Contents

Chapter 1

Preamble

This econometric guide is intended for internal use of the Research Computing Services at Harvard Business School. I heavily borrowed materials from books, such as Davidson and MacKinnon's "Econometric Theory and Methods", Chris Baum's "An Introduction to Modern Econometrics using Stata", etc., and some materials from the web. It is not intended for publication.

I try to include some sample codes after the concepts, in R or Stata, depending on which one has the procedure implemented, or which one is easier to use.

Chapter 2

OLS

A linear model with n observations and k regressors can be written as (in vector forms)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

The idea of least-squares is to choose β to minimize the residual sum of squares (SSR),

$$SSR = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}'\mathbf{y} - \mathbf{2}\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta$$

The first-order condition to minimize SSR are:

$$\frac{\partial(SSR)}{\partial\beta} = -\mathbf{2}\mathbf{X}'\mathbf{y} + \mathbf{2}\mathbf{X}'\mathbf{X}\beta = \mathbf{0}$$

This generates the so-called normal equations

$$(\mathbf{X}'\mathbf{X})\beta = \mathbf{X}'\mathbf{y}$$

Therefore, Ordinary Least Squares (OLS) estimator of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

Fitness of OLS

We can see y vector as the part explained by the regression and the unexplained part,

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{u} = \mathbf{X}\beta + \mathbf{u}$$

Therefore

$$\mathbf{y}'\mathbf{y} = (\hat{\mathbf{y}} + \mathbf{u})'(\hat{\mathbf{y}} + \mathbf{u}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{u}'\mathbf{u} = \beta'\mathbf{X}'\mathbf{X}\beta + \mathbf{u}'\mathbf{u}$$

Subtracting $n\bar{y}^2$ ($\bar{y}$ is the sample mean) from both sides,

$$\mathbf{y}'\mathbf{y} - \mathbf{n}\bar{\mathbf{y}}^{\mathbf{2}} = (\beta'\mathbf{X}'\mathbf{X}\beta - \mathbf{n}\bar{\mathbf{Y}}^{\mathbf{2}}) + \mathbf{u}'\mathbf{u}$$

We decompose the total sum of squares into two parts: sum of squares due to error (noise), and sum of squares explained by the linear regression.

The R^2 is defined by

$$R^2 = 1 - \frac{SSR}{SST}$$

SST is the total sum of squares, which is the total variance of y, $\mathbf{y}'\mathbf{y} - \mathbf{n}\bar{\mathbf{y}}^{\mathbf{2}}$.

R^2 is simply the proportion of the variation of Y that can be attributed to the variation of X. R^2, however, will never decrease with the addition of any variable to the set of regressors. If the added variable is totally irrelevant then R^2 will stay the same. The adjusted R^2, however, takes account of the addition of any new regressors:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k)}{SST/(n-1)}$$

Example

Let's look at an example of OLS of car's mpg on disp, hp and wt.

```
# load libraries I'll need later.
library(car)
library(stats4)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(mvtnorm)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(AER)
```

```
## Error in library(AER): there is no package called 'AER'
```

```
library(sem)
library(gmm)
```

```
## Loading required package: sandwich
##
## Attaching package: 'gmm'
##
## The following object is masked from 'package:sem':
##
##     tsls
```

```
lm1 <- lm(mpg~disp+hp+wt, data=mtcars)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.891 -1.640 -0.172  1.061  5.861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
## disp        -0.000937   0.010350  -0.091  0.92851
## hp          -0.031157   0.011436  -2.724  0.01097 *
## wt          -3.800891   1.066191  -3.565  0.00133 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.639 on 28 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8083
## F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

The Geometry of Least Squares

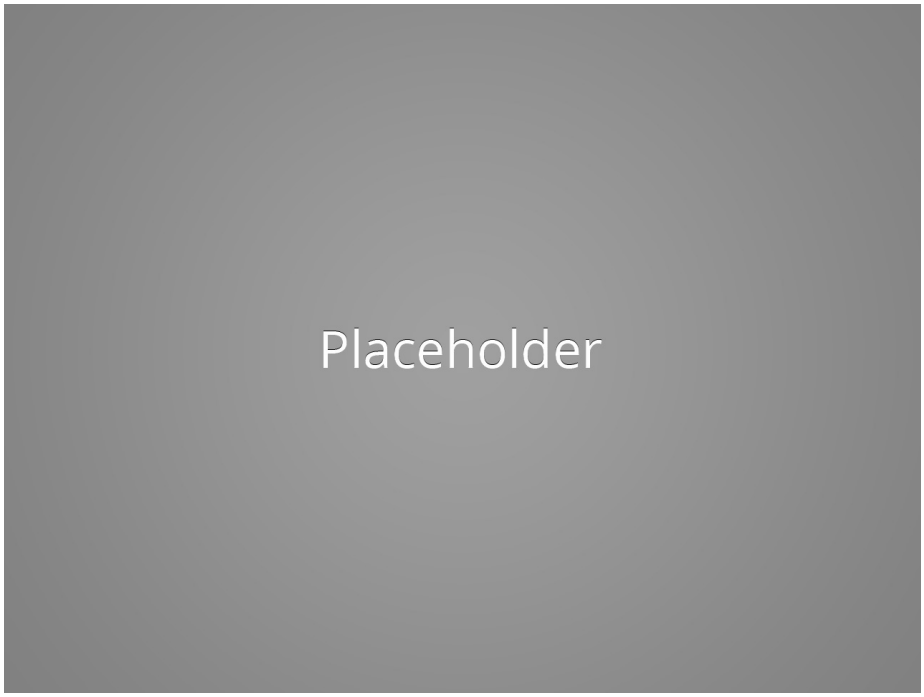


Figure 2.1:

For simplicity, let's assume there are two explanatory variables x_1 and x_2. x_1 and x_2 form a plane. What OLS does is to project y onto this plane.

In mathematical terms, all linear combinations of these two vectors define a two-dimensional subspace of Euclidean space $\mathbf{E}^n$. This is called the column space of $\mathbf{X}$. The least-square principle is to choose β to make $\hat{y}$, which belongs to the subspace of $\mathbf{X}$, as close as possible to y.

When we estimate a linear regression model, we map the y into a vector of fitted values $\mathbf{X}\hat{\beta}$ and a vector of residuals $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$. Geometrically, these are examples of orthogonal projections. A projection is a mapping that takes each

point of $\mathbf{E}^n$ into a point in a subspace of $\mathbf{E}^n$. An orthogonal projection maps any point into the point of the subspace that is closest to it.

An orthogonal projection can be performed by premultiplying the vector to be projected by a projection matrix. In the case of OLS, the two projection matrices that yield the vector of fitted values and the vector of residuals, are

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$$

$$\mathbf{M_X} = \mathbf{I} - \mathbf{P_X} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$$

From this, we see that the effects of two projection matrices.

$$\mathbf{P_X y} = \mathbf{X}\hat{\beta} = \hat{\mathbf{y}}$$

$$\mathbf{M_X y} = (\mathbf{I} - \mathbf{P_X})\mathbf{y} = \mathbf{y} - \mathbf{P_X y} = \mathbf{y} - \mathbf{X}\hat{\beta} = \hat{\mathbf{u}}$$

That is, $\mathbf{P_X y}$ projects $\mathbf{y}$ onto $\mathbf{X}$ and makes it $\hat{\mathbf{y}}$; $\mathbf{M_X y}$ makes it $\hat{\mathbf{u}}$.

In the picture, that corresponds to the fact that the projection of $\mathbf{y}$ onto the plane of $\mathbf{X}$ creates two parts: $\hat{\mathbf{y}}$ and $\hat{\mathbf{u}}$.

Properties of OLS estimator

Unbiasedness

In a linear model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u},$$

the OLS estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

Since $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$,

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{u}.$$

This makes

$$\mathrm{E}(\hat{\beta}) = \beta + (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathrm{E}(\mathbf{u}|\mathbf{X}).$$

The condition that makes the OLS estimator unbiased is:

$$\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0},$$

that is, all explanatory variables which form the columns of $\mathbf{X}$ are exogenous. This condition is weaker than the independence condition that u and X are

independent. This says that given $\mathbf{X}$, the expected value of $\mathbf{u}$ is zero; it implies that the model is correctly specified. That is, $\mathbf{y}$ is a linear function of $\mathbf{X}$.

In the context of cross-sectional data, this assumption is plausible. However, when we have time series data, the assumption becomes strong, because it assumes that the entire series of $\mathbf{X}$ has no relationship with the error term. In a time series context, this is hard to satisfy. The OLS estimator is biased if this condition is not satisfied.

For example, suppose we have a model

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

In this simple model, even if we assume that y_{t-1} and u_t are uncorrelated, OLS estimator is still biased. That is because $\text{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ is not satisfied: y_{t-1} depends on u_{t-1}, u_{t-2} and so on.

There is a weaker condition:

$$\text{E}(\mathbf{X}\mathbf{u}) = \mathbf{0},$$

Once we know that the model is correctly specified, then this equation can be used to derive results, such as GMM estimators.

Consistency

For OLS estimator to be consistent, a much weaker condition is needed:

$$\text{E}(u_t|X_t) = 0,$$

This condition is much weaker since it only assumes that the mean of current error term does not depend on the current predictors. Even a model with lagged dependent variable can easily satisfy this condition. This condition is called predeterminedness condition, or say regressors are predetermined. So in the time series example, OLS estimator is biased, but can be consistent, if we are willing to assume no contemporaneous correlation.

Chapter 3

The Method of Maximum Likelihood

If we make the assumption that the error terms are normally distributed, the maximum likelihood estimators (MLE) coincide with the least square estimators. But MLE can also be applied to a wide variety of models other than regression models, and it generally generates estimators with excellent asymptotic properties. The major disadvantage of MLE is that it requires stronger distributional assumptions than does the method of moments.

Basic Concepts

Maximum likelihood estimation is to find the set of parameters which makes the current sample most likely given the statistical model.

Suppose we have a sample of n independent and identically distributed (IID) observations. Each of them comes from some distribution function $f(\cdot, \theta)$, where parameters θ is fixed. The joint distribution of all the data points would be

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta).$$

It is referred to as the likelihood function of the model for the given data set.

Then it's natural to estimate θ by picking the θ that maximize this joint distriubtion function. A parameter vector $\hat{\theta}$ at which the likelihood takes on its maximum value is called a maximum likelihood estimate, or MLE, of the parameters.

Usually the log form of the likelihood function is preferred for computational purposes.

MLE for a linear model

Consider the classical normal linear model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$$

In this case, we have two parameters to estimate, σ and β. The distribution of $\mathbf{y}|\mathbf{X}$ is

$$f(\mathbf{y}, \beta, \sigma|\mathbf{X}) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(\mathbf{y} - \mathbf{X}\beta)^2}{2\sigma^2})$$

The log-likelihood function is

$$l(\mathbf{y}, \beta, \sigma) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log 2\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

The maximization of the log-likelihood function will lead to the same estimator as the OLS estimator, for β. For σ^2, the MLE estimator is slightly different in limited sample, but asymptotically the same. This is strictly based on the assumption of normally distributed error term. Otherwise, these two estimators will be different.

Asymptotic Properties of MLE

1. Consistency
$$\text{plim}(\hat{\theta}) = \theta.$$

This is to say that MLE estimator can get arbitrarily close to the true parameter if the sample size gets to infinity.

2. Asymptotic Normality
$$\hat{\theta} \sim^a N(\theta, \mathbf{I^{-1}}(\theta)).$$

This states that the asymptotic distribution of $\hat{\theta}$ is normal with mean θ and variance given by the inverse of $\mathbf{I}(\theta)$, the information matrix, defined by

$$\mathbf{I}(\theta) = E[(\frac{\partial l}{\partial \theta})(\frac{\partial l}{\partial \theta})'] = -E[(\frac{\partial^2 l}{\partial \theta \partial \theta'})]$$

The information matrix is the expected value of the inner product of gradient of the log-likelihood function, or the negative value of the expected value of the Hessian (second derivative of the log-likelihood function.

3. Asymptotic efficiency.

If $\hat{\theta}$ is the MLE estimator of θ, $\mathbf{V}$ denotes the variance covariance matrix of $\hat{\theta}$, then

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow^d N(\mathbf{0}, \mathbf{V})$$

If $\tilde{\mathbf{V}}$ denotes the variance matrix of any other consistent, asymptotically normal estimator, then $\tilde{\mathbf{V}} - \mathbf{V}$ is a positive semidefinite matrix. That means MLE estimator is the best (in terms of variance) consistent and asymptotically normal estimator.

4. Invariance.

If $\hat{\theta}$ is the MLE of θ and $g(\theta)$ is a continuous function of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Example

Let's look at the example of regression of car's mpg on disp, hp and wt. This time we use MLE. (R has a mle() function which I cannot make it work on this example. So I am using optim().)

```
ols.lf <- function(theta, y, X) {
  beta <- theta[-1]
  sigma2 <- theta[1]
  if (sigma2 <= 0) return(NA)
  n <- nrow(X)
  e <- y - X%*%beta
  logl <- ((-n/2)*log(2*pi)) - ((n/2)*log(sigma2)) - ((t(e)%*%e)/(2*sigma2))
  return(-logl)
}
X <- cbind(1,mtcars[,c('disp','hp','wt')])
y <- mtcars$mpg

# note that I have a rough idea of the starting values from OLS esimators first; otherwise it may
optim(c(1,1,-1,-1,-1), method="L-BFGS-B", fn=ols.lf, lower=c(1e-6,-Inf,-Inf,-Inf,-Inf), upper=rep
```

```
## $par
## [1]  6.0935580702 37.1058998652 -0.0009352804 -0.0311594264 -3.8010038304
##
## $value
## [1] 74.32149
##
```

```
## $counts
## function gradient
##      126      126
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
```

```
# it should match the OLS estimator

lm1 <- lm(mpg~disp+hp+wt, data=mtcars)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.891 -1.640 -0.172  1.061  5.861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
## disp        -0.000937   0.010350  -0.091  0.92851
## hp          -0.031157   0.011436  -2.724  0.01097 *
## wt          -3.800891   1.066191  -3.565  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.639 on 28 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8083
## F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

Hypotheses Testing

Linear Hypotheses Testing

Linear hypothesis can be framed as

$$\mathrm{H}_0: \quad \mathbf{R}\beta - \mathbf{r} = \mathbf{0},$$

where $\mathbf{R}$ is a $q \times k$ matrix of known constants, with $q < k$, and $\mathbf{r}$ is a q-vector of known constants.

We assume that u_i are normally distributed.

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$$

Even if this assumption does not hold, we still have asympototic normality for $\mathbf{u}$.

