Advanced Econometrics – Part 1

Stefan Sperlich

Université de Genève

May 12, 2020

Literature for Part 1

- Amemiya, T. (1985) Advanced Econometrics, Basil Blackwell Ltd., Oxford.
- Berndt, E. R. (1996) The practice of econometrics: classic and contemporary, Addison Wesley.
- Florens, J.-P., Marimoutou, V., Péguin-Feisolle, A. (2004) Econometrie, Armand Colin, Paris.
- Gourieroux, Ch. (1989) Econometrie des variables qualitatives, Economica, Paris.
- Greene, W. (2003) Econometric Analysis, Prentice Hall.
- Judge, G., Hill, R., Griffiths, W., Lütkepohl, H. (1988) Introduction to the Theory and Practice of Econometrics, New York: Wiley.
- Maddala, G. S. (1986) Limited dependent and Qualitative Variables in Econometrics.
- Wooldridge, J. (2002) Econometric Analysis of Cross Section and Panel Data, MIT Press.

Topics

- Binary Choice (Qualitative Variables)
- Counting Data models
- Multiple and Sequential Choice Models (QV continued)
- Generalized Linear Models and exponential families
- The equivalence of GMM, GLS and Pseudo ML
- Limited dependent variables: Selection Problems, Censoring and Truncation
- Diagnostics: testing in QV models

Binary Choice Models

- Logit, Probit, etc. vs LPM
- Estimation via MLE
- Latent Variable Models

Theoretical Framework, Basics

- Consider a binary dependent variable y, which has only two possible outcomes (0 and 1), and a vector of explanatory variables x thought to influence the realization of y.
- The unconditional expectation of the binary variable y is by definition a probability:

$$E(y) = P(y = 1)$$

- Further, let the set of explanatory variables x influence the outcome of y. Then, the conditional expectation of y given x is:

$$E(y|x) = P(y = 1|x)$$

Towards Regression

Relate this term to the standard regression analysis: for

$$y = F(x, \beta) + u$$

the conditional expectation

$$E(y|x) = E(F(X, \beta) + u|x) = F(x, \beta) + E(u|x) = F(x, \beta)$$

- Hence, the standard regression functional $F(x, \beta)$ is a representation of the conditional expectation of y given x.
- If the dependent variable in a regression relationship is binary, then the regression functional $F(x, \beta)$ equates directly to the conditional probability of observing $y = 1$.
- Thus, the characteristics of binary choice models crucially depend on the way we specify the regression functional $F(x, \beta)$.

The Linear Probability Model

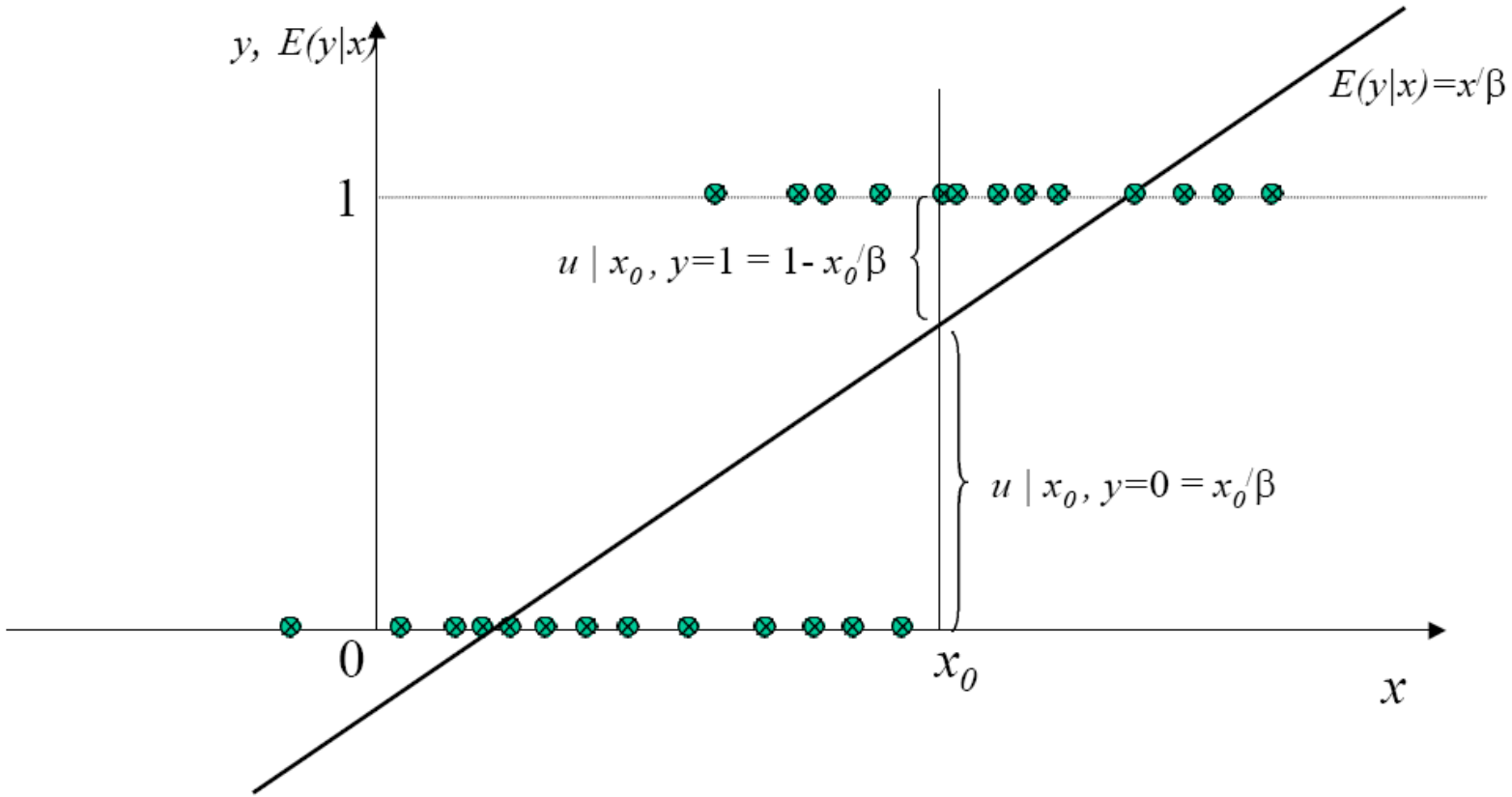
- Consider a binary dependent variable y and a (k–dim.) vector of explanatory variables x
- We may specify the conditional probability directly as:

$$P(y = 1|x) = F(x, \beta) = x'\beta$$

- Introducing random disturbances, we have

$$y = x'\beta + u,$$

  where u represents the stochastic disturbance term in the relationship, $f(u)$ represents its density and $E(u|x) = 0$ by definition.
- For a sample of n observations $\{y_i, x_i\}$ drawn at random from a population, $y_i = x_i'\beta + u_i$
- OLS-estimation procedures may be applied.
- Known as the **Linear Probability Model** (LPM)


y, E(y|x)
E(y|x)=x'β
1
u | x0, y=1 = 1- x0'β
u | x0, y=0 = x0'β
0
x0
x

Problems with the LPM

- **disturbance terms are non-normal**

  $u_i = 1 - x_i'\beta$ with probability $f(u_i) = x_i'\beta$
  $u_i = -x_i'\beta$ with probability $f(u_i) = 1 - x_i'\beta$

- **disturbance terms are heteroskedastic**

$$\begin{aligned} Var(u_i) = E(u_i^2) &= (-x_i'\beta)^2 \cdot (1 - x_i'\beta) + (1 - x_i'\beta)^2 \cdot (x_i'\beta) \\ &= (x_i'\beta) \cdot (1 - x_i'\beta) \\ &= P(y_i = 1|x_i) \cdot P(y_i = 0|x_i) \end{aligned}$$

- **the conditional expectation is not bounded between zero and one**

$$E(y_i|x_i) = P(y_i = 1|x_i) = x_i'\beta,$$

  which is defined over the entire real line

Probability Models

- In general,

$$E(y_i|x_i) = P(y_i = 1|x_i) = F(x_i, \beta)$$

- For the Linear Probability Model,

$$F(x_i, \beta) = x_i'\beta$$

To solve the probability problem, constrain the outcome $F(x_i, \beta)$ to the interval $[0, 1]$

Which alternatives do we know?

Transformations via Link Functions

- For the **Probit**

$$F(x_i, \beta) = \Phi(x_i'\beta)$$

where Φ is the cumulative distribution of the standard normal density

- For the **Logit**

$$F(x_i, \beta) = \Lambda(x_i'\beta)$$

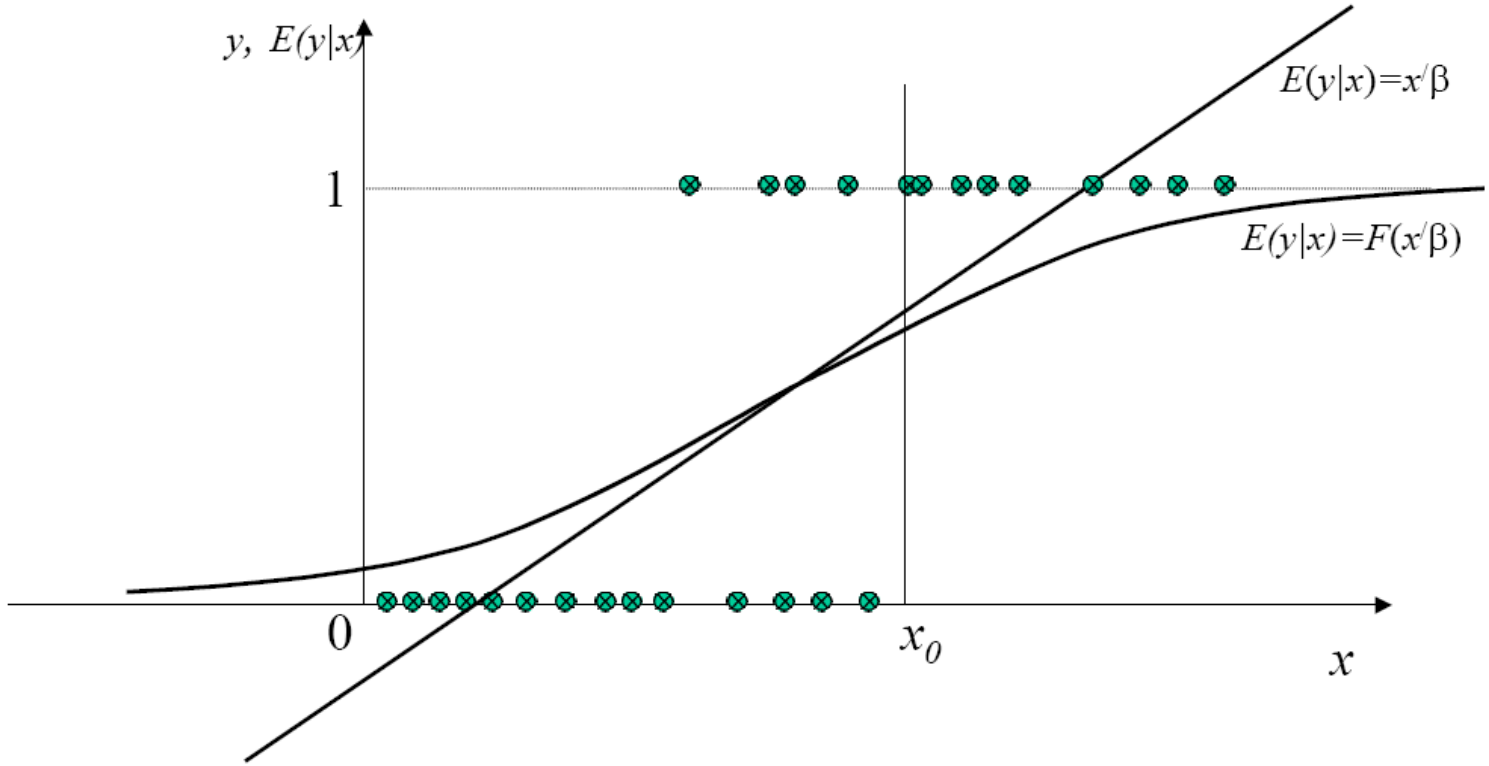
where

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + \exp(-z)}$$

represents the logistic function

- For the **Gompit** (Gompertz distribution, typical for extreme values)

$$F(x_i, \beta) = \exp\{-exp(-x_i'\beta)\}$$


y, E(y|x)
1
0
x_0
x
E(y|x)=x'β
E(y|x)=F(x'β)

The Inverse of a Link

- As already indicated, F is also often called link, or moreover 'inverse link', as its inverse - if exists - links the conditional expectation to the index
$$F^{-1}(E[Y|X = x]) = x'\beta$$
- Realize that $Y \sim B(1, p(X))$, for linear index $p(x) = p(x'\beta)$
- Consequently, $x'\beta = F^{-1}(p[x'\beta])$, s.th. $F^{-1}(\cdot)$ quantile function of $p(\cdot)$ and is indeed invertible
- for Gaussian, clearly $F^{-1} = \Phi^{-1}$, and
- for logit, $F^{-1} = \log\left(\frac{p}{1-p}\right)$, i.e. the logit function, where $\frac{p}{1-p}$ is known as 'odds ratio', frequently used concept in statistics in medicine, technometrics, etc., maybe less so in econometrics

Back to the general case – 'Latent' Variables

- Assume that there is some underlying (and unobserved) latent *propensity* variable y^* where $y^* \in (-\infty, \infty)$
- Whilst we do not observe y^* directly, we do observe a binary outcome y such that

$$y = \mathbb{1}\{y^* > 0\}$$

where $\mathbb{1}$ is termed the **indicator function**, taking the value 1 if the condition within parentheses is satisfied, and 0 otherwise

- Define (for simplicity) the latent equation in linear form:

$$y^* = x'\beta + u,$$

where u is random with symmetric density f and corresponding cumulative density function F (cf. above !?)

'Latent' Variables continued

- We now have that

$$\begin{aligned} E(y|x) &= P(y = 1|x) = P(y^* > 0|x) \\ &= P(x'\beta + u > 0) = P(u > -x'\beta) \\ &= 1 - F(-x'\beta) = F(x'\beta) \end{aligned}$$

- By specifying an appropriate distribution function for u, we can derive the Probit and logit models.
- When u is assumed normally distributed, parameters must be scaled to force the variance of u to $\sigma^2 = Var(u) = 1$. Discuss why!

$$\begin{aligned} P(y = 1|x) &= P(u > -x'\beta) = P(u/\sigma > -x'(\beta/\sigma)) \\ &= P(z > -x'(\beta/\sigma)) = \Phi(x'(\beta/\sigma)) \end{aligned}$$

From reduced to structured Models

- Suppose, $y = 1$ represents a person that works, and $y = 0$ one that doesn't.
- Consider state-specific utilities U_y^*:

$$U_{y=1}^* = x'\beta_1 + u_1, \quad U_{y=0}^* = x'\beta_0 + u_0$$

- This is an indirect utility depending on
  - individual characteristics (incl. income), and
  - alternative (mode) characteristics (incl. prices)

  i.e. characteristics summarized in x_i
- Note that the errors (unexplained heterogeneities) u_1 and u_0 may have different distributions and may be correlated.
- What we need are assumptions on its differences! (see below)

The structured Model

- Participation in the work force requires that $U^*_{y=1} > U^*_{y=0}$, such that

$$\begin{aligned} y &= \mathbb{1}\{U^*_{y=1} > U^*_{y=0}\} = \mathbb{1}\{x'\beta_1 + u_1 > x'\beta_0 + u_0\} \\ &= \mathbb{1}\{u_1 - u_0 > -x'(\beta_1 - \beta_0)\} \end{aligned}$$

- Identify the difference $\beta_1 - \beta_0$. Hence,

$$y = \mathbb{1}\{y^* > 0\}$$

where you have, see above,

$$y^* = x'(\beta_1 - \beta_0) + (u_1 - u_0) =: x'\beta + u$$

Now assume $$(u_0, u_1) \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\right]$$

Then, for example, we have a probit model

The reduced Model

- The observable model is $y_i|x_i \sim B\left(1, p(x_i, \theta)\right)$

$$\text{where} \qquad \theta = (\beta_0, \beta_1, \sigma_0^2, \sigma_1^2, \sigma_{01})$$

$$\text{and} \qquad p(x_i, \theta) = \Phi\left(\frac{x_i(\beta_1 - \beta_0)}{(\sigma_0^2 + \sigma_1^2 - 2\sigma_{01})^{1/2}}\right)$$

- So what is β ?
- What can be identified ?
- What would happen for other distribution assumptions ?
- What changes if there is no correlation between errors ($\sigma_{01} = 0$) ?
- What if we don't want to specify the (joint) distribution ?

The Maximum Likelihood approach

- Consider a sample of n independent observations $\{(y_i, x_i)\}_{i=1}^n$
- where y_i is binary; so assume $y_i = \mathbb{1}\{y_i^* > 0\}$ for $y_i^* = x_i'\beta + u_i$.
- i.e. $y_i|x_i \sim B(1, P[x_i, \beta])$, $i = 1, ..., n$
- For any vector β, the probability of observing y_i conditional on x_i is:

$$L(\beta|x_i) = \prod_{i=1}^{n} P(y_i|x_i, \beta) = \prod_{i=1}^{n} P(y_i = 0|x_i, \beta)^{1-y_i} \cdot P(y_i = 1|x_i, \beta)^{y_i}$$

- which is called the likelihood (take density for continuous Y).
- Taking logarithms you obtain

$$\ln L(\beta|x_i) = \sum_{i=1}^{n} \{(1 - y_i) \cdot \ln P(y_i = 0|x_i, \beta) + y_i \cdot \ln P(y_i = 1|x_i, \beta)\}$$

- For the Probit model this is

$$P(y_i = 1|x_i, \beta) = \Phi(x_i'\beta),$$
$$P(y_i = 0|x_i, \beta) = 1 - \Phi(x_i'\beta)$$

giving a log-likelihood of the form

$$\ln L(\beta|x_i) = \sum_{i=1}^{n}\{(1 - y_i) \cdot \ln(1 - \Phi(x_i'\beta)) + y_i \cdot \ln \Phi(x_i'\beta)\}$$

- For the logit model it is

$$P(y_i = 1|x_i, \beta) = \Lambda(x_i'\beta) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}$$
$$P(y_i = 0|x_i, \beta) = 1 - \Lambda(x_i'\beta) = \frac{1}{1 + \exp(x_i'\beta)}$$

to give a log-likelihood

$$\ln L(\beta|x_i) = \sum_{i=1}^{n}\{(1 - y_i) \cdot \ln(1 - \Lambda(x_i'\beta)) + y_i \cdot \ln \Lambda(x_i'\beta)\}$$

First Order Conditions

- Parameters which maximise the general log likelihood require that

$$S(\beta) = \frac{\partial \ln L(\beta|x_i)}{\partial \beta} = \sum_{i=1}^{n} x_i \frac{f(x_i'\beta)}{F(x_i'\beta)(1 - F(x_i'\beta))}[y_i - F(x_i'\beta)] = 0$$

- For the Probit,

$$S(\beta) = \sum_{i=1}^{n} \frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta) \cdot (1 - \Phi(x_i'\beta))} \cdot \phi(x_i'\beta) \cdot x_i$$

- For the logit,

$$S(\beta) = \sum_{i=1}^{n} \left[y_i - \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \right] \cdot x_i$$

- Solution to ML is obtained by finding parameters for which $S(\beta) = 0$

Asymptotic Behaviour

- For the LPM, estimated standard errors are easily derived and evaluated.
- But remember, LPM is heteroskedastic.

