Advanced Microeconometrics

course notes

Kristian Olesen Larsen*

University of Copenhagen

December 29, 2017

Abstract

These notes cover in brief the contents of the course advanced microeconometrics taught at the UCPH department of economics in the fall of 2017. The course covers topics in frequentist and bayesian estimation, including the probit/logit and tobit model, as well as nonparametric methods and various simulated estimators.

These course notes are written as part of my personal studies for an exam, and should be taken only to reflect my understanding of the topic. I cannot guarantee that everything in these notes is correct, much less that the explanations provided here are better than those that others have already provided. To fully understand the topics covered i suggest following a course in microeconometrics yourself.

Take note that the mathematics used are inherently multidimensional, and as such most expressions should be interpreted as vectors and matrices to properly understand the topic. The dimensions of objects are often omitted due to space concerns.

Contents

*kuol@econ.ku.dk. Please share these notes with as many people as you feel like.

I. Extremum estimators - a general framework for frequentist estimation

Frequentist estimators are those most often encountered in undergraduate studies including basic statistics. The term frequentist hints at the estimators utilizing the frequency with which observations appear to estimate underlying models. Extremum estimation is a generalized framework to study these estimators. Common for all extremum estimators is a *stochastic criterion function* $Q_N(\theta)$ which is to be minimized, as well as data $w_i = \{y_i, x_i'\}_{i=1}^N$. Using these ingredients, as well as a set of assumptions (i.e. $E[\epsilon|x] = 0$) and a parametric representation of the model. An extremum estimator is formally then an estimator which solves

$$\hat{\theta}_E = \arg\min_{\theta} Q_N(\theta) \tag{1}$$

In these notes all of the frequentist estimators studied will be asymptotically normal, with limit distribution (more on this later)

$$\hat{\theta}_N \overset{a}{\sim} \mathcal{N}(\theta_0, A_0^{-1} B_0 A_0^{-1}) \tag{2}$$

Usually the two matrices A_0 and B_0 can be estimated by the hessian and the outer product of the gradients respectively. It is the criterion function $Q_N(\cdot)$ that essentially defines the properties of an estimator. A subgroup of extremum estimators is the M estimators, defined by solving the problem

$$\hat{\theta}_N = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} q(w_i, \theta) \tag{3}$$

In other words M-estimators are extremum estimators where $Q_N(\theta) = \frac{1}{N}\sum_{i=1}^N q(w_i, \theta)$. The advantage to working with M-estimators compared to the more general extremum estimators, is that they allow for the application of a law of large number and the central limits theorem.

i. Identification

Identification essentially ensures that one can estimate a unique solution for each of the model parameters. Mathematically θ_1 is identified if $P_{\theta_1}(w) = P_{\theta_2}(w), \forall w \in \mathcal{W}$ iff $\theta_1 = \theta_2$. In the case of extremum estimators this is equivalent with the requirement that

$$Q(\theta_0) < Q(\theta), \quad \forall \theta \in \Theta : \theta \neq \theta_0 \tag{4}$$

In the case of M-estimators this can be reduced to

$$E[q(w_i, \theta_0)] < E[q(w_i, \theta)], \quad \forall \theta \in \Theta : \theta \neq \theta_0$$

ii. Consistency

Assuming the specified model is correct compared to the underlying datagenerating process (DGP), and thus is identified we can derive consistency on the basis of the criterion function. Consistency is the notion that for large values of the sample size N the estimator should converge in probability to θ_0, formally a convergent estimator $\hat{\theta}_N$ has the property that

$$\text{plim}\ \hat{\theta}_N = \theta_0 \tag{5}$$

or equivalently

$$\lim_{N\to\infty} \Pr\left(|\hat{\theta}_N - \theta_0| > \epsilon\right) = 0, \quad \forall \epsilon > 0 \tag{6}$$

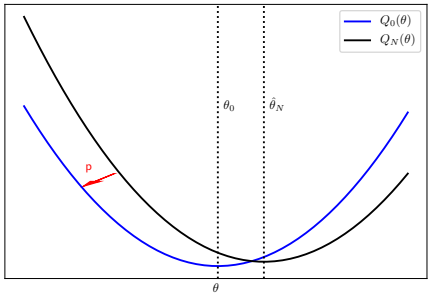
In simple cases, such as for OLS we know how to express the model in simple terms that allow for the derivation of consistency, but in the general case we know nothing else than the criterion function $Q_N(\cdot)$. The concept is illustrated in figure 1.

We can think of $Q_N(\cdot)$ as the sample criterion function, which when minimized yields a sample parameter estimate $\hat{\theta}_0$, and likelise the population criterion $Q_0(\cdot)$ yields the true parameter θ_0. The goal in showing consistency is to show that the sample criterion converges in probability to $Q_0(\cdot)$. For M-estimators applying a law of large numbers gives that

$$\frac{1}{N}\sum_{i=1}^{N} q(w_i, \theta) \underset{N\to\infty}{\to} E[q(w_i, \theta)] \equiv Q_0(\theta) \tag{7}$$

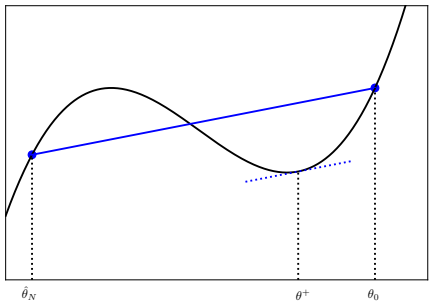
but this is not possible in the case of extremum estimators. Instead, c.f theorem 5.1 of C&T the following assumptions will ensure that an estimator $\hat{\theta}_N$ is consistent:

Figure 1: *Concept of consistency in extremum estimators*



As N increases consistency will mean for the criterion function $Q_N(\theta)$ to converge in probability to the actual criterion function $Q_0(\theta)$, thus yielding a consistent estimate of $\hat{\theta}_N$.

Figure 2: *Mean Value Theorem illustration*



The mean value theorem essentially tells us that any continous function $f(z)$ over the interval $\mathcal{I}$ will have at least one point in $\mathcal{I}$ where $\frac{\partial f(z)}{\partial z}$ equals the slope between the endpoints of $\mathcal{I}$.

- The parameter space Θ is a compact subset of $\mathbb{R}^q$, compact meaning closed an bounded. (Note that $\mathbb{R}$ is not compact, making this asusmptions untrue for most estimation methods in practise).
- $Q_N(\theta)$ is measurable and continous for all $\theta \in \Theta$.
- $Q_N(\theta)$ converges uniformly in probability to a nonstochastic function $Q_0(\theta)$, which has a unique global maximum at θ_0.

Assuming continuity and measurability of $Q_N(\theta)$ and compactness of Θ allows to invoke the extreme value theorem, which simply guarantees that a minimum and amaximum of $Q_N(\cdot)$ over a closed interval $[a, b]$ will exist.

iii. Asymptotic distribution

To derive the limit distribution of $\hat{\theta}_N$ we will invoke the *mean value theorem*, which has a very intuitive graphical presentation shown in figure 2. Mathematically we simply equate the derivative of a function, measured in a point x^+ with the slope of a line connecting two points $[x_0, x_1]$. To use the theorem it is required that $Q_N(\cdot)$ is twice differentiable around θ_0. Applying the mean value theorem on $\frac{\partial Q_N(\theta)}{\partial \theta}$ on the interval $[\hat{\theta}_N, \theta_0]$ will yield

$$\left.\frac{\partial Q_N(\theta)}{\partial \theta}\right|_{\hat{\theta}_N} = \left.\frac{\partial Q_N(\theta)}{\partial \theta}\right|_{\theta_0} + \left.\frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'}\right|_{\theta^+} (\hat{\theta}_N - \theta_0) \quad (8)$$

Where by definition $\left.\frac{\partial Q_N(\theta)}{\partial \theta}\right|_{\hat{\theta}_N} = 0$ so rearranging leaves us with

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \left(\left.\frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'}\right|_{\theta^+}\right)^{-1} \times \left(\left.\frac{\partial Q_N(\theta)}{\partial \theta}\right|_{\theta_0} \sqrt{N}\right) \quad (9)$$

It is then assumed that the following matrices exist and, that the terms converge towards

$$\left.\frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'}\right|_{\theta^+} \xrightarrow{p} A_0$$
$$\left.\frac{\partial Q_N(\theta)}{\partial \theta}\right|_{\theta_0} \sqrt{N} \xrightarrow{d} \mathcal{N}(0, B_0)$$

If A_0 and B_0 exist, we can then see from the expression in (9) that asymptotically a consistent M estimator $\hat{\theta}_N$ has distribution

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}(0, A_0^{-1} B_0 A_0^{-1}) \quad (10)$$

In practise of course A_0 and B_0 are unknown, and will have to be estimated, usually with the following empirical matrices

$$\hat{A} = \frac{\partial^2 Q_N(\theta)}{\partial\theta\partial\theta'}\bigg|_{\hat{\theta}}$$
$$\hat{B} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial q(w_i,\theta)}{\partial\theta}\bigg|_{\hat{\theta}}\frac{\partial q(w_i,\theta)}{\partial\theta'}\bigg|_{\hat{\theta}}$$

iii.1 Maximum likelihood estimators

Maximum likelihood estimators has the property that $Q_N(\theta)$ is specified so that

$$Q_N(\theta) = -\sum_{i=1}^{N}\log L_i(\theta|x_i,y_i) \tag{11}$$

Because of this the sandwich expression for the variance derived above collapses to an even simpler expression where $\hat{\theta}_{ML} \sim \mathcal{N}(0,[\mathcal{I}(\theta)]^{-1})$ where $\mathcal{I}(\theta)$ is the fischer information, defined as

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2\log L_i(\theta|x_i,y_i)}{\partial\theta\partial\theta'}\right] \tag{12}$$

The proof of this is as follows

Use that $\log L_i(\theta|x_i,y_i) \propto f(y_i|\theta)$, that is the likelihood is proportional to the density of the data. Starting here, we know that

$$\int_{\mathbb{R}} f(y_i|\theta) = 1 \tag{13}$$

and since the bound of the integral does not depend on θ taking the derivative w.r.t it on both sides gives

$$\int_{\mathbb{R}}\frac{\partial}{\partial\theta}f(y_i|\theta) = 0 \tag{14}$$

Now apply that $\frac{\partial\log f(x)}{\partial x} = \frac{\partial f(x)}{\partial x}[f(x)]^{-1}$ to get

$$\int_{\mathbb{R}}\frac{\partial\log f(y_i|\theta)}{\partial\theta}f(y_i|\theta) = 0 \tag{15}$$