Since linear combination of normal variables are also normally distributed,

$$\hat{\beta} \sim \mathbf{N}(\beta, \sigma^{\mathbf{2}}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}})\mathbf{R}\hat{\beta} \sim \mathbf{N}(\mathbf{R}\beta, \sigma^{\mathbf{2}}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{R}')\mathbf{R}(\hat{\beta}-\beta) \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{R}')$$

Under null hypothesis,

$$(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{R}')$$

Therefore,

$$(\mathbf{R}\hat{\beta} - \mathbf{r})'[\sigma^{\mathbf{2}}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{R}']^{-\mathbf{1}}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(q)$$

Here we don't know σ^2. However, it can be shown that

$$\frac{\hat{\mathbf{u}}'\mathbf{u}}{\sigma^{\mathbf{2}}} \sim \chi^2(n-k)$$

Then we have

$$\frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{R}']^{-\mathbf{1}}(\mathbf{R}\hat{\beta} - \mathbf{r})/\mathbf{q}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}/(\mathbf{n}-\mathbf{k})} \sim F(q, n-k)$$

This is an example of the F test for any linear hypotheses testing. t test will be a special case of this.

Test of Structural Change (Chow test)

Chow test is a test between two groups of observations. For example, we have data on consumption function of prewar period and postward period. It is natural to think of a test on whether consumption function parameters differ between prewar and postwar periods. The null hypothesis is that there is no difference.

Let $\mathbf{y_i}$, $\mathbf{X_i}$ $(i = 1, 2)$ indicate the appropriate partitioning of the data. The unrestricted model may be written

$$\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix} = \begin{bmatrix}\mathbf{X_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{X_2}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \mathbf{u} \quad \mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$$

The null hypothesis of no structural break is

$$\mathrm{H}_0: \quad \beta_{\mathbf{1}} = \beta_{\mathbf{2}}.$$

Running OLS on two equations separately, we have

$$\begin{bmatrix} \hat{\beta}_{\mathbf{1}} \\ \hat{\beta}_{\mathbf{2}} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1'X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2'X_2} \end{bmatrix}^{-\mathbf{1}} \begin{bmatrix} \mathbf{X_1'y_1} \\ \mathbf{X_2'y_2} \end{bmatrix} = \begin{bmatrix} (\mathbf{X_1'X_1})^{-\mathbf{1}}\mathbf{X_1'y_1} \\ (\mathbf{X_2'X_2})^{-\mathbf{1}}\mathbf{X_2'y_2} \end{bmatrix}$$

Restricted model under null hypothesis

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} \beta + \mathbf{u}$$

The test of the null hypothesis is given by

$$\mathbf{F} = \frac{(\hat{\mathbf{u}}_*'\hat{\mathbf{u}}_* - \hat{\mathbf{u}}'\hat{\mathbf{u}})/\mathbf{k}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}/(\mathbf{n}-\mathbf{2k})} \sim F(k, n-2k)$$

where $\hat{u}_*$ is the residual of the restricted model, $\hat{u}$ is the stacked residual of two unrestricted models.

For example, for prewar period (suppose there is n observations):

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

For postwar period (suppose there is m observations):

$$y_1 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2$$

The null hypothesis of no structural break is

$$\mathrm{H}_0: \quad \beta_{\mathbf{0}} = \gamma_{\mathbf{0}}, \beta_{\mathbf{1}} = \gamma_{\mathbf{1}}, \beta_{\mathbf{2}} = \gamma_{\mathbf{2}}.$$

Running OLS on two equations separately, we have two sums of squares of error (SSR_1 and SSR_2). The unrestricted error sum of squares is

$$SSR_U = SSR_1 + SSR_2$$

Then run the regression on the stacked sample of $n+m$ observations and get SSR_R.

Then

$$\frac{(SSR_R - SSR_U)/k}{SSR_U/(n+m-2k)} \sim F_{k,n+m-2k}.$$

k is number of restrictions; in this example, it is 3.

The other way to do the same test is easier: simply include interaction terms between group dummy and dependent variables in the regression.

For our earlier example, we would estimate the model:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x2 + \eta_0 d + \eta_1 x_1 d + \eta_2 x_2 d$$

Here d is a dummy variable for postward period. The Chow test is simply a test on the hypothesis that all the coefficients involving d are zero. A regression of the above model with $n+m$ observations and a F test are easy to do.

$$\mathrm{H}_0: \quad \boldsymbol{\eta_0 = \eta_1 = \eta_2 = 0}.$$

An example of Chow test in R

Here we use an example from the "strucchange" library.

```
## Example 7.4 from Greene (1993), "Econometric Analysis"
## Chow test on Longley data
data("longley")
library(strucchange)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## use structural change test from the library.
sctest(Employed ~ Year + GNP.deflator + GNP + Armed.Forces, data = longley,
  type = "Chow", point = 7)
```

```
##
##  Chow test
##
## data:  Employed ~ Year + GNP.deflator + GNP + Armed.Forces
## F = 3.9268, p-value = 0.06307
```

```
## which is equivalent to segmenting the regression via
fac <- factor(c(rep(1, 7), rep(2, 9)))
## here fac is the factor for segmenting it at point 7.
fm0 <- lm(Employed ~ Year + GNP.deflator + GNP + Armed.Forces, data = longley)
fm1 <- lm(Employed ~ fac/(Year + GNP.deflator + GNP + Armed.Forces), data = longley)
## Here anova is returning the F test (Chow test). Equivalent to sctest results.
anova(fm0, fm1)
```

```
## Analysis of Variance Table
##
## Model 1: Employed ~ Year + GNP.deflator + GNP + Armed.Forces
## Model 2: Employed ~ fac/(Year + GNP.deflator + GNP + Armed.Forces)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     11 4.8987
## 2      6 1.1466  5    3.7521 3.9268 0.06307 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## estimates from Table 7.5 in Greene (1993)
summary(fm0)
```

```
##
## Call:
## lm(formula = Employed ~ Year + GNP.deflator + GNP + Armed.Forces,
##     data = longley)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9058 -0.3427 -0.1076  0.2168  1.4377
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.169e+03  8.359e+02   1.399  0.18949
## Year         -5.765e-01  4.335e-01  -1.330  0.21049
## GNP.deflator -1.977e-02  1.389e-01  -0.142  0.88940
## GNP           6.439e-02  1.995e-02   3.227  0.00805 **
## Armed.Forces -1.015e-04  3.086e-03  -0.033  0.97436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6673 on 11 degrees of freedom
## Multiple R-squared:  0.9735, Adjusted R-squared:  0.9639
## F-statistic: 101.1 on 4 and 11 DF,  p-value: 1.346e-08
```

```
summary(fm1)
```

```
##
## Call:
## lm(formula = Employed ~ fac/(Year + GNP.deflator + GNP + Armed.Forces),
##     data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47717 -0.18950  0.02089  0.14836  0.56493
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.678e+03  9.390e+02   1.787  0.12413
## fac2                2.098e+03  1.786e+03   1.174  0.28473
## fac1:Year          -8.352e-01  4.847e-01  -1.723  0.13563
## fac2:Year          -1.914e+00  7.913e-01  -2.419  0.05194 .
## fac1:GNP.deflator -1.633e-01  1.762e-01  -0.927  0.38974
## fac2:GNP.deflator -4.247e-02  2.238e-01  -0.190  0.85576
## fac1:GNP           9.481e-02  3.815e-02   2.485  0.04747 *
## fac2:GNP           1.123e-01  2.269e-02   4.951  0.00258 **
## fac1:Armed.Forces -2.467e-03  6.965e-03  -0.354  0.73532
## fac2:Armed.Forces -2.579e-02  1.259e-02  -2.049  0.08635 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4372 on 6 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9845
## F-statistic: 106.9 on 9 and 6 DF,  p-value: 6.28e-06
```

Heteroskedasticity and Autocorrelation Consistent Standard Errors

Variance estimator under homoscedasticity

In OLS regression, the variance-covariance matrix of $\hat{\beta}$ is

$$\text{var}(\hat{\beta}) = \text{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \text{E}[(\mathbf{X'X})^{-1}\mathbf{X'uu'X}(\mathbf{X'X})^{-1}] = \sigma^2(\mathbf{X'X})^{-1}$$

if

$$\text{var}(\mathbf{u}) = \sigma^2\mathbf{I}$$

.

That is, the error term has mean zero and constant variance (homoscedasticity). The pairwise correlation between error terms is always zero (no serial correlation).

The estimation of σ^2:

$$s^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}$$

is an unbiased estimator of σ^2 (proof omitted).

The standard estimate of the variance-covariance matrix of the OLS parameter estimates under the assumption of IID errors is

$$\hat{\text{Var}}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$$

White's estimator

We made strong assumption that the error terms of the regression model are IID when we estimate the variance-covariance matrix of OLS estimator. Under this assumption, the usual estimator of variance-covariance matrix of $\hat{\beta}$ is consistent. Now let's relax this assumption to only independent but not identically distributed. Again, the linear regression models is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \text{E}(\mathbf{u}) = \mathbf{0}, \quad \text{E}(\mathbf{u}\mathbf{u}') = \mathbf{\Omega},$$

where $\mathbf{\Omega}$ is the error variance-covariance matrix with diagonal elements being σ_t^2 for t^{th} element, off-diagonal elements being zero. In other words, the error terms are heteroscedastic.

The variance-covariance matrix of the OLS estimator $\hat{\beta}$ is equal to

$$\text{Var}(\hat{\beta}) = \text{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = [(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'(\text{E}(\mathbf{u}\mathbf{u}'))\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}] = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$$

If we know σ_t^2, then we would be able to estimate this "sandwich covariance matrix". But we don't.

$$\text{Var}(\hat{\beta}) = \frac{\mathbf{1}}{\mathbf{n}}[\frac{\mathbf{1}}{\mathbf{n}}(\mathbf{X}'\mathbf{X})]^{-\mathbf{1}}[\frac{\mathbf{1}}{\mathbf{n}}\mathbf{X}'\mathbf{\Omega}\mathbf{X}][\frac{\mathbf{1}}{\mathbf{n}}(\mathbf{X}'\mathbf{X})]^{-\mathbf{1}}$$

Let y_t denote the tth observation on the dependent variable, and $x_t' = [1 x_{2t} \cdots x_{kt}]$ denote the tth row of the $\mathbf{X}$ matrix. Then

$$\mathbf{X}'\mathbf{\Omega}\mathbf{X} = \sum_{\mathbf{t=1}}^{\mathbf{n}} \sigma_{\mathbf{t}}^{\mathbf{2}}\mathbf{x_t}\mathbf{x_t'}$$

The White estimator replaces the unknown σ_t^2 by $\hat{u}_t^2$, the estimated OLS residuals. This provides a consistent estimator of the variance matrix for the OLS coefficient

vector and is particularly useful since it does not require any specific assumptions about the form of the heteroscedasticity.

Therefore,

$$\hat{\mathrm{Var}}(\hat{\beta}) = \mathbf{\frac{1}{n}}[\mathbf{\frac{1}{n}}(\mathbf{X'X})]^{-\mathbf{1}}[\mathbf{\frac{1}{n}}\mathbf{X'\hat{\Omega}X}][\mathbf{\frac{1}{n}}(\mathbf{X'X})]^{-\mathbf{1}}\mathbf{\hat{\Omega}} = \mathrm{diag}(\hat{u}_1^2, \hat{u}_2^2, \cdots, \hat{u}_n^2)$$

Newey-West estimator

White's estimator deals with the situation that we have heteroskedasticity (a diagonal Σ) of unknown form. When we have serial correlation of unknown form (a non-diagonal Σ), we can estimate the variance-covariance matrix by a heteroskedasticity and autocorrelation consistent, or HAC, estimator. Newey-West estimator is the most popular HAC estimator.

Given a time series data set, suppose we are interested in estimating the mean vector (suppose we have more than one variable) and its variance. We know that given IID data, we can apply central limit theorem: sample mean is a consistent estimator of the population mean and it's variance can be calculated since asymptotically the sample mean conforms to a normal distribution and the variance can be estimated, relatively easily. However, in the case of time series data, autocorrelation usually exists. We may be concerned the CLT may not work in this case.

Fortunately, as proved in Hamilton (1994), if $\mathbf{y_t}$ is a covariance-stationary (meaning that the covariance is not a function of time) vector process, then the sample mean satisfies:

$$\mathbf{\bar{y}_t} \to \mu,$$

$$\mathbf{S} = \lim_{T\to\infty} \mathbf{T} \cdot \mathbf{E}[(\mathbf{\bar{y}_T} - \mu)(\mathbf{\bar{y}_T} - \mu)'] = \sum_{\mathbf{v}=-\infty}^{\infty} \mathbf{\Gamma_v}.$$

where $\mathbf{\Gamma_v}$ is the variance-covariance matrix for $\mathbf{y_t}$ and $\mathbf{y_{t-v}}$.

The first one says for a covariance-stationary vector process, the law of large numbers still holds. The second one is used to calculate the standard error.

If the data were generated by a vector MA(q) process, then

$$\mathbf{S} = \sum_{v=-q}^{q} \Gamma_v.$$

A natural estimate is

$$\mathbf{\hat{S}} = \mathbf{\hat{\Gamma}_0} + \sum_{\mathbf{v}=\mathbf{1}}^{\mathbf{q}} (\mathbf{\hat{\Gamma}_v} + \mathbf{\hat{\Gamma}'_v}),$$

where

$$\hat{\mathbf{\Gamma}}_{\mathbf{v}} = (\mathbf{1}/\mathbf{T}) \sum_{\mathbf{t=v+1}}^{\mathbf{T}} (\mathbf{y_t} - \bar{\mathbf{y}})(\mathbf{y_{t-v}} - \bar{\mathbf{y}}).$$

This gives a consistent estimate of $\mathbf{S}$; however, it sometimes is not positive semidefinite.

Newey-West (1987) suggested putting in a weight:

$$\hat{\mathbf{S}} = \hat{\mathbf{\Gamma}}_{\mathbf{0}} + \sum_{v=1}^{q} (1 - \frac{v}{q+1})(\hat{\mathbf{\Gamma}}_{\mathbf{v}} + \hat{\mathbf{\Gamma}}'_{\mathbf{v}}),$$

where q is from the MA(q) process.

Consider a linear regression model:

$$y_t = \mathbf{x}'_{\mathbf{t}}\beta + u_t$$

Suppose we have the OLS estimator $\mathbf{b_T}$, then

$$\sqrt{T}(\mathbf{b_t} - \beta) = [(1/T)\sum_{t=1}^{T} \mathbf{x_t}\mathbf{x}'_{\mathbf{t}}]^{-1}[(\sqrt{T})\sum_{t=1}^{T} \mathbf{x_t}\mathbf{u}'_{\mathbf{t}}]$$

The first term converges in probability to some constant. The second term is the sample mean of the vector $\mathbf{x_t u_t}$.

