

SE Correlation & regression

PART 1 introduction to concepts



Regression controls values of all other factors.
Can be used to demonstrate causation.
Correlation does not control values of all other factors.
Can be suggestive, but it does not demonstrate causation.

Beware of medical studies that show links between factors, they are almost always just correlation studies.

Q: Do they control all other factors?
A: Rarely.

e.g., compliance (drug vial cap), diet, drug use, etc.

How do we know that sun exposure causes skin cancer?
Correlation between sun exposure and cancer rates.
Regression left and right arms in US and Australia.



Case study: Gordon & Brieva, 2012. *Unilateral Dermatochloasis*. *New England Journal of Medicine*, 366:e25. doi.org/10.1056/NEJMc1104059

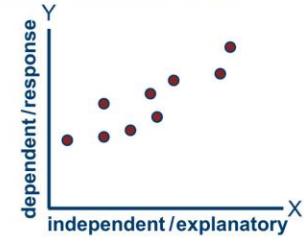
A 69-year-old man presented with a 25-year history of gradual, asymptomatic thickening and wrinkling of the skin on the left side of his face. The physical examination showed hyperkeratosis with accentuated ridging, multiple open comedones, and areas of nodular elastosis. Histopathological analysis showed an accumulation of keratinous material in the infundibula and the formation of cysts within the infundibula. Findings were consistent with the **Ferre-Rouchaud syndrome of photoaged skin**, known as **dermatochloasis**. The patient reported that he had driven a delivery truck for 20 years. Ultraviolet A (UVA) rays filtered through window glass, penetrating the windshield and upper layers of dermis. Chronic UVA exposure can result in thickening of the epidermis and collagen crosslinking, as well as destruction of elastic fibers. This phototoxic effect of UVA is consistent with photoaged skin. This phototoxic exposure to ultraviolet B (UVB) rays is linked to a higher rate of photoaged skin. UVA has also been shown to induce substantial DNA mutations and direct toxicity, leading to the formation of skin cancer. The use of sun protection and topical retinoids and periodic monitoring for skin cancer were recommended for the patient.



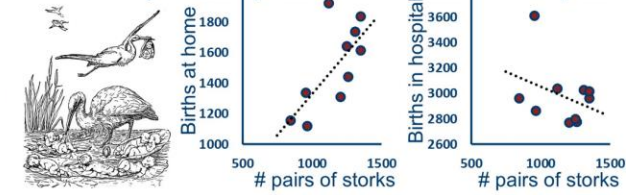
Regression/correlation
For looking for a possible relationship between X and Y.

The usual idea is to put the "causing" variable as the X and the "caused" variable as the Y.

X: independent or explanatory
Y: dependent or response

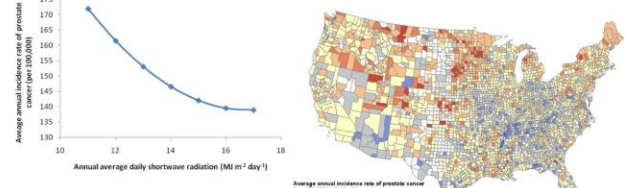


Sometimes the driving factor is subtle. This is **real data** from Germany.



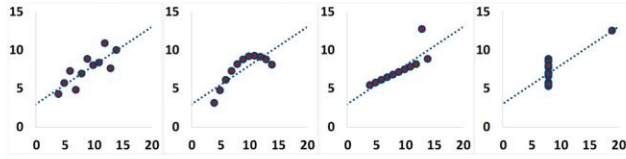
Hofer et al., 2004. *New evidence for the theory of the stork*. *Paediatric and perinatal epidemiology* Vol.18, Iss. 1 pp. 88-92. <https://doi.org/10.1111/j.1365-3016.2003.00534.x>

The relationship between sun exposure and prostate cancer?



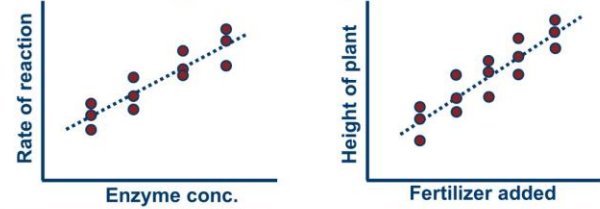
St-Hilaire et al. 2010. *Correlations between meteorological parameters and prostate cancer*. *International Journal of Health Geographics* vol 9, article: 19.

Caution: always plot data before starting stats. Consider Anscombe's quartet.



