multiple regression

Adam Okulicz-Kozaryn
adam.okulicz.kozaryn@gmail.com

this version: Tuesday 13th February, 2024 17:58

outline

misc

intuition

trivariate

multiple

go and regress

F-tests

outline

ps1 comments

- get in habit of citing your data: name and url at least
- be clear bout u/a: alaways helps to state it explicitly
- keep it clean! the fewer the files in dropbox the better!

mechanics, again

◇ read carefully slides

◇ make sure you understand **everything crystal clear**

◇ if any sligtest doubts, mark it up, stop by my office

◇ unlike most other classes, some stuff is non-intitive

◦ let it digest, set aside, come back to it several times

◦ practice, practice, practice

outline

bivariate vs multivariate

- ◇ so far we have looked at the bivariate relationships
- ◇ today relax the very limiting assumption that the dv is predicted by only one iv
- ◇ and extend math to deal with more than one iv
- ◇ get into 'art' part and away from 'technical' part:
- ◦ more thinking, less math and plugging in numbers

multivariate/multiple OLS

- ◇ it's arguably the most common tool in social science
- ◇ finds effect of var of interest (main/key iv) on dv controlling/holding constant other vars
- ◇ a statistical trick that makes sample equal on all characteristics that we control for and "imitates" experimental setting (randomization)
  explain/draw picture
  - ○ in experiment you randomize into treatment and control groups so that both groups are on average the same and then we apply treatment (eg drug) to treatment group and see if had effect as compared to control group

Multivariate OLS

- ◇ most of the time we cant use experiment!
  - ○ cant tell some people to smoke and some not to
  - ○ cant tell some people to get education and others not to
  - ○ we can only use regression
- ◇ eg the effect of education (IV) on income (DV)
  - ○ may not be the same for males and females, and hence, we control for gender in regression
- ◇ as if everybody had the same gender! gender doesn't matter anymore!

multivariate OLS

- $X \rightarrow Y$
- $Y = f(X)$
- $Y = f(X_1, X_2, ..., X_n, u)$

yet, world is always more complicated than any OLS

◇ the world is more complicated than you can model

◇ soc sci relationships more complex than nat sci

∘ easy to predict what would make an airplane fly (speed, wings shape, and few more things)

∘ but what would make an economy grow ? almost infinite number of things

◇ your model oversimplifies world (that's why it's called a model)

cps example

- let's have a look at the relationships between wages, gender, experience, and marriage
- again, before regression *always* des sta
- great descriptive statistics is graphs!
- one of the most useful graphs is bar chart
- dofile: cps

a “complete” explanation

- wage=f(ability, education, age, gender, race, height, weight, strength, attitudes, neighb, family, interactions of the above...)
- multiple regression will tell you the effect of one variable while controlling for the effect of other variables (again, as if everybody was the same on other vars)
- $wage_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_n X_{ni} + u_i$
- look at cross-correlation matrix `pwcorr x1 x2 xk`

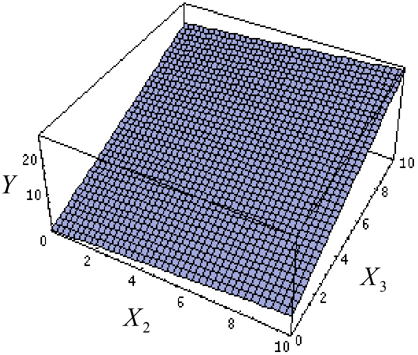
outline

trivariate regression

- (again, bivariate regression always biased)
- trivariate:

$$E(Y_i|X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$Y_i = E(Y_i|X_{2i}, X_{3i}) + u_i$$
$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

regression plane



- demonstartion:
  room's edges as axes

and sheet of paper as 3d

- $Y_i = 2 + 0.5X_{2i} + 2X_{3i} + u_i$
- $\hat{\beta}_2 = \frac{\Delta Y_i}{\Delta X_{2i}} = 0.5$
- $\hat{\beta}_3 = \frac{\Delta Y_i}{\Delta X_{3i}} = 2$
- we hold the other variable constant
- points above the plane are the positive residuals; below, negative residuals

adding assumption

- X's are not perfectly correlated
  - (squared term is not perfectly corr with regular term)
  - (they must be linearly realted, not non-linearly)
- example when they are?

what happens to rss?