Using again the rewrite of the derivative of the log of a function gives

$$\int_{\mathbb{R}}\frac{\partial}{\partial\theta}\frac{\partial\log f(y_i|\theta)}{\partial\theta}f(y_i|\theta) = 0 \tag{16}$$

Which when written out gives

$$\int_{\mathbb{R}}\frac{\partial\log f(y_i|\theta)}{\partial\theta}\frac{\partial f(y_i|\theta)}{\partial\theta} + \frac{\partial^2\log f(y_i|\theta)}{\partial\theta\partial\theta'}f(y_i|\theta) = 0$$

Using the log-derivative trick one final time gives the desired result

$$\int_{\mathbb{R}}\frac{\partial f(y_i|\theta)}{\partial\theta}\frac{\partial f(y_i|\theta)}{\partial\theta}f(y_i|\theta) + \frac{\partial^2\log f(y_i|\theta)}{\partial\theta\partial\theta'}f(y_i|\theta) = 0$$

Or written in terms of expectations

$$E\left[\frac{\partial f(y_i|\theta)}{\partial\theta}\frac{\partial f(y_i|\theta)}{\partial\theta}\right] = -E\left[\frac{\partial^2\log f(y_i|\theta)}{\partial\theta\partial\theta'}\right] \tag{17}$$

Comparing this expression to the $A_0^{-1}B_0A_0^{-1}$ expression derived above should make it clear why the asymptotic variance of maximum likelihood estimators simplifies to the fischer information.

II. Numerical optimization

The motivation behind numerical optimization is simple: often we come across functions that are difficult to optimize analytically, and in any case computers are not very good at analytical math. Instead of solving problems exactly, we would like to have methods which come close to the true solution without being to computationally intensive.

In the lectures a host of topics including steepest descent optimization and non-gradient based methods were covered, but here we'll only give a brief overview of the Newton Rhapson algorithm.

i. The Newton Rhapson algorithm

The idea of the Newton Rhapson algorithm is to intiate the maximization of a function $f(x)$ at some starting guess x_0, where a second order Taylor polynomial $p(x_0)$ is constructed. The next point in the procedure will then be the $x = x^1$ where $p(x_1)$ is minimized. The

algorithm is generalized to the multivariate case, but the intuition nonetheless remains the same. Let $x_o \in \mathbb{R}^p$ then the second order Taylor approximation of $f(\cdot)$ around x_0 is

$$p(x) = f(x_0) + \nabla f(x_0)(x - x_0) + (x - x_0)' \frac{\nabla^2 f(x_0)}{2}(x - x_0) \quad (18)$$

for which there is naturally only one solution when solving $p'(x) = 0$, which is the one where

$$x = x_0 - [\nabla^2 f(x_0)]^{-1} \nabla f(x_0) \quad (19)$$

This equation gives the next point x to approximate as a function of the current point of approximation x_0. It should be noted that the second term is essentially *slope over curvature*. Generalizing the above equation to an iterative algorithm leads to the Newton Rhapson equation

$$x_{n+1} = x_n - s_n [\nabla^2 f(x_n)]^{-1} \nabla f(x_n) \quad (20)$$

where s_n is some arbitrarily chosen step size, which ensures the algorithm doesn't get stuck in a steady state, where it jumps back and forth between two suboptimal points.

i.1 Berndt-Hall-Hall-Hausman (BHHH)

Computationally estimating the hessian at every step might be to costly, so as an alternative the BHHH algorithm utilizes the results derived for ML estimators above, and estimates the hessian as a product of gradients, that is

$$\frac{\partial^2 Q_N(x)}{\partial x \partial x'} \approx \sum_{i=1}^{N} \frac{\partial q_i(x)}{\partial x} \frac{\partial q_i(x)}{\partial x'} \quad (21)$$

with this procedure it is only neccesary to estimate the gradient, which is much less costly than computing the hessian. On the other hand, if the model is wrongly specified, or the sample is small BHHH will perform poorly.

III. Nonlinear least squares

Nonlinear least squares is, as the name suggests similar to OLS, but with the added ability to handle nonlinear relations. Before jumping to use NLS, it should however be considered if the relationship of interest is nonlinear even when considering variable transformations. If this not is the case, NLS might be the right choice of estimator. We begin by defining the model, noting that NLS is a M-estimator, due to the form of it's criterion function.

$$y_i = g(x_i, \beta) + u_i, \qquad E[y_i | x_i] = g(x_i, \beta) \quad (22)$$

This is very similiar to OLS, but with the addition of a link function $g(\cdot)$, which alters the estimation problem to

$$\hat{\beta}_{NLS} = \arg\max_{\beta} \sum_{i=1}^{N} (y_i - g(x_i, \beta))^2 \quad (23)$$

Taking the first order conditions is straight forward from here, simply solve $\frac{\partial Q_N(\beta)}{\partial \beta} = 0$. Results on the asymptotic distribution of $\hat{\beta}_{NLS}$ follows from theory on M-estimators.

