

Economics 722: Econometrics
Harry Kelejian

Matthew Chesnes

Updated: May 11, 2006

1 Lecture 1: January 26, 2006

1.1 Topics in Linear Systems: Truncated 2SLS and 3SLS

- Sometimes our X matrix, $T \times K$, is such that $T < K$, so $X'X$ is singular and 2SLS cannot be implemented.
- A researcher may not even know the entire list of predetermined variables in a system, or data on these variables may not be available.
- In such cases, we choose a subset of our entire list of predetermined variables, and run Truncated 2SLS, (T2SLS).
- Consider the i^{th} equation of a linear system:

$$y_i = Y_i \gamma_i + X_i \delta_i + \epsilon_i, \quad \epsilon_i \sim (0, \sigma_{ii} I), \quad i = 1 \dots G,$$

where Y_i is a $T \times G_i$ matrix of endogenous variables in the i^{th} equation and X_i is a $T \times K_i$ matrix of predetermined variables. Rewrite the equation as:

$$y_i = Z_i B_i + \epsilon_i, \quad B_i' = (\gamma_i', \delta_i'), \quad Z_i = (Y_i, X_i).$$

- The entire set of predetermined variables is $X = [X_i, X_i^*]$, where X_i^* are those predetermined variables not in equation i . Take a subset of X_i^* , called Q , which is $T \times R$. Let $W = [X_i, Q]$. Regressing the endogenous Y 's on W yields:

$$\hat{\Pi}_w = (W'W)^{-1}W'Y_i,$$

and fitted values:

$$\begin{aligned} \hat{Y}_i &= W \hat{\Pi}_w \\ &= W(W'W)^{-1}W'Y_i \\ &= R_w Y_i \end{aligned}$$

where R_w is the projection matrix. $\hat{\Pi}_w$ can be partitioned into $(\hat{\Pi}_{X_i}, \hat{\Pi}_Q)'$, where $\hat{\Pi}_Q$ is $R \times G_i$.

- We assume $\hat{\Pi}_w$ is consistent and the rank of Π_Q is full, so we have at least as many excluded predetermined variables as we have included endogenous variables. Rank condition.
- Then the T2SLS estimator is as follows:

$$\hat{B}_i = (\hat{Z}_i' \hat{Z}_i)^{-1} \hat{Z}_i' y_i,$$

where $\hat{Z}_i = (\hat{Y}_i, X_i)$.

- Given the conditions of the Schonfeld CLT, we get the following results:
 - (A.1) $\hat{B}_i = (\hat{Z}'_i Z_i)^{-1} \hat{Z}'_i y_i$.
 - (A.2) $\sqrt{T}(\hat{B}_i - B_i) \rightarrow^d N(0, \sigma_{ii} \text{plim } T(\hat{Z}'_i \hat{Z}_i)^{-1})$.
 - (A.3) $\text{plim } T(\hat{Z}'_i \hat{Z}_i)^{-1} - \text{plim } T(\tilde{Z}'_i \tilde{Z}_i)^{-1}$ is positive definite, where \tilde{Z}_i is formed using a submatrix of those columns of X not already accounted for by W .

So what does all this mean? (A.1) says that \hat{B}_i can be expressed in instrumental variables form. (A.2) says the estimator is asymptotically normal. And (A.3) says that if you use \tilde{Z} based on everything that \hat{Z} has AND MORE, then you get efficiency gains.

- See HK notes for proofs.

2 Lecture 2: January 31, 2006

2.1 More on T2SLS and T3SLS

- The formulas for truncated two and three stage least squares are the same (strangely) though you lose efficiency by only doing 2SLS.
- The conflict is that asymptotically, as you add instruments, the large sample variance/covariance matrix gets smaller. But in small samples, as you add instruments, $\hat{Y}_i \rightarrow Y_i$ and 2SLS \rightarrow OLS, which is inconsistent.

T3SLS

- Consider the i^{th} equation:

$$y_i = Z_i B_i + \epsilon_i, \quad i = 1, \dots, L < G.$$

So, even though there may be G equations in the system, we only use L of them. Everything is ok with estimating the partial system though there are efficiency gains from estimating all G equations (if possible).

- Note $Z_i = [Y_i, X_i]$, where X_i are the predetermined variables: exogenous and lagged endogenous.
- If $L = 1$, we have 2SLS. If $L > 1$, we have 3SLS.
- Stack the L equation system to form:

$$y = ZB + \epsilon,$$

where $y = (y_1, \dots, y_L)'$, and Z is block diagonal as usual.

- Let $E[\epsilon\epsilon'] = \Sigma_\epsilon \otimes I = \Omega_\epsilon$, where Σ_ϵ is $L \times L$.
- The T3SLS estimator is obtained as follows.
 - (1) Using a set of instruments, $H = [X_1, \dots, X_L, \Psi]$, where Ψ are other predetermined variables possibly not in any of the L equations, form the fitted values:

$$\hat{Y}_i = R_H Y_i,$$

and,

$$\hat{Z}_i = (\hat{Y}_i, X_i).$$

- (2) Estimate each equation by 2SLS to get \hat{B}_i and form:

$$\hat{\epsilon}_i = y_i - Z_i \hat{B}_i,$$

$$\hat{\sigma}_{ij} = T^{-1} \hat{\epsilon}_i' \hat{\epsilon}_j,$$

$$\hat{\Sigma}_\epsilon = \{\hat{\sigma}_{ij}\}_{L \times L},$$

$$\hat{\Omega}_\epsilon = \hat{\Sigma}_\epsilon \otimes I.$$

– (3) Finally form:

$$\hat{B}_{T3SLS} = (\hat{Z}'\hat{\Omega}_\epsilon^{-1}\hat{Z})^{-1}\hat{Z}'\hat{\Omega}_\epsilon^{-1}y.$$

- We have similar results as in the two stage case where \hat{B}_{T3SLS} can be expressed as an IV estimator, and:

$$\sqrt{T}(\hat{B}_{T3SLS} - B) \rightarrow^d N(0, plim T(\hat{Z}'\hat{\Omega}_\epsilon^{-1}\hat{Z})^{-1}).$$

2.2 Weak Instruments

- Sometimes we may have a very large sample, but our R^2 's remain small and our χ^2 values do not tend to infinity, but rather hover around their critical values. Why is this case with such a large sample?
- If our IVs are correlated with the endogenous variables and not with the errors, we're golden. All is consistent.
- But if our IVs are correlated with the errors, we get inconsistency and bias which is even more severe when the IVs are weakly correlated with the endogenous variables.
- Consider the model:

$$y_i = X_i b + u_i, \quad i = 1 \dots N, \quad (X_i, u_i) \sim iid(0, \Sigma),$$

but $\sigma_{12} \neq 0$. So the covariance between X and u is not zero so OLS would be inconsistent.

- Suppose we have an IV, Z_i which is iid $(0, \sigma_z^2)$. Form our estimator:

$$\hat{b}_{IV} = (Z'X)^{-1}Z'y = b + \frac{\sum Z_i u_i}{\sum Z_i X_i} = b + \left(\frac{N^{-1} \sum Z_i u_i}{N^{-1} \sum Z_i X_i} \right) \frac{\sigma_z \sigma_u \sigma_x}{\sigma_z \sigma_u \sigma_x}.$$

Then,

$$\hat{b}_{IV} \rightarrow b + \frac{\rho_{zu} \sigma_u}{\rho_{zx} \sigma_x}.$$

So if the correlation between Z and u is large or the correlation between Z and X is small, we have bias problems. This is exactly the problem of Weak Instruments.

- Now consider the two equation system:

$$\underbrace{Y_1}_{N \times 1} = \alpha Y_2 + \epsilon,$$

$$Y_2 = C_n \underbrace{X}_{N \times K} + v,$$

where $C_N = N^{-1/2}C \rightarrow 0$. This last condition means that the X 's are weak instruments. The coefficient on X in the second equation goes to zero as we increase N .

- Assume $(\epsilon_i, v_i) \sim iid N(0, \Omega)$, with $\sigma_{v\epsilon} \neq 0$.
- Assume X is exogenous and $N^{-1}X'X \rightarrow Q_x$ and Q_x^{-1} exists as usual.
- Then the population version of the χ^2 test for the significance of X is:

$$\chi^2 = \frac{C'_N X' X C_N}{\sigma_v^2} \rightarrow \frac{C' Q_x C}{\sigma_v^2} < \infty.$$

So if the X 's are weak instruments, the χ^2 statistic is limited to be less than infinity as N gets large.

- Note that if $C_N = C \neq 0$, the $\chi^2 \rightarrow \infty$ as expected. But this is only if the X 's are NOT weak instruments.
- Consider the 2SLS estimator of α :

$$\hat{\alpha}_{2SLS} = (\hat{Y}'_2 \hat{Y}_2)^{-1} \hat{Y}'_2 Y_1.$$

We can rewrite this as:

$$\hat{\alpha}_{2SLS} = \alpha + (\hat{Y}'_2 \hat{Y}_2)^{-1} \hat{Y}'_2 \epsilon,$$

where $\hat{Y}_2 = R_x Y_2$, with projection matrix: $R_x = X(X'X)^{-1}X'$, so:

$$\hat{Y}_2 = X(X'X)^{-1}X'Y_2 = X\hat{C}_N.$$

- We can also rewrite \hat{C}_N as:

$$\hat{C}_N = (X'X)^{-1}X'Y_2 = C_N + (X'X)^{-1}X'v.$$

Since $C_N = N^{-1/2}C$, $C_N = O(N^{-1/2})$, that is C_N is of order $N^{-1/2}$. So,

$$(X'X)^{-1}X'v = N^{-1/2} \underbrace{N(X'X)^{-1}}_{\rightarrow Q_x^{-1}} \underbrace{N^{-1/2}X'v}_{\rightarrow^d N(0, \sigma_v^2 Q_x)},$$

which means $(X'X)^{-1}X'v \sim O_p(N^{-1/2})$, [“Order in Probability”]. Thus the overall estimator,

$$\hat{C}_N \sim O_p(N^{-1/2}).$$

- Returning to the terms of $\hat{\alpha}$, we have:

$$\hat{Y}'_2 \hat{Y}_2 = \hat{C}'_N X' X \hat{C}_N \sim O_p(1) \neq \infty.$$

So the first term is at most in probability of order 1. Also,

$$\hat{Y}'_2 \epsilon = \hat{C}'_N X' \epsilon \sim O_p(1) \neq 0.$$

Thus,

$$plim \hat{\alpha} \neq \alpha.$$

- Punchline. 2SLS is inconsistent with weak instruments!
- We can also write the bias as a function of K , the number of instruments we use. It can be shown that as we increase K , the bias associated with α goes to zero. But we can't include too many instruments or 2SLS will become OLS as before. So usually, as our sample size increases, we pick more and more instruments. Write this as:

$$K_N = F(N).$$

Then,

$$K_N(\hat{\alpha} - \alpha) \rightarrow^d N\left(\frac{\sigma_{v\epsilon}}{s}, \frac{\sigma_\epsilon^2}{s}\right),$$

where,

$$s = \lim_{N \rightarrow \infty} \frac{1}{K_N^2} C_N' \frac{X'X}{N} C_N \neq 0.$$

So (somehow) this says we have large sample consistency, but in small samples, we'll get a bias:

$$\hat{\alpha} \approx N\left(\alpha + \frac{\sigma_{v\epsilon}}{sK_N}, \frac{\sigma_\epsilon^2}{sK_N^2}\right),$$

which means our bias term is:

$$BIAS = \frac{\sigma_{v\epsilon}}{sK_N}.$$

This goes to zero as $K_N \rightarrow \infty$, but in the nonlimiting case, we have a bias using these weak instruments.

3 Lecture 3: February 2, 2006

3.1 Final Note on IV Selection

- Consider the model:

$$y_{t1} = c_0 + c_1 y_{t2} + c_2 x_{t1} + \epsilon_{t1},$$

$$y_{t2} = d_0 + d_1 y_{t1} + d_2 x_{t2} + \epsilon_{t2}.$$

Under usual assumptions, both equations are identified since there is one pre-determined variable outside of each equation to identify the single endogenous variable.

- But suppose we didn't have data on x_{t2} . Can we still identify the first equation? If we assume that the x 's are serially correlated, which for time series is reasonable, then $x_{t-1,2}$ and x_{t2} are correlated and $x_{t-1,2}$ is correlated with $y_{t-1,2}$ if you solve for the reduced form of the system. Thus, using $y_{t-1,2}$ as an instrument is one option.
- What if the errors are $MA(1)$? We can't use $y_{t-1,2}$ because it would be correlated with the errors, but we could use $y_{t-2,1}$.
- What if the errors are an AR process? Then we can't use lagged y 's because they are ALL correlated with the errors, but $x_{t-1,1}$ might be an option.
- In general, if you suspect the errors are serially correlated, don't use lagged y 's as instruments.

3.2 Bayesian Econometrics

- In Bayesian analysis, a subjective view of probability is taken. Suppose we have a Keynesian consumption function and we want to estimate the marginal propensity to consume. In general this is a parameter that is bounded between zero and one. A classical statistician would stop there and say our parameter is equally likely to come from anywhere in that interval. A Bayesian would say the probability that it falls in the range, say, from 0.75 to 0.80 is much higher than the probability it falls in the interval 0.00 to 0.05, so we SHOULD use this information. If we have a prior, use it.
- Recall from 623,

$$f(y|x) = \frac{f(x,y)}{f(x)}.$$

Thus, we can write:

$$f(x,y) = f_2(y|x)f_1(x) = f_4(x|y)f_3(y),$$

so,

$$f_4(x|y) = \frac{f_2(y|x)f_1(x)}{f_3(y)}.$$

Let $x = \theta$, a parameter and assume y is some known data:

$$\underbrace{f(\theta|y)}_{\text{Posterior}} = \frac{\overbrace{f(y|\theta)}^{\text{Likelihood}} \overbrace{f(\theta)}^{\text{Prior}}}{\underbrace{f(y)}_{\text{Const of Int}}}$$

Thus, we see that the posterior probability is proportional to the likelihood of the data (given our parameter) multiplied by the prior probability. Or:

$$Post(\theta|y) \propto \mathcal{L}(y|\theta) * Prior(\theta).$$

- **Remark** An important formula for completing the square with matrices:

$$X'AX - 2B'X = (X - A^{-1}B)'A(X - A^{-1}B) - B'A^{-1}B, \text{ if } A' = A, \text{ and } A^{-1} \text{ exists.}$$

- **Example of Learning.** Suppose we have a parameter θ and data on y_1 and y_2 . From our formula,

$$Post(\theta|y_1, y_2) \propto \mathcal{L}(y_1, y_2|\theta) * Prior(\theta).$$

Which we can write:

$$Post(\theta|y_1, y_2) \propto \underbrace{Prior(\theta) * \mathcal{L}_1(y_1|\theta)}_{Post(\theta|y_1)} * \mathcal{L}_2(y_2|y_1, \theta).$$

So the first term on the RHS is the posterior probability after the “learning” we do when realizing y_1 . Thus,

$$Post(\theta|y_1, y_2) \propto \underbrace{Post(\theta|y_1)}_{\text{New Prior}} * \mathcal{L}_2(y_2|y_1, \theta).$$

Again the first term on the RHS now becomes our NEW prior when considering y_2 .

- So the prior is something we have to bring to the table. How do people choose the prior in practice?
 - (1) Ideally, the prior should describe what you know before looking at the data. It should be your belief about the parameter of interest.
 - (2) Sometimes it is chosen to be compatible with the likelihood function. If the likelihood is normal, we often choose the prior to be normal to make the analysis easier. When we do this, we call it a Natural Conjugate Prior.
 - (3) We could also assume an “Uninformed Prior” or the “Assumption of Ignorance.” Assume $prior(\theta) \sim$ uniform over some interval. The problem with this is that if you know nothing about θ , then you should also be ignorant about functions of θ , say $\gamma = e^\theta$. But using a standard transformation technique, you won’t get a uniform for γ . We’ll ignore this issue.

Application: Matrix Notation

- Suppose our model is:

$$Y = XB + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

and our prior is:

$$Prior(B) \sim N(B_0, \sigma_0^2 I),$$

where B_0 and σ_0^2 are both known. For now we will also assume σ^2 is known.

- Thus, our posterior is:

$$Post(B|Y) \propto \exp\left(-\frac{1}{2\sigma^2}(Y - XB)'(Y - XB)\right) \exp\left(-\frac{1}{2\sigma_0^2}(B - B_0)'(B - B_0)\right),$$

where the first term is the likelihood of Y given B and the second is the prior probability of B . Note everything else is constant so it gets thrown into the constant of proportionality.

- Consider the likelihood function:

$$\begin{aligned} \mathcal{L}(Y|B) &\propto \exp\left(-\frac{1}{2\sigma^2}(Y - XB)'(Y - XB)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(Y - X\hat{B} + X\hat{B} - XB)'(Y - X\hat{B} + X\hat{B} - XB)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\hat{\epsilon} + X(\hat{B} - B))'(\hat{\epsilon} + X(\hat{B} - B))\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\underbrace{\hat{\epsilon}'\hat{\epsilon}}_{const} + (\hat{B} - B)'X'X(\hat{B} - B))\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\hat{B} - B)'X'X(\hat{B} - B)\right) \end{aligned}$$

From notes: “The posterior can now be determined by completing the quadratic [using the formula in the remark above]; if this is done, that posterior will turn out to be a normal distribution.”

Application: Joint Posterior

- Now we assume that both B and σ^2 need to be estimated. Our model is thus:

$$Y = XB + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

and our prior is:

$$Prior(B, \sigma) \propto \frac{1}{\sigma},$$

where he tried to explain this prior, but it was unclear. For now we will also assume σ^2 is known.

- Somehow (??):

$$Post(B, \sigma | data) \propto \frac{1}{\sigma^{T+1}} \exp\left(-\frac{1}{2\sigma^2}(Y - XB)'(Y - XB)\right).$$

And applying the same technique as in the last application,

$$Post(B, \sigma | data) \propto \frac{1}{\sigma^{T+1}} \exp\left(-\frac{1}{2\sigma^2}(\hat{\epsilon}'\hat{\epsilon} + (\hat{B} - B)'X'X(\hat{B} - B))\right).$$

This is expressible as the product of a marginal, $g_1(\sigma)$ and a conditional, $g_2(B, \sigma)$. Which implies:

$$f(B, \sigma) \sim N(\hat{B}, \sigma^2(X'X)^{-1}),$$

$$g(\sigma) \propto \frac{1}{\sigma^{T+1-K}} \exp\left(-\frac{1}{2\sigma^2}\hat{\epsilon}'\hat{\epsilon}\right).$$

- He then goes on to show that the marginal of B given the data is a multivariate t which is shown on page 5 and 6 of the notes. I'm confident that no one followed any of this.

Application: Scalar Notation

- Suppose our model is:

$$y_t = x_t b + \epsilon_t, \quad \epsilon_t \sim iidN(0, \sigma^2), \quad t = 1 \dots T,$$

with x_t non-stochastic and our prior is:

$$Prior(b) \sim N(b_0, \sigma_0^2),$$

with b_0 , σ_0^2 , and σ^2 all known.

- Posterior becomes:

$$Post(b) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_t (y_t - x_t b)^2\right) \exp\left(-\frac{1}{2\sigma_0^2} (b - b_0)^2\right).$$

Which after some algebra:

$$Post(b) \propto \exp\left(-\frac{1}{2\sigma_3^2} (b - \bar{b})^2\right).$$

See notes for definitions of σ_3^2 and \bar{b} .

- Limiting cases.
 - (1) If $T \rightarrow \infty$, our estimator becomes OLS (because the sample size dominates) and the variance is the standard, $\sigma^2(X'X)^{-1}$.
 - (2) If $\sigma_0^2 \rightarrow \infty$, then our prior is poor, so we are just doing normal OLS.

– (3) If $\sigma_0^2 \rightarrow 0$, our prior is golden, so we should just use b_0 for our estimate.

4 Lecture 4: February 7, 2006

4.1 Pretest Estimators

- Two of the biggest empirical problems in economics are aggregation and pretest issues. The former is the idea that we never have fine enough data to reflect the individual choices that are being made, but instead just get some overall picture. The latter is the problem that when we formulate a model, estimate the model, and then based on the results, reformulate the model, we get biases. It happens all the time (every time really), but it does create biases.
- First a preliminary result. Suppose Z is a discrete random variable with density $f_1(Z)$. Z takes on values C_1, \dots, C_N with probabilities P_1, \dots, P_N . Let X be another RV with density $f_2(X)$, which is possibly continuous. Finally let $h(X)$ be a function such that:

$$E[h(X)] = \int_{-\infty}^{\infty} h(X)f_2(X)dX \text{ exists.}$$

THEN:

$$E[Zh(X)] = \sum_{i=1}^N C_i P_i E[h(X)|Z = C_i].$$

So we sum over the realization of Z , multiplied by the conditional expectation of $h(X)$ and weight by the probability of each Z .

- Recall the laws for iterated expectations:

$$E[Y] = E_x[E[Y|X]],$$

$$E[Y|X] = E[E(Y|X, W)|X],$$

and so on.

- So what is the pretest issue. Suppose our true model is:

$$[M1] : Y = XB + \epsilon,$$

with all the classical assumptions satisfied. Suppose we estimate (unknowingly):

$$[M2] : Y = XB + Z\gamma + \epsilon.$$

If we estimate M2, we might test $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$. If we accept the null, then re-estimate B via M1, if we reject the null, estimate B via M2.

- But what have we done? Suppose we test H_0 via the F ratio. Let $v = 1$ if we accept H_0 , ie if $F < F_{0.95}$. Let $v = 0$ if we reject. Then our estimator of B is:

$$\tilde{B} = v\hat{B}_1 + (1 - v)\hat{B}_2 = \hat{B}_2 + v(\hat{B}_1 - \hat{B}_2),$$

where the subscripts refer to the estimators from each model. Note, under classical assumptions:

$$E[\hat{B}_1] = E[\hat{B}_2] = B.$$

Adding an additional (unneeded) variable does not create a bias. However(!):

$$\begin{aligned} E[\tilde{B}] &= E[\hat{B}_2] + E[v(\hat{B}_1 - \hat{B}_2)] \\ &= B + 1 * E[(\hat{B}_1 - \hat{B}_2)|v = 1] * Prob\{v = 1\} + 0 * E[(\hat{B}_1 - \hat{B}_2)|v = 0] * Prob\{v = 0\} \\ &= B + E[(\hat{B}_1 - \hat{B}_2)|v = 1] * Prob\{v = 1\} \\ &= B + E[(\hat{B}_1 - \hat{B}_2)|F < F_{0.95}] * Prob\{F < F_{0.95}\} \\ &= B + E[(\hat{B}_1 - \hat{B}_2)|F < F_{0.95}] * 0.95 \neq B \end{aligned}$$

So we've introduced a bias! By reformulating after seeing the results, we have imposed some kind of circular reasoning which is unfortunate.

- There are two other examples in the notes of how pretest estimators will create biases. If we know that one coefficient should be negative for example by economic theory and it keeps coming out positive in our regressions, but we reformulate until the coefficient is negative, we've created biases.
- Another useful result:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = Pr\{X < 0\} \frac{\int_{-\infty}^0 xf(x)dx}{Pr\{X < 0\}} + Pr\{X > 0\} \frac{\int_0^{\infty} xf(x)dx}{Pr\{X > 0\}},$$

or,

$$E[X] = Pr\{X < 0\}E[X|X < 0] + Pr\{X > 0\}E[X|X > 0].$$

- So in general, consider the battle between purists and pre-testers. The purist goes away for 2 years and studies everything there is to know about a model and comes up with the "true" model of what he's trying to estimate. He runs one regression and that's it. The pre-tester starts on day one with a set of facts about this coefficient and that one. He immediately starts running regressions and constantly reformulates until the facts closely resemble what is shown in the regression. The pre-tester creates biases for sure, but the purist dies in the library before even running a regression. Who do you side with?
- One problem with the pre-tester is that he ALWAYS gets what he wants, because if he doesn't, he just reformulates. This is why economics doesn't move forward very quickly, compared to other sciences.
- There may, however, be gains from using pre-test estimators, at least in terms of the Mean Square Error (MSE) of your estimator. Consider a RV, $Z \sim \text{bernoulli}$ with $f(Z = 1) = P$ and $f(Z = 0) = 1 - P$. Then $E[Z] = P$. Suppose our model is:

$$y = X\beta + Z\gamma + \epsilon,$$

where the true value of β is β_0 . Suppose we test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ and let $v = 1$ if H_0 is accepted via an F test and $v = 0$ otherwise. Then our pre-test estimator is:

$$\tilde{\beta} = v\beta_0 + (1 - v)\hat{\beta},$$

where $\hat{\beta}$ is our OLS estimator. Then, rearranging:

$$\tilde{\beta} - \beta_0 = v\beta_0 - \beta_0 + \hat{\beta} - v\hat{\beta},$$

$$\tilde{\beta} - \beta_0 = (1 - v)(\hat{\beta} - \beta_0).$$

So,

$$\begin{aligned} MSE(\tilde{\beta}) &= (\tilde{\beta} - \beta_0)'(\tilde{\beta} - \beta_0) \\ &= \underbrace{(1 - v)^2}_{\in\{0,1\}}(\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0) \\ &\leq (\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0) \\ &= MSE(\hat{\beta}) \end{aligned}$$

So the MSE of our pre-test estimator is smaller than the MSE of the OLS estimator!

4.2 Section 3 - Nonparametrics: The Method of Kernels

- Suppose we seek an estimate of the density of a random variable. We have a random sample X_1, \dots, X_n from a continuous PDF, $f(x)$ with CDF, $F(x)$ and we seek to estimate $\hat{f}_n(x)$. We would like the estimator to be unbiased but in general this is NOT possible. Thus, we'll seek a consistent estimator such that:

$$plim_{n \rightarrow \infty} \hat{f}_n(x) = f(x) \quad \forall x.$$

- One naive estimator is the empirical distribution function:

$$F_n(x) = \frac{\text{number of observations} \leq x}{n}.$$

This is like the histogram approach where we choose a window size $h_n > 0$ and just count the observations that fall in each box. This induces PDF:

$$f_n(x) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n}.$$

As the sample size grew, we might want to make the window size smaller and smaller. However, with this approach, suppose we want to estimate the density at x , we place equal weight on all observations in the interval $(x - h_n, x + h_n)$. This doesn't seem like a good idea. We can do better with kernels.

- First, let's try to represent the naive estimator above with a kernel. Let:

$$K(x) = \frac{1}{2}, \text{ if } x \in [-1, 1),$$

and $K(x) = 0$, else. So this is called the rectangular kernel. Thus our PDF can be written:

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right).$$

This density function is often the first approach we use (our straw-man which we'll knock down shortly).

- To see the above density, consider:

$$\begin{aligned} f_n(x) &= \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \\ &= \frac{1}{nh_n} \sum_{i=1}^n \frac{1}{2} I(x - h_n < X_i < x + h_n) \\ &= \frac{1}{2nh_n} \sum_{i=1}^n [F_n(x + h_n) - F_n(x - h_n)] \\ &= \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n} \end{aligned}$$

which is what we had above.

- In general, continuous kernel density functions, $K(x)$, have to satisfy a certain set of conditions:
 - (c1) $K(x) > 0$.
 - (c2) $\int K(x)dx = 1$.
 - (c3) $\sup|K(x)| = c_K < \infty$.
 - (c4) $\int K^2(x)dx < \infty$.

These conditions imply that $K(x)$ is a valid kernel density function. We will assume that all kernels we deal with in what follows satisfy these conditions.

5 Lecture 5: February 9, 2006

5.1 More on the Method of Kernels

- Five kernels which satisfy (c1-c4):
 - (1) Naive: $K_1(x) = \frac{1}{2}$ if $|x| < 1$, $K_1 = 0$ else.
 - (2) Bartlett: $K_2(x) = 1 - |x|$ if $|x| < 1$, $K_2 = 0$ else.
 - (3) Gaussian: $K_3(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ if $-\infty < x < \infty$.
 - (4) Quartic: $K_4(x) = (15/16)(1 - x^2)^2$ if $|x| < 1$, $K_4 = 0$ else.
 - (5) Epanechnikov: $K_5(x) = (3/4)(1 - x^2)$ if $|x| < 1$, $K_5 = 0$ else.
- To proceed, we need some further assumptions. Let X be an $r \times 1$ random vector with density $f(x)$ and the window width, h_n , is such that:

$$h_n \rightarrow 0, \text{ and } nh_n^r \rightarrow \infty \text{ as } n \rightarrow \infty.$$

So the speed that the window width goes to zero is inversely related to the length of the vector, X . Our estimated density:

$$\hat{f}_n(x) = \frac{1}{nh_n^r} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right),$$

is such that:

$$\sup_x |f(x)| = c_f < \infty, \text{ and } \hat{f}_n(x) \rightarrow^p f(x),$$

so the density is bounded and the estimated density is consistent.

- **Proof of the Consistency of our Kernel Estimator** We will follow (T)Chebychev, and show that the expectation is the actual density and the variance of our estimator goes to zero. First the expectation:

$$\begin{aligned} E[\hat{f}_n(x)] &= \int \frac{1}{nh_n^r} \sum_{j=1}^n K\left(\frac{x - x_j}{h_n}\right) f(x_j) dx_j \\ &= \int \frac{1}{nh_n^r} \sum_{j=1}^n K\left(\frac{x - z}{h_n}\right) f(z) dz \\ &= \int \frac{1}{h_n^r} K\left(\frac{x - z}{h_n}\right) f(z) dz \\ &\quad \text{because the random sample is iid} \\ &= \int \frac{1}{h_n^r} K(w) f(x - h_n w) h_n^r dw \\ &= \int K(w) f(x - h_n w) dw \end{aligned}$$

The last two lines follow from a change of variables. We let $w = (x - z)/h_n$ so that $z = x - h_n w$. The jacobian is thus:

$$\left| \frac{\partial z}{\partial w} \right| = | -h_n I_r | = h_n^r,$$

noting that x, w , and z are all $r \times 1$ vectors. Substitute in $dz = h_n^r dw$ and you're there. So the next step is to take the limit. We can take the limit inside the integral by the Dominated Convergence theorem of Billingsly. As long as the integrand is bounded by a function which is integrable, we can do this. Thus,

$$\begin{aligned} \lim_{n \rightarrow \infty} E[\hat{f}_n(x)] &= \lim_{n \rightarrow \infty} \int K(w) f(x - h_n w) dw \\ &= \int K(w) \lim_{n \rightarrow \infty} f(x - h_n w) dw \\ &= \int K(w) f(x) dw \\ &= f(x) \underbrace{\int K(w) dw}_{=1} = f(x) \end{aligned}$$

Again, noting that $h_n \rightarrow 0$ as $n \rightarrow \infty$. So our estimated density is asymptotically

unbiased. Now we'll show the variance goes to zero. Consider the variance:

$$\begin{aligned}
\text{Var}[\hat{f}_n(x)] &= n^{-2}h_n^{-2r} \sum_{j=1}^n \text{var}\left[K\left(\frac{x-x_j}{h_n}\right)\right] \\
&= n^{-1}h_n^{-2r} \text{var}\left[K\left(\frac{x-z}{h_n}\right)\right] \\
&\quad \text{because the random sample is iid} \\
&= n^{-1}h_n^{-2r} \left\{ E\left[\left(K\left(\frac{x-z}{h_n}\right)\right)^2\right] - \left(E\left[K\left(\frac{x-z}{h_n}\right)\right]\right)^2 \right\} \\
&\quad \text{same substitution: } w = (x-z)/h_n, dz = h_n^r dw \\
&= n^{-1}h_n^{-2r} \left\{ E[(K(w))^2] - (E[K(w)])^2 \right\} \\
nh_n^r \text{Var}[\hat{f}_n(x)] &= h_n^{-r} E[(K(w))^2] - h_n^{-r} (E[K(w)])^2 \\
&= h_n^{-r} \int K^2(w)f(x-h_nw)h_n^r dw - h_n^{-r} \left[\int K(w)f(x-h_nw)h_n^r dw \right]^2 \\
&= \int K^2(w)f(x-h_nw)dw - h_n^r \left[\int K(w)f(x-h_nw)dw \right]^2 \\
&\quad \text{Using Dominated Convergence} \\
nh_n^r \text{Var}[\hat{f}_n(x)] &\rightarrow \int K^2(w)f(x)dw - \underbrace{\lim(h_n^r)}_0 \underbrace{\left[\int K(w)f(x)dw \right]^2}_{<\infty} \\
nh_n^r \text{Var}[\hat{f}_n(x)] &\rightarrow f(x) \int K^2(w)dw < \infty
\end{aligned}$$

So since $\lim_{n \rightarrow \infty} nh_n^r \text{Var}[\hat{f}_n(x)] < \infty$ and $nh_n^r \rightarrow \infty$, it must be that:

$$\text{Var}[\hat{f}_n(x)] \rightarrow 0.$$

QED. The estimator is asymptotically unbiased and its variance goes to zero. By Chebychev, the estimator is consistent.

A Kernel Estimator of a Regression Function

- So in this section we will apply the above method to a regression model. Suppose we want to estimate a regression function:

$$r(x) = E[Y|X] = \int yf(y|x)dy,$$

so we're essentially solving for the function $\hat{y} = X\hat{\beta}$. Recall:

$$f(y|x) = \frac{f(x, y)}{f(x)},$$

so in order to find the conditional, we'll need to estimate the joint, estimate the marginal, divide, and finally integrate to get the conditional expectation.

- Assume X is $k \times 1$ and Y is scalar. Assume we have a joint random sample. Let $K_*(y, x)$ be a kernel joint density function such that:

$$\int y K_*(y, x) dy = 0,$$

$$\int K_*(y, x) dy = K(x),$$

so hitting our kernel with y and integrating gives us zero (just a scale assumption) and integrating out the y gives us a marginal for x . We maintain the assumption above that:

$$h_n \rightarrow 0, nh_n^k \rightarrow \infty \text{ as } n \rightarrow \infty.$$

Denote $E[Y^2|X = x] = \sigma_y^2(x)$. Note this isn't a conditional variance unless the mean of Y is zero. We also assume everything is well-behaved and bounded.

- Our joint estimator is:

$$\hat{f}_n(y, x) = n^{-1} h_n^{-(k+1)} \sum_{j=1}^n K_*\left(\frac{y - Y_j}{h_n}, \frac{x - X_j}{h_n}\right).$$

We raise h_n to the $k + 1$ because X is $k \times 1$ and Y is 1×1 . This is just to keep things scaled correctly. Also recall the marginal for X is:

$$\hat{f}_n(x) = n^{-1} h_n^{-k} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right).$$

- One nice result that we can prove is that if we integrate out the Y of our estimated joint density, we'll get our estimated marginal for X . That is:

$$\hat{f}_n(x) = \int \hat{f}_n(y, x) dy.$$

Proof:

$$\begin{aligned}
\int \hat{f}_n(y, x) dy &= \int n^{-1} h_n^{-(k+1)} \sum_{j=1}^n K_* \left(\frac{y - Y_j}{h_n}, \frac{x - X_j}{h_n} \right) dy \\
&\text{same substitution: } z_j = (y - Y_j)/h_n, dy = h_n dz_j \\
&= \int n^{-1} h_n^{-(k+1)} \sum_{j=1}^n K_* \left(z_j, \frac{x - X_j}{h_n} \right) h_n dz_j \\
&= n^{-1} h_n^{-k} \sum_{j=1}^n \int K_* \left(z_j, \frac{x - X_j}{h_n} \right) dz_j \\
&= n^{-1} h_n^{-k} \sum_{j=1}^n K \left(\frac{x - X_j}{h_n} \right) = \hat{f}_n(x)
\end{aligned}$$

QED.

- Next, lets see what the estimator of the regression function, $\hat{r}_n(x)$, looks like. Note:

$$\hat{r}_n(x) = \hat{E}(Y|X = x) = \int y \frac{\hat{f}_n(y, x)}{\hat{f}_n(x)} dy.$$

So lets do a crazy amount of algebra:

$$\begin{aligned}
\hat{r}_n(x) \hat{f}_n(x) &= \int y \hat{f}_n(y, x) dy \\
&= \int \sum_{j=1}^n \frac{y}{n h_n^{k+1}} K_* \left(\frac{y - Y_j}{h_n}, \frac{x - X_j}{h_n} \right) dy \\
&\text{same substitution: } z_j = (y - Y_j)/h_n, dy = h_n dz_j \\
&= \int \sum_{j=1}^n \frac{h_n z_j + Y_j}{n h_n^{k+1}} K_* \left(z_j, \frac{x - X_j}{h_n} \right) h_n dz_j \\
&= \frac{h_n}{n h_n^k} \sum_{j=1}^n \underbrace{\int z_j K_* \left(z_j, \frac{x - X_j}{h_n} \right) dz_j}_{=0 \text{ by assump.}} + \sum_{j=1}^n \frac{Y_j}{n h_n^k} \underbrace{\int K_* \left(z_j, \frac{x - X_j}{h_n} \right) dz_j}_{K((x - X_j)/h_n)} \\
&= \sum_{j=1}^n \frac{Y_j}{n h_n^k} K \left(\frac{x - X_j}{h_n} \right)
\end{aligned}$$

Thus:

$$\begin{aligned}
\hat{r}_n(x) &= \frac{1}{f_n(x)} \sum_{j=1}^n \frac{Y_j}{nh_n^k} K\left(\frac{x - X_j}{h_n}\right) \\
&= \left(n^{-1} h_n^{-k} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right) \right)^{-1} \sum_{j=1}^n \frac{Y_j}{nh_n^k} K\left(\frac{x - X_j}{h_n}\right) \\
&= \left(\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right) \right)^{-1} \sum_{j=1}^n Y_j K\left(\frac{x - X_j}{h_n}\right) \\
&= \frac{\sum_{j=1}^n Y_j K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)} \\
&= \sum_{j=1}^n Y_j \left[\frac{K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)} \right]
\end{aligned}$$

This regression function is called the Nadaraya-Watson (NW) kernel regression function estimator. The regression function is just a weighted average of the Y 's.

- See Kelejian's notes for a proof that the NW estimator is CONSISTENT. That is:

$$\hat{r}_n(x) \rightarrow^p r(x).$$

It's tedious, but the same method that was applied above (twice) is used again to show (via Chebychev) that the estimator is asymptotically unbiased and the variance goes to zero. Done.

6 Lecture 6: February 14, 2006

6.1 More on the Method of Kernels

- Some kernel terminology. h_n in the previous derivations is called the “bandwidth” or the “smoothing-parameter.” For a given n , if h is small, the kernel makes things relatively rough while if h is large, the kernel smooths things out. If you talk about 2 times h_n , this is called the “window-width.”

Large Sample Normality

- As shown in the notes, we can get a normality result if you make some assumptions about the boundedness of the variance, the $2 + \delta$ moment, the density function, etc, and have the following two important conditions:

$$(f) : h_n^2(nh_n^k)^{1/2} \rightarrow \mu, \text{ where } 0 \leq \mu < \infty.$$

$$(g) : nh_n^k \rightarrow \infty.$$

Then it can be shown:

$$(nh_n^k)^{1/2}[\hat{r}(x) - r(x)] \rightarrow^d N\left(\mu \frac{b(x)}{f_x(x)}, \text{var}_{\hat{r}}(x)\right).$$

- So note the normalizing factor out front depends on k ! Usually we don't have this and it will lead to a problem in a moment. The variance of this distribution is relatively easy to estimate, but the mean turns out to be very difficult. The $b(x)$ function depends on second partials of unknown functions! So that's not good.
- One way to estimate the large sample moments is the following. Note in condition (f) above, μ could be zero. So if we can choose just the right bandwidth, so that $h_n^2(nh_n^k)^{1/2} \rightarrow 0$, while still maintaining condition (g), we're golden because the large sample mean just becomes zero! Consider the following choice for h :

$$h_n^{2+k/2} = n^{-1/2-\delta},$$

where $0 < \delta < 2/k$. This implies:

$$h_n^{\frac{4+k}{2}} = n^{\frac{-1-2\delta}{2}}$$

$$h_n = n^{\frac{-1-2\delta}{2} \frac{2}{4+k}}$$

$$h_n = n^{\frac{-2-4\delta}{2(4+k)}}$$

$$h_n = n^{\frac{-1}{4+k}} n^{\frac{-2\delta}{4+k}}$$

Thus,

$$nh_n^k = n * n^{\frac{-k-2\delta k}{4+k}}$$

$$nh_n^k = n^{\frac{4+k-k-2\delta k}{4+k}}$$

We need this expression to go to infinity, so we need the exponent to be positive. So:

$$\frac{4+k-k-2\delta k}{4+k} > 0$$

$$4-2\delta k > 0$$

$$\delta k < 2$$

$$\delta < \frac{2}{k}$$

So as long as $\delta \in (0, 2/k)$, conditions (f) and (g) are satisfied and $\mu = 0$. Thus we can use the large sample distribution to do inference.

- Finally, note the normalizing factor which depends on k . The limiting speed of convergence is inversely related to the number of regressors, $k!$ the larger is k , the slower the rate of convergence. This is called the ‘‘Curse of Dimensionality.’’

6.2 Section 9: Nonlinear Systems

- First, lets review some of the elements of linear systems. Consider the structural system:

$$Y_t \Gamma + X_t \Delta = u_t,$$

where X_t are predetermined variables, $u_t \sim iid(0, \Sigma_u)$ are the structural disturbances. This implies a reduced form:

$$Y_t = X_t \Pi + v_t, \quad \Pi = -\Delta \Gamma^{-1}, \quad v_t = u_t \Gamma^{-1}.$$

Thus,

$$E[Y_t | X_t] = X_t \Pi,$$

is the conditional mean prediction equation and,

$$v_t = Y_t - X_t \Pi,$$

is the prediction error.

- For a linear system, this implies that since u is iid, so is v , the reduced form is linear in u , and we can get the prediction equation by simply solving for Y and setting the structural disturbance vector to zero.
- Now consider the following linear system:

$$Y_{t1} = a_1 + a_2 Y_{t2} + a_3 X_t + \epsilon_{t1},$$

$$Y_{t2} = b_1 + b_2 X_t + \epsilon_{t2},$$

with $E[\epsilon_{ti}\epsilon_{tj}] \neq 0$. Estimating the first equation via OLS would be inconsistent due to the correlation between the errors. The first equation is NOT identified because the order condition is violated.

- What about 2SLS? If we regress Y_{t2} on a constant and X_t , we'll get a fitted value $\hat{Y}_{t2} = \hat{\Pi}_0 + \hat{\Pi}_1 X_t$, but the second stage regressor matrix contains a constant, \hat{Y}_{t2} , and X_t , which will have rank 2! So we can't do 2SLS.
- What about introducing a nonlinear instrument like X_t^2 ? While this WON'T work in a linear system like we have, it WILL work in a non-linear system.
- First we need to make some additional assumptions:
 - (1) $|X_t| < M < \infty$ for all $t \geq 1$.
 - (2) $T^{-1}\bar{X}'_K \bar{X}_K \rightarrow Q_K$ where Q_K^{-1} exists, where:

$$\bar{X}_{tK} = (1, X_t, X_t^2, \dots, X_t^K),$$

$$\bar{X}_K = (\bar{X}_{1K}, \dots, \bar{X}_{TK})'.$$

Via Lindberg-Feller CLT, this implies:

$$T^{-1/2}\bar{X}'_k \epsilon_1 \rightarrow^d N(0, \sigma_{11} Q_k).$$

- So consider doing 2SLS with X_t and X_t^2 as instruments. Then:

$$\hat{Y}_{t2} = \hat{\Pi}_0 + \hat{\Pi}_1 X_t + \hat{\Pi}_2 X_t^2.$$

In this case, the second stage regressor matrix will have full rank so the 2SLS estimator is defined. However, $plim T^{-1}\hat{Z}'_1 \hat{Z}_1$, will have rank 2 so the 2SLS estimator is NOT consistent! We could also write this as:

$$plim \hat{\Pi}_2 = 0.$$

- Now lets consider a simple NON-linear model:

$$Y_{t1} = a_1 + a_2 e^{Y_{t2}} + a_3 X_t + \epsilon_{t1},$$

$$Y_{t2} = b_1 + b_2 X_t + \epsilon_{t2}.$$

- The reduced form is thus:

$$Y_{t1} = a_1 + a_2 (e^{b_1 + b_2 X_t + \epsilon_{t2}}) + a_3 X_t + \epsilon_{t1},$$

$$Y_{t2} = b_1 + b_2 X_t + \epsilon_{t2}.$$

- So first note the reduced form equations are NOT linear in all the disturbance terms, unlike in linear systems. What about the prediction equations? Consider:

$$E[Y_{t1}|X_t] = a_1 + a_2 e^{b_1 + b_2 X_t} E[e^{\epsilon_{t2}}] + a_3 X_t = a_1 + a_2 e^{b_1 + b_2 X_t} K + a_3 X_t,$$

$$E[Y_{t2}|X_t] = b_1 + b_2 X_t.$$

By Jensen's inequality:

$$K = E[e^{\epsilon_{t2}}] \neq e^{E(\epsilon_{t2})} = e^0 = 1.$$

Since $K \neq 1$, we cannot get the prediction equation just by setting the errors to zero.

- More next time on prediction error.

7 Lecture 7: February 16, 2006

7.1 More Nonlinear Systems

- Recall the system of structural equations from last lecture:

$$Y_{t1} = a_1 + a_2 e^{Y_{t2}} + a_3 X_t + \epsilon_{t1},$$

$$Y_{t2} = b_1 + b_2 X_t + \epsilon_{t2}.$$

Which induced a reduced form:

$$Y_{t1} = a_1 + a_2 (e^{b_1 + b_2 X_t} e^{\epsilon_{t2}}) + a_3 X_t + \epsilon_{t1},$$

$$Y_{t2} = b_1 + b_2 X_t + \epsilon_{t2}.$$

- If $K = E[e^{\epsilon_{t2}}]$, then we can write:

$$e^{\epsilon_{t2}} = K + v_t,$$

where v_t is mean zero. Substituting this into the first RF equation above:

$$Y_{t1} = a_1 + a_2 e^{b_1 + b_2 X_t} (K + v_t) + a_3 X_t + \epsilon_{t1},$$

or:

$$Y_{t1} = \underbrace{a_1 + a_2 e^{b_1 + b_2 X_t} K + a_3 X_t}_{\text{deterministic}} + \underbrace{w_t}_{\text{stochastic}},$$

with $w_t = \epsilon_{t1} + a_2 e^{b_1 + b_2 X_t} v_t$. w_t is the PREDICTION ERROR!

- Clearly $E[w_t] = 0$, and:

$$\text{Var}[w_t] = E[w_t^2] = \sigma_{11} + a_2^2 e^{2b_1 + 2b_2 X_t} \sigma_v^2 + 2a_2 e^{b_1 + b_2 X_t} \text{cov}(\epsilon_t, v_t) = h(X_t).$$

Therefore w_t is HETEROSKEDASTIC! and so it is NOT iid. Thus the general form of the prediction error:

$$w_t = Y_{t1} - E[Y_{t1}|X_t]$$

is NOT iid, or linear in structural disturbance terms.

Generalization of the Results

- Let an N equation nonlinear system for the $N \times 1$ endogenous vector, Y_t and the $g \times 1$ nonstochastic vector, X_t be,

$$f_1(Y_t, X_t, \epsilon_{t1}) = 0$$

⋮

$$f_N(Y_t, X_t, \epsilon_{tN}) = 0$$

Or just:

$$F(Y_t, X_t, \epsilon_t) = 0_{Nx1},$$

where F is the $Nx1$ vector of functions.

- The reduce form is:

$$Y_t = G(X_t, \epsilon_t),$$

which is in general nonlinear in the disturbances.

- Prediction Equations:

$$E[Y_t|X_t] = E[G(X_t, \epsilon_t)|X_t] \neq G(X_t, E[\epsilon_t]),$$

since G is nonlinear. We CANNOT move the expectation inside and thus we can't just set the errors to zero to solve for these equations.

- Finally the prediction error, $v_t = Y_t - E[Y_t|X_t] = Y_t - E[G(X_t, \epsilon_t)|X_t]$, will be heteroskedastic.

Estimation Issues with Nonlinear Systems

- Consider estimating the nonlinear system above using the following set of instruments:

$$\bar{X}_{t2} = [1, X_t, X_t^2].$$

- First write the first structural equation as:

$$Y_1 = Z_1 A_1 + \epsilon_t,$$

with:

$$Z_1 = \begin{bmatrix} 1 & e^{Y_{12}} & X_1 \\ \vdots & \vdots & \vdots \\ 1 & e^{Y_{T2}} & X_T \end{bmatrix}.$$

- Suppose we regress $e^{Y_{t2}}$ on a constant, X_t , and X_t^2 and form our fitted values, $\widehat{e^{Y_{t2}}}$. Note(!!!):

$$\widehat{e^{Y_{t2}}} \neq e^{\widehat{Y_{t2}}}.$$

So we could write:

$$\widehat{e^{Y_{t2}}} = \hat{\Pi}_0 + \hat{\Pi}_1 X_t + \hat{\Pi}_2 X_t^2.$$

- Then our second stage regressor matrix becomes:

$$\hat{Z}_1 = \begin{bmatrix} 1 & \overbrace{e^{Y_{12}}} & X_1 \\ \vdots & \vdots & \vdots \\ 1 & \overbrace{e^{Y_{T2}}} & X_T \end{bmatrix}.$$

- We could also write:

$$\hat{Z}_1 = \underline{\overline{X}}_2 \hat{P}_1 = [1, X_t, X_t^2] * \begin{bmatrix} 1 & \hat{\Pi}_0 & 0 \\ 0 & \hat{\Pi}_1 & 1 \\ 0 & \hat{\Pi}_2 & 0 \end{bmatrix}.$$

And if $R_2 = \underline{\overline{X}}_2 (\underline{\overline{X}}_2' \underline{\overline{X}}_2)^{-1} \underline{\overline{X}}_2'$, the projection matrix, then,

$$\hat{Z}_1 = R_2 Z_1.$$

- We will show in a moment that if $plim \hat{\Pi}_i = \Pi_i$ for $i = 0, 1$, and 2 , then in general $\Pi_2 \neq 0$. This is all we need to get \hat{P}_1 to have full rank and thus we get a consistent estimator.
- Given this, we have:

$$\hat{A}_1 = (\hat{Z}_1' \hat{Z}_1)^{-1} \hat{Z}_1' Y_1 = A_1 + (\hat{Z}_1' \hat{Z}_1)^{-1} \hat{P}_1' \underline{\overline{X}}_2' \epsilon_1,$$

so,

$$\sqrt{T}(\hat{A}_1 - A_1) \rightarrow^d N(0, \sigma_{11}(P_1' Q_2 P_1)^{-1}),$$

or,

$$\sqrt{T}(\hat{A}_1 - A_1) \rightarrow^d N(0, \sigma_{11} plim T(\hat{Z}_1' \hat{Z}_1)^{-1}).$$

Thus, \hat{A}_1 is consistent.

Proof that $\Pi_2 \neq 0$

- Note,

$$e^{Y_{t2}} = e^{b_1 + b_2 X_t + \epsilon_{t2}} = e^{b_1 + b_2 X_t} e^{\epsilon_{t2}}.$$

- Using $e^{t^2} = K + v_t$,

$$\begin{aligned} e^{Y_{t2}} &= e^{b_1 + b_2 X_t} [K + v_t]. \\ e^{Y_{t2}} &= K e^{b_1 + b_2 X_t} + e^{b_1 + b_2 X_t} v_t = f_t + \Phi_t, \end{aligned}$$

where f_t is deterministic and nonlinear in X_t , while Φ_t is stochastic with mean zero and $Var[\Phi_t] = e^{2b_1 + 2b_2 X_t} \sigma_v^2$ which is UNIFORMLY BOUNDED! This will be important in a moment.

- If we run e^{Y_2} on \bar{X}_2 , we get:

$$\hat{\Pi} = (\bar{X}'_2 \bar{X}_2)^{-1} \bar{X}'_2 e^{Y_2}.$$

So, $E[\hat{\Pi}] = (\bar{X}'_2 \bar{X}_2)^{-1} \bar{X}'_2 f$, and,

$$VC(\hat{\Pi}) = (\bar{X}'_2 \bar{X}_2)^{-1} \bar{X}'_2 \Omega_{\Phi} \bar{X}_2 (\bar{X}'_2 \bar{X}_2)^{-1},$$

or,

$$VC(\hat{\Pi}) = \underbrace{T^{-1}}_{\rightarrow 0} \underbrace{[T(\bar{X}'_2 \bar{X}_2)^{-1}]}_{\rightarrow Q_2^{-1}} \underbrace{[T^{-1} \bar{X}'_2 \Omega_{\Phi} \bar{X}_2]}_{\rightarrow M < \infty} [T(\bar{X}'_2 \bar{X}_2)^{-1}] \rightarrow 0.$$

Note the Ω_{Φ} matrix is diagonal and uniformly bound.

- Thus, by Chebychev,

$$plim \hat{\Pi} = lim (\bar{X}'_2 \bar{X}_2)^{-1} \bar{X}'_2 f,$$

which looks a lot like a regression of f on \bar{X}_2 . So $plim \hat{\Pi} = (plim \hat{\Pi}_0, plim \hat{\Pi}_1, plim \hat{\Pi}_2)'$. So this last coefficient is the one on X^2 . Note that since f is nonlinear, the best fitting polynomial to a nonlinear function will in general NOT be linear! So $lim \hat{\Pi}_2 \neq 0$ in general!

- The next thing we want to consider is using higher order X 's beyond X^2 . Why not include higher degrees of the X 's as instruments. As long we have enough observations, this cannot hurt! In our example, if we use the instruments \bar{X}_K , then all we need to get a matrix of rank 3 is for ONE of the higher order (two or above) X 's to have a non-zero coefficient. We'll get consistency if this is satisfied.
- So by similar reasoning to above,

$$\sqrt{T}(\hat{A}_1^K - A_1) \rightarrow^d N(0, \sigma_{11} plim T(\hat{Z}_1^{K'} \hat{Z}_1^K)^{-1}).$$

- **Remark** If $K \leq L$, then,

$$plim T(\hat{Z}_1^{K'} \hat{Z}_1^K)^{-1} - plim T(\hat{Z}_1^{L'} \hat{Z}_1^L)^{-1}$$

is POSITIVE semi-definite! Thus, by including more and more powers of the X 's, we get a more efficient estimator. As we estimate with a higher order polynomial, our approximation of the nonlinear f , becomes better. See Kelejian's note for a very confusing proof of this. Note this works for the model because the nonlinearity comes in the form of an exponential function which has a best approximation of an infinite order polynomial.

8 Lecture 8: February 21, 2006

8.1 More Nonlinear Systems

- The lower bound for the VC matrix for the 2SLS estimator in a non-linear system is as follows:

$$VC^* = \sigma_{11} plim T[(EZ_1)'(EZ_1)]^{-1}.$$

Proof: If we use the instruments \bar{X}_K , the VC matrix can be written:

$$\sigma_{11} plim T[(EZ_1)'R_K(EZ_1)]^{-1}.$$

So rewrite part of VC^* above as (identity):

$$(EZ_1)'(EZ_1) = (EZ_1)'R_K(EZ_1) + (EZ_1)'[I - R_K](EZ_1).$$

Or just $A = B + C$. Note that B is positive definite and C is positive semi-definite. Thus we have the result that $B^{-1} - A^{-1}$ is positive semidefinite, or:

$$[(EZ_1)'R_K(EZ_1)]^{-1} - [(EZ_1)'(EZ_1)]^{-1} \geq 0.$$

Which proves our result.

- We can also do all this non-linear stuff in more general form. Consider the i^{th} equation of an M equation system:

$$y_i = f_i(B_i) + u_i = f_i + u_i,$$

where y_i is $T \times 1$ and f_i is a $T \times 1$ vector of values relating to the RHS of the equation. We could write the model like this if we had something like:

$$Y_{t1} = b_1 + b_2 X_{t1} + b_3 e^{b_4 Y_{t2}} + u_{t1}.$$

- Still more generally is when we can't separate out the LHS endogenous variable. In this case we write the model as:

$$F_i(B_i) = u_i.$$

See notes for an example involving heteroskedastic errors, which once corrected for, yields a LHS variable which is a function of the predetermined variables and parameters.

8.2 Useful Formulas and Definitions

- If Y is $M \times 1$ and X is $N \times 1$, then $\frac{\partial Y}{\partial X}$ is $M \times N$. Of course:

$$\left(\frac{\partial Y}{\partial X}\right)' = \underbrace{\frac{\partial Y}{\partial X'}}_{N \times M}.$$

- If $Y = AX$ and X is a vector and A is a matrix, then $\frac{\partial Y}{\partial X} = A$.
- If $Y = AX$ and the elements of X are functions of a vector, α , then,

$$\frac{\partial Y}{\partial \alpha} = \frac{\partial Y}{\partial X} \frac{\partial X}{\partial \alpha} = A \frac{\partial X}{\partial \alpha}.$$

So that's nice.

- If $Y = Z'AX$, a scalar, then clearly, $Y = Y' = X'A'Z$, where both X and Z are functions of α but $A = A'$ is not, then:

$$\frac{\partial Y}{\partial \alpha} = X'A' \frac{\partial Z}{\partial \alpha} + Z'A \frac{\partial X}{\partial \alpha}.$$

And if $Z = X$,

$$\frac{\partial Y}{\partial \alpha} = 2X'A' \frac{\partial X}{\partial \alpha}.$$

The Considered Model

- Consider:

$$y_{it} = f_i(Y_{it}, X_{it}, B_i^0) + u_{it}, \quad i = 1, \dots, M; \quad t = 1, \dots, T.$$

Lets write this as:

$$y_{it} = f_{it}(B_i^0) + u_{it},$$

or even more simply as:

$$y_i = f_i(B_i^0) + u_i,$$

where y_i is stacked with dimension $T \times 1$.

- Let $W_{T \times r}$ be our matrix of instruments such that $r \geq K_i$, ie, equation i is identified. W could contain a constant and powers of the X 's.
- So we now will state the distribution of the two-stage nonlinear least squares estimator for B_i^0 . We need some assumptions:
 - (a.1) The parameters space is open and convex.

- (a.2) $u_{it} \sim iid(0, \sigma_{ii})$. So the errors are iid within equations. Not necessarily across equations.
 - (a.3) The elements of W are bounded and $T^{-1}W'W$ converges to something of full rank.
 - (a.4) The first derivatives of the f_i function exist and is continuous in the neighborhood of the true parameter value.
 - (a.5) We can expand the f_i function around B_i^0 such that $G_* = \left. \frac{\partial f_i}{\partial B} \right|_{B^*}$. Then $T^{-1}W'G_* \rightarrow^p M(B)$, a matrix of full column rank.
 - (a.6) Second partials exist and are continuous and bounded.
- Then if we minimize the objective function:

$$\text{Min}_B \{Q_T(B) = [y_i - f_i(B)]' R_W [y_i - f_i(B)]\},$$

where $R_W = W(W'W)^{-1}W'$, then there IS A ROOT of the minimization (possibly one of many), call it \hat{B}_i , such that:

$$\sqrt{T}(\hat{B}_i - B_i^0) \rightarrow^d N\left(0, \sigma_{ii} \text{plim } T \left[\left(\frac{\partial f_i(B)}{\partial B} \right)' R_W \left(\frac{\partial f_i(B)}{\partial B} \right) \right]_{\hat{B}_i}^{-1}\right).$$

So \hat{B}_i is the NL2SLS estimator for B_i^0 .

- Insights from all this: \hat{B}_i is consistent. Small sample guidance:

$$\hat{B}_i \approx N\left(B_i^0, \hat{\sigma}_{ii} \left[\left(\frac{\partial f_i(B)}{\partial B} \right)' R_W \left(\frac{\partial f_i(B)}{\partial B} \right) \right]_{\hat{B}_i}^{-1}\right).$$

With:

$$\hat{\sigma}_{ii} = T^{-1}[y_i - f_i(\hat{B}_i)]'[y_i - f_i(\hat{B}_i)].$$

- See Amemiya for a proof of the consistency of the estimator. There is some nice intuition on page 21 of Kelejians notes which shows that:

$$\text{if } B = B_i^0, \text{ then: } T^{-1}Q_T(B) \rightarrow^p 0,$$

but,

$$\text{if } B \neq B_i^0, \text{ then: } T^{-1}Q_T(B) \rightarrow^p \gamma > 0.$$

9 Lecture 9: February 23, 2006

9.1 More Nonlinear Systems

A Useful Preliminary

- Suppose we have a consistent estimator, $\hat{\theta} \xrightarrow{p} \theta_0 \in S$, which is open and convex. Consider a function of our estimator, $g_T(\hat{\theta})$, and assume:

$$plim g_T(\theta_0) = \mu.$$

- Also assume:

- $\frac{\partial g_T(\theta)}{\partial \theta}$ exists and is continuous.
- $|\frac{\partial g_T(\theta)}{\partial \theta}| < c_g < \infty$.

Then:

$$plim g_T(\hat{\theta}) = plim g_T(\theta_0) = \mu.$$

- So the function evaluated at the estimator converges to the function evaluated at the true parameter value.

Asymptotic Normality of the NL2SLS Estimator

- See notes for this. Pg 22-26.
- Our NL2SLS estimator is consistent and asymptotically normal.
- This section also includes a note about the minimum VC matrix for these types of models:

$$VC_{min} = \sigma_{ii} plim T \left[\left(\frac{E \partial f_i(B_i^0)}{\partial B} \right)' \left(\frac{E \partial f_i(B_i^0)}{\partial B} \right) \right]^{-1}.$$

But our usual actual VC matrix is:

$$VC = \sigma_{ii} plim T \left[E \left(\frac{\partial f_i(B_i^0)}{\partial B} \right)' R_W E \left(\frac{\partial f_i(B_i^0)}{\partial B} \right) \right]^{-1}.$$

We would want to include high order polynomials in our instruments, W , to make this as small as possible.

NL3SLS Estimator

- Consider the model:

$$Y_i = f_i^0 + u_i, \quad i = 1, \dots, M,$$

where M is the number of equations we estimate. The system may be larger but we truncate the system at M equations.

- All usual assumptions hold and now we assume $E[u_i u_j'] = \sigma_{ij} I$, for $i, j = 1, \dots, M$.
- Consider stacking the system:

$$y = f^0 + u,$$

where y is $MT \times 1$ for example.

- In this case we can write:

$$E[uu'] = \Sigma_u \otimes I_T.$$

- Then our Three Stage Non-linear Least Squares estimator of B^0 is \hat{B} where \hat{B} minimizes:

$$\Phi(B) = T^{-1}(Y - f)' [\hat{\Sigma}_u^{-1} \otimes W(W'W)^{-1}W'] \underbrace{(Y - f)}_u,$$

where $\hat{\Sigma}_u^{-1}$ is the estimator of Σ_u based on NL2SLS. Again W is our matrix of instruments which clearly will contain all predetermined variables (and squares, etc) in the M equation system, but also may contain X 's from outside the system.

- So the large sample distribution of our NL3SLS estimator is as follows:

$$\sqrt{T}(\hat{B} - B_0) \rightarrow^d N\left(0, \text{plim } T \left[\left(\frac{\partial f}{\partial B} \right)' [\Sigma_u^{-1} \otimes R_W] \left(\frac{\partial f}{\partial B} \right) \right]^{-1} \right).$$

- This induces small sample guidance:

$$\hat{B} \approx N\left(B_0, \left[\left(\frac{\partial f}{\partial B} \right)'_{\hat{B}} [\hat{\Sigma}_u^{-1} \otimes R_W] \left(\frac{\partial f}{\partial B} \right)_{\hat{B}} \right]^{-1} \right).$$

- Example. Consider a linear model: $f_i = Z_i B_i$ so $\frac{\partial f_i}{\partial B_i} = Z_i$. Note that $\hat{Z}_i = R_W Z_i$, so $Z_i' R_W Z_j = \hat{Z}_i' \hat{Z}_j'$, the VC matrix becomes:

$$\begin{aligned} VC &= \left[\left(\frac{\partial f}{\partial B} \right)'_{\hat{B}} [\hat{\Sigma}_u^{-1} \otimes R_W] \left(\frac{\partial f}{\partial B} \right)_{\hat{B}} \right]^{-1} \\ &= \left[Z' [\hat{\Sigma}_u^{-1} \otimes R_W] Z \right]^{-1} \\ &= \left[\hat{Z}' [\hat{\Sigma}_u^{-1} \otimes I] \hat{Z} \right]^{-1} \end{aligned}$$

Consistency of the NL3SLS Estimator via Amemiya's Theorem 4.1.2.

- Recall Amemiya's Theorem: Suppose (A) the parameter space is open and convex, (B) the first partial of the objective function exists and is continuous in the neighborhood around the true parameter value, and (C) $T^{-1}Q_T(B) \xrightarrow{p} Q(B)$ uniformly for B in an open set about B_i^0 and $Q(B)$ has a unique minimum at B_i^0 . Given these conditions, there is a CONSISTENT root of:

$$\frac{\partial Q_T(B)}{\partial B} = 0.$$

- So we have assumed (A) and (B), so we will now show (C) holds.
- First note:

$$plim T^{-1}u'(\hat{\Sigma}_u^{-1} \otimes R_W)u = plim \sum_{j=1}^m \sum_{i=1}^m (T^{-1}u'_i W) \underbrace{(T(W'W)^{-1})}_{\rightarrow \Omega^{-1}} \underbrace{(T^{-1}W'u_j)}_{\rightarrow 0} \hat{\sigma}^{ij} = 0.$$

This means that all terms involving a u will plim to zero.

- Recall our model: $Y = f^0 + u$, so write:

$$Y - f = Y - f^0 + f^0 - f = u + \underbrace{f^0 - f}_{\Delta} = u + \Delta.$$

Back to our objective function:

$$\Phi(B) = T^{-1}(Y - f)'[\hat{\Sigma}_u^{-1} \otimes W(W'W)^{-1}W'](Y - f).$$

Or,

$$\begin{aligned} \Phi(B) &= T^{-1}(u + \Delta)'[\hat{\Sigma}_u^{-1} \otimes R_W](u + \Delta). \\ \Phi(B) &= T^{-1} \sum_{j=1}^m \sum_{i=1}^m (u_i + \Delta_i)' W(W'W)^{-1}W'(u_j + \Delta_j) \hat{\sigma}^{ij}. \end{aligned}$$

Take the limit noting that all terms involving the u 's all go to zero:

$$plim \Phi(B) = plim T^{-1} \sum_{j=1}^m \sum_{i=1}^m (u_i + \Delta_i)' W(W'W)^{-1}W'(u_j + \Delta_j) \hat{\sigma}^{ij}.$$

$$plim \Phi(B) = plim T^{-1} \sum_{j=1}^m \sum_{i=1}^m \Delta_i' W(W'W)^{-1}W'\Delta_j \hat{\sigma}^{ij}.$$

- So consider the term $T^{-1}W'\Delta_j$:

$$\begin{aligned}
T^{-1}W'\Delta_j &= T^{-1}W'[f_j^0 - f_j] \\
&= T^{-1}W'[f_j^0 - f_j^0 - \frac{\partial f_j}{\partial B_j} \Big|_{\tilde{B}_j}](B_j - B_j^0) \\
&\quad \text{via the mean value theorem} \\
&= -T^{-1}W' \frac{\partial f_j}{\partial B_j} \Big|_{\tilde{B}_j} (B_j - B_j^0) \\
&\rightarrow M_j(B_j)(B_j - B_j^0) \equiv C_j < \infty
\end{aligned}$$

Note that $C_j \neq 0$ unless $B_j = B_j^0$.

- So back in our limiting objective function:

$$plim \Phi(B) = \sum_{j=1}^m \sum_{i=1}^m C_i' [\Sigma_u^{-1} \otimes \Omega_W^{-1}] C_j > 0.$$

Thus $plim \Phi(B)$ is UNIQUELY minimized at $C_i = 0, i = 1, \dots, M$ which requires that $B = B_0$, consistent!

- A final note on the minimum VC matrix in three stage. The best we can normally do is:

$$VC_{min} = plim T \left[E \left(\frac{\partial f}{\partial B} \right)' [\Sigma_u^{-1} \otimes I_T] E \left(\frac{\partial f}{\partial B} \right) \right]_{B_0}^{-1}.$$

But our usual actual VC matrix is:

$$VC = plim T \left[E \left(\frac{\partial f}{\partial B} \right)' [\Sigma_u^{-1} \otimes R_W] E \left(\frac{\partial f}{\partial B} \right) \right]^{-1}.$$

9.2 Section 10: Qualitative and Limited Dependent Variable Models

- Consider a dependent variable, Y_t , that takes on two values, 0 or 1. Examples might be “a firm merges or doesn’t”, “a woman works or doesn’t”, “a person votes or doesn’t”, etc.
- Consider the linear model:

$$Y_t = X_t B + \epsilon_t = f_t + \epsilon_t,$$

with $E[\epsilon_t] = 0$. Clearly $E[Y_t] = X_t B = f_t$.

- In general, Y has density, $g(Y_t = 1) = P_t$ and $g(Y_t = 0) = 1 - P_t$, so $E[Y_t] = P_t$. $E[Y_t^2] = P_t$, so:

$$\sigma_{Y_t}^2 = E[Y_t^2] - [E(Y_t)]^2 = P_t - P_t^2 = P_t(1 - P_t).$$

- So we might interpret $X_t B^{OLS}$ as the estimate of the probability that Y_t is equal to one. Or $X_t B^{OLS} = f_t = P_t$.
- Clearly there are problems.
 - $X_t B^{OLS}$ is NOT bounded by zero and one, as a probability should be.
 - ϵ_t is not homoskedastic.

To see the second problem, note since $g(Y_t = 1) = f_t$ and $g(Y_t = 0) = 1 - f_t$, then:

$$Y_t = 1 \implies \epsilon_t = 1 - f_t \implies g(\epsilon_t) = f_t,$$

$$Y_t = 0 \implies \epsilon_t = -f_t \implies g(\epsilon_t) = 1 - f_t.$$

Then:

$$E[\epsilon_t] = (1 - f_t)f_t - f_t(1 - f_t) = f_t - f_t = 0,$$

$$Var[\epsilon_t] = E[\epsilon_t^2] = (1 - f_t)^2 f_t + f_t^2(1 - f_t) = (1 - f_t)f_t,$$

which is NEGATIVE when $f_t > 1$ or $f_t < 0$. Not a good thing for a variance.

- More next time.

10 Lecture 10: February 28, 2006

10.1 More Qualitative Models

- Suppose instead of letting $P_t = X_t B$, we let $P_t = F(X_t B)$, where F is a CDF of something. Then this would map positive and negative values into something between zero and one as required.
- The problem: which CDF to use and how do we formulate X_t ?
- There are two main CDFs which are often used:
 - Probit Model - $N(0, 1)$:

$$P_t = \text{Prob}(Y_t = 1) = F(X_t B) = \int_{-\infty}^{X_t B} (2\pi)^{-1/2} e^{-1/2u^2} du.$$

- Logit Model:

$$P_t = \text{Prob}(Y_t = 1) = F(X_t B) = \frac{e^{X_t B}}{1 + e^{X_t B}}.$$

The logit has slightly heavier tails than the probit. Note also,

$$\text{Prob}(Y_t = 0) = 1 - F(X_t B) = 1 - \frac{e^{X_t B}}{1 + e^{X_t B}} = \frac{1 + e^{X_t B}}{1 + e^{X_t B}} - \frac{e^{X_t B}}{1 + e^{X_t B}} = \frac{1}{1 + e^{X_t B}}.$$

- A couple things about the logit model. First, it's density:

$$f(Z) = \frac{dF}{dZ} = \frac{e^Z}{(1 + e^Z)^2}.$$

Note that it's also symmetric:

$$f(-Z) = \frac{e^{-Z}}{(1 + e^{-Z})^2} = \frac{e^{-Z}}{(1 + e^{-Z})^2} \frac{e^{2Z}}{e^{2Z}} = \frac{e^Z}{1 + 2e^Z + e^{2Z}} = f(Z).$$

Finally, it can be shown that if Z follows the logit $f(Z)$ above,

$$Z \sim \left(0, \frac{\pi^2}{3}\right).$$

- Now consider the probit model. Should we always use the standard normal, or should we parameterize the mean and variance? Suppose we have the following:

$$X_t B = b_0 + X_{t1} b_1.$$

Then,

$$F(X_t B) = \text{Pr}(u \leq b_0 + X_{t1} b_1),$$

where u is standard normal. What if we assumed $Z \sim N(\mu, \sigma^2)$, and we wrote:

$$\begin{aligned}
P_t = \text{Prob}(Y_t = 1) = F(X_t B) &= \text{Pr}(Z \leq b_0 + X_{t1} b_1) \\
&= \text{Pr}(Z - \mu \leq b_0 - \mu + X_{t1} b_1) \\
&= \text{Pr}\left(\frac{Z - \mu}{\sigma} \leq \frac{b_0 - \mu}{\sigma} + X_{t1} \frac{b_1}{\sigma}\right) \\
&= \text{Pr}(u \leq b_0^* + X_{t1} b_1^*)
\end{aligned}$$

where u is standard normal again!

So we might as well just work with a $N(0, 1)$. By including an intercept, we can absorb any non-zero mean and the variance just scales things. So WLOG, we can always include an intercept and work with a standard normal.

- But what should we report in a paper? The coefficients we estimate don't have much meaning, but we could estimate an elasticity of P_t wrt X_{tk} , such as (for the logit):

$$\begin{aligned}
\frac{\partial P_t}{\partial X_{t5}} \frac{X_{t5}}{P_t} &= \frac{(1 + e^{X_t B}) B_5 e^{X_t B} - e^{X_t B} B_5 e^{X_t B}}{(1 + e^{X_t B})^2} \frac{X_{t5} (1 + e^{X_t B})}{e^{X_t B}} \\
&= \frac{B_5 e^{X_t B}}{(1 + e^{X_t B})^2} \frac{X_{t5} (1 + e^{X_t B})}{e^{X_t B}} \\
&= \frac{B_5 X_{t5}}{1 + e^{X_t B}}
\end{aligned}$$

Of course this can be evaluated for any t , but usually it's evaluated at the mean or at several different quantiles.

- So what about X_t ? What types of things should we put inside? We may have a model to predict the probability that Y_t is one, but it also may be an ad-hoc formulation.

Illustration

- Consider the choice to drive to work by a group of individuals. Let Z_{t1} be characteristics of driving a car for the t^{th} person (how long it takes, how much it costs, etc).
- Let X_t be the characteristics of person t (income, age, etc). Let Z_{t0} be a vector of variables on an alternative form of transport.
- Consider the utility model with:

$$\text{Utility if No Drive: } u_{t0} = \alpha_0 + Z_{t0} B + X_t \gamma_0 + \epsilon_{t0}.$$

$$\text{Utility if Drive: } u_{t1} = \alpha_1 + Z_{t1} B + X_t \gamma_1 + \epsilon_{t1}.$$

Note in this model, they assumed that B was the same in both utility functions, but clearly it shouldn't be.

- Thus we have:

$$\begin{aligned}
Pr(Y_t = 1) &= Pr(Drive_t) \\
&= Pr(u_{t1} > u_{t0}) \\
&= Pr(\alpha_1 + Z_{t1}B + X_t\gamma_1 + \epsilon_{t1} > \alpha_0 + Z_{t0}B + X_t\gamma_0 + \epsilon_{t0}) \\
&= Pr(\epsilon_{t0} - \epsilon_{t1} < (\alpha_1 - \alpha_0) + (Z_{t1} - Z_{t0})B + X_t(\gamma_1 - \gamma_0)) \\
&\quad \text{if } \epsilon_{t0} - \epsilon_{t1} \sim N(\mu, \sigma^2) \\
&= Pr\left(\frac{\epsilon_{t0} - \epsilon_{t1} - \mu}{\sigma} < \frac{(\alpha_1 - \alpha_0 - \mu)}{\sigma} + (Z_{t1} - Z_{t0})\frac{B}{\sigma} + X_t\frac{(\gamma_1 - \gamma_0)}{\sigma}\right) \\
&= Pr\left(\frac{\epsilon_{t0} - \epsilon_{t1} - \mu}{\sigma} < b_0 + (Z_{t1} - Z_{t0})B^* + X_t\gamma^*\right) \\
&= F(b_0 + (Z_{t1} - Z_{t0})B^* + X_t\gamma^*)
\end{aligned}$$

- A note on stochastic utility models. As above, we can write:

$$u_{ti} = u(X_t, Z_{ti}, H_{ti}, \epsilon_t),$$

where H_{ti} are either unknown or more likely, we know about them but do not have data. The other variables have been described before. Thus we view H as random and condition on X and Z :

$$E[u_{ti}|X_t, Z_{ti}] = \bar{u}(X_t, Z_{ti}, \lambda_i),$$

where we observe X and Z . Then a first order polynomial expansion yields:

$$\bar{u}(X_t, Z_{ti}, \lambda_i) = \alpha(\lambda_i) + X_t\gamma(\lambda_i) + Z_{ti}B(\lambda_i),$$

or,

$$\bar{u}(X_t, Z_{ti}, \lambda_i) = \alpha_i + X_t\gamma_i + Z_{ti}B_i,$$

and note the B is indexed by i !! So we assumed this away before but it should have been different in each case.

Estimation of Probits and Logits

- Proofs for consistency, efficiency, and asymptotic normality are assumed though unavailable.
- Note if $Y_t = 1$ with probability F_t and $Y_t = 0$ with probability $1 - F_t$, we can write:

$$f(Y_t) = F_t^{Y_t}(1 - F_t)^{1-Y_t}.$$

- We can estimate the model several ways, but the first and most common is maximum likelihood.

- (1) **Maximum Likelihood Estimation.** Consider the likelihood function:

$$\mathcal{L} = \prod_{t=1}^T F_t^{Y_t} (1 - F_t)^{1 - Y_t},$$

or often written:

$$\mathcal{L} = \prod_{t \in T_1} F_t \prod_{t \in T_0} 1 - F_t,$$

where T_1 is the set where the Y_t 's are 1 and T_0 is the set where the Y_t 's are 0. This induces small sample guidance:

$$\hat{B}_{MLE} \approx N(B, \hat{R}^{-1}),$$

where,

$$\hat{R} = - \left. \frac{\partial^2 \log(\mathcal{L})}{\partial B \partial B'} \right|_{\hat{B}}.$$

Following Amemya, we can show that the ML estimator is consistent. See notes. The objective function will be uniquely maximized at the true parameter.

- (2) **Non-Linear Least Squares.** Write the model as:

$$Y_t = F(X_t B) + \epsilon_t,$$

with $E[\epsilon_t] = 0$ and,

$$var(\epsilon_t) = F_t(1 - F_t).$$

Thus,

$$Y_t - F(X_t B) = \epsilon_t,$$

$$\underbrace{\frac{Y_t - F(X_t B)}{\sqrt{F_t(1 - F_t)}}}_{g_t} = \underbrace{\frac{\epsilon_t}{\sqrt{F_t(1 - F_t)}}}_{v_t}.$$

Thus $v_t \sim (0, 1)$. If we then minimize the sum of the squared v_t 's, we get:

$$\hat{B}_{NLLS} \approx N\left(B, \left[\left(\frac{G}{B}\right)' \left(\frac{G}{B}\right)\right]_{\hat{B}}^{-1}\right),$$

where $G' = (g_1, \dots, g_T)$.

- (3) **GMM.** See notes. This isn't as good as ML.
- (4) **NLLS without Heteroskedasticity Correction.** See notes. This isn't as good as ML but would be computationally simpler.
- So in sum, unless you're having trouble, do maximum likelihood.
- We'll next move to measures of fit and model selection. Which should we use, probit or logit? How do we assess if our model is doing a good job at prediction?

11 Lecture 11: March 2, 2006

11.1 More Qualitative Models

- Suppose we have our data on Y_t which are zeros and ones and we estimate (by logit or probit), $E[Y_t] = \hat{P}_t = X_t\hat{B}$. We might create something like:

$$\hat{Y}_t = 1 \text{ if } \hat{P}_t \geq \frac{1}{2},$$

$$\hat{Y}_t = 0 \text{ if } \hat{P}_t < \frac{1}{2}.$$

And then compute:

$$(IP/N) = \frac{1}{N} \sum_{t=1}^N (Y_t - \hat{Y}_t)^2,$$

which is the proportion of incorrect predictions (scaled by N). The lower the value, the better is our model.

- Note the MINIMUM MSE predictor of Y_t is $E[Y_t|X_t] = P_t$. For any other predictor, say the one above, \hat{P}_t , is must be:

$$E(Y_t - \hat{Y}_t)^2 > E(Y_t - P_t)^2,$$

ie the conditional mean is the best we can do. But it still might seem intuitive to say that $\hat{Y}_t = 1$ if our estimate of the probability was 0.99. This is easily verified for a simple example in the notes.

- Another way to access our model's worth is the Error Sum of Squares:

$$ESS/N = \frac{1}{N} \sum_{t=1}^N (Y_t - \hat{P}_t)^2.$$

- A third way to measure fit is the square of the sample correlation coefficient between Y_t and the estimated probability \hat{P}_t . Specifically:

$$\hat{\rho}_{Y,\hat{P}}^2 = \frac{[\sum_{t=1}^N (Y_t - \bar{Y})(\hat{P}_t - \bar{\hat{P}})]^2}{\sum_{t=1}^N (Y_t - \bar{Y})^2 \sum_{t=1}^N (\hat{P}_t - \bar{\hat{P}})^2}.$$

Clearly the higher is this correlation, the better our model does.

Sample Development of $R^2 = \rho^2$ in the Logit Model

- This is important so review before exams.

- Consider a bernoulli variable, Y_t , with $f(Y_t = 1) = P_t$ and $f(Y_t = 0) = 1 - P_t$. Thus $E[Y_t] = P_t$ and $Var(Y_t) = P_t(1 - P_t)$. Write:

$$Y_t = P_t + \epsilon_t, \epsilon_t \sim (0, P_t(1 - P_t)).$$

Note,

$$\bar{Y} = \bar{P} + \bar{\epsilon},$$

so,

$$Y_t - \bar{Y} = P_t - \bar{P} + \epsilon_t - \bar{\epsilon}.$$

- Then the correlation between Y_t and P_t is:

$$\begin{aligned}
\hat{\rho}^2 &= \frac{[\sum_{t=1}^N (Y_t - \bar{Y})(P_t - \bar{P})]^2}{\sum_{t=1}^N (Y_t - \bar{Y})^2 \sum_{t=1}^N (P_t - \bar{P})^2} \\
&= \frac{[\sum (Y_t - \bar{Y})(P_t - \bar{P})/N]^2}{\sum \frac{(Y_t - \bar{Y})^2}{N} \sum \frac{(P_t - \bar{P})^2}{N}} \\
&= \frac{[\sum (P_t - \bar{P} + \epsilon_t)(P_t - \bar{P})/N]^2}{\sum \frac{(P_t - \bar{P} + \epsilon_t)^2}{N} \sum \frac{(P_t - \bar{P})^2}{N}} \\
&= \frac{[\sum (P_t - \bar{P})^2/N + \epsilon_t(P_t - \bar{P})/N]^2}{\sum \frac{1}{N} [(P_t - \bar{P})^2 + 2\epsilon_t(P_t - \bar{P}) + \epsilon_t^2] \frac{1}{N} \sum (P_t - \bar{P})^2} \\
&\quad \text{since } N^{-1} \sum (P_t - \bar{P})\epsilon_t \rightarrow 0 \text{ and} \\
&\quad E[\epsilon_t^2] = P_t(1 - P_t) \\
\hat{\rho}^2 &\rightarrow \frac{[\sum (P_t - \bar{P})^2/N]^2}{\sum \frac{1}{N} [(P_t - \bar{P})^2 + \epsilon_t^2] \frac{1}{N} \sum (P_t - \bar{P})^2} \\
\hat{\rho}^2 &\rightarrow \frac{[\sum (P_t - \bar{P})^2/N]^2}{\sum \frac{1}{N} (P_t - \bar{P})^2 [\frac{1}{N} \sum (P_t - \bar{P})^2 + \frac{1}{N} \sum \epsilon_t^2]} \\
\hat{\rho}^2 &\rightarrow \frac{\sum (P_t - \bar{P})^2/N}{\frac{1}{N} \sum (P_t - \bar{P})^2 + \frac{1}{N} \sum \epsilon_t^2} \\
\hat{\rho}^2 &\rightarrow \frac{N^{-1} \sum (P_t - \bar{P})^2}{N^{-1} \sum (P_t - \bar{P})^2 + N^{-1} \sum P_t(1 - P_t)} \\
\hat{\rho}^2 &\rightarrow \frac{N^{-1} \sum (P_t - \bar{P})^2}{N^{-1} \sum [P_t^2 - 2P_t\bar{P} + \bar{P}^2 + P_t - P_t^2]} \\
\hat{\rho}^2 &\rightarrow \frac{N^{-1} \sum (P_t - \bar{P})^2}{-2\bar{P}N^{-1} \sum P_t + \bar{P}^2 + N^{-1} \sum P_t} \\
\hat{\rho}^2 &\rightarrow \frac{N^{-1} \sum (P_t - \bar{P})^2}{-2\bar{P}^2 + \bar{P}^2 + \bar{P}} \\
\hat{\rho}^2 &\rightarrow \frac{N^{-1} \sum (P_t - \bar{P})^2}{\bar{P} - \bar{P}^2}
\end{aligned}$$

So $\hat{\rho}^2$ depends on the sample variance of P_t , \bar{P} and \bar{P}^2 .

A Probit Model in a Panel Data Framework - Random Effects

- Suppose we have a latent variable, y_{it}^* which we do not observe such as the expected value of an investment. What we do observe is y_{it} which is either 1 or 0 if the individual i in period t chooses to invest or not invest. Thus,

$$y_{it}^* = x_{it}\beta + \epsilon_{it}, \quad t = 1, \dots, T_i, \quad i = 1, \dots, N.$$

$$y_{it} = 1 \text{ if } y_{it}^* > 0,$$

$$y_{it} = 0 \text{ if } y_{it}^* \leq 0.$$

- Assume we have the following error component structure:

$$\epsilon_{it} = v_{it} + u_i, \quad v_{it} \sim iid N(0, 1), \quad u_i \sim iid N(0, \sigma_u^2),$$

and the $\{v_{it}\}$ and $\{u_i\}$ processes are independent. Since u is only independent over the i 's and not the t 's, the ϵ 's are not iid over both i and t so we're going to have problems. Note since v is $N(0, 1)$, then,

$$\epsilon_{it}|u_i \sim N(u_i, 1).$$

- If all was well behaved and the ϵ 's were iid over both i and t , we could do a probit:

$$Pr(y_{it} = 1) = Pr(x_{it}\beta + \epsilon_{it} > 0) = Pr(\epsilon_{it} < x_{it}\beta) = F(x_{it}\beta).$$

- So consider the likelihood for the sample:

$$L = \prod_{i=1}^N L_i,$$

$$L_i = \int_{Low_{i,T_i}}^{Up_{i,T_i}} \cdots \int_{Low_{i,1}}^{Up_{i,1}} f(\epsilon_{i1}, \dots, \epsilon_{iT_i}) d\epsilon_{i1} \cdots d\epsilon_{iT_i},$$

where,

$$\text{if } y_{it} = 0, \text{ then } Low_{it} = -\infty, \text{ and } Up_{it} = -x_{it}\beta$$

$$\text{if } y_{it} = 1, \text{ then } Low_{it} = -x_{it}\beta, \text{ and } Up_{it} = \infty$$

- This is clearly a mess. We have for each equation, i , T_i integrals over a joint density which doesn't factor. So consider the following simplification:

$$f(\epsilon_{i1}, \dots, \epsilon_{iT_i}, u_i) = f(\epsilon_{i1}, \dots, \epsilon_{iT_i}|u_i)f(u_i).$$

So,

$$f(\epsilon_{i1}, \dots, \epsilon_{iT_i}) = \int_{-\infty}^{\infty} f(\epsilon_{i1}, \dots, \epsilon_{iT_i}|u_i)f(u_i)du_i,$$

if we integrate out the u_i 's. By independent of the ϵ 's across time:

$$f(\epsilon_{i1}, \dots, \epsilon_{iT_i}) = \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(\epsilon_{it}|u_i) \right] f(u_i) du_i.$$

- Now substitute this into the above likelihood:

$$L_i = \int_{Low_{i,T_i}}^{Up_{i,T_i}} \cdots \int_{Low_{i,1}}^{Up_{i,1}} \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} f(\epsilon_{it}|u_i) \right] f(u_i) du_i d\epsilon_{i1} \cdots d\epsilon_{iT_i}.$$

And rearrange:

$$L_i = \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \int_{Low_{i,t}}^{Up_{i,t}} f(\epsilon_{it}|u_i) d\epsilon_{it} \right] f(u_i) du_i.$$

So note the density in this last equation is $f(\epsilon_{it}|u_i)$ and we have already noted that $\epsilon_{it}|u_i \sim N(u_i, 1)$. Thus let $z_{it} = \epsilon_{it} - u_i$ via a change of variables to get $z_{it} \sim N(0, 1)$ and our likelihood becomes:

$$L_i = \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \Phi(Up_{it} - u_i) - \Phi(Low_{it} - u_i) \right] \exp[-(u_i^2/(2\sigma_u^2))] (2\pi)^{-1/2} \sigma_u^{-1} du_i.$$

where $\Phi(\cdot)$ is the CDF of the $N(0, 1)$. Finally let $h_i = u_i/\sqrt{2}\sigma_u$,

$$L_i = \pi^{-1/2} \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \Phi(Up_{it} - \sqrt{2}\sigma_u h_i) - \Phi(Low_{it} - \sqrt{2}\sigma_u h_i) \right] \exp[-h_i^2] dh_i.$$

So that's clearly trivial (!) We can then use numerical techniques to maximize this likelihood.

- So if we have a Random Effects model, we pretty much have to do a Probit Model due to the complexity of the problem. If we have a Fixed Effects model (next), then we have to do a Logit.

A Logit Model in a Panel Data Framework - Fixed Effects

- Consider the model:

$$Prob(Y_{it} = y_{it}) = \frac{\exp(y_{it}[\alpha_i + x_{it}\beta])}{1 + \exp(\alpha_i + x_{it}\beta)}, \quad y_{it} = (0, 1); \quad t = 1, \dots, T_i, \quad i = 1, \dots, N.$$

Thus, as in the standard logit:

$$Prob(y_{it} = 0) = \frac{1}{1 + \exp(\alpha_i + x_{it}\beta)},$$

$$Prob(y_{it} = 1) = \frac{\exp(\alpha_i + x_{it}\beta)}{1 + \exp(\alpha_i + x_{it}\beta)}.$$

- Because of the α 's, we have too many parameters to estimate so again we'll have to do something special to our likelihood. Consider the case where T_i is small and N is large. Denote:

$$F_{it} = \frac{\exp(\alpha_i + x_{it}\beta)}{1 + \exp(\alpha_i + x_{it}\beta)}.$$

And note we can write:

$$Prob(Y_{it} = y_{it}) = F_{it}^{y_{it}} (1 - F_{it})^{1-y_{it}}.$$

Thus we can express the likelihood function as:

$$L = \prod_{i=1}^N \prod_{t=1}^{T_i} F_{it}^{y_{it}} (1 - F_{it})^{1-y_{it}}.$$

Note the number of parameters is $k + N \rightarrow \infty$ as $N \rightarrow \infty$. This is called the "Incidental Parameter Problem."

- So we need a way to eliminate the fixed effects! Let,

$$S_i = \sum_{t=1}^{T_i} y_{it}$$

be the sum of the realized "ones" in the i^{th} unit. Now consider the CONDITIONAL likelihood:

$$L_c = \prod_{i=1}^N L_{ci},$$

$$L_{ci} = Prob\left(Y_{i1} = y_{i1}, \dots, Y_{iT_i} = y_{iT_i} \mid \sum_{t=1}^{T_i} Y_{it} = S_i\right),$$

which we will show is equal to:

$$L_{ci} = \frac{\exp[\sum_{t=1}^{T_i} y_{it}x_{it}\beta]}{\prod_{\sum_t d_{it}=S_i} \exp(\sum_{t=1}^{T_i} d_{it}x_{it}\beta)}, d_{it} = 0, 1.$$

- So the product in the denominator is over all sequences d_{i1}, \dots, d_{iT_i} such that the sum is S_i . As long as T_i is small, this maximization is feasible.
- ILLUSTRATION. Suppose $T_i = 2$ so that $S_i = 0, 1, \text{ or } 2$. Then the possible probability

statements underlying L_{ci} are:

$$P(0, 0|0) = \text{Prob}(Y_{i1} = 0, Y_{i2} = 0 | \sum_{i=1}^2 Y_{it} = 0) = 1$$

$$P(1, 1|2) = \text{Prob}(Y_{i1} = 1, Y_{i2} = 1 | \sum_{i=1}^2 Y_{it} = 2) = 1$$

$$P(0, 1|1) = \text{Prob}(Y_{i1} = 0, Y_{i2} = 1 | \sum_{i=1}^2 Y_{it} = 1) = ?$$

$$P(1, 1|1) = \text{Prob}(Y_{i1} = 1, Y_{i2} = 0 | \sum_{i=1}^2 Y_{it} = 1) = ?$$

Note for $P(0, 0|0)$, then $y_{i1} = y_{i2} = S_i = d_{i1} = d_{i2} = 0$, so, our equation for L_{ci} reduces to:

$$\frac{\exp[\sum_{t=1}^{T_i}(0)x_{it}\beta]}{\prod_{\sum_t d_{it}=0} \exp(\sum_{t=1}^{T_i}(0)x_{it}\beta)} = \frac{\exp(0)}{\exp(0)} = 1.$$

- So finally, what about the 3rd and 4th probabilities above? Recall $F_{it} = \frac{\exp(\alpha_i + x_{it}\beta)}{1 + \exp(\alpha_i + x_{it}\beta)}$.

Consider the 3rd equation:

$$\begin{aligned}
&= P(0, 1|1) \\
&= \text{Prob}(Y_{i1} = 0, Y_{i2} = 1 | \sum_{i=1}^2 Y_{it} = 1) \\
&= \text{Prob}(Y_{i1} = 0, Y_{i2} = 1, \sum_{i=1}^2 Y_{it} = 1) / \text{Prob}(\sum_{i=1}^2 Y_{it} = 1) \\
&= \text{Prob}(Y_{i1} = 0, Y_{i2} = 1) / \text{Prob}(\sum_{i=1}^2 Y_{it} = 1) \\
&= \text{Prob}(Y_{i1} = 0) \text{Prob}(Y_{i2} = 1) / [\text{Prob}(Y_{i1} = 0) \text{Prob}(Y_{i2} = 1) + \text{Prob}(Y_{i1} = 1) \text{Prob}(Y_{i2} = 0)] \\
&= \frac{(1 - F_{i1}) F_{i2}}{(1 - F_{i1}) F_{i2} + F_{i1} (1 - F_{i2})} \\
&= \frac{(1 - \frac{\exp(\alpha_i + x_{i1}\beta)}{1 + \exp(\alpha_i + x_{i1}\beta)}) \frac{\exp(\alpha_i + x_{i2}\beta)}{1 + \exp(\alpha_i + x_{i2}\beta)}}{(1 - \frac{\exp(\alpha_i + x_{i1}\beta)}{1 + \exp(\alpha_i + x_{i1}\beta)}) \frac{\exp(\alpha_i + x_{i2}\beta)}{1 + \exp(\alpha_i + x_{i2}\beta)} + (\frac{\exp(\alpha_i + x_{i1}\beta)}{1 + \exp(\alpha_i + x_{i1}\beta)}) (1 - \frac{\exp(\alpha_i + x_{i2}\beta)}{1 + \exp(\alpha_i + x_{i2}\beta)})} \\
&= \frac{1}{1 + \exp(\alpha_i + x_{i1}\beta)} \frac{\exp(\alpha_i + x_{i2}\beta)}{1 + \exp(\alpha_i + x_{i2}\beta)} \\
&= \frac{1}{1 + \exp(\alpha_i + x_{i1}\beta)} \frac{\exp(\alpha_i + x_{i2}\beta)}{1 + \exp(\alpha_i + x_{i2}\beta)} + (\frac{\exp(\alpha_i + x_{i1}\beta)}{1 + \exp(\alpha_i + x_{i1}\beta)}) \frac{1}{1 + \exp(\alpha_i + x_{i2}\beta)} \\
&= \frac{\exp(\alpha_i + x_{i2}\beta)}{\exp(\alpha_i + x_{i2}\beta) + \exp(\alpha_i + x_{i1}\beta)} \\
&= \frac{\exp(x_{i2}\beta)}{\exp(x_{i2}\beta) + \exp(x_{i1}\beta)} \\
&=? \frac{\exp(x_{i2}\beta)}{\exp(x_{i1}\beta) + \exp(x_{i2}\beta)}
\end{aligned}$$

So the α_i 's cancel and we can estimate the model. This last equation is in the form of L_{ci} above.

12 Lecture 12: March 7, 2006

12.1 More Qualitative Models

A Dynamic Panel Data Probit Model with Random Effects

- Here we consider a model where the initial time period is essentially exogenously determined, and we have a lagged dependent in the model for all future periods. Consider:

$$Y_{i0} = 1 \text{ if } x_{i0}\beta + v_{i0} > 0,$$

and zero otherwise,

$$Y_{it} = 1 \text{ if } x_{it}\beta + \gamma y_{it-1} + z_i\delta + \eta_i + v_{it} > 0, \quad t = 1, \dots, T,$$

and zero otherwise.

- Assumptions on errors:

$$\eta_i \sim iid N(0, \sigma_\eta^2),$$

$$v_{i0} \sim iid N(0, \sigma_0^2),$$

$$v_{it} = \rho v_{it-1} + w_{it}, \quad t = 1, \dots, T; \quad |\rho| < 1; \quad w_{it} \sim iid N(0, \sigma_w^2),$$

over both i and t .

- Then normalize such that: $\sigma_\eta^2 + \sigma_w^2 = 1$ and $\sigma_0^2 = \frac{\sigma_w^2}{1 - \rho^2}$. This implies:

$$Var(v_{i1}) = \rho^2 Var(v_{i0}) + Var(w_{i1}) = \rho^2 \frac{\sigma_w^2}{1 - \rho^2} + \sigma_w^2 = \frac{\sigma_w^2}{1 - \rho^2} = Var(v_{i0}).$$

So v_{it} is stationary!

- So what's our likelihood look like in this model:

$$L = \prod_{i=1}^N L_i,$$

$$L_i = Pr(Y_{i0} = y_{i0}) * Pr(Y_{i1} = y_{i1} | Y_{i0} = y_{i0}) * \dots * Pr(Y_{iT} = y_{iT} | Y_{iT-1} = y_{iT-1}),$$

where the small y 's are the observed data (zeros and ones).

- As an illustration, suppose we observed $y_{i0} = 1$ and $y_{i1} = 0$, then:

$$\begin{aligned} Pr(Y_{i0} = 1) &= Pr(v_{i0} > -x_{i0}\beta) \\ Pr(Y_{i1} = 0 | Y_{i0} = 1) &= \frac{Pr(Y_{i0} = 1, Y_{i1} = 0)}{Pr(Y_{i0} = 1)} \\ &= \frac{Pr(v_{i0} > -x_{i0}\beta, \eta_i + v_{i1} < -x_{i1}\beta - \gamma - z_i\delta)}{Pr(v_{i0} > -x_{i0}\beta)} \end{aligned}$$

Note for the numerator of the second probability, we need the joint density (which is a bivariate normal).

- See notes for an example of a random effects probit where the initial value depends on the random effect. Not responsible for exam.

Polychotomous Dependent Variable Models

- Before our dependent variable only had two possible values, zero or one, yes or no, right or wrong, etc. Now lets relax this so that:

$$Y_t = 0, 1, \dots, M,$$

so there are $M + 1$ possible discrete categories.

- First we consider **Ordered Response Models** where we order the categories in terms of intensity. See G-12.1 and G-12.2. For the poly case, we have:

$$\begin{aligned} Pr(Y_t = M) &= F(X_t B) &= P_{Mt} \\ Pr(Y_t = M - 1) &= F(X_t B + \alpha_1) - F(X_t B) &= P_{M-1,t} \\ Pr(Y_t = M - 2) &= F(X_t B + \alpha_1 + \alpha_2) - F(X_t B + \alpha_1) &= P_{M-2,t} \\ &\vdots \\ Pr(Y_t = 0) &= 1 - F(X_t B + \alpha_1 + \alpha_2 + \dots + \alpha_{M-1}) &= P_{0,t} \end{aligned}$$

- Some examples.
 - (1) Purchase decisions on a car: $Y_t = 0$ if you spend less than 1000 dollars, $Y_t = 1$ if you spend between 1000 and 2000 dollars, $Y_t = 2$ if you spend between 2000 and 5000 dollars on a car, $Y_t = 3$ if you spend more than 5000 dollars on a car.
 - (2) Smoke, smoke a little, smoke a lot.
 - (3) Accidents: no injury, minor injuries, critical injuries, death.
 - (4) Olympics: no medal, bronze, silver, gold.
 - (5) Grad school: rejected, accepted with no help, accepted and on waitlist for help, accepted with assistance.

The key thing is that your categories should be both **Mutually Exclusive and Exhaustive!**

- Likelihood function:

$$\mathcal{L} = \prod_{t=1}^T P_{0t}^{C_{0t}} \dots P_{Mt}^{C_{Mt}},$$

where $C_{it} = 1$ if $Y_t = i$ and $C_{it} = 0$ otherwise.

- In some circumstances, we can reduce the number of parameters we need to estimate if the underlying variable is cardinal. In our car example, we can do this. We end up only estimating B and σ^2 .
- Next we consider **Unordered Response Models** where the categories are not related to intensity.
- Some examples.
 - (1) $Y_t = 1$ if you drive when taking a trip, $Y_t = 2$ if you travel by bus, $Y_t = 3$ if you travel by train, $Y_t = 4$ otherwise.
 - (2) Occupational choice.
 - (3) Corporate form - tax structure.
 - (4) Voting choices.

Suppose in general there are K categories.

- Denote:

$$P_{jt} = Pr(Y_t = j), \quad j = 1, \dots, K.$$

If we assume $P_{Kt} \neq 0 \forall t$, then we utilize the following mechanical method to generate our probabilities (which we plug into our likelihood).

$$\begin{aligned} \frac{P_{1t}}{P_{1t} + P_{Kt}} &= F(X_t B_1) = F_{1t} \\ \frac{P_{2t}}{P_{2t} + P_{Kt}} &= F(X_t B_2) = F_{2t} \\ &\vdots \\ \frac{P_{K-1,t}}{P_{K-1,t} + P_{Kt}} &= F(X_t B_{K-1}) = F_{K-1,t} \\ \sum_{j=1}^K P_{jt} &= 1 \end{aligned}$$

Note that our probabilities are all between zero and one.

- Consider some manipulations:

$$\begin{aligned} \frac{P_{jt}}{P_{jt} + P_{Kt}} &= F_{jt} \\ P_{jt} &= F_{jt}(P_{jt} + P_{Kt}) \\ \frac{P_{jt}}{P_{Kt}} &= F_{jt} \frac{P_{jt} + P_{Kt}}{P_{Kt}} \\ \frac{P_{jt}}{P_{Kt}} &= \frac{F_{jt}}{1 - F_{jt}} \equiv G_{jt} \end{aligned}$$

Also,

$$\begin{aligned}\sum_{j=1}^{K-1} \frac{P_{jt}}{P_{Kt}} &= \frac{1}{P_{Kt}} \sum_{j=1}^{K-1} P_{jt} \\ &= \frac{1}{P_{Kt}} (1 - P_{Kt}) \\ &= \frac{1}{P_{Kt}} - 1\end{aligned}$$

So:

$$\begin{aligned}\sum_{j=1}^{K-1} G_{jt} &= \sum_{j=1}^{K-1} \frac{P_{jt}}{P_{Kt}} \\ &= \frac{1}{P_{Kt}} - 1 \\ &\text{Now rearrange:} \\ P_{Kt} &= \frac{1}{1 + \sum_{j=1}^{K-1} G_{jt}}\end{aligned}$$

And since $P_{jt} = G_{jt}P_{Kt}$,

$$P_{jt} = G_{jt} \frac{1}{1 + \sum_{j=1}^{K-1} G_{jt}} = \frac{G_{jt}}{1 + \sum_{j=1}^{K-1} G_{jt}}, \quad j = 1, \dots, K - 1.$$

- So if we assume a logit model, with:

$$F(X_t B_j) = F_{jt} = \frac{e^{X_t B_j}}{1 + e^{X_t B_j}},$$

Then,

$$G_{jt} = \frac{F_{jt}}{1 - F_{jt}} = \frac{e^{X_t B_j} / (1 + e^{X_t B_j})}{1 / (1 + e^{X_t B_j})} = e^{X_t B_j}.$$

And thus,

$$P_{jt} = \frac{G_{jt}}{1 + \sum_{j=1}^{K-1} G_{jt}} = \frac{e^{X_t B_j}}{1 + \sum_{j=1}^{K-1} e^{X_t B_j}}, \quad j = 1, \dots, K - 1.$$

And,

$$P_{Kt} = 1 - P_{1t} - \dots - P_{K-1,t}.$$

- These are our probabilities which we would plug into our likelihood:

$$\mathcal{L} = \left(\prod_{t \in T_1} P_{1t} \right) \left(\prod_{t \in T_2} P_{2t} \right) \cdots \left(\prod_{t \in T_K} P_{Kt} \right).$$

- More next time on variations of the polychotomous model.

13 Lecture 13: March 9, 2006

13.1 More Qualitative Models

Variations of the Polychotomous Model

- Consider this more generalized version of the unordered model:

$$P_{jt} = \frac{e^{Z_{jt}\alpha + X_t B_j}}{\sum_{i=1}^M e^{Z_{it}\alpha + X_t B_i}},$$

where t is the individual and j is the category (of which there are M).

- So consider a few cases:
 - (a) If $\alpha = 0$, the model is called the multinomial model. This model would be useful in determining the probability a NEW person will select a given category.
 - (b) If $B_j = 0 \forall j$, the model is called the conditional logit. In this case we could have a variable number of categories per individual:

$$P_{jt} = \frac{e^{Z_{jt}\alpha}}{\sum_{i=1}^{M_t} e^{Z_{it}\alpha}}.$$

This model would be useful to predict if an existing person would choose a NEW category.

- (c) If $\alpha \neq 0$ and $B_j \neq 0$, the model is called the mixed model.
 - (d) Note that Z_{jt} cannot contain a constant term (it would cancel) but X_t may.
 - (f) If $M = 2$, $B_1 = 0$, $B_2 = B$, and $\alpha = 0$, the model is called the binary model.
 - (g) We can also derive the probability model above using a stochastic utility approach.
 - (j) In the model above, we normalize on the K^{th} or M^{th} category so everything is relative to this omitted category. Clearly we could easily normalize on ANY category and get the same results (though clearly the coefficients will be different).
- So how do we measure goodness of fit in a polychotomous model? Recall in a dichotomous model, we just used the square of the correlation coefficient between the observed y 's and our fitted y 's. However, now the y 's are just categories and the values are not bounded between zero and one. So we'll need a new measure.
 - In usual OLS, we often use the ratio of the likelihoods under the null and alternative as an estimate of R^2 . Consider our likelihood:

$$\mathcal{L} = \prod_{t=1}^T P_{1t}^{Y_{1t}} \dots P_{Kt}^{Y_{Kt}},$$

and compute:

$$R^2 = 1 - \left(\frac{\hat{\mathcal{L}}_0}{\hat{\mathcal{L}}_1} \right)^{2/T}.$$

The null is that all the X 's are unimportant.

- But there is a problem with this in that each P_{jt} is less than or equal to 1 so R^2 is strictly bounded below one. Maddala's suggested alternative:

$$\underline{R}^2 = \frac{1 - \left(\frac{\hat{\mathcal{L}}_0}{\hat{\mathcal{L}}_1} \right)^{2/T}}{1 - \hat{\mathcal{L}}_0^{2/T}}.$$

Now for some super sweet X 's, the R^2 can be one.

Multinomial Interpretation

- Recall the following from 623. Consider an experiment with K outcomes that is repeated N times. Let Y_i be the number of times outcome i occurs across all repetitions. So,

$$Y_1 + \dots + Y_K = N,$$

and the multinomial density is:

$$f_N(Y_1, \dots, Y_{K-1}) = \frac{N!}{Y_1! \dots Y_K!} P_1^{Y_1} \dots P_K^{Y_K}.$$

If $N = 1$,

$$f(Y_1, \dots, Y_{K-1}) = P_1^{Y_1} \dots P_K^{Y_K}.$$

Here $P(Y_i = 1) = P_i$. The marginal density of Y_1 is:

$$f_1(Y_1) = P_1^{Y_1} (1 - P_1)^{1-Y_1}, \quad Y_1 \in \{0, 1\}.$$

So clearly $P(Y_1 = 1) = P_1$ so the multinomial and "point normal" (Bernoulli) are consistent.

- Now consider the multinomial logit. In this model we specify probabilities:

$$P(Y_t = i) = P_{ti}, \quad i = 1, \dots, K.$$

so Y_t could be $1, 2, \dots, K$, for the K categories.

- The density of Y_t could be expressed as follows

$$f(Y_t) = P_{t1}^{I(Y_t=1)} \dots P_{tK}^{I(Y_t=K)},$$

where $I(Y_t = i) = 1$ if $Y_t = i$ and zero otherwise. Denote $I(Y_t = i) = Y_{ti}$. Thus $Y_{ti} \in \{0, 1\}$.

- Now consider the joint density of the Y_{ti} 's:

$$g(Y_{t1}, \dots, Y_{t,K-1}) = P_{t1}^{Y_{t1}} \dots P_{tK}^{Y_{tK}},$$

which is a multinomial with $N = 1$. And the marginal density of Y_{t1} is:

$$g_1(Y_{t1}) = P_{t1}^{Y_{t1}} (1 - P_{t1})^{1-Y_{t1}}.$$

- So this is all consistent with the multinomial formulation however we specify P_{t1} in the last two equations differently in practice. Unlike the 623 notes above, the estimates of P_{t1} would be different in the joint and marginal densities. See notes.

Red Bus / Blue Bus Problem

- Consider the probability model above which can be expressed as:

$$P_{jt} = \frac{e^{\bar{u}_{tj}}}{\sum_{i=1}^M e^{\bar{u}_{ti}}},$$

which implies:

$$\frac{P_{jt}}{P_{it}} = \frac{e^{\bar{u}_{tj}}}{e^{\bar{u}_{ti}}}$$

- So this says that when considering the relative probabilities that individual t chooses category i compared to j , the alternatives do NOT matter! "Independence of Irrelevant Alternatives."
- Lets use an example to show just how wrong this is. Consider an individual that can choose from the following three categories to get to work in the morning:

$Y = 1$ if by red bus

$Y = 2$ if by blue bus

$Y = 3$ if by car

- Suppose the individual is indifferent between the two buses and also indifferent between driving a car and taking a bus. Thus:

$$P(Y = 1|X_1, X_2) = P(Y = 2|X_1, X_2) = \frac{1}{2},$$

because taking a red bus given the option of both buses is just as good as taking the blue bus given the two options.

- Also,

$$P(Y = 1|X_1, X_3) = P(Y = 2|X_2, X_3) = \frac{1}{2},$$

because he's indifferent between buses and cars.

- However, given the option of all three, the individual would group the two buses together as one *meaningful* alternative so:

$$P(Y = 1|X_1, X_2, X_3) = P(Y = 2|X_1, X_2, X_3) = \frac{1}{4}.$$

Why? Because the person would group the red and blue bus together and the 1/2 choice of taking a bus would be split 50/50 between red and blue buses.

- So all this implies in terms of relative probabilities:

$$\frac{P(Y = 1|X_1, X_3)}{P(Y = 3|X_1, X_3)} = \frac{1/2}{1/2} = 1.$$

And:

$$\frac{P(Y = 1|X_1, X_2, X_3)}{P(Y = 3|X_1, X_2, X_3)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

So there is NOT an independence of irrelevant alternatives in this case!!

- Punchline: when choosing your categories, do not choose categories that are “too close” to each other that an individual will have trouble discerning one from another. Categories must be mutually exclusive as we’ve said.

14 Lecture 14: March 14, 2006

14.1 More Qualitative Models

A Note on the Omitted Category in the Unordered Model

- Suppose we normalize on category $j = K$:

$$P_{jt} = \frac{e^{x_t\beta_j}}{1 + \sum_{i=1}^{K-1} e^{x_t\beta_i}}, \quad j = 1, \dots, K, \quad \beta_K = 0$$

$$P_{Kt} = \frac{1}{1 + \sum_{i=1}^{K-1} e^{x_t\beta_i}}.$$

- But we could also normalize on a different category, say $j = 1$:

$$P_{jt} = \frac{e^{x_t\gamma_j}}{1 + \sum_{i=2}^K e^{x_t\gamma_i}}, \quad j = 2, \dots, K, \quad \gamma_1 = 0$$

$$P_{1t} = \frac{1}{1 + \sum_{i=2}^K e^{x_t\gamma_i}}.$$

- The relationship between the coefficients is thus:

$$\gamma_i = \beta_i - \beta_1.$$

Disturbance Terms in the Red Bus / Blue Bus Problem

- Recall we had a problem with specifying our categories too finely that an individual would be indifferent between them. Suppose the t^{th} person has utility for the i^{th} category of:

$$u_{ti} = X_t B_{ti} + Z_{ti} \gamma_{ti} = W_{ti} \alpha_{ti}.$$

- Now suppose $\alpha_{ti} = \alpha_i + \phi_t$, ie, there are two components of this coefficient. Suppose $E[\phi_t] = 0$, $E[\phi_t \phi'_s] = 0$ and $E[\phi_t \phi'_t] = V$. Then, we could write our model as:

$$u_{ti} = W_{ti} \alpha_{ti} = W_{ti} \alpha_i + W_{ti} \phi_t = W_{ti} \alpha_i + \epsilon_{ti}.$$

- Clearly $E[\epsilon_{ti}] = 0$ but,

$$E[\epsilon_{ti}^2] = W_{ti} V W'_{ti},$$

and,

$$E[\epsilon_{ti} \epsilon_{tj}] = W_{ti} V W'_{tj} \neq 0.$$

This is especially true when $W_{ti} = W_{tj}$, ie, the possibilities are similar. Thus the covariance is not zero so the errors are NOT iid across categories as we need.

A Probit Model with Covariance

- Recall the following.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \begin{pmatrix} 0, & \sigma_{11} & \sigma_{12} \\ 0 & \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Then:

$$\begin{pmatrix} Z_1 = X_1/\sqrt{\sigma_{11}} \\ Z_2 = X_2/\sqrt{\sigma_{22}} \end{pmatrix} \sim N \begin{pmatrix} 0, & 1 & \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} \\ 0 & \sigma_{12}/\sqrt{\sigma_{22}\sigma_{21}} & 1 \end{pmatrix} \equiv N \begin{pmatrix} 0, & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$$

which is the standard bivariate normal.

- So suppose there are individuals (indexed by t) and 3 categories to choose from ($i = 1, 2, 3$). Thus,

$$u_{ti} = \bar{u}_{ti} + \epsilon_{ti},$$

so we decompose the utility into its mean and deviation. Let $\epsilon_t = (\epsilon_{t1}, \epsilon_{t2}, \epsilon_{t3})'$ and assume:

$$\epsilon_t \sim N(0, \Omega), \quad \Omega = \{\sigma_{ij}\}.$$

- Define:

$$\bar{u}_{tij} = \bar{u}_{ti} - \bar{u}_{tj},$$

$$\bar{\epsilon}_{tij} = \bar{\epsilon}_{ti} - \bar{\epsilon}_{tj}.$$

- Thus the probability of selecting category 1 is:

$$\begin{aligned} P_{t1} &= Pr(u_{t1} > u_{t2} \text{ and } u_{t1} > u_{t3}) \\ &= Pr(u_{t1} - u_{t2} > 0 \text{ and } u_{t1} - u_{t3} > 0) \\ &= Pr(\bar{u}_{t1} + \epsilon_{t1} - \bar{u}_{t2} - \epsilon_{t2} > 0 \text{ and } \bar{u}_{t1} + \epsilon_{t1} - \bar{u}_{t3} - \epsilon_{t3} > 0) \\ &= Pr(\bar{u}_{t12} + \bar{\epsilon}_{t12} > 0 \text{ and } \bar{u}_{t13} + \bar{\epsilon}_{t13} > 0) \\ &= Pr(\bar{\epsilon}_{t12} > -\bar{u}_{t12} \text{ and } \bar{\epsilon}_{t13} > -\bar{u}_{t13}) \\ &= Pr(\bar{\epsilon}_{t21} < \bar{u}_{t21} \text{ and } \bar{\epsilon}_{t31} < \bar{u}_{t31}) \end{aligned}$$

which is a JOINT NORMAL!

- So given the parameters of the joint normal, (see notes), we could write:

$$P_{t1} = \int_{-\infty}^{\bar{u}_{t13}/\sqrt{\gamma_{22}}} \int_{-\infty}^{\bar{u}_{t12}/\sqrt{\gamma_{11}}} f_1(X_1, X_2) dX_1 dX_2.$$

$$P_{t2} = \int_{-\infty}^{\bar{u}_{t21}/\sqrt{\alpha_{11}}} \int_{-\infty}^{\bar{u}_{t23}/\sqrt{\alpha_{22}}} f_1(X_1, X_2) dX_1 dX_2.$$

- But then P_{t1}/P_{t2} depends on the characteristics of case 3!! So we don't have the independence of irrelevant alternatives in this model.

The Mixed Logit Model - (Random Parameter Logit)

- Say the utility that person n gets from alternative j in time t is:

$$u_{njt} = B'_n X_{njt} + \epsilon_{njt}, \quad n = 1, \dots, N, \quad t = 1, \dots, T, \quad j = 1, \dots, K.$$

- Assume X is non-stochastic and observable. B_n is a RANDOM parameter and we do NOT estimate it. Instead we estimate the distribution of B_n , ie we estimate θ in the pdf, $f(B_n|\theta)$.
- Assume ϵ_{njt} is iid extreme value and independent of B_n .
- Suppose that conditional on B_n ,

$$Pr(\text{Person } n \text{ chooses } i \text{ in period } t) = \frac{e^{X_{nit}B_n}}{\sum_j e^{X_{njt}B_n}} \equiv L_{nit}(B_n).$$

So this is our conditional likelihood.

- Integrate out B_n :

$$Q_{nit}(\theta) = \int L_{nit}(B_n) f(B_n|\theta) dB_n.$$

So this is our UNconditional likelihood.

- Now let i_{nt} be the alternative that person n actually chooses in period t . The probability of person n 's observed sequence is thus:

$$S_n(B_n) = \prod_{t=1}^T L_{ni_{nt}}(B_n).$$

- Again, integrate out B_n :

$$P_n(\theta) = \int S_n(B_n) f(B_n|\theta) dB_n.$$

Which yields log-likelihood:

$$\log(\mathcal{L}) = \sum_{n=1}^N \log(P_n(\theta)).$$

- To maximize the log-likelihood above is computationally intensive. A simulated approach is often the only way to do it. Essentially its a bootstrap method and is outlined in the notes. Under certain conditions, the simulated likelihood estimator we find will be consistent and asymptotically normal.

Non-Categorical Discrete Variables

- Suppose our categories are actually cardinal. Maybe Y_t is the number of magazine subscriptions a person has, or how many accidents they get into. In this case we might estimate the following Poisson model:

$$f(Y_t) = \frac{e^{-\lambda_t} \lambda_t^{Y_t}}{Y_t!}, \quad Y_t = 0, 1, \dots$$

$$\lambda_t = e^{B_0 + X_t B}.$$

- A very interesting result is that this Poisson distribution results whenever the Poisson Postulates are satisfied. Let $P(j, \Delta t)$ be the probability of j events in a period of length Δt . The Postulates are as follows:
 - (1) Events are independent.
 - (2) $P(1, \Delta t) = \lambda \Delta t + o(\Delta t)$ where $\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$ and Δt is small.
 - (3) $\sum_{j=2}^{\infty} P(j, \Delta t) = o(\Delta t)$.

Count Data Model with Endogeneity

- Assume we have a cardinal variable, Y_t , to be explained and suppose X is some vector of variables that we do observe. However, there are also variables, W , which we do NOT observe that explain Y and are correlated with X . Classic omitted variables problem.
- Example. Y_t is the number of cigarettes smoked, X includes income, age, location, etc, but W contains the number of smoking friends a person has (not observed).
- Assume:

$$E[Y|X, W, \alpha, \gamma] = e^{X\alpha} W\gamma,$$

where α and γ are constant vector and $E[W\gamma] = 1$ by a normalization. So this is our assumed model though I'm not sure why it's appropriate.

- Decomposing Y , we have:

$$Y = e^{X\alpha} W\gamma + \epsilon,$$

where ϵ is (conditional) mean zero.

- Add and subtract:

$$Y = e^{X\alpha} + [e^{X\alpha}][W\gamma - 1] + \epsilon,$$

which is our non-linear regression model.

- Assume (!) $E[W\gamma - 1|X] = g(X)$ and assume we have instruments, Z , such that:

$$E[\epsilon|X, Z] = 0, \quad E[W\gamma - 1|Z] = 0,$$

so our instruments are uncorrelated with W and ϵ .

- Thus, multiplying our model by the inverse of the exponential, we have:

$$Ye^{-X\alpha} = 1 + [W\gamma - 1] + e^{-X\alpha}\epsilon$$

$$Ye^{-X\alpha} - 1 = [W\gamma - 1] + e^{-X\alpha}\epsilon$$

Take expectations:

$$\begin{aligned} E[Ye^{-X\alpha} - 1|Z] &= E[W\gamma - 1 + e^{-X\alpha}\epsilon|Z] \\ &= E[e^{-X\alpha}\epsilon|Z] \\ &= E[E[e^{-X\alpha}\epsilon|X, Z]|Z] \\ &= E[e^{-X\alpha} \underbrace{E[\epsilon|X, Z]}_0|Z] \\ &= 0 \end{aligned}$$

Thus we can estimate α by GMM.

Sequential Models

- Now suppose we have a model which corresponds to a sequence of events. Suppose:

$Y_t = 1$ if t^{th} person did not finish high school

$Y_t = 2$ if t^{th} person finished HS but not college

$Y_t = 3$ if t^{th} person finished college but not grad

$Y_t = 4$ if t^{th} person finished graduate school

- Then our probabilities become:

$$P_{t1} = Pr(Y_t = 1) = F(X_t B_1) = F_{t1}$$

$$P_{t2} = Pr(Y_t = 2) = (1 - F_{t1})F(X_t B_2)$$

$$P_{t3} = Pr(Y_t = 3) = (1 - F_{t1})(1 - F_{t2})F(X_t B_3)$$

$$P_{t4} = Pr(Y_t = 4) = (1 - F_{t1})(1 - F_{t2})(1 - F_{t3})$$

Clearly $P_{t1} + P_{t2} + P_{t3} + P_{t4} = 1$.

- Our likelihood function for this model is thus:

$$\mathcal{L} = \left[\prod_{t \in T_1} F_{t1} \right] * \left[\prod_{t \in T_2} (1 - F_{t1})F_{t2} \right] * \left[\prod_{t \in T_3} (1 - F_{t1})(1 - F_{t2})F_{t3} \right] * \left[\prod_{t \in T_4} (1 - F_{t1})(1 - F_{t2})(1 - F_{t3}) \right].$$

Note that B_i only appears in F_{ti} so we can separate out our likelihood as follows:

$$\mathcal{L}_1(B_1) = \prod_{t \in T_1} F_{t1} \prod_{t \in T_2, T_3, T_4} (1 - F_{t1}),$$

$$\mathcal{L}_2(B_2) = \prod_{t \in T_2} F_{t2} \prod_{t \in T_3, T_4} (1 - F_{t2}),$$

$$\mathcal{L}_3(B_3) = \prod_{t \in T_3} F_{t3} \prod_{t \in T_4} (1 - F_{t3}).$$

Note, eg, \mathcal{L}_1 is a probit for finish high school versus did not. And we can then write:

$$Max \mathcal{L} = [Max_{B_1} \mathcal{L}_1] * [Max_{B_2} \mathcal{L}_2] * [Max_{B_3} \mathcal{L}_3].$$

That is, we can estimate a sequential model by a series of simple probit models!!

15 Lecture 15: March 16, 2006

15.1 More Qualitative Models

Generalized Sequential Model - Mobarak

- Consider a sequential probability model with students applying to graduate school.
- In stage 1, they apply and are either Accepted with Financial aid (AF), Accepted with No Financial aid (ANF), or are Rejected (R). So:

$$Pr(AF) = F(X_t B),$$

$$Pr(ANF) = F(X_t B + \alpha_1) - F(X_t B),$$

$$Pr(R) = 1 - F(X_t B + \alpha_1).$$

- In stage 2, they can Accept the Offer (AO) or Decline the Offer (DO). So:

$$Pr(DO) = G(X_t \gamma),$$

$$Pr(AO) = 1 - G(X_t \gamma).$$

- In stage 3, they either Fail their comps (F), or they succeed (S):

$$Pr(F) = H(X_t C),$$

$$Pr(S) = 1 - H(X_t C).$$

- Note F, G , and H are all CDFs which COULD BE THE SAME! We could make the model more general by allowing for X_t to be different in each stage.

Tobit Models (James Tobit)

- A Tobit model considers a dependent variable which is equal to zero with positive probability and otherwise is positive and continuous.
- Example might be the length of time a machine is broken down or the length of unemployment. Need:

$$Pr(Y_t = 0) \neq 0.$$

- Preliminary: If $f(X)$ is the pdf of X , then:

$$f(X|X \in A) = \frac{f(X)}{Pr(X \in A)}.$$

So if $X \sim N(0, \sigma^2)$,

$$f(X|X > -a) = \frac{e^{-0.5\sigma^{-2}X^2}}{\sqrt{2\pi\sigma^2}Pr(X > -a)}.$$

And clearly,

$$E[X|X > -a] = \int_{-a}^{\infty} X \frac{e^{-0.5\sigma^{-2}X^2}}{\sqrt{2\pi\sigma^2}Pr(X > -a)} dX.$$

And if $Z = e^{-0.5\sigma^{-2}X^2}$, by a change of variables:

$$\begin{aligned} E[X|X > -a] &= -\frac{1}{Pr(X > -a)} \int_{e^{-0.5a^2\sigma^{-2}}}^0 \sigma(2\pi)^{-1/2} dZ \\ &= \frac{-1}{Pr(X > -a)} (-\sigma(2\pi)^{-1/2} e^{-0.5\sigma^{-2}a^2}) \\ &= \sigma((2\pi)^{-1/2} e^{-0.5\sigma^{-2}a^2}) \frac{1}{Pr(X > -a)} \\ &= \sigma\phi_n(-a/\sigma) \frac{1}{1 - \Phi_n(-a/\sigma)} \end{aligned}$$

where ϕ_n and Φ_n are the standard normal pdf and cdf.

- One illustration (shown in the notes) is for home sales data. Either a person purchases a house (for some positive amount) or they do not (in which case $Y_t = 0$).
- Tobit models are estimated by ML. If we observe the values of X_t when Y_t is equal to zero, we have a censored model. If we do not observe X_t when Y_t is zero, we have a truncated model. The likelihood function is a bit different in each case. The two statistics often reported for Tobits are the conditional expectation, $E[Y_t|Y_t > 0]$, and the unconditional expectation, $E[Y_t]$.
- Its is important for a Tobit model to use the Normal CDF with mean zero but variance, σ^2 , not 1! Restricting the variance to be one is too restrictive.

Key Results on Conditional Means and Variances

- Suppose $X \sim N(0, 1)$. Then:

$$\begin{aligned}
 E[X|X \geq C_1] &= \frac{\phi(C_1)}{1 - \Phi(C_1)} = M_1 \\
 \text{Var}[X|X \geq C_1] &= 1 - M_1(M_1 - C_1) \\
 E[X|X \leq C_2] &= \frac{-\phi(C_2)}{\Phi(C_2)} = M_2 \\
 \text{Var}[X|X \leq C_2] &= 1 - M_2(M_2 - C_2) \\
 E[X|C_1 \leq X \leq C_2] &= \frac{\phi(C_1) - \phi(C_2)}{\Phi(C_2) - \Phi(C_1)} = M \\
 \text{Var}[X|C_1 \leq X \leq C_2] &= 1 + M^2 + \frac{C_1\phi(C_1) - C_2\phi(C_2)}{\Phi(C_2) - \Phi(C_1)}
 \end{aligned}$$

- A nice application. Suppose $X \sim N(\mu, \sigma^2)$. Then,

$$\begin{aligned}
 E[X|X > a] &= E[X - \mu + \mu|X - \mu + \mu > a] \\
 &= \sigma E\left[\frac{X - \mu}{\sigma} + \frac{\mu}{\sigma} \mid \frac{X - \mu}{\sigma} > \frac{a - \mu}{\sigma}\right] \\
 &= \sigma \frac{\mu}{\sigma} + \sigma E\left[\frac{X - \mu}{\sigma} \mid \frac{X - \mu}{\sigma} > \frac{a - \mu}{\sigma}\right] \\
 &= \mu + \sigma E[Z|Z > \frac{a - \mu}{\sigma}] \\
 &= \mu + \sigma E[Z|Z > C_1] \\
 &= \mu + \sigma \frac{\phi(C_1)}{1 - \Phi(C_1)}
 \end{aligned}$$

Switching Model

- The basic structure of a switching model is you have different regimes and depending on the levels of some unknown parameters of the model, you either observe one regime or another.
- Consider the following standard example. Let W_i^m be the market wage of an individual, W_i^r is their reservation wage, and W_i^o is the observed wage. If a person's reservation wage is smaller than or equal to the market wage, they work and we observe their

market wage. Otherwise, we may not observe the individual at all! So,

$$\begin{aligned} W_i^o &= W_i^m & \text{if } & W_i^m > W_i^r \\ W_i^o &= X_{1i}B_1 + u_{1i} & \text{if } & X_{1i}B_1 + u_{1i} > X_{2i}B_2 + u_{2i} \\ W_i^o &= X_{1i}B_1 + u_{1i} & \text{if } & u_{1i} - u_{2i} > X_{2i}B_2 - X_{1i}B_1 \\ W_i^o &= X_{1i}B_1 + u_{1i} & \text{if } & u_{3i} > Z_i\gamma \end{aligned}$$

- More next time.

16 Lecture 16: March 28, 2006

16.1 More Qualitative Models

Heckman's Sample Selection Model

- This is a model where you observe one regime depending on the outcome of another relationship. Consider:

$$Y_{1i} = X_{1i}B_1 + u_{1i},$$

$$Y_{2i} = X_{2i}B_2 + u_{2i},$$

where we observe Y_{1i} only if $Y_{2i} > 0$. Thus,

$$Y_{1i} = X_{1i}B_1 + u_{1i}, \text{ if } u_{2i} > -X_{2i}B_2.$$

- One example of this might be a model of earnings of workers who have migrated to a certain region. You wouldn't want to estimate that regression and then say that anyone who migrated would earn such and such a wage. The people who didn't migrate are different (via the second equation) than those that did.
- We could have more than one regime of course. A 2-regime model would be:

$$Y_i = X_{1i}B_1 + u_{1i}, \text{ if } u_{3i} \leq Z_i\gamma,$$

$$Y_i = X_{2i}B_2 + u_{2i}, \text{ if } u_{3i} > Z_i\gamma.$$

- How do we estimate model like these? We can either use ML or a 2-step method. The ML estimation can be difficult but the 2-step method is less efficient.
- One way to write the likelihood is:

$$\mathcal{L} = \prod_{i=1}^T \left[\int_{-\infty}^{Z_i\gamma} g(u_{1i}, u_{3i}) du_{3i} \right]^{I_i} * \left[\int_{Z_i\gamma}^{\infty} f(u_{2i}, u_{3i}) du_{3i} \right]^{1-I_i},$$

where $I_i = 1$ if $u_{3i} \leq Z_i\gamma$ and f and g are joint normal densities.

- Another way to write the likelihood is to group the regime one observations into T_1 and the regime two observations into T_2 :

$$\mathcal{L} = \prod_{i \in T_1} \left[\int_{-\infty}^{Z_i\gamma} g(u_{1i}, u_{3i}) du_{3i} \right] * \prod_{i \in T_2} \left[\int_{Z_i\gamma}^{\infty} f(u_{2i}, u_{3i}) du_{3i} \right].$$

- Thus, if λ is our parameter vector, large sample theory tells us:

$$\sqrt{T}(\hat{\lambda} - \lambda) \rightarrow^d N\left(0, \text{plim } T \left[-\frac{\partial \log \mathcal{L}}{\partial \lambda \partial \lambda'} \Big|_{\hat{\lambda}} \right]^{-1}\right).$$

- As a special case, we can assume that u_{1i} and u_{2i} are both independent of u_{3i} . This is a silly assumption though because the thing that makes you switch regimes should be related to the regime itself!?. But if you did, things simplify quite a bit and the likelihood becomes a product of a probit and two OLS maximizations. See HK notes.
- For the two step method, we use the formulas from a few lectures back to write:

$$E[u_{1i}|u_{3i} \leq Z_i\gamma] = -\sigma_{13} \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \equiv -\sigma_{13}W_{1i},$$

$$E[u_{2i}|u_{3i} > Z_i\gamma] = \sigma_{23} \frac{\phi(Z_i\gamma)}{1 - \Phi(Z_i\gamma)} \equiv \sigma_{23}W_{2i}.$$

- Note that W_{2i}^{-1} is sometimes called the “Mills Ratio.” Then write u_{1i} and u_{2i} as mean plus deviation from mean:

$$u_{1i} = -\sigma_{13}W_{1i} + \epsilon_{1i},$$

$$u_{2i} = \sigma_{23}W_{2i} + \epsilon_{2i},$$

but note that ϵ_{1i} and ϵ_{2i} are NOT iid. They are heteroskedastic.

- Thus we could rewrite the model as:

$$Y_i = X_{1i}B_1 - \sigma_{13}W_{1i} + \epsilon_{1i}, \text{ if } I_i = 1,$$

$$Y_i = X_{2i}B_2 + \sigma_{23}W_{2i} + \epsilon_{2i}, \text{ if } I_i = 0.$$

- Then our two-step method involves first estimating γ via a probit, plugging $\hat{\gamma}$ into our W ratios and plugging the \hat{W} 's into our rewritten model and estimating by OLS. This has some nice properties but isn't quite as efficient as MLE.
- There is also a section in the notes about situations when you don't observe your regime. If that's the case, you can't use the conditional densities, but must instead integrate out u_{3i} and form the unconditional density of Y_i . Then the likelihood is just the product of these unconditional densities.

One Regime Case

- One interesting side note is the case of one regime. Suppose:

$$Y_{1i} = X_{1i}B + \epsilon_i, \text{ if } u_i < Z_i\gamma.$$

Thus this IS an issue that we have to address since we don't have Y_{1i} data if $u_i > Z_i\gamma$. If we just had censoring on X_{1i} , we would be ok since that isn't related to the error term. We could still just do regular OLS to estimate consistently.

- But here we might do ML using the conditional density, $f(\epsilon_i|u_i < Z_i\gamma)$. Form the likelihood:

$$\mathcal{L} = \prod_i f(\epsilon_i|u_i < Z_i\gamma) = \prod_i \left[\int_{-\infty}^{Z_i\gamma} \frac{g(\epsilon_i, u_i)}{\Pr(u_i < Z_i\gamma)} du_i \right].$$

Randomized Data Models

- The final (!!) section in the qualitative notes that we will cover deals with situations where people don't always answer questions honestly and how we can estimate consistently in the face of these issues.
- Some questions might be: do you take drugs, do you cheat on your taxes, do you cheat on your wife, etc.
- Here's a fun example. Suppose an urn contains R red balls, B blue balls, and W white balls. Then a person selects a ball at random and reports:

$Y_i = 1$ if they chose a red ball

$Y_i = 0$ if they chose a blue ball

$Y_i = 1$ if they chose a white ball and they take drugs

$Y_i = 0$ if they chose a white ball and they do not take drugs

- So clearly,

$$P_R = \frac{R}{R + B + W},$$

$$P_B = \frac{B}{R + B + W},$$

$$P_W = 1 - P_R - P_B.$$

- Also, if X_i are the attributes of individual i :

$$\Pr(i^{th} \text{ person takes drugs}) = \Pi(X_i B) = \Pi_i = \frac{e^{X_i B}}{1 + e^{X_i B}},$$

ie, a logit model.

- So what do we have:

$$\begin{aligned}
Pr(Y_i = 1) &= Pr(R \text{ or } (W \text{ and Drug Taker})) \\
&= P_R + Pr(W \text{ and Drug Taker}) \\
&= P_R + Pr(W|Drug Taker) * Pr(Drug Taker) \\
&= P_R + P_W * Pr(Drug Taker) \\
&= P_R + (1 - P_R - P_B) * \Pi_i \\
&\text{and similarly:} \\
Pr(Y_i = 0) &= P_B + (1 - P_R - P_B) * (1 - \Pi_i)
\end{aligned}$$

- So if we let T_0 be the set of points where $Y_i = 0$ and T_1 is the set where $Y_i = 1$, then we could maximize our likelihood:

$$\mathcal{L} = \prod_{i \in T_0} \left[P_B + (1 - P_R - P_B) * (1 - \Pi_i) \right] * \prod_{i \in T_1} \left[P_R + (1 - P_R - P_B) * \Pi_i \right].$$

Or,

$$\mathcal{L} = \prod_{i \in T_0} \left[P_B + (1 - P_R - P_B) \frac{1}{1 + e^{X_i B}} \right] * \prod_{i \in T_1} \left[P_R + (1 - P_R - P_B) \frac{e^{X_i B}}{1 + e^{X_i B}} \right],$$

which we could maximize wrt B .

- So the idea is that we have given the individual a way to disguise the fact that he takes drugs. Observing a $Y_i = 1$ could mean that the person takes drugs OR that he has chosen a white ball. Hopefully this will induce truthful reporting and we can then estimate the B consistently.

17 Lecture 17: March 30, 2006

17.1 Section 11: Rational Expectations

- **Definition** A Rational Expectation is an expectation that is consistent with the corresponding distribution implied by the model. If Y_t is a variable and our information set is Λ_t so Y has conditional density, $f(Y|\Lambda_t)$, then,

$$E[Y_t|\Lambda_t] = \int_{-\infty}^{\infty} Y f(Y|\Lambda_t) dY.$$

- Notation:

$$E[X_t|\Lambda_{t-1}] = X_{t,t-1}.$$

Example 1: Simple Exogenous Type Model

- Consider the model:

$$Y_t = \alpha X_{t,t-1} + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma^2),$$
$$X_{t,t-1} = w_{t-1}\gamma.$$

- We can write the X term as mean plus deviation from mean as usual:

$$X_t = X_{t,t-1} + \eta_t = w_{t-1}\gamma + \eta_t,$$

where η_t is conditional mean zero.

- Thus the “empirical form of the model” is:

$$Y_t = \alpha w_{t-1}\gamma + \epsilon_t,$$

$$X_t = w_{t-1}\gamma + \eta_t.$$

which is clearly nonlinear in the parameters.

- Note that η_t may be heteroskedastic but will not be autocorrelated.

Example 2

- Consider the model:

$$Y_t = \alpha Y_{t,t-1} + X_t b + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2),$$

$$X_t = \rho_1 X_{t-1} + \dots + \rho_q X_{t-q} + \phi_t, \quad \phi_t \sim iid(0, \sigma_\phi^2).$$

- So our information set is:

$$\Lambda_t = (Y_t, Y_{t-1}, \dots, X_t, X_{t-1}, \dots).$$

- Then taking expectations of the Y_t equation,

$$E[Y_t|\Lambda_{t-1}] = \alpha E[Y_{t,t-1}|\Lambda_{t-1}] + E[X_t b|\Lambda_{t-1}] + E[\epsilon_t|\Lambda_{t-1}]$$

$$Y_{t,t-1} = \alpha Y_{t,t-1} + X_{t,t-1} b.$$

So:

$$Y_{t,t-1} = \frac{1}{1-\alpha} X_{t,t-1} b.$$

- Now denote:

$$\bar{X}_{t-1} = (X_{t-1}, \dots, X_{t-q}),$$

$$\rho' = (\rho_1, \dots, \rho_q).$$

So:

$$X_{t,t-1} = E[X_t|\Lambda_{t-1}] = E[\rho_1 X_{t-1} + \dots + \rho_q X_{t-q} + \phi_t|\Lambda_{t-1}] = \bar{X}_{t-1} \rho.$$

- Then we can rewrite the model (in empirical form):

$$\begin{aligned} Y_t &= \alpha Y_{t,t-1} + X_t b + \epsilon_t \\ &= \alpha \frac{1}{1-\alpha} X_{t,t-1} b + X_t b + \epsilon_t \\ &= \frac{\alpha}{1-\alpha} \bar{X}_{t-1} \rho b + X_t b + \epsilon_t \\ X_t &= \rho_1 X_{t-1} + \dots + \rho_q X_{t-q} + \phi_t \\ &= \bar{X}_{t-1} \rho + \phi_t \end{aligned}$$

And again there are nonlinearities, but we don't have heteroskedasticity or autocorrelation.

Example 3: A model with auto, hetero, and a systems problem

- Consider the model:

$$Y_t = \alpha Y_{t+1,t} + X_t b + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2), \quad |\alpha| < 1,$$

$$X_t = \bar{X}_{t-1} \rho + \phi_t, \quad \phi_t \sim iid(0, \sigma_\phi^2).$$

- Now update the Y_t equation one period:

$$Y_{t+1} = \alpha Y_{t+2,t+1} + X_{t+1} b + \epsilon_{t+1}.$$

And take expectations conditional on Λ_t :

$$Y_{t+1,t} = \alpha E[Y_{t+2,t+1}|\Lambda_t] + X_{t+1,t} b.$$

And what's this crazy second term?

$$\begin{aligned}
 E[Y_{t+2,t+1}|\Lambda_t] &= E[E(Y_{t+2}|\Lambda_{t+1})|\Lambda_t] \\
 &= E[E(Y_{t+2}|\Lambda_t, (\Lambda_{t+1} - \Lambda_t))|\Lambda_t] \\
 &\quad \text{because you never forget anything!} \\
 &= E[Y_{t+2}|\Lambda_t] \\
 &= Y_{t+2,t}
 \end{aligned}$$

So we have:

$$Y_{t+1,t} = \alpha Y_{t+2,t} + X_{t+1,t}b.$$

- We could do a similar analysis to show that:

$$Y_{t+j,t} = \alpha Y_{t+j+1,t} + X_{t+j,t}b, \quad j \geq 1.$$

And by repeated substitution we could write:

$$Y_{t+1,t} = f(\bar{X}_t|\rho, \alpha, b),$$

which is VERY nonlinear in the parameters. However it does NOT depend on Y . We only know this by solving out the model (ie, doing repeated substitutions or solving a difference equation).

- So the empirical form of our model becomes:

$$\begin{aligned}
 Y_t &= \alpha Y_{t+1,t} + X_t b + \epsilon_t \\
 Y_t &= \alpha f(\bar{X}_t|\rho, \alpha, b) + X_t b + \epsilon_t \\
 X_t &= \bar{X}_{t-1}\rho + \phi_t.
 \end{aligned}$$

- The problem with this is estimating. The nonlinearities are huge and we might have trouble getting the thing to converge. An alternative is available which people still do but involves hetero, auto, and systems problems.
- This model is due to McCullagh (?) and it starts by expressing Y_{t+1} as mean plus deviation:

$$Y_{t+1} = Y_{t+1,t} + \eta_{t+1},$$

so,

$$Y_{t+1,t} = Y_{t+1} - \eta_{t+1}.$$

- Replace this in our model to get:

$$\begin{aligned}
 Y_t &= \alpha Y_{t+1,t} + X_t b + \epsilon_t \\
 Y_t &= \alpha(Y_{t+1} - \eta_{t+1}) + X_t b + \epsilon_t
 \end{aligned}$$

$$Y_t = \alpha Y_{t+1} + X_t b + v_t, \quad v_t = \epsilon_t - \alpha \eta_{t+1}.$$

- This model would feature auto, hetero, and a systems problem. To see the systems problem, note:

$$E[Y_{t+1}v_t] = E[(Y_{t+1,t} + \eta_{t+1})(\eta_t - \alpha\eta_{t+1})] \neq 0.$$

And similarly for the autocorrelation problem:

$$E[v_t v_{t-1}] = E[(\epsilon_t - \alpha\eta_{t+1})(\epsilon_{t-1} - \alpha\eta_t)] \neq 0.$$

- So to estimate the model, you would need to account for all these issues. But it is still done.

Estimating Rational Expectations Models

- We now show that many Rational Expectations (RE) models can be estimated by NLLS, MLE, 2SLS, 3SLS, and GMM.
- Recall the method of moments (MM). This is when you have exactly the same number of parameters as you have moment conditions. Moments may be polynomial functions like $g_i(x) = x^i$. We just take the sample counterparts and choose parameters to match the population moments.
- OLS is equivalent to the method of moments. If $Y = XB + \epsilon$, then premultiplying everything by X' ,

$$X'Y = X'XB + X'\epsilon.$$

So,

$$E[X'Y] = E[X'X]B + E[X'\epsilon].$$

$$E[X'Y] = X'XB.$$

$$B = (X'X)^{-1}E[X'Y],$$

And we replace this last term by the same moment, $X'Y$, which yields:

$$B = (X'X)^{-1}X'Y = B_{ols}!$$

- GLS is also the MM. See HK notes.
- MLE is also the MM. See HK notes.
- Generalized Method of Moments (GMM). Often times, we have more moments than parameters, in which case we do GMM. In this situation, we just minimize some function of the difference between our sample and population moments, with appropriate weighting. If h is a $rx1$ vector of population moments and g is a $rx1$ vector of sample moments and θ is a $Kx1$ vector of parameters of interest (with $r > K$), consider:

$$d = g - h.$$

- We could minimize:

$$F(d) = d'd,$$

but we could do better minimizing:

$$F(d) = d'Ad,$$

where A is some positive definite matrix. Which positive definite matrix to choose? The logical choice is the inverse of the variance/covariance matrix of d . Hence, we minimize:

$$F(d) = d'\Omega_d^{-1}d.$$

- Then, suppose $\hat{\theta}$ minimizes this last equation and let:

$$\hat{d} = g - h(\hat{\theta}).$$

Then we have a nice result which says that under H_0 (ie, the model is true):

$$\hat{F} = \hat{d}'\Omega_d^{-1}\hat{d} \rightarrow^d \chi_{r-K}^2.$$

- So if we choose our moments, form the $F(d)$ quadratic form, minimize it, we can then test to see if we're really explaining the model with a simple χ^2 test. Brilliant.

18 Lecture 18: April 4, 2006

18.1 More on Rational Expectations

The Hal White VC Adjustment

- Consider the model,

$$Y_t = X_t B + \epsilon_t, \quad t = 1, \dots, T.$$

Or in matrix form:

$$Y = XB + \epsilon.$$

- Assume:

- (A1) X is nonstochastic and uniformly bounded.
- (A2) $T^{-1}X'X \rightarrow Q_x \neq 0$.
- (A3) $\epsilon_t = g_t v_t$ with g_t nonrandom but unknown.
- (A4) g_t bounded away from zero and infinity and let $G = \text{diag}_{t=1}^T(g_t)$; $G^2 = \text{diag}_{t=1}^T(g_t^2)$; such that:

$$T^{-1}X'G^2X \rightarrow h \neq 0,$$

where h is finite.

- (A5) $v_t \sim iid(0, 1)$, with finite fourth moment.
- Given all these assumptions, we have the following results:
 - (1) B_{OLS} is consistent and $VC(B_{OLS}) = (X'X)^{-1}X'G^2X(X'X)^{-1}$.
 - (2) B_{OLS} is asymptotically normal:

$$\sqrt{T}(B_{OLS} - B) \rightarrow^d N(0, Q_x^{-1}hQ_x^{-1}).$$

- So we would like to get a consistent estimator for: $Q_x^{-1}hQ_x^{-1} = \Omega$. One such estimator is:

$$\hat{\Omega} = T(X'X)^{-1}X' \text{diag}(\hat{\epsilon}_t^2) X(X'X)^{-1},$$

where $\hat{\epsilon}_t = Y_t - X_t B_{OLS}$, the OLS residuals. This is our Hal White estimator.

- Proof that Hal White is consistent. Note $B_{OLS} = B + B_{OLS} - B \equiv B + \Delta$, where $\Delta \rightarrow^p 0$. Thus

$$\hat{\epsilon}_t = Y_t - X_t B_{OLS} = Y_t - X_t [B + \Delta] = \epsilon_t - X_t \Delta.$$

Thus,

$$\begin{aligned}
\hat{\Omega} &= [T(X'X)^{-1}][T^{-1}X' \text{diag}(\hat{\epsilon}_t^2) X][T(X'X)^{-1}] \\
&= [T(X'X)^{-1}][T^{-1} \sum X_t^2 \hat{\epsilon}_t^2][T(X'X)^{-1}] \\
&= [T(X'X)^{-1}][T^{-1} \sum X_t^2 (\epsilon_t - X_t \Delta)^2][T(X'X)^{-1}] \\
&= [T(X'X)^{-1}][T^{-1} \sum X_t^2 (\epsilon_t^2 - 2\epsilon_t X_t \Delta + X_t^2 \Delta^2)][T(X'X)^{-1}] \\
&= [T(X'X)^{-1}][T^{-1} \sum X_t^2 \epsilon_t^2 - 2\epsilon_t X_t^3 \Delta + X_t^4 \Delta^2][T(X'X)^{-1}] \\
\hat{\Omega} &\rightarrow [Q_x^{-1}] \underbrace{[plim T^{-1} \sum X_t^2 \epsilon_t^2 - 0 + 0]}_{\delta} [Q_x^{-1}]
\end{aligned}$$

But note that:

$$E[\delta] = T^{-1} \sum X_t^2 E[\epsilon_t^2] = T^{-1} \sum X_t^2 g_t^2 = T^{-1} X' G^2 X \rightarrow h,$$

and:

$$Var[\delta] = T^{-2} \sum X_t^4 [E(\epsilon_t^4) - (E(\epsilon_t^2))^2] \rightarrow 0.$$

Thus,

$$\hat{\Omega} \rightarrow Q_x^{-1} h Q_x^{-1}.$$

QED.

- Thus, our small sample guidance becomes:

$$B_{OLS} \approx N(B, (X'X)^{-1} X' \text{diag}(\hat{\epsilon}_t^2) X (X'X)^{-1}).$$

GMM in Detail

- Consider a system of M possibly nonlinear equations and a sample size of T :

$$q_{ti}(B_i^0) = u_{ti}, \quad t = 1, \dots, T, \quad i = 1, \dots, M.$$

Stacking the system, we have:

$$q_i(B_i^0) = u_i.$$

- Suppose B_i^0 is $K_i x 1$ for the i^{th} equation, and the total number of distinct parameters is:

$$K = K_1 + \dots + K_M.$$

- Let A be an instrument matrix of order Txr . We will assume (strongly) that:

$$E[A'u_i] = 0 \quad \forall i = 1, \dots, M.$$

These are our $r * M$ moment conditions. Need $r * M \geq K$.

- Premultiply our model by $T^{-1}A'$ yields:

$$T^{-1}(I_M \otimes A')q(B^0) = \underbrace{T^{-1}(I_M \otimes A')u(B^0)}_{g_T(B^0)}.$$

Also let:

$$g_T(B) = T^{-1}(I_M \otimes A')u(B),$$

for any given parameter vector, B .

- Note we can also write:

$$g_T(B^0) = T^{-1} \sum_{t=1}^T (u'_t \otimes a'_t).$$

So our moment condition is:

$$E[g_T(B^0)] = 0.$$

- Then the GMM estimator is as follows:

$$\hat{B}_{GMM} = \arg \min_B \{g_T(B)'C_T g_T(B)\},$$

where C_T is some positive semidefinite matrix. Of course it will be the inverse of the VC matrix of orthgonality conditions.

- Now, to get the large sample properties of our estimator, we need some (strong) assumptions. Some of these are:

- (A1) $g_T(B) \xrightarrow{p} g_0(B)$ uniformly where g_0 is C^1 .
- (A2) $g_0(B) = 0$ is uniquely defined at $B = B^0$.
- (A5) The matrix of first partials of g_T with respect to B exists, is continuous, and has full column rank for B in the neighborhood of B^0 .

- Given these assumptions (and more: see HK notes), we have:

$$\sqrt{T}(\hat{B}_{GMM} - B^0) \rightarrow^d N(0, V_0),$$

with:

$$V_0 = (G'_0 C_0 G_0)^{-1} G'_0 C_0 \Lambda_0 C_0 G_0 (G'_0 C_0 G_0)^{-1},$$

where $G_0 = G_0(B^0)$ and $\Lambda_0 = \Lambda_0(B_0)$.

- **Remark** If $C_T \rightarrow C_0 = \Lambda^{-1}$, then:

$$V_0 = (G'_0 C_0 G_0)^{-1},$$

which is very nice. This also implies that \hat{B}_{GMM} is efficient (in its class).

- **Remark** Note that $\Lambda_0(B^0)$ is the VC matrix of the orthogonality conditions. So if $g_{*t}^0 = u_t' \otimes a_t'$, then,

$$g_T(B^0) = T^{-1} \sum_t g_{*t}^0,$$

and:

$$VC[\sqrt{T}g_T(B^0)] = T^{-1}E[(\sum g_{*t}^0)(\sum g_{*t}^0)'] \rightarrow \Lambda_0(B^0).$$

- **Remark** Small sample guidance:

$$\hat{B}_{GMM} \approx N(B^0, T^{-1}\hat{V}_0),$$

$$\hat{V}_0 = \left[\left(\frac{\partial g_T(B)}{\partial B} \right)'_{\hat{B}_{GMM}} \hat{C}_T \left(\frac{\partial g_T(B)}{\partial B} \right)_{\hat{B}_{GMM}} \right]^{-1}.$$

- So what is this \hat{C}_T term? See notes on page 19 of HK for the Newey-West approach of weighting the different moments where more weight is placed on those that are closer together. Essentially,

$$\hat{C}_T^{-1} = \hat{V}C_T = T^{-1} \left[\sum_{t=1}^T g_{*t}(\hat{B}_{GMM}) \right] \left[\sum_{t=1}^T g_{*t}(\hat{B}_{GMM}) \right]'$$

- **Remark** Note that \hat{C}_T depends on \hat{B}_{GMM} and in turn, the GMM estimator depends on \hat{C}_T . Thus, an iterative approach is recommended. First estimate B with any consistent method (like IV), use this to estimate \hat{C}_T , re-estimate \hat{B}_{GMM} , and repeat until convergence.

Makeup Lecture Notes: April 16, 2006

Finishing Rational Expectations

More on GMM

- **Remark** If $E[u'_t u_s] = 0$ for all $t \neq s$, so that there is NO autocorrelation in the disturbance terms, then:

$$\hat{C}_T^{-1} = T^{-1} \sum_{t=1}^T \hat{g}_t \hat{g}'_t.$$

- **Remark** If there is no autocorrelation OR heteroskedasticity in the disturbances, then:

$$\hat{C}_T^{-1} = \hat{\Sigma}_u \otimes T^{-1} \sum_{t=1}^T a'_t a_t.$$

with,

$$\hat{\Sigma}_u = T^{-1} \sum_{t=1}^T \hat{u}'_t \hat{u}_t.$$

An Interpretation of the VC Formula for the GMM Estimator

- Suppose:

$$Y = XB^0 + \epsilon, \quad V_\epsilon = \Omega.$$

Then:

$$Tg_T(B) = A'(Y - XB).$$

- The objective function is:

$$\begin{aligned} Q &= (Y - XB)' A C_T A' (Y - XB) \\ &= Y' A C_T A' Y - 2Y' A C_T A' X B + B' X' A C_T A' X B \\ \frac{\partial Q}{\partial B} = 0 &\rightarrow \hat{B} = (X' A C_T A' X)^{-1} X' A C_T A' Y \\ VC(\hat{B}) &= (X' A C_T A' X)^{-1} X' A C_T A' \Omega A C_T A' X (X' A C_T A' X)^{-1} \end{aligned}$$

- In the notation above, we have, due the linearity of the model:

$$\frac{\partial g_T(B^0)}{\partial B} = A' X T^{-1} \rightarrow G_0(B^0),$$

and,

$$T^{-1} E[g_T(B^0) g_T(B^0)'] \rightarrow \Lambda_0.$$

- Note that OLS, IV, and 3SLS are all special cases of GMM. See page 23 of the HK notes. See page 25 for an application of GMM to a 2SLS problem with autocorrelation and/or hetero. Review this before final.

A Test of the Model

- Also called a test of an over-identifying restriction. The GMM estimator is defined as:

$$\text{Min}_B \{Q = g_T'(B)C_T g_T(B)\}, \quad C_T^{-1} \rightarrow \Lambda_0.$$

- Suppose B is $K \times 1$ and $g_T(B)$ is $rM \times 1$. Then under H_0 (the model is correct):

$$\hat{Q} = g_T'(\hat{B})C_T g_T(\hat{B}) \rightarrow^d \chi_{rM-K}^2.$$

- If the model is not correct, we expect $E[g_T(B^0)] \neq 0$ so that there is a violation of the orthogonality conditions. Thus the test:

Reject H_0 if: $\hat{Q} > \chi_{rM-K}^2(0.95)$ at the 5% level.

19 Lecture 19: April 18, 2006

19.1 Spatial Econometrics (Section 13)

Background and Introduction

- Some examples of economic models where space is important:
 - (1) Gas tax issues: when DC sets their gas tax, they must consider the taxes set in VA and MD.
 - (2) Police expenditures: knowing the crime rates and expenditures in neighboring communities will influence your own rates and needed funding.
 - (3) Infrastructure productivity: roads transverse various states.
 - (4) Volatility of GDP: if there is a jump in German GDP, France might also experience some effects due to spillovers, trade, etc.
 - (5) Welfare benefits: in order to not attract the lowest income earners to your state, you may want to have welfare benefits just below your neighbors.
- In general, anytime we have spill-overs, externalities, geographic proximity issues, etc, we have spatial interaction and models like the ones we will present become essential.
- Example of geographic neighbors:

<i>NN</i>	<i>NN</i>	<i>NN</i>	<i>NN</i>	<i>NN</i>
<i>NN</i>	<i>N</i>	<i>N</i>	<i>N</i>	<i>NN</i>
<i>NN</i>	<i>N</i>	<i>i</i>	<i>N</i>	<i>NN</i>
<i>NN</i>	<i>N</i>	<i>N</i>	<i>N</i>	<i>NN</i>
<i>NN</i>	<i>NN</i>	<i>NN</i>	<i>NN</i>	<i>NN</i>

The pattern with just the close neighbors (*N*'s) is called a Queen and with the *NN*'s is called a Double Queen.

- **Definition** Neighboring Units: units that interact in a meaningful way.
- **Definition** Weighting Matrix: a matrix that selects neighbors. Suppose $Y = (Y_1, \dots, Y_5)$ are observations on the GDP of the UK, France, Germany, China and Russia. W is a 5×5 weighting matrix that chooses each countries "neighbors." The weights for the neighbors will be non-zero and those that are not neighbors will be zero. Suppose X is a vector of unemployment rates in each of the 5 countries, then WX will be in our regression and W might look like:

$$W = \begin{bmatrix} 0 & w_{12} & w_{13} & 0 & 0 \\ w_{21} & 0 & w_{13} & 0 & 0 \\ w_{31} & w_{32} & 0 & w_{14} & 0 \\ 0 & 0 & w_{43} & 0 & w_{45} \\ 0 & 0 & 0 & w_{54} & 0 \end{bmatrix}.$$

So the UK interacts with France and Germany, China interacts with Germany and Russia, etc. Note the w_{ii} terms are ALWAYS zero! We include the individual effects separately. So the sum of the rows might be normalized to one though not always and sometimes we'll show that it's not appropriate to normalize.

- Simple example:

$$Y_i = b_0 + b_1 X_i + b_2 W_i X + \epsilon_i,$$

where $X = (X_1, \dots, X_n)$. So X_i is the within unit effect. Note we can also write the model:

$$Y_i = b_0 + b_1 X_i + b_2 \sum_{j=1}^n w_{ij} X_j + \epsilon_i,$$

or,

$$Y_i = b_0 + b_1 X_i + b_2 \bar{X}_i + \epsilon_i,$$

where \bar{X}_i is called the **Spatial Lag of X** .

- So how do we specify the weighting matrix.
 - (1) Row normalized. If Y_i has 5 neighbors, set all non-zero weights in row i equal to $\frac{1}{5}$. \hat{b}_2 becomes the average effect of your neighbors.
 - (2) Distance measures. Let d_{ij} be the distance between i and j . So let $w_{ij} = 1/d_{ij}$ which implies the weight goes to zero as the distance increases. We can also normalize the distance weight as:

$$w_{ij} = \frac{1/d_{ij}}{\sum_{r=1}^n (1/d_{ir})}.$$

- (3) Economic distance. If Q_j is the income per capita on cross-sectional unit j , then form weights as:

$$w_{ij} = |Q_i - Q_j|^{-1}.$$

So if our incomes are similar, the weights are larger. Other Q 's might be "average level of education," "ethnic group composition," or "trade shares between countries."

- (4) Euclidean distance. Suppose we actually had several (r) different variables that we thought were important for our weights. We could form:

$$w_{ij} = \frac{1}{1 + [(z_{i1} - z_{j1})^2 + \dots + (z_{ir} - z_{jr})^2]^{1/2}}.$$

Note that if you normalize here, be sure to "scale-normalize" because you don't want one variable that has large units to wipe out another that has small units.

Cliff and Ord Models

- Consider the following spatial model:

$$Y_i = a + \underbrace{X_i}_{1 \times K} \underbrace{B_1}_{K \times 1} + \rho_1 \left(\underbrace{W_i}_{1 \times N} \underbrace{Y}_{N \times 1} \right) + \left(\underbrace{W_i}_{N \times K} \underbrace{X}_{N \times K} \right) B_2 + u_i,$$

$$u_i = \rho_2 W_i u + \epsilon_i, \quad |\rho_1| < 1, \quad |\rho_2| < 1, \quad \epsilon_i \sim iid(0, \sigma_\epsilon^2).$$

So note we have a spatial lag of the dependent variable and the independent variables. We use the SAME weighting matrix in both case and also in our residuals which we say are “Spatial AR(1)” or just SAR(1).

- Write the model in matrix form:

$$Y = ae_n + XB_1 + \rho_1 WY + WX B_2 + u, \quad u = \rho_2 W u + \epsilon.$$

- Solve the model:

$$Y = (I - \rho_1 W)^{-1} [ae_n + XB_1 + WX B_2 + u], \quad u = (I - \rho_2 W)^{-1} \epsilon.$$

And we assume both of these inverses exist. Note because the inverses depend on N , the vectors Y and u are really triangular arrays. See notes.

- **Remark** If W is row-normalized then $(I - aW)$ is singular at $a = 1$.
- **Remark** If W is row-normalized, then $(I - aW)^{-1}$ exists for all $|a| < 1$. See proof in HK notes by Gershgorin.

20 Lecture 20: April 20, 2006

20.1 More Spatial Econometrics

More Cliff/Ord

- Consider a weighting matrix, W , with $w_{ii} = 0$ for all i . Let:

$$r = \text{Max}_i \sum_j |w_{ij}|, \quad c = \text{Max}_j \sum_i |w_{ij}|,$$

so r and c are the largest row and column sums respectively. Let:

$$\alpha = \min(r, c).$$

Then, assuming that the elements of W are nonnegative, $(I - aW)$ will be nonsingular for all:

$$|a| < \frac{1}{\alpha}.$$

This result is forthcoming in the JoE and it says that instead of always row normalizing your weighting matrix (even when the economics says you shouldn't do so), you can use the parameter space defined by this condition instead.

- Consider the following model:

$$Y = XB + \rho_1 WY + \epsilon = XB + \rho_1 \alpha \frac{W}{\alpha} Y + \epsilon.$$

Then if $\rho_1^* = \rho_1 \alpha$ and $W^* = W/\alpha$, we have:

$$Y = XB + \rho_1^* W^* Y + \epsilon.$$

Note α is defined above as the minimum of the maximal row and column sums. Then, using the above result, if we estimate the starred model, with ρ_1^* as the parameter, the inverse, $(I - aW)^{-1}$, will exist for all $|\rho_1^*| < 1$. We could then form:

$$|\hat{\rho}_1| = \frac{\hat{\rho}_1^*}{\alpha}.$$

Estimation of Cliff/Ord Models

- Consider the following model:

$$Y = XB_1 + \rho_1 WY + WXB_2 + u, \quad u = \rho_2 Wu + \epsilon, \quad |\rho_1| < 1, \quad |\rho_2| < 1.$$

- If $\rho_1 = \rho_2 = 0$, the model just has spatial lags on X and the disturbances are good ones. We can just do OLS.

- What if $\rho_1 = 0$, $\rho_2 \neq 0$, and $|\rho_2| < 1$. Our model becomes:

$$Y = ZB + u, \quad u = \rho_2 W u + \epsilon.$$

- Properties of u . Since $\epsilon \sim (0, \sigma_\epsilon^2 I)$,

$$u \sim (0, \sigma_\epsilon^2 (I - \rho_2 W)^{-1} (I - \rho_2 W')^{-1}) \equiv (0, \sigma_\epsilon^2 \Omega_u).$$

So the elements of u are heteroskedastic and spatially correlated.

- One estimation idea is GLS:

$$\hat{B}_{GLS} = (Z' \Omega_u^{-1} Z)^{-1} Z' \Omega_u^{-1} Y,$$

but this is NOT feasible since ρ_2 is generally not known. We'll turn to ML and a Cochrane-Orcutt procedure next.

Aside on Elasticities

- Consider the following model:

$$Y_i = X_{1i} b_1 + X_{2i} b_2 + \lambda \sum_{j=1}^N w_{ij} Y_j + \epsilon_i.$$

Or,

$$Y = Xb + \lambda W Y + \epsilon.$$

Solving the model:

$$Y = (I - \lambda W)^{-1} [Xb + \epsilon].$$

Thus, $E[Y] = (I - \lambda W)^{-1} Xb$, or,

$$E[Y_j] = (I - \lambda W)_j^{-1} Xb = G_j \cdot Xb = \sum_{i=1}^n G_{ji} [X_{1i} b_1 + X_{2i} b_2].$$

- So we have for $j = 2, \dots, N$,

$$\frac{\partial E[Y_j]}{\partial X_{11}} = G_{j1} b_1.$$

Thus, due to spillovers, the average Y in ALL countries is effected by the X_1 in country 1. We could then calculate:

$$\eta_{j1} = G_{j1} b_1 \frac{X_{11}}{Y_j},$$

as our elasticity.

- Clearly if $\lambda = 0$, these spillovers disappear.

Estimating with Maximum Likelihood

- So we're still in the case where there is no spatial lag on Y , there is one on X , and the disturbances are spatially correlated.
- Since typically, $u \sim N(0, \sigma_\epsilon^2 \Omega_u)$,

$$Y \sim N(ZB, \sigma_\epsilon^2 \Omega_u).$$

- Now we'll write our likelihood, take logs, and simplify:

$$\begin{aligned} \mathcal{L} &= \frac{(\sigma_\epsilon^2)^{-N/2} (2\pi)^{-N/2} \exp\left(-0.5\sigma_\epsilon^{-2}[Y - ZB]'\Omega_u^{-1}[Y - ZB]\right)}{|I - \rho_2 W|_+^{-1}} \\ \ln \mathcal{L} &= \underbrace{-\frac{N}{2}\ln(\sigma_\epsilon^2) - \frac{N}{2}\ln(2\pi)}_{\psi} - 0.5\sigma_\epsilon^{-2}[Y - ZB]'\Omega_u^{-1}[Y - ZB] + \ln|I - \rho_2 W|_+ \\ &= \psi - 0.5\sigma_\epsilon^{-2}[Y - ZB]'\Omega_u^{-1}[Y - ZB] + \ln|I - \rho_2 W|_+ \\ &= \psi - 0.5\sigma_\epsilon^{-2}[Y - ZB]'[(I - \rho_2 W)^{-1}(I - \rho_2 W')^{-1}]^{-1}[Y - ZB] + \ln|I - \rho_2 W|_+ \\ &= \psi - 0.5\sigma_\epsilon^{-2}[Y - ZB]'(I - \rho_2 W)(I - \rho_2 W')[Y - ZB] + \ln|I - \rho_2 W|_+ \\ &= \psi - 0.5\sigma_\epsilon^{-2}[(I - \rho_2 W)(Y - ZB)]'[(I - \rho_2 W)(Y - ZB)] + \ln|I - \rho_2 W|_+ \\ &= \psi - 0.5\sigma_\epsilon^{-2}[Y^*(\rho_2) - Z^*(\rho_2)B]'[Y^*(\rho_2) - Z^*(\rho_2)B] + \ln|I - \rho_2 W|_+ \end{aligned}$$

So how did we get that last line? Note:

$$(I - \rho_2 W)(Y - ZB) = (I - \rho_2 W)Y - (I - \rho_2 W)ZB = Y^* - Z^*B,$$

which is our Spatial Cochrane Orcutt Transformation.

- So is this a possible maximization? The only troublesome term is the $\ln|I - \rho_2 W|_+$ term. This is the log of the absolute value of the determinant of an $N \times N$ matrix! It's a huge mess because it depends on ρ_2 which we don't know! Often times, for reasonably sized N , we won't be able to do maximum likelihood.

Estimating with Feasible GLS - Using Cochrane Orcutt

- So where are we heading? Recall our model:

$$Y = ZB + u, \quad u = \rho_2 W u + \epsilon,$$

so we still don't have a spatial lag on Y . We'll get a consistent estimate of B , use it to obtain some \hat{u} 's, then use \hat{u} to get a $\hat{\rho}_2$, then use $\hat{\rho}_2$ to do a spatial C-O procedure, and then reestimate B via OLS (which overall is a FGLS procedure).

- Some preliminaries.

- (1) An $N \times N$ matrix, A , is absolutely summable if, FOR ALL $N \geq 1$, the row and column sums are all finite.
- (2) If A and B , both $N \times N$, are absolutely summable, then $D = AB$ is too.

Proof:

$$\begin{aligned}
 d_{ij} &= \sum_{r=1}^N a_{ir} b_{rj} \\
 \sum_{j=1}^N |d_{ij}| &\leq \sum_{j=1}^N \sum_{r=1}^N |a_{ir}| |b_{rj}| \\
 &= \sum_{r=1}^N \sum_{j=1}^N |a_{ir}| |b_{rj}| \\
 &= \sum_{r=1}^N |a_{ir}| \sum_{j=1}^N |b_{rj}| \\
 &\leq c_a c_b
 \end{aligned}$$

where c_a and c_b are the respective maximal column sums of A and B . A similar analysis shows that the columns of D will have finite sums. Thus D is absolutely summable. QED.

- (3) If A is absolutely summable, its elements are bounded.
- (4) If A is absolutely summable, and $Z_{N \times K}$ has bounded elements, then the elements of $Z'AZ$ are at most of order N . See notes.

- Some assumptions:

- (1) $\epsilon_i \sim iid(0, \sigma_\epsilon^2)$, finite fourth moment.
- (2) $|\rho_2| < 1$.
- (3) $P = (I - \rho_2 W)$ is nonsingular at the true value of ρ_2 .
- (4) $w_{ii} = 0 \forall i$.
- (5) W and P^{-1} are absolutely summable.
- (6) Z is nonstochastic, bounded, and full column rank (K).
- (7) $\lim N^{-1} Z'Z = Q_z$, nonsingular.
- (8) $\lim N^{-1} Z'\Omega_u Z = Q_1$, nonsingular.
- (9) $\lim N^{-1} Z'\Omega_u^{-1} Z = Q_2$, nonsingular.

- Note that assumption (5) means that your weighting matrix must not have elements that do NOT trail off to zero at some point. As you add neighbors, say, these new observations must not ALL be significant for everyone already in the model. For instance, if you had a central unit in your model where all things were neighbors to

it, your W matrix may not be absolutely summable because the weights for everyone would not trail off to zero at some point.

- More next time.

21 Lecture 21: April 25, 2006

21.1 Spatial Econometrics - Cliff/Ord

Feasible GLS

- Recall our model:

$$Y = ZB + u, \quad u = \rho_2 W u + \epsilon,$$

so we don't have a spatial lag on Y , only on X (included in the ZB term).

- Basic Results:

$$(R1) \quad VC(u) = \sigma_\epsilon^2 \Omega_u, \quad \Omega_u = (I - \rho_2 W)^{-1} (I - \rho_2 W')^{-1}.$$

$$(R2) \quad \hat{B} = (Z'Z)^{-1} Z'Y \text{ is consistent!}$$

So we can just do OLS on our model to get a consistent (and unbiased) estimator for B .

- Note that we can write $\hat{B} = B + (Z'Z)^{-1} Z'u$, so $E[\hat{B}] = B$ and:

$$VC(\hat{B}) = \sigma_\epsilon^2 (Z'Z)^{-1} Z' \Omega_u Z (Z'Z)^{-1} \rightarrow 0,$$

so by Chebychev, \hat{B} is consistent.

- Consider our OLS residuals:

$$\hat{u} = Y - Z\hat{B} = Y - ZB - Z\hat{B} + ZB = u + Z(B - \hat{B}) = u + Z\Delta_N \rightarrow u.$$

So we now assume that we know the disturbances, u , and we will now find a Generalized Moments Estimator (GME) for ρ_2 .

- Consider our disturbances:

$$u = \rho_2 W u + \epsilon,$$

or,

$$u - \rho_2 W u = \epsilon.$$

Thus,

$$W u - \rho_2 W^2 u = W \epsilon.$$

- Denote $\bar{u} = W u$, $\bar{\bar{u}} = W^2 u$, and $\bar{\epsilon} = W \epsilon$. Then consider again the two equations:

$$u - \rho_2 W u = \epsilon. \quad (1)$$

$$\bar{u} - \rho_2 \bar{\bar{u}} = \bar{\epsilon}. \quad (2)$$

- We consider now three algebraic exercises on (1) and (2):

- (1) Square (1), sum, and divide by N :

$$N^{-1} \sum u_i^2 + N^{-1} \rho_2^2 \sum \bar{u}_i^2 - 2N^{-1} \rho_2 \sum u_i \bar{u}_i = N^{-1} \sum \epsilon_i^2. \quad (3)$$

- (2) Square (2), sum, and divide by N :

$$N^{-1} \sum \bar{u}_i^2 + N^{-1} \rho_2^2 \sum \bar{\bar{u}}_i^2 - 2N^{-1} \rho_2 \sum \bar{u}_i \bar{\bar{u}}_i = N^{-1} \sum \bar{\epsilon}_i^2. \quad (4)$$

- (3) Multiply (1) by (2), sum, and divide by N :

$$N^{-1} \sum u_i \bar{u}_i + N^{-1} \rho_2^2 \sum \bar{u}_i \bar{\bar{u}}_i - N^{-1} \rho_2 [\sum \bar{u}_i \bar{\bar{u}}_i + \sum \bar{u}_i^2] = N^{-1} \sum \epsilon_i \bar{\epsilon}_i. \quad (5)$$

- Note the RHS of (3):

$$N^{-1} \sum \epsilon_i^2 \rightarrow^p \sigma_\epsilon^2,$$

by Chebychev.

- Note the RHS of (4):

$$N^{-1} \sum \bar{\epsilon}_i^2 = N^{-1} \epsilon' W' W \epsilon \rightarrow^p \sigma_\epsilon^2 \lim N^{-1} Tr(W' W),$$

by the usual trace and expected value rules.

- Note the RHS of (5):

$$N^{-1} \sum \epsilon_i \bar{\epsilon}_i = N^{-1} \epsilon' W \epsilon \rightarrow^p 0.$$

- So we can write the RHS terms as:

$$N^{-1} \sum \epsilon_i^2 = \sigma_\epsilon^2 + \delta_1,$$

$$N^{-1} \sum \bar{\epsilon}_i^2 = \sigma_\epsilon^2 N^{-1} Tr(W' W) + \delta_2,$$

$$N^{-1} \sum \epsilon_i \bar{\epsilon}_i = 0 + \delta_3,$$

where all the δ terms plim to zero.

- Ignoring the fact that we know one of the unknown parameters is ρ_2^2 , let $r = \rho_2^2$ and denote our parameter vector:

$$\lambda' = [r, \rho_2, \sigma_\epsilon^2].$$

- Thus write the system of 3 equations and three unknowns in matrix form as:

$$\begin{bmatrix} N^{-1} \sum \bar{u}_i^2 & -2N^{-1} \sum u_i \bar{u}_i & -1 \\ N^{-1} \sum \bar{\bar{u}}_i^2 & -2N^{-1} \sum \bar{u}_i \bar{\bar{u}}_i & -N^{-1} Tr(W' W) \\ N^{-1} \sum \bar{u}_i \bar{\bar{u}}_i & -N^{-1} [\sum u_i \bar{\bar{u}}_i + \sum \bar{u}_i^2] & 0 \end{bmatrix} \begin{bmatrix} r \\ \rho_2 \\ \sigma_\epsilon^2 \end{bmatrix} = \begin{bmatrix} -N^{-1} \sum u_i^2 \\ -N^{-1} \sum \bar{u}_i^2 \\ -N^{-1} \sum u_i \bar{u}_i \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

Or more compactly:

$$A_1\lambda = A_2 + \delta.$$

- Thus,

$$(\text{plim } A_1)\lambda = \text{plim } A_2 + \text{plim } \delta = \text{plim } A_2.$$

- Thus if u were observed, a consistent estimator of λ would be the (over parameterized) OLS estimator:

$$\hat{\lambda} = A_1^{-1}A_2.$$

It's over parameterized because we don't recognize that $r = \rho_2^2$.

- A better approach than OLS would be to recognize that $r = \rho_2^2$ and do NLLS, ie:

$$\text{Min}_{\rho_2, \sigma_\epsilon^2} [A_1\lambda - A_2]'[A_1\lambda - A_2].$$

- But through all of this, we have assumed we know u , so this clearly is NOT feasible. To make it FGLS, we replace u by the OLS residuals \hat{u} . Or, better yet, do NLLS above to get an estimator for λ , which includes the ρ_2 term which we are really interested in. We can then show that:

$$\hat{\lambda} = \hat{A}_1^{-1}\hat{A}_2 \rightarrow^p \lambda.$$

See page 16 - 18 of the HK notes for the consistency proof.

- **Remark** How could we estimate in TSP? Given some \hat{u} 's, form the \hat{A} matrices, and use the LSQ command on:

$$\hat{A}_2 = \hat{A}_1\lambda - \delta.$$

- So overall, what do we have. Given a consistent estimator for ρ_2 , let:

$$\hat{\Omega}_u = (I - \hat{\rho}_2 W)^{-1}(I - \hat{\rho}_2 W')^{-1}.$$

Then our True and Feasible estimators are:

$$\hat{B}_{GLS} = (Z'\Omega_u^{-1}Z)^{-1}Z'\Omega_u^{-1}Y,$$

$$\hat{B}_{FGLS} = (Z'\hat{\Omega}_u^{-1}Z)^{-1}Z'\hat{\Omega}_u^{-1}Y.$$

- Asymptotics:

$$\sqrt{N}(\hat{B}_{FGLS} - B) \rightarrow^d N(0, \sigma_\epsilon^2 \text{plim } N(Z'\Omega_u^{-1}Z)^{-1}),$$

$$\sqrt{N}(\hat{B}_{FGLS} - \hat{B}_{GLS}) \rightarrow^p 0,$$

which is nice. True and Feasible GLS estimators are asymptotically equivalent (see notes for a simple proof). To show the asymptotic distribution of the FGLS estimator, you need a CLT for triangular arrays which is not so nice.

- An IV/2SLS method to estimate this type of model is also outlined in the HK notes but it is not consistent and should be avoided.

Spatial Lags on Y

- So now we finally relax the assumption that we only have a spatial lag on the X 's. Consider the model:

$$Y = XB_1 + \rho_1 WY + WXB_2 + u, \quad u = \rho_2 Wu + \epsilon, \quad |\rho_1| < 1, \quad |\rho_2| < 1.$$

If $\rho_1 \neq 0$ and $\rho_2 \neq 0$, write the model:

$$Y = ZB + \rho_1 WY + u, \quad u = \rho_2 Wu + \epsilon.$$

- The WY term is now endogenous, so we'll need instruments. Possible instruments include Z, WZ, W^2Z, W^3Z , etc.
- More next time.

22 Lecture 22: April 27, 2006

22.1 Spatial Econometrics - Cliff/Ord

More on Spatial Lags on Y

- Recall our model:

$$Y = ZB + \rho_1 WY + u, \quad u = \rho_2 Wu + \epsilon.$$

If W is row-normalized then $(I - \rho_1 W)^{-1}$ exists for $|\rho_1| < 1$. If it is NOT row-normalized, then do the transformation above deflating by α .

- Solve for Y :

$$Y = (I - \rho_1 W)^{-1}(ZB + u).$$

So,

$$\begin{aligned} E[WY] &= W(I - \rho_1 W)^{-1}ZB \\ &= W(I + \rho_1 W + \rho_1^2 W^2 + \dots)ZB \\ &= WZB + \rho_1 W^2 ZB + \rho_1^2 W^3 ZB + \dots \end{aligned}$$

So since we need instruments for WY , clearly the best ones would be: Z, WZ, W^2Z, W^3Z , etc. We usually can stop after the W^2Z term. The gains from using the others are small.

- But we cannot just do 2SLS using the instruments just listed. The u 's are still bad disturbances. Thus we need to do a Spatial Cochrane Orcutt procedure (spatial C-O).
- Consider premultiplying by $(I - \rho_2 W)$:

$$\begin{aligned} Y &= ZB + \rho_1 WY + u \\ (I - \rho_2 W)Y &= (I - \rho_2 W)ZB + (I - \rho_2 W)\rho_1 WY + \underbrace{(I - \rho_2 W)u}_{\epsilon} \\ Y - \rho_2 WY &= (Z - \rho_2 WZ)B + \rho_1 W(Y - \rho_2 WY) + \epsilon \\ Y^*(\rho_2) &= Z^*(\rho_2)B + \rho_1 WY^*(\rho_2) + \epsilon \end{aligned}$$

- So do 2SLS of Y^* on Z^* and WY^* using the instruments above for WY^* . See notes for some confusing notation on the top of page 24. The usual asymptotic distribution results.

Weighting Matrix Selection

- Suppose we had two weighting matrices that we thought might be correct. We could write the model:

$$Y = XB + \lambda_1 W_1 Y + \lambda_2 W_2 Y + \epsilon,$$

and test:

$$H_0 : \lambda_2 = 0, \text{ vs } H_1 : \lambda_2 \neq 0.$$

We still need to do 2SLS but we might use instruments: $X, W_1 X, W_1^2 X, W_2 X,$ and $W_2^2 X$.

- What about having parameters in a weighting matrix that need to be estimated? Suppose:

$$w_{ij} = \gamma_0 d_{1ij}^{-\gamma_1} \cdots d_{rij}^{-\gamma_r},$$

so we have r measures of distance between two cities (say geographic, income, housing ownership, unemployment, etc), and we need to estimate the gammas. There are NO formal parametric results on this. We'll turn soon to non-parametric estimation.

- Case (1): What if our model was:

$$Y = XB + u, \quad u = \rho_2 W u + \epsilon, \quad W = W(\gamma).$$

So we have a parameterized weighting matrix but the problem is that γ is not identified if $\rho_2 = 0$! One thing that is often done is to assume that $\rho_2 \neq 0$ and then do OLS to get \hat{B} , and \hat{u} . Then do a GM procedure to estimate ρ_2 and γ . The GM procedure would involve 6 equations instead of the 3 we had last time because of the additional parameter we need to estimate. See notes.

- Case (2): What if our model was:

$$Y = XB_1 + WXB_2 + u, \quad u = \rho_2 W u + \epsilon, \quad W = W(\gamma).$$

Here since the W also hits the X , we would need to do ML. If N is large, this won't work because we would have to invert a huge matrix. So avoid this if your dataset is large.

Other Spatial Autocorrelation Models

- Consider a $SAR(q)$:

$$u = r_1 W_1 u + r_2 W_2 u + \cdots + r_q W_q u + \epsilon.$$

So W_1 might be the matrix that selects the nearest neighbors, W_2 selects the next nearest neighbors, etc. See G-22.1.

- There are also $SARMA(p, q)$ models which are not very common.

Moran I Test for Spatial Correlation

- Use this test ONLY for models with NO spatial lag on Y .
- Consider the model:

$$Y = ZB + u, \quad u = \rho_2 W u + \epsilon.$$

- We would like to test if $\rho_2 = 0$ or not.
- Consider the following statistic:

$$I = \frac{N \hat{u}' W \hat{u}}{s \hat{u}' \hat{u}},$$

with $s = \sum_i \sum_j w_{ij}$, $\hat{u} = Y - Z\hat{B}$, and $\hat{B} = (Z'Z)^{-1}Z'Y$.

