Charlie Thomas

Git hub link: charliethomasct82/Bike_sharing_regression (github.com)

Assignment -based Subjective Questions

Question1:From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?(3marks)

Answer1:

weekday:

weekday variable shows very constant trend across all days of the week having 13.5%-14.8% of total booking each day. ALL days have medians between 4000 to 5000 bookings. Weekday can have some or no influence on the target variable.

I will either drop this variable or will keep it, if it shows any effect as pair with other variable.

month:

Almost 10-15% of the bike booking were happening in the months May, June, July, August and September with a median of over 4000 booking per month. This indicates, month has some trend for bookings and can be a good predictor for the dependent variable.

weather_condition(Weathersit):

Almost 67% of the bike booking were happening during weathers_condition is clean with a median of close to 5000 booking (for the period of 2 years). Cloudy weather_condition with 30% of total booking. This indicates, weather_condition does have effect on target variable and can be a good predictor for target variable.

holiday:

Almost 97% of the bike booking were happening when it is not a holiday which means data is clearly biased. This indicates, holiday cannot be a good predictor for the dependent variable.

season:

Almost 30% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years).This was followed by season2 & season4 with 28% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable

working_day:

Almost 70% of the bike booking were happening on working day with a median of close to 5000 booking (for the period of 2 years).This indicates, working_day can be a good predictor for the dependent variable

Question2:Why is it important to use drop_first=True during dummy variable creation?(2 mark)

Answer2:

The regression model contains dummy variables of categorical data after using one-hot encoding. The variables are highly correlated with each other which means one variable can predict from other variables. In the regression model, this variable creates a trap which is called the dummy variable trap. Including all variable result in redundant data. If you do not use drop_first=True ,then even your VIF value will be infinite.

The solution of the Dummy variable trap is to drop/remove one of the dummy variables. If there are p categories than p-1 dummy variable should use. The model should exclude one dummy variable. This Overall approach reduces Multi-collinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression. Hence, it is important to use drop_first=true during dummy variable creation.

Question3:Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?(1 mark)

Answer3: temperature_felt (atemp) has the highest correlation with target variable with correlation value of 0.63 .Temperature_actual (temp) is close second with value of 0.62 correlation value.

Question4:How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer 4:

1) Linearity

This assumes that there is a linear relationship between the predictors and the response variable. This also assumes that the predictors are additive.

How to validate it: I have used a scatterplot to plot predicted values versus the actual values of target variable. The points lie on or around a diagonal line on the scatter plot. Hence, Assumption is validated.

2) Normality of the residuals

More specifically, this assumes that the residuals of the model are normally distributed.

How to Validate it: There are a variety of ways to do so, but we plotted a histogram as well as Q-Q plot on the residuals. Histogram gave a normal distribution of the training dataset. In Q-Q plot, Most of the points lie on the diagonal line with little deviation at start and end of the tails. Hence assumption is validated that residuals of the model are normally distributed.

3) No Autocorrelation of the Residuals.

This assumes no autocorrelation of the residuals. Autocorrelation being present typically indicates that we are missing some information that should be captured by the model.

How to detect it: We performed a Durbin-Watson test on the residuals to determine if either positive or negative correlation is present. Durbin Watson statistics always assume value between 0-4.

If the value is 2, then there is no Auto-correlation.

If the value 0-2 then there is positive auto-correlation

If the value is 2-4 then there is negative auto-correlation

Hypothesis for Durbin-Watson Test:

Null hypothesis,H0: First order autocorrelation does not exist.

Alternative hypothesis.H1:First order autocorrelation exists.

Value of Durbin-Watson test: 1.9095304087780651

we fail to reject the null hypothesis. Hence, we validate the assumption there is Little to no autocorrelation of the residuals

4) Homoscedasticity

This assumes homoscedasticity, which is the constant variance within our residuals.

How to Validate it: Plotted scatter plot on the residuals and the variance of the residuals appears to be uniform.Hence, the assumption is validated.

5) There is No Multicollinearity between the predictor variables

I utilised Variance Inflation factor. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

For example: I have kept VIF value of 5.0 as threshold in my assignment. Any features with value above this threshold was dropped from the model.

