

Quantitative Methods

End Term Project: Group 10

**Quantitative Data Analysis of car data in market to draw inferences from attributes and predict user choices**

**A research project submitted in partial fulfilment of the requirements for the**

**QUANTITATIVE METHODS**

**course for the degree of**

**MASTER OF BUSINESS ADMINISTRATION**

**at**

**INDIAN INSTITUTE OF MANAGEMENT, AMRITSAR**

**by**

**GARIMA SHARMA (MBA/06/131)**

**MADHAV SHARMA (MBA/06/140)**

**AKASH R (MBA/06/154)**

**SHREYANS JAIN (MBA/06/163)**

**VIDHI KOCHAR (MBA/06/171)**

**INDIAN INSTITUTE OF MANAGEMENT, AMRITSAR**

**OCTOBER 2020**

**Introduction:**  
 The automobile Industry is a very huge industry. Imagine if someone wanted to enter into the car making business and the person has no clue how to start his research. We are analyzing the past trends in the automobile industry and are trying to predict how the future trends will be like, so that the person who is going to enter the market can get maximum profit.

**Problem Statement:**

To analyze the automobile industry with the given data and to find on which parameters does the demand grow in the past.

**Methodology:**

We collected data from multiple sources (AI hub, Kaggle, etc) and we wanted to identify the relationship between the parameters like the Engine power, Mileage, so on and so forth. The data which we collected had to be cleaned to remove the incomplete data fields, which was done by using the simple excel tools. So, we started by performing Hypothesis testing, proceeded with Sampling, Regression analysis and then the correlation testing.

**Hypothesis Testing:**

Hypothesis testing is a prominent way in statistics to test the results of a survey or experiment to see if you have meaningful results. In this testing whether your results are valid by figuring out the odds that your results have happened by chance. If the results may have happened by chance, the experiment won’t be repeatable and so it has little use.

Hypothesis testing is used to check the plausibility of given hypothesis by using the sample data. This kind of data may come from a larger population, or from a data-generating process.

**How Hypothesis Testing Works**

In the hypothesis testing, we test a statistical sample, with the goal of checking evidence on the plausibility of the null hypothesis. Statistical data users and analysts test a hypothesis by measuring and examining the random sample of the population that is being analysed. All such analysts use the random population sample in order to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis (called the H0) is usually a hypothesis of equality between population or sample parameters; e.g., a null hypothesis may state that population means return is equal to zero.

The alternative hypothesis (called the Ha) on the other hand is effectively the opposite of a null hypothesis; e.g., the population means return is not equal to zero. Thus, these two are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

We either reject null hypothesis or we fail to reject it.

**Four Steps of Hypothesis Testing**

All hypotheses are tested using the below mentioned four-step process:

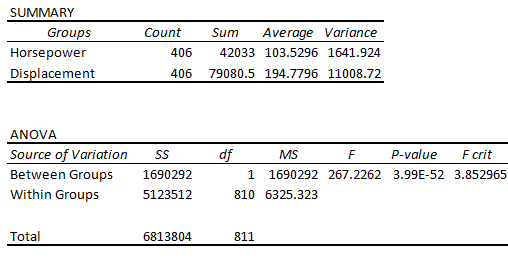
* The first step is to state the two hypotheses so that only one can be proved.
* The next step is to formulate an analysis plan, which depicts how the data will be evaluated.
* The third step is to carry out the plan and actually analyze the sample data.
* The fourth and last step is to analyze the results and either we reject the null hypothesis, or state that the null hypothesis is plausible, given the population data.

**Implementation of Hypothesis**

In our project, we applied two hypothesis testing for two different scenarios.

1. We tested a hypothesis weather cars with more horsepower have more displacement or not?

Considering the two hypothesis-

* Null Hypothesis - Cars with more horsepower do not have more Displacement
* Alternate Hypothesis - Cars with More Horsepower have more Displacement

F is greater than F Critical, so our Null hypothesis is rejected. A Smaller P values means the results are significant.

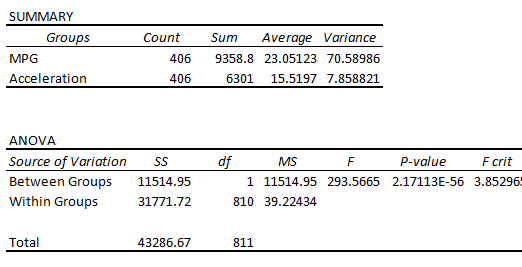
The F value in one-way ANOVA is a tool to help you answer the question “Is the variance between the means of two populations significantly different?” The F value in the ANOVA test also determines the P value; The P value is the probability of getting a result at least as extreme as the one that was actually observed, given that the null hypothesis is true.

Here our P value is very less as compared to level of significance alpha so our **null hypothesis is rejected.**

**Proving - Cars with More Horsepower have more Displacement**

1. We tested a hypothesis Whether cars with higher MPG have higher Acceleration or not?

Considering the two hypothesis-

* Null Hypothesis - Cars with higher MPG (miles per gallon) do not have higher Acceleration
* Alternate Hypothesis - Cars with higher MPG have higher Acceleration

F is greater than F Critical, so our Null is false. A Smaller P values means the results are significant.

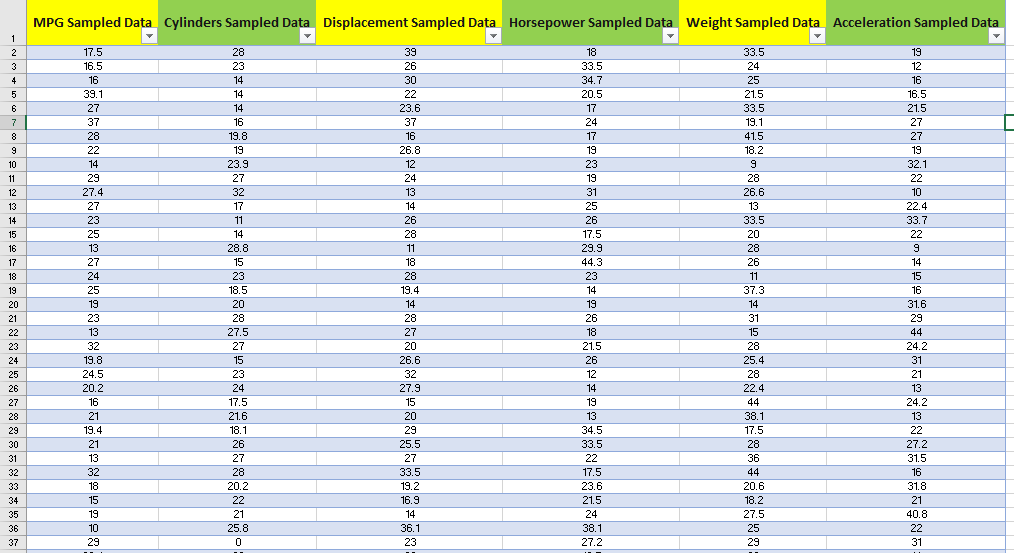
Reject the null when your p value is smaller than your alpha level. You should not reject the null if your critical f value is smaller than your F Value, unless you also have a small p-value.

Our P value is less than the level of significance a. So, we can say that our **null hypothesis is rejected.**

**Proving - Cars with higher MPG have higher Acceleration**

**Sampling:**

Sampling is the technique which is used majorly in statistical analysis having a countable number of predetermined observations recorded or taken from a larger population and then this sampled into smaller chunks of units having lesser number of observations as compared to the original population observations.

To draw meaningful and valid conclusions from the results, we have to cautiously decide how we should select sample which is representative of the whole group from the population set.

There are two types of sampling methods – Probability Sampling and Non-probability sampling.

We can have mainly 4 types for probability sample –

Simple Random Sampling, Systematic Sampling, Stratified Sampling, Cluster Sampling

**In the Project:**

We have used Simple Random Sampling which was applied on the dataset available to us having different Metadata defining columns and around 406 rows filled with data.

We choose Simple Random Sampling and output data was a set of around 25, 50, 75 and 100 observations. For each of the output sample size we checked the variance and standard deviation of parameters from the actual data consisting of 406 rows and it was very large.

From this we can see that even though Simple Random Sampling might seem to be representative of the whole group but still from the inference perspective it might not be always feasible to sample and rather if the population data is manageable we should do the survey/research study on the basis of that only to get a true picture to draw conclusions from!

Periodic sampling was in the same sense not even comparable to the results from that produced by Simple Random so it was not preferred while carrying out the study in this project. A point to note here is that even though the dataset values were not sorted and filtered still the data available is not fairly distributed as there is no comprehensive way to distribute the values fairly across all columns for the defined rows, in this case was for the data for cars distributed across columns for different parameters some of them being horsepower of car, displacement, number of cylinders etc.

Regression and correlation when done on the samples were also quite far varied when compared with the same operations performed on the actual population data set.

**Correlation:**

Since, Correlation gives us an indication of as to how strongly two variables are related, so we performed correlation with different set of variables in order to draw to some meaningful conclusions and figure out how various parameters are related to each other.

So, these are some of the parameters which have strong positive and negative correlation:

**Positive Correlation:**

**Negative Correlation:**

For e.g. when **Cylinders and Weight** of the car when fed to check correlation states that although the weight is… to a great extent is a factor of engine but number of cylinders present in the car also has a **strong positive correlation with the weight of the car.**

Similarly, **Correlation between Displacement and MPG is negative** which suggests that otherwise stated and implied Displacement must be positively correlated in the same direction i.e. with increase in Miles per gallon, displacement should increase but the dataset proves that statement wrong.

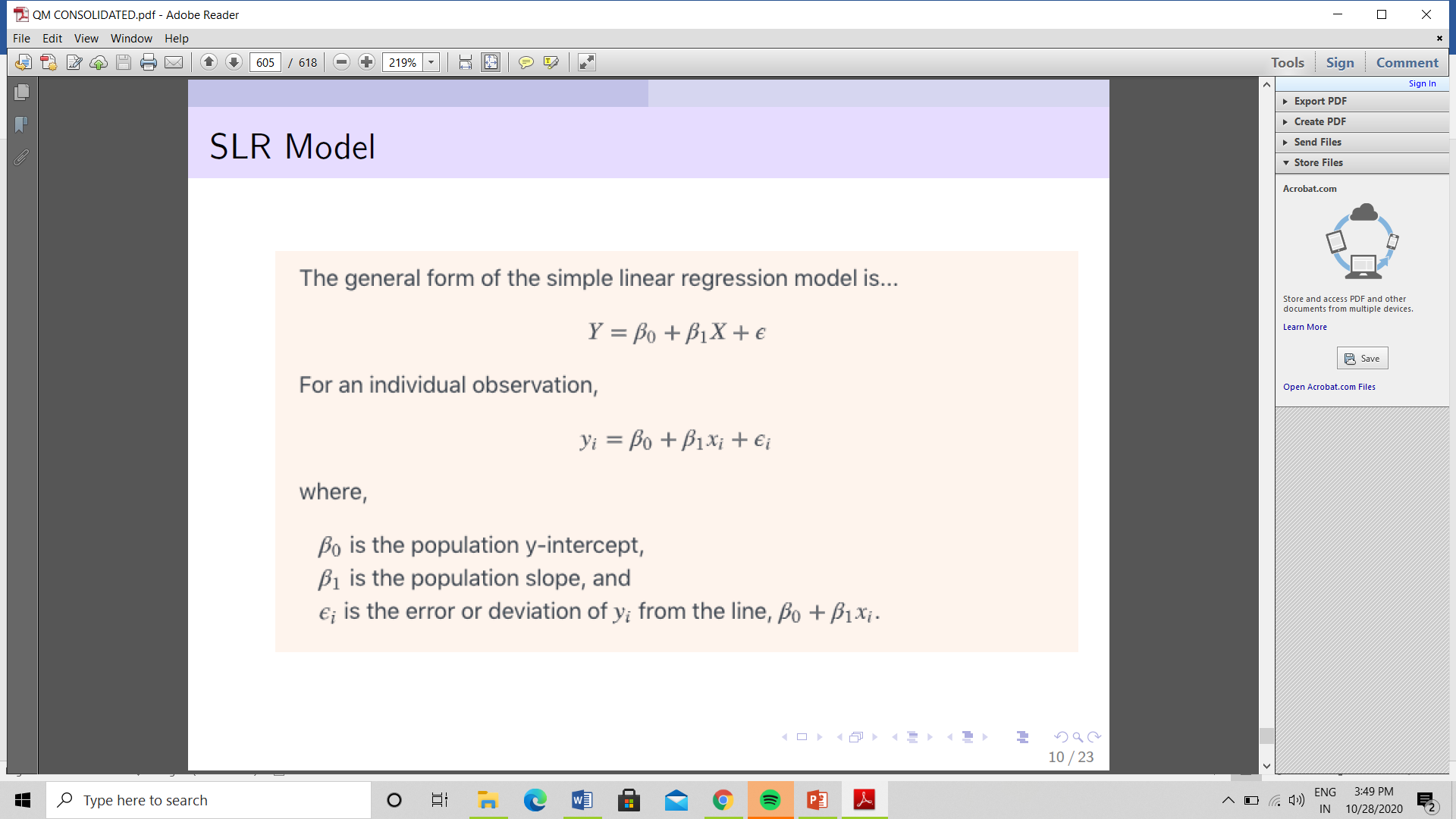
Likewise, we were able to draw such conclusions as to how different variables are related to each other through this analysis.

**Regression:**

Regression Analysis is a tool to investigate how two or more variables are related.

It is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

**The general form of the simple linear regression model is:**



**ASSUMPTIONS OF SIMPLE LINEAR REGRESSION:**

1. **Linearity**: The relationship between X and Y must be linear.

We checked the linearity through the scatter plot.

1. **Independence of Errors**: There is no relationship between the residuals and the Y variable. Hence Y is independent of errors.
2. **Equal Variances:** The variance of the residuals is the same for all values of X.

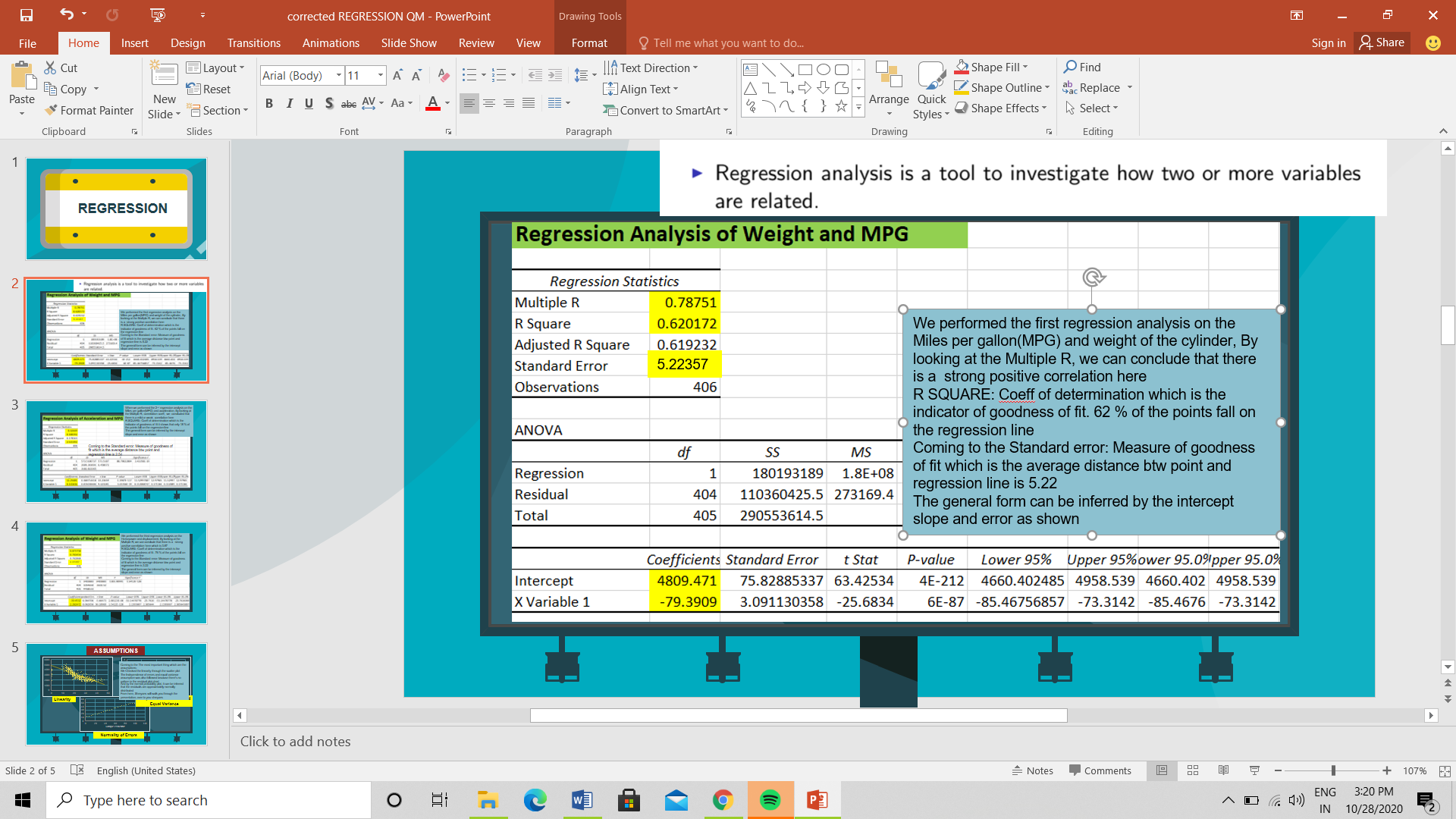
Independence of Errors and Equal Variance assumption was also followed because there’s no pattern in the residual plot chart.

(4) **Normality of Errors:** The residuals must be approximately normally distributed.

By the Normal probability plot, it can be inferred that the residuals are approximately normally distributed.

**In the project**, we performed the first regression analysis on:

1. **Miles per gallon (MPG) and the Weight of the cylinder**



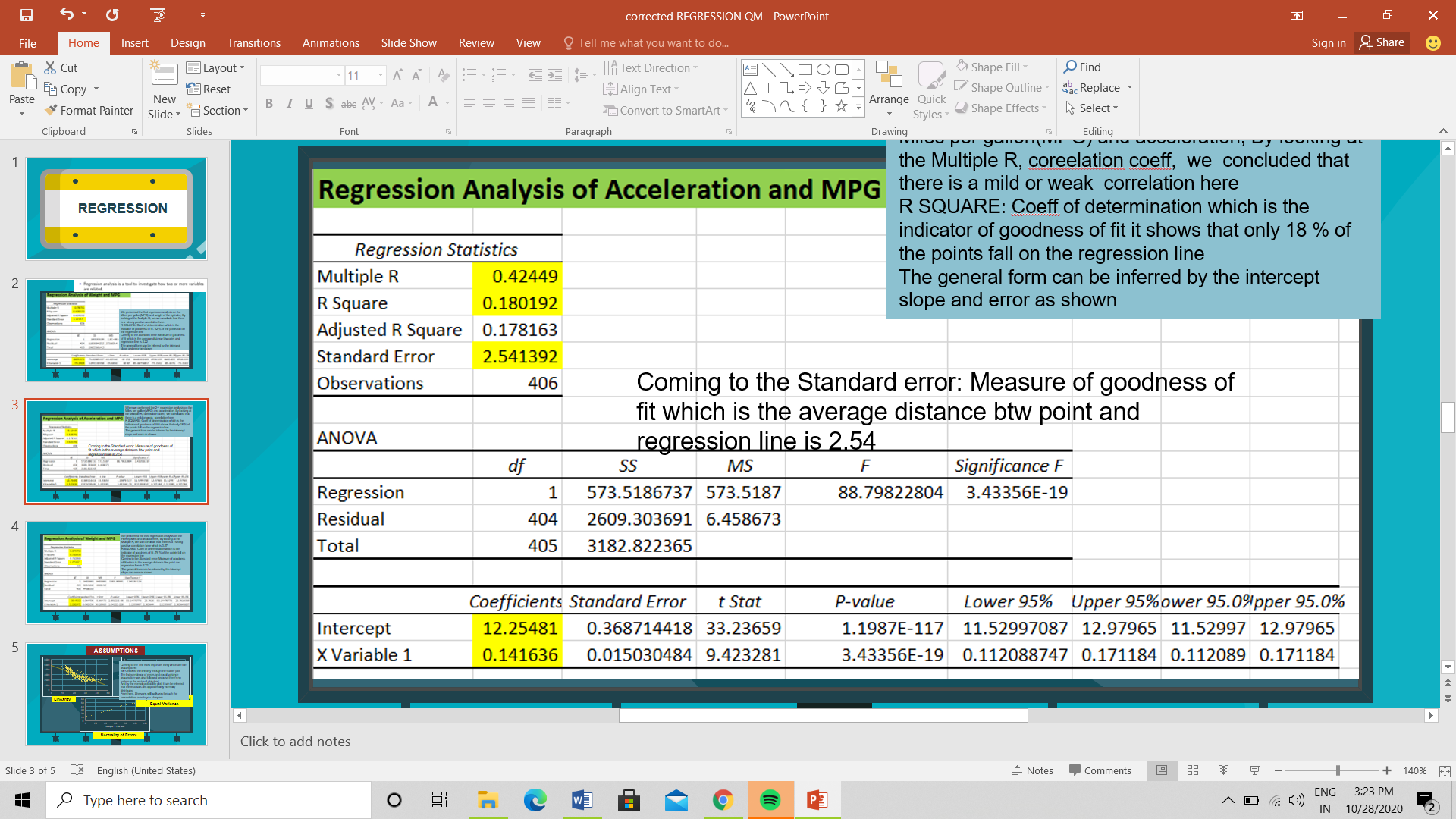
By observing the **Multiple R**, we can conclude that there is a strong positive correlation (0.78) between MPG and Weight of the cylinder.

**R Square**: The Coefficient of determination/ The indicator of goodness of fit indicates that 62 % of the points fall on the regression line

**Standard error**: Measure of goodness of fit which is the average distance btw point and regression line is 5.22

The general form can be inferred by the intercept, slope and error in the excel sheet.

1. **Acceleration and Miles per gallon (MPG)**



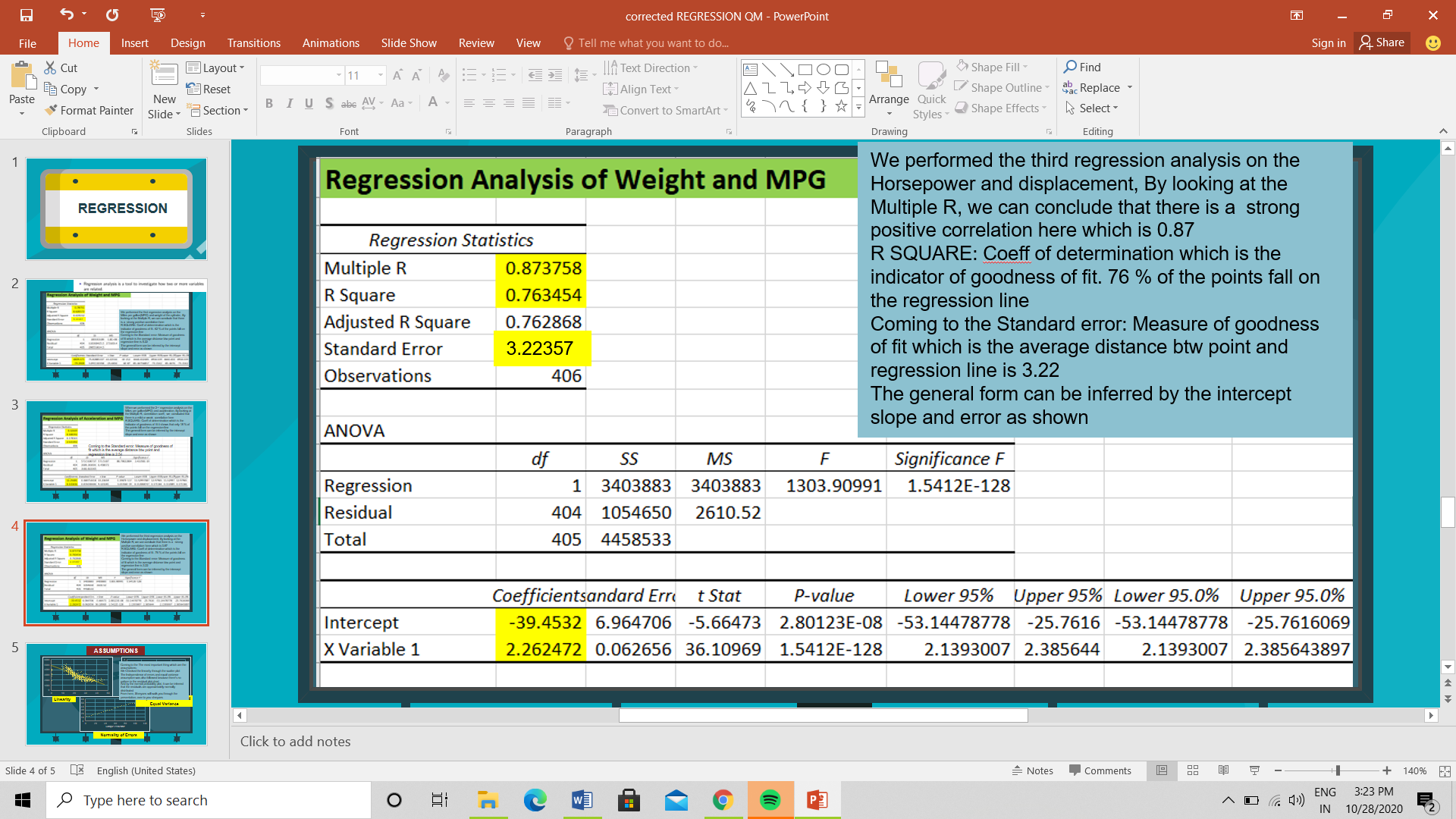
By observing the **Multiple R**, we can conclude that there is a ***mild or weak correlation (0.42)*** between Acceleration and Miles per gallon (MPG).

**R Square**: The Coefficient of determination/ The indicator of goodness of fit indicates that 18 % of the points fall on the regression line

**Standard error**: Measure of goodness of fit which is the average distance btw point and regression line is 2.45

The general form can be inferred by the intercept, slope and error in the excel sheet.

1. **Horsepower and displacement**



By observing the **Multiple R**, we can conclude that there is a ***positive correlation (0.87)*** between Horsepower and displacement

**R Square**: The Coefficient of determination/ The indicator of goodness of fit indicates that 76 % of the points fall on the regression line

**Standard error**: Measure of goodness of fit which is the average distance btw point and regression line is 3.22

The general form can be inferred by the intercept, slope and error in the excel sheet.

**Conclusion:**

From the analysis which we performed, we found that

1) The hypothesis testing was performed and the null hypothesis was rejected in both the cases which was checked after different steps just for sanity check, analysis which were expected considering the dataset and the questions we asked and assumed at the first step of Hypothesis testing, proving that cars with high MPG have higher acceleration and cars with more horsepower have more displacement!

2) Random Sampling was done from the dataset with 25, 50, 75 and 100 data and regression and correlation was performed, but the results were not congruent with the analysis performed with the whole data which was not expected as random sampling is preferred and is generally representative of whole set but still the variances were huge so we finally did the analysis and inference from mathematical results were taken performed on actual data available having 406 rows and multiple columns.

3) Correlation Analysis was performed on the data and we found out which parameters have a positive correlation and which parameters have negative correlation. Major key takeaway we take from the correlation performed on this data after observing negative correlation between miles per gallon i.e., MPG vs Displacement is that although they should be positively correlated and that too with a high factor but still there are other factors which might affect these factors hence they are not truly independent so it is important to understand the parameters first, how they change, what are key factors affecting those and how can these be dependent or independent and then further analysis should be carried out.

4) The regression analysis was done after checking the necessary conditions required to perform Regression Analysis with multiple parameters having independent inputs and dependent outputs and based on the R-value; we were able to conclude the relationship between the attributes such as Weight of the car being independent input and MPG being the dependent output, Acceleration being the independent input and MPG being the dependent output and Horsepower being the independent output. We observed that weight and MPG have a close interknitted connection and can be used to predict the user behavior and trend analysis. Similarly, we worked for others and meaningful conclusions were drawn from the same to check how one parameter is triggering change in other.

**Individual Contribution:**

**Garima Sharma (MBA/06/131):** Brainstorming the ideas, performing null and alternative Hypothesis as per the scenarios and datasets relational linking, analysis as per the Hypothesis and the inference that can be drawn from them.

**Madhav Sharma (MBA/06/140):** Brainstorming the ideas, Sampling of data from the given 406 columns having different metadata from varied parameters associated with car models, comparison of results obtained from Random sampling, Periodic Sampling with those from the actual datasets available as per the deviations and what meaningful insights can be drawn from those.

**Akash R (MBA/06/154):** Collection, Cleaning, check and bounds performed on different datasets available from open sources. Selecting 3 datasets, making them comparable and collating them for analysis and presentation purpose. Datasets were obtained from AI Hub, Statistica and data hub.

**Shreyans Jain (MBA/06/163):** Careful analysis of the parameters, performing correlation on different parameters keeping in mind how directly and indirectly can they affect each other. Some results were astonishing as they were exactly opposite to the general view we hold about them.

**Vidhi Kochar (MBA/06/171):** Mathematical use of data in performing Regression analysis, not only linear one but we also tried multiple parameters to see the trigger they cause in our desired output parameter. Necessary conditions were checked and then proceeded for the inference and analysis part.

**Dataset Sources:**

<https://www.kaggle.com/jessicali9530/stanford-cars-dataset>

<https://ai.stanford.edu/~jkrause/cars/car_dataset.html>

<https://aihubprojects.com/used-cars-price-estimation-a-regression-model/>

<https://towardsdatascience.com/exploring-the-us-cars-dataset-dbcebf954e4a>