Under general conditions,

$$\sqrt{T}(\mathbf{b_t} - \beta) \rightarrow^{L} N(0, Q^{-1}SQ^{-1})$$

where S can be estimated by

$$\hat{\mathbf{S}}_{\mathbf{T}} = \hat{\mathbf{\Gamma}}_{\mathbf{0T}} + \sum_{v=1}^{q} (1 - \frac{v}{q+1})(\hat{\mathbf{\Gamma}}_{\mathbf{v,T}} + \hat{\mathbf{\Gamma}}'_{\mathbf{v,T}}),$$

where

$$\hat{\Gamma}'_{v,T} = (1/T) \sum_{t=v+1}^{T} (x_t \hat{u}_{t,T} \hat{u}_{t-v,T} x'_{t-v}),$$

where $\hat{u}_{t,T}$ is the OLS residual for data t in a sample of size T.

Overall, the variance of $\mathbf{b_T}$ is approximated by

$$\hat{\Sigma}_{NW} = [\sum_{t=1}^{T} x_t x'_t][\sum_{t=1}^{T} \hat{u}_t^2 x_t x'_t + \sum_{v=1}^{q} (1 - \frac{v}{q+1}) \sum_{t=v+1}^{T} (x_t \hat{u}_{t,T} \hat{u}_{t-v,T} x'_{t-v} + x_{t-v} \hat{u}_{t-v,T} \hat{u}_{t,T} x'_t)][\sum_{t=1}^{T} x_t x'_t]^{-1}$$

This estimation obviously depends on the selection of q, the lag length beyond which we are willing to assume that the autocorrelation of $x_t u_t$ and $x_{t-v} u_{t-v}$ is essentially zero. The rule of thumb for the selection of q is $0.75 \cdot T^{\frac{1}{3}}$. Newey-West (1994) has suggested a way to automatically select the bandwidth q. Here we omitted the discussion. Both Stata and R now also implement Newey-West (1994) estimator, with no need to specify q.

Chapter 4

GLS

Introduction

Since heteroscedasticity and serial correlation affect both linear and nonlinear regression models in the same way, there is no harm in limiting our attention to the simpler linear case.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{u}) = \mathbf{0}, \quad \mathrm{E}(\mathbf{u}\mathbf{u}') = \mathbf{\Omega},$$

where $\mathbf{\Omega}$, the variance-covariance matrix of the error term, is a positive definite $n \times n$ matrix. If $\mathbf{\Omega}$ is equal to $\sigma^2\mathbf{I}$, then it is just the linear regression model without heteroscedasticity and serial correlation. If $\mathbf{\Omega}$ is diagonal with nonconstant diagonal elements, then the error terms are still uncorrelated, but there is heteroscedasticity. If $\mathbf{\Omega}$ is not diagonal, then u_i and u_j are correlated. In next section, we obtain an efficient estimator for the vector β by transforming the regression so that it satisfies the conditions of Gauss-Markov Theorem. This efficient estimator is called the Generalized Least Squares, or GLS, estimator.

The GLS Estimator

Since $\mathbf{\Omega}$, the variance-covariance matrix of the error term, is a positive definite $n \times n$ matrix, there always exist full-rank $n \times n$ matrices (usually triangular) $\mathbf{\Psi}$ such that

$$\mathbf{\Omega^{-1}} = \mathbf{\Psi}\mathbf{\Psi}'.$$

Premultiplying (4) by $\mathbf{\Psi'}$ gives

$$\mathbf{\Psi}'\mathbf{y} = \mathbf{\Psi}'\mathbf{X}\beta + \mathbf{\Psi}'\mathbf{u}.$$

The OLS estimator from regression (4) is

$$\hat{\beta}_{\mathbf{GLS}} = (\mathbf{X'\Psi\Psi'X})^{-\mathbf{1}}\mathbf{X'\Psi\Psi'y} = (\mathbf{X'\Omega^{-1}X})^{-\mathbf{1}}\mathbf{X'\Omega^{-1}y}$$

The transformed error term has an identity variance-covariance matrix:

$$\mathrm{E}(\mathbf{\Psi' uu'\Psi}) = \mathbf{\Psi'}\mathrm{E}(\mathbf{uu'})\mathbf{\Psi} = \mathbf{\Psi'\Omega\Psi} = \mathbf{I}$$

Since the transformed model satisfies OLS assumptions, we have

$$\mathrm{Var}(\hat{\beta}_{\mathbf{GLS}}) = (\mathbf{X'\Psi\Psi'X})^{-\mathbf{1}} = (\mathbf{X'\Omega X})^{-\mathbf{1}}$$

Weighted Least Squares

It is easy to obtain GLS estimates when the error terms are heteroscedastic but uncorrelated, which means $\mathbf{\Omega}$ is diagonal. Let ω_t^2 denote the tth diagonal element of $\mathbf{\Omega}$. Then $\mathbf{\Psi}$ can be chosen as the diagonal matrix with tth diagonal element ω_t^{-1}. For a typical observation, regression can be written as

$$\omega_t^{-1} y_t = \omega_t^{-1}\mathbf{X_t}\beta + \omega_t^{-1} u_t.$$

This is called weighted least squares, or WLS. The weight given to each observation is ω_t^{-1}. Observations of which variance of the error term is large are given low weights, and observations for which it is small are given high weights.

Feasible Generalized Least Squares

In many cases it is reasonable to suppose that $\mathbf{\Omega}$ depends in a known way on a vector of unknown parameters γ. If so, it may be possible to estimate γ consistently, so as to obtain $\mathbf{\Omega}(\hat{\gamma})$. This type of procedure is called feasible generalized least squares, or feasible GLS. But we'll have to specify the error term as some function of some known variables.

Chapter 5

Endogeneity

When we discuss the linear model we assume that the regressors are exogenous, meaning that they are independent of or uncorrelated with the error term. However, there could be reasons to believe that some regressors are correlated with the error term. In that case we call those regressors endogenous.

Under the classical assumptions OLS estimators are unbiased and consistent. One key assumption is that the regressors have to be uncorrelated with the error term. If this condition does not hold, OLS estimators are biased and inconsistent.

When one independent variable does not satisfy this condition, we say this variable is endogenous. When one variable is endogenous, the estimates of other coefficients will also be biased (or inconsistent).

The most popular cure for endogeneity is to use instrumental variables.

Instrumental Variables (IV)

The basic idea of IV is to use an exogenous variable (or exogenous variables) which is correlated with the endogenous independent variable to as an "instrument" for the endogenous variable.

Suppose the linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^{\mathbf{2}}\mathbf{I},$$

at least one of the explanatory variables in the $n \times k$ matrix $\mathbf{X}$ is assumed not to be predetermined with respect to the error terms, or say, endogenous.

Suppose we have a set of variables $\mathbf{Z}$, an $n \times l$ matrix of instruments , which satisfies the moment condition

$$\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}.$$

That is, $\mathbf{Z}$ is uncorrelated with the error term.

Here is what we do for two-stage least squares (2sls):

Stage 1: Regress each of the variables in the $\mathbf{X}$ matrix on $\mathbf{Z}$ to obtain a matrix of fitted values $\hat{\mathbf{X}}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P_Z X}$$

This is essentially to get the part of $\mathbf{Z}$ that is correlated with $\mathbf{X}$. Or to say, to project $\mathbf{X}$ on to $\mathbf{Z}$.

Stage 2: Regress $\mathbf{y}$ on $\hat{\mathbf{X}}$ to obtain the estimated β

$$\hat{\beta}_{\mathbf{2sls}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}'\mathbf{y}) = (\mathbf{X}'\mathbf{P_Z X})^{-1}(\mathbf{X}'\mathbf{P_Z y}) = \hat{\beta}_{\mathbf{IV}}$$

Standard errors:

$$Var[\hat{\beta}_{\mathbf{2sls}}] = \hat{\sigma}^{\mathbf{2}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-\mathbf{1}} = \hat{\sigma}^{\mathbf{2}}(\mathbf{X}'\mathbf{P_Z X})^{-\mathbf{1}}$$

where $\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/N$.

Note that $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{2sls}}$.

If we do it manually by two steps, the second step will report a wrong standard error. That is because the second stage regression will report standard errors based on $\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{X}}\hat{\beta}_{\mathbf{2sls}}$. Therefore, it's always recommanded to ask the statistical program to do a 2sls for you, since presumably that will give you correct standard errors.

Geometry Illustration:

Suppose the simplest case: $y = \beta_0 + \beta_x + u$, where x can be decomposed into x_1, which is exogenous and x_2, which is endogenous. Therefore x_2 is parallel to u and x_1 perpendicular to u. Suppose z is a vector that is perpendicular to x_2 or u, but not perpendicular to x_1. Then z can be an instrument for x. The way instrumental variable works: Regress x on z, suppose $\hat{x}_1$ is the projection. Regress y on z, $\hat{y}_2$ be the projection. Then the result of $\beta_2 = \hat{y}_2/\hat{x}_1$ is the same as $\beta_1 = \hat{y}_1/x_1$ (since the two triangles are similar). Here $\hat{y}_1$ is the projection of y on x_1, which is hypothetical since we have no way to decompose x into x_1 and x_2; otherwise, we would not need instrumental variables.

```
## DGP: data$y <- data$x + data$z + data$u
library(MASS)
set.seed(66)
nobs=10000
nDim = 3
sdxx = 1
```

```
sdww=1
sdzz=1

## here we have three variables x,z,w.
## z is the omitted variable,x and z are correlated,  w is the instrument, which is correlated wi
crxz=.6
crzw=0
crxw=.8

covarMat = matrix( c(sdxx^2, crxz^2, crxw^2, crxz^2, sdzz^2, crzw^2,  crxw^2, crzw^2, sdww^2 ) ,
covarMat
```

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.36 0.64
## [2,] 0.36 1.00 0.00
## [3,] 0.64 0.00 1.00
```

```
data  = data.frame(mvrnorm(n=nobs, mu=rep(0,nDim), Sigma=covarMat ))
names(data) <- c('x','z','w')
data$u <- rnorm(nobs,0,1)
# dgp
data$y <- data$x +  data$z + data$u

lm <- lm(y~x, data=data)
lm.full <- lm(y~ x + z, data=data)
tsls.model <- tsls(y ~ x , ~ w , data=data)
# lm is biased
summary(lm)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3948 -0.9103 -0.0037  0.9175  5.4364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003581   0.013723  -0.261    0.794
## x            1.358930   0.013969  97.281   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.372 on 9998 degrees of freedom
## Multiple R-squared:  0.4863, Adjusted R-squared:  0.4862
## F-statistic:  9464 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```
# lm.full is good
summary(lm.full)
```

```
##
## Call:
## lm(formula = y ~ x + z, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7295 -0.6642 -0.0178  0.6750  3.5957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.001834   0.009973   0.184    0.854
## x           0.992439   0.010868  91.317   <2e-16 ***
## z           1.013403   0.010723  94.504   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.997 on 9997 degrees of freedom
## Multiple R-squared:  0.7287, Adjusted R-squared:  0.7286
## F-statistic: 1.342e+04 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```
# tsls is good.
summary(tsls.model)
```

```
##
## Call:
## tsls(g = y ~ x, x = ~w, data = data)
##
##
## Method:  Two Stage Least Squares(Meat type = Classical)
##
## Coefficients:
##              Estimate    Std. Error  t value     Pr(>|t|)
## (Intercept)   0.0060083   0.0142452   0.4217749   0.6731894
## x             0.9724165   0.0231256  42.0493890   0.0000000
##
## J-Test: degrees of freedom is 0
##                 J-test               P-value
```

```
## Test E(g)=0:    1.0653868599088e-26  *******
##
##
##  No first stage F-statistics (just identified model)
```

Control Function Approach

A second way to do an IV regression is also two-step approach: Regress $\mathbf{X}$ (endogenous) on $\mathbf{Z}$, get the residual: $\hat{\mathbf{v}} = \mathbf{X} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$, then regress y on $\mathbf{X}$ and $\hat{v}$ to get $\hat{\beta}_{IV}$.

The difference between this approach and the 2sls approach is that in 2sls we regress y on $\hat{\mathbf{X}}$; in the control function approach, we regress y on $\mathbf{X}$ and $\hat{v}$. They should both give you the same coefficient estimates. There are advantages of using the control function approach.

An example to show how to do 2sls manually, or using control function approach.

```
## DGP: data$y <- data$x + data$z + data$u

lm1 <- lm(y~x, data=data)
lm2 <- lm(x~w, data=data)
tsls.manual <- lm(data$y ~ lm2$fitted.values)
summary(tsls.manual)
```

```
##
## Call:
## lm(formula = data$y ~ lm2$fitted.values)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2338 -1.2292 -0.0247  1.2216  6.4869
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.006008   0.018194    0.33    0.741
## lm2$fitted.values 0.972417   0.029536   32.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.818 on 9998 degrees of freedom
## Multiple R-squared:  0.09781,    Adjusted R-squared:  0.09772
## F-statistic:  1084 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
# control function approach
tsls.control <- lm(data$y ~ data$x + lm2$residual)
summary(tsls.control)
```

```
##
## Call:
## lm(formula = data$y ~ data$x + lm2$residual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6276 -0.9006  0.0160  0.8860  5.0761
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.006008   0.013386   0.449    0.654
## data$x       0.972417   0.021730  44.750   <2e-16 ***
## lm2$residual 0.636578   0.027887  22.827   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.337 on 9997 degrees of freedom
## Multiple R-squared:  0.5117, Adjusted R-squared:  0.5116
## F-statistic:  5238 on 2 and 9997 DF,  p-value: < 2.2e-16
```

This approach can be used to do a simple endogeneity test. First, you can do a simple test of endogeneity of $\mathbf{X}$. For example:

```
reg x_endog x* z*
predict v_x, resid
reg y x_endog x* v_x
test v_x
```

Obviously, this test is based on $\mathbf{Z}$ being exogenous.

Secondly, it works for some non-linear models, such as logit, probit, poisson, or any other glm models. That is, if you have an endogenous variable in a glm model, you can regress that endogenous variable on instruments, get the residual, then run the glm model with the original regressors, plus the residual from the first stage. For example, the "eteffects" procedure in Stata (version 14) uses control function approach to get endogenous treatment effects for different types of outcomes.

Durbin-Wu-Hausman Test

Idea

In econometric modeling, there are often questions on endogeneity. Do we know how to test whether an independent variable is endogenous statistically? The answer is: sort of, but not really. We cannot do endogeneity test without a valid instrument. Therefore, we have to have strong argument for a valid instrument first before we can do endogeneity test.

With endogenous variables on the right-hand side of the equation, we need to use instrumental variable (IV) regression for consistent estimation. However, with IV regression, we lose efficiency: the asymptotic variance of the IV estimator is larger, and can be much larger than the OLS estimator. Therefore, we gain consistency, but lose efficiency, by using IV estimator when there is an endogeneity problem.

Now we have a familiar scenario (if you are familiar with Hausman test for fixed effect and random effect estimator for panel data): Suppose we have the null hypothesis as the regressor being exogenous. We have an efficient estimator under null hypothesis yet inconsistent under alternative hypothesis (OLS estimator). We also have a consistent estimator under both null and alternative (IV estimator).

Similar to panel data setting, we have the Hausman test statistic as:

$$H = (\hat{\beta}_c - \hat{\beta}_e)' D^{-} (\hat{\beta}_c - \hat{\beta}_e)$$

where $D = \text{Var}[\hat{\beta}_c] - \text{Var}[\hat{\beta}_e]$, $^{-}$ is the generalized inverse, $\hat{\beta}_c$ is the consistent estimator (in this case the IV estimator) and $\hat{\beta}_e$ is the efficient estimator (in this case OLS estimator).

H conforms to χ^2_k asymptotically, where k is the number of endogenous variables.

This test is to compare the IV estimator and the OLS estimator: if it's close, then OLS estimator is fine (fail to reject null that OLS is consistent, or say the variable is exogenous). If it's large, then IV estimator is needed, although we lose some efficiency. This test is based on the assumption that the instruments are exogenous. If that is in question, then it's pointless to do the test, since the IV estimator cannot guarantee consistency either.

Implementation in Stata

In Stata, there are different ways to do it:

Do a regular Hausman test:

```
ivreg y x1 (x2=x3 x4)
estimates store iv
reg y x1 x2
hausman iv ., constant sigmamore
```

Or, simply use "ivendog" in Stata.

Identification

Identification in a regression equation means that all parameters can be uniquely estimated. A necessary condition of that is to have at least as many instruments as the number of endogenous variables. That is, $l \geq k$ in our example above. If $l = k$ we have exact-identification. If $l > k$, we have over-identification.

Over-identification generates more efficient estimates, given the assumption of instruments being exogenous. The other advantage of over-identification is that over-identification tests can be done to test the adequacy of instruments.

Under the null hypothesis that all the instruments are uncorrelated with the error term, an LM statistic $N \times R^2$ conforms to $\chi^2(r)$ distribution, $r = l - k$, the number of excess instruments, or say, the number of excluded restrictions. If we reject the null, then we should be concerned about the exogeneity of the whole set of the instruments. This test is called Sargan's test in IV context, and (Hansen's) J test in GMM context.

What the J test or Sargan's test does is to test the whole set of instruments being exogenous or not. There is another test for testing exogeneity for a subset of instruments. It's call a C test or a difference-in-Sargan test. The idea is to calculate the difference between two Sargan's statistics (or Hansen's J in GMM setting); one is with the whole set of instruments, the other one without the suspected instruments. The null is that the suspect instruments are exogenous; or orthogonal to the error term. Obviously to conduct the C test, we'll have to have at least one extra instrument more than the number of endogenous variables.

To understand it better, we look at how to implement Sargan's test manually: For the 2SLS estimator, the test statistic is Sargan's statistic, typically calculated as $N \times R^2$ from a regression of the IV residuals on the full set of instruments.

```
. ivregress 2sls rent pcturban (hsngval = faminc i.region)
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(3) =  11.2877  (p = 0.0103)
  Basmann chi2(3)        =  12.8294  (p = 0.0050)
```

```
. predict res, residual

. reg res pcturban faminc i.region

. disp e(N)*e(r2)
11.287665
```

Implementation in Stata

In Stata, there are different ways to do over-identification test, ivreg2 reports a comprehensive set of tests; overid command does the over-identification test after the ivreg command.

ivreg2 with gmm option returns J test; it reports Sargan's test without this option.

ivreg2 also reports C test statistic, with ortho(). If the C test rejects the null, and J test without the suspect instruments fail to reject null, then the suspect instruments are indeed the ones are not exogenous.

Weak Instruments

Problem with the cure

An instrument needs to satisfy to criteria: orthogonality and relevance. We need instruments to be orthogonal to the error term. We can verify the orthogonality condition by Sargan's test if there are extra instruments.

It turns out instrument relevance is important too: if instruments are weak, then the regular large sample properties of IV or GMM estimators do not hold any more. The estimators are inconsistent or biased.

To see the problem, suppose

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma_{\mathbf{u}}^{\mathbf{2}}\mathbf{I},$$

$$\mathbf{X} = \mathbf{Z}\mathbf{\Pi} + \mathbf{v}, \quad \mathrm{E}(\mathbf{v}\mathbf{v}') = \sigma_{\mathbf{v}}^{\mathbf{2}}\mathbf{I},$$

and

$$\mathrm{E}(\mathbf{Z}\mathbf{u}) = \mathbf{0}.$$

We can see here $\mathbf{Z}$ is exogenous. However, the model does not say anything about relevance. To illustrate the problem caused by weak instrument, suppose we have only one endogenous variable and one instrument.

$$\hat{\beta}_{2sls} = \frac{\mathbf{Z}'\mathbf{y}}{\mathbf{Z}'\mathbf{X}} = \frac{\mathbf{Z}'(\mathbf{X}\beta + \mathbf{u})}{\mathbf{Z}'\mathbf{X}} = \beta + \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{X}}.$$

If $\mathbf{Z}$ is irrelevant, or, $\Pi = 0$, then

$$\hat{\beta}_{2sls} - \beta = \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{v}} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{N} Z_i u_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{N} Z_i v_i} \xrightarrow{d} \frac{z_u}{z_v},$$

where

$$\begin{bmatrix} z_u \\ z_v \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}).$$

Therefore, if $\mathbf{Z}$ is irrelavent, β_{2sls} is inconsistent. Also, the distribution of the bias is Cauchy-like (the ratio of correlated normals).

This is a case where the cure might be worse than the disease itself: the bias can be big comparing to the bias an OLS estimate suffers.

Weak Instrument Tests

There are a variety of weak-instruments tests proposed. Most of them are based on so-called weak-instruments asymptotics and a new parameter called "concentration parameter" $\mu^2 = \Pi' Z' Z \Pi / \sigma_v^2$. Sample size only enters the distribution through μ^2.

With weak-instruments asymptotics, IV estimators are no longer consistent, and they are not normal asymptotically. Most test statistics (J test, etc.) do not have normal or χ^2 distributions anymore.

Now I list the following tests in the order of recommended level by James Stock:

1. Moreira (2003) conditional likelihood ratio test (CLR).

Advantages of this test: a. Uniformly most powerful tests among valid tests. b. Implemented in Stata as condivereg.

Disadvantages:

a. Complicated. Only developed so far for one endogenous variable case.

2. Stock-Yogo bias method and size method.

Stock and Yogo (2005) provide critical values for both methods: one is to control the size of bias, the other one is to control the size of a Wald test of $\beta = \beta_0$. Bias method is more frequently used. In the case of multiple endogenous variables, the Craigg-Donald statistics is used to compare with the critical values. It is implemented in Stata as part of the *{ivreg2} command, but it's only available for the situation there are at least two excluded variables (meaning the number of instruments minus the number of endogenous variables).*

2. Anderson-Rubin confidence intervals.

In the model of

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathrm{E}(\mathbf{uu}') = \sigma_{\mathbf{u}}^{\mathbf{2}}\mathbf{I},$$

$$\mathbf{X} = \mathbf{Z\Pi} + \mathbf{v}, \quad \mathrm{E}(\mathbf{vv}') = \sigma_{\mathbf{v}}^{\mathbf{2}}\mathbf{I},$$

The null hypothesis $H_0 : \beta = \beta_0$. Anderson-Rubin statistic is the F statistic in the regression of $y - X\beta_0$ on Z, the F test on Π being zero:

$$AR(\beta_0) = \frac{(y - X\beta_0)' P_Z (y - X\beta_0)/k}{(y - X\beta_0)' M_Z (y - X\beta_0)/(N - k)}$$

The idea of AR confidence interval is to construct an interval for all possible values of β to fail to reject $\Pi = 0$.

3. First-stage F test.

The rule of thumb for first-stage F test is $F > 10$ for a single instrument case, the more instruments, the higher it gets.

4. Kleibergen's LM test.

This test is dominated by the CLR test, thus no longer the optimal test to use.

5. First-stage R^2, or partial R^2, etc., are not recommended.

When is endogeneity a problem and what can we say from an IV regression

Let's consider a model of housing price on no. of rooms and square footage.

$$P = \alpha R + \beta F + \epsilon$$

Should square footage really be there? We are not interested in, given square footage, how much money we need to pay for an additional room. We are generally interested in how much money we need to pay for an additional footage. However, if we do not include footage, then we "sort of" have endogeneity problem due to omitted variable.

Another example would be education's impact on wage. Say IQ is an omitted variable. Do we want to say something about education's impact on wage controlling on IQ, or not?

So it depends on what questions to ask. If I am a builder and ask what an extra bedroom on the already built house would bring to me, I need to control for footage. If I am generally asking what an extra bedroom would cost, then I am implying what that bedroom and associated footage would cost me. In that case, without controlling for footage is not a problem, since the question is: what an additional bedroom (AND the associated average square footage) would cost me.

In the wage and education example, if I am asking: what's the effect of education on wage? I implicitly ask what would be my pay raise if I go to college (given IQ, of course), for example. Then we are concerned that this model of

$$wage = \alpha edu + \epsilon$$

will generate biased estimate of α, because we are not controlling for IQ. But if I am an employer and interested in hiring a person, what would I have to pay extra to hire a collage graduate vs. a high school graduate? Then I don't have to include that IQ variable, because implicitly I am asking: what extra I need to pay for that college graduate who comes with a higher IQ, and everything else that is associated with higher education.

In addition, what an IV regression really gives us is part of the causal inference of X (suspected to be endogenous) on y, namely the part induced by Z. For example, we are interested in the relationship between colleage education and wage. If we use distance to college as an instrument, then our inference is the effect of college education from those who decide to go because of nearbyness of the college.

Chapter 6

General Method of Moments (GMM)

The Method of Moments (MOM)

A population moment γ can be defined as the expectation of some continuous function g of a random variable x:

$$\gamma = \mathrm{E}[g(x)]$$

On the other hand, a sample moment is the sample version of the population moment in a particular sample:

$$\hat{\gamma} = \frac{1}{n}\sum[g(x)]$$

OLS as a moment problem

Consider the simple linear regression

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u} \sim \mathbf{IID}(\mathbf{0}, \sigma^{\mathbf{2}}).$$

If the model is correctly specified, then

$$\mathrm{E}(\mathbf{X}'\mathbf{u}) = \mathbf{0}.$$

The MOM principle suggests that we replace the left-hand side with its sample analog $\frac{1}{n}\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)$.

Since we know that the true β sets the population moment equal to zero in expectation, it seems reasonable to assume that a good choice of $\hat{\beta}$ would be one that sets the sample moment to zero. The MOM procedure suggests an estimate of β that solves

$$\frac{1}{n}\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}.$$

The MOM estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y},$$

which is the same as the OLS estimator.

IV as a moment problem

Consider the simple linear regression

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u} \sim \mathbf{IID}(\mathbf{0}, \sigma^{\mathbf{2}}).$$

If the model is mis-specified, then

$$\mathrm{E}(\mathbf{X}'\mathbf{u}) \neq \mathbf{0}.$$

We have to find an instrumental variable $\mathbf{Z}$ which is

$$\mathrm{E}(\mathbf{Z}'\mathbf{u}) = \mathbf{0}.$$

Or,

$$\mathrm{E}(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)) = \mathbf{0}.$$

The sample analogy of this is

$$\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}.$$

That gives us the IV estimator

$$\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-\mathbf{1}}\mathbf{Z}'\mathbf{y}.$$

The Generalized Method of Moments

The expectation $\mathrm{E}(Y^r)$ for any $r = 1, 2, \ldots$ is called the r^{th} (raw) moment of Y. The expectation $\mathrm{E}[(Y - \mathrm{E}(Y))^r]$ is called the r^{th} centered moment of Y.

The mean is the first raw moment.

The variance is the second centered moment.

The third centered moment measures the skewness of the distribution.

The fourth centered moment measures the kurtosis of the distribution. Interpreted as a measure of "fatness of tails".

The standardized kurtosis is

$$k = \frac{E[(Y - E(Y))^4]}{E[(Y - E(Y))^2]^2}.$$

For a normal distribution, $k = 3$.

For a t distribution with $v \geq 5$ *degrees of freedom,* $k = 3 + 6/(v - 4) > 4$. *i.e., the t distribution has fatter tails than a normal distribution.*

The distribution function of a random variable captures all information about the random variable. It can be shown using all moments also captures all information.

This distinction underlies the relative strengths and weaknesses of ML and GMM.

GMM

The statistical model takes the general form

$$E[m(Y_i; \theta_0)] = 0$$

where

- $Y_1, \cdots, Y_2$ *are random variables from which the sample* $y_1, \cdots, y_n$ *is drawn,*
- $m(Y, \theta)$ *is a function specifying the model,*
- θ_0 *is the "true value" of the parameter.*

$E[m(Y_i; \theta_0)] = 0$ *are called the population moment conditions.*\

Two ideas behind GMM:

1. *Replace the population mean* $E[.]$ *with the sample mean calculated from the observed sample* $y_1, \cdots, y_n$.
2. *Since* $E[m(Y_i; \theta_0)] = 0$, *choose* $\hat{\theta}_{GMM}$ *to make* $\frac{1}{n}\sum_{i=1}^{n} m(y_i; \hat{\theta}_{GMM})$ *as close to zero as possible.*

Define the notation

$$\bar{m}(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(y_i; \theta).$$

$\hat{\theta}_{GMM}$ is chosen to make $\bar{m}(\theta)'\bar{m}(\theta)$ as close to zero as possible.

More generally, $\hat{\theta}_{GMM}$ is chosen to minimize $\bar{m}(\theta)'W\bar{m}(\theta)$ for some weighting matrix W.

An example

Let's see an example with GMM, using the same simulated data as before. We have the same situation as before, X is endogenous. We are doing GMM version of 2sls.

Here I use R's "gmm" library which makes things easy. It expects two arguments: "g" and "x", which correponds to u and W here. The moment condition is $E(Wu) = 0$ in this example. W is the instrument, and u is the residual from regressing y on endogenous X.

