DAO2702 Programming for Business Analytics

Group Members: A02

CHUA MIN YI	**A0187724U**
LIM DING NENG	**A0183970W**
SOUMIL BANERJEE	**A0176385R**

Group Project

Bike Sharing in London

21st November 2019 | Semester 1 - AY 2019/2020

Introduction

Being a well-known bike sharing company in London, ShareBike, for 3 years, our vision is to provide the best rider's experience. This aligns with our primary objective to build the trust that our consumers have in us to provide a high standard of bikes and never fail to provide a bike when they need one.

There are several elements or considerations we need to take into account to sustain our high-quality standard and deliver our promise. For instance, regular maintenance, ensures a sufficient number of bikes at each period and reducing wasted bikes during low peak seasons.

In recent months, there is a rise in the price of public transport that resulted in an **increased demand** for bike sharing services. Even though it's an opportunity for us, competition arise as more firms entered this market. Despite the rising competition, our firm believes that our consumers would continue to use our service if we are able to maintain **the trust built** over the years.

However, in the recent months, many consumers have given us feedback that the frequency of our bikes having faulty brakes and flat tires has been increasing and we aim to resolve these problems promptly. To do so, we plan to **replace the old maintenance schedule**. Thus, we will analyze the data to effectively predict the patterns of the usage of our bike sharing service to **come up with a more well-defined schedule.**

Business Question To Be Analyzed

1. Whether weekdays or weekends are more feasible to carry out extensive maintenance?
2. Which time period within a day should the maintenance be?
3. Which season of the year and what percentage of the bike should be stored, if they are idle?

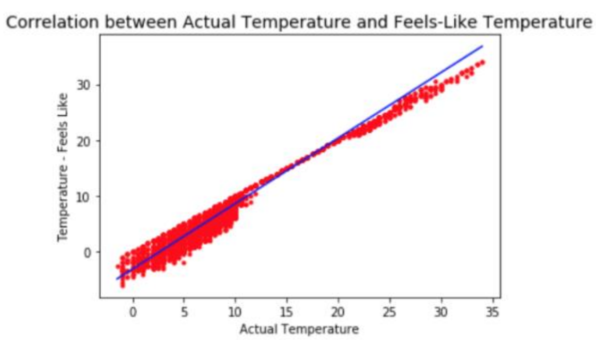
Data Source

The CSV file was sourced from Kaggle and it is a historical record of bike sharing data over a period of 2 years. The data is acquired from 3 credible sources and were vouched to be accurate given that they are of credible institutions. The data set consists of a sufficient sample size of 17,414 for our data analysis. The link for the data is: https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset

Reasoning for Variable Selection

To start off, we conducted a regression analysis (Appendix 1) where the dependent variable (Y) is the count of bikes used in a given hour and the predictors are all the variables given to us, namely season, weather, humidity percentage, holiday indicator, weekend indicator, time grouped by 6 hour indicators, feels-like temperature and real temperature. These are the factors that could possibly affect the bike usage. With the **F-Statistic P-value less than 0.05**, it shows that this is a good fit model. However, there are a few weather indicators which are not statistically significant and thus, causing collinearity of variables. Hence, we hypothesize that the weather variables are correlated with season. This correlation between weather and seasons can be understood intuitively as a weather condition such as snow would most likely occur during winter, with rain in fall or winter and clear skies or scattered clouds in summer. Therefore, **they are not very good indicators of count of bikes used** and are removed from the next regression.

Additionally, we also believe that there might be a **high correlation between the actual temperature and the 'feels-like' temperature**, suggesting that there is no additional benefit from adding an extra variable to the regression. Thus, we visualize the relationship between temperature and temperature feels-like.



We calculate the correlation coefficient to be 0.988. Thus, we remove the 'feels-like' temperature and the weather indicators to get the regression results as seen in Appendix 2. Although the R-square falls, all the **coefficients are all statistically significant**, thereby suggesting that removing the weather variables improves the predictability of the bikes used in a given hour. Removing the temperature feels-like helps us **escape the problem of multicollinearity**.

