Micro Econometrics

Sol Yates

February 14, 2024

Contents

Lecture 2: Standard Errors

Sun 04 Feb 17:54

1 Introduction

- after a point estimate, we want to know the statistical significance
- requiring the standard error and the distribution
- if the standard errors are wrong we cannot use the usual t-dost statistics for drawing inference
- We either have too large SEES,
  - zero might be included in the CI when it should not be
  - there is a risk of not detecting an effect even there was
- Or, too small Sees
  - zero might not be in the CI when it should be
  - we may claim the existence of an effect when in reality there is none
- **wrong** SE can lead to **wrong** conclusions!
- Robust SE

  - traditional inference assumes homoskedacity
  - but the variance of error terms might be different for different observations depending on their characteristics
  - heteroskedacity robust SE to the rescue

- SE
  - traditional estimation relied on random sampling
  - in the case of data with a group structure, the error terms might be correlated
  - to account we use clustered SE

- bootstrap
  - bootstrap is a re sampling method that offers an alternative to inference based on asymptotic formulas convenient in cases where the sampling distribution is unknown

Note.
1. if we can estimate a model parameter consistently, why do we care about inference?
2. do heteroskedatic errors or clustering affect the OLS point estimate for model parameters
3. an example where heteroskedacity / clustering occurs
4. when would bootstrap be useful?

2 Heteroskedacity - Robust Standard Errors

2.1 Heteroskedacity Problems

- traditional inference assumes homoskedatic errors $V(u|x) = \sigma^2$
- this implies that the variance of the unobserved error a, is constant for all possible values of all the regressor x's
- since the proofs for unbiasedness and consistency do not depend on this assumption we still obtain unbiased and consistent OLS estimates
- however, if this is not true (σ_i^2) then the errors are called **heteroskedatic** and traditional variance estimators are biased
- heteroskedacity robust SE specifically in the CS case
- if the degree of heteroskedacity is low ,the traditional variance estimator might be less biased

Example (Returns to education)**.** if we regress $wage \sim educ$

it is reasonable to believe that the variance is unobserved factors hidden in the error term differers by educational attainment

individuals with higher education : potentially more diverse interests and more job opportunities affecting their wage

individuals with very low education : fewer opportunities and often must work at the minimum

> wage, the error variance is lower

variance estimation with heteroskedacity

simple regression : $y = \beta_0 + \beta_1 x + u$
we know $\hat{b}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x} u_i)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
which is a function of the error terms
therefore :

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sigma_i^2}{SST_x^2}$$

where $SST_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$

- where σ_i^2 is conditional variance of error term (depending on each individual)
- if $\sigma_i^2 = \sigma^2$ the formula reduces to the traditional (OLS variance) formula : $V(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$
- we have to estimate the conditional variance of the error, we do this by tang the residuals of OLS, squaring them and replacing them in the following formula for the error variance
- this leads to the following heteroskedacity robust estimator (simple regression model) :

$$\hat{V}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \overline{u}_i^2}{SST_x^2}$$

where u_i^2 are the OLS residuals

Generalisation

the formula generalises to

$$\hat{V}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

where the σ_i^2 are replaced by residuals soured from OG regression and the $\hat{r_{ij}}$ are the residuals from regressing x_j on all other independent variables.

where $\hat{r_{ij}}$ is the i-th residual from regressing x_j on all other independent variables and SSR_j the sum of squared residuals from this regression

- robust to heteroskedacity **of any form** (inc homoskedacity)
- often also called white, huber, eicker SE
- sometimes degrees of freedom adjustment by multiplying $\frac{n}{n-k-1}$
- but with **drawback** that it only has asymptotic justification (need large sample for it to be valid)

Matrix Representation - Asymptotic Variance
model : $y = X\beta + U$

We know $\sqrt{n}(\hat{\beta} - \beta) \vec{d} \mathcal{N}(0, V)$ Where

$$V = E[E[X'X]]^{-1}[E[X'Xu^2]][E[X'X]]^{-1}$$
$$\text{with fixed regressors(replace with sample analog)}$$
$$= [\frac{1}{n}X'X]^{-1}[\frac{1}{n}X'\psi X][\frac{1}{n}X'X]^{-1}$$
$$\text{And the variance-covaraince matrix}\psi$$
$$\psi = \begin{bmatrix} V(u_1|x) & 0 & \dots & 0 \\ 0 & V[u_2|x] & \dots & 0 \\ \dots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & V[u_n|X] \end{bmatrix}$$

Eventually, $\frac{V}{n} = AV(\hat{\beta})$

Matrix Representation - Estimation

- we can then find an estimate the middle term by : $\frac{1}{n}\sum_{i=1}^{n} \hat{u_i^2 x_i'} x_i = \frac{1}{n} X\hat{\psi}X$
- Where $\hat{\psi} = diag[\hat{u_i^2}, \dots,]$
$$\hat{V} = [\frac{1}{n}X'X]^{-1}\frac{1}{nX'\hat{\psi}}X[\frac{1}{n}X'X]^{-1}$$
- In order to estimate the Asymptotic Variance (AV) $\hat{\beta}_j$, we need to remove the asymptotic normalisation by dividing by n
- Resulting Estimator :
$$\hat{AV} = n[X'X]^{-1}\frac{\sum_{i=1}^{n} \hat{u}_i^2 x_i' x_i}{n}[X'X]^{-1}$$
- sometimes corrected by the degrees of freedom $n/n - k - 1$ to improve finite sample properties
- SEs : square root of the diagonal elements
- Recall that under homoskedacity, we obtain $\sigma^2(X'X)^{-1}$

Example. Returns to Education reg1 = lm(wage ∼ educ, data =wage1)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.21158 -0.36393 -0.07263  0.29712  1.52339 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
educ        0.082744   0.007567  10.935  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4801 on 524 degrees of freedom
Multiple R-squared:  0.1858,	Adjusted R-squared:  0.1843 
F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16

> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 0.5837727  0.0982339  5.9427 5.118e-09 ***
educ        0.0827444  0.0077389 10.6920 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: R regression output

In which we have used coeftest and vcovHC HC1 variance covariance matrix for one form of the robust one. We have obtained the estimates in both cases.