- we hope that the new variable explains more of the variance in Y, but suppose $\hat{\beta}_3 = 0$
- $\sum e_i^2 = \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - [0] X_{3i})^2 = \sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i})^2$
  - same as the bivariate case!
- since ols minimizes rss, 3-var regression result will never have rss higher than the bivariate model
- rss will be lower, even if x_3 is random noise (eg bananas production in la will explain some of deaths on us hwys)

RSS declines, therefore R^2 Improves

- $(\sum e_i^2)^{trivariate} \leq (\sum e_i^2)^{bivariate}$
- the TSS is unchanged, so if RSS declines, the ESS (explained sum of squares) must increase
- so R^2 will improve:
- $R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} \qquad \frac{\text{declines}}{\text{no change}}$
- again, this is true even if X_3 is random noise or an irrelevant variable

how about estimate of uncertainty?

- $s = \sqrt{\frac{\sum e_i^2}{n-3}}$ $\frac{declines}{declines}$ so, what happens to s?
- bivariate: $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_i^2}}$
- trivariate: $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_{2i}^2(1-r_{23}^2)}}$

  $s_{\hat{\beta}_3} = \frac{s}{\sqrt{\sum x_{3i}^2(1-r_{23}^2)}}$
- $r_{23} = corr(X_2, X_3)$
- $-1 < r_{23} < 1$
- $0 \leq r_{23}^2 < 1$
- hence, in addition to the usual things, the variance of the slope depends on the corr between Xs

correlation between x's matters

- ◇ if $r_{23}^2 = 0$ then $s_{\hat{\beta}_2}$ is the same as in bivariate case
- ◇ if $r_{23}^2 = 1$ then $s_{\hat{\beta}_2}$ cannot be computed, cannot divide by 0
- ◦ this is why we assume no perfect corr between Xs
- ◦ non-perfect corr only makes the std err bigger
- ◇ as corr goes 0 to 1, or 0 to -1, denominator shrinks
- ◦ std err of the slope and CI inflate
- ◦
- ◇ so called collinearity and most of time
- ◦ the best thing: do nothing
- ◦ the worst thing: drop a var
- ◇ dofile: trivariate

collinearity

- collinearity/multicollinearity = corr among RHS vars
- don't do anything about it
- the problem of collinearity is that CI are wider
- but this is the nature of the data
- not a problem with your model
- conceptually same problem as "micronumerosity" (also wider CI)

calculations

- let's have a closer look at the regressions we just ran

hypothesis testing

$$wage_i = \underset{\substack{(1.219) \\ t=-4.02}}{-4.90} + \underset{\substack{(0.081) \\ t=11.38}}{0.93}(educ_i) + \underset{\substack{(0.017) \\ t=6.11}}{0.11}(exp_i)$$

$$H_0 : \beta_2 = \$0$$

$$H_A : \beta_2 \neq \$0$$

$$\alpha = 0.05$$

$$DOF = n - k = 531$$

$$\text{Reject } H_0 \text{ if } |t| > 1.96$$

$$t = \frac{0.93 - 0}{0.081} = 11.38$$

outline

the k-variable model

$$X_{2i}, X_{3i}, ..., X_{ki}$$
$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$$

- ◇ $e_i = Y_i - \hat{Y}_i$
- ◇ choose $\hat{\beta}_1, ..., \hat{\beta}_k$ to minimize $\sum e_i^2$
- ◇ k-var model not conceptually different from 3-var model
- ◇ just now more general assumption:
- ◇ no perfect correlation between any combination of X's
- ○ say cant include all races; need to omit one

the true meaning of multiple regression

- ◇ we say that beta is the effect "controlling" for the other variables
- ◦ but what does that really mean?
- ◦ how it controls for the other variables?
- ◦ dofile: truth

lovb (easy to see true meaning!)

- true model:

$$Y_i = \beta_1 + \beta_2 INCL + \beta_3 EXCL + u_i$$

- we estimate:

$$Y_i = \alpha_1 + \alpha_2 INCL + v_i$$

$$\mathrm{E}[\hat{\alpha}_2] = \alpha_2 = \beta_2 + \beta_3\left((\rho_{EI})\left(\frac{\sigma_E}{\sigma_I}\right)\right)$$

What you estimate using the 2 variable regression

The unbiased coefficient

The coefficient on the left out variable

rho is the bivariate correlation of the included and excluded variables

sign of bias: $\beta_3 * \rho_{EI}$

wages example

```
. sum educ exp
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        educ |       534    13.01873    2.615373          2         18
         exp |       534     17.8221    12.37971          0         55
```

$$\hat{\alpha}_2 = \hat{\beta}_2 + \hat{\beta}_3 \left((r) \left(\frac{s_{excluded}}{s_{included}} \right) \right)$$
$$= 0.93 + 0.11 \left(-0.35 \left(\frac{12.4}{2.6} \right) \right)$$
$$= 0.93 + 0.11(-1.669)$$
$$= 0.93 - 0.18$$
$$= 0.75$$

If experience didn't effect wage, OR if experience was uncorrelated with education, there would be no left out variable bias.

