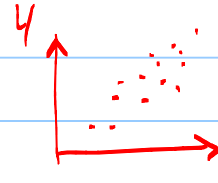
Lecture 11 (ch. 3)

Last time:

1) scatterplots for seeing the relationship between 2 continuous r.v.'s. Learn the types



2) Correlation as a summary measure for the strength of the association. "skinniness"

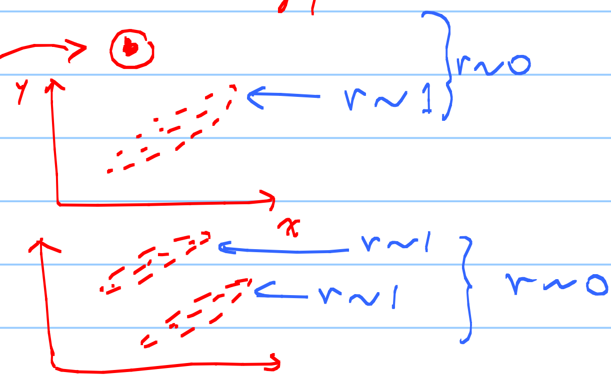
$$r = \frac{1}{n-1} \sum_{i}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

But, every summary measure (r, $\bar{x}$, s, ...) can be misleading.

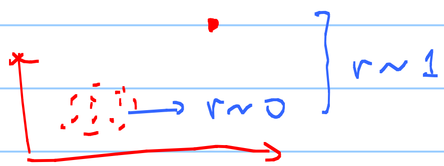
When you see r = large (e.g. 0.9) or r = small (0.1), you should wonder if r is lying to you.

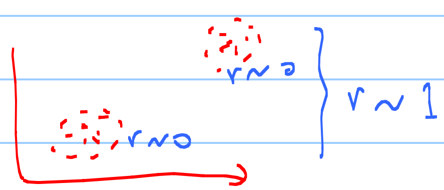
⇒ There are situations which make r "artificially" small:
misleadingly

1) when there is a nonlinear rel.
2) when there are outliers
3) when there are clusters



⇒ There are situations which make r "artificially" large:



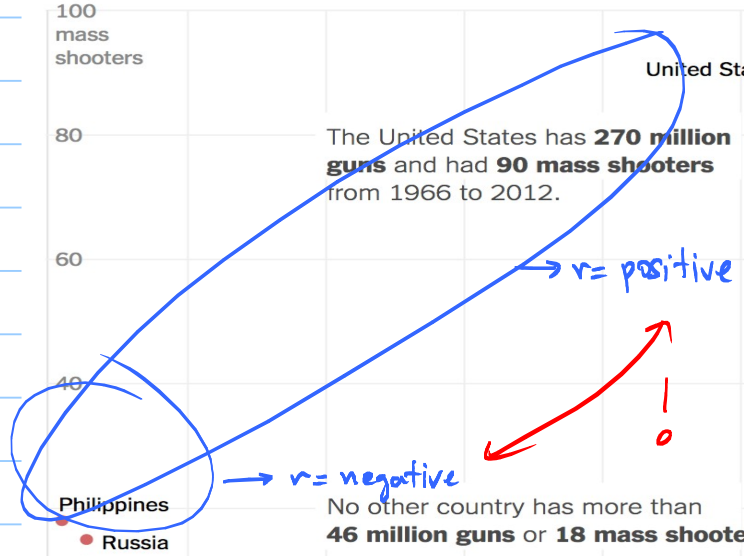


Also see "ecological correl" in lab.

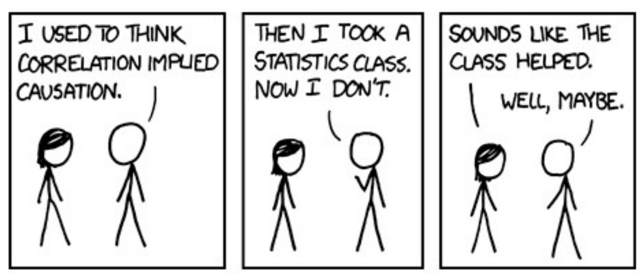
<u>Moral:</u> r (like any other summary measure) can be misleading if the data have clusters, outliers, ... So, regardless of the r value you get in your problem, look at the scatterplot, too.

There are some situations where even the scatterplot can fail in capturing all the facets of the relationship. E.g.

Not only r is sensitive to outliers (it can go from negative to positive), but the very notion of association (eg.when viewed through your eyes) is sensitive to outliers. In this example, even forgetting r altogether, because of a single observation, the association goes from a negative one to a positive one. As such, the question of whether gun ownership and mass shootings are related does not have a yes/no answer at all! And it doesn't matter what measure of association you use.



Finally,



Important:
Association/correlation does not imply Causation.
Even if there is a strong correlation or association between 2 variables, that does not mean one causes the other. E.g. Shoe size and reading ability are associated. But I cannot increase my reading ability by wearing a larger shoe.

