MACHINE LEARNING

Bike Sharing Assignment

Submitted By– Sumit Raghuvanshi

SUBJECTIVE QUESTIONS

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

Ans:

Categorical variables like yr., season, weather sit, month, holiday, weekday and working day are the predictions of how many bikes on rent out with the dependent variables. It effects on the dependent variable by directly proportional with categorical variable and rest are inversely proportional to dependent variable. while taking into account all the other variables may help visualize the “true nature of the relationship” between variables. The effects of relationship is not linear between the categorical and the dependent variables.

2. **Why is it important to use drop_first=True during dummy variable creation?**
Ans:
Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
Ans:
In the pair-plot among the numerical variables indicates that “registerted" has the highest correlation with the target variable.

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

Ans:

Validating the assumption of Linear Regression Model :

•**Linear Relationship:** Linear regression assumes that there exists a linear relationship between the dependent variable and the predictors. In our case, represents the relationship between the model and the predictor variables. As we can see, linearity is well preserved.

•**Homoscedasticity:** Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable. In our case, there is no visible pattern in residual values, thus homoscedastic is well preserved.

•**Absence of Multicollinearity:** Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). While it may not be important for non-parametric methods, it is primordial for parametric models such as linear regression. In our case, All the predictor variables have VIF value less than 5. So we can consider that there is insignificant multicollinearity among the predictor variables.

•**Independence of residuals:** Autocorrelation refers to the fact that observations' errors are correlated. To verify that the observations are not auto-correlated, we can use the **Durbin-Watson test**. The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables (0–2: positive auto-correlation, 2–4: negative auto-correlation). In our case, Durbin Watson= 1.977 so, here is positive auto-correlation.

•**Normality of Errors:** If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased. In our case, Based on the histogram, we can conclude that error terms are following a normal distribution.

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

Ans:

Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **Year (yr.)**

A coefficient value of '0.247' indicated that a year wise the rental numbers are increasing.

2. **Season(season_summer)**

A coefficient value of '0.135' indicated that a season_summer has significant impact on bike rents.

3.**Light Snow (weathersit =3)**

A coefficient value of '-0.295' indicated that the light snow deters people from renting out bikes.

GENERAL SUBJECTIVE QUESTIONS

1.**Explain the linear regression algorithm in detail.**

Ans:

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

- **Simple Linear Regression**: With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance. All of the data must be available to traverse and calculate statistics.

Simple linear regression uses traditional slope-intercept from, where 'm' and the 'b' are the variables our algorithm learn to produce the most accurate predictions *'x'* represents our input data and '*y*' represents our predictions.

$$y=mx+c$$

- **Multivariable regression**

A more complex, multi-variable linear equation might look like this, where we represents the coefficients, or weights, our model will try to learn.

$$f(x,y,z)=w1x+w2y+w3zf(x,y,z)=w1x+w2y+w3z$$

The variables x,y,zx,y,z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

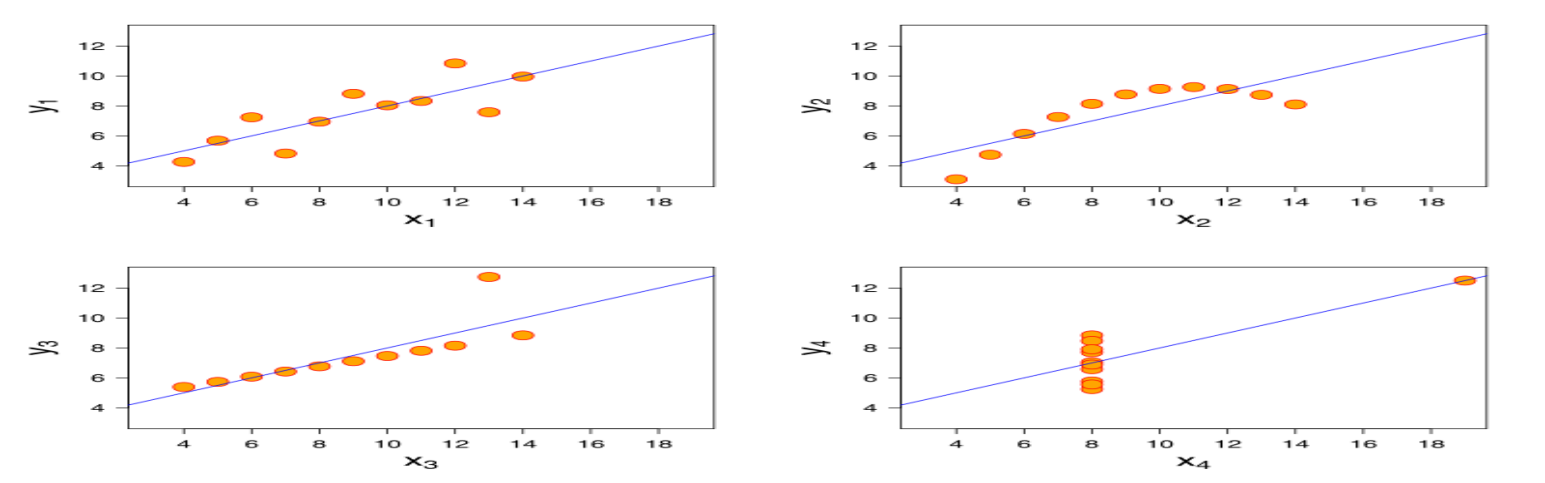
$$Sales=w1Radio+w2TV+w3News$$

2.**Explain the Anscombe's quartet in detail.**

Ans:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 and other influential observations on statistical properties. He described the article as being intended 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.
- _The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line .
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

