TEXAS TECH UNIVERSITY SYSTEM™

Creating ACD Models for Boeing Duration Data

Data Analysis Project

Andras Horvath (Presenter)
Graduate Student in Statistics

STAT 6351/MATH 5099

December 5, 2023

Background

- For high frequency financial data a common goal is to fit a duration model in order to model the time between transactions.
- The most successful duration model has been the Autoregressive Conditional Duration (ACD) model from Engle and Russell (1998).
- General form of an ACD(r,s) model, as given by Tsay (2010) is:

$$x_i = \psi_i \epsilon_i, \qquad \psi_i = \omega_i + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j},$$

and $\{\epsilon_i\}$ is a sequence of iid nonnegative random variables such that $\mathrm{E}(\epsilon_i) = 1$.

Data

- The data is from Boeing stock trades for a 5-day period spanning from December 1st to 5th in 2008.
- The stock is traded on the NYSE which is open on weekdays from 9:30 to 16:00.
- There are 224,326 individual observations which represent the trades/transactions.
- The first 5 transactions are given below:

	year	month	day	hour	minute	second	price	volume	time
1	2008	12	1	9	30	2	41.64	3441	2008-12-01 09:30:02
2	2008	12	1	9	30	2	41.64	3441	2008-12-01 09:30:02
3	2008	12	1	9	30	12	41.43	100	2008-12-01 09:30:12
4	2008	12	1	9	30	16	41.82	100	2008-12-01 09:30:16
5	2008	12	1	9	30	16	41.83	100	2008-12-01 09:30:16

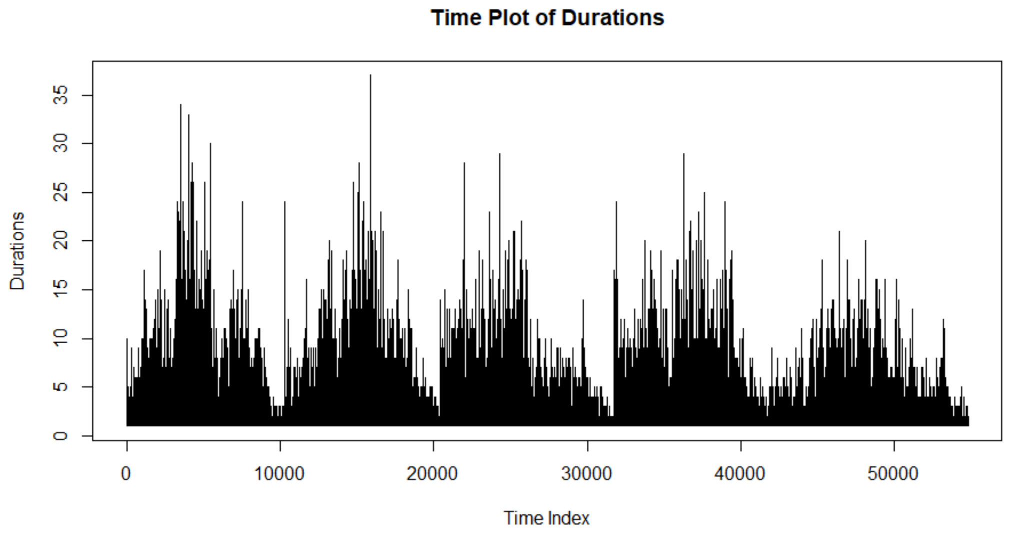
Data

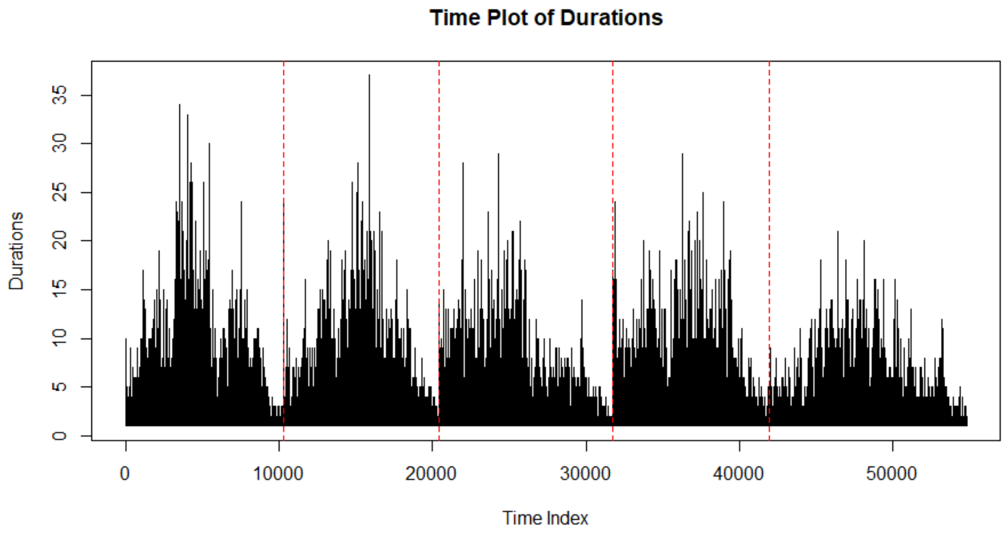
- The goal of this project is to use an ACD model to analyze the durations of the Boeing stock data.
- Using the ACDm package in R, we can extract the durations data. The first 5 are given below:

	time	price	volume	Ntrans	durations
1	2008-12-01 09:30:02	41.64000	6882	2	2
2	2008-12-01 09:30:12	41.43000	100	1	10
3	2008-12-01 09:30:16	41.91116	2016	11	4
4	2008-12-01 09:30:17	41.97000	1000	3	1
5	2008-12-01 09:30:27	41.50500	200	2	10

Exploring Diurnal Cycle

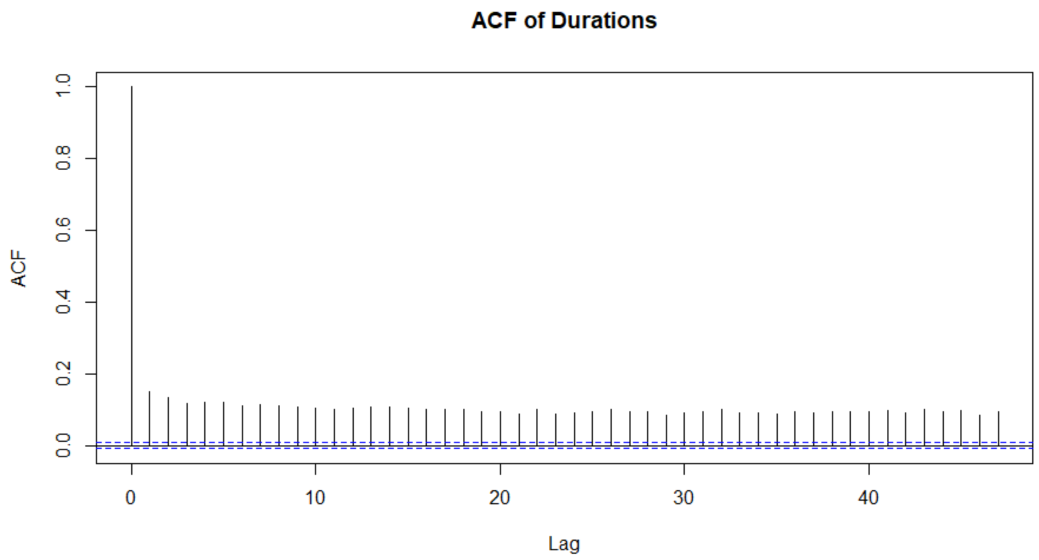
- Before we proceed with modeling, determine if a diurnal cycle exists. If so, this must be extracted out.
- Time plots of the positive durations for the five trading days and ACF plots have been used to determine this.





- The time plots both suggest a diurnal pattern is taking place.

Exploring Diurnal Cycle



- The ACF shows no signs of decaying, and all lags have bars exceeding the bands.
- Conclusion: Serial correlations do exist in the durations i.e. a diurnal cycle/systematic trend is present.

Removing Cycle

- Two options:

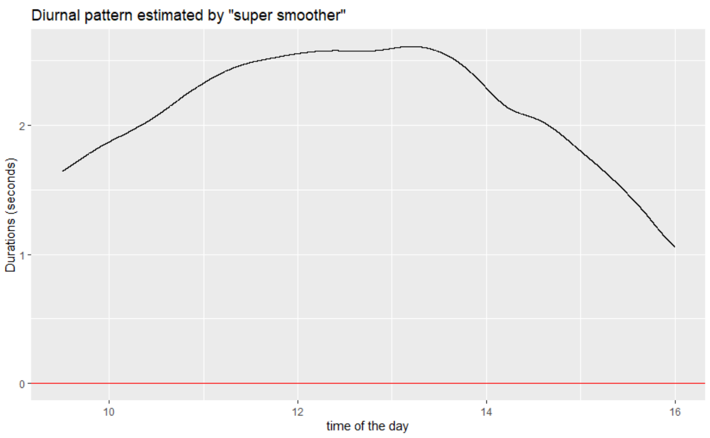
1. Construct trend functions using specific knowledge about the data.
2. Nonparametric smoothing using the diurnalAdj function in the ACDm package.

