# Final Review

# ECON 1123

**Sebastian James**

**(Slides prepared by Bo Jackson and Hanley Chiang)**

# I. The Big Picture

- For much of this course, our aim has been to estimate the *causal effect* of X on Y
  - Causal effect: Effect that would be measured in an ideal randomized controlled experiment

- We choose an estimator $\hat{\beta}$, a function of the observations in our sample, to estimate the true effect $\beta$

- Every estimator has a *sampling distribution*
  - Over repeated samples, the estimator produces a distribution of estimates
  - This distribution of estimates has a mean, variance, etc.

- We want the sampling distribution of the estimator to have certain nice properties:
  - Unbiased: $E(\hat{\beta}) = \beta$. Over repeated samples (each of size $n$), the mean of our estimates would be the true population value $\beta$.
  - Consistent: $\hat{\beta}$ converges in probability to $\beta$. As $n$ increases, the distribution of $\hat{\beta}$ becomes narrower and narrower around $\beta$.
  - Small variance.

# II. Multiple Linear Regression

To estimate the multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K + u,$$

the Ordinary Least Squares (OLS) estimators

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$$

minimize the sum of squared residuals ("prediction mistakes").

- Usually, there is one *independent variable of interest* ($X_1$). The other regressors ($X_2$, $X_3$,…, $X_K$) are *control variables*.
- Which controls should you include? Any variable whose omission would cause *omitted variable bias* in $\hat{\beta}_1$:
  - The variable affects Y; and
  - The variable is correlated with $X_1$.

- If including a certain variable doesn't change the coefficient of interest, then excluding that variable won't cause OVB. Which model below would be your preferred model? Model 3

**TABLE 5.2** Results of Regressions of Test Scores on the Student-Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

**Dependent variable: Average test score in the district.**

| Regressor | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Student-teacher ratio ($X_1$) | −2.28** (0.52) | −1.10* (0.43) | −1.00** (0.27) | −1.31** (0.34) | −1.01** (0.27) |
| Percent English learners ($X_2$) | | −0.650** (0.031) | −0.122** (0.033) | −0.488** (0.030) | −0.130** (0.036) |
| Percent eligible for subsidized lunch ($X_3$) | | | −0.547** (0.024) | | −0.529** (0.038) |
| Percent on public income assistance ($X_4$) | | | | −0.790** (0.068) | 0.048 (0.059) |
| Intercept | 698.9** (10.4) | 686.0** (8.7) | 700.2** (5.6) | 698.0** (6.9) | 700.4** (5.5) |
| **Summary Statistics** | | | | | |
| *SER* | 18.58 | 14.46 | 9.08 | 11.65 | 9.08 |
| $\overline{R}^2$ | 0.049 | 0.424 | 0.773 | 0.626 | 0.773 |
| $n$ | 420.0 | 420.0 | 420.0 | 420.0 | 420.0 |

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

OLS Standard Errors

- Standard error of $\hat{\beta}_1$ is our estimate for the *spread of the sampling distribution* of $\hat{\beta}_1$
  - More spread → more uncertainty in our estimate
  - Standard error is proportional to $1/\sqrt{n}$

- Standard error formulas differ according to whether they account for possible *heteroskedasticity* in the error term
- Heteroskedasticity: $Var(u \mid X_1,..., X_K)$ is a function of the X's
  - Intuition: Divide the population into subpopulations such that members of same subpopulation have the same values for the X's. Heteroskedasticity means that the variance of the dependent variable differs by subpopulation.

<u>Single Hypotheses</u>

- Example: $TestScore = \beta_0 + \beta_1 \ln(SPC) + \beta_2 \ln(TPC) + u$
  where SPC = students per class, TPC = teachers per class
- To test $\beta_1 + \beta_2 = 0$, rearrange the equation to make $\beta_1 + \beta_2$ a coefficient on one of the variables:

$$TestScore = \beta_0 + (\beta_1 + \beta_2)\ln(SPC) + \beta_2[\ln(TPC) - \ln(SPC)] + u$$

- Run new regression, do t-test on coefficient of ln(SPC)

<u>Joint Hypotheses</u>: Requires the conjunction **and**:

- In the above example, to test hypothesis that student-teacher ratio does not matter at all, test $\beta_1 = 0$ <u>and</u> $\beta_2 = 0$
- q = # of "restrictions" = # of statements separated by *and*
- Use F-test. Doing multiple t-tests rejects the null too often.
- F-statistic is distributed as $\chi_q^2 / q$

## Internal Validity: When is OLS consistent and unbiased?

Boils down to one main question:

- Is $E(u \mid X_1,\ldots,X_K)=0$? Are the independent variables uncorrelated with the error term?
- ***Actually***, we only need the *variable of interest* to be uncorrelated with the error term after holding constant the control variables. Called *conditional mean independence*.

**Conditional Mean Independence**

If the variable of interest is uncorrelated with the error term after holding constant the control variables, then the OLS estimator for the causal effect of $X_1$ is consistent and unbiased.

