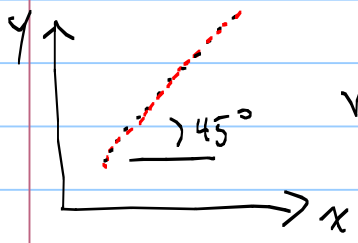
Lecture 10 (Ch.3)

Last time we learned about the corr. coeff.

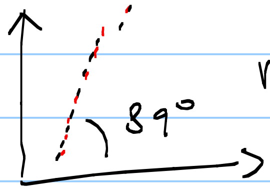
$$r = \frac{1}{n-1}\sum_{i}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

[See last hw for other forms.

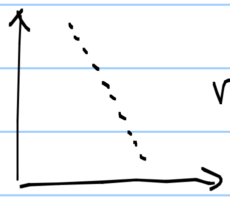
r museum :



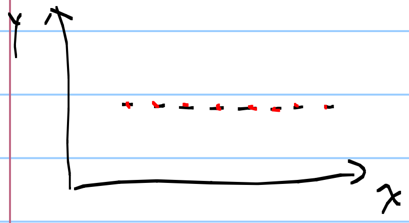
$r = +1$



$r = +1$

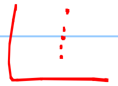


$r = -1$

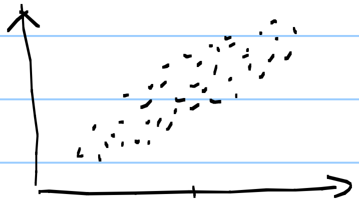


$r = 0$

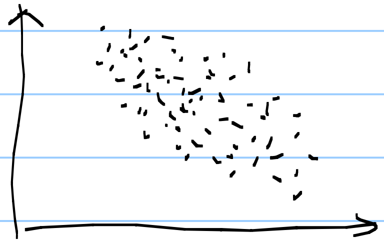
[This involves some limits



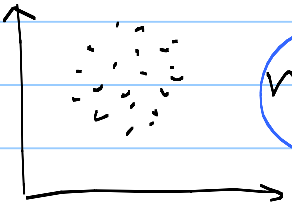
$r=0$



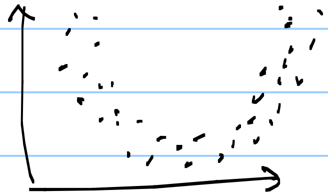
$r \sim 0.7, 0.8$



$r \sim -0.7$
~ -0.6



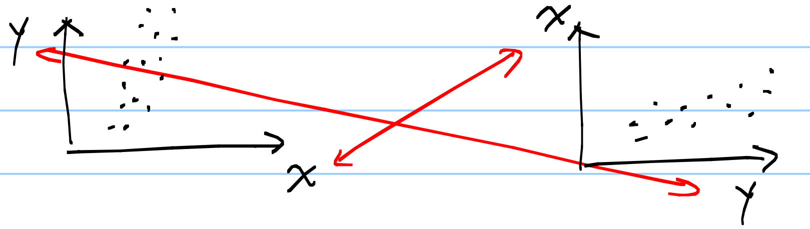
$r \sim 0$



$r \sim 0 \Rightarrow$ r is a measure of linear association

Important: r is a summary measure of a scatterplot. As such, some info is lost when you look only at r. Look at the scatterplot (too)!

Q How does switching x and y affect r ?

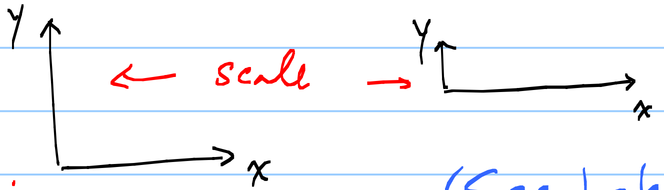


$r_{xy} = r_{yx}$

(because sh $z_x z_y$ in r.)

(See Lab)

Q How does scaling (ie. multiplying all x or y values by some number) affect r ?



It does not!

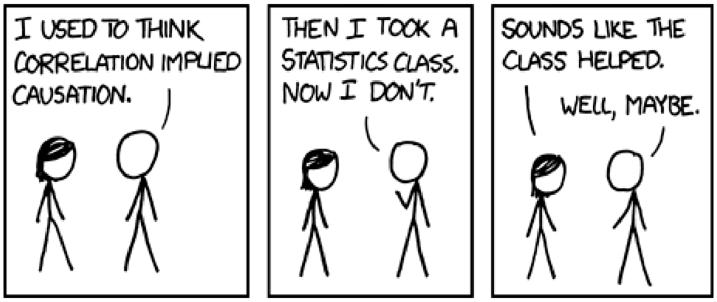
r is invariant under scaling.