Another example: ability and education. Will there be a bias? In which direction?

neg bias

(-.18): biv coef .75 smaller than should be (true) (trivariate) .93

one more example

- ◇ counties with more Polish per 1k have more crime per 1k
- ◇ so regress crime/1k on Polish/1k and get + eff and sig
- ◇ Polish cause more crime?
- ◦ but lovb: educ, so whats the bias? (Polish don't get much education)

standardized coefficients

- $z_Y = \frac{Y_i - \bar{Y}}{s_Y} \quad z_{X2} = \frac{X_{2i} - \bar{X}_2}{s_{X_2}} \quad z_{X3} = \frac{X_{3i} - \bar{X}_3}{s_{X_3}}$
- regress: $\hat{z}_Y = \hat{\beta}_1^* + \hat{\beta}_2^* z_{x_2} + \hat{\beta}_3^* z_{x_3}$
- each β represents the effect on Y (measured in standard deviations of Y) of a one standard deviation change in each X variable – so you can compare the magnitudes of the coefficients

the 'beta' option

```
. sum wage educ exp
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       534    9.023939    5.138876          1       44.5
        educ |       534    13.01873    2.615373          2         18
         exp |       534     17.8221    12.37971          0         55
```

```
. reg wage educ exp, beta
      Source |       SS       df       MS              Number of obs =     534
-------------+------------------------------           F(  2,   531) =   67.22
       Model |  2843.72544     2  1421.86272           Prob > F      =  0.0000
    Residual |   11231.763   531   21.152096           R-squared     =  0.2020
-------------+------------------------------           Adj R-squared =  0.1990
       Total |  14075.4884   533  26.4080458           Root MSE      =  4.5991
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |    .925947   .0813995    11.38   0.000                 .4712502
         exp |   .1051282   .0171967     6.11   0.000                 .2532571
       _cons |  -4.904318   1.218865    -4.02   0.000                        .
------------------------------------------------------------------------------
```

$$\hat{\beta}_2^* = \hat{\beta}_2 \frac{s_X}{s_Y} = 0.926\left(\frac{2.615}{5.139}\right) = 0.471 \quad \hat{\beta}_3^* = \ldots$$

outline

academic research: how?

- ◇ have a research idea: a problem/question/hypothesis
- ◇ read about it, mostly peer reviewed (goog sch)
- ◇ find data with vars to test your hypotheses
- ○ des sta
- ◇ build model based on literature and your research idea
- ○ defend it: robustness/contribution/novelty
- ◇ interpret your results and discuss them
- ◇ regression is driven by correlation, but correlation by itself is never sufficient to prove causation – what do you need?

paper

- try to start getting at analyses that make research sense
- to do that, you need to read literature!
- will be back and forth:
- read lit, draft paper, run analyses
- reuse code from class! each class i give you dofile
- just copy paste and run on your vars :)
- and don't forget about:

  https://stats.oarc.ucla.edu/stata/webbooks/reg/

  http://www.princeton.edu/~otorres/Regression101.pdf

descriptive statistics! always do these!

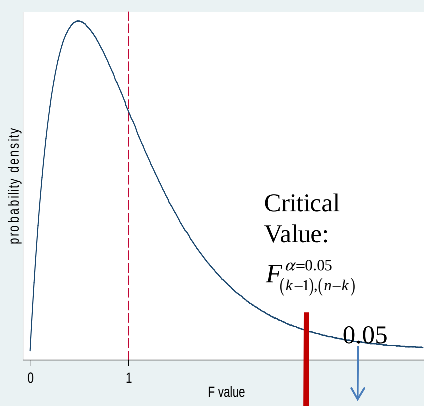
- histograms of dv and key ivs: hist x1; hist x2
- tabs and crosstabs (ordinal/nominal): tab x1; tab x1 x2
- cross correlation table: pwcorr x1 x2 y
- scatterplots: scatter y x
- and barcharts over various categories like we did with wages over gender over married!

