14.661: Recitation 1

Chris Walters

September 17, 2010

1 Probit Maximum Likelihood

1.1 MLE background

Maximum likelihood estimation is a technique used to estimate the parameters of a model when we have a parametric model for the data generating process. Formally, we have data on realizations of some random variable or vector x_i, $i = 1...N$, and we know that the probability density or mass function associated with x is

$$f(x;\theta)$$

for some parameter vector θ. The idea of maximum likelihood estimation is to pick the parameter vector $\hat{\theta}$ that maximizes the probability of observing the sample that we actually observed. The likelihood of observing the x_i that we saw for one observation if the parameter vector is θ is

$$\mathcal{L}_i = f(x_i;\theta)$$

and (with independent and identically distributed data) the likelihood for the whole sample is just the product of likelihoods across all observations:

$$\mathcal{L} = \prod_i \mathcal{L}_i = \prod_i f(x_i;\theta)$$

The maximum likelihood estimator (MLE) is just the parameter vector that maximizes this likelihood:

$$\hat{\theta}_{ML} = \arg\max_\theta \prod_i f(x_i;\theta)$$

We often find it convenient to work with the log of the likelihood function. Maximizing a function is equivalent to maximizing its log since log is a strictly increasing function, and taking logs converts the product over individual likelihoods into a sum. We can therefore write

$$\hat{\theta}_{ML} = \arg\max_\theta \sum_i \log f(x_i;\theta)$$

As long as we've specified the distribution of the data correctly, under weak additional conditions the ML estimator will be:

1. Consistent

2. Efficient
3. Asymptotically normal

You'll work through the details of MLE in your econometrics classes. For now, it is enough to understand the intuition for the procedure, to know these properties, and to know how to apply it in some particular cases.

1.2 Example: Probit

"Probit" is a particular example of maximum likelihood. It is used for situations involving a binary choice between two alternatives, which we can call alternative 1 and alternative 0. The model is:

$$y_i^* = X_i'\beta - \epsilon_i,\ \epsilon_i|X_i \sim N\left(0, \sigma^2\right)$$

Here, y_i^* is a latent (unobserved) preference for alternative 1 relative to alternative 0; if y_i^* is positive, the agent likes alternative 1 best. X_i is a vector of observed characteristics. We want to estimate β to determine how the X's affect the agent's choice. Instead of observing y_i^*, we observe

$$y_i = 1\left\{y_i^* > 0\right\}$$

That is, we see each person's choice between the two possible alternatives, along with the vector X_i.

Note that we have assumed a parametric form for ϵ_i – the normal distribution. This is what makes the model a probit. We could make other assumptions; if ϵ_i follows a type I extreme value distribution the model is called "logit." Note also that we have made an assumption about the error distribution conditional on X_i; we will actually be doing conditional MLE, as we aren't specifying the distribution of X.

Conditioning on X_i, we can use our distributional assumption to work out the probability of each choice as a function of the parameter vector β. This is

$$Pr\left[y_i = 1|X_i\right] = Pr\left[y_i^* > 0|X_i\right]$$

$$= Pr\left[X_i'\beta > \epsilon_i|X_i\right]$$

$$= Pr\left[\tfrac{X_i'\beta}{\sigma} > \tfrac{\epsilon_i}{\sigma}\right]$$

$$= \Phi\left(\tfrac{X_i'\beta}{\sigma}\right)$$

where Φ is the standard normal cdf. This works because ϵ/σ follows a standard normal distribution conditional on X_i. Similarly, the probability of an outcome of zero is

$$Pr\left[y_i = 0|X_i\right] = 1 - \Phi\left(\tfrac{X_i'\beta}{\sigma}\right)$$

The likelihood for an individual observation, which is the probability of observing the value of y_i that we actually observed for a given set of parameters, is then

$$\mathcal{L}_i(\beta, \sigma) = \begin{cases} \Phi\left(\tfrac{X_i'\beta}{\sigma}\right), & y_i = 1 \\ 1 - \Phi\left(\tfrac{X_i'\beta}{\sigma}\right), & y_i = 0 \end{cases}$$

The likelihood for the whole sample is just the product of the individual likelihoods:

$$\mathcal{L}(\beta,\sigma)=\prod_{y_i=1}\Phi\left(\tfrac{X_i'\beta}{\sigma}\right)\cdot\prod_{y_i=0}\left[1-\Phi\left(\tfrac{X_i'\beta}{\sigma}\right)\right]$$

which we can also write as

$$\mathcal{L}(\beta,\sigma)=\prod_i\Phi\left(\tfrac{X_i'\beta}{\sigma}\right)^{y_i}\left[1-\Phi\left(\tfrac{X_i'\beta}{\sigma}\right)\right]^{1-y_i}$$

Note that β and σ are not separately identified; they always appear in a ratio. We can therefore normalize $\sigma=1$ and interpret β relative to the standard deviation of ϵ_i. Taking logs, we can define the maximum likelihood estimator of β as

$$\hat{\beta}_{ML}=\arg\max_{\beta}\sum_i\left(y_i\cdot\log\Phi\left(X_i'\beta\right)+(1-y_i)\cdot\log\left[1-\Phi\left(X_i'\beta\right)\right]\right)$$

The FOCs of this program with respect to the elements of β are

$$\sum_i\left(\frac{y_i\cdot\phi\left(X_i'\hat{\beta}\right)}{\Phi(X_i'\hat{\beta})}\cdot X_i-\frac{(1-y_i)\cdot\phi\left(X_i'\hat{\beta}\right)}{1-\Phi(X_i'\hat{\beta})}\cdot X_i\right)=0$$

which can be re-written as

$$\sum_i\frac{\left(y_i-\Phi\left(X_i'\hat{\beta}\right)\right)}{\Phi(X_i'\hat{\beta})\left(1-\Phi\left(X_i'\hat{\beta}\right)\right)}\cdot\phi\left(X_i'\hat{\beta}\right)X_i=0$$

This equation can be solved numerically for $\hat{\beta}$.

One more note on probit: The β's that we get from maximum likelihood estimation are NOT the effects of the X's on the probability of choosing one alternative rather than the other. Instead, the marginal effect of a given element of X on the probability of choosing alternative 1 is

$$\frac{\partial Pr\left[y_i=1|X_i\right]}{\partial X_{ik}}=\phi\left(X_i'\beta\right)\cdot\beta_k$$

Unlike the marginal effects we usually get from regression, this derivative is not constant as a function of X_i. We can therefore look at a number of different marginal effects. Two quantities that are of particular interest are

$$\text{Average marginal effect: }E\left[\tfrac{\partial Pr[y_i=1|X_i]}{\partial X_{ik}}\right]=E\left[\phi\left(X_i'\beta\right)\cdot\beta_k\right]$$

$$\text{Marginal effect at the average: }\frac{\partial Pr\left[y_i=1|X_i=\bar{X}\right]}{\partial X_{ik}}=\phi\left(\bar{X}'\beta\right)\cdot\beta_k$$

These need not be the same, though usually they will be close. Give estimates of β, it is straightforward to estimate either one.

2 Properties of the expenditure function

In consumer theory, the expenditure function is defined by

$$e(p_1, .., p_N, \bar{u}) = \min_{x} \sum_{i} p_i x_i$$

s.t.

$$u\left(x_1, ..., x_N\right) \geq \bar{u}$$

That is, the expenditure function is the "minimized minimand" or the expenditure minimization problem; for given prices, it tells us the minimum amount of income needed to achieve a given utility level. The minimizers of this problem are compensated demands, $x_i^c(p, \bar{u})$.

Today we will prove a couple of useful properties of the expenditure function.

2.1 Property 1: Shephard's Lemma

Shephard's Lemma says that

$$\frac{\partial e(p, \bar{u})}{\partial p_j} = x_j^c(p, \bar{u})$$

In words, the derivative of the expenditure function with respect to the price of good j is equal to the hicksian (compensated) demand for good j.

