A Short Monograph on Regression Analysis

TO SERVE AS A REFRESHER FOR LINEAR REGRESSION

Index

Contents

List of Figures

List of Tables

1. Introduction

Today, we are living in a world dominated by data. Thanks to impressive technology and data tracking systems, computers are capable of capturing and storing massive amounts of data. For example, a car sales company like Maruti is in a position to collect enormous amount of data on different aspects of its business, such as car sales, economy of the country, stock market prices, crude oil prices etc. Environment activists are able to record proportions of different greenhouse gases in the atmosphere in minute details and are interested in understanding how the fluctuations of these gases in different times of day in different seasons are responsible for climatic pollution and human health.

The immediate next step after data collection is data analysis and interpretation. Consider the car sales example of Maruti. The question of how the economy of a country affects car sales, and possibly vice versa, is a rather important question. For this, the industry needs to understand the dependence between car sales and various macro-economic factors. For environmentalists, knowing how the average global temperature fluctuates with fluctuations in the proportions of the greenhouse gases, is a question of crucial importance.

Also, whether there is at all any direct relation among these features, or the relationship is controlled by one or more unobserved factors, need to be objectively determined. In both the cases mentioned above, it is required to provide a definitive relationship between the variable of interest (or *response*) and the other variables (or *predictors*) which are used to understand it. This helps in clarifying the dependence between the variables concerned.

Another very important aspect of data analysis, particularly in the above two examples cited, is **prediction**. Maruti would like to predict car sales for future, given the expected measure of economy growth or recession that year, so that it can take informed decisions regarding its business. Similarly, environmentalists would be interested in knowing how reduction in global warming will be effected by decreasing concentrations of the greenhouse gases and by how much. For both these objectives, one needs to form a model which explicitly relates the response to the predictors.

Linear regression is one of the simplest statistical tools used to analyse the dependence of a response on one or more predictors. In this monograph, we discuss how it works, what information it gives, and what one needs to be careful about while performing linear regression.

2. What is Regression?

Regression analysis is one of the most commonly used tools to find a relationship (linear or non-linear) between a response and one or more predictors and exploit that relationship in predicting the expected value of the response for certain values of the predictor(s) with maximum accuracy possible.

Two important terms and corresponding notations are given below:

Dependent variable or Response: It is the variable of interest that one wants to model or predict using one or more variables whose values are known.

Independent variable(s) or Predictor(s): It is assumed that the response depends on one or more predictors. These variables are independent and a model is formulated identifying the explicit relationship between the response and the predictor(s).

Types of Regression: In this monograph two types of regression are discussed in detail:

I. **Simple Linear Regression:** When the response is assumed to have a linear dependence on one <u>single</u> predictor.

II. **Multiple Linear Regression:** When the response is assumed to have a linear dependence on <u>multiple</u> predictors.

Important Note: Regression is a widely applicable tool which can and does incorporate all types of non-linear relationships also. Linear regression is only a very small subset of all possible regression functions applicable in various domains.

Case Study:

A top wine manufacturer wants to invest in new technologies to improve its wine quality. Wine quality is directly dependent on the amount of alcohol in wines and the smoothness which, in turn, are controlled by various chemicals either directly added during the manufacturing process or generated through various chemical reactions. Wine certification and quality assessment are key elements for wine gradation and its price ticket. Wine certification is determined by various physiochemical elements in the wine. Therefore, the company wants to estimate the percentage (%) of alcohol in a bottle of wine as a function of various chemical components of the wine.

Statement of the Problem: Regress alcohol percentage on the chemical components present in the wines.

Description of the data (Data Dictionary):

Variables	**Description**
fixed acidity (FA)	Number of grams of Tartaric acid per cubic decimetre, $(g(tartaric\ acid)/dm^3)$
volatile acidity (VA)	Number of grams of Acetic acid per cubic decimetre, $(g(acetic\ acid)/dm^3)$
citric acid (CA)	Number of grams of Citric acid per cubic decimetre, (g/dm^3)
residual sugar (RS)	Number of grams of Residual sugar per cubic decimetre, (g/dm^3)
Chlorides	Number of grams of Sodium chloride per cubic decimetre, $(g(sodium\ chloride)/dm^3)$
free sulphur dioxide (FSD)	Number of milligram of Free sulphur dioxide per cubic decimetre, (mg/dm^3)
total sulphur dioxide (TSD)	Number of milligram of total sulphur dioxide per cubic decimetre, (mg/dm^3)
Density	Number of grams per cubic centimetre (g/cm^3)
Ph	pH is a scale of acidity from 0 to 14.
Sulphates	Number of grams of Potassium sulphate per cubic decimetre, $(g(potassium\ sulphate)/dm^3)$
Brand (categorical)	three different brands of wine are considered where 1 represents "Grover Zampa", 2 represents "Seagram" and 3 represents "Sula Vineyards".
Alcohol (response)	percentage volume of alcohol in wine $(\%\ vol.)$

It is possible to regress alcohol on each of the 10 continuous predictors one at a time and explore their individual relationships. However, in order to avoid repetition, only one predictor (density) has been considered.

Before a regression model of alcohol on the predictors can be built, it is necessary to investigate whether there exists any dependence among the observed variables.

3. What is Pairwise Correlation?

The formal definition of Correlation: The pairwise correlation coefficient (denoted as r) measures the degree of relatedness between two variables.

The correlation coefficient (r) developed by Karl Pearson is termed as **Pearson product-moment correlation coefficient**. The term r is a measure of the strength and the direction of the linear relationship between two **continuous** variables. Mathematically it can be expressed as:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{Cov\ (X,Y)}{SD(X)\ SD(Y)}$$

Here, X and Y are the pair of attributes measured on n units. From the i-th unit a pair of observations (x_i, y_i) is obtained, and $\bar{x}$ and $\bar{y}$ are the respective means. The value of r ranges from -1 to +1, both values included.

$\boldsymbol{r = -1}$: denotes a perfect negative correlation between X and Y. If (x, y) pairs are plotted, the points would fall on a straight line with negative slope. This indicates that X and Y are perfectly linearly related but, if one variable increases, the other decreases.

$\boldsymbol{r = 0}$: denotes that no correlation exists between X and Y.

$\boldsymbol{r = +1}$: denotes a perfect positive correlation between X and Y data. If (x, y) pairs are plotted, the points would fall on a straight line with positive slope. This indicates that X and Y are perfectly linearly related and if one variable increases, the other also increases.

Often the plot of (x, y) pairs is known as *scatterplot.*

In Fig 1, the first scatterplot corresponding to r = 0 clearly shows that there is no discernible pattern in the plot. The other two scatterplots exhibit moderate correlations of equal magnitude (0.6) but in opposite directions.

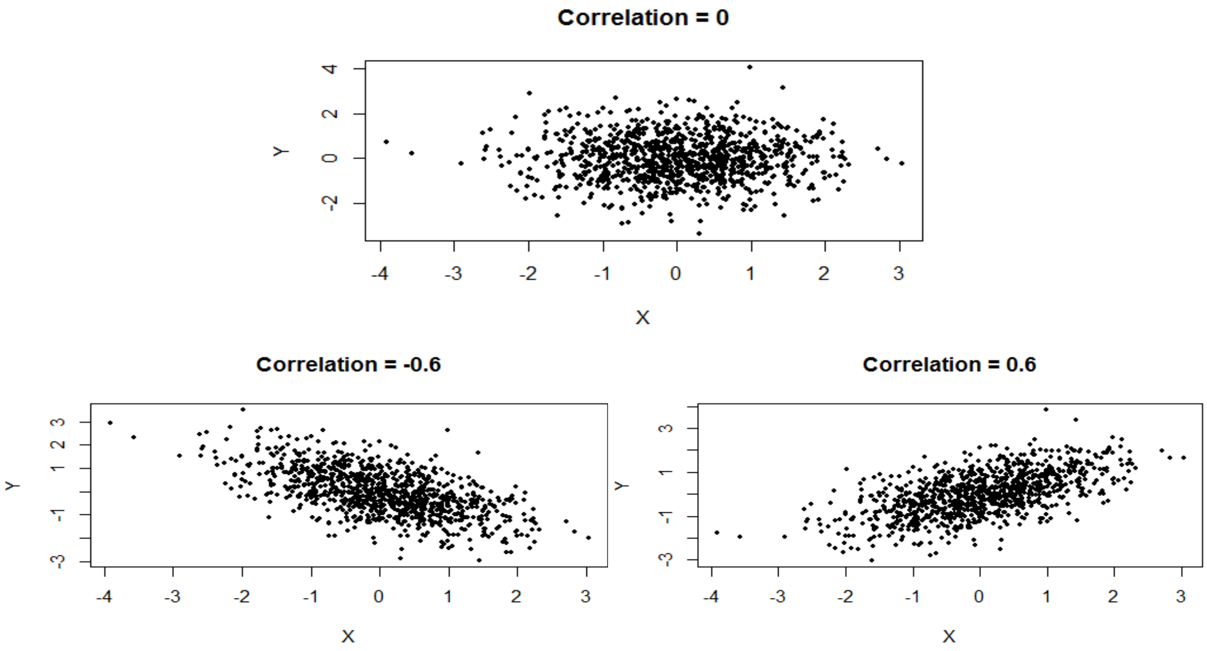


Fig. 1: Three different correlations

Case Study continued.

The primary objective here is to determine whether any dependence exists between alcohol (%) and three selected chemical components, FA, pH and density individually.
Exploratory Analysis (EDA) on Wine Data

We first upload the "WineData" dataset into Python.

```
#Step 1: Import important packages into python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import pearsonr
from statsmodels.formula.api import ols
from statsmodels.graphics.gofplots import ProbPlot
import statsmodels.api as sm
from scipy import stats

#Step 2: Read the dataset into python using read_csv
WineData = pd.read_csv('WineData.csv')

WineData.shape
(1599, 13)

WineData.columns
```

```
Index(['ID', 'Brand', 'FA', 'VA', 'CA', 'RS', 'chloride', 'FSD', 'TSD',
'density', 'sulphate', 'pH', 'alcohol'],dtype='object')
```

```
WineData.head()
```

ID	Brand	FA	VA	CA	RS	chloride	FSD	TSD	density	pH	sulphate	alcohol
1	Seagram	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
2	Seagram	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8
3	Seagram	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8
4	Sula Vineyards	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8
5	Seagram	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4

The dataset contains 1599 observations on 13 variables. The first column is “ID”, which is just a label and will not be used in the analysis. The second column is “Brand”, which is the only categorical variable present in the data.

The five number summary of each of the quantitative variables is presented below.

For Brand the proportions in each category is shown.

```
# Summary of all variables
WineData[WineData.columns[2:13]].describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
FA	1599.0	8.32	1.74	4.60	7.10	7.90	9.20	15.90
VA	1599.0	0.53	0.18	0.12	0.39	0.52	0.64	1.58
CA	1599.0	0.27	0.19	0.00	0.09	0.26	0.42	1.00
RS	1599.0	2.54	1.41	0.90	1.90	2.20	2.60	15.50
chloride	1599.0	0.09	0.05	0.01	0.07	0.08	0.09	0.61
FSD	1599.0	15.87	10.46	1.00	7.00	14.00	21.00	72.00
TSD	1599.0	46.47	32.90	6.00	22.00	38.00	62.00	289.00
density	1599.0	1.00	0.00	0.99	1.00	1.00	1.00	1.00
ph	1599.0	3.31	0.15	2.74	3.21	3.31	3.40	4.01
sulphate	1599.0	0.66	0.17	0.33	0.55	0.62	0.73	2.00
alcohol	1599.0	10.42	1.07	8.40	9.50	10.20	11.10	14.90

```
WineData['Brand'].value_counts()
 Seagram   Sula Vineyards    Grover Zampa
      633        553              413
Name: Brand, dtype: int64
```

The minimum % of alcohol in wine is 8.40 and maximum is 12.90 while the mean value is 10.45. We can see there are 3 brands of alcohol viz. “Grover Zampa”, “Seagram” and “Sula Vineyards”.

To visually explore if there is any association between alcohol and the three selected variables respectively correlation charts for (alcohol, FA), (alcohol, pH) and (alcohol, density) are obtained.

Note that correlation can only be defined for two variables which are continuous (or at least ordinal). ***Hence correlation cannot be defined between alcohol and Brand***.

```
# Correlation: Alcohol and Fixed Acidity
g = sns.scatterplot(data=WineData, x='FA', y='alcohol')
r = np.round(pearsonr(WineData['FA'], WineData['alcohol'])[0],3)
g.text(x = 10,y = 14.5,s = "rho = "+str(r))
plt.show()
```

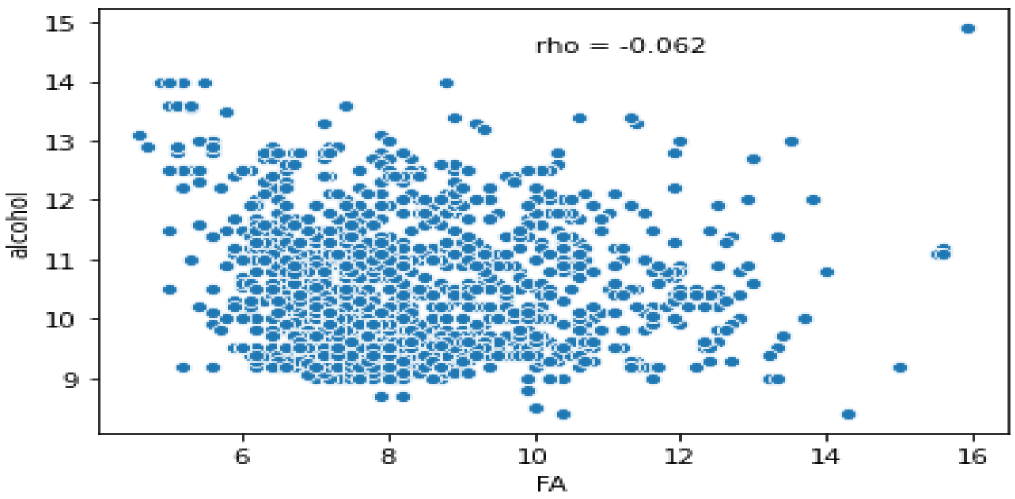


Fig. 2: Correlation between alcohol and FA

```
#  Correlation: Alcohol and pH
g = sns.scatterplot(data=WineData, x='pH', y='alcohol')
r = np.round(pearsonr(WineData['pH'], WineData['alcohol'])[0],3)
g.text(x = 3.4,y = 14.5,s = "rho = "+str(r))
plt.show()
```

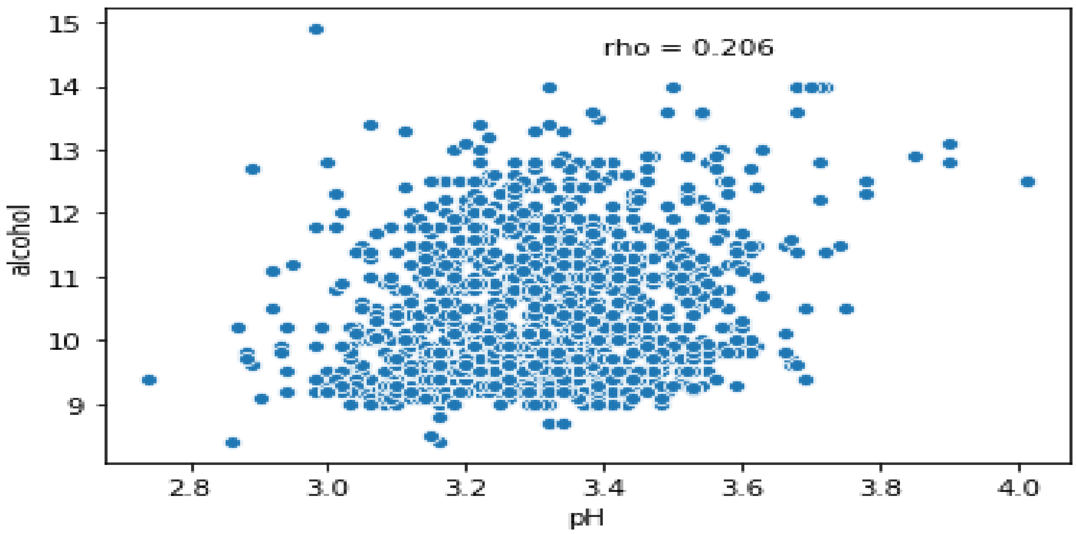


Fig. 3: Correlation between alcohol and pH

```
# Correlation: Alcohol and density
g = sns.scatterplot(data=WineData, x='density', y='alcohol')
r = np.round(pearsonr(WineData['density'], WineData['alcohol'])[0],3)
g.text(x = 0.996,y = 14.5,s = "rho = "+str(r))
plt.show()
```

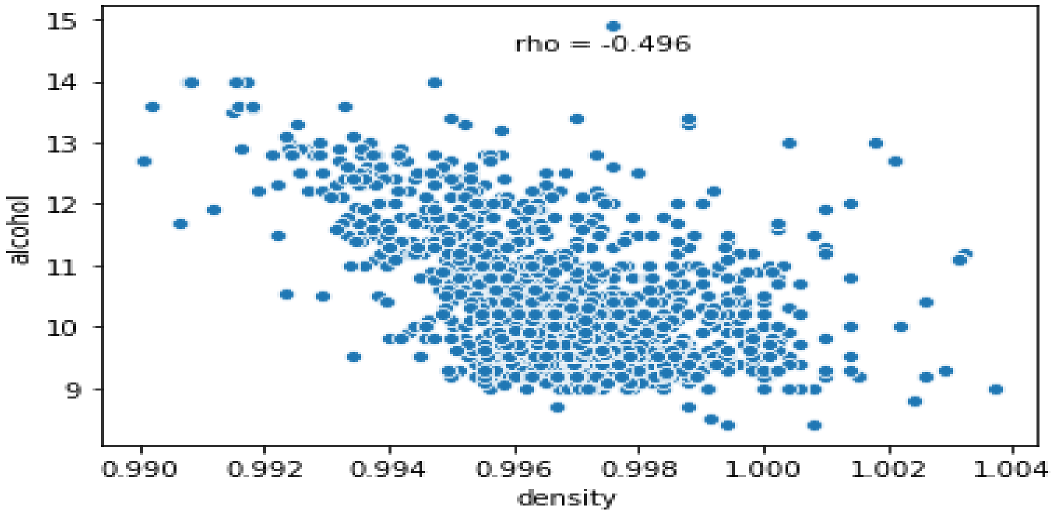


Fig. 4: Correlation between alcohol and density

Interpretations of Fig. 2 − 4 are given below

i) Fig. 1: Correlation between alcohol (Y) and X_1 (FA) is $r = -0.062$. Though the correlation is negative and statistical significance is also there, for all practical purpose correlation between alcohol and FA is taken to be non-existent.

ii) Fig. 2: Correlation between alcohol (Y) and X_2 (pH) is $r = 0.206$. This indicates positive dependence but the numerical value is very small.

iii) Fig. 3: Correlation between alcohol (Y) and X_3 (density) is $r = -0.496$. This indicates moderate dependence between alcohol and density and as density increases, alcohol (%) reduces.

NOTE: Correlation is only useful in determining linear relationship between two variables. Zero correlation may imply no linear dependence, but that does not preclude any other form of dependence (such as, polynomial) between the variables concerned.

Further, correlation does not indicate any cause and effect relationship. Correlation simply quantifies how well two variables are related, if at all, and its direction (i.e. positive or negative).

Another important observation about correlation is that even when correlation is significantly different from zero, it may not have any practical significance. Typically, the rule of thumb for correlation values is as follows:

i) r between −0.4 and +0.4 indicates absence of linear dependence

ii) r between −0.7 and −0.4 or r between +0.4 and +0.7 indicates moderate linear dependence, the sign indicating its direction
iii) r less than −0.7 or r greater than +0.7 indicates strong linear dependence

Correlation coefficient, however, cannot determine what value the response will take for a given value of the predictor(s). Regression model building is necessary for that.

4. Simple Linear Regression (SLR)

4.1 Definition

The formal definition of Simple Linear Regression: Let n pairs of observations (x_i, y_i), i = *1, 2, …, n*, be available on two features, one of which is assumed to depend on the other. Typically, y denotes the dependent variable and X the independent variable. Let the quantity $E(Y)$ denote the expected or mean value of Y.
Simple linear regression relates the response Y to the single predictor variable X through a straight line. The mathematical formulation of the simple linear regression line is:

$$E(Y) = \beta_0 + \beta_1 X$$

where,

y: is the value of the continuous response (or dependent) variable,

β_0 and β_1: are intercept and slope coefficients, respectively, and known as the regression parameters.

X: represents the independent (predictor) variable continuous in nature.

It is assumed that the expected value of the response is a linear function of the predictor. When β_0 and $\beta_1$are known, a given value of the predictor will specify the expected value of the response.

However, since Y is a random variable, not all values y_i will be equal to $E(Y)$. Another form of SLR equation is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

ϵ: represents the unobservable error term. Note that error here does not indicate any mistake, simply the difference between the expected and observed values of the response.

The error terms provide crucial insight into the regression process.

4.2 The method of Ordinary Least Square (OLS)

The simple linear regression model involves unknown parameters β_0 and β_1, which need to be estimated from data. There are several different methods of estimating the parameters. The simplest and the most widely used method is known as the Ordinary Least Squares method (OLS).
Given n pairs of observations on Y and X, the objective is to minimise the sum of squared errors and thus get appropriate estimates of β_0 and β_1. We want to minimise

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Explicitly minimizing the above equation, the following estimates of β_0 and β_1 are available.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}, \qquad \text{and}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$\hat{\beta}_1$ has an equivalent representation $\hat{\beta}_1 = \frac{r s_y}{s_x}$ where r is the correlation coefficient between X and Y, and s_x and s_y are the respective standard deviations of X and Y. Since s_x and s_y are both positive quantities, $\hat{\beta}_1$ has the same sign as r. So for two positively correlated variables, the $\hat{\beta}_1$ will be positive and for two negatively correlated variables, $\hat{\beta}_1$ will be negative.

Case Study continued.

The objective here is to determine the dependence of alcohol (%) on density. Though any one of the 10 continuous predictors may have been used as predictor, the choice is made based on a higher correlation between the response and the predictor. A scatterplot of alcohol (%) versus density helps to get a visual impression about whether a linear function of density will at all be suitable to describe alcohol (%).

```
# Scatter plot of alcohol vs. density
a4_dims = (10,5)
fig, ax = plt.subplots(figsize=a4_dims)
a = sns.scatterplot(x="density", y="alcohol", data=WineData)
plt.show()
```

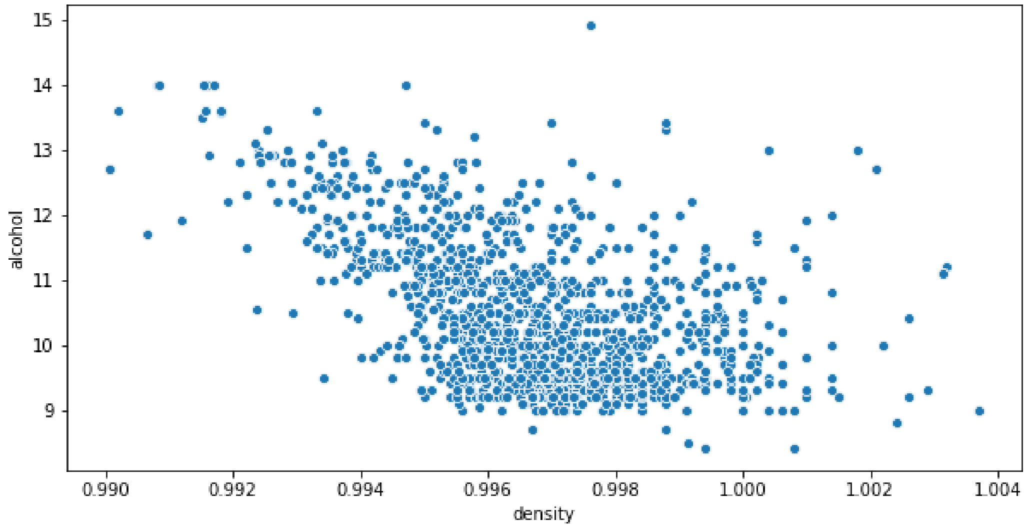


Fig. 5: Scatterplot of density and alcohol (%)

The scatterplot above suggests that a linear relationship between alcohol and density may exist since the majority of the points seem to fall on a straight line. We also expect the slope to be negative and hence, increase in density is expected to decrease alcohol (%).

Recall that the correlation between alcohol (%) and density is -0.496.

A linear regression model is fit with alcohol (%) as the response and density as the predictor.

```
# Regression: Alcohol on density
mod = ols('alcohol ~ density', data = WineData).fit()
intercept , density_slope = mod.params
equation = "\n Y = {}".format(round(density_slope,2))+"*X +"+" {}".format(round(intercept,2))
print(equation)
Y = -280.16*X + 289.68
```

Thus the OLS line has the form

$$\widehat{Alcohol}\,(\%) = 289.68 - 280.16 \times density$$

The *hat* symbol is used to indicate that the regression gives an estimate of the response.

We first note that the sign of $\hat{\beta}_1$ is negative. This shows that the two variables are inversely related, that is, if one increases, the other decreases. This confirms our expectation that the variables alcohol (%) and density increase/decrease in opposite directions and we get a straight line with negative slope.

The sign of the regression slope and the correlation coefficient will always be the same.

The regression slope is the measure of change in the response with one unit change in predictor. The sign of the regression slope indicates the direction of the change.

The value of $\hat{\beta}_1$ indicates that if density increases by 1 unit, the estimated alcohol (%) decreases by 280.16 unit. However, the response being in percentage, its values are bounded between 0 and 100. So is our interpretation misleading? Actually, from the EDA, it can be seen that the minimum and maximum values of density are mostly between 0.99 and 1.00. Hence the range of density is approximately 0.01. While interpreting a regression parameter, it is important to pay heed to the range of values of the predictor. In this case density cannot be expected to change by 1 unit, but the changes in the density will be in the order of 1/1000.

Hence, the following statement will be appropriate in this case:

If density increases by 0.001 unit, then alcohol (%) decreases by 0.28 unit.

The intercept term is the estimated value of the response when the predictor is 0. However, the intercept term is not always interpretable, such as in this case.

The following graph shows the OLS regression line (in blue) through the scatterplot.

```
# Plotting
a4_dims = (10,5)
fig, ax = plt.subplots(figsize=a4_dims)
a = sns.regplot(x="density", y="alcohol", data=WineData)
a.set_title('Model:'+equation,fontsize=15)
# Displaying the plot
```

```
plt.show()
```

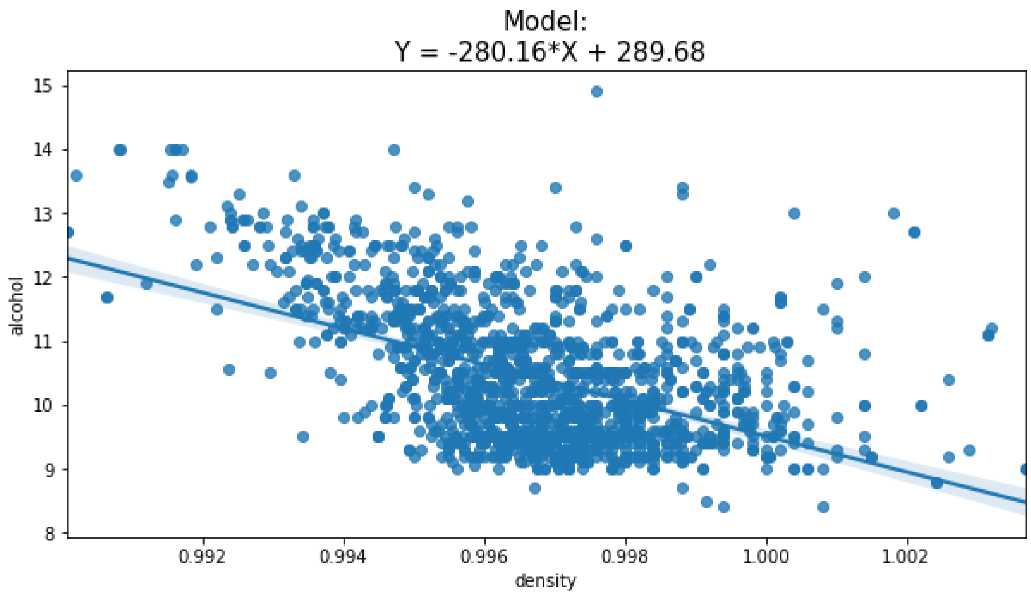


Fig. 6: Regression of alcohol (%) on density

Given a particular value of density, using the estimated values of β_0 and β_1, estimated or fitted values of alcohol (%) can be obtained. The fitted values may or may not be equal to the observed value of the response.

Let us look at alcohol (%) in the wines at several levels of density.

Observation	Density	Observed response (Y)	Estimated response ($\hat{Y} = 289.68 - 280.16 \times density$)	Residual ($\hat{\epsilon} = Y - \hat{Y}$)
355	0.9912	11.9	11.98	−0.08
481	1.0026	9.2	8.79	0.41
609	1.0026	10.4	8.79	1.61
954	0.99458	12.1	11.04	1.06
1198	0.99458	9.8	11.04	−1.24
1201	0.99458	9.8	11.04	−1.24
1460	0.99458	11.9	11.04	0.86

Table 1: Observed response, estimated response (based on density) and residuals

It is clear from the above table that the pairs (x_i, y_i) is not unique but $(x_i, \hat{y}_i)$ are. At a given level of density, there may be several observed values of alcohol (%). When density is 0.99458, alcohol content may range from 9.8% to 12.1%. On the other hand, 11.9% alcohol may be found when density is 0.9912 as well as when density is 0.99458. However, estimated alcohol content is uniquely defined through the regression equation for a given value of density.

The difference between the observed and fitted values of the response is called residual. Residuals may be both positive or negative.

4.2.1 Assumptions of OLS method

The linear regression model is defined on 4 important assumptions, often referred to as (***LINE***)
Assumption 1: The regression model is ***linear*** in the parameters i.e. $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$,

Assumption 2: The observations Y_i (or the error termsϵ_i) are ***independent***

Assumption 3: The error variables ϵ_i are ***normally*** distributed.

Assumption 4: The errors have no bias (that is,$E(\epsilon_i) = 0$) and they are *homoscedastic*, that is, they have ***equal variance***.

4.3 Examining the statistical significance of regression model

Fitting the regression model or simply estimating the regression coefficients is not enough. It is important to check if the regression slope is significantly different from 0 in the population.

4.3.1 Significance of regression slope

Estimating the regression coefficients is only the first step of regression model fitting. OLS will produce estimates of regression intercept and regression slope given any sample data. However, it is of vital importance to know whether the *population* regression slope is significantly different from 0. If in the population the regression slope is not significantly different from 0, then there is no regression of *Y* on *X*.
We now need to know, is density at all statistically significant in explaining alcohol (%)?

Let us consider the test of hypothesis $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ where β_1 is the regression coefficient of density when alcohol (%) is regressed on density.

```
# Significance of regression
print(mod.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                alcohol   R-squared:                       0.246
Model:                            OLS   Adj. R-squared:                  0.246
Method:                 Least Squares   F-statistic:                     521.6
Date:                Wed, 22 Jan 2020   Prob (F-statistic):          3.94e-100
Time:                        16:49:11   Log-Likelihood:                -2144.1
No. Observations:                1599   AIC:                             4292.
Df Residuals:                    1597   BIC:                             4303.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    289.6753     12.227     23.691      0.000     265.692     313.659
```

```
density     -280.1638     12.267     -22.838      0.000    -304.226    -256.102
==============================================================================
Omnibus:                      147.785   Durbin-Watson:                   1.460
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              193.960
Skew:                           0.768   Prob(JB):                     7.62e-43
Kurtosis:                       3.743   Cond. No.                     1.06e+03
==============================================================================
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.06e+03. This might indicate that there are strong mu
lticollinearity or other numerical problems.
```

Observe that the p-value corresponding to density is very small and thus the null hypothesis $H_0\colon \beta_1 = 0$ is rejected which in turn indicates that density is significant in explaining alcohol (%)

Statistical significance alone, however, is not enough to decide whether the predictor is useful in explaining the variability in the response. Is density enough to explain a large part of variation in alcohol? This leads us to the concept of coefficient of determination, R^2.

4.3.2 The coefficient of determination $\mathbf{R^2}$

The coefficient of determination R^2 is a summary measure that explains how well the sample regression line fits the data. The rationale and computation of R^2 is discussed below:

Let $\hat{Y_\iota} = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. $\hat{Y_\iota}$ is the estimated value of the response, for a given value of the predictor X_i. Fig 6 and Table 1 show that not all predicted values of the response will be equal to the observed value Y_i. In fact, it may well happen that none of the estimated values of the response coincides with the corresponding observed values.

The difference between the observed and the estimated values of the response is called residual. Residual is the estimated value of the unobserved error component in the regression equation.

$\hat{\epsilon}_i$ = Y_i - $\hat{Y}_i$

Residuals are very important part of regression and they have many useful properties. In fact, it can be shown that $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$, if the estimation method is OLS.

$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$: is the total variation of the actual Y about their sample mean and termed as the ***total sum of squares (SST)***. This is closely linked to the sample variance of Y.

$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$: is the sum of squares due to regression and is called ***regression sum of squares (SSR)***

$\sum_{i=1}^{n} \hat{\epsilon}_i{}^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$: is the sum of squared differences between the observed and the predicted values of the response. This is known as the ***residual sum of squares or the error sum of squares (SSE)***

$$SST = SSR + SSE$$

$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$ (on dividing SST on both sides)

We now define R^2 as : $\mathrm{R}^2 = \frac{SSR}{SST}$ or $\mathrm{R}^2 = 1 - \frac{SSE}{SST}$

The numerical value of R^2 is equal to the square of the correlation coefficient between the response and the predictor *only if* there is a single predictor in the linear regression model, *i.e.* for simple linear regression only.

R^2 measures the proportion of the total variation in Y that is explained by the regression model. It ranges from $0 \leq R^2 \leq 1$. The higher the value of R^2, the more powerful is the predictor to predict the response. A regression model with high R^2 value indicates that the model fits the data well. In that case a high proportion of variance in the response is explained by the dependence of the response on the predictor.

For the current model, the value of R^2 turns out to be approximately 25% . This means, density is able to explain around 25% of the total variation in alcohol (%).

It is clear that, even though, the regression of alcohol (%) on density is significant (p-value < 0.05), density by itself is not sufficient to explain the variation in the response to a satisfactory degree. Later we will investigate if the other predictors in the data set will help in explaining more variation in the response (Multiple Linear Regression).

4.4 Residual Plot

Residuals are estimated errors and are defined as $\hat{\epsilon}_i = Y_i - \hat{Y}_i$.

Residuals have many important properties and are employed to check various regression assumptions.

Recall that $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$.

If the assumptions (Sec 4.2.1) of linear regression are satisfied, the residuals will form a band around the 0 line when plotted against the fitted values. The residuals should be randomly distributed around the horizontal line (that represents zero residual errors) i.e. there should not be a distinct trend in the distribution of points.

If the scatterplot shows any systematic deviations from this band shape, they may indicate various departures from the assumptions. For example, if the residuals show any funnel-out or funnel-in shape, then one may conclude that the error variances are not all equal. If the scatterplot shows sinusoidal pattern, that may indicate serial dependence among the observations.

Normal Q-Q plot provides evidence whether the errors are normally distributed. If the plotted residuals form a straight line making 45^0 angle with x-axis, then one may conclude that the errors are normally distributed. Even when the angle is not preserved, but the residuals form a straight line, we may conclude that the data is approximately normal.

Case Study continued.

Below we have shown the residual plot for the regression of alcohol (%) on density.

```
# Residual plot
def regression_plots(model,data):
    # model values
    model_fitted_y = model.fittedvalues
    # model residuals
    model_residuals = model.resid
    # normalized residuals
    model_norm_residuals = model.get_influence().resid_studentized_internal
    # absolute squared normalized residuals
    model_norm_residuals_abs_sqrt = np.sqrt(np.abs(model_norm_residuals))
    # absolute residuals
    model_abs_resid = np.abs(model_residuals)
    # leverage, from statsmodels internals
    model_leverage = model.get_influence().hat_matrix_diag
    # cook's distance, from statsmodels internals
    model_cooks = model.get_influence().cooks_distance[0]

    def graph(formula, x_range, label=None):
        x = x_range
        y = formula(x)
        plt.plot(x, y, label=label, lw=1, ls='--', color='red')
    ########################################################################
    fig, axes = plt.subplots(nrows=2,ncols=2)
    fig.set_size_inches(10, 10)
    a = sns.residplot(model_fitted_y, 'alcohol', data=data,lowess=True,scatter_kws={'alpha': 0.5},
    line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8} ,ax=axes[0][0])
    a.set_title("Residuals vs Fitted ",fontsize=15)
    a.set_xlabel('Fitted Values')
    a.set_ylabel('Residuals')
    ########################################################################
    axes[0][1].set_title("Normal Q-Q",fontsize=15)
    QQ = ProbPlot(model_norm_residuals)
    b = QQ.qqplot(line='45', alpha=0.5, color='#4C72B0', lw=1,ax=axes[0][1],ylabel = 'Standardized R
esiduals')
    ########################################################################
    c= sns.scatterplot(model_fitted_y, model_norm_residuals_abs_sqrt, alpha=0.5,ax=axes[1][0])
    sns.regplot(model_fitted_y, model_norm_residuals_abs_sqrt,
    scatter=False,ci=False,lowess=True,line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8},ax=axes[1][0]);
    c.set_title('Scale-Location',fontsize=15)
    c.set_xlabel('Fitted values')
    c.set_ylabel('$\sqrt{|Standardized Residuals|}$');
    ###########################################################################
    d= sns.scatterplot(model_leverage, model_norm_residuals, alpha=0.5,ax=axes[1][1]);
    sns.regplot(model_leverage, model_norm_residuals,
    scatter=False,ci=False,lowess=True,line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8},ax=axes[1][1]);
    d.set_xlim(0, max(model_leverage))
    d.set_ylim(-3, 5)
    d.set_title('Residuals vs Leverage',fontsize=15)
    d.set_xlabel('Leverage')
    d.set_ylabel('Standardized Residuals')
    p = len(model.params) # number of model parameters
    graph(lambda x: np.sqrt((0.5 * p * (1 - x)) / x),np.linspace(0.001, 0.200, 50),'Cook\'s distance') # 0
.5 line
    graph(lambda x: np.sqrt((1 * p * (1 - x)) / x),np.linspace(0.001, 0.200, 50)) # 1  line
    d.legend()
    plt.show()