outline

F-test

- $F = \frac{\text{explained variation per regressor}}{\text{Residual variation per degree of freedom}} = \frac{ESS/(k-1)}{RSS/(n-k)}$
- $F = \frac{\sum(\hat{Y}_i - \bar{Y})^2/(k-1)}{\sum e_i^2/(n-k)}$
- $F = \frac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} = \frac{R^2/(k-1)}{1-R^2/(n-k)}$
- (k=#of betas (incl interecept))
- see dofile for by-hand calculations

F-test

- $H_o : \beta_2 = \beta_3 = ... = \beta_k = 0$
- H_A : At least one $\beta \neq 0$
- 

- assuming the Null is true, the expected value of F is 1

F-test for restrictions

- UR: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$
- R: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + [0] X_{4i} + [0] X_{5i} + u_i$
- $H_0 : \beta_4 = \beta_5 = 0$
- H_A : at least one $\beta \neq 0$
- $F = \frac{ESS_U - ESS_R / m}{RSS_U / (n-k)}$ $\quad \frac{\text{m=\# of restrictions}}{\text{k=\#of betas (incl interecept) in UR}}$
- critical F: $(m, n-k)$
- dofile:F

chow test (F-test)

- ◇ chow test: just F-test for stability of betas across groups
- ○ eg: male v female; black vs white; before 00 v after 00
- ◇ run model and get RSS: RSS_R
- ◇ run the same model for each group separately:
- ○ $RSS_U = RSS_{male} + RSS_{female}$
- ◇ $F = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n-2k)}$
- ◇ dofile:chow

stata output

. regress Y X_2 X_3 … X_k , [beta]

Source	SS	df	MS
Model	$ESS=\sum\left(\hat{Y}_i-\bar{Y}\right)^2$	$k-1$	$\dfrac{ESS}{k-1}$
Residual	$RSS=\sum e_i^2$	$n-k$	$s^2=\dfrac{RSS}{n-k}$
Total	$TSS=\sum\left(Y_i-\bar{Y}\right)^2$	$n-1$	$s_Y^2=\dfrac{TSS}{n-1}$

Number of obs = n

$F(1,n-2)$ = $F=\dfrac{ESS/(k-1)}{RSS/(n-k)}$

Prob > F = p value for the model

R-squared = $R^2=1-\dfrac{RSS}{TSS}$

Adj R-Squared = $\bar{R}^2=1-\dfrac{RSS/(n-k)}{TSS/(n-1)}$

Root MSE = s

Y	Coef.	Std.Err.	t	P>|t|	[95% Conf.	Interval]	[Beta]
X_2	$\hat{\beta}_2$	$s_{\hat{\beta}_2}$	$\hat{\beta}_2/s_{\hat{\beta}_2}$	$\mathrm{H}_0:\beta_2=0$	$\hat{\beta}_2-t_{0.025}s_{\hat{\beta}_2}$	$\hat{\beta}_2+t_{0.025}s_{\hat{\beta}_2}$	$\hat{\beta}_2\left(s_{X_2}/s_Y\right)$
X_3	$\hat{\beta}_3$	$s_{\hat{\beta}_3}$	$\hat{\beta}_3/s_{\hat{\beta}_3}$	$\mathrm{H}_0:\beta_3=0$	$\hat{\beta}_3-t_{0.025}s_{\hat{\beta}_3}$	$\hat{\beta}_3+t_{0.025}s_{\hat{\beta}_3}$	$\hat{\beta}_2\left(s_{X_3}/s_Y\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$	$\hat{\beta}_k/s_{\hat{\beta}_k}$	$\mathrm{H}_0:\beta_k=0$	$\hat{\beta}_k-t_{0.025}s_{\hat{\beta}_k}$	$\hat{\beta}_k+t_{0.025}s_{\hat{\beta}_k}$	$\hat{\beta}_2\left(s_{X_k}/s_Y\right)$
$\_cons$	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$\hat{\beta}_1/s_{\hat{\beta}_1}$	$\mathrm{H}_0:\beta_1=0$	$\hat{\beta}_1-t_{0.025}s_{\hat{\beta}_1}$	$\hat{\beta}_1+t_{0.025}s_{\hat{\beta}_1}$	.