Linear Regression Assignment :-

Bike Sharing Case Study- Subjective Questions

Submitted By - Kunal Arneja

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)

- Bookings are more in Fall, Summers when compared to Spring while they are average in Winters.
- Bookings are higher on non-holidays compared to holidays.
- Bookings are significantly lower in case of Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
- Bookings show an increasing Trend from JAN to JUL, flatten from JUL to SEP and then drop from OCT to DEC.

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

It helps in reducing the extra column when we one hot encode a categorical variable.
For example if a categorical feature has 4 unique labels, we would need one 3 new columns.
These three columns all being 0 will imply the value being of 4th label.

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

Temperature has highest correlation as the graph has a linear trend.

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

Linear regression has five key assumptions:

- Linear relationship - Looking at the scatter Plots, Box Plots. We saw linear relationship between most of the dependent variables and our target variable.
- Multivariate normality - The variable are Multivariate normal looking at the histograms of the variables. Most of them follow similar to a Normal Distribution.
- No or little multicollinearity- Independent variables are not highly correlated with each other, looking at the correlation Matrix.

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

Top 3 features Contributing towards demand are :-

- Temperature
- Year
- Weathersit_ Light Snow, Light Rain

General Subjective Questions

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

Linear regression is a statistical technique to understand the relationship between one dependent variable and one or more independent variables. The objective is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Y = B0 + B1X1 + B2X2 + B3X3 ….

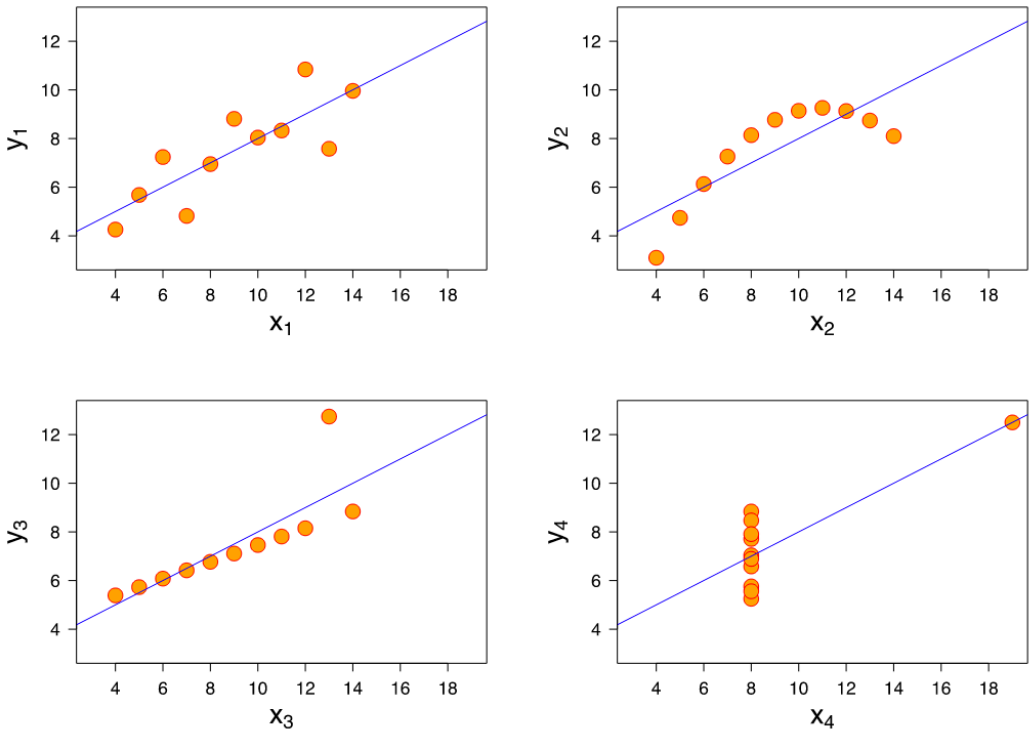
The overall idea of regression is to examine two things:
(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
(2) (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

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

Anscombe's is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed.
All the summary statistics you'd think to compute are close to identical:

- The average *x* value is 9 for each dataset
- The average *y* value is 7.50 for each dataset
- The variance for *x* is 11 and the variance for *y* is 4.12
- The correlation between *x* and *y* is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation *y = 0.5x + 3*

These four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:

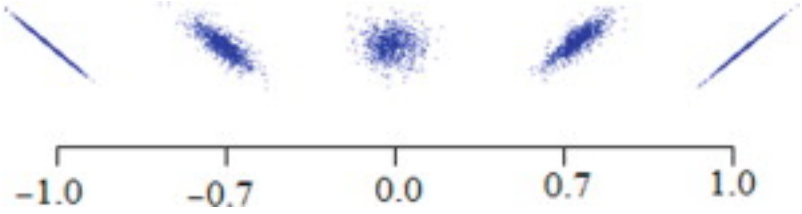


Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier. Dataset IV looks like *x* remains constant, except for one outlier as well.
Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

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

The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between – 1 and 1, where 0 is no correlation, 1 is total positive correlation, and – 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

Considering the two variables "age" and "salary," a strong positive correlation between the two would be expected: as people get older, they tend to earn more money. Therefore, the correlation between age and salary probably gives a value over 0.7. Figure 6.2 illustrates pairs of numerical variables plotted against each other, with the corresponding correlation value between the two variables shown on the x-axis. The right-most plot shows a perfect positive correlation of 1.0, whereas the middle plot shows two variables that have no correlation whatsoever between them. The left-most plot shows a perfect negative correlation of – 1.0.



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

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

- Ease of interpretation
- Faster convergence for gradient descent methods You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

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-square, etc.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
To check this sort of relations between variables, we use VIF. VIF basically helps explaining the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity.
In our dataset, temp and temp showed almost linear perfect correlation, The VIF in the case would have even close to infinity as R2 would have been close to 1.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
This helps in a scenario of linear regression 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.
Few advantages:
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
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.