# ECON7772: Econometric Methods: Econometric Methods LECTURE NOTES PART 2

Arthur Lewbel, Department of Economics, Boston College

January 11, 2023

## Contents

# 1 Lecture 08. GLS and non i.i.d Errors

Readings for this lecture are: Greene Chapters 9, and 20.

## 1.1 OLS with Heteroskedasticity and/or Autocorrelation

Suppose that $Y = X\beta + e$ and that all the G-M assumptions hold, except that instead of $E(ee') = \sigma^2 I$, we now have

$$E(ee') = \sigma^2 \Omega$$

where $\Omega$ is an $n \times n$ symmetric, positive definite matrix. We can without loss of generality assume the normalization $tr(\Omega) = n$. If the diagonal of $\Omega$ is not all ones, then we have heteroskedasticity. If any off diagonal elements of $\Omega$ are nonzero, then we have autocorrelation.

What properties will the OLS estimator $\hat{\beta} = (X'X)^{-1} X'Y$ have now? with $E(ee') = \sigma^2\Omega$ we get

$$\begin{aligned} E\left(\hat{\beta}\right) &= \beta + (X'X)^{-1} X'E(e) \\ &= \beta \end{aligned}$$

and

$$\begin{aligned} var\left(\hat{\beta}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right] \\ &= E\left[(X'X)^{-1} X'ee'X (X'X)^{-1}\right] \\ &= (X'X)^{-1} X'E(ee') X (X'X)^{-1} \\ &= (X'X)^{-1} X'\sigma^2\Omega X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'\Omega X (X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1} \frac{X'\Omega X}{n}\left(\frac{X'X}{n}\right)^{-1} \end{aligned}$$

1. $\hat{\beta}$ is unbiased: even if we have autocorrelation or heteroskedasticity, we still have unbiasedness.

2. The usual formula for standard errors, the square root of the diagonal of $\widehat{var(\hat{\beta})} = s^2(X'X)^{-1}$, is wrong.

3. The OLS estimator $\hat{\beta}$ is still linear in $e$, so $\hat{\beta}$ is normal if $e$ is multivariate normal.

4. Consistency? we can still get consistency by the following argument: We have already assumed that $Q$, defined by

   $$\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = \lim_{n\to\infty} Q_n = Q$$

   is nonsingular. Now assume in addition that

   $$\lim_{n\to\infty} \frac{\sigma^2}{n}\frac{X'\Omega X}{n} = 0.$$

Then

$$\lim_{n\to\infty} var\left(\hat{\beta}\right) = \lim_{n\to\infty} Q_n^{-1}\frac{\sigma^2 X'\Omega X}{n^2}Q_n^{-1} = 0$$

So $\hat{\beta} \xrightarrow{ms} \beta$ and therefore $\hat{\beta} \xrightarrow{p} \beta$, so $\hat{\beta}$ is consistent.

5. Is OLS asymptotically normal here? If we only have heteroskedasticity and no autocorrelation, then $\Omega$ is a diagonal matrix, but not the identity matrix, and $\lim X'\Omega X/n$ is what we have previously referred to as the matrix $R$. In this heteroskedasticity only case, the errors are inid, and so the Lindeberg-Feller central limit theorem can still holds. However, if we have autocorrelation, then the errors are correlated, and the Lindeberg-Feller CLT no longer applies. There do exist Central Limit Theorems that can be applied with dependent data, so we can still get asymptotic normality, but they requires assumptions that are stronger than those needed just for consistency above.

6. Example of inconsistency: suppose $x_i = 1$, $E(e_i^2) = \sigma^2$ and $E(e_i e_j) = 1$ for $i \neq j$. Then $X$ is an $n \times 1$ vector of ones, $Q_n = 1$, and $\sigma^2\Omega$ equals a matrix with every element on the diagonal being $\sigma^2$, and ones everywhere else. This makes

$$var\left(\hat{\beta}\right) = Q_n^{-1}\frac{\sigma^2 X'\Omega X}{n^2}Q_n^{-1} = 1 * \frac{n\left(\sigma^2 - 1\right) + n^2}{n^2} * 1$$

and so $\lim var\left(\hat{\beta}\right) = 1$. Since this limit is not zero, $\hat{\beta}$ does not converge in mean square, violating the above assumption. It can be shown that this example also does not converge in probability, and so is inconsistent. The problem here is that correlations across the errors are too large, making $X'\Omega X$ grow too fast..

## 1.2 GLS - Generalized Least Squares

Continue to assume that $Y = X\beta + e$ and that all the G-M assumptions hold, except that $E(ee') = \sigma^2\Omega$ with $\Omega \neq I$, so we have heteroskedasticity or autocorrelation or both. The previous section shows that OLS is still unbiased, but might no longer be BLUE, since it no longer has the same variance. In this case, can we construct a better estimator than OLS? One that is BLUE?

The answer is yes, if we know $\Omega$. The way to do so is to transform the original data $Y$ and $X$ into new data $Y^*$ and $X^*$, where $Y^* = X^*\beta + e^*$ for some new errors $e^*$ (with the same $\beta$ as the original regression) and where this new regression does satisfy all the G-M assumptions.

Here's how: Given any $n \times n$ symmetric positive definite matrix $\Omega$, one can show there exists an $n \times n$ matrix $P$ such that

$$P\Omega P' = I \text{ and } \Omega^{-1} = P'P.$$

In matrix algebra this is known as the "spectral decomposition."

Now construct new data $Y^*$ and $X^*$, and new errors $e^*$ by

$$Y^* = PY, \quad X^* = PX, \text{ and } e^* = Pe.$$

Then

$$\begin{aligned} & Y = X\beta + e \\ & \Rightarrow PY = PX\beta + Pe \\ & \Rightarrow Y^* = X^*\beta + e^* \end{aligned}$$

and now note that

$$\begin{aligned} E\left(e^* e^{*\prime}\right) &= E\left(Pee'P'\right) \\ &= PE\left(ee'\right)P' \\ &= P\sigma^2\Omega P' \\ &= \sigma^2 P\Omega P' = \sigma^2 I \end{aligned}$$

so the new regression model $Y^* = X^*\beta + e^*$ satisfies all of the G-M assumptions (note that $X$ constant implies that $X^*$ is also constant).

The GLS estimator $\hat{\beta}_{GLS}$ is defined to be just OLS on the new regression model, so

$$\hat{\beta}_{GLS} = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}Y^*.$$

This $\hat{\beta}_{GLS}$ is BLUE because $Y^* = X^*\beta + e^*$ satisfies the G-M assumptions.

What other properties does $\hat{\beta}_{GLS}$ have? If $e$ is normal, then $e^*$ is normal (because $e^*$ is linear in $e$), and in that case $\hat{\beta}_{GLS}$ is normal. If the errors are not normal but the sample size is large, then $\hat{\beta}_{GLS}$ will be asymptotically normal if $\lim X^{*\prime}X^*/n$ is nonsingular and if $X^{*\prime}e^*$ satisfies a CLT.

We haven't said how to construct $P$, and hence how to construct $\hat{\beta}_{GLS}$. But it turns out you don't actually need to construct $P$, because

$$\begin{aligned} \hat{\beta}_{GLS} &= \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}Y^* \\ &= \left(X'P'PX\right)^{-1} X'P'PY \\ &= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y \end{aligned}$$

and

$$\begin{aligned} var\left(\hat{\beta}_{GLS}\right) &= \sigma^2\left(X^{*\prime}X^*\right)^{-1} \\ &= \sigma^2\left(X'\Omega^{-1}X\right)^{-1}. \end{aligned}$$

So you don't need $P$, you just need to know $\Omega$ to construct $\hat{\beta}_{GLS}$ and its variance.

In practice, we may not know $\Omega$, but we may have a way to estimate it. In that case, given a consistent estimate $\hat{\Omega}$, we can apply what is known as the Feasible GLS estimator

$$\hat{\beta}_{FGLS} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}Y$$

Now $\hat{\beta}_{FGLS}$ may no longer have the finite sample properties of GLS (due to estimation error in $\hat{\Omega}$), but the asymptotic properties of GLS can still hold.

## 1.3 Heteroskedasticity

Suppose our regression equation is

$$Y_i = a + bX_i + cZ_i + e_i$$

and the errors $e_i$ are heteroskedastic but not autocorrelated, so $E\left(e_i e_j\right) = 0$ for $i \neq j$ but $E\left(e_i^2\right) = \sigma_i^2$ which varies by $i$, and otherwise the G-M assumptions hold.

Suppose we can find constants $w_1$,$w_2$,...,$w_n$ such that

$$var\left(\frac{e_i}{w_i}\right) = \kappa$$

for some constant $\kappa$. This means that each $w_i$ is proportional to $\sigma_i$. Since we never actually observe the variance of each error $e_i$, it may be hard to find such a $w_i$ constants.

In this model the GLS transformation (converting the regression to one that satisfies G-M) consists of dividing every variable in the model, including the constant, by $w$:

$$\frac{Y_i}{w_i} = a\frac{1}{w_i} + b\frac{X_i}{w_i} + c\frac{Z_i}{w_i} + \frac{e_i}{w_i}$$

which we can rewrite as

$$Y_i^* = aU_i^* + bX_i^* + cZ_i^* + e_i^*$$

This regression satisfies G-M, and so is BLUE if we estimate it by OLS. This estimator is called Weighted Least Square (WLS). This is a special case of GLS, where

$$\Omega = \begin{bmatrix} w_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n^2 \end{bmatrix} \text{ and } P = \begin{bmatrix} w_1^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n^{-1} \end{bmatrix}$$

Intuitively, OLS is inefficient under heteroskedasticity because it weighs all observations equally. More efficient is to give more weight to the observations that have low variance, because those observations have a higher probability of being close to the true line. This is what WLS does.

In practice, often we don't do WLS, but instead use the WLS idea to construct a better model that is less heteroskedastic. For example, suppose each observation $i$ is a country, the variables in the model include gross domestic product (GDP) and the consumer price indes (CPI), and $w_i$ is the population of country $i$. WLS would divide every regressor, including the constant, by population. More sensible might be to just divide variables like GDP by population, and not variables like CPI, or the constant term. This would then change the model to one that both makes economic sense and may suffer from less heteroskedasticity.

### 1.3.1 White's Test

A few different tests exist for heteroskedasticity. One of the commonest ones is White's test, which works as follows.

1. Estimate the original model by OLS, say $Y_i = b_0 + b_1X_i + b_2Z_i + e_i$, to obtain residuals $\hat{e}_i$.

2. Run the auxiliary regression

$$\hat{e}_i^2 = a_0 + a_1X_i + a_2Z_i + a_3X_i^2 + a_4Z_i^2 + a_5X_iZ_i + u_i$$

3. Test the joint significance of all coefficients except for the constant term.

$$H_0 : a_1 = a_2 = \cdots = a_5 = 0$$

This test is asymptotic. The Lagrange Multiplier test of this null has a particularly simple form - the test statistic is just $nR^2$, where $n$ is the sample size and $R^2$ is the $R^2$ statistic from this auxiliary regression. Under the null:

$$nR^2 \xrightarrow{d} \chi_5^2$$

More generally, the auxiliary regression includes all the variables in the original regression, along with all the squares and cross products of those variables. The degrees of freedom is the number of coefficients in the auxiliary regression (not including the constant term). This is White's test.

Note that this tends to be a low power test: rejection implies heteroskedasticity but non-rejection doesn't tell us much.

### 1.3.2 White Corrected Standard Errors

If we have heteroskedasticity, the best solution is to fix the problem using something like WLS. But often we can't fix it (e.g., we don't know weights $w$), and so are left doing OLS. In that case, we know our usual variance formula $\frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}$ is no longer correct. Instead, to construct correct (consistent) standard errors we need to estimate

$$var\left(\hat{\beta}\right) = \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\Omega X}{n}\left(\frac{X'X}{n}\right)^{-1}$$

The problem is that this formula depends on the term $\sigma^2 X'\Omega X/n$, and we don't know $\Omega$ (if we knew it, we could be efficient and do GLS).

Fortunately, we don't need to know or estimate $\Omega$ itself, all we need is to estimate $\sigma^2 X'\Omega X/n$, and that turns out to be easier. In particular, if we only have heteroskedasticity and not autocorrelation, then $\Omega$ is diagonal, and we have

$$\sigma^2\frac{X'\Omega X}{n} = \frac{1}{n}\sum\nolimits_i X_i X_i'\sigma_i^2.$$

Under some general conditions it can be shown that

$$plim\left(\frac{1}{n}\sum\nolimits_i X_i X_i'\hat{e}_i^2\right) = \lim_{n\to\infty}\frac{1}{n}\sum\nolimits_i X_i X_i'\sigma_i^2$$

where $\hat{e}_i$ are the residuals from doing OLS. The intuition is that $\sigma_i^2$ is the mean of $e_i^2$, and the $x_i$'s are constants, so what the sample average of $x_i x_i' e_i^2$ estimates is the average of $x_i x_i'\sigma_i^2$.

We therefore get a consistent estimator of $var\left(\hat{\beta}\right)$ given by

$$\widehat{var\left(\hat{\beta}\right)} = \frac{1}{n}\left(\frac{\sum_i X_i X_i'}{n}\right)^{-1}\left(\frac{1}{n}\sum\nolimits_i X_i X_i'\hat{e}_i^2\right)\left(\frac{\sum_i X_i X_i'}{n}\right)^{-1} \xrightarrow{p} var\left(\hat{\beta}\right)$$

The square root of the elements on the diagonal of $\widehat{var\left(\hat{\beta}\right)}$ are the standard errors of the elements of $\hat{\beta}$. These are called "White corrected standard errors" or "Heteroskedasticity robust standard errors." or "Heteroskedasticity consistent standard errors."

Note: This is an asymptotic derivation. White standard errors could give bad estimates of the true standard errors with small sample sizes. And even with moderate sample sizes, if the errors actually have little or no heteroskedasticity, then White standard errors would still be consistent, but they would be inefficient. That is, if the true errors are homoskedastic, then the usual standard error formula will be more accurate (having less estimation error) than White standard errors.

## 1.4 Autocorrelation

Autocorrelation in errors $e$ is when the errors for different observations are correlated with each other, so $E(e_i e_j) \neq 0$ for at least some $i \neq j$.

Any kind of data can have autocorrelation. For example, in cross sections you could have 'network correlation,' e.g., your $e_i$ could be correlated with your friend's $e_j$'s. Or we could have 'spatial correlation' e.g., your $e_i$ could be correlated with the $e_j$'s of others that live near you.

Autocorrelation is particularly common in time series data: whatever is left out of the model in one time period $t$, and so appears in the error $e_t$, may correlate with what's left out in the previous or next period, making $e_t$ correlate with $e_{t-1}$ or $e_{t+1}$. We will focus here on autocorrelation in time series data, mainly because most models of time series dependence are simpler than models of dependence in cross section, panel, or more complicated DGP's.

The model: $Y_t = X_t'\beta + e_t$ for time periods $t = 1, ..., T$.
Assume all the G-M assumptions are satisfied except that $E(e_t e_s) \neq 0$ for at least some $s \neq t$.

Now $E(ee') = \sigma^2\Omega$, and some or all of the off diagonal elements of $\Omega$ can be non-zero. So $\Omega$ can have very many elements we'd need to know to FGLS, or estimate to do FGLS. It helps to make some assumptions that reduce the number of unknowns.

### 1.4.1 Stationarity

A random sequence $Z_1$, $Z_2$, $Z_3$,... is defined to be strictly stationary, or strongly stationary, if, for any $s$, the joint distribution function of $Z_t$, $Z_{t+1}$,... $Z_{t+s}$ does not depend on $t$.

For example, if $Z_1$, $Z_2$, $Z_3$,... is strictly stationary, then the joint distribution of $(Z_1, Z_2, Z_3)$ is the same as the joint distribution of $(Z_{50}, Z_{51}, Z_{52})$.

A random sequence $Z_1$, $Z_2$, $Z_3$,... is defined to be weakly stationary (or covariance stationary) if, $E(Z_t)$ is the same for all $t$, $var(Z_t)$ is finite for all $t$, and and for any $s$, $cov(Z_t, Z_{t+s})$ does not depend on $t$. This last condition means that the covariance or correlation between any two elements $Z_t$ and $Z_{t+s}$ only depends on $s$, which is how far apart they are in time.

For example, if $Z_1$, $Z_2$, $Z_3$,... is stationary (either weakly or strongly), then $cov(Z_1, Z_2) = cov(Z_{10}, Z_{11})$.

Consider regression errors. The definition of the correlation between two errors $e_t$ and $e_{t-s}$ is

$$corr(e_t, e_{t-s}) = \frac{cov(e_t, e_{t-s})}{stddev(e_t) * stddev(e_{t-s})} \text{ for any } t.$$

Recall $E(ee') = \sigma^2\Omega$. Assume errors $e$ are homoscedastic and stationary.

Then $cov(e_t, e_{t-s})$ only depends on $s$, and $stddev(e_t) * stddev(e_{t-s}) = \sigma^2$ so we can let $\rho_s = corr(e_t, e_{t-s})$, which doesn't depend on $t$.

Also, having $e$ homoscedastic and stationary implies equality along diagonals of the matrix $\Omega$. For example, if $T = 4$ (recalling that $e$ is now a vector of length $T$) then:

$$\Omega = \frac{E(ee')}{\sigma^2} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

Look at the diagonals. A similar structure holds for any $T$.

### 1.4.2 Correlogram and Ljung Box Q statistic

Assume we have stationary errors $e$. Since errors have mean zero the numerator equals $cov(e_t, e_{t-s}) = E(e_t e_{t-s})$ which is the same for all $t$, and $var(e_t) = E(e_t^2) = \sigma^2$ which is the same all $t$. We can therefore estimate the correlation $\rho_s$ by

$$r_s = \widehat{corr(e_t, e_{t-s})} = \frac{\frac{1}{T-s}\sum_{t=s+1}^{T} \hat{e}_t \hat{e}_{t-s}}{\frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2}$$

where each $\hat{e}_t$ is the residual from OLS. Recalling that correlations always lie between -1 and +1, we can calculate $r_s$ for some low integers $s$, and see if any are small (close to 0) or large (close to -1 or +1). If the data have no autocorrelation, then we would expect $r_s$ for all integers $s$ to be small.

Note we only have $T - s$ pairs of observations to estimate the numerator of $r_s$, so to apply a LLN we want $s$ to not be too big relative to $T$.

A Correlogram is a bar chart, where the bars are the values of $r_1$, $r_2$, $r_3$,...$r_L$, for some chosen integer $L$. If the correlogram has some large elements, then we suspect autocorrelation. Patterns in the correlogram will suggest different possible types or models of autocorrelation. For example, if autocorrelation is present, it is common to see errors that are closer together in time having higher correlations than errors further apart in time. This would then yield a correlogram where $r_s$ is large for small values of $s$, and tend to decline as $s$ increases.

We often pay particular attention to $\rho_1$ and its estimate, $r_1$. Autocorrelation is defined to positive if $\rho_1 > 0$ and negative if $\rho_1 < 0$.

One way to test for autocorrelation is to test the null hypothesis $H_0$: $\rho_1 = \rho_2 = ... = \rho_L = 0$, for some chosen integer $L$. One good test is the Ljung-Box (1978) Q statistic:

$$Q = \sum_{s=1}^{L} \frac{(T+2)T}{T-s} r_s^2$$

Under null hypothesis $H_0$ we get $Q \xrightarrow{d} \chi_L^2$. This is an asymptotic test - it's based on applying a central limit theorem to the averages in each $r_s$.

Note: we need to choose an $L$. If $L$ is too small, then we could miss some significant correlation. But if $L$ is too large, then the test could have low power. $L$ needs to be much smaller than $T$, since otherwise $r_L$ and hence $Q$ will be very poorly estimated.

### 1.4.3 First-order autocorrelation

A model for autocorrelation is a model for the off diagonal elements of $\Omega$. If the errors are stationary, then a model for autocorrelation is a model of $\rho_1$,$\rho_2$,... $\rho_n$. There are many possible models for autocorrelation. Here we define one such model, called first order autocorrelation. This model is also called autoregressive of order one, and can be denoted as the AR(1) model.

The AR(1) model for errors is

$$\begin{aligned} Y &= X\beta + e \\ e_t &= \rho e_{t-1} + v_t \end{aligned}$$

where $|\rho| < 1$ and

$$\begin{aligned} E(v_t) &= 0 \\ E(v_t v_{t-s}) &= 0 \text{ for all } s \neq 0 \\ E\left(v_t^2\right) &= \sigma_v^2 \end{aligned}$$

These imply that $E(vv') = \sigma_v^2 I$. If $v$ and not $e$ were the errors in the regression model, then the G-M conditions would be satisfied.

Let's now see what the AR(1) model implies about $e$. First

$$\begin{aligned} e_t &= \rho e_{t-1} + v_t \\ &= \rho(\rho e_{t-2} + v_{t-1}) + v_t \\ &= \rho(\rho(\rho e_{t-3} + v_{t-2}) + v_{t-1}) + v_t \\ &= \sum\nolimits_{s=0}^{\infty} \rho^s v_{t-s} \end{aligned}$$

So $e_t$ is a weighted average of $v_t$, $v_{t-1}$, $v_{t-2}$... with weights 1, $\rho$, $\rho^2$,... From this, we can calculate the mean and variance of $e_t$:

$$\begin{aligned} E(e_t) &= \sum\nolimits_{s=0}^{\infty} \rho^s E(v_{t-s}) = 0 \\ var(e_t) &= E\left[\left(\sum\nolimits_{s=0}^{\infty} \rho^s v_{t-s}\right)^2\right] = E\left(\sum\nolimits_{s=0}^{\infty} \rho^{2s} v_{t-s}^2\right) \\ &= \sum\nolimits_{s=0}^{\infty} \rho^{2s} E\left(v_{t-s}^2\right) = \left(\sum\nolimits_{s=0}^{\infty} \rho^{2s}\right) \sigma_v^2 \\ &= \frac{\sigma_v^2}{1-\rho^2} = \sigma^2 \end{aligned}$$

which is a constant, so $e_t$ is homoscedastic, with a variance we can denote $\sigma^2$.

Now look at autocorrelations. Consider

$$\begin{aligned} E(e_t e_{t-1}) &= E\left[(\rho e_{t-1} + v_t) e_{t-1}\right] \\ &= \rho E\left(e_{t-1}^2\right) + E(v_t e_{t-1}) \\ &= \rho\sigma^2 + 0 = \rho\sigma^2 \end{aligned}$$

Note that $E(v_t e_{t-1}) = 0$ because $e_{t-1}$ is linear in $v_{t-1}$, $v_{t-2}$,... and $v_t$ is uncorrelated with these lagged values of $v$. We therefore have $cov(e_t, e_{t-1}) = \rho\sigma^2$, so

$$\rho_1 = corr(e_t, e_{t-1}) = \rho.$$

What about $\rho_2$?

$$\begin{aligned} E(e_t e_{t-2}) &= E[(\rho e_{t-1} + v_t) e_{t-2}] \\ &= \rho E(e_{t-1} e_{t-1}) + E(v_t e_{t-2}) \\ &= \rho\left(\rho\sigma^2\right) + 0 = \rho^2\sigma^2 \end{aligned}$$

So

$$\rho_2 = corr(e_t, e_{t-2}) = \rho^2.$$

Continuing in the same way, we get for any integer $s$

$$\rho_s = corr(e_t, e_{t-s}) = \rho^s$$

This implies that, if the first order autocorrelation model is correct, that we should see Correlogram estimates $r_s$ that decline exponentially. That is one way to assess if this AR(1) model is correct. One could also use estimates $r_s$ to test if $\rho_s = \rho_1^s$ for a few values of $s$.

These calculations also show that $\Omega$ has the form

$$E(ee') = \sigma^2\Omega = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \vdots \\ \vdots & \vdots & & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{bmatrix}$$

If we knew $\rho$, we'd know the entire $\Omega$ matrix.

We want $|\rho| < 1$ because it means that the further apart in time two errors are, the lower is their correlation. If we instead had $|\rho| > 1$, then the errors in observations that were a hundred years apart in time would be more highly correlated than errors one day apart.

### 1.4.4 GLS for first order autocorrelated errors

Recall that the logic of GLS is that it transforms the data to create a new model that has the same coefficients as the original model, but satisfies all the GM conditions. What transformation works for AR(1) errors? Equivalently, what transformation of the data is implied by doing GLS with the AR(1) $\Omega$ matrix?

