Class 2, Bivariate Regression

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

(note: for official utd business use ajo021000@utdallas.edu)

this version: Monday 9th February, 2015 18:05

outline

misc

bivariate regression

hands-on: dofile

other interesting properties

outline

misc

bivariate regression

hands-on: dofile

other interesting properties

math

- today we will start doing some math
- it is important that you understand it
- again, memorizing formulas is not enough to pass this class
- again, ask questions
- there is a ps due next week
- it's a good idea to rework math after the class...
- ... there may be a quizz next week

math

- don't worry - we won't be increasing the amount of math anymore
- notation: note hats
- notation: later instead of $\sum_{i=1}^{n}$ i will just use $\sum$

credits

- this class is based on prof. Jargowsky's class

outline

why regression?

- ◇ ols regression is the most fundamental technique for social science – if you are a social scientist you need to know it...
  - (things like anova or t-test or (partial) correlations are just done with regression)
- ◇ regression is useful – if you want to figure out what predicts something you use regression
  - e.g. what will make you live longer, or which year wine is good

examples

- ◇ see some of the useful things you can predict
- · e.g. lexp=weighted avg(diet, exercise, smoking, etc)
- · e.g. 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://islandia.law.yale.edu/ayres/predictionTools.htm

“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

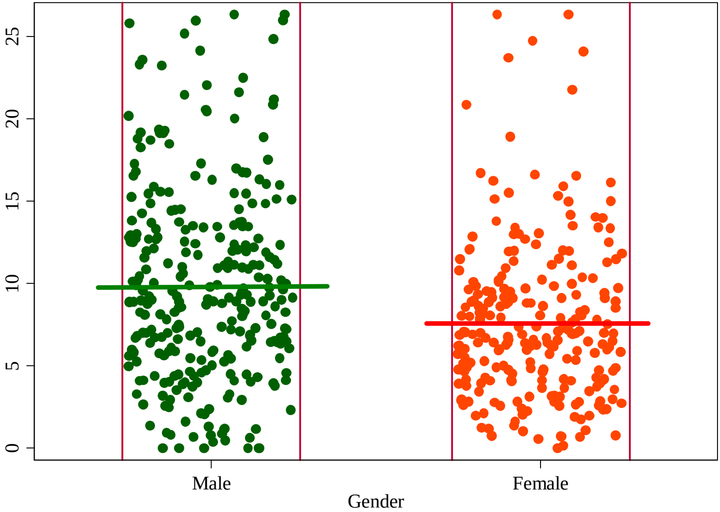
  - sounds complicated but it’s easy, too

  draw a picture

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

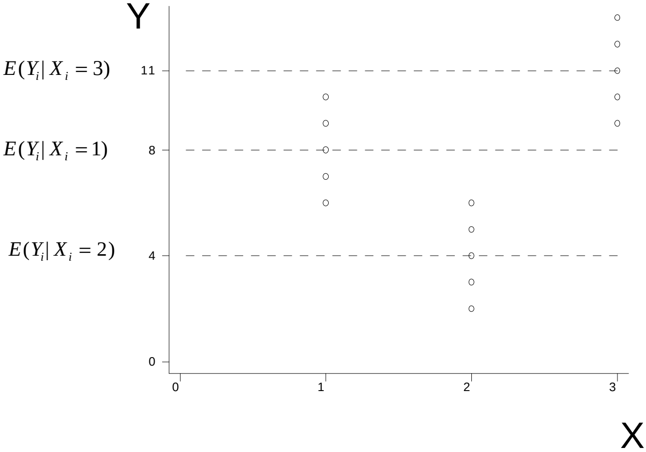
  and stata code that does the math....

