Kaustav Roy
Assignment:
Submission date: 13.03.2024

Topic: Bike Sharing Assignment

Assignment-based 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. There are 7 categorical columns in the raw dataset

season yr mnth holiday weekday workingday weathersit

Inference drawn :

a) Season : among the categories of *season*, spring or season_map_Spring shows most statistically significant regression coefficient of **-0.043**

b) yr : has a regression coefficient of **0.234**

c) mnth : showed significantly poor statistics like high p-value and high VIF values,hence not used in final model feature variables.

d) holiday : has a regression coefficient of **-0.101**

e) weekday : among the categories of *weekday*, Monday or weekday_map_Monday shows most statistically significant regression coefficient of **-0.043**

f) workingday : showed lower statistical significance than holiday, hence not used in final model feature variables.

g) Weathersit : among the categories of *wearhersit*, Lightsnow or *season_Lightsnow* shows most statistically significant regression coefficient of -**0.154**

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

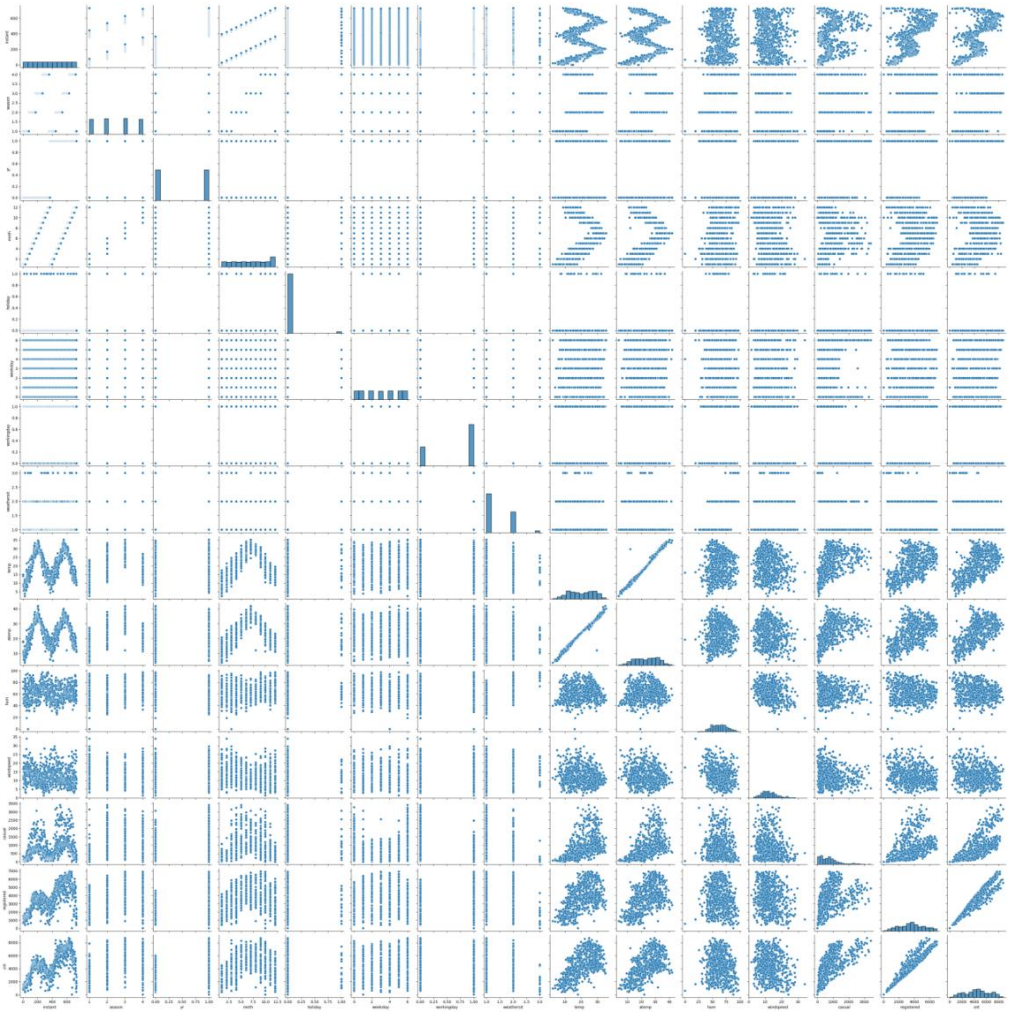
Ans. It is import to set drop_dirst=True when creating dummy variable as it ensures that only n-1 dummy variables are created for a categorical variable with n categories.By dropping one dummy variable, the multicollinearity is eliminated and the dummy variable trap can be avoided.
To avoid dummy trap means to make sure there exists no scenario where two or more dummy variables are perfectly correlated and hence leading to multicollinearity. Avoiding multicollinearity ensures reliable and interpretable results from regression model.