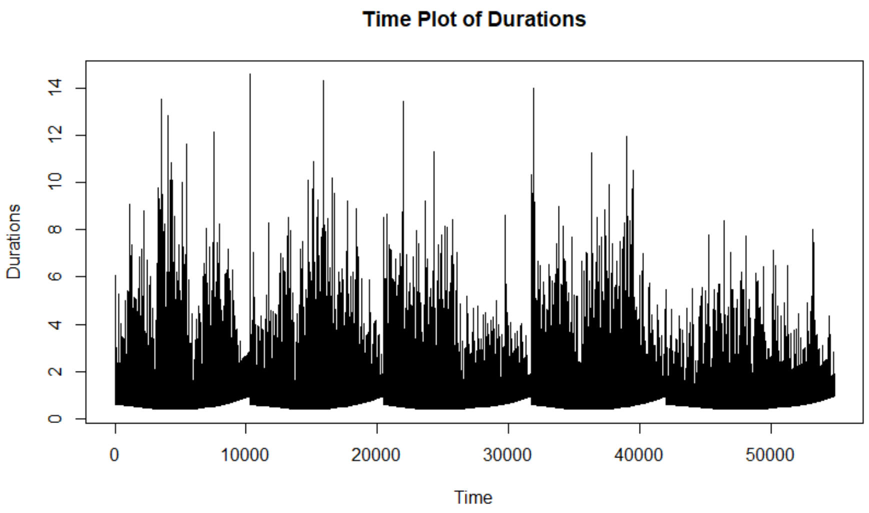
- The latter case has been chosen here as no underlying information was known.

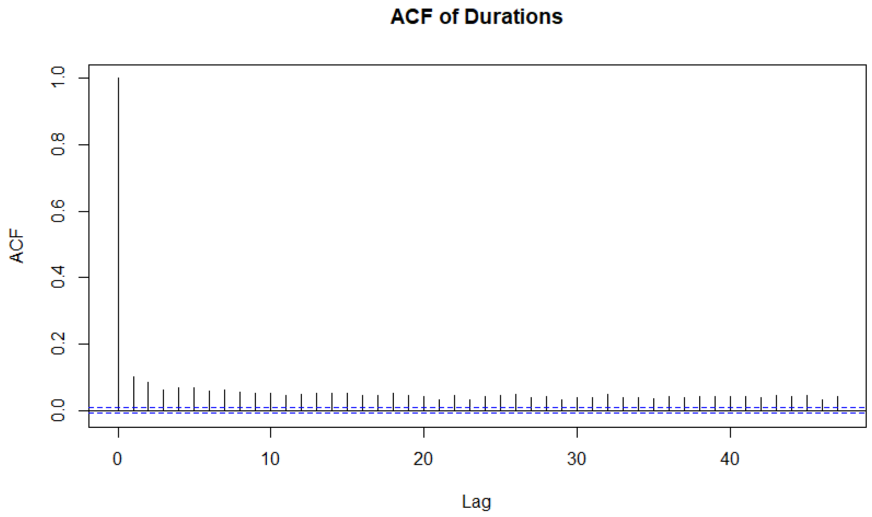
- Some options within diurnalAdj:
- <u>Methods</u> to remove the daily seasonal component include Cubic Splines, Friedman's SuperSmoother, a Smoothed version of the Cubic Spline, and Flexible Fourier Form.
- <u>Aggregation</u> options include "weekdays", "all", or "none". The "all" option smooths each day in the same way, the "none" option smooths each day separately. Only these two were used.
- <u>Nodes</u> must be used for the cubic and smooth spline options. Signifies the start and end of the trading day (in minutes after midnight) and intervals of a given length within them. The nodes use the means of the intervals.

Removing Cycle

- Results:
- Smoothing each day the same way (i.e. aggregation = “all”) does not resolve the diurnal cycle.
- The plots below are for a Friedman’s SuperSmoother with aggregation = “all”, and span = “cv” which effects the smoothness of the curve.