It can be shown that NLS will be consistent, but not nessecarily efficient if $E[u_i | x_i] = 0$, as this assumption alone doesn't account for heteroscedasticity. To account for this, the idea is to implement a weight in the optimization problem. We can show that the optimal weight is the actual variance-covariance matrix Ω_0

$$V[y_i | x_i] = V[u_i | x_i] = E[uu' | x] = \Omega_0 \quad (24)$$

Of course Ω_0 is unknown, so instead of using this, we begin by simply guessing a matrix Σ and estimating $\hat{\beta}_{WNLS}$ as

$$\arg\min_{\beta} (y - g(x, \beta))' \Sigma^{-1} (y - g(x, \beta)) \quad (25)$$

One way to select Σ is then to assume that it depends on a set of parameters γ and estimate $\hat{\Sigma} = \Sigma(\hat{\gamma})$. In practise this is implemented in steps

1. Do regular NLS to estimate $\hat{\beta}_{NLS}$
2. compute the residuals $\hat{u}_i = y_i - g(x_i, \hat{\beta}_{NLS})$ and regress their square on a guessed variance matrix structure and covariates γ to get $\hat{\Sigma}$.

3. Implement the estimated $\hat{\Sigma}$ and conpute the WNLS estimator.

This procedure can theoretically be iterated over multiple times, each time refining the estimate of $\hat{\gamma}$ and thus of $\hat{\Sigma}$.

IV. Binary response models - probit and logit

Binary response models are designed to be useful whenever the outcome of interest can take only one of two values, i.e. $y_i \in \{0, 1\}$, where $y_i = 0$ happens with probability $1 - p_i$ and conversely $\Pr(y_i = 1) = p_i$. To include observable characteristics we parametrize the probability by a linear relation of dependents and a link function $F(\cdot)$ such that

$$p_i \equiv \Pr(y_i = 1|x_i) = F(x_i'\beta) \tag{26}$$

Since $F(\cdot)$ must be confined in the interval $[0, 1]$ and represents a probability, we will usually select $F(\cdot)$ to be a CDF. Common choices are the logistic CDF $\Lambda(\cdot)$ and the normal CDF $\Phi(\cdot)$.

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp z}$$
$$\Phi(z) = \int_{-\infty}^{z} \phi(w)\mathrm{d}w$$

with $\phi(\cdot)$ begin the density of a gausian variable with mean 0 and variance 1.

i. Maximum Likelihood estimation

Notice that conditional on x_i the outcome of y_i is essentially a bernoulli trial, so the density of y_i will be

$$f(y_i|\theta) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{27}$$

Inserting the link function and assuming independence of observations allow us to construct a full sample likelihood given by

$$L_N(\theta) = \prod_{i=1}^{N} F(x_i'\beta)^{y_i}(1 - F(x_i'\beta))^{1-y_i}$$

The log likelihood function is then easily derived by taking the log, which transforms the product into a sum, draws down the exponents,

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log L_i(\theta|x_i, y_i) \tag{28}$$

From here it's possible to derive the scores $s_N(\theta)$ by taking the first derivative, and with a lot of work the hessian is also analytically derivable. The equation system $s_N(\theta) = 0$ is however not analytically solveable but requires numerical methods.

ii. Latent variable representation

An alternative to the derivations above is to assume the existence of a latent, unobservable variable. This variable y^* acts as an underlying continous variable, which indirectly determines y through a cutoff value, that is

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & y_i^* \leq 0 \end{cases} \tag{29}$$

Where we assume y_i* to be linear, i.e.

$$y_i^* = x_i'\beta + \epsilon_i \tag{30}$$

The interpretation of y_i^* can be as a utility variable, where individuals gaining more than $y_i^* = 0$ utility will make a certain choice, which we record as $y_i = 1$. Now instead of assuming a certain link function, we will make a distributional assumption on the error term ϵ_i. In the following we assume that $\epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ to derive the probit model. First of calculate the probability of observing $y_i = 1$

$$\begin{aligned} p_i \equiv \mathrm{P}(y_i = 1|x_i) &= \mathrm{P}(y_i^* > 0|x_i) \\ &= (P)(x_i'\beta + \epsilon_i > 0|x_i) \\ &= 1 - \mathrm{P}(\epsilon_i \leq -x_i'\beta|x_i) \end{aligned}$$

To finalize this we first need to normalize ϵ_i to be drawn from a standard normal distribution

$$p_i = 1 - \mathrm{P}\left(\frac{\epsilon_i - \mu}{\sigma} \leq -\frac{x_i'\beta + \mu}{\sigma}|x_i\right)$$

Which then is rewriteable into the final expression, by using that ϵ_i is normally distributed,

and that this distribution is symmetric meaning $1 - \Phi(-z) = \Phi(z)$

$$p_i = \Phi\left(\frac{x_i'\beta + \mu}{\sigma}\right) \tag{31}$$

This p_i is the same one defined and parametrized as $F(\cdot)$ in the non-latent interpretation, showing us that whether we choose a latent-variable aproach to the model or not, we end up with the same conclusions. From the expression of p_i it's clear that neither β or μ are going to be present on their own in the likelihood function, as they are always divided by σ. Thus identification requires fixing σ. Usually we set $\sigma = 1$ to get the simplest expression of β's $\Phi(x_i'\beta + \mu)$ with $\mu = 0$ usually also implemented as the intercept of the model is otherwise not identified. The logit model has a natural scale of $\mu = 0$ and $\sigma^2 = \pi^2/3$, so results between the two models are not directly comparable.

