Lecture 11 (Ch. 3)

We did regression (fitting) by assuming a model for data (x_i, y_i):

$$\underset{\text{obs. } y \text{ at } x_i}{y_i} = \underbrace{\alpha + \beta x_i}_{y \text{ of line } y(x) = \alpha + \beta x \text{ at } x = x_i} + \epsilon_i \leftarrow \text{error/residual.}$$

To find the "best" α, β (i.e. line), we minimized SSE:

$$SSE = \sum_i^n \epsilon_i^2 = \sum_{i=1}^n \Big[\underset{\text{obs}}{y_i} - \underbrace{(\alpha + \beta x_i)}_{\text{pred.}}\Big]^2$$

Compare!

Compare!

and got $\hat\beta = \dfrac{\overline{xy} - \bar x \bar y}{\overline{x^2} - \bar x^2}$, $\hat\alpha = \bar y - \hat\beta \bar x$.

Then, the eqn of the "best fit" is $\hat y(x) = \hat\alpha + \hat\beta x$

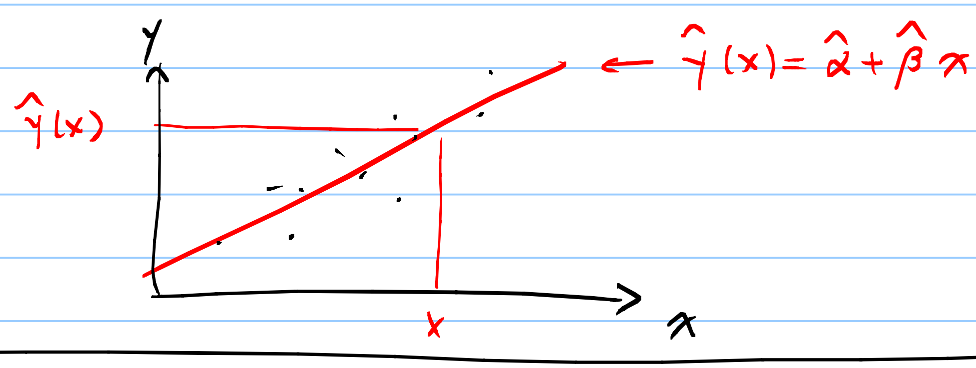
Note: $\hat y(x_i) = \hat\alpha + \hat\beta x_i$ (no ϵ!)

$\hat y(x_i)$ is sometimes written as $\hat y_i$.

I forgot to mention that regression is most useful when x is easy to measure, but y is hard to measure.

E.g. x = Blood flow velocity (FV) with ultra sound.
y = Intracranial Pressure (ICP).

When $\hat\alpha, \hat\beta$ are obtained from regression, then, given $\underset{\text{FV}}{x}$, we can predict $\underset{\text{ICP}}{y}$ from $\hat y(x) \equiv \hat\alpha + \hat\beta x$. (No ϵ_i!)



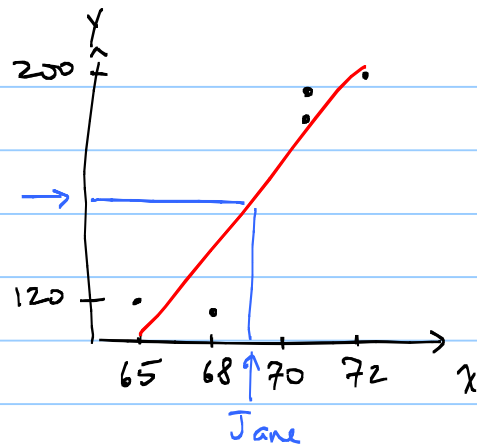
Also $\hat\beta = \dfrac{S_{xy}}{S_{xx}}$ where $S_{xx} \equiv \sum_i (x_i - \bar x)^2$ $S_{xy} \equiv \sum_i (x_i - \bar x)(y_i - \bar y)$

