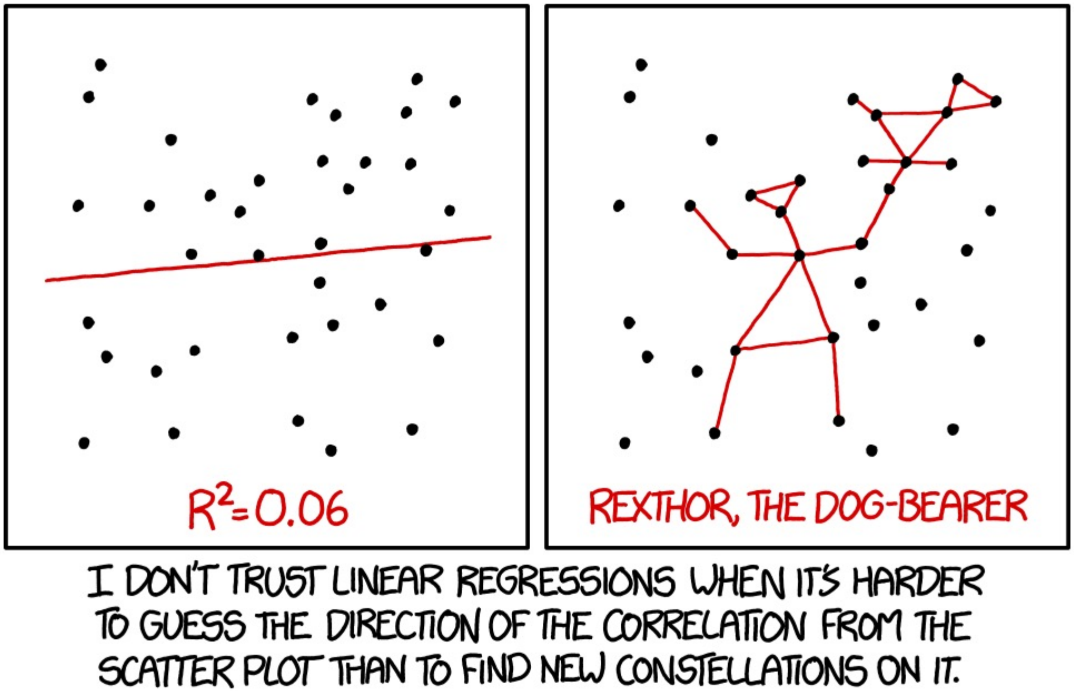
Stat 88: Probability & Mathematical Statistics in Data Science



https://xkcd.com/1725/

Lecture 27 : 4/28/2022

Finishing Chapter 11, and some of chapter 12

Correlation, Regression

Mathematical derivation of the formulas for a and b OPTIONAL

- As usual, $E(X) = \mu_X, SD(X) = \sigma_X; E(Y) = \mu_Y, SD(Y) = \sigma_Y$ r = correlation(X, Y)
- (X, Y) are our random variables, that we *think* are related by a linear function, perhaps with some error: $Y = aX + b + error$
- We want to estimate the equation of the line, that is, find $\hat{Y}$ such that $\hat{Y} = aX + b$

  $\hat{Y} = aX + b$ (a, b are unknown)

- Find the a and b by minimizing the mean square error, where error is the difference between our estimate $\hat{Y}$ and the original random variable Y.

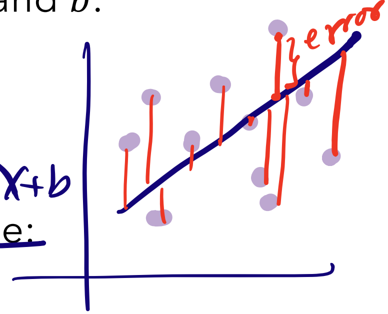
  Error = $Y - \hat{Y}$, SQUARED ERROR $= (Y - \hat{Y})^2$

  MEAN SQUARED ERROR $= E[(Y - \hat{Y})^2]$

- Notice that the mean squared error will be a function of a and b:

$$MSE(a, b) = E\left((Y - \hat{Y})^2\right) = E\left(\left(Y - \underline{(aX + b)}\right)^2\right)$$

observed value — value on regression line $\hat{Y} = aX + b$



- First, we can look for the best intercept for some fixed slope:

  Slope = a, Intercept = b.

Mathematical derivation of the formulas for a and b OPTIONAL

- Looking for the best intercept for some fixed slope, that is, fix a, and then see, for this *given* value of a, what would be the b that minimizes the MSE?

$$MSE(a,b) = \mathbb{E}[(Y-\hat{Y})^2] = \mathbb{E}[(Y-(aX+b))^2]$$

- We can write out the MSE as a function of b, take the derivative, and set it equal to 0, and look for the best b.