conditional mean

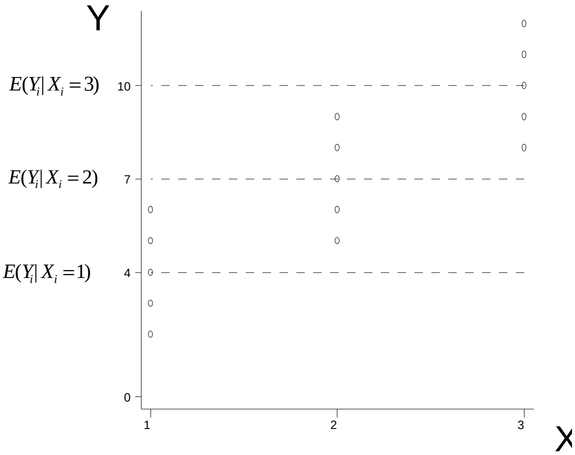


conditional mean of y depends on x

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



conditional mean may be a linear function

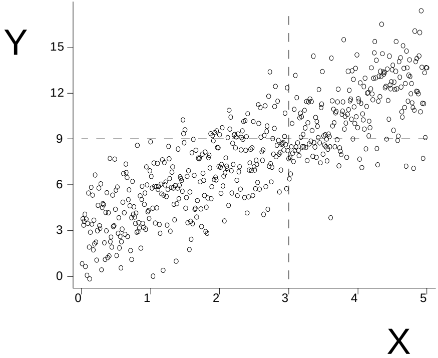


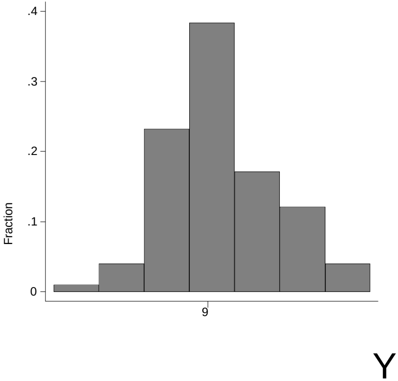
- alternative notations:

$$Y = b + mx = \alpha + \beta x = \beta_0 + \beta_1 x = \beta_1 + \beta_2 x$$

distribution of Y around the Expected Value

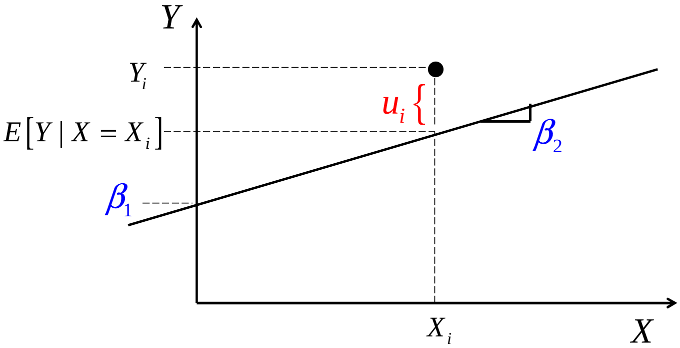
◇ e.g: $E(Y|X=3) = 9$; Values of Y cluster around 9





PRF population regression function

- PRF $Y_i = E(Y|X_i + u_i = \beta_1 + \beta_2 X_i + u_i)$



- the PRF is the true relationship – we don't observe it
- the goal is to obtain the best possible estimate of the PRF

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 all other variables not in the model
- random events that affect the outcome
- errors of measurement in Y and X

the mean of u_i is zero

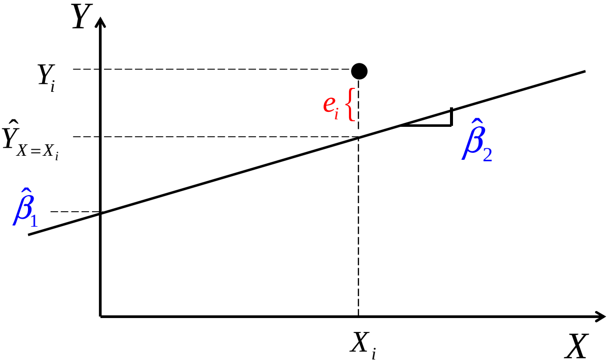
- by assumption, the u_i are the deviations of Y_i from the mean (expectation) of Y given X
- so the expectation of u_i given any particular X_i is zero, because the sum of deviations from a mean is always zero
- $E(u_i) = E[u_i|X_i] = 0$
  (convince yourself – substract mean from every obs and add it up)

the variance of the disturbances

- $var(u_i) = E[(u_i - E[u_i])^2] = E(u_i^2) = \sigma^2$
  (exp val of dist is 0, hence 2nd term drops out)
- if we assume that the variance of the disturbance is constant across all i, then it is a single number
- we can give that number a name or symbol, e.g. Fred or Ω, but the convention is to call it σ^2
- note there is *no* subscript i, because we are assuming (for now) a constant variance
- this is a measure of the degree of random variation in the outcome variable

