Panel Data Econometrics Summary

Rebecca Sela

September 28, 2005

1 General Statistical Models

Throughout this, we assume that both response variables (y) and covariates (x) are random. Therefore, most of the models we use will be defined by the mean of the response variable conditional on the covariates, $E(y|x)$.

Definition Suppose $E(y|x) = g(x)$. The *linear Taylor series approximation* of the model is given by:

$$\hat{g}(x) = g(x_0) + \sum_{k=1}^{K} \frac{\partial}{\partial x_k} g(x)|_{x=x_0}(x_k - x_{k0}) = \delta_0 + \sum_{k=1}^{K} \delta_k(x_k - x_{k0})$$

where x_0 is any point (usually a central value like the mean).

Definition Suppose $E(y|x) = g(x)$. The *linear projection* of the model is given by:

$$g^*(x) = E(y) + \sum_{k=1}^{K} \frac{Cov(y, x_k)}{Var(x_k)}(x_k - E(x_k)) = \gamma_0 + \sum_{k=1}^{K} \gamma_k(x_k - E(x_k))$$

Note that these two linearizations are generally not the same (though they are identical if $g(x)$ is linear). Regression models are estimates of the linear projection. Note that the linear projection does not exist if the second moments do not exist.

Theorem 1.1 The Law of Iterated Expectations. $E_X(E(Y|X)) = E(Y)$.

Theorem 1.2 The Law of Iterated Variances. $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$.

Theorem 1.3 The Law of Iterated Covariances. $Cov(X, Y) = Cov(X, E(Y|X)) = E_X(X \cdot E(Y|X)) - E(X)E_X(E(Y|X))$.

Definition The *marginal effect* of a covariate is the effect of changes in that variable on the conditional mean. In the case of continuous covariates (and a differentiable conditional expectation), the marginal effect is given by $\delta(x) = \frac{\partial E(y|x)}{\partial x}$. In the case of a covariate that is a dummy variable, the marginal effect is $E(y|x, d = 1) - E(y|x, d = 0)$.

There are other measures of the effects of variables on the response, such as the elasticity (defined as $\epsilon(x) = \delta(x)\frac{x}{E(y|x)}$). In most cases, the marginal effects depend on x. To remedy this, we use estimated average partial effects.

Definition The *average partial effect* is given by:

$$APE_x = E_q(\delta(x,q)) = \int_R \delta(x,q)f(q)dq$$

where $f(q)$ is the density of all the other variables q.

Since $f(x)$ is generally unknown, we may estimate the average partial effects using the empirical distribution: $\hat{APE} = \frac{1}{N}\sum_{j=1}^{N}\hat{\delta}(x_j)$. We may also estimate the average partial effect by finding the marginal effect at the mean: $\hat{APE} = \hat{\delta}(\bar{x})$. A Taylor series shows that this approximation is close:

$$\begin{aligned} \delta(x) &= \delta(\mu) + \delta'(\mu)(x-\mu) + \frac{1}{2}\delta''(\mu)(x-\mu)^2 + O((x-\mu)^3) \\ APE &= E(\delta(x)) = \delta(\mu) + \frac{1}{2}\delta''(\mu)Var(x) + O((x-\mu)^3) \approx \delta(\mu) \end{aligned}$$

This version has a simpler standard error as well.

Sometimes, we are given structural models in which not all the parameters can be estimated. In this case, we may convert them to reduced form models, but the best we will be able to do is estimate some functions of the parameters.

1.1 Generalized Method of Moments

Suppose we have a model in a K-dimensional β that includes M orthogonality conditions, $E(g(\beta, X)) = 0$. We consider their sample counterparts, $\overline{g}(\beta) = \frac{1}{N}\sum_{i=1}^{N} g(\beta, X) = 0$. If $M = K$, then the model is exactly identified and there is a unique solution. If $M < K$, then there is not enough information to solve the model (and more conditions are needed). If $M > K$, then the model is *overidentified*, and we will not be able to find a solution that fulfills all of the equations at once. Note that these conditions need not be linear.

Definition Let A be a symmetric positive definite matrix. A *minimum distance estimator* of β is found by minimizing $\overline{g}(\beta)'A^{-1}\overline{g}(\beta)$.

For any positive definite A, $\hat{\beta}$ is consistent and asymptotically normal. The asymptotic variance is given by:

$$AsyVar(\beta) = \left(\frac{\partial\overline{g}(\beta)}{\partial\beta}\right)' A\left(AsyVar(\overline{g}(\beta))\right) A\left(\frac{\partial\overline{g}(\beta)}{\partial\beta}\right)$$

Definition The minimum distance estimator with weighting matrix $A = (AsyVar(\overline{g}(\beta)))^{-1}$ is the *generalized method of moments* (GMM) estimator.

This is the minimum-variance estimator in the class of minimum distance estimators with these orthogonality conditions. To implement this:

1. Use the identity as a weighting matrix to estimate β. This is used to estimate the residuals and the weighting matrix, $AsyVar(\overline{g}(\beta))$.

2. Use this weighting matrix with the same orthogonality conditions.

If $M > K$, we have *over-identifying restrictions* which may be used to test the assumption that all of the restrictions hold. Under the null hypothesis that all the restrictions are correct, that is, $E(\overline{g}(\beta)) = 0$, we have the test statistic

$$q = \overline{g}(\hat{\beta})'(\hat{AsyVar}(\overline{g}(\hat{\beta})))^{-1}\overline{g}(\hat{\beta}) \to \chi^2(M-K)$$

A rejection of the null hypothesis does not identify which restriction is being violated.

To test restrictions on β (not on moment conditions), we have:

$$\begin{aligned} q_R &= \overline{g}(\hat{\beta}_R)'(\hat{AsyVar}(\overline{g}(\hat{\beta}_R)))^{-1}\overline{g}(\hat{\beta}_R) \to \chi^2(M-K_R) \\ q_U &= \overline{g}(\hat{\beta}_U)'(\hat{AsyVar}(\overline{g}(\hat{\beta}_U)))^{-1}\overline{g}(\hat{\beta}_U) \to \chi^2(M-K_U) \\ q_R - q_U &\to_D \chi^2(K_U - K_R) \end{aligned}$$

For this test to be valid, the same weighting matrix must be used for both estimations.

1.2 M Estimation

Definition Let $q(y_i, x_i, \theta)$ be any function of the data and the parameter, such that $E(q(y_i, x_i, \theta))$ is minimized by the true value of θ. The *M estimator* is $\hat{\theta}$ which minimizes $\overline{q}(\theta) = \frac{1}{n}\sum_{i=1}^{n} q(y_i, X_i, \theta)$.

Let θ_0 be the true parameter value. By the weak law of large numbers, $\overline{q}(\theta) \to E(q(y, X, \theta))$. Furthermore, $\hat{\theta} \to \theta_0$ if θ_0 is unique. (This rules out perfect collinearity, indeterminacy (where some parameters are irrelevant under certain values of other parameters), and parameters that need to be normalized.) For other results, we assume that:

- $q(y, X, \theta)$ continuous in θ for all y, X.
- $\frac{\partial}{\partial\theta}q$ exists and its continuous.
- q is twice differentiable (though the second derivatives need not be consistent).

Theorem 1.4 *Under these assumptions, $\hat{\theta}$ is asymptotically normal.*

Proof Let $\frac{\partial}{\partial\theta}q = g$. Then,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} q(y_i, Z_i, \theta) \\ &= \frac{1}{n}\sum_{i=1}^{n} g(y_i, X_i, \theta) \\ &= \overline{g}(y, X, \theta) \end{aligned}$$

which is asymptotically normal by the Lindberg-Feller Central Limit Theorem. Using a Taylor Series expansion, we find that:

$$0 = \overline{g}(y, X, \hat{\theta}) = \overline{g}(y, X, \theta_0) + \overline{H}(\tilde{\theta})(\hat{\theta} - \theta_0)$$

where $\overline{h} = \frac{1}{n}\sum_{i=1}^{N}\frac{\partial^2}{\partial\theta^2}q(y, X, \theta)$ and $\tilde{\theta}$ lies between $\hat{\theta}$ and θ_0. Then,

$$(\hat{\theta} - \theta_0) = \overline{H}(\tilde{\theta})^{-1}\overline{g}(y, X, \theta_0)$$

and $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normal, with $E(\hat{\theta}) = \theta_0$ and $AsyVar(\hat{\theta}) = \overline{H}(\theta_0)^{-1}Var(\overline{g}(y, X, \theta_0))\overline{H}(\theta_0)^{-1}$. $Var(\overline{g}(y, X, \theta_0)) = \frac{1}{n}E(g(y, X, \theta_0)g(y, X, \theta_0)')$, which can be estimated (and works out to the White estimator in the linear least squares case). ∎

Note that the calculated M estimator might not converge or might converge to the wrong value (such as a local maximum or minimum). Trying different starting values may help.

M estimation is special case of GMM where the model is exactly identified (and therefore the weighting matrix does not matter).

1.2.1 Hypothesis testing

Suppose we have the null hypothesis, $c(\theta) = 0$ which has J functions (and therefore J restrictions). Assume $R(\theta) = \frac{\partial}{\partial\theta'}c(\theta)$ has rank J. In the *Wald Test*, we have the test statistic:

$$(c(\hat{\theta}) - c(\theta))'(R(\theta)Var(\hat{\theta})R(\theta)')^{-1}(c(\hat{\theta}) - c(\theta))'$$

which has a $\chi^2(J)$ distribution asymptotically.

We may also calculate the criterion function, $\overline{q}$, in the restricted and unrestricted cases. Under the null hypothesis, the test statistic, $2n(\overline{q}^R - \overline{q}^U)$, also has a $\chi^2(J)$ distribution asymptotically. (When this is used with the maximum likelihood estimates, this is the likelihood ratio test.)

Finally, we may use the *Score Test*, in which we calculate whether the derivatives, $\overline{g}$, are close to 0 when they are evaluated at the restricted estimates. (This also leads to a $\chi^2(J)$ distribution asymptotically.)

1.2.2 Non-Linear Least Squares

In non-linear least squares, we minimize $q(y_i, X_i, \theta) = (y_i - m(X_i, \theta))^2$, where $E(y|X) = m(x, \theta_0)$.

Gauss-Marquadt Algorithm for NLLS:

- Step 0: Choose $\hat{\theta}^{(0)}$. (This may come from a linear regression, where many of the parameters are set to 0 to avoid non-linearity.)
- Step 1: Set $q_i = m(x_i, \theta)$.
- Step 2: Then, $g_i = \frac{\partial}{\partial \theta} m(x_i, \theta) = X_i^0$. The x_i^0 are called the *pseudo-regressors*.
- Step 3: Set $\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + ((X_k^0)' X_k^0)^{-1} (X_k^0)' e_k^0$, where e_k^0 are the residuals, $y_i - m(x_i, \hat{\theta}_k)$.
- Step 4: Continue until the $\hat{\theta}$ converge.

The conditional variance estimator for NLLS is $\frac{1}{n-k} \sum_{i=1}^n (y_i - m(X_i, \hat{\theta}))^2 ((X^0)' X^0)^{-1}$.

1.2.3 Maximum Likelihood Estimation

Suppose the density of y given x is fully specified. Then, we may maximize the likelihood function, which is simply the joint density of the observations as a function of the parameters:

$$\begin{aligned} L(\beta|Y, X) &= \prod_{i=1}^n f(y_i|X_i, \beta) \\ l(\beta|Y, X) &= \sum_{i=1}^n \log f(y_i|X_i, \beta) \end{aligned}$$

Note that the likelihood is conditional on X. Therefore, we assume that the distribution of X depends only on parameters that are not in $f(y_i|X_i, \beta)$. Then, we have:

$$l(\beta, \delta|Y, X) = \sum_{i=1}^n \log f(y_i|X_i, \beta) + \sum_{i=1}^n \log g(x_i|\delta)$$

and we use only the first term.

Suppose we have two sets of parameters, α, β, and we have a closed form solution of $\frac{\partial \log L}{\partial \alpha} = 0$ for α in terms of β. Then, the *concentrated log likelihood* is given by $\log L_C(\beta, \alpha(\beta))$. We may maximize the concentrated log likelihood in terms of β (which may be easier) and then find the MLE of α as $\hat{\alpha} = \alpha(\hat{\beta})$.

Note that the MLE is an M-estimator, so all the consistency, normality, and testing results apply. Furthermore, under certain regularity conditions,

$$\begin{aligned} Var(\frac{\partial}{\partial\theta}\log L) &= -E(\frac{\partial^2 \log L}{\partial\theta'\partial\theta}) \\ AsyVar(\hat{\theta}) &= \left(-E\left(\frac{\partial^2 \log L}{\partial\theta'\partial\theta}\right)\right)^{-1} Var(\frac{\partial \log L}{\partial\theta})\left(-E\left(\frac{\partial^2 \log L}{\partial\theta'\partial\theta}\right)\right)^{-1} \\ &= \left(-E\left(\frac{\partial^2 \log L}{\partial\theta'\partial\theta}\right)\right)^{-1} \end{aligned}$$

This leads to three variance estimators:

- BHHH (an estimate based on the first derivatives): $(\sum_{i=1}^{n} \frac{\partial \log L_i}{\partial\theta}\frac{\partial \log L_i}{\partial\theta}')$.
- An estimate based on the second derivatives: $(-\sum_{i=1}^{n} H_i)^{-1}$.
- An estimate based on the expectations of the second derivatives (if it can be computed): $(E(-\sum_{i=1}^{n} H_i))^{-1}$

Because of the Cramer-Rao lower bound, the MLE is asymptotically efficient among all consistent and asymptotically normal estimators for the density of the data. Furthermore, the MLE is invariant; if g is a continuous function and $\hat{\theta}$ is an MLE of θ, then the MLE of $g(\theta)$ is $g(\hat{\theta})$.

Does two step estimation (front of p. 36) live here?

1.3 Bayesian Methods

In Bayesian econometrics, one formulates theory and begins with *priors* which assemble and form beliefs based on existing evidence. Then, evidence is collected and *posteriors* combine the prior beliefs with new evidence in order to revise beliefs about the theory. (In contrast, classical econometrics formulates the theory, gathers evidence, and then accepts or rejects the theory.) In general, Bayesians tend to use uninformative priors, instead of having real prior beliefs.

Bayesians see the likelihood as a function containing all the current information about the parameters and the data. According the the *likelihood principle*, any two proportional likelihoods have the same information (for example, the binomial and the negative binomial).

Bayesians understand randomness as uncertainty about the state of the world, instead of as a random process that governs nature.

Bayesian "estimation" studies the characteristics of the posterior distribution. The Bayesian estimator is (usually) the mean of the posterior distribution. According to a theorem by Bernstein and Von Mises, in large samples, the posterior will be approximately normal with the mean equal to the maximum likelihood estimator.

In linear regression with normally distributed errors, we have the likelihood:

$$L(\beta, \sigma^2|y, x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta)\right)$$

We may use a uniform (improper) prior for β and a gamma prior for σ^2 (which is the conjugate prior). If we integrate σ^2 out of the joint posterior, then $f(\beta|y,x)$ turns out to the be multivariate t distribution with mean $\hat{\beta}$ and covariance matrix $\frac{N-K}{N-K-2}s^2(X'X)^{-1}$.

