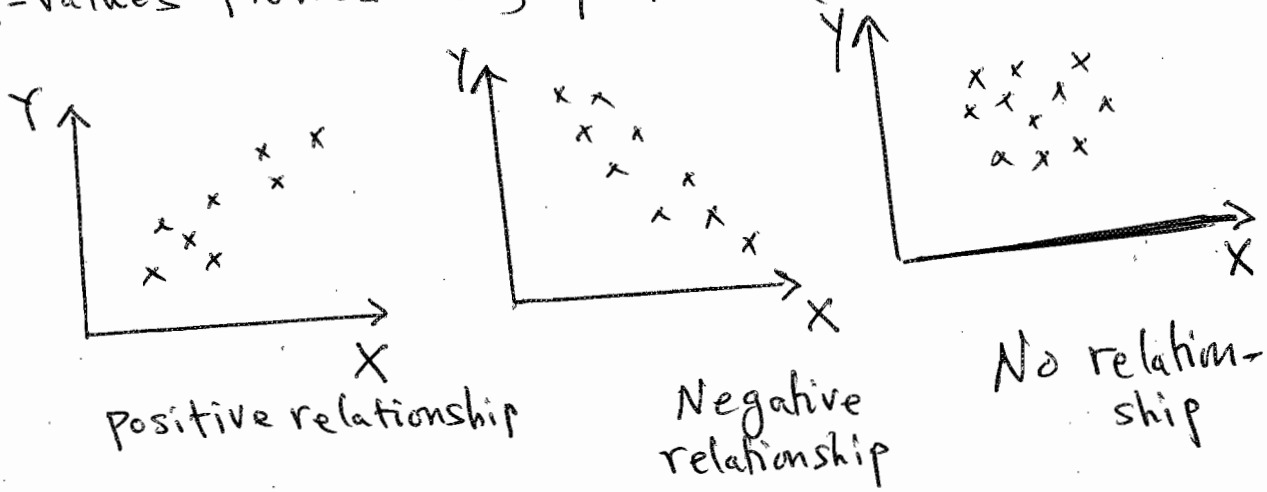
Chapter 11 [Simple Linear Regression]

Graphing Bivariate Relationships

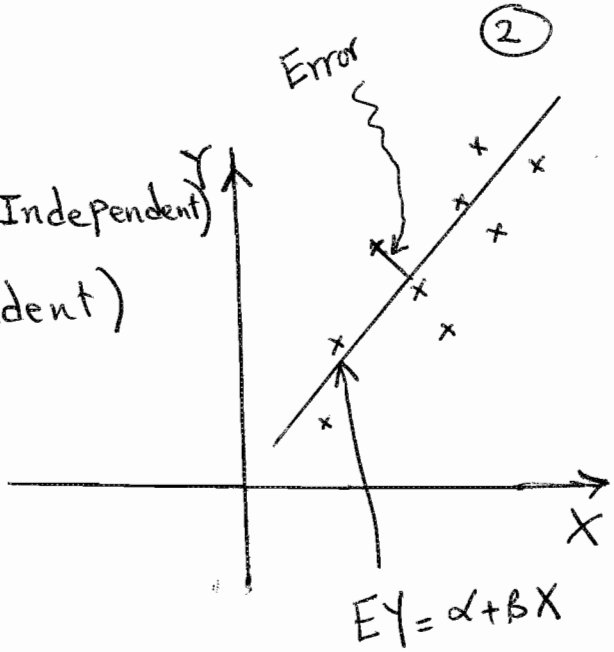
To describe the relationship between 2 quantitative variables X & Y, we plot the data in a scatter plot which is a 2-dimensional plot; with X-values plotted along X-axis (Horizontal Axis) and Y-values plotted along Y-Axis (Vertical Axies).



Goal

- Introduce the straight-line linear regression model for relating one variable X to another variable Y.
- Introduce the correlation coefficient as a tool for measuring the relationship between X and Y.
- Assess how well the simple linear regression model fits the data.
- Use the simple linear regression model to predict the value of one variable based on the value of another variable.

Probabilistic Models



X : Your score in secondary school (Independent)

Y : " " " University (Dependent)

Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

The probabilistic model is

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Where α = Y-intercept

β = slope of the line

ϵ_i = random error

$E(\epsilon_i) = 0$

Note:

① $EY = \alpha + \beta X + E(\epsilon_i)$

$= \alpha + \beta X$

② α and β are population parameters.

③ To estimate EY, we need to estimate α and β as $\hat{\alpha}$ & $\hat{\beta}$

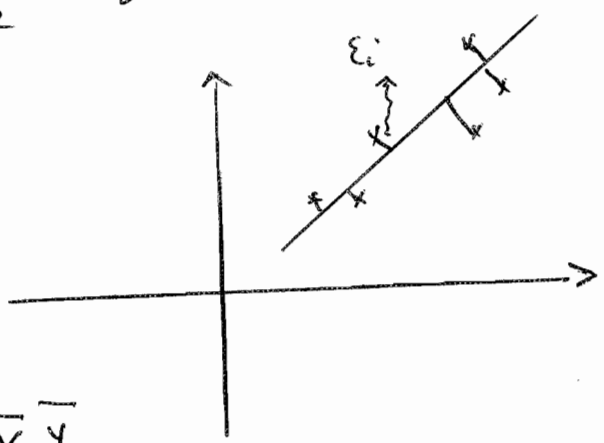
$$\widehat{EY} = \hat{\alpha} + \hat{\beta} X$$

④ EY is the mean response

⑤ β is the change increase in EY when X is increased one unit.

Least squares Estimates

The estimates of α & β are obtained by minimizing the sum of error squares $\sum_{i=1}^{n} \epsilon_i^2$



~~We get~~

$$\sum \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

$$\Rightarrow \hat{\beta} = \frac{SS_{xy}}{SS_x} = \frac{\sum x_i Y_i - n\bar{X}\bar{Y}}{\sum x_i^2 - n\bar{X}^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Note: $SS_{xy} = \sum x_i Y_i - n\bar{X}\bar{Y}$

$SS_{xx} = \sum x_i^2 - n\bar{X}^2$

$SS_{YY} = \sum Y_i^2 - n\bar{Y}^2$

- $SS_{YY} = \sum Y_i^2 - n\bar{Y}^2$: Variation of the response variable Y

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

$$= \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

$$SS_{YY} = SSE + SSR$$

SSE: residual sum of squares

SSR: regression sum of squares (importance of X-variable)

$$SS_R = \frac{SS_{xy}^2}{SS_{xx}}, \quad SSE = SS_{yy} - SSR$$

Example : Find the regression line for the data on income and food expenses.

	Income	Food Expenses			
	X	Y	X^2	Y^2	XY
	28	12	784	144	336
	50	18	2500	324	900
	20	8	400	64	160
	45	20	2025	400	900
	55	21	3025	441	1155
	60	25	3600	625	1500
	42	15	1764	225	630
	38	18	1444	324	684
	32	16	1024	256	512
	35	22	1225	484	770
Total	405	175	17791	3287	7547

$$\bar{X} = \frac{405}{10} = \boxed{40.5}, \quad \bar{Y} = \frac{175}{10} = \boxed{17.5}$$

$$SS_{xx} = \sum X_i^2 - n\bar{X}^2 = 17791 - 10(40.5)^2 = \boxed{1388.5}$$

$$SS_{yy} = \sum Y_i^2 - n\bar{Y}^2 = 3287 - 10(17.5)^2 = \boxed{224.5}$$

$$SS_{xy} = \sum X_i Y_i - n\bar{X}\bar{Y} = 7547 - 10(40.5)(17.5) = \boxed{459.5}$$

$$\hat{\beta} = \frac{SS_{xy}}{SS_{xx}} = \frac{459.5}{1388.5} = \underline{0.33}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 17.5 - (0.33)(40.5) = \underline{4.14}$$

The fitted linear regression line is

$$\hat{Y} = 4.14 + 0.33X$$

The predicted value of the expenses when the income $X=43$ is $\hat{Y} = 4.14 + (0.33)(43) = \boxed{18.33}$

* How much the mean response change when X is increased by one unit?

$\hat{\beta} = 0.33$ for 1 unit increase

$\hat{\beta} = 0.66$ = 2 units =

Coefficients of Correlation & Coefficients

r = Corr. Coefficient [measuring the strength of the linear relationship between X & Y]

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

In our previous example,

$$r = \frac{459.5}{\sqrt{(1388.5)(224.5)}} = \frac{459.5}{558.3} = \boxed{0.82}$$ positive strong relationship

$$-1 < r < 1$$

$r \approx 0$ Weak relationship

$r \approx 1$ strong = (positive)

$r \approx -1$ = = (negative)

Coefficient of Determination

$$SS_y = SSE + SSR$$

(Error sum of squares) (regression sum of squares)

R^2 = Coeff. of Determination $= \frac{SSR}{SS_y}$

$$SSR = \frac{SS_{xy}^2}{SS_x} = \frac{(459.5)^2}{1388.5} = \frac{211140.25}{1388.5} = \boxed{152.06}$$

$$R^2 = \frac{152.06}{224.5} = 0.68$$

Explaination: 68% of the total variations of Y: expenses are explained by X: income. That is, 32% of the total variations are explained by other variables not mentioned.

What is the percentage of variation of the expenses explained (or due to) income? $R^2 = 0.68$