Linear Regression Model

Dependent Variable - focus of study; want to know how other factors (called regressors, "independent" variables, exogenous variables, or covariates) affect the dependent variable; also called **endogenous variable** or **regressand**

Exogenous Variables - have no control over them at time of decision; also called **characteristic variables**, **independent variables**, or **regressors**

- **Demand Example** - exogenous variables in theory include prices and income, but gets more complicated in real life:
  - **Household vs. Individual** - number and age of kids influence demand
  - **Other Factors** - education, religion, gender, ethnicity
- **Supply** - complications arise here too: capital structure, organizational structure, technology, location
- **Result** - empirical work needs to worry about all possible exogenous variables

Specify Model - $y = f(\mathbf{x})$; where $\mathbf{x}$ is a column vector of exogenous variables

- **Specification** - economic theory gets us an approximation of f, but we never actually know f because we can't know or collect data on all possible exogenous variables
- **Data** - collect sample data to check if $y_i = f(\mathbf{x}_i)\ \forall\ i = 1,..., N$
- **Problem** - we'll never find f that works for all data (e.g., two individuals can have same characteristics $\mathbf{x}_i$, but different demand y because of utility function (individual taste) which depends on unobserved factors
  - **Add Error Term** - $y = f(\mathbf{x}, u)$; 4 explanations (3 & 4 are intuitive explanations for undergrads, not this course)
    1. Unobserved Factors - heterogeneity; problem described above of data points with same characteristic variables resulting in different demand (i.e., same $\mathbf{x}$ yields different y)
    2. Approximation - wrong functional form; don't really know f, just estimating as close as we can
    2. Errors in Data - either recorded wrong or reported incorrectly
    3. Average vs. Individual - no way to forecast individual demand so only focus on average demand for given characteristics; this is "good enough" for firms and government because they're worried about market/population demand; simply multiply average demand by number in population

Transition - steps to get from theoretical model to econometric model:

1. **Add Error Term**
2. **Proxy Variables** - if we can't observe variables from theoretical model we use a proxy that is highly correlated to theoretical variable it replaces
   - **"Health" Example** - "health status" is not clear or could be too subjective so use proxy variable: # times sick, # doctor visits, etc.
   - **"Quality" Example** - # defects per million units of output could be a proxy

$y^* = f(\mathbf{x}^*, u) \Rightarrow y = g(\mathbf{x}, u)$... here y and $\mathbf{x}$ are proxy for y^* and $\mathbf{x}^*$

Econometrics - unification of economic theory, mathematics, and statistics

- **Theory** - used to develop the model and again to interpret and discuss the results
  - **Note:** if there's no theoretical model, rely on common sense or institutional knowledge (mechanism that generated the data) to develop a model; this could lead to a new theory (depending on results you get)
- **Statistics/Math** - used to solve the model and derive the results

Linear Regression - refers to linear in parameters (β_j), not necessarily the regressors (exogenous variables)

Basic Model - 3 ways to write it: k parameters; N observations

$$\boxed{y_i = \mathbf{x}_i{}'\boldsymbol{\beta} + u_i} \quad \text{or} \quad \boxed{y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k + u_i} \quad \text{or} \quad \boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}}$$

$$\mathbf{x}_i = \underset{k \text{ x } 1}{\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{bmatrix}} \quad \boldsymbol{\beta} = \underset{k \text{ x } 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}} \quad \mathbf{Y} = \underset{N \text{ x } 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1{}' \\ \mathbf{x}_2{}' \\ \vdots \\ \mathbf{x}_k{}' \end{bmatrix} = \underset{N \text{ x } k}{\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{kN} \end{bmatrix}} \quad \mathbf{U} = \underset{N \text{ x } 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}}$$

$\mathbf{x}_i$ is column vector of exogenous variables; can include squared terms, interaction terms, dummy variables, etc.; when there's more than one exogenous variable, this is called **multiple regression**

$\mathbf{X}$ - matrix of all observed exogenous variables; each row consists of a single observation (i) of the exogenous variables (for a total of N rows); there are k columns, one for each exogenous variable

x_{ji} - the i^{th} observation of the j^{th} exogenous variable; elements of the $\mathbf{X}$ matrix; note the <u>backwards notation</u> here... i refers to the row, but it's the second number

$\boldsymbol{\beta}$ is column vector of parameters

u_i is uncontrolled factor or error term

Functional Form - could be step function, polynomial, log, etc.; which is best is topic for a more advanced course

Linear - most important because all functions can be locally approximated by a linear function

$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k + u_i$

Non-Linear - allows interactions between variables (x_1x_2) and polynomial terms (x_1^2); important because Taylor Series allows up to approximate any function with a polynomial... in theory, of course; it could be too difficult to interpret in practice if it's a very high order polynomial

$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_3 + x_{1i}^2\beta_4 + u_i$

Step Function - uses dummy variables to represent a step function

$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + d_{1i}\beta_3 + u_i$, where $d_{1i} = 1$ if $x_{3i} < 10$; 0 otherwise

Log-Linear - usually used when left hand side variables (y_i) is always > 0

$y_i = \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + u_i$ (Cobb-Douglas)

Trans-Log - allows higher order terms since many people argue Cobb-Douglas isn't realistic for production functions)

$y_i = \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3(\ln x_{1i})(\ln x_{2i}) + \beta_3(\ln x_{1i})^2 + u_i$

Building and Interpreting Model -

Sales Pitch - need to say why you're interested in y and why other people should care

Justification - why is each x included in the model; most criticism of empirical work comes from which variables are used (or not used) in the model

Objective - estimate $\boldsymbol{\beta}$ and make inferences; want to know how characteristic variables affect the average y

Interpretation of Coefficients - depends on functional form

Demand Example - $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k + u_i$, where y_i is demand for good 1 for individual i; x_{ji} is price of good j ($j \neq k$) and x_{ki} is income for individual i; when other variables don't change, how does price of good 1 affect demand? Marginal effect of x_1 on y... $\partial y/\partial x_1 = \beta_1$

Warning - this isn't always the case; examples where it doesn't happen:

$D_i = \beta_1 + \beta_2 P_i + \beta_3 I_i + \beta_4 I_i^2 + u_i$... here $\partial y/\partial I \neq \beta_3$ because of I_i^2 term

$D_i = \beta_1 + \beta_2 P_i + \beta_3 I_i + \beta_4 P_i I_i + u_i$... here $\partial y/\partial P \neq \beta_2$ and $\partial y/\partial I \neq \beta_3$ (have $P_i I_i$ term)

Still OK - can still answer question of marginal effect on demand; it's just not as simple as a single coefficient

Estimating Parameters

Multiple Regression - statistical technique to isolate how individual x affects y with others unchanged; works even if x_i are correlated (just not highly correlated; "high" is relative to sample size; larger sample allows higher correlation); **Note:** we're still doing linear regression; it's called multiple regression in the general case when there's more than one regressor

4 Assumptions - the first 2 are essential; the second 2 are to make computations easier and aren't required unless you're using a statistical package that uses them

1. **Error Uncorrelated to Contemporary Regressors** - $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k + u_i$, we want to change x_{1i} while holding all other x_{ji}'s constant; we look at change in y_i and use that to estimate β_1... but only if u_i doesn't change either (i.e., x_{1i} & u_i not correlated); since x_{1i} & u_i are both from the ith observation, the assumption deals with "contemporary" regressors (all the x_{ji} are from the same observation); there are a couple ways to write this assumption:
   - $E(u_i x_{ji}) = 0 \ \forall\ i = 1, ..., N; j = 1, ..., k$
   - $E(u_i \mathbf{x}_i) = \mathbf{0} \ \forall\ i = 1, ..., N$ ($\mathbf{x}_i$ a column vector shown on previous page)

   Zero Error on Average - if we use a constant term, we basically have a column of 1s in the $\mathbf{X}$ matrix (i.e., $\exists\ j$ such that $x_{ji} = 1 \ \forall\ i = 1, ..., N$); in such a case, we get the implicit assumption that the expected value of the error term is zero:
   - $E(u_i) = 0 \ \forall\ i = 1, ..., N$

   Stronger Condition - some textbooks start with $E(u \mid x_1, x_2, ..., x_k) = 0$; this is a much stronger condition as we'll see in a minute, but it implies the basic assumption
2. **$\mathbf{X'X}$ and $E(\mathbf{X'X}/N)$ Nonsingular** - estimate for parameters:

   Start with assumption: $E(u_i \mathbf{x}_i) = \mathbf{0}$

   Substitute model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$: $E((y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i) = \mathbf{0}$

   Multiply that out and break up the expectation: $E(\mathbf{x}_i y_i) - E(\mathbf{x}_i \mathbf{x}_i')\boldsymbol{\beta} = \mathbf{0}$

   We pulled $\boldsymbol{\beta}$ out of the expectation because it's a constant

   Now solve for $\boldsymbol{\beta}$: $\boldsymbol{\beta} = \left[E(\mathbf{x}_i \mathbf{x}_i')\right]^{-1} E(\mathbf{x}_i y_i)$

   This brings us to the theoretical part of assumption 2; there are a couple ways to write it:
   - $E(\mathbf{x}_i \mathbf{x}_i')$ is nonsingular
   - $E\left(\dfrac{\mathbf{X'X}}{N}\right)$ is nonsingular

   There's a problem with using $E((y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i) = \mathbf{0}$: we can't really get expected values so we try to approximate using sample average:

$\frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i \approx \mathbf{0}$ ... to get = $\mathbf{0}$ we use $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$: $\frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i = \mathbf{0}$

Now we can solve for $\hat{\boldsymbol{\beta}}$: $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i$

This brings us to the practical part of assumption 2 (you'll know it doesn't hold if the computer crashes when you try to estimate the parameters):

- $\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'$ is nonsingular
- $\mathbf{X}'\mathbf{X}$ is nonsingular

Meaning - practically, what this assumption means is that no two rows (or columns) in $\mathbf{X}$ can be identical or no row (or column) in $\mathbf{X}$ can be a linear combination of the other rows (or columns)... in practical terms: each exogenous variable must bring some new information (i.e., can't just repeat what other variables tell you)

Estimating $\hat{\boldsymbol{\beta}}$ - there are several ways to write it out:

$$\boxed{\begin{array}{ll} \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i & \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i u_i \\ \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} & \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U} \end{array}}$$

To get from estimate on left to the one on the right, we substitute for y_i (or $\mathbf{Y}$):

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i(\mathbf{x}_i'\boldsymbol{\beta} + u_i) =$$

Break up the second summation: $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left[\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)\boldsymbol{\beta} + \sum_{i=1}^{N}\mathbf{x}_i u_i\right]$

Multiply it out (note the inverse cancels in the first term):

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i u_i$$

Intuition - $\hat{\boldsymbol{\beta}}$ close to $\boldsymbol{\beta}$ if sample average is close to population average; we know from statistics that sample average is consistent and unbiased estimator of population average

Consistent - $\hat{\boldsymbol{\beta}}$ is consistent (i.e., $\hat{\boldsymbol{\beta}} \xrightarrow{\text{P}} \boldsymbol{\beta}$ as $N \to \infty$); consistent means that as N gets larger, $\hat{\boldsymbol{\beta}}$ is "more likely" to be close to $\boldsymbol{\beta}$

Identification Conditions - assumptions 1 & 2 combined are called the identification or regularity conditions; they guarantee that we can calculate $\hat{\boldsymbol{\beta}}$ and that it is consistent

Normal Distribution - assumptions 1 & 2 can be combined with other technical statistics stuff (like central limit theorem; specifics not important to this course) to say that:

$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, (\mathbf{X'X})^{-1}\boldsymbol{\Omega}(\mathbf{X'X})^{-1})$

Note: we did not have to assume u_i ~ Normal

Variance term comes from $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'U}$ and the fact that $Var(z) = E(z^2) - [E(z)]^2$:

$$Var((\mathbf{X'X})^{-1}\mathbf{X'U}) = E\left(\left[(\mathbf{X'X})^{-1}\mathbf{X'U}\right]^2\right) - \left[E\left((\mathbf{X'X})^{-1}\mathbf{X'U}\right)\right]^2$$

In the right term, we can pull $(\mathbf{X'X})^{-1}$ out of the expectation because it's known from the sample; with $E(\mathbf{X'U})$ a little work with the matrices will show:

$$E(\mathbf{X'U}) = E\begin{bmatrix} \sum_{i=1}^{N} x_{1i}u_i \\ \vdots \\ \sum_{i=1}^{N} x_{1i}u_i \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} E(x_{1i}u_i) \\ \vdots \\ \sum_{i=1}^{N} E(x_{1i}u_i) \end{bmatrix} = \mathbf{0} \text{ (from assumption 1)}$$

Now we have $Var((\mathbf{X'X})^{-1}\mathbf{X'U}) = E\left(\left[(\mathbf{X'X})^{-1}\mathbf{X'U}\right]^2\right) = E\left[(\mathbf{X'X})^{-1}\mathbf{X'UU'X}(\mathbf{X'X})^{-1}\right]$

We can pull the $(\mathbf{X'X})^{-1}$ out of the expectation and get $(\mathbf{X'X})^{-1}E(\mathbf{X'UU'X})(\mathbf{X'X})^{-1}$, so if we let $\boldsymbol{\Omega} = E(\mathbf{X'UU'X})$, we get the variance term we had above

If we work out the matrix stuff we can find that

$$\boldsymbol{\Omega} = E(\mathbf{X'UU'X}) = E\left([\mathbf{X'U}]^2\right) = E\left(\left[\sum_{i=1}^{N}\mathbf{x}_i u_i\right]^2\right) = E\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j' u_i u_j\right)$$

(at least that's what I wrote down in class... looks hard to prove)

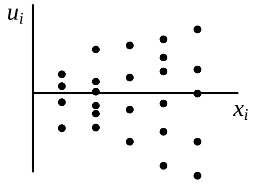
3. **No Autocorrelation** - this is a very technical assumption that is purely statistical to simplify calculations; it doesn't really have economic meaning and we don't need it to get a good estimate of to $\boldsymbol{\beta}$; basically this assumption means that any two error terms (u_i and u_j) are uncorrelated (e.g., any two firms or individuals don't affect each other); there are several ways to write this, each slightly different; they're listed here from weakest to strongest

3a. $E(\mathbf{x}_i\mathbf{x}_j' u_i u_j) = \mathbf{0} \;\forall\; i \neq j$ - this allows us to simplify the variance term:

$$\boldsymbol{\Omega} = E\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{x}_i\mathbf{x}_j' u_i u_j\right) = \sum_{i=1}^{N} E(\mathbf{x}_i\mathbf{x}_i' u_i^2) \quad \text{(all the terms with } i \neq j \text{ go away)}$$

3b. $E(u_i u_j \mid \mathbf{x}_i, \mathbf{x}_j) = 0$

3c. $E(u_i \mid \mathbf{x}_i) = 0 \;\forall\; i = 1, ..., N$ and u_i and u_j are independent - this is the strongest version used in classical linear regression; it means that u_i and $\mathbf{x}_i$ are <u>orthogonal</u>; in a graph, it means that the average of the error terms for each value of the exogenous variables is zero



Average u in each column is zero

Unbiased - $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$; doesn't depend on sample size; doesn't mean estimate is close to true value; single estimate can be very wrong, but average of estimates will be close to $\boldsymbol{\beta}$

Start with $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i u_i$

Now take the expectation: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + E\left[\left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i\right]$

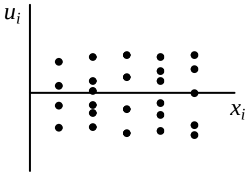
In order for $\hat{\boldsymbol{\beta}}$ to be unbiased, that term on the right must equal zero; it's too complicated to work with so we (obviously) use the iterative expectation rule: $E(UV) \equiv E_V(E_U(U|V))$

$$E\left[\left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i\right] = E\left[\left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i E\left(u_i \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)\right]$$

$\therefore$ $\hat{\boldsymbol{\beta}}$ is unbiased if $E(u_i|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = 0$ $\forall$ i = 1, 2, ..., N (i.e., **orthogonal**) that is, u_i is uncorrelated to all $\mathbf{x}$, not just $\mathbf{x}_i$ like we need for $\hat{\boldsymbol{\beta}}$ to be consistent

Not Required - an unbiased estimator is better to have for small sample sizes, but a consistent estimator is better for large samples; large samples are now common so consistency is better (also allows less restrictive assumptions about data)

4. **Homoskedasticity** - to simplify the computations even further, the next assumption says that the error term u_i has the same variance for all i (i.e. there are no patterns in the error terms with respect to $\mathbf{x}_i$... doesn't get bigger or smaller)



Variance of u in each column is the same

$$\boldsymbol{\Omega} = \sum_{i=1}^{N} E(\mathbf{x}_i \mathbf{x}_i' u_i^2) = \sigma^2 \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i' \text{ which can also be written } \sigma^2 \mathbf{X}'\mathbf{X}$$

Going back to the whole point of simplifying these calculations, look at the distribution now with assumptions 3 & 4:

$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1} \boldsymbol{\Omega} (\mathbf{X}'\mathbf{X})^{-1})$ or

$$\boxed{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2) \quad \text{or} \quad \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left(\mathbf{0}, \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sigma^2\right)}$$

Summary of Assumptions - these are added to the assumptions that we have the correct model and that it's linear in the parameters:

1. $E(u_i \mathbf{x}_i) = \mathbf{0}$ $\forall$ i = 1, ..., N - error uncorrelated to contemporary regressors
2. $\mathbf{X}'\mathbf{X}$ and $E(\mathbf{X}'\mathbf{X}/N)$ Nonsingular - combined with assumption 1, these are the identification or regularity conditions (i.e., we can find $\boldsymbol{\beta}$ in theory or $\hat{\boldsymbol{\beta}}$ in practice)

   1 & 2 with some technical stat stuff say $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left[\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1} E(\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\right]$
3. $E(\mathbf{x}_i \mathbf{x}_j' u_i u_j) = \mathbf{0}$ $\forall$ $i \neq j$ - error terms are unrelated to each other
4. $E(u_i^2) = \sigma^2$ - homoskedasticity (error terms have the same variance)

   3 & 4 say $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$

Testing Parameters

What to Estimate - most computer packages start with the four assumptions we just covered so they use $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$; the only thing we need to estimate here is

$\sigma^2 = Var(u_i) = E(u_i^2)$ ... replace with sample average... $\frac{1}{N}\sum_{i=1}^{N} u_i^2$

Estimate *u_i* - now problem is that we don't know $u_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ ... and problem there is that we don't know $\boldsymbol{\beta}$

Regression Residuals - estimate u_i with $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$

Estimate Variance - $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2$

Sample Variance - $\boxed{S^2 = \frac{1}{N-k}\sum_{i=1}^{N} e_i^2}$, where k= # of parameters (β's)

Note: both $\hat{\sigma}^2$ and S^2 converge to σ^2 (as $N \to \infty$), but S^2 gives better estimates for small sample size; most packages use S^2

Check Assumptions - if there's enough evidence in the data to think 3 or 4 don't hold, then we can't use the results from the package

Time Series - u_i serially correlated which violates assumption 3

Refresher - here's what we're working with in order to test the parameters:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \qquad S^2(\mathbf{X}'\mathbf{X})^{-1} = Cov(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_k) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k, \hat{\beta}_1) & Cov(\hat{\beta}_k, \hat{\beta}_2) & \cdots & Var(\hat{\beta}_k) \end{bmatrix}$$

Standard Error - also called standard deviation; $\sqrt{Var(\hat{\beta}_m)}$

***t* Test** - use to look at single restriction on parameters

Standardized Parameter - $\hat{\beta}_m / \sqrt{Var(\hat{\beta}_m)} \sim t_{N-k}$; distributed as a t distribution with N - k degrees of freedom under the null hypothesis that $\beta_m = 0$

Hypothesis Test -

H₀: $\beta_m = 0$

H_a: $\beta_m \neq 0$

Rejection Region: if $\boxed{\left| \frac{\hat{\beta}_m}{\sqrt{Var(\hat{\beta}_m)}} \right| > t_{\alpha, N-k}}$, then reject H₀

If we can't reject we say (a) β_m is statistically insignificant; (b) regressor x_m has no statistically significant effect on y; (c) there's not enough evidence in the data to suggest there's an association between x_m and y; or (d) the m^{th} exogenous variable doesn't provide any useful information for explaining the variation in y

***p*-value** - gives level at which β_m is statistically insignificant so we don't need to go to tables to get critical value of t; if p < 0.05 (or desired level), reject H_0

$$\Pr\left[\left|t_{\alpha,N-k}\right| > \left|\frac{\hat{\beta}_m}{\sqrt{Var(\hat{\beta}_m)}}\right| > \right]$$

Parameters Equal to Other Values -

H_0: $\beta_m = a$

H_a: $\beta_m \neq a$

Rejection Region: $\left|\frac{\hat{\beta}_m - a}{\sqrt{Var(\hat{\beta}_m)}}\right| > t_{\alpha,N-k}$

Test Multiple Parameters - single test involving more than one β

H_0: $\beta_1 + \beta_2 = a$

H_a: $\beta_1 + \beta_2 \neq a$

Hard Way -

Rejection Region: $\left|\frac{\hat{\beta}_1 + \hat{\beta}_2 - a}{\sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2Cov(\hat{\beta}_1, \hat{\beta}_2)}}\right| > t_{\alpha,N-k}$

"Easy" Way - re-run the regression by enforcing the restriction:

$$y_i - ax_{1i} = x_{1i}(\beta_1 + \beta_2 - a) + (x_{2i} - x_{1i})\beta_2 + x_{3i}\beta_3 + \cdots + x_{ki}\beta_k + u_i$$

$$\tilde{y}_i = x_{1i}\tilde{\beta}_1 + (x_{2i} - x_{1i})\beta_2 + x_{3i}\beta_3 + \cdots + x_{ki}\beta_k + u_i$$

Use regular t-test to check H_0: $\tilde{\beta}_1 = 0$

Non-Linear Restriction - can't use the easy trick from before because it's not linear in $\boldsymbol{\beta}$

H_0: $\beta_1^2 + \beta_2^2 = a$

General Case - H_0: $h(\boldsymbol{\beta}) = 0$, where h is a continuous function of $\boldsymbol{\beta}$

Delta Method - $h(\hat{\boldsymbol{\beta}}) \overset{A}{\sim} N\left(0, \frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} Cov(\hat{\boldsymbol{\beta}}) \frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)$

Rejection Region: $\left|\frac{h(\hat{\boldsymbol{\beta}})}{\sqrt{\frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}'} Cov(\hat{\boldsymbol{\beta}}) \frac{\partial h(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}}}\right| > t_{\alpha,N-k}$

Problem - right hand side regressor may be scaled (e.g., per capita GDP in 1,000s of dollars)

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \cdots + x_{ki}\beta_k + u_i$$

$$y_i = x_{1i}\beta_1 + \frac{x_{2i}}{1000}(1000\beta_2) + x_{3i}\beta_3 + \cdots + x_{ki}\beta_k + u_i$$

t ratio (standardized parameter) won't change, but delta method doesn’t work

Wald Test - use to test multiple restrictions on the parameters; we might need to do that because a parameter may be insignificant individually, but significant jointly

General - $\mathbf{R\beta} = \mathbf{r}$

Examples -

(a) H$_0$: $\beta_1 = 0$, $\beta_2 = 0$ $\quad \mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

H$_a$: $\beta_1 \neq 0$ or $\beta_2 \neq 0$

(b) H$_0$: $\beta_1 + 2\beta_3 = 0$, $\beta_2 = 0$ $\quad \mathbf{R} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix}$, $\mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

H$_a$: $\beta_1 + 2\beta_3 \neq 0$ or $\beta_2 \neq 0$

$\mathbf{R\hat{\beta}} - \mathbf{r} \sim N(\mathbf{0}, \mathbf{R}Cov(\hat{\boldsymbol{\beta}})\mathbf{R}')$

Weighted Quadratic Distance - $(\mathbf{R\hat{\beta}} - \mathbf{r})'(\mathbf{R}Cov(\hat{\boldsymbol{\beta}})\mathbf{R}')^{-1}(\mathbf{R\hat{\beta}} - \mathbf{r}) \sim \chi^2_m$ (m = # restrictions)

Prof Ai doesn’t like the Wald Test... too much work

***F* Test** - use to test multiple restrictions on the parameters just like with the Wald Test; steps:

1. Start with original model

$$y_i = x_{1i}\hat{\beta}_1 + x_{2i}\hat{\beta}_2 + x_{3i}\hat{\beta}_3 + \cdots + x_{ki}\hat{\beta}_k + \hat{u}_i$$

2. Solve for a different $\hat{\beta}$ in each of the m restrictions (using $\mathbf{R}$)

(a) $\hat{\beta}_1 = \hat{\beta}_2 = 0$ $\qquad$ (b) $\hat{\beta}_1 = -2\hat{\beta}_3$ and $\hat{\beta}_2 = 0$

3. Plug into the original model

(a) $y_i = x_{1i}(0) + x_{2i}(0) + x_{3i}\hat{\beta}_3 + \cdots + x_{ki}\hat{\beta}_k + \hat{u}_i$

(b) $y_i = x_{1i}(-2\beta_3) + x_{2i}(0) + x_{3i}\beta_3 + \cdots + x_{ki}\beta_k + u_i$

4. Combine terms

(a) $y_i = x_{3i}\hat{\beta}_3 + \cdots + x_{ki}\hat{\beta}_k + \hat{u}_i$

(b) $y_i = (x_{3i} - 2x_{1i})\beta_3 + \cdots + x_{ki}\beta_k + u_i$

In general will have $\tilde{y}_i = \tilde{x}_{1i}\tilde{\beta}_1 + \cdots + \tilde{x}_{ki}\tilde{\beta}_k + \tilde{u}_i$

5. Re-run regression

Test Statistic - $\boxed{F = \dfrac{\sum_{i=1}^{N}\left(\tilde{u}_i^2 - \hat{u}_i^2\right)/m}{\sum_{i=1}^{N}\hat{u}_i^2/(N-k)} \sim F_{m,N-k}}$

Ai's Favorite - "very good finite sample properties"

Problem - assumptions 3 & 4 (no autocorrelation and homoskedasticity); F test doesn't actually use the sample variance, but it relies on these assumptions (other tests will be fine as long as we can compute the new variance)

Sum of Squares

Start with basic model: $y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{u}_i$

Now subtract means: $y_i - \bar{y} = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{u}_i - (\bar{\mathbf{x}}'\hat{\boldsymbol{\beta}} + \bar{u}) = (\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}} + \hat{u}_i$ (since $\bar{u} = 0$)

Now square it: $(y_i - \bar{y})^2 = \left((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}} + \hat{u}_i\right)^2 = \left((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}}\right)^2 + \hat{u}_i^2 + 2\left((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}}\right)\hat{u}_i$

Now add over all *i*: $\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}\left((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}}\right)^2 + \sum_{i=1}^{N}\hat{u}_i^2$

Note: $\sum_{i=1}^{N} 2\left((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}}\right)\hat{u}_i = 0$ because of assumption 1

Sum of Squared Residuals (SSR) - $\sum_{i=1}^{N}\hat{u}_i^2$; also called sum of squared error (SSE)

Sum of Squared Model (SSM) - $\sum_{i=1}^{N}\left((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\boldsymbol{\beta}}\right)^2$; also called sum of squared regression (SSR)

Sum of Squared Total (SST) - $\sum_{i=1}^{N}(y_i - \bar{y})^2$

Centered R^2 - $1 - \frac{\text{SSR}}{\text{SST}}$; how much variation in y (in % terms) is explained by the model; could be low because (a) large variation in $\hat{u}_i$, (b) bad model (wrong functional form or missing regressors)

Cross-Sectional Data - big difference among individuals so there's lots of heterogeneity in data; R^2 > 0.3 is usually pretty good

Time Series - R^2 > 0.8

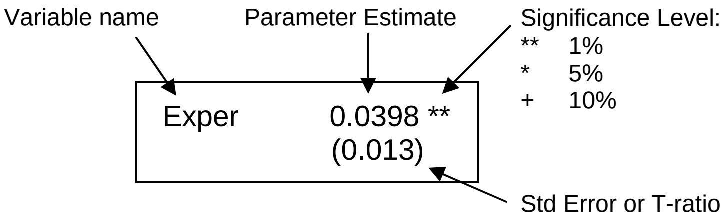
Economic Theory - R^2 is increasing function of k ($k\uparrow \Rightarrow R^2\uparrow$) and decreasing function of N ($N\uparrow \Rightarrow R^2\downarrow$); statisticians basically add and remove variables to improve R^2; econometricians don't do that because economic theory tells us which variables to include

Adjusted R^2 - $1 - \frac{\text{SSR}/(N-k)}{\text{SST}/(N-1)}$; solves problem of $k\uparrow \Rightarrow R^2\uparrow$

Uncentered R^2 - $1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$; used when we don't use a constant term

Reporting

Convention for reporting a parameter:



Failure of Assumptions

Review...

Basic Model - 3 ways to write it: k parameters; N observations

$$\boxed{y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i} \quad \text{or} \quad \boxed{y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k + u_i} \quad \text{or} \quad \boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}}$$

$$\mathbf{x}_i = \underset{k \times 1}{\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{bmatrix}} \quad \boldsymbol{\beta} = \underset{k \times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}} \quad \mathbf{Y} = \underset{N \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_k' \end{bmatrix} = \underset{N \times k}{\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{kN} \end{bmatrix}} \quad \mathbf{U} = \underset{N \times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}}$$

Estimating $\hat{\boldsymbol{\beta}}$ - there are several ways to write it out:

$$\boxed{\begin{array}{ll} \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i & \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i u_i \\ \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} & \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U} \end{array}}$$

Assumptions - fall into three categories: regressors (2), error terms (3 & 4), or both (1)

1. $E(u_i \mathbf{x}_i) = \mathbf{0}$ $\forall$ i = 1, ..., N - error uncorrelated to contemporary regressors
2. $\mathbf{X}'\mathbf{X}$ and $E(\mathbf{X}'\mathbf{X}/N)$ Nonsingular - combined with assumption 1, these are the identification or regularity conditions (i.e., we can find $\boldsymbol{\beta}$ in theory or $\hat{\boldsymbol{\beta}}$ in practice)

   1 & 2 with some technical stat stuff say $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left[\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1} E(\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\right]$
3. $E(\mathbf{x}_i \mathbf{x}_j' u_i u_j) = \mathbf{0}$ $\forall$ $i \neq j$ - error terms are unrelated to each other
4. $E(u_i^2) = \sigma^2$ - homoskedasticity (error terms have the same variance)

   3 & 4 say $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$

Assumptions on regressors and relationship between regressors and error terms (i.e., assumptions 1 & 2) are required for $\hat{\boldsymbol{\beta}}$ to be consistent

Assumptions on error terms (i.e., 3 & 4) are mainly just to simply calculations for $Var(\hat{\boldsymbol{\beta}})$

Best Estimate - As long as the four assumptions hold, $\hat{\boldsymbol{\beta}}$ is our best estimate given the data set (i.e., has the lowest variance); this is true even if we had additional info such as σ^2 = 3

Basic Proofs - almost all proofs in econometrics rely on just two things:

- Sample averages converges to population mean
- Sample average over square root of N is normally distributed (central limit theorem)

Heteroskedasticity - $Var(u_i)$ is not constant so $E(u_i^2\mathbf{x}_i\mathbf{x}_i') \neq \sigma^2 E(\mathbf{x}_i\mathbf{x}_i')$ which we used to simplify the calculations to find $Var(\hat{\boldsymbol{\beta}})$; more realistic because we wouldn't expect a big firm to have the same variation as a small firm (or big state versus small state or rich person versus poor person); since we can't simplify, we have $E(u_i^2\mathbf{x}_i\mathbf{x}_i') \neq E(\sigma_i^2\mathbf{x}_i\mathbf{x}_i')$

White Heteroskedasticity Consistent Covariance Estimator -

$$\boxed{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left[\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}\right]}$$

Note 1: $\hat{u}_i^2$ is not good estimate for σ_i^2, but $\sum_{i=1}^{N}\hat{u}_i^2$ is OK for $\sum_{i=1}^{N}\sigma_i^2$

Note 2: still have assumption 3 (no correlation between u_i and u_j)

Note 3: it's safer to use the WHCCE; t ratio and Wald Test are valid even when using WHCCE; still need to check for homoskedasticity before using F test though

Homoskedasticity - purely statistical assumption

Detecting Heteroskedasticity - $E(u_i^2 \mid \mathbf{x}_i) = \sigma_i^2$ so the variance is not constant; we don't need to know what σ_i^2 is (or its distribution) to detect that it's not constant

Informal Way -

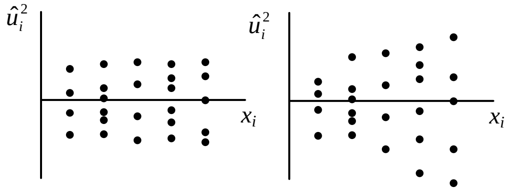
Run regression

Save residuals ($\hat{u}_i$)

Square them

Plot $\hat{u}_i^2$ against each regressor

Look for patterns



Looks good

Looks suspect ($\hat{u}_i^2$ is increasing with x_i)

Lagrange Multiplier Test - formal way

Do informal way and let $\mathbf{z}_i$ be column vector of regressors that are correlated with $\hat{u}_i^2$ (note that $\mathbf{z}_i \subset \mathbf{x}_i$)

Linear Functional Form - assume $\sigma_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} = \alpha_0 + \alpha_1 z_{1i} + \cdots + \alpha_m z_{mi}$

We want to test H$_0$: $\alpha_1 = 0$, ..., $\alpha_m = 0$ vs. H$_a$: some $\alpha_i \neq 0$ (don't care which one); this is just the F-test that's reported when we run a regression

More General - under H$_0$, we're actually testing if $\sigma_i^2 = h(\mathbf{z}_i'\boldsymbol{\alpha})$ where h is <u>any</u> function because under H$_0$, h(0) is a constant (α_0); $\therefore$ using linear functional form is fine for detecting heteroskedasticity

Generalized Least Squares - we could use the WHCCE mentioned earlier to adjust for heteroskedasticity or we could get fancy; here's the theory:

We start with the basic model: $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $\sigma_i^2 = E(u_i^2 \mid \mathbf{z}_i)$

Divide both sides by σ_i: $\dfrac{y_i}{\sigma_i} = \left(\dfrac{\mathbf{x}_i'}{\sigma_i}\right)\boldsymbol{\beta} + \dfrac{u_i}{\sigma_i} \;\Rightarrow\; \tilde{y}_i = \tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \tilde{u}_i$

This eliminates the heteroskedasticity and preserves the other assumptions

<u>Proof</u>: $Var(\tilde{u}_i^2 \mid \mathbf{z}_i) = E\left[\left(\dfrac{u_i}{\sigma_i}\right)^2 \middle| \mathbf{z}_i\right] = \dfrac{1}{\sigma_i^2}E(u_i^2 \mid \mathbf{z}_i) = \dfrac{\sigma_i^2}{\sigma_i^2} = 1$ (constant!)

