**DATA VISUALIZATION**

**MAJOR PROJECT**

**By: Glenn Solomon**

**Student Number: 040908930**

**Contents**

[**1. Introduction: Discovery/Description** 4](#_Toc503386308)

[**1.1 Problem/Challenges:** 4](#_Toc503386309)

[**1.2 Motivation:** 4](#_Toc503386310)

[**1.3 Expected Outcome:** 4](#_Toc503386311)

[**1.4 Dataset Used:** 5](#_Toc503386312)

[**2. Bonus Material Content:** 5](#_Toc503386313)

[**3. Data Preparation:** 5](#_Toc503386314)

[**4. Model:** 6](#_Toc503386315)

[**4.1 Software Used:** 6](#_Toc503386316)

[**4.2 Regression Model:** 6](#_Toc503386317)

[**4.2.1 Mean:** 6](#_Toc503386318)

[**4.2.2 Standard Deviation:** 6](#_Toc503386319)

[**4.2.3 Minimum and Maximum:** 7](#_Toc503386320)

[**4.2.4 Dependent and Independent Variables:** 7](#_Toc503386321)

[**4.2.5 Variance:** 8](#_Toc503386322)

[**4.2.6 F Value:** 8](#_Toc503386323)

[**4.2.7 DF:** 8](#_Toc503386324)

[**4.2.8 Root MSE:** 9](#_Toc503386325)

[**4.2.9 Dependent Mean:** 9](#_Toc503386326)

[**4.2.10 Coefficient of Variance:** 9](#_Toc503386327)

[**4.2.11 R-Squared:** 10](#_Toc503386328)

[**4.2.12 Adjusted R-Squared:** 10](#_Toc503386329)

[**4.2.13 Intercept:** 11](#_Toc503386330)

[**4.2.14 Parameter estimate:** 11](#_Toc503386331)

[**4.2.15 Residuals are measured as follows:** 12](#_Toc503386332)

[**4.2.16 Normal Probability (Quantile QQ) plots of Residual:** 13](#_Toc503386333)

[**4.2.17 Cook’s Distance:** 13](#_Toc503386334)

[**4.3 Process Followed:** 15](#_Toc503386335)

[**5. Results:** 16](#_Toc503386336)

[**5.1 Description of results:** 16](#_Toc503386337)

[**5.2 Comparison with 10% Sample:** 17](#_Toc503386338)

[**5.2.1 Comparison:** 19](#_Toc503386339)

[**5.3 Comparison with 10% Sample:** 20](#_Toc503386340)

[**5.3.1 Comparison:** 22](#_Toc503386341)

[**6. Next Steps:** 23](#_Toc503386342)

[**7. Appendices:** 23](#_Toc503386343)

[**7.1 Figure Table:** 23](#_Toc503386344)

[**7.2 Data-set:** 24](#_Toc503386345)

[**7.3 References:** 25](#_Toc503386346)

# **1. Introduction: Discovery/Description**

## **1.1 Problem/Challenges:**

Plastics are inexpensive, lightweight and durable materials, which can readily be moulded into a variety of products that find use in a wide range of applications. As a consequence, the production of plastics has increased markedly over the last 60 years. However, current levels of their usage and disposal generate several environmental problems. Around 4 per cent of world oil and gas production, a non-renewable resource, is used as feedstock for plastics and a further 3–4% is expended to provide energy for their manufacture. A major portion of plastic produced each year is used to make disposable items of packaging or other short-lived products that are discarded within a year of manufacture. These two observations alone indicate that our current use of plastics is not sustainable. In addition, because of the durability of the polymers involved, substantial quantities of discarded end-of-life plastics are accumulating as debris in landfills and in natural habitats worldwide.

## **1.2 Motivation:**

Recycling is one of the most important actions currently available to reduce these impacts and represents one of the most dynamic areas in the plastics industry today. Recycling provides opportunities to reduce oil usage, carbon dioxide emissions and the quantities of waste requiring disposal.

So, the dataset that is being used takes into consideration the strength of the plastics at different temperatures and pressures to examine which best suits for the purpose of recycling of the plastic polymers.

## **1.3 Expected Outcome:**

By doing the Multiple Regression, an attempt is made to co-relate the strength of the plastic with respect to the temperature and pressure and then plot the graph and analyze at what temperature and pressure the plastic would be best to be recycled.

In general sense, regression analysis is also used to understand which among the independent variables (temperature and pressure) are related to the dependent (strength) variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

In other words, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) and one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable). More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

## **1.4 Dataset Used:**

As mentioned above, the dataset deals all about plastic. It has 3 variables – strength of the plastic (force per unit area), temperature in degree Celsius, and pressure in Pascal. The dataset takes 200 different cases into consideration for corresponding temperature and pressure.

The source of the data is as follows:

<http://www.itl.nist.gov/div898/education/reg/plastic1.dat>

OR

<http://www.itl.nist.gov/div898/education/datasets.htm#regression>

# **2. Bonus Material Content:**

SAS is the environment that is used to conduct the Multiple Regression Analysis. Unlike excel, the code was written to generate the values for the co-relation between the dependent variables and the independent variables.

# **3. Data Preparation:**

The data was taken from the source which is mentioned above. The data deals with the strength of the plastic with respect to the corresponding temperature and pressures. The dataset was very exhaustive containing about more than 1600 field and because of it the result was not clean. So, the data was cleaned and unnecessary data was removed from the source dataset. Only about 200 fields of data were kept for regression analysis. The column temperature is in degree Celsius and pressure is in Pascal. Besides no other issues arose related to data. Moreover, various other datasets were taken into consideration for regression analysis before finding the plastic dataset, but there was no good co-relation found between the variable. One such dataset found was about the medals won by countries in the summer and winter Olympics in the year 1992 and 1994 respectively, with respect to population of those countries and latitude. But no specific co-relation was found between the medals won in summer and winter Olympics with respect to either population or latitude.

# **4. Model:**

## **4.1 Software Used:**

SAS is the software used for performing regression analysis.

Below are some of the benefits of SAS with regards to perform analysis work:

1. SAS has masterful data extraction and analysis capability. It can read virtually any data source. For example, it can easily join your native Excel table with a SAS dataset.
2. SAS data set can be as large as the operating system will allow. While the new Excel does perform better with [increased limits](http://msdn.microsoft.com/en-us/library/ff700514.aspx) (starting in Excel 2007, the maximum number of rows per worksheet increased from 65,536 to over 1 million) these are limits nevertheless on large datasets. SAS is perfect for big data or high-performance analytics (while Excel wasn’t really designed for this type of extreme analysis).
3. SAS automatically documents your work creating an audit trail without your intervention.

