REGRESSION ANALYSIS FOR PREDICTION

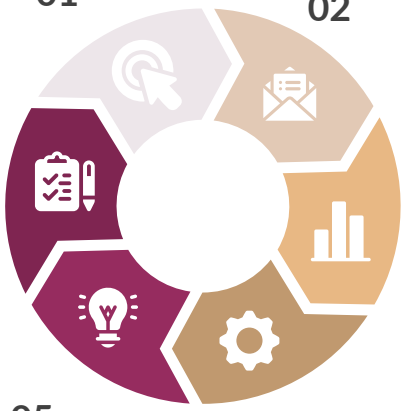
PREDICTION

- Historical (sample) data analyze/train with statistical/machine learning model
- Algorithms perform statistical analysis and/or data mining to determine patterns and trends in data
- Predict future events or behavior

PREDICTION MODEL PROCESS

Business Understanding
- Know what data to use while building the model
- Define model objectives and goals

01



02

Model Specification
- Determine Dependent and Independent Variables
- Choice of Independence variables
- Variables significance (correlation & multicollinearity)

03

Choice of Algorithm
- Statistical process
- Machine Learning process

04

Model Building
- Understand model Output
- Define model parameter
- Construct model equation

05

Model Diagnostic
- ML Model Validation
- Statistical significance test on model and variables significance.
- Percentage of model fitness
- Model Error rate (standard error and RMSE)

06

Model Deployment
- Feedback on model performance to inline with business needs

REGRESSION

- Statistical method used to analyze and predict the relationship between variables.

- Application of Regression:

Forecasting, Time series modelling and Finding the causal effect relationship between the variables.

- In Machine Learning, it falls under Supervised Learning and widely used Algorithm for prediction

VARIABLES FOR REGRESSION ANALYSIS

Dependent (response) variable (DV), *y* – a variation wish to explain

- The response variable is plotted on the vertical axis.

Independent (explanatory) variable (IV), x – used to explain variation in dependent variable.

- The explanatory variable is plotted on the horizontal axis

TYPE OF REGRESSION

SIMPLE REGRESSION ANALYSIS: When we want to make a prediction based on 2 variables (IV & DV) only

When the 2 variables (IV & DV) have linear relationship (as shown from the scatter plot) → it can be referred as simple linear regression

MULTIPLE REGRESSION ANALYSIS: When we want to make a prediction based on more than 2 variables (with more IVs & one DV)

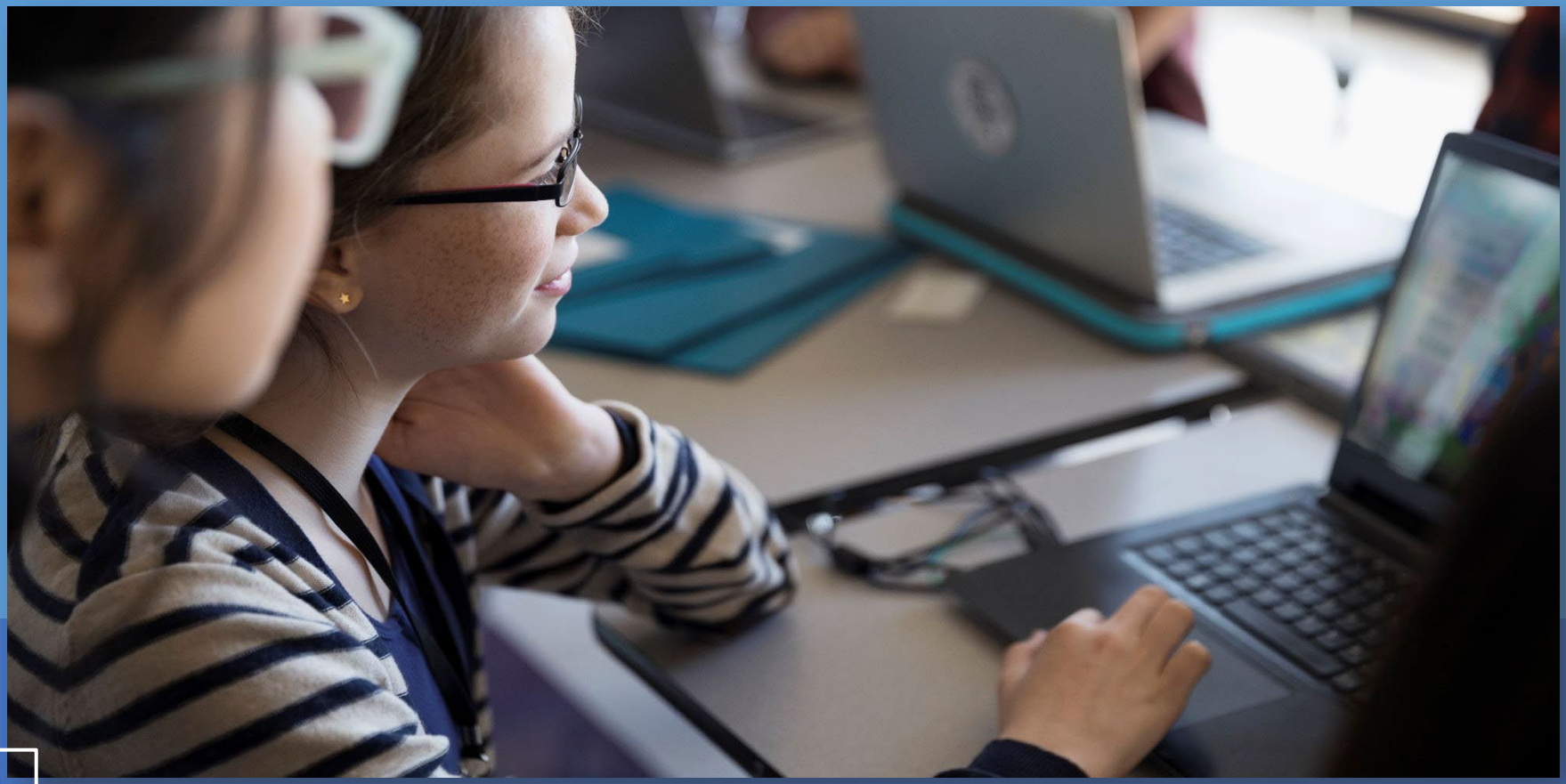
NON-LINEAR/POLYNOMIAL REGRESSION ANALYSIS: When we want to make a prediction based on 2 or more than 2 variables and the relationship between the variables is non-linear as shown from the scatter plot.

REGRESSION MODEL COMPONENT

REGRESSION MODEL SPECIFICATION

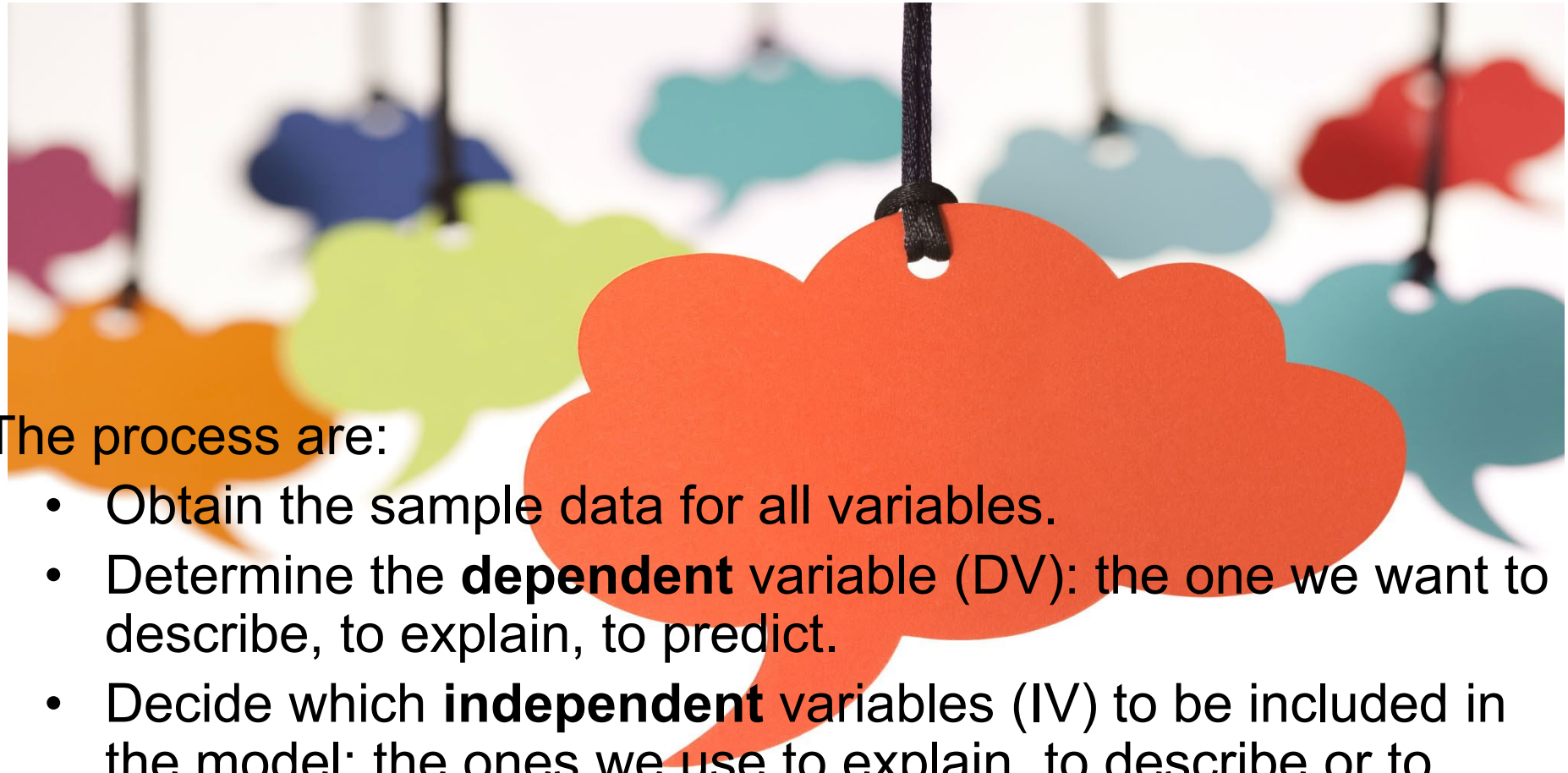
REGRESSION MODEL BUILDING

REGRESSION MODEL DIAGNOSIS



1

REGRESSION MODEL SPECIFICATION



The process are:

- Obtain the sample data for all variables.
- Determine the **dependent** variable (DV): the one we want to describe, to explain, to predict.
- Decide which **independent** variables (IV) to be included in the model: the ones we use to explain, to describe or to predict the dependent variable(s).
- Construct Scatter plot to identify the pattern (linear or non-linear).
- Decide which variables to include in the model and which to exclude. One tool that will aid us in this decision is the *correlation analysis*

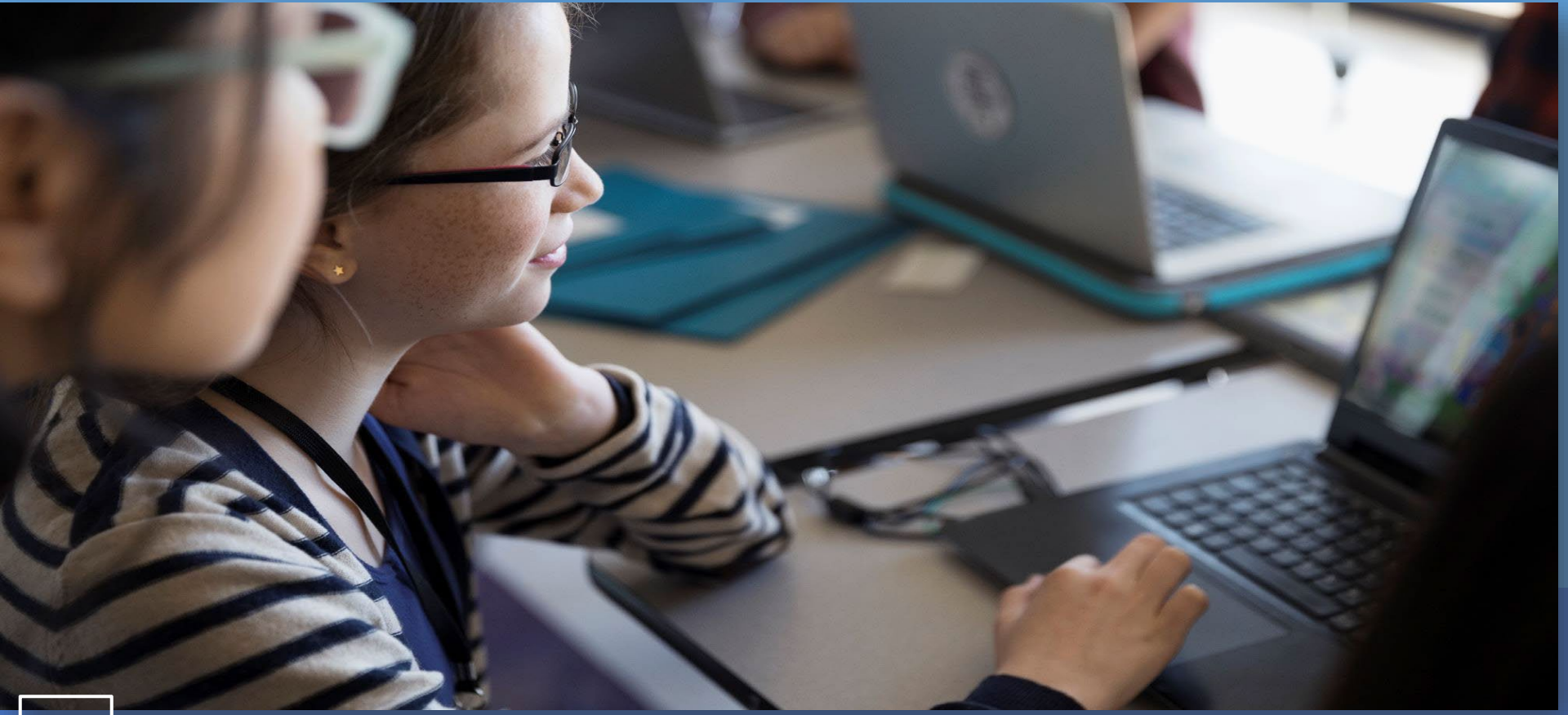
2

REGRESSION MODEL BUILDING

The process of constructing a mathematical equation in which **some or all** the independent variables are used to explain the variation in the dependent variable.

Model Parameter : the dependent variable(s) is linked to the explanatory ones through a mathematical equation (the model) that involves quantities

- Example:
  - for SIMPLE REGRESSION: Dependent_var = **intercept** + **slope***indep_var
  - for MULTIPLE REGRESSION: Dependent_var = **intercept** + **slope***indep_var_1+ **slope***indep_var_2 **+ slope***indep_var_3+…..+ **slope***indep_var_n
- Computations behind statistical modeling allow the **estimation of model parameters** and further predictions of the dependent variable.



3

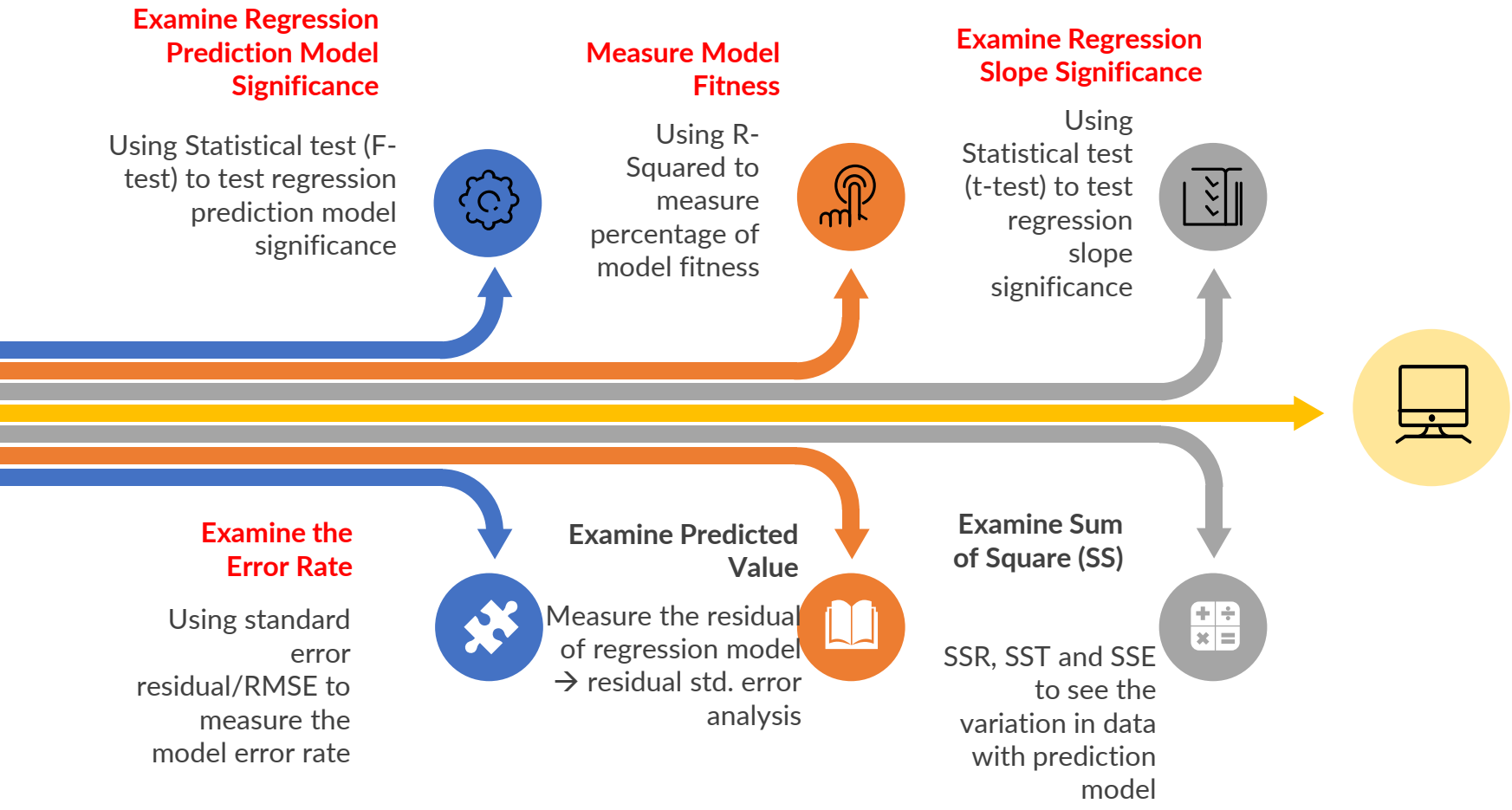
REGRESSION MODEL DIAGNOSTIC

Sophisticated model does not necessary will produce an acceptable result.

