Linear Regression

Contingency Table Tests allow us to explore association between two categorical variable:

$R_1 = \{Blue, Green, Red, male, Turkish\}$

Regression analysis also allow us to explore association between two variables but instead two numeric variables:

$Y = \{2.18, 4.28, 7.92 \dots\}$

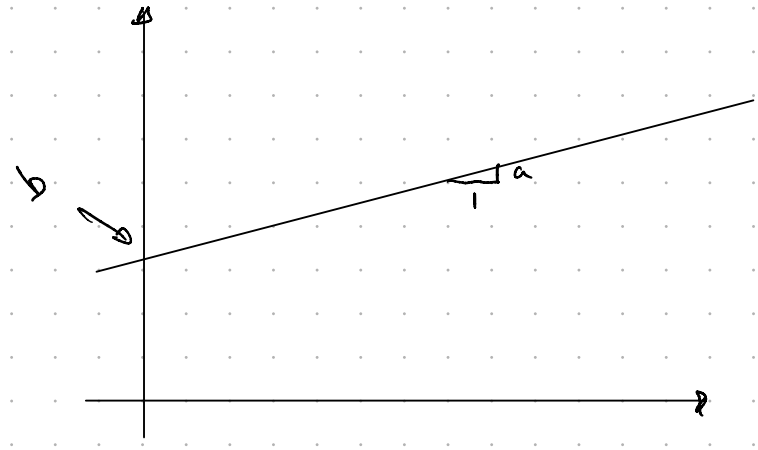
$X = \{1.1, 1.2, 1.4, \dots\}$

R.V. → Discrete: Contingency tables

R.V. → Continuous: Regression

$y = ax + b$

Dependent — Independent



$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots$

The association is called correlation between X and y:

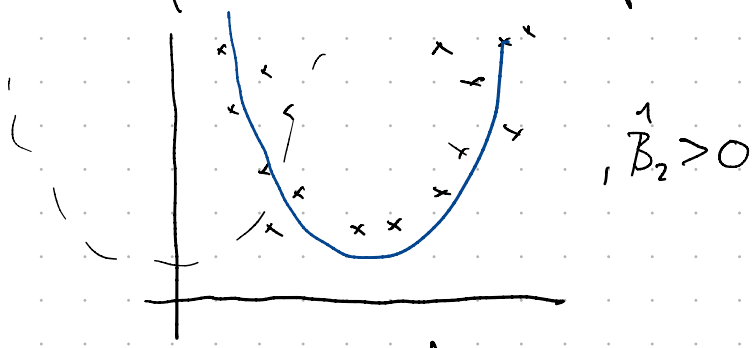
	x_1	x_2	x_3	x_4	x_5
P_1	16	87	122	.	.
P_2	18	72	:	'	'
P_3	20	94	.	'	'
P_4	15	31	'	'	

$X =$

$$X = \begin{bmatrix} 1 & 16 & 87 & 122 \\ 1 & 18 & 72 & \dots \\ 1 & 20 & 94 & \dots \\ 1 & 15 & 31 & \dots \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \end{bmatrix} \quad , \; X \cdot \beta$$

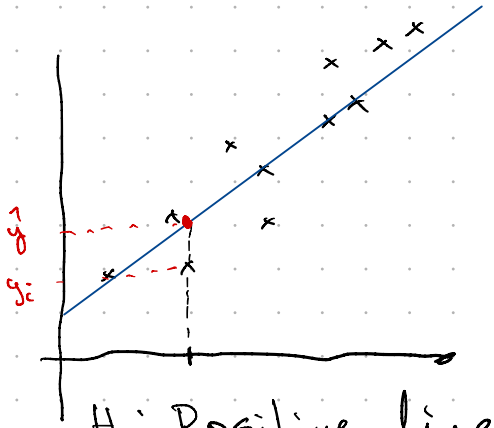
① Scatter Plot of X and Y to visually inspect relationship.



, $\hat{\beta}_2 > 0$

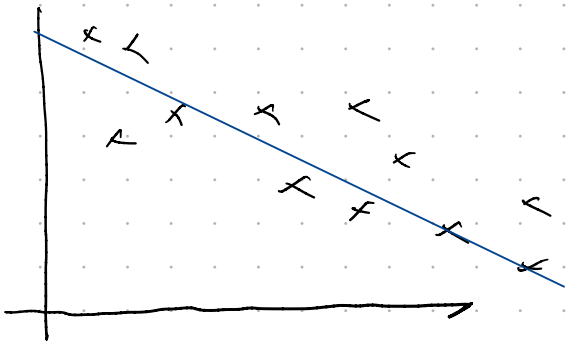
H_0: Quadratic

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1^2$



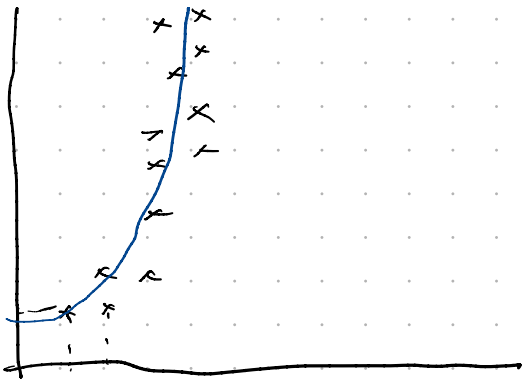
H_0: Positive linear

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X \; , \; \hat{\beta}_1 > 0$



H_0: Negative linear

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X \; , \; \hat{\beta}_1 < 0$



H_0: exponential

$\hat{y} = \hat{\beta}_0 \cdot \hat{\beta}_1^{X}$

2. Remove outliers. In regression this is done qualitatively

3. Determine regression equation, i.e. estimat β_0 and β_1:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\left.\begin{array}{l} \hat{\beta}_0 = ? \\ \hat{\beta}_1 = ? \end{array}\right\} \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$EY = \beta_0 + \beta_1 \cdot E[X] + E(\varepsilon)$$

$$= \beta_0 + \beta_1 \cdot E[X]$$

$$\beta_0 = EY - \beta_1 \cdot E[X]$$

$$Cov(X,Y) = (X,\ \beta_0 + \beta_1 X + \varepsilon)$$

$$= \beta_0 (cov(X,1) + \beta_1 (cov(X,X) + cov(X,\varepsilon)$$

$$= 0 + \beta_1 \cdot Cov(X,X) + 0$$

$$= \beta_1 \cdot Var(X)$$

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} \quad , \beta_0 = EY - \beta_1 E(X)$$

$$\bar{y} \qquad \bar{x}$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad , \quad Cov(X,Y) = \frac{1}{n-1} \cdot \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$\bar{y} = \frac{y_1 + y_2 - y_3 \cdots - y_n}{n} \qquad Var(x) = \frac{1}{n-1} \cdot \sum (x_i - \bar{x})^2$$

$$num = S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$den = S_{xx} = \sum (x_i - \bar{x})^2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$$

4. Check assumptions that errors are normally distributed (normal probability plot)

Residual (error)

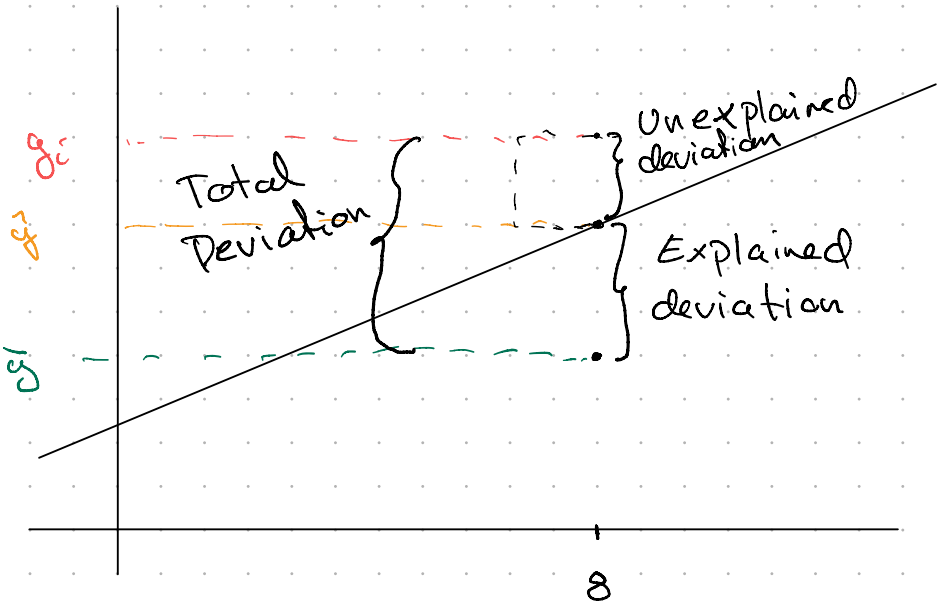
$$e_i = y_i - \hat{y}_i$$

$y_i = 9$ → observed

$\hat{y}_i = 7$ → predicted

$\bar{y} = 4$ → average

$x = 8$.



Total: $(y_i - \bar{y}_i)$

Unexp: $(y_i - \hat{y}_i)$

expl: $(\hat{y}_i - \bar{y}_i)$

Total sum of squares:

$$SS_T = \Sigma(y_i - \bar{y})^2$$

Regression sum of squares:

$$SS_R = \Sigma(\hat{y} - \bar{y})^2$$

Residual sum of squares:

$$SS_E = \Sigma(y_i - \hat{y})^2$$ ← Minimize this!

5. Assess adequacy of model:

a) $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Test static: $T_0 = \dfrac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$, $\hat{\sigma}^2 = \dfrac{SS_E}{n-2}$

b) Determine correlation:

1) $r = \dfrac{\Sigma(Z_x \cdot Z_y)}{n-1}$ Z scores of all x and y

2)

⋮

4) $r = \dfrac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}} = \dfrac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

$|r| > 0.6$: good

$|r| \leq 0.6$: High

$0.4 \leq |r| < 0.6$: OK

c) Find Correlation of Determination:

r^2 = square of r

$r^2 = \dfrac{SS_R}{SS_T}$ ← Expl / ← Total

} amount of variance my model is able to explain.

6) Find confidence intervals for β_0 and β_1