Simple Linear Regression

I- Regression Basics

↳ Simple Regression is the study of the relationship between two variables (Dependent and Independent)

↳ Dependent Variable (DV): Estimated Variable - y -
Independent Variable (IV): Used to estimate(DV) - x -

↳ The Totat Variation (SST) is separated into Explained (SSR) - due to regression - and Unexplained (SSE) - errors - variations

↳ With linear regression, we are trying to reduce unexplained variation, therefor to find the 'best fitting line' trought data that minimizes the SSE (difference between the observed and predicted value)

II- Regression

→ Least Squares Linear Regression (best-fit line)

$$\text{DV} \rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$$

($\hat{y}_i$: DV; $\hat{\beta}_0$: y-intercept; $\hat{\beta}_1$: slope; x_i: IV; ε_i: error)

→ Finding Coefficient of the best-fit lin ($\hat{\beta}_0$ and $\hat{\beta}_1$)

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

or

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$

- $\bar{y} = \frac{\sum y_i}{n}$ - $\bar{x} = \frac{\sum x_i}{n}$

(n: nb of observation)

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

→ SST / SSR / SSE

SST
(Total Sum of Squares)

SSR (Sum of squares due regression) | SSE (Sum of squares due error)

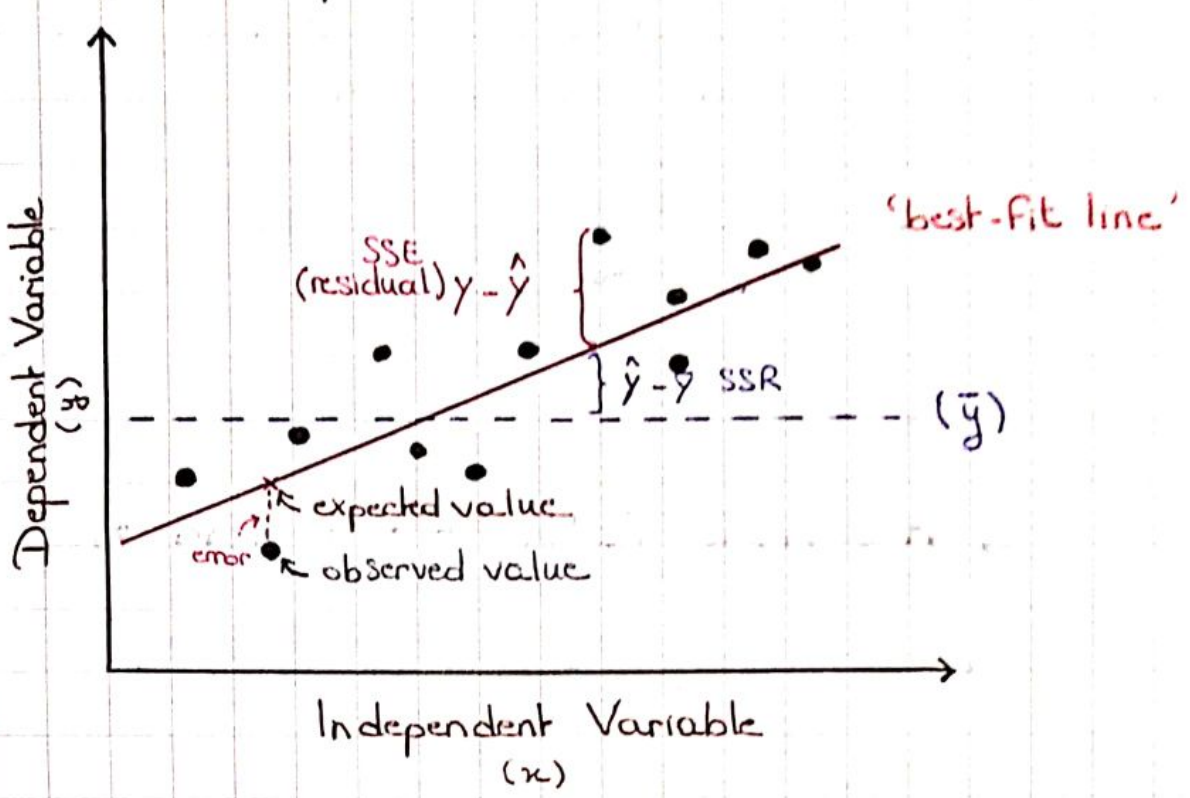
↳ $SSR = \sum (\hat{y}_i - \bar{y})^2$

↳ $SSE = \sum (y_i - \hat{y}_i)^2$

↳ $SST = SSR + SSE = \sum (y_i - \bar{y}_i)^2$

!! Error Variance Estimator: $s^2 = \frac{SSE}{df} = \frac{SSE}{n-2}$

→ Graphical Representation



III. Regression Result

↳ Testing how well the estimated regression equation fit our data.

→ Coefficient of Determination (R^2)

↳ Proportion of the variation in the D.V. (y) that is explained by the variation in the I.V. (x)

↳ $R^2 = \frac{SSR}{SST}$

↳ R^2 range between 0 and 1

↳ SSE ↘
R^2 ↗

↳ The closer R^2 is to 1, the more the data are align with the 'best-fit line'

→ Standard Error of Estimate (S)

↳ Mesure of dispersion of the observed values around the line of regression

↳ $S = \sqrt{\frac{SSE}{n-2}}$

↳ Smaller S => better fit

→ Hypothesis Testing for β_0 and β_1

↳ Assessing whether the estimated regression coefficient (slope and intercept) are significantly different from zero

↳ Understand whether the independent variable has a statistically significant effect on the dependent variable