More generally, this is how one uses Bayesian estimation:

1. Parameterize the model.
2. Compute the likelihood conditional on the parameters, $L(data|\theta)$.
3. Develop a joint prior for the parameters, $p(\theta)$.
4. Compute the posterior, which is proportional to the product of the likelihood and the prior: $L(data|\theta)p(\theta)$.
5. Since calculation of statistics like $E(\beta|data) = \int \beta \frac{f(data|\beta)p(\beta)}{f(data)} d\beta$ in closed form is often impossible, we generally use simulation, in which we deduce the posterior, draw random samples from it, and compute statistics based on the sample.

Note that both the prior and the posterior are joint distributions. For inference about individual parameters, we need to use the marginals, $p(\beta|data)$ and $p(\sigma^2|data)$. However, we sometimes only know $p(\beta|data, \sigma^2)$ and $p(\sigma^2|data, \beta)$. To do this, we use the Gibbs sampler. If a posterior cannot be sampled from directly, then we may use the *Metropolis-Hastings Algorithm.*

2 The Linear Model

In the linear model, we have $Y = X\beta + \epsilon$, where X is a random matrix with K variables (possibly including a constant) for each of N observations of full rank and reasonable moment conditions. In this model, we assume that:

- $E(\epsilon|x) = E(\epsilon) = 0$ so that $Cov(\epsilon, x) = 0$.
- $E(Y|X) = X\beta$ (if we are just using a linear approximation, this is a linear projection, not a Taylor series)

We estimate the parameters by least squares:

$$\begin{aligned} b &= (X^T X)^{-1} X^T y \\ s^2 &= \frac{e^T e}{N} \approx \frac{e^T e}{N-K} \end{aligned}$$

Under some conditions,

- $plim(b) = \beta$, and the estimator is consistent.
- $\sqrt{N}(b-\beta) \rightarrow_d Normal(0, \sigma^2(plim(\frac{X^T X}{N}))^{-1})$, and the estimator is asymptotically normal.

- In finite samples, we may use the approximation, $b \to_a Normal(\beta, \frac{\sigma^2}{N}(plim(\frac{X^TX}{N}))^{-1})$.
- The estimated asymptotic variance of b is given by $s^2(X^TX)^{-1}$.

Most hypothesis testing is in the context of nested models. In this case, we have J restrictions of the form $r(\beta, q) = 0$, where $\frac{\partial r(\beta,q)}{\partial \beta} = R(\beta, q)$ is a matrix of full rank. We have three tests of the null hypothesis that the restrictions hold:

- *Wald test*: We estimate the unrestricted model and find b. Under the null hypothesis, we must have $r(b,q)^T Var(r(b,q))^{-1} r(b,q) \to_d \chi^2(J)$. We use the delta method to estimate the variance and find that $Var(r(b,q)) \approx R(b,q)Var(b)R(b,q)^T \approx R(b,q)s^2(X^TX)^{-1}R(b,q)^T$.
- *Likelihood Ratio Test*: For each model, we have $logL(\hat{\beta}, \hat{\sigma}^2) = -\frac{N}{2}(1 + \log(2\pi) + \log(\hat{\sigma}^2))$. Under the null hypothesis, $2(logL(\hat{\beta}_{unrestricted}) - logL(\hat{\beta}_{restricted})) \to_d \chi^2(J)$.
- *Score/LM Test*: We estimate the restricted model based on only X_1, to find the residuals, $\tilde{u}$. Under the null hypothesis, these residuals are unrelated to X_2. Let R_u^2 be from the regression of $\tilde{u}$ on X_1, X_2. Under the null hypothesis, $LM = NR_u^2 \sim \chi_j^2$. To make this heteroskedasticity-robust, let $\hat{r}$ be the matrix of residuals from the regression of X_2 on X_1. Then, the test statistic is

$$LM = (\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{r_i}'\tilde{u_i})'(\frac{1}{N}\sum_{i=1}^{N} \tilde{u}_i^2 \hat{r}_i'\hat{r}_i)^{-1}(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{r_i}'\tilde{u_i})$$

  which is related to the regression of 1 on $\tilde{u}\hat{r}$ with no constant.

In the linear case, the Wald statistic is always the largest and the Lagrange multiplier statistic is always the smallest.

2.1 Heteroskedasticity

Suppose we have a linear equation, $Y = X'\beta + \epsilon$, with $Var(\epsilon_i|X) = \sigma_i^2$ and no covariances. That is, $Var(\epsilon|X) = \sigma^2\Omega$, where Ω is diagonal but not the identity. Then, the asymptotic variance of the coefficients estimator is:

$$AsyVar(b) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

The OLS estimator of the coefficients is consistent, but the standard errors are wrong. Instead, we may use the *White Estimator*:

$$\hat{\sigma}^2 X'\hat{\Omega}X = \sum_{i=1}^{N} e_i^2 x_i' x_i$$

This estimator is heteroskedasticity robust, but the estimation is not efficient.

2.2 Spatial Autocorrelation

Definition *Spatial autocorrelation* occurs when the value in one location is correlated with the values in nearby locations. With positive spatial autocorrelation, like values tend to cluster. With negative spatial autocorrelation, a checkerboard-type pattern may emerge.

To test for this, we use *Moran's I Spatial Autocorrelation Statistic*:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N} z_i^2}$$

where $z_i = x_i - \mu_i$ and $w_{ij} = 1$ if i and j are contiguous (but not equal).

We may also use the model:

$$Y - \mu 1 = \lambda W(Y - \mu 1) + \epsilon$$

where W is the *contiguity matrix* with the w_{ij} above. This must be specified in advance. λ is the spatial autocorrelation parameter. Solving this, we find:

$$(Y - \mu 1) = (I - \lambda W)^{-1}\epsilon$$

and $Y \sim [\mu 1, \sigma_\epsilon^2((I - \lambda W)^T(I - \lambda W))^{-1}]$.

2.3 Instrumental Variables

Definition Suppose we have a model $Y = X'\beta + \epsilon$. If $E(\epsilon_{it}|X_{it}) = 0$ then we say that X is *exogenous*. If the expectation is non-zero, then we say that X is *endogenous*.

In the case that X is endogenous, the OLS estimate of β is biased, with $plim(b) = \beta + plim(\frac{X'X}{N})^{-1}(\frac{X'\epsilon}{N})$.

Definition Suppose we have a model $Y = X\beta + \epsilon$ with $E(\epsilon|X_k) \neq 0$ (so that only that last variable is endogenous). Suppose there exists a variable, Z, such that $E(X_k|X_1, ..., X_{k-1}, Z) \neq E(X_k|X_1, ..., X_{k-1})$ and $E(\epsilon|X_1, ..., X_{k-1}, Z) = 0$. Then, Z is an *instrumental variable*.

Let $X = X_1, ..., X_k$ and $Z = X_1, ..., X_{k-1}, Z_k$. We define the instrumental variables estimator by $\hat{\beta} = (Z'X)^{-1}Z'y$. Note that this does *not* just replace X by Z. This estimator is consistent, with variance:

$$\hat{AsyVar}(\hat{\beta}) = \frac{\sum_{i=1}^{N}(y_i - x_i'\beta)^2}{N}(Z'X)^{-1}Z'Z(X'Z)^{-1}$$

If Z is not very correlated with X, then $Z'X \approx 0$, and the variance is quite large. In general, the IV estimate can be quite imprecise, and the OLS estimator may have a smaller MSE in finite samples. In addition, if there is even a small

covariance of Z with the error term, a small covariance between X and Z will magnify the bias this causes, since $plim(\hat{\beta}_K) = \beta_K + \frac{Cov(Z,u)}{Cov(Z,X_k)}$.

Suppose we have $Y = X_1'\beta_1 + X_2'\beta_2 + \epsilon$, where K_1 variables satisfy $Cov(X_1, \epsilon) = 0$ but K_2 variables have $Cov(X_2, \epsilon) \neq 0$. Suppose there is a set of $M \geq K_2$ variables, W, such that W are exogenous and correlated with X_2. Let $Z_1 = X_1$ and $Z_2 = WP$, where P is any matrix that creates K_2 linear combinations of W. Then, we may use IV on these variables, so that $\hat{\beta} = (Z'X)^{-1}Z'Y$. Then, $\hat{\sigma}_\epsilon^2 = \frac{1}{N}(Y - X\hat{\beta})'(Y - X\hat{\beta})$ and $\hat{Var}(\hat{\beta}) = \hat{\sigma}_\epsilon^2(Z'X)^{-1}Z'Z(X'Z)^{-1}$. The optimal P (and the optimal member of this class of estimators) is found in two-stage least squares:

1. Regress X_1 and X_2 on X_1 and W and compute the predicted values. Note that

$$\begin{aligned} \hat{X}_1 &= (X_1, W)P_1 = (X_1, W)\begin{pmatrix} I \\ 0 \end{pmatrix} = X_1 \\ \hat{X}_2 &= (X_1, W)P_2 = Z(Z'Z)^{-1}Z'X \end{aligned}$$

   Note that $\hat{X}_2$ is a linear combination of X_1 and W.

2. Regress Y on $\hat{X} = (\hat{X}_1, \hat{X}_2)$ to estimate β. That is $\hat{\beta} = (\hat{X}'X)^{-1}\hat{X}'Y$.

Note that:

$$\begin{aligned} \hat{X}'\hat{X} &= (Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X) \\ &= X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X \\ &= X'Z(Z'Z)^{-1}Z'X \\ &= \hat{X}'X \end{aligned}$$

The variance of the residuals is $\hat{\sigma}_\epsilon^2 = \frac{1}{N}(Y - X\hat{\beta})'(Y - X\hat{\beta})$, and $\hat{Var}(\hat{\beta}) = \hat{\sigma}_\epsilon^2(\hat{X}'\hat{X})^{-1}$.

The Wald test can still be used (provided that the correct standard error is used). However, the F statistic must be calculated differently:

$$\begin{aligned} SSU &= (Y - X\hat{\beta}_U)'(Y - X\hat{\beta}_U) \\ SSR &= (Y - \hat{X}\hat{\beta}_R)'(Y - \hat{X}\hat{\beta}_R) \\ S\hat{S}U &= (Y - \hat{X}\hat{\beta}_U)'(Y - \hat{X}\hat{\beta}_U) \\ F &= \frac{(S\hat{S}R - S\hat{S}U)/J}{SSU/(N-K)} \sim F(J, N-K) \end{aligned}$$

The White estimator for heteroskedasticity is:

$$Asy\hat{V}ar(\hat{\beta}) = (\hat{X}'\hat{X})^{-1}(\sum_{i=1}^{N}(y_i - x_i'\hat{\beta})^2\hat{x}_i'\hat{x}_i)(\hat{X}'\hat{X})^{-1}$$

OLS is inconsistent but may have a smaller MSE. In addition, 2SLS or IV may be biased in finite samples.

We may also test for endogeneity (to avoid IV/2SLS is we can). If X_2 is endogenous, then the coefficient β_3 in this regression is non-zero: $Y = \beta_1' X_1 + \beta_2' X_2 + \beta_3 \hat{X}_2 + \epsilon$.

2.3.1 Instrumental Variables and GMM

The orthogonality conditions to use GMM with IV are $E(Z(Y - X'\beta)) = 0$. In this case, we have $\frac{\partial \overline{g}(\beta)}{\partial \beta} = Z'X$. Note that this leads to $AsyVar(\hat{\beta}_{GMM}) = AsyVar(\hat{\beta}_{2SLS})$, and 2SLS is asymptotically equivalent to GMM under homoskedasticity. In the case of heteroskedasticity, $AsyVar(\overline{g}(\beta)) = \frac{1}{N^2}\sum_{i=1}^{N} \sigma_i^2 Z_i Z_i'$, which gives us a weighting matrix of $\frac{1}{N^2}\sum_{i=1}^{N} e_i^2 Z_i Z_i'$.

2.4 Systems of Equations and Seemingly Unrelated Regressions

Consider the M-equation system:

$$\begin{aligned} y_{i1} &= \beta_1' x_{i1} + \epsilon_{i1} \\ &... \\ y_{iM} &= \beta_M' x_{iM} + \epsilon_{iM} \end{aligned}$$

If the covariances of the error terms across equations are non-zero, then running equation-by-equation OLS is inefficient. Instead, we use the seemingly unrelated regressions (SUR) procedure:

1. Estimate $\beta_1, ..., \beta_M$ using OLS for each equation.
2. Estimate the cross-equation covariances by $\hat{\sigma}_{jk} = \frac{1}{M} e_j' e_k$.
3. Use the estimated covariances for FGLS (Zellner's method).
4. The residuals from the FGLS can be used to re-estimate the covariance matrix for another round of FGLS. The iterations of this will converge to the MLE.

Suppose we have a system of equations,

$$\begin{aligned} Y_1 &= X_1'\beta_1 + \epsilon_1 \\ &... \\ Y_G &= X_G'\beta_G + \epsilon_G \end{aligned}$$

where each equation holds for N observations and the X_g may be endogenous. As with SUR's, the equations can be fit separately, but this may be inefficient if the errors are correlated or if the β_m have elements in common (and inconsistent

if there is endogeneity). Suppose each equation has $L_g \geq K_g$ instruments, Z_g, with $E(\epsilon_{ig} z_{ig}) = 0$. Then, we may write:

$$
\begin{aligned}
Y_i &= \begin{pmatrix} y_{i1} \\ y_{i2} \\ ... \\ y_{iG} \end{pmatrix} \\
X_i &= \begin{pmatrix} x'_{i1} & 0 & ... & 0 \\ 0 & x'_{i2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & x'_{iG} \end{pmatrix} \\
\beta &= \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ ... \\ \beta_{iG} \end{pmatrix} \\
Z_i &= \begin{pmatrix} z'_{i1} & 0 & ... & 0 \\ 0 & z'_{i2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & z'_{iG} \end{pmatrix}
\end{aligned}
$$

Note that X_i has $K_1 + ... + K_G$ columns, and Z_i has $L_1 + ... + L_G$ columns. This gives us the orthogonality conditions:

$$
0 = E(\frac{1}{N} \sum_{i=1}^{N} Z_i \epsilon_i) = E\left(\frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} z'_{i1} \epsilon_{i1} \\ ... \\ z_{iG} \epsilon_{iG} \end{pmatrix} \right)
$$

As before, we use a weighting matrix. If the disturbances across equations are uncorrelated, then the weighting matrix is block-diagonal (with one block per equation). If there is correlation across equations, then the weighting matrix becomes:

$$
\hat{W} = \left(\frac{1}{N^2} \begin{pmatrix} \sum_{i=1}^{N} \hat{\epsilon}_{i1}^2 z_{i1} z'_{i1} & ... & \sum_{i=1}^{N} \hat{\epsilon}_{i1} \hat{\epsilon}_{iG} z_{i1} z'_{iG} \\ ... & & ... \\ \sum_{i=1}^{N} \hat{\epsilon}_{iG} \hat{\epsilon}_{i1} z_{iG} z'_{i1} & ... & \sum_{i=1}^{N} \hat{\epsilon}_{iG}^2 z_{iG} z'_{iG} \end{pmatrix} \right)^{-1}
$$

This gives us the quadratic form:

$$
q = \sum_{g=1}^{G} \sum_{h=1}^{G} \left(\frac{1}{N} \sum_{i=1}^{N} z_{ig} (y_{ig} - x'_{ig} \beta_g) \right) \hat{w}_{gh} \left(\frac{1}{N} \sum_{i=1}^{N} z_{ih} (y_{ih} - x'_{ih} \beta_h) \right)
$$

As before, we may use 2SLS to estimate W and then minimize the quadratic form. We may also use this framework to impose constraints on β or include non-linear functions of β.

3 Models with Individual Effects

Suppose we have a model $y_{it} = x'_{it}\beta + c_i + \epsilon_{it}$, for $i = 1, ..., N$ and $t = 1, ..., T_i$ (if $T_i = T$ for all i, then this is called a *balanced panel*). In this model, c_i is an *unobservable individual effect*. We are interested in estimating $E(y_{it}|x_{it}, c_i)$.

Definition We define the group and time means for a variable in a panel by:

$$\begin{aligned} \overline{Z_{i\cdot}} &= \frac{1}{T_i}\sum_{t=1}^{T_i} Z_{it} \\ \overline{Z_{\cdot t}} &= \frac{1}{N_t}\sum_{i=1}^{N_t} Z_{it} \end{aligned}$$

We assume that the full data vector, X has full column rank, and we assume strict exogeneity, in which $Cov(\epsilon_{it}, x_s) = 0$ for all t, s. (This excludes the case of lagged dependent variables.)

Throughout this, we use "fixed T" asymptotics, in which we hold T_i fixed and let $N \to \infty$.

3.1 Pooled Regression

We may simply regress Y on X without considering the panel structure. Then, our estimated slope coefficients are:

$$\begin{aligned} b &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + C + \epsilon) \\ &= \beta + (\frac{1}{N}\sum_{i=1}^{N} X'_iX_i)^{-1}(\frac{1}{N}\sum_{i=1}^{N} X'_ic_i) + (\frac{1}{N}\sum_{i=1}^{N} X'_iX_i)^{-1}(\frac{1}{N}\sum_{i=1}^{N} X'_i\epsilon_i) \\ plim(b) &= \beta + plim((\frac{1}{N}\sum_{i=1}^{n} X'_iX_i)^{-1})Cov(\overline{X_i}, c_i) \end{aligned}$$

if we assume that the ϵ_{it} do not depend on X. The coefficients of the pooled OLS regression are biased if the individual effects are related to the regressors. If there is no relationship, then the pooled OLS estimates are consistent, but not efficient. Furthermore, the group effects will cause serial correlation, and a robust variance estimator, the *cluster estimator*, must be used:

$$(\sum_{i=1}^{N} Z'_iZ_i)^{-1}(\sum_{i=1}^{N} Z'_i\hat{w}_i\hat{w}'_iZ_i)(\sum_{i=1}^{N} Z'_iZ_i)^{-1}$$

3.2 First Differences

If $y_{it} = x'_{it}\beta + c_i + \epsilon_{it}$, then $\Delta y_{it} = (\Delta x'_{it})\beta + (\epsilon_{it} - \epsilon_{i,t-1})$. In this case, the errors are autocorrelated, but the OLS estimators would be consistent, and FGLS or Newey-West could be used to correct autocorrelation.

In first differences, any time trend will become a constant, and any time dummy variables will become sequences of +1, -1, and 0.

Difference in Differences

Suppose $T_i = 2$ for all observations, and a subsample is in a "treatment" in the second period. Then, we have the model $\Delta y_i = \delta_0 + (\Delta x_i)'\beta + \delta D_i + u_i$, where D_i is the treatment dummy. If there are no other regressors, then $d_1 = \overline{\Delta y_{treatment}} - \overline{\Delta y_{control}}$ is the *difference in differences estimator*.

Note that sometimes one should control for "regression to the mean" by adding $\overline{X}$ to the regression.

3.3 Fixed Effects Estimators

In the *fixed effects model*, c_i is allowed to be arbitrarily correlated with the regressors (but we still require that $E(\epsilon_{it}|x_{it}, c_i) = 0$). Let $d_1, ..., d_N$ be dummy variables for each group. Then we use OLS to estimate the equation

$$y_i = x_i\beta + \sum_{i=1}^{N} \alpha_i d_i + \epsilon_i$$

With so many variables, the estimation may take up too much memory. Therefore, we use an equivalent method that avoids estimating the dummy variable coefficients.

Theorem 3.1 Frisch-Waugh. *To estimate only β in the equation,*

$$y = [X, D] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \epsilon$$

we use the estimate $b = (X'M_DX)^{-1}(X'M_Dy)$, where

$$M_D = \begin{bmatrix} I_{T_1} - \frac{1}{T_1} d_1 d_1' & 0 & ... & 0 \\ 0 & I_{T_2} - \frac{1}{T_2} d_2 d_2' & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & I_{T_1} - \frac{1}{T_N} d_N d_N' \end{bmatrix}$$

Definition The *within transformation* of panel data subtracts the group means from each observation. That is, we have $\ddot{y}_{it} = y_{it} - \overline{y_{i.}}$. The *between transformation* is just the group mean.

Note that the total variation is the sum of the within-group variation and the between-group variation:

$$\sum_{i=1}^{N}\sum_{t=1}^{T_i}(z_{it} - \overline{z_{..}})^2 = \sum_{i=1}^{N}\sum_{t=1}^{T_i}(z_{it} - \overline{z_{i.}})^2 + \sum_{i=1}^{N} T_i(\overline{z_{i.}} - \overline{z_{..}})^2$$

The matrix, M_D, is just the within transformation (and is idempotent), so we may estimate β by running the OLS regression of $\ddot{y}$ on $\ddot{x}$. This is called the *least squares dummy variables* (LSDV) estimator. We then back out $\alpha_i = \frac{1}{T_i}\sum_{t=1}^{T_i}(y_{it} - x'_{it}b) = \overline{e_i}$.

The asymptotic variance and estimate of the variance are:

$$\begin{aligned} AsyVar(b) &= \frac{\sigma_\epsilon^2}{\sum_{i=1}^N T_i} plim(\frac{1}{\sum_{i=1}^N T_i}\sum_{i=1}^N X_i' M_{Di} X_i)^{-1} \\ \hat{\sigma}_\epsilon^2 &= \frac{\sum_{i=1}^N \sum_{t=1}^{T_i}(y_{it} - a_i - x'_{it}b)^2}{\sum_{i=1}^N T_i - N - K} \end{aligned}$$

Note that the estimates of α_i are unbiased but not consistent; they have an asymptotic variance on the order of $\frac{1}{T_i}$, since future observations are for different groups and therefore will not give more information about previously sampled groups. For a similar reason, the degrees of freedom correction in the denominator of the variance must include the $-N$, or the variance will be asymptotically biased. (In fact, the MLE of the variance is asymptotically biased; this is a case of the Incidental Parameters Problem.)

Under the Gauss-Markov and fixed effects assumptions, the LSDV estimator is consistent and efficient. If $Var(\epsilon_i) = \Omega_i \neq \sigma_\epsilon^2 I_{T_i}$, then the slope estimator is consistent but not efficient, and we must estimate the asymptotic variance using a generalization of the White estimator (by Arellano) that deals with both heteroskedasticity and serial correlation:

$$\hat{AsyVar}(b) = (\sum_{i=1}^N X' M_{Di} X_i)^{-1} (\sum_{i=1}^N (X_i' M_{Di}) e_i e_i' (M_{Di} X_i)) (\sum_{i=1}^N X' M_{Di} X_i)^{-1}$$

To test the null hypothesis of pooled OLS versus the alternative of fixed effects, one may use an F test.

Note that time invariant regressors are perfectly collinear with the dummy variables, and therefore their impact cannot be estimated. We could run a two stage regression, in which the fixed effects are estimated and are then regressed on the time-invariant variables. This assumes that the classical assumptions hold in the second regression (and that we correct the variance of $\hat{c}_i$, since it is inversely proportional to T_i).

We may also use *two-way fixed effects*, in which there are both group-specific and time-specific effects. This may be estimated by demeaning the data with respect to both effects:

$$\ddot{y}_{it} = y_{it} - \overline{y_{i\cdot}} - \overline{y_{\cdot t}} + \overline{y_{\cdot\cdot}}$$

If the panel is balanced and T is relatively small, it is sometimes easier to just add the $T-1$ time dummies directly. If only the coefficients on the other covariates are estimated, then the fixed effects and means are:

$$\begin{aligned} \hat{\mu} &= \overline{y_{\cdot\cdot}} - \overline{x_{\cdot\cdot}}' b \\ \hat{\alpha}_i &= (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}}) - (\overline{x_{i\cdot}} - \overline{x_{\cdot\cdot}})' b \\ \hat{\gamma}_i &= (\overline{y_{\cdot t}} - \overline{y_{\cdot\cdot}}) - (\overline{x_{\cdot t}} - \overline{x_{\cdot\cdot}})' b \end{aligned}$$

In addition, there can be a *nested effects model*, such as:

$$y_{ijkt} = x'_{ijkt}\beta + u_{ijk} + v_{ij} + w_i + \epsilon_{ijkt}$$

where there is a fixed effect for each layer. (Estimation of this is more tractable with maximum likelihood, but carefully constructed dummy variables can also be used.)

Definition The *between groups estimator* is estimated from the regression $\overline{y_i} = \overline{x_i}'\beta + (c_i + \overline{\epsilon_i})$, where both c_i and ϵ_i are in the error term. (Note that the error term may be correlated with x under the fixed effects assumptions, in which case this estimator is inconsistent.)

3.4 Random Effects Model

In the random effects model, we have:

$$\begin{aligned} y_{it} &= x'_{it}\beta + c_i + \epsilon_{it} \\ E(c_i|X_i) &= 0 \\ E(\epsilon_{it}|X_i, c_i) &= 0 \end{aligned}$$

Under these assumptions, the pooled OLS regression of Y on X is unbiased and consistent, because the group effects in the error are uncorrelated with the regressors. However, because the errors within groups are autocorrelated, a robust estimator for the covariance matrix should be used, and the regression is not efficient.

In terms of the regression model orthogonality conditions, we may let $f_i = \frac{T_i}{\sum_{j=1}^N T_j}$, to find that:

$$\begin{aligned} plim\frac{1}{\sum_{i=1}^N T_i}X'w &= plim\frac{1}{\sum_{i=1}^N T_i}\sum_{i=1}^N X'_i w_i \\ &= plim\frac{1}{\sum_{i=1}^N T_i}\sum_{i=1}^N X'_i(\epsilon_i + u_i 1) \\ &= plim(\sum_{i=1}^N f_i\frac{X'_i\epsilon_i}{T_i} + \sum_{i=1}^N f_i\frac{x'_i 1}{T_i}u_i) \\ &= plim(\sum_{i=1}^N f_i\frac{X'_i\epsilon_i}{T_i} + \sum_{i=1}^N f_i\overline{X_i}u_i) \\ &= 0 \end{aligned}$$

(The two terms correspond to the two different types of errors and the assumptions about their covariances.) In addition, $\frac{X'X}{\sum_{i=1}^N T_i} = \sum_{i=1}^N f_i\frac{X'_i X_i}{T_i}$. The

covariance matrix of the errors in the i^{th} group is of the form:

$$\Omega_i = Var(\epsilon_i + u_i 1) = \begin{pmatrix} \sigma_\epsilon^2 + \sigma_u^2 & \sigma_u^2 & ... & \sigma_u^2 \\ \sigma_u^2 & \sigma_\epsilon^2 + \sigma_u^2 & ... & \sigma_u^2 \\ ... & ... & ... & ... \\ \sigma_u^2 & \sigma_u^2 & ... & \sigma_\epsilon^2 + \sigma_u^2 \end{pmatrix} = \sigma_\epsilon^2 I_{T_i} + \sigma_u^2 1_{T_i} 1'_{T_i}$$

The covariance matrix, Ω, of the entire regression is a block diagonal matrix with blocks having the form above (in an unbalanced panel, the blocks will have different sizes). We may also write

$$\frac{X'\Omega X}{\sum_{i=1}^N T_i} = \sum_{i=1}^N f_i \frac{X_i'\Omega_i X_i}{T_i} = \sigma_\epsilon^2 \sum_{i=1}^N f_i \frac{X_i'X_i}{T_i} + \sigma_u^2 \sum_{i=1}^N f_i \overline{x_{i.}} \overline{x_{i.}}'$$

This allows us to estimate the OLS variance correctly using the *cluster estimator*:

$$\frac{X'\hat{\Omega}X}{\sum_{i=1}^N T_i} = \sum_{i=1}^N f_i \frac{X_i \hat{w}_i \hat{w}_i' X_i}{T_i}$$

where the $\hat{w}_i$ are the estimated residuals.

The inverse of each block is of the form

$$\Omega_i^{-1} = \frac{1}{\sigma_\epsilon^2}(I_{T_i} - \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T_i\sigma_u^2} 1_{T_i} 1'_{T_i})$$

We may also calculate the "square root" of the inverse as:

$$\Omega_i^{-1/2} = \frac{1}{\sigma_\epsilon}(I - \theta_i(1_{T_i}(1'_{T_i}1_{T_i})^{-1}1'_{T_i}))$$

where $\theta_i = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T_i\sigma_u^2}}$.

This form can be used in GLS:

$$\begin{aligned} \hat{\beta} &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y) \\ &= (\sum_{i=1}^N X_i'\Omega_i^{-1}X_i)^{-1}(\sum_{i=1}^N X_i'\Omega_i^{-1}y_i) \\ &= (\sum_{i=1}^N (\Omega_i^{-1/2}X_i)'(\Omega_i^{-1/2}X_i))^{-1}(\sum_{i=1}^N (\Omega_i^{-1/2}X_i)'(\Omega_i^{-1/2}y_i)) \end{aligned}$$

This is a form of autocorrelation; however, the autocorrelation does not fade over time (*equicorrelation*). Note that we may write $y_{it}^* = \Omega_i^{-1/2} y_i = \frac{1}{\sigma_u}(y_i - \theta_i \overline{y_i} 1)$, and the GLS regression is equivalent to the OLS regression of y_{it}^* on x_{it}^*. If $\theta_i = 1$, then this is the fixed effects specification; if $\theta_i = 0$, this is the pooled specification.

For feasible GLS, we must estimate σ_ϵ^2 and σ_u^2. Note that the variance of the residuals from pooled OLS is $\sigma_\epsilon^2 + \sigma_u^2$ while the variance of the residuals

from LSDV is σ_ϵ^2. Thus, we may estimate the two and take their difference to estimate σ_u^2. This difference may be negative in some cases; this suggests that there may be autocorrelation in the residuals or some other problem with the model. (We may also estimate σ_u^2 by taking the estimated covariance between residuals from the same group in different periods, as suggested by Wooldridge.)

We may use a Lagrange Multiplier test to test the null hypothesis of pooled OLS versus the alternative of random effects. The test statistic is:

$$LM = \frac{NT}{2(T-1)} \left(\frac{\sum_{i=1}^N (T\overline{e_{i.}})^2}{\sum_{i=1}^N \sum_{t=1}^T e_{it}^2} - 1 \right)^2$$

We may also use maximum likelihood on the random effects model. To do this, we assume that $\epsilon_{it} \sim Normal(0, \sigma_\epsilon^2)$ and $u_i \sim Normal(0, \sigma_u^2)$. Then, the total error for the i^{th} group, $w_i = \epsilon_i + u_i 1$, has a $Normal(0, \Omega_i)$ distribution. We use this to construct the likelihood function:

$$\begin{aligned} \log L &= \sum_{i=1}^N \log L_i \\ \log L_i(\beta, \sigma_\epsilon^2, \sigma_u^2) &= -\frac{1}{2}\left(T_i \log(2\pi) + \log|\Omega_i| + (y_i - x_i\beta)'\Omega_i^{-1}(y_i - x_i\beta)\right) \end{aligned}$$

If this is maximized symbolically, then the result will be FGLS, with some estimates of σ_ϵ^2 and σ_u^2. Alternatively, it can be maximized iteratively, according to the following algorithm:

1. Start with the FGLS estimates of $\beta, \sigma_\epsilon^2, \sigma_u^2$.
2. Compute $\hat{\beta}_{k+1}$ using FGLS, holding $\hat{\sigma}_{\epsilon,k}^2$ and $\hat{\sigma}_{u,k}^2$ fixed.
3. Compute $\hat{\sigma}_{\epsilon,k+1}^2 = \frac{1}{\sum_{i=1}^N (T_i - 1)} \sum_{i=1}^N \hat{\epsilon}_{i,k+1}' M_D^i \hat{\epsilon}_{i,k+1}$ and $\hat{\sigma}_{u,k+1}^2 = \frac{1}{N} \sum_{i=1}^N \overline{\epsilon_{i,k+1}}^2$.
4. Iterate until these estimates converge.

