Lecture Contents

Lecture Contents

1. Intro, statistics
2. Statistics
3. Simple linear regression, OLS, method of moments
4. Properties of OLS, goodness of fit
5. Unit changes, SLR 1- 5, variance of β_1
6. Standard error, Gauss-Markov theorem (BLUE), Non-linear transformation
7. Omitted variable bias, MLR goodness of fit
8. MLR interpretation, properties, partialling out, MLR 1 - 4
9. MLR interpretation with different regressor choice, MLR 5, Variance of slope estimator, select regressors, quadratic transformation
10. Interaction terms, chi-squared, MLR 6
11. Confidence Intervals, hyphothesis testing

12. Statistical significance, test multiple restrictions, test linear coombinations of params
13. OLS Consistency, asymptotic distribution,
14. Dummy variables, differences across groups, dummy variable trap
15. More dummy variable trap, graphing, multiple sets of dummy variables, dummy log-linear, dummy for ordinal data, Linear Probability models
16. Heteeroskedasticity & robust SE, Robust standard errors e.g., testing for heteroskedasticity
17. Motivating WLS, WLS, FWLS
18. Measurement error, non-random samples, outliers
19. Time series models, assumptions, Time trends, prediction
20. Difference in difference, random effects and clustered errors
21. Fisrt differences and fixed effects, panel data e.g.
22. Instrumental variables,Estimatin IV with 2SLS
23. LPM revisited, Non-linear probability models

Statistics

> **Lec 1**

Joint probability density functions

f(x,y) = P(X = x, Y = y)

marginal pdf of X: f(x) = $\sum_y f(x,y)$

$\sum_x \sum_y f(x,y) = 1$

Conditional probability density functions

f(y | x) = f(y | X = x) = P (Y = y | X = x)

$\sum_y P(y|x) = 1$

> Lec 2

Population and sample mean

Expected value (population mean or population average)

A measure of the **central tendency** of the random variable X

E(X) = $\sum_x xf(X)$

E(X) = μ = μ_x highlight the fact as a population parameter

Properties:

- E(c) = c
- E(aX + b) = aE(X) + b
- E(X + Y) = E(X) + E(Y)

Sample mean

real world, estimate the population mean using the sample mean

$\overline{X} = \frac{1}{n} \sum_{i=1} nX_i$

Where X_i is a simple random sample and E(X_i) = μ

Properties:

- unbiased
- $\overline{b} = b$
- $\overline{aX} = a\overline{X}$
- $\overline{X = Y} = \overline{X} + \overline{Y}$
- $\sum_{i=1}^{n} (X_i - \overline{X}) = 0$

Check unbiased

Suppose μ is the population mean

check whether $E(\overline{X}) = \mu$

Population and sample variance

Population variance

$Var(X) = E(X-\mu)^2 = \sum x(x-\mu)^2 f(x) = E(X^2) - \mu_x^2$

Properties:

- Var(c) = 0
- $Var(aX+b) = a^2 Var(X)$
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- Variance of sample mean: assume $Var(X_i) = \sigma^2$, then $Var(\overline{x}) = \frac{\sigma^2}{n}$

The sample variance

$S_n^2 = \overline{X^2} - \overline{X}^2$, elegant biased

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, inelegant, unbiased

Notes:

- $S^2 = \frac{n}{n-1} S_n^2$
- $E(S_n^2) = \frac{n-1}{n}\sigma^2$
- $E(S^2) = \sigma^2$

Covariance and correlation

$Cov(X,Y) = E[(X - E(x))(Y - E(Y))]$

If X and Y are on the same side of their mean-> cov is positive

Properties:

- $Cov(X,X) = Var(X)$
- $Cov(a_1X + b_1, a_2Y + b_2) = a_1 a_2 Cov(X,Y)$

Correlation coeefficient, always between -1 and 1, unit free masurement of association

$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$

Property: $\rho(aX+b, Y) = \rho(X,Y)$

Reasons for correlation:

- Spurious relationship: variables are unrelated in the population
- Causal relationship: variables related in expected direction
- Reverse Causality: variables related but contrary to expected direction
- Simulataneity: causality in both directions
- Omitted variables: third factor affect multiple variables

Normal Distribution

$X \sim N(\mu, \sigma^2)$: x is normally distributed with mean μ and variance σ^2

Properties:

- depends only on mean and variance (mean + variance -> everything)
- Linear transformation -> still normal
- sum or normal r.v. is still normal $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2), Cov(X_1, X_2) = \sigma_1 2$ Let $Y = X_1 + X_2$, then $E(Y) = \mu_1 + \mu_2, Var(Y) = \sigma_1^2 + \sigma_2^2 + 2\sigma_1 2$
  $Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\sigma_1 2)$

Conditional Expectations

$E(Y|X = x) = \sum_y yP(Y = y|X = x)$

Properties:

- $E[b(X)|X] = b(X)$
- $E[Y + b(X)|X] = E[Y|X] + b(X)$
- $E[b(X)Y|X] = b(X)E[Y|X]$
- law of iterated expectations: $E[Y] = E[E[Y|X]]$

OLS Simple Linear Regression

> Lec 3

Intro

$y = \beta_0 + \beta_1 x + u$ u: error term or disturbance

Assumption: E(u|x) = 0 knowing x tells us nothing about the expected value of the error

