Linear Regression

Contingency Table Tests allow to explore association between two categorical variable

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

Regression analysis also allows us to explore association between two variables but instead two <u>numeric</u> variables:

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

$$X = \{1.1,\ 1.2,\ 1.4.\ 1.8 \ldots\}$$

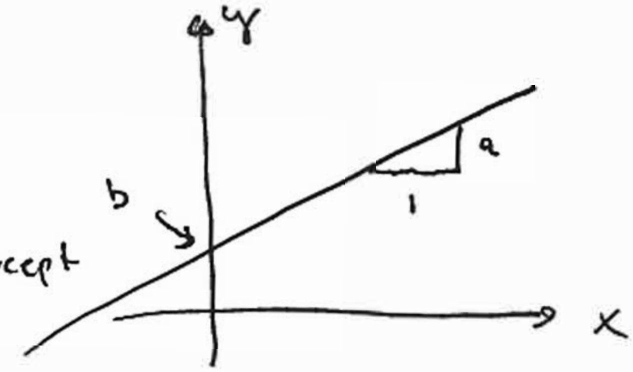
R.V.
- Discrete → Contingency tables
- continuous → Regression.

More specifically it is the type of the dependent variable that decides:

$y = ax + b$

Slope

Intercept



x is called the independent variable.

y is called the dependent or respons variable.

This association is called the <u>correllation</u> between X and Y.

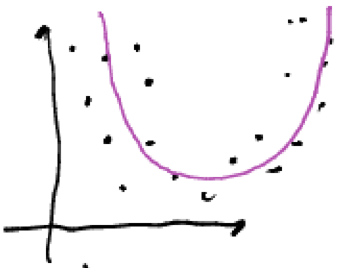
<u>Simple Linear Regression:</u>

One dependent and one independent.

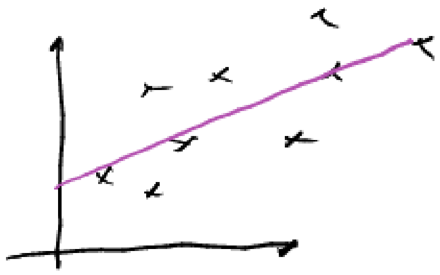
- It there relation between x and y?
  - How strong is this relation?

<u>A regression Analysis includes:</u>

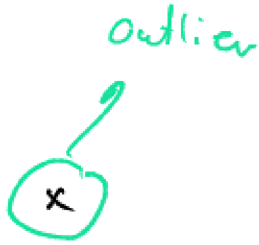
1) scatter plot of x and y to visually inspect relationship

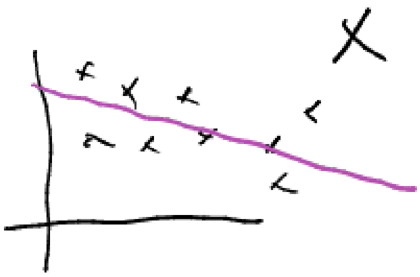


H_0: Quadratic

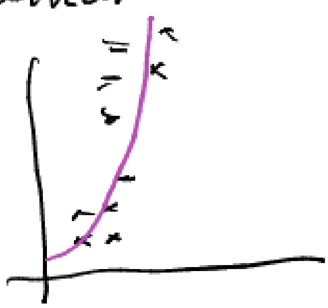


H_0: positive linear





H_0: Negative linear



H_0: exponential

2) Remove any outliers. In regression this is qualitative assessment.

Visual inspection.

3) Determine the regression equation, i.e. estimate β_0 and β_0:

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon} \quad (y = b + ax)$$

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

$$Ey = \beta_0 + \beta_1 EX + E[\varepsilon]$$

$$= \beta_0 + \beta_1 EX$$

$$\Downarrow$$

$$\beta_0 = EY - \beta_1 EX$$

Now we need $Cov(X, Y)$

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

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

$$= 0 + \beta_1 Var X + 0$$

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

$$\bar{X} = \frac{x_1 + x_2 + \dots x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \dots y_n}{n}$$

$$Cov(X,Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$Var(x) = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

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

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

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

Based on $Cov(x, y)$, $Var(x)$

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

At exam I will ask for S_{xy} and S_{xx}.

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

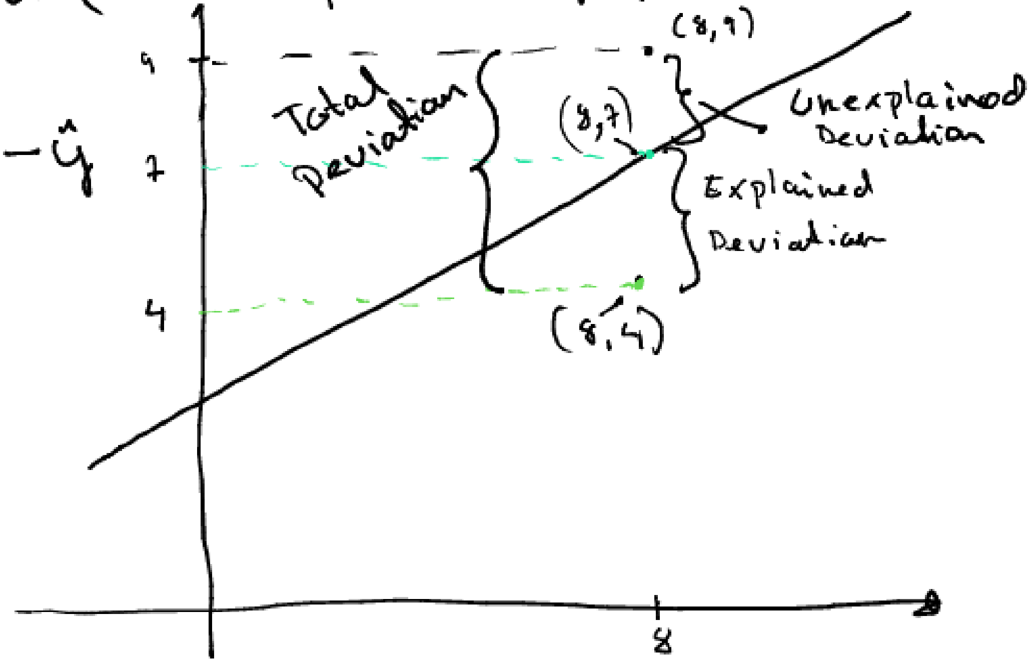
Residual:

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

$y_i = 9$ → Observed

$\hat{y} = 7$ → Predicted

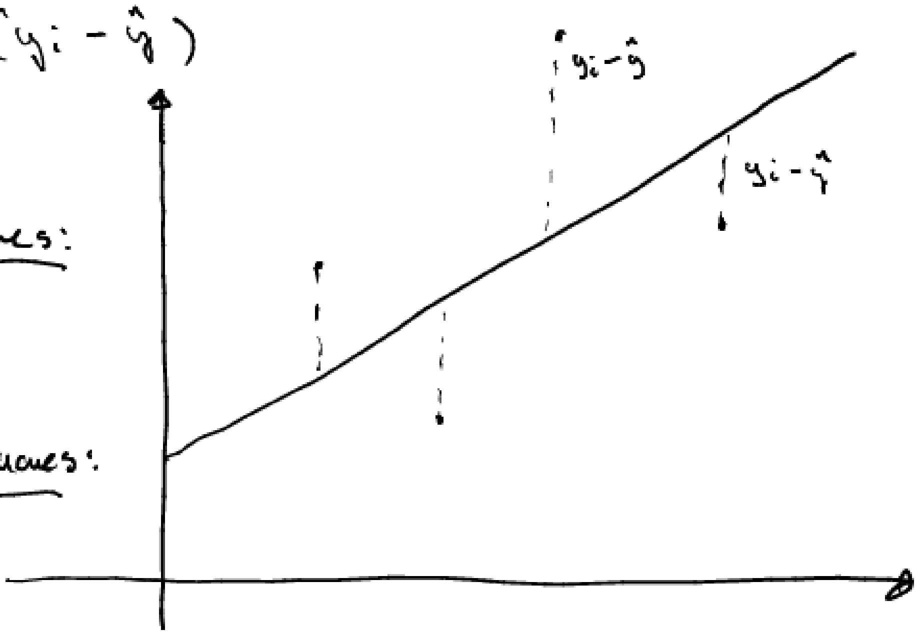
$\bar{y} = 4$ → Average



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

Explained : $(\hat{y} - \bar{y})$

Unexplained (Residual) : $(y_i - \hat{y})$



Total Sum of squares:

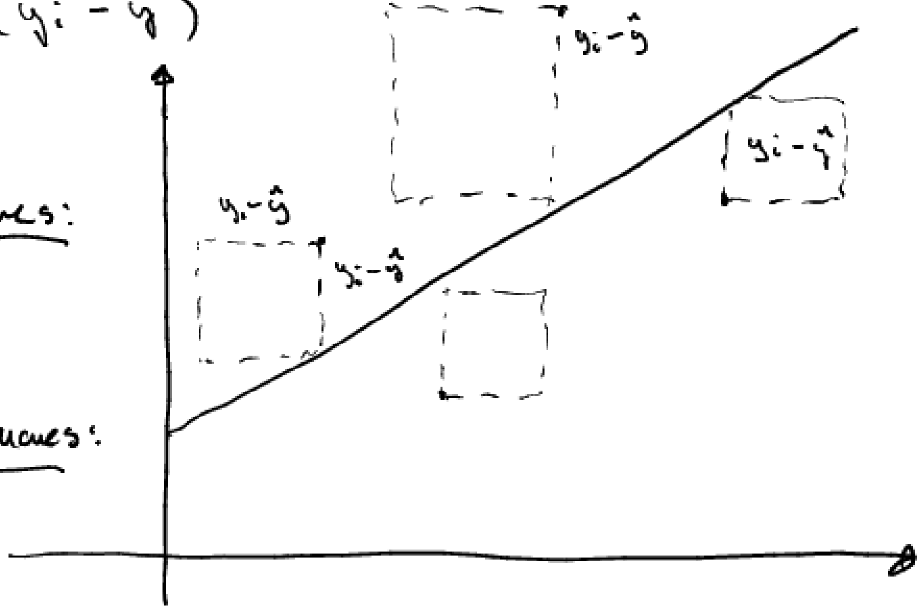
$$SS_t = \sum (y_i - \bar{y})^2$$

Regression sum of squares:

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

explained : $y - y$

Unexplained : $(y_i - \hat{y})$
(Residual)



<u>Total sum of squares:</u>

$$SS_t = \sum (y_i - \bar{y})^2$$

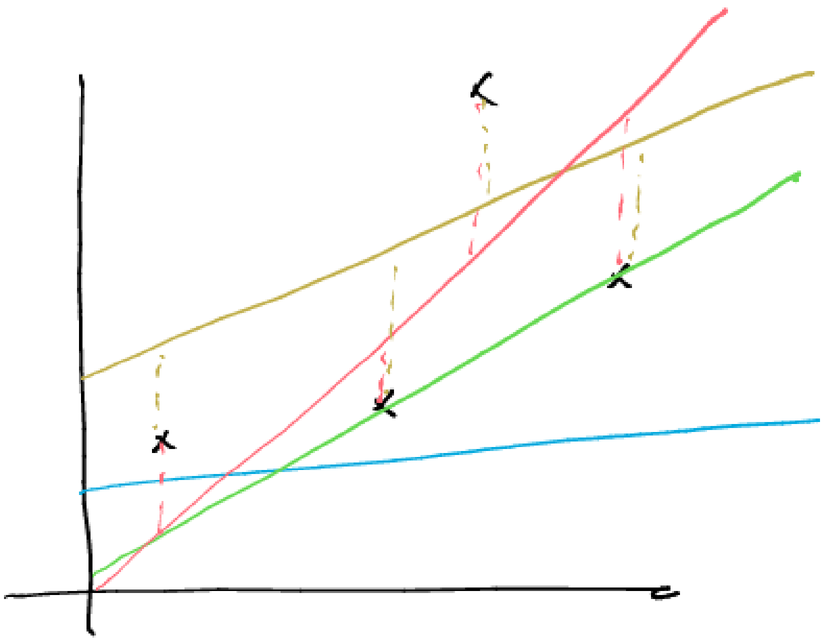
<u>Regression sum of squares:</u>

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

<u>Residual sum of squares:</u>

$$SS_E = \sum (y_i - \hat{y})^2$$

The goal of Linear Regression is to **<u>minimize</u>** the area of all of these squares



5) Assess adequacy of Model

ⓐ Test: $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

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

ⓑ Determine correlation

1) $r = \dfrac{\sum (z_x \cdot z_y)}{n-1}$, z scores for all sample values

2) $r = \dfrac{n \cdot \sum xy - \sum x \cdot \sum y}{\sqrt{n \cdot \sum x^2 - (\sum x)^2} \sqrt{n \cdot \sum y^2 - (\sum y)^2}}$

3) $r = \sqrt{\dfrac{SS_R}{SS_T}} = \sqrt{1 - \dfrac{SS_E}{SS_T}}$

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

$r > |0.6|$ → Good correlation

$r > |0.8|$ → High correlation

ⓒ Find Correlation of Determination:

r^2 = square of r

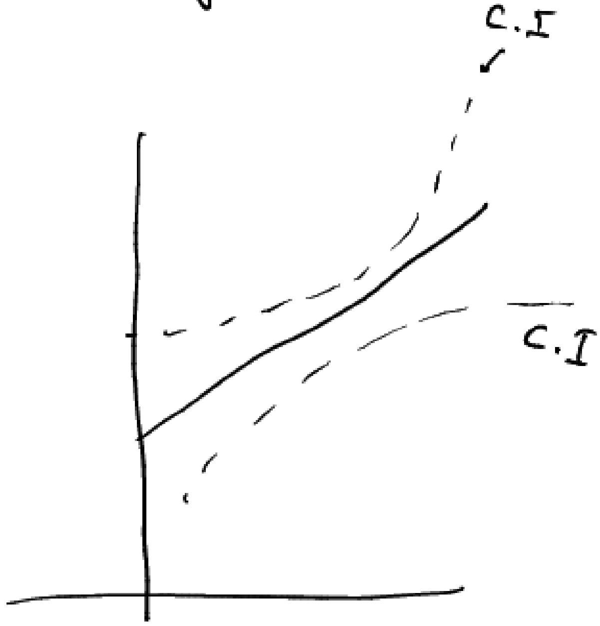
$r^2 = \dfrac{SS_R}{SS_T}$ ← Explained, ← Total } r^2 tells me the amount of variance that my model is able to explain.

$r^2 = 0.78$ → r^2 shows poor

- r^2 tells me something about my model's Predictive Power (Proportion of explained error)

- r tells me something about how well x and y co-vary

6) Find Confidence Intervals for slope and Intercept



Error for slope:

$$E_s = t_{\alpha/2,\,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$S(b_1) \rightarrow$ standard error of slope

Error for intercept:

$$E_I = t_{\alpha/2,\,n-2}\underbrace{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}}$$

$S(b_0) \rightarrow$ standard error of intercept.

Prediction Intervals:

$$E_p = t_{\alpha/2,\,n-2}\cdot\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

x_0 is the x-value we want to find $\hat{y}$ at.