Regression Analysis

Regression Analysis : **Regression analysis is a statistical** technique for investigating and **modeling** the relationship between variables.

Mathematical Model:

– Equation of a straight line *y = mx +b*

We usually write this $$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \varepsilon$$

where ε represents **error**

- it is a random variable that accounts for the failure of the model to fit the data *exactly.*

Simple Linear Regression

Simple Linear Regression Model :

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

y – Dependent (response) variable

x – Independent (regressor/predictor) variable

β0 – **Intercept: if *x = 0 is in the range,*** β0 is the expected value of the response *y, when x = 0;*

β1 – **Slope: change in the expected value of the response produced** by a unit change in *x.*

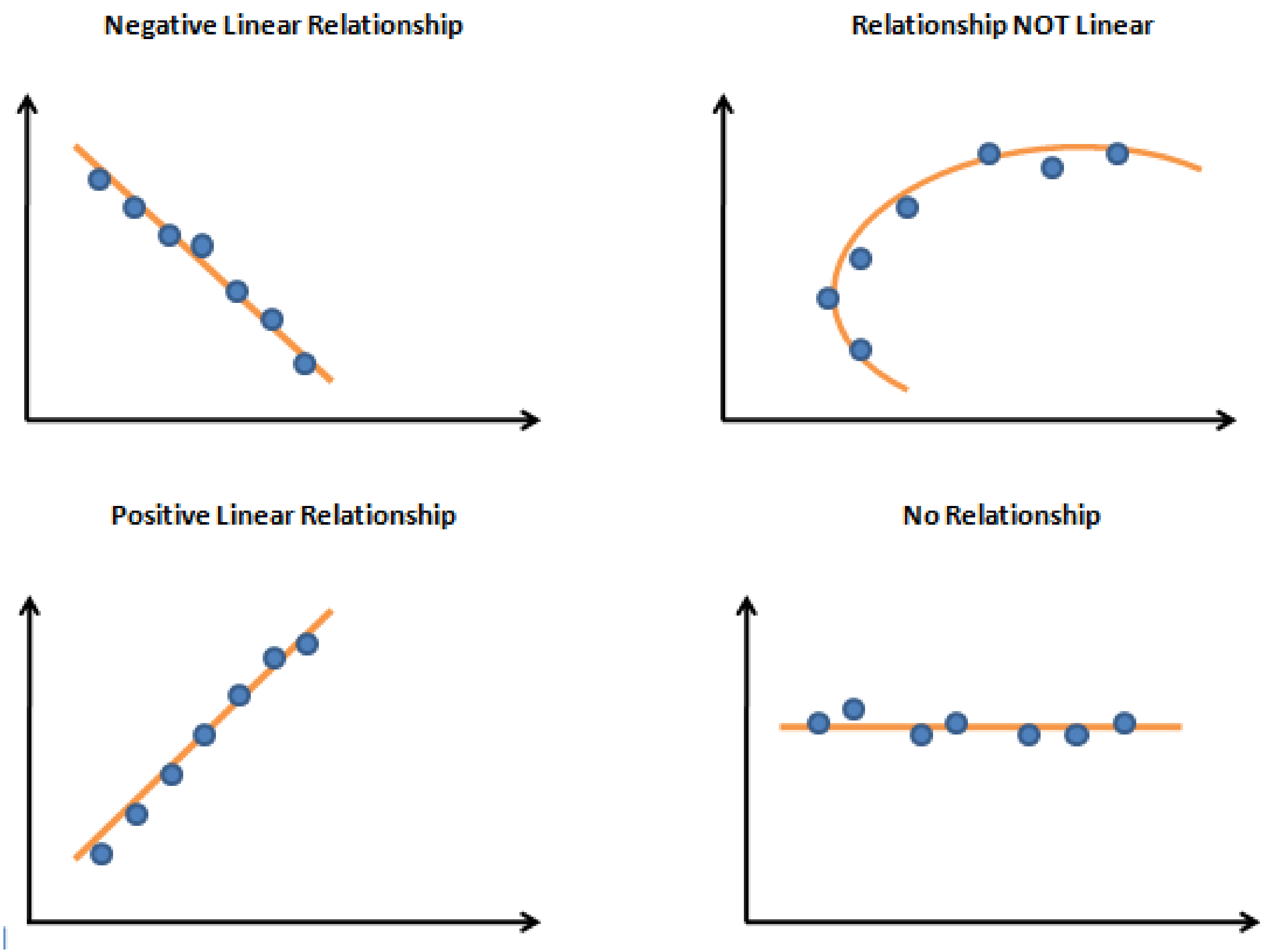
ε - Random error term

Simple Linear Regression

Here are some examples of Linear Regression

- We use product price to predict the number of sales.
- We predict annual sales from advertising budget.
- We use rainfall amount to predict the fruits yield.
- We use Parent height to predict child height.
- We use sales-rep commission to predict products sales.

