# Linear regression

# Subjective Questions

Submitted by

Vedavyas Burli

# Assignment based questions:

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

Based on our analysis and early assumption, categorical variables require special attention in regression analysis. Unlike continuous variable they cannot be entered into equation. There are many coding method used to convert the categorical variables in understandable text then perform regression analysis. Dummy coding is used mostly to compare each level of variables.

For example:

For the given dataset of bike rental, Season, month and weather which are categorical variables have positive effect on dependent variables. A unit change in categorical variables can have certain coefficient value of increase /decrease in dependent value

- season_Summer: A coefficient value indicates that a unit increase in season_Summer variable, increases dependent value numbers by 0.0812
- season_Winter: A coefficient value indicates that a unit increase in temp variable, increases the bike hire numbers by 0.1261
- mnth_sep: A coefficient value indicates that a unit increase in temp variable, increases the bike hire numbers by mnth_Sep 0.0895
- weathersit_Light Snow/Rain: A coefficient value indicates that a unit decreases in weathersit_Light Snow/Rain variable, decrease the bike hire numbers by -0.2535

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

If you perform Drop_first=True- it will reduce the correlations created among dummy variables and as it helps in reducing the extra column created during dummy variable creation.

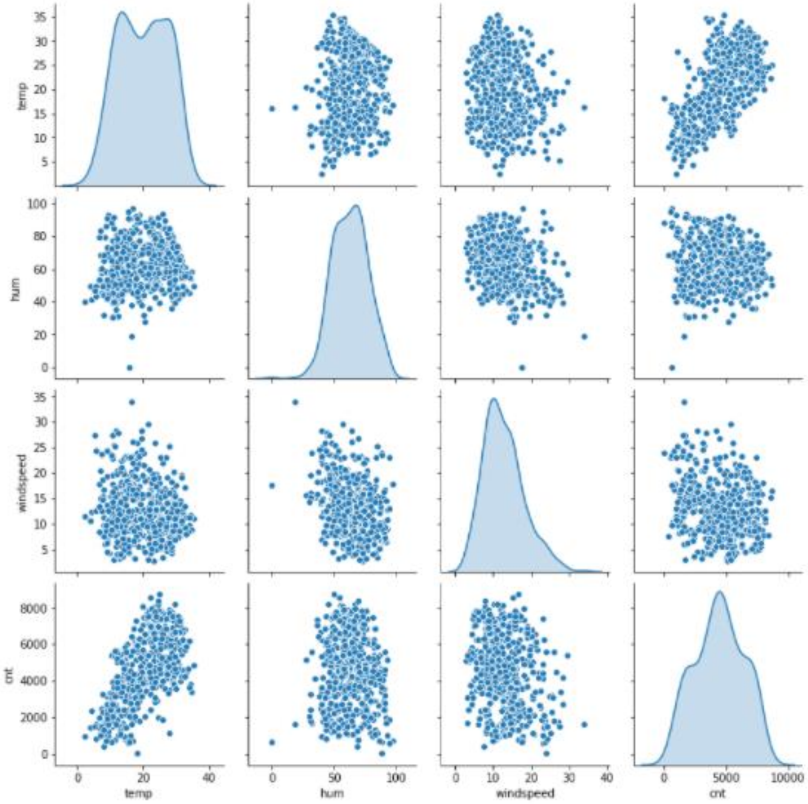
For example, if we have 4 type of values in categorical column:

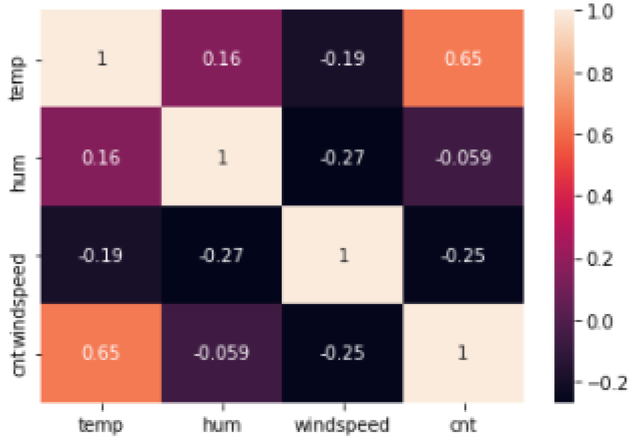
- Let’s say we have 4 types of values in Categorical column and we want to create dummy variable for that column.
- If one variable is not fall, Summer and Winter, then It is obvious spring. So we do not need 4rd variable to identify the spring.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Using pair plot we can observe from the below image that **temp** variable has linear relation with target variable ‘cnt’

