Regression

Simple Linear Regression

Simple Linear Regression

أكاديمية طويق
Tuwaiq Academy

$$\hat{y} = b_0 + b_1 X_1$$

Dependent variable

y-intercept (constant)

Slope coefficient

Independent variable

Simple Linear Regression

أكاديمية طويق
Tuwaiq Academy

Linear regression tries to model the relationship between an independent variable (x1, e.g., Nitrogen Fertilizer) and a dependent variable (y, e.g., Potato Yield) using a straight line.

Points in the Plot:

- Each blue dot represents a separate observation or harvest.
- The model attempts to fit the best line that minimizes the error (distance) between the actual data points and the predicted values.

Regression Line:

- The grey line is the fitted regression line.
- It shows the predicted relationship between x1 (fertilizer) and y (yield).

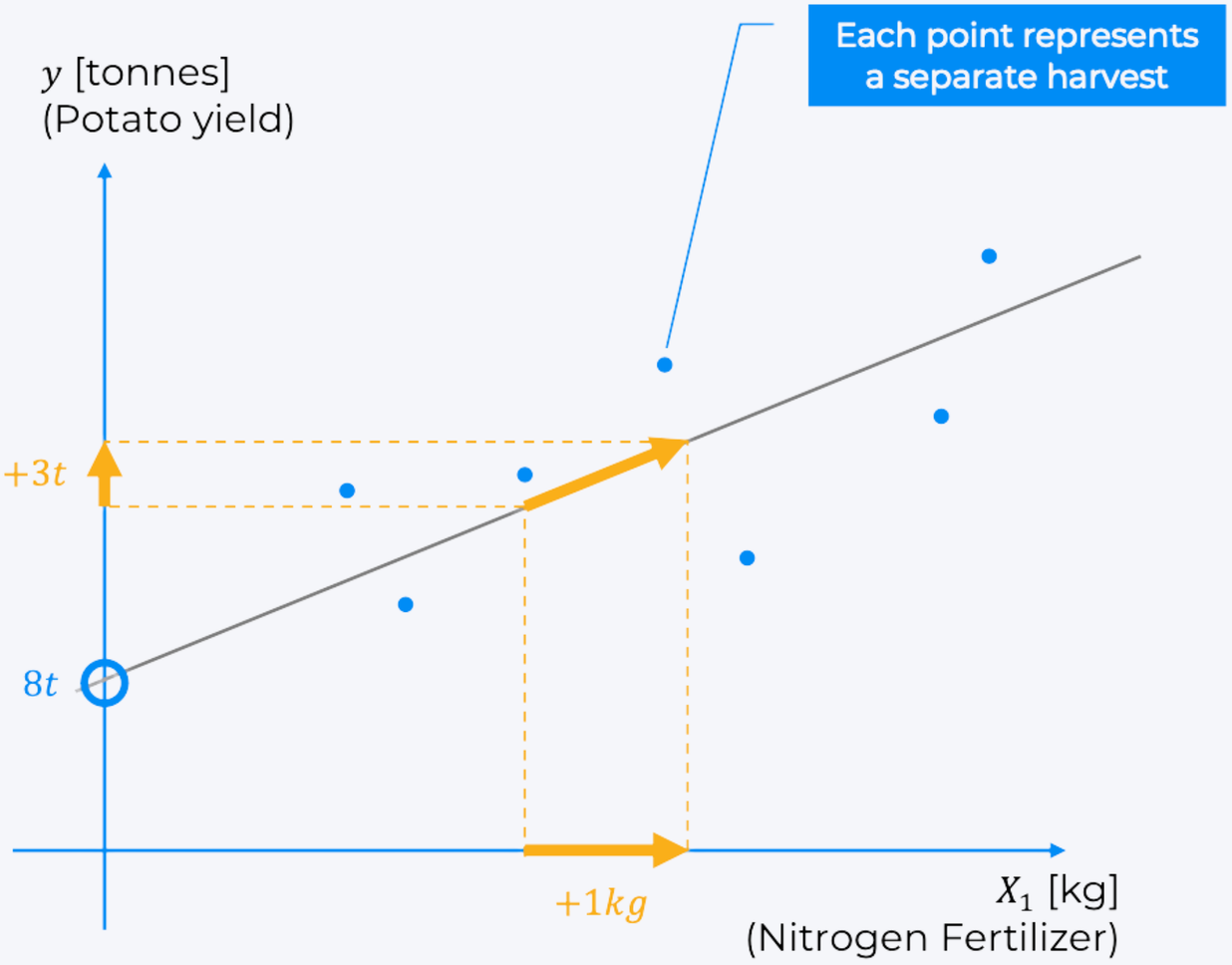
Slope of the Line:

- The slope (+3t) indicates that **for every 1 kg increase in Nitrogen Fertilizer, the Potato Yield increases by 3 tonnes.**
- This is the coefficient of x1 in the regression equation:

y=8+3x1

Intercept (Baseline Value):

- The intercept (8t) represents the predicted Potato Yield when no fertilizer (x1=0) is used.



Simple Linear Regression

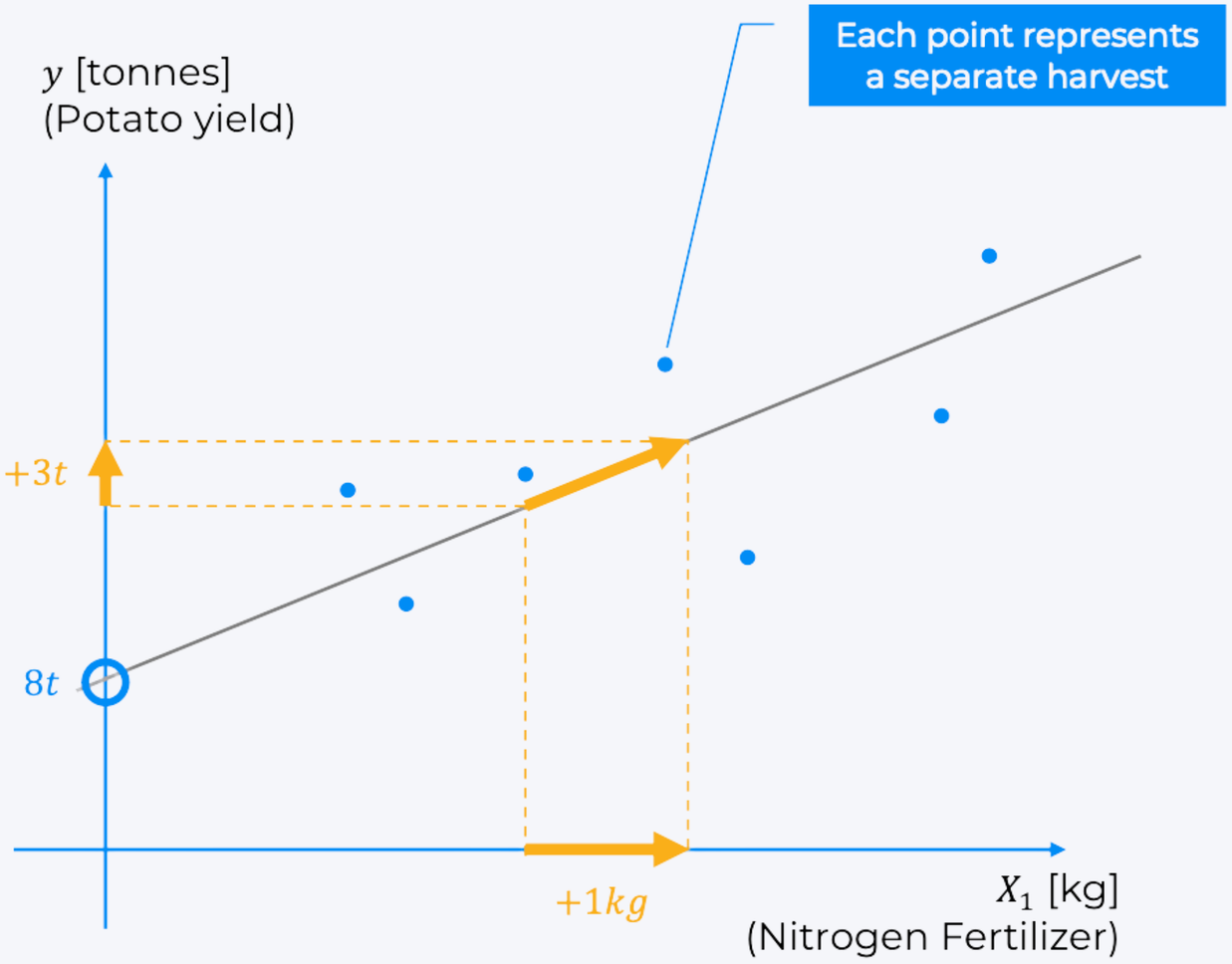
أكاديمية طويق
Tuwaiq Academy

$$\hat{y} = b_0 + b_1 X_1$$

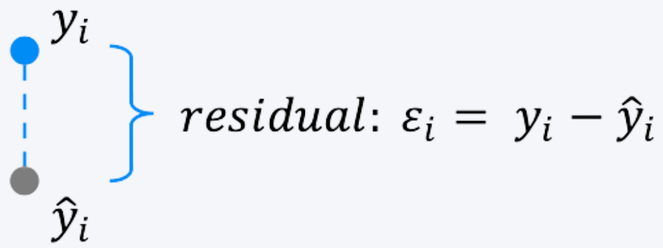
$$Potatoes[t] = b_0 + b_1 \times Fertilizer[kg]$$

