Lecture 14 (Ch. 3)

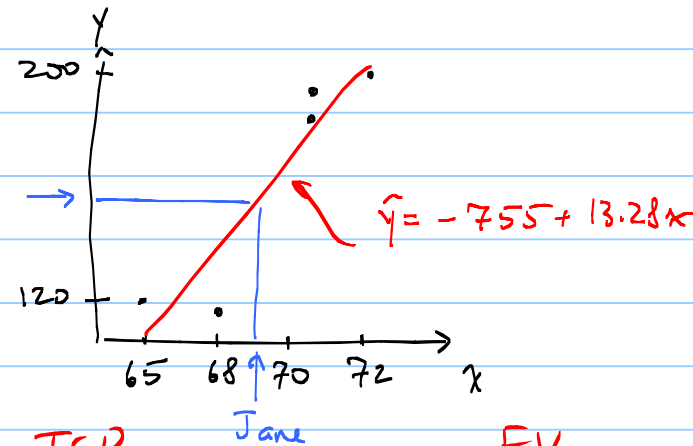
Last time we learned about regression (or fitting).
We learned that given data on x and y, the "best" fit is

$$f(x) = \hat{\alpha} + \hat{\beta} x \quad \text{where} \quad \hat{\beta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Here was the example:

Example

	height (x)	weight (y)
	72	200
Joe:	70	180
	65	120
	68	118
	70	190



⇒ We can now predict everyone's ~~weight~~ ICP, from their ~~height~~ FV:

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

	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

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

⇒ For people outside the data set (e.g. 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.

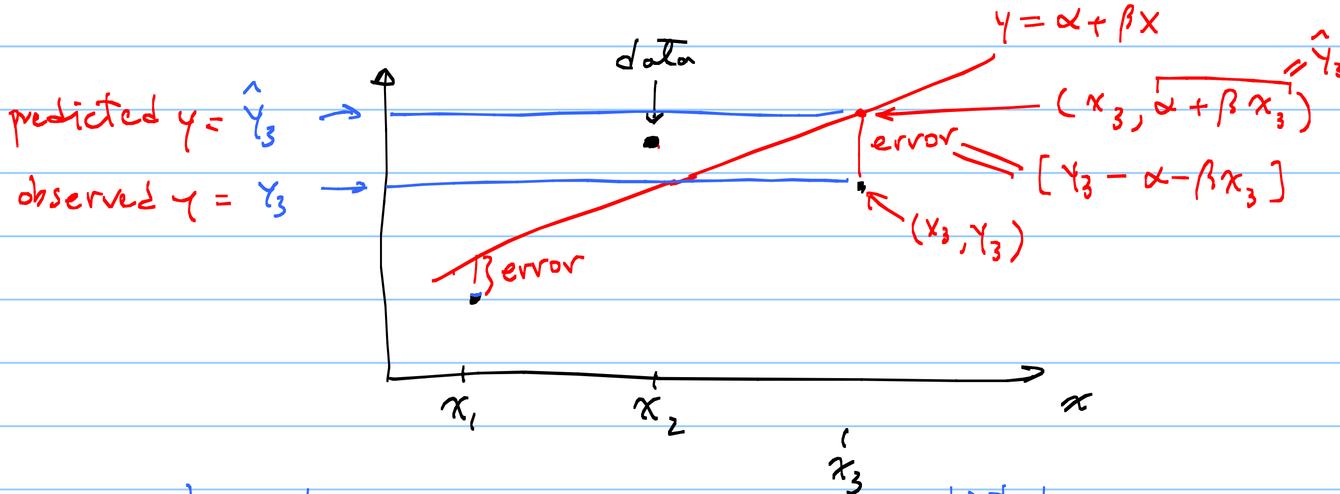
Now, let's derive the eqns for $\hat{\alpha}, \hat{\beta}$:

Derivation:

Called Ordinary Least Squares (OLS)

One very common selection criterion is to take the fit (line) that has the smallest Sum of Squared Errors (SSE)
or equivalently Mean " " " $(MSE = \frac{1}{n} SSE)$

