ANALYTIXLABS

Linear Regression

Some of the Key analytics techniques & business problems

	Business Examples	Analytics Techniques
Regression Problems Predicting a value	• Predict CTR for each ad to determine placement, based on Historical CTR, Keyword match etc • Which of the retail image levers drives footfalls or conversions? • What drives satisfaction among branch users? • What causes high performance of bank branch on the basis of financial parameters?	• Linear Regression / GLM • Decision Trees (CART) • Support vector regression • Ensembles (RF, GBM, BAGGING) • ANN
Classification Problems Predicting an event or category	• Credit approval: classify credit application as low risk, high risk, or average risk • Fraud Detection: Fraud Vs. Not Fraud • Collections: Identify cardholders that are likely to default and thus need collection effort (Payment Projection Models) • Insurance: Identify claims that are Fraud or Not Fraud • campaigns(Response/Non Response, Buying/Not Buying • Operations: Models to identify to employees who attrite(Attrition/ Retention)	• Logistic Regression /Probit Regression • Naïve Bayes/KNN • Decision Trees (CART, CHAID, C5.0) • Support Vector Machines (SVM • ANN • Ensembles (RF, GBM, BAGGING)
Segmentation Problems Classifying data into unknown number of groups	• Improve customer retention by providing products tailored for specific segments • Increase profits by leveraging disposable incomes and willingness to spend • Grow you business quicker by focusing marketing campaigns on segments with higher propensity to buy • Improve customer lifetime value by identifying purchasing patterns and targeting customers when they are in the market	• K-means/K-Medians • Hierarchical clustering • Spectral clustering • DBSCAN
Forecasting Problems Predicting Future value (depends on time)	• Call volume demand in call centers • Average handle time trends • Demand for seasonal maintenance • Event based demand for field services • Estimation of cash requirement in ATMs and Branches • Number of transactions for tellers	• Averages / Smoothening • Decomposition • ARIMA/SARIMA • ARIMAX • ARCH/GARCH/VAR

ANALYTIXLABS

Linear Regression

ANALYTIXLABS

Business Problem

I am the CEO of a hypermarket chain “Safegroceries” and I want to open new store which should give me the best sales . I am hiring “Alabs” to help me figure out a location where to open the new store

What should ALABS do ?

Additional Information about Safe groceries:

- Safegroceries has more than 5000 stores across the world
- It is upstream hypermarket store catering to high end products
- There are more than 100 locations he needs to choose from ?

ANALYTIXLABS

What could impact sales ?

- ✓ Population Density in the area
- ✓ Disposable Income
- ✓ Demographics of the region
- ✓ Parking size of the location
- ✓ No of other grocery stores in around (3km)
- ✓ Credit card usage
- ✓ Internet penetration/usage
- ✓ Average no of cars/household
- ✓ Avg family size/household
- ✓ No of working people/household
- ✓ ……………………
- ✓ …………………..

ANALYTIXLABS

Relationship between Sales and Variables

- ✓ Sales = function (X1, X2, X3, X4,X5,X6……)
- ✓ Sales = 10X1 + 20X2 +0.5X3 + 8X4 +…………
- ✓ If the function is linear we call it linear regression

This was a case of prediction . How about doing root cause analysis ?

Now CEO wants to improve the performance of the existing stores and wants to increase sales ?

Decision – Prediction vs Inference(root causal)

ANALYTIXLABS

Regression

Regression Analysis

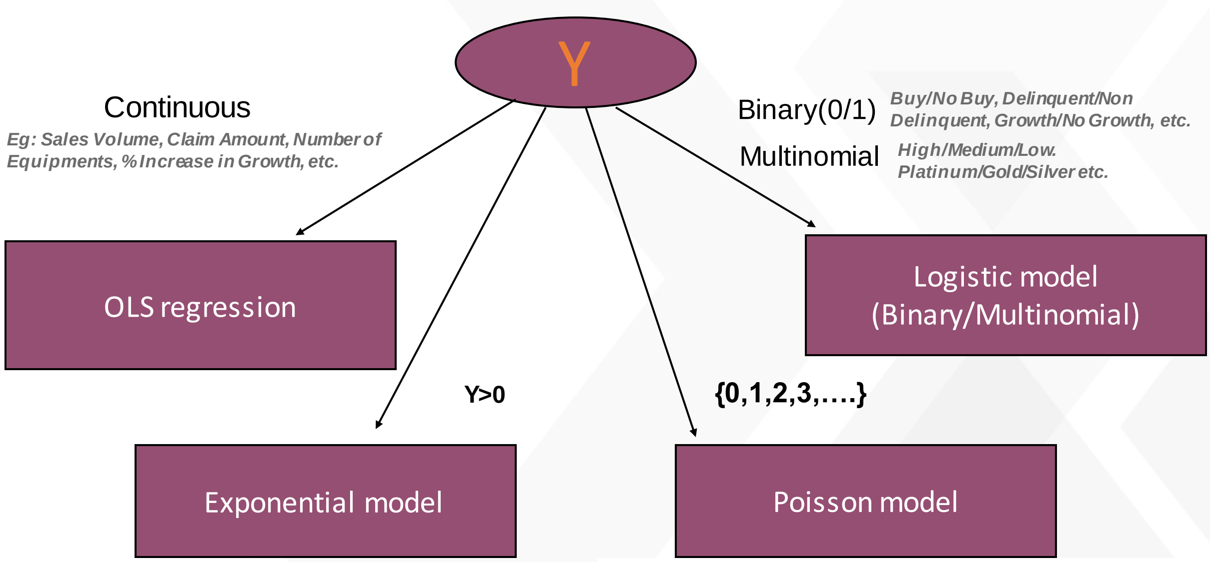
“Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another”

Regression modeling

Establishing a functional relationship between a set of Explanatory or Independent variables $X_1, X_2, \ldots, X_p$ with the Response or Dependent variable Y.

$$Y = f\,(X_1, X_2, \ldots, X_p)$$

Types of Regression Models

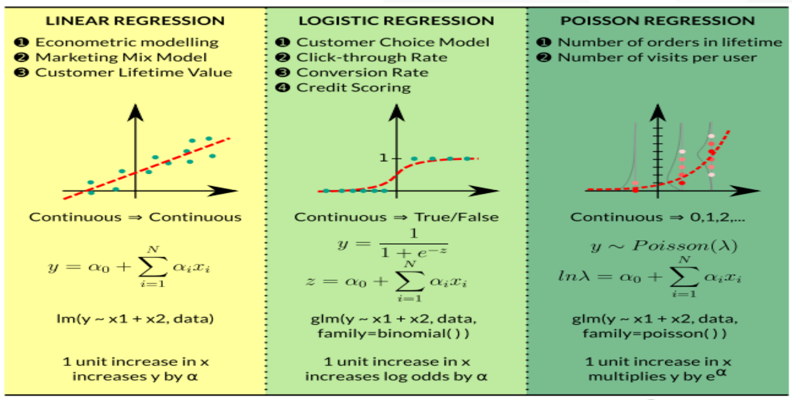


Three Regression Types (GLM)

Generalized linear models extend the ordinary linear regression and allow the response variable y to have an error distribution other than the normal distribution.

GLMs are:

- A. Easy to understand
- B. Simple to fit and interpret in any statistical package
- C. Sufficient in a lot of practical applications



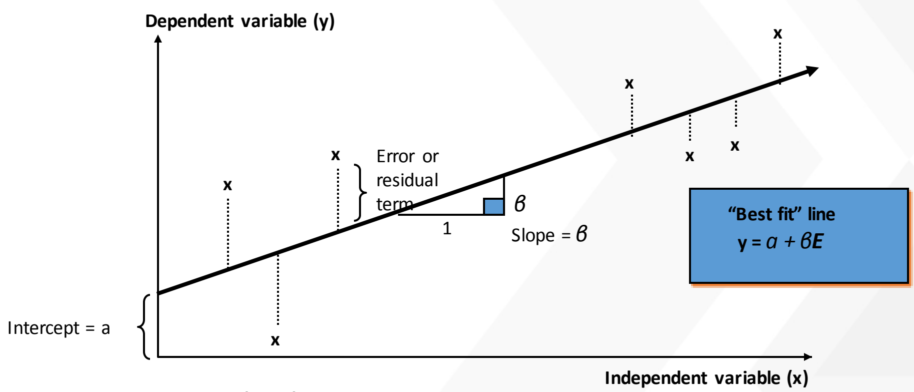
ANALYTIXLABS

Ordinary Least Square Regression(OLS)

ANALYTIXLABS

What is OLS REGRESSION ANALYSIS?

OLS Regression basically try to draw the best fit regression line - a line such that the sum of the squared deviations of the distances of all the points to the line is minimized.



Ordinary Least Squares (OLS) linear regression assumes that the underlying relationship between two variables can best be described by a line.

ANALYTIXLABS

Regression-Step-0

Step-0:

Identification of Dependent Variable
Example: Expected revenue from telecom license

Step-1:

Once we have selected the dependent variable we wish to predict, the first step before running a regression is to identify what independent variables might influence the magnitude of the dependent variable and why.

ANALYTIXLABS

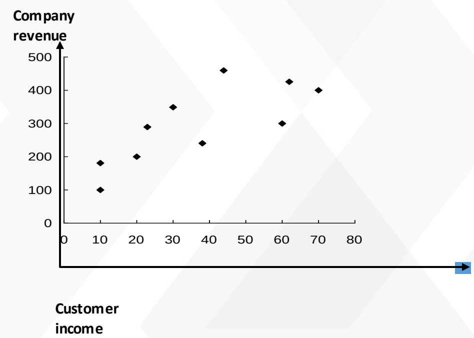
Regression-Step-1

COLLECTING AND GRAPHING THE DATA

The first step is to collect the necessary information and to enter it in a format that allows the user to graph and later "regress" the data.

(Y) Company revenue	(X) Customer income
180	10
100	10
200	20
290	23
350	30
240	38
460	44
300	60
425	62
400	70

Plotting the data allows us to get a "first look" at the strength of our relationship

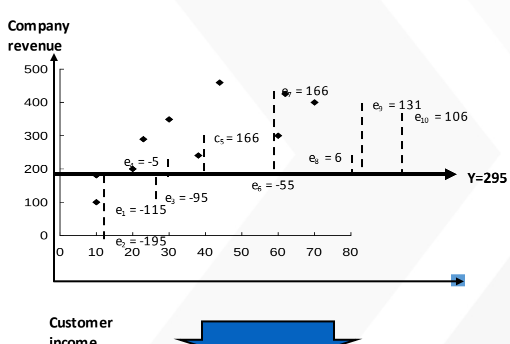


Regression-Step-2

The way linear regression "works" is to start by naively fitting a horizontal no-slope (slope = A=0) line to the data. The y-intercept B of this line is simply the arithmetic average of the collected values of the dependent variable.

(Y) Company revenue	(X) Customer income
180	10
100	10
200	20
290	23
350	30
240	38
460	44
300	60
425	62
400	70

Average Y value = 295

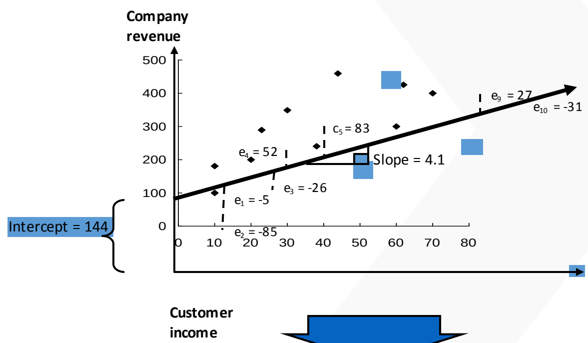


The sum of the squared residuals, $S_{no\text{-}slope}$ gives us a measure of how well the horizontal line fits the data

$$S_{no\text{-}slope} = (-115)^2 + (-195)^2 + (-95)^2 + (-5)^2 + \ldots (106)^2 = 121{,}523$$

Regression-Step-3

If we allow the line to vary in slope and intercept, we should be able to find that line which minimizes the sum of squared residuals.



"Best fit" sloped line =

Revenue = 144 + 4.1 x (income)

The new sum of squared residuals, S_{slope}, should be lower than $S_{no\text{-}slope}$, if the new line provides a better fit to the data

$$S_{slope} = (-5)^2 + (-85)^2 + (-26)^2 + \ldots + (-31)^2 = 49{,}230$$

ANALYTIXLABS

Critical Elements of linear Regression

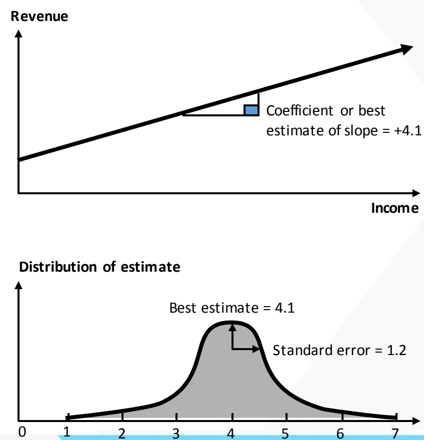
Since software packages like SAS/R will regress any stream of data regardless of its integrity, it is critical that we review the regression results first to determine if a meaningful relationship exists between the two variables before drawing any conclusions.

- **Sign and magnitude of coefficients**
- **T-statistics**
- **R2-statistics**

ANALYTIXLABS

Interpreting the coefficient – Sign test

The coefficient of the independent variable represents our best estimate for the change in the dependent variable given a one-unit change in the independent variable.



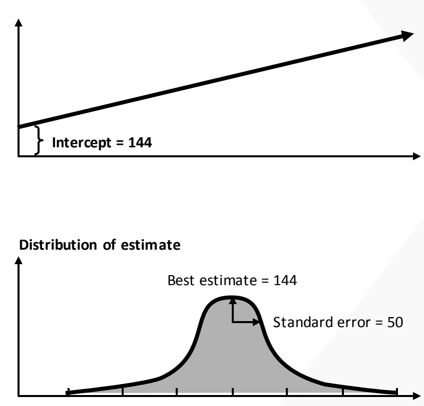
If the sign of the resulting coefficient does not match the anticipated change in the dependent variable

- Data may be corrupt (or incomplete) preventing the true relationship from appearing
- True relationship between variables may not be as strong as initially thought
- Counter-intuitive relationship might exist between variables

ANALYTIXLABS

Interpreting the coefficient

Similarly, the intercept represents our best estimate for the value of the dependent variable when the value of the independent variable is zero.



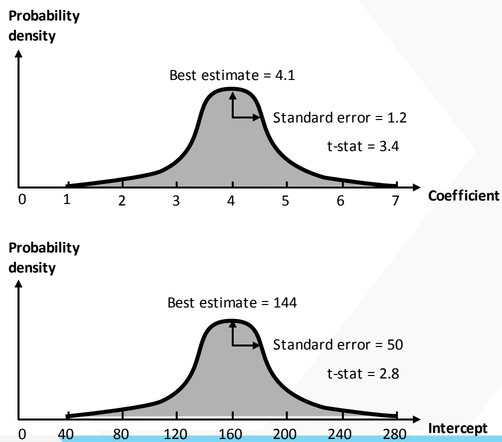
If the sign of the intercept does not match your expectation, data may be corrupt or incomplete

In some cases, it is appropriate to force the regression to have an intercept of 0, if, for instance, no meaningful value exists if the independent variable is 0

ANALYTIXLABS

T-Statistics

If the regression has passed the sign test, the single most important indicator of how strong the data supports an underlying linear relationship between the dependent and independent variables is the t-statistic.

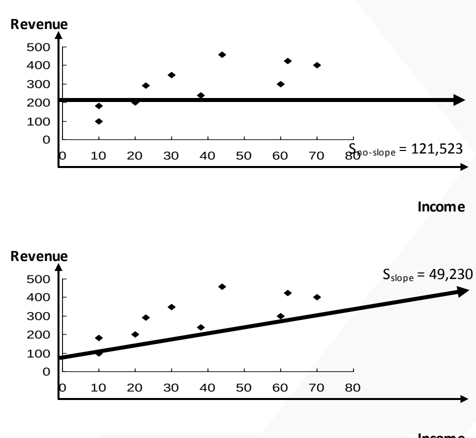


In general, a t-statistic of magnitude equal or greater than 2 suggests a statistically significant relationship between the 2 variables

ANALYTIXLABS

Interpreting R^2-Statistic

If we are comfortable with the sign and magnitude of the coefficient and intercept, and our t-statistic is sufficiently large to suggest a statistically significant relationship, then we can look at the R2-statistic.



The R^2-statistic is the percent reduction in the sum of squared residuals from using our best fit sloped line vs. a horizontal line

$$R^2 = \frac{S_{no\text{-}slope} - S_{slope}}{S_{no\text{-}slope}}$$

$$R^2 = \frac{121{,}523 - 49{,}230}{121{,}523}$$

$R^2 = 0.59$

If the independent variable does not drive (or is not correlated) with the dependent variable in any way, we would expect no consistent change in "y" with consistently changing "x." This is true when the slope is zero or $S_{slope} = S_{no\text{-}slope}$ which makes $R^2 = 0$

ANALYTIXLABS

Multiple Regression

Multiple regression allows you to determine the estimated effect of multiple independent variables on the dependent variables.

Dependent variable: Y

Independent variables:
$X_1, X_2, X_3, \ldots, X_n$

Relationship:
$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n$

Multiple regression programs will calculate the value of all the coefficients (a_0 to a_n) and give the measures of variability for each coefficient (i.e., R^2 and t-statistic)

Tests for multiple regressions

- Sign test – check signs of coefficients for hypothesized change in dependent variable
- T-statistic – check t-stat for each coefficient to establish if t>2 (for a "good fit")
- R^2, adjusted R^2
  – R^2 values increase with the number of variables; therefore check adjusted R^2 value to establish a good fit (adjusted R^2 close to 1)

Multiple Regression

If you can dream up multiple independent variables or "drivers" of a dependent variable, you may want to use multiple regression.

Independent variable	Dependent variables	Slopes	Intercept
y	x_1	a_1	b
	x_2	a_2	
	⋮	⋮	
	x_i	a_i	

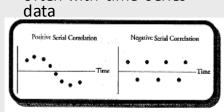
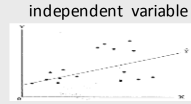
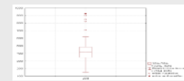
$$y = a_1x_1 + a_2x_2 \ldots + a_ix_i + b$$
$$= b + \sum_i a_1x_1$$

Multiple regression notes

- Having more independent variables always makes the fit better – **even** if it is not a statistically significant improvement. So:

1. Do the sign check for **all** slopes and the intercept
2. Check the t-stats (should be >2) for **all** slopes and the intercept
3. Use the adjusted R^2 which takes into account the false improvement due to multiple variables

Multiple regression – 4 primary issues

	Multicollinearity	Serial correlation/ Autocorrelation	Heteroscedasticity	Outlier
What is it?	• High correlation among two or more of the independent variable 	• Residual terms are correlated with one another. It occurs most often with time series data 	• Variance of the residual term increases as the value of the independent variable 	• If some values are markedly different from the majority of the values 
Effect	• Distorts the standard error of coefficient. This will lead to greater probability of incorrectly concluding that a variable is not statistically significant (Type II error)	• Coefficient standard error too large or too small leading to erroneous t-statistic	• Standard errors will be different for different sets of independent variable	• Prediction line gets pulled-up /down in presence of outlier(s) and R-sq dips
Detection	• R-square is high, F test is statistically significant but t-tests indicate that none of the individual coefficients is significantly different than zero, VIF and CI is very high	• Scatter plot of residuals or run Durbin Watson statistic	• Examine scatter plot of residuals or run Breusch-Pagan test	• Examine from scatter plot or do an univariate analysis and look at 5,10,90,95,98,99,100 percentiles to detect outlier or check from box-plot
Correction	• Run correlation matrix and drop one of the correlated variable	• Adjust coefficient standard error using Hansen method (SAS/SPSS). This will help in correct hypothesis testing of the regression coefficient	• Calculate robust standard errors (also called White-corrected standard errors) to recalculate t-statistics	• Either drop the values or cap it by the closest observation/ replace by mean

ANALYTIXLABS

Steps in Regression Model building

1. Converting business problem into statistical problem - Identifying type of problem
2. Define hypothetical relationship (Defining Y & X variables)
3. Collect the data from across sources
4. Aggregation-getting data at same level (depends on type of problem)
5. Data Audit Report - Meta data level - table level - individual variable level
6. Data preparation
   a. Exclusions - Based business rules
   b. Data type conversions
   c. Outliers
   d. Fill rate – Missing's
   e. Derived variable creation - New variable creation - Binning of variables
   f. dummy variable creation
7. Data preparation (based on technique)
   Check the Assumptions (Y- Normal, Y & X linear)
   Transformations
   Multi-collinierity
8. Split the data into training & testing data sets(70:30)
9. Build the model on training
10. Interpreting the model - by checking few set of metrics
11. Validate the model using testing data
    1. Re-run the model
    2. Scoring the model
    3. K-Fold validation(cross Validation)
12. Preparing the final reports to share the results
13. Identify the limitations of Model
14. Converting statistical solution into Business Solution – Implementation

ANALYTIXLABS

Development of the model

Identify	Explanatory and Response variables
Decide on type of model	Here the type of model is OLS
Variable Selection	Forward Backward Stepwise
Check Multicollinearity	VIF Condition index Variance proportions
Run model	OLS
Diagnostics	For OLS
Model unsatisfactory?	Try transformations Log, sqrt, Inverse, Box-Cox etc.

ANALYTIXLABS

Diagnostics for OLS Model

Is the model satisfactory?

- ✓ R2 = proportion of variation in the response variable explained by the model
  - check R2 >50%
- ✓ Plots of Standardized Residual (= (Actual – Predicted)/SD)
  - vs predicted values
  - vs X variables
  - check if there is no pattern
  - check for homoscedasticity
- ✓ Significance of parameter estimates
  - check if p-value<0.01
- ✓ Stability of parameter estimates:
  - Take a random subsample from the development sample
  - Obtain a new set of parameter estimates from the sub sample
  - Check if the parameter estimates got from development sample and the subsample differ by less than 3 standard deviations
- ✓ Rank ordering:
  - order data in descending order of predicted values
  - Break into 10 group
  - check if average of actual is in the same order as average predicted

ANALYTIXLABS

Validation

On the validation sample

- Stability of parameter estimates:
  - -Obtain a new set of parameter estimates from the validation sample
  - -check if the new parameter estimates differ from that got from development sample by less than 3 standard deviations
- Compare Predicted vs Actual values

ANALYTIXLABS

Regression-Best practices

1. Check for the collinearity (by finding correlation between all the variables and keeping only 1 of the variables which is highly correlated)
2. Transform data as applicable – e.g., income should be transformed by taking log of that
3. Do not run regression on categorical variables, recode them into dummy variables
4. Check the directionality of the variables
5. Following methods should be used under different situations

- **Enter Method :** To get the coefficient of each and every variable in the regression
- **Back ward method :** When the model is exploratory and we start with all the variables and then remove the insignificant ones
- **Forward Method:** Sequentially add variables one at a time based on the strength of their squared semi-partial correlations (or simple bivariate correlation in the case of the first variable to be entered into the equation)
- **Step wise method :** A combination of forward and backward at each step one can be entered (on basis of greatest improvement in R2 but one also may be removed if the change (reduction) in R2 is not significant (In the Bordens and Abbott text it sounds like they use this term to mean Forward regression)

ANALYTIXLABS

Contact us

Visit us on: http://www.analytixlabs.in/

For course registration, please visit: http://www.analytixlabs.co.in/course-registration/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 88021-73069

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/

ANALYTIXLABS