SUBJECTIVE QUESTIONS

Linear Regression Assignment

C44 Group

Shrinivas Bhat
s.shrinivas.bhat@gmail.com

Table of Contents

Assignment-based Subjective Questions

Categorical Variable Inferences

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer

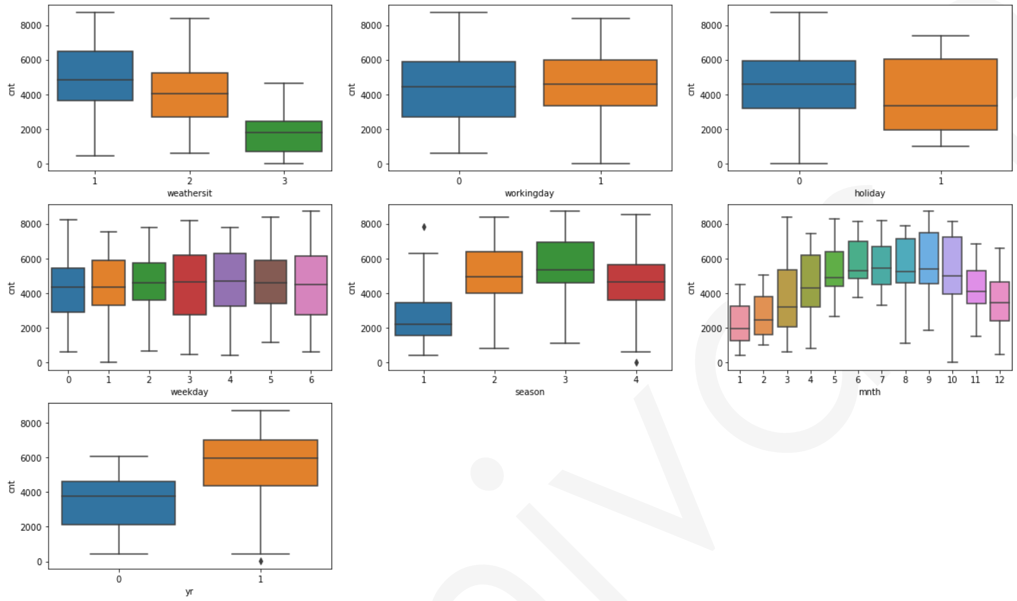


Figure 1 Impact of Categorical Variables on the 'Cnt'

