multiple regression

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

this version: Monday 5$^{\text{th}}$ February, 2024   16:50

outline

misc

intuition

trivariate

multiple

go and regress

F-tests

outline

misc

intuition

trivariate

multiple

go and regress

F-tests

ps1 comments

$\diamond$ get in habit of citing your data: name and url at least

$\diamond$ be clear bout u/a: alaways helps to state it explicitly

$\diamond$ keep it clean! the fewer the files in dropbox the better!

mechanics, again

$\diamond$ read carefully slides

$\diamond$ make sure you understand **everything crystal clear**

$\diamond$ if any sligtest doubts, mark it up, stop by my office

$\diamond$ unlike most other classes, some stuff is non-intitive

$\circ$ must let it digest, set it aside, come back to it several
times

$\circ$ practice, practice, practice

outline

bivariate vs multivariate

◇ so far we have looked at the bivariate relationships

◇ today we will relax the very limiting assumption that the dependent variable is predicted by only one independent variable

◇ and we will extend the math to deal with more than one independent variables

◇ we also start to 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

$\diamond$ 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

$\circ$ we can only use regression

$\diamond$ eg the effect of education (IV) on income (DV)

$\diamond$ may not be the same for males and females, and hence, we control for gender in regression

$\diamond$ The effect is as if everybody had the same gender ! gender doesn't matter anymore !

multivariate OLS

$\diamond\ X \rightarrow Y$

$\diamond\ Y = f(X)$

$\diamond\ Y = f(X_1, X_2, ..., X_n, u)$

yet, world is always more complicated than <u>any</u> OLS

$\diamond$ the world is more complicated than you can model

$\diamond$ social science relationships are more complex than natural
  science relationships

$\circ$ it is easy to predict what would make an airplane fly
  (speed, wings' shape, and few more things)

$\circ$ but what would make an economy grow ? there is an
  almost infinite number of things...

$\diamond$ 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 running regressions *always* do descriptive statistics

◇ great descriptive statistics is graphs!

◇ one of the most useful graphs is bar chart

◇ dofile: cps

a "complete" explanation

◇ wage=f(native ability, education, family background, age, gender, race, height, weight, strength, attitudes, neighborhood influences, family connections, interactions of the above, chance encounters,...)

◇ 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

misc

intuition

trivariate

multiple

go and regress

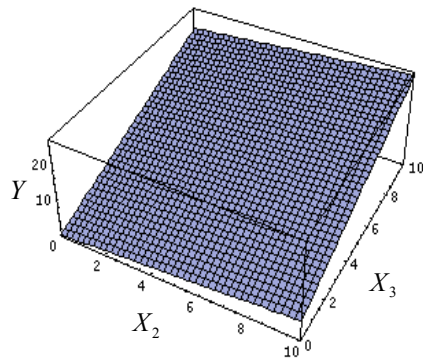F-tests

trivariate regression

$\diamond$ bivariate regression will be always biased

$\diamond$ a trivariate regression:

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

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

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

regression plane

Y, X_2, X_3 axes

$\diamond$ demonstartion:
room's edges as axes

and sheet of paper as 3d

$\diamond$ $Y_i = 2 + 0.5 X_{2i} + 2 X_{3i} + u_i$

$\diamond$ $\hat{\beta}_2 = \frac{\Delta Y_i}{\Delta X_{2i}} = 0.5$

$\diamond$ $\hat{\beta}_3 = \frac{\Delta Y_i}{\Delta X_{3i}} = 2$

$\diamond$ we hold the other
variable constant

$\diamond$ points above the plane
are the positive residuals;
below, negative residuals

adding assumption

$\diamond$ X's are not perfectly correlated

$\circ$ (squared term is not perfectly corr with regular term)

$\circ$ (they must be linearly realted, not non-linearly)

$\diamond$ 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 (try it–bananas production in la will explain some of deaths on us hwys)

RSS declines, therefore R^2 Improves

$\diamond$ $(\sum e_i^2)^{trivariate} \leq (\sum e_i^2)^{bivariate}$

$\diamond$ the TSS is unchanged, so if RSS declines, the ESS (explained sum of squares) must increase

$\diamond$ so R^2 will improve:

$\diamond$ $R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$ $\quad$ $\frac{\text{declines}}{\text{no change}}$

$\diamond$ 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

$\diamond$ if $r_{23}^2 = 0$ then $s_{\hat{\beta}_2}$ is the same as in bivariate case

$\diamond$ if $r_{23}^2 = 1$ then $s_{\hat{\beta}_2}$ cannot be computed, cannot divide by 0

$\circ$ this is why we assume no perfect corr between Xs

$\circ$ note that non-perfect correlation only makes the std. error of coefficient bigger...

correlated X's as a problem...

⋄ as correlation goes from 0 to 1, or 0 to -1, the term in the denominator shrinks

○ so the standard error of the slope "inflates."

○ larger variance of the slope coefficients means less precise estimates, wider confidence intervals, and higher p values on hypothesis tests

⋄ so called collinearity and most of time

○ the best thing to do is to do nothing

○ and the worst thing to do is to drop a variable

⋄ dofile: trivariate

collinearity

$\diamond$ collinearity/multicollinearity simply means correlation among RHS vars

$\diamond$ don't do anything about it

$\diamond$ the problem of collinearity is that CI are wider

$\diamond$ but this is the nature of the data

$\diamond$ not a problem with your model

$\diamond$ conceptually it is the same problem as "micronumerosity" (wider CI)

calculations

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

hypothesis testing

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

$H_0 : \beta_2 = \$0$

$H_A : \beta_2 \neq \$0$

$\alpha = 0.05$

$DOF = n - k = 531$

Reject H_0 if $|t| > 1.96$

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

<u>outline</u>

misc

intuition

trivariate

multiple

go and regress

F-tests

the k-variable model

$\diamond$ now we will extend the model to k-variables:

$X_{2i}, X_{3i}, ..., X_{ki}$

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

$\diamond$ $e_i = Y_i - \hat{Y}_i$

$\diamond$ choose $\hat{\beta}_1, ..., \hat{\beta}_k$ to minimize $\sum e_i^2$

$\diamond$ the k variable model is not conceptually different from the 3 variable model

adding a new assumption

◇ no perfect correlation between any combination of X's

the true meaning of multiple regression

◇ we say that beta is the effect "controlling" for the other variables

○ but what does that really mean?

○ in what way does it control for the other variables?

○ dofile: truth

partial correlation

◇ the partial correlation of Y and X_2 controlling for X_3 is the
correlation of Y and X_2 that is separate and distinct from
the correlation of Y and X_3

$r_{YX_2|X_3}$ or $r_{YX_2.X_3}$ or $r_{YX_2X_3}$

◇
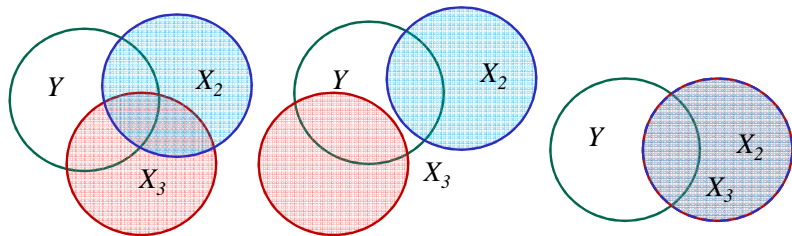
true meaning, conclusion

◇ β_2 in a bivariate regression reflects the linear correlation of the two variables

$\hat{\beta}_2 = r_{YX}(\frac{s_Y}{s_X})$

◇ β_2 in a 3-var regression reflects the correlation of X_2 and Y when both variables are purged of correlation with X_3 as we have just seen

$\hat{\beta}_2 = r_{YX_2|X_3}(\frac{s_Y}{s_X})$

◇ β_2 in k-var regression reflects the "partial correlation" of X_2 and Y controlling for $X_3...X_k$

$\hat{\beta}_2 = r_{YX_2|X_3...X_k}(\frac{s_Y}{s_X})$

◇ regression is driven by correlation, but correlation by itself is never sufficient to prove causation – what do you need?

standardized coefficients

◇ $z_Y = \frac{Y_i - \bar{Y}}{s_Y}$ $z_{X2} = \frac{X_{2i} - \bar{X}_2}{s_{X_2}}$ $z_{X3} = \frac{X_{3i} - \bar{X}_3}{s_{X_3}}$