Features	VIF
season_spring	2.65
temperature_felt	2.56
season_winter	1.96
month_Nov	1.55
month_Mar	1.12
month_Sep	1.06
weather_condition_Light Rain	1.05
weekday_work_Thu_not_workingday	1.05
weather_condition_Cloudy	1.0
year_2019	1.02

There is low to no correlation within predictor variables. Hence, the assumption was established.

Question5 : Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?(2 marks)

Answer 5 :

As per our final Model, the top 3 predictor variables that influences the bike booking are:

1.temperature_felt: A coefficient value of '0.4399' indicated that a unit increase in temp variable, increases the bike hire numbers by 0.4399 units.

2.week_work_thu_not_workingday:

A coefficient value of '-0.3140' indicated that a unit increase in year variable, decreases the bike hire numbers by -0.3140 units.

3.weather_condition_Light_Rain:

A coefficient value of '-0.2977' indicated that a unit increase in weather_condition_Light_Rain variable decreases the bike hire numbers by 0.2977 units.

Company can predict demand of rental bikes using following multiple linear regression model:

count_rental_bikes = 0.2216 + (0.4399 * temperature_felt) - (0.1316 * season_spring) + (0.0744 * season_winter) + (0.2335 * year_2019) + (0.0449 * month_Mar) - (0.0390 * month_Nov) + (0.0706 * month_Sep) + (0.0604 * month_Sep) - (0.0667 * weather_condition_Cloudy) - (0.2977 * weather_condition_Light_Rain)-(0.3140 * weekday_work_Thu_not_workingday)

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.(4 Marks)

Answer1: Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.

The whole idea of the linear Regression is to find the best fit line, which has very low error(cost function).This line is also called Least Square Regression Line. Cost Function is basically the calculation of the error between predicted values and expected values. The Cost Function of a linear Regression is taken to be Mean Squared Error or Root Mean Square Error.

Linear regression is of the 2 types:

a. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$Y=\beta0+\beta1X1 +\epsilon$

b. **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.

Formula for the Multiple Linear Regression:

Y=β0+β1X1+β2X2+…+βpXp+ϵ

The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by the following two methods:

· Differentiation

· Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression.

Where β0 or a and β1.. βn or b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a\,(intercept) = \frac{n\sum y - b\left(\sum x\right)}{n}$$

Here,

X or x is the dependent variable we are trying to predict

Y or y is the dependent variable we are using to make predictions.

b or β1 is the slope of the regression line which represents the effect X has on Y

a or β0 is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

For example: Linear equation from my assignment.

count_rental_bikes = 0.2216 + (0.4399 * temperature_felt) - (0.1316 * season_spring) + (0.0744 * season_winter) + (0.2335 * year_2019) + (0.0449 * month_Mar) - (0.0390 * month_Nov) + (0.0706 * month_Sep) + (0.0604 * month_Sep) - (0.0667 * weather_condition_Cloudy) - (0.2977 * weather_condition_Light_Rain)-(0.3140 * weekday_work_Thu_not_workingday)

Interpretation of Coefficients:

Intercept: The Constant value of '0.2216' indicated that, in the absence of all other predictor variables (i.e. when x1,x2...xn =0), The bike rental can still increase by 0.2216 units.

temperature_felt: A coefficient value of '0.4399' indicated that a unit increase in temp variable, increases the bike hire numbers by 0.4399 units.

season_spring: A coefficient value of '-0.1316' indicated that a unit increase in season_spring variable, decreases the bike hire numbers by 0.1316 units.

season_winter: A coefficient value of '0.0744 ' indicated that a unit increase in season_winter variable increases the bike hire numbers by 0.0744 units.

year: A coefficient value of '0.2335' indicated that a unit increase in year variable, increases the bike hire numbers by 0.2335 units.

month_Mar: A coefficient value of '0.0499' indicated that, a unit increase in workingday variable increases the bike hire numbers by 0.0499 units.

month_Nov: A coefficient value of '-0.0390' indicated that, a unit increase in month_Nov variable decreases the bike hire numbers by 0.0390 units.

month_Sep: A coefficient value of '0.0706' indicated that a unit increase in month_Sep variable increases the bike hire numbers by 0.0706 units.