```
## DGP: data$y <- data$x + data$z + data$u

set.seed(66)
nobs=10000
nDim = 3
sdxx = 1
sdww=1
sdzz=1

## here we have three variables x,z,w.
## z is the omitted variable,x and z are correlated,  w is the instrument, which is co
crxz=.6
crzw=0
crxw=.8

covarMat = matrix( c(sdxx^2, crxz^2, crxw^2, crxz^2, sdzz^2, crzw^2,  crxw^2, crzw^2,
covarMat
```

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.36 0.64
## [2,] 0.36 1.00 0.00
## [3,] 0.64 0.00 1.00
```

```
data  = data.frame(mvrnorm(n=nobs, mu=rep(0,nDim), Sigma=covarMat ))
names(data) <- c('x','z','w')
data$u <- rnorm(nobs,0,1)
# dgp
data$y <- data$x +  data$z + data$u
```

```
gmm.fit = gmm(data$y~data$x, data$w)
summary(gmm.fit)
```

```
##
## Call:
## gmm(g = data$y ~ data$x, x = data$w)
##
##
## Method:  twoStep
##
## Kernel:  Quadratic Spectral
##
## Coefficients:
##              Estimate    Std. Error  t value     Pr(>|t|)
## (Intercept)   0.0060083   0.0141464   0.4247205   0.6710404
## data$x        0.9724165   0.0234731  41.4268398   0.0000000
##
## J-Test: degrees of freedom is 0
##                 J-test               P-value
## Test E(g)=0:    1.0653868599088e-26  *******
```

```
# It returns the same estimates as the 2sls results.
tsls.model <- tsls(y ~ x , ~ w , data=data)
summary(tsls.model)
```

```
##
## Call:
## tsls(g = y ~ x, x = ~w, data = data)
##
##
## Method:  Two Stage Least Squares(Meat type = Classical)
##
## Coefficients:
##              Estimate    Std. Error  t value     Pr(>|t|)
## (Intercept)   0.0060083   0.0142452   0.4217749   0.6731894
## x             0.9724165   0.0231256  42.0493890   0.0000000
##
## J-Test: degrees of freedom is 0
##                 J-test               P-value
## Test E(g)=0:    1.0653868599088e-26  *******
##
##
##  No first stage F-statistics (just identified model)
```

In OLS case, it would be $E(Xu) = 0$.

```
gmm.ols = gmm(data$y~data$x, data$x)
summary(gmm.ols)
```

```
##
## Call:
## gmm(g = data$y ~ data$x, x = data$x)
##
##
## Method:  twoStep
##
## Kernel:  Quadratic Spectral
##
## Coefficients:
##              Estimate    Std. Error  t value     Pr(>|t|)
## (Intercept)  -0.0035807   0.0136163  -0.2629680   0.7925753
## data$x        1.3589304   0.0137738  98.6604351   0.0000000
##
## J-Test: degrees of freedom is 0
##                 J-test                 P-value
## Test E(g)=0:    5.19457810483928e-25  *******
```

```
# It returns the same estimates as the OLS results.
ols <- lm(y ~ x, data=data)
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3948 -0.9103 -0.0037  0.9175  5.4364
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003581   0.013723  -0.261    0.794
## x            1.358930   0.013969  97.281   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 9998 degrees of freedom
## Multiple R-squared:  0.4863, Adjusted R-squared:  0.4862
## F-statistic:  9464 on 1 and 9998 DF,  p-value: < 2.2e-16
```

A few concepts of conditioning

Independence

If X and Y are independent then

$$f(x, y) = f(x)f(y)$$

and hence

$$f(y|x) = f(y).$$

If X and Y are independent then

$$E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$$

and hence

$$Cov[g(X), h(Y)] = 0.$$

i.e. all functions of X and Y are uncorrelated.

Law of Iterated Expectations

$$E[Y] = E[E(Y|X)].$$

Dependence Concepts

X, Y independent:

$$Cov[g(X), h(Y)] = 0$$

X, Y uncorrelated:

$$Cov[X, Y] = 0$$

$E[Y|X] = 0$:

$$Cov[g(X), Y] = 0$$

Regression

A regression model is a model of $E[Y_i|X_i]$. For example,

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where $E[u_i|X_i] = 0$. \

GMM regression

The regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad E[u_i|X_i] = 0$$

implies the moment condition

$$E[u_i] = 0 \quad \textit{and} \quad E[X_i u_i] = 0$$

That is,

$$E[Y_i - \beta_0 - \beta_1 X_i] = 0$$
$$E[X_i(Y_i - \beta_0 - \beta_1 X_i)] = 0$$

The sample moment conditions are

$$\frac{1}{n}\sum_{i=1} n(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{1}{n}\sum_{i=1} n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

These are just normal equations for OLS.

A characteristic of GMM: the specification of the model generates the estimator. i.e. only $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$ is assumed.

Note there are no assumptions that u_i is homoscedastic, not autocorrelated or normally distributed. These properties affect the statistical properties of the GMM estimator, not its definition.

Chapter 7

Censored, truncated or selected data

Compare three different cases

A sample is truncated if some observations are systematically excluded from the sample. For example, suppose we are interested in relationship between people's income(y) and education(x). If we have observations of both y and x only for people whose income is above $20,000 per year, then we have a truncated sample. A sample is censored if no observations have been systematically excluded but some of information contained in them has been suppressed. In other words, in a truncated sample, you don't have observations which have been truncated; while in a censored sample, you do have observations but you don't have full information. The idea of "censoring" is that some data above or below a threshold are mis-reported at the threshold. If we use the same example of people's income, if we don't observe people's income (y) and education(x) with income lower than $20,000 per year, then we have a truncated sample. If we have observations of x for people whose income above AND below that limit, but no information on exactly how much their income is for those whose income below the limit (only thing we know is that it's lower than $20,000 per year), then we have a censored sample.

The problem of sample selection arises when no observations have been systematically excluded but some information has been suppressed, just like censoring. However, in sample selection problems, we have information on how the "selection" happens. In censoring, usually a constant "threshold" is set. Observations on dependent variable for those above (or below) that criteria are missing. Using the income example again, the threshold is set to be $20,000, we have no information on exactly how much their income is for those whose income below the limit (only thing we know is that it's lower than $20,000 per year). When we have

a third variable that determines the censoring situation, then we call it sample selection. For example, we have observed only union member's information on income, but we observe information on education for everybody, including union members or non-members. In this example, union membership is an additional variable that "selects" the sample which are involved in the estimation of effect of education on income.

An illustration for truncated or censored data

Consider the model

$$y_t^0 = \beta_1 + \beta_2 x_t + u_t,\ u_t \sim NID(0, \sigma^2),$$

where y_t^0 is a latent variable. We actually observe y_t, which differs from y_t^0 , because it's either truncated or censored.

Suppose that censorship or truncation occurs whenever y_t^0 is less than 0. Clearly, the larger the error term u_t, the larger is y_t^0 and thus the greater must be the probability that $y_t^0 \geq 0$. This probability also depends on x_t. So for the sample we observe, u_t does not have conditional mean 0 and is not uncorrelated with x_t. OLS using truncated or censored samples (OLS of y_t on x_t) yields biased and inconsistent estimators. Normally we would like to draw inference on the population which is represented by the full sample. Shown in figure 1, ideally we would have the OLS regression line (if we had the full sample), which is very close to the "true" regression line (which is the mechnism generates the data). We would have the "OLS censored sample" line if we run OLS on the censored sample; we would have "OLS truncated sample" if we run OLS on the truncated sample. It shows that both of them are severely biased from the "true" regression line.

Here is a graph to illustrate from Davidson and MacKinnon.

Truncated Models

For truncated data, a consistent estimator comes from maximum likelihood estimator (MLE). If we assume the error terms in the latent variable model has a known distribution, then MLE estimator can be applied. The most popular choice is Gaussian.

$$\begin{aligned} Pr(y_t^0 \geq 0) &= Pr(\mathbf{X}_t\beta + u_t \geq 0) \\ &= 1 - Pr(u_t/\sigma < -\mathbf{X}_t\beta/\sigma) \\ &= 1 - \Phi(-\mathbf{X}_t\beta/\sigma) = \Phi(\mathbf{X}_t\beta/\sigma) \end{aligned} \tag{7.1}$$

Figure 7.1: Effects of truncation and censoring from Davidson and MacKinnon

The density of y_t is proportional to the density of y_t^0 when $y_t^0 \geq 0$ and y_t is observed. It is 0 elsewhere. The factor of proportionality, which is needed to ensure that the density integrates to unity, is the inverse of the probability that $y_t^0 \geq 0$. The density of y_t can be written as

$$\frac{\sigma^{-1}\phi((y_t - \mathbf{X}_t\beta)/\sigma)}{\Phi(\mathbf{X}_t\beta/\sigma)}$$

This implies the log-likelihood function,

$$\ell(\mathbf{y}, \beta, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \mathbf{X}_t\beta)^2 - \sum_{t=1}^{n}\log\Phi(\mathbf{X}_t\beta/\sigma).$$

This can be estimated by MLE. In Stata, truncreg is the command to do truncated regression model.

Censored Models

The most popular censored model is Tobit model.

$$\begin{aligned} y_t^0 &= \mathbf{X}_t + u_t, \quad u_t \sim NID(0, \sigma^2) \\ y_t &= y_t^0 \text{ if } y_t^0 > 0;\ y_t = 0 \text{ otherwise.} \end{aligned} \tag{7.2}$$

We see that

$$\begin{aligned} Pr(y_t = 0) &= Pr(y_t^0 \leq 0) = Pr(\mathbf{X}_t\beta + u_t \leq 0) \\ &= Pr(u_t/\sigma < -\mathbf{X}_t\beta/\sigma) \\ &= \Phi(-\mathbf{X}_t\beta/\sigma) \end{aligned} \tag{7.3}$$

The contribution to the log-likelihood function made by observations with $y_t = 0$ *is*

$$\ell_t(y_t, \beta, \sigma) = \log \Phi(-\mathbf{X}_t\beta/\sigma).$$

If y_t *is positive, the contribution to the log-likelihood is the logarithm of the density,*

$$\log(\frac{1}{\sigma}\phi((y_t - \mathbf{X}_t\beta)/\sigma)).$$

The log-likelihood function of the tobit model is

$$\sum_{y_t=0} \log \Phi(-\mathbf{X}_t\beta/\sigma) + \sum_{y_t>0} \log(\frac{1}{\sigma}\phi((y_t - \mathbf{X}_t\beta)/\sigma)))$$

This can be estimated by MLE. In Stata, tobit or intreg can be used for censoring models.

Sample Selection

The sample selection models differ from censored model in that it involves a different variable (from y itself) to determine the censorship (selection).

Suppose that y_t^0 *and* z_t^0 *are two latent variables, generated by the bivariate process*

$$\begin{bmatrix} y_t^0 \\ z_t^0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_t\beta \\ \mathbf{W}_t\gamma \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \ \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim NID(\mathbf{0}, \begin{bmatrix} \sigma^2 & \rho \\ \rho & 1 \end{bmatrix})$$

We observe y_t *and* z_t*:*

$y_t = y_t^0$ *if* $z_t^0 > 0$; y_t *unobservable otherwise*; $z_t = 1$ *if* $z_t^0 > 0$; $z_t = 0$ *otherwise.*

There are two types of observations, ones we observe $y_t = y_t^0$ *and* $z_t = 1$*, along with both* $\mathbf{X_t}$ *and* $\mathbf{W_t}$*, and ones we observe only* $z_t = 0$ *and* $\mathbf{W_t}$*.*

Each observation contributes to the likelihood function by

$$I(z_t = 0)Pr(z_t = 0) + I(z_t = 1)Pr(z_t = 1)f(y_t^0|z_t = 1),$$

Under the normality assumption,

$$u_t = \rho v_t + e_t$$

where e_t is independent of $v_t \sim N(0,1)$. A useful fact about the standard normal distribution is that

$$\mathrm{E}(v_t | v_t > -x) = \lambda(x) = \frac{\phi(x)}{\Phi(x)}$$

and the function $\lambda(x)$ is called the inverse Mills ratio.

The log-likelihood function can be shown to be

$$\sum_{z_t=0} \log \Phi(-\mathbf{W}_t\gamma) + \sum_{z_t=1} \log(\frac{1}{\sigma}\phi((y_t - \mathbf{X}_t\beta)/\sigma)) + \sum_{z_t=1} \log \Phi(\frac{\mathbf{W}_t\gamma + \rho(y_t - \mathbf{X}_t\beta)/\sigma}{(1-\rho^2)^{1/2}})$$

So this model can be estimated by MLE.

It is also popular to use Heckman's two-step method.

Heckman's method is based on the fact that the original latent model can be rewritten as

$$y_t = \mathbf{X}_t\beta + \rho v_t + e_t$$

Here the error term u_t is divided into two parts, one perfectly correlated with v_t, and one independent of v_t.

In the first step, an ordinary probit model is used to obtain consistent estimates $\hat{\gamma}$ of the parameters of the selection equation.

In the second step, the unobserved v_t is replaced by the selectivity regressor $\frac{\phi(\mathbf{W}_t\hat{\gamma})}{\Phi(\mathbf{W}_t\hat{\gamma})}$.

Therefore, the regression becomes

$$y_t = \mathbf{X}_t\beta + \rho\frac{\phi(\mathbf{W}_t\hat{\gamma})}{\Phi(\mathbf{W}_t\hat{\gamma})} + e_t$$

Or,

$$y_t = \mathbf{X}_t\beta + \rho\hat{\lambda}_t + e_t$$

One thing to note in Heckman model is that in many situations, $\hat{\lambda}$ is believed to be highly collinear with $\mathbf{X_t}$, if $\mathbf{W_t}$ is the same as $\mathbf{X_t}$. However, the model can still be estimated due to the nonlinearity of the model. Nevertheless, it becomes a regular practice to require $\mathbf{W_t}$ contains at least one extra variable than $\mathbf{X_t}$, which is sometimes called exclusion restriction. In many situations, $\mathbf{W_t}$ contains all $\mathbf{X_t}$ variables, and at least one more variable. The reason to contain all $\mathbf{X_t}$ variables is because the selection is "endogenous" in the sense that y_t is a factor in determining selection; this is modeled by including all the $\mathbf{X_t}$'s.

In Stata, heckman is the command to do sample selection models. It has options to do a Heckman two-step estimation or MLE estimation. Usually MLE is preferred.

Switching Regression (Treatment-Effects Model)

There is another situation that is similar to sample selection model: we observe y for $z = 1$ and $z = 0$. In the example of effect of education on income, we observe both union member's income and non-member's income. In this case, we have switching regression model or treatment-effect model. The treatment effect model estimates the effect of an endogenous binary treatmentz_t (treatment, program participation, etc.) on a continuous, fully-observed variable,y_t, conditional on the independent variables x_t and w_t.

Suppose z_t^0 is a latent variable.

$$\begin{bmatrix} y_t \\ z_t^0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_t\beta \\ \mathbf{W}_t\gamma \end{bmatrix} + \delta \begin{bmatrix} z_t \\ 0 \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \ \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim NID(\mathbf{0}, \begin{bmatrix} \sigma^2 & \rho \\ \rho & 1 \end{bmatrix})$$

We observe y_t and z_t:

$$z_t = 1 \text{ if } z_t^0 > 0; \ z_t = 0 \text{ otherwise.} \tag{7.4}$$

Notice that the only difference between the switching regression and selection model is that we observe y_t when $z_t = 0$ and $z_t = 1$. It's not a selection, but a regime switching.

Both Stata (etregress in version 14) and SAS (qlim) can estimate the switching regression. The default is by MLE.

Chapter 8

Discrete and Limited Dependent Variables

Binary Response Models

Probit and Logit

Let P_t denote the probability that $y_t = 1$ conditional on the information set Ω, which consists of exogenous and predetermined variables. A binary response model serves to model this conditional expectation. Since the values are 0 or 1, it is clear that P_t is also the expectation of y_t conditional on Ω_t:

$$P_t \equiv Pr(y_t = 1|\Omega_t) = E(y_t|\Omega_t).$$

Since $0 \leq P_t \leq 1$, $X_t\hat{\beta}$ needs to be in this interval too, X_t is a set of regressors.

We ensure that $0 \leq P_t \leq 1$ by specifying that

$$P_t \equiv Pr(y_t = 1|\Omega_t) = F(\mathbf{X_t}\beta).$$

$F(x)$ is a transformation function, which has the same characteristics as the CDF of a probability distribution.

Two popular choices of $F(x)$ are Gaussian (probit) and Logistic (logit).

The less familiar logistic function is

$$\Lambda(x) = \frac{e^x}{1 + e^x}$$

The logit model is most easily derived by assuming that

$$\log(\frac{P_t}{1 - P_t}) = \mathbf{X_t}\beta$$

which says the logarithm of the odds (the ratio of the two probabilities) is equal to $\mathbf{X_t}\beta$*. Therefore,*

$$P_t = \frac{\exp(\mathbf{X}_t\beta)}{1+\exp(\mathbf{X}_t\beta)} = \Lambda(\mathbf{X}_t\beta)$$

MLE for binary data

The likelihood for an observation t *is the probability that* $y_t = 1$ *if* $y_t = 1$*, or the probability that* $y_t = 0$ *if* $y_t = 0$*. The logarithm of the appropriate probability is then the contribution to the loglikelihood made by observation* t*. Therefore, if* $\mathbf{y}$ *is an* n*-vector with typical element* y_t*, the loglikelihood function for* $\mathbf{y}$ *can be written as*

$$\ell(\mathbf{y},\beta) = \sum_{t=1}^{n}(y_t \log F(\mathbf{X_t}\beta) + (1-y_t)\log(1-F(\mathbf{X_t}\beta))$$

For the logit and probit models, this function is globally concave with respect to β*. This implies that the first-order conditions, or likelihood equations, uniquely define the MLE estimator* $\hat{\beta}$*. These likelihood equations can be written as*

$$\sum_{t=1}^{n}\frac{(y_t - F(\mathbf{X_t}\beta))F(\mathbf{X_t}\beta)x_{ti}}{F(\mathbf{X_t}\beta)(1-F(\mathbf{X_t}\beta))} = 0,\ i = 1,\ldots,k.$$

Newton's Method can be used to find $\hat{\beta}$*.*

Models for More than Two Discrete Responses

The Ordered Probit

Ordered Probit can be easily derived from a latent variable model.

$$y_t^0 = \mathbf{X}_t\beta + u_t,\ u_t \sim NID(0,1)$$

Suppose we observe y_t *with three values.*

$$y_t = 0 \quad \textit{if } y_t^0 < \gamma_1; y_t = 1 \quad \textit{if } \gamma_1 \le y_t^0 < \gamma_2; y_t = 2 \quad \textit{if } \le y_t^0 \ge \gamma_1.$$

Therefore,

$$Pr(y_t = 0) = Pr(y_t^0 < \gamma_1) = Pr(\mathbf{X}_t\beta < \gamma_1) = Pr(u_t < \gamma_1 - \mathbf{X}_t\beta) = \Phi(\gamma_1 - \mathbf{X}_t\beta)$$

Similarly,

$$Pr(y_t = 2) = \Phi(\mathbf{X}_t\beta - \gamma_2)$$

$$Pr(y_t = 2) = \Phi(\gamma_2 - \mathbf{X}_t\beta) - \Phi(\gamma_1 - \mathbf{X}_t\beta)$$

These probabilities depend solely on the value of the index function and on the two threshold parameters.

The loglikelihood function is

$$\ell(\beta, \gamma_1, \gamma_2) = \sum_{y_t=0} \log(\Phi(\gamma_1 - \mathbf{X}_t\beta)) + \sum_{y_t=2} \log(\Phi(\mathbf{X}_t\beta - \gamma_2)) + \sum_{y_t=1} \log(\Phi(\gamma_2 - \mathbf{X}_t\beta) - \Phi(\gamma_1 - \mathbf{X}_t\beta))$$

The Multinomial Logit

When responses are unordered, the popular choice is multinomial logit.

Suppose there are $J + 1$ responses, for $J \geq 1$.

$$Pr(y_t = l) = \frac{\exp(\mathbf{W}_{tl}\beta^l)}{\sum_{j=0}^{J} \exp(\mathbf{W}_{tl}\beta^l)} \quad \textit{for } l = 0, \ldots, J.$$

Here $\mathbf{W}_{tj}$ is a row vector of length k_j of observations on variables that belong to the information set of interest, and β^j is a k_j-vector of parameters, usually different for each $j = 0, \ldots, J$.

Chapter 9

Count data models

Poisson Model

If the interested variable is a count data variable, it is natural to model it as a Poisson process. The number of occurance follows a Poisson process:

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!}$$

One of Poisson's properties is that it has equal variance as its mean:

$$\mathrm{E}[Y] = \mathrm{Var}[Y] = \mu$$

In a count data model such as Poisson regression model, we are modeling log of the expected counts:

$$\log(\mathrm{E}(Y)) = \mathbf{X_i'}\beta$$

The log-likelihood function is

$$\log L(\beta) = \sum_{i=1}^{N} -\exp(X_i'\beta) + y_i X_i'\beta - \ln y_i!$$

To maximize it, the first order condition is

$$\sum_{i=1}^{N} (y_i - \exp(X_i'\beta))X_i' = 0$$

We can use Newton-Raphson or other optimization algorithms to find the solution for the first order condition.

QMLE Poisson

Model

Note that in equation 9, if X_i' includes a constant, then $(y_i - \exp(X_i'\beta))$ sums to zero. If $\mathrm{E}[y_i|X_i] = \exp(X_i'\beta)$, then the summation on the left-hand side has expectation of zero. Hence the only specification needed to apply equation 9 is the conditional expectation of Y given X. Even the data is not Poisson-distributed, the estimator by equation 9 is still consistent. Therefore, this estimator is called quasi_ML (QML) Poisson estimator.

Although the QML Poisson estimator is consistent under relatively weak condition, the regular variance estimator for the coefficients are not valid anymore.

If a stronger assumption is made that the data follows a Poisson distribution, then the error term has a variance which equals to the mean of Y. The estimator $\hat{\beta}_P$ (ML estimator) follows a normal distribution asymptotically with a variance matrix

$$\mathrm{Var}_{ML}[\hat{\beta}_P] = (\sum_{i=1}^{N} \mu_i X_i X_i')^{-1}$$

On the other hand, the variance matrix for the QML Poisson estimator $\hat{\beta}_{QML}$ is

$$\mathrm{Var}_{QML}[\hat{\beta}_P] = (\sum_{i=1}^{N} \mu_i X_i X_i')^{-1} (\sum_{i=1}^{N} \omega_i X_i X_i')^{-1} (\sum_{i=1}^{N} \mu_i X_i X_i')^{-1}$$

where $\omega_i = \mathrm{Var}[y_i|X_i]$ is the conditional variance of y_i.

Implementation

Poisson estimation is implemented in almost every statistical package. However, some of them may not work if you have a continuous dependent variable.

Stata's implementation of Poisson model: "poisson" and "xtpoisson" do take continuous dependent variable. Note that standard errors reported without any specification in "vce" will be likely downward biased. Therefore, always use "robust" option at least. Bootstrapped standard errors are also preferred. However, if you intend to use it as QMLE-Poisson, standard errors need to be adjusted. Those two procedures do not adjust for standard errors. A user-written program called xtpqml calls for xtpoisson and it calculates robust standard error which is suggested by Wooldridge (1999).

Fixed-effect Poisson model

We specify a panel data count model as:

$$E[y_{it}|\alpha_i, X_{it}] = \alpha_i \exp(X'_{it}\beta) = \exp(\gamma_i + X'_{it}\beta).$$

It turns out that for fixed-effect Poisson model, just like in linear regression case, there is NO incidental parameter problem. Poisson MLE is the same for conditional and unconditional likelihood.(See Cameron and Trivedi, 2005, Page 805).

Therefore, the coefficient estimates are the same for conditional or unconditional fixed-effect Poisson model. That is, in Stata, "xtpoisson, fe" will return the same results as "xi: poisson i.group", is the same as "xtpqml, fe". The only difference is that "xtpqml, fe" returns "correct" standard errors for QMLE Poisson. You can also calculate standard errors with the other commands, using clustered standard errors, or bootstrapped standard errors.

Conditional fixed effect negative binomial model is also possible to be consistent for some specifications. But Poission fixed effect is more popular since it is QMLE.

Negative Binomial model

The Poisson model has a restrictive property that the conditional variance equals the conditional mean. Often we'll see "overdispersion", that is, the variance exceeds the mean, in real data. The negative binomial model is a generalization of poisson model which allows for overdispersion by introducing an unobserved heterogeneity term. In a negative binomial model,

$$\mathrm{E}[Y] = \mu\tau = e^{X'\beta} \cdot \tau$$

This extra τ term follows a $Gamma(\theta, \theta)$ distribution, with $E(\tau) = 1$ and $Var(\tau) = 1/\theta$.

Conditional on X and τ, the distribution of Y is still poisson:

$$\Pr[Y = y|X, \tau] = \frac{e^{-\mu\tau}(\mu\tau)^y}{y!}$$

Conditional on X only, the distribution of Y is negative binomial:

$$\Pr[Y = y|X] = \frac{\theta^\theta \mu^y \Gamma(\theta + y)}{\Gamma(y + 1)\Gamma(\theta)(\mu + \theta)^{\theta+y}}$$

This distribution still has conditional mean of μ, but variance is $\mu(1 + (1/\theta)\mu)$.

When $\alpha = 1/\theta$ approaches zero, the negative binomial model converges to poisson model.

Models for truncated counts

In some situations, all we observe are non-zero counts, because of the way data was collected. For example, we only observe people's visits to a park, only if they visit. To model how many times they visit, based on this kind of data set, we need to use zero-truncated count models.

$$\Pr[Y = y|y > 0, X] = \frac{\Pr(y|X)}{\Pr(y > 0|X)} = \frac{\Pr(y|X)}{1 - \exp(-\mu)}$$

Thus we are modeling the counts given that we have a positive outcome. Both zero-truncated poisson ("ztp" in Stata) and zero-truncated negative binomial ("ztnb" in Stata) are estimated by Maximum-likelihood method.

We need to pay extra attention to the "over-dispersion" problem when dealing with zero-truncated data. When data is not truncated, coefficients estimated by poisson model is consistent, even if data is "overdispersed". However, if we have zero-truncated sample, then "over-dispersion" results in inconsistent coefficient estimation (Grogger and Carson, 1991). Therefore, "over-dispersion" needs to be tested before using truncated poisson model or truncated negative binomial model.

The hurdle regression model

Although negative binomial model relaxes the assumption of equal mean and variance, some researchers prefer modeling the process of generating zeros differently from the process of generating other values. Suppose zero counts are generated by a binary process:

$$\Pr[y = 0|X] = \frac{\exp(X\gamma)}{1 + \exp(X\gamma)} = \pi$$

Positive counts are generated by a truncated count process.

In this model, zero is a "hurdle" to get past before reaching positive counts. The hurdle model is estimated by two seperate equations. It is easy to estimate. However, the marginal effect is tricky to calcuate, since and independent variable, if apprearing in both equations, will have effect through both equations.

$$\mathrm{E}(y|X) = [\pi \times 0] + (1 - \pi) \times \mathrm{E}(y|y > 0, X) = (1 - \pi) \times \mathrm{E}(y|y > 0, X)$$

In stata, there are commands such as "hplogit", which is a hurdle model with first equation logit, and second equation truncated poisson. It is the same as two seperate estimations; namely "logit" first, then "ztp" on positive counts.

The difference between hurdle model and a Heckman selection model is that in a Heckman model, two equations need to be jointly estimated. There must be exlusion condition to make the selection equation identifiable. That is, the second stage equation need to take account of the selection bias; therefore, some instrument for selection is needed for the first stage equation. On the other hand, a hurdle model assumes that the two processes are seperate; there is no selection.

Zero-inflated count models

A zero-inflated count model also models two different processes. One is a binary process: generating zero or non-zeros. The second one is a count model. Note there is a different between zero-inflated and hurdle model: In hurdle model, the second process is a truncated count model (positive counts). In a zero-inflated model, a zero can come from either the binary process, or from the count process.

A second difference is that the estimation of zero-inflated model is a joint estimation of the two processes.

$$y_i \sim \left\{ \begin{array}{c} 0 \text{ with probability } \psi \\ g(y_i|X_i) \text{with probability } 1-\psi \end{array} \right.$$

Then

$$P(Y_i = y_i|X_i, Z_i) = \left\{ \begin{array}{c} \psi(\gamma' Z_i) + (1-\psi(\gamma' Z_i))g(0|X_i) \text{ if } y_i = 0 \\ (1-\psi(\gamma' Z_i))g(y_i|X_i) \text{ if } y_i > 0 \end{array} \right.$$

This says that every non-zero observation follows a count process $g(y_i|X_i)$. Every zero observation has two possible sources: from a binary process (with probability $\psi(\gamma' Z_i)$) and from a count process (with probability $1-\psi(\gamma' Z_i)$ into the count process and then with probability of $g(0|X_i)$ to be a zero). The count process can be a Poisson process ("zip" in Stata) or a Negative Binomial process ("zinb" in Stata).

An example of extra-zero count data model comparison

http://hbs-rcs.github.io/blog/2014/09/17/poisson-models/

How to interpret coefficients and calculate marginal effects in Discrete Choice Models

Binary Response Models

In a linear model

$$y_i = X_i'\beta + \epsilon,$$
$$E(y_i) = X_i'\beta$$

When y_i is binary (meaning it takes two values, 1 or 0), We can also do an OLS regression on it, but a more popular way is to use a non-linear model. The most popular choices for modeling binary response are logit model and probit model. Both of them use the same idea: use a link function to map the binary variable into a continuous variable which is a linear function of the predictors.

Suppose p_t is the probability that $y_i = 1$. Then

$$p_i = E(y_i) = g^{-1}(X_i'\beta),$$

where $g^{-1}()$ is a function that maps from the linear predictor to y_i, which is often called index function or inverse link function.

This says that the expectation of y_i (equivalently, p_i) is not a linear function of x_i's, but by using the link function, we are still able to model a transformed p_i as a linear function of x_i's:

$$g(p_i) = X_i'\beta.$$

$g^{-1}()$ can be a normal CDF then we have a probit model:

$$g^{-1}(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp[-(t^2/2)]dt,$$

or it can be a logit:

$$g^{-1}(z) = \Lambda(z) = e^z/(1+e^z).$$

Marginal Effects

Often times we are interested in the marginal effects of the predictors. In a linear model, it's easy: the coefficient on x_1 means when x_1 increase by one unit, how much will y increase.

In a binary response model,

$$\frac{\partial p}{\partial x} = \frac{\partial g^{-1}(X_i'\beta)}{\partial x}$$

In the case of probit:

$$\frac{\partial p}{\partial x} = \frac{\partial \Phi(X_i'\beta)}{\partial x} = \phi(X_i'\beta)\beta,$$

where $\phi()$ is the density function of normal distribution.

In the case of logit:

$$\frac{\partial p}{\partial x} = \frac{\partial \Lambda(X_i'\beta)}{\partial x} = \gamma(X_i'\beta)\beta,$$

where $\gamma()$ is the density function for logistic distribution:

$$\gamma(X_i'\beta) = \Lambda(X_i'\beta)(1 - \Lambda(X_i'\beta)) = p(1-p).$$

Therefore for logit, it's easier to calculate the marginal effect from coefficient estimate β, all you have to do is to plug in sample proportion mean $\bar{p}$ and coefficient estimate $\hat{\beta}$:

$$\frac{\hat{\partial p}}{\partial x} = \bar{p}(1-\bar{p})\hat{\beta}$$

The logit model can also be formulated as

$$\log(\frac{p}{1-p}) = \mathbf{X_i'}\beta$$

which says the logarithm of the odds (the ratio of the two probabilities) is equal to $\mathbf{X_i'}\beta$. Therefore,

$$p = \frac{\exp(X_i'\beta)}{1 + \exp(\mathbf{X}_t\beta)} = \Lambda(X_i'\beta)$$

Often times we have discrete valued predictors, such as gender, or other dummy variables. In that case, it makes more sense to ask what's the effect of gender being 1 vs. 0 (female vs. male). What this question is can be formulated as:

$$E(p|x_1 = 1) - E(p|x_1 = 0).$$

In empirical analysis, it's often calculated by calculating the predicted probability by setting x_1 to 1 and by setting x_1 to 0. Then calculate the difference.

Count Data Models

In a count data model such as Poisson regression model or Negative Binomial model, we are modeling log of the expected counts:

$$\log(\mathrm{E}(y)) = \mathbf{X}_\mathbf{i}'\beta$$

Therefore the marginal effect of x_1*, for example, on* $\mathrm{E}(y)$ *is*

$$\frac{\partial \log \mathrm{E}(y)}{\partial x} = \hat{\beta},$$

which means β *is the marginal effect of* x *on* log *of the expected counts. Suppose the expected counts at* x *is* μ_x*, then*

$$\frac{(\partial \mu_x)/\mu_x}{\partial x} = \hat{\beta},$$

which says that β *represents the marginal effect of percentage change in* μ_x *with respect to* x*.*

Another way to formulate this is to use the Incidence Rate Ratio Interpretation. Suppose the expected counts at x *is* μ_x*, and expected counts at* $x+1$ *is* μ_{x+1}*.*

$$\log(\mu_x) = x'\beta.$$

Therefore,

$$\mu_x = \exp(x'\beta),$$

and

$$\frac{\mu_{x+1}}{\mu_x} = \exp(\beta).$$

This says that the exponentiated β *is the incidence rate ratio of the expected counts, if* x *increases by* 1*.*

To calculate the marginal effect of x *on* y *(or* $\mathrm{E}(y)$*), we need to calculate*

$$\frac{\partial \mu_x}{\partial x} = \hat{\beta}\mu_x = \hat{\beta}\exp(X_i'\hat{\beta}),$$

if we evaluate μ_x *at predicted value.*

How to calculate marginal effects in Stata

Stata's margins command is a powerful command to get the predicted value or marginal effects after an estimation command. Check Stata's manual for more details.

Chapter 10

Panel Data Models

Background

Panel data are repeated observations on the same "unit" (we call that cross-sectional units) over time. Panel data models are used since it has advantages over cross-sectional models or time-series models. The major advantage it has over cross-sectional models is: {**Controlling for individual heterogeneity**}.

For example, Baltagi and Levin (1992) studies cigarette demand across 46 states for the years 1963-88. Consumption is modeled as a function of lagged consumption, price and income. These variables vary across states and time. However, there are other variables that may be state-invariant (z_i or time-invariant (w_t). Examples of z_i are religion and education. For example, Utah, as a Mormon state, has very low cigarette demand due to religious reason. Generally we consider it does not change over time or change very little over time. Examples of w_i include cigarette commercials on national TV or radio. Panel data models are able to control for individual heterogeneity (or cross-sectional heterogeneity) while cross-sectional models are not. In many (or most) social science studies, there is "unobserved effects" which is embedded in the error term. In other words, we have "omitted variable" problem. Without controlling for it, the estimation results are generally biased and inconsistent.

Panel data give more informative data, more variability and less collinearity among variables. Often time series data suffers from multicollinearity, while panel data has more variability from its cross-sectional units.

The basic unobserved effects model can be written as:

$$y_{it} = \mathbf{x_{it}}\beta + c_i + u_{it}$$

where i indexes "units" (can be people, firms, or households, etc.), an t indexes time periods.

Comparing to regular cross-sectional models, there is an extra term c_i. c_i can be treated as a random effect or fixed effect. Traditionally it is distinguished by whether it is estimated as a random variable or a parameter. Modern econometrics tends to distinguish by the correlation between c_i and $\mathbf{x_{it}}$. If there is no correlation between c_i and $\mathbf{x_{it}}$, then it's a random effect; otherwise, it's a fixed effect.

It is true that in some cases it is hard to justify that If there is no correlation between c_i and $\mathbf{x_{it}}$, which is one reason that people in economics tend to use fixed-effect models. However, fixed-effect models have its own limitations. One of them is: It is hard to justify that ALL the cross-sectional units' characteristics other than those already in the model do not change over time.

Random Effect Methods

A random effect model puts c_i into the error term, then estimate by FGLS (feasible GLS). Random effect model is sometimes called "Error-components model" since the overall error term is divided into an individual level error term (c_i) and individual-time level error term (u_{it}).

Under a random effect model, the variance-covariance matrix of the error term becomes:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \cdots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \cdots & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

The random effect estimator is:

$$\hat{\beta}_{\mathbf{RE}} = (\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i'} \hat{\mathbf{\Omega}}^{\mathbf{-1}} \mathbf{X_i})^{-1} (\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i'} \hat{\mathbf{\Omega}}^{\mathbf{-1}} \mathbf{y_i})$$

It is a special case of GLS. Before calculating $\hat{\beta}_{RE}$, σ_c^2 and σ_u^2 need to be estimated. This is generally done by using pooled OLS estimates.

Fixed Effect Methods

Random effect methods assume that c_i be orthogonal to $\mathbf{x_{it}}$. In many applications, the whole point of using panel data is to allow c_i be correlated with $\mathbf{x_{it}}$. We need fixed-effect models in those cases.

For the same model as in equation 10, the fixed-effect methods try to eliminate c_i using fixed effects transformation, or "within transformation". The FE transformation is to "de-mean" each observation by subtracting the group mean. Basically subtracting equation 10 from the following equation:

$$\bar{y}_i = \bar{\mathbf{x}}_\mathbf{i}\beta + c_i + \bar{u}_i$$

where $\bar{y}_i$, etc., means group means. For example, if we have 50 states with state level data of cigarette consumption across years, then groups means states, and we have 50 group means which are state level average cigarette consumption across years.

The reason for "demeaning" is to remove c_i from final estimation. The "demeaned" regression equation is the final regression used in estimation:

Chapter 11

Survival Models

Survival function

Survival analysis is also called time to event analysis or event history analysis. We use survival models to analyze data that have the following characteristics:

1. **the dependent variable is the waiting time till the occurrence of some event;**
2. **some observations are censored; that is, some units are not in the data set anymore for some reason before an event happens. For some units we observe the event and the exact waiting time; for others it has not occurred, and all we know is that the waiting time exceeds the observation time. However, we believe if they stayed in the study, sooner or later an event would happen.**
3. **we are trying to explain (predict) the occurrence of an event with some predictors (explanatory variables).**

Let T denote the time variable, a nonnegative, continuous random variable with PDF $f(t)$ and CDF $F(t)$, where t is a realization of T. The survival function is defined by

$$S(t) = Pr(T > t) = 1 - F(t)$$

The graph of $S(t)$ against t is known as the survival curve. If there is no censoring, then the survival function is easy to estimate by the ratio of number of units survives at t vs. total number of units at risk at t. The most common method to estimate the survival function with

a survival data set containing censored observations is the Kaplan-Meier method. The basic idea is to use the product of a series of conditional probabilities.

First sort the event times from the smallest to the largest:

$$t_{(1)} \leq t_{(2)} \leq ... \leq t_{(r)}$$

Then the survival curve is estimated by:

$$\hat{S}(t) = \prod_{j|t(j)\leq t} (1 - \frac{d_j}{r_j}),$$

where r_j is the number of units at risk at $t_{(j)}$ and d_j is the number of "failure" or events at $t_{(t)}$. Note that units censored at $t_{(j)}$ are included in r_j.

The variance of this estimator can be estimated by:

$$\hat{var}(S(t)) = (\hat{S}(t))^2 \sum_{j|t(j)\leq t} \frac{d_j}{r_j(r_j - d_j)}.$$

Hazard function

We are often interested in which periods have the highest or lowest change of "death" or "failure" or other events. That is, we may be interested in the probability that a state ends between t and $t + \Delta t$ conditional on having reached t in the first place.

The probability is

$$\mathbf{Pr}(t < T \leq t + \Delta t | T \geq t) = \frac{F(t + \Delta t) - F(t)}{S(t)}$$

The hazard function is defined by

$$h(t) = \lim_{\Delta t \to 0} \mathbf{Pr}(t < T \leq t + \Delta t | T \geq t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

One of the simplest functional forms is the exponential distribution

$$f(t, \theta) = \theta e^{-\theta t}$$

Therefore, the hazard function is

$$h(t) = \theta$$

A much more flexible functional form is Weibull

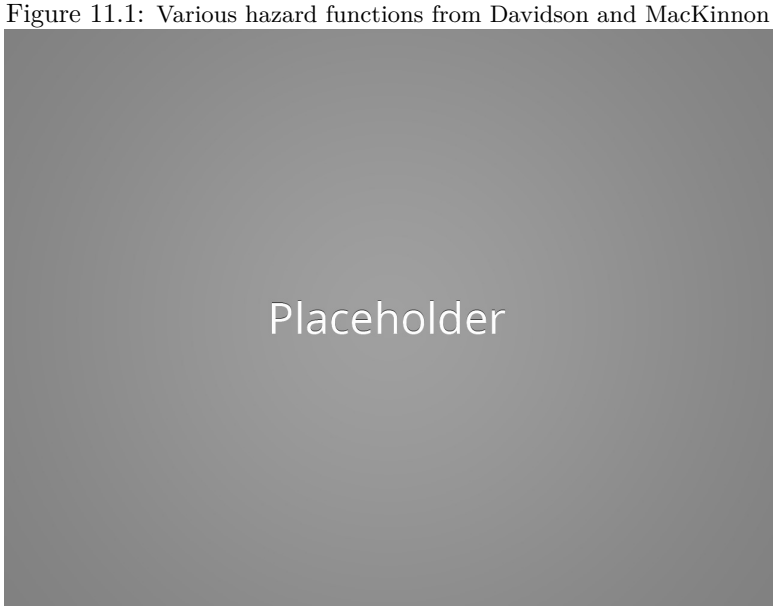
$$f(t, \theta, \alpha) = 1 - \exp(-(\theta t)^{\alpha}).$$

Hazard function can be shown to be

$$h(t) = \alpha \theta^{\alpha} t^{\alpha - 1}$$

However, Weibull distribution does not allow for the possibility that the hazard may first increase then decrease over time. Log-normal distribution does allow this.

Figure 11.1: Various hazard functions from Davidson and MacKinnon



Log-likelihood function

Suppose we have n units with a survival function $S(t)$, density $f(t)$, and hazard $h(t)$. If for unit i, the event happens at t_i, it's contribution to the likelihood function is the density at that point:

$$L_i = f(t_i) = S(t_i)h(t_i)$$

If no event at t_i, all we know is that the lifetime exceeds t_i. The probability is

$$L_i = S(t_i),$$

which is the contribution to the likelihood function from a censored observation.

Therefore, if U denotes the set of uncensored observations, the log-likelihood function for the entire sample can be written as

$$\ell(t,\theta) = \sum_{i\in U} \log h(t_i|\mathbf{X}_i,\theta) + \sum_{i=1}^{n} \log S(t_i|\mathbf{X}_i,\theta)$$

Proportional Hazard Models

Model and likelihood

One class of models that is quite widely used is the proportional hazard models (based on Cox (1972)).

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i'\beta),$$

In this model $h_0(t)$ is called baseline hazard rate; it describes the risk for units with $\mathbf{x_i} = \mathbf{0}$. $\exp(\mathbf{x}_i'\beta)$ represents the relative risk, a proportionate increase or decrease in risk, due to $\mathbf{x_i}$. The interpretation of β is that $\exp(\beta_i)$ gives the relative risk change associated with an increase of one unit in x_i, all other explanatory variables remaining constant.

β is estimated by maximizing a partial likelihood function. It is partial likelihood function since the baseline hazard is factored out. Suppose one event happens at time t_j. Conditional on this event the probability that case i dies (meaning this event happened to case i, not other cases in the risk set) is

$$L_i = \frac{h_0(t)\exp(x_i'\beta)}{\sum_l I(T_l \geq t)h_0(t)\exp(x_l'\beta)} = \frac{\exp(x_i'\beta)}{\sum_l I(T_l \geq t)\exp(x_l'\beta)}$$

We can see here that Cox's model is for continuous time survival data. It is assumed there is no ties at time t_j (in continuous time is not possible for two events to happen at the same time). In reality, we observe ties of event time, because in many situations, we observe grouped data. For example, we observe case m and n have events at the same time j. In that case, in the denominator, whether we include case m in the calculation of the likelihood of case n is ambiguous. In that case, special treatment is needed. A standard approximation (Breslow and Peto) is to let

$$L_{m\in D(t_j)} \simeq \frac{\prod_{m\in D(t_j)} \exp(x'_m\beta)}{[\sum_{l\in R(t_j)} \exp(x'_l\beta)]^{d_j}}$$

where $D(t_j)$ is the set of cases that die at time t_j and d_j denotes the number of cases that die at time t_j. This approximation works well with small number of ties relative to total number of cases at risk.

One important feature (disadvantage?) for the proportional hazard model is that for all units, the effect is the same at all time t. To see this:

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t)\exp(\mathbf{x_1}'_i\beta)}{h_0(t)\exp(\mathbf{x_2}'_i\beta)} = \frac{\exp(\mathbf{x_1}'_i\beta)}{\exp(\mathbf{x_2}'_i\beta)},$$

which does not depend on t.

One advantage of Cox's model is that it is considered as a semi-parametric model since it does not have to specify the distribution of the survival time. The estimation process relies only on the order in which events occur, not the exact time they occur. Parameter estimates are derived assuming continuous survival times.

Interpretation

In equation 11, if we would like to calculate the partial effect of x_j on h, we have:

$$\partial h(t|\mathbf{x}_i)/\partial x_j = h_0(t)\exp(\mathbf{x}'_i\beta)\beta_j = \beta_j h(t|\mathbf{x}_i),$$

If we do it the other way, plug in $x_j + 1$ will generate

$$h_0(t)\exp(\mathbf{x}'_i\beta + \beta_j) = \exp(\beta_j)h(t|\mathbf{x}'_i\beta)$$

Therefore, one unit change in x_j will have a change of $1-\exp(\beta_j)$ times the original hazard.

Example in R

```
library(ggfortify)
```

```
## Error in library(ggfortify): there is no package called 'ggfortify'
```

```
library(survival)
fit <- survfit(Surv(time, status) ~ sex, data = lung)
autoplot(fit)
```

```
## Error in eval(expr, envir, enclos): could not find function "autoplot"
```

Censored observations are denoted by crosses. This is the so-called Kaplan-Meier curve.

Now let's do a Cox model example.

```
fit2 <- coxph(Surv(time, status) ~ sex + age, data = lung, method="breslow")
summary(fit2)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + age, data = lung,
##     method = "breslow")
##
##   n= 228, number of events= 165
##
##          coef exp(coef)  se(coef)      z Pr(>|z|)
## sex -0.512565  0.598957  0.167462 -3.061  0.00221 **
## age  0.017013  1.017158  0.009222  1.845  0.06506 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## sex     0.599     1.6696    0.4314    0.8316
## age     1.017     0.9831    0.9989    1.0357
##
## Concordance= 0.603  (se = 0.026 )
## Rsquare= 0.06   (max possible= 0.999 )
## Likelihood ratio test= 14.08  on 2 df,   p=0.0008741
## Wald test            = 13.44  on 2 df,   p=0.001208
## Score (logrank) test = 13.69  on 2 df,   p=0.001067
```

Examples of parametric survival models: exponential, Weibull, or loglog:

```
exp <-  survreg(Surv(time, status) ~ sex + age, data = lung, dist="exponential")
summary(exp)
```

```
##
```

```
## Call:
## survreg(formula = Surv(time, status) ~ sex + age, data = lung,
##     dist = "exponential")
##               Value Std. Error     z        p
## (Intercept)  6.3597    0.63547 10.01 1.41e-23
## sex          0.4809    0.16709  2.88 4.00e-03
## age         -0.0156    0.00911 -1.72 8.63e-02
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -1156.1   Loglik(intercept only)= -1162.3
##  Chisq= 12.48 on 2 degrees of freedom, p= 0.002
## Number of Newton-Raphson Iterations: 4
## n= 228
```

```
weibull <-  survreg(Surv(time, status) ~ sex + age, data = lung, dist="weibull")
summary(weibull)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ sex + age, data = lung,
##     dist = "weibull")
##               Value Std. Error     z        p
## (Intercept)  6.2749    0.48137 13.04 7.69e-39
## sex          0.3821    0.12748  3.00 2.72e-03
## age         -0.0123    0.00696 -1.76 7.81e-02
## Log(scale)  -0.2823    0.06188 -4.56 5.07e-06
##
## Scale= 0.754
##
## Weibull distribution
## Loglik(model)= -1147.1   Loglik(intercept only)= -1153.9
##  Chisq= 13.59 on 2 degrees of freedom, p= 0.0011
## Number of Newton-Raphson Iterations: 5
## n= 228
```

```
loglog <-  survreg(Surv(time, status) ~ sex + age, data = lung, dist="loglogistic")
summary(loglog)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ sex + age, data = lung,
##     dist = "loglogistic")
```

```
##              Value Std. Error     z        p
## (Intercept)  5.922    0.53269 11.12 1.03e-28
## sex          0.478    0.14036  3.40 6.69e-04
## age         -0.014    0.00771 -1.82 6.95e-02
## Log(scale)  -0.570    0.06543 -8.71 3.05e-18
##
## Scale= 0.566
##
## Log logistic distribution
## Loglik(model)= -1152.9   Loglik(intercept only)= -1160.9
##  Chisq= 16.07 on 2 degrees of freedom, p= 0.00032
## Number of Newton-Raphson Iterations: 4
## n= 228
```

Discrete-time Survival Models

Cox (1972) proposed a discrete-time model by modeling the odds of dying at time t_j given survival up to that point.

$$\frac{h(t_j|\mathbf{x}_i)}{1-h(t_j|\mathbf{x}_i)} = \frac{h_0(t_j|\mathbf{x}_i)}{1-h_0(t_j|\mathbf{x}_i)}\exp(\mathbf{x}_i'\beta),$$

which is similar to Cox's proportion hazard model for continuous time. If we take logs on both sides, we have

$$\text{logit}(h(t_j|\mathbf{x}_i)) = \alpha_j + \mathbf{x}_i'\beta,$$

This looks very similar to a logit model. In fact, we can fit a discrete-time proportional hazard model by running a logistic regression on a set of pseudo observations generated by "filling in" the observations between the start time to the event time or time at the end of the study period (censored). The outcome of this process is 0 unless there is an event, then it turns 1. It is treated as independent Bernoulli observations with probability given by hazard h_{ij} for individual i at time t_j. The likelihood function for the discrete-time survival model coincide with the binomial likelihood of independent binomial process ([?] has more details).

Do it in Stata

To do a discrete-time survival model in Stata, your data set needs to be set up as by unit-time. For example, observations by company month. The event (or failure) variable needs to be set to zero until

the event happens. If no event until the end of the study period, then it is censored. A data set by unit-time can be analyzed by logit model, optionally with clustered standard error. To set up a structure like this, user-written programs called *dthaz* **and** *prsnperd* **(means person-period) can be used.**

To do a continuous-time survival model in Stata, your data can be either set up as discrete-time case (in which case you can use a time-varying explanatory variable) or a single-record-per-person data set. Stata has build-in command *stset* **to set up a survival analysis structure. Then you can run** *stcox* **for Cox's proportional hazard model.**

Chapter 12

Dynamic Panel Data

when is it a problem

Model setup:

$$y_{i,t} = \gamma y_{i,t-1} + X\beta + C_i + \epsilon_{i,t}$$

Pooled regression is biased and inconsistent. The problem is:

$$Cov(y_{i,t-1}, (C_i + \epsilon_{i,t})) \approx \frac{\sigma_c^2}{(1-\gamma)}$$

Random effects model is biased and inconsistent, for the same reason that $y_{i,t-1}$ is necessarily correlated with C_i, which is part of the composite error term in the random effect model.

Fixed effect model setup:

$$y_{i,t} - \bar{y}_i = (X_{i,t} - \bar{X}_i)'\beta + \gamma(y_{i,t-1} - \bar{y}_i) + (\epsilon_{i,t} - \bar{\epsilon}_i)$$

Anderson and Hsiao (1981) show that

$$Cov((y_{i,t-1} - \bar{y}_i), (\epsilon_{i,t} - \bar{\epsilon}_i)) \approx \frac{\sigma_\epsilon^2}{T(1-\gamma)^2}[\frac{(T-1) - T\gamma + \gamma^T}{T}]$$

which indicates that the correlation between the demeaned lagged y is correlated with the demeaned error term. The correlation may be large if T is small, but when T is big, then the correlation goes down to near zero. When T is big, we don't have a problem with fixed effect model.

how big is the bias

When T is small, which is typically the case in many microeconomic settings, then we say the fixed effect model with a lagged dependent variable is inconsistent. But how bad is it?

The limit of $\hat{\gamma} - \gamma$ is approximately $-\frac{(1+\gamma)}{T-1}$. When $T = 10$ and $\gamma = .5$, the bias is about -0.167 which is $1/3$ of the true value.

But a simulation study (Judson and Owen 1999) shows that the bias can be as big as 20% of the true coefficient even when $T = 30$.

Chapter 13

Anderson and Hsiao estimator

Anderson and Hsiao (1981) suggested looking at the first difference estimator:

$$y_{i,t} - y_{i,t-1} = (X_{i,t} - X_{i,t})'\beta + \gamma(y_{i,t-1} - y_{i,t-2}) + (\epsilon_{i,t} - \epsilon i, t-1)$$

or

$$\Delta y_{i,t} = \Delta X_{i,t}\beta + \gamma \Delta y_{i,t-1} + \Delta \epsilon_{i,t}$$

This does not solve the endogeneity problem, since $\Delta y_{i,t-1}$ is still correlated with the error term. AH's idea is to instrument $\Delta y_{i,t-1}$ with the past level $y_{i,t-2}$, or past difference $y_{i,t-2} - y_{i,t-3}$. This estimator is consistent, since neither of these instruments is correlated with $\Delta \epsilon_{i,t}$, assuming error is not auto-correlated.

Chapter 14

Arellano-Bond estimator

Arellano and Bond (1991) expanded the idea by using additional lags of the dependent variable as instruments. For example, both $y_{i,t-2}$ and $y_{i,t-3}$ can be used as instruments. In fact, as t increases, the number of instruments avialable also increases. In period 3 only $y_{i,1}$ is available. In period 4 $y_{i,1}$ and $y_{i,2}$ are available. In period 5 $y_{i,1}$ and $y_{i,2}$ and $y_{i,3}$ are availabe, and so on. In other words, we'll have an instrument matrix with one row for each time period that we are instrumenting:

$$Z_i = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ y_{i,1} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & y_{i,1} & y_{i,2} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & y_{i,1} & y_{i,2} & y_{i,3} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & y_{i,1} & \cdots & y_{i,T-2} \end{bmatrix}$$

This is the so called difference GMM estimator.

Chapter 15

Blundell and Bond estimator

AB model has a problem: when the instruments are weak, the estimator is not good. That will happen when y follows random walk or near random walk. In that case the past levels won't be a good predictor of the future changes.

BB model comes in to instrument levels with differences. In stead of differencing to remove the fixed effects, it keeps the fixed effects and difference the instruments to make them exogenous to the fixed effects. The basic idea is that $\Delta y_{i,t}$ is not correlated with C_i. Therefore it can be used as instruments in the fixed effects model without demeaning the variables.

This is implemented by so called system GMM. In this case, we have fixed effects in the model, but do not attempt to purge the fixed effect by demeaning or first differencing. Instead, we use past differences of y as instruments, assuming $\Delta y_{i,t}$ is not correlated with C_i. In this case, time-invariant variables can be used in the model.

Chapter 16

Missing data

Data set like:

y x z w q

1 4 2 ? ?

4 ? 1 2 ?

2 ? ?

? 2 1

? ? ?

Different cases under different assumptions

1. **Missing completely at Random (MCAR): the occurrence of missing data is not related to the missing value, the values of any other variables, or the pattern of missingness in other variables. Too good to be true situation.**
2. **Missing at Random (MAR): the occurrence of missing values for a variable is random, contingent on the value or missingness of observable variables. Or, the missingness can be modeled.**
3. **Missing Not at Random (MNAR): the occurrence of missing values is systematically related to unknown or unmeasured covariate factors. Hopeless situation.**

The MAR case is the one we are interested in.

For MCAR: you lose efficiency if you simply drop the data. For MAR: if you don't model it, you suffer bias and efficiency.

If we model missingness, for both MCAR, you gain efficiency. For MAR, you correct bias and gain efficiency.

Old methods

1. **Listwise deletion**

2. **Mean imputation.**

Problems: * understates variability in the imputed variable * does not recover associations between variables.

So standard errors are in general under-estimated.

3. **Regression-based imputation**

Use a model such as $x = \alpha_0 + \alpha_1 z + \alpha_2 y$. But that introduce extra noise that are not accounted for (similar to generated variable problem).

4. **Interpolation of panel data**

That is, to use the observation from last period or a linear interpolation for the same unit.

Modern methods

- **Account for uncertainty in imputed variable**
- **Use a model to predict missing observation.**
- **Instead of picking one, pick many**
- **uncertainty is represented by VCV matrix of the coefficients used to predict missing values.**

Basic ideas

Imputation model: $x = \alpha_0 + \alpha_1 z + \alpha_2 y$

Main model: $y = \beta_0 + \beta_1 x + \beta_2 z$

1. **Pick m values of α out of the asympototic distribution, the multivariate normal, using α and VCV $\hat{\Sigma}$ for the mean and VCV of the distribution $\Phi(\alpha\Sigma)$.**
2. **predict m values of the missing values, creating m data sets.**
3. **calculate m new estimates of $\tilde{\beta} = \sum_{m=1}^{M} \tilde{\beta}_m$ using each of the imputed data sets. then calculate the standard error of $\tilde{\beta}$.**

$V_\beta = W + (1 + 1/m)B$

where $W = \frac{1}{m}\sum_{m=1}^{M} s_m^2$, $B = \frac{1}{m=1}^{M}(\tilde{\beta}_m - \tilde{\beta})^2$, in other words, the within-imputation and between-imputation variation.

MI through Chained Equations (MICE) (by Buuren)

1. **Discard observations with all missing.**
2. **Fill in the missing data with random draws from the observed values.**
3. **Move through the columns and perform single-variable imputation using some method.**
4. **Replace the original replacements with the fitted replacements. Repeat 3 for a large number of times, or with a convergence criteria.**
5. **Do 1-4 m times to create m imputed data sets.**

Many ways to implement MICE.

- **Regression (linear, logit, multinomial), get $\hat{w}$ or $f(\hat{w})$ sample.**
- **The default is Predictive Mean Matching (PMM)**

a. **create predicted value**

b. **pick three cases that have the closed predicted values in terms of Euclidean distance.**

c. **randomly choose one of the three values to impute.**

Bayesian Data Augmentation

- **MICE is Markov Chain**
- **Build a missing data model into a Bayesian model, treating missing values as another parameter to estimate by drawing out of its posterior distribution.**

$f(\beta, y_{miss}|y_{obs}) \sim f(y_{obs}|\beta, y_{miss})f(\beta, y_{miss})$

- **We integrate out the missing values by sampling from the total distribution, then averaging out the beta distributions over the space of missing data points.**
- **As long as data is MAR, the likelihood of missingness is not related to β (ignorability), this is fine.**

FIML

If there are no missing data, the likelihood function is

$L(\mu, \Sigma) = \prod_{i=1} f(y_i|\mu, \Sigma)$

If there are missing value,

$L(\mu, \Sigma) = \prod_{i=1} f(y_i|\mu_i, \Sigma_i)$

mi in stata

```
log using mi_10_model2.log, replace

clear
set more off

set matsize 4000

use "CROSSED_ContestUserActivity FEB 2014 temp.dta", clear
*sample 10

mi set flong
*local dummies "x1 x2"
local continuous "CulturalDist_KS5D_ctr tight_ctr targ_tight_ctr user_country_openness

mi register imputed `continuous'
mi impute chained  (regress) `continuous' , add(10) rseed(1)

mi estimate, cmdok: heckprob has_won c.CulturalDist_KS5D_ctr##c.tight_ctr targ_tight_ct
```

This is a sample code. In stata, you have to do "mi set", then "mi register" to register variables you need to impute. Then "mi impute chained" if you have multiple variables to impute. Then the imputation will take all variables you registered (imputed or regular) in the prediction model unless you specify otherwise. Then do "mi estimate".