For the Probit and logit models,

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \overset{\text{asym.}}{\sim} N(0, I(\beta)^{-1}), \tag{1}$$

where $I(\beta)$ is the Fisher-Information $E[(\frac{\partial}{\partial\beta} logL(\beta|X))^2|\beta]$

- Computer software for ML estimation approximates the variance-covariance matrix $V(\hat{\beta})$ directly.
- Hence, statistical inference and hypothesis testing can be carried out using standard inferential techniques.

The Covariance Matrix

The Fisher Information is (with one observation) is

$$\begin{aligned}
(\textit{generally}) \qquad & E_\theta \left[\frac{\partial \log f(y_i|x_i;\beta)}{\partial \beta} \frac{\partial \log f(y_i|x_i;\beta)}{\partial \beta'} | x_i \right] \\
(\textit{here}) \qquad & = E_\theta \left[x_i x_i' \frac{f^2(y_i|x_i;\beta)}{F^2(x_i'\beta)[1-F(x_i'\beta)]^2} [y_i - F(x_i'\beta)]^2 | x_i \right] \\
& = x_i x_i' \frac{f^2(y_i|x_i;\beta)}{F(x_i'\beta)[1-F(x_i'\beta)]}
\end{aligned}$$

(comment on notation) and thus the estimated covariance matrix is

$$\widehat{Var}(\hat{\beta}) = \left(\sum_{i=1}^{n} x_i x_i' \frac{f^2(y_i|x_i;\hat{\beta})}{F(x_i'\hat{\beta})[1-F(x_i'\hat{\beta})]} \right)^{-1} \tag{2}$$

For the logit: $\left(\sum_{i=1}^{n} x_i x_i' F(x_i'\hat{\beta})[1-F(x_i'\hat{\beta})] \right)^{-1}$

cf. remark about weights in GLS

assumptions, consistency and asymptotic properties of ML estimation

Proof at the green board if time

Recall that $(u/v)' = (u'v - v'u)/v^2$ and consider the score function

$$
\begin{aligned}
E\left[\frac{\delta lnf}{\delta\theta}\right] &= \int \frac{\delta f(x)}{\delta\theta} f^{-1}(x) dF_x = \frac{\delta}{\delta\theta}\int dF_x = 0 \\
E\left[\frac{\delta^2 lnf}{\delta\theta^2}\right] &= E\left[\frac{f''f - (f')^2}{f^2}\right] \\
&= \int f'' dx - E\left[\frac{(f')^2}{f^2}\right] = 0 - E\left[\left(\frac{\delta lnf}{\delta\theta}\right)^2\right]
\end{aligned}
$$

extension to multivariate case is straight forward, just notationally more challenging

Similarities to GLS

- If we take $\hat{u}_i = y_i - P(x_i'\hat{\beta})$ as residual
- then the FOC are an orthogonality condition between the regressor x and the residuals
- but, with weights (see last slight) which are inverse proportional to the conditional variance
- In other words, this is like a GLS
- and (also) therefore, we speak of Generalized Linear Models

... might want to recall proof(s) for statistical behavior of ML estimators

More general: Binary Choice Models

- Let's keep concentrating on the following Binary Choice models:

$$\begin{aligned} \textbf{LPM} \quad & P(y_i = 1|x_i, \beta) = x_i'\beta \\ \textbf{Probit} \quad & P(y_i = 1|x_i, \beta) = \Phi(x_i'\beta) \\ \textbf{Logit} \quad & P(y_i = 1|x_i, \beta) = \Lambda(x_i'\beta). \end{aligned}$$

- If β_j is positive (negative), then $P(y_i = 1|x_i, \beta) = F(x_i'\beta)$ will increase (decrease) with an increase in x_j.

Marginal Effects

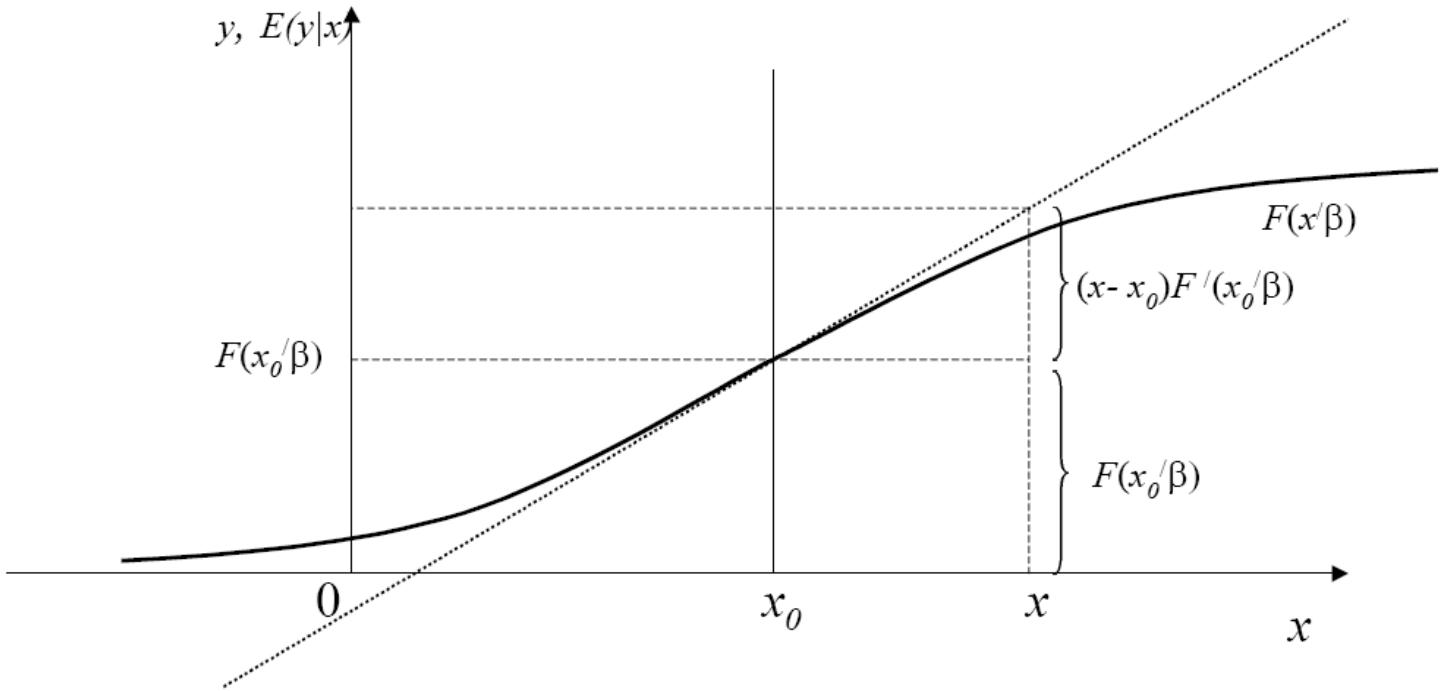
LPM
$$\frac{\partial P(y_i = 1|x_i, \beta)}{\partial x_{ij}} = \beta_j$$

Probit
$$\frac{\partial P(y_i = 1|x_i, \beta)}{\partial x_{ij}} = \phi(x_i'\beta) \cdot \beta_j$$

Logit
$$\frac{\partial P(y_i = 1|x_i, \beta)}{\partial x_{ij}} = \frac{\exp(x_i'\beta)}{(1 + \exp(x_i'\beta))^2} \cdot \beta_j$$

Implications

- Slope estimates are **not** directly comparable.
- E. g. variance of disturbances in logit model and the Probit model are different.
- Hence the *parameters* are also scaled differently.


y, E(y|x)
F(x'β)
(x- x0)F '(x0'β)
F(x0'β)
F(x0'β)
0
x0
x
x

Differences in Marginal Effects due to the Link

Notice also:

- The marginal effects in the LPM are constant (i. e. independent of the data).
- The marginal effects in the Probit and logit models depend on x_i.

A popular transformation:

- $\hat{\beta}_{LPM} \approx 0.25\hat{\beta}_L$ for the slopes, and
- $\hat{\beta}_{LPM} \approx 0.25\hat{\beta}_L + 0.5$ for the intercept.
- $\hat{\beta}_P \approx 0.625\hat{\beta}_L$.

An Empirical Example: childcare take-up estimates

	Parameter Estimates		
Variable	LPM	Probit	Logit
single woman	-0.059	-0.184	-0.310
other children aged 5+	-0.101	-0.318	-0.540
woman works	0.152	0.430	0.713
left school at 18	0.109	0.310	0.520
attended college/uni	0.160	0.458	0.757
youngest child aged 2	0.186	0.556	0.928
youngest child aged 3-4	0.309	0.882	1.458
receives maintenance	0.089	0.264	0.432
constant	0.153	-0.995	-1.645

- Data source: 1991/92 General Household Survey, from which a random sample of $n = 1288$ women was taken, which are responsible for at least one child in pre-school-age.

An Empirical Example: childcare take-up estimates

- The dependent variable is 1, if the woman pays for childcare, else 0.
- The reference in all cases is a married woman who doesn't work, has left school at 16, has one child aged less than two and who receives no maintenance.

For the reference household, all explanatory variables take a value of 0, which leads to probability estimates in each model of:

$$\textbf{LPM} \qquad \hat{P}(y_i = 1|x_i) = x_i'\hat{\beta} = 0.153,$$

$$\textbf{Probit} \qquad \hat{P}(y_i = 1|x_i) = \Phi(x_i'\hat{\beta}) = \Phi(-0.995) = 0.161,$$

$$\textbf{Logit} \qquad \hat{P}(y_i = 1|x_i) = \Lambda(x_i'\hat{\beta}) = \frac{\exp(-1.645)}{1 + \exp(-1.645)} = 0.162.$$

An Empirical Example: childcare take-up estimates

How, for example, does the probability change for women who attend university?

LPM
$$\hat{P}(y_i = 1|x_i) = x_i'\hat{\beta} = 0.153 + 0.160 = 0.313,$$

Probit
$$\begin{aligned}\hat{P}(y_i = 1|x_i) &= \Phi(x_i'\hat{\beta}) \\ &= \Phi(-0.995 + 0.458) = \Phi(-0.537) = 0.296,\end{aligned}$$

Logit
$$\begin{aligned}\hat{P}(y_i = 1|x_i) &= \Lambda(x_i'\hat{\beta}) \\ &= \frac{\exp(-1.645 + 0.757)}{1 + \exp(-1.645 + 0.757)} = 0.291.\end{aligned}$$

Goodness–of–fit

- Let L_{UR} represent likelihoods for the full model.
- Let L_R represent likelihoods for a restricted model estimated on an intercept alone.
- Then the formulation for two proposed measures are as follows:

$$\textbf{Cragg–Uhler}\text{–pseudo–}R^2 = \frac{L_{UR}^{2/n} - L_R^{2/n}}{1 - L_R^{2/n}},$$

$$\textbf{McFadden}\text{ –pseudo–}R^2 = 1 - \frac{\ln L_{UR}}{\ln L_R}.$$

Remember: the classical and adjusted R^2 can't be used. (Why?)

Goodness–of–fit

An alternative outcome-based measure: the *proportion of correct predictions*.

- For $\hat{P}_i = \hat{P}(y_i = 1|x_i)$, eg. $\Phi(x_i'\hat{\beta})$ (Probit), let $\hat{y}_i = \mathbb{1}\{\hat{P}_i > 0.5\}$
- Define the proportion of correct predictions as

$$\mathrm{P} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i = \hat{y}_i\}.$$

- In many statistic computer programs, you can see tables of predicted and observes binary values:

observed	predicted 0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

- This measure should be avoided. It doesn't make sense if x are weak predictors and one of the answers dominates in the sample.

Testing the Overall Significance of the Regression

- Let L_{UR} represent likelihoods for the full model.
- Let L_R represent likelihoods for a restricted model.
- r represents the number of restrictions imposed.

Then:

$$-2\ln(L_R/L_{UR}) = 2(\ln L_{UR} - \ln L_R) \sim \chi^2_{restr.}.$$

For example:

$$H_0: \quad \beta_2 = \beta_3 = \cdots = \beta_J = 0$$
$$H_A: \quad \text{at least one} \quad \beta_j \neq 0, \quad j = 2, \ldots, J,$$

where β_1 is the intercept

Count Data Models

- Poisson Regression & Pseudo-Poisson
- Negative Binomial Regression
- Zero inflation models

Examples

- number of visits to a medical doctor in a month
- number of claims in a year, e.g. for a car insurance
- number of days of strike in a year
- number of "failures" in a year (hedge funds, machines, ...)
- number of patents of an individual firm
- etc ...

Poisson Distribution and Regression

Let us denote the Poisson distribution by $Poiss(\lambda)$,

$$P(Y = y) = \exp(-\lambda)\frac{\lambda^y}{y!}, \qquad \lambda \geq 0$$

with mean $E(Y) = \lambda$ and variance $Var(Y) = \lambda$

The Poisson regression model: $P(Y = y|X = x) \sim Poiss(\lambda(x;\beta))$, i.e. face different parameter values according to individual characteristics

Generally one works with parametrization $0 \leq \lambda(x;\beta) = \exp(x'\beta)$ to ensure the positivity, and to get results easy to interpret

$$P(Y = y|x) = \exp(-\exp(x'\beta))\frac{\exp(x'\beta)^y}{y!}, \qquad \lambda \geq 0$$

with mean $E(Y) = \exp(x'\beta)$ and variance $Var(Y) = \exp(x'\beta)$

Max.Likelihood Poisson Estimation

$$l(\beta; y) := \ln L(\beta; y) = \ln \prod_{i=1}^{n} \exp[-\exp(x_i'\beta)] \frac{\exp(y_i x_i'\beta)}{y_i!}$$

$$= \sum_{i=1}^{n} y_i x_i'\beta - \exp(x_i'\beta) - \ln(y_i!)$$

The First Order Condition(s) [FOC]

$$\frac{\partial l(\beta; y)}{\partial \beta} = 0 \quad \Leftrightarrow \quad \sum_i \left(y_i - \exp(x_i'\beta)\right) x_i = 0$$

$$\Leftrightarrow \quad \sum_i \tilde{u}_i x_i = 0 \tag{3}$$

with 'residual' $\tilde{u}_i = y_i - \exp(x_i'\beta)$: orthogonality (moment) condition !

$$\frac{\partial^2 l(\beta; y)}{\partial \beta \partial \beta'} = -\sum_i x_i x_i' \exp(x_i'\beta)$$

Relationship to Binary Model Case

The count data are often analyzed in a first step by simply distinguishing 0-1. E.g. in case of insurance claims you may think of

- the existence of a claim:

$$a\ 0-1\ variable\ Y^{\star}\ \begin{cases} no & claim \\ at & least \quad one \end{cases}$$

- the number of claims, if at least one claim: quantitative variable

The models considered in the following are (also) valid if the observed number of claims is pretty small but the probabilities of 0, 1 or 2 claims are significant

Extreme case of Binary Model Case

- Note that in this situation you may have only "a few" zeros
- More specifically:
- $Y^\star = \mathbb{1}\{Y > 0\}$ with Y being $Poiss(\lambda)$, $\lambda(x;\beta) = exp(x'\beta)$

$$P(Y^\star = 1|x) = 1 - P(Y = 0|x) = 1 - exp[-exp(x'\beta)]$$

- This is our Gompit regression model from last chapter:

$$\ln L(y^\star;\beta) = \sum_{i=1}^{n} y_i^\star \ln\{1 - exp[-exp(x_i'\beta)]\} - (1 - y_i^\star)exp(x_i'\beta)$$

Remark: On last slide we saw relation to GLS, therefore the name *Generalized Linear Model (GLM)*

More motivation of modelling

- Why $\lambda(x;\beta) = exp(x'\beta)$ should be a good idea?
- Recall that regression with large output are often modeled by log-log-linear models like wage, expenditure, income, production function

$$Y = \alpha L^{\beta_2} C^{\beta_3} e, \qquad E[e|L, C] = 1$$

- Then, working with particularly small values, one has the tendency to apply the inverse transform, i.e. taking the exponential.
- but this explains only the exponential transform (which is related to the quite popular Laplace transformation and moment generating functions), not the linear index
- There is certainly a trade-off between the transform and the index modelling

The Attractiveness of Poisson Pseudo ML

- One often works with

$$\ln Y = \beta_1 + \beta_2 \ln L + \beta_3 \ln C + u, \;\; \beta_1 = \ln \alpha, \;\; E[u|L, C] = 0$$

- But if e.g. $Var[e|L, C] \neq const$ then $E[u|L, C] \neq const \Rightarrow$ inconsistency $\Rightarrow$ this transformation requires independence in the original equation
- Might work with $E[Y|L, C] = exp(\beta_1 + \beta_2 \ln L + \beta_3 \ln C)$ instead, leading us to the so-called *Poisson Pseudo ML*
- Recalling the FOC (3): these give valid moment conditions which are even efficient iff $E[Y|L, C] \propto V[Y|L, C]$

Example:

Consider the production function from above and imagine e is log-normal with $E[e|L, C] = 1$ and $V[e|L, C] = g(L, C)$. Then it is easy to check that $E[u|L, C] = E[\ln e|L, C] = -\frac{1}{2} \ln(1 + g(L, C))$ which implies endogeneity unless $g(L, C)$ is constant.

Limitation of the Poisson regression model

- Recall that despite the modelling of λ, we still have

$$\lambda(x;\beta) \;=\; E[Y|x] \;=\; V[Y|x] \left(\;=\; \exp(x\beta)\right)$$

- whereas in practice we generally observe conditional overdispersion:

$$E[Y|x] \;<\; V[Y|x]$$

- Even worse, mean and variance are not just in a deterministic relationship, e.g. up to an appropriate multiplicative or additive constant, they are equal.
- I.e., the Poisson distribution is a one-parameter distribution.
- switch therefore to a tow-parameter counting data distribution
- A different problem is that in practice we often find zero inflation

Basic Parameterizations of The negative Binomial

- Let us consider only response Y for a moment
- Then, the first version considers the number of the trial at which the v'th success occurs, typically modeled as

$$P(Y_1 = y|p, v) = \left(\begin{array}{c} y-1 \\ v-1 \end{array} \right) p^v (1-p)^{y-v}$$

for integer $y \geq v$, $0 < p < 1$ and v a positive integer. p corresponds to the chance of a success.

