Linear Regression – Assignment Q&A's

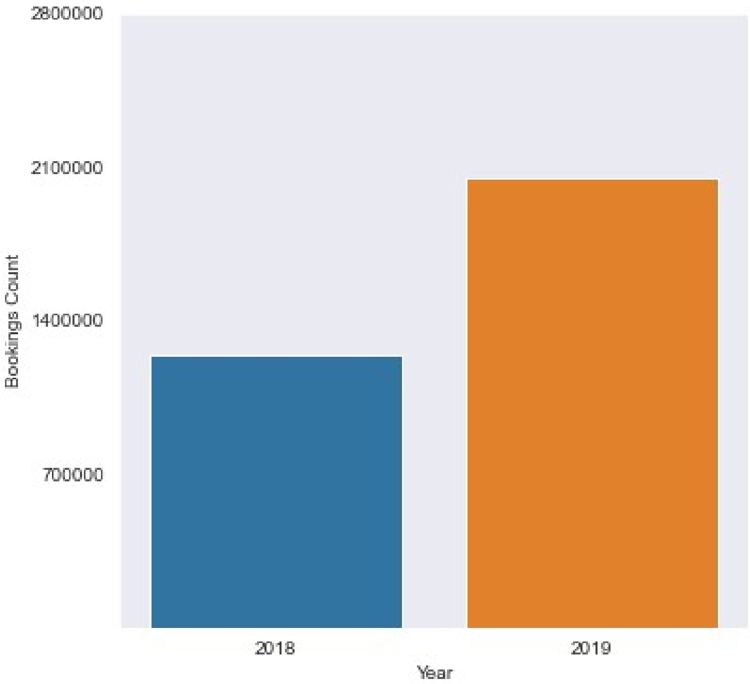
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

Categorical variables like

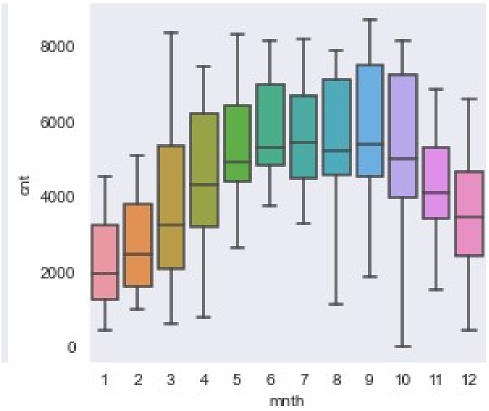
'yr' – helps us in evaluating the demand growth in given year: 2018, 2019.

Inference: Demand has risen very well in 2019 when compared to 2018

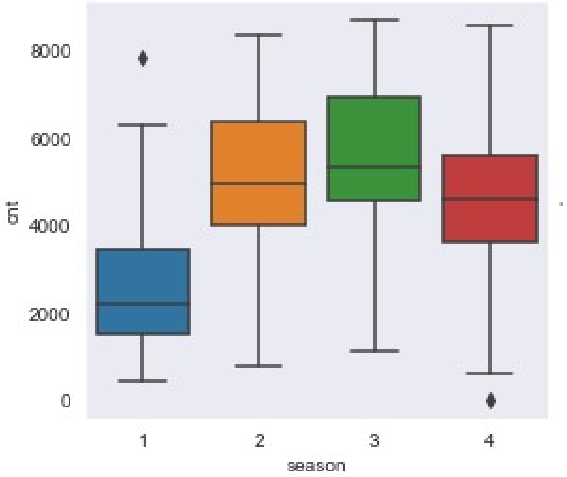


'month' - helps us to check demand variation with each month and then we can correlate to seasonal attribute to get high level inferences.

Inference: The demand has risen towards mid-year and tapered towards the end.



