Linear Regression Subjective Answers

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

Categorical Variable	Effect on demand
Season	Summer and Fall season sees the most demand.
Month	Corresponding to season, the months from June to September has the highest demand
Holiday	People do not use a lot of bikes on holiday
Weekday	Saturdays and Sundays have a high demand, but Tuesday sees the lowest.
Working Day	Bikes are less in demand on working days
Weather Situation	Clear and Misty weather sees more demand.

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

- Drop_First helps in reducing the number of dummy variables by reducing the number of columns generated.
- For e.g. if there are 3 levels in a certain categorical variable (A,B,C) it can be easily represented by 2 dummy variables.

- When we are dealing with a large number of categorical variables, the number of features when building dummy variables will be very large. Using these will involve more computational power. This step slightly helps in reducing the number of features

B	C	Result
0	0	A
1	0	B
0	1	C

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

- The Target variable is cnt (total count) of users.
- The features that have the highest correlation are :
  - **Temperature** (temp)
  - **Feeling Temperature**(atemp)
- Both have a **correlation value of 0.63**
- The values casual and registered also are highly correlated but that is only because cnt = casual + registered.

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

Assumptions	Method	Code

Linear Relationship between Target and Independent variables.	Using a pairplot or separate Scatterplots, we can get check the linearity.
	Regplot will try to plot line of best fit between the two variable.
		Plots:
		Sns.scatterplot(X,y)
		Sns.pairplot(X,y)
		Sns.regplot(X,y)

Error terms should be normally distributed	A distribution plot will indicate whether the error terms are normally distributed or not.	Res = y_train – y_pred
		Sns.distplot(res)

Mean of Residuals should be zero.	A histogram or even a mean calculation will show if the errors are have mean = 0	Res = y_train – y_pred
		Sns.distplot(res)
		
		Np.mean(res)

Heteroscedascity	A scatterplot between error terms vs y_pred will show if there is any pattern present. In case, no patterns are found, the model has heteroscedascity.	Sns.scatterplot(y_pred,residuals)
No Multicollinearity	A heatmap between the dependent variables will show if they are multicollinear or not.	Sns.heatmap(df.corr())

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

- In my model, the most significant features are:
  - Feeling Temperature (*atemp*)
    - The temperature that customer feels will dictate if he wants to take a bike
  - Humidity (*hum*)
    - Humidity will affect usage as bikers will sweat and will be uncomfortable in a humid weather.
  - Season (*season*)
    - Winter and Summer seasons sees most demand.

Explain the linear regression algorithm in detail

- Linear Regression is a type of machine learning algorithm which belongs to the class of Supervised Learning. This means that the dataset needs to be clearly labelled in order to build the model.
- The target variable in Linear Regression is a continuous variable. E.g. sales figure, estimated costs etc.
- This regression technique finds out a linear relationship between x (input) and y(output). This line is called the Line of Best Fit.

$$y = mX + c$$

Here X = input variable

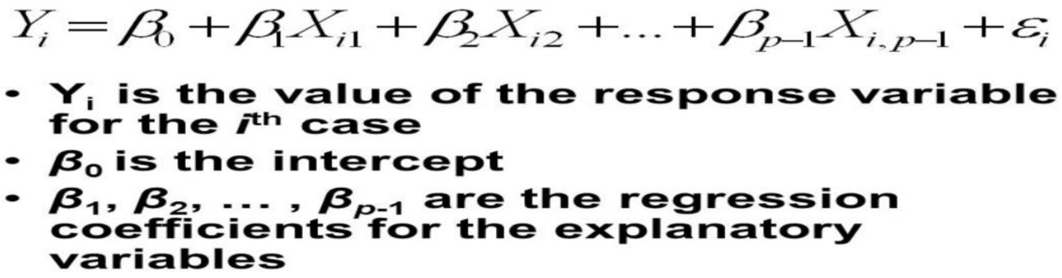
And y = target variable (always singular),

m = slope of the line

c = y-intercept

The model tries to find the best coefficient(s) for the input variables.

This equation hold true for even more number of input variables – hweover the equation slightly changes:

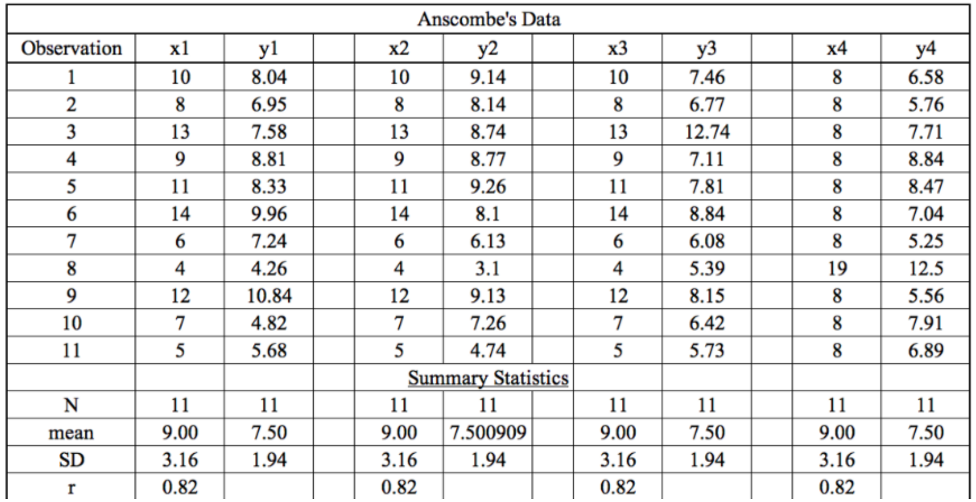


$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Y_i is the value of the response variable for the i^{th} case
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables

Explain the Anscombe’s quartet in detail.

- Anscombe’s Quartet is a group of four datasets which are identical in manner of simple descriptive statistics which are meant to delude the regression model.
- They have very different distributions and appear differently when plotted via a scatterplot.
- Invented by French statistician, Francis Anscombe, the quartet was designed to illustrate the importance of plotting graphs before modelling. Basically, Anscombe’s Quartet is meant to show the importance of Exploratory Data Analysis (EDA) and Data Visualization.
- It can be seen that the 4 sets of data have the same mean, SD and other summary statistics. However, the story is different when we plot these points.
- Only X1-Y1 set can be handled by a Linear Regression.
- X2-Y2 is non linear.
- X3-y3 has outliers that can’t be handled by Linear Regression.
- X4-Y4 has a linear shape but can’t be handled by a LR model.



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
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

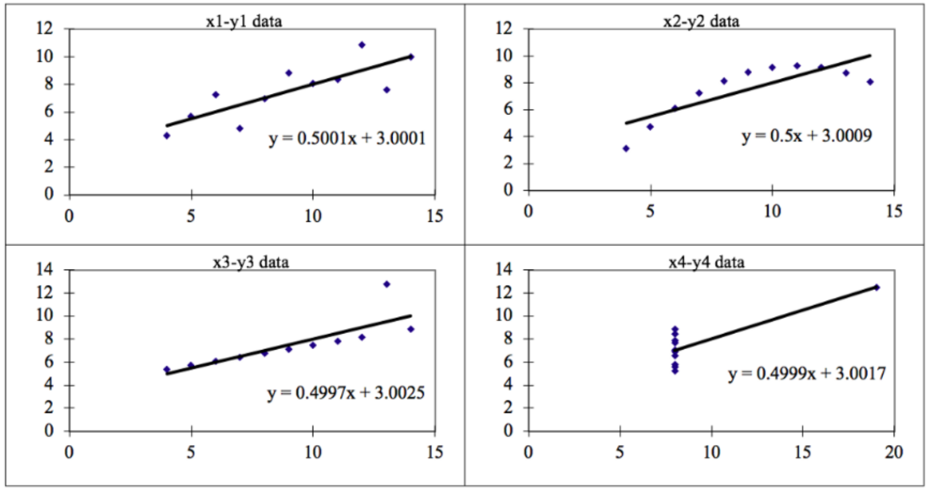
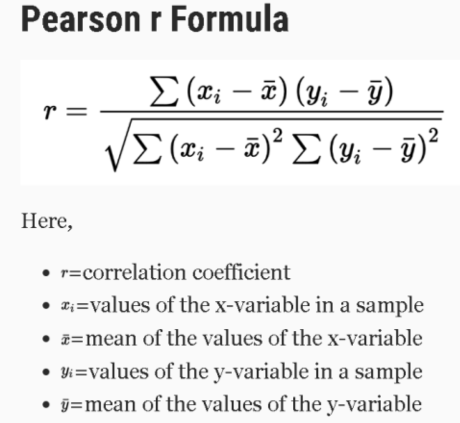


Image by Author

What is Pearson’s R?

- Named after Karl Pearson, the Pearson’s Correlation Coefficient is a statistic that measures the linear relationship between two variables.
- It has a numerical range -1 to +1.
- Interpretation:
    - The closer the absolute value is to 1, the more highly correlated are the two variables.
    - The sign signifies the slope:
        - +1 : perfectly linear with +ve slope (both variables change in the same direction:
        - -1 : perfectly linear with –ve slope (both variables change in opposite directions)
        - 0: no correlation



Pearson r Formula

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

Here,

- r=correlation coefficient
- x_i=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

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

- Scaling is a step in pre-processing stage of the model building which is applied to the predictor variables to normalize the data within a particular range.
- Most of the times, the data used to model vary in values, units and range. Without scaling, the model consumes the data with its magnitude and will give incorrect predictions.
- Thus scaling is used to bring all the data to similar magnitudes.

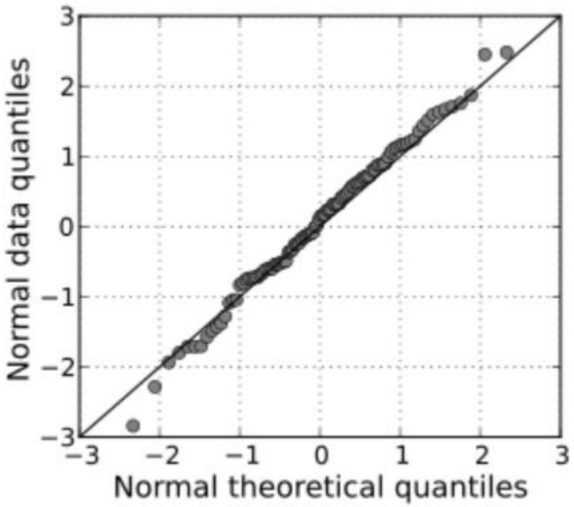
- There are two types of Scaling - Normalization and Standardization:
  - Normalization:
    - Also Known as Min-Max Scaling, this method brings all the predictor variables in a range of 0 to 1.
    - This is helpful as the scale of the features doesn’t matter since the model will consume all values within 0 to 1.
    - $x = \frac{x - mean(x)}{\max(x) - \min(x)}$
  - Standardization:
    - This method involves replacing their values by their Z-Scores. This method brings all the data into a standard normal distribution which has mean = 0 and S.D = 1.
    - $x = \frac{X - mean(x)}{sd(x)}$

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

- Variable Inflation factor (VIF) provides a measure of how one independent variable is explained by all the other independent variables combined.
- $VIF = \frac{1}{1-R^2}$
- VIF can only be infinity when R^2 is 1 i.e. all the independent variables are perfectly correlated to each other.

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

- Quantile-Quantile (Q-Q) plot, is a graph that checks if a set of data came from same distribution. It also helps to determine if two data sets come from same population.
- This helps in 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.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
- If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.


Normal data quantiles
Normal theoretical quantiles