Suppose we have n cases of data: (x_i, y_i) $i = 1, 2, 3, \ldots, n$



observed y predicted y

$$MSE = \frac{1}{n} SSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

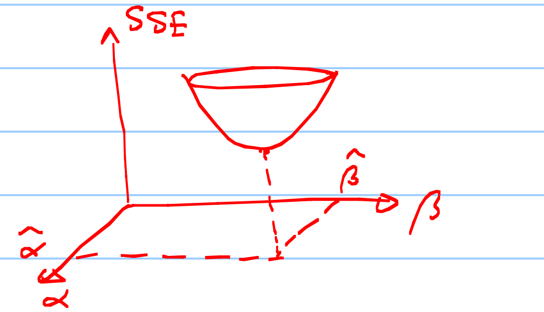
# of cases

Minimize MSE $\Longrightarrow$ differentiate w.r.t. α, β; set to zero; solve for the critical values of α, β $\Longrightarrow$ $\boxed{\hat{\alpha}, \hat{\beta}}$

The specific values of α, β that minimize SSE are called OLS estimates of α, β, and denoted $\hat{\alpha}, \hat{\beta}$:

$$\frac{\partial}{\partial \alpha} MSE(\alpha, \beta) \Big|_{\alpha = \hat{\alpha},\ \beta = \hat{\beta}} = 0$$

$$\frac{\partial}{\partial \beta} MSE(\alpha, \beta) \Big|_{\alpha = \hat{\alpha},\ \beta = \hat{\beta}} = 0$$



If you are not familiar with partial derivatives, $\frac{\partial}{\partial \alpha}$, Then just Think of Them as total derivatives. Let's do one:

Walk Thru These

$$\frac{\partial}{\partial \beta} MSE = \frac{1}{n} \sum_i \frac{\partial}{\partial \beta} [y_i - \alpha - \beta x_i]^2$$

$$= \frac{1}{n} 2 \sum_i [y_i - \alpha - \beta x_i]^1 [-x_i]$$

$$= -\frac{2}{n} \sum_i [x_i y_i - \alpha x_i - \beta x_i^2]$$

$$= -2 \left[\underbrace{\frac{1}{n} \sum_i x_i y_i}_{} - \alpha \underbrace{\frac{1}{n} \sum_i x_i}_{} - \beta \underbrace{\frac{1}{n} \sum_i x_i^2}_{} \right]$$

$$= -2 [\overline{xy} - \alpha \bar{x} - \beta \overline{x^2}]$$

$$\therefore \boxed{\overline{xy} - \hat{\alpha} \bar{x} - \hat{\beta} \overline{x^2} = 0}$$

That's 1 eqn for 2 unknowns $(\hat{\alpha}, \hat{\beta})$. But There is $\frac{\partial}{\partial \alpha}$:

$$\frac{\partial}{\partial \alpha} MSE \Big|_{\hat{\alpha}, \hat{\beta}} = 0 \implies \boxed{\bar{y} - \hat{\alpha} - \hat{\beta} \bar{x} = 0}$$ See how, below.

Now we have 2 eqns for 2 unknowns. Solve!

$$\boxed{\hat{\beta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}}$$

Normal equations of regression.

R: lm(y ~ x)

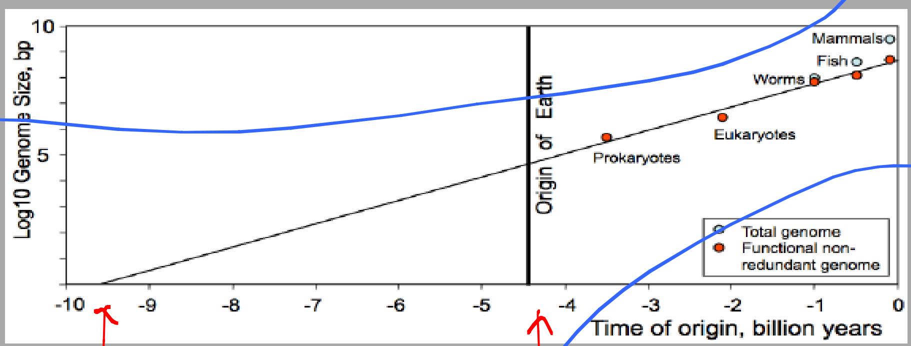
$\boxed{\text{Q1:}}$ For The above example, with $\hat{y} = -755 + 13.28x$, The SSE is given by $(-1.5)^2 + (5.1)^2 + \cdots + (15.1)^2$. The SSE of any other line $(\hat{y} = \cdots)$ will be

$\boxed{\text{A) greater}}$ B) smaller C) equal to D) Will depend! I.e. none of The above.

↑ We minimized SSE to get $\hat{y} = -755 + 13.28x$.

Here is an example of regression where The x-intercept is of interest. Sharov & Gordon (2013) "Life Before Earth":

What is most interesting in this relationship is that it can be extrapolated back to the origin of life. Genome complexity reaches zero, which corresponds to just one base pair, at time ca. 9.7 billion years ago (Fig. 1). A sensitivity analysis gives a range for the extrapolation of ±2.5 billion years (Sharov, 2006). Because the age of Earth is only 4.5 billion years, life could not have originated on Earth even in the most favorable scenario (Fig. 2). Another complexity measure yielded an estimate for the origin of life date about 5 to 6 billion years ago, which is similarly not compatible with the origin of life on Earth (Jørgensen, 2007). Can we take these estimates as an approximate age of life in the universe? Answering this question is not easy because several other problems have to be addressed. First, why the increase of genome complexity follows an exponential law instead of fluctuating erratically? Second, is it reasonable to expect that biological evolution had started from something equivalent in complexity to one nucleotide? And third, if life is older than the Earth and the Solar System, then how can organisms survive interstellar or even intergalactic transfer? These problems as well as consequences of the exponential increase of genome complexity are discussed below.



Confidence Bands (Ch. 11)

origin of life

Earth's age

From This They conclude That Life predates Earth, and That life must have been formed on some other planet, Then transported to Earth.

In a follow-up paper (Marzban et al. (2014): "Earth Before Life", Biology Direct 9:1) we showed That There are (at least) 2 problems with That analysis

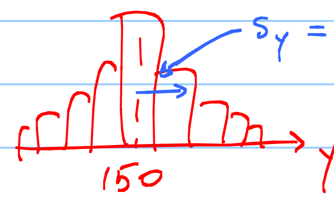
1) Extrapolation is bad!
2) Uncertainty ("Confidence Bands", in Ch. 11) must be considered.

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 Table Length, y.

→ Repeat, and histogram:



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

→ One may report:

True length = 150 ± 10 cm

→ 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.

The (scary) math will be in the next lecture.

hw_lect 14-1 : Show that $\frac{\partial}{\partial \alpha} MSE|_{\hat{\alpha},\hat{\beta}} = 0$ implies $\bar{y} - \hat{\alpha} - \hat{\beta}\,\bar{x} = 0$

hw_lect 14-2 : Show that $\hat{\beta}$ as defined by $\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$ or $\frac{S_{xy}}{S_{xx}}$ can be written as $\hat{\beta} = r\frac{S_y}{S_x}$ where S_x = sample std. dev. of x.
S_y = " " " " y.

hw_lect 14-3 :

Suppose data on x and y fall on a straight line $y_i = b + m x_i$.
If we perform a linear fit $y = \alpha + \beta x$ to this data,
What is the value of the OLS estimate of β ?

hw_lect 14-4 According to the OLS fit, what is the predicted value of y, when $x = \bar{x}$ (ie. when x = sample mean of x).
Hint: All you need is $\hat{\alpha}$.

hw_lect 14-5 Come-up with another example of x and y (like FV and ICP), where regression can help in predicting y from x in a situation where without regression the "cost" of measuring y directly is extremely high (like ICP).