Example

or FV and ICP

Data

height (x)	weight (y)	xy	x^2
72	200		
Joe: 70	180		
65	120		
68	118		
70	190		
$\bar{x}$	$\bar{y}$	$\overline{xy}$	$\overline{x^2}$



$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{11224.8 - 69(161.6)}{4766.6 - 69(69)} = 13.28$$

Interpret: A change of 1 in. is associated with an avg. change of 13.28 pounds.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 161.6 - 13.28(69) = -755$$

lm(y ~ x) $\Rightarrow$ $\hat{\beta} = 13.3$, $\hat{\alpha} = -755.11$ $\Rightarrow$ $\hat{y}(x) = -755 + 13.28x$

$\Rightarrow$ E.g. Joe's predicted weight, according to his height, is

$\hat{y} = 13.28(70'') - 755.11 \approx 174.9$ pounds.

$\Rightarrow$ We can now predict everyone's weight (or ICP), from their height (or FV):

Height (x)	Weight (y)	...	$\hat{y}$	$(y - \hat{y})$
72	200		201.5	−1.5
Joe = 70	180		174.9	5.1
65	120		108.5	11.5
68	118		148.3	−30.3
70	190		174.9	15.1

$\hat{y} = \hat{\alpha} + \hat{\beta}x$ predicted y

any other fit will have a larger SSE.

$\Rightarrow$ For the people in the data set, we can also find their $\boxed{\text{error/residual}}$

$\Rightarrow$ For people outside the data set (eg. Jane) we can predict their y from their x, but we cannot compute error, because we don't know their true y. In Ch.11, we'll address this issue.

However, be WARNED if you extrapolate

$x = 0 \Rightarrow y = -755$ pounds!

Shifting gears again.

There is a different (more useful) way of looking at regression, via Variance. This way, we will arrive at quantities called R^2 an s_e. which together assess how good the fit is.

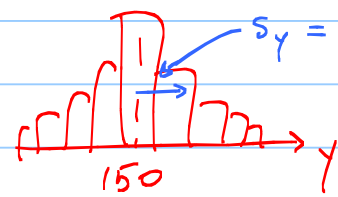
Let me motivate it:

→ Suppose we measure my Tablet's Length.

→ Repeat, and histogram:

→ One may report:

True length = 150 ± 10 cm



$s_y = \sqrt{\frac{S_{yy}}{n-1}} \sim 10$

→ Now, suppose you are unhappy with the large s_y. low precision.

→ You may wonder, could some of that variability be due to something else that is varying everytime you

→ make a measurement of y. x = temperature? humidity?

If so, then by measuring y and x, we may be able to reduce the ± of our report, by specifying y at a given x.

Analysis of variance (ANOVA) approach to regression:

Q How much of the variation in y is due to the (linear) relationship between y and x?

Table length ... temperature.

A variance of $y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

OLS

$+\hat{y}_i - \hat{y}_i$, $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

OLS

E for Error, not for Explained

$$\underbrace{S_{yy} = \sum_i (y_i - \bar{y})^2}_{SS_{total}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SS_{explained}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SS_{unexplained}}$$

total variation in y.

variation in y explained by (or due to) x

variation in y unexplained by x

$$\underset{\sim (10)^2}{SST} = SS_{explained} + \underset{\sim (3)^2}{SSE}$$

Variability is reduced from $\pm(10)^2$ to something smaller, say $\pm(3)^2$.

Therefore, $\underbrace{\frac{SS_{expl}}{SST} \times 100}$, called R^2, measures how good the fit is.

percent variation in y, explained by x.

(Bad Model/fit) $0 < R^2 < 1$ (Good Model/fit)

The other piece, $SS_{unexpl.} = SSE$, is a sum (of squares), and so can be can be "Averaged" to provide a measure of typical error

$$\sqrt{\frac{SSE}{n-2}} = \sqrt{\underbrace{\frac{1}{n-2}\sum_{i}^{n}}_{\text{"funny Avg."}} \underbrace{(y_i - \hat{y}_i)}_{\text{error}}^2} = S_e \sim \text{std. dev. of errors}$$

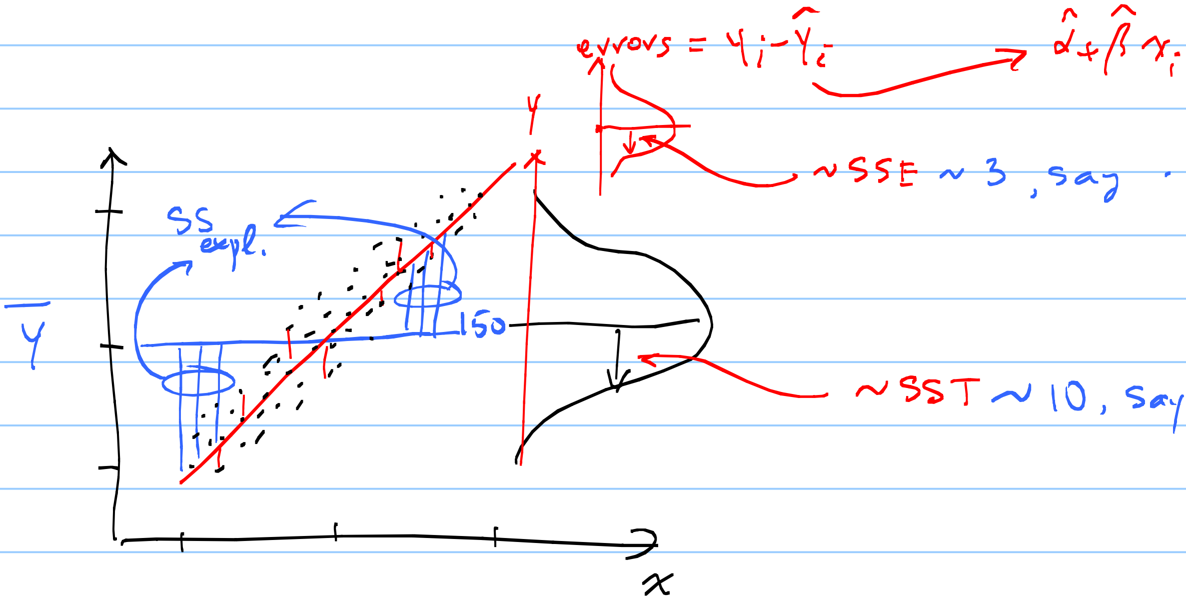
$\sim$ typical error.

Compare with $S_y^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$

$$\boxed{\text{Report } \hat{y}(x) \pm S_e}$$

FYI

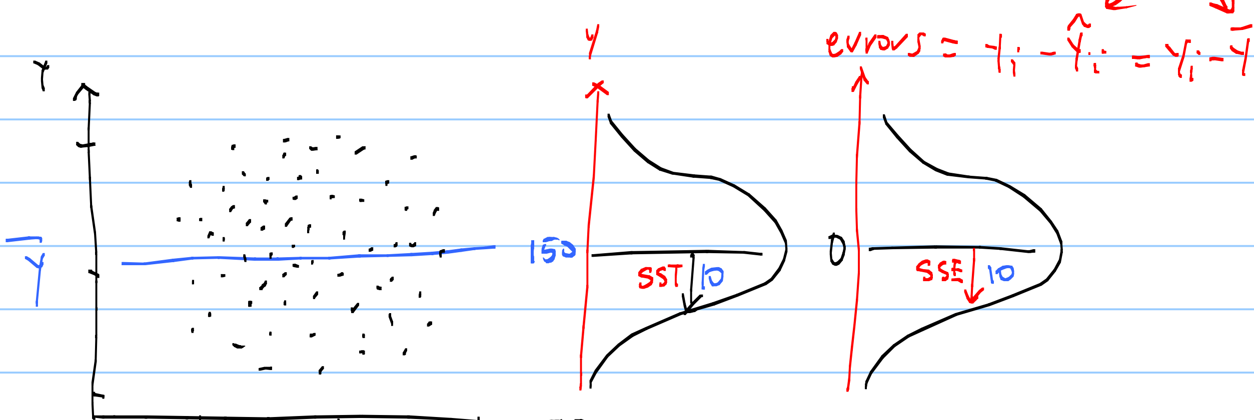
Picture for the ANOVA decomposition:



So, when there is a (linear) relationship between x & y, then some portion of the variation in y can be attributed to (or explained by) x. That portion is $SS_{exp.}$, and the (unexplained) rest is $SS_{unexp} = SSE$.

So the variability in y, SST $(10)^2$, is reduced to SSE $(3)^2$.

When there is no relationship between x and y, then the fig looks like below. Note that this situation is equivalent to the situation where we have data only on y, and not on x at all. In that case the best prediction for every case is $\bar{y}$ (see hw):



In this case there is no reduction in SST at all, as expected.

Example (same as in last few lectures):

$SST = \sum_i (y_i - \bar{y})^2 = \cdots = 6251.2$

$SSE = \sum_i (y_i - \hat{y}_i)^2 =$ last column in table in prev. lecture.

$= (-1.5)^2 + (5.1)^2 + (11.5)^2 + (-30.3)^2 + (15.1)^2 = 1307$

$\Rightarrow R^2 =$ Coef. of det. $= \dfrac{SST - SSE}{SST} = \dfrac{6251.2 - 1307}{6251.2} = \boxed{0.79.}$

Conclusion (Meaning): 79% of the variability (or variation) in y (weight, or Tablet length) is due to (can be explained by) the linear relation with x (height, or temperature).

The other piece of the decomposition:

$\Rightarrow S_e = \sqrt{\dfrac{1307}{5-2}} = 20.9$ pounds

Conclusion (Meaning): The typical deviation of the y values (weight / Tablet length) (ie. error or residual) about the fit is about 21 pounds.

Report weight (or Tablet length) or ICP or ... : $\hat{y} \pm 20.9$ with $R^2 = 0.79$

$\hat{y} = -755 + 13.3 x$ ← height or FV or ...

$\boxed{Q1:}$ In the prev. clicker qz we found that if $y(x) = \beta$, then the OLS estimate of β is $\hat{\beta} = \bar{y}$. Then S_e is (proportional to)

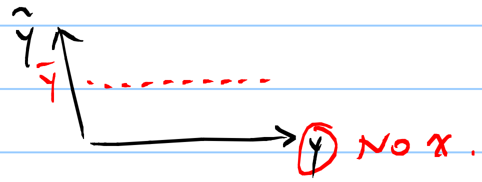
A) 0 B) S_x $\boxed{C)\ S_y}$ D) None of the above.

$S_e = \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{n-2}} = \sqrt{\dfrac{n-2}{n-1}}\, S_y \Rightarrow \hat{y}(x) \pm S_e = \bar{y} \pm S_y$ Make sense?

BTW: for this $\hat{y}(x) = \hat{\beta}$ example, $R^2 = 1 - \dfrac{SSE}{SST} = 1 - \dfrac{SST}{SST} = 0$

FYI

We learned from the last 2 clicker questions, that if there is no x data, then the OLS prediction $\hat{y}$ is just $\bar{y}$.

Ie. 

What are the R^2 and S_e ?

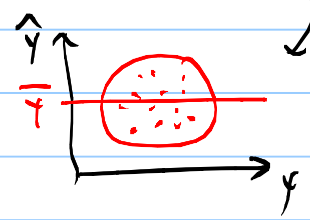
$\hat{y}_i = \bar{y}$ defn. of s_y^2.

$$S_e^2 = \frac{SSE}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i (y_i - \bar{y})^2}{n-2} = \left(\frac{n-1}{n-2}\right) S_y^2 \Rightarrow S_e \sim S_y$$

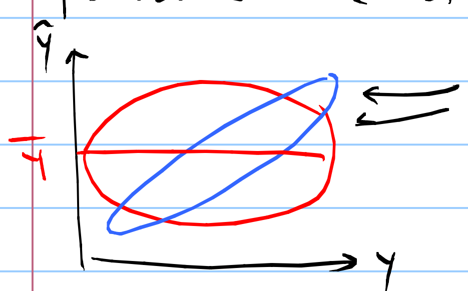
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 0 \Rightarrow R^2 = 0.$$

∴ So, if we use $\bar{y}$ as our prediction, then $R^2 = 0$ (Bad), and $S_e \sim S_y$, ie. the typical error ~ typical dev. in y, ie. nothing gained.

