



Regression Analysis is a powerful statistical method that allows you to examine relationship between two or more variable of interest.



# MEMBERS

DIYA HALDANKAR

PARVATHY NAIR

SHRENIKA BHOSLE



# WHAT IS REGRESSION?

Regression is defined as a statistical method that helps us to analyze and understand the relationship between two or more variables of interest. The process that is adapted to perform regression analysis helps to understand which factors are important, which factors can be ignored, and how they are influencing each other. In regression, we normally have one dependent variable and one or more independent variables. Here we try to “regress” the value of the dependent variable “Y” with the help of the independent variables. In other words, we are trying to understand, how the value of ‘Y’ changes w.r.t change in ‘X’. For the regression analysis to be a successful method, we understand the following terms:

- *Dependent Variable*: This is the variable that we are trying to understand or forecast.
- *Independent Variable*: These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.



# What is Regression Analysis?

Regression analysis is used for prediction and forecasting. This has substantial overlap with the field of machine learning. This statistical method is used across different industries such as,

- **Financial Industry-** Understand the trend in the stock prices, forecast the prices, and evaluate risks in the insurance domain
- **Marketing-** Understand the effectiveness of market campaigns, and forecast pricing and sales of the product.
- **Manufacturing-** Evaluate the relationship of variables that determine to define a better engine to provide better performance
- **Medicine-** Forecast the different combinations of medicines to prepare generic medicines for diseases.



# TERMINOLOGIES USED IN REGRESSION ANALYSIS

## Outliers

Suppose there is an observation in the dataset that has a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier. In simple words, it is an extreme value. An outlier is a problem because many times it hampers the results we get.

## Multicollinearity

When the independent variables are highly correlated to each other, then the variables are said to be multicollinear. Many types of regression techniques assume multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on its importance, or it makes the job difficult in selecting the most important independent variable.

## Heteroscedasticity

When the variation between the target variable and the independent variable is not constant, it is called heteroscedasticity.

Example-As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times, eat expensive meals. Those with higher incomes display a greater variability of food consumption.

## Underfit and Overfit

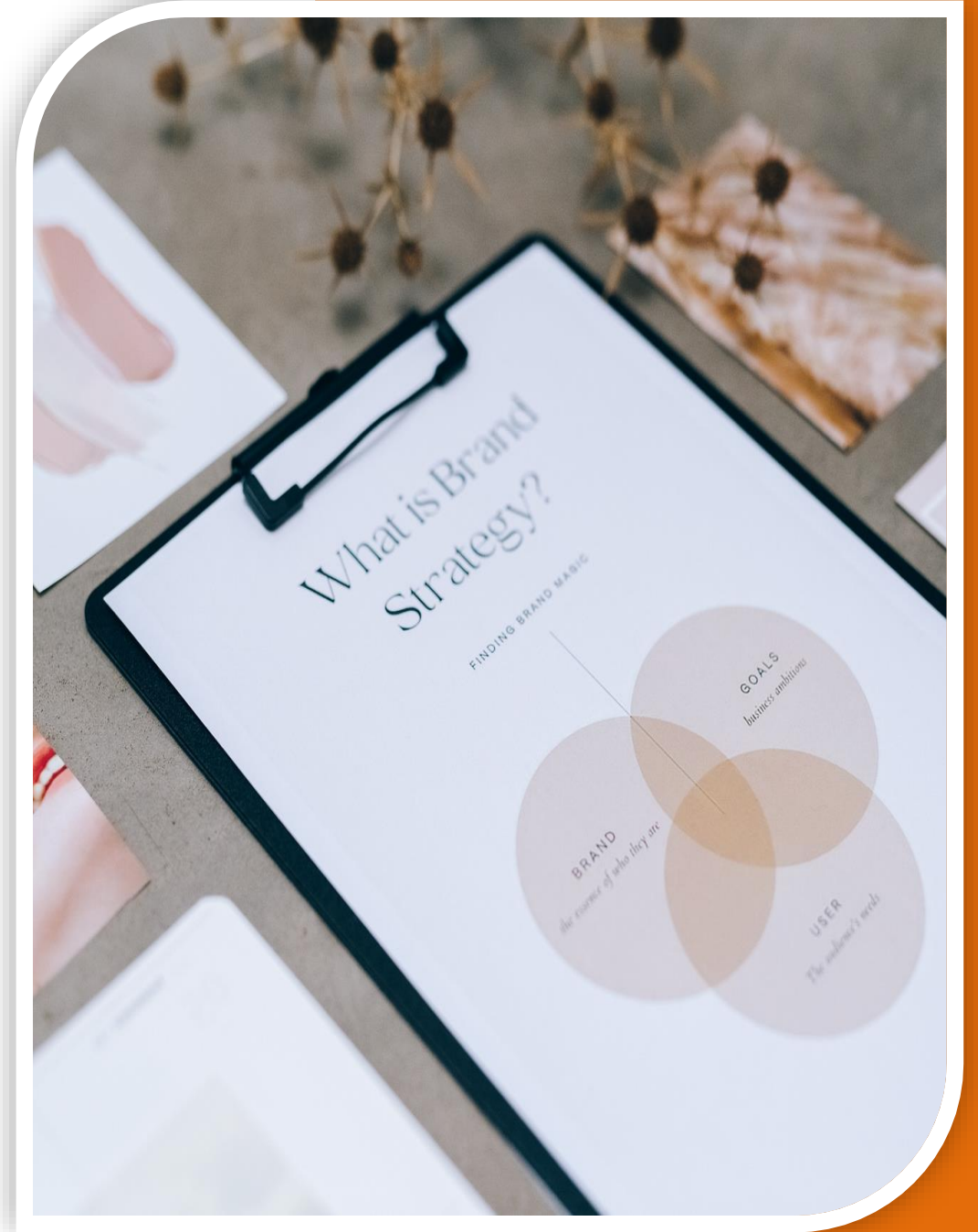
When we use unnecessary explanatory variables, it might lead to overfitting. Overfitting means that our algorithm works well on the training set but is unable to perform better on the test sets. It is also known as a problem of **high variance**.

When our algorithm works so poorly that it is unable to fit even a training set well, then it is said to underfit the data. It is also known as a problem of **high bias**.

## Types of Regression

For different types of Regression analysis, there are assumptions that need to be considered along with understanding the nature of variables and their distribution.

- Linear Regression
- Polynomial Regression
- Logistic Regression





# What is Linear Regression?

Linear Regression is a predictive model used for finding the **linear** relationship between a dependent variable and one or more independent variables.

Here, 'Y' is our dependent variable, which is a continuous numerical and we are trying to understand how 'Y' changes with 'X'.

## Simple Linear Regression

$$X \dashrightarrow Y$$

As the model is used to predict the dependent variable, the relationship between the variables can be written in the below format

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where,

$Y_i$  – Dependent variable

$\beta_0$  -- Intercept

$\beta_1$  – Slope Coefficient

$X_i$  – Independent Variable

$\varepsilon_i$  – Random Error Term

The diagram shows the linear regression equation  $\hat{y} = \beta_0 + \beta_1 X$  enclosed in a light gray box. Below the equation, four red arrows point upwards to the terms:  $\hat{y}$ ,  $\beta_0$ ,  $\beta_1$ , and  $X$ . Below each arrow is a label: "Predicted value" for  $\hat{y}$ , "Intercept" for  $\beta_0$ , "Slope" for  $\beta_1$ , and "Predictor" for  $X$ .

$$\hat{y} = \beta_0 + \beta_1 X$$

Predicted value    Intercept    Slope    Predictor

# Assumptions

Since Linear Regression assesses whether one or more predictor variables explain the dependent variable and hence it has 5 assumptions:

- Linear Relationship
- Normality
- No or Little Multicollinearity
- No Autocorrelation in errors
- Homoscedasticity

With these assumptions considered while building the model, we can build the model and do our predictions for the dependent variable. For any type of machine learning model, we need to understand if the variables considered for the model are correct and have been analysed by a metric. In the case of Regression analysis, the statistical measure that evaluates the model is called the coefficient of determination which is represented as  $r^2$ .