<u>Threats to Internal Validity of Multiple Regression</u>

$E(u \mid X_1,\ldots,X_K) \neq 0$

1. Omitted Variable Bias from *Unobserved* Variables
   - To predict whether an unobserved omitted variable causes positive or negative bias in $\hat{\beta}_1$, figure out:

     (1) Does variable have (+) or (-) effect on Y?

     (2) Does variable have (+) or (-) correlation with $X_1$?

     If (1) and (2) have same sign → positive bias

     If (1) and (2) have different sign → negative bias

2. Simultaneous Causality Bias

3. Errors in variables bias: Measurement Error in the X's
   - Random measurement error in Y is <u>not</u> a threat

4. Sample Selection Bias

- Your sampling procedure systematically overlooks certain values of the dependent variable.
- Example from Problem Set 9:
  - If you exclude out-of-labor-force people, then you will understate the effect of alcoholism on job loss.
  - Why? The people whose work capacity is most affected by alcoholism drop out of the labor force, and therefore drop out of your sample!

5. Wrong functional form.

- Is a form of omitted variable bias -> you've left out variables that capture nonlinear relationships between X and Y
- Correctible by using nonlinear regression or probit/logit.

# III. Nonlinear Regression with OLS

Nonlinear Functions of Single Independent Variable

If the marginal effect of X on Y differs according to different values of X, then use either:

- Polynomials: $TestScore = \beta_0 + \beta_1 STR + \beta_2 STR^2 + u$
  - When STR=15, the marginal effect of STR on test score is $[\beta_0 + \beta_1(16) + \beta_2(16)^2] - [\beta_0 + \beta_1(15) + \beta_2(15)^2]$

OR

- Logarithms

1) $Y = \beta_0 + \beta_1 \ln(X) + u$ (X up by 1% → Y up by .01$\beta_1$ units)
2) $\ln(Y) = \beta_0 + \beta_1 X + u$ (X up by 1 unit → Y up by 100$\beta_1$%)
3) $\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$ (X up by 1% → Y up by $\beta_1$%)

- To interpret coefficients in log equations, only memorize:

$$\boxed{\text{1\% change in X is approximately a 0.01 change in ln(X).}}$$

- Choosing between polynomial or logarithmic functions:
  - Polynomials allow marginal effect of X to change sign; logarithms do not.

<u>Interactions Between Independent Variables</u>

- If the marginal effect of $X_1$ on Y differs according to values of *another* variable $X_2$, then include interaction term $X_1 \times X_2$.
- Example: Let SouthCA = 1 if district is in Southern California

$$\boxed{TestScore = \beta_0 + \beta_1 STR + \beta_2 SouthCA + \beta_3 (STR \times SouthCA) + u}$$

  - $\beta_1$ = effect of STR in Northern CA (set SouthCA=0)
  - $\beta_1 + \beta_3$ = effect of STR in Southern CA (set SouthCA=1)
  - $\beta_3$ = *additional* effect of STR in south relative to north

# IV. Panel Data Regression

Cross-sectional data is a snapshot of n entities and panel data is a sequence of snapshots over time. We describe panel data by using two subscripts: often i for entities (individuals, firms, states, etc.) and t for time. N entities each time can be different or same. If follows up the same entities over time. We often call it longitudinal data.

In a panel setting the regression error often has the form

$$error_{it} = \alpha_i + u_{it}$$

Regarding the forms of error term, there are three different situations we handle differently

- $\alpha_i$ is zero. For example, if we have different car models for each year during 5-year period and pool them together, there is no common component across time. The pooled data (stack data over time) OLS can give us unbiased estimates.

- $\alpha_i$ is not zero, and it is correlated with X, we can use fixed-effect methodology. When we only have two-time period, fixed-effect is equivalent with differencing strategy (also called "before and after" methodology)

- $\alpha_i$ is not zero, but it is not correlated with X. Not really covered in this course. Random Effects or Clustered SE

Note that in a fixed effects model it is possible that $Cov(u_{i,t}, u_{i,t-1}) \neq 0$
In this case we should use clustered standard errors even though we have included fixed effects.

We focus on the second situation ($\alpha_i$ is not zero, and it is correlated with X). The goal is to deal with the omitted factor $\alpha_i$ that is constant within an entity but doesn't vary over time. (example: a persons gender, race, or country of birth does not change over time)

*A fixed effects regression uses the variation within an entity to make identification in order to control for unobserved differences across entities, rather than comparing apples and oranges. (Ex: compare Arizona today to Arizona tomorrow rather than comparing Arizona to New York)

The aim of a fixed effects estimator is to remove (or estimate) this linear entitiy specific fixed effect.