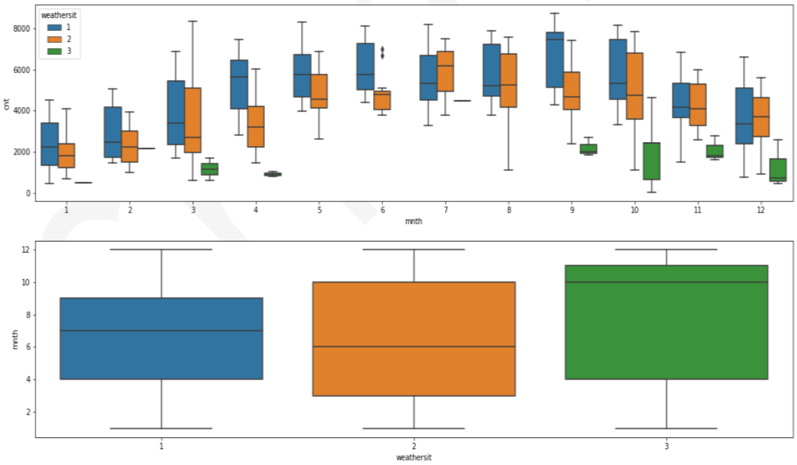


Figure 2 Weather situations across various months

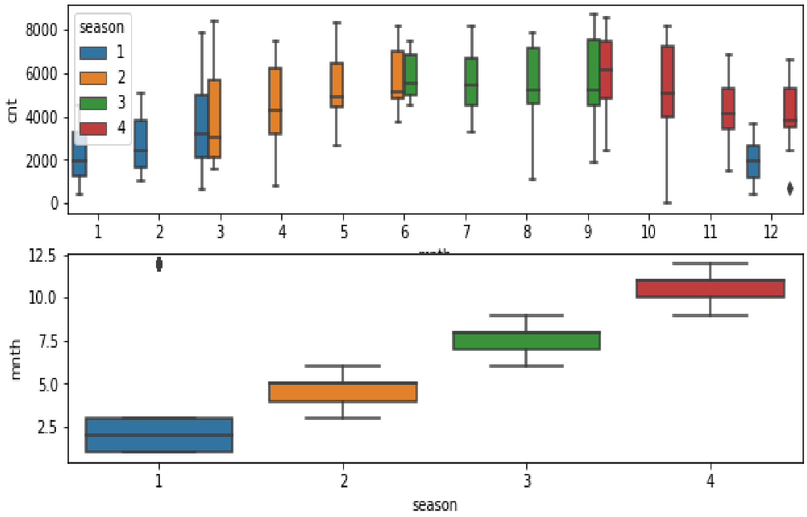


Figure 3 Seasons across months

From the box plot shown in Figure 1 we can infer that

- Rental bike usage reduces when the weather situation becomes harsh.
  - ‘weathersit’ 1 seems to be more favourable for usage of rental bikes compared to ‘weathersit’ 2 and 3. i.e Clear clouds or Sparse clouds seems to be more favourable for people’s usage of rental bikes.
  - ‘weathersit’ 2 is more favorable than ‘weathersit’ 3. i.e Misty+ Cloudy weather more favourable compared to Light snow or Light rain for the usage of rental bikes.
  - There are no data available for ‘weathersit’ 4 in the dataset.
- Median demand for rental bikes seems to show weak trend (not much variation) with ‘weekday' and 'workingday' features’ the demand for the rental bikes does not seem to vary much with respect to the specific days of a week or whether it is a working day or not.
- Median demand for rental bikes seems to be lesser on a 'holiday', even though the peak demand (75%) does not change if it is a holiday or not.
- Even though the dataset has been sampled only for 2 years there is an increasing trend in the demand for rental bikes year on year (between 2018 and 2019). But since the samples are not available for multiple years extrapolation of data for a different year may be risky. Model is built considering 2 Situations
  - Considering that there is an linearly increasing trend year on year we can map the ‘yr’ feature (0=2018,1=2019,2=2020…and so on) and build the model. The coeff obtained from this fit can be considered as a ratio of increase in the demand for bikes year on year.
  - Consider ‘yr’ feature for only 2018 and 2019 as available in dataset the coefficients are determined for 2018 and 2019 by doing a one-hot encoding or dummy feature creation for the ‘yr’ feature.

- The 'season' fall and summer (2 and 3) seems more favourable for usage in bikes compared winter and spring.

- From the 'mnth' categorical variable it can be inferred that
  - Demand for the bikes seems to be showing an increasing trend between months 1 to 6.
  - Demand for bikes seems to be showing a linearly decreasing trend between months 9 to 12.
  - Demand for bikes is at the peak and constant between months 6 to 9.
  - These variations in demands can be partially attributed to seasons and weather situations variations happening across those months. (As shown in Figure 2 and Figure 3)
    - Decreasing demand in bikes from months 1-3 may be possibly due to spring season occurring in those months. We had already seen that spring seasons demand for the bikes is lesser.
    - The months 6-9 have a 'season' 3 which is fall and favourable for bike renting.
    - Decreasing demand in bikes from months 9-12 may be attributed to combination of winter season and the harsher weather situation (weathersit=3) occurring in those months. We had already seen that winter seasons demand for the bikes is lesser.

Dummy Variable Importance

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer

Its important to use drop_first=True as it helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.

