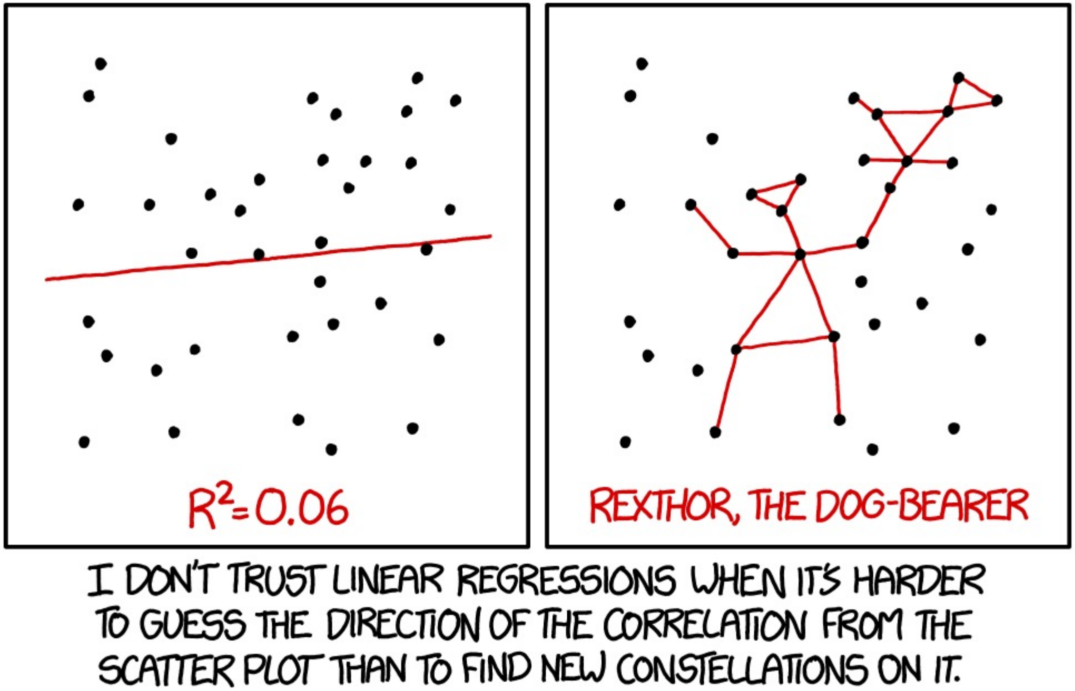
Stat 88: Probability & Mathematical Statistics in Data Science

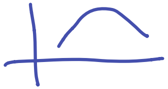

R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Lecture 37 : 4/23/2021

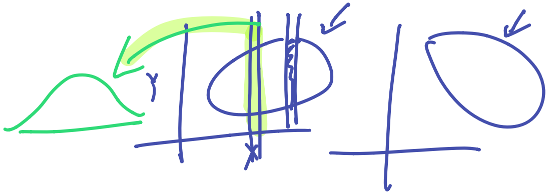
Chapter 11

Least Squares

https://xkcd.com/1725/



Simple Linear Regression



- One of the most used statistical techniques, used for summarizing a scatterplot, and sometimes drawing inferences about the data
- We will be revisiting simple linear regression, but using random variable notation
- Basically, given a random variable pair (X, Y), we want a model that describes the relationship between the predictor (X) and response (Y) variables. That is, can we express the relationship mathematically? Perhaps as

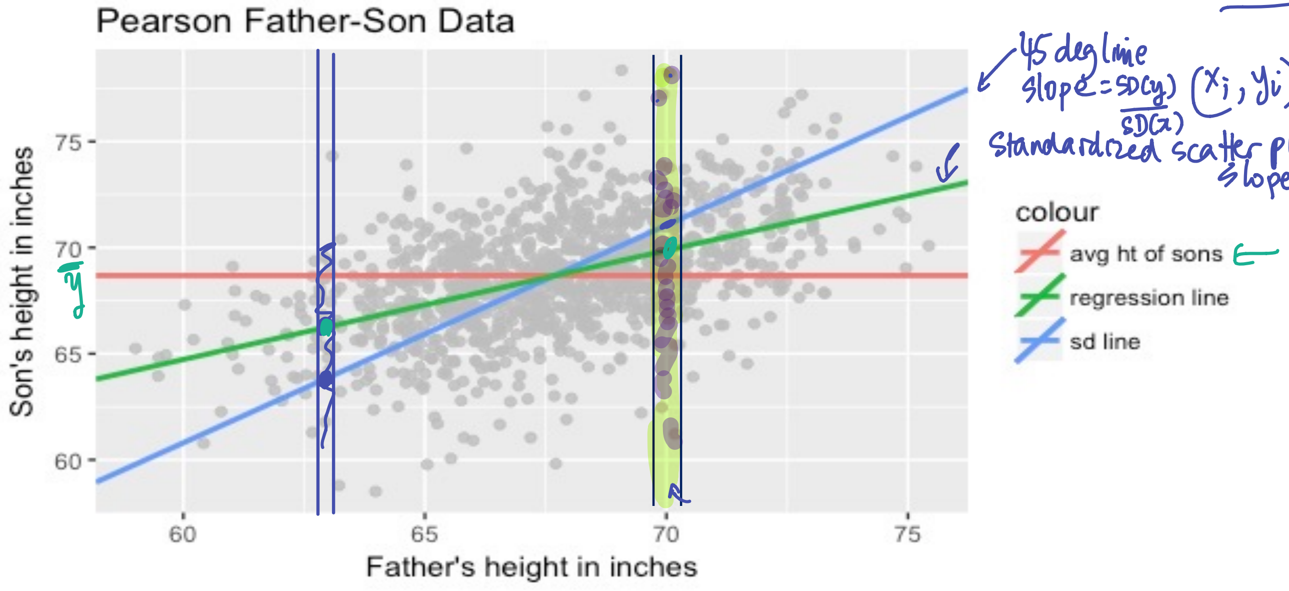
$$Y = f(X) \text{ or } Y = f(X) + random\ error$$

- Want to use a linear function of X to estimate Y, say $\overset{\downarrow}{a}X + \overset{\downarrow}{b}$
- "Best" line for these data.

$$Y = \overset{\downarrow}{a}X + \overset{\downarrow}{b} + error$$

Father-son data

Karl Pearson Francis Galton
MacTutor



45 deg line
slope $= \frac{SD(y)}{SD(x)}$ (x_i, y_i) $1 \le i \le n$
Standardized scatter plot (Data8)
slope blue line $= 1$

Want to predict y from x. Could use:

- Average of y (so don't use x at all)
- The SD (diagonal) line: better, but not so good (too steep)
- Much better, if the scatter plot shows a linear relationship, to use the **regression method**, which incorporates the correlation.

The regression method for data

- The regression method is used to draw the regression line which can be used for prediction.

$\hat{y}_i = a x_i + b$

error $= y_i - \hat{y}_i$

- It is also called the **least squares line** because it minimizes mean squared error. By *error* we mean the vertical difference between the y-value for some x, and the height of the regression line at that x.

$$e_i = y_i - (ax_i + b), i = 1, 2, \ldots, n$$

- From Data 8, do you recall the slope of the regression line? What about the intercept?

a

b

$$a = r \cdot \frac{SD(y)}{SD(x)}$$

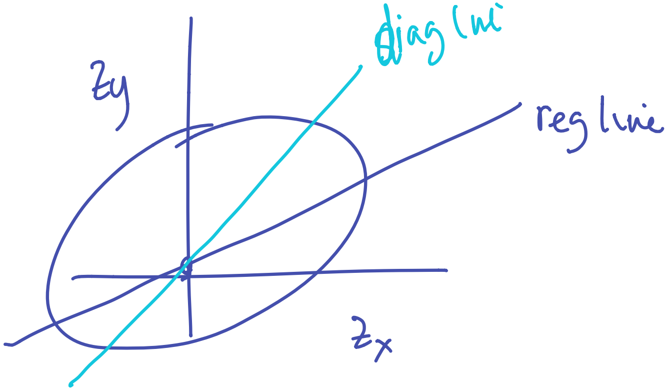
$$b = \bar{y} - a\bar{x}$$

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The regression line: least squares line

$\bar{x}$ } sample
$\bar{y}$ } avgs

- You found the slope of the regression line in Data 8 by:
  - using the geometry of the shape of the scatter plot (putting everything in standard units, and looking for the slope of the line that went through the centers of the vertical strips)



Slope of reg line = r

$$z_{\hat{y}} = r z_x$$

$$\boxed{\frac{\hat{y}-\bar{y}}{SD(y)} = r \cdot \frac{(x-\bar{x})}{SD(x)}}$$

$$\hat{y} = r \cdot \frac{SD(y)}{SD(x)} x + \bar{y} - \left(\frac{r SD(y)}{SD(x)}\right) \bar{x}$$

  - And also by minimizing (numerically) the ***mean squared error:***

  - The regression line is the *unique* straight line that minimizes the mean squared error of estimation among all straight lines, which is why it is called the “Least Squares” line

SD of sample data

$$\hat{y} = r \cdot \frac{SD(y)}{SD(x)} x + \left(\bar{y} - \frac{r SD(y)}{SD(x)} \bar{x}\right)$$

Mathematical derivation of the formulas for a and b

- As usual: $E(X) = \mu_X, SD(X) = \sigma_X; E(Y) = \mu_Y, SD(Y) = \sigma_Y$
- (X, Y) are our random variables that we think are related by a linear function, with some error: $Y = aX + b + error$
- We want to estimate the equation of the line, that is, find $\hat{Y}$ such that $\hat{Y} = aX + b$
- We will do this by minimizing the mean squared error, where error is the difference between our estimate $\hat{Y}$ and the original random variable Y: $error = Y - \hat{Y}$
- Note that the mean squared error will be a function of a and b:

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

- First, we can look for the best intercept for some fixed slope. That is, given a, what would be the b that minimizes the MSE?
- Let's write it out as a function of b, take the derivative and set it to 0

$$MSE_a(b) = E\left[\left(Y - (aX+b)\right)^2\right] = E\left[\left((Y-aX) - b\right)^2\right]$$

$$= E\left[(Y-aX)^2 - 2b(Y-aX) + b^2\right]$$

$$= E\left[(Y-aX)^2\right] - 2b\,E(Y-aX) + b^2$$

$$\frac{d MSE_a(b)}{db} = -2E(Y-aX) + 2b = 0$$

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

Best slope

First, for any fixed slope, plug in the corresponding $\hat{b}_a$ intercept so that everything is in terms of a.

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

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

$$= E\left[((Y-\mu_Y) - a(X-\mu_X))^2\right]$$

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

σ_Y^2

$$\begin{aligned} MSE(a) &= E\left[D_Y^2 - 2aD_YD_X + a^2D_X^2 \right] \\ &= E(D_Y^2) - 2aE(D_YD_X) + a^2E(D_X^2) \\ &= \sigma_Y^2 - 2aE(D_YD_X) + a^2\sigma_X^2 \\ &= Var(Y) - 2a\,E(D_YD_X) + a^2 Var(X) \end{aligned}$$

$$\frac{\overset{\downarrow}{d}}{da} MSE(a) = -2E(\overset{\downarrow}{D}_Y D_X) + 2a\sigma_X^2 = 0$$

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

This has r (the correlation) in there:

$$r = r(X,Y) = E\left[\frac{(\overset{D_X}{X-\mu_X})(\overset{D_Y}{Y-\mu_Y})}{\sigma_X \sigma_Y} \right]$$

$$E\left[D_XD_Y\right] = r\sigma_X\sigma_Y$$

$$\hat{a} = \frac{E(D_XD_Y)}{\sigma_X^2} = \frac{r\cancel{\sigma_X}\sigma_Y}{\sigma_X^{\cancel{2}}} = r\frac{\sigma_Y}{\sigma_X}$$

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

regression estimate

$$\hat{Y} = \hat{a} X + \hat{b} \quad , \quad \hat{a} = r\frac{\sigma_Y}{\sigma_X} \quad , \quad \hat{b} = \mu_Y - \hat{a}\mu_X$$

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

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

Correlation

$$Cov(X,Y) = E(D_X D_Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

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

expectation of product of Z-scores of X, Y

Show that $-1 \leq r \leq 1$