- The second version counts the number of failures before the v'th success, which is basically the same but with $Y_2 = Y_1 - v$,

$$P(Y_2 = y|p, v) = \left(\begin{array}{c} v+y-1 \\ y \end{array} \right) p^v (1-p)^{y}$$

for integer y, $0 < p < 1$ and v a positive integer

- If Y_1 is a negative binomial random variable according to the first definition, then also Y_2 is a negative binomial (2nd definition)

Standardised version on the second parametrisation

- One advantage to the second version is that the range of y is all non-negative integers
- Also, the definition can be more easily extended to all positive real values of v since there is no factor of v in the bottom of the binomial coefficient
- Rather than parameterizing the negative binomial in terms of v and p, we could consider **mean** μ and **variance** σ^2 in terms of v and p
- In this case

$$v = \frac{\mu^2}{\sigma^2 - \mu} \quad \text{and} \quad p = \frac{v}{v + \mu} \tag{4}$$

- A little bit of algebra shows that

$$\mu = v\frac{1-p}{p} \quad \text{and} \quad \sigma^2 = \mu + \frac{1}{v}\mu^2 \tag{5}$$

Consequences and Comments

- Thus the variance is always larger than the mean,
- and we have at hand two parameters to determine.
- Note that if X_j are NB(v_j, p), then $\sum_j X_j \sim$NB$(\sum_j v_j, p)$
- Factor $1/v$ is a sort of "clumping" parameter. As $v \to \infty$, the negative binomial converges in distribution to the Poisson.
  This fact is suggested by the variance approaching the mean then.
- Viewing μ and σ as primary, we ignore the combinatorial motivation for defining the negative binomial and instead view it simply as a model for count data.

Extension to the Regression case

- First, as we want to allow for any real $v \geq 0$, we write

$$\begin{pmatrix} v+y-1 \\ y \end{pmatrix} = \frac{\Gamma(v+y)}{\Gamma(y+1)\Gamma(v)}$$

recall that for positive integers $m > k$

$$\Gamma(m+1) = m! \quad \text{and} \quad \begin{pmatrix} m \\ k \end{pmatrix} = \frac{m!}{k!(m-k)!}$$

- With (4) and (5) we get

$$P(Y=y|X=x) = \frac{\Gamma(y+v)}{\Gamma(y+1)\Gamma(v)} \left(\frac{v}{v+\mu}\right)^v \left(\frac{\mu}{v+\mu}\right)^y \qquad (6)$$

with $E[Y|X=x] = \mu(x)$, $Var[Y|X=x] = \mu(x) + \mu^2(x)v^{-1}$

- Often in econometrics you find parametrization in terms of $a = v^{-1}$, *negative Binomial of type I*, cf eqn (5)
- Alternatively, one finds $a = \mu/v$, called *exposure measure*

Motivating the "exposure" measure

- Recall the linear regression model

$$y = x_1'\beta + x_2'\theta + \epsilon = x_1'\beta + u + \epsilon = x_1'\beta + \tilde{u}$$

  where u could be considered as random effect.
- If $E[\tilde{u}|x_1] = 0$ we get consistent estimates from reduced form
- However, the impact of ϵ and u cannot be separated.
  It is therefore impossible to identify them without further assumptions
- In nonlinear models this is different: Imagine (with $\mathcal{P} =$ Poisson)

$$y|(x_1, x_2) \sim \mathcal{P}[\exp(x_1'\beta + x_2'\theta)] \Rightarrow y|x_1 \sim \mathcal{P}[\exp(x_1'\beta + u) = e \cdot \exp(x_1'\beta)]$$

- So we have two types of uncertainty:
  1. the basic one (ϵ in the linear model), ie. drawing from Poisson
  2. the unobserved heterogeneity e ($= exp(u)$) or u respectively, cf (7)

From unobserved heterogeneity to overdispersion

Assume $y_i \sim Poiss(\lambda_i)$, where $\lambda_i = \exp(x_i'\beta + u_i) = e_i \cdot \exp(x_i'\beta)$

with $E[e_i] = 1$, $Var[e_i] = a$, being independent from x

then we get for the conditional distribution of y given x

$$\begin{aligned} E[y|x] &= E\left[E(y|x,e)|x\right] = E\left[e \cdot \exp(x'\beta)|x\right] \\ &= \exp(x'\beta) \end{aligned}$$

and

$$\begin{aligned} Var[y|x] &= E\left[Var[y|x,e]|x\right] + Var\left[E[y|x,e]|x\right] \\ &= E\left[e \cdot \exp(x'\beta)\right] + Var\left[e \cdot \exp(x'\beta)\right] \\ &= \exp(x'\beta) + a \cdot \exp(2x'\beta) \end{aligned}$$

This holds without further specification of distribution of e.

For moment estimator fine, for ML we need entire distribution.

From Poisson to Negative Binomial

- NB (6) has same first two moments. The entire distribution comes via distribution of e_i:
- With $e_i \sim Gamma(v_1; v_2)$, $E[e_i] = v_1 v_2 = 1$, so use $v := v_1 = v_2^{-1} > 0$, and thus $Var[e_i] = v_1 v_2^2 = v^{-1}$
- I.e. for $Gamma(v; v^{-1})$ the density is

$$f(e) = v^v e^{v-1} \exp(-v \cdot e)/\Gamma(v) \ , \qquad \Gamma(v) = \int_0^\infty \exp(-e) e^{v-1} de$$

- With $y|(x, e)$ Poisson, and e gamma, the conditional distribution of $y|x$ is obtained by integrating over e

$$\int f(y, e|x) de = \int Poiss(y|x, e) f(e) de =$$

$$\frac{\Gamma(y+v)}{\Gamma(y+1)\Gamma(v)} \frac{[\exp(x'\beta)/v]^y}{[1+\exp(x'\beta)/v]^{y+v}} = NB(\lambda = \exp(x'\beta), v)$$

Generalised Linear Regression with NB (type I)

- Assuming /defining a multiplicative conditional mean by

$$\tilde{\mu}(x, e) = E[Y|x, e] = e \cdot \exp(x'\beta) = \exp(x'\beta + u) \tag{7}$$

with a "measure of exposure" e
(like a random effect to model additional heterogeneity)

with $E[e] = 1$, $Var[e] = a$, independent from x

for unexplained heterogeneity (recall NB-I, $a = 1/v$), see also below

- one can use (log-)likelihood to get β and a (see probability formula)

$$\sum_i \left\{ \sum_{r=1}^{y_i - 1} \log(1 + ar) \right\} - \log(y_i!) + y_i \log(\mu_i) - (y_i + a^{-1}) \log(1 + a\mu_i)$$

with $\mu_i = \exp(x_i'\beta)$ and using $\sum_i \log m(i) = \log \prod_i m(i)$

Maximum Likelihood Estimation

- Maximizing, i.e. taking first derivatives and setting them to zero
- gives first order conditions (still $\mu_i = \exp(x_i'\beta)$), ie.

$$\frac{\partial l(\beta, a)}{\partial \beta} = \sum_i \frac{(y_i - \mu_i)x_i}{1 + a\mu_i}$$

$$\frac{\partial l(\beta, a)}{\partial a} = \sum_i \left\{ \sum_{r=1}^{y_i - 1} \frac{r}{1 + a \cdot r} \right\} + a^{-2} \log(1 + a\mu_i) - \frac{(y_i + a^{-1})\mu_i}{1 + a\mu_i}$$

being equal to zero, etc.

- For a one could also take the moment estimator, solving

$$\sum_i \frac{(y_i - \mu_i)^2}{\mu_i(1 + a\mu_i)} = n - \dim(\beta)$$

Example: The Bonus-Malus Scheme

Consider individual claim histories for car insurance:

- $y_{i1}, y_{i2}, y_{i3}, \cdots$, where y_{it} is number of claims for individual i in year t
- imagine we have panel data with individual length
- may assume that $y_{it}|(x_{it}, e_i) \sim Poiss(e_i \cdot \exp(x_{it}'\beta))$
- where e_i is a time independent random effect reflecting the uncertainty due to unobserved individual characteristics
- Questions:
  How to "price" the observed (identifiable) heterogeneity?
  How to "price" the unobserved heterogeneity?
- Answer:
  Consider the learning on the unobserved heterogeneity by means of the *posteriori distribution*
  $$f(e_{i,T}|y_{i1}, \ldots, y_{iT}, x_{i1}, \ldots, x_{iT}) = \frac{f(y_{i1},\ldots,y_{iT}|e_{i,T},x_{i1},\ldots,x_{iT}) f(e_{i,T}|x_{i1},\ldots,x_{iT})}{f(y_{i1},\ldots,y_{iT}|x_{i1},\ldots,x_{iT})}$$

Incorporating the learning effect

The posteriori distribution for individual i at time T is

$$e_{i,T} \sim Gamma\left(v + \sum_{t=1}^{T} y_{i,t} \; ; \; \left\{v + \sum_{t=1}^{T} \exp(x'_{i,t}\beta)\right\}^{-1}\right)$$

If the expected CHF amount per accident is C_{T+1} for year $T+1$, then the *pure premium* is

$$\begin{aligned} P_{i,T} &= C_{T+1} \cdot E\left[Y_{i,T+1} | y_{it}, (t = 1, \ldots, T), x_{it}, (t = 1, \ldots, T+1)\right] \\ &= C_{T+1} \exp(x'_{i,T+1}\beta) E\left[e_i | y_{it}, x_{it}, (t = 1, \ldots, T)\right] \\ &= C_{T+1} \exp(x'_{i,T+1}\beta) \frac{v + \sum_{t=1}^{T} y_{it}}{v + \sum_{t=1}^{T} \exp(x'_{it}\beta)}, \end{aligned}$$

Bonus – Malus

The change of pure premium can therefore be decomposed as

$$\frac{P_{i,T+1}}{P_{i,T}} = \frac{C_{T+2}}{C_{T+1}} \frac{\exp(x'_{i,T+2}\beta)}{\exp(x'_{i,T+1}\beta)} \frac{E[e_{i,T+1}]}{E[e_{i,T}]}$$

where the

- first fraction is the effect of inflation
- second fraction stands for the change of characteristics
- third fraction is the effect of learning on e_i

i.e. a clear distinction between the components of the pure premium

Remark:
have still to add the effect of variance to account for risk premia in pricing

Summarizing the main results

Let us now consider time invariant individual covariates $x_{it} = x_i$, $\forall t$

$$E[e_{i,T}] = \frac{v + \sum_{t=1}^{T} y_{it}}{v + T \exp(x_i'\beta)} = \frac{v/T + \frac{1}{T}\sum_{t=1}^{T} y_{it}}{v/T + \exp(x_i'\beta)}$$

- At time 0, no claim record, $P_{i,0} = C_1 \exp(x_i'\beta)$
  say, the insurance between individuals is a flat rate
- When time increases, we learn more and more on the true individual risk (on individual i or its e_i)
- At time $T = \infty$, v constant, i.e. $v/T \approx 0$

$$\begin{aligned} E[e_{i,T}] &\approx E[Y_{it}]/\exp(x_i'\beta) = E[e_i] \\ P_{iT} &\approx C_T \exp(x_i'\beta) E[e_i] \end{aligned}$$

Zero Inflation: The general idea

- Recall the slide on the limitations of Poisson
- Assume now we face count data with a serious number of zeros
- Make sure where this is due to, e.g. censoring, truncation, certain sequential decision; if it is we have to model them accordingly
- A straight forward modelling approach is then to consider adequate mixing distributions
- In the case of 'zero inflation' such a mixing says that a certain percentage [which can again be modeled as a conditional probability based on observed covariates] of people has chosen for none (zero), and the other part of the population follows e.g. a Poisson or negative binomial
- As the principle is always the same but notation complicates for more complex distributions, we stick here to a zero-inflated Poisson

Zero-inflated Poisson (ZIP) / Zero-altered Poisson (ZAP)

Let Y be ZIP with

$$P(Y=y|x,z) = \begin{cases} p(z) + (1-p(z))\exp\{-\lambda(x)\} & \text{for } y=0 \\ (1-p(z))\exp\{-\lambda(x)\}\frac{\lambda(x)^y}{y!} & \text{for } y>0 \end{cases}$$

where x and z can overlap (no exclusion restriction), and
$p(z)$ the probability of 'not participating' (first decision if sequential)

$$E[Y] = (1-p)\lambda = \mu \qquad Var[Y] = \mu + \frac{p}{1-p}\mu^2$$

- Often one models $\lambda = \exp\{x'\beta\}$ and $\frac{p}{1-p} = \exp\{z'\gamma\}$
- Sometimes, if $x = z$ one sets $\beta = \tau\gamma$ where only β and τ are to be estimated then.

Use ML methods for estimation

A Re-parametrization

For propagation models the following is used:

$$P(Y=y|x,z) = \begin{cases} \pi(z) & \text{for } y=0 \\ \frac{1-\pi(z)}{(1-\exp\{-\lambda(x)\})} \exp\{-\lambda(x)\}\frac{\lambda(x)^y}{y!} & \text{for } y>0 \end{cases}$$

where it's assumed that participation always entails a positive output y

- Again, $\pi(z)$ is the probability to not participate.
- Expectation and Variance can be obtained from above by plugging in $\pi = p + (1-p)/e^{\lambda}$
- Typical modelling proposals for λ and π are $\lambda = \exp\{x'\beta\}$ and $\ln\{-\ln\pi\} = z'\gamma$
- and if x and z coincide, like before $\beta = \tau\gamma$

ZIP can be extended, e.g. to zero inflated negative binomial models

Multiple Discrete Choice Models

- Ordered Probit and Ordered Logit
- The Multinomial and the Conditional Logit
- Sequential Discrete Choice models
- The Bivariate Probit model
- Nested Logit Models

Multiple responses revisited

- As already seen, sometimes a simple binary choice model is inappropriate:
  - eg. model of labour market status
  - degree of satisfaction
  - choice between different items
- Each of these examples involves more than two possible outcomes.
- but are clearly not count data
- If they are still ordered, a possible model specification are the **Ordered Probit or Logit** models:
  - appropriate when discrete outcomes have a **natural (ordinal) ranking**
  - major advantage: the resulting model is relatively easy to estimate
  - down-side: the behavioural model may be considered too restrictive
  - the large number of unknown parameter complicates (numerical) identification and interpretation

From the Data to the Model

- Consider an independent sample of data $\{y_i, x_i\}$ of size n.
- Let y_i have M possible outcomes $y_i = m$ for $m = 1, ..., M$ and
- natural ordering (e.g. $m + 1$ is in some sense better than m).
- Consider a latent variable y_i^* where

$$y_i^* = x_i'\beta + u_i \quad \text{for} \quad i = 1, ..., n$$

- Define the following observability criterion:

$$y_i = m \quad \text{if } \alpha_{m-1} \leq y_i^* \leq \alpha_m \text{ for } m = 1, ..., M,$$
$$\alpha_0 < \alpha_1 < \alpha_2 < ... < \alpha_M,$$
$$\alpha_0 = -\infty \quad \text{and} \quad \alpha_M = \infty$$

Toward an ordered Probit/Logit Model

- The conditional probability of observing $y_i = m$ is

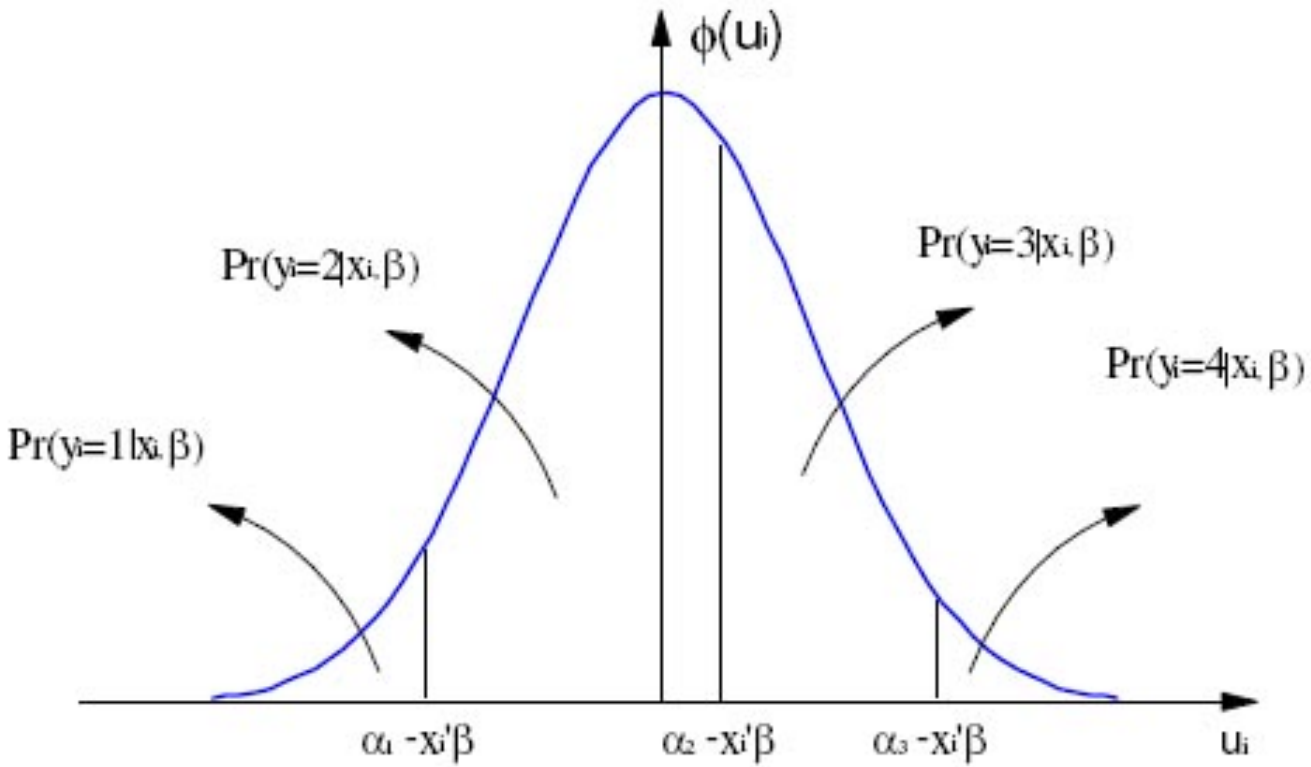
$$\begin{aligned} P(y_i = m|x_i) &= P(\alpha_{m-1} \leq y_i^* \leq \alpha_m|x_i) \\ &= P(\alpha_{m-1} \leq x_i'\beta + u_i \leq \alpha_m) \end{aligned}$$

- Rearranging gives for $m = 1, ..., M$

$$\begin{aligned} P(y_i = m|x_i) &= P(\alpha_{m-1} - x_i'\beta \leq u_i \leq \alpha_m - x_i'\beta) \\ &= P(u_i \leq \alpha_m - x_i'\beta) - P(u_i \leq \alpha_{m-1} - x_i'\beta) \end{aligned}$$

- Need a distribution for u_i
  - u_i std normal gives the Ordered Probit
  - u_i logistic gives the Ordered Logit

Ordered Probit: graphical representation



- eg. let $u_i \sim N(0,1)$. Then

$$P(y_i = m|x_i) = \Phi(\alpha_m - x_i'\beta) - \Phi(\alpha_{m-1} - x_i'\beta)$$

Methods of Estimation I

- Estimate this non-linear model by maximum likelihood:
  - let $z_{im} = \mathbb{1}(y_i = m)$, for $m = 1, ..., M$,
  - then the ith likelihood contribution is

$$L_i = \prod_{m=1}^{M} P(y_i = m|x_i)^{z_{im}}$$
$$= \prod_{m=1}^{M} [\Phi(\alpha_m - x_i'\beta) - \Phi(\alpha_{m-1} - x_i'\beta)]^{z_{im}}$$

- The full likelihood function becomes

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{m=1}^{M} [\Phi(\alpha_m - x_i'\beta) - \Phi(\alpha_{m-1} - x_i'\beta)]^{z_{im}}$$

Methods of Estimation II

- Taking logs,

$$l = \sum_{i=1}^{n} \sum_{m=1}^{M} z_{im} \ln[\Phi(\alpha_m - x_i'\beta) - \Phi(\alpha_{m-1} - x_i'\beta)]$$

- For ML estimates, solve

$$\frac{\partial l}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial l}{\partial \beta} = 0.$$

- discuss conditions
- discuss consequences

Unordered Discrete Responses: Multinomial Model

- Simplest model for **unordered** discrete choices
- where **covariates do not vary with** m.
- Example: public transport choice.
- Consider M discrete alternatives

$$P_{mi} = P(y_i = m) \quad \text{for} \quad m = 1, ..., M.$$

with $\sum_m P_{mi} = 1$ for all i.

- Thinking again of latent variables, here utilities;

$$U_{im}^* = x_i'\beta_m + u_{im}$$

we get

$$P_{mi} = P\left(U_{im}^* > U_{ij}^*, \ \forall j \neq m\right)$$

General Multinomial Probability Model

- Let us derive the probability model generally:
- For the Multinomial Model , $m = 1, ..., M-1$ let us consider

$$\frac{P_m}{P_m + P_M} =: F(x'\beta_m)$$

(see next slides),
for a benchmark probability P_M, i.e. M is the benchmark choice

- Notice that this implies

$$\frac{P_m}{P_M} = \frac{F(x'\beta_m)}{1 - F(x'\beta_m)} =: \lambda(x'\beta_m) \tag{8}$$

which reminds us of the logit distribution.

- Will see that $F(\cdot)$ is cdf. indeed

The Multinomial Probability continued

- Since $P_m \in (0,1)$, we have that

$$\frac{P_m}{P_m + P_M} \to 0 \quad \text{as} \quad P_m \to 0,$$
$$\frac{P_m}{P_m + P_M} \to 1 \quad \text{as} \quad P_m \to 1.$$

- Can even imagine $F(.)$ as a monotone increasing function,

$$F(u) \to 0 \quad \text{as} \quad u \to -\infty,$$
$$F(u) \to 1 \quad \text{as} \quad u \to \infty.$$

- Note that, since $\sum_{m=1}^{M} P_m = 1$, we have

$$\sum_{j=1}^{M-1} \frac{P_j}{P_M} = \frac{1 - P_M}{P_M} = \frac{1}{P_M} - 1. \tag{9}$$

The Multinomial Logit

- Hence, for all $m = 1, ..., M-1$, combining equations (9), (8)

$$P_M = \left[1 + \sum_{j=1}^{M-1} \frac{P_j}{P_M}\right]^{-1} = \left[1 + \sum_{j=1}^{M-1} \lambda(x'\beta_j)\right]^{-1}$$

$$\Rightarrow P_m = \frac{\lambda(x'\beta_m)}{1 + \sum_{j=1}^{M-1} \lambda(x'\beta_j)} := P(Y = m|X = x)$$

- For the Multi Logit Model one sets $\lambda(u) = \exp(u)$, c.f. binary logit
- Alternatives are possible but rarely used.
- Max.Lik.: $log \prod_{i=1}^{n} \prod_{m=1}^{M} P(y_i = m|x_i)^{\mathbb{1}\{y_i=m\}}$
- discuss interpretation ... what is the marginal impact of x_j

The Independence of Irrelevant Alternatives

- Recall the formulae for the probabilities,

$$P_m = \frac{\exp(x'\beta_m)}{1 + \sum_{j=1}^{M-1} \exp(x'\beta_j)}$$

  for all $m = 1, ..., M-1$.

- However, comparing two items

$$\frac{P_j}{P_k} = \frac{exp(x'\beta_j)}{\exp(x'\beta_k)}$$

  we notice that this ratio is independent of the probability of any other outcome.

- This is called the *assumption of independence of irrelevant alternatives*.

- Not always reasonable, e.g. in context of sequential decisions!

A Conditional Logit Model

- before discussing those, consider situation
- where now **covariates vary over** m
- Example: distance to store.
- Consider M discrete alternatives

$$P_{mi} = P(y_i = m) \quad \text{for} \quad m = 1, ..., M.$$

- Thinking again of latent variables, e.g. utilities:

$$U^*_{im} = z'_{im}\gamma + u_{im}$$

slopes γ **now fixed for identification**. Again we have

$$P_{mi} = P\left(U^*_{im} > U^*_{ij},\ \forall j \neq m\right)$$

- Have no benchmark, similar derivation leads to

$$P_{mi} = \frac{\exp(z'_{im}\gamma)}{\sum_{j=1}^{M} \exp(z'_{ij}\gamma)}$$

Mixing individual and mode effects

- Certainly, one may want to include both, individual characteristics x_i with mode-dependent slopes β_m and characteristics that change over slopes z_{im} with fixed slopes γ

$$P(y_i = m|x_i, z_{im}) = \frac{\lambda(x_i'\beta_m + z_{im}'\gamma)}{\sum_{j=1}^{M} \lambda(x_i'\beta_j + z_{ij}'\gamma)}$$

- Note that this implies the inclusion of mode specific characteristics z_m, which is the most typical case for conditional logits
- Can be extended to more levels than individual, and mode specific characteristics
- Discuss inclusion of mode specific intercepts versus mode specific z versus none

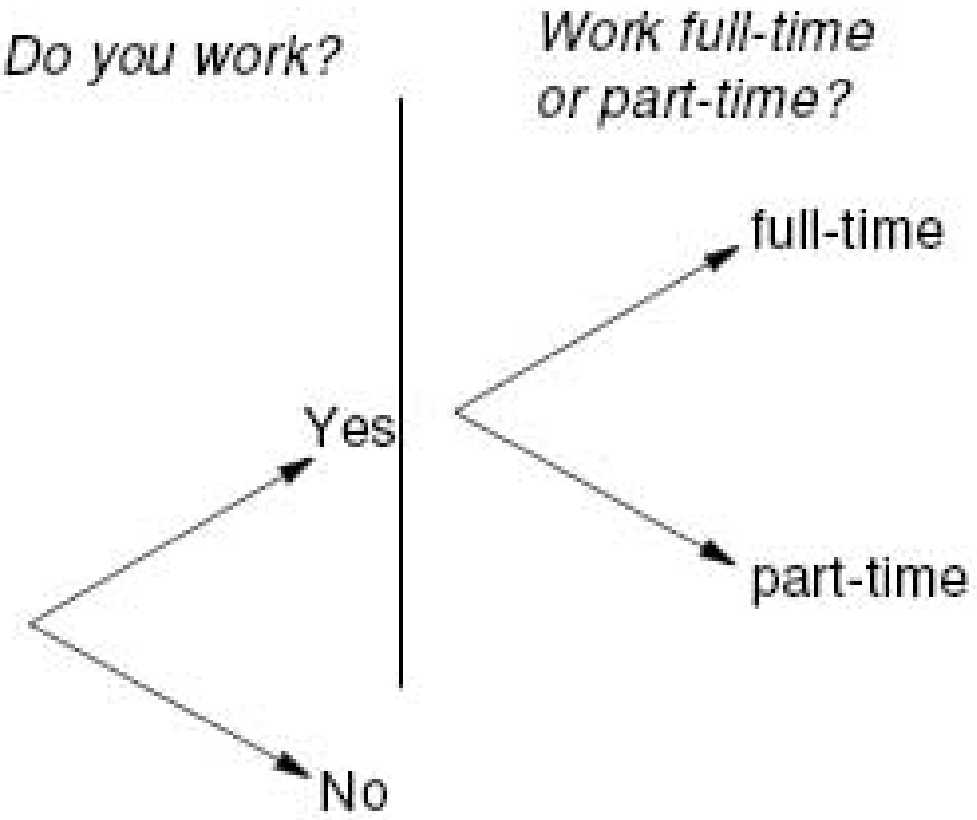
Max.Lik.: $log \prod_{i=1}^{n} \prod_{m=1}^{M} P(y_i = m|x_i, z_{im})^{\mathbb{1}\{y_i=m\}}$

The Problem of dependent - eg. sequential decisions

- What if decisions / alternatives are not independent?
- Take as an example a sequential decision rule (see next slide)
- Can be used e.g. when dependent variable can be separated into a sequence of binary choices
- For the simplest sequential model, we also assume u_i independent.
- Some examples:

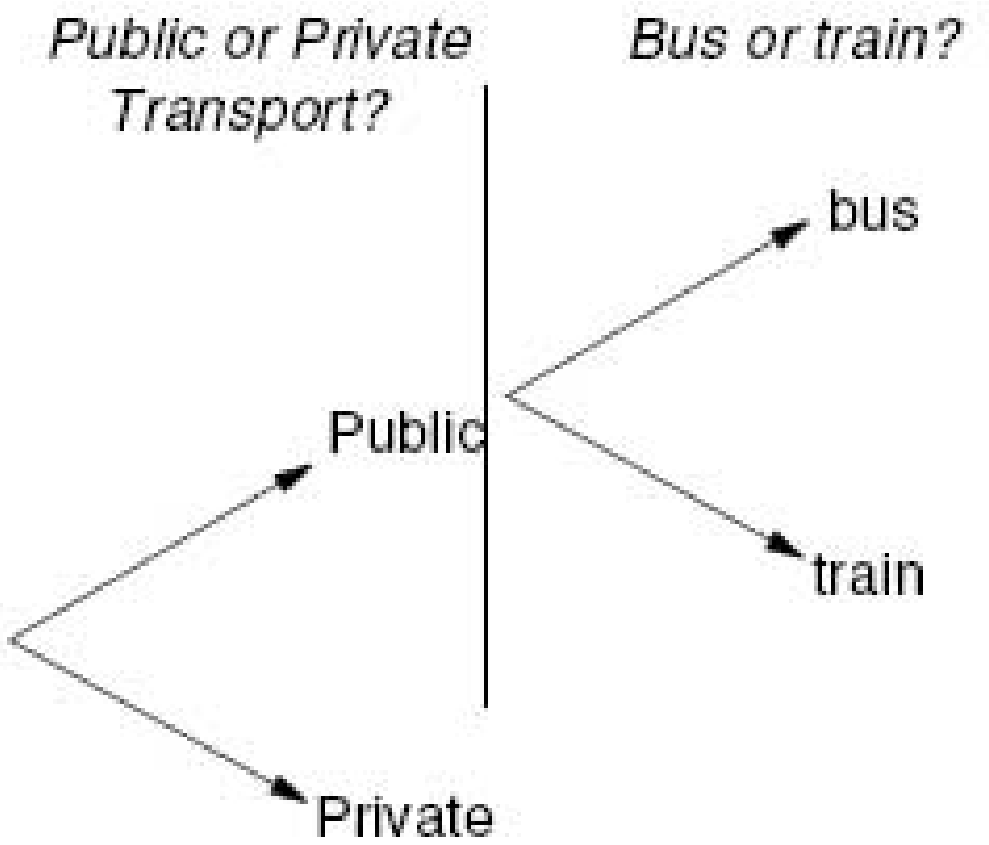
Sequential decisions: Example 1

- labour force status



Sequential decisions: Example 2

- transport mode



The Data: Notation

- Consider a sample of data $\{y_{Ai}, y_{Bi}, x_{Ai}, x_{Bi}\}$.
- Let y_{Ai} represent a binary indicator variable for some discrete choice.
- Let y_{Bi} represent a second discrete choice, observed only when $y_{Ai} = 1$.
- Let the k_A explanatory variables x_{Ai} influence the first choice.
- Let the k_B explanatory variables x_{Bi} influence the conditional choice.
- For the first stage, assume with $u_{Ai} \sim N(0,1)$ iid (recall our discussions about latent models for binary choice outcome)

$$y_{Ai}^* = x_{Ai}'\beta_A + u_{Ai}$$

The sequential Probit/Logit Model

- Observe $y_{Ai} = \mathbb{1}(y_{Ai}^* > 0)$.
- Hence $P(y_{Ai} = 1|x_{Ai}) = \Phi(x_{Ai}'\beta_A)$.
- Estimation by standard Probit MLE on the full sample.
- For the second stage, note first that

$$P(y_{Ai} = 1, y_{Bi} = 1) = P(y_{Ai} = 1) * P(y_{Bi} = 1|y_{Ai} = 1).$$

- Hence, select a sample of the n_B observations for which $y_{Ai} = 1$.
- Define for $u_{Bi} \sim N(0, 1)$ iid **and independent from** u_{Ai}

$$y_{Bi}^* = x_{Bi}'\beta_B + u_{Bi}$$

Resulting Probabilities

- For the second stage $y_{Bi} = \mathbb{1}(y_{Bi}^* > 0)$. So,

$$P(y_{Bi} = 1|x_{Bi}) = \Phi(x_{Bi}'\beta_B).$$

- Estimation by standard Probit MLE on the selected sample.
- The overall probabilities of the three possible outcomes are

$$\begin{aligned}
P(y_{Ai} = 0|x_{Ai}) &= 1 - \Phi(x_{Ai}'\beta_A) \\
P(y_{Ai} = 1, y_{Bi} = 0|x_{Ai}, x_{Bi}) &= \Phi(x_{Ai}'\beta_A) * [1 - \Phi(x_{Bi}'\beta_B)] \\
P(y_{Ai} = 1, y_{Bi} = 1|x_{Ai}, x_{Bi}) &= \Phi(x_{Ai}'\beta_A) * \Phi(x_{Bi}'\beta_B)
\end{aligned}$$

- Upside: easy to estimate
- Downside: ignores a potential correlation between u_{Ai} and u_{Bi}.

A Bivariate Distribution Approach

- Binary decisions may form part of a **system of choices** rather than a sequence, eg. simultaneous decisions of work and take-up of paid childcare.
- For convenience, generalize also to possibility for choice if $Y_{iA} = 0$
- Can apply the Bivariate Probit in these circumstances:
- Consider $\{y_{Ai}, y_{Bi}, x_{Ai}, x_{Bi}\}$ for $i = 1, ..., n$.
- Here, y_{Ai} and y_{Bi} represent two binary indicator variables.
- Assume an underlying system of propensities:

$$y^*_{Ai} = x'_{Ai}\beta_A + u_{Ai},$$
$$y^*_{Bi} = x'_{Bi}\beta_B + u_{Bi}.$$

The resulting Model

- The observability criteria:

$$y_{Ai} = \mathbb{1}(y_{Ai}^* > 0),$$
$$y_{Bi} = \mathbb{1}(y_{Bi}^* > 0).$$

- For a **Bivariate Probit** model, u_{Ai} and u_{Bi} are bivariate normal:

$$\phi_2(u_A, u_B; \rho) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} * \exp\Big(- \frac{u_A^2 + u_B^2 - 2\rho u_A u_B}{1-\rho^2}\Big)$$
$$\Phi_2(u_A, u_B; \rho) = \int_{-\infty}^{u_B}\int_{-\infty}^{u_A} \phi_2(u, v; \rho)\partial u \partial v$$

- Note that when $\rho = 0, \phi_2(u_A, u_B; 0) = \phi(u_A) * \phi(u_B)$.
- so that
  $\Phi_2(u_A, u_B; 0) = \int_{-\infty}^{u_B}\int_{-\infty}^{u_A} \phi(u_A) * \phi(u_B)\partial u \partial v = \Phi(u_A) * \Phi(u_B)$.

Estimating a Bivariate Probit

- Derive probabilities P_{jk} for $j, k = 0, 1$.
- For example,

$$\begin{aligned} P_{00i} &= P(y_{Ai} = 0, y_{Bi} = 0 | x_{Ai}, x_{Bi}) = P(y_{Ai}^* \leq 0, y_{Bi}^* \leq 0 | x_{Ai}, x_{Bi}) \\ &= P(u_{Ai} \leq -x_{Ai}'\beta_A, u_{Bi} \leq -x_{Bi}'\beta_B) = \Phi_2(-x_{Ai}'\beta_A, -x_{Bi}'\beta_B; \rho) \end{aligned}$$

- Similarly,

$$\begin{aligned} P_{11i} &= P(y_{Ai} = 1, y_{Bi} = 1 | x_{Ai}, x_{Bi}) = \Phi_2(x_{Ai}'\beta_A, x_{Bi}'\beta_B; \rho) \\ P_{01i} &= P(y_{Ai} = 0, y_{Bi} = 1 | x_{Ai}, x_{Bi}) = P(y_{Bi} = 1 | x_{Ai}, x_{Bi}) - P_{11i} \\ &= \Phi(x_{Bi}'\beta_B) - P_{11i} \\ P_{10i} &= P(y_{Ai} = 1, y_{Bi} = 0 | x_{Ai}, x_{Bi}) = P(y_{Ai} = 1 | x_{Ai}, x_{Bi}) - P_{11i} \\ &= \Phi(x_{Ai}'\beta_A) - P_{11i} \end{aligned}$$

- where in the sequential case we merge P_{00i} & P_{01i} to $P(y_{Ai} = 0 | x_{Ai})$

Likelihood Estimation for the Bivariate Probit

- Estimation then follows by ML:

$$L(\beta,\rho)=\prod_{i=1}^{n} P_{00i}^{(1-y_{Ai})(1-y_{Bi})}\cdot P_{01i}^{(1-y_{Ai})y_{Bi}}\cdot P_{10i}^{y_{Ai}(1-y_{Bi})}\cdot P_{11i}^{y_{Ai}y_{Bi}}$$

- Taking logs,

$$\begin{aligned}\ln L(\beta_A,\beta_B,\rho)=\sum_{i=i}^{n}\Big\{&(1-y_{Ai})(1-y_{Bi})\ln P_{00i}\\&+(1-y_{Ai})y_{Bi}\ln P_{01i}\\&+y_{Ai}*(1-y_{Bi})\ln P_{10i}\\&+y_{Ai}y_{Bi}\ln P_{11i}\Big\}\end{aligned}$$

Independence of irrelevant alternatives assumption revisited

- We defined the multinomial logit model but pointed out its weakness when some of the alternatives is similar
- as then the IoIA is likely to be violated
- The nested or nonindependet logit model alleviates that weakness
- For the sake of presentation start with trichotomous model
- Consider the transport example with red.bus, blue.bus and car

$$U_j = \mu_j + e_j, \qquad j = 0, 1, 2$$

  be the utilities associated with **car**, **red.bus** and **blue.bus**
- and think of e_1, e_2 having thus a joint (random) component

The Gumbel's Type B bivariate extreme-value distribution

- Clearly, the IoIA assumption is unreasonable for e_1, and e_2 although e_0 may be assumed independent of the others
- A convenient way to account for the correlation is

$$F(e_1, e_2) = \exp\left\{-\left[\exp(-\rho^{-1}e_1) + \exp(-\rho^{-1}e_2)\right]^{\rho}\right\} \tag{10}$$

  for $0 < \rho \leq 1$.
- This is known as the Gumbel's Type B bivariate extreme-value distribution.
- The correlation coefficient can be shown to be $1 - \rho^2$
- If $\rho = 1$, ie. in case of independence between e_1 and e_2, the $F(e_1, e_2)F(e_0)$ becomes the multinomial logit model