## **4.2 Regression Model:**

Components Calculated for Regression Analysis:

### **4.2.1 Mean:**

In [probability](https://en.wikipedia.org/wiki/Probability) and [statistics](https://en.wikipedia.org/wiki/Statistics), population **mean** and [expected value](https://en.wikipedia.org/wiki/Expected_value) are used synonymously to refer to one measure of the [central tendency](https://en.wikipedia.org/wiki/Central_tendency) either of a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) or of the [random variable](https://en.wikipedia.org/wiki/Random_variable) characterized by that distribution.

In other words, the mean is the average of the numbers: a calculated "central" value of a set of numbers.

In [statistics](https://en.wikipedia.org/wiki/Statistics), **regression toward** **the mean** is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement—and if it is extreme on its second measurement, it will tend to have been closer to the average on its first.

### **4.2.2 Standard Deviation:**

The **standard deviation** is a measure that is used to quantify the amount of variation or [dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion) of a set of data values. A low standard deviation indicates that the data points tend to be close to the [mean](https://en.wikipedia.org/wiki/Mean) (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

The standard deviation of a [random variable](https://en.wikipedia.org/wiki/Random_variable), [statistical population](https://en.wikipedia.org/wiki/Statistical_population), [data set](https://en.wikipedia.org/wiki/Data_set), or [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) is the [square root](https://en.wikipedia.org/wiki/Square_root) of its [variance](https://en.wikipedia.org/wiki/Variance).

### **4.2.3 Minimum and Maximum:**

In [statistics](https://en.wikipedia.org/wiki/Statistics), the **sample maximum** and **sample minimum,** also called the **largest observation** and **smallest observation,** are the values of the greatest and least elements of a [sample](https://en.wikipedia.org/wiki/Sample_(statistics)).

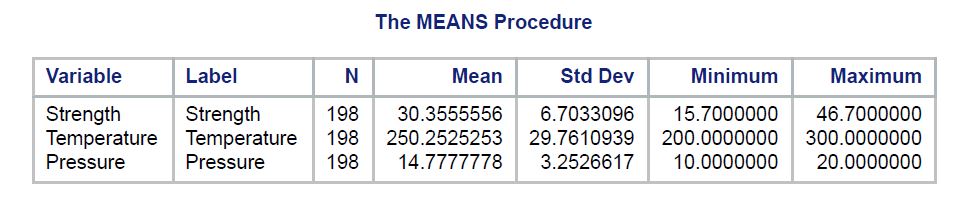


Figure 1: Means Procedure

### **4.2.4 Dependent and Independent Variables:**

The values of **dependent variables** depend on the values of **independent variables**.

The dependent variables represent the output or outcome whose variation is being studied.

The independent variables represent inputs or causes, i.e., potential reasons for variation or, in the experimental setting, the variable controlled by the experimenter.

Models and experiments test or determine the effects that the independent variables have on the dependent variables.

In our case, dependent variable is ***strength*** and the independent variables are ***temperature*** and ***pressure.***

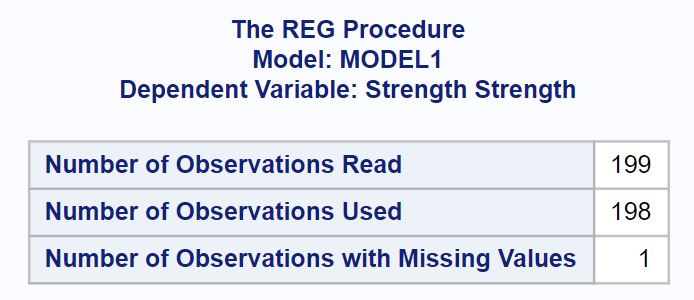


Figure 2: Model1 Dependent Variable Strength

The above figure displays about the number of observations read which is 199, number of observations used i.e. 198 because one row is blank (null) and the number of observations with missing values i.e. the null value is 1.

### **4.2.5 Variance:**

Variance is the [expectation](https://en.wikipedia.org/wiki/Expected_value) of the squared [deviation](https://en.wikipedia.org/wiki/Deviation_(statistics)) of a [random variable](https://en.wikipedia.org/wiki/Random_variable) from its [mean](https://en.wikipedia.org/wiki/Mean). Informally, it measures how far a set of (random) numbers are spread out from their average value.

 A value of zero means that there is no variability; All the numbers in the data set are the same.

The variance for a population is calculated by:

1. Finding the mean (the average).
2. Subtracting the mean from each number in the data set and then squaring the result. The results are squared to make the negatives positive. Otherwise negative numbers would cancel out the positives in the next step. It’s the distance from the mean that’s important, not positive or negative numbers.
3. Averaging the squared differences.

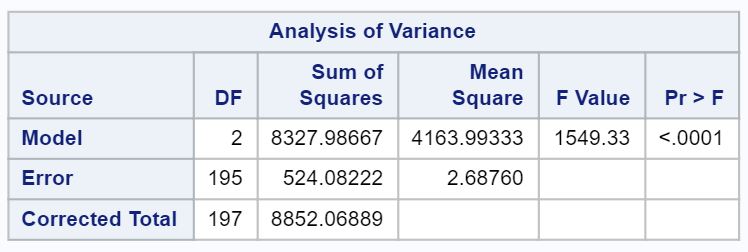


Figure 3: Analysis of Variance

### **4.2.6 F Value:**

The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero).

### **4.2.7 DF:**

The table entries are the critical values (percentiles) for the distribution. The column headed DF gives the degrees of freedom for the values in that row.

In linear regression, the [degrees of freedom](https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)) of the residuals is:

df=n−k∗

Where k∗ is the numbers of parameters we're estimating INCLUDING an intercept. (The residual vector will exist in an n−k∗ dimensional [linear space](https://en.wikipedia.org/wiki/Vector_space).)



Figure 4: Analysis of Variance Contd.

### **4.2.8 Root MSE:**

Root Mean Square Error (RMSE) is the standard deviation of the [residuals](http://www.statisticshowto.com/residual/) (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the [line of best fit](http://www.statisticshowto.com/line-of-best-fit/). Root mean square error is commonly used in climatology, forecasting, and [regression analysis](http://www.statisticshowto.com/probability-and-statistics/regression-analysis/) to verify experimental results.

