Regression Notes: Strong / perfect Regression line.

⇓

its explotary Variable.

مينفعش يكونوا مرتبطين ببعض

Independent Correlated

=

Multi Correniality.

Auto Correlation ⟹ Relationship Between Variable & itself.

Muli Correlation ⟹ Relationship Between explatory Variable of Regression line.

Correlation ⟹ Relationship Between two independent Variables.

* the Regression line "perfect" ⟹ Should build on Independent Variables.

* لو فيه 5 او 6 وفيه علاقة بينهم لازم نشيل الى عامل مشكلة قبل ما ابدأ اعمل Regression Model

* علشان اقدر اشوف العلاقة بين الـ Variables اعمل Scatterplot وشوفة الى شيحددلى " المفروض يطلع " Randomized " دة كدة معناه ان مفيش علاقة.

* Scatterplot ⟹ use to find Relationship Between variables.
⟹ use to Detect the outliers.

* اشوف لو فعلاً مؤثرين من اللى عملت اعمل الـ Regresson Model ولا موجودين.
* بقى الـ Model الأول مرة بيهم و مرة من غيرهم واشوف لو فيه اختلاف فى الـ output بقى شيلهم لو هو هو بقى شيلهم ولما كدة مش مؤثرين.

Nonlinear ← بيقيس على

Correlations Linear Relationships :∴

Correlation ⇓ Strength of Linear Relationships Bet. 2 variables.

VS

Covariance ⇓ Measure of association. Between two variables. (numerator.)

لدى قوة الـ Strength بتستخدم ↑

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})(n-1)}{S_x S_y}$$

* الاثنين دول بيقيسوا الـ Strength بس مقدر شي اعتمد على الـ Covariance وانا بقارن بين pairs of variables

لما اجي اقارن بين اكثر من variables ← استخدم الـ Correlation احسن.

* **Simple Linear Regression** :∴ this line quantifies Relationships Between two variables $x \rightarrow y$ (explatory ← x, dependent ← y)

عن طريق الـ Scatterplot

يعدل نقطة من النقط line

(✓) Datapoints بتاعت الـ Deviation الـ Min يعدل line * بحاول ادور على

(✗) Datapoints عشان ان بحاول اجيب line يقع عليه كل او معظم الـ

Fitted Value $(\hat{y})$ Magnitude Value of point if this point on the Straight line.

* بحسب قيمة لو كانت على الخط كانت هتبقى قيمتها كام.

$(\hat{y} - \hat{y})$ ⇒ Residual. (الباقي) if (-ve) ∴ point under the line. if (+ve) ∴ point over the line.

Least Squares Estimation :∴ بتعدل الـ Errors الـ Square دى الـ (+ve) و الـ (-ve)

$\sum ( )^2$

* the Best Fitting Line ⇒ Line with the Smallest Sum of Squared Residuals

Imp

$\sum e_i^2$

Regression eq

output (dependent).

$y = a + b x$ → Feature (explatory).

Intercept, Slop.

slop: (+ve), (-ve)

$b = r_{xy} \frac{S_y}{S_x}$

$a = \bar{y} - b\bar{x}$

Imp for exam

observed value = Fitted value + Residual.

Residual ⇒ Diffrence Between Actual & Fitted value of Dependent variable. $(y - \hat{y})$

Standard error of estimate ⇒ Magnitude of Residuals. the Standard deviation of Residuals

$$Se = \sqrt{\frac{\Sigma e_i^2}{n-2}}$$

R^2 ⇒ Imp to Measure goodness of fit between (0 - 1).

No Relationship ↙ ↘ perfect.

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma (y_i - \bar{y})^2}$$

Note: In Simple Linear Regression R^2 is Square of Correlation Between Dependent var (y) and explatory var (x).

Residual = (Actual - predicted)

Standardized Residual = $\frac{\text{residual}}{\text{Standard deviation}}$ ↘ if > (±2 or ±3) ∴ it is outliers.

Checking Assumptions

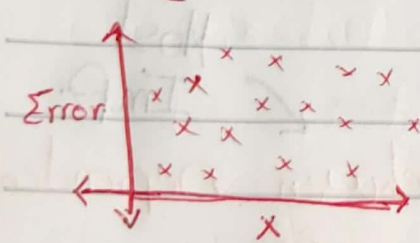
لازم نعمل check على كذا حاجة قبل ما نبدأ نعمل الـ Regression Model.

Linearity

Linear ولا Non linear

1) Scatterplot. ← Linear
2) Residual plot.

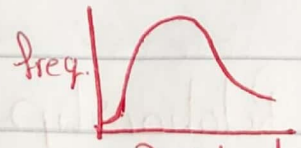
لازم يطلع Random



Normality of Errors

1) Histogram of Standarized Residuals.

لازم يطلع Normal.