We may also transform the parameters to simplify the maximization:

$$\begin{aligned} \theta &= \frac{1}{\sigma_\epsilon^2} \\ \tau &= \frac{\sigma_u^2}{\sigma_\epsilon^2} \\ R_i &= T_i\tau + 1 \\ Q_i &= \frac{\tau}{R_i} \\ \log L_i &= \frac{1}{2}(\theta(\epsilon_i'\epsilon_i - Q_i(T_i\overline{\epsilon_i})^2) + \log R_i + T_i \log\theta + T_i \log(2\pi)) \end{aligned}$$

After optimization with these transformed variables, we must use the delta method to estimate the variance of the original parameter estimates.

Mundlak's Estimator

Suppose $c_i = \overline{x_i}'\delta + u_i$, where the u_i are uncorrelated with the group means. Then, we have the regression model $y_{it} = x_{it}'\beta + \overline{x_i}'\delta + u_i + \epsilon_{it}$, which is just the random effects model. Thus, adding in the group means reduces a fixed effects model to random effects, under these assumptions.

We may also add proxy variables for the unobserved effects. If the remaining unobserved effects are uncorrelated with the other variables, then we may again reduce the problem to random effects.

3.5 Fixed Effects versus Random Effects

Since fixed effects must estimate one parameter for each group, it is inefficient relative to random effects. Furthermore, time-invariant variables cannot be used with fixed effects. However, the assumption of random effects is quite strong, and random effects will be biased and inconsistent if it is violated.

Hausman's Test is used to test the null hypothesis that the group effects are uncorrelated with the regressors. Under the null hypotheses, both estimators are consistent, but only the FGLS estimator is efficient. Under the alternative hypothesis, LSDV is consistent but FGLS is inconsistent. To test this, we check whether the test statistic, $\hat{q} = \hat{\beta}_{FE} - \hat{\beta}_{RE} = 0$, is close to 0.

Lemma 3.2 *Under the hypothesis of random effects,*

$$\begin{aligned}\sqrt{NT}(\hat{\beta}_{RE} - \beta) &\to_D Normal(0, V_{RE}) \\ \sqrt{NT}(\hat{\beta}_{FE} - \beta) &\to_D Normal(0, V_{FE})\end{aligned}$$

and these two estimators have an asymptotic covariance of 0.

This allows us to calculate $\hat{Var}(\hat{q}) = \hat{V}_{FE} - \hat{V}_{RE}$. (For this difference to be guaranteed to be positive definite, we must use the same value of σ_ϵ^2; the estimate from LSDV is preferable.) We then have a Wald test, $W = \hat{q}\hat{Var}(\hat{q})^{-1}\hat{q}$, which is distributed $\chi^2(k)$, where k is the number of time-varying variables in the regression.

Alternatively, we can use the *variable addition test*. Under the null hypothesis of random effects, $Cov(\overline{x_i.}, u_i) = 0$. Therefore, we may fit the original model plus the group means of all the (time-varying) variables with FGLS and test the restriction that the coefficients on all the group means are zero.

3.6 Violations of the Assumptions

Let Z_i be the regressors (including the dummy variables in the fixed effects case), w_i be the errors, and θ be the parameters (in either fixed or random effects). Then the robust covariance matrix is $(\sum_{i=1}^N Z_i'Z_i)^{-1}(\sum_{i=1}^N Z_i'\hat{w}_i\hat{w}_i'Z_i)(\sum_{i=1}^N Z_i'Z_i)^{-1}$ (this is the cluster estimator that is used with pooled OLS).

3.6.1 Heteroskedasticity

Note that the worst cases of heteroskedasticity in the two models are $E(\epsilon_{it}^2|Z_i) = \sigma_{\epsilon,it}^2$ (where each error has its own variance) for fixed effects and $E(\epsilon_{it}^2|Z_i) = \sigma_{\epsilon,it}^2$ and $E(u_i^2|Z_i) = \sigma_{u,i}^2$ for random effects. Notice that, in both cases, heteroskedasticity is detectable only if it is related to the variables in the model, which may include the group identifiers.

For the fixed effects model, options include:

- Using the robust covariance matrix with usual LSDV. This does not take advantage of the possible relationship of heteroskedasticity to groups or other variables.
- Assume that $E(\epsilon_{it}^2|X_i) = \sigma_{\epsilon,i}^2$. Then, we have the robust variance estimator,

$$\hat{Var}(b|X) = \left(\sum_{i=1}^{N} X_i' M_D^i X_i\right)^{-1} \left(\sum_{i=1}^{N} \frac{\sum_{t=1}^{N} e_{it}^2}{T} X_i' M_D^i X_i\right) \left(\sum_{i=1}^{N} X_i' M_D^i X_i\right)^{-1}$$

- We may also do FGLS under the assumption of group-specific heteroskedasticity. Then, the GLS estimate of β is

$$\hat{\beta} = \left(\sum_{i=1}^{N} \frac{1}{\sigma_{\epsilon,i}^2} (X_i' M_D^i X_i)\right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{\sigma_{\epsilon,i}^2} X_i' M_D^i y_i\right)$$

  We do FGLS using $\hat{\sigma}_{\epsilon,i}^2 = \frac{1}{T_i} \sum_{i=1}^{T_i} e_{it}^2$, where the residuals are from regular LSDV. (Note that this weighting does not affect the estimates of the group-specific intercepts.)

- We may also run FGLS using other models for the residuals, such as $\sigma_{\epsilon,it}^2 = \sigma_\epsilon^2 f(z'\delta)$. We then use regress the initial estimates of the residuals on the independent variables, and run FGLS with the estimated variances.
- We may run standard FGLS on the demeaned variables (after dropping one period of each, to avoid serial correlation in the error terms caused by demeaning).

For the random effects model, there are two different variances that may be heteroskedastic. Options include:

- Using the cluster estimator (from above) is valid:

$$\hat{Var}(b|X) = (X'X)^{-1} (\sum_{i=1}^{N} (\sum_{t=1}^{T} X_{it}' e_{it})' (\sum_{t=1}^{T} X_{it}' e_{it})) (X'X)^{-1}$$

  (Using just the White estimator is incorrect because it ignores the cross-observation correlation from the random effects.)

- The matrix estimated above could be used for FGLS as well, but it involves many more parameter estimates, which will lead to bad finite sample properties.
- We cannot use GLS if $\sigma^2_{u,i}$ depends only on i, since there is only one observation of each u_i. Its variance could be modeled as a function of the regressors, though.
- We may also model $\sigma^2_{u,i}$ and $\sigma^2_{\epsilon,i}$ jointly as a function of the regressors.

In both of these cases, robust OLS or simple FGLS are probably close enough.

3.6.2 Autocorrelation

If there is autoregressive autocorrelation of size ρ in the ϵ_{it} (beyond the equicorrelation induced by random effects), then we may run fixed effects or random effects on $y_{it} - \rho y_{i,t-1}$ and $x_{it} - \rho x_{i,t-1}$. However, the errors induced by this often outweigh the benefit if $\hat{\rho} < 0.3$.

3.6.3 Measurement Error

Suppose we have a model in which we can only measure the regressor with error:

$$\begin{aligned} y_{it} &= x^*_{iy}\beta + c_i + \epsilon_{iy} \\ x_{it} &= x^*_{it} + h_{it} \end{aligned}$$

Then, the OLS (non-LSDV) estimate of β based on a regression of y_{it} on x_{it} is biased, with:

$$plim(\hat{\beta}) = \beta \left(\frac{Var(x^*_{it})}{Var(x^*_{it}) + Var(h_{it})} \right) + \frac{Cov(x^*_{it}, c_i)}{Var(x^*_{it}) + Var(h_{it})}$$

If this is a random effects model, $\hat{\beta}$ is always biased toward zero, called *attenuation error*. An estimate of this bias is the *reliability ratio*, which is based on an estimate of $\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_u}$. If there are additional variables or a violation of the random effects assumption, then all the coefficients are biased in unpredictable ways (this is known as *smearing*).

Instrumental variables can be useful in dealing with measurement error.

3.7 Modeling Panel Data with Multiple Equations

3.7.1 Chamberlain's Estimator

We may treat a panel of data as a set of seemingly unrelated set of regressions, with one for each time period. In this, we assume that the X are strictly

exogenous, conditional on the group effects. Assuming we have a balanced panel, and no time-invariant X, we have the model:

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}$$

for each $t = 1, ..., T$. Suppose $\alpha_i = \alpha_0 + \sum_{t=1}^{T} X'_{it}\delta_t + w_i$, so that $Cov(w_i, X_i) = 0$. Then we may rewrite the equations above as:

$$\begin{aligned} y_{it} &= \alpha_0 + \sum_{s=1}^{T} x'_{is}\delta_t + x'_{it}\beta + \epsilon_{it} + w_i \\ &= \alpha_0 + x'_i\pi_t + v_{it} \\ \pi_t &= \begin{pmatrix} \delta_1 \\ ... \\ \delta_t + \beta \\ ... \\ \delta_T \end{pmatrix} \\ E(v_{it}v_{is}|x_i) &= \sigma_w^2 + Cov(\epsilon_{it}, \epsilon_{is}) \end{aligned}$$

We may run seemingly unrelated regressions on the set of equations:

$$\begin{aligned} y_{i1} &= \alpha_0 + x'_i\pi_1 + v_{i1} \\ &... \\ y_{iT} &= \alpha_0 + x'_i\pi_T + v_{iY} \end{aligned}$$

where the covariance matrix of the errors is unrestricted, but the relationship among the coefficients in $\pi_1, ..., \pi_T$ are restricted. We estimate the covariances by:

$$\begin{aligned} \hat{\sigma}_{ts} &= \frac{1}{N}\sum_{i=1}^{N}(y_{it} - x'_i\hat{\pi}_t)(y_{is} - x'_i\hat{\pi}_{is}) \\ \hat{\Sigma} &= \frac{1}{N}(Y - X\hat{\Pi})'(Y - X\hat{\Pi}) \end{aligned}$$

To estimate this system of equations, we may use OLS (which gives $T(T-1)$ different estimates of β and $T-1$ estimates for each δ_t by using all the different estimates; this is consistent but inefficient) or FGLS. We may also use the minimum distance estimator, in which we choose $(\alpha_0, \beta, \delta_1, ..., \delta_T)$ to minimize the distance from the Π based on them to the OLS estimates, based on some weighting matrix, as in GMM; by the strict exogeneity of X, the X from all the time periods are used as instruments.

We may also use maximum likelihood, if we are willing to assume normally distributed errors. In this case, we write $v'_i = y'_i - x'_i\Pi$, and the log likelihood is $\log L - \frac{NT}{2}(\log(2\pi) + \log|\Sigma| + trace(\Sigma^{-1}\Sigma))$. We may minimize this function by setting $\hat{\Sigma}$ equal to the matrix of estimated covariances of the residuals. We may

then write the log likelihood in terms of $\beta, \delta_1, ..., \delta_T$ only (this is called *concentrating the likelihood*), and maximize this. The MLE has the same asymptotic properties as the minimum distance estimator.

3.7.2 Covariance Structures

Suppose we have a balanced panel. Then, we may run seemingly unrelated regressions, with one for each period, and with $E(\epsilon_{it}\epsilon_{js}|X) = \sigma_{ij}I(t=s)$ (this is covariance across individuals, because of individual effects). Then we may use robust OLS with:

$$\begin{aligned} s_{ij} &= \hat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T} e_{it}e_{jt} \\ \hat{Var}(b|X) &= (\sum_{i=1}^{N} X_i'X_i)^{-1}(\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}X_i'X_j)(\sum_{i=1}^{N} X_i'X_i)^{-1} \end{aligned}$$

FGLS can also be based on the s_{ij}, but the matrix, $S = [s_{ij}] = \frac{1}{T}\sum_{t=1}^{T} e_t e_t'$ has $rank(S) \leq \min(N,T)$, which means that FGLS in this form requires that $T \geq N$. Furthermore, FGLS requires $\frac{N(N+1)}{2}$ estimated parameters which may also inflate the variance.

3.8 IV, GMM and Panel Data

If we write the equations in panel data as a system of T equations, each with L instrumental variables, then we have $T \times L$ moment equations. Furthermore, under the assumption that $E(\epsilon_{it}Z_{is}) = 0$ for all t, s, we have T^2L moment equations. With this many moment equations, we often reject the null hypothesis that all the conditions hold.

3.8.1 Hausman and Taylor Method

Suppose we have the model:

$$\begin{aligned} y_{it} &= x1_{it}'\beta_1 + x2_{it}'\beta_2 + z1_i'\alpha_1 + z2_i'\alpha_2 + u_i + \epsilon_{it} \\ E(u_i|x1_{it}, z1_i) &= 0 \\ E(u_i|x2_{it}, z2_i) &\neq 0 \\ E(\epsilon_{it}|x1_{it}, x2_{it}, z1_i, z2_i) &= 0 \\ Var(u_i|x1, x2, z1, z2) &= \sigma_u^2 \\ Var(\epsilon_i|x1, x2, z1, z2) &= \sigma_\epsilon^2 \end{aligned}$$

We may use LSDV to consistently estimate β_1 and β_2. This means that $\ddot{x1}_{it}$ and $\ddot{x2}_{it}$ are valid instruments. By assumption, $z1_i$ is a valid instrument and $\overline{x1}_{i\cdot}$ are valid instruments as well. This provides a set of instruments for $x2$ and $z2$, if there are more variables in $x1$ than in $z2$ (and if there is some partial correlation between the variables in $z2$ and the means of $x1$).

Extending this method, we have Hausman and Talyor's FGLS (or *Generalized IV*) estimator:

1. Find the LSDV estimates of β_1, β_2, and σ_ϵ^2.

2. Let

$$\begin{aligned} Z_i^* &= [Z_{i1}', Z_{i2}']1_T \\ W_i &= \begin{pmatrix} z1_i' & x1_{i1}' \\ ... & ... \\ z1_i' & x1_{iT_i} \end{pmatrix} \\ e_i^* &= (\overline{e}_{i.}, ..., \overline{e}_{i.}) \end{aligned}$$

   Run an IV regression of e^* on Z^* with instruments W to consistently estimate α_1, α_2.

3. Calculate the residual variance, to compute $\hat{\theta}_i = 1 - \sqrt{\frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\epsilon^2 + T_i\hat{\sigma}_u^2}}$.

4. Create new instruments, $w_i^* = (x1_{it}, x2_{it}, z1_i, z2_i) - \hat{\theta}_i(\overline{x1_{i.}}, \overline{x2_{i2.}}, z1_i, z2_i)$, and run a 2SLS regression of $y_{it}^* = y_{it} - \hat{\theta}_i\overline{y_{i.}}$ on them, using the instruments from above $(\ddot{x1}, \ddot{x2}, z1, \overline{x1})$.

This combines random effects FGLS with an IV estimate.

3.8.2 Arellano, Bond and Bover Formulation

In the set-up above, we may also run GMM using the same instruments, $W = [(x1 - \overline{x1_{i.}})', (x2 - \overline{x2_{i.}}), z1_i', \overline{x1_{i.}}']$. We may estimate this in a single step, using:

$$\hat{\delta} = [(\sum X_i'H_i'Z_i)(\sum Z_i'H_i\hat{\Omega}_iH_i'Z_i)^{-1}(\sum Z_i'H_iX_i)]^{-1}(\sum X_i'H_i'Z_i)(\sum Z_i'H_i\hat{\Omega}_iH_i'Z_i)^{-1}(\sum Z_i'H_iy_i)$$

where H_i is M_d^i with the last column replaced by a column of $\frac{1}{T_i}$. As before, we may estimate $\hat{\Omega}_i$ based on a preliminary estimate that uses the identity matrix.

3.9 Dynamic Linear Panel Data Models

Definition If $E(\epsilon_{it}|X_{i1}, ..., X_{iT}) = 0$, then X is *strictly exogenous.* If $E(\epsilon_{it}|X_{i1}, ..., X_{it}) = 0$, then we say that X is *sequentially exogenous.*

Suppose we have the model:

$$\begin{aligned} y_{it} &= x_{it}'\beta + \delta y_{i,t-1} + c_i + \epsilon_{it} \\ E(\epsilon_{it}|x_{it}, c_i) &= 0 \\ E(\epsilon_{it}^2|x_{it}, c_i) &= \sigma_\epsilon^2 \\ E(\epsilon_{it}\epsilon_{is}|x_{it}, c_i) &= 0 \\ E(c_i|X_i) &= g(X_i) \end{aligned}$$

The regressors are no longer exogenous, because $Cov(y_{i,t-1}, c_i) = \frac{\sigma_c^2}{1-\delta} \neq 0$. This makes LSDV inconsistent, since $Cov(y_{i,t-1} - \overline{y_{i.}}, \epsilon_{it} - \overline{\epsilon_{i.}}) \approx \frac{\sigma_\epsilon^2(T-1-T\delta+\delta')}{T(1-\delta)^2}$

The *Anderson-Hsaio IV Estimator* notes that the first differences are of the form:

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + \delta(y_{i,t-1} - y_{i,t-2}) + (\epsilon_{it} - \epsilon_{i,t-1})$$

Notice that $(y_{i,t-1} - y_{i,t-2})$ is correlated with $y_{i,t-2}$ but $(\epsilon_{it} - \epsilon_{i,t-1})$ is not (this trick may be used for variables in X as well). This suggests that $y_{i,t-2}$ is a possible IV. In this case, note that the residuals are $MA(1)$, and we must use GLS, but the matrix is of a known form (2 along the main diagonal, -1 along the two second diagonals).

In the *Arellano and Bond Estimator*, we assume that the X are predetermined at period t, so that $y_{i,t-2}, ..., y_{i1}, x_{i,t-1}, ..., x_{i1}$ are all valid instruments for the equation at period t. Furthermore, if the X are strictly exogenous, then we may use all the periods of X as instruments, and $y_{i,t-2}, ..., y_{i1}, x_{iT}, ..., x_{i1}$ are valid instruments at period t. Note that the number of instruments increases each period. Estimation can be done with a robust error matrix or GLS (to correct for the autocorrelation in the residuals). GMM can also be used.

We may also use the *Hausman and Taylor* method and treat $y_{i,t-1}$ as part of the set of time-varying, endogenous variables.

Ahn and Schmidt propose an estimator of the model

$$y_{it} = \delta y_{i,t-1} + x1_{it}'\beta_1 + x2_{it}'\beta_2 + z1_i'\alpha_1 + z2_i'\alpha_2 + u_i + \epsilon_{it}$$

which uses moment conditions including:

- An initial condition of $y_{i0} = x_{i0}'\lambda + \epsilon_{i0}$, with $E(y_{i0}\epsilon_{it}) = 0$.
- $E(y_{is}(\epsilon_{it} - \epsilon_{i,t-1})) = 0$ for $t = 2, ..., T$ and $s = 0, ..., T-2$.
- $E(\epsilon_{iT}(\epsilon_{it} - \epsilon_{i,t-1})) = 0$ for $t = 2, ..., T-2$.

As before, this leads to a very large number of moment conditions (which can lead to bad finite sample properties).

Panel data with long time series can exhibit the problems of non-stationary time series data. To fix these, first differences or the removal of common trends may be helpful.

3.10 Linear Models with Parameter Heterogeneity

Individual heterogeneity can be caused by:

- Observable differences across individuals.
- Choice strategy, where people have different underlying frames.
- Structural differences, where models differ across individuals

- Parameter differences, where the model is same and the parameters differ across individuals.

Heterogeneity can be discrete (where the population is a mixture of a finite number of types) or continuous (where a random process assigns a parameter vector).

3.10.1 The Random Parameters Model

Suppose we have the model:

$$\begin{aligned} y_{it} &= x_{it}'\beta_i + \epsilon_{it} \\ \beta_i &= \beta + u_i \\ E(u_i|X_i) &= 0 \\ Var(u_i|X_i) &= \Gamma \end{aligned}$$

(If $E(u_i|X_i) \neq 0$, then pooled OLS will not be consistent.) Notice that, if we use pooled OLS, we find consistent but inefficient estimators:

$$\begin{aligned} Y_i &= X_i\beta_i + \epsilon_i = X_i\beta + (X_iu_i + \epsilon_i) \\ E(X_iu_i + \epsilon_i|X_i) &= 0 \\ Var(X_iu_i + \epsilon_i|X_i) &= X_i\Gamma X_i' + \sigma_{\epsilon,i}^2 I \\ b &= \beta + (\sum_{i=1}^N X_i'X_i)^{-1}(\sum_{i=1}^N X_i'(X_iu_i + \epsilon_i)) \\ Var(b|X) &= (\sum_{i=1}^N X_i'X_i)^{-1}Var(X_iu_i + \epsilon_i|X_i)(\sum_{i=1}^N X_i'X_i)^{-1} \\ &= \sigma_\epsilon^2(\sum_{i=1}^N X_i'X_i)^{-1} + (\sum_{i=1}^N X_i'X_i)^{-1}(\sum_{i=1}^N (X_i'X_i)\Gamma(X_i'X_i))(\sum_{i=1}^N X_i'X_i)^{-1} \end{aligned}$$

We estimate the variance of this estimator by $(\sum_{i=1}^N X_i'X_i)^{-1}(\sum_{i=1}^N X_i\hat{w}_i\hat{w}_i'X_i)(\sum_{i=1}^N X_i'X_i)^{-1}$, and use robust standard errors.

Alternatively, we may use GLS, using the matrix $Var(X_iu_i + \epsilon_i|X_i) = \Omega_i = (X_i\Gamma X_i' + \sigma_\epsilon^2)$. We use equation-by-equation estimation to estimate $\hat{\sigma}_{\epsilon,i}^2 = \frac{1}{T_i - K}\sum_{t=1}^{T_i}(y_{it} - x_{it}'b_i)^2$, which is unbiased. We then estimate Γ:

$$\begin{aligned} Var(b_i) &= Var_X(E(b_i|X_i)) + E_X(Var(b_i|X_i)) \\ &= 0 + E_X(\Gamma + \sigma_{\epsilon,i}^2(X_i'X_i)^{-1}) \\ &= \Gamma + E_X(\sigma_{\epsilon,i}^2(X_i'X_i)^{-1}) \end{aligned}$$

Then, we may estimate $\hat{Var}(b_i) = \frac{1}{N}\sum_{i=1}^N (b_i - \bar{b})'(b_i - \bar{b})$ and subtract off the estimate $\hat{\sigma}_{\epsilon,i}^2$. However, the difference may not be positive definite, in which case we may use just $\hat{Var}(b_i)$, or another method (like Bayesian shrinkage or

ML). It turns out that the GLS estimate of β is a weighted average of the OLS slope estimates:

$$\hat{\beta}_{GLS} = \sum_{i=1}^{N} W_i b_{i,OLS} = \sum_{i=1}^{N} \left((\sum_{j=1}^{N} (\Gamma + \sigma^2_{\epsilon,j}(X_j' X_j))) \right)^{-1} (\gamma + \sigma^2_{\epsilon,i}(X_i' X_I')^{-1}) b_{i,OLS}$$

Given the GLS estimates, we estimate β_i as a weighted average of $\hat{\beta}$ and $b_{i,OLS}$:

$$\hat{\beta}_i = A_i \hat{\beta}_{GLS} + (I - A_i) b_{i,OLS}$$

where $A_i = (\Gamma^{-1} + \sigma^2_{\epsilon,i}(X_i' X_i)^{-1})^{-1} \Gamma^{-1}$.

We may also think of nested models, in which the β_i are functions of other variables, just as we may regress estimated fixed effects on time-invariant variables.

More generally (for non-linear models), we may write the random parameters model as:

$$\begin{aligned} f(y_{it}|x_{it}, \beta_{it}) &= g(y_{it}|x_{it}, \beta_i, \theta) \\ f(\beta_i|z_i) &= h(\beta_i, z_i, \Omega) \\ f(y_{it}|x_{it}, z_i, \theta, \Omega) &= \int_{\beta_i} f(y_{it}|x_{it}\beta_i, \theta) h(\beta_i, z_i, \Omega) d\beta_i \end{aligned}$$

A simpler form might specify $\beta_i = \overline{\beta} + \xi z_i + u_i$. There might also be heterogeneity in the variance of the parameters:

$$\begin{aligned} Var(u_{ik}|z_i) &= \phi_{ik} = \phi_k \exp(z_i' \delta_k) \\ Var(u_i|z_i) &= \Phi_i = diag(\phi_{ik}) \end{aligned}$$

where k ranges over the different parameters. We may also choose to model correlation among the parameters.

We may use maximum simulated likelihood to find the likelihood:

$$\log L(\theta, \omega) = \sum_{i=1}^{N} \log \int_{\beta_i} f(y_{it}|x_{it}\beta_i, \theta) h(\beta_i, z_i, \Omega) d\beta_i$$

by integrating out the unobserved parameters, β_i.

Alternatively, if $T_i > K$ for all i, and we have a linear projection of $E(u_i|X_i)$ on X_i, then running OLS or GLS for each observation individually is unbiased, and $\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_i$ is consistent for β, even though $\hat{\beta}_i$ is not consistent (since T is fixed).

For *Partial Fixed Effects*, we allow only some parameters to vary across individuals. Then, we may estimate:

$$\begin{aligned} y_i &= Z_i \alpha_i + X_i \beta + \epsilon_i \\ \hat{\beta} &= (\sum_{i=1}^{N} X_i' M_Z^i X_i)^{-1} (\sum_{i=1}^{N} X_i' M_Z^i y_i) \\ M_Z^i &= I - Z_i (Z_i' Z_i)^{-1} Z_i' \\ \hat{\alpha} &= (Z_i' Z_i)^{-1} Z_i' (y_i - X_i' \hat{\beta}) \end{aligned}$$

We may also think about running regressions involving group means (between estimators) or time means. These require some tricks to make them consistent, if they work at all.

3.10.2 Latent Class Variation

Discrete parameter variation may occur...

- when there is mixing in the population (for example, "zero inflation", where one part of the population is always zero, and everyone else is drawn from some distribution),
- as a discrete approximation to a continuous distribution, or
- when a mixture of normals is used to approximate a non-normal distribution.

Then the population is a mixture of J groups, but group membership is not observed. Within each group, we have the parameters (β_j, σ_j). Before we observe anything, each individual has a probability π_j of being in group j (these are called the *mixing probabilities*). Note that π_j may be constant or may be a function of covariates, in the form:

$$P(class = q|z_i) \quad = \quad \pi_{iq} = \frac{\exp(z_i'\delta_q)}{\sum_s \exp(z_i'\delta_s)}$$

(this assumes a logistic model for class membership).

Then, we have conditional and unconditional densities:

$$\begin{aligned} f(y_{i1}, ..., y_{iT_i}|X_i, \beta_j, \sigma_j) &= \prod_{t=1}^{T_i} f(y_{it}|x_{it}, \beta_j, \sigma_j) \\ f(y_{i1}, ..., y_{iT_i}|X_i) &= \sum_{j=1}^{N} \pi_j \prod_{t=1}^{T_i} f(y_{it}|x_{it}, \beta_j, \sigma_j) \end{aligned}$$

We may then use maximum likelihood to estimate the parameters $\pi_1, ..., \pi_J, \beta_1, ..., \beta_J, \sigma_1, ..., \sigma_J$. Once we have estimated the parameters, we have posterior probabilities for each individual's group membership:

$$P(j|data_i) = \frac{\pi_j \prod_{t=1}^{T_i} f(y_{it}|X_{it}, \beta_j, \sigma_j)}{\sum_{k=1}^{N} \pi_k \prod_{t=1}^{T_i} f(y_{it}|x_{it}, \beta_k, \sigma_k)}$$

In a latent class regression, the likelihoods are of the form:

$$f(y_{it}|j) = \frac{1}{\sigma_j}\phi\left(\frac{y_{it} - x_{it}'\beta_j}{\sigma_j}\right)$$

This is more of a problem when the classes are "close together". Furthermore, we may need to choose the number of classes. Using an information criterion may be the best option for this.

3.10.3 Bayesian random parameters

Bayesians don't need to distinguish between fixed effects and random effects (and similarly for random parameters).

Suppose that parameters vary across individuals. Then, we may use a *hierarchical Bayes model* to estimate the parameters. This may take the form:

$$\begin{aligned} \beta_i &\quad Normal(\overline{\beta}, V_\beta) \\ \overline{\beta} &\quad Normal(\beta^*, aV_\beta) \\ V_\beta^{-1} &\quad Wishart(\nu_0, V_0) \end{aligned}$$

where the Wishart distribution is a multivariate generalization of the Gamma distribution. Note that the first distribution is true of the population, while the latter two are prior distributions. If a and V_0 are large, then the priors are less informative. (In contrast, classical random parameters only includes the first distribution.)

We then use Gibbs sampling for one parameter at a time (the order does not matter).

Bayesian methods can also be used to estimate fixed effects models (though the incidental parameters problem means that the priors for the effects must be informative).

3.11 Non-Linear Models with Panel Data

Many non-linear models are of the form $E(y|x) = m(x, \theta)$, with $\theta \in \Theta$ (the parameter space). We estimate θ based on the observed y, x. For these purposes, we define a *non-linear model* as a model in which we can only define the estimator implicitly. However, we have $h(y, X, \hat{\theta}) = 0$, for some function h.

In panels, there may be relationships among the observations of a single individual in different periods. Then, the correct likelihood for the i^{th} individual is:

$$\log L_i = \log f(y_{i1}, ..., y_{iT_i}|X_i, \theta)$$

which is the joint likelihood for all T_i periods. In many cases, we may use the *pseudo-likelihood* based on the marginal densities, $f(y_{it}|x_{it}\theta)$. Though this will (usually) give a consistent estimate of $\hat{\theta}$, the standard errors will be incorrect. Robust standard errors are:

$$\hat{Var}(g) = \sum_{i=1}^{n}(\sum_{t=1}^{T} g_{it})(\sum_{t=1}^{T} g_{it})'$$

Dynamic Models

If there is a lagged response variable as a predictor, then we have *state dependence*. As before, if there is also an individual-specific intercept, then there is a correlation between the individual effect and the lagged variable. In addition, there is the *initial conditions problem*, where the starting point might affect

future outcomes (especially if there is a strong tendency to stay at one's current state).

