Machine Learning – I

Assumptions of Regression and Model Evaluation

Agenda

- ○ Assumptions of Linear Regression

- ○ Tests for Assumptions of Linear Regression

- ○ Model Evaluation Metrics

- ○ Presence of Categorical Variables

- ○ Interaction Effect

Assumptions of Linear Regression

Assumptions of linear regression

- The dependent variable must be numeric

- Linear relationship between dependent and independent variables

- Predictors must not show multicollinearity

- Independence of observations should exist (Absence of Autocorrelation)

- The error terms should be homoscedastic

- The error terms must follow normal distribution

Assumptions of linear regression

- Numeric dependent variable

- Absence of Multicollinearity

Assumption to be checked before building a model

Build a model

Assumption to be checked after building a model

- Linear relationship

- Absence of Autocorrelation

- Error terms should be homoscedastic

- Error terms must follow $N(0,\sigma^2)$

Tests for Assumptions of Linear Regression

Tests before model building

- The dependent variable must be numeric

- Predictors must not show multicollinearity

Is the dependent variable numeric?

- Regression Analysis requires the target variable to be numeric in nature

- For example: returns, sales of a product, yield of a crop, risk in financial services

- In context with our example, we see that Premium is numeric

Mileage	Premium (in dollars)
15	392.5
14	46.2
17	15.7
7	422.2
10	119.4
7	170.9
20	56.9
21	77.5
18	214
11	65.3
7.9	250
8.6	220
12.3	217.5
17.1	140.88
19.4	97.25

Q & A

Question:

If our target variable is: 0, 1, 1, 0, 0, where 0 indicates presence of a disease and 1 indicates absence.

Is it appropriate to use regression to find the whether the person has a disease?

Q & A

Question:
If our target variable is: 0, 1, 1, 0, 0, where 0 indicates presence of a disease and 1 indicates absence. Is it appropriate to use regression to find the whether the person has a disease?

Answer:
No. Because the target variable is a categorical variable. Thus, it is a classification problem.

Tests before model building

- The dependent variable must be numeric

- Predictors must not show multicollinearity

What is multicollinearity?

- Multicollinearity arises when the independent variables have high correlation among each other

- Multicollinearity may be introduced if there exists empirical relationship among variables such as income = expenditure + saving

- In presence of it, the best fit line obtained from OLS method is no more "best"

- Also, the confidence interval obtained for β's is wider since the SE(β) becomes large

Multicollinearity detection

- Determinant of correlation mat[rix]

- Condition Number (CN)

Is there multicollinearity present?

- Correlation matrix

- Variance Inflation Number (VIF)

Which variables are involved in multicollinearity?

Is there multicollinearity?

Determinant of the correlation matrix:

Let D be the determinant of correlation matrix. Then 0 < D < 1
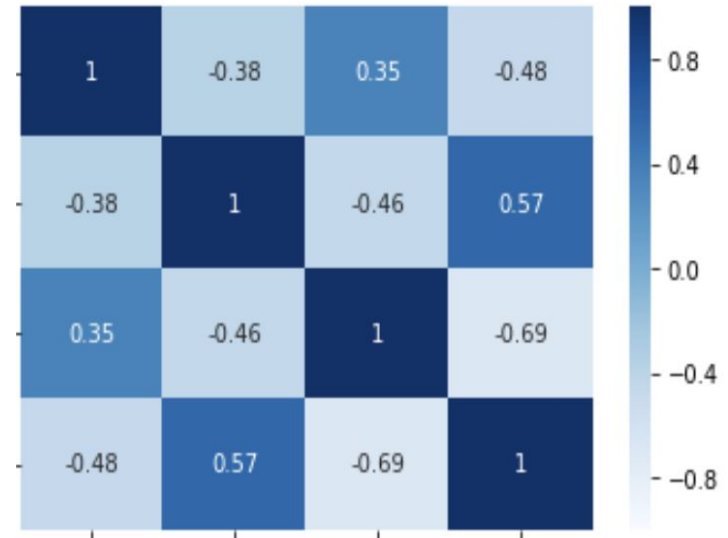
D=0	High multicollinearity
D=1	No multicollinearity

Condition Number (CN):

CN > 1000	Severe multicollinearity
100 < CN < 1000	Moderate multicollinearity
100 < CN	No multicollinearity

Which variables are involved in multicollinearity?