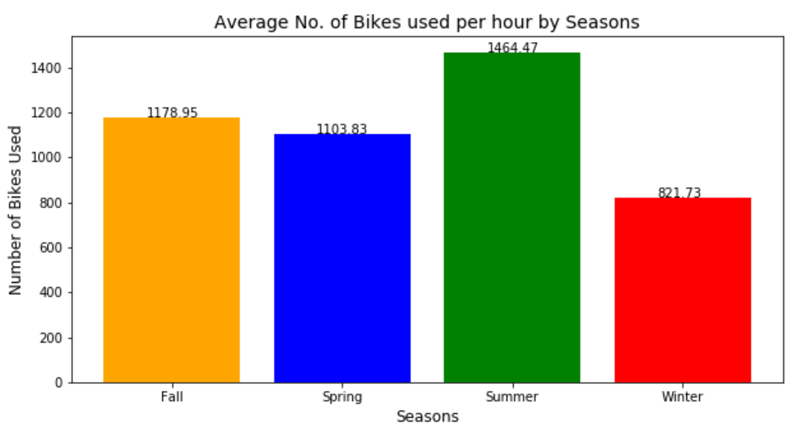
Next, we decide to try and break up the time periods further to see if the predictability increases. We run a regression by breaking up time into 4 hours each and thus have 6 groups within a day as shown in Appendix 3. Although R-squared increases drastically, we run into a huge standard deviation on the coefficient on the Summer Indicator variable suggesting that this grouping is not optimal. Hence, we look for a different grouping and group the hours into **8 groups of 3 hours** each and this regression is seen in Appendix 4.

On adding interaction terms between the categorical terms, we get large p-values and hence these interaction terms are not statistically significant. Therefore, we believe that the regression in Appendix 4 is optimal, as **all the coefficient on the variables are statistically significant and the R-square is maximized.**

However, as our business problem states, to effectively plan out our schedule of maintenance, we would only focus on seasons, days of the week and time period as these variables are easier to predict and have lesser variability. Although humidity, temperature and wind speed are good regressors, **these factors are very volatile**, making it difficult to carry out business decisions based on them. On the other hand, we can anticipate the season of the year perfectly while we are always certain of the day of the week that we are in. Thus, having solutions and suggestions based on these variables would lead to better business decisions.

Data Visualization and Analysis

1. Number of Bikes Used per Hour by Seasons

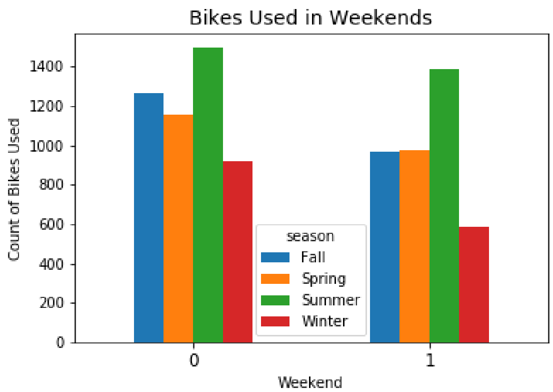


Based on our findings, bike sharing peaks during the summer and troughs during the winter period. Thus, we believe the company would benefit by **storing bikes when they are not used as frequently**. The specific percentage to be stored in each season would be calculated arithmetically, by finding the number of bikes that are not used in each season. The maximum number of bikes used in an hour in a season is multiplied by 1.05 to provide a buffer value to **ensure that the supply of bikes would not fall below the demand of bikes** at any point in time. This value is then divided against the maximum number of bikes ever rented in an hour to get the percentage of bikes that is able to be stored.

$$\%\ \boldsymbol{of\ bikes\ to\ store} = \frac{Max\ bikes\ used\ in\ an\ hour - Max\ bikes\ used\ in\ an\ hour(season) \times 1.05}{Max\ bikes\ used\ in\ an\ hour}$$

