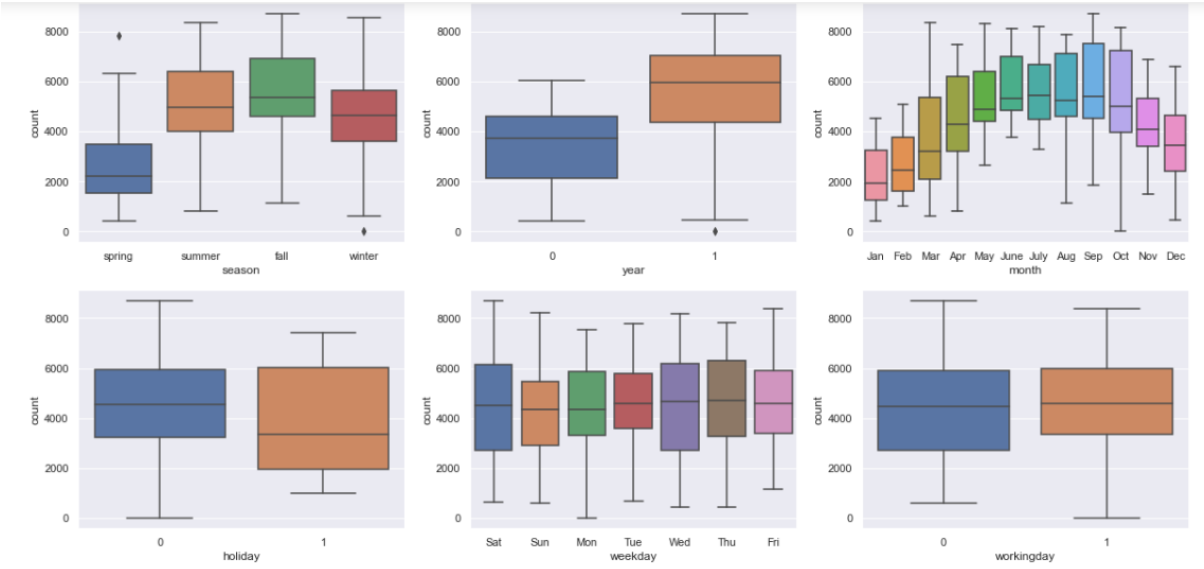
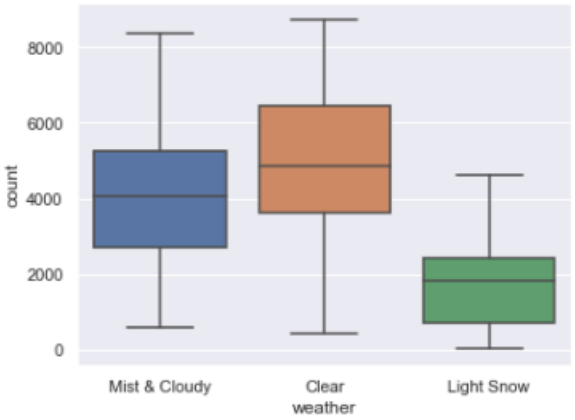
Assignment-based Subjective Questions

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

Answer: Categorical Variable is dataset are as below

- **Season**
- **Year**
- **Month**
- **Holiday**
- **Weekday**
- **Working day**
- **Weather**





Inference: -

- Fall season (season 3) has Highest demand for rental bikes while, spring seaso n has lowest.
- It's observed that demand for next year (from 2018 to 2019) has increased.
- Bike Sharing Demand is continuously increasing from January to June, while September month has highest demand among all. After September demand start decreasing.
- Whenever there is holiday started demand has been decreased.
- Weekday has almost same demand. It's not much concluding demand during weekday
- Its looks like similar demand during working day and non-working day, but stil l working day have more demand
- During September month Bike sharing demand is more. While during year beginning and end its less demand. It could be due to Extreme weather conditions.

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

Answer: drop_first=True helps to reduce extra created column during dummy v ariable creation, which helps in reducing redundancy among dummy variables while
Correlation.

- If we have n level of categorical variables then we need to use n-1 colum ns to represent the dummy variable.
- For example:
  While finding the house is furnished, semi-furnished or unfurnished, wit hout
  Using drop_first = True we needed three columns as shown below

```
status = pd.get_dummies(housing_data['furnishingstatus'])
status.head()
```

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

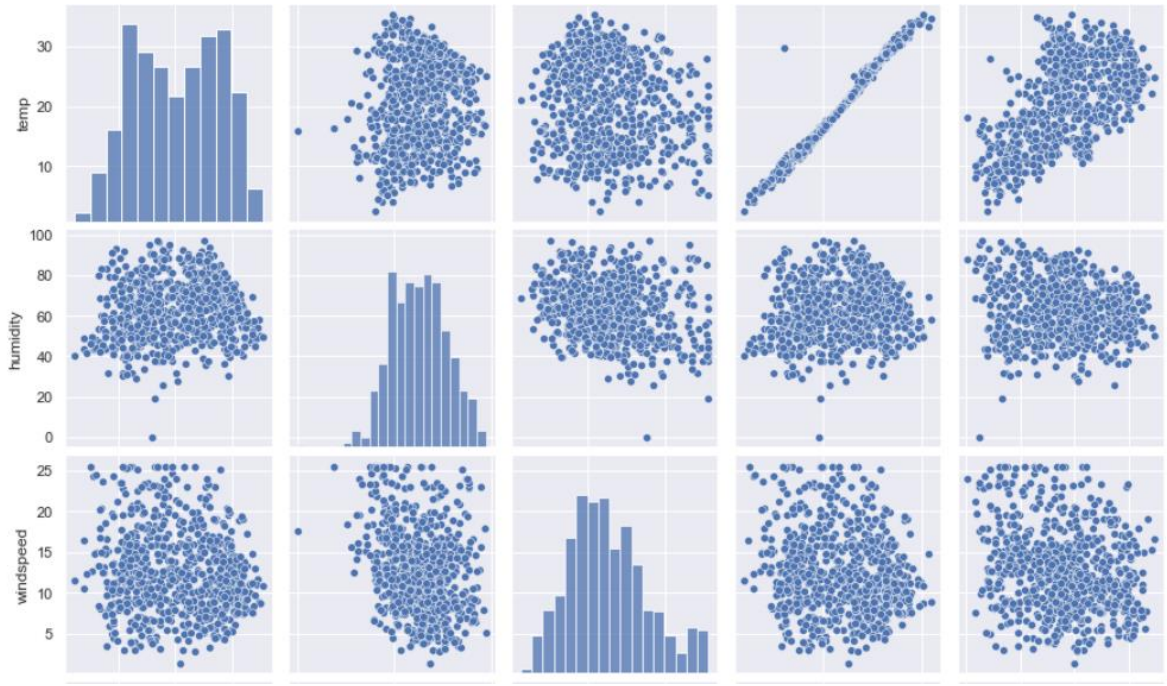
- Here we don't need unfurnished columns as Furnished and Semi-furnished together can represent status of unfurnished which is '00'
- After Using drop_first = True

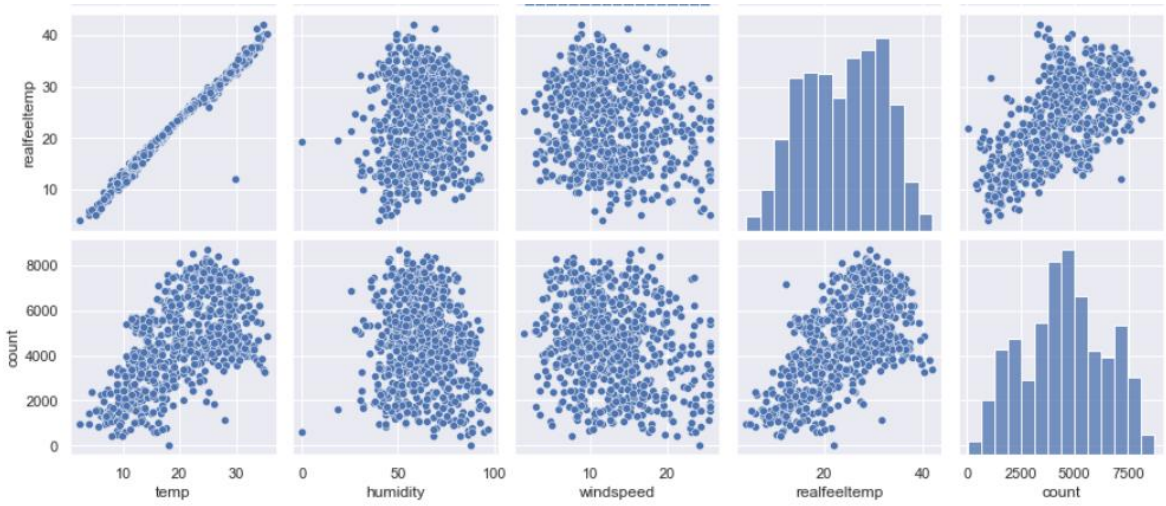
```
_dummy_vars = pd.get_dummies(housing_data['furnishingstatus'], drop_first=True)
_dummy_vars.head()
```

	semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0

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

Answer:



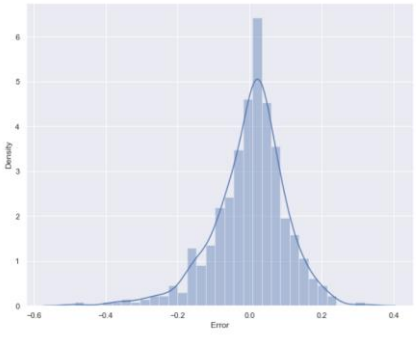


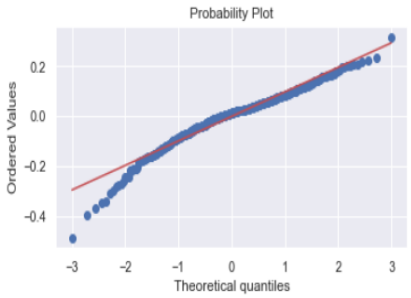
From above pair plot we can clearly conclude that there is highest correlation between "temp", "atemp" with "count". They are almost linear to the target variable.

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

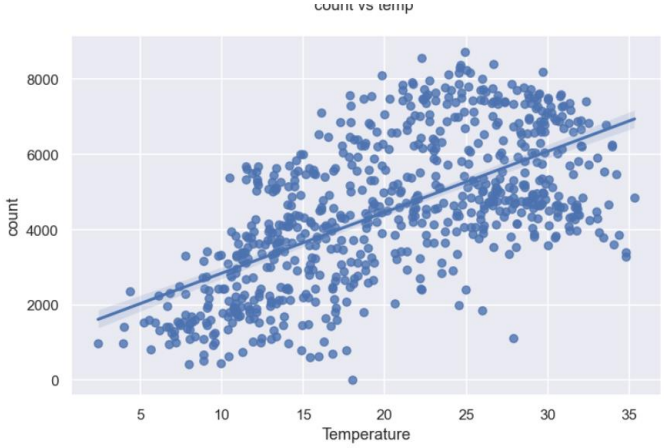
Answer: There are 5 assumptions of the linear regression

a) Normality: The residual should be normally distributed. Same is shown below in the qq plot:

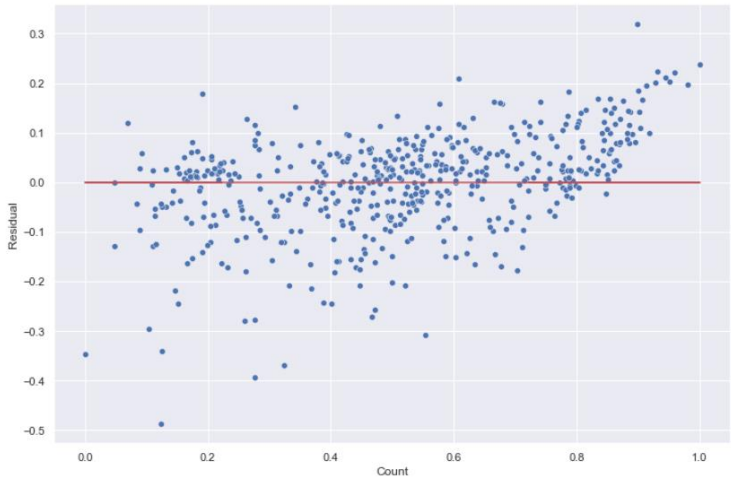




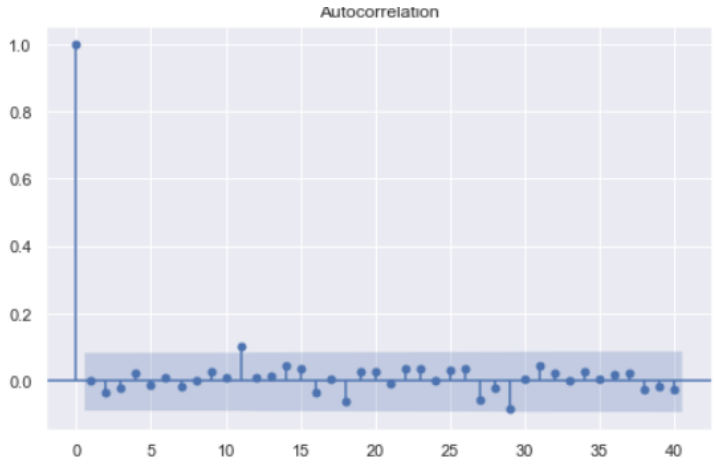
b) **Linearity**: The relationship between dependent variable and independent variable should be linear.
As shown below in scatter plot:



c) **Homoscedasticity**: Error should be constant across all the independent variable.



d) **Auto-correlation**: There should be no auto correlation between the errors.



e) **Multicollinearity**: There should be no correlation between independent variables. It can be verify using VIF.

	Features	VIF
8	windspeed	4.81
7	temp	4.07
6	year	2.03
4	spring	1.74
5	winter	1.48
0	Dec	1.21
2	Sun	1.17
1	Sep	1.15
3	Light Snow	1.05

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

Answer: Based on final model (model 7) top three features contributing significantly towards explaining the demand are:

1. temp
2. Weather Situation: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light and Rain + Scattered clouds
3. Year

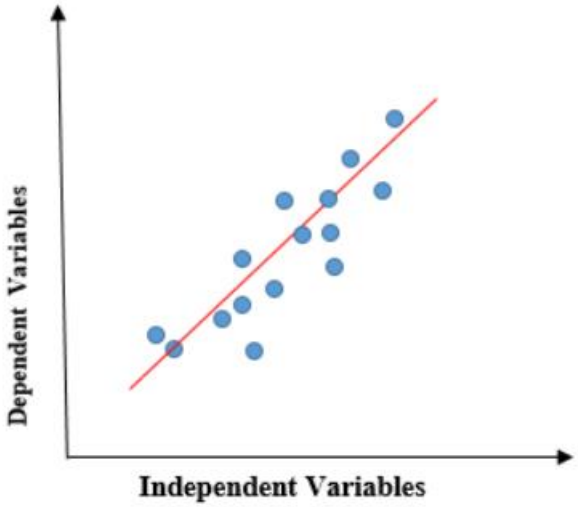
General Subjective Questions

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

Linear Regression: Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.

The two types of regression are:

a. **Simple linear Regression**: If there is a single input variable (x), such linear regression is called **simple linear regression**
b. **Multiple Linear regression:** if there is more than one input variable, such linear regression is called **multiple linear regression**.



- The above graph presents the linear relationship between the dependent variable and independent variables.
- When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing.
- The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best
- to calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = \beta_0 + \beta_1 X + \varepsilon$$

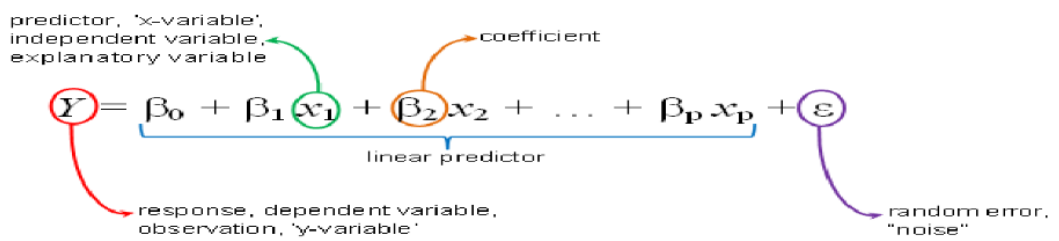
Y = dependent Variable or Target Variable. B_0 = Intercept of line.

X = independent variable or predictor variable. B_1 = linear regression coefficient.

ε = random error.

- For multiple linear regression:

It's represent as



2. Explain the Anscombe's quartet in detail.

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.

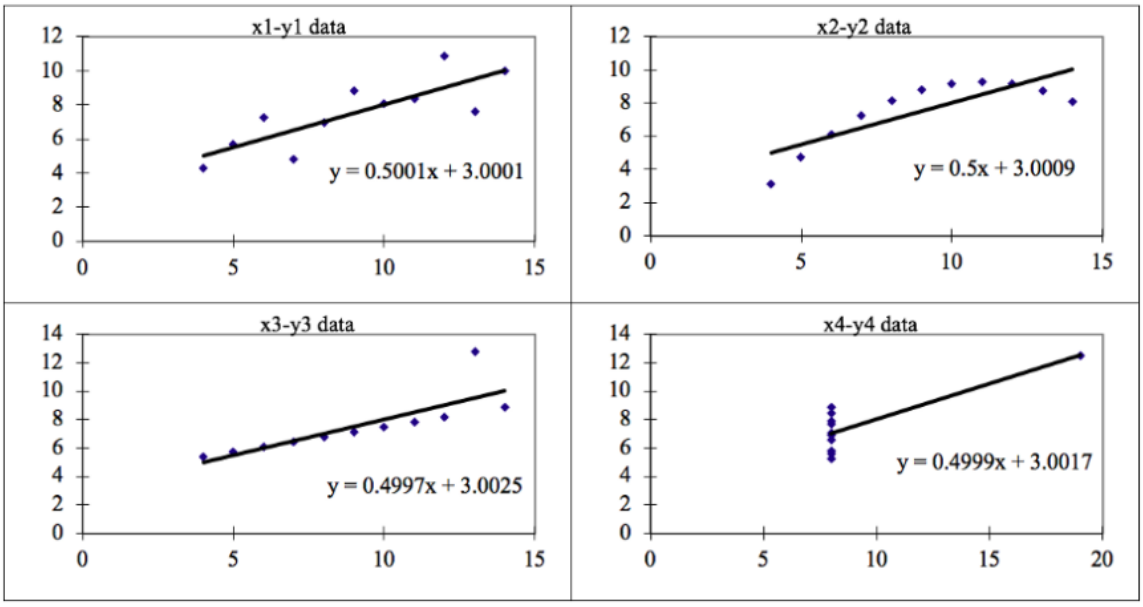
These four plots can be defined as follows:

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

- The statistical information for all these four datasets are approximately similar and can be computed as follows:

				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	

- When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

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

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Requirements for Pearson's Correlation Coefficient

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

Where,

N = the number of pairs of scores

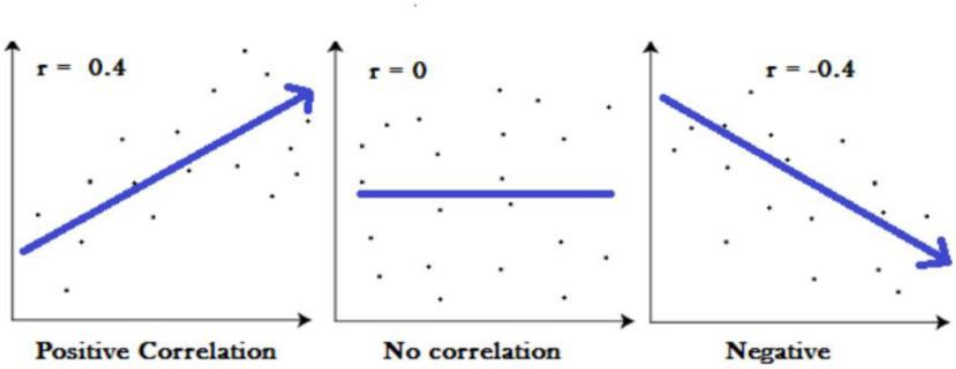
Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores



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

Scaling (Feature 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 it is performed: Real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

- **Techniques to perform Feature Scaling**

The two most important ones scaling techniques are as below:

- **Min-Max Normalization**: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization**: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

Normalized scaling	Standardized scaling
Used when features are of different scale	Used when we want to ensure 0 mean and unit std
Scales value between [0,1] or [-1,1]	It's not bounded to any range.
It's affected by outliers	Very less affected by outliers
It's also called as scaling normalization	It's called as z-score normalization
it useful when we have no idea about distribution	Useful when features distribution is normal/gausian
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
In python we use MinMaxScaler()	In python we use Standardscaler()

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

   Whenever there is perfect correlation then VIF will be infinite.
   From the formulae if R^2 will be 1 then VIF will be infinity. In this case we have to drop the variable to avoid multicollinearity.

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

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

- The advantages of the q-q plot are:
  - The sample sizes do not need to be equal.
  - Many distributional aspects can be simultaneously tested

- Below code is use to plot Q-Q plot.

In [592]:
```
# plotting qq plot

sm.qqplot(res, fit=True, line='45')
plt.show()
```