GLS Estimator - $\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$ (also called Weighted Least Squares Estimator)

In Practice - sounds good, but we don't know σ_i

We could assume $\sigma_i = \sigma_0 s_i$ (that is, there's some constant variance σ_0 in all the error terms and they vary by some scalar multiple of that variance), but this is pretty risky because we could have $\sigma_i = \sigma_0 s_i + \sigma_1$, $\therefore$ we'll make a more general assumption: $\sigma_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} = \alpha_0 + \alpha_1 z_{1i} + \cdots + \alpha_m z_{mi}$

We'll run an OLS regression on $\hat{u}_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} = \alpha_0 + \alpha_1 z_{1i} + \cdots + \alpha_m z_{mi}$ and then let $\hat{\sigma}_i^2$ be the predictions from that model: $\hat{\sigma}_i^2 = \mathbf{z}_i'\hat{\boldsymbol{\alpha}} = \hat{\alpha}_0 + \hat{\alpha}_1 z_{1i} + \cdots + \hat{\alpha}_m z_{mi}$

Then we run the **feasible generalized least squares**: $\dfrac{y_i}{\hat{\sigma}_i} = \left(\dfrac{\mathbf{x}_i'}{\hat{\sigma}_i}\right)\boldsymbol{\beta} + \dfrac{u_i}{\hat{\sigma}_i}$

which will be the same as GLS for large samples

Problem - no guarantee that $\hat{\sigma}_i^2 > 0$ so we cheat: $\hat{\sigma}_i^2 = \max[0.01, \mathbf{z}_i'\hat{\boldsymbol{\alpha}}]$ (or whatever number you decide is small enough; there should only be a few observations that this is an issue for; if there are many, the functional form for $\hat{\sigma}_i^2$ may be wrong

Correlated Error Terms - we assumed $E(\mathbf{x}_i\mathbf{x}_j'u_iu_j) = \mathbf{0}$

Time Series - usually get error terms correlated sequential, hence **serial correlation**:

$u_i = \rho_1 u_{i-1} + \rho_2 u_{i-2} + \gamma_i$

Cross Section - i and (i - 1) doesn't mean anything; usually called **spatial**, **network**, or **cluster** correlation (e.g., firms next to each other; family members; groups of friends/similar interests)

Network Model - suppose M groups: G_1, G_2, ..., G_M are sets containing index of observations in each group; each group can have a different number of observations

Stata - G_i is represented by a single variable with values: 1, 1, 1, 2, 2, 2, 3, 3, etc. denoting which cluster each observation belongs to

Basic Idea - no correlation in error terms between groups, but error terms within group are correlated at a constant rate... $E(u_iu_j) = \rho$ if $i, j \in G_k$

Problem - $E(\mathbf{UU}') \neq \sigma^2\mathbf{I}$; main diagonal is still σ^2 (assuming no heteroskedasticity); problem is off diagonal terms; some are 0 (like they should be); others are ρ

$$\mathbf{V} = E(\mathbf{UU}') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1,u_N) \\ E(u_2,u_1) & E(u_2^2) & \cdots & E(u_2,u_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_N,u_1) & E(u_N,u_2) & \cdots & E(u_N^2) \end{bmatrix}$$

Estimate $\boldsymbol{\sigma}^2$ - can still use sample variance: $s^2 = \dfrac{1}{N-K}\sum_{i=1}^{N}\hat{u}_i^2$

Estimate ρ - two cases:

ρ **Different for Each Group** - $\hat{\rho}_i = \dfrac{\sum_{i,j\in G_i} \hat{u}_i \hat{u}_j}{\binom{|G_i|}{2}}$ (divided by # pairs in Gi);

need each group to have a large sample; this is what Stata uses

ρ **Same for Each Group** - $\hat{\rho} = \dfrac{\sum_{i,j\in G_1} \hat{u}_i \hat{u}_j + \sum_{i,j\in G_2} \hat{u}_i \hat{u}_j + \cdots + \sum_{i,j\in G_N} \hat{u}_i \hat{u}_j}{\text{Total \# pairs}}$

In both cases, we use σ^2 and ρ to estimate $\hat{\mathbf{V}}$

Cholesky Decomposition - $\hat{\mathbf{V}} = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}'$; hard to do in Stata

Transform Data - $\hat{\mathbf{\Gamma}}^{-1}\mathbf{Y} = \hat{\mathbf{\Gamma}}^{-1}\mathbf{X\beta} + \hat{\mathbf{\Gamma}}^{-1}\mathbf{U}$

GLS Estimator - $\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$

Potential Problem - still need $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ to be consistent; example where it's <u>not</u>:

$h_m = \beta_0 + \beta_1 w_m + \beta_2 h_f + u_m$

$h_f = \alpha_0 + \alpha_1 w_f + \alpha_2 h_m + u_f$ ... if u_m and u_f are correlated, then h_f and u_m are correlated

Time Series - many ways for error terms to be correlated:

(a) $u_i = \rho u_{i-1} + \varepsilon_i$

(b) $u_i = \rho_1 u_{i-1} + \rho_2 u_{i-2} + \varepsilon_i$

(c) $u_i = \rho_1 u_{i-1} + \rho_2 u_{i-2} + \rho_3 u_{i-4} + \varepsilon_i$ (can be any previous error term)

Problem - could lead to RHS regressor correlated with u which violates identification condition so $\hat{\boldsymbol{\beta}}$ is not consistent (e.g., $y_i = \beta_0 + \beta_1 y_{i-1} + \beta_2 x_i + u_i$ if u_i is serially correlated, it's probably correlated to y_{i-1})

Detecting -

1. supposed $\hat{\boldsymbol{\beta}}$ is consistent; estimate $\hat{u}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$
2. Run regression based on how you think error terms are correlated:
   a. $\hat{u}_i = \rho\hat{u}_{i-1} + \varepsilon_i$ ... check t-ratio for ρ
   b. $\hat{u}_i = \rho_1\hat{u}_{i-1} + \rho_2\hat{u}_{i-2} + \varepsilon_i$ ... check the F-test for all parameters jointly (use F-test for any multiple lag problem like this)

Durbin-Watson Statistic - same as case a above (i.e., first order serial correlation); very high or very low values indicate correlation present; problem with this statistic is that we don't know the distribution so we don't know the idea value

Stata - generate lagged variables: `generate lagy = y[_n-1]`

Fixing - if ρ is (are) significant modify variables:

a. $y_i - \hat{\rho}y_{i-1} = (\mathbf{x}_i - \hat{\rho}\mathbf{x}_{i-1})'\boldsymbol{\beta} + \varepsilon_i$

b. $y_i - \hat{\rho}_1 y_{i-1} - \hat{\rho}_2 y_{i-2} = (\mathbf{x}_i - \hat{\rho}_1\mathbf{x}_{i-1} - \hat{\rho}_2\mathbf{x}_{i-2})'\boldsymbol{\beta} + \varepsilon_i$

Heteroskedasticity & **Correlated Error Terms** - two problems; exact solution depends on type of correlation; assume $u_i = \rho u_{i-1} + \varepsilon_i$ and $E(\varepsilon_i^2 \mid \mathbf{x}_i) = \sigma_i^2 = \mathbf{z}_i'\boldsymbol{\alpha}$ (that that heteroskedasticity of ε_i causes heteroskedasticity of u_i); $\mathbf{z}_i$ is column vector of regressors that are correlated with $\hat{\varepsilon}_i^2$ (see heteroskedasticity section); steps:

1. Regress y_i on x_i and get $\hat{\boldsymbol{\beta}}_{\text{OLS}}$
2. Regress $\hat{u}_i^2$ on $\hat{u}_{i-1}^2$ and get $\hat{\rho}$ and $\hat{\varepsilon}_i = \hat{u}_i - \hat{\rho}\hat{u}_{i-1}$
3. Regress $\hat{\varepsilon}_i^2$ on $\mathbf{z}_i$ and get $\hat{\boldsymbol{\alpha}}$
4. Modify variables and rerun regression: $\dfrac{y_i - \hat{\rho} y_{i-1}}{\sqrt{\mathbf{z}_i'\boldsymbol{\alpha}}} = \dfrac{(\mathbf{x}_i - \hat{\rho}\mathbf{x}_{i-1})'\boldsymbol{\beta}}{\sqrt{\mathbf{z}_i'\boldsymbol{\alpha}}} + \dfrac{u_i - \hat{\rho} u_{i-1}}{\sqrt{\mathbf{z}_i'\boldsymbol{\alpha}}}$

Multicollinearity - RHS regressors are extremely correlated

Pure Multicollinearity - have redundant regressor (i.e., it's a linear combination of the other regressors); $\mathbf{X}'\mathbf{X}$ is not invertible

Stata - will automatically drop the problem regressor (and tell you)

Near Multicollinearity - $\mathbf{X}'\mathbf{X}$ is invertible but have at least one eigenvalue close to zero (should all be > 0)

Problem - $\hat{\boldsymbol{\beta}}$ will be unstable (could switch sign if we remove a regressor); hypothesis tests and interpretation of $\hat{\boldsymbol{\beta}}$ can't be trusted; no problems for forecasting though

Cause - probably have too many proxy variables for the same thing

Detecting - t-ratios are small so regressors seem insignificant; no unique rule or procedure to detect near multicollinearity

Ai's Method - regress each regressor on the other regressors; if R^2 > 0.95 we should be concerned; t-ratio tells which repressors are correlated

Solutions -

1. Increase sample size (may just have a bad sample)
2. Consider dropping problem variable... could lead to problems with economic theory; safe thing to do is re-run regression and make sure parameters of uncorrelated variables don't change... example:

   Assume we want to run: $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k + u_i$

   We run $x_{1i} = \alpha_0 + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki} + \eta_i$ and get R^2 = 0.98 and x_{4i},..., x_{ki} are significant; x_{2i} and x_{3i} aren't correlated to x_{1i}

   Drop x_{1i} and run $y_i = \tilde{\beta}_0 + x_{2i}\tilde{\beta}_2 + \cdots + x_{ki}\tilde{\beta}_k + \tilde{u}_i$, where $\tilde{\beta}_j = \beta_j + \alpha_j\beta_1$; effect of x_{1i} will be "picked up" by the new parameters; for those regressors that weren't correlated with x_{1i} we'd expect the parameter not to change much: $\tilde{\beta}_2 = \beta_2 + \alpha_2\beta_1 \approx \beta_2$ because $\alpha_2 \approx 0$ since x_{1i} and x_{2i} were not highly correlated

   Bias - $\tilde{\beta}_j$ (or β_j depending on how you look at it) could be biased if $|\alpha_j\beta_1|$ is large; don't know if it's to high or too low because we don't know β_j

Measurement Error - very common

Additive - just one type of error (easiest to deal with)

True Model - $\tilde{y} = \tilde{\mathbf{x}}'\boldsymbol{\beta} + u = \beta_0 + \beta_1\tilde{x}_1 + \cdots + \beta_{k-1}\tilde{x}_{k-1} + u$ ... note, we're leaving off the index of the observations to keep the notation simple (i.e., not writing $\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \cdots$)

Random Error - $y = \tilde{y} + v$ and $\mathbf{x} = \tilde{\mathbf{x}} + \boldsymbol{\varepsilon}$ (i.e., $x_i = \tilde{x}_i + \varepsilon_i$); we assume the error is not deliberate (i.e., it's random) so we have $E(v) = 0$, $E(\varepsilon_i) = 0$, $E(\tilde{y}v) = 0$ and $E(\tilde{x}_i\varepsilon_i) = 0$... actually we can go farther and assume none of the regressors is correlated with any of the error terms

Observed Model - solve the error equations for $\tilde{y}$ and $\tilde{\mathbf{x}}$ and substitute into the model: $y - v = (\mathbf{x} - \boldsymbol{\varepsilon})'\boldsymbol{\beta} + u$, which can be re-written: $y = \mathbf{x}'\boldsymbol{\beta} + [u + v - \boldsymbol{\varepsilon}'\boldsymbol{\beta}]$

Observed Error - $\eta = u + v - \boldsymbol{\varepsilon}'\boldsymbol{\beta}$

Problem - $E(x_i\eta) = E[x_i(u + v - \boldsymbol{\varepsilon}'\boldsymbol{\beta})] = E[(\tilde{x}_i - \varepsilon_i)(u + v - \boldsymbol{\varepsilon}'\boldsymbol{\beta})]$... even when we assume $E(\tilde{x}_i u) = 0$, $E(\tilde{x}_i v) = 0$, and $E(\tilde{x}_i\varepsilon_j) = 0$, we have $E(x_i\eta) = -E(\varepsilon_i^2)\beta_i$ ... i.e., we could have regressors correlated with the error term so $\hat{\boldsymbol{\beta}}$ may not be consistent

Effect on $\hat{\boldsymbol{\beta}}$ - $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta}$... $\mathbf{X}'\mathbf{X}$ is a positive definite matrix so the only way $\hat{\boldsymbol{\beta}}$ is unbiased is if $\mathbf{X}'\boldsymbol{\eta} = \mathbf{0}$, we'll actually look at $\boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\eta}}{N}$ (the N's cancel; we divide by N so the term converges as sample size increases [$N\uparrow$])

$$\frac{\mathbf{X}'\boldsymbol{\eta}}{N} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\eta_i \\ \frac{1}{N}\sum_{i=1}^{N}x_{1i}\eta_i \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}x_{(k-1)i}\eta_i \end{bmatrix} = \begin{bmatrix} E(\eta_i) \\ E(x_{1i}\eta_i) \\ \vdots \\ E(x_{(k-1)i}\eta_i) \end{bmatrix} = \begin{bmatrix} E(\eta_i) \\ -E(\varepsilon_1^2)\beta_1 \\ \vdots \\ -E(\varepsilon_{k-1}^2)\beta_{k-1} \end{bmatrix}$$

Case 1 - if there is no measurement error in the regressors (i.e., $\varepsilon_i = 0 \;\forall\; i$), then $\hat{\boldsymbol{\beta}}$ is still consistent (i.e., measurement error in dependent variable (y) doesn't matter)

Case 2 - only a single regressor has measurement error (e.g., $\varepsilon_i = 0 \;\forall\; i \neq 1$)... all parameter estimates are affected so $\hat{\boldsymbol{\beta}}$ is not consistent

Direction of Bias - $E(\varepsilon_1^2) > 0$ so $-E(\varepsilon_i^2)\beta_i$ has opposite sign of β_i, so all $\hat{\boldsymbol{\beta}}$ are biased toward the origin

Case 3 - two or more regressors have measurement error; still have all $\hat{\boldsymbol{\beta}}$ biased, but can't determine direction of bias

Multiplicative - $y = \tilde{y}v$ and $x_i = \tilde{x}_i\varepsilon_i$

Random Error - $E(v) = 1$, $E(\varepsilon_i) = 1$, $E(\tilde{y}v) = 0$ and $E(\tilde{x}_i\varepsilon_i) = 0$; we'll also assume the errors are independent of their corresponding variables

Observed Model - start with true model: $\tilde{y} = \tilde{\mathbf{x}}'\boldsymbol{\beta} + u = \beta_0 + \beta_1\tilde{x}_1 + \cdots + \beta_{k-1}\tilde{x}_{k-1} + u$

$\pm y$ to left sides: $y + (\tilde{y} - y) = \beta_0 + \beta_1\tilde{x}_1 + \cdots + \beta_{k-1}\tilde{x}_{k-1} + u$

$\pm\beta_i x_i$ on right:

$$y+(\tilde{y}-y)=\beta_0+\beta_1 x_1+\beta_1(\tilde{x}_1-x_1)+\cdots+\beta_{k-1}x_{k-1}+\beta_{k-1}(\tilde{x}_{k-1}-x_{k-1})+u$$

Move all () terms to end:

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_{k-1}x_{k-1}+\left[u+(y-\tilde{y})+\beta_1(\tilde{x}_1-x_1)+\cdots+\beta_{k-1}(\tilde{x}_{k-1}-x_{k-1})\right]$$

Observed Error - $\eta=u+(y-\tilde{y})+\beta_1(\tilde{x}_1-x_1)+\cdots+\beta_{k-1}(\tilde{x}_{k-1}-x_{k-1})$

Sub $y=\tilde{y}v$ and $x_i=\tilde{x}_i\varepsilon_i$: $\eta=u+(\tilde{y}v-\tilde{y})+\beta_1(\tilde{x}_1-\tilde{x}_1\varepsilon_1)+\cdots+\beta_{k-1}(\tilde{x}_{k-1}-\tilde{x}_{k-1}\varepsilon_{k-1})$

Gather terms: $\eta=u+\tilde{y}(v-1)+\beta_1\tilde{x}_1(1-\varepsilon_1)+\cdots+\beta_{k-1}\tilde{x}_{k-1}(1-\varepsilon_{k-1})$

Problem - $E(x_i\eta)=E(x_iu)+E(x_i\tilde{y}(v-1))+\beta_1E(x_i\tilde{x}_1(1-\varepsilon_1))+\cdots+\beta_{k-1}E(x_i\tilde{x}_{k-1}(1-\varepsilon_{k-1}))$

Use independence:

$$E(x_i\eta)=E(x_iu)+E(x_i\tilde{y})E(v-1)+\beta_1E(x_i\tilde{x}_1)E(1-\varepsilon_1)+\cdots+\beta_{k-1}E(x_i\tilde{x}_{k-1})E(1-\varepsilon_{k-1})$$

Case 1 - if there is no measurement error in the regressors (i.e., $\varepsilon_i=1\ \forall\ i$), then $E(x_i\eta)=E(x_iu)+E(x_i\tilde{y})E(v-1)=0$ so $\hat{\boldsymbol{\beta}}$ is still consistent (same result as additive error)... but now we have <u>heteroskedasticity</u> ($Var(u)$ depends on $Var(\tilde{y})$)

Case 2 - only a single regressor has measurement error (e.g., $\varepsilon_i=1\ \forall\ i\neq 1$), then

$E(x_1\eta)=E(x_1u)+\beta_1E(x_1\tilde{x}_1(1-\varepsilon_1))=\beta_1E(x_1\tilde{x}_1(1-\varepsilon_1))$ (assumed $E(x_iu)=0$)

Substitute $x_1=\tilde{x}_1\varepsilon_1$:

$$E(x_i\eta)=\beta_1E(\tilde{x}_1\tilde{x}_1\varepsilon_1(1-\varepsilon_1))=\beta_1E(\tilde{x}_1\tilde{x}_1e_1-\tilde{x}_1\tilde{x}_1e_1^2)=-\tilde{x}_1^2E(\varepsilon_1^2)\beta_1$$

Difference from additive is $\tilde{x}_1^2>0$ so just like before all parameter estimates are affected and $\hat{\boldsymbol{\beta}}$ is not consistent (biased toward the origin)... but now we have <u>heteroskedasticity</u> ($Var(u)$ depends on $Var(\tilde{x}_1)$)

Multiplicative Error in ln Model - $\ln\tilde{y}=\beta_0+\beta_1\ln\tilde{x}_1+\cdots+\beta_{k-1}\ln\tilde{x}_{k-1}+u$

Additive Error - $y=\tilde{y}v$ and $x_i=\tilde{x}_i\varepsilon_i$ become $\ln y=\ln\tilde{y}+\ln v$ and $\ln x_i=\ln\tilde{x}_i+\ln\varepsilon_i$ so multiplicative error in $\ln$ model becomes same as additive error (eliminates heteroskedasticity problem)

Omitted Variable Bias

True Model - $y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+u$

Observed Model - if we leave out x_k: $y=\beta_0+\beta_1x_1+\cdots+\beta_{k-1}x_{k-1}+\left[u+\beta_kx_k\right]$

Problem - error term $\eta=u+\beta_kx_k$ could be correlated with regressors (i.e., $\hat{\boldsymbol{\beta}}$ not consistent):

$E(x_i\eta)=E(x_i(u+\beta_kx_k))=E(x_iu)+\beta_kE(x_ix_k)$ ... we know $E(x_iu)=0$ by assumption, but in general $E(x_ix_k)\neq 0$

Best Case - none of the regressors are correlated so this isn't a problem

Next Best - only one regressor, say x_1, is correlated with x_k; all parameter estimates are biased and the direction depends on $E(x_1x_k)$

Solution - x_k could be missing because there's no data; easy fix is to find a proxy variable

Proxy Variable

Proxy - z is proxy variable for x_k if $E(y \mid x_1, \ldots, x_k, z) = E(y \mid x_1, \ldots, x_k)$ (i.e., z doesn't contain additional information on y); we'd rather use x_k if we had it, but z will work

Assumption - $x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \alpha_k z + \varepsilon$

Plug that into model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \beta_k (\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \alpha_k z + \varepsilon) + u$$

Gather terms:

$$y = (\beta_0 + \alpha_0 \beta_k) + (\beta_1 + \alpha_1 \beta_k) x_1 + \cdots + (\beta_{k-1} + \alpha_{k-1} \beta_k) x_{k-1} + \alpha_k \beta_k z + (u + \varepsilon \beta_k)$$

Problem - could bias all coefficients (depends on corresponding α)

Good Proxy - want $\alpha_{i \neq k} \approx 0$ (i.e., want x_k to be correlated to z only and not any of the other regressors); in this case significance test on $\alpha_k \beta_k$ is "good enough" (i.e., roughly the same) as test on β_k

Multiple Proxies - $x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \alpha_k z_1 + + \alpha_{k+1} z_2 + \varepsilon$

Model becomes:

$$y = (\beta_0 + \alpha_0 \beta_k) + (\beta_1 + \alpha_1 \beta_k) x_1 + \cdots + (\beta_{k-1} + \alpha_{k-1} \beta_k) x_{k-1} + \alpha_k \beta_k z_1 + + \alpha_{k+1} \beta_k z_2 + (u + \varepsilon \beta_k)$$

Now do <u>joint test</u> on parameters for z_1 and z_2 to test if β_k is significant

Problem - since z_1 and z_2 are (hopefully) highly correlated to x_k, they're probably correlated to each other so we could have near multicollinearity

Redundant Regressors

No effect on $\hat{\boldsymbol{\beta}}$ (i.e., still consistent)

Problems - lose efficiency, could have near multicollinearity, more likely to have regressor correlated to error term

Restricted Regression

If there are restrictions on the parameters, we can enforce them in the regression by:

1. Run unrestricted regression and compute $\hat{u}_i$
2. Solve the for as many parameters as there are restrictions
3. Substitute these into the original model
4. Collect terms; terms that do not have a parameter are moved to the left hand side
5. Run restricted regression and compute $\tilde{u}_i$
6. New F-test: $\dfrac{\left(\sum \tilde{u}_i^2 - \sum \hat{u}_i^2\right)/2}{\sum \hat{u}_i^2 / (N-4)} \sim F_{2,N-4}$

Example -

Model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$

Restrictions: (1) $\beta_1 + \beta_2 + \beta_3 = 1$

(2) $\beta_2 - \beta_3 = 2$

<u>Step 1</u>: solve (2) for β_2: $\beta_2 = 2 + \beta_3$

Substitute that into (1): $\beta_1+(2+\beta_3)+\beta_3=1$

Solve for β_1: $\beta_1=-1-2\beta_3$

Step 2: solve these into original model: $y_i=\beta_0+(-1-2\beta_3)x_{1i}+(2+\beta_3)x_{2i}+\beta_3 x_{3i}+u_i$

Step 3: collect terms: $\underbrace{y_i+x_{1i}-2x_{2i}}_{\tilde{y}_i}=\beta_0+\beta_3(\underbrace{-2x_{1i}+x_{2i}+x_{3i}}_{\tilde{x}_i})+u_i$

Developing the Restrictions - trans-log cost function requires $C(Q,KP_1,KP_2)=KC(Q,P_1,P_2)$

$$\ln C(Q,P_1,P_2)=\beta_0+\beta_1\ln P_1+\beta_2\ln P_2+r_1\ln Q+\lambda_{11}(\ln P_1)^2++\lambda_{12}\ln P_1\ln P_2+\lambda_{22}(\ln P_2)^2+\delta_0(\ln Q)^2+\delta_1\ln Q\ln P_1+\delta_2\ln Q\ln P_2$$

$$\ln C(Q,KP_1,KP_2)=\beta_0+\beta_1\ln K+\beta_1\ln P_1+\beta_2\ln K+\beta_2\ln P_2+r_1\ln Q+\lambda_{11}(\ln K)^2+2\lambda_{11}\ln K\ln P_1+\lambda_{11}(\ln P_1)^2+\lambda_{12}(\ln K)^2+\lambda_{12}\ln K\ln P_1+\lambda_{12}\ln K\ln P_2+\lambda_{12}\ln P_1\ln P_2+\lambda_{22}(\ln K)^2+\lambda_{22}\ln K\ln P_2+\lambda_{22}(\ln P_2)^2+\delta_0(\ln Q)^2+\delta_1\ln Q\ln K+\delta_1\ln Q\ln P_1+\delta_2\ln Q\ln K+\delta_2\ln Q\ln P_2$$

Restriction: $\ln C(Q,KP_1,KP_2)=\ln K+\ln C(Q,P_1,P_2)$

Look at terms that cancel

$$\ln C(Q,P_1,P_2)=\cancel{\beta_0}+\cancel{\beta_1\ln P_1}+\cancel{\beta_2\ln P_2}+\cancel{r_1\ln Q}+\cancel{\lambda_{11}(\ln P_1)}^2++\cancel{\lambda_{12}\ln P_1\ln P_2}+\cancel{\lambda_{22}(\ln P_2)}^2+\cancel{\delta_0(\ln Q)}^2+\cancel{\delta_1\ln Q\ln P_1}+\cancel{\delta_2\ln Q\ln P_2}$$

$$\ln C(Q,KP_1,KP_2)=\cancel{\beta_0}+\beta_1\ln K+\cancel{\beta_1\ln P_1}+\beta_2\ln K+\cancel{\beta_2\ln P_2}+\cancel{r_1\ln Q}+\lambda_{11}(\ln K)^2+2\lambda_{11}\ln K\ln P_1+\cancel{\lambda_{11}(\ln P_1)}^2+\lambda_{12}(\ln K)^2+\lambda_{12}\ln K\ln P_1+\lambda_{12}\ln K\ln P_2+\cancel{\lambda_{12}\ln P_1\ln P_2}+\lambda_{22}(\ln K)^2+\lambda_{22}\ln K\ln P_2+\cancel{\lambda_{22}(\ln P_2)}^2+\cancel{\delta_0(\ln Q)}^2+\delta_1\ln Q\ln K+\cancel{\delta_1\ln Q\ln P_1}+\delta_2\ln Q\ln K+\cancel{\delta_2\ln Q\ln P_2}$$

That means

$$\boxed{\beta_1\ln K}+\boxed{\beta_2\ln K}+\boxed{\lambda_{11}(\ln K)^2}+\boxed{2\lambda_{11}\ln K\ln P_1}+\boxed{\lambda_{12}(\ln K)^2}+\boxed{\lambda_{12}\ln K\ln P_1}+\boxed{\lambda_{12}\ln K\ln P_2}+\boxed{\lambda_{22}(\ln K)^2}+\boxed{\lambda_{22}\ln K\ln P_2}+\delta_1\ln Q\ln K+\delta_2\ln Q\ln K=\ln K$$

$$\left.\begin{array}{lll}\text{Collect } \ln K \text{ terms:} & \beta_1+\beta_2=1\\ \text{Collect } (\ln K)^2 \text{ terms:} & \lambda_{11}+\lambda_{12}+\lambda_{22}=0\\ \text{Collect } \ln K\ln P_1 \text{ terms:} & \lambda_{11}+\lambda_{12}=0\\ \text{Collect } \ln K\ln P_2 \text{ terms:} & \lambda_{12}+\lambda_{22}=0\\ \text{Collect } \ln Q\ln K \text{ terms:} & \delta_1+\delta_2=0\end{array}\right\}\text{5 restrictions}$$

Regressors Correlated with Error Terms - $E(u_i\mathbf{x}_i) \neq \mathbf{0}$ for some i

Detecting - how do we know if x_i is correlated with u; rule of thumb

Simultaneous Decision (from Economic Theory) - think about LHS variable and RHS variable jointly determined by individual (or household)

Example - $h_f = \alpha_0 + \alpha_1 h_h + \alpha_2 w_f + \alpha_3 w_h + u$... may be maximizing joint household utility function so h_f and h_h are correlated... that means h_h and u are correlated

Example - $D_{\text{chicken}} = \alpha_0 + \alpha_1 D_{\text{beef}} + \alpha_2 P + \alpha_3 I$... demand for chicken and beef determined jointed because they're substitutes so it's likely that D_{beef} and u are correlated

Example - $S_1 = \alpha_0 + \alpha_1 S_2 + u_1$ and $S_2 = \beta_0 + \beta_1 S_1 + u_2$... firm 1 can't select S_2, but it can affect it by changing S_1

Omitted Variable - ColGPA = α_0 + α_1Attrte + α_2HSGPA + u... CollGPA also depends on Ability (unobserved variable) which is also correlated to HSGPA $\therefore$ HSGPA could be correlated with u

Constraint - LHS and RHS related by constraint

Example - $D_i = \alpha_0 + \alpha_1 P_i + u_i$... $D_i = S_i$ (supply and demand)

Example - $S_i = \beta_0 + \beta_1 P_i + u_i$... firms select S_i and P_i determined

Consequence -

Theoretical - $E(u_i\mathbf{x}_i) \neq \mathbf{0}$... don't have k equations to solve for k unknowns in $\boldsymbol{\beta}$

Practical - $E(u_i\mathbf{x}_i) \neq \mathbf{0} \Rightarrow \frac{1}{N}\sum_{i=1}^{N} x_{ji}u_i \neq 0$ (for some j) $\Rightarrow \frac{1}{N}\sum_{i=1}^{N} x_{ji}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \neq 0$... that's using the true value of $\boldsymbol{\beta}$... but our formula for $\hat{\boldsymbol{\beta}}$ imposes $\frac{1}{N}\sum_{i=1}^{N} x_{ji}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = 0$ so we'll end up with $\hat{\boldsymbol{\beta}}$ being <u>biased</u>

Correction - use instrumental variable

Testing - haven't covered yet

Instrumental Variable (IV) Estimations

Simple Case - only 1 variable correlated to the error term:

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$, with $E(x_{1i}u_i) = \cdots = E(x_{k-1,i}u_i) = 0$ and $E(x_{ki}u_i) \neq 0$

Problem - we find $\hat{\boldsymbol{\beta}}$ (k unknowns) by solving k equations in $E(\mathbf{x}_i u_i) = \mathbf{0}$, but in this case we only have k-1 equations because $E(x_{ki}u_i) \neq 0$

Goal - want to find a variable, w_i, that is correlated to x_k, but not correlated to u_i:

$E(w_i x_{ki}) \neq 0$ and $E(w_i u_i) = 0$

How to Do It - define $\mathbf{z}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{k-1,i} \\ w_i \end{bmatrix}$, now use $E(\mathbf{z}_i u_i) = \mathbf{0}$ to get estimate for $\boldsymbol{\beta}$

Sub $u_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$: $E(\mathbf{z}_i u_i) = E\big(\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\big) = \mathbf{0}$

Multiply it out and move $E(\mathbf{z}_i y_i)$ to other side: $E(\mathbf{z}_i \mathbf{x}_i')\boldsymbol{\beta} = E(\mathbf{z}_i y_i)$

Solve for $\boldsymbol{\beta}$: $\boldsymbol{\beta} = \left(E(\mathbf{z}_i \mathbf{x}_i')\right)^{-1} E(\mathbf{z}_i y_i)$

Making it Practical - now it's possible to find $\boldsymbol{\beta}$ (in theory), but we don't know expected values so we have to substitute sample averages:

Instrumental Variable Estimator - $\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i y_i = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$

Note the difference from OLS: $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$

CAUTION - we're not regressing y_i on $\mathbf{z}_i$: $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{k-1} x_{k-1,i} + \beta_k w_i + u_i$ ...

that would give us $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i \neq \hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ ... <u>not the same thing</u>

Instrumental Variable - now ready for official definition; w_i is an instrumental variable for x_{ki} if the following hold:

1) $E(w_i x_{ki}) \neq 0$ and $E(w_i u_i) = 0$ (i.e., w_i correlated to x_{ki}, but not to u_i)
2) $E(\mathbf{z}_i \mathbf{x}_i')$ is nonsingular (required for theory to identify $\boldsymbol{\beta}$)
3) $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'$ is nonsingular (required in practice to calculate $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$)

Multiple Variables with Single Instruments - now look at same example with $E(x_{k-1,i} u_i) \neq 0$ and $E(x_{ki} u_i) \neq 0$ so we have two variables correlated to the error term and assume we have two instruments, w_{1i} and w_{2i}

Not Assigned - the IVs are not assigned to the correlated variables; each IV could be correlated to one or both of the regressors; for this section we're only concerned with the fact that we have the same number of instruments as we do regressors correlated to the error term (later we'll look at having more IVs... fewer IVs is not possible)

Same Solution - still use $\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$, using $\mathbf{z}_i' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & w_{1i} & w_{2i} \end{bmatrix}$

All Regressors - this works even if all regressors are correlated: $E(x_{1i} u_i) \neq 0$, $E(x_{2i} u_i) \neq 0$, ..., $E(x_{ki} u_i) \neq 0$; use same $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ with $\mathbf{z}_i' = \begin{bmatrix} w_{1i} & w_{2i} & \cdots & w_{ki} \end{bmatrix}$

Stata - `ivreg y1 x1 x2 (x3 = z1 z2 z3)`; can use `robust` to correct for heteroskedasticity like before; if no heteroskedasticity and no correlated error terms, then standard error calculated the same way as OLS

What's a Good Instrument? Need to determine if w_i independently explains x_i

Bad Instrument - if correlation between w_i and x_i is week, result is $\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i{}'$ being "near singular".... like having near multicollinearity

Detecting - 2 methods

(1) regress each problem regressor (i.e., those correlated to the error term) against all remaining regressors and all IVs (and a constant term if not already there)

Example - from 1 variable example, regress x_{ki} on x_{1i}, x_{2i}, ..., $x_{k\text{-}1,i}$, w_i

Example - from 2 variable example, regress x_{ki} on x_{1i}, x_{2i}, ..., $x_{k\text{-}2,i}$, w_{1i}, w_{2i} and regress $x_{k\text{-}1,i}$ on x_{1i}, x_{2i}, ..., $x_{k\text{-}2,i}$, w_{1i}, w_{2i}

3 Checks - (a) $R^2 < 0.1$ means weak instrument

(b) coefficient on w_i is not significant means weak instrument

(c) coefficient on w_i is "too small" (even if significant) means weak instrument... "too small" depends on units of x_i and w_i

(2) better method... (given as example with x_{1i} being correlated with error term)

(a) regress x_{1i} on 1, x_{2i}, x_{3i} and save residuals in r_{1i}

(b) regress w_i on 1, x_{2i}, x_{3i} and save residuals in r_{2i}

(c) regress r_{1i} on r_{2i} ... $R^2 < 0.1$ means weak instrument

Multiple Instruments - having more instruments than there are regressors correlated to the error terms

Single Regressor - use same model with only 1 variable correlated: $E(x_{ki}u_i) \neq 0$, but this time assume we have 2 instruments: w_{1i} and w_{2i}

Problem - $\mathbf{z}_i{}' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & w_{1i} & w_{2i} \end{bmatrix}$ is (k+1)x1 so $\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i{}'$ is not a square matrix (it's (k+1)xk)... can't be inverted to calculate $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$

Solution - define $\mathbf{v}_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{k-1,i} \\ \mathbf{z}_i{}'\boldsymbol{\delta} \end{bmatrix}$, where $\mathbf{z}_i{}'\boldsymbol{\delta} = \delta_1 x_{1i} + \cdots + \delta_{k-1}x_{k-1,i} + \delta_k w_{1i} + \delta_{k+1}w_{2i}$

Want instrument ($\mathbf{z}_i{}'\boldsymbol{\delta}$) to be highly correlated to x_{ki} so we want

$\min_{\boldsymbol{\delta}} \sum_{i=1}^{N}(x_{ki} - \mathbf{z}_i{}'\boldsymbol{\delta})^2$ ... that means $\boldsymbol{\delta}$ is least squares estimate:

$$\hat{\boldsymbol{\delta}} = \left(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i{}'\right)^{-1}\sum_{i=1}^{N}\mathbf{z}_i x_{ki}$$

Two-Stage Least Squares - we can estimate $\hat{x}_{ki} = \mathbf{z}_i{}'\hat{\boldsymbol{\delta}}$ and use

$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i\mathbf{x}_i{}'\right)^{-1}\sum_{i=1}^{N}\hat{\mathbf{x}}_i y_i$ ... where $\hat{\mathbf{x}}_i{}' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & \hat{x}_{ki} \end{bmatrix}$; special property (we won't prove) says $\sum_{i=1}^{N}\hat{\mathbf{x}}_i\mathbf{x}_i' = \sum_{i=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i'$ so we have $\hat{\boldsymbol{\beta}}_{\text{2SLS}} = \left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i'\right)^{-1}\sum_{i=1}^{N}\hat{\mathbf{x}}_i y_i$ ... that's OLS for regressing y_i on $\hat{\mathbf{x}}_i$

In Practice -

1) Run each regressor that is correlated to the error term on the entire IV list (i.e., $\mathbf{z}_i$) and save fitted values to generate $\hat{\mathbf{x}}_i$
2) Run y_i on $\hat{\mathbf{x}}_i$

Example - consider 2 regressors correlated to error term: $E(x_{k-1,i}u_i) \neq 0$ and $E(x_{ki}u_i) \neq 0$; and four IVs: w_{1i}, w_{2i}, w_{3i}, and w_{4i}

1) Regress $x_{k-1,i}$ on $x_{1i}, x_{2i}, \ldots, x_{k-2,i}, w_{1i}, w_{2i}, w_{3i}, w_{4i}$ and generate $\hat{x}_{k-1,i}$

   Regress $x_{k,i}$ on $x_{1i}, x_{2i}, \ldots, x_{k-2,i}, w_{1i}, w_{2i}, w_{3i}, w_{4i}$ and generate $\hat{x}_{k,i}$
2) Regress y_i on $x_{1i}, x_{2i}, \ldots, x_{k-2,i}, \hat{x}_{k-1,i}, \hat{x}_{ki}$

Note1: if there is one instrument for each correlated regressor, then $\hat{\boldsymbol{\beta}}_{\text{2SLS}} = \hat{\boldsymbol{\beta}}_{\text{IV}}$

Note2: this procedure gives the correct $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$, but the wrong standard error... Stata does it right

Stata - `ivreg y x1 x2 ... xk-2 (xk-1 xk = w1 w2 w3 w4)`

Testing Parameters - $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ is best estimator when some regressors are correlated to the error term, but other assumptions hold (no heteroskedasticity, no correlated error terms); Wald and *t* tests are OK, but *F*-test is not valid; the *F*-test is based on regression residuals being orthogonal to the regressors, but for $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$, the residuals are orthogonal to $\hat{\mathbf{x}}_i$, not necessarily $\mathbf{x}_i$

Finding Instruments - ad hoc rules... didn't cover them yet

Testing Instruments - in order to have a specified model (i.e., same number of equations [$E(x_i u) = 0$] and unknowns [parameters]), we must have at least 1 valid instrument for each regressor that's correlated to the error term; this test only looks at additional instruments

Hausman Test - used to test the additional instruments

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$ and $E(x_{3i}u_i) \neq 0$

Suppose w_{1i} and w_{2i} are instruments and w_{2i} is known to be a good instrument

H_0: $E(w_1 u) = 0$ and H_1: $E(w_1 u) \neq 0$

Define $\hat{\boldsymbol{\beta}}$ as the 2SLS estimator using both w_{1i} and w_{2i} for x_{3i}

Define $\tilde{\boldsymbol{\beta}}$ as the 2SLS estimator using only w_{2i} for x_{3i}

Under H_0, both $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent so $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \mathbf{0}$; if H_0 fails $\hat{\boldsymbol{\beta}}$ is biased

Test Statistic - $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' Cov(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \sim \chi_k^2$ (*k* = # coefficients)

Example - `reg hours kidslt6 educ wage hushrs faminc unem, robust`

`hushrs` (husband hours) is probably a joint decision when deciding the wife's hours (`hours`), so it's probably correlated with the error term

Suppose `huseduc` is known to be a good instrument; test if `huswage` is also a good instrument:

```
ivreg hours kidslt6 educ wage famine unem (hushrs = huseduc),
   robust
hausman, save
ivreg hours kidslt6 educ wage famine unem (hushrs = huswage
   huseduc), robust
hausman
```

Other Uses of Hausman Test

Endogeneity - used Hausman Test earlier to test additional instruments, but can also use it to test endogeneity (i.e., test whether regressor is correlated with the error term)

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$ and $E(x_{3i}u_i) \neq 0$

Suppose z_i is an instrument

History - technically was first discovered by Durbin, then rediscovered by Hausman and Wu (independently in 1950s)... called Hausman test because he's from MIT

H_0: $E(x_{3i}u_i) = 0$ and H_1: $E(x_{3i}u_i) \neq 0$

Two Estimators - Hausman tests requires two estimators ($\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$) such that

(1) Both are consistent under H_0 and one is best

(2) Under H_1, only one is consistent (i.e., other is biased under H_1)

Idea - if H_0 is true, $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \mathbf{0}$ ... technically, it converges to zero (order doesn't matter)

Quadratic Distance - like Wald Test, measure distance between the vectors with

$(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' Cov(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \sim \chi^2_k$ (k = rank of Cov matrix)

If this distance is "close to zero" then there's no evidence to reject H_0

Rank Issues - Usually rank of Cov matrix = # parameters; if rank < # parameters then we can't take the inverse so use generalized inverse (didn't cover this in class, but Ai said software packages will do it automatically)

Order Matters - order doesn't matter for differences as long as they're the same (these are being squared so negative goes away), but order does matter for Cov term...

