LINEAR REGRESSION

Rishitha Pallala

2 August 2023

1 Introduction

Linear regression is a statistical analysis to determine the relation between dependent variables and one or more independent variables.It can be divided into simple and multiple linear regressions.

2 simple linear regression

In simple regression we deal with one independent variable which can be plotted on the x-axis and dependent variable which can be plotted on the y-axis. finding a "best fit line" will help in quantifying the relationship between the variables.

2.1 best fit line

$$y_i = ax + b \tag{1}$$

the y_i ,a,b in the equation(1) are the predicted value of the dependent variable ,slope and y-axis intercept of the line respectively.

2.2 residuals and least squares method

the residuals is the difference between the predicted y-values and observed y values.the E_i in the equation(2) is the residual.

$$E_i = y_p - y_o \tag{2}$$

The line with parameters a,b which has the least residual sum of squares (given by equation(3))will be considered the best fit line. this method of finding the best fit line is called 'least squares method'.

$$rss = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{3}$$

3 Evaluation metrics

To analyse the ability of the model to describe the relation , we will find the R^2 and p-value of the model.

3.1 Finding R^2

R^2 explains the correlation between the variables.its values range from 0 to 1.values closer and equal to 1 indicate high correlation and vica versa for values near 0. for example:if R^2=0.91, tells that the independent variable explains about 91% of the total variation.

$$rss = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{4}$$

$$tss = \sum_{i=1}^{n} (y_{mean} - y_i)^2 \tag{5}$$

$$R^2 = 1 - (rss/tss) \tag{6}$$

$$\tag{7}$$

3.2 Finding F-value and p -value

if we know the parameters in best fit line(p_{fit}) and mean line(p_{mean}), along with rss and tss,F-value of dataset can be calculated by (8)

$$F - value = \frac{\frac{tss - rss}{p_{fit} - p_{mean}}}{\frac{rss}{n - p_{fit}}} \tag{8}$$

we generate the CDF of F-distribution and calculate the "probability of extreme values" which is equal to "p-value"

$$p - value = pr(F - values >= F - values_{originaldata}) \tag{9}$$

if p-value <0.05, than the model has less impact due to random chances and R^2 also becomes significant.

4 Multiple regression

This model is used when more than one independent variables are present.The best fit line is given by equation (10) where a,b,c are parameters for the best fit line.

$$y = a + bx_1 + cx_2.. \tag{10}$$

4.1 considerations of multiple regression

Although all the calculations made for simple linear regression are valid for mutiple also, some extra points are:

**overfitting and underfitting:
Overfitting causes the data to be perfect with training data but not with the testing data as model uses all data including pseudo patterns caused by noise.can be reduced by adding clean and relevant data while removing unnecessary features.
In underfitting the model is not able to predict patterns in both training and test data,solutions are increasing features,complexity and reducing noise.

**Multicollinearity:
when a lot of feature variables are present many variables may be correlated and increase the bias of the model.doing pairwise correlation of the variables and removing few variables will help.

4.2 Assumptions:

1)Linearity of residuals:
The relation between the dependent and independent variables is linear.

2)Independence of residuals:
The errors should not be dependent on others. so ,no patterns should be observed in the residuals.

3)Normal distribution of errors:
If the distribution is not normal with mean around zero, then there might be unusual data points which manipulate the model.

4)Constant Variance:
this property is called Homoscedasticity and the lack of it is due to outliers.