Correlation matrix:

If the off-diagonal values tend to ±1 then it indicates high correlation between the variable pair. However this inspection is not enough

Which variables are involved in multicollinearity?

Variance Inflation Factor (VIF):

$$VIF = \frac{1}{1-R^2}$$

Where R^2 is obtained by regressing a predictor variable over all the other predictors in the model

Value	Interpretation
VIF > 5	High correlation
5 > VIF >1	Moderate correlation
VIF = 1	No correlation

Tests after model building

- Linear relationship between dependent and independent variables

- Independence of observations should exist (i.e. Absence of Autocorrelation)

- The error terms should be homoscedastic

- The error terms must follow normal distribution

Tests after model building

- Linear relationship between dependent and independent variables

- Independence of observations should exist (i.e. Absence of Autocorrelation)

- The error terms should be homoscedastic

- The error terms must follow normal distribution

Assumption of linearity

- An assumption of linear regression is that it should be linear in the parameter

- The independent variables must have a linear relationship with the dependent variable

- The residuals and the fitted values should be independent

Existence of linear relationship

An assumption of linear regression is that it should be linear in the parameter

Linear Relationship

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \varepsilon$$

$$y = \beta_0 - \beta_1 \log(x_1) + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 - \beta_3 x_1 x_2 + \varepsilon$$

Nonlinear Relationship

$$y = \beta_0 - e^{\beta_1 x_1} + \varepsilon$$
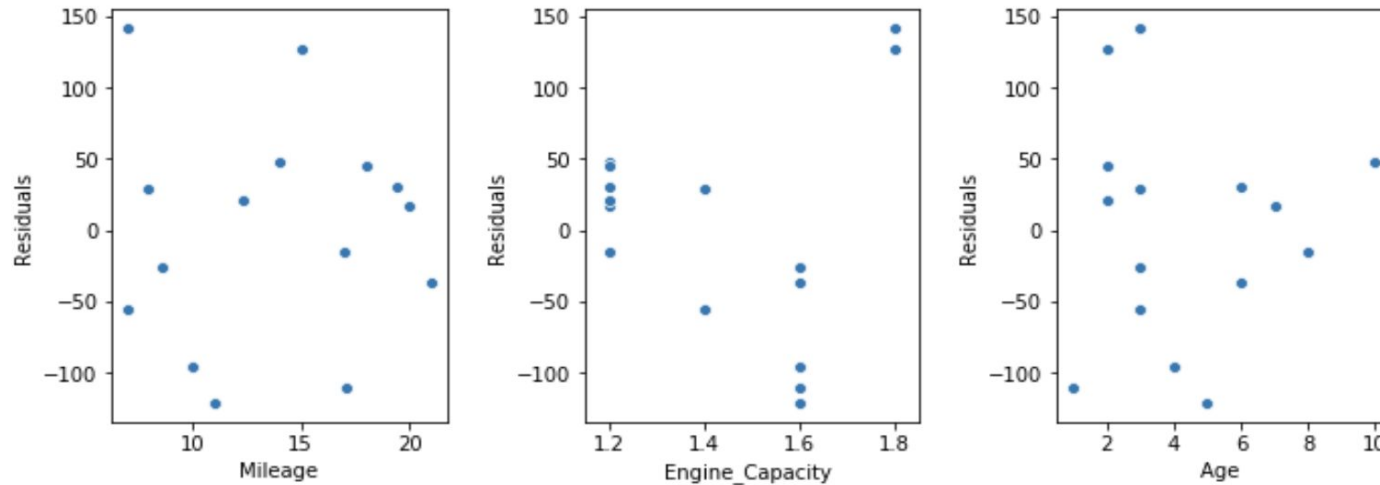
$$y = \beta_0 x_1 / \beta_1 x_1 + \varepsilon$$

$$y = \beta_0 + x_1^{\beta_1} . x_2^{\beta_2} + \varepsilon$$

Existence of linear relationship

- The independent variables must have a linear relationship with the dependent variable