Another situation when nothing is gained is if we make random predictions, e.g. $\hat{y}_i$ = random. Suppose the mean and the var. of these random predictions are the same as those of observations, ie. $\hat{y}_i$ = random with $\bar{\hat{y}} = \bar{y}$, $S_{\hat{y}} = S_y$. The picture is



But now, something strange happens:
Although one can use the formula for R^2 to arrive at a number, that number does not have the usual interpretation (ie. percentage of var. in y, explained by x), because $\hat{y}_i$ = random are not OLS predictions. So, we don't have the ANOVA decomposition at all. Same objection applies to se.
Again, the ANOVA decomposition is correct only for OLS $\hat{y}$; $\hat{y}$ = random are not OLS predictions.



These both have equal/comparable $S_{\hat{y}}$ (ie. R^2).
But the blue one has lower S_e.
This doesn't contradict ANOVA, because the red $\hat{y}$ is not OLS.

In short, both have equal precision, but blue is more accurate hw-lect11-1

For the data shown here:
x= 45, 58, 71, 71, 85, 98, 108
y = 3.20, 3.40, 3.47, 3.55, 3.60, 3.70, 3.80
a) Compute the eq. of the OLS fit.
b) Compute the total variation, SST.
c) Decompose it into explained and unexplained.
d) Compute R2 and interpret it (in English),
e) Compute the std. dev of errors, s_e, and interpret it (in English).

All by hand. You may use R to compute sums, means, std. deviations, but not a function that does regression or analysis of variance.

hw-lect11-2 Consider The following decomposition:

By R

$$\sum_i (y_i-\bar{y})^2 = \sum_i \left[(\hat{y}_i-\bar{y}) + (y_i-\hat{y}_i)\right]^2$$

$$= \sum_i (\hat{y}_i-\bar{y})^2 + \sum_i (y_i-\hat{y}_i)^2 + 2\underbrace{\sum_i (\hat{y}_i-\bar{y})(y_i-\hat{y}_i)}$$

In past hws I have asked students to prove That The last term is zero if $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, with $\hat{\alpha}, \hat{\beta}$ being The OLS estimates (ie. $\hat{\alpha}, \hat{\beta}$ given in lects, book). Unfortunately, it's a long calculation; so This time we'll try to show That it's zero using simulation in R. Write code to

a) generate a sample of size 100 from The unif dist. between -1 and +1. call it x.

b) generate y such That $y_i = 2 + 3x_i + \epsilon_i$ with ϵ_i having a normal distr. with $\mu = 0$, $\sigma = 0.5$.

c) Do regression on x, y, and call The predictions $\hat{y}$.

d) compute $\sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$. It should be (very) zero!

hw-lect11-3

$SS_{exp.}$ can be computed from its defining relation $\left(\sum_i (\hat{y}_i - \bar{y})^2\right)$ or from $(SST - SSE)$, or from $\hat{\beta}$ and S_{xx}, as follows.

Explain what has happened at every step.

$$SS_{exp} = \sum_i (\hat{y}_i - \bar{y})^2$$

?

$$= \sum_i (\hat{\alpha} + \hat{\beta} x_i - \bar{y})^2$$

?

$$= \sum_i (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i - \bar{y})^2$$

?

$$= \sum_i (\hat{\beta})^2 (x_i - \bar{x})^2$$

?

$$= (\hat{\beta})^2 \sum_i (x_i - \bar{x})^2$$

?

$$= (\hat{\beta})^2 S_{xx}$$

If you would like, print out this page, and write your answers in the space here.