SRF sample regression function

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



- SRF is an estimate PRF
- (e_i) are errors of prediction

disturbances$\neq$residuals

disturbance	residuals
$u_i = Y_i - \beta_1 - \beta_2 X_i$	$e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$
other influences on Y_i	errors of prediction
unknown	based on estimates $\hat{\beta}_1, \hat{\beta}_2$
unobservable	observable

parameters vs 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

parameters vs estimators

- estimators are based on samples
- parameters are fixed (and usually unknown)
- estimators have sampling distributions
- what are the characteristics of good estimators?
- how can we get good estimators, given a sample?

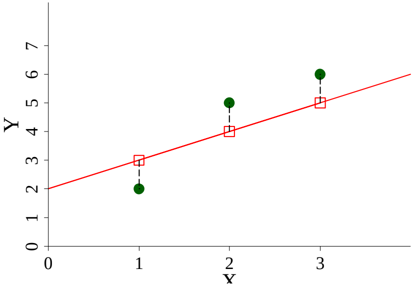
estimation methods

- ◇ guess (not very scientific, prone to bias)
- ◇ minimize the sum of all residuals
  - · doesn't work as expected, because positives and negatives cancel out
- ◇ minimize sum of abs(e) (mad)
  - · ok in theory, (used to be) difficult in practice
- ◇ minimize the sum of squared residuals (ols)
- ◇ maximize the likelihood of the sample (mle)
- ◇ method of moments
- ◇ we will only do ols in this class

first guess

◇

Y_i	X_i
2	1
5	2
6	3

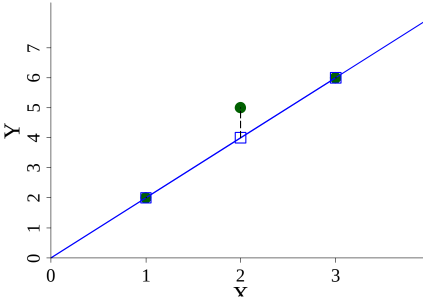


◇ (1) $Y_i = 2 + X_i \rightarrow \sum e_i^2 = 3$

second guess

◇

Y_i	X_i
2	1
5	2
6	3



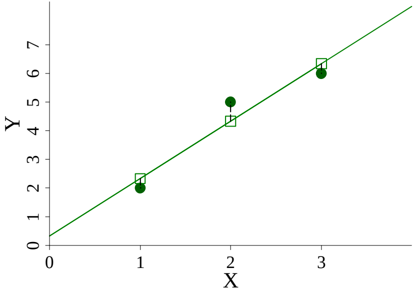
◇ (1) $Y_i = 2 + X_i \rightarrow \sum e_i^2 = 3$

◇ (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

ols

- SRF: $Y_i = \hat{\beta}_1 - \hat{\beta}_2 X_i + e_i \rightarrow e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$
- 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: $\boldsymbol{\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}}$

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

slope

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

the solution

- These formulas produce the estimates of the slope and intercept that minimize the sum of the squared residuals, given the sample. They can be easily calculated from the sample data, without guessing or searching for an answer. The next few slides show the algebra, but the formulas above are the bottom line

some intuitive algebra

$$\begin{aligned}\sum(Y_i-\bar{Y})(X_i-\bar{X}) &= \sum(Y_iX_i - Y_i\bar{X} - \bar{Y}X_i + \bar{Y}\bar{X})\\ &= \sum Y_iX_i - \bar{X}\sum Y_i - \bar{Y}\sum X_i + n\bar{Y}\bar{X}\\ &= \sum Y_iX_i - n\bar{Y}\bar{X} - n\bar{Y}\bar{X} + n\bar{Y}\bar{X}\\ &= \sum Y_iX_i - n\bar{Y}\bar{X}\end{aligned}$$

