Lecture 15 (Ch. 3)

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

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

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

$$SSE = \sum_i^n \epsilon_i^2$$

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

Compare!

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

i.e. $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

= "best" prediction.

Then, I started to tell you about a very different, but more useful (and more difficult to understand) way of doing regression.

We begin by looking at the spread in y_i, and ask if we can reduce that variability in y by taking into account the variability of other quantities. E.g. can we reduce the variability in our table length by accounting for the variability in temperature? Yes, This is how

Analysis of variance (ANOVA) approach to regression:

Table length

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

temperature.

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

$\boxed{\text{OLS}}$

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

E for Error, not for Explained

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

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)

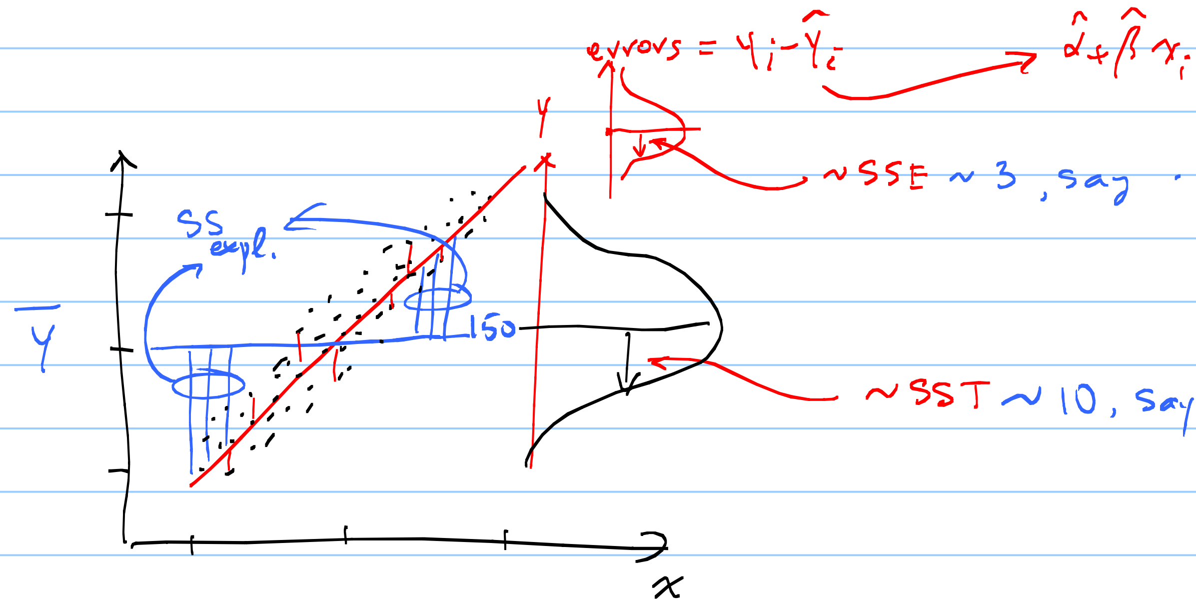
The other piece, $SS_{unexpl.} = SSE$, is a sum (of squares), and so can be can be "Averaged" to provide a measure of typical error

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

"Avg." error $\sim$ typical error.

Typo in book on p.121

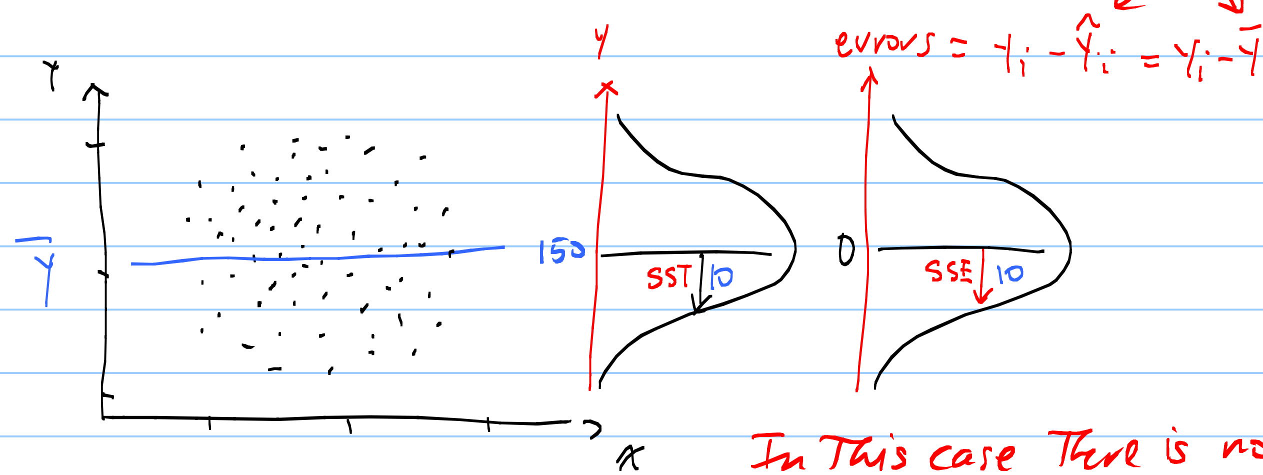
Picture for The above decomposition: (ANOVA)



