Linear Regression

Contents

1. Discussion Questions
2. Linear Regression fundamentals
3. Performance measures
4. Statistical assumptions
5. Pros & cons

Questions to discuss

1. What is the relationship between Machine Learning and Supervised Learning?
2. What is Linear Regression and how does it work?
3. What are the evaluation metrics for regression?
4. What are the assumptions of linear regression?
5. What are the pros and cons of linear regression?

What is the relationship between ML and SL?

Machine Learning (ML)

- Machine Learning is the ability of a computer to do some task without being explicitly programmed.
- The ability to do the tasks comes from the underlying model which is the result of the learning process.
- The model is generated by learning from huge volumes (both in breadth and depth) of historical data reflecting the real world in which the processes are performed.

Examples of what machine learning algorithms can do

- Search through the data to look for patterns in the form of trends, cycles, associations, etc.
- Express these patterns as mathematical structures (model)
- Using those patterns to test the unseen data

Supervised Learning (SL)

- It builds a mathematical model using data that contains both the inputs and the desired output (labels or ground truth)
- There are basically two types of supervised learning:
  - Regression - where the desired output is in the form of continuous values
    - **e.g.** predicting the house prices based on some features like area, the number of rooms, etc.
  - Classification - where the desired output is in the form of categories
    - **e.g.** predicting if the person is likely to default on a loan based on the features like age, past transactions, etc.
- The model learns from the training data using these ‘target variables’ as reference variables.
- The model thus generated is used to make predictions about data not seen by the model before.

What is Linear Regression and how does it work?

- Linear regression is a way to identify a relationship between the independent variable(s) and dependent variable

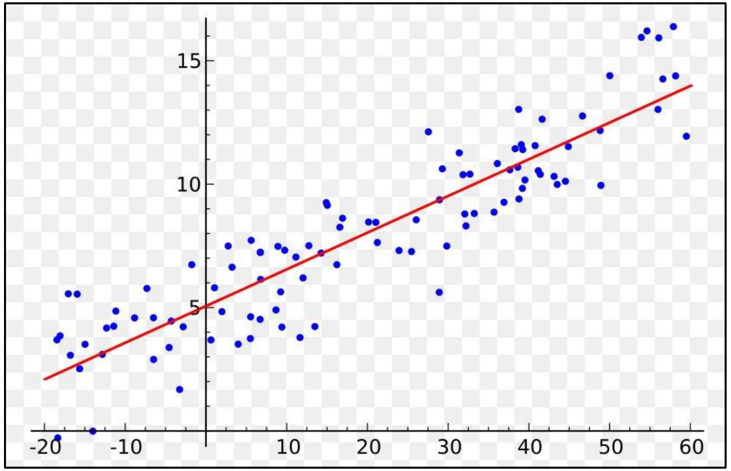
- We can use these relationships to predict values for one variable for the given value(s) of other variable(s)

- It assumes the relationship between variables can be modeled through a linear equation or an equation of a line.

- The variable, which is used in prediction is termed as independent/explanatory/regressor where the predicted variable is termed as dependent/target/response variable.

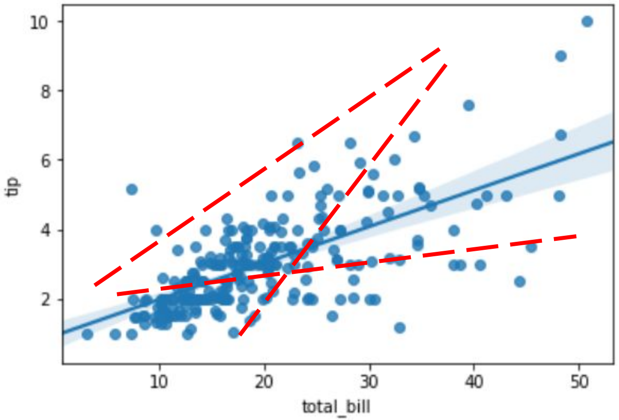
- In the case of linear regression with a single explanatory variable, the linear combination can be expressed as :

response = intercept + constant * explanatory variable



What is the best fit line in linear regression?

- Learning from the data, the model generates a line that fits the data.
- Our aim is to find a regression line that best fits the data
- By best fit, it means that the line will be such that the cumulative distance of all the points from the line is minimized
- Mathematically, the line that minimizes the sum of squared error of residuals is called Regression Line or the Best Fit Line.