Suppose the model is

$$Y_t = a + bX_t + e_t$$

Consider lagging this by one period

$$Y_{t-1} = a + bX_{t-1} + e_{t-1}$$

and then multiplying both sides by $\rho$, we have

$$\rho Y_{t-1} = \rho a + \rho b X_{t-1} + \rho e_{t-1}$$

Now subtract this equation from the original equation. We get

$$\begin{aligned} Y_t - \rho Y_{t-1} &= (1-\rho)\, a + b\,(X_t - \rho X_{t-1}) + (e_t - \rho e_{t-1}) \\ Y_t^* &= a^* + bX_t^* + v_t \end{aligned}$$

where $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$. This construction of $Y_t^*$ and $X_t^*$ is called quasi-differencing. The equation $Y_t^* = a^* + bX_t^* + v_t$ satisfies the G-M assumptions.

What GLS does in this case is quasi-difference the data. In particular, if we applied the spectral decomposition to construct the matrix $P$ that corresponds to the matrix $\Omega$ in the previous section we would get

$$P = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & & \\ \vdots & & & \ddots & \\ 0 & \cdots & & -\rho & 1 \end{bmatrix}$$

If we apply the GLS transformation: $PY = PX\beta + Pe$ with this matrix $P$, we will exactly get the quasi-differenced regression $y_t^* = a^* + bx_t^* + v_t$ for $t = 2, 3, ...T$. But what about $t = 1$? we can't quasi-difference the first observation because we don't observe $y_t$ and $x_t$ for $t = 0$. What GLS does instead is just multiply the first observation by $\sqrt{1-\rho^2}$, which then makes the variance of the resulting first error equal $\sigma_v^2$ (this turns out to be more efficient than just discarding the first observation).

### 1.4.5 Feasible GLS for first order autocorrelated errors

In practice we don't know $\rho$, so we need to estimate it. One method of doing so is the following, known as the Cochrane-Orcutt procedure:

1. Obtain an estimate $\widehat{\beta}$ by doing OLS on the model $Y = X\beta + e$
2. Construct residuals $\hat{e} = Y - X\widehat{\beta}$
3. Estimate $\hat{\rho}$ by regressing $\hat{e}_t$ on $\hat{e}_{t-1}$, that is, estimate by OLS $\hat{e}_t = \rho\hat{e}_{t-1} + v_t$, using observations $t = 2, ..., T$.
4. Use the $\hat{\rho}$ from step 3 to construct an estimate of $\Omega$, and re-estimate the original model using FGLS
5. Go back to step 2, using the new FGLS estimate $\widehat{\beta}$ from step 4.
6. Keep iterating the procedure until $\hat{\rho}$ doesn't change.

Other methods are to do a grid search for $\hat{\rho}$, or to assume the errors are normal and do MLE, simultaneously estimating $\rho$ and $\beta$ (note this also gives a standard error for $\hat{\rho}$).

### 1.4.6 Other models of autocorrelation

There exist many other common models of Autocorrelation. Some of them are:

1. The second order autocorrelation model, also called AR(2):

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + v_t$$

This has a correlogram where $r_1$ is almost any value, but then $r_2$, $r_3$, etc. decline exponentially. We can write $\Omega$ in terms of $\rho_1$ and $\rho_2$, and we can estimate $\rho_1$ and $\rho_2$ as in the Cochrane Orcutt procedure, in each step regressing $\widehat{e}_t$ on both $\widehat{e}_{t-1}$ and $\widehat{e}_{t-2}$. AR(3), AR(4), etc. models are analogous.

2. Moving average processes. The MA(1) model is

$$e_t = v_t + \theta v_{t-1}$$

This model has a very different correlogram from AR models. in particular,

$$\begin{aligned} e_t &= v_t + \theta v_{t-1} \\ E\left(e_t e_{t-1}\right) &= E\left[\left(v_t + \theta v_{t-1}\right)\left(v_{t-1} + \theta v_{t-2}\right)\right] \\ &= \theta E v_{t-1}^2 = \theta \sigma_v^2 \\ E\left(e_t e_{t-j}\right) &= 0 \text{ for } j > 1 \end{aligned}$$

So $\rho_1 = \theta$ but $\rho_s = 0$ for all $s > 1$. Each error $e_t$ is correlated with the adjacent errors $e_{t-1}$ and $e_{t+1}$, but has zero correlation with the errors in all other time periods. The correlogram has single large value for $r_1$, but then $r_2$, $r_3$, etc should be near zero.

3. The MA(2) model:

$$e_t = v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2}$$

Here $\rho_1$ and $\rho_2$ are nonzero, but $\rho_s = 0$ for all $s > 2$. Again MA(3), MA(4), etc., are analogous.

4. Mixtures of AR and MA models. For example, the ARMA(1,1) model:

$$e_t = \rho_1 e_{t-1} + v_t + \theta v_{t-1}$$

It is possible to fit AR or MA models to time series data even if we have no $X$'s. In this case $e_t$ would just equal $Y_t$ (or $Y_t - \overline{Y}$, or $Y_t - Y_{t-1}$). These ARMA models are sometimes used for forecasting without regressors.

It also possible to extend these models to vectors of random variables in each time period. For example, suppose $Y_t$ is a $J$ vector of variables in each time period $t$, and $v_t$ is a vector of non autocorrelated errors. A first order vector autoregressive model, denoted VAR(1), would be $Y_t = AY_{t-1} + v_t$, where $A$ is now a $J \times J$ matrix of coefficients instead of the single scalar $\rho$. VAR models are commonly used for forecasting in macroeconomics.

### 1.4.7 Newey-West or HAC Standard errors

As with heteroskedasticity, the best (most efficient) solution to autocorrelation is to fix the problem. If we're lucky, a simple AR or MA model will suffice. But often we can't fix the problem, or we can't completely fix it. For example, we might fit an ARMA model, but find even after doing so that the remaining correlogram still has a few large $r_s$ estimated autocorrelations.

Also as with heteroskedasticity, the second best solution is to do OLS, but fix the associated standard errors to account for the autocorrelation we weren't able to model. Again we have the correct general variance formula being

$$var\left(\hat{\beta}\right) = \frac{\sigma^2}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'\Omega X}{T}\left(\frac{X'X}{T}\right)^{-1}$$

(now with sample size denoted $T$ instead of $n$). But now the unknown term the term $\sigma^2 X'\Omega X/T$ is much harder to estimate than in the heteroskedasticity case, because now $\Omega$ has nonzero off diagonal terms.

With autocorrelation we have

$$\frac{\sigma^2 X'\Omega X}{T} = \left(\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t x_t' var\left(e_t\right)\right) + \frac{1}{T}\sum\nolimits_{s=1}\sum\nolimits_{t=s+1}^{T}\left(x_t x_{t-s}' + x_{t-s}x_t'\right) cov\left(e_t, e_{t-s}\right)$$

Where the first sum is the heteroskedasticity term and the second double sum is the autocorrelation terms. White standard errors replace the unknown $var\left(e_t\right)$ with $\widehat{e}_t^2$, and we could consister similarly replacing $cov\left(e_t, e_{t-s}\right)$ with $\widehat{e}_t\widehat{e}_{t-s}$. This works for small values of $s$, but for large $s$, we only have $T-s$ terms to estimate $cov\left(e_t, e_{t-s}\right)$, which means that this covariance cannot be consistently estimated for $s$ close to $T$.

In practice we can pick $L$ lags for some $L$ much smaller than $T$, and estimate this equation just letting $s$ go from 1 to L. This then only corrects the standard errors allowing for arbitrary autocorrelations $\rho_1$,...,$\rho_L$, but not higher order lags (technically, a semiparametric version of this estimator assumes that $L$ itself is a slowly increasing function of $T$, to obtain standard error estimates that are consistent with any length lags, but that topic is too advanced for now).

One other adjustment we need to make is to account for the variation in the efficiency with which each autocovariance can be estimated, by weighting each term in the double sum by a function of $s$ and $L$. The resulting estimator is:

$$\sigma^2\widehat{\frac{X'\Omega X}{T}} = \left(\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t x_t'\widehat{e}_t^2\right) + \frac{1}{T}\sum\nolimits_{s=1}^{L}\sum\nolimits_{t=s+1}^{T}\left(1-\frac{s}{L+1}\right)\left(x_t x_{t-s}' + x_{t-s}x_t'\right)\hat{e}_t\hat{e}_s$$

Using this expression we then get the Newey-West (1987) estimator

$$\widehat{var\left(\hat{\beta}\right)} = \frac{1}{T}\left(\frac{X'X}{T}\right)^{-1}\left(\sigma^2\widehat{\frac{X'\Omega X}{T}}\right)\left(\frac{X'X}{T}\right)^{-1} \longrightarrow^{p} var\left(\hat{\beta}\right)$$

The square roots of the elements on the diagonal of this matrix are called Newey-West standard errors or HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors. Unlike White standard errors, Newey-West requires that we choose a number of lags $L$.

The same comments about only being asymptotically valid, and potential inefficiency, of White standard errors apply even more so to Newey-West standard errors. They will often be poor estimates of the true standard errors unless the sample size is very large.

# 2 Lecture 09. Dynamic Models and Time Series Models

Readings for this lecture is: Greene Chapter 20

## 2.1 Lag Models

Consider the "distributed lag" model

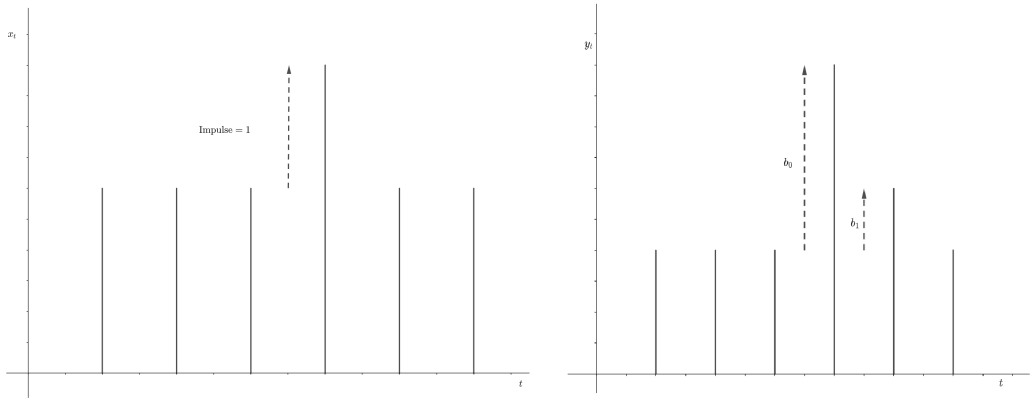
$$y_t = a + \beta_0 x_t + \beta_1 x_{t-1} + e_t$$

which can satisfy all the GM assumptions, but might still have multicollinearity. In this model, how do we interpret $\beta_0$ and $\beta_1$? $\beta_0$ is the effect of change in $x$ on $y$ in the same period that $x$ changes. Call that the short run effect. $\beta_1$ is the effect of $x$ on $y$ one period later. In this model, the total effect of a change in $x$ on $y$ in both current and later periods is $\beta_0 + \beta_1$. That's called the long response or total response.

Consider holding $x$ fixed at some value for many time periods, then in just one time period increase $x$ by 1, and then in all later periods return $x$ to its original value. This is called an impulse, and the resulting effect on $y$ over all time periods is called the impulse response.

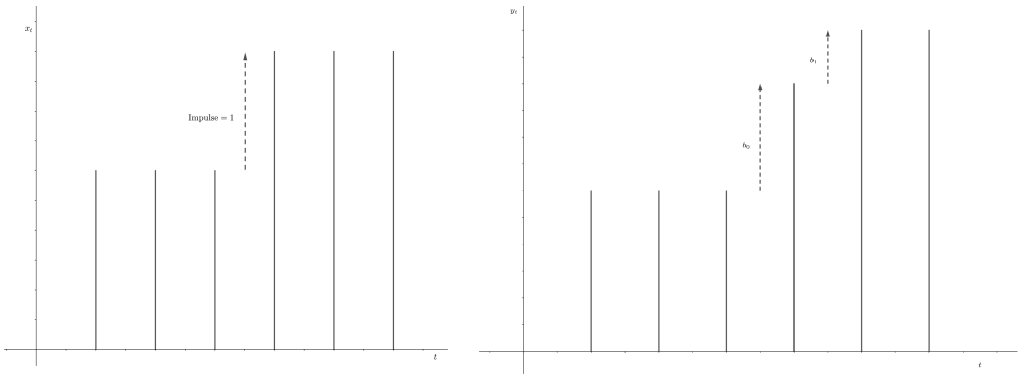
Here is the impulse response for the above model. Fix $x$ at some value $x^*$ for many periods. Then (ignoring the error term) $y$ will be fixed at $y^* = a + \beta_0 x^* + \beta_1 x^*$. If the impulse in $x$ happens in period $t$, then $x_t = x^* + 1$ while $x_s = x^*$ for all time periods $s \neq t$. The impulse response is that $y_t$ increases by $\beta_0$ to $y^* + \beta_0$ in period $t$, and $y_{t+1}$ increases by $\beta_1$ to $y^* + \beta_1$ in period $t + 1$. Then the impulse has no more effect on $y$, so $y_{t+2}$, $y_{t+3}$, etc all go back to equaling, $y^*$.

Impulse response:




The permanent response function instead looks at what happens if we fix $x$ at some value $x^*$ for many periods (up until period $t$), and then for all time periods $s \geq t$, we set $x$ to $x^* + 1$. In other words, we permanently increase $x$ from $x^*$ to $x^* + 1$. The permanent response is the resulting effect on $y$ in each time period. The permanent response here is that $y_t$ increases by $\beta_0$ to $y^* + \beta_0$ in period $t$, and $y_{t+1}$ increases by an additional $\beta_1$ to $y^* + \beta_0 + \beta_1$ in period $t + 1$, and then $y$ remains at this new higher level, so $y_{t+2}$, $y_{t+3}$, etc all equal $y^* + \beta_0 + \beta_1$.

Permanent response:




Now consider a model with more lags:

$$y_t = a + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_K x_{t-k} + e_t$$

The long run (total) response is $\beta_0 + \cdots + \beta_K$. This is the sum of all the effects on $y$ in all time periods of a unit change in $x$ in one time period.

The total response time (the number of time periods in which $y$ is affected by a unit change in $x$ in one time period) is $K$. The median lag time $t_{1/2}$ is defined such that

$$\frac{\sum_{t=0}^{t_{1/2}} \beta_t}{\sum_{t=0}^{\infty} \beta_t} = 0.5$$

This is the amount of time that goes by until $y$ experiences half of the total response $\sum_{t=0}^{\infty} \beta_t$. We can analogously define the mean lag time as

$$t = \frac{\sum_{t=0}^{\infty} t\beta_t}{\sum_{t=0}^{\infty} \beta_t} = \frac{0\beta_0 + 1\beta_1 + 2\beta_2 + \cdots + K\beta_K}{\beta_0 + \cdots + \beta_K}$$

The shorter is the mean lag or median lag, the more rapidly changes in $x$ affect $y$.

Actually estimating this model could be a problem due to multicollinearity among all the regressors. One solution to this problem is to impose some restriction on the shape of the impulse response function. Examples:

1. Estimate the model imposing the constraint that $\beta_k$ coefficients are a polynomial in $k$ (these are called polynomially distributed lags).

2. Use some variable selection technique like LASSO to set some of the $\beta_k$ coefficients to zero.

3. Replace the regressors $x_t$, $x_{t-1}$,... $x_{t-k}$ with a lagged dependent variable $y_t$. As we will see, this makes the impulse response follow a geometric decay function.

## 2.2 The lagged dependent variable model

Instead of including lags of $x_t$ as regressors, what happens if we include a lag of $y_t$? This gives us the model:

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + e_t$$

This is called a lagged dependent variable model, or the Koyck lag model. Assume $0 < \gamma < 1$.

Note 1: if we dropped $x_t$, this would be an AR(1) model of $y$. If $e_t$ is autocorrelated, then this could be an ARMA model for $y$.

Note 2: This model violates the GM assumptions, because the regressor $y_{t-1}$ is random, not constant. But it can still have good asymptotic properties, though we need $e_t$ to not be autocorrelated. Why? Because $y_{t-1}$ depends on $e_{t-1}$ the same way $y_t$ depends on $e_t$. If $e_t$ also depends on $e_{t-1}$, then $y_{t-1}$ will correlate with $e_t$ (because both depend on $e_{t-1}$), and so we'll have the problem of $E(y_{t-1}e_t) \neq 0$.

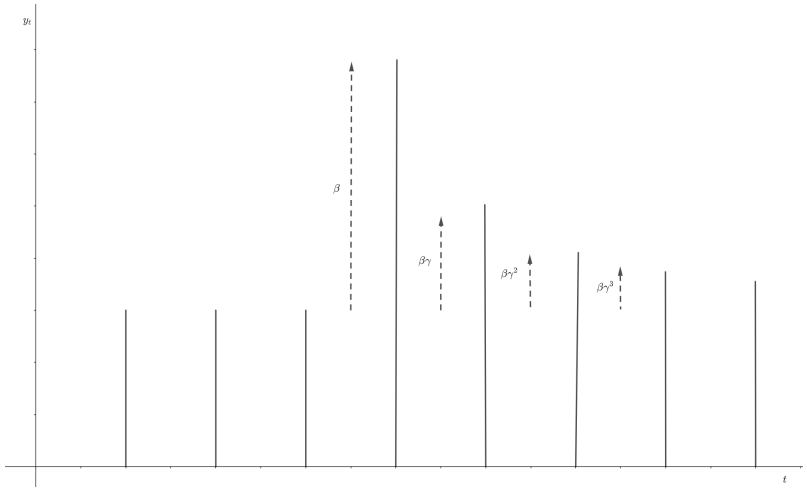
Note 3: Recursive substitution gives

$$\begin{aligned} y_t &= \alpha + \beta x_t + \gamma(\alpha + \beta x_{t-1} + \gamma y_{t-2} + e_{t-1}) + e_t \\ &= \alpha\left(1 + \gamma + \gamma^2 + \cdots\right) + \beta x_t + \beta\gamma x_{t-1} + \beta\gamma^2 x_{t-2} + \cdots + e_t + \gamma e_{t-1} + \gamma^2 e_{t-2} + \cdots \\ &= a + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \varepsilon_t \end{aligned}$$

This is a distributed lag model where the number of lags $K$ is infinite, with $a = \alpha(1 + \gamma + \gamma^2 + \cdots)$, $\varepsilon_t = e_t + \gamma e_{t-1} + \gamma^2 e_{t-2} + \cdots$, and each lag coefficient $\beta_k = \beta\gamma^k$ for $k = 0, 1, 2...$

If $e_t$ is not autocorrelated, then $\varepsilon_t$ will be autocorrelated. Flipping this, if we assume $\varepsilon_t$ is not autocorrelated, then the original $e_t$ must have a specific autocorrelation structure.

Having $\beta_k = \beta\gamma^k$ for $k = 0, 1, 2...$ means that the total response time $K = \infty$, the short run response is $\beta$, the long run (total) response is $\sum_{s=0}^{\infty} \beta\gamma^s = \beta/(1-\gamma)$, and the mean lag is $\gamma/(1-\gamma)$. The closer $\gamma$ is one, the larger is the mean lag and the total response. The impulse response is a geometric decay:




## 2.3 The Stock Adjustment model

A variety of economic behavioral models can generate lagged dependendent variable regressions. One example is the "stock adjustment" model.

Suppose in each time period $t$ we have desired target level of the outcome $y_t$. For example, $y_t$ could be firm's inventory (stock) of some good. Call this target level $y_t^*$. Assume a regression model for $y_t^*$:

$$y_t^* = \alpha + \beta x_t + e_t$$

Suppose changes in $y$ from one period to the next are costly. To reduce these costs, instead of changing $y$ all the way from $y_{t-1}$ to $y_t^*$, we just set $y_t$ using the rule:

$$y_t - y_{t-1} = (1-\gamma)(y_t^* - y_{t-1})$$

So, e.g., suppose $\gamma = 1/2$ and $y_t^* - y_{t-1} = 10$. That means you'd like to increase $y_{t-1}$ by 10 units. But to reduce these costs of changing, you'd instead only increase by $(1-\gamma)*10$, which is 5 units. The more costly changing the level of $y$ is, the closer you set $\gamma$ to one, and hence the less you change $y$ by each period.

We don't observe $y_t^*$, but if you just substitute the first equation for $y_t^*$ into the second, and add $y_{t-1}$ to both sides, you get the lagged dependent variable model

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + e_t.$$

Another model that can give rise to the lagged dependent variable model is the "adaptive expectations," model, also sometimes called a habit model. In this model, you don't fully adjust $y$ to a change in $x$ because of habits. For example, if $x$ is a price and $y$ is how much you usually purchase of a good, you don't immediately fully adjust your purchase demand in response to a price change, but instead adjust slowly over time. The larger $\gamma$ is, the stronger is the purchasing habit, and hence the slower is the adjustment.

## 2.4 More regressors

One can include in the model lags of $x$ in addition to a lagged $y$. For example, we could have the model

$$y_t = \alpha + \beta x_t + \delta x_{t-1} + \gamma y_{t-1} + e_t.$$

This will now be equivalent to a distributed lag model where the coefficient of $x_{t-1}$ can be anything, but then the coefficients of $x_{t-2}$, $x_{t-3}$, etc. decline geometrically. Once again, multicollinearity could become a problem.

We can also consider a model with more regressors, like

$$y_t = \alpha + \beta x_t + \delta z_t + \gamma y_{t-1} + e_t.$$

This will now be equivalent to a distributed lag model where both $x_t$ and $z_t$ have infinite length distributed lags that decline geometrically, and both decline at the same rate (determined by $\gamma$).

## 2.5 The Lag and Difference Operators

- The Lag Operator $L$ is defined by:

$$\begin{aligned} LX_t &= X_{t-1} \\ L^2X_t &= LLX_t = LX_{t-1} = X_{t-2} \\ L^kX_t &= X_{t-k} \end{aligned}$$

- The Difference Operator $\Delta$ is defined by

$$\begin{aligned} \Delta X_t &= (1-L)X_t = X_t - X_{t-1} \\ \Delta^2 X_t &= (1-L)^2 X_t = (1-2L+L^2)X_t \\ &= X_t - 2X_{t-1} + X_{t-2} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \end{aligned}$$

- We can write the distributed lag model

$$y_t = a + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_K x_{t-k} + e_t$$

using the lag operator, as

$$y_t = a + B(L) x_t + e_t$$

where $B(L)$ is the $K$'th order polynomial function

$$B(L) = \beta_0 + \beta_1 L + \cdots + \beta_K L^K$$

- In many ways, we can treat $L$ as if it were a variable. For example, the long run response is obtained by setting $L = 1$, giving $B(1) = \beta_0 + \beta_1 + \cdots + \beta_K$.

- Another example: Let $B'(L) = \partial B(L)/\partial L$. Then $B'(L) = \beta_1 + 2\beta_2 L \cdots + (K-1)\beta_K L^{K-1}$, and the formula we gave earlier for the mean lag is just $B'(1)/B(1)$.

- The lag and difference operators allow us to readily calculate things that we before derived by repeated substitution. For example consider the lagged dependent variable model

$$Y_t = a + bX_t + \gamma Y_{t-1} + e_t$$

Instead of repeatly substituting in for $Y_{t-1}$, then $Y_{t-2}$, etc, to obtain the equivalent distributed lag representation of this model, we can use the lag operator. Rewrite the model as $(1 - \gamma L)Y_t = bX_t + e_t$. Again treating $L$ as if it was a variable, we have

$$\begin{aligned}
(1-\gamma L)Y_t &= bX_t + e_t \\
Y_t &= \frac{b}{1-\gamma L}X_t + \frac{1}{1-\gamma L}e_t \\
\text{where } \frac{b}{1-\gamma L} &= b(1 + \gamma L + (\gamma L)^2 + ...) \\
\text{so } \frac{b}{1-\gamma L}X_t &= bX_t + b\gamma X_{t-1} + b\gamma^2 X_{t-2} + ...
\end{aligned}$$

## 2.6 Nonstationary data, Integrated Processes, and Cointegration

A non-stationary random series $Z_t$ is said to be "integrated of order one," denoted I(1), if $\Delta Z_t$ is stationary. Similarly, $Z_t$ is "integrated of order k," I(k), if $\Delta^k Z_t$ is stationary. Example: suppose $Z_t$ is given by the random walk model $Z_t = Z_{t-1} + V_t$, where $V_t$ is iid. Then $Z_t$ is an I(1) series. More generally, if $V_t$ is stationary then $Z_t = Z_{t-1} + V_t$ is I(1).

A nonstationary series $Z_t$ will generally grow over time, and/or have a variance that grows over time. For example, the random walk has $E(Z_t) = E(Z_{t-1}) + E(V_t)$, and so will be trending up or down if $E(V_t) \neq 0$, and $var(Z_t) = var(V_t) + var(Z_{t-1})$ and so has a growing variance if $V_t$ is random rather than constant.