The formula is:



Where,

* f = forecasts (expected values or unknown results),
* o = observed values (known results).

### **4.2.9 Dependent Mean:**

The dependent t-test (also called the paired t-test or paired-samples t-test) compares the means of two related groups to determine whether there is a statistically significant difference between these means.

You need one dependent variable that is measured on an interval or ratio scale.

You also need one categorical variable that has only two related groups.

### **4.2.10 Coefficient of Variance:**

A coefficient of variation (CV) is a statistical measure of the [dispersion](https://www.investopedia.com/terms/d/dispersion.asp) of data points in a data series around the mean. It is calculated as follows: (standard deviation) / (expected value). The coefficient of variation represents the ratio of the [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp) to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another. It is calculated as follows:

Coefficient Of Variation (CV)

### **4.2.11 R-Squared:**

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

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

* 0% indicates that the model explains none of the variability of the response data around its mean.
* 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data.

### **4.2.12 Adjusted R-Squared:**

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it’s usually not.  It is always lower than the R-squared.

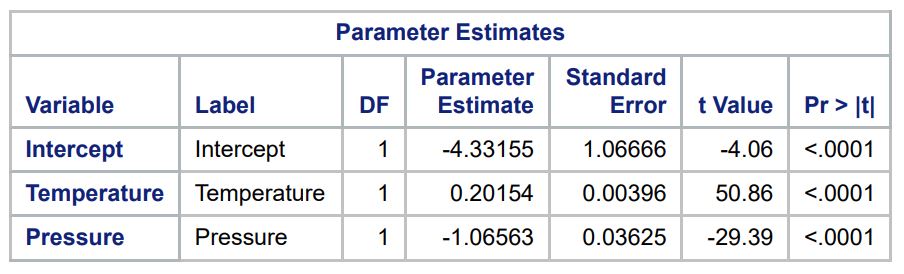


Figure 5: Parameter Estimates

### **4.2.13 Intercept:**

The intercept (often labeled the constant) is the expected mean value of Y when all X=0, where Y is dependent variable and X is independent variable.

### **4.2.14 Parameter estimate:**

Parameter estimates (also called coefficients) are the change in the response associated with a one-unit change of the predictor, all other predictors being held constant.

Regression Diagnostics:

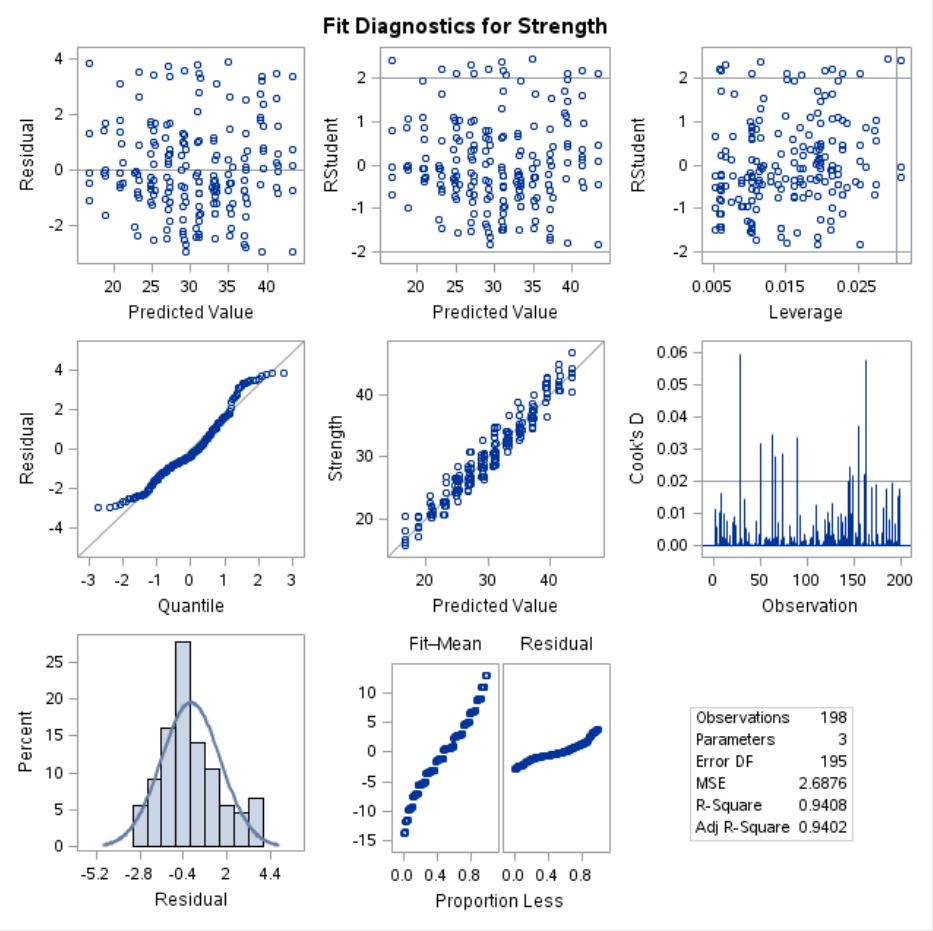


Figure 6: Fit Diagnostics for dependent Variable - Strength

### **4.2.15 Residuals are measured as follows:**

residual =  observed y   –   model-predicted y

The plot of residuals versus predicted values is useful for checking the assumption of linearity and homoscedasticity. If the model does not meet the linear model assumption, we would expect to see residuals that are very large (big positive value or big negative value). To assess the assumption of linearity we want to ensure that the residuals are not too far away from 0 **(standardized** values less than -2 or greater than 2 are deemed problematic). To assess if the homoscedasticity assumption is met we look to make sure that there is no pattern in the residuals and that they are equally spread around the y = 0 line.

### **4.2.16 Normal Probability (Quantile QQ) plots of Residual:**

The **normal probability plot**is a graphical tool for comparing a data set with the [normal distribution](http://www.r-tutor.com/node/58). We can use it with the [standardized residual](http://www.r-tutor.com/node/98) of the [linear regression model](http://www.r-tutor.com/node/91) and see if the error term *ϵ*is actually normally distributed.

### **4.2.17 Cook’s Distance:**

It is a measure of the influence of each observation on the regression coefficients. The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in model estimates when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation.

