Data Mining

Mohammed Almulla

College of I.T

Regression and Correlation methods

Regression Models

Regression analysis is a form of predictive modelling technique or supervised learning which investigates the relationship between a dependent (target) and independent variable(s) (predictor). It used for prediction or estimation, time series modelling, and finding the causal effect relationship or causation between the variables (any change in the value of one variable will cause a change in the value of another variable). Beside regression models, we have other probabilistic models. One of them is correlation models, and correlation means a statistical technique which tells us how strongly the pair of variables are linearly related and change together. Back to regression model, it works in steps as follows:

1. Hypothesize Deterministic Component that Estimate Unknown Parameters.
2. Specify Probability Distribution of Random Error Term by Estimate Standard Deviation of Error.
3. Evaluate the fitted Model.
4. Use Model for Prediction and Estimation.

There are two types of regression models: simple that is when we have 1 independent variable, and multiple when we have 2 or more independent variables. Each one of them divides into linear and non-linear. This teleology is based on the number of explanatory variables & nature of relationship between X & Y. Figure 65 show types of regression models.

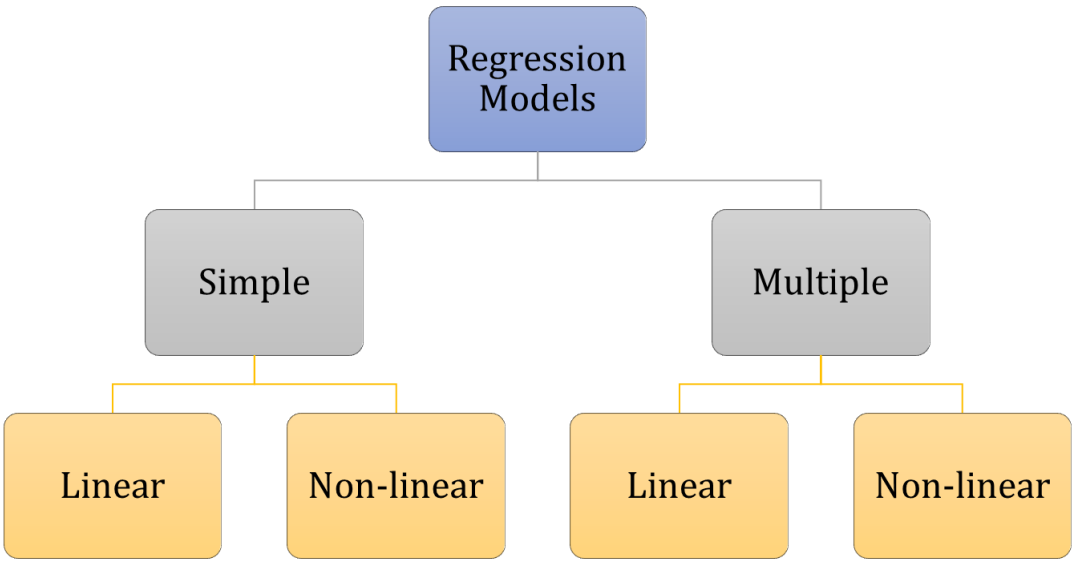


Figure 65: Types of regression models

Linear Regression Model

Linear regression follows the linear mathematical model for determining the value of one dependent variable from value of one given independent variable.

$$y = mx + b$$

Where y is the dependent variable, m is slope, x is the independent variable, and c is the intercept for a given line. The same definition we will implement it in regression form.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where Y_i is the dependent variable, β_0 is population Y-intercept, β_1 is population slope, X_i is the independent variable, and ε_i is random error or the failure of data to lie on the straight line and represents the difference between the true and observed or predicted realization of y.

That equation is use in population regression models, but when we have a sample from that population, we will use the following equation.

$$Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\varepsilon_i}$$

Figure 66 shows population and sample regression model.

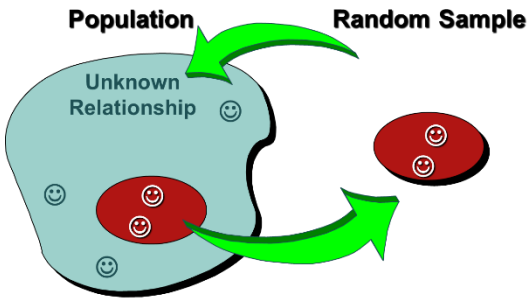


Figure 66: Population and sample regression model

Least Squares Method

There are many methods for estimating the parameters β_0 , and β_1. Among them is least squares method. We will scatter plot that is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It suggests how will model will fit. Figure 67 shows a scatter plot.

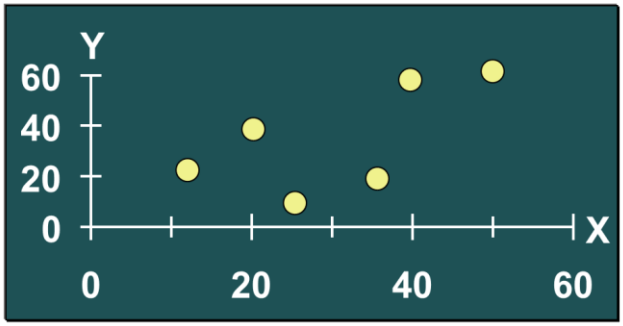


Figure 67: Scatter Plot

We can draw many lines between points in the scatter plot, but how we will know that a specific line is the best fit line. And here the job of least squares come in. It will calculate the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. 'Best Fit' Means Difference Between Actual Y Values and Predicted Y Values Are a Minimum. But Positive Differences Off-Set Negative. And least square minimizes the Sum of the Squared Differences (errors) (SSE).

$$SSE = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \widehat{\varepsilon_i}^2$$

Figure 68 shows least squares method.

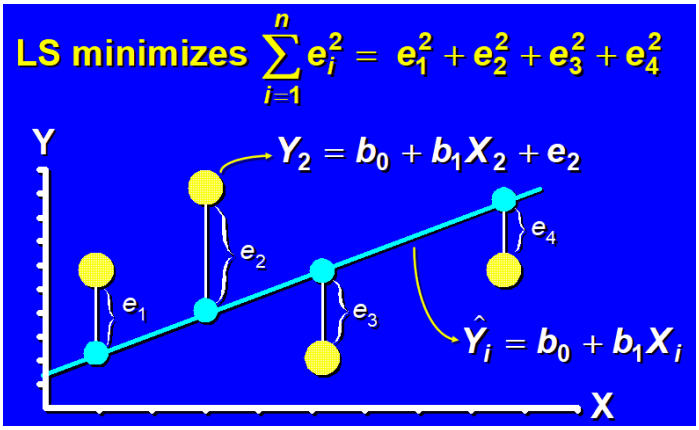


Figure 68: least squares method

Some of coefficient equations as follows:

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample Slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{[\sum x_i y_i] - n\bar{x}\bar{y}}{\sum x_i^2 - (n)(\bar{x})^2}$$

- Sample Y-intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Covariance
  - For population

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

  - For samples

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

  - For two jointly distributed real-valued random variables X and Y

$$cov(x, y) = E[(x - E[x])(y - E[y])] = E[xy] - E[x]E[y]$$

Where E is the expected value or the mean value.

The interpretation of coefficients we have first the slope $\hat{\beta}_1$ that estimated Y changes by $\hat{\beta}_1$ for each 1 unit increase in X. E.g., if $\hat{\beta}_1 = 2$, then Y Is Expected to Increase by 2 for Each 1 Unit Increase in X. The second is Y-intercept $\hat{\beta}_0$ that is average value of Y When X = 0. E.g., if $\hat{\beta}_0 = 4$, then Average Y Is Expected to Be 4 When X Is 0. Moreover, to calculate the regression parameters it will take time so use the

computation table in figure 69 so you can easily calculate the linear regression parameters and get the final equation.