- This can be checked plotting a scatter plot of residuals vs predictors

- A scatter plot depicting no pattern indicates that the variable has a linear relationship with the response variable
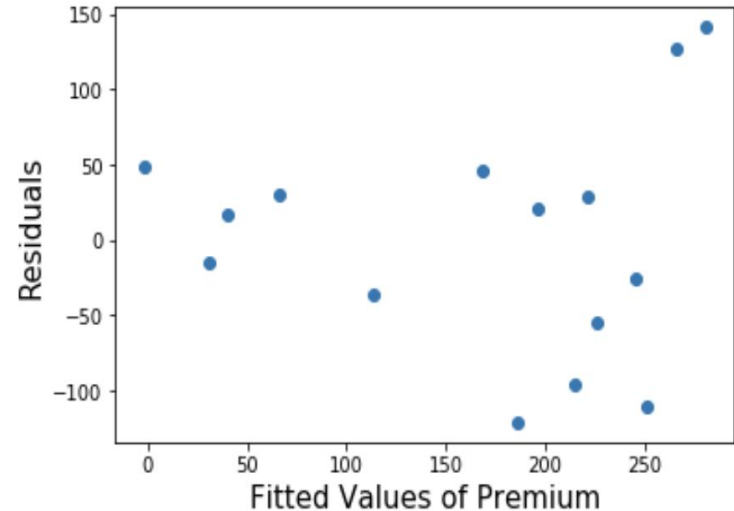
Existence of linear relationship

In context with our data, we see a random pattern in all the three plots. Hence, we may say that, the predictors are linearly related with the response variable.

Existence of linear relationship

- The plot of residuals against the fitted tells the presence of linear relationship

- For linear relationship, the points must be at random, i.e., it should not exhibit much distinctive pattern, no non-linear trends or changes in variability

Tests after model building

- Linear relationship between dependent and independent variables

- Independence of observations should exist (i.e. Absence of Autocorrelation)

- The error terms should be homoscedastic

- The error terms must follow normal distribution

Assumption of autocorrelation

$$\left\{ Auto \right\} correlation$$

self

- Assumption of autocorrelation is violated when residuals are correlated within themselves, i.e. they are serially correlated

- Autocorrelation does not impact the regression coefficients but the associated standard errors are reduced

- This reduction in standard error leads to a reduction in associated p-value

- It incorrectly concludes that a predictor is statistically significant

Causes of autocorrelation

- Some important variables are not considered in the data

- If the relationship between the target and predictor variables is non-linear and is incorrectly considered linear

- Presence of carry over effect

  Example: The additional expenses from the budget for last month are carried over in creating the budget for next month

Durbin - Watson Test

- To test whether the error terms are autocorrelated, we Durbin-Watson test

- We test whether autocorrelation is present or not

- The hypothesis is given by:

H_0:The error terms are not autocorrelated

against

H_1:The error terms are autocorrelated

- Failing to reject H_0, will imply that the error terms are autocorrelated

Durbin - Watson test

The test statistic is given by

$$d = \frac{\sum \hat{e}_t - \hat{e}_{t-1}}{\sum \hat{e}_t^2}$$

Residual of tth observation

Residual of t-1th observation

d ∈ [0,4]

Value	Interpretation
0 < d < 2	Positive autocorrelation
d = 2	No autocorrelation
2 < d < 4	Negative autocorrelation

Tests after model building

- Linear relationship between dependent and independent variables

- Independence of observations should exist (i.e. Absence of Autocorrelation)

