multiple regression 1

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

this version: Tuesday 3rd March, 2015 10:20

outline

misc

intuition

trivariate

multiple

go and regress

outline

logs again

- logs may be confusing...
- let's revisit slides from last week about logs again!

looking ahead

- we will have intensive next 2 weeks, and then less work...
- by the end of 1sf half of semester we will be done with most of the class material
- then practice and revision
- next week we will talk more about multiple regression
- note lots of extra materials for the next class
- then we will have left violations of assumptions
- and this will be it

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 can be predicted by only one independent variable
- and we will extend the math to deal with more than one independent variables

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 (e.g. 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

- $X \rightarrow Y$
- $Y = f(X)$
- $Y = f(X_1, X_2, ..., X_n, u)$
- If the variable X determines Y, we say that Y is a function of X. However, in the realm of social sciences, all variables of interest are functions of multiple explanatory variables, as well as random components.

still, the world is more complicated than any OLS

- the idea is that the world is more complicated than you think
- 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 in last class in 1st half of semester

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(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)
- $wage_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_n X_{ni} + u_i$

outline

misc

intuition

trivariate

multiple

go and regress

trivariate regression

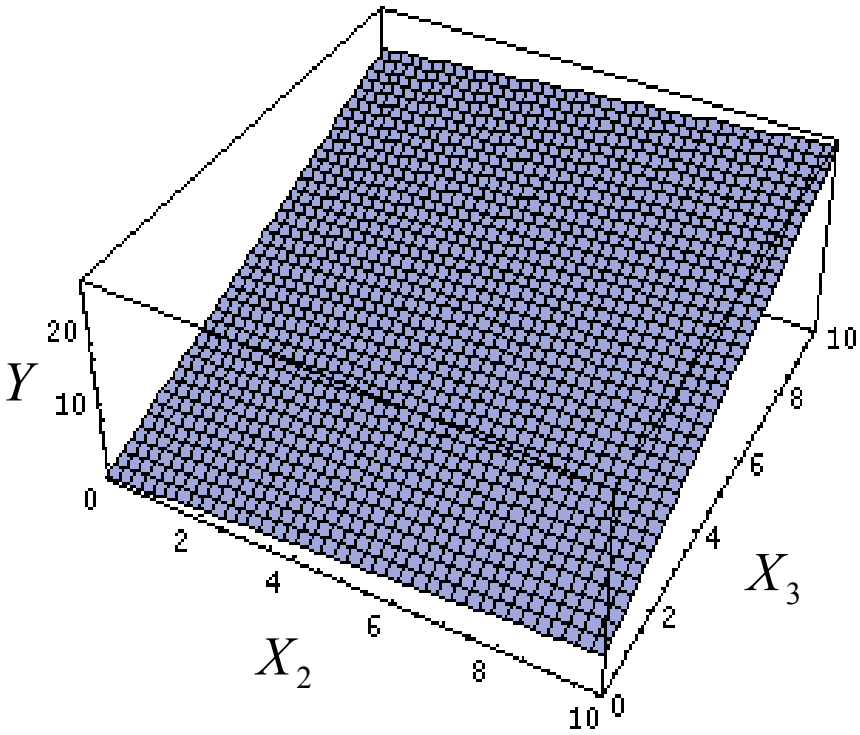
- virtually always bivariate regression will be biased
  - the disturbance term includes the effects of other variables, which leads to a correlation between the disturbance term and the X in a bivariate regression
  - this violates the assumption needed for the coefficient to be unbiased
- we begin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

- demonstartion:

adding assumption

- X's are not perfectly correlated

SRF

- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
- $e_i = Y_i - \hat{Y} = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}) =$
  $Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}$

the solution

- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$
- $\hat{\beta}_2 = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum y_i x_{3i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}$
- $\hat{\beta}_3 = \frac{\sum y_i x_{3i} \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}$
- adding one more variable, the formulas become horribly more complicated, even in deviations notation! (it gets much worse very quickly)
- each slope coefficient depends on a complex mingling of both X_2 and X_3.
- still, like the bivariate case, the estimators are just combinations of deviations; the covariance between X_2 and X_3 shows up in both slopes

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

how about estimate of uncertainty?

- $s = \sqrt{\frac{\sum e_i^2}{n-3}}$ $\frac{\textit{declines}}{\textit{declines}}$ 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 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

calculations

- let's have a closer look at the regressions we just ran
- and let's calculate by hand the output numbers
- (this is an excellent quiz question)

hypothesis testing

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

$H_0 : \beta_2 = \$0$

$H_A : \beta_2 \neq \$0$

$\alpha = 0.05$

$DOF = n - k = 531$

Reject H_0 if $|t| > 1.96$

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

$H_0 : \beta_2 = \$1$

$H_A : \beta_2 \neq \$1$

$\alpha = 0.05$

$DOF = n - k = 531$

Reject H_0 if $|t| > 1.96$

$$t = \frac{0.93 - 1}{0.081} = 0.86$$

comparing the anova tables ['1-' in rsq calc]

```
. reg wage educ
      Source |       SS       df       MS              Number of obs =     534
-------------+------------------------------           F(  1,   532) =   90.86
       Model |  2053.22494     1  2053.22494            Prob > F      =  0.0000
    Residual |  12022.2635   532  22.5982396           R-squared     =  0.1459
-------------+------------------------------           Adj R-squared =  0.1443
       Total |  14075.4884   533  26.4080458           Root MSE      =  4.7538

.reg wage educ exp
      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
```

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

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

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.