For example, consider a categorical variable with 3 possible values ‘a’,’b’,’c’

When the dummy encoding is done for these categorical variables, it creates 3 new columns/features ‘a’,’b’ and ‘c’ as shown below with ‘a’ =1 for when categorical variable is ‘a’ : ‘b’ = 1 when categorical variable is ‘b’ : similarly for ‘c’=1 when categorical variable is ‘c’. But one of these columns is redundant i.e categorical variable with value ‘a’ can be represented with ‘b’ and ‘c’ = 0, For the Linear regression model the coefficient multiplied ‘b’ and ‘c’ columns can be considered incremental with the coefficient for ‘a’ captured into the Constant (Intercept) of the regression model. Since the first column becomes redundant it can be dropped using the drop_first=true.

Following code-snippet shows the dummy creation where columns ‘a’,’b’,’c’ are created.

```
pd.get_dummies(pd.Series(list('abcaa')))
   a  b  c
0  1  0  0      ‘a is explicit’
1  0  1  0      ‘b is explicit’
2  0  0  1      ‘c is explicit’
3  1  0  0      ‘a is explicit’
4  1  0  0      ‘a is explicit’
```

Following code-snippet shows the dummy creation where columns ’b’,’c’ are created and ‘a’ is implied.

```
pd.get_dummies(pd.Series(list('abcaa')), drop_first=True)
   b  c
0  0  0        ‘a is implied’
1  1  0
2  0  1
3  0  0        ‘a is implied’
4  0  0        ‘a is implied’
```

Correlation with Target Variable

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer

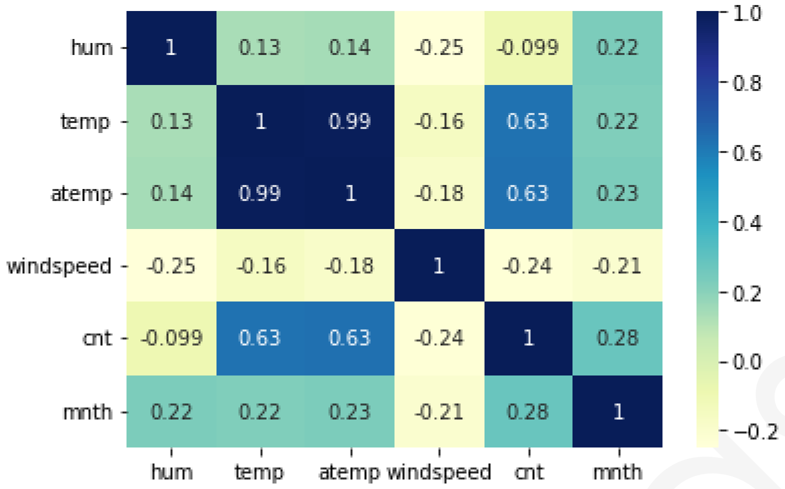


Figure 4 Heatmap for numeric variables

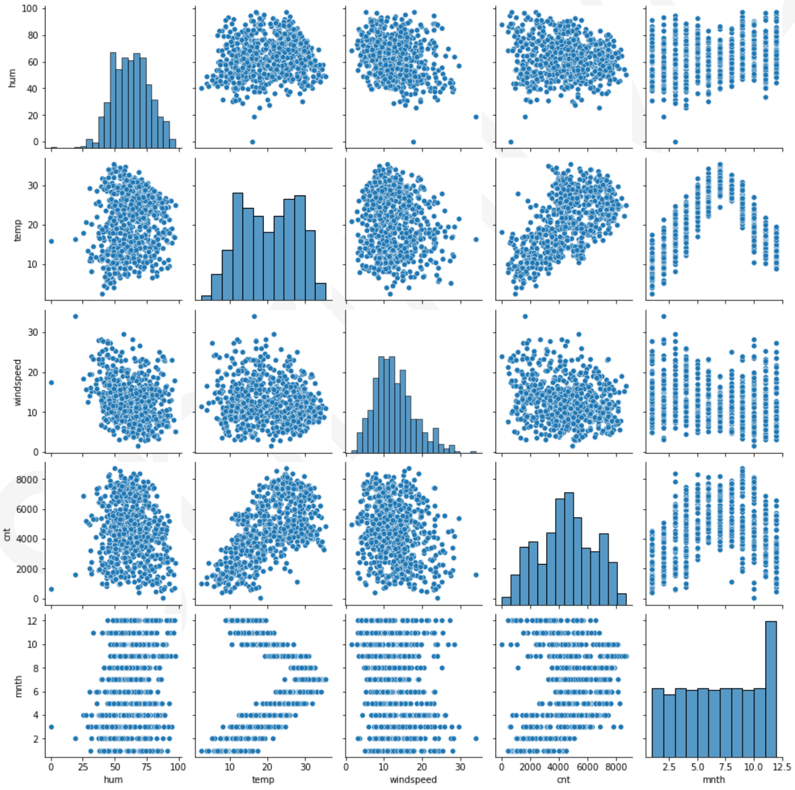


Figure 5 Pair-plot for numeric variables

- 'temp' variable seems to have the highest correlation with the target variable 'cnt'.

- Since ‘temp’ and ‘atemp’ are collinear variable with 0.99 Correlation ‘atemp’ is dropped.

Assumptions of Linear Regression

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer

1. Linearity assumptions was validated by
   - Observing fitted data (on both training and test data) versus the training/test data and ensure that they are linear. (y_train vs y_pred and y_test vs y_pred look linear)

2. Zero-mean assumption, Constant variance assumption and independent error assumptions of the residual error
   - From the residual plots versus training /test data ensure that
     - Residual error has zero mean and very low mean square error
     - Residual error has constant variance (Homoscedasticity) can be seen from residual plots of fig-6 and fig-7.
     - Residual error is normally distributed. We can look at the randomness of the error from the residual error plot in the fig-6 and fig-7 for training and test data respectively.

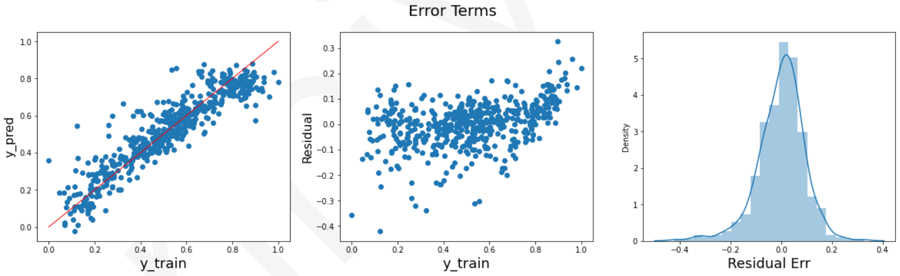


Figure 6 Linearity test with predicted data and training data (ytrain vs ypred ,Residual Err plot and distribution)

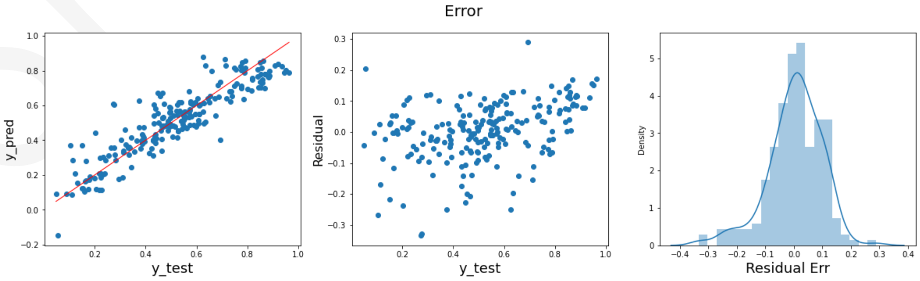
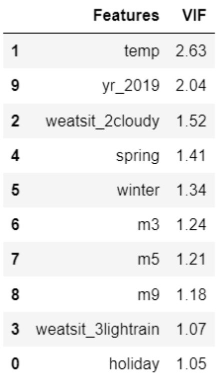


Figure 7 Linearity test with predicted data and test data (ytest vs ypred , Residual Err plot and distribution)

3. No Multicollinearity between predictor variables i.e the predictors chosen for the linear regression are linearly independent of each other. This can be done by looking at the VIF for each predictor variables. VIF of the predictor variables should be less than 5.

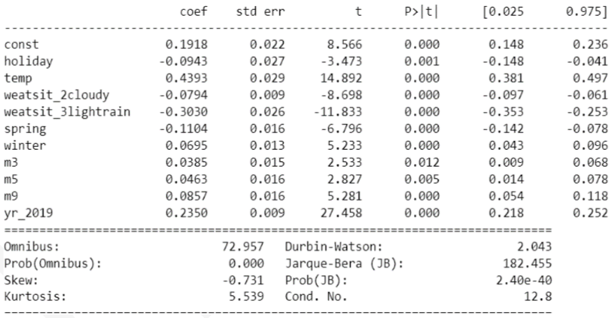


	Features	VIF
1	temp	2.63
9	yr_2019	2.04
2	weatsit_2cloudy	1.52
4	spring	1.41
5	winter	1.34
6	m3	1.24
7	m5	1.21
8	m9	1.18
3	weatsit_3lightrain	1.07
0	holiday	1.05

Note: Above VIF is when "windspeed" is dropped

Figure 8 VIF of predictor variables.

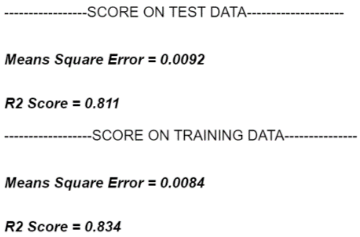
4. Check the significance of the coefficients using the p-value for each predictor. The p-value indicates the probability of the null hypothesis that the predictor variable is independent of the target variable. (i.e coeff for predictor variable is zero) p-value < 5% is used to reject the null hypothesis. To check if the overall fit is significant or not the primary parameter is ***F-static.***



```
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.1918      0.022      8.566      0.000       0.148       0.236
holiday              -0.0943      0.027     -3.473      0.001      -0.148      -0.041
temp                  0.4393      0.029     14.892      0.000       0.381       0.497
weatsit_2cloudy      -0.0794      0.009     -8.698      0.000      -0.097      -0.061
weatsit_3lightrain   -0.3030      0.026    -11.833      0.000      -0.353      -0.253
spring               -0.1104      0.016     -6.796      0.000      -0.142      -0.078
winter                0.0695      0.013      5.233      0.000       0.043       0.096
m3                    0.0385      0.015      2.533      0.012       0.009       0.068
m5                    0.0463      0.016      2.827      0.005       0.014       0.078
m9                    0.0857      0.016      5.281      0.000       0.054       0.118
yr_2019               0.2350      0.009     27.458      0.000       0.218       0.252
==============================================================================
Omnibus:                       72.957   Durbin-Watson:                   2.043
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              182.455
Skew:                          -0.731   Prob(JB):                     2.40e-40
Kurtosis:                       5.539   Cond. No.                         12.8
------------------------------------------------------------------------------
```

Figure 9 Linear regression output

5. Finally evaluate the performance of the Linear regression i.e the goodness of the fit with R^2 Score and Adjusted-R^2 Score.It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.
R^2 Score is determined both for training data and the test-data.