Even more important:
Even a non-causal association can be useful; for example, it can be used to predict one from the other. You can predict reading ability from shoe size.

Switching gears (even though it may not seem so, make sure you see the change we are making: we are going from correlation to regression)

Q What is an association between 2 vars. good for?

A 1) It can help in building theories.

2) It sets the stage for building predictive models, where one predicts one variable from the other.

Note: prediction is not in time.

Q Can we use r itself for making predictions?

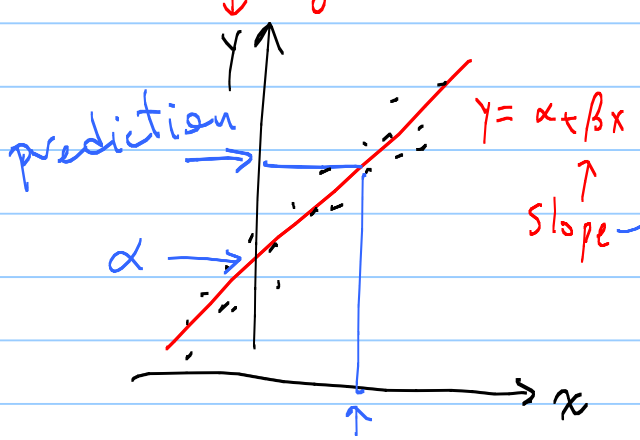
A No. We need a fit, e.g. a line (ie. regression model)

But you do not need a line for computing r.

Keep 2 things apart: Whereas r has (almost) nothing to do with the slope of anything, the fit does. After all, a fit is an equation e.g. $y = \alpha + \beta x$, and so, the slope is very present

BTW, in statistics, (line or curve) fitting is called regression

e.g. Intracranial pressure (ICP) [HARD to measure]



meaning: how much does y change with 1 unit change in x?

Blood flow measured with Ultrasound [EASY to measure]

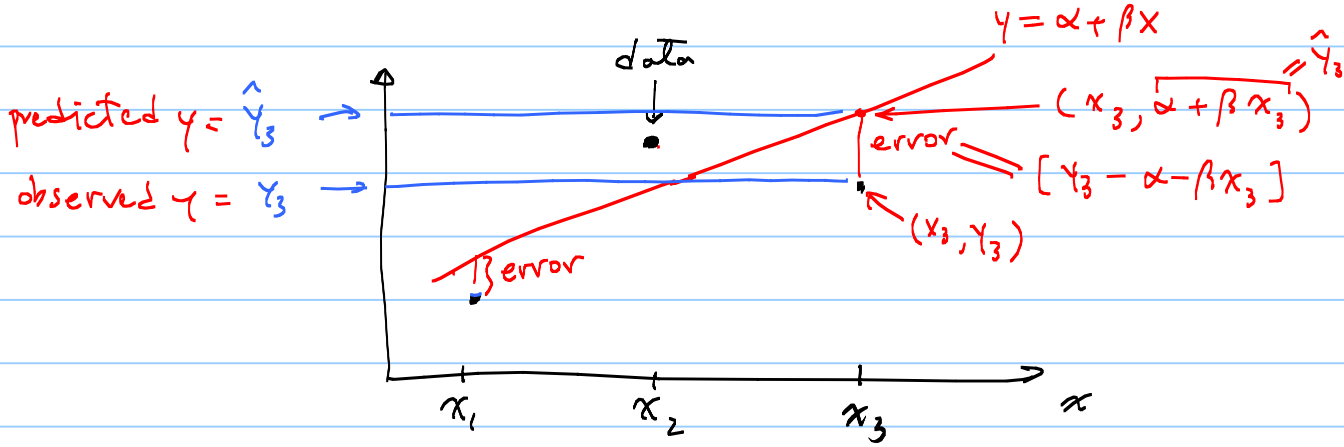
Q: For finite points on a scatterplot, there are lots of possible fits. Which one do we pick?

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 " " " $\left(MSE = \frac{1}{n} SSE\right)$

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



observed y predicted y

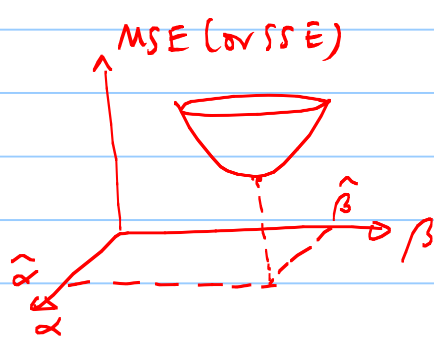
$$MSE = \frac{1}{n} SSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

# of cases

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

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

$$\left(\begin{array}{l} \frac{\partial}{\partial \alpha} MSE(\alpha, \beta) \Big|_{\alpha = \hat{\alpha},\ \beta = \hat{\beta}} = 0 \\ \frac{\partial}{\partial \beta} MSE(\alpha, \beta) \Big|_{\alpha = \hat{\alpha},\ \beta = \hat{\beta}} = 0 \end{array} \right.$$



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

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

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

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

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

$$= -2\left[\overline{xy} - \alpha\,\overline{x} - \beta\,\overline{x^2}\right]$$

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

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

$$\left.\frac{\partial}{\partial\alpha} MSE\right|_{\hat\alpha,\hat\beta} = 0 \implies \boxed{\overline{y} - \hat\alpha - \hat\beta\,\overline{x} = 0}$$

See how, below.