A simple case with two X variables:
First, differencing:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + error_{it}$$
$$Y_{i,t-1} = \beta_0 + \beta_1 X_{1,i,t-1} + \beta_2 X_{2,i,t-1} + error_{i,t-1}$$
$$Y_{it} - Y_{i,t-1} = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + error_{it} - (\beta_0 + \beta_1 X_{1,i,t-1} + \beta_2 X_{2,i,t-1} + error_{i,t-1}) = \beta_1 (X_{1,it} - X_{2,i,t-1}) +$$
$$\beta_1 (X_{2,it} - X_{2,i,t-1}) + (error_{it} - error_{i,t-1})$$

This is differencing methodology – when we regress change in Y on change in X, we cancel out the troublemaking term $\alpha_i$

$$error_{it} - error_{i,t-1} = (\alpha_i + u_{it}) - \left(\alpha_i + u_{i,t-1}\right) = u_{it} - u_{i,t-1}$$

So in OLS we had $\hat{\beta}_1 = \frac{\text{cov}(Xit, Yit)}{\text{var}(Xit)} + \frac{\text{cov}(Xit, \alpha i + uit)}{\text{var}(Xit)}$

Thus $E[\hat{\beta}_1] = \hat{\beta}_1 + \frac{\text{cov}(Xit, \alpha i + uit)}{\text{var}(Xit)}$

Since cov(αi,Xit)=0 then $E[\hat{\beta}_1] = \beta_1 + \frac{\text{cov}(Xit, \alpha i)}{\text{var}(Xit)}$

But in the differenced model the alpha is differenced away so the error term is just Δuit. Thus …

$$E[\hat{\beta}_1] = \beta_1 + \frac{\text{cov}(\Delta Xit, \Delta\alpha i + \Delta uit)}{\text{var}(\Delta Xit)} = \beta_1$$

**With more than two time periods…**

Instead of differencing, we use a fixed effects estimator [instead of removing the fixed effect, we estimate it using dummy variables] We know there is $\alpha_i$, but don't know what they are. To remove the bias term include a variable to identify each entity, so we regress Y on X controlling for n-1 entity dummies. <u>(recall dummy variable trap)</u>

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \delta_1 D_{1i} + \delta_2 D_{2i} + .... + \delta_{n-1} D_{n-1,i} + u_{it}$$

or you can write as

$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \beta_2 X_{2,it} + \alpha_i + u_{it}$, in which $\alpha_1,...,\alpha_n$ are treated as unknown intercepts to be estimated, one for each i.

For example, if i stands for state, the interpretation of $\alpha_i$ is a state-specific intercept. If the dependent variable is crime, then a large $\alpha_i$ means that state I has a lot of crime on average, and a low $\alpha_i$ means state I has less crime than average.

Also time fixed effects are useful in dealing with omitted variables that don't vary across entities yet vary over time.

*Note: One should use an F-test in order to determine the significance of the dummy variables.

**In summary, the fixed effects regression model**:

$Y_{it} = (\beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + ... + \beta_k X_{kit}) + (\gamma_1 D_1 + \gamma_2 D_2 + ... + \gamma_{N-1} D_{N-1}) + (\delta_1 B_1 + \delta_2 B_2 + ... + \delta_{T-1} B_{T-1}) + u_{it}$ for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$

$X_{2it}$ means variable $X_2$ for individual $i$ at time $t$.

$D_1, D_2, ...D_{N-1}$ are the individual dummies and the $\gamma_i$ coefficients are individual fixed effects.

$B_1, B_2, ...B_{T-1}$ are the time dummies and the $\delta_t$ coefficients are time fixed effects.

You use $N-1$ individual dummies and $T-1$ time dummies to avoid perfect multicollinearity. Also, you cannot use any variable that stays constant for each individual or that varies linearly with time.

Sample Question:

We have data for 366 student-athletes from a large university for fall and spring semesters during one academic year. The research question is this: Do athletes perform more poorly in school during the semester their sport is in season?

Y: GPA
X's: spring, SAT, hsperc(high school percentile), female, black, white, frstsem, season
We pool data across semesters and use OLS to get the following estimation:
GPA=-1.75-0.058spring + 0.0017sat-0.0087 hsperc +0.35female -0.254black+0.023white -0.027 season

standard errors for each coefficients are 0.35, 0.48, 0.00015, 0.001, 0.052, 0.123, 0.117, and 0.0049.

(1) Explain the coefficient of season.
(Other things fixed, an athlete's team GPA is about 0.27 points lower when his/her sport is in season. The estimate is statistically significant. )

Now suppose ability is correlated with season, then you can use fixed effects to solve the problem (to control for the unobserved ability).

(2) Now if we used a fixed effects estimator, which variables will drop out?

(variables that don't vary over time will drop out, including sat, hsperc, female, black, and white).

(3) If we simply included a dummy variable for each student will the standard errors in part(2) be valid?

(there may exist serial correlation between the error terms for each individual)

(4) What can we do to address the problem in part(3)?

The issue in part(3) is that the error term may be correlated across time. The best way to address this is to use HAC standard errors.

(5) Can you think of one potential internal threat after using differencing strategy (or individual fixed effect) – individual fixed effect deals with individual-specific and time-constant omitted variable, but one or more potentially important, time-varying variables might be omitted from the analysis. This will bias our estimation. One example is the measure of course load. If some fraction of student-athletes take a lighter load during the season, then term GPA may tend to be higher, other things equal. This would bias the results away from finding an effect of season on term GPA.

