bivariate regression

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Tuesday 23rd January, 2018 17:18

outline

bivariate regression

stat significance (hypothesis testing)

basic measurement

math

- today we start some math
- important you understand it
- memorizing formulas is not enough to pass this class
- again, ask questions early!
- good idea to go over slides again after the class
- note hats: $\hat{\beta}$ v β
- instead of $\sum_{i=1}^{n}$ i may just use $\sum$

looking ahead

- repeat today's material next week
  - and extend a bit: esp stat significance
  - and talk about measurment time
  - over time, esp after midterm, class will get more applied
  - and we will have more examples

outline

bivariate regression

stat significance (hypothesis testing)

basic measurement

the idea

- $Y \leftarrow\ X$, there is a directional relationship
- like in correlation, but here there is a direction
  - (almost causality, but to argue causality you need also research design!)
- so we have outcome, or dependent variable predicted or affected by:
  - independent variable (does not depend on the dependent variable),

why regression?

- ols is the most fundamental technique for soc sci
  - and quite powerful
  - things like anova, t-test, z-test, chi-sq test, etc are obsolete!
- just run regression! indeed, no studies use these anymore
  - the only thing to remember from qm1 is descriptive stats, esp graphs

- if you want to figure out what predicts something, run regression
  - eg what will make you live longer, or which year wine is good

examples

◇ see some of the useful things you can predict

- eg lexp=weighted avg(diet, exercise, smoking, etc)
- eg lexp=50+2*(veggie serv/day)+3*(hrs at gym)-10*(packs of cigarettes per day)
  life expectancy http://www.northwesternmutual.com/learning-center/the-longevity-game.aspx

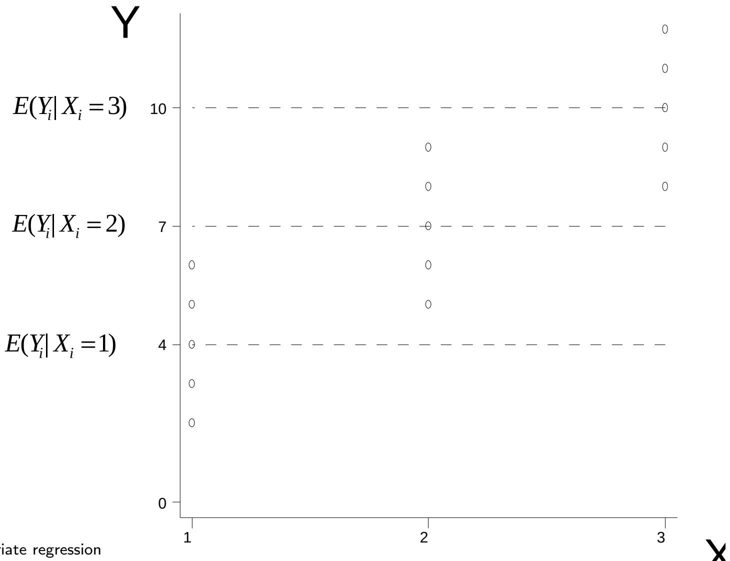
  [*]http://ianayres.yale.edu/prediction-tools

conditional mean

◇ (useful way to start thinking about regression)

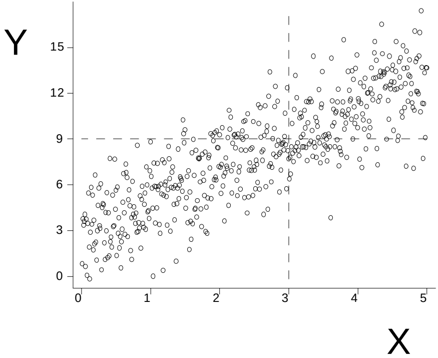
conditional mean of y depends on x

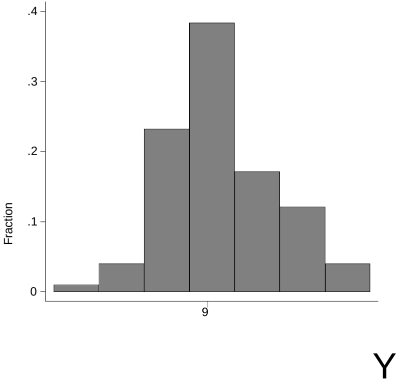
◇ for each value of x(1,2,3) E(y) is different