◇

$$\begin{aligned}\sum(X_i-\bar{X})^2 &= \sum(X_i^2 - 2\bar{X}X_i + \bar{X}^2)\\ &= \sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\\ &= \sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\\ &= \sum X_i^2 - n\bar{X}^2\end{aligned}$$

alternative expressions for the slope

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

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

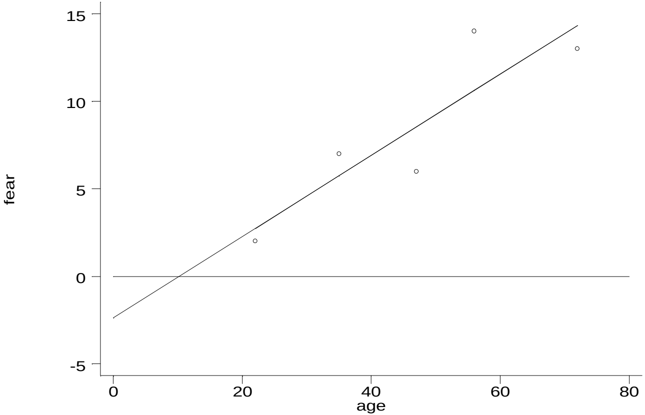
$$\bar{Y} = \frac{42}{5} = 8.4$$

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

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

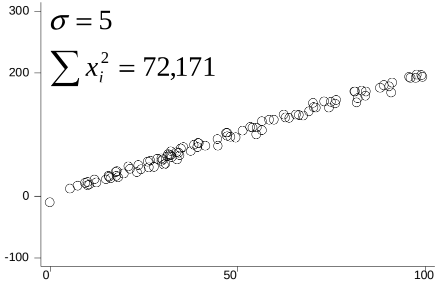
- $s^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$
- $s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$
  again, the mean of the residuals is zero (hence, $\bar{e}$ drops out)
- why divide by n-2?
- s^2 and s are measures of the spread of the points around the estimated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

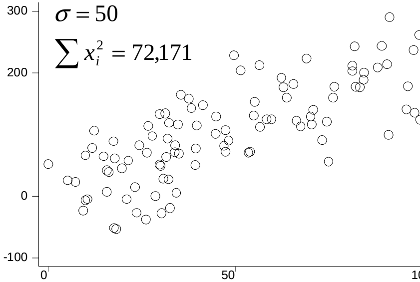
how good are ols estimators?

- ◇ Are they unbiased? (And just what does unbiased mean?)
- ◇ How reliable are they, i.e. how much do they vary from sample to sample?
- * for elaboration and derivations see gujarati...

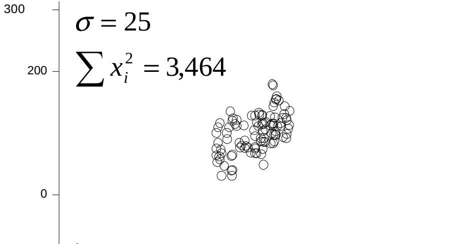
Standard Error of the Slope Coefficient

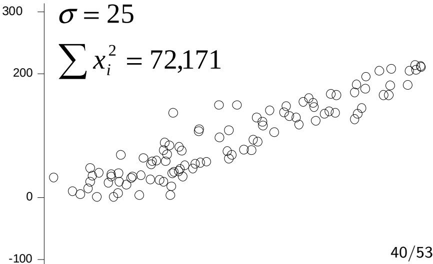
Numerator -- variance of disturbance term





Denominator -- variation in X





from predicted values to std err

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!!