- Diurnal cycle has not been resolved. This was the typical case when aggregation = “all” was chosen. Most plots were very similar to the above 3 plots.

Removing Cycle

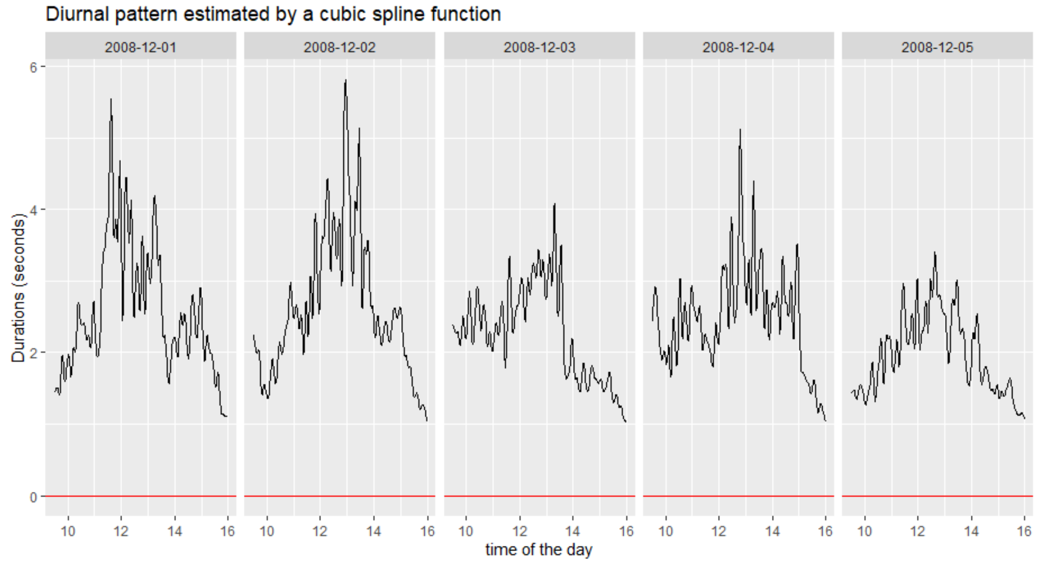
- Results:
- The best smoothers all came from using aggregation = “none” where each day was smoothed differently. The three best smoothers were:

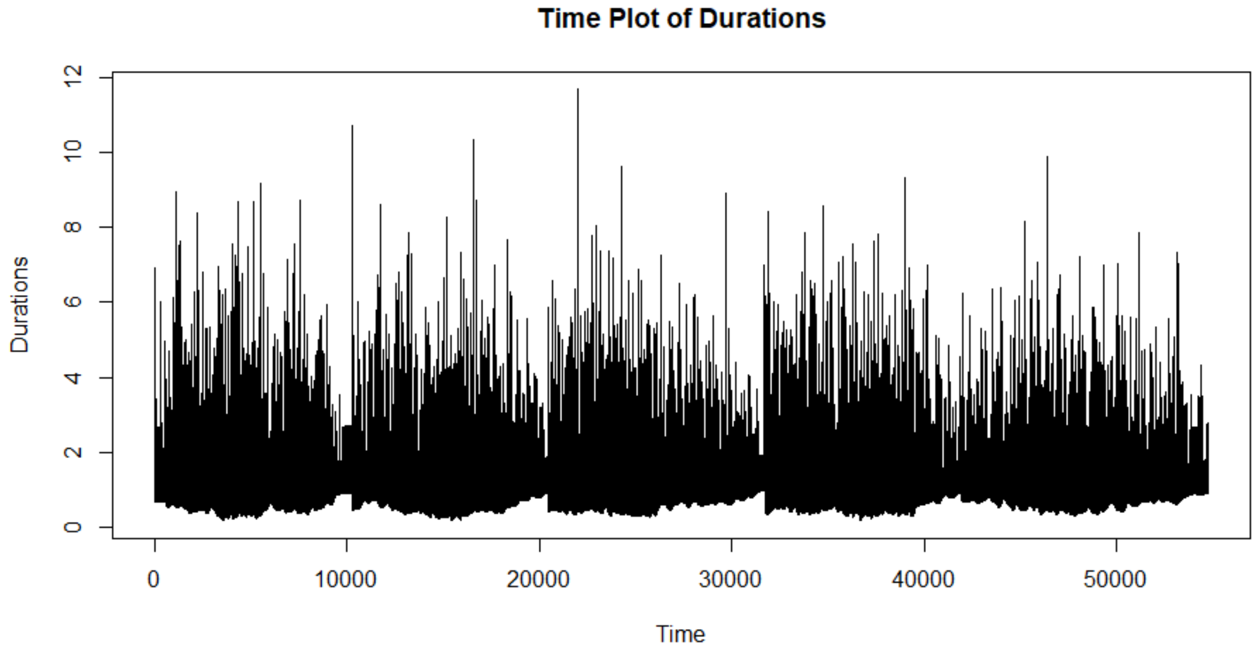
1. Smoothed version of the Cubic Spline using 5-minute time intervals for the nodes.
2. Cubic Spline using 5-minute time intervals for the nodes.
3. Flexible Fourier Form using Q = 50 (which is the number of trigonometric function pairs).