distribution of Y around the Expected Value

- e.g: $E(Y|X=3)=9$; Values of Y cluster around 9





“regression” sounds scary

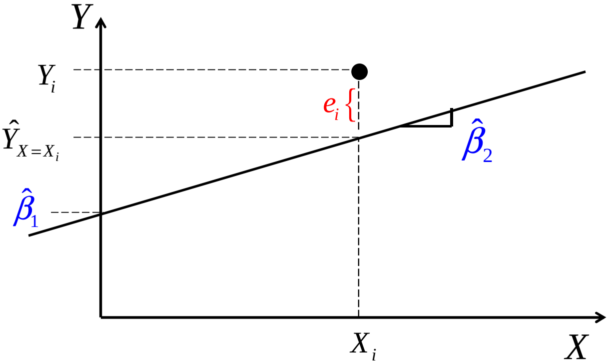
◇ regression is easy (yes, we will do all the tedious calculations), but all that regression does it fits a line that ...
minimizes the sum of the squared vertical distances in a scatter plot; hence “OLS” !

· sounds complicated but it’s easy, too

◇ that’s it ! we will be just showing some math that can fit this line

regression function

- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$



- (e_i) are errors of prediction

what are the disturbance terms?

- $Y_i = \beta_1 + \beta_2 X_i + u_i$
- $u_i = Y_i - \beta_1 - \beta_2 X_i = Y_i - E(Y|X_i)$
- the combined effect of all other variables not in the model
- random events that affect the outcome
- errors of measurement in Y and X

parameters v estimators

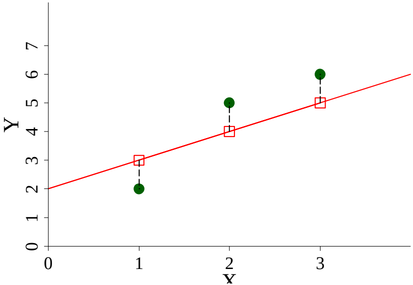
parameters (PRF)	estimators (SRF)
β_1	$\hat{\beta}_1$
β_2	$\hat{\beta}_2$
μ	$\bar{X}$
p	$\hat{p}$
σ	s
μ_i	e_i

- estimators are based on samples
- parameters are fixed (and usually unknown)
- estimators have sampling distributions

first guess

◇

Y_i	X_i
2	1
5	2
6	3

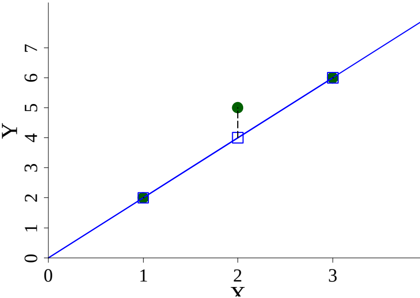


◇ (1) $Y_i = 2 + X_i \rightarrow \sum e_i^2 = 3$

second guess

◇

Y_i	X_i
2	1
5	2
6	3



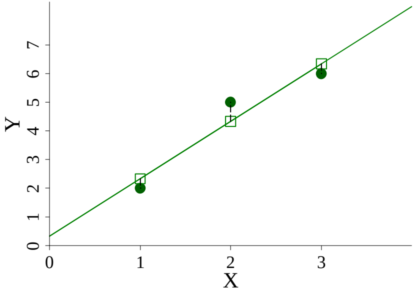
◇ (1) $Y_i = 2 + X_i \rightarrow \sum e_i^2 = 3$

◇ (2) $Y_i = 0 + 2X_i \rightarrow \sum e_i^2 = 1$

example – you cannot beat ols!

◇

Y_i	X_i
2	1
5	2
6	3



◇ (1) $Y_i = 2 + X_i \rightarrow \sum e_i^2 = 3$

◇ (2) $Y_i = 0 + 2X_i \rightarrow \sum e_i^2 = 1$

◇ (3) $Y_i = 0.33 + 2X_i \rightarrow \sum e_i^2 = 0.67$

◇ dofile: guessing; then can use these est to predict like in lexp eg

ols

- $Y_i = \hat{\beta}_1 - \hat{\beta}_2 X_i + e_i \rightarrow e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$
- chose estimators to minimize $\sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$

* for elaboration and derivations see gujarati...

intercept

◇

Intercept: $\boxed{\hat{\beta}_{\mathbf{1}} = \bar{\mathbf{Y}} - \hat{\beta}_{\mathbf{2}}\bar{\mathbf{X}}}$

