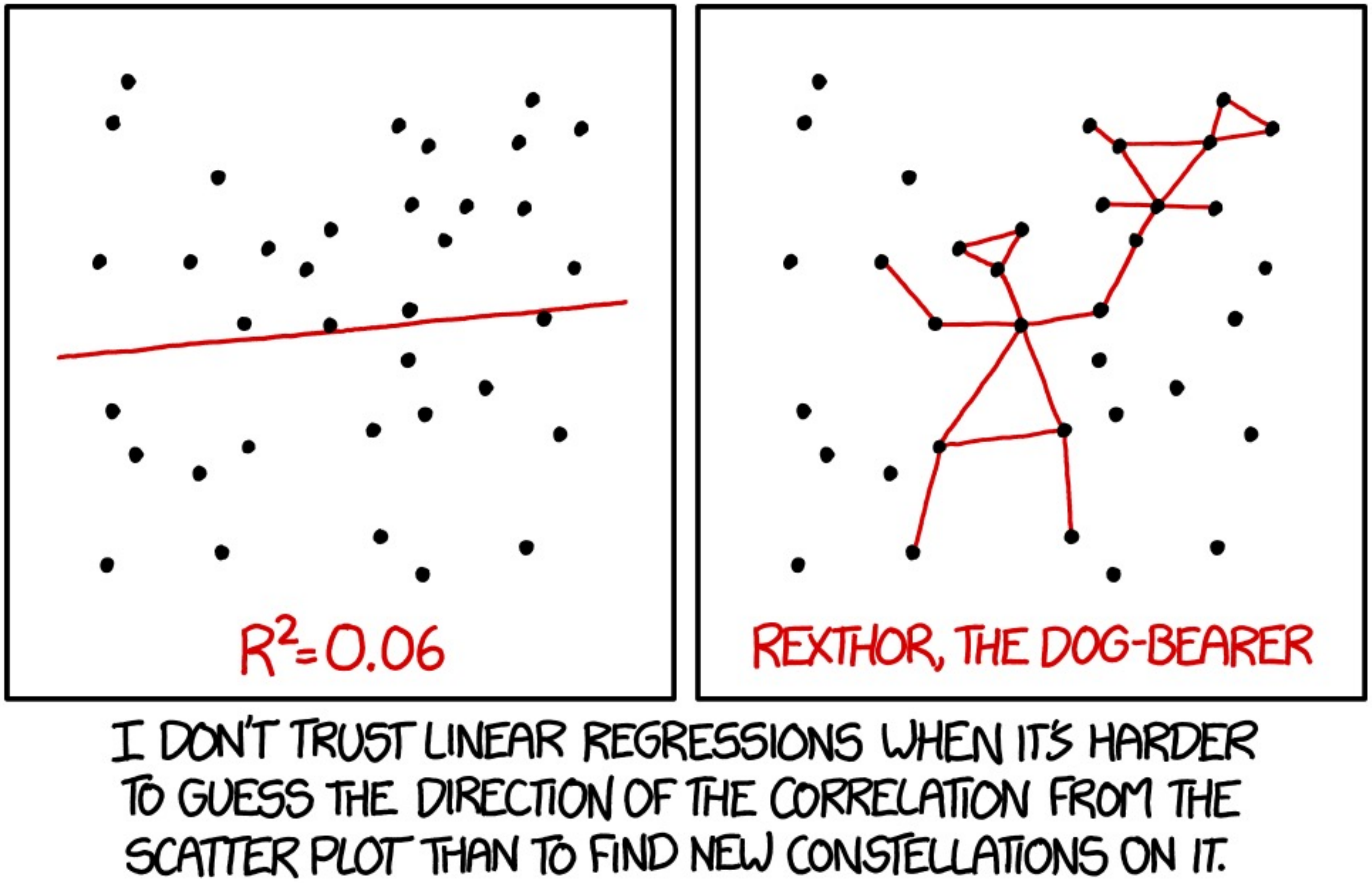
Stat 88: Probability & Mathematical Statistics in Data Science


$R^2 = 0.06$
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Lecture 38 : 4/26/2021

Chapter 11

Correlation

https://xkcd.com/1725/

Equation of the regression line

- $\hat{Y} = \hat{a}X + \hat{b}$

- $\hat{Y}$ is called the fitted value of Y, $\hat{a}$ is the slope, $\hat{b}$ is the intercept where:

- $\hat{a} = \frac{r\sigma_Y}{\sigma_X}, r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X \times Z_Y)$

- $\hat{b} = \mu_Y - \hat{a}\ \mu_X$

Correlation

- The expected product of the deviations of X and Y, $E(D_X D_Y)$ is called the ***covariance*** of X and Y.
- The problem with using covariance is that the units are multiplied *and* the value depends on the units
- Can get rid of this problem by dividing each deviation by the SD of the corresponding SD, that is, put it in standard units. The resulting quantity is called the ***correlation coefficient*** of X and Y:
- $r(X,Y) =$
- Note that it is a pure number with no units, and now we will prove that it is always between -1 and 1.

Bounds on correlation

- $r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X Z_Y)$
- (Note that this implies that $E(D_X D_Y) = r\sigma_X\sigma_Y$. We will use this later.)

Errors in regression

- The error in regression $D = Y - \hat{Y}$

- What is $E(D)$? $Var(D)$?

- Note that we made no assumptions on the distributions of X & Y. This means that the residuals average to 0, *no matter what the joint distribution of* X & Y.

- What does the expectation of the error being 0 imply for the residuals?

Correlation as a measure of linear association

- $D = Y - \hat{Y},\ E(D) = 0,\ Var(D) = (1 - r^2)\sigma_Y^2$
- What if the correlation is very close to 1 or -1? What does this tell you about X & Y?
- What about if the correlation is close to 0? What does this tell you about X & Y?

Residual is uncorrelated with X

- What about $r(D, X), D = Y - \hat{Y}$?
- Intuitively, what should this be? Why?
- What should your residual (diagnostic) plot look like?

The Simple Linear Regression Model

- Regression model from data 8
- Model has two variables: response (Y) & (x) predictor/covariate/feature variable