In the example here, you can see a scatter plot between the *tip* amount and the *total_bill* amount

We can see that there is a positive correlation between these two - as the bill amount increases, the tip increases

The line in blue that you see is the ‘best fit’ line - those in red are some examples of all other lines that are not the ‘best fit’

What is Multiple Linear Regression?

- This is just the extension of the concept of simple linear regression with one variable
- In the real world, any phenomenon or outcomes could be driven by many different independent variables
- Therefore the need to have a mathematical model that can capture this relationship
  - Ex: predicting the price of a house, we need to consider various attributes related to the house, such as area, number of rooms, number of kitchens, etc.
- Such a regression problem is an example of multiple regression.
- It can be represented by :

  target = intercept + constant1*feature1 + constant2*feature2 + constant3*feature3 + ...

- The model aims to find the constants and intercept such that this line is the best fit.

What are evaluation metrics?

- Evaluating a model is very important as it helps us understand the model performance.
- Evaluation metrics allow us to quantify our model's performance using a single number.
- Comparing the metric values for train and test sets helps us get an idea about the fit of the model.
  - If the model performance is low on the train and test sets, then the model is said to underfit the data
  - If the model performance is high on the train set but low on the test set, then the model is said to overfit the data.
- The aim is to find the model which best fits our data.

What are the evaluation metrics for regression?

R-squared	Adjusted R-squared	Mean Absolute Error	Root Mean Square Error
● Measure of the % of variance in the target variable explained by the model ● Generally the first metric to look at for linear regression model performance ● Higher the better	● Conceptually, very similar to R-squared but penalizes for the addition of too many variables ● Generally used when you have too many variables as adding more variables always increases R^2 but not Adjusted R^2 ● Higher the better	● Simplest metric to check prediction accuracy ● Same unit as the dependent variable ● Not sensitive to outliers i.e. errors doesn’t increase too much if there are outliers ● Difficult to optimize from a mathematical point of view (pure maths logic) ● Lower the better	● Another metric to measure the accuracy of prediction ● Same unit as the dependent variable ● Sensitive to outliers - errors will be magnified due to the square function ● But has other mathematical advantages that will be covered later ● Lower the better

What are the assumptions of Linear Regression?

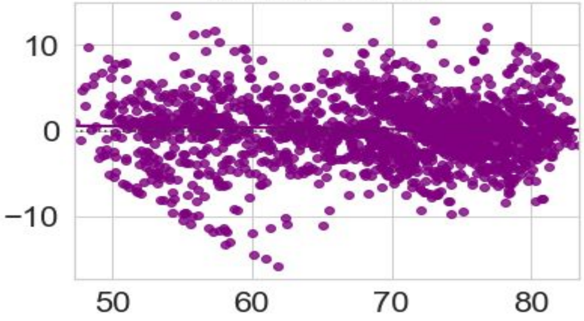
Assumption	How to test	How to fix
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance inflation factor)	Remove correlated variables or merge them
There should be a linear relationship between dependent and independent variables	Plot residuals vs. fitted values and check the plot	Transform variables that appear non-linear (log, square root, etc.)
The residuals should be independent of each other	Plot residuals vs. fitted values and check the plot	Transform variables (log, square root, etc.)
Residuals must be normally distributed	Plot residuals or use Q-Q plot	Non-linear transformation of the independent or dependent variable
No heteroscedasticity, i.e., residuals should have constant variance	Use statistical test (like goldfeldquandt test)	Non-linear transformation of the dependent variable or add other important variables

Testing Multicollinearity using VIF

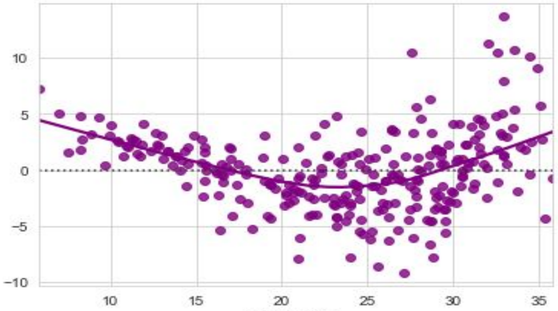
- Multicollinearity occurs when predictor variables in a regression model are correlated.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.
- We can detect or test for multicollinearity using the Variance Inflation Factor or VIF.
- Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors.
  - If VIF is 1, then there is no correlation between the selected predictor and the remaining predictor variables, and hence, the variance of its coefficient is not inflated at all.
- General Rule of thumb:
  - 1 < VIF <= 5: There is low multicollinearity
  - 5 < VIF <= 10: There is moderate multicollinearity
  - VIF > 10: There is high multicollinearity

Testing Linearity and Independence using residuals vs fitted values plot

- Predictor variables must have a linear relation with the dependent variable.
- If the residuals are not independent, then the confidence intervals of the coefficient estimates will be narrower and make us incorrectly conclude a parameter to be statistically significant.
- We can check for linearity and independence by checking a plot of fitted values vs residuals.
  - If they don't follow any pattern, then we say the model is linear and residuals are independent.
  - Otherwise, the model is showing signs of non-linearity and residuals are not independent.



