# Time Series Analysis of UBER Travel Time in Los Angeles

Yuhe Tian 305243288
Hongyi Wang 405242855
Xinyuan Yang 305255240
Kemo Zhou 405060826

## Introduction

Uber trips occur all over cities. Analyzing Uber travelling data can help learn the traffic flows in a specific city as well as its economy level. Further, estimating travelling time is also meaningful in practice in an individual level by contributing to optimize schedule.

The aims of our study are to identify a model best fitting the daily travel time data of Uber ride and find out the specific patterns of the travel time across different times of day or days of the week. Our project will focus on zone-to-zone travel times based on daily mean travel time of rides to investigate the travel pattern and users' travel habits as well.

## Data

The dataset being examined in this project consist Uber date and a set of dummy variables. We captured the Daily Average Travel Time (from downtown to LAX) from the Uber Movement platform. The original data consists the mean, lowest and highest time need to travel from downtown to LAX in different time of day (morning, midday, evening, early morning). We selected the data in the morning (early peak time) as our main subject to observe. There are 365 observations from 04/01/2017 to 03/31/2018 in Travel Time variable. We defined it as a one-year-long daily time series variable.

We also want to include the possible influential variables into our dataset. It is reasonable to suppose the weather situation will affect the heaviness of the traffic. Thus, we captured the precipitation as Los Angeles to describe the weather situation. A series dummy variables (Rainfall, Holiday and 7-day in week) are also included in the dataset.

There are 13 missing values (out of 365 obs) in the Travel Time variable. We fill in value for those variables by the average value of one-day-before and one-day-after. There are also several missing values in the Precipitation variable (27 out of 365). We fill in zeros for those NAs.

*Table 1.* Summary of Variables

| Names | Category | Value |
|---|---|---|
| Travel Time | Time series | Mean travel time in seconds |
| Precipitation | Times series | Volume Precipitation in millimeter |
| Rainfall | Dummy | Rainfall = 1 if precipitation >= 0.01<br>Rainfall = 0 if otherwise |

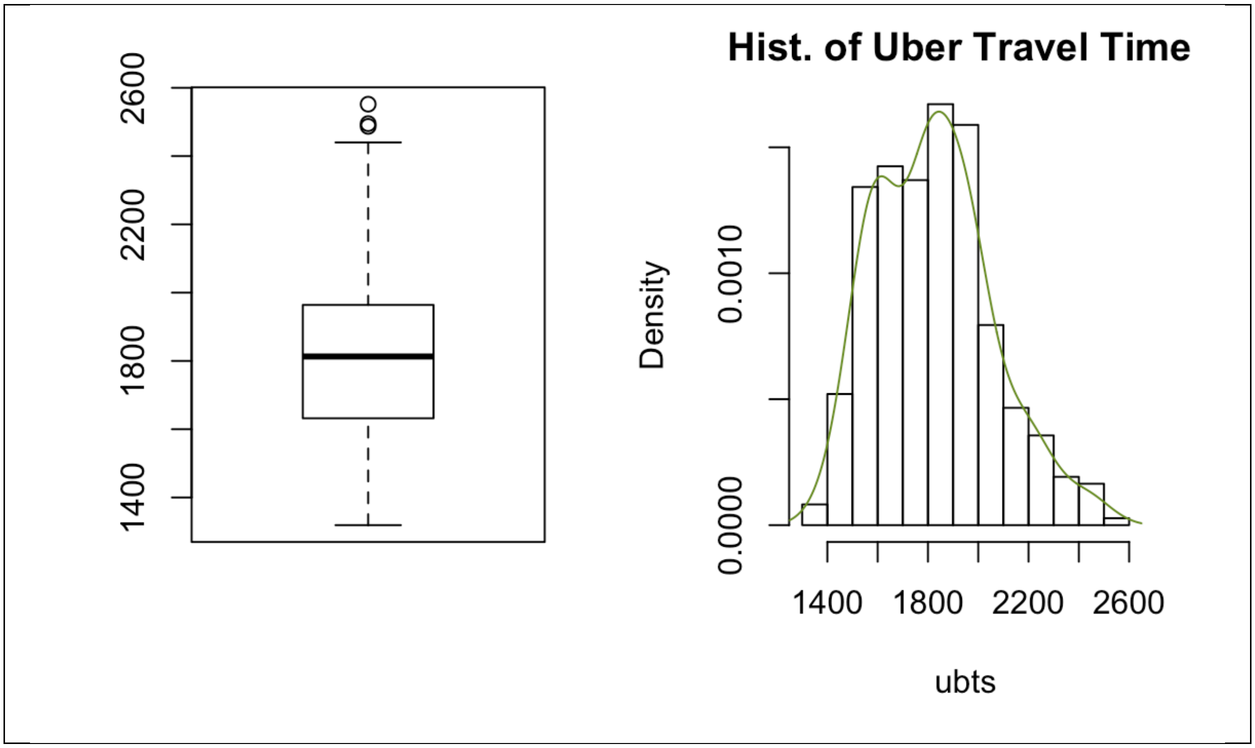
| | | |
|---|---|---|
| Holiday | Dummy | Holiday = 1 if the day is a holiday<br>Holiday = 0 if otherwise |
| Mon | Dummy | Mon = 1 if the day is a Monday<br>Mon = 0 if otherwise |
| Tue | Dummy | Tue = 1 if the day is a Monday<br>Tue = 0 if otherwise |
| Wed | Dummy | Wed = 1 if the day is a Monday<br>Wed = 0 if otherwise |
| Thur | Dummy | Thur = 1 if the day is a Monday<br>Thur = 0 if otherwise |
| Fri | Dummy | Fri = 1 if the day is a Monday<br>Fri = 0 if otherwise |
| Sat | Dummy | Sat = 1 if the day is a Monday<br>Sat = 0 if otherwise |
| Sun | Dummy | Sun = 1 if the day is a Monday<br>Sun = 0 if otherwise |