Implications:

- E(u) = 0
- E(xu) = 0
- Cov(x,u) = 0

Conditional expectation: $E(y|x) = \beta_0 + \beta_1 x$ E(y|x = 0) = β_0 intercept parameter, constant term

$\frac{E(y|x)}{x} = \beta_1$ Sloppy parameter

Terminology: Y: dependent variable, explained variable, predicted variable, regressand X: independent variable, explanatory variable, predictor variable, regressor

OLS - Ordinary least squares

True line: $E(y_i|x_i) = \beta_0 + \beta_1 x_i$ True data not fall on the line: $y_i = \beta_0 + \beta_1 x_i + u_i$ Estimate the **fitted value**: $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ **Residual**: $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Minimizing the sum of squared residuals $min \sum_{i=1}^{n} \hat{u}^2 = min \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ Method: differentiating w.r.t $\hat{\beta}_0$ and $\hat{\beta}_1$ Solved to: $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} = \frac{Cov(\hat{x,y})}{Var\hat{(}x)}$

Note: variance > 0, so the sign of $\hat{\beta}_1$ depends on covariance between x and y

Method of moments

Using E(u) = E(ux) = 0, rewrite the equations with sample analogs unobserable population expectations are set equal to their sample analogs These equationas are then solved to derive estimators

> Lec 4

Properties of OLS

$\sum_{i=1}^{n} \hat{u_i} = 0$ -> $\overline{\hat{u}}$ = 0

$\sum_{i=1}^{n} x_i \hat{u_i} = 0$

$Cov(\hat{x, \hat{u}}) = 0$

$\overline{y} = \hat{\beta_0} + \hat{(\beta_1)}\overline{x}$ -> OLS regressioon line must go through the point $(\overline{x}, \overline{y})$

Goodness of fit

Measure how well the model fits the data

Total sum of squares (**SST**) : $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$ meeasure the total variability in y (opposed to the average)

Explained sum of squares (**SSE**): $SSE = \sum_{i=1}^{n}(\hat{y_i} - \overline{y})^2$ measure how much variability in y **is** explained by the regressor

Residual sum of squares (**SSR**): $SSR = \sum_{i=1}^{n} \hat{u_i}^2$ measure how much avriability in y **is not** explained by the regressor

$d_i = x_i - \overline{x}$

SST = SSE + SSR

$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ R-squared: the fraction of the total sample variation in y that is explained by x a possible measure for goodness of fit betweeen 0 and 1 Interpretation: $100 * R^2$ precent of the sample variation in y has been explained by x Not a subjective assessment orjudgement of the quality of the regression alone correlation coefficient: $\hat{\rho}^2 = R^2$

> Lec 5

Unit Changes

Interpretation is unaffected

R-squared is the same: the fraction of variation of y explained by x should not depend on unit, and R-squared is equal to the square of correlation coefficient, which is unit-free

Dependent variable - y

If $y_i' = cy_i$, then $\beta_1' = c\beta_1$, $\beta_0' = c\beta_0$

Independent variable - x

If $x_i' = cx_i$, then $\beta_1' = \beta_1/c$, $\beta_0' = \beta_0$

Evaluate estimators

Unbiasedness and variance (proof of unbiadness is in p7-8 on the handout)

SLR 1 - 5

SLR.1 (Linear in Parameters) The population model (true model) can be written as: **y = β0 + β1x + u** where β0 and β1 are the population intercept and slope parameters, respectively, and u is the unobservable random error.

SLR.2 (Random Sampling) We have a simple random sample of size n, {(xi,yi) : i = 1, 2, . . . , n}, following the population model defined in SLR.1.

SLR.3 (No Perfect Collinearity) The sample outcomes of x, namely {xi : i = 1, ..., n }are not all the same value.

SLR.4 (Zero Conditional Mean) The error term (u) has an expected value of zero given any value of the explanatory variable. **E(u|x) = 0**.

SLR.5 (Homoskedasticity) The error term (u) has the same variance given any 2 value of the explanatory variable. In other words, $Var(u|x) = \sigma^2$

SLR 1-4 -> unbiasedness

SLR 1- 5 -> Simple classical lenear regression model

Variance of OLS Slope estimator

$Var(\hat{\beta_1 | x_1, \ldots, x_n}) = \frac{\sigma^2}{SST_x} = \frac{\sigma^2}{nV\hat{ar}(x)}$

> Lec 6

Standard error

$Var(u|x) = \sigma^2$

$\hat{\sigma^2} = \frac{\sum_{i-1}^{n}(u_i = \overline{u})^2}{n-1}$ not able to proceed

Instead, use: $\hat{\sigma^2} = \frac{SSR}{n-2} = \frac{\sum_{i=1}^{n} \hat{u_i}^2}{n-2}$

Take root to achieve standard error

Gauss-Markov theorem

SLR 1- 5 -> BLUE (best linear unbiased estimator) Linear: linear function of the dependent variable (y) Best: minimum variance estimator

Proof in p3-4 in the handout

Non-linear transformations

here: regress ln(y) on x

Reasons:

- do not systematically over/under predict for some x
- make homoskedasticity more realsitic
- make the error term more normally distributed
- interpretation of the coefficients

