Boom bikes

Assignment

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

Season and weathersit have an impact on cnt.

- The data indicates that the probability of people hiring a bike increases when the weather is clear and decreases when it is Misty and reduces when it is Snowing.
- The data indicates that the probability of people hiring a bike increases when the season is summer or fall and reduces when it is spring.

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

- We need only n-1 features to represent a categorical variable with n values. This is to avoid multi-collinearity.
- pd.get_dummies is a method to get the dummies on the dataframe. When we pass drop_first, it creates only n-1 features for the categorical variables.

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

After deleting registered and casual, atemp has the highest correlation with cnt.

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

To check if Linear regression assumptions are met

- Linear relationship - plotted a graph of predicted values vs actual values. This shows a linear relationship.
- No autocorrelation - Residual vs Fits plot is unstructured. All the model error terms are centered around 0 and have an approximately normal distribution. The errors are randomly scattered and centered around zero. So this can be considered as white noise. Durbin-Watson is close to 2. It is within the limit, indicating that we have homoscedasticity, or an even distribution of errors throughout our data.
- No multicollinearity - The VIF values are all below 5. This indicates that the independent variables are not correlated.
- Homoscedasticity - Residual vs Fitted plot does not have a prominent pattern.
- The distribution plot of residuals i.e error terms have approximately normal distribution.

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

Top 3 features of the final model

- 'atemp'
- ‘yr
- 'spring'

Explain the linear regression algorithm in detail.

Regression is the most commonly used **supervised** predictive analysis model to predict a **continuous** variable.

Method - generate a relationship between 2 or more variables and access the strength of the relationship.

Linear regression is a supervised machine learning method in which we find if the target variable has a linear relationship with the predictor variables.

Step 1:- Simple and multiple linear regression techniques to understand the relationship between the variables. The objective of this is to **find an equation that can determine the target variable based on the independent variables**.

- Confidence interval - None of the parameters' confidence interval should contain 0
- All the parameters should have statistical significance P>|t| should be less than 0.05
- The VIF values should be below 5

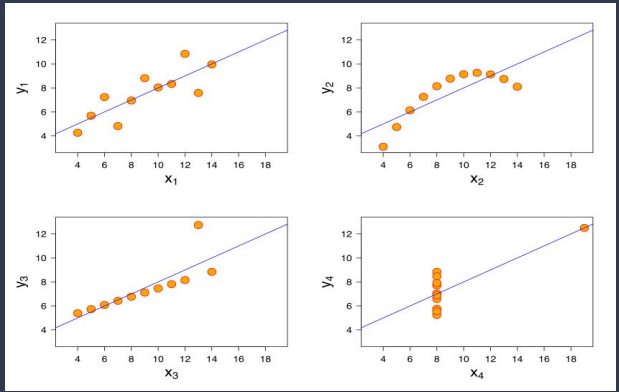
Check if Linear regression assumptions are met as mentioned in slide 5.

Step 2:- Access the strength of the model. Compute the degrees of freedom, R square, adjusted r-square and RSE and compare

Step 3:- Compare the models using AIC, BIC.

Based on all the parameters select the model. The final equation is derived based on the predictor variables coefficient.

Explain the Anscombe's quartet in detail.



Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

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 when analyzing it, and the effect of outliers and other influential observations on statistical properties.

The intention was to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.

- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

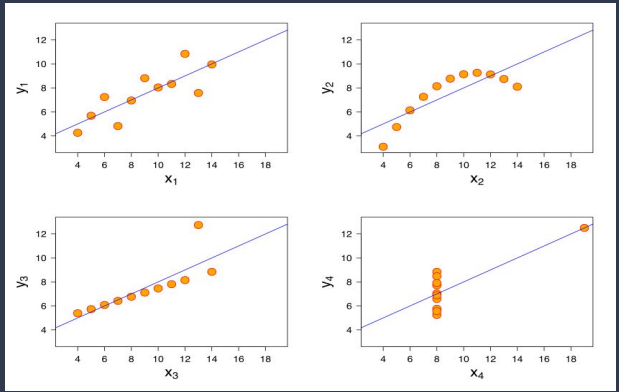
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Visualizing our data is important as summary statistics can be the same, while data distributions can be very different

What is Pearson's R?



A correlation coefficient is applied to measure a degree of association in variables and is usually called Pearson's correlation coefficient

Definition:- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

It is 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.

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to $a + bX$ and transform Y to $c + dY$, where a, b, c, and d are constants with b, $d > 0$, without changing the correlation coefficient. (This holds for both the population and sample Pearson correlation coefficients.)

The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1.[15] A value of 0 implies that there is no linear dependency between the variables.

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

It is important to have the predictor variables in comparable scale else some of the coefficients obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation.

So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. This is called scaling.

Min-Max scaling	Standardisation (mean-0, sigma-1)
It compress the data between a particular range.	It doesn't compress the data between a particular range.
This is useful when upper and lower boundaries are well known	This is useful for features that follow a normal distribution

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

Variance Inflation Factor(VIF) is used to examine if the features are correlated with each other . The formula for calculating VIF is (1/(1-R^2))

A large value of VIF indicates that there is a correlation between the variables. If the variable is perfectly predictable with other predictor variables, then VIF is infinity.

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

A **Q–Q plot** is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right.

* If the two distributions being compared are identical, the Q–Q plot follows the 45° line $y = x$.

*If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line $y = x$.

* if the general trend of the Q–Q plot is steeper than the line $y = x$, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis.

* Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

* Although a Q–Q plot is based on quantiles, in a standard Q–Q plot it is not possible to determine which point in the Q–Q plot determines a given quantile.

* The intercept and slope of a linear regression between the quantiles gives a measure of the relative location and relative scale of the samples. If the median of the distribution plotted on the horizontal axis is 0, the intercept of a regression line is a measure of location, and the slope is a measure of scale.

* The distance between medians is another measure of relative location reflected in a Q–Q plot. The "probability plot correlation coefficient" (PPCC plot) is the correlation coefficient between the paired sample quantiles.

Another common use of Q–Q plots is to compare the distribution of a sample to a theoretical distribution, such as the standard normal distribution $N(0,1)$, as in a normal probability plot. As in the case when comparing two samples of data, one orders the data (formally, computes the order statistics), then plots them against certain quantiles of the theoretical distribution.[3]