**REGRESSION ANALYSIS**

**The most common regression algorithms are**

1. Simple linear regression
2. Multiple linear regression
3. Polynomial regression
4. Multivariate adaptive regression splines
5. Logistic regression
6. Maximum likelihood estimation (least squares)

**Simple Linear Regression**

* Simple linear regression is the simplest regression model which involves only one predictor(X).
* This model assumes ` between the dependent

variable and the predictor variable as shown in Figure below.



Fig. Simple Linear Regression

***Example of simple regression***

A college professor believes that if the grade for internal examination is high in a class, the grade for external examination will also be high. A random sample of 15 students in that class was selected, and the data is given below:

To build a simple linear regression model for a given problem, it is time to summarize the algorithm.



***OLS algorithm***

Step 1: Calculate the mean of *X* and *Y*

Step 2: Calculate the errors of *X(X-X\_mean)* and *Y(Y-Y\_mean)*

Step 3: Get the product *(X-X\_mean) \* (Y-Y\_mean)*

Step 4: Get the summation of the products ∑*(X-X\_mean) \* (Y-Y\_mean)*

Step 5: Square the difference of *X (X-X\_mean)2*

Step 6: Get the sum of the squared difference ∑*(X-X\_mean)2*

Step 7: Divide output of step 4 by output of step 6 to calculate ‘*b*’

Step 8: Calculate ‘*a*’ using the value of ‘*b*’



***8.3.1.3 Error in simple regression***  above slope–intercept format in algorithms. X and Y values are provided to the machine, and it identifies the values of a (intercept) and b (slope) by relating the values of X and Y.

* However, identifying the exact match of values for a and b is

not always possible. There will be some error value (ɛ)

associated with it. This error is called marginal or residual

error.

*Y* = (*a* + *bX*) + *ε*

**Residual** is the distance between the predicted point (on the

regression line) and the actual point as depicted in Figure below



**Multiple Linear Regression**

In a multiple regression model, two or more independent

variables, i.e. predictors are involved in the model.

The simple linear regression model and the multiple regression model assume that the dependent variable is continuous.

The following expression describes the equation involving

the relationship with two predictor variables, namely *X1* and

*X2* .

*Ŷ* = *a* + *b1 X1* + *b2 X2*

The model describes a plane in the three-dimensional space

of *Ŷ*, *X* , and *X* . Parameter ‘*a*’ is the intercept of this plane.

Parameters ‘*b1* ’ and ‘*b2* ’ are referred to as **partial regression**

**coefficients**. Parameter *b1* represents the change in the mean

response corresponding to a unit change in *X1* when *X2* is held

constant. Parameter *b2* represents the change in the mean

response corresponding to a unit change in *X2* when *X1* is held

constant.

Multiple regression for estimating equation when there are

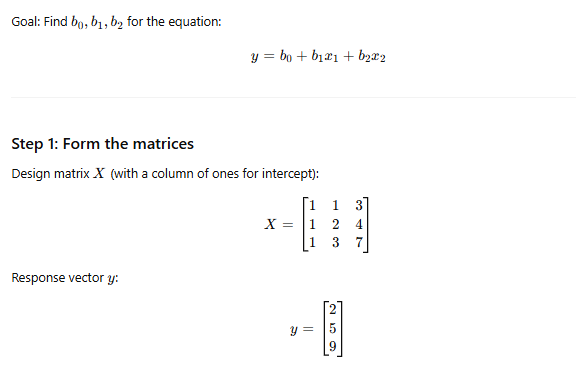
‘*n*’ predictor variables is as follows:



**Solve a Multiple Linear Regression** manually using **Matrix Algebra** , based on this small dataset.

You gave:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Size (x₁)** | **Bedrooms (x₂)** | **Price (y)** | | 1 | 3 | 2 | | 2 | 4 | 5 | | 3 | 7 | 9 | |



A math equation on a white background

AI-generated content may be incorrect.

A screenshot of a math problem

AI-generated content may be incorrect.

A screenshot of a math problem

AI-generated content may be incorrect.

A math problem with numbers and equations

AI-generated content may be incorrect.

A screenshot of a math test

AI-generated content may be incorrect.

