General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical method used for predictive modelling and analysis. It aims to model the relationship between a dependent variable (also called the response variable or output) and one or more independent variables (also known as predictors or inputs). The key idea is to find the best-fitting straight line (in the case of simple linear regression) or hyperplane (in the case of multiple linear regression) that can predict the dependent variable based on the independent variables.

Key Concepts of Linear Regression

1. .**Dependent and Independent Variables:**
   Dependent Variable (Y): This is the variable we are trying to predict or explain.
   Independent Variable(s) (X): These are the variables that we use to make predictions about the dependent variable. In simple linear regression, there is one independent variable, while in multiple linear regression, there are multiple independent variables.

2. **Equation of the Linear Model:**

   Simple Linear Regression:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept (the value of Y when $X = 0$).
- β_1 is the slope coefficient (the change in Y for a one-unit change in X).
- ϵ represents the error term, capturing the difference between the actual and predicted values.

- **Multiple Linear Regression:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

- Y is the dependent variable.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables.
- ϵ is the error term.

3. **Objective**
   The goal of linear regression is to find the values of the coefficients (β0,β1,…,βn\beta_0, \beta_1, \dots, \beta_nβ0 ,β1,…,βn) that minimize the sum of the squared differences between the actual and predicted values of the dependent variable. This method is called Ordinary Least Squares (OLS).

4. **Cost Function**
   The cost function used to measure the accuracy of the model is the Mean Squared Error (MSE), which is the average of the squared differences between the actual and predicted values

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- n is the number of observations.
- Y_i is the actual value of the dependent variable.
- $\hat{Y}_i$ is the predicted value of the dependent variable.

5. **Fitting the Model**

   To fit a linear regression model, the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are estimated in such a way that the MSE is minimized. This is usually done using optimization techniques like gradient descent or by directly solving the OLS equations.

6. **Assumptions of Linear Regression**
   1.Linearity: The relationship between the dependent and independent variables is linear.
   2.Independence: The residuals (errors) are independent.
   3.Homoscedasticity: The residuals have constant variance.
   4.Normality: The residuals of the model are normally distributed.
   5.No Multicollinearity: In multiple linear regression, the independent variables should not be highly correlated with each other.

7. **Model Evaluation:**
   After fitting the model, it's important to evaluate its performance. Common metrics include:
   1. R-squared (R^2): Represents the proportion of variance in the dependent variable that can be explained by the independent variables.
   2. Adjusted R-squared: Adjusts the R^2 value for the number of predictors, providing a more accurate measure in multiple regression.
   3. Root Mean Squared Error (RMSE): The square root of MSE, providing an estimate of the standard deviation of the prediction errors.
   4. F-statistic: Tests whether at least one predictor variable has a non-zero coefficient.

8. **Extensions of Linear Regression:**
    1. Polynomial Regression: When the relationship between the dependent and independent variables is non-linear, polynomial regression (which is a special case of linear regression) can be used by adding polynomial terms of the independent variables.
    2. Ridge and Lasso Regression: These are regularization techniques that add a penalty to the cost function to prevent overfitting by shrinking the coefficients. Ridge regression uses L2 regularization, while Lasso uses L1 regularization.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that are famous for illustrating the importance of data visualization in statistical analysis. The quartet was created by the British statistician Francis Anscombe in 1973 to demonstrate how different datasets can have identical or very similar summary statistics, yet be strikingly different when graphed. The purpose of Anscombe's quartet is to emphasize that relying solely on summary statistics (like mean, variance, correlation, etc.) without visualizing the data can lead to misleading conclusions.
The Four Datasets of Anscombe's Quartet
Each dataset in Anscombe's quartet consists of 11 data points (pairs of x and y values). Despite their different distributions, all four datasets share the following identical summary statistics:

- Mean of x: 9.0
- Mean of y: 7.5
- Variance of x: 11.0
- Variance of y: 4.125
- Correlation between x and y: Approximately 0.816
- Linear regression line (y = mx + c): y=3+0.5xy = 3 + 0.5xy=3+0.5x

Lessons from Anscombe's Quartet

1. Importance of Data Visualization: Simply relying on summary statistics can be misleading. Visualizing data helps to understand its true nature and avoid incorrect conclusions.

2. Influence of Outliers: Outliers can have a significant impact on statistical measures like the mean, variance, and correlation. In regression, they can disproportionately influence the slope and intercept of the fitted line.

2. Non-Linearity: Not all relationships are linear. Data that appears to have a strong linear correlation might actually follow a non-linear trend, which could be missed if only summary statistics are considered.