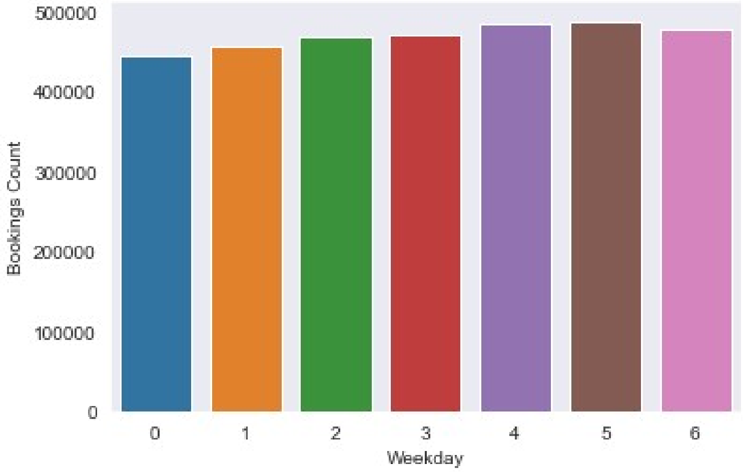
'season' – helps us to check the demand variance across spring (1), summer (2), fall (3) and winter (4).
Inference: The demand is high in summer and fall. Demand is lowest in spring.



'holiday' – helps us to check the demand on working day and holidays.
Inference: On holidays the demand is low when compared to other days.

'weekday' – helps us to find demand across week in general and predict rush days with respect to weekly demands as well.
Note: With given dataset:
Weekday:
Monday - 6
Tuesday - 0
Wednesday - 1
Thursday - 2
Friday - 3
Saturday - 4
Sunday – 5

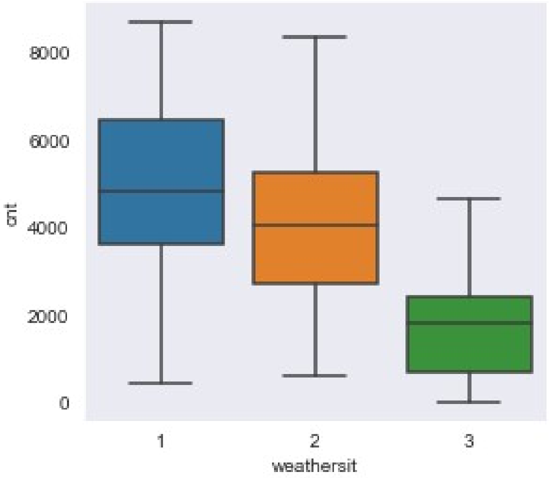
Inference: The below graph clearly shows the demand is high on day 4 and 5 which are Saturday and Sunday.



'weathersit' – helps us to check demand rise and fall during weather conditions like Clear (1), Misty (2), Snow (3) and Heavy Rains (4) etc.

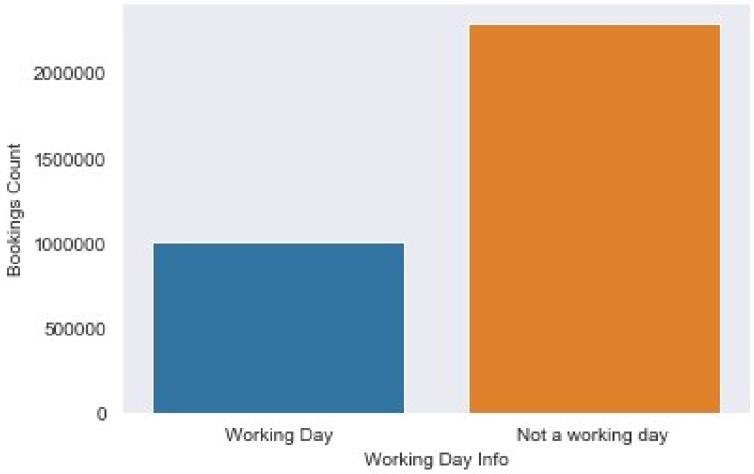
Note: We don't have data for Heavy Rain (4) days.

Inference: It is evident that on clear days (1) the demand is high and on Snow (3) days the demand is lowest



'workingday' – helps us to determine whether it's a holiday or not, based on this we can check the demand variance.

Inference: The demand is high on non-working day when compared to working days.



2. Why is it important to use drop_first=True during dummy variable creation?

Answer: drop_first=True can be used while creating dummy variable as it helps in reducing one column and hence directly reduces the correlations for the created dummy variables.

For example: If there' a column which denotes male-student (1), female-student (2) and staff (3) and we want to create dummy variables for this column. If we have 0 for male-student and 0 for female-student, then that row of data signifies that person is staff member and not a student. So, we may not need 3rd variable to identify the staff.

