STUTERN

Regression (Multivariate)

SGA07_DATASCI

13th February 2020

Module Overview

- Multiple Regression Model
- Overfitting & Multicollinearity
- Scatterplot & Correlation Matrix
- Metrics: p-value, Standard error, R square, R square (predicted)
- Dummy Variables for Categorical Attributes

Book Keeping

- Resources for API Development
- Upload of module slides along with video & audio
- Morning challenge as regards SQL scripting
- 50% of module covered 💪💪💪

Outcome

After this Module, you will;

- Understand how to extend the linear model for multiple independent variables
- Review some concepts that may cause modelling errors such as overfitting and multicollinearity
- Review how to use visualisation (scatterplot) and statistics (correlation) to build intuition on attribute relationships
- Overview of performance metrics (such as p-value, standard error, r square) for model selection
- Understand how to engineer dummy variables used as features to replace nominal and ordinal attributes in a multiple regression model

Linear Regression (Formula)

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

$\hat{y} =$ Expected value of dependent variable
$x =$ Independent variable
$\beta_1 =$ Slope of line
$\beta_0 =$ y-intercept given by $x = 0$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$x_i =$ Observed value of independent variable
$\bar{x} =$ Mean value of independent variable
$y_i =$ Observed value of dependent variable
$\bar{y} =$ Mean value of dependent variable

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Multivariate Regression Model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

$\beta_n =$ Estimated change in y to a one unit change in x_n, when all other x_i are held constant

> This is an extension of the linear regression model that considers more than one independent variable as its input for prediction.

STUTERN

GoKada Logistics

Let's assume that following Lagos State ban on Okada & Keke, GoKada now starts a logistic delivery business given their existing infrastructure. They want you to build a model that predicts hours a day's delivery operation will take given data on distance covered, fuel price and number of delivery for some past days operations

GoKada Data

Day_ID	Distance Covered	Fuel Price (thousand	No. Of Deliveries	Hours Travelled (Hrs)
1	59	3.84	4	7
2	66	3.19	1	5.4
3	78	3.78	3	6.6
4	111	3.89	6	7.4
5	44	3.57	1	4.8
6	77	3.57	3	6.4
7	80	3.03	3	7
8	66	3.51	2	?
9	109	3.54	5	?
10	76	3.25	3	?

Overfitting

- Contains more parameters than can be justified by the data
- The model has learned the noise instead of the signal
- Always split your data into train and test sets
- Check relationship between each independent and dependent variables

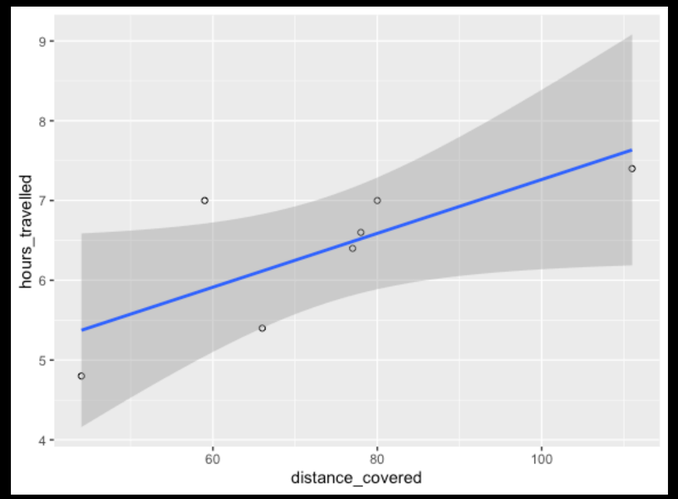
> The production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably

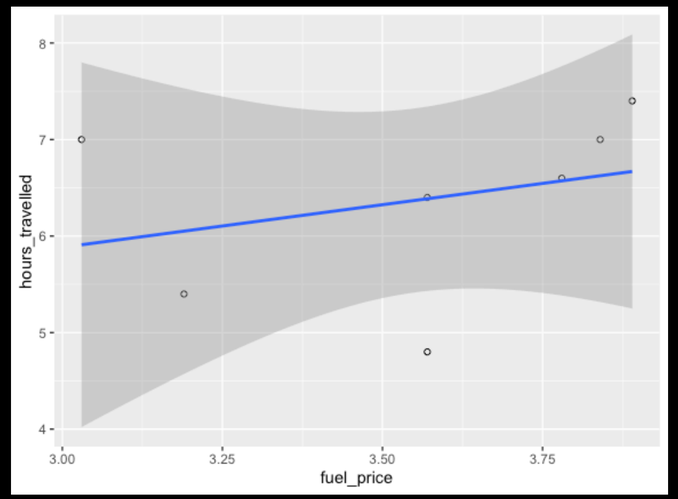
Multicollinearity

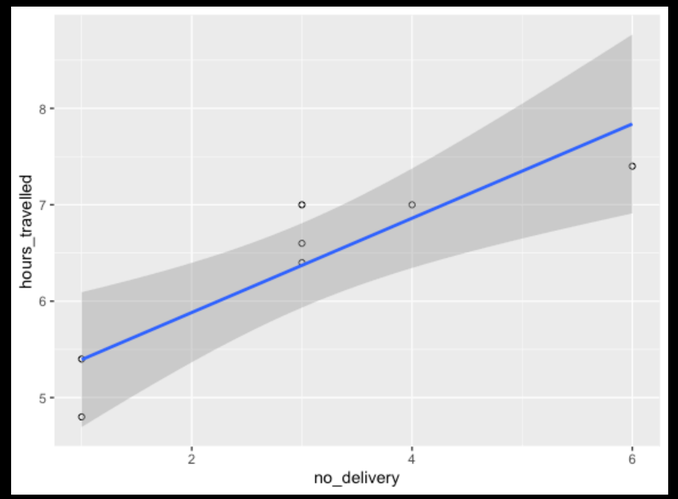
- Independent variables should be independent.
- Reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model
- Check relationship between each independent variables