$\tilde{\boldsymbol{\beta}}$ is the inefficient estimator (i.e., not best under H_0)

Cov Matrix - $Cov(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = Cov(\hat{\boldsymbol{\beta}}) + Cov(\tilde{\boldsymbol{\beta}}) - 2Cov(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$ ... this is an asymptotic result so it's not guaranteed for small samples; first two terms come directly from the regressions used to estimate the parameters; the third term is complicated

Hausman's Trick - by imposing condition that $\hat{\boldsymbol{\beta}}$ is best under H_0, we can use

$Cov(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}})$

$\therefore$ test statistic becomes: $\boxed{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' \left[Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}})\right]^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \sim \chi^2_k}$

From Example - $\hat{\boldsymbol{\beta}}$ is OLS estimator; $\tilde{\boldsymbol{\beta}}$ is 2SLS (IV) estimator with z_i as instrument for x_{3i}

Problem 1 - if heteroskedasticity, $\hat{\boldsymbol{\beta}}$ is not best so $Cov(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}) \neq Cov(\tilde{\boldsymbol{\beta}})-Cov(\hat{\boldsymbol{\beta}})$

Problem 2 - if $Cov(\tilde{\boldsymbol{\beta}})-Cov(\hat{\boldsymbol{\beta}})<\mathbf{0}$ could be several reasons: (a) heteroskedasticity or some other failure that makes $\hat{\boldsymbol{\beta}}$ not best; (b) sample problem (not big enough or just a "bad" sample

Heteroskedasticity - $y_i = \mathbf{x}_i'\boldsymbol{\beta}+u_i$ with $E(u_i^2 \mid \mathbf{x}_i)=\mathbf{x}_i'\boldsymbol{\alpha}$; define estimators as $\hat{\boldsymbol{\beta}}$ is GLS estimator and $\tilde{\boldsymbol{\beta}}$ is OLS estimator... both are consistent under H_0 and H_1 ∴ <u>cannot</u> use Hausman test... heteroskedasticity affects computation of variance of estimates, it doesn't affect the fact that they are consistent

Serial Correlation - $y_i = \mathbf{x}_i'\boldsymbol{\beta}+u_i$ with $u_i = \rho u_{i-1}+\varepsilon_i$; H_0: $\rho = 0$, H_1: $\rho \neq 0$; define estimators as $\hat{\boldsymbol{\beta}}$ is OLS estimator and $\tilde{\boldsymbol{\beta}}$ is GLS estimator... both are consistent under H_0 and H_1 (as long as y_{i-1} is not a regressor) ∴ <u>cannot</u> use Hausman test

Exception - if we change model it include lagged dependent variable (e.g., $y_i = \beta_0+\beta_1 y_{i-1}+\beta_2 x_i+u_i$), then Hausman test is valid using $\hat{\boldsymbol{\beta}}$ is OLS estimator and $\tilde{\boldsymbol{\beta}}$ is 2SLS estimator using x_{i-1} as instrument for y_{i-1}; under H_0, $\hat{\boldsymbol{\beta}}$ is consistent and best and $\tilde{\boldsymbol{\beta}}$ is consistent (but inefficient); under H_1, $\hat{\boldsymbol{\beta}}$ is not consistent and $\tilde{\boldsymbol{\beta}}$ is

Coefficients - $y_i = \beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+u_i$; H_0: $\beta_2 = 0$ and $\beta_3 = 0$, H1: $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Define $\tilde{\boldsymbol{\beta}}$ as unrestricted OLS and $\hat{\beta}_1$ comes from the restricted OLS (i.e., $y_i = \beta_1 x_{1i}+u_i$)

so $\hat{\boldsymbol{\beta}}$ is $\begin{bmatrix}\hat{\beta}_1\\0\\0\end{bmatrix}$; now use $Cov(\hat{\boldsymbol{\beta}}) = \begin{bmatrix}Var(\hat{\beta}_1) & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$... Hausman Test works!

***F*-Test** - can do this same test with an *F*-test... it's better (and easier)

Limited Information

$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1+\mathbf{x}_{2i}'\boldsymbol{\beta}_2+u_i$ with $E(\mathbf{x}_{1i}u_i)=\mathbf{0}$ and $E(\mathbf{x}_{2i}u_i)\neq\mathbf{0}$ (i.e., $\mathbf{x}_{2i}$ are regressors correlated with error term)

Need instrument for each regressor; suppose: $\mathbf{x}_{2i} = \boldsymbol{\alpha}_1 y_i+\mathbf{x}_{1i}'\boldsymbol{\alpha}_2+\mathbf{z}_i'\boldsymbol{\alpha}_3+\varepsilon_i$

Practice - economic theory may specify y_i, but it doesn't care about $\mathbf{x}_{2i}$ (i.e., partial [one market] equilibrium vs. total equilibrium)

Structural Equations - y_i and $\mathbf{x}_{2i}$ for system of simultaneous equations... full information

Reduced Form - sub y_i into $\mathbf{x}_{2i}$ to get $\mathbf{x}_{2i} = \mathbf{x}_{1i}'\boldsymbol{\pi}_1+\mathbf{z}_i'\boldsymbol{\pi}_2+v_i$... limited information; run 2SLS using $\mathbf{x}_{2i} = \mathbf{x}_{1i}'\boldsymbol{\pi}_1+\mathbf{z}_i'\boldsymbol{\pi}_2+v_i$ for first stage and plugging that into y_i for second stage; given limited information (and homoskedasticity), 2SLS is best estimator

Maximum Likelihood - could also estimate jointly if we assume $(u_i, v_i) \sim N(0, \mathbf{\Omega})$ (multivariate normal).... do limited information maximum likelihood estimator (LIML); it's also best

2SLS - is just as good as LIML (i.e., best), but "better" because it's less complicated and we don't need to include all instruments

Two-Stage Least Squares

Refresher - $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$

1) Run regression to estimate error term and determine which regressors are correlated to the error term
2) Create vector comprised of all regressors that are NOT correlated to the error term and all the instrumental variables (Note: must have at least 1 instrument for each regressor that was not included [i.e., correlated to the error term]; must include constant term too)

   e.g., $\mathbf{z}_i' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & w_{1i} & w_{2i} \end{bmatrix}$ (only x_{ki} is correlated and there are 2 instruments)
3) **1st Stage** - estimate each correlated regressor by regressing it on $\mathbf{z}_i$

   e.g., $\hat{x}_{ki} = \mathbf{z}_i'\hat{\boldsymbol{\delta}}$

   Define $\hat{\mathbf{x}}_i$ as vector of uncorrelated regressors and estimated values for correlated regressors

   e.g., $\hat{\mathbf{x}}_i' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & \hat{x}_{ki} \end{bmatrix}$
4) **2nd Stage** - Regress y_i on $\hat{\mathbf{x}}_i$

F-Test - need to modify the F-test to work with 2SLS

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) \neq 0$ and $E(x_{2i}u_i) = E(x_{3i}u_i) = 0$ and w_i is instrument for x_{1i}

Two restrictions: H$_0$: $\beta_1 = 0$ and $\beta_2 = 0$

Define $\mathbf{z}_i' = \begin{bmatrix} w_i & x_{2i} & x_{3i} \end{bmatrix}$

1st Stage - regress x_{1i} on $\mathbf{z}_i$ and generate $\hat{x}_{1i}$ (can repeat for other regressors, but they'll be perfect fits so we can skip that step and just use $\hat{x}_{2i} = x_{2i}$ and $\hat{x}_{3i} = x_{3i}$)

Behind the Scenes - we're really assuming $x_{1i} = \mathbf{z}_i'\boldsymbol{\delta}_1 + \boldsymbol{\gamma}_{1i}$, $x_{2i} = \mathbf{z}_i'\boldsymbol{\delta}_2 + \boldsymbol{\gamma}_{2i}$, and $x_{3i} = \mathbf{z}_i'\boldsymbol{\delta}_3 + \boldsymbol{\gamma}_{3i}$... when we sub those into the original equation we get:

$$y_i = \beta_1(\mathbf{z}_i'\boldsymbol{\delta}_1) + \beta_2(\mathbf{z}_i'\boldsymbol{\delta}_3) + \beta_3(\mathbf{z}_i'\boldsymbol{\delta}_3) + \left[u_i + \beta_1\boldsymbol{\gamma}_{1i} + \beta_2\boldsymbol{\gamma}_{2i} + \beta_3\boldsymbol{\gamma}_{3i}\right]$$

so the error term when we go to the second stage is not actually estimating u_i, but that long thing in the brackets... that's why the F-test needs to be modified

2nd Stage - regress y_i on $\hat{\mathbf{x}}_i$ to get 2SLS estimators $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ and the residuals $\hat{u}_i = y_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1\hat{x}_{1i} - \hat{\beta}_2\hat{x}_{2i} - \hat{\beta}_3\hat{x}_{3i}$ (which we just showed are not consistent so we have to get a consistent estimate for u_i: $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

Restricted Regression - regress y_i on $\hat{x}_{3i}$ (recall restriction above says $\beta_1 = 0$ & $\beta_2 = 0$); generate residuals $\tilde{u}_i = y_i - \tilde{\beta}_3 \hat{x}_{3i}$

F-Statistic - $F = \dfrac{\left(\sum \tilde{u}_i^2 - \sum \hat{u}_i^2\right)/m}{\sum e_i^2 / n-k}$, as before, *m* = # restrictions (2 in this case), *n* = # observations and *k* = # parameters in full model (3 in this case)

Heteroskedasticity in 2SLS - since we're running 2SLS, we know $E(u_i \mid \mathbf{x}_i) \neq 0$, but $E(u_i \mid \mathbf{z}_i) = 0$; we should have $E(u_i^2 \mid \mathbf{z}_i) = \sigma^2$, but if we don't, there's heteroskedasticity (i.e., $E(u_i^2 \mid \mathbf{z}_i) \neq \sigma^2$)

Detecting - (1) run 2SLS and get $\hat{\boldsymbol{\beta}}$;

(2) compute <u>consistent</u> residuals: $e_i = y_i - \mathbf{x}_i{}'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

(3) regress e_i^2 on $(1 \ \mathbf{z}_i)$ (i.e., be sure to include a constant term if it's not already in $\mathbf{z}_i$)

(4) do overall test of significance (i.e., standard *F*-test to check if all parameters are simultaneously equal to zero)... if regression is significant, there's heteroskedasticity

Correcting - (1) save fitted value of $\hat{e}_i^2$ (from regression in step (3) above)

(2) transform model: $\dfrac{y_i}{\hat{e}_i} = \dfrac{\mathbf{x}_i{}'}{\hat{e}_i}\boldsymbol{\beta} + \dfrac{u_i}{\hat{e}_i} = \beta_1 \dfrac{x_{1i}}{\hat{e}_i} + \hat{\beta}_2 \dfrac{x_{2i}}{\hat{e}_i} + \hat{\beta}_3 \dfrac{x_{3i}}{\hat{e}_i} + \dfrac{u_i}{\hat{e}_i}$

(3) do 2SLS on the transformed model; can use $\mathbf{z}_i = \begin{bmatrix} w_i \\ x_{2i} \\ x_{3i} \end{bmatrix}$ or $\mathbf{z}_i = \begin{bmatrix} w_i / \hat{e}_i \\ x_{2i} / \hat{e}_i \\ x_{3i} / \hat{e}_i \end{bmatrix}$... will give different results, but both have same statistical properties

Serial Correlation in 2SLS -

Detecting - same as before (e.g., use $e_i = \rho e_{i-1} + \boldsymbol{\gamma}_i$ to estimate $\hat{\rho}$; if it's significantly different than zero, there's serial correlation)

Correcting -

(1) transform model:

$$(y_i - \hat{\rho} y_{i-1}) = \beta_1 (x_{1i} - \hat{\rho} x_{1i-1}) + \beta_2 (x_{2i} - \hat{\rho} x_{2i-1}) + \beta_3 (x_{3i} - \hat{\rho} x_{3i-1}) + (u_i - \hat{\rho} u_{i-1})$$

(2) do 2SLS on the transformed model; can use $\mathbf{z}_i - \hat{\rho}\mathbf{z}_{i-1}$, $\mathbf{z}_i$, or $\mathbf{z}_{i-1}$ (will have same statistical properties)

Another Test for Endogeneity

Endogeneity - Hausman Test not valid with heteroskedasticity so here's another way to test for endogeneity (i.e., test whether regressor is correlated with the error term)

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$ and $E(x_{3i}u_i) \neq 0$

Suppose z_i is an instrument

1. Estimate reduced for x_{3i}: $x_{3i} = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 z_i + \varepsilon_i$
2. Get predictions for error term: $\hat{\varepsilon}_i = x_{3i} - \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 z_i$
3. Plug prediction into original model: $y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 \hat{\varepsilon}_i + u_i$
4. H$_0$: x_{3i} is exogenous is equivalent to H$_0$: $\alpha_3 = 0$

Regressors Correlated with Error Terms - $E(u_i \mathbf{x}_i) \neq \mathbf{0}$ for some i

Detecting - how do we know if x_i is correlated with u; rule of thumb

Simultaneous Decision (from Economic Theory) - think about LHS variable and RHS variable jointly determined by individual (or household)

Example - $h_f = \alpha_0 + \alpha_1 h_h + \alpha_2 w_f + \alpha_3 w_h + u$... may be maximizing joint household utility function so h_f and h_h are correlated... that means h_h and u are correlated

Example - $D_{\text{chicken}} = \alpha_0 + \alpha_1 D_{\text{beef}} + \alpha_2 P + \alpha_3 I$... demand for chicken and beef determined jointed because they're substitutes so it's likely that D_{beef} and u are correlated

Example - $S_1 = \alpha_0 + \alpha_1 S_2 + u_1$ and $S_2 = \beta_0 + \beta_1 S_1 + u_2$... firm 1 can't select S_2, but it can affect it by changing S_1

Omitted Variable - ColGPA = $\alpha_0 + \alpha_1$Attrte + α_2HSGPA + u... CollGPA also depends on Ability (unobserved variable) which is also correlated to HSGPA ∴ HSGPA could be correlated with u

Constraint - LHS and RHS related by constraint

Example - $D_i = \alpha_0 + \alpha_1 P_i + u_i$... $D_i = S_i$ (supply and demand)

Example - $S_i = \beta_0 + \beta_1 P_i + u_i$... firms select S_i and P_i determined

Consequence -

Theoretical - $E(u_i \mathbf{x}_i) \neq \mathbf{0}$ ... don't have k equations to solve for k unknowns in $\boldsymbol{\beta}$

Practical - $E(u_i \mathbf{x}_i) \neq \mathbf{0} \;\Rightarrow\; \frac{1}{N}\sum_{i=1}^{N} x_{ji} u_i \neq 0$ (for some j) $\Rightarrow\; \frac{1}{N}\sum_{i=1}^{N} x_{ji}(y_i - \mathbf{x}_i{}'\boldsymbol{\beta}) \neq 0$ ... that's using the true value of $\boldsymbol{\beta}$ ... but our formula for $\hat{\boldsymbol{\beta}}$ imposes $\frac{1}{N}\sum_{i=1}^{N} x_{ji}(y_i - \mathbf{x}_i{}'\hat{\boldsymbol{\beta}}) = 0$ so we'll end up with $\hat{\boldsymbol{\beta}}$ being <u>biased</u>

Correction - use instrumental variable

Testing - haven't covered yet

Instrumental Variable (IV) Estimations

Simple Case - only 1 variable correlated to the error term:

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$, with $E(x_{1i}u_i) = \cdots = E(x_{k-1,i}u_i) = 0$ and $E(x_{ki}u_i) \neq 0$

Problem - we find $\hat{\boldsymbol{\beta}}$ (k unknowns) by solving k equations in $E(\mathbf{x}_i u_i) = \mathbf{0}$, but in this case we only have k-1 equations because $E(x_{ki}u_i) \neq 0$

Goal - want to find a variable, w_i, that is correlated to x_k, but not correlated to u_i:

$E(w_i x_{ki}) \neq 0$ and $E(w_i u_i) = 0$

How to Do It - define $\mathbf{z}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{k-1,i} \\ w_i \end{bmatrix}$, now use $E(\mathbf{z}_i u_i) = \mathbf{0}$ to get estimate for $\boldsymbol{\beta}$

Sub $u_i = y_i - \mathbf{x}_i{}'\boldsymbol{\beta}$: $E(\mathbf{z}_i u_i) = E\big(\mathbf{z}_i (y_i - \mathbf{x}_i{}'\boldsymbol{\beta})\big) = \mathbf{0}$

Multiply it out and move $E(\mathbf{z}_i y_i)$ to other side: $E(\mathbf{z}_i \mathbf{x}_i')\boldsymbol{\beta} = E(\mathbf{z}_i y_i)$

Solve for $\boldsymbol{\beta}$: $\boldsymbol{\beta} = \left(E(\mathbf{z}_i \mathbf{x}_i')\right)^{-1} E(\mathbf{z}_i y_i)$

Making it Practical - now it's possible to find $\boldsymbol{\beta}$ (in theory), but we don't know expected values so we have to substitute sample averages:

Instrumental Variable Estimator - $\hat{\boldsymbol{\beta}}_{\text{IV}} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i y_i = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$

Note the difference from OLS: $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$

CAUTION - we're not regressing y_i on $\mathbf{z}_i$: $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{k-1} x_{k-1,i} + \beta_k w_i + u_i$ ...

that would give us $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i \neq \hat{\boldsymbol{\beta}}_{\text{IV}}$ ... <u>not the same thing</u>

Instrumental Variable - now ready for official definition; w_i is an instrumental variable for x_{ki} if the following hold:

1) $E(w_i x_{ki}) \neq 0$ and $E(w_i u_i) = 0$ (i.e., w_i correlated to x_{ki}, but not to u_i)
2) $E(\mathbf{z}_i \mathbf{x}_i')$ is nonsingular (required for theory to identify $\boldsymbol{\beta}$)
3) $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'$ is nonsingular (required in practice to calculate $\hat{\boldsymbol{\beta}}_{\text{IV}}$)

Multiple Variables with Single Instruments - now look at same example with $E(x_{k-1,i} u_i) \neq 0$ and $E(x_{ki} u_i) \neq 0$ so we have two variables correlated to the error term and assume we have two instruments, w_{1i} and w_{2i}

Not Assigned - the IVs are not assigned to the correlated variables; each IV could be correlated to one or both of the regressors; for this section we're only concerned with the fact that we have the same number of instruments as we do regressors correlated to the error term (later we'll look at having more IVs... fewer IVs is not possible)

Same Solution - still use $\hat{\boldsymbol{\beta}}_{\text{IV}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$, using $\mathbf{z}_i' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & w_{1i} & w_{2i} \end{bmatrix}$

All Regressors - this works even if all regressors are correlated: $E(x_{1i} u_i) \neq 0$, $E(x_{2i} u_i) \neq 0$, ..., $E(x_{ki} u_i) \neq 0$; use same $\hat{\boldsymbol{\beta}}_{\text{IV}}$ with $\mathbf{z}_i' = \begin{bmatrix} w_{1i} & w_{2i} & \cdots & w_{ki} \end{bmatrix}$

Stata - `ivreg y1 x1 x2 (x3 = z1 z2 z3)`; can use `robust` to correct for heteroskedasticity like before; if no heteroskedasticity and no correlated error terms, then standard error calculated the same way as OLS

What's a Good Instrument? Need to determine if w_i independently explains x_i

Bad Instrument - if correlation between w_i and x_i is week, result is $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i{}'$ being "near singular".... like having near multicollinearity

Detecting - 2 methods

(1) regress each problem regressor (i.e., those correlated to the error term) against all remaining regressors and all IVs (and a constant term if not already there)

Example - from 1 variable example, regress x_{ki} on x_{1i}, x_{2i}, ..., $x_{k\text{-}1,i}$, w_i

Example - from 2 variable example, regress x_{ki} on x_{1i}, x_{2i}, ..., $x_{k\text{-}2,i}$, w_{1i}, w_{2i} and regress $x_{k\text{-}1,i}$ on x_{1i}, x_{2i}, ..., $x_{k\text{-}2,i}$, w_{1i}, w_{2i}

3 Checks - (a) R^2 < 0.1 means weak instrument

(b) coefficient on w_i is not significant means weak instrument

(c) coefficient on w_i is "too small" (even if significant) means weak instrument... "too small" depends on units of x_i and w_i

(2) better method... (given as example with x_{1i} being correlated with error term)

(a) regress x_{1i} on 1, x_{2i}, x_{3i} and save residuals in r_{1i}

(b) regress w_i on 1, x_{2i}, x_{3i} and save residuals in r_{2i}

(c) regress r_{1i} on r_{2i} ... R^2 < 0.1 means weak instrument

Multiple Instruments - having more instruments than there are regressors correlated to the error terms

Single Regressor - use same model with only 1 variable correlated: $E(x_{ki}u_i) \neq 0$, but this time assume we have 2 instruments: w_{1i} and w_{2i}

Problem - $\mathbf{z}_i{}' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & w_{1i} & w_{2i} \end{bmatrix}$ is (k+1)x1 so $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i{}'$ is not a square matrix (it's (k+1)xk)... can't be inverted to calculate $\hat{\boldsymbol{\beta}}_{\text{IV}}$

Solution - define $\mathbf{v}_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{k-1,i} \\ \mathbf{z}_i{}'\boldsymbol{\delta} \end{bmatrix}$, where $\mathbf{z}_i{}'\boldsymbol{\delta} = \delta_1 x_{1i} + \cdots + \delta_{k-1} x_{k-1,i} + \delta_k w_{1i} + \delta_{k+1} w_{2i}$

Want instrument ($\mathbf{z}_i{}'\boldsymbol{\delta}$) to be highly correlated to x_{ki} so we want

$\min_{\boldsymbol{\delta}} \sum_{i=1}^{N} \left(x_{ki} - \mathbf{z}_i{}'\boldsymbol{\delta}\right)^2$ ... that means $\boldsymbol{\delta}$ is least squares estimate:

$$\hat{\boldsymbol{\delta}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i{}'\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i x_{ki}$$

Two-Stage Least Squares - we can estimate $\hat{x}_{ki} = \mathbf{z}_i{}'\hat{\boldsymbol{\delta}}$ and use

$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \mathbf{x}_i{}'\right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i y_i$ ... where $\hat{\mathbf{x}}_i{}' = \begin{bmatrix} x_{1i} & \cdots & x_{k-1,i} & \hat{x}_{ki} \end{bmatrix}$; special property (we won't prove) says $\sum_{i=1}^{N}\hat{\mathbf{x}}_i\mathbf{x}_i' = \sum_{i=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i'$ so we have $\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}} = \left(\sum_{i=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i'\right)^{-1}\sum_{i=1}^{N}\hat{\mathbf{x}}_i y_i$ ... that's OLS for regressing y_i on $\hat{\mathbf{x}}_i$

In Practice -

1) Run each regressor that is correlated to the error term on the entire IV list (i.e., $\mathbf{z}_i$) and save fitted values to generate $\hat{\mathbf{x}}_i$
2) Run y_i on $\hat{\mathbf{x}}_i$

Example - consider 2 regressors correlated to error term: $E(x_{k-1,i}u_i) \neq 0$ and $E(x_{ki}u_i) \neq 0$; and four IVs: w_{1i}, w_{2i}, w_{3i}, and w_{4i}

1) Regress $x_{k-1,i}$ on $x_{1i}, x_{2i}, \ldots, x_{k-2,i}, w_{1i}, w_{2i}, w_{3i}, w_{4i}$ and generate $\hat{x}_{k-1,i}$

   Regress $x_{k,i}$ on $x_{1i}, x_{2i}, \ldots, x_{k-2,i}, w_{1i}, w_{2i}, w_{3i}, w_{4i}$ and generate $\hat{x}_{k,i}$
2) Regress y_i on $x_{1i}, x_{2i}, \ldots, x_{k-2,i}, \hat{x}_{k-1,i}, \hat{x}_{ki}$

Note1: if there is one instrument for each correlated regressor, then $\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}} = \hat{\boldsymbol{\beta}}_{\mathrm{IV}}$

Note2: this procedure gives the correct $\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$, but the wrong standard error... Stata does it right

Stata - `ivreg y x1 x2 ... xk-2 (xk-1 xk = w1 w2 w3 w4)`

Testing Parameters - $\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$ is best estimator when some regressors are correlated to the error term, but other assumptions hold (no heteroskedasticity, no correlated error terms); Wald and *t* tests are OK, but *F*-test is not valid; the *F*-test is based on regression residuals being orthogonal to the regressors, but for $\hat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$, the residuals are orthogonal to $\hat{\mathbf{x}}_i$, not necessarily $\mathbf{x}_i$

Finding Instruments - ad hoc rules... didn't cover them yet

Testing Instruments - in order to have a specified model (i.e., same number of equations [$E(x_i u) = 0$] and unknowns [parameters]), we must have at least 1 valid instrument for each regressor that's correlated to the error term; this test only looks at additional instruments

Hausman Test - used to test the additional instruments

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$ and $E(x_{3i}u_i) \neq 0$

Suppose w_{1i} and w_{2i} are instruments and w_{2i} is known to be a good instrument

H_0: $E(w_1 u) = 0$ and H_1: $E(w_1 u) \neq 0$

Define $\hat{\boldsymbol{\beta}}$ as the 2SLS estimator using both w_{1i} and w_{2i} for x_{3i}

Define $\tilde{\boldsymbol{\beta}}$ as the 2SLS estimator using only w_{2i} for x_{3i}

Under H_0, both $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent so $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \mathbf{0}$; if H_0 fails $\hat{\boldsymbol{\beta}}$ is biased

Test Statistic - $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})'Cov(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \sim \chi^2_k$ (*k* = # coefficients)

Example - `reg hours kidslt6 educ wage hushrs faminc unem, robust`

`hushrs` (husband hours) is probably a joint decision when deciding the wife's hours (`hours`), so it's probably correlated with the error term

Suppose `huseduc` is known to be a good instrument; test if `huswage` is also a good instrument:

```
ivreg hours kidslt6 educ wage famine unem (hushrs = huseduc),
   robust
hausman, save
ivreg hours kidslt6 educ wage famine unem (hushrs = huswage
   huseduc), robust
hausman
```

Other Uses of Hausman Test

Endogeneity - used Hausman Test earlier to test additional instruments, but can also use it to test endogeneity (i.e., test whether regressor is correlated with the error term)

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$ and $E(x_{3i}u_i) \neq 0$

Suppose z_i is an instrument

History - technically was first discovered by Durbin, then rediscovered by Hausman and Wu (independently in 1950s)... called Hausman test because he's from MIT

H_0: $E(x_{3i}u_i) = 0$ and H_1: $E(x_{3i}u_i) \neq 0$

Two Estimators - Hausman tests requires two estimators ($\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$) such that

(1) Both are consistent under H_0 and one is best

(2) Under H_1, only one is consistent (i.e., other is biased under H_1)

Idea - if H_0 is true, $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \mathbf{0}$ ... technically, it converges to zero (order doesn't matter)

Quadratic Distance - like Wald Test, measure distance between the vectors with

$(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' Cov(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \sim \chi^2_k$ (k = rank of Cov matrix)

If this distance is "close to zero" then there's no evidence to reject H_0

Rank Issues - Usually rank of Cov matrix = # parameters; if rank < # parameters then we can't take the inverse so use generalized inverse (didn't cover this in class, but Ai said software packages will do it automatically)

Order Matters - order doesn't matter for differences as long as they're the same (these are being squared so negative goes away), but order does matter for Cov term...

$\tilde{\boldsymbol{\beta}}$ is the inefficient estimator (i.e., not best under H_0)

Cov Matrix - $Cov(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = Cov(\hat{\boldsymbol{\beta}}) + Cov(\tilde{\boldsymbol{\beta}}) - 2Cov(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$ ... this is an asymptotic result so it's not guaranteed for small samples; first two terms come directly from the regressions used to estimate the parameters; the third term is complicated

Hausman's Trick - by imposing condition that $\hat{\boldsymbol{\beta}}$ is best under H_0, we can use

$Cov(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}})$

$\therefore$ test statistic becomes: $\boxed{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})' \left[Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}})\right]^{-1} (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \sim \chi^2_k}$

From Example - $\hat{\boldsymbol{\beta}}$ is OLS estimator; $\tilde{\boldsymbol{\beta}}$ is 2SLS (IV) estimator with z_i as instrument for x_{3i}

Problem 1 - if heteroskedasticity, $\hat{\boldsymbol{\beta}}$ is not best so $Cov(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}) \neq Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}})$

Problem 2 - if $Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}}) < \mathbf{0}$ could be several reasons: (a) heteroskedasticity or some other failure that makes $\hat{\boldsymbol{\beta}}$ not best; (b) sample problem (not big enough or just a "bad" sample

Heteroskedasticity - $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $E(u_i^2 \mid \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\alpha}$; define estimators as $\hat{\boldsymbol{\beta}}$ is GLS estimator and $\tilde{\boldsymbol{\beta}}$ is OLS estimator... both are consistent under H_0 and H_1 ∴ cannot use Hausman test... heteroskedasticity affects computation of variance of estimates, it doesn't affect the fact that they are consistent

Serial Correlation - $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $u_i = \rho u_{i-1} + \varepsilon_i$; H_0: $\rho = 0$, H_1: $\rho \neq 0$; define estimators as $\hat{\boldsymbol{\beta}}$ is OLS estimator and $\tilde{\boldsymbol{\beta}}$ is GLS estimator... both are consistent under H_0 and H_1 (as long as y_{i-1} is not a regressor) ∴ cannot use Hausman test

Exception - if we change model it include lagged dependent variable (e.g., $y_i = \beta_0 + \beta_1 y_{i-1} + \beta_2 x_i + u_i$), then Hausman test is valid using $\hat{\boldsymbol{\beta}}$ is OLS estimator and $\tilde{\boldsymbol{\beta}}$ is 2SLS estimator using x_{i-1} as instrument for y_{i-1}; under H_0, $\hat{\boldsymbol{\beta}}$ is consistent and best and $\tilde{\boldsymbol{\beta}}$ is consistent (but inefficient); under H_1, $\hat{\boldsymbol{\beta}}$ is not consistent and $\tilde{\boldsymbol{\beta}}$ is

Coefficients - $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$; H_0: $\beta_2 = 0$ and $\beta_3 = 0$, H1: $\beta_2 \neq 0$ or $\beta_3 \neq 0$

Define $\tilde{\boldsymbol{\beta}}$ as unrestricted OLS and $\hat{\beta}_1$ comes from the restricted OLS (i.e., $y_i = \beta_1 x_{1i} + u_i$)

so $\hat{\boldsymbol{\beta}}$ is $\begin{bmatrix} \hat{\beta}_1 \\ 0 \\ 0 \end{bmatrix}$; now use $Cov(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} Var(\hat{\beta}_1) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$... Hausman Test works!

***F*-Test** - can do this same test with an *F*-test... it's better (and easier)

Limited Information

$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_i$ with $E(\mathbf{x}_{1i}u_i) = \mathbf{0}$ and $E(\mathbf{x}_{2i}u_i) \neq \mathbf{0}$ (i.e., $\mathbf{x}_{2i}$ are regressors correlated with error term)

Need instrument for each regressor; suppose: $\mathbf{x}_{2i} = \boldsymbol{\alpha}_1 y_i + \mathbf{x}_{1i}'\boldsymbol{\alpha}_2 + \mathbf{z}_i'\boldsymbol{\alpha}_3 + \varepsilon_i$

Practice - economic theory may specify y_i, but it doesn't care about $\mathbf{x}_{2i}$ (i.e., partial [one market] equilibrium vs. total equilibrium)

Structural Equations - y_i and $\mathbf{x}_{2i}$ for system of simultaneous equations... full information

Reduced Form - sub y_i into $\mathbf{x}_{2i}$ to get $\mathbf{x}_{2i} = \mathbf{x}_{1i}'\boldsymbol{\pi}_1 + \mathbf{z}_i'\boldsymbol{\pi}_2 + v_i$... limited information; run 2SLS using $\mathbf{x}_{2i} = \mathbf{x}_{1i}'\boldsymbol{\pi}_1 + \mathbf{z}_i'\boldsymbol{\pi}_2 + v_i$ for first stage and plugging that into y_i for second stage; given limited information (and homoskedasticity), 2SLS is best estimator

Maximum Likelihood - could also estimate jointly if we assume $(u_i, v_i) \sim N(0, \mathbf{\Omega})$ (multivariate normal).... do limited information maximum likelihood estimator (LIML); it's also best

2SLS - is just as good as LIML (i.e., best), but "better" because it's less complicated and we don't need to include all instruments

Two-Stage Least Squares

Refresher - $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$

1) Run regression to estimate error term and determine which regressors are correlated to the error term
2) Create vector comprised of all regressors that are NOT correlated to the error term and all the instrumental variables (Note: must have at least 1 instrument for each regressor that was not included [i.e., correlated to the error term]; must include constant term too)
   e.g., $\mathbf{z}_i' = [x_{1i} \quad \cdots \quad x_{k-1,i} \quad w_{1i} \quad w_{2i}]$ (only x_{ki} is correlated and there are 2 instruments)
3) **1st Stage** - estimate each correlated regressor by regressing it on $\mathbf{z}_i$

   e.g., $\hat{x}_{ki} = \mathbf{z}_i'\hat{\boldsymbol{\delta}}$

   Define $\hat{\mathbf{x}}_i$ as vector of uncorrelated regressors and estimated values for correlated regressors

   e.g., $\hat{\mathbf{x}}_i' = [x_{1i} \quad \cdots \quad x_{k-1,i} \quad \hat{x}_{ki}]$
4) **2nd Stage** - Regress y_i on $\hat{\mathbf{x}}_i$

F-Test - need to modify the F-test to work with 2SLS

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) \neq 0$ and $E(x_{2i}u_i) = E(x_{3i}u_i) = 0$ and w_i is instrument for x_{1i}

Two restrictions: H$_0$: $\beta_1 = 0$ and $\beta_2 = 0$

Define $\mathbf{z}_i' = [w_i \quad x_{2i} \quad x_{3i}]$

1st Stage - regress x_{1i} on $\mathbf{z}_i$ and generate $\hat{x}_{1i}$ (can repeat for other regressors, but they'll be perfect fits so we can skip that step and just use $\hat{x}_{2i} = x_{2i}$ and $\hat{x}_{3i} = x_{3i}$)

Behind the Scenes - we're really assuming $x_{1i} = \mathbf{z}_i'\boldsymbol{\delta}_1 + \boldsymbol{\gamma}_{1i}$, $x_{2i} = \mathbf{z}_i'\boldsymbol{\delta}_2 + \boldsymbol{\gamma}_{2i}$, and $x_{3i} = \mathbf{z}_i'\boldsymbol{\delta}_3 + \boldsymbol{\gamma}_{3i}$... when we sub those into the original equation we get:

$$y_i = \beta_1(\mathbf{z}_i'\boldsymbol{\delta}_1) + \beta_2(\mathbf{z}_i'\boldsymbol{\delta}_3) + \beta_3(\mathbf{z}_i'\boldsymbol{\delta}_3) + [u_i + \beta_1\boldsymbol{\gamma}_{1i} + \beta_2\boldsymbol{\gamma}_{2i} + \beta_3\boldsymbol{\gamma}_{3i}]$$

so the error term when we go to the second stage is not actually estimating u_i, but that long thing in the brackets... that's why the F-test needs to be modified

2nd Stage - regress y_i on $\hat{\mathbf{x}}_i$ to get 2SLS estimators $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ and the residuals $\hat{u}_i = y_i - \hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1\hat{x}_{1i} - \hat{\beta}_2\hat{x}_{2i} - \hat{\beta}_3\hat{x}_{3i}$ (which we just showed are not consistent so we have to get a consistent estimate for u_i: $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

Restricted Regression - regress y_i on $\hat{x}_{3i}$ (recall restriction above says $\beta_1 = 0$ & $\beta_2 = 0$); generate residuals $\tilde{u}_i = y_i - \tilde{\beta}_3 \hat{x}_{3i}$

F-Statistic - $F = \dfrac{\left(\sum \tilde{u}_i^2 - \sum \hat{u}_i^2\right)/m}{\sum e_i^2 / n-k}$, as before, m = # restrictions (2 in this case), n = # observations and k = # parameters in full model (3 in this case)

Heteroskedasticity in 2SLS - since we're running 2SLS, we know $E(u_i \mid \mathbf{x}_i) \neq 0$, but $E(u_i \mid \mathbf{z}_i) = 0$; we should have $E(u_i^2 \mid \mathbf{z}_i) = \sigma^2$, but if we don't, there's heteroskedasticity (i.e., $E(u_i^2 \mid \mathbf{z}_i) \neq \sigma^2$)

Detecting - (1) run 2SLS and get $\hat{\boldsymbol{\beta}}$;

(2) compute <u>consistent</u> residuals: $e_i = y_i - \mathbf{x}_i{}'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

(3) regress e_i^2 on $(1 \;\; \mathbf{z}_i)$ (i.e., be sure to include a constant term if it's not already in $\mathbf{z}_i$)

(4) do overall test of significance (i.e., standard F-test to check if all parameters are simultaneously equal to zero)... if regression is significant, there's heteroskedasticity

Correcting - (1) save fitted value of $\hat{e}_i^2$ (from regression in step (3) above)

(2) transform model: $\dfrac{y_i}{\hat{e}_i} = \dfrac{\mathbf{x}_i{}'}{\hat{e}_i}\boldsymbol{\beta} + \dfrac{u_i}{\hat{e}_i} = \beta_1 \dfrac{x_{1i}}{\hat{e}_i} + \hat{\beta}_2 \dfrac{x_{2i}}{\hat{e}_i} + \hat{\beta}_3 \dfrac{x_{3i}}{\hat{e}_i} + \dfrac{u_i}{\hat{e}_i}$

(3) do 2SLS on the transformed model; can use $\mathbf{z}_i = \begin{bmatrix} w_i \\ x_{2i} \\ x_{3i} \end{bmatrix}$ or $\mathbf{z}_i = \begin{bmatrix} w_i / \hat{e}_i \\ x_{2i} / \hat{e}_i \\ x_{3i} / \hat{e}_i \end{bmatrix}$... will give different results, but both have same statistical properties

Serial Correlation in 2SLS -

Detecting - same as before (e.g., use $e_i = \rho e_{i-1} + \boldsymbol{\gamma}_i$ to estimate $\hat{\rho}$; if it's significantly different than zero, there's serial correlation)

Correcting -

(1) transform model:

$$(y_i - \hat{\rho} y_{i-1}) = \beta_1 (x_{1i} - \hat{\rho} x_{1i-1}) + \beta_2 (x_{2i} - \hat{\rho} x_{2i-1}) + \beta_3 (x_{3i} - \hat{\rho} x_{3i-1}) + (u_i - \hat{\rho} u_{i-1})$$

(2) do 2SLS on the transformed model; can use $\mathbf{z}_i - \hat{\rho}\mathbf{z}_{i-1}$, $\mathbf{z}_i$, or $\mathbf{z}_{i-1}$ (will have same statistical properties)

Another Test for Endogeneity

Endogeneity - Hausman Test not valid with heteroskedasticity so here's another way to test for endogeneity (i.e., test whether regressor is correlated with the error term)

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$ and $E(x_{3i}u_i) \neq 0$

Suppose z_i is an instrument

1. Estimate reduced for x_{3i}: $x_{3i} = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 z_i + \varepsilon_i$
2. Get predictions for error term: $\hat{\varepsilon}_i = x_{3i} - \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 z_i$
3. Plug prediction into original model: $y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 \hat{\varepsilon}_i + u_i$
4. H$_0$: x_{3i} is exogenous is equivalent to H$_0$: $\alpha_3 = 0$

Special Topics

I. Use Instrumental Variable to Fix Specification Problem

(e.g., omitted variable)

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$

Assume we don't have data on x_{4i}

If x_{4i} is correlated to x_{1i}, x_{2i}, or x_{3i}, then could have regressors correlated with error term if x_{4i} is missing from the regression

Traditional Solution - proxy variable: $w_i = \delta_0 + \delta_1 x_{4i} + \varepsilon_i \;\Rightarrow\; x_{4i} = \dfrac{w_i - \delta_0 - \varepsilon_i}{\delta_1}$

Plug that into original model: $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 \left(\dfrac{w_i - \delta_0 - \varepsilon_i}{\delta_1} \right) + u_i$

To be more general, assume $x_{1i} = 1$ (i.e., constant):

$$y_i = \left(\beta_1 - \frac{\delta_0}{\delta_1} \right) + \beta_2 x_{2i} + \beta_3 x_{3i} + \frac{\beta_4}{\delta_1} w_i + \left(u_i - \frac{\varepsilon_i}{\delta_1} \right)$$

This is the actual equation we run the regression on

Problem - w_i is correlated with ε_i so all the estimators will be biased; in fact, the less correlated w_i is to x_{4i} (i.e., more correlated to ε_i), the worse the problem is; if w_i is perfectly correlated to x_{4i} this problem doesn't occur

Fixing It - easiest way is to test H$_0$: $\beta_4 = 0$ with standard *t*-test by using the *t*-ratio for β_4 / δ_1 ... if we can't reject H$_0$, then we can remove the proxy and run the model without x_{4i}

Use Instrumental Variable - add another proxy $w_{1i} = \lambda_0 + \lambda_1 x_{4i} + \varepsilon_{1i}$

If ε_{1i} is not correlated to ε_i, then we can use w_{1i} as instrument for w_i

II. Generated Regressors

$y_i = \beta_1 x_{1i} + \beta_2 \hat{x}_{2i} + \beta_3 \hat{x}_{3i} + u_i$, where $\hat{x}_{2i}$ and $\hat{x}_{3i}$ are estimates for x_{2i} and x_{3i} (i.e., $\hat{x}_{2i} = x_{2i} + \eta_{2i}$ and $\hat{x}_{3i} = x_{3i} + \eta_{3i}$)

Example - have to estimate expected price in model to forecast GDP

Key - as sample size gets larger, η_{2i} and η_{3i} "go away" (i.e. $\hat{x}_{2i}$ and $\hat{x}_{3i}$ are better estimates for x_{2i} and x_{3i})

Problem - OLS estimates ($\hat{\boldsymbol{\beta}}$) are OK, but statistical tests aren't valid because the standard error computed by software packages is incorrect; the correct version is very complicated (i.e., "beyond the scope of this course")

Exception 1 - *F*-test and Wald Test for $\beta_2 = \beta_3 = 0$ are OK (because if you fail to reject, you can drop the variables and not worry about generating them!)

Exception 2 - technically we use generated regressors in 2SLS, but the problem above (i.e., standard error not being correct) doesn't apply because we know how the instruments are generated

III. Generated Instruments

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ with $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$, $E(x_{3i}u_i) \neq 0$; w_i instrument for x_{3i}

If $\hat{w}_i$ is consistent estimate of w_i it can be used and there's no effect on 2SLS (see exception 2 in previous section)

IV. Testing Nonlinearity (Functional Form)

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$

Assume we want to know if model is misspecified (e.g., should there be a higher order term or an interaction term?)

Run OLS and get $\hat{\boldsymbol{\beta}}$ and generate residuals: $\hat{u}_i = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

If model is misspecified $\hat{u}_i$ will be correlated with higher order (or interaction) terms

Method 1 - run $\hat{u}_i = b_0 + b_1 x_{1i}^2 + b_2 x_{2i}^2 + b_3 x_{2i}^2 + b_4 x_{1i} x_{2i} + b_5 x_{1i} x_{2i} + b_6 x_{2i} x_{3i} +$ etc.; could include cubed terms (or higher)... usually end up with too many terms to check, but this is technically better than the Ramsey Test

Method 2 (Ramsey Test) -

(1) Regress $\hat{u}_i$ on x_{1i}, x_{2i}, x_{3i}... this is restricted regression; technically it should be insignificant because $\hat{u}_i$ is orthogonal to the regressors (by design)

(2) Generate fitted values: $\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$

(3) Run $\hat{u}_i = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 x_{3i} + \delta_4 \hat{y}_i^2 + \delta_5 \hat{y}_i^3 + \delta_6 \hat{y}_i^4 + \varepsilon_i$... this is unrestricted regression (note also that we're using generated regressors)

(4) Use F-test to test $\delta_4 = \delta_5 = \delta_6 = 0$... if fail to reject, then there probably aren't any higher order terms

(5) Optional - can look at δ_4, δ_5, δ_6 individually to get hint on correct functional form; technically can't use *t*-test, but if *t*-ratio is very small (or very big), we're probably safe to say it's not (or is) significant... have to use some judgment there

V. Difference in Difference

Cross section over time where we observe 2 groups at 2 times

Group A	y_{A1}	y_{A2}
Group B	y_{B1}	y_{B2}
	$t = 1$	$t = 2$

Treatment Group - group A; receives treatment (i.e., policy change, training, reorganization, etc.)

Control Group - group B; don't receive treatment

Observations - y_{ij} is group *i*'s average at time *j*

Buried Treatment Effects - two of them:

(1) $y_{\mathrm{A}2} - y_{\mathrm{A}1}$ - includes treatment effect and other factors (e.g., tastes change over time); doesn't account for control

(2) $y_{\mathrm{A}2} - y_{\mathrm{B}2}$ - includes treatment effect, but because it includes different groups there are other factors (e.g., different tastes between groups)

Assumption - in order to isolate the actual treatment effect we have to assume that if behavior (e.g., tastes or preferences) changes over time, it changes the same way for both groups

Actual Treatment Effect - $y_{\mathrm{A}2} - y_{\mathrm{A}1} - \left(y_{\mathrm{B}2} - y_{\mathrm{B}1}\right)$ (second term accounts for change in behavior)

CAUTION - this is only valid for a <u>controlled</u> experiment; otherwise we need more complicated techniques (will cover in second year course)

VI. Seemingly Unrelated Regression Estimate (SURE) Model

Pooled Cross-Section Time-Series Data - collect different cross-section data over time; may not necessarily be from same source (i.e., might not be same people in sample from year to year)

Panel Data - random selection of individuals but all future observations come form same individuals... we'll study this later

Pooled Regression - combines more than 1 regression model $y_{1i} = \beta_1 x_{1i} + u_{1i}$ and $y_{2i} = \beta_2 x_{2i} + u_{2i}$; could have different or same regressors (x_{1i} and x_{2i}) or regressands (y_{1i} and y_{2i}); examples:

(a) demand equation for year 1 and year 2 (*y*'s same; *x*'s same)

(b) demand for pork and demand for beef (*y*'s different; *x*'s same [some of them anyway])

(c) demand for pork vs. price of pork and demand for beef vs. price of beef (*y*'s different; *x*'s different)

Why Use Pooled Regression -

(1) might be able to get better estimate by increasing sample size

(2) might be able to get a "better" standard error

(3) to test cross-equation restrictions (e.g., $\beta_1 = \beta_2$)

Solving - can solve each regression separately by simple OLS or pool them together and apply OLS of "stacked data"

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} x_{11} & 0 \\ x_{12} & 0 \\ \vdots & \vdots \\ x_{1n_1} & 0 \\ 0 & x_{21} \\ 0 & x_{22} \\ \vdots & \vdots \\ 0 & x_{2n_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1n_1} \\ u_{21} \\ u_{22} \\ \vdots \\ u_{2n_2} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{1i} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{1i} \\ \mathbf{u}_{2i} \end{bmatrix} \quad \text{or} \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Theoretical Mathematics - this "stacking" is done for mathematical reasons; there's no economic theory at work here... in fact, the new "dependent variable" could be gibberish from an economic perspective

Parameter Estimates - $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ ... will be the <u>same</u> as running the regressions separately

Problems - pooled regression estimates $\mathrm{var}(\mathbf{u})$, but single regressions estimate $\mathrm{var}(u_{1i})$ and $\mathrm{var}(u_{2i})$ and these are <u>not the same</u> (if u_{1i} and u_{2i} are correlated)... that means standard error and *t*-ratios will not be the same

Solving Correlation Problem - use GLS estimation

Define **variance-covariance matrix** $\mathbf{\Omega} = E\begin{bmatrix} u_{1i}^2 & u_{1i}u_{2i} \\ u_{1i}u_{2i} & u_{2i}^2 \end{bmatrix}$

Use **Cholesky Decomposition** to break it down into a triangular matrix $\mathbf{\Omega} = \mathbf{\Gamma\Gamma}'$ (apparently, the exact technique for this isn't important... statistical package will do it)

"Crop" data - pair data so we set $n = \min(n_1, n_2)$

Rewrite model: $\mathbf{\Gamma}^{-1}\mathbf{Y} = \mathbf{\Gamma}^{-1}\mathbf{X\beta} + \mathbf{\Gamma}^{-1}\mathbf{u}$

Which can be written: $\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{x}}_{11i} & \tilde{\mathbf{x}}_{12i} \\ \tilde{\mathbf{x}}_{21i} & \tilde{\mathbf{x}}_{22i} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{u}}_{1i} \\ \tilde{\mathbf{u}}_{2i} \end{bmatrix}$ or $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{\beta} + \tilde{\mathbf{u}}$

$$\begin{bmatrix} \tilde{x}_{11,1} & \tilde{x}_{12,1} \\ \tilde{x}_{11,2} & \tilde{x}_{12,2} \\ \vdots & \vdots \\ \tilde{x}_{11,n} & \tilde{x}_{12,n} \\ \tilde{x}_{21,1} & \tilde{x}_{22,1} \\ \tilde{x}_{21,2} & \tilde{x}_{22,2} \\ \vdots & \vdots \\ \tilde{x}_{21,n} & \tilde{x}_{22,n} \end{bmatrix}$$

Note: $\tilde{\mathbf{X}}$ is not block diagonal like $\mathbf{X}$ is

Now have $\text{var}(\tilde{u}_{1i}) = \text{var}(\tilde{u}_{2i}) = 1$ and $\text{cov}(\tilde{u}_{1i}, \tilde{u}_{2i}) = 0$

New Estimates - $\hat{\mathbf{\beta}}_{\text{GLS}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$

Estimating $\mathbf{\Omega}$

1. Run each regression individually
2. Save residuals: $\hat{u}_{1i} = y_{1i} - \beta_1 x_{1i}$ and $\hat{u}_{2i} = y_{2i} - \beta_2 x_{2i}$
3. $\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \hat{u}_{1i}^2 & \hat{u}_{1i}\hat{u}_{2i} \\ \hat{u}_{1i}\hat{u}_{2i} & \hat{u}_{2i}^2 \end{bmatrix}$
4. Use statistical package to do the decomposition: $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}'$

Wasted Effort? - $\hat{\mathbf{\beta}}_{\text{OLS}}$ and $\hat{\mathbf{\beta}}_{\text{GLS}}$ will be (asymptotically the same) in 2 cases:

(1) u_{1i} and u_{2i} are <u>not</u> correlated (i.e., $\text{cov}(u_{1i}, u_{2i}) = 0$ or $\mathbf{\Omega}$ is a diagonal matrix)

(2) $x_{1i} = x_{2i}$ (i.e., same values for regressors)

If either of these is not the case, $\hat{\mathbf{\beta}}_{\text{GLS}}$ will be more efficient

Problem - procedure doesn’t handle heteroskedasticity of serial correlation... if either of those exists, you have to do it by hand (kind of)
Create new variable to run clustered regression (see Stata notes)

$$\mathbf{d} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \\ 1 \\ 2 \\ \vdots \\ n \end{bmatrix}$$

VII. SURE with Endogenous Regressors (3SLS)

Endogeneity - at least one regressor is correlated with the error term; those regressors that are correlated at called **endogenous regressors**

Specification - if a model has one or more endogenous regressors, it is **under specified** (there are more unknowns that equations so the model can't be solved); if there is one IV for each endogenous regressor, the model is **exactly specified**; if there are more IVs than endogenous regressors, the model is **over specified**

Model - $y_{1i} = \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i} + u_{1i}$

$y_{2i} = \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i} + u_{2i}$

$E(x_{1i}u_{1i}) = E(x_{1i}u_{2i}) = 0$ and $E(u_{1i}) = E(u_{2i}) = 0$ ... no problem there

$E(x_{2i}u_{1i}) \neq 0$ and $E(x_{2i}u_{2i}) \neq 0$

z_i is an instrument for x_{2i} (i.e. $E(z_i u_{1i}) = E(z_i u_{2i}) = 0$ and $E(x_{2i}z_i) \neq 0$)

2SLS - can estimate coefficients by applying 2SLS to each equation individually:

1. Regress x_{2i} on 1, x_{1i}, z_i and predict $\hat{x}_{2i}$
2. Run $y_{1i} = \beta_{10} + \beta_{11}x_{1i} + \beta_{12}\hat{x}_{2i}$ and $y_{2i} = \beta_{20} + \beta_{21}x_{1i} + \beta_{22}\hat{x}_{2i}$

Pooled Regression (3SLS) - combines 2SLS with SURE model; would do it for same reasons covered in previous section (better estimates, "better" standard error, cross-equation restrictions)

1. Regress x_{2i} on 1, x_{1i}, z_i and predict $\hat{x}_{2i}$
2. Run $y_{1i} = \beta_{10} + \beta_{11}x_{1i} + \beta_{12}\hat{x}_{2i}$ and $y_{2i} = \beta_{20} + \beta_{21}x_{1i} + \beta_{22}\hat{x}_{2i}$ to get
   a. coefficient estimates: $\hat{\beta}_{10}$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, $\hat{\beta}_{20}$, $\hat{\beta}_{21}$, $\hat{\beta}_{22}$
   b. residuals: $\hat{u}_{1i} = y_{1i} - \hat{\beta}_{10} + \hat{\beta}_{11}x_{1i} + \hat{\beta}_{12}\hat{x}_{2i}$ and $\hat{u}_{2i} = y_{2i} - \hat{\beta}_{20} + \hat{\beta}_{21}x_{1i} + \hat{\beta}_{22}\hat{x}_{2i}$

3rd Stage -

Theory -

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \hat{x}_{21} & 0 & 0 & 0 \\ 1 & x_{12} & \hat{x}_{22} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \hat{x}_{2n} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{11} & \hat{x}_{21} \\ 0 & 0 & 0 & 1 & x_{12} & \hat{x}_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & x_{1n} & \hat{x}_{2n} \end{bmatrix} \begin{bmatrix} \beta_{10} \\ \beta_{11} \\ \beta_{12} \\ \beta_{20} \\ \beta_{21} \\ \beta_{22} \end{bmatrix} + \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1n} \\ u_{21} \\ u_{22} \\ \vdots \\ u_{2n} \end{bmatrix} \quad \text{or}$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 & \hat{\mathbf{x}}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{x}_1 & \hat{\mathbf{x}}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$$

Problem - pooled regression estimates $\text{var}(\mathbf{u})$, but single regressions estimate $\text{var}(u_{1i})$ and $\text{var}(u_{2i})$ and these are <u>not the same</u> (if u_{1i} and u_{2i} are correlated)... that means standard error and *t*-ratios will not be the same

Variance-Covariance Matrix - $\mathbf{\Omega} = E\begin{bmatrix} u_{1i}^2 & u_{1i}u_{2i} \\ u_{1i}u_{2i} & u_{2i}^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix} \neq \mathbf{I}\sigma^2$

Cholesky Decomposition - same as before; break matrix down into a triangular matrix $\mathbf{\Omega} = \mathbf{\Gamma\Gamma}'$

"Crop" data - if had different amount of data for each equation, we have to pair data and set $n = \min(n_1, n_2)$ ... drop extra data points just like we covered in the previous section

Rewrite Model - $\mathbf{\Gamma}^{-1}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{\Gamma}^{-1}\begin{bmatrix} \mathbf{1} & \mathbf{x}_1 & \hat{\mathbf{x}}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{x}_1 & \hat{\mathbf{x}}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \mathbf{\Gamma}^{-1}\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$

$E\left(\mathbf{\Gamma}^{-1}\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}\right) = \mathbf{I}$ (i.e., $\text{var}(\tilde{u}_{1i}) = \text{var}(\tilde{u}_{2i}) = 1$ and $\text{cov}(\tilde{u}_{1i}, \tilde{u}_{2i}) = 0$)

New Estimates - $\hat{\boldsymbol{\beta}}_{3\text{SLS}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$

Practice - note that there was only 1 endogenous variable above (the same one in both equations); to make this more general, we'll now look at two endogenous variables, one in each equation; since we have two of them, we need at least to IVs:

Model - $y_{1i} = \beta_{10} + \beta_{11} x_{1i} + \beta_{12} x_{2i} + u_{1i}$ (same as before)

$y_{2i} = \beta_{20} + \beta_{21} x_{3i} + \beta_{22} x_{4i} + u_{2i}$ (two new variables)

$E(x_{1i} u_{1i}) = E(x_{3i} u_{2i}) = 0$ and $E(u_{1i}) = E(u_{2i}) = 0$ ... no problem there

$E(x_{2i} u_{1i}) \neq 0$ and $E(x_{4i} u_{2i}) \neq 0$

z_{1i} and z_{2i} are instruments for x_{2i} and x_{4i}

Stage 1 - use IVs to get fitted values for endogenous variables

a. Regress x_{2i} on 1, x_{1i}, z_{1i}, z_{2i} and predict $\hat{x}_{2i}$

b. Regress x_{4i} on 1, x_{3i}, z_{1i}, z_{2i} and predict $\hat{x}_{4i}$

Stage 2 - plug fitted values into original equations to get residuals

a. Run $y_{1i} = \beta_{10} + \beta_{11} x_{1i} + \beta_{12} \hat{x}_{2i}$ to get $\hat{u}_{1i} = y_{1i} - \hat{\beta}_{10} + \hat{\beta}_{11} x_{1i} + \hat{\beta}_{12} \hat{x}_{2i}$

b. Run $y_{2i} = \beta_{20} + \beta_{21} x_{3i} + \beta_{22} \hat{x}_{4i}$ to get $\hat{u}_{2i} = y_{2i} - \hat{\beta}_{20} + \hat{\beta}_{21} x_{3i} + \hat{\beta}_{22} \hat{x}_{4i}$

Stage 3 -

a. Estimate $\hat{\mathbf{\Omega}} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \hat{u}_{1i}^2 & \hat{u}_{1i} \hat{u}_{2i} \\ \hat{u}_{1i} \hat{u}_{2i} & \hat{u}_{2i}^2 \end{bmatrix}$

b. Use statistical package to do the decomposition: $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Gamma}} \hat{\mathbf{\Gamma}}'$

c. Run regression with transformed data to get $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$

Wasted Effort? - $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$ will be numerically the same in 2 cases (same as previous section):

(1) u_{1i} and u_{2i} are <u>not</u> correlated (i.e., $\rho = \text{cov}(u_{1i}, u_{2i}) = 0$; $\mathbf{\Omega}$ is a diagonal matrix)

(2) same values for regressors

3SLS Better - If above cases don't hold (and below case doesn't hold), then $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$ are asymptotically the same, but $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$ will be better estimate (i.e., have smaller variance) if the model is over specified... that's good to know for the Hausman test which requires 2 estimates [1 better than the other]

Idea - extra information from data in second model could help get better predictions

3SLS Worse - 3SLS is less robust; if we're interested in the first equation, but the second equation is misspecified (e.g., (a) think there are no endogenous variables, but there are; (b) functional form is wrong; (c) missing variables, etc.), then $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$ will be <u>biased</u>

Iteration - after stage 3, we can use $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$ to estimate new residuals, then repeat stage 3 (i.e., reestimate $\hat{\mathbf{\Omega}}$ and $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$); eventually $\hat{\boldsymbol{\beta}}_{\text{3SLS}}$ will converge to the maximum likelihood estimate

Full Information Maximum Likelihood Estimation (FIML) - specify everything (e.g., $(u_{1i}, u_{2i}) \sim$ joint normal); 3SLS is the same as FIML (but easier); 2SLS is same as LIML (limited information)

When is System Identified -

Linear System - $D_i = \beta_0 + \beta_1 p_i + u_i$ and $S_i = \alpha_0 + \alpha_1 p_i + \varepsilon_i$

$D_i = S_i = Q_i$ so Q_i and p_i are jointly determined... p_i is endogenous (correlated to error term)

Need instruments for p_i

If we add income to demand equation, we can use I_i as an IV for price in supply equation, but we still have D_i under specified because we don't have an instrument for price in the demand equation; if we add wage to the supply equation, we can use it as the IV in the demand equation

$D_i = \beta_0 + \beta_1 p_i + \beta_2 I_i + u_i$ and $S_i = \alpha_0 + \alpha_1 p_i + \alpha_2 w_i + \varepsilon_i$

Nonlinear Endogenous Variable - $D_i = \beta_0 + \beta_1 p_i + \beta_2 I_i + u_i$ and $S_i = \alpha_0 + \alpha_1 \ln p_i + \varepsilon_i$

Second equation is OK because we can use income from the first one as an IV for $\ln p_i$ in the second one

Finding IV for first equation is tricky... looks like we don't have one because there are no extra variables in the second equation; here's the trick:

$D_i = \beta_0 + \beta_1 p_i + \beta_2 I_i + u_i = S_i = \alpha_0 + \alpha_1 \ln p_i + \varepsilon_i$ ... we don’t have to solve for p_i to know it won't be a linear function of income $\therefore$ use I_i^2 or $\ln I_i$ as IV for price in first equation

Serial Correlation - $D_i = \beta_0 + \beta_1 p_i + \beta_2 I_i + u_i$ and $S_i = \alpha_0 + \alpha_1 p_i + \varepsilon_i$ with $\varepsilon_i = \rho\varepsilon_{i-1} + \gamma_i$

As in previous case, second equation is OK because we can use income from the first one as an IV for p_i in the second one

$D_i = S_i \Rightarrow \beta_0 + \beta_1 p_i + \beta_2 I_i + u_i = \alpha_0 + \alpha_1 p_i + \rho\varepsilon_{i-1} + \gamma_i$

Solve for p_i: $p_i = \dfrac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} - \dfrac{\beta_2}{\beta_1 - \alpha_1} I_i + \dfrac{\rho\varepsilon_{i-1} + \gamma_i - u_i}{\beta_1 - \alpha_1}$

$\therefore$ can use ε_{i-1} as IV for p_i in first equation

Lessons -

1. For linear model, need one IV for each endogenous variable; if a separate variable is not available, the IV could be one of the variables form the other equation
2. For model with nonlinear endogenous regressor will always be identified (exactly or over specified) because there will be plenty of IVs to use [just use functions of the exogenous variables]
3. Dynamic system (includes serial correlation or any lagged regressor or regressand) will always be identified

Need Right Number of IVs -

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$ with $E(x_i u_i) \neq 0$

Assume z_i is IV for x_i

Wrong Way - use 2SLS

1. Regress x_i on 1, z_i and get $\hat{x}_i$

2. Run $y_i = \beta_0 + \beta_1 \hat{x}_i + \beta_2 (\hat{x}_i)^2 + u_i$

This is <u>not</u> consistent (i.e., $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ will be biased)

Right Way - need 2 IVs... since z_i is correlated to x_i, we know z_i^2 is correlated to x_i^2

1a. Regress x_i on 1, z_i, z_i^2 and get $\hat{x}_i$

1b. Regress x_i^2 on 1, z_i, z_i^2 and get $\hat{x}_i^2$ (this is <u>not</u> the same as $(\hat{x}_i)^2$)

2. Run $y_i = \beta_0 + \beta_1 \hat{x}_i + \beta_2 \hat{x}_i^2 + u_i$

Panel Data

Pooled Time Series - (review) cross-section data that is collected over time from different people in each period

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}$$

$t = 1, \ldots, T$ (# Periods)

$i = 1, \ldots, N_t$ (# Individuals, could vary by period)

Pooled Implicit Assumption - if we run all data together in a pooled regression (SURE model) we are assuming individuals behave the same over time (i.e., same β_j regardless of individual or time period)

Individual Regressions - if rather than using the SURE model, we run a separate regression for each time period, we're effectively assuming different parameters for each time period:

$$y_{it} = \beta_{0t} + \beta_{1t} x_{1it} + \beta_{2t} x_{2it} + \cdots + \beta_{kt} x_{kit} + u_{it}$$

In Between - if we set up the data correctly (i.e., keep track of time period), we can use the SURE model with GLS (or 3SLS) to impose the restriction that some coefficients don't change (but we can let others change over time)

Panel Data - repeated observations on the same sample

Pooled OK - can still use the pooled approach, but with panel data there's more reason to believe coefficients are the same because people are less likely to change tastes from year to year (unless the time period covers a longer period like decades)

Less Restrictive - can allow parameters to change based on individual (index i):

$$y_{it} = \beta_{0i} + \beta_{1i} x_{1it} + \beta_{2i} x_{2it} + \cdots + \beta_{ki} x_{kit} + u_{it}$$

Problem - nice in theory that we can do that, but we usually don't have enough data on individuals

Alternative - assume slope coefficients are the same, but intercept can change:

$$y_{it} = \beta_{0i} + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}$$

More General - could also run model with some parameters constant and others changing... more on this later

Balanced Panel - same number of observations for each individual

Panel Regression Model - assuming we only let the intercept term change, β_{0i} captures effect of things (observed and unobserved) that don't vary over time (e.g., ethnicity, gender, firm's management style)

Theory - $\beta_{0i} = \beta_0 + \eta_i$ (a constant, average plus an unobserved component that changes over time); add this to the model:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \varepsilon_{it}, \text{ where } \varepsilon_{it} = u_{it} + \eta_i$$

Regression - since we have all the same parameters now, we could use the SURE model to solve the model... but there's a problem

Problem 1 - ε_{it} are correlated (e.g., $\varepsilon_{it_1} = u_{it_1} + \eta_i$ and $\varepsilon_{it_2} = u_{it_2} + \eta_i$)

Problem 2 - people are different from each other; chance that difference is based on other regressors (x_{1it}, x_{2it}, etc.) $\therefore$ could have regressors correlated to η_i which means they're also correlated with error term ε_{it}

Solution - only use x_{1it}, x_{2it}, etc. for variables that change over time (for at least some people), then break β_{0i} into variables that are constant for individuals over time and an unobserved component: $\beta_{0i} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$

Examples - $\mathbf{z}_i$ could contain data such as ethnicity, gender, education, religion

Random Effect Model - sub $\beta_{0i} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$ into original model:

$y_{it} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \varepsilon_{it}$, where $\varepsilon_{it} = u_{it} + \eta_i$

Solving - can run individual regression for each time period or use pooled approach (i.e., SURE model)

Assumptions - from the original model we have: $E(u_{it}) = E(u_{it} x_{jit}) = 0$

New - if we account for all observable factors, then we also have $E(\eta_i) = E(\eta_i x_{jit}) = 0$ and $E(u_{it}\mathbf{z}_i) = E(\eta_i \mathbf{z}_i) = \mathbf{0}$ (i.e., random [unobserved] part [η_i] is not correlated with regressors and both the random part and the error term are not correlated to the regressors that don't change over time [$\mathbf{z}_i$])... this is a very strong assumption (i.e., unrealistic)

Real World - η_i will probably be correlated (especially with $\mathbf{z}_i$); if η_i is correlated with at least one regressor, we need to use the fixed effect model (can't do least squares regression)

Fixed Effect Model -

$$y_{it} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + (u_{it} + \eta_i) \quad \text{or}$$

$$y_{it} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + (u_{it} + \eta_i) \quad \text{or}$$

$$y_{it} = \beta_{0i} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it} \quad \text{where } \beta_{0i} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$$

$\mathbf{x}_{it}$ is a vector of observed regressors that do change for a given individual over time

$\mathbf{z}_i$ is a vector of observed regressors that do <u>not</u> change for a given individual over time

η_i captures unobserved qualities for a given individual that do not change over time... this will lead to problems because we can't tell it apart from α_0 or $\mathbf{z}_i$

Assumptions -

1. u_{it} uncorrelated with time varying regressors in each period: $E(u_{it}\mathbf{x}_{is}) = \mathbf{0}$ $\forall$ s = 1, ..., T

   A stronger way to write this is $E(u_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}) = 0$

2. η_i is allowed to be correlated with regressors: $E(\eta_i \mathbf{x}_{it}) \neq \mathbf{0}$ and $E(\eta_i \mathbf{z}_i) \neq \mathbf{0}$

3a. u_{it} is serially uncorrelated: $E(u_{it}u_{is}) = 0 \ \forall \ s \neq t$, and

3b. u_{it} is not correlated between individuals: $E(u_{it}u_{jt}) = 0 \ \forall \ i \neq j$

   This is not a critical assumption; the first is frequently violated for time series and the second would be violated by having businesses in the same location

4. Homoskedasticity: $E(u_{it}^2) = \sigma^2$

Note: no restriction imposed on correlation between η_i and regressors (#2)... this is what makes fixed effect different than random effect estimation

Identification - α_0 and $\boldsymbol{\alpha}$ are not identified $\therefore$ define fixed effect

Fixed Effect - $\beta_{0i} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$; individual specific intercept; note that α_0 will be absorbed by η_i so we can't tell them apart in practice

Book Assumption - some books list $E(\beta_{0i}\mathbf{x}_{it}) \neq \mathbf{0}$ as the assumption for a fixed effect model, but what they really mean is assumption 2 above

Still Not Identified - $\boldsymbol{\alpha}$ still aren't identified; for example, we can "shift it around" and have any $\boldsymbol{\alpha}$ we want: $\beta_{0i} = \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i = \mathbf{z}_i'\overline{\boldsymbol{\alpha}} + (\eta_i - \mathbf{z}_i'(\boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}}))$

Purpose of Study - if the focus of the study is $\boldsymbol{\alpha}$ (e.g., discrimination study), we can't use fixed effect estimation; if the focus is on $\boldsymbol{\beta}$, fixed effect estimation is better because it has fewer assumptions than random effect model (2 uses ≠ instead of =)

Method - add dummy variable for each individual; example for person 1 (will have N_1 of these):

$$d_{1it} \begin{cases} 1 \text{ for } i = 1 \\ 0 \text{ otherwise} \end{cases}$$

Note: normally for dummy variables, we add 1 less than the number needed, but in this case there is no constant term (other than β_{0i} which is different for each individual) so we need to add a dummy for each individual

Example - assume 3 years of data (i.e., $T = 3$)

y_{it}	x_{1it}	...	x_{kit}	d_{1it}	d_{2it}	...	d_{Nit}
y_{11}	x_{111}	...	x_{k11}	1	0	...	0
y_{12}	x_{112}	...	x_{k12}	1	0	...	0
y_{13}	x_{113}	...	x_{k13}	1	0	...	0
y_{21}	x_{121}	...	x_{k21}	0	1	...	0
y_{22}	x_{122}	...	x_{k22}	0	1	...	0
y_{23}	x_{123}	...	x_{k23}	0	1	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
y_{N3}	x_{1N3}	...	x_{kN3}	0	0	...	1

Data Warning - several warnings actually:

1. Need to have data that is complete for the entire time period for any given individual; if data on any regressor is mission, that entire observation should be removed
2. Need to have at least 2 repeat observations for each individual (i.e., two time periods); they don't have to be consecutive, but need more than one or else β_{0j} for that individual can't be identified

   Example - consider data for 3 regressors below for individual 1 over time $t = 1,2,3$.