To show Shephard's Lemma, we can use the following theorem:

Envelope Theorem for Constrained Optimization: Consider the problem

$$V(\alpha) = \min_{x} f(x; \alpha)$$

s.t.

$$g(x; \alpha) \geq 0$$

Let $x^*(\alpha)$ be the minimizer. Then

$$\frac{\partial V(\alpha)}{\partial \alpha} = \frac{\partial f\left(x^*(\alpha); \alpha\right)}{\partial \alpha} - \lambda \frac{\partial g\left(x^*(\alpha); \alpha\right)}{\partial \alpha}$$

where λ is the Lagrange multiplier for the problem.

Let's prove this theorem. The Lagrangian is

$$\mathcal{L} = f(x; \alpha) - \lambda g(x; \alpha)$$

So the FOCs are

$$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x}$$

$$g(x;\alpha) = 0$$

We want to get the derivative of the objective function. We have

$$V(\alpha) = f\left(x^*(\alpha);\alpha\right)$$

so

$$\frac{\partial V}{\partial \alpha} = \frac{\partial f}{\partial x} \cdot \frac{dx^*}{d\alpha} + \frac{\partial f}{\partial \alpha}$$

Differentiating the second FOC, we know

$$\frac{\partial g}{\partial x} \cdot \frac{dx^*}{d\alpha} + \frac{\partial g}{\partial \alpha} = 0$$

$$\Longrightarrow \frac{dx^*}{d\alpha} = -\frac{\partial g/\partial \alpha}{\partial g/\partial x}$$

Plugging this and the first FOC into the expression of interest yields

$$\frac{\partial V}{\partial \alpha} = -\lambda \frac{\partial g}{\partial x} \cdot \frac{\partial g/\partial \alpha}{\partial g/\partial x} + \frac{\partial f}{\partial \alpha}$$

or

$$\frac{\partial V}{\partial \alpha} = \frac{\partial f}{\partial \alpha} - \lambda \cdot \frac{\partial g}{\partial \alpha}$$

which is what we wanted to prove. I've done this for the scalar case but it works just as well for vectors.

We can then apply this to the expenditure function example: Here,

$$f(x;\alpha) = \sum_i p_i x_i$$

$$g(x;\alpha) = u(x) - \bar{u}$$

$$V(\alpha) = e(p,\bar{u})$$

The α we are interested in is p_j, the price of good j. This does not appear in g, so we can just take the partial derivative of f to obtain

$$\boxed{\frac{\partial e}{\partial p_j} = x_j^c}$$

This is Shephard's lemma. The intuition is pretty straightforward. If the price of good j increases by 1 dollar, the first effect is to make the bundle the consumer is already consuming more expensive by x_j^c. The consumer can also re-optimize after the price change, but since she is already at a maximum, small changes in her consumption bundle have no first-order effect on the value of the objective function.

2.2 Property 2: Concavity in prices

The second property we want to show is that the expenditure function is concave in prices. Let's work with vectors and dot products instead of sums for ease of notation. Concavity of the expenditure function means that for any two price vectors p_1 and p_2, and any $\alpha \in (0, 1)$,

$$e\left(\alpha p_1 + (1-\alpha)p_2, u\right) \geq \alpha e\left(p_1, u\right) + (1-\alpha)e(p_2, u)$$

We'll prove it by contradiction. Suppose this isn't true. Then for some p_1, p_2 and α,

$$e\left(\alpha p_1 + (1-\alpha)p_2, u\right) < \alpha e\left(p_1, u\right) + (1-\alpha)e(p_2, u)$$

By definition of the expenditure function, this says

$$\left[\alpha p_1 + (1-\alpha)p_2\right] \cdot x^c\left(\alpha p_1 + (1-\alpha)p_2, u\right) < \alpha x^c(p_1, u) + (1-\alpha)x^c(p_2, u)$$