Suppose we have a regression model $Y_t = a + bX_t + e_t$, where $Y_t$ is non-stationary. Then either $X_t$, or $e_t$, or both must also be non-stationary. Recall

$$\hat{b} = b + \frac{\frac{1}{T}\sum(x_t - \bar{x})e_t}{\frac{1}{T}\sum(x_t - \bar{x})^2}$$

If $X_t$ is nonstationary, then the denominator (which equals the sample variance of $X$) will be growing as the sample size grows. If at the same time $e_t$ is stationary, then the denominator will tend to grow quickly relative to the numerator, making $\hat{b}$ converge to $b$ at a rate faster than root-n. In this case we say that $X_t$ and $Y_t$ are "cointegrated." Our usual CLT for $\hat{b}$, and hence our usual standard error formulas, do not apply. We saw an example like this back in lecture 5, where $x_t = t^{1/2}$.

But suppose $e_t$ is not stationary. Then the asymptotic behavior of $\hat{b}$ might be very bad, possibly inconsistent, depending on how $e_t$ evolves over time compared to how $X_t$ evolves. In practice, it's hard to know if $e_t$ is stationary or not, so hard to know if we are in the fast converging case of cointegration, or in one of these possibly inconsistent situations. Also, spurious correlation is commong with nonstationary data. For example, any two time series that are both trending upward will have a positive sample correlation, even if they're unrelated.

Often, to deal with this issue, we difference data before running the regression. E.g., if $X_t$ and $Y_t$ are both $I(1)$, then we may run the regression $\Delta Y_t = a + b\Delta X_t + \varepsilon_t$. This now can satisfy our usual root n asymptotics. We may have sacrificed precision if $Y$ and $X$ had been cointegrated, but we ensure ourselves against problems if $e_t$ had been non-stationary.

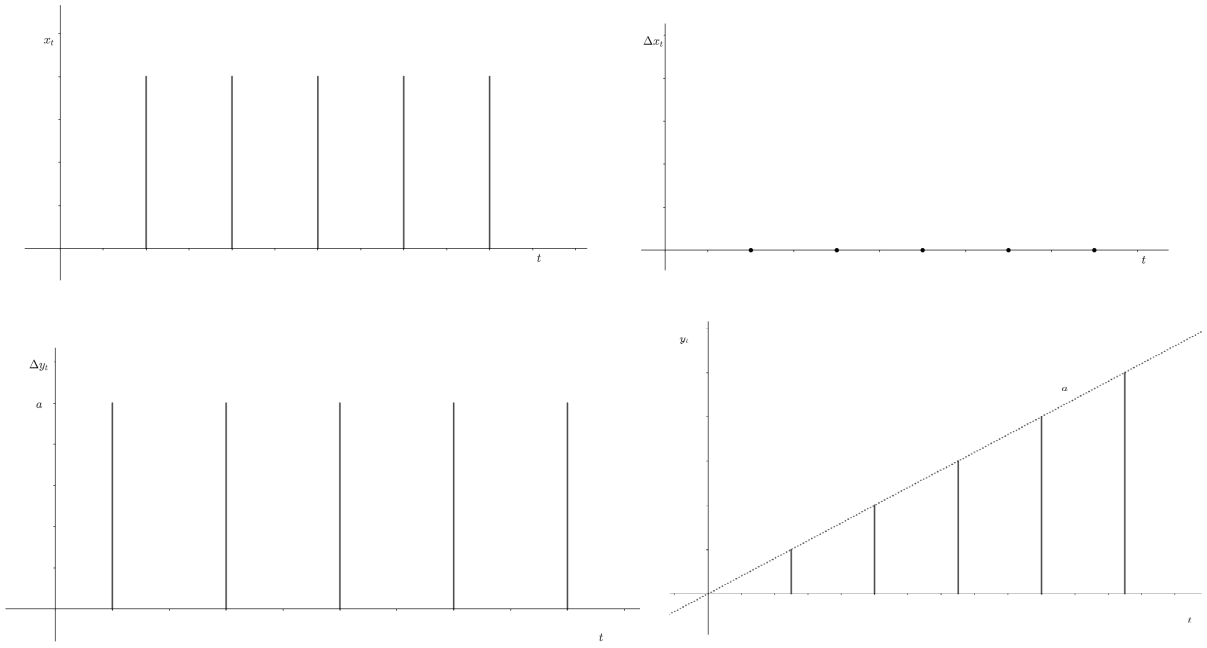
## 2.7 Interpreting a regression with differenced data

If we run the regression

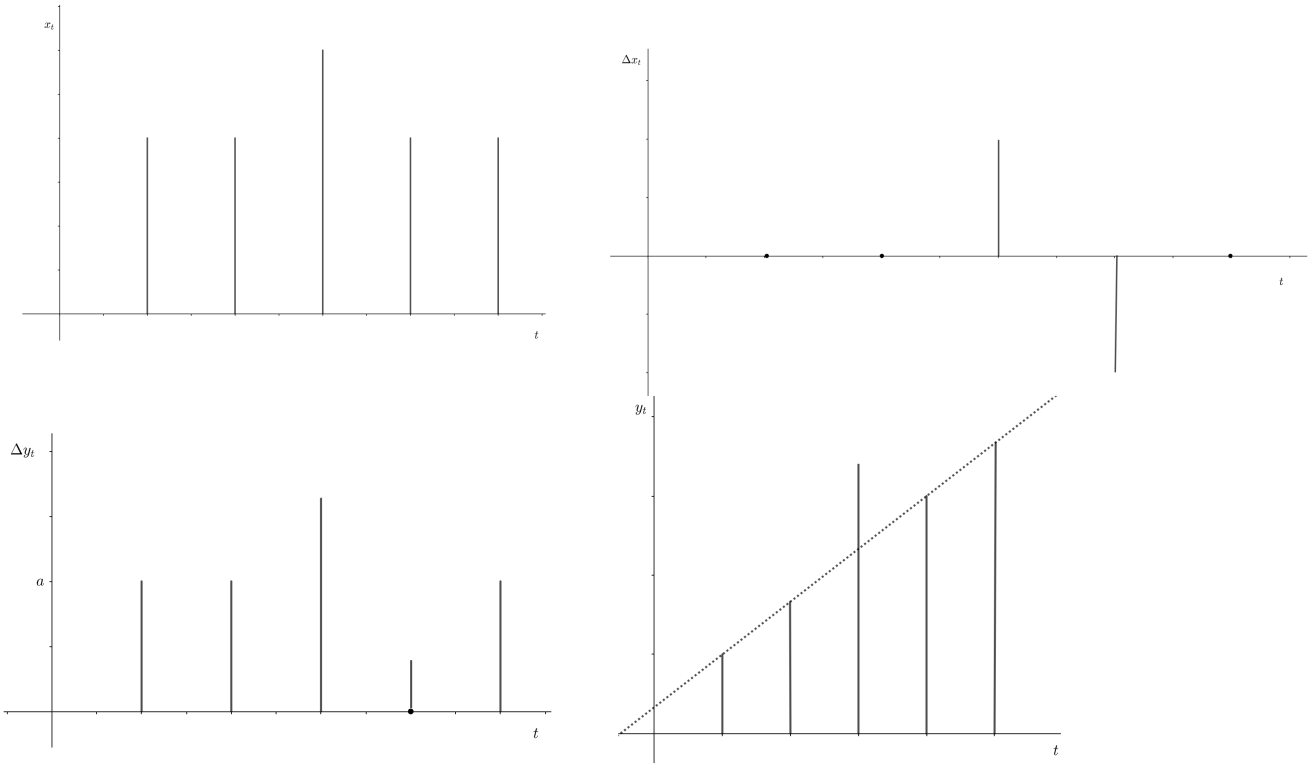
$$\Delta Y_t = a + b\Delta X_t + \varepsilon_t$$

what do the coefficients mean? Consider the impulse response function. Ignore $\varepsilon_t$, and look at what happens to $\Delta Y_t$, and then to $Y_t$, if we apply an impulse to $X_t$.

First look at no impulse. If for some constant $x^*$ we had $X_s = x^*$ for all time periods $s$, then $\Delta X_s = 0$, so $\Delta Y_s = a$, and $Y_s = Y_0 + as$ for all time periods $s$. Here's what $\Delta X$, $X$, $\Delta Y$, and $Y$, look like:



Now consider an impulse in time period $t$. We now let $X_s = x^*$ for all $s \neq t$, and $X_t = x^* + 1$. Then $\Delta X_t = 1$, but notice that $\Delta X_{t+1} = -1$, and then $\Delta X_s = 0$ for all $s \geq t + 2$. The impulse response on $\Delta Y$ is then $\Delta Y_t = a + b$, $\Delta Y_{t+1} = a - b$, and $\Delta Y_s = a$ for all $s \geq t + 2$. The resulting response on $Y$ is $Y_t = Y_0 + at + b$, while $Y_s = Y_0 + as$ for all $s \neq t$. Now $\Delta X$, $X$, $\Delta Y$, and $Y$ look like this:




The end result is that $b$ has the same interpretation in the differenced model as in an undifferenced model: it is the change in $Y_t$ resulting from a unit change $X_t$. But now, even if $X_t$ were held fixed over time, $Y_t$ would be trending upward, increasing by $a$ every time period (or trending down if $a$ is negative). So $b$ is the size of the departure of $Y_t$ from that trend line in period $t$ (and only period $t$), if $X$ was increased by one only in period $t$.

# 3 Lecture 10. IV and 2SLS Estimation, Endogeneity and Simultaneity

Readings for this lecture are: Greene Chapter 8, and 10.

## 3.1 Endogeneity

A key linear regression assumption is $cov(X, e) = 0$. A regressor $X$ is 'endogenous' if $cov(X, e) \neq 0$ (we'll give a more formal definition of endogeneity later). A regressor can only be endogenous if it's random, not a fixed constant, because the covariance of a constant with any random variable is zero.

If a regressor is endogenous, then OLS is inconsistent: For the model, $Y_i = a + bX_i + e_i$, we have the OLS estimate

$$\begin{aligned} plim\left(\hat{b}\right) &= b + \frac{plim\frac{1}{n}\sum_i \left(X_i - \overline{X}\right) e_i}{plim\frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2} \\ &= b + \frac{cov(X, e)}{var(X)} \neq b \text{ if } cov(X, e) \neq 0 \end{aligned}$$

Why might $cov(X, e) \neq 0$?

1. Model misspecified: E.g., omitted variable bias. If the left out variable is correlated with included regressors, then get $cov(X, e) \neq 0$.
2. Simultaneity: If the value of $X$ partly depends on $Y$, then $X$ partly depends on $e$, making it correlated with $e$.
3. Measurement error in $X$.

## 3.2 Measurement Error in a Regressor

Why should measurement error in $X$ make $cov(X, e) \neq 0$?

Suppose we have a linear regression model $Y_i = bX_i^* + U_i$, which satisfies GM with $E(U) = 0$ and $E(UX^*) = 0$. However, $X_i^*$ is not observed perfectly. Instead of observing $X^*$ we observe $X$ where

$$X_i = X_i^* + V_i$$

Here $V$ is the measurement error. The classical measurement error assumptions are $E(V) = 0$, $E(VU) = 0$ and $E(VX^*) = 0$. This is a best case scenario, where the measurement error is mean zero and is uncorrelated both with the the model error $U$ and with true value of the regressor $X^*$.

What happens if we do OLS using the observed, mismeasured $X$ instead of the true $X^*$?

$$\begin{aligned} Y_i &= bX_i^* + U_i \\ &= b(X_i - V_i) + U_i \\ &= bX_i + (U_i - bV_i) \\ &= bX_i + e_i \end{aligned}$$

So if we regressed $Y$ on $X$, the regression error $e$ would be given by $e = U_i - bV_i$. Note $E(e) = E(U_i - bV_i) = E(U_i) - bE(V_i) = 0$. However, GM can't hold because the regressor $X$ is a random variable. Even if the true $X^*$ is fixed, the observed $X$ is still random because it depends $V$, and measurement error $V$ is random. More seriously,

$$\begin{aligned} cov(e, X) &= cov[(U - bV), (X^* + V)] \\ &= cov(U, X^*) + cov(U, V) - bcov(V, X^*) - bvar(V) \\ &= -bvar(V) \neq 0 \end{aligned}$$

So OLS is inconsistent. How bad is the inconsistency?

$$\begin{aligned} plim\left(\hat{b}\right) &= b + \frac{-bvar(V)}{var(X)} \\ &= b - b\frac{var(V)}{var(X^* + V)} \text{ given } EX = 0 \text{ in demeaned data} \\ &= b\left(1 - \frac{var(V)}{var(V) + var(X^*)}\right) \\ &= b\frac{var(X^*)}{var(V) + var(X^*)} \end{aligned}$$

As long as there is any random measurement error, so $var(V) > 0$, then $plim\left(\hat{b}\right)$ will be closer to zero than $b$. Informally, people say that this means $\hat{b}$ is "biased towards zero," or "downward biased," though technically this difference between $plim\left(\hat{b}\right)$ and $b$ is inconsistency, not bias. This inconsistency of the coefficient, due to measurement error in the regressor, is called "attenuation bias." The ratio $var(X^*)/var(V)$ is called the signal to noise ratio. The larger is the signal to noise ratio, the smaller is the attenuation bias.

If we have more regressors, e.g., $Y_i = bX_i^* + cZ_i + U_i$, and we replace $X_i^*$ with $X$, then $\hat{b}$ will still be biased downward (though the formula for the bias becomes more complicated), and $\widehat{c}$ will also be biased (unless $X_i^*$ and $Z_i$ are uncorrelated), though direction of the bias will depend on how $X_i^*$ and $Z_i$ are correlated.

If more than one regressor is measured with error, then it is possible for the mismeasured regressor coefficients to be biased upwards instead of downwards, but that is unusal. Typically, measurement error biases coefficients towards zero.

## 3.3 Instrumental variables (IVs)

### 3.3.1 Basic idea of IV

We gave three reasons why we might get $cov(X, e) \neq 0$ (misspecification, simultaneity, or measurement error in a regressor). For now, let's ignore why $cov(X, e) \neq 0$, and just consider what happens when $X$ is correlated with $e$ for any reason. When that happens, OLS is inconsistent, and typically biased as well. So, if we want a consistent estimator, we will need to do something other than OLS.

Let's assume for now that $Y$ and $X$ both have mean zero, so we can ignore constant terms and just think about the coefficient $b$ in a regression $Y = Xb + e$ for scalar RVs $Y$ and $X$. One way to think about why OLS is consistent when $cov(X, e) \neq 0$ is to multiply both sides by $X$ and take expectations:

$$\begin{aligned} Y &= Xb + e \\ XY &= X^2 b + Xe \\ E(XY) &= E\left(X^2\right) b + E(Xe) \end{aligned}$$

Then, if we had $E(Xe) = 0$, we'd get $b = E(XY)/E(X^2)$, and OLS just replaces these expectations with sample averages. This procedure breaks down if $E(Xe) \neq 0$, that is, if $X$ is correlated with $e$.

But what if we could find some other variable $Z$ that
1. is correlated with $X$, so $cov(Z, X) \neq 0$, and
2. is uncorrelated with $e$, so $cov(Z, e) = 0$.

If we could find such a $Z$, then instead of multiplying both sides of the regression by $X$, we could multiply both sides by $Z$ and take expectations:

$$\begin{aligned} Y &= Xb + e \\ ZY &= ZXb + Ze \\ E(ZY) &= E(ZX) b + E(Ze) = E(ZX) b \end{aligned}$$

So $b = E(ZY)/E(ZX)$. The corresponding estimator that replaces these expectations with sample average is called the Instrumental Variables (or IV) estimator, and $Z$ is called the instrument:

$$\widehat{b}_{IV} = \frac{\sum_i Z_i Y_i}{\sum_i Z_i X_i}$$

Plugging $Y_i = X_i b + e_i$ into this expression and taking plim's shows that $\widehat{b}_{IV}$ is consistent.

Notes:
1. IV is consistent, but it's not in general unbiased. So it's mainly useful in large samples.
2. OLS is a special case of IV, where the instrument $Z$ is the regressor $X$.
3. A variable $Z$ is said to be a 'valid' instrument if $cov(Z, e) = 0$ and $cov(Z, X) \neq 0$.
4. Finding valid instruments is hard! We'll discuss how to find valid instruments later.

Terminology note: there's a literature on what is known as 'local average treatment effect' estimation, or LATE estimation, that also uses instruments and the above $\widehat{b}_{IV}$ formula. But that literature uses different assumptions, and the definition of a valid instrument in that literature is different from what calling a valid instrument. Don't mix them up! whenever you read anything about instrument validity be sure to know which literature, and hence which definition, is being used!

### 3.3.2 IV estimation in matrix form

The model is $Y = X\beta + e$. Recall $X$ is an $n \times K$ matrix, where $K$ is the number of regressors (including the constant term, i.e., the first column of $X$, the vector $X_1$, is an $n$ vector of ones).

For IV estimation, we need an $n \times K$ matrix $Z$ of instruments. We basically want an instrument $Z_{ki}$ for each regressor $X_{ki}$, for $k = 1, ..., K$. If, for a given variable $k$, $E(X_{ki}e_i) = 0$ then we can let $Z_{ki} = X_{ki}$ for that $k$. In short, if $X_{ki}$ is not one of the problem regressors, not one of the regressors that is correlated with the errors, then $X_{ki}$ can be the instrument for itself. An example is that we can let $Z_{1i} = X_{1i} = 1$, i.e., the constant is an instrument for itself.

For every variable $k$ that is a problem, meaning $E(X_{ki}e_i) \neq 0$, we need to find a valid instrument; a $Z_{ki}$ that has $cov(Z_{ki}, X_{ki}) \neq 0$ and $E(Z_{ki}e_i) = 0$.

Actually, what we really need is an asymptotic, matrix version of these conditions: that $plim\frac{Z'X}{n}$ exists and is nonsingular, and that $plim\frac{Z'e}{n} = 0$. Note that since $(Z'X)/n$ and $(Z'e)/n$ are averages, this means we're going to want a LLN to hold.

The IV estimator is then

$$\widehat{\beta}_{IV} = (Z'X)^{-1}(Z'Y)$$

To see that $\widehat{\beta}_{IV}$ is consistent, we have

$$\begin{aligned}
\widehat{\beta}_{IV} &= (Z'X)^{-1}(Z'(X\beta + e)) \\
&= \beta + (Z'X)^{-1}(Z'e) \\
&= \beta + \left(\frac{Z'X}{n}\right)^{-1}\left(\frac{Z'e}{n}\right).
\end{aligned}$$

So, using the theorem that the plim of a (smooth) function equals the function of the plims, we get:

$$\begin{aligned}
plim\left(\widehat{\beta}_{IV}\right) &= \beta + \left(plim\frac{Z'X}{n}\right)^{-1}\left(plim\frac{Z'e}{n}\right) \\
&= \beta
\end{aligned}$$

### 3.3.3 The Limiting Distribution of IV

Since

$$\widehat{\beta}_{IV} = \beta + \left(\frac{Z'X}{n}\right)^{-1}\left(\frac{Z'e}{n}\right)$$

we have

$$\sqrt{n}(\widehat{\beta}_{IV} - \beta) = \underbrace{\left(\frac{Z'X}{n}\right)^{-1}}_{\text{Term 1}}\underbrace{\sqrt{n}\left(\frac{Z'e}{n}\right)}_{\text{Term 2}}$$

Let $\Sigma_{ZX} = plim\frac{Z'X}{n}$, which we assumed exists and is nonsingular. Then the probability limit of term 1 is $\Sigma_{ZX}^{-1}$.

We now need to make another assumption: that $\sqrt{n}\frac{Z'e}{n}$ satisfies a CLT. Let us also assume for now that we don't have any other regression problems, so assume that $E(ee' \mid Z) = E(ee') = \sigma^2 I_n$. Finally, assume $\frac{Z'Z}{n}$ satisfies a LLN, so let $\Sigma_{ZZ} = plim\frac{Z'Z}{n}$, which exists, is finite and equals $E\left(\frac{Z'Z}{n}\right)$.

Making these assumptions, we need to calculate the variance of term 2:

$$\begin{aligned} var\left(\sqrt{n}\left(\frac{Z'e}{n}\right)\right) &= E\left[\left(\frac{\sqrt{n}}{n}\sum \mathbf{z}_i e_i\right)\left(\frac{\sqrt{n}}{n}\sum \mathbf{z}_i e_i\right)'\right] \\ &= E\left[\frac{1}{n}\sum_i\sum_j \mathbf{z}_i\mathbf{z}_j E[e_i e_j \mid Z]\right] \\ &= E\left[\frac{Z'Z}{n}\sigma^2\right] = \sigma^2\mathbf{\Sigma}_{ZZ} \end{aligned}$$

And so, applying a CLT, we get for term 2.

$$\sqrt{n}\left(\frac{Z'e}{n}\right) \xrightarrow{d} N(0, \sigma^2\mathbf{\Sigma}_{ZZ})$$

Finally, we can apply our rule about the limiting distibution of a product of two terms, where one term has a plim and the other converges in distribution, to get

$$\sqrt{n}(\widehat{\beta}_{IV} - \beta) \xrightarrow{d} N\left(0, \sigma^2\mathbf{\Sigma}_{ZX}^{-1}\mathbf{\Sigma}_{ZZ}\mathbf{\Sigma}_{ZX}^{-1}\right)$$

We can then use this root-n CAN convergence in distribution to write the asymptotic approximation for the distribution of $\widehat{\beta}_{IV}$ itself as

$$\widehat{\beta}_{IV} \overset{a}{\sim} \left(\beta, \frac{s^2}{n}\left(\frac{Z'X}{n}\right)^{-1}\left(\frac{Z'Z}{n}\right)\left(\frac{Z'X}{n}\right)^{-1}\right)$$

The standard errors of each element of $\widehat{\beta}_{IV}$ are then equal to the square root of the diagonal elements of the above variance matrix, which we can calculate from data.

## 3.4 Two Stage Least Squares

Return to the single regressor case, with all variables having zero means, so $Y_i = bX_i + e_i$.

What if we have two valid instruments, say $Q_i$ and $R_i$? (Note more common is the opposite problem - it can be hard to find even one valid instrument).

We could use each, and get two different IV estimates

$$\hat{b} = \frac{\sum_i Q_i Y_i}{\sum_i Q_i X_i} \text{ and } \check{b} = \frac{\sum_i R_i Y_i}{\sum_i R_i X_i}$$

each of which is consistent and asymptotically normal. In fact, any linear combination $Z_i = a_0 + a_1 Q_i + a_2 R_i$ will also be a valid instrument. We could do IV using $Z_i$ as the instrument for any choice of the constants $a_0$, $a_1$, and $a_2$. What choice is best?

$$\begin{aligned} \hat{b}_{IV} &= \frac{\sum_i Z_i Y_i}{\sum_i Z_i X_i} \\ &= b + \frac{\sum_i Z_i e_i}{\sum_i Z_i X_i} \end{aligned}$$

so we want to make either: (1) $\sum_i Z_i e_i$ as small as possible or (2) $\sum_i Z_i X_i$ as big as possible. We can't do option 1 because we don't know $e_i$. What about option (2)?

We could make the denominator big by just replacing $Z_i$ with $1000Z_i$, but that doesn't help, because then the numerator would be increased by the same amount. What we really need is to choose $a_0$, $a_1$, and $a_2$ to make the correlation between $Z$ and $X$ as large as possible.

Another way to think about it: if it were true that $X$ and $e$ were uncorrelated, we would do OLS, and that would be efficient. And OLS is the special case of IV where $Z = X$. So to make the IV estimator as efficient as possible, we want to choose a $Z$ that's as close to equalling $X$ as possible, while still being a valid instrument. This means we want to choose $a_0$, $a_1$, and $a_2$ to make $a_0 + a_1 Q_i + a_2 R_i$ be as close to $X_i$ as possible.

How do we do that? By using OLS! Run the regression $X_i = a_0 + a_1 Q_i + a_2 R_i + error_i$, and use the fitted values as the instrument $Z_i$.