- The error terms should be homoscedastic

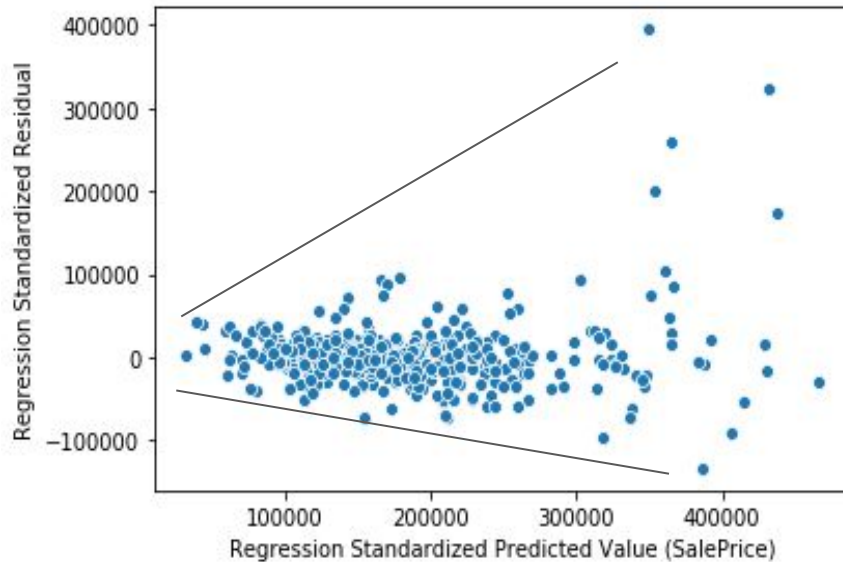- The error terms must follow normal distribution

Homoscedasticity assumption
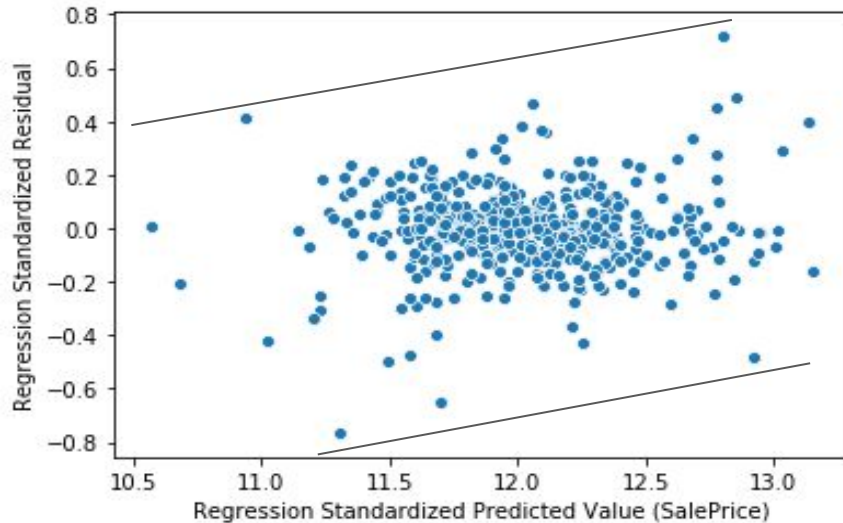
Homoscedasticity

Same   Variance

- Variance of the residual is assumed to be independent of the explanatory

  variables

- Heteroscedasticity: non-constant variance of residuals

- It happens due to the presence of extreme values

Heteroscedasticity

- Funnel type shape is seen in the graph

- Hence we can say that there is a presence of "Heteroscedasticity"

Homoscedasticity

- There is no visible funnel or bow type pattern in the plot

- We can see presence of "Homoscedasticity"

Residual vs Fitted plot

- The plot of residuals against the fitted values tells whether the error terms have equal variance