Just re-arranging terms, this says

$$\alpha \cdot \left(p_1 x^c\left(\alpha p_1 + (1-\alpha)p_2, u\right) - p_1 x^c(p_1, u)\right) + (1-\alpha) \cdot \left(p_2 x^c\left(\alpha p_1 + (1-\alpha)p_2, u\right) - p_2 x^c(p_2, u)\right) < 0$$

For this whole quantity to be less than zero, at least one of the two terms must be negative. But neither can be; each yields utility u, and by definition of the expenditure function the quantity $p_i x^c(p_i, u)$ must be smaller than the cost of any other bundle that yields utility u at prices p_i. So we have a contradiction, and the expenditure function must be concave.

This makes intuitive sense. Using our Shephard's lemma result, note that the second derivative of the expenditure function is the first derivative of Hicksian demand. Concavity means that the own-price derivative of compensated demand is negative; if price goes up, compensated demand must go down. To put it another way, when price increases, the substitution effect leads to a decrease in consumption.

3 Regression review

3.1 Bivariate OLS

Suppose we want to estimate the parameters of the bivariate model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

The OLS estimator is given by

$$\hat{\beta}_{OLS} = \frac{\frac{1}{N}\sum(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{N}\sum(x_i - \bar{x})^2} = \frac{\widehat{Cov(x_i, y_i)}}{\widehat{Var}\left(x_i\right)}$$

The law of large numbers tells us that sample moments converge to population moments as long as the population moments exist, so

$$\boxed{plim\hat{\beta}_{OLS} = \frac{Cov\left(x_i, y_i\right)}{Var\left(x_i\right)}}$$

This is always true for a bivariate regression; you should remember this formula.

3.2 Partialling Out: The Frisch-Waugh Theorem

Most of the regressions you will run in your life will include multiple explanatory variables, but the bivariate regression formula is still relevant for such situations due to the following result.

Suppose we want to run the regression

$$y_i = \alpha + \beta x_i + z_i'\gamma + \epsilon_i$$

Let's define this as a regression in the population, so that $Cov(x_i, \epsilon_i) = Cov\left(z_{ik}, \epsilon_i\right) = 0$. We can obtain our estimate of β in two steps:

1. First, run the regression $x_i = \theta_0 + z_i'\theta_1 + \eta_i$. Compute the residuals from this regression: $\tilde{x}_i = x_i - \hat{\theta}_0 - z_i'\hat{\theta}_1$.
2. Run the regression $y_i = \alpha + \beta\tilde{x}_i + \epsilon_i$.

The estimate of β from step 2 will be algebraically identical to what we would have gotten by running the full multivariate regression. This is the Frisch-Waugh Theorem. Why does this work? Let's pretend we have data on the whole population so we can just work with population quantities (the same thing holds exactly in finite samples). The 2-step regression gives us

$$\beta_{FW} = \frac{Cov\left(\tilde{x}_i, y_i\right)}{Var\left(\tilde{x}_i\right)}$$

$$= \frac{Cov\left(\tilde{x}_i, \alpha + \beta x_i + z_i'\gamma + \epsilon_i\right)}{Var\left(\tilde{x}_i\right)}$$

$$= \beta\frac{Cov\left(\tilde{x}_i, x_i\right)}{Var\left(\tilde{x}_i\right)} + \frac{Cov\left(\tilde{x}_i, z_i'\gamma\right)}{Var\left(\tilde{x}_i\right)} + \frac{Cov\left(\tilde{x}_i, \epsilon_i\right)}{Var\left(\tilde{x}_i\right)}$$

As a residual from a regression on z_i, $\tilde{x}_i$ is uncorrelated with linear functions of z_i by construction. In addition, since $\tilde{x}_i$ is just a linear function of x_i and z_i, it must be uncorrelated with the population residual ϵ_i by construction. This leaves us with

$$= \beta \cdot \frac{Cov\left(\tilde{x}_i, \tilde{x}_i + \hat{x}_i\right)}{Var\left(\tilde{x}_i\right)}$$

$$= \beta$$

since $\hat{x}_i$ and $\tilde{x}_i$ are uncorrelated by construction.