# V. Binary Dependent Variables

## *1.1 Linear Probability Model*

$Y_i = \beta_0 + \beta_1 X_i + u_i$ $\quad$ $Y_i$: 0 or 1

$\beta_1$ = change in probability that $Y = 1$ for a given $\Delta x$:

$$\beta_1 = \frac{\Pr(Y=1 \mid X = x + \Delta x) - \Pr(Y=1 \mid X = x)}{\Delta x}$$

- Advantages:
  - simple to estimate and to interpret
  - inference is the same as for multiple regression (need heteroskedasticity robust standard errors)
- Disadvantages:
  - Does it make sense that the probability should be linear in X?
  - Predicted probabilities can be <0 or >1!

Why the probability is linear in X? How about an "S-curve", then it can satisfy:

- $0 \leq \Pr(Y = 1|X) \leq 1$ for all $X$
- $\Pr(Y = 1|X)$ to be increasing in $X$ (for $\beta_1>0$)

Solutions: Probit or Logit Model: Non-linear function bounded between zero and one

## *1.2 Probit Model and Logit Model*

- ***Probit regression*** models the probability that $Y$=1 using the cumulative standard normal distribution function, evaluated at $z = \beta_0 + \beta_1 X$:

  $\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$

  $\Phi$ is the cumulative normal distribution function.
- ***Logit regression*** models the probability that $Y$=1 using the logit distribution:

  $\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$ , where $F(\beta_0 + \beta_1 X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$.

  Why bother with logit if we have probit?
  - Historically, numerically convenient
  - In practice, very similar to probit
- Both Probit and Logit regression estimates use maximum likelihood
- Usual assumptions → consistent, efficient estimator
- For the Probit and Logit, you need to know how to calculate predicted values:
  - First, compute $z_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - Second, use cdf: $\Pr(Y_i \mid X_i) = \Phi(z_i)$ (Probit) or

    $\Pr(Y_i \mid X_i) = \frac{1}{1+e^{-z_i}}$ (Logit)

- Marginal Effect: since the cdf is a non-linear function, marginal effect depends on the level of X, if $\hat{\beta}_1>0$ → X has a positive effect on Y.

**Logit/Probit practice question** (From 2004 final, from Ashenfelter, Levine and Zimmerman, "Statistics and Econometrics")

Perhaps the welfare system creates a form of intergenerational dependency. Perhaps children born into households that receive welfare are more likely to receive welfare when they become adults. We have observations on families over time. Let $mother_i$ be a dummy equal to 1 if the mother received welfare, and zero otherwise. Let $daughter_i$ be a dummy equal 1 if the daughter received welfare when she attained adulthood. Recall the logit function $G(x)=\frac{1}{1+e^{-x}}$.

Consider the ols results

$$daughter_i = .069 + .161 mother_i + \text{residual}$$

with standard errors .005 and .020.

(a) Test the null that the mother's welfare status has no effect on the daughter's status.

[ANSWER: we calculate t-statistics for the coefficient of $mother_i$, .161/.020 = 8.05 > 1.96, so we reject the null, it suggests that the mother's welfare status does have effect on the daughter's status.]

(b) Next we estimate the logit model:

$$\Pr[daughter=1|mother]=G(-2.60+1.39mother)$$

with standard errors .076 and .135. Calculate $\Pr[daughter=1|mother=1]$- $\Pr[daughter=1|mother=0]$, the effect of the mother's welfare status on the daughter. Compare this to the ols results in part (a).

ANSWER:

$\Pr[daughter=1|mother=1] = G(-2.60+1.39 \text{ x } 1) = G(-1.21)$

$\Pr[daughter=1|mother=0] =G(-2.60)$

$\Pr[daughter=1|mother=1]$- $\Pr[daughter=1|mother=0] = \frac{1}{1+e^{-1.21}} - \frac{1}{1+e^{-2.60}} = 0.1606$

The effect of the mother's welfare status on the daughter from the logit model is almost the same as the ols results in part (a), which is 0.161. So we know the simple linear model actually makes a good approximation to the logit answer.

(b2) Do you think that this estimate is biased? How and Why?

[probably an upwards bias on the effect of having a mother on welfare since this is going to be correlated with socioeconomic status which is a determinant of being on welfare]

(c) Let $faminc_i$ denote the log income of the mother's family. We estimate the logit model

$$\Pr[daughter=1|mother]=G(-2.94+.855mother-.578faminc)$$

with standard errors .793, .165, .083. Is it still the case that the mother's welfare status affects whether her daughter is on welfare? Take an educated guess about the sign of the correlation between *faminc* and *mother*. Do you think they are positively or negatively correlated? Explain your answer.

ANSWER: Yes, the mother's welfare status still affects whether her daughter is on welfare, but the effect is lower. It is reasonable to guess that the correlation between faminc and mother is negative, since the lower the family income the more likely the mother would receive welfare benefits. Since -.578<0, so we can derive that if we omit *faminc*, the OVB >0, so we get higher estimates for variable *mother* in part b.

# VI. Instrumental Variables Regression

Motivation: Some threats to internal validity cannot be eliminated by adding more controls or running fixed effects regressions.

- Simultaneous causality bias
- Omitted variable bias from unobserved, time-varying variables
- Errors-in-variables bias

Setup: $Y = \beta_0 + \beta_1 X + \beta_2 W + u$

- X is an endogenous variable of interest: $Corr(X, u) \neq 0$
- W is an exogenous control variable: $Corr(W, u) = 0$
- Z will be denote the instrumental variable.

Instruments:

- There are many *reasons* why X varies from observation to observation. Some sources of variation are uncorrelated with the error term, while others are correlated with the error term.

- An instrument Z is a ***source of variation in X*** that is ***uncorrelated with omitted factors influencing Y.***

- Two conditions for a valid instrument:
  - **Instrument Relevance:** $Corr(Z, X) \neq 0$. Means that Z is a source of variation in X.
  - **Instrument Exogeneity:** $Corr(Z, u) = 0$. Means that Z does not directly influence Y (except by affecting X) and Z is uncorrelated with other omitted variables affecting Y.

Famous Instrument for Student-Teacher Ratio

- Hoxby (2000): Random fluctuations in the size of age groups within each school district (due to random timing of births)
  - Example: If there is random spike in the number of children born in 1995, then this district's student-teacher ratio for kindergarteners will be higher in 2000 than in 1999.

Mechanics of IV Regression: **Two Stage Least Squares**

1) Estimate **first-stage regression** by OLS:

$$X = \delta_0 + \delta_1 Z + \delta_2 W + v$$

2) Obtain $\hat{X}$, the predicted value of X from first-stage regression.
3) Estimate **second-stage regression** of Y on $\hat{X}$ and W by OLS.
4) Coefficient on $\hat{X}$ in second stage is the 2SLS estimator.

- 2SLS regression ***isolates the variation in X due to Z*** and only uses that variation to estimate the effect of X on Y.

## Assessing Internal Validity of IV Regression

Boils down to one question: Are your instruments valid?

## Testing Instrument Relevance

- In first-stage regression, do F-test for joint hypothesis that the coefficients on all of the instruments are equal to zero.

- F >10 → relevance satisfied.
- F < 10 → *weak instruments* problem. When instruments are weak, 2SLS estimator is biased even in large samples, and your confidence intervals and hypothesis tests will be invalid.
  - Possible solutions: Drop weakest instruments (if model is overidentified), find stronger instruments, or forget about estimating $\beta_1$ and just obtain confidence intervals through Anderson-Rubin method

<u>Testing Instrument Exogeneity:</u> Overidentifying Restrictions Test

- Can only use this test when model is overidentified:

| |
|---|
| # of instruments > # of endogeneous X's, i.e. |
| m > k |

- Intuition of Overidentifying Restrictions Test:
  - Run 2SLS with only k instruments ***<u>assumed</u>*** to be exogenous. By assumption, these estimates are "valid".
  - If the other m-k instruments are exogenous, they should not be correlated with residuals from "valid" regression.
- Actual mechanics of the test: See Lecture 17, page 25
  - If J-statistic is too far out in the tails of the $\chi^2_{m-k}$ distribution, then reject null that all instruments exogenous.
  - If reject, use reasoning to figure out which IV's invalid.
  - If fail to reject, you have not ***proven*** that all IV's are exogenous b/c you ***assumed*** k instruments were exogenous

# VII. Program Evaluation

- Aim: Study effects of policies, programs, interventions
- Two basic methods: Experiments and Quasi-Experiments

Experiments

- Treatment status often denoted by binary variable

| | |
|---|---|
| X | = 1 if treated |
| | = 0 if not treated |

- Differences Estimator: If X is randomly assigned, then difference in mean outcomes of treated and untreated groups, $\bar{Y}^{X=1} - \bar{Y}^{X=0}$, is unbiased estimator of treatment effect.
- Example: TN Class Size Experiment (X = 1 if small class size)
  - Estimator for treatment effect: Difference between avg test score in small classes and avg test score in regular classes.

Using Regression to Estimate Causal Effects in Experiments

- OLS Regression: $Y = \beta_0 + \beta_1 X + u$.
  - $\hat{\beta}_1$ = differences estimator.
- Often, control variables (W's) included as regressors. Why?

1) **Conditional Randomization**: Sometimes treatment is only randomly assigned ***conditional on*** certain factors. In TN experiment, treatment was randomized ***within*** schools.

- Both probability of treatment and unobserved determinants of test scores differ ***across*** schools, so $Corr(X, u) \neq 0$.
- **But** if school dummies are included, remaining variation in X is within schools, so $Corr(X, u) = 0$.
- Is another example of ***conditional mean independence***: X is uncorrelated with u only after conditioning on W's.

2) Including control variables reduces standard errors of $\hat{\beta}_1$.

3) Check for randomization: If X is truly randomized, including controls should not alter estimate of treatment effect.

<u>Potential Problems with Experiments</u> See pp. 377-382 in text.

- One easily resolvable problem: Partial compliance
  - If not everyone complies with his/her assignment, then estimate $Y = \beta_0 + \beta_1 X + u$ with 2SLS, where:

Variable of interest:
X = 1 if actually got treated
= 0 if actually did not get treated
Instrument:
Z = 1 if initially assigned to treatment group
= 0 if initially assigned to control group

Quasi-Experiments

- Treatment is determined by markets or governments – but in a way that mimics random assignment (i.e. treatment is independent of other factors influencing dependent variable)
- Two main methods for estimating treatment effects in quasi-experiments: 1) IV; 2) Differences-in-Differences

Using Instrumental Variables in Quasi-Experiments

- Some *part*, but not all, of the variation in treatment status mimics random assignment. The instrument is a source of variation in treatment status that is "as if" randomly assigned.
  - Going back to Hoxby (2000): Whether you are born in a year with a spike in births is "as if" randomly assigned.
  - To estimate effect of class size, only use variation in class size due to this "as if" randomly assigned instrument.

## Using Differences-in-Differences in Quasi-Experiments

- Motivation: Sometimes treated and untreated groups differ on unobservable characteristics influencing dependent variable.
  → Cannot just compare $\bar{Y}^{treated}$ and $\bar{Y}^{untreated}$

- Setup of Differences-in-Differences

| |
|---|
| Two periods (pre-treatment and post-treatment) |
| Two groups (treatment and control) |
| Pre-treatment Period: Neither group is treated |
| Post-treatment Period: Only treatment group is treated |

- Basic idea of diffs-in-diffs: If unobserved characteristics of treatment & control groups are not changing in different ways, then avg change in Y (from pre to post) should differ between the two groups ***only as a result of the treatment effect***.

- Diffs-in-Diffs Estimation: Use OLS to estimate

$$Y = \beta_0 + \beta_1 Post + \beta_2 TreatGrp + \beta_3 (Post \times TreatGrp) + u$$

Post = post period dummy; TreatGrp = treatment group dummy.

- $\hat{\beta}_3$ is the differences-in-differences estimator. Why?
  - $\beta_1$ = avg pre-to-post change in Y for control group
  - $\beta_1 + \beta_3$ = avg pre-to-post change in Y for treatment group
  - $\beta_3$ = difference in the changes: due to treatment effect!
- With panel data, estimate $\boxed{\Delta Y = \delta_0 + \delta_1 TreatGrp + u}$.
- Example: In 1986, tax refunds to workers were expanded for low-income <u>parents</u>, but not for <u>childless</u> adults. What is the effect of this tax change on probability being in labor force?

$$\Pr(LF = 1) = \Phi[\beta_0 + \beta_1 Post + \beta_2 Parent + \beta_3 (Post \times Parent)]$$

From Eissa and Liebman (1996):

Table IV
PROBIT RESULTS: CHILDREN VERSUS NO CHILDREN
DIFFERENT SUBSAMPLES

| Variables | Sample | | | | |
|---|---|---|---|---|---|
| | Less than high school (1) | High school (2) | Beyond high school (3) | Predicted earned income in EITC range (4) | Predicted earned income above EITC range (5) |
| Coefficient estimates | | | | | |
| *Kids* ($\gamma_0$) | −0.663 (.202) | −1.551 (.164) | −1.352 (.264) | −1.427 (.126) | −1.071 (.357) |
| *Post86* ($\gamma_1$) | −0.232 (.126) | −0.040 (.105) | 0.188 (.158) | −0.022 (.078) | −0.151 (.221) |
| *Kids × Post86* ($\gamma_2$) | *0.181 (.083)* | *0.103 (.062)* | *0.030 (.098)* | *0.137 (.049)* | *−0.048 (.119)* |
| Log likelihood | −5052 | −7723 | −3380 | −13845 | −2612 |
| Number of observations | 9354 | 26,229 | 31,514 | 51,535 | 15,562 |
| *Predicted participation response for treatment group* | *.061 (.024)* | *.026 (.014)* | *.004 (.011)* | *.036 (.012)* | *−.007 (.016)* |

Data are from survey years 1985–1987 and 1989–1991 of the March CPS. The dependent variable is labor force participation. It equals one if the woman worked at least one hour during the tax year. *Post86* equals one for tax years 1988, 1989, and 1990. *Kids* equals one if the tax filing unit contained at least one child. In addition to the variables shown, all regressions include all the variables from the specification in column (5) of Table III. Standard errors are in parentheses. Regressions are weighted with CPS March supplement weights.

Probit coefficient for treatment effect
Marginal effect derived from the probit coefficient

Heterogeneous Treatment Effects

$Y_i = \beta_0 + \beta_{1i} X_i + u_i$, where $\beta_{1i}$ = person $i$'s response to treatment

- **If we run OLS of Y on X, and X is not randomly assigned**, $\hat{\beta}$ estimates avg effect of treatment ***on the treated***. Why? We only observe treatment group's response; we don't observe how control group would have responded to treatment!
- **But if X is randomly assigned**, OLS also estimates ***average treatment effect*** in whole population. Why? Under random assignment, avg characteristics of treated are identical to avg characteristics of whole population!
- **If we instrument for X with an instrument Z**, the 2SLS coefficient on X estimates the treatment effect for people whose value of X is most affected by the instrument Z: called the ***local average treatment effect***.

# VIII. Time series

Topics

- Forecasting, AR&ADL models, Prediction errors, overfitting and Information Criteria
- Omitted trends and dynamic causal effects
- Newey-West standard errors
- Forecasting volatility

## 1 Forecasting

We are interested in using past data to forecast what will happen in the future. Absent any complicated theory we use statistical models to estimate the expected value of the variables of interest in the future. There are two types of models that are commonly used for forecasting - AR and ADL.

### 1.0.1 The Autoregressive model (AR)

This is just a statistical model that tries to predict the value of a variable using the past values of that variable. The simplest model is called AR(1). AR(1) means that we use only one lag to predict a variable:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

A more general model is the AR(p) model, where p stands for the number of lags used in the prediction. An AR(p) model looks:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_{p-1} Y_{t-p+1} + \beta_p Y_{t-p} + u_t$$

What is the prediction using an AR(1) model for future quantities?

$$E[Y_{t+1} \mid t] = \beta_0 + \beta_1 Y_t$$

This stems from the fact that our prediction for the error is zero since the expectation of the error under our standard assumption is zero.

We can go on and try to predict $Y_{t+2}$ :

$$E[Y_{t+2} \mid t] = \beta_0 + \beta_1 E[Y_{t+1} \mid t] + E[u_t \mid t]$$

$$E[Y_{t+2} \mid t] = \beta_0 + \beta_1 [\beta_0 + \beta_1 Y_t] = \beta_0 + \beta_1\beta_0 + \beta_1^2 Y_t$$

And so on... The reason this is our prediction is again because our prediction of all future u's is zero.

## 1.1 The Autoregressive Distributed Lags model (ADL)

We can predict variables using not only their lags but lags of other variables. The Simple autoregressive distributed lags model is ADL(1,1) which include one lag of the LHS variable and one lag of an additional variable:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \theta_1 X_{t-1} + u_t$$

for example: in predicting prices of M&M.s I might consider including past prices of M&M.s and the amount of advertisement last period (since it can change the demand for M&M.s today and hence impact prices). A generalization of this model to more periods is called ADL(p,q), where p is the number of lags of the LHS variable and q is the number of lags of the other predictor:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_{p-1} Y_{t-p+1} + \beta_p Y_{t-p} + \theta_1 X_{t-1} + \theta_{p-1} X_{t-p+1} + \theta_p X_{t-p} + u_t$$

Otherwise, the model are similar (the statistical inference, the calculation of predictions, etc.).

## 2.2 Prediction errors and overfitting

Suppose I did predict the prices using the equation above, what is the prediction error? Plug the relevant variables into the equations above and the prediction error we will make is (the difference between the actual value tomorrow and the forecast today of the value tomorrow):

$$Y_{t+1} - \widehat{Y_{t+1}} = \alpha + \beta_1 Y_t + u_{t+1} - \widehat{\alpha} - \widehat{\beta_1} Y_t = \left[\alpha - \widehat{\alpha} + (\beta_1 - \widehat{\beta_1}) Y_t\right] + [u_{t+1}]$$

You can see that the prediction error has two parts –

1. the first is related to how well we have estimated the model
2. the second stems from the idiosyncratic shocks.

(when we do an in-sample prediction our estimates of the parameters are the correct values for the sample so this estimation error is not of concern , but when we are making an out of sample prediction this matters)

This highlights the trade-off we face when we use statistical models to predict future values. If we include more variables in the prediction given the same data we might explain more of the shock $u_{t+1}$ but we will also have harder time estimating the parameters and we will make more mistakes. Another way to look at this, is that when we include more variables we might get a better fit in the sample but we can make more mistakes out of the sample.

(For example if we just regress on junk it will probably explain some of the sample data, but will do very badly if we used this junk to predict the future)

2.3 Number of lags and the information criteria

As a result of the overfitting problem we face we need to choose carefully how many lags to include. The tool that is widely used is the information criterion. We learned one such criterion; the BIC (**Bayes Information Criteria**). The idea of this procedure is to balance the two possible problems – including too many lags which leads to overfitting or not including enough lags and then leaving a large variation of the LHS variable unexplained.

The criteria is: $BIC(p) = \log(SSR(p)/T) + (p + 1)(\log (T))/T$

p is the number of lags SSR is the sum of squared residuals and T is the size of the sample (number of time periods). The first term will always decrease as we add lags. The second term will increase. This balances the two problems we face appropriately (**a proof is given in the appendix 12.5 in S&W**). We choose the number of lags that minimizes this quantity.

## 3 Omitted trends and dynamic causal effects (A side note)

One phenomenon in time series data is that we might observe a high correlation between two time series but there is no actual causal effect in the relationship between the two variables. The problem is that there is some trend in the series over time due to factors that are unrelated. However, when we measure the correlation we find a strikingly strong relationship (which can quickly disappear with more data). This is called spurious regressions and we need to be aware of this phenomenon. The way to address this problem is to add a time trend and try to see if the variables are related to each other when we actually look at only the variation around the trend.

### Dynamic Causal Effects

When we consider the relationship between two series over time, we might not only be interested in the contemporaneous relationship but also in how past values of other variables impact the LHS variable of interest (like an ADL model but without the autoregressive part and with causality and not forecasting in mind).