iii. Marginal effects

The regression coefficients in the probit and logit model are uninformative as they depend on the scale of errors, and are not in any simple way related to the observed y_i's. Instead one may study the marginal effects of changing a covariate x_{ij} on the probability that $y = 1$, that is

$$\frac{\partial \Pr(y_i = 1|x_i)}{\partial x_{ij}} = F'(x_i'\beta)\beta_j \tag{32}$$

Often this partial derivative is denoted $\delta_j(x_i)$. In the case of dummy variables, δ_j is simply calculated as the difference in probability with the dummy on and off, keeping all other covariates constant. Marginal effects depend nonlinearly on x_i, and thus change depending on the observation for which they are calculated. In practise there are a number of ways to handle this. Some researchers choose a specific individual i (i.e. the median individual) and compute marginal effects only for this observation, or for the 'mean' individual where $\bar{x} = \frac{1}{N}\sum_i x_i$ is used as covariates. Others compute marginal effects for all observations and report their average.

iii.1 The Delta method

One issue with the marginal effects derived above is that while they depend on β, and so in practice also on $\hat{\beta}$, we cannot derive an expression for their standard errors. To get an estimate of these one option is the Delta method, which in general derives approximate standard errors for function of variables with known errors. An alternative approach is to bootstrap sample from the dataset, which is covered later in this note. In the most general case the delta method can be used to show that

$$V[h(\hat{\beta})] \approx \frac{\partial h(\hat{\beta})}{\partial \beta'} V[\hat{\beta}] \frac{\partial h(\hat{\beta})}{\partial \beta} \tag{33}$$

where $h(\cdot)$ is any function of β.

In the case of the probit model begin by noting that the the marginal effects are rewriteable as

$$\delta_k(x_i) = F'(x_i'\beta)|_{x_i = x^0}\beta_k = g(x^0\beta)\beta_k \tag{34}$$

where $g(\cdot)$ is simply shorthand for F' evaluated in x^0. δ_k is a function of parameters which we will denote $d_k(\beta)$ and the estimates partial effects are $\hat{\delta}_k = d_k(\hat{\beta})$. Stacking these in a column vector $\mathbf{d}(\hat{\beta})$ gives is a $K \times 1$ column, for which we can do the following Taylor expansion around β

$$\mathbf{d}(\hat{\beta}) \approx \mathbf{d}(\beta) + \nabla_\beta \mathbf{d}(\beta)(\beta - \hat{\beta}) \tag{35}$$

Rewriting that $\mathbf{d}(\hat{\beta}) = \hat{\delta}$ we get that

$$\hat{\delta} - \delta \approx \nabla_\beta \mathbf{d}(\beta)(\beta - \hat{\beta}) \tag{36}$$

Multiplying by $\sqrt{N}$ on both sides, and noting that $\hat{\beta} \sim \mathcal{N}(0, V[\beta])$ we get that

$$\hat{\delta} \overset{a}{\sim} \mathcal{N}(0, [\nabla_\beta \mathbf{d}(\beta)] V[\beta] [\nabla_\beta \mathbf{d}(\beta)]') \tag{37}$$

Now we dont know $\nabla_\beta \mathbf{d}(\beta)$ yet, but it is easily derived. Use the definition of δ_k from above and we have that

$$\begin{aligned}
\nabla_\beta \mathbf{d}(\beta) &= \frac{\partial}{\partial \beta} g(x^0\beta)\beta \\
&= \beta \frac{\partial}{\partial \beta} [g(x_0\beta)] + I_k g(x^0\beta) \\
&= \beta g'(x^0\beta)x^0 + I_k g(x^0)
\end{aligned}$$

where I_k is the identity matrix. This is as far as we can get without further assumptions, but in the probit case we can utilize that $g(\cdot)$ is the normal pdf $\phi(\cdot)$, for which it holds that $\phi'(z) = -z\phi(z)$[1]. When this is the case simply inserting $\phi'(x^0\beta) = -(x^0\beta)'\phi(x^0\beta)'$ and factoring out will give the desired result shown in the lecture.

V. Censored regression models

The Tobit model is designed to handle data with unnatural cutoffs, either due to censoring or traits of the DGP. An important distinction to make is between *censoring*, which keeps all observations, but censors certain of them to an arbitrary value and *truncation* which remove all observations at or below a certain threshold from the dataset. The Tobit model is formulated in terms of a latent variable y^*, so to begin with let

$$y_i^* = x_i'\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{38}$$

The observational rule is slightly modified compared to the binary case, as we now have $y_i = \max\{0, y_i^*\}$.

i. Maximum Likelihood estimation

In the case of the Tobit model the likelihood is multiplicatively separable into the likelihood from observations where $y_i = 0$ and the likelihood from observations where $y_i > 0$, that is

$$f(y_i|x_i) = \begin{cases} f_0(y_i|x_i), & \text{if } y_i = 0 \\ f_1(y_i|x_i), & \text{if } y_i > 0 \end{cases} \tag{39}$$

Which, as will later come in handy, can be condensed to a product of factors raised to indicators for the case. In the face of $f_0(\cdot)$ this is equal to the probability that $y_i = 0$. Following the same steps as in the probit example, inserting y_i^* and isolating ϵ_i will provide that

$$f_0(y_i|x_i) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right) \tag{40}$$