PURPOSE OF REGRESSION MODEL DIAGNOSTIC

Analyze the **quality of the model** you have constructed by determining how well a specified model fits the data you just gathered.

PROCESS IN REGRESSION MODEL DIAGNOSTIC



EXAMINE REGRESSION PREDICTION MODEL

The hypothesis statement:

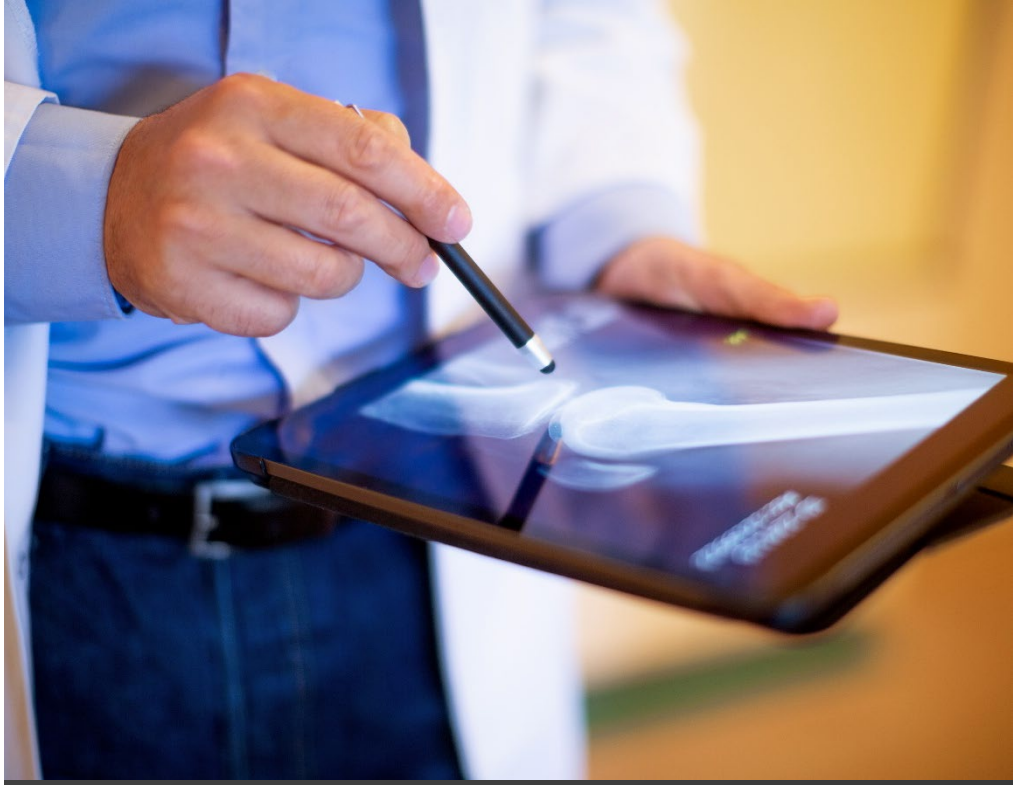
H_0: $\beta_j = 0$ (no variation in the model; model not significance)

H_A:$\beta_j \neq 0$ (exists variation in the model; model significance)

Using the p-value of the F-test and compare with significance value (α):

“the smallest the p-value compare with the significance value (α), then we **REJECT** the h-null”

Therefore, it can conclude that the regression model is significant



F-test

Statistical test for testing whether the regression model explains a significant proportion of the variation in the dependent variable (and whether the overall model is significant).

MEASURE REGRESSION MODEL FITNESS

R^2 value:

- The value is between 0 and 1.0
- R^2 with value 1.0 would respond to a situation in which the regression line would pass through each of the points in the scatter plot. (** the lower the residuals, the higher the R^2 statistic.)

Coefficient of Determination, R^2

$$R^2 = \frac{SSR}{SST}$$

HOW THE DECISION MAKERS USE R^2

To indicate how well the regression line fits the data points. The better the fit, the closer R^2 will be to 1.0. R^2 will be close to 0 when there is a weak relationship.

Coefficient of Determination (R-Squared) → R^2

Used to determine the proportion of variation in the dependent variable that is explained by the dependent variable's relationship to all the independent variables in the model.

EXAMINE REGRESSION SLOPE SIGNIFICANCE

The hypothesis statement:

- Hypothesis Statement for simple regression:

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

- Hypothesis Statement for multiple regression:

$H_0: \beta_j = 0$

H_A: at least one $\beta_j \neq 0$

Using the p-value of the t-test and compare with significance value (α):

"the smallest the p-value compare with the significance value (α), then we **REJECT** the h-null"

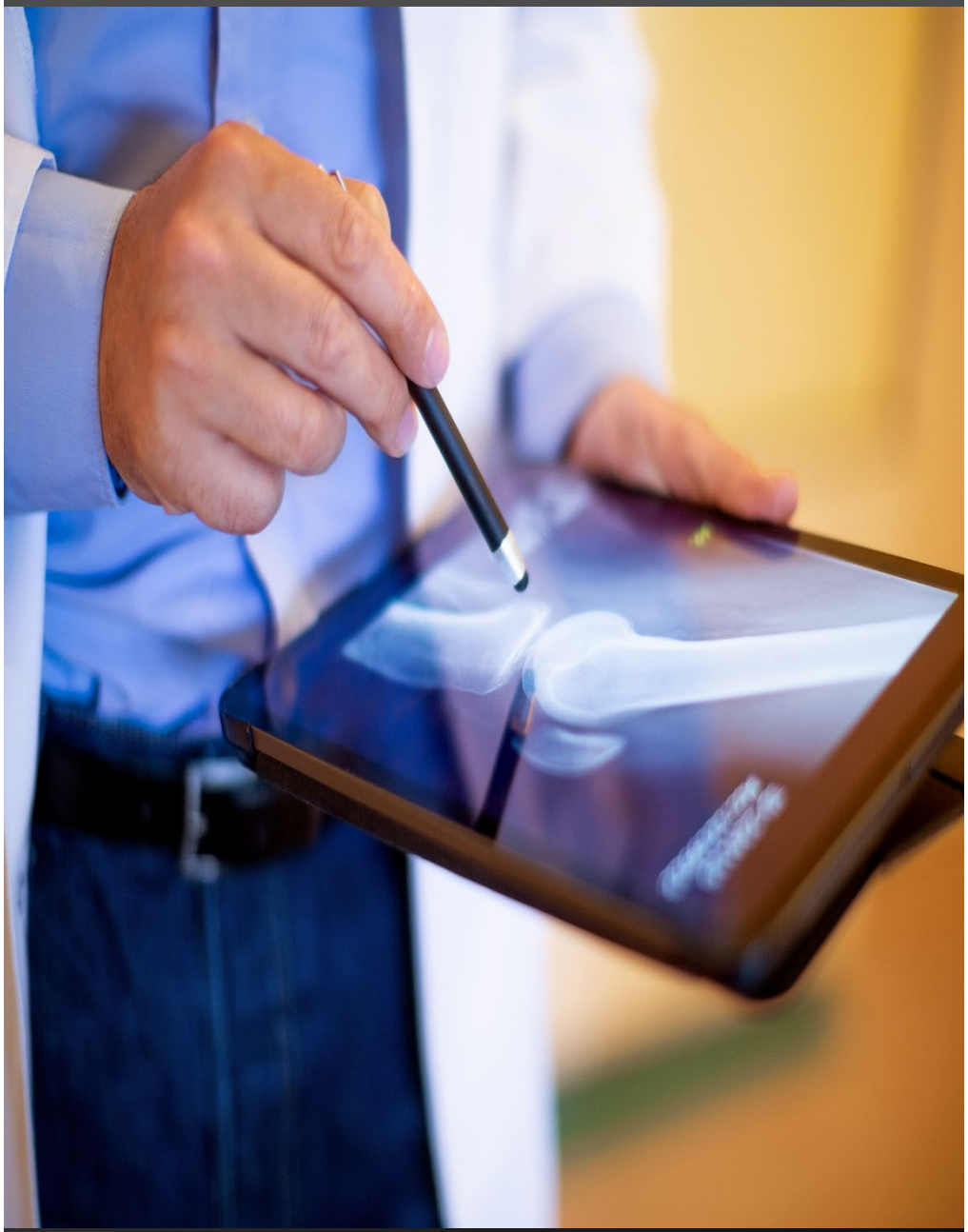
Therefore, it can be concluded that the regression coefficient (independent variable) is significant (can be included in the model)

t-test

Statistical test for testing whether the regression coefficient can be included/excluded in the regression model.

- A slope of 0 would imply that there is no relationship between *x* and *y* variables

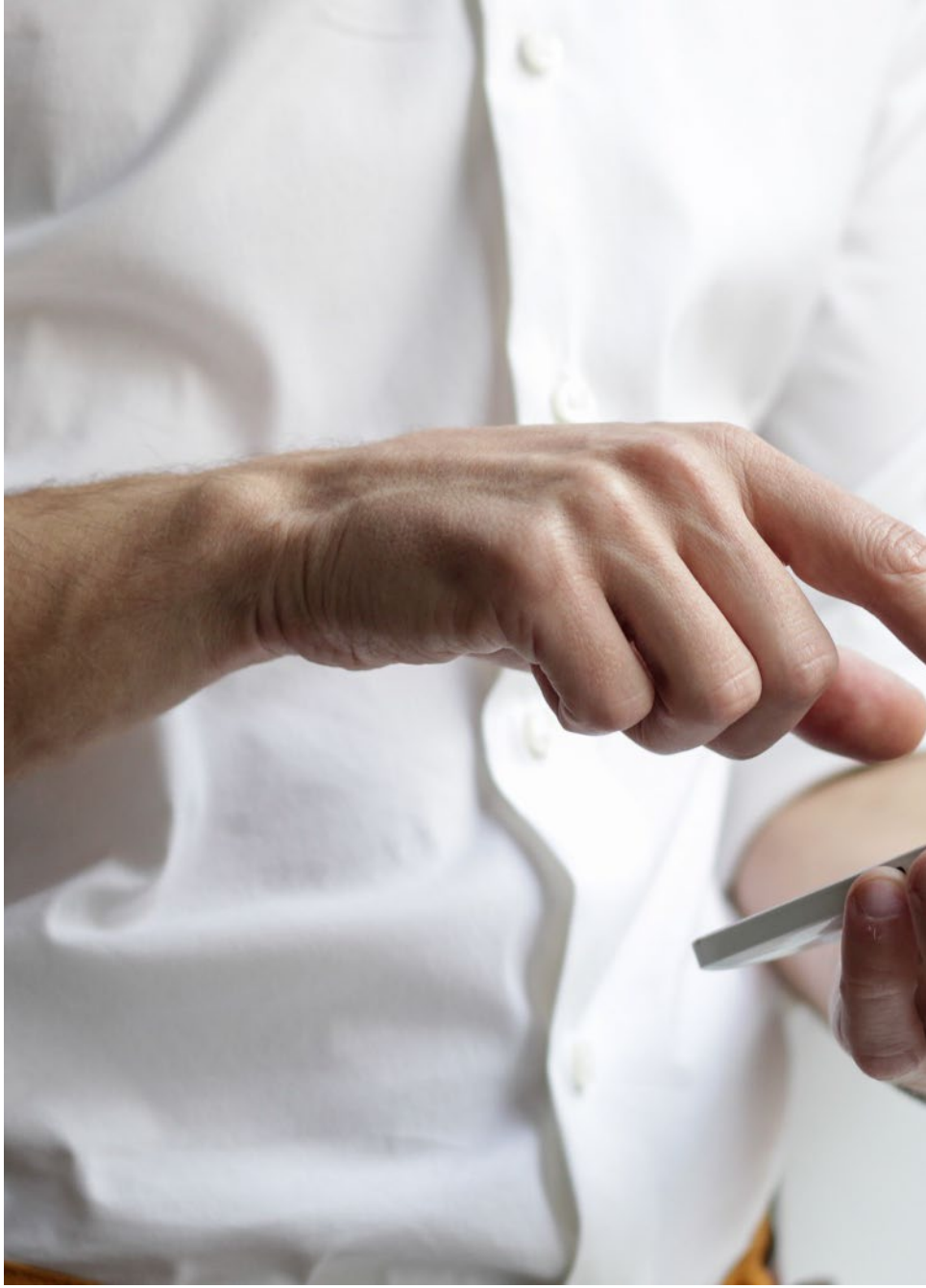
- For simple linear regression model, regression slope = 0, meaning the regression model is equal to constant value of intercept.



EXAMINE THE ERROR RATE OF REGRESSION PREDICTED MODEL

- Root Mean Square Error (RMSE) - square root of the variance of the residuals.
- Lower values of RMSE indicate better fit.

- RMSE is a good measure of how accurately the model predicts the response and is the most important criterion for fit if the main purpose of the model is prediction.




SIMPLE REGRESSION

- This Regression model is used for predicting continuous values of dependent variable with only one independent variable.

- When the 2 variables (IV & DV) have linear relationship (as shown from the scatter plot) it can be referred as simple linear regression

SIMPLE REGRESSION

THE SIMPLE LINEAR REGRESSION MODEL & ASSUMPTIONS

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

where:

y = Value of the dependent variable
x = Value of the independent variable
β_0 = Population's y intercept
β_1 = Slope of the population regression line
ε = Random error term

Estimated Population regression coefficient

Linear Component

Random error component - maybe positive, zero or negative

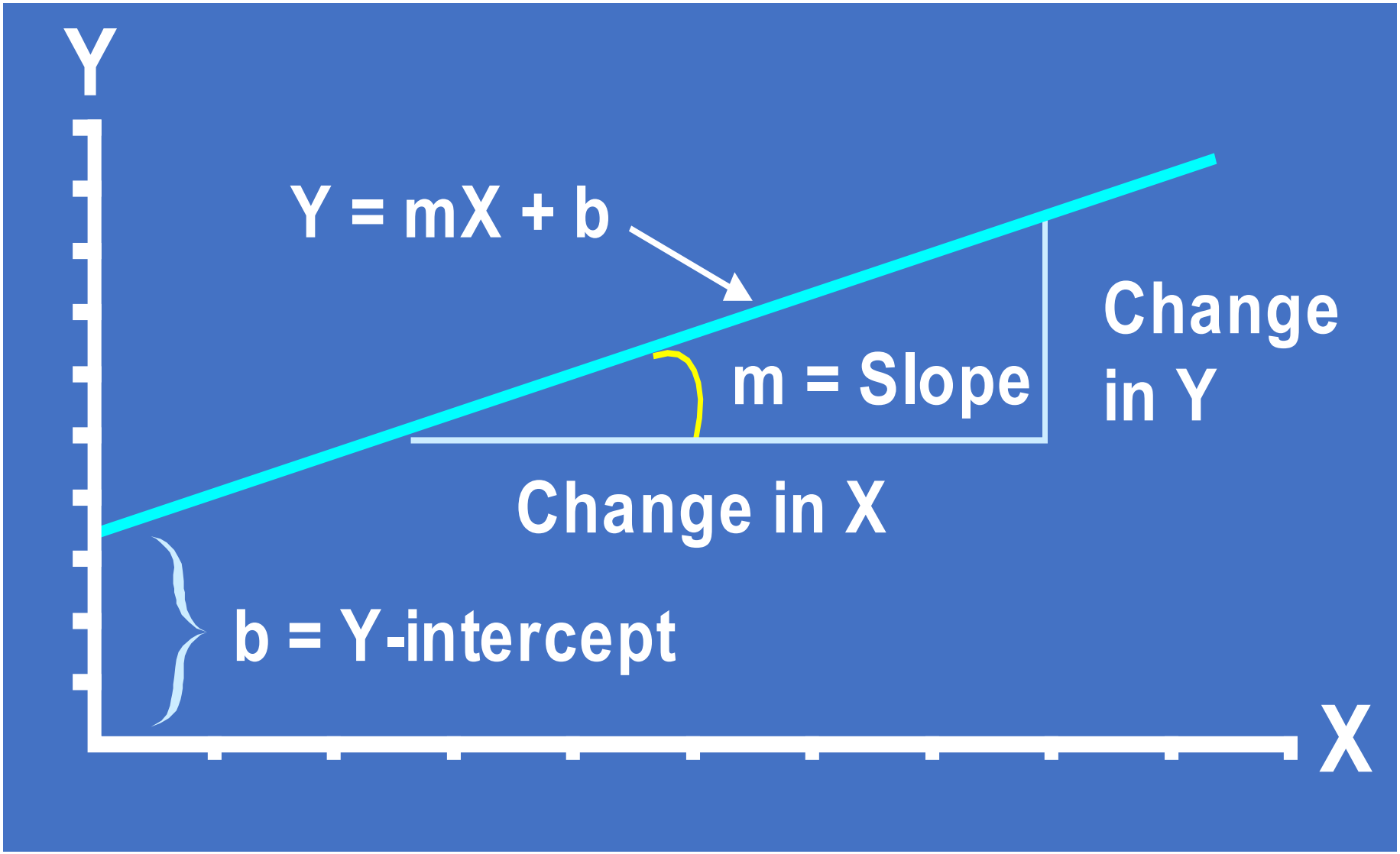
β_0(intercept) : estimated population intercept