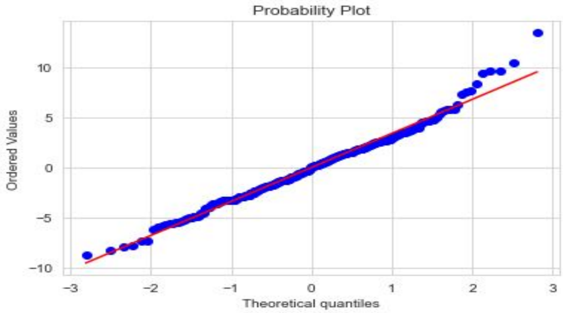
No pattern spotted



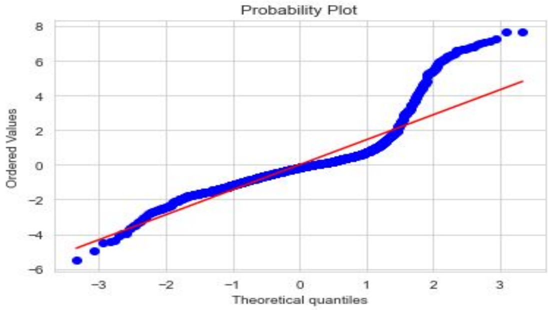
Some non-linearity spotted

Testing Normality using QQ plots

- If the error terms are not normally distributed, the confidence intervals of the coefficient estimates may become too wide or narrow.
- Non-normality suggests that there are a few unusual data points that must be studied closely to make a better model.
- The shape of the histogram of residuals can give an initial idea about the normality.
- We can also check normality via a Q-Q plot of residuals.
  - If the residuals follow a normal distribution, they will make a straight line plot, otherwise not.



Close to normal

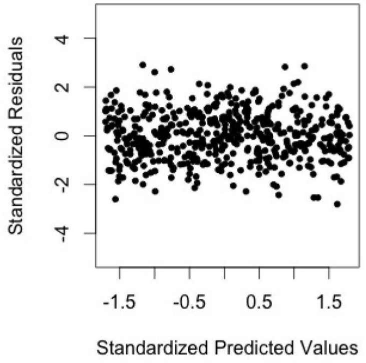


Not normal

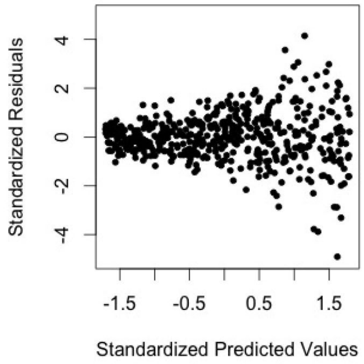
Testing Homoscedasticity using goldfeldquandt test

- If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic. Else, they are heteroscedastic.
- Generally, non-constant variance arises in presence of outliers.
- The residual vs fitted values plot can be looked at to check for homoscedasticity.
    - In the case of heteroscedasticity, the residuals can form an arrow shape or any other non-symmetrical shape.
- The goldfeldquandt test can also be used.
    - If we get a p-value > 0.05 we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.
    - Null hypothesis: Residuals are homoscedastic
    - Alternate hypothesis: Residuals have heteroscedasticity

Homoscedasticity



Heteroscedasticity



What are the pros and cons of linear regression?

Pros:

- Simple to implement and easier to interpret the outputs coefficient.
- Helpful if the relationship between the independent and dependent variable is linear

Cons:

- Has a lot of statistical assumptions which are not always true for real-world data
- Outliers can have huge effects on regression
- Assumes that the input variables are independent. It gets highly affected by multicollinearity.

greatlearning

Power Ahead

Happy Learning !

Appendix: Regression Model Evaluation Metrics

Metric	Formula
R-squared	$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Adjusted R-squared	$Adj.R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$
Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$