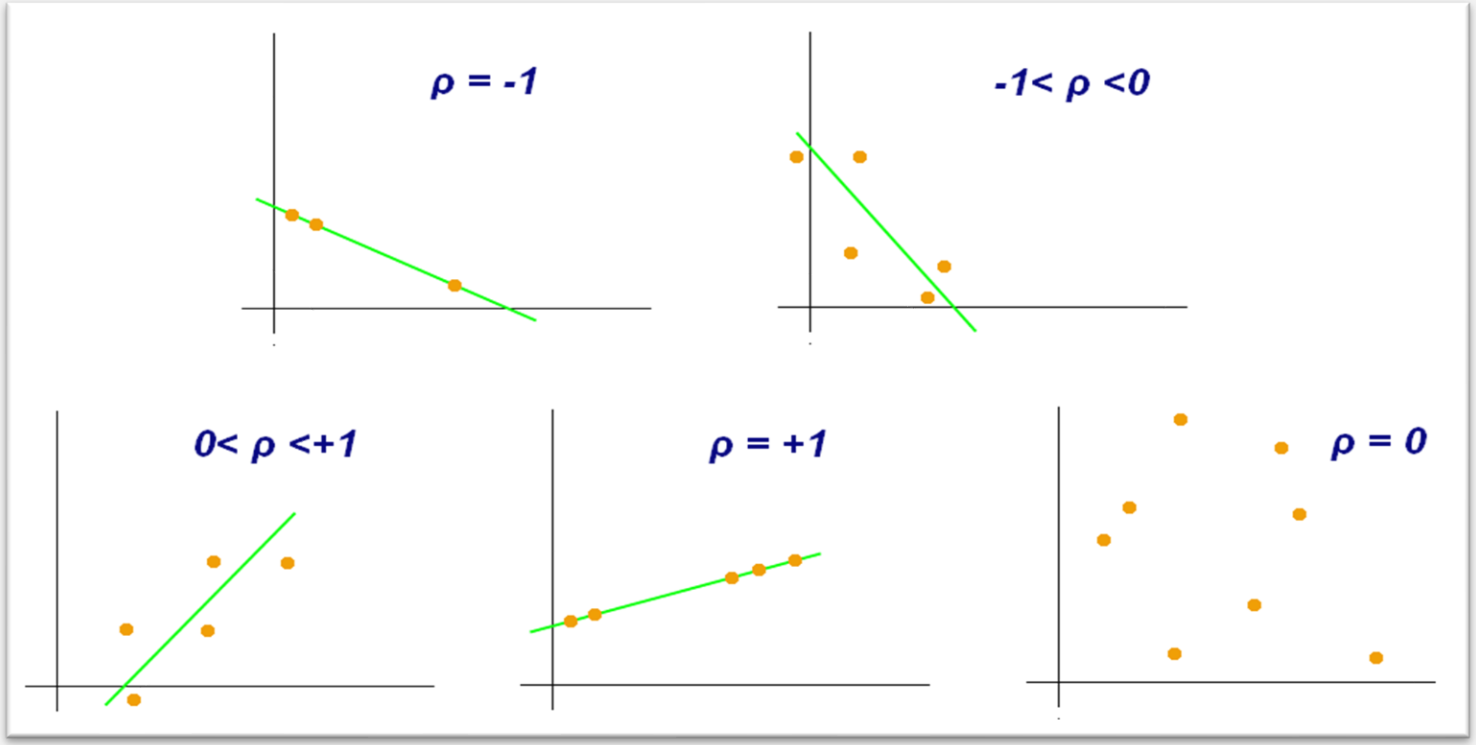


3. What is Pearson's R?

Ans:

The Pearson correlation coefficient , also referred to as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation is a statistic that measures linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Examples of scatter diagrams with different values of correlation coefficient (ρ).



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

Ans:

Scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

Scaling is performed because 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.

Below are the few ways we can do feature scaling:

MinMaxScaling: 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**).

Standardization: x= x-mean(x)/sd(x)

The difference between two most discussed scaling methods are: Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Ans:

VIF(Variance Inflation Factor)- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent 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).

When our model value of r-squared reaches to one then VIF reaches infinity. So, it is totally depend on the data and there r-squared.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The Q-Q plot use in linear regression in a scenario when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Some other uses and importance of a Q-Q plot in Linear Regression are:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

I. Come from populations with a common distribution
II. Have common location and scale
III. Have similar distributional shapes
IV. Have similar tail behavior

THANK YOU!