# Econometrics Notes

Preliminary and very incomplete

T. Lamadon (lamadon@uchicago.edu)

October 14, 2020

## Contents

# 1 Labor supply and the linear model

Consider that we are given data on wages, hours and consumption, together with a linear tax schedule. Consider further that we would like to know the impact on revenue labor supply of changing that tax schedule. How would we proceed?

We are given a random sample $(H_i, C_i, W_{i,} X_i)$ for a given tax linear schedule $\rho$. The sample tax revenue is given by $\sum_i (1-\rho) W_i H_i$. Can we form an estimates of the average revenue per individual? A natural guess is

$$\frac{1}{n}\sum_{i=1}^{n} \rho w_i c_i$$

but what does it mean for us to think that this is a good estimate of the actual average revenue per individual? This object of interest that lives not in the sample but in the "population" that generated the sample is often referred to as the **estimand**.

Now that we need to construct an object that we do not directly observe, we need to introduce additional notation. We need to be able to refer to the **population**. The population is what generates the sample. In general it can be fought of as the joint distribution of all the variable of the sample: $\Pr[H_1, C_1, W_{1,} X_1 ... H_n, C_n, W_{n,} X_n]$ for any given $n$.

Consider for a second that each of this draws are indeed drawn from an infinite population with distribution $F_0 = \Pr[H, C, W, X]$. We can precisely define our estimand:

$$\begin{aligned} R &= \int (1-\rho) W H \mathrm{d} F_0(W, H) \\ &= \mathbb{E}_0 (1-\rho) W_i H_i \\ &= (1-\rho) \mathbb{E}_0 W_i H_i . \end{aligned}$$

Note how this is defined using the population object. Given this estimand of interest we want to design an **estimator**. Such an object is a function of the sample. In our case a natural sample analog of our estimand is

$$R_n = \frac{1}{n}\sum_{i=1}^{n} (1-\rho) W_i C_i .$$

We can then check the properties of our estimator. Is it **unbiased**? Is if **consistent**?

We finally consider inference. Given a point estimates $R_n$we might want to create what is called coverage intervals, or even evaluate some statistical test.

### Finite sample normal assumption

We can if we are willing to make additional assumptions. For instance if we are willing to assume that $e_i = w_i c_i$ is iid with a normal distribution. Then we can actually compute the distribution of $R_n$. It is normal with mean centered on $R$ and variance given by

$$\begin{aligned} Var(R_n) &= \frac{1}{n^2}(1-\rho)^2 \sum_n Var(E_i) \\ &= \frac{(1-\rho)^2 \sigma_E^2}{n} . \end{aligned}$$

from this we can compute confidence interval for $R$ since we know that

$$\Pr[\frac{R_n - R}{\frac{(1-\rho)\sigma_E}{\sqrt{n}}} \leq x] = \Phi(x)$$
$$\Pr[R_n - R \leq x\frac{(1-\rho)\sigma_E}{\sqrt{n}}] = \Phi(x)$$

hence choosing $a_n = R_n - \frac{(1-\rho)\sigma_E}{\sqrt{n}}\Phi^{-1}(0.975)$ we get

$$\begin{aligned}\Pr[a_n \leq R] &= \Pr[R_n - \frac{(1-\rho)\sigma_E}{\sqrt{n}}\Phi^{-1}(0.975) \leq R] \\ &= \Pr[\frac{R_n - R}{\frac{(1-\rho)\sigma_E}{\sqrt{n}}} \leq \Phi^{-1}(0.975)] = 0.975\end{aligned}$$

and similarly with $b_n = R_n + \frac{(1-\rho)\sigma_E}{\sqrt{n}}\Phi^{-1}(0.975)$

$$\begin{aligned}\Pr[R \leq b_n] &= \Pr[R \leq R_n - \frac{(1-\rho)\sigma_E}{\sqrt{n}}\Phi^{-1}(0.975)] \\ &= \Pr[\frac{R - R_n}{\frac{(1-\rho)\sigma_E}{\sqrt{n}}} \leq \Phi^{-1}(0.975)] = 0.975\end{aligned}$$

and hence we get that

$$\Pr\left[a_n \leq R \leq b_n\right] = 0.95$$

**Central limit theorem**

What about in the case where we don't want to assume joint normality? We have the central limit theorem. we have the following two results. Given a sequence of random variable iid with with mean $\mu$ and variance $\sigma^2$ then we have that

$$\frac{1}{\sqrt{n}}\left(\sum_i X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Hence we can construct an asymptotic confidence interval in the exact same way as before, defining $a_n$ and $b_n$ in the same way we get that

$$\lim_{n\to\infty} \Pr\left[a_n \leq R \leq b_n\right] = 0.95$$

Off course here, for a given $n$ we do not know the exact bound, we only know that asymptotically they will vanish.