- Because the Flexible Fourier form used a Q much larger than the default (Q = 4) this model was not considered.
- The AIC and BIC for ACD models produced by the Smoothed Cubic Spline with 5-minute intervals was lower than the regular Cubic Spline. Therefore, only the former is shown in the upcoming plots.

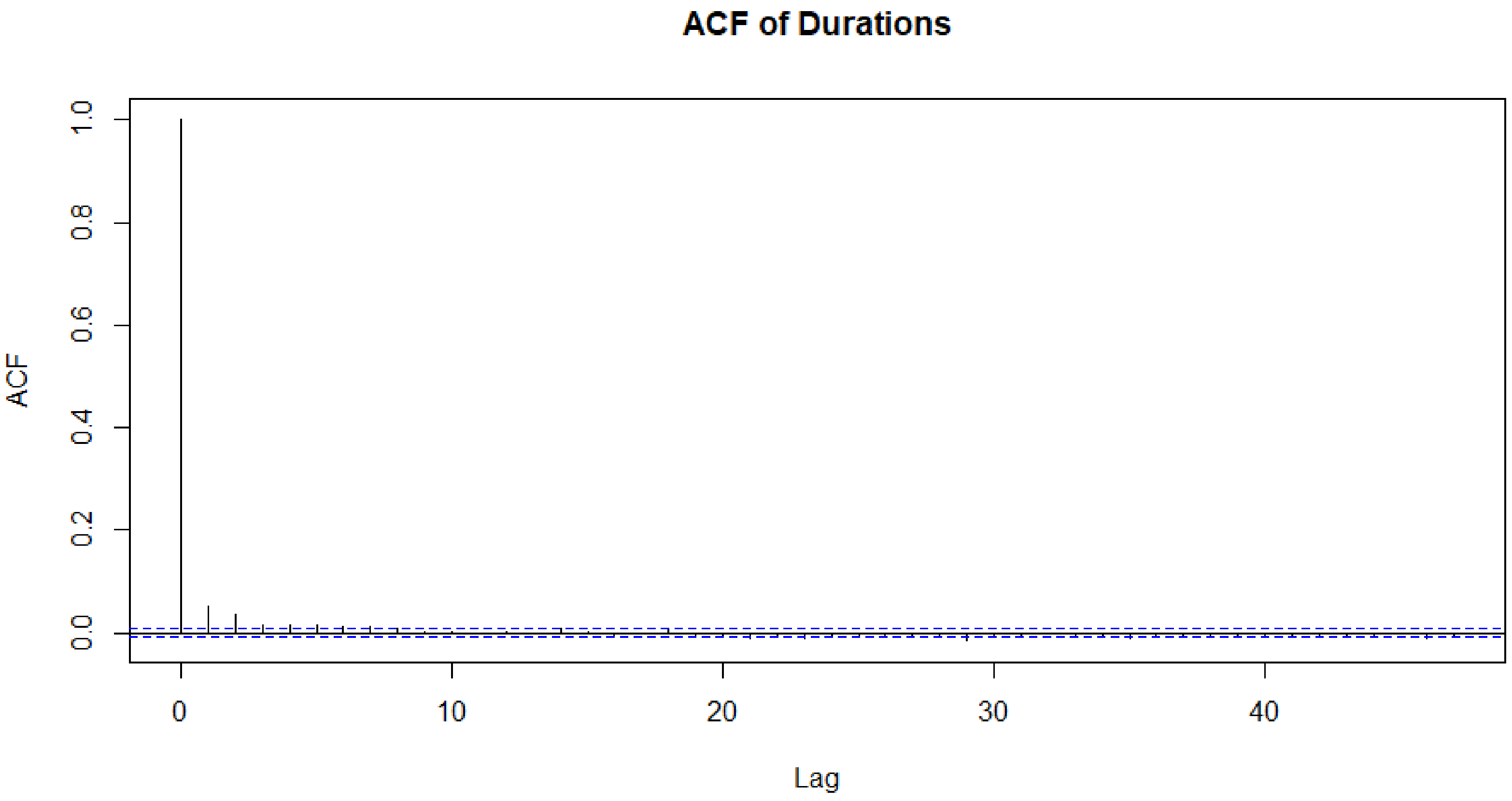
Removing Cycle