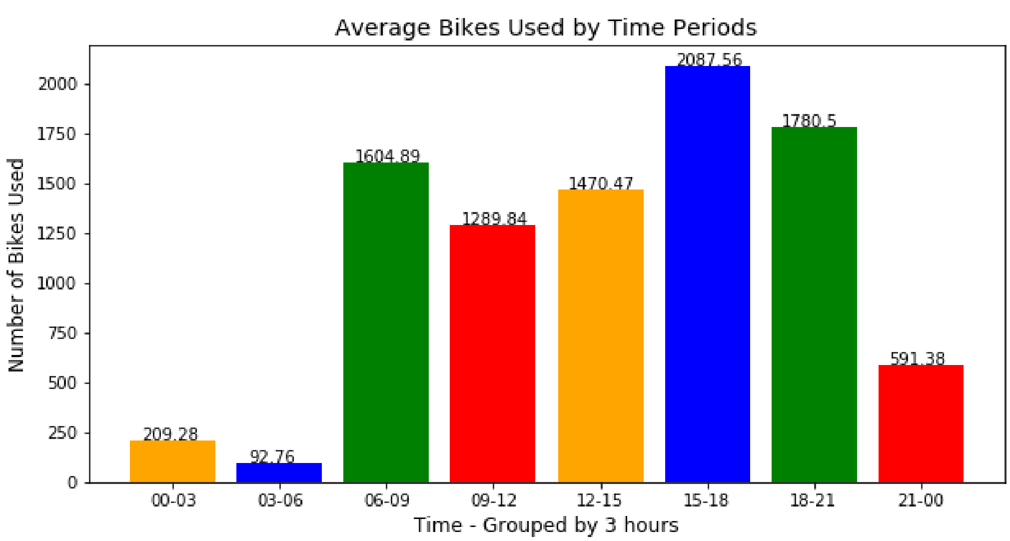
This is to prevent some of the bikes being damaged by the weather as they are being kept in the open. Therefore, we are able to conclude that in winter we would store some of the bikes since they are not being used and conduct longer more intensive maintenance session. Since consumption is the most in Summer, we will take that as our base and **not store any bikes** in summer. Similarly, we store 27.57% of the bikes in Fall, 28.90% of bikes in Spring and 41.02% of bikes in Winter.

2. Bikes Used in Weekends



From the graph, it can be observed that during the weekends, there are **lesser bikes rented**. A reason for this could be that people are only renting the bikes to get to and fro from work, thus are unlikely to have a need to use the bikes on weekends. As such, we are likely to choose to **remove the bikes on weekends** to send them for repairs as the demand for bikes are lower than that of demand for bikes on weekdays.

3, Average Bikes Used by Time Periods



This graph plots the number of bikes used against time which is **grouped into three-hour intervals**. 1 represents time from 0000 to 0300, 2 represents time from 0300 to 0600 and so on. It would not be meaningful to view all the timings at hourly intervals to decide which time should the bike be taken for checks and maintenance as such checks would definitely take more than an hour. By grouping the timings into three-hour intervals, we would be able to analyze when would be the best time range to collect the bikes.

Based on the graph itself, it suggests that either time period 1,2 or 8 would be suitable to collect the bikes for checks and repair as the number of usages is relatively lower compared to the rest. However, we would **choose time periods 1 and 8** as there will be a buffer time for us in cases of emergency or error during the maintenance process. We will just need to ensure that bikes are returned before period 3 to not cause any disruption in the supply of bikes. Nevertheless, we will not be able to conclude the exact number of bikes to be collected respectively hence we will draw conclusions from the regression analysis.

