earnings:
  - Increase mean earnings.
  - Increases the earnings at the bottom of the distribution. Individuals are less likely to have really low wages.
  - Increases the earnings at the top of the distribution as well.
- Ambiguous effect a priori on the dispersion (inequality) in wages.

Return to Experience in the US

- Buchinsky (1994) *Econometrica* "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression"
- Uses the March Current Population Survey (March CPS) for 1964 through 1988. Contains between 10,000 and 34,000 observations each year. Sample consists of black and white males between 18 and 70.
- Model:

$$\text{Log Income}_{i,t} = X_{i,t}\beta_{t,\tau} + u_{i,t,\tau}$$

where $X_{i,t}$ contains a constant, years of schooling, experience, experience squared, race.

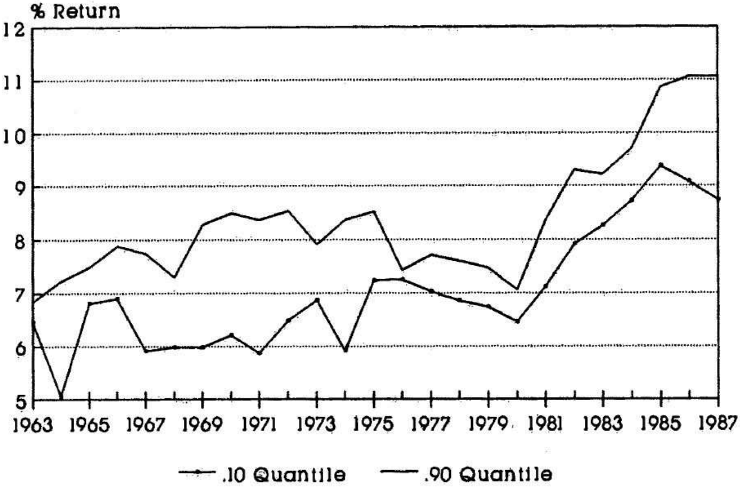
Return to Education in the US

TABLE I

ONE-GROUP MODEL—RETURN TO EDUCATION
MEAN, .10, .25, .50, .75, .90 QUANTILES, AND RESTRICTED ESTIMATES
DEPENDENT VARIABLE: LOG OF TOTAL WEEKLY INCOME FROM WAGES AND SALARIES

	Regressions						
Year	Mean	.10 Qnt.	.25 Qnt.	.50 Qnt.	.75 Qnt.	.90 Qnt.	Restricted
1963	6.65	6.45	6.23	6.35	6.58	6.84	6.48
	(0.22)	(0.54)	(0.45)	(0.29)	(0.37)	(0.42)	(0.25)
1964	6.51	5.07	6.44	6.41	6.46	7.22	6.39
	(0.22)	(0.57)	(0.48)	(0.30)	(0.34)	(0.43)	(0.25)
1965	6.84	6.81	6.82	6.60	6.61	7.48	6.67
	(0.13)	(0.42)	(0.27)	(0.16)	(0.20)	(0.31)	(0.15)
1966	6.85	6.90	6.02	6.44	7.04	7.88	6.73
	(0.19)	(0.55)	(0.42)	(0.24)	(0.29)	(0.48)	(0.22)
1967	6.82	5.91	6.54	6.57	6.88	7.74	6.72
	(0.13)	(0.40)	(0.27)	(0.17)	(0.19)	(0.33)	(0.15)
1968	6.53	5.98	6.19	6.26	6.82	7.30	6.45
	(0.13)	(0.42)	(0.25)	(0.16)	(0.19)	(0.31)	(0.15)
1969	7.05	5.99	6.48	6.89	7.52	8.28	7.14
	(0.14)	(0.43)	(0.27)	(0.17)	(0.21)	(0.28)	(0.16)
1970	7.32	6.21	6.44	7.18	7.81	8.50	7.43
	(0.14)	(0.44)	(0.30)	(0.18)	(0.21)	(0.31)	(0.16)

Return to Education in the US



- General increase in the return to schooling over 1963-1987.
- However, similar increase at all quantiles.

Quantile Treatment Effects

- Analogous to the Treatment Effect literature in models with heterogenous effects.
- Define the potential outcome of being treated as $Y(1)$ and the of not being treated as $Y(0)$.
- The quantile treatment effect is defined as:

$$\Delta_\tau = q_{1,\tau} - q_{0,\tau}$$

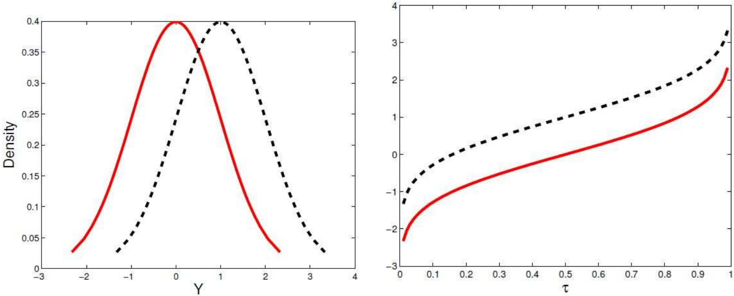
  with

$$q_{j,\tau} : Prob(Y(j) \leq q) = \tau, \qquad j = 0, 1$$

- Quantile regression with the treatment dummy:

$$Q_{Y_i}(\tau | T_i) = \alpha(\tau) + \delta(\tau) T_i$$

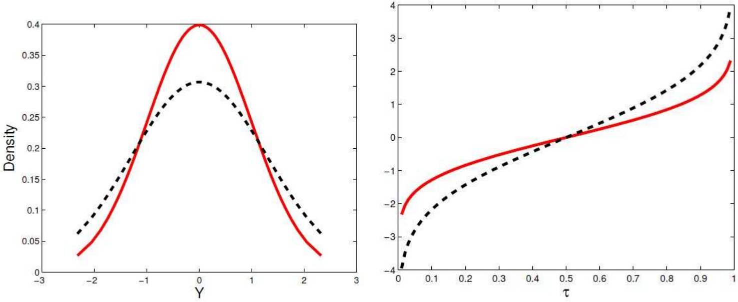
Quantile Treatment Effect: Location Shift



- The effect of the treatment is just to shift the entire distribution to the right, uniformly.

$$\delta(\tau) = \delta, \qquad \forall \tau$$

Quantile Treatment Effect: Scale Shift



- The effect of the treatment is to increase the variance, without a change in mean.

$$\delta(\tau) < 0 \qquad \forall \tau < 0.5$$
$$\delta(\tau) > 0 \qquad \forall \tau > 0.5$$

Example: Effect of Parental Death on Children

- Adda et al (2011) investigate the effect of parental death on children's long-term outcomes.
- Extensive administrative data on parental death, income, cognitive and non cognitive outcomes.
- They look at mean effect by gender and by parental death. One issue may also be that parental death affects not only the mean but the whole distribution of outcomes. They also present results on log earnings:

$$\log Earnings_i = \alpha_{0,\tau} + \beta_\tau Death_i + \gamma_\tau X_i + u_{i,\tau}$$

- $Death_i = 1$ if the child lost one of his/her parent before turning 18.

Example: Effect on Log Earnings

Section 5

Estimation in non-linear regression models

Estimation in non-linear regression models

- An obvious extension to the linear regression model studied so far is the non-linear regression model:

$$E[Y|X = x] = g(x, \theta)$$

equivalently, in regression function plus error form:

$$\begin{aligned} Y &= g(x, \theta) + \varepsilon \\ E[\varepsilon|X = x] &= 0. \end{aligned}$$

Consider M-estimation and in particular the non-linear least squares estimator obtained as follows.

$$\hat{\theta} = \arg\min_{\theta^*} n^{-1} \sum_{i=1}^{n} (Y_i - g(x_i; \theta^*))^2$$

- For now we just consider how a minimising value $\hat{\theta}$ can be found. Many of the statistical software packages have a routine to conduct non-linear optimisation and some have a non-linear least squares routine. Many of these routines employ a variant of Newton's method.

Numerical optimisation: Newton's method and variants

- Write the minimisation problem as:

$$\hat{\theta} = \arg\min_{\theta^*} Q(\theta^*).$$

  Newton's method involves taking a sequence of steps, $\theta_0, \theta_1, \ldots,$ $\theta_m, \ldots \theta_M$ from a starting value, θ_0 to an approximate minimising value θ_M which we will use as our estimator $\hat{\theta}$.
- The starting value is provided by the user. One of the tricks is to use a good starting value near to the final solution. This sometimes requires some thought.

Numerical optimisation: Newton's method and variants

- Suppose we are at θ_m. Newton's method considers a quadratic approximation to $Q(\theta)$ which is constructed to be an accurate approximation in a neighbourhood of θ_m, and moves to the value θ_{m+1} which minimises this quadratic approximation.
- At θ_{m+1} a new quadratic approximation, accurate in a neighbourhood of θ_{m+1} is constructed and the next value in the sequence, θ_{m+2}, is chosen as the value of θ minimising this new approximation.
- Steps are taken until a convergence criterion is satisfied. Usually this involves a number of elements. For example one might continue until the following conditions is satisfied:

$$Q_\theta(\theta_m)' Q_\theta(\theta_m) \leq \delta_1, \qquad |Q(\theta_m) - Q(\theta_{m-1})| < \delta_2.$$

Convergence criteria vary form package to package. Some care is required in choosing these criteria. Clearly δ_1 and δ_2 above should be chosen bearing in mind the orders of magnitude of the objective function and its derivative.

Numerical optimisation: Newton's method and variants

- The quadratic approximation used at each stage is a quadratic Taylor series approximation. At $\theta = \theta_m$,

$$Q(\theta) \simeq Q(\theta_m) + (\theta - \theta_m)' \, Q_\theta(\theta_m) + \frac{1}{2} (\theta - \theta_m)' \, Q_{\theta\theta'}(\theta_m) \, (\theta - \theta_m) = Q^a(\theta, \theta_m).$$

The derivative of $Q^a(\theta, \theta_m)$ with respect to θ is

$$Q^a_\theta(\theta, \theta_m) = Q_\theta(\theta_m) + Q_{\theta\theta'}(\theta_m) \, (\theta - \theta_m)$$

and θ_{m+1} is chosen as the value of θ that solves $Q^a_\theta(\theta, \theta_m) = 0$, namely

$$\theta_{m+1} = \theta_m - Q_{\theta\theta'}(\theta_m)^{-1} Q_\theta(\theta_m).$$

Numerical optimisation: Newton's method and variants

There are a number of points to consider here.

1. Obviously the procedure can only work when the objective function is twice differentiable with respect to θ.
2. The procedure will stop whenever $Q_\theta(\theta_m) = 0$, which can occur at a maximum and saddlepoint as well as at a minimum. The Hessian, $Q_{\theta\theta'}(\theta_m)$, should be positive definite at a minimum of the function.
3. When a minimum is found there is no guarantee that it is a global minimum. In problems where this possibility arises it is normal to run the optimisation from a variety of start points to guard against using an estimator that corresponds to a local minimum.
4. If, at a point in the sequence, $Q_{\theta\theta'}(\theta_m)$ is not positive definite then the algorithm may move away from the minimum and there may be no convergence. Many minimisation (maximisation) problems we deal with involve globally convex (concave) objective functions and for these there is no problem. For other cases, Newton's method is usually modified, e.g. by taking steps

$$\theta_{m+1} = \theta_m - A(\theta_m)^{-1} Q_\theta(\theta_m)$$

   where $A(\theta_m)$ is constructed to be positive definite and in cases in which $Q_{\theta\theta'}(\theta_m)$ is in fact positive definite, to be a good approximation to $Q_{\theta\theta'}(\theta_m)$.

Numerical optimisation: Newton's method and variants

5. The algorithm may "overstep" the minimum to the extent that it takes an "uphill" step, i.e. so that $Q(\theta_{m+1}) > Q(\theta_m)$. This is guarded against in many implementations of Newton's method by taking steps

$$\theta_{m+1} = \theta_m - \alpha(\theta_m) A(\theta_m)^{-1} Q_\theta(\theta_m)$$

where $\alpha(\theta_m)$ is a scalar step scaling factor, chosen to ensure that $Q(\theta_{m+1}) < Q(\theta_m)$.

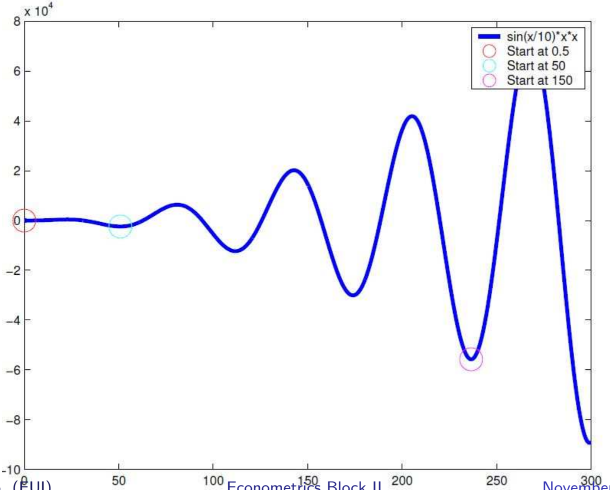
6. In practice it may be difficult to calculate exact expressions for the derivatives that appear in Newton's method. In some cases symbolic computational methods can help. In others we can use a numerical approximation, e.g.

$$Q_{\theta_i}(\theta_m) \simeq \frac{Q_\theta(\theta_m + \delta_i e_i) - Q_\theta(\theta_m)}{\delta_i}$$

where e_i is a vector with a one in position i and zeros elsewhere, and δ_i is a small perturbing value, possibly varying across the elements of θ.

Numerical Optimisation: Example

- Function $y = sin(x/10) * x^2$.
- This function has many (an infinite) number of local minimas.
- Start off the nonlinear optimisation at various points.



Section 6

Maximum Likelihood Methods

Maximum Likelihood Methods

- Some of the models used in econometrics specify the complete probability distribution of the outcomes of interest rather than just a regression function.
- Sometimes this is because of special features of the outcomes under study - for example because they are discrete or censored, or because there is serial dependence of a complex form.
- When the complete probability distribution of outcomes given covariates is specified we can develop an expression for the probability of observation of the responses we see as a function of the unknown parameters embedded in the specification.
- We can then ask what values of these parameters maximise this probability for the data we have. The resulting statistics, functions of the observed data, are called *maximum likelihood estimators*. They possess important optimality properties and have the advantage that they can be produced in a rule directed fashion.

Estimating a Probability

- Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with $P[Y_i = 1] = p$, $P[Y_i = 0] = 1 - p$ for all i.
- We might use such a model for data recording the occurrence or otherwise of an event for n individuals, for example being in work or not, buying a good or service or not, etc.
- Let $y_1, \ldots, y_n$ indicate the data values obtained and note that in this model

$$
\begin{aligned}
P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n, p] &= \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)} \\
&= p^{\sum_{i=1}^{n} y_i}(1-p)^{\sum_{i=1}^{n}(1-y_i)} \\
&= \pounds(p; y).
\end{aligned}
$$

  With any set of data $\pounds(p; y)$ can be calculated for any value of p between 0 and 1. The result is the probability of observing the data to hand for each chosen value of p.
- One strategy for estimating p is to use that value that maximises this probability. The resulting estimator is called the *maximum likelihood estimator* (MLE) and the maximand, $\pounds(p; y)$, is called the *likelihood function*.