$$b_0 = 8[t]$$

$$b_1 = 3[\frac{t}{kg}]$$



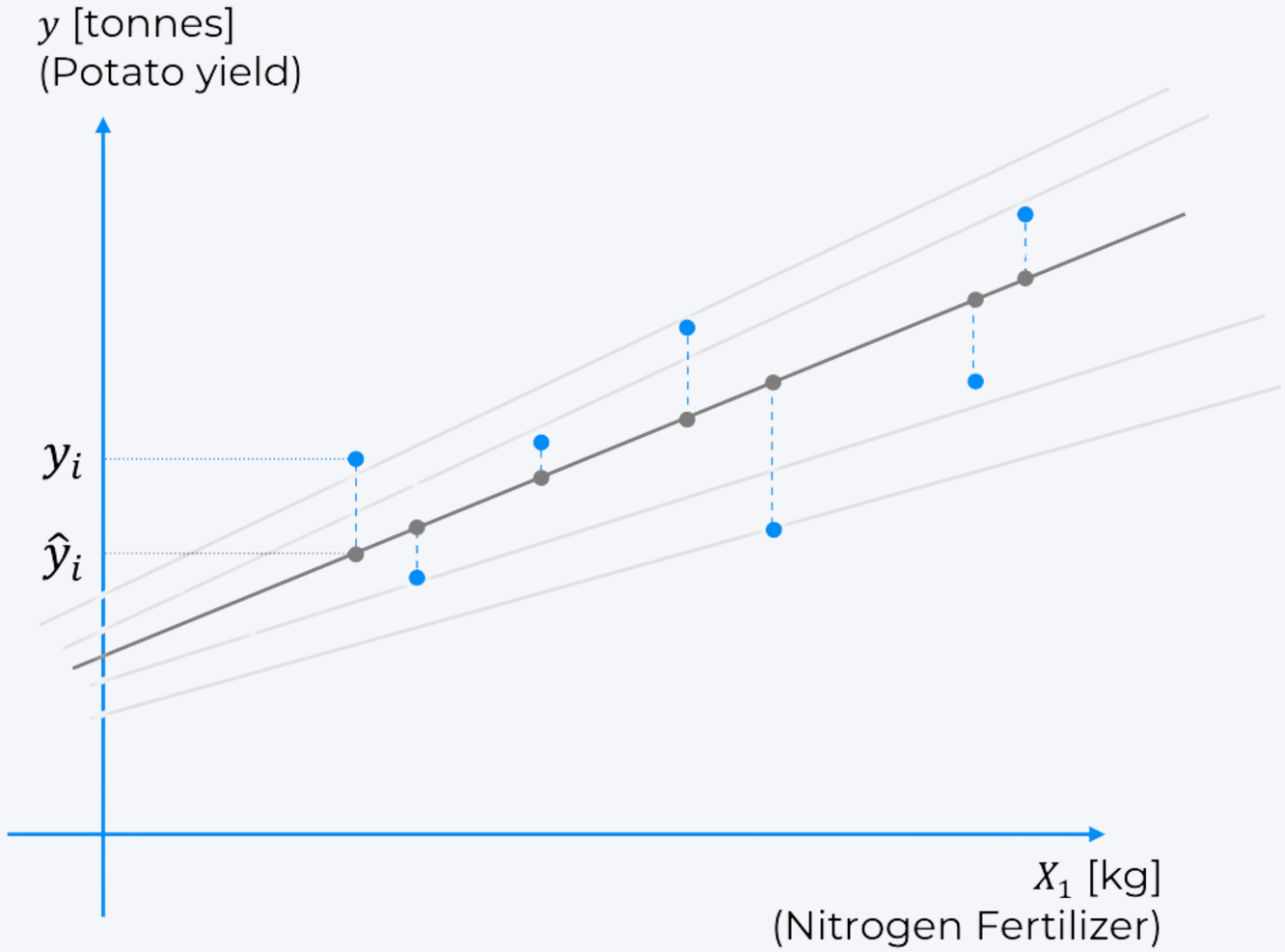
Ordinary Least Squares



$$\hat{y} = b_0 + b_1 X_1$$

b_0, b_1 such that:

$SUM(y_i - \hat{y}_i)^2$ is minimized



Multiple Linear Regression

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

- $\hat{y}$: Dependent variable
- b_0: y-intercept (constant)
- b_1: Slope coefficient 1
- X_1: Independent variable 1
- b_2: Slope coefficient 2
- X_2: Independent variable 2
- b_n: Slope coefficient n
- X_n: Independent variable n

Multiple Linear Regression

أكاديمية طويق
Tuwaiq Academy

Definition: Multiple linear regression models the relationship between one dependent variable (y) and two or more independent variables (x1,x2,…,xn), using the equation:

$$Potatoes[t] = 8t + 3\frac{t}{kg} \times Fertilizer[kg] - 0.54\frac{t}{^\circ C} \times AvgTemp[^\circ C] + 0.04\frac{t}{mm} \times Rain[mm]$$

Purpose: It predicts the dependent variable (y) based on multiple factors (independent variables), allowing a more comprehensive analysis of real-world scenarios.

R Squared

R Squared

أكاديمية طويق
Tuwaiq Academy

R-Squared is a statistical measure that explains how well the regression line fits the observed data.

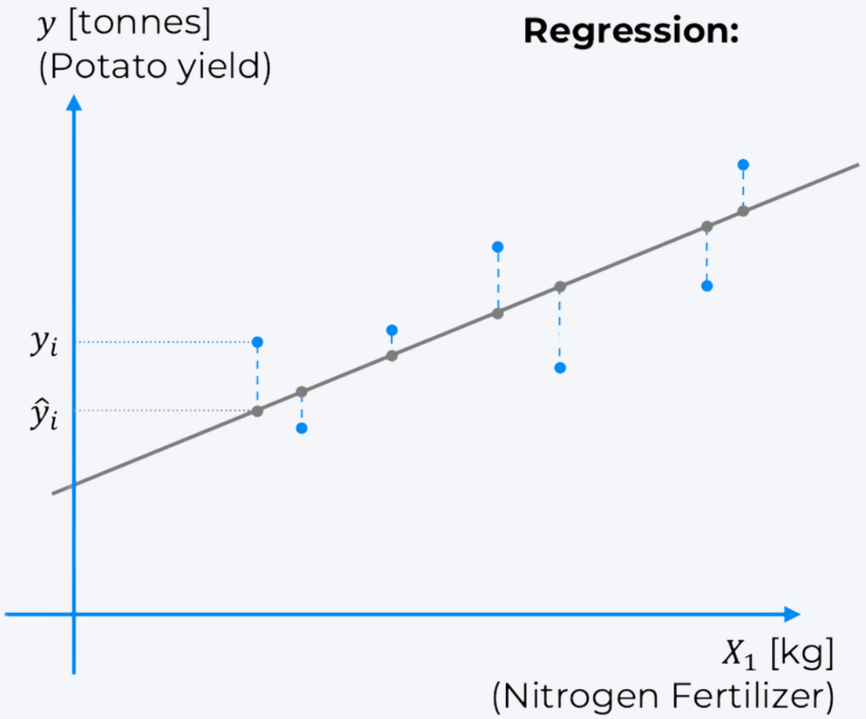
- It represents the proportion of variance in the dependent variable (y) that is predictable from the independent variable (x1).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

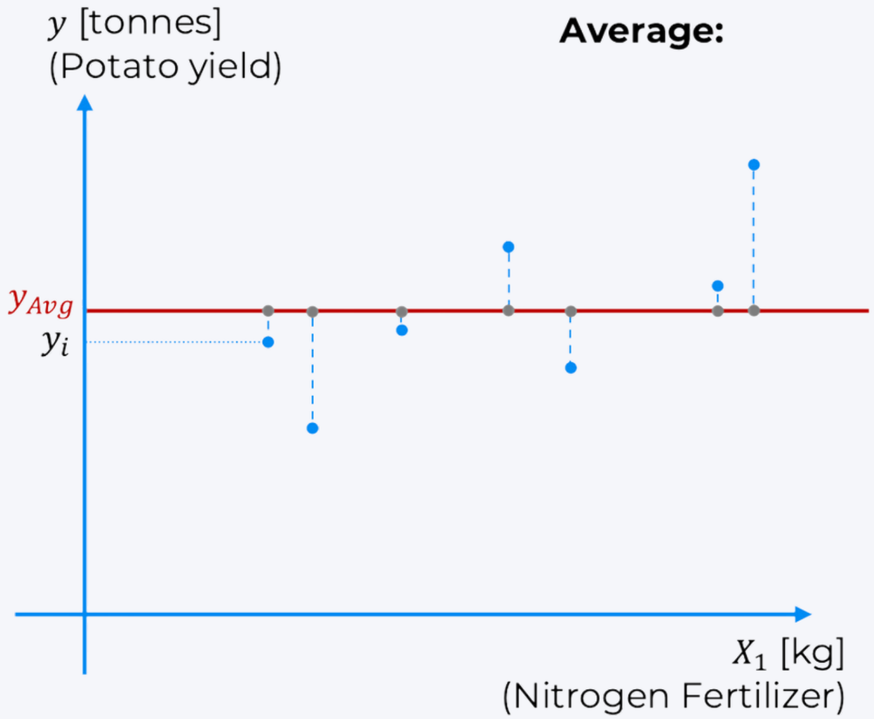
Rule of thumb (for our tutorials)*:
1.0 = Perfect fit (suspicious)
~0.9 = Very good
<0.7 = Not great
<0.4 = Terrible
<0 = Model makes no sense for this data

- **SSres:** Residual Sum of Squares (difference between actual values yi and predicted values y^i).
- **SStot:** Total Sum of Squares (difference between actual values yi and the mean of y (yavg)).

R Squared



$$SS_{res} = SUM(y_i - \hat{y}_i)^2$$



$$SS_{tot} = SUM(y_i - y_{avg})^2$$

Adjusted R Squared

أكاديمية طويق
Tuwaiq Academy

Problem with R-Squared:

- R-Squared **always increases or stays the same when you add more predictors** (x3,x4) to the model, even if the new predictors don't contribute meaningful information.

- This happens because adding predictors reduces **SSres** (Residual Sum of Squares) or keeps it the same, but **SStot** remains constant.

Issue with Overfitting:

- Adding irrelevant predictors makes the model more complex without improving its performance or explanatory power, leading to overfitting.

- **R-Squared** cannot penalize for unnecessary predictors, so it can give a false impression of better performance.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

Problem:

$\hat{y} = b_0 + b_1X_1 + b_2X_2$ ← $+ b_3X_3$

SS_{tot} doesn't change
SS_{res} will decrease or stay the same

$$SS_{res} = SUM(y_i - \hat{y}_i)^2$$

(This is because of Ordinary Least Squares: SS_{res}-> Min)

Adjusted R Squared

أكاديمية طويق
Tuwaiq Academy

Solution: Adjusted R-Squared:

- Adjusted R-Squared penalizes for adding predictors that don't improve the model.
- It provides a more realistic measure of how well the model explains the variability in the data.

Solution:

$$Adj\ R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-k-1}$$

k – number of independent variables
n – sample size

Assumptions Of Linear Regression

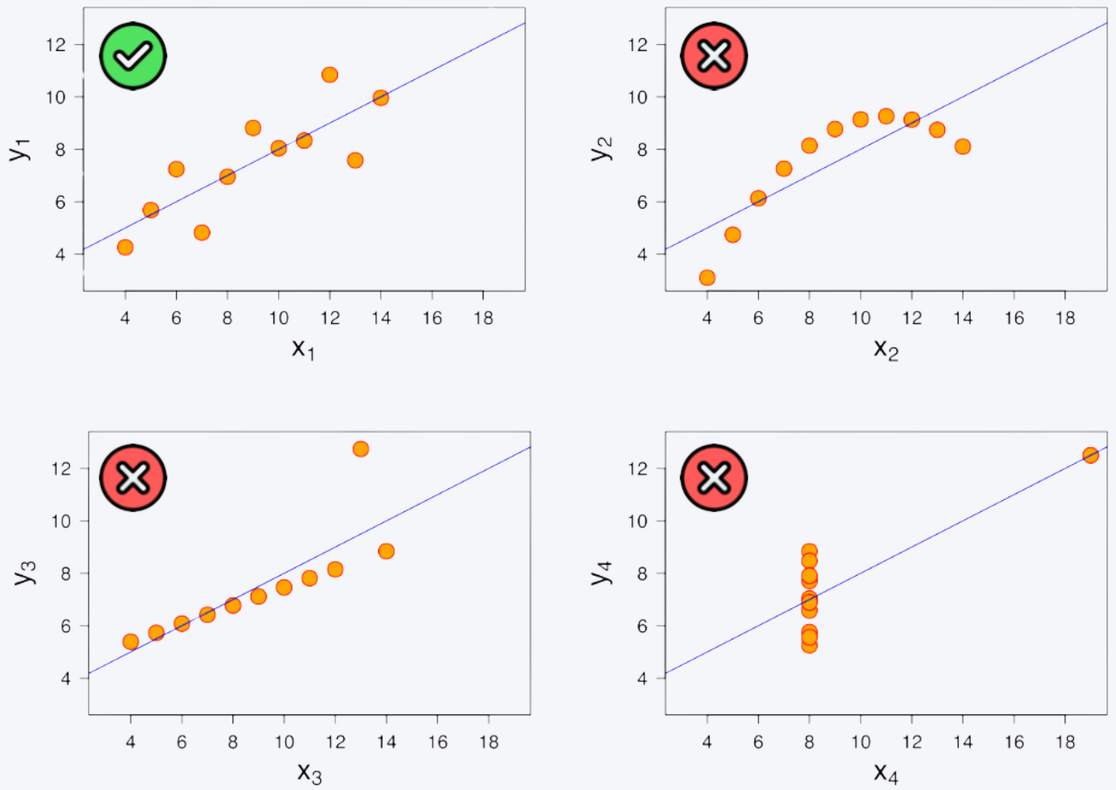
Assumptions Of Linear Regression

أكاديمية طويق
Tuwaiq Academy

Linear regression relies on several key assumptions to ensure accurate predictions and meaningful results.

linearity assumes a straight-line relationship between the dependent variable (y) and each independent variable (x1,x2,...).

If the relationship is **non-linear**, linear regression will not capture it correctly.