Comparing, we have the **same estimate**, however the **SE** in the robust case are slightly bigger. This isn't a great example since it doesn't change significance however it shows both estimators can give different SE, but the estimate from OLS remains the same.

2.2 Breusch-Pagan Test for Heteroskedacity

- testing hypothesis
$$H_0 : V(u|x_1, \ldots, x_k) = E(u^2|x_1, \ldots, x_k) = \sigma^2$$
where $V[u|x] = E[u^2|x] - \underbrace{0} E[u|x]$

- assume a linear relationship :
$$u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v, E[v|x_1, \ldots, x_k] = 0$$

- since we cannot observe the errors (u^2), we replace them with the residuals and estimate the regression

1. estimate
$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + error$$
Recover the $R^2_{\hat{u}}$

2. Hypotheses : $\delta_1 = \ldots = \delta_k = 0$

3. test stat : $F = \frac{R^2_{\hat{u}}/k}{(1-R^2_{\hat{u}})/(n-k-1)} \sim^{H_0} \mathcal{F}_{k,n-k-1}$ Or $LM = nR^2_{\hat{u}^2} \sim^{H_0} \chi^2_k$ where k dof, F follows fisher dist, LM follows chi squared dist.

4. Decision : if the p-value is small enough (typically < 0.05), we **reject** the null of homoskedacity

Exercise 1 (Heteroskedacity with 2 Categories). model $y_i = \beta_0 + \beta_1 d_i + u_i$, $i = 1, \dots, n$ where d_i is a binary variable

let $n_1 = \sum_i d_i$, $n_0 = \sum_i (1 - d_i)$, $n = n_1 + n_0$ and $p = \frac{n_1}{n}$ (probability of being trated, share of treated ind in samp / n)

we have seen that $\hat{\beta}_1 = \overline{y}_1 - \overline{y_0}$ and $\hat{\beta}_0 = \overline{y_0}$ (differences in group mean outcomes) ($\hat{\beta}_0$ is intercept, mean of untreated)

under homoskedacity in small sample conventional t statistic has a t-distribution

Heteroskedacity here means that the variances in the $d_i = 1$ and $d_i = 0$ population are different: the exact small sample distribution for this problem is unknwon

differences in the standard error formulae depend on how the variance in d_i is modelled (residual as difference between outcome and group mean outcome)

- note $\hat{u}_i = y_i = \overline{y_i}$ for $d_i = I$, $I \in \{0, 1\}$
- Define $s_I^2 = \sum_{i:d=I} (y_i - \overline{y_I})^2$ (which is the estimated sum of squared residuals in each group)
- Under conventional SEs: $\hat{\sigma^2}(X'X)^{-1}$ with estimate of $\hat{\sigma^2} = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u_i}^2$
- where $\sum_{i=1}^n \hat{u_i}^2 = \sum_{i:d=1} \hat{u_i}^2 + \sum_{i:d=0} \hat{u_i}^2 = s_1^2 + s_0^2$ (sum od squared resid = sum of residuals squared for treated and untreated ind)
- hence $\hat{\sigma}^2 = \frac{s_1^2 + s_0^2}{n-2}$ (is equal to n-2 since have single regressor and intercept)
- now, $(X'X)^{-1}_{[2,2]} = \frac{n}{nn_1 - n_1^2}$ (if interested in slope, take X and 2,2 element equal to this expression, using this we can take estimator for variance)
- hence $\hat{V}(\hat{\beta}_1)_c = \frac{n}{n_1 n_0} \frac{s_1^2 + s_0^2}{n-2}$ (conventional variance estimator if replace elements by percentage shares)
- it can be shown that $\hat{V}(\hat{\beta}_1)_c = \frac{1}{np(1-p)} \frac{s_1^2 + s_0^2}{n-2}$
- for robust SEs : $\hat{\sigma^2}(X'X)^{-1}(X\hat{\psi}Z)(X'X)^{-1} \to \hat{V}(\hat{\beta}_1)_r = \frac{s_1^2}{n_1^2} + \frac{s_0^2}{n_0^2}$
- when $\frac{s_1^2}{n_1} = \frac{s_0^2}{n_0}$, both estimates coincide (for large n)
- when $n_1 = n_0 = \frac{n}{2}$ they also coincide, when the data are balanced, the robust SE won't differ much from the traditional one under heteroskedacity
- if both groups variances are the same, then both estimates coincide, because then we have homoskedacity
- also if we have the same indiviudals for treated and untreated groups, then they also coincide, so if we have very balanced data (2 cat) the robust SE won't differ much from the traditional one

BP test

- interpretation of the BP test

- recall the regression $\hat{u_i}^2 = \delta_0 + \delta_1 d_i + v$

$$\hat{\delta}_0 = \frac{\sum_{i:d=0} \hat{u_i}^2}{n_0} = \frac{s_0^2}{n_0}\hat{\delta}_1 = \frac{\sum_{i:d=1} \hat{u_i}^2}{n_1} - \frac{\sum_{i:d=0}}{\hat{u_i}^2} n_0 = \frac{s_1^2}{/} n_1 - \frac{s_0^2}{n_0}$$