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

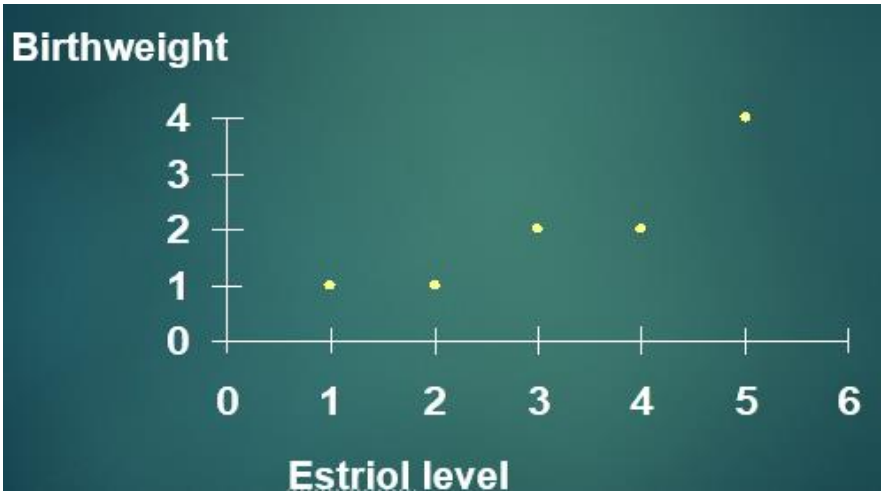
Figure 69: Computation Table

Let us take an example on how to use the laws.

Example: Obstetrics What is the relationship between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	Birthweight (g/1000)
1	1
2	1
3	2
4	2
5	4

If we draw a scatterplot based on the above data, we will get as in the figure below.



Then, we draw the computation table for the data above.

	X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1	1
2	1	4	1	2	
3	2	9	4	6	
4	2	16	4	8	
5	4	25	16	20	
Sum	15	10	55	26	37

The mean values will be as follows.

$$\overline{x} = {}^{15}/_{5} = 3$$

$$\overline{y} = {}^{10}/_{5} = 2$$

Now, let us use the laws above to calculate the parameters.

$$\widehat{\beta_1} = \frac{(\sum_{i=1}^{n} X_iY_i) - n\overline{XY}}{(\sum_{i=1}^{n} X_i^2) - n\overline{X}^2} = \frac{37 - 5 \times 3 \times 2}{55 - 5 \times 3^2} = 0.7$$

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X} = 2 - 0.7 \times 3 = -0.1$$

Then, the linear regression equation will be:

$$Y_i = \widehat{\beta_0} + \widehat{\beta_1}x_i = -0.1 + 0.7x_i$$

Finally, if we want to understand this example based on the slope $\widehat{\beta_1}$ the birthweight Y is expected to increase by 0.7 units for each 1 unit increase in Estriol X. Also, to understand the example based on the intercept $\widehat{\beta_0}$ the average birthweight Y is (-0.1) units when Estriol level X is 0. But for the intercept is difficult to explain because the birthweight should always be positive.

Model Evaluation

To quantify the quality of the regression model we use Residual Standard Error (RSE).

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where RSS is the residual sum of squares. Also, the smaller RSE the better.

Also, we can use R-squared R^2 is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.

$$R^2 = \frac{TSS - RSS}{TSS}$$

Where TSS is total sum of squares.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

It ranges in [0,1] and the higher R^2 the better.

Example: go back to our example that if we get the prediction values $\hat{y}_i$ then calculate RSE, and R-squared. Note that $\hat{y}_i$ we get it from the equation we found by calculating the parameters.

$$\hat{y}_i = -0.1 + 0.7x_i$$

X_i	Y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	1	0.6	0.16
2	1	1.3	0.09
3	2	2	0
4	2	2.7	0.49
5	4	3.4	0.36
Sum			1.1

$$RSE = \sqrt{\frac{1}{5-2} \times 1.1} = 0.605$$

To calculate R^2.

Where $RSS = 1.1$, as we found it to find RSE.

And $\bar{y} = 2$

X_i	Y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y}_i)^2$
1	1	0.6	0.16	1
2	1	1.3	0.09	1
3	2	2	0	0
4	2	2.7	0.49	0
5	4	3.4	0.36	4
Sum			1.1	6

We get $TSS = 6$

$$R^2 = \frac{6 - 1.1}{6} = 0.81$$

Data Preprocessing: Outliers

Rarely in our data set we have extreme values may distort the outcome that could be an error or a very important observation, it is called outliers. The outlier is more than 3 standard deviations from the mean. If you see one, just check if it is a mistake.