-----------------SCORE ON TEST DATA--------------------

Means Square Error = 0.0092

R2 Score = 0.811

------------------SCORE ON TRAINING DATA---------------

Means Square Error = 0.0084

R2 Score = 0.834

Feature Selection

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer

- Model-1: With Min-Max Scaling

```
-------------------------------------------------------
      Final Features and Coefficients
      ------------------------------
      temp                      0.447379
      weatsit_3lightrain       -0.287028
      yr_2019                   0.234598
      const                     0.214403
      windspeed                -0.151694
      m9                        0.091051
      holiday                  -0.090982
      winter                    0.086565
      spring                   -0.082791
      weatsit_2cloudy          -0.079341
      m5                        0.064284
      m3                        0.054170
      m4                        0.053375
      m6                        0.036216
```

- Model-2: With No Scaling

```
-------------------------------------------------------
      Final Features and Coefficients
      ------------------------------
      weathsit_3lightrain       -2669.849887
      yr_2019                   2053.088906
      const                     1434.605972
      spring                   -1025.859679
      m9                         757.881984
      weathsit_2cloudy           -686.199809
      winter                     439.059989
      m5                         422.225315
      m3                         360.672893
      m10                        352.148076
      w6                         174.738815
      temp                       113.114149
```

The top 3-features explaining the demand of the shared bikes

1. Demand for shared bikes seems to increase by 0.45 times normalized 'temp'
2. Demand for shared bikes seems to decrease by -0.28 when 'weathersit_3' (light_rain)
3. Demand for shared bikes seems to increase by +0.23 when 'yr_2019'

   Note: In Model-2 even though the coefficient of temp is lowest the range of the temp is higher than any of the other variables hence for identifying the significantly contributing variable it is better to consider Model-1 where the features are normalized/scaled.

General Subjective Questions

Linear Regression Algorithm Details

Q1. Explain the linear regression algorithm in detail.

Answer:

Regression is a statistical technique to derive an algebraic relationship between two or more variables. Based on this algebraic relationship one can estimate the value of a dependent variable, given the values of the other independent variables. If a correlation finds any relationship between the variables, regression is used to find the degree of relationships that can be then used for prediction. Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear.

Linear Regression is a machine learning algorithm based on supervised learning. Linear regression is a quiet and simplest statistical regression method used for predictive analysis of a target variable based on independent variables and shows the relationship between the dependent and independent variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called ***simple linear regression***. And if there is more than one input variable, such linear regression is called ***multiple linear regression***. The linear regression model gives a sloped straight line describing the relationship within the variables.

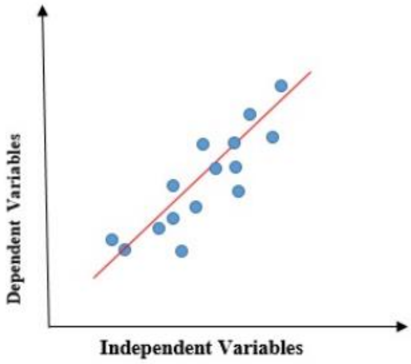
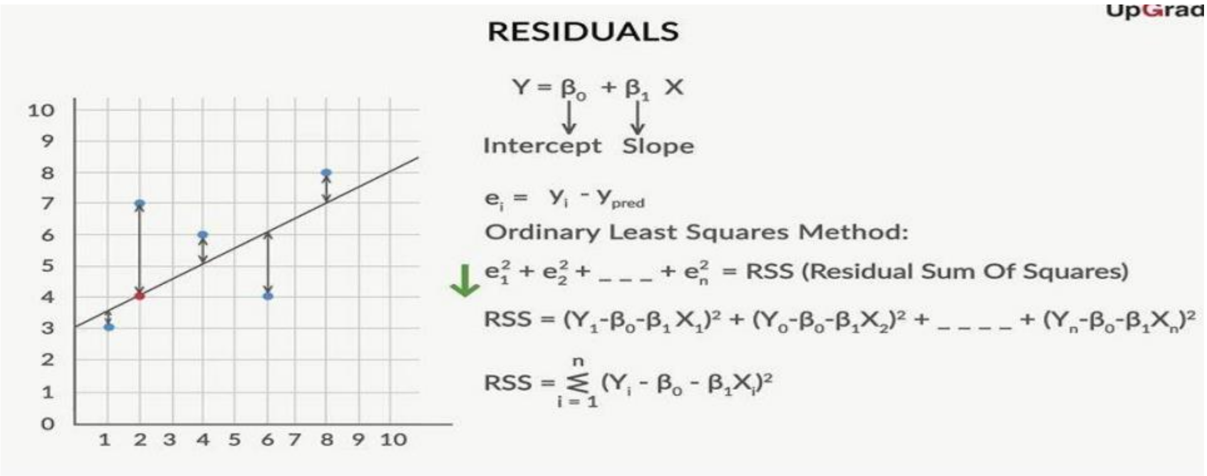


