



ANALYTIXLABS

## Linear Regression

Disclaimer: This material is protected under copyright act AnalytixLabs ©, 2011-2018. Unauthorized use and/ or duplication of this material or any part of this material including data, in any form without explicit and written permission from AnalytixLabs is strictly prohibited. Any violation of this copyright will attract legal actions

# Linear Regression

## 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 ?

## 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
- ✓ .....
- ✓ .....

## Relationship between Sales and Variables

- ✓ Sales = function (X1, X2, X3, X4, X5, X6.....)
- ✓ Sales =  $10X1 + 20X2 + 0.5X3 + 8X4 + \dots$
- ✓ 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)**

# 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”

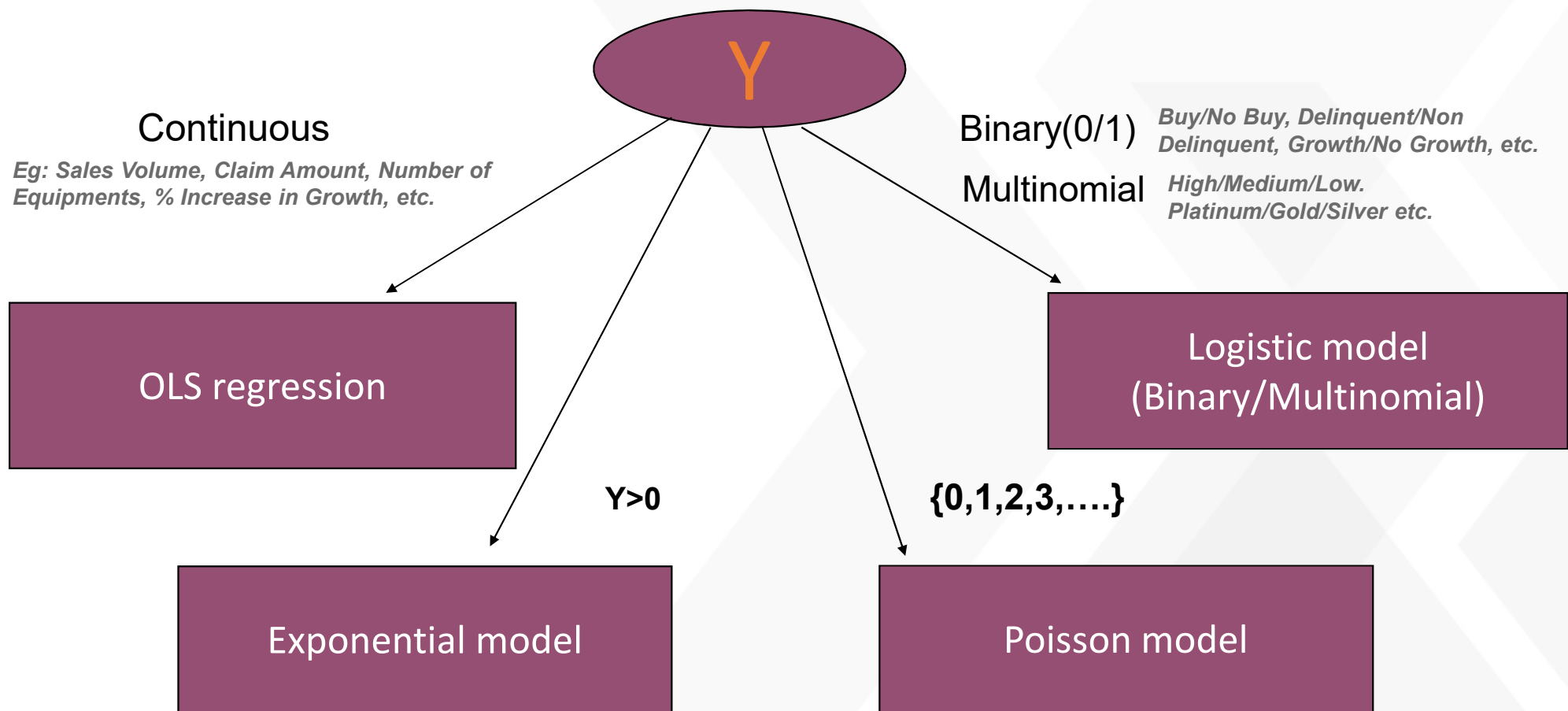
## Regression modeling

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

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



# Types of Regression Models

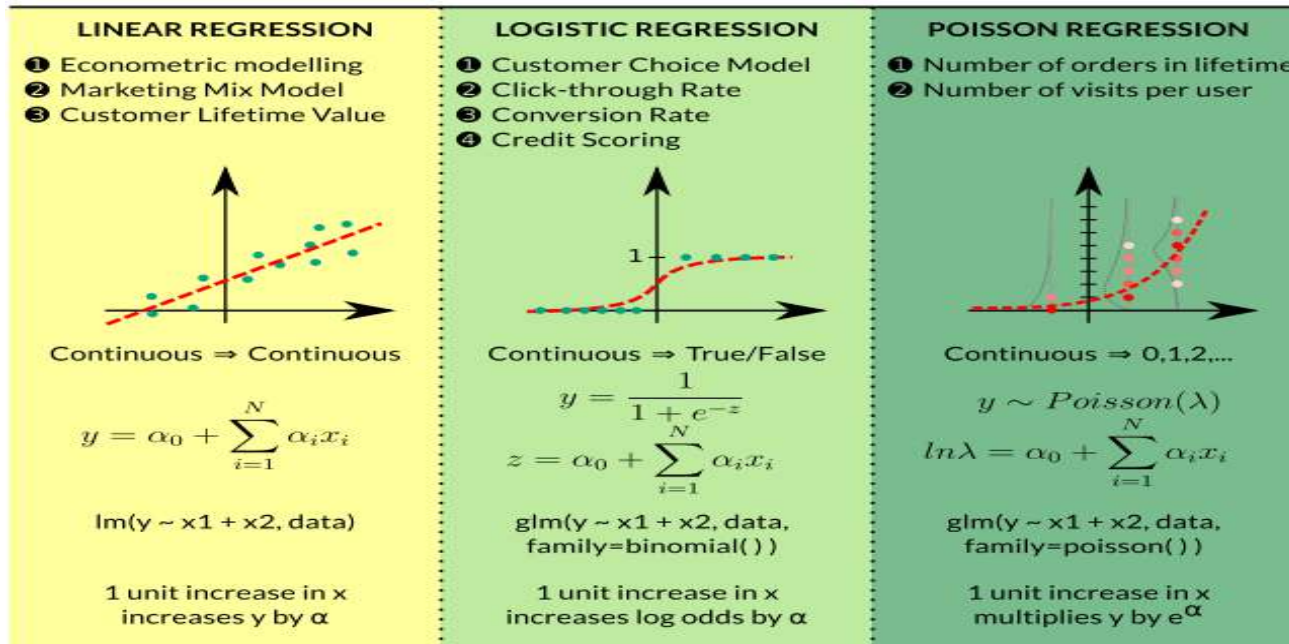


# Three Regression Types (GLM)

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

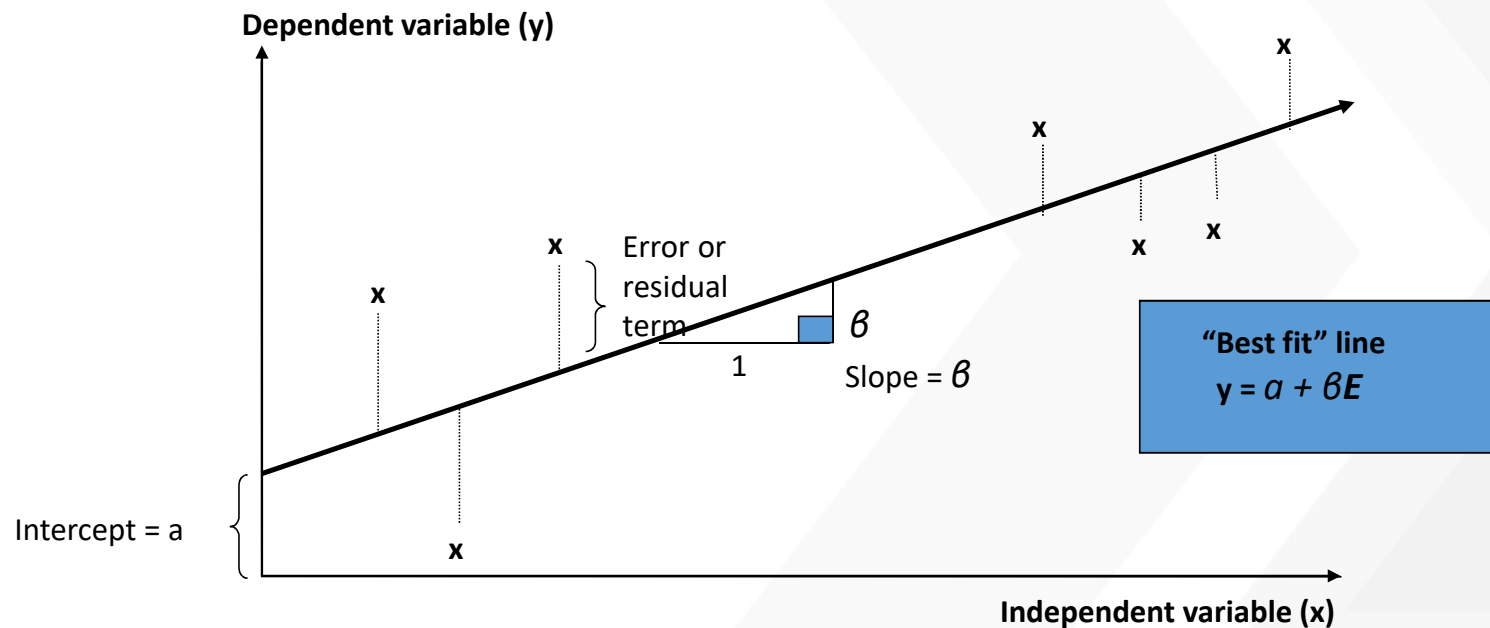




# Ordinary Least Square Regression(OLS)

# 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.

# 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.

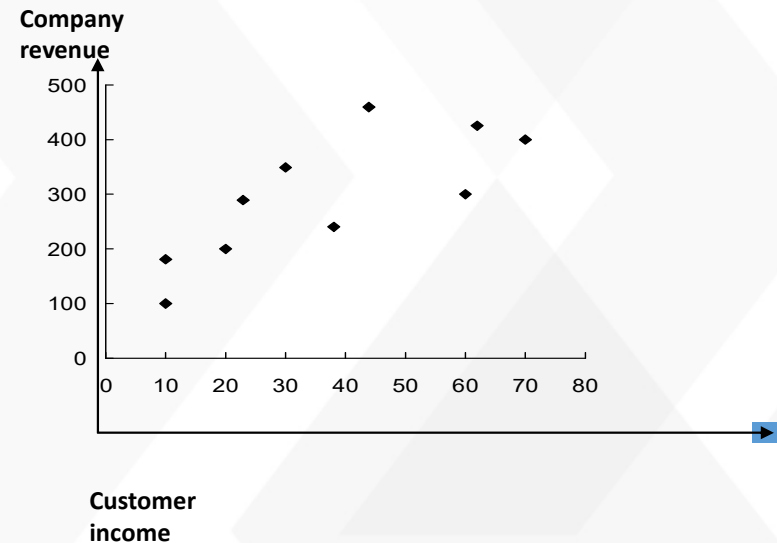
# Regression-Step-1

## 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

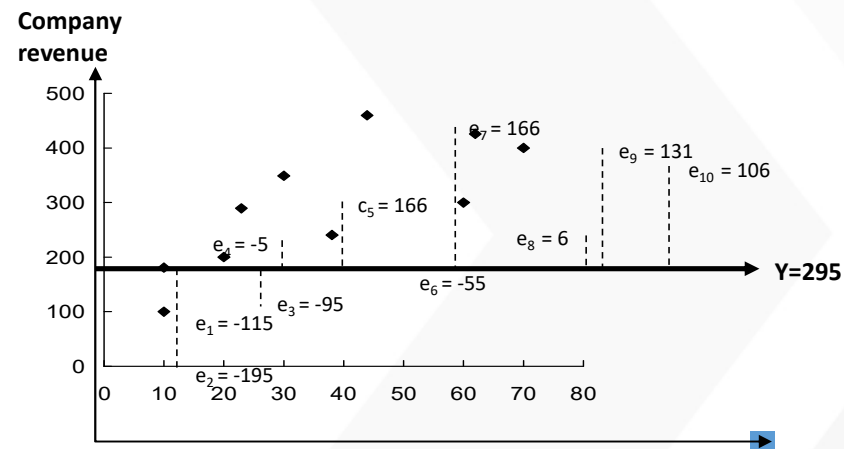


## Regression-Step-2

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

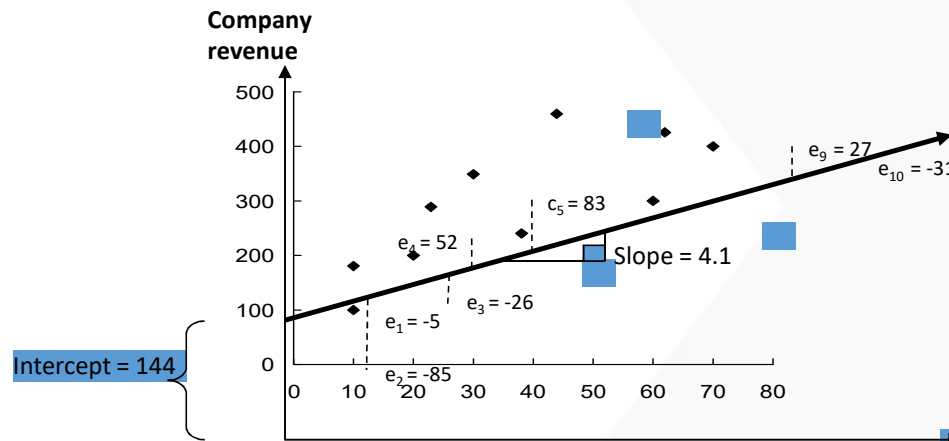


The sum of the squared residuals,  $S_{\text{no-slope}}$  gives us a measure of how well the horizontal line fits the data

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

## 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 =

$$\text{Revenue} = 144 + 4.1 \times (\text{income})$$

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

Customer  
income



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

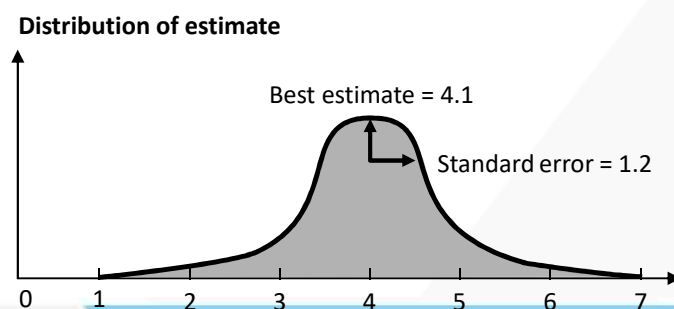
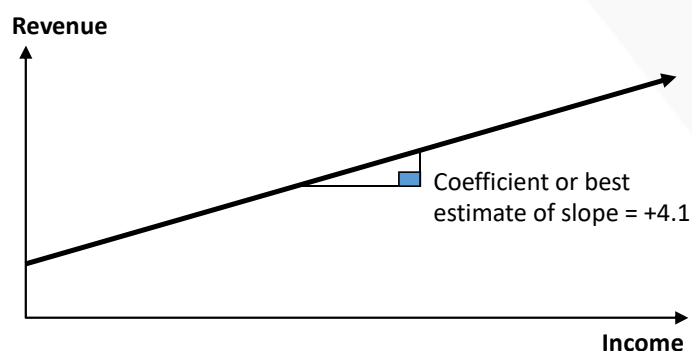
## Critical Elements of linear Regression

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
- R<sup>2</sup>-statistics

## 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.



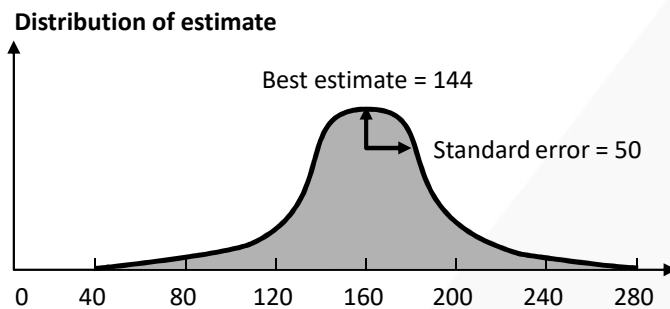
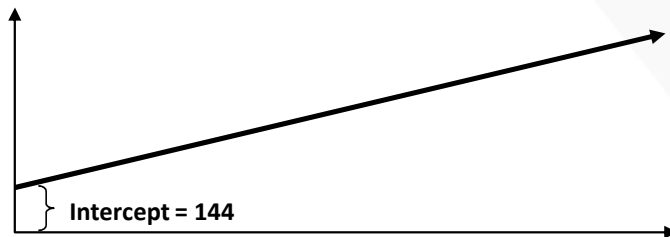
**If the sign of the resulting coefficient does not match the anticipated change in the dependent variable**

- 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



## 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.

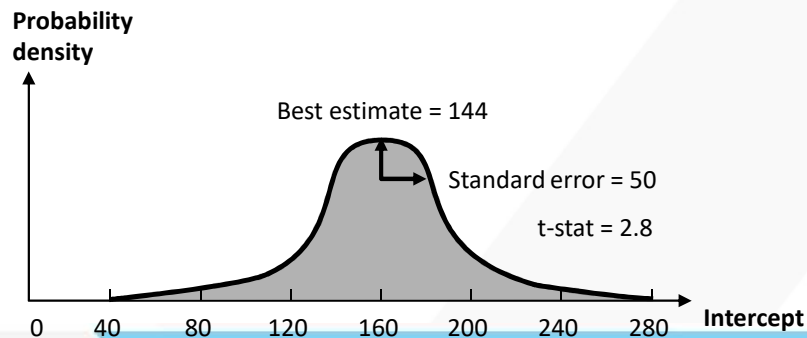
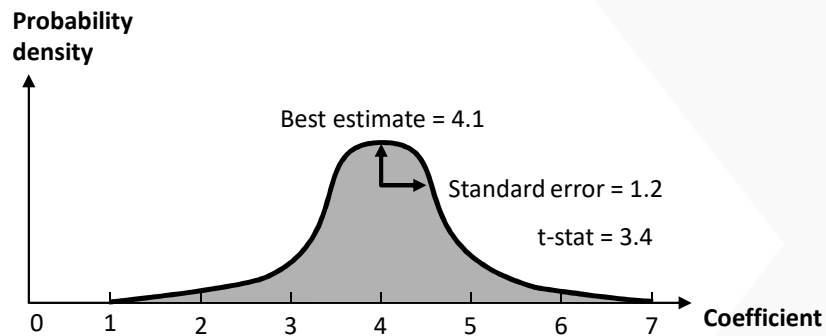


If the sign of the intercept does not match your expectation, data may be corrupt or incomplete

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

## T-Statistics

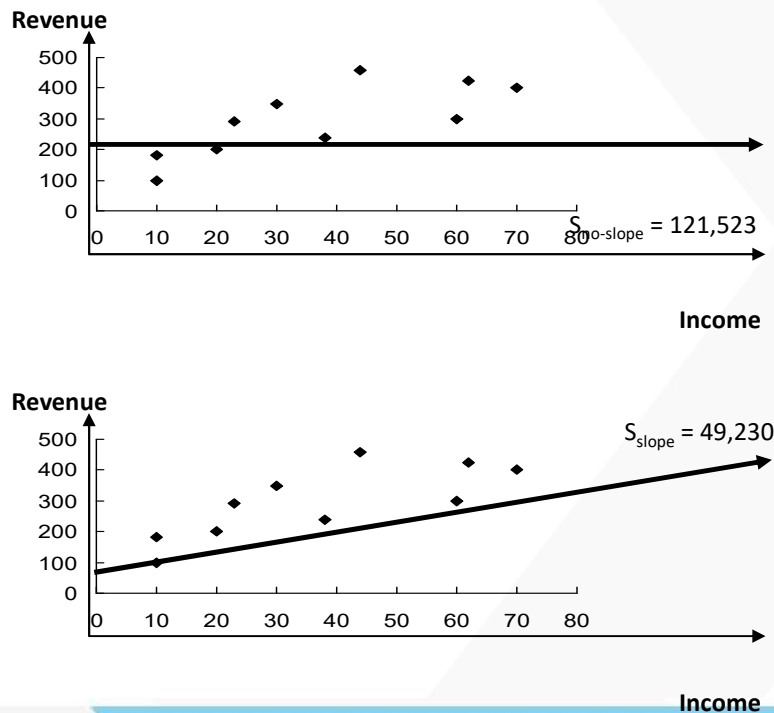
If the regression has passed the sign test, the single most important indicator of how strong the data supports an underlying linear relationship between the dependent and independent variables is the t-statistic.



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

## Interpreting R<sup>2</sup>-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 R<sup>2</sup>-statistic.



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

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

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

$$R^2 = 0.59$$

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  $S_{\text{slope}} = S_{\text{no-slope}}$  which makes  $R^2 = 0$

# Multiple Regression

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

Dependent variable:  $Y$

Independent variables:

$X_1, X_2, X_3, \dots, X_n$

Relationship:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + \dots + a_n X_n$$

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.

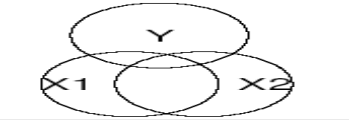
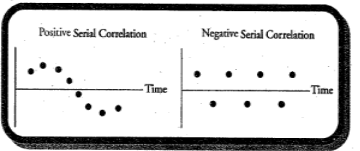
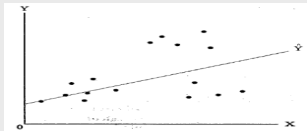
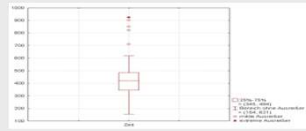
Independent variable	Dependent variables	Slopes	Intercept
y	$x_1$	$a_1$	b
	$x_2$	$a_2$	
	•	•	
	•	•	
	$x_i$	$a_i$	

$$y = a_1 x_1 + a_2 x_2 \dots + a_i x_i + b$$
$$= b + \sum_i a_i x_i$$

## 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?	<ul style="list-style-type: none"> <li>High correlation among two or more of the independent variable</li> </ul> 	<ul style="list-style-type: none"> <li>Residual terms are correlated with one another. It occurs most often with time series data</li> </ul> 	<ul style="list-style-type: none"> <li>Variance of the residual term increases as the value of the independent variable</li> </ul> 	<ul style="list-style-type: none"> <li>If some values are markedly different from the majority of the values</li> </ul> 
Effect	<ul style="list-style-type: none"> <li>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)</li> </ul>	<ul style="list-style-type: none"> <li>Coefficient standard error too large or too small leading to erroneous t-statistic</li> </ul>	<ul style="list-style-type: none"> <li>Standard errors will be different for different sets of independent variable</li> </ul>	<ul style="list-style-type: none"> <li>Prediction line gets pulled-up /down in presence of outlier(s) and R-sq dips</li> </ul>
Detection	<ul style="list-style-type: none"> <li>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</li> </ul>	<ul style="list-style-type: none"> <li>Scatter plot of residuals or run Durbin Watson statistic</li> </ul>	<ul style="list-style-type: none"> <li>Examine scatter plot of residuals or run Breusch-Pagan test</li> </ul>	<ul style="list-style-type: none"> <li>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</li> </ul>
Correction	<ul style="list-style-type: none"> <li>Run correlation matrix and drop one of the correlated variable</li> </ul>	<ul style="list-style-type: none"> <li>Adjust coefficient standard error using Hansen method (SAS/SPSS). This will help in correct hypothesis testing of the regression coefficient</li> </ul>	<ul style="list-style-type: none"> <li>Calculate robust standard errors (also called White-corrected standard errors) to recalculate t-statistics</li> </ul>	<ul style="list-style-type: none"> <li>Either drop the values or cap it by the closest observation/ replace by mean</li> </ul>

# 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

## Development of the model

Identify

Decide on type of model

Variable Selection

Check Multicollinearity

Run model

Diagnostics

Model unsatisfactory ?

Explanatory and Response variables

Here the type of model is OLS

Forward  
Backward  
Stepwise

VIF  
Condition index  
Variance proportions

OLS

For OLS

Try transformations Log, sqrt,  
Inverse, Box-Cox etc.



# Diagnostics for OLS Model

## Is the model satisfactory ?

- ✓  $R^2$  = proportion of variation in the response variable explained by the model
  - check  $R^2 > 50\%$
- ✓ Plots of Standardized Residual ( $= (\text{Actual} - \text{Predicted})/\text{SD}$ )
  - vs predicted values
  - vs X variables
  - check if there is no pattern
  - check for homoscedasticity
- ✓ Significance of parameter estimates
  - check if  $p\text{-value} < 0.01$
- ✓ 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

# 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

## 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  $R^2$  but one also may be removed if the change (reduction) in  $R^2$  is not significant (In the Borden and Abbott text it sounds like they use this term to mean Forward regression)

# Ten assumptions of linear regression

# Assumption 1: Regression model is linear in parameters

An example of model equation that is linear in parameters

$$Y = a + (\beta_1 * X_1) + (\beta_2 * X_2^2)$$

Though, the  $X_2$  is raised to power 2, the equation is still linear in beta parameters. So the assumption is satisfied in this case.

## Assumption 2: The mean of residuals is zero

### How to check?

Check the mean of the residuals. If it zero (or very close), then this assumption is held true for that model. This is default unless you explicitly make amends, such as setting the intercept term to zero.

```
mod <- lm(dist ~ speed, data=cars)
mean(mod$residuals)
#=> 2.442491e-17
```

Since the mean of residuals is approximately zero, this assumption holds true for this model.

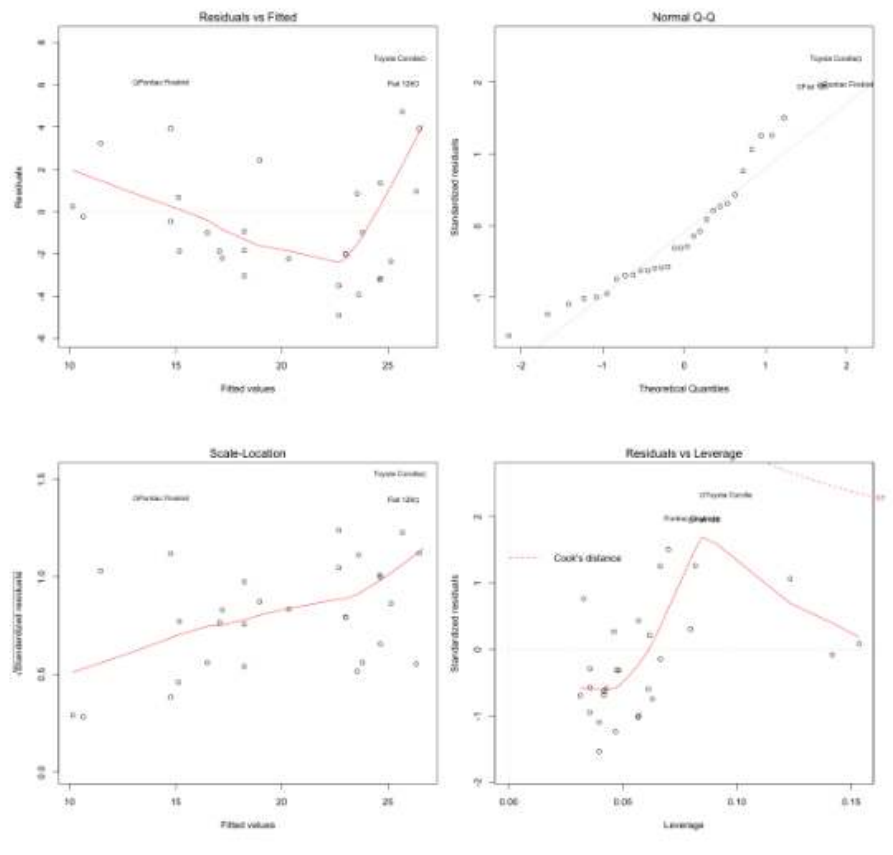
# Assumption 3: Homoscedasticity of residuals or equal variance

## How to check?

Once the regression model is built, set `par(mfrow=c(2, 2))`, then, plot the model using `plot(lm.mod)`. This produces four plots. The top-left and bottom-left plots shows how the residuals vary as the fitted values increase.

```
par(mfrow=c(2,2)) # set 2 rows and 2 column plot layout
mod_1 <- lm(mpg ~ disp, data=mtcars) # linear model
plot(mod_1)
```

## Assumption 3: Homoscedasticity of residuals or equal variance



From the first plot (top-left), as the fitted values along x increase, the residuals decrease and then increase. This pattern is indicated by the red line, which should be approximately flat if the disturbances are homoscedastic.

The plot on the bottom left also checks this, and is more convenient as the disturbance term in Y axis is standardized. In this case, there is a definite pattern noticed. So, there is heteroscedasticity.



## Assumption 4: No autocorrelation of residuals

This is applicable especially for time series data.

Autocorrelation is the correlation of a time Series with lags of itself. When the residuals are autocorrelated, it means that the current value is dependent of the previous (historic) values and that there is a definite unexplained pattern in the Y variable that shows up in the disturbances.

How to rectify?

Add lag1 of residual as an X variable to the original model. This can be conveniently done using the slide function in DataCombine package.

```
library(DataCombine)  
econ_data <- data.frame(economics, resid_mod1=lmMod$residuals)  
econ_data_1 <- slide(econ_data, Var="resid_mod1", NewVar = "lag1", slideBy = -1)  
econ_data_2 <- na.omit(econ_data_1)  
lmMod2 <- lm(pce ~ pop + lag1, data=econ_data_2)
```

# Assumption 5: The X variables and residuals are uncorrelated

## How to check?

Do a correlation test on the X variable and the residuals

```
mod.lm <- lm(dist ~ speed, data=cars)
cor.test(cars$speed, mod.lm$residuals) # do correlation test
#=> Pearson's product-moment correlation
#=>
#=> data: cars$speed and mod.lm$residuals
#=> t = -8.1225e-17, df = 48, p-value = 1
#=> alternative hypothesis: true correlation is not equal to 0
#=> 95 percent confidence interval:
#=> -0.2783477 0.2783477
#=> sample estimates:
#=>      cor
#=> -1.172376e-17
```

## Assumption 6: The number of observations must be greater than number of Xs

**This can be directly observed by looking at the data.**

## Assumption 7: The variability in X values is positive

This means the X values in a given sample must not all be the same (or even nearly the same).

How to check?

```
var(cars$speed)  
#=> [1] 27.95918
```

The variance in the X variable above is much larger than 0. So, this assumption is satisfied.

## Assumption 8: The regression model is correctly specified

This means that if the Y and X variable has an inverse relationship, the model equation should be specified appropriately.

$$Y = \beta_1 + \beta_2 * \left( \frac{1}{X} \right)$$

# Assumption 9: No perfect multicollinearity

**There is no perfect linear relationship between explanatory variables.**

## **How to check?**

Using Variance Inflation factor (VIF).

VIF is a metric computed for every X variable that goes into a linear model. If the VIF of a variable is high, it means the information in that variable is already explained by other X variables present in the given model, which means, more redundant is that variable. So, lower the VIF ( $<2$ ) the better.

## **How to rectify?**

1. Either iteratively remove the X var with the highest VIF or,
2. See correlation between all variables and keep only one of all highly correlated pairs.

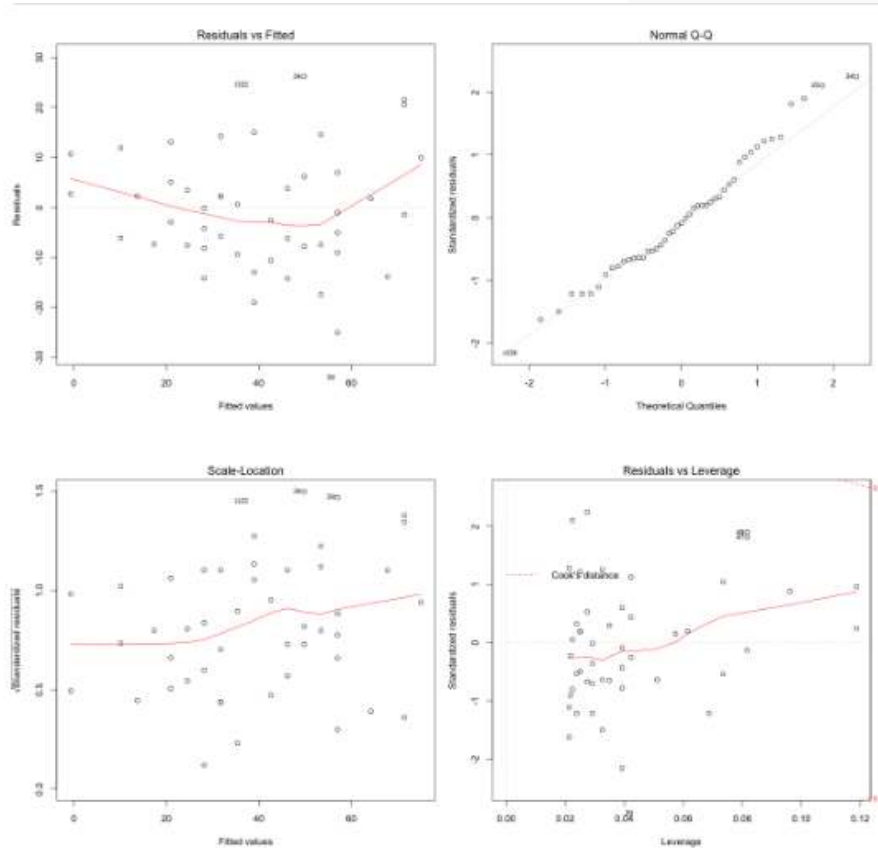
# Assumption 10: Normality of residuals

The residuals should be normally distributed. If the maximum likelihood method (not OLS) is used to compute the estimates, this also implies the Y and the Xs are also normally distributed.

This can be visually checked using the `qqnorm()` plot (top right plot).

```
par(mfrow=c(2,2))  
mod <- lm(dist ~ speed, data=cars)  
plot(mod)
```

# Assumption 10: Normality of residuals



The `qqnorm()` plot in top-right evaluates this assumption. If points lie exactly on the line, it is perfectly normal distribution. However, some deviation is to be expected, particularly near the ends (note the upper right), but the deviations should be small, even lesser that they are here.



## 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](mailto:info@analytixlabs.co.in)

Call us we would love to speak with you: +91 95-55-219007

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/>