

Programming Assignment

Data Preparation

1. **Categorical Variables:** Some columns like 'weathersit' and 'season' have values as 1, 2, 3, 4, which should be converted into categorical string values to avoid the assumption of order. Refer to the data dictionary for details.
2. **'yr' Column:** While the 'yr' column only has two values (0 and 1) representing the years 2018 and 2019, it might still be valuable for prediction since the demand for shared bikes is increasing each year. Consider keeping it in the dataset.

Model Building

3. **Target Variable:** The target variable is 'cnt', which represents the total number of bike rentals, including both casual and registered users.

Model Evaluation

4. After building your multiple linear regression model and performing residual analysis, use the following code to calculate the R-squared score on the test set:

```
python
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

Ensure that you perform this step, as the R-squared score on the test set holds some marks. You should replace `y_test` and `y_pred` with your actual test data and predicted values.

Submission

Your submission should include the following:

1. **Python Notebook:** Create a Jupyter notebook containing the entire linear model, including data preprocessing, model building, predictions, and evaluation.
2. **Subjective Questions PDF:** Answer the general subjective questions related to linear regression in a separate PDF file. The questions are listed below.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more independent variables. Here's a detailed explanation:

- **Objective:** Linear regression aims to find the best-fitting linear relationship (line) between the independent variables (features) and the dependent variable (target) by minimizing the sum of squared errors.
- **Equation:** The linear regression equation is of the form:

- makefile
- $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_n \cdot X_n$
 - - Y represents the predicted target variable.
 - b_0 is the intercept.
 - b_1, b_2, \dots, b_n are the coefficients of the independent variables X_1, X_2, \dots, X_n .
 - **Fitting:** The coefficients ($b_0, b_1, b_2, \dots, b_n$) are estimated using techniques like Ordinary Least Squares (OLS), which minimizes the sum of squared differences between the actual and predicted values.
 - **Assumptions:** Linear regression assumes that there is a linear relationship between the variables, the errors are normally distributed, and there is no multicollinearity or heteroscedasticity.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It illustrates the importance of visualizing data and not relying solely on summary statistics. Key points:

- All four datasets have the same mean, variance, correlation, and linear regression line.
- However, when you plot these datasets, you'll see significant differences in their distributions and relationships.
- Anscombe's quartet highlights that summary statistics alone can be misleading and may not capture the full complexity of the data.

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient (often denoted as Pearson's R) is a statistic used to measure the linear relationship between two continuous variables. It ranges from -1 to 1:

- If $R = 1$, it indicates a perfect positive linear relationship.
- If $R = -1$, it indicates a perfect negative linear relationship.
- If $R = 0$, it indicates no linear relationship.

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It quantifies the strength and direction of a linear association between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling in the context of machine learning refers to the process of transforming features to have a consistent scale. It is performed to:

- Ensure that different features contribute equally to the model.
- Prevent features with larger scales from dominating those with smaller scales.
- Facilitate the convergence of gradient-based optimization algorithms.

Normalized scaling (also known as Min-Max scaling) scales features to a specific range, typically $[0, 1]$. It is calculated as:

```
scss
X_normalized = (X - X_min) / (X_max - X_min)
```

Standardized scaling (Z-score scaling) scales features to have a mean of 0 and a standard deviation of 1. It is calculated as:

```
makefile
X_standardized = (X - X_mean) / X_std
```

Normalized scaling retains the original distribution but scales it, while standardized scaling centers the data around the mean and adjusts for the spread.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF (Variance Inflation Factor) measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity in the data. VIF can become infinite when a predictor variable can be perfectly predicted by a linear combination of other predictor variables.

In other words, when one predictor variable is a linear combination of others, its VIF becomes infinite because its variance is infinitely inflated by the multicollinearity. This happens because the matrix used in the VIF calculation becomes singular, making it impossible to compute the exact value of VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, usually the normal distribution. Here's how it works:

- In a Q-Q plot, the x-axis represents the quantiles (sorted values) from the theoretical distribution (e.g., a normal distribution), and the y-axis represents the quantiles of the observed data.
- If the points on the Q-Q plot approximately follow a straight line (the 45-degree line), it suggests that the data closely follows the theoretical distribution.

Importance in linear regression:

- Q-Q plots are used to check the assumption of normality of residuals in linear regression. Residuals should ideally follow a normal distribution for the linear regression model to be valid.
- If the Q-Q plot deviates from a straight line, it indicates that the residuals are not normally distributed. This can help identify potential issues in the model, such as heteroscedasticity or outliers.
- Q-Q plots are valuable for diagnosing and improving the accuracy of linear regression models by highlighting deviations from normality assumptions in the residuals.

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? (3 marks) 2. Why is it important to use `drop_first=True` during dummy variable creation? (2 mark) 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark) 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks) 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ChatGPT

Here are the answers to the 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? (3 marks)

In the analysis of categorical variables, we typically perform one-hot encoding or create dummy variables to convert categorical data into a format suitable for regression models. Here's what can be inferred about their effect on the dependent variable:

- **Effect of Categorical Variables:** The effect of categorical variables on the dependent variable (demand for shared bikes) can be inferred by examining the coefficients of the dummy variables in the linear regression model. Each coefficient represents the change in the target variable associated with a one-unit change in the corresponding categorical variable, while holding all other variables constant.
- **Significance:** The significance of each categorical variable can be determined by the p-value associated with its coefficient. A low p-value (typically less than 0.05) suggests that the categorical variable is statistically significant in predicting the target variable.
- **Magnitude and Direction:** The magnitude (positive or negative) of the coefficients indicates whether a particular category of a categorical variable has a positive or negative impact on bike demand compared to the reference category (the category dropped during dummy variable creation).
- **Interpretation:** For example, if 'season' is a categorical variable with four categories (spring, summer, fall, winter), and we create dummy variables, we can interpret the coefficients as follows: A positive coefficient for 'summer' indicates that, on average, there is an increase in bike demand during the summer season compared to the reference season (e.g., spring). Conversely, a negative coefficient for 'winter' suggests a decrease in bike demand during the winter season compared to the reference season.

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

Using `drop_first=True` during dummy variable creation is important for avoiding multicollinearity in regression models. Here's why:

- **Multicollinearity:** Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. This can cause several issues:
 - It makes it challenging to isolate the individual effect of each independent variable on the target variable.

- It inflates the standard errors of the coefficient estimates, leading to unreliable p-values and confidence intervals.
- It can result in unstable and less interpretable models.
- **The Role of `drop_first=True`:** When you set `drop_first=True` during dummy variable creation, it automatically drops one of the dummy variables for each categorical feature. This is done to avoid perfect multicollinearity, where one dummy variable can be perfectly predicted from the others.
- **Example:** Suppose you have a categorical variable 'season' with four categories (spring, summer, fall, winter). If you create dummy variables without dropping the first one, you'll have four dummy variables (e.g., 'season_spring', 'season_summer', 'season_fall', 'season_winter'). By setting `drop_first=True`, one dummy variable (e.g., 'season_spring') is dropped, and the model is fitted with three dummy variables. This prevents perfect multicollinearity because you can always deduce the value of the dropped category from the others (e.g., if 'season_summer', 'season_fall,' and 'season_winter' are all 0, it implies 'season_spring' is 1).

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

To identify which numerical variable has the highest correlation with the target variable 'cnt,' you can visually inspect the pair-plot or calculate the correlation coefficients. The variable with the highest absolute correlation coefficient is the one with the strongest linear relationship with 'cnt.'

You can use the following code to calculate and print the correlation coefficients with 'cnt' for each numerical variable:

```
python
import pandas as pd

# Assuming df is your DataFrame containing the dataset
correlations = df.corr()['cnt'].abs().sort_values(ascending=False)
print(correlations)
```

The variable with the highest absolute correlation coefficient (closer to 1) is the one with the highest correlation with 'cnt.'

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

Validating the assumptions of Linear Regression is crucial to ensure the model's reliability. After building the model on the training set, you can validate the assumptions using several techniques:

- **Residual Analysis:**
 - **Check Linearity:** Plot the residuals against the predicted values (fitted values). The residuals should be randomly scattered around 0; there should be no discernible pattern.
 - **Check Homoscedasticity:** Plot the residuals against the predicted values or independent variables. Look for consistent spread of residuals across the range of predicted values; there should be no fan or cone shape.

- **Check Normality:** Create a Q-Q plot or histogram of residuals to check if they approximately follow a normal distribution. Alternatively, perform a statistical test like the Shapiro-Wilk test.
- **Multicollinearity:** Calculate Variance Inflation Factors (VIF) for each independent variable. High VIF values indicate multicollinearity, which can be problematic. Consider removing highly correlated variables or using dimensionality reduction techniques.
- **Independence of Errors:** Ensure that there is no autocorrelation in the residuals. You can use the Durbin-Watson statistic to check for autocorrelation.
- **Outliers and Influential Points:** Identify outliers and influential data points by examining studentized residuals, leverage, and Cook's distance. Outliers can distort the model, and influential points can heavily influence the regression coefficients.
- **Heteroscedasticity:** Use statistical tests like the Breusch-Pagan test or White test to check for heteroscedasticity. If detected, consider applying heteroscedasticity-robust standard errors or transforming the data.
- **Feature Engineering:** If necessary, perform feature engineering to create polynomial features or interactions to better capture non-linear relationships.
- **Cross-Validation:** Apply k-fold cross-validation to evaluate the model's performance on multiple subsets of the data. This helps assess the model's generalizability.
- **R-squared and Residual Sum of Squares:** Calculate the R-squared score and residual sum of squares on a validation set to assess how well the model fits the data.

Address any violations of assumptions by making necessary adjustments to the model. Iterative testing and refinement of the model may be needed.

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

To identify the top 3 features contributing significantly to explaining the demand for shared bikes, you can examine the coefficients of the features in your final linear regression model. The magnitude of the coefficients indicates the strength of their contribution. Here's how you can determine the top 3 features:

```
python
# Assuming you have fitted a linear regression model named 'model'
coefficients = model.coef_
feature_names = df.columns[:-1] # Exclude the target variable 'cnt'
# Create a DataFrame to pair features with their coefficients
coef_df = pd.DataFrame({'Feature': feature_names, 'Coefficient':
coefficients})

# Sort the DataFrame by absolute coefficient values in descending order
```