Lecture 12 (Ch. 3)

Last time we did Ordinary Least Squares (OLS) regression, also known as curve/line fitting. It's a dangerously intuitive method. But at the 390 level, you are expected to know about a level of detail that is devilishly complicated.

I.e. The devil is in the details.

Look!

In regression, we assume that data (x_i, y_i) follow a model:

Model: $y_i = \alpha + \beta x_i + \epsilon_i$ ← error/residual.

obs. y at x_i ; y of line $y(x) = \alpha + \beta x$ at $x = x_i$

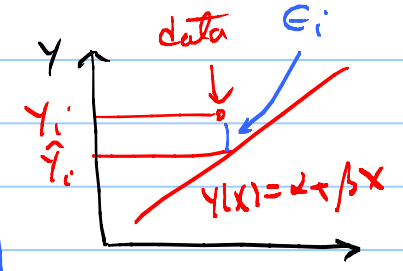
To find the "best" line (i.e. OLS α, β), we minimized SSE:

$$SSE = \sum_i^n \epsilon_i^2 = \sum_{i=1}^n \left[y_i - (\alpha + \beta x_i) \right]^2$$

obs ; $\hat{y}_i$ = pred

Compare!

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



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

$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$

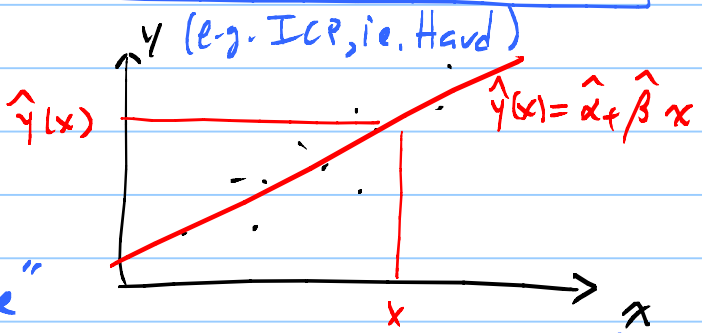
Compare!

When $\hat{\alpha}, \hat{\beta}$ are obtained from regression, Then, given x, we can predict y from the eqn. of the "best" (i.e. OLS) fit:

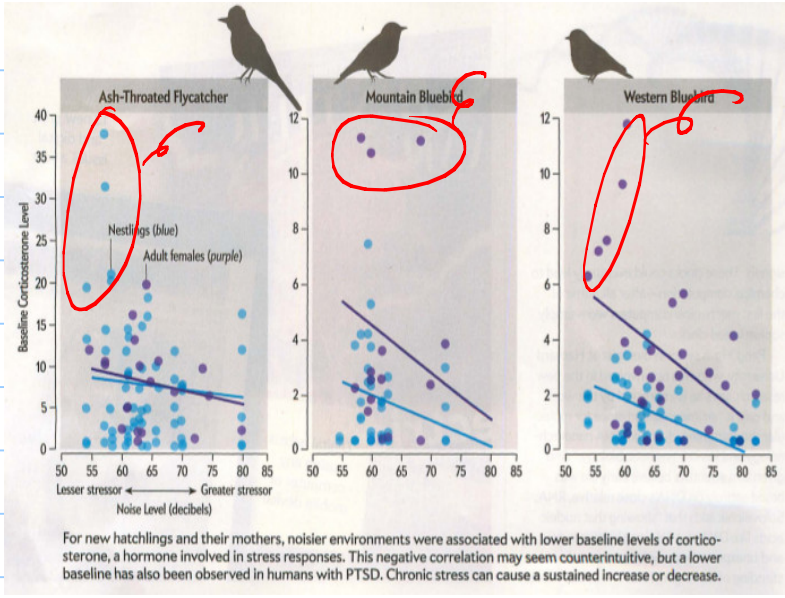
$\hat{y}(x) \equiv \hat{\alpha} + \hat{\beta} x$.

$\hat{y}(x_i) \equiv \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ (No ϵ_i)

The "prediction", or "predicted value" or "fitted value", ...



Examples of BAD regression.

