# CHAPTER 4: Regression Machine learning Algorithms

## Introduction

In the previous chapter, we studied what is feature scaling and feature selection. In this chapter, we are going to look at different regression algorithms that you can spot-check on your dataset.

**Regression Analysis**

**Regression analysis is a set of statistical processes for estimating the relationship between a dependent variable (often called the ‘outcome variable’ or ‘target variable’) and one or more independent variables (often called ‘predictors’ or ‘features’).**

Regression analysis is an important tool for analyzing and modeling data. Here, we fit a line to the data points, in such a manner that the difference between the distances of the actual data points from the plotted line is minimum.

**The use of Regression**

Regression analyses the relationship between two or more features. Let’s take an example:

Let’s suppose we want to make an application that predicts the chances of admission of a student to a foreign university. In that case, the

The benefits of using Regression analysis are as follows:

* It shows the significant relationships between the label (dependent variable) and the feature (independent variable).
* It shows the extent of the impact of multiple independent variables on the dependent variable.
* It can also measure these effects even if the variables are on a different scale.

These features enable the data analysts to find the best set of independent variables for predictions.

## Linear Regression

Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with.

Building blocks of a Linear Regression Model are:

* Discrete/Continuous independent variable
* A best-fit regression line
* Continuous dependent variable. I.e., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

**Y = MX + C**

Chart, diagram, line chart

Description automatically generated

Where,

* Y = Dependent variable
* M = Slope
* X = Independent variable
* C = Constant or Intercept

**Problem Statement:**

This data is about the amount spent on advertising through different channels like TV, Radio, and newspapers. The goal is to predict how the expense on each channel affects the sales and is there a way to optimize the sales?

Importing the necessary packages

Text

Description automatically generated

Loading and exploring the data

Text

Description automatically generated with low confidence

Table

Description automatically generated

What are the **features**?

1. TV: Dollars spent on TV ads for a single product in each market (in thousands of dollars).
2. Radio: Dollars spent on Radio ads.
3. Newspaper: Dollars spent on Newspaper ads.

What is the **response**?

* Sales: sales of a single product in each market (in thousands of widgets)

Dimensions of the data

Graphical user interface, text

Description automatically generated

Find the missing values from different columns

Text

Description automatically generated

Let’s showcase the relationship between the feature and target variables

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

From the relationship diagrams above, it can be observed that there seems to be a linear relationship between the features TV ad, Radio ad, and the sales is almost a linear one. A linear relationship typically looks like this:

Shape

Description automatically generated

Hence, we can build a model using the Linear Regression Algorithm.

## Simple Linear Regression

**Simple Linear Regression**

**A simple linear regression is a method for predicting a quantitative response using a single feature (“input variable”).**

The mathematical equation is:

***𝑦* =*𝛽*0 + *𝛽*1*𝑥***

What do terms represent?

* Y: Response or Target Variable
* X: Feature Variable
* 𝛽1: Coefficient of X
* 𝛽0: Intercept

𝛽0 and 𝛽1 are the **model coefficients**. To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the Sales!

**Estimating (“Learning”) Model Coefficients**

**The coefficients are estimated using the Ordinary Least Squares estimates criterion (OLS], i.e., the best fit line must be calculated that minimizes the sum of squared residuals (or “Sum of Squared Error”)**

### Ordinary Least Squared Estimation

Chart, line chart

Description automatically generated

* In each two-dimensional space, an infinite number of lines can be plotted through the scatter diagram between two variables.
* Calculate the distance between the observed values and predicted values, which are called Errors or Residuals.
* Errors can be positive and negative, therefore square the error term (squared Error).
* Repeat the process with all the infinite number of lines and identify the line with a minimum sum of squared error.
* Hence the name “Ordinary least Squared”.

### The mathematics involved

Take a quick look at the plot created. Now consider each point and know that each of them has a coordinate in the form (X, Y). Now draw an imaginary line between each point and the current "best-fit" line. We'll call the distance between each point and the current best-fit line as D. To get a quick image of what we're trying to visualize, look at the picture below:

Chart, line chart

Description automatically generated

What elements are present in the diagram?

* The red points are the observed values of X and Y.
* The blue line is the least square line.
* The green lines are the residuals, which is the distance between the observed values and the least squared line.

Before, we're labeling each green line as having a distance D, and each red point as having a coordinate of (X, Y). Then we can define our best fit line as the lines having the property were:

𝐷21+𝐷22+𝐷23+𝐷24+....+𝐷2𝑁

So how do we find this line? The least-square line approximating the set of points:

**(𝑋,𝑌)1,(𝑋,𝑌)2,(𝑋,𝑌)3,(𝑋,𝑌)4,(𝑋,𝑌)5,**

Has the equation:

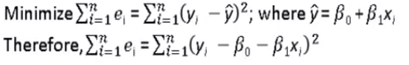
***𝑦* =*𝛽*0 + *𝛽*1*𝑥***

This is just a rewritten form of the standard equation for a line:

**Y = MX + C**

### Derivation of OLS by Minimizing Errors

Minimize the sum of squared error term by substituting as below



Solving the above equation by calculus – partial differencing by ***𝛽*0 and *𝛽*1** respectively and solving for the two variables we get the below equations

Text, letter

Description automatically generated

Building Simple Linear Regression Model to predict the sales based on TV ads

Graphical user interface, text, application, email

Description automatically generated

### Interpreting the model

How do we interpret the coefficient for spends on TV ads(β1)?

* A “unit” increase in spending on a TV ad is **associated with** a 0.04753 “unit” increase in sales.
* Or, an additional $1,000 on TV ads is translated to an increase in sales by $47.53.

### Prediction using the model

If the expense on a TV ad is $50000, what will be the sales prediction for that market?

***𝑦* =*𝛽*0 + *𝛽*1*𝑥***

Y = 7.032594 + 0.047537 \* (50)

Graphical user interface, text, application

Description automatically generated

Thus, we would predict Sales of 9,409 widgets in that market.

Let’s do the same thing using code.

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Plotting the Least Squares Line

Text

Description automatically generated

A picture containing application

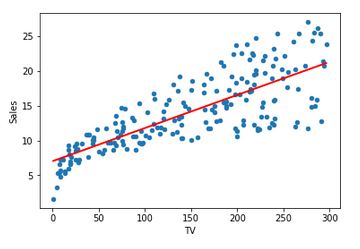
Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated



### Model Confidence

**Question:** Is linear regression a low bias/high variance model or a high bias/low variance model?

**Answer:** It's a High bias/low variance model. Even after repeated sampling, the best fit line will stay roughly in the same position (low variance), but the average of the models created after repeated sampling won't do a great job in capturing the perfect relationship (high bias). Low variance is helpful when we don't have less training data!

If the model has calculated a 95% confidence for our model coefficients, it can be interpreted as follows: If the population, from which this sample is drawn, is **sampled 100 times**, then approximately **95 (out of 100) of those confidence intervals** shall contain the "true" coefficients.

We will be discussing more in detail about bias and variance in the coming sections.

Text

Description automatically generated

Table

Description automatically generated with low confidence

### Hypothesis Testing and p-values

**Hypothesis testing** is closely related to confidence intervals. We start with a **null hypothesis** and an **alternate hypothesis** (that is opposite to the null). Then, we check whether the data **reject the null hypothesis** or **fails to reject the null hypothesis**.

The conventional hypothesis test is as follows:

* **Null hypothesis:** No relationship exists between TV advertisements and Sales (and hence *𝛽*1 equals zero).
* **Alternative hypothesis:** There exists a relationship between TV advertisements and Sales (and hence, *𝛽*1 is not equal to zero).

How do we test this? We reject the null hypothesis (and thus believe the alternative hypothesis) if the 95% confidence interval **does not include zero**. The **p-value** represents the probability of the coefficient is zero.

Text

Description automatically generated

Table

Description automatically generated with medium confidence



Chart

Description automatically generated with medium confidence

If the 95% confidence interval **includes zero**, the p-value for that coefficient will be **greater than 0.05**. If the 95% confidence interval **does not include zero**, the p-value will be **less than 0.05**.

Thus, a p-value of less than 0.05 is a way to decide whether there is any relationship between the feature in consideration and the response or not. Using 0.05 as the cutoff is just a convention.

In this case, the p-value for TV ads is way less than 0.05, and so we **believe** that there is a relationship between TV advertisements and Sales.

Note that we generally ignore the p-value for the intercept.

## Multiple Linear Regression

**Multiple Linear Regression**

**A multiple linear regression is a method for predicting a quantitative response using a multiple feature (“input variable”).**

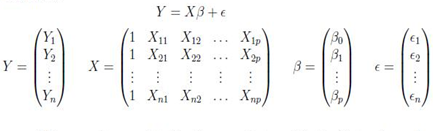
Till now, we have studied models based on only one feature. Now, we’ll include multiple features and create a model to see the relationship between those features and the target column. This is called **Multiple Linear Regression**.

**A picture containing text, antenna

Description automatically generated**

### Estimation of model parameters

Consider the model where



The columns of X are each covariate for the n patients, with the first column being all 1’s to include the intercept in the model.

Based on this model we get the following expansion for the first subject:



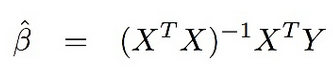
Then using matrix calculus we find that the least square estimate for β is given by



Hence, the least-squares regression line is



The beta values are obtained by calculating below equation



Note:

1. (XTX)-1 should be a non-singular matrix, otherwise, we cannot calculate beta
2. Beta cannot be calculated if columns are linear combinations of others.

We know that

Text, letter

Description automatically generated

H is called a Hat matrix

Properties of Hat matrix:

1. H is symmetric; HT = H
2. H is idempotent; HH = H

## Assumptions of OLS Regression

Regression is a parametric approach. ‘Parametric’ means it makes assumptions about data for analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver good results with data sets that don’t fulfill its assumptions. Therefore, for a successful regression analysis, it’s essential to validate these assumptions.

So, how would you check (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical tests.

Let’s look at the important assumptions in regression analysis:

1. There should be a linear and additive relationship between the dependent (response) variable and the independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X1 is constant, regardless of the value of X1. An additive relationship suggests that the effect of X1 on Y is independent of other variables.
2. There should be no correlation between the residual (error) terms. The absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. The absence of this phenomenon is known as multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to as heteroskedasticity.
5. The error terms must be normally distributed.

**What if these assumptions get violated?**

Let’s dive into specific assumptions and learn about their outcomes (if violated):

1. **Linear and Additive:**  If you fit a linear model to a non-linear, non-additive dataset, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen dataset.

**How to check:** Look for residual vs fitted value plots (explained below).

1. **Autocorrelation:** The presence of correlation in error terms drastically reduces a model's accuracy. This usually occurs in time series models where the next instant is dependent on the previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it causes confidence intervals and prediction intervals to be narrower. A narrower confidence interval means that a 95% confidence interval would have a lesser probability than 0.95 that it would contain the actual value of coefficients. Let’s understand narrow prediction intervals with an example:

For example, the least square coefficient of X1 is 15.02 and its standard error is 2.08 (without autocorrelation). But in presence of autocorrelation, the standard error reduces to 1.20. As a result, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

**How to check:** Look for Durbin – Watson (DW) statistics. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, 0 < DW < 2 implies positive autocorrelation while 2 < DW < 4 indicates negative autocorrelation. Also, you can see residual vs time plots and look for the seasonal or correlated pattern in residual values.

1. **Multicollinearity:** This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of predictors with response variables. In other words, it becomes difficult to find out which variable is contributing to predicting the response variable.

Another point, with the presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.

Also, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, you’ll end up with an incorrect conclusion that a variable strongly / weakly affects the target variable. Since, even if you drop one correlated variable from the model, its estimated regression coefficients would change. That’s not good!

**How to check:** You can use a scatter plot to visualize the correlation effect among variables. Also, you can use the VIF factor. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

1. **Heteroscedasticity:** The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influencing the model’s performance. When this phenomenon occurs, the confidence interval for out-of-sample prediction tends to be unrealistically wide or narrow.

**How to check**: You can look at the residual vs fitted values plot. If heteroscedasticity exists, the plot would exhibit a funnel shape pattern. Also, you can use Breusch-Pagan / Cook – Weisberg test or the White general test to detect this phenomenon.

1. **Normal Distribution of error terms:** If the error terms are non-normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. The Presence of non–normal distribution suggests that there are a few unusual data points that must be studied closely to make a better model.

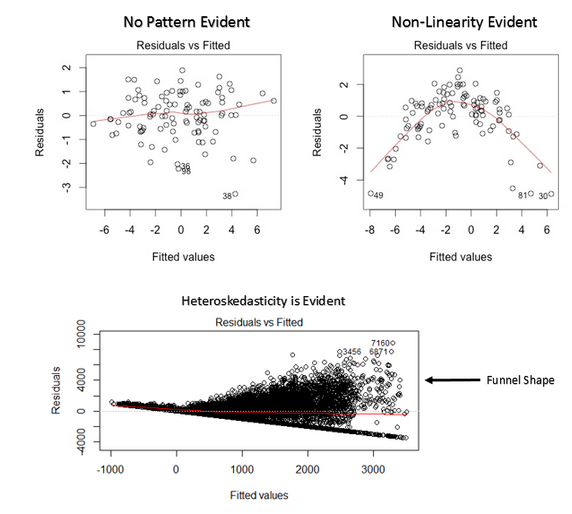
**How to check:** You can look at the QQ plot (shown below). You can also perform statistical tests of normality such as the Kolmogorov-Smirnov test, Shapiro-Wilk test.

## Interpretation of Regression Plots

Until here, we’ve learned about the important regression assumptions and the methods to undertake, if those assumptions get violated.

But that’s not the end. Now, you should know the solutions also to tackle the violation of these assumptions. In this section, I’ve explained the 4 regression plots along with the methods to overcome limitations on assumptions.

**Residual vs Fitted Values**



This scatters plot shows the distribution of residuals (errors) vs fitted values (predicted values). It is one of the most important plots which everyone must learn. It reveals various useful insights including outliers. The outliers in this plot are labeled by their observation numbers which make them easy to detect.

There are two major things which you should learn:

If there exists any pattern (may be, a parabolic shape) in this plot, consider it as signs of non-linearity in the data. It means that the model doesn’t capture non-linear effects.

If a funnel shape is evident in the plot, consider it as the signs of non-constant variance i.e. heteroscedasticity

**Solution:** To overcome the issue of non-linearity, you can do a nonlinear transformation of predictors such as log (X), √X, or X² transform the dependent variable. To overcome heteroscedasticity, a possible way is to transform the response variable such as log(Y) or √Y. Also, you can use a weighted least square method to tackle heteroscedasticity.

Normal Q-Q Plot

Chart, line chart

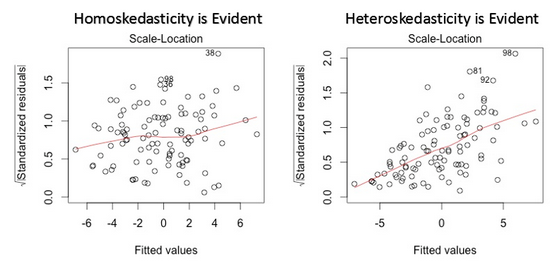
Description automatically generated

This q-q or quantile-quantile is a scatter plot that helps us validate the assumption of normal distribution in a data set. Using this plot, we can infer if the data comes from a normal distribution. If yes, the plot would show a straight line. The absence of normality in the errors can be seen with deviation in the straight line.

If you are wondering what ‘quantile’ is, here’s a simple definition: Think of quantiles as points in your data below which a certain proportion of data falls. Quantile is often referred to as percentiles. For example: when we say the value of the 50th percentile is 120, it means half of the data lies below 120.

**Solution:** If the errors are not normally distributed, non–a linear transformation of the variables (response or predictors) can bring improvement in the model.

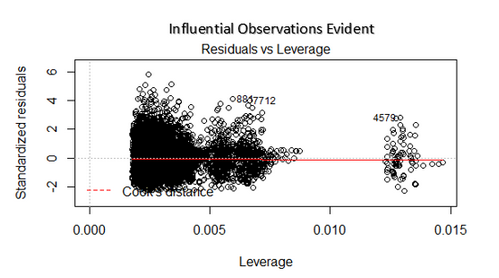
Scale Location Plot



This plot is also used to detect homoscedasticity (assumption of equal variance). It shows how the residuals are spread along with the range of predictors. It’s similar to the residuals vs fitted value plot except it uses standardized residual values. Ideally, there should be no discernible pattern in the plot. This would imply that errors are normally distributed. However, if the plot shows any discernible pattern (probably a funnel shape), it will imply non-normal distribution of errors.

Solution: Follow the solution for heteroskedasticity given in plot 1.

Residuals vs Leverage Plot



It is also known as Cook’s Distance plot. Cook’s distance attempts to identify the points which have more influence than other points. Such influential points tend to have a sizable impact on the regression line. In other words, adding or removing such points from the model can completely change the model statistics.

But, can these influential observations be treated as outliers? This question can only be answered after looking at the data. Therefore, in this plot, the large values marked by the cook’s distance might require further investigation.

**Solution:** For influential observations which are nothing but outliers, if not many, you can remove those rows. Alternatively, you can scale down the outlier observation with maximum value in data or else treat those values as missing values.

## Hands-on

**Problem Statement**

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems 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.

A US bike-sharing provider BikeIndia has recently suffered considerable dips in its revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain itself in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.

In such an attempt, BikeIndia aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all around and stand out from other service providers and make huge profits.

They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

Which variables are significant in predicting the demand for shared bikes. How well those variables describe the bike demands Based on various meteorological surveys and people's styles, the service provider firm has gathered a large dataset on daily bike demands across the American market based on some factors.

**Business Goal:**

We are required 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.

**Step1: Importing the Required Libraries**

**A picture containing text

Description automatically generated**

**Step2: Importing the dataset**



A picture containing table

Description automatically generated

Let’s perform some basic descriptive statistics

Table

Description automatically generated

Text

Description automatically generated with medium confidence

Graphical user interface, table

Description automatically generated

Finding:

* Dataset has 730 rows and 16 columns
* Except for one column, all others are float or integer types.
* One column is a date type.
* Looking at the data, there seems to be some categorical but treated as float/integer type.
* We will analyze and finalize whether to convert them to categorical or treat them as an integer.

**Step3: Data Quality Check**

**Check for NULL/MISSING values**

Graphical user interface, text, application

Description automatically generated

There are no missing/null values in a dataset.

**Checking for Duplicate values**

Graphical user interface, text, application

Description automatically generated

Application

Description automatically generated with medium confidence

**Insights:** The shape after running the drop duplicate command is the same as the original data frame. Hence, we can conclude that there were zero duplicate values in the dataset.

**Step4: Removing redundant & unwanted columns**

Based on the high-level look at the data and the data dictionary, the following variables can be removed from further analysis:

1. **instant**: It’s only an index value
2. **today**: This has the date since we already have separate columns for 'year' & 'month', hence, we could live without this column.
3. **casual & registered**: Both these columns contain the count of bikes booked by different categories of customers. Since our objective is to find the total count of bikes and not by specific category, we will ignore these two columns. Moreover, we have created a new variable to have the ratio of these customer types.
4. We will save the new data frame as bike\_new, so that the original dataset is preserved for any future analysis/validation.

Text

Description automatically generated

A picture containing chart

Description automatically generated

Table

Description automatically generated

**Step5: Creating Dummy Variables**

We will create Dummy variables for 4 categorical variables ‘mnth’, ‘weekday’, ‘season’ & ‘weathersit’. Before creating dummy variables, we will have to convert them into ‘category’ data types.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated with medium confidence

We are going to create dummy variable, drop the original variable for which the dummy was created and drop the first dummy variable for each set of dummies created.



Table

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

**Step6: Splitting the Data**

* Splitting the data to Train and Test – We will now split the data into Train and Test (70:30 ratio)
* We will use train\_test\_split method from the sklearn package for this

Text, letter

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Step7: Exploratory Data Analysis**

Let’s perform EDA on the Train dataset.

**Visualizing Numeric Variables**

Create a new data frame contains only numeric variables.

A picture containing diagram

Description automatically generated

A picture containing text, tree, map

Description automatically generated

The above Pair-Plot tells us that there is a Linear Relation between ‘temp’, ‘atemp’, and ‘cnt’

**Visualizing Categorical Variables**

Let’s build a boxplot of all categorical variables (before creating dummies) against the target variable 'cnt' to see how each of the predictor variables stack up against the target variable.

Text

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

There were six categorical variables in the dataset. We used a Box plot to study their effect on the dependent variable (‘cnt’).

The inference that We could derive was:

* Season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 bookings (for 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates season can be a good predictor for the dependent variable.
* mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 bookings per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
* weathersit: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 bookings (for 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
* holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is biased. This indicates holidays CAN NOT be a good predictor for the dependent variable.
* weekday: weekday variable shows a very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence on the predictor. I will let the model decide if this needs to be added or not.
* workingday: Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 bookings (for 2 years). This indicates working day can be a good predictor for the dependent variable

**Correlation Matrix**

Let's check the correlation coefficients to see which variables are highly correlated.

Note: here we are considering only those variables (dataframe: bike\_new) that were chosen for analysis

Text, letter

Description automatically generated

Chart, timeline, treemap chart

Description automatically generated

**Insights:**

* The heatmap clearly shows which all variables are multicollinear, and which variables have high collinearity with the target variable.
* We will refer to this map back-and-forth while building the linear model so as to validate different correlated values along with VIF & p-value, for identifying the correct variable to select/eliminate from the model.

**Step8: Rescaling the Features**

Graphical user interface, text, application, email

Description automatically generated

Apply scaler() to all the numeric variables

Diagram

Description automatically generated with low confidence

Graphical user interface, application, table

Description automatically generated

**Step9: Building A Linear Model**

Dividing into X and Y sets for the model building

Text

Description automatically generated with medium confidence

Recursive Feature Elimination: We will be using the LinearRegression function from Scikit Learn for its compatibility with RFE (which is a utility from sklearn)

Graphical user interface, text

Description automatically generated

Table

Description automatically generated with low confidence

Text, letter

Description automatically generated

Text

Description automatically generated

Creating X\_test data frame with RFE selected variables.



**Step10: Building Linear Model using ‘STATS MODEL’**

**Model 1**

**VIF check**

Text

Description automatically generated

Table

Description automatically generated with low confidence

Text, letter

Description automatically generated

Check the parameters obtained

Table

Description automatically generated

Print a summary of the linear regression model obtained

Table

Description automatically generated

**Model 2**

Removing the variable ‘atemp’ based on its high p-value & high VIF

A picture containing text

Description automatically generated

**VIF Check**

Text

Description automatically generated

Table

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

Print a summary of the linear regression model obtained



Table

Description automatically generated

**Model 3**

* Removing the variable ‘hum’ based on its Very High ‘VIF’ value.
* Even though the VIF of ‘hum’ is second highest, we decided to drop ‘hum’ and no ‘temp’ based on business knowledge that temperature can be an important factor for a business like bike rentals, and wanted to retain ‘temp’.



**VIF Check**

Text

Description automatically generated

Table

Description automatically generated with medium confidence

Table

Description automatically generated

Print a summary of the linear regression model obtained

Table

Description automatically generated

**Model 4**

* Removing the variable ‘season3’ based on its very high VIF value.



**VIF Check**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Text

Description automatically generated with medium confidence

Print a summary of the Linear Regression model obtained

Table

Description automatically generated

**Model 5**

Removing the variable ‘mnth\_10’ based on the very high p-value.



**VIF Check**

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Text, email

Description automatically generated

Print a summary of the linear regression model obtained

Table

Description automatically generated

**Model 6**

Removing the variable ‘mnth\_3’ based on high p-value

A picture containing text

Description automatically generated

**VIF Check**

Text

Description automatically generated

Table

Description automatically generated

The Key Value pair of the of features and their coefficients.

Graphical user interface, text, email

Description automatically generated

Table

Description automatically generated

**Insights**

This model looks good, as there seems to be Very Low Multicollinearity between the predictors and the p-values for all the predictors seems to be significant. For now, we will consider this as our final model (unless the Test data metrics are not significantly close to this number).

**Step11: Final Model Interpretation**

**Hypothesis Testing**

Hypothesis testing states that:

* H0:B1=B2=...=Bn=0
* H1: at least one Bi!=0

**lr6 model coefficient values**

* const 0.084143
* yr 0.230846
* workingday 0.043203
* temp 0.563615
* windspeed -0.155191
* season\_2 0.082706
* season\_4 0.128744
* mnth\_9 0.094743
* weekday\_6 0.056909
* weathersit\_2 -0.074807
* weathersit\_3 -0.306992

**Insights**

* From the lr6 model summary, it is evident that all our coefficients are not equal to zerowhich means **We REJECT the NULL HYPOTHESIS**

**F Statistics**

F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant the Model is.

* F-statistic: 233.8
* Prob (F-statistic): 3.77e-181

The F-Statistics value of 233 (which is greater than 1) and the p-value of '~0.0000' states that the overall model is significant

**The equation of best fitted surface based on model Ir6:**

cnt = 0.084143 + (**yr** × 0.230846) + (**workingday** × 0.043203) + (**temp** × 0.563615) − (**windspeed** × 0.155191) + (**season2** × 0.082706) + (**season4** ×0.128744) + (**mnth9** × 0.094743) + (**weekday6** ×0.056909) − (**weathersit2** × 0.074807) − (**weathersit3** × 0.306992)

**Interpretation of Coefficients**

**temp:** A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable, increases the bike hire numbers by 0.5636 units.

**weathersit\_3:** A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable, decreases the bike hire numbers by 0.3070 units.

**yr:** A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2308 units.

**season\_4:** A coefficient value of ‘0.128744’ indicated that w.r.t season\_1, a unit increase in season\_4 variable increases the bike hire numbers by 0.128744 units.

**windspeed:** A coefficient value of ‘-0.155191’ indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.155191 units.

**workingday:** A coefficient value of ‘0.043203’ indicated that, a unit increase in workingday variable increases the bike hire numbers by 0.043203 units.

**season\_2:** A coefficient value of ‘0.082706’ indicated that w.r.t season\_1, a unit increase in season\_2 variable decreases the bike hire numbers by 0.082706 units.

**mnth\_9:** A coefficient value of ‘0.094743’ indicated that w.r.t mnth\_1, a unit increase in mnth\_9 variable increases the bike hire numbers by 0.094743 units.

**weekday\_6:** A coefficient value of ‘0.056909’ indicated that w.r.t weekday\_1, a unit increase in weekday\_6 variable increases the bike hire numbers by 0.056909 units.

**weathersit\_2:** A coefficient value of ‘-0.074807’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit2 variable, decreases the bike hire numbers by 0.074807 units.

**const:** The Constant value of ‘0.084143’ indicated that, in the absence of all other predictor variables (i.e. when x1,x2...xn =0), The bike rental can still increase by 0.084143 units.

**Step12: Assumptions**

Error terms are normally distributed with mean zero (not X and Y).

Chart, histogram

Description automatically generated

**Insights**

From the above histogram, we could see that the residuals are normally distributed. Hence, our assumption for Linear Regression is valid.

**Linear assumption between X & Y**

Graphical user interface

Description automatically generated with low confidence

A picture containing shoji, building, sofa, crossword puzzle

Description automatically generated

**Insight**

Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor ‘cnt’

**Multicollinearity between the Predictors**

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

**Insight**

From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5

**Dividing into X\_test and y\_test**

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Selecting the variables that were part of final model.



Adding constant variable to test data frame

A picture containing text

Description automatically generated

Text, table

Description automatically generated with medium confidence

Making predictions using the final model (lr6)

Graphical user interface, text

Description automatically generated

**Step13: Model Evaluation**

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

**R2 Value for Test**

Graphical user interface, text, application

Description automatically generated

**Adjusted R2 Value for Test**



Text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

**Result Comparison**

* Train R^2 :0.824
* Train Adjusted R^2 :0.821
* Test R^2 :0.820
* Test Adjusted R^2 :0.812
* This seems to be a good model that can very well 'Generalize' various datasets.

**Step14: Final Report**

As per our final Model, the top 3 predictor variables that influences the bike booking are:

* **Temperature (temp)** - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
* **Weather Situation 3 (weathersit\_3)** - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
* **Year (yr)** - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

#### So, it's suggested to consider these variables utmost importance while planning, to achive maximum Booking

The next best features that can also be considered are

* **season\_4:** - A coefficient value of ‘0.128744’ indicated that w.r.t season\_1, a unit increase in season\_4 variable increases the bike hire numbers by 0.128744 units.
* **windspeed:** - A coefficient value of ‘-0.155191’ indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.155191 units.

#### NOTE:

* + The details of weathersit\_1 & weathersit\_3
  + **weathersit\_1:** Clear, Few clouds, Partly cloudy, Partly cloudy
  + **weathersit\_3:** Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

The details of season1 & season4

* **season1:** spring
* **season4:** winter

**Summary**

We have discussed, what is simple linear regression, multiple linear regression, and their assumptions. Then, we have seen the practical implementation to validation of assumptions and finally we build a model on the training set and evaluated the performance of the model using R squared and Adjusted R squared.

## Assessment

**Choose the appropriate option**

1. **Which One of the following are regression tasks?**
   1. Predict the age of a person
   2. Predict the country from where the person comes from
   3. Predict whether the price of petroleum will increase tomorrow
   4. Predict whether a document is related to science
2. **Which of the following plots is used for normality test?**
   1. Scatter plot
   2. Bar plot
   3. qqplot
   4. None of these
3. **Which of the following tests is used for heteroscedasticity?**
   1. AD
   2. Ljung-Box
   3. Breusch-Pagan
   4. All of the above
4. **Which of the following tests is used for autocorrelation?**
   1. AD
   2. Ljung-Box
   3. Breusch-Pagan
   4. White test
5. **VIF > 10 is said to be**
   1. No Multicollinearity
   2. Less Multicollinearity
   3. High Multicollinearity
   4. None of the above

**Fill in the spaces with appropriate answers**

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ diagrams are graphs of the data that are helpful in displaying the relationship between variables.
2. The SSR is sometimes referred to as the variability in Y explained by \_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. If the adjusted R2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_ when a new variable is added, it would be an indication that the variable should not remain in the model.
4. Regression analysis is sometimes called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
5. Complete the following equation: SST = SSR +

**True or False**

1. Simple Linear regression is built between more than two variables.
   1. True
   2. False
2. Multiple Linear regression is built between two or more than two variables
   1. True
   2. False
3. Autocorrelation is having relationship between observation
   1. True
   2. False
4. We drop the variable from the model if p-value is insignificant
   1. True
   2. False
5. Difference between actual and predicted values is called errors (or) residuals
   1. True
   2. False

## Programming Assessment

Using the data in the below URL, Perform the following tasks

https://github.com/fenago/MLBook/blob/main/Chapter%204/Code/datasets/mtcars.csv

1. Import the data
2. Perform Data Cleaning
3. Perform EDA
4. Build the Regression Model
5. Validate the Assumptions

## Assessment Solutions

**Choose the appropriate options**

1. A
2. C
3. C
4. B
5. C

**Fill in the spaces with appropriate answers**

1. Scatter
2. The regression equation
3. Decreases
4. Least-Squares Regression
5. SSE

**True or False**

1. False
2. True
3. True
4. True
5. True