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

Answer: 'temp' and 'atemp' has very high correlation with target variable 'cnt'.

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

Answer: I have stats model and applied Multiple Linear Regression for bike share dataset.

I have validated by looking at p-value and Variance Inflation Factor (VIF) value for a feature before dropping them.

High - p, High - VIF -> should be dropped
High – Low:
    High - p, Low - VIF -> remove these first and re-run the model
    High - VIF, Low - p -> remove these after the above case.
Low - p, Low - VIF -> we can keep this feature

Conditions applied: Feature with below values are dropped after applying above conditions:

- VIF > 6.5 dropped (Industry standard practice few consider below 10 is good, but conservative models look at 5 as critical value for banking and healthcare related ones. Here I have chosen average at 6.5 as critical value for VIF)
- p-Value > 0.05 dropped

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

Answer:

1. atemp – feels like temperature has highest positive correlation
2. windspeed – has highest negative correlation
3. yr – year attribute too has correlation as the demand shot up sharply in 2019 compared to 2018.

General Subjective Q&A

1. Explain the linear regression algorithm in detail.

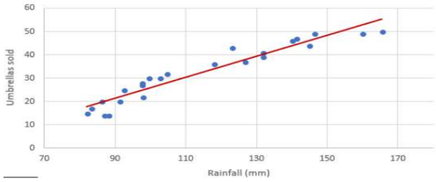
   Answer: Linear regression is a machine learning model classified as Supervised Learning.

   Linear Regression performs a task to help us in predict the value of target (dependent) attribute by forming a linear equation with one or many independent attributes. Thus, helping us to predict the future values by learning on current data set.

   Broadly Linear Regression models are of two types:

   1. Simple Linear Regression (SLR) – Deals with one predictor (independent) variable and a target variable.
      Equation: $Y = a + bX$
      Where X -> independent variable
                b -> coefficient or slope of the line
                a -> intercept value

Y -> target or dependent variable



2. Multiple Linear Regression (MLR) – Deals with multiple predictor (independent) variable and a target variable.

   Equation: $Y = a + b_1X_1 + b_2X_2 + \ldots. + b_nX_n$
   Where X_1 - X_n -> independent variable
   b -> coefficient or slope of the line
   a -> intercept value
   Y -> target or dependent variable

We can build Linear Regression models using

1. statsmodel api functions
2. Scikit Linear Regression Model using RFE

Stats Model is more from statistics perspective which gives in detail attributes like:

a) R-Square
b) Adjusted R-Square
c) F-Statistics
d) AIC
e) BIC

SciKit Linear Regression Model used mostly by programmers and not statisticians. But its all dependent on organization with their decision on which method to opt for.

Further points to be considered as part of linear regression:

High - p, High - VIF -> feature should be dropped
High – Low:
High - p, Low - VIF -> remove feature first and re-run the model
High - VIF, Low - p -> remove feature after the above case.
Low - p, Low - VIF -> we can keep this feature.

General boundaries:

1. p-value > 0.05 would be removed
2. VIF > 10 is considered to have high multi-collinearity and to be removed. Conservative modelling would consider VIF > 5 should be removed in case of highly sensitive domain like: Finance, Healthcare etc.

Scaling the data for better performance and comparison is essential. We can use one of the following:

1. Min-Max Scaling – formulae: (x - x_min) / (x_max - x_min)
2. Standardization – sklearn.preprocessing.StandardScaler

General Steps followed in Linear Regression Modelling:

- Step 1: Reading and Understanding Data
- Step 2: Visualizing the Data
- Step 3: Prepare the data for Modelling
  - Encoding -> One Hot Encoding technique can be used
  - Categorical variable to dummy variables
  - Split the data to train and test set
  - Rescaling the features (Min-Max Scaler)
- Step 4: Train the model (Stats Model can be used)
  - You can add all relevant attributes for the modelling
  - Generate the linear regression model, evaluate and then remove an independent variable and repeat the process.
  - Till all boundary conditions of p-value and VIF satisfied.
- Step 5: Residual Analysis
  - The error should be normal distribution
- Step 6: Prediction and Evaluation on Test data set
- Step 7: R-Sqaure value of train and test set should be close enough to term this as a successful model.