Eg,if color has three categories(n=3) red, blue and green, having two dummy variables (n-1=2) is sufficient. Since, absence of red or blue will imply green. To avoid the use of third dummy variable of green we will use *drop_first=True.*

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

Ans. Looking at the generated pair-plot,we can say that temperature and cnt (target) are most linearly correlated as per business logic.

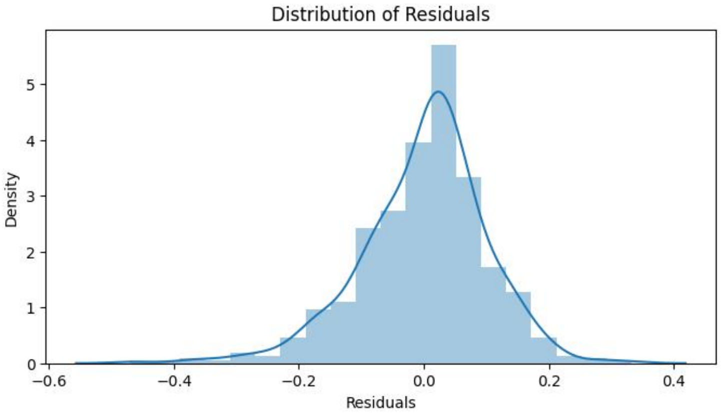
By simple visualization observation, we can say that temp and atemp are the most correlated.



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

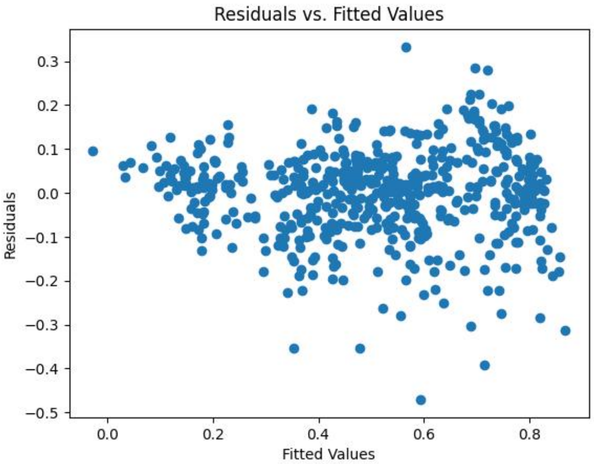
Ans. Assumptions of linear regression model are as follows,
a) We assume distribution of errors as a normal distribution
b) We assume that errors are independent of each other
c) We assume that errors have a constant variance.

In our linear regression model we validate the assumption by plotting the residuals/errors.



This is the distribution of the residuals from our trained model. We can observe that the distribution is similar to a normal distribution, hence validating our assumption.

Similarly, a scatter plot of residuals vs predictor can be observed to understand independence of errors.



As we can clearly see there exists no patter between residuals and predictors, it validates the assumption that errors are independent of each other

Similarly, the fact that the residual vs scatter plot is centered around 0, it validates the assumption of errors having constant covariance.

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

Ans. After evaluation best model, the top 3 features that contribute significantly towards explaining the demand of shared bikes are:

a) Temp : it has a regression coefficient of 0.381
It goes on to explain that a rise in temperature might be an outcome of people enjoying a nice bike ride on a sunny/warm day

b) Weather (Lightsnow) : it has a regression coefficient of -0.251
Inference can be drawn that among all weathers Lightsnow shows statistically significant regression coefficient. The negative nature implies that people do not like to share bikes if it likely to snow

c) Yr : it has a regression coefficient of 0.234
Inference can be drawn that given all other features constant, there has been ≈ 23% increase in shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans. Linear regression is a statistical technique to model a linear relationship between a target variable (dependant variable) and one or more independent variables (predictor variables).

