Linear Regression Assignment

-Anil Kumar Narayanan

Date: July 31’ 2024

Assignment-based Subjective Question

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

   Based on the box plots and the correlations observed, here are some reasons about the effect of categorical variables on the dependent variable 'cnt':

a. Season: Seasonality plays a significant role, with fall (season_3) showing the highest rentals and winter (season_1) the lowest. This suggests people rent bikes more in pleasant weather.
b. Year: The upward trend in rentals from 2018 (yr_0) to 2019 (yr_1) indicates a growing popularity of bike rentals over time.
c. Month: Rentals peak in summer months (June-August) and decline in winter months, further supporting the influence of weather on bike usage.
d. Holiday: Slightly lower rentals on holidays (holiday_1) suggest people might prefer leisure activities other than biking on holidays.
e. Weekday: Consistent rentals across weekdays imply that daily commuting might not be the primary driver of bike rentals.
f. Working Day: Slightly higher rentals on working days (workingday_1) could indicate some usage for work commutes, but the effect isn't very strong.
g. Weather: Adverse weather conditions (weathersit_3) significantly reduce rentals, highlighting the importance of good weather for bike riding.

   In conclusion, weather (represented by season, month, and weathersit) appears to be the most dominant factor influencing bike rentals.

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

   Setting drop_first=True helps in preventing multicollinearity. In the bike-sharing case-study, the categorical variable season has four values like 1 (Spring), 2 (Summer), 3 (Fall), and 4 (Winter). With drop_first=True, the resulting data frame included dummy columns for 2 (Summer), 3 (Fall), and 4 (Winter) thereby dropping the column for 1 (Spring).

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

Based on the pair plots and correlation matrix among the numerical variables and the target variable (cnt) it visually appears that temp and atemp have a strong linear relationship with cnt.

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

   After building the model, the assumptions were validated by performing residual analysis.

- **Residuals vs. Predicted Values Plot:** This plot checked for linearity and it does not show a clear non-linear pattern and the residuals are spread around the horizontal line (y=0) indicating the assumptions are met.
- **Distribution of Residuals:** By plotting the histogram of residuals, we checked for normality. The plot shows that the residuals follow a roughly normal distribution with a bell-curve indicating this assumption is met.

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

   Overall, the model identified the below three features:

- temperature and weather conditions as significant predictors of bike demand. This indicates that bike usage is highly sensitive to changes in temperature and weather, suggesting that people are more likely to use bikes in favourable weather conditions.
- Seasonal variations significantly affect bike demand, with higher demand in certain seasons. This implies that bike usage patterns change throughout the year, peaking during warmer or more pleasant weather months.
- Working days and holidays have a notable impact on bike demand. Noticed higher demand on working days for commuting purposes and different on holidays for leisure activities.

General Subjective Questions

6. Explain the linear regression algorithm in detail?

Linear Regression is mapping of linear relationship between two variables i.e a method to predict dependent variable (Y) based on values of independent variables (X) and estimate the strength and direction of the relationship between two or more variables.

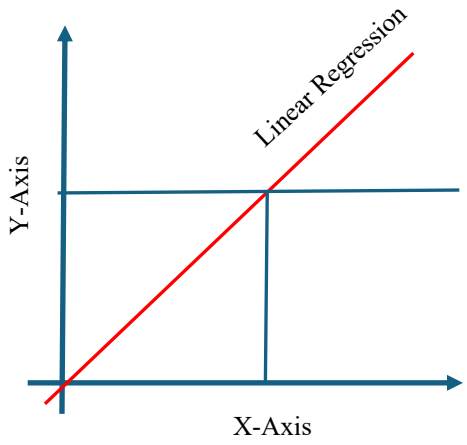

Fig. 1

The method of **least squares** is the standard approach in regression analysis. In this approach, the sum of the squares of the residuals made in the results of every single equation. The objective consists of adjusting the parameters of a model function to best fit a data set. The formula for the residual for each observation is:

Residual = observed - predicted

It is an iterative process to find the most optimal coefficients using optimization techniques such as gradient descent. This involves adjusting the coefficients to minimize the cost function, usually the Mean Squared Error (MSE).

7. Explain the Anscombe's quartet in detail?

Anscombe's quartet is a collection of four datasets that illustrate the importance of data visualization in statistical analysis.

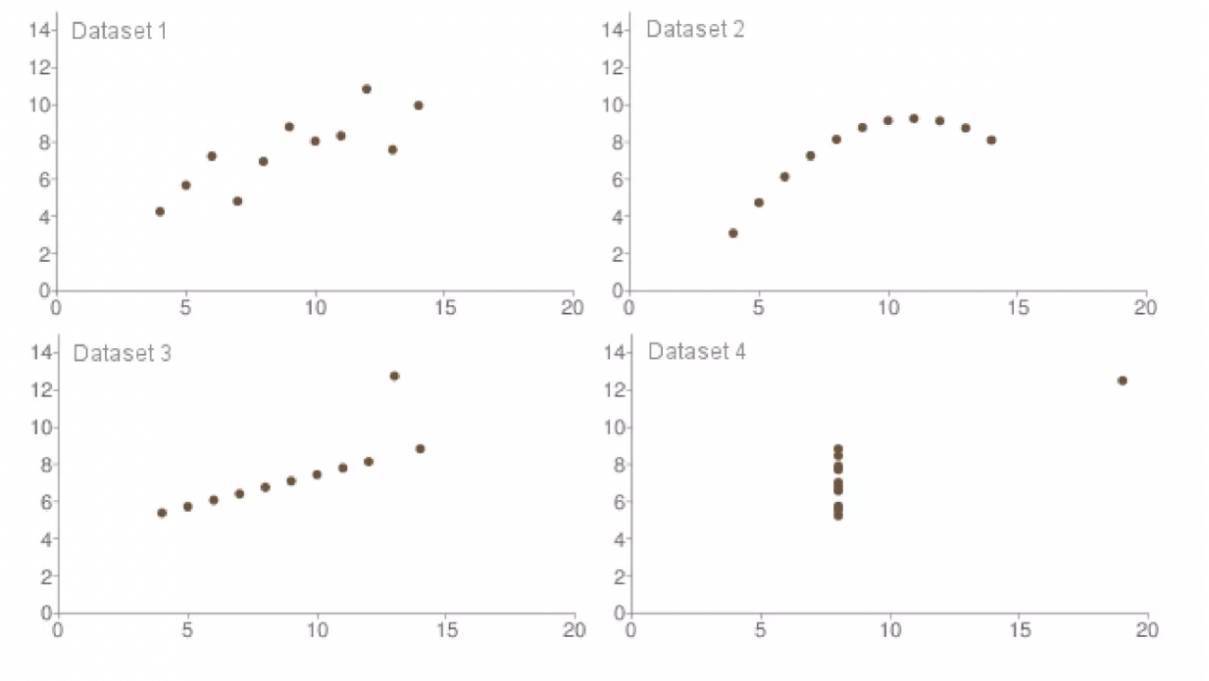

Fig. 2

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.80

For all four datasets:

N = 11
mean of *x* values = 9.0
variance of *x* values = 11.0
mean of *y* values= 7.5
variance of *y* values = 4.12
correlation between x & y = 0.816
regression line: $y = 3 + 0.5x$

Fig. 3

As noticed in the above figures, the statistical properties identical are Mean, Variance of x, Mean of y, Variance of y, Correlation between x & y and linear regression line. These identical statistics can lead to misleading conclusions if one only relies on numerical summaries without visualizing the data. To elaborate this further, in the above provided four datasets:

Dataset 1:

This dataset shows a strong linear relationship between x and y. The points cluster around a straight line, making it suitable for linear regression analysis.

Dataset 2:

This dataset exhibits a clear non-linear relationship. The points form a curve, indicating that a linear model would not adequately describe the relationship between x and y.

Dataset 3:

In this dataset, there is a linear relationship, but with an outlier. Most points lie along a line, but the outlier affects the slope of the regression line.

Dataset 4:

Like Dataset 3, this dataset features a linear relationship but is affected by a high-leverage point.

8. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables and is commonly used in understanding relationships between predictors and outcomes. The formula used is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Fig. 4

- r = Coefficient of correlation
- xbar = Mean of x-variable
- ybar = Mean of y-variable.
- xi yi = Samples of variable x, y

The value of Pearson's R ranges from -1 to +1:

- +1: Perfect positive correlation (both variables increase together)
- -1: Perfect negative correlation (one variable increases while the other decreases)
- 0: No correlation (no linear relationship)

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

   Scaling is a technique to standardize the independent features present in the data. It is performed during the data pre-processing phase before creating a machine learning model. Scaling is performed when:

- The real-world datasets include features that highly vary in magnitudes, units and range
- When a feature in the dataset is big in scale compared to others in algorithms where Euclidean distance is measured which can become a dominating factor and needs to be normalized.
- Algorithms relying on gradient descent converge faster with features scaled.

There are two scaling techniques:

1. Standardization: In this technique, the values are centered around the mean with a unit standard deviation. The values are replaced by their Z scores. The formula is as below where x_i is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature:

$$z = \frac{x_i - \mu}{\sigma}$$

Fig. 5

2. Normalization: Here, values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. The formula is as below where x is the original value, x_{min} is the minimum value of the feature, and x_{max} is the maximum value of the feature.

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

Fig. 6

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

    Variance Inflation Factor is infinite when there is perfect multicollinearity between two or more independent variables in a regression model. The regression equation cannot uniquely estimate the coefficients for these variables.

    It can also occur if the dataset includes redundant variables i.e variables that provide the same information.

    Another reason is when the number of predictors exceeds the number of observation where R^2 can reach 1 for many variables.

    Additionally, data preprocessing steps can inadvertently introduce perfect multicollinearity. For example, when creating dummy variables for categorical variables without dropping one category (the “dummy variable trap”), we may end up with perfect correlations.

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

    A Q-Q plot stands for quantile-quantile plot and is a probability plot for assessing how closely two datasets agree. The Q-Q plot is formed with:

- dataset 1 on the vertical axis and
- dataset 2 on the horizontal axis.

    Both the above given axes are plotted based on the units of datasets in the given system.

    The plots are used to assess the normality assumption of the residuals. If the residuals are normally distributed, the points in the Q-Q plot should approximately form a straight line.

    The importance of checking the normality assumption using Q-Q plots lies in the following:

- Validity of Inference: If the residuals are not normally distributed, the standard errors and confidence intervals may not be valid, leading to incorrect inferences about the model parameters.
- Outlier Detection: These plots can help identify outliers in the data by revealing points that deviate significantly from the expected straight line.
- Transformation Guidance: If the plot suggests non-normality, it can guide the choice of transformations to improve the normality of the residuals.