(See Lab)

Because $z_i = \frac{x_i - \bar{x}}{s_x}$ " " " "

e.g. $x_i \to c\,x_i$: $\frac{\cancel{c} x_i - \cancel{c}\bar{x}}{\cancel{c} s_x} = \frac{x_i - \bar{x}}{s_x}$

$$s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \longrightarrow \frac{1}{n-1} \sum_i (c\,x_i - c\bar{x})^2 = c^2 s_x^2$$


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Important: Association/Correlation $\not\Rightarrow$ Causation.

Even if there is a strong correlation or association between 2 vars, that does not mean one causes the other.

Shoe size and reading ability are associated

But 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.

Generally, r has The following properties:

$-1 \leq r \leq +1$, $r_{xy} = r_{yx}$, measures linear association $\Updownarrow$ skinniness

unaffected by scaling or shifting

BUT, it can be misleading:

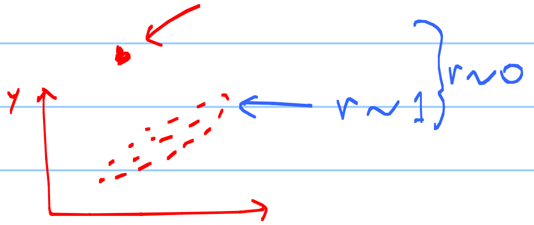
When you see $r =$ large (e.g. 0.9) or $r =$ small (0.1), you should wonder if r is lying to you.

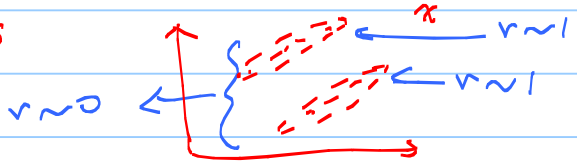
$\Rightarrow$ There are situations which make r "artificially" small:
misleadingly

1) when there is a nonlinear rel.

2) when there are outliers

3) when there are clusters

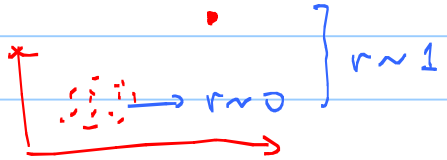


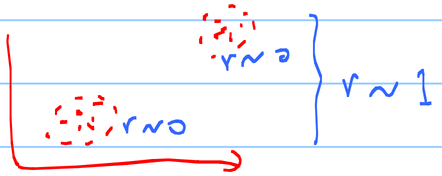


Also keep in mind that $r \neq \varphi$

even if $r = 0.9$, φ may still be 0 . And vice versa

$\Rightarrow$ There are situations which make r "artificially" large:





Also "ecological correl" in lab.

Moral: r is misleading if The scatterplot has clusters, outliers, ... So, regardless of The r value you get in your problem, look at The scatterplot, too.

Switching gears (even though it may not seem so).

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.

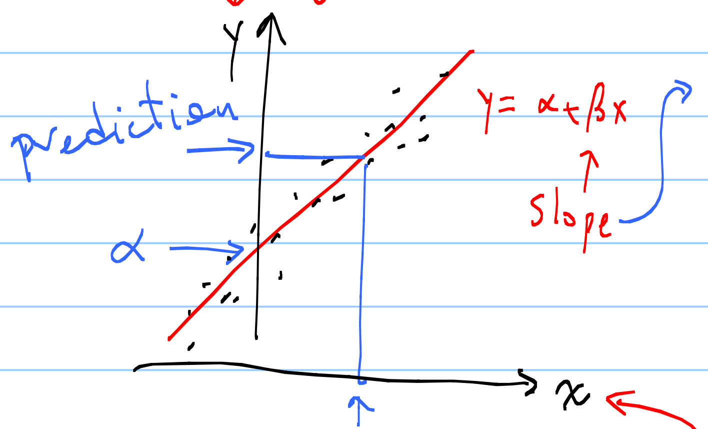
Note: prediction is not in time.

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.

e.g. Intracranial pressure (ICP) Hard to measure



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

anything easy to measure.

e.g. Arterial Blood Pressure (ABP) or Flow Velocity (FV), or ...

