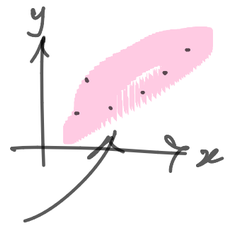
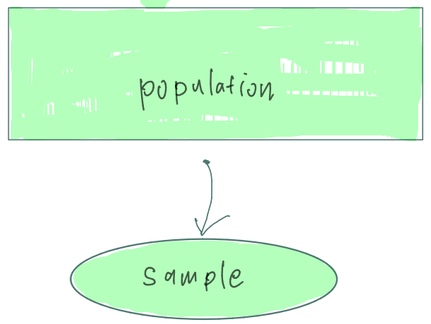
Bivariate data : (x, y)

independent (explanatory) ← x ; y → dependent variable (regressor)



Correlation Coefficient (r) measures linear association = how tightly points are clustered around a line

Module 2: Modelling Data



T4 : Normal Model

T5 : Linear Model

LO5: Model the relationship between 2 (quantitative) variables using linear regression.

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

$= r_{sample}$

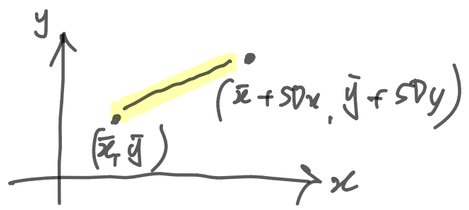
$-1 \le r \le 1$

classic mistakes with r

eg $r = 0 \neq$ no association

(4 numerical summaries)

SD Line



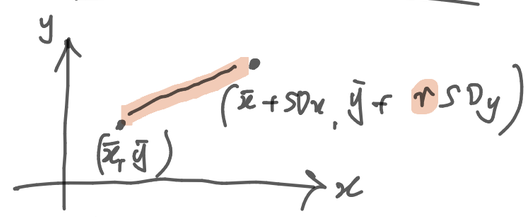
$\hat{y} = a + bx$

$b = \frac{SD_y}{SD_x}$

(5 numerical summaries)

$\bar{x}\ \bar{y}\ SD_x\ SD_y\ r$

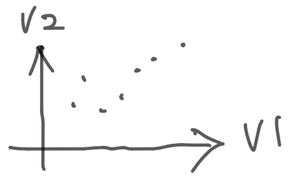
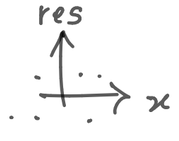
Regression Line



$\hat{y} = a + bx$

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

(for $r \geq 0$)

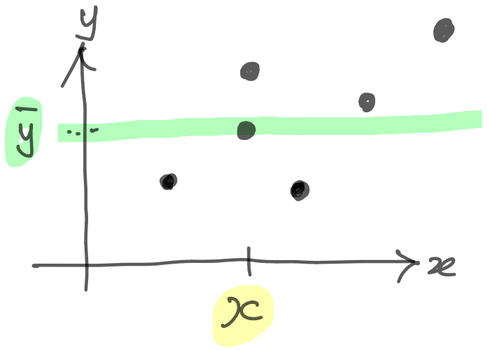
Model	Variable	When?	What?	Diagnostics	Predictions
Linear	2 Quant	 looks linear	① linear regression line $\hat{y} = a + bx$ ② correlation coefficient $-1 \le r \le 1$	① Does the scatter plot look linear?  ② Does the residual plot look random? 	① Baseline ② Strip ③ Regression Line ④ Percentile Rules ⑤ Strips

Predictions based on the Linear Model

RMS Error = "SD of Line" = RMS of (gaps from the line)
$= \sqrt{1-r^2}\, SD_y$
$\Rightarrow SD_y$ (Base Line)

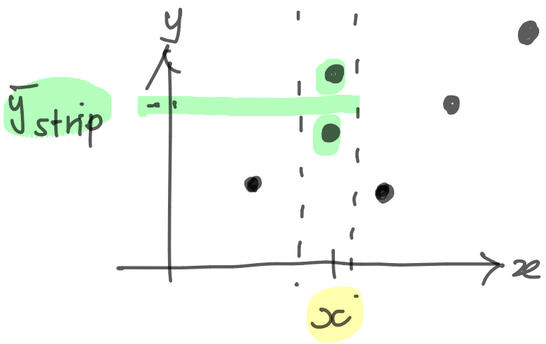
(DV15)

① Base Line



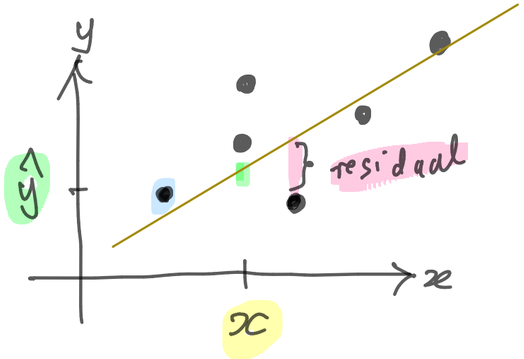
for any x,
we predict $\bar{y}$

② Strip



we take the average of the y_s in the x strip

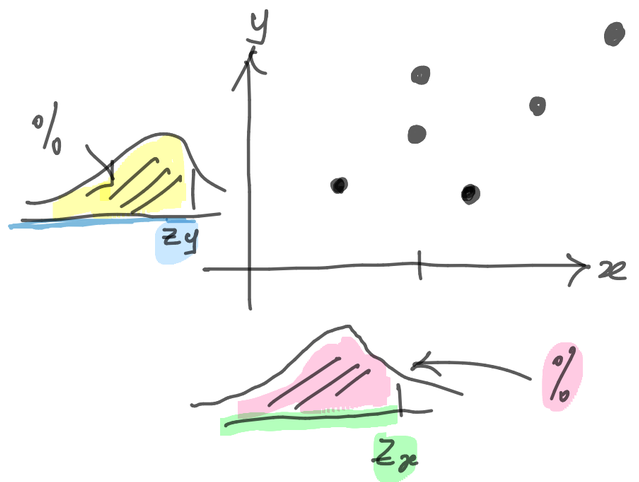
③ Regression Line



$\hat{y}$ = the point corresponding to $\bar{x}$ on the line

Residual Plot
random

④ Percentile Ranks



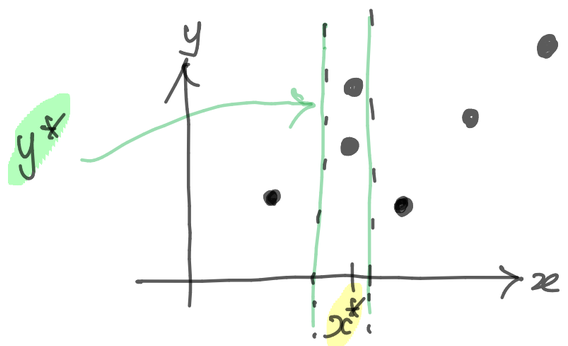
Summary: pnorm (r * qnorm (%))

Steps

1. Convert % in x direction to z score
   z_x = qnorm (%)
2. Calculate $z_y = r\, z_x$
3. Convert z score in y direction to %
   pnorm (z_y)

(DV16)

⑤ Strips (Homoscedastic)



$\bar{y}^* = \bar{y} + r z_{x^*} SD_y$

SD_{y^*} = RMS Error

$y^* \sim N(\bar{y}^*, SD_{y^*})$