- Smoothed Cubic Spline using 5-minute time intervals for the nodes:



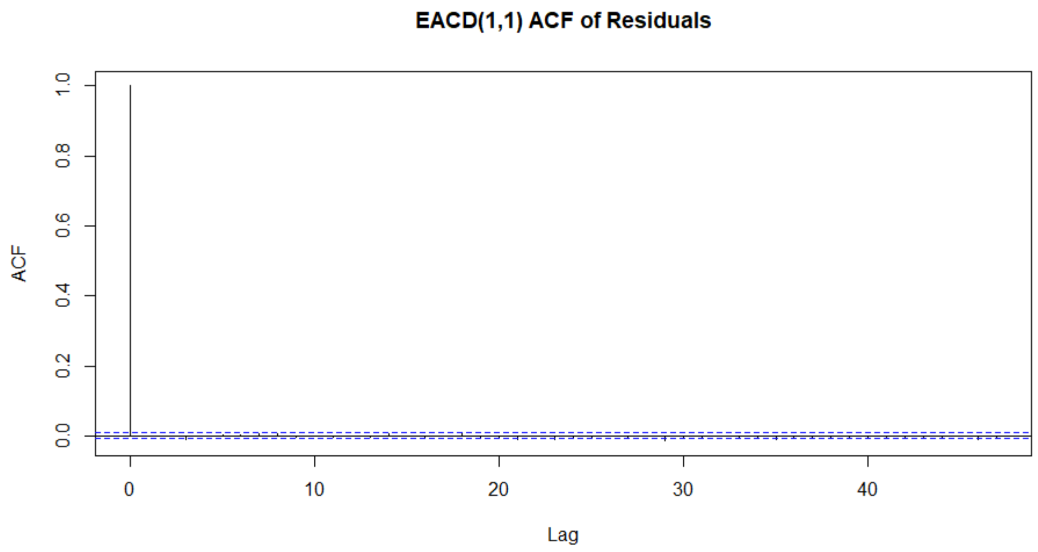


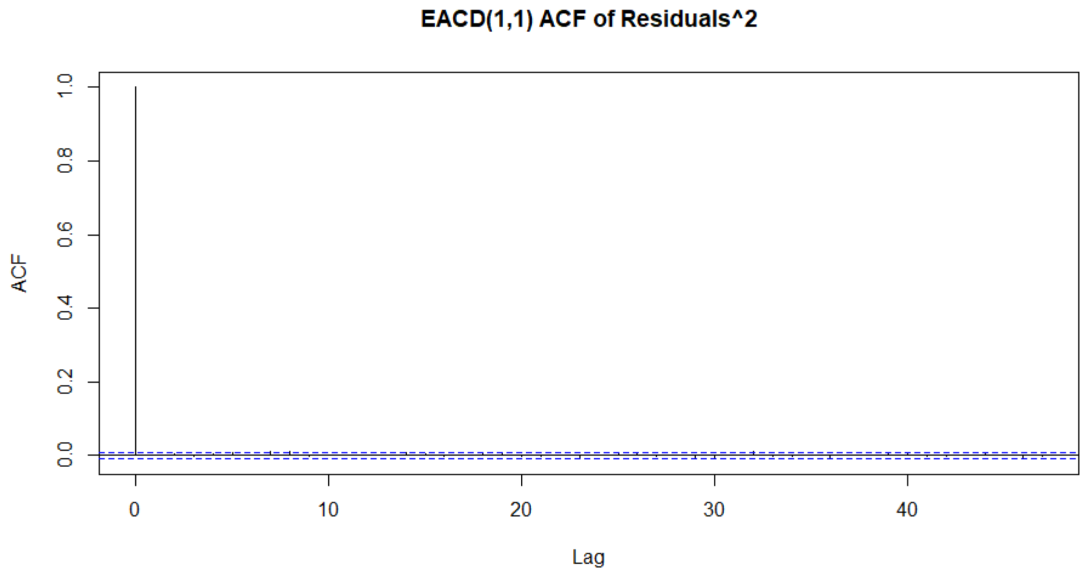
Removing Cycle



Fitting an ACD

- Numerous ACD models were fitted, including Exponential, Weibull, and Generalized Gamma for the assumed error term distribution ϵ_i.
- Only two models passed Ljung-Box and confirmed lack of serial dependence.
- EACD(1,1)

