upGrad

Linear Regression Assignment – Bike Sharing

Subjective Questions

upGrad

Course : MS AI/ ML

Case Study : Bike Sharing

Project By : Varjit Gupta

upGrad

Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer> Categorical columns were analysed using Box Plot and Bar Plot, inferences are mentioned below

a) Year on Year demand has increased

b) Demand is higher when weather situation is (weathersit) is clear

c) Demand seems to be increasing till July, no clear ternd for Aug and Sep (When checked for both years) and decreases from oct to dec

d) Monday seems to be the day with highest demand, overall not much can be said about the relation of week days with demand

e) Demand increases during holiday

f) In terms of season, fall has seen highest demand

g) There is slightly more demand on working days

Q2) Why is it important to use drop_first=True during dummy variable creation?

Answer>

Any categorical variable having n values can be explained by n-1 dummy variables.

For example if we have a variable that has 3 values A,B,C. For an observation if the variable is not A or B then it has to be C

drop_first=True drops 1 dummy variable and gives n-1 dummy variables for a variable having n values. This helps in reducing the number of variables in the model and reduces complexity

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer>

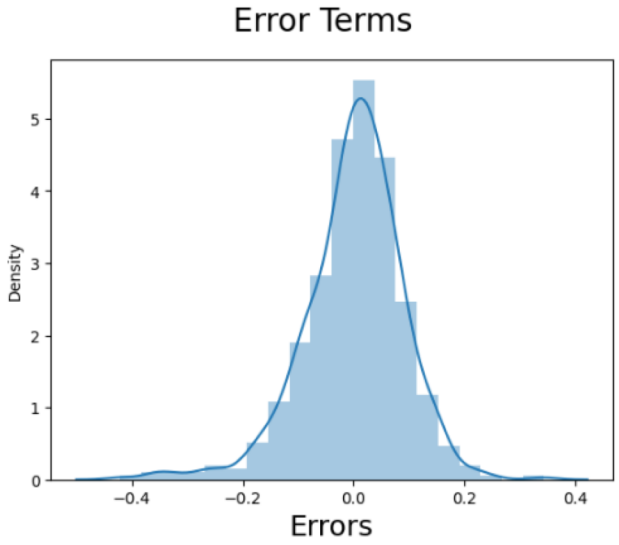
Temp variable has the highest correlation with target variable

Q4) . How did you validate the assumptions of Linear Regression after building the model on the training set?