Log Likelihood Function

- The maximum of the *log likelihood function*, $L(p; y) = \log \pounds(p, y)$, is at the same value of p as is the maximum of the likelihood function (because the log function is monotonic).
- It is often easier to maximise the log likelihood function (LLF). For the problem considered here the LLF is

$$L(p; y) = \left(\sum_{i=1}^{n} y_i\right) \log p + \sum_{i=1}^{n} (1 - y_i) \log(1 - p).$$

Let

$$\hat{p} = \arg\max_{p} L(p; y) = \arg\max_{p} \pounds(p; y).$$

On differentiating we have the following.

$$\begin{aligned} q(p; y) &= \frac{1}{p} \sum_{i=1}^{n} y_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - y_i) && \textit{score} \\ Q(p; y) &= -\frac{1}{p^2} \sum_{i=1}^{n} y_i - \frac{1}{(1-p)^2} \sum_{i=1}^{n} (1 - y_i) && \textit{hessian} \end{aligned}$$

Likelihood Functions and Estimation in General

- Let Y_i, $i = 1, \ldots, n$ be continuously distributed random variables with joint probability density function $f(y_1, \ldots, y_n, \theta)$.
- The probability that Y falls in infinitesimal intervals of width $dy_1, \ldots dy_n$ centred on values $y_1, \ldots, y_n$ is

  $$A = f(y_1, \ldots, y_n, \theta) dy_1 dy_2 \ldots dy_n$$

  Here only the joint density function depends upon θ and the value of θ that maximises $f(y_1, \ldots, y_n, \theta)$ also maximises A.
- In this case the likelihood function is defined to be the joint *density* function of the Y_i's.
- When the Y_i's are discrete random variables the likelihood function is the joint probability mass function of the Y_i's, and in cases in which there are discrete and continuous elements the likelihood function is a combination of probability density elements and probability mass elements.
- In all cases the likelihood function is a function of the observed data values that is equal to, or proportional to, the probability of observing these particular values.

Likelihood Functions and Estimation in General

- When Y_i, $i = 1, \ldots, n$ are *independently* distributed the joint density (mass) function is the *product* of the marginal density (mass) functions of each Y_i, the likelihood function is

$$\pounds(y;\theta) = \prod_{i=1}^{n} f_i(y_i;\theta),$$

  There is a subscript i on f to allow for the possibility that each Y_i has a distinct probability distribution.

- This situation arises when modelling conditional distributions of Y given some covariates x. In particular, $f_i(y_i;\theta) = f_i(y_i|x_i;\theta)$, and often $f_i(y_i|x_i;\theta) = f(y_i|x_i;\theta)$.

- In time series and panel data problems there is often dependence among the Y_i's. For any list of random variables $Y = \{Y_1, \ldots, Y_n\}$ define the $i-1$ element list $Y_{i-} = \{Y_1, \ldots, Y_{i-1}\}$. The joint density (mass) function of Y can be written as

$$f(y) = \prod_{i=2}^{n} f_{y_i|y_{i-}}(y_i|y_{i-}) f_{y_1}(y_1),$$

Invariance

- Note that (parameter free) monotonic transformations of the Y_i's (for example, a change of units of measurement, or use of logs rather than the original y data) usually leads to a change in the value of the maximised likelihood function when we work with continuous distributions.
- If we transform from y to z where $y = h(z)$ and the joint density function of y is $f_y(y;\theta)$ then the joint density function of z is

$$f_z(z;\theta) = \left| \frac{\partial h(z)}{\partial z} \right| f_y(h(z);\theta).$$

- For any given set of values, y^*, the value of θ that maximises the likelihood function $f_y(y^*, \theta)$ also maximises the likelihood function $f_z(z^*;\theta)$ where $y^* = h(z^*)$, so the maximum likelihood estimator is invariant with respect to such changes in the way the data are presented.
- However the maximised likelihood functions will differ by a factor equal to $\left| \frac{\partial h(z)}{\partial z} \right|_{z=z^*}$.

Maximum Likelihood: Properties

- Maximum likelihood estimators possess another important *invariance property*. Suppose two researchers choose different ways in which to parameterise the same model. One uses θ, and the other uses $\lambda = h(\theta)$ where this function is one-to-one. Then faced with the same data and producing estimators $\hat{\theta}$ and $\hat{\lambda}$, it will always be the case that $\hat{\lambda} = h(\hat{\theta})$.
- There are a number of important consequences of this:
  - For instance, if we are interested in the ratio of two parameters, the MLE of the ratio will be the ratio of the ML estimators.
  - Sometimes a re-parameterisation can improve the numerical properties of the likelihood function. Newton's method and its variants may in practice work better if parameters are rescaled.

Maximum Likelihood: Improving Numerical Properties

- An example of this often arises when, in index models, elements of x involve squares, cubes, etc., of some covariate, say x_1. Then maximisation of the likelihood function may be easier if instead of x_1^2, x_1^3, etc., you use $x_1^2/10$, $x_1^3/100$, etc., with consequent rescaling of the coefficients on these covariates. You can always recover the MLEs you would have obtained without the rescaling by rescaling the estimates.

Maximum Likelihood: Improving Numerical Properties

- There are some cases in which a re-parameterisation can produce a globally concave likelihood function where in the original parameterisation there was not global concavity.
- An example of this arises in the "Tobit" model.
  - This is a model in which each Y_i is $N(x_i'\beta, \sigma^2)$ with negative realisations replaced by zeros. The model is sometimes used to model expenditures and hours worked, which are necessarily non-negative.
  - In this model the likelihood as parameterised here is not globally concave, but re-parameterising to $\lambda = \beta/\sigma$, and $\gamma = 1/\sigma$, produces a globally concave likelihood function.
  - The invariance property tells us that having maximised the "easy" likelihood function and obtained estimates $\hat{\lambda}$ and $\hat{\gamma}$, we can recover the maximum likelihood estimates we might have had difficulty finding in the original parameterisation by calculating $\hat{\beta} = \hat{\lambda}/\hat{\gamma}$ and $\hat{\sigma} = 1/\hat{\gamma}$.

Properties Of Maximum Likelihood Estimators

- First we just sketch the main results:
  - Let $L(\theta; Y)$ be the log likelihood function now regarded as a random variable, a function of a set of (possibly vector) random variables $Y = \{Y_1, \ldots, Y_n\}$.
  - Let $q(\theta; Y)$ be the gradient of this function, itself a vector of random variables (scalar if θ is scalar) and let $Q(\theta; Y)$ be the matrix of second derivatives of this function (also a scalar if θ is a scalar).
  - Let

$$\hat{\theta} = \arg\max_{\theta} \quad L(\theta; Y).$$

In order to make inferences about θ using $\hat{\theta}$ we need to determine the distribution of $\hat{\theta}$. We consider developing a large sample approximation. The limiting distribution for a quite wide class of maximum likelihood problems is as follows:

$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_0)$$

where

$$V_0 = -\operatorname*{plim}_{n\to\infty}(n^{-1}Q(\theta_0; Y))^{-1}$$

and θ_0 is the unknown parameter value. To get an approximate distribution that can be used in practice we use $(n^{-1}Q(\hat{\theta}; Y))^{-1}$ or some other consistent estimator of V_0 in place of V_0.

Variance of ML Estimator

- An alternative way of "estimating " V_0, is:

$$\hat{V}_0^o = \left\{ n^{-1} q(\hat{\theta}; Y) q(\hat{\theta}; Y)' \right\}^{-1}$$

  which compared with

$$\tilde{V}_0^o = \left\{ -n^{-1} Q(\hat{\theta}; Y) \right\}^{-1}$$

  has the advantage that only first derivatives of the log likelihood function need to be calculated. Sometimes $\hat{V}_0^o$ is referred to as the "outer product of gradient" (OPG) estimator.
- Maximum likelihood estimators possess optimality property, namely that, among the class of consistent and asymptotically normally distributed estimators, the variance matrix of their limiting distribution is the smallest that can be achieved in the sense that other estimators in the class have limiting distributions with variance matrices exceeding the MLE's by a positive semidefinite matrix.

Section 7

Discrete Choice Models

Estimating a Conditional Probability

- Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with

$$\begin{aligned} P[Y_i &= 1|X = x_i] = p(x_i, \theta) \\ P[Y_i &= 0|X = x_i] = 1 - p(x_i, \theta). \end{aligned}$$

This is an obvious extension of the model in the previous section.

- The likelihood function for this problem is

$$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n|x] &= \prod_{i=1}^{n} p(x_i, \theta)^{y_i}(1 - p(x_i, \theta))^{(1-y_i)} \\ &= \pounds(\theta; y). \end{aligned}$$

where y denotes the complete set of values of y_i and dependence on x is suppressed in the notation. The log likelihood function is

$$L(\theta; y) = \sum_{i=1}^{n} y_i \log p(x_i, \theta) + \sum_{i=1}^{n}(1 - y_i) \log(1 - p(x_i, \theta))$$

and the maximum likelihood estimator of θ is

$$\hat{\theta} = \arg\max_{\theta} \quad L(\theta; y).$$

So far this is an obvious generalisation of the simple problem met in the last section.

Estimating a Conditional Probability

- To implement the model we choose a form for the function $p(x, \theta)$, which must of course lie between zero and one.
  - One common choice is

$$p(x, \theta) = \frac{\exp(x'\theta)}{1 + \exp(x'\theta)}$$

    which produces what is commonly called a *logit model*.
  - Another common choice is

$$\begin{aligned} p(x, \theta) &= \Phi(x'\theta) = \int_{-\infty}^{x'\theta} \phi(w) dw \\ \phi(w) &= (2\pi)^{-1/2} \exp(-w^2/2) \end{aligned}$$

    in which Φ is the standard normal distribution function. This produces what is known as a *probit model*.
- Both models are widely used. Note that in both cases a single index model is specified, the probability functions are monotonic increasing, probabilities arbitrarily close to zero or one are obtained when $x'\theta$ is sufficiently large or small, and there is a symmetry in both of the models in the sense that $p(-x, \theta) = 1 - p(x, \theta)$.
- Any or all of these properties might be inappropriate in a particular application but there is rarely discussion of this in the applied econometrics literature.

More on Logit and Probit

- Both models can also be written as a linear model involving a latent variable.
- We define a **latent variable** Y_i^*, which is unobserved, but determined by the following model:

$$Y_i^* = X_i\theta + \varepsilon_i$$

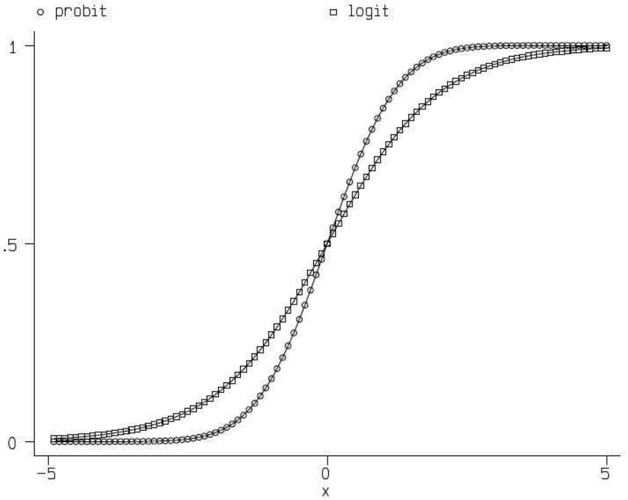
We observe the variable Y_i which is linked to Y_i^* as:

$$\begin{cases} Y_i = 0 & \text{if } \; Y_i^* < 0 \\ \\ Y_i = 1 & \text{if } \; Y_i^* \geq 0 \end{cases}$$

- The probability of observing $Y_i = 1$ is:

$$\begin{aligned} p_i = P(Y_i = 1) &= P(Y_i^* \geq 0) \\ &= P(X_i\theta + \varepsilon_i \geq 0) \\ &= P(\varepsilon_i \geq -X_i\theta) \\ &= 1 - F_\varepsilon(-X_i\theta) \end{aligned}$$

Shape of Logit and Probit Models



Odds-Ratio

- Define the ratio $p_i/(1-p_i)$ as the **odds-ratio**. This is the ratio of the probability of outcome 1 over the probability of outcome 0. If this ratio is equal to 1, then both outcomes have equal probability ($p_i = 0.5$). If this ratio is equal to 2, say, then outcome 1 is twice as likely than outcome 0 ($p_i = 2/3$).
- In the logit model, the log odds-ratio is linear in the parameters:

$$\ln \frac{p_i}{1-p_i} = X_i\theta$$

- In the logit model, θ is the marginal effect of X on the log odds-ratio. A unit increase in X leads to an increase of θ % in the odds-ratio.

Marginal Effects

- Logit model:

$$\begin{aligned}\frac{\partial p_i}{\partial X} &= \frac{\theta\exp(X_i\theta)(1+\exp(X_i\theta)) - \theta\exp(X_i\theta)^2}{(1+\exp(X_i\theta))^2} \\ &= \frac{\theta\exp(X_i\theta)}{(1+\exp(X_i\theta))^2} \\ &= \theta p_i(1-p_i)\end{aligned}$$

  A one unit increase in X leads to an increase in the probability of choosing option 1 of $\theta p_i(1-p_i)$.

- Probit model:

$$\frac{\partial p_i}{\partial X_i} = \theta\phi(X_i\theta)$$

  A one unit increase in X leads to an increase in the probability of choosing option 1 of $\theta\phi(X_i\theta)$.

Maximum Likelihood in Single Index Models

- We can cover both cases by considering general single index models, so for the moment rewrite $p(x, \theta)$ as $g(w)$ where $w = x'\theta$.
- The log-likelihood is then:

$$L(\theta, y) = \sum_{i=1}^{n} y_i \log(g(w_i)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - g(w_i))$$

- The first derivative of the log likelihood function is:

$$\begin{aligned} q(\theta; y) &= \sum_{i=1}^{n} \frac{g_w(x_i'\theta) x_i}{g(x_i'\theta)} y_i - \frac{g_w(x_i'\theta) x_i}{1 - g(x_i'\theta)} (1 - y_i) \\ &= \sum_{i=1}^{n} (y_i - g(x_i'\theta)) \frac{g_w(x_i'\theta)}{g(x_i'\theta)(1 - g(x_i'\theta))} x_i \end{aligned}$$

Here $g_w(w)$ is the derivative of $g(w)$ with respect to w.

Maximum Likelihood in Single Index Models

- The expression for the second derivative is rather messy. Here we just note that its expected value given x is quite simple, namely

$$E[Q(\theta; y)|x] = -\sum_{i=1}^{n} \frac{g_w(x_i'\theta)^2}{g(x_i'\theta)\,(1 - g(x_i'\theta))} x_i x_i',$$

the negative of which is the Information Matrix, $I(\theta)$ for general single index binary data models.

Asymptotic Properties of the Logit Model

- For the logit model there is major simplification

$$\begin{aligned} g(w) &= \frac{\exp(w)}{1+\exp(w)} \\ g_w(w) &= \frac{\exp(w)}{(1+\exp(w))^2} \\ \Rightarrow \frac{g_w(w)}{g(w)\,(1-g(w))} &= 1. \end{aligned}$$

Therefore in the logit model the MLE satisfies

$$\sum_{i=1}^{n}\left(y_i - \frac{\exp(x_i'\hat{\theta})}{1+\exp(x_i'\hat{\theta})}\right)x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n}\frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2}x_i x_i',$$

Asymptotic Properties of the Logit Model

- the MLE has the limiting distribution

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V_0)$$

$$V_0 = \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2} x_i x_i' \right)^{-1},$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta_0) \simeq N(0, V_0)$$

using the estimator

$$\hat{V}_0 = \left(n^{-1} \sum_{i=1}^{n} \frac{\exp(x_i'\hat{\theta})}{\left(1+\exp(x_i'\hat{\theta})\right)^2} x_i x_i' \right)^{-1}$$

when producing approximate hypothesis tests and confidence intervals.

Asymptotic Properties of the Probit Model

- In the probit model

$$\begin{aligned} g(w) &= \Phi(w) \\ g_w(w) &= \phi(w) \\ \Rightarrow \frac{g_w(w)}{g(w)\,(1-g(w))} &= \frac{\phi(w)}{\Phi(w)(1-\Phi(w))}. \end{aligned}$$

Therefore in the probit model the MLE satisfies

$$\sum_{i=1}^{n}\left(y_i - \Phi(x_i'\hat{\theta})\right)\frac{\phi(x_i'\hat{\theta})}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))}x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n}\frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))}x_i x_i',$$

Asymptotic Properties of the Probit Model

- the MLE has the limiting distribution

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V_0)$$

$$V_0 = \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))} x_i x_i' \right)^{-1},$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta_0) \simeq N(0, V_0)$$

using the estimator

$$\hat{V}_0 = \left(n^{-1} \sum_{i=1}^{n} \frac{\phi(x_i'\hat{\theta})^2}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))} x_i x_i' \right)^{-1}$$

when producing approximate tests and confidence intervals.

Example: Logit and Probit

- We have data from households in Kuala Lumpur (Malaysia) describing household characteristics and their concern about the environment. The question is
  "Are you concerned about the environment? Yes / No".
  We also observe their age, sex (coded as 1 men, 0 women), income and quality of the neighborhood measured as air quality. The latter is coded with a dummy variable *smell*, equal to 1 if there is a bad smell in the neighborhood. The model is:

$$Concern_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 \log income_i + \beta_4 smell_i + u_i$$

Example: Logit and Probit

- We estimate this model with three specifications, linear probability model (LPM), logit and probit:

Probability of being concerned by Environment

Variable	LPM		Logit		Probit	
	Est.	t-stat	Est.	t-stat	Est.	t-stat
age	.0074536	3.9	.0321385	3.77	.0198273	3.84
sex	.0149649	0.3	.06458	0.31	.0395197	0.31
log income	.1120876	3.7	.480128	3.63	.2994516	3.69
smell	.1302265	2.5	.5564473	2.48	.3492112	2.52
constant	-.683376	-2.6	-5.072543	-4.37	-3.157095	-4.46
	Some Marginal Effects					
Age	.0074536		.0077372		.0082191	
log income	.1120876		.110528		.1185926	
smell	.1302265		.1338664		.1429596	

Goodness of Fit

- As in linear models, we are generally *not* interested in the fit of the model, but rather to test whether some variables are significant or not.
- However, researchers have designed a number of ways to assess the fit of a discrete choice model:
  - **Percent correctly predicted:** For each observation, we compute the predicted probability that $Y_i = 1$, given X_i. If that probability is higher than $1/2$ we say that the outcome for individual i is correctly predicted and vice versa. However, there is no reason to focus on $Y_i = 1$ as an outcome, and we can look at $Y_i = 0$ as well. Or take an average of these two measures of goodness of fit.
  - **pseudo R-squared:** There are several ways to compute such a measure
    - $R = 1 - L_{UR}/L_o$, where L_{UR} is the log-likelihood of the model and L_o is the log-likelihood in a model with only a constant.
    - $1 - SSR_{UR}/SSR_o$, where SSR_{UR} is the sum of squared residuals $(\hat{u}_i = Y_i - g(X_i\theta)$ and SSR_o is the total sum of squares of Y_i.

Neglected Heterogeneity

- We consider the probit case which is the simplest to analyze in this context.
- Suppose we are interested in the following model:

$$P(Y_i = 1) = \Phi(X_i\theta + \gamma c_i)$$

  where c_i is a scalar.
- We are interested in the partial effect of X_i, holding c_i fixed. We assume that $E(c_i) = 0$ without loss of generality if X_i contains a constant.
- We assume that c_i is independent of X_i and that $c_i \sim N(0, \tau^2)$.

Neglected Heterogeneity

- We can write our model in latent variable form:

$$Y_i^* = X_i\theta + \gamma c_i + e_i, \qquad e_i \sim N(0,1)$$

- Hence $\gamma c_i + e_i$ is normally distributed with mean 0 and variance $\sigma^2 = 1 + \gamma^2\tau^2 > 1$.
- If we neglect c_i in our regression, the probit estimation gives:

$$P(Y_i = 1|X_i) = P(\gamma c_i + e_i > -X_i\theta|X_i) = \Phi(X_i\theta/\sigma) = \Phi(X_i\tilde{\theta})$$

- In this case the estimation gives us $\tilde{\theta} < \theta$ as $\sigma > 1$. Hence our estimates are biased towards zero.
- However, the sign of the coefficient is correct, and so is the ratio of two elements of θ. We may not be interested in the exact magnitude of the effect of X_i on Y_i.

Neglected Heterogeneity and Marginal Effects

- We would like to estimate the marginal effect :

$$\frac{\partial P(Y_i = 1|X_i, c_i)}{\partial X_{ij}} = \theta_j \phi(X_i\theta + \gamma c_i)$$

but this is impossible as we do not observe c_i.

- We can evaluate the marginal effect at $c_i = 0$ which is $\theta_j \phi(X_i\theta)$. However, what we identify with our data is: $\tilde{\theta}\phi(X_i\tilde{\theta})$. The direction of the bias is ambiguous because $\tilde{\theta} < \theta$, but $\phi(X_i\tilde{\theta}) > \phi(X_i\theta)$
- However, we may be interested in the **average partial effect**:

$$E_c[\theta_j \phi(X_i\theta + \gamma c_i)] = \frac{\theta_j}{\sigma}\phi(X_i\theta/\sigma) = \tilde{\theta}_j \phi(X_i\tilde{\theta})$$

Hence, the probit, with neglected heterogeneity, consistently estimates the average partial effect, even though the marginal effect conditional on a given c_i is biased.

Endogenous Explanatory Variables

- Consider the (latent) model:

$$\begin{aligned} &Y_1^* = Z_1\delta_1 + \alpha_1 Y_2 + u_1 \qquad var(u_1) = 1 \\ &Y_2 = Z_1\delta_{21} + Z_2\delta_{22} + v_2 = Z\delta_2 + v_2 \\ &Y_1 = I(Y_1^* > 0) \end{aligned}$$

  where u_1 and v_2 have zero mean are are drawn from a bivariate normal distribution, independent of Z

- In this model, Y_2 is endogenous if u_1 and v_2 are correlated. (We restrict our analysis to the case where Y_2 is continuous.)
- The object of interest is the coefficient α_1.

Endogenous Explanatory Variables

- One way to proceed is to estimate the first equation using a LPM and 2SLS, with Z as an instrument. This is fairly straightforward, and should give a consistent estimate of the average effect.
- We can also deal with the endogeneity directly in a probit framework. This requires quite strong restrictions on the model.

Endogenous Explanatory Variables

- Write u_1 as a function of v_2 and a shock independent of Z and v_2:

$$u_1 = \theta_1 v_2 + e_1 \qquad \theta_1 = \eta_1/\tau_2^2, \quad \eta_1 = cov(v_2, u_1), \quad \tau_2^2 = var(v_2)$$

$$Var(e_1) = Var(u_1) - \eta_1^2/\tau_2^2 = 1 - \rho_1^2$$

- Write the model as:

$$Y_1^* = Z_1\delta_1 + \alpha_1 Y_2 + \theta_1 v_2 + e_1$$

$$\begin{aligned} P(Y_1 = 1 | Z, Y_2, v_2) &= \Phi\left(\frac{Z_1\delta_1 + \alpha_1 Y_2 + \theta_1 v_2}{\sqrt{1-\rho_1^2}}\right) \\ &= \Phi(Z_1\tilde{\delta}_1 + \tilde{\alpha}_1 Y_2 + \tilde{\theta}_1 v_2) \end{aligned}$$

- As $\rho_1^2 < 1$, the tilded coefficients are larger than the untilded ones, unless the correlation between v_2 and u_1 is zero (no endogeneity).

Endogenous Explanatory Variables

- The estimation proceeds in two steps:
  - First run an OLS regression of Y_2 on Z and get an estimate of the residual $\hat{v}_2$.
  - Run the probit Y_1 on Z_1, Y_2 and $\hat{v}_2$, to get consistent estimators of $\tilde{\delta}_1$, $\tilde{\alpha}_1$ and $\tilde{\theta}_1$.
- Under the null of no endogeneity ($H_0 : \theta_1 = 0$), we can use the usual probit t statistic for a test of endogeneity.

Endogenous Explanatory Variables

- How do we compute average partial effects?
- It can be shown that:

$$\begin{aligned} E_{v_2}[P(Y_1 = 1|Z, Y_2, v_2)] &= E_{v_2}[\Phi(Z_1\tilde{\delta}_1 + \tilde{\alpha}_1 Y_2 + \tilde{\theta}_1 v_2)] \\ &= \Phi(Z_1\tilde{\tilde{\delta}}_1 + \tilde{\tilde{\alpha}}_1 Y_2) \end{aligned}$$

with $\tilde{\tilde{\alpha}}_1 = \tilde{\alpha}_1/\sqrt{\tilde{\theta}_1^2\hat{\tau}_2^2 + 1}$, so that:

$$E_{v_2}\left[\frac{\partial P(Y_1 = 1|Z, Y_2, v_2)}{\partial Y_2}\right] = \tilde{\tilde{\alpha}}_1\phi(Z_1\tilde{\tilde{\delta}}_1 + \tilde{\tilde{\alpha}}_1 Y_2)$$

- Another way is to compute:

$$\frac{1}{N}\sum_{i=1}^{N}\Phi(Z_{i1}\tilde{\delta}_1 + \tilde{\alpha}_1 Y_{i2} + \tilde{\theta}_1\hat{v}_{i2})$$

using the residuals of the first stage and averaging across the whole sample.
- Note that standards errors are complicated to get, because of the two step nature of the method.

Endogenous Explanatory Variables: Example

- Data drawn from the Swedish Survey of Living Conditions, years 1981, 1988, 1995.
- We are interested in the effect of education on smoking behavior.
- *Smoker* is a dummy variable equal to 1 if the individual is or is has been a regular smoker. We also control for age and sex, and we are interested in the effect of years of education.

$$Smoker_i^* = \delta_{10} + \delta_{11} age_i + \delta_{12} sex_i + \alpha_1 educ_i + u_1$$

$$educ_i = \delta_{21} age_i + \delta_{22} sex_i + \delta_{23} educFather_i + \delta_{21} educMother_i + v_2$$

where *educFather* and *educMother* are the education level of the father and mother, taken as instrument for the child's education. How valid are these instrument a priori?

Example: ignoring endogeneity

```
. *** Exogeneity assumed
. probit smoker age sex educa

Iteration 0:   log likelihood =  -27433.08
Iteration 1:   log likelihood = -26924.137
Iteration 2:   log likelihood = -26924.098
Iteration 3:   log likelihood = -26924.098

Probit regression                                 Number of obs   =      39578
                                                  LR chi2(3)      =    1017.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -26924.098                       Pseudo R2       =     0.0186
```

smoker	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0017906	.0003403	-5.26	0.000	-.0024575	-.0011237
sex	.3927524	.012725	30.86	0.000	.3678119	.4176929
educa	.0029786	.0041954	0.71	0.478	-.0052444	.0112015
_cons	-.1192115	.024657	-4.83	0.000	-.1675382	-.0708847

```
. margins, predict(pr) dydx(educa)

Average marginal effects                          Number of obs   =      39578
Model VCE    : OIM

Expression   : Pr(smoker), predict(pr)
dy/dx w.r.t. : educa
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
educa	.0011643	.0016399	0.71	0.478	-.0020499	.0043784

Example: first stage

```
. ** First stage
. reg educa age sex educF educM

      Source |       SS       df       MS              Number of obs =   11081
-------------+------------------------------           F(  4, 11076) =  297.10
       Model |  2676.98298     4  669.245744           Prob > F      =  0.0000
    Residual |  24950.1034 11076  2.25262761           R-squared     =  0.0969
-------------+------------------------------           Adj R-squared =  0.0966
       Total |  27627.0864 11080  2.49341935           Root MSE      =  1.5009

------------------------------------------------------------------------------
       educa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.016948   .0008455   -20.05   0.000    -.0186053   -.0152907
         sex |   .0382278   .0285503     1.34   0.181    -.0177358    .0941915
       educF |   .1517326   .0112084    13.54   0.000     .1297623     .173703
       educM |   .0384227   .0129562     2.97   0.003     .0130263    .0638191
       _cons |    3.64479   .0542462    67.19   0.000     3.538458    3.751122
------------------------------------------------------------------------------

. predict v2,res
(28497 missing values generated)
```

Example: second stage

```
. probit smoker age sex educa v2

Iteration 0:   log likelihood = -7680.0812
Iteration 1:   log likelihood = -7575.1409
Iteration 2:   log likelihood = -7575.1219
Iteration 3:   log likelihood = -7575.1219

Probit regression                               Number of obs   =      11081
                                                LR chi2(4)      =     209.92
                                                Prob > chi2     =     0.0000
Log likelihood = -7575.1219                     Pseudo R2       =     0.0137
```

smoker	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0042454	.001149	3.69	0.000	.0019934	.0064975
sex	.1900386	.0240682	7.90	0.000	.1428658	.2372114
educa	-.1469944	.0428318	-3.43	0.001	-.2309432	-.0630457
v2	.1144017	.0435772	2.63	0.009	.0289921	.1998114
_cons	.1829667	.1856595	0.99	0.324	-.1809192	.5468526

```
. margins, predict(pr) dydx(educa)

Average marginal effects                        Number of obs   =      11081
Model VCE    : OIM

Expression   : Probability of positive outcome, predict(pr)
dy/dx w.r.t. : educa
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
educa	-.057544	.0152761	-3.77	0.000	-.0874847	-.0276034

Multinomial Logit

- The logit model was dealing with two qualitative outcomes. This can be generalized to multiple outcomes:
  - choice of transportation: car, bus, train...
  - choice of dwelling: house, apartment, social housing.
- The multinomial logit: Denote the outcomes as $j = 1, \ldots, J$ and p_j the probability of outcome j.

$$p_j = \frac{\exp(X\theta^j)}{\sum_{k=1}^{J} \exp(X\theta^k)}$$

where θ^j is a vector of parameter associated with outcome j.

Identification

- If we multiply all the coefficients by a factor λ this does not change the probabilities p_j, as the factor cancel out. This means that there is under identification. We have to normalize the coefficients of one outcome, say, J to zero. All the results are interpreted as deviations from the baseline choice.
- We write the probability of choosing outcome $j = 1, \ldots, J-1$ as:

$$p_j = \frac{\exp(X\theta^j)}{1 + \sum_{k=1}^{J-1} \exp(X\theta^k)}$$

- We can express the logs odds-ratio as:

$$\ln \frac{p_j}{p_J} = X\theta^j$$

- The odds-ratio of choice j versus J is only expressed as a function of the parameters of choice j, but not of those other choices: Independence of Irrelevant Alternatives (IIA).

Independence of Irrelevant Alternatives

An anecdote which illustrates a violation of this property has been attributed to Sidney Morgenbesser:

After finishing dinner, Sidney Morgenbesser decides to order dessert. The waitress tells him he has two choices: apple pie and blueberry pie. Sidney orders the apple pie.

After a few minutes the waitress returns and says that they also have cherry pie at which point Morgenbesser says "In that case I'll have the blueberry pie."

Independence of Irrelevant Alternatives

- Consider travelling choices, by car or with a red bus. Assume for simplicity that the choice probabilities are equal:

$$P(car) = P(\text{red bus}) = 0.5 \implies \frac{P(car)}{P(\text{red bus})} = 1$$

- Suppose we introduce a blue bus, (almost) identical to the red bus. The probability that individuals will choose the blue bus is therefore the same as for the red bus and the odd ratio is:

$$P(\text{blue bus}) = P(\text{red bus}) \implies \frac{P(\text{blue bus})}{P(\text{red bus})} = 1$$

- However, the IIA implies that odds ratios are the same whether of not another alternative exists. The only probabilities for which the three odds ratios are equal to one are:

$$P(car) = P(\text{blue bus}) = P(\text{red bus}) = 1/3$$

However, the prediction we ought to obtain is:

$$P(\text{red bus}) = P(\text{blue bus}) = 1/4 \quad P(car) = 0.5$$

Marginal Effects: Multinomial Logit

- θ^j can be interpreted as the marginal effect of X on the log odds-ratio of choice j to the baseline choice.
- The marginal effect of X on the probability of choosing outcome j can be expressed as:

$$\frac{\partial p_j}{\partial X} = p_j[\theta^j - \sum_{k=1}^{J} p_k \theta^k]$$

  Hence, the marginal effect on choice j involves not only the coefficients relative to j but also the coefficients relative to the other choices.
- Note that we can have $\theta^j < 0$ and $\partial p_j / \partial X > 0$ or vice versa. Due to the non linearity of the model, the sign of the coefficients does not indicate the direction nor the magnitude of the effect of a variable on the probability of choosing a given outcome. One has to compute the marginal effects.

Example

- We analyze here the choice of dwelling: house, apartment or low cost flat, the latter being the baseline choice. We include as explanatory variables the age, sex and log income of the head of household:

Variable	Estimate	Std. Err.	Marginal Effect
		Choice of House	
age	.0118092	.0103547	-0.002
sex	-.3057774	.2493981	-0.007
log income	1.382504	.1794587	0.18
constant	-10.17516	1.498192	
		Choice of Apartment	
age	.0682479	.0151806	0.005
sex	-.89881	.399947	-0.05
log income	1.618621	.2857743	0.05
constant	-15.90391	2.483205	

Ordered Models

- In the multinomial logit, the choices were not ordered. For instance, we cannot rank cars, busses or train in a meaningful way. In some instances, we have a natural ordering of the outcomes even if we cannot express them as a continuous variable:
  - Yes / Somehow / No.
  - Low / Medium / High
- We can analyze these answers with ordered models.

Ordered Probit

- We code the answers by arbitrary assigning values:

$$Y_i = 0 \;\text{ if No}, \quad Y_i = 1 \;\text{ if Somehow}, \quad Y_i = 2 \;\text{ if Yes}$$

- We define a latent variable Y_i^* which is linked to the explanatory variables:

$$Y_i^* = X_i'\theta + \varepsilon_i$$

$$\begin{array}{ll} Y_i = 0 & \text{if } \; Y_i^* < 0 \\ Y_i = 1 & \text{if } \; Y_i^* \in [0, \mu[\\ Y_i = 2 & \text{if } \; Y_i^* \geq \mu \end{array}$$

  μ is a threshold and an auxiliary parameter which is estimated along with θ.

- We assume that ε_i is distributed normally.
- The probability of each outcome is derived from the normal cdf:

$$\begin{array}{l} P(Y_i = 0) = \Phi(-X_i'\theta) \\ P(Y_i = 1) = \Phi(\mu - X_i'\theta) - \Phi(-X_i'\theta) \\ P(Y_i = 2) = 1 - \Phi(\mu - X_i'\theta) \end{array}$$

Ordered Probit

- Marginal Effects:

$$\frac{\partial P(Y_i = 0)}{\partial X_i} = -\theta\phi(-X_i'\theta)$$
$$\frac{\partial P(Y_i = 1)}{\partial X_i} = \theta\left(\phi(X_i'\theta) - \phi(\mu - X_i'\theta)\right)$$
$$\frac{\partial P(Y_i = 2)}{\partial X_i} = \theta\phi(\mu - X_i'\theta)$$

- Note that if $\theta > 0$, $\partial P(Y_i = 0)/\partial X_i < 0$ and $\partial P(Y_i = 2)/\partial X_i > 0$:
  - If X_i has a positive effect on the latent variable, then by increasing X_i, fewer individuals will stay in category 0.
  - Similarly, more individuals will be in category 2.
  - In the intermediate category, the fraction of individual will either increase or decrease, depending on the relative size of the inflow from category 0 and the outflow to category 2.

Ordered Probit: Example

- We want to investigate the determinants of health.
- Individuals are asked to report their health status in three categories: poor, fair or good.
- We estimate an ordered probit and calculate the marginal effects at the mean of the sample.

Variable	Coeff	sd. err.	Marginal Effects			Sample
			Poor	Fair	Good	Mean
Age 18-30	-1.09**	.031	-.051**	-.196**	.248**	.25
Age 30-50	-.523**	.031	-.031**	-.109**	.141**	.32
Age 50-70	-.217**	.026	-.013**	-.046**	.060**	.24
Male	-.130**	.018	-.008**	-.028**	.037**	.48
Income low third	.428**	.027	.038**	.098**	-.136**	.33
Income medium third	.264**	.022	.020**	.059**	-.080**	.33
Education low	.40**	.028	.031**	.091**	-.122**	.43
Education Medium	.257**	.026	.018**	.057**	-.076**	.37
Year of interview	-.028	.018	-.001	-.006	.008	1.9
Household size	-.098**	.008	-.006**	-.021**	.028**	2.5
Alcohol consumed	.043**	.041	.002**	.009**	-.012**	.04
Current smoker	.160**	.018	.011**	.035**	-.046**	.49
cut1	.3992**	.058				
cut2	1.477**	.059				

Ordered Probit: Example

Age group	Proportion		
	Poor Health	Fair Health	Good Health
Age 18-30	.01	.08	.90
Age 30-50	.03	.13	.83
Age 50-70	.07	.28	.64
Age 70 +	.15	.37	.46

Ordered Probit: Example

- Marginal Effects differ by individual characteristics.
- Below, we compare the marginal effects from an ordered probit and a multinomial logit.

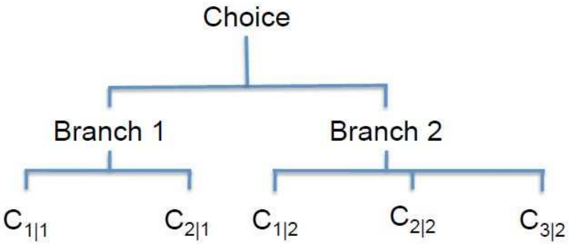
	Marginal Effects for Good Health			
Variable	Ordered Probit at mean	X	Ordered Probit at X	Multinomial Logit at X
Age 18-30	.248**	1	.375**	.403**
Age 30-50	.141**	0	.093**	.077**
Age 50-70	.060**	0	.046**	.035**
Male	.037**	1	.033**	.031**
Income low third	-.136**	1	-.080**	-.066**
Income medium third	-.080**	0	-.071**	-.067**
Education low	-.122**	1	-.077**	-.067**
Education Medium	-.076**	0	-.069**	-.064**
Year of interview	.008	1	.006	.003
Household size	.028**	2	.023**	.020**
Alcohol consumed	-.012**	0	-.010**	-.011**
Current smoker	-.046**	0	-.041**	-.038**

Models without IIA

Here we discuss 3 ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the ε_{ij}:

- nested logit model: the researcher groups together sets of choices. This allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation across nests.
- unrestricted multinomial probit model: with no restrictions on the covariance between unobserved components, beyond normalizations.
- mixed or random coefficients logit: where the marginal utilities associated with choice characteristics vary between individuals, generating positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

Nested Logit Models



- Partition the set of choices $\{0, 1, ..., J\}$ into S sets $B_1, ..., B_S$.
- The idea is that we still maintain the IIA within each branch, but we allow the variance to differ across branches.
- Note that we can generalize this approach by having many more layers.

Link with Utility Maximization

- Link with the random utility model: Consumer i who opts for choice j gets the following utility:

$$U_{ij} = \alpha_j + X_{ij}\beta_j + Z_i\gamma_j + \epsilon_{ij}$$

- α_j is a particular attribute of choice j, constant across all agents.
- X_{ij} are variables that vary by choice and individual, for instance the distance of i to the choice j, or specific costs and so on.
- Z_i are variables that describe the individual, such as income, sex, or education.
- The error term $\epsilon_{i1}, \ldots, \epsilon_{iJ}$ follow the Generalized Extreme-value (GEV) distribution. This is a generalization of the extreme value distribution that allows for alternatives within branches to be correlated.

More on GEV Distribution

- For your information, the GEV distribution takes the form:

$$F_{T,R}(\epsilon) = \exp\left[-\sum_{k \in T} \left(\sum_{l \in R_k} \exp(-\epsilon_{kl}/\rho_k \right)^{\rho_t} \right]$$

Nested Logit Models

- Let the conditional probability of choice j given that your choice is in the set B_b (branch b) , be equal to

$$Pr(Y_i = j | X_i, Y_i \in B_b) = \frac{\exp(X'_{bj}\beta_j/\rho_b)}{\sum_{l \in B_b} \exp(X'_{bl}\beta_j/\rho_b)}$$

- So within a branch, the formula looks like the multinomial logit, except that we are scaling the parameters with the coefficient ρ_b.
- The probability to choose branch b is:

$$P(\text{branch} = b) = \frac{\left(\sum_{m \in R_b} \exp(\eta_{bm}/\rho_b)\right)^{\rho_b}}{\sum_{k \in T}\left(\sum_{m \in R_k} \exp(\eta_{km}/\rho_k)\right)^{\rho_k}}$$

with $\eta_{bm} = X_{bm}\beta_m + Z_b\gamma_b$.

Nested Logit Models

- Noting that

$$\begin{aligned}
\left(\sum_{m\in R_b} \exp(\eta_{bm}/\rho_b)\right)^{\rho_b} &= \left(\sum_{m\in R_b} \exp(\frac{Z_b\gamma_b + X_{bm}\beta_m}{\rho_b})\right)^{\rho_b} \\
&= \exp(Z_b\gamma_b)\left(\sum_{m\in R_b} \exp(X_{bm}\beta_m/\rho_b)\right)^{\rho_b} \\
&= \exp(Z_b\gamma_b + \rho_b I_b)
\end{aligned}$$

- We have defined the inclusive value I_b:

$$I_b = \ln\left(\sum_{m\in R_b} \exp(X_{bm}\beta_m/\rho_b)\right)$$

Nested Logit Models

- The probability of choosing branch b is then expressed in a simpler way as the function of the inclusive values:

$$P(\text{branch} = b) = \frac{\exp(Z_b \gamma_b + \rho_b I_b)}{\sum_{k \in T} \exp(Z_k \gamma_k + \rho_k I_k)}$$

Marginal Effects

- As usual in non-linear models, the coefficients are not directly interpretable.
- The marginal effects are:

$$\frac{\partial \ln P(choice = m, branch = b)}{\partial X(k) \text{in choice } M \text{and branch} B} = \beta_k \Big[I_{b=B}(I_{m=M} - P_{M|B}) + \rho_B [I_{b=B} - P_B] P_{M|B}) \Big]$$

Estimation of the Nested Logit Model

- Maximization of the likelihood function is difficult (see below).
- Another method is to proceed in two steps. This is called the Limited Information Maximum Likelihood, and is less efficient than the Full Information Maximum Likelihood. However, it is rather intuitive:
  - Note that within a branch, we have a simple conditional logit model with coefficients β_j/ρ_b. We can directly estimate $\widehat{\beta_j/\rho_b}$ by using only information for that branch.
  - In a second step, we can compute the inclusive values and look at the probability of a particular set B_b to estimate ρ_b. This is also a multinomial logit:

$$P(\text{branch} = b) = \frac{\exp(Z_b\gamma_b + \rho_b I_b)}{\sum_{k \in T} \exp(Z_k\gamma_k + \rho_k I_k)}$$

Maximum Likelihood of the Nested Logit

- To illustrate, we consider only two layers.

$$
\begin{aligned}
\log l &= \sum_{i=1}^{N} \sum_{b \in T} \sum_{m \in B_b} y_{ibm} \log[P(B_i = b)P(C_i = m|B_i = b)] \\
&= \sum_{i=1}^{N} \sum_{b \in T} \sum_{m \in B_b} y_{ibm} \left[Z_{ib}\gamma_b + \rho_b I_{ib} \right. \\
&\quad - \log(\sum_{l \in T} \exp(Z_{il}\gamma_l + \rho_l I_{il})) + \\
&\quad \left. X_{ibm}\beta_m/\rho_b - \log(\sum_{l \in B_b} \exp(X_{ibl}\beta_m/\rho_b)) \right]
\end{aligned}
$$

Nested Logit Model

- We can extend the model to many layers of nests.
- The results may be sensitive to the specification of the nest structure.
- The procedure is ad-hoc as it is up to the researcher to first choose the nests and which products are close substitutes.
- If we want to predict market shares for a new product, we have to decide in which nest it belongs.

Example: Restaurant Choice

- The data comes from the Stata website (webuse restaurant).
- Data on 300 families with their choice of restaurant, with information on family size, income, distance to restaurants, price of menu and rating of restaurant.
- Not all restaurants are alike. It is possible that families do not substitute at random between them.
- We can classify them in several groups:
  - Fast food restaurants.
  - Family restaurants.
  - Fancy restaurants.

Multinomial Probit with Unrestricted Covariance Matrix

- A second possibility is to unrestrict the covariance matrix of the error terms. It is easier to do this in a multinomial probit casr.
- Consider

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X_{i0}'\beta + \epsilon_{i0} \\ X_{i1}'\beta + \epsilon_{i1} \\ \vdots \\ X_{iJ}'\beta + \epsilon_{iJ} \end{pmatrix} \qquad \epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} |X_i \sim N(0, \Omega)$$

where Ω is a $(J+1)$x$(J+1)$ covariance matrix.

Multinomial Probit with Unrestricted Covariance Matrix

- The maximization of the log likelihood function requires the evaluation of a $J+1$ order integral, which is infeasible if $J \geq 3, 4$.
- This requires usually simulated methods.
- The advantage of the model is that we do not have to specify how close a choice is to the other. However, it requires a lot of data to precisely estimate the covariance matrix which can involve a huge number of parameters.
- To predict the effect of a new good, the researcher has to specify all the correlations with the other goods.

Random Effects Models

- A third possibility is to allow for unobserved heterogeneity in the slope coefficients. We typically think that individuals who have a taste for a particular attribute may have a similar taste for a close substitute.
- We can model this by allowing the marginal utilities to vary at the individual level:

$$U_{ij} = X_{ij}'\beta_i + \epsilon_{ij} = X_{ij}'\bar{\beta} + \nu_{ij} \qquad \nu_{ij} = \epsilon_{ij} + X_{ij}(\beta_i - \bar{\beta})$$

- In this formulation the ν_{ij} are not independent across choices.

Random Effects Models

- One possibility to implement this is to assume the existence of a finite number of types of individuals: $\beta_i \in \{b_0, b_1, \ldots, b_K\}$, with

$$Pr(\beta_i = b_k | Z_i) = p_k$$

- We can then construct the likelihood and integrate out the unobserved types using these weights. The maximization of the log-likelihood is done over the coefficients of the model *and* the $\{p_k, b_k\}$.

Example

Econometrica, Vol. 63, No. 4 (July, 1995), 891–951

PRODUCT DIFFERENTIATION AND OLIGOPOLY IN INTERNATIONAL MARKETS: THE CASE OF THE U.S. AUTOMOBILE INDUSTRY

BY PINELOPI KOUJIANOU GOLDBERG[1]

- Develops and estimates a model of the US Automobile Industry.
- Equilibrium oligopoly model with product differentiation.
- Used to evaluate the effect of trade-policies.
- Analysis is done in two steps:
  - Demand side: estimation of demand for automobiles using micro-data.
  - Supply side: Static model of firms that takes micro-demand as given.

The Model

- Consumer j maximizes an indirect utility of the form:

$$U_j^h = \bar{V}_j^h + \varepsilon_j^h \qquad \text{for vehicle } h$$

- $\bar{V}_j^h$ is a deterministic component that is a function of the vehicle attributes (power, engine size) and consumer's characteristics.
- ε_j^h: captures unmeasured variables, taste shocks...
- Each car is characterized by four elements:
    - n: newness.
    - c: market segment.
    - o: origin.
    - m: make.
- The utility is expressed as:

$$U_{b,n,c,o,m}^h = \bar{V}_{b,n,c,o,m}^h + \varepsilon_{b,n,c,o,m}^h$$

The Model

- The deterministic part is a linear function of consumer and vehicle characteristics

$$
\begin{aligned}
U^h_{b,n,c,o,m} &= \alpha' B^h_b + \beta' N^h_{b,n} + \gamma' C^h_{b,n,c} \\
&\quad + \delta' O^h_{b,n,c,o} + \zeta M^h_{b,n,c,o} + \varepsilon^h_{b,n,c,o,m}
\end{aligned}
$$

- B^h_b varies only with the decision to purchase, $N_{b,n}$ captures the utility of new cars, C is segment, O origin and M make.
- ε follows an extreme value distribution, and the choices are nested.

Decision Tree

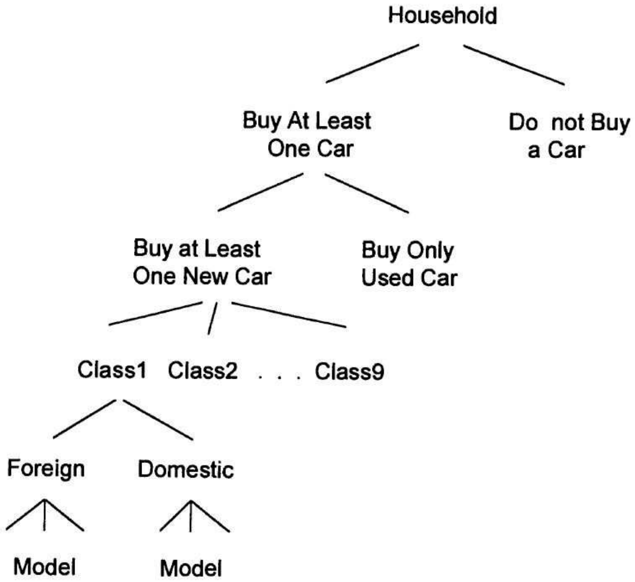


FIGURE 1.—Automobile choice model.

Choice Probabilities

- The joint probability of choosing a new vehicle type (b, n, c, o, m) is

$$P^h_{b,n,c,o,m} = P^h_b . P^h_{n|b} . P^h_{c|n,b} . P^h_{o|c,n,b} . P^h_{m|o,c,n,b}$$

- At each node s of the tree, the marginal probability of purchasing a car has the form:

$$P^h_{i_s|j_{s-1}} = \frac{\exp(X^h_{i_s}\theta_s/\lambda_{j_{s-1}} + I^h_{i_s}\lambda_{i_s}/\lambda_{j_{s-1}})}{\sum\limits_{k \in C_{j_{s-1}}} \exp(X^h_{k_s}\theta_s/\lambda_{j_{s-1}} + I^h_{k_s}\lambda_{k_s}/\lambda_{j_{s-1}})}$$

with

$$I^h_{i_s} = \log\left[\sum_{p \in C_{i_s}} \exp(X^h_{p_{s+1}}\theta_{s+1}/\lambda_{i_s})\right]$$

- $I^h_{i_s}$ is the inclusive value, i.e. the expected aggregate utility of subset i_s and $\lambda_{i_s}/\lambda_{j_{s-1}}$ reflect the dissimilarity of alternatives belonging to a particular subset.
- Model estimated by sequential maximum likelihood in five stages.

Price Effects

- Prices enter the specification of the bottom nest. They only have an effect in other choices through the inclusive values.

$$\zeta' M^h_{b,n,c,o,m} = \zeta'_1 M1^h_{b,n,c,o,m} + \zeta_2(INC)(INC^h - PRICE_{b,n,c,o,m})$$

- Price (and income) effects are allowed to vary with income, so that rich can be less price sensitive.

Producers' Problem

- The producer's problem is to maximize expected profits with respect to price.
- Uncertainty comes from the demand side, as there is randomness in choices.
- Producers play a Nash game.
- Foreign producers may be facing quota restrictions, imposed by the US.
- Crucial part of the model is how consumers react to prices and substitute to other types of car. Important to relax the IIA assumption.

Data Set

- The data comes from the Consumer Expenditure Survey (CES).
- Each quarter, about 4500 households are interviewed.
- Representative of the US population.

TABLE A1

CES HOUSEHOLD CHARACTERISTICS
(Sample Size: 20,571. Years covered: 1983–87)

Characteristics	Means	Standard Deviations
Age of Household Head	43.6	14.7
Income ($/year, in 82 dollars)	21,104	14,231
Family Size	2.67	1.44
Number of Earners	1.52	0.50
Cars/Household	1.57	1.12
Percent w/College Education	48	30
Ethnic Composition (%)		
Caucasian	88	29
Black/Hispanic	9	24
Asian	3	14
Geographic Composition (%)		
Northeast	20	42
North Central	25	43
West	26	40
South	29	40

Data Set

TABLE A4

NUMBER OF ALTERNATIVES BY CLASS[a]

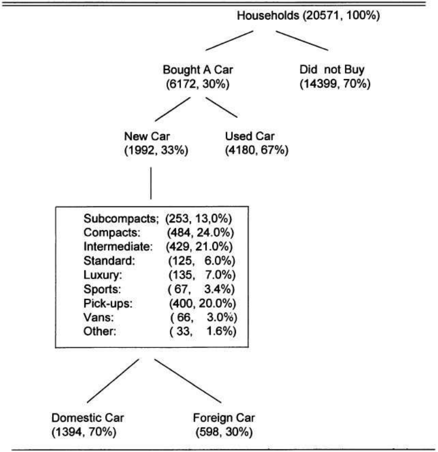
Class	Origin	# of Models	Class	Origin	# of Models
1. Subcompacts	Domestic	16	6. Sports	Domestic	6
	Foreign	28		Foreign	7
2. Compacts	Domestic	14	7. Pick-ups	Domestic	30
	Foreign	19		Foreign	7
3. Intermediate	Domestic	30	8. Vans	Domestic	14
	Foreign	4		Foreign	2
4. Standard	Domestic	17	9. Other	Domestic	8
	Foreign	0		Foreign	2
5. Luxury	Domestic	14			
	Foreign	10			

[a] The number of models in a class is an average over 1983–87.

Data Set

TABLE A3

CAR PURCHASES OF CES HOUSEHOLDS[a]



[a] The figures in parentheses refer to number and percentage of households respectively. The sample period is 1983–87. Approximately 9,000 households with missing or invalid responses were eliminated from the sample.

Data Set

TABLE A5

MEANS OF VEHICLE CHARACTERISTICS BY CLASS[a]

Class	Ori	Price	Leng	Wid	HP	Wei	Cyl	MPG
1. Subcompact	Dom	9504	174.4	67.4	84.5	2.5	4.1	26.4
		(2126)	(7.5)	(2.9)	(13.1)	(0.3)	(0.3)	(6.5)
	For	10390	163.8	64.1	79.4	2.1	4.0	27.1
		(2674)	(7.2)	(1.6)	(9.8)	(0.2)	(0.0)	(3.6)
2. Compact	Dom	9577	174.1	66.3	86.3	2.4	4.1	24.0
		(2374)	(7.1)	(1.7)	(13.3)	(0.2)	(0.3)	(3.2)
	For	11284	170.4	65.1	87.6	2.3	4.3	25.4
		(2814)	(7.5)	(1.1)	(23.5)	(0.3)	(0.6)	(4.0)
3. Intermed.	Dom	11933	191.5	70.7	102.7	2.9	4.9	19.8
		(3101)	(9.1)	(2.0)	(13.6)	(0.3)	(1.1)	(2.4)
	For	16585	169.3	69.7	103.8	2.9	4.5	18.5
		(5301)	(35.1)	(1.1)	(7.1)	(0.2)	(0.4)	(1.7)
4. Standard	Dom	13782	204.5	73.9	124.7	3.3	6.2	18.4
		(2658)	(10.0)	(2.1)	(14.3)	(0.3)	(0.7)	(1.4)
5. Luxury	Dom	20524	205.1	72.8	132.1	3.7	7.2	17.4
		(4996)	(14.2)	(3.5)	(15.8)	(0.5)	(1.2)	(0.8)
	For	27615	190.8	67.8	126.8	3.1	4.8	17.5
		(10446)	(7.2)	(1.8)	(18.6)	(0.5)	(0.7)	(3.2)
6. Sports	Dom	14754	186.0	70.8	148.6	3.1	6.3	19.9
		(5203)	(10.2)	(1.0)	(45.6)	(0.6)	(1.4)	(2.9)
	For	17849	172.2	66.5	136.6	2.7	4.8	19.8
		(7454)	(4.5)	(1.2)	(33.5)	(0.3)	(0.9)	(2.1)
7. Pick-up	Dom	12078	193.0	75.6	132.6	6.0	6.4	16.5
		(2318)	(11.5)	(4.4)	(22.7)	(1.0)	(1.2)	(2.4)
	For	8705	174.6	64.1	98.2	4.4	4.0	23.0

Results

TABLE B1

MODEL CHOICE: SMALL CARS
Number of Observations: 707
Log of Likelihood Function: −517.9

Variable	Parameter Estimate	Standard Error
PP10	−4.747	0.862
PP20	−4.501	0.356
PP11	−2.927	0.328
PP21	−2.755	1.277
TRANS	3.516	0.225
PS	0.615	0.202
AIRC	5.777	0.255
HPD	−0.018	0.588
HPDYOUNG	−0.203	0.903
FUELC	−7.143	0.740
+15 Brand Dummies		
(All of them highly significant)		

TABLE B2

MODEL CHOICE: BIG CARS
Number of Observations: 980
Log of Likelihood Function: −414.5

Variable	Parameter Estimate	Standard Error
PP10	−4.445	0.602
PP20	−3.745	0.332
PP11	−3.076	0.649
PP21	−2.171	0.396
TRANS	0.877	0.281
PS	5.525	0.364
AIRC	8.956	0.429
HPD	3.580	0.864
HPDYOUNG	0.275	1.760
FUELC	−1.381	0.744
+16 Brand Dummies		
(5 of them significant at the 10% level)		

Results

TABLE B4

FOREIGN VS. DOMESTIC[a]
Number of Observations: 1867
Log of Likelihood Function: −413.9
0: Domestic; 1: Foreign

Variable	Parameter Estimate	Standard Error
INCL1	0.891	0.024
INCL2	0.988	0.023
INCL3	0.199	0.100
C1	−0.165	0.499
AGE1	−1.193	0.340
EDUC1	0.791	0.197
NE1	0.127	0.243
NC1	−0.435	0.261
WE1	0.460	0.246
ASIAN1	0.584	0.652
BLUEC1	−0.381	0.257
INCOM1	0.347	0.180
D841	0.255	0.359
D851	−0.199	0.367
D861	0.508	0.365
D871	1.743	0.371
CC21	0.147	0.269
CC31	−3.367	0.421

Results

TABLE B5

CLASS CHOICE
Number of Observations: 1992
Log of Likelihood Function: −2115.1
1−9: Class 1−Class 9

Variable	Parameter Estimate	Standard Error
CINCL	0.944	0.024
C2	1.122	0.363
AGE2	−0.363	0.424
INCOM2	0.037	0.208
FAMSIZE2	−0.234	0.118
PERSLT182	0.180	0.170
C3	−10.793	0.469
AGE3	2.365	0.393
INCOM3	0.146	0.209
FAMSIZE3	−0.299	0.107
PERSLT183	0.448	0.156
C4	−17.199	0.672
AGE4	2.891	0.489
INCOM4	0.404	0.278
FAMSIZE4	−0.674	0.153
PERSLT184	0.795	0.225
C5	−5.027	0.506

Section 8

Censored Regression Models

Examples

- This chapter deals with the estimation of models in which observations are censored or for which there is a corner solution.
- We also look at cases where data is missing, creating a sample bias.

Example 1: Top Coding

- In many surveys recording information on income and wealth, information on those variables are not fully reported for the super-rich, to protect their identity. The variable is top-coded, i.e. all income above a given value is replaced by that value.
- Nonetheless, we may want to relate (true) income (Y^*) to a number of characteristics (X).

$$E(Y^*|X) = X\beta$$

- What we observe instead is $Y = \min(Y^*, \bar{Y})$. Hence, we have to estimate the following model:

$$Y = \min(\bar{Y}, X\beta + u)$$

$$-(Y - \bar{Y}) = \max(0, -\bar{Y} - X\beta - u)$$

Example 2: Corner Solution

- Suppose an agent is maximising utility over two goods, own consumption c and charitable contributions q:

$$\max_{c,q} u(c,q) = \max_{c,q} c + a_i \log(1+q)$$

$$c + pq = m$$

- The solution to this problem is:

$$\begin{cases} q_i = 0 & \text{if } a_i/p_i \leq 1 \\ q_i = a_i/p_i - 1 & \text{if } a_i/p_i > 1 \end{cases} \quad \text{or } 1 + q_i = \max(1, a_i/p_i)$$

$$\log(1+q_i) = \max(0, Z_i\gamma - \log p_i + u_i) \qquad \text{if } a_i = \exp(Z_i\gamma + u_i)$$

Tobit Model

- First proposed by Tobin (1958), [1]
- We define a latent (unobserved) variable Y^* such as:

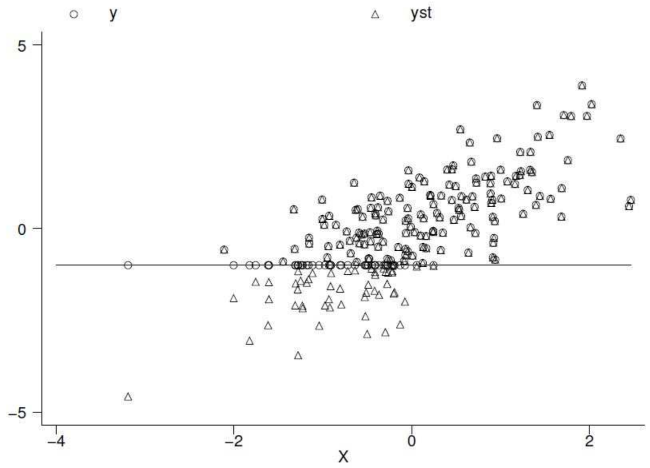
$$Y^* = X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

- We only observe a variable Y which is related to Y^* such as:

$$\begin{array}{lll} Y = Y^* & \text{if} & Y^* > a \\ Y = a & \text{if} & Y^* \leq a \end{array}$$

[1] Tobin, J. (1958), ŤEstimation of Relationships for Limited Dependent VariablesŤ, *Econometrica 26, 24-36.*

Example



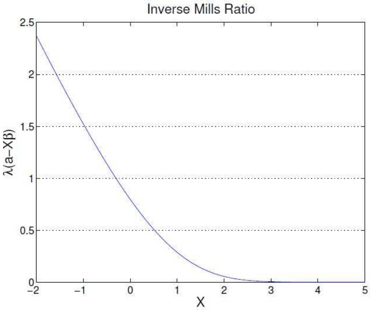
Truncation Bias

- The conditional mean of Y given X takes the form:

$$E[Y|Y^* > a, X] = X\beta + \sigma \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$$

with $\alpha = \frac{a - X\beta}{\sigma}$.

- The ratio $\phi(\alpha)/(1 - \Phi(\alpha))$ is called the inverse Mills ratio.



- Intuitively, the second term is there because the conditional expectation of the error term is not equal to zero, but positive.

Truncation Bias: Proof

- Proof: Note that the conditional c.d.f of $Y^*|Y^* > a$ is:

$$
\begin{aligned}
H(y|Y^* > a, X) &= P(Y^* \leq y|Y^* > a) = \frac{P(a < Y^* \leq y)}{P(Y^* > a)} \\
&= \frac{P(a - X\beta < \varepsilon \leq y - X\beta)}{P(\varepsilon > a - X\beta)} \\
&= \frac{\Phi(\frac{y - X\beta}{\sigma}) - \Phi(\frac{a - X\beta}{\sigma})}{1 - \Phi(\frac{a - X\beta}{\sigma})}
\end{aligned}
$$

so that the conditional distribution is:

$$
h(y|Y^* > a, X) = \frac{\partial H(y|Y^* > a, X)}{\partial y} = \frac{\phi(\frac{y - X\beta}{\sigma})}{\sigma(1 - \Phi(\frac{a - X\beta}{\sigma}))}
$$

Proof Continued

$$
\begin{aligned}
E[Y|Y^* > a, X] &= \int_a^{+\infty} yh(y|Y^* > a, X)dy \\
&= \frac{1}{\sigma(1-\Phi(\alpha))}\int_a^{+\infty} y\phi(\frac{y - X\beta}{\sigma})dy \\
&= \frac{1}{1-\Phi(\alpha)}\int_{(a-X\beta)/\sigma}^{+\infty} (X\beta + \sigma z)\phi(z)dz \\
&= X\beta - \frac{1}{1-\Phi(\alpha)}\sigma\int_{(a-X\beta)/\sigma}^{+\infty} \phi'(z)dz \\
&= X\beta + \sigma\frac{\phi(\alpha)}{1-\Phi(\alpha)} \\
&= X\beta + \sigma\lambda(\alpha)
\end{aligned}
$$

Inconsistency of OLS

- OLS using the entire sample or the the subsample for which $Y^* > a$ yields inconsistent estimators of β.
- OLS on the subsample with $Y > a$:

$$E[Y|X, Y > a] = X\beta + E(u|u > a - X\beta) = X\beta + \sigma\lambda(\alpha)$$

the OLS parameters estimate of β will be biased and inconsistent as we omit the term $\sigma\lambda(\alpha)$.

- OLS on the whole data:

$$\begin{aligned} E(Y|X) &= P(Y > a|X)E(Y|X, Y > a) + aP(Y < a|X) \\ &= (1 - \Phi(\alpha))(X\beta + \sigma\lambda(\alpha)) + a\Phi(\alpha) \end{aligned}$$

Not much hope here as well...

Likelihood for Tobit Model

- The conditional c.d.f of Y given X is:

$$\begin{aligned} G(y|X,\beta,\sigma) &= P(Y \leq y|X] \\ &= P(Y \leq y|X, Y > a)P(Y > a|X) \\ &\quad +P(Y \leq y|X, Y = a)P(Y = a|X) \\ &= I(y > a)H(y|Y > a, X)(1 - \Phi(\frac{a - X\beta}{\sigma})) \\ &\quad +I(y = a)\Phi(\frac{a - X\beta}{\sigma}) \end{aligned}$$

where $I(.)$ is the indicator function: $I(true) = 1$, $I(false) = 0$.

- The corresponding conditional density is:

$$\begin{aligned} g(y|X,\beta,\sigma) &= I(y > a)h(y|Y > a, X)(1 - \Phi(\frac{a - X\beta}{\sigma})) + I(y = a)\Phi(\frac{a - X\beta}{\sigma}) \\ &= I(y > a)\frac{\phi(\frac{Y - X\beta}{\sigma})}{\sigma} + I(y = a)\Phi(\frac{a - X\beta}{\sigma}) \end{aligned}$$

Likelihood for Tobit Model

- The log-likelihood function of the Tobit model is:

$$
\begin{aligned}
l(\beta,\sigma) &= \sum_{i=1}^{n} \log(g(Y_i|X_i,\beta,\sigma)) \\
&= \sum_{i=1}^{n} I(y_i > a)(\log(\phi(\frac{Y_i - X_i\beta}{\sigma})) - log(\sigma)) \\
&\quad + \sum_{i=1}^{n} I(y_i = a)\log(\Phi(\frac{a - X\beta}{\sigma})) \\
&= \sum_{i=1}^{n} I(y_i > a)\left(-\frac{1}{2}(Y_i - X_i\beta)^2/\sigma^2 - 2\log(\sigma) - \log(\sqrt{2\pi})\right) \\
&\quad + \sum_{i=1}^{n} I(y_i = a)\log(\Phi(\frac{a - X_i\beta}{\sigma}))
\end{aligned}
$$

- This can be maximised with respect to β, σ or $\gamma = 1/\sigma$ and $\lambda = \beta/\sigma$.

Tobit Model: Marginal Effects

- How do we interpret the coefficient β?

$$\beta_j = \frac{\partial E(Y^*|X)}{\partial X_j}$$

  This is the marginal effect of X on the (latent) variable Y^*. This has a direct interpretation in the case of censored data.

- For corner solutions, what we care about is:

$$\begin{aligned}\frac{\partial E(Y|X, Y > a)}{\partial X_j} &= \frac{\partial}{\partial X_j}(X\beta + \sigma\lambda(\alpha)) = \beta_j + \sigma\frac{\partial \alpha}{\partial X_j}\frac{\partial \lambda(\alpha)}{\partial \alpha} \\ &= \beta_j\left[1 - \lambda(\alpha)^2 + \alpha\lambda(\alpha)\right]\end{aligned}$$

- Another marginal effect is:

$$\frac{\partial E[Y|X]}{\partial X} = \beta(1 - \Phi(\alpha))$$

Example: Willingness To Pay for better

- The WTP is censored at zero. We can compare the two regressions:

OLS: $$WTP_i = \beta_0 + \beta_1 lny + \beta_2 age_i + \beta_3 smell_i + u_i$$

Tobit: $$WTP_i^* = \beta_0 + \beta_1 lny + \beta_2 age_i + \beta_3 smell_i + u_i$$
$$WTP_i = WTP_i^* \quad \text{if} \quad WTP_i^* > 0$$
$$WTP_i = 0 \quad \text{if} \quad WTP_i^* < 0$$

	OLS		Tobit		
Variable	Estimate	t-stat	Estimate	t-stat	Marginal effect
lny	2.515	2.74	2.701	2.5	2.64
age	-.1155	-2.00	-.20651	-3.0	-0.19
sex	.4084	0.28	.14084	0.0	.137
smell	-1.427	-0.90	-1.8006	-0.9	-1.76
constant	-4.006	-0.50	-3.6817	-0.4	

Endogenous Explanatory Variables

- Suppose that we allow one of the explanatory variables to be endogenous:

$$y_1 = \max(0, z_1\delta_1 + \alpha_1 y_2 + u_1)$$
$$y_2 = Z\delta_2 + v_2 = z_1\delta_21 + z_2\delta_{22} + v_2$$

- If u_1 and v_2 are correlated, then y_2 is endogenous. We assume that (u_1, v_2) are zero-mean normally distributed.
- We can write $u_1 = \theta_1 v_2 + e_1$, where $\theta_1 = corr(u_1, v_2)$, so that:

$$y_1 = \max(0, z_1\delta_1 + \alpha_1 y_2 + \theta_1 v_2 + e_1)$$

  If we observed v_2, we could include it in the regression to get a consistent estimator of δ_1 and α_1.
- Two step procedure 1) Regress y_2 on Z and get $\hat{v}_2$, 2) Include $\hat{v}_2$ in the tobit equation.

Sample Selection

- In many cases, in (micro) econometrics, we have to face the problem of sample selection. Our estimators will be consistent estimators for the population under consideration. However, we often want to say something about a much more general population.
- Sample selection can take many forms.
  - Sample selected on the basis of an explanatory variable. For instance we may want to investigate the effect of age and experience on wages, but only have individuals between 20 and 30.
  - Sample selected based on the endogenous variable: We only observe wages of those who work, which limits our ability to study the return to education for instance.

Notations

- The population is represented by a random vector (x, Y, z), where z is a vector of instruments.

$$Y = x\beta + u, \qquad E(u|z) = 0$$

- Rather than using a random sample, we use only data satisfying $s = 1$, where s is an indicator of selection.

When Can Sample Selection be Ignored?

- Suppose that
$$E(u|z,s) = 0$$
- This can happen when s is a deterministic function of z, the selection is a fixed rule involving only the exogenous variables z, or if selection is independent of (z, u).
- When x is exogenous and the selection is based solely on the explanatory variable, then OLS is consistent in the selected sample.
- When x is endogenous, and the selection is based solely on exogenous variables, 2SLS is consistent in the selected sample.
- However, if the selection operates through the endogenous variable, then $E(u|z,s) \neq 0$ and neither OLS or 2SLS are consistent.

Selection on the Basis of the Response Variable

- We assume that y_i is continuous and that the selection takes the following form:

$$s_i = I[a_1 < Y_i < a_2]$$

- where a_1 and a_2 are known constants. The cdf of Y_i conditional on $(x_i, s_i = 1)$ is:

$$P(Y_i \leq y|x, s_i = 1) = \frac{P(Y_i \leq y, s_i = 1|x)}{P(s_i = 1|x)}$$

$$\text{and } \; P(s_i = 1|x) = P(a_1 < Y_i < a_2|x) = F(a_2|x) - F(a_1|x)$$

$$P(Y_i \leq y, s_i = 1|x) = P(a_1 < Y_i \leq y|x_i) = F(y|x_i) - F(a_1|x_i)$$

- Taking the derivative with respect to y, we get the pdf:

$$f(y|x_i, s_i = 1) = \frac{f(y|x_i)}{F(a_2|x_i) - F(a_1|x_i)}$$

which can be used in a Maximum Likelihood framework.

Probit Selection Equation

- So far, the selection took a deterministic form. We now turn to a stochastic form of sample selection, where participation is determined by a probit equation.
- Example: Labor supply.

$$\max_h u(w_i h + a_i, h)$$

We observe individual i working if the wage is above the reservation wage, w_i^R.

$$\begin{aligned} w_i &= \exp(x_{1i}\beta_1 + u_{i1}) \\ w_i^R &= \exp(x_{2i}\beta_2 + \gamma_2 a_i + u_{i2}) \end{aligned}$$

but the wage is only observed if

$$\log w_i - \log w_i^R = x_{i1}\beta_1 - x_{i2}\beta_2 - \gamma_2 a_i + u_{i1} - u_{i2} = x_i\delta_2 + v_{i2} > 0$$

- The model can be written more compactly as

$$y_1 = x_1\beta_1 + u_1$$
$$y_2 = I[x\delta_2 + v_2 > 0]$$

- We assume that we always observe (x, y_2). y_1 is only observed when $y_2 = 1$.
- We assume that (u_1, v_2) is independent of x with zero mean, $v_2 \sim N(0,1)$ and $E(u_1|v_2) = \gamma_1 v_2$.
- There are more cases than the labor supply example. For instance, we may wish to model the fact that some individuals drop out of a program that we are evaluating.
- We can hope to estimate $E(y_1|x, y_2 = 1)$ and $P(y_2 = 1|x)$.

- The conditional mean of y_1 can be expressed as:

$$
\begin{aligned}
E(y_1|x, y_2) &= x_1\beta_1 + E(u_1|x, y_2) \\
&= x_1\beta_1 + \gamma_1 E(v_2|x, y_2) \\
&= x_1\beta_1 + \gamma_1 E(v_2|v_2 > -x\delta_2) \\
&= x_1\beta_1 + \gamma_1 \lambda(x\delta_2) \qquad \text{with } \lambda(.) = \phi(.)/\Phi(.)
\end{aligned}
$$

- Hence, we can regress y_1 (the ones we observe) on x_1 and the inverse Mills ratio and get consistent estimates of β_1 and γ_1.
- We can get a consistent estimator for δ_2 (and $\lambda(x\delta_2)$) from the probit equation in a first step. This procedure is sometimes called Heckit.
- Note that the model is identified even if $x = x_1$, due to the nonlinearity of $\lambda()$. However, this all due to the parametric assumptions (normality). A more convincing identification scheme would be to have exclusion restrictions.

Maximum Likelihood Estimation

- The model can be estimated in two-steps, a probit first and then OLS. This is more robust, but there is an issue about standard errors.
- If we assume that (u_1, v_2) is bivariate normal with mean zero, $\text{Var}(u_1) = \sigma_1^2$ and $\text{cov}(u_1, v_2) = \sigma_{12}$, $\text{Var}(v_2)$=1, then we can estimate the model in one step, through (partial) maximum likelihood.
- One can generalize the estimation method in order not to make any distributional assumptions (Ahn and Powell (1993)).

Section 9

Likelihood Based Hypothesis Testing

Likelihood Based Hypothesis Testing

Likelihood Based Hypothesis Testing

- We now consider test of hypotheses in econometric models in which the complete probability distribution of outcomes given conditioning variables is specified.
- There are three natural ways to develop tests of hypotheses when a likelihood function is available.
    1. Is the unrestricted ML estimator significantly far from the hypothesised value? This leads to what is known as the Wald test.
    2. If the ML estimator is restricted to satisfy the hypothesis, is the value of the maximised likelihood function significantly smaller than the value obtained when the restrictions of the hypothesis are not imposed? This leads to what is known as the likelihood ratio test.
    3. If the ML estimator is restricted to satisfy the hypothesis, are the Lagrange multipliers associated with the restrictions of the hypothesis significantly far from zero? This leads to what is known as the Lagrange multiplier or score test.

Likelihood Based Hypothesis Testing

- In the normal linear regression model all three approaches, after minor adjustments, lead to the same statistic which has an $F^{(j)}_{(n-k)}$ distribution when the null hypothesis is true and there are j restrictions.
- Outside that special case, in general the three methods lead to different statistics, but in large samples the differences tend to be small.
- All three statistics have, under certain weak conditions, $\chi^2_{(j)}$ limiting distributions when the null hypothesis is true and there are j restrictions.
- The exact distributional result in the normal linear regression model fits into this large sample theory on noting that $\operatorname{plim}_{n\to\infty}\left(jF^{(j)}_{(n-k)}\right) = \chi^2_{(j)}$.

Test of Hypothesis

- We now consider tests of a hypothesis $H_0 : \theta_2 = 0$ where the full parameter vector is partitioned into $\theta' = [\theta_1' \vdots \theta_2']$ and θ_2 contains j elements. Recall that the MLE has the approximate distribution

$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_0)$$

where

$$V_0 = -\operatorname*{plim}_{n \to \infty} (n^{-1} Q(\theta_0; Y))^{-1} = \bar{I}(\theta_0)^{-1}$$

and $\bar{I}(\theta_0)$ is the asymptotic information matrix per observation.

Wald Test

- This test is obtained by making a direct comparison of $\hat{\theta}_2$ with the hypothesised value of θ_2, zero.
- Using the approximate distributional result given above leads to the following test statistic.

$$S_W = n\hat{\theta}_2' \widehat{V}_{22}^{-1} \hat{\theta}_2'$$

where $\widehat{V}_{22}$ is a consistent estimator of the lower right hand $j \times j$ block of V_0.

- Under the null hypothesis $S_W \xrightarrow{d} \chi^2_{(j)}$ and we reject the null hypothesis for large values of S_W.
- Using one of the formulas for the inverse of a partitioned matrix the Wald statistic can also be written as

$$S_W = n\hat{\theta}_2' \left(\widehat{\bar{I}}(\hat{\theta})_{22} - \widehat{\bar{I}}(\hat{\theta})_{21}' \widehat{\bar{I}}(\hat{\theta})_{11}^{-1} \widehat{\bar{I}}(\hat{\theta})_{12} \right) \hat{\theta}_2'$$

where the elements $\widehat{\bar{I}}(\hat{\theta})_{ij}$ are consistent estimators of the appropriate blocks of the asymptotic Information Matrix per observation evaluated at the (unrestricted) MLE.

The Score - or Lagrange Multiplier - test

- Sometimes we are in a situation where a model has been estimated with $\theta_2 = 0$, and we would like to see whether the model should be extended by adding additional parameters and perhaps associated conditioning variables or functions of ones already present.
- It is convenient to have a method of conducting a test of the hypothesis that the additional parameters are zero (in which case we might decide not to extend the model) without having to estimate the additional parameters. The *score test* provides such a method.

The Score - or Lagrange Multiplier - test

- The score test considers the gradient of the log likelihood function evaluated at the point
$$\hat{\theta}^R = [\hat{\theta}_1^{R\prime}, 0]'$$
and examines the departure from zero of that part of the gradient of the log likelihood function that is associated with θ_2.
- Here $\hat{\theta}_1^R$ is the MLE of θ_1 when θ_2 is restricted to be zero. If the unknown value of θ_2 is in fact zero then this part of the gradient should be close to zero. The score test statistic is
$$S_S = n^{-1} q(\hat{\theta}^R; Y)' \widehat{\bar{I}}(\hat{\theta}^R)^{-1} q(\hat{\theta}^R; Y)$$
and $S_S \xrightarrow{d} \chi^2_{(j)}$ under the null hypothesis. There are a variety of ways of estimating $\widehat{\bar{I}}(\theta_0)$ and hence its inverse.
- Note that the complete score (gradient) vector appears in this formula. Of course the part of that associated with θ_1 is zero because we are evaluating at the restricted MLE. That means the score statistic can also be written, using the formula for the inverse of a partitioned matrix, as the algebraically identical

Likelihood ratio tests

- The final method for constructing hypothesis tests that we will consider involves comparing the value of the maximised likelihood function at the restricted MLE ($\hat{\theta}^R$) and the unrestricted MLE (now written as $\hat{\theta}^U$).
- This likelihood ratio test statistic takes the form

$$S_L = 2\left(L(\hat{\theta}^U; Y) - L(\hat{\theta}^R; Y)\right)$$

and it can be shown that under H_0, $S_L \xrightarrow{d} \chi^2_{(j)}$.

Specification Testing

- Maximum likelihood estimation requires a complete specification of the probability distribution of the random variables whose realisations we observe.
- In practice we do not *know* this distribution though we may be able to make a good guess. If our guess is badly wrong then we may produce poor quality estimates, for example badly biased estimates, and the inferences we draw using the properties of the likelihood function may be incorrect.
- In regression models the same sorts of problems occur. If there is heteroskedasticity or serial correlation then, though we may produce reasonable point estimates of regression coefficients if we ignore these features of the data generating process, our inferences will usually be incorrect if these features are not allowed for, because we will use incorrect formulae for standard errors and so forth.
- It is important then to seek for evidence of departure from a model specification, that is to conduct *specification tests*.
- In a likelihood context the score test provides an easy way of

Detecting Heteroskedasticity

- We consider one example here, namely detecting heteroskedasticity in a normal linear regression model.
- In the model considered, $Y_1, \ldots, Y_n$ are independently distributed with Y_i given x_i being $N(x_i'\beta, \sigma^2 h(z_i'\alpha))$ where $h(0) = 1$ and $h'(0) = 1$, both achievable by suitable scaling of $h(\cdot)$.
- Let $\theta^U = [\beta, \sigma^2, \alpha]$ and let $\theta^R = [\beta, \sigma^2, 0]$. A score test of $H_0 : \alpha = 0$ will provide a specification test to detect heteroskedasticity.
- The log likelihood function when $\alpha = 0$, in which case there is homoskedasticity, is as follows.

$$L(\theta^R; y|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

whose gradients with respect to β and σ^2 are

$$\begin{aligned} q_\beta(\theta^R; y|x) &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)\,x_i \\ q_{\sigma^2}(\theta^R; y|x) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \end{aligned}$$

Detecting Heteroskedasticity

- The log likelihood function for the unrestricted model is

$$L(\theta^U; y|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\log h(z_i'\alpha) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2}{h(z_i'\alpha)}$$

whose gradient with respect to α is

$$q_\alpha(\theta^U; y|x) = -\frac{1}{2}\sum_{i=1}^{n}\frac{h'(z_i'\alpha)}{h(z_i'\alpha)}z_i + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2\, h'(z_i'\alpha)}{h(z_i'\alpha)^2}z_i$$

which evaluated at the restricted MLE (for which $\alpha = 0$) is

$$\begin{aligned} q_\alpha(\hat\theta^R; y|x) &= -\frac{1}{2}\sum_{i=1}^{n}z_i + \frac{1}{2\hat\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i'\hat\beta\right)^2 z_i \\ &= \frac{1}{2\hat\sigma^2}\sum_{i=1}^{n}\left(\hat\varepsilon_i^2 - \hat\sigma^2\right)z_i. \end{aligned}$$

- The specification test examines the correlation between the squared OLS residuals and z_i. The score test will lead to rejection when this correlation is large.

Information Matrix Tests

- We have seen that the results on the limiting distribution of the MLE rest at one point on the Information Matrix Equality

$$E[q(\theta_0, Y)q(\theta_0, Y)'] = -E[Q(\theta_0, Y)]$$

  where $Y = (Y_1, \ldots, Y_n)$ are n random variables whose realisations constitute our data.

- In the case relevant to much microeconometric work the log likelihood function is a sum of independently distributed random variables, e.g. in the continuous Y case:

$$L(\theta, Y) = \sum_{i=1}^{n} \log f(Y_i, \theta),$$

  where $f(Y_i, \theta)$ is the probability density function of Y_i. Here the Information Matrix Equality derives from the result

$$E[\frac{\partial}{\partial\theta} \log f(Y, \theta) \frac{\partial}{\partial\theta'} \log f(Y, \theta) + \frac{\partial^2}{\partial\theta\partial\theta'} \log f(Y, \theta)] = 0.$$

- Given a value $\hat{\theta}$ of the MLE we can calculate a sample analogue of

Section 10

Panel Data

Introduction

- Panel data consist of multiple observations over time of the same individuals. e.g. wages over time for many workers.
- Usually, we have small T (time-dimension) and large N (cross-section dimension), so it is not feasible to run separate regressions for each individuals.
- So the focus is more on the heterogeneity across individuals than complex modeling of the time-series dynamic.
- However, overlooking unobserved heterogeneity often leads to conclude (wrongly) that state dependence is important. Both phenomenons imply different types of policy.

Introduction

- The basic model is a regression model of the form:

$$\begin{aligned} y_{it} &= x_{it}\beta + z_i\alpha + \varepsilon_{it} \\ &= x_{it}\beta + c_i + \varepsilon_{it} \end{aligned}$$

- There are K regressors in x_{it}, not including a constant. The heterogeneity, or individual effect is $z_i\alpha = c_i$. Note that if z_i is observed, this is a classic estimation problem, OLS is consistent. The problem arises if c_i is unobserved.

Introduction

- The main objective is the consistent and efficient estimation of the partial effects:

$$\beta = \frac{\partial E[y_{it}|x_{it}]}{\partial x_{it}}$$

- We assume strict exogeneity for the independent variables:

$$E[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots] = 0$$

Model Structures

- **Pooled regressions**: If z_i contains only a constant term, then OLS provides consistent and efficient estimates of α and β.
- **Random Effects:** the unobserved individual heterogeneity is assumed to be *uncorrelated* with x_{it}. Simpler but more difficult as an assumption.
- **Fixed Effects:** if z_i is unobserved but correlated with x_{it}, then OLS is inconsistent.

Pooled Regression Model

$$y_{it} = X_{it}\beta + v_{it} \qquad \text{with} \quad v_{it} = c_i + u_{it}$$

- Under certain assumptions the pooled OLS estimator can be used to obtain a consistent estimator of β.
- We have to assume that:

$$E(X_{it}' u_{it}) = 0$$

$$E(X_{it}' c_i) = 0$$

- The unobserved heterogeneity introduces **autocorrelation**:

$$E[v_{it} v_{is}] = \sigma_c^2 \qquad \text{if } t \neq s$$

- OLS may be consistent, but inefficient. We need to use a robust variance matrix estimator.

Random Effects Methods

- The random effect method puts also c_i into the error term. It assumes:

$$E(u_{it}|X_i, c_i) = 0, \qquad t = 1, \ldots T$$

$$E(c_i|X_i) = E(c_i) = 0$$

- Here we assume that the unobserved effect c_i is orthogonal to all elements in X_i. This may be a very strong assumption. Note that it is a stronger assumption than the one we made for the pooled OLS. This is because the method explicitly deals with the autocorrelation induced by the unobserved effect in a GLS framework.

Random Effect Methods

- Write the model for all T time periods as:

$$y_i = X_i\beta + v_i$$

- The variance matrix of v_i (of dimension $T \times T$) is:

$$\Omega = E(v_i v_i') = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \dots & \vdots \\ \vdots & & & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

- Ω depends only on two parameters.

Random Effect Estimator

- The random effects estimator is:

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \hat{\Omega}^{-1} y_i \right)$$

- As with FGLS, we need first an estimate of $\hat{\Omega}$. As OLS is consistent with the assumptions made above, we can back out the estimates from the residual of a pooled OLS regression. Denote $\hat{\hat{v}}_i$ such a residual. An estimator for $\hat{\sigma}_v$ is:

$$\hat{\sigma}_v^2 = \frac{1}{NT - K} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\hat{v}}_{it}^2$$

Random Effect Estimator

- However, we need an estimator for both σ_c and σ_u. Recall that $\sigma_c^2 = E(v_{it}v_{is})$ for $t \neq s$.

$$\begin{aligned} E(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} v_{it}v_{is}) &= \sum_{t=1}^{T-1}\sum_{s=t+1}^{T} E(v_{it}v_{is}) \\ &= \sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \sigma_c^2 \\ &= \sigma_c^2 \frac{T(T-1)}{2} \end{aligned}$$

- Hence, a consistent estimator is

$$\hat{\sigma}_c^2 = \frac{1}{(NT(T-1)/2 - K)} \sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \hat{\hat{v}}_{it}\hat{\hat{v}}_{is}$$

- Note that we can also test for the presence of random effects. The absence of random effects can be tested with $H_0 : \sigma_c^2 = 0$.

Testing for Random Effects

- One way to test for random effects is to test for serial correlation in the residual, with an AR(1) test. This test is valid as the residuals are uncorrelated under the null of no random effects.
- We can test for random effects by testing more directly:

$$H_0: \quad \sigma_c^2 = 0 \qquad H_1: \sigma_c^2$$

the test statistic is:

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{v}_{it}\hat{v}_{is}}{\left(\sum_{i=1}^{N}\left(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{v}_{it}\hat{v}_{is}\right)^2\right)^{1/2}} \sim N(0,1) \text{ under the null}$$

This test is quite general, so it is not clear why we reject it when we find significant serial correlation.

Fixed Effect Methods

- We assume $E(u_{it}|X_i, c_i) = 0 \qquad t, = 1, 2, \ldots, T$ (strict exogeneity of $\{X_{it}\}$.
- However, $E(c_i|X_{it})$ can be whichever function of X_{it}.
- As c_i is correlated with X_{it}, we have to get rid of it. There are several ways to do so:
  - fixed effect transformation (within transformation).
  - first differences.
- The fixed effect transformation is obtained by averaging the model over time periods

$$\bar{y}_i = \bar{X}_i\beta + c_i + \bar{u}_i$$

- If we substract this equation from the original one, we get:

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + u_{it} - \bar{u}_i$$

$$\ddot{y}_{it} = \ddot{X}_{it}\beta + \ddot{u}_{it}$$

Fixed Effect Methods

$$\ddot{y}_{it} = \ddot{X}_{it}\beta + \ddot{u}_{it}$$

- This auxiliary model can be estimated by pooled OLS, as $E(\ddot{X}_{it}'\ddot{u}_{it}) = 0$.
- The estimator is:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X}_{it}'\ddot{X}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X}_{it}'\ddot{y}_{it}\right)$$

- The fixed effect is consistent. It is also efficient if

$$E(u_i u_i'|X_i, c_i) = \sigma_u^2 I_T$$

i.e. constant variance across t and not serially correlated.

Fixed Effect GLS

- If we do not want to assume that $E(u_i u_i'|X_i, c_i) = \sigma_u^2 I_T$, then we can relax this assumption and posit instead that $E(\ddot{u}_i \ddot{u}_i'|X_i, c_i) = \Omega$ (which is equivalent to $E(u_i u_i'|X_i, c_i) = \Lambda$).

$$\hat{\beta}_{FEGLS} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{X}_{it}' \hat{\Omega}^{-1} \ddot{X}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{X}_{it}' \hat{\Omega}^{-1} \ddot{y}_{it} \right)$$

with

$$\hat{\Omega} = N^{-1} \sum_{i=1}^{N} \hat{\hat{u}}_i \hat{\hat{u}}_i'$$

Between Estimator

- The between estimator is the OLS applied to the time-averaged equation.

$$\bar{y}_i = \bar{X}_i \beta + c_i + \bar{u}_i$$

$$\hat{\beta}_{BE} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \bar{X}_i' \bar{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \bar{X}_i' \bar{y}_i \right)$$

- Note that this estimator is not necessarily consistent because $E(u_{it}|X_i, c_i) = 0$ does not guarantee that $E(\bar{X}_i' c_i) = 0$.

Estimation with First Differences

- We can eliminate c_i by lagging the model one period and subtracting it
$$\Delta y_{it} = \Delta X_{it} \beta + \Delta u_{it}$$
- With this procedure, we loose one wave of data.
- We have now to deal with an error term Δu_{it} which is autocorrelated (moving average of order one).
- The first difference estimator is the pooled OLS estimator from the regression of Δy_{it} on ΔX_{it}.
- The estimator is consistent as we have assumed $E(\Delta X'_{it} \Delta u_{it}) = 0$.

Fixed-Effect versus First-Differences

- If $T = 2$, the fixed effect and first differences estimators are identical.
- If $T > 2$, the choice between First Differences and Fixed Effects depends on the assumptions about the error term u_{it}.
  - The fixed effect estimator is more efficient if the error term is serially uncorrelated.
  - The first difference estimator is more efficient if the error term follows a random walk.

Return to Experience

Adda et al. (2013) "Career Progression, Economic Downturns, and Skills", NBER Working Paper No. 18832.

- They use administrative data from Germany (IAB) which record wages and work experience for a very large random sample of workers.
- Focus on low skilled individuals (about 75% of a birth cohort) who either leave school at age 16 or go through vocational training.
- Observe all wages earned in all jobs for up to 20 years.

Example: Return to Experience

Example: Return to Experience

- Use administrative data from Germany (IAB) which record wages and work experience for a very large random sample of workers.

	OLS	GLS Random Effect	Between Effects	Within
experience	0.033	0.03	0.054	0.030
	(0.0009)	(0.0002)	(0.002)	(0.0002)
Apprentice	0.055	0.051	0.018	-
	(0.015)	(0.013)	(0.014)	
constant	4.43	4.43	4.30	4.500
	(0.015)	(0.013)	(0.016)	(0.0014)
Number of obs		81442		

Random Effects and Fixed Effect Estimators

- The goal of the following derivation is to find a transformation under which we get rid of the heteroskedasticity.
- Suppose we can find a matrix C_T such that $C_T \Omega C_T = \sigma_u^2 I_T$. Then we can apply C_T to our model:

$$\begin{aligned} C_T y_i &= C_T X_i \beta + C_T \nu_i \qquad \nu_i = c_i + u_i \\ \check{y}_i &= \check{X}_i \beta + \check{\nu}_i \end{aligned}$$

  This new (transformed) model has a homoskedastic error term.

- We'll see that $C_T = I_T - \lambda P_T$ with $P_T = \begin{bmatrix} T^{-1} & \dots & T^{-1} \\ \vdots & \dots & \vdots \\ T^{-1} & \dots & T^{-1} \end{bmatrix}$ and $\lambda = 1 - \sqrt{\frac{1}{1+T(\sigma_c^2/\sigma_u^2)}}$

- We now apply OLS to get the Random Effect estimator:

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \breve{X}_{it}' \breve{X}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \breve{X}_{it}' \breve{y}_{it} \right)$$

- Note that the transformed model can be written:

$$y_{it} - \lambda \bar{y}_i = (X_{it} - \lambda \bar{X}_i)\beta + \nu_{it} - \lambda \bar{\nu}_i$$

- The random effect estimator is obtained by quasi-time demeaning. We remove a fraction of the time average from the left and right hand side.
  - If λ is close to one, random effects and fixed effect are similar. This happens either when T is large or when $\sigma_c^2 >> \sigma_u^2$.
  - If λ is close to 0, then Random Effects is close to pooled OLS.

Proof

- Define J_T as a Tx1 vector of ones. We can express the covariance matrix of $c_i + u_i$ as:

$$\begin{aligned}\Omega &= \sigma_u^2 I_T + \sigma_c^2 J_T J_T' = \sigma_u^2 I_T + T\sigma_c^2 J_T (J_T' J_T)^{-1} J_T' \\ &\quad \text{(where we have used the fact that } J_T' J_T = T) \\ &= \sigma_u^2 I_T + T\sigma_c^2 P_T \qquad \text{with } P_T = J_T (J_T' J_T)^{-1} J_T' \\ &= (\sigma_u^2 + T\sigma_c^2)(P_T + \eta Q_T) \qquad \textit{with } \; Q_T = I_T - P_T \\ &\quad \text{where } \eta = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2}\end{aligned}$$

- Note that $P_T P_T = P_T$ and $Q_T Q_T = Q_T$ (These matrices are idempotent).