Probabilities for the Nested Logit Model

- As for $F(e_0) = \exp[-\exp(-e_0)]$ we get with equation (10)

$$P(y=0) = \frac{\exp(\mu_0)}{\exp(\mu_0) + [\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)]^{\rho}} \quad (11)$$

$$P(y=1|y\neq 0) = \frac{\exp(\rho^{-1}\mu_1)}{\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)} \quad (12)$$

- All probabilities can be deduced from these, and thus the likelihood!
- By dividing the numerator and the denominator of (11) by $exp(\mu_0)$
- and those of (12) by $exp(\rho^{-1}\mu_0)$
- We see that the probabilities depend on $\mu_2 - \mu_0$, $\mu_1 - \mu_0$ and ρ
- Typically one specifies $\mu_j = x_j'\beta$ for $j = 0, 1, 2$
- (12) $\Rightarrow$ choice between two similar alternatives is made according to a binary logit

Derivation of (11)

- (11) suggests that choice between car and non-car is like logit model
- except that a certain kind of a weighted average of $exp(\mu_1)$ and $exp(\mu_2)$ is used

$$
\begin{aligned}
&= P(U_0 > U_1, U_0 > U_2) = P(\mu_0 + e_0 > \mu_1 + e_1, \mu_0 + e_0 > \mu_2 + e_2) \\
&= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{e_0+\mu_0-\mu_1} \left[\int_{-\infty}^{e_0+\mu_0-\mu_2} \exp(-e_0) \cdot \exp[-\exp(-e_0)] f(e_1, e_2) de_2 \right] de_1 \right\} de_0 \\
&= \int_{-\infty}^{\infty} \exp\left(-e_0 - \exp(-e_0) - \{\exp[(\mu_1 - \mu_0 - e_0)/\rho] + \exp[(\mu_2 - \mu_0 - e_0)/\rho]\}^{\rho}\right) de_0 \\
&= \int_{-\infty}^{\infty} \exp(-e_0) \exp[-\alpha \exp(-e_0)] de_0 = \alpha^{-1}
\end{aligned}
$$

giving $P(y = 0) = 1/\alpha = \{1 + \exp(-\mu_0)[\exp(\rho^{-1}\mu_1) + \exp(\rho^{-1}\mu_2)]^{\rho}\}^{-1}$

Derivation of (12)

First observe that (last equation follows from distribution)

$$P(y=1|y\neq 0) = P(U_1 > U_2 | U_1 > U_0 \text{ or } U_2 > U_0) = P(U_1 > U_2)$$

$$\begin{aligned} &= P(\mu_1 + e_1 > \mu_2 + e_2) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{e_1+\mu_1-\mu_2} f(e_1, e_2) de_2 \right] de_1 \\ &= \int_{-\infty}^{\infty} \exp(-e_1) \left(1 + \exp\{-\rho^{-1}[\mu_1 - \mu_2]\}\right)^{\rho-1} \cdot \exp\left(- \exp[-e_1] \cdot \right. \\ &\quad \left. \cdot \left[1 + \exp\{-\rho^{-1}(\mu_1 - \mu_2)\}\right]^{\rho} \right) de_1 \;=\; \left(1 + \exp[-\rho^{-1}\{\mu_1 - \mu_2\}]\right)^{-1} \end{aligned}$$

where in the third line we used

$$\frac{\partial F(e_1, e_2)}{\partial e_1} = \left(\exp\{-\rho^{-1} e_1\} + \exp\{-\rho^{-1} e_2\}\right)^{\rho-1} \cdot \exp(-\rho^{-1} e_1) F(e_1, e_2)$$

Remark: H_0: IoIA $\iff$ H_0: $\rho = 1$

Generalized Linear Models

—

Exponential Families

Generalized Linear Models and beyond

Model, notations

$$E(Y|x) = G(x'\beta)$$

or, more generally,

$$\mu = G(\eta)$$

being the components of a GLM.

Discussion of what we have done so far.

Consider now

- distribution of Y (exponential family)
- still denote *link* function G (note that other authors call G^{-1} the link function)

allows us to treat a large family in one shot

Exponential Family

$$f(y, \theta, \psi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi) \right\}$$

where

- f is the density of Y (continuous)
- f is the probability function of Y (discrete)

note that

- in GLM $\theta = \theta(\beta)$
- θ is the parameter of interest
- ψ is a nuisance parameter (as σ in regression)

Example 1

- Let $Y \sim N(\mu, \sigma^2)$.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\}$$
$$= \exp\left\{y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right\}$$

exponential family with

$$a(\psi) = \psi^2$$
$$b(\theta) = \frac{\theta^2}{2}$$
$$c(y, \psi) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$$

where $\psi = \sigma$ and $\theta = \mu$

Example 2

- Y is Bernoulli:

$$f(y) = P(Y = y) = p^y (1-p)^{1-y} = \begin{cases} p & \text{if} \quad y = 1 \\ 1-p & \text{if} \quad y = 0 \end{cases}$$

- Transform into

$$P(Y = y) = \left(\frac{p}{1-p}\right)^y (1-p) = \exp\left\{y \log\left(\frac{p}{1-p}\right)\right\}(1-p)$$

- Define *logit*:

$$\theta = \log\left(\frac{p}{1-p}\right) \quad \Longleftrightarrow \quad p = \frac{e^\theta}{1+e^\theta}$$

Example 2

- Thus, we arrive at

$$P(Y = y) = \exp\{y\theta + \log(1 - p)\}$$

- The parameters of the exponential family are

$$\begin{aligned} a(\psi) &= 1 \\ b(\theta) &= -\log(1 - p) = \log(1 + e^{\theta}) \\ c(y, \psi) &= 0. \end{aligned}$$

- This is a one-parameter exponential family as there is no nuisance parameter ψ

Example 3

- For Y being conditional Poisson :

$$\begin{aligned} f(y|x) &= P(Y=k|x) = \exp\left\{\frac{k\theta(x)-b(\theta(x))}{a(\psi(x))}+c(k,\psi(x))\right\} \\ &= \exp\left\{k\theta(x)-e^{\theta(x)}+c(k)\right\} = (e^{\theta(x)})^k \cdot e^{-\exp\{\theta(x)\}} \cdot e^{c(k)} \\ &= \mu^k(x)\cdot e^{-\mu(x)}\cdot e^{-\ln(k!)} = \mu^k(x)/(k!)\cdot e^{-\mu(x)} \end{aligned}$$

- with canonical link *log* (or ln) such that $\eta = \theta$

$$\mu(x) = e^{\theta(x)} = e^{x'\beta}$$

- when having linear index $\eta = x'\beta$
- This is the typical way to model a regression for counting data

Properties of the exponential family

$$0 = E\left\{\frac{Y - b'(\theta)}{a(\psi)}\right\}$$

$$E\left\{\frac{-b''(\theta)}{a(\psi)}\right\} = -E\left\{\frac{Y - b'(\theta)}{a(\psi)}\right\}^2$$

- From this we conclude

$$E(Y) = \mu = b'(\theta)$$
$$Var(Y) = V(\mu)a(\psi) = b''(\theta)a(\psi)$$

- The expectation of Y only depends on θ whereas the variance of Y depends on θ and ψ.
- We assume that $a(\psi)$ is constant (there are modifications using prior weights).

Link Functions

$$\mu = G(\eta), \quad \eta = x'\beta$$

- The *canonical* link is given when

$$x'\beta = \eta = \theta$$

- **Example**: Y Bernoulli
  - *logit* $\mu = 1/\{1 + exp(-\eta)\}$ (canonical)
  - *probit* $\mu = \Phi(\eta)$
  - *complementary log-log link* $\eta = \log\{-\log(1-\mu)\}$.
- **Example**: *power* link
  - $\eta = \mu^\lambda$ (if $\lambda \neq 0$) and $\eta = \log \mu$ (if $\lambda = 0$)

	Notation	Range of y	$b(\theta)$	$\mu(\theta)$	Canonical link $\theta(\mu)$	Variance $V(\mu)$	$a(\psi)$
Bernoulli	$B(1,\mu)$	$\{0,1\}$	$\log(1+e^\theta)$	$e^\theta/(1+e^\theta)$	logit	$\mu(1-\mu)$	1
Binomial	$B(k,\mu)$	$\{0,1,\ldots,k\}$	$k\log(1+e^\theta)$	$ke^\theta/(1+e^\theta)$	logit	$\mu\left(1-\frac{\mu}{k}\right)$	1
Poisson	$P(\mu)$	$\{0,1,2,\ldots\}$	$\exp(\theta)$	$\exp(\theta)$	log	μ	1
Geometric	$GE(\mu)$	$\{0,1,2,\ldots\}$	$-\log\left(1-e^\theta\right)$	$e^\theta/(1-e^\theta)$	$\log\left(\frac{\mu}{1+\mu}\right)$	$\mu+\mu^2$	1
Negative Binomial	$NB(\mu,k)$	$\{0,1,2,\ldots\}$	$-k\log\left(1-e^\theta\right)$	$ke^\theta/(1-e^\theta)$	$\log\left(\frac{\mu}{k+\mu}\right)$	$\mu+\frac{\mu^2}{k}$	1
Normal	$N(\mu,\sigma^2)$	$(-\infty,\infty)$	$\theta^2/2$	θ	identity	1	σ^2
Exponential	$Exp(\mu)$	$(0,\infty)$	$-\log(-\theta)$	$-1/\theta$	reciprocal	μ^2	1
Gamma	$G(\mu,\nu)$	$(0,\infty)$	$-\log(-\theta)$	$-1/\theta$	reciprocal	μ^2	$1/\nu$
Inverse Gaussian	$IG(\mu,\sigma^2)$	$(0,\infty)$	$-(-2\theta)^{1/2}$	$-(-2\theta)^{-1/2}$	squared reciprocal	μ^3	σ^2

Table: Characteristics of some GLM distributions

Maximum-Likelihood Algorithm

- Observations $Y = (Y_1, \ldots, Y_n)'$,

$$E(Y_i|x_i) = \mu_i = G(x_i'\beta).$$

- We maximize the log-likelihood of the sample, which is

$$\ell(\mu, Y, \psi) = \sum_{i=1}^{n} \log f(Y_i, \theta_i, \psi).$$

- Alternatively, one may minimize the deviance function

$$D(Y, \mu, \psi) = 2\left\{\ell(Y, Y, \psi) - \ell(\mu, Y, \psi)\right\}.$$

Log-Likelihood and Exponential Family

$$\ell(\mu, Y, \psi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\psi)} - c(Y_i, \psi) \right\}$$

- Neither $a(\psi)$ nor $c(Y_i, \psi)$ have an influence on the maximization w.r.t. β, hence we maximize only

$$\widetilde{\ell}(\mu, Y, \psi) = \sum_{i=1}^{n} \{ Y_i \theta_i - b(\theta_i) \}$$

- Gradient of $\widetilde{\ell}$:

$$\mathcal{D}(\beta) = \frac{\partial}{\partial \beta} \widetilde{\ell}(\mu, Y, \psi) = \sum_{i=1}^{n} \{ Y_i - b'(\theta_i) \} \frac{\partial}{\partial \beta} \theta_i$$

Log-Likelihood and Exponential Family

- We have to solve

$$\mathcal{D}(\beta) = 0$$

  - *Newton-Raphson* step

$$\widehat{\beta}_{new} = \widehat{\beta}_{old} - \left\{\mathcal{H}(\widehat{\beta}_{old})\right\}^{-1} \mathcal{D}(\widehat{\beta}_{old})$$

  - *Fisher scoring* step

$$\widehat{\beta}_{new} = \widehat{\beta}_{old} - \left\{E\mathcal{H}(\widehat{\beta}_{old})\right\}^{-1} \mathcal{D}(\widehat{\beta}_{old})$$

- and know that $\frac{\partial \tilde{\ell}}{\partial \beta} = \frac{\partial \tilde{\ell}}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta}$ with $\frac{\partial \mu}{\partial \eta} = G'(\eta)$, $\frac{\partial \eta}{\partial \beta} = x$ and

$$\frac{\partial \mu}{\partial \theta} = b''(\theta) = V(\mu) \Leftrightarrow \frac{\partial \theta}{\partial \mu} = 1/V(\mu)$$

Gradient and Hessian after some calculation

$$\mathcal{D}(\beta) = \sum_{i=1}^{n} \{Y_i - \mu_i\} \frac{G'(\eta_i)}{V(\mu_i)} x_i$$

- Newton-Raphson

$$\begin{aligned}
\mathcal{H}(\beta) &= \sum_{i=1}^{n} \left\{ -b''(\theta_i) \left(\frac{\partial}{\partial \beta} \theta_i \right) \left(\frac{\partial}{\partial \beta} \theta_i \right)' - \{Y_i - b'(\theta_i)\} \frac{\partial^2}{\partial \beta \beta'} \theta_i \right\} \\
&= \sum_{i=1}^{n} \left\{ -\frac{G'(\eta_i)^2}{V(\mu_i)} + \{Y_i - \mu_i\} \frac{G''(\eta_i) V(\mu_i) - G'(\eta_i)^2 V'(\mu_i)}{V(\mu_i)^2} \right\} x_i x_i'
\end{aligned}$$

- Fisher scoring

$$E\mathcal{H}(\beta) = \sum_{i=1}^{n} \left\{ -\frac{G'(\eta_i)^2}{V(\mu_i)} \right\} x_i x_i'$$

Simpler presentation of algorithm (Fisher scoring)

$$W = diag\left(\frac{G'(\eta_1)^2}{V(\mu_1)}, \ldots, \frac{G'(\eta_n)^2}{V(\mu_n)}\right)$$

$$\widetilde{Y} = \left(\frac{Y_1 - \mu_1}{G'(\eta_1)}, \ldots, \frac{Y_n - \mu_n}{G'(\eta_n)}\right)', \quad X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}$$

- Each iteration step for β can be written as

$$\begin{aligned} \beta_{new} &= \beta_{old} + (X'WX)^{-1}X'W\widetilde{Y} \\ &= (X'WX)^{-1}X'WZ \end{aligned}$$

with adjusted dependent variables

$$Z_i = x_i'\beta_{old} + (Y_i - \mu_i)\{G'(\eta_i)\}^{-1}.$$

Remarks

- Canonical link: Newton-Raphson = Fisher scoring
- Initial values
  - For all but binomial models:
    $\mu_{i,0} = Y_i$ and $\eta_{i,0} = G^{-1}(\mu_{i,0})$
  - For binomial models:
    $\mu_{i,0} = (Y_i + \frac{1}{2})/(w_{x_i} + 1)$ and $\eta_{i,0} = G^{-1}(\mu_{i,0})$.
    (w_{x_i} denotes the binomial weights, i.e. $w_{x_i} = 1$ in the Bernoulli case.)
- Convergence is controlled by checking the relative change in β and/or deviance

Asymptotics

Theorem *Denote*

$$\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)' = \{G(x_1'\widehat{\beta}), \ldots, G(x_n'\widehat{\beta})\}'$$

then for $n \to \infty$

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{L} N_p(0, \Sigma)$$

and it holds approximately

$$D(Y, \widehat{\mu}, \psi) \sim \chi^2_{df} \quad \textit{and} \quad 2\{\ell(Y, \widehat{\mu}, \psi) - \ell(Y, \mu, \psi)\} \sim \chi^2_{dim(\beta)}.$$

- The asymptotic covariance of $\widehat{\beta}$ can be estimated by

$$\widehat{(n Var(\hat{\beta}))}^{-1} = \widehat{\Sigma}^{-1} = -E\mathcal{H}(\beta_{last}) = \sum_{i=1}^{n} \left\{ \frac{G'(\eta_{i,last})^2}{V(\mu_{i,last})} \right\} x_i x_i'$$

with *last* denoting the values from the last iteration step

The equivalence of GMM, GLS and (Pseudo-)ML

Introduction to the GMM idea

- The GMM estimation technique is a direct extension of the method of moments technique.
- Gives a consistent (LLN) estimator
- To capture the main ideas let us first concentrate on a simple problem such as parameter estimation for a density
- Assume the data $y = (y_1, \ldots, y_n)$ are generated by a probability distribution indexed by a parameter vector θ with L elements.
- Estimate θ by computing K sample moments of y, setting them equal to population moments, and solving for θ.
- Because sample moments are generally consistent estimators of population moments (under some regularity conditions), $\hat{\theta}$ will be consistent for θ (if it exists).

Examples

Example (Sample Mean)

The model specified for the random variable y_i implies certain expectations, for example

$$E(y_i) = \mu$$

where μ is the mean of the distribution of y_i. Moment condition:

$$E(y_i - \mu) = 0.$$

Sample counterpart:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0.$$

The estimator is the value of $\hat{\mu}$ that satisfies the sample moment condition.

Examples

Example (Ordinary Regression)

Model:

$$y_i = \beta' x_i + \varepsilon_i.$$

Among the assumptions of the model is

$$E(x_i \varepsilon_i) = 0.$$

Sample analog is

$$\frac{1}{n}\sum_{i=1}^{n} x_i \hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n} x_i (y_i - \hat{\beta}' x_i) = \frac{1}{n} X' \hat{\varepsilon} = 0.$$

The estimator of $\hat{\beta}$ is the one that satisfies this moment equation.

Examples

Example (Instrumental Variables)

Counterpart for the instrumental variables estimator is

$$E(z_i \varepsilon_i) = 0.$$

Sample analog is

$$\frac{1}{n} \sum_{i=1}^{n} z_i \hat{\varepsilon}_i = \frac{1}{n} \sum_{i=1}^{n} z_i (y_i - \hat{\beta}' x_i) = \frac{1}{n} Z' \hat{\varepsilon} = 0.$$

The estimator $\hat{\beta}_{IV}$ is the one that satisfies these moment equation.

Examples

Example (Generalized Least Squares)

The GLS estimator is defined by the normal equations

$$E(X'\Omega^{-1}\varepsilon) = 0.$$

which can be transformed to

$$E(X^{*\prime}\varepsilon^*) = 0 \quad \text{or} \quad E(x_i^*\varepsilon_i^*) = 0$$

for one observation. The sample analog that produces the GLS estimator is

$$\frac{1}{n}\sum_{i=1}^{n} x_i^*\hat{\varepsilon}_i^* = \frac{1}{n}X^{*\prime}\hat{\varepsilon}_i^* = \frac{1}{n}X'\Omega^{-1}\hat{\varepsilon} = 0.$$

Examples

Example (Maximum Likelihood)

The moment condition is obtained by equating the derivatives of a log–likelihood to zero. For regular problems we have

$$E\left(\frac{\partial \ln f(y_i, x_i|\theta)}{\partial \theta}\right) = 0.$$

The ML–estimator is obtained by equating the sample analog to zero:

$$\frac{1}{n}\frac{\partial \ln L}{\partial \hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f(y_i, x_i|\hat{\theta})}{\partial \hat{\theta}} = 0.$$

The Method of Moments

- In the given examples there were exactly as many moment equations as there were parameters to be estimated: $K = L$ (exactly identified).
- Suppose that the model involves L parameters,

$$\theta = (\theta_1, \ldots, \theta_L)$$

and that it implies a set of $K = L$ moment condition

$$E[m_k(w_i, \theta)] = 0$$

with $w_i = (y_i, x_i)$ for $k = 1, \ldots, K$.