- It should look random, i.e., it should not exhibit much distinctive pattern, no non-linear trends or changes in variability
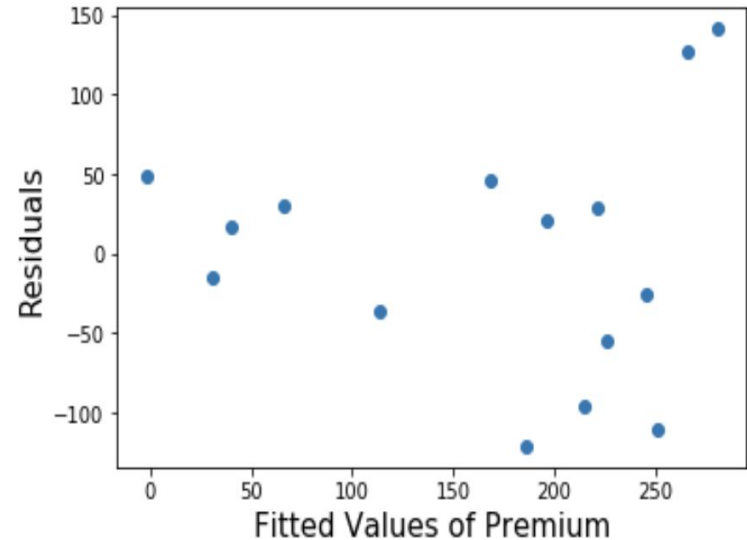
Homoscedasticity

The statistical test to test for the homoskedasticity of the errors are

- Goldfeld Quandt test

- Breusch Pagan test

Goldfeld-Quandt test

- For presence of a constant variance of error terms, i.e. to test

H_0: The errors terms are homoskedastic **against** H_1: The errors terms are heteroskedastic

- Decision rule: Reject H_0, if the p-value associated with test statistic is less than α (level of significance), which implies that there is heteroskedastic, i.e. the error terms have do not equal variance

Breusch Pagan test

- For presence of a constant variance of error terms, i.e. to test

H_0: The errors terms are homoskedastic

against

H_1: The errors terms are heteroskedastic

- Decision rule: Reject H_0, if the p-value associated with test statistic is less than α (level of significance), which implies that there is heteroskedastic, i.e. the error terms have do not equal variance

Tests after model building

- Linear relationship between dependent and independent variables

- Independence of observations should exist (i.e. Absence of Autocorrelation)

- The error terms should be homoscedastic

- The error terms must follow normal distribution

Normality test

- Parametric statistical methods assume that the underlying data has a normal distribution

- Normality tests are used to determine if a data set is well-modeled by a normal distribution

Normality testing techniques

- Quantile-Quantile Plot

- Jarque-Bera (JB) Test

- Shapiro-Wilk Test

Quantile-Quantile Plot (QQ plot)

- Used to determine whether two datasets follow the same distribution

- The quantiles of two datasets are plotted against each other

- A reference line is plotted at 45^0

- If the points lie on the reference line we conclude they follow the same distribution

Normal QQ plot
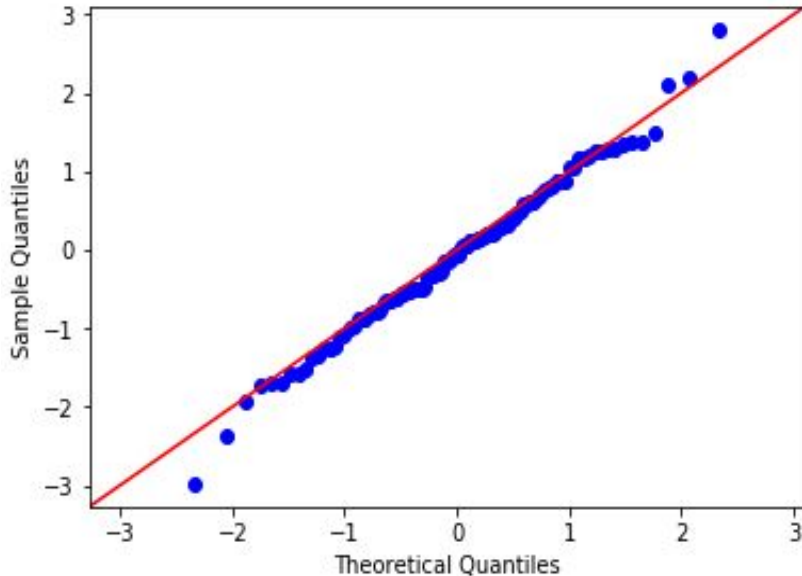
- The x axis has points from a theoretically calculated normal distribution