One method of dealing with this is:

- Step 1: Write the joint likelihood, conditioning on individual effects, as well as the initial condition and the other predictors.
- Step 2: Assume that u_i depends on the initial state and choose a distribution. For example, $h(u_i|y_{i0}, z_i)$ $Normal(\alpha + \theta y_{i0} + z_i'\delta, \sigma_u^2)$. (The z_i might be group means, for example.)
- Step 3: Integrate out the individual effects. This can be converted into a reduced form for all the other parameters.

4 Limited Dependent Variables Models

In a general limited dependent variable mode1, we begin with a latent regression, $y^* = x'\beta + \epsilon$. Then, we observe the transformed variable, $y = T(y^*)$. This may include *censoring* (top-coding or bottom-coding certain values), *truncation* (omitting certain values), or *sample selection* (choosing only some of the data). In all of these cases, OLS based on y tends to be biased.

4.1 Binary Choice Models

Suppose we have a utility model, $U = \alpha + \beta' x + \epsilon$, and we observe a choice Y, which is either 0 or 1. Then, we model $Y = I(U > 0) = P(\epsilon > \alpha + \beta' x)$. This is a *binary choice model*. The probability function depends on the assumed distribution for ϵ; it is a probit model if $\epsilon \sim Normal$ and a logit model if ϵ has a logistic distribution. (It may also be Gompertz, semiparametric, or something else.) This choice will affect the results (at least slightly). The coefficients of the logit model are generally 1.6 times the coefficients for the probit model (because the logistic density evaluated at 0 is about 1.6 times the normal density evaluated at 0). However, the marginal effects are approximately the same (usually). The likelihood is derived as:

$$\begin{aligned} P(y_i = 1|X_i) &= F(X_i'\beta) \\ f(y_i|x_i) &= (1 - F(X_i'\beta))^{1-y_i} F(x_i'\beta)^{y_i} \\ \frac{\partial}{\partial \beta} \log L &= \sum_{i=1}^{N} \left(\frac{y_i}{F(X_i'\beta)} - \frac{1 - y_i}{1 - F(X_i'\beta)} \right) \end{aligned}$$

If the density is symmetric about 0, we may simplify these expressions by noting that $1 - F(x'\beta) = F(-x'\beta)$. Note that the Hessian is negative definite at all points for the probit and logit models, so that the model is globally concave in the parameters and there is a global maximum.

The marginal effect of a dummy variable in this model is:

$$\hat{\delta} = P(y_i = 1|x_i, d_i = 1) - P(y_i = 1|x_i, d_i = 0)$$

The marginal effect of a continuous variable is:

$$\hat{\delta} = \frac{\partial}{\partial x} F(\alpha + \beta' x) = f(\alpha + \beta' x)\beta$$

Note that both of these depend on the location. Partial effects are usually computed at the means (which is simple and has well-defined inference) or is the average of the partial effects over all observations (which has only asymptotic standard errors). Note that the marginal effects tend to have large standard errors, since they include the standard errors of all the parameters; this means that coefficients might be significant while the corresponding partial effects are not. To find the standard errors for marginal effects, we may use the delta method, $\hat{AsyVar}(\hat{\delta}) = G(\hat{\beta}, x)\hat{V}G(\hat{\beta}, x)'$. We may also use the method of Krinsky and Robb, which uses simulation.

The mean of the predicted probabilities is always equal to the proportion of successes in the data:

$$\hat{\overline{F}} = \frac{1}{N}\sum_{j=1}^{N} F(\hat{\beta}' x_j) = \frac{1}{N}\sum_{j=1}^{N} y_j$$

4.1.1 Model Fits and Hypothesis Testing

Definition The *Likelihood Ratio Index (LRI)* by McFadden or *Pseudo-R-Squared*, is defined as $1 - \frac{\log L_0}{\log L}$, where L_0 is the likelihood of the model with all the slopes set to 0, and L is the likelihood of the model of interest.

We may also measure the fit of a model by checking how many outcomes would have been correctly predicted. One method, by Cramer, sets:

$$\hat{\lambda} = mean(\hat{F}|y = 1) - mean(\hat{F}|y = 0)$$

This is the difference in the estimated probabilities for the successes and failures. Alternatively, we may compute the proportion of observations correctly predicted using a prediction rule (such as predicting a success when $\hat{F} > \frac{1}{2}$, or equivalently when $\hat{\beta}' x > 0$). These measures also show the usefulness of additional variables. Note that probit and logit do not try to maximize the number of correct predictions; an estimator which does this is called an *m-score* estimator, and it converges as $\sqrt[3]{n}$.

To test nested hypotheses, we may use likelihood ratio tests, Lagrange multiplier tests, and Wald tests, and everything is only asymptotic now. (Also, the order from the linear case, with $Wald > LR > LM$, might not hold anymore.)

4.1.2 Heteroskedasticity

Suppose $Var(\epsilon_i) = \exp(\gamma' z_i)^2$. Then, the probit model is given by:

$$P(y_i = 1|x_i, z_i) = \Phi\left(\frac{\beta' x_i}{\exp(\gamma' z_i)}\right)$$

This changes the functional form of the model and makes the partial effects more complicated. It is not enough to fix the standard errors, since estimating the model without the denominator would be inconsistent.

4.1.3 Endogenous Regressors

Suppose we have $y^* = \beta' x + \gamma z + \epsilon$, with $E(\epsilon|z) \neq 0$. Then, we may write $z = w'\delta + u = x'\delta + h'\delta + u$, where $Cov(h, \epsilon) = 0$. Then, the reduced form is:

$$\begin{aligned} y^* &= x'\beta + \gamma(w'\delta + u) + \epsilon \\ &= x'(\beta + \gamma\delta_1) + h'\gamma\delta_2 + (\gamma u + \epsilon) \end{aligned}$$

Using probit, we estimate: $\frac{\beta+\gamma\delta_1}{\sqrt{1+\gamma^2\sigma_u^2+2\gamma\rho\sigma_u}}, \frac{\gamma\delta_2}{\sqrt{1+\gamma^2\sigma_u^2+2\gamma\rho\sigma_u}}$, where $\rho = Corr(u, \epsilon)$. Though we can estimate $\delta_1, \delta_2, \sigma_u^2$ from the OLS regression of z on w, we cannot estimate ρ.

As a specific case, we may have an endogenous binary variable, with the model:

$$\begin{aligned} y &= 1(x'\beta + \gamma z + \epsilon > 0) \\ z &= 1(x'\delta + u > 0) \\ Cov(\epsilon, u) &= \rho \end{aligned}$$

Then, we may analyze y and z jointly, since $P(y = 1, z = 1) = P(y = 1|z = 1)P(z = 1)$ is a bivariate probit model.

4.1.4 Effects Models in Binary Choice

Suppose utility is of the form $y_{it}^* = \alpha + \beta' x_{it} + \gamma' z_i + u_i + \epsilon_{it}$, where x_{it} are the attributes of a particular decision, and z_i are individual characteristics. Since the scale is not observed, we set $Var(\epsilon_{it}) = 1$. In binary choice, one option is chosen if $y_{it}^* > 0$ (and the other if the utility is less than 0). This leads to the probability model:

$$\begin{aligned} P(y_{it} = 1) &= P(y_{it}^* > 0) \\ &= P(\epsilon_{it} > -(\alpha + \beta' x_{it} + \gamma' Z_i + u_i)) \end{aligned}$$

In the case where we have strict exogeneity and unobserved effects which are uncorrelated with the right-hand-side variables, we may estimate a pooled model:

$$\begin{aligned} P(y_{it} = 1) &= P(\epsilon_{it} > -(\alpha + \beta' x_{it} + \gamma' Z_i + u_i)) \\ &= P(\epsilon_{it} + u_i > -(\alpha + \beta' x_{it} + \gamma' Z_i)) \\ &= F(\frac{x_{it}'\beta}{\sqrt{1 + \sigma_u^2}}) \\ &= F(x_{it}'\delta) \end{aligned}$$

Note that the coefficient estimates are attenuated and no longer consistent. However, the partial effects are less attentuated, because $f(x_{it}'\delta) > f(x_{it}'\beta)$. Thus, we worry less about the partial effects (which are what we care about anyway).

In pooled estimation, we are now using a partial pseudo-log-likelihood for estimation:

$$\text{“}\log L\text{”} = \sum_i \sum_t = (1 - y_{it}) \log(1 - F(x_{it}'\beta)) + y_{it} \log(F(x_{it}'\beta))$$

This will work if the marginals for y_{it} are correct; however, we should be considering joint likelihoods of $y_{i1}, ..., y_{iT}$ instead. This is an M-estimator, so it is consistent, even if it isn't efficient. A “panel probit model” can be estimated using methods like SUR:

$$\begin{aligned} y_{it}^* &= x_{it}'\beta + \epsilon_{it} \\ y_{it} &= 1(y_{it}^* > 0) \\ \begin{pmatrix} \epsilon_{i1} \\ ... \\ \epsilon_{iT} \end{pmatrix} & \quad Normal\left(0, \begin{pmatrix} 1 & \rho_{12} & ... & \rho_{1T} \\ \rho_{12} & 1 & ... & \rho_{2T} \\ ... & ... & ... & ... \\ \rho_{1T} & \rho_{2T} & ... & 1 \end{pmatrix}\right) \end{aligned}$$

(We assume that the variance is 1 because only the sign matters.)

Full information maximum likelihood would be written as:

$$\log L = \sum_{i=1}^{n} \log Prob(y_{i1}, ..., y_{iT})$$

Estimation with this method is hard.

We may also use GMM. The obvious set of orthogonality conditions is:

$$E((y_{it} - \Phi(x_{it}'\beta))x_{it}) = 0$$

Under strict exogeneity, we have TK orthogonality conditions for β, since we may use all periods as orthogonality conditions:

$$E((y_{it} - \Phi(x_{it}'\beta))x_{is}) = 0$$

(This is probably overkill.)

To implement GMM:

1. Pool the data and estimate $\hat{\beta}$ to get the initial weighting matrix:

$$W = \frac{1}{n^2} \sum_{i=1}^{n} \begin{pmatrix} (y_{i1} - \Phi(x_{i1}'\beta))x_{i1} \\ ... \\ (y_{iT} - \Phi(x_{iT}'\beta))x_{iT} \end{pmatrix} \begin{pmatrix} (y_{i1} - \Phi(x_{i1}'\beta))x_{i1} & ... & (y_{iT} - \Phi(x_{iT}'\beta))x_{iT} \end{pmatrix}$$

2. Minimize the GMM criterion, $q = \overline{g}(\beta)' W^{-1} \overline{g}(\beta)$, where $\overline{g}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} (y_{i1} - \Phi(x_{i1}'\beta))x_{i1} \\ ... \\ (y_{iT} - \Phi(x_{iT}'\beta))x_{iT} \end{pmatrix}$.

Using a random effects type of model may be inaccurate when $Cov(x_{it}, u_i) \neq 0$, but it is easier to compute and does not suffer from the incidental parameters problem. In this case, we have the model:

$$U_{it} = \alpha + \beta' x_{it} + (\epsilon_{it} + \sigma_v v_i)$$

where $Var(v_i) = 1$. This leads to a likelihood of:

$$\begin{aligned} f(y_i|v_i) &= \prod_{t=1}^{T} F(\alpha + \beta' x_{it} + \sigma_v v_i) \\ L &= \prod_{i=1}^{N} \left(\int_{-\infty}^{\infty} \prod_{t=1}^{T} F(\alpha + \beta' x_{it} + \sigma_v v_i) \right) g(v_t) dv_i \\ \log L &= \sum_{i=1}^{N} \log \left(\int_{-\infty}^{\infty} \prod_{t=1}^{T} F(\alpha + \beta' x_{it} + \sigma_v v_i) \right) g(v_t) dv_i \end{aligned}$$

To evaluate this likelihood, we must assume a distribution, g, for the effects (usually, the standard normal distribution). We then must maximize this with respect to α, β, σ_v. Since there is no closed form for the integral (in realistic cases), we replace the integral by a sum using Hermite quadrature or simulation. In this model, $\rho = \frac{\sigma_v^2}{1+\sigma_v^2}$ is the off-diagonal correlation.

Using a fixed effects type of model, we will always have inconsistent estimates because of the incidental parameters problem. (In general, people assume β is biased upward in this case, with a bigger bias when T is small.) Note that differencing the data will not remove this problem.

To estimate the model anyway, we may do conditional estimation based on sufficient statistics. In this case, we assume that $f(y_{i1}, ..., y_{iT}|g(Y))$ does not depend on the fixed effects. This leads to a conditional logit:

$$P(Y_{i1}, ..., Y_{iT}| \sum_{t=1}^{T} y_{it} = S_i, X) = \frac{\exp(\sum_t y_{it} x'_{it} \beta)}{\sum_{\sum d_{it} = S_i} \exp(\sum_t d_{it} x'_{it} \beta)}$$

That is, we condition on the sum and see which values of the parameters make the order of the results most likely. In this estimation method, only individuals that have both successes and failures contribute to the estimates (otherwise, the conditional probability is 1). Once β has been estimated, we estimate the fixed effects from:

$$0 = \sum_{t=1}^{T} (y_{it} - p_{it}) = \sum_{t=1}^{T} \left(y_{it} - \frac{\exp(\alpha_i + \hat{\beta}' x_{it})}{1 + \exp(\alpha_i + \hat{\beta}' x_{it})} \right)$$

However, the estimates of the fixed effects are not consistent (and do not even exist when y_{it} is constant for all t). Values of the fixed effects may be necessary to calculate the partial effects.

To estimate standard errors, we may use the diagonal elements of $(-H)^{-1}$, which ignores the problem, or use the cluster estimator.

In multinomial (panel) logit, we have the model:

$$P(choice = j|x_{itj}) = \frac{\exp(\alpha_j + \beta_j' x_{itj})}{\sum_{j=1}^{J_{it}} \exp(\alpha_j + \beta_j x_{itj})}$$

In this model, we have *independence from irrelevant altervatives*, where P_{ijt}/P_{ikt} depends only on j, k. (This may not hold in some cases, like the red bus/blue bus example.) In applications of this model, we may be interested in elasticities, which are the effect of an increase in the price of one brand on the probability of choice of each other brand. This is given by:

$$\frac{\partial \log P_{ijt}}{\partial \log x_{k,imt}} = \beta_k(I(j = m) - P_j)$$

Note that the relative elasticities of the brands with no price increase does not change (because of the independence from irrelevant alternatives).

4.2 Ordered Probability and Interval Censored Data

In the *ordered probability model*, we have the latent variable, $y_{it}^* = x_{it}'\beta + \epsilon_{it}$, and we then observe:

$$y_{it} = \begin{cases} 0 & \text{if } y_{it}^* \leq 0 \\ 1 & \text{if } 0 \leq y_{it}^* \leq \mu_1 \\ ... & ... \\ J-1 & \text{if } \mu_{J-1} \leq y_{it}^* \leq \mu_J \\ J & \text{if } \mu_J \leq y_{it}^* \end{cases}$$

(The constant term takes care of the normalization to 0.) We must estimate the μ_j; they are not observed. Thus model assumes that μ_j is constant across all individuals. However, the distance between the cutoffs need not be constant. Then,

$$\begin{aligned} P(y_{it} = j) &= P(\mu_{j-1} \leq y_{it}^* \leq \mu_j) \\ &= P(\mu_{j-1} \leq x_{it}'\beta + \epsilon_{it} \leq \mu_j) \\ &= P(\epsilon_{it} \leq \mu_j - x_{it}'\beta) - P(\epsilon_{it} \leq \mu_{j-1} - x_{it}'\beta) \end{aligned}$$

After we assume a particular distribution for ϵ, we find a likelihood function:

$$\log L = \sum_{i=1}^{N} \sum_{j=0}^{J} 1(y_{it} = j) \log(P(y_{it} = j))$$

We must normalize this by setting $Var(\epsilon_{it}) = 1$. Then we estimate β and the $J - 1$ cutoffs.

