

Lecture 10: Prediction Based on Regression Line

Ideal situation:

$(x_i, y_i), i = 1, \dots, n$ — data. Plot shows football-shape.

Five summary statistics: \bar{x} , SD_x , \bar{y} , SD_y and r

$$r = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{SD_x} \frac{y_i - \bar{y}}{SD_y}$$

$$SD_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad SD_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Three lines:

1. SD line: center of the football

- passes (\bar{x}, \bar{y})
- slope $\frac{r}{|r|} \frac{SD_y}{SD_x}$
- line equation: $\frac{y - \bar{y}}{x - \bar{x}} = \frac{r}{|r|} \frac{SD_y}{SD_x}$ or $y = \bar{y} + \frac{r}{|r|} SD_y \frac{x - \bar{x}}{SD_x}$

2. Regression line to predict y from x

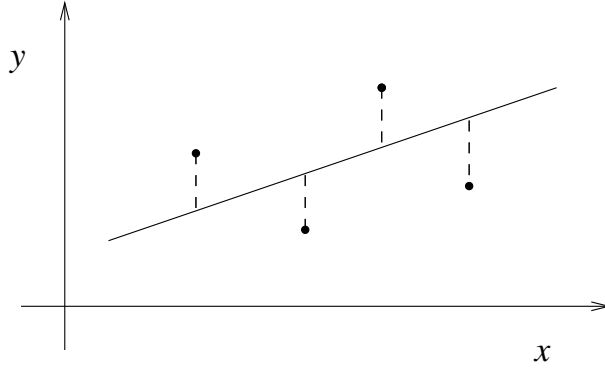
- passes (\bar{x}, \bar{y})
- slope $r \frac{SD_y}{SD_x}$
- equation: $\frac{y - \bar{y}}{x - \bar{x}} = r \frac{SD_y}{SD_x}$ or $y = \bar{y} + r SD_y \frac{x - \bar{x}}{SD_x}$

It is the least squares (LS) line which minimizes the vertical distance² of the data points to any line.

That is, $\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ is minimized when

$$\alpha = \bar{y} - r \frac{SD_y}{SD_x}$$

$$\beta = r \frac{SD_y}{SD_x}$$



Robust regression line is obtained if we minimize $\sum_{i=1}^n |y_i - \alpha - \beta x_i|$ instead.

Proof: Let $f(\alpha, \beta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ quadratic, so unique minimum achieved by the root of $\frac{\partial f}{\partial \alpha} = 0$, $\frac{\partial f}{\partial \beta} = 0$.

$$\begin{cases} \frac{\partial f}{\partial \alpha} = -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial f}{\partial \beta} = -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) x_i = 0 \end{cases}$$

leads to

$$\bar{y} = \alpha + \beta \bar{x} \tag{1}$$

$$\frac{1}{n} \sum_{i=1}^n x_i y_i = \alpha \bar{x} + \beta \frac{1}{n} \sum_{i=1}^n x_i^2 \tag{2}$$

Plug (1) into (2):

$$\frac{1}{n} \sum_{i=1}^n x_i y_i = (\bar{y} - \beta \bar{x}) \bar{x} + \beta \frac{1}{n} \sum_{i=1}^n x_i^2$$

$$\begin{aligned} \text{so } \beta &= \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\text{SD}_x^2} \\ &= \frac{\text{SD}_y \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\text{SD}_x \text{SD}_x \text{SD}_y} \\ &= r \frac{\text{SD}_y}{\text{SD}_x} \end{aligned}$$

(This explains using n in SD_x^2).

$$\alpha = \bar{y} - r \frac{\text{SD}_y}{\text{SD}_x} \bar{x}$$

3. Regression line to predict x from y

- passes (\bar{x}, \bar{y})
- slope $r \frac{SD_x}{SD_y}$

It is the least squares (LS) line which minimizes the horizontal distance² of the data points to any line.

That is, $\sum_{i=1}^n$

$(x_i - a - by_i)^2$ is minimized when

$$a = \bar{x} - rSD_x \frac{SD_y}{SD_x}, \quad b = r \frac{SD_x}{SD_y}$$

Example: IQ scores at 18 and 35.

$$\begin{aligned} 18: & \bar{x} = 100, \quad SD_x = 15, \\ 35: & \bar{y} = 100, \quad SD_y = 15, \quad r = 0.8 \end{aligned}$$

i.e. X - IQ at 18, y - IQ at 35. Estimate the average score at age 35 for those who scored 115 at age 18. $x_0 = 115$, $y_0 = ?$

- $\frac{x_0 - \bar{x}}{SD_x} = \frac{115 - 100}{15} = 1$ std unit above average
- $\times r$ for regression effect. 0.8 std units above y average.
- convert to units of y :

$$y_0 = 100 + (0.8 \times 15) = 112 < 115$$

So people get dumber when they get older?

Regression Effect

In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test — and the top group will on average fall back.

Regression Fallacy

Thinking that the regression effect may be due to something important, not just the spread around the line.

The RMS error in Regression

The regression predictions differ from actual values. By how much?

Intuition: the closer $|r|$ to 1, the smaller the error.

$$\text{(Root Mean Square) RMS} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$\hat{y}_i = \alpha + \beta x_i$ — regression line or predicted value at x_i , $y_i - \hat{y}_i$ — residual

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 &= \frac{1}{n} \sum_{i=1}^n \left(y_i - \bar{y} - r \frac{\text{SD}_y}{\text{SD}_x} (x_i - \bar{x}) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \bar{y}}{\text{SD}_y} - r \frac{x_i - \bar{x}}{\text{SD}_x} \right]^2 \text{SD}_y^2 \\ &= \left[\frac{1}{n} \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{\text{SD}_y^2} - 2r \frac{1}{n} \sum_{i=1}^n \frac{y_i - \bar{y}}{\text{SD}_y} \frac{x_i - \bar{x}}{\text{SD}_x} + \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\text{SD}_x^2} r^2 \right] \text{SD}_y^2 \\ &= [1 - 2r \cdot r + r^2] \text{SD}_y^2 \\ &= (1 - r^2) \text{SD}_y^2 \end{aligned}$$

$\implies \text{RMS} = \sqrt{1 - r^2} \text{SD}_y$ — average error measure for regression prediction.

When the scatterplot is football shaped the $(x_i, i = 1, \dots, n)$'s histogram is quite Normal and so is y 's.

Moreover, for each vertical strip the histogram is quite normal too. This normal is centered as the predicted value and with an $\text{SD} = \text{RMS} = \sqrt{1 - r^2} \text{SD}_y$.

How about the horizontal strips?

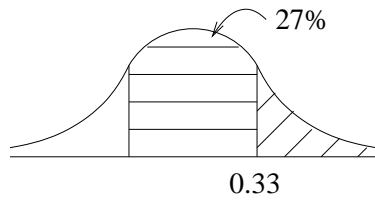
- center: $\bar{x} + r \frac{\text{SD}_x}{\text{SD}_y} (y - \bar{y})$
- $\text{SD} = \sqrt{1 - r^2} \text{SD}_x$

Example: IQ data.

If someone has an IQ score of 115 at age 18, what is the probability that the IQ score will be above 115 at age 35?

For vertical strip at $x_0 = 115$

- center $y_0 = 112$
- $SD = \sqrt{1 - 0.8^2} SD_y = 0.6 \times 15 = 9$



$$\frac{115 - 112}{9} = \frac{3}{9} = 0.33$$

$$\text{prob} = \frac{1 - 0.27}{2} = 0.365 < 0.5$$