regression_plots(mod,WineData)
```

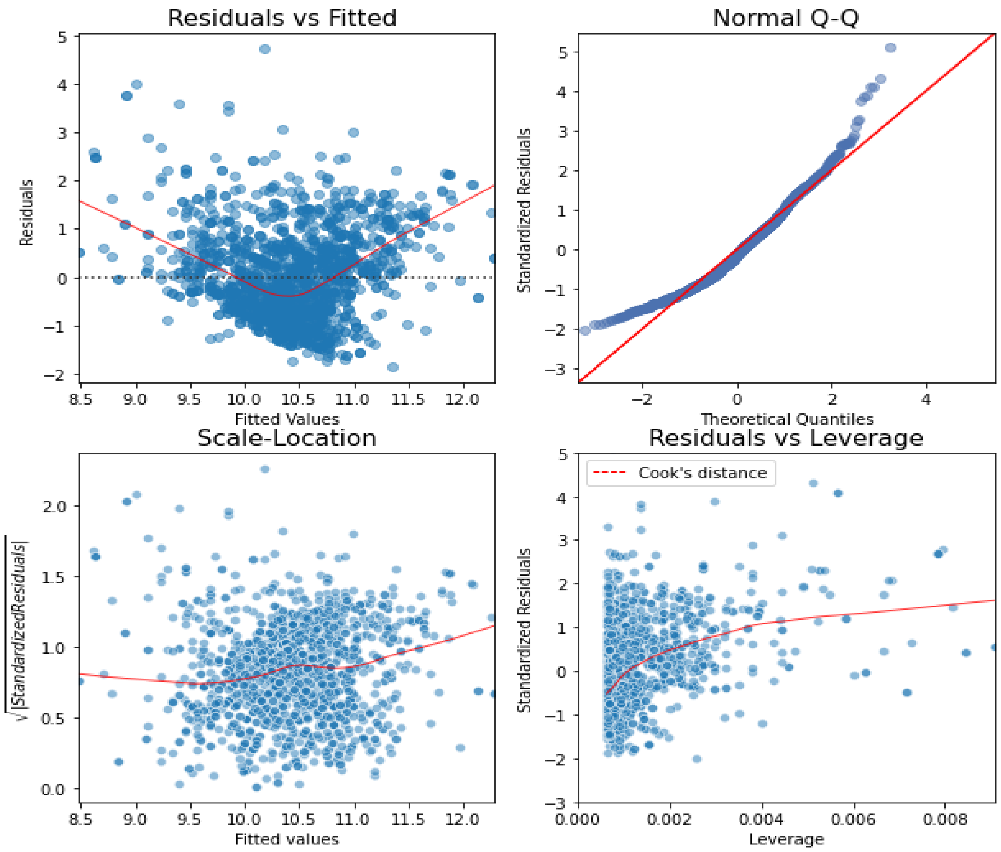


Fig. 7: Residual plot for alcohol vs. density

Using the above figure, we may infer that the regression assumptions are not violated.

It is clear from the discussions above that density, in spite of having significant slope when alcohol (%) is regressed on it, can explain only 25% of the total variability in the response. It is a natural step then to examine whether inclusion of the other predictors contribute towards explanation of the variability in the response, and if so, to what degree.

5. Multiple Linear Regression Analysis

In the previous sections we have investigated how one single predictor variable may be used to predict a response. In reality, the response depends on multiple variables. For example, the volume of car sales of a company depends not only on the GDP but also on other factors like crude oil prices, confidence index of the population and many other considerations. In general, the response variable is found to depend on several predictors simultaneously.

Multiple linear regression is one of the simplest statistical tools that is used to analyse the relationship of the response with several predictors. Like simple linear regression, this procedure fits a linear function of the predictors to the response.

5.1 Definition of Multiple Linear Regression Analysis

The formal definition of Multiple Linear Regression: Multiple regression is a statistical technique used to analyse relationship between a single dependent variable and several predictors simultaneously. Assume *n* (multivariate) sample observations $(x_{1i}, x_{2i}, \dots x_{ki}, y_i)$, *i = 1, 2, …, n*, are available. Y denotes the dependent variable and$X_1, X_2, \dots, X_k$ the independent variables.

The mathematical formulation of multiple linear regression line is:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where,

Y: is the value of the continuous response (or dependent) variable,

β_0 : is the intercept

X_j: represents the j^{th} independent (predictor) variable continuous in nature. *j = 1, …, k*

β_j: represents the coefficient of the j^{th} independent (predictor) variable.

It is assumed that the expected value of the response is a linear function of all the *k* predictors. When the regression coefficients β_0 and $\beta_j, j = 1, \dots, k$ are all known, or estimated, a given value of the predictor combination will specify the expected value of the response.

However, since Y is a random variable, not all values y_i will be equal to $E(Y)$. Another form of MLR equation is

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$, *i = 1, 2, …,n*

ε: represents the unobservable error term.

Formally the parameters $\beta_1, \beta_2, \dots, \beta_k$ are called **partial regression coefficients**, since the coefficient β_k measures the change in the value of the response due to unit change in the value of X_k, keeping all other variables fixed. We will refer to the $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ simply as regression coefficients.

5.2 The method of ordinary least squares

In case of multiple regression also the regression coefficients are estimated by minimizing the error sum of squares (SSE) $\sum_{i=1}^{n} \epsilon_i^2$.

Case Study continued.

Now we consider all the physiochemical liquids together and evaluate the regression model that would predict the alcohol (%) in wine. Before, running regression model it is important to look at correlations of all variables with respect to each other.

```
# Correlation: Alcohol versus all independent variables
plt.figure(figsize=(12,7))
corr = WineData[WineData.columns[1:13]].corr()
cmap = sns.diverging_palette(230, 20, as_cmap=True)
mask = np.triu(np.ones_like(corr, dtype=bool))
sns.heatmap(corr,annot=True,mask = mask,cmap = cmap)
plt.show()
```

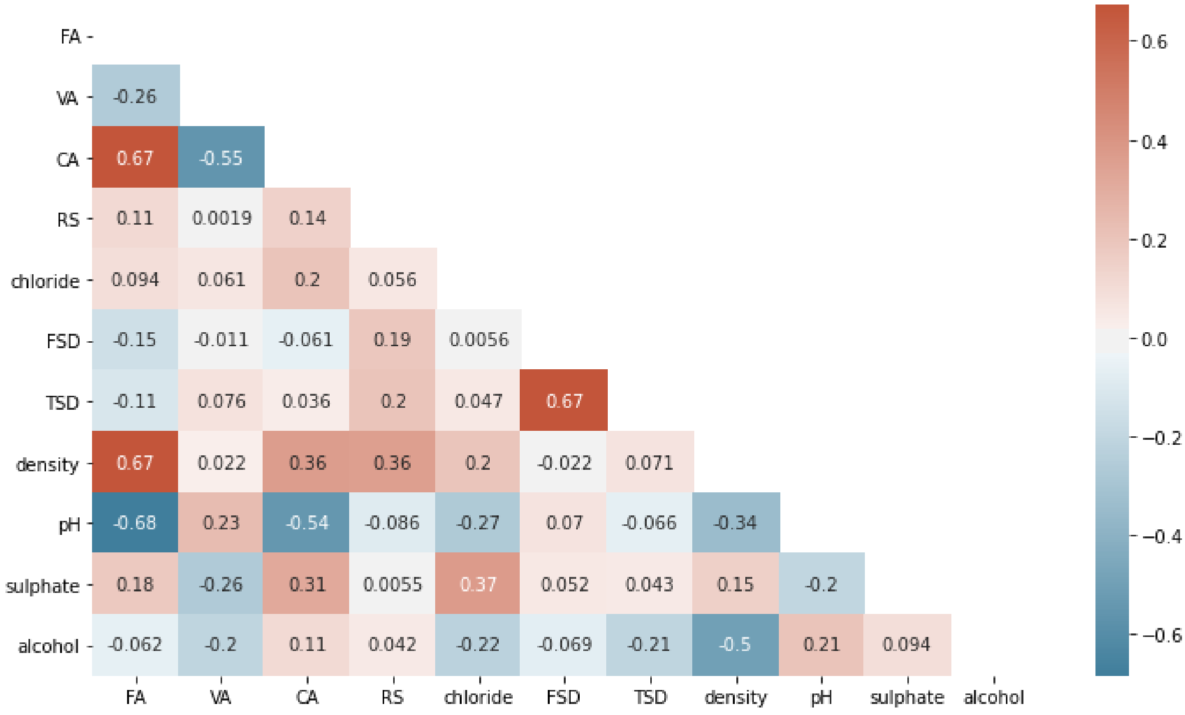


Fig. 8: Heatmap showing correlations among all variables

It may be observed that FA is positively correlated with CA with correlation coefficient 0.67. FA is also positively correlated with density (67%) and negatively correlated (−68%) with pH. All these are moderately high correlations. Similarly, FSD and TSD has moderately high correlation. Likewise, other correlations can also be observed.

Although almost all correlations have been shown to be statistically significant, we will treat only those which are above 0.4 or below −0.4 to be of any importance. Once we agree to impose this restriction, only a few variable pairs show substantial correlation.

5.3 When the predictor is categorical

Before the problem of fitting a multiple linear regression model is taken up, the case of categorical predictor needs to be explicitly discussed.

So far we have tacitly assumed that the response as well as the predictors are all continuous. The case of categorical response is not covered under MLR. Let us discuss the situation when one or more predictors are categorical. A set of ***dummy variables*** or ***indicator variables*** is introduced corresponding to each categorical variable. A discrete variable with m categories is represented by a set of $m - 1$ dummy variables.

Indicator coding: It is the most common format of dummy variable coding in which each category of the nominal variable is represented by either 1 or 0.

The equation of regression model and concept of the indicator coding using dummy variable is explained based on our case study.

If $X_1, \ldots, X_k$ are continuous then we can write the regression model of Y on X as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

But, when a predictor variable X is a categorical variable with m levels, small modification is required for the above.

Brand is a categorical predictor with three levels Grover Zampa, Seagram and Sula Vineyards. Hence $m - 1 = 3 - 1 = 2$ dummy variables need to be introduced in the model. The modified equation is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} D_1 + \beta_{k+2} D_2 + \epsilon$$

where, D_1 and D_2 are dummy variables taking values 1 or 0 as per the observation representation.

Case Study continued.

For an explicitly declared categorical variable, the corresponding dummy variable assignment is automatically done. The default choice of the baseline is that level whose name comes first in lexicographic ordering. The default ordering can be changed. In this case the baseline is taken as Grover Zampa since Grover Zampa, Seagram and Sula Vineyards are arranged in increasing lexicographic order.

Next we perform ***multiple linear regression*** in Python and the output is presented below.

```
# Multiple Linear Regression of alcohol on all variables
```

```
mod = ols('alcohol ~ FA+VA+CA+RS+chloride+FSD+TSD+density+sulphate+pH+Brand', data = Wine
Data).fit()
coefficients = mod.params
print(coefficients)
Intercept                   585.744496
Brand[T.Seagram]             -0.235467
Brand[T.Sula Vineyards]      -0.002991

FA                   0.509555
VA                   0.423617
CA                   0.823537
RS                   0.272455
chloride            -1.303246
FSD                 -0.002695
TSD                 -0.001520
density           -595.004881
sulphate             1.084578
pH                   3.617430
dtype: float64
```

Regression coefficients corresponding to Brand actually acts as constants. For three levels of Brand, three different regression equations are obtained.

BrandSeagram takes the value 1 if Brand = Seagram, 0 otherwise

BrandSula Vineyards takes the value 1 if Brand = Sula Vineyards, 0 otherwise

When Brand = Grover Zampa, then both these variables take the value 0.

Explicit form of the regression equations are:

(1) Brand = Grover Zampa:

$\hat{Y} = \underline{585.7} + 0.51FA + 0.424VA + 0.824CA + 0.272RS - 1.303Chloride - 0.003FSD + -0.002TSD - 595.005density + 1.085sulphate + 3.617pH$

(2) Brand = Seagram

$\hat{Y} = 585.7 + 0.51FA + 0.424VA + 0.824CA + 0.272RS - 1.303Chloride - 0.003FSD + -0.002TSD - 595.005density + 1.085sulphate + 3.617pH - 0.23$

or

$\hat{Y} = \underline{585.47} + 0.51FA + 0.424VA + 0.824CA + 0.272RS - 1.303Chloride - 0.003FSD + -0.002TSD - 595.005density + 1.085sulphate + 3.617pH$

(3) Brand = Sula Vineyard

$\hat{Y} = 585.7 + 0.51FA + 0.424VA + 0.824CA + 0.272RS - 1.303Chloride - 0.003FSD + -0.002TSD - 595.005density + 1.085sulphate + 3.617pH - 0.003$

or

$\hat{Y} = \underline{585.697} + 0.51FA + 0.424VA + 0.824CA + 0.272RS - 1.303Chloride - 0.003FSD + -0.002TSD - 595.005density + 1.085sulphate + 3.617pH$

The only difference among the three regression equations is in the intercept terms. Slopes corresponding to the continuous predictors remain the same. Sign of the slopes indicates in which direction the response will change, given the predictor *increases* by a unit amount.

Any positive coefficient means that a unit increase in the corresponding predictor increases the response by the numerical value of the coefficient, provided all other predictors are held at constant level. Any negative coefficient means that a unit increase in the corresponding predictor decreases the response by the value of the coefficient, provided all other predictors are held at constant level.

If FA (fixed acidity) increases by one unit, other predictors remaining constant, estimated alcohol (%) will increase by 0.51.

If FSD (free sulphuric acid) increases by one unit, other predictors remaining constant, estimated alcohol (%) will decrease by 0.003.

Note also that the estimated slope parameter corresponding to density has not changed sign, but the numerical value is very different. In general, whether the sign of the regression coefficient of a predictor will remain unchanged in both SLR and MLR cannot be determined beforehand. The sign depends on the correlations among the predictors. Sufficiently high correlations among the predictors can result in disturbance in the sign of the regression coefficient.

Note also that the change in estimated response is identical for all continuous predictors at all three levels of the categorical predictor Brand.

5.4 Multi-collinearity

In multiple regression, if one or more pairs of explanatory variables is highly correlated among themselves, then the phenomenon is known as multi-collinearity.
Effects of Multi-collinearity: Multi-collinearity is not desirable. It leads to inflated standard errors of the estimates of the regression coefficients, which in turn affects significance of the regression parameters. Often the signs of the regression coefficients may also change. As a result, the regression model becomes non-reliable or lacks interpretability.

The first thing one should do in multiple linear regression is, to check if multi-collinearity is present in the data.

Detection of Multi-collinearity: There are some ways of detecting (or testing) multi-collinearity such as:

- **Correlation Matrix**: We can start by computing the pairwise correlations among all the independent variables. The independent variables should not be highly (positive or negative) correlated. But this itself is not enough as the correlation matrix only detects high pairwise correlations. It is possible that even when no pairwise correlations are high, several moderately correlated pairs may give rise to multi-collinearity.
- **Variance Inflation factor**: Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among

the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is *"inflated"* by the existence of correlation among the predictor variables in the model.

General Rule of thumb: If VIF is 1 then there is no correlation among the k^{th}predictor and the remaining predictor variables, and hence the variance of $\hat{\beta}_k$ is not inflated at all. Whereas if VIF exceeds 5 or is close to exceeding 5, we say there is moderate VIF and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.

Case Study continued.

We now calculate the VIF of each predictor variable. Before calculating VIF, we need to create the dummy variable for the Brand as it is categorical variable.