- They are compared with sample data on the y axis

- If the sample data has a normal distribution the points lie on the reference line

Normal QQ plot
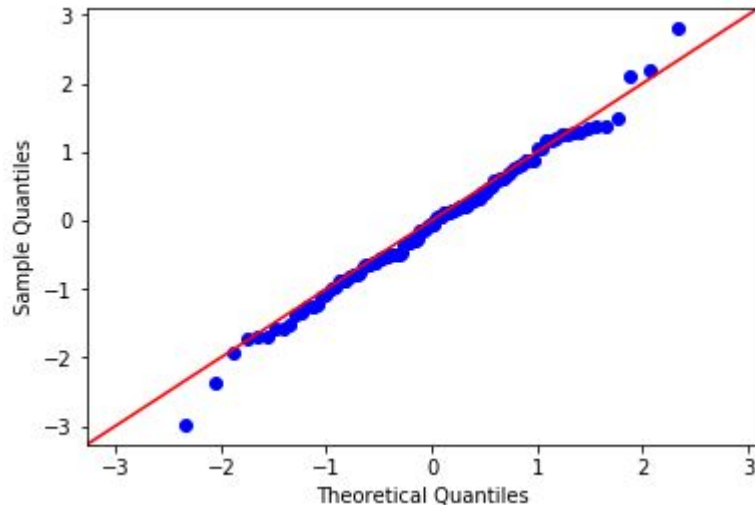
- The x axis has points from a theoretically calculated normal distribution

- They are compared with sample data on the y axis

- If the sample data has a normal distribution the points lie on the reference line

JB test

- To test whether the data follows normal distribution, we test whether the skewness and kurtosis of the data are same as that of the normal distribution, i.e. to test

  H_0: Skewness (S) = 0 and Kurtosis (K) = 0

  against

  H_1: Skewness (S) ≠ 0 and Kurtosis (K) ≠ 0

- Failing to reject H_0, implies that the data does not follow normal distribution

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K - 3)^2\right)$$

JB test

- The test statistics for n observations is given by

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

Kurtosis coefficient

Sample skewness

- The test statistics asymptotically follows chi squared distribution with 2 degrees of freedom (χ^2_2)

Shapiro-Wilk test

- To test whether the data follows normal distribution, i.e. to test

$$H_0: \text{ The data is normally distributed}$$

against

$$H_1: \text{ The data is not normally distributed}$$

- Failing to reject H_0, implies that the data does not follow normal distribution

Shapiro-Wilk test statistic

The test statistic is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

n = sample size

a_i = values computed from n samples (of size n each) from normal distribution based on their means, covariance matrix

x_i = i^{th} ordered sample values

$\bar{x}$ = sample mean

Model Evaluation Metrics

Model evaluation metrics

The model evaluation metrics are

- R^2

- Adjusted R^2

- The F test for overall significance

R-squared

- The R^2 value gives the percentage of variation in the response variable explained by the predictor variables

- If the values of R^2 = 0.87, it implies that 87% of variation in the response variable is explained by the predictor variables

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{SSR}}{\text{SST}}$$

Adjusted R-squared

- Adjusted R^2 gives the percentage of variation explained by independent variables that actually affect the dependent variable

- If the values of R^2 = 0.87, it implies that 87% of variation in the response variable is explained by the predictor variables

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

F test

- To check the significance of the regression model we use the F test

- It is similar to ANOVA for regression

- The test hypothesis is given by

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ = \beta_n = \ 0$$

against

$$H_1 : \beta_i > 0 \ \ or \ \ \beta_i < 0$$

for at least one of the i values

- Failing to reject H_0, implies that the model is not significant

F statistic

- The test statistics is given by

$$Fstat = \frac{(\text{SST-SSE})/k}{\text{SSE}/(n-k-1)}$$

$$n = \text{sample size}$$

$$k = \text{number of predictor variables}$$

- Decision rule: Reject H_0, if $F_0 > F_{(k,n-k-1),\alpha}$ or if the p-value is less than the α (level of significance)

