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?

A1: The most important factor is the “feels like temperature” or atemp. As cycling is an outdoor activity that offers little shield from environmental factors, an ideal temperature tolerable by North Americans in particular and Humans in general would see a more hike in rentals.

The second factor is the wind speed. Strong headwinds or tailwinds can alter the cyclable experience as gusts of wind lead to an decreased enjoyability in cycling.

Certain seasons such as summer would boost the dopamine derived from a cycling experience in a North American environment.

Year: It is seen that with the increase in year the number of users open to cycling and renting cycles has a positive effect. Environmental consciousness and peer pressure could be a factor. As the saying goes correlation is not causation.

Certain months see an increase in cycling activity. These months could be linked with the seasons in general.

Weather conditions: Adverse weather conditions have a negative effect on rentals.

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

A2: drop_first implicitly drops the first dummy variable created for a categorical value that has n features. As we know the number of dummy variables needed to determine the outcome for a categorical nominal feature with n levels is n -1 . Think of the nth feature as a bit wise operation of all the n – 1 dummy variables.

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

A3: The temperature or actual temperature has the highest correlation at 0.67.

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

A4: As taught by our Gurus at UpGrad and IIT, I did a

1. Analysis of the Error Terms. It was a Normal Distribution with mean at zero and SD of 1
2. There was a linear relationship between the influencing features and the outcome

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

1. Actual Temperature
2. Windspeed
3. Weather Conditions

General Subjective Questions

Q1: Explain the linear regression algorithm in detail

A1:

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Usually a Regression analysis is utilized for the following uses cases:

- Determine effect of Input variables on Target variable.
- Effect of change in Target variable with respect to one or more input variable.

Using a Linear Regression model we can predict the influence of one or more factors - termed as features - on the outcome of a situation.

The effect of the features will usually have a positive or negative effect on the outcome.

In simple terms suppose the input factors are the amount of alcohol consumed by a person and the smoking patterns of a set of individuals. Given a truth table of the alcohol intake and smoking habits we would like to predict the longevity of a person.

A typical equation to model this situation would be

age = ***min_years*** + (**ca** * ***alcohol_units_per_day***) + (cs * ***cigarettes_per_day***)

where

ca is a factor that influences the ***alcohol_units_per_day*** consumed by the person

cs is a factor that influences the ***cigarettes_per_day*** smoked by the person

Normal Human intuition suggests that alcohol and cigarette smoking decreases longevity. But can be find any patterns to establish the same?

Using Linear Regression we can come up with an Algebraic model for the same.

The formulae for a Linear Regression consists usually of

- A constant factor which has no influence of the input factors. In this case the ***min_years***. I.e to irrespective of whether a person who smokes and drinks will live at least the above said years.
- Some factors (features) that affect the outcome : - ***alcohol_units_per_day , cigarettes_per_day***
- An effect (weightage) of the factors : **ca** , **cs**

min_years is termed as a intercept

alcohol_units_per_day , cigarettes_per_day are known as the independent variables.

ca and **cs** are termed coeffients.

age is the target variable

When only one independent variable affects the outcome, it is termed as Single Linear Regression. When two or more independent variables affect the outcome, it is termed Multiple Linear Regression.

To carry out a Linear Regression we start with an arbitrary intercept and coefficients . We then do what is known as a **Gradient Descent** using Calculus to arrive at the value of the outcome given a small change in the values of the coefficients.

In Supervised learning we already have a set of data that has the outcome of the features given values of input features. We need to have a Loss Function that determines the outcome of our assumptions of our coefficients.

In this case suppose we assumed positive values for our **ca** and **cs.** This would have an effect of increasing the longevity(age). But the observed actual data would suggest that we have deviated from the actual value. We would the adjust the coefficients to be lower than our initial coefficients so that the deviation from the observed value is minimal for each data point.

Repeating a process of continuously trying to decrease the loss over all points of our data set, we would eventually reach a point where the difference between the actual observation and the outcome of our formulae is the least. This is when we would have reach what is known as a **Global Minima.**

It is pertinent to know that in a Multiple Linear Regression, there could be multiple points where we could arrive at the lowest point for a given set of values. This is known as hitting the **Local Minima.**

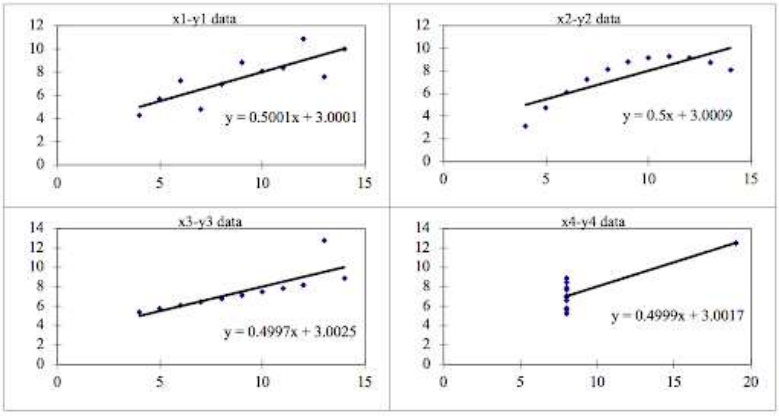
The subject of avoiding a **Local Minima** instead of a **Global Minima** is beyond the scope of this Question. But some factors that influence the same are choosing an approprpiate learning rate and using other advanced methods that have not yet been covered in this course , as of now.

Q2: Explain the Anscombe's quartet in detail.

A1: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

These models when plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm. (Figure from Wikipedia)



Q3: What is Pearson's R?

A3: a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

A4: Normally the data used for regression is in different orders of magnitude. So that a particular feature does not affect or dominate the outcome, we use scaling to bring down the magnitude of all the features into the same level.

There are two types of Scaling

- Standardization

  Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

- Normalization

  Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].

As a Thumb rule, Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true.

Q5: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.