For example, a model with dynamic causal effects is:

$$Y_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + u_t$$

A shock to the values of $X_{t-1}$ will impact the value of Yt and if it is persistent it will also have an effect through $X_t$ (see example in the practice question). You should know how to test for dynamic causal effect and how to interpret the impact of changes in the variables.

4 Newey-West Standard Errors [ HAC standard errors]

We have usually assumed that when we run an OLS (or an IV) regression we might have heteroskedasticity but that

$cov(u_i, u_j) = 0$

The problem with this assumption in time series is that we actually think that many factors the determine the dependent variable and are not observable (in the error) are persistent. Hence, we usually should check carefully if this assumption is reasonable. If not our measure of the variance is biased and any inference we try make using the S.E is wrong.

The Newey-West standard errors or HAC is a tool that helps us address this problem. These standard errors account for the serial correlation in the errors by actually estimating these correlations. However, if we try to estimate all possible correlations between all possible lags of the errors we run into the problem that we don't have enough observations to do that accurately. The formula for the Newey-West standard errors takes that into account

A. It limits the number of lags actually used (depending on the size of the data set) - we can use a rule of thumb for that. [Recall the rule of thumb formula for the truncation parameter]

B. It puts more weight on lags that are closer in time for which we have more observations.

## 5 Conditional Heteroskedasticity Models (ARCH and GARCH)

In many time series we notice the volatility is clustered - some times are more volatile and others are less volatile. That is, we can observe a number of subsequent years/periods with high volatility and some subsequent periods with low volatility. This is called volatility clustering or conditional heteroskedasticity.

There are two commonly used models to predict volatility. Suppose we want to predict the volatility of the stock market and we estimate a regression of stock returns and get some measure of the errors - for example, suppose I run an ADL(1,1) model of returns on past returns and past average price earnings ratios:

$$R_t = \beta_0 + \beta_1 R_{t-1} + \theta_1 PE_{t-1} + u_t$$

and that we can't predict the returns (we find that $\beta_1 = \theta_1 = 0$). But, we can try to check if we can predict the volatility. The ARCH model tries to predict the conditional variance of ut, call it $\sigma^2_t$ using the values of past squared residuals

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_p u_{t-p}^2$$

which is an ARCH(p) model.

The GARCH(p,q) model is a generalization of the ARCH(p) model and uses the lags of the variance in addition to the squared errors as predictors:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \ldots + \alpha_p u_{t-p}^2 + \theta_1 \sigma_{t-1}^2 + \theta_2 \sigma_{t-2}^2 + \ldots + \theta_q \sigma_{t-q}^2$$

If the coefficients are statistically significant then we conclude that the errors are conditionally heteroskedastic. We can calculate the predicted values using the coefficients estimated and the residuals from the regression of the stock returns. The coefficients are estimated using a Maximum likelihood estimation procedure of the time series regression together with the prediction for the volatility (we use a likelihood function that incorporates the equation for the variance as the variance of the errors)

**Practice question on time series (from the fall 2004 .nal):**

Suppose we want to study the effect of air pollution on health. We focus specifically on the effect of total suspended pediculates (TSP) on the infant mortality rate (number of infant deaths per 100K live births). We have the following data - quarterly data on average national TSP concentration and national infant mortality 1970-1990. This data contains n=84 observations (21 years x 4 quarters per year). Consider the following model:

Infant Mortality $Rate_t = \beta_0 + \beta_1 TSP_t + \beta_2 TSP_{t-1} + \beta_3 TSP_{t-2} + \beta_4 TSP_{t-3} + \beta_5 TSP_{t-4} + u_t$

(t is the quarter of the observation).

1) What is the number of observations available for estimation of the regression by OLS? (80)

2) Suppose TSP falls by 1 unit in the current quarter, and that this reduction is permanent. What is the expected reduction in mortality rate, 2 quarters hence (t + 2)?
$(\beta_1 + \beta_2 + \beta_3)$

3) Suppose someone wanted to estimate the regression with OLS and asked your advice about how to compute the standard errors. What advice would you give, why?
(Newey-West S.E.)

3a) How many lags would you use to estimate the standard errors. i.e. What is the truncation parameter you would use?
(Use rule of thumb $m = 0.75T^{1/3} = 0.75*80^{\wedge}(1/3) = 3.23$ thus 3 or 4 lags.)

4) The estimates indicate a very large effect of TSP on infant mortality – the decline in TSP from 1970-1990 explains almost all of the decline in the infant mortality (and significant at the 1% level). In your judgment, would this be compelling evidence that reductions in TSP produced the fall in infant mortality?
(No! Spurious regression)

5) Suppose you are sure that the causality is correct. Furthermore, you get an R2 = 0.95 in the above regression. How confident are you that the

$$E[\text{Infant Mortality Rate}_t] = \beta'_0 + \beta'_1 TSP_t + \beta'_2 TSP_{t-1} + \beta'_3 TSP_{t-2} + \beta'_4 TSP_{t-3} + \beta'_5 TSP_{t-4}$$

(We might be overfitting)