Differences in interpreetation

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	β_1 = change/change
log-log	ln(y)	ln(x)	β_1 = percentage change / percentage change
level-log	y	ln(x)	$\beta_1/100$ = change/percentage change
log-level	ln(y)	x	$100 * \beta_1$ = percentage change/change

Lec 7

Omitted variable Bias

underspecified model and true model

Assume the true model is $y = \beta_0 + \beta_1 x + \beta_2 z + v$ a

nd satisfied SLR 1 - 4: variation in both regressors, simple random samples, zero conditional mean E(v|x,z) = 0

Run the underspecified model: $y = \beta_0 + \beta_1 x + u$

In underspecified model, we treated the omitted variable in the error

$/u = \beta_2 x_2 + v$ where x_2 is the ommited variable

SLR 1-3 satisfied,

E(u|x) = \beta_2E(z|x) + E(v|x)$

E(v|x) = 0 as E(v|x,z) = 0

Satisfies SLR 4 if E(z|x) = 0

i.e., whether knowing x does not help predict z (the omitted variable)

If the omitted variable belong in the true model and is correlated with x, then our estimator will be biased

Formula for the bias: Assume: $z = \delta_0 + delta_1 x + \epsilon$ Then: $E[\hat{\beta}_1 | x, z] = \beta_1 + \beta_2 \delta_1$ Special case:

- $\beta_2 = 0$ z is not an omitted variable
- $\delta_1 = 0$ Knowing x does not help predict z, ok to omit z

Signing:

bias = E(estimator) - parameter = $E(\hat{\beta}_1) - \beta_1 = \beta_2 \delta_1$

When β_2 (the relationship betweeen omitted variable and dependent variable) and δ_1 (the relationship between the omitted variable and the independent variable) has the same sign, the bias is positive

- Negative: downward bias $E(\hat{\beta}_1) < \beta_1$
- Positive: upward bias $E(\hat{\beta}_1) > \beta_1$

To eliminate the bias: run the multiple regression.

OLS Multiple Linear Regression

> Lec 7

$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + u_i$

Goodness of fit

the definitions of SST, SSE, SSR and R-squared are unchanged

Interpretation of R-squared: what proportion of the variation in y is explained by all our regressors

When adding a regressor, R-squared can not decrease and may increase (more regressors help us explain variations in y than before)

Similarly: after adding one regressor, $SSR_{new} \le SSR_{old}$

Problems of maximizing R-squared by adding regressors:

- too complicated, and may let the model become over-fitting
- Comes at a cost (in lec 8)
- estimate all the params with less precision -> every variable becomes less important

Solutions:

- Adjusted R-squared $\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$
  - when the #regressors increase -> k increases -> n-k-1 goes down
  - a race between the two effects (the R-squared can go in either direction)
  - Note: if thoose the highest adjusted R-squared, still have chances to pick the wrong model
- Pay less attention to goodness of fit

> Lec 8

Interpretation

$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$

β_0 is the estimated y for .... without.....

Thee expected y is estimated to be β_1 higher/lower per (unit) increase of x_1 **holding ..... constant**

Properties of OLS for multiple regression

$\sum_{i=1}^{n} \hat{u_i} = 0$

$Cov(\hat{x_i, \hat{u}}) = 0$

$\overline{y} = \hat{\beta_0} + \hat{\beta_1}\overline{x_1} + ... + \hat{\beta_k}\overline{x_k}$

$Cov(\hat{\hat{y}, \hat{u}}) = 0$ and $\sum_{i=1}^{n} \hat{y_i}\hat{u_i} = 0$ (covariance between thee fitted values and the redisuals is 0, applied to OLS for SLR)

Partialling out

Basic insight: modify x_1 so that it does not correelated with the other x's, then there would be no omitted variable bias

Run a regression of x_1 on all the other x's and divide it into two parts

- $x_1 = \alpha_0 + \alpha_2 x_2 + ... + \alpha_k x_k + r_1$ $
- x_1 = \hat{x_1} + \hat{r_1}$
- dependent variable = fitted value + residual
- Fitted value: x that can be explained by other x's
- **Residual**: x that **can not** be explained by other x's <- use this

Run the regression: $y = \beta_0 + \beta_1 \hat{r_1} + u$

Now, $\beta_1 = \frac{\sum_{i=1}^{n} \hat{r_{1i}} y_i}{\sum_{i=1}^{n} \hat{r_{1i}}^2}$ is also a multiple regression OLS slope estimator

Partialling out and multiple regression turn out to be **equivalent**

MLR 1 -4

MLR.1 (Linear in Parameters)

The population model (in other words, the true model) can be written as:

y=β0 +β1x1 +···+βkxk +u

where β0, . . . , βk are the population parameters and u is the unobservable random error.

MLR.2 (Random Sampling) We have a simple random sample of size n,

{(yi,x1i,...,xki): i = 1, 2, . . . , n}, following the population model defined in MLR.1.

MLR.3 (No Perfect Collinearity) In the sample, there are no exact linear relation-

ships among the independent variables (including the constant term).

MLR.4 (Zero Conditional Mean) The error term (u) has an expected value of zero

given any value of the explanatory variables. In other words, $E(u|x_1,...,x_k) = 0$.

MLR 1- 4 -> unbiasedness

> Lec 9

More about Interpretations

Suppose three regressors $x_1 + x_2 = x_3$, use two of them when running the regression to avoid perfect collinearity.