- Corresponding sample mean for $l = 1, \ldots, L$:

$$\bar{m}_k = \frac{1}{n} \sum_{i=1}^{n} m_k(w_i, \theta) = 0$$

- and it is clear that we can solve this problem exactly if there is no case of redundant (moment) conditions

- Due to the Taylor expansion we know then that

$$0 = \bar{m}(\hat{\theta}) \approx \bar{m}(\theta_0) + \dot{\bar{m}}(\theta_0)'(\hat{\theta} - \theta_0) \text{ where } \dot{\bar{m}} = \frac{\partial \bar{m}}{\partial \theta'}$$

$$\textit{heuristically} \sim \sqrt{n}(\hat{\theta} - \theta_0) \approx -[\dot{\bar{m}}(\theta_0)]^{-1}\sqrt{n}\bar{m}(\theta_0)$$

  if everything correct $\Leftrightarrow$ $Var(\hat{\theta}) \approx \left\{\dot{\bar{m}}'(\theta_0) Var^{-1}[\bar{m}(\theta_0)]\dot{\bar{m}}(\theta_0)\right\}^{-1}$

- Since the moments are sums of the observable statistics, they are random variables whose variances are estimable.
- It is clear that under the assumption that the observations are uncorrelated, we can estimate

$$n\ \widehat{Var}[\bar{m}] = \frac{1}{n}\sum_{i=1}^{n} m'(w_i) \cdot m(w_i) \tag{13}$$

  or more generally

$$n\ \widehat{Cov}[\bar{m}_j, \bar{m}_k] = \frac{1}{n}\sum_{i=1}^{n}(m_j(w_i) - \bar{m}_j) \cdot (m_k(w_i) - \bar{m}_k)$$

From MM to GMM: Generalizing he Method of Moments

- But there are cases in which there are more moment equations than parameters, $L < K$ (overidentified or overdetermined system).
- A system of K equations in L unknown parameters will not have a unique solution (assume that K is strictly less than n).
- It will be necessary to reconcile the $\binom{K}{L}$ different sets of estimates
- Furthermore, by the same logic that makes GLS preferable to OLS, it should be beneficial to use a weighted criterion in which the weights are inversely proportional to the variances of the moments.
- In sum, the idea is then to minimize the weighted sum of squares

$$\bar{m}(\theta)' A \, \bar{m}(\theta)$$

where A is any positive definite matrix that is not a function of θ but can be used for optimization of the estimator's efficiency

- For asymptotics of Minimum distance estimators, see Amemiya (1985)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \bar{m}(\theta)' A\, \bar{m}(\theta)$$

- By virtue of the CLT applied to the sample moments and the Slutsky theorem follows the asymptotic distribution of the GMM–estimator:

$$\hat{\theta}_{GMM} \overset{a}{\sim} N(\theta, V_{GMM}) \text{ where}$$

$$V_{GMM} = \frac{1}{n}[\dot{\bar{m}}' A \dot{\bar{m}}]^{-1} \dot{\bar{m}}' A V_m A \dot{\bar{m}} [\dot{\bar{m}}' A \dot{\bar{m}}]^{-1}$$

because the first order condition is now

$$0 = 2\dot{\bar{m}}(\hat{\theta})' A \bar{m}(\hat{\theta}) \underset{Taylor}{\Leftrightarrow} \dot{\bar{m}}(\hat{\theta})' A \left[\bar{m}(\theta_0) + \dot{\bar{m}}(\theta_0)(\hat{\theta} - \theta_0)\right] \approx 0$$

and therefore

$$(\hat{\theta} - \theta_0) \approx - \left[\dot{\bar{m}}(\hat{\theta})' A \dot{\bar{m}}(\theta_0)\right]^{-1} \dot{\bar{m}}(\hat{\theta})' A \bar{m}(\theta_0)$$

- Note:
$A^{-1} = V_m := V_{asympt}[\sqrt{n}\bar{m}]$ gives $V_{GMMopt} = \frac{1}{n}[\dot{\bar{m}}' V_m^{-1} \dot{\bar{m}}]^{-1}$

Examples

Example (Gamma distribution (cf. Greene))

We use a sample of 20 observations on income:

20.5	31.5	47.7	26.2	44.0	8.28	30.8	17.2	19.9	9.96
55.8	25.2	29.0	85.5	15.1	28.5	21.4	17.7	6.42	84.9

and assume a Gamma distribution with density

$$f(x) = \frac{\lambda^P}{\Gamma(P)} \exp(-\lambda x) x^{P-1}$$

with $x \geq 0, P > 0, \lambda > 0$ and $\Gamma(P) = \int_0^\infty t^{P-1} \exp(-t) dt$.

Example (continued)

We obtain for example the following four moment equations:

$$E\left(x_i - \frac{P}{\lambda}\right) = 0, \quad E\left(x_i^2 - \frac{P(P+1)}{\lambda^2}\right) = 0,$$
$$E\left(\frac{1}{x_i} - \frac{\lambda}{P-1}\right) = 0, \quad E\left(\ln(x_i) - \Psi(P) + \ln(\lambda)\right) = 0$$

where $\Psi(P) = \Gamma'/\Gamma$ is the digamma function. Let $x_1 = x$, $x_2 = x^2$, $x_3 = \ln(x)$, and $x_4 = 1/x$. Then:

$$\bar{m}_1(P,\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{1i} - \frac{P}{\lambda}\right) = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \mu_1) = \bar{x}_1 - \mu_1(P,\lambda)$$

and likewise for $\bar{m}_2(P,\lambda)$, $\bar{m}_3(P,\lambda)$, $\bar{m}_4(P,\lambda)$.

Example (continued)

- For an initial set of estimates, we will use Ordinary Least Squares. Optimization problem:

$$\begin{aligned}\underset{P,\lambda}{\text{Min}} \sum_{l=1}^{4} \bar{m}_l(P,\lambda)^2 &= \sum_{l=1}^{4} [\bar{x}_l - \mu_l(P,\lambda)]^2 \\ &= \bar{m}(P,\lambda)' \bar{m}(P,\lambda).\end{aligned}$$

- This nonlinear optimization problem must be solved iteratively. It turns out that for purposes of the nonlinear optimization procedure, a rescaling of the moment conditions greatly eases convergence. Thus by multiplying the moment conditions by $[\lambda, \lambda^2, 1, (P-1)]$, respectively, the four moment equations are

$$\begin{aligned}\lambda \bar{x}_1 - P = 0, \quad \lambda^2 \bar{x}_2 - P(P+1) = 0, \\ \bar{x}_3 - \Psi(P) + \ln(\lambda) = 0, \quad (P-1)\bar{x}_4 - \lambda = 0.\end{aligned}$$

Example (continued)

- Minimizing the numerically simplified version

$$q = \bar{m}(P,\lambda)' H' H \, \bar{m}(P,\lambda) = \bar{m}^*(P,\lambda)' \bar{m}^*(P,\lambda)$$

  with $H = diag(\lambda, \lambda^2, 1, (P-1))$ gives $\hat{P}_H = 2.1799$ and $\hat{\lambda}_H = 0.0691$.
- else get *tiny subject to starting values* $\hat{P} = 2.0583$ and $\hat{\lambda} = 0.0658$.
- We now look for the efficient weight V_m to get the 'standard' GMM estimator:

$$q_{GMM} = \bar{m}(P,\lambda)' \hat{V}_m^{-1} \bar{m}(P,\lambda)$$

  requires the first-step consistent estimates to compute $\hat{V}_m$.
- With the above estimates and (13)

$$20 \cdot \widehat{Var}(\bar{m}) = \hat{V}_m = \frac{1}{20} \sum_{i=1}^{20} \begin{bmatrix} x_{i1} - \hat{P}/\hat{\lambda} \\ x_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ x_{i3} - \Psi(\hat{P}) + \ln(\hat{\lambda}) \\ x_{i4} - \hat{\lambda}/(\hat{P}-1) \end{bmatrix} \begin{bmatrix} x_{i1} - \hat{P}/\hat{\lambda} \\ x_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ x_{i3} - \Psi(\hat{P}) + \ln(\hat{\lambda}) \\ x_{i4} - \hat{\lambda}/(\hat{P}-1) \end{bmatrix}$$

Example (continued)

- The two estimates are $\hat{P}_{GMM} = 3.3589$ and $\hat{\lambda}_{GMM} = 0.1245$. The value of the function at these two points is $q_{GMM} = 1.9752$.
- Recalculate with these estimates results along Greene, 6th edition

$$\frac{1}{20}\hat{V}_m = \begin{bmatrix} 24.7051 & & & \\ 2307.126 & 229609.5 & & \\ 0.6974 & 58.8148 & 0.2302 & \\ -0.02881 & -2.14227 & -0.0011 & 0.000065414 \end{bmatrix}$$

  which we use to estimate the variance of $\hat{\theta}_{GMM} = (\hat{P}_{GMM}, \hat{\lambda}_{GMM})$
- To complete the computation, we need the derivatives matrix $\dot{\bar{m}}(\hat{\theta})$

Example (continued)

$$\dot{\bar{m}}'(\hat{\theta}) = \begin{bmatrix} \partial m_1/\partial P & \partial m_2/\partial P & \partial m_3/\partial P & \partial m_4/\partial P \\ \partial m_1/\partial \lambda & \partial m_2/\partial \lambda & \partial m_3/\partial \lambda & \partial m_4/\partial \lambda \end{bmatrix}$$
$$= \begin{bmatrix} -1/\lambda & -(2P+1)/\lambda^2 & -\Psi'(P) & \lambda/(P-1)^2 \\ P/\lambda^2 & 2P(P+1)/\lambda^3 & 1/\lambda & -1/(P-1) \end{bmatrix}$$

$$\hat{\dot{\bar{m}}}' = \begin{bmatrix} -8.0328 & -498.01 & -0.34635 & 0.022372 \\ 216.74 & 15178.2 & 8.0328 & -0.42392 \end{bmatrix}.$$

Finally, the estimated asymptotic covariance matrix is

$$\frac{1}{20}[\hat{\dot{\bar{m}}}' V_m^{-1} \hat{\dot{\bar{m}}}]^{-1} = \begin{bmatrix} 0.202084 & 0.0117284 \\ 0.0117284 & 0.000846541 \end{bmatrix}$$

Testing the validity of the moment restrictions

- By construction,

$$q = \bar{m}(\hat{\theta})'[V_{asympt}(\bar{m}(\hat{\theta}))]^{-1}\bar{m}(\hat{\theta})$$

  is a Wald statistic.
- Therefore, under the hypothesis of the model,

$$q \xrightarrow{d} \chi^2_{K-L} \qquad (K > L)$$

Example (continued)

- For the different method used to recompute the GMM estimates, we obtained always $q > 6$. This value exceeds the critical value $\chi^2_{2,0.05} = 5.99$.
- The hypothesis that all four moment equations are satisfied is rejected.

- Only when the restrictions overidentify the parameters, q can be used to make inference about the specification of the model.
- Otherwise $q = 0$ and all the information in the sample is consumed in estimation of the parameters.

Example (continued)

The implications of the rejection are many

- The hypothesis that the data were drawn from an underlying Gamma distribution should be rejected.
- The method used to compute the asymptotic covariance matrix is incorrect.
- The moment equations themselves should be deemed invalid.

- This result is in accordance with the fact that in this example the ML and GMM where significantly different (along calculations in Greene, 6th edition).

Maximum Likelihood Estimation as GMM

- Generalization of estimation of simple parameters in an unconditional distribution (e.g. of μ, σ^2 in ϕ) to a conditional distribution $f(y_1, \ldots, y_n | X; \theta)$.
- The Likelihood function under independence is

$$L(\theta|y, X) = \prod_{i=1}^{n} L(\theta|y_i, X_i) = \prod_{i=1}^{n} f(y_i|X_i; \theta).$$

- The procedure is

$$\max_{\theta \in \Theta} \ln L(\theta) = \max_{\theta \in \Theta} \sum_{i=1}^{n} \ln L_i(\theta).$$

- Set first derivative (score function) to zero:

$$\sum_{i=1}^{n} s_i(\theta) = \sum_{i=1}^{n} \frac{\partial \ln L_i(\theta)}{\partial \theta}\bigg|_{\hat{\theta}} = 0.$$

Recall Properties of MLE

Properties of the MLE:

- Consistency ($plim\hat{\theta} = \theta$)
- Asymptotic efficiency (smallest variance)
- Asymptotically normal: $\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, V)$
- $V = -E^{-1}\left[\frac{\partial^2 \ln L_i(\theta)}{\partial\theta\partial\theta'}\right] = I^{-1}$ (I Fisher information).
  Thus, reaches asymptotically Cramér-Rao-bound.
- Estimate V by $\hat{V}^{-1} = -\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \ln L_i(\theta)}{\partial\theta\partial\theta'}\right|_{\hat{\theta}}$
- Note, that $\hat{V}^{-1} = \frac{1}{n}\sum_{i=1}^{n} s_i(\hat{\theta})s_i'(\hat{\theta})$. if distribution is correctly specified.
- We have seen that for members of the exponential family these statements can be extended; in particular, a correct specification of the distribution is not needed for consistent estimation but an issue for correct inference.

Pseudo (or Quasi) Maximum-Likelihood

- So what happens, if the distribution was not correctly specified?
- Often it is plausible, that $E[s_i(\theta)] = 0$ even though $L(\theta) \neq f(y|X;\theta)$, i.e. wrong distribution.
- As always use the sample analog $\frac{1}{n}\sum_{i=1}^{n} s_i = 0$
- With the GMM or minimum distance (MD) idea we get ML-like estimates, although $f(\cdot)$ may be misspecified:

$$\sqrt{n}(\hat{\theta} - \theta) \;\to\; N(0, V) \quad \text{with} \quad V = I^{-1} J I^{-1}$$

  where $J = E[s_i(\theta)s_i'(\theta)]$ and $I = E[-\dot{s}_i(\theta)]$,
- Note that for correct specification $I = J$ and we get MLE.
- Here we see the aforementioned potentials and problems regarding inference if correct distribution is not known.

Revise Generalized Least Squares Estimation

- Let us allow for non-linear LS
- and any kind of weight
- finally you average over the data
- which is equivalent to moment estimation
- you also may find a corresponding distribution such that it corresponds to Pseudo ML estimation

Limited Dependent Variable Models

- Censored and truncated samples
- Sample selection bias and Mills ratio
- Truncated regression
- The Tobit I Model
- The Selectivity Tobit II Model
- The Selectivity Tobit III Model
- The Double Hurdle Model

Censored and Truncated Samples: Examples

- One may not always observe data on continuous dependent and explanatory variables over the entire population.
- **Examples:**
  - Wages are only observed for people with job
  - Same happens for working hours
  - Individuals whose incomes fall short of some poverty line are assigned incomes equal to that poverty line
  - for records there may also exist an upper bound (highest taxable income)
  - Only people with children may need child care
  - etc.
- We define samples as either truncated or censored depending on the nature of the limitations of the data.

Censored and Truncated Samples: Definition

- Truncated samples:
  - A sample is truncated if data is only available on a subset of the whole population.
  - In many samples, only people having a job are recorded
  - For truncated samples, data is simply not available to the researcher
  - Could be characterized as a sample defect.
- Censored samples:
  - A sample is censored if data is re-coded for a subset of the population.
  - Also censored data can be characterized as a sample defect.
  - Recall example of minimum wages
- **Further Examples:**
  - A model of the demand for tickets to a concert.
  - A model for labor supply (in hours) for labor force survey
  - A model for bilateral trade (before 1990)
  - A wage equation for the entire population

Observability Criteria

- from the dependent variable's perspective
- Consider
  - an observed dependent variable y,
  - a set of explanatory variables x,
  - a latent variable y^*.
- **Truncated Samples:**
  - T1: $y = y^*$ if $y^* > b$, not observed otherwise,
  - T2: $y = y^*$ if $y^* < d$, not observed otherwise,
  - T3: $y = y^*$ if $c < y^* < d$, not observed otherwise.
- **Censored Samples:**
  - C1: $y = y^* \cdot \mathbb{1}(y^* > b) + b \cdot \mathbb{1}(y^* \leq b)$,
  - C2: $y = y^* \cdot \mathbb{1}(y^* < d) + d \cdot \mathbb{1}(y^* \geq d)$,
  - C3: $y = y^* \cdot \mathbb{1}(b < y^* < d) + b \cdot \mathbb{1}(y^* \leq b) + d \cdot \mathbb{1}(y^* \geq d)$.

The Selection Bias: truncation (T1)

- Consider the latent relationship $y^* = x'\beta + u$ with $E(u|x) = 0$.
- Suppose that the observed dependent variable y is truncated from below at zero (T1 from above), i.e. $b = 0$.
- Then, the conditional expectation

$$\begin{aligned} E(y|x, y^* > 0) &= E(y^*|x, y^* > 0) = x'\beta + E(u|x, y^* > 0) \\ &= x'\beta + E(u|u > -x'\beta) \\ &= x'\beta + \alpha \cdot \psi(x'\beta). \end{aligned} \tag{14}$$

(discuss why not / if better $\psi(x)$)

- Suppose we would simply estimate β from model

$$y_i = x_i'\beta + u_i$$

with a truncated sample for $y_i \leq 0$ (from below at zero).

A theoretical model behind ...

- Problem: the OLS or GLS estimates $\hat{\beta}^c$ for that model **may not** converge to the (true) parameters of the underlying latent relationship

$$y^* = x'\beta + u$$

- In a sense, the model above suffers from *omitted variables bias*. The correct model should have been

$$y_i = x_i'\beta + \alpha \cdot \psi(x_i'\beta) + v_i$$

with $E[v_i|x_i, y_i^* > 0] = 0$, cf. equation (14).