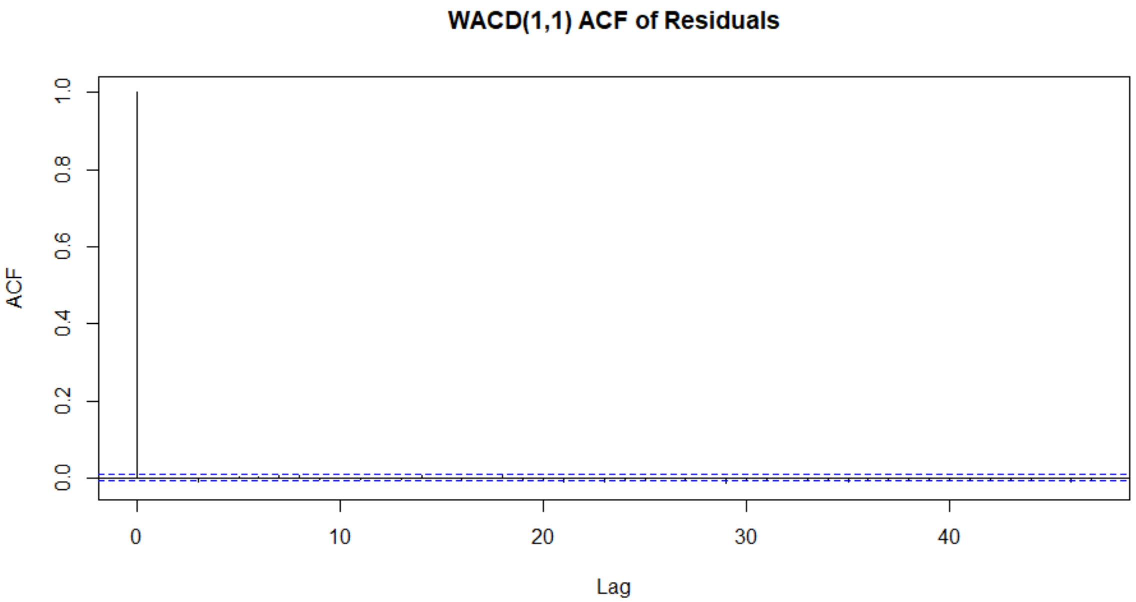


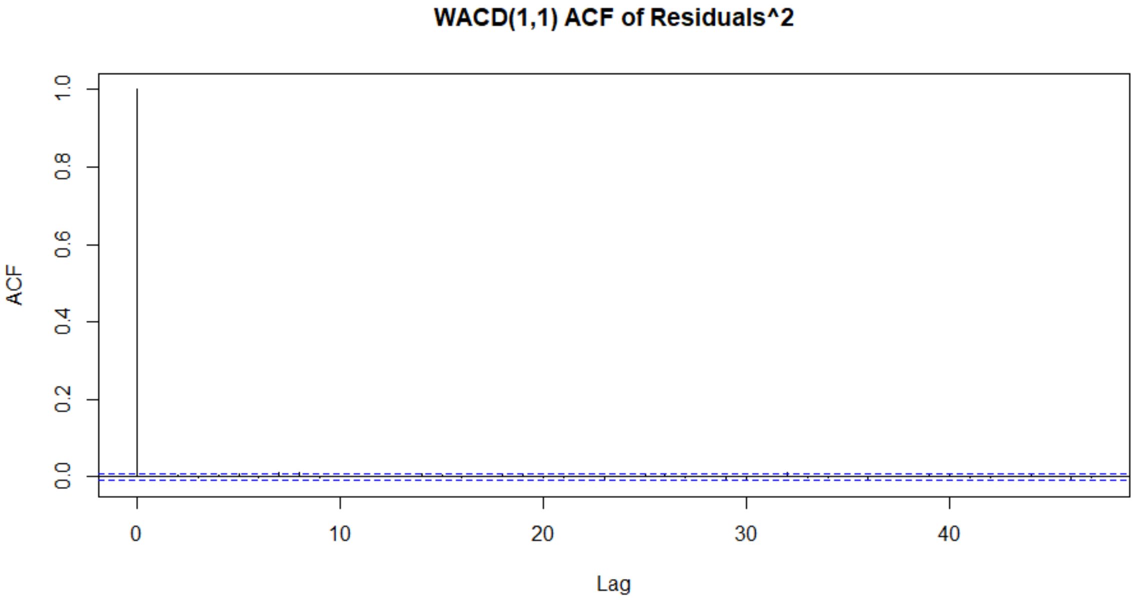


- Ljung-Box statistics p-values > 0.05 for Q(10) and Q(20) for the residuals, and the p-values > 0.05 for Q(10) and Q(20) for the squared residuals.

Fitting an ACD

- WACD(1,1)





- Ljung-Box statistics p-values > 0.05 for Q(10) and Q(20) for the residuals, and the p-values > 0.05 for Q(10) and Q(20) for the squared residuals.

Fitting an ACD

- When determining the “best” model, the AICs and BICs for each respective model were compared below:

Model	AIC	BIC
EACD(1,1)	1.091146e+05	1.091414e+05
WACD(1,1)	9.839056e+04	9.842621e+04

- The AIC and BIC was lower for the Weibull ACD model. A WACD(2,2) was also checked but did not provide a lower AIC or BIC.

Fitting an ACD

- Best model is a WACD(1,1).
- Fitted model:

$$x_i = \psi_i \epsilon_i, \qquad \psi_i = 0.2946 + 0.0537 x_{i-1} + 0.6553 \psi_{i-1}, \qquad \epsilon_i \sim \text{Standardized Weibull}(\alpha = 1.4051)$$

Prediction

- In the ACDm package look at acdFit-methods.

`acdFit-methods` *Methods for class acdFit*

Description

`residuals.acdFit()` returns the residuals and `coef.acdFit()` returns the coefficients of a fitted ACD model of class 'acdFit', while `print.acdFit()` prints the essential information. `predict.acdFit()` predicts the next N durations by thier expected value.

Usage

```
## S3 method for class 'acdFit'
residuals(object, ...)
## S3 method for class 'acdFit'
coef(object, returnCoef = "all", ...)
## S3 method for class 'acdFit'
print(x, ...)
## S3 method for class 'acdFit'
predict(object, N = 10, ...)
```


Arguments

`object`	the fitted ACD model of class 'acdFit' (as returned by the function acdFit).