Models are interchangeable with whatever two are chosen, R-squared unaffected.

Interpretation changes - when using x_3 with either of the other two (e.g. use x_1, x_3): The estimate of y increases/decreases by ... unit when swapping one unit of x_2 with one unit of x_2 (as the total, x3, is held constant)

Choose the model based on interests: absolute or relative

MLR 5

MLR.5 (Homoskedasticity) The error term (u) has the same variance given any

2 value of the explanatory variables. In other words, $Var(u|x_1, ..., x_k) = \sigma^2$.

MLR 1 - 5 -> Classical Linear Regression Model, BLUE

Variance of slope estimator

$Var(\hat{\beta}_1|x) = \frac{\sigma^2}{SSR_1} = \frac{\sigma^2}{(1-R_1^2)SST_1} = \frac{\sigma^2}{(1-R^2)nV\hat{ar}(x_1)}$

- If $R_1^2 = 1$ perfect collinearity , MLR3 violated
- If R_1^2 is close to 1, MLR 3 not violated, but there is **Multicollinearity**
- If $R_1^2 = 0$, simple regression formula, MLR 3 is fine

Select regressors

Suppose selecting z as regressor, know z is correlated with x, but unsure whether it is relevant (bias=variance tradeoff)

- If include but it is irrelevant, increce the $Var(\beta_1)$
- If exclude but its is relevant, be biased estimates of β_1

Endogenous regressor E.g. of Lipitor,

- Model: liveExpec = $\beta_0 + \beta_1 LipitorDose + \beta_2 LDLcholesterol + u$
- Lipitor increase life expectancy by reducing LDL cholesterol (that reduce life expectancy)
- Interpret β_1: holding LDL cholesterol constant. In this way, the effect of Liptor is 0 LDL cholesterol becomes a endogenous regressor.
- Instead, run the model: liveExpec = $\beta_0 + \beta_1 LipitorDose + u$

Exclude the regressor if:

- sure enough that it is irrelevant
- sure enough that it is uncorrelated with other regressors
- descriptive analysis, uninterested in holding the variable constant
- endogenous variable

Non-linear transformations

here: quadratic

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$

$\frac{\partial y}{\partial x_1} = \beta_1 + 2\beta_2 x_1$

Interpretation:

- β_1: the slope when x1 = 0
- β_2:the change in the slope as x1 increase by one half unit
- Slope: return to x1

> Lec 10

Interaction Terms

Suppose the model: $econlbs = \beta_0 + \beta_1 ecoPrice + \beta_2 FamIncome + u$

The model does not allow to test: high income households are less responsive to price changes

Add interactive term:
$econlbs = \beta_0 + \beta_1 ecoPrice + \beta_2 FamIncome + \beta_3 ecoPrice * FamIncome + u$

Now we have: $\frac{\partial ecolbs}{\partial ecoPrice} = \beta_1 + \beta_3 famIncome$

$\frac{\partial ecolbs}{\partial famIncome} = \beta_2 + \beta_3 ecoPrice$

If we have $\frac{\partial ecolbs}{\partial ecoPrice} = -1.61 + 0.015 famIncome$, we know that household with higher income is less responsive to price increases.

Chi-squared distribution

z-distribution $z \sim N(0, 1)$

Possible outcomes: z-scores (the outcome is z-score standard deviatioons above the mean)

Chi-squared with one degree of freedom $z^2 \sim x_1^2$

$\sum z^2 \sim x^2_{number\ of\ squared\ normals\ summed\ up}$

$SSR/\sigma^2 \sim x^2_{n-k-1}$

MLR 6

MLR.6 (Normality) $u \sim N(0, \sigma^2)$

MLR 1 - 6: classical normal linear regression model

$\hat{\beta}_j \sim N(\beta_j, \frac{\sigma^2}{(1-R_j^2)SST_j})$

$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$

$se(\hat{\beta}_j) = (\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}})sd(\hat{\beta}_j)$

$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \frac{z}{\sqrt{\frac{x^2_{n-k-1}}{n-k-1}}} = \frac{standardnormal}{chi-squared_{n-k-1}/(n-k-1)}$

$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$

> Lec 11

Confidence Intervals

an interval estimator where we provide a range that's like ly to contain the true value instead of a point estimator

Want to estimate θ,assume $\hat{\theta} \sim N(\theta, \sigma^2_{\hat{\sigma}})$ unbiased, known variance, normal

$P(Lower\ cutoff < \theta < Upper\ cutoff) = 1 - \alpha$

- $1-\alpha$: confidence level
- α: significance level

Decide the cutoffs:

$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0,1)$

Use table: culmulative areas under the standard normal distribution

Find the area value closest to $\alpha/2$ and get the z value

Now we have: $Pr = (-c_{z,\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < c_{z,\alpha/2}) = 1-\alpha$,with probability $1-\alpha$, the true θ lies in $[\hat{\theta} - c_{z,\alpha/2}\ \sigma_{\hat{\theta}}\, , \hat{\theta} + c_{z,\alpha/2}\ \sigma_{\hat{\theta}}]$

Application: handout p4-8 for sample mean and regression params

If not normal but has t-distribution, use table: critical values of the t distribution, 2 tailed

Interpretation: the ..%. confidence interval estimator contains the true return to x ...% of the time; tofr this sample, our estimate of this interval is [...,...]