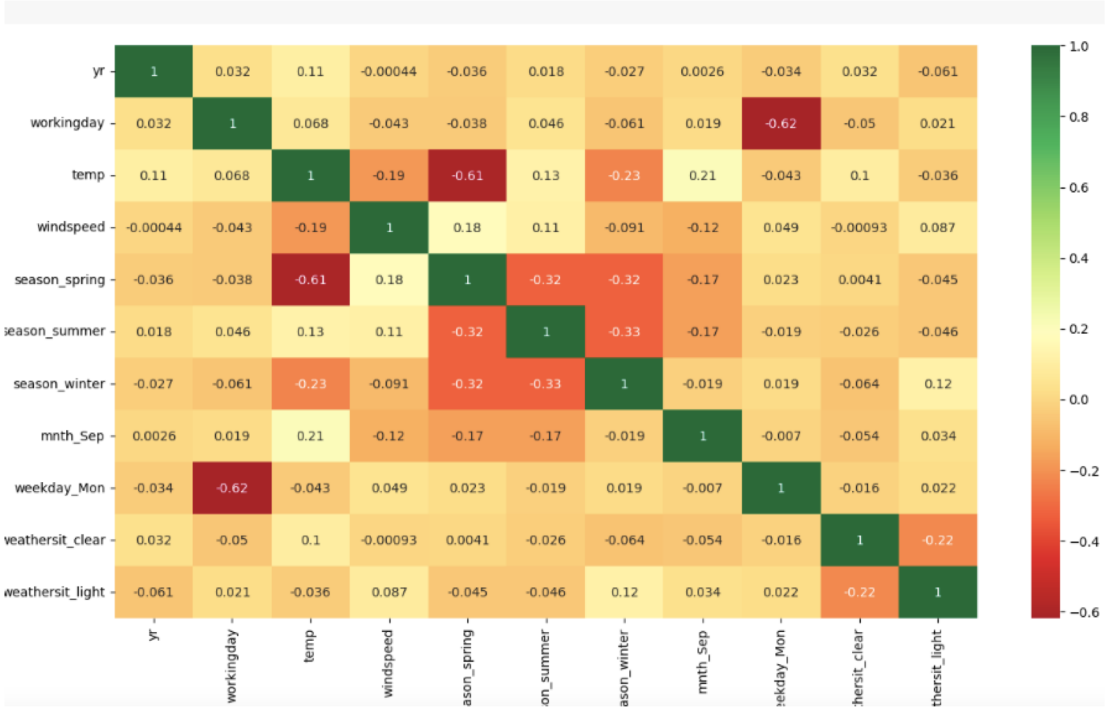
Answer>

a) Normality of error terms – This validated by plotting error terms in histogram. We can see error terms are normally distributed

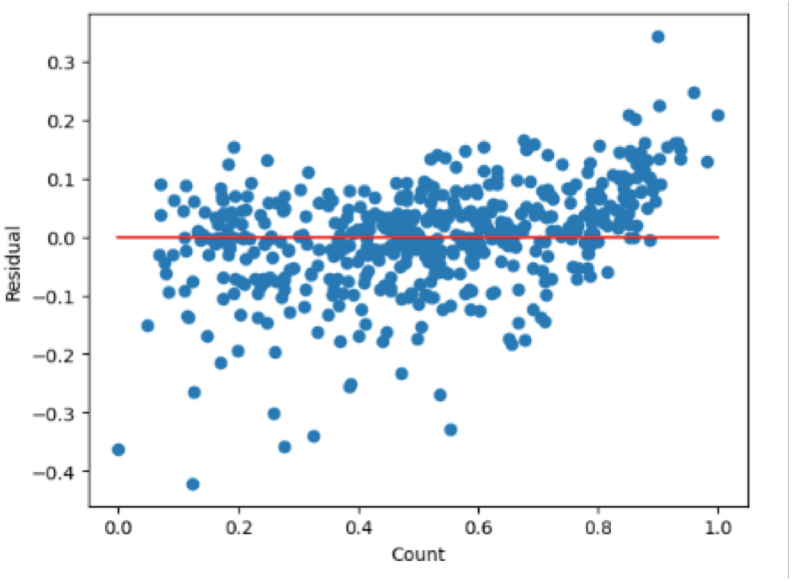
b) Residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.



c) Multicollinearity Check – There should be insignificant collinearity among the independent variables in the final model.
This was tested using Heat Map



d) Homoscedasticity – Checked using scatter plot residuals and cnt variable.
This was tested using Heat Map



e) Independence of variables – Checked using Durbin-Watson statistic value of the model

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	233.6
Date:	Wed, 14 Jun 2023	Prob (F-statistic):	1.42e-188
Time:	13:30:27	Log-Likelihood:	502.47
No. Observations:	510	AIC:	-980.9
Df Residuals:	498	BIC:	-930.1
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0602	0.031	1.961	0.050	-0.000	0.120
yr	0.2344	0.008	28.655	0.000	0.218	0.250
workingday	0.0555	0.011	4.996	0.000	0.034	0.077
temp	0.4796	0.033	14.678	0.000	0.415	0.544
windspeed	-0.1500	0.025	-6.007	0.000	-0.199	-0.101
season_spring	-0.0554	0.021	-2.692	0.007	-0.096	-0.015
season_summer	0.0626	0.014	4.447	0.000	0.035	0.090
season_winter	0.0958	0.017	5.788	0.000	0.063	0.128
mnth_Sep	0.0873	0.016	5.423	0.000	0.056	0.119
weekday_Mon	0.0667	0.014	4.665	0.000	0.039	0.095
weathersit_clear	0.0804	0.009	9.241	0.000	0.063	0.097
weathersit_light	-0.2089	0.025	-8.372	0.000	-0.258	-0.160

Omnibus:	76.073	Durbin-Watson:	2.083
Prob(Omnibus):	0.000	Jarque-Bera (JB):	187.745
Skew:	-0.765	Prob(JB):	1.71e-41
Kurtosis:	5.548	Cond. No.	20.4