◇ 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                        .
------------------------------------------------------------------------------
```

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

lovb

◇ 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;

coefficient is smaller than should (true)

outline

now you can predict anything !

$\diamond$ remember examples of predictions from the first class

$\circ$ airfare price

$\circ$ life expectancy

$\circ$ wine quality

$\diamond$ these days you can get data to study almost anything

$\diamond$ (avoid time series; try to have DV continuous)

paper, can do a lot with multiple regression

$\diamond$ it is really high time now to start your empirical paper due
  at the end of the class
$\diamond$ if you are stuck and cannot start email me
$\diamond$ if you started but have questions, email me
$\diamond$ you can test complex hypotheses
$\circ$ you can test interesting hypotheses
$\circ$ and contribute to the literature
$\diamond$ remember, world is always more complicated than your
  model

academic research: how?

◇ have a research idea: a problem/question/hypothesis
◇ read about it, mostly peer reviewed articles (goog sch)
○ write literature review
◇ find data that has vars that can be used to test your hypotheses
○ write about your data and show des stats
◇ build your model based on literature AND your research idea
○ write about your model and defend it robustness/contribution/novelty
◇ interpret your results and discuss them

paper

$\diamond$ try to start getting at analyses that make research sense

$\diamond$ to do that, you need to read literature!

$\diamond$ will be back and forth:

$\circ$ read lit, draft paper, run analyses

$\diamond$ reuse code from class! each class i give you dofile

$\circ$ just copy paste and run on your vars :)

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
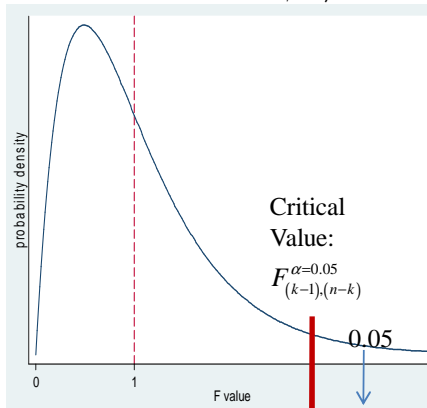
⬦ scatterplots: scatter y x

outline

F-test

$\diamond$ $F = \dfrac{\text{explained variation per regressor}}{\text{Residual variation per degree of freedom}} = \dfrac{ESS/(k-1)}{RSS/(n-k)}$

$\diamond$ $F = \dfrac{\sum(\hat{Y}_i - \bar{Y})^2/(k-1)}{\sum e_i^2/(n-k)}$

$\diamond$ $F = \dfrac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} = \dfrac{R^2/(k-1)}{1-R^2/(n-k)}$

F-test

◇ $H_o : \beta_2 = \beta_3 = ... = \beta_k = 0$

◇ $H_A :$ At least one $\beta \neq 0$

◇

◇ assuming that the Null is true, the expected value of F is 1

F-test for restrictions

$\diamond$ UR: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$

$\diamond$ R: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + [0]X_{4i} + [0]X_{5i} + u_i$

$\diamond$ $H_0 : \beta_4 = \beta_5 = 0$

$\diamond$ H_A : at least one $\beta \neq 0$

$\diamond$ $F = \frac{ESS_U - ESS_R/m}{RSS_U/(n-k)}$ m=# of restrictions
                                        k=#of betas (incl interecept) in UR

$\diamond$ critical F: $(m, n-k)$

$\diamond$ blackboard: draw a real example like in exam

$\diamond$ dofile:F

chow test (F-test)

◇ chow test is just an F-test that tests stability of betas across groups

○ eg: male v female; black vs white; before 00 v after 00

◇ first, run a model and get RSS – it will be your RSS_R

◇ second, run the same model for each group separately and get:

○ $RSS_U = RSS_{male} + RSS_{female}$

◇ $F = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n-2k)}$

◇ dofile:chow

adj Rsq

- $R^2 = 1 - \frac{RSS}{TSS}$ $adj.R^2 = \bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{s^2}{s_Y^2}$

- regular R^2 always increases when new variables added, even if they are just noise

- Adj. R^2 "corrects" for degrees of freedom

- can decline, or even become negative

- widely used, but not very useful

- neither accurate as a description nor a valid test statistic for some hypothesis

- don't use it

- if you see it ignore it and complain

- if you are concerned about the significance of a variable or variables, look to t and F tests

stata output

`. regress Y X₂ X₃ … Xₖ , [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 $= n$

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

Prob > F $\quad = \quad$ p value for the model

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

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

Root MSE $\quad = \quad s$

Y	Coef.	Std.Err.	t	P>|t|	[95% Conf.	Interval]	[Beta]
X_2	$\hat{\beta}_2$	$s_{\hat{\beta}_2}$	$\hat{\beta}_2 / s_{\hat{\beta}_2}$	$H_0 : \beta_2 = 0$	$\hat{\beta}_2 - t_{0.025} s_{\hat{\beta}_2}$	$\hat{\beta}_2 + t_{0.025} s_{\hat{\beta}_2}$	$\hat{\beta}_2\left(s_{X_2}/s_Y\right)$
X_3	$\hat{\beta}_3$	$s_{\hat{\beta}_3}$	$\hat{\beta}_3 / s_{\hat{\beta}_3}$	$H_0 : \beta_3 = 0$	$\hat{\beta}_3 - t_{0.025} s_{\hat{\beta}_3}$	$\hat{\beta}_3 + t_{0.025} s_{\hat{\beta}_3}$	$\hat{\beta}_2\left(s_{X_3}/s_Y\right)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_k	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$	$\hat{\beta}_k / s_{\hat{\beta}_k}$	$H_0 : \beta_k = 0$	$\hat{\beta}_k - t_{0.025} s_{\hat{\beta}_k}$	$\hat{\beta}_k + t_{0.025} s_{\hat{\beta}_k}$	$\hat{\beta}_2\left(s_{X_k}/s_Y\right)$
_cons	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$\hat{\beta}_1 / s_{\hat{\beta}_1}$	$H_0 : \beta_1 = 0$	$\hat{\beta}_1 - t_{0.025} s_{\hat{\beta}_1}$	$\hat{\beta}_1 + t_{0.025} s_{\hat{\beta}_1}$	.