Final Project DSC 520

Reenie Christudass

2022-07-31

Contents

Introduction - MULTIPLE LINEAR REGRESSION

Any organization can make better decisions by using regression techniques for predictive analysis. This statistical technique can help organizations to make better decisions. Relationships between data can transform analysis into actionable information. Linear regression can fit in one dependent and one independent variable and has significant limits. In comparison, Multiple regression can overcome and fit in single dependent and multiple independent variables. This statistical techniques is used for forecasting, time series modelling and find relationship between variable.

Research questions:

Identify potential predictor variable and dependent variable

```
One dependent variable
Multiple Independent variable
```

Identify data set and import

```
Use appropriate library to read data
```

Summarize the variable (univariate analysis)

```
Identify the columns and type, number of rows, summarize to understand the min, max, Q1, Q2, Q3
```

Data transformation and cleaning

```
If the values are larger to fit to a plot, transform to log
Clean the data, remove nulls
```

Check the relationship using scatter plot and correlations

```
Create scatter plot to understand the strength of relationship
Identify the correlation, if they are strong or week
```

Check for multicollinearity

```
Check for predictor variable before fitting the model
```

Selection of predictor variable

```
Add and drop as per need
```

Fit the model

```
Use the appropriate library (lm)
```

Hypothesis Testing

```
H0 = Ha or H0 not equal to Ha
```

Visualize data

```
Create plot to show the current vs predicted values
```

Predict analysis

```
Are there any outliers?
Are there missing values? Remove or find mean ?
Handle categorical variables?
```

Dataset

Medical insurance (https://www.kaggle.com/datasets/mirichoi0218/insurance)

Does the cost of the insurance is related with age, habits (smoking vs non smoking), BMI , region etc. Predict the cost increase / decrease on the cost of the insurance?

Death rate from Cancer (https://data.world/nrippner/cancer-trials)

Who is prevailing the trails? Is it been driven by the environment conditions or wealth

Vehinle data (https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho)

Is the price of the car increase or decrease as the mileage goes up or because of accident etc? Any other factor.

Library packages

library(readxl) library(ggplot2) library(dplyr) library(caTools) library(tidyverse) library(car)

Visualize data

Scatter plot - Predicted value vs Actual value - Plot the residual vs fitted Quantile-Quantile plots -determining if two data sets come from populations with a common distribution. Histogram - to find residuals

References