   In this case, we'd drop the observation for $t = 2$, but keep individual 1 in the data because we have 2 good time periods.

t	x_{11t}	x_{21t}	x_{31t}
1	2	5	7
2	·	5	6
3	3	4	6

Model - $y_{it} = \sum_{j=1}^{N} \beta_{0j} d_{jit} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$

Solving - because of assumptions about u_{it}, we don't need to use the SURE model; we can solve this using OLS

Problem - number of coefficients is not fixed ($k+N$ of them); if $N \to \infty$ then the number of coefficients also $\to \infty$ so asymptotic properties are not consistent for β_{0j}, but will be consistent for $\boldsymbol{\beta}$ ($\beta_1, \ldots, \beta_k$). Usually we're only interested in $\boldsymbol{\beta}$ anyway so this isn't a big issue

In Practice -

Start with: $y_{it} = \sum_{j=1}^{N} \beta_{0j} d_{jit} + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}$

Take average wrt time: $y_{i\bullet} = \sum_{j=1}^{N} \beta_{0j} d_{ji\bullet} + \beta_1 x_{1i\bullet} + \beta_2 x_{2i\bullet} + \cdots + \beta_k x_{ki\bullet} + u_{i\bullet}$

where $y_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $d_{ji\bullet} = \frac{1}{T}\sum_{t=1}^{T} d_{jit} = d_{jit}$, $x_{mi\bullet} = \frac{1}{T}\sum_{t=1}^{T} x_{mit}$ $(m = 1, \ldots, k)$

Get rid of β_{0j} by subtracting second model from first:

$$(y_{it} - y_{i\bullet}) = (\beta_1 x_{1it} - \beta_1 x_{1i\bullet}) + \cdots + ((\beta_k x_{kit} - \beta_k x_{ki\bullet}) + (u_{it} - u_{i\bullet})$$

Now run OLS on the transformed data (don't need SURE or GLS)

With Vectors - same thing, different notation:

Start with: $y_{it} = \sum_{j=1}^{N} \beta_{0j} d_{jit} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$

Take average wrt time: $y_{i\bullet} = \sum_{j=1}^{N} \beta_{0j} d_{ji\bullet} + \mathbf{x}_{i\bullet}'\boldsymbol{\beta} + u_{i\bullet}$

where $y_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $d_{ji\bullet} = \frac{1}{T}\sum_{t=1}^{T} d_{jit} = d_{jit}$, $\mathbf{x}_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$

Get rid of β_{0j} by subtracting: $(y_{it} - y_{i\bullet}) = (\mathbf{x}_{it} - \mathbf{x}_{i\bullet})'\boldsymbol{\beta} + (u_{it} - u_{i\bullet})$

Within Regression - this technique is called within regression because it uses the variation over time for each individual; OLS estimate of $\hat{\boldsymbol{\beta}}$ in the within regression model is the fixed effect estimate

Problem - OLS will be numerically the same as OLS on big model with all the dummy variables, but the standard error will be wrong because packages will use the wrong degrees of freedom (if you run OLS on within regression model rather than use specific panel data regression command); there are NT observations* ; transformed data appears to only have k parameters so standard software packages will use $NT-k$ df, but should be using $NT-(k+N)$ df (the previous df will give standard errors that are too small)

Fix for Stata - take OLS standard error and multiply by $\sqrt{\frac{NT-k}{NT-k-N}}$

(but if we use `xtreg`, Stata does this automatically)

* assuming we have same # observations per individual and same number of individuals each time period; technically could have $\sum_{i=1}^{m} T_i$ (i.e., add up time periods for each individual)

Statistical Properties -

1. For $\hat{\boldsymbol{\beta}}$ to be consistent, we must have $E[(\mathbf{x}_{it} - \mathbf{x}_{i\bullet})(u_{it} - u_{i\bullet})] = \mathbf{0} \ \Rightarrow$ $E[u_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots \mathbf{x}_{iT_i}] = 0$ (i.e., each error term for a given time period for an individual is uncorrelated with all the regressors for that individual over all time periods... very strong assumption)... this is addressed in difference approach p.9
2. For consistency, also can't use any lagged dependant variable (y_{it}) as a regressor (on RHS)

Stata - `xtreg scrap grant, fe i(fcode)`

For more info on Stata command see Stata Notes

Time Dummies - can include time dummies to capture macro economic trend (if there aren't too many time periods)

e.g., `xtreg scrap grant d88 d89, fe i(fcode)`

Heteroskedasticity - assumption 4 is homoskedasticity: $E(u_{it}^2) = \sigma^2$

Heteroskedasticity - $E(u_{it}^2 \mid \mathbf{h}_{it}) = \sigma_{it}^2$ (could depend on time or individual); $\mathbf{h}_{it}$ is vector of regressors that are suspected to be correlated with u_{it}

Easy Solution - there's a formula similar to the White Heteroskedasticity Consistent Estimator... but Stata doesn't have it so we have to do it the hard way

Detecting -

1. Run xtreg to get $\hat{u}_{it}$ ... this is actually $\widehat{u_{it} - u_{i\bullet}}$, but the notation is easier with $\hat{u}_{it}$
2. Regress $\hat{u}_{it}^2$ on 1, $\mathbf{h}_{it}$ and test coefficients

Fixing -

1. Transform data:

$$\frac{y_{it}}{\sqrt{\hat{a}_0 + \mathbf{h}_{it}'\hat{\boldsymbol{\alpha}}}} = \frac{\mathbf{x}_{it}'}{\sqrt{\hat{a}_0 + \mathbf{h}_{it}'\hat{\boldsymbol{\alpha}}}}\boldsymbol{\beta} + \frac{\beta_{0i}}{\sqrt{\hat{a}_0 + \mathbf{h}_{it}'\hat{\boldsymbol{\alpha}}}}$$

Note: this last term won't be constant anymore

Note: GLS fix shown below deals with heteroskedasticity wrt time only (not individual)

Serial Correlation - assumption 3a: $E(u_{it}u_{is}) = 0 \ \forall \ s \neq t$

Detecting -

1. Run xtreg to get $\hat{u}_{it}$
2. Run $\hat{u}_{it} = \rho \hat{u}_{it-1} + \varepsilon_{it}$

   Book Way - add lagged residual to original model as a regressor:

   $y_{it} = \beta_{0i} + \mathbf{x}_{it}'\boldsymbol{\beta} + \rho \hat{u}_{it}$ ... ρ should be insignificant

Note: GLS fix shown below deals with correlation wrt time only (not individual)

Joint Hypothesis Testing - H$_0$: $\mathbf{R\beta} = \mathbf{r}$

1. Impose restrictions and obtain fixed effect estimate $\tilde{\boldsymbol{\beta}}$
2. Generate restricted residuals: $\widetilde{u_{it} - u_{i\bullet}}$
3. Do same for unrestricted model
4. Compute F-statistic: $$F = \frac{\sum_i \sum_t \left[(\widetilde{u_{it} - u_{i\bullet}})^2 - (\widehat{u_{it} - u_{i\bullet}})^2 \right] \Big/ \overset{\text{\# restrictions}}{m}}{\sum_i \sum_t (\widehat{u_{it} - u_{i\bullet}})^2 \Big/ \left(\underset{\text{\# observations}}{NT} - \underset{\text{\# parameters}}{(N+k)} \right)}$$

Generalized Least Squares - used to fix heteroskedasticity (wrt t) &/or serial correlation (wrt t)

Assumption - $Cov(\hat{\boldsymbol{\beta}}_{\text{FE}}) = \left[\sum_i \sum_t (\mathbf{x}_{it} - \mathbf{x}_{i\bullet})(\mathbf{x}_{it} - \mathbf{x}_{i\bullet})' \right]^{-1} \sigma^2$, but if heteroskedasticity or serial correlation are present, this method for calculating $Cov(\hat{\boldsymbol{\beta}}_{\text{FE}})$ will be wrong; serial correlation is very likely

New Notation - $\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} & \mathbf{x}_{i2} & \cdots & \mathbf{x}_{iT} \end{pmatrix} = \begin{bmatrix} x_{1i1} & x_{1i2} & \cdots & x_{1iT} \\ x_{2i1} & x_{2i2} & \cdots & x_{2iT} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ki1} & x_{ki2} & \cdots & x_{kiT} \end{bmatrix}$ and $\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}$

a $k \times T$ matrix listing regressors by row and time periods by column for individual i and a $T \times 1$ vector listing dependent variable for individual i

Now we have $Cov(\hat{\boldsymbol{\beta}}_{\text{FE}}) = \left[\sum_i \underset{k\times T}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})} \underset{T\times k}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})'} \right]^{-1} \sigma^2$

Fixed $Cov(\hat{\boldsymbol{\beta}}_{\text{FE}})$ - similar to White Heteroskedasticity Consistent Covariance Estimator

$$Cov(\hat{\boldsymbol{\beta}}_{\text{FE}}) = \left[\sum_i \underset{k\times T}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})} \underset{T\times k}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})'} \right]^{-1} \left[\sum_i \underset{k\times T}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})} \underset{T\times T}{\boldsymbol{\Omega}} \underset{T\times k}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})'} \right] \left[\sum_i \underset{k\times T}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})} \underset{T\times k}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})'} \right]^{-1}$$

$\boldsymbol{\Omega} = E(\mathbf{u}_i \mathbf{u}_i')$ ... where $\mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}$

Standard Assumption (from above) is $\boldsymbol{\Omega} = E(\mathbf{u}_i \mathbf{u}_i') = \sigma^2 \mathbf{I}_T$

GLS Estimate - supposed to be better than fixed effect estimate

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\sum_i \underset{k\times T}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})} \underset{T\times T}{\boldsymbol{\Omega}^{-1}} \underset{T\times k}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})'} \right]^{-1} \left[\sum_i \underset{k\times T}{(\mathbf{X}_i - \mathbf{X}_{i\bullet})} \underset{T\times T}{\boldsymbol{\Omega}^{-1}} \underset{T\times 1}{(\mathbf{y}_i - \mathbf{y}_{i\bullet})} \right]$$

Run SURE Model - combine all data for individual i: $(\mathbf{y}_i - \mathbf{y}_{i\bullet}) = (\mathbf{X}_i - \mathbf{X}_{i\bullet})'\boldsymbol{\beta} + (\mathbf{u}_i - \mathbf{u}_{i\bullet})$

$T \times 1$ $T \times k$ $k \times 1$ $T \times 1$

Stata - sureg $([y_{i1} - y_{i\bullet}]\ [\mathbf{x}_{i1} - \mathbf{x}_{i\bullet}])\ ([y_{i2} - y_{i\bullet}]\ [\mathbf{x}_{i2} - \mathbf{x}_{i\bullet}])\ \ldots\ T$ equations

** Need to add restriction that coefficients are the same... Ai: "it's not going to be easy"

Example - (rough coding)

```
generate scrap_idot = average(scrap) if fcode == i
```
(finding $y_{i\bullet}$)

```
generate dm_scrap_i1 = scrap - scrap_idot if time == 1
```
(computing $y_{i1} - y_{i\bullet}$)

```
sureg (dm_scrap_i1 dm_grant_i1 dm_price_i1) (dm_scrap_i2
   dm_grant_i2 dm_price_i2)
```
... etc.

Cluster - another way to fix serial correlation, is to use regular OLS on de-meaned variables with cluster (grouped by individual)

```
regress dm_scrap dm_grant dm_price, cluster(fcode)
```

Limitation - must have same number of observations for each individual (balanced panel)

Extension of Fixed Effect Model

- problem with FE estimate is that we can't consistently estimate coefficient of time constant regressors ($\mathbf{z}_i$); but for some studies (e.g., discrimination), these are the regressors we're most interested in; the solution is to impose an additional restriction to the FE model:

$$y_{it} = \beta_{0i} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it} \ \Rightarrow\ y_{it} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + (u_{it} + \eta_i)$$

Additional Restrictions -

1. u_{it} uncorrelated with time constant regressors: $E(u_{it}\mathbf{z}_i) = \mathbf{0}$
2. η_i is cannot be correlated with time constant regressors: $E(\eta_i\mathbf{z}_i) = \mathbf{0}$

   This is a modificaiton of FE assumption 2; still allow $E(\eta_i\mathbf{x}_{it}) \neq \mathbf{0}$ which makes this more general than random effect model

Estimating $\boldsymbol{\alpha}$ -

1. Get $\hat{\boldsymbol{\beta}}_{\text{FE}}$
2. Regress $y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{\text{FE}}$ (dependent variable) on 1, $\mathbf{z}_i$ (regressors plus a constant) to get $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$

Problem - OLS estimates for $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ are consistent, but using $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$ standard error isn't right

Solution - "next semester" (generalized method moment)

$$\mathbf{x}_{it} = \begin{bmatrix} p_{it} \\ I_{it} \end{bmatrix} \qquad \mathbf{w}_i = \begin{bmatrix} \left.\begin{matrix} p_{i1} \\ I_{i1} \end{matrix}\right\} \mathbf{x}_{i1} \\ \left.\begin{matrix} p_{i2} \\ I_{i2} \end{matrix}\right\} \mathbf{x}_{i2} \\ \vdots \end{bmatrix}$$

Other Specification - define individual profile: $\underset{Tk \times 1}{\mathbf{w}_i} = [\underset{1 \times k}{\mathbf{x}_{i1}'}\ \ \mathbf{x}_{i2}'\ \ \cdots\ \ \mathbf{x}_{iT}']'$

Rewrite Model - $\beta_{0i} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha}_1 + \mathbf{w}_i'\boldsymbol{\alpha}_2 + \varepsilon_i$, so now

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha}_1 + \mathbf{w}_i'\boldsymbol{\alpha}_2 + (\varepsilon_i + u_{it})$$

Assumptions - $E(\varepsilon_i\mathbf{z}_i) = \mathbf{0}$ and $E(\varepsilon_i\mathbf{w}_i) = \mathbf{0}$

Ai - "don't see single application of this"; very similar to random effect model; this model implies fixed effect model uses omitted variable $\mathbf{w}_i$

Random Effect Model -

$$y_{it} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + (u_{it} + \eta_i)$$

Assumptions -

1. u_{it} uncorrelated with all regressors: $E(u_{it}\mathbf{x}_{it}) = \mathbf{0}$ and $E(u_{it}\mathbf{z}_i) = \mathbf{0}$
2. η_i is uncorrelated with all regressors: $E(\eta_i\mathbf{x}_{it}) = \mathbf{0}$ and $E(\eta_i\mathbf{z}_i) = \mathbf{0}$

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{z}_i' & \mathbf{x}_{i1}' \\ 1 & \mathbf{z}_i' & \mathbf{x}_{i2}' \\ \vdots & \vdots & \vdots \\ 1 & \mathbf{z}_i' & \mathbf{x}_{iT}' \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} u_{i1} + \eta_i \\ u_{i2} + \eta_i \\ \vdots \\ u_{iT} + \eta_i \end{bmatrix}$$

Problem - looks like SURE model and like that model, the error terms are correlated:

By assumption: $E(u_{it} + \eta_i) = 0$, $E(u_{it}u_{is}) = 0$, $E(u_{it}^2) = \sigma_u^2$, $E(\eta_i^2) = \sigma_\eta^2$

Correlation $E[(u_{it} + \eta_i)(u_{is} + \eta_i)] = E(u_{it}u_{is}) + E(\eta_i u_{is}) + E(u_{it}\eta_i) + E(\eta_i^2) = E(\eta_i^2) = \sigma_\eta^2$

Variance $E[(u_{it} + \eta_i)^2] = E(u_{it}^2) + 2E(u_{it}\eta_i) + E(\eta_i^2) = \sigma_u^2 + \sigma_\eta^2$

$$Cov\begin{bmatrix} u_{i1} + \eta_i \\ u_{i2} + \eta_i \\ \vdots \\ u_{iT} + \eta_i \end{bmatrix} = \begin{bmatrix} \sigma_u^2 + \sigma_\eta^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_u^2 + \sigma_\eta^2 & \cdots & \sigma_\eta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\eta^2 & \sigma_\eta^2 & \cdots & \sigma_u^2 + \sigma_\eta^2 \end{bmatrix}$$

Trade off - fixed effect model always gives consistent $\hat{\boldsymbol{\beta}}$; random effect model allows us to get $\hat{\boldsymbol{\alpha}}$ but will not give consistent $\hat{\boldsymbol{\beta}}$ if η_i correlated with any regressors; no way tot test for this, but could tell which technique gives better estimator...

Hausman Test - $\hat{\boldsymbol{\beta}}_{\text{FE}}$ always consistent; when consistent $\hat{\boldsymbol{\beta}}_{\text{RE}}$ (using GLS) is best... can do this in Stata

Difference Approach - in practice, there are lots of complaints about the fixed effect technique

Measurement Error - if time varying regressors ($\mathbf{x}_{it}$) don't change much over time, de-meaned values are very small numbers; measurement error becomes more pronounced ($\hat{\boldsymbol{\beta}}_{\text{FE}}$ will be biased)

Difference Approach - only use first and last period for an individual; idea is that there will be a bigger difference over time so measurement error won't be a big a deal

$\Delta y_{it} = \Delta\mathbf{x}_{it}'\boldsymbol{\beta} + \Delta u_{it}$ or $y_{iT} - y_{i1} = (\mathbf{x}_{iT} - \mathbf{x}_{i1})'\boldsymbol{\beta} + (u_{iT} - u_{i1})$

Too Long - want different in time to be big enough to overpower measurement error, but if time period is too long there could be fundamental changes that invalidate model

Another Reason - for $\hat{\boldsymbol{\beta}}_{\text{FE}}$ to be consistent, we must have $E[u_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots \mathbf{x}_{iT_i}] = 0$ and $E[u_{it}\mathbf{x}_{it}] = \mathbf{0}$; the first one is often is not satisfied

Dynamic Model - if there is any lagged dependent variable as a regressor (i.e., y_{it-1} on RHS), then within regression doesn't work because $(\mathbf{x}_{it} - \mathbf{x}_{i\bullet})$ and $(u_{it} - u_{i\bullet})$ will be correlated

Solution - find instrument for lagged dependent variable: $\mathbf{w}_{it}$ (may have some of the original $\mathbf{x}_{it}$ regressors, but has instruments for those that are correlated to the error term [included lagged dependent variable])... this technique actually solves problem of lagged dependent variables and endogeneity of $\mathbf{x}_{it}$

IV Assumption - $E[\mathbf{w}_{it}(u_{it} - u_{i\bullet})] = 0$ ... this is implied by $E[u_{it} \mid \mathbf{w}_{i1}, \mathbf{w}_{i2}, \ldots \mathbf{w}_{iT}] = 0$

Difference Approach - it's usually very hard to find IVs that satisfy this condition so use difference approach instead of de-mean (within regression): $\Delta y_{it} = \Delta \mathbf{x}_{it}'\boldsymbol{\beta} + \Delta u_{it}$ and use $\mathbf{w}_{it}$ as IV for $\Delta \mathbf{x}_{it}$ ... need to have $E(\mathbf{w}_{it}\Delta \mathbf{x}_{it}) \neq 0$ and $E(\mathbf{w}_{it}\Delta u_{it}) = 0$

Stata - `xtreg` has difference option

Summary $y_{it} = \beta_{0i} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$ where $\beta_{0i} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$

(If there are inconsistencies between notes and this page, trust this page)

Fixed Effect Model -

Assumptions -

1. u_{it} uncorrelated with past, current, and future time varying regressors: $E(u_{it}\mathbf{x}_{is}) = \mathbf{0}\ \forall\ s$ = 1, ..., T (stronger way to write this is $E(u_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}) = 0$)
2. η_i can be correlated with regressors: $E(\eta_i \mathbf{x}_{it}) \neq \mathbf{0}$ and $E(\eta_i \mathbf{z}_i) \neq \mathbf{0}$

Technique - Within Regression - OLS on de-meaned data:
$(y_{it} - y_{i\bullet}) = (\mathbf{x}_{it} - \mathbf{x}_{i\bullet})'\boldsymbol{\beta} + (u_{it} - u_{i\bullet})$ ... but need to fix standard error by multiplying by: $\sqrt{\dfrac{NT-k}{NT-k-N}}$

Others -

GLS - solve heteroskedasticity (wrt *t*) and/or serial correlation (wrt *t*) by running SURE model on de-meaned data treating each individual as a separate equation, but force all equations to use the same parameters

Cluster - run OLS on de-meaned data and cluster on individuals

Good - $\hat{\boldsymbol{\beta}}_{\text{FE}}$ consistent as long as assumption 1 holds

Bad - can't consistently estimate β_{0j} (i.e., $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$)

Extension of Fixed Effect Model -

Additional Assumption -

1. u_{it} & η_i uncorrelated with time constant regressors: $E(u_{it}\mathbf{z}_i) = \mathbf{0}$ & $E(\eta_i \mathbf{z}_i) = \mathbf{0}$ required to identify α_0 & $\boldsymbol{\alpha}$; η_i can be correlated with $\mathbf{x}_{it}$: $E(\eta_i \mathbf{x}_{it}) \neq \mathbf{0}$

Technique -

1. Get $\hat{\boldsymbol{\beta}}_{\text{FE}}$ (same as fixed effect model so we still need FE assumption 1)
2. Regress $y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{\text{FE}}$ on 1 and $\mathbf{z}_i$ to get $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ (using OLS)

Good - $\hat{\boldsymbol{\beta}}_{\text{FE}}$ consistent if FE assumption 1 holds; $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ consistent as long as additional assumptions hold

Bad - $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ have wrong standard error (not as easily fixed as $\hat{\boldsymbol{\beta}}_{\text{FE}}$ standard error)

Random Effect Model -

Assumptions -

1. u_{it} uncorrelated with all regressors: $E(u_{it}\mathbf{x}_{it}) = \mathbf{0}$ and $E(u_{it}\mathbf{z}_i) = \mathbf{0}$
2. η_i uncorrelated with all regressors: $E(\eta_i \mathbf{x}_{it}) = \mathbf{0}$ and $E(\eta_i \mathbf{z}_i) = \mathbf{0}$
3. For GLS also need FE assumption 1

Technique - run OLS or GLS on $y_{it} = \alpha_0 + \mathbf{z}_i'\boldsymbol{\alpha} + \mathbf{x}_{it}'\boldsymbol{\beta} + (u_{it} + \eta_i)$

Good - $\hat{\boldsymbol{\beta}}_{\text{RE}}$, $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ consistent with OLS as long as assumptions 1 and 2 hold (will have correlated errors though); all estimates are consistent and best with GLS

Bad - If either assumption 1 or 2 fails, no estimates are consistent

Other Assumptions - all three models require these for computing standard error (*t*-ratios), but not for consistent estimates

3a. u_{it} is serially uncorrelated: $E(u_{it}u_{is}) = 0\ \forall\ s \neq t$, and

3b. u_{it} is not correlated between individuals: $E(u_{it}u_{jt}) = 0\ \forall\ i \neq j$

4. Homoskedasticity: $E(u_{it}^2) = \sigma^2$

4.7. Consider estimating the effect of personal computer ownership, as represented by a binary variable, PC, on college GPA, $colGPA$. With data on SAT scores and high school GPA you postulate the model

$$ColGPA = \beta_0 + \beta_1 hsGPA + \beta_2 SAT + \beta_3 PC + u$$

a. Why might u and PC be positively correlated?

b. If the given equation is estimated by OLS using a random sample of college students, is $\hat{\beta}_3$ likely to have an upward or downward asymptotic bias?

c. What are some variables that might be good proxies for the unobservables in u that are correlated with PC?

a. PC is most likely positively correlated with family income which isn't included in the model.

b. p. 149 of Green: "If more than one variable is included, then the terms in the omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations.".. that means it's hard to figure out what the bias will be in the coefficients. Best guess is that it'll be biased upward (too high) because it'll also capture the effects of the missing variables that are correlated with PC.

c. Other variables that might be correlated with PC:

Family Income
Student Income
Student Loan (binary)... money to buy a computer
Foreign Student (binary)... less likely to have a computer?

Example 4.3. (*Using IQ as a Proxy for Ability*): We apply the proxy variable method to the data on working men in NLS80.RAW, which was used by Blackburn and Neumark (1992), to estimate the structural model

$$\log(wage) = \beta_0 + \beta_1 \exp er + \beta_2 tenure + \beta_3 married + \beta_4 south + \beta_5 urban$$
$$+ \beta_6 black + \beta_7 educ + \gamma abil + u \qquad (4.29)$$

where $exper$ is labor market experience, $married$ is a dummy variable equal to unity if married, $south$ is a dummy variable for the southern region, $urban$ is a dummy variable for living in an SMSA, $black$ is a race indicator, and $educ$ is years of schooling. We assume that IQ satisfies the proxy variable assumptions: in the linear projection $abil = \theta_0 + \theta_1 IQ + r$, where r has zero mean and is uncorrelated with IQ, we also assume that r is uncorrelated with experience, tenure, education, and other factors appearing in equation (4.29). The estimated equations without and with IQ are

$$\log(wage) = \underset{(0.11)}{5.40} + \underset{(.003)}{.014}\, exper + \underset{(.002)}{.012}\, tenure + \underset{(.039)}{.199}\, married$$

$$- \underset{(.026)}{.091}\, south + \underset{(.027)}{.184}\, urban - \underset{(.038)}{.188}\, black + \underset{(.006)}{.065}\, educ$$

N = 935, R^2 = .253

$$\begin{array}{lllll}
\log(wage) = & 5.18 & +\ .014\ exper & +\ .011\ tenure & +\ .200\ married \\
 & (0.13) & (.003) & (.002) & (.039)
\end{array}$$

$$\begin{array}{lllll}
-\ .080\ south & +\ .182\ urban & -\ .143\ black & +\ .054\ educ & +\ .0036\ IQ \\
(.026) & (.027) & (.039) & (.007) & (.0010)
\end{array}$$

$N = 935$, $R^2 = .263$

Notice how the return to schooling has fallen from about 6.5 percent to about 5.4 percent when *IQ* is added to the regression. This is what we expect to happen if ability and schooling are (partially) positively correlated. Of course, these are just the finding from one sample. Adding *IQ* explains only one percentage point more of the variation in log(*wage*), and the equation predicts that 15 more *IQ* points (one standard deviation) increases wage by about 5.4 percent. The standard error of the return to education has increased, but the 95 percent confidence interval is still fairly tight.

The data set NLS80.RAW also contains each man's score on the knowledge of the world of work (*KWW*) test. Problem 4.11 asks you to reestimate equation (4.29) when *KWW* and *IQ* are both used as proxies for ability.

4.11. a. In example 4.3, use *KWW* and *IQ* simultaneously as proxies for ability in equation (4.29). Compare the estimated return to education without a proxy for ability and with *IQ* as the only proxy for ability.
b. Test *KWW* and *IQ* for joint significance in the estimated equation from part a.
c. When *KWW* and *IQ* are used as proxies for *abil*, does the wage differential between nonblacks and blacks disappear? What is the estimated differential?
d. Add the interactions $educ(IQ - 100)$ and $educ(KWW - \overline{KWW})$ to the regression from part a, where $\overline{KWW}$ is the average score in the sample. Are these terms jointly significant using a standard *F* test? Does adding them affect any important conclusions?

a. The return to education drops quite a bit form the original model without a proxy for ability (from 0.065 to 0.049). We can say this is about a 25% drop (or only 1.6% points, depending on what we're trying to get across... I love letting statistics lie for me!). The difference isn't as big compared to the regression that already included *IQ* as a proxy. It's important to note, however, that the R^2 value barely increased (from 0.263 to 0.266)... just for grins, the adjusted R^2 also increased (from 0.2564 to 0.2591). Fortunately, it appears *IQ* is still significant with roughly the same effect (0.0036 vs. 0.0031) and the other parameters did not change too much (as they would if there was a lot of correlation between them and *KWW*).

```
use NLS80
regress lwage exper tenure married south urban black educ iq kww
```

```
Source          SS        df        MS         Number of obs  =     935
                                               F(  9,    925)  =   37.28
Model      44.0967944     9  4.89964382        Prob > F        =  0.0000
Residual   121.559489   925  .131415664        R-squared       =  0.2662
                                               Adj R-squared   =  0.2591
Total      165.656283   934  .177362188        Root MSE        =  .36251

lwage          Coef.    Std. Err.     t    P>|t|    [95% Conf. Interval

exper       .0127522  .0032308     3.95  0.000    .0064117   .0190927
tenure      .0109248  .0024457     4.47  0.000     .006125   .0157246
married     .1921449  .0389094     4.94  0.000    .1157839   .2685059
south      -.0820295  .0262222    -3.13  0.002   -.1334913  -.0305676
urban       .1758226  .0269095     6.53  0.000    .1230118   .2286334
black      -.1303995  .0399014    -3.27  0.001   -.2087073  -.0520917
educ        .0498375   .007262     6.86  0.000    .0355856   .0640893
iq          .0031183  .0010128     3.08  0.002    .0011306   .0051059
kww          .003826  .0018521     2.07  0.039    .0001911   .0074608
_cons       5.175644   .127776    40.51  0.000    4.924879   5.426408
```

b. The F-test is very significant (p-value = 0.0002) $\therefore$ IQ and KWW jointly add something to the model.

```
test iq kww

( 1)  iq = 0
( 2)  kww = 0

      F(  2,   925) =     8.59
           Prob > F =    0.0002
```

c. The estimated differential is still significant (p-value 0.001). The differential is -0.13 which suggests blacks get paid 13% less than nonblacks.

d. The terms are jointly significant at the 98% level (p-value 0.0154), although *educiq* is not significant on its own (p-value 0.788). Adding these interactions further reduced the impact of *educ* on ln(*wage*). Most of the parameters are fairly close to their values in part a, except for *kww* which changed signs. Even though we differenced the mean, *kww* is highly correlated with *educkww* (R^2 of 0.97!) so the parameters are suspect (i.e., not good for interpretation).

```
generate educiq = educ*(iq-100)
egen meankww = mean(kww)
generate educkww = educ*(kww - meankww)
regress lwage exper tenure married south urban black educ iq kww educiq
   educkww
```

```
Source          SS        df        MS         Number of obs  =     935
                                               F( 11,   923)  =   31.48
Model      45.1916885   11   4.10833532        Prob > F       =  0.0000
Residual   120.464595  923   .130514187        R-squared      =  0.2728
                                               Adj R-squared  =  0.2641
Total      165.656283  934   .177362188        Root MSE       =  .36127

lwage          Coef.    Std. Err      t     P>|t|    [95% Conf. Interval]

exper        .0121544  .0032358    3.76   0.000     .005804   .0185047
tenure       .0107206  .0024383    4.40   0.000    .0059353    .015506
married      .1978269  .0388272    5.10   0.000    .1216271   .2740267
south       -.0807609  .0261374   -3.09   0.002   -.1320565  -.0294652
urban         .178431   .026871    6.64   0.000    .1256957   .2311664
black       -.1381481  .0399615   -3.46   0.001   -.2165741  -.0597221
educ          .045241  .0076469    5.92   0.000    .0302338   .0602483
iq           .0048228  .0057333    0.84   0.400    -.006429   .0160745
kww         -.0248007  .0107382   -2.31   0.021   -.0458749  -.0037266
educiq      -.0001138  .0004228   -0.27   0.788   -.0009436   .0007161
educkww       .002161  .0007957    2.72   0.007    .0005994   .0037227
_cons        6.080005  .5610875   10.84   0.000    4.978849    7.18116

test educiq educkww

( 1)  educiq = 0
( 2)  educkww = 0

      F(  2,   923) =    4.19
           Prob > F =    0.0154
```

Example 4.4. (*Effects of Job Training Grants on Worker Productivity*): The data in JTRAIN1.RAW are for 157 Michigan manufacturing firms for the years 1987, 1988, and 1989. These data are from Holzer, Block, Cheatham, and Knott (1993). The goal is to determine the effectiveness of job training grants on firm productivity. For this exercise, we use only the 54 firms in 1988 which reported nonmissing values of the scrap rate (number of items out of 100 that must be scrapped). No firms were awarded grants in 1987; in 1988, 19 of the 54 firms were awarded grants. If the training grant has the intended effect, the average scrap rate should be lower among firms receiving a grant. The problem is that the grants were not randomly assigned: whether or not a firm received a grant could be related to other factors unobservable to the econometrician that affect productivity. In the simplest case, we can write (for the 1988 cross section)

$$\log(scrap) = \beta_0 + \beta_1 grant + \gamma q + u$$

where u is orthogonal to grant but q contains unobserved productivity factors that might be correlated with $grant$, a binary variable equal to unity if the firm received a job training grant. Since we have the scrap rate in the previous year, we can use $\log(scrap_{-1})$ as a proxy variable for q:

$$q = \theta_0 + \theta_1 \log(scrap_{-1}) + r$$

where r has zero mean and, by definition, is uncorrelated with $\log(scrap_{-1})$. We hope that r has no or little correlation with $grant$. Plugging in for q gives the estimable model

$$\log(scrap) = \delta_0 + \beta_1 grant + \gamma\theta_1 \log(scrap_{-1}) + r + u$$

From this equation, we see that β_1 measures the proportionate difference in scrap rates for two firms having the *same* scrap rates in the previous year, but where one firm received a grant and the other did not. This is intuitively appealing. The estimated equations are

$$\log(scrap) = \underset{(0.240)}{.409} + \underset{(.406)}{.057}\, grant$$

$N = 54$, $R^2 = .0004$

$$\log(scrap) = \underset{(0.240)}{.021} - \underset{(.406)}{.254}\, grant + \underset{(.044)}{.831} \log(scrap_{-1})$$

$N = 54$, $R^2 = .873$

Without the lagged scrap rate, we see that the grants appear, if anything, to reduce productivity (by increasing the scrap rate), although the coefficient is statistically insignificant. When the lagged dependent variable is included, the coefficient on grant changes signs, becomes economically large--firms awarded grants have scrap rates about 25.4 percent less than those not given grants--and the effect is significant at the 5 percent level against a one-sided alternative. [The more accurate estimate of the percentage effect is $100\cdot[\exp(-.254) - 1] = -22.4\%$; see Problem 4.1(a).]

4.12. Redo Example 4.4, adding the variable *union*--a dummy variable indicating whether the workers at the plant are unionized--as an additional explanatory variable.