- Testing $H_0 : \delta_1 = 0$ is equivalent to testing $\sigma_1^2 = \sigma_0^2$

Example (Housing Price Equation). Log is sometimes used to get rid of heteroskedacity

model : price $= \beta_0 + \beta_1$lotsize $+ \beta_2$sqrft $+ \beta_3$bdrms $+ u$

where price is the housing price, lotsize the size of the lot, size size of house in sq ft

we want to estimate teh above regression and test for heteroskedacity and see whether using logs in the dependent variable changes our conclusion

```
Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-120.026  -38.530   -6.555   32.323  209.376

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,	Adjusted R-squared:  0.6607
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```

Figure 2: Housing Price Equation Output 1

We cannot really learn much about heteroskedacity, although lot and size is statistically significant

Testing for heteroskedacity using BP test, predicting residuals from previous regression and squared them, then we take the squared residuals and regress on independent variables

```
Call:
lm(formula = u.hat2 ~ lotsize + sqrft + bdrms, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -9044  -2212  -1256    -97  42582

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.523e+03  3.259e+03  -1.694  0.09390 .
lotsize      2.015e-01  7.101e-02   2.838  0.00569 **
sqrft        1.691e+00  1.464e+00   1.155  0.25128
bdrms        1.042e+03  9.964e+02   1.046  0.29877
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6617 on 84 degrees of freedom
Multiple R-squared:  0.1601,    Adjusted R-squared:  0.1301
F-statistic: 5.339 on 3 and 84 DF,  p-value: 0.002048

>
> # Compute F-stat by hand: recover the R^2:
> R_u2 = summary(reg.res)$r.squared
> df = reg.res$df # n-k-1
> k = 3
>
> # F-stat:
> F = (R_u2/k) / ((1-R_u2)/(df))
> F
[1] 5.338919
```

Figure 3: Housing Price equation output 2

we obtain the f stat, testing for joint normality of parameter estimate, 5.3 with p value <0.05, testing for heteroskedacity using BP test leads us to reject the null of homoskedacity

```
Call:
lm(formula = log(price) ~ log(lotsize) + log(sqrft) + bdrms,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68422 -0.09178 -0.01584  0.11213  0.66899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.29704    0.65128  -1.992   0.0497 *
log(lotsize)  0.16797    0.03828   4.388 3.31e-05 ***
log(sqrft)    0.70023    0.09287   7.540 5.01e-11 ***
bdrms         0.03696    0.02753   1.342   0.1831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1846 on 84 degrees of freedom
Multiple R-squared:  0.643,     Adjusted R-squared:  0.6302
F-statistic: 50.42 on 3 and 84 DF,  p-value: < 2.2e-16
```

Figure 4

does our question change if we use logs? running the regression we obtain the above, not telling us much again, but helps us to predict residuals based on this regression, then we can test for heteroskedacity

```
Call:
lm(formula = u.hat.ln2 ~ log(lotsize) + log(sqrft) + bdrms, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05601 -0.03011 -0.01687  0.00523  0.40978

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.509994   0.257857   1.978   0.0512 .
log(lotsize) -0.007016   0.015156  -0.463   0.6446
log(sqrft)   -0.062737   0.036767  -1.706   0.0916 .
bdrms         0.016841   0.010900   1.545   0.1261
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07309 on 84 degrees of freedom
Multiple R-squared:  0.04799,   Adjusted R-squared:  0.01399
F-statistic: 1.411 on 3 and 84 DF,  p-value: 0.2451
```

Figure 5

> doing the same as before (without logs) we take our residuals, square the, then regress on individual variables. Giving us f stat of 1.4, which given the p val of 0.2 leads us to *failing to reject the null of homoskedacity* . Thus our initial SE werent very useful, but using the logs we can assume homoskedacity.

Heteroskedacity Conclusion

- use robust SE when heteroskedatic errors
- but there is a danger of small sample bias from robust SE (arising from asymptotic justification)
- under homoskedacity or little heteroskedacity, it might be preferable to use the traditional OLS variance estimator
- it is recommended to report both the robust and conventional standard error and suggest to take the maximum of both for inference
- white test for heteroskedacity includes the squares and cross-products of the independent variables
- LPM : built in heteroskedacity $\rightarrow$ need to compute robust SEs
- using logs in the dependent variables has been seen to improve in terms of heteroskedacity in many applications

3 Clustered Standard Errors

Illustration of Moulton Problem

- Pillar assumption is random sampling
- there is potential dependence of data within a group structure
  - exam grades of children from same class or school : grades are correlated because of the same school, teacher and background / class environment
  - health outcomes in the same village, Errors are correalted because of the same medical and food supply and similar cultural background
  - earning in the same region might be correlated because of the same industrial structure
  - analysing workers in firms (earnings, tenure, promotion) will suffer from common firm effects

The problem

- illustration using a simple model with a group structure
- intuitively, effect of a macro variable on an individual level outcomes
  - effect of school-type on exam-grades
  - effect of regional unemployment on individuals' wages
- model

$$y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$$

  - with $g = 1, \ldots, G$ and $i = 1, \ldots, n$

  - y_{ig} is the outcome for individual I in group g
  - here x_g varies only at the group level

Note. Lecture : if we estimate a model parameter consistently, why do we care about inference?

- we would like to investigate a problem
- policymaker would like to know whether to implement school building program
- but what is decision rule? Typically, think about Statistical significance and sufficient magnitude then the policymaker wants to adopt the program, if not then not adoptable.
- we need CI or at least a statistical test. For this we need SE and distribution
- need to estimate SE correctly to get correct CI, if we have too large CI (SE wrong), the implication/ error is that we risk not detecting an effect, when there is
- but the other way around too small SE (forgot to cluster), might think building schools help and invest a lot of money, but the effect is 0
- this is a danger and the problem is *incorrect standard errors lead to incorrect confidence intervals*

Note. Lecture : Once we have accounted for clustering using the Moulton approach compared to the standard errors, is it more likely that the clustered standard errors are larger or smaller than the OLS

> **Note.** larger

Note. Lecture : what solutions exist to account for clustering

- group averages (only valid for regressors that don't vary within each individual within a group)
- parametric - estimate the Moulton factor
- clustering SE
- block bootstrap

Moulton Problem - Cont

Data Structure

- Recall $E[e_{ig}] = 0$ & $v(E_{ig}] = \sigma_e^2$
- Recall correlation: $\rho_e = \frac{Cov(e_{ig}, e_{ig})}{sd(e_{ig})sd(e_{ig})}$
- Likely : for individual I and j from the same group g :

$$Cov[e_{ig}, e_{jg}] = \rho_E \sigma_e^2 > 0$$

Additive Random Effects

- group correlation often modelled using additive random effects, assume $e_{ig} = v$
- v_g : group specific error term which captures all the within-group correlation with $E[v_g] = 0$ & $V(v_g] = \sigma_b^2$
- n_{ig} : individual level specific error term with $E[n_{ig}] = 0$ & $V(n_{ig}) = \sigma_n^2$
- assuming v_g and n_{ig} are uncorrelated
- we note that n_{ig} and n_{jg} are uncorrelated

Data Structure

- Recall $E[e_{ig}] = 0$ & $v(E_{ig}] = \sigma_e^2$
- Recall correlation: $\rho_e = \frac{Cov(e_{ig}, e_{ig})}{sd(e_{ig})sd(e_{ig})}$
- Likely : for individual I and j from the same group g :

$$Cov[e_{ig}, e_{jg}] = \rho_E \sigma_e^2 > 0$$

Additive Random Effects

- group correlation often modelled using additive random effects, assume $e_{ig} = v$
- v_g : group specific error term which captures all the within-group correlation with $E[v_g] = 0$ & $V(v_g] = \sigma_b^2$
- n_{ig} : individual level specific error term with $E[n_{ig}] = 0$ & $V(n_{ig}) = \sigma_n^2$
- assuming v_g and n_{ig} are uncorrelated
- we note that n_{ig} and n_{jg} are uncorrelated

$$Cov(e_{ig}, e_{jg}) = E[(v_g, n_{ig})(v_g + n_{jg})] = E[v_g^2] = \sigma_v^2$$
$$V[e_{ig}] = E[(v_g + n_{ig})^2] = E(v_g^2 + n_{ig}^2) = \sigma_v^2 + \sigma_n^2$$

Intraclass Correlation Coefficient