2. Explain the Anscombe's quartet in detail.

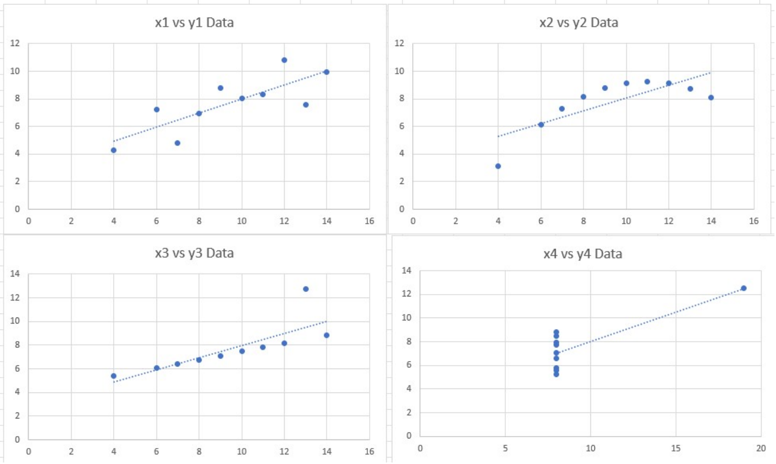
Answer:

Anscombe's quartet can be defined as four data set have similar statistical metrics, but there are some anomalies between them, that needs to be figured by plotting it in a graph before applying any kind of algorithm out in the market. They could have different distribution plots or appear differently when plotted against each other.

We will illustrate this with following example:

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
Stats								
Size	10	10	10	10	10	10	10	10
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Std Dev	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94

But when we plot the above pairs: x1-y1, x2-y2, x3-y3, x4-y4
It would look like:



This shows us even though statistics metrics look similar the plots and distributions are very different. If we happen to apply any regression models on this without looking at distribution and plot, we could be easily fooled and lead to lot of errors.

3. What is Pearson's R?
Answer:
Pearson' R is a statistical value that measures the linear correlation between two attributes. Its in range of 1.0 to -1.0

There are few constraints for Pearson' R coefficient:

1. There should be no outliers
2. Attributes should be normally distributed
3. Scale of measurement should be in ratio
4. The association should be linear

Formula is given by:

Pearson's R =
N * SUM(X*Y) – (SUM(X) * SUM(Y)) / SQRT((N * SUM(X^2) – SUM(X)^2) * (N * SUM(Y^2) – SUM(Y)^2)

N = number of pairs of scores
X = X scores
Y = Y scores

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:
Scaling: Scaling is a technique to bring the attributes to normalized range.

Why Scaling? – We do have multiple attributes to consider in our statistical analysis that can range from person age, salary, loan amount, number of cars he owns etc. Its very difficult to compare if we take all these values in their raw form. Few primary reasons for scaling:

1. It normalizes the data to be in range say 0 to 1 in case of Min-Max scaling so computation becomes easier and yield faster outputs.
2. Analyst doesn't have to deal with large coefficients thus complexity reduces to a great extent.

Difference between Normalized and Standardized Scaling:

Normalization	Standardization
Normalization does bring the attributes to a range of 0,1	Standardization doesn't have any range bound.
It is adversely impacted by outliers. If your data set has very high percent of outliers this normalization technique could causes issues	Standardization is impacted less by outliers
Normalization technique is useful when we don't know distribution pattern	Standardization technique is helpful when your data or feature is normally distributed.
Formula: X_val = (X - X_min) / (X_max - X_min)	Formula: X_val = (X - mean) / Std dev.

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

Answer: This happens due to perfect collinearity between independent variables. In this case R^2 gets to 1 and thus making 1 / (1 – R^2) to ∞.
We can solve this problem by removing the one of the features in data set and try running the model again.

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

Answer: Q-Q plot is a graphical tool which helps in assessing the data set if its origin from any of the theoretical distribution like Normal, Exponential, Uniform distribution.

Use and Importance of Q-Q Plot:

1. If we receive two data sets, separately and to determine whether it belong to populations from common distribution.
2. Helps in detecting shifts in scale, symmetry and outliers detection etc.

   In Python statsmodel api provides qqplot feature.