R^2 improved slightly and the coefficient for $\log(scrap_{-1})$ didn't change much (dropped from 0.831 to 0.821). The grant appears more effective now with 28.5 percent less scrap (vs. 25.4 percent before adding the *union* term). The interesting finding from including *union* is the huge negative effect union has on productivity. (Here, I'm assuming *union* = 1 means the workers are unionized.) Having a union shop means 25.8 percent more scrap (although it's only significant at 9%)... how did the union let this data set get out?

```
use JTRAIN1
regress lscrap grant if year == 1988 & scrap != .
regress lscrap grant lscrap_1 if year == 1988 & scrap != .
regress lscrap grant lscrap_1 union if year == 1988 & scrap != .
```

```
Source          SS        df        MS         Number of obs  =      54
                                               F(  3,    50)  =  122.33
Model       92.7289733    3   30.9096578       Prob > F       =  0.0000
Residual    12.6336868   50   .252673735       R-squared      =  0.8801
                                               Adj R-squared  =  0.8729
Total        105.36266   53   1.98797472       Root MSE       =  .50267

lscrap         Coef.    Std. Err.     t    P>|t|    [95% Conf. Interval]

grant      -.2851103   .1452619   -1.96   0.055   -.5768775   .0066568
lscrap_1    .8210298    .043962   18.68   0.000    .7327295     .90933
union       .2580653   .1477832    1.75   0.087   -.0387659   .5548964
_cons      -.0477754   .0958824   -0.50   0.620   -.2403608     .14481
```

4.13. Use the data in CORNWELL.RAW (from Cornwell and Trumball, 1994) to estimate a model of county level crime rates, using the year 1987 only.
a. Using logarithms of all variables, estimate a model relating the crime rate to the deterrent variables *prbarr*, *prbconv*, *prbpris*, and *avgsen*.
b. Add $\log(crmrte)$ for 1986 as an additional explanatory variable, and comment on how the estimated elasticities differ from part a.
c. Compute the F statistic for joint significance on all of the wage variables (again in logs), using the restricted model from part b.
d. Redo part c but make the test robust to heteroskedasticity of unknown form.

a. The regression is very significant (p-value for F-test is 0.0000), but only explains 40 percent of the variation in the crime rate (R^2 = 0.4162). Only two of the regressors are significant based on their t-ratios: *lprbarr* and *lprbconv*. These both have a negative sign like we'd expect (i.e., people being more likely to be arrested or convicted lowers the crime rate). The other regressors (*lprbpris* and *lavgsen*) have the opposite sign from what we'd expect, but they are not statistically significant.

```
use CORNWELL
regress lcrmrte lprbarr lprbconv lprbpris lavgsen if year == 87

Source          SS        df        MS          Number of obs  =       90
                                                F(  4,    85)  =    15.15
Model       11.1549601     4   2.78874002       Prob > F       =   0.0000
Residual    15.6447379    85    .18405574       R-squared      =   0.4162
                                                Adj R-squared  =   0.3888
Total        26.799698    89   .301120202       Root MSE       =   .42902

lcrmrte       Coef.    Std. Err.     t     P>|t|    [95% Conf. Interval]

lprbarr    -.7239696   .1153163   -6.28   0.000    -.9532493  -.4946899
lprbconv   -.4725112   .0831078   -5.69   0.000    -.6377519  -.3072706
lprbpris    .1596698   .2064441    0.77   0.441    -.2507964    .570136
lavgsen     .0764213   .1634732    0.47   0.641    -.2486073   .4014499
_cons      -4.867922   .4315307  -11.28   0.000    -5.725921  -4.009923
```

b. The biggest change is a huge jump in R^2 from 0.4162 to 0.8715. That means adding the lagged crime rate helps explain crime rate more than those other variables did... makes sense since a high crime area will probably still be high crime next year. Another important thing to note is that all the deterrent variables now have the proper sign (negative) although *lprbconv* and *lprbpris* aren't significant (p-values 0.409 and 0.204 respectively). The parameters for *lprbarr* and *lprbconv* are much smaller than before... of course analyzing these parameter estimates isn't too important since they probably aren't consistent (i.e., are biased) because we have a lagged regressor (*lcrmrte_1*) that is most likely correlated with the error term.

```
generate lcrmrte_1 = lcrmrte[_n-1]
regress lcrmrte lprbarr lprbconv lprbpris lavgsen lcrmrte_1 if year
   == 87
```

```
Source          SS        df        MS          Number of obs   =       90
                                                F(  5,     84)  =   113.90
Model      23.3549731      5   4.67099462       Prob > F        =   0.0000
Residual    3.4447249     84    .04100863       R-squared       =   0.8715
                                                Adj R-squared   =   0.8638
Total       26.799698     89   .301120202       Root MSE        =   .20251

lcrmrte      Coef.    Std. Err.     t    P>|t|    [95% Conf. Interval]

lprbarr    -.1850424  .0627624   -2.95   0.004    -.3098523  -.0602325
lprbconv   -.0386768  .0465999   -0.83   0.409    -.1313457   .0539921
lprbpris   -.1266874  .0988505   -1.28   0.204    -.3232625   .0698876
lavgsen    -.1520228  .0782915   -1.94   0.056    -.3077141   .0036684
lcrmrte_1   .7798129  .0452114   17.25   0.000     .6899051   .8697208
_cons      -.7666256  .3130986   -2.45   0.016    -1.389257  -.1439946
```

c. There are no wage variables in (b). The data appears to have nine wage variables though so I ran the regression in (b) with those. Combined the wage variables are not significant (p-value 0.1643). None of them is significant individually either.

```
regress lcrmrte lprbarr lprbconv lprbpris lavgsen lcrmrte_1  lwcon
   lwtuc lwtrd lwfir lwser lwmfg lwfed lwsta lwloc if year == 87
```

```
Source          SS        df        MS          Number of obs   =       90
                                                F( 14,     75)  =    43.81
Model      23.8798774     14   1.70570553       Prob > F        =   0.0000
Residual   2.91982063     75   .038930942       R-squared       =   0.8911
                                                Adj R-squared   =   0.8707
Total       26.799698     89   .301120202       Root MSE        =   .19731

lcrmrte      Coef.    Std. Err.     t    P>|t|    [95% Conf. Interval]

lprbarr    -.1725122  .0659533   -2.62   0.011    -.3038978  -.0411265
lprbconv   -.0683639   .049728   -1.37   0.173    -.1674273   .0306994
lprbpris   -.2155553  .1024014   -2.11   0.039    -.4195493  -.0115614
lavgsen    -.1960546  .0844647   -2.32   0.023     -.364317  -.0277923
lcrmrte_1   .7453414  .0530331   14.05   0.000     .6396942   .8509887
lwcon      -.2850008  .1775178   -1.61   0.113    -.6386344   .0686327
lwtuc       .0641312   .134327    0.48   0.634    -.2034619   .3317244
lwtrd        .253707  .2317449    1.09   0.277    -.2079524   .7153665
lwfir      -.0835258  .1964974   -0.43   0.672    -.4749687   .3079171
lwser       .1127542  .0847427    1.33   0.187    -.0560619   .2815703
lwmfg       .0987371  .1186099    0.83   0.408    -.1375459   .3350201
lwfed       .3361278  .2453134    1.37   0.175    -.1525615   .8248172
lwsta       .0395089  .2072112    0.19   0.849    -.3732769   .4522947
lwloc      -.0369855  .3291546   -0.11   0.911    -.6926951    .618724
_cons      -3.792525  1.957472   -1.94   0.056    -7.692009   .1069592
```

```
test  lwcon lwtuc lwtrd lwfir lwser lwmfg lwfed lwsta lwloc

( 1)  lwcon = 0
( 2)  lwtuc = 0
( 3)  lwtrd = 0
( 4)  lwfir = 0
( 5)  lwser = 0
( 6)  lwmfg = 0
( 7)  lwfed = 0
( 8)  lwsta = 0
( 9)  lwloc = 0

      F(  9,    75) =    1.50
           Prob > F =    0.1643
```

d. The standard *F*-test isn't supposed to be valid under heteroskedasticity and using the White Heteroskedasticity Consistent Covariance Estimator (i.e., `robust` in Stata) doesn't change that fact. According to the book, however, Stata runs a modification of the Wald test which is valid under heteroskedasticity. With the correction, the nine wage variables are jointly significant (p-value 0.0319), but only one is significant on it's own: *lwcon*... the construction wage (higher wage results in lower crime rate).

```
regress lcrmrte lprbarr lprbconv lprbpris lavgsen lcrmrte_1  lwcon
   lwtuc lwtrd lwfir lwser lwmfg lwfed lwsta lwloc if year == 87,
   robust

Regression withrobust standard errors     Number of obs  =      90
                                          F( 14,    75)  =  110.75
                                          Prob > F       =  0.0000
                                          R-squared      =  0.8911
                                          Root MSE       =  .19731
                        Robust
lcrmrte      Coef.     Std. Err.    t    P>|t|    [95% Conf. Interval]

lprbarr   -.1725122  .0831236   -2.08   0.041   -.3381028  -.0069215
lprbconv  -.0683639  .0874696   -0.78   0.437   -.2426123   .1058844
lprbpris  -.2155553  .0895319   -2.41   0.019   -.3939121  -.0371986
lavgsen   -.1960546  .0976231   -2.01   0.048   -.3905298  -.0015795
lcrmrte_1  .7453414  .1594535    4.67   0.000    .4276937   1.062989
lwcon     -.2850008  .1276141   -2.23   0.029   -.5392212  -.0307805
lwtuc      .0641312  .1108165    0.58   0.565   -.1566265    .284889
lwtrd       .253707  .1712913    1.48   0.143   -.0875227   .5949368
lwfir     -.0835258  .1477461   -0.57   0.574    -.377851   .2107995
lwser      .1127542  .0715635    1.58   0.119   -.0298077    .255316
lwmfg      .0987371  .1083497    0.91   0.365   -.1171065   .3145807
lwfed      .3361278  .4416827    0.76   0.449   -.5437491   1.216005
lwsta      .0395089  .1829791    0.22   0.830   -.3250042    .404022
lwloc     -.0369855  .2825442   -0.13   0.896   -.5998425   .5258714
_cons     -3.792525  3.383901   -1.12   0.266    -10.5336   2.948551
```

```
test  lwcon lwtuc lwtrd lwfir lwser lwmfg lwfed lwsta lwloc

( 1)  lwcon = 0
( 2)  lwtuc = 0
( 3)  lwtrd = 0
( 4)  lwfir = 0
( 5)  lwser = 0
( 6)  lwmfg = 0
( 7)  lwfed = 0
( 8)  lwsta = 0
( 9)  lwloc = 0

      F(  9,    75) =    2.19
           Prob > F =    0.0319
```

4.14. Use the data in ATTEND.RAW to answer this question.
a. To determine the effects of attending lecture on final exam performance, estimate a model relating *stndfnl* (the standardized final exam score) to *atndrte* (the percent of lectures attended). Include the binary variables *frosh* and *soph* as explanatory variables. Interpret the coefficient on *atndrte*, and discuss its significance.
b. How confident are you that the OLS estimates from part a are estimating the causal effect of attendance? Explain.
c. As proxy variables for student ability, add to the regression *priGPA* (prior cumulative GPA) and *ACT* (achievement test score). Now what is the effect of *atndrte*? Discuss how the effect differs from that in part a.
d. What happens to the significance of the dummy variables in part c as compared with part a? Explain.
e. Add the squares of *priGPA* and *ACT* to the equation. What happens to the coefficient on *atndrte*? Are the quadratics jointly significant?
f. To test for a nonlinear effect of *atndrte*, add its square to the equation from part e. What do you conclude?

a. The model is pretty pathetic. Despite the significant *F*-test, the variables combine to only explain less than 3 percent of the variation in stndfnl (R^2 = 0.0290). Although *atndrte* is significant (p-value 0.000), it's effect on *stndfnl* is very small: 0.008 means each 1 percent increase in attendance rate improves final exam performance by less than 1 percent of the exam standard deviation.

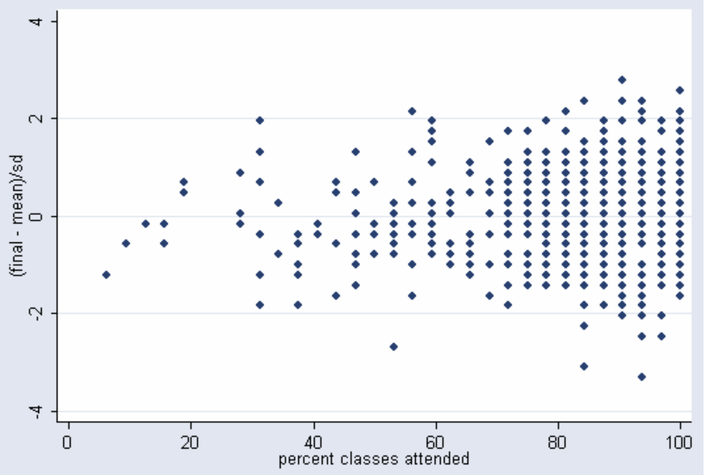
```
use ATTEND
regress stndfnl atndrte frosh soph

Source         SS        df        MS        Number of obs  =     680
                                             F(  3,   676)  =    6.74
Model      19.3023776     3    6.43412588    Prob > F       =  0.0002
Residual    645.46119   676   .954824246    R-squared      =  0.0290
                                             Adj R-squared  =  0.0247
Total      664.763568   679   .979033237    Root MSE       =  .97715
```

```
stndfnl       Coef.    Std. Err.     t    P>|t|    [95% Conf. Interval]

atndrte    .0081634   .0022031    3.71   0.000    .0038376   .0124892
frosh     -.2898943   .1157244   -2.51   0.012   -.5171168  -.0626719
soph      -.1184456   .0990267   -1.20   0.232   -.3128824   .0759913
_cons     -.5017308    .196314   -2.56   0.011   -.8871893  -.1162724
```

b. According to (a), *atndrte* has very little effect on *stndfnl*. Whether it's causal or not is irrelevant for such a small effect. Looking at a plot of *stndfnl* vs. *atndrte* shows that's there's not a very strong relationship between the two. Even if there is a relationship, the parameter estimate is probably not consistence (i.e., biased) because there are other variables that we probably omitted and should've included (student ability for example, which we add in (c)).



```
scatter stndfnl atndrte
```

c. Adding *priGPA* and *ACT* had an immense improvement on the model. The *F* statistic jumped by a factor of 5 and the R^2 value by 7. Still the model doesn't do much to explain the variation in performance ($R^2 = 0.2058$). Now *atndrte* appears to have even less effect on *stndfnl* (0.005 instead of 0.008).

```
regress stndfnl atndrte frosh soph priGPA ACT

Source          SS        df        MS         Number of obs  =      680
                                               F(  5,   674)  =    34.93
Model       136.801957     5   27.3603913      Prob > F       =   0.0000
Residual    527.961611   674   .783325833      R-squared      =   0.2058
                                               Adj R-squared  =   0.1999
Total       664.763568   679   .979033237      Root MSE       =   .88506

stndfnl       Coef.    Std. Err.     t    P>|t|    [95% Conf. Interval]

atndrte    .0052248   .0023844    2.19   0.029     .000543   .0099065
frosh     -.0494692   .1078903   -0.46   0.647   -.2613108   .1623723
soph      -.1596475   .0897716   -1.78   0.076   -.3359132   .0166181
priGPA     .4265845   .0819203    5.21   0.000    .2657348   .5874342
ACT        .0844119   .0111677    7.56   0.000    .0624843   .1063395
_cons     -3.297342    .308831  -10.68   0.000   -3.903729  -2.690956
```

d. We went from having *frosh* significant and *soph* not to neither being significant. There may have been some information about student ability captured in these variables, but it was pretty weak and once we included better proxies for that, they lost their significance to the model. In fact, I have no idea why it was suggested to include them in the first place.

e. The coefficient for *atndrte* increased slightly from 0.005 to 0.006. The quadratic terms are jointly significant (p-value 0.0000).

```
generate priGPA2 = priGPA^2
generate ACT2 = ACT^2
regress stndfnl atndrte frosh soph priGPA ACT priGPA2 ACT2

Source          SS       df       MS         Number of obs  =     680
                                             F(  7,   672)  =   28.94
Model      153.9743097  21.9963299           Prob > F       =  0.0000
Residual   510.789259   672   .760103064     R-squared      =  0.2316
                                             Adj R-squared  =  0.2236
Total      664.763568   679   .979033237     Root MSE       =  .87184

stndfnl      Coef.    Std. Err.    t    P>|t|    [95% Conf. Interval]

atndrte    .0062317  .0023583    2.64   0.008    .0016011   .0108623
frosh     -.1053368  .1069747   -0.98   0.325   -.3153817   .1047081
soph      -.1807289  .0886354   -2.04   0.042   -.3547647  -.0066932
priGPA     -1.52614  .4739715   -3.22   0.001   -2.456783  -.5954967
ACT       -.1124331   .098172   -1.15   0.253   -.3051938   .0803276
priGPA2    .3682176  .0889847    4.14   0.000    .1934961   .5429391
ACT2       .0041821  .0021689    1.93   0.054   -.0000766   .0084408
_cons      1.384812  1.239361    1.12   0.264   -1.048674   3.818298

test priGPA2 ACT2

( 1)  priGPA2 = 0
( 2)  ACT2 = 0

      F(  2,   672) =   11.30
           Prob > F =    0.0000
```

f. The squared term (*atndrte2*) is not significant (p-value 0.971).

```
generate atndrte2 = atndrte^2
regress stndfnl atndrte frosh soph priGPA ACT priGPA2 ACT2 atndrte2
```

```
Source          SS        df        MS          Number of obs  =     680
                                                F(  8,   671)  =   25.28
Model      153.975323    8    19.2469154        Prob > F       =  0.0000
Residual   510.788245   671   .761234344        R-squared      =  0.2316
                                                Adj R-squared  =  0.2225
Total      664.763568   679   .979033237        Root MSE       =  .87249

stndfnl        Coef.    Std. Err.     t     P>|t|     [95% Conf. Interval]

atndrte     .0058425   .0109203    0.54   0.593   -.0155996    .0272847
frosh      -.1053656   .1070572   -0.98   0.325   -.3155729    .1048418
soph       -.1808403   .0887539   -2.04   0.042    -.355109   -.0065716
priGPA     -1.524803    .475737   -3.21   0.001   -2.458915   -.5906903
ACT        -.1123423   .0982764   -1.14   0.253   -.3053087     .080624
priGPA2     .3679124   .0894427    4.11   0.000    .1922911    .5435337
ACT2        .0041802   .0021712    1.93   0.055   -.0000829    .0084433
atndrte2    2.87e-06   .0000787    0.04   0.971   -.0001517    .0001574
_cons       1.394292   1.267186    1.10   0.272   -1.093835     3.88242
```

Documentation.

I used my notes, the text book, and lots of help features in Stata.

5.2. Consider a model for the health of an individual:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u_1 \quad (5.53)$$

where *health* is some quantitative measure of the person's health; *age*, *weight*, *height*, and *male* are self-explanatory, *work* is weekly hours worked, and *exercise* is the hours of exercise per week.
a. Why might you be concerned about *exercise* being correlated with the error term u_1?
b. Suppose you can collect data on two additional variables, *disthome* and *distwork*, the distances from home and from work to the nearest health club or gym. Discuss whether these are likely to be uncorrelated with u_1.
c. Now assume that *disthome* and *distwork* are in fact uncorrelated with u_1, as are all variables in equation (5.53) with the exception of *exercise*. Write down the reduced form for *exercise*, and state the conditions under which the parameters of equation (5.53) are identified.
d. How can the identification assumption in part c be tested?

a. Rules of thumb for regressors being correlated to the error term: (i) LHS and RHS variables determined by simultaneous decision (e.g., Q^D_{chicken} as function of Q^D_{beef} and other factors; since chicken and beef are substitutes people's decision on how much to consume is a joint decision), (ii) omitted variable (i.e., regressor left out is captured by error term so if that omitted variable is correlated to any of the regressors in the model, the error term will be correlated to those regressors), (iii) LHS and RHS variables related by a constant (e.g., two equations for Q^D and Q^S, both as function of price; because equilibrium has $Q^D = Q^S$, price is automatically determined).

   In this case, one could argue either case (i) or (ii). For the first one, health and exercise could be jointly determined: if a person is not feeling well, he may not work out as much. In the second case, we can easily think of variables that were omitted: family history (for genetic illnesses), occupation (e.g., teachers exposed to more illnesses).

b. I can't think of any reason why *disthome* and *distwork*, would be correlated to the error term. On the other hand, there's probably a strong correlated between these variables and *exercise*, because having a health club or gym nearer to home or work would make it more likely for someone to workout (assuming they workout in a gym... I don't). A better option may be *gymonway* set to 1 if there is a gym located between work and home.

c. Structural equations (full information):

   $health = \beta_0 + \beta_1\, age + \beta_2\, weight + \beta_3\, height + \beta_4\, male + \beta_5\, work + \beta_6\, exercise + u_1$
   $exercise = \alpha_0 + \alpha_1\, age + \alpha_2\, weight + \alpha_3\, height + \alpha_4\, male + \alpha_5\, work + \alpha_6\, disthome + \alpha_7\, distwork + \alpha_8\, health + \varepsilon$

   Reduced form (sub health equation into exercise equation):

   $exercise = \pi_0 + \pi_1\, age + \pi_2\, weight + \pi_3\, height + \pi_4\, male + \pi_5\, work + \pi_6\, disthome + \pi_7\, distwork + v$

   This reduced form equation is what we use to estimate *exercise* in the first stage of 2SLS. Assuming *disthome* and *distwork* correlated to *exercise* and not correlated to u_1, then the estimate for *exercise* will not be correlated to u_1. When we plug that in for *exercise* in the second stage of 2SLS, the model (5.53) is identified (actually since we have two IVs it'll be over specified).

d. We can run a regression on the reduced form and check if *disthome* and *distwork* are significant. If at least one of them is, the we probably have a valid instrument so the model is identified. (There's no way to test if *disthome* and *distwork* are correlated to u_1.)

5.3. Consider the following model to estimate the effects of several variables, including cigarette smoking, on the weight of newborns:

$$\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + u \qquad (5.54)$$

where *male* is a binary indicator equal to one if the child is male; *parity* is the birth order of this child; *faminc* is family income; and *packs* is the average number of packs of cigarettes smoked per day during pregnancy.
a. Why might you expect *packs* to be correlated with *u*?
b. Suppose that you have data on average cigarette price in each woman's state of residence. Discuss whether this information is likely to satisfy the properties of a good instrumental variable for packs.
c. Use the data in BWGHT.RAW to estimate equation (5.54). First, use OLS. Then, use 2SLS, where *cigprice* is an instrument for packs. Discuss any important differences in the OLS and 2SLS estimates.
d. Estimate the reduced form for *packs*. What do you conclude about identification of equation (5.54) using *cigprice* as an instrument for *packs*? What bearing does this conclusion have on your answer from part c?

a. Smoking could be related to other "bad" habits that affect the weight of the child (drinking or drugs). These habits are picked up by the error term because they're not in the model so the error term could be correlated to *packs*.
b. Price of cigarettes is not related (strongly) to other factors that influence birth weight, but is probably related to how much someone smokes... of course, if the person doesn't smoke (i.e., *packs* = 0), then the price is uncorrelated.
c. The coefficient of packs changes dramatically from OLS to 2SLS (from -0.084 to 0.797), although the 2SLS coefficient is not statistically significant. It also doesn't make sense since one would expect cigarette smoking to lead to reduced birth weight... problem is investigated in part d.

```
use bwght
regress lbwght male parity lfaminc packs

Source          SS        df        MS         Number of obs  =    1388
                                               F(  4,  1383)  =   12.55
Model      1.76664363    4   .441660908        Prob > F       =  0.0000
Residual    48.65369   1383  .035179819        R-squared      =  0.0350
                                               Adj R-squared  =  0.0322
Total      50.4203336  1387  .036352079        Root MSE       =  .18756
```

```
lbwght       Coef.    Std. Err.     t     P>t     [95% Conf. Interval]

male       .0262407  .0100894    2.60   0.009    .0064486   .0460328
parity     .0147292  .0056646    2.60   0.009    .0036171   .0258414
lfaminc    .0180498  .0055837    3.23   0.001    .0070964   .0290032
packs     -.0837281  .0171209   -4.89   0.000   -.1173139  -.0501423
_cons      4.675618  .0218813  213.68   0.000    4.632694   4.718542
```

```
ivreg lbwght male parity lfaminc (packs = cigprice)
```

```
Instrumental variables (2SLS) regression
Source         SS        df        MS         Number of obs  =    1388
                                              F(  4,  1383)  =    2.39
Model      -91.350027    4   -22.8375067     Prob > F       =  0.0490
Residual   141.770361  1383  .102509299      R-squared      =       .
                                              Adj R-squared  =       .
Total      50.4203336  1387  .036352079      Root MSE       =  .32017

lbwght       Coef.    Std. Err.     t     P>t     [95% Conf. Interval]

packs      .7971063  1.086275    0.73   0.463   -1.333819   2.928031
male       .0298205   .017779    1.68   0.094   -.0050562   .0646972
parity    -.0012391  .0219322   -0.06   0.955    -.044263   .0417848
lfaminc     .063646  .0570128    1.12   0.264   -.0481949   .1754869
_cons      4.467861  .2588289   17.26   0.000    3.960122   4.975601
Instrumented:  packs
Instruments:   male parity lfaminc cigprice
```

d. It seems *cigprice* is not a good IV for *packs* because the coefficient in the reduced form equation is not statistically significant. In class Prof. Ai said if the IV (*cigprice*) and the problem regressor (*packs*) are not highly correlated (or at least $\rho > 0.1$) then $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i'$ would be near singular (same problem as near multicollinearity) so the estimates from 2SLS are bad. In this case, the correlation is only 0.0097. As mentioned in part b, the problem here is that the data contains many nonsmokers (packs = 0 for 1176 out of 1388 data points: 85%!) so the price of cigarettes has nothing to do with the number of packs smoked.

```
regress packs male parity lfaminc cigprice
```

```
Source         SS        df        MS         Number of obs  =    1388
                                              F(  4,  1383)  =   10.86
Model      3.76705108    4    .94176277      Prob > F       =  0.0000
Residual   119.929078  1383  .086716615      R-squared      =  0.0305
                                              Adj R-squared  =  0.0276
Total      123.696129  1387  .089182501      Root MSE       =  .29448
```

```
packs        Coef.    Std. Err.     t     P>t     [95% Conf. Interval]

male      -.0047261  .0158539   -0.30   0.766   -.0358264   .0263742
parity     .0181491  .0088802    2.04   0.041    .0007291   .0355692
lfaminc   -.0526374  .0086991   -6.05   0.000   -.0697023  -.0355724
cigprice    .000777  .0007763    1.00   0.317   -.0007459   .0022999
_cons      .1374075  .1040005    1.32   0.187   -.0666084   .3414234

correlate packs cigprice

   (obs=1388)
                packs    cigprice
   packs        1.0000
   cigprice     0.0097   1.0000

count if packs == 0

   1176
```

Consider again the omitted variable model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \gamma q + v \tag{5.45}$$

where q represents the omitted variable and $E(v \mid \mathbf{x}, q) = 0$.

5.7. Consider model (5.45) where v has zero mean and is uncorrelated with x_1, ..., x_K and q. The unobservable q is thought to be correlated with at least some of the x_j. Assume without loss of generality that $E(q)$ = 0.

You have a single indicator of q, written as $q_1 = \delta_1 q + a_1$, $\delta_1 \neq 0$, where a_1 has zero mean and is uncorrelated with each of x_j, q, and v. In addition, z_1, z_2, …, z_M is a set of variables that are (1) redundant in the structural equation (5.45) and (2) uncorrelated with a_1.
a. Suggest an IV method for consistently estimating the β_j. Be sure to discuss what is needed for identification.
b. If equation (5.45) is a $\log(wage)$ equation, q is ability, q_1 is IQ or some other test score, and z_1, …, z_M are family background variables, such as parents' education and number of siblings, describe the economic assumptions needed for consistency of the IV procedure in part a.
c. Carry out this procedure using the data in NLS80.RAW. Include among the explanatory variables *exper*, *tenure*, *educ*, *married*, *south*, *urban*, and *black*. First use *IQ* as q_1 and then *KWW*. Include in the z_h the variables *meduc*, *feduc*, and *sibs*. Discuss the results.

a. "The solution that would follow from Section 5.1.1 is to put q in the error term, and then to find instruments for any element of $\mathbf{x}$ that is correlated with q. It is useful to think of the instruments satisfying the following requirements: (1) they are redundant in the structural model $E(y \mid \mathbf{x}, q)$; (2) they are uncorrelated with the omitted variable, q; and (3) they are sufficiently correlated with the endogenous elements of $\mathbf{x}$ (that is, those elements that are correlated with q). Then 2SLS applied to equation (5.45) with $u \equiv \gamma q + v$ produces consistent and asymptotically normal estimators." (p.105)

In English, that means we solve $q_1 = \delta_1 q + a_1$ for q and plug it into equation (5.45). We know q_1 will be uncorrelated with the error term (v) and it is correlated to q so it is a good IV. The family background variables (z_1, z_2, …, z_M) could be used at IVs fro the IV (q_1) since they are uncorrelated with a_1.

b. The assumption necessary is that the family background variables (z_1, z_2, …, z_M) are correlated with $\log(wage)$, but only in the way that ability (q) is. That is, family background is redundant once we account for ability. Since we don’t know ability, it's good enough if the family background variables are correlated with IQ (q_1).

c. Three regressions listed before. First is OLS with missing variable (no ability), then using *meduc*, *feduc*, and *sibs* as IVs for *IQ* (as a measure of ability), then the same IVs for *KWW*. The first thing to note is the change in the sample size (from 935 to 722). This is because there is missing data on *meduc* and *feduc*. Although *educ* has a small, but statistically significant affect on *lwage*, in both 2SLS regressions *educ* becomes insignificant (and has an even smaller impact). In fact, many variables that are significant before accounting for ability seem to lose their significance when using the IVs. Even so there are inconsistencies between both 2SLS. For example, *exper* is significant in the *IQ* 2SLS, but not in the *KWW* 2SLS. The opposite occurs with *south* (not significant for *IQ*, but is for *KWW*). This suggests that there's either something wrong with the IVs (*meduc*, *feduc*, and *sibs*) or the proxies for ability (*IQ* or *KWW*). Given the strong correlation between the IVs and *IQ* (and *KWW*), the problem is likely to be the IVs correlated to the error term (which we can't verify).

```
use nls80
regress lwage exper tenure educ married south urban black

Source         SS      df       MS        Number of obs   =      935
                                          F(  7,   927)   =    44.75
Model      41.8377619   7   5.97682312    Prob > F        =   0.0000
Residual   123.818521  927  .133569063    R-squared       =   0.2526
                                          Adj R-squared   =   0.2469
Total      165.656283  934  .177362188    Root MSE        =   .36547

lwage       Coef.    Std. Err.    t     P>t    [95% Conf. Interval]

exper      .014043   .0031852   4.41   0.000    .007792    .020294
tenure    .0117473    .002453   4.79   0.000   .0069333   .0165613
educ      .0654307   .0062504  10.47   0.000   .0531642   .0776973
married   .1994171   .0390502   5.11   0.000   .1227801    .276054
south    -.0909036   .0262485  -3.46   0.001   -.142417  -.0393903
urban     .1839121   .0269583   6.82   0.000   .1310056   .2368185
black    -.1883499   .0376666  -5.00   0.000  -.2622717  -.1144281
_cons     5.395497    .113225  47.65   0.000    5.17329   5.617704

ivreg lwage exper tenure educ married south urban black (iq = meduc
   feduc sibs)
```

```
Instrumental variables (2SLS) regression
Source          SS       df        MS        Number of obs =      722
                                             F(  8,   713) =    25.81
Model      19.6029198    8   2.45036497      Prob > F      =   0.0000
Residual   107.208996   713  .150363248      R-squared     =   0.1546
                                             Adj R-squared =   0.1451
Total      126.811916   721  .175883378      Root MSE      =   .38777

lwage        Coef.    Std. Err.     t      P>t     [95% Conf. Interval]

iq         .0154368   .0077077    2.00   0.046    .0003044   .0305692
exper      .0162185   .0040076    4.05   0.000    .0083503   .0240867
tenure     .0076754   .0030956    2.48   0.013    .0015979   .0137529
educ       .0161809   .0261982    0.62   0.537    -.035254   .0676158
married    .1901012   .0467592    4.07   0.000    .0982991   .2819033
south      -.047992   .0367425   -1.31   0.192   -.1201284   .0241444
urban      .1869376   .0327986    5.70   0.000    .1225442   .2513311
black      .0400269   .1138678    0.35   0.725   -.1835294   .2635832
_cons      4.471616    .468913    9.54   0.000       3.551   5.392231
Instrumented:  iq
Instruments:   exper tenure educ married south urban black meduc
   feduc sibs
```

```
ivreg lwage exper tenure educ married south urban black (kww = meduc
   feduc sibs)
```

```
Instrumental variables (2SLS) regression
Source          SS       df        MS        Number of obs =      722
                                             F(  8,   713) =    25.70
Model       19.820304    8     2.477538      Prob > F      =   0.0000
Residual   106.991612   713  .150058361      R-squared     =   0.1563
                                             Adj R-squared =   0.1468
Total      126.811916   721  .175883378      Root MSE      =   .38737

lwage        Coef.    Std. Err.     t      P>t     [95% Conf. Interval]

kww        .0249441   .0150576    1.66   0.098   -.0046184   .0545067
exper      .0068682   .0067471    1.02   0.309   -.0063783   .0201147
tenure     .0051145   .0037739    1.36   0.176   -.0022947   .0125238
educ       .0260808   .0255051    1.02   0.307   -.0239933   .0761549
married    .1605273   .0529759    3.03   0.003    .0565198   .2645347
south      -.091887   .0322147   -2.85   0.004   -.1551341  -.0286399
urban      .1484003   .0411598    3.61   0.000    .0675914   .2292093
black     -.0424452   .0893695   -0.47   0.635   -.2179041   .1330137
_cons      5.217818   .1627592   32.06   0.000    4.898273   5.537362
Instrumented:  kww
Instruments:   exper tenure educ married south urban black meduc
   feduc sibs
```

```
regress iq meduc feduc sibs
```

```
Source           SS        df        MS          Number of obs   =     722
                                                 F(   3,    718)  =   51.21
Model      27728.3381     3    9242.77937    Prob > F         =  0.0000
Residual   129598.655   718    180.499519    R-squared        =  0.1762
                                             Adj R-squared    =  0.1728
Total      157326.993   721    218.206648    Root MSE         =  13.435

iq             Coef.    Std. Err.     t      P>t      [95% Conf. Interval]

meduc       .8355582   .2210192    3.78    0.000     .401637    1.269479
feduc       .9454379   .1857649    5.09    0.000    .5807306    1.310145
sibs       -1.186177   .2330697   -5.09    0.000   -1.643757   -.7285974
_cons       86.78424   2.423919   35.80    0.000    82.02542    91.54306

regress kww meduc feduc sibs

Source           SS        df        MS          Number of obs   =     722
                                                 F(   3,    718)  =   29.66
Model      4687.71235     3    1562.57078    Prob > F         =  0.0000
Residual   37827.0688   718    52.6839399    R-squared        =  0.1103
                                             Adj R-squared    =  0.1065
Total      42514.7812   721    58.9664094    Root MSE         =  7.2584

kww            Coef.    Std. Err.     t      P>t      [95% Conf. Interval]

meduc       .3562475   .1194074    2.98    0.003    .1218181    .5906768
feduc       .2761705    .100361    2.75    0.006    .0791345    .4732065
sibs       -.6485913   .1259178   -5.15    0.000   -.8958023   -.4013803
_cons       31.04776   1.309541   23.71    0.000    28.47677    33.61875
```

6.3. Consider a model for individual data to test whether nutrition affects productivity (in a developing country):

$$\log(produc) = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 educ + \alpha_1 calories + \alpha_2 protein + u_1 \qquad (6.35)$$

where *produc* is some measure of worker productivity, *calories* is caloric intake per day, and *protein* is a measure of protein intake per day. Assume here that *exper*, *exper*^{2}, and *educ* are all exogenous. The variables *calories* and *protein* are possibly correlated with u_1 (see Strauss and Thomas, 1995, for discussion). Possible instrumental variables for *calories* and *protein* are regional prices of various goods such as grains, meats, breads, dairy products, and so on.
a. Under what circumstances do prices make good IVs for *calories* and *proteins*? What if prices reflect quality of food?
b. How many prices are needed to identify equation (6.35)?
c. Suppose we have M prices, p_1, ..., p_M. Explain how to test the null hypothesis that *calories* and *protein* are exogenous in equation (6.35).

a. A good IV should be correlated to the problem regressor (i.e., the one correlated to the error term) and the IV should be uncorrelated to the error term. In this case, if we assume prices reflect quality of food and quality means more calories and more protein, then we could probably use prices of food.
b. If we're looking to replace both *calories* and *protein*, we need at least two prices.

c. We run a regression on the reduced form for both *calories* and *protein*:

$$calories = \lambda_{\mathrm{a}} + \lambda_{\mathrm{b}}\,exper + \lambda_{\mathrm{c}}\,exper^2 + \lambda_{\mathrm{d}}\,educ + \sum_{i=1}^{M} \lambda_i p_i + v_1$$

$$protein = \pi_{\mathrm{a}} + \pi_{\mathrm{b}}\,exper + \pi_{\mathrm{c}}\,exper^2 + \pi_{\mathrm{d}}\,educ + \sum_{i=1}^{M} \pi_i p_i + v_2$$

Method from book... we didn't cover this in class

i. Get the residuals for the reduced for equations above $\hat{v}_1$ and $\hat{v}_2$

ii. Run $\log(produc) = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 educ + \gamma_1\hat{v}_1 + \gamma_2\hat{v}_2$

iii. Do a joint significance test: $\gamma_1 = \gamma_2 = 0$

Prof Ai said to use the Hausman test:

i. Get $\hat{\boldsymbol{\beta}}$ and $Cov(\hat{\boldsymbol{\beta}})$ from OLS of 6.35

ii. Get $\tilde{\boldsymbol{\beta}}$ and $Cov(\tilde{\boldsymbol{\beta}})$ from 2SLS of 6.35 using $p_1, ..., p_M$ as instruments for *calories* and *protein*

iii. Compute test statistic $(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\left[Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}})\right]^{-1}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \sim \chi_k^2$

6.8. The data in FERTIL1.RAW are a pooled cross section on more than a thousand U.S. women for the even years between 1972 and 1984, inclusive; the data set is similar to the one used by Sander (1992). These data can be used to study the relationship between women's education and fertility.

a. Use OLS to estimate a model relating number of children ever born to a woman (*kids*) to years of education, age, region, race, and type of environment reared in. You should use a quadratic in age and should include year dummies. What is the estimated relationship between fertility and education? Holding other factors fixed, has there been any notable secular change in fertility over the time period?

b. Reestimate the model in part a, but use *meduc* and *feduc* as instruments for *educ*. First check that these instruments are sufficiently partially correlated with *educ*. Test whether *educ* is in fact exogenous in the fertility equation.

c. Now allow the effect of education to change over time by including interaction terms such as y74-educ, y76-educ, and so on in the model. Use interactions of time dummies and parents' education as instruments for the interaction terms. Test that there has been no change in the relationship between fertility and education over time.

a. According to the OLS regression, *educ* is statistically significant with coefficient -0.128. That means for each year of education, a woman is less likely to have children (or will have fewer of them). This doesn't really measure fertility though because we're regressing the number of kids (fertile women may not want kids or may not be trying yet)... this whole model is suspect if we're trying to talk about fertility.