## 1.1 A second object of interest

Let's now assume that we are interested in the revenue generated for a different taxation level $\rho$. Things are a bit more difficult now because we don't have really any data directly observable for what would wages and labor supply be under a different value of $\rho$.

We can think of a particular population model as

$$\Pr[H, C, W, X|\rho],$$

which then allows us to define our estimand for any value of $\rho$. Indeed

$$\begin{aligned} R(\rho) &= \mathbb{E}_0 \left[\rho W_i H_i | \rho\right] \\ &= \rho \mathbb{E}_0 \left[W_i H_i | \rho\right] \end{aligned}$$

But the problem here is how can we learn about $\mathbb{E}_0 \left[W_i H_i | \rho_1\right]$ if we only observe $\Pr[H, C, W, X|\rho_2]$? In this particular case it seems hopeless without additional assumptions. However if one is willing to assume that $\mathbb{E}_0 \left[W_i H_i | \rho_1\right] = \mathbb{E}_0 \left[W_i H_i\right]$ then a simple estimate is

$$R_n(\rho_1) = \frac{1}{n} \sum_{i=1}^{n} (1 - \rho_1) W_i C_i,$$

however is seems unreasonable to think that people will continue to work even if you tax away all their income. One needs to make weaker assumption. Unfortunately, we see that we then enter the game of finding and defending assumptions. Later we will try to think about whether some assumptions can tested or not in the data. Some unfortunately won't be testable using a given dataset.

### State level data

Imagine that we are given some variation in our data. For instance people in different state face different level of income tax. Our data now looks likes $W_i, H_i, \rho_i$. We can still compute

$$R = \mathbb{E}(1 - \rho_i) W_i H_i$$

and indeed I have a joint distribution of $(\rho_i, W_i, H_i)$. Imagine that I am now interested in what would be the IRS revenue in the case where $\rho_i = \rho$. I guess I have two potential candidates. I could use $\mathbb{E}(1 - \rho) W_i H_i$ or I could use $\mathbb{E}\left[(1 - \rho_i) W_i H_i | \rho_i \rho\right]$ . But neither seems very satisfactory, because neither tells us what hour each individual would chose when faced with a different tax $\rho$. What we would really need to, not only to have the joint distribution of $(\rho_i, W_i, H_i)$ but actually, we would like to have for each individual, the value of $E_i = W_i H_i$ for each potential value of $\rho$. If we had such $E_i(\rho)$ we could then compute

$$\mathbb{E}(1 - \rho) E_i(\rho)$$

and what we observe is

$$\mathbb{E}(1 - \rho_i) E_i(\rho_i)$$

### Adding structure: Choice of hours and consumptions

We then make the randomness explicitly. We want to think about the sample and the population. We then define the population as what we construct the sample from. Under this iid assumption, the population is simply the

We consider the following model of labor supply decision

$$\max_{c,h} \frac{c^{1+\eta}}{1+\eta} - \beta \frac{h^{1+\gamma}}{1+\gamma}$$
$$\text{s.t. } c = \rho w h + r$$

where we further specify that $\log \beta = \nu x + \epsilon$. The FOC condition gives us that

$$\rho w c^{\eta} = \beta h^{\gamma} \text{ or } \rho w \left(\rho w h + r\right)^{\eta} = \beta h^{\gamma}$$

from which we can derive the chosen hours and consumption for a given individual $w_i(\rho), c_i(\rho)$ and $h_i(\rho)$ for any level of taxation $\rho$. In this context this is a deterministic function of $x_i, \epsilon_i$ and the parameters $\nu, \eta, \gamma$.This is sometimes called a potential outcome.

If we are willing to assume this model, and if we know the parameters we can then express our object of interest:

$$\begin{aligned} R(\rho) &= \mathbb{E}_0 \left[\rho W_i H_i | \rho\right] \\ &= \rho \mathbb{E}_0 \left[W_i(\rho) H_i(\rho) | \rho\right] \\ &= \rho \mathbb{E}_0 \left[E(X_i, \epsilon_i, \rho) | \rho\right] \end{aligned}$$