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer>

a) Temp
b) Weathersit
c) Windspeed

General Subjective Questions

Q1) Explain the linear regression algorithm in detail

Answer>

Linear regression is a supervised learning algorithm used for predicting continuous values based on the relationship between input variables (features) and an output variable. It assumes a linear relationship between the independent variables and the dependent variable. Detailed overview of the linear regression algorithm is explained below,

Let's consider a simple case with one independent variable, often called a simple linear regression. The goal is to find a line that best fits the data points, minimizing the difference between the predicted values and the actual values. The line is represented by the equation:

$y = mx + c$

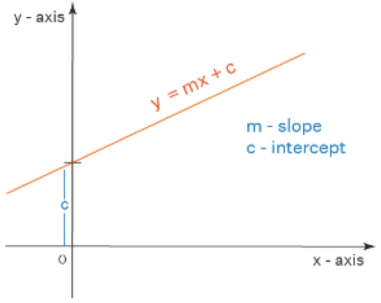
where:

y is the dependent variable

x is the independent variable (the input feature)

m is the slope of the

c is the y-intercept



The main steps of the linear regression algorithm are as follows:

Data Preparation:

- Gather a dataset consisting of input features (x) and corresponding output values (y).
- Split the dataset into a training set and a test set to evaluate the model's performance.

Model Training:

- During training, the algorithm aims to find the optimal values of m and c that minimize the difference between the predicted and actual values.
- The model tries to fit the training data by adjusting the slope and intercept values.
- The common method for finding the optimal values is called the "least squares" method, which minimizes the sum of squared differences between the predicted and actual values.
- This can be achieved using various optimization algorithms like gradient descent.

Prediction:

- Once the model is trained, it can be used to make predictions on new, unseen data.
- Given an input feature x, the model calculates the corresponding predicted value y using the equation $y = mx + c$.

Model Evaluation:

The performance of the linear regression model is assessed using evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or R-squared.

These metrics quantify the difference between the predicted values and the actual values, providing a measure of how well the model fits the data.

Linear regression can also handle multiple independent variables, known as multiple linear regression. In this case, the equation for predicting y becomes:

$y = b0 + b1x1 + b2x2 + ... + bnxn$

where:

b0 is the y-intercept (the point where the line intersects the y-axis)

b1, b2, ..., bn are the coefficients that determine the relationship between each independent variable (x1, x2, ..., xn) and the dependent variable (y)

The training process and prediction in multiple linear regression are similar to simple linear regression, but the equation is more complex due to the inclusion of multiple independent variables.

Q2) Explain the Anscombe's quartet in detail.

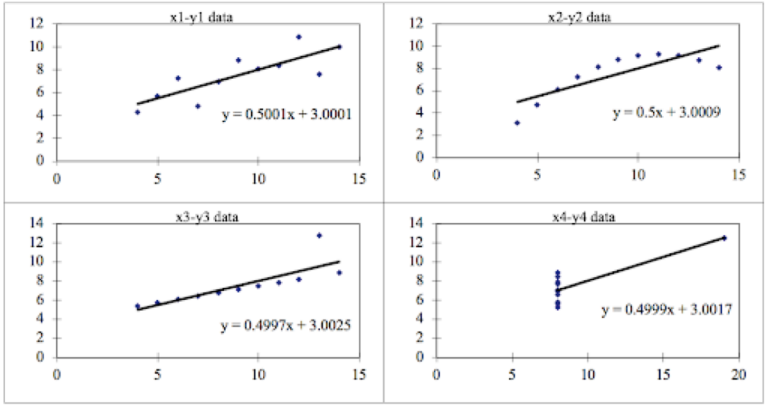
Answer>

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises of four datasets, each containing eleven (x, y) pairs. Thing that should be noted about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

Anscombe's Data									
Observation	x1	y1		x2	y2	x3	y3	x4	y4
1	10	8.04		10	9.14	10	7.46	8	6.58
2	8	6.95		8	8.14	8	6.77	8	5.76
3	13	7.58		13	8.74	13	12.74	8	7.71
4	9	8.81		9	8.77	9	7.11	8	8.84
5	11	8.33		11	9.26	11	7.81	8	8.47
6	14	9.96		14	8.1	14	8.84	8	7.04
7	6	7.24		6	6.13	6	6.08	8	5.25
8	4	4.26		4	3.1	4	5.39	19	12.5
9	12	10.84		12	9.13	12	8.15	8	5.56
10	7	4.82		7	7.26	7	6.42	8	7.91
11	5	5.68		5	4.74	5	5.73	8	6.89
				Summary Statistics					
N	11	11		11	11	11	11	11	11
mean	9.00	7.50		9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94	3.16	1.94	3.16	1.94
r	0.82			0.82		0.82		0.82	

We can see the summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the standard deviation of x is 3.16 and of y is 1.94 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.82 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the visualised data gives better and clear picture of the dataset.

Q3) What is Pearson's R?

Answer>

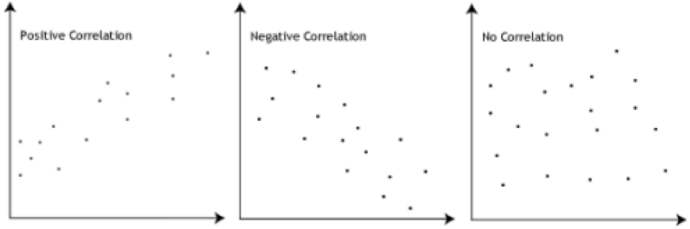
R, also known as Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistical analysis to assess the degree of association between variables.

The Pearson correlation coefficient, denoted as r, ranges between -1 and +1. The value of r indicates the strength and direction of the linear relationship between the variables:

If r = +1, it indicates a perfect positive linear relationship. As one variable increases, the other variable also increases in a linear fashion.

If r = -1, it indicates a perfect negative linear relationship. As one variable increases, the other variable decreases in a linear fashion.

If r = 0, it indicates no linear relationship or a very weak linear relationship between the variables.



Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer>

Scaling, in the context of linear regression, refers to the technique of transforming the values of variables to a specific range or distribution. It is performed to ensure that all variables are on a similar scale, which can have several benefits during data analysis and modelling.

The primary reasons for performing scaling are as follows:

Comparable Scale: Scaling allows variables with different units and ranges to be directly comparable. When variables have vastly different scales, those with larger values can dominate the analysis or modelling process.

Optimization Algorithms: Many machine learning algorithms are sensitive to the scale of variables. Variables with larger scales can have a more significant impact on the model's optimization process, leading to biased results. Scaling helps to prevent this.

Interpretability: Scaling enhances the interpretability of coefficients or feature importance measures. When variables are on different scales, their coefficients or importance values cannot be directly compared. Scaling allows for a fair interpretation of the magnitude and direction of the relationships between variables and the target variable.

Normalized Scaling (Min-Max Scaling)	Standardized Scaling (Z-score Scaling)
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	Outliers effect is less
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer>

Infinite value for the Variance Inflation Factor (VIF) typically occurs in situations of perfect multicollinearity. Perfect multicollinearity refers to an extreme cases where one or more independent variables in a regression model can be expressed as a perfect linear combination of other independent variables.

The VIF is a measure used to assess multicollinearity in a regression modelThe formula for calculating the VIF of an independent variable is:

$VIF = 1 / (1 - R^2)$

where R^2 is the coefficient of determination,

In the case of perfect multicollinearity, one or more independent variables can be expressed as a linear combination of the other variables.

This means that the coefficient of determination (R^2) for those variables becomes 1, indicating that they can be perfectly predicted using the other variables. Consequently, when R^2 equals 1, the denominator ($1 - R^2$) becomes zero, resulting in an infinite VIF value.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer>

A Q-Q (quantile-quantile) plot, also known as a quantile plot, is a graphical tool used to assess whether a dataset follows a specific probability distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically the normal distribution. The Q-Q plot provides a visual representation of the agreement or deviation between the dataset and the theoretical distribution.

The use of a Q-Q plot in linear regression are as follows:

Normality Assumption: The Q-Q plot helps to assess whether the residuals of a linear regression model follow a normal distribution. If the residuals are not normally distributed, it indicates a violation of the normality assumption.

Outliers and Skewness: The Q-Q plot can reveal the presence of outliers or skewness in the residuals.

Model Evaluation: The Q-Q plot is a useful tool to evaluate the goodness-of-fit of a linear regression model. If the residuals closely follow a straight line in the Q-Q plot, it suggests that the model assumptions are satisfied, and the model provides a good fit to the data.

Importance of Q-Q plot:
Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

upGrad

Thank You!