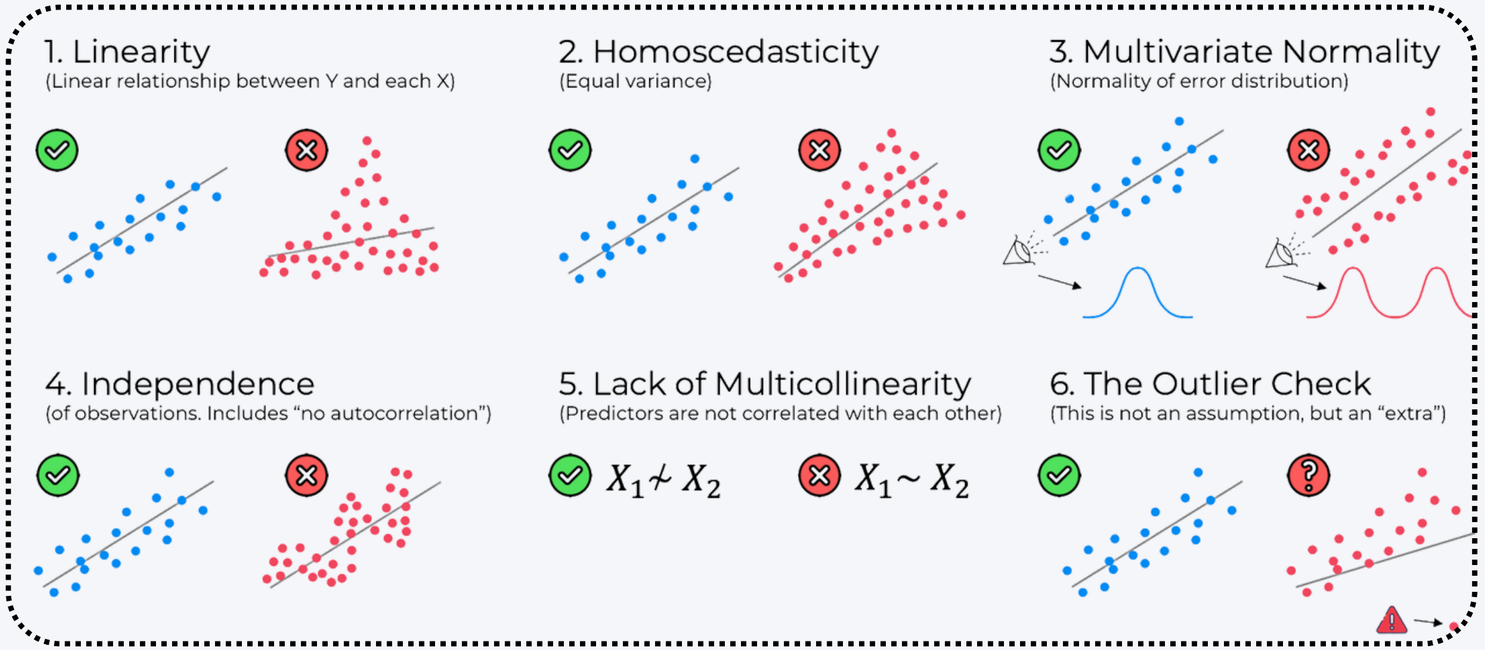
Assumptions Of Linear Regression

أكاديمية طويق
Tuwaiq Academy

Linear regression works best when certain conditions are met, ensuring accurate and reliable results.

These conditions include:

1. **Linear Relationship:** The dependent variable (y) should have a straight-line relationship with the independent variables (x1,x2, etc.).
2. **Equal Spread of Errors:** The variation in prediction errors should stay consistent across all values of the independent variables.
3. **Normal Distribution of Errors:** The errors (differences between actual and predicted values) should follow a normal distribution.
4. **Independence**: Each observation in the data should be independent of the others.
5. **No Strong Correlation Between Variables**: Independent variables should not be too closely related to each other, as it can confuse the model.
6. **Check for Outliers:** Outliers can heavily influence the regression line and should be addressed if found.



If these conditions aren't met, the model might not provide the best predictions, and adjustments or different techniques may be needed.

Dummy Variables

Dummy Variables

أكاديمية طويق
Tuwaiq Academy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + \ ???$$

Dummy Variables

أكاديمية طويق
Tuwaiq Academy

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 \quad b_3*x_3 \quad + b_4*D_1$$

Dummy Variables Trap

Dummy Variables Trap

أكاديمية طويق
Tuwaiq Academy

Dummy Variables:

- "**State**" is converted into two dummy variables:
  - **New York (D_1):** 1 if the state is New York, 0 otherwise.
  - **California (D_2)**: 1 if the state is California, 0 otherwise.

Dummy Variable Trap:

- Notice that the **two dummy variables are highly dependent:**
  - if $D_1 = 1$ (New York), D_2 must be 0 (California), and vice versa.

This creates a problem of multicollinearity in the regression model because one dummy variable can be perfectly predicted from the other (e.g., D2=1−D1).

Multicollinearity makes it hard for the model to estimate coefficients correctly.

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70			ia	0	1
191,050.39	153,441.51			ia	0	1
182,901.99	144,372.41			ork	1	0
166,187.94	142,107.34			ia	0	1

$D_2 = 1 - D_1$

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

Dummy Variables Trap

أكاديمية طويق
Tuwaiq Academy

Solution:

- To avoid the trap, **drop one dummy variable (e.g., D_2) from the model.**

Now, the model only uses D_1 to represent the "State" variable, while the dropped category (e.g., California) becomes the baseline.