Note: sum of the residuals is zero: $\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)$

slope

- $\hat{\beta}_2 \quad \boxed{= \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y}}{(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)}}$
- $\hat{\beta}_2 = \frac{\sum Y_i X_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2}$
- $\hat{\beta}_2 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$
- $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} \quad y_i = Y_i - \bar{Y} \;\; x_i = X_i - \bar{X}$
- Another way to look at the slope coefficient is the covariance of Y and X divided by the variance of X. Since the variance is always positive, the numerator (the covariance) will determine the sign of the slope.

solving the problem

	Y_i	X_i	$(Y_i - \bar{Y}) = y_i$	$(X_i - \bar{X}) = x_i$	y_i^2	x_i^2	$y_i x_i$
	2	1	−2.33	−1	5.53	1	2.33
	5	2	0.67	0	0.45	0	0
	6	3	1.67	1	2.79	1	1.67
Σ	13	6	0	0	8.67	2	4
mean	4.33	2					

- $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{4}{2} = 2$
- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 4.33 - (2)(2) = 0.33$

example: age and fear

- In this example, imagine that we have some sort of survey that measures people's fear of crime, and that our hypothesis is that fear of crime increases with age. Assume the fear measure is an index ranging from 0 to 15.
- First, we calculate the means. Second, we calculate the deviations from the means and the their squares for each observation, as well as the co-product of the X and Y deviations. Finally, we sum these up.
- blackboard! all steps!

example: age and fear

The Data

obs	X_i	Y_i
1	22	2
2	35	7
3	47	6
4	56	14
5	72	13
$\sum$	232	42

$$\bar{X} = \frac{232}{5} = 46.4$$

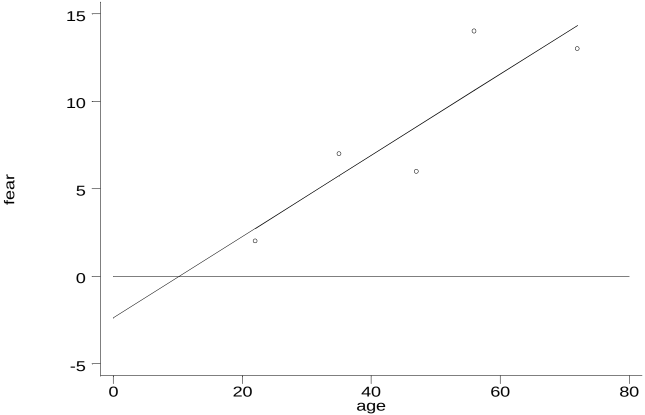
$$\bar{Y} = \frac{42}{5} = 8.4$$

Deviations from the means

Obs	x_i	x_i^2	y_i	y_i^2	$x_i y_i$
1	–24.4	595.36	–6.4	40.96	156.16
2	–11.4	129.96	–1.4	1.96	15.96
3	0.6	0.36	–2.4	5.76	–1.44
4	9.6	92.16	5.6	31.36	53.76
5	25.6	655.36	4.6	21.16	117.76
$\sum$	0	1473.2	0	101.2	342.2

-
- $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{342}{1473} = .232$
- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.4 - (.232)(46.4) = -2.365$
- $\hat{Y}_i = \hat{\beta}_1 + \beta_2 X_i = -2.365 + .232 X_i$
- how would you interpret this?

the estimated regression line



variance and std error of regression

- ok, we know how to calculate betas and fit the line (that min the sum of the squared resid)
- but there are lines that fit better and lines that fit worse in different samples

  draw good and bad fits with same betas
- we need a measure of uncertatinty, i.e. how well our line fit the data...
- and the fit is measured by residuals...
- ... so our measure of uncertainty has to do with residuals !

variance and std error of regression

- $s^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$
- $s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$
  again, the mean of the residuals is zero (hence, ē drops out)
- why divide by n-2?
- s^2 and s are measures of the spread of the points around the estimated regression line.
- they are estimators of the variance and standard deviation of the disturbance terms: σ^2 and σ

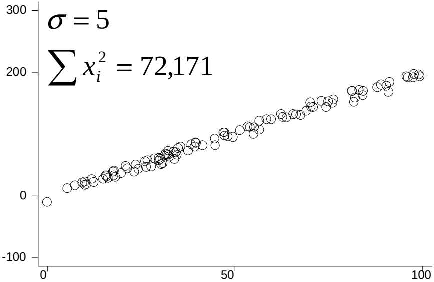
from $\hat{Y}$ to s (se of reg) to $s_{\hat{\beta}_2}$ (se of slope)

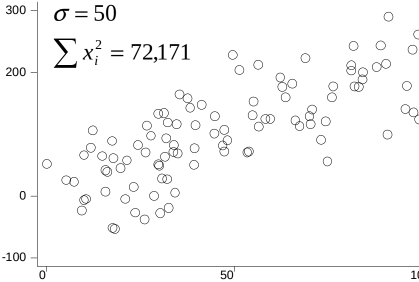
i	$\hat{Y}_i$	e_i	e_i^2
1	2.739	–0.739	0.546
2	5.755	1.245	1.556
3	8.539	–2.539	6.447
4	10.627	3.373	11.377
5	14.339	–1.339	1.793
Σ		0	21.713

- $s = \sqrt{\frac{\sum_{i=1}^{5} e_i^2}{n-2}} = \sqrt{\frac{21.7}{3}} = 2.7$
- what is it measuring?
- $s_{\hat{\beta}_2} = \frac{s}{\sum_{i=1}^{5} x_i^2} = \frac{2.7}{\sqrt{1473}} = .07$
- how does it differ from s?
- calc yhats and se of beta!!
- yahts are important! like our lexp we tried to predict earlier

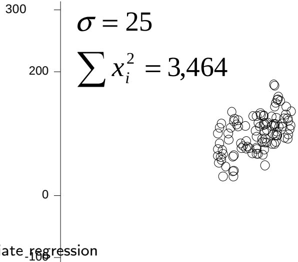
Standard Error of the Slope Coefficient

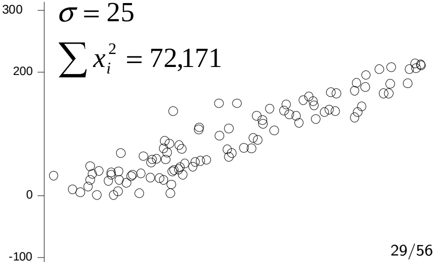
Numerator -- variance of disturbance term





Denominator -- variation in X





ucla: hands-on dofile

- ◇ `https://stats.idre.ucla.edu/stata/webbooks/reg`
- ◇ let's just see a first reg output (you'll do it for ps2)
  - · what is bivariate regression command?
  - · where is β_1 and β_2
- ◇ excellent for self study!!
- ◇ do it at home; and do ask me questions about it if any
- ◇ this is especially an excellent resource for final paper

outline

bivariate regression

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					

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

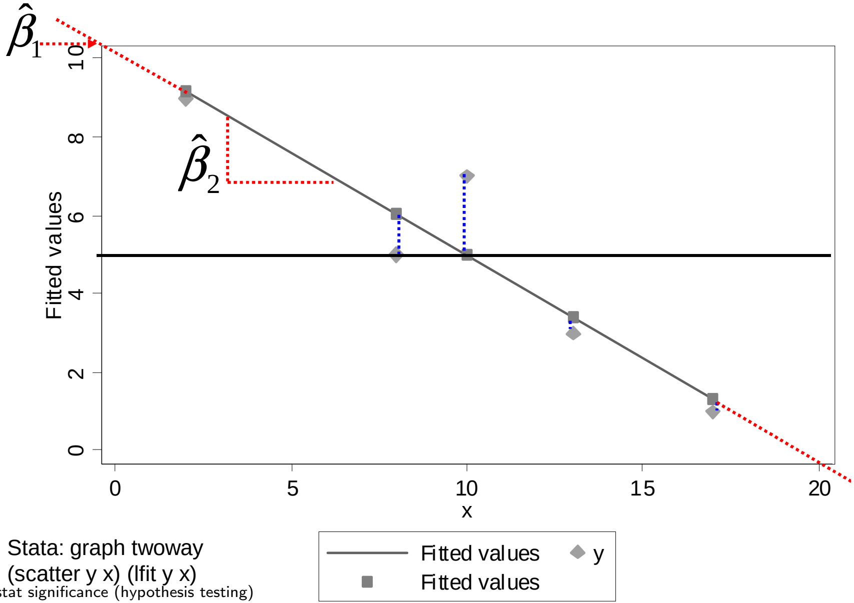
the coefficients–interpretation

- 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



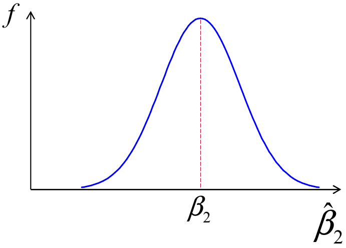
Stata: graph twoway
(scatter y x) (lfit y x)

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: $\hat{\beta}_2 \pm (t_{n-2,\frac{\alpha}{2}})(s_{\hat{\beta}_2})$

accounting for variation in Y blackboard in 3 colors

- ◇ before regression $E[Y] = \bar{Y}$
  - · TSS total sum of squares

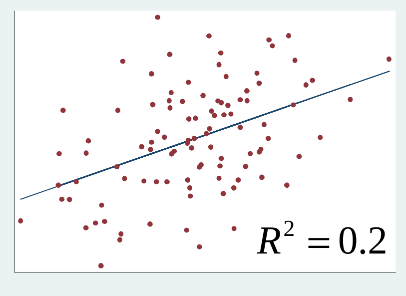
    $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- ◇ after regression

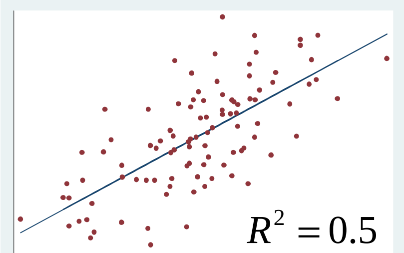
  $E[Y|X_i] = \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
  - · ESS explained sum of squares

    $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
  - · RSS residual sum of squares

    $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$
- ◇ $TSS = ESS + RSS$

R^2 variation explained





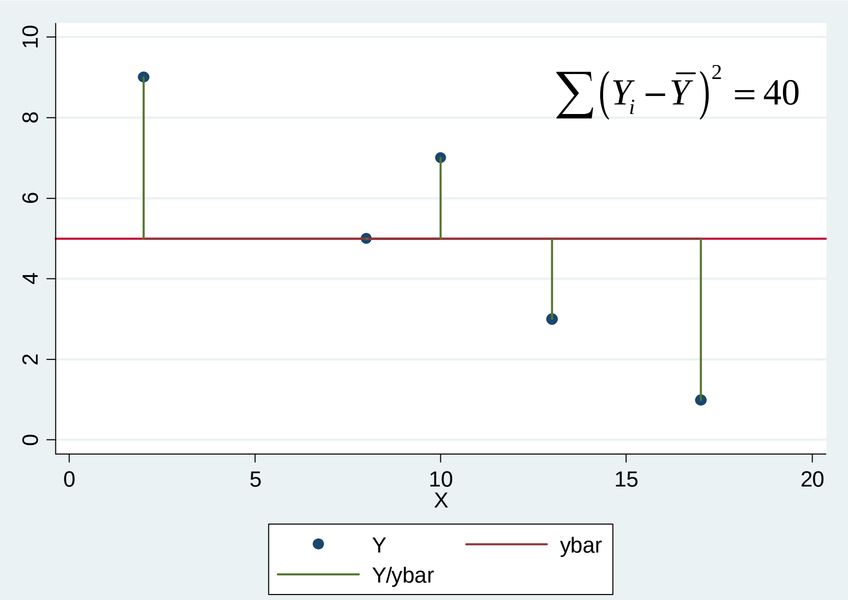
- $TSS = ESS + RSS$
- $1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$
- $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{y_i^2}$
- R^2 : the percent of the variance in the dependent variable explained by the model

partitioning variance in Y dofile

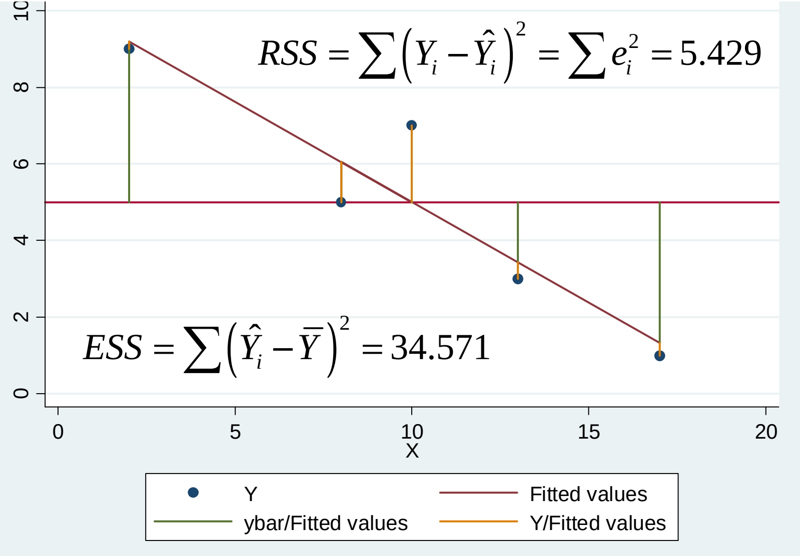
- before regression $E[Y_i] = \bar{Y}$
  - $TSS = \sum(Y_i - \bar{Y})^2 = \sum y_i^2 = 40$
- after regression $E[Y_i|X_i] = \hat{Y}_i$
  - $RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2 = 5.43$
  - $ESS = TSS - RSS = 40 - 5.4 = 34.57$

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

TSS



RSS


$RSS = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum e_i^2 = 5.429$
$ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 = 34.571$
10
8
6
4
2
0
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

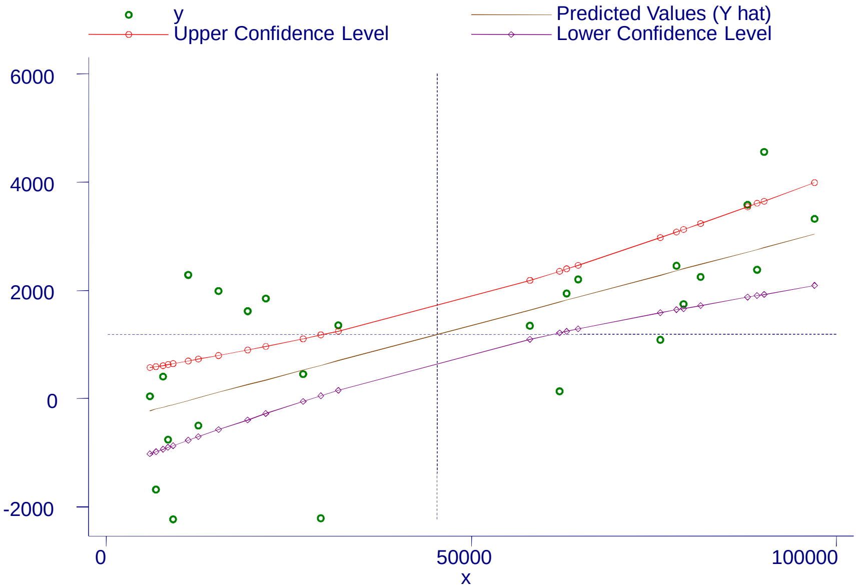
```
------------------------------------
       price |      Coef.   Std. Err.
-------------+----------------------
      weight |   2.044063    .3768341
```

- 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



anatomy of stata output dofile: outlier

```
. regress DV IV
```

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

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 $\beta_1 = 0$	$\hat{\beta}_1 - t_{0.025} s_{\hat{\beta}_1}$	$\hat{\beta}_1 + t_{0.025} s_{\hat{\beta}_1}$

outline

bivariate regression

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

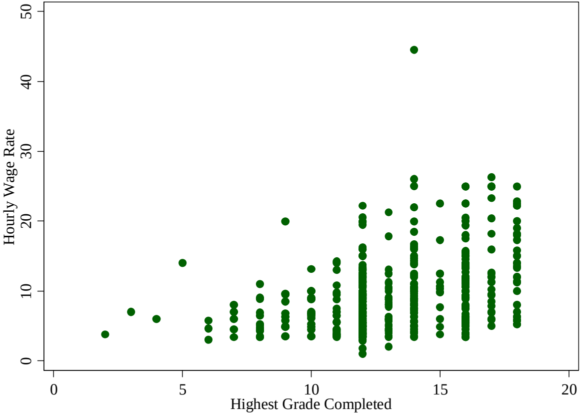
multiply X or Y by constant (say months, not years)

- ◇ 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

- $r = \frac{\sum y_i x_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$ $\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2}$
- 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