From there we can write down a DGP. We draw $X$ and $\epsilon$ independently and we draw $W$ correlated with X. From there we can compute $H_i, C_i$ using the FOC. This defines a complete DGP. Of course only observe $X, W, H$. The question for the econometrician, even in the case where where the model is true is then to figure out if the parameters can be recovered from available data. The first order condition gives us the following linear equation:

$$\log w = \gamma h - \eta c - \log \rho + \nu x + \epsilon$$

This equation looks very promising as all parameter of interests appear. This is very familiar, as it is the structure of an OLS! Let's ask ourself when is that this regression delivers correct estimates for $\gamma, \eta$

**First case,** $\eta = 0$

In this case, we the following linear equation:

$$h = -\frac{1}{\gamma} \log \rho + \frac{1}{\gamma} w - \frac{\nu}{\gamma} x - \epsilon$$

Can we use this equation to recover $\nu, \gamma$? To find out we need to check that the condition for the OLS do hold. This conditions are that

$$\mathbb{E}[\epsilon | w, x] = 0$$

as it turns out, here they do hold since $\epsilon$is independent of $w$ and $x$.

**Back to the general case**

Looking back at $\log w = \gamma h - \eta c - \log \rho + \nu x + \epsilon$, we can establish that $\mathbb{E}[\epsilon|h, c, x] = 0$, however can establish a different conditional mean expression. Let's look at:

$$\begin{aligned}\mathbb{E}[\epsilon|w, r, x] &= 0 \\ &= \mathbb{E}[\log w - \gamma h + \eta c + \log \rho - \nu x|w, r, x]\end{aligned}$$

This true because of the assumption of orthogonality of $\epsilon$ with respect to $w, r, x$. This is exactly the form of an Instrumental Variable approach. Here the instrument are $w, r, x$ and the regressor are $w, h, c$ and the dependent variable is $w$. The rank condition for identification here is that the matrix $\mathbb{E}\left[(w, r, x)' \times (w, h, c)\right]$ of size 3x3 is full rank.

**A causal parameter interpretation**

We see that the key element of the counter-factual we are interested in is linked to causal parameter of interest. This parameter is the effect of a change in wages on hours. This is usually what people refer to when talking about labor supply.

This is about the change in $h_i$ induced by a change in $w_i$. In the previous example, the OLS has a causal interpretation, because the structural model we considered did indeed satisfy $\mathbb{E}[\epsilon|w, x] = 0$. Hence the regression coefficient actually delivers the "causal effect" of change wages on hours. In what situation might this not reflect the causal effect?

## 1.2 Potential threat to causal interpretation of OLS

Let's consider 3 potential threats to the causal interpretation of $OLS$:

- omitted variable problem, imagine that part of X is not observed, for instance imagine that education is not observed. Education affect the wage, and

  $$h = c + \frac{1}{\gamma}w - \frac{\nu}{\gamma}x_1 - \underbrace{\frac{\nu}{\gamma}x_2 - \epsilon}_{v_i}$$

  and we can ask, do we have $\mathbb{E}\left[\frac{\nu}{\gamma}x_2 - \epsilon|w, x_1\right] = 0$? but there could be many reason for why that might not be the case. Indeed consider the OLS estimates of $h$ on $w$:

  $$\beta_{OLS} = \frac{cov(h_i, w_i)}{var(w_i)} = \frac{cov(\frac{1}{\gamma}w - \frac{\nu}{\gamma}x, w_i)}{var(w_i)} = \frac{1}{\gamma} + \frac{cov(-\frac{\nu}{\gamma}x, w_i)}{var(w_i)}$$

  and we see that bias can go in either direction.

- measurement error, what if wages are not reported precisely. Consider the case where the wage is realized ex-post. For instance the individual decides to work or not work based on some expected wages, but then collect more or less because of some random outcome. For example imagine a cab driver who leases a cab for a fixed number of hours, then weather realizes. In this case

  $$h_i = c + \frac{1}{\gamma}w^* - \epsilon$$

but we only observe $w_i = w_i^* + v_i$. What happens in this case? We want to compare the estimator to the estimand:

$$\beta_{OLS} = \frac{cov(h_i, w_i)}{var(w_i)} = \frac{var(w_i^*)}{var(w_i^*) + \sigma_v^2} \frac{1}{\gamma}$$

Let's go back to our model and consider the presence of endogeneity. For instance imagine that non labor income $r$ is also correlated with x.