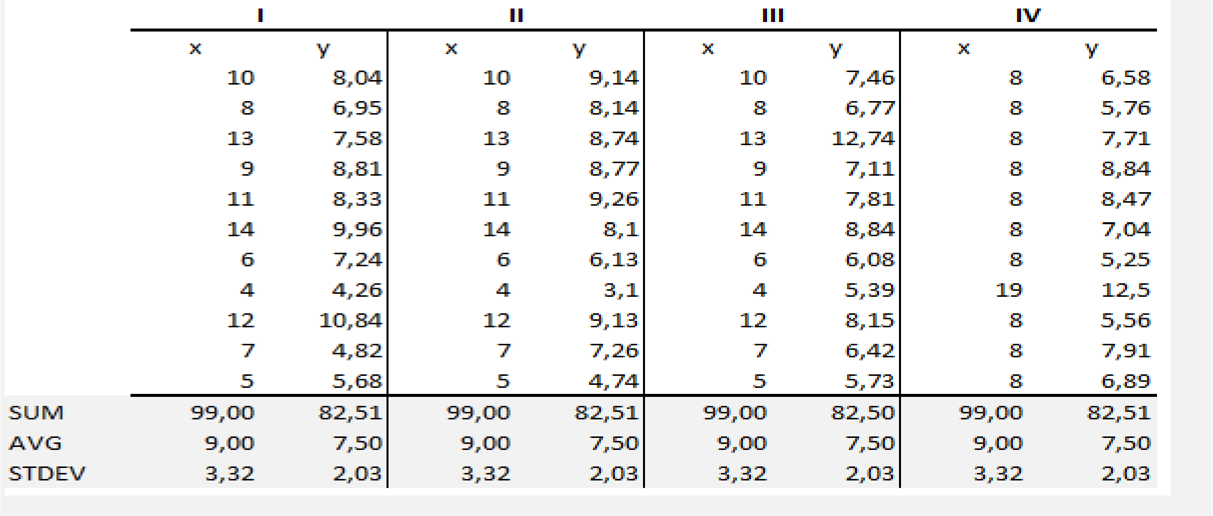
weather_condition_Cloudy: A coefficient value of '-0.0667' indicated that a unit increase in weather_condition_Cloudy variable decreases the bike hire numbers by 0.0667 units.

weather_condition_Light_Rain: A coefficient value of '-0.2977' indicated that a unit increase in weather_condition_Light_Rain variable decreases the bike hire numbers by 0.2977 units.

weekday_work_thu_not_workingday: A coefficient value of '-0.3140' indicated that a unit increase in weather_condition_Light_Rain variable decreases the bike hire numbers by 0.3140 units.

Question 2: Explain the Anscombe's quartet in detail.

Answer2: Anscombe's Quartet was developed by statistician **Francis Anscombe**. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

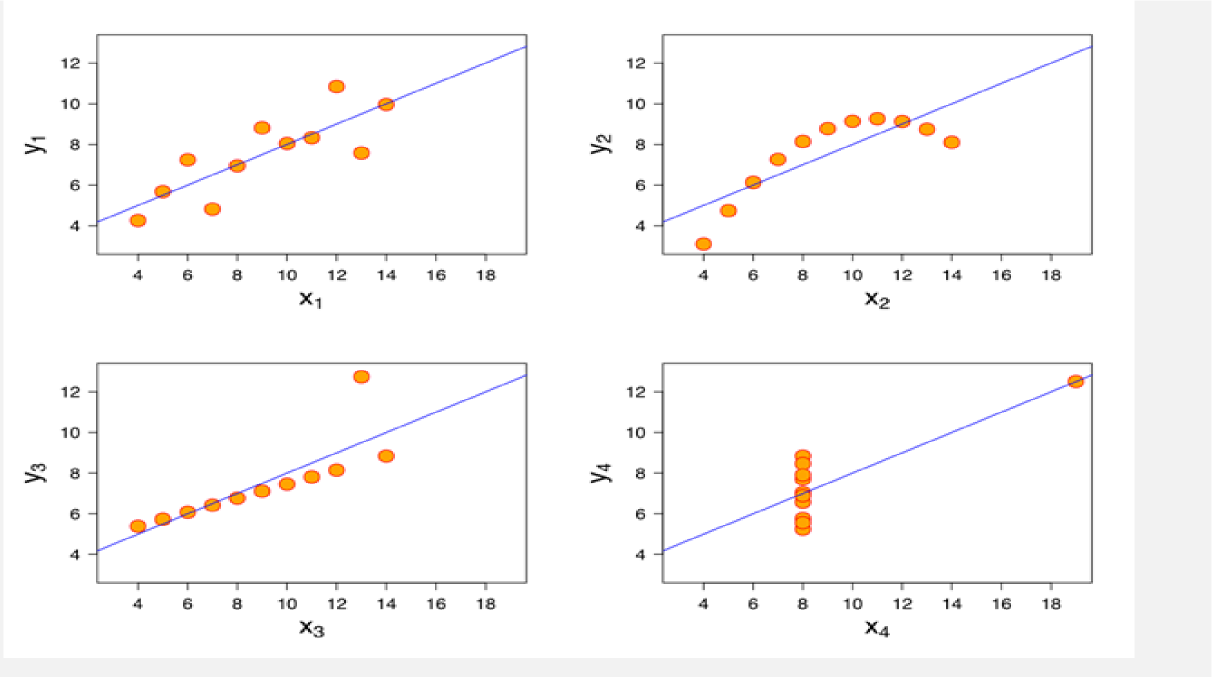
The summary statistics show that the means and the variances were identical for x and y across the groups:

Mean of x is 9 and mean of y is 7.50 for each dataset.

·Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

·The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 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.

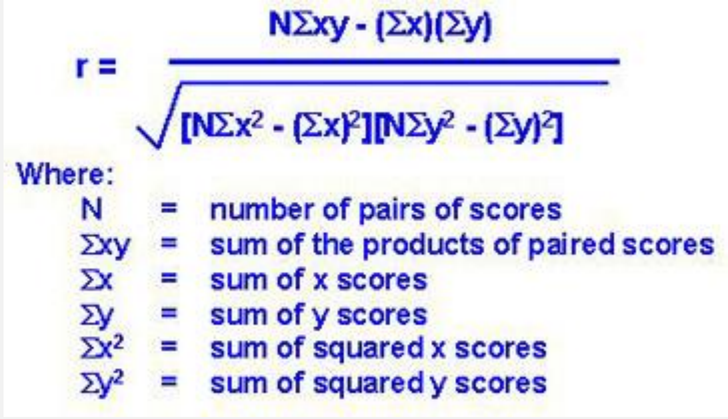
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset. Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets.

Question 3: What is Pearson's R?

Answer: Pearson's R was developed by **Karl Pearson** and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Mathematically, Pearson's correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations.

Formula:



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = number of pairs of scores

Σxy = sum of the products of paired scores

Σx = sum of x scores

Σy = sum of y scores

Σx^2 = sum of squared x scores

Σy^2 = sum of squared y scores

Example:

· Statistically significant relationship between age and height.

· Relationship between temperature and ice cream sales.

· Relationship among job satisfaction, productivity, and income.

· Which two variables have the strongest co-relation between age, height, weight, size of family and family income.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer4:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalized scaling:

- It brings all of the data in the range of 0 and 1.
- **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

Minmax Scaling:x = (x-min(x))/(max(x)-min(x))

Standardization Scaling:

- Standardization replaces the values by their Z- scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

  Standardisation: x= (x-mean(x))/sd(x)
- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

Answer5:

The value of VIF is calculated by the below formula:

VIFi =1/(1-Square(Ri))

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

For example, after one hot encoding ,if you do not drop the first peer variable then VIF will give infinite value as these peer variables are highly correlation to each other.

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

Answer 6: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.