Presence of categorical variable

Linear regression of categorical variable

- The regression method fails in presence of categorical variable

- Thus we need to convert the categorical variable to numeric variable

- In order to so, we use N - 1 dummy encoding

N-1 dummy encoding

- Dummy variables are binary variables used to represent categorical data

- For a categorical variable that can take k values k-1 dummy variables need to be created

- Dummy variable is assigned 1 if it takes a particular value else it is assigned 0

Dummy variable example

Consider a variable, Gender, used to represent the gender of a citizen during the census

Gender: Male, Female

Since Gender takes 2 values it can be represented with 1 dummy variable D_1 as:

Value	D_1
Male	0
Female	1

Data

Let us consider a categorical variable Manufacturer in the data and find out how it behaves.

Mileage	Manufacturer	Premium (in dollars)
15	Ford	392.5
14	Honda	46.2
17	Tata	15.7
7	Ford	422.2
10	Ford	119.4
7	Tata	170.9
20	Tata	56.9
21	Honda	77.5
18	Honda	214
11	Tata	65.3
7.9	Ford	250
8.6	Tata	220
12.3	Tata	217.5
17.1	Ford	140.88
19.4	Honda	97.25

Example

- In context with our example, the categorical variable Manufacturer takes values Ford, Honda and Tata

- Since Manufacturer takes 3 values, two dummy variables Mfr_Honda and Mfr_Tata are created

Value	Mfr_Honda	Mfr_Tata
Ford	0	0
Honda	1	0
Tata	0	1

Model with categorical variable

Now our model is

$$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage} + \beta_2 \text{ Mfr\_Honda} + \beta_3 \text{ Mfr\_Tata} + \varepsilon$$

Parameter	Description
β_0	Premium value where the best fit line cuts the Y-axis (Premium)
β_1	Regression coefficient of the variable Mileage
β_2	Regression coefficient of the dummy variable Mfr_Honda
β_3	Regression coefficient of the dummy variable Mfr_Tata

Linear regression model (dummy variable)

Mileage	Manufacturer	Premium (in dollars)
15	Ford	392.5
14	Honda	46.2
17	Tata	15.7
7	Ford	422.2
10	Ford	119.4
7	Tata	170.9
20	Tata	56.9
21	Honda	77.5
18	Honda	214
11	Tata	65.3
7.9	Ford	250
8.6	Tata	220
12.3	Tata	217.5
17.1	Ford	140.88
19.4	Honda	97.25

Based on the data, the β parameters are:

$\beta_0 = 368.93$, $\beta_1 = -9.117$,

$\beta_2 = -95.174$ and $\beta_3 = -129.216$

Thus the model is

$Y = 368.93 - 9.117\ x_1 - 95.174\ x_2 - 129.216\ x_3$

That is,

Premium = 368.93 - 9.117 Mileage - 95.174 Mfr_Honda - 129.216 Mfr_Tata

Regression line (dummy variable)

The regression line:

$$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage} + \beta_2 \text{ Mfr\_Honda} + \beta_3 \text{ Mfr\_Tata} + \varepsilon$$

If the manufacturer is Honda, the regression line becomes:

$$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage} + \beta_2 \text{ Mfr\_Honda} + \beta_3 \text{ Mfr\_Tata}$$
$$= \beta_0 + \beta_1 \text{ Mileage} + \beta_2 (1) + \beta_3 (0)$$
$$= \beta_0 + \beta_1 \text{ Mileage} + \beta_2 + 0$$
$$= (\beta_0 + \beta_2) + \beta_1 \text{ Mileage}$$

Value	Mfr_Honda	Mfr_Tata
Ford	0	0
Honda	1	0
Tata	0	1

Note the change in the intercept value.

Regression line (dummy variable)

The regression line:

$$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage} + \beta_2 \text{ Mfr\_Honda} + \beta_3 \text{ Mfr\_Tata} + \varepsilon$$

Value	Mfr_Honda	Mfr_Tata
Ford	0	0
Honda	1	0
Tata	0	1

For manufacturer = Ford,

$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage}$

Actual intercept

For manufacturer = Honda,

$\text{Premium} = (\beta_0 + \beta_2) + \beta_1 \text{ Mileage}$

Change in intercept

For manufacturer = Tata,

$\text{Premium} = (\beta_0 + \beta_3) + \beta_1 \text{ Mileage}$

Change in intercept

Interaction Effect

Interaction effect

Sentiment

Interaction

- An interaction effect occurs when the effect of one variable depends on another variable. This combined effect may or may not improve the performance of the model

- Note: It does not imply that the predictor variables are collinear

  Example: Salary of an employee increases with experience, but this may vary based whether the person has completed additional courses like MBA

Interaction Effect

- In context with our example, we shall consider the interaction effect of variables Engine_Capacity and Mileage

- We obtained Int_EC_Mil by taking the product of Mileage and Engine_Capacity

- Let us check whether the interaction term is adding value to our model

Mileage	Engine_Capacity	Int_EC_Mil	Age	Premium (in dollars)
15	1.8	27	2	392.5
14	1.2	16.8	10	46.2
17	1.2	20.4	8	15.7
7	1.8	12.6	3	422.2
10	1.6	16	4	119.4
7	1.4	9.8	3	170.9
20	1.2	24	7	56.9
21	1.6	33.6	6	77.5
18	1.2	21.6	2	214
11	1.6	17.6	5	65.3
7.9	1.4	11.06	3	250
8.6	1.6	13.76	3	220
12.3	1.2	14.76	2	217.5
17.1	1.6	27.36	1	140.88
19.4	1.2	23.28	6	97.25

Interaction Effect

Now our model is

$$\text{Premium} = \beta_0 + \beta_1 \text{ Mileage} + \beta_2 \text{ Engine\_Capacity} + \beta_3 \text{Age} + \beta_4 \text{ Int\_EC\_Mil} + \varepsilon$$

Parameter	Description
β_0	Premium value where the best fit line cuts the Y-axis (Premium)
β_1	Regression coefficient of the variable Mileage
β_2	Regression coefficient of the variable Engine_Capacity
β_3	Regression coefficient of the variable Age
β_4	Regression coefficient of the variable Int_EC_Mil

Mileage	Engine_Capacity	Int_EC_Mil	Age	Premium (in dollars)
15	1.8	27	2	392.5
14	1.2	16.8	10	46.2
17	1.2	20.4	8	15.7
7	1.8	12.6	3	422.2
10	1.6	16	4	119.4
7	1.4	9.8	3	170.9
20	1.2	24	7	56.9
21	1.6	33.6	6	77.5
18	1.2	21.6	2	214
11	1.6	17.6	5	65.3
7.9	1.4	11.06	3	250
8.6	1.6	13.76	3	220
12.3	1.2	14.76	2	217.5
17.1	1.6	27.36	1	140.88
19.4	1.2	23.28	6	97.25

Linear regression model (interaction effect)

Based on the data, the β parameters are:

β_0 = - 502.011, β_1 = 40.306, β_2 = 568.723,

β_3 = - 25.781and β_4 = - 30.547

Thus the model is

$Y = 502.011 + 40.306 \, x_1 + 68.723 \, x_2$
$- 25.781 x_3 - 30.547 \, x_4$

That is,

Premium = 502.011 + 40.306 Mileage + 68.723 Engine_Capacity - 30.54 Age -
25.781 Int_EC_Mil

Mileage	Engine_Capacity	Int_EC_Mil	Age	Premium (in dollars)
15	1.8	27	2	392.5
14	1.2	16.8	10	46.2
17	1.2	20.4	8	15.7
7	1.8	12.6	3	422.2
10	1.6	16	4	119.4
7	1.4	9.8	3	170.9
20	1.2	24	7	56.9
21	1.6	33.6	6	77.5
18	1.2	21.6	2	214
11	1.6	17.6	5	65.3
7.9	1.4	11.06	3	250
8.6	1.6	13.76	3	220
12.3	1.2	14.76	2	217.5
17.1	1.6	27.36	1	140.88
19.4	1.2	23.28	6	97.25

Thank You