```
# VIF
df_new=pd.get_dummies(WineData, drop_first=True)
def vif_cal(input_data, dependent_col):
   x_vars=input_data.drop([dependent_col], axis=1)
   xvar_names=x_vars.columns
   for i in range(0,xvar_names.shape[0]):
      y=x_vars[xvar_names[i]]
      x=x_vars[xvar_names.drop(xvar_names[i])]
      rsq=ols(formula="y~x", data=x_vars).fit().rsquared
      vif=round(1/(1-rsq),2)
      print (xvar_names[i], "VIF = " , vif)
vif_cal(input_data=df_new.iloc[:, 1:14], dependent_col="alcohol")

FA VIF =  5.65
VA VIF =  1.79
CA VIF =  3.07
RS VIF =  1.31
chloride VIF =  1.48
FSD VIF =  1.98
TSD VIF =  2.24
density VIF =  2.91
pH VIF =  2.49
sulphate VIF =  1.4
Brand_Seagram VIF =  1.91
Brand_Sula Vineyards VIF =  1.59
```

Note that Brand is a nominal variable, so the notion of correlation is difficult to define in this case. However, we still output a VIF for Brand which needs to be ignored.

We observe that among all continuous predictors, only FA has a sufficiently high VIF (5.65) indicating it is substantially correlated with the other predictor variables. FA is removed from the model.

```
# MLR with FA removed
```

```
mod2 = ols('alcohol ~ VA+CA+RS+chloride+FSD+TSD+density+sulphate+pH+Brand', data = WineDat
a).fit()
coeff_MLR_wine2 = mod2.params
print(coeff_MLR_wine2)

Intercept              364.757981
Brand[T.Seagram]          -0.416173
Brand[T.Sula Vineyards]     -0.066258
VA                       0.759100
CA                       2.454701
RS                       0.199998
chloride                   -4.397360
FSD                      0.002859
TSD                     -0.006540
density                 -361.063201
sulphate                   0.985574
pH                       1.257011
dtype: float64
```

Note that the coefficients of the different predictor values have changed. We check the VIFs of the new predictors.

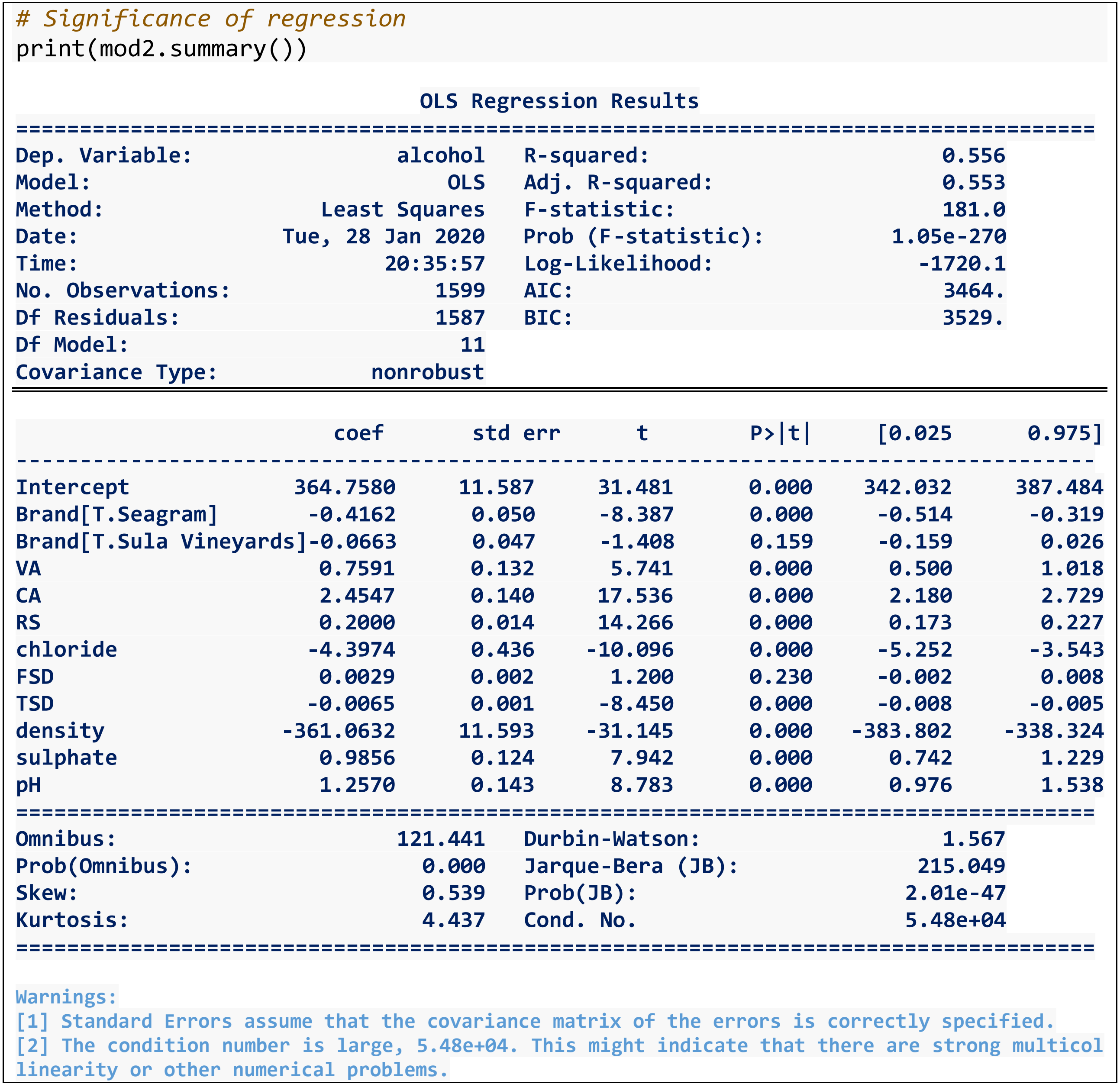
```
# VIF after removing FA
vif_cal(input_data=df_new.iloc[:, 2:14], dependent_col="alcohol")

VA VIF =  1.77
CA VIF =  2.34
RS VIF =  1.23
chloride VIF =  1.32
FSD VIF =  1.96
TSD VIF =  2.04
density VIF =  1.51
pH VIF =  1.54
sulphate VIF =  1.39
Brand_Seagram VIF =  1.86
Brand_Sula Vineyards VIF =  1.58
```

We can see that after removing FA, all the predictors have low VIF (at most 2.34). So the problem of multi-collinearity has been eliminated. In all our subsequent discussions, we will consider the multiple linear regression model with FA removed.

5.4.1 Examining significance of multiple regression model

It is not enough to merely fit a multiple regression model to the data, it is necessary to check whether all regression coefficients are significant or not. Significance here means whether the population regression parameters are significantly different from zero. For the *j*-th slope parameter, the null hypothesis of interest is: H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$. This test is done separately for each regression coefficient, including the parameters corresponding to the dummy variables, if necessary. The test may not always make sense for the intercept parameters.

Case Study continued.



```
# Significance of regression
print(mod2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                alcohol   R-squared:                       0.556
Model:                            OLS   Adj. R-squared:                  0.553
Method:                 Least Squares   F-statistic:                     181.0
Date:                Tue, 28 Jan 2020   Prob (F-statistic):         1.05e-270
Time:                        20:35:57   Log-Likelihood:                -1720.1
No. Observations:                1599   AIC:                             3464.
Df Residuals:                    1587   BIC:                             3529.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================

                            coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept               364.7580     11.587     31.481      0.000     342.032     387.484
Brand[T.Seagram]         -0.4162      0.050     -8.387      0.000      -0.514      -0.319
Brand[T.Sula Vineyards]-0.0663      0.047     -1.408      0.159      -0.159       0.026
VA                        0.7591      0.132      5.741      0.000       0.500       1.018
CA                        2.4547      0.140     17.536      0.000       2.180       2.729
RS                        0.2000      0.014     14.266      0.000       0.173       0.227
chloride                 -4.3974      0.436    -10.096      0.000      -5.252      -3.543
FSD                       0.0029      0.002      1.200      0.230      -0.002       0.008
TSD                      -0.0065      0.001     -8.450      0.000      -0.008      -0.005
density                -361.0632     11.593    -31.145      0.000    -383.802    -338.324
sulphate                  0.9856      0.124      7.942      0.000       0.742       1.229
pH                        1.2570      0.143      8.783      0.000       0.976       1.538
==============================================================================
Omnibus:                      121.441   Durbin-Watson:                   1.567
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              215.049
Skew:                           0.539   Prob(JB):                     2.01e-47
Kurtosis:                       4.437   Cond. No.                     5.48e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.48e+04. This might indicate that there are strong multicol
linearity or other numerical problems.
```

From the above it may be noted that the regression coefficients corresponding to FSD and Brand = Sula Vineyards are not statistically significant at level α=0.05. In other words, the regression coefficients corresponding to these two are not significantly different from 0 in the population.

Hence, FSD may be eliminated from the multiple regression model. However, it is not recommended to eliminate the Brand level Sula Vineyards directly, because it is part of a system of dummy variable. The adjustment required is explained in Section 5.4.3.

5.4.2 R^2 and Adjusted R^2

The coefficient of determination, R^2 is defined for multiple linear regression similarly, as it is defined for SLR.

$$R^2 = 1 - SSE/SST$$

where SST $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and SSE $= \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$.

Recall that in simple linear regression, R^2 is the square of the pairwise correlation coefficient between the single predictor X and the response Y. In MLR no such interpretation of R^2 holds.

A general formulation valid for both simple and multiple regression is that R^2 equals the square of the correlation between the observed response Y and estimated/predicted response $\hat{Y}$.

Hence a high numerical value of R^2 gives an intuitive justification that the model works well since the observed response and the predicted responses are close.

One limitation of the coefficient of determination is that, its value increases as the number of independent variables in the model increases. A good regression model should include only those predictors that are significantly different from 0. However, the numerical value of R^2 is non-decreasing even if non-significant predictors are included in the model. Addition of non-significant predictors adversely affect the predictive quality of the model. Therefore, there is a need of another measure of model adequacy.

Adjusted R^2 measure involves an adjustment based on the number of predictors relative to the sample size. Adjusted R^2 is defined as:

$$\text{Adj } R^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i{}^2/(n-p)}{\sum_{i=1}^{n}(Y_i - Y)^2/(n-1)}$$

Here p is the number of parameters in the regression model including intercept term.

Case Study continues

$R^2$and adjusted R^2 are available from the output of summary(). Values of these two statistics are almost equivalent (55%) albeit adjusted R^2 is slightly smaller than R^2.

5.4.3 Handling the categorical predictor

In the regression equation including all predictors, it was noted that Brand = Sula Vineyards is non-significant. Before any action is taken regarding this, let us understand what significance in case of categorical data means in this situation.

While dealing with a categorical predictor with *m* nominal levels, the set of *m – 1* indicator variables implicitly fixes one level as baseline, when all *m-1* indicator variables take the value 0. In this case study the baseline is Brand = Grover Zampa.

Significance of regression coefficients for categorical variable means that the levels are significantly different from the baseline.

Case Study continues

If the regression coefficient for Brand = Sula Vineyards is non-significant, it means that the effect of Brand = Sula Vineyards statistically indistinguishable from the baseline level Grover Zampa. This is also evident from the box plot (Fig 9).