## Methodology

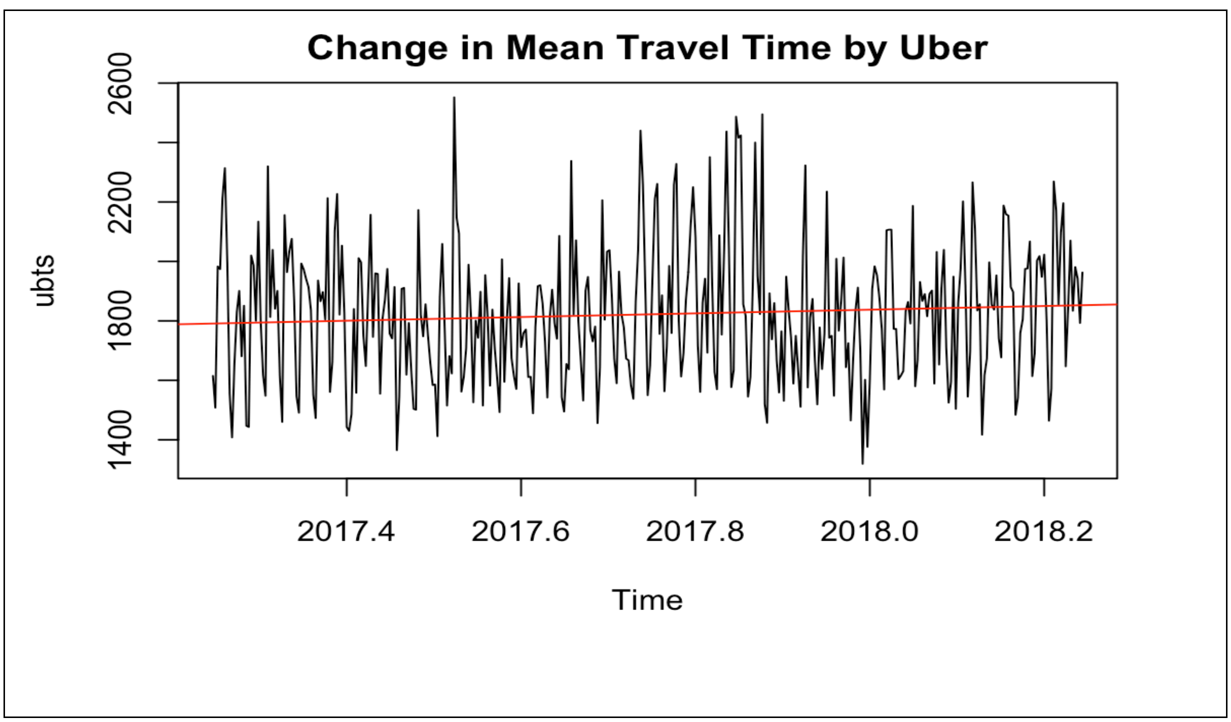
By analyzing daily mean travel time over time, we can reliably estimate how long it takes to get from one area to another like during Monday rush hour. To make the estimation, we applied ARIMA and other forecasting model to analysis this time series.
Intuitively, we expected to find a weekly seasonality since people's preference on traveling may varies in weekdays and weekends. We also expected to find increasing trend based on increasing traffic problem in LA. There may exist co-movement pattern in traveling data and weather data. A researching involving LA's weather data is considered but still has to be decided due to the data availability.

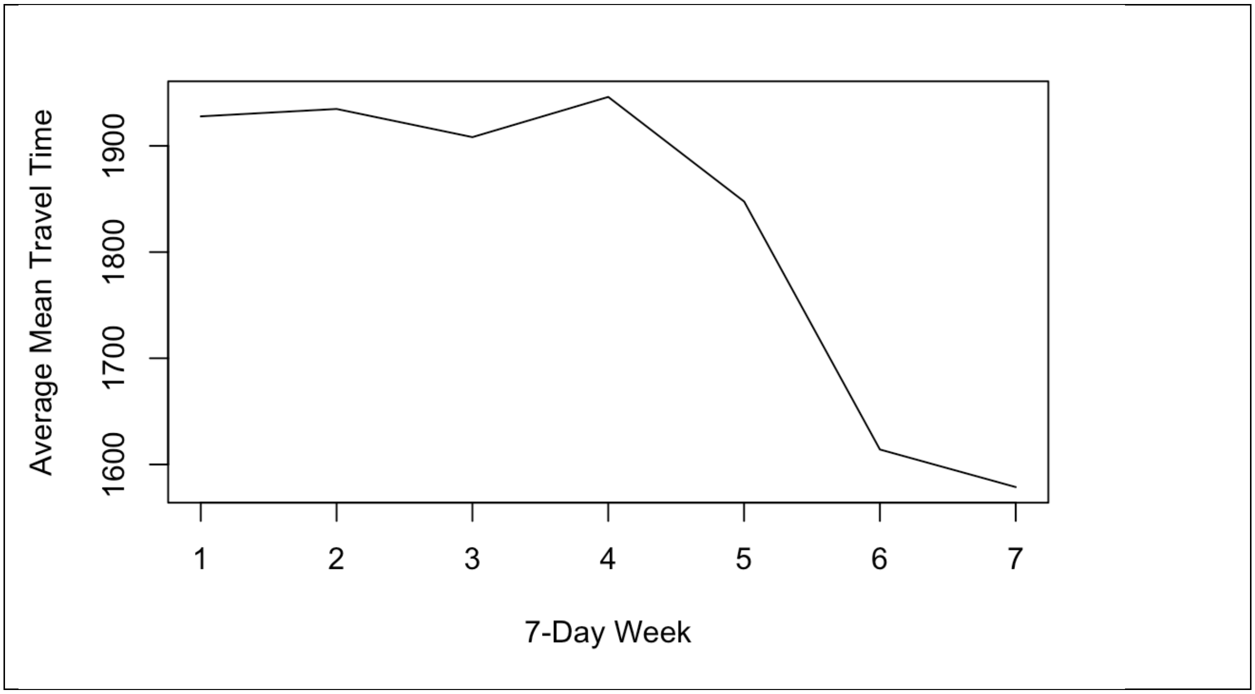
We will also try to fit a VAR model to the data with another time series of weather condition (also with daily frequency) to delve into how travel times are impacted by bad weather condition. The data we used is mean travel time from 2016 to 2018, which have been documented daily by "Uber Movements".

Exploratory Data Analysis

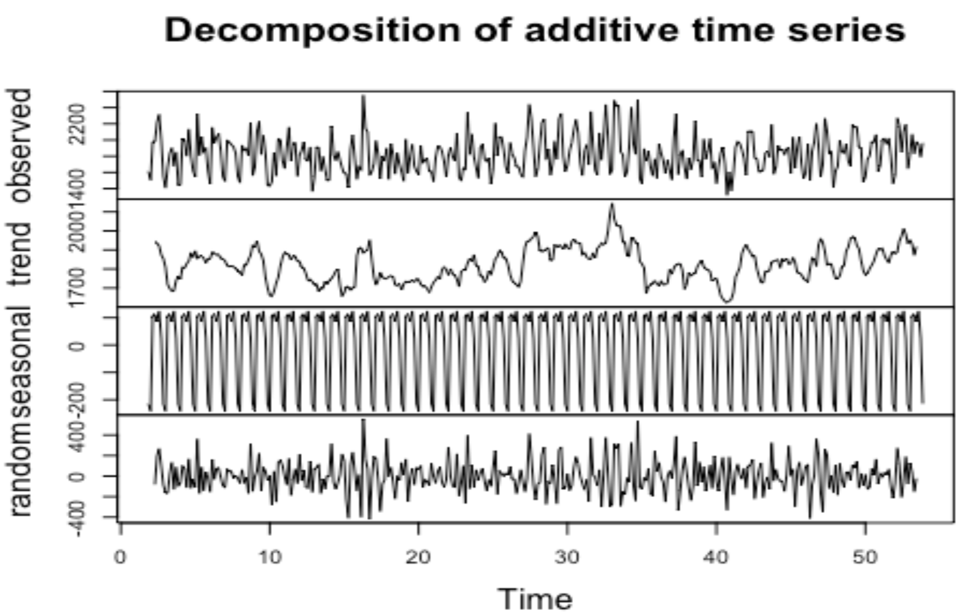



The plot shows：
1 1600~2200
2 right skew





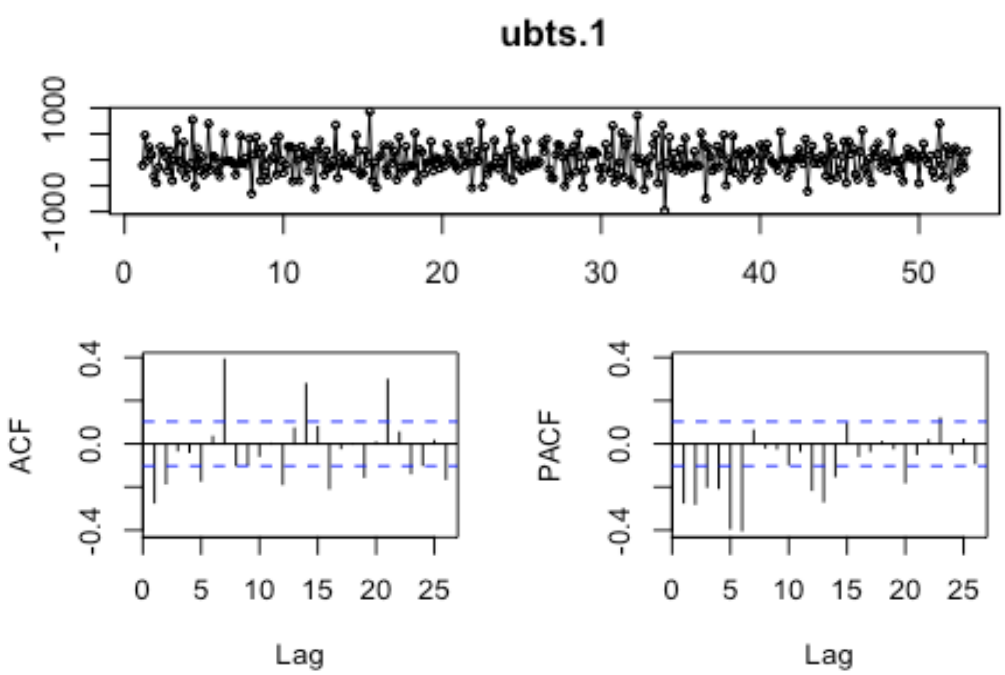

Travel Time Forecasting：ARIMA Model
We start with defining the time series with frequency of seven and decompose the time series to look at the trend, seasonality, etc. From the time series movement and decomposition plots, we can hardly find a trend for the Uber travel time. But a significant seasonal pattern does exists.



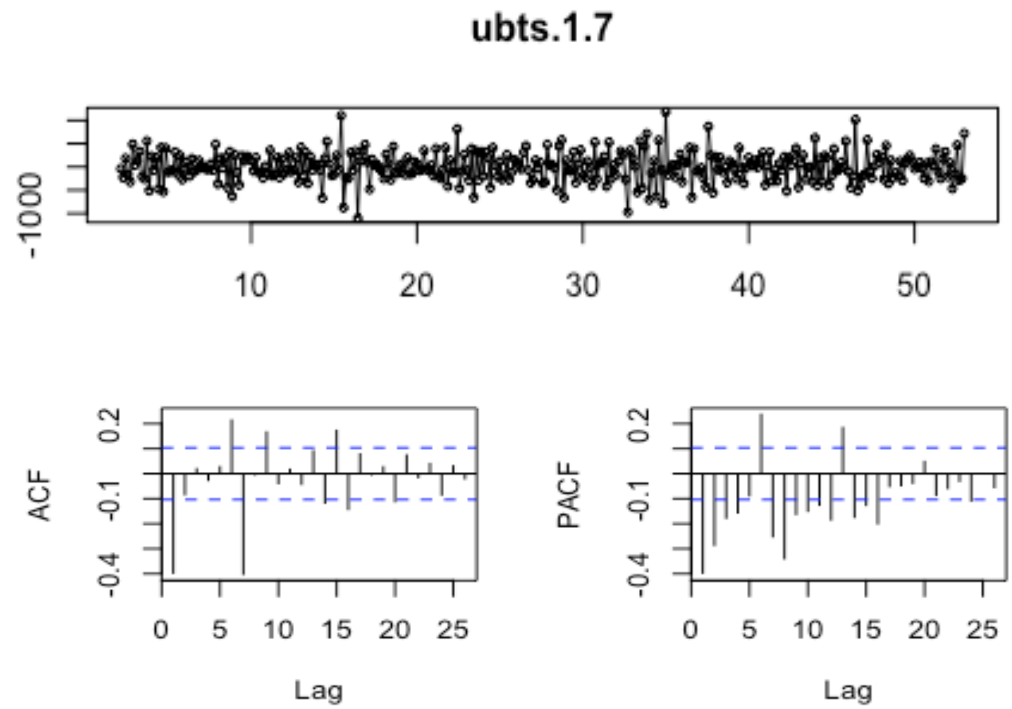

Before processing with the model, we checked the stationarity of the time series with unit root test. Which indicates that it's a non-stationary one, so we take the first difference to transform it into a stationary one,

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-774.48 -165.51    3.64  179.40  926.72

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
z.lag.1    -0.008176   0.007905  -1.034    0.302
z.diff.lag -0.268038   0.050746  -5.282 2.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 275.8 on 361 degrees of freedom
Multiple R-squared:  0.07679,	Adjusted R-squared:  0.07168
F-statistic: 15.01 on 2 and 361 DF,  p-value: 5.453e-07


Value of test-statistic is: -1.0344

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

So we took the first difference to transform it into a stationary one and this time passed the stationary test. ACF, PACF plot after stationary transformation is as follows.

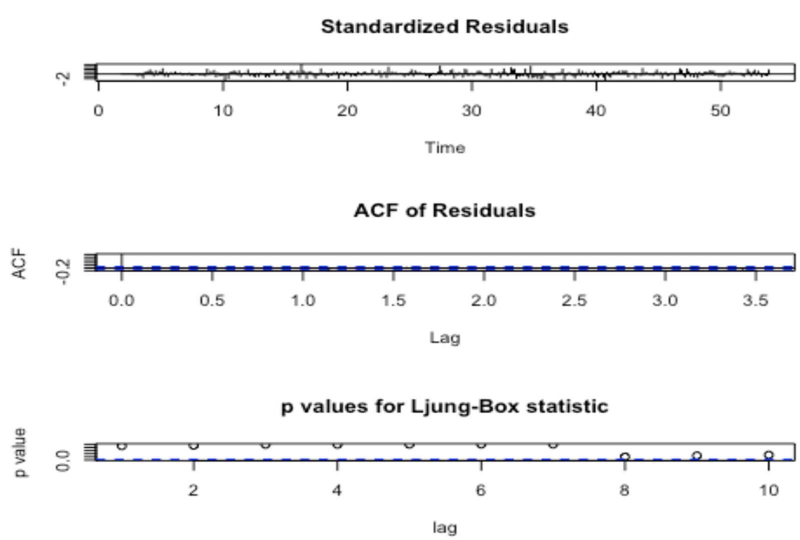



We can see regular spikes at 7, 14 and 21in ACF and PACF plots. PACF decays at the corresponding position. So we consider pattern of weekly seasonality and took 7 steps difference to check seasonality.




After taking seven step difference, the ACF and PACF plots do not show that strong seasonality se we moved on to fitting the model. It is hard to define the optimal parameters (p, d, q) only given the ACF PACF plots, after several trials with the auto ARIMA function, the finalized model is ARIMA(1,1,1)(1,1,1)

Next, to verify the efficiency of the model, we plot the Diagnostic Plots for Time-Series Fits. ACF plot of residuals shows there’s no spike afterwards which means no autocorrelation in the errors.

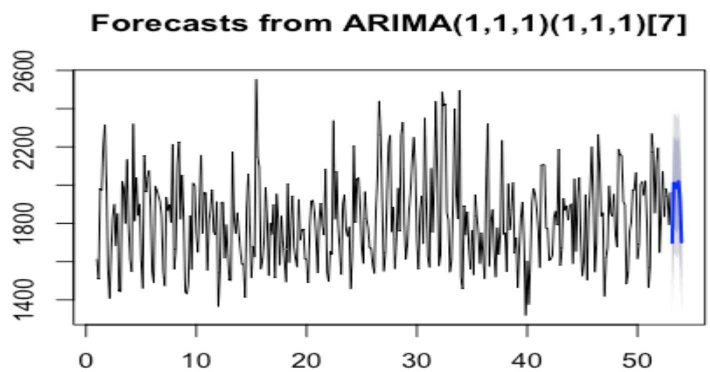



Ljung-Box test result with P-value>0.5 Further verify that the residuals are generally white noise.

```
Box.test(fit.b.1$residuals, type = "Ljung-Box")
##
##  Box-Ljung test
##
## data:  fit.b.1$residuals
## X-squared = 0.021971, df = 1, p-value = 0.8822
```

So we suppose the series should be stationary and modelled correctly. And the plot is showed below.

According to the model.AR parts indicates travel time is auto correlated with last period.MA parts indicate travel time is affected by stochastic disturbance from last period.




Adjustment for Heteroscedasticity：ARCH & GRACH Model

1. GARCH model

Although ACF & PACF of residuals have no significant lags, the time series plot of residuals shows some cluster of volatility. As we know, ARIMA provides best linear forecast for the series and thus plays little role in forecasting model nonlinearly. In order to model volatility, we use GARCH to reflect recent changes and fluctuations in the series to test whether we could make the predictions better.

