

multiple regression

Adam Okulicz-Kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Tuesday 21st February, 2017 10:18

outline

misc

intuition

trivariate

multiple

go and regress

outline

misc

intuition

trivariate

multiple

go and regress

ps1 comments

- ◇ note, i am saving commenst as ans* in dropbox
- ◇ and then search that dofile (ctrl-f)
- ◇ for " //aok:"
- ◇ do stuff in stata like descriptive stats for full sample, not just 4 obs
 - more interesting that way! :)
- ◇ always have **clear** and **replace**
 - so that it doesnt break on 2nd run
- ◇ get in habit of citing your data: name and url at least
- ◇ be clear bout u/a: always helps to state it explicitly
- ◇ keep it clean! the fewer the files in dropbox the better!

note: departure from pre2017

- ◇ dropped:
 - F-test incl chow test
 - adjRsqr
- ◇ they won't appear in exams, neither PhD Quant Qualifier

outline

misc

intuition

trivariate

multiple

go and regress

bivariate vs multivariate

- ◇ so far we have looked at the bivariate relationships
- ◇ today we will relax the very limiting assumption that the dependent variable can be 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 OLS

- ◇ Multiple (multivariate) regression is arguably the most common quantitative tool in social science
- ◇ The idea is to find effect of a variable of interest on the dependent variable **controlling/holding constant other vars**
- ◇ It is 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 cannot use experiment—we cannot tell some people to smoke and some not to; we cannot tell some people to get education and others not to
 - ... we can only use regression
- ◇ For instance, we investigate the effect of education (IV) on income (DV)
- ◇ But it may not be the same for males and females, and hence, we control for gender in regression
- ◇ The effect is as if everybody had the same gender !
gender doesn't matter anymore !

multivariate OLS

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

yet, world is always more complicated than any OLS

- ◇ the idea is that the world is more complicated than you can model
- ◇ social science relationships are more complex than natural science relationships
 - it is easy to predict what would make an airplane fly (speed, wings' shape, and few more things)
 - but what would make an economy grow ? there is an almost infinite number of things...
- ◇ your model oversimplifies world (that's why it's called a model)
- ◇ we will talk more about this later

cps example

- ◇ let's have a look at the relationships between wages, gender, experience, and marriage
- ◇ again, before running regressions **always** do descriptive statistics
- ◇ again, a great way to produce descriptive statistics is to use graphs
- ◇ one of the most useful graphs is bar chart
- ◇ dofile: cps

a “complete” explanation

- ◇ $wage = f(\text{native ability, education, family background, age, gender, race, height, weight, strength, attitudes, neighborhood influences, family connections, interactions of the above, chance encounters, ...})$
- ◇ 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)
- ◇ $wage_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_n X_{ni} + u_i$

outline

misc

intuition

trivariate

multiple

go and regress

trivariate regression

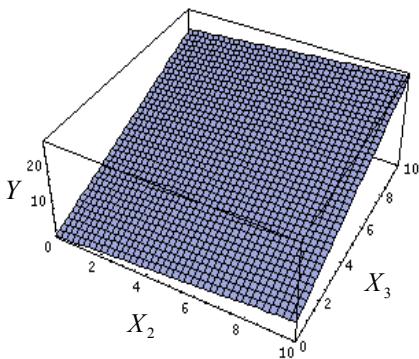
- ◇ virtually always bivariate regression will be biased
 - the disturbance term includes the effects of other variables, which leads to a correlation between the disturbance term and the X in a bivariate regression
 - this violates the assumption needed for the coefficient to be unbiased
- ◇ we begin with 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



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

◇ demonstration:

adding assumption

- ◇ X 's are not perfectly correlated
 - (note squared term is not perfectly corr with regular term)
 - (they must be linearly related, not non-linearly)
- ◇ give me 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
- ◇ virtually always, rss will be lower, even if x_3 is random noise (try it – again, bananas production in la will explain a big portion of deaths on US highways)

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
- ◇ as a consequence, R^2 will improve:
- ◇ $R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$ declines
no change
- ◇ again, this is true even if X_3 is random noise or an irrelevant variable

how about estimate of uncertainty?

◇ $s = \sqrt{\frac{\sum e_i^2}{n-3}}$ *declines declines* so, what happens to s ?

◇ bivariate: $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_i^2}}$

◇ trivariate: $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_{2i}^2(1-r_{23}^2)}}$

$$s_{\hat{\beta}_3} = \frac{s}{\sqrt{\sum x_{3i}^2(1-r_{23}^2)}}$$

• $r_{23} = \text{corr}(X_2, X_3)$

• $-1 < r_{23} < 1$

• $0 \leq r_{23}^2 < 1$

• hence, in addition to the usual things, the variance of the slope depends on the correlation between the X variables

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, because you cannot divide by 0
- and this is why we assume no perfect correlation between X's
- 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, thus...
 - 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
- ◇ this is 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

- ◇ collinearity/multicollinearity simply means correlation 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 it is the same problem as “micronumerosity” (wider CI)

calculations

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

hypothesis testing

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

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

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

$$\alpha = 0.05$$

$$DOF = n - k = 531$$

Reject H_0 if $|t| > 1.96$

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

comparing sums of squares ['1-' in rsq calc]

. reg wage educ

Source	SS	df	MS
Model	2053.22494	1	2053.22494
Residual	12022.2635	532	22.5982396
Total	14075.4884	533	26.4080458

Number of obs = 534
 F(1, 532) = 90.86
 Prob > F = 0.0000
 R-squared = 0.1459
 Adj R-squared = 0.1443
 Root MSE = 4.7538

. reg wage educ exp

Source	SS	df	MS
Model	2843.72544	2	1421.86272
Residual	11231.763	531	21.152096
Total	14075.4884	533	26.4080458

Number of obs = 534
 F(2, 531) = 67.22
 Prob > F = 0.0000
 R-squared = 0.2020
 Adj R-squared = 0.1990
 Root MSE = 4.5991

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} = \frac{11232}{14075} = 0.2020 \quad s = \sqrt{\frac{\sum e_i^2}{n-3}} = \sqrt{\frac{11232}{531}} = 4.599$$

$$s_{\hat{\beta}_2} = \frac{s}{\sqrt{(\sum x_{2i}^2)(1-r_{23}^2)}} = \frac{4.599}{\sqrt{(3645)(1-0.123)}} = 0.081$$

calculations for denominator

The quantities referred to in the denominator of the standard error of the slope on the previous page are not in the regression output, but are readily available.

$$r_{23} = -0.35 \rightarrow r_{23}^2 = 0.123$$

$$s_{X_2} = \sqrt{\frac{\sum (X_{2i} - \bar{X}_2)^2}{n-1}} \rightarrow (n-1)s_{X_2}^2 = \sum (X_{2i} - \bar{X}_2)^2 = \sum x_{2i}^2$$

$$\sum x_{2i}^2 = (534-1)(2.615^2) = 3645$$

The numbers come from the output from `sum` and `corr` on an earlier slide.

outline

misc

intuition

trivariate

multiple

go and regress

the k-variable model

- ◇ now we will extend the model to k-variables:

$$X_{2i}, X_{3i}, \dots, X_{ki}$$

- ◇ $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$

- ◇ $e_i = Y_i - \hat{Y}_i$

- ◇ choose $\hat{\beta}_1, \dots, \hat{\beta}_k$ to minimize $\sum e_i^2$

- ◇ the solution is not possible to write in general form with algebra

- ◇ still, 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

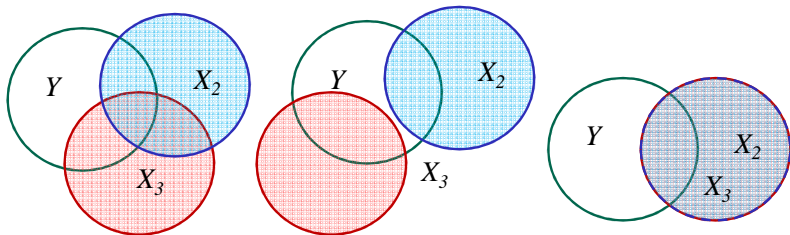
the true meaning of multiple regression

- ◇ 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

$$r_{YX_2|X_3} \text{ or } r_{YX_2 \cdot X_3} \text{ or } r_{YX_2X_3}$$



true meaning, conclusion

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

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

- ◇ 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

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

- ◇ β_2 in k-var regression reflects the “partial correlation” of X_2 and Y controlling for $X_3 \dots X_k$

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

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

standardized coefficients

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

$$\diamond \text{regress: } \hat{z}_Y = \hat{\beta}_1^* + \hat{\beta}_2^* z_{X2} + \hat{\beta}_3^* z_{X3}$$

- \diamond 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	
Model	2843.72544	2	1421.86272	F(2, 531) =	67.22
Residual	11231.763	531	21.152096	Prob > F =	0.0000
				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	.

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

lovb

- ◇ true model:

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

- ◇ we estimate:

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

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

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

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	534	13.01873	2.615373	2	18
exp	534	17.8221	12.37971	0	55

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

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)

a special case

- ◇ sysuse auto, clear
corr price wei len
reg price wei
//so wei has positive bias so should be overest
//but the result is opposite!
reg price wei len

cont

- ◇ the effect of the excluded variable on DV should be anticipated from the full “true” model as in few slides above: β_3 , not from the bivariate relationship ! but in this case it is impossible to predict that it would be negative ! so this is rather a bad example—a very specific case where sign flips in trivariate regression—it rarely happens!

cont

- ◇ sometimes, as in my example, it is impossible to guess what the effect would be in the true regression based just on correlations—in other words, sometimes, it does not help to know how each of the variables is related to one another to predict sign of the bias

outline

misc

intuition

trivariate

multiple

go and regress

you are a phd student

- ◇ so you must have some research ideas
- ◇ otherwise you should not be here!
- ◇ phd=research (not: civic engagement, social work, community development, etc)
- ◇ regression is an excellent tool for research
- ◇ use it (or lose it)
- ◇ sure, can do qualitative research...but

Lord Kelvin

◇ https://en.wikiquote.org/wiki/William_Thomson

now you can predict anything !

- ◇ remember examples of predictions from the first class
 - airfare price
 - life expectancy
 - wine quality
- ◇ you can use regression to predict anything !
- ◇ most of the time regression predictions will be more accurate than expert predictions
- ◇ these days you can get data to study almost anything
- ◇ (avoid time series; try to have DV continuous)

paper

- ◇ it is really high time now to start your empirical paper due at the end of the class
- ◇ if you are stuck and cannot start email me
- ◇ if you started but have questions, email me
- ◇ you will present your paper at the end of semester
- ◇ it's only few weeks

you can do a lot with multiple regression

- ◇ you can test complex hypotheses
 - you can test interesting hypotheses
 - and contribute to the literature
- ◇ remember, world is always more complicated than your model
- ◇ the most interesting are interactions (next wk)
 - interactions are a great way to get closer to the real complexity

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