- the intraclass correlation coefficient as the proportion of variation in $(v + n)$ due to v :

$$\rho_e = \frac{Cov(e_{ig}, e_{jg})}{sd(e_{ig})sd(e_{jg})} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_n^2}$$

- When the regressor of interest varies only at group level, then this error structure can increase standard errors sharply
- By how much is the conventional variance of the OLS estimate inflated?
- let $V_c(\hat{\beta}_1)$ denote the conventional OLS variance expression and $V(\hat{\beta}_1)$ be the correct sampling variance with this error structure

- depending on the data structure. There are various versions to quantify $\frac{V(\hat{\beta}_1)}{V_c[\hat{\beta}_{s1:1}]}$

For the following data structure :

- nonstochastic regressors fixed at the group level (that is, all regressors are the same for each individual in a group)
- Equal group sizes $N = n_1 = \ldots = n_G$ with total sample size $n = G * N$

$$\frac{V[\hat{\beta}_{s1:1}]}{V_c[\hat{\beta}_{s1:1}]} = 1 + (N-1)\rho_e$$
$$\text{Moulton Factor} \quad \sqrt{1 + (N-1)\rho_e}$$

Which quantifies how much we over estimate precision by ignoring intraclass correlation

Remark.
- Conventional standard errors become increasingly misleading as group size N and / or ρ_e increase
- if there is no error correlation ($\rho_e = 0$) , there is no overestimation
- if $\rho_e = 1$ (or $n_{ig} = 0$), then within a group, all 's are the same : the conventional variance is scaled up by (N-1) since we copy each information N times without generating new information
- with the total sample size fixed, increasing the group sizes N just decreases the number of clusters which leads to less independent information
- the Moulton factor can be very big even with a small correlation. Assume 100 observations per group and a $\rho_e = 0.1$ leads to a Moulton factor of 3.3. The conventional standard errors are only roughly $\frac{1}{3}$ of what they should be

Generalisations

The most general form where x varies by g and I with variations in g :

$$\frac{V[\hat{\beta}_{s1:1}]}{V_c[\hat{\beta}_{s1:1}]} = 1 + [\frac{V(N_g)}{\overline{n}_g} + \overline{n}_g - 1]\rho_r \rho_x$$

where rho_x is the within cluster correlation coefficient for x:

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j} (x_{ig} - \overline{x})(x_{jg} - \overline{x})}{V[x_g] \sum_g n_g(n_g - 1)}$$

- ρ_x is a generic measure of the correlation of the regressors within the group. If this correlation is zero, the Moulton effect disappears
- clustering has a bigger impact on standard errors with variable group sizes and when ρ_x is large
- If the group size is fixed but x varies by g and I, the Moulton factor becomes the square root of $1 + (N-1)\rho_E \rho_x$

Solutions

model $y = \beta_0 + \beta_1 x_{ig} + e_{ig}$ with $g = 1, \ldots, G$

1. Parametric approach
   - Fix the conventional standard errors using the general formula for the Moulton factor by estimating the intraclass correlations ρ_e and ρ_x
2. Cluster standard errors
   (a) Generalisation of white's robust covariance matrix

   $$\hat{AV}(\hat{\beta}_{s1:1}) = (X'X)^{-1}(a\sum_{g=1}^{G} X_g'\hat{e}_g\hat{e}_g'X_g)(X'X)^{-1}$$

   (b) where $\hat{e}_g$ is a n_g x 1 vector of redials for observations in the g-the cluster and X_g is a n_g x k matrix of regressors for observations in the g-the cluster
   (c) typically, there is a degrees of freedom adjustment $a = \frac{G(n-1)}{(G-1)(n-k)}$
   (d) consistent if number of cluster is large but not consistent with fixed number of groups (even when group sizes tend to ∞)
   (e) no assumptions about within-group correlation structure (not just parametric such as in the additive error structure)
   (f) if each individual is his own group (I = g and G = n) then the formula collapses back the robust estimator
3. use group averages instead of microdata
   (a) model : $y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}$, $g = 1, \ldots, G$
   (b) we estimate $\overline{y}_g = \beta_0 + \beta_1 x_g + \overline{e}_g$ by weighted least squares using n_g as weights
   (c) However, neglecting heteroskedacity unless the group sizes are equal
   (d) relying on asymptotics for group number, not group sizes
   (e) with modest group sizes, it is expected to have good finite sample properties of regressions with normal errors
   (f) and is likely to be more reliable than clustered standard errors with few clusters
   (g) but does not work if x varies within groups and ignores any other micro-level covariates
   (h) but there exists a 2 step approach to include micro level covariates (A & P)
4. block bootstrap
   (a) to be discussed
5. GLS or Max Likelihood approaches

   GLS :
   (a) in some cases is possible to estimate GLS or maximum likelihood model
   (b) requires a model for error structure

Example (Star Experiment). Krueger (1999) uses IV to estimate the effect of class size on students' achievements y_{ig} is the test score of student I in class g and class size x_g

Students were randomly assigned to each class but data are unlikely to be independent across observations.

Test scores in the same classes are correlated because students in the same class share background characteristics and are exposed to the same teacher and classroom environment

It is likely for students I and j from the same class g :

$$E[e_{ig}, e_{jg}] = \rho_r \sigma_e^2 > 0$$

The estimation strategy is for now not in our focus, thought we can compare the different standard error estimates

Standard errors for class size effects in the STAR data (318 clusters)

Variance Estimator	Std. Err.