> A phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy

Scatterplot Visualisation



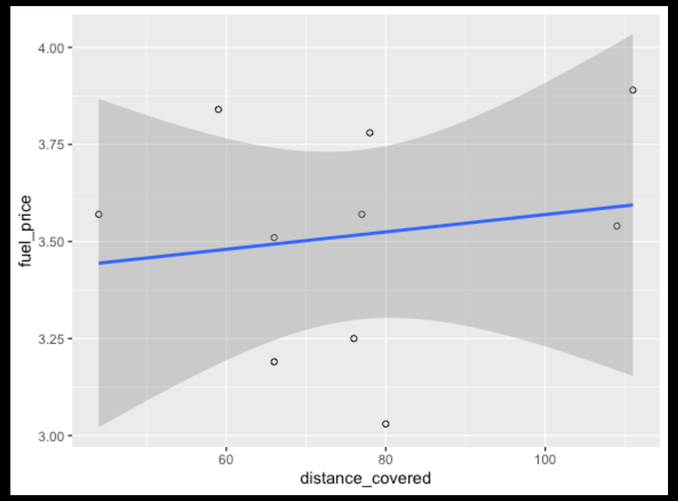


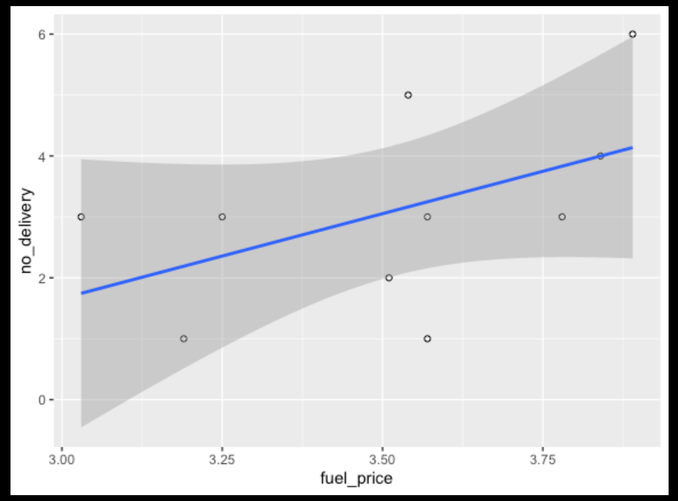


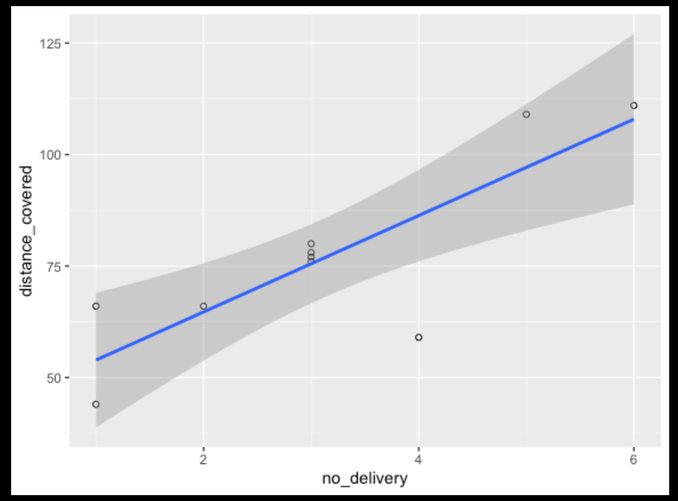
Test for Overfitting

STUTERN

Scatterplot Visualisation





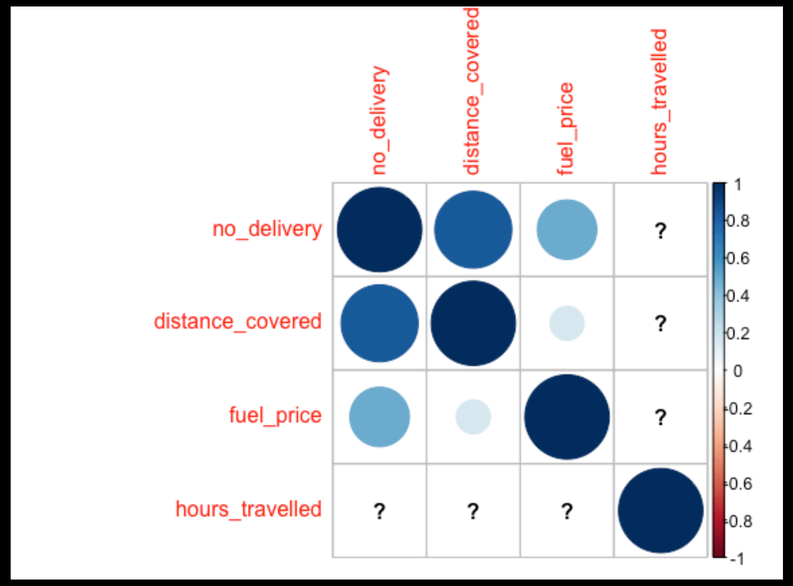


Test for Mulitcollinearity

STUTERN

Correlation Analysis

	no_delivery	distance_covered	fuel_price	hours_travelled
no_delivery	1.0000000	0.8338111	0.4982422	NA
distance_covered	0.8338111	1.0000000	0.1618352	NA
fuel_price	0.4982422	0.1618352	1.0000000	NA
hours_travelled	NA	NA	NA	1



STUTERN

Multiple Regression Performance

Predictor	R squared	R squared (Adj)	p-value	SE	F-stat	R square (diff)
Distance	0.56	0.47	0.052	0.68	6.381	0.09
Delivery	0.81	0.77	0.005	0.45	21.25	0.04
Fuel	0.10	-0.09	0.500	0.98	0.53	0.19
Dist + Del	0.81	0.72	0.035	0.49	8.70	0.09
Dist + Fuel	0.58	0.37	0.171	0.74	2.83	0.21
Del + Fuel	0.88	0.82	0.015	0.40	14.36	0.06
Dist + Del + Fuel	0.89	0.77	0.063	0.45	7.75	0.12

Multiple Regression (Qualitative)

- Feature engineering of dummy variables
- Works like a coding system to transform qualitative variables into a series of variables which can then be entered into the regression model
- Avoid dummy variable trap by encoding into just k-1 variables for qualitative attribute with k-levels

> A dummy variable (aka, an indicator variable) is a numeric variable that represents categorical data, such as gender, race, political affiliation, etc.

STUTERN

Dummy Variable (Nominal)

North = 0
West = 1
South = 2
East = 3

Day_ID	Distance Covered (KM)	Fuel Price (thousand Naira)	No. Of Deliveries	Hours Travelled (Hrs)	Region
1	59	3.84	4	7	East
2	66	3.19	1	5.4	West
3	78	3.78	3	6.6	East
4	111	3.89	6	7.4	North
5	44	3.57	1	4.8	South
6	77	3.57	3	6.4	North
7	80	3.03	3	7	East
8	66	3.51	2	?	North
9	109	3.54	5	?	West
10	76	3.25	3	?	South

Day_ID	Distance Covered (KM)	Fuel Price (thousand Naira)	No. Of Deliveries	Hours Travelled (Hrs)	Region
1	59	3.84	4	7	3
2	66	3.19	1	5.4	1
3	78	3.78	3	6.6	3
4	111	3.89	6	7.4	0
5	44	3.57	1	4.8	2
6	77	3.57	3	6.4	0
7	80	3.03	3	7	3
8	66	3.51	2	?	0
9	109	3.54	5	?	1
10	76	3.25	3	?	2