Before the calculations, we assume that the bikes **would only require maintenance every 2 weeks** hence only half of the total number of bikes will be sent to repair every week. The calculations process are as follows:

STEP 1:

-Find the % of unused in period 1 & 8 (lowest two period) and use period 6 as denominator (most period)

$$\% \; of \; Unused \; Bikes \; Period_{x \; or \; y} = 1 - \left(\frac{No. of \; Bikes \; used \; in \; Period_{x \; or \; y}}{No. of \; Bikes \; used \; in \; the \; highest \; Period} * 100\% \right)$$

x = The period with the lowest number of bikes used
y = The period with the second lowest number of bikes used

STEP 2:

-Calculate the % of bikes send for each period

$$\% \; of \; Bikes \; send \; for \; Period_{x \; or \; y} = \frac{\% \; of \; Unused \; Bikes \; Period_{x \; or \; y}}{\% \; of \; Unused \; Bike \; Period_{x} + \% \; of \; Unused \; Bike \; Period_{y}}$$

STEP 3:

-Calculate the actual number of bikes send for each period

$$No. of \; Bikes \; send \; for \; Period_{x \; or \; y} = \frac{\% \; of \; Bikes \; send \; for \; Period_{x \; or \; y} * \; Total \; number \; of \; Bikes}{2}$$

* Computing with Data from OLS REGRESSION RESULT (Appendix 4),

STEP 1:

$$\%\ of\ \boldsymbol{Unused}\ Bikes\ Period_1 = 1 - \frac{1412.9960}{1412.9960+1494.1755} = 51.4\ \%\ (3.s.f)$$

$$\%\ of\ \boldsymbol{Unused}\ Bikes\ Period_8 = 1 - \frac{1412.9960+284.4080}{1412.9960+1494.1755} = 41.6\%\ (3.s.f)$$

STEP 2:

$$\%\ of\ Bikes\ send\ for\ Period_1 = \frac{51.4\%}{51.4\%+41.6\%} = 55.3\%\ (3.s.f)$$

$$\%\ of\ Bikes\ send\ for\ Period_8 = \frac{41.6\%}{51.4\%+41.6\%} = 44.7\%\ (3.s.f)$$

STEP 3:

$$No.of\ Bikes\ send\ for\ Period_1 = \frac{55.3\% * Total\ no.of\ bikes}{2} = 27.65\%\ of\ Total\ no.\ of\ bikes$$

$$No.of\ Bikes\ send\ for\ Period_8 = \frac{44.7\% * Total\ no.of\ bikes}{2} = 22.35\%\ of\ Total\ no.\ of\ bikes$$

Conclusion & Limitations

In summary, we have decided that repairs would be **conducted during weekends** where there is lesser usage of bikes. However, we have to look deeper to determine which time period of the weekends would be ideal for us to schedule the maintenance. Through the regression analysis and data visualization aids, we determine that the whole process would be spread across 2 time periods which are time period 1 (0000 to 0300) and time period 8 (2100 to 0000). These two time periods would be the optimal timing where the bikes would be collected to be sent for maintenance and returned before period 3 where the demand would spike again.

Furthermore, through our data analysis, we recognize that bike sharing is the **least popular during winter**. With this knowledge, we would store some of our bikes using the formula above to reduce unnecessary wear and tear due to weather conditions. However, we have to keep in mind that the number of bikes to store is dynamic and will need to be recalculated annually.

However, the study does possess its limitations. Firstly, the accuracy of the findings is **highly dependent on the data sources**. For instance, the number of bikes, temperature, humidity etc. Secondly, the suggestions that we are making are **dynamic** which means that it will change according to the data. The numbers of users would change year after year; hence it would be best to review the schedule every year to ensure its effectiveness. Lastly, there is **a lack of data** which leads us to make a few assumptions. For example, we are unaware of the location points where the bikes will be parked at, hence we will assume that the whole maintenance process would only take 3 hours regardless of their location. Furthermore, when calculating the number of bikes that is required to send for repair at each period, we could not conclude the exact number as **we do not know the total number of bikes.**