These four data sets have the same:
mean X, var X, mean Y, var Y, correlation coefficient, coefficient of determination, best fit slope and Y-intercept

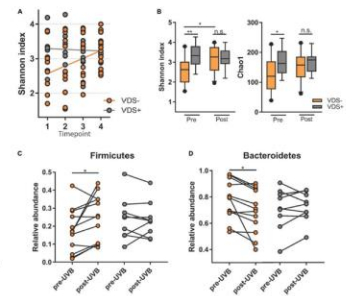
Regression can imply causation



Pro: if we control everything, we can demonstrate causation.
Con: relatively difficult to do.

The relationship between sun exposure and prostate cancer?

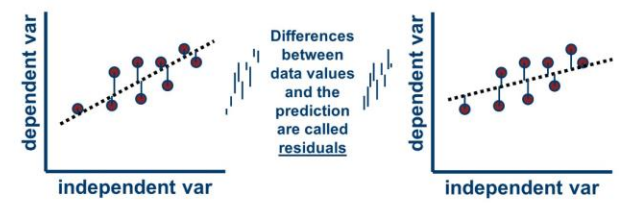
Baseline vitamin D
Added vitamin D



Correlation is not useless. It is often the first step in further studies.

Bosman et al. 2019. *Skin Exposure to Narrow Band Ultraviolet (UVB) Light Modulates the Human Intestinal Microbiome*. *Frontiers in Microbiology* Vol 10, pp. 2410. doi.org/10.3389/fmicb.2019.02410

The criterion for the "best" fit line for data.



Square the residuals to make the magnitudes all positive. "Best fit" is the line with smallest sum of these squares.

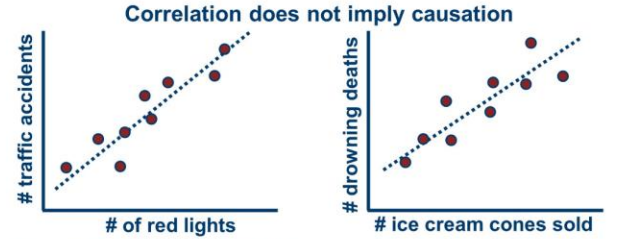
Regression/correlation: two purposes

(1) **Establish causality.** If the only differences are different X values, then, differences in Y are due to X. Requires detailed knowledge.

(2) **Make predictions.** $Y = a + bX$ (not: $y=mx+b$)

PREDICTION ≠ CAUSALITY

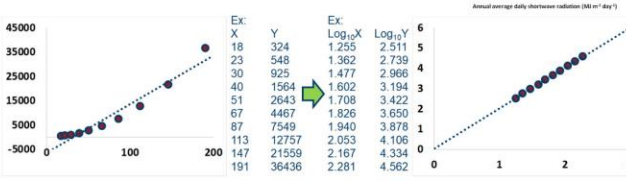
Correlation does not imply causation



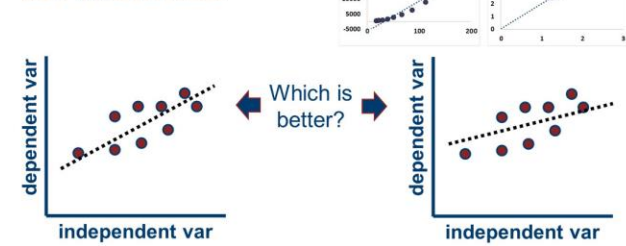
Pro: relatively easy to get data.
Con: other factors can cause correlations.

Q: what do we do about curved relationships?

A: perform a transformation on the data.



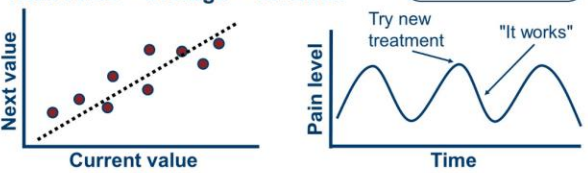
What criterion for the "best" fit line for data?



THE FALLACY OF REGRESSION

Current value = average + variation

Next value = average + variation



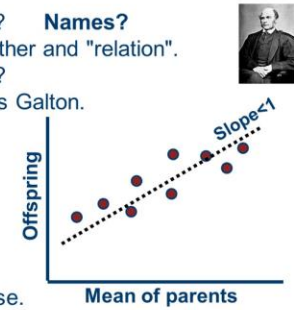
Amazing Good
Excellent Average Good
Good Average Good
Excellent Average Good
Good Average Average
Good Bad Excellent

Q: Why is this called correlation? **Names?**
A: From Latin, "co-" means together and "relation".
Q: Why is this called regression?
A: From Darwin's cousin, Francis Galton.

He worried that offspring of very "good" (e.g., tall, high IQ) parents "regressed" back to the mean.

Truth: goes both ways.

Trait = genes + environment/noise.



StatsExamples.com