Tut10: Ramsey RESET, Endogeneity

Anh Le

Ramsey RESET test

What's the null hypothesis of the RESET test?

```
library(lmtest)
x <- c(1:30)
y1 <- 1 + x + x^2 + rnorm(30)
y2 <- 1 + x + rnorm(30)
resettest(y1 ~ x , power=2, type="fitted")
```

```
##
##  RESET test
##
## data:  y1 ~ x
## RESET = 131800, df1 = 1, df2 = 27, p-value < 2.2e-16
```

```
resettest(y2 ~ x , power=2, type="fitted")
```

```
##
##  RESET test
##
## data:  y2 ~ x
## RESET = 0.062376, df1 = 1, df2 = 27, p-value = 0.8047
```

Endogeneity

The boogeyman for social scientists. But what exactly does endogeneity mean?

Good seminar manner is not to just say you have an endogeneity problem, but spell out what kind it is and what is causing it.

We'll examine different causes of endogeneity below.

Reverse causality / Simultaneity bias

$$police = \alpha_0 + \alpha_1 * crime + u$$
$$crime = \beta_0 + \beta_1 * police + v$$

We plug in the function for `crime` into the first equation

$$\begin{aligned} police &= \alpha_0 + \alpha_1 * (\beta_0 + \beta_1 * police + v) + u \\ &= \alpha_0 + \alpha_1\beta_0 + \alpha_1\beta_1 police + \alpha_1 v + u \\ (1 - \alpha_1\beta_1)police &= (\alpha_0 + \alpha_1\beta_0) + (\alpha_1 v + u) \\ police &= \frac{\alpha_0 + \alpha_1\beta_0}{1 - \alpha_1\beta_1} + \frac{\alpha_1 v + u}{1 - \alpha_1\beta_1} \end{aligned}$$

As we can see, *police* is correated with v. That's endogeneity.

```r
a0 <- 1 ; a1 <- 1; b0 <- 2; b1 <- -2
u <- rnorm(1000)
v <- rnorm(1000)

police <- (a0 + a1 * b0 + a1 * v + u) / (1 - a1 * b1)
crime <- b0 + b1 * police + v

lm(crime ~ police)
```

```
##
## Call:
## lm(formula = crime ~ police)
##
## Coefficients:
## (Intercept)       police
##      0.4367      -0.4638
```

Real examples: - Development and Institution. Solution: settler mortality (Acemoglu, Johnson, Robinson 2001), Vietnam's / Peru's boundary (Melissa Dell) - Supply and Demand. Solution: storm that affects shrimp supply - Employment rate and minimum wage. Solution: compare McDonals across states' line with different mimimum wage, everything else the same (Card and Krueger)

Omitted variable bias

If x is correlated with z, that's endogeneity.

$$y = \beta_0 + \beta_1 x + (\beta_2 z + u)$$

Real example: - Corruption and culture. Omitted: Law. Solution: UN workers parking tickets in NYC (Fisman, Miguel) - Income and army experience. Omitted: Background, Income expectation. Solution: Vietnam draft lottery (Angrist)

A note on sample selection bias vs selection bias

These two are different!

- Sample selection bias: when you only observe a non-representative sample of the population (i.e. your sample is biased). Ex: We want to study the effect of civil war on GDP. But war torn countries can't collect GDP data, so you only observe non-war-torn countries. Thus your sample is biased.

Sample selection bias is a bias because your coefficient estimate is only true for your sample and different from the true parameter value of the population. But it's NOT endogeneity!

- Selection bias: your sample is representative of the population (i.e. it's different from sample selection bias). Instead, the problem is that your units self-select into the treatment for some unobserved and thus uncontrolled reasons.

Ex: Continuing the example above, let's say that you manage to collect GDP for all countries (again, so it's NOT a sample selection bias). But the problem is that countries "self-select" into wars if they have interethnic conflict, which we don't observe and control for. In this case, GDP is outcome, war is treatment, interethnic conflict is the omitted variable.