key ols assumptions

- ◇ $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$ Choose $\hat{\beta}_1, \hat{\beta}_2$ to minimize $\sum e_i^2$
- ◇ the true model is linear $Y_i = \beta_1 + \beta_2 X_i + u_i$
- ◇ The true model has a stochastic disturbance term, u, with the following properties:
  - $E[u_i = 0]$ expected value of u is zero
  - $cov[X_i u_i] = 0$ X and u are not correlated
  - $var[u_i] = \sigma^2$ constant variance
  - $cov[u_i u_j] = 0$ for all $i \neq j$
- ◇ if true, then BLUE: Best Linear Unbiased Estimators
- ◇ there will be more assumptions later

outline

misc

bivariate regression

hands-on: dofile

other interesting properties

ucla

- just see `http://www.ats.ucla.edu/stat/stata/webbooks/reg/`
- 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

misc

bivariate regression

hands-on: dofile

other interesting properties

assumptions about the model

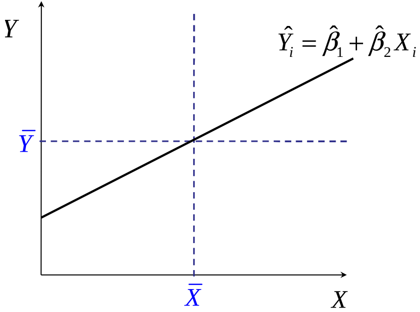
- We assume a model that is linear in the parameters and has an additive disturbance term:
- $Y_i = \beta_1 + \beta_2 X_i + u_i$
- Linear in the parameters means that the betas have a power of one and only one beta appears in each term. The following are not linear models:
- $Y_i = \frac{X_i}{\beta_2} + u_i$
- $Y_i = (\beta_1 + \beta_2 X_i) u_i$
- but this one is linear in parameters:
- $Y_i = \beta_1 + \beta_2 X_i^9 + u_i$

assumptions about X

- X varies, i.e. $\sum x_i^2 > 0$
- if X does not vary, the slope is undefined
- $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$
- also, remember from rd "constant cannot explain variance"

fitted line goes through $\bar{X}$ and $\bar{Y}$

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



mean of predictions

$$
\begin{aligned}
\bar{\hat{Y}}_i &= \\
&= \frac{\sum_{i=1}^{n} \hat{Y}_i}{n} \\
&= \frac{\sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 X_i)}{n} \\
&= \frac{n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i}{n} \\
&= \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \\
&= \bar{Y}
\end{aligned}
$$

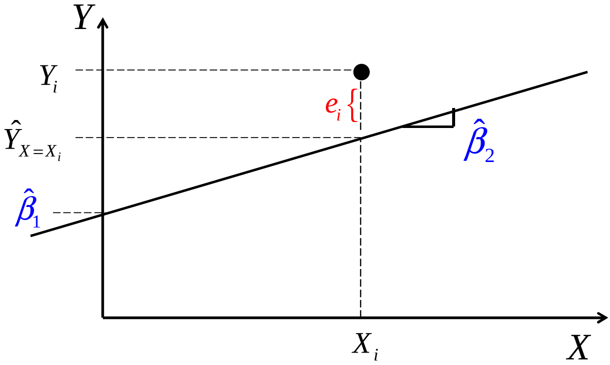
$\diamond$

mean of residuals

$$
\begin{aligned}
\bar{e}_i &= \\
&= \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n} \\
&= \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)}{n} \\
&= 0
\end{aligned}
$$

recall SRF

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

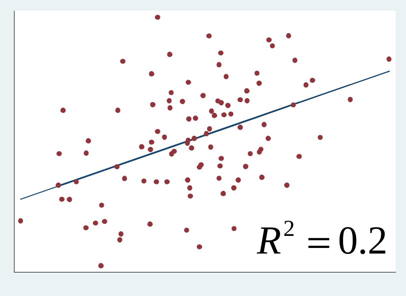


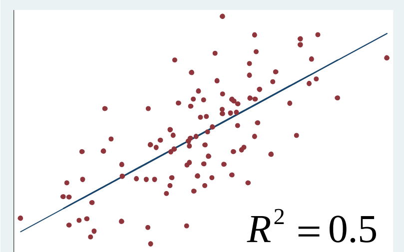
◇ SRF is generated from estimates of the PRF's parameters. Every SRF has residuals (e_i), i.e. errors of prediction.

accounting for variation in Y

- 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