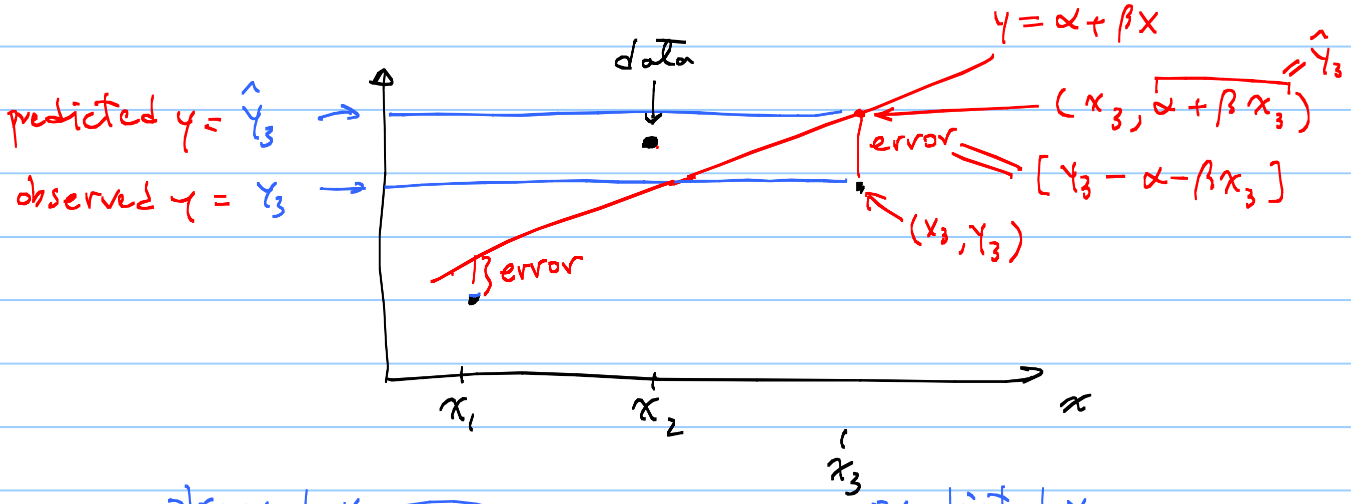
Q: For finite points on a scatterplot, there are lots of possible fits. Which one do we pick?

A: Next.

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) $i = 1, 2, 3 \cdots, n$



observed y predicted y

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

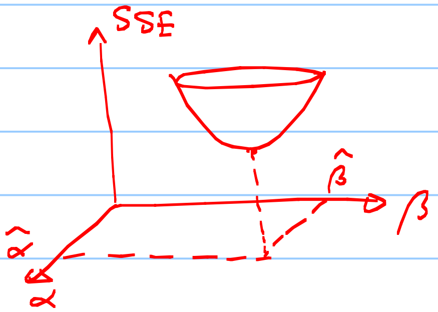
# of cases

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:

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

Walk thru these

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

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

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

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

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

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

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

Now we have 2 eqns for 2 unknowns. Solve!

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

Normal equations of regression.

R: lm(y~x)

Q1: Consider a problem wherein $SSE = \sum_{i=1}^{n} (y_i - \beta)^2$

I.e. The prediction for every y_i is a constant β.

Find $\hat{\beta}$ s.t. SSE is minimized.

A) $\hat{\beta} = 0$ B) $\hat{\beta} = \beta$ C) $\boxed{\hat{\beta} = \overline{y}}$ D) $\hat{\beta} = \infty$

$$\frac{\partial SSE}{\partial \beta} \sim \sum_i (y_i - \beta) \implies \sum_i (y_i - \hat{\beta}) = 0 \implies \sum_i y_i = \sum_i \hat{\beta}$$

$$\sum y_i = n \hat{\beta} \implies \hat{\beta} = \overline{y}$$

In The book, $\hat{\alpha}, \hat{\beta}$ are written as a, b (in italic). But I can't write in italic, and without italic The parameter a gets mixed-up with The English article a! Hence, $\hat{\alpha}, \hat{\beta}$.

The book also introduces The notation:

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

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

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

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

hw-lect 10-1

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 computer.
b) Make a boxplot of MOE, and of Strength. By computer.
c) Make a qqplot of MOE, and of Strength. By computer.
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). You may use the computer to compute sample means of necessary quantities.

hw-lect 10-2 : Show that $\frac{\partial}{\partial\alpha} \mathrm{MSE}\big|_{\hat{\alpha},\hat{\beta}} = 0$ implies $\bar{y} - \hat{\alpha} - \hat{\beta}\,\bar{x} = 0$

hw_lect10-3 :

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_lect10-4 Prove That The OLS fit goes Through The point $(\bar{x}, \bar{y})$.

hw_lect10-5 Show That $\hat{\beta}$ as defined by $\dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$ or $\dfrac{S_{xy}}{S_{xx}}$ can be written as $\hat{\beta} = r\,\dfrac{S_y}{S_x}$ where S_x = sample std. dev. of x.

S_y = " " " " y.