The coefficients are not easy to interpret, but we can compute the marginal effects on the probabilities, find predictions, and find measures of fit. The marginal effect is given by:

$$\frac{\partial}{\partial x} P(y_{it} = j) = \beta(f(\mu_j - x_{it}'\beta) - f(\mu_{j-1} - x_{it}'\beta))$$

If β_k is positive, then the marginal effect of x_k on the probability of the smallest outcome is always negative and on the probability of the largest outcome is always positive. However, the effect on the intermediate probabilities is indeterminate; it depends on $f(\mu_j - x'_{it}\beta) - f(\mu_{j-1} - x'_{it}\beta)$. The marginal effects always sum to zero (to keep the total probability equal to 1), and they switch sign exactly once. For prediction, we may choose the cell with the highest probability.

To add individual heterogeneity, we may allow $\mu_1, ..., \mu_J$ to depend on the variables. However, μ_{ij} cannot depend linearly on the same variables, since we cannot identify effects if we have $P(z'_i\delta_j - x'_i\beta \leq \epsilon)$. We must also ensure that $\mu_j \leq \mu_{j+1}$ for all individuals; one way to do this is to have $\mu_{ij} = \exp(\theta_j + z'_i\delta)$. This also allows variables to overlap, since the relationship is nonlinear.

If there is omitted heterogeneity, there is attenuation bias in the coefficients:

$$\begin{aligned} y^*_{it} &= x'_{it}\beta + u_i + \epsilon_{it} \\ P(y_{it} = j) &= P\left(\epsilon_{it} \leq \frac{\mu_j - x'_{it}\beta}{\sqrt{1+\sigma^2_u}}\right) - P\left(\epsilon_{it} \leq \frac{\mu_{j-1} - x'_{it}\beta}{\sqrt{1+\sigma^2_u}}\right) \end{aligned}$$

There is likely to be a smaller impact on the marginal effects.

4.2.1 Zero inflated ordered probit

Suppose "non-participants" always report zero, while "participants" report $0, ..., J$. Then, we may wish to combine a probit model for participation, p_{it}, with an ordered probit model for the participants, y_{it}. Then,

$$\begin{aligned} P(y_{it} = 0) &= P(p_{it} = 0) + P(p_{it} = 1)P(y_{it} = 0|p_{it} = 1) \\ P(y_{it} = j) &= P(p_{it} = 1)P(y_{it} = j|p_{it} = 1) \end{aligned}$$

The two models may have some variables in common.

4.2.2 Interval censored data

For the *interval censored data model*, the cutoffs between the groups are known to be $a_0, ..., a_J$. In this case, the model is:

$$\begin{aligned} y^*_{it} &= x'_{it}\beta + \epsilon_{it} \\ y_{it} &= j \text{ if } a_{j-1} \leq y^*_{it} \leq a_j \\ P(y_{it} = j) &= P(\frac{a_j - x'_{it}\beta}{\sigma} > \epsilon_{it}) - P(\frac{a_{j-1} - x'_{it}\beta}{\sigma} > \epsilon_{it}) \end{aligned}$$

Maximum likelihood estimation basically runs a regression based on $E(y^*_{it}|y_{it} = j)$ instead of y^*_{it} directly. This changes the estimation procedure and allows us to estimate the scale of ϵ.

4.3 Censoring and Truncation

In the censoring model, we observe the transformation of the latent variable:

$$T(y^*) = \begin{cases} 0 & \text{if } y^* \leq 0 \\ y^* & \text{otherwise} \end{cases}$$

There are other forms of censoring, where there is a maximum value or a different minimum, in which case we may transform the observed variable to return to this model. There may also be censoring at both ends or person-specific censoring, which are more complicated. This is similar to a corner solution (where some people choose a value and other just choose zero) but the underlying theory is different.

In the Tobit model, we have $Y^* = X'\beta + \epsilon$ and we assume that ϵ $Normal(0, \sigma^2)$. In this case, we have the conditional means:

$$\begin{aligned} E(Y^*|X) &= X'\beta \\ E(Y|Y>0,X) &= X'\beta + \sigma\frac{\phi(X'\beta/\sigma)}{\Phi(X'\beta/\sigma)} \\ E(Y|X) &= P(Y=0|X) * 0 + P(Y>0|X)E(Y|Y>0,X) \\ &= \Phi(\frac{X'\beta}{\sigma})(X'\beta + \sigma\frac{\phi(X'\beta/\sigma)}{\Phi(X'\beta/\sigma)}) \\ &= \Phi(X'\beta/\sigma)X'\beta + \sigma\phi(X'\beta/\sigma) \end{aligned}$$

This shows why OLS is biased: the coefficients are attenuated. In general, the slopes in OLS approximate the derivatives:

$$\frac{\partial}{\partial x}E(Y|X) = \beta\Phi(x'\beta/\sigma)$$

Predicting y^* is irrelevant, though either $E(Y|X)$ or $E(Y|Y>0,X)$ may be useful. In addition, we consider generalized residuals (Cheshire and Irish) around zero:

$$gr_i = \frac{\partial \log L_i}{\partial \beta_0} = \begin{cases} y_i - x_i\beta & y_i > 0 \\ g(x_i,\beta,\sigma) & y_i = 0 \end{cases}$$

As a form of R^2, we may generate predictions from the model and then find the squared correlation between the predictions and the true values.

For estimation, we have the log likelihood:

$$\log L = \sum_{i=1}^{n}\left(1(y_i=0)\log\Phi(-x_i'\beta/\sigma) + 1(y_i>0)\log\left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right)\right)$$

To simplify estimation, we may use the Olsen transformation of the variables,

$\theta = 1/\sigma$ and $\gamma = \beta/\sigma$. This yields the likelihood and derivative:

$$\begin{aligned}
\log L &= \sum_{i=1}^{n}(1(y_i = 0)\log\Phi(-x_i'\gamma) + 1(y_i > 0)\log(\theta\phi(\theta y_i + x_i'\gamma))) \\
\frac{\partial}{\partial\gamma}\log L &= \sum_{i=1}^{n}(1(y_i = 0)\phi(x_i'\gamma)/\Phi(x_i'\gamma) - 1(y_i > 0)e_i)x_i \\
\frac{\partial}{\partial\theta}\log L &= \sum_{i=1}^{n}1(y_1 > 0)(\frac{1}{\theta} - e_i y_i)
\end{aligned}$$

This simplifies the Hessian; we may then use the delta method to work out the standard errors for the original coefficients.

The marginal effects are given by:

$$\begin{aligned}
\frac{\partial}{\partial x}E(y|x) &= \beta\Phi(x'\beta/\sigma) \\
\frac{\partial}{\partial x}E(y|x, y > 0) &= \beta(1 - \lambda(a)a + \lambda(a)^2)
\end{aligned}$$

where $\lambda(a) = \frac{\phi(x'\beta/\sigma)}{\Phi(x'\beta/\sigma)}$. Note that the coefficient is attenuated in both cases. This means that discarding the *limit data* (the zero observations) will not produce a consistent estimate of β.

The McDonald and Moffit marginal effects are given by:

$$\frac{\partial}{\partial x}E(y|x) = P(y > 0|x)\frac{\partial}{\partial x}E(y|x, y > 0) + E(y|x, y > 0)\frac{\partial}{\partial x}P(y > 0|x)$$

We may estimate β/σ using the probit model of whether y is non-zero. This also provides a specification test for the model. Another specification test is a truncated regression, omitting the zero values. One may specify this more generally as a two part model, where a probit is first fit, and then a truncated regression is fit to the remaining data. The Tobit model is a restriction of this two-part model that forces certain coefficients to agree; this provides another specification test.

If we have individual-specific effects that are orthogonal to X, then the estimates of β are attenuated, but the marginal effects, $\frac{\beta}{\sigma_\epsilon^2+\sigma_c^2}\Phi(\frac{x'\beta}{\sigma_\epsilon^2+\sigma_c^2})$ are consistently estimated. Thus, pooling will work, but cluster estimators will be necessary for standard errors.

Random effects including the group means (in an extension of the Mundlak method to probit models) can allow for a more general model.

5 Count Data

The Poisson model for count data is:

$$\begin{aligned}
P(y_i = j|x) &= \frac{1}{j!}\exp(-\lambda_i)\lambda_i^j \\
\lambda_i &= E(y_i|x_i) = \exp(x_i'\beta)
\end{aligned}$$

If the counts are observed over intervals of different lengths, then the log of the length of time should be added, with a coefficient of 1. In this case, the partial effects are $\frac{\partial}{\partial x_i} = \lambda_i \beta$.

In the Poisson model, we must have the variance equal to the mean. Suppose $Var(y|x) \neq E(y|x)$. Then, we have *overdispersion*. This may occur because of misspecification or because of omitted heterogeneity. If we actually have $\lambda = \exp(x'\beta + u)$ where e^u *Gamma*, then $y|x$ has a negative binomial distribution. The dispersion parameter is $\frac{1}{\alpha}$, where $f(e^u) = \frac{\alpha^\alpha}{\Gamma(\alpha)} \exp(-\alpha u) u^{\alpha-1}$ (this is a Gamma distribution with mean 1).

To test for overdispersion, we may regress $(y_i - \lambda_i)^2$ on λ_i and the other variables and test that (1) the coefficient on λ_i is 1, and (2) the coefficients on all the other variables are insignificant. We may also test against the specific hypothesis of the negative binomial model, which has $Var(y|x) = E(y|x) + \sigma^2 E(y|x)^2$ by inserting the sample moments and testing whether $\sigma^2 = 0$. (This is more powerful against the alternative of the Negative Binomial.)

Poisson data has consistent pseudo-likelihoods. That is, we may estimate using the likelihood function $f(y_{it}) = \exp(-\lambda_{it}) \lambda_{it}^{y_{it}} / y_{it}!$, with $\lambda_{it} = \exp(\beta' x_{it})$, even though we should really have $\lambda_{it} = \exp(\beta' x_{it} + \epsilon_{it})$ with $\exp(\epsilon_{it})$ $\Gamma(1, \theta)$. In this case, the true likelihood is negative binomial, but $\hat{\beta}$ is consistent (though inefficient). In this case, we should use a sandwich estimator for the standard errors, since the information matrix will be wrong.

5.1 Zero Inflation Poisson Model

In the Zero Inflation Poisson (ZIP) model, we have:

$$\begin{aligned} P(p_i = 1) &= \frac{\exp(z_i'\alpha)}{1 + \exp(z_i'\alpha)} \\ y_i &\begin{cases} 0 & \text{if } p_i = 0 \\ Poisson(\lambda_i) & \text{if } p_i = 1 \end{cases} \end{aligned}$$

Note that the Poisson model is not nested in the ZIP model, since that would require $\alpha = \pm\infty$. To choose between the two models, we use the *Vuong Statistic*. To do this, we find the log likelihoods of the two models for each individual (???) and compute:

$$\begin{aligned} a_i &= \log L_{i0} - \log L_{i1} = \log(f_0(y_i|x_i, \theta_0)/f_1(y_i|x_i, \theta_1)) \\ V &= \frac{\overline{a}}{s_a/\sqrt{n}} \end{aligned}$$

Under some conditions, V is normally distributed. Therefore, if $V > 1.96$, we choose model 0, if $V < -1.96$, we choose model 1, and otherwise the models are not significantly different.

5.2 Panel Data

The conditional Poisson, based on $\sum Y_i$, is identical to the unconditional (brute force) Poisson, so there is no incidental parameters problem using fixed effects.

For random effects, using heterogeneity with a gamma distribution allows for estimation with the negative binomial. One can also use normally distributed heterogeneity, but estimation is more complicated.

6 Duration/Survival Models

Suppose we observe either the time until an event occurred, T, or that an event has not happened yet. In this case T is the random variable, with density $f(t)$, cdf $F(t)$, and a *survival function*, $S(t) = 1 - F(t)$. For small $\Delta > 0$, define:

$$\begin{aligned} h(t) &= P(T \in [t, t+\Delta] | T > t) \\ &= \frac{F(t+\Delta) - F(t)}{1 - F(t)} \\ \lim_{\Delta \to 0} h(t) &= \lambda(t) = \frac{f(t)}{S(t)} \end{aligned}$$

We called $\lambda(t)$ the *hazard function*. Then

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t)) \\ F(t) &= 1 - \exp(-\int_0^t \lambda(s)ds) \\ \frac{dF}{dt} &= \lambda(t)\exp(-\int_0^t \lambda(s)ds) \end{aligned}$$

for any $F()$.

If $\lambda(t) = \lambda$ is constant, then duration does not matter, and we have the *exponential model*:

$$\begin{aligned} S(t) &= \exp(-\lambda t) \\ F(t) &= 1 - \exp(-\lambda t) \\ f(t) &= \lambda \exp(-\lambda t) \end{aligned}$$

Otherwise, we have *duration dependence* in survival (for example, negative duration dependence means that the longer that it has been before the event has happened, the longer we expect to wait). One model with duration dependence is the *Weibull distribution*, where $\lambda(t) = \lambda p(\lambda t)^{p-1}$ (the exponential is a special case with $p = 1$). Other distributions include the log-logistic, the log-normal, and the Gompertz distributions.

If the event has not occurred yet, then we have *censoring*. We may transform this into an analog of the Tobit model.

In a *split population model*, the event (usually failure, in this case) may never occur; this is a latent class model.

In an *accelerated failure model*, a set of covariates modeifes the hazard function. In the Weibull case, we have:

$$\lambda(t|x) = \exp(x'\beta)p(\exp(x'\beta)t)^{p-1}$$

More generally, in the *proportional hazards model*, we have $\lambda(t|x) = g(x)\lambda(t)$, so all the hazard functions are proportional to some baseline. We estimate this with maximum likelihood:

$$\begin{aligned} g(t|x) &= (\frac{f(t|x)}{S(t|x)})^{1(\text{not censored})} S(t|x) = (\lambda(t|x))^{1(\text{not censored})} S(t|x) \\ \log L &= \sum_{i=1}^{n} 1(\text{not censored}) \log \lambda(t_i|x_i) + \log S(t_i|x_i) \end{aligned}$$

It is possible that x is observed multiple times between the starting time and the event (so that a different number might be observed for each individual). To use such data, one might create an observation for each time x is observed and make the new observations censored if failure did not occur before x was observed again.

If there is unobserved heterogeneity, then we may have $\lambda(t|x,u) = u\lambda(t|x)$. In the case of Weibull proportional hazards, we have $\lambda(t|x,\epsilon) = \exp(x'\beta)\exp(\epsilon)\lambda(t)$. If $\exp(\epsilon)$ *Gamma*, then there is a closed form for $f(t|x)$. In other cases, heterogeneity may be estimated numerically.