The resulting estimator is called 2SLS or TSLS, which stands for Two Stage Least Squares. The 2SLS estimator is:

1. Using OLS, regress $X$ on 1, $Q$, and $R_i$. Let $\widehat{X}_i$ be the fitted values from this regression
2. Do IV, using the instrument $Z_i = \widehat{X}_i$, so

$$\hat{b}_{IV} = \frac{\sum_i \widehat{X}_i Y_i}{\sum_i \widehat{X}_i X_i}$$

Note: Some textbooks say the second stage is to regress $Y_i$ on $\widehat{X}_i$ using OLS, that is,

$$\hat{b}_{IV} = \frac{\sum_i \widehat{X}_i Y_i}{\sum_i \widehat{X}_i \widehat{X}_i}$$

It turns out that both of these are numerically the same, that is, $\sum_i \widehat{X}_i X_i = \sum_i \widehat{X}_i \widehat{X}_i$, but this is just a mathematical trick. You should NOT think of 2SLS this way. Think of the second stage as being IV. This is because:

1. The first stage is really creating a valid instrument, not a valid regressors, and,
2. To get the standard errors correct, you need to apply the IV standard error formula to the second stage, not the OLS standard error formula

Question: why does the IV standard error formula still work, even though $Z$ is now estimated? Why don't we need to account for the estimation error in the first stage $a_0$, $a_1$, and $a_2$ coefficients? Intuitively, it's because $Z_i$ is a valid instrument for whatever numbers we choose for $a_0$, $a_1$, and $a_2$. In this case, we just happen to be choosing them by estimating them. We're then conditioning on that choice in the second stage.

## 3.5 2SLS in Matrix Form

The model is $Y = X\beta + e$. Recall $X$ is an $n \times K$ matrix, where $K$ is the number of regressors. We now assume we have $L$ instruments, where $L \geq K$. Let $Q$ be an $n \times L$ matrix, where the columns of $Q$ are these $L$ instruments.

The first stage of 2SLS is regressing each column of $X$, i.e., each of the regressors, on all of the instruments, and getting the fitted values. This is just: $Z = \widehat{X} = Q\left(Q'Q\right)^{-1}Q'X$

To see what's going on, let $W$ be one of the regressors in the model, so $W$ is one of the columns of $X$. Then you recognize $\left(Q'Q\right)^{-1}Q'W$ as just the coefficients of regressing $W$ on $Q$, and $Q\left(Q'Q\right)^{-1}Q'W$ would then be the column vector of fitted values $\widehat{W}$.

When we write $Z = \widehat{X} = Q\left(Q'Q\right)^{-1}Q'X$, we're doing this same regression on every column of $X$, and so we end up with the matrix of fitted values $\widehat{X}$.

Once we have this $Z$, then we do IV. So the 2SLS estimator is

$$\begin{aligned} Z &= Q\left(Q'Q\right)^{-1}Q'X \\ \widehat{\beta}_{2SLS} &= (Z'X)^{-1}(Z'Y) \end{aligned}$$

With the same limiting distribution as $\widehat{\beta}_{IV}$, using this $Z$ as instruments.

Notes:Recall that we need to find at least one instrument for each model regressor (i.e., for each column of $X$).

But when we run 2SLS, it doesn't matter which instrument goes with which model regressor. The first stage regresses every column of $X$ on all of the instruments $Q$, to get a fitted value. So we never need to actually assign each instrument to a regressor. We just need enough instruments in total, and together they need to explain enough of the variation in $X$ so that $X'Z$ is nonsingular.

For validity we also want $Z$ and $e$ to be uncorrelated. If any columns of $X$ are is uncorrelated with $e$, we can just include those columns in $Q$, and we then only need to find valid instruments for the other columns.

Let $S$ be one of the columns (i.e. one of the model regressors) that is uncorrelated with $e$, and so is included in $Q$. When we run the first stage regression of $X$ on $Q$, that will include regressing $S$ on $Q$. And if $Q$ contains $S$, then the fitted value will just equal $S$ (e.g., if you ran the regression $w = a + bw + cq + e$, the fitted values would just be $b = 1$, $a = c = 0$, because that makes $e = 0$ and thereby minimizes the sum of squared errors).

In short, when some regressor $S$ is the instrument for itself, what the first stage of 2SLS will do is include that $S$ in the $Z$ matrix, like it should.

## 3.6 Testing in IV regression

### 3.6.1 Hausman Test

In $Y = X\beta + e$, how do we know if $E\left(Xe\right) \neq 0$? Usually, it's because economic or statistical theory tells us so. For example, we know it happens when regressors are mismeasured.

Can we test if $E\left(Xe\right) \neq 0$? Sometimes. Suppose we have instruments $Q$, and we know they're valid, so $E\left(Qe\right) = 0$. Suppose the number of instruments $L$ is bigger than the number of regressors $K$. Then we could do 2SLS, but, if it were true that $E\left(Xe\right) = 0$, then it would be better (more efficient) to do OLS. In this situation, it's possible to test if $E\left(Xe\right) = 0$.

Here's one possible test: Let $\widetilde{X}$ be the matrix consisting of the columns of $X$ that we think might be correlated with $e$. The null hypothesis is $H_0$: $E\left(\widetilde{X}e\right) = 0$.

Step 1: Let $\widetilde{Z} = \widehat{\widetilde{X}} = Q\left(Q'Q\right)^{-1}Q'\widetilde{X}$. This is the first stage of 2SLS, but only for the variables in $\widetilde{X}$. So $\widetilde{Z}$ is the matrix of first stage estimated instruments for $\widetilde{X}$.

Step 2: Using OLS, estimate the regression $Y = X\delta + \widetilde{Z}\gamma + error$. Test if $\gamma = 0$ (can use a standard F-test or a chi-squared test). If you reject $\gamma = 0$, then you reject $E\left(\widetilde{X}e\right) = 0$.

Why does this work? By construction, we have that $E\left(\widetilde{Z}e\right) = 0$ (if this isn't true, then $Q$ isn't a matrix of valid instruments).

Now, suppose that $H_0$ is true. Then $E\left(\widetilde{X}e\right) = 0$, and therefore $E\left(Xe\right) = 0$. This means that the regression in Step 2 will be the same as the original regression $Y = X\beta + e$, with some extra useless regressors $\widetilde{Z}$ added in (if these aren't useless regressors, then the original regression must have been misspecified). Therefore, in this case, $\widehat{\delta}$ must go to $\beta$ and $\widehat{\gamma}$ goes to zero, asymptotically.

Alternatively if $H_0$ is false, then the estimates of $\widehat{\delta}$ not converge to $\beta$, due to the correlation of $\widetilde{X} - \widetilde{Z}$ with $e$, and and therefore $\widetilde{Z}$ can have some additional explanatory power, making $\widehat{\gamma}$ asymptotically nonzero.

This test often has very low power, in part because, under the alternative, $\gamma$ might not be far from zero, and in part because $X$ and $\widetilde{Z}$ are, by construction, multicollinear. Note this test requires $L > K$ (more instruments than regressors); it can be shown to have zero power when $L = K$.

This test is an example of a general class of tests called Hausman tests. Hausman tests apply more generally to situations where you have one estimator that is consistent under both null and alternative (here that's 2SLS), and another estimator that is efficient under the null and inconsistent under the alternative (here that's OLS).

### 3.6.2 Overidentifying restriction test (Hansen-Sargan J test)

The Hausman test assumes all the instruments in $Q$ are valid. Can we test that? In particular, can we test if $E\left(Qe\right) = 0$? Yes, but again it requires $L > K$ (more instruments than regressors). Here's one way to test:

Step 1: Do 2SLS using all the instruments $Q$, and get the residuals get $\hat{e} = Y - X\hat{\beta}_{2SLS}$

Step 2: regress the residuals $\hat{e}$ on $Q$ using OLS, i.e., estimate $\delta$ in the regression $\hat{e} = Q\delta + error$. Get the $R^2$ from this regression

Step 3: If $H_0$: $E\left(Qe\right) = 0$, then $\delta = 0$. We can test if $\hat{\delta} = 0$ using the test statistic $nR^2 \overset{a}{\sim} \chi^2_{L-K}$.

The intuition is that, if $E\left(Qe\right) = 0$, then a regression of $e$ on $Q$ should have no significant coefficients. But notice that the degrees of freedom of the test is $L - K$, even though $\delta$ is a $L$ vector. This is roughly because we used up $K$ degrees of freedom in estimating $\hat{e}$.

## 3.7 Finding instruments

How do we find instruments? Mostly it depends on why $X$ and $e$ are correlated to begin with.

Example: Suppose we have a regressor $S$ that is mismeasured. Then the source of correlation between $S$ and $e$ is that both $S$ and $e$ contain the measurement error. In this case, a valid instrument is some variable that is correlated with the true $S$, but is uncorrelated with the measurement error.

To illustrate, suppose $S$ is someone's reported income. They might lie or misremenber their true income. In this case a valid instrument might be the value of their house that you look up online (but not what they say if you ask them the house value - if e.g., they exaggerated their income, they might also exaggerate their house value, and then the instrument would be correlated with the measurement error, and hence invalid).

Sometimes, lagged variables, like $X$ in some previous time period, might be a valid instrument.

Often, if some variable $R$ is a valid instrument, then functions of $R$ like $R^2$ or $\log R$ might also be valid.

We will talk later in the term about the treatment effects or causal effects literature. Instruments in that literature are often generated by random events or natural experiments. The idea is that a random event (e.g., a natural disaster) couldn't be determined by someone's behavior, and so must be uncorrelated with $e$. (But recall the earlier warning about the definition of instrument validity being somewhat different in that literature).

Some terminology: We use IV's to identify and estimate coefficients. In a model with K regressors, if we only have exactly K instruments (and so do IV estimation), we say the model is exactly identified.
If we have $L > K$ instruments (and so do 2SLS instead of IV), then we say the model is over identified.
If we don't have enough instruments, we say the model is under identified or just not identified.

In the next subsection, we will talk about the commonest reason why $X$ and $e$ may be correlated: Simultaneity. And we'll discuss how you find instruments under simultaneity.

## 3.8 Simultaneity

### 3.8.1 Structural models, reduced forms, and instruments

Given a model, we say a variable is endogenous if it's value is determined (or at least partly determined) by the model. Otherwise, it is called exogenous.

If the model $y = xb + e$ satisfies the GM assumptions, then $y$ is endogenous and $x$ is exogenous.

But suppose instead that the model is $y = xb + e \quad \text{and} \quad x = g(y, z, u)$ The function $g$ is some unknown function, that depends on $y$, on a vector $z$ of exogenous variables we observe, and vector $u$ of errors we don't observe. For now we don't care about estimating $g$.

Example, $y$ could be a person's demand for coffee, and $x$ could be their demand for milk. Each depends on how much you consume of the other.

Now both $y$ and $x$ are determined by the model, and they are both determined simultaneously (each affects the other). So now $x$ and $e$ will be correlated, because $x$ depends in part on $y$, and $y$ depends in part on $e$. Here both $y$ and $x$ are endogenous, and $z$ is exogenous.

Simultaneity of $y$ and $x$ is another, common source of the endogeneity problem $cov(x, e) \neq 0$.

Example: A Demand and Supply system. Let $Q_t$ be the quantity of a good sold in time $t$, $P_t$ be the price in time $t$, and $Y_t$ be income in time $t$. Supply and demand curves are:

$$S : Q_t = \alpha_1 + \alpha_2 P_t + \varepsilon_t$$
$$D : Q_t = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_t$$

$Q$ and $P$ are endogenous (determined by the system), $Y$ is exogenous.

These equations, which describe the behavior of economic agents, are together called a structural model. Suppose we want to estimate price effects, i.e., the slope of the supply curve $\alpha_2$ and the slope of the demand curve $\beta_2$.

For simplicity, let's demean all the variables, so we can ignore constants, and get

$$S : q_t = \alpha_2 p_t + \varepsilon_t$$
$$D : q_t = \beta_2 p_t + \beta_3 y_t + u_t$$

We can approach this like microeconomists, and ask first, what are the equilibrium prices and quantities? Solving this system of equations gives

$$p_t = \frac{\beta_3}{\alpha_2 - \beta_2} y_t + \frac{u_t - \varepsilon_t}{\alpha_2 - \beta_2}$$

$$q_t = \frac{\alpha_2 \beta_3}{\alpha_2 - \beta_2} y_t + \frac{\alpha_2 u_t - \beta_2 \varepsilon_t}{\alpha_2 - \beta_2}$$

which we can rewrite as

$$p_t = \Pi_1 y_t + v_{1t}$$
$$q_t = \Pi_2 y_t + v_{2t}$$

for constants $\Pi_1$ and $\Pi_2$, and errors $v_{1t}$ and $v_{2t}$.

These last equations express the endogenous variables $p$ and $q$ as functions of exogenous variables $y$ and errors. Equations that give the endogenous variables in terms of exogenous variables and errors is called the 'reduced form'.

Notes about the reduced form equations.

1. Reduced form equations have only exogenous variables on the right, so they can be estimated by OLS.

2. If one's main interest is just in predicting or fitting the endogenous variables, then one could just use the reduced form and not bother with identifying and estimating the structural model. Structural model estimates are needed to explain underlying behavior (like learning the supply and demand curves).

3. For the most part, computer science techniques like neural networks and machine learning, are devoted to finding reduced forms, not structural models.

Suppose we want to estimate $\alpha_2$, the slope of the supply equation. What happens if we estimate the supply equation $q_t = \alpha_2 p_t + \varepsilon_t$ using OLS?

$$\hat{\alpha} = \frac{\sum_t p_t q_t}{\sum_t p_t^2} = \alpha_2 + \frac{\frac{1}{n}\sum_t p_t \varepsilon_t}{\frac{1}{n}\sum_t p_t^2}$$

and

$$plim\left(\hat{\alpha}_2\right) = \alpha_2 + \frac{cov\left(p_t, \varepsilon_t\right)}{var\left(p_t\right)} \neq \alpha_2 \text{ as } cov\left(p_t, \varepsilon_t\right) \neq 0$$

So OLS is inconsistent. Why is $cov\left(p_t, \varepsilon_t\right) \neq 0$. Look at the reduced form price equation $p_t = \Pi_1 y_t + v_{1t}$. The error is $v_{1t} = \left(u_t - \varepsilon_t\right) / \left(\alpha_2 - \beta_2\right)$, which has $\varepsilon_t$ in it.

So we have the endogeneity problem. We need to do IV, which means we need an instrument.

Here $y_t$ might be a valid instrument, so we can do the IV estimator

$$\hat{\alpha}_2 = \frac{\sum_t y_t q_t}{\sum_t y_t p_t}.$$

Why might $y_t$ be a valid instrument?

1. It's exogenous (came from outside the system), so it's reasonable to assume that $cov\left(y_t, \varepsilon_t\right) = 0$.

2. We can tell from the reduced form $p_t = \Pi_1 y_t + v_{1t}$ that $cov\left(y_t, p_t\right) \neq 0$ as long as $\Pi_1 \neq 0$. Note $\Pi_1 = \beta_3 / \left(\alpha_2 - \beta_2\right)$ and so nonzero as long as $\beta_3 \neq 0$. Having $\beta_3 \neq 0$ ensures that $y$ doesn't drop out of the system, and so affects the equilibrium outcome.

The supply equation is exactly identified. We have $K = 1$ regressors and $L = 1$ instrument.

In a simultaneous system, a way to find possible instruments is to look for exogenous variables in the other equations in the systems.

---

Suppose we wanted to estimate the demand function coefficients $\beta_2$ and $\beta_3$. Now we have two regressors, $p_t$ and $y_t$, but we only have one possible instrument, $y_t$. The demand function is under identified

Could we use some other function of $y_t$, like maybe $y_t^2$, as another instrument? $y_t^2$ should also be uncorrelated with $\varepsilon_t$. But, the reduced form shows that no additional function of $y_t$ will have explanatory power. Consider what the first stage of 2SLS would be (note we're still dropping the constant everywhere):

$$p_t = a_1 y_t + a_2 y_t^2 + error$$

Comparing this to the reduced form

$$p_t = \Pi_1 y_t + v_{1t}$$

shows that, asymptotically $a_1$ would go to $\Pi_1$ and $a_2$ would go to zero.

This also shows that, in a linear model, for an endogenous regressor, the reduced form *is* the first stage of 2SLS. The exogenous variables are all potential instruments, and the reduced form expresses each endogenous variable in terms of all of the exogenous variables.

In our supply and demand model, the demand function is identified because we have an instrument, which was an exogenous variable that was in the supply equation and not the demand equation. The supply function was not identified, because we do not have an instrument.

Suppose we instead had the model:

$$S : q_t = \alpha_2 p_t + \alpha_3 x_t + \varepsilon_t$$
$$D : q_t = \beta_2 p_t + \beta_3 y_t + u_t$$

For some exogenous $x_t$. Now both equations are identified.
For supply: regress $q$ on $p$ and $x$, using $y$ and $x$ as instruments.
For demand: regress $q$ on $p$ and $y$, using $y$ and $x$ as instruments.

Note: suppose we had $x_t = p_{t-1}$. Maybe supply takes time to manufacture, so quantity depends in part on previous period's price. In time period $t-1$, the price $p_{t-1}$ must be endogenous. However, this price is not endogenous in period $t$, since it's already been determined in period $t-1$. This kind of variable (one determined by the system in a previous period) is called a predetermined variable. Usually, predetermined variables can be used like exogenous variables.

At the opposite extreme, suppose we had the model

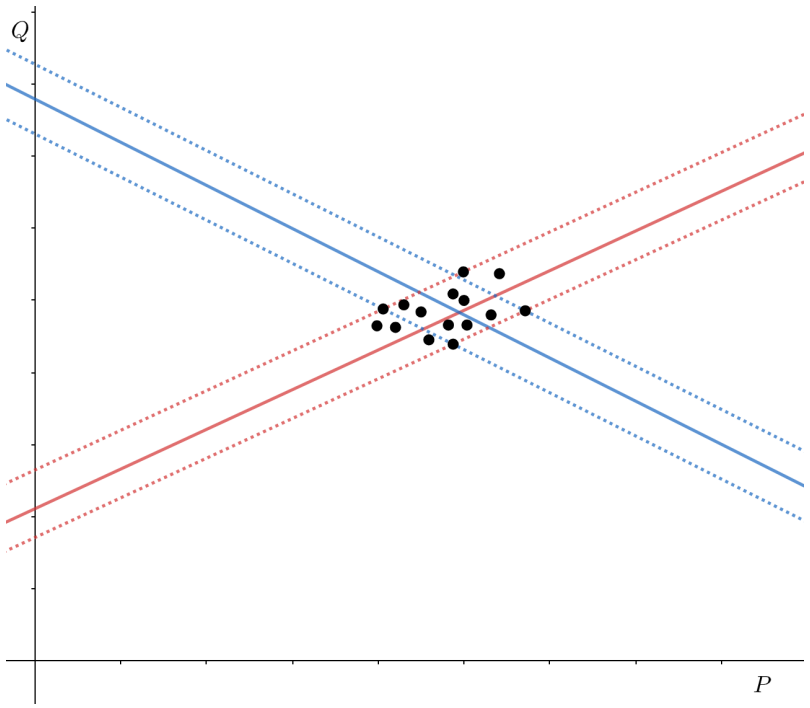
$$S : q_t = \alpha_2 p_t + \varepsilon_t$$
$$D : q_t = \beta_2 p_t + u_t$$

Now there are no instruments, so neither equation can be identified.

To see what's going graphically, let's put the constants back in. First consider the model

$$S : Q_t = \alpha_1 + \alpha_2 P_t + \varepsilon_t$$
$$D : Q_t = \beta_1 + \beta_2 P_t + u_t$$

The slope of the supply curve is $\alpha_2$, and the intercept is $\alpha_1 + \varepsilon_t$.
The slope of the demand curve is $\beta_2$, the intercept is $\beta_1 + u_t$.
It looks like this:

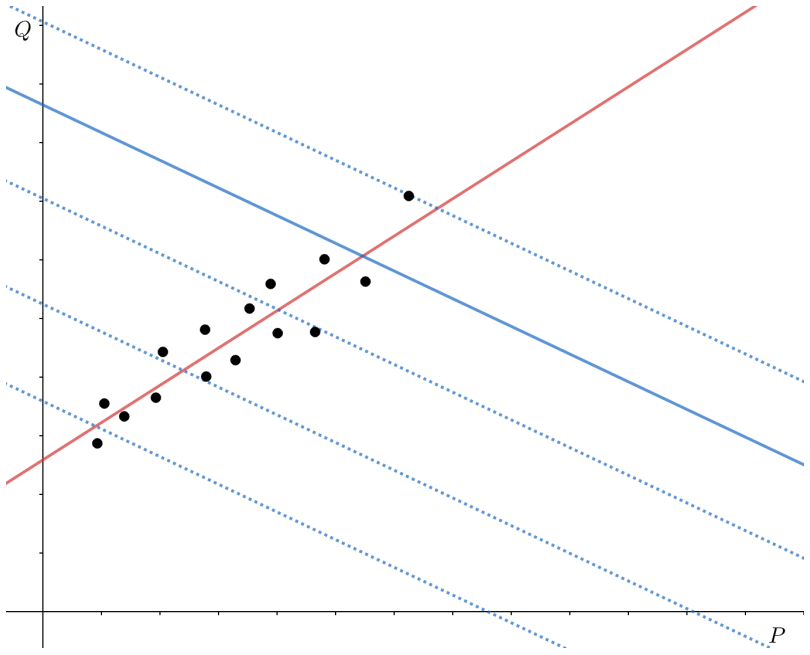



The only thing that moves the curves is the errors $\varepsilon_t$ and $u_t$, so our data just end up being a cloud of points, from which we couldn't deduce supply or demand curves. Neither is identified.

But now let's let demand depend on that exogenous variable, income.

$$S : Q_t = \alpha_1 + \alpha_2 P_t + \varepsilon_t$$
$$D : Q_t = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_t$$

The slope of the supply curve is still $\alpha_2$, and the intercept is still $\alpha_1 + \varepsilon_t$.
The slope of the demand curve is $\beta_2$, but now the intercept is $\beta_1 + \beta_3 Y_t + u_t$.
Now it looks like this:



As $Y_t$ changes, it shifts the demand curve up and down, and the intersections between demand and supply trace out the the supply curve. Exogenous variation in $Y_t$, and hence in the demand curve, allows us to identify the supply curve.

Key points about finding instruments in simultaneous systems:
1. The instrument must be an exogenous or predetermined variable.
2. Candidate instruments for endogenous regressors are non-endogenous variables that are in other equations in the system.

Note that we do *not* need to know the functional form or what all the variables are in the other equations. Go back to the original example in this section: suppose $y = xb + e$ and $x = g(y, z, u)$. To identify and estimate $b$, all we need to know about the the function $g$ is that it has the exogenous variable $z$ in it. We can therefore use $z$ as the instrument for $x$ to estimate $b$.

### 3.8.2 Seemingly unrelated regression (SUR, Zellner estimator)

At the end of the previous section, we saw that to estimate one equation in a system by IV or 2SLS, we didn't need to know the functional forms of the other equations. All we needed was to find exogenous variables in those other equations to use as instruments.

This then begs the question: Can we do better if we do know the other equations? Can we get better estimates than those from 2SLS, by using more information from other equations?

The answer is yes. To see why, let's first focus on a simpler model. What if we have a two equation system but without any simultaneity problem? Consider these two regression equations, and assume that each equation satisfies the GM assumptions

$$\begin{cases} y_{1i} = a + bx_i + u_{1i} \\ y_{2i} = c + dz_i + u_{2i} \end{cases}$$

We could estimate each by OLS, and each by itself would be BLUE.

But suppose $cov(u_{1i}, u_{2i}) \neq 0$. More specifically, assume that $E(u_{1i}u_{2j}) = 0$ if $i \neq j$ and $E(u_{1i}u_{2j}) = \sigma_{12}$ for some constant $\sigma_{12}$ if $i = j$. For example, $y_{1i}$ and $y_{2i}$ could be person $i$'s demands for two different goods, and some unobserved characteristics of person $i$ could affect both. This is now additional information, and the general rule in economics is that, if you have additional information, then you can use it to increase efficiency. The way we can use this additional information is do combine both regressions into one big regression, and apply GLS to estimate the single big regression efficiently.

Here's how to write the two separate regression as one big regression in matrix form:

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} a + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} c + \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} b + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ z_1 \\ \vdots \\ z_n \end{bmatrix} d + \begin{bmatrix} u_{11} \\ \vdots \\ u_{1n} \\ u_{21} \\ \vdots \\ u_{2n} \end{bmatrix}$$

$$\Rightarrow \mathbf{Y} = \mathbf{X}\beta + u$$

Now in this big regression we get

$$E(uu') = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} \\ \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 \end{bmatrix} \equiv \sigma^2 \Omega$$

We can therefore implement the following feasible GLS estimator:

1. do OLS and get $\hat{u}_{1i}$ and $\hat{u}_{2i}$
2. estimate $s_1^2 = \hat{\sigma}_1^2$, $s_2^2 = \hat{\sigma}_2^2$ and $s_{12} = \hat{\sigma}_{12} = \widehat{cov}(\hat{u}_1, \hat{u}_2)$
3. do GLS

This estimator is called SUR, the "Seemingly Unrelated Regressions" estimator. It's also sometimes called Zellner's estimator.

