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)

   [Ans:]

   - Bike rentals are good on non-Holidays.
   - Bike rentals are slightly preferred more on working days.
   - During falls bike rentals are more preferred.
   - Clear weather is more preferrable for bike rentals.

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

[Ans:] This helps in reducing the number of dummy variables as one of variable can be inferred from values of other variables. For example: Color of Bike can be red, green or blue.

For this only 2 variables can be used and presence of third one can be inferred. So if red=0 and blue=0 then it can be inferred that bike color is green. In case of lot of features/dummy-variables in the dataset **drop_first=True** helps in reducing the number of predictors.

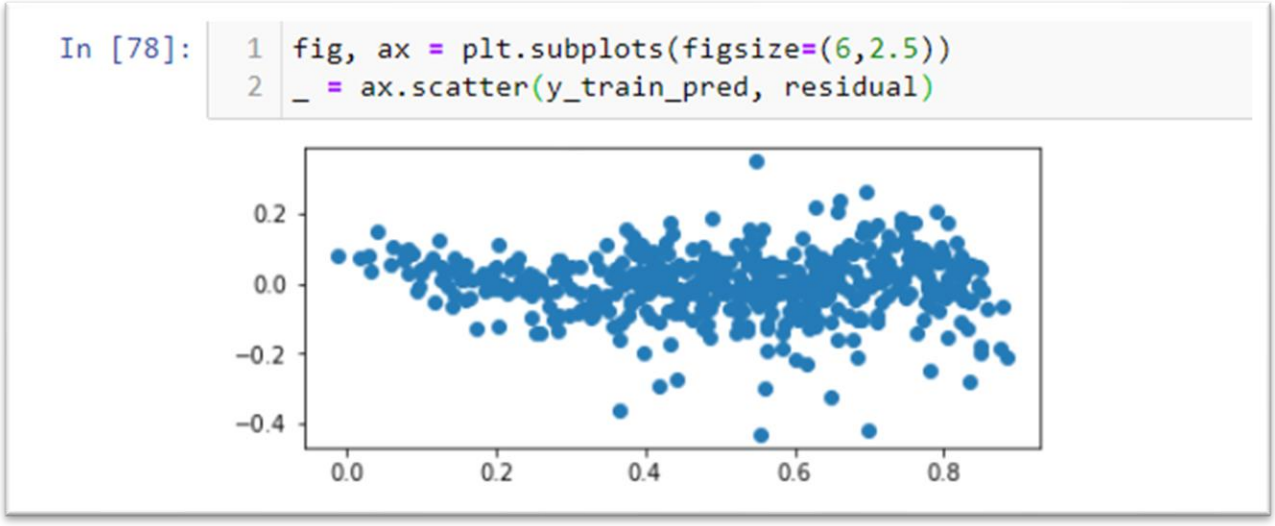
2. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   [Ans] : temp

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

   [Ans]:

   1. Analyzing the scatter plot between features and target variable.
   2. Plotting a scatter plot between predicted target variable and error term and verifying error term is spread uniform across all values of target variable.



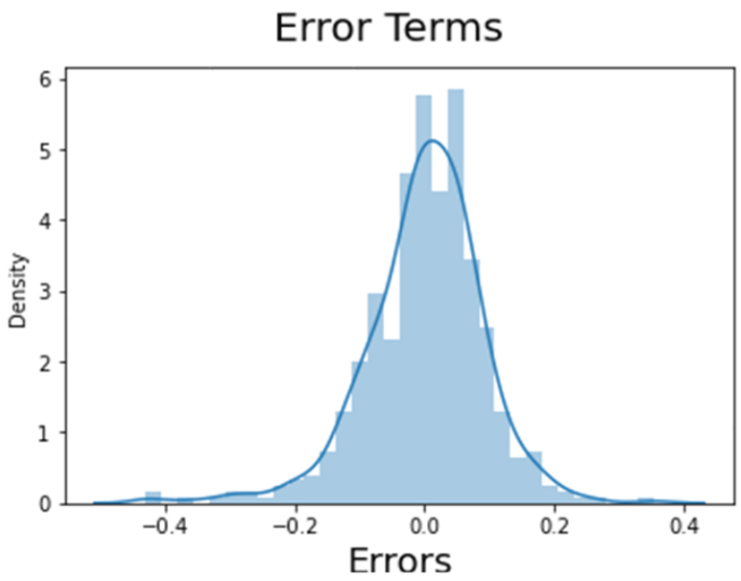
   3. Checking the distribution of error term to follow normal distribution.

In [71]:

```
# Plot the histogram of the error terms
fig = plt.figure()
residual=y_train - y_train_pred

sns.distplot((residual))
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

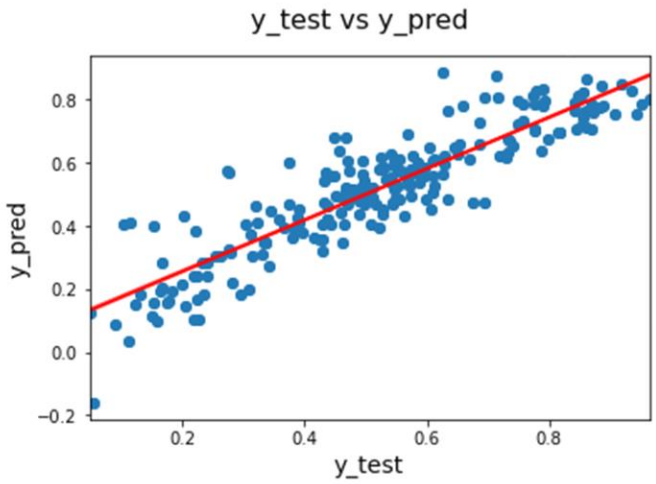
Out[71]: Text(0.5, 0, 'Errors')



4. Checking linear relation between predicted target variable and actual test target variable to follow linear model.

In [65]:

```
fig = plt.figure()
sns.regplot(x=y_test, y=y_pred, ci=52, fit_reg=True, line_kws={"color": "red"})
plt.scatter(y_test, y_pred)
fig.suptitle('y_test vs y_pred', fontsize = 16)
plt.xlabel('y_test', fontsize = 14)
plt.ylabel('y_pred', fontsize = 14)
plt.show()
```



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

- Temperature
- Year
- Season

General Subjective Questions

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

   [Ans] : Linear regression algorithm is a type of supervised Machine learning algorithm which provides a linear relationship between dependent and one of more independent variables. As the name suggests the variables should be linearly correlated which means it provides a sloped straight line representing the relationship between variables.

   There are 2 types of linear regression models

   a. Simple Linear regression - linear relation of one feature variable with target variable.
   b. Multiple linear regression - linear relation of more than one feature variable with target variable.

   Linear equation of regression line is represented using following formula :

   Simple Linear regression : $Y = \beta_0 + \beta_1 x$

   Multiple Linear regression : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_n x_n$

   Where β_0 = Intercept and β_1 = Slope.

Which means the residuals should be minimized to get a best fit line. Residual is the difference of predicted target variable and actual target variable.

This used where the target variable is a continuous variable.

2. Explain the Anscombe's quartet in detail. (3 marks)

   [Ans:] Anscombe's quartet states that visualization helps in identifying anomalies in the data which may not be visible using statistical methods. It emphasizes on using data visualization to analyze data before starting statistical analysis. It consists of 4 data set with each data set consisting of 11 (x,y) points. The peculiar properties of these data-sets is that they exhibit same statistical data like mean, standard deviation, etc but graph shows different behavior. Btw, does not seem to be covered in any session.

3. What is Pearson's R? (3 marks)

   [Ans:] Pearson's R or Pearson's correlation coefficient is a method of measuring a linear correlation. It tells how strongly variables are linearly associated. It varies between -1 and 1 and the values tells their association strength.

   A values closer to -1 shows that variables are inversely proportional (increase in one variable causes decrease in another) also called negatively correlated.

   A values closer to 1 shows that variables are directly proportional (increase in one variable causes increase in another) also called positively correlated.

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

   [Ans] Scaling is a method of bringing the value of each variable to same scale before feeding the data for model building. Its performed so that coefficients do not vary drastically between variables which will cause incorrect analysis of variables.

   Normalized scaling also called min-max scaling scales/standardizes the value between 0 and 1.

   Standardized scaling also called z-score scaling scales the values as per their z-score.

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

   [Ans] VIF of infinite means variables are strongly correlated. VIF = $1/(1-R^2)$ and when variables are stringly correlated then R^2 tends to go closer to 1 sure to which $1-R^2$ tends towards 0 hence ViF gets higher value. In case of perfect correlation $R^2 = 1$ hence VIF = ∞.

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

   [Ans] Q-Q plot (Quantiles-to-Quantiles plot) is a scatter plot to plot quantiles of two distributions with respect to each other. It helps determine if a dataset follows normal, uniform or exponential distribution. In one sense it checks how much skewed is the data from normality. In q-q plot x-axis represent the theoretical quantiles and y-axis actual quantiles.