Type of Linear Regression

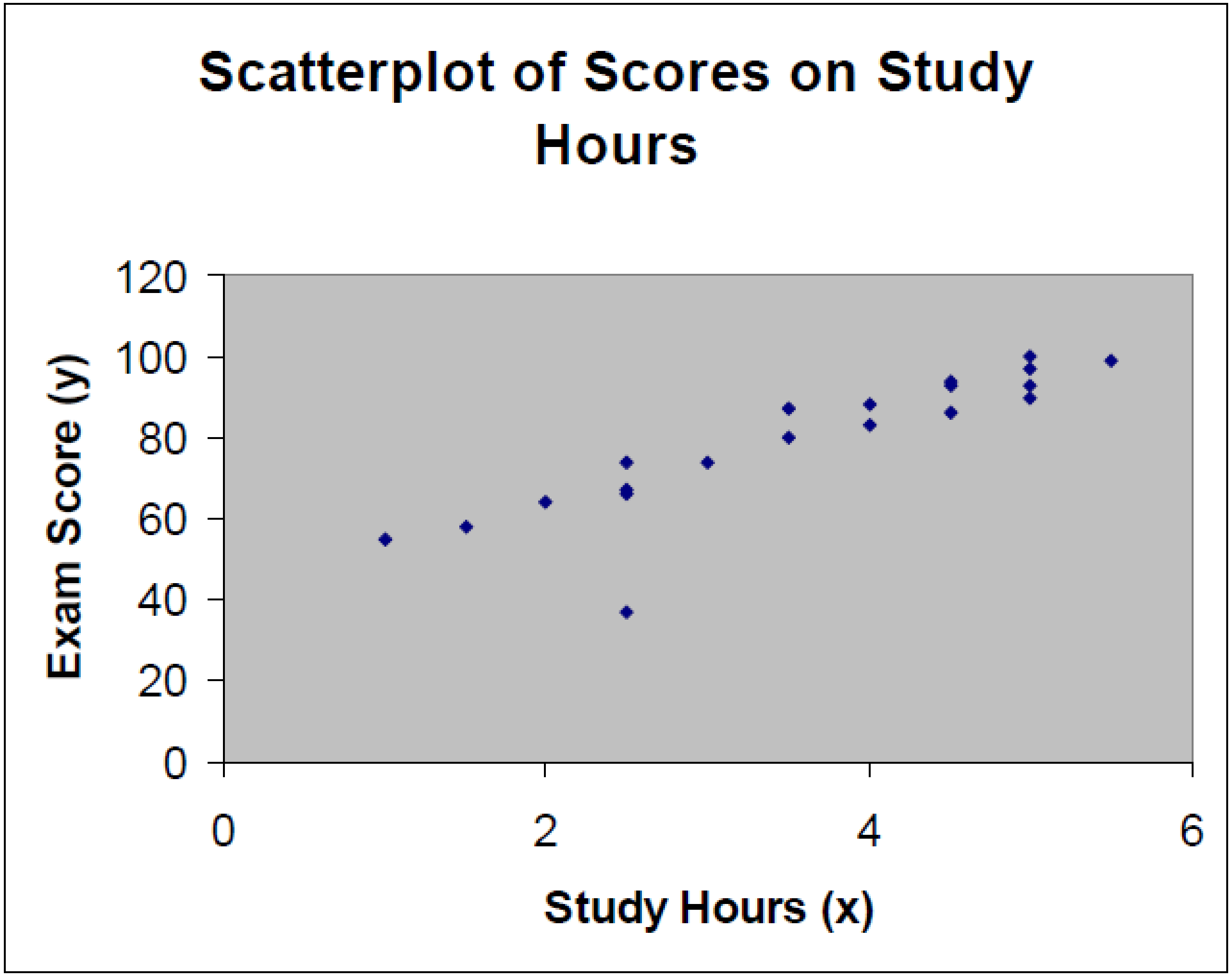


Simple Linear Regression

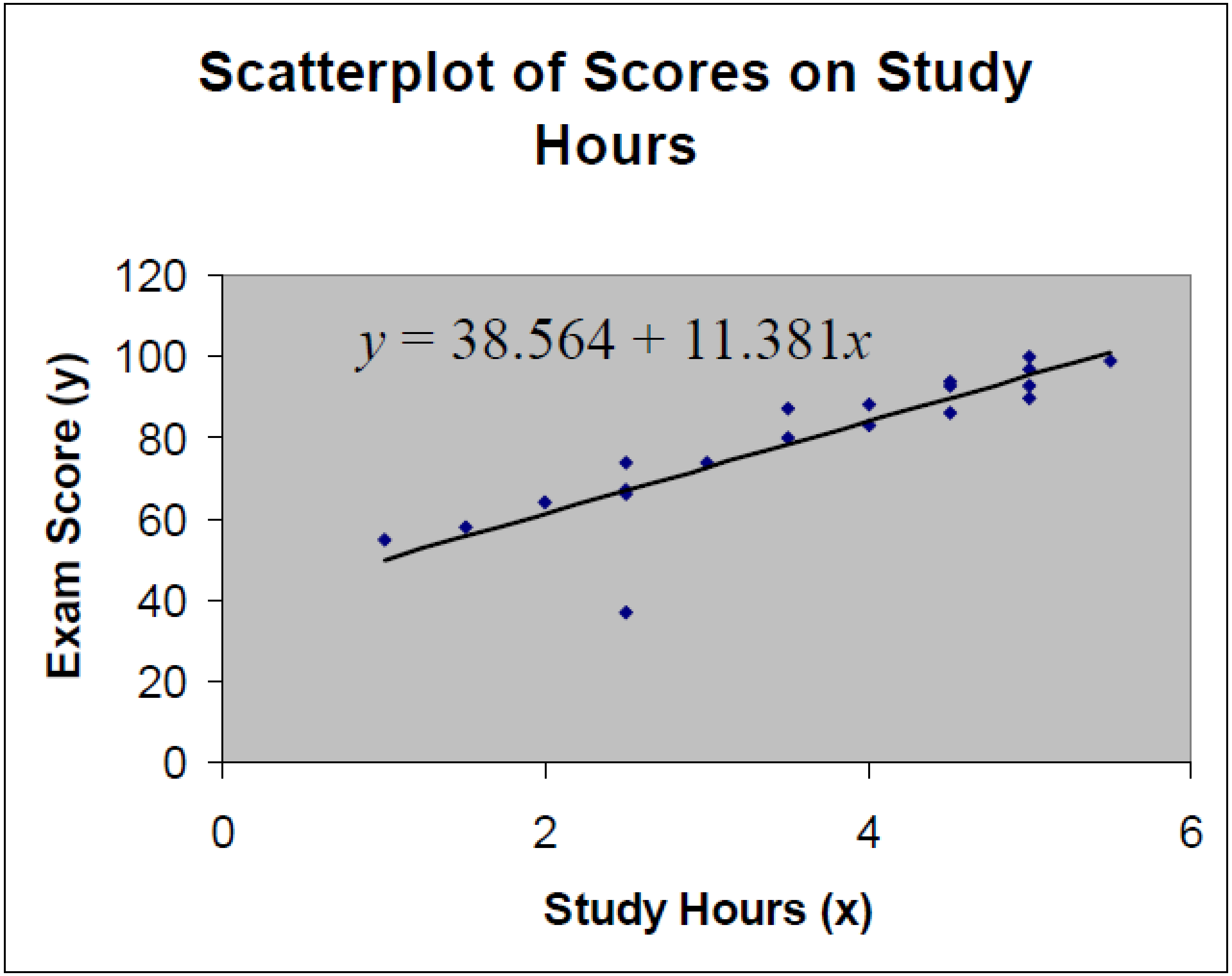
Example : Use **study hours** to predict **test score**

		y	x
		Response	**Regressor / Predictor**
		Dependent Variable	**Independent Variable**
		Exam Score	**Study Hours**
Data Point 1	Student 1	93	4.5
Data Point 2	Student 2	37	2.5
Data Point 3	Student 3	93	5
Data Point 4	Student 4	67	2.5
Data Point 5	Student 5	87	3.5
Data Point 6	Student 6	100	5
Data Point 7	Student 7	90	5
Data Point 8	Student 8	94	4.5
Data Point 9	Student 9	88	4
Data Point 10	Student 10	74	2.5
Data Point 11	Student 11	99	5.5
Data Point 12	Student 12	74	3
Data Point 13	Student 13	64	2
Data Point 14	Student 14	97	5
Data Point 15	Student 15	83	4
Data Point 16	Student 16	55	1
Data Point 17	Student 17	80	3.5
Data Point 18	Student 18	58	1.5
Data Point 19	Student 19	86	4.5
Data Point 20	Student 20	66	2.5

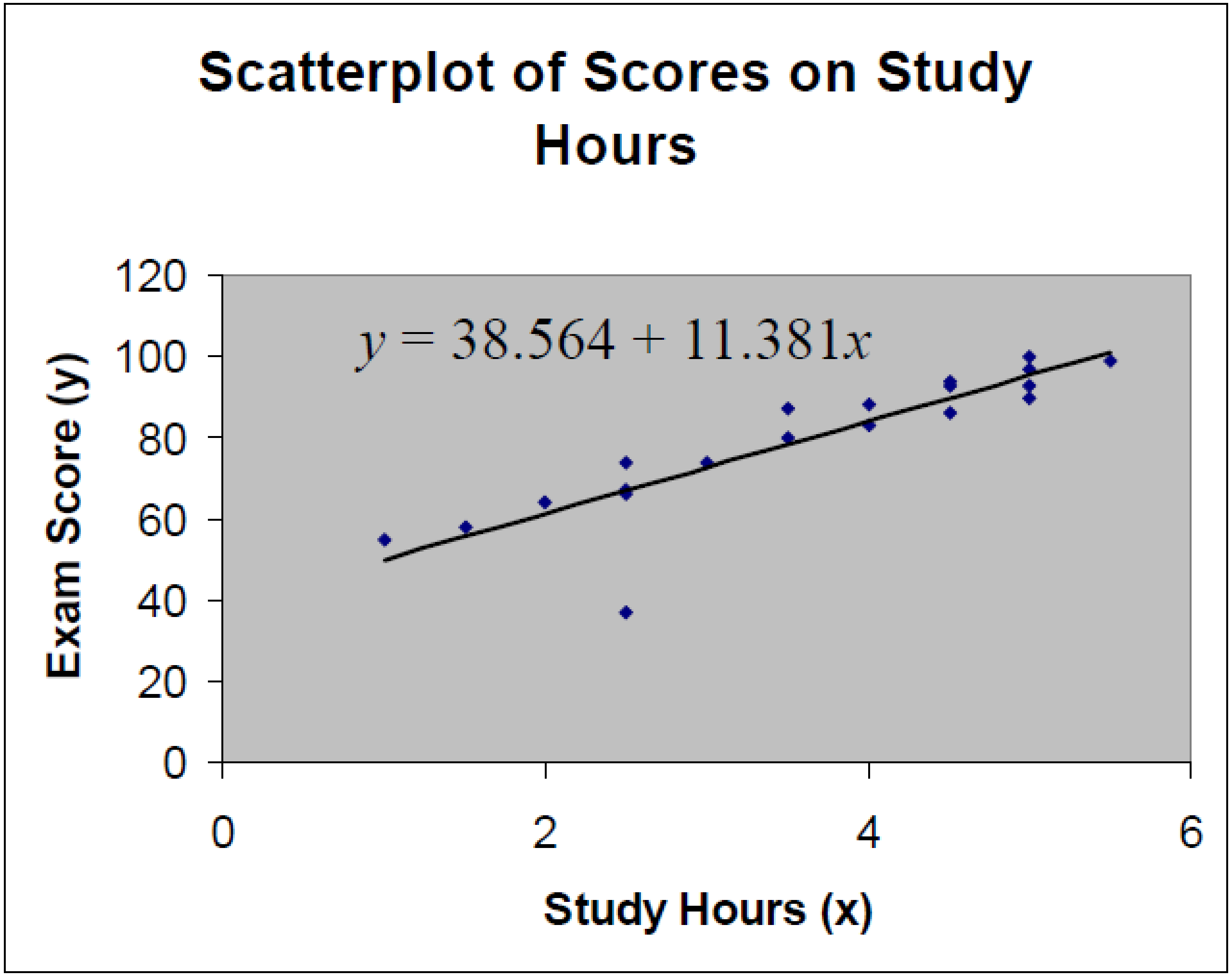
Simple Linear Regression



Simple Linear Regression



Simple Linear Regression

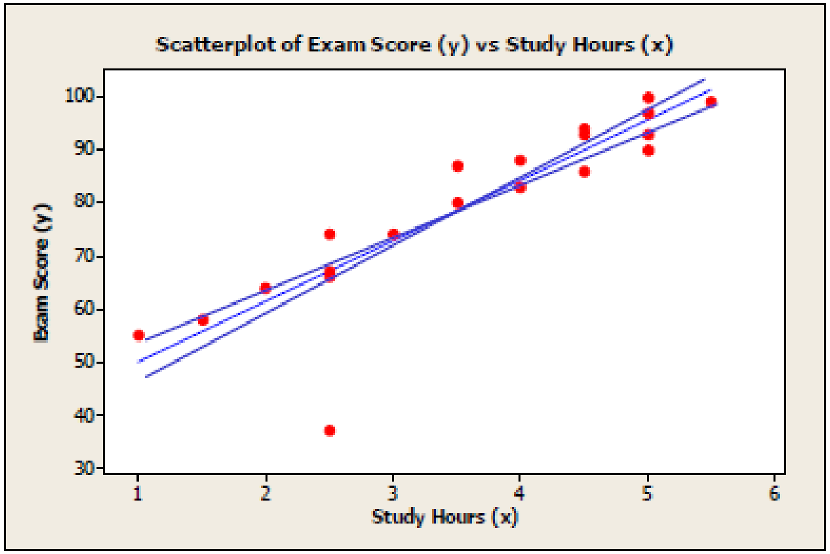


Parameter Estimation

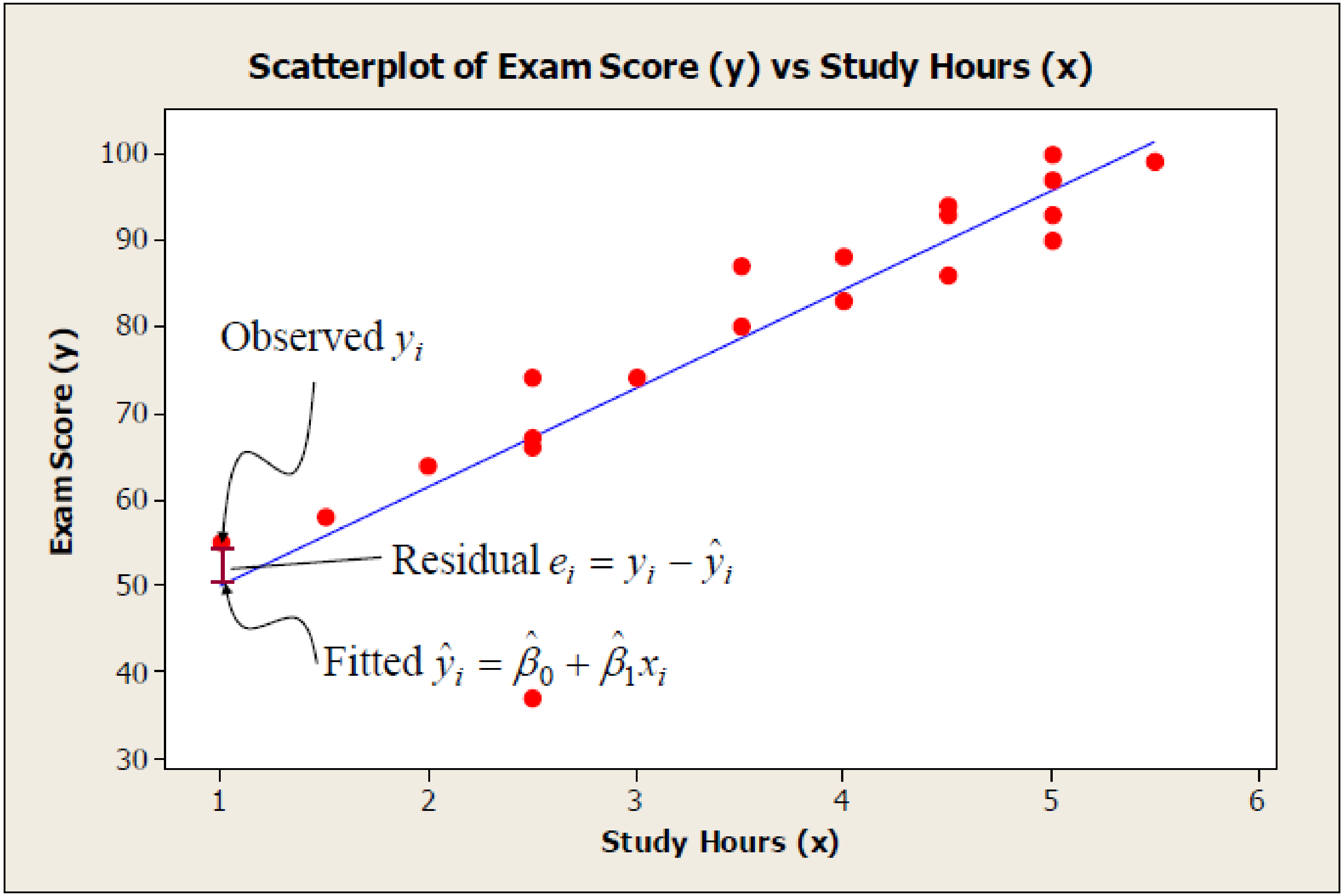
Parameter Estimation

β_0 – **intercept**
β_1 – **slope**

Which set of estimates is the best?
I.e., which is the best fitting line?



Parameter Estimation

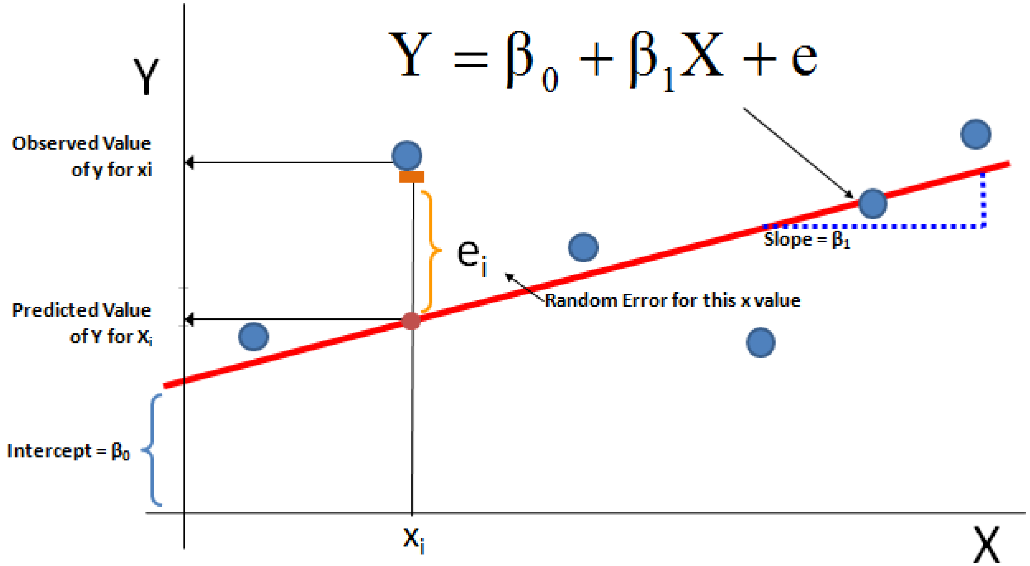


Ordinary Least-Squares Estimation

OLE seeks β0 and β1 to minimize the sum of squares of the differences between the observed response, Yi, and the straight line.

OLE solves an optimization problem to find the best straight line that fits the data.

$$
\begin{aligned}
\sum e^2 &= \sum (y - \hat{y})^2 \\
&= \sum (y - (\beta_0 + \beta_1 x))^2
\end{aligned}
$$



Ordinary Least-Squares Estimation

- Least-squares criteria:

$$\min \sum_{i-1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Based on this criteria, Gauss says the following least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ best fit the data.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Simple Linear Regression Example

Regression Model - Regressing the carbon footprint of cars versus their fuel economy in City driving conditions.

Carbon (Response) - Carbon footprint in tones per year

City_mileage (Predictor) - Fuel economy in City driving conditions in miles per gallon

Model	Cylinders	Litres	Barrels	City_mileage	Highway	Cost	Carbon
Chevrolet Aveo	4	1.6	12.2	25	34	1012	6.6
Chevrolet Aveo 5	4	1.6	12.2	25	34	1012	6.6
Chevrolet Cobalt	4	2.2	12.7	24	33	1049	6.8
Chevrolet Colorado 2WD	4	2.9	17.1	18	24	1418	9.2
Chevrolet Colorado 2WD	5	3.7	18.0	17	23	1491	9.6
Chevrolet Colorado Cab Chassis inc 2WD	5	3.7	20.1	15	20	1667	10.8
Chevrolet Colorado Crew Cab 2WD	4	2.9	17.1	18	24	1418	9.2
Chevrolet Colorado Crew Cab 2WD	5	3.7	18.0	17	23	1491	9.6
Chevrolet HHR FWD	4	2.0	14.9	19	29	1233	8.0
Chevrolet HHR Panel FWD	4	2.0	14.9	19	29	1233	8.0
Chevrolet Malibu	4	2.4	13.2	22	33	1091	7.1
Chevrolet Malibu	4	2.4	13.7	22	30	1134	7.3
Chevrolet Malibu Hybrid	4	2.4	11.8	26	34	978	6.3
Chrysler PT Cruiser	4	2.4	16.3	19	24	1349	8.7

Simple Linear Regression Example

Numerical Measures of Covariability

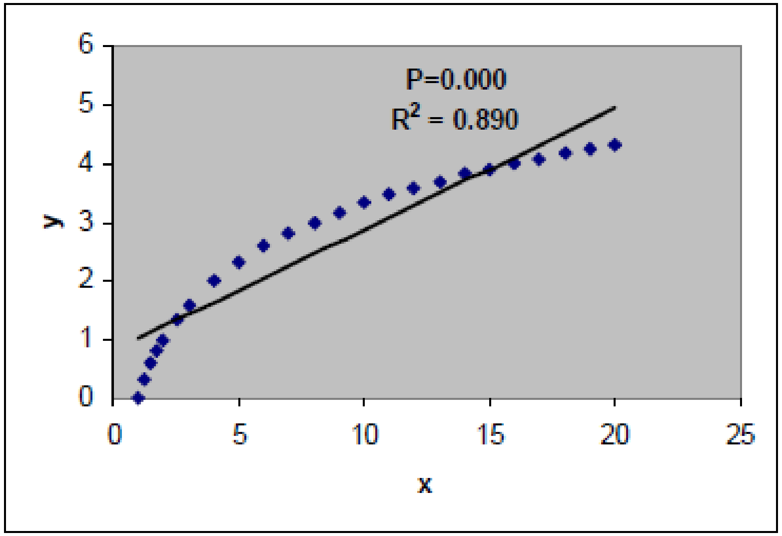
- The R^2 statistic is the correlation squared.
- It defines the fraction of the uncertainty about the y variable that is explained by the x variable.

$$R^2 = \rho_{xy}^2$$

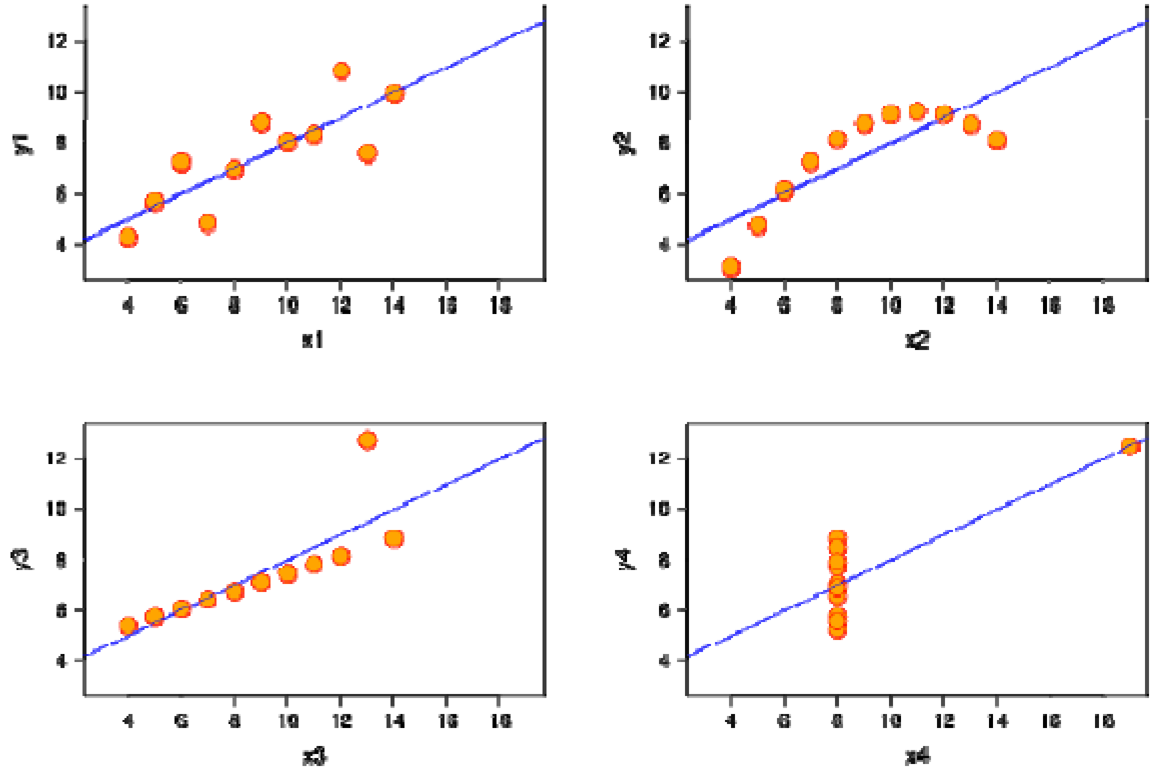
Simple Linear Regression Example

Caution

- Regression analysis is perhaps the most widely used statistical technique, and probably the most widely misused.
- Just because you *can* fit a linear model to a set of data, does not mean you should.



Simple Linear Regression Example



All four cases have significant slopes and the same R-square (0.7), and regression line ($y = 3 + 0.5x$).

Source: Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, 27, 17–21.

Simple Linear Regression Example

Model Adequacy (Diagnostic) Check

- Checking *p*-value and R^2 only is not sufficient.
- We also need to validate some underlying model assumptions:
  - Linear relationship (at least approximately).
  - The error (residual) follows a normal distribution with a (nearly) constant variance.
- Look for potential outliers.
- We need to check residual plots to validate underlying assumptions:
  1. Relationship between response and regressor is **linear** (at least approximately).
  2. Error term, $\boldsymbol{\varepsilon}$ has zero mean
  3. Error term, $\boldsymbol{\varepsilon}$ has **constant variance**
  4. Errors are **normally distributed** (required for tests and intervals)

Simple Linear Regression Example

How to know if the model is best fit for your data?

The most common metrics to look at while selecting the model are:

STATISTIC	CRITERION
R-Squared	Higher the better *(> 0.70)*
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
AIC	Lower the better
BIC	Lower the better
Mallows cp	Should be close to the number of predictors in model
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))	Higher the better

Box Cox Transformation

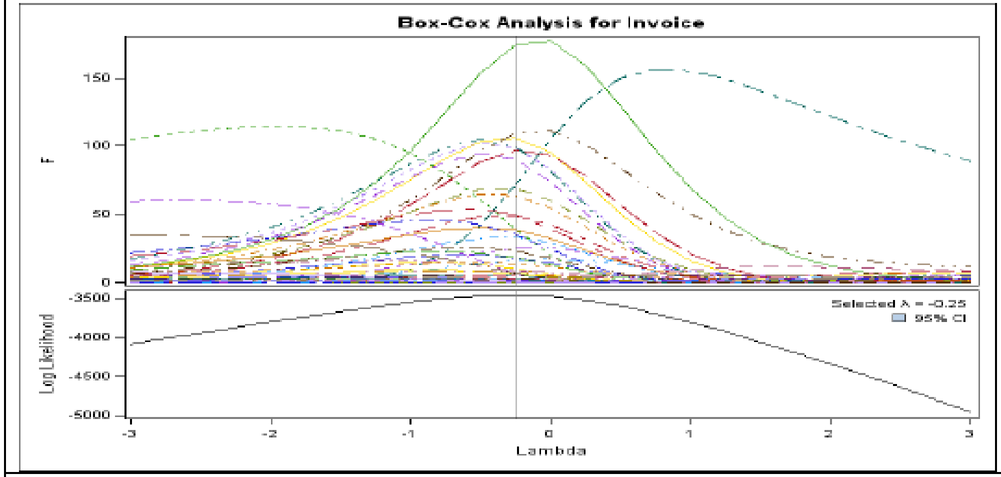
The Box-Cox transformation of the variable x is also indexed by λ, and is defined as

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda}. \quad \text{(Equation 1)}$$

OUTPUT

The TRANSREG Procedure Hypothesis Tests for BoxCox(Invoice)

Root MSE	0.01002	R-Square	0.9350
Dependent Mean	3.68434	Adj R-Sq	0.9260
Coeff Var	0.27188	Lambda	-0.2500 ❶



❶ Based on the Recommended Transformation Chart below, Lambda equals -0.2500 suggesting natural log transformation for invoice (y).

Recommended Transformation	Equation	Lambda
Square	Y^2	1.5 to 2.5
None	Y	0.75 to 1.5
Square-root	$Y^{1/2}$	0.25 to 0.75
Natural log	Ln(Y)	-0.25 to 0.25
Inverse square-root	$1/Y^{1/2}$	-0.75 to -0.25
Reciprocal	1/Y	-1.5 to -0.75
Inverse square	$1/Y^2$	-2.5 to -1.5

("Box-Cox Method")

Multiple Linear Regression

Use Case: Product Sales prediction based on advertising expenses:

The **Advertising data** set consists of the **sales of a product in 200 different markets**, along with advertising budgets for the product in each of those markets for three different media: **TV, radio, and newspaper**.
In this case the advertising budgets are input variables while sales is an output variable.

Analytics goal is to recommend the right media with advertising budgets to improve the sales of a that product by analyzing the historical data.

Predictors or Features or dependent Variable

TV : advertising budgets(in thousands of dollars) spent on TV ads for a single product in a market.
Radio : advertising budgets(in thousands of dollars) spent on Radio ads.
Newspaper : advertising budgets(in thousands of dollars) spent on Newspaper ads.

Response or Independent Variable

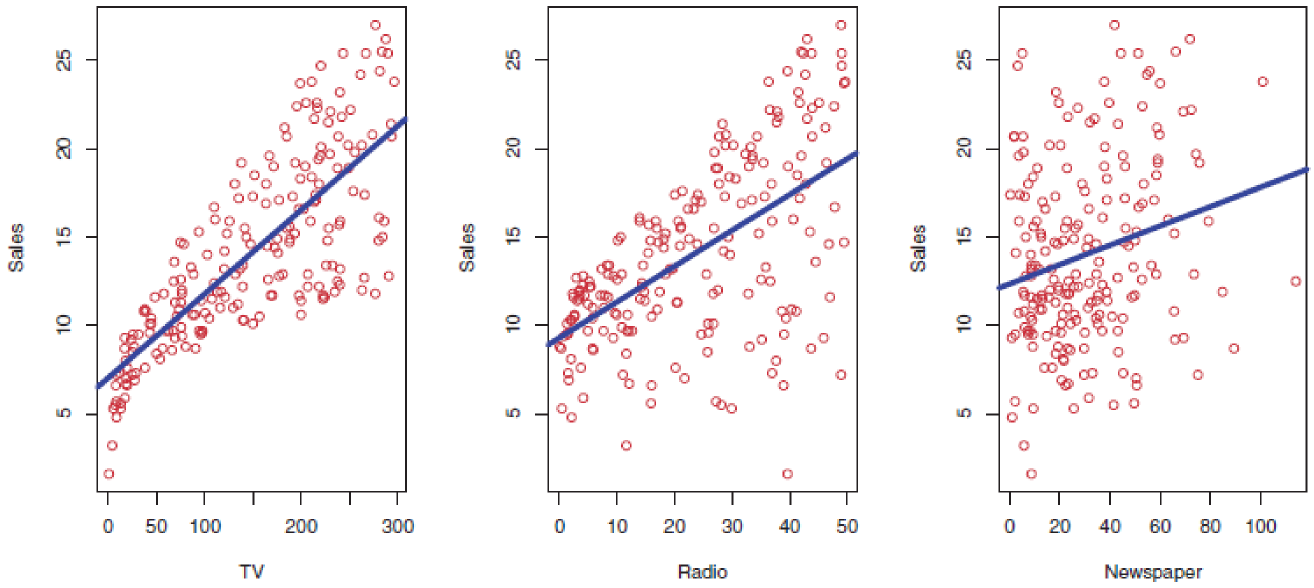
Sales : Sales of a single product in a given market (in thousands units).

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

Multiple Linear Regression

Use Case: Product Sales prediction based on advertising expenses:

Scatter plot to visualize the relationship between different advertisement and sales.



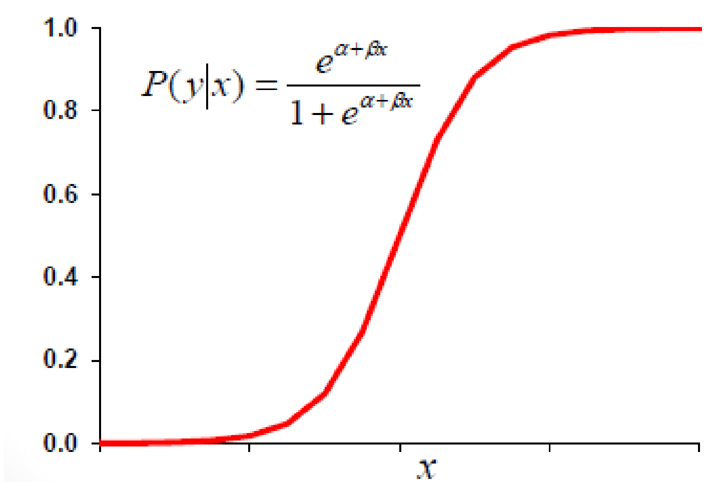
Correlation.

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

Logistic Regression Analysis

Logistic Regression Analysis : Extension of Regression analysis to the situations where the **Response variable is categorical.**

- In logistic regression, instead of using Y as the response variable, we use the logit function (odds = P/(1-P)).
- In logistic Regression two steps involved :
  (i) Estimate the probabilities belonging to each class
  (ii) Use the cutoff values on these probabilities to classify in one of the classes.
- Probability = 1 / [1 + exp(β0 + β1 X)] or $\log_e$[P/(1-P)] = β0 + β1 X
- The Function or $\log_e$[P/(1-P)] is called the logistic function.



Logistic Mathematical Model

$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\text{odds} = P/(1-P)$$

Estimating the Logistic model

- In logistic regression, the relation between Y and the beta parameters is **nonlinear.**
- The beta parameters are estimated using **Maximum Likelihood.**
- For all methods, the contribution to the model is measures by model Deviance or AIC.
- A better model will have a lower Deviance/AIC.
- Deviance is calculated from Maximum-likelihood estimation (MLE).
- MLE is an interactive procedure that successively tries works to get closer and closer to the correct answer.
- A perfect model will have MLE = 0.
- **Cutoff value** can be chosen to **maximize** the classification accuracy.

Logistic model – Variable Requirements

- Logistic regression analysis requires that the **dependent variable** be categorical.

- Logistic regression analysis requires that the independent variables be numerical or categorical.

- If an independent variable is categorical, we need to **dummy code** the variable.

- Logistic regression **does not make any assumptions** of normality, linearity, and homogeneity of variance for the independent variables.

Machine Learning Classification Examples

Logistics Regression Classification Examples:

- Fraud Identification – Fraud vs Non-fraud
- Credit Card and Loans – Default vs Non-default
- Marketing – Response vs Non-response
- Sales – Buying vs Non-buying
- Gaming – Win vs Loss
- Website – Click vs No-click
- Healthcare - Cure vs Non-cure