### 3.8.3 3 stage least squares (3SLS)

Now suppose we have a simultaneous system of equations, e.g., the demand and supply equations we looked at earlier. Or we could have a system of equations like this

$$\begin{cases} y_{1i} = a + bx_i + \beta y_{2i} + u_{1i} \\ y_{2i} = c + dz_i + \gamma y_{1i} + u_{2i} \end{cases}$$

This has simultaneity since $y_{1i}$ depends on $y_{2i}$ and $y_{2i}$ depends on $y_{1i}$. We could estimate each equation using IV, with instruments 1, $x$, and $z$. But as with SUR, we can get a more efficient by combining both equations. Basically 3SLS combines 2SLS with SUR.

It can be shown that, if errors are normal, then 3SLS is asymptotically equivalent to maximum likelihood estimation of this system of equations, and so is asymptotically efficient.

## 3.9 Endogeneity from AR errors and lag dependent variables

Consider the model with both a lagged dependent variable and autocorrelated errors. Suppose $x$ is exogenous.

$$y_t = a + bx_t + \gamma y_{t-1} + e_t$$

$$cov(e_t, e_{t-1}) \neq 0$$

Now the regressor $y_{t-1}$ most likely correlates with $e_t$. This is because, by the model $y_{t-1} = a + bx_{t-1} + \gamma y_{t-2} + e_{t-1}$, so $y_{t-1}$ depends on $e_{t-1}$, and $e_{t-1}$ correlates with $e_t$.

But we have an instrument. The equation for $y_{t-1}$ depends on $x_{t-1}$, so we can use $x_{t-1}$ as an instrument for $y_{t-1}$. Further lags like $x_{t-2}$ can also be used as instruments, since e.g., $x_{t-2}$ correlated with $y_{t-2}$, which correlates with $y_{t-1}$.

In time series data, lagged values of exogenous regressors will often be useful as instruments.

# 4 Lecture 11. Nonlinear Models, Extremum Estimators, and GMM

Readings for this lecture are: Greene Chapter 7, 12, and 13, and the Newey-McFadden handbook chapter.

## 4.1 Nonlinear Least Square (NLS)

### 4.1.1 Model

Consider a nonlinear regression model

$$y_i = g(x_i, \theta) + e_i$$

Here $g$ is a known function, and $\theta$ is a vector of unknown parameters. Assume that $E(e \mid x) = 0$. Let $\theta_0$ denote the true value of $\theta$, and $\Theta$ is the set of all possible values of $\theta$.

With linear regression, the OLS estimator minimized the sum of squared residuals. Consider using the same least square criteria for the above nonlinear model. This is called nonlinear least squares (NLS). The NLS estimator is

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n \quad \text{where } Q_n = \frac{1}{n} \sum\nolimits_i (y_i - g(x_i, \theta))^2$$

This minimizing $\hat{\theta}$ satisfies the first order conditions

$$\frac{1}{n} \sum\nolimits_i 2 \left( y_i - g\left( x_i, \hat{\theta} \right) \right) \frac{dg\left( x_i, \hat{\theta} \right)}{d\theta} = 0$$

NLS is an extremum estimator, so we can apply our consistency theorem. The identification condition in that theorem requires that the plim of the objective function be minimized at the true $\theta_0$. Assuming a LLN holds we have.

$$Q_0 = plim Q_n = E\left[ (Y - g(X, \theta))^2 \right]$$

The first order condition for minimizing $Q_0$ is

$$E\left( 2(Y - g(X, \theta)) \frac{dg(X, \theta)}{d\theta} \right) = 0$$

(assuming enough regularity to take the derivative inside the expectation; applying the dominated convergence theorem). We can verify this holds at the true $\theta_0$, because

$$\begin{aligned} & E\left( 2(Y - g(X, \theta_0)) \frac{dg(X, \theta_0)}{d\theta} \right) \\ = \; & E\left( 2e \frac{dg(X, \theta_0)}{d\theta} \right) = E\left( E\left( 2e \frac{dg(X, \theta_0)}{d\theta} \mid X \right) \right) \\ = \; & E\left( 2E(e \mid X) \frac{dg(X, \theta_0)}{d\theta} \right) = 0 \end{aligned}$$

where we used the law of iterated expectations, and the assumption that $E(e \mid X) = 0$. So the necessary condition that $Q_0$ be optimized at the true $\theta_0$ holds.

However, the identification condition also requires that there not be any other $\theta$ that also minimizes $Q_0$. We therefore require that

$$E\left[2\left(Y - g\left(X, \theta\right)\right)\frac{dg\left(X, \theta\right)}{d\theta}\right] \neq 0 \text{ for any } \theta \in \Theta \text{ other than } \theta = \theta_0.$$

Does this condition hold? Maybe. The answer depends on exactly what the function $g$ and the distribution of $Y$ are.

### 4.1.2 Numerical estimation of NLS

How can the computer find the NLS $\hat{\theta}$? Taylor expanding $g\left(x, \hat{\theta}\right)$, we have

$$\begin{aligned} g\left(x, \hat{\theta}\right) &= g\left(x, \theta\right) + \frac{dg\left(x, \tilde{\theta}\right)}{d\theta}\left(\hat{\theta} - \theta\right) \text{ for some } \tilde{\theta} \text{ between } \theta \text{ and } \hat{\theta} \\ &\approx g\left(x, \theta\right) + \frac{dg\left(x, \hat{\theta}\right)}{d\theta}\left(\hat{\theta} - \theta\right) \end{aligned}$$

so that

$$g\left(x, \theta\right) \approx g\left(x, \hat{\theta}\right) - \frac{dg\left(x, \hat{\theta}\right)}{d\theta}\left(\hat{\theta} - \theta\right)$$

and since the model is $y_i = g\left(x_i, \theta\right) + e_i$, we have

$$y_i \approx g\left(x_i, \hat{\theta}\right) - \frac{dg\left(x_i, \hat{\theta}\right)}{d\theta}\left(\hat{\theta} - \theta\right) + e_i$$

So for $\theta = \theta_0$

$$y_i - g\left(x_i, \hat{\theta}\right) + \frac{dg\left(x_i, \hat{\theta}\right)}{d\theta}\hat{\theta} \approx \frac{dg\left(x_i, \hat{\theta}\right)}{d\theta}\theta_0 + e_i$$

Given some initial estimate $\hat{\theta}$ define

$$y_i^* = y_i - g\left(x_i, \hat{\theta}\right) + \frac{dg\left(x_i, \hat{\theta}\right)}{d\theta}\hat{\theta} \text{ and } x_i^* = \frac{dg\left(x_i, \hat{\theta}\right)}{d\theta}$$

Then

$$y_i^* \approx x_i^{*\prime}\theta_0 + e_i$$

This suggests the following iterative numerical procedure (called iterated least squares) to estimate $\theta$:

1. start with an initial guess/estimate $\hat{\theta}$, and use it to construct $y_i^*$ and $x_i^*$ as above
2. Run a linear OLS regression of $y_i^*$ on $x_i^*$. The estimated coefficient vector is your new estimate of $\hat{\theta}$.

   Repeat the above steps until $\hat{\theta}$ converges.

Potential problems: This might not converge, or it is possible that there are multiple values of $\theta$ that statisfy the first order conditions, and $\hat{\theta}$ could converge to the wrong one.

## 4.2 More general NLS

Above had $y = g(x, \theta) + e$. One could more generally have $G(y, x, \theta) = e$. Could still estimate by least squares:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum\nolimits_i [G(y_i, x_i, \theta)]^2$$

Though now it's numerically harder to find $\hat{\theta}$; we can't use the above iterated linear least squares method.

Examples:

1. $y = a + bx + cx^2 + e$ This is nonlinear in $x$, but can still estimate it using linear OLS, by just letting $z = x^2$, and linearly regress $y$ on a constant, $x$, and $z$.

2. $y = a + c^3 x + e$ This is nonlinear in the parameters, but can still estimate it using linear OLS, by just letting $b = c^3$ so $y = a + bx + e$. Linearly regress $y$ on a constant and on $x$, and then let $\widehat{c} = \widehat{b}^{1/3}$. We can apply the delta method to do inference on $\widehat{c}$.

3. $y = \dfrac{1 + bx}{c + dx} + e$ Can do NLS with $\theta = (b, c, d)'$ and $g(x, \theta) = \dfrac{1 + bx}{c + dx}$

4. $\dfrac{y + bx + cxy}{y + dx} = e$ Can do more general NLS with $\theta = (b, c, d)'$ $G(y, x, \theta) = \dfrac{y + bx + cxy}{y + dx}$

## 4.3 Extremum Estimators

The NLS estimators given above are extremum estimators. Earlier in the term we saw a general theorem for consistency of an extremum estimator. Here we'll provide a general theorem for the limiting distribution of an extremum estimator:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta).$$

This will require 5 assumptions, the final one being a particularly strong, high level assumption.

1. $\hat{\theta}$ is a consistent estimator for $\theta_0$ (our earlier Theorem in lecture 3 is sufficient for this).

2. $\theta_0 \in \Theta$, where $\theta_0$ is in the interior of $\Theta$.

3. $Q_n(\theta)$ is a twice continuously differentiable function.

4. $\sup_{\theta \in \Theta} \left| \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'} - H(\theta) \right| \xrightarrow{p} 0$, where $H(\theta) = plim \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'}$ (uniform convergence).

   The matrix function $H(\theta)$ is bounded, continuous, and nonsingular for all $\theta$ in a neighborhood of $\theta_0$. Let $H_0 = H(\theta_0)$.

5. $\partial Q_n(\theta_0)/\partial\theta$ is asymptotically linear. This means there exists some $S_i$ (which is a function of the data and the model) such that

$$\sqrt{n}\left(\frac{\partial Q_n(\theta_0)}{\partial\theta} - \frac{1}{n}\sum S_i\right) \xrightarrow{p} 0$$

$$\frac{1}{\sqrt{n}}\sum S_i \xrightarrow{d} N\left(0, \mathbf{\Sigma}_0\right)$$

Basically, this condition says that $\partial Q_n(\theta_0)/\partial\theta$ is asymptotically the same as an average that satisfies a CLT.

Theorem: If conditions 1 to 5 above hold, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H_0^{-1}\Sigma_0 H_0^{-1})$$

Proof: By the first order condition of the extremum estimator we have that $\frac{\partial Q_n(\widehat{\theta})}{\partial\theta} = 0$. We can perform a Taylor expansion on this. For some $\widetilde{\theta}$ between $\theta_0$ and $\widehat{\theta}$ we have

$$0 = \sqrt{n}\left(\frac{\partial Q_n(\widehat{\theta})}{\partial\theta}\right) = \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} + \frac{\partial^2 Q_n(\widetilde{\theta})}{\partial\theta\partial\theta'}\sqrt{n}(\widehat{\theta} - \theta_0).$$

Solving for $\sqrt{n}(\widehat{\theta} - \theta_0)$ gives

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \underbrace{\left[-\frac{\partial^2 Q_n(\widetilde{\theta})}{\partial\theta\partial\theta'}\right]^{-1}}_{\text{Term 1}} \underbrace{\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}}_{\text{Term 2}}.$$

By Assumption 4, $\frac{\partial^2 Q_n(\widetilde{\theta})}{\partial\theta\partial\theta'} \xrightarrow{p} H(\widetilde{\theta})$ and by consistency of $\widehat{\theta}$, $H(\widetilde{\theta}) \xrightarrow{p} H(\theta_0) = H_0$. Since the plim of a function is the function of the plim, term 1 above converges to $-H_0^{-1}$. Next, by assumption 5, term 2 converges in distribution to $N\left(0, \mathbf{\Sigma}_0\right)$. So$\sqrt{n}(\widehat{\theta} - \theta_0)$ is the product of two terms, one that converges in probability to $-H_0^{-1}$ and the other that converges in distribution to $N\left(0, \mathbf{\Sigma}_0\right)$. Applying the rule about product of such convergences, we get

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N\left(0, H_0^{-1}\mathbf{\Sigma}_0 H_0^{-1}\right).$$

Note the "sandwich" form of the variance. This is common in nonlinear estimators.

Example: Suppose $Q_n = \frac{1}{n}\sum_i R\left(Z_i, \theta\right)$ for some function $R$. Nonlinear least squares has this form. So does maximum likelihood estimation (MLE) with iid data. Suppose $Z_i$ are iid. Then $S_i = \frac{\partial R(Z_i,\theta_0)}{\partial\theta}$, $\mathbf{\Sigma}_0 = var\left(S\right)$, and $H_0 = E\left(\frac{\partial^2 R(Z_i,\theta_0)}{\partial\theta^2}\right)$. We then have the limiting distribution for MLE or NLS as:

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, E\left(\frac{\partial^2 R\left(Z_i, \theta_0\right)}{\partial\theta^2}\right)^{-1} var\left(\frac{\partial R\left(Z_i, \theta_0\right)}{\partial\theta}\right) E\left(\frac{\partial^2 R\left(Z_i, \theta_0\right)}{\partial\theta^2}\right)^{-1}\right).$$

In the special case of MLE, both $\mathbf{\Sigma}_0$ and $H_0$ equal the matrix we called $J_0$.

## 4.4 Numerical issues of Extremum Estimators

How does the computer numerically find $\widehat{\theta} = \arg\max_{\theta} Q_n(\theta)$ ?

One set of methods are called gradient (hill climbing) routines.

Let $\widetilde{\theta}$ be a starting guess for $\theta_0$. look at a mean value (Taylor) expansion of the first order condition that the unknown $\widehat{\theta}$ should satisfy, around the given value $\widetilde{\theta}$

$$0 = \frac{\partial Q_n(\widehat{\theta})}{\partial \theta} \approx \frac{\partial Q_n(\widetilde{\theta})}{\partial \theta} + \frac{\partial^2 Q_n(\widetilde{\theta})}{\partial \theta} \left(\widehat{\theta} - \widetilde{\theta}\right)$$

so

$$\widehat{\theta} \approx \widetilde{\theta} - \left(\frac{\partial^2 Q_n(\widetilde{\theta})}{\partial \theta}\right)^{-1} \frac{\partial Q_n(\widetilde{\theta})}{\partial \theta}$$

We call $-\left(\frac{\partial^2 Q_n(\widetilde{\theta})}{\partial \theta}\right)^{-1} \frac{\partial Q_n(\widetilde{\theta})}{\partial \theta}$ the step. Starting from $\widetilde{\theta}$, we add the step, and call the result a new $\widetilde{\theta}$. Repeat (iterate) this procedure until the step is numerically very close to zero. The result is then our estimate $\widehat{\theta}$.

Think of the $Q_n$ function as a (high dimensional) mountain, and we're trying to climb to the top. But imagine being blindfolded. We start at point $\widetilde{\theta}$. By feeling the slope of the mountain under our feet, we take a step in the direction we think is uphill (i.e., has a positive slope).

We expect each estimate of $\widetilde{\theta}$ to be better than the previous one, that is we expect $Q_n(\widetilde{\theta}+step) > Q_n(\widetilde{\theta})$, meaning that the step takes us uphill rather than down. But it's possible that we, e.g. step in hole, or off a cliff. What if that happens? Then you can take a smaller step: let your new estimate be $\widetilde{\theta} + step * \lambda$ for some positive $\lambda < 1$. And if that doesn't work, you can try an even smaller step: $\widetilde{\theta} + step * \lambda^2$. This process of looking at smaller and smaller steps until one leads to an improved estimate, is called, "squeezing."

To implement this procedure we must give the computer:

1. the function $Q_n(\theta)$ and its first and second derivatives (or let the computer numerically calculate those derivatives. Numerical is slower, and introduces an additional source of numerical errors).

2. a 'starting value,' that is, an initial guess of $\widetilde{\theta}$.

3. a 'convergence criterion,' that is, how small a change (in absolute value or percentage terms) in $\widetilde{\theta}$ and in $Q_n(\widetilde{\theta})$ counts as having converged.

4. a maximum number of iterations: We can't let the computer run forever, so at some point you may need to stop the search even if it hasn't yet converged.

5. a maximum number of squeezes: We can't let the computer keep reducing the step size forever - at some point the steps would get smaller than the roundoff errors of the computer calculations.

What can go wrong when trying to find $\widehat{\theta}$?

1. The process might converge to some 'wrong' value: a local maximum instead of the global maximum, or an inflection point or saddle point instead of a max. To be reasonably sure you've found a global maximum, you may need to repeat the search using many different starting values.

2. Could fail to converge: $\frac{\partial^2 Q_n(\widetilde{\theta})}{\partial \theta}$ could be too close to singular to invert (this is analogous to high multicollinearity in a linear model, where $X'X$ is close to singular).

3. The squeeze limit is exceeded: can't find a small enough step size to improve the estimate, but the estimate hasn't converged.

4. The maximum number of iterations is exceeded. The objective function could have a difficult shape to maximize (the shape of curved ridge).

There exist other, non-gradient based, methods to find the maximizing $\widehat{\theta}$, like a grid search and so-called genetic algorithms. These are generally very much slower, taking far more computing time and power, but can often succeed in finding a maximizing value $\widehat{\theta}$ when gradient based methods fail.

## 4.5 Generalized Method of Moments (GMM)

### 4.5.1 Method of Moments (MM)

Suppose, given observations of a random sequence $Z_1$,...,$Z_n$ we want to estimate $\theta_0 = E[h(Z)]$ for some known function $h$. The obvious estimator, based on a LLN, is $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$.

We can equivalently say we have a model $E[h(Z) - \theta_0] = 0$, and we estimate the parameter $\theta_0$ by finding the value $\widehat{\theta}$ such that $\frac{1}{n}\sum_{i=1}^{n}\left[h(Z_i) - \widehat{\theta}\right] = 0$.

We can apply the same logic to estimating $\theta_0$ defined by the model $E[G(Z, \theta_0)] = 0$ for some known function $G$. A sensible estimator is the value $\widehat{\theta}$ such that $\frac{1}{n}\sum_{i=1}^{n}\left[G\left(Z_i, \widehat{\theta}\right)\right] = 0$. The previous example is just a special case where $G(Z, \theta)$ is the function $h(Z) - \theta$.

This is called the Method of Moments, or MM, estimator. Note that $h$, $G$ and $\theta$ can all be vectors. But they all must have the same dimension $K$. The MM estimator then consists of solving $K$ equations in $K$ unknowns to get $\widehat{\theta}$.

Example 1: suppose $Z_i$ are iid with $E(Z_i) = \mu$ and $var(Z_i) = \sigma^2$. To estimate these moments, we can let

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \quad G(Z_i, \theta) = \begin{pmatrix} Z_i - \mu \\ (Z_i - \mu)^2 - \sigma^2 \end{pmatrix}$$

With this definition of $G$, we have by the definitions of $\mu$ and $\sigma^2$ that $E[G(Z, \theta_0)] = 0$, and so a corresponding estimator is the MM, which here would be the $K = 2$ equations

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Z_i - \widehat{\mu} \\ (Z_i - \widehat{\mu})^2 - \widehat{\sigma}^2 \end{pmatrix} = 0$$

Example 2: Consider regular maximum likelihood estimation with iid data. This is a special case of an MM estimator, where $G(Z, \theta)$ is the score function.

Suppose $\widehat{\theta}$ is given by the MM estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left[G\left(Z_i,\widehat{\theta}\right)\right]=0$$

for some $K$ vector valued function $G$. Another way to write the same estimator is

$$\widehat{\theta}=\arg\min\left(\frac{1}{n}\sum_{i=1}^{n}\left[G\left(Z_i,\theta\right)\right]\right)'W\left(\frac{1}{n}\sum_{i=1}^{n}\left[G\left(Z_i,\theta\right)\right]\right)$$

where $W$ is any positive definite matrix (e.g., $W$ could be the identity matrix). This works because the above expression is a quadratic in $\frac{1}{n}\sum_{i=1}^{n}\left[G\left(Z_i,\widehat{\theta}\right)\right]$. Quadratics are nonnegative, so the minimizing value of the quadratic must be zero, and the quadratic can be zero only if $\frac{1}{n}\sum_{i=1}^{n}\left[G\left(Z_i,\widehat{\theta}\right)\right]$ is zero.

The advantage of writing our MM estimator in the above quadratic form is that it makes the MM estimator be an extremum estimator, and we can then apply our theorems for consistency and asymptotic normality of extremum estimators to the MM estimator.

### 4.5.2 OLS and IV Linear Regressions as MM estimators

Example 1: Consider the linear regression $Y_i=X_i'\theta+e_i$ where $\theta$ is a $K$ vector of coefficients. We usually express the OLS estimator as

$$\widehat{\theta}_{OLS}=\arg\min\sum_{i=1}^{n}e_i^2=\arg\min\sum_{i=1}^{n}\left(Y_i-X_i'\theta\right)^2$$

which has the first order conditions

$$\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-X_i'\widehat{\theta}_{OLS}\right)=0$$

The OLS estimator is an example of an MM estimator, where $Z_i=(Y_i,X_i)$ and the function $G$ is $G\left(Z_i,\theta\right)=X_i\left(Y_i-X_i'\theta\right)$, so the $K-$ vector of OLS moments are

$$E\left[X_i\left(Y_i-X_i'\theta_0\right)\right]=0$$

Notice these moments are equivalent to $E\left(X_ie_i\right)=0$. That is, OLS is an MM estimator where the moments are that the regressors are uncorrelated with the errors.

By observing that OLS is an MM estimator, we get an alternative way to write the OLS estimator:

$$\widehat{\theta}_{OLS}=\arg\min\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-X_i'\theta\right)X_i'\right)W\left(\frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i-X_i'\theta\right)\right)$$

Example 2: Consider again the linear regression $Y_i=X_i'\theta+e_i$ where $\theta$ is a $K$ vector of coefficients. But now suppose $X_i$ is endogenous, so we instead do instrumental variables estimation, using a $K$ vector of instruments. IV is also a MM estimator, which is equivalent to the moments that the instruments and the errors are uncorrelated.

Example 3: Once more start with the linear regression $Y_i = X_i'\theta + e_i$ where $\theta$ is a $K$ vector of coefficients and $X_i$ is endogenous. But now suppose we are overidentified. We have an $L$ vector of instruments $Q_i$ where $L > K$. Since these are all valid instruments, we have the $L$ moments $E(Qe) = 0$, which would imply the MM estimator $\frac{1}{n}\sum_{i=1}^n Q_i\left(Y_i - X_i'\widehat{\theta}\right) = 0$.

But this MM estimator won't work! It consists of $L$ equations, and $\widehat{\theta}$ only has $K$ elements. In general, we cannot find $K$ constants that will simultaneously solve $L$ equations! Asymptotically, our assumption is that the true $K$-vector $\theta_0$ does solve the $L$ equations $E[Q(Y - X'\theta_0)] = 0$, but in general we won't be able to find $\widehat{\theta}$ that makes all $L$ sample moments be exactly zero.

### 4.5.3 GMM for Linear Regressions With Instruments

How do we solve the problem of having more instruments than coefficients? One solution is 2SLS: from the $L$ instruments $Q$, first construct just $K$ instruments, and then do IV (or equivalently, MM) using just the $K$ moments that say each constructed instrument is uncorrelated with $e$.

But there's another possible solution. Consider the alternative way of writing the MM estimator, as minimizing a quadratic. For the moments $E(Qe) = 0$, this becomes

$$\widehat{\theta} = \arg\min\left(\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\theta)\, Q_i'\right) W \left(\frac{1}{n}\sum_{i=1}^n Q_i (Y_i - X_i'\theta)\right)$$

Notice that $W$ is now an $L \times L$ positive definite matrix. Unlike MM, this $\widehat{\theta}$ can't make the $L$ vector of sample moments $\frac{1}{n}\sum_{i=1}^n Q_i\left(Y_i - X_i'\widehat{\theta}\right)$ all equal zero, because no choice of a $K$ vector $\widehat{\theta}$ can do that. But by minimizing this quadratic, we get a $\widehat{\theta}$ that is in some way as close as we can get to making the $L$ sample moments equal zero. This is an example of the Generalized Method of Moments, or GMM estimator.

Unlike MM, the GMM estimator $\widehat{\theta}$ does depend on what $W$ matrix you choose. In particular, a large element on the diagonal of $W$ puts a lot of weight on that corresponding moment, so $\widehat{\theta}$ will be chosen to make that element much closer to zero, at the expense of letting other moments be further from zero.

It can be shown that 2SLS is the special case of the above GMM estimator where $W = \left(\frac{1}{n}\sum_{i=1}^n Q_i Q_i'\right)^{-1}$. Other choices of $W$ yield estimators for $\theta$ that are not the same as 2SLS. We'll discuss the general choice of $W$ later.

### 4.5.4 Other Examples of GMM

The previous subsection gave the GMM estimator for linear regression with instruments, based on moments $E(Qe) = 0$. More generally, we can write a GMM estimator for moments of the general form $E[G(Z, \theta_0)] = 0$, where the function $G$ is a known $L$-vector valued function, $\theta$ is an unknown $K$ vector of parameters, the model is the $L$ equations $E[G(Z, \theta_0)] = 0$, and $K \leq L$. MM is the special case of GMM where $K = L$. The GMM estimator is

$$\widehat{\theta}_{GMM} = \arg\min\left(\frac{1}{n}\sum_{i=1}^n [G(Z_i, \theta)]\right)' W \left(\frac{1}{n}\sum_{i=1}^n [G(Z_i, \theta)]\right)$$

Example: Suppose we have the nonlinear instrumental variables regression model

$$Y_i = g(X_i, \theta) + e_i$$

where $\theta$ is an unknown $K$ vector of parameters, and $Q_i$ is an $L$ vector of instruments, with $L > K$, that satisfy $E(Qe) = 0$. We don't want to do something like 2SLS, because in a nonlinear model theres no obvious way to choose linear functions of $Q$ to use as instruments for a second stage IV. But we can instead do GMM estimation where $G(Z, \theta) = Q(Y - g(X, \theta))$.

Example: Rational Expectations models. Consider choosing consumption this period to maximize lifetime expected utility

$$E_t\left[\sum_{s=t}^{T} \beta^s U(C_s, \theta)\right]$$

under a lifetime budget constraint

$$\sum_{s=t}^{T} R_s (C_s - Y_s) + A_t = 0$$

Here $C$ is consumption, $\beta$ is the personal rate of time preference, $R$ is the discount rate, $Y$ is income, $A$ is assets, and $U$ is the utility function. We want to estimate parameters of the utility function $\theta$ (these could include objects like relative risk aversion) and maybe $\beta$ as well.

The Euler equation for this maximization is

$$\frac{\partial U(C_t, \theta)}{\partial C_t} = \beta E_t\left[\frac{\partial U(C_{t+1}, \theta)}{\partial C_{t+1}} R_t\right]$$

Let $S_t$ be a set of variables known at time $t$. Then

$$E\left[\left(\frac{\partial U(C_t, \theta)}{\partial C_t} - \beta \frac{\partial U(C_{t+1}, \theta)}{\partial C_{t+1}} R_t\right) S_t\right] = 0$$

This is now a set of moments we can estimate using GMM.

An attractive feature of this example is that it only uses what economic theory tells us. For example, it doesn't require guessing what the distribution of variables like $C_{t+1}$ and $R_t$ are, which we would need for something like maximium likelihood.

Historic note MM was invented by Karl Pearson in 1894.

GMM was invented by Lars Peter Hansen in 1982, and first applied to a rational expectations model by Hansen and Singleton in 1982. Hansen won the nobel prize in part for this work in 2013.

### 4.5.5 GMM Limiting Distribution

We can show consistency of the GMM estimator using our theorem for consistency of extremum estimators. Some of the conditions for that theorem were discussed earlier. Now we'll derive its asymptotic distribution.

The GMM estimator is:

$$\widehat{\theta} = \arg\min \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \theta)\right)' \widehat{W} \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \theta)\right)$$

Notice we're now allowing for the possibility that the matrix $W$ is estimated. Assume $\widehat{W} \xrightarrow{p} w$ for some postive definite $w$. The first order condition, multiplied by $\sqrt{n}$, is

$$0 = \left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta}\right)' \widehat{W} \left(\sqrt{n}\frac{1}{n}\sum\nolimits_i G(Z_i,\widehat{\theta})\right)$$

Notice the dimensions of these terms: $\frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta}$ is $L \times K$, $\widehat{W}$ is $L \times L$, and $G(Z_i,\widehat{\theta})$ is $L \times 1$, so the first order condition is that a $K \times 1$ vector equals zero, giving $K$ equations in the $K$ unknowns $\widehat{\theta}$.

Start by taking a Taylor expansion of the second term in parentheses:

$$\begin{aligned}
\sqrt{n}\frac{1}{n}\sum\nolimits_i G(Z_i,\widehat{\theta}) &= \sqrt{n}\frac{1}{n}\sum\nolimits_i \left(G(Z_i,\theta_0) + \frac{\partial G(Z_i,\widetilde{\theta})'}{\partial \theta}(\hat{\theta}-\theta_0)\right) \\
&= \left(\sqrt{n}\frac{1}{n}\sum\nolimits_i G(Z_i,\theta_0)\right) + \left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widetilde{\theta})}{\partial \theta}\right)' \sqrt{n}(\hat{\theta}-\theta_0)
\end{aligned}$$

where $\widetilde{\theta}$ lies between $\hat{\theta}$ and $\theta_0$. Plugging this expression back in we get

$$0 = \left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta}\right)' \widehat{W} \left(\left(\sqrt{n}\frac{1}{n}\sum\nolimits_i G(Z_i,\theta_0)\right) + \left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widetilde{\theta})}{\partial \theta}\right)' \sqrt{n}(\hat{\theta}-\theta_0)\right)$$

Solving this for $\sqrt{n}(\hat{\theta}-\theta_0)$ gives

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &= \left[-\left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta}\right)' \widehat{W} \left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widetilde{\theta})}{\partial \theta}\right)\right]^{-1} \\
&\quad \left(\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta}\right)' \widehat{W} \left(\sqrt{n}\frac{1}{n}\sum\nolimits_i G(Z_i,\theta_0)\right)
\end{aligned}$$

Let $r = E\left(\frac{\partial G(Z_i,\theta_0)}{\partial \theta}\right)$. Using assumptions like those that we made for maximum likelihood, we can show

$$\frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta} \xrightarrow{p} E\left(\frac{\partial G(Z_i,\theta_0)}{\partial \theta}\right) = r$$

For example, we could first expand $\frac{\partial G(Z_i,\widehat{\theta})}{\partial \theta}$ around $\frac{\partial G(Z_i,\theta_0)}{\partial \theta}$, show the difference between the two has a probability limit of zero (using consistency of $\widehat{\theta}$), and then apply a LLN to the average of $\frac{\partial G(Z_i,\theta_0)}{\partial \theta}$. In the same way we can show that $\frac{\partial G(Z_i,\widetilde{\theta})}{\partial \theta} \xrightarrow{p} r$.

So we get that $\sqrt{n}(\hat{\theta}-\theta_0)$ equals a term that goes in probability to $(-r'wr)^{-1}r'w$ times $\sqrt{n}\frac{1}{n}\sum_i G(Z_i,\theta_0)$. Next assume that $\sqrt{n}\frac{1}{n}\sum_i G(Z_i,\theta_0)$ satisfies a Central Limit Theorem, so

$$\sqrt{n}\frac{1}{n}\sum\nolimits_i G(Z_i,\theta_0) \xrightarrow{d} N(0,\Omega)$$

for some variance matrix $\Omega$. This normal is mean zero because the GMM moments are $E[G(Z_i,\theta_0)] = 0$. If $Z_i$ is iid then this can just be the Lindeberg-Levy CLT and $\Omega = var(G(Z_i,\theta_0)) = E[G(Z_i,\theta_0)G(Z_i,\theta_0)']$.

Finally, we can apply the rule about a product of sequences where one converges in distribution and the other converges in probability to get

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, (r'wr)^{-1} r'w\Omega wr (r'wr)^{-1}\right)$$

Notice the sandwich form of the variance. Note also that $r$ is $L \times K$, while $w$ and $\Omega$ are $L \times L$ matrices.

### 4.5.6 Efficient Two Step GMM

Recall that $w$ was any chosen positive definite matrix. It turns out that the most efficient choice of $w$ is $\Omega^{-1}$. With this choice of $w$, the limiting variance simplifies, and we get:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \left(r'\Omega^{-1}r\right)^{-1}\right)$$

The problem is, we don't know $\Omega$. However, we can consistently estimate it, and our derivation allows us to use an estimated $\widehat{W}$. This yields the following efficient Two Step GMM estimator:

1. Let $\widetilde{\theta}$ be the GMM estimate of $\theta$, using the identity matrix (or any other convenient choice) of weighting matrix. That is, let

$$\widetilde{\theta} = \arg\min \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \theta)\right)' \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \theta)\right)$$

2. Let $\widehat{W}$ be a consistent estimator of $\Omega^{-1}$, using $\widetilde{\theta}$. In particular, with iid data we have

$$\widehat{W} = \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \widetilde{\theta}) G(Z_i, \widetilde{\theta})'\right)^{-1}$$

Then let $\widehat{\theta}$ be the GMM estimator, using this $\widehat{W}$, so

$$\widehat{\theta} = \arg\min \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \theta)\right)' \widehat{W} \left(\frac{1}{n}\sum\nolimits_i G(Z_i, \theta)\right)$$

This estimator then has the limiting distribution

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \left(r'\Omega^{-1}r\right)^{-1}\right)$$

from which we get the asymptotic approximation

$$\hat{\theta} \sim^a N\left(\theta_0, \frac{1}{n}\left(\widehat{r}'\widehat{W}\widehat{r}\right)^{-1}\right)$$

where

$$\widehat{r} = \frac{1}{n}\sum\nolimits_i \frac{\partial G(Z_i, \widehat{\theta})}{\partial \theta}$$

Example: Return to the linear regression with instruments model $Y_i = X_i'\theta + e_i$ with instruments $Q_i$. In this model $G(Z_i, \theta) = Q_i(Y_i - X_i'\theta)$. It can be shown that in this model 2SLS is the

special case of GMM where $\widehat{W} = \left(\frac{1}{n}\sum_{i=1}^{n} Q_i Q_i'\right)^{-1}$. If the errors $e_i$ are homoskedastic and not autocorrelated, then this $\widehat{W}$ is efficient. But if the errors are heteroskedastic or autocorrelated, then two step GMM will be more efficient than 2SLS. In particular, if we have heteroskedasticity but not autocorrelation, then the optimal $\widehat{W}$ will resemble the matrix we constructed for the White corrected standard errors. But unlike in OLS, where the White correction only affects the standard errors, here in GMM it affects the estimated coefficients as well.

Why did we stop at two steps? We could have used our $\hat{\theta}$ to construct a new $\widehat{W}$, and kept iterating unti convergence. That alternative estimator is sometimes used. It's called the Continuously Updated Estimator, or CUE GMM. But it turns out to be unnecessary. The two step GMM is just as efficient as the CUE GMM. Both have the same limiting distribution given above.

# 5 Lecture 12. Nonparametric Estimators

All of our discussion of estimation so far has focused on estimating a finite set of parameters, like a vector $\theta$. Nonparametric estimation is estimation of an infinite number of parameters, like an infinite vector, or a function.

## 5.1 Introduction

- Definitions of parametric, nonparametric, and semiparametrics:

  Parametric estimation is estimation of a finite number of parameters. Examples:

  Linear regression: $Y_i = X_i'\beta + e_i$

  nonlinear regression: $Y_i = G(X_i, \theta) + e_i$

  these have finite dimensional vectors $\theta$, $\beta$

  Nonparametric estimation is estimation of an infinite number of parameters. Example

  nonparametric regression: $Y_i = m(X_i) + e_i$

  $X$ can take on an infinite number of values, and so the unknown function $m()$ consists of an infinite number of parameters: one value of $m(X)$ for each possible value of $X$.

  Semiparametric estimation: estimation having a finite number of parameters of interest, but other components that are infinite. Example

  Average elasticities: $\alpha = E\left(\frac{\partial \ln m(X_i)}{\partial \ln X}\right)$. $\alpha$ is a finite vector, $m()$ is an unknown function.

  Partly linear model: $Y_i = m(X_i) + Z_i'\gamma + e_i$. $\gamma$ is a finite vector, $m()$ is an unknown function.

- Why do nonparametrics estimation?

  Many economic theories do not imply specific functional forms. E.g., we often assume linear models, but rarely have a good reason for linearity.

  Can provide guidance for parametric models.

- Why not always do nonparametrics? E.g., if we can estimate $Y_i = m(X_i) + e_i$ without knowing $m$, why should we risk specification error by assuming a functional form like linearity?

  Curse of Dimensionality: nonparametric estimation is usually much, much less efficient than parametric estimation, with convergence rates slower than root-n, and with rates that get slower as the number of covariates increases.

- Why use semiparametrics estimation?

  Sometimes can overcome the curse of dimensionality, with the parametric part converging at rate root-n .

  More popular in econometrics than in statistics (we have finite dimensional features of interest).

## 5.2 Estimation

### 5.2.1 Empirical Distribution Function

Consider estimation of a distribution function.

$X$ is a r.v. (random variable), for now a scalar. $x$ is a value $X$ could take on. $X_i$ for $i = 1, ..., n$ are iid rv's. $x_i$ for $i = 1, ..., n$ are the corresponding realizations (observations, our sample). Each $X_i$ is drawn from a distribution function $F(X)$, each $X_i$ is a r.v. with distribution $F$.

$F(x) = \Pr(X \leq x)$ – the true distribution function of $X$, evaluated at a point $x$. If $X$ continuous, $F(x)$ is usually an $S$ shaped curve. It has a range of values from 0 to 1.

Note that $F()$ is a function. But for a given number $x$, $F(x)$ is a scalar, the function evaluated at this one point. By giving an estimator for $F(x)$ that works for every separate value $x$, we get a (pointwise) nonparametric estimator for the function.

Recall $I()$ is the indicator function, that equals one if what's written in the parentheses is true, and zero otherwise. For any given value $x$, define

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

This $\widehat{F}(x)$ is just the fraction of our sample that is less than or equal to the value $x$.
$\widehat{F}()$ is called the empirical distribution function.
$\widehat{F}(x)$ is the empirical distribution function evaluated at the value $x$.

Graph $\widehat{F}(x)$ against $x$: a step function with $n$ steps.
$\widehat{F}(x)$ is a nonparametric estimator of $F(x)$.
What are its properties?

### 5.2.2 Bias

Is $\widehat{F}(x)$ an unbiased estimator of $F(x)$?

Recall what unbiasedness means:
Suppose we had an estimator $\widehat{\theta}$ of a vector $\theta$.
Imagine we had very many data sets, instead of just one.
Calculate $\widehat{\theta}$ using each data set separately.
$\widehat{\theta}$ is unbiased if the average of the estimates $\widehat{\theta}$ across all data sets equals the true value $\theta_0$. This needs to hold for each element $\theta_j$ of the vector $\theta$.

The nonparametric estimator $\widehat{F}(x)$ is unbiased if the average of $\widehat{F}(x)$ across an infinite number of data sets equals the true $F(x)$ for every possible $x$.

Here $x$ is like an index, it just refers to one 'element' of the function $F()$ (the element we happen to be estimating), just like $j$ in $\theta_j$ indexes one element of a vector $\theta$ in a parametric model.

For every real number $x$, $\widehat{F}(x)$ is unbiased. Heres the derivation:

$$E\left[\widehat{F}(x)\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right]$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} E\left[I(X_i \leq x)\right] \\
&= E\left[I(X_i \leq x)\right] = E\left[I(X \leq x)\right] \\
&= \int_{-\infty}^{\infty} I(X \leq x) f(X)\, dX = \int_{-\infty}^{x} f(X)\, dX \\
&= F(X)\,|_{-\infty}^{x} = F(x) - F(-\infty) \\
&= F(x) - 0 = F(x)
\end{aligned}$$

Notice the $X_i$'s are the random variables we are averaging over by taking the expectation. We are *not* averaging across $x$ values. Again, $x$ is like an index, it just refers to an 'element' of the function $F()$, just like $j$ in $\theta_j$ indexes one element of a vector $\theta$.

So we have shown $\widehat{F}(x)$ is unbiased:

$$E\left[\widehat{F}(x)\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)\right] = F(x)$$

### 5.2.3 Variance and limiting distribution

Using the same kind of algebra we had above for bias, we can show that the variance of $\widehat{F}(x)$ is given by:

$$\begin{aligned}
var\left[\widehat{F}(x)\right] &= E\left[\left(\widehat{F}(x) - F(x)\right)^2\right] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left[I(X_i \leq x) - F(x)\right]\right)^2\right] \\
&= E\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[I(X_i \leq x) - F(x)\right]\left[I(X_j \leq x) - F(x)\right]\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} E\left(\left[I(X_i \leq x) - F(x)\right]^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} E\left(\left(I(X_i \leq x)\right)^2 - 2I(X_i \leq x)F(x) + F(x)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} E\left(\left(I(X_i \leq x)\right) - 2I(X_i \leq x)F(x) + F(x)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} E\left(F(x) - 2F(x)F(x) + F(x)^2\right) \\
&\quad \frac{1}{n}F(x)\left[1 - F(x)\right]
\end{aligned}$$

Note above uses the fact that $\left(I(X_i \leq x)\right)^2 = I(X_i \leq x)$ because it can only equal zero or one.

Combining unbiasedness and the above variance formula, we get that $\widehat{F}(x)$ converges in mean square to $F(x)$, and so $\widehat{F}(x)$ consistently estimated $F(x)$, i.e., $\widehat{F}(x) \xrightarrow{p} F(x)$.

What about asymptotic distribution? $\widehat{F}(x)$ is the average of $I(X_i \leq x)$. Since $X_i$ is iid, $I(X_i \leq x)$ is also iid (and is bounded since it only equals zero or one). So we can apply the Lindeberg-Levy CLT to get,

$$\sqrt{n}\left(\widehat{F}(x) - F(x)\right) \xrightarrow{d} N[0, F(x)[1 - F(x)]]$$

$\widehat{F}(x)$ is root-n-CAN: consistent, asymptotically normal, at rate root-n. True at every point (every real number) $x$. So we have a nonparametric root-n-CAN estimator.

We have shown above that $\widehat{F}(x)$ converges pointwise. In fact, it is also possible to prove that $\widehat{F}(x)$ converges uniformly to $F(x)$ meaning that:

$$\sup_x |\widehat{F}(x) - F(x)| \xrightarrow{p} 0$$

This result is the Glivenko–Cantelli theorem.

## 5.3 Kernel Density Estimation

Now suppose $X$ is continuously distributed, and we want to estimate $f(x)$, the pdf (probability density function) of $X$, at each point $x$. Assume this pdf is a nice smooth function (twice continuously differentiable).

If one can parameterize it as $f(x, \theta)$, then we can do maximum likelihood estimation. For example, if we assumed $X$ was normal, then $\theta$ could be the mean and variance, and we could construct an MLE to estimate $\theta$, or just use our usual unbiased estimators of mean and variance.

Another alternative is to approximate $f(x)$ by a histogram:

For a small number (called the binwidth) $h$, look at

$$\frac{1}{n}\sum_{i=1}^{n} I(x - h \leq X_i \leq x + h)$$

This is the fraction of observations in our sample that are near $x$ (within the distance $h$ of $x$). This gives us the rough, histogram bar estimate of the density at $x$.

### 5.3.1 The Kernel Estimator

We will now construct a formal estimator of $f$, similar to the histogram. We can start by observing that the bar of the histogram for $f(x)$ is actually the difference of the empirical distribution function, evaluated at $x - h$ and $x + h$.

$$\begin{aligned}
\widehat{F}(x + h) - \widehat{F}(x - h) &= \frac{1}{n}\sum_{i=1}^{n} I(x - h \leq X_i \leq x + h) \\
&= \frac{1}{n}\sum_{i=1}^{n} I\left(|\frac{x - X_i}{h}| \leq 1\right)
\end{aligned}$$

Now compare that to the definition of the pdf $f(x)$ relative to the cdf $F(x)$.

$$\begin{aligned} f(x) &= \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} \\ &\approx \frac{F(x+h) - F(x-h)}{2h} \quad \text{for small } h \end{aligned}$$

This suggests the following estimator: Pick a small value of $h$ and let

$$\widehat{f}_h(x) = \frac{\widehat{F}(x+h) - \widehat{F}(x-h)}{2h} = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I\left( \left| \frac{x - X_i}{h} \right| \leq 1 \right)$$

We call this estimator $\widehat{f}_h(x)$ instead of just $\widehat{f}(x)$, because it depends on the number $h$ that we choose. Another way to write this estimator is

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - X_i}{h} \right) \qquad \text{where} \qquad K(u) = \frac{1}{2} I(|u| \leq 1)$$

so $K(u)$ equals a uniform density on $[-1, 1]$.

This estimator is a little ugly in that it's not continuous in $x$, because $K(u)$ discontinuously jumps from $1/2$ to zero where $|u| = 1$.

This $K(u)$ function also makes the estimator inefficient, because it gives equal weight to all observations $x_i$ within $h$ of $x$ and zero weight to those outside the distance $h$ from $x$. It will generally be more efficient to give the greatest weight to observations that are closest to $x$, because they are the most informative points about the density at $x$.

One can consider replacing the uniform density with some other $K(u)$ function to get a smoother and more efficient estimator. In particular, we can choose a $K(u)$ that is smooth without discontinuities, and one that is larger the closer $u$ is to zero, meaning the closer an observation $x_i$ is to $x$.

For a general choice of the function $K(u)$, this is called the Rosenblatt and Parzen kernel density estimator, or just a kernel density estimator:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - X_i}{h} \right)$$

The chosen constant $h$ is called the binwidth or the bandwidth, and $K(u)$ is called the kernel function (or window function). To implement this estimator you must choose both a kernel function $K(u)$ and a bandwidth $h$. The kernel function $K(u)$ should be continuous almost everywhere and satisfies $\int_{-\infty}^{\infty} K(u)\, du = 1$. (these are also properties that ordinary density functions have). $K$ is usually also chosen to be a symmetric function, meaning $K(u) = K(-u)$, and to have a mode at zero.

Popular choices of kernel functions are the gaussian (normal) kernel and the Epanechnikov (quadratic) kernel:

$$\begin{aligned} K(u) &= \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \\ K(u) &= \frac{3}{4}\left(1-u^2\right) I\left(|u| \leq 1\right) \end{aligned}$$

### 5.3.2 Bias

Now let's start figuring out the properties of kernel density estimators. First, is $\widehat{f_h}(x)$ unbiased like $\widehat{F}(x)$ was? Assume $K(u) = K(-u)$, i.e. a symmetric kernel.

$$E[\hat{f}_h(x)] = E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] = E\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right] = \int_{-\infty}^{\infty} \frac{1}{h}K\left(\frac{x-X}{h}\right) f(X)\, dX$$

If we let $u = \frac{X-x}{h}$ so $X = x + uh$ we can perform a change of variables in the above integral to get:

$$E[\hat{f}_h(x)] = \int_{-\infty}^{\infty} \frac{1}{h}K(-u) f(x+uh) h\, du = \int_{-\infty}^{\infty} K(u) f(x+uh)\, du$$

Note that the jacobian term for the change of variables replaces $dX$ with $hdu$, and that $h$ cancels out the leading $\frac{1}{h}$ term.

Now we'll do a second order Taylor expansion of $f(x+uh)$ around $x$. For some $\tilde{x}$ between $x$ and $x+uh$, we have

$$\begin{aligned} E[\hat{f}_h(x)] &= \int_{-\infty}^{\infty} K(u)\left[f(x) + uh\frac{\partial f(x)}{\partial x} + \frac{u^2h^2}{2}\frac{\partial^2 f(\tilde{x})}{\partial x}\right] du \\ &= \left(f(x)\int_{-\infty}^{\infty} K(u)\, du\right) + \left(h\frac{\partial f(x)}{\partial x}\int_{-\infty}^{\infty} uK(u)\, du\right) + \left(h^2\int_{-\infty}^{\infty} \frac{\partial^2 f(\tilde{x})}{\partial x^2}\frac{u^2K(u)}{2}\, du\right) \end{aligned}$$

Note that we've taken the $f(x)$ functions outside of the integrals in the first two terms above, but we can't take the $f(\tilde{x})$ term out in the last integral, because $\tilde{x}$ depends on $u$.

The first integral above equals one by the properties of $K(u)$, the and the second equals zero by the symmetry of $K(u)$, so

$$E[\hat{f}_h(x)] = f(x) + h^2\int_{-\infty}^{\infty} \frac{\partial^2 f(\tilde{x})}{\partial x^2}\frac{u^2K(u)}{2}\, du$$

and therefore the bias is

$$\begin{aligned} bias\left[\widehat{f_h}(x)\right] &= E[\hat{f}_h(x)] - f(x) \\ &= h^2\int_{-\infty}^{\infty} \frac{\partial^2 f(\tilde{x})}{\partial x^2}\frac{u^2K(u)}{2}\, du \\ &\approx h^2 b(x), \text{ where } b(x) = \frac{1}{2}\frac{d^2 f(x)}{dx^2}\int_{u=-\infty}^{\infty} u^2 K(u)\, du \end{aligned}$$

It's possible to be a little more precise about the above approximation, saying (in little O notation) that

$$bias\left[\widehat{f_h}(x)\right] = h^2 b(x) + o\left(h^2\right)$$

This shows the kernel density estimator is biased. The bias is approximately proportional to $h^2$, so the smaller the $h$ we choose, the less biased is $\widehat{f_h}(x)$. The size of the bias also depends on the choice of the kernel function, and on the second derivative of the density function.