`x`	same as `object`, ie. an object of class 'acdFit'.
`returnCoef`	on of "all", "distribution", or "model". Specifies whether all estimated parameters should be returned or only the distribution parameters or the model (for the conditional mean duration) parameters.
`N`	the number of the predictions in `predict`.
`...`	additional arguments to `print`.

- I was not able to use predict(model, N = 3) as I could not figure out the 3rd parameter and repeatedly got an error.
- There is a package racd which may be able to do ACD forecasting but there is very little information on it and no R Cran documentation.
- Literature: I was not able to find anything related to out of sample forecasts for an ACD model.
- Conclusion: Not able to do forecasting with the given WACD(1,1) model.

Resources

- *Analysis of Financial Time Series*, by Ruey Tsay (2010, 3rd ed.), Wiley.
- https://cran.r-project.org/web/packages/ACDm/ACDm.pdf

Code

```
# Code of ACD modeling highlights
setwd(...)
library(ACDm)

# Preparing data:
Bt=read.csv("Boeing.csv")
head(Bt, 5)
dur=computeDurations(Bt, open = "09:30:00", close = "16:00:00", rm0dur = TRUE, type = "trade")
head(dur, 5)

# Analysis of diurnal cycle
ts.plot(dur$durations, ylab = "Durations", main = "Time Plot of Durations", xlab = "Time Index")
abline(v = 10274, col = "red", lty = 2)
abline(v = 20432, col = "red", lty = 2)
abline(v = 31708, col = "red", lty = 2)
abline(v = 41968, col = "red", lty = 2)
acf(dur$durations, main = "ACF of Durations")

# Nonparametric Smoothing
data_new2 <- diurnalAdj(dur = dur, method = "supsmu", aggregation = "all", span = "cv")
ts.plot(data_new2$adjDur, ylab = "Durations", main = "Time Plot of Durations")
acf(data_new2$adjDur