Power calculation

$$x \sim N(0, sd = 1)$$
$$y = 2x + N(1, sd = 10)$$

```
x <- rnorm(50)
y <- 2 * x + rnorm(50, mean = 1, sd = 10)
summary(lm(y ~ x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.807  -4.837  -0.986   5.249  24.723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3457     1.3166   1.022    0.312
## x             0.5031     1.3439   0.374    0.710
##
## Residual standard error: 9.224 on 48 degrees of freedom
## Multiple R-squared:  0.002911,   Adjusted R-squared:  -0.01786
## F-statistic: 0.1401 on 1 and 48 DF,  p-value: 0.7098
```

```
number_of_simulations <- 1000
pvalues <- rep(NA, number_of_simulations)
for (i in 1:number_of_simulations) {
  n <- 100
  x <- rnorm(n)
  y <- 2 * x + rnorm(n, mean = 1, sd = 10)
  pvalues[i] <- coef(summary(lm(y ~ x)))[2, 4]
}
mean(pvalues < 0.05)
```

```
## [1] 0.503
```

```
pvalues <- rep(NA, number_of_simulations)
for (i in 1:number_of_simulations) {
  n <- 200
  x <- rnorm(n)
  y <- 2 * x + rnorm(n, mean = 1, sd = 10)
  pvalues[i] <- coef(summary(lm(y ~ x)))[2, 4]
}
mean(pvalues < 0.05)
```

```
## [1] 0.789
```

Balance table

Basically t-tests for all covariates.

```r
mtcars$treatment <- sample(x = c(0, 1), size = nrow(mtcars), prob = c(0.5, 0.5),
                           replace = TRUE)

t.test(mpg ~ treatment, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by treatment
## t = -0.45717, df = 29.624, p-value = 0.6509
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.325901  3.378450
## sample estimates:
## mean in group 0 mean in group 1
##        19.57333        20.54706
```

```r
t.test(cyl ~ treatment, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  cyl by treatment
## t = 1.0426, df = 29.89, p-value = 0.3055
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6244216  1.9263824
## sample estimates:
## mean in group 0 mean in group 1
##        6.533333        5.882353
```

I don't think there's a package to automate this into a table yet. The closest is `MatchBalance`, whose results then need to be extracted and put into a table.

```r
library("Matching")
MatchBalance(treatment ~ mpg + cyl + disp + I(mpg^2), data = mtcars)
```

```
##
## ***** (V1) mpg *****
## before matching:
## mean treatment........ 20.547
## mean control.......... 19.573
## std mean diff......... 14.441
##
## mean raw eQQ diff..... 2.2533
## med  raw eQQ diff..... 2.2
## max  raw eQQ diff..... 4.6
##
## mean eCDF diff........ 0.11843
## med  eCDF diff........ 0.11373
## max  eCDF diff........ 0.2549
##
## var ratio (Tr/Co)..... 1.628
## T-test p-value........ 0.65089
## KS Bootstrap p-value.. 0.468
## KS Naive p-value...... 0.67847
```

```
## KS Statistic.......... 0.2549
##
##
## ***** (V2) cyl *****
## before matching:
## mean treatment........ 5.8824
## mean control.......... 6.5333
## std mean diff......... -33.683
##
## mean raw eQQ diff..... 0.66667
## med  raw eQQ diff..... 0
## max  raw eQQ diff..... 2
##
## mean eCDF diff........ 0.1085
## med  eCDF diff........ 0.054902
## max  eCDF diff........ 0.27059
##
## var ratio (Tr/Co)..... 1.4635
## T-test p-value........ 0.30551
## KS Bootstrap p-value.. 0.176
## KS Naive p-value...... 0.60392
## KS Statistic.......... 0.27059
##
##
## ***** (V3) disp *****
## before matching:
## mean treatment........ 223.49
## mean control.......... 238.91
## std mean diff......... -11.068
##
## mean raw eQQ diff..... 49.233
## med  raw eQQ diff..... 47
## max  raw eQQ diff..... 115.8
##
## mean eCDF diff........ 0.12694
## med  eCDF diff........ 0.10196
## max  eCDF diff........ 0.32941
##
## var ratio (Tr/Co)..... 1.6591
## T-test p-value........ 0.72751
## KS Bootstrap p-value.. 0.21
## KS Naive p-value...... 0.3528
## KS Statistic.......... 0.32941
##
##
## ***** (V4) I(mpg^2) *****
## before matching:
## mean treatment........ 464.97
## mean control.......... 409.18
## std mean diff......... 19.291
##
## mean raw eQQ diff..... 88.339
## med  raw eQQ diff..... 93.61
## max  raw eQQ diff..... 178.87
```

```
##
## mean eCDF diff........ 0.11843
## med  eCDF diff........ 0.11373
## max  eCDF diff........ 0.2549
##
## var ratio (Tr/Co)..... 1.346
## T-test p-value........ 0.56223
## KS Bootstrap p-value.. 0.468
## KS Naive p-value...... 0.67847
## KS Statistic.......... 0.2549
##
##
## Before Matching Minimum p.value: 0.176
## Variable Name(s): cyl  Number(s): 2
```