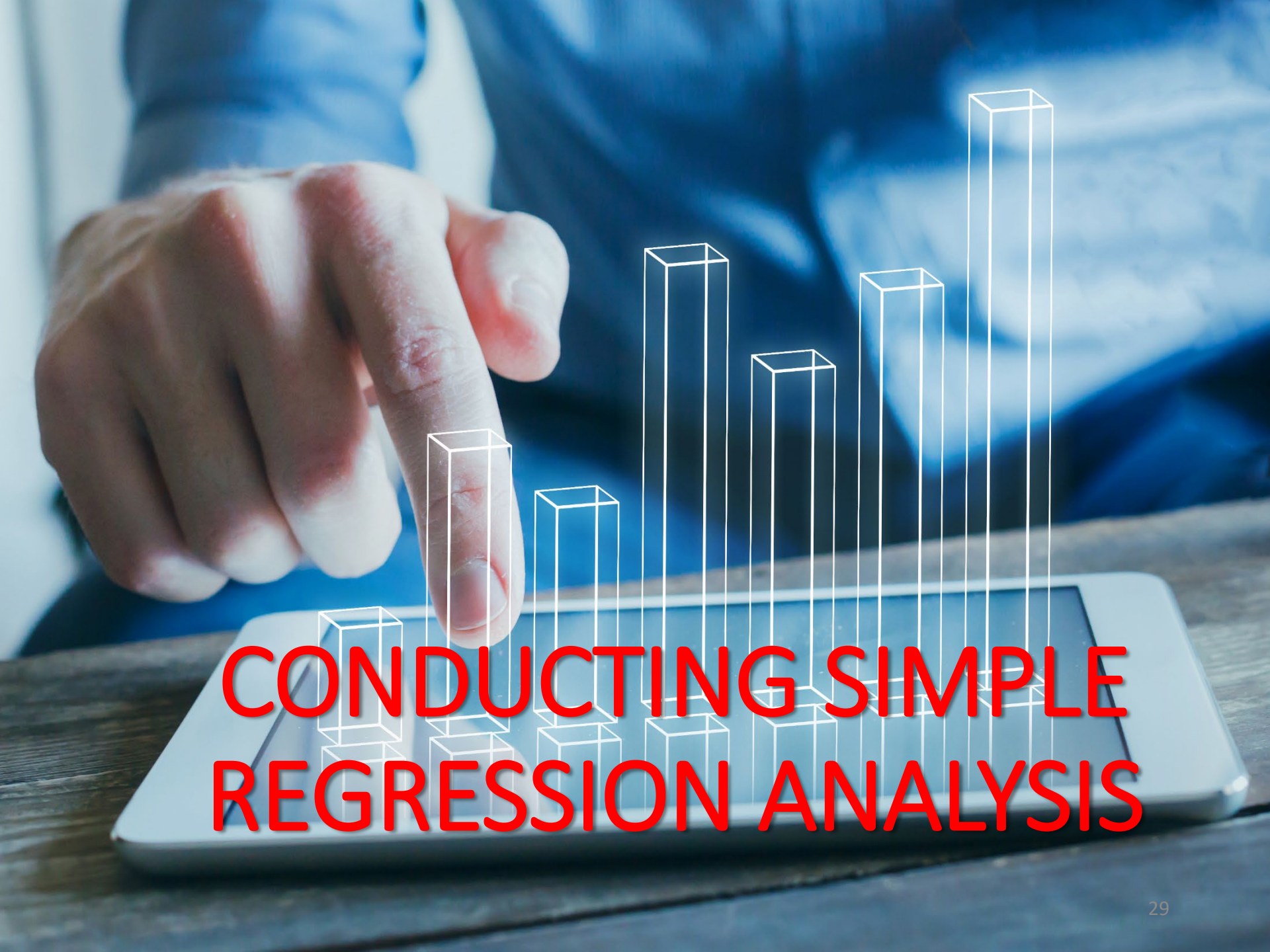
: The *y*-intercept of any line is the point where the graph intersects the vertical axis. It is found by letting $x = 0$ in an equation and solving for *y*. i.e the constant value .

β_1(slope): estimated population slope (slope coefficient)

:can be either positive, zero or negative.

LINEAR EQUATIONS




CONDUCTING SIMPLE
REGRESSION ANALYSIS

PREDICTING PATIENT RECOVERY

A regional hospital wants to optimize patient care by predicting **patient recovery time (in days)** based on different patient characteristics and treatment variables. The hospital's analytics department is interested in using **regression analysis** to identify the most significant factors influencing recovery duration and to forecast expected recovery times for future patients.

Data Set:
Patient_Recovery_data_large.xlsx

Question:

1. Can we predict Recovery Time using Physical Activity Score?
2. What is the strength and direction of the relationship?

MODEL SPECIFICATION

Variables

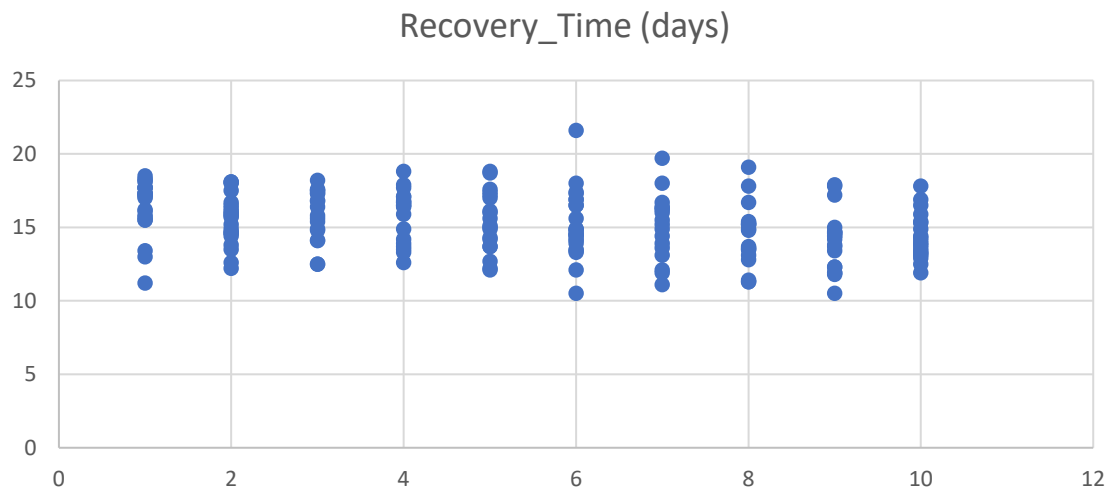
DV: ?, IV:?

Scatter Plot: Data Visualization

Relationship

Correlation Analysis: Assuming the model fitness based on the analysis.

Scatter Plot: Homogeneity pattern



relatively spread evenly across activity levels — indicating **homogeneity of variance** (homoscedasticity).

Correlations

		Physical_Activity_Score	Recovery_Time (days)
Physical_Activity_Score	Pearson Correlation	1	-.336**
	Sig. (2-tailed)		<.001
	N	200	200
Recovery_Time (days)	Pearson Correlation	-.336**	1
	Sig. (2-tailed)	<.001	
	N	200	200

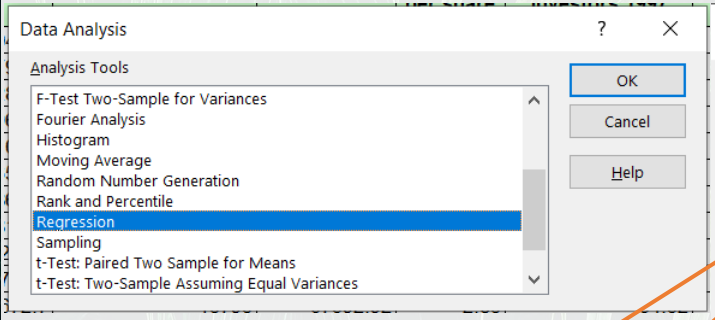
**. Correlation is significant at the 0.01 level (2-tailed).

The correlation analysis yields a coefficient of r = - 0.336, suggesting a weak negative relationship between the Recovery Time and Physical Activity Score

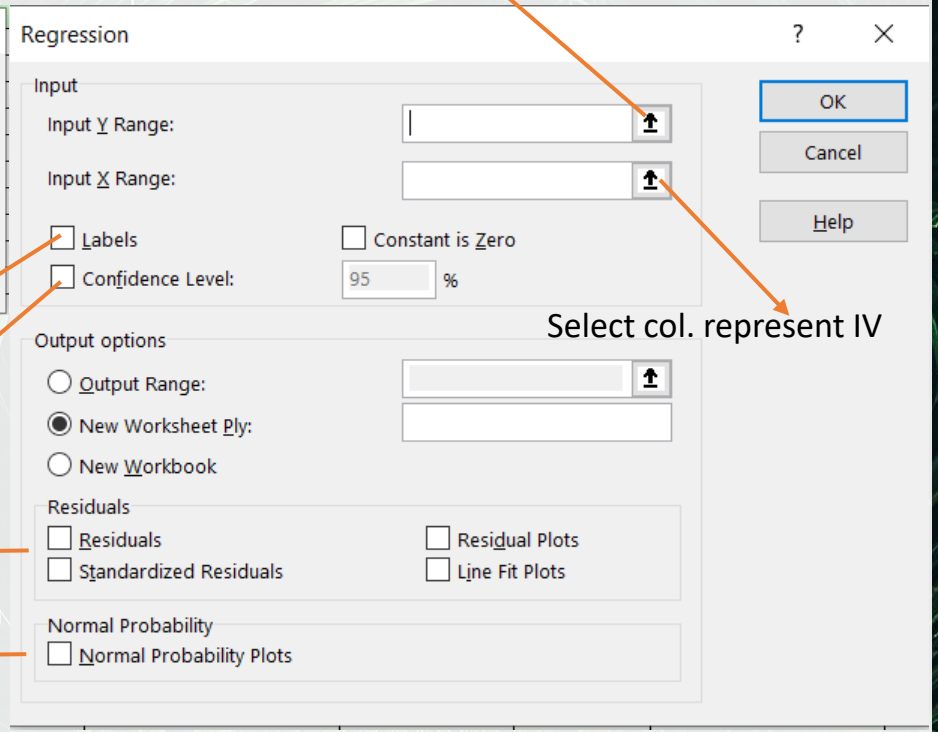
The correlation analysis indicates that higher physical activity levels are **associated with shorter recovery times**, though the strength of this relationship is weak. This could suggest that while being active helps recovery, it's only **one of many factors** influencing recovery duration.

