upGrad

#LifeKoKaroLift

Bike Sharing Assignment

Course : Machine Learning

Lecture On : Bike Sharing Pre-assignment

Instructor : Manish Kumar

Today's Agenda

- Problem Statement walkthrough
- The objectives for the assignment
- Pre-Modelling Steps
- Model Building
- Model Evaluation
- Q & A

What is BoomBikes ?

A bike-sharing system in which bikes are made available for shared use to individuals on a short term basis for a price or free. It allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.

Business objective: The objective is to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Bike Sharing Assignment

upGrad

Requirements

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands.

What you need to do?

- Create a linear model that describe the effect of various features on demand.
- Th model should be interpretable so that the management can understand it.

Bike Sharing Assignment

Assignment Steps

- Identify different variable types
- Drop Unnecessary variables: ex. ‘instant’ , ‘dteday’ , ‘casual’ and ‘registered’ etc.
- Make necessary changes as per the different features

DEMAND DATA

“day.csv”

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	01-01-2018	1	0	1	0	6	0	2	14.11	18.181	80.6	10.749882	331	654	985
2	02-01-2018	1	0	1	0	0	0	2	14.9	17.687	69.6	16.652113	131	670	801
3	03-01-2018	1	0	1	0	1	1	1	8.051	9.4703	43.7	16.636703	120	1229	1349
4	04-01-2018	1	0	1	0	2	1	1	8.2	10.606	59	10.739832	108	1454	1562
5	05-01-2018	1	0	1	0	3	1	1	9.305	11.464	43.7	12.5223	82	1518	1600
6	06-01-2018	1	0	1	0	4	1	1	8.378	11.66	51.8	6.0008684	88	1518	1606

Assignment Steps

Data Visualization

- Perform EDA to understand various variables.
- Check the correlation between the variables.

Data Preparation

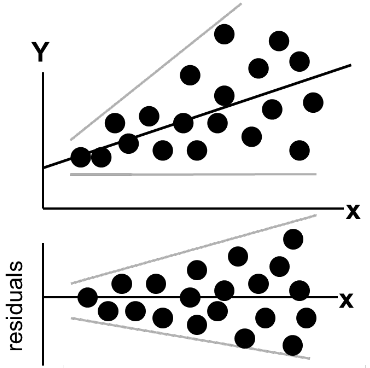
- Create dummy variables for all the categorical features.
- Divide the data to train & Test.
- Perform Scaling.
- Divide data into dependent & Independent variables.

Data Modelling & Evaluation

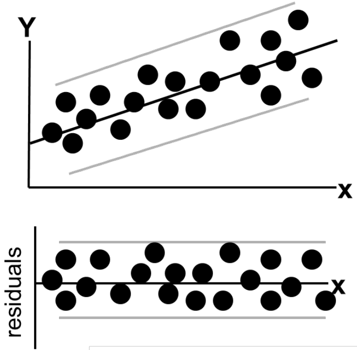
- Create Linear Regression model using mixed approach (RFE & VIF/p-value).
- Check the various assumptions.
- Check the Adjusted R-Square for both train & Test data.
- Report the final model.

Assumptions for residual analysis

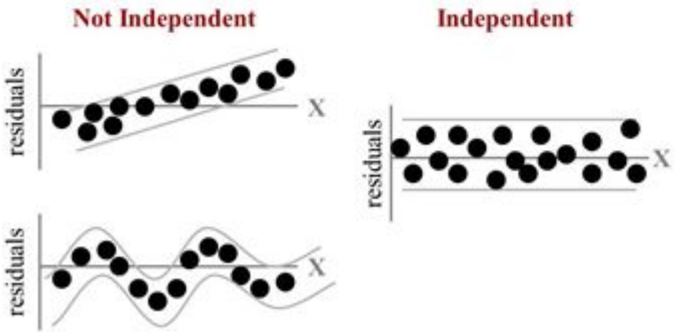
- Normality of Error
  - Error values (ε) are normally distributed for any given value of X
- Homoscedasticity
  - The probability distribution of the errors has constant variance
- Independence of Errors
  - Error values are statistically independent



Non-constant variance



Constant variance


Not Independent

Independent

Subjective Questions

Steps to complete the subjective answers

- Answer all the questions.
- For subjective answers, use DOC file. If you wish to add images you can. But convert the document to PDF format before submitting.
- You can use images and plots to support your answers.
- Make sure the question is answered in a concise and effective manner: No word limit
- Please do not copy from online available literature. You are free to refer any resources. Write the answers in your own words

End note

Tips

- Add comments/observations after every cell of code. So that we can understand your approach and method.
- Describe the results
- Create only once Jupyter Notebook.
- Submit once zip file with the code and the PDF
- Use stack Overflow
- Post on the discussion forums for resolving any doubts
- Finally, Write the code manually instead of copy-pasting from the in-content notebooks provided. It's fine to look and write

upGrad

#LifeKoKaroLift

Thank You!

Descriptive Statistics

upGrad

" A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. "

Measurement level of Data

Cardinal Numbers

A Cardinal Number says how many of something there are, such as one, two, three .. It answers the question "How Many?" , Example: here are **five** coins

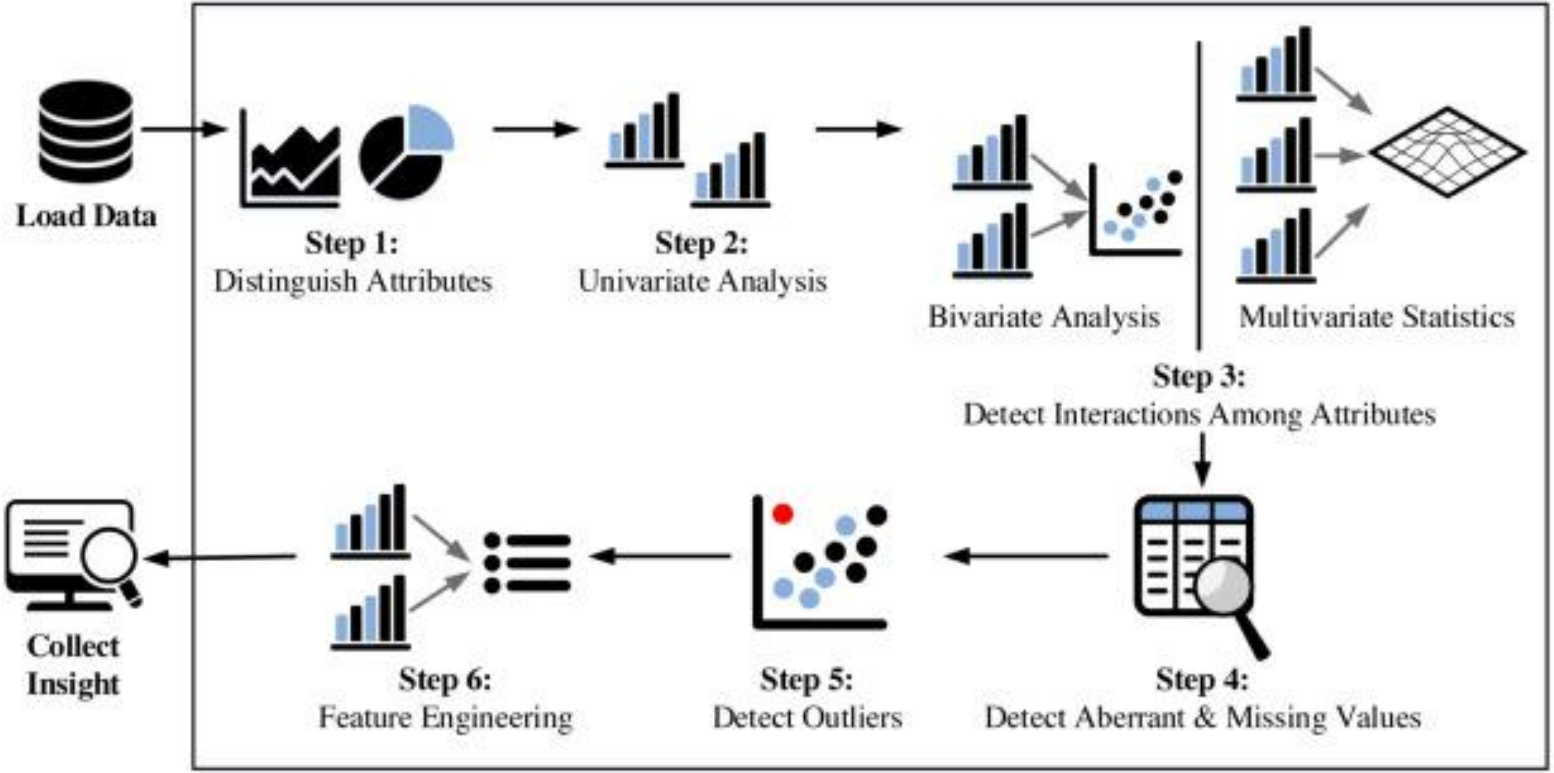
Ordinal Numbers

An Ordinal Number tells us the position of something in a list. **1st, 2nd, 3rd, 4th, 5th** and so on, eg: Grades, Star Reviews, Position in Race, Date

Nominal Numbers

A Nominal Number is a number used only as a name, or to identify something (not as an actual value or position) eg: Brand Name, ZipCode, Gender

Fundamental steps of EDA process

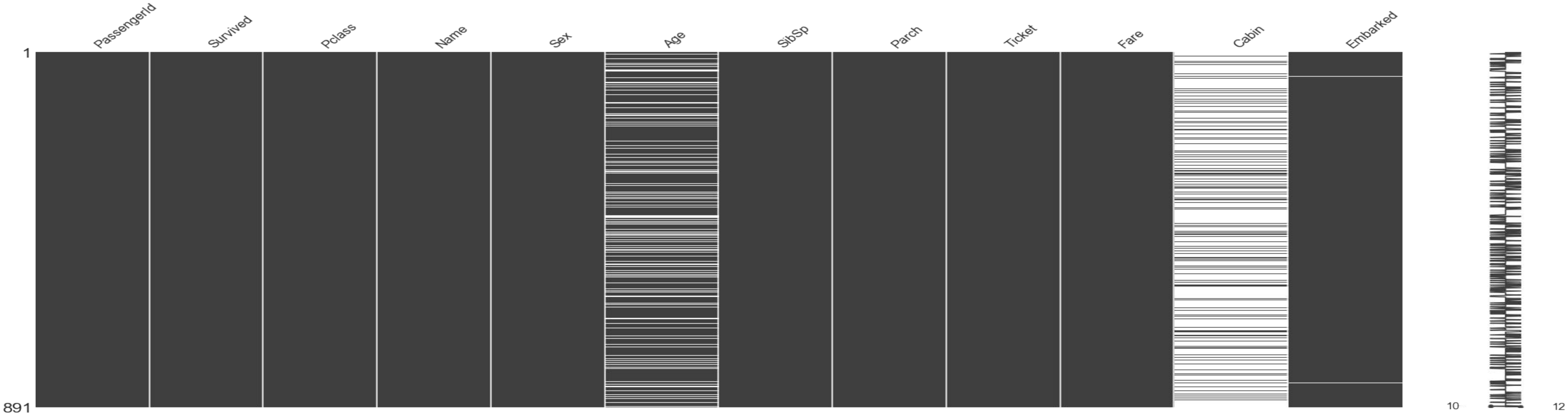


Missing Value Treatment

upGrad

1. **Data Extraction**: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.

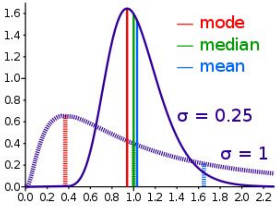
2. **Data collection**: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:



Methods to treat missing values

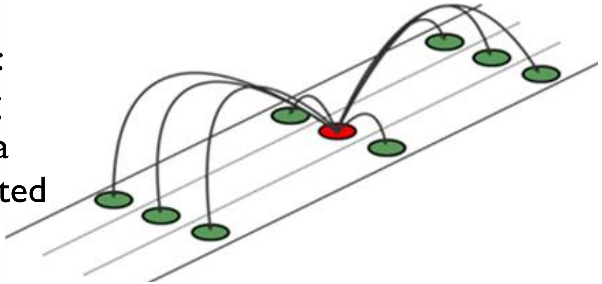
upGrad

Data – deletion Deletion methods are used when the nature of missing data is “**Missing completely at random**” or we have good amount of data and the data loss would be really low ,else non-random missing values can bias the model output



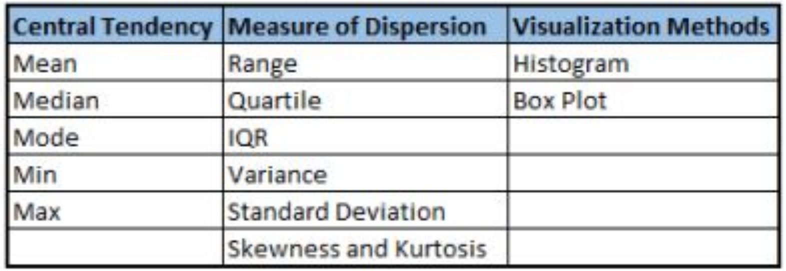
Mean/ Mode/ Median Imputation Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

Prediction Model we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable.

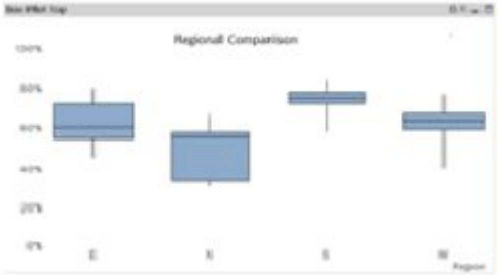


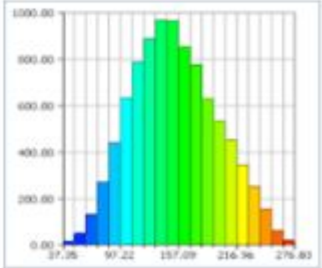
Univariate Analysis

At this stage, we explore variables one by one. Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.



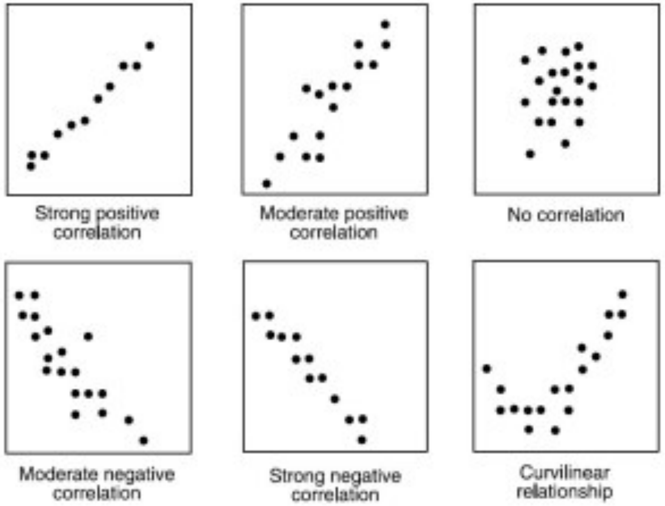
Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	



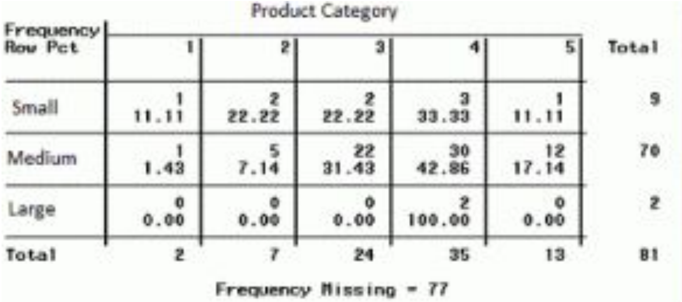


Continuous vs continuous

While doing bi-variate analysis between two continuous variables, we should look at scatter plots. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.

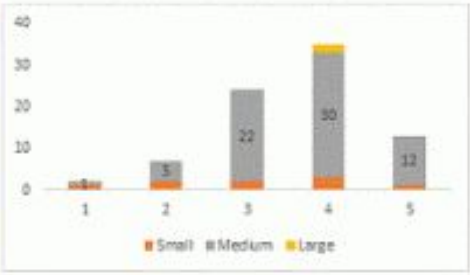


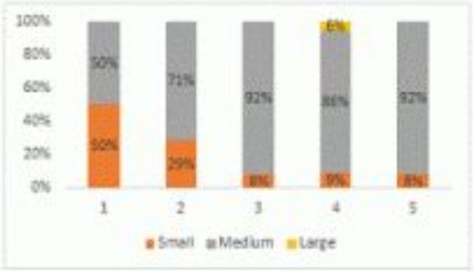
Categorical & Categorical



Frequency Row Pct	Product Category					
	1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77



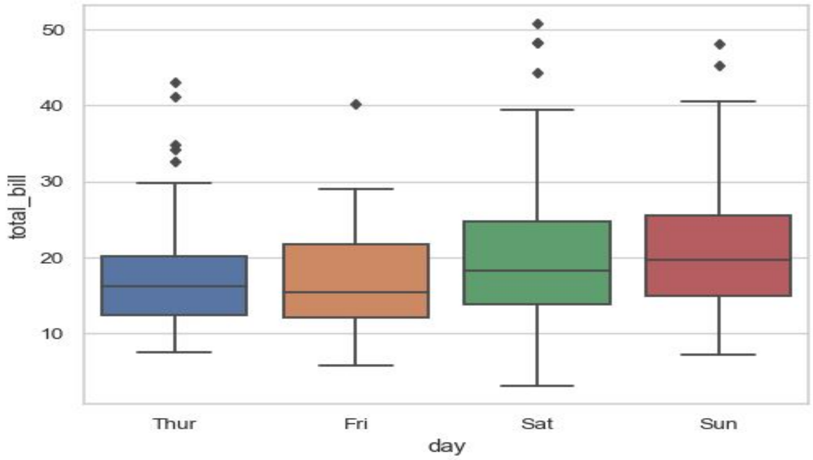


Two-way table: We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.

Stacked Column Chart: This method is more of a visual form of Two-way table.

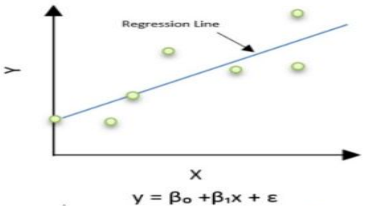
Categorical & continuous

upGrad



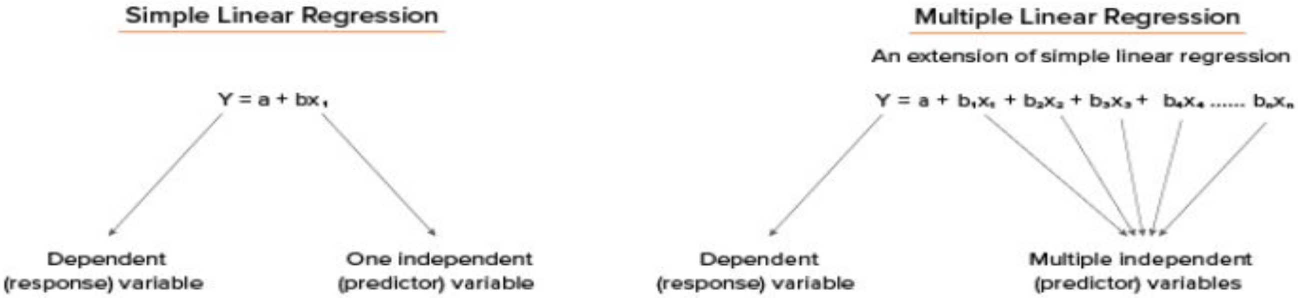
While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables

Linear Regression



What Is Linear Regression?

It is the simplest form of regression. It is a technique in which the **dependent variable is continuous** in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature.



Simple Linear Regression

$Y = a + bx_1$

Dependent (response) variable

One independent (predictor) variable

Multiple Linear Regression

An extension of simple linear regression

$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \ldots\ldots b_nx_n$

Dependent (response) variable

Multiple independent (predictor) variables

***Simple linear regression**, gets its adjective 'simple', because it concerns the study of only one predictor variable, similarly **multiple linear regression** concerns the study of two or more predictor variables.*

Linear Regression

Residual sum of squares (RSS)

This gives information about how much the target value varies around the regression line (predicted value).

$$\text{Error} = \sum_{i=1}^{n} (Actual\_output - predicted\_output)\text{\textasciicircum}2$$

The best-fit line is obtained by minimizing a quantity called Residual Sum of Squares (RSS) which could be optimized using Gradient Descent to get parameters of the best fit line.

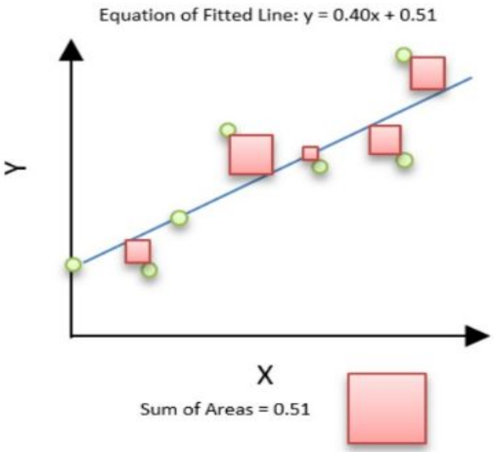
Total sum of squares (TSS)

This tells how much the data point move around the mean.

$$\text{Error} = \sum_{i=1}^{n} (\text{Actual\_output} - \text{average\_of\_actual\_output})\text{\textasciicircum}2$$

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination

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



$$\textbf{total variation} = \sum (y - \bar{y})^2$$

$$\textbf{explained variation} = \sum (\hat{y} - \bar{y})^2$$

$$\textbf{unexplained variation} = \sum (y - \hat{y})^2$$

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

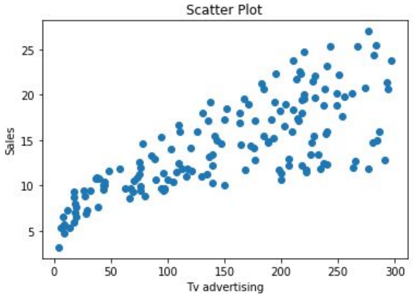
N = Total sample size.

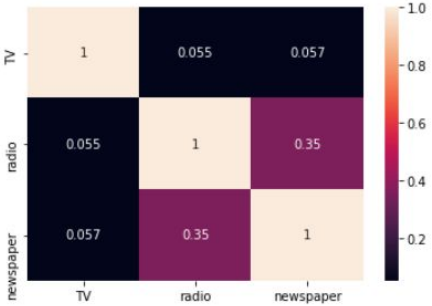
Assumptions of Linear Regression

upGrad

Linear Relationship between the features and target

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.





Little or no Multicollinearity between the features

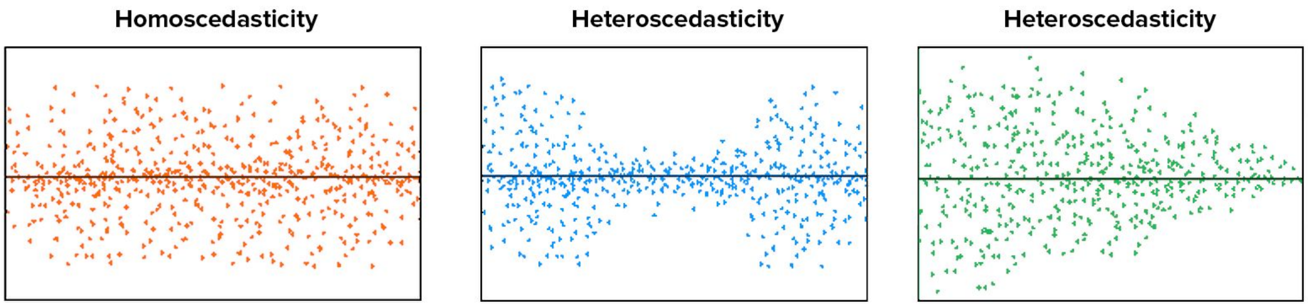
Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heat maps(correlation matrix) can be used for identifying highly correlated features.

http://people.duke.edu/~rnau/testing.htm
https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

Assumptions of Linear Regression

Homoscedasticity Assumption

Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables



Homoscedasticity

Heteroscedasticity

Heteroscedasticity

Random Cloud (No Discernible Pattern)

Bow Tie Shape (Pattern)

Fan Shape (Pattern)

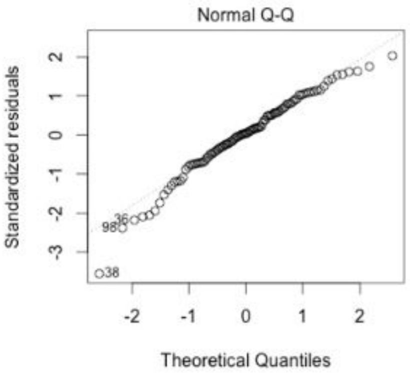
Assumptions of Linear Regression

Error terms are normally distributed with mean zero

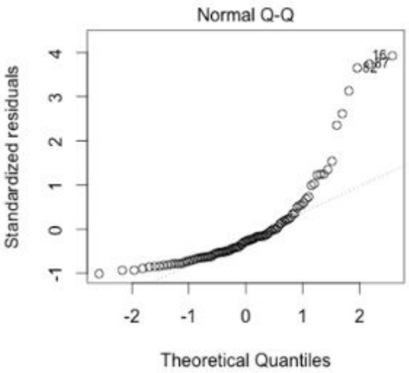
The fourth assumption is that the error(residuals) follow a normal distribution. One particular repercussion of the error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable.

Normal distribution of the residuals can be validated by plotting a q-q plot.

Normal Distribution Evident

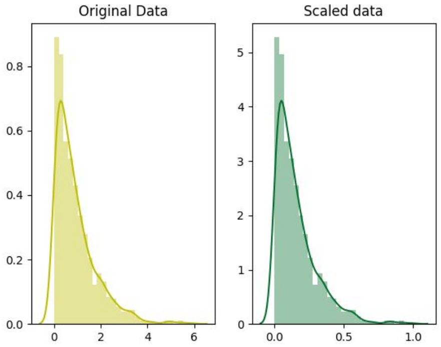


Not Normally Distributed



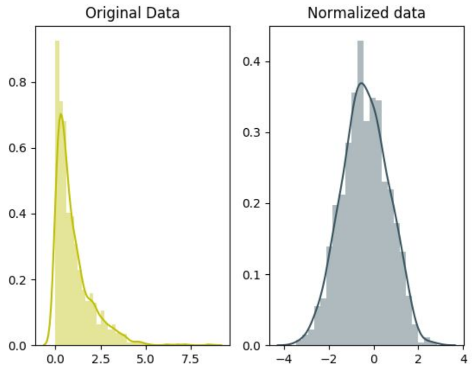
Scaling vs Standardization

Scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].



Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important. For example, in the dataset containing prices of products; without scaling, SVM might treat 1 USD equivalent to 1 INR though 1 USD = 65 INR.

Standardization *(also called **z-score normalization**)* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1



You need to normalize our data if you're going use a machine learning or statistics technique that assumes that data is normally distributed e.g. t-tests, ANOVAs, linear regression, linear discriminant analysis (LDA) and Gaussian Naive Bayes.

upGrad

#LifeKoKaroLift

Thank You!