- Define $S_T = P_T + \eta Q_T$, then $S_T^{-1} = P_T + 1/\eta Q_T$ as $S_T S_T^{-1} = I_T$.
- It can also be shown that $S_T^{-1/2} = P_T + \frac{1}{\sqrt{\eta}} Q_T$, as $S_T^{-1/2} S_T^{-1/2} = S^{-1}$ (check this).
- Further, we can write $S_T^{-1/2} = (1-\lambda)^{-1}(I_T - \lambda P_T)$, $\lambda = 1 - \sqrt{\eta}$
- Therefore, we can get an expression for $\Omega^{-1/2}$:

$$\begin{aligned}
\Omega^{-1/2} &= (\sigma_u^2 + T\sigma_c^2)^{-1/2}(1-\lambda)^{-1}(I_T - \lambda P_T) \\
&= \frac{1}{\sigma_u}(I_T - \lambda P_T)
\end{aligned}$$

- and

$$\lambda = 1 - \frac{\sigma_u}{\sqrt{\sigma_u^2 + T\sigma_c^2}}$$

Dynamic Panel-Data Models

- We are interested in estimating the parameters of models of the form

$$y_{it} = \gamma y_{it-1} + X_{it}\beta + c_i + u_{it}$$

  for $i = 1, \dots, N$ and $t = 1, \dots, T$ using datasets with large N and fixed T.
- By construction, y_{it-1} is correlated with the unobserved individual-level effect c_i.
- Removing c_i by the within transform (removing the panel-level means) produces an inconsistent estimator with T fixed.
- First difference both sides and look for instrumental-variables (IV) and generalized method-of-moments (GMM) estimators .

The Anderson and Hsiao Estimator

- First differencing the model equation yields:

$$\Delta y_{it} = \Delta y_{it-1}\gamma + \Delta X_{it}\beta + \Delta u_{it-1}$$

- We have eliminated c_i, but y_{it-1} in Δy_{it-1} is a function of u_{it-1} in Δu_{it-1}.

$$cov(\Delta y_{it-1}, \Delta u_{it}) = -\sigma_u^2$$

- Anderson and Hsiao (1981) suggest a 2SLS estimator based on further lags of Δy_{it} as instruments for Δy_{it-1}. For instance, if u_{it} is iid over i and t, then y_{it-2} would be a valid instrument for Δy_{it-1}.

Arellano-Bond Estimator

- Arellano and Bond (1991) show how to construct estimators based on moment equations constructed from further lagged levels of y_{it} and the first-differenced errors.
- There are in fact a huge number of instruments to be used. The exact number depends on the assumption we are willing to make on the exogeneity of X_{it}.
  - Strict exogeneity: $E(u_{it}|X_{is}, c_i) = 0, \;\; s = 1, \ldots, T$.
  - Predetermined variables: $E(u_{it}|X_{is}, c_i) = 0, \;\; s = 1, \ldots, t$.

Arellano-Bond Estimator

- Write $\eta_{it} = c_i + u_{it}$
- Suppose we only have two periods of data. We can write the following moment conditions:

$$E\left[\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix}(\eta_{i1} - \bar{\eta}_i)\right] = 0 \qquad E\left[\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix}(\eta_{21} - \bar{\eta}_i)\right] = 0$$

$$E\left[\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix}\bar{\eta}_i\right] = 0$$

- To use more compact notations, rewrite the model such as:

$$\tilde{y}_{it} = \tilde{X}_{it}\theta + \tilde{\eta}_{it}$$

where $\tilde{y}_{it} = \Delta y_{it}$ and so on, and $\theta = [\gamma, \beta']$.
- Stacking these quantities gives the following matrices:

$$\tilde{y}_i = \begin{bmatrix} \Delta y_{i3} \\ \Delta y_{i4} \\ \vdots \\ \Delta y_{iT} \end{bmatrix} \qquad \tilde{X}_i = \begin{bmatrix} \Delta y_{i2} & \Delta X'_{i3} \\ \Delta y_{i3} & \Delta X'_{i4} \\ \vdots & \\ \Delta y_{iT-1} & \Delta X'_{iT} \end{bmatrix}$$

Matrix of Instruments

$$Z_i = \begin{bmatrix} y_{i,1}, X'_{i,1}, \ldots, X'_{i,T} & 0 & \ldots & 0 \\ 0 & y_{i,1}, y_{i,2}, X'_{i,1}, \ldots, X'_{i,T} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \begin{matrix} y_{i,1}, y_{i,2}, \ldots, y_{iT-2}, \\ X'_{i,1}, \ldots, X'_{i,T} \end{matrix} \end{bmatrix}$$

- This matrix of instrument is valid under strict exogeneity of the variable $\{X_{it}\}$.
- If $\{X_{it}\}$ is predetermined, then we would not use all the Xs, but only those up to period t.

Estimator

$$\hat{\theta}_{IV} = \left[\left(\sum_{i=1}^{n} \tilde{X}_i' Z_i\right)\left(\sum_{i=1}^{n} Z_i' Z_i\right)^{-1}\left(\sum_{i=1}^{n} Z_i' \tilde{X}_i\right)\right]^{-1} \cdot \left[\left(\sum_{i=1}^{n} \tilde{X}_i' Z_i\right)\left(\sum_{i=1}^{n} Z_i' Z_i\right)^{-1}\left(\sum_{i=1}^{n} Z_i' \tilde{y}_i\right)\right]$$

- This estimator is an instrumental variable estimator. We can generalize this method to get an estimator which is more efficient, by using GMM.

GMM Estimator

- The optimal GMM estimator is obtained as:

$$
\begin{aligned}
\hat{\theta}_{GMM} &= \left[\left(\sum_{i=1}^{n} \tilde{X}_i' Z_i \right) \left(\sum_{i=1}^{n} Z_i' \Sigma Z_i \right)^{-1} \left(\sum_{i=1}^{n} Z_i' \tilde{X}_i \right) \right]^{-1} \\
&\quad \cdot \left[\left(\sum_{i=1}^{n} \tilde{X}_i' Z_i \right) \left(\sum_{i=1}^{n} Z_i' \Sigma Z_i \right)^{-1} \left(\sum_{i=1}^{n} Z_i' \tilde{y}_i \right) \right]
\end{aligned}
$$

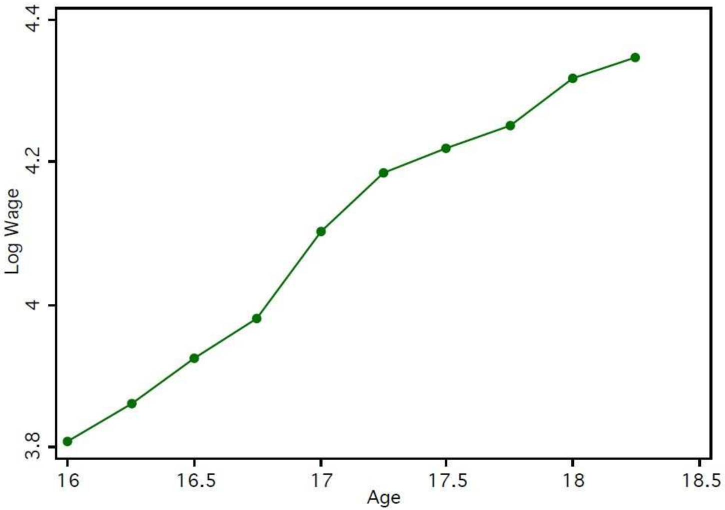
with $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} Z_i' \hat{\eta}_i \hat{\eta}_i' Z_i$

- To get an estimator of $\hat{\Sigma}$ we need to compute the predicted residuals, and we therefore need a first guess for $\hat{\theta}$. We can start with $\hat{\theta}_{IV}$ and then go on and do GMM.

Example: Dynamic of Wages

- We have a sample of young unskilled Germans.
- We observe their log wage at a quarterly frequency, together with the amount of work experience they get.
- The sample contains 419 individuals, observed for two years, between age 16 and 19.
- In total, we have 2106 observations.

Wage Path



Empirical Question:

- We are interested in exploring the determinants of wages in this very homogenous sample.
- In particular, what is the role of work experience?
- How persistent are wages?
- To this end, we write down the model for log wages:

$$w_{it} = \gamma w_{it-1} + \beta exp_{it} + c_i + u_{it}$$

- The return to experience is measured by β, and the persistence by γ.
- We allow for unobserved heterogeneity, which may be due to differences in ability.

Estimation Results

	OLS	RE	FD	FE	AB 1	AB 2
Lagged wage	0.965	0.965	-0.04	0.765	0.767	0.695
	(0.005)	(0.005)	(0.025)	(0.015)	(0.11)	(0.07)
Experience	0.011	0.011	0.147	0.042	0.028	0.038
	(0.005)	(0.005)	(0.014)	(0.006)	(0.016)	(0.011)

RE: Random Effect, FD: First Difference, FE: Fixed effect, AB1,2: Arellano-Bond, without or with predetermined experience.

- Note that the persistence γ decreases when we allow for unobserved heterogeneity.

Example: Return to Seniority

- Altonji and Shakotko (1987) "Do Wages Rise with Job Seniority?", *Review of Economic Studies*.
- Wages of individuals who stay longer in a given firm appears to be higher. Why?
- Is this due to firm specific human capital? Is this a statistical artefact?
- Matters to inform policy: for instance, a number of labor market policies try to provides temporary jobs to unemployed. Their effect depends on the return to human capital acquired in firms. Matters also to understand job to job mobility.

The Empirical Model

For individual i in job j in period t:

$$w_{ijt} = b_0 X_{ijt} + b_1 T_{ijt} + b_2 T_{ijt}^2 + b_3 O_{ijt} + \varepsilon_{ijt}$$

$$\varepsilon_{ijt} = \varepsilon_i + \varepsilon_{ij} + \eta_{ijt}$$

- w_{ijt} is the log real wage, X_{ijt} is a vector of characteristics of the person, the job and labor market experience.
- T_{ijt} is the duration in job j, O_{ijt} is a dummy for $T > 1$.
- The model includes an individual fixed effect as well as a job-person fixed effect.

Sources of Bias

- The tenure variables are likely to be correlated with the error term:
  - High productivity individuals (ε_i) are probably less likely to experience layoffs or quits. Health problems are likely to be positively correlated with quits and negatively correlated with tenure, productivity and wages.
  - Tenure is likely to be correlated to the person-firm match effect ε_{ij}. Individuals who have a good match are less likely to move to another job.
  - Individuals who move to a new job usually do so because the new job pays better.
- Removing the individual fixed effect does not solve the endogeneity problem completely as there is still endogeneity associated with the match effect ε_{ij}.

Econometric Methodology

- One way to estimate the model is to employ a within estimator:

$$\begin{aligned} w_{ijt} - \bar{w}_{ijt} &= b_0(X_{ijt} - \bar{X}_{ij}) + b_1(T_{ijt} - \bar{T}_{ij}) + b_2(T_{ijt}^2 - \bar{T}_{ij}^2) \\ &\quad + b_3(O_{ijt} - \bar{O}_{ij}) + \varepsilon_{ijt} - \bar{\varepsilon}_{ij} \end{aligned}$$

- However, with this specification, we cannot separately identify b_0 and b_1 (why?).
- Hence, there is a need for an alternative estimation method.

Econometric Methodology

- Construct an instrument for tenure: deviation from the mean of tenure over an employment spell:

$$\tilde{T}_{ijt} = T_{ijt} - \bar{T}_{ij}$$

- $\tilde{T}_{ijt}$ sums to zero over the periods in job j. Hence it is orthogonal to the individual and firm specific shocks ε_i and ε_{ij}.
- It is also correlated with tenure by construction.
- However, there may still be some endogeneity. Experience may not be independent of the individual fixed effect. Not taken care of in the study.
- The fixed effects introduce serial correlation. Also try a GLS method

Data

- 1968-1981 waves of the Panel Study of Income Dynamics.
- Focus on white males, head of households, between 18 and 60.
- not retired, not disabled, not self-employed, not employed by the government.
- In total, about 15000 observations on 2163 individuals and 4334 job matches.

Results: OLS and IV

		OLS		IV_1		GLS	IV_1-GLS	IV_2	IV_3
Variable[a]	Mean (St Dev)	1	2	3	4	5	6	7	8
Education	12·5	0·0198	0·0179	0·0163	0·0155	0·0451	0·0462	0·0156	0·0162
	(2·91)	(0·0191)	(0·0191)	(0·0199)	(0·0198)	(0·0153)	(0·0154)	(0·0198)	(0·0197)
Education^2	165·1	0·0017	0·0017	0·0019	0·0019	0·0008	0·0008	0·0019	0·0019
	(70·1)	(0·0007)	(0·0007)	(0·0007)	(0·0007)	(0·0006)	(0·0006)	(0·0007)	(0·0007)
Time	1975·7	0·0090	0·0087	0·0092	0·0092	0·0076	0·0087	0·0091	0·0089
	(3·7)	(0·0011)	(0·0011)	(0·0011)	(0·0011)	(0·0008)	(0·0009)	(0·0011)	(0·0011)
Experience	17·8	0·0397	0·0335	0·0480	0·0449	0·0518	0·0589	0·0451	0·0448
	(10·8)	(0·0065)	(0·0066)	(0·0070)	(0·0071)	(0·0041)	(0·0044)	(0·0068)	(0·0067)
Experience^2/10	43·1	−0·0150	−0·0119	−0·0150	−0·0135	−0·0168	−0·0185	−0·0137	−0·0142
	(44·5)	(0·0030)	(0·0030)	(0·0032)	(0·0033)	(0·0017)	(0·0019)	(0·0031)	(0·0031)
Experience^3/100	123·0	0·0016	0·0012	0·0014	0·0012	0·0017	0·0019	0·0012	0·0013
	(169·1)	(0·0005)	(0·0005)	(0·0005)	(0·0005)	(0·0003)	(0·0003)	(0·0005)	(0·0005)
Ed · Exper	214·0	0·0004	0·0004	0·0005	0·0005	0·0001	0·0002	0·0005	0·0004
	(133·4)	(0·0003)	(0·0003)	(0·0003)	(0·0003)	(0·0002)	(0·0002)	(0·0003)	(0·0003)
T[b]	7·7	0·0280	0·0178	0·0001	−0·0041	0·0044	−0·0043	−0·0036	−0·0002
	(8·0)	(0·0026)	(0·0031)	(0·0021)	(0·0022)	(0·0016)	(0·0019)	(0·0022)	(0·0021)
$T^2/10$[b]	12·3	−0·0055	−0·0027	0·0008	0·0018	0·0006	0·0018	0·0018	0·0017
	(21·6)	(0·0010)	(0·0011)	(0·0008)	(0·0008)	(0·0006)	(0·0007)	(0·0007)	(0·0007)
OLDJOB[b]	0·778		0·1113		0·0501	0·0742	0·0470	0·0485	0·0570
	(0·42)		(0·0126)		(0·0094)	(0·0077)	(0·0088)	(0·0093)	(0·0093)
S.E.[d]		0·403	0·402	0·411	0·410	0·410	0·417	0·408	0·407
Effect of 10 years of T on log wage[b]		0·2419	0·2627	0·0074	0·0268	0·1241	0·0227	0·0307	0·0719
		(0·0180)	(0·0176)	(0·0159)	(0·0162)	(0·0101)	(0·0134)	(0·0160)	(0·0158)
Effect of 10 years of Experience on log wage for high school grads[c]		0·3069	0·2756	0·4008	0·4300	0·3817	0·4416	0·3847	0·3715
		(0·0318)	(0·0319)	(0·0335)	(0·0340)	(0·0196)	(0·0204)	(0·0347)	(0·0324)

Results: Fixed Effects

	Coefficient on					
	$EXP + T + \text{Time}$	T^2	$OLDJOB$	$Ed \cdot EXP$	$EXP^2/10$	$EXP^3/100$
Estimator:						
JOB EFFECTS	0·0642	0·00004	0·0461	−0·00026	−0·0153	0·00153
	(0·0052)	(0·00003)	(0·0072)	(0·00025)	(0·0020)	(0·00031)
$\text{IV}_1 - \text{GLS}$	0·0644	0·00018	0·0470	0·00016	−0·0185	0·00187
	(0·0045)	(0·00007)	(0·0088)	(0·00018)	(0·0019)	(0·00028)