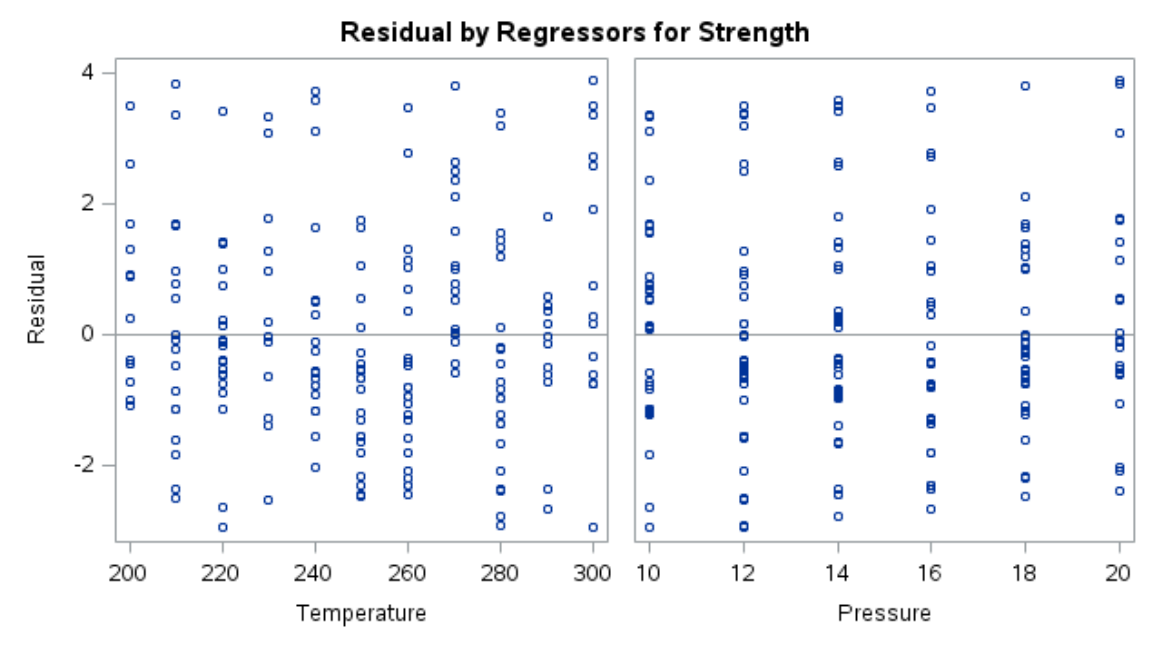


Figure 7: Residuals by Regressors for Strength with respect to temperature and pressure

Scatter-Plots:

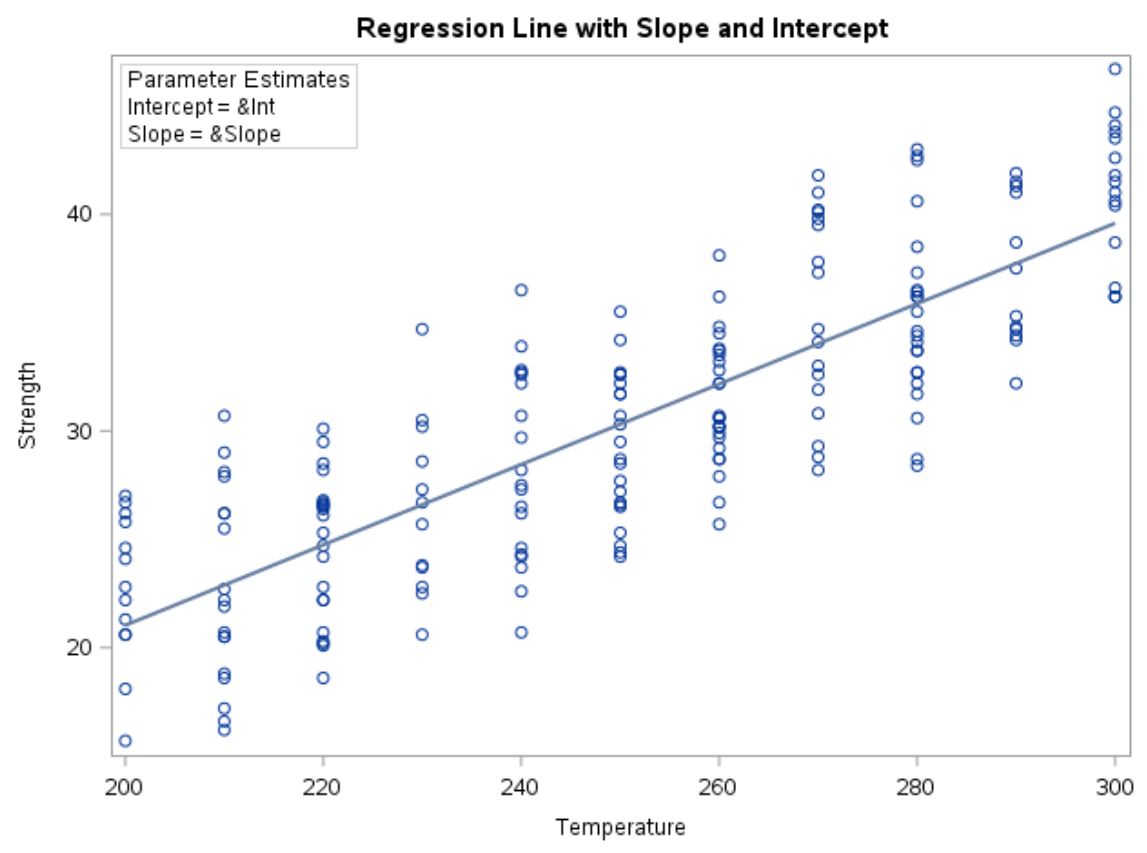


Figure 8: Scatter-Plot with respect to Temperature

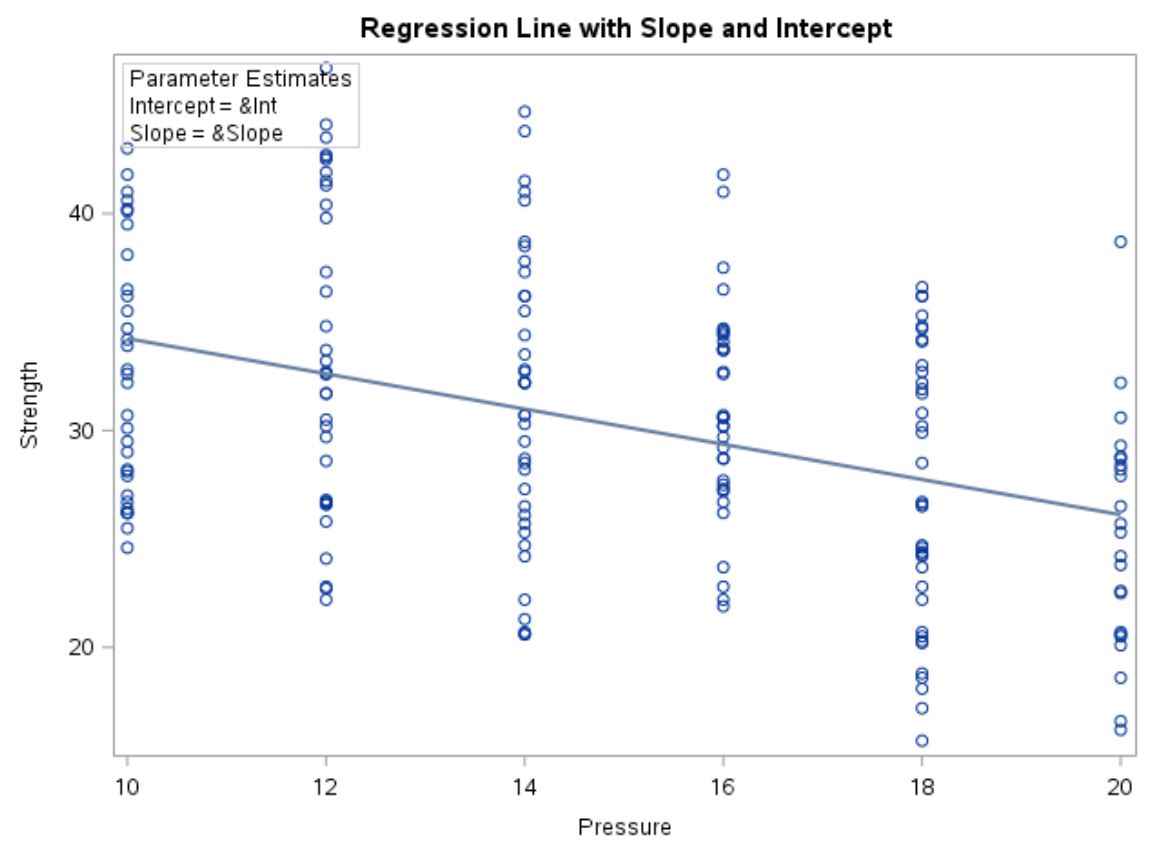


Figure 9: Scatter-Plot with respect to Pressure

## **4.3 Process Followed:**

The regression analysis is done in SAS with the help of a piece of code which processes the data and produces the relevant output.

The code is as follows:

*proc print data=WORK.IMPORT;*

*run;*

*proc means data=WORK.IMPORT;*

*var Strength Temperature Pressure;*

*run;*

*proc reg data=WORK.IMPORT;*

*model Strength = Temperature Pressure;*

*run;*

The WORK.IMPORT is the table on which the whole operation is conducted.

The ‘model’ command does the work of doing the regression by specifying the one on the left-hand side as dependent variable and the ones on the right-hand side as the independent variable.

The code for generating Scatter-Plot is as follows:

*proc sgplot data=WORK.IMPORT noautolegend;*

*title "Regression Line with Slope and Intercept";*

*reg y=Strength x=Temperature;*

*inset "Intercept = &Int" "Slope = &Slope" /*

*border title="Parameter Estimates" position=topleft;*

*run;*

*proc sgplot data=WORK.IMPORT noautolegend;*

*title "Regression Line with Slope and Intercept";*

*reg y=Strength x=Pressure;*

*inset "Intercept = &Int" "Slope = &Slope" /*

*border title="Parameter Estimates" position=topleft;*

*run;*

The ‘sgplot’ command generates the scatter-plot.

Such simple it is to generate regression analytical tables in SAS. Just what is required is command of code and all the required parameters are calculated inherently and regression is done.

# **5. Results:**

## **5.1 Description of results:**

The whole purpose of doing the regression analysis was to know how the strength of plastics is related and changed with respect to temperature and pressure. We did this to most favourable strengths of plastics which are easily recyclable.

The outcome came as expected. The strength of the plastics is co-related with temperature and pressure to a much higher degree. So, by just knowing the appropriate temperature and pressure, the plastic’s strength can be determined which suits best for recycling.

Moreover, the R-Squared value is much closer to 1 (0.9408) from which we can deduce that variables are highly co-related to each other.

## **5.2 Comparison with 10% Sample:**

The calculated values of different components are as follows:

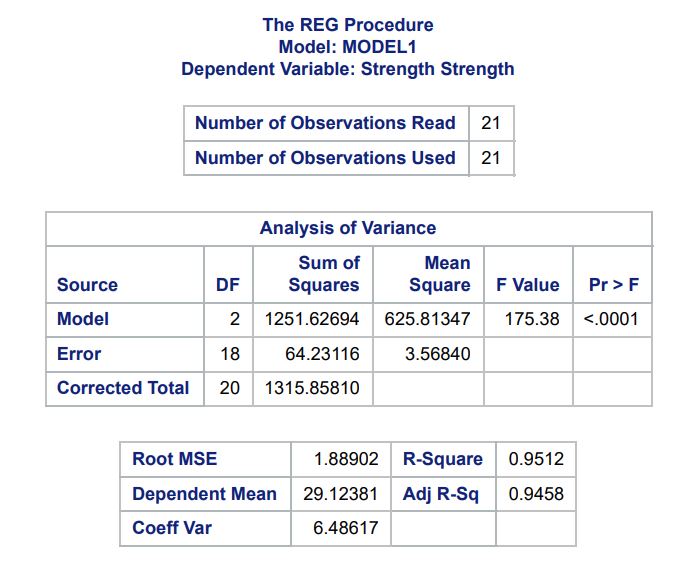


Figure 10: Calculated Components for Regression of 10% random sample from original

The scatter-plot (with respect to temperature and pressure) for the same 10% data is:

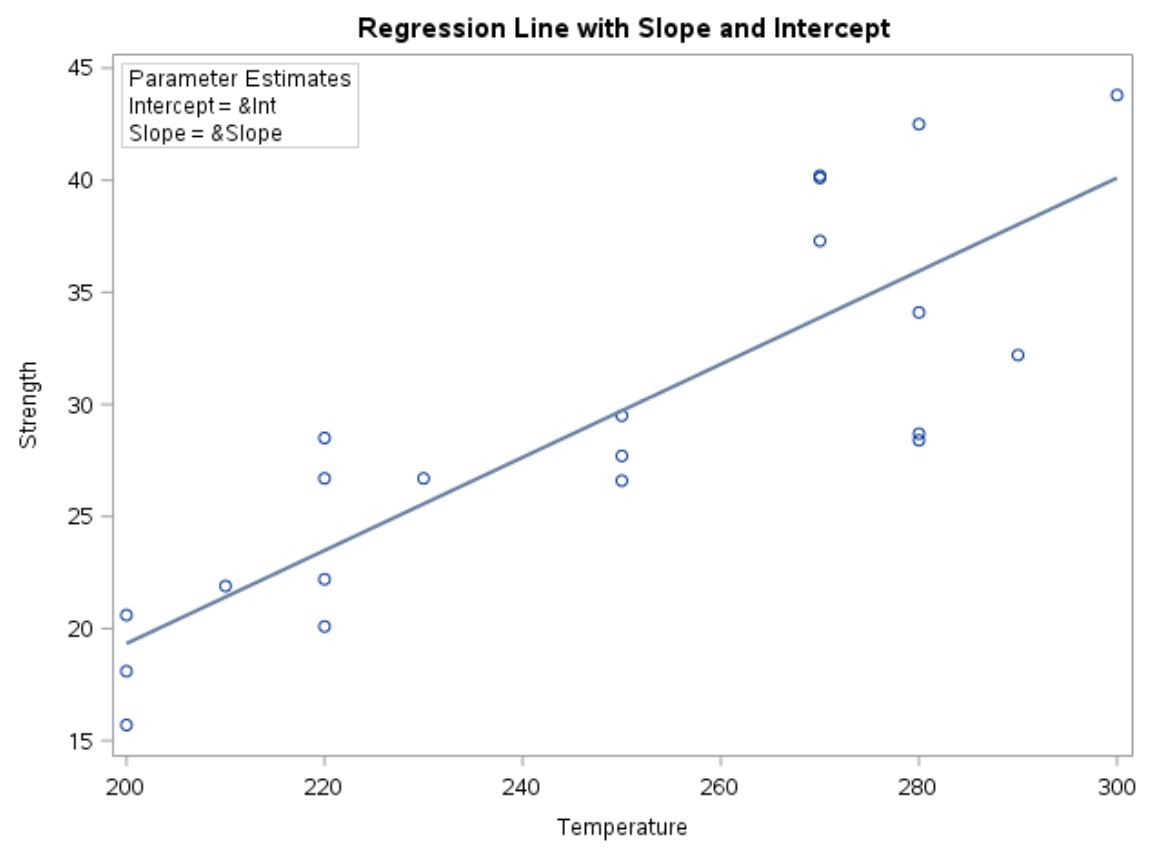


Figure 11: With Respect to Temperature

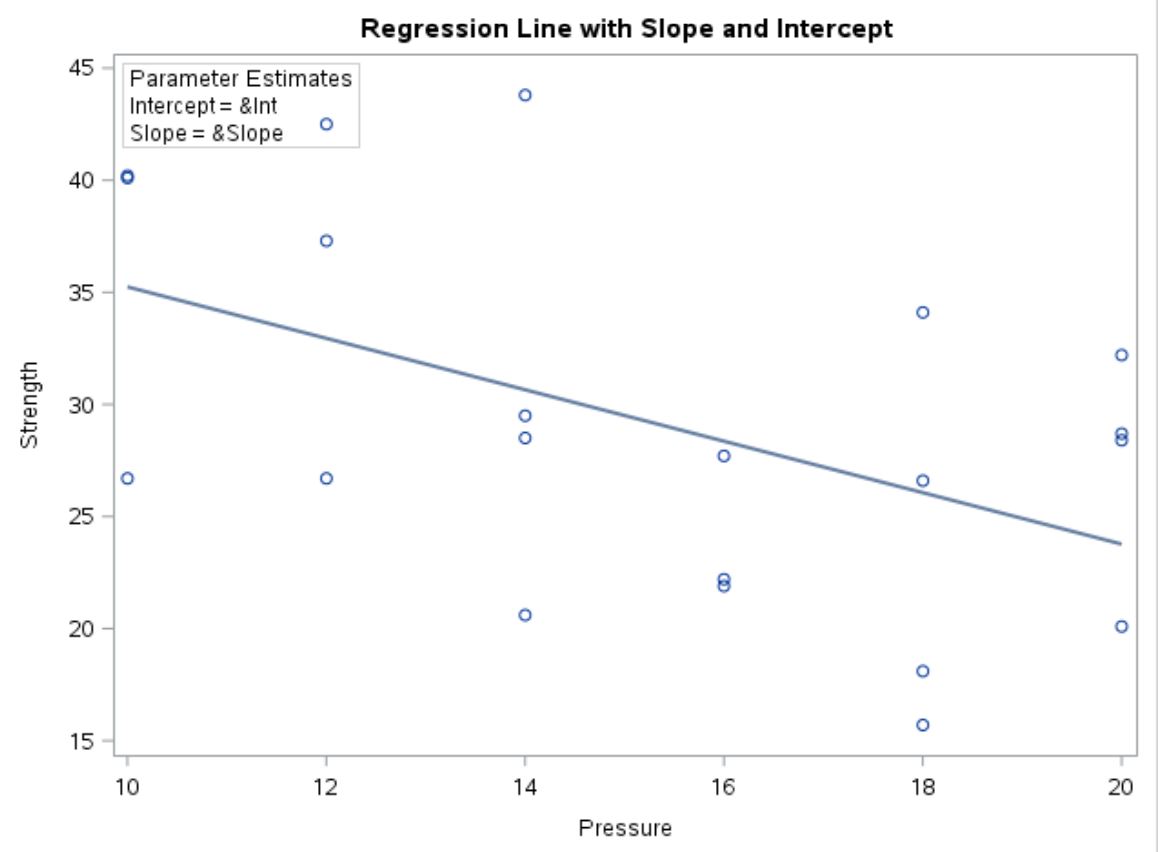


Figure 12: With respect to Pressure

### **5.2.1 Comparison:**

Comparing Original Sample data with 10% data:

We can clearly see from the above figures 4 and 10, that the R-Squared values are 0.9408 and 0.9512 for original sample data and 10% sample data respectively.

