

PES University, Bangalore

Established under Karnataka Act No. 16 of 2013

UE21CS342AA2 - Data Analytics - Worksheet 2a - Linear and Logistic Regression
Designed by Aaditya S Goel, Dept. of CSE - aadityasgoel@gmail.com

Welcome to DATA Motors

India is poised to become the third largest economy in the world. To fuel and sustain this growth, Indian businesses are looking to increase their footprint and expand into different markets across the world.

DATA Motors is the leading automotive manufacturer in the country and they're now looking to enter the second largest auto market in the world, the United States of America.

But there's a catch! Pricing of cars in USA seems to be very different to that in India. Now DATA motors wants to enter this market with a bang and they need to get their pricing spot-on. So they have hired you, a consultant at the prestigious Bangalore Consulting Group. Now the onus is on you to understand what factors drive the pricing models of the most successful car companies currently in the market. Let's get to work!

Regression

Regression is a statistical method used to model the connection between variables, understanding how changes in one influence another. It's vital for predicting outcomes, finding patterns, and making informed decisions.

Regression is essential across diverse fields like economics and medicine due to its ability to quantify relationships and make predictions for new data. Its popularity arises from its simplicity, adaptability, and its central role in data-driven decision-making.

In this worksheet we will be exploring 3 concepts. Namely:

- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

Before we go any further, let's have a look at the dataset and it's different columns

Data Dictionary

price: price of the car in dollars
fuel_type: gas or diesel
CompanyName: name of the manufacturer
aspiration: std (standard or naturally aspirated engine) or turbo (turbocharged engine)
doornumber: number of doors in the car
carbody: type of car (sedan, wagon, hatchback, convertible, hardtop)
drivewheel: rwd (Rear-wheel drive) or fwd(front-wheel drive)
enginelocation: front or rear
wheelbase: distance between front and rear axles in inches
carlength: length of car in inches
carwidth: width of car in inches

carheight: height of car in inches
 curbweight: weight of car with a full tank and standard equipment
 cylindernumber: number of cylinders in the engine
 horsepower: power generated by the engine in horsepower (hp)
 mpg: fuel economy of car in miles per gallon

Data Visualising

Let's visualize this all in the form of a Data Frame

```
cars <- read.csv('Dataset_2a.csv')
head(cars)
```

##	price	car_ID	fueltype	CompanyName	aspiration	doornumber	carbody
## 1	13495	1	gas	alfa-romero	std	two	convertible
## 2	16500	2	gas	alfa-romero	std	two	convertible
## 3	16500	3	gas	alfa-romero	std	two	hatchback
## 4	13950	4	gas	audi	std	four	sedan
## 5	17450	5	gas	audi	std	four	sedan
## 6	15250	6	gas	audi	std	two	sedan

##	drivewheel	engine	location	wheelbase	carlength	carwidth	carheight	curbweight
## 1	fwd		front	88.6	168.8	64.1	48.8	2548
## 2	rwd		front	88.6	168.8	64.1	48.8	2548
## 3	rwd		front	94.5	171.2	65.5	52.4	2823
## 4	fwd		front	99.8	176.6	66.2	54.3	2337
## 5	fwd		front	99.4	176.6	66.4	54.3	2824
## 6	fwd		front	99.8	177.3	66.3	53.1	2507

##	cylindernumber	horsepower	mpg
## 1	four	111	27
## 2	four	111	27
## 3	six	154	26
## 4	four	102	30
## 5	five	115	22
## 6	five	110	25

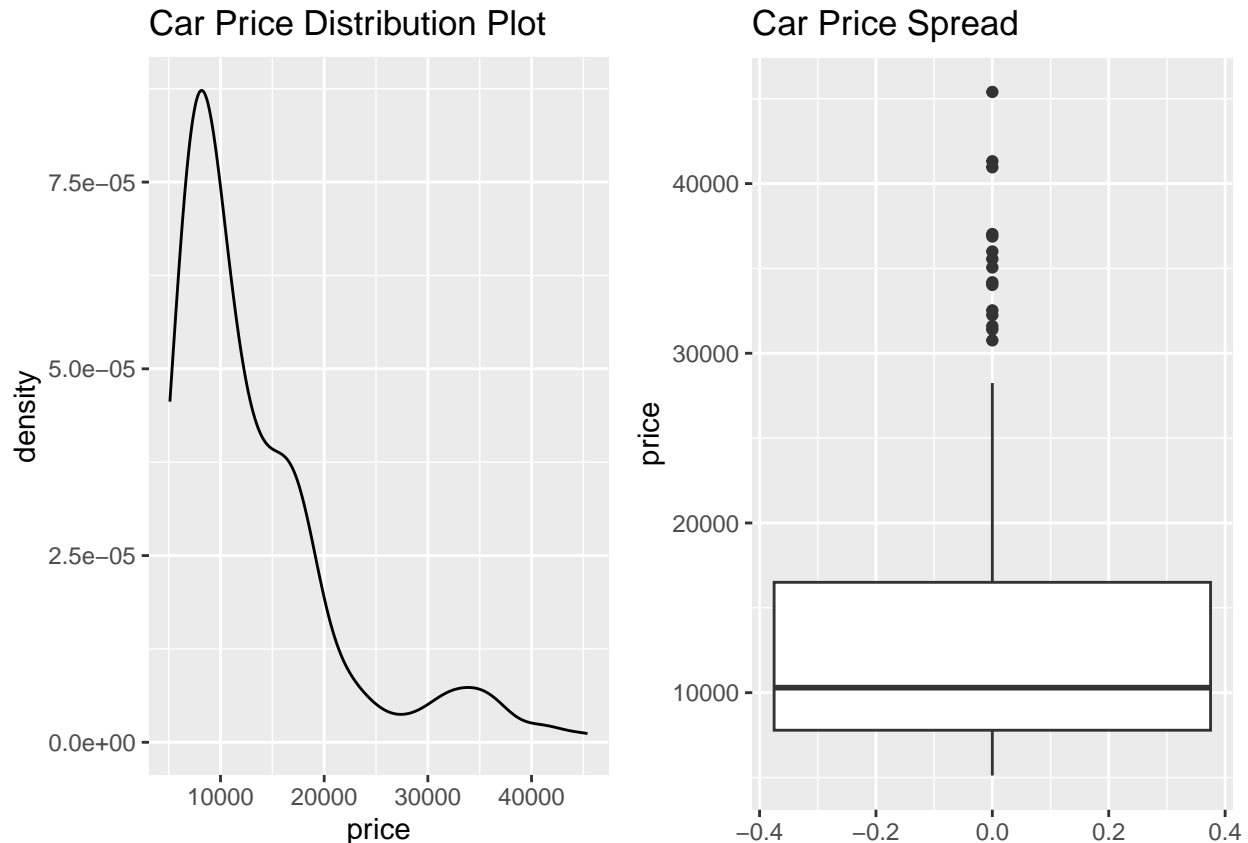
Let us plot the distribution of car prices and see what the spread looks like.

```
library(ggplot2)
library(gridExtra)

plot1 <- ggplot(cars, aes(x = price)) +
  geom_density() +
  labs(title = "Car Price Distribution Plot")

# Create the second plot (Car Price Spread)
plot2 <- ggplot(cars, aes(y = price)) +
  geom_boxplot() +
  labs(title = "Car Price Spread")

# Combine the plots and display
grid.arrange(plot1, plot2, ncol = 2)
```



We can clearly see that the prices are heavily right-skewed with some outliers. This seems to explain the exclusive, luxurious vehicles only affordable for a few.

Regression Analysis

Before proceeding to a full analysis, your client DATA Motors have some questions they want you to answer.

1. Simple Linear Regression

From experience, they have understood that the more powerful their car is, the higher they are able to price it at to the public. They want to know if this trend holds perfectly in this new market too. Have a look at the data, pick the right variables and find the if this relationship is true. Create a scatter plot between the dependent and independent variable with the best-fit line passing through. (Hint: use the ggplot library)

Write your code here

Ensure you have installed all the necessary packages required for plotting

What do you infer from your graph? The results don't seem to be very surprising. But there's something that's off about the scatter plot itself. Try plotting the residuals and analyzing if it's only white noise.

Write your code here

How will you tackle this problem? (Hint: Think about the different kind of transformations you've learnt in class)

Write you code here

2. Logistic Regression

Logistic regression is an algorithm that estimates the parameters, or coefficients, of the linear combination of the logit model. The logistic or logit model is used to predict the probability 'p' of a binary dependent variable taking on one of two possible outcomes. This feature makes Logistic Regression useful even in problems of binary classification

DATA motors currently only build vehicles with rear-wheel drive. In America however, front-wheel drive is known to be quite popular too. Development of this technology will require significant investments into Research & Development. The client wants to know if they can recover costs quickly by charging a premium on front-wheel drive vehicles.

Analyze the price at which these two types of cars are sold and try to find out if Front-wheel Drive cars are indeed the premium variety in the market, or if rear-wheel drive vehicles can fetch high rates.

```
## Write your code here
```

Is this good news or bad news for the client? As with most things, it's a bit of both. Go ahead and think about why that might be the case here.

Meanwhile let us try and see how good our logistic regression models are performing on the data. (Hint: Use the inbuilt functions in the PROC library)

```
## Write your code here
```

Those are striking numbers. What does it say about our the drivewheel variable that our Logistic Regression models are able to achieve such high scores across metrics?

3. Multiple Linear Regression

For our Multiple Linear Regression models, we could use all the attributes and try to predict the price. But the aim is to always predict the maximum variation in the target, with the minimum variables.

Thus, it's important to identify which features are most important to predict our target variable. Use the help of a correlogram to visually analyze the correlation between different independent variables and the one dependent variable. (Don't forget to keep an eye on the correlation between independent variables. Try and identify why it is important to do this.)

```
## Before constructing a correlogram, you will have to convert all categorical values to numerical values
## To achieve this, you can use the mutate function in the dplyr library
```

```
## Then proceed to create the correlogram
```

We can now see that there are features positively correlated to price, and features negatively correlated to price. Let us use all the significant variables we have noticed in the correlogram in our Multiple Linear regression model.

Use different variables to create the Multiple Linear Regression model and analyze the difference in residual values and F-statistic scores between each of them.

```
## Write your code here
```

What can you infer about the fit of Multiple Linear Regression on to the given dataset?

Which are the most important variables to predict the price of the car?

How many variables did you use in your best fitting model? Which ones were they?

Good job with the analysis! DATA motors and Bangalore Consulting Group have both picked up valuable information from the work you just did.

The methods used in this worksheet form the fundamental basis for many more complex techniques and algorithms. As internship season is upon is, those of you who get to work in Data Science, Analytics etc will find yourselves using these very same techniques to answer the business questions posed by your organizations.

In a world where ChatGPT and DALL-E get all the Spotlight, classic ML techniques like Linear Regression still form the backbone of real world Analytics. The simplicity and interpretability of these models have made these models invaluable in providing insights to business owners across industries make informed, data-driven decisions.

Happy Learning!