REGRESSION ANALYSIS WITH MICROSOFT EXCEL

Ms Excel -> Data -> Data Analysis



Select col. represent DV



Select col. represent IV

Make 1st row as label

Changing the confidence level

Output for error (residual) diagnosis)

To see either normal or not normal distribution

Regression Statistics	
Multiple R	0.335957021
R Square	0.11286712
Adjusted R Square	0.108386651
Standard Error	1.909377289
Observations	200

R-squared: 0.113 → About 11.3% of the variance in recovery time is explained.

p-value: < 0.001 → Prediction model is Statistically significant.

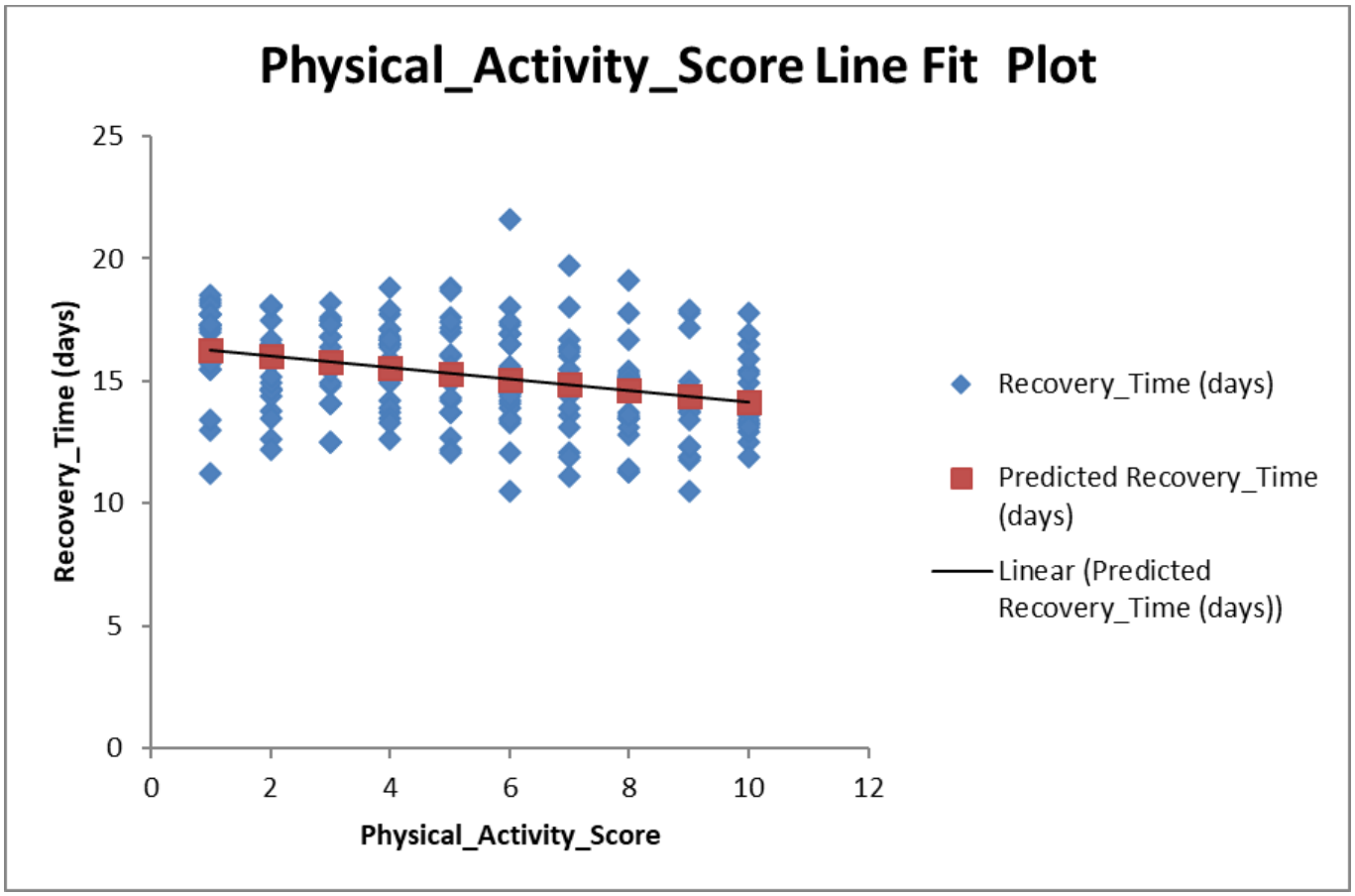
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	91.83907	91.83907	25.19092	1.15344E-06
Residual	198	721.8529	3.645722		
Total	199	813.692			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	16.47643241	0.286263	57.5569	1.4E-125	15.91191601	17.04095
Physical_Activity_S	-0.233320885	0.046487	-5.01906	1.15E-06	-0.324994086	-0.14165

p-value (Physical_activity_score): < 0.001 → The variable physical_activity is statistically significant in predicting the recovery time of patients.

Equation:

Recovery_Time = 16.48 - 0.23 * Physical_Activity_Score

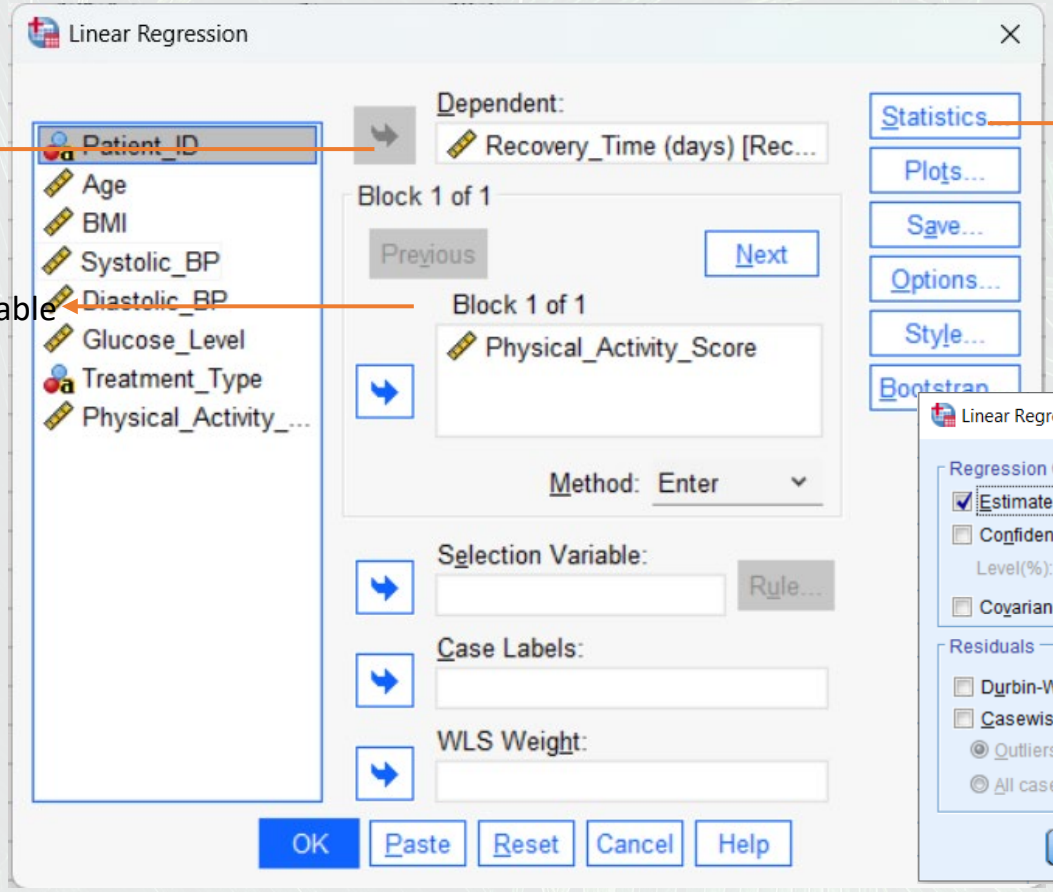

Physical_Activity_Score Line Fit Plot
Recovery_Time (days)
Physical_Activity_Score
0
5
10
15
20
25
0
2
4
6
8
10
12
Recovery_Time (days)
Predicted Recovery_Time (days)
Linear (Predicted Recovery_Time (days))

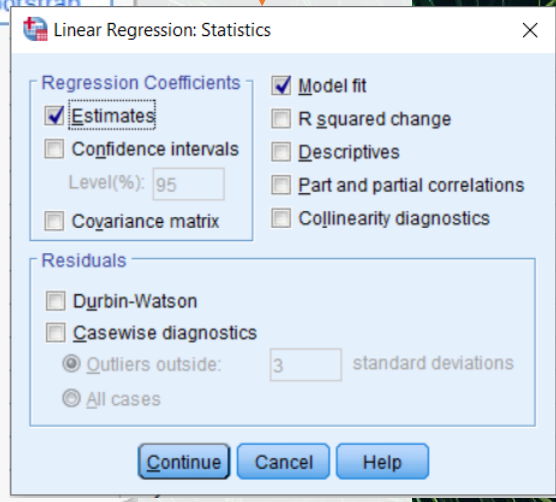
REGRESSION ANALYSIS WITH SPSS

SPSS-> Analyze -> Regression -> Linear..