↳ Hypotheses :

For β_1	For β_0
$H_0 : \beta_1 = 0$	$H_0 : \beta_0 = 0$
$H_1 : \beta_1 \neq 0$	$H_1 : \beta_0 \neq 0$

Test Statistic

T-test for significance

$$t = \frac{\hat{\beta}_1}{S\hat{\beta}_1}$$

- $S\hat{\beta}_1$: Standard deviation of the slope

$$S\hat{\beta}_1 = \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$t = \frac{\hat{\beta}_0}{S\hat{\beta}_0}$$

- $S\hat{\beta}_0$: Standard deviation of the y-intercept

$$S\hat{\beta}_0 = S \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

Degree of Freedom

$df = n - 2$

Critical Region

Choose a significance (α) - commonly 0.05

Decision Rule

Find t-critical value from t-table (two tail)

if $|t| > t\text{-critical}$ => reject H_0

if $|t| \leq t\text{-critical}$ => fail to reject H_0

Interpretation :

H_0 rejected: Slope significantly different from zero

=> Existance of linear rlt between x and y

H_0 not rejected: not enough support to claim the existance of linear rlt between x and y

H_0 rejected: the intercept is statistically significant

H_0 not rejected: the intercept is not statistically significant

→ Confidence Interval of $\hat{\beta}_1$ and $\hat{\beta}_0$

↳ Range of values within wich we are reasonably confident, the true population value lies

↳ CI: $\hat{\beta}_1 \pm t \times S\hat{\beta}_1$

$\hat{\beta}_1$: estimated Slope
t : critical value
$S\hat{\beta}_1$: Standard deviation of the slope

↳ CI: $\hat{\beta}_0 \pm t \times S\hat{\beta}_0$

$\hat{\beta}_0$: estimated intercept
t : critical value
$S\hat{\beta}_0$: Standard deviation of the y-intercept

→ Hypothesis testing for Slope Using Anova

↳ Anova Table

Source of Variation	Sum of Squares	Degrees of freedom	Means of Squares	F
Regression	SSR	1	MSR = SSR/1	MSR/MSE
Residuals	SSE	n-2	MSE = SSE/n-2	
Total	SST	n-1		

↳ Hypothesis

$H_0 : \hat{\beta}_1 = 0$ (There is no relationship between x and y)

$H_1 : \hat{\beta}_1 \neq 0$ (There is a linear relationship between x and y)

↳ Test Statistic (F-Statistic)

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$$

↳ Critical Value

F-table : $\alpha = 0.05$

$df = n - 2$

↳ Decision Rule

if $F > F$ critical => reject H_0

→ P-value of the model

If $P < 0.05$ => there is a statistically significant relationship between IV and DV