- To estimate β from the correct model need to know: What form should $\psi(x_i'\beta)$ take?
- if we know, least squares estimation is possible
- must be related to truncated distribution of u

Conditional Normal Density

- Let $u_i \sim N(0, \sigma^2)$. Then, we may write the density $f(u)$ as

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp(-\frac{u^2}{2\sigma^2})$$
$$= \frac{1}{\sigma} * \frac{1}{\sqrt{2\pi}} * \exp(-\frac{1}{2}(\frac{u}{\sigma})^2) \;\; = \frac{1}{\sigma} * \phi(\frac{u}{\sigma})$$

- Also,

$$P(u > c) \;\; = P(\frac{u}{\sigma} > \frac{c}{\sigma}) \;\; = 1 - \Phi(\frac{c}{\sigma}) \;\; = \Phi(-\frac{c}{\sigma})$$

- Hence,

$$f(u|u > c) = \frac{\frac{1}{\sigma} * \phi(\frac{u}{\sigma})}{1 - \Phi(\frac{c}{\sigma})} = \frac{\frac{1}{\sigma} * \phi(\frac{u}{\sigma})}{\Phi(-\frac{c}{\sigma})} \qquad \text{if } u > c, \text{ else } 0$$

Conditional Expectation for a Normal Random Variable

Let $u_i \sim N(0, \sigma^2)$ (homoscedastic!).
Then the conditional expectation $E(u|u > c)$ may be derived as

$$
\begin{aligned}
E[u|u>c] &= \int_c^\infty u * f(u|u>c)\partial u = \int_c^\infty u * \frac{\frac{1}{\sigma} * \phi(\frac{u}{\sigma})}{1-\Phi(\frac{c}{\sigma})}\partial u \\
&= \frac{1}{1-\Phi(\frac{c}{\sigma})} * \int_c^\infty \frac{u}{\sigma} * \phi(\frac{u}{\sigma})\partial u \\
&= \frac{\sigma}{1-\Phi(\frac{c}{\sigma})} * \int_{c/\sigma}^\infty z * \phi(z)\partial z = \frac{\sigma}{1-\Phi(\frac{c}{\sigma})} * \Big[-\phi(z)\Big]_{c/\sigma}^\infty \\
&= \frac{\sigma * \phi(\frac{c}{\sigma})}{1-\Phi(\frac{c}{\sigma})} .
\end{aligned}
$$

Inverse Mills Ratio

- To determine $\alpha\psi$ need to assume distribution for u.
- Take normal distribution (ϕ, Φ) with variance σ^2.
- Then, with straight forward integration

$$E[u|y^* > 0] = E[u|u > -x'\beta] =$$

$$= \int_{-x'\beta}^{\infty} \sigma \frac{\phi(\frac{u}{\sigma})}{\left\{1 - \Phi(\frac{-x'\beta}{\sigma})\right\}} du = \sigma \frac{\phi(z)}{\Phi(z)}$$

  where $z = x'\beta/\sigma$
- Therefore $\alpha = \sigma$
- and $\psi(x'\beta) = \phi(z)/\Phi(z) = \lambda(z)$ *called the inverse Mills ratio*
- sometimes we find in the literature $\lambda(x'\beta)$

Truncated Regression Model (General Case)

- What if u is not normal?
- Recall **Conditional distribution / density**
- The conditional density for a random variable u (recall, given x, we need only to consider distribution of error term) with
  - unconditional density $f(.)$
  - cumulative distribution $F(.)$
  - for truncation from below at c
- can be written as

$$f(u|u > c) = \frac{f(u)}{P(u > c)} = \frac{f(u)}{1 - P(u \leq c)} = \frac{f(u)}{1 - F(c)}. \quad (15)$$

- is still a density: clearly ≥ 0; integrating over $[c, \infty)$ gives

$$\int_c^\infty f(u|u > c)\partial u = \int_c^\infty \frac{f(u)}{1 - F(c)}\partial u$$

$$= \frac{1}{1 - F(c)} * \int_c^\infty f(u)\partial u = \frac{1}{1 - F(c)} * (1 - F(c)) = 1.$$

A Model for Truncated Regression

- Then we can calculate

$$\begin{aligned} E[u|u>c] &= \int_c^\infty u \cdot f(u|u>c)\partial u = \int_c^\infty u \cdot \frac{f(u)}{1-F(c)}\partial u \\ &= \frac{1}{1-F(c)} * \int_c^\infty uf(u)\partial u \;=\; \textit{to be calculated} \end{aligned}$$

- So still consider a truncated sample where $y = y^*$ if $y^* > 0$, but y not observed otherwise.
- The model you would like to estimate is

$$y_i^* = x_i'\beta + u_i \;, \quad u \sim f(\cdot)$$

but you have only truncated data

- Rest as before ...

When knowing Distribution: why not ML?

- If we assumed the unconditional distribution of u to be normal and homoscedastic:

$$u_i \sim N(0, \sigma^2) \;\Rightarrow\; Y^*|X \sim N(X'\beta, \sigma^2)$$

- if y_i is truncated from below at 0, then we only observe u_i over the limited range for which $u_i > -x_i'\beta$ – and so is $y_i - x_i'\beta$
- Then, conditional density of u_i is, compare eqn (15),

$$f(u_i|u_i > c) = \frac{\frac{1}{\sigma} * \phi(\frac{u_i}{\sigma})}{1 - \Phi(-z_i)} = \frac{\frac{1}{\sigma} * \phi(\frac{u_i}{\sigma})}{\Phi(z_i)}$$

where $z_i = (x_i'\beta)/\sigma$.

$$f(y_i|x_i, u_i > -x_i'\beta) = \frac{1}{\sigma} \cdot \phi(\frac{y_i - x_i'\beta}{\sigma})/\Phi(x_i'\beta/\sigma)$$

ML-Estimation

- For each observation in the truncated sample, the likelihood contribution is precisely the conditional density

$$L_i(\beta,\sigma) = \frac{\frac{1}{\sigma} * \phi(\frac{y_i - x_i'\beta}{\sigma})}{\Phi(x_i'\beta/\sigma)} \Rightarrow L(\beta,\sigma) = \prod_{i=1}^{N} L_i(\beta,\sigma) \ .$$

- The log likelihood is therefore

$$l(\beta,\sigma|y,x) = \sum_{i=1}^{N} -ln(\sigma) + ln\{\phi(\frac{y_i - x_i'\beta}{\sigma})\} - ln\{\Phi(x_i'\beta/\sigma)\}$$

- Maximization gives ML-parameter estimates for β and σ.
- **Popular Alternative:** Two step Heckman estimator

Interpretation

- The marginal effect of a change in x_i on the conditional expectation:

$$E(y_i|x_i, y_i > 0)$$

  is not simply β.
- Comparing with eqn (14), it is rather

$$\frac{\partial E(y_i|x_i, y_i > 0)}{\partial x_i} = \beta + \alpha * \frac{\partial \psi(z_i)}{\partial x_i}$$

- Nevertheless, also β has a clear interpretation; actually

$$\frac{\partial E(y_i^*|x_i)}{\partial x_i} = \beta$$

- **Discuss!**
- Need to know ψ (inverse Mills ratio, commonly denoted by λ).

The Tobit Model (Tobit I)

- *Censored regression model* or **Tobit model** (attributable to Tobin, 1958, gives Tobin-probit).
- The difference here is that data is available on the entire sample.
- But the dependent variable is censored at some value (say, 0).
- As before, the censoring (or truncation) is determined by the equation / model of interest.
- Consider again the latent relationship

$$y_i^* = x_i'\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

  with a censored variable $y_i = y_i^* * \mathbb{1}(y_i^* > 0)$.
- Will derive MLE making use what we have learned above (inverse Mills ration)

Estimating the Tobit I Model

- Given normality for u_i, the *probability* of observing a censored observation is

$$\begin{aligned} P(y_i = 0|x_i) = P(y_i^* \leq 0|x_i) &= P(u_i \leq -x_i'\beta) \\ &= \Phi(-z_i) = 1 - \Phi(z_i) \end{aligned}$$

  where $z_i = x_i'\beta/\sigma$.
- For uncensored observations, we write the density of y_i in the normal way as

$$f(y_i) = \frac{1}{\sigma} * \phi(\frac{y_i - x_i'\beta}{\sigma}).$$

- Then, the sample likelihood function can be compiled as

$$L(\beta, \sigma) = \prod_{y_i=0} 1 - \Phi(\frac{x_i'\beta}{\sigma}) \prod_{y_i>0} \frac{1}{\sigma} * \phi(\frac{y_i - x_i'\beta}{\sigma}).$$

Interpreting the Tobit I

- Given a latent relationship of the form

$$y_i^* = x_i'\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

with an observability criterion

$$y_i = y_i^* * \mathbb{1}(y_i^* > 0)$$

we have that

$$P(y_i = 0|x_i) = 1 - \Phi(x_i'\beta/\sigma)$$
$$P(y_i > 0|x_i) = \Phi(x_i'\beta/\sigma).$$

- The expected value of the *observed* dependent variable y_i (censored at zero) is

$$\begin{aligned} E(y_i|x_i) &= P(y_i > 0|x_i) * E(y_i|x_i, y_i > 0) \\ &= \Phi(x_i'\beta/\sigma) * [x_i'\beta + \sigma * \lambda(x_i'\beta/\sigma)] \end{aligned}$$

where $\lambda(z) = \phi(z)/\Phi(z)$.

Marginal Effects

- On the latent variable:
$$\frac{\partial E(y_i^*|x_i)}{\partial x_i} = \beta$$
- On the *observed dependent variable*, **censored at zero**, for the whole sample:
$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \beta * \Phi(x_i'\beta/\sigma)$$
- On the **non-censored** *observed dependent variable*:
$$\frac{\partial E(y_i|x_i, y_i > 0)}{\partial x_i} = \beta + \sigma * \frac{\partial \lambda}{\partial x_i}$$

Testing and Miscellaneous

- Specification tests can be derived easily due to the use of maximum likelihood methods: think of LM-tests, Likelihood ratio test, etc.
- Extensions to models and estimators without the normality assumption are available though not trivial. For example, with logit, the inverse Mills ratio looks more complicated.
- Moreover, semiparametric methods to avoid distribution assumptions can get complicated. Can be constructed e.g. from Methods of Moments, see future chapters.
- If, as will happen on the next slides, the selection rule is based on a different model, then a two step estimator is thinkable, too.

Alternative Selection Rules

Recall that so far we have assumed that censoring and truncation are driven by the model of interest. What happens if there is a simultaneous or sequential decision process behind?

- Consider a structural latent relationship:

$$y_i^* = x_i'\beta + u_i$$

and a latent, possibly censored, observability relationship for y_i:

$$I_i^* = z_i'\gamma + v_i.$$

- Disturbance terms u_i and v_i are (jointly) normal and homoscedastic with covariance $cov(u_i, v_i) = \sigma_{uv}$.
- Without loss of generality, for identification set $\sigma_v = 1$ (discuss)

Tobit II, III, etc.

Then, we get the following **observability criteria**

$$\begin{aligned}
\text{Truncated Regression} \quad & y_i = y_i^* \quad \text{if} \quad y_i^* > 0 \quad \text{(not observed else)} \\
\text{Tobit I Model} \quad & y_i = y_i^* * \mathbb{1}(y_i^* > 0) \\
\text{Selectivity / Tobit II Model} \quad & y_i = y_i^* * \mathbb{1}(I_i^* > 0) \\
\text{Selectivity / Tobit III Model} \quad & y_i = y_i^* * \mathbb{1}(I_i^* > 0) \qquad \text{with } I_i \text{ censored} \\
\text{Double Hurdle Model} \quad & y_i = y_i^* * \mathbb{1}(I_i^* > 0 \quad \text{and} \quad y_i^* > 0)
\end{aligned}$$

(Further assumption: $y_i \geq 0$)
Note that we will need multivariate distributions now.
For the normal distribution, it is helpful to know

$$\begin{aligned}
(x, y) \quad &\sim \quad N\left[(\mu_x, \mu_y), \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix}\right] \quad \rightarrow \\
x|y \quad &\sim \quad N\left[\mu_x + V_{xy} V_{yy}^{-1}(y - \mu_y), V_{xx} - V_{xy} V_{yy}^{-1} V_{yx}\right]
\end{aligned}$$

Selectivity Tobit II Model

- For the Selectivity Tobit II model,

$$y_i^* = x_i'\beta + u_i, \qquad I_i^* = z_i'\gamma + v_i.$$

  i.e. z has no effect on $y^*|x$, and x no one on $I^*|z$
- Observability criterion:

$$y_i = y_i^* * \mathbb{1}(I_i^* > 0).$$

- General likelihood for the Selectivity model:

$$\begin{aligned} L(\beta,\gamma,\Sigma) &= \prod_{y_i=0} P(I_i=0|z_i) * \prod_{y_i>0} P(I_i=1|z_i) f(y_i|x_i,z_i,I_i=1) \\ &= \prod_{y_i=0} P(I_i=0|z_i) * \prod_{y_i>0} \int_0^\infty f(I^*,y_i|x_i,z_i)dI^* \qquad (16) \end{aligned}$$

$$\begin{aligned} \text{with} \quad P(I_i=0|z_i) &= P(I_i^*<0|z_i) = P(v_i<-z_i'\gamma) \\ &= \Phi(-z_i'\gamma) = 1-\Phi(z_i'\gamma). \end{aligned}$$

Selectivity Model

For the uncensored observations, need to calculate integral of

$$f(I^*, y_i|x_i, z_i) = f(I_i^*|y_i, x_i, z_i) * f(y_i|x_i, z_i).$$

For bivariate normal disturbances u and v, $(\sigma_{uv} = \rho\sigma_u\sigma_v)$

$$I^*|(y_i, x_i, z_i) \sim N\left[z_i'\gamma + \frac{\sigma_{uv}}{\sigma_u^2}(y_i - x_i'\beta), \sigma_v^2 - (\frac{\sigma_{uv}}{\sigma_u})^2\right]$$

$$\int_0^\infty f(I_i^*|y_i, x_i, z_i)dI^* = P(I_i^* > 0|y_i, x_i, z_i) = \Phi\left(\frac{z_i'\gamma + \frac{\sigma_{uv}}{\sigma_u^2}(y_i - x_i'\beta)}{\sqrt{1 - (\frac{\sigma_{uv}}{\sigma_u})^2}}\right)$$

or if $\sigma_{uv} = 0$, $P(I_i^* > 0|y_i) = \Phi(z_i'\gamma)$. As second part only runs over $y_i > 0$, for those $y_i = y_i^*$ such that $f(y_i|x_i, z_i) = \phi(\{y_i - x_i'\beta\}/\sigma_u)/\sigma_u$.

Putting both together you have (16).

Selectivity Tobit III Model

- Imagine the Selectivity model is a Tobit I
- For example, we have labor supply measured in hours.
- this is only considered if also the selectivity model is of interest, else reduction to binary response would be fine.
- MLE: where f is the density of (I^*, y^*)

$$L(\beta, \gamma, \Sigma) = \prod_{y_i=0} P(I_i = 0|z_i) * \prod_{y_i>0} f(I_i, y_i|x_i, z_i)$$

  which you can easily write down when (I_i^*, y_i^*) are jointly normal distributed, and knowing that for $y_i > 0$, $(I_i, y_i) = (I_i^*, y_i^*)$.
- Alternative (inefficient) **two & three-step** methods
  1. Estimate selectivity model by Tobit Probit to get $\hat{\alpha} := \hat{(}\gamma/\sigma_v)$.
  2. If Probit, then LS on $I_i = z_i'\gamma + \sigma_v\lambda(z_i'\hat{\alpha}) + v_i$
  3. In any case, LS on $y_i = x_i'\beta + \sigma_{uv}/\sigma_v\lambda(z_i\hat{\alpha}) + u_i$

Why Double-Hurdle?

- An alternative to the selectivity model, proposed by Cragg (1971).
- So-called because of two hurdles to be overcome before observing a non-censored observation.
- **Examples**:
  1. Labour Supply:
     1. Do you want to work?
     2. Given that you choose to look for work, can you find a job?
  2. Credit constraints:
     1. Do you want to buy the good?
     2. Given that you want to buy the good, can you obtain credit?

A Proper Model

- For the Double Hurdle model,

$$y_i = y_i^* * \mathbb{1}(l_i^* > 0 \quad \text{and} \quad y_i^* > 0),$$

where, as before, with $(u, v) \sim N_2(0, \Sigma)$

$$y_i^* = x_i'\beta + u_i,$$
$$l_i^* = z_i'\gamma + v_i.$$

- For non-censored observations, the likelihood contribution is similar to the Selectivity Tobit II model:

$$L(\beta, \gamma, \Sigma) = \prod_{y_i=0} P(y_i = 0|x_i, z_i) \prod_{y_i>0} P(y_i > 0|x_i, z_i) f(y_i|y_i > 0, x_i, z_i)$$

using slightly different notation, compare above

Probabilities for the double hurdle model

- The probability of observing a *non-censored* observation is

$$P(y_i > 0|x_i, z_i) = P(I_i^* > 0 \quad \text{and} \quad y_i^* > 0)$$

$$= P(v_i > -z_i'\gamma \quad \text{and} \quad u_i > -x_i'\beta) = \Phi_2\left(z_i'\gamma, \frac{x_i'\beta}{\sigma_u}; \rho\right)$$

- Consequently, for censored observations we have

$$P(y_i = 0|x_i, z_i) = 1 - \Phi_2\left(z_i'\gamma, \frac{x_i'\beta}{\sigma_u}; \rho\right).$$

- furthermore

$$f(y_i|y_i > 0, x_i, z_i) = \int_{-z_i'\gamma}^{\infty} f(u_i) f(v_i|u_i) dv \;\; \Phi_2^{-1}\left(z_i'\gamma, \frac{x_i'\beta}{\sigma_u}; \rho\right)$$

$$= \frac{1}{\sigma_u}\phi\left(\frac{u_i}{\sigma_u}\right)\Phi\left(\frac{z_i'\gamma + \frac{\sigma_{uv}}{\sigma_u^2}(y_i - x_i'\beta)}{\sqrt{1 - (\frac{\sigma_{uv}}{\sigma_u})^2}}\right)\Phi_2^{-1}\left(z_i'\gamma, \frac{x_i'\beta}{\sigma_u}; \rho\right)$$

The Likelihood, Estimation

$$L(\beta,\gamma,\Sigma) = \prod_{y_i=0}[1-\Phi_2\left(z_i'\gamma, \frac{x_i'\beta}{\sigma_u};\rho\right)] \prod_{y_i>0}\Phi\left(\frac{z_i'\gamma+\frac{\sigma_{uv}}{\sigma_u^2}(y_i-x_i'\beta)}{\sqrt{1-(\sigma_{uv}/\sigma_u)^2}}\right)\frac{1}{\sigma_u}\phi\left(\frac{y_i-x_i'\beta}{\sigma_u}\right)$$

- A simplification is assuming no correlation, so that

$$L(\beta,\gamma,\sigma_u) = \prod_{y_i=0}[1-\Phi(z_i'\gamma)\Phi(\frac{x_i'\beta}{\sigma_u})] \prod_{y_i>0}\Phi(z_i'\gamma)\frac{1}{\sigma_u}\phi\left(\frac{y_i-x_i'\beta}{\sigma_u}\right)$$

- **discuss:** scaling and identification problems (i.e. we still work with $\sigma_v = 1$)
- **discuss:** heteroscedasticity ?!

Diagnostic and Specification Tests in QV and LDV Models

- Lagrange Multiplier test
- The extension to QV models
- The extension to Tobit I models
- An alternative LR testing framework

Lagrange Multiplier Method

- The LM method is based on the restricted model.
  Imagine we maximize log-likelihood subject to the set of J *uncorrelated* constraints $c(\theta) - q = 0$, $\theta \in \boldsymbol{R}^k$:

$$max\ lnL^*(\theta) = lnL(\theta) - \lambda'\{c(\theta) - q\}, \lambda \in \boldsymbol{R}^J$$

$$\begin{aligned} \frac{\partial lnL^*(\theta_R)}{\partial \theta_R} &= \frac{\partial lnL(\theta_R)}{\partial \theta_R} - C'\lambda &= 0 \\ \frac{\partial lnL^*(\theta_R)}{\partial \lambda} &= c(\theta_R) - q &= 0 \end{aligned}$$

  C is $J \times k$ matrix of the derivatives of the constraints w.r.t. θ.
- If the restrictions R are valid, then imposing them will not lead to significant differences in the maximized value of the likelihood function. In particular, λ will be small. (vector of *shadow prices*)

Lagrange Multiplier Test

- If the restrictions are valid the log-likelihood evaluated at the restricted parameter vector will be approx. zero, i.e. (under say J constraints)

$$\hat{s}_R = \frac{\partial lnL(\widehat{\theta}_R)}{\partial \hat{\theta}_R} = \hat{\mathrm{C}}' \hat{\lambda} \underset{H_0}{\approx} 0$$

  The vector of first derivatives of the log-likelihood is the vector of the efficient scores s_R. Therefore, it is called **score test** as well as **Lagrange multiplier test**.

- The variance of the first derivative vector is the Fisher information matrix

$$LM = \left(\frac{\partial lnL(\hat{\theta}_R)}{\partial \hat{\theta}_R} \right)' [I(\hat{\theta}_R)]^{-1} \frac{\partial lnL(\hat{\theta}_R)}{\partial \hat{\theta}_R} \underset{H_0}{\sim} \chi^2_J \qquad (17)$$

  because $Var[\hat{s}_R] = E[\hat{s}_R^2] - E^2[\hat{s}_R] = I(\hat{\theta}_R) - 0$

Performing the Lagrange Multiplier Test

- Recall that we can estimate the variance via the product of the scores.
- The LM statistic has therefore a useful form: Let the $\hat{s}_{iR}$ denote the ith term in the gradient of the log-likelihood, and $\hat{S}_R$ be the matrix with n rows, the ith row equal to s_{iR} such that $\hat{S}_R'\mathbf{1} = \hat{s}_R$ for $\mathbf{1} = (1, \cdots, 1)' \in \boldsymbol{R}^n$. If we use the outer product of gradients to estimate the Hessian, then for J constraints

$$LM = \mathbf{1}'\hat{S}_R[\hat{S}_R'\hat{S}_R]^{-1}\hat{S}_R'\mathbf{1} = nR^2_{aux} \underset{H_0}{\sim} \chi^2_J$$

  with R^2_{aux} from an uncentered regression of 1s on the derivatives of the log-likelihood function (i.e. the scores) computed at the restricted estimator ($\hat{\theta}_R$).
- **Interpretation LM**: the higher the shadow prices, the less probable the restrictions

Extension to QV: Introduction

- Distributional assumptions have been used for the specification of the majority of QV and LDV models.
- However, we know that therefore ML estimation is pre-conditioned on the assumption of known distributions
- If the assumption is violated, one may derive wrong conclusions – they ought to be tested.
- Similarly, we would like to test for heteroskedasticity, omitted variables, or any other possible endogeneity.
- We will introduce score tests and LR tests in analogue to the LM testing idea discussed above

Recall ML concept - but modified for our purposes

- Let $y_i = m(x_i) + u_i$ and assume that $u_i \sim N(0, \sigma^2)$.
- Suppose we would like to test different assumptions being made implicitly in this model.
- We have a sample of data $\{y_i, x_i\}$ for $i = 1, ..., n$.
- The log likelihood function given k parameters $\theta = (\beta, \sigma)$ may be written as

$$l(\theta) = \sum_{i=1}^{n} \ln \left[\frac{1}{\sigma} * \phi \left(\frac{y_i - m(x_i)}{\sigma} \right) \right] = \sum_{i=1}^{n} l_i$$

- Solve for ML estimates $\tilde{\theta} = (\tilde{\beta}, \tilde{\sigma})$, using the following FOC, $\theta \in \boldsymbol{R}^k$:

$$\begin{aligned} \frac{\partial l(\tilde{\theta})}{\partial \theta} &= \sum_{i=i}^{n} \frac{\partial l_i(\tilde{\theta})}{\partial \theta} = \sum_{i=1}^{n} s_i(\tilde{\theta}) \\ &= \sum_{i=i}^{n} \left\{ \frac{\partial l_i(\tilde{\theta})}{\partial \theta_1}, \frac{\partial l_i(\tilde{\theta})}{\partial \theta_2}, ..., \frac{\partial l_i(\tilde{\theta})}{\partial \theta_k} \right\} = 0. \end{aligned}$$

Simple Example: a Linear Regression Function

- Suppose that $y = x_i'\beta + u_i$ and set $\epsilon_i = u_i/\sigma = (y_i - x_i'\beta)/\sigma$.
- Then,

$$l_i = \ln\left[\frac{1}{\sigma} * \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]$$

and consequently

$$\begin{aligned}
\frac{\partial l_i(\theta)}{\partial \beta} &= \sigma * \phi(.)^{-1} * \frac{1}{\sigma} * \frac{\partial[\phi(.)]}{\partial \beta_1} \\
&= -\phi(.)^{-1} * \phi(.) * \frac{y_i - x_i'\beta}{\sigma} * \frac{x_i}{\sigma} = \frac{-x_i\epsilon_i}{\sigma} \\
\frac{\partial l_i(\theta)}{\partial \sigma} &= \sigma * \phi(.)^{-1} * \left[\frac{1}{\sigma} * \frac{\partial[\phi(.)]}{\partial \sigma} - \phi(.)\frac{1}{\sigma^2}\right] = \frac{1}{\sigma} * \left[\epsilon_i^2 - 1\right].
\end{aligned}$$

are our score functions for the unrestricted case under normality

First order conditions for our linear regression example

- At Maximum Likelihood estimates $\tilde{\theta} = (\tilde{\beta}, \tilde{\sigma})$ the first order conditions require that (in terms of scaled residuals $\tilde{\epsilon}_i$)

$$n^{-1} \sum_{i=1}^{n} \tilde{\epsilon}_i = 0$$

$$n^{-1} \sum_{i=1}^{n} x_i \tilde{\epsilon}_i = 0$$

$$n^{-1} \sum_{i=1}^{n} \tilde{\epsilon}_i^2 - 1 = 0$$

  where $\tilde{\epsilon}_i = (y_i - x_i' \tilde{\beta}) * \tilde{\sigma}^{-1}$.

- As we know, normality is hard to test with these FOC as they correspond to the general moment conditions for linear models no matter what the error distribution is

The Lagrange Multiplier (LM) Framework - example continued

- Under the null, $y_i = x_i'\beta + u_i$, where $u_i \sim N(0, \sigma^2)$.
- Consider now the alternative, $y_i = x_i'\beta + z_i'\gamma + u_i$ for $\gamma \neq 0$
- Here, z_i represents a vector of additional regressors, and γ denotes the corresponding vector of parameters.
- now we introduce our modifications (which are actually only notational ones!!) because above we started from full model likelihood (above L, below L_A), and then reduced it to the restricted (above L_R, below -i.e."now"- L_0)
- denote the likelihood under the null as $L_0(\theta)$
- denote the likelihood under the alternatives as $L_A(\theta, \gamma)$ for parameters $\theta = (\beta, \sigma)$ and γ

Lagrange Multiplier (LM) Score Test - example continued

- if the null is correct, then $L_0(\theta) = L_A(\theta, 0) = L_R(\theta)$
- The score test approach involves:
  - maximise $L_0(\theta)$ to obtain $\hat{\theta}$; you may say $\hat{\theta}_R$ because obtained for $\gamma \equiv 0$
  - calculating $\frac{\partial ln\ L_A(\hat{\theta},\gamma)}{\partial \gamma}|_{\gamma=0} = \frac{\partial ln\ L_R(\hat{\theta})}{\partial \gamma}|_{\gamma=0}$
- LM test for $H_0 : \gamma = 0$: for $S_R := \left(\frac{\partial ln\ L_A(\hat{\theta},0)}{\partial \theta} \ , \ \frac{\partial ln\ L_A(\hat{\theta},0)}{\partial \gamma} \right)$

$$LM = \mathbf{1}' S_R (S_R' S_R)^{-1} S_R' \mathbf{1}$$

  should be distributed as chi-squared with $dim(z)$ degrees of freedom under the null, where $\mathbf{1} = (1, \ldots, 1)' \in \boldsymbol{R}^n$.
- Now discuss how/why this actually corresponds to the LM test (17)

Further LM applications in Practice

Testing for Normality

- Statistically, the characteristics of the normal distribution are well-known.
- In particular, if we let ϵ represent a standard normally distributed random variable, then we may list properties of the *moments* $E(\epsilon^j)$ as follows:

$$\begin{aligned} E(\epsilon^1) &= 0 \quad \text{(mean)} \\ E(\epsilon^2) - 1 &= 0 \quad \text{(variance)} \\ E(\epsilon^3) &= 0 \quad \text{(skewness)} \\ E(\epsilon^4) - 3 &= 0 \quad \text{(kurtosis)} \end{aligned}$$

- discuss why this is considered as a test for normality

Scores for different LM tests

A **Score Test for Normality** could be checking the first four moments

$$S_{iR} = \left[\frac{\partial l_i(\tilde{\beta}, \tilde{\sigma})}{\partial \tilde{\beta}'}, \frac{\partial l_i(\tilde{\beta}, \tilde{\sigma})}{\partial \tilde{\sigma}}, \tilde{\epsilon}_i^3, \tilde{\epsilon}_i^4 - 3 \right]$$

A **Score Test for Heteroscedasticity** could be

$$S_{iR} = \left[\frac{\partial l_i(\tilde{\beta}, \tilde{\sigma})}{\partial \tilde{\beta}'}, \frac{\partial l_i(\tilde{\beta}, \tilde{\sigma})}{\partial \tilde{\sigma}}, (\tilde{\epsilon}_i^2 - 1) * z_i' \right]$$

where you actually test whether z_i and $\tilde{\epsilon}_i$ are (un-)correlated

A **Score Test for Omitted Variables** is simply

$$S_{iR} = \left[\frac{\partial l_i(\tilde{\beta}, \tilde{\sigma})}{\partial \tilde{\beta}'}, \frac{\partial l_i(\tilde{\beta}, \tilde{\sigma})}{\partial \tilde{\sigma}}, \tilde{\epsilon}_i * z_i' \right]$$

Qualitative Variable Models – no natural residuals

- When confronted with QV or LDV models, score tests aren't so easy.
- Why? One is unable to generate *continuous* realisations of the underlying disturbances or residuals.
- Consider

$$y_i^* = x_i'\beta + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$, with an observability criterion of the form

$$y_i = \mathbb{1}(y_i^* > 0).$$

Given ML estimates $\tilde{\beta}$, but given the limited observability of y_i we can't solve for the residuals $\tilde{\epsilon}_i$.

- We produce 'estimates' of the residuals $\tilde{\epsilon}_i$ through knowledge of the fact that, for $y_i = 0, \epsilon_i < -x_i'\beta$.

Looking at Qualitative Variable Models

- For the "simple case" of binary responses we already discussed the problem of / with residuals.
- The above procedure could be a remedy. Let us consider the case of Probit models, i.e. assuming / working with normality
- For the ease of calculation will first split the two cases ($Y = 0/1$)
- Specifically, the *expected value* of ϵ_i conditional on $y_i = 0$ is

$$E(\epsilon_i|y_i = 0) = E(\epsilon_i|\epsilon_i < -x_i'\beta) = \frac{-\phi(x_i'\beta)}{1 - \Phi(x_i'\beta)}$$

- Similarly, still under normality and conditional on $y_i = 1$

$$E(\epsilon_i|y_i = 1) = E(\epsilon_i|\epsilon_i > -x_i'\beta) = \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)}$$

- it is straight forward to do the same for the Logit model
- or other qualitative (discrete) responses
- or for more complex "index" functions

Calculating Generalized Residuals

- Let us continue with the Probit:
- The *generalised error* $\epsilon_i^{(1)}$ may the therefore be defined as $E(\epsilon_i|y_i)$,

$$\begin{aligned}\epsilon_i^{(1)} &= E(\epsilon_i|y_i) \\ &= (1-y_i) * E(\epsilon_i|y_i=0) + y_i * E(\epsilon_i|y_i=1) \\ &= (1-y_i) * E(\epsilon_i|\epsilon_i < -x_i'\beta) + y_i * E(\epsilon_i|\epsilon_i > -x_i'\beta) \\ &= -(1-y_i) * \frac{\phi(x_i'\beta)}{1-\Phi(x_i'\beta)} + y_i * \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)}.\end{aligned}$$

- Generate generalised residuals $\widetilde{\epsilon}_i^{(1)}$ by replacing β by ML estimate
- For higher order moments may use the following recursion [for GAUSSIAN]

$$E[\epsilon^{(j)}|\epsilon > -a] = (j-1)E[\epsilon^{(j-2)}|\epsilon > -a] + (-a)^{j-1}\phi(a)/\Phi(a) \quad (18)$$

Calculating Generalized Residuals – continued

- Given our results for the Probit model above, with an observation **for which** $y_i = 1$
- we get thus for the higher order moments (to check variance, symmetry, kurtosis)

$$\epsilon_i^{(2)} = E(\epsilon_i^2|y_i) = 1 - (x_i'\beta) * \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)}$$

$$\epsilon_i{}^{(3)} = \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} * \left[2 + (x_i'\beta)^2\right]$$

$$\epsilon_i{}^{(4)} = 3 - \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} * \left[3x_i'\beta + (x_i'\beta)^3\right]$$

by using the recursion formula given in equation (18)

- similar formulae are / can be derived for other distributions

Calculating Generalized Residuals 3

- Similarly, **for** $y_i = 0$, we take advantage of a similar recursion formula for the GAUSSIAN

$$E[\epsilon^{(j)}|\epsilon < -a] = (j-1)E[\epsilon^{(j-2)}|\epsilon < -a] - (-a)^{j-1}\phi(a)/[1-\Phi(a)]$$

- This gives - after some tedious calculation -

$$\epsilon_i{}^{(1)} = E(\epsilon_i|y_i) = \frac{-\phi(x_i'\beta)}{1-\Phi(x_i'\beta)}$$

$$\epsilon_i{}^{(2)} = E(\epsilon_i^2|y_i) = 1 + (x_i'\beta) * \frac{\phi(x_i'\beta)}{1-\Phi(x_i'\beta)}$$

$$\epsilon_i{}^{(3)} = \frac{-\phi(x_i'\beta)}{1-\Phi(x_i'\beta)} * \left[2 + (x_i'\beta)^2\right]$$

$$\epsilon_i{}^{(4)} = 3 + \frac{\phi(x_i'\beta)}{1-\Phi(x_i'\beta)} * \left[3x_i'\beta + (x_i'\beta)^3\right]$$

Extension to Tobit I Models

Recall ML Estimation

$$\ln L(\beta, \sigma^2) = \sum_{y_i=0} \ln P(y_i = 0) + \sum_{y_i>0} [\ln f(y_i|y_i > 0) + \ln P(y_i > 0)]$$

for Gaussian errors $e_i \sim N(0, \sigma^2)$ this gave us

$$= \sum_{y_i=0} \ln \left[1 - \Phi(\frac{x_i'\beta}{\sigma})\right] + \sum_{y_i>0} \ln \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{-(y_i - x_i'\beta)^2}{2\sigma^2}\}\right]$$

Will again define generalised residuals via the first order conditions.

$$0 = \sum_{i=1}^{n} \tilde{\epsilon}_i^{(1)} x_i = \sum_{y_i=0} \frac{-\phi(x_i'\hat{\beta}/\hat{\sigma})}{1 - \Phi(x_i'\hat{\beta}/\hat{\sigma})} x_i + \sum_{y_i>0} \frac{\hat{e}_i}{\hat{\sigma}} x_i$$

$$0 = \sum_{i=1}^{n} (\tilde{\epsilon}_i^{(2)} - 1) = \sum_{y_i=0} \frac{x_i'\hat{\beta}}{\hat{\sigma}} \frac{\phi(x_i'\hat{\beta}/\hat{\sigma})}{1 - \Phi(x_i'\hat{\beta}/\hat{\sigma})} + \sum_{y_i>0} \frac{\hat{e}_i^2}{\hat{\sigma}^2} - 1$$

Testing for Omitted Variables and Heteroscedasticity

With these generalized residuals we can apply LM-testing idea:

Compare with binary response and linear models:

Testing for Omitted Variables

- want to test $H_0 : Z$ *insignificant* vs $H_1 : Z$ *significant*
- estimate model without Z and check $\sum_{i=1}^{n} \tilde{\epsilon}_i^{(1)} z_i = 0$

Testing for Homo- vs Heteroscedasticity

- For $V[e_i] = \sigma^2 h(z_i'\alpha)$, $h(0) = 1$ want to test $H_0 : h(z_i'\alpha) \equiv 1$
- estimate model and check $\sum_{i=1}^{n} (\tilde{\epsilon}_i^{(2)} - 1) z_i = 0$

Note that in both cases, for $z_i \in \boldsymbol{R}^J$, $nR^2_{aux} \overset{H_0}{\sim} \chi^2_J$, where R^2_{aux} from regressing 1's on $\tilde{\epsilon}_i^{(1)} x_i$, $(\tilde{\epsilon}_i^{(2)} - 1)$, and $\tilde{\epsilon}_i^{(1)} z_i$ or $(\tilde{\epsilon}_i^{(2)} - 1) z_i$

An alternative LR testing framework

Construction of score tests is sometimes complicated

$\Rightarrow$ alternative tests for non-normality, heteroskedasticity, omitted variables.

$\Rightarrow$ still using the Likelihood principle (as everything is distribution based here) we might think of *Likelihood Ratio* tests.

- Similar to regression tests of heteroskedasticity or serial correlation
- May be viewed as variable addition tests in which the significance of an extra set of regressors is tested by means of asymptotically valid ratio test statistics to resolve hypotheses under investigation.
- Advantage: Tests are often easier to implement than score test.

LR example 1: Testing for Non-Normality in "Probit" Models

- Estimate $y_i^* = x_i'\beta + \epsilon_i$ to obtain ML estimates $\hat{\beta}$ and maximised log-Likelihood $\log L_0$.
- Add test variables $(x_i'\hat{\beta})^2$ and $(x_i'\hat{\beta})^3$ to an auxiliary regression $y_i^* = x_i'\beta + \delta_1(x_i'\hat{\beta})^2 + \delta_2(x_i'\hat{\beta})^3 + \epsilon_i$.
- Obtain maximised log-Likelihood $\log L$ from the auxiliary regression.
- The test statistic $2(\log L - \log L_0)$ should be distributed as a χ^2_2 under the null of normality.
- You may alternatively say that you test for misspeficication of the index function, similar to the RESET test

LR example 2: Testing for Heteroskedasticity in Probit Models

- A test you find in the literature proposes to:
- Estimate $y_i^* = x_i'\beta + \epsilon_i$ to obtain ML estimates $\hat{\beta}$ and maximised log-Likelihood $\log L_0$.
- Add test variables $(x_i'\hat{\beta}) * z_i$ to an auxiliary regression $y_i^* = x_i'\beta + \delta_1(x_i'\hat{\beta}) * z_i + \epsilon_i$ where z_i represents an m-vector of variables which may potentially cause the heteroskedasticity.
- Obtain maximised log-Likelihood $\log L$ from the auxiliary regression.
- The test statistic $2(\log L - \log L_0)$ should be distributed as a χ^2_m under the null of homoskedasticity.
- You actually test for a very specific interaction which is here interpreted as a scedasticity function linear in z_i

LR example 3: Testing for Omitted Variables in Probit Models

- Estimate $y_i^* = x_i'\beta + \epsilon_i$ to obtain ML estimates $\hat{\beta}$ and maximised log-Likelihood $\log L_0$.
- Add test variables z_i to an auxiliary regression $y_i^* = x_i'\beta + \delta_1 z_i + \epsilon_i$ where z_i represents an m-vector of potentially omitted variables.
- Obtain maximised log-Likelihood $\log L$ from the auxiliary regression.
- The test statistic $2(\log L - \log L_0)$ should be distributed like a χ^2_m under the null of no incorrect omission.

discuss further testing problems ...