- Then, given $W_z = I - R_z = I - Z(Z'Z)^{-1}Z'$, under $H_0 : u \sim N(0, \sigma^2 I)$,

$$E[I] = \frac{-N \text{Tr}(R_z W)}{s(N - K)},$$

$$\sigma_I^2 = \frac{N^2[\text{Tr}(W_z W W_z W') + \text{Tr}(W_z W)^2 + [\text{Tr}(W_z W)]^2]}{s^2(N - K)(N - K - 2)} - [E(I)]^2.$$

- So our test becomes, under $H_0 : \rho_2 = 0$,

$$\eta = \frac{I - E[I]}{\sigma_I} \rightarrow^d N(0, 1).$$

So reject the null in favor of spatial correlation at the five percent level if $|\eta| > 1.96$. Note I , $E[I]$, and σ_I do NOT depend on any unknown parameters.

- **Theorem:** If $\epsilon_i \sim iid N(0, 1)$ and $h(\epsilon_1, \dots, \epsilon_N)$ is a scale free ($homo(0)$) function, then h and $Q = \sum_{i=1}^N \epsilon_i^2$ are independent. See notes for proof.

The Kelejian/Prucha Generalization of the Moran I Test for Limited Dependent Variable Models

- Consider the (possibly) nonlinear model:

$$y_i = f(x_i, \beta) + \epsilon_i, \quad i = 1, \dots, N, \quad \epsilon_i \sim (0, \sigma_i^2).$$

So our errors are possibly heteroskedastic with $\sigma_i^2 = h(x_i, \beta)$.

- We would like to test the hypothesis:

$H_0 : \epsilon_i$ are independently distributed

$H_1 : \epsilon_i$ are spatially correlated

- Suppose we are considering a weighting matrix, W , which is absolutely summable.
- Form:

$$I = \frac{\hat{\epsilon}'W\hat{\epsilon}}{\hat{\sigma}_Q},$$

with:

$$\hat{\sigma}_Q^2 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2 \hat{\sigma}_i^2 \hat{\sigma}_j^2,$$

$$\hat{\sigma}_i^2 = h(x_i, \hat{\beta}).$$

Then under H_0 ,

$$I \rightarrow^d N(0, 1).$$

- So in the notes, we have examples of the KP test applied to a tobit model, a dichotomous dependent variable model, a sample selection model, and a polychotomous model. We'll consider the second and see the notes for the rest.

The Kelejian/Prucha Test Applied to a Dichotomous Dependent Variable Model

- Suppose:

$$y_i^* = x_i\beta + \eta_i, \quad i = 1, \dots, N,$$

$$y_i = 1 \text{ if } y_i^* \geq 0,$$

and zero else.

- Under H_0 , η_i is iid with zero mean and CDF, F .
- So write y_i as mean plus deviation:

$$y_i = F(x_i\beta) + \epsilon_i, \quad \epsilon_i \sim (0, \sigma_i^2),$$

where,

$$\sigma_i^2 = F(x_i\beta)[1 - F(x_i\beta)].$$

- Then let $\hat{\beta}$ be the MLE of β and form:

$$\hat{\sigma}_i^2 = F(x_i\hat{\beta})[1 - F(x_i\hat{\beta})].$$

$$\hat{\epsilon}_i = y_i - F(x_i\hat{\beta}).$$

- Then given our weighting matrix under consideration, calculate:

$$I = \frac{\hat{\epsilon}'W\hat{\epsilon}}{\hat{\sigma}_Q} \sim N(0, 1),$$

and reject iid disturbances in favor of spatially correlated errors if $|I| > 1.96$.

- Next we'll move to non-parametric estimation.

23 Lecture 23: May 2, 2006

23.1 Spatial Econometrics - Non-Parametric Estimation

HAC Estimation

- Here we consider the nonparametric estimation of a Heteroskedastic Autocorrelation estimator, or HAC estimator.

- Consider the model:

$$y_n = X\beta + \lambda W_n y_n + u_n = Z_n \gamma + u_n, \quad u_n = \rho W_n u_n + \epsilon_n.$$

- So the y_n equation is from economic theory but formulating u_n as a $SAR(1)$ process is unfounded.

- If we assume $\epsilon_n \sim (0, \Sigma_n)$, then,

$$u_n \sim (0, \Sigma_n) \equiv (0, \sigma^2 (I_n - \rho W_n)^{-1} (I_n - \rho W_n')^{-1}),$$

and u_n generally suffers from auto and hetero.

- To estimate the model, we do 2SLS using IV matrix, H_n and the second stage regressor matrix is:

$$\hat{Z} = H_n (H_n H_n)^{-1} H_n' Z_n.$$

- Thus,

$$n^{1/2}(\hat{\gamma} - \gamma) = \underbrace{(n^{-1} \hat{Z}' \hat{Z})^{-1} (n^{-1} Z_n' H_n)}_{\rightarrow^{pM}} \underbrace{(n^{-1} H_n H_n)^{-1} n^{-1/2} H_n' u_n}_{\rightarrow^{dN(0, \Omega)}}.$$

Where,

$$\Omega = \lim [n^{-1} H_n' \Sigma_n H_n].$$

- So we want to estimate Ω and if we assume a $SAR(1)$, we're fine. But what if we don't specify a specific form for the errors?
- We'll show our HAC estimator of Ω is consistent and positive semi-definite. We also will allow for multiple distance measures and measurement error for our weighting matrix.

Basic Model of a HAC Estimator

- Consider the model:

$$y_n = Z_n \gamma + u_n, \quad u_n = R_n \epsilon_n, \quad \epsilon_n \sim (0, I_n).$$

Thus,

$$VC(u_n) = R_n R_n' = \Sigma_n.$$

We assume R_n is nonstochastic but unknown and is absolutely summable. A subset of this type of model is the $SAR(1)$, as well as more complicated $SARMA(p, q)$ models.

- So, as above, the IV estimator will involve a term in the VC matrix like:

$$\Psi_n = n^{-1} H_n' \Sigma_n H_n,$$

and we need to estimate the $n(n+1)/2$ unknowns in Σ_n . So we need some assumptions.

- Assumptions.

- (0) Distances between locations satisfy $d_{ij,n} = d_{ji,n} \geq 0$, which the researcher measures possibly with error as:

$$d_{ij,n}^* = d_{ji,n}^* \geq 0.$$

- (1) $\epsilon \sim iid(0, \sigma^2)$ with finite fourth moments (at least).
- (2) R_n and R_n^{-1} non singular and absolutely summable.
- (3) H_n is uniformly bounded.
- (4) So this one is important. For each location, the researcher specifies a cut-off distance $d_n > 0$ such that if $d_{ij,n} < d_n$, then j is a neighbor to i , and otherwise, i and j are not neighbors. l_n is the maximum number of neighbors any location has. Then we must have:

$$l_n < n^{1/3},$$

ie any location can't have too many neighbors.

- (5) The distances observed:

$$d_{ij,n}^* = d_{ij,n} + v_{ij,n},$$

have bounded errors, ie $|v_{ij,n}| \leq c$.

- (6) Complicated. See HK notes. Involves something similar to the OLS trick of writing:

$$\hat{u}_i = u_i - x_i(\hat{\beta} - \beta).$$

- (7) Kernels. So we have a kernel, $K : \mathfrak{R} \mapsto [-1, 1]$ with $K(0) = 1$, $K(x) = K(-x)$, and $K(x) = 0$ for $|x| > 1$. One might be:

$$K\left(\frac{d_{ij,n}}{d_n}\right) = 1 - \frac{d_{ij,n}}{d_n}.$$

- HAC ESTIMATOR. So the (r, s) element of Ψ_n looks like:

$$\psi_{rs,n} = n^{-1} \sum_{i=1}^n \sum_{j=1}^n h_{ir,n} h_{js,n} \sigma_{ij,n},$$

which we estimate with the (r, s) element of $\hat{\Psi}_n$ which looks like:

$$\hat{\psi}_{rs,n} = n^{-1} \sum_{i=1}^n \sum_{j=1}^n h_{ir,n} h_{js,n} \hat{u}_{i,n} \hat{u}_{j,n} K(d_{ij,n}^*/d_n).$$

- **Theorem 1:** $\hat{\Psi}_n - \Psi_n \rightarrow^p 0$, consistent.
- The paper goes on to do a slight variation of all of this based on multiple distance measures. This technique has never been implemented (hell, this paper isn't even published yet), so it might be something to return to later. The idea is that you might have several ways to measuring distance between unit and you're not quite sure which to use. Income per capita, geographic, etc. So you define them all and if the distance between i and j for ANY of the distances falls below some cut off, then count i and j as neighbors.
- Similar assumptions as before reflecting the multiple distance measures but we again get a consistent HAC estimator out of all of this which looks like:

$$\hat{\psi}_{rs,n} = n^{-1} \sum_{i=1}^n \sum_{j=1}^n h_{ir,n} h_{js,n} \hat{u}_{i,n} \hat{u}_{j,n} K(\min_m \{d_{ij,m,n}^*/d_{m,n}\}),$$

where we have m measures of distance.

- Application to a general linear spatial model:

$$y_n = X_n \beta_0 + \lambda_0 W_n y_n + Y_n \gamma_0 + u_n, \quad u_n = R_n \epsilon_n, \quad \epsilon_n \sim (0, I_n).$$

Write, $y_n = Z_n \delta_0 + u_n$.

- To estimate this, we need instruments for both $W_n y_n$ and the endogenous (non-spatially lagged) Y_n 's. So we have:

$$\hat{Z}_n = H_n (H_n' H_n)^{-1} H_n' Z_n,$$

and,

$$\hat{\delta}_{2SLS,n} = (\hat{Z}_n' Z_n)^{-1} \hat{Z}_n' y_n.$$

Then let:

$$\hat{u}_n = y_n - Z_n \hat{\delta}_{2SLS,n},$$

and use these to form our HAC estimator of $\Psi_n = n^{-1} H_n' \Sigma_n H_n$.

- Given our model, we have:

$$n^{1/2}(\hat{\delta}_n - \delta_0) \rightarrow^d N(0, M \Psi M'),$$

which we estimate with $\hat{\Psi}_n$ as above (using the residuals and the Kernel) and:

$$\hat{M}_n = n(\hat{Z}_n' \hat{Z}_n)^{-1} \hat{Z}_n' H_n (H_n' H_n)^{-1}.$$

24 Lecture 24: May 4, 2006

24.1 Time Series Econometrics

ARCH/GARCH

- (Generalized) Auto Regressive Conditional Heteroskedasticity models or (G)ARCH models allow for variances to depend on the values of a lagged disturbances (and possibly lagged variances). Perfect for financial instruments.
- See G-24.1. Consider the ARCH model:

$$Y_t = X_t B + \epsilon_t, \quad t = 1, \dots, T.$$

Assume X_t non-stochastic, $T^{-1}X'X \rightarrow Q_{xx}$ with Q_{xx}^{-1} exists. Then the disturbances are assumed to be conditional mean zero and:

$$E[\epsilon_t^2 | \epsilon_{t-1}, \epsilon_{t-2}, \dots] = h(\epsilon_{t-1}) > 0,$$

where the last holds for an ARCH(1) model.

- The usual specification is:

$$\epsilon_t = u_t [\alpha_0 + \alpha_1 \epsilon_{t-1}^2]^{1/2}, \quad \alpha_0 > 0, \quad \alpha_1 \in [0, 1), \quad u_t \sim iid N(0, 1).$$

Note we can assume that u_t has variance 1 because making it more general would leave the variance unidentified anyway, so we might as well work with a $N(0, 1)$.

- The α_1 term is the so called “ARCH effect.” We will be primarily interested in seeing if it is nonzero.

- Properties of the error term:

$$\begin{aligned} \text{Conditional Mean: } E[\epsilon_t | \epsilon_{t-1}] &= E[u_t][\alpha_0 + \alpha_1 \epsilon_{t-1}^2]^{1/2} \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Unconditional Mean: } E[\epsilon_t] &= E[u_t]E[(\alpha_0 + \alpha_1 \epsilon_{t-1}^2)^{1/2}] \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Conditional Variance: } E[\epsilon_t^2 | \epsilon_{t-1}] &= E[u_t^2][\alpha_0 + \alpha_1 \epsilon_{t-1}^2] \\ &= \alpha_0 + \alpha_1 \epsilon_{t-1}^2 \end{aligned}$$

$$\begin{aligned} \text{Unconditional Variance: } E[\epsilon_t^2] &= E[u_t^2]E[\alpha_0 + \alpha_1 \epsilon_{t-1}^2] \\ &= \alpha_0 + \alpha_1 E[\epsilon_{t-1}^2] \\ &\quad \text{let } d_t = E[\epsilon_t^2] \end{aligned}$$

$$d_t = \alpha_0 + \alpha_1 d_{t-1}$$

$$d_t(1 - \alpha_1 L) = \alpha_0$$

$$d_t = \frac{\alpha_0}{1 - \alpha_1 L} = \frac{\alpha_0}{1 - \alpha_1}$$

- So unconditionally,

$$\epsilon_t \sim \left(0, \frac{\alpha_0}{1 - \alpha_1}\right) \equiv (0, \sigma_\epsilon^2),$$

homoskedastic!

- But conditionally,

$$\epsilon_t | \Lambda_{t-1} \sim (0, \alpha_0 + \alpha_1 \epsilon_{t-1}^2),$$

heteroskedastic!

- Unfortunately for us, we need use the conditional form of the errors if we want to see what the α parameters look like. Otherwise they are not identified. Using the unconditional form means that OLS is BLUE.
- Note also the covariance, $E[\epsilon_t \epsilon_s] = 0$ for $t \neq s$.
- So to model things using the conditional errors, we need to do a nonlinear estimation so we'll use ML.

Non-Linear Estimation of an ARCH(1) with Maximum Likelihood

- Given the joint density of the errors as $f(\epsilon_1, \dots, \epsilon_T | \epsilon_0)$, write our likelihood:

$$\mathcal{L} = f_1(\epsilon_1 | \epsilon_0) f_2(\epsilon_2 | \epsilon_0, \epsilon_1) \cdots f_T(\epsilon_T | \epsilon_0, \epsilon_1, \dots, \epsilon_{T-1}),$$

where $\epsilon_t = Y_t - X_t B$ and we observe the initial values, Y_0 and X_0 .

- Given that $\epsilon_t | \Lambda_{t-1} \sim N(0, \alpha_0 + \alpha_1 \epsilon_{t-1}^2)$, the likelihood becomes:

$$\mathcal{L} = \left[(2\pi(\alpha_0 + \alpha_1 \epsilon_0^2))^{-1/2} \exp\left(-0.5 \frac{\epsilon_1^2}{\alpha_0 + \alpha_1 \epsilon_0^2}\right) \right] * \left[(2\pi(\alpha_0 + \alpha_1 \epsilon_1^2))^{-1/2} \exp\left(-0.5 \frac{\epsilon_2^2}{\alpha_0 + \alpha_1 \epsilon_1^2}\right) \right] \\ * \dots * \left[(2\pi(\alpha_0 + \alpha_1 \epsilon_{T-1}^2))^{-1/2} \exp\left(-0.5 \frac{\epsilon_T^2}{\alpha_0 + \alpha_1 \epsilon_{T-1}^2}\right) \right].$$

Or in slightly more compact notation:

$$\mathcal{L} = L(\alpha_0, \alpha_1, B | data).$$

- So just do ML on that beast (which is already canned in all statistical software) and you get your ML estimates of B along with the α terms in the ARCH disturbances. You could then test for ARCH effects by testing if $\alpha_1 = 0$ versus $\alpha_1 > 0$.

Non-Linear Estimation of an ARCH(p) with Maximum Likelihood

- Suppose the errors looked like:

$$\epsilon_t = u_t [\alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_p \epsilon_{t-p}^2]^{1/2}, \quad \alpha_0 > 0, \quad \alpha_i \in [0, 1) \quad i = 2, \dots, p, \quad u_t \sim iid N(0, 1).$$

- Once again we have the following properties:

$$\begin{aligned} E[\epsilon_t | \Lambda_{t-1}] &= 0 \\ E[\epsilon_t] &= 0 \\ E[\epsilon_t^2 | \Lambda_{t-1}] &= \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_p \epsilon_{t-p}^2 \\ E[\epsilon_t^2] &= \frac{\alpha_0}{1 - \alpha_1 - \dots - \alpha_p} \end{aligned}$$

We again estimate by ML and test for ARCH effects using a Wald test under the null of NO ARCH effects:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0.$$

GARCH(1,1) Model

- Consider the model:

$$Y_t = X_t B + \epsilon_t, \quad t = 1, \dots, T. \\ \epsilon_t | \Lambda_{t-1} \sim N(0, \sigma_t^2),$$

where,

$$\sigma_t^2 = P_1(L) \sigma_t^2 + P_2(L) \epsilon_t^2.$$

- The usual specification:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \alpha_2 \epsilon_{t-1}^2, \quad \alpha_1 \in [0, 1),$$

a GARCH(1,1) model. So we've added a dependence on past variances.

- If we iteratively substitute and solve for σ_t^2 just in terms of ϵ_t 's, we'll still be left with a σ_0^2 term. There are several ways to deal with this (view it as a parameter to be estimated, use the unconditional variance, etc), but asymptotically it really doesn't matter.
- Estimate the GARCH model also with ML.

The ARCH-M Model: ARCH in the Mean

- Consider the model:

$$y_t = x_t' \beta + h_t \phi + \epsilon_t, \quad t = 1, \dots, T,$$

$$\epsilon_t | \Lambda_{t-1} \sim N(0, h_t^2),$$

$$h_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \dots + \alpha_p \epsilon_{t-p}^2.$$

In this model the standard deviation, h_t , enters the mean of the regression function and so noise effects both the mean and variance of the dependent variable.

- See notes for the Bera-Ra LM test which allows one to test for ARCH effects under this specification. The problem is that under the null of no ARCH effects, there are two constants in the model (one in β and α_0) and so ϕ is not identified.
- Done.

25 Lecture 25: May 9, 2006

25.1 Time Series Econometrics

General Introduction

- Consider the series:

$$Y_t = \rho Y_{t-1} + u_t, \quad |\rho| < 1, \quad u_t \sim iid(0, \sigma_u^2).$$

- Properties of this stationary time series:

- (1) $E[Y_t] = 0$.
- (2) $\sigma_Y^2 = \frac{\sigma_u^2}{1 - \rho^2}$.
- (3) Series can be written: $Y_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots$.
- (4) $Cov(Y_t, Y_{t-k}) = \rho^k \sigma_Y^2 \rightarrow 0$ as $k \rightarrow \infty$.
- (5) $Corr(Y_t, Y_{t-k}) = \rho^k \rightarrow 0$ as $k \rightarrow \infty$.
- (6) The expected length of time between crossing the mean (0) is finite (see Engle and Granger, *Econometrica* (1987)).

- What if $\rho = 1$? We have a random walk:

$$Z_t = Z_{t-1} + u_t, \quad u_t \sim (0, \sigma_u^2).$$

- Properties of a non-stationary times series:

- (1) $E[Z_t] = 0$.
- (2) $\sigma_Z^2 = t\sigma_u^2 \rightarrow \infty$ as $t \rightarrow \infty$.
- (3) Series can be written: $Z_t = u_t + u_{t-1} + \dots + u_1$.
- (4) $Cov(Z_t, Z_{t-k}) = (t-k)\sigma_u^2 \rightarrow \infty$ as $t-k \rightarrow \infty$.
- (5) $Corr(Z_t, Z_{t-k}) = \frac{\sqrt{t-k}}{\sqrt{t}} \rightarrow 1$ as $t \rightarrow \infty$ for all k .
- (6) The expected length of time between crossing the mean (0) is infinite (see Engle and Granger, *Econometrica* (1987)).

- To test for a unit root, use a Dickey Fuller test. The standard large sample theory does NOT apply to unit root processes.

- Other problems with unit roots.

- (1) If Y_t is a random walk, it can be written:

$$Y_t = \epsilon_t + \dots + \epsilon_1 = \bar{\epsilon}_t.$$

If you regress Y_t on a constant, a time trend, and $\bar{\epsilon}_t$, you'll get significance for the constant and time trend fairly often due to spurious correlation. R^2 's will also be large.

- (2) If Y_t is regressed on a constant, time trend, and an exogenous X , the coefficient on X will also come out significant.

- So clearly unit roots are a problem. Check your series first for unit roots and proceed once you have eliminated them (via next section).
- Suppose your model is:

$$Y_t = \alpha v_t + bz_t + \epsilon_t,$$

where $v_t \sim \text{iid}$, z_t is a unit root process, and $\epsilon_t \sim \text{iid}$. If we just run OLS on this model, our large sample distribution of α will NOT depend on z_t so we don't need to worry about the presence of z_t in the model.

Integrated Variables

- If you take a non-stationary series, Y_t , and difference it d times and end up with a stationary series, then Y_t is said to be integrated of order d , or $Y_t \sim I(d)$.
- Example. $Y_t = Y_{t-1} + \epsilon_t$, $Y_0 = 0$, and $\epsilon_t \sim \text{iid}(0, \sigma^2)$, then $\Delta Y_t = \epsilon_t$, stationary. So $Y_t \sim I(1)$. Note $\epsilon_t \sim I(0)$.
- Example. If $Y_t = f(Z_t)$ where $Z_t \sim I(2)$, then $Y_t \sim I(2)$. Ie $\Delta^2 Y_t \sim I(0)$.
- It is possible for a non-stationary process to NOT be intergrated of ANY order! Eg, $Y_t = Y_{t-1} + \epsilon_t$, but $\epsilon_t \sim (0, \sigma_t^2)$, heteroskedastic.
- In general if the RHS of an equation is $I(d)$, then so is the LHS.
- If $X_t \sim I(d)$ and $Z_t \sim I(d + \alpha)$, then

$$Y_t = X_t + Z_t \sim I(d + \alpha),$$

ie, the big term dominates.

- Consider the following model:

$$Y_t = \beta v_t + \gamma z_t + \epsilon_t,$$

where $v_t \sim I(0)$, $z_t \sim I(1)$ and $\epsilon_t \sim I(0)$, then we can still estimate $\hat{\beta}$ consistently via OLS for the same reason as above. The $I(1)$ process does not affect the $I(0)$.

Cointegration

- Typically economists deal with $I(1)$ variables (GDP, income, prices, interest rates, etc). Suppose X_t is an $N \times 1$ vector of $I(1)$ variables.

- If there is a vector, α , $N \times 1$, such that:

$$\alpha'X_t \sim I(0),$$

then the variables in X_t are said to be “cointegrated” and α is the cointegrating vector.

- We can write $\alpha'X_t = \epsilon_t \sim I(0)$, where ϵ_t reflects deviations from equilibrium at time t .
- More next time.

26 Lecture 26: May 11, 2006

26.1 Time Series Econometrics

More on Cointegration

- **The following is material beyond the final and also beyond comprehension.**
- Some examples of economic variables that are cointegrated:

- (1) Consumption, C_t , and income, Y_t , are suggested to be $I(1)$. Thus,

$$C_t = \lambda Y_t + \epsilon_t, \quad \epsilon_t \sim I(0),$$

where $\alpha = [1, -\lambda]$ is our cointegrating vector.

- (2) Interest rates of various time horizons tend to be cointegrated.

- **Remarks**

- (a) Variables integrated of higher orders than one can be cointegrated, but usually we only work with $I(1)$'s.
- (b) If C_t and Y_t in our example above are cointegrated, then the non-stationary components must cancel if λY_t is subtracted from C_t . See notes.
- (c) In the literature, non-stationary and unit root are used interchangeably.
- (d) More than one cointegrating vector is possible and linear combinations of cointegrating vectors are cointegrating vectors.
- (e) If X_t is a $N \times 1$ vector of $I(1)$ elements, there are AT MOST $N - 1$ cointegrating vectors.

A Test for Cointegration

- Suppose you suspect that Y_t and C_t are cointegrated. Ie, there exists a λ such that:

$$C_t - \lambda Y_t = \epsilon_t \sim I(0).$$

- Hypotheses:

$$H_0 : \epsilon_t \sim I(1) \text{ (No Cointegration)} \text{ vs } H_1 : \epsilon_t \sim I(0) \text{ (Cointegration)}.$$

- To run this, just do OLS of C_t on Y_t to get an estimate for $\hat{\lambda}$. Then do a DF type test on the residuals to test for a unit root. If the resids are $I(0)$, reject the null.

The Error Correction Model

- Suppose $Y_t \sim I(1)$ and Y_t is $N \times 1$.
- Suppose Y_t satisfies:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \alpha + \epsilon_t, \quad \epsilon_t \sim (0, \Omega),$$

where $\epsilon_t \sim I(0)$.

- Add and subtract $\phi_3 Y_{t-2}$:

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \alpha + \epsilon_t + \phi_3 Y_{t-2} - \phi_3 Y_{t-2} \\ &= \phi_1 Y_{t-1} + (\phi_2 + \phi_3) Y_{t-2} - \phi_3 \Delta Y_{t-2} + \alpha + \epsilon_t \end{aligned}$$

- Add and subtract $(\phi_2 + \phi_3) Y_{t-1}$:

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + (\phi_2 + \phi_3) Y_{t-2} - \phi_3 \Delta Y_{t-2} + \alpha + \epsilon_t + (\phi_2 + \phi_3) Y_{t-1} - (\phi_2 + \phi_3) Y_{t-1} \\ &= (\phi_1 + \phi_2 + \phi_3) Y_{t-1} - (\phi_2 + \phi_3) \Delta Y_{t-1} - \phi_3 \Delta Y_{t-2} + \alpha + \epsilon_t \end{aligned}$$

- Subtract Y_{t-1} from both sides:

$$\begin{aligned} Y_t - Y_{t-1} &= (\phi_1 + \phi_2 + \phi_3) Y_{t-1} - (\phi_2 + \phi_3) \Delta Y_{t-1} - \phi_3 \Delta Y_{t-2} + \alpha + \epsilon_t - Y_{t-1} \\ \Delta Y_t &= (\phi_1 + \phi_2 + \phi_3 - I) Y_{t-1} - (\phi_2 + \phi_3) \Delta Y_{t-1} - \phi_3 \Delta Y_{t-2} + \alpha + \epsilon_t \\ &= b_0 Y_{t-1} + a_1 \Delta Y_{t-1} + a_2 \Delta Y_{t-2} + \alpha + \epsilon_t \end{aligned}$$

So now everything is $I(0)$, as long as $b_0 Y_{t-1} \sim I(0)$. But this means that b_0 is a cointegrating vector for Y_{t-1} !

- The notes go on to show that you can rewrite that b_0 term to involve fewer parameters making use of the fact that b_0 is non-singular (due to remark (e) above). A two variable system is shown on page 27 and an estimation procedure by Engle-Granger is laid out on page 28. OLS can be applied in two stages to estimate the model in error correction form.
- See Hamilton.