### 5.3.3 Variance and bandwidth choice

Using the same change of variables as we did for the bias calculation, and just a first order Taylor expansion, we can show that

$$\begin{aligned} var\left[\widehat{f_h}(x)\right] &= \frac{1}{nh} v(x) + O(h) \approx \frac{1}{nh} v(x), \\ \text{where } v(x) &= f(x) \int_{u=-\infty}^{\infty} [K(u)]^2 du \end{aligned}$$

The variance is approximately proportional to $\frac{1}{nh}$, so as usual a bigger sample size means smaller variance. But also the smaller the bandwidth $h$ we choose, the bigger the variance.

The intuition for how bias and variance depend on $h$ is similar to that of a histogram: small $h$ is like thin bars in a histogram. The thinner the bars, the closer a histogram is, asymptotically, to a density (meaning less bias), but also the thinner the bars, the fewer observations we have in each bar, and so the higher variance.

If we want, we can choose $h$ to minimize the mean squared error (MSE), which equals bias squared plus variance:

$$\begin{aligned} \text{MSE} &\approx h^4 b^2(x) + \frac{1}{nh} v(x) \\ \text{best } h &\approx \frac{v(x)}{4b^2(x)} n^{-1/5} \end{aligned}$$

The best $h$ is found by minimizing the MSE: taking the derivative of the approximate MSE with respect to $h$, setting it equal to zero, and solving for the resulting $h$.

We were describing $h$ as a constant you choose. But notice that the best (minimum MSE) $h$ depends on $x$ and on $n$. In particular, the best $h \to 0$ as $n \to \infty$. Again, as with a histogram, the larger your sample size, the thinner you want to make your histogram bars be. But notice the best $h$ shrinks much more slowly than $n$ grows.

You can check that if we let $h$ be proportional to $n^{-1/5}$, then both the bias and variance of $\widehat{f_h}(x)$ go to zero in the limit as $n \to \infty$, making $\widehat{f_h}(x)$ converge in mean square to $f(x)$, and therefore $\widehat{f_h}(x)$ is a consistent estimator of $\widehat{f_h}(x)$.

How do we choose $h$? There are some rules of thumb (e.g., choose the $h$ that would be optimal if $f$ was normal. Silverman's rule is based on this idea). Or one could do two step estimation: choose a rule of thumb $h$, estimate $f(x)$, use that estimate to estimate $b(x)$ and $v(x)$, then reestimate $f(x)$ with the estimated optimal $h$.

### 5.3.4 Asymptotic Distribution

$\widehat{f_h}(x)$ is an average, so apply the Lindeberg Levy central limit theorem (CLT)

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$$n^{1/2}\left[\widehat{f_h}(x) - E\left(\widehat{f_h}(x)\right)\right] \xrightarrow{d} N\left[0, var\left(\frac{1}{h}K\left(\frac{x-X}{h}\right)\right)\right]$$

$$n^{1/2}\left[\widehat{f_h}(x) - f(x) - h^2 b(x)\right] \approx N\left[0, \frac{1}{h}v(x)\right]$$

If $h$ is a fixed constant, then asymptotically we have the bias term $h^2 b(x)$, so our usual confidence interval wouldn't be centered around the true $f(x)$.

But if we let $h$ be a function of $n$ (like using the formula for $h$ that minimizes the MSE), where $h \to 0$ as $n \to \infty$, then the variance term $v(x)/h$ blows up to infinity. To fix the problem, consider multiplying both sides by $h^{1/2}$

$$(nh)^{1/2}\left[\widehat{f_h}(x) - f(x) - h^2 b(x)\right] \approx N[0, v(x)]$$

Using a more complicated CLT, we can show that this formally convergences in distribution. Note that the rate of convergence now is not $n^{1/2}$, it's $(nh)^{1/2}$. We are CAN (consistent, asymptotically normal), but not root-n CAN.

What is the rate of convergence? The minimum MSE had $h$ proportional to $n^{-1/5}$, so $(nh)^{1/2}$ is proportional to $n^{2/5}$. Under some smoothness assumptions, $n^{2/5}$ is the fastest possible "rate of convergence" for nonparametric density estimation.

Note: for more rigor, we would need to keep track of and bound remainder terms, and use a CLT that allows $h$ to depend on $n$.

### 5.3.5 Extensions

- Type of convergence:

  Pointwise convergence (consistency): $plim|\widehat{f_h}(x) - f(x)| = 0$

  Uniform convergence (consistency): $plim \sup_x |\widehat{f_h}(x) - f(x)| = 0$

  We proved pointwise convergence, but kernel estimators can be shown to converge uniformly also. The uniform rate of convergence is slower than $n^{2/5}$ though.

- Density derivatives. Suppose we wanted to estimate the derivative of a density:

$$\frac{d\widehat{f_h}(x)}{dx} = \frac{1}{nh}\sum_{i=1}^{n}\frac{dK\left(\frac{x-X_i}{h}\right)}{dx} = \frac{1}{nh^2}\sum_{i=1}^{n} K'\left(\frac{x - X_i}{h}\right)$$

We can do all the same kinds of derivations we did before. Notice this has a $\frac{1}{nh^2}$ in front instead of $\frac{1}{nh}$. This will end up requiring that we shrink $h$ more slowly as $n \to \infty$, and so ends up having a slower optimal rate of convergence than $\widehat{f}_h(x)$, specifically, $n^{1/3}$ instead of $n^{2/5}$.

- Bias reduction

  We assumed that $f$ was twice continuously differentiable. If we assume $f$ is smoother (e.g., four times differentiable), then we can construct estimators that converge more quickly. Recall we had

$$\begin{aligned} E\left[\widehat{f}_h(x)\right] &\approx \int_{-\infty}^{\infty} K(u)\left[f(x) + uh\frac{\partial f(x)}{\partial x} + \frac{u^2h^2}{2}\frac{\partial^2 f(\tilde{x})}{\partial x}\right] du \\ &\approx f(x) + h^2 b(x) \end{aligned}$$

  Symmetry of $K(u)$ made $\int_{u=-\infty}^{\infty} uK(u)\,du = 0$, which made the first term in the expansion drop out, so the bias only depended on the second term. If you look at the formula for $b(x)$, you can see it contains $\int_{u=-\infty}^{\infty} u^2K(u)\,du$. Suppose we could choose a kernel $K(u)$ that made this integral equal zero? Then (assuming $f$ is smooth enough), we could do a higher order Taylor series expansion, and the $h^2$ term would also drop out. Moreover, symmetry will make $\int_{u=-\infty}^{\infty} u^3K(u)\,du$ equal zero, and so we could do a fourth order Taylor expansion, and get $bias \approx h^4B(x)$ for some function $B$. Variance is still $n^{-1}h^{-1}v(x)$, and best MSE is now rate $n^{-4/9}$. Definition: A kernel is defined to be $P$'th order kernel if $\int_{u=-\infty}^{\infty} u^pK(u)\,du = 0$ for all positive integers $p < P$. Our usual symmetric kernel function, which has $P = 2$, is a second order kernel. $P > 2$ is called a higher order kernel. Note that since $u^2$ is always positive for $u \neq 0$, having $\int_{u=-\infty}^{\infty} u^2K(u)\,du = 0$ (and therefore all higher order kernels) require that $K(u)$ be negative for some values of $u$. As a result, higher order kernels often lead to high variance, and typically behave poorly in empirical applications numerical applications unless the sample size $n$ is huge.

- Multivariate Density Estimation

  The joint density $f(y,x)$ of two random variables $Y$ and $X$ can be estimated as

$$\widehat{f}_h(y,x) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right)K\left(\frac{x - X_i}{h}\right)$$

  (this is analogous to a two dimensional histogram). We can estimate a $J$ dimensional joint density similarly for any integer $J$. When we do, the bias will still be proportional to $h^2$, but the variance will be proportional to $n^{-1}h^{-J}$. This then makes the optimal MSE choice of $h$ (no using higher order kernels) be proportional to $n^{-1/(J+4)}$, and corresponding optimal rate of convergence is then $n^{2/(J+4)}$.

- The "curse of dimensionality"

  For parametric models, the more parameters we estimate, the bigger the variance, but the rate of convergence is always $n^{1/2}$. But for nonparametric estimation, the higher the dimension of the function we estimate, the slower is the rate of convergence. To estimate a $J$ dimensional

function (without higher order kernels), the rate of convergence is $n^{2/(J+4)}$. This is the "curse of dimensionality."

For example, going from 2 to 4 variables in a linear regression only needs a modest amount of additional data for coefficients to stay significant. But nonparametrically estimating a joint density of 4 vs 2 variables requires tremendously more data to stay significant, since the rate of convergence drops from $n^{1/3}$ to $n^{1/4}$.

## 5.4 Nonparametric Kernel Regression

Suppose we have iid $(Y_i,\ X_i)$ draws from joint density $f(Y,X)$. Let $m(x) = E(Y \mid X = x)$ (reminder: this means the expected value of $Y$ when the random variable $X$ equals the particular value $x$.

Then by the definition of $m$, $Y = m(x) + e$ where $E(e \mid X = x) = 0$.

In linear regression, we assume $m$ is linear, i.e., $m(x) = a + bx$, so we only need to estimate two parameters $a$ and $b$. Or we might specify a polynomial like a quadratic, with $m(x) = a + bx + cx^2$. These are parametric models.

But suppose now we can't parameterize $m(x)$. We don't know what the functional form is. All we know is that $m(x)$ is a smooth function, e.g., twice continuously differentiable.

Our goal is to estimate the function $m(x)$ at any point $x$.

If $x$ is discretely distributed, then estimation of $m(x)$ is easy. We can just let

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i I(X_i = x)}{\sum_{i=1}^{n} I(X_i = x)}$$

The denominator here is the number of observations where $X_i = x$, and the numerator is the sum of $Y_i$'s for every observation $i$ that has$X_i = x$. So what this equation is doing is just calculating the average value of $Y_i$ for every observation $i$ in our sample that has $X_i = x$.

But suppose $x$ is continuously distributed. Then we'll see at most one data point that has $X_i = x$, and for most possible values of $x$, there won't be any observations that have $X_i = x$. So we can't use the above estimator. But what we can do instead is just look at all the observations that have $X_i$ close to $x$, and average the $Y_i$'s for those observations. The logic is that, for a small value $h$,

$$m(x) = E(Y \mid X = x) \approx E(Y \mid x - h \leq X \leq x + h).$$

This gives the following "local average" estimator:

$$\widehat{m}_h(x) \approx \frac{\sum_{i=1}^{n} Y_i I(x - h \leq X_i \leq x + h)}{\sum_{i=1}^{n} I(x - h \leq X_i \leq x + h)} = \frac{\sum_{i=1}^{n} Y_i I\left(\left|\frac{x - X_i}{h}\right| \leq 1\right)}{\sum_{i=1}^{n} I\left(\left|\frac{x - X_i}{h}\right| \leq 1\right)}$$

The numerator here is the sum of $Y_i$ for all observations that have $X_i$ within the distance $h$ of $x$, and the denominator is number of such observations.

Like our kernel density estimator, we can replace the functions $I\left(\left|\frac{x - X_i}{h}\right| \leq 1\right)$ with a general kernel function $K()$. This then produces what's called the Nadayara-Watson kernel regression estimator:

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

This is just a weighted average of the $Y_i$'s where the weights are the kernel functions.

If we choose the $K$ function to have a single mode at zero, then the closer $X_i$ is to $x$, the larger will be the weight $K\left[\left(x - X_i\right)/h\right]$. And choosing $K$ to be symmetric means the weight an observation $i$ gets depends only on how close $X_i$ is to $x$, not which side of $x$ it is on.

Another way to derive this kernel regression estimator is to use the definition of the expectation in $E\left(Y \mid X = x\right)$, which depends on the density of $y$ and $x$, and replace those true densities with kernel density estimates:

$$\begin{aligned} m\left(x\right) &= E\left(Y \mid X = x\right) = \int_{y=-\infty}^{\infty} y f_{y|x}\left(y \mid x\right) dy \\ &= \int_{y=-\infty}^{\infty} y \frac{f_{y,x}\left(y, x\right)}{f_x\left(x\right)} dy = \frac{\int_{y=-\infty}^{\infty} y f_{y,x}\left(y, x\right) dy}{f_x\left(x\right)} \end{aligned}$$

$$\begin{aligned} \widehat{m}_h\left(x\right) &= \frac{\int_{y=-\infty}^{\infty} y \widehat{f}_{y,x}\left(y, x\right) dy}{\widehat{f}_x\left(x\right)} = \frac{\int_{y=-\infty}^{\infty} y \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{y-Y_i}{h}\right) K\left(\frac{x-X_i}{h}\right) dy}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \\ &= \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \left[\frac{1}{h} \int_{y=-\infty}^{\infty} y K\left(\frac{y-Y_i}{h}\right) dy\right]}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \\ &= \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \left[\frac{1}{h} \int_{u=-\infty}^{\infty} \left(hu + Y_i\right) K\left(u\right) h du\right]}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \end{aligned}$$

$$\text{and} \quad \frac{1}{h} \int_{u=-\infty}^{\infty} \left(hu + Y_i\right) K\left(u\right) h du = Y_i$$

so we get just

$$\widehat{m}_h\left(x\right) = \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

**Properties of kernel regressions:**

$$m\left(x\right) = E\left(Y \mid X = x\right), \qquad \widehat{m}\left(x\right) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

The bandwidth $h$ determines the smoothness of $\widehat{m}\left(x\right)$. It's roughly the size of the neighborhood around $x$ over which data are averaged.

If $h \to \infty$ then $\widehat{m}\left(x\right) \to \overline{Y}$. Complete averaging.

If $h \to 0$, then $\widehat{m}\left(x\right) = \frac{0}{0}$ when $x \neq X_i$ for any $i$, otherwise $\widehat{m}\left(x\right) \to Y_i$. No averaging.

With large $h$, $\widehat{m}\left(x\right)$ is close to a flat line at $\overline{Y}$. With a tiny $h$, $\widehat{m}\left(x\right)$ erratically jumps around, almost passing through each point $\left(Y_i, X_i\right)$ in the sample.

A good value of $h$ will give a function that is neither too jumpy nor too flat.

Just like in kernel density estimation case,

$$bias = E\left[\widehat{m}_h(x)\right] - m(x) \approx h^2 B(x)$$

for some $B(x)$, and

and

$$var\left[\widehat{m}_h(x)\right] \approx n^{-1}h^{-1}V(x)$$

for some $V(x)$. In particular,

$$V(x) = \left(\int_{u=-\infty}^{\infty} \left[K(u)\right]^2 du\right) E\left(e^2 \mid X = x\right) / f_x(x)$$

All the bias, variance, consistency, and limiting distribution analyses we did for kernel density estimation, we can similarly do for kernel regression estimation. In particular, we can get the limiting distribution

$$(nh)^{1/2}\left[\widehat{m}_h(x) - m(x) - h^2 B(x)\right] \approx N\left[0, V(x)\right].$$

As before the optimal (in MSE sense) $h$ is proportional to $n^{-1/5}$ so the resulting optimal rate $(nh)^{1/2}$ is $n^{2/5}$.

As before, these give pointwise confidence intervals, and it is also possible to calculate uniform confidence intervals.

As before, with more smoothness we can use higher order kernels to get faster rates of convergencer in theory, though again these usually don't work well numerically unless $n$ is huge.

**Multivariate kernel regression**

iid $(Y_i, X_i, Z_i)$ — draws from joint density $f(Y, X, Z)$
Let $m(X, Z) = E(Y \mid X, Z)$.

$$\widehat{m}_h(x, z) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right) K\left(\frac{z - Z_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) K\left(\frac{z - Z_i}{h}\right)}$$

For $Y$ conditional on a $J$ vector (above is $J = 2$) as with joint density estimation get bias still proportional to $h^2$, variance is proportional to $n^{-1}h^{-J}$, optimal MSE has $h$ proportional to $n^{-1/(J+4)}$, and optimal rate of convergence is $n^{2/(J+4)}$.

The "curse of dimensionality" is back.

**Bandwidth choice for kernel regression:**
Method of cross-validation.
If we chose $h$ to minimize sum of squared errors $\sum_{i=1}^n \left(Y_i - \widehat{m}_h(X_i)\right)^2$, would get $h = 0$.

Instead, we can used the following method, called cross-validation, to get a good value of $h$:
First, for a given bandwidth $h$, define $\widehat{m}_{hi}(x)$ to be the kernel regression of $Y$ on $X$, leaving out observation $i$. So, for example for $i = 1$ we get

$$\widehat{m}_{h1}(x) = \frac{\sum_{i=2}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=2}^n K\left(\frac{x - X_i}{h}\right)}$$

Similarly define $\widehat{m}_{h2}(x)$, $\widehat{m}_{h3}(x)$, etc.

Now, for each observation $i$, evaluate the 'leave $i$ out' kernel regression $\widehat{m}_{hi}(x)$ at the point $X_i$. The error in the kernel regression at the point $i$ is $Y_i - \widehat{m}_{hi}(X_i)$. Cross validation then says to choose $h$ to minimize the sum of squares of these errors. That is, we choose $h$ by

$$h = \arg\min \sum_{i=1}^{n} \left(Y_i - \widehat{m}_{hi}(X_i)\right)^2$$

Essentially, the idea is to use all the data except observation $i$ to predict observation $i$, and choose $h$ to minimize the sum of squares of these prediction errors.

- Extension: varying bandwidths

  Can let $h$ vary by $x$. Where data are sparse (e.g., in the tails of the density of $X$), choose large $h$. where data are dense, choose a smaller $h$ .

- Example: "k nearest neighbor" estimation:

  Pick an integer $k$. At each $x$, choose $h$ so that only the $k$ observations having $X_i$ closest to $x$ are used.

- Extension: regression derivatives:

  $\frac{d\widehat{m}_{hi}(x)}{dx}$ is an estimator of $\frac{dm(x)}{dx} = \frac{dE(Y|X=x)}{dx}$.

  If $Y$ and $X$ are log data, this is the expected elasticity of $Y$ with respect to $X$ at point $x$. Like density derivatives, nonparametric regression derivatives have a slower rate of convergence than the nonparametric regression itself.

- Extension: local linear estimation

  Kernel regression is equivalent to a weighted average regression of $Y$ on a constant, where the weights are the kernel function. In other words, for any $x$, the kernel regression estimator $\widehat{m}_h(x)$ at the point $x$ equals the $m$ that minimizes

  $$\sum_{i=1}^{n} (Y_i - m)^2 K\left(\frac{x - X_i}{h}\right)$$

  To see this, observe that the first order condition is

  $$\sum_{i=1}^{n} -2(Y_i - m) K\left(\frac{x - X_i}{h}\right) = 0$$

  Solving this expression for $m$ makes $m = \widehat{m}_h(x)$

Instead of regressing $Y$ on a constant, we could do a weighted least squares regression of $Y$ on a constant and on $X$, again using the kernel function as the weights. That is, for a given $x$, let $\widehat{M}(x)$ and $\widehat{D}(x)$ equal the constants $M$ and $D$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - M - (x - X_i) D\right)^2 K\left(\frac{x - X_i}{h}\right)$$

This is called the "local linear" estimator. One can then show that, if $h \to 0$ at the proper rate as $n \to \infty$, we get, $\widehat{M}(x) \xrightarrow{p} m(x)$ and $\widehat{D}(x) \xrightarrow{p} \frac{dm(x)}{dx}$.

- Local polynomial estimators:

  We can comparably construct "local quadratic" and higher order "local polynomial" estimators. With sufficient smoothness conditions, these can be shown to obtain rates of convergence that are the same as those from higher order kernels .

  Local linear and local polynomial regressions work especially well if the true model is close to linear or polynomial. They'e also convenient in that they automatically provide derivative estimates.

  Disadvantages of local linear and local polynomial regressions are that, like higher order kernels, they generally require larger sample sizes than ordinary kernel regressions to work well. They are also more complicated, and can be be more numerically unstable, such as being more sensitive to outliers.

## 5.5 Series, Sieves, and Neural Nets: Other Forms of Nonparametric Regression

Let $m(x) = E(Y \mid X = x)$, so $Y = m(x) + e$ where $E(e \mid X = x) = 0$.

Polynomial Series Estimator: Suppose $m(x)$ is a very smooth function, so that there exists constants $a_0$, $a_1$, $a_2$, ... such that $m(x) = \sum_{j=0}^{\infty} a_j x^j$. For example, if $m(x)$ is what's called an analytic function, then it is infinitely differentiable, and the $a_j$ coefficients can be the terms in an infinite order Taylor expansion of $m(x)$ around $x = 0$.

Consider estimating $m(x)$ by picking an integer $J$, and letting

$$\widehat{m}_J(x) = \sum_{j=0}^{J} \widehat{a}_j x^j$$

where the $\widehat{a}_j$ coefficients are estimated by an ordinary least squares regression of $Y_i$ on 1, $X_i$, $X_i^2$,...,$X_i^J$. So we are simply approximating $m(x)$ by a $J'th$ order polynomial. This is called a polynomial series estimator.

This estimator will be biased, because it leaves out all of the higher order terms in the expansion. But suppose we let $J$ get bigger as $n$ gets bigger, just like we let the bandwidth $h$ be a function of the sample size $n$ in kernel regression. In particular, suppose we let $J \to \infty$ but $n/J \to \infty$, so we are adding terms slowly as $n \to \infty$.

Then the average number of observations per coefficient we have, $n/J$, is growing to infinity, so we we can get $\widehat{m}_J(x)$ to consistently estimate $m(x)$. Like a bandwidth, we will want to choose a $J$ trades off bias and variance, e.g., we could choose $J$ by cross validation.

Careful interpreting the estimates! what each $\widehat{a}_j$ means depends on $J$. For example, when $J = 1$, $\widehat{m}_J(x) = \widehat{a}_0 + \widehat{a}_1 x$, so $\widehat{a}_1$ is the marginal effect of $x$ on $Y$. But when $J = 2$, then $\widehat{m}_J(x) = \widehat{a}_0 + \widehat{a}_1 x + \widehat{a}_2 x^2$, and $\widehat{a}_1$ is no longer the marginal effect of $x$ on $Y$, instead the estimated marginal effect becomes $d\widehat{m}_J(x)/dx = \widehat{a}_1 + 2\widehat{a}_2 x$.

Best to focus on objects like $\widehat{m}_J(x)$ or $\frac{d\widehat{m}_J(x)}{dx}$, not on each $\widehat{a}_j$.

- Why use a polynomial series estimator?
  - Matches what practitioners often do - fit a line if have a small data set, try adding quadratic terms with larger $n$., cubic with even more data, etc.
  - It's easy to do and understand: it's ordinary least squares regression.

Why not do series estimation?

- Asymptotic distribution theory is not as clean as for kernel estimation.
- In kernel regression, a small change in the bandwidth $h$ produces a small change in the resulting estimated function $\widehat{m}_h$. But in series estimation, a single change in $J$ to $J + 1$ can completely change the fitted curve (consider, e.g., the difference between a quadratic and a cubic function). So series estimators can be much more sensitive to choice of $J$ .

Extension: multivariate polynomial series regressions:

$$m(x, z) = E(Y \mid X = x, Z = z)$$

$$\widehat{m}_J(x) = \sum_{j=0}^{J} \sum_{k=0}^{J} \widehat{a}_{jk} x^j z^k$$

Extension: other series estimators:

$$\widehat{m}_J(x) = \sum_{j=0}^{J} \widehat{a}_j \phi_j(x)$$

$\phi_1(x)$, $\phi_2(x)$, $\phi_3(x)$,... are Fourier or other series (formally, basis functions that span the space).

Extension: Sieve estimators:

Let $A_1$, $A_2$,., $A_J$,.. be a sequence of vectors that get longer as $J$ increases.

Let $\Phi_J(A_J, x)$ for $J = 1, 2, ...$ be a sequence of functional forms that get more complicated as $J$ grows.

Assume there exists a sequence of values for $A_J$ such that

$$m(x) = \lim_{j \to \infty} \Phi_j (A_j, x)$$

Then we choose a value $J$ and let $\widehat{m}_J(x) = \Phi_j \left( \widehat{A}_j, x \right)$, where $\widehat{A}_j$ obtained by regressing (linear or nonlinear least squares) $Y_i$ on $\Phi_j (A_j, x)$.

Example: any series estimator is a special case of a sieve estimator. E.g., the polynomial Series $\widehat{m}_J(x) = \Phi_j \left( \widehat{A}_j, x \right) = \sum_{j=0}^{J} \widehat{a}_j x^j$.

Example of a sieve estimator: Neural Networks. This is what's called a single layer neural net:

$$\widehat{m}_J(x) = \sum_{j=0}^{J} \widehat{a}_j g \left( x' \widehat{b}_j \right)$$

And this a double layer neural net:

$$\widehat{m}_J(x) = \sum_{k=0}^{K} \widehat{c}_k g \left( \sum_{j=0}^{J} \widehat{a}_{jk} g \left( x' \widehat{b}_j \right) \right)$$

The function $g$ is a 'squasher' function (like arctan, or a distribution function) that maps the real line into a zero-one interval. Each $g$ function represents a brain neuron. The parameter vectors $a$, $b$, and $c$ are estimated by nonlinear least squares regression.

What people call 'training' or 'learning' in a neural networks is just updating the estimated nonlinear least squares coefficients when you get more data.

Why do something complicated like a neural net instead of just polynomials? One can show neural nets are less sensitive to overfitting and to outliers, and have some better properties for approximating messy looking functions.

Other more complicated statistical procedures like Random Forests and Machine Learning are mostly just various versions of nonparametric regressions.

# 6 Lecture 13. Causal Models and Treatment Effects

## 6.1 History of Identifying a Treatment Effect

Identification based on randomization: The fundamental idea is randomly assigning some people to get a treatment, and others not (the treatment group and the control group). The difference in average outcomes between the groups is an estimate of the causal effect of treatment.

Early work on making this idea rigorous: Jerzy Neyman (1923), David R. Cox (1958), Donald B. Rubin (1978), many others.

In contrast to random assignment, econometricians historically focused on cases where selection (who is treated or observed) and outcomes are correlated. Sources of correlation:

- Simultaneity as in Trygve Haavelmo (1943).
- Optimizing self selection as in Andrew D. Roy (1951).
- Survivorship bias as in Abraham Wald (1943) - treatment assignment is random, but sample attrition is correlated with outcomes (WW II planes are hit randomly, but only those hit in survivable locations return to be observed).
- General models where selection and outcomes are correlated - James J. Heckman (1978).

## 6.2 Treatment Effect (TE)

$Y$ is an observable outcome (e.g. a wage). $T$ is a treatment indicator: a binary variable that equals 1 if treated, 0 otherwise. (e.g., $T = 1$ if finished high school, otherwise $T = 0$).

The Rubin notation for potential outcomes is $Y(t)$, defined as what the outcome $Y$ would be if given treatment $T = t$. So $Y(1)$ is what your outcome would be if you were treated, $Y(0)$ is what your outcome would be if you were not treated.

By the definitions of $Y(t)$ and $T$, we have

$$Y = Y(1)T + Y(0)(1 - T) = Y(0) + [Y(1) - Y(0)]T \tag{1}$$

This holds because if you plug $T = 1$ in you get $Y = Y(1)$ and if you plug $T = 0$ in you get $Y = Y(0)$.

The treatment effect ($TE$ for short) for an individual is defined as

$$TE = Y(1) - Y(0).$$

For any individual, the TE is the difference in outcome he or she would experience between being treated and not being treated. So a person's TE is the difference between what his or her $Y$ would equal if assigned $T = 1$ versus getting $T = 0$.

$Y(1)$, $Y(0)$, $T$, and $TE$ are random variables, just like $Y$ itself is a random variable.
Let $y$ and $t$ denote a realization of $Y$ and $T$.

**Example:**

Suppose $T_i$ and $Y_i$ are the random variables denoting possible treatments and outcomes for individual $i$.

Then $t_i$ and $y_i$ would be the treatment and outcome that actually occurred for individual $i$.

Our data consists of realizations $t_i$ and $y_i$ for a sample of individuals $i$.

Imagine realizations of the random variables $Y_i(1)$ and $Y_i(0)$. We denote these $y_i(1)$ and $y_i(0)$.

$y_i(t_i) = y_i$ is the outcome for individual $i$ that actually happened

$y_i(1 - t_i)$ is called the realized *counterfactual outcome* for individual $i$.

Example: if $t_i = 1$, person $i$ was treated.

Then $y_i = y_i(1)$ is the outcome that actually happened.

In this case $y_i(0)$ is the realized counterfactual outcome - what would have happened if person $i$ were not treated.

**Problem:**

The primary obstacle to identifying treatment effects: we can't observe counterfactual realizations.

So, can't observe the realized treatment effect $y_i(1) - y_i(0)$ for any person $i$.

Instead, we only try to identify and estimate some kind of average of treatment effects across people.

## 6.3 Average Treatment Effect (ATE)

One common goal of program evaluation is estimation of the average treatment effect (ATE).

The ATE is the expected value of people's TE:

$$ATE = E(TE) = E[Y(1) - Y(0)]$$

If we could observe counterfactuals, then identification and estimation of the ATE would be easy. We could then just take a sample average of $y_i(1) - y_i(0)$ over every individual $i$ in some sample.

How do we deal with the fact that we can't observe counterfactuals? To answer, first consider things we know we can estimate.

Begin by dividing our sample into two groups: people who are treated and those who aren't. For now, let's not worry about how it was decided who got treatment and who didn't. Consider the mean outcome in each of the two groups (which we can consistently estimate by taking the sample average of outcomes in the two groups):

The mean outcome of the treatment group is:

$$E(Y \mid T=1) = E(Y(1) \mid T=1)$$

and the mean outcome of the control group is:

$$E(Y \mid T=0) = E(Y(0) \mid T=0)$$

Differencing them gives:

$$E(Y \mid T=1) - E(Y \mid T=0) = E(Y(1) \mid T=1) - E(Y(0) \mid T=0)$$

This corresponds to the obvious estimator for the ATE (average outcome of the treated minus average outcome of the untreated). But as this equation shows, this doesn't automatically equal the ATE. That's because we need to think about how $T$ was determined.

Suppose that the following condition, called "unconditional mean unconfoundedness" holds:

$$E(Y(0) \mid T=0) = E(Y(0)) \qquad \text{and} \qquad E(Y(1) \mid T=1) = E(Y(1))$$

This says, "The average value of $Y(0)$ among the untreated is the same as the average value of $Y(0)$ for everyone," and "The average value of $Y(1)$ among the treated is the same as the average value of $Y(1)$ for everyone."

If "unconditional mean unconfoundedness" holds, then

$$E(Y \mid T=1) - E(Y \mid T=0) = E[Y(1) - Y(0)] = ATE$$

Unconditional mean unconfoundedness is the minimal condition we need for the obvious estimator (mean of the treated group minus mean of the control group) to work.

Suppose treatment is randomly assigned. Then

$$(Y(0), Y(1)) \perp T$$

This says, "The random variable $T$ is independent of the random variables $Y(0)$ and $Y(1)$."

This condition is called "unconditional unconfoundedness."

This implies the weaker "unconditional mean unconfoundedness," which was what we needed.

The stronger "unconditional unconfoundedness" condition implies that we can estimate the ATE for any function of the outcome. E.g., we can get the ATE for Y = wages or for Y = ln(wages).

"unconditional unconfoundedness" basically means either treatment was really randomly assigned, or it was assigned in a way that's just as good as random. For example, if I gave treatment only to short people, that wouldn't be random. But if height is completely unrelated to the outcome $Y$, then it's just as good as random, meaning "unconditional unconfoundedness" holds.

Neither unconditional unconfoundedness nor mean unconditional unconfoundedness will hold if the decision of who is treated depends on expected outcomes conditional on treatment. For example, suppose treatment is only given to those people who are expected to benefit from treatment. Then these unconfoundedness conditions don't hold. If treatment is a medicine, and only people who think they might get sick buy and use the medicine, then unconfoundedness does not hold.

## 6.4 Conditional ATE (CATE)

Sometimes randomization happens only within groups.

Example: suppose $X = 1$ for high school graduates, $X = 0$ for dropouts.

Suppose we first sort people into graduates and dropouts. Then we randomly assign graduates to treatment and control groups, and separately randomly assign dropouts to treatment and control groups. Then

$$(Y(0), Y(1)) \perp T \mid X$$

This says, "The random variable $T$ is conditionally independent of the random variables $Y(0)$ and $Y(1)$, conditioning on the observed covariates $X$."

This condition is called "conditional unconfoundedness."

Given conditional unconfoundedness, the previous results hold, conditional on $X$, meaning the previous results hold for each group (i.e., each value of $X$) separately.

Define CATE (Conditional ATE) by $E[Y(1) - Y(0) \mid X] = CATE(X)$

We can estimate averages of treated and untreated within each group. That is, we can easily estimate

$$E(Y \mid T = 1, X) = E(Y(1) \mid T = 1, X)$$

and

$$E(Y \mid T = 0, X) = E(Y(0) \mid T = 0, X)$$

If unconfoundedness holds then

$$E(Y \mid T = 1, X) - E(Y \mid T = 0, X) = E[Y(1) - Y(0) \mid X] = CATE(X)$$

But do we want to know CATE? Sometimes. For example, if we're considering treatment like a job training program specifically for dropouts, then we want $CATE(0)$.

What about ATE? If we can identify and estimate $CATE(X)$ for each possible value of $X$, then we can average CATE across all $X$ values to get ATE: $ATE = E(CATE(X))$, where the expectation is over the random variable $X$.

Note that this expectation must be taken over the population distribution of $X$. Example, we might have collected data on 100 dropouts and 100 high school grads. In the US, about 3 times as many people graduate high school as dropout. So the average we take would need to be

$$ATE = \frac{CATE(0) + 3CATE(1)}{4}$$

Definition: the "Overlap condition" holds if, for every value $x$ that $X$ can take on, there are people in the treated group who have $X = x$, and there are people in the control group who have $X = x$.

To estimate ATE given CATE, we need the overlap condition to hold.

Example: if no dropouts were treated, we wouldn't be able to estimate $E(Y \mid T = 1, X = 0)$, so we couldn't get the CATE for dropouts, and so we couldn't estimate the ATE.

## 6.5 Estimating ATE

Given unconfoundedness and overlap, we can identify and consistently estimate ATE. The estimator based on the derivations above is

1. For each value $x$ that $X$ can take on, calculate the average outcome of all treated people $i$ who have $x_i = x$.

2. For each value $x$ that $X$ can take on, calculate the average outcome of all untreated (control group) people $j$ who have $x_j = x$.

3. Take the difference between the above averages to get CATE(x).

4. Average CATE(X) over the population distribution of X to get ATE.

This works for discretely distributed $X$, but we must modify it a bit for continuous $X$, by doing "local" averaging. That is, steps 1 and 2 can be replaced by nonparametric regressions of $Y$ on $X$ in the treated group, and in the control group.

Many other estimators also exist. Here are some:

## 6.6 Propensity score based estimation of ATE

The function $p(X) = E(T \mid X)$ is called the propensity score. It equals the probability of being treated, conditional on $X$. One possible estimator of ATE, based on the propensity score, is this:

1. Estimate (either parametrically or nonparametrically) the propensity score function $p(X)$.

2. Let $p_i = p(x_i)$ for each person $i$ in the sample. Note that each $p_i$ is an observed value of the random variable $P$.

3. Apply the previous estimator, using $P$ in place of $X$.

Advantage: $P$ is one dimensional, while $X$ may be high dimensional. All of the information about $X$ that one needs to condition on to estimate ATE turns out to be contained in $P(X)$.

Disadvantage: Need to estimate $p(X)$. Good if you have a parametric model for $p(X)$, (e.g., probit), but bad if that model is misspecified. Otherwise, you need to estimate $p(X)$ nonparametrically.

## 6.7 Estimation of ATE based on Matching

Many treatment effect estimators are based on the idea of matching treated and untreated individuals. The simplest version of matching is this:

1. Draw a value of $x$ randomly from the population distribution of $X$.

2. given that value $x$, randomly draw from the treatment group a person $i$ who has $x_i = x$, and randomly draw a person $j$ from the control group who has $x_j = x$.

3. For those people $i$ and $j$, calculate $y_i - y_j$.

4. Repeat steps 1, 2, and 3 many times. An estimate of ATE is the average of $y_i - y_j$ over all these draws.

Note that as described above, this will only work if $X$ is discrete. If $X$ is continuous, you need to modify this procedure, so that in each repeat you draw people who have an $x_i$ and $x_j$ that is within some bandwidth $h$ of each other. Inference on the resulting estimated ATE will then entail asymptotic theory like that of nonparametric regression, where the bandwidth shrinks to zero as the sample size grows.

## 6.8 Estimation of ATE based on Propensity score matching

It's possible to do matching based on propensity scores, as follows.

1. Estimate (either parametrically or nonparametrically) the propensity score function $p(X)$.

2. Let $p_i = p(x_i)$ for each person $i$ in the sample. Note that each $p_i$ is an observed value of the random variable $P$.

3. Apply the matching estimator for ATE, using $p_i$ and $p_j$ instead of $x_i$ and $x_j$.

Note that this only requires matching scalars $p_i$ and $p_j$, instead of vectors $x_i$ and $x_j$, but it does require first estimating the propensity score function.

## 6.9 Average Treatment Effect on the Treated (ATT)

The ATE, $E[Y(1) - Y(0)]$, answers the question: what would be the difference in mean outcomes if everyone were treated, vs if no one were treated.

A different question one could ask is, what is the average benefit of treatment just for those who are actually treated? This is called the ATT or ATOT: the average effect of treatment on the treated.

ATT is the expected value of the TE of people who get treatment:

$$\begin{aligned} ATT &= E(TE \mid T = 1) \\ &= E[Y(1) - Y(0) \mid T = 1] \\ &= E(Y \mid T = 1) - E(Y(0) \mid T = 1) \end{aligned}$$

Note here $E(Y(1) \mid T = 1) = E(Y \mid T = 1)$, because if $T = 1$, meaning that you're treated, then your outcome $Y$ is the same as your potential outcome if treated, $Y(1)$.

$E(Y \mid T = 1)$ is easy to identify. A consistent estimator is just the sample average of $Y$ over everyone in the treatment group.

The difficult part is $E(Y(0) \mid T = 1)$, this is the expectation of a counterfactual. We need some form unconfoundedness to identify and estimate this.

Example: if we assume $Y(0) \perp T$, or the weaker condition $E(Y(0) \mid T = 1) = E(Y(0))$, then the average outcome of the untreated will consistently estimate the counterfactual $E(Y(0) \mid T = 1)$.

Many of the ATE estimators we discussed can be suitably modified to estimate ATT instead.

## 6.10 Local Average Treatment Effect (LATE)

What if $T$ is not randomly assigned, but some other related variable $Z$ is randomly assigned?

Define compliers to be people for whom $T$ and $Z$ are the same random variable. A complier either:
has $T = Z = 1$, and would have had $T = 0$ if he had $Z = 0$, OR
has $T = Z = 0$, and would have had $T = 1$ if he had $Z = 1$.

LATE: Local Average Treatment Effect: $E[Y(1) - Y(0) \mid \text{being a complier}]$.

Imbens and Angrist (1984) showed that, under some conditions:

$$LATE = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(T \mid Z = 1) - E(T \mid Z = 0)}$$

For binary $Z$, the sample version of this formula (replacing the expectations with sample averages) is numerically identical to the coefficient of $T$ in an instrumental variables regression of $Y$ on a constant and on $T$, using the constant and $Z$ as the instruments.

Advantages: Often in economics treatments are not randomly assigned. But if we can find some variable that is correlated with treatment and is randomly assigned, we can use that as the instrument $Z$ and estimate LATE. For example, If we want to know the effect of a lower vs higher income on health $Y$, we can't randomly assign some people to be poorer and others richer (unless we the experimenters have a very large amount of money to give away!). That's $T$. But we can look at which were hit by an unexpected natural disaster and which ones weren't. That's $Z$.

Disadvantages: The definition of a complier depends on counterfactuals, so we can't know who the compliers are. If compliers are very different from other people, then knowing their ATE doesn't say much about the treatment effects in the general population.

## 6.11 Difference in Difference Estimation

Difference-in-difference (DiD) is a model where the treatment (e.g., an event like a natural disaster) happens that affects one group of people and not another. Let group one be the group of people who are treated (e.g., live where the natural disaster occurred), and group zero be the control group. We have some outcome $Y$. We create two dummy variables, $D$ and $T$:

$$D_i = \begin{cases} 1 & \text{for people } i \text{ in group 1} \\ 0 & \text{for people } i \text{ in group 0} \end{cases}$$

$$T_i = \begin{cases} 1 & \text{people } i \text{ who are observed after the event} \\ 0 & \text{people } i \text{ who are observed before the event} \end{cases}$$

The DiD model is then just a regression with these dummy variables and their interaction:

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 D_i + \beta_4 T_i D_i + e_i$$

$\beta_2$ is the average change in outcome (before vs after the event) in the control group.
$\beta_2 + \beta_4$ is the average change in outcome (before vs after the event) in the treatedl group.

DiD makes the "parallel trends" assumption: Suppose the average change in outcome $Y$ over time would have been the same in both groups, if neither had been treated.

If parallel trends is true, then $\beta_4$ equals the difference between the average outcome in the treated group, and what the average outcome in the treated group would have been, had they had not been treated. This is the ATT.

Let's look at this formally, using potential outcome notation. Drop the $i$ subscript for now. Suppose the treatment happens between time periods $t-1$ and $t$.

$Y_{t-1}$ = an individual's outcome in time $t-1$. Everyone is untreated.

$Y(0)_t$ = an individual's outcome in time $t$ if the individual is untreated.

$Y(1)_t$ = an individual's outcome in time $t$ if the individual is treated.

$D$ = indicator of whether an individual is in the treatment group or the control group.

$$\begin{aligned} ATT &= E\left[Y(1)_t - Y(0)_t \mid D=1\right] \\ &= E\left[Y(1)_t \mid D=1\right] - E\left[Y(0)_t \mid D=1\right] \end{aligned}$$

Now add and subtract $E\left[Y_{t-1} \mid D=1\right]$:

$$ATT = E\left[Y(1)_t - Y_{t-1} \mid D=1\right] - E\left[Y(0)_t - Y_{t-1} \mid D=1\right]$$

The parallel trends assumption is:

$$E\left[Y(0)_t - Y_{t-1} \mid D=1\right] = E\left[Y(0)_t - Y_{t-1} \mid D=0\right]$$

This says that if the treated group, $D=1$, had been untreated, they would have had the same expected change in outcome as the untreated group. Plug this in to the above ATT expression:

$$\begin{aligned} ATT &= E\left[Y(1)_t - Y_{t-1} \mid D=1\right] - E\left[Y(0)_t - Y_{t-1} \mid D=0\right] \\ &= E\left[Y_t - Y_{t-1} \mid D=1\right] - E\left[Y_t - Y_{t-1} \mid D=0\right] \end{aligned}$$

This shows why it's called difference in difference.

Now add back in the $i$ subscript. Observe that $Y_{i,t-1}$ corresponds to $T_i = 0$ and $Y_{it}$ corresponds to $T_i = 1$, so we can rewrite the regression

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 D_i + \beta_4 T_i D_i + e_i$$

as

$$Y_{i.t-1} = \beta_1 + \beta_3 D_i + e_{i,t-1} \quad \text{and} \quad Y_{it} = \beta_1 + \beta_2 + (\beta_3 + \beta_4) D_i + e_{it}$$

This implies

$$Y_{it} - Y_{i.t-1} = \beta_2 + \beta_4 D_i + e_{it} - e_{i,t-1}$$

And so, assuming $e_{it} - e_{i,t-1}$ has mean zero and is uncorrelated with $D_i$ (this is basically the same as parallel trends) we get

$$\begin{aligned} E\left[Y_t - Y_{t-1} \mid D=1\right] &= \beta_2 + \beta_4 \\ E\left[Y_t - Y_{t-1} \mid D=0\right] &= \beta_2 \end{aligned}$$

$$ATT = \beta_4$$

Note: The original way we wrote the regression,

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 D_i + \beta_4 T_i D_i + e_i$$

assumes that each individual is only observed once, either in time period $t-1$ or in time period $t$ (corresponding to $T_i$ being either zero or one). But more generally we might see the same person twice, once in each period. So we could instead write the model as

$$Y_{is} = \beta_1 + \beta_2 T_{is} + \beta_3 D_i + \beta_4 T_{is} D_i + e_{is}$$

where time period $s$ can be either $t-1$ or $t$, so $T_{i,t-1} = 0$ and $T_{it} = 1$.

If all individuals $i$ are observed in both time periods, then we can relax the error assumption by including an individual specific fixed effect $\theta_i$:

$$Y_{is} = \theta_i + \beta_1 + \beta_2 T_{is} + \beta_3 D_i + \beta_4 T_{is} D_i + e_{is}$$

Differencing over time gives

$$Y_{it} - Y_{i.t-1} = \beta_2 + \beta_4 D_i + e_{it} - e_{i,t-1}$$

So we can again estimate $\beta_4$.

## 6.12 Regression Discontinuity Design - RDD

$Y$ is an outcome
$X$ is a continuous random variable, called the "running" variable.
$c$ is a "cutoff," a threshold value of $X$.
$T$ is the binary treatment indicator.
In sharp RD, $T$ is one if $X \geq c$, else $T$ is zero.

Example: Goodman (2008):
$Y = 1$ if go to a 4 year public Massachusetts college, $Y = 0$ if go to a private 4 year college.
$X$ is score on the Massachusetts Comprehensive Assessment System (MCAS) test.
$T$ is qualify for a scholarship to a Massachusetts public college.
$T$ is one if and only if score $X$ exceeds a given cutoff $c$.

We want a treatment effect: The change in $Y$ from a change in $T$.

Problem 1: $T$ is not randomly assigned.

Problem 2: $Y$ depends on $T$, but also directly on $X$, and $T$ depends on $X$.

How to separate the effects of $T$ and $X$?

Idea: Exact test score depends on ability/education/knowledge, but also depends a little on luck.

Identifying Assumption: For someone just smart/educated enough to score around $X = c$, it's pure chance whether they score $X$ just above $c$, or $X$ just below $c$.

Solution: Compare average $Y$ of people with $X$ just above $c$ to those with $X$ just below $c$.

Is like random assignment to treatment, among people who all have pretty much the same ($X \approx c$) level of ability/education/knowledge.

Let $\tau = E[Y(1) - Y(0) \mid X = c]$ is ATE for people who are at the cutoff.

Our estimate (sharp) RD ATE $\widehat{\tau}$ is:
average $Y$ for people with grades at or slightly above $c$,
minus average $Y$ for people with grades slighly below $c$,

Formally:

$$\begin{aligned} \tau &= \lim_{\varepsilon \to 0} E[Y(1) \mid c \leq X \leq c + \varepsilon] - E[Y(0) \mid c - \varepsilon \leq X < c] \\ &= \lim_{\varepsilon \to 0} E[Y \mid c \leq X \leq c + \varepsilon] - E[Y \mid c - \varepsilon \leq X < c] \end{aligned}$$

So $\widehat{\tau} = \widehat{E}[Y \mid c \leq X \leq c + \varepsilon] - \widehat{E}[Y \mid c - \varepsilon \leq X < c]$ for small $\varepsilon$, where $\widehat{E}$ means estimated conditional expectation.

In Goodman (2008), $E(Y)$ is probability of going to a private Massachusetts college. $\widehat{\tau}$, the increase in this probability from qualifying for the scholarship (among people who are smart enough to just qualify), was 7.6%.

Sharp RD above assumed $T$ is one if and only if $X \geq c$.

What if some people got the scholarship with a lower score, and others didn't get it even with a higher score?

In this case, define $Z$ to be one if and only if $X \geq c$.

Now $T$ is not the same as $Z$, but it is highly correlated with $Z$.

Recall a complier is someone who either:
has $T = Z = 1$, and would have had $T = 0$ if he'd had $Z = 0$, OR
has $T = Z = 0$, and would have had $T = 1$ if he'd had $Z = 1$.

Fuzzy RD is the RD version of LATE.

The Fuzzy RD ATE $= E[Y(1) - Y(0) \mid$ being a complier and $X = c]$.

Under some conditions:

$$\text{Fuzzy RD ATE} = \frac{\lim_{\varepsilon \to 0} E[Y \mid c \leq X \leq c + \varepsilon] - E[Y \mid c - \varepsilon \leq X < c]}{\lim_{\varepsilon \to 0} E[T \mid c \leq X \leq c + \varepsilon] - E[T \mid c - \varepsilon \leq X < c]}$$