Robust (HC_1)	.090
Parametric Moulton correction (using Moulton intraclass correlation)	.222
Parametric Moulton correction (using Stata intraclass correlation)	.230
Clustered	.232
Block bootstrap	.231
Estimation using group means (weighted by class size)	.226

Notes: The table reports standard errors for the estimates from a regression of kindergartners' average percentile scores on class size using the public use data set from Project STAR. The coefficient on class size is $-.62$. The group level for clustering is the classroom. The number of observations is 5,743. The bootstrap estimate uses 1,000 replications.

Figure 6: Robust standard errors after correcting for clustering

Lecture 3: Third Lecture

Tue 13 Feb 15:20

3.1 Basic Intro to Bootstrap

Based on the data we have we simulate and pretend we many more 'made-up' datasets we didn't have previously. Runs into issues when estimating min or max, rather than mean and under non-Gaussian distribution. When asymptotically normal or ..., bootstrap good choice.

OvB only creates bias if correlation with regressors, if going to argue variable is non-correlated, it is fine. But including too many regressors may be problematic too, end up including too many variables correlated with regressor, on top of fact it creates noise.

Attenuation Bias - if measurement error, nothing can do about it. As long as variance and this measurement error, it exists. But if less variance in measurement error, then it disappears. Of course, provided error isn't systematic.

- another method for estimating variance, CI and dist on statistic
- often used when exact distribution is unknown
- different versions but non parametric most common

3.2 Non-parametric Bootstrap

- X is distributed according to some distribution F: $X \sim F$
- $x = (x_1, \ldots, x_n)$ represents an iid sample from this variable
- suppose we want to estimate the variance and the distribution of a statistic $T_n = g(x_1, \ldots, x_n)$
- ultimately interested in variance of distribution of this statistic $T_n = g(x_1, \ldots, x_n)$

NP Bootstrap - Variance

- let V_F denote the variance of T_n where the subscript F indicates that the variance is a function of F
- if we knew F, we could compute the variance
- for example for $T_n = \frac{1}{n}\sum_{i=1}^{n} x_i$,

  $$V_F(T_n) = \frac{V(x)}{n} = \frac{\int x^2 dF(x) - (\int x dF(x))^2}{n}$$

  where $dF(x)$ is the pdf in integral form (2nd term)
- which is a function of F
- idea is to estimate $V_F(T_n)$ with $V_{\hat{F}}(T_n)$
- Or, *use a plug in estimator* of the variance
- since $V_{\hat{F}}(T_n)$ may be difficult to compute, we approximate it with a simulation estimate denoted by v_{boot}

Key Idea

Put the initial sample $x = (x_1, \ldots, x_n)$ into an urn

1. draw n observations from x with replacement
   - each observation has the probability of $\frac{1}{n}$ of being drawn
   - gives each bootstrap sample $x_1^* = (x_{11}^*, \ldots, x_{n_1}^*)$
2. based on the single bootstrap sample, we estimate (compute bootstrap statistic)

   $$T_{n_1}^* = g(x_{11}^*, \ldots, x_{n_1}^*)$$

3. repeat steps 1 and 2 B times to get $T^*_{n_1}, \ldots, T^*_{nB}$ where :

$$T^*_{nb} = g(x*_{1b}, \ldots, x^*_{nb}) \; for \; b = 1, \ldots, B$$

Where B is the number of bootstrap replications

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(T^*_{nb} - \frac{1}{B}\sum_{r=1}^{B} T^*_{nr})$$

(then take sample analog of variance)

Then by the law of large numbers $v_{boot} \xrightarrow{a} V_{\hat{F}(T_n)}$ as $B \to \infty$ (bootstrap variance tends to variance of stat we were after)

Need to reiterate quite often, in real world we have initial sample from true distribution F which gies us stat T_n in bootstrap world we have bootstrap sample which comes from resampling our initial sample which gives us our bootstrap stat : T^*_n

Imagine in real world, initial sample with 4 obs, giving us stat which is a function of these 4 obs (say the average over these 4 obs). In order to get into boostrap world, we place sample in urn, we draw b times 4 observations each time with replacement

$$\begin{array}{rll}
1^{st} \text{ draw:} & x^*_1 = \{1,3,1,2\} & \to g(1,3,1,2) = T^*_{n1} \\
2^{nd} \text{ draw:} & x^*_2 = \{1,4,4,4\} & \to g(1,4,4,4) = T^*_{n2} \\
\ldots & & \\
b^{th} \text{ draw:} & x^*_b = \{2,4,1,1\} & \to g(2,4,1,1) = T^*_{nb} \\
\ldots & & \\
B^{th} \text{ draw:} & x^*_B = \{1,3,2,4\} & \to g(1,3,2,4) = T^*_{nB}
\end{array}$$

Figure 7: Boostrap World

When we do the 1-st draw we get 1, then since draw with replacement, it could happen we draw this again, second is 3, we also put this back, we do this even further then we got again observation with 1.

Then eventually we get the observation with 2 then we can compute the bootstrap statistic b times to obtain bootstrap samples

Use of the Bootstrap

- The empirical distribution of the B bootstrap samples gives us the approximated distribution / moments of T_n
- EEG standard Errors : $\hat{se} = \sqrt{v_{boot}}$
- approximate the CDF of T_n. Let $G_n(t) = (T_n < t)$ be the CDF of T_n
- the bootstrap appropriate to G_n is

$$\hat{G}^*_n(t) = \frac{1}{B}\sum_{b=1}^{B} 1_{\{T^*_{nb} \leq t\}}$$

where the binary variable obtains probability

- confidence intervals based on SE or quantiles

- normal interval :
$$T_n \pm z_{\frac{a}{2}} \hat{se}_{boot}$$
- where $\hat{se}_{boot}$ is the bootstrap estimate for the SE
- where $z_{\frac{a}{2}}$ is the $\frac{a}{2}$ quantile of the standard normal distribution
- the interval is not accurate unless the distribution of T_n is close to normal