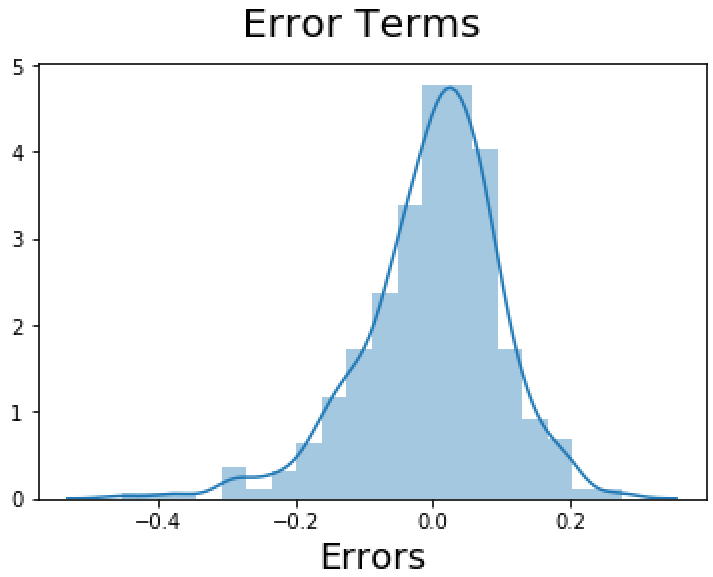




| | temp | hum | windspeed | cnt |
|---|---|---|---|---|
| temp | 1 | 0.16 | -0.19 | 0.65 |
| hum | 0.16 | 1 | -0.27 | -0.059 |
| windspeed | -0.19 | -0.27 | 1 | -0.25 |
| cnt | 0.65 | -0.059 | -0.25 | 1 |

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

After model is build to validate assumptions we need to perform residual analysis of training data.



An important step in regression analysis whether simple or multiple model has achieved its goal to explain as much variation as possible in a dependent variable while respecting the underlying assumption, is to check the *residuals* of a regression.

In other words, having a detailed look at what is ***left over*** after explaining the variation in the dependent variable using independent variables i.e the unexplained variation.

From the above graph we can conclude that residuals are normally distributed. Our assumption for linear regression is valid. Using the final model, we are can start predicting.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Based on our analysis, assumptions and final model, we have 3 features which are contributing towards the demand for these shared bikes in the American market.

   - Temp
   - year (yr)
   - Season

   - Above mentioned variables are significant in predicting the demand for shared bikes.
   - As Temperature(temp), year(yr) and season must be given high importance while planning to achieve max booking
   - Other features like month, weather, windspeed, holiday need to be considered before planning to change the strategy.

## General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

Linear regression is a method and an attempt to explain the relationship between a dependent and an independent variable using a straight line.

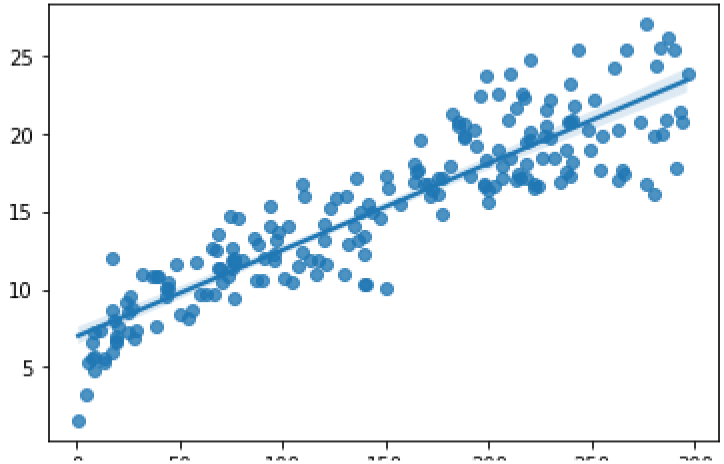
Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: When the number of independent variables is one
- Multiple linear regression: When the number of independent variables is more than one

Independent variable is known as predictor variable and dependent variable is known as output variables.

$Y = \beta0 + \beta1 (X)$ - simple linear regression

$Y = \beta0 + \beta1(X_1) + \beta2(X_2)+.... \beta pXp + E$ - Multiple linear regression



**Cost function:** To get the best possible values for  β0 , β2, β3… which would provide best fit line for the data points. For fitting a straight line, the cost function was the sum of squared errors, but it will vary from algorithm to algorithm. Most of the time you will have to minimize the cost function. Minimizing and maximizing a function is very simple.

- If the function value of the double differential should be greater than 0.
- If the function value of the double differential should be less than 0.

Best-fit line is obtained by minimizing a quantity called Residual Sum of Squares (RSS) -  It is the  calculation by finding the distance from each actual data point to the regression line (predicated values), square it, and sum all of the squared errors together.

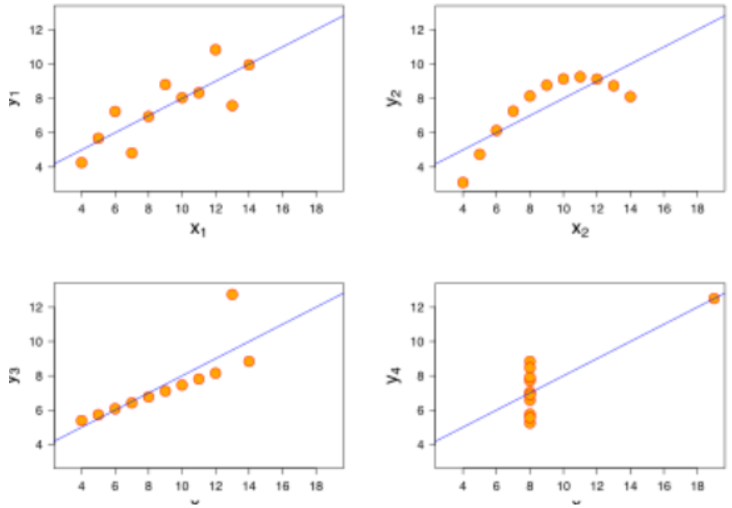
The strength of a linear regression model is mainly explained by $R^2$, where $R^2 = 1 - (RSS / TSS)$

RSS: Residual Sum of Squares

TSS: Total Sum of Squares

2. **Explain the Anscombe’s quartet in detail.**

Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics yet have different distributions and appear very different when graphed. It is developed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

### 3. What is Pearson’s R?

**Pearson’s R or bivariate correlation or Pearson correlation coefficient** is a measure of linear correlation between two sets of data. It is the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \text{(Eq.1)}$$

where:

- cov is the covariance
- σx is the standard deviation of X
- σy is the standard deviation of Y

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

**Scaling:**

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

**Why scaling is performed?**

If scaling is not done right , then a ML algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

- Standardization - 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). - sklearn.preprocessing.scale helps to implement standardization in python.
- 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.

Note: 1) Categorical variables cannot used as they are so they are converted to numeric format.2) Scaling affects the coefficients

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

Variance Inflation Factor (VIF) detects multicollinearity in regression analysis. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

It is calculated as— VIFi=1/(1−Ri 2 )

When corresponding variable is being expressed exactly by a linear combination of other independent variables, R2 becomes 1 hence denominator of VIF equation (1-R 2 ) = 0 so VIF is infinite.

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

Quantile-quantile 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 datasets are from populations with same distributions.

Importance of Q-Q plot in linear regression:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, 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 answer the following scenarios - if two data sets:

- Come from populations with a common distribution
- Have common location and scale
- Have similar distributional shapes
- Have similar tail behavior

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.