Over time, only the last two time periods are statistically significant; still all but the first have a negative coefficient which implies that (all things equal) women are having less children (than 1972).

```
use fertil1
generate age2 = age^2
regress kids educ age age2 east northcen west black farm othrural
   town smcity y74 y76 y78 y80 y82 y84
```

```
Source          SS        df        MS         Number of obs  =    1129
                                               F( 17,  1111)  =    9.72
Model      399.610888    17   23.5065228      Prob > F       =  0.0000
Residual   2685.89841  1111   2.41755033      R-squared      =  0.1295
                                               Adj R-squared  =  0.1162
Total       3085.5093  1128   2.73538059      Root MSE       =  1.5548

kids           Coef.    Std. Err.     t     P>t     [95% Conf. Interval]

educ       -.1284268   .0183486   -7.00   0.000    -.1644286   -.092425
age         .5321346   .1383863    3.85   0.000     .2606065   .8036626
age2        -.005804   .0015643   -3.71   0.000    -.0088733  -.0027347
east         .217324   .1327878    1.64   0.102    -.0432192   .4778672
northcen     .363114   .1208969    3.00   0.003      .125902   .6003261
west        .1976032   .1669134    1.18   0.237    -.1298978   .5251041
black       1.075658   .1735356    6.20   0.000     .7351631   1.416152
farm       -.0525575     .14719   -0.36   0.721    -.3413592   .2362443
othrural   -.1628537    .175442   -0.93   0.353    -.5070887   .1813814
town        .0843532    .124531    0.68   0.498    -.1599893   .3286957
smcity      .2118791    .160296    1.32   0.187    -.1026379   .5263961
y74         .2681825    .172716    1.55   0.121    -.0707039   .6070689
y76        -.0973795   .1790456   -0.54   0.587     -.448685   .2539261
y78        -.0686665   .1816837   -0.38   0.706    -.4251483   .2878154
y80        -.0713053   .1827707   -0.39   0.697      -.42992   .2873093
y82        -.5224842   .1724361   -3.03   0.003    -.8608214   -.184147
y84        -.5451661   .1745162   -3.12   0.002    -.8875846  -.2027477
_cons      -7.742457   3.051767   -2.54   0.011    -13.73033  -1.754579
```

b. Simple check of the correlation shows that *meduc* and *feduc* each is correlated with *educ* (about 0.46). Looking at the regression for the reduced form of *educ*, both *meduc* and *feduc* are statistically significant and have the largest coefficients. So *meduc* and *feduc* are at least correlated to *educ*, we'll assume they're good instruments and then check for endogeneity. There are a couple of ways to do it. The book's method is to save the residuals for the reduced from OLS and then plug them in place of *educ* in the original model and check the significance of their coefficients. The coefficient is in fact significant which suggests *educ* is endogenous.

Moving on to the Hausman test we did in class, if we use H_0: there is no endogeneity problem, the OLS estimates are efficient, but they're biased if H_0 doesn't hold. The 2SLS estimates are consistent in both cases, but inefficient under H_0. The result of the test statistic shows that we cannot reject ∴ *educ* is exogenous (or at least, we couldn't prove that it wasn't).

Note: this test doesn’t really add that much since the coefficient on *educ* is pretty close and very significant in both OLS and 2SLS (-0.13 vs. -0.15).

```
estimates store ols
```
(This is Stata 8 version for Hausman test)

```
correlate educ meduc feduc

   (obs=1129)
             educ      meduc    feduc
   educ      1.0000
   meduc     0.4671    1.0000
   feduc     0.4714    0.6380   1.0000
```

```
regress educ age age2 east northcen west black farm othrural town
   smcity y74 y76 y78 y80 y82 y84 meduc feduc
```

```
Source          SS         df        MS        Number of obs =    1129
                                               F( 18,  1110) =   24.82
Model       2256.26171    18   125.347873      Prob > F      =  0.0000
Residual    5606.85432  1110   5.05122011      R-squared     =  0.2869
                                               Adj R-squared =  0.2754
Total       7863.11603  1128   6.97084755      Root MSE      =  2.2475

educ           Coef.    Std. Err.     t     P>t     [95% Conf. Interval]

age        -.2243687   .2000013   -1.12   0.262    -.616792   .1680546
age2        .0025664   .0022605    1.14   0.256    -.001869   .0070018
east        .2488042   .1920135    1.30   0.195   -.1279462   .6255546
northcen    .0913945   .1757744    0.52   0.603   -.2534931   .4362821
west        .1010676   .2422408    0.42   0.677   -.3742339   .5763691
black       .3667819   .2522869    1.45   0.146   -.1282311    .861795
farm       -.3792615   .2143864   -1.77   0.077   -.7999099   .0413869
othrural    -.560814   .2551196   -2.20   0.028   -1.061385   -.060243
town        .0616337   .1807832    0.34   0.733   -.2930816    .416349
smcity      .0806634   .2317387    0.35   0.728   -.3740319   .5353587
y74         .0060993    .249827    0.02   0.981   -.4840872   .4962858
y76         .1239104   .2587922    0.48   0.632   -.3838667   .6316874
y78         .2077861   .2627738    0.79   0.429   -.3078033   .7233755
y80         .3828911   .2642433    1.45   0.148   -.1355816   .9013638
y82         .5820401   .2492372    2.34   0.020    .0930108   1.071069
y84         .4250429   .2529006    1.68   0.093   -.0711741     .92126
meduc       .1723015   .0221964    7.76   0.000    .1287499   .2158531
feduc       .2074188   .0254604    8.15   0.000    .1574629   .2573747
_cons       13.63334   4.396773    3.10   0.002    5.006421   22.26027
```

Book's Method

```
predict u_ed,  residuals
regress kids age age2 east northcen west black farm othrural town
   smcity y74 y76 y78 y80 y82 y84 u_ed
```

```
Source          SS         df        MS        Number of obs  =     1129
                                               F( 17,  1111)  =     8.74
Model       364.084469    17   21.4167335      Prob > F       =   0.0000
Residual    2721.42483  1111    2.4495273      R-squared      =   0.1180
                                               Adj R-squared  =   0.1045
Total        3085.5093  1128   2.73538059      Root MSE       =   1.5651
```

```
kids          Coef.     Std. Err     t      P>t     [95% Conf. Interval]

age         .5774619   .1391459    4.15   0.000    .3044435   .8504803
age2       -.0062687   .0015732   -3.98   0.000   -.0093554  -.0031819
east        .1579961   .1333905    1.18   0.236   -.1037296   .4197219
northcen    .3033986   .1213904    2.50   0.013    .0652182    .541579
west        .1445866   .1678405    0.86   0.389   -.1847335   .4739068
black       1.089947   .1746674    6.24   0.000    .7472321   1.432662
farm        .0765636    .146992    0.52   0.603   -.2118496   .3649767
othrural    .0082479   .1748758    0.05   0.962   -.3348761   .3513719
town        .0977874    .125337    0.78   0.435   -.1481365   .3437114
smcity      .2086013   .1613519    1.29   0.196   -.1079875   .5251902
y74         .2473348   .1738287    1.42   0.155   -.0937348   .5884043
y76        -.1123343    .180213   -0.62   0.533   -.4659304   .2412618
y78        -.1289488   .1826757   -0.71   0.480   -.4873771   .2294795
y80        -.1666899   .1834634   -0.91   0.364   -.5266636   .1932838
y82        -.6612469   .1724218   -3.84   0.000   -.9995559  -.3229378
y84        -.6709152   .1747332   -3.84   0.000    -1.01376  -.3280709
u_ed       -.1216021   .0209017   -5.82   0.000   -.1626133  -.0805908
_cons      -10.39001   3.048196   -3.41   0.001   -16.37088  -4.409146
```

Hausman Test in Stata 8

```
regress kids educ age age2 east northcen west black farm othrural
   town smcity y74 y76 y78 y80 y82 y84
estimates store ols
ivreg kids age age2 east northcen west black farm othrural town
   smcity y74 y76 y78 y80 y82 y84 (educ = meduc feduc)
estimates store two_sls
hausman two_sls ols   (supposed to list inefficient first)
```

```
             ---- Coefficients ----
               (b)          (B)          (b-B)    sqrt(diag(V_b-V_B))
             two_sls        ols        Difference        S.E.

educ        -.1527395    -.1284268    -.0243126        .0346668
age          .5235536     .5321346     -.008581        .0134128
age2         -.005716     -.005804      .000088          .00014
east         .2285554      .217324     .0112314        .0168588
northcen     .3744188      .363114     .0113048        .0168171
west         .2076398     .1976032     .0100366        .0157715
black        1.072952     1.075658    -.0027052        .0079033
farm        -.0770015    -.0525575    -.0244441        .0353343
othrural    -.1952451    -.1628537    -.0323914        .0466997
town           .08181     .0843532    -.0025432        .0061361
smcity       .2124996     .2118791     .0006205        .0064334
y74          .2721292     .2681825     .0039467        .0088768
y76         -.0945483    -.0973795     .0028311        .0081823
y78         -.0572543    -.0686665     .0114121        .0177998
y80          -.053248    -.0713053     .0180574        .0267476
y82         -.4962149    -.5224842     .0262693        .0380707
y84         -.5213604    -.5451661     .0238057        .0346383
```

```
            b = consistent under Ho and Ha; obtained from regress
B =inconsistent under Ha, efficient under Ho; obtained from ivreg

Test:  Ho:    difference in coefficients not systematic

   chi2(16)    = (b-B)'[(V_b-V_B)^(-1)](b-B)
               =        0.49
   Prob>chi2   =      1.0000
```

Hausman Test... pre Stata 8

```
ivreg kids age age2 east northcen west black farm othrural town
   smcity y74 y76 y78 y80 y82 y84 (educ = meduc feduc)
hausman, save
regress kids educ age age2 east northcen west black farm othrural
   town smcity y74 y76 y78 y80 y82 y84
hausman
```

Gives same result as in Stata 8

c. Now nothing in the regression is significant (except the regression itself). A joint test of the education-time interactions shows that they are not statistically different than zero. So there has been no change in the relationship between fertility and education over time.

```
generate y74educ = y74*educ
generate y76educ = y76*educ
generate y78educ = y78*educ
generate y80educ = y80*educ
generate y82educ = y82*educ
generate y84educ = y84*educ
generate y74meduc = y74*meduc
generate y76meduc = y76*meduc
generate y78meduc = y78*meduc
generate y80meduc = y80*meduc
generate y82meduc = y82*meduc
generate y84meduc = y84*meduc
generate y74feduc = y74*feduc
generate y76feduc = y76*feduc
generate y78feduc = y78*feduc
generate y80feduc = y80*feduc
generate y82feduc = y82*feduc
generate y84feduc = y84*feduc
ivreg kids age age2 east northcen west black farm othrural town
   smcity y74 y76 y78 y80 y82 y84 (educ y74educ-y84educ = y74meduc-
   y84meduc y74feduc-y84feduc)
```

```
Instrumental variables (2SLS) regression
Source          SS        df       MS         Number of obs =    1129
                                              F( 23,  1105) =    5.83
Model      321.975637    23   13.9989407      Prob > F      =  0.0000
Residual   2763.53366  1105   2.50093544      R-squared     =  0.1044
                                              Adj R-squared =  0.0857
Total       3085.5093  1128   2.73538059      Root MSE      =  1.5814

kids          Coef.    Std. Err.    t     P>t     [95% Conf. Interval]

educ      -.2665368  5.709934   -0.05   0.963   -11.47007     10.937
y74educ      .16957  5.677001    0.03   0.976   -10.96935   11.30849
y76educ    .0229305  5.647092    0.00   0.997    -11.0573   11.10316
y78educ   -.0611455  5.717204   -0.01   0.991   -11.27895   11.15665
y80educ     .187394   5.66007    0.03   0.974    -10.9183   11.29309
y82educ    .0313912  5.620736    0.01   0.996   -10.99713   11.05991
y84educ    .0997358  5.613407    0.02   0.986    -10.9144   11.11388
age        .4988042  .3400548    1.47   0.143   -.1684219    1.16603
age2      -.0054645  .0034644   -1.58   0.115    -.012262   .0013329
east       .2643142  .8105764    0.33   0.744   -1.326128   1.854757
northcen   .3975034  .5414025    0.73   0.463   -.6647896   1.459796
west        .240049  1.261689    0.19   0.849   -2.235527   2.715625
black      1.057324  .6727885    1.57   0.116   -.2627628   2.377411
farm      -.1007702  .2580721   -0.39   0.696   -.6071368   .4055964
othrural  -.2229637  .4759143   -0.47   0.640   -1.156761    .710834
town       .0832029  .2492774    0.33   0.739   -.4059076   .5723134
smcity     .2198452  .4889677    0.45   0.653   -.7395646   1.179255
y74       -1.797701  69.03333   -0.03   0.979   -137.2489   133.6535
y76       -.3679201  68.66399   -0.01   0.996   -135.0944   134.3586
y78        .7673704  69.59531    0.01   0.991   -135.7865   137.3212
y80       -2.384622  68.82471   -0.03   0.972   -137.4265   132.6572
y82       -.7889115  68.14246   -0.01   0.991   -134.4921   132.9143
y84       -1.726933  68.25031   -0.03   0.980   -135.6418   132.1879
_cons     -5.276107  76.48671   -0.07   0.945   -155.3517   144.7995
Instrumented: educ y74educ y76educ y78educ y80educ y82educ y84educ
Instruments:  age age2 east northcen west black farm othrural town
              smcity y74 y76 y78 y80 y82 y84 y74meduc y76meduc
              y78meduc y80meduc y82meduc y84meduc y74feduc y76feduc
              y78feduc y80feduc y82feduc y84feduc

test y74educ y76educ y78educ y80educ y82educ y84educ

( 1)    y74educ = 0
( 2)    y76educ = 0
( 3)    y78educ = 0
( 4)    y80educ = 0
( 5)    y82educ = 0
( 6)    y84educ = 0

   F(  6,  1105) =    0.78
        Prob > F =    0.5829
```

$$\begin{array}{lllll} \log(durant) = 1.126 & +\ 0.0077\ afchnge & +\ 0.256\ highearn & +\ 0.191\ afchnge \cdot highearn & (6.33) \\ \qquad\qquad\quad (0.031) & \quad (0.0447) & \quad (0.047) & \quad (0.069) & \end{array}$$

$N = 5626, \quad R^2 = 0.021$

6.9. Use the data in INJURY.RAW for this question.
a. Using the data for Kentucky, reestimate equation (6.33) adding as explanatory variables male, married, and a full set of industry- and injury-type dummy variables. How does the estimate on afchnge-highearn change when these other factors are controlled for? Is the estimate still statistically significant?
b. What do you make of the small R-squared from part a? Does this mean the equation is useless?
c. Estimate equation (6.33) using the data for Michigan. Compare the estimate on the interaction term for Michigan and Kentucky, as well as their statistical significance.

a. The coefficient for *afchnge·highearn* actually increases and become more statistically significant. The standard error didn't see to change (still 0.069).

```
use injury
regress ldurat afchnge highearn afhigh male married head neck upextr
   trunk lowback lowextr occdis manuf construc if ky
```

```
Source          SS         df        MS        Number of obs  =    5349
                                               F( 14,  5334)  =   16.37
Model      358.441793    14   25.6029852       Prob > F       =  0.0000
Residual   8341.41206  5334   1.56381928       R-squared      =  0.0412
                                               Adj R-squared  =  0.0387
Total      8699.85385  5348   1.62674904       Root MSE       =  1.2505

ldurat        Coef.     Std. Err.     t      P>t     [95% Conf. Interval]

afchnge     .0106274   .0449167    0.24   0.813   -.0774276   .0986824
highearn    .1757598   .0517462    3.40   0.001    .0743161   .2772035
afhigh      .2308768   .0695248    3.32   0.001    .0945798   .3671738
male       -.0979407   .0445498   -2.20   0.028   -.1852766  -.0106049
married     .1220995   .0391228    3.12   0.002    .0454027   .1987962
head       -.5139003   .1292776   -3.98   0.000   -.7673372  -.2604634
neck        .2699126   .1614899    1.67   0.095   -.0466737   .5864988
upextr      -.178539   .1011794   -1.76   0.078    -.376892   .0198141
trunk       .1264514   .1090163    1.16   0.246   -.0872651    .340168
lowback    -.0085967   .1015267   -0.08   0.933   -.2076305   .1904371
lowextr    -.1202911   .1023262   -1.18   0.240   -.3208922   .0803101
occdis      .2727118    .210769    1.29   0.196   -.1404816   .6859052
manuf      -.1606709   .0409038   -3.93   0.000   -.2408591  -.0804827
construc    .1101967   .0518063    2.13   0.033    .0086352   .2117581
_cons       1.245922   .1061677   11.74   0.000     1.03779   1.454054
```

b. The R^2 value is pretty small (0.04), but it's still twice as large as (6.33) which only had 0.021. This means that most of the variation in *ldurat* is explained by unobserved (or not included) variables. Although the coefficients are statistically significant, the confidence intervals of the predictions for *ldurat* will be very wide. If the model's purpose is to predict *ldurat*, it's not very good.

c. The value of the coefficient on $afchnge{\cdot}highearn$ is practically the same between KY and MI, but it's not statistically significant in the MI regression... possibly because of the difference in sample size (5349 vs. 1524)

```
regress ldurat afchnge highearn afhigh if mi

Source          SS        df        MS         Number of obs   =    1524
                                               F(   3,  1520)  =    6.05
Model       34.3850177     3    11.4616726     Prob > F        =  0.0004
Residual    2879.96981  1520    1.89471698     R-squared       =  0.0118
                                               Adj R-squared   =  0.0098
Total       2914.35483  1523    1.91356194     Root MSE        =  1.3765

ldurat         Coef.    Std. Err.     t      P>t     [95% Conf.Interval]

afchnge     .0973808   .0847879    1.15   0.251   -.0689329   .2636945
highearn    .1691388   .1055676    1.60   0.109   -.0379348   .3762124
afhigh      .1919906   .1541699    1.25   0.213   -.1104176   .4943988
_cons       1.412737   .0567172   24.91   0.000    1.301485   1.523989
```

Example 7.7 (Effects of Job Training Grants on Firm Scrap Rates): Using the data from JTRAIN1.RAW (Holzer, Block, Cheatham, and Knott, 1993), we estimate a model explaining the firm scrap rate in terms of grant receipt. We can estimate the equation for 54 firms and three years of data (1987, 1988, and 1989). The first grants were given in 1988. Some firms in the sample in 1989 received a grant only in 1988, so we allow for a one-year-lagged effect:

$$\begin{array}{lllll} \log(scrap_{it}) = 0.597 & - \ 0.239\, d88_t & - \ 0.497\, d89_t & + \ 0.200\, grant_{it} & + \ 0.049\, grant_{i,t\text{-}1} \\ \quad\quad\quad\quad (.203) & \quad (.311) & \quad (.388) & \quad (.338) & \quad (.436) \end{array}$$

$N = 54, \quad T = 3, \quad R^2 = .0173$

where we have put i and t subscripts on the variables to emphasize which ones change across firm or time. The R-squared is just the usual one computed form the pooled OLS regression.

In this equation, the estimated grant effect has the wrong sign, and neither the current nor lagged grant variable is statistically significant. When a lag of $\log(scrap_{it})$ is added to the equation, the estimates are notably different. See Problem 7.9

7.9. Redo Example 7.7 but include a single lag of $\log(scrap_{it})$ in the equation to proxy for omitted variables that may determine grant receipt. Test for AR(1) serial correlation. If you find it, you should also compute the fully robust standard errors that allow for arbitrary serial correlation across time and heteroskedasticity.

```
use jtrain1
regress lscrap d89 grant grant_1 lscrap_1 if year != 1987
```

```
Source          SS        df        MS         Number of obs  =     108
                                               F(  4,   103)  =  153.67
Model      186.376973    4    46.5942432       Prob > F       =  0.0000
Residual   31.2296502   103   .303200488       R-squared      =  0.8565
                                               Adj R-squared  =  0.8509
Total      217.606623   107   2.03370676       Root MSE       =  .55064

lscrap        Coef.    Std. Err.     t      P>t     [95% Conf. Interval]

d89       -.1153893   .1199127   -0.96   0.338    -.3532078    .1224292
grant     -.1723924   .1257443   -1.37   0.173    -.4217765    .0769918
grant_1   -.1073226   .1610378   -0.67   0.507     -.426703    .2120579
lscrap_1   .8808216   .0357963   24.61   0.000      .809828    .9518152
_cons     -.0371354   .0883283   -0.42   0.675    -.2123137     .138043
```

Because we're using a lagged variable now, we have to drop the data from 1987 (there's no lagged data for that year... if you leave out the `if year := 1987` the results would be the same). Neither *grant* nor its lag are statistically significant

```
predict uhat, residuals
generate uhat_1 = uhat[_n-1] if d89
regress uhat uhat_1 if d89
```

```
Source          SS        df        MS         Number of obs =      54
                                               F(  1,    52) =    3.64
Model      1.13722287    1    1.13722287       Prob > F      =  0.0619
Residual   16.2366222   52    .312242735       R-squared     =  0.0655
                                               Adj R-squared =  0.0475
Total      17.3738451   53    .327808398       Root MSE      =  .55879

uhat          Coef.    Std. Err.     t      P>t     [95% Conf. Interval]

uhat_1     .2864883   .1501171    1.91   0.062-.0147438    .5877204
_cons      2.41e-09   .0760413    0.00   1.000-.1525879    .1525879
```

Looks like there may be AR(1). Note, solution checks a different way:

```
   regress lscrap grant grant_1 lscrap_1 uhat_1 if d89
```

```
regress lscrap d89 grant grant_1 lscrap_1 if year != 1987, robust
   cluster(fcode)
```

```
Regression with robust standard errors         Number of obs  =     108
                                               F(  4,    53)  =   77.24
                                               Prob > F       =  0.0000
                                               R-squared      =  0.8565
Number of clusters (fcode) = 54                Root MSE       =  .55064
```

```
                         Robust
lscrap        Coef.     Std. Err.     t      P>t     [95% Conf. Interval]

d89        -.1153893   .1145118   -1.01   0.318   -.3450708   .1142922
grant      -.1723924   .1188807   -1.45   0.153   -.4108369   .0660522
grant_1    -.1073226   .1790052   -0.60   0.551   -.4663616   .2517165
lscrap_1    .8808216   .0645344   13.65   0.000    .7513821   1.010261
_cons      -.0371354   .0893147   -0.42   0.679    -.216278   .1420073
```

The coefficients don't change much and *grant* and its lag are still not statistically significant.

7.11. Use the data in CORNWELL.RAW for this question; see Problem 4.13.
a. Using the data for all seven years, and using the logarithms of all variables, estimate a model relating the crime rate to *prbarr*, *prbconv*, *prbpris*, *avgsen*, and *polpc*. Use pooled OLS and include a full set of year dummies. Test for serial correlation assuming that the explanatory variables are strictly exogenous. If there is serial correlation, obtain the fully robust standard errors.
b. Add a one-year lag of log(*crmrte*) to the equation from part a, and compare with the estimates from part a.
c. Test for first-order serial correlation in the errors in the model from part b. If serial correlation is present, compute the fully robust standard errors.
d. Add all of the wage variables (in logarithmic form) to the equation from part c. Which ones are statistically and economically significant? Are they jointly significant? Test for joint significance of the wage variables allowing arbitrary serial correlation and heteroskedasticity.

a. There is very strong evidence for AR(1) serial correlation in the error terms (coefficient of 0.79 and *t*-ratio of 28!)

```
regress lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87

Source          SS        df        MS         Number of obs  =     630
                                               F( 11,   618)  =   74.49
Model       117.644669    11   10.6949699      Prob > F       =  0.0000
Residual     88.735673   618   .143585231      R-squared      =  0.5700
                                               Adj R-squared  =  0.5624
Total       206.380342   629   .328108652      Root MSE       =  .37893

lcrmrte       Coef.     Std. Err.     t      P>t     [95% Conf. Interval]

lprbarr    -.7195033   .0367657  -19.57   0.000   -.7917042  -.6473024
lprbconv   -.5456589   .0263683  -20.69   0.000   -.5974413  -.4938765
lprbpris    .2475521   .0672268    3.68   0.000    .1155314   .3795728
lavgsen    -.0867575   .0579205   -1.50   0.135   -.2005023   .0269872
lpolpc      .3659886   .0300252   12.19   0.000    .3070248   .4249525
d82         .0051371    .057931    0.09   0.929   -.1086284   .1189026
d83         -.043503   .0576243   -0.75   0.451   -.1566662   .0696601
d84        -.1087542    .057923   -1.88   0.061    -.222504   .0049957
d85        -.0780454   .0583244   -1.34   0.181   -.1925835   .0364927
d86        -.0420791   .0578218   -0.73   0.467     -.15563   .0714718
d87        -.0270426    .056899   -0.48   0.635   -.1387815   .0846963
_cons      -2.082293   .2516253   -8.28   0.000   -2.576438  -1.588149
```

```
predict uhat, residuals
generate uhat_1 = uhat[_n-1] if year > 81
regress uhat uhat_1

Source         SS        df       MS         Number of obs =     540
                                             F(  1,    538) =  831.46
Model       46.6680407    1    46.6680407    Prob > F      =  0.0000
Residual    30.1968286   538   .056127934    R-squared     =  0.6071
                                             Adj R-squared =  0.6064
Total       76.8648693   539   .142606437    Root MSE      =  .23691

uhat           Coef.    Std. Err.     t      P>t      [95% Conf. Interval]

uhat_1      .7918085     .02746    28.84   0.000     .7378666    .8457504
_cons       1.74e-10   .0101951     0.00   1.000    -.0200271    .0200271

regress lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc d82-d87,
   robust cluster(county)

Regression with robust standard errors       Number of obs  =     630
                                             F( 11,     89)  =   37.19
                                             Prob > F        =  0.0000
                                             R-squared       =  0.5700
Number of clusters (county) = 90             Root MSE        =  .37893

                           Robust
lcrmrte        Coef.    Std. Err.     t      P>t      [95% Conf.Interval]

lprbarr    -.7195033   .1095979   -6.56   0.000    -.9372719   -.5017347
lprbconv   -.5456589   .0704368   -7.75   0.000    -.6856152   -.4057025
lprbpris    .2475521   .1088453    2.27   0.025     .0312787    .4638255
lavgsen    -.0867575   .1130321   -0.77   0.445    -.3113499    .1378348
lpolpc      .3659886    .121078    3.02   0.003     .1254092    .6065681
d82         .0051371   .0367296    0.14   0.889    -.0678438    .0781181
d83         -.043503    .033643   -1.29   0.199    -.1103509    .0233448
d84        -.1087542   .0391758   -2.78   0.007    -.1865956   -.0309127
d85        -.0780454   .0385625   -2.02   0.046    -.1546683   -.0014224
d86        -.0420791   .0428788   -0.98   0.329    -.1272783    .0431201
d87        -.0270426   .0381447   -0.71   0.480    -.1028353    .0487502
_cons      -2.082293   .8647054   -2.41   0.018    -3.800445   -.3641423
```

b. The lagged crime rate is very significant (*t*-ratio of 43!). The coefficients of the other variables are all smaller now except for *lavgsen* which is up slightly (and is now significant).

```
generate lcrmrt_1 = lcrmrte[_n-1] if year > 81
regress lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc d83-d87
   lcrmrt_1
```

```
Source         SS        df       MS         Number of obs =      540
                                             F( 11,    528)  =  464.68
Model      163.287174    11   14.8442885     Prob > F        =  0.0000
Residual   16.8670945   528   .031945255     R-squared       =  0.9064
                                             Adj R-squared   =  0.9044
Total      180.154268   539   .334237975     Root MSE        =  .17873

lcrmrte      Coef.     Std. Err.     t      P>t      [95% Conf. Interval]

lprbarr   -.1668349   .0229405    -7.27   0.000    -.2119007   -.1217691
lprbconv  -.1285118   .0165096    -7.78   0.000    -.1609444   -.0960793
lprbpris  -.0107492   .0345003    -0.31   0.755     -.078524    .0570255
lavgsen   -.1152298    .030387    -3.79   0.000     -.174924   -.0555355
lpolpc      .101492   .0164261     6.18   0.000     .0692234    .1337606
d83       -.0649438   .0267299    -2.43   0.015    -.1174537   -.0124338
d84       -.0536882   .0267623    -2.01   0.045    -.1062619   -.0011145
d85       -.0085982   .0268172    -0.32   0.749    -.0612797    .0440833
d86        .0420159    .026896     1.56   0.119    -.0108203    .0948522
d87        .0671272   .0271816     2.47   0.014     .0137298    .1205245
lcrmrt_1   .8263047   .0190806    43.31   0.000     .7888214    .8637879
_cons     -.0304828   .1324195    -0.23   0.818    -.2906166     .229651
```

c. It is no significant evidence for AR(1) serial correlation.

```
drop uhat uhat_1
predict uhat, residuals
generate uhat_1 = uhat[_n-1] if year > 82
regress uhat uhat_1
```

```
Source         SS        df       MS         Number of obs   =      450
                                             F(  1,    448)  =     1.11
Model       .037059214    1   .037059214     Prob > F        =   0.2916
Residual    14.8943441  448   .033246304     R-squared       =   0.0025
                                             Adj R-squared   =   0.0003
Total       14.9314033  449   .033254796     Root MSE        =   .18234

uhat         Coef.     Std. Err.     t      P>t      [95% Conf. Interval]

uhat_1    -.0533265   .0505088    -1.06   0.292      -.15259     .045937
_cons      2.95e-11   .0085954     0.00   1.000    -.0168923    .0168923
```

Don't know why book does it this way:

```
   regress lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc d84-d87
      lcrmrt_1 uhat_1
```

d. None of the wage variables is statistically significant. The are not jointly significant either.

```
regress lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc d83-d87
   lcrmrt_1 lwcon-lwloc

Source        SS       df       MS        Number of obs  =     540
                                          F( 20,   519)  =  255.32
Model     163.533423   20   8.17667116    Prob > F       =  0.0000
Residual  16.6208452  519    .03202475    R-squared      =  0.9077
                                          Adj R-squared  =  0.9042
Total     180.154268  539   .334237975    Root MSE       =  .17895

lcrmrte     Coef.   Std. Err.     t    P>t     [95% Conf. Interval]

lprbarr  -.1746053  .0238458   -7.32   0.000  -.2214516  -.1277591
lprbconv -.1337714  .0169096   -7.91   0.000   -.166991  -.1005518
lprbpris -.0195318  .0352873   -0.55   0.580  -.0888553   .0497918
lavgsen  -.1108926  .0311719   -3.56   0.000  -.1721313   -.049654
lpolpc    .1050704  .0172627    6.09   0.000    .071157   .1389838
d83      -.0729231  .0286922   -2.54   0.011  -.1292903  -.0165559
d84      -.0652494  .0287165   -2.27   0.023  -.1216644  -.0088345
d85      -.0258059  .0326156   -0.79   0.429  -.0898807    .038269
d86       .0263763  .0371746    0.71   0.478  -.0466549   .0994076
d87       .0465632  .0418004    1.11   0.266  -.0355555   .1286819
lcrmrt_1  .8087768  .0208067   38.87   0.000    .767901   .8496525
lwcon    -.0283133  .0392516   -0.72   0.471  -.1054249   .0487983
lwtuc    -.0034567  .0223995   -0.15   0.877  -.0474615   .0405482
lwtrd     .0121236  .0439875    0.28   0.783  -.0742918    .098539
lwfir     .0296003  .0318995    0.93   0.354  -.0330676   .0922683
lwser      .012903  .0221872    0.58   0.561  -.0306847   .0564908
lwmfg    -.0409046  .0389325   -1.05   0.294  -.1173893   .0355801
lwfed     .1070534  .0798526    1.34   0.181  -.0498207   .2639275
lwsta    -.0903894  .0660699   -1.37   0.172  -.2201867    .039408
lwloc     .0961124  .1003172    0.96   0.338  -.1009652     .29319
_cons    -.6438061  .6335887   -1.02   0.310   -1.88852   .6009076

test lwcon lwtuc lwtrd lwfir lwser lwmfg lwfed lwsta lwloc

 ( 1)  lwcon = 0
 ( 2)  lwtuc = 0
 ( 3)  lwtrd = 0
 ( 4)  lwfir = 0
 ( 5)  lwser = 0
 ( 6)  lwmfg = 0
 ( 7)  lwfed = 0
 ( 8)  lwsta = 0
 ( 9)  lwloc = 0

       F(  9,   519) =    0.85
            Prob > F =    0.5663
```

Documentation

5.2.iii - Prof Ai covered reduced form during our review session... it didn't help
5.3.i - Katie said this during our review session & Prof Ai said it was right
5.3.ii - Prof Ai gave this answer during our review session
6.3.c - Prof Ai said to run the Hausman test to test if calories and protein are exogenous (OLS is best under H_0)
6.8.a-c - Prof Ai walked me through how to do this problem
7.9 - I pulled this out of the solution manual; have no idea what it is because we didn’t cover this

ECO 7424 Practice Questions 1 Len Cabrera

1. You are estimating a macroeconomic equation of the from $y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t$ under the usual assumptions. Originally, you run the regression on quarterly data 1965-1974 where due to recent economic troubles the $\mathbf{X}$ factors for t =1965:1 to t = 1974:4 are not at all collinear (i.e., multicollinearity is not a problem). You then consider extending your estimation to the years t = 1961:1 to t = 1964:4. While you are confident that the specification of the economic relationship is the same over these earlier years, you notice that the $\mathbf{X}$'s have a high degree of multicollinearity due to the smooth running of the economy in these years. Is it a good idea to add these additional data since it is often noted that multicollinearity of the right hand side variables leads to large standard errors?

Near multicollinearity doesn't matter for forecasting, but we (economists) are usually more interested in parameter estimates so we should worry about it

We always want more data because it reduces sample error, but this can be offset by near multicollinearity. In this case, the multicollinearity introduced in the 61-64 data will probably be offset by the 65-74 data because there's much more data that is not correlated. If we were only running the 61-64 data by itself, there may be a problem.

Note: going the other way would also be better to add the data (i.e., if 65-74 had the near multicollinearity problem, adding 61-64 data that was not correlated would help fix the multicollinearity problem).

2. In a two-factor model, the elasticity of substitution between capital and labor is defined as the elasticity of the capital-labor ratio, K/L, with respect to the factor price ratio, r/w. Set up a regression model using K/L to form the left hand side variable and r/w to form the right hand side variable. Specify the model so that the elasticity of substitution β is a constant. Show how to test the hypothesis $\beta = 1$. What conclusion would you draw if $\hat{\beta} = 0.90$ with standard error 0.40?

$\ln(K/L) = \alpha + \beta \ln(r/w) + u$

H$_0$: $\beta = 1$; H$_1$: $\beta \neq 1$

Use t-test: $\dfrac{\hat{\beta} - 1}{\sqrt{Var(\hat{\beta})}} = \dfrac{0.90 - 1}{0.40} = -0.25$

p-value depends on N, but this is very small t-ratio; fail to reject (not enough evidence to contradict $\beta = 1$)

3. Suppose that an econometric model is given by

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + U_i,\ i = 1,2,\ldots,100$$
$$Y_i = X_{1i}\beta_3 + X_{2i}\beta_4 + U_i,\ i = 101,\ldots,200$$

Construct test statistics for each of the following sets of hypotheses under various conditions on the error term. Simplify your results as much as possible.