- For example:
  - **If $D_1 = 1 \rightarrow$ State is New York.**
  - **If $D_1 = 0 \rightarrow$ State is California (by default).**

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

Always omit one dummy variable

Regression Equation:

- The equation includes:
  - Numerical variables (x1,x2,x3) like R&D Spend, Admin, Marketing.
  - A single dummy variable (D1) representing the state.

- This avoids the dummy variable trap while still capturing the effect of the state on profit.

Building A Model

Building A Model

The idea here is to explain **different methods of building regression models** by selecting the most **relevant predictors** (independent variables).

These methods aim to **balance simplicity and accuracy in the model**, ensuring only significant variables are included

Building A Model

All in one

Use all the predictors without elimination.

- When to use:
    - You have prior knowledge that all variables are important.
    - It's required to include all variables (e.g., for regulatory reasons).
    - You're preparing for Backward Elimination and want to start with all predictors.

Building A Model

Backward Elimination

Start with all predictors and **remove the least significant one (based on p-value)** until only significant predictors remain.

Steps:

1. Set a significance level (e.g., SL=0.05).
2. Fit a full model with all predictors.
3. Remove the predictor with the highest p-value if p>SL, then refit the model.
4. Repeat until all remaining predictors have p<SLp<SL.

Useful when starting with many predictors and you want to eliminate irrelevant ones.

Building A Model

Forward Selection

Start with no predictors and **add the most significant one step-by-step.**

Steps:

1. Set a significance level for inclusion (e.g., SL=0.05).
2. Fit models with each predictor separately and select the one with the **lowest p-value.**
3. Add the selected variable and repeat by testing additional predictors one by one.
4. Stop when no remaining variable has p<SL.

Useful when starting with no predictors and gradually adding relevant ones.

Building A Model

Bidirectional Elimination

Combine F**orward Selection and Backward Elimination** by adding and removing predictors dynamically.

Steps:

1. Set significance levels for adding (SLENTER) and removing (SLSTAY).
2. Add new predictors with **p<SLENTER** (like Forward Selection).
3. Remove existing predictors with **p>SLSTAY** (like Backward Elimination).
4. Repeat until no variables can be added or removed.

Useful when you want a flexible approach that checks both directions.

Building A Model

Why SL=0.05 (Significance Level) ?

The significance level (SL) is a threshold used in statistical tests to decide whether a variable is significant enough to be included in a regression model.

Standard Practice (Default Value):

- **SL=0.05:** This is a common **default** value used in most statistical analyses.
    - It corresponds to a **5% risk of incorrectly rejecting a variable that is actually significant (Type I error).**
    - Widely accepted in fields like science, engineering, and social sciences.

Building A Model

Why SL=0.05 (Significance Level) ?

Based on Desired Confidence:

- Relationship to Confidence Level:
- **SL=1−Confidence Level.**

For example:

- SL=0.05SL=0.05 → 95% Confidence Level.
- SL=0.01SL=0.01 → 99% Confidence Level (more stringent).
- SL=0.10SL=0.10 → 90% Confidence Level (less stringent).

Domain-Specific Considerations:

- **High-Stakes Decisions**: Use a smaller SL (e.g., SL=0.01) in fields where errors are costly, such as medicine, finance, or safety-critical industries.

- **Exploratory Analysis:** Use a larger SL (e.g., SL=0.10) if you're exploring data and want to include more predictors for further analysis.

Building A Model

أكاديمية طويق
Tuwaiq Academy

Why SL=0.05 (Significance Level) ?

Size of Dataset and Model Complexity:

- **Large Datasets:** A smaller SL (e.g.,SL=0.01) might be more appropriate since larger datasets can detect even small effects.

- **Small Datasets:** A larger SL (e.g.,SL=0.10) might be acceptable because small datasets may lack the power to detect subtle relationships.

Practical Tips:

- Start with **SL=0.05** as a general rule.

Adjust it based on:

  - The importance of avoiding errors in your context.
  - The size and complexity of your data.
  - Your domain knowledge and goals.

أكاديمية طويق
Tuwaiq Academy

Hands-On Code

Simple Linear Regression

أكاديمية طويق
Tuwaiq Academy

Hands-On Code

Multiple Linear Regression

Polynomial Regression