7 Sample Selection

7.1 Linear Models

Suppose we have a regression model, $y^* = X'\beta + \epsilon$, and $d^* = Z'\gamma + u$, where $y = y^*$ is observed if $d^* > 0$ and y is not observed otherwise. We assume that we observe X, Z for all individuals. Then,

$$E(y^*|X, d = 1) = X'\beta + E(\epsilon|X, d = 1) = X'\beta + E(\epsilon|X, u > -Z'\gamma)$$

This shows that sample selection is a problem if the error term in the regression equation is correlated with the error term in the selection model. Note that X and Z may contain some or all of the same variables. Assuming that the probability of sample selection is non-linear, everything is still identified, even if the two sets of variables are identical. This selection is not based on the value of y and is therefore called *incidental truncation*.

A special case is *Heckman's Model*, where $(\epsilon_i, u_i) \sim Normal(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$. In this case, $E(y_i^*|x_i, d_i = 1) = x_i'\beta + \rho\sigma\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}$, and the second term would be an omitted variable in the original model, unless $\rho = 0$.

We may estimate Heckman's Model using two-step-least-squares:

- Estimate the probit model, $d_i^* = z_i'\gamma + u_i$, $d_i = 1(d_i^* > 0)$. Based on the estimates, create $\hat{\lambda}_i = \phi(z_i'\hat{\gamma})/\Phi(z_i'\hat{\gamma})$. This is called *Heckman's* λ.
- Regress Y on $X, \hat{\lambda}$.
- Fix the standard errors to account for the fact that $\hat{\lambda}$ had to be estimated.

- One may also estimate ρ and σ based on the coefficient $\hat{\lambda}$ in the second step and on the residual variance.

Though this method is easy to understand, the two-step nature makes it inefficient. As before, X and Z may overlap, since λ is a non-linear function of Z. However, if they are identical, $\hat{\lambda}_i$ and X will be highly correlated.

Full information maximum likelihood (FIML) is the efficient method of estimating sample selection models. For this, we compute:

$$\log L = \sum_{d=0} \log \Phi(-z_i'\gamma) + \sum_{d=1} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(y_i - x_i'\beta)^2}{2\sigma^2} \right) \Phi\left(\frac{z_i'\gamma + \rho(y_i - x_i'\beta)/\sigma}{\sqrt{1-\rho^2}} \right) \right)$$

This is an efficient estimator. To simplify this expression, we may use the Olsen reparameterization:

$$\begin{aligned} \theta &= \frac{1}{\sigma} \\ \delta &= -\frac{\beta}{\sigma} \\ \tau &= \frac{\rho}{\sqrt{1-\rho^2}} \end{aligned}$$

In this case, the inverse Mills ratio is irrelevant (only the conditional mean, not the likelihood, depends on it).

To extend this to panel data, we may run a random effects probit (with effect variance η) combined with a fixed effects regression. In this case, the marginal likelihood is:

$$\log L = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} \Phi \left((2d_{it} - 1) \frac{z_{it}'\gamma + \Delta_{it} + u_{i1} + d_i u_{i2}}{\sqrt{\sigma_\eta^2 (1 - d_{it}\rho^2)}} \right) f(u_{i1}, u_{i2}) du_{i1} du_{i2}$$

where $u_i 1, u_i 2$ are the error terms in the two equations and $\Delta_{it} = \frac{\rho}{\sigma_\epsilon} d_{it}((y_{it} - \overline{y_i}) + (x_{it} - \overline{x_i})'\beta)$. We may also use the Mundlak method to fix the assumptions of the random effects probit, use a similar method to add random parameters, or use two-step methods.

This may also be used for a *treatment effects model*, where individual are assigned to treatment groups based on $d_i^* = z_i'\gamma + u_i$ and we then observe $y_i^* = x_i'\beta + \delta d_i + \epsilon_i$, but d_i is endogenous. In this case, we have:

$$\begin{aligned} E(y_i^* | x_i, d_i = 1) &= x_i'\beta + \rho\sigma \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)} \\ E(y_i^* | x_i, d_i = 0) &= x_i'\beta + \rho\sigma \frac{-\phi(z_i'\gamma)}{\Phi(z_i'\gamma)} \end{aligned}$$

and a similar procedure (but where everyone is observed) can be used.

This may also be used for *binary data*, where we observe pairs of decisions. In this case, we have the model:

$$\begin{aligned} d_i^* &= z_i'\gamma + u_i \\ d_i &= 1(d_i^* > 0) \\ y_i^* &= x_i'\beta + \epsilon_i \\ y_i &= 1(y_i^* > 0) \end{aligned}$$

7.2 Survival Analysis

Suppose we have the proportional hazard function, $\lambda(t, x|d = 1) = h(t)f(x'\beta + \epsilon)$, with the sample selection model $d^* = z'\gamma + u$ with selection if $d^* > 0$. Suppose (u, ϵ) $Normal$ with correlation ρ. We only observe T if $d^* > 0$.

In general, we need not assume normality. However, we must know the distribution of $u_i|\epsilon_i$.

Because of the relationship between the random component in the selection model and the random component in the hazard function, the hazard function differs depending on whether an individual was selected or not. This is an omitted variable problem.

This model of sample selection will not allow selection to change the sign of the Weibull model's duration dependence.

8 Stochastic Frontiers Models

Suppose that production of Y depends on two inputs, X_1 and X_2. The producer might not be efficient (based on the isoquant); production might be require only $\theta(X_1, X_2)$ for the same amount. Alternatively, the ratio of the inputs might not be efficient either (since the production should be at the tangent of an isocost curve and an isoquant). So the cost might only need to be αX_A. This is *allocative inefficiency.*

Definition The *technical efficiency* of an amount of output and inputs is given by:

$$TE(y, x) = \min\{\theta : \theta x \in L(y)\}$$

where $L(y)$ is the isoquant of y. Note that $TE(y, x) \leq 1$.

We model production in the i^{th} firm by:

$$\begin{aligned} y_i &= f(x_i, \beta)TE_i \\ \ln y_i &= \ln f(x_i, \beta) - u_i \end{aligned}$$

Because technical efficiency is constrained to be less than one, we must have $u_i > 0$. Thus, this is not a regression. We could add a constant term (equal to $E(u_i)$) so that we may assume a disturbance with a zero mean. Alternatively, we could model one-sided residuals using either a specific distribution (such as

gamma or half-normal) or using *Data Envelopment Analysis* (which uses linear programming to find a hull that encompasses all the points).

In a *stochastic frontier model*, the frontier is also randomly determined; it is not just determined by efficiency. (For example, mismeasurement of quality or human capital may cause some randomness.) That is,

$$y_i = f(x_i)TE_i \exp(v_i)$$

where the frontier is $f(x_i)\exp(v_i)$. We have v_i $Normal$ and $u_i \geq 0$ as before. Thus, $v_i - u_i$ is non-normal and has a non-zero mean. Assuming a symmetric distribution for v_i, then the skewness shows the importance of u_i in determining production.

OLS estimation will be unbiased and consistent for the slope parameters, but the constant will be biased.

For maximum likelihood estimation, we have

$$\begin{aligned} \log L &= -N \ln \sigma - constant + \sum_{i=1}^{N} \ln \Phi(-\frac{\epsilon_i \lambda}{\sigma} + \frac{1}{2}(\frac{\epsilon_i}{\sigma})^2) \\ \epsilon_t &= \alpha + \beta' x_t - \ln y_i \\ \lambda &= \frac{\sigma_u}{\sigma_v} \end{aligned}$$

If $\lambda > 1$, then the inefficiency is a more important factor than the noise. Then,

$$\begin{aligned} E(u_i|\epsilon_i, data) &= (\frac{\sigma\lambda}{1+\lambda^2}) \\ z_i &= -\frac{\epsilon_i \lambda}{\sigma} \end{aligned}$$

We may also think in terms of a cost function:

$$C(y, w) = \min(w'x : f(x) \geq y)$$

Then, we have $C_i = \frac{1}{TE_i} e^{-v_i} c(w_i)$.

To apply this to panel data, we must consider whether u_i is a fixed or random effect and whether it depends on t.

9 Computational Methods

Sometimes, one wants to impose restrictions on the parameters in a likelihood equation. One method is to ignore the restriction and optimize, and then move any parameters back to the boundary if needed. Alternatively, one may reparameterize in a way so that the new parameter is unrestricted. For example, if we must have $-1 < \rho < 1$, then $\theta = \ln(\frac{1+\rho}{1-\rho})$ may take on any value while keeping ρ in the correct interval. We then use the chain rule for derivatives to find the first order conditions.

9.1 Generating Random Numbers

Suppose a random variable X has CDF F. If U is a random variable with a $Uniform(0,1)$ distribution, then $X = F^{-1}(U)$ has the desired distribution. Thus, it is sufficient to be able to generate uniform random numbers.

Computer-generated uniform random numbers are a Markov chain, which means that they can be replicated. Basically, they begin with a large odd number and then use modular arithmetic on the unit interval.

In quasi-Monte Carlo integration, we take advantage of the fact that it is more important to cover the interval of interest than the have true randomness. In this case, the draws are not random (or are less random), so we need fewer of them.

9.2 Method of Krinsky and Robb

Suppose we have $\hat{\beta}$ $Normal(\beta, \Sigma)$ (we might use an estimate, $\hat{\Sigma}$, instead). Suppose we wish to compute the mean and variance of $g(\hat{\beta})$. We may then take a sample of size R from this distribution, compute $g(\hat{\beta}_r)$ for each $r = 1, ..., R$, and compute the mean and variance of these.

To take this sample:

1. Compute the Cholesky decomposition of $\hat{\Sigma}$: $\hat{\Sigma} = QQ'$.
2. Draw standard normal random numbers, v_r.
3. Set $\hat{\beta}_r = \hat{\beta} + Qv$.

This also allows a more detailed look at the distribution of $g(\hat{\beta})$. (By the Slutsky Theorem, it is approximately normal.)

9.3 Metropolis-Hastings Algorithm

- Suppose we wish to sample from $p(\beta_i|b, \Gamma) \propto L(data|\beta_i)g(\beta_i|b, \Gamma)$.
- Let β_{i0} be the previous draw, Γ the diagonal matrix of standard deviations, and σ the tuning constant that controls the acceptance rate. Draw v_r $Normal(0, I_k)$ and set $d_r = \sigma\Gamma v_r$.
- Compute the new trial value, $\tilde{\beta}_{i1} = \beta_{i0} + d_r$. Compute $R = p(\tilde{\beta}_{i1}|b, \Gamma)/p(\beta_{i0}|b, \Gamma)$.
- Draw a U $Uniform(0,1)$. If $U < R$, then set $\beta_{i1} = \tilde{\beta}_{i1}$. Otherwise, set $\beta_{i1} = \beta_{i0}$.

9.4 Gibbs sampling algorithm

- Suppose we wish to sample from $f(x_1, x_2)$, and we know $f(x_1|x_2)$ and $f(x_2|x_1)$.
- Choose any x_{10} in the range of X_1.

- Draw $x_{2,n}$ $f(x_2|x_{1,n})$.
- Draw $x_{1,n+1}$ $f(x_1|x_{2,n})$.

The initial draws should be thrown out, because they depend on the initial choice. These are called the "burn-in" period.

9.5 Optimization Algorithms

Most optimization algorithms are iterative, with $\theta^{(k+1)} = \theta^{(k)} + Update^{(k)}$. In derivative-based methods, the update is a function of the gradient of the function, $g^{(k)}$, which points in the direction of a better solution; often, the Hessian, $H^{(k)}$, is also required. Some particular methods include:

- Steepest Ascent: $Update^{(k)} = -\frac{g^{(k)T}g^{(k)}}{g^{(k)T}H^{(k)}g^{(k)}}g^{(k)}$. This is a slow method.
- Newton-Raphson Method: $Update^{(k)} = -(H^{(k)})^{-1}g^{(k)}$.
- Method of Scoring: $Update^{(k)} = -(E(H^{(k)}))^{-1}g^{(k)}$. (May be computationally easier than Newton-Raphson, but may also be slower.)
- BHHH Method for the MLE: $Update^{(k)} = -(\sum_{i=1}^{n} g_i^{(k)}g_i^{(k)T})^{-1}g^{(k)}$.
- Line Search Methods: Multiply each step by some scalar $\lambda^{(k)}$ (called the *step size*) to get larger improvements (by moving a different length in the same direction). Particular methods include:
  - Squeezing: Set $\lambda^{(k)}$ equal to decreasing powers of 2 until the improvements stop growing.
  - Golden Section: Use $\lambda^{(k-1)}$ and interpolate to find the next step size.
- Quasi-Newton Methods: We may multiply the gradient by other weighting matrices. For example, in the Davidon-Fletcher-Powell method, we multiply by $W^{(k)} = W^{(k-1)} + a^{(k-1)}a^{(k-1)T}$. (WHAT IS a?)

We must also decide when the iterations have converged. Two common stopping criteria are (1) testing how close the derivatives are to zero and (2) testing the absolute change in the parameters. Both of these methods depend on the scales of the variables, which can cause problems (if estimates are very large or very small). A scale-free test can be based on $\Delta = g^{(k)T}(H^{(k)})^{-1}g^{(k)}$.

9.6 Maximum Simulated Likelihood

Often, it is easier to write a likelihood conditional on unknown variables (such as group effects). However, we wish to maximize the unconditional likelihood. To do this, we must integrate over all possible values of the unknown variables, assuming they are normally distributed; that is, we calculate $\log L(\beta) = E(\log L(\beta|v)) = \int_{-\infty}^{\infty} \log L(\beta|v)e^{-v^2}\,dv$. In most cases, this must be done numerically.

9.6.1 Gauss-Hermite Quadrature

We estimate $\int_{-\infty}^{\infty} e^{-v^2} g(v)dv \approx \sum_{h=1}^{H} w_h g(a_h)$, where the w_h are the Hermite weights and the a_h are the Hermite nodes; H determines how good the approximation is. We then maximize this sum.

For panel data econometrics, we wish to evaluate:

$$\log L = \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} g(\sqrt{2}u) \exp(-u^2) du$$

Then, we may use Hermite quadrature, to estimate:

$$\log L \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} \log \sum_{h=1}^{H} w_h g(\sqrt{2}z_h)$$

. This reduces the integral to a sum, which is possible to maximize numerically.

9.6.2 Integration by Simulation (Monte Carlo Integration)

We may estimate $E(g(v))$ by sampling from the distribution of v, so that $E(g(v)) \approx \frac{1}{R} \sum_{i=1}^{R} g(v_i)$. The same $v_1, ..., v_R$ can be used to maximize over all the parameters and must also be used to compute the derivatives, so that the standard errors are correct. Note that non-random but properly chosen $v_1, ..., v_R$ may make the estimates converge faster.

More generally, suppose we wish to evaluate:

$$\log L = \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} f(v)g(v)dv$$

where v has density $g(v)$. Then, we may draw R independent random numbers from the distribution of v, compute $f(v)$ for each draw, and then take the average. This gives:

$$\log L \approx \sum_{i=1}^{N} \log \frac{1}{R} \sum_{j=1}^{R} f(v_r)$$

We may then maximize the sum with respect to the parameters, holding the draws fixed. (Note that we need a few hundred draws for each individual, but fewer if the draws are not random but chosen properly.) The same draws must be used to calculate standard errors and anything else as well.

9.7 EM Algorithm

The EM algorithm is applied in cases where there is missing data (which includes unknown intermediate parameters, like class membership). It has two steps which alternate:

- Expectation: Compute an expected log likelihood (across the missing data) given the data and the previous estimates for the parameters.
- Maximization: Maximize the expected log likelihood by adjusting the parameter values.