

bivariate regression

`adam.okulicz.kozaryn@gmail.com`

this version: Monday 29th January, 2024 17:17

outline

bivariate regression

stat significance (hypothesis testing)

basic measurement

discussions

- everyone saw discussion i posted on canvas?
- post stuff too!

math

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

outline

bivariate regression

stat significance (hypothesis testing)

basic measurement

the idea

- $Y \leftarrow X$, there is a directional relationship, an effect
- like correlation, but here there is a direction
 - (almost causality, but causality need research design!)
- outcome Y , dependent variable (dv)
- predictor X , independent variable (iv)

why regression?

- ols is the most fundamental technique for soc sci
- anova, t-test, z-test, chi-sq test, etc are obsolete!
- just run regression!
- from qm1 only use des sta esp graphs
- want to figure out what predicts something? ols!
- eg what will make you live longer
- $\text{lexp} = f(\text{diet, exercise, smoking, etc})$
- $\text{lexp} = 50 + 2 * (\text{veggie serv/day}) + 3 * (\text{hrs at gym}) - 10 * (\text{packs of cigarettes per day})$

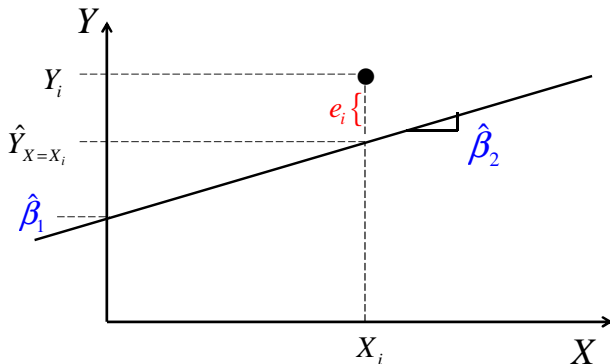
<http://ianayres.yale.edu/prediction-tools>

“regression” sounds scary

- its easy (yes, tedious math), but all all it does it fits a line that
 - minimizes the sum of the squared vertical distances in a scatter plot
 - hence “OLS”
- that's it! we will just use some math to fit this line

regression function

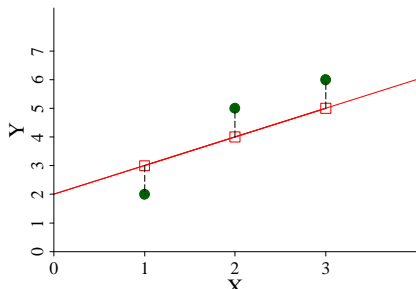
- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$



- (e_i) are errors of prediction

first guess

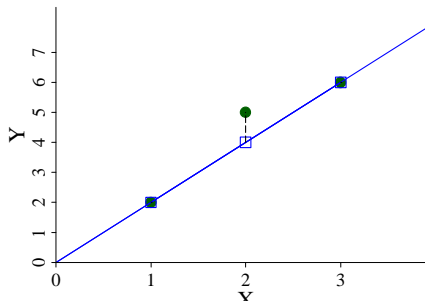
Y_i	X_i
2	1
5	2
6	3



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

second guess

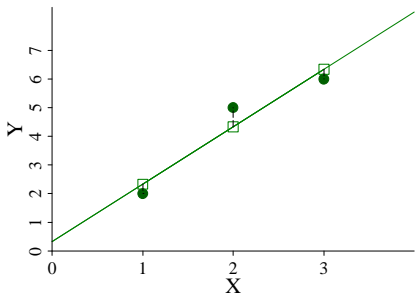
Y_i	X_i
2	1
5	2
6	3



- (1) $Y_i = 2 + X_i \rightarrow \sum e_i^2 = 3$
- (2) $Y_i = 0 + 2X_i \rightarrow \sum e_i^2 = 1$

ols – cannot beat it!

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** can use these est to predict like lexp ex

- $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \rightarrow e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$
- 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 and slope

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n Y_i X_i - n \bar{X} \bar{Y}}{(\sum_{i=1}^n X_i^2 - n \bar{X}^2)}$$

$$\hat{\beta}_2 = \frac{\sum Y_i X_i - n \bar{Y} \bar{X}}{\sum X_i^2 - n \bar{X}^2}$$

$$\hat{\beta}_2 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} \quad y_i = Y_i - \bar{Y} \quad x_i = X_i - \bar{X}$$

slope is the covariance of Y and X divided by the variance of X; variance is always positive, so the numerator (the covariance) determines the sign of the slope

solving the problem [blackboard]

	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
<i>mean</i>	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(18-80) and fear(0-15)

[NOblackboard]

The Data		
obs	X_i	Y_i
1	22	2
2	35	7
3	47	6
4	56	14
5	72	13
Σ	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
Σ	0	1473.2	0	101.2	342.2