The definition of the model is mathematically given as

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \dots + X_n\beta_n + \varepsilon$$

Here the linear regression model establishes relationship between the Target variable Y with independent variables $X_1, X_2, \dots X_n$. β_0 , β_1 , … β_n are regression coefficients or parameter coefficients. Each β value represents the change in target variable Y if the X corresponding to the β increases by one unit, given all other parameters remain constant. ε is the error term representing deviation of observed model from true model.

Objection of linear regression model is to estimate the regression coefficients β that minimizes sum of squares between observed values and predicted values of the dependent variable. Typically done by least square method. The estimated coefficients can be obtained using various techniques, for example Gradient Descent optimization. Once the coefficients are estimated, the models performance is evaluated using various metrics such as R^2 , adjusted R^2 , mean squared error.These assess the goodness of fit and its predictive capability.

In summary, linear regression is a statistical tool that models relationship by fitting a linear equation to observed data. It provides valuable insights into the underlying relationship and enables predictions and inference tasks in various domains.

In python, linear regression can be modeled as follows (assume the data is already fit and doesn't require scaling or creation of dummy variables:

```
#import libraries
import numpy as np
import statsmodels.api as sm
from sklearn.metrics import r2_score

#adding constant for linear regression model of statsmodel
Xtrain_sm = sm.add_constant(Xtrain)
lr = sm.OLS(y_train, Xtrain_sm)
lr_model = lr.fit()
#predicting on test
ytest_pred = lr_model.predict(Xtest)
#Evaluation results
print("Training R^2 score: ", lr_model.rsquared)
print("Test R^2 score: ", r2_score(y_true = ytest, y_pred = ytest_pred
```

2. Explain the Anscombe’s quartet in detail.

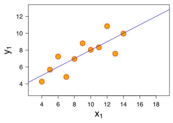
Ans. Anscombe’s quartet is set of 4 different datasets that have nearly identical summary statistics. The mean, standard deviation, correlations, regression lines were all similar but these datasets exhibited drastic graphical differences.

It was introduced by Francis J.Anscombe in 1973 emphasizing the usefulness of graphs and restraining from drawing conclusion based on summary statistics solely.

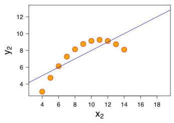
In his paper ‘Graphs in statistical Analysis’, Anscombe criticizes the following notions:
1) ‘numerical calculations are exact, but graphs are rough;’
2) ‘for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis’
3) ‘performing intricate calculation is virtuous, whereas actually looking at data is cheating.

To summarize, Anscombe emphasizes that visual representation of data shows underlying insights correctly which would be overlooked if one just relied on statistical conclusion.

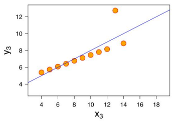
An Anscombe’s quartet has 4 datasets.
a) Dataset 1: Linearly related points with some noise.Best for linear regression model



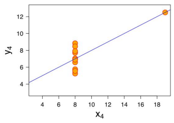
b) Dataset 2: Non linear points but linear regression line seems to be a good fit because of similar mean and variance of the variables



c) Dataset 3: Linearly related points but with an outlier that significantly impacts regression analysis



d) Dataset 4: Perfectly non-linear relationship between variables, obscured by the presence of an outlier



All 4 datasets are statistically similar put graphically vastly different. Hence, in summary the Asncombe’s quartet exemplifies the necessity of complementing traditional statistics with visual exploration of data.This can be seen in mordern day practice of Exploratory Data Analysis.

3. What is Pearson’s R?

Ans. Pearson’s R is a correlation coefficient denoted by *r* to measure linear relationship between two variables. It quantifies the strength and direction of a linear association between two continuous variables. Measure of degree of association.