We tried number from 0 to 8 and compute the AICs by using likelihood function. From the table we could see that GARCH(1,3) has the lowest AICs while correct with most parameters significant. From the Ljing-box test, we could conclude that such model perfectly reflects the residuals.

```
Coefficient(s):
         mu        omega       alpha1        beta1        beta2        beta3
-0.00376519   0.01127952   0.19055243   0.38822250   0.00000001   0.37653535

Std. Errors:
 based on Hessian

Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)
mu     -3.765e-03   8.482e-03   -0.444 0.657123
omega   1.128e-02   2.826e-03    3.991 6.58e-05 ***
alpha1  1.906e-01   2.791e-02    6.828 8.63e-12 ***
beta1   3.882e-01   1.038e-01    3.741 0.000184 ***
beta2   1.000e-08   1.723e-01    0.000 1.000000
beta3   3.765e-01   1.279e-01    2.944 0.003240 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, from output for GARCH(1,3), we could see that full GARCH model could be written as

$$h_t = -0.003 + 0.19055\varepsilon_{t-1} + 0.011279 + 0.38822\sigma_{t-1|t-2} + 0.000001\sigma_{t-2|t-3}$$

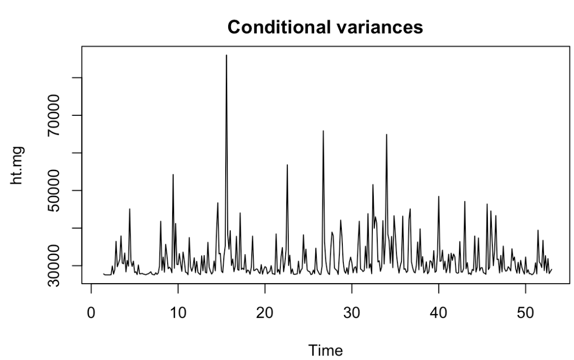
2. ARIMA-GARCH model performance

The best ARIMA model for Uber data as the above displayed, is ARIMA(1,1,1). Thus, we will compare the results from our original ARIMA model and the combined ARIMA-GARCH model.

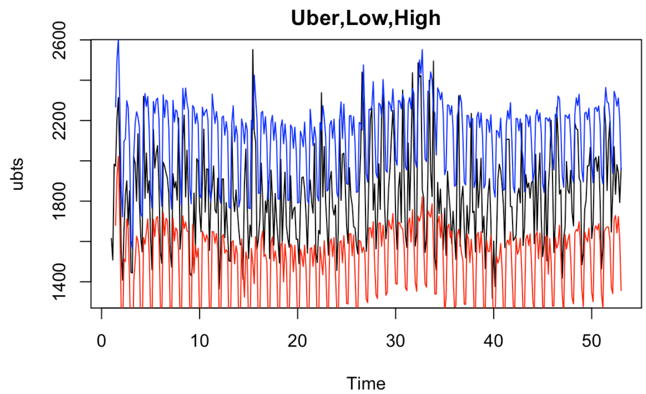
Including the ARIMA model, the mixed model could be written as:

$$Y_t - Y_{t-1} = 0.1188(Y_{t-1} - Y_{t-2}) - 0.9465\varepsilon_{t-1} + \varepsilon_t - 0.003 + 0.19055\varepsilon_{t-1} + 0.011279 + 0.38822\sigma_{t-1|t-2} + 0.000001\sigma_{t-2|t-3}$$

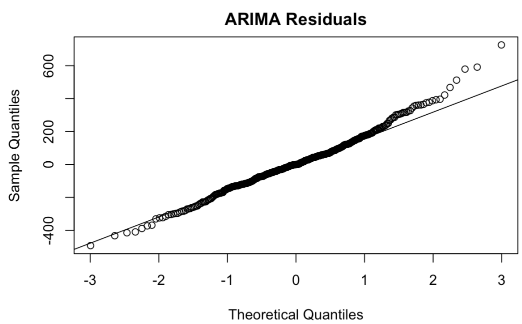
We use the ARIMA forecast obtained from R, and then add GARCH to ARIMA forecast. The condition variances are plotted and it successfully reflects the volatility of the time series over the entire period since high volatility is closely related to period where travel times tumbled.

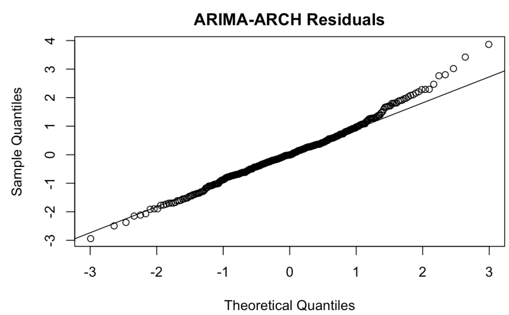



The 95% forecast interval of travel times is shown below.




The final check on the model is to look at Q-Q Plot of residuals of ARIMA-GARCH model.




The plot shows that residuals seem to be roughly normally distributed although some points remain off the line. However, compared to residuals of ARIMA model, those of mixed model are more normally distributed.

3. Conclusion

Time domain method is a useful way to analyze the financial time series. There are some points in forecasting based on ARIMA- GARCH model that need to take into account. ARIMA model focuses on analyzing time series linearly and it does not reflect recent changes as new information is available. Therefore, as a method to measure volatility of the series, GARCH incorporates new information and analyzes the series based on conditional variances where users can forecast future values with up-to-date information. Generally speaking, the forecast interval for the mixed model is closer than that of ARIMA-only model.
However, I want to mention that in our case, the original does not suffer much problem from residuals so that the improvement we bring might not outweigh the new error it bring.

<u>Causality Exploring: Analysis with Dummy Sets and VAR Model</u>

In linear regression models, we introduce six seasonal dummies into the seasonal dummy model, aiming to analyze the relationship between mean travel time, weekdays, weekends, holidays and rainfall.

1. Pure season dummy

We assume that there is a seasonal pattern in a week, and we introduce six seasonal dummies into the seasonal dummy model. From the p value of the whole model, we can see that the model is significant. Additionally, the dummy Monday, Sunday and Saturday have significant influence on the mean travel time. We guess this is because Monday and weekends have more traffic than other days.

```
## Call:
## lm(formula = ubts ~ Mon + Sun + Sat + Fri + Thur + Wed)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -528.64 -110.76  -10.08   84.31  647.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1927.779     25.611  75.271   <2e-16 ***
## Mon         -349.087     36.220  -9.638   <2e-16 ***
## Sun         -313.694     36.048  -8.702   <2e-16 ***
## Sat          -80.139     36.220  -2.213   0.0276 *
## Fri           18.303     36.220   0.505   0.6136
## Thur         -19.558     36.220  -0.540   0.5896
## Wed            6.981     36.220   0.193   0.8473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184.7 on 358 degrees of freedom
## Multiple R-squared:  0.3908, Adjusted R-squared:  0.3806
## F-statistic: 38.28 on 6 and 358 DF,  p-value: < 2.2e-16
```

2. Add holiday and rainfall dummies

Secondly, we introduce the holiday dummy into the original seasonal dummy model. From the p value of the whole model, we can conclude this model is significant as well. From the p value of the holiday dummy, the holiday variable has significant influence on the mean travel time.

```
Call:
lm(formula = ubts ~ Mon + Tue + Sat + Sun + Thur + Wed + Fri +
    holiday)

Residuals:
    Min      1Q  Median      3Q     Max
-528.64 -111.24  -14.24   84.13  647.36

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1946.082     25.400  76.616  < 2e-16 ***
Mon         -358.321     36.085  -9.930  < 2e-16 ***
Tue           -6.212     36.212  -0.172  0.86390
Sat          -98.442     35.921  -2.740  0.00644 **
Sun         -329.031     35.769  -9.199  < 2e-16 ***
Thur         -34.838     35.940  -0.969  0.33303
Wed           -8.299     35.940  -0.231  0.81751
Fri               NA         NA      NA       NA
holiday     -157.184     59.546  -2.640  0.00866 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 183.2 on 357 degrees of freedom
Multiple R-squared:  0.4025,	Adjusted R-squared:  0.3908
F-statistic: 34.35 on 7 and 357 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = ubts ~ Mon + Tue + Sat + Sun + Thur + Wed + Fri +
    rainfall)

Residuals:
    Min      1Q  Median      3Q     Max
-526.86 -109.57   -9.75   86.09  649.14

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1943.110     25.869  75.112  < 2e-16 ***
Mon         -366.201     36.263 -10.098  < 2e-16 ***
Tue          -18.897     36.242  -0.521  0.60240
Sat          -97.254     36.263  -2.682  0.00766 **
Sun         -331.358     36.072  -9.186  < 2e-16 ***
Thur         -37.266     36.242  -1.028  0.30452
Wed           -9.539     36.298  -0.263  0.79285
Fri               NA         NA      NA       NA
rainfall      30.901     37.106   0.833  0.40552
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 184.8 on 357 degrees of freedom
Multiple R-squared:  0.392,	Adjusted R-squared:  0.3801
F-statistic: 32.88 on 7 and 357 DF,  p-value: < 2.2e-16
```

Thirdly, instead of holiday dummy, we introduce rainfall dummy into the seasonal dummy model. This model is significant as a whole, but the p value of rainfall is too large that rainfall doesn’t have influence on travel time in this model.

Through employing VAR model, here we want to analyze how well the precipitation can influence the travel time.

```
## Estimation results for equation prec:
## =====================================
## prec = prec.l1 + ubts.l1 + prec.l2 + ubts.l2 + prec.l3 + ubts.l3 + prec.l4 + ubts.l4 + const
##
##           Estimate Std. Error t value Pr(>|t|)
## prec.l1  0.5110361  0.0533216   9.584  < 2e-16 ***
## ubts.l1  0.0007117  0.0005850   1.217 0.224569
## prec.l2 -0.2050054  0.0592538  -3.460 0.000607 ***
## ubts.l2  0.0010439  0.0006039   1.729 0.084745 .
## prec.l3  0.1682997  0.0589933   2.853 0.004589 **
## ubts.l3 -0.0001164  0.0006050  -0.192 0.847495
## prec.l4  0.0116481  0.0528936   0.220 0.825829
## ubts.l4 -0.0001154  0.0005877  -0.196 0.844420
## const   -1.4166064  1.8763355  -0.755 0.450762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results shown above, we can see that there is no causality between travel time and precipitation. The reason behind needs further research. It may due to the data quality since we only have 27 dates defines as 'rainfall' in the year investigated.

The time series is autocorrelated and can be forecasted from historical data. A significant weekly seasonal pattern and holiday dummy are found. But there is no causality between travel time and precipitation.

Model Comparison

| **Model** | **AIC** |
|---|---|
| ARMA(1,6) | 4806.1 |
| ARIMA(1,1,1) | 4766.72 |
| ARIMA(1,1,1)+GARCH(1,3) | 4816.45 |
| ARIMA + Kalman Filter | 4983.61 |

As the AIC test indicated, ARIMA(1,1,1) model provides best fitness to Uber ride travel time and successfully reduced the residuals to white noise.

## Economic Insights Derived and Conclusions

### Q&A

Using of Differencing in ARIMA model

We can see regular spikes at 7, 14 and 21in ACF and PACF plots. PACF decays at the corresponding position. So we consider pattern of weekly seasonality and took 7 steps difference to check seasonality.

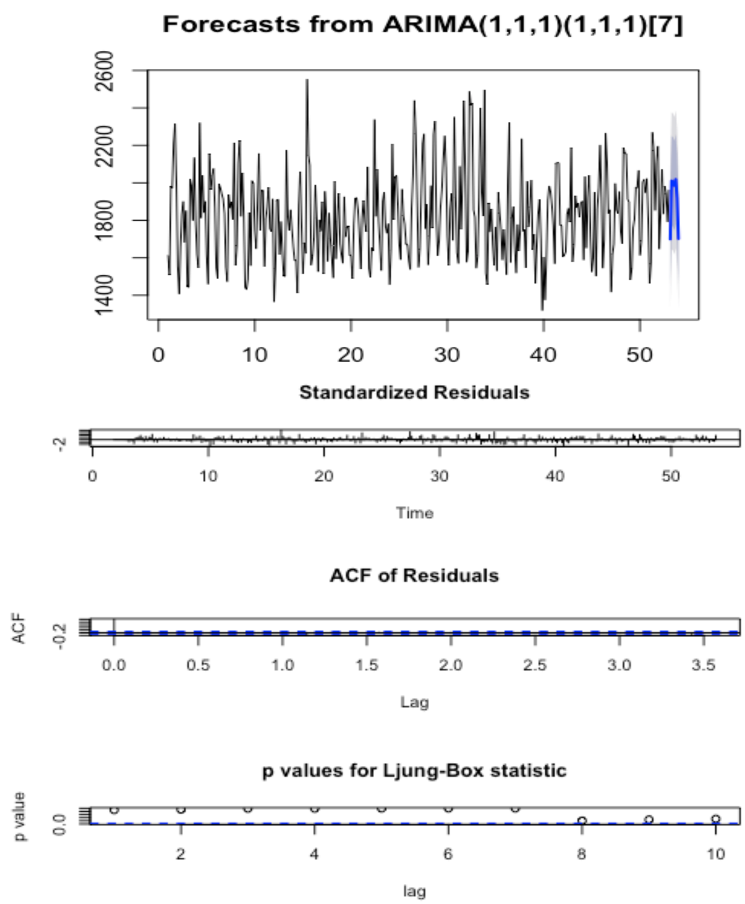
Reasoning behind the Relationship between Travel Time and Rainfall

The reason behind needs further research. It may due to the data quality since we only have 27 dates defines as 'rainfall' in the year investigated. We can consider expanding the data scope or select other city's data to  analyze.

<u>Traveling Time Daily Change Trend</u>
According to the simple EDA analysis, there are differences in time and time. The differences behind needs further research.

## Summary and Future Work

The time series Uber ride average travel time is autocorrelatedand and can be forecasted from historical data. A significant weekly seasonal pattern is also found. Differing from the intuition, there is no causality between travel time and precipitation. The reason behind needs further research. It may due to the data quality since we only have 27 days defined as "rainfall" out of 365 observations. We can consider expanding the data scope or select other city's data to analyze. According to the simple EDA analysis, there are differences in time and time. The differences behind needs further research.



**Reference:**

Uber Movement, (c) 2019 Uber Technologies, Inc., https://movement.uber.com
U.S. Climate Data, Your Weather Service, https://www.usclimatedata.com/

Appendix 1: Sample of Original Data

| Appendix 1A. Sample of Selected Uber Movement Data | | | | |
|---|---|---|---|---|
| Date | mean | lower | upper | holiady |
| 04/01/2017 | 1615 | 1354 | 1926 | 0 |
| 04/02/2017 | 1508 | 1275 | 1783 | 0 |
| 04/03/2017 | 1983 | 1507 | 2609 | 0 |
| 04/04/2017 | 1974 | 1418 | 2747 | 0 |
| 04/05/2017 | 2213 | 1562 | 3134 | 0 |
| 04/06/2017 | 2314 | 1869 | 2864 | 0 |
| 04/07/2017 | 1999 | 1568 | 2549 | 0 |
| 04/08/2017 | 1555 | 1252 | 1930 | 0 |
| 04/09/2017 | 1408 | 1148 | 1727 | 0 |
| 04/10/2017 | | | | 0 |
| 04/11/2017 | 1830 | 1455 | 2300 | 0 |
| 04/12/2017 | 1901 | 1357 | 2664 | 0 |
| 04/13/2017 | 1681 | 1294 | 2183 | 0 |
| 04/14/2017 | 1851 | 1457 | 2350 | 0 |
| 04/15/2017 | 1447 | 1206 | 1735 | 0 |
| 04/16/2017 | 1443 | 1175 | 1772 | 0 |
| 04/17/2017 | 2020 | 1593 | 2561 | 0 |
| 04/18/2017 | 1987 | 1496 | 2638 | 0 |
| 04/19/2017 | 1800 | 1505 | 2153 | 0 |
| 04/20/2017 | 2134 | 1595 | 2855 | 0 |
| 04/21/2017 | 1795 | 1277 | 2523 | 0 |
| 04/22/2017 | 1617 | 1327 | 1970 | 0 |
| 04/23/2017 | 1548 | 1224 | 1959 | 0 |
| 04/24/2017 | 2320 | 1688 | 3190 | 0 |
| 04/25/2017 | 1813 | 1400 | 2347 | 0 |
| 04/26/2017 | 2039 | 1497 | 2776 | 0 |
| 04/27/2017 | 1841 | 1425 | 2379 | 0 |
| 04/28/2017 | 1901 | 1486 | 2433 | 0 |
| 04/29/2017 | 1617 | 1215 | 2151 | 0 |

| Appendix 2A. Sample of Weather Data | | | |
|---|---|---|---|
| **Day** | **High** | **Low** | **Precip** |
| 4/1/17 | 69.1 | 51.1 | 0 |
| 4/2/17 | 73 | 55 | 0 |
| 4/3/17 | 64.9 | 54 | 0 |

| | | | |
|---|---|---|---|
| 4/4/17 | 68 | 55.9 | 0 |
| 4/5/17 | 81 | 55 | 0 |
| 4/6/17 | 73.9 | 55 | 0 |
| 4/7/17 | 66.9 | 54 | 0 |
| 4/8/17 | 64.9 | 54 | 0.2 |
| 4/9/17 | 68 | 51.1 | 0 |
| 4/10/17 | 75 | 54 | 0 |
| 4/11/17 | 73 | 52 | 0 |
| 4/12/17 | 69.1 | 53.1 | 0 |
| 4/13/17 | 66 | 53.1 | 0 |
| 4/14/17 | 66 | 55 | 0 |
| 4/15/17 | 70 | 52 | 0 |
| 4/16/17 | 70 | 54 | 0 |
| 4/17/17 | 71.1 | 53.1 | 0 |
| 4/18/17 | 73.9 | 57.9 | 0 |
| 4/19/17 | 68 | 57.9 | 0 |

Appendix 2: R code