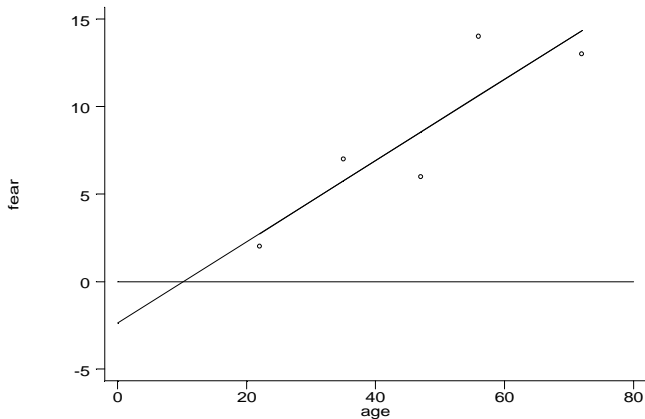
$$\hat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{342}{1473} = .232$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.4 - (.232)(46.4) = -2.365$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = -2.365 + .232 X_i$$

interpretation?

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 some lines fit better and some worse
- need a measure of uncertainty, ie how line fits
- the fit is measured with residuals
- so our measure of uncertainty has to do with residuals!
- $s^2 = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$
- $s = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$
the mean of the residuals is 0 so \bar{e} drops out
- s measures spread of the points around the regression line

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

- $s = \sqrt{\frac{\sum_{i=1}^5 e_i^2}{n-2}} = \sqrt{\frac{21.7}{3}} = 2.7$

- $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum_{i=1}^5 x_i^2}} = \frac{2.7}{\sqrt{1473}} = .07$

-

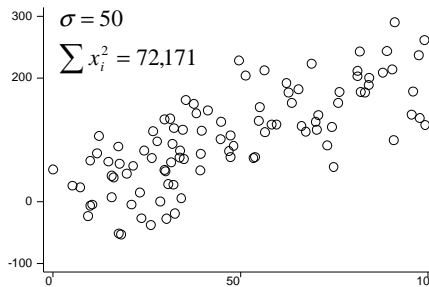
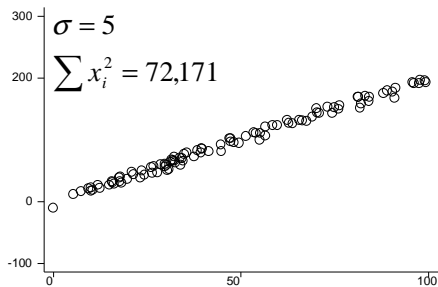
calc yhats and se of beta

- yhats important! like our lexp we predicted earlier

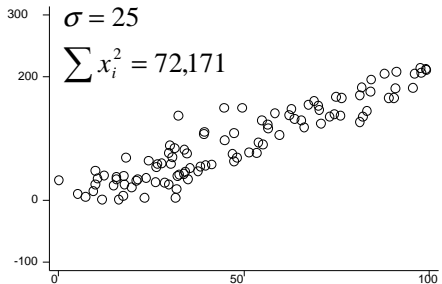
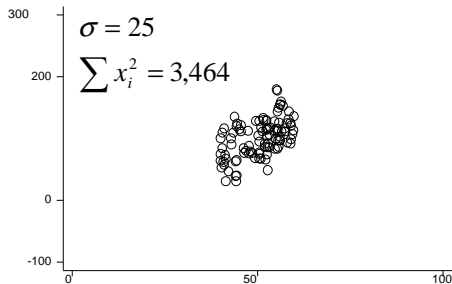
- $t = \frac{\hat{\beta}}{s_{\hat{\beta}}}$

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
- this is especially an excellent resource for final paper

finish first class here

- finish first class here

outline

bivariate regression

stat significance (hypothesis testing)

basic measurement

calculations again **blackboard; dofile**

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

Sum:

25 50

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

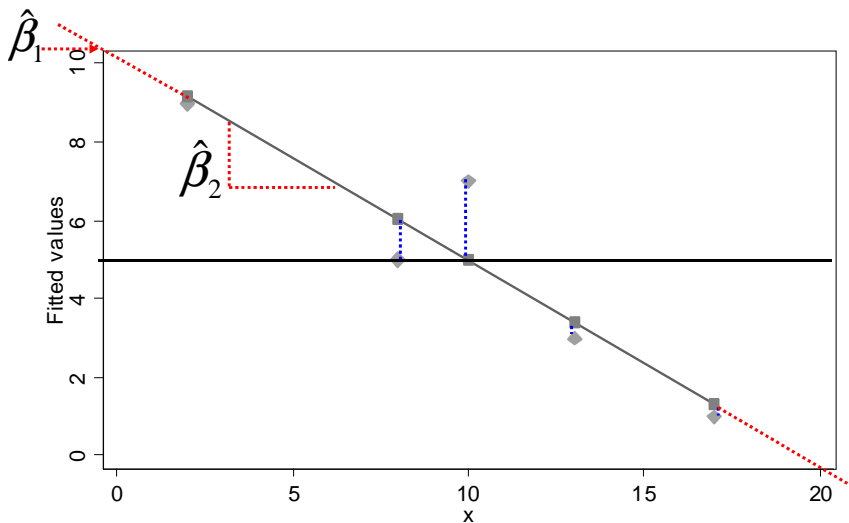
the coefficients—interpretation

- $\hat{\beta}_2$ is the slope coefficient. Thus, a one unit change in X leads to a 0.524 decrease in Y . $\hat{\beta}_1$ is the intercept term. It is the predicted value for Y when X is equal to zero.

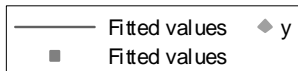
Y	X	Y hat	e	e ²
1	17			
3	13			
5	8			
7	10			
9	2			

- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
- 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)

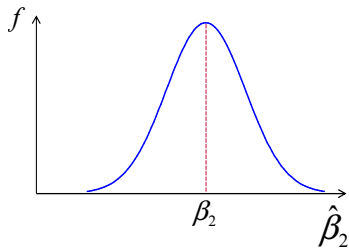


se of the slope **blackboard; dofile**

- $\sum e_i^2 = 5.42$
- $s = \sqrt{\frac{\sum e_i^2}{n-2}} =$
- $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_i^2}}$
- 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:



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

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: $\hat{\beta}_2 \pm (t_{n-2, \frac{\alpha}{2}})(s_{\hat{\beta}_2})$
- lets do it and calculate all by hand! incl crit val

accounting for variation in Y

- before regression $E[Y] = \bar{Y}$
- TSS total sum of squares [like RSS before reg; we were off by this much!]

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

- after regression

$$E[Y|X_i] = \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

- ESS explained sum of squares

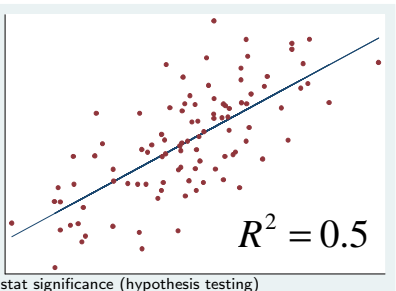
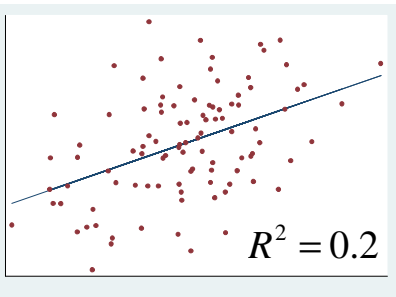
$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

- 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

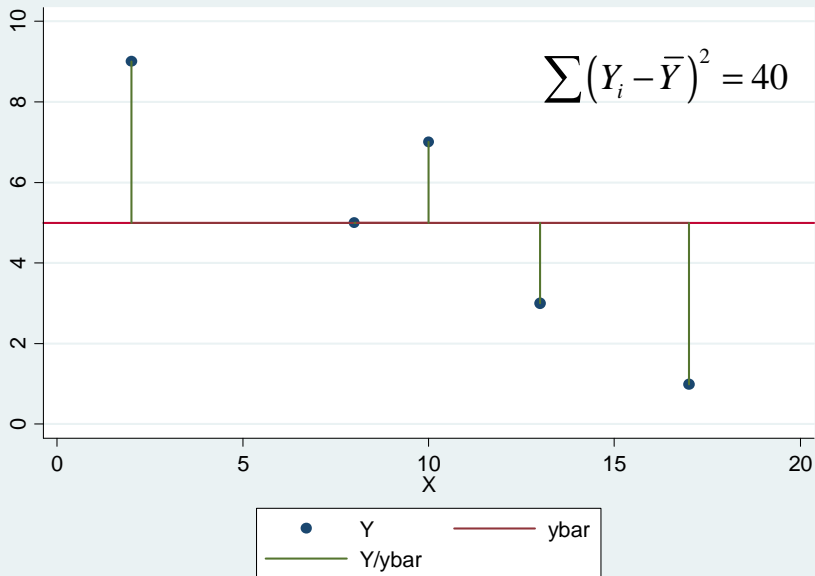


- $TSS = ESS + RSS$
- $1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$
- $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{y_i^2}$
- R^2 : the fraction of the variance in the dependent variable explained by the model

