Bike Sharing Case Study & Linear Regression

Vasudha Srinivasaiah
IIIT-B Machine Learning & AI Program -Dec-2023
3/13/24

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?

- The demand for bikes grew in 2019 compared to 2018.
- Bicycle demand is highest between June and September. January is the lowest demand month.
- Bike demand is lower during vacations than during non-holiday periods.
- The demand for bikes is high in Summer and lower in Spring
- The demand for bikes is almost constant throughout the week.
- There is no substantial difference in bike demand between working and non-working days.
- Bike demand is high when the weather is clear and there are few clouds, but it is low when there is light snow and rain. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow, and Fog, so we cannot draw any conclusions. Perhaps the company is not working on those days, or there is little demand for the bike.

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

To prevent multicollinearity problems in regression analysis and to guarantee accurate interpretation of the model coefficients, the drop_first=True parameter must be used while creating dummy variables.

An effective reference category is created by removing one of the dummy variables, which is typically the first one, to which the remaining categories are compared. This facilitates a more intuitive interpretation of the dummy variable coefficients. The remaining categories' coefficients show how they vary from the discarded category, which now serves as the baseline or reference level.

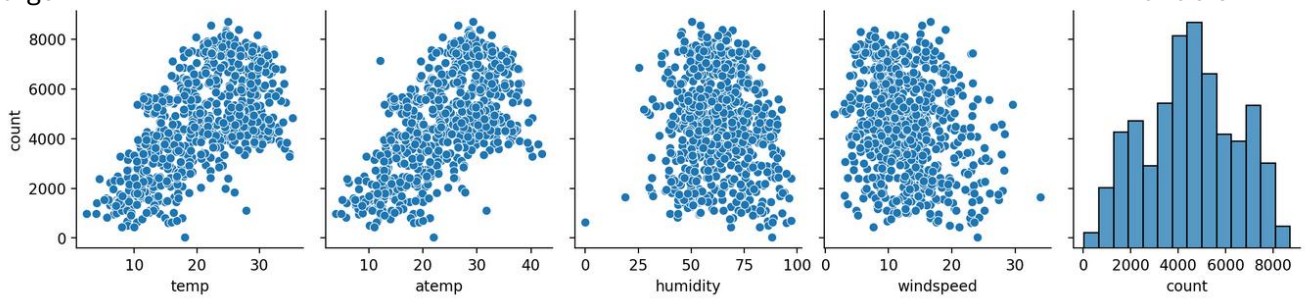
For example, consider the 'BloodGroup' variable with categories 'A', 'B', 'AB', and 'O'. There will be four dummy variables in total—one for each category—if you create new ones without deleting the initial one. Instead, you establish three dummy variables ('BloodGroup_B, 'BloodGroup_AB, and 'BloodGroup_O') that indicate whether the blood group is 'B, 'AB, or 'O' in relation to the reference category 'A' by removing the first one.

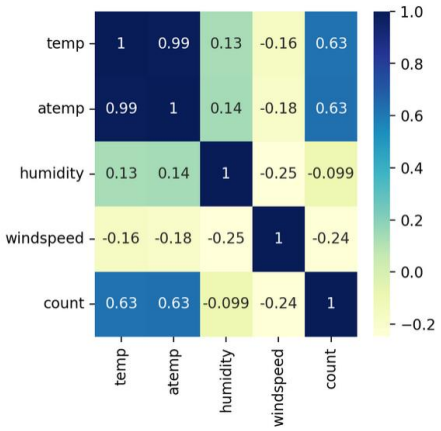
This guarantees that multicollinearity problems are avoided while maintaining the model's identity and interpretability. Additionally, it makes the interpretation of the model coefficients easier by offering a distinct benchmark for comparison.

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

From the pair plots, it's observed that there is a linear relation between temp, atemp, and count. Numerical variables, *atemp*, and *temp* has the highest correlation coefficient value =0.63 with the

target variable.





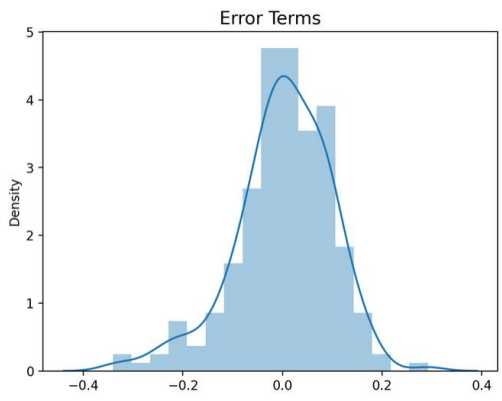
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

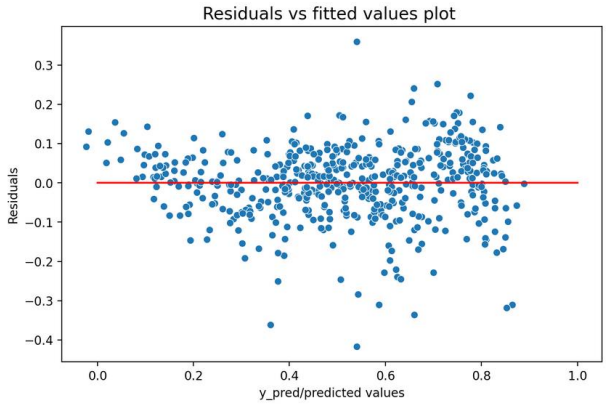
Primary assumptions of linear regression are:

a) There is a linear relationship between the dependent and independent variables. And the independent variables are not highly correlated with each other.
b) Residual Errors have a mean value of zero
c) Residual Errors have constant variance
d) **Homoscedasticity**: The variance of the errors is constant across all levels of the independent variables.

In the study, the assumptions were checked using the

- **Residual vs Fitted plot**. Ideally, this plot would not have a pattern where the red line is approximately horizontal at zero. In this case it's pretty flat, which provides evidence that a linear model is reasonable. If the regression assumptions are met, the plot to get a flat line, as any slice of your residuals should be mean zero (and often normally distributed). The same is observed in the plots.
- **Normal Distribution Plot**. The Error terms are checked with a normal distributed plot which should ideally have the mean value at zero.
- **Out[63] and Out[69] shows the above plots**. And this seems to validate the assumptions.





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

Based on the R-squared and adj R-squared values of the train and test datasets, we may deduce that the variables mentioned above account for about 80% of the demand for bicycles. The variables' coefficients explain the factors influencing the demand for bikes.

- According to the final model, the Top Predictor variables influencing bike booking are :
  - Temperature (temp)
  - Year
  - Seasons Summer, Winter, Spring
  - Windspeed

- Overall variables influencing the bike booking are. Windspeed, Temperature (temp), Year, Seasons Spring, Months September, Seasons Summer, Season Winter, If its a holiday, If the weather is Clear.

To maximize demand, it is advised that these aspects be given the utmost consideration throughout planning.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as y) and one or more independent variables (often denoted as X). Linear regression aims to find the best-fitting linear relationship between the independent variables and the dependent variable. This relationship is represented by a linear equation of the form:

$$y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\epsilon$$

where:

- y is the dependent variable (the variable we want to predict),
- X_1, X_2,..., X_n are the independent variables (features or predictors),
- $\beta_0,\beta_1,...,\beta_n$ are the coefficients (parameters) of the model,
- ϵ is the error term (residuals), representing the difference between the observed and predicted values.

The coefficients $\beta_0,\beta_1,...,\beta_n$ are estimated from the training data using a method called ordinary least squares (OLS) or other optimization techniques. The objective of linear regression is to minimize the sum of the squared differences between the observed values of the dependent variable and the values predicted by the linear equation.

The process included under the linear regression algorithm is listed below:

1. **Data Preparation**: It is necessary to compile the dataset, which includes observations of the independent variables (X) and dependent variables (y). The variables are scaled suitably when necessary, the dataset is cleaned, and missing values are handled.
2. **Model Fitting**: Using a dataset comprising m observations and n independent variables, the linear regression algorithm calculates the coefficients ($\beta_0,\beta_1,...,\beta_n$) that reduce the total squared residuals. The ordinary least squares (OLS) approach is commonly used for this, as it identifies the coefficients that minimize the sum of the squared differences between the observed and predicted values.
3. **Model Evaluation**: After the model has been fitted to the training set, assess its performance with a variety of measures, including mean absolute error (MAE), root mean square error (RMSE), adjusted R-squared, and R-squared. These metrics evaluate the model's capacity to generalize to new data and how well it fits the training set.
4. **Prediction**: Utilize the fitted model to forecast fresh, unobserved data. Using the learnt coefficients, the model forecasts the dependent variable's values based on the values of the independent variables.
5. **Model Interpretation**: To comprehend the link between the independent and dependent variables, interpret the linear regression model's coefficients. When all other variables are held constant, the coefficients show how the dependent variable changes for every unit change in the corresponding independent variable.

A popular and easily understood technique for simulating the relationship between variables is linear regression. In a variety of disciplines, including economics, finance, the social sciences, and machine learning, it is frequently utilized for predictive modeling, hypothesis testing, and comprehending the underlying correlations between variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a well-known statistical example that highlights the value of data visualization as well as the drawbacks of depending only on summary statistics. It comprises of four datasets that, when displayed visually, show radically distinct patterns despite having almost similar summary statistics (mean, variance, correlation, and linear regression coefficients).

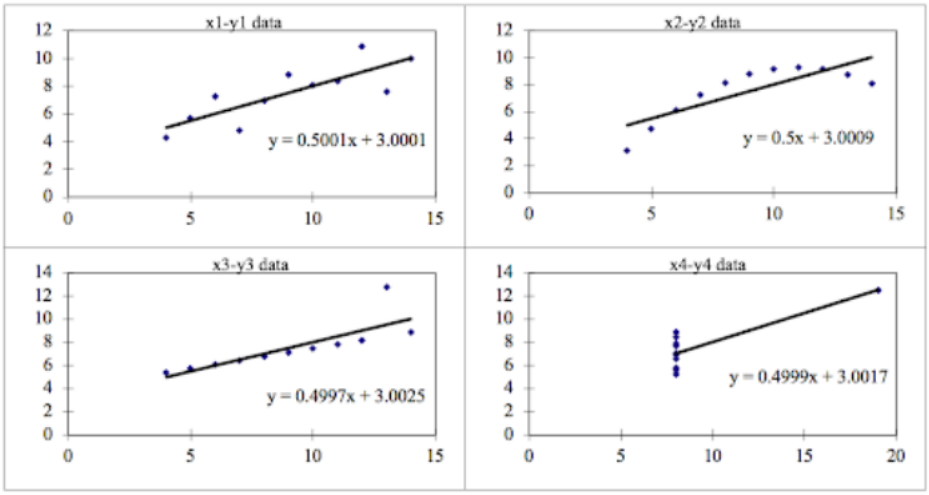
The statistician Francis Anscombe created the quartet in 1973 to highlight the risks associated with making decisions based only on numerical summaries without physically inspecting the facts. Anscombe's quartet is frequently cited as an example of the ideas behind outliers, pivotal moments, and the importance of exploratory data analysis.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Here are the characteristics of each dataset in Anscombe's quartet:

- **Dataset I**:
    a. Linear relationship: The data points form a perfect linear relationship.
    b. Summary statistics: Mean of x = 9, mean of y = 7.50, variance of x ≈ 11, variance of y ≈ 4.12, correlation coefficient ≈ 0.816, linear regression line: y=3+0.5xy=3+0.5x.
    c. Graphical pattern: The data points follow a clear linear trend, with minimal variability around the regression line.
- **Dataset II**:
    a. Non-linear relationship: The data points form a non-linear relationship with a clear outlier.
    b. Summary statistics: Mean of x = 9, mean of y = 7.50, variance of x ≈ 11, variance of y ≈ 4.12, correlation coefficient ≈ 0.816, linear regression line: y=3+0.5xy=3+0.5x.
    c. Graphical pattern: Despite having the same summary statistics as Dataset I, Dataset II shows a non-linear relationship with most data points clustered around a quadratic curve. One outlier significantly influences the regression line.
- **Dataset III**:
    a. Linear relationship with an outlier: The data points form a linear relationship with one outlier.
    b. Summary statistics: Mean of x = 9, mean of y = 7.50, variance of x ≈ 11, variance of y ≈ 4.12, correlation coefficient ≈ 0.816, linear regression line: y=3+0.5xy=3+0.5x.
    c. Graphical pattern: Similar to Dataset I, but with one outlier that influences the regression line and correlation coefficient.
- **Dataset IV**:
    a. No apparent relationship: The data points do not exhibit a clear linear relationship.
    b. Summary statistics: Mean of x = 9, mean of y = 7.50, variance of x ≈ 11, variance of y ≈ 4.12, correlation coefficient ≈ 0.817, linear regression line: y=3+0.5xy=3+0.5x.

c. Graphical pattern: Despite having similar summary statistics and linear regression parameters as Datasets I, II, and III, Dataset IV consists of points with widely varying x and y values, indicating no meaningful relationship.



Learnings from Anscombe's quartet include :

- The significance of visualizing data to comprehend its underlying patterns is emphasized in Anscombe's quartet.
- Although the datasets' summary statistics are comparable, their graphical patterns are very different.
- It is possible to get the wrong inferences about the links between the data if you only rely on summary statistics.
- To find trends and comprehend the subtleties of the data, exploratory data analysis—including data visualization—is crucial.

3. What is Pearson's R?

Pearson's correlation coefficient often denoted as **r** or **Pearson's r**, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It measures the degree to which two variables are linearly related to each other.

Pearson's correlation coefficient r ranges from -1 to 1:

- r=1: Perfect positive linear correlation. As one variable increases, the other variable also increases proportionally.
- r=−1: Perfect negative linear correlation. As one variable increases, the other variable decreases proportionally.
- r=0: No linear correlation. There is no linear relationship between the variables.

The formula for Pearson's correlation coefficient between two variables X_i and YY is:

$$r = \frac{\sum_{i-1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i-1}^{n}(X_i-\bar{X})^2}\times\sqrt{\sum_{i-1}^{n}(Y_i-\bar{Y})^2}}$$

Where:

- X_i and Y_i are the individual data points of variables XX and YY, respectively.
- X^- and Y^- are the means of variables X and Y, respectively.
- n is the number of data points.

To determine the degree of link between variables, Pearson's correlation coefficient is widely utilized in a variety of domains, including statistics, the social sciences, economics, and machine learning. It is very helpful for evaluating the predictive capacity of models and figuring out the direction and strength of linear correlations between continuous variables. It is crucial to remember that Pearson's rr can only assess linear relationships; it may not be able to identify other kinds of associations or non-linear relationships.

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

Scaling is the process of converting the values of variables in a dataset into a given range or distribution. It is a popular preprocessing procedure for features (independent variables) in machine learning and statistical modelling. Scaling guarantees that all variables have similar scales and distributions, which can boost the performance and stability of different algorithms.

Most of the time, the acquired data set includes features with widely variable magnitudes, units, and ranges. If scaling is not done, the method simply considers magnitude rather than units, resulting in erroneous modelling. To solve this problem, we need to scale all of the variables to the same magnitude. Scaling only impacts the coefficients and not the other factors such as t-statistic, F-statistic, p-values, R-squared, etc.

The significance of Scaling to the model includes below listed points:

- **Improves Algorithm Performance:** The size of features affects several machine learning methods. Biased model training might result from features with larger scales predominating over those with lower sizes. Scaling helps to ensure that all features contribute equally to the model.
- **Speeds up Convergence:** By preventing features with enormous scales from controlling the optimization process, scaling can aid gradient-based optimization algorithms in convergent convergence more quickly. This holds special significance for techniques like gradient descent.
- **Interpretability:** In linear models, scaling facilitates the interpretation of the coefficients or feature weights. Comparing the coefficient magnitudes becomes difficult when features are on different scales.
- **Regularization:** For regularization strategies like L1 and L2 regularization, scaling can be crucial. Scaling guarantees that the penalties are applied consistently to all features because these methods punish large coefficients.

Difference between Normalized Scaling and Standardized Scaling

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in Python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Normalization preserves the relative relationships between the values of each feature but does not handle outliers well

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**).

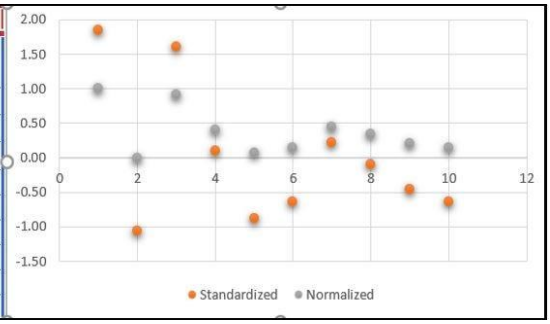
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in Python.
- Standardization ensures that features have a similar scale and distribution, making it robust to outliers. It does not preserve the original scale of the data but is commonly used in practice due to its robustness.

One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Example:
Below is an example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

The Variance Inflation Factor (VIF) assesses the level of multicollinearity in a regression model. Multicollinearity happens when the independent variables in a regression model are substantially linked with one another. It quantifies how much the variance of the estimated coefficients is inflated due to multicollinearity among the predictor variables.

When the value of VIF is infinite, it signals a serious multicollinearity issue. This usually occurs when one or more independent variables in the model are fully correlated with one another, which means they can be written exactly as a linear combination of the other variables.

In other words, perfect multicollinearity occurs when there is a linear relationship among the independent variables that can be expressed as $X_i = c_1X_1 + c_2X_2 + ... + c_iX_i$ for some constants $c_1, c_2, ..., c_i$ where X_i is one of the independent variables and $X_1, X_2, ..., X_i - 1$ are the other independent variables in the model.

When the value of VIF is infinite, it indicates perfect multicollinearity between the predictor variable X_i and other predictor variables in the model. Perfect multicollinearity occurs when one or more predictor variables can be exactly predicted by a linear combination of the other predictor variables.

There are several reasons why the VIF may be infinite:

- A Perfect Linear Relationship occurs when one or more predictor variables in a model are completely linearly related to one another. For example, one variable can be described as a linear combination of the others, whereas two variables are identical or almost similar.
- Redundant Variables: Predictor variables that don't add anything to the model's existing information. This can occur when one variable is a constant multiple of another, or when there is significant redundancy across predictor variables.
- Rounding errors or floating-point accuracy restrictions in calculations can result in infinite VIF values. This is less prevalent, although it can happen in some computing contexts.

  When VIF values are infinite, it indicates a severe multicollinearity problem that needs to be addressed. Perfect multicollinearity can lead to unstable coefficient estimates and inflated standard errors, interpreting the regression results as unreliable. To address this issue, it may be necessary to identify and remove redundant variables or transform them to reduce multicollinearity before fitting the regression model.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. In statistics, it is a probability plot that compares two probability distributions by showing their quantiles against one another. A point (x, y) on the plot represents one of the second distribution's quantiles (y-coordinate) displayed against the corresponding quantile from the first distribution (x-coordinate). This defines a parametric curve, with the quantile interval index serving as the parameter. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, allowing one to visually inspect how closely the data aligns with the theoretical distribution

The use and importance of Q-Q plots in linear regression include:

- Q-Q plots check if a regression model's residuals (errors) follow a specific distribution, such as the normal distribution. Linear regression models frequently assume that the residuals are normally distributed, so verifying this assumption is critical to the validity of the regression results.
- Visually inspecting the Q-Q graphic reveals deviations from the assumed distribution. If the points depart greatly from the identity line, it indicates that the residuals do not conform to the stated distribution. This implies that the regression model may have difficulties, such as heteroscedasticity or non-normality of errors.
- Q-Q charts are one of a variety of diagnostic methods used to assess the appropriateness of a regression model. They are used in conjunction with other diagnostic approaches such as residual plots, leverage plots, and Cook's distance to discover flaws with model assumptions and influential data points.
- If the Q-Q plot shows deviations from the predicted distribution, you may need to consider model changes, such as modifying the response variable or adding more predictors to account for errors' non-normal distribution.

To sum up, Q-Q plots are useful instruments for evaluating the distributional assumptions of linear regression models and identifying any model problems. They support the validity and dependability of regression analysis by giving a visual depiction of how well the observed data matches the theoretical distribution.