Hyphothesis testing

Errors:

- Type 1 error: reject a null that is true <- focused more
- Type 2 error: fail to reject a null that is false
- Move the threshoold higher -> harder to reject, more type 2 errors, less type 1 errors

Approach:

- decide on an acceptable leve lof type 1 error (significance leve, size of test, α)
- pick a threshold (critical value) s.t. the type 1 error is equal to α
- this threshold implies a certain level of type 2 error

Process: e.g. p10

1. assuming the null is true H0, generate a test statistic with a known distribution
2. draw a conclusion (reject or fail)
3. interpret

P(test statistic in the rejection region|H0 is true) = α

If the sign of the test statistics matter: use one tail

P-value:

- the probabiltiy that we'd observe a |test statistic| greater than we did assuming the null is true (in the two tails of the distribution);

- the lowest significant value at which we can reject

E.g. p13 - 16

> Lec 12

Statistical significance

- **Eonomic significance**: whether the association between x and y is large enough to be meaningful to the world, based on human
- **Statistical significance**: whether the association between y and x is likely to have arisen by chance

Possible to have one without the other

A regression coefficient is said to be statistically significant if run the two tailed test and reject the null with a 5% size of test

Test multiple restrictions

Joint hypothesis test

- H0: $\beta_0 = 1\ and\ \beta_1 = 0\ and\ \beta_3 = 0$
- H1: $\beta_0 \neq 1\ or\ \beta_1 \neq 0\ or\ \beta_3 \neq 0$, at least one of these coefficients is significant
- **Restricted model**: the model without regressors to be tested
- Unrestricted model: the originla model

Test based on the SSRs:

- If the H0 is true, the SSRs should be the same
- If the H0 is false, then the SSR for the unrstricted model should be smaller. i.e.,the unrestricted model should do a better job explaining y

Propose a new test statistic $\frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$

- q: restrictions being tested
- k: the total regressors in the unrestricted model

$\frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}$

Use the table: **5% critical values of the F distribution**

Test all of the regressors:

Use the test statistics: $\frac{R^2/k}{(1-R^2)/(n-k-1)}$

Test linear combinations of Params

Whether two regressors has the same impact on y

H0: $\beta_1 = \beta_2$

Method 1: modified F-test

H0: $\beta_1 - \beta_2 = 0$

test statistics: $\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)}}$

Method 2: transform the regression

Use x3 = x1 + x2, x1 = x3 - x2

Run the model: $y = \theta_0 + \theta_1 x_1 + \theta_2 x_3 + u$

Here we have: $\theta_1 = \beta_1 - \beta_2$

We can test with H0: $\theta_1 = 0$

Method 3: F-test for multiple restrictions

unrestricted model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

restricted model: $y = \theta_0 + \theta_1 x_3 + u$

Test with SSR and F test.

> Lec 13

Consistency

- large sample properties: n-> infinity, approximations
- small sample properties: for all n, always hold

Consistency: estimator -> parameter when n->infinity or plim(estimator) = parameter

Law of large numbers: $plim(\overline{Y}_n) = E(Y) = \mu$

The sample mean is a consistent estimator of the population mean

OLS consistency: $plim(\hat{\beta}_j) = \beta_j$ for all j

SLR 1-4 or MLR 1-4 -> OLS is consistent

Asymptotic distribution

Central limit theorem: $\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ ("a" on tilde, aymptotically distributed as, or conver to... when n->infinity)

MLR1 - 5 ->$\frac{\hat{\beta}_j-\beta_j}{se(\hat{\beta}_j)} \sim N(0,1)$

> Lec 14

Dummy variables

make qualitative information quantitative

e.g. $wage = \beta_0 + \beta_1 female + \beta_2 exper$

To test the gender gap in wage with the samle level of experience: H0: $\beta_1 = 0$

Differences across Groups

The previous model, the return to experiences is the same

Method 1: Use two separate regression models

To test **difference in the entire wage profile**

- $E(wage|female, exper) = \beta_0^f + \beta_1^f exper$ ->SSR1
- $E(wage|male, exper) = \beta_0^m + \beta_1^m exper$ ->SSR2
- SSRu = SSR1 + SSR2
- $wage = \beta_0 + \beta_1 exper$ -> SSRp = SSRr

Chow Test: $\frac{[SSRp-(SSR1+SSR2)/(k+1)]}{(SSR1+SSR2)/(n-2k-2)}$ tested with **F distribution** table

Not able to only test return to experience as we do not have covariance between $\beta_1^f \ and \ \beta_1^m$

Method 2: interaction term

To test **difference in the entire wage profile**

$wage = \beta_0 + \beta_1 female + \beta_2 exper + \beta_3 female * xexper$

Ho: $\beta_1 = 0 \ and \ \beta_3 = 0$

F-test

Test **only return to experience**, also use this model

H0: $\beta_3 = 0$

Dummy variable trap

male + female = constant -> perfect collinearity **identification problem**

1. have a **omitted group**,
   - the constant term represent the y for this group
   - slope estimators for other dummies represent the gaps between the dummy and the omitted group
2. do not have the contant term
   - each coefficients represent the estimation for the corresponding dummy

> Lec 15

Dummy variable trap e.g. p1-2, graphing p3-4

Multiple sets of Dummies

add interaction term to match coefficients and #group.

E.g. need 4 coefficient if there are two sets, and each set has two dummies (4 combinations) Otherwise, params are **implicitly restricted**

Can also generate a single categorical variable, break down into smaller groups

Log-linear models

$ln(wage) = \beta_0 + \beta_1 female + \beta_2 exper$

when take derivatives w.r.t. dummy (Female), thinking about small changes to a continuous variable, but not applied to dummy variables

- Approximation: β_1 * 100%
- Actually estimates: $(e^{\beta_1} - 1) * 100\%$

Ordinal Data

Before: cardinal

Ordinal: e.g. attractiveness: 1=homely, 2=plain, 3=average, r=attractive.....

Problem: .e.g. increasing from 1-2 is the same as 3-4, not what we want.

Solution: create dummies for each value L1, L2, L3, L4, L5

Maybe not create separate dummy for every category:

- some categories are too rare
- too many categories

LPM - linear probability model

Dummy dependent variable

Response probability: E(y|x) = Pr(y=1|x)

Interpretation: the probability of y instead of the expected value of y

> Lec 23

Drawbacks of LPM

1. always violate MLR5 homoskedasticity -> use robust SE
2. predicted probabilities can take on any value (<0 or >1)
    - probability are limited dependent variable
    - e.g. p2-4

Non-linear probability model

add a function that has an ragne from 0 to 1: **infinate domain and non-decreasing**

$$Pr(y_i \hat{=} 1|X) = G(\hat{\beta}_0 + \beta_1 x_i)$$

estimate using maximum likelihood

2 models are similar, interchangeable

Probit model

cdf of normal distribution

$$Pr(y_i \hat{=} 1|X) = \Phi(\hat{\beta}_0 + \beta_1 x_i)$$

Logit

$G(z) = \frac{exp(z)}{1+exp(z)}$

Interpretation

$\frac{\partial Pr(y=1|x)}{\partial x} = G'(\beta_0 + \beta_1 x)\beta_1$

β_1 can not be interpreted as a marginal effect

must estimate marginal effects depend on the value of x

common approact (stata): calculates the marginal effect at the sample means of the covariates

LPM estimates are often close to marginal effects from probit and login, but need to check

Decide on LPM or probit/logit

- is it fine to have some observation <0 or >1?
- estimate the marginal effect, sufficient to use LPM
- predict probabilities, use probit/logit

Buttom lines:

- Use robust SE
- concerned about nonsensical predicted probabiliteis, use probit/logit

Note: Maybe other cases, not dummy, where the dependent variable has a limited range, use the limited dependent variable models than OLS

> Lec 16

Heteroskedasticity

$Var(u|x) \neq \sigma^2$

Violate MLR 5-> not BLUE, no test statistics and condifence intervals

Robust standard errors

Solution: repeat measure variance of u for every x -> **robust standard errors**

$Var(\hat{\hat{\beta}}_1|x) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 u_i^2}{(\sum_{i=1}^{n} \hat{r}_{ij}^2)^2}$

Test statistics: $\frac{\hat{\beta}_j - \beta_j}{robust\_se(\hat{\beta}_j)}$

	distribution for test statistic for all n	... when n is large	divided by
MLR 1 - 6, sigma known	N(0,1)	N(0,1)	Sd
MLR 1 - 6	t_{n-k-1}	approximately N(0,1)	Se
MLR 1 - 5	Unknown	Approximately N(0,1) by CLT	Se
MLR 1 - 4	Unknown	approximately N(0,1) by CLT	robust se

Examples

robust SE usually larger than plain SE, coefficients are the same

If heteroskedastic, use robust SE

If homoskedastic, both are valid (consistent)

Testing heteroskedasticity

run the regression, get the residuals, square the residuals, regress the squared residuals on the regressors, a test of joint significance of all the regressors

Test statistics: $\frac{R^2_{\hat{u}^2}/k}{(1-R^2_{\hat{u}^2})/(n-k-1)}$

Reject -> heteroskedasticity, proceed as if there is heteroskedasticity, robust SE or FWLS

Fail to reject-> nothing can conclude, often taken as evidence that assuming homoskedasticity is reasonable (plain SE)

Breusch-Pagan

Test the linear regression, regress the squared residual on the original regressors

White Test

Also test the nonlinear relationship,

regress the squared residual on regressors, their squares, and a full set of interaction terms

WLS and FWLS

> Lec 17

WLS - weighted least squares

Heteroskedasticity -> OLS not blue, find a more efficient estimator

WLS: weight the observations according to their error variances,

- unbiased and consistent
- BLUE with heteroskedasticity

Idea: give more weights to those observations with lower error variance

Assume: we know the formula of variance of the error.

Weighted everything by **the square root of Var(u|x)**

WLS and OLS(robust SE) has similar coefficient and WLS has smaller errors (more efficient)

FWLS - feasible weighted least squares

WLS works if we know:

- there is heteroskedasticity
- exactly the form of Var(u|x)

However, Var(u|x) is usually unknown Var(u|x) = h(x)

process:

1. regress y on the x's and get the residuals
2. regress the logged squared residuals on x's and use the resulting exponentialed predicteed values as the estimates of the error variance
3. use the estimates of the error variance to perform WLS

Use logged and exponentialed... in step 2 to avoid negative estimates of the error variance

As we only reduced the extent of the problem but not eliminated it entirely, use rebust SE after implementing FWLS

FWLS is more efficient than OLS and OLS is easier to implement.

