bivariate regression 2

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Saturday 9th December, 2017 08:20

outline

misc

stat significance (hypothesis testing)

basic measurement

outline

misc

stat significance (hypothesis testing)

basic measurement

ps1

- ◇ note: on 1-31 i beefed up a bit ps1
- ◇ still need stata lab next week?
- ◇ ps1: finding data: be opportunistic:
  - cannot find data for neighborhoods?
  - study cities, counties, or states
- ◇ do not have exact variable you need?
  - study sth similar!
- ◇ read literature! and follow it! (in terms of data too)
  - your first studies should closely follow published examples
  - just tweak them a little bit

anatomy of stata output **dofile: outlier**

```
. regress DV IV
```

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

Number of obs	=	n
$\mathrm{F}(1, n-2)$	=	....
Prob > F	=	....
R-squared	=	r^2
Adj R-Squared	=	....
Root MSE	=	s

DV	Coef.	Std.Err.	t	P>|t|	[95% Conf.	Interval]
IV	$\hat{\beta}_2$	$s_{\hat{\beta}_2}$	$\left(\frac{\hat{\beta}_2}{s_{\hat{\beta}_2}}\right)$	p val. for H_0 that $\beta_2 = 0$	$\hat{\beta}_2 - t_{0.025} s_{\hat{\beta}_2}$	$\hat{\beta}_2 + t_{0.025} s_{\hat{\beta}_2}$
Intercept	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$\left(\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right)$	p val. for H_0 that $\beta_1 = 0$	$\hat{\beta}_1 - t_{0.025} s_{\hat{\beta}_1}$	$\hat{\beta}_1 + t_{0.025} s_{\hat{\beta}_1}$

today and looking ahead

- ◇ let's begin by repeating key stuff from last class
- ◇ and then we'll add stat significance
- ◇ next week we will start multiple regression
  - over time, esp after midterm, class will get more applied
  - and we will have more examples

outline

misc

stat significance (hypothesis testing)

basic measurement

basic calculations blackboard; dofile

Y	X	y	y2	x	x2	xy
1	17					
3	13					
5	8					
7	10					
9	2					

Sum:
25 50

$\bar{Y} = 5 \quad \bar{X} = 10$

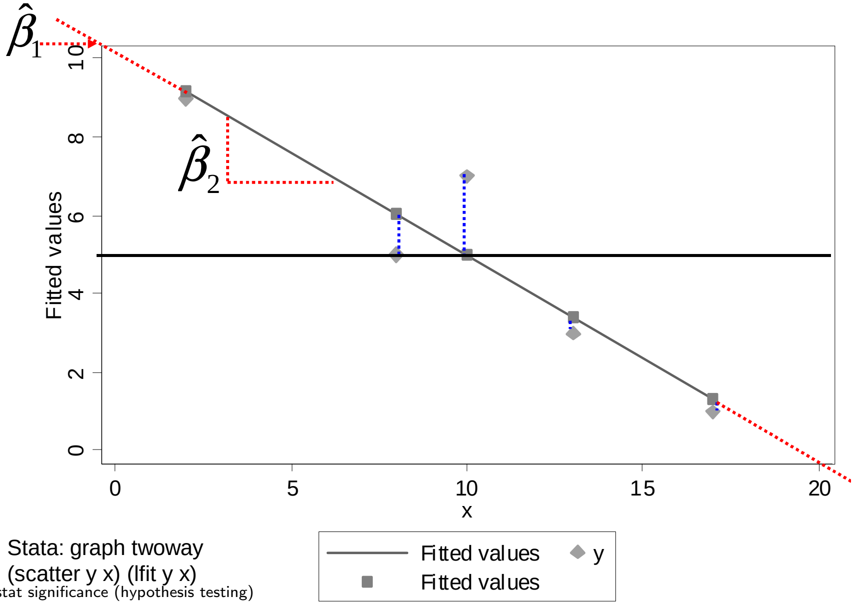
the coefficients–interpretation

- Beta hat two is the slope coefficient. Thus, a one unit change in X leads to a 0.524 decrease in Y. Beta hat one is the intercept term. It is the predicted value for Y when X is equal to zero.

predicted val and resid **blackboard; dofile**

Y	X	Y_hat	e	e^2
1	17			
3	13			
5	8			
7	10			
9	2			
- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
- for obs 1:
- $\hat{Y}_1 = 10.24 + (-0.524)(17) = 1.332$
- $e_1 = 1 - 1.33 = -0.33$

regression plot again



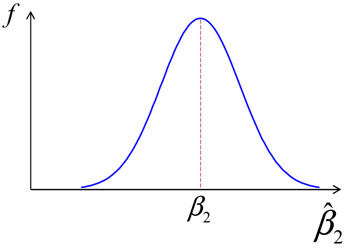
Stata: graph twoway
(scatter y x) (lfit y x)

se of the slope **blackboard; dofile**

- $\sum e_i^2 = 5.42$
- $s = \sqrt{\frac{\sum e_i^2}{n-2}} =$
- $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_i^2}}$
  - it gives us info about reliability (like sd or se) of slope

sampling distribution of the slope

probability distribution of $\hat{\beta}_2$ is centered on the true value of the parameter (i.e. unbiased) and is normally distributed with variance:



- $s^2_{\hat{\beta}_2} = \frac{s^2}{\sum x_i^2}$
- $s_{\hat{\beta}_2} = \sqrt{\frac{s^2}{\sum x_i^2}} = \frac{s}{\sqrt{\sum x_i^2}}$

hypothesis test dofile

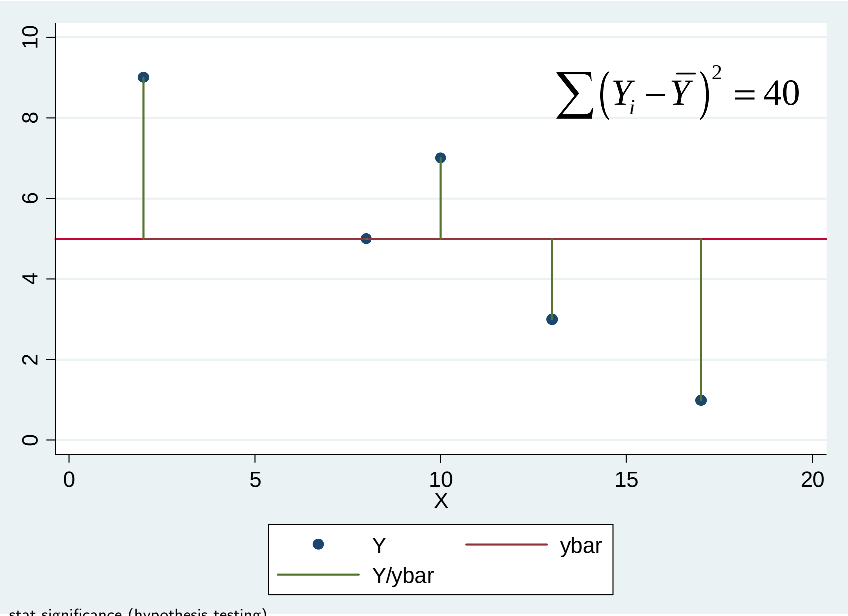
- the null is that slope ("the unobserved true parameter")
  - is zero (ie no effect)
- $H_0 : \beta_2 = 0$
- $H_A : \beta_2 \neq 0$
- $t = \frac{\hat{\beta}_2 - \beta_2}{s_{\hat{\beta}_2}} = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}}$
- CI: $\hat{\beta}_2 \pm (t_{n-2,\frac{\alpha}{2}})(s_{\hat{\beta}_2})$

partitioning variance in Y dofile

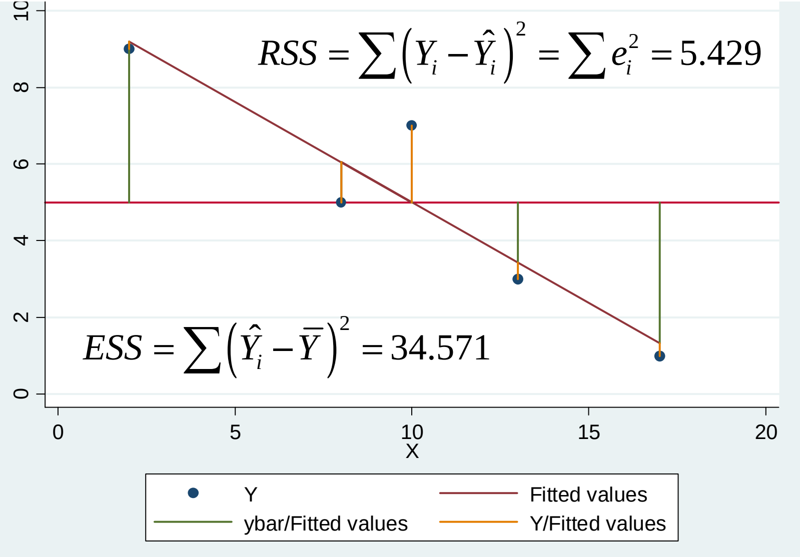
- before regression $E[Y_i] = \bar{Y}$
  - $TSS = \sum(Y_i - \bar{Y})^2 = \sum y_i^2 = 40$
- after regression $E[Y_i|X_i] = \hat{Y}_i$
  - $RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2 = 5.43$
  - $ESS = TSS - RSS = 40 - 5.4 = 34.57$

- $R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$
- proportion of the total variance in the Y explained by Xs
- $0 \leq R^2 \leq 1$

TSS



RSS



$RSS = \sum \left(Y_i - \hat{Y}_i \right)^2 = \sum e_i^2 = 5.429$
$ESS = \sum \left(\hat{Y}_i - \overline{Y} \right)^2 = 34.571$
0
2
4
6
8
10
0
5
10
15
20
X
Y
Fitted values
ybar/Fitted values
Y/Fitted values

exercise 1 dofile

- you regressed car's price on its weight

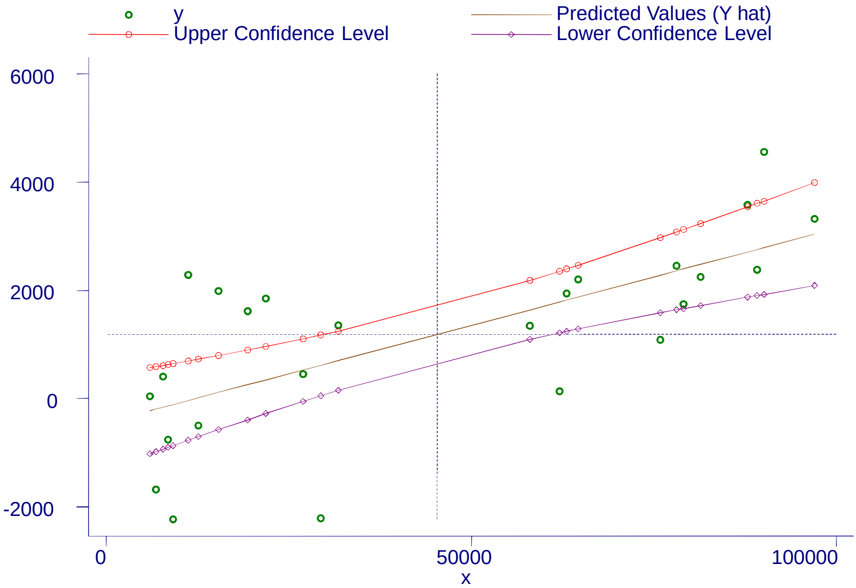
```
------------------------------------
       price |      Coef.   Std. Err.
-------------+----------------------
      weight |   2.044063   .3768341
```

- interpret the coefficient
- is it significant ?
- calculate 95% CI

reliability of predict val (se of $E(Y|X)$)

- We have discussed the fact that parameter estimates are random variables, and so they have standard errors. Predicted values are also random variables because they are linear combinations of the coefficients.
- The further from the mean of X, the wider the confidence interval around the predicted value.
- leave it to software, no need to know the formula

se of $E(Y|X$ illustration dofile



outline

misc

stat significance (hypothesis testing)

basic measurement

intuition

- ◇ what happens to betas if we change variables' measurement?
  - millions of dollars as opposed to dollars
  - curved grades (each person gets extra 10 points)
  - proportion of people in poverty v percent in poverty
- ◇ income per capita v income per 100k people

add constant c to X or Y (say curved grades)

- if you add c to each obs, mean of var would change by that much
- but demeaned var doesn't change:
- $x_i^{'} = (X_i^{'} - \bar{X}^{'}) = [(X_i + c) - (\bar{X} + c)] = x_i$ same for Y
- $\hat{\beta}_2 = \frac{\sum y_i x_i^{'}}{\sum x_i^{'2}} = \frac{\sum y_i x_i}{\sum x_i^2}$ only demeaned vars so no change
- and nobody cares about intercept anyway, so let's spare our brain

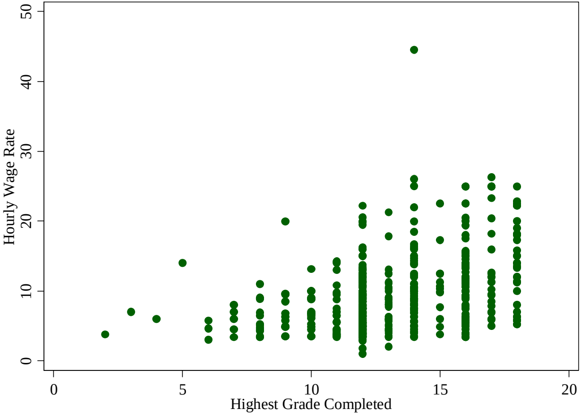
multiply X or Y by constant (say months, not years)

- ◇ think about it, assume some example
  - say year of educ produces \$2 increase in wage
- ◇ how about a month of educ? should be 1/12 of \$2 !
- ◇ to convert yr to mo, multiply years by 12, right?
  - if a person has 2yr of educ, that's 24mo
- ◇ so if i multiply X by c, say 12, I need to divide $\hat{\beta}_2$ by 12
- ◇ what if multiply Y?
  - again, say year of educ produces \$2 increase in wage
  - ...or 200 cent increase in wage
- ◇ to get cents from dollars, I multiply dollars by 100
  - so if I multiply Y by 100, i get β_2 100x bigger

fun fact1: correlation v bivariate regression

- $r = \frac{\sum y_i x_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \quad \hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2}$
- bivariate slope equals corr coef scaled by std dev of Y and X:
  $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = r\left(\frac{s_Y}{s_X}\right)$

education and wages dofile




```
. corr wage educ
(obs=534)

             |     wage     educ
-------------+------------------
        wage |   1.0000
        educ |   0.3819   1.0000
```

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

education and wages dofile

```
. regress 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

------------------------------------------------------------------------------
    wage |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    educ |   .7504488     .07873      9.532   0.000       .5957891    .9051086
   _cons |   -.745949   1.045404     -0.714   0.476      -2.799576    1.307678
------------------------------------------------------------------------------
```

The estimated regression line:

$$\widehat{wage}_i = \hat{\beta}_1 + \hat{\beta}_2 educ_i = -0.75 + 0.75 educ_i$$

Interpret the coefficients.

fun fact2: Z scores bivariate regression=correlation

- $z_{Yi} = \beta_1 + \beta_2 z_{Xi} + u_i$
  $z_{Xi} = \frac{X_i - \bar{X}}{s_X} = \frac{x_i}{s_x}$
  $z_{Yi} = \frac{Y_i - \bar{Y}}{s_U} = \frac{y_i}{s_Y}$
- z scores always have a mean of 0 and a variance (and standard deviation) of 1
- $\hat{\beta}_2 = r_{Z_Y Z_X} \frac{s_{Z_Y}}{s_{Z_X}} = r_{YX}$
  $\hat{\beta}_1 = \bar{z}_Y - \hat{\beta}_2 \bar{z}_X = 0 - r(0) = 0$
- Thus, a regression of the z scores of Y on the z scores of X produces a slope equal to the correlation coefficient of X and Y and a zero intercept.

exercise 2: if no time do at home: see **dofile**

- ◇ confirm the above in stata using our simple data we started today's lecture with
- ◇ run regression of Y on X
- ◇ modify X or Y and check what happened