# 2 Heckman Selection Model

### Interior solution again

We consider again our static labor supply model, but with an additional twist: the individual can choose to spend his hours at home, in which case he receives an equivalent wage $r$.

$$\begin{aligned} \sup_{c,h,e} \quad & \frac{c^{1+\eta}}{1+\eta} - \beta \frac{h^{1+\gamma}}{1+\gamma} \\ s.t. \quad & c = e \cdot w \cdot h + (1-e) \cdot r \cdot h \end{aligned}$$

Conditional on working ($e = 1$) then

$$\frac{\beta h^{\gamma}}{c^{\eta}} = w = \frac{\beta h^{\gamma}}{(wh)^{\eta}}$$

as before, which gives the following relationship in logs between hours and wages:

$$\ln h = \frac{1+\eta}{\gamma - \eta} \ln w - \frac{1}{\gamma - \eta} \ln \beta$$

where the term on the wage is the Marshallian elasticity. Focusing on this equation, and specifying further

$$\log \beta_i = \nu x_i + \epsilon_i$$

we get the following expression in logs

$$\tilde{h} = \frac{1+\eta}{\gamma - \eta} \tilde{w} - \frac{1}{\gamma - \eta} x_{it} \alpha + \frac{1}{\gamma - \eta} \epsilon_{it}$$

Under the assumption that $E(\epsilon_i | \tilde{w}_i, x_i) = 0$ we can recover the Marshaling elasticity $\frac{1+\eta}{\gamma-\eta}$ by running an OLS regression.

## Participation decision

The individual also decides to participate in the labor market at wage $w$ or use his time to produce at home at equivalent rate $r$. $r$ is directly interpretable as a reservation wage since the agent will never accept a wage less than $r$. In the same way we parametrized $\beta$ let's also expand the reservation wage:

$$\begin{aligned}\log \beta_i &= \nu x_i + \epsilon_i \\ \log r_i &= \delta z_i + u_i \\ \tilde{w}_i = \log w_i &= \theta x_i + v_i\end{aligned}$$

and let's assume that $(\epsilon_{i,} u_i)$ are joint normal and correlated with each other, but that $v_i \perp \epsilon_{i,} u_i$. This gives the following hours and participation decision:

$$\begin{aligned}\tilde{h}_i &= \frac{1+\eta}{\gamma-\eta}\tilde{w}_i - \frac{\nu}{\gamma-\eta}x_{it} + \frac{1}{\gamma-\eta}\epsilon_{it} \text{ and } e_i = 1 \qquad \text{if } \tilde{w} \geq \delta z_i + u_i \\ \tilde{h}_i &= \frac{1+\eta}{\gamma-\eta}\tilde{r}_i - \frac{\nu}{\gamma-\eta}x_{it} + \frac{1}{\gamma-\eta}\epsilon_{it} \text{ and } e_i = 0 \qquad \text{if } \tilde{w} < \delta z_i + u_i\end{aligned}$$

Call $\tilde{h}^*, \tilde{w}^*$ the observed hours and consumption, which the econometrician only sees when $e = 1$ then we get

$$\begin{aligned}\tilde{h}_i^* = \frac{1+\eta}{\gamma-\eta}\tilde{w}_i - \frac{\nu}{\gamma-\eta}x_i + \frac{1}{\gamma-\eta}\epsilon_i \qquad &\text{if } \tilde{w}_i \geq \delta z_i + u_i \\ \tilde{h}_i^* = -\infty \qquad &\text{if } \tilde{w}_i < \delta z_i + u_i\end{aligned}$$

jjCan we still credibly assume that $E(\epsilon_i|\tilde{w}, x_i, z_i, e_i = 1) = 0$? To find out let's derive it given the primitives of the model and given $e_i = 1[\tilde{w} \geq \delta z_i + \nu_i]$. We get that:

$$\begin{aligned}E(\tilde{h}_i^*|\tilde{w}, x_i, z_i, e_i = 1) &= \frac{1+\eta}{\gamma-\eta}w_i - \frac{\nu}{\gamma-\eta}x_i + \frac{1}{\gamma-\eta}E(\epsilon_i|\tilde{w}_i, x_i, z_i, e_i = 1) \\ &= \frac{1+\eta}{\gamma-\eta}w_i - \frac{\nu}{\gamma-\eta}x_i + \frac{1}{\gamma-\eta}E(\epsilon_i|\tilde{w}_i, x_i, z_i, \tilde{w}_i \geq \delta z_i + u_i) \\ &= \frac{1+\eta}{\gamma-\eta}w_i - \frac{\nu}{\gamma-\eta}x_i + \frac{1}{\gamma-\eta}E(\epsilon_i|u_i \leq \delta z_i - \tilde{w}_i)\end{aligned}$$

but since $\epsilon$ and $u$ are jointly normal we can write $\epsilon_i = \sigma u_i + \xi_i$ with $\xi_i$ independent and mean zero, so

