# Lecture 11 (Ch. 3)

In the bivariate world, we have learned about

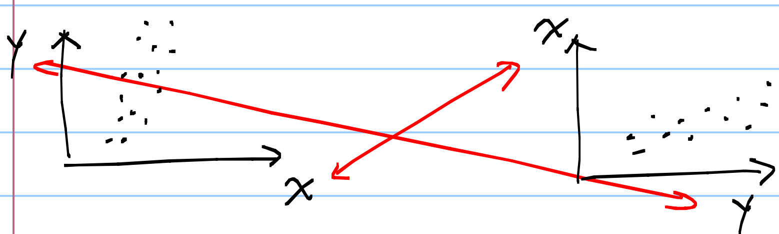
scatterplots : for visualizing the association between $x$ and $y$.

Correlation (coeff.) : for quantifying the strength of the association.

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

= measure of "skinniness" of scatterplot.

Q How does switching $x$ and $y$ affect $r$?



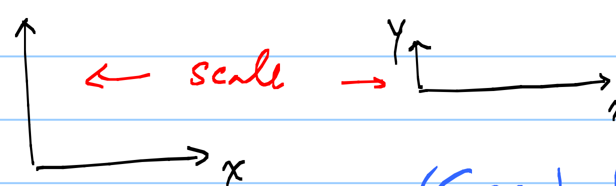

$r_{xy} = r_{yx}$

(because $z_x z_y$ in $r$.)

(See Lab)

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

It does not!



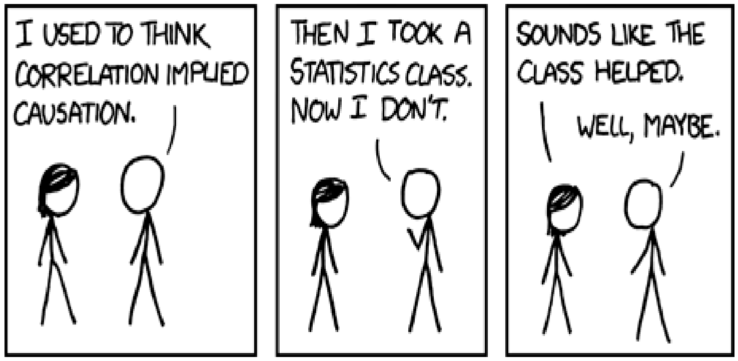

$r$ is invariant under scaling.

(See Lab)

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

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

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




Important: Relation $\not\Rightarrow$ Causation.

Even if there is a strong (linear) relationship between 2 variables, that does not mean that one causes the other.

Shoe size and reading ability are correlated.

But even an acausal relationship can be used for predicting one from the other.

You can predict reading ability from shoe size.

---

Generally, r has the following properties:

- $-1 \leq r \leq +1$
- $r_{xy} = r_{yx}$
- measure of linear assoc. (spread about line)  "skinniness"
- unaffected by scaling, shifting, ...
- misleading!

We have learned that $r$ (Pearson's correlation coefficient) is a measure of the strength of the linear relationship between 2 continuous variables, with "strength" measured by "skinniness". But $r$ 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.

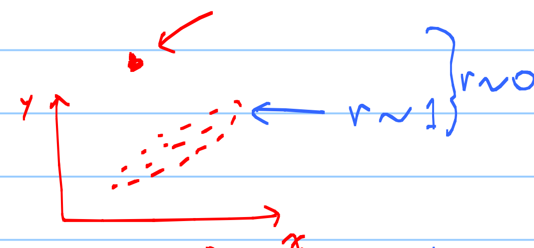
$\Rightarrow$ There are situations which make $r$ "artificially" small:

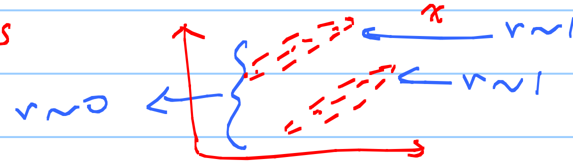
misleadingly

1) when there is a nonlinear rel,

2) when there are outliers

3) when there are clusters

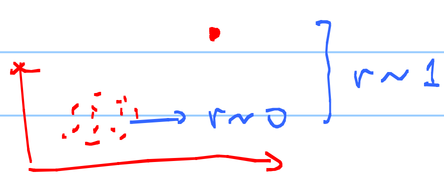


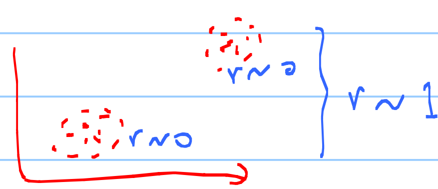




Also keep in mind that $r \neq \rho$

even if $r = 0.9$, $\rho$ 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.

**Q1:** How does $r$ change if $x$ (and/or $y$) are shifted by $c$?

$\boxed{\text{A) } r \to r}$ B) $r \to r+c$ C) $r \to rc$ D) $r \to r+c^2$

$$z_x = \frac{x_i - \bar{x}}{s_x} \to \frac{x_i \cancel{+c} - \bar{x} \cancel{-c}}{s_x} = z_x$$

A linear association

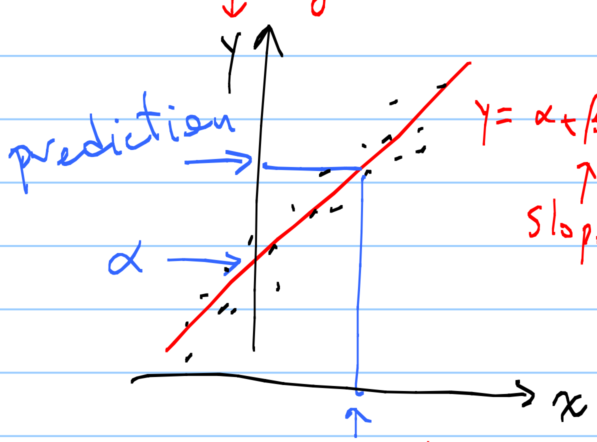
1) can help in building theories.
2) sets the stage for building predictive models, where one predicts one variable from the other.

Note: prediction is not (necessarily) 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) $\boxed{\text{Hard}}$ to measure




$Y = \alpha + \beta x$

slope

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

anything $\boxed{\text{easy}}$ to measure.

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

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

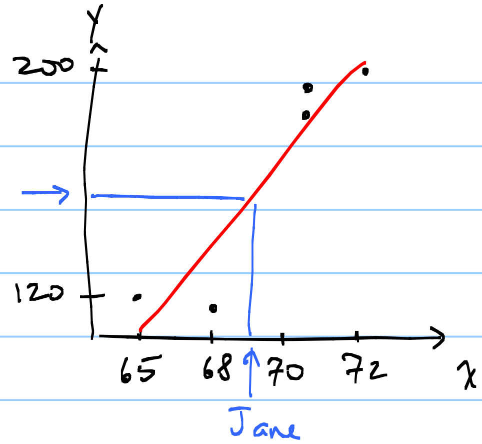
**A**: The line $f(x) = \hat{\alpha} + \hat{\beta} x$ where $\hat{\beta} = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

I'll derive these next time.

## Example

Data

| | height ($x$) | weight ($y$) | $xy$ | $x^2$ |
|---|---|---|---|---|
| | 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~x) $\Longrightarrow$ $\hat{\beta} = 13.3$, $\hat{\alpha} = -755.11$ $\Longrightarrow$ $\hat{y} = -755 + 13.28x$

$\Longrightarrow$ E.g. Joe's predicted weight, according to his height, is

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

$\Longrightarrow$ We can now predict everyone's weight: $\hat{y} = \hat{\alpha} + \hat{\beta}x$ 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 |

Errors (or residuals, deviations, ...)

$\Longrightarrow$ even if they are not in the data set. Need $x$, though.

However, be WARNED if you extrapolate. See next lect. for another example of Bad extrapolation.

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

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:

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

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

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

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

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

Finally, note the resemblance between the formulas for $\hat{\beta}$ and r. But their meaning is completely different.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

---

hw-lect 13-1 (Revised 3.21)

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.