Multivariate Regression

If we have a set of independent variables x and a dependent variable, y then we need to find the relationship between them same as we did for the binary regression. But now we have multiple independent variables or more than 2.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

For $i = 1, \dots, n$.

We can express it in matrix form as shown in figure 70.

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}$$

n x 1 n x (k+1) (k+1) x 1 n x 1

Figure 70: Matrix form of multivariate regression

So, our equation will be as

$$Y = X\beta + u$$

Where u is the error. Hence the error will be as follows.

$$u = Y - X\beta$$

The squared error will be:

$$u^T u = (Y - X\beta)^T (Y - X\beta)$$

Example: If you have the data set as below, how the matrix form will be for the multivariate regression.

Height X_1	Weight X_2	Sugar Y_i
111	33	2
123	43	3
144	44	4
167	66	2.5

The number of instances $n = 4$. And $K = 2$. So, the matrix form will be as follows.

Sugar Y_i		Coefficient of β_0	Height X_1	Weight X_2		Sugar Y_i		Error
2	=	1	111	33	X	β_0	+	u_1
3		1	123	43		β_1		u_2
4		1	144	44		β_2		u_3
2.5		1	167	66				u_4
nx1 4x1		nx(k+1) 4x3				(k+1)x1 3x1		nx1 4x1

Least Squared Residual Approach

The least squared estimator for the multivariate regression linear model in matrix form. We call it as the Ordinary Least Squared (OLS) estimator. It minimizes the sum of squared error.

$$\min u^T u = (Y - X\beta)^T (Y - X\beta)$$

Where $u^T u$ is scalar value, (1xn) x (nx1).

$$\beta = (X^T X)^{-1} X^T Y$$

Also, we can use R^2 and RSE to evaluate the model.

The model interpretation for the multivariate regression is that we interpret β_i as the average effect on Y of a one unit increase in X_i holding all other predictors fixed.

Logistic Regression

It takes a linear regression values and put it in a heuristic function to get result in binary form either 0 or 1. Also known as binary logistic because the dependent variable is nominal.

$$Y = \beta_0 + \beta_1 X + e$$

Where $Y = [0, 1]$. The predicted probabilities can be greater than 1 or less than 0. Also, heteroskedastic refers to a condition in which the variance of the residual term, or error term, in a regression model varies widely. Figure 71 shows heteroskedastic condition.

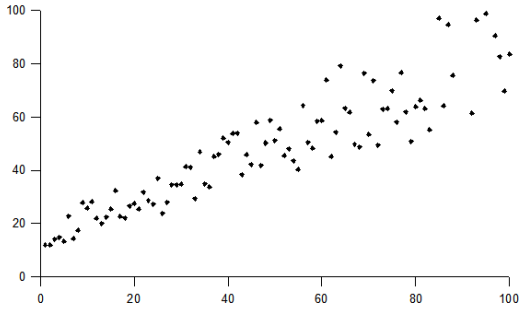


Figure 71: Heteroskedastic condition

The problem of linear regression is that predicted values are outside the 0,1 range. So, logistic model solves this problem by applying logarithmic function as follows.

$$\ln \frac{p}{1-p} = \alpha + \beta X + e$$

Where p is the probability that the event Y occurs, $\frac{p}{1-p}$ is the odds ratio, and $\ln \frac{p}{1-p}$ is the log odds ratio or also known as logit.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

The logistic distribution constrains the estimated probabilities to lie between 0 and 1. The estimated probability is based on the linear model output.

$$p_i = \frac{1}{1 + e^{-y_i}}$$

Where $y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$. If you let $\beta_0 + \beta_1 X_1 = 0$, then $p = 0.50$. Whereas $\beta_0 + \beta_1 X_1$ gets really big, p approaches 1. And as $\beta_0 + \beta_1 X_1$ gets really small, p approaches 0. The prediction will be as if $p > 0.5$ then prediction is 1, else prediction is 0. Most ML libraries allow having the predicted value to be nominal, (e.g., yes, no). Figure 72 shows comparing linear probability model and logit model.