$$E(\tilde{h}_i^*|\tilde{w}, x_i, z_i, e_i = 1) = \frac{1+\eta}{\gamma-\eta}w_i - \frac{\nu}{\gamma-\eta}x_i + \frac{\sigma}{\gamma-\eta}E(u_i|u_i \leq \delta z_i - \tilde{w}_i) + \frac{1}{\gamma-\eta}\underbrace{E(\xi_i|u_i \leq \delta z_i - \tilde{w}_i)}_{0}$$

we then note that for a mean zero, variance $\sigma$ normal distribution we have that $E(X|X < a) = -\sigma\frac{\Phi'(a/\sigma)}{\Phi(a/\sigma)} = \lambda(a)$ where $\Phi$ is the CDF for a normal distribution. $\lambda$ is called the inverse Mills ratio. This means that we have the following expression:

$$\begin{aligned}\tilde{h}_i^* &= \frac{1+\eta}{\gamma-\eta}w_i - \frac{\nu}{\gamma-\eta}x_i + \frac{\sigma}{\gamma-\eta}\lambda_i + \xi_i \\ with \qquad & E(\xi_i|\tilde{w}, x_i, z_i, e_i = 1) = 0\end{aligned}$$

where $\lambda_i = -\frac{\Phi'(\frac{\delta}{\sigma}z_i - \frac{1}{\sigma}\tilde{w}_i)}{\Phi(\frac{\delta}{\sigma}z_i - \frac{1}{\sigma}\tilde{w}_i)}$. Of course $\tilde{w}_i$ is not directly observed for non-employed people, but we have postulated a model for it. We use the fact that $\tilde{w}_i = \log w_i = \theta x_i + \zeta_i$ to re-express $\tilde{w}_i$ in terms of observables $x_i$. So estimation requires the following steps:

1. using a Probit regression, estimate the participation equation using $x_i$ and $z_i$. This gives estimates of $\frac{\theta}{\sigma}$ and $\frac{\delta}{\sigma}$ where $\sigma$ is the sum of the variances of $\nu_i$ and $\zeta_i$. Indeed remember that we have the following Probit structure:

$$\begin{aligned} e_i &= 1[\tilde{w} \geq \delta z_i + \nu_i] \\ &= 1[\theta x_i + v_i \geq \delta z_i + u_i] \\ &= 1[\theta x_i - \delta z_i \geq u_i - v_i] \end{aligned}$$

   h

2. next predict the values, and construct $\lambda_i = -\frac{\Phi'(\frac{\theta}{\sigma}x_i - \frac{\delta}{\sigma}z_i)}{\Phi(\frac{\theta}{\sigma}z_i - \frac{\delta}{\sigma}\tilde{w}_i)}$

3. regress observed hours $\tilde{h}_i$ on $x_i, \tilde{w}_i, \lambda_i$ for employed individuals and recover the Marshallian elasticity $\frac{1+\eta}{\gamma-\eta}$.

### 2.1 Application Blundell, Duncan, and Meghir (1998).

The paper proposes a group estimator. We consider an additional source of endogeneity. For instance, imagine that $\zeta_i$ also enters the $\beta_i$ equation. We then specify

$$\log \beta_i = \nu x_i + \nu_2 \zeta_i + \epsilon_i$$

We also introduce a tax parameter $\rho$. The first order condition is given by:

$$\begin{aligned} \tilde{h}_i^* &= \frac{1+\eta}{\gamma-\eta} w_i - \frac{\nu}{\gamma-\eta} x_i + \frac{\rho}{\gamma-\eta}\lambda_i + \xi_i \\ with & \quad E(\xi_i | \tilde{w}, x_i, z_i, e_i = 1) = 0 \end{aligned}$$

# References

Blundell, R., A. Duncan, and C. Meghir (1998): "Estimating Labor Supply Responses Using Tax Reforms," *Econometrica*, 66(4), 827–861.

# A Additional notes

## A.1 o and $o_p$ notations

$X_n = o(R_n)$ means that $X_n = Y_n R_n$ and $Y_n \xrightarrow{p} 0$
also we have that $(1+x)^\alpha = 1 + \alpha x + o(x^2)$

## A.2 KL divergence

Let's show that it is always positive

$$\begin{aligned} KL(p,q) &= \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \\ &= \mathbb{E}_p \log \frac{p(x)}{q(x)} \\ &= -\mathbb{E}_p \log \frac{q(x)}{p(x)} \\ &\geq -\log \mathbb{E}_p \frac{q(x)}{p(x)} \\ &= -\log \int q(x) \mathrm{d}x = 0 \end{aligned}$$

## A.3 Linear algebra refresher

- Matrices, product, rank, eigen value decompositions

## A.4 Bellman principle of optimality

TBD