On the other hand when $y_i > 0$ we know that we're observing the actual latent variable y^*, and from the definition of this variable it's clear that $y_i^*|x_i \sim \mathcal{N}(x_i'\beta, \sigma^2)$. Thus[2]

$$f_1(y_i|x_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) \tag{41}$$

Now these expression can be joined as follows

$$f(y_i|x_i) = \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{\mathbb{1}_{(y_i=0)}} \times \left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]^{\mathbb{1}_{y_i>0}} \tag{42}$$

The indvidual (log) likelihood contributions as well as the overall (log) likelihood function are then easy to derive. The score is difficult to find, and the first order condition has no analytical solution, so numerical methods must be used. Notice there are a number of ways to answer questions about the expected value of our model, specifically we might be interested in $E[y^*|x]$, $E[y|x]$ or $E[y|y > 0, x]$. The first one is simple, $E[y^*|x] = x'\beta$, but the other two are slightly less straight forward. In all cases the same method is applied, namely substituting in $y^* = x'\beta + \epsilon$ and isolating stochastic terms. For example

$$\begin{aligned} E[y|y > 0, x] &= E[x'\beta + \epsilon | x'\beta + \epsilon > 0, x] \\ &= x'\beta + E[\epsilon | \epsilon > -x'\beta, x] \\ &= x'\beta + \sigma E\left[\frac{\epsilon}{\sigma} | \frac{\epsilon}{\sigma} > \frac{x'\beta}{\sigma}, x\right] \\ &= x'\beta + \sigma\lambda\left(\frac{x'\beta}{\sigma}\right) \end{aligned} \tag{43}$$

Where a number of tricks are used. In line 3 we multiply and divide by σ to get a standard normal in the expectation, and below in line 4 that for any $X \sim \mathcal{N}(\mu, \sigma)$ we have that the inverse Mills ratio $\lambda(z)$ is $E[X|X > \alpha] = \frac{\phi(z)}{1-\Phi(z)}$

[1] **Proof**: let $\phi(z) = (2\pi)^{-1/2}\exp(\frac{-z^2}{2})$, then by differentiating $\phi'(z) = (2\pi)^{-1/2}\exp(\frac{-z^2}{2}) \cdot (\frac{-2}{2} \cdot z)$ which clearly reduces to $\phi'(z) = -z\phi(z)$

[2] the expression of f_1 is simply a clever way to write a non-standard normal distribution without having to write the entire pdf. Be aware that $\phi(z)$ is always the pdf of a standard normal.

where $z = \frac{\alpha-\mu}{\sigma}$. To derive $E[y|x]$ first notice that this must be $\Pr(y = 0) \cdot 0 + \Pr(y > 0) \cdot E[y|y > 0, x]$ and $\Pr(y > 0)$ is

$$\begin{aligned}\Pr(y > 0) &= \Pr(x'\beta + \epsilon > 0) \\ &= \Pr(\epsilon > -x'\beta) \\ &= 1 - \Pr\left(\frac{\epsilon}{\sigma} \leq \frac{-x'\beta}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{-x'\beta}{\sigma}\right) \\ &= \Phi\left(\frac{x'\beta}{\sigma}\right)\end{aligned} \tag{44}$$

Since $\lambda(z) \equiv \frac{\phi(z)}{\Phi(z)}$ is is then clear how

$$E[y|x] = x'\beta\Phi\left(\frac{x'\beta}{\sigma}\right) + \sigma\phi\left(\frac{x'\beta}{\sigma}\right) \tag{45}$$

Marginal effects are ususally calculated on these quantities, and should be straight forward to derive, simply by calculating $\frac{\partial}{\partial x}$ in one of the above expectations or probabilities.

VI. Multinomial Logit models

Multinomial choice models are models concerned with ordered or unordered choices, i.e. the choice of education, which car to buy or similar. The common trait of all problems which can be modelled by multinomial models is the discrete choice levels available. The following models are all concerned with unordered choices, (i.e. not education, which is a 'ladder-choice'). Our goal is to specify and estimate a utility function $u : (x_{ij}, \theta) \mapsto \mathbb{R}$ and estimate this mapping such that $\forall j, \theta' : u(x_{ij}, \theta) > u(x_{ij}, \theta')$. In the following we will study three variations of the model class *Random Utility Models* (RAM), which is generally defined by the following construct of u_i

$$y_i = \underset{j\in\{1,2,\ldots,J\}}{\arg\max}\, u_{ij} \tag{46}$$

where $u_{ij} = v_{ij} + \epsilon ij$ consists of a deterministic/observable term v and a stochastic/unobservable term ϵ. The parametrization of v will determine the type of RAM estimated amongst the conditional- (CL) multinnomial (MNL) and mixed logit models.

