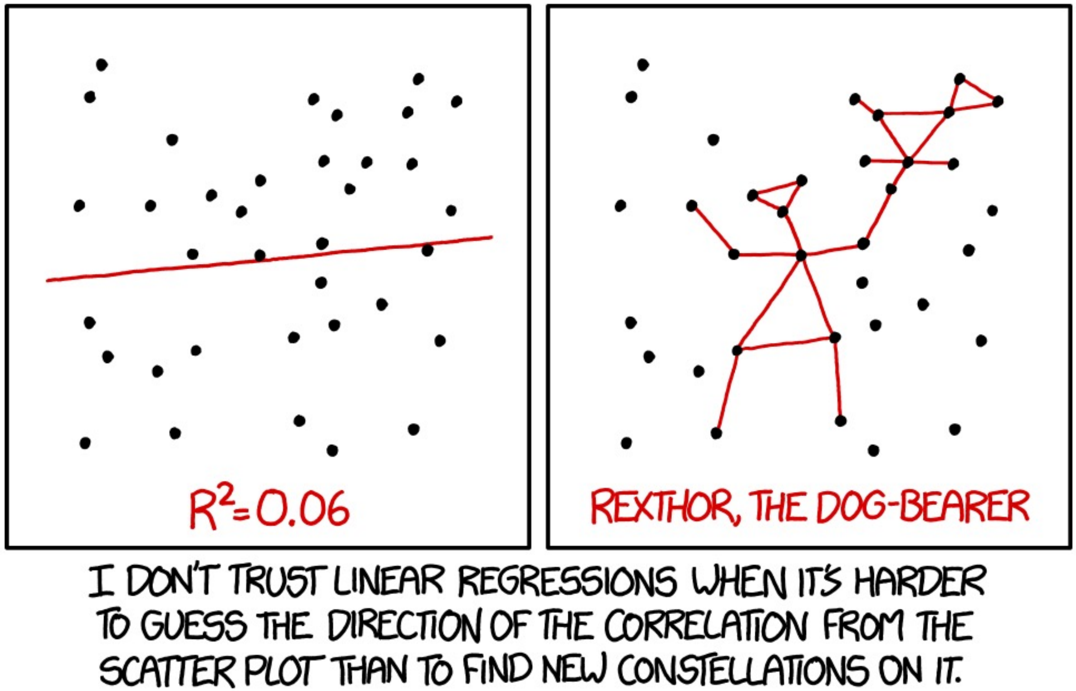
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 38 : 4/26/2021

Chapter 11

Correlation

https://xkcd.com/1725/

Equation of the regression line

$Y = aX + b + \text{error}$

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

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

X^* Y^* ← Text notation

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

$Z_X = \frac{X - \mu_X}{\sigma_X}$

$r(X, Y)$

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

$Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$

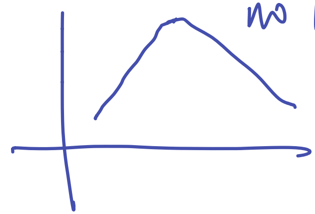
$D = Y - \hat{Y}$

Correlation

$r = E(Z_X \cdot Z_Y)$

- 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:
- $r(X, Y) =$ $E\left(\frac{D_X D_Y}{\sigma_X \sigma_Y}\right) \Rightarrow E(D_X D_Y) = \text{cov}(X,Y) = r\sigma_X \sigma_Y$
- Note that it is a pure number with no units, and now we will prove that it is always between -1 and 1.

$$E\left(\underline{(X - \mu_X)(Y - \mu_Y)}\right)$$



no linear relationship

Bounds on correlation : $-1 \le r \le 1$

$$Z_X = \frac{X-\mu_X}{\sigma_X}$$

$$E(Z_X^2) = \frac{E(X-\mu_X)^2}{\sigma_X^2}$$

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

$$0 \le E[(Z_X + Z_Y)^2] = E[Z_X^2 + 2Z_XZ_Y + Z_Y^2]$$

$$= \underbrace{E(Z_X^2)}_{=1} + 2E(Z_XZ_Y) + \underbrace{E(Z_Y^2)}_{=1}$$

$$0 \le 1 + 2\underbrace{E(Z_XZ_Y)}_{r} + 1$$

$$0 \le 2 + 2r$$

$$\Rightarrow \quad r \ge -1 \quad ☆$$

$$0 \le E[(Z_X - Z_Y)^2]$$

$$E(Z_X^2 + Z_Y^2 - 2Z_XZ_Y)$$

$$= E(Z_X^2) + E(Z_Y^2) - 2E(Z_XZ_Y)$$

$$= 1 + 1 - 2r$$

$$0 \le 2 - 2r$$

$$2r \le 2$$

$$r \le 1 \quad ☆☆$$

☆ & ☆☆ imply

$$-1 \le r \le 1$$

$$D_X = X - \mu_X$$

$$D_Y = Y - \mu_Y$$

$$Z_X = \frac{D_X}{\sigma_X} = \frac{X-\mu_X}{\sigma_X}$$

$$E(D_X^2) = E\left((X-\mu_X)^2\right) = Var(X) = \sigma_X^2$$

$$E(Z_X^2) = E\left[\left(\frac{D_X}{\sigma_X}\right)^2\right] = \frac{1}{\sigma_X^2} E(D_X^2) = \frac{1}{\sigma_X^2}\cdot \sigma_X^2 = 1$$

Exercise Suppose $Y = aX + b$, $a > 0$

$\mu_Y = E(Y) = a\mu_X + b$, $Var(Y) = a^2\sigma_X^2$

$\sigma_Y = a\sigma_X$

$$r(X,Y) = E(Z_X Z_Y)$$

$$= E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\underset{a\sigma_X}{\sigma_Y}}\right)\right]$$

$Y = aX+b$

$\mu_Y = a\mu_X + b$

$Y - \mu_Y = aX + \cancel{b} - (a\mu_X + \cancel{b})$

$= aX - a\mu_X = a(X-\mu_X)$

$$= \frac{1}{\sigma_X \cdot a \cdot \sigma_X} E\left[(X-\mu_X)(aX + \cancel{b} - a\mu_X - \cancel{b})\right]$$

$$= \frac{1}{a\sigma_X^2} E\left[a\cdot(X-\mu_X)^2\right] = \frac{a\sigma_X^2}{a\cdot\sigma_X^2} = 1$$

Exercise ① Show that if $Y = aX+b$, $a<0$

$r(X,Y) = -1$

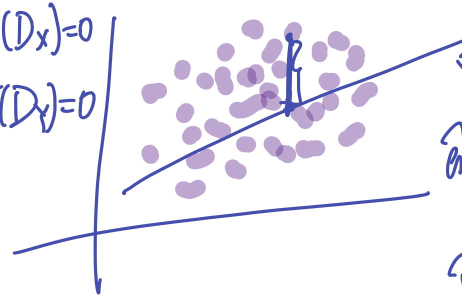
② $r(aX+b, cY+d) = \{r(X,Y),\ ac>0$

$\} -r(X,Y),\ ac < 0$

Errors in regression

$D_X = X - \mu_X,\ E(D_X) = 0$

$D_Y = Y - \mu_Y,\ E(D_Y) = 0$



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

$\hat{a} = r\frac{\sigma_Y}{\sigma_X}$

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

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

$$E(D) = E(Y - \hat{Y})$$

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

$$\hat{Y} = \hat{a}X + \hat{b} = \hat{a}X + \mu_Y - \hat{a}\mu_X = \hat{a}(X - \mu_X) + \mu_Y = \hat{a}D_X + \mu_Y$$

$$D = Y - \hat{Y} = Y - \hat{a}D_X - \mu_Y = D_Y - \hat{a}D_X$$

- 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?

$$E(D) = E(D_Y - \hat{a}D_X) = E(D_Y) - \hat{a}E(D_X) = 0 = \mu_D$$

$$Var(D) = E((D - \mu_D)^2) = E((D - 0)^2) = E(D^2)$$

$$Var(D) = \underbrace{E(D^2)}$$

$$E((Y - \hat{Y})^2) = \text{Mean squared error for regression}$$

$D^2 = $ MSE for regression $= \mathrm{Var}(D)$

$$D = D_Y - \hat{a} D_X$$

$$E(D^2) = E[(D_Y - \hat{a}D_X)^2] = E[D_Y^2 - 2\hat{a} D_X D_Y + \hat{a}^2 D_X^2]$$

$$= \underbrace{E[D_Y^2]}_{\sigma_Y^2} - 2\hat{a}\underbrace{E(D_X D_Y)}_{r\sigma_X\sigma_Y} + \hat{a}^2 \underbrace{D_X^2}_{\sigma_X^2} \qquad \hat{a} = \frac{r\sigma_Y}{\sigma_X}$$

$$E(D^2) = \sigma_Y^2 - 2r\frac{\sigma_Y}{\cancel{\sigma_X}} \cdot r\cancel{\sigma_X} \cdot \sigma_Y + \frac{r^2\sigma_Y^2}{\cancel{\sigma_X^2}} \cdot \cancel{\sigma_X^2}$$

$$= \sigma_Y^2 - 2r^2\sigma_Y^2 + r^2\sigma_Y^2 = \sigma_Y^2 - r^2\sigma_Y^2$$

$\mathrm{Var}(D) = E(D^2) = $ MSE for regression $= (1-r^2)\sigma_Y^2$

$$\boxed{SD(D) = \sqrt{1-r^2}\cdot\sigma_Y}$$

$D = Y - \hat{Y}$ = error in regression estimate is called the residual.

spread in residuals is smaller than spread in Y.

Correlation as a measure of linear association

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

We know $E(D) \simeq 0$ & if SD(D) close to 0

This tells you Y close to $\hat{Y} = \hat{a}X + \hat{b}$ $\Rightarrow$ Y is close to being a linear function of X.

(if $r = \pm 1$, then Y is exactly a linear function of X)

- What about if the correlation is close to 0? What does this tell you about X & Y?

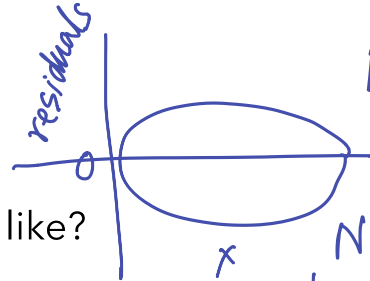
If $r \approx 0$, then $1 - r^2 \approx 1$

$SD(D) \approx SD(Y)$

then the linear relationship of X & Y is weak. May as well just use μ_Y to predict 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?



$E(D) = 0$

No linear trend.

$\hat{Y} = \hat{a}X + \hat{b}$ $\quad$ $Y - \hat{Y}$ removes the linear trend with X.

<u>Exercise</u>: Show that $r(D, X) = E\left[\frac{(D-0)}{\sigma_D} \frac{(X-\mu_X)}{\sigma_X}\right]$

$= 0$

$E(D \cdot D_X) = ?$

The Simple Linear Regression Model

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