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

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

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

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: $X_{2i}, X_{3i}, ..., X_{ki}$
- SRF: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki} = \sum_{j=1}^{k} \hat{\beta}_j X_{ji} \quad X_{1i} = 1$
- $e_i = Y_i - \hat{Y}_i = Y_i - \sum_{j=1}^{k} \hat{\beta}_j X_{ji}$
- choose $\hat{\beta}_1, ..., \hat{\beta}_k$ 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

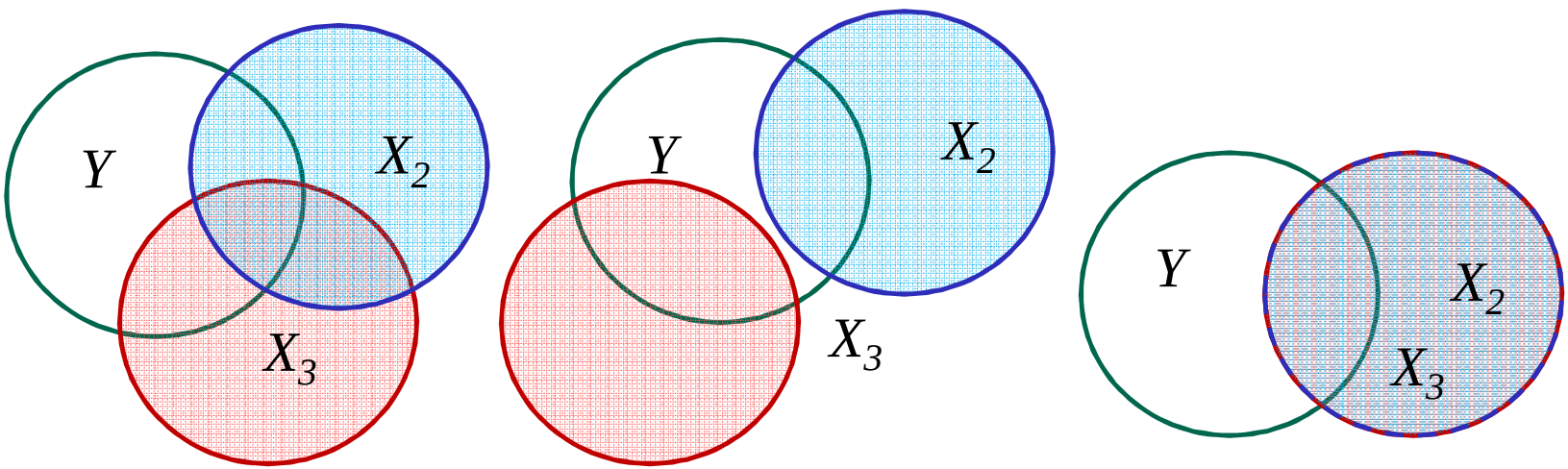
the true meaning of multiple regression

- 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

  $r_{YX_2|X_3}$ 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

- $z_Y = \frac{Y_i - \bar{Y}}{s_Y} \quad z_{X2} = \frac{X_{2i} - \bar{X}_2}{s_{X_2}} \quad z_{X3} = \frac{X_{3i} - \bar{X}_3}{s_{X_3}}$
- 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
- $\hat{\beta}_1^* = \bar{z}_Y - \hat{\beta}_2^* \bar{z}_{X_2} - \hat{\beta}_3^* \bar{z}_{X_3} = 0$
- $\hat{\beta}_2^* = r_{X_2,Y|X_3}(\frac{s_Y}{s_{X_2}}) = r_{X_2,Y|X_3}(\frac{1}{1}) = r_{X_2,Y|X_3}$
- $\hat{\beta}_3^* = r_{X_3,Y|X_2}(\frac{s_Y}{s_{X_3}}) = r_{X_3,Y|X_2}(\frac{1}{1}) = r_{X_3,Y|X_2}$

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                        .
------------------------------------------------------------------------------
```

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

anatomy of lovb

- true model:

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

- we estimate:

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

- write *EXCL* as a function of *INCL*:

$$EXCL_i = \gamma_1 + \gamma_2 INCL_i + \epsilon_i \quad \hat{\gamma}_2 = r_{EI} \frac{s_E}{s_I}$$

- so, in the 2 var model we actually estimate this:

$$Y_i = \beta_1 + \beta_2 INCL + \beta_3(\gamma_1 + \gamma_2 INCL + \epsilon_i) + u_i$$

$$Y_i = (\beta_1 + \beta_3\gamma_1) + (\beta_2 + \beta_3\gamma_2) INCL + (u_i + \beta_3\epsilon_i)$$

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

lovb

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

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

- since the standard deviations are always positive, the sign of the bias is determined by
  1) the coefficient of the excluded variable and
  2) the correlation of the included and excluded variables

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

misc

intuition

trivariate

multiple

go and regress

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 that expert predictions
- these days you can get data to study almost anything

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 July
- it's only 5 weeks ...