```
> set.seed(1000)
> n = 100
> mu = 5
> sigma2= 2
> X = rnorm(n,mu,sd = sqrt(sigma2))
> #Tn = mean(X)
>
> B = 500
> TnStar = c()
> for (i in 1:B){
+   XStar = sample(X, size = n, replace = TRUE)
+   TnStar[i] = mean(XStar)
+ }
> # Bootstrap Mean:
> mean(TnStar)
[1] 5.031754
> # True mean: 5
>
> # Bootstrap Variance:
> var(TnStar)
[1] 0.02003969
> #True Variance:
> V = sigma2/n
> V
[1] 0.02
```

Figure 8: Boostrap Code

set seed to ensure RV is same on diff computers, then 100 obs, mean = 5, variance = 2

we want 500 bootstrap replications, then we initiate an empty vector t_n^* to collect bootstrap replications, then iterate over 500 i's. For each I in 1:500 we sample from our initial vector, with replacement 100 observations, giving us bootstrap sample, then we take mean to obtain bootstrap mean

T_n^* has 500 bootstrap means, then we take mean over these 500 and compare to true expectation. 5.03 is very close to the true mean,

We proceed the same to estimate variance based on bootstrap replications, we are also close to variance also.

Practically, it depends on the situation to normalise test stat (demean or standardise in order to ensure normal distribution)

Regression Estimates

procedure quite similar, but with at least 2 characteristics for each individual parameters

Instead of drawing directly from RV, we draw pairs of $\{y_i, x_i\}$ to

- sometimes called the *pairs bootstrap*
- instead of drawing directly from the random variable, you would sample the indices of the observations

Empirical, non parametric, standard, pairs.

Wild Bootstrap

relies on assumption that error term at disposal

- model $y_i = \beta_0 + \beta_1 x_i + u_i$ (one regressor)
- preserves heteroskedatic behaviour since don't destroy link between x's and error terms
- initial sample $z = [(y_1, x_1) \dots (y_{n,x_n})]$ with outcome and regressors for each individual

Methodology
quite similar but main difference that it is residual bootstrap but keep regressors fixed

1. estimate $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ for all $i = 1, \dots, n$
2. randomly create a bootstrap residual (weights)

$$\text{weights:} \quad w_i = \begin{cases} 1 \ , & \text{with probability } \frac{1}{2} \\ -1 \ , & \text{with probability } \frac{1}{2} \end{cases}$$

   Bootstrap residuals $\hat{u}_i^* - w_i \hat{u}_i$, for all $i = 1, \dots, n$

3. Compute the bootstrap dependent variables (essentially changed sign of original residual)

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

   for all $i = 1, \dots, n$ Gives : a single bootstrap sample : $z_1^* = [(y_{11}^*, x_1), (y_{n1}^*, x_n)]$

Repeat 1-3 B times to obtain B wild bootstrap samples : $z_b^*[(y_{1b}^*, x_1), \dots, (y_{nb}^*, x_n)]$ for $b = 1, \dots, B$

Clustered Bootstrap

Under similar randomisation, you do not preserve the dependence (unobserved factors relating to micro-data) structure in the data

To fix this, we draw blocks of data defined by the groups g. Say block bootstrap by re sampling entire classes instead of individual students, to keep structure of correlation intact.

Can also have cluster 1 bootstrap, maybe you have stratified sampling such that while you sampled you made sure have say gender quota or certain subset, we would need to do bootstrap for this.

The way you sample data structure, try to mimic through the bootstrap exactly this structure. That is, replicate DGP as close as you can (provided we know about it)

Exercise 2 (Algorithm to obtain SE using clustered bootstrap).

set the seed

if we have 1 village, should we have randomly different households in village 1? No

village 1, 2, 3 and households a, b, c (1), def, (2), ghi (3). Everything within the villages is kept fixed

we then draw 1 (abc) , 3(ghi) , 3 (ghi) for 1-st bootstrap

then 2, (ghi) , 1 (abc) , 1 (abc) for 2-nd bootstrap

need large sample.

based on practice not theory

advantages - as opposed to random draw, forget about cluster, end up doing randomly a,d,f,c and estimate OLS $\hat{\beta}_1$ and g,h,c,f and estimate $\hat{\beta}_2$, you have no attachment to group and lose correlation within each group

need to re merge together 'blocks' into single data set since we have individually sampled blocks

have to store estimator so on

Exercise 3 (Effect of schooling on wages, use father educ as instrument for years of educ)**.**

conditions of good instrument?

Exclusion restriction (orthogonality - instrument cannot be correlated with error term), Relevance restriction ($Cov(x_1, z) = 0$)

maybe since there is push to education, maybe with time this effect is fading, but likely still relevant, but maybe in other countries this is deterministic and is something we can test

exclusion restriction - 1. We can control for this, if this is not part of the model. 2. Might be violated if we can argue ability for singers, parents can sing, inherit singing talent so opera hires, this might be correlated with number of years taking singing lessons but choosing to take singing lessons due to natural talent suggests violation of exclusion criteria

Exercise 4 (Is month of birth good instrument for years of education)**.**

Exclusion Restriction - it is pretty random when you are born, there is no reason to believe the error term left over when explaining wages is correlated with when you are born

some spikes of the births (seasonality etc) though, could this be an issue?

Relevance condition - Structure of Education System : cutoff for year of schooling in the year, therefore matters for when say leave school at 16

usa : start in September, able to leave when 16, people born earlier get extra months of schooling

does extra month of education have strong effect? No

F stat - very low in this case, but we wont a large F-stat

if we have a small f stat we have very low relevance, but also $\beta = cov(y, x)/cov(x, z)$

bias end up having depends on the covariance between y and u and x and z

if low relevance, then $\frac{cov(z,u)}{cov(x,z)}$ 'explodes'

to test, there is no proper overall applicable test for exclusion restriction, it is something you have to argue for.

4 Instrumental Variables

Motivation

Let's say we are interested in identifying the causal effect of years of schooling on wage, we estimate the model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

- One of the key assumptions for unbiasedness is the *homogeneity of regressors* : $E[u|x_1, \ldots, x_k] = 0$

- Indeed, the problem arises when the regression error is correlated with a regressor : ie $E[u|x_k] \neq 0$
- there are three broad reasons for *Endogeneity* :
  1. Omitted variable bias
  2. measurement error
  3. simultaneous equations