Data Issues

> Lec 18

Measurement Error

Dependent variable

Assume measurement error = observed value - true value: $e_0 = y - y^*$

The regression we are running: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (u + e_0)$

MLR 1-3 still fulfilled

MLR4: $E(u + e_0|x) = E(u|x) + E(e_0|x) = E(e_0|x)$

Question: if measuremenet error in y zero mean conditional on x?

If it is, then MLR 4 is fine; otherwise, MLR4 is violated

If $E(e_0|x) = 0$

- x does not tell use anything about the expected measurement error in y
- measurement error is just a new part of the error term
- OLS unbiased

If $E(e_0|x) \neq 0$

- measuremeent error becomes an omitted variable (correlated with regressors)
- OLS may be biased, direction unknown

Independeent variable

Assume measurement error = value observed - true value: $e_1 = x_1 - x_1^*$

The regression we are running: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (u - \beta_1 e_1)$

Assume the measurement error has zero mean conditional on the true value of the regressor

$E(e_1|x_1^*) = 0$ (Classical measureement error) Imply: $Cov(x_1^*, e_1) = 0$

However, the observed must be correlated: $Cov(x_1, e_1) = \sigma_{e_1}^2$

$$plin\hat{\beta}_1 = \beta_1(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2})$$

Ratio = variation in true x/variation in observed x (variation in true x1, variation due to the measurement error) = **signal/(signal + noise)**

- If all of the variation in x is due to the measurement erro, then this ration = 0
- If **none** of the variation in x is due to the measurement erro, then this ration = 1

expect β_1 to converge to a value between zero and truth.

- The estimator will no longer be conssitent
- This is called **attneuation bias** (bias toward zero)

If the meausrement error is related to the true x (not classical ME as assumed), the bias could go in either direction

Missing or Non-random samples

1. get biased result
2. conditional on the random data

Internal valid:good job estimating the returns to x on y

external validation: also applicable to other contexts

Outliers and LAD

Enormous impact on regression results

To fix: use Least Absolute Deviation (LAD)

- OLS give a lot of impact to the outliers
- taking the absolute rather than squared residuals moderates their impact
- Calculation is harder

OLS Time Series Models

> Lec 19

Model

Temporally ordered, all observations are associated with the same sampling unit

e.g. inflation and employment in different years

Static model: $y = \beta_0 + \beta_1 x_t + u_t$

Finite distributed lag model of order 1: $y = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + u_t$

Finite distributed lag model of order 2: $y = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + u_t$

->approprate when the independent variable has a lagged effect on the dependent variable

Temporary change: goes up some unit in a year, and go back in the year after

- impact propensity: β_1
- Estimator represents the lag

Permanent Change: change in a year and stays

- Long-run propensity: $\beta_1 + \beta_2 + \beta_3$

Static model can not measure the impact of time lag.

- With lagged regressors, more lag -> increase chances of multicollearity in some cases
- e.g. regress feretility on personal tax exemption, when we use lagged regressors, as the tax policies does not change a lot, there are large chances to have the problem of multicollinearity

TS 1-5

TS.1 (linear in Parameters) $y = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + u_t$

TS.2 (No perfect collinearity)

TS.3 (Zero conditional mean) $E(u_t|X) = 0$

- x's are contemporaneously exogenous
- across-observation part $E[u_t|\{(x_{1s}, ..., x_{ks}) for\ all\ s \neq t\}] = 0$
- x's are strictly exogenous -> include too few lags, violate TS.3
- **strict exogeneity assumption**

TS.4 (Homoskedasticity)

TS.5 (No serial correlation) $Corr(u_t, u_s|X) = 0$ for all t not equal to s

TS.6 (Normality) the errors are independent of X and are i.i.d. as Normal(0, σ^2)

TS 1- 3 -> OLS is unbiased

TS 1-5 ->BLUE

Time Trends

omitted variable is time: spurious regression problem**

2 options yield similar coefficieent estimates

Option 1: control for time

add a regressor of time to the model

Option 2: detrend the data

1. detrend each and every independent variable
    - regress each one on the time index t
    - the residual from the regression is the detrended var
2. run the regression using the detrended vars

Linear, quadratic, exponentially detrend...

Seasonality

monthly time series data

similar approach as before

Prediction

handout p15-18

generate prime variables and run the modified regression

give the confidence interval

Pooled cross-section data

repeated cross-section

> lec 20

Cross-sectional: random sample at a single pooint in time

time series: data for a sequence of time points

pooled cross-section: random samples drawn from the same population at different times how relationship has changed over time

- No worry for serial correlation/ violations of strict exogeneity data within the same year is drawn randomly
- randomly selected data from one year should not have impact on the randomly selected data from another year

Pooled regression (Add all data from different years together) does not represent the trends in diffeerent year separately.

Solution: **add a year dummy with an interaction term for the year dummy**

Estimator for the year dummy: difference in return to x in different years

Test of **structual change**:

- Entire change: H0: coefficients for the year dummy and interaction terms =0
- Only test the impact of x differ or not. H0: coefficients for the interaction terms = 0

Difference in Difference estimator

Program evaluation

Handout p4-6

treatment variable: year when the treatment started

treated group: the group impacted by the treatment

control group: the group not impacted by the treatment

Dnd estimator: the coefficient on the interaction, unbiased estimator of the truee average treatment effect if

- treatment only affects the treated group (no spillovers)
- in the absence of the treatmeent, there would have been parallel trends

Panel Data

> lec 20

panel data: logitudinal data, a time series for every observation of a cross-section

not random sample, observation for the same sampling unit are called a group or cluster

Problem: **serial correlation**

Violate: $cov(v_{it}, v_{js}) = 0$ when i=j, t != s same sampling unit in different time

Will have: $cov(v_{it}, v_{js}) = Cov(a_i + u_{it}, a_i = u_{is}) \neq 0$, where a is time invariant and u is time varying

Clustered standard errors

not assuming anything about $cov(v_{it}, v_{js})$ or Var(v)

estimate them instead

point estimates: the same as robust SE or regular SE

Usually error: regular <= robust <= clustered

Random effects

not stick with OLS, more efficient estimator

Assume

- $cov(v_{it}, v_{js}) = 0$ for all t != s
- $cov(a_i, v_{js}) = 0$ for all t
- $Var(a_i) = \sigma_a^2$
- $var(u_{it}) = \sigma_u^2$

Imply: $cov(v_{it}, v_{js}) = \sigma_a^2$

differences in point estimates from OLS,

Errors: usually RE<= clustered

> Let 21

Time Invariant omitted variable

benefits for the panel data: time-invariant omitted var

The omitted variables in the error term,

- decompose the error term: $u_{it} = a_i + u_{it}$
- a_i contains omitted variables, correlated with y and x,
- If it is time-invariant, possible to comsistently estimate with the model below

First Differecnce

$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + uit$

$y_{it-1} = \beta_0 + \beta_1 x_{it-1} + a_i + uit - 1$

Substracted the second from the first, we get

First difference of variables: $\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta u_{it}$

Fixed effects

substract the original model by the goup mean: $\overline{y}_i = \beta_0 + \beta_1\overline{x}_i + a_i + \overline{u}_i$

group demented variables: $\ddot{y} = \beta_1\ddot{x}_{it} + \ddot{u}_{it}$

Comparison

Both FD and FE

- Both FD and FE provides a way forward when there is a time invariant omitted variable
- Unbiased if assume E[u|x] = 0
- Neighter can estimate the effect of a time-invariant regressor
- If litter variation over time, make matter worse (e.g. height as x does not change much over time)

FD or FE

- When T=2, FD = FE
- When T>2, they will differ
- Both are good to use, FE more popular

FE/FE or RE

- If ai is uncorrelated with x
    - serial correlation
    - OLS with clustered SE or RE
- If ai is correlated iwth x
    - OLS and RE are biased
    - FE/FD

e.g. p4-8 lec 21 for FE/FD stata + time as explanatory variable

Instrumental variables

> Lec 21

Problem with SLR/MLR 4(x uncorrelated with u) that maybe due to omitted variables

- try to include all the variables -> always sth. not included

- differencee in difference model -> maybe data does not exist
- FE/FD
  - can not estimate the effect of time-invariant omitted variables
  - can noat deal with time variant omitted variables

New option: Instrumental variables z

- uncorrelated with u, Cov(z,u) = 0 -> valid
  - referred as statistically exogenous
  - excludes the possibility that z impacts y via u -> satisfy the exclusion restriction
- correlated with x, Cov(z,x) != 0->relevant

E.g. when estimating wages with education, ability is omitted. Can use number of siblings as the instrument

- correlated with education
- probably not correlated with ability

Testing Relevance

regress x on z: $x = \pi_0 + \pi_1 z + u$

H0: $\pi_1 = 0$ -> Cov(z.x)= 0, not relevant

Testing Validity

Problem: test the relationship between z and the error while the error is unobserved

Can not be directly tested. -> central dilemma

The estimator is **biased**

But when it is valid and relevant, it is **consistent**

Variance of the IV estimator >= the variance of the OLS estimator

If x and u are uncorrelated:

- OLS and IV are consistent, OLS has lower variance
- use OLS

If x and u are correlated:

- OLS is inconsistnet, IV is consistent
- Use IV

Weak instrument:

- Cov(x,z) is small, z explains only a small fraction of the variation in x
- variance will be large

Nearly exogenous, Cov (z,u) is close to 0

- small degreee of asymptotic bias
- if we have a weak instrument, large asymptotic bias

Multiple regressors:

- x correlated with error: endogenous regressor
- x uncorrelated with error: exogenous regressor
- Test relevant: regress the endogenous regressor on otherssss regressors and the IV

Estimating IV using 2SLS

Steps p8, e.g. p9-10:

1. First stage: purges x1 of its correlation with u
    - regress the endogenous regressor on the IV and exogenous regressors
    - Take the fitted value from the regression
2. Second stage: explains y using the remaining variation in x1
    - swap the endogenous regress with the fitted value from the original model

k endogenous regressors and i IV

- k>i, unidentified
- k=i, fine
- k<i, over-identification