Thanks to the Frisch-Waugh Theorem, we can therefore just use the bivariate regression formula assuming that we've "partialled out" any other explanatory variables we want to include in this way. Note that we could partial out the additional explanatory variables from y also, or not – it doesn't matter. This occurs because

$$y = \hat{y} + \tilde{y}$$

and

$$Cov(y_i, \tilde{x}_i) = Cov(\hat{y}_i + \tilde{y}_i, \tilde{x}_i) = Cov(\tilde{y}_i, \tilde{x}_i)$$

since $\hat{y}$ is a projection onto the space spanned by z_i and (as a residual from a regression on z_i) $\tilde{x}_i$ is therefore uncorrelated with $\hat{y}_i$ by construction.

3.3 OLS Problems

There are a number of situations in which OLS will fail to estimate the parameters of interest. In such cases, we have to appeal to alternative econometric techniques. Three such cases are covered below.

3.3.1 Omitted Variable Bias

Suppose the regression we want to run is

$$y_i = \alpha + \beta x_i + \gamma z_i + \eta_i$$

with $Cov(x_i, \eta_i) = Cov(z_i, \eta_i) = 0$. However, we can't observe z_i, so we omit it from our regression and instead run:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

In this case, the probability limit of our OLS estimator is

$$plim \hat{\beta}_{OLS} = \frac{Cov(x_i, y_i)}{Var(x_i)}$$

$$= \frac{Cov\left(x_i, \alpha + \beta x_i + \gamma z_i + \eta_i\right)}{Var(x_i)}$$

$$\boxed{\implies plim \hat{\beta}_{OLS} = \beta + \gamma \cdot \frac{Cov(x_i, z_i)}{Var(x_i)}}$$

You should remember this formula. Note that the bivariate OLS estimator is consistent if z_i has a coefficient of zero in the full regression, or if x_i and z_i aren't correlated.

3.3.2 Measurement Error

Suppose we want to run the regression

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

with $Cov\left(x_i^*, \epsilon_i\right) = 0$. We observe y_i^*, but instead of observing x_i^* we instead observe

$$x_i = x_i^* + v_i$$

with $Cov(v_i, x_i^*) = Cov(v_i, \epsilon_i) = 0$. What happens if we regress y_i^* on x_i? Note that we can write

$$y_i^* = \alpha + \beta x_i + (\epsilon_i - \beta v_i)$$

Since v_i shows up in both the error term and in our regressor, we are in trouble. The result of running this regression is

$$plim \hat{\beta}_{OLS} = \frac{Cov\left(x_i, y_i^*\right)}{Var\left(x_i\right)}$$

$$= \frac{Cov\left(x_i^* + v_i, \alpha + \beta x_i^* + \epsilon_i\right)}{Var(x_i)}$$

$$= \beta \cdot \frac{Var\left(x_i^*\right)}{Var\left(x_i\right)}$$

Under the conditions assumed above this is

$$\boxed{\implies plim\hat{\beta}_{OLS} = \beta \cdot \frac{Var\left(x_i^*\right)}{Var\left(x_i^*\right) + Var(v_i)}}$$

The quantity

$$\lambda = \frac{Var\left(x_i^*\right)}{Var\left(x_i^*\right) + Var(v_i)}$$

is called the "reliability ratio;" λ is the proportion of the variance in the observed x_i due to the variance of the true variable of interest ("signal" rather than "noise"). Since $\lambda \in (0,1)$, we have

$$|plim\hat{\beta}_{OLS}| < |\beta|$$

That is, random measurement error causes our parameter estimate to be too close to zero. This is called "attenuation bias." Note that if you are only interested in testing the null hypothesis that $\beta = 0$, then if you can reject this hypothesis the possibility of measurement error strengthens your conclusion.