```
#Brand
pd.DataFrame(WineData['Brand'].value_counts()).transpose()
Seagram          Sula Vineyards  Grover Zampa
633              553                       413

round(WineData.groupby("Brand")["alcohol"].mean(),2)
Grover Zampa     10.88
Seagram          9.91
Sula Vineyards   10.67
Name: alcohol, dtype: float64

# Box plot showing alcohol values at different levels of Brand
a4_dims = (10,5)
fig, ax = plt.subplots(figsize=a4_dims)
a = sns.boxplot(x= "Brand", y = 'alcohol' , data = WineData)
plt.xlabel('Independent variable: Brand')
plt.ylabel('% of alcohol')
plt.show()
```

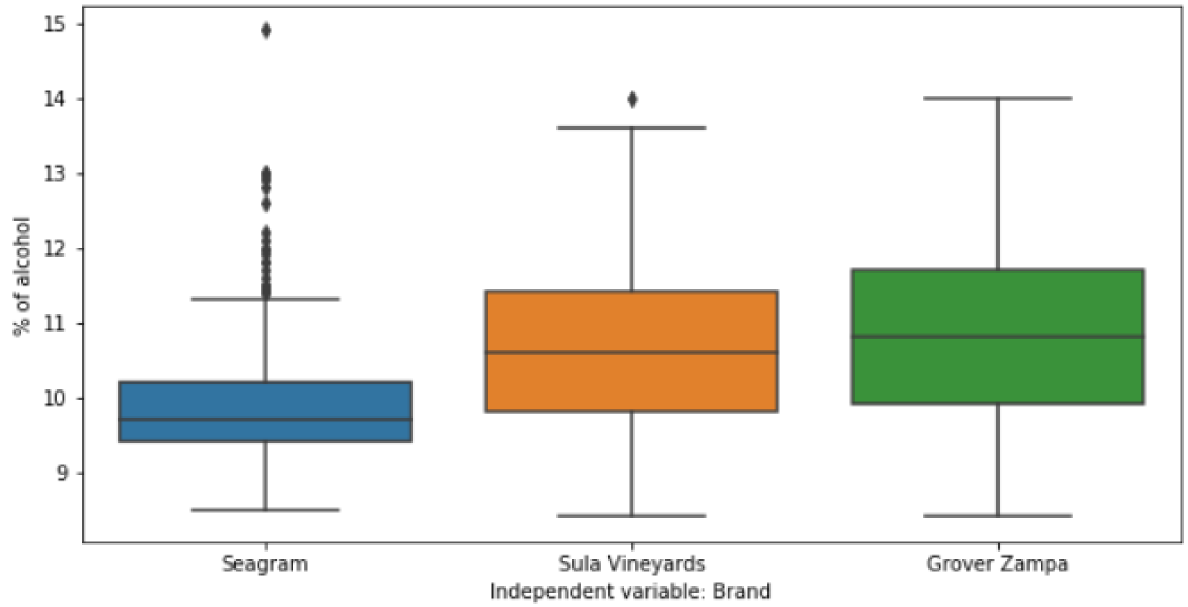


Fig. 9: Box plot for alcohol (%) at different levels of Brand

If a continuous predictor is not significant, i.e. if the p-value in the regression table is less than a pre-fixed level α, we simply eliminate the variable from the regression equation. However, we cannot do that in case of a categorical predictor.

Instead, the non-significant level is merged into the baseline and a new baseline is created. Consequently, the number of levels is also reduced, along with the degrees of freedom that the categorical variable carries.

```
# Defining binary predictor new_Brand
x = np.array([0])
new_brand = np.repeat(x,len(WineData), axis = 0)
for i in range(len(WineData)):
   if(WineData.iloc[i,1] == "Seagram"):
       new_brand[i]=1
WineData['new_brand'] = new_brand
```

Brands Sula Vineyards and Grover Zampa are merged and we consider only the indicator variable for Brand = Seagram as our predictor. A three category variable is reduced to a binary variable. Let us define a new variable (new_Brand) that takes the value 1 if the brand is Seagram otherwise takes the value 0. Note that this variable takes the same value when the brand is either Grover Zampa or Sula Vineyards and thus combines these two brands together. So now instead of Brand as a predictor we use new_Brand.

Now we perform MLR of alcohol (%) on the modified set of predictors. The results are shown below.

```
# Multiple linear regression on modified set of predictors

MLR_new = ols('alcohol ~ VA+CA+RS+chloride+TSD+density+sulphate+pH+new_brand', data = WineD
ata).fit()
```

```
print(MLR_new.summary())

                            OLS Regression Results                            
==============================================================================
Dep. Variable:                alcohol   R-squared:                       0.556
Model:                            OLS   Adj. R-squared:                  0.553
Method:                 Least Squares   F-statistic:                     220.7
Date:                Tue, 28 Jan 2020   Prob (F-statistic):          2.31e-272
Time:                        20:52:58   Log-Likelihood:                -1721.7
No. Observations:                1599   AIC:                             3463.
Df Residuals:                    1589   BIC:                             3517.
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    365.5364     11.582     31.560      0.000     342.818     388.255
VA             0.7443      0.131      5.698      0.000       0.488       1.000
CA             2.4537      0.138     17.770      0.000       2.183       2.725
RS             0.2025      0.014     14.534      0.000       0.175       0.230
chloride      -4.4167      0.435    -10.158      0.000      -5.270      -3.564
TSD           -0.0060      0.001    -10.320      0.000      -0.007      -0.005
density     -361.9534     11.586    -31.240      0.000    -384.679    -339.228
sulphate       1.0068      0.124      8.152      0.000       0.765       1.249
pH             1.2808      0.142      8.994      0.000       1.001       1.560
new_brand     -0.3784      0.040     -9.360      0.000      -0.458      -0.299
==============================================================================
Omnibus:                      118.826   Durbin-Watson:                   1.567
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              208.529
Skew:                           0.533   Prob(JB):                     5.23e-46
Kurtosis:                       4.412   Cond. No.                     5.24e+04
==============================================================================
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.24e+04. This might indicate that there are strong mult
icollinearity or other numerical problems.
```

Notice that now, all the predictors are statistically significant with very low p-values. The value of R^2 has not decreased much and adjusted R^2 is almost same as R^2 (around 55%). Thus we select this as our final regression model.

We now present the regression coefficients for this final model which will help us to write the regression equation.

```
# Coefficients
coeff_MLR = MLR_new.params
print(coeff_MLR )

Intercept    365.536428
VA             0.744273
CA             2.453709
RS             0.202516
chloride      -4.416669
TSD           -0.005956
density     -361.953393
sulphate       1.006848
pH             1.280760
```

```
new_brand    -0.378445
dtype: float64
```

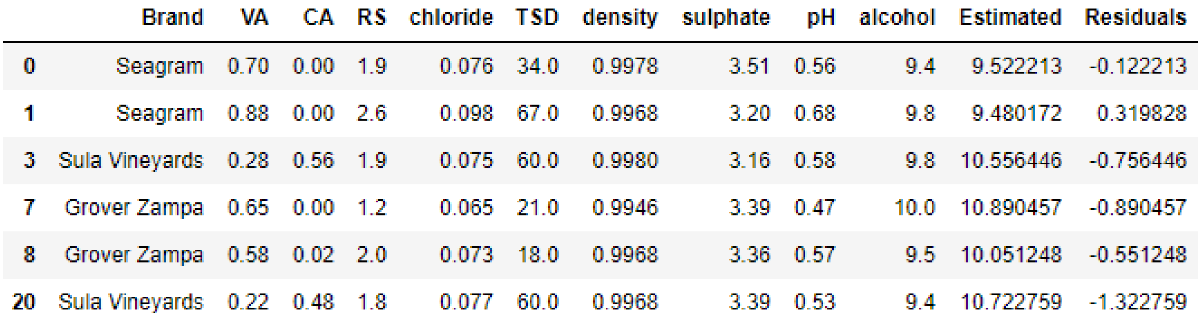
Thus, the final regression equation becomes

$$\widehat{\text{alcohol}}(\%) = 365.54 + 0.74\text{VA} + 2.45\text{CA} + 0.2\text{RS} - 4.42\text{chloride} - 0.006\text{TSD} - 361.95\text{density} + 1.007\text{sulphate} + 1.280\text{pH} - 0.378\ \text{I}(\text{Brand} = \text{Seagram})$$

The fitted values of the response and the residuals can be extracted directly from the model.

```
# Extraction of fitted response
model_fitted_y = pd.DataFrame(mod.fittedvalues,columns= ['Estimated'])
# Extraction of residuals
model_residuals = pd.DataFrame(mod.resid , columns= ['Residual'])
d1 = pd.concat([WineData, model_fitted_y,model_residuals], axis=1, ignore_index=True)
d1.columns  = ['ID', 'Brand', 'FA', 'VA', 'CA','RS','chloride','FSD','TSD','density','sulphate','pH','alcohol','new_brand', 'Estimated','Residuals']
d1.loc[[0,1,3,7,8,20], ['Brand',  'VA',  'CA','RS','chloride','TSD','density','sulphate','pH','alcohol','Estimated','Residuals']]
```

Below we present some particular values of the predictors, the observed response, predicted response and the residuals.



	Brand	VA	CA	RS	chloride	TSD	density	sulphate	pH	alcohol	Estimated	Residuals
0	Seagram	0.70	0.00	1.9	0.076	34.0	0.9978	3.51	0.56	9.4	9.522213	-0.122213
1	Seagram	0.88	0.00	2.6	0.098	67.0	0.9968	3.20	0.68	9.8	9.480172	0.319828
3	Sula Vineyards	0.28	0.56	1.9	0.075	60.0	0.9980	3.16	0.58	9.8	10.556446	-0.756446
7	Grover Zampa	0.65	0.00	1.2	0.065	21.0	0.9946	3.39	0.47	10.0	10.890457	-0.890457
8	Grover Zampa	0.58	0.02	2.0	0.073	18.0	0.9968	3.36	0.57	9.5	10.051248	-0.551248
20	Sula Vineyards	0.22	0.48	1.8	0.077	60.0	0.9968	3.39	0.53	9.4	10.722759	-1.322759

Table 2: Observed response, estimated response (based on final MLR model) and residuals

5.4.4 Residual Analysis

Residual plots are shown below. They exhibit no specific pattern and are more or less normally distributed. So we can assume they satisfy the properties of linear regression.

```
# Residual plot
regression_plots(MLR_new,WineData)
```

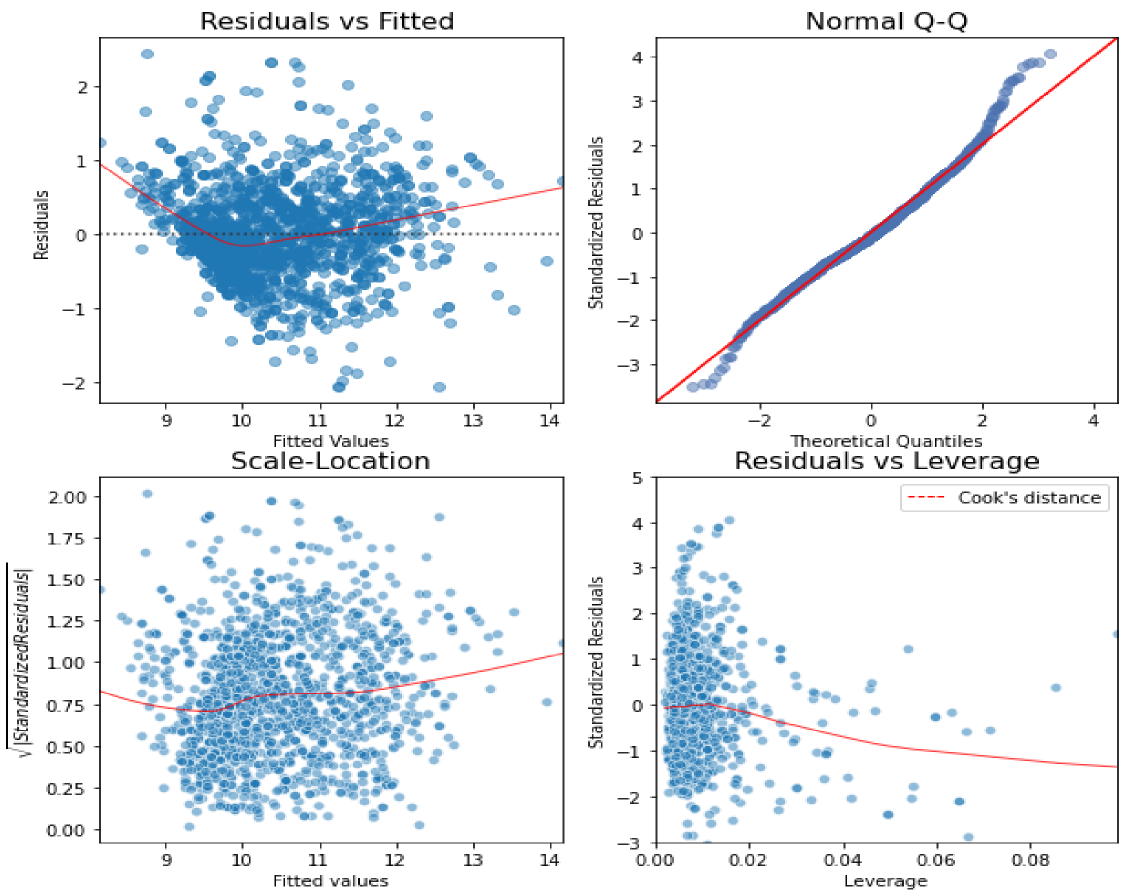


Fig. 10: Residual Plot of alcohol vs all variables

We also check the VIFs of the final model which are all under 3 and thus there is no issue of multicollinearity.

```
# VIF
vif_cal(input_data=WineData.iloc[:, 3:14], dependent_col="alcohol")
VA VIF =  1.76
CA VIF =  2.32
RS VIF =  1.23
chloride VIF =  1.32
FSD VIF =  1.94
TSD VIF =  2.04
density VIF =  1.51
pH VIF =  1.54
sulphate VIF =  1.38
new_brand VIF =  1.24
```

5.5 Transformation of the response

After fitting the linear regression models, a residual analysis was performed to check if the regression assumptions remained valid. Recall that one of the assumptions was normality. Another important assumption was that the variances of the response were constant.

If any or both of these two assumptions are violated, which can be detected from the residual plot, a transformation of the response may be necessary. Many transformations of the response are possible, but the most popular choice of transformations are given by the Box-Cox family of transformations.

Let λ be a constant, $\lambda \neq 0$. The response Y is transformed to a new variable Z (say) where Z = $(y^{\lambda} - 1) / \lambda$. If $\lambda = 0$ then the response is changed to $\log(y)$. The transformed response Z fulfils the regression assumptions. However, it is clear from the functional form of the transformation, that Z can be simplified as Y^{λ}. This simplification works because $1 / \lambda$ being a constant, gets absorbed into the intercept term, and the regression slope parameters are multiplied (scaled) by the constant λ. If any predictor was significant in predicting $(y^{\lambda} - 1) /$ λ then it is expected to be significant in predicting y^λ. Therefore, the inference remains unchanged whether we use y^{λ} or $(y^{\lambda} - 1) / \lambda$.

That value of λ is chosen such that the log-likelihood curve reaches its maximum. This will be explained in further detail through the case study.

Case Study continues

Let us first determine the value λ to see whether any transformation is necessary.

```
lmbdas = np.linspace(-3, 2)
llf = np.zeros(lmbdas.shape, dtype=float)
for ii, lmbda in enumerate(lmbdas):
   llf[ii] = stats.boxcox_llf(lmbda, MLR_new.fittedvalues)


lmbda_optimal = llf.max()
fig = plt.figure()
ax = fig.add_subplot(111)
ax.plot(lmbdas, llf, 'b.-')
ax.axhline(lmbda_optimal, color='r')
ax.set_xlabel('lambda parameter')
ax.set_ylabel('log-likelihood')
```

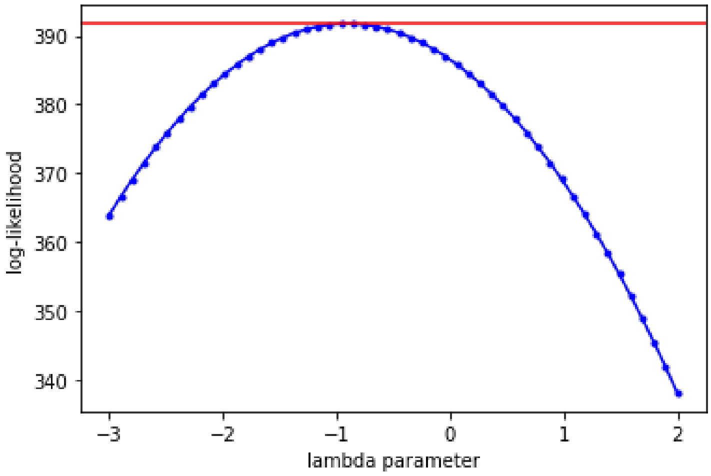


Fig. 11: Log likelihood vs Lambda Parameter

That value(s) of λ is (are) chosen for which the graph plotted above reaches its maximum. For this data the maximum is attained between -1 and 0 (closer to -1 than 0). For clarity of interpretation, any fractional value of λ is ignored.

Taking $\lambda = -1$, the response is transformed to get new_resp1 and is modelled by MLR on the predictors from the training dataset.

```
#lambda=-1

new_resp1=1/WineData['alcohol']

WineData['new_resp'] = new_resp1

MLR_wine1 = ols(formula="new_resp ~
VA+CA+RS+chloride+TSD+density+sulphate+pH+new_brand",data=WineData).fit()


regression_plots(MLR_wine1, WineData)
```

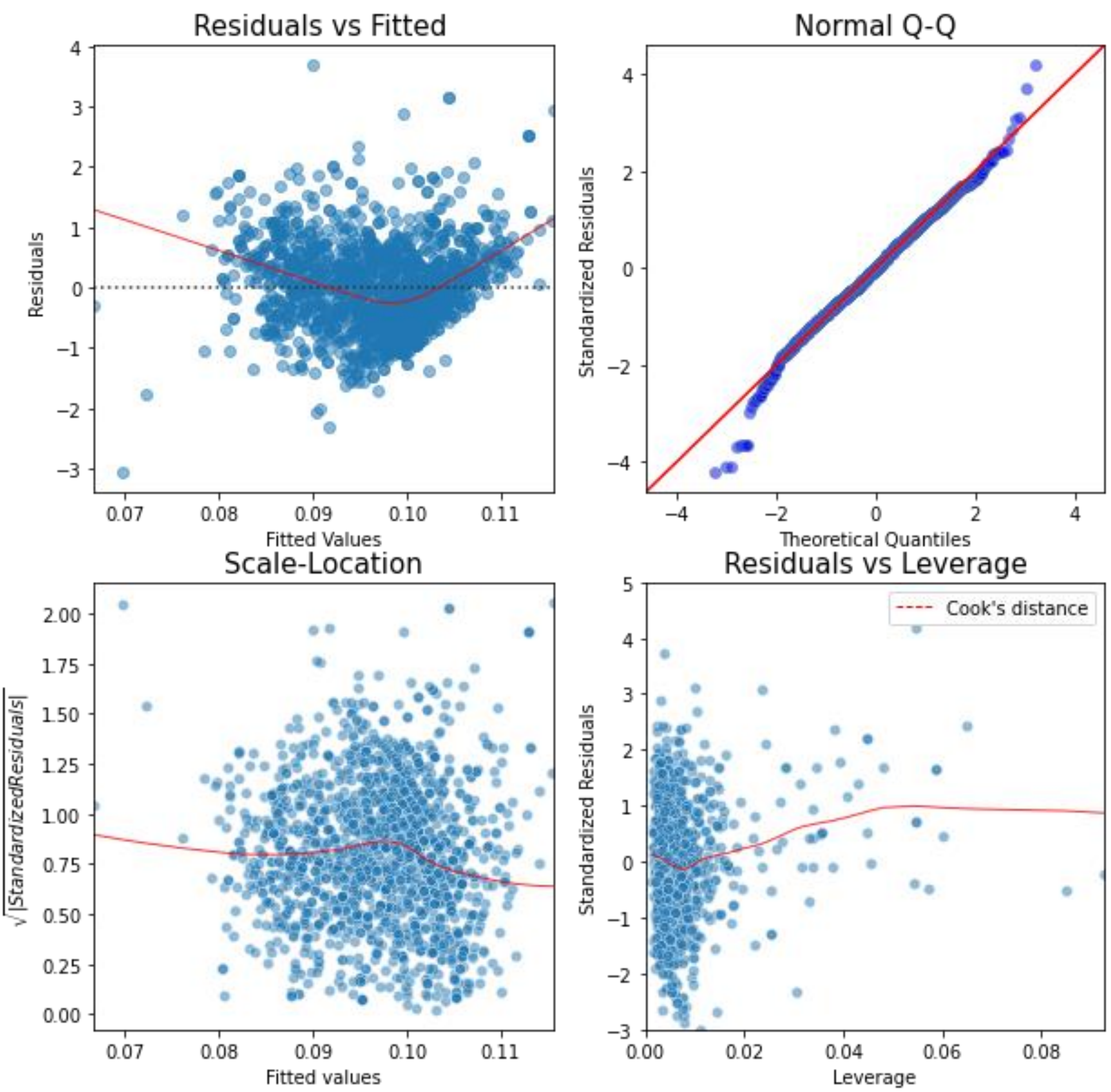


Fig. 12: Residual Plot of alcohol vs all variables

It is clear that the regression assumptions are not violated. Next, $\lambda = 0$ transformation is considered.

```
#lambda=0

new_resp2 = np.log(WineData['alcohol'])

WineData['new_resp'] = new_resp2

MLR_wine2 = ols(formula="new_resp ~
VA+CA+RS+chloride+TSD+density+sulphate+pH+new_brand",data=WineData).fit()
```

```
regression_plots(MLR_wine2, WineData)
```

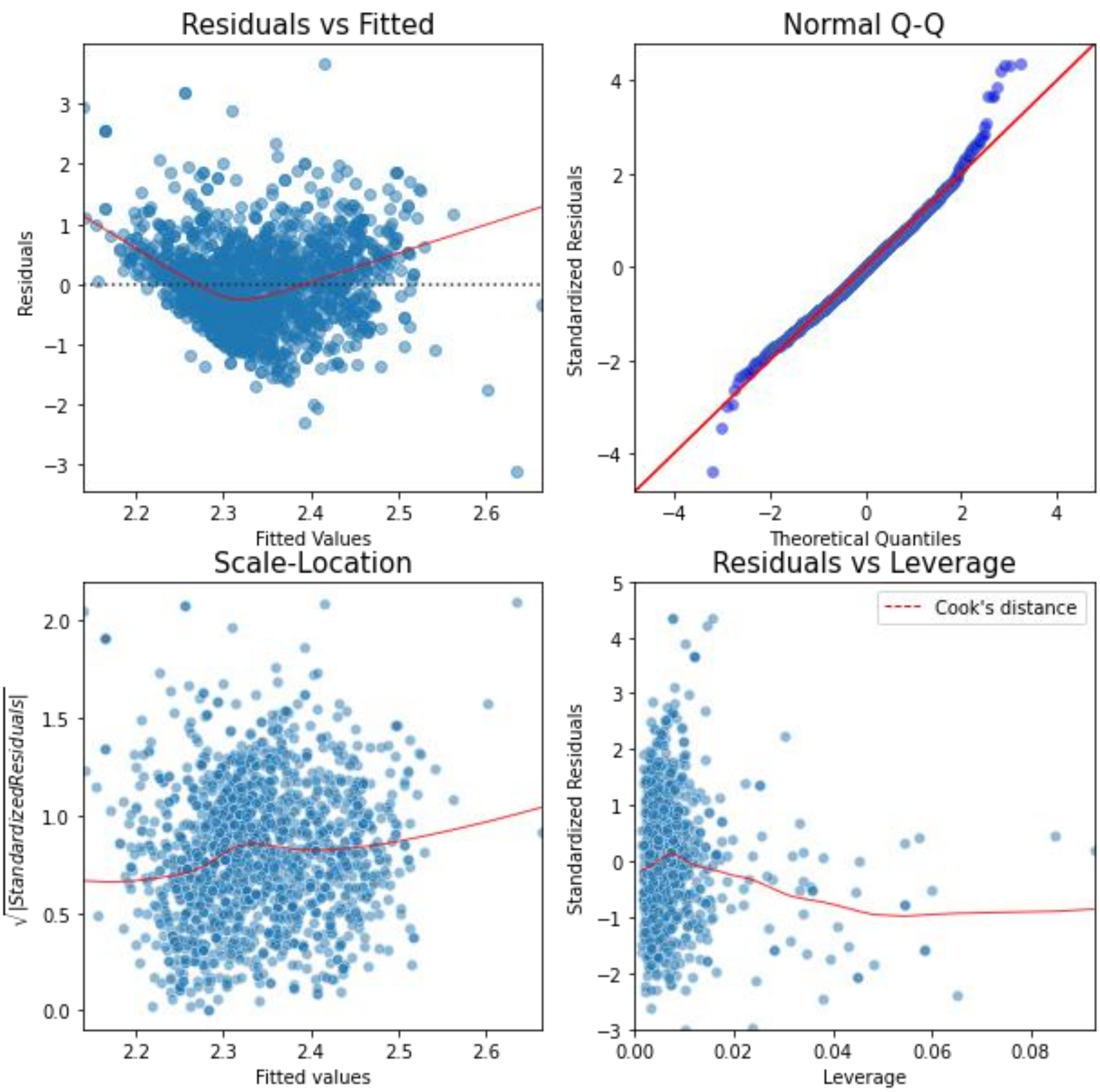


Fig. 13: Residual Plot of alcohol vs all variables

The regression assumptions are not violated in this case also. So both the transformations with λ=−1 and λ=0 may be considered. Whereas λ=−1 involves taking the reciprocal of the response (alcohol), λ=0 implies taking the logarithm of the response (alcohol). As the maximum value of log-likelihood was obtained near λ=−1, λ=−1 transformation is chosen.

```
#lambda=-1

new_resp = 1/WineData['alcohol']

WineData['new_resp'] = new_resp

MLR_wine = ols(formula="new_resp ~
VA+CA+RS+chloride+TSD+density+sulphate+pH+new_brand",data=WineData).fit()

print(MLR_wine.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               new_resp   R-squared:                       0.547
Model:                            OLS   Adj. R-squared:                  0.544
Method:                 Least Squares   F-statistic:                     213.2
Date:                Thu, 17 Jun 2021   Prob (F-statistic):          9.26e-266
Time:                        01:38:10   Log-Likelihood:                 5848.0
No. Observations:                1599   AIC:                        -1.168e+04
Df Residuals:                    1589   BIC:                        -1.162e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -2.8919      0.102    -28.402      0.000      -3.092      -2.692
VA            -0.0066      0.001     -5.735      0.000      -0.009      -0.004
CA            -0.0210      0.001    -17.313      0.000      -0.023      -0.019
RS            -0.0017      0.000    -13.867      0.000      -0.002      -0.001
chloride       0.0407      0.004     10.646      0.000       0.033       0.048
TSD         5.53e-05   5.07e-06     10.899      0.000    4.53e-05    6.52e-05
density        3.0474      0.102     29.920      0.000       2.848       3.247
sulphate      -0.0089      0.001     -8.223      0.000      -0.011      -0.007
pH            -0.0111      0.001     -8.880      0.000      -0.014      -0.009
new_brand      0.0034      0.000      9.588      0.000       0.003       0.004
==============================================================================
Omnibus:                       35.464   Durbin-Watson:                   1.554
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               54.819
Skew:                          -0.209   Prob(JB):                     1.25e-12
Kurtosis:                       3.805   Cond. No.                     5.24e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.24e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

We get the following regression equation:

$$1/Alcohol = -2.8919 - 0.0066 \times VA - 0.0210 \times CA - 0.0017 \times RS + 0.0407 \times chloride + 0.00001 \times TSD + 3.0474 \times density - 0.0089 \times sulphate - 0.0111 \times pH + 0.0034 \times I(Brand = new\_brand)$$

6. Further Discussions and Considerations

6.1 Regression ANOVA

The problem of regression is also a technique of reducing variation in data. Total variation in the response is being explained by the regression on one or more predictors. The ANOVA table describes the partition of the variance.

Corresponding to each partition of the total variance, there is an associated ***degrees of freedom*** (df). The total degrees of freedom in a sample of size n is always $n - 1$. df for the partitioned sums of squares depend on the number of parameters to be estimated. If there are p parameters to be estimated in the regression equation, then the df corresponding to SSR will be $p - 1$. df corresponding to SSE is $(n - 1) - (p - 1) = n - p$.

We present the ANOVA table corresponding to the final model in our case study.

```
# ANOVA table
aov_tbl = sm.stats.anova_lm(MLR_new)
print(aov_tbl)

                df      sum_sq     mean_sq            F        PR(>F)
VA             1.0   74.260974   74.260974   146.311486   2.773338e-32
CA             1.0    0.009037    0.009037     0.017805   8.938658e-01
RS             1.0    3.435694    3.435694     6.769121   9.361018e-03
chloride       1.0   86.123751   86.123751   169.683930   6.277644e-37
TSD            1.0   72.561124   72.561124   142.962383   1.300526e-31
density        1.0  620.963182  620.963182  1223.442687  3.003708e-199
sulphate       1.0   60.471165   60.471165   119.142337   8.482902e-27
pH             1.0   45.968786   45.968786    90.569259   6.367156e-21
new_brand      1.0   44.467580   44.467580    87.611532   2.622912e-20
Residual    1589.0  806.503244    0.507554          NaN            NaN
```

We can see that each continuous predictor has 1 degree of freedom, and since our modified categorical predictor is new_Brand with 2 levels, it enjoys (2 - 1 =)1 degree of freedom. The residual sum of squares has 1589 degrees of freedom, and the total sum of squares has 1598 degrees of freedom.

6.2 Leverage points

In the context of regression, an observation is called a **leverage point** if it has an unusually high or low value for the predictor(s) and/or for the response. All leverage points are outliers on at least one of the variables. However, all outliers may not be leverage points. Presence of any leverage point can substantially shift the OLS regression line towards itself. The following figure illustrates this.

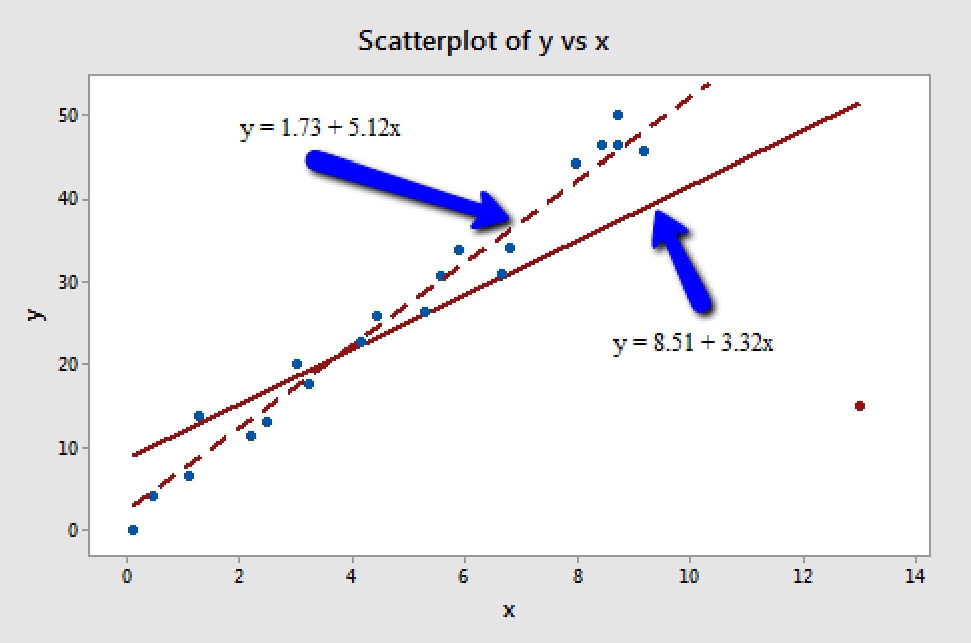


Fig. 14: Action of leverage point in shifting the regression line (ref: https://newonlinecourses.science.psu.edu/stat501/node/337/ **)**

The red point shown to the right of the figure is a leverage point. Note that this regressor value is unusually large compared to the remaining values but the corresponding response value is not an outlier. The red line is OLS line obtained by regression including the leverage value and the dashed line is the OLS line obtained by regression excluding the leverage value. We can notice how the leverage point pulls the regression line towards itself.

The flowchart below explains the main steps of the regression.

Regression model building will be discussed in a subsequent monograph.

6.3 Information Criteria: AIC and BIC

Once several candidate models are found, it is necessary to be able to compare among them. There are several options for comparison. Two of those, namely R^2 and adjusted R^2, have already been introduced and their merits and demerits have been discussed.

In this section two more criteria are introduced both of which are based on information lost in fitting a model. A model is a simplification of the process from which the observed data is generated. The closer the model is to the real process; the more information it contains. Nevertheless, no model will ever be able to emulate the real process and hence, some amount of information will always be lost. The errors or residuals of the model fit provide quantification of the information lost.

Both the information criteria, AIC and BIC, are popular for comparison of models. Both criteria are based on the error sum of squares and both penalize models on the number of predictors included. In a way, they try to strike a balance between bias and variance.

Consider the multiple linear regression model with p parameters (including intercept term) on n data points and let the residual sum of squares be denoted by SSE.

The ***Akaike Information Criteria (AIC)*** is defined by

$$AIC = n\log(SSE) - n\log(n) + 2p$$

However, the AIC tends to overfit models. This criticism of AIC has led to the development of BIC.

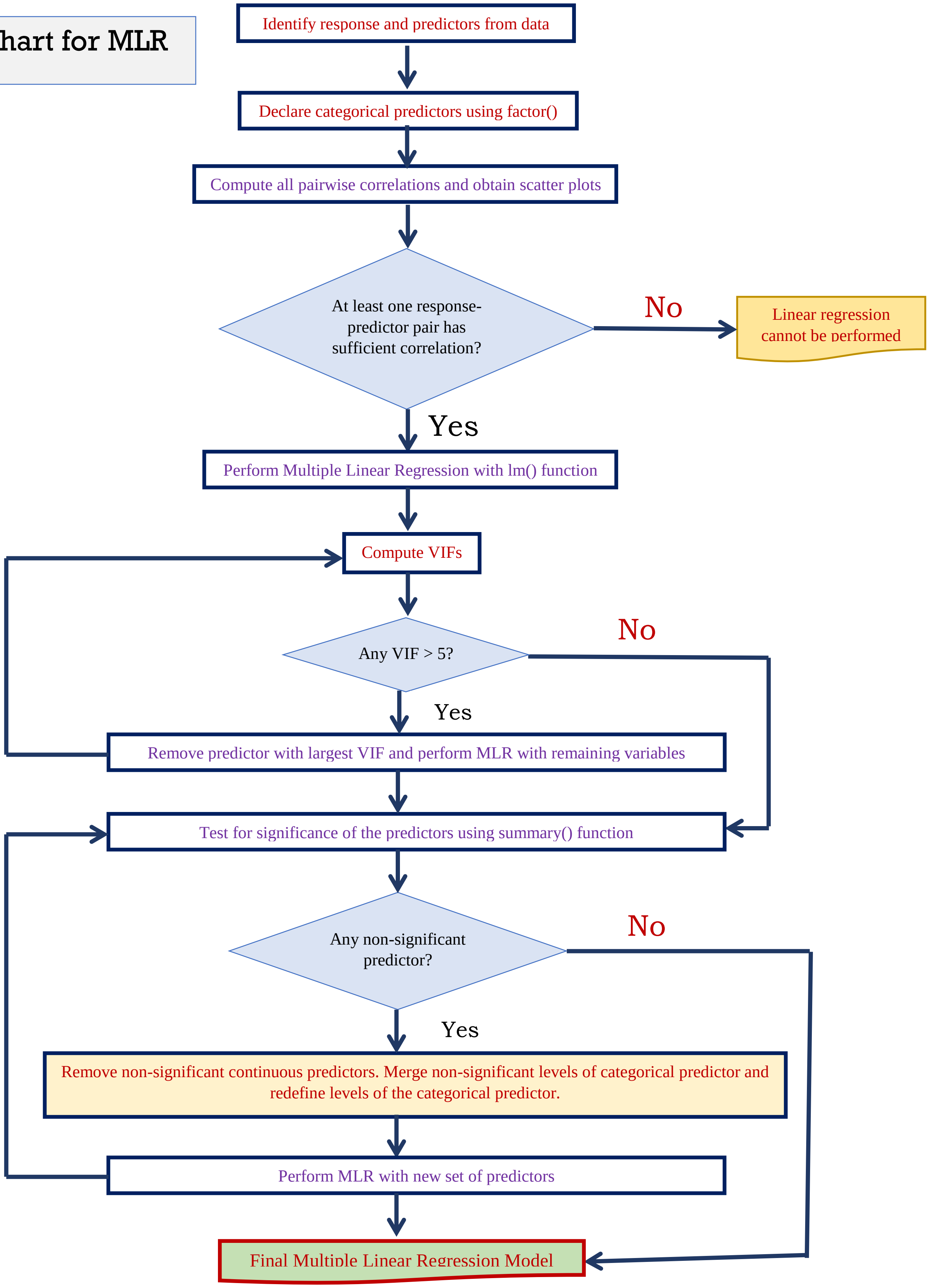
The ***Bayesian Information Criteria (BIC)*** is defined by

$$BIC = n\log(SSE) – n\log(n) + p\log(n)$$

Since $\log(n)$ is usually much larger than 2, it follows that BIC imposes greater penalty if the number of parameters, p, is too large. Thus BIC maintains a greater balance in the number of parameters used to fit the model. In general, BIC is preferred to AIC in model building exercises.

Naturally, the smaller the value of AIC or BIC, the better is the model, since the model with minimum value of these criteria identifies the smallest quantity of information lost.

Flowchart for MLR


Identify response and predictors from data
Declare categorical predictors using factor()
Compute all pairwise correlations and obtain scatter plots
At least one response-predictor pair has sufficient correlation?
No
Linear regression cannot be performed
Yes
Perform Multiple Linear Regression with lm() function
Compute VIFs
Any VIF > 5?
No
Yes
Remove predictor with largest VIF and perform MLR with remaining variables
Test for significance of the predictors using summary() function
Any non-significant predictor?
No
Yes
Remove non-significant continuous predictors. Merge non-significant levels of categorical predictor and redefine levels of the categorical predictor.
Perform MLR with new set of predictors
Final Multiple Linear Regression Model

References

Draper, N. R., Smith, H. (1998). Applied Regression Analysis. Wiley Series in Probability and Statistics.

Neter, J., Wasserman, W., Kutner, M. H. (1983). Applied Linear Regression Models. Richard D. Irwin, Inc.

Seber, G. A. F., Lee, A. J. (2003). Linear Regression Analysis. Wiley Series in Probability and Statistics.

https://newonlinecourses.science.psu.edu/stat501/node/2/

https://online-learning.harvard.edu/course/data-science-linear-regression