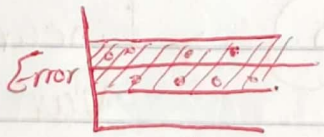
Homoscedasticity

1) Variation about Regression line is constant

⇓

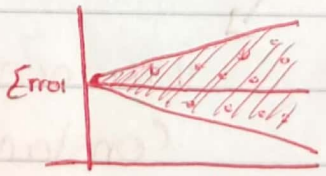
Residual plot.

لازم يطلع Constant



(✓)

لو كده مش Constant الـ Errors بتزيد.



(✗)

Independence of Errors.

Successive observations should be not Related.

Imp when Independent variable is time.

Multiple Regression ⇒ Filling a plane to Data. in 3D. space.
⇒ Slope term for each explantory vari
In equation but interpretation of
these terms are different.
⇒ R^2 Same as Simple. Regression.

$$y = a + b_1 X_1 + b_2 X_2 + \cdots + b_K X_K$$

$$S_e = \sqrt{\frac{\Sigma e_i^2}{n-K-1}}$$

↳ No of explantory var.

Check Assumptions. → Linear Relationship Between Dependent Var & explantory vars. (Most Imp.)

- Errors prob. Independent
- Dependent Var is Normaly Distributed.
- variance of Dependant var is constant for any explantory vars.

(Note) More common problem is. Multicorrllinearity, where explantory vars. are Highly correlated.

VIF grater then 10. ∴ large Correlation

(from table) VIF ⇒ (1 - 5) No correlation
⇒ (1-5) Moderate correlation.

Cross Sectional Data ⇓ usually taken for granted.

Time Series Data ⇓ Assumption often violated. → Bec auto correlation.

(Note) Durbin-watson (DW) Statistic.
(Measure auto correlation).
(always bet (0 - 4))
2 ⇒ No auto correlation.
0 - 2 ⇒ Positive. Correlation.
2 - 4 ⇒ negative Correlation

Measure the Fit of Regression Model

SST (=) SSE (+) SSR.

SST ⇓ Sum square of total. $\sum (y-\bar{y})^2$

SSE ⇓ Sum Square Error $\sum e^2$ (=) $\sum (y-\hat{y})^2$

SSR ⇓ Sum Square of Regression. $\sum (\hat{y}-\bar{y})^2$.

Note $\hat{y}$ ⇒ value which I get. (from Reg equation).

$\bar{y}$ ⇒ Avarage of all values

y ⇒ Real value I have.

$$r^2 = \frac{SSR}{SST} \text{ or } 1 - \frac{SSE}{SST}.$$

Anova table

	df	SS	MS	F	Significance.
Regression.	K	SSR	$\frac{SSR}{K}$	$\frac{MSR}{MSE}$	$p(F > \frac{MSR}{MSE})$
Residual	n-K-1	SSE	$\frac{SSE}{n-K-1}$		
total.	n-1	SST			

p value for calculated F value from table.

we Do hypothesis test ↖ → help Determine if the values are Meaningful.

F-test from Anova table.

testing Model for Significance.

- test Significance for overall Model.
- test Significance of coeff of Model.

Notes: when Sample Size (n) ∴ toosmall ∴ good values of MSE ↙ ↘ r^2 even if No Relationship bet vars

F-test → Ho $B_1 = 0$ ∴ No Relationship Bet x, y.
→ H1 $B_1 \neq 0$ ∴ Here is linear Relationship.

if very Little Error ∴ MSE (Small) & F (Large)

if F (Large) ∴ p-value (low)

↳ we Can Reject Null hypothesis & Accept linear Relationship Bet x-y & MSE & r^2 are Meaningful.

∴ Steps of Hypothesis test ∴

1- Ho : $B_1 = 0$.
H1 : $B_1 \neq 0$.

2- Select Level of Significance (α) (given).

3- Calculate F-value = $\frac{MSR}{MSE}$

4- Reject if F calculated > F_{α, df_1, df_2} (df_1 → k, df_2 → n-k-1)
or Reject if p value < α.

Choose any Method ⇒

Multicollinearity

* When there is Strong Linear Relationships Among Set of explantory variables.

* في الحالة هنا ال (B_i) الى هو ال Slop معناه هو الوحيد الى بيعبر عن مدى تعلق (X) بال (Y) لأن فيه تانيه ظهرت بتأثر فال (Y).

* there are various degrees of Multicollinearity, but in each of them there is Linear.

Include / Exclude Decisions

=> we use t-value. & p-value.

↳ Finding Best X_k to include them in Regression equation.

Notes # t value & p value => if p-value Above (α).
∴ Candicate for exclusion.

if t value less than or ~~greater than (1)~~
∴ excluded from eq.

Stepwise Regression

→ Stepwise.
↳ Same the forward. except also considers possible deletions. along the way.

forward
↳ begins with No explantory var then Add. one By one. until No Remain. variables found.

Backward.
↳ begins with all. Variables. & Delete them one By one. until further Delete Do harm good.