APPENDIX 1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.446
Model:                            OLS   Adj. R-squared:                  0.445
Method:                 Least Squares   F-statistic:                     776.4
Date:                Sun, 17 Nov 2019   Prob (F-statistic):               0.00
Time:                        23:37:24   Log-Likelihood:            -1.4129e+05
No. Observations:               17414   AIC:                         2.826e+05
Df Residuals:                   17395   BIC:                         2.828e+05
Df Model:                          18
Covariance Type:            nonrobust
=======================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
Intercept                            1128.9100     77.299     14.604      0.000     977.396    1280.424
C(weather_code)[T.Clear]              -13.3995     18.875     -0.710      0.478     -50.396      23.597
C(weather_code)[T.Cloudy]             -94.6865     25.702     -3.684      0.000    -145.065     -44.308
C(weather_code)[T.Rain]              -232.8299     22.578    -10.312      0.000    -277.085    -188.575
C(weather_code)[T.Scattered_Clouds]    46.6941     19.280      2.422      0.015       8.903      84.485
C(weather_code)[T.Snow]                24.7712    105.900      0.234      0.815    -182.803     232.345
C(weather_code)[T.Thunderstorm]      -963.0748    216.706     -4.444      0.000   -1387.839    -538.310
C(season)[T.Spring]                   -60.5385     18.908     -3.202      0.001     -97.600     -23.477
C(season)[T.Summer]                   -81.2868     19.960     -4.072      0.000    -120.411     -42.162
C(season)[T.Winter]                   -72.5607     20.217     -3.589      0.000    -112.189     -32.933
C(group1)[T.2]                       1153.0792     18.088     63.750      0.000    1117.626    1188.533
C(group1)[T.3]                       1212.0896     20.944     57.873      0.000    1171.038    1253.142
C(group1)[T.4]                        824.7013     18.146     45.448      0.000     789.134     860.269
t1                                     77.5997      8.523      9.105      0.000      60.894      94.306
t2                                    -30.3683      6.847     -4.435      0.000     -43.789     -16.947
hum                                   -15.1283      0.654    -23.149      0.000     -16.409     -13.847
wind_speed                             -9.6820      0.924    -10.477      0.000     -11.493      -7.871
is_holiday                           -300.8112     42.296     -7.112      0.000    -383.716    -217.906
is_weekend                           -219.1489     13.694    -16.004      0.000    -245.990    -192.308
==============================================================================
Omnibus:                     4939.006   Durbin-Watson:                   0.847
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13344.897
Skew:                           1.524   Prob(JB):                         0.00
Kurtosis:                       6.017   Cond. No.                     2.72e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.72e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

APPENDIX 2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.439
Model:                            OLS   Adj. R-squared:                  0.439
Method:                 Least Squares   F-statistic:                     1240.
Date:                Sun, 17 Nov 2019   Prob (F-statistic):               0.00
Time:                        23:41:03   Log-Likelihood:            -1.4138e+05
No. Observations:               17414   AIC:                         2.828e+05
Df Residuals:                   17402   BIC:                         2.829e+05
Df Model:                          11
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept              1471.3776     60.148     24.462      0.000    1353.481    1589.275
C(season)[T.Spring]     -86.0152     18.900     -4.551      0.000    -123.062     -48.969
C(season)[T.Summer]     -74.6545     19.793     -3.772      0.000    -113.451     -35.858
C(season)[T.Winter]     -77.7594     20.224     -3.845      0.000    -117.400     -38.119
C(group1)[T.2]         1170.3368     17.754     65.921      0.000    1135.538    1205.136
C(group1)[T.3]         1199.2931     20.270     59.165      0.000    1159.561    1239.025
C(group1)[T.4]          820.4269     18.157     45.186      0.000     784.838     856.016
t1                       38.8444      1.789     21.708      0.000      35.337      42.352
hum                     -18.3615      0.558    -32.880      0.000     -19.456     -17.267
wind_speed               -9.6562      0.843    -11.452      0.000     -11.309      -8.004
is_holiday             -304.8534     42.504     -7.172      0.000    -388.166    -221.541
is_weekend             -221.8805     13.714    -16.179      0.000    -248.761    -195.000
==============================================================================
Omnibus:                     4928.312   Durbin-Watson:                   0.835
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13286.356
Skew:                           1.522   Prob(JB):                         0.00
Kurtosis:                       6.009   Cond. No.                         768.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

APPENDIX 3

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.539
Model:                            OLS   Adj. R-squared:                  0.539
Method:                 Least Squares   F-statistic:                     1568.
Date:                Sun, 17 Nov 2019   Prob (F-statistic):               0.00
Time:                        23:41:34   Log-Likelihood:            -1.3967e+05
No. Observations:               17414   AIC:                         2.794e+05
Df Residuals:                   17400   BIC:                         2.795e+05
Df Model:                          13
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept            1206.8817     55.206     21.861      0.000    1098.671    1315.092
C(season)[T.Spring]   -65.1598     17.136     -3.803      0.000     -98.748     -31.572
C(season)[T.Summer]     2.0655     18.016      0.115      0.909     -33.247      37.378
C(season)[T.Winter]  -115.3945     18.347     -6.289      0.000    -151.357     -79.432
C(group2)[T.2]        386.1991     19.394     19.913      0.000     348.185     424.213
C(group2)[T.3]       1376.7622     19.831     69.424      0.000    1337.891    1415.633
C(group2)[T.4]       1007.1765     21.507     46.829      0.000     965.020    1049.333
C(group2)[T.5]       1798.0414     20.907     86.001      0.000    1757.061    1839.021
C(group2)[T.6]        422.6369     19.600     21.563      0.000     384.220     461.054
t1                     32.4398      1.628     19.923      0.000      29.248      35.631
hum                   -14.0124      0.514    -27.276      0.000     -15.019     -13.005
wind_speed            -10.8022      0.765    -14.115      0.000     -12.302      -9.302
is_holiday           -320.5025     38.526     -8.319      0.000    -396.017    -244.988
is_weekend           -226.3873     12.430    -18.213      0.000    -250.752    -202.023
==============================================================================
Omnibus:                     4790.641   Durbin-Watson:                   0.886
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13524.822
Skew:                           1.455   Prob(JB):                         0.00
Kurtosis:                       6.190   Cond. No.                         793.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

APPENDIX 4

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.531
Model:                            OLS   Adj. R-squared:                  0.530
Method:                 Least Squares   F-statistic:                     1311.
Date:                Sun, 17 Nov 2019   Prob (F-statistic):               0.00
Time:                        23:41:55   Log-Likelihood:            -1.3984e+05
No. Observations:               17414   AIC:                         2.797e+05
Df Residuals:                   17398   BIC:                         2.798e+05
Df Model:                          15
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept            1412.9960     56.320     25.089      0.000    1302.602    1523.390
C(season)[T.Spring]   -79.0536     17.305     -4.568      0.000    -112.974     -45.134
C(season)[T.Summer]   -60.2380     18.218     -3.307      0.001     -95.947     -24.529
C(season)[T.Winter]   -83.5886     18.532     -4.510      0.000    -119.914     -47.263
C(group3)[T.03-06]    -58.9550     22.643     -2.604      0.009    -103.337     -14.573
C(group3)[T.06-09]   1413.4061     22.574     62.613      0.000    1369.159    1457.653
C(group3)[T.09-12]    882.0714     23.141     38.117      0.000     836.713     927.430
C(group3)[T.12-15]    889.4202     24.339     36.543      0.000     841.713     937.128
C(group3)[T.15-18]   1494.1755     24.371     61.309      0.000    1446.405    1541.946
C(group3)[T.18-21]   1317.2565     23.347     56.420      0.000    1271.494    1363.019
C(group3)[T.21-00]    284.4080     22.689     12.535      0.000     239.935     328.881
t1                     38.1024      1.646     23.146      0.000      34.876      41.329
hum                   -17.2436      0.522    -33.060      0.000     -18.266     -16.221
wind_speed             -9.4767      0.773    -12.255      0.000     -10.992      -7.961
is_holiday           -308.7921     38.902     -7.938      0.000    -385.044    -232.540
is_weekend           -223.3855     12.552    -17.797      0.000    -247.988    -198.783
==============================================================================
Omnibus:                     3770.991   Durbin-Watson:                   0.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9590.146
Skew:                           1.184   Prob(JB):                         0.00
Kurtosis:                       5.759   Cond. No.                         836.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```