Figure 10 Depiction of Simple Linear regression

<u>Simple Linear Regression Algorithm</u>

The mathematical equation can be given as: Y = β0 + β1*x Where Y is the response or the target variable x is the independent feature ***β1 is the coefficient*** of x ***β0 is the intercept***. β0 and β1 are the model coefficients (or weights).

While training the model we are given: x: input training data (univariate – one input variable(parameter)) Y: labels to data (supervised learning).When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best β0 and β1.



RESIDUALS

$Y = \beta_o + \beta_1 X$

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$RSS = (Y_1-\beta_o-\beta_1X_1)^2 + (Y_o-\beta_o-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_o-\beta_1X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2$

Figure 11 Residual

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

In linear regression, ***Gradient descent*** is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.

Gradient Descent starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value. It repeats the following optimization equation until convergence

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).x_j^{(i)} \text{ for } j = 1,2,\ldots,n$$

Figure 12 Optimization

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
   R^2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1.
   Mathematically $R^2 = 1 - \frac{RSS}{TSS}$
   RSS (Residual Sum of Squares) defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output.
   $$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$
   TSS (Total sum of squares): defined as the sum of errors of the datapoints from mean of response variable given by

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

2. Residual Standard Error (RSE)

Linear regression is based on the following assumptions and these assumptions needs to be validated after building the model.

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Note: No assumption on the distribution of X and Y, just that the error terms have a normal distribution.

Hypothesis testing for coefficients

After fitting the straight line on the data, we need to determine if coefficients determined are the significant fit for the data. For this hypothesis testing on the βs are done.

Null and alternate hypothesis are (considering only one feature this can be repeated to all features)

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

For which the test statistics is given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

The p-value is then calculated on this test statistic in order to determine whether the coefficients are significant or not.

Finally, the parameters to access the model fit are

- t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
- F statistic: Used to assess whether the overall model fit is significant or not.
- R^2 : After it has been concluded that the model fit is significant, the R^2 value tells the extent of the fit, i.e. how well the straight line describes the variance in the data.

Multiple Linear regressions Algorithm (MLR)

In case of Multiple regression we have multiple βs instead of a single β as shown below equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

And in this case the model fits a ***hyperplane*** instead of a line like in simple regression.

The assumptions of simple linear regressions like zero-mean, independent, normally distributed error terms still hold and coefficients are obtained by minimizing sum of square error using gradient descent method.

Additional considerations in Multiple regressions

1) Issues of overfitting the model to the training set with no generalization. It will not fit the test-set.
    a. When you add more and more variables, for example, let's say you keep on increasing the degree of the polynomial function fitting the data, your model might end up memorizing all the data points in the training set.
2) Issues of Multicollinearity
    a. Multicollinearity refers to correlation between the independent features/variables. This can be identified by looking at a parameter called VIF (Variable inflation factor) for each independent feature/variable and iteratively dropping the variables having higher VIF > 5.
    b. This can also be identified in the pre-model building process by studying the correlation and pair plots between the independent variables.
3) Feature scaling
    a. Independent variables in a model might be on very different scales which will lead a model with very weird coefficients leading to interpretation issues and issues in convergence gradient descent algorithms. Feature scaling of the independent variables needs to be performed before building the model to mitigate these issues
    b. Two forms of feature scaling
        i. Standardizing: Feature scaled to ensure zero-mean and unit standard deviation for that feature.
        ii. Min-Max Scaling: Feature values scaled in such a way that they lie between zero and one.

4) Feature selection process (Algorithms)
    a. Try all possible combinations (2 p models for p features) ○
        i. Time consuming and practically unfeasible
    b. Manual Feature Elimination
        i. Build model
        ii. Drop features that are least helpful in prediction (high p-value)
        iii. Drop features that are redundant (using correlations, VIF)
        iv. Rebuild model and repeat
    c. Automated Approach ○ Recursive Feature Elimination (RFE)
        i. Forward/Backward/Stepwise Selection based on AIC /BIC
5) Categorical feature
    a. In MLR Categorical variables that might turn out to be useful for the model.
    b. Create one hot encoding or dummy variable creation for handling the categorical variables.

Anscombe’s Quartet

Q2. Explain the Anscombe’s quartet in detail.

Answer :

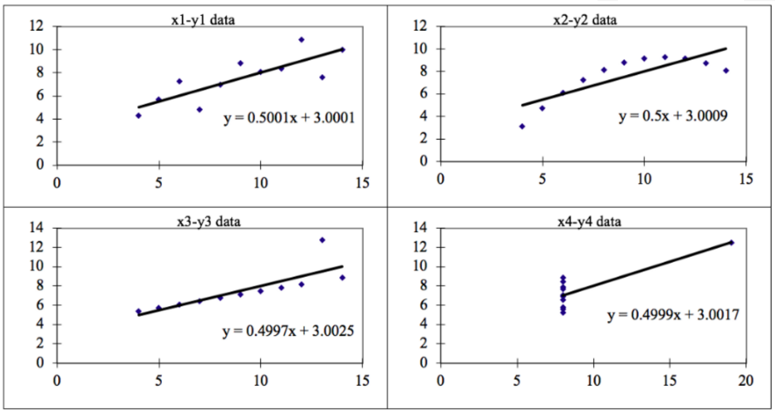
Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before model building.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms to build models, to see the distribution of the samples to identify the various anomalies present in the data like outliers, diversity of the data, linear separability etc

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable/differentiable by any regression algorithm.



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Conclusion: *All the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.*

Pearson’s R

Q3. What is Pearson’s-R?

Answer:

Pearson's R is a coefficient which represents the strength of the linear association between the variables/features. This is applicable mostly on continuous numeric features or variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. This is used in linear regression models. If the relationship between variables is non-linear this is not a good measure.

The Pearson's correlation coefficient varies between -1 and +1 where:

R= 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

R = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

R = 0 means there is no linear association

|R| > 0 < 0.5 means there is a weak association

|R| > 0.5 < 0.8 means there is a moderate association

|R| > 0.8 means there is a strong association

Feature Scaling

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

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.

We need to scale features because of two reasons:

1. Ease of interpretation

2. 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. Min-Max 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

Variance Inflation factor

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

VIF also called Variance inflation factor is used to check the presence of multicollinearity in the dataset. It is given by

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

Here VIF_i is the Inflation factor for the i^{th} independent variable and R_i^2 is the R^2 of the i^{th} independent variable w.r.t another independent variable. If there is a perfect Correlation between the supposed independent variables VIF becomes infinity. In the case of perfect correlation, we get R_i^2 =1, which lead to 1/(1-R_i^2) infinity. In such cases for a Linear regression problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q-Q Plot

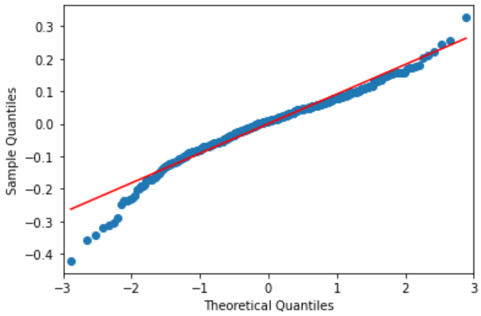
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data 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.

In linear regression when we have training and test data set received separately and using this tool we can confirm that both the data sets are from populations with same distributions.

A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45-degree reference line:



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

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. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

In summary Q-Q Plots are used to find 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 behaviour