(a) H$_0$: $\beta_1 = \beta_3$; H$_1$: $\beta_1 \neq \beta_3$

(b) H$_0$: $\beta_1 = \beta_3$ and $\beta_2 = \beta_4$; H$_1$: $\beta_1 \neq \beta_3$ of $\beta_2 \neq \beta_4$ (test for structural change)

(a) Option 1: Wald Test... do it in Stata

Option 2: F-test... need u_i to satisfy usual conditions

Run unrestricted regression:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{100} \\ y_{101} \\ \vdots \\ y_{200} \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,100} & x_{2,100} & 0 & 0 \\ 0 & 0 & x_{1,101} & x_{2,101} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & x_{1,200} & x_{2,200} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_{100} \\ u_{101} \\ \vdots \\ u_{200} \end{pmatrix}$$

Run restricted regression:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{100} \\ y_{101} \\ \vdots \\ y_{200} \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} & 0 \\ \vdots & \vdots & \vdots \\ x_{1,100} & x_{2,100} & 0 \\ x_{1,101} & 0 & x_{2,101} \\ \vdots & \vdots & \vdots \\ x_{1,200} & 0 & x_{2,200} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_{100} \\ u_{101} \\ \vdots \\ u_{200} \end{pmatrix}$$

$$F = \frac{(SSR_{res} - SSR_{unres})/m}{SSR_{unres}/(N-k)} = \frac{\left(\sum \hat{u}_i^2 - \sum \tilde{u}_i^2\right)/m}{\sum \hat{u}_i^2/(N-k)}$$

m = # restrictions (1); N = # observations (200); k = # parameters in unrestricted (4)

(b) Can't use Wald Test in this case; have to use F-test

Run same unrestricted regression as above

Run restricted regression:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{100} \\ y_{101} \\ \vdots \\ y_{200} \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{2,1} \\ \vdots & \vdots \\ x_{1,100} & x_{2,100} \\ x_{1,101} & x_{2,101} \\ \vdots & \vdots \\ x_{1,200} & x_{2,200} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_{100} \\ u_{101} \\ \vdots \\ u_{200} \end{pmatrix}$$

m = 2; N = 200; k = 4

Note1: for unrestricted regression, we could run two separate regressions and combine the SSR for each regression

Note2: this method is better than using a dummy variable to indicate the model because that assumes same slopes with different intercepts.

Note3: General case $N = N_1 + N_2$ and $k = k_1 + k_2$. Will have $k_1 = k_2$, but can have $N_1 \neq N_2$. Must have enough observations in each sub-sample to run unrestricted regression. If not (e.g., $N_2 < k_2$), then don't include that sub-sample in the unrestricted regression and modify the F-test:

$$F=\frac{\left(\sum \hat{u}_i^2-\sum \tilde{u}_i^2\right)/m}{\sum \hat{u}_i^2/(N_1-k_1)}$$

4. "Collinearity is always something to be avoided in data; the best right-hand side variables are mutually orthogonal." Discuss, considering in particular the problems of estimating sums and differences of the regression coefficients.

Multicollinearity doesn't matter unless it's perfect (can't get parameter estimate) or it's near perfect (R^2 > 0.95)...

$$y_i=\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}+u_i$$

This is k equations and k unknowns. With perfect multicollinearity, there can be more than 1 solution; one way to get unique solution is to put restriction on parameters

Example: $\beta_1+\beta_2$ so we run $y_i=(\beta_1+\beta_2)x_{1i}+\beta_2(x_{2i}-x_{1i})+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i$

5. A firm is accused of discriminating against women. You propose to test this by running the regression:

(1) $W=\alpha+\beta Q+\lambda S+u$

where W is wages; Q is a qualifications measure; S is 0 for females and 1 for males.

(a) Briefly compare this as a test of discrimination to estimating separate regressions for males and females and testing to see if the regressions are different. (do not get into heavy detail. Just indicate what the tests mean in each case. When you want to do each?)

(b) The firm hires someone who contends that (1) obscures the true structure of the way things work. He (in circumstances it is unlikely to be "she") claims that the true structure is:

(2) $W=\alpha+\beta Q+\lambda S+\delta J+u$

and

(3) $J=\theta_0+\theta_1 Q+\theta_2 S+\varepsilon$

where J is a measure of job placement which equals 0 for blue-collar jobs and 1 for clerical jobs and 2 for managerial jobs. The opposing expert claims that (3) shows how people are assigned to jobs and (2) shows how they are assigned wages given jobs. He claims that a positive estimate for λ shows discriminating in wages given jobs while a positive estimate for θ_2 shows discrimination in job assignment. Estimation of (2) and (3) results in positive estimates of these two coefficients but neither estimate is significantly different form zero. The opponent claims this means there is no convincing evidence of discrimination. Criticize this procedure.

(a) H$_0$: $\lambda=0$; H$_1$: $\lambda>0$ (1 sided test)... use t-test

Problem with this test is that is assumes return to experience (β) is the same for men and women (i.e., same slope)

Better test... similar to #3:

$W_i=\alpha_m+\beta_m Q_i+u_i$, $i=1,\ldots,n$ (men)

$W_i=\alpha_f+\beta_f Q_i+u_i$, $i=n+1,\ldots,N$ (women)

H$_0$: $\alpha_m=\alpha_k$ and $\beta_m=\beta_k$; H$_1$: $\alpha_m\neq\alpha_k$ or $\beta_m\neq\beta_k$... use F-test to look for different slope return on experience and intercept

(b) Combine (2) and (3):

$$W = \alpha + \beta Q + \lambda S + \delta(\theta_0 + \theta_1 Q + \theta_2 S + \varepsilon) + u =$$

$$(\alpha + \delta\theta_0) + (\beta + \delta\theta_1)Q + (\lambda + \delta\theta_2)S) + (u + \delta\varepsilon)$$

So original test actually looked at $(\lambda + \delta\theta_2)$ not just λ

- Could get different result testing λ and θ_2 jointly rather than individually
- More likely, may have introduced multicollinearity so error increased (parameters not significant)

1. Consider a single model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u$$

where PC is a binary variable indicating PC ownership.
(i) Why might PC ownership be correlated with u?
(ii) Explain why PC is likely to be related to parent's annual income. Does this mean parental income is a good IV for PC? Why or why not?
(iii) Suppose that, four years ago, the university gave grants to buy computers to roughly one half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for PC.

(i) There are many potential factors that influence GPA. The effects of these potential regressors have to be captured wither by PC or u. If any of these is also correlated to PC (such as family income), the error term could also be correlated with PC.
Ai - you can always argue an omitted variable; give a story; e.g., PC could be correlated to family income; higher income makes it more likely to have private tutors; So maybe what's really going on is GPA being explained by tutors, not PC.

(ii) Parents who have higher income probably also have higher disposable income and can afford to buy a PC for their kids at college. For a good IV, we want something that is correlated with PC, but not correlated with u. In the case of income, there is still a chance that it is correlated with u. For example, income could be correlated with regional effects (better public school districts) which are captured by u.
Ai - income could be correlated to parent's education (omitted variable) or some other unobserved characteristic of the student (e.g., may be more motivated [or pressured] in school)

(iii) Satisfies 2 conditions: correlated to PC and not correlated to error term (since being selected for the grant was random, it shouldn't be related to any other parameters, so even if they are omitted and correlated to the error term, $Grant$ will not be)

Define $\mathbf{x}_i = \begin{bmatrix} 1 \\ PC_i \end{bmatrix}$, $\mathbf{z}_i = \begin{bmatrix} 1 \\ Grant_i \end{bmatrix}$

Three options:

1. **Directly** - $\hat{\boldsymbol{\beta}}_{\text{IV}} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i ' \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$

2. **2SLS** -

   a. Regress PC on $\mathbf{z}_i$

   b. Generate $P\hat{C} = \hat{\delta}_0 + \hat{\delta}_1 Grant$

   c. Regress GPA on $P\hat{C}$

3. **Stata** - `ivreg GPA (PC = GRANT)`

Ai - another potential instrument would be PC price (determined by market so it's probably not related to other factors that are student dependent)

2. In a recent article, Evans and Schwab (1995) studied the effects of attending a Catholic high school on the probability of attending college. For concreteness, let $college$ be a binary variable equal to unity if a student attends college, and zero otherwise. Let $CathHS$ be a binary variable equal to one if the student attends a Catholic high school. A linear probability model is

$$college = \beta_0 + \beta_1 CathHS + other\_factors + u$$

where the other factors include gender, race, family income, and parental education.
(i) Why might $CathHS$ be correlated with u?
(ii) Evans and Schwab have data on a standardized test score taken when each student was a sophomore. What can be done with this variable to improve the ceteris paribus estimate of attending a Catholic high school?
(iii) Let $CathRel$ be binary variable equal to one if the student is Catholic. Discuss the two requirements needed for this to be a valid IV for $CathHS$ in the preceding equation. Which of these can be tested?
(iv) Not surprisingly, being Catholic has a significant effect on attending a Catholic high school. Do you think $CathRel$ is a convincing instrument for $CathHS$?

(i) Reasons for regressor being correlated to error term: (a) simultaneous decision [LHS and RHS variables being jointly determined], (b) omitted variable, or (c) constraint relating LHS and RHS variables. Given that Catholic schools are private, a student who attends one probably have parents who are more concerned about their child's education and will push harder for them to attend college. In such a situation, one could argue that $college$ and $CathHS$ are jointly determined.

Ai - possible omitted variable for ability; "self-select"... better students go to private schools

(ii) If we assume students at Catholic high schools score better (or worse) on average than other students, we may be able to use the standardized test score as an instrumental variable for $CathHS$. The score is not jointly determined by the parents so it may solve the problem discussed in (i).

(iii) Two requirements is the IV being (highly) correlated to the regressor and being uncorrelated to the error term. The first one can be tested by regressing $CathHS$ on $CathRel$ and look for R^2 > 0.1 and significant coefficient on $CathRel$. Also want to check the impact (magnitude) of the coefficient on $CathRel$ (i.e., check size becase even if it's significant at 99.99%, a value of 0.1 doesn't mean much)

Ai - Can't test the second one unless we have another instrument that we know is good; then model is over identified and we can use the Hausman test

(iv) No. We have to consider the direction of the relationship... there percentage of students who attend Catholic high school that are Catholic may be high, but the percentage of Catholics who attend Catholic high school may not be. (Kind of the smoking-lung cancer problem... % how have lung cancer that smoke is high, but not the other way around.)

Ai - $CathRel$ may be related to error term... didn't really cover why

3. For a large university, you are asked to estimate the demand for tickets to women's basketball games. You can collect time series data over 10 seasons, for a total of about 150 observations. One possible model is

$$\ln attend_t = \beta_0 + \beta_1 \ln price_t + \beta_2 winperc_t + \beta_3 rival_t + \beta_4 weekend + \beta_5 t + u_t$$

where *price* is the price of admission, probably measured in real terms, *winperc*, is the team's current winning percentage, *rival*, is a dummy variable indicating a game against a rival, and *weekend*, is a dummy variable indicating whether the game is on a weekend.
(i) Why is it a good idea to have a time trend in the equation?
(ii) The supply of tickets is fixed by the stadium capacity; assume this has not changed over the 10 years. This means that quantity supplied does not vary with price. Does this mean that price is necessarily exogenous in the demand equation?
(iii) Suppose that the nominal price of admission changes slowly. The athletic office chooses price based partly on last season's average attendance, as well as last season's team success. Under what assumptions is last season's winning percentage a valid instrumental variable for price?
(iv) Does it seem reasonable to include the (log of the) real price of men's basketball games in the equation? Can you think of another variable related to men's basketball that might belong in the women's attendance equation?
(v) If some games are sold out, what problems does this cause for estimating the demand function?

(i) Demand grows over time because of population growth. Since there is no variable for population in the model, including time may work (assuming steady, linear population growth).
Ai - *t* may capture macroeconomic events: population growth, income growth over time, bigger pool for alumni

(ii) Exogenous means $E[price \cdot u] = 0$ (i.e., uncorrelated to error term); since capacity is fixed, Q^D does not have to equal to Q^S, but school is still trying to maximize profit which is based on the capacity ∴ price is not exogenous

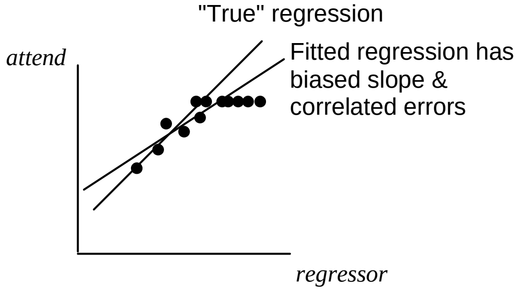
(iii) Want $winperc_{t\text{-}1}$ to be correlated to $price_t$, but not to error term u_t

$$\ln attend_t = \beta_0 + \beta_1(\delta_0 + \delta_1 attend_{t-1} + \delta_2 winperc_{t-1}) + \beta_2 winperc_t +$$
$$\beta_3 rival_t + \beta_4 weekend + \beta_5 t + u_t$$

Ai - $price_t$ depends on $attend_{t\text{-}1}$ and $winperc_{t\text{-}1}$

(iv) The price of men's basketball games could make sense in the sense that men's games could be viewed as a substitute for the women's games. Unless the games are on the same night, however, the correlation may not be as strong. Another variable related to men's basketball that would be better is a binary variable: 1 if there is a men's game (home or away) at the same time as the women's game.
Ai - relative winning percentage of men vs. women (i.e., which team is doing better)

(v) Linear regression wouldn't work well because *attend* would not be linear... it will result in biased coefficient estimates and possibly correlated error terms



4. Discuss test and correction for heteroskedasticity and error term correlation in the 2SLS framework.

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$

Heteroskedasticity in 2SLS - i.e., $E(u_i^2 \mid \mathbf{z}_i) \neq \sigma^2$

Detecting -

(1) run 2SLS and get $\hat{\boldsymbol{\beta}}$

(2) compute <u>consistent</u> residuals: $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

(3) regress e_i^2 on $(1\ \mathbf{z}_i)$ (i.e., be sure to include a constant term if it's not already in $\mathbf{z}_i$)

(4) do overall test of significance (i.e., standard F-test to check if all parameters are simultaneously equal to zero)... if regression is significant, there's heteroskedasticity

Correcting -

(1) save fitted value of $\hat{e}_i^2$ (from regression in step (3) above)

(2) transform model: $\dfrac{y_i}{\hat{e}_i} = \dfrac{\mathbf{x}_i'}{\hat{e}_i}\boldsymbol{\beta} + \dfrac{u_i}{\hat{e}_i} = \beta_1 \dfrac{x_{1i}}{\hat{e}_i} + \hat{\beta}_2 \dfrac{x_{2i}}{\hat{e}_i} + \hat{\beta}_3 \dfrac{x_{3i}}{\hat{e}_i} + \dfrac{u_i}{\hat{e}_i}$

(3) do 2SLS on the transformed model; can use $\mathbf{z}_i = \begin{bmatrix} w_i \\ x_{2i} \\ x_{3i} \end{bmatrix}$ or $\mathbf{z}_i = \begin{bmatrix} w_i / \hat{e}_i \\ x_{2i} / \hat{e}_i \\ x_{3i} / \hat{e}_i \end{bmatrix}$... will give different results, but both have same statistical properties

Serial Correlation in 2SLS -

Detecting -

(1) same (1) and (2) as heteroskedasticity

(3) run $e_i = \rho e_{i-1} + \boldsymbol{\gamma}_i$ (or any other form); if $\hat{\rho}$ is significantly different than zero, there's serial correlation

Correcting -

(1) transform model:

$$(y_i - \hat{\rho} y_{i-1}) = \beta_1 (x_{1i} - \hat{\rho} x_{1i-1}) + \beta_2 (x_{2i} - \hat{\rho} x_{2i-1}) + \beta_3 (x_{3i} - \hat{\rho} x_{3i-1}) + (u_i - \hat{\rho} u_{i-1})$$

(2) do 2SLS on the transformed model; can use $\mathbf{z}_i - \hat{\rho}\mathbf{z}_{i-1}$, $\mathbf{z}_i$, or $\mathbf{z}_{i-1}$ (will have same statistical properties)

Using Stata

Welcome to Stata, one of the most un-user friendly programs ever created. Although later versions have some features to make it easier to use, they more than make up for it by not being 100% compatible with previous versions... and some even generate code through the "helpful" dialog boxes that cause errors. This tutorial is intended to give you the very basic tools needed to get things done in Stata.

How this document is organized:

- `Courier font` is what you type or see in Stata.
- Things in `[brackets]` are optional arguments.
- `Blue text` are reserved commands in Stata; the underlined part of the next is how you can abbreviate the command.
- `Red text` are Stata error messages.
- File extensions are listed (e.g., `.do`, `.dta`, `.log`) to let you know what Stata is expecting. Stata doesn't not require the extensions to be used if the file is of the type expected.

How Stata Works

Stata uses a combination of command-line and menu-driven inputs along with a very basic spreadsheet-style data editor. There are several types of files you can work with in Stata, but the basic one is the Stata data set file (`.dta`). This file contains all your data as well as variable names.

Extension	File Type
.dta	Stata data file
.log	Log file (explained below)
.raw	ASCII (text) file

PgUp and PgDn buttons scroll through commands in the Review window (i.e., writes them in the Stata Command window for you)

Data Sets

You can view the data Stata is working with by opening the data editor or data browser. Both of these work similar to a spreadsheet with the variables listed in columns. The only difference is that the editor lets you change values and the browser doesn't. If you insist on using the command line, you can use the list command. Although it's old school, list could help find your problem areas when used in conjunction with if. For example:

```
list varname if varname > 5000
```

will list all observations of the variable varname that are greater then 5000. The good thing is that each one has the observation number. You can jot that down and look up the data points in the data editor or you can go back to your original data source to track down potential problems.

Save Data

```
save filename.dta [,*]
```

Options:

- nolabel - omits value labels; still saves associations between variables and value label names (just not the labels themselves).
- replace - allows you to overwrite the existing file; prevents "file filename.dta already exists" error.
- orphans - saves all value labels, including those not attached to any variables
- emptyok - allows you to save an empty data set to prevent "no variables defined" error. (Used for programming.)
- intercooled - makes Stata/SE save in Intercooled Stata format.

Load Data (.dta Files)

```
use filename.dta [,*]
```

Options:

- clear - Stata will not let you load a data file if you already have a data in memory. Using the clear option removes any data from memory (even if it hasn't been saved) to allow you to load the data file.

Load Data (Other Sources)

Formatted Text File - one observation per line; values are tab or comma delimited; can have variable names in the first line (optional); if you don't include file extension, .raw is assumed.

```
insheet [varlist] using filename.raw [,*]
```

Options:

- varlist - list variables names separated by spaces (not commas)
- double - forces Stata to store variables as doubles (rather than floats)
- no[names] - informs Stata whether variable names are included; Stata will figure it out on its own, but this option will allow the file to open faster
- comma - specify comma delimited (not required)
- tab - specify tab delimited (not required)

`delimiter("*")` - specify a different delimiter in the data (e.g., `delimiter(";")`)

`clear` - removes any data from memory (even if it hasn't been saved) to allow you to load the data file.

Examples:

```
insheet using newdata
insheet using newerdata.txt, clear
insheet using weirddata.txt, clear delimiter("&")
insheet height gender mom dad using heights.dat
```

Log Files

These keep track of everything that happens during your Stata session by recording everything that appears in the Stata Results window (the one with the black background). A log file can be handy for tracking down errors when you're running a `.do` file. If you specify the `.log` extension, the file is saved in ASCII (text) format which means the colors are not saved, but it's pretty easy to tell the difference between your commands and Stata output because commands are preceded by a period (`.`). There's a different format for the Stata viewer, but it's not really any better than a text file. There are also other options for a log file than aren't covered here, but this section should give you all you need to know. The only tricky part is deciding where (or if) to turn the log file on or off during execution of your .do file or Stata session. You don't really need to close the log file to be able to read it.

```
log using filename.log [,*]
log off
log on
log close
```

Options:

`replace` - overwrites current log file

`append` - adds this session to the end of the log file

text |

Examples:

```
log using newlog.log, replace
log using "file with spaces.log"
log close
```

Commands

Stata commands are the things that get things done in Stata. They are how you tell Stata to do what it is you want done.

Note: `exp` refers to any expression, logical or mathematical; the type should be clear in the context; if `exp` is written twice in a single line, it does not imply that it is the same expression. Expressions use the following operators:

Arithmetic		Logical		Relational	
+	addition	~	not	>	greater than
-	subtraction	!	not	<	less than
*	multiplication	|	or (shift \)	>=	> or equal
/	division	&	and	<=	< or equal
^	power			==	equal
+	string concatenation			~=	not equal
				!=	not equal

Generate - creates a new variable based on `exp`

```
generate [type] newvar[:lblname] = exp [if exp]
```

Options:

`type` - specifies the variable type; if none is specified, Stata will automatically select `float` for numeric data and `str` for text

Examples:

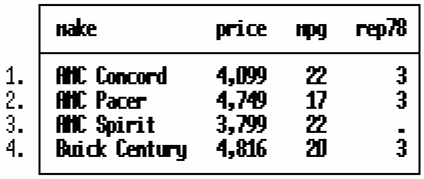
```
generate age2 = age*age
generate biginc = income>100000 & income!=.
gen double unitpr = cost/quantity
gen byte biginc = income>100000 & income!=.
gen xlag = x[_n-1]
```

List - prints data on the screen

```
list [varlist] [if exp] [, *]
```

Options:

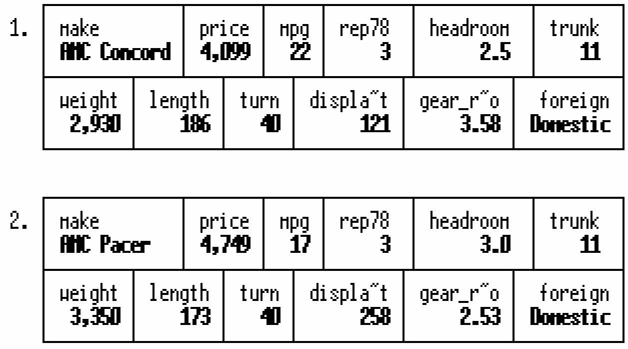
`table` - lists variables vertically, one observation per row



```
       make            price   mpg   rep78

  1.   AMC Concord     4,099    22       3
  2.   AMC Pacer       4,749    17       3
  3.   AMC Spirit      3,799    22       .
  4.   Buick Century   4,816    20       3
```

display - lists observations together; useful if there are a lot of variables to keep it from wrapping around the screen



```
1. | make        | price | mpg | rep78 | headroom | trunk |
   | AMC Concord | 4,099 |  22 |     3 |      2.5 |    11 |
   |-------------------------------------------------------|
   | weight | length | turn | displa~t | gear_r~o | foreign |
   |  2,930 |    186 |   40 |      121 |     3.58 | Domestic|

2. | make        | price | mpg | rep78 | headroom | trunk |
   | AMC Pacer   | 4,749 |  17 |     3 |      3.0 |    11 |
   |-------------------------------------------------------|
   | weight | length | turn | displa~t | gear_r~o | foreign |
   |  3,350 |    173 |   40 |      258 |     2.53 | Domestic|
```

Replace - changes the contents of an existing variable

```
replace oldvar = expression1 [if expression] [, nopromote]
```

Options:

oldvar - name of a variable that already exists in the data set

nopromote - prevents replace from promoting the variable type to accommodate the change (e.g., if you replace an integer variable with data containing 3.14 and prevent the type to promote, you'll end up with 3)

Examples:

```
replace income=. if income<=0
replace age = 25 in 1007
```

Set Memory - specifies how much system memory you want to be dedicated to Stata ; **Note:** typing memory without set before it will display a report of Stata's memory usage

```
set memory #[b|k|m|g] [, permanently ]
```

Options:

# - amount of memory to set; specified in terms of bytes (b), kilobytes (k), megabytes (m), or gigabytes (g)

permanently - specifies that in addition to making the change right now, Stata will remember the new limit and use it in the future when you open Stata

Examples:

```
set memory 5m
```

Set Type - specifies the default data type assigned to new variables (such as by generate) when the storage type is not explicitly specified

```
set type *
```

where * is either a numeric storage type listed here or a string explained below the table

Numeric Storage Type	Bytes	Minimum	Maximum	Closets to 0 without being 0
`byte`	1	-127	100	+/-1
`int`	2	-32,767	32,740	+/-1
`long`	4	-2,147,483,647	2,147,483,620	+/-1
`float`	4	-1.70141173319*10^38	1.70141173319*10^36	+/-10^-36
`double`	8	-8.9884656743*10^307	8.9884656743*10^308	+/-10^-323

Precision for float is 3.795x10^-8
Precision for double is 1.414x10^-16

Character strings are specified by `str#`, where # gives the maximum length of the string (ranges from 1 to 80). Each character reserved by a string takes one byte regardless of the data stored in the string (e.g., "`it`" stored in a variable of type `str80`, still takes up 80 bytes).

Summarize

```
summarize [varlist] [if expression] [, detail]
```

Options:

- `varlist` - list of variables, separated by spaces (not commas); if you don't indicate a variable list, Stata will summarize all the variables in the data set
- `if expression` - allows you to specify a subset of the data to be summarized
- `detail` - standard summarize command lists number of observations, mean, standard deviation, minimum and maximum; specifying `detail` adds 1, 5, 10, 15, 75, 90, 95, 99th percentiles, variation, skewness, and kurtosis

Functions

Functions are actually series of embedded commands designed to accomplish a specific task. They make working with Stata a little easier because you don't have to program them in yourself. This is just a subset of frequently used functions. You can get more functions by using the online help in Stata and searching for these.

Type of function	See help
Mathematical Functions	`mathfun`
Probability Functions	`probfun`
Random Numbers	`random`
String Functions	`strfun`
Programming Functions	`progfun`
Date Functions	`datefun`
Time-series Functions	`tsfun`

Mathematical Functions

`abs(x)` returns the absolute value of x

`exp(x)` returns the ex

`int(x)` returns the integer obtained by truncated x towards zero

`ln(x)` or `log(x)` returns the natural logarithm of x

`log10(x)` returns the base 10 logarithm of x

`max(x1,x2,...,xn)` returns the maximum of x1, x2, ..., xn (missing values are ignored)

`min(x1,x2,...,xn)` returns the minimum of x1, x2, ..., xn (missing values are ignored)

`round(x,y)` returns x rounded off to units of y

`sqrt(x)` returns the square root of x

Probability Functions

`binomial(n,k,p)` returns the probability of k or more successes in n trials when the probability of a success on a single trial is p

`chi2(n,x)` returns the cumulative chi-squared distribution with n degrees of freedom

`chi2tail(n,x)` returns the reverse cumulative (upper-tail) chi-squared distribution with n degrees of freedom; `chi2tail(n,x) = 1 - chi2(n,x)`

`F(n1,n2,f)` returns the cumulative F distribution with n1 numerator and n2 denominator degrees of freedom

`Fden(n1,n2,f)` returns the probability density function for the F distribution with n1 numerator and n2 denominator degrees of freedom

`Ftail(n1,n2,f)` returns the reverse cumulative (upper-tail) F distribution with n1 numerator and n2 denominator degrees of freedom; `Ftail(n1,n2,f) = 1 - F(n1,n2,f)`

`invbinomial(n,k,P)` returns the inverse binomial: for P<=0.5, probability p such that the probability of observing k or more successes in n trials is P; for P>0.5, probability p such that the probability of observing k or fewer successes in n trials is 1-P.

`invchi2(n,p)` returns the inverse of `chi2()`; if `chi2(n,x) = p`, then `invchi2(n,p) = x`

`invF(n1,n2,p)` returns the inverse cumulative F distribution; if `F(n1,n2,f) = p`, then `invF(n1,n2,p) = f`

`invnorm(p)` returns the inverse cumulative standard normal distribution; if `norm(z) = p`, then `invnorm(p) = z`

`norm(z)` returns the cumulative standard normal distribution

`normden(z)` returns the standard normal density

`normden(x,m,s)` returns the normal density with mean m and standard deviation s; `normden(x,m,s) = normden((x-m)/s)/s`

`tden(n,t)` returns the probability density function of Student's t distribution with n > 0 degrees of freedom

`ttail(n,t)` returns the reverse cumulative (upper-tail) Student's t distribution with n > 0 degrees of freedom

Random Numbers

`uniform()` returns uniformly distributed pseudo-random numbers on the interval [0,1)

`invnorm(uniform())` returns normally distributed random numbers with mean zero and standard deviation one

String Functions

Programming

Data Functions

Time-series Functions

Matrix Functions

set seed #
uniform()
invnorm(uniform())

sum(x)
sum(x!=.)

Regression

Basic Regression

```
regress depvar [varlist] [,*]
```

Options:

`depvar` - name of dependent variable

`varlist` - list of independent variables, separated by spaces (not commas)

`level(#)` - specifies the confidence level (e.g., 95) for confidence intervals of the coefficients

`noconstant` - suppresses the constant (intercept) term

`robust` - uses the White Heteroskedasticity Consistent Covariance Estimator; results in higher standard errors and lower *t*-ratios

Examples:

```
regress y x1 x2
reg height gender mom dad, level(95)
reg consumption output, noconstant
```

Using Results

Parameter Estimates - returns the estimated coefficient for `regressorname`

```
_b[regressorname]
```

Predict - generates a new variable that stores the designated prediciton based on the last regression run by Stata

```
predict newvarname [,statistic]
```

Statistic:

`xb` - fitted values; sample point estimate; this is the default so you don't need to include it

`residuals` - residuals (dependent variable minus ybar)

`rstandard` - standardized residuals

`stdp` - standard error of each predicted value (i.e., $Stdev(\hat{y}_i)$)

`stdf` - standard error of each forecasted value

`stdr` - standard error of each residual

Variance - displays the variance-covariance matrix (i.e., $Var(\hat{\boldsymbol{\beta}})$)

```
vce
```

Testing Linear Hypotheses After Estimation

`test coeflist` - test that coefficients are equal 0; list coefficients separated by spaces
`test exp = exp [= ...]` - test that linear expressions are equal

Options:
`accumulate` - adds test to previous test(s) in memory making a joint test

Note: This performs the Wald Test... approximated with an F distribution instead of chi-square

***F*-Test** - to do a real F-test of m restrictions:

1. Run the unrestricted regression: $y_i = x_{1i}\hat{\beta}_1 + x_{2i}\hat{\beta}_2 + x_{3i}\hat{\beta}_3 + \cdots + x_{ki}\hat{\beta}_k + \hat{u}_i$
2. Record SSR (just on paper if you want)
3. Run the restricted regression: $\tilde{y}_i = \tilde{x}_{1i}\tilde{\beta}_1 + \cdots + \tilde{x}_{ki}\tilde{\beta}_k + \tilde{u}_i$
4. `generate F = ((RstctdSSR - UnrstctdSSR)/m)/(UnrstctdSSR/(N-k))`
5. Compare that to an F(2,N-k)... display `Ftail(m,N-k,F)`

Example -

```
regress lwage educ huswage city unem exper expersq
```

Using Wald Test:

```
test uduc-expr = 0
test city + unem = 0, accumulate
```

Returns 3.96... p-value 0.0199

Using F-Test:

```
generate edex = educ - exper
generate ctun = city + unem
regress lwage edex huswage ctun expersq
generate F = ((190.12475-186.55)/2)/(186.55/(428-7))
```

Returns 4.022... p-value 0.0186

Advanced Regression Techniques

Options:

`beta` - requests that normalized beta coefficients be reported instead of confidence intervals, if the original model is $y = x_1\beta_1 + x_2\beta_2 + u$, beta alters the model to be

$$\frac{y-\bar{y}}{Stdev(y)} = \frac{x_1-\bar{x}_1}{Stdev(x_1)}\tilde{\beta}_1 + \frac{x_2-\bar{x}_2}{Stdev(x_2)}\tilde{\beta}_2 + \tilde{u}$$, where $\tilde{\beta}_i = \beta_i Stdev(x_i)$

`cluster [varname]` - `varname` describes ID variable to allow correlation between errors within a cluster

Heterskedasticity - here's a series of commands to deal with heteroskedasticity; assume only x2 and x3 are correlated to the error terms

```
regress y x1 x2 x3
predict e, residuals
generate e2 = e^2
regress e2 x2 x3
predict sigma2
```

Method 1 - Transform Model

```
generate newy = y/sqrt(sigma2)
generate newx1 = x1/sqrt(sigma2)
etc.
regress newy newx1 newx2 newx3
```

Method 2 - Weights

```
regress y x1 x2 x3 [weight = sigma2]
```

generating lagged variables - `generate lagy = y[_n-1]`

Regressors Correlated with Error Terms - use instrumental variable estimation and the Hausman test

```
ivreg depvar [varlist] (varlist2 = varlist_iv) [,*]
```

Options:

- `varlist2` - list of independent variables that are correlated with the error term
- `varlist_iv` - list of instrument variables used in place of the variables in `varlist2`
- Other options are same are `regress` command

`hushrs` (husband hours) is probably a joint decision when deciding the wife's hours (`hours`), so it's probably correlated with the error term; suppose `huseduc` is known to be a good instrument; test if `huswage` is also a good instrument:

```
ivreg hours kidslt6 educ wage famine unem (hushrs = huseduc),
   robust
hausman, save
ivreg hours kidslt6 educ wage famine unem (hushrs = huswage
   huseduc), robust
hausman
```

Seemingly Unrelated Regression (SUR) - simultaneous equations using pooled data (i.e., cross-section data over time that may not necessarily be from same source)

```
sureg (depvar1 varlist1 [,noconstant]) (depvar2 varlist2) ...
```

Options:

- `noconstant` - omits constant term for specified equations `isure` - iterate over the estimated disturbance covariance matrix and parameter estimates until the parameter estimates converge; better finite sample properties

`dfk` - use alternate divisor in computing the covariance matrix for the equation errors; better estimates for small samples

`small` - specifies that small sample statistics are to be computed; shifts test statistics from chi-squared and Z statistics to F statistics and t-statistics

Examples:

3 simultaneous equations:

```
sureg (price foreign weight length) (mpg foreign weight)
  (displ foreign weight)
```

Test if coefficient for `foreign` is zero is all equations:

```
test foreign
```

Test across equations

```
test [price] foreign = [mpg] foreign
```

Problem with Heteroskedasticity or Serial Correlation -

Run simple OLS on stacked data (use $n = \min(n_1 n_2)$; drop extra data)

Create new variable to account for pairs

```
regress y x1 x2, cluster[d] robust
```

$$\mathbf{d} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \\ 1 \\ 2 \\ \vdots \\ n \end{bmatrix}$$

Fixed Effect Regression with Panel Data -

```
xtreg depvar [varlist], type i(varname)
```

Options:

Type is one of the followed depending on which estimation technique is used:

`be` - between-effects estimator: $y_{i\bullet} = \beta_{0i} + \mathbf{x}_{i\bullet}'\boldsymbol{\beta} + u_{i\bullet}$

`fe` - fixed-effects estimator: $(y_{it} - y_{i\bullet}) = (\mathbf{x}_{it} - \mathbf{x}_{i\bullet})'\boldsymbol{\beta} + (u_{it} - u_{i\bullet})$

where $y_{i\bullet} = \dfrac{1}{T}\sum_{t=1}^{T} y_{it}$, $x_{mi\bullet} = \dfrac{1}{T}\sum_{t=1}^{T} x_{mit}$

`re` - GLS random-effects estimator

`pa` - GEE population-averaged estimator

`mle` - maximum-likelihood random-effects estimator

`i(varname)` - specifies the variable corresponding to an independent unit (e.g., a subject id); this variable represents the i in x_{it} (similar to `cluster`)

Output:

Reports # Observations, # Groups (individuals), min, max and avg Obs/Group

`R-Sq`... only care about overall... that's the one based on the original model:

$$y_{it} = \sum_{j=1}^{N} \beta_{0j} d_{jit} + \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$$

`F(##, ###)` (above table) testing H$_0$: $\beta_i = 0$ (i.e., all parameters are zero)... this is the standard F-test checking all parameters simultaneously for a regression

`Const` = $\hat{\beta}_0 = \dfrac{1}{N}\sum_{i=1}^{n} \hat{\beta}_{0i}$

`Corr(u_i,xb)` = Corr($\hat{\beta}_{oi}, \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}$)... this is to check the assumption of the random effect model which assumes $E(\beta_{0i}\mathbf{x}_{it}) = \mathbf{0}$

`Sigma_u` = standard deviation of u_{it}

`Sigma_e` = standard deviation of $\hat{\beta}_{0i}$

`F(##, ###)` (below table) testing H$_0$: $Var(\beta_{0i}) = 0$; i.e., whether individual effect is correlated with regressors (or all the same); numerator degrees of freedom is N + k; denominator is NT - (N + k) (assuming same number of time periods per individual)... another way to think of this test is a test on whether N - 1 dummy variables are simultaneously equal to zero (1 dummy is left out and captured with the constant term in the regression)

Programming

General Program

Specify Stata Version - some commands and formats are specific to the version of Stata (so copying someone else's code made not work in your version of Stata). If the person wrote it in a previous version, you may be able to get away with a simple command that allows the older code to work. Type the version number you want to emulate at the beginning of the file:

```
version 8
```

Comments

`*` Used at the beginning of a line; the line is ignored.

`/* */` Used in the middle of a line; everything between `/*` and `*/` is ignored.

`//` Used at the beginning or end of a line (must be preceded by one or more blanks if at the end); everything on the line after `//` is ignored.

`///` Instructs Stata to view from `///` to the end of a line as a comment, and to join the next line with the current line; must be preceded by one or more blanks; used make long lines more readable.