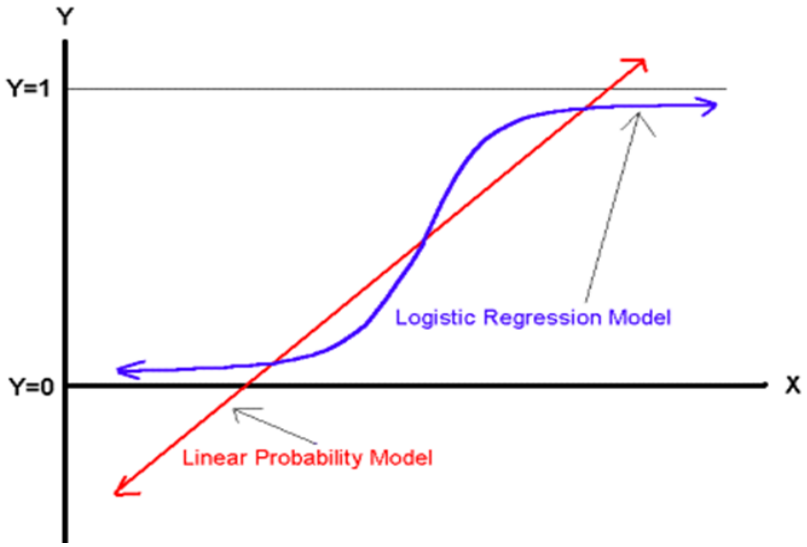


Figure 72: Comparing linear probability model and logit model

As we say before that there is binary or binomial logistic regression that predicts a binomial probability for each input example. There is multinomial logistic regression that predicts a multinomial probability or more than 2 classes for each input example. The multinomial logistic has two types as follows.

- One-vs-One scheme that generate binary logistic models for each pair of classes. The prediction will be majority vote. For N classes, we generate $\frac{N(N-1)}{2}$ binary models.
- One-vs-Rest scheme that generate binary logistic models for each class and the rest of classes. For N classes, we generate N binary models.

Figure 73 shows an example on multinomial logistic regression.

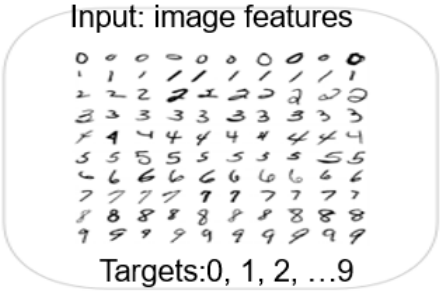


Figure 73: Example on multinomial logistic regression

Figure 74 shows One-vs-One and One-vs-Rest schemes.

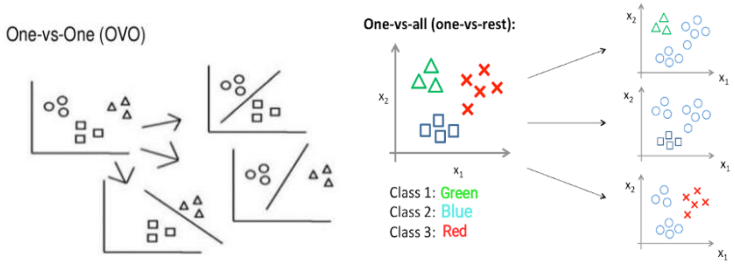


Figure 74: One-vs-One and One-vs-Rest schemes

References & Useful Resources:

- CSBP 320: Data Mining, UAE University
- Textbook: Hongbo Du.*Data Mining Techniques and Applications: an introduction*, 1st Edition
- https://en.wikipedia.org/wiki/Covariance
- https://slideplayer.com/slide/2770955/
- https://medium.com/edureka/least-square-regression-40b59cca8ea7
- http://home.iitk.ac.in/~shalab/regression/Chapter2-Regression-SimpleLinearRegressionAnalysis.pdf
- https://towardsdatascience.com/linear-regression-model-899558ba0fc4
- https://dev.to/apoorvtyagi/understanding-linear-regression-42o5
- http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm