Assignment-based subjective questions:

1. Categorical variables that have positive effects on the dependent variables:

Workingday, summer, winter, 2019, May, September, Sunday, weathersit1, weathersit2, week2 and week3

Categorical variables that have negative effects on the dependent variables:

Holiday, spring, December, Febuary, January, and November

2. During one-hot encoding, we only need to create n-1 dummy variables since the last dummy variable can be derived from others. This way, we can save computation power.

3. atemp

4. (1) plot the histogram of the error terms to verify normal distribution with mean of zero

(2) Calculate the homoscedasticity of error terms and see if the residuals have relative constant variance to verify the homoscedasticity assumption

5. (1) whether the year is 2019 (2) whether the weathersit is 1 (3) whether the weathersit is 2

General subjective questions

1. Linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent and dependent variable. Based on the given data points, we try to plot a line that models the points the best.

The motive of the linear regression algorithm is to find the best values for the intercept and parameters. To achieve this, we minimize the cost function (Mean Squared Error).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Or using gradient descent which updates parameters to reduce the cost function. The idea is that we start with some values for the parameters and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

2. Anscombe's Quartet demonstrates the dangers of summary statistics. It's 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 are close to identical:

-The average x value is 9 for each dataset

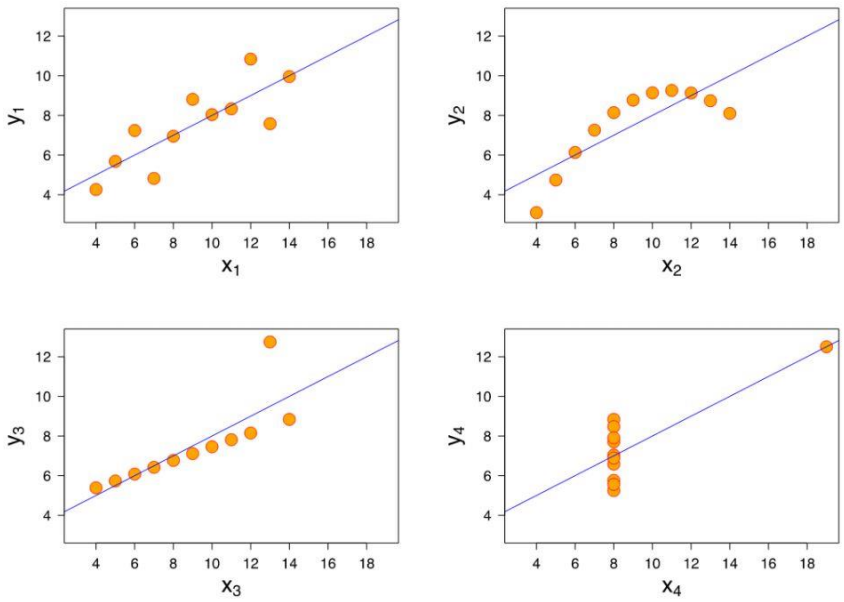
-The average y value is 7.50 for each dataset

-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

The plot shows as follows:



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. Pearson's R is a correlation coefficient commonly used in linear regression which measures how strong a relationship is between two variables.

Formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

The formulas return a value between -1 and 1, where:

-1 indicates a strong positive relationship.

- -1 indicates a strong negative relationship.

- A result of zero indicates no relationship at all.

4. Feature scaling is a method used to normalize the range of independent variables or features of data.

Most of the times, the dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem.

If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Normalized scaling is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardized scaling transforms the data to have zero mean and a variance of 1, they make our data unitless.

Normalization: min-max scaler standardization: standard scaler

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad x_{new} = \frac{x - \mu}{\sigma}$$

5. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

6. 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. Also, it helps to determine if two data sets come from populations with a common 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.

Importance:

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.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis