# 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)**

Dependent variable is cnt (sales of bikes)

The categorical variables effects are:

* **Season** – Spring season seems to have lowest sales. Summer and Fall have good sales. Winter season has some reduced sales.
* **yr** - There is a significant difference in cnt from 2018 to 2019
* **mnth** – These follow almost the same pattern as the season that they represent. That is months of spring (starting January) shows the lowest sales. Months of Fall (9 and 10) show the highest sales.
* **holiday** - Distribution of cnt on holidays seems to have a different distribution than regular days.
* **weekday** - Distribution of cnt is almost similar for all weekdays. So no significant impact.
* **workingday**- Distribution of cnt is almost similar for all workingday. So no significant impact.
* **weathersit** - Worse weather conditions impact the cnt of bikes sold.

**2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)**

Will explain this with example:

Consider a field with values for Male / Female. When we create dummy variables, we will get two columns eg.,

|  |  |
| --- | --- |
| Male | Female |
| 0 | 1 |
| 1 | 0 |
| 1 | 0 |

But we only need 1 column to indicate if the variable was male or female. **Hence, we can drop the first column**. The value of 1 indicates it is female otherwise 0 indicates male.

|  |
| --- |
| Female |
| 1 |
| 0 |
| 0 |

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

The variables **temp** and **atemp** both have a correlation factor of **0.63** with cnt. Both temp and atemp are very strongly correlated variables (practically 1.0). Hence, they can both be considered almost the same variable.

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

We validate the assumptions of Linear Regression Model by showing following properties of our model hold:

|  |  |
| --- | --- |
| **Linear Relationship**: Our model’s prediction values (Ypred) correctly follow the same pattern as the actual Y’s. That is, the residuals (Y – Ypred) should be within a narrow range evenly along the entire range of X. No discernible patterns visible in the residuals-vs-predictors plot as seen below. |  |
| **Homoscedasticity**:  The variance of “standardized” residuals is nearly same for any value of X. |  |
| **Absence of Multicollinearity**: The VIF values of all our model variables was < 5.0. Hence, no significant multicollinearity is present among these variables. | **Features VIF.**  4 spring 2.84  3 temp 2.67  5 winter 1.70  7 Rainy 1.05  1 yr 1.03  6 Cloudy 1.03  2 holiday 1.01 |
| **No Auto-Correlation (Independence of residuals)**: Residuals are independent of each other & randomly scattered along the x-axis. There is no dependence of any residual on previous value. We should **not** see patterns like below.    *Autocorrelation may happen if* **(1)** *Model has missing relevant variables.* **(2)** *Model has incorrect function form (e.g., should have X2 or log(x) instead of xi ).* | Also, Durbin-Watson statistic is **2.01** (hence no auto-correlation) |
| **Normality of Errors**: The errors are distributed in a normal (bell-shaped) curve around mean=0. |  |

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



Based on the model above, the top 3 variables impacting significantly the demand of bikes are:

1. Rainy Day -- Negative impact.
2. Year – Positive impact.
3. Spring Season – Negative impact.

# General Subjective Questions

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

Primary idea is that the prediction values based on our model is not very different than what the actual value should have been.

Assuming, that all the data has been cleaned, we fit a linear model based on our final parameters.

**Ypred = B0 + B1.x1 + B2.x2 + B3.x3 + …..**

Where B0 is a fixed constant. Remaining B# are coefficients (weights) for each of our parameters. The goal is to find the vector [B1, B1, B2, B3 …. ]

We can use the statsmodel.api function **OLS(target, X).fit()** to get a linear model. (Internally, this uses a gradient descent function – explained at end).

We can iteratively refine our model by adding/deleting parameters to our model and checking:

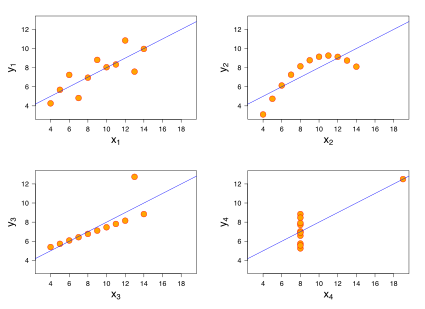
* Does it improve the R2 and adjusted R2 values.
* Is F-statistic a high value and Prob (F-Statistic) should be very low (almost 0).
* P-value for a parameter should be less than 0.05 (otherwise they are a candidate to be removed from our model).
* Coefficient value for our parameter should be significant (i.e., not almost 0). Else they could possibly be removed.

We repeat the above steps until we can no longer add/remove any parameter. That gives us the final regression model.

|  |
| --- |
| **Gradient Descent**: (in brief)   * It tries to minimize the cost (penalty) function of:   ∑i (yi – (β0 + β1X))2   * Above cost function is dependent on B0 and B1 (for example). We partially differentiate the above function w.r.t B0 and B1 (separately) and find the slope at a particular X value. * We refine the B0 (or B1) by a small learning factor (typically 0.01 or 0.001). * We can then recalculate our cost function to see if it is much different. When the slope value is almost Zero, we have achieved a minimal cost value (i.e., in further iterations, there will not be any change in cost function). * Hence, at that values, we have achieved our desired B0 and B1.   This method can be generalized to more than 1 parameter i.e, if we have B0, B1, B2, B3 …. |

1. Explain the Anscombes quartet in detail. (3 marks)

Anscombe defined a group of 4 data sets which are nearly identical in descriptive statistics (mean, std-deviation etc), but they actually have very different distributions when visualized on a scatter plot.



* Graph 1 – Indeed a linear relationship.
* Graph 2 – Non-Linear relationship
* Graph 3 – Regression line is erroneous due to outlier.
* Graph 4 – One high-leverage point produces a high correlation coefficient, even though other data points indicate there is no relationship between X and Y.

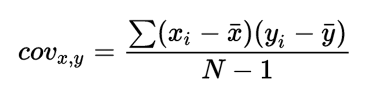
These plots demonstrate both the **importance of graphing data** before analysing it, and the **effect of outliers** and other influential observations on statistical properties.

1. What is Pearsons R? (3 marks)

To understand correlation between independent and dependent variables X & Y, we need to derive 2 things:

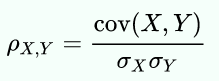
1. Direction of relationship (does Y increase/decrease when X increases).
2. Strength of the relationship (“how much” does Y vary when X varies).

Covariance is given by formula:



But covariance only gives the **direction** of the relationship between X and Y

Pearson’s R (related to covariance above) is given by:



The correlation coefficient R [or ρ (rho)] is essentially a normalization of the covariance by dividing it by standard deviations of X and Y. But with Pearson’s coefficient, we can know the “**Strength**” of the relationship where (-1 < ρ < +1).

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

**The What**: Scaling is the process of converting all variables to be compressed within similar ranges. (eg., between 0 and 1).

**The Why**: There are 2 reasons for scaling to be performed when there are multiple variables with possible co-linearity.

1. If no scaling is done, the regression model may produce imprecise coefficients (because of the vastly different scales of the parameters) and make it more difficult to choose the correct model. For example: One coefficient might be something like 456789.45 and another might end up being 0.000034. Hence, the parameter with extremely small coefficient might end up being unfairly rejected from the model.
2. When the parameters are scaled to similar scales, the regression algorithm will converge much faster (as the cost function in gradient descent will reduce similarly across all the parameters).

**The How**: Scaling can be done by either normalized or standardized scaling.

|  |  |
| --- | --- |
| **Normalized**: rescales the values into a range of [0,1] using a MinMax Scaler. This basically rescales using: |  |
| **Standardized**: typically means rescales data to have a mean of 0 and a standard deviation of 1 using: |  |

1. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

VIF = 1 / (1 - R2)

If R2 is practically 1, then VIF is infinity. R2 will be almost 1 when there is a very strong (almost perfect) correlation between 2 variables.

1. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

This Q-Q or **Quantile-Quantile** is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we **can infer if the data comes from a normal distribution**. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.