By formula,

$$\mathrm{p(X,Y)} = \sum \frac{(\mathrm{X} - \bar{\mathrm{X}}) \cdot (\mathrm{Y} - \bar{\mathrm{Y}})}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}}$$

This can be simplified as,

$$\mathrm{p(X,Y)} = \frac{\mathrm{cov(X,X)}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{x}}}$$

There are two extreme cases,
a) X and Y are independent => p=0, I.e they are uncorrelated. Note, converse of this is not true, that is p= 0 but are X and Y are dependent.
b) X and Y are as correlated they can be, I,e Y=X.Then our equation becomes as follows,

$$\mathrm{p(X,X)} = \frac{\mathrm{cov(X,X)}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{x}}}$$

$$=> \mathrm{p(X,X)} = \frac{\mathrm{var(X)}}{\sigma_{\mathrm{X}}^{2}} = 1$$

Hence we see, Pearson’s r ranges from ±1 to 0. Having a r value of ±1 means that there exists a deterministic relationship between X and Y.

Pearson’s r makes 2 assumptions about the variables.
a) It assumes existence of a linear relationship between X and Y in the for of Y=aX+b
b) It assumes variables are approximately normally distributed.

A big advantage of Pearson’s r is that the correlation coefficient,if we change the units of the random variables it would not affect the correlation value with other variables. This is because mathematically the units in the numerator and denominator cancel out each other.This reflects the fact that Pearson’s r is dimensionless.

Hence it is a widely used measure of linear relationship between two continuous variables.

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

Ans. The process of transforming data to a common scale is known as scaling.It means to transform data to a specific range.

Scaling ensures that features contribute equally to the model in order to achieve an optimal model. Without scaling, a feature would impact just due to its magnitude of feature value, hence disturbing uniformity among features.

Machine learning algorithms are sensitive to the scale of input features. Scaling helps algorithms to converge faster and perform more effectively.

Normalized and standard both are scaling techniques. Normalized scaling is described as

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This results in features to be mapped between 0 and 1.
Standard scaling on the other hand transforms the data such that the mean of the data is 0 and the standard deviation is 1. The formula is give as

$$X_{std} = \frac{X - \mu}{\sigma}$$

μ is the mean of the data and σ is the variance of the data.

---

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

Ans. The VIF (Variance Inflation Factor) is calculated by the formula,

$$\mathrm{VIF}(\mathrm{X_i}) = \frac{1}{1 - \mathrm{R}^2_{\mathrm{X_j}}}$$

Infinite VIF means that $(1 - \mathrm{R}^2_{\mathrm{X_j}}) \rightarrow$ or = 0

Therefore, $\mathrm{R}^2_{\mathrm{X_j}}$ has to be 1

The value $\mathrm{R}^2_{\mathrm{X_j}}$ implies that there exists perfect multicollinearity.

A real example of where we encounter this scenario is when we do not drop one of the dummy variables created.

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

Ans. A Q-Q plot stands for Quantile-Quantile plot.It compares quantiles of a dataset to quantiles of the theoretical distribution, generally a normal distribution.

The Q-Q plot can be comprehended in the following manner -
* if the dataset follows the assumed theoretical distribution,points on the Q-Q plot will fall on the straight line.
* any deviation of the points from the straight line,it means dataset does not follow the assumed distribution.

Use and importance of Q-Q plot in linear regression is as follows:
* Assumption Checking: In linear regression we assume the assumption that the residuals are normally distributed. Hence, we can plot Q-Q plot to check whether the residuals for a straight line (I.e do they follow our assumption).
* Outlier detection: Q-Q points can be used as outlier detection. Points that will deviate significantly on the Q-Q plot,are the outliers.
* Model Improvement: Checking how the residuals of the model are plotted in a Q-Q plot,I.e how close their distribution is to that of a normal distribution.

In python we can observe a Q-Q plot as follows(assuming model is already trained and predicted):

```
res = ytrain - ytrain_pred
stats.probplot(res,dist=[“norm”], plot=plt
plt.title(“Q-Q plot”)
plt.xlabel(“Theoretical_Quantiles”)
plt.ylabel(“Sample_Quantiles”)
plt.show()
```