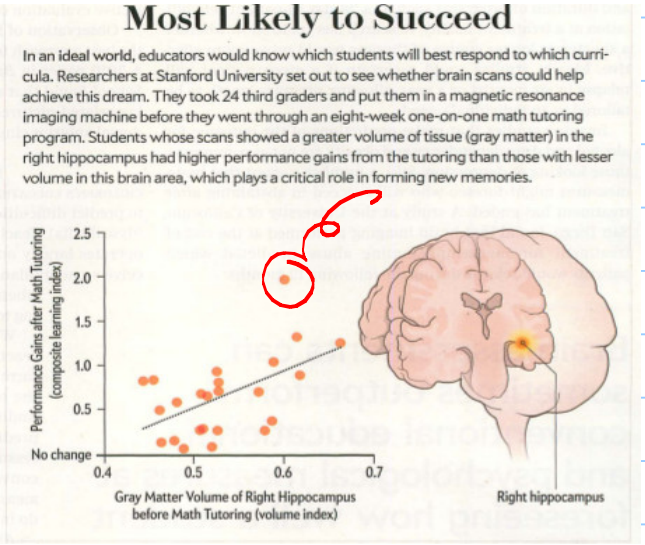


For new hatchlings and their mothers, noisier environments were associated with lower baseline levels of corticosterone, a hormone involved in stress responses. This negative correlation may seem counterintuitive, but a lower baseline has also been observed in humans with PTSD. Chronic stress can cause a sustained increase or decrease.

Scientific Amer. April 2018

Authors claim that levels of some hormone in the brain (y-axis) fall as the distance between bird nests and sources of sound increase.

Ignore the regression line; would you agree with the conclusion, based on the scatterplot? What if we remove the 2 or 3 circled birds?

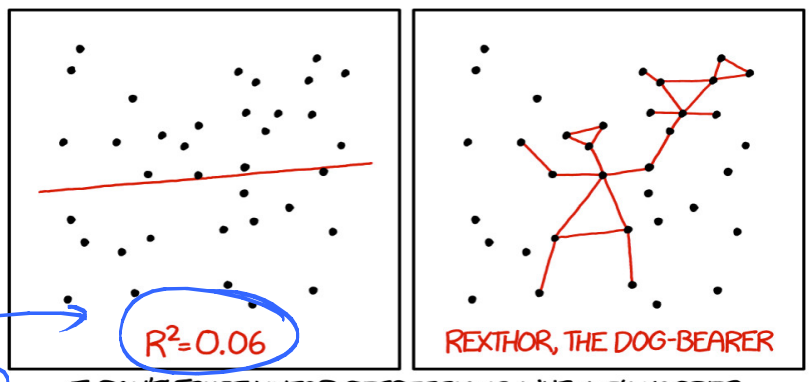


Scientific Amer. March 2018

Authors claim the bigger a brain (x-axis) the more the gain from tutoring (y-axis).

Ignore the regression line; would you agree with the conclusion, based on the scatterplot? What if we remove the 1 or 2 circled patients.

Both of these studies are highly sensitive to sampling. And even if the regression lines are truly sloped, given the huge spread in the scatterplots, do we care that the slopes are non-zero?



below

I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Regression is a very powerful (sharp) tool. But, if you are not careful, you can do serious damage!

Complete change of perspective !

So far, we have viewed regression as "curve-fitting."

Data: (x_i, y_i)

Model: $y_i = \alpha + \beta x_i + \epsilon_i$

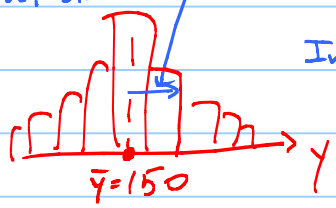
Minimize: $SSE = \sum \epsilon_i^2 \Rightarrow \hat{\alpha}, \hat{\beta} \Rightarrow$ predict: $\hat{y}(x) = \hat{\alpha} + \hat{\beta} x$

($\hat{\alpha}, \hat{\beta}$: OLS estimates of α, β.)

→ But regression can also be viewed as a method for error reduction.
→ This other way of viewing regression will look very different, and it's also less intuitive/geometrical.
→ But it is the more useful and generalizable view. In this view, the focus is on the variance of the y's,
→ and as a result, the method is called The Analysis of Variance (ANOVA). It leads to quantities called R^2 and s_e which together quantify how good is the regression model.