So, we can say that it doesn’t create any significant difference in the co-relation between the dependent and independent variable which is the strength and temperature and pressure respective. Besides, the Adjusted R-Squared values and the Coefficient of Variance values doesn’t pose a major factor of difference between the two samples.

When we look at the scatter-plots of both the samples as shown in figures 8 and 9 for original sample and figures 11 and 12 for the 10% sample then we see that for some of the data for 20% samples, they are dispersed somewhat far from the intercept-line for the one with respect to pressure. But still that doesn’t create any exception for those data to be outliers because it follows the straight-line equation. The scatter-plot shown in figure 11, the values are properly graphed close to the straight line giving a surety of high co-relation.

So, we can conclude that for the above case of comparison between the original sample data and the 20% sample data, there is no significant difference in the co-relation of dependent and independent variables.

## **5.3 Comparison with 10% Sample:**

The calculated values of different components are as follows:

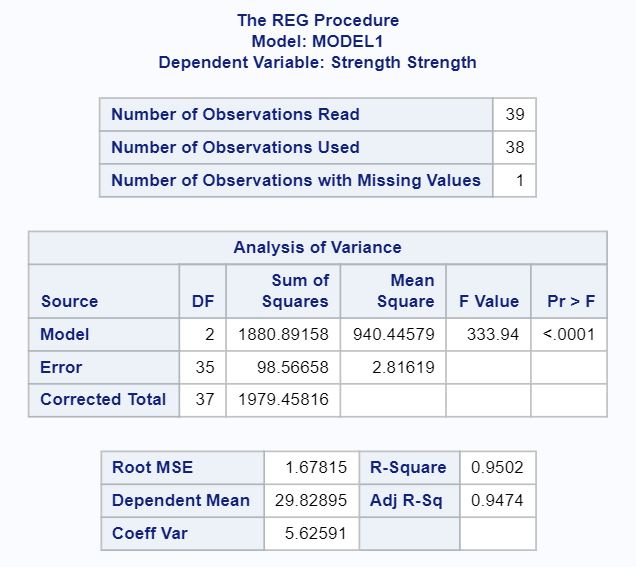


Figure 13: Calculated Components for Regression of 20% random sample from original

The scatter-plot (with respect to temperature and pressure) for the same 10% data is:

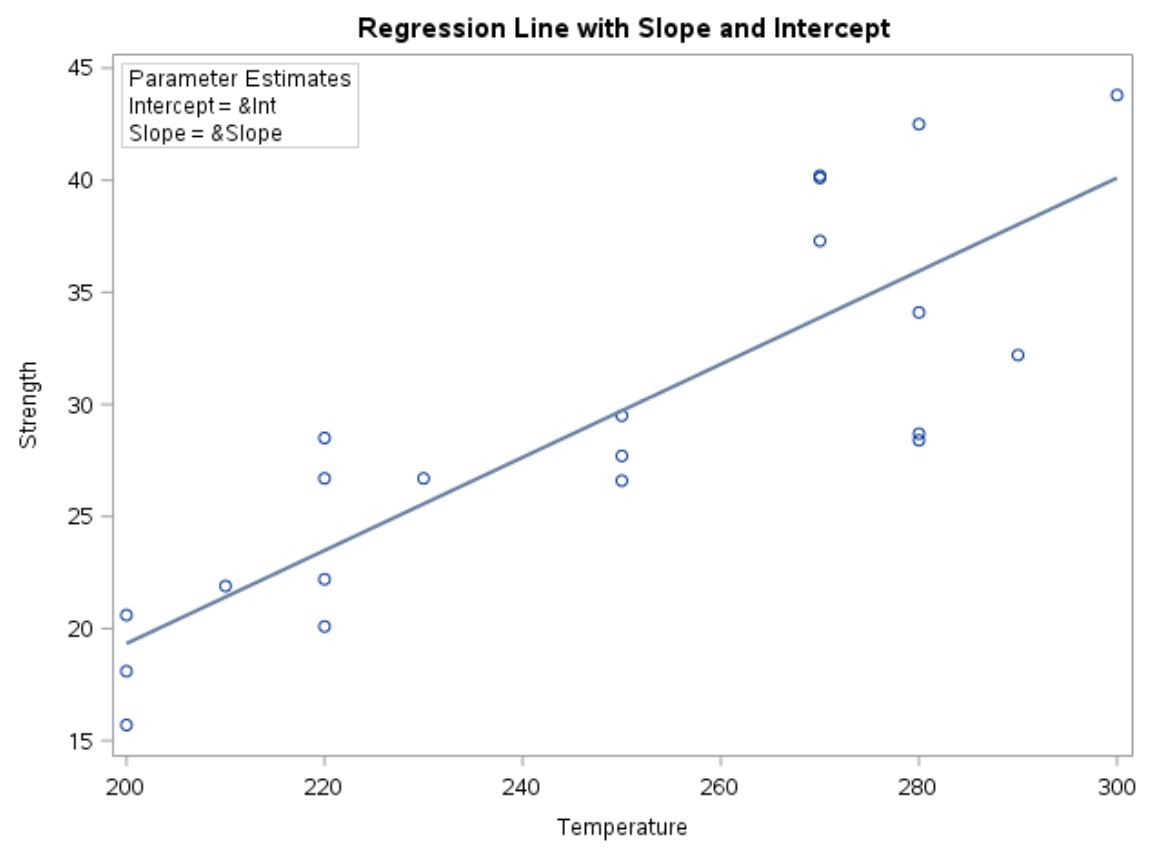


Figure 14: With Respect to Temperature

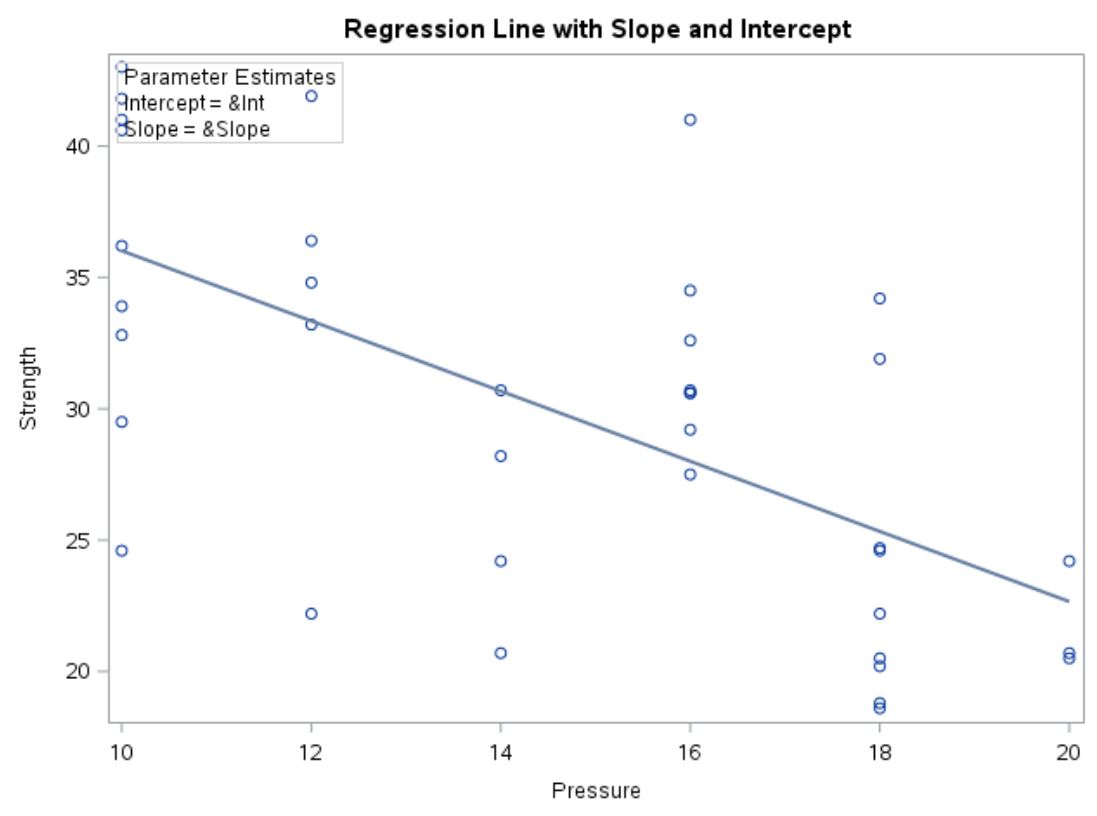


Figure 15: With respect to Pressure

### **5.3.1 Comparison:**

Comparing 20% data with Original and 10% sample data:

By looking the figures 4, 10, and 13, we see the R-Squared values increases as the sample data decreases; which in turn indicates that the co-relation increases with low sample data for this particular type of data.

Moreover, the coefficient of variance also increases as the sample data size decreases.

By looking at the scatter-plot for the sample data of 20%, the data values are evenly dispersed for both the temperature and pressure scatter-plots. Hence, we can conclude that for both the variables the strength of the plastic is evenly co-related.

Therefore, we can finally conclude that all the datasets don’t show any significant difference in co-relation. If we analyze statistically, then we see that co-relation increase as sample data decreases for this particular type of data.

# **6. Next Steps:**

The dataset involved plastic’s strength in general and its relation with the different values of temperature and pressure.

Now the next step to be recommended to improve the analysis should be to bring in the different plastic types into consideration. The data must be collected for the strength of different types of plastics with respect to temperature and pressure. For example, if we take Acetal Copolymer, Acrylic, Nylon6, Polyamide-Imide, polyethylene, and other different types of plastics into consideration, then how will they behave in given temperature and pressure should be observed and noted for analysis.

Furthermore, the dataset must be increased and the temperature and pressure range must also be increased to see the effect. By doing these, we can know whether the particular type of plastic is decomposable/recyclable. We can also add the monetary criterion to examine the expense in recycling such particular plastics at particular temperatures and pressure to analyse whether the type of plastic is worth recycling or not.

# **7. Appendices:**

## **7.1 Figure Table:**

[Figure 1: Means Procedure 6](#_Toc503386133)

[Figure 2: Model1 Dependent Variable Strength 6](#_Toc503386134)

[Figure 3: Analysis of Variance 7](#_Toc503386135)

[Figure 4: Analysis of Variance Contd. 8](#_Toc503386136)

[Figure 5: Parameter Estimates 10](#_Toc503386137)

[Figure 6: Fit Diagnostics for dependent Variable - Strength 11](#_Toc503386138)

[Figure 7: Residuals by Regressors for Strength with respect to temperature and pressure 12](#_Toc503386139)

[Figure 8: Scatter-Plot with respect to Temperature 13](#_Toc503386140)

[Figure 9: Scatter-Plot with respect to Pressure 14](#_Toc503386141)

[Figure 10: Calculated Components for Regression of 10% random sample from original 16](#_Toc503386142)

[Figure 11: With Respect to Temperature 17](#_Toc503386143)

[Figure 12: With respect to Pressure 18](#_Toc503386144)

[Figure 13: Calculated Components for Regression of 20% random sample from original 19](#_Toc503386145)

[Figure 14: With Respect to Temperature 20](#_Toc503386146)

[Figure 15: With respect to Pressure 21](#_Toc503386147)

[Figure 16: Data-Set up to 32 data 23](#_Toc503386148)

## **7.2 Data-set:**

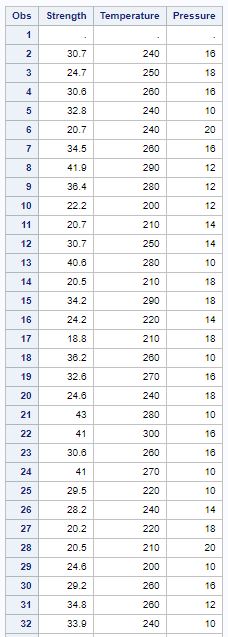


Figure 16: Data-Set up to 32 data

## **7.3 References:**

‘Regression Analysis’: <https://en.wikipedia.org/wiki/Regression_analysis>

‘Variance’: <http://www.statisticshowto.com/probability-and-statistics/variance/>

‘Root Mean Square Error’: <http://www.statisticshowto.com/rmse/>

‘Co-Efficient of Variance’: <https://www.investopedia.com/terms/c/coefficientofvariation.asp>

‘Regression Analysis: How do I assess R-Squared’: <https://www.linkedin.com/pulse/regression-analysis-how-do-i-interpret-r-squared-assess-gaurhari-dass>

‘Regression Diagnostics’: <http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html>

‘Normal Probability Plots of Residuals’: <http://www.r-tutor.com/elementary-statistics/simple-linear-regression/normal-probability-plot-residuals>

‘Compressive Strength Testing of Plastics’: <http://www.matweb.com/reference/compressivestrength.aspx>