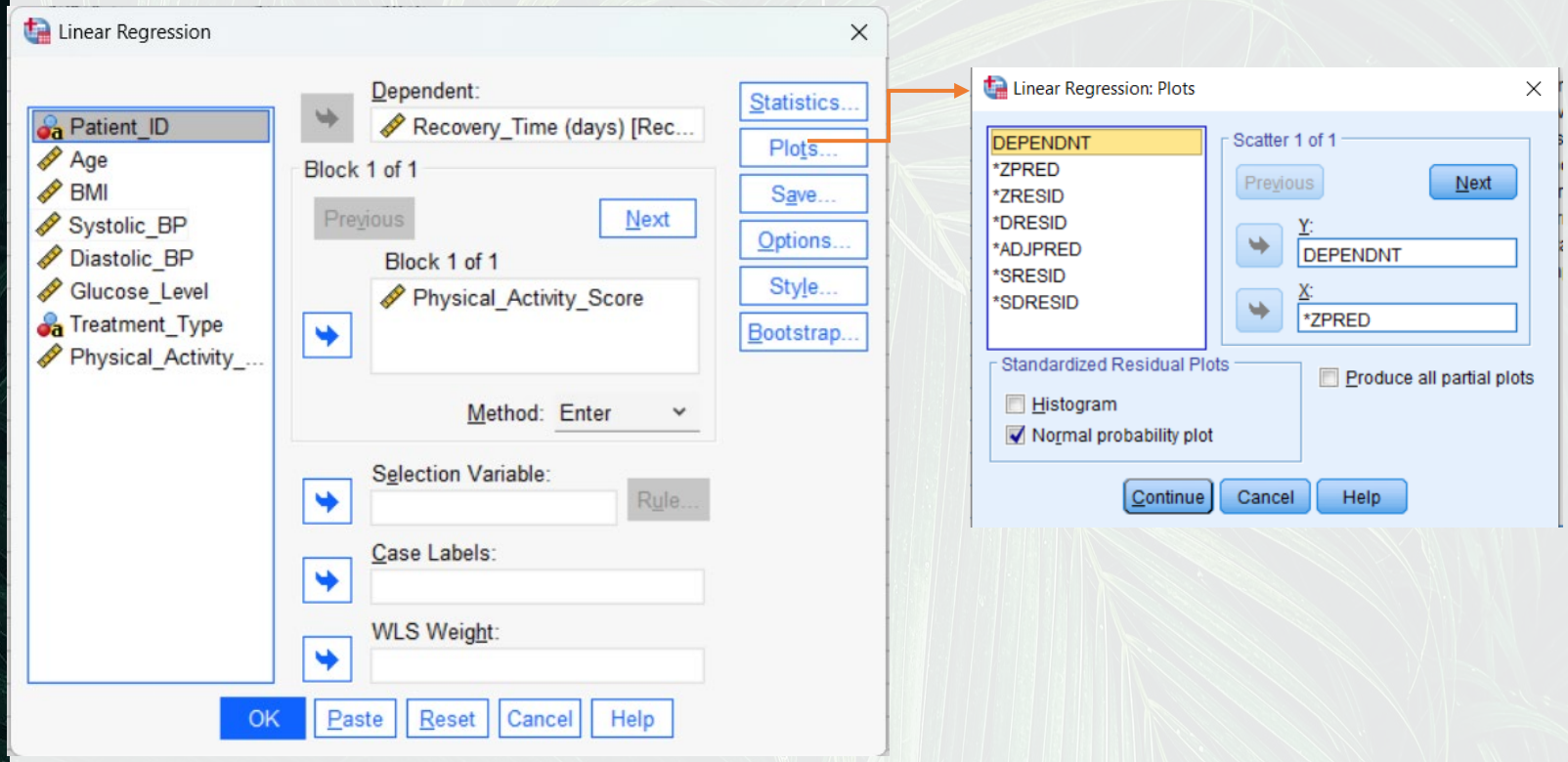
Dependent (Y) variable

Independent (X) variable

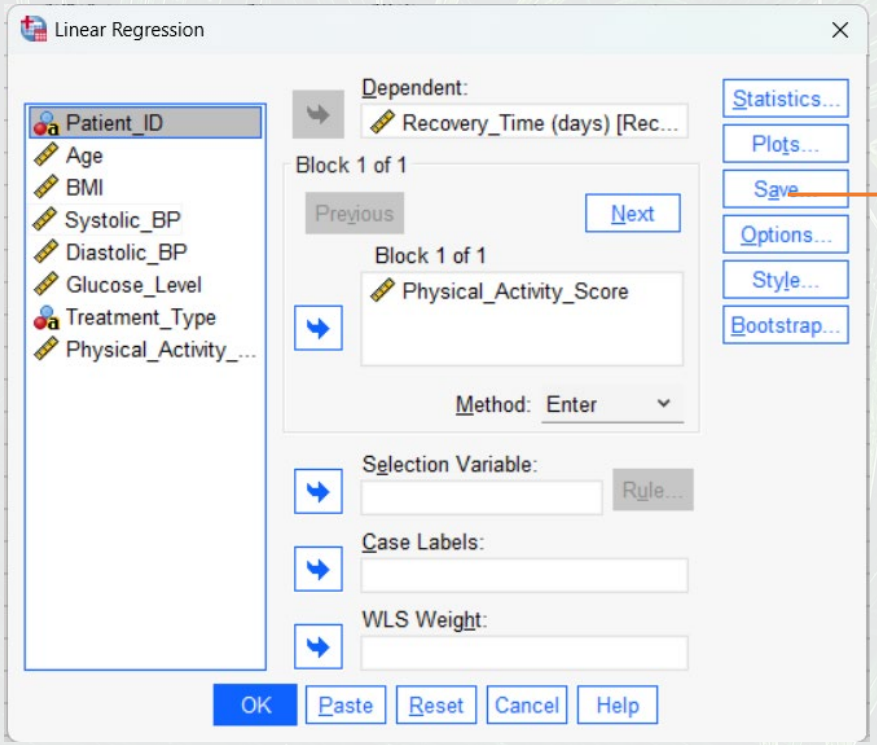


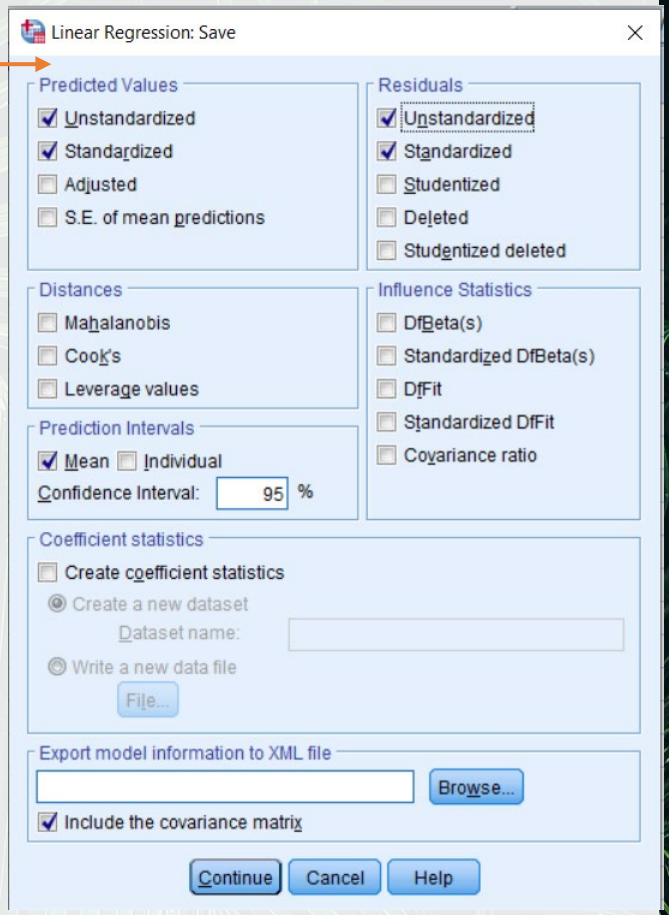


SPSS-> Analyze -> Regression -> Linear..



SPSS-> Analyze -> Regression -> Linear..





Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.336[a]	.113	.108	1.9094

a. Predictors: (Constant), Physical_Activity_Score

b. Dependent Variable: Recovery_Time (days)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91.839	1	91.839	25.191	<.001[b]
	Residual	721.853	198	3.646		
	Total	813.692	199			

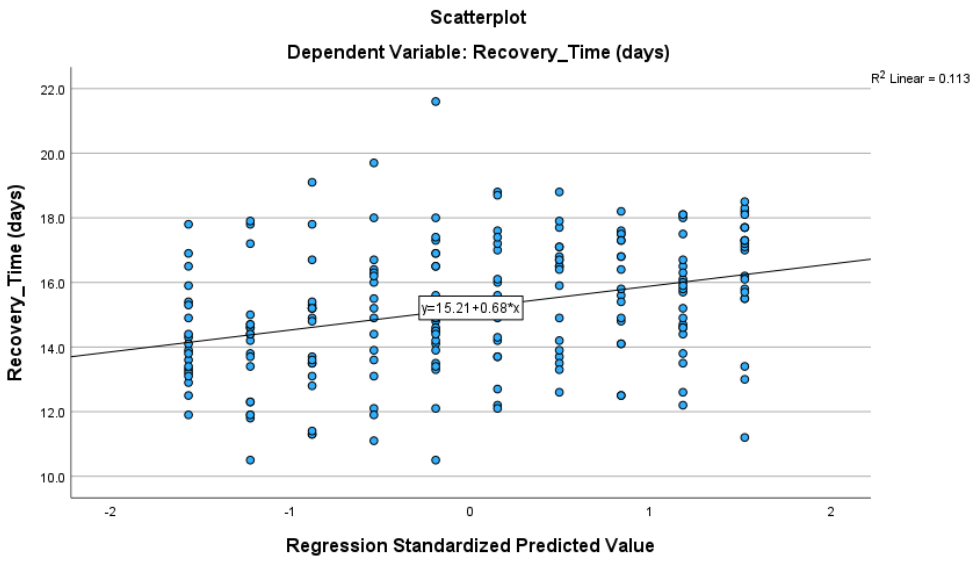
a. Dependent Variable: Recovery_Time (days)

b. Predictors: (Constant), Physical_Activity_Score

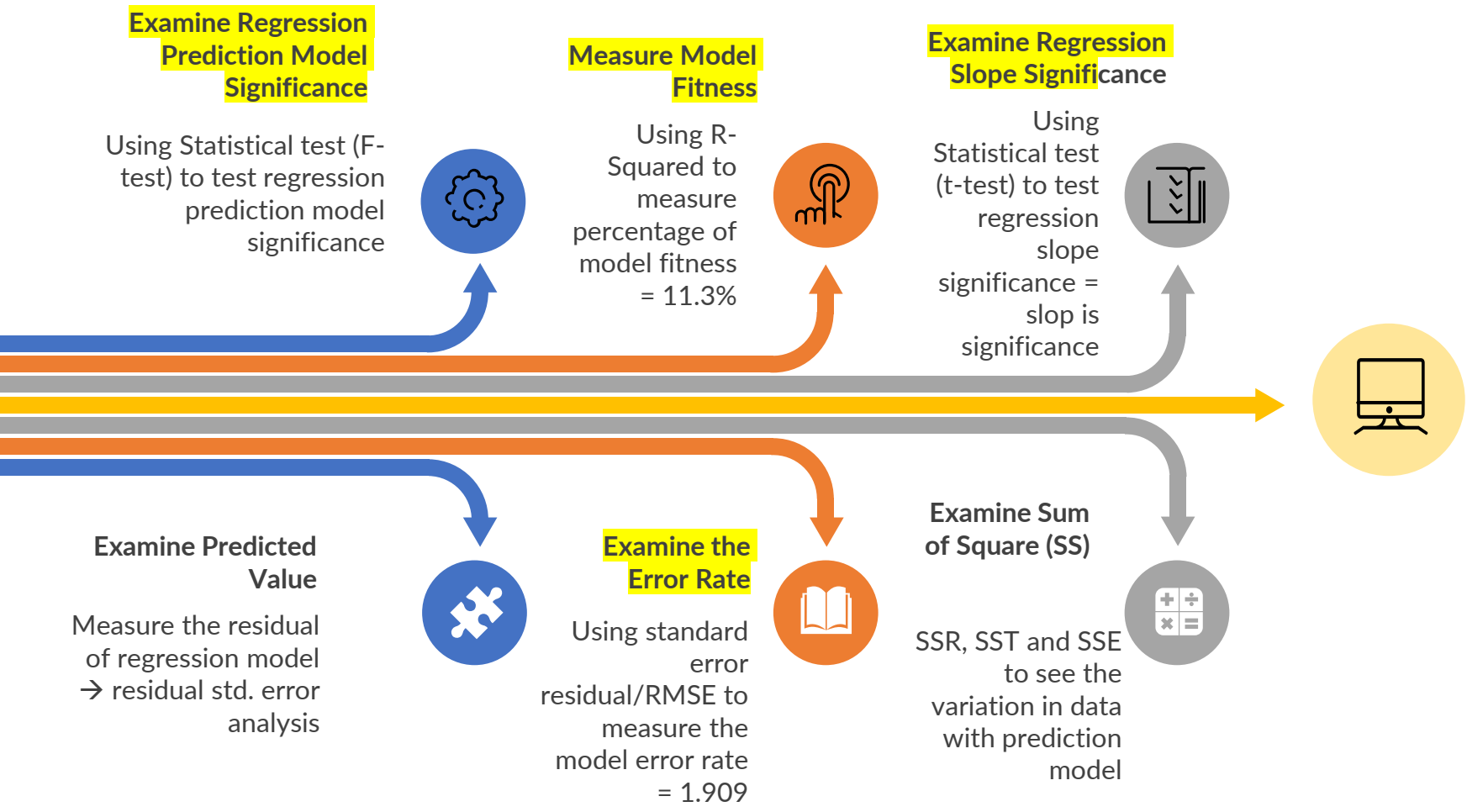
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	16.476	.286		57.557	<.001
	Physical_Activity_Score	-.233	.046	-.336	-5.019	<.001

a. Dependent Variable: Recovery_Time (days)



PROCESS IN REGRESSION MODEL DIAGNOSTIC



TAKE HOME EXERCISE: SIMPLE REGRESSION ANALYSIS

PREDICTING PARTICIPATION IN HEALTH CAMPAIGNS

The Ministry of Health wants to understand how **health literacy** influences public participation in national health campaigns (e.g., vaccination, screening, or fitness drives). This helps in designing better awareness materials and outreach strategies.

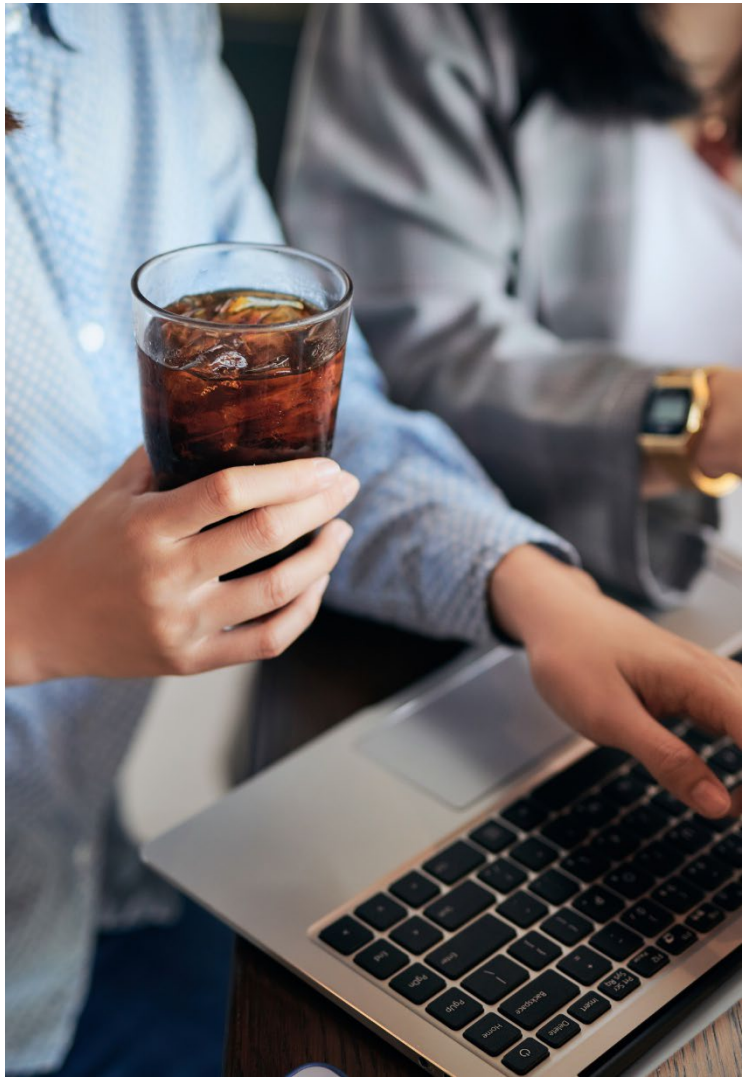
Data Set: Health_Campaign_Participation.xlsx

📂 **Dataset Description:**

- Health_Literacy_Score (0–100 scale): Measures an individual's understanding of health-related information.
- Participation_Rate_% (0–100): Percentage likelihood of participating in health campaigns.

Question:

1. Is there a positive relationship between Health Literacy Score and Participation Rate?
2. What is the slope of the regression line?
3. How strong is the relationship (R^2 value)?
4. Is the model statistically significant (p-value)?
5. Can we predict future participation rates given a health literacy score?



THANK YOU