Let me motivate it:

→ Suppose we want to know a table's length.
→ We measure it, repeated (say 101 times): y_i
→ make the histogram



$s_y = \sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2} = 10$ cm

$\equiv s_{yy} = 10^4$ cm^2

Important (see next page)

So, if we use $\bar{y}$ as our estimate of the true table length, then s_y (or s_{yy}) gives us a measure of our uncertainty. And so, we can report Table length: 150 ± 10 cm $(\bar{y} \pm s_y)$.

(low precision.)

→ Now, suppose you are unhappy with the large s_y.
→ You may wonder, could some of that variability be due to something else that is varying everytime you
→ make a measurement of y. E.g. temperature? ← call this x.
→ 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. Here is the math ↴

The Analysis of Variance (ANOVA) Decomposition:

How much of The variation in y (Table length) is due to the (linear) relationship between y and x (temperature)?

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

$\sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy}$ ← Let's call This "variation".

(See hw) $\left\{ + \hat{y}_i - \hat{y}_i \; , \; \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \right.$ (OLS)

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

- SS_{total}: total variation in y.
- $SS_{explained}$: variation in y due to (or explained by) x.
- $SS_{unexplained}$: variation in y unexplained by x. (Error; E for Errors, not for Explained)

SST = $SS_{explained}$ + SSE

e.g. 10^4 = something + something less Than 10^4.

So, if we use $\hat{y}$ as our estimate of True length, Then The uncertainty is reduced from S_{yy} (or SST) to SSE.

⇒ $SS_{explained}$ is generally reported as a percentage of SST:

$$R^2 = \frac{SS_{expl}}{SST} (\times 100) =$$ percentage of The variability in y (e.g. Table length) That is due to (or can be explained by) x (temperature)

meaning of R^2

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

⇒ SS_{unexp} is reported as an "average": $s_e^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_i^n (y_i - \hat{y}_i)^2$ (funny Avg.; error)

meaning of s_e = "typical error"

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

⇒ Altogether, for y (Table length):

Report $\hat{y}(x) \pm s_e$

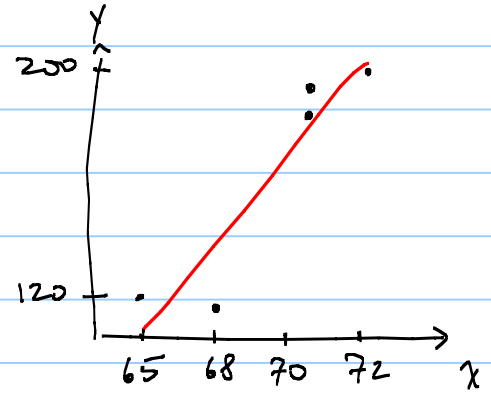
The ANOVA decomposition (above) assures $s_e < s_y$ (error reduction)

Jargon: R^2 = Coeff. of determination
s_e = sample std. dev. of errors

Last example: $\ln(y \sim x) \Rightarrow \hat{y}(x) = -7.55 + 13.28x$

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

$\bar{y} = 161.6$, $s_y = 39.5$



⇒ Without regression, if I wanted to predict the next person's weight, I would say: $\bar{y} \pm s_y = 161.6 \pm 39.5$ (*) (see bottom)

⇒ With regression:

$SST = \sum_i (y_i - \bar{y})^2 = \cdots = 6251$ ← variability in y is reduced from to →

$SSE = \sum_i (y_i - \hat{y}_i)^2 = (-1.5)^2 + (5.1)^2 + (11.5)^2 + (-30.3)^2 + (15.1)^2 = 1307$

⇒ $R^2 = \frac{SS_{expl}}{SST} = \frac{SST - SSE}{SST} = \frac{6251.2 - 1307}{6251.2} = 0.79.$

Meaning: 79% of the variability (or variation) in y (weight, ...) is due to (can be explained by) x (height, ...).
↑ or Table length or ... ↑ or temperature, or ... .