**Polynomial Regression Model**

**Polynomial regression model** is the extension of the simple

linear model by adding extra predictors obtained by raising

(squaring) each of the original predictors to a power. For

example, if there are three variables, *X*, *X2* , and *X3* are used as

predictors. This approach provides a simple way to yield a

non-linear fit to data.

*f*(*x*) = *c0* + *c1* .*X1* + *c2X 2*+ *c 3*.*X3*

In the above equation, *c0*, *c1* , *c2* , and *c* are the coefficients.

A screenshot of a computer

AI-generated content may be incorrect.

**Typical applications of regression can be seen in**

* Demand forecasting in retails
* Sales prediction for managers
* Price prediction in real estate
* Weather forecast
* Skill demand forecast in job market

**Logistic Regression**

* Logistic regression is both classification and regression

technique depending on the scenario used. Logistic regression

(logit regression) is a type of regression analysis used for

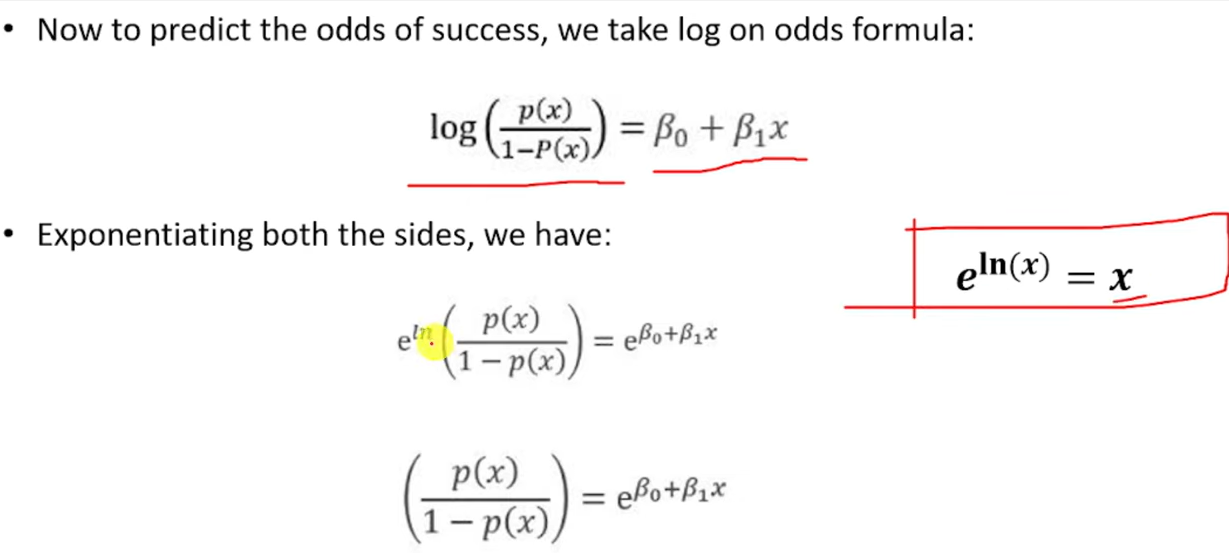
predicting the outcome of a categorical dependent variable

similar to OLS regression.

* In logistic regression, dependent variable (*Y*) is binary (0,1) and independent variables (*X*) are continuous in nature.
* The probabilities describing the possible outcomes (probability that *Y* = 1) of a single trial are modelled as a logistic function of the predictor variables.
* The goal of logistic regression is to predict the likelihood that *Y* is equal to 1 (probability that *Y* = 1 rather than 0) given certain values of *X*.
* The logistic formulae are stated in terms of the probability that *Y* = 1, which is referred to as *P*. The probability that *Y* is 0 is 1

–A math equation with numbers and symbols

AI-generated content may be incorrect.



**A graph of a function

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Assumptions in logistic regression**

The following assumptions must hold when building a logistic

regression model:

* There exists a linear relationship between logit function and independent variables
* The dependent variable *Y* must be categorical (1/0) and take binary value,e.g. if pass then *Y* = 1; else *Y* = 0
* The data meets the ‘iid’ criterion, i.e. the error terms, ε, are independent from one another and identically distributed
* The error term follows a binomial distribution [*n*, *p*]

*n* = # of records in the data, *p* = probability of success (pass, responder)

Reference:<https://medium.com/analytics-vidhya/a-comprehensive-guide-to-logistic-regression-e0cf04fe738c>