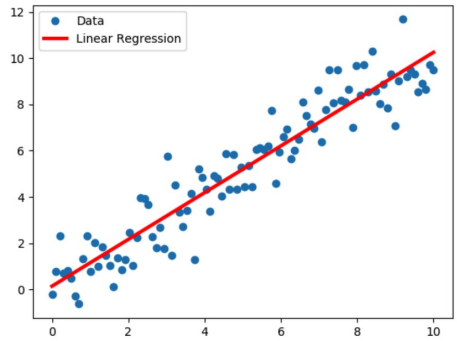
MULTIPLE LINEAR REGRESSION IN R :: CHEAT SHEET



Multiple linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response variable) and two or more independent variables (also known as predictor variables or features). It extends the concept of simple linear regression, which deals with the relationship between two variables, to a scenario where multiple predictors are considered simultaneously.

The goal is to find the best-fitting linear equation that describes how the combination of predictor variables influences the dependent variable. This equation takes the form:

$Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$

Where:

Y is the dependent variable (response).

X1,X2,...,Xp are the independent variables (predictors).

β0,β1,β2,...,βp are the regression coefficients, representing the impact of each predictor.

ϵ is the error term, accounting for unexplained variability.

Multiple linear regression is a technique used in various fields, including economics, social sciences, engineering, and natural sciences. It provides a way to analyze complex relationships and make informed decisions based on data-driven insights.

Load the necessary libraries

```
library(tidyverse)
library(tidymodels)
library(car)
library(GGally)
```

Data Preparation

Load your dataset (replace 'data.csv' with your file)

```
data <- read.csv("data.csv")
```

Check structure of the data

```
str(data)
```

Split data into predictors (X) and response (Y)

```
X <- data[, c("x1", "x2", "x3")]  # Select predictor variables
Y <- data$y
# Select response variable
```

Data Exploration

Obtain graphical and numerical summaries to describe pairwise relationships

```
data %>%
  select(Y, x1, x2,x3) %>%
  ggpairs()
```

Fit Multiple Linear Regression Model

Create a linear regression model specification

```
lm_spec <- linear_reg() %>%
  set_mode("regression") %>%
  set_engine("lm")
```

Fit the linear regression model

```
model<- lm_spec %>%
  fit(Y ~ x1 + x2 + x3, data = data)
```

Display a tidy summary of the model

```
tidy(model)
```

Model Evaluation

R-squared value (proportion of variance explained by the model)

```
rsquared <- summary(model)$r.squared
cat("R-squared:", rsquared, "\n")
```

Adjusted R-squared value (accounts for number of predictors)

```
adj_rsquared <- summary(model)$adj.r.squared
cat("Adjusted R-squared:", adj_rsquared, "\n")
```

Coefficient Interpretation

Display coefficients with their standard errors and p-values

```
coef_summary <- summary(model)$coefficients
print(coef_summary)
```

Predictions

Make predictions using the model

```
new_data <- data.frame(x1 = c(10, 20, 30), x2 = c(5, 8, 12), x3 = c(2, 3, 4))
predictions <- predict(model, newdata = new_data)
print(predictions)
```

Residual Analysis

Residual plot

```
plot(model, which = 1)
```

Normality Q-Q plot

```
plot(model, which = 2)
```

Scale-location plot

```
plot(model, which = 3)
```

Model Assumptions

Test for multicollinearity (Variance Inflation Factor)

```
vif(model)
```

Test for heteroscedasticity (Breusch-Pagan test)

```
vif(model)
```

Test for normality of residuals (Shapiro-Wilk test)

```
shapiro.test(residuals(model))
```