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 how):



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.

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

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

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

The other piece of the decomposition:

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

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

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

$\hat{y} = -755 + 13.3x$

$\boxed{\text{Summary}}$

⇒ Suppose we have data on variable y, only.
What's the "best" number for predicting y? | see hw (below)

Sample mean of y, ie. $\bar{y}$ report $\bar{y} \pm S_y$

⇒ Suppose we also have data on x, which is related to y.
What's the "best" number for predicting y?

The fitted value $\hat{y} = \hat{\alpha} + \hat{\beta} x$. report $\hat{y} \pm S_e$

Here, the number depends on x.

We will improve on these, later.

⇒ How?

Given data (x_i, y_i) $i = 1, 2, \ldots, n$

assume $y = \alpha + \beta x$,

which means $y_i = \alpha + \beta x_i + \epsilon_i$ ← errors.

minimize $SSE = \sum_{i=1}^{n} \epsilon_i^2$

to get $\hat{\alpha}, \hat{\beta}$ OLS estimates of α, β

predict: $\hat{y} = \hat{\alpha} + \hat{\beta} x$ OLS fit to data.

Decompose Var. $S_{yy} = SST = SS_{exp} + SS_{unexp}$

$\frac{SS_{exp}}{SST} = R^2$

$S_e = \sqrt{\frac{SS_{unexp}}{n-2}}$ = std. dev. of errors.

$\boxed{\text{Typo p.121}}$

Note: The idea of minimizing SSE (in fitting) translates to maximizing $SS_{explained}$ (in ANOVA of regression)

I say $\hat{\alpha}, \hat{\beta}$, SSE, book says a, b, SSResid (and SSE)

Summary continued

⇒ In the decomposition, $SS_{explained}$ is converted to a %, and reported as R^2.

Note $SS_{exp.} = \sum_i (\hat{y}_i - \bar{y})^2$, but sometimes it's easier to compute it as $SST - SSE$, and sometimes from $\hat{\beta}$:

$$SS_{exp.} = \hat{\beta} S_{xy} = \hat{\beta}\, n\,(\overline{xy} - \bar{x}\,\bar{y}) = \cdots$$

⇒ SSE (SS_{unexp}) is not reported "raw" either. It is converted to a standard deviation, and it's called std. dev. about regression, or std. dev. of errors.

Interpretation

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2} \sim \text{typical error}$$

Compare with $S_y = \sqrt{\frac{1}{n-1} \sum_i (y_i - \bar{y})^2}$ ~ typical deviation of y from its mean.

⇒ These quantities are generally formatted in an ANOVA Table. Look at p. 121 and learn how to read the outputs to identify what you need. For example, some computer outputs may call R^2, Coeff. of determ., or r^2, R-sqd ...

Also, they may give RMSE, instead of S_e:

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\underbrace{\frac{1}{n-2} \sum_i^n}_{\text{funny mean}} (\underbrace{y_i - \hat{y}_i}_{\text{error}})^2}$$

Root — squared

see examples in Lab.

hw-lect 15-1

Suppose all we have are data on a single variable y: y_i, $i=1,2,3,\dots,n$. (No x, at all) Show that the predictor that minimizes SSE is the sample mean $\bar{y}$. Hint: let γ denote the prediction, and then minimize SSE.

hw-lect 15-2 Consider the following decomposition:

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

By R

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

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 $y = 2 + 3x + \epsilon$ with ϵ 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_lect15-3

For the data shown in problem 3.22

a) compute the eqn of the OLS fit

b) Compute the total variation, SST.

c) Decompose it into explained and unexplained.

d) Compute R^2, and interpret (in English).

e) Compute the std. dev. of errors, and interpret (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.