⇒ The other piece of the decomposition: $s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1307}{5-2}} \simeq 21$ pounds

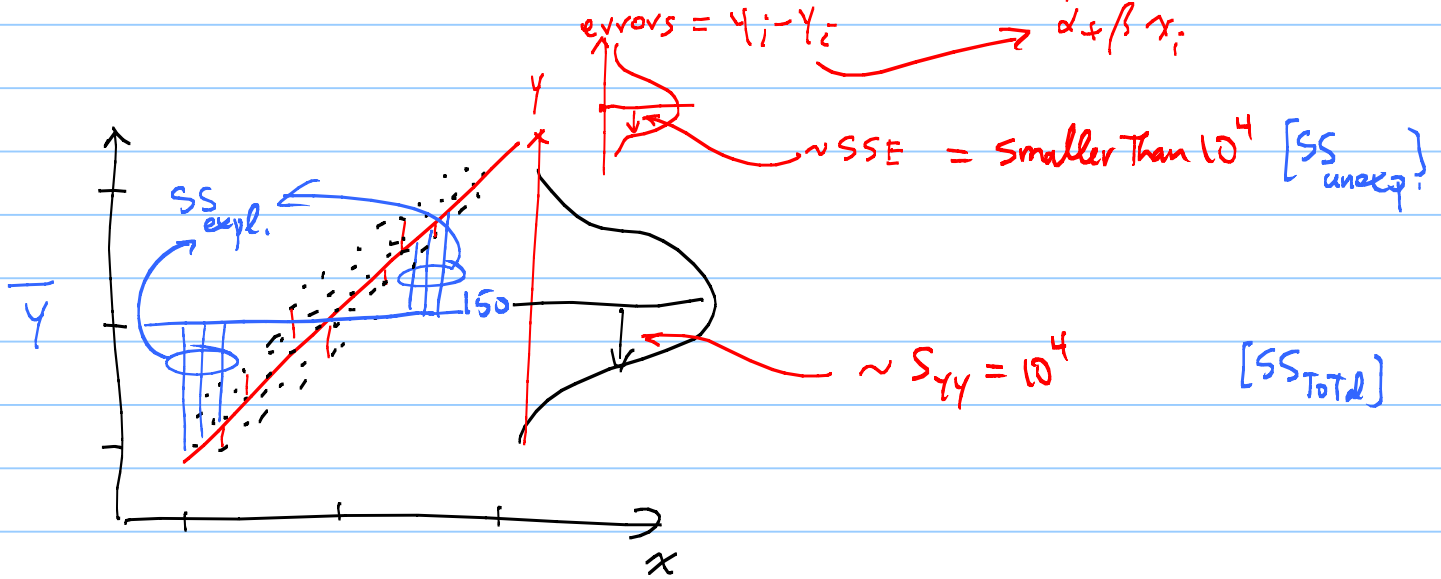
Meaning: The typical error (ie. deviation of y values from the fit) is about 21 pounds.

⇒ Report: weight (or Table length): $\hat{y} \pm 21$ (*) with $R^2 = 0.79$
or ICP or ...
$\hat{y} = -7.55 + 13.3(x)$ ← height or FV or ...

(*) Note that the initial uncertainty (39.5) has been reduced (21) by doing regression.

FYI

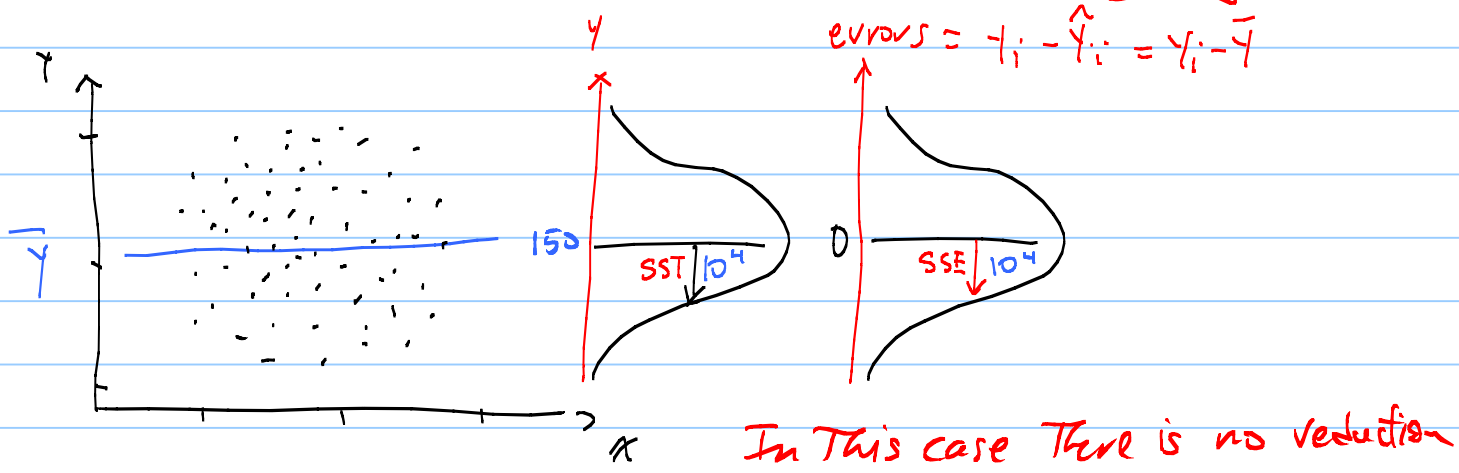
Picture for the ANOVA decomposition:



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 (S_{yy}) is reduced to SSE.

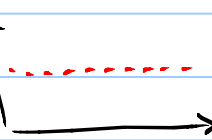
When there is no relationship between x and y, then the fig looks like below. Note that this situation is equivalent to the situation where we have data only on y, and not on x at all. In that case the best prediction for every case is $\bar{y}$ (see hw):



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

FYI

If there is no x data, then the OLS prediction $\hat{y}$ is just $\bar{y}$.

I.e. 

What are the R^2 and s_e?

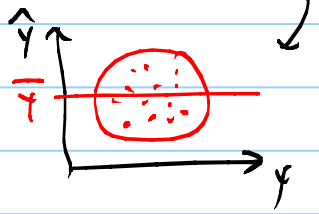
$\hat{y}_i = \bar{y}$; defn. of s_y^2.

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

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

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.

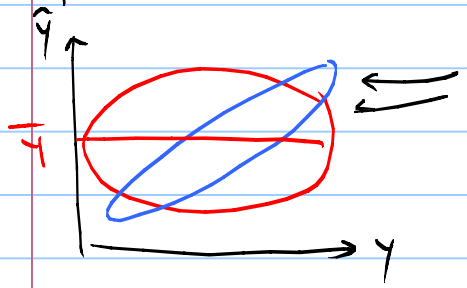
Another situation when nothing is gained is if we make random predictions, e.g. $\hat{y}_i$ = random. Suppose the mean and the var. of these random predictions are the same as those of observations, ie. $\hat{y}_i$ = random with $\bar{\hat{y}} = \bar{y}$, $s_{\hat{y}} = s_y$. The picture is



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 s_e.

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_lect12-1

a) Show that the sample mean of the predictions, ie. $\bar{\hat{y}} \equiv \frac{1}{n}\sum_i \hat{y}_i$ is equal to the sample mean of the y observations, ie. $\bar{y}$.
Hint: use $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

b) Then, use part a to show that the sample mean of errors, $\bar{\hat{\epsilon}} \equiv \frac{1}{n}\sum_i (y_i - \hat{y}_i)$, is equal to zero. Note: $\hat{\epsilon}_i \equiv y_i - \hat{y}_i$.
(FYI: This is why s_e is defined as $\sqrt{\frac{\sum (y_i-\hat{y}_i)^2}{n-2}}$; otherwise, it's zero!)

c) A property of OLS is $\sum_i \hat{\epsilon}_i x_i \equiv (y_i - \hat{y}_i)\, x_i = 0$. You don't need to prove it; it's essentially $\frac{\partial}{\partial \beta} SSE = 0$ written differently. But do use that result to show that the correl. coeff. between the errors ($\hat{\epsilon}$) and the predictions ($\hat{y}$) is zero. Hint: The numerator of the r of u and v is $\sum_i (u_i - \bar{u})(v_i - \bar{v})$. Also use the results of parts a and b.

hw_lect12-2

Here are 3 important facts about OLS (Do not derive them):

$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ ← prediction eqn.

$\sum_i \hat{\epsilon}_i \equiv \sum_i (y_i - \hat{y}_i) = 0$ ← This is $\frac{\partial}{\partial \alpha} SSE$, written differently.

$\sum_i \hat{\epsilon}_i x_i \equiv \sum_i (y_i - \hat{y}_i) x_i = 0$ ← This is $\frac{\partial}{\partial \beta} SSE$, written differently.

Now, consider the following identity (that we used in lect):

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

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

Use the above 3 facts to show that the last term (cross-term) is zero

get used to this phrase.

hw_lect12_3:
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 SST 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.