Now we have 2 eqns for 2 unknowns. Solve!

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

Normal equations of regression.

R: lm(y~x)

Notation:

→ What I call $\hat\alpha, \hat\beta$ are denoted a and b in book.

SSE is " SSResid "

(I have good reasons for using my notation!)

→ The book also introduces the notation:

$$S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad S_{xy} = \sum_i (x_i-\bar{x})(y_i-\bar{y})$$

$$S_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2$$

Numerators of sample var. s_x^2, s_y^2.

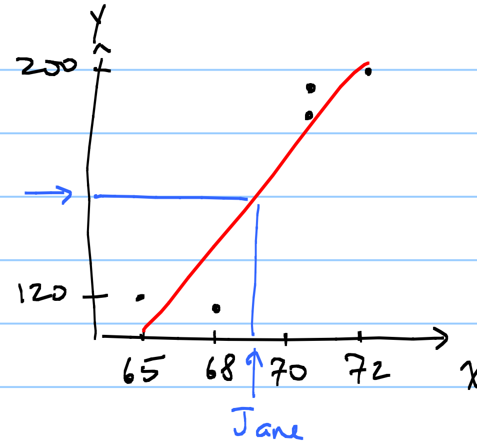
in which case it's easy to show that $\hat\beta = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{S_{xy}}{S_{xx}}$

Example

Height or Blood Flow (Easy)

Weight or ICP (Hard)

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

$\Rightarrow$ E.g. Joe's predicted y based on his x.

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

$\Rightarrow$ We can now predict everyone's y from their x.

	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.

$\Rightarrow$ Finally, be WARNED if you extrapolate

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

hw_lect11_1
I encourage you to come-up with other examples with "Easy" and "Hard" variables, because you're likely to come across something practically useful. But, even if you don't want to do that, do develop a regression model for predicting one of the continuous variables in your collected data from the other continuous variable. By R. Include your code, and report and interpret the slope parameter.

hw_lect11-2 Show that $\frac{\partial}{\partial \alpha} MSE \Big|_{\hat{\alpha}, \hat{\beta}} = 0$ implies $\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$

hw_lect11-3 Prove that the Ordinary Least Square (OLS) fit, (i.e. the one described in this lecture) goes through the point $(\bar{x}, \bar{y})$. Hint: All you need is the Normal eqn. for $\hat{\alpha}$.

hw_lect11_4 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_lect11_5
Values of modulus of elasticity (MoE, the ratio of stress, i.e., force per unit area, to strain, i.e., deformation per unit length, in GPa) and flexural strength (a measure of the ability to resist failure in bending in MPa) were determined for a sample of concret beams of a certain type, resulting in the following data (read from a graph in the article "Effects of Aggregates and Microfillers on the Flexural Propertie of Concrete," Magazine of Concrete Research, 1997 8198):
MoE:
29.8 33.2 33.7 35.3 35.5 36.1 36.2 36.3 37.5 37.7 38.7 38.8 39.6 41.0 42.8 42.8 43.5 45.6 46.0 46.9 48.0 49.3 51.7 62.6 69.8 79.5 80.0
Strength:
5.9 7.2 7.3 6.3 8.1 6.8 7.0 7.6 6.8 6.5 7.0 6.3 7.9 9.0 8.2 8.7 7.8 9.7 7.4 7.7 9.7 7.8 7.7 11.6 11.3 11.8 10.7

a) Plot a scatterplot of Strength vs. MOE. By R.
b) Make a boxplot of MOE, and of Strength. By R.
c) Make a qqplot of MOE, and of Strength. By R.
d) Compute the correlation coefficient between MOE and Strength. By hand. You may use the computer to compute sample means of necessary quantities,but you must use one of the formulas for r.
e) Compare it with the correlation you get from cor() in R.
f) Compute the equation of the OLS fit (i.e., the intercept and slope). By hand.You may use the computer to compute sample means of necessary quantities,but you must use the formulas for OLS intercept and slope).
g) Interpret the slope.
h) Predict Strength when MoE is 39.0 . By hand.
i) Compute the sum squared error (SSE, or SSResid). By hand, but you may use R to compute sample means of necessary quantities.