STUTERN

Dummy Variable (Nominal)

Day_ID	Distance Covered (KM)	Fuel Price (thousand Naira)	No. Of Deliveries	Hours Travelled (Hrs)	Region
1	59	3.84	4	7	3
2	66	3.19	1	5.4	1
3	78	3.78	3	6.6	3
4	111	3.89	6	7.4	0
5	44	3.57	1	4.8	2
6	77	3.57	3	6.4	0
7	80	3.03	3	7	3
8	66	3.51	2	?	0
9	109	3.54	5	?	1
10	76	3.25	3	?	2

Day_ID	Distance Covered (KM)	Fuel Price (thousand Naira)	No. Of Deliveries	Hours Travelled (Hrs)	Region (North)	Region (West)	Region (South)	Region (East)
1	59	3.84	4	7	0	0	0	1
2	66	3.19	1	5.4	0	1	0	0
3	78	3.78	3	6.6	0	0	0	1
4	111	3.89	6	7.4	1	0	0	0
5	44	3.57	1	4.8	0	0	1	0
6	77	3.57	3	6.4	1	0	0	0
7	80	3.03	3	7	0	0	0	1
8	66	3.51	2	?	1	0	0	0
9	109	3.54	5	?	0	1	0	0
10	76	3.25	3	?	0	0	1	0

STUTERN

Dummy Variable (Ordinal)

Small = 0
Medium = 1
Large = 2

Day_ID	Distance Covered (KM)	Fuel Price (thousand Naira)	No. Of Deliveries	Hours Travelled (Hrs)	Package Size
1	59	3.84	4	7	Small
2	66	3.19	1	5.4	Medium
3	78	3.78	3	6.6	Medium
4	111	3.89	6	7.4	Large
5	44	3.57	1	4.8	Large
6	77	3.57	3	6.4	Small
7	80	3.03	3	7	Medium
8	66	3.51	2	?	Small
9	109	3.54	5	?	Large
10	76	3.25	3	?	Small

Day_ID	Distance Covered (KM)	Fuel Price (thousand Naira)	No. Of Deliveries	Hours Travelled (Hrs)	Package Size
1	59	3.84	4	7	0
2	66	3.19	1	5.4	1
3	78	3.78	3	6.6	1
4	111	3.89	6	7.4	2
5	44	3.57	1	4.8	2
6	77	3.57	3	6.4	0
7	80	3.03	3	7	1
8	66	3.51	2	?	0
9	109	3.54	5	?	2
10	76	3.25	3	?	1

Practice Lab

Build a predictive multiple regression model using R

Use the following Instructions:

- Use the mtcar data in r with ***mpg*** as the dependent variable
- Divide into train (70%) and test (30%)
- Explore the data (Univariate & Bivariate)
- Build a multiple regression model
- Apply linear model to test data to validate model

Recap/Summary

At the end of this Module, you should understand;

- Understand how to extend the linear model for multiple independent variables
- Review some concepts that may cause modelling errors such as overfitting and multicollinearity
- Review how to use visualisation (scatterplot) and statistics (correlation) to build intuition on attribute relationships
- Overview of performance metrics (such as p-value, standard error, r square) for model selection
- Understand how to engineer dummy variables used as features to replace nominal and ordinal attributes in a multiple regression model

Suggested Material

- O'Reilly Doing Data Science by Carthy O'Neil and Rachel Schutt Pages 55 - 71
- https://www.youtube.com/playlist?list=PLleGtxpvyG-lqjoU8liF0Yu1WtxNq_4z-
- https://elitedatascience.com/overfitting-in-machine-learning
- https://en.wikipedia.org/wiki/Overfitting
- https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/
- https://en.wikipedia.org/wiki/Multicollinearity
- https://rpubs.com/davoodastaraky/mtRegression

Suggested Material (Logistic Regression)

- O'Reilly Doing Data Science by Carthy O'Neil and Rachel Schutt Pages Chapter 5 (113 - 129)

Suggested Material (API Development)

- https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-apis-application-programming-interfaces-5-apis-a-data-scientist-must-know/
- https://medium.com/better-practices/api-driven-analytics-d980b28cb15e
- https://towardsdatascience.com/deploying-a-machine-learning-model-as-a-rest-api-4a03b865c166