CL: $v_{ij} = x_{ij}\beta$
MNL: $v_{ij} = x_i\beta_j$
Mixed: $v_{ij} = x_{ij}\beta + w_i\gamma_j$

In the conditional logit x will contain information on the alternatives in J, while in the multinomial it will contain information on the individual i. To determine if all models are identified we need to first specify the likelihood function.

i. Maximum Likelihood estimation

Each individual makes only one choice, and so each individual will only contribute likelihood for this choice. To capture this idea define $d_{ij} = \mathbb{1}_{(y_i=j)}$, which will be 1 for individual i only if individual i made choice j, and 0 otherwise. To ensure that each individual cannot choose multiple values in J we will require

$$\forall i : \sum_{j=1}^{J} d_{ij} = 1 \tag{47}$$

With this notation we can define the probability that i chooses j as

$$p_{ij} \equiv \Pr(d_{ij} = 1, d_{ik} = 0\ \forall k \neq j) \tag{48}$$

The individuals likelihood contribution will then be the product of these p_{ij}'s, but only counting the one for the exact value of j what individual i actually chose, so

$$\begin{aligned}\ell_i(\theta) &= \prod_{j=1}^{J} p_{ij}^{d_{ij}} \\ \log \ell(\theta) &= \sum_{j=1}^{J} d_{ij} \log(p_{ij})\end{aligned} \tag{49}$$

i.1 Specifying p_{ij}

To finish we need an expression for p_{ij}, it can be shown that it's distribution will depend on the difference $\epsilon_{ij} - \epsilon_{ik}$ and further that assuming $\epsilon \sim$ Gumbel will result in the difference of two epsilons being logistically distributed. Thus we assume

$$\begin{aligned}F(\epsilon) &= \exp(-\exp(-\epsilon)) \\ f(\epsilon) &= \exp(-\epsilon - \exp(-\epsilon))\end{aligned} \tag{50}$$

This assumption implies that

$$p_{ij} = \frac{\exp(v_{ij})}{\sum_{k=1}^{J} \exp(v_{ik})} \tag{51}$$

where v_{ij} will vary depending on the model specification.

With this information we can complete the likelihood, by taking the product over the individual likelihood contributions, or equivalent in log-likelihood terms

$$\log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \log \left(\frac{\exp(v_{ij})}{\sum_{k=1}^{J} \exp(v_{ik})} \right)$$

By inserting the relevant specification of v_{ij} and rewriting the log-term this expression simplifies slightly.

ii. Identification

The derived expression of p_{ij} implies specific requirements must be fulfilled to properly identify parameters in the model. What requirements exactly will depend on the specification of v_{ij}. In the conditional logit model adding an intercept β_0 to the model is problematic since

$$\frac{\beta_0 + \exp(v_{ij})}{\sum_{k=1}^{J} \exp(\beta_0 + v_{ik})} = \frac{\exp(v_{ij})}{\sum_{k=1}^{J} \exp(v_{ik})} \tag{52}$$

so to identify the CL model the intercept needs to be fixed to 0. In the MNL model a constant δ can be added to β_j in a similar fashion. The solution is then to fix the β_j of one alternative as 1, making the specific alternative a baseline, against which other β's should be compared. Naturally in the mixed model, both problems are present, and both restrictions need to be implemented.

iii. Marginal effects

Marginal effects will depend on the model specification, beginning with the conditional logit, we have to consider two cases, namely that we take the derivative w.r.t the x such that $j = k$ as well as the alaternative $j \neq k$. Beginning with the case of $j = k$ we calculate

$$\begin{aligned} \frac{\partial \Pr(y_i = j|x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \frac{\exp(x_{ij}\beta)}{\sum_{k=1}^{J} \exp(x_{ik}\beta)} \\ &= p_{ij}(1 - p_{ik})\beta \end{aligned} \tag{53}$$

To get this result simply apply the derivative-of-a-fraction rule naively and compare the nominator and denominator in the resulting expression. You should be able to rewrite parts of the derivative as p_{ik} and p_{ij}'s. The only tricky part is that although $j = k$ we need to keep the notation distinct to later accommodate the case where $j \neq k$. When this is the case the derivative changes slightly as the nominators derivative w.r.t x_k is now 0, and there's an actual difference between differentiating w.r.t x_k or x_j. Applying the derivate-of-a-fraction rule will now yield

$$\frac{\partial \Pr(y_i = j|x)}{\partial x_k} = -p_{ij}p_{ik}\beta \tag{54}$$

Combining these two expression is easily done with an indicator function, giving the final derivative

$$\frac{\partial \Pr(y_i = j|x)}{\partial x_k} = p_{ij}(\mathbb{1}_{(j=k)} - p_{ik})\beta \tag{55}$$

In the case of the multinomial logit model the expression of v_{ij} is different, and what's more important x_{ij} is now constant across all alternatives j so $x_{ij} = x_i$. Because of this, taking the derivative w.r.t some "x_j" is now meaningless, as there is no variation. A result is that as the sum in the denominator loops over β_j's, each iteration will produce a non-0 result under differentiation. Again simply apply the same rule of differentiation to get the desired result, keeping in mind that x is fixed across all alternatives, resulting in

$$\frac{\partial \Pr(y_i = j|x)}{\partial x} = p_{ij} \left(\beta_j - \sum_{l=1}^{J} p_{il}\beta_l \right) \tag{56}$$

Since the marginal effects are quite complicated and potentially non-linear, the log-odds ratio is often considered as an alternative. Defined simply as $\log(p_{ij}/p_{i1})$ where p_{i1} is the

probability of the baseline alternative, it is rather eqasy to show that this is equal to $x_i'\beta_j$ in the MNL case.

iii.1 Independence of irrelative alternatives

A major drawback of the above models is that the relative probability of two choices p_{ij}/p_{ik} does not depend on any other alternatives than j and k, severely restricting how individuals can substitute between alternatives in this setup.