2. Misleading Statistics: Identical summary statistics can be observed in very different datasets, highlighting the limitation of using these statistics alone to understand data.

3.What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It is one of the most widely used correlation coefficients in statistics.
Definition
Pearson's R quantifies how well two variables are linearly related. It ranges from -1 to 1, where:

1. +1 indicates a perfect positive linear relationship (as one variable increases, the other increases proportionally).
2. -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases proportionally).
3. 0 indicates no linear relationship between the variables.

Formula

Pearson's R is calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Where:

- X_i and Y_i are the individual data points for the variables X and Y.
- $\overline{X}$ and $\overline{Y}$ are the means of the variables X and Y.
- n is the number of data points.

Interpretation

- r=1r = 1r=1: Perfect positive correlation. Every increase in one variable is associated with an exact proportional increase in the other variable.
- r=−1r = -1r=−1: Perfect negative correlation. Every increase in one variable is associated with an exact proportional decrease in the other variable.
- r=0r = 0r=0: No correlation. There is no linear relationship between the variables.
- 0<r<10 < r < 10<r<1: Positive correlation. As one variable increases, the other variable tends to increase as well.
- −1<r<0-1 < r < 0−1<r<0: Negative correlation. As one variable increases, the other variable tends to decrease.

'Assumptions and Considerations

1. Linearity: Pearson’s R assumes that the relationship between the two variables is linear. If the relationship is non-linear, Pearson's R may not accurately reflect the strength of the relationship.
2. Scale: Pearson’s R is sensitive to the scale of the data. Both variables should be continuous and measured on an interval or ratio scale.
3. Outliers: Pearson’s R can be significantly affected by outliers. A single outlier can greatly influence the correlation coefficient, leading to misleading results.
4. Bivariate Normality: Ideally, both variables should follow a bivariate normal distribution, although this is more critical when conducting significance tests related to Pearson's R.

Application

Pearson's R is commonly used in various fields, including social sciences, natural sciences, and finance, to measure the degree of linear association between two variables. For example, it can be used to determine how closely related students' test scores are to the amount of time they spend studying, or to analyse the relationship between stock returns and market indices.

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

Scaling is the process of adjusting the range of data values so that different features (variables) have comparable ranges. In machine learning and statistical analysis, scaling is crucial because many algorithms are sensitive to the scale of the input data. For example, features with larger ranges can disproportionately influence the model, leading to suboptimal performance.

Why is Scaling Performed?

Scaling is performed for several reasons:

1. **Improved Model Performance**: Some machine learning algorithms, such as gradient descent-based methods (e.g., linear regression, logistic regression, neural networks), distance-based algorithms (e.g., K-Nearest Neighbors, K-Means clustering), and support vector machines, perform better when features are on a similar scale. This is because these algorithms are sensitive to the relative magnitudes of feature values.
2. **Faster Convergence**: In optimization algorithms like gradient descent, scaling can lead to faster convergence. Without scaling, features with larger ranges can cause the algorithm to oscillate, leading to slower convergence.
3. **Fair Contribution of Features**: Scaling ensures that all features contribute equally to the model. Without scaling, features with larger ranges could dominate the learning process, leading to biased models.
4. **Prevention of Numerical Instability**: In some cases, very large or very small feature values can cause numerical instability, leading to errors in computation. Scaling helps mitigate this risk.

Key Differences Between Normalization and Standardization

Aspect	Normalization (Min-Max Scaling)	Standardization (Z-Score Scaling)
Range of Values	Typically [0, 1] or [-1, 1]	Mean of 0, standard deviation of 1
Formula	$\frac{X - X_{\min}}{X_{\max} - X_{\min}}$	$\frac{X - \mu}{\sigma}$
Effect on Outliers	Sensitive to outliers; outliers can dominate the scaling	Less sensitive to outliers compared to normalization
Use Case	When a specific range is needed, like in KNN, neural networks	When features are normally distributed, as in regression models
Effect on Distribution	Changes the distribution shape (doesn't preserve normality)	Preserves the distribution shape (centers around 0)

When to Use Normalization vs. Standardization

- **Normalization**: Use when you need to bound data within a specific range, especially in cases where the data is not normally distributed and you do not require a mean-centered distribution. Also, when you expect the data to have outliers that should influence the scaling minimally.
- **Standardization**: Use when your data follows a normal distribution or when the algorithm assumes a normal distribution. It is also appropriate when you are dealing with algorithms that are sensitive to the mean and variance of the data.

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

A Variance Inflation Factor (VIF) can be infinite when there is perfect multicollinearity between variables, or when one variable can be expressed as a linear combination of other variables:

Perfect multicollinearity

When the regressor is equal to a linear combination of other regressors, the VIF tends to infinity.

Linear combination

When a variable can be expressed as a linear combination of other variables, the corresponding variable and other variables will have an infinite VIF.

A large VIF indicates a correlation between variables, and the higher the VIF, the greater the degree of multicollinearity. For example, a VIF of 4 means that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity. While there is no precise rule for deciding when a VIF is too high, values above 10 are often considered a strong hint that reducing multicollinearity might be worthwhile

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

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically a normal distribution. The purpose of a Q-Q plot is to assess whether a given dataset follows a specific distribution.

What is a Q-Q Plot?

- Definition: A Q-Q plot plots the quantiles of the sample data against the quantiles of a theoretical distribution. If the data follow the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.
- Axes:
  - X-axis: Theoretical quantiles (based on the assumed distribution, e.g., normal distribution).
  - Y-axis: Sample quantiles (calculated from the actual data).

Key Assumptions in Linear Regression

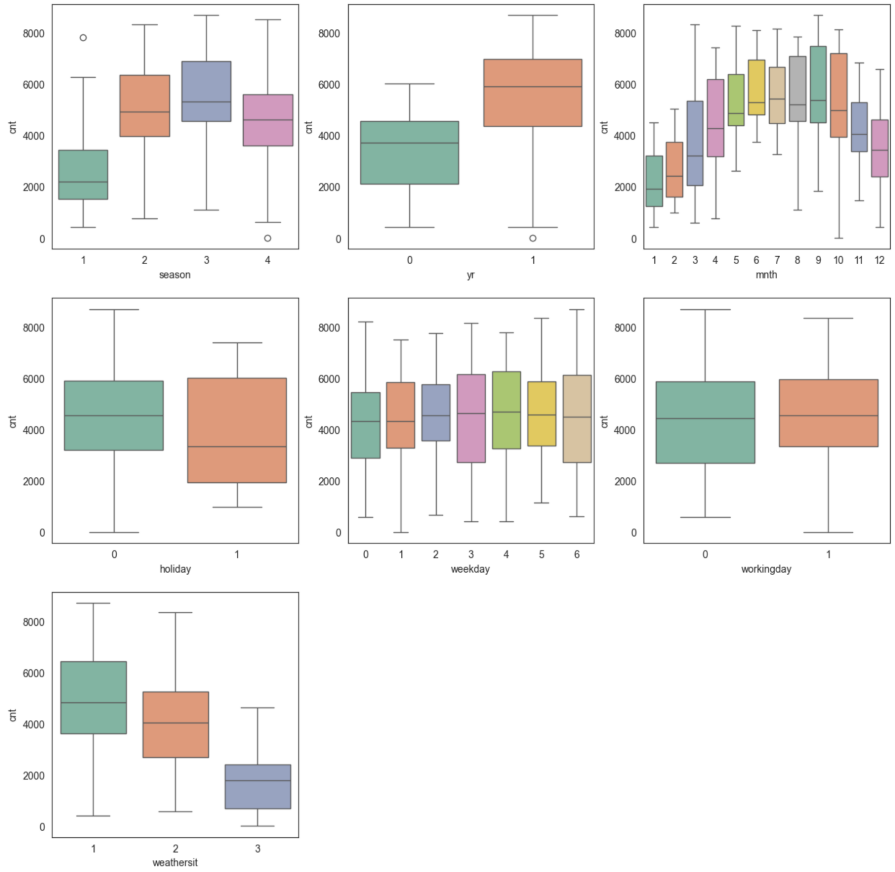
1. Linearity: The relationship between the independent and dependent variables should be linear.
2. Independence: Observations should be independent of each other.
3. Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.
4. Normality of Residuals: The residuals should be normally distributed.

Importance of a Q-Q Plot in Linear Regression

1. Checking Normality of Residuals: One of the key assumptions in linear regression is that the residuals are normally distributed. A Q-Q plot of the residuals can help you visually assess this assumption. If the residuals are normally distributed, the points on the Q-Q plot will lie close to a straight line.
2. Identifying Deviations from Normality: Deviations from the straight line in the Q-Q plot can indicate issues such as skewness, heavy tails, or the presence of outliers in the residuals. These deviations suggest that the residuals are not normally distributed, which could invalidate some of the inferences drawn from the regression model (e.g., confidence intervals, hypothesis tests).
3. Model Diagnostics: If the Q-Q plot indicates non-normality of residuals, it may suggest the need for transforming the dependent variable, adding polynomial terms, or using a different regression model altogether (e.g., robust regression) to improve model fit.
4. Detection of Outliers: The Q-Q plot can also help in identifying outliers, which appear as points that are far from the 45-degree line. Detecting and handling outliers is crucial for improving model accuracy and reliability.

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?



cnt = (0.0839) + (yr * 0.2337) + (temp * 0.5719) - (windspeed * 0.1474) - (weathersit_Light_Snow * 0.2502) + (season_summer * 0.0824) + (season_winter * 0.1251) + (mnth_sept * 0.0868)

1. Season: 3:fall has highest demand for rental bikes
2. Bike renting has increased in 2019 when compared to 2018, It indicated positive growth in business in 2019 compared to 2018.
3. Demand is continuously growing each month till June. September month has highest demand. After September, It decline as the climatic condition getting deteriorated
4. Bike renting is observed making good business during weekdays/working days when compared to Weekends/Holidays. This indicates Its an good/convenient commute option for people to travel during working days
5. The demand of the bike renting is almost similar across all the working days,
6. Bike renting seems to be a popular option during the good climatic condition.
7. During September, bike sharing is more. During the year end and beginning, it is less, could be due to extreme weather conditions.
8. Bikes are rented during moderate Windspeed and low Humidity

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

Reducing Redundancy: By discarding one dummy variable, we effectively remove the redundancy in the model caused by perfect multicollinearity. This allows the regression model to estimate the coefficients of the remaining dummy variables accurately and interpret their effects on the outcome variable independently

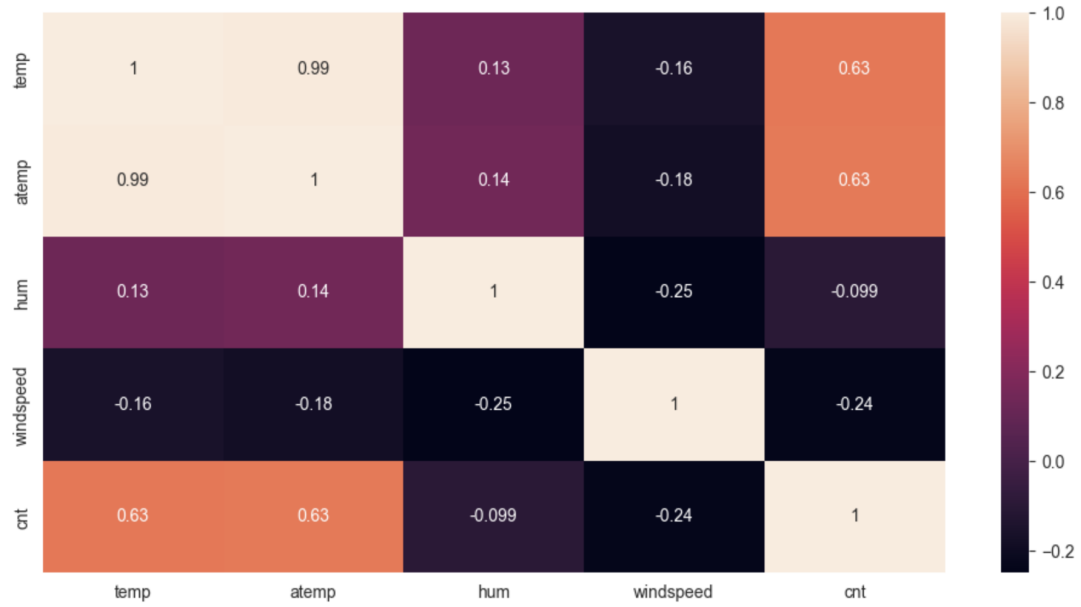
1. It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