3.3.3 Simultaneity Bias

A regression can also fail to recover parameters of interest when the right and left-hand side variables are jointly determined. This is best explained by way of an example. Suppose we have the supply and demand equations (respectively)

$$q_i = \alpha + \beta p_i + \epsilon_i$$

$$p_i = \omega + \gamma q_i + \eta_i$$

with $Cov(\epsilon_i, \eta_i) = 0$. What happens if we regress quantity on price in an attempt to recover the parameters of supply? Note that we can write

$$p_i = \omega + \gamma\left(\alpha + \beta p_i + \epsilon_i\right) + \eta_i$$

$$\implies p_i = \frac{\omega + \gamma\alpha}{1 - \gamma\beta} + \frac{\gamma\epsilon_i + \eta_i}{1 - \gamma\beta}$$

Again, since the error from the supply equation appears in p_i due to feedback through the demand equation, a regression is not going to give us the parameters of interest. Running this regression gives

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(p_i, q_i\right)}{Var\left(p_i\right)}$$

$$= \frac{Cov\left(p_i, \alpha + \beta p_i + \epsilon_i\right)}{Var\left(p_i\right)}$$

$$= \beta + \frac{Cov\left(p_i, \epsilon_i\right)}{Var\left(p_i\right)}$$

$$= \beta + \frac{\gamma(1-\gamma\beta)Var\left(\epsilon_i\right)}{\gamma^2 Var\left(\epsilon_i\right) + Var\left(\eta_i\right)} < \beta$$

since $\beta > 0$, $\gamma < 0$.

A lot of what we do in future recitations will be to review methods that solve these problems.

4 Panel Data Methods

In many data sets, we get to see repeated observations on the same units (people, firms, countries, etc.) over time. This is called "panel data." Formally, our data include $i = 1...N$ units and $t = 1...T$ time periods. Suppose we want to estimate a model like the following:

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it}$$

In such models we often decompose the error term ϵ_{it} into a permanent individual-specific component θ_i and an idiosyncratic error term η_{it}:

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \eta_{it}$$

with

$$Cov(\alpha_i, \eta_{it}) = 0$$

$$Cov\left(\eta_{it}, \eta_{ks}\right) = 0 \ \forall i \neq j, \ t \neq s$$

In addition let's suppose that

$$Cov(\eta_{it}, x_{it}) = 0$$

In this case the only bias we are worried about comes from θ_i; for now, we are assuming that the only potential omitted variables are things that are fixed over timeThere are two standard ways to proceed from here.

4.1 Random Effects

Suppose we are comfortable making the assumption that

$$Cov\left(x_{it}, \theta_i\right) = 0$$

In this case, we have

$$Cov\left(x_{it}, \theta_i + \eta_{it}\right) = 0$$

so there is no correlation between our right-hand side variable and the composite error term. Then we know that OLS will be consistent! However, given the nature of our data, we can actually do even better than OLS. Note that

$$Cov\left(\theta_i + \eta_{it}, \theta_i + \eta_{is}\right) = Var\left(\theta_i\right),$$

so we have autocorrelation in the unobserved part of the model. Furthermore, we know the structure of this correlation – there is a common covariance between the error terms for observations on the same individual, and no other autocorrelation. In situations with a known non-spherical error structure, the most efficient estimator is Generalized Least Squares (GLS). For this panel model, GLS is called "Random Effects." You'll learn how to do GLS in econometrics.

4.2 Fixed Effects

In most cases, we won't be comfortable making the assumption that $Cov\left(x_{it}, \theta_i\right) = 0$. Instead, we view θ_i as a potential source of omitted variable bias. Fortunately, we can deal with this by simply controlling for θ_i! We can do this by directly including a vector of person-dummies in our regression:

$$y_{it} = \beta x_{it} + \sum_{j=1}^{N} \theta_j D_{ij} + \eta_{it}$$

Here D_{ij} is a dummy variable that is one if $i = j$ and zero else; each person gets their own dummy variable (note that I've now excluded the constant). We can just run this with OLS, knowing that including the person-dummies has eliminated any bias due to permanent unobserved characteristics. This procedure is called Fixed Effects (FE).