VII. Quantile regression

Quantiles are loosely for any given distribution the "x-values" matching up to a certain probability mass, covered by the distribution. In mathematical terms, with τ being some fraction of the total probability of 1, the quantile related to τ is μ_τ which is given implicitly in

$$\tau = \Pr(y \leq \mu_\tau) = F_y(\mu_\tau) \tag{57}$$

Normally μ_τ is called the τ^{th} quantile, for example $\mu_{0.5}$ would be the 0.5^{th} quantile, or simply the median. Rewriting the above it's clear that $\mu_\tau = F_y^{-1}(\tau)$. A parallel definition is that of conditional quantiles. These are simply taken to be conditional on some other variables x, for which the conditional distribution of y is known, and possibly a set of parameters θ, that is

$$\mu_\tau(x, \theta) = F_{y|x}^{-1}(\tau) \tag{58}$$

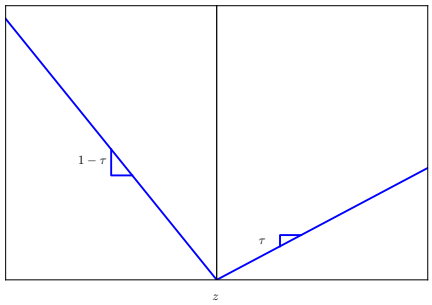
How $F^{-1}(\cdot)$ is specified determines the amount of freedom in shaping the quantile lines. For example if $\mu(x, \beta_\tau) = x'\beta_\tau$ each quantile must be linear, but can have a unique slope.

i. Estimation

Quantiles can also be defined as solutions to a minimization problem, specifically by defining $\rho(\cdot)$ as

$$\rho(z) = (\tau - \mathbb{1}_{z<0}) \cdot z \tag{59}$$

Figure 3: *ρ function for quantile regression*



Notice how ρ "weights" observations on either side of $z = 0$ differently. This weighting ensures the propper result from the optimization problem.

minimizing the expectation of $\rho(y_i - \mu_\tau)$ will produce the τ^{th} quantile. Empirically we replace the expectation with a sum, and thus end up with the following estimator for $\hat{\mu}_\tau$

$$\hat{\mu}_\tau = \arg\min_{\mu_\tau \in \mathbb{R}} \sum_{i=1}^{N} \rho(y_i - \mu_\tau) \tag{60}$$

Of course we're generally not interested in estimating constants. The mosst commonly used approach is to estimate conditional quantiles, whereby the value of μ depends on x and parameters θ, which are then the goal of optimization. In other words one normally solves

$$\hat{\theta}_\tau = \arg\min_{\theta_\tau \in \mathbb{R}} \sum_{i=1}^{N} \rho(y_i - \mu(x_i, \theta_\tau)) \tag{61}$$

The lecture slides covers a heteroscedasticity model, as well as an example of quantile regression applied to birth weight. These are left out here.

VIII. Non- and semiparametric estimation methods

As the name suggest non- and semiparametric estimation is a branch of econometrics concerned with estimating curves, densities etc. without assuming parametric functions for the relationships. Common for the nonarametric

methods covered is that they essentially all are based off of generalizations of discrete counting estimators either in the form of histograms or binned scatterplots.

i. Kernel Density estimators

Begin by observing that a histogram can be constructed by counting the number of observations within a fixed distance from each x_0, so we can mathematically describe the value of a histogram at x_0 as

$$\begin{aligned}\hat{f}_{\text{hist}}(x_0) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(x_0-h\leq x_i\leq x_0+h)}}{2h} \\ &= \frac{1}{N\cdot h}\sum_{i=1}^{N}\mathbb{1}_{\left(-1\leq\frac{x_i-x_0}{h}\leq 1\right)}\end{aligned} \tag{62}$$

where $2h$ (later respecified simply as h) is the bandwidth, i.e. the left-right distance from x_0 within which we count each x_i. The reason we divide with $2h$ and not just h is that we count within a distance of h on both the left and right side of x_0. Generalizing this to a kernel density estimator (KDE) is then as simple as moving from counting observations with the $\mathbb{1}$-function, to computing a weighted sum using a kernel $K(z)$, thus in general we have

$$\hat{f}(x_0) = \frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x_i-x_0}{h}\right) \tag{63}$$

where $K(z)$

A) Is continuous and symmetric around 0, i.e. $K(z) = K(-z)$.
B) Integrates to one: $\int K(z)\mathrm{d}z = 1$ and is mean zero: $\int zK(z)\mathrm{d}z = 0$ (this is already given by the symmetry requirement).
C) Is at least convergent to 0, for $|z| \to \infty$.
D) Has non-infinite variance and constant, $\int z^2K(z)\mathrm{d}z = \kappa$, where κ is some constant in $\mathbb{R}$.

The choice of kernel is not trivial, and several options exist. Often a regular gaussian distribution is used, but this has the disadvantage of assigning non-zero weight to all observations, meaning computations increase drastically with the number of observations. As alternatives triangle, uniform or Epanechnikov kernels might be used.

i.1 Bias of the KDE estimator