The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable. A higher value of  $r^2$  better is than the model with the independent variables being considered for the model.

$$r^2 = \frac{SSR}{SST}$$

**Note: The value of  $r^2$  is the range of  $0 \leq r^2 \leq 1$**



# Logistic Regression

Logistic Regression is also known as Logit, Maximum-Entropy classifier is a supervised learning method for classification. It establishes a relation between dependent class variables and independent variables using regression.

The dependent variable is categorical i.e. it can take only integral values representing different classes. The probabilities describing the possible outcomes of a query point are modelled using a logistic function. This model belongs to a family of discriminative classifiers. They rely on attributes which discriminate the classes well. This model is used when we have 2 classes of dependent variables. When there are more than 2 classes, then we have another regression method which helps us to predict the target variable better.

There are two broad categories of Logistic Regression algorithms

Binary Logistic Regression when the dependent variable is strictly binary

Multinomial Logistic Regression is when the dependent variable has multiple categories.

**There are two types of Multinomial Logistic Regression**

Ordered Multinomial Logistic Regression (dependent variable has ordered values)

Nominal Multinomial Logistic Regression (dependent variable has unordered categories)

# Assumptions

The dependent variable is categorical. Dichotomous for binary logistic regression and multi-label for multi-class classification

Attributes and log odds i.e.  $\log(p / 1-p)$  should be linearly related to the independent variables

Attributes are independent of each other (low or no multicollinearity)

In binary logistic regression class of interest is coded with 1 and other class 0

In multi-class classification using Multinomial Logistic Regression or OVR scheme, class of interest is coded 1 and rest 0 (this is done by the algorithm)

## Some examples where this model can be used for predictions.

**Predicting the weather:** You can only have a few definite weather types. Stormy, sunny, cloudy, rainy and a few more.

**Medical diagnosis:** Given the symptoms predicted the disease patient is suffering from.

**Credit Default:** If a loan has to be given to a particular candidate depends on his identity check, account summary, any properties he holds, any previous loan, etc

**Elections:** Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign and the amount of time spent campaigning negatively

# The Six Assumptions of Linear Regression

## 1) The population model (or the *true* model) is linear in its parameters.

Below is a simple regression model, where  $Y$  is the target variable,  $X$  is the independent variable, and epsilon is the error term (randomness not captured by the model).

$$Y = B_0 + B_1X + \epsilon$$

What we mean by ‘linear in its parameters’ is that the population model may have a mathematical transformation (a square root, a logarithm, a quadratic) on the target variable, or the independent variables, but not on the parameters.

Thus, changes in our independent variables will have the same marginal effect regardless of their value.



## 2) We have a random sample of $n$ observations.

This assumption assures us that our sample is representative of the population. More specifically, it assures us that the sampling method does not affect the characteristics of our sample.

## 3) No perfect collinearity.

The independent variables do not share a perfect, linear relationship. They can be related in some fashion — indeed, we would not include variables in our regression model if they were entirely *unrelated*— however, we should not be able to write one variable as a linear combination of another variable.

$RSS_1$  = Residual Sum of Squares of fitted model 1

$RSS_2$  = Residual Sum of Squares of fitted model 2

$$F \text{ statistic} = \frac{\left(\frac{RSS_1 - RSS_2}{k_2 - k_1}\right)}{\left(\frac{RSS_2}{n - k_2}\right)}$$

## 4) Zero conditional mean

The error term, epsilon, conditional on the independent variables, equals zero, on average. That is, the error term is unrelated to our independent variables.

$$E(\epsilon | x_1, x_2, \dots, x_k) = 0$$

## 5) Homoskedasticity

In statistics parlance, homoskedasticity is when the variance of a random variable is constant.

Linear regression assumes that the error term, conditional on the independent variables, is homoscedastic. That is,  $Var(\epsilon | x_1, x_2, \dots, x_k) = 0$



## 6) Normality

The error term is normally distributed with zero mean and a constant variance

### What Happens When We Break Some of These Assumptions? (And What Can We Do About It?)

#### 1) *The true model is non-linear*

Specifically, suppose the true model was of the form

$$Y = \ln(B_0 + B_1X + \epsilon)$$

but we estimated a linear model.

Our parameter estimates would be biased, and our model would make poor predictions.

There are two ways to determine if you should use a linear function or a non-linear function to model the relationship in the population.

You could create a scatter plot between the two variables and see if the relationship between them is linear or non-linear. You can then compare the performance between a linear regression and a non-linear regression, and choose the function that performs best.

A second method is to fit the data with a linear regression, and then plot the residuals. If there is no obvious pattern in the residual plot, then the linear regression was likely the correct model. However, if the residuals look non-random, then perhaps a non-linear regression would be the better choice.

#### 2) *Our sample is non-random*

Say that you want to test if lock downs have impacted consumer spending decisions. You happened to have a survey of online shoppers prior to the pandemic, and you decide to take a second survey of online shoppers three months into the pandemic.

This is an example of non-representative sampling. Not everyone likes to shop online, and people who do shop online may have traits that are absent from people who refuse to shop online. As a researcher, it is impossible for you to know if it is these unobserved traits, or if it is the lock downs, that changed consumer spending decisions.

Of course, perfect random sampling is impossible. Thus, it is good practice to do some EDA prior to building a regression model to confirm that the two groups are not drastically different.



### 3) We Have Perfect Collinearity

Suppose that

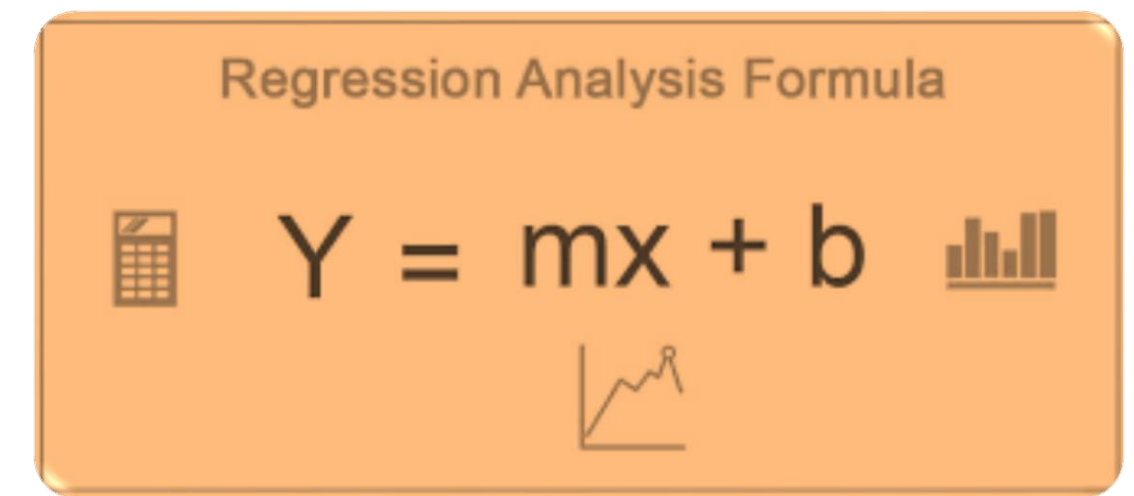
$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \epsilon$$

and that  $X_3 = X_1 + X_2$

In that case, our third variable is a linear combination of the first two variables.

Recall that multiple linear regression estimates the effect of one variable by holding all other variables constant. However, this *all else equal* assumption is impossible in the above regression model. If we change one variable, the first variable, for example, then that changes the third variable. Similarly, if we change the third variable, then that changes the first variable or the second variable (or both).

The solution to perfect collinearity is to drop one of the variables (in the above example, we would drop the third variable).



### 4) Our Error Term is Correlated with One of Our Independent Variables

This occurs if our regression model differs from the true model. For example, we might think that the true model is  $Y = B_0 + B_1X + \epsilon$  when the true model is, in fact,

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \epsilon$$

The exclusion of the second and third independent variables causes **omitted variable bias**. Our slope estimate,  $B1$ , will either be larger or smaller, on average, than the true value of  $B1$ .

There are two solutions. First, if you know the variables that should be included in the true model, then you can add these variables to the model you are building. This is the best solution; however, it is also unrealistic because we can never truly know what variables are in the true model.

The second solution is to conduct a Randomized Control Trial (RCT). In an RCT, the researcher randomly allocates participants to the treatment group or the control group. Because the treatment is given randomly, the relationship between the error term and the independent variables is equal to zero.

6) Our Errors are Non-Normal

Similar to what occurs if assumption five is violated, if assumption six is violated, then the results of our hypothesis tests and confidence intervals will be inaccurate.

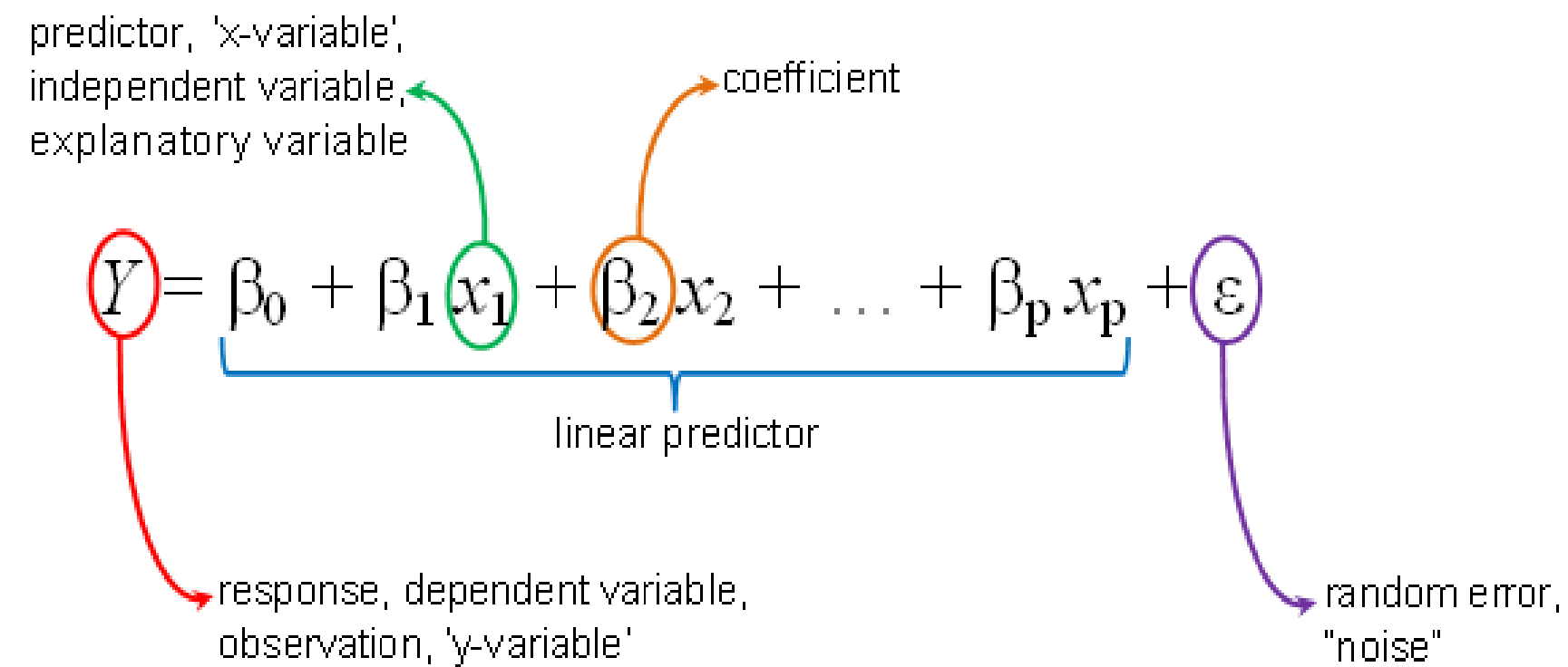
One solution is to transform your target variable so that it becomes normal. This can have the effect of making the errors normal, as well. The log transform and square root transform are most common. If you want to get fancy, then you can also use a Box-Cox transformation

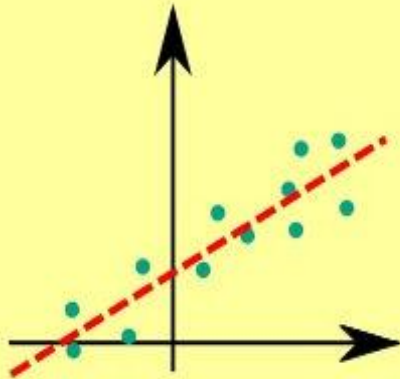
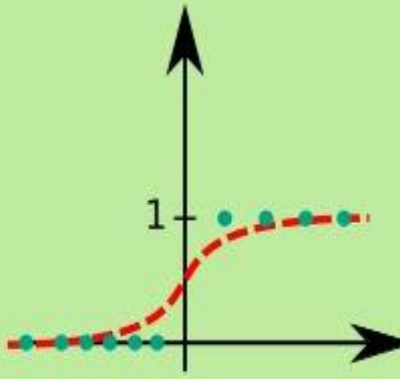
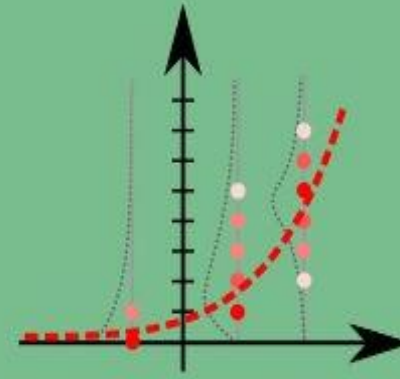
Bonus

The above assumptions only hold true if we are working with cross-sectional data. Linear regression requires different assumptions if we have panel data or time series data.

Conclusion

Now you know the six assumptions of linear regression, the consequences of violating these assumptions, and what to do if these assumptions *are* violated.



LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION
<ul style="list-style-type: none"><li>1 Econometric modelling</li><li>2 Marketing Mix Model</li><li>3 Customer Lifetime Value</li></ul>	<ul style="list-style-type: none"><li>1 Customer Choice Model</li><li>2 Click-through Rate</li><li>3 Conversion Rate</li><li>4 Credit Scoring</li></ul>	<ul style="list-style-type: none"><li>1 Number of orders in lifetime</li><li>2 Number of visits per user</li></ul>
		
Continuous ⇒ Continuous	Continuous ⇒ True/False	Continuous ⇒ 0,1,2,...
$y = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y = \frac{1}{1 + e^{-z}}$ $z = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y \sim \text{Poisson}(\lambda)$ $\ln \lambda = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$
<code>lm(y ~ x1 + x2, data)</code>	<code>glm(y ~ x1 + x2, data, family=binomial())</code>	<code>glm(y ~ x1 + x2, data, family=poisson())</code>
1 unit increase in x increases y by $\alpha$	1 unit increase in x increases log odds by $\alpha$	1 unit increase in x multiplies y by $e^\alpha$



# REGRESSION ANALYSIS

PDF

[using python.pdf](#)



[using R.pdf](#)





# CONCLUSION

From the output (of R-code) we can see the following:

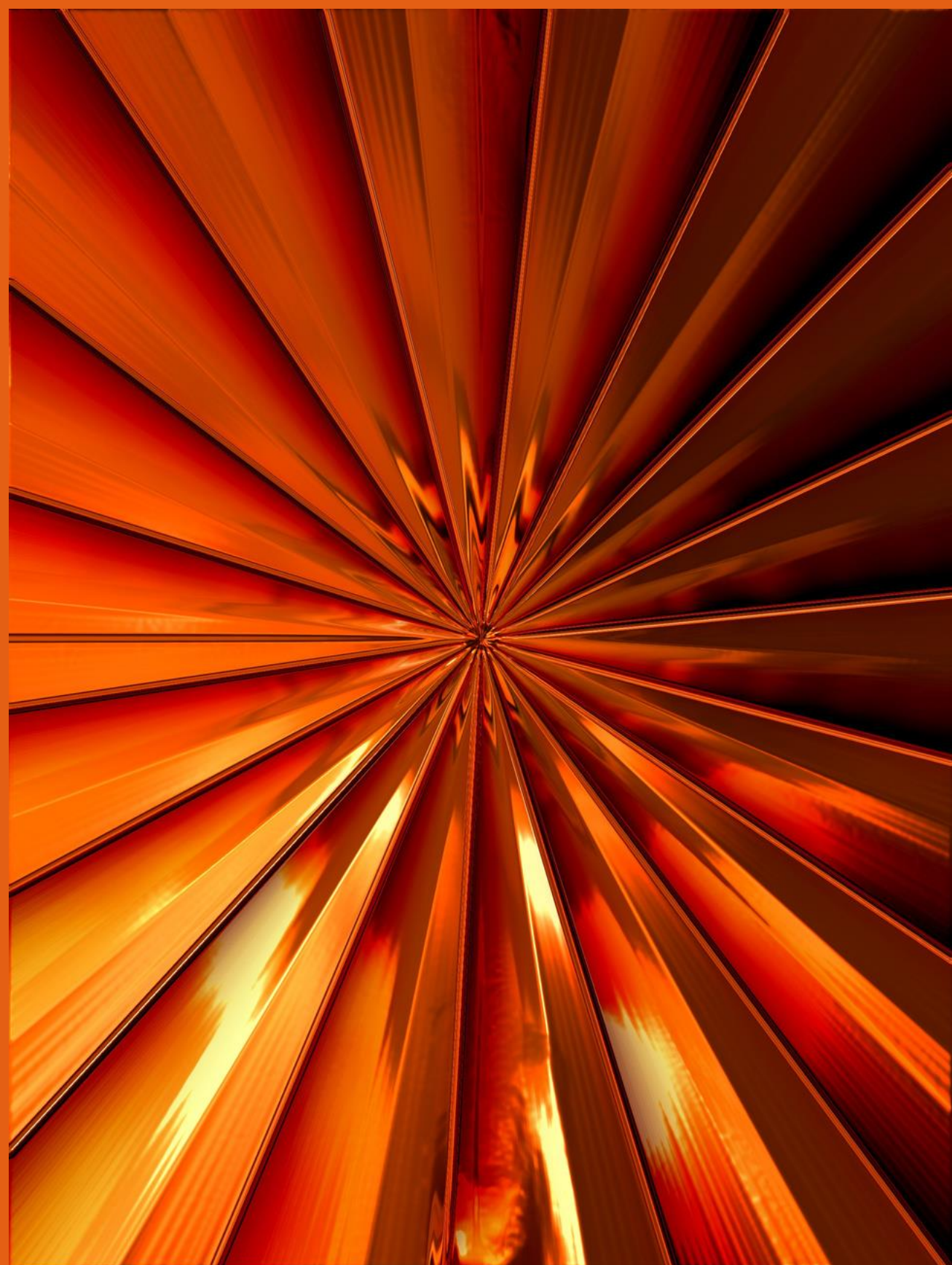
The overall F-statistic of the model is 32.15 and the corresponding p-value is 3.28e-09. This indicates that the overall model is statistically significant. In other words, the regression model as a whole is useful.

disp is statistically significant at the 0.10 significance level. In particular, the coefficient from the model output tells is that a one unit increase in disp is associated with a -0.019 unit decrease, on average, in mpg, assuming hp and drat are held constant.

hp is statistically significant at the 0.10 significance level. In particular, the coefficient from the mFrom the output we can see the following:

The overall F-statistic of the model is 32.15 and the corresponding p-value is 3.28e-09. This indicates that the overall model is statistically significant. In other words, the regression model as a whole is useful.





THANK YOU  
SO MUCH!