It is worth thinking more about how to interpret Fixed Effects estimates. Recall from the Frisch-Waugh theorem that we can obtain our estimates by first partialling out the person-dummies and then regressing y on the resulting residuals. For once, it will be easier to use matrices. Let's order the data with our T observations on person 1 first, followed by our T observations on person 2, etc. Let $X_{NT\times 1}$ be the vector containing the x_{it}. Then the coefficient vector from regressing X on the person-dummies is

$$(D'D)^{-1}D'X$$

and the residuals are given by

$$\tilde{X} = X - D\left(D'D\right)^{-1} D'X$$

where

$$D_{NT\times N} = \begin{bmatrix} 1_T & 0_T & \cdots & 0_T \\ 0_T & 1_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & 1_T \end{bmatrix}$$

is our matrix of person dummies. Here 1_T is a column-vector of T 1's. Writing this out yields

$$\tilde{X} = X - D\left(\begin{bmatrix} 1_T' & 0_T' & \cdots & 0_T' \\ 0_T' & 1_T' & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T' & \cdots & 0_T' & 1_T' \end{bmatrix} \cdot \begin{bmatrix} 1_T & 0_T & \cdots & 0_T \\ 0_T & 1_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & 1_T \end{bmatrix}\right)^{-1} D'X$$

$$= X - D \begin{bmatrix} T & 0 & \cdots & 0 \\ 0 & T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & T \end{bmatrix}^{-1} D'X$$

$$= X - D \begin{bmatrix} \frac{1}{T} & 0 & \cdots & 0 \\ 0 & \frac{1}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{T} \end{bmatrix} D'X$$

$$= X - \begin{bmatrix} 1_T & 0_T & \cdots & 0_T \\ 0_T & 1_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & 1_T \end{bmatrix} \begin{bmatrix} \frac{1}{T} & 0 & \cdots & 0 \\ 0 & \frac{1}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{T} \end{bmatrix} D'X$$

$$= X - \begin{bmatrix} \frac{1_T}{T} & 0_T & \cdots & 0_T \\ 0_T & \frac{1_T}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_T & \cdots & 0_T & \frac{1_T}{T} \end{bmatrix} \begin{bmatrix} 1'_T & 0'_T & \cdots & 0'_T \\ 0'_T & 1'_T & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0'_T & \cdots & 0'_T & 1'_T \end{bmatrix} X$$

$$= X - \begin{bmatrix} \frac{1_{T\times T}}{T} & 0 & \cdots & 0 \\ 0 & \frac{1_{T\times T}}{T} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1_{T\times T}}{T} \end{bmatrix} \begin{bmatrix} X_{1_{T\times 1}} \\ X_{2_{T\times 1}} \\ \vdots \\ X_{N_{T\times 1}} \end{bmatrix}$$

$$= X - \bar{X}$$

where $\bar{X}$ is a matrix where each individual observation has been replaced with the mean for the relevant person. Then fixed effects is equivalent to estimating the regression

$$y_{it} - \bar{y}_i = \beta (x_{it} - \bar{x}_i) + \omega_{it}$$

That is, fixed effects estimates the model using deviations from person-means. This is called the "within" model because it uses only variation within persons and does not use the variation in mean x's and mean y's across people. Other things to know about fixed effects:

1. Fixed effects cannot be used to estimate the coefficients on time-invariant variables – there is no variation left in such variables once we take out the individual-specific means
2. With only 2 time periods, fixed effects is equivalent to first differences (with no constant; fixed effects with a time dummy is equivalent to first differences with a constant)
3. Fixed effects and differencing can make measurement error a lot worse. We will see this in a future recitation.

14.661 Labor Economics I
Fall 2010