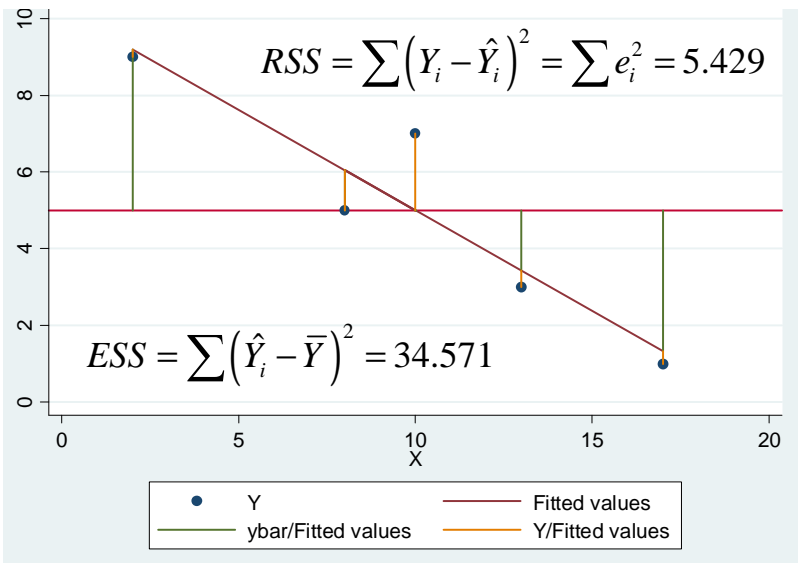
partitioning variance in Y dofile

- 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



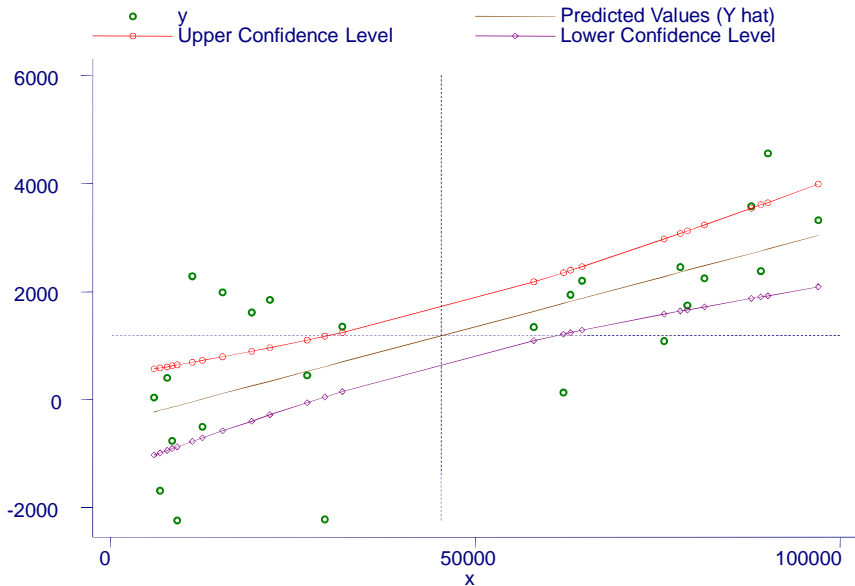
RSS



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

- 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 [biv] **dofile: outlier**

. regress DV IV

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

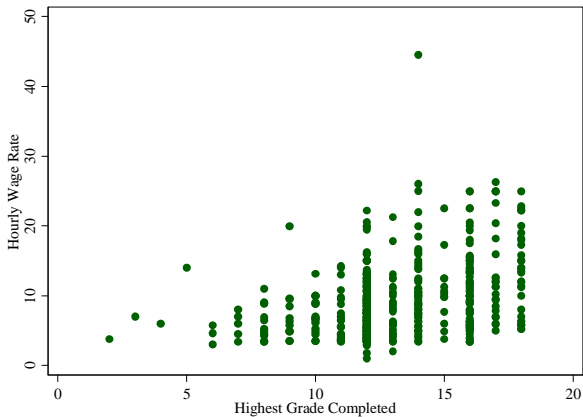
multiply X or Y by constant (say months, not years)

- 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
 - 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		
Model	2053.22494	1	2053.22494	F(1, 532)	=	90.86
Residual	12022.2635	532	22.5982396	Prob > F	=	0.0000
Total	14075.4884	533	26.4080458	R-squared	=	0.1459
				Adj R-squared	=	0.1443
				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:

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

Interpret the coefficients.

fun fact2: Z scores bivariate regression=correlation

- $z_{Yi} = \beta_1 + \beta_2 z_{Xi} + u_i$
 $z_{Xi} = \frac{X_i - \bar{X}}{s_X} = \frac{x_i}{s_x}$
 $z_{Yi} = \frac{Y_i - \bar{Y}}{s_Y} = \frac{y_i}{s_y}$
- z scores always have a mean of 0 and a variance and std dev 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