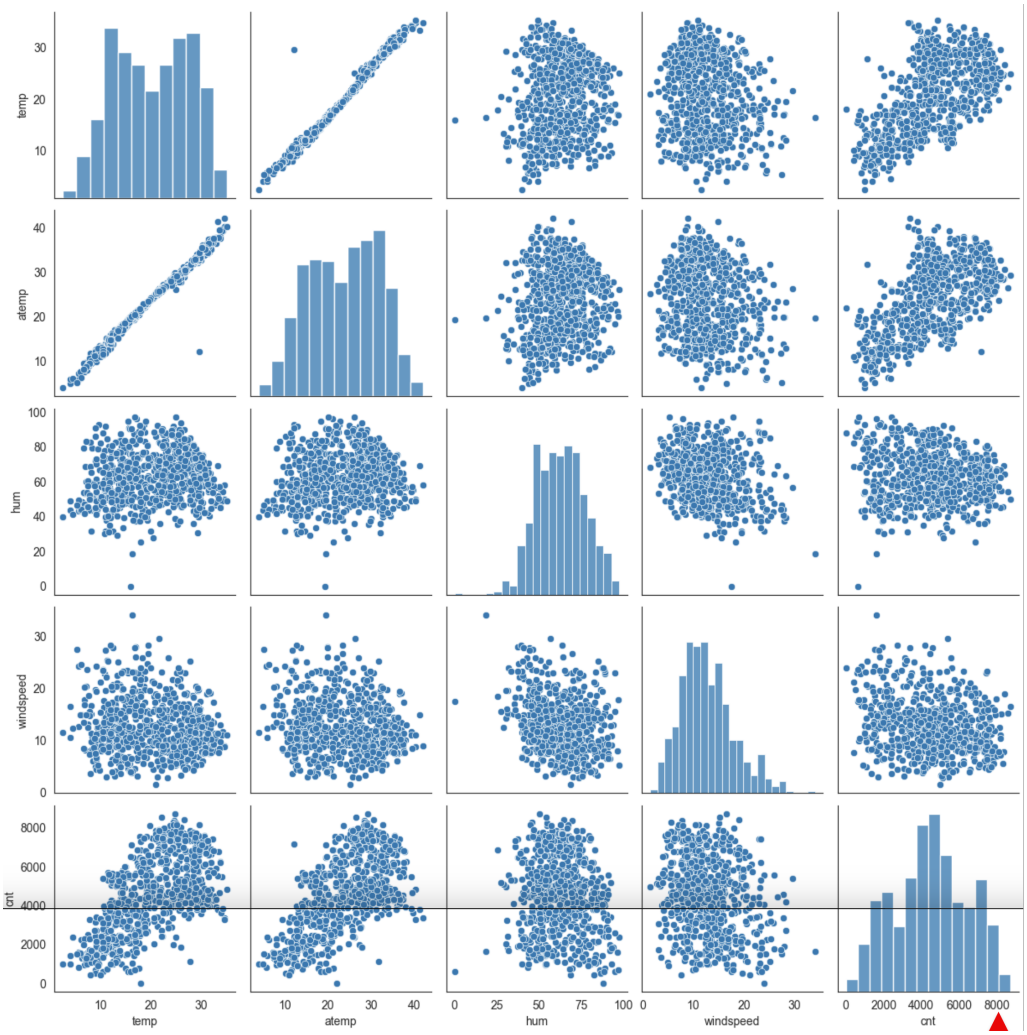
2. For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

3. It is also used to reduce the collinearity between dummy variables .

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

atemp and **temp** both have same correlation with target variable of **0.63** which is the highest among all numerical variables





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

I have validated the assumption of Linear Regression Model based on below 5 assumptions –

1. Normality of error terms
   - Error terms should be normally distributed
2. Multicollinearity check
   - There should be insignificant multicollinearity among variables.
3. Linear relationship validation
   - Linearity should be visible among variables
4. Homoscedasticity
   - There should be no visible pattern in residual values.
5. Independence of residuals
   - No auto-correlation

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes
1. temp
2. year
3. Season Winter

cnt = (0.0839) + **(yr * 0.2337)** + **(temp * 0.5719)** - (windspeed * 0.1474) - (weathersit_Light_Snow * 0.2502) + (season_summer * 0.0824) + **(season_winter * 0.1251)** + (mnth_sept * 0.0868)

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.800
Model:	OLS	Adj. R-squared:	0.797
Method:	Least Squares	F-statistic:	287.1
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	5.34e-171
Time:	12:22:42	Log-Likelihood:	449.48
No. Observations:	510	AIC:	-883.0
Df Residuals:	502	BIC:	-849.1
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0839	0.018	4.790	0.000	0.049	0.118
yr	0.2337	0.009	25.908	0.000	0.216	0.251
temp	0.5719	0.022	26.576	0.000	0.530	0.614
windspeed	-0.1474	0.027	-5.366	0.000	-0.201	-0.093
weathersit_bad	-0.2502	0.027	-9.314	0.000	-0.303	-0.197
season_summer	0.0824	0.011	7.317	0.000	0.060	0.105
season_winter	0.1251	0.011	11.004	0.000	0.103	0.147
mnth_sept	0.0868	0.017	5.035	0.000	0.053	0.121

Omnibus:	64.591	Durbin-Watson:	1.985
Prob(Omnibus):	0.000	Jarque-Bera (JB):	112.216
Skew:	-0.778	Prob(JB):	4.29e-25
Kurtosis:	4.691	Cond. No.	9.78