- in our example, x_k is said to be endogenous, meaning the years of schooling might be correlated with innate and unobserved ability
- the OLS estimator β_k is *biased* and *inconsistent*
- One approach to deal with this issue is to use instrumental variables

4.1 Forms of Endogeneity

- Omitted variables
  1. arises in cases when one fails to control for a regressor that is correlated with other regressors
  2. often due to self selection : if an agent chooses the value of the regressor, this might depend on factors that we cannot observe
  3. that is, *unobserved heterogeneity*
- Measurement Error
  1. Occurs when we can only observe an imperfect measure of a variable
  2. Depending on how the observed and true variable are related, we might have endogeneity
- simultaneity
  1. Occurs when dependent and independent variables are simultaneously determined
  2. if x is partially determined by y, then the error might be correlated with x

Though this is not to say there exist sharp distinctions

> **Example.** Effect of alcohol consumption on worker productivity (measured by wages)
>
> Alcohol usage correlated wuth unobserved factors such as family background, which may also have an effect on wage. Leading to an Omitted variable problem
>
> Alcohol demand can depend on income, leading to the simultaneity problem
>
> There also exists possibility of mismeasurement of alcohol consumption

Omitted Variable Bias (OVB)

- Long regression - true model is : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$
- then, assuming we cannot observe x_2 but only x_1
- Or, in short : $y = \delta_0 + \delta_1 x_1 + u$ with $u = \beta_2 x_2 + e$

- we know the population parameter can be expressed as :

$$\delta_1 = \frac{cov(y, x_1)}{v[x_1]}$$

replacing y from the true model:

$$= \frac{Cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + e, x_1)}{V[x_1]}$$
$$= \beta_1 v[x_1] + \beta_2 Cov(x_2, x_1) + cov(e, x_1)/V[x_1]$$
$$= \beta_1 + \beta_2 \frac{cov(x_2, x_1)}{v(x_1)}$$

Defining τ_1 as the parameter in the population model that relates x_1 to x_2 :

$$x_2 = \tau_0 + \tau_1 x_1 +' error'$$

We therefore estimate $\delta_1 = \beta_1 + \beta_2 \tau_1$

However, our OLS estimate is *inconsistent* (asymptotically biased)

$$\underset{n \to \infty}{plim} \hat{\delta_1} - \beta_1 = \beta_2 \frac{Cov(x_1, x_2)}{v(x_1)} = \beta \tau_1$$

with

$$Bias(\hat{\delta_1}) = E[\hat{\delta_1}] - \beta_1 = \beta_2 \hat{\tau_1}$$

Where thinking about the direction of the correlation helps us think about the direction of the bias

Essentially, if the omitted variable is related to the included regressor, then the parameter in the short regression will not identify the parameter in the long regression.

With more regressors, the formula changes but the principle remains the same

> **Example.** omitted variable bias
>
> let y be the wages, x_1 years of education and x_2 ability
>
> regressing wages on years of education alone delivers a biased estimate $\hat{\delta_1}$
>
> we would expect both years of education and ability to have a positive impact on average earnings (that is $\{\beta_{1/2} > 0, \}$)
>
> but we also expect both regressors to be *positively correlated* as individuals with more *innate ability* tend to choose / acquire more education ($\tau_1 > 0$)
>
> therefore, $\hat{\tau_1}$ likely overestimates the value of education, since in our education regressor we have not controlled for the correlation with ability and thus include more than the effect of education in this estimate, here thinking about the direction of the correlation has helped us to identify the sign of the bias

Measurement Error in y

Situation 1 : Measurement error in the dependent variable (y)

true model $y = \beta_0 + \beta_1 x_1 + \varepsilon$, $E[\varepsilon | x_1] = 0$

We can only observe $\tilde{y}$ which measures the unobserved y with an error $\tilde{y} = y + e$

We regress $\tilde{y} = \tilde{\beta_0} + \tilde{\beta_1} x_1 + \tilde{\varepsilon}$

$$\tilde{\beta_1} = \frac{Cov(\tilde{y}, x_1)}{V[x_1]} = \frac{Cov(y+e, x_1)}{V[x_1]} = \beta_1 + \frac{Cov(e, x_1)}{V(x_1)}$$
$$\tilde{\beta_0} = E(\tilde{y}) - \beta_1 E(x_1) = E[y] + E[e] - \beta_1 E[x_1] = \beta_0 + E(e)$$

However, this can cause bias and inconsistency. Although it vanishes if the measurement error is statistically independent of each explanatory variable. We note the usual OLS inference procedures are asymptotically valid.

Situation 2 : Measurement error in the regressor (x)

true model $y = \beta_0 + \beta_1 x_1 + \varepsilon$, $E(\varepsilon|x_1) = 0$

Where we can only observe $\tilde{x_1}$, a measure of the unobserved x_1 with an error $\tilde{x_1} = x_1 + \varepsilon$

We then regress $y = \tilde{\beta_0} + \tilde{\beta_1}\tilde{x_1} + \tilde{\varepsilon}$

$$\tilde{\beta_1} = \frac{Cov(y, \tilde{x_1})}{V[x_1]} = \frac{Cov(\beta_0 + \beta_1 x_1 + e, x_1 + e)}{V[\tilde{x_1}]} = \beta_1 \frac{V(x_1)}{V(\tilde{x_1})}$$
$$\tilde{\beta_0} = E(\tilde{y}) - \beta_1 E(x_1) = E[y] + E[e] - \beta_1 E[x_1] = \beta_1 \frac{V(x_1)}{V(x_1) + V(e)} = \beta_1 \lambda$$

With the key assumptions that $Cov(e, x_1) = 0$, $cov(e, \varepsilon) = 0$, and $E[e] = 0$

In which we can show $plim_{n\to\infty}\hat{\tilde{\beta_1}} = \beta_1 \lambda$, where $\lambda \in \{0, 1\}$: $\hat{\tilde{\beta_1}}$ underestimates β_1, this is attenuation bias. Though as $V(e)$ shrinks relative to $V(x_1)$, the attenuation bias disappears.

In the general model, it is not the variance of the true regressor that affects the consistency but the variance in the true regressor after netting out the other explanatory variables

Simultaneity / Reverse Causality

Problem : y_1 and y_2 are simultaneously determined.

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1, E[z_1|u_1] = 0$$
$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2, E[z_2|u_2] = 0$$

A classic example is when y_1 is price and y_2 is the quantity and both equations are demand and supply. But note the intercept is suppressed for simplicity.

Focusing on the 1-st equation, to show that $Cov(y_2, u_1) \neq 0$:

$$y_2 = \alpha_2[\alpha_1 y_2 + \beta_1 z_1 + u_1] + \beta_2 z_2 + u_2$$
$$= \frac{\beta_1 \alpha_2}{1 - \alpha_1 \alpha_2} z_2 + \frac{\alpha_2}{1 - \alpha_1 \alpha_2} u_1 + \frac{1}{1 - \alpha_1 \alpha_2} u_2$$

Assuming that $\alpha_1 \alpha_2 \neq 0$, $Cov(u_1, u_2) = 0$ and $cov(z_2, u_1) = cov(z_2, u_2) = 0$, thus violating exogeneity

$$Cov(y_2, u_1) = \frac{\alpha_2}{1 - \alpha_1 \alpha_2} V(u_1 \neq 0)$$

Although, without additional controls (z_1), it can be shown that the inconcsisnency has the same sign as $\frac{\alpha_2}{1-\alpha_1\alpha_2}$

5 IV Estimator

Properties of an instrument to overcome Endogeneity