Fix a, $MSE_a(b)$ as a function of b for this fixed a.

$$MSE(b) = \mathbb{E}[(Y-(aX+b))^2] = \mathbb{E}[(\underbrace{Y-aX}-\underbrace{b})^2]$$

$$= \mathbb{E}[(Y-aX)^2 - 2(Y-aX)b + b^2]$$

$$= \mathbb{E}[(Y-aX)^2] - 2b\,\mathbb{E}(Y-aX) + b^2$$

[a, b not random variable]

Differentiate with respect to b (treat a as a constant)

$$\frac{d(MSE(b))}{db} = -2\mathbb{E}(Y-aX)+2b \overset{\text{set}=0}{=} 0$$

$$\hat{b} = \mathbb{E}(Y-aX) = \mu_Y - a\mu_X$$

so, for fixed a, "best" $b = \hat{b} = \boxed{\mu_Y - a\mu_X}$

<u>best slope</u> plug in $\hat{b} = \mu_Y - a\mu_X$

$$MSE(a) = \mathbb{E}\left[(Y-(aX+\hat{b}))^2\right]$$

$$= \mathbb{E}\left[(Y-aX-\hat{b})^2\right]$$

$$= \mathbb{E}\left[(Y-aX-\mu_Y+a\mu_X)^2\right]$$

$$= \mathbb{E}\left[(\underbrace{(Y-\mu_Y)}_{D_Y} - a\underbrace{(X-\mu_X)}_{D_X})^2\right]$$

Y-deviation from mean $\rightarrow D_Y$

$$MSE(a) = \mathbb{E}\left[(D_Y - aD_X)^2\right]$$

$$= \mathbb{E}\left[D_Y^2 - 2aD_XD_Y + a^2D_X^2\right]$$

$$= \underbrace{\mathbb{E}(D_Y^2)}_{Var(Y)} - 2a\,\mathbb{E}(D_XD_Y) + a^2\underbrace{\mathbb{E}(D_X)^2}_{Var(X) = \sigma_X^2}$$

$$\frac{d(MSE(a))}{da} = 0 - \cancel{2}\mathbb{E}(D_XD_Y) + \cancel{2}a\sigma_X^2 = 0$$

set $=0$ to solve for a

$$\mathbb{E}(D_X D_Y) = \hat{a}\sigma_X^2$$

$$\boxed{\hat{a} = \frac{\mathbb{E}[(Y-\mu_Y)(X-\mu_X)]}{\sigma_X^2}}$$

$$Z_X = \frac{X-\mu_X}{\sigma_X} \qquad Z_Y = \frac{Y-\mu_Y}{\sigma_Y}$$

covariance (X,Y)

$$\underbrace{\text{Correlation}(X,Y)}_{\boxed{r(X,Y)}} = \frac{\boxed{\mathbb{E}[\overbrace{(X-\mu_X)(Y-\mu_Y)}^{\mathbb{E}(D_X D_Y)}]}}{\sigma_X\sigma_Y} = \mathbb{E}(Z_X Z_Y)$$

So $\mathbb{E}(D_X D_Y) = \underbrace{r(X,Y)}_{=r} \cdot \sigma_X\sigma_Y$

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

$$\hat{a} = \frac{r\cdot\cancel{\sigma_X}\cdot\sigma_Y}{\sigma_X^{\cancel{2}}} = \frac{r\cdot\sigma_Y}{\sigma_X}$$

$\text{Cov}(X,X) = \text{Var}(X) = \mathbb{E}(D_X^2)$

$$\boxed{\hat{a} = r\frac{\sigma_Y}{\sigma_X}} \qquad \boxed{\hat{b} = \mu_Y - \hat{a}\mu_X}$$

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

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

reg. estimate

Equation of the regression line

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

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

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

$= E\left(\frac{D_X}{\sigma_X} \cdot \frac{D_Y}{\sigma_Y}\right) = E(Z_X Z_Y)$

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

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

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

Bounds on correlation

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

$Z_Y = \frac{D_Y}{\sigma_Y}$

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

$0 \le \mathbb{E}\left[(Z_X+Z_Y)^2\right] = \mathbb{E}\left[Z_X^2 + 2Z_XZ_Y + Z_Y^2\right]$

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

$= 1 + 2\underbrace{\mathbb{E}(Z_XZ_Y)}_{} + 1$

$= 1 + 2r + 1$

$0 \le 2 + 2r$

$-2 \le 2r$

$\boxed{r \ge -1}$ ☆

$Z_X = \frac{D_X}{\sigma_X}$

$Z_X^2 = \frac{D_X^2}{\sigma_X^2}$

$\mathbb{E}(Z_X^2) = \mathbb{E}\left(\frac{D_X^2}{\sigma_X^2}\right)$

$= 1$

$= \frac{\mathbb{E}(D_X^2)}{\sigma_X^2} = \frac{\sigma_X^2}{\sigma_X^2}$

$$0 \le (z_X - z_Y)^2$$

so $\quad 0 \le \mathbb{E}[(z_X - z_Y)^2]$

$$0 \le \mathbb{E}[z_X^2 - 2z_Xz_Y + z_Y^2] = \underbrace{\mathbb{E}(z_X^2)}_{1} - 2\underbrace{\mathbb{E}(z_Xz_Y)}_{r} + \underbrace{\mathbb{E}(z_Y^2)}_{=1}$$

$$0 \le 1 - 2r + 1 = 2 - 2r$$

$$\Rightarrow \boxed{r \le 1} \quad ✿$$

$$★ + ✿ \longrightarrow \boxed{-1 \le r \le 1}$$

<u>Exercise</u>: Let $Y = aX + b$, $a < 0$

1. Show that $r = -1$

2. $r(aX+b, cY+d)$?? $(ac > 0)$

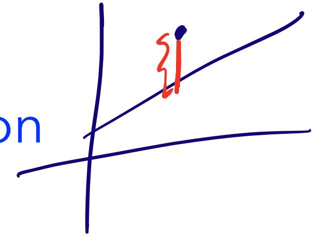
3. Let $D = Y - \hat{Y}$

$$\mathbb{E}(\hat{Y}) = \mathbb{E}(\hat{a}D_X + \mu_Y)$$
$$= \mathbb{E}(\hat{a}D_X) + \mu_Y$$
$$= \hat{a}\underbrace{\cancel{\mathbb{E}(D_X)}}_{0} + \mu_Y = \mu_Y$$

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

$$\mathbb{E}(Y - \hat{Y}) = \mathbb{E}(Y) - \mathbb{E}(\hat{Y})$$
$$= \mu_Y - \mu_Y = 0$$

<u>Exercise</u>. Show $\mathrm{Var}(D) = \mathbb{E}(D^2) = (1 - r^2)\sigma_Y^2$



Correlation as a measure of linear association

vertical error

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

If $r \approx \pm 1$ $Var(D) = (1-r^2)\sigma_Y^2 \approx 0$

This tells you that Y is very close to $\hat{Y}$ and Y is close to being a linear function of X (If $r = \pm 1$, 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?

$Var(D) = (1-r^2)\sigma_Y^2 \approx \sigma_Y^2$ (b/cos $r^2 \approx 0$)

In this case X gives no information about Y & may as well just use μ_Y to predict Y.

$D = [Y - (\hat{a}X + \hat{b})]$ residual

$D = Y - \hat{Y}$

= observed value − fitted value

Residual is uncorrelated with X

D is called the residual of the regression

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

Want $\mathrm{Cor}(D, X) = 0$ (once we subtract off the linear relationship with X what's left should not be correlated with X)

$$r(D,X) = \mathbb{E}(z_D \cdot z_X)$$

$$= \mathbb{E}\left(\left(\frac{D-0}{\sigma_D}\right) \cdot \left(\frac{X-\mu_X}{\sigma_X}\right)\right) = \frac{1}{\sigma_D \sigma_X} \mathbb{E}\Big(D \underbrace{(X-\mu_X)}_{D_X}\Big)$$

$$D = Y - \hat{Y} = Y - (\hat{a} D_X + \mu_Y) = (Y - \mu_Y) - \hat{a} D_X$$

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

$$r(D,X) = \frac{1}{\sigma_D \sigma_X} \mathbb{E}\left(\left(D_Y - \hat{a} D_X\right) \cdot D_X\right)$$

$$= \frac{1}{\sigma_D \sigma_X} \cdot \mathbb{E}\left(D_Y D_X - \hat{a} D_X^2\right)$$

$$\boxed{\begin{aligned}&\mathbb{E}(D_X D_Y) \\ &= r\sigma_Y \sigma_X\end{aligned}}$$

Covariance

$$= \frac{1}{\sigma_D \sigma_X} \left[\underbrace{\mathbb{E}(D_Y D_X)}_{r\sigma_Y\sigma_X} - \hat{a} \underbrace{\mathbb{E}(D_X^2)}_{\sigma_X^2} \right]$$

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

$$r(D,X) = \frac{1}{\sigma_D \sigma_X} \left[r\sigma_Y\sigma_X - r\frac{\sigma_Y}{\cancel{\sigma_X}} \cdot \sigma_X^{\cancel{2}} \right] \to 0$$

$$= 0$$

The Simple Linear Regression Model

- Regression model from data 8
- Model has two variables: response (Y) & (x) predictor/covariate/feature variable
- **Assumptions**: response is a linear function of the predictor (signal) + random error (noise), where the noise has a **normal** distribution, centered at 0. The signal is not random, but the response is, because the noise is random:

$Y = \beta_0 + \beta_1 x + \text{error}$

response = signal + noise

Noise $\sim N(0, \sigma^2)$
= error

- In mathematical language:

$i = 1, 2, \ldots n$

$(x_1, Y_1), (x_2, Y_2) \ldots (x_n, Y_n)$

$E(Y_i) = E(\beta_0 + \beta_1 x_i + \underset{\varepsilon_i}{\text{error}})$

$\boxed{E(Y_i) = \beta_0 + \beta_1 x_i}$

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Y : response
x : signal
error : noise.
β_0, β_1 : intercept, slope of regression

Y_i : response

Model we assume $\quad Y_i = f(x_i) + \underbrace{\text{noise}}_{\varepsilon_i}$

$\hookrightarrow Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2)$

Fitted values: $\hat{Y}_i = \beta_0 + \beta_1 x_i \leftarrow$ line

Unknown parameters : $\beta_0, \beta_1, \sigma^2$

Y_i is considered a r.v.

x_i is not

(Y response is a r.v.
ε noise is a r.v.
x signal is not)

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$$

$$\mathbb{E}(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$$

$$\text{Var}(Y_i) = \text{Var}(\underbrace{\beta_0 + \beta_1 x_i}_{\text{no variance}} + \underbrace{\varepsilon_i}_{\sigma^2})$$

$$= \sigma^2$$

$$\text{Var}(\bar{Y}) = \sigma^2/n$$

The regression line

- For each i, we want to get as close as we can to the *signal* $\boxed{\beta_0 + \beta_1 x_i}$
- There is some "true" regression line $\beta_0 + \beta_1 x$ that we cannot observe since there is noise. We estimate this line by minimizing the squared observed error.
- Estimate of the line given the data is $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope, respectively, given the data.
- We will investigate the distribution of the slope estimate (why is it random?) after looking at the individual and average response.

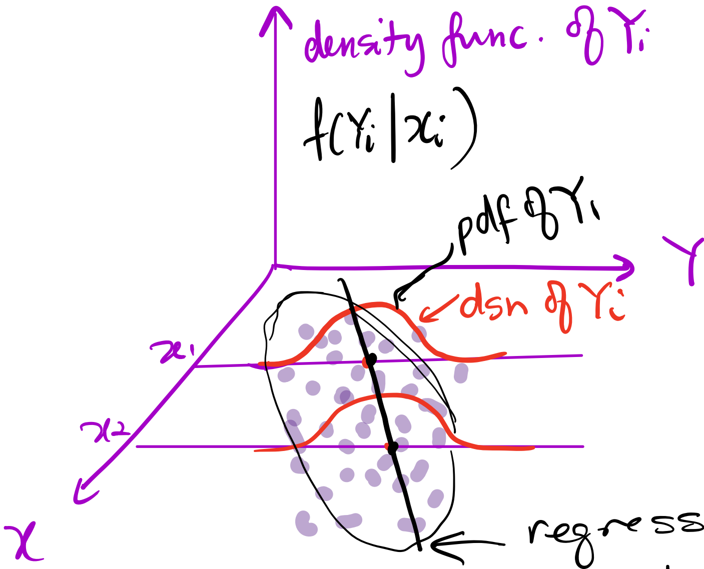
$(x_i, Y_i) \leftarrow$ Y_i is random by assumption

$(x_i, y_i) \leftarrow$ observed values

$\varepsilon_i \sim (\text{iid})\, N(0, \sigma^2)$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{\beta}_0, \hat{\beta}_1$ are r.v bcos they are estimated from the sample.



$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$\sigma^2 = Var(\varepsilon_i)$

regression line goes through the means of each of dsns of $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{} + \underbrace{\varepsilon_i}_{}$$

The individual response Y_i and the average response $\bar{Y}$

ε_i iid

- For any fixed i, Y_i is the sum of the signal and the noise.

- The signal is not random, but the noise is random with $\epsilon_i \sim N(0, \sigma^2)$
- Therefore what is the distribution of the Y_i ? $\sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

- What can you say about the independence and distribution of each of the Y_i ? Are they iid? Y_i NOT identically distributed but yes, independent

- Let $\bar{Y}$ be the average response. What would be its distribution?
- $E(\bar{Y}) =$ $\beta_0 + \beta_1 \bar{x}$

- $Var(\bar{Y}) =$ $\frac{\sigma^2}{n}$

The individual response Y_i and the average response $\bar{Y}$

- Y_i are normal with expectation $\beta_0 + \beta_1 x_i$ and variance σ^2
- Note that the individual responses are independent of each other.

- Let $\bar{Y}$ be the average response.
- $E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ (the expected average response is the *signal* at the average value of the predictor variable)

- $Var(\bar{Y}) = \frac{\sigma^2}{n}$ (only involves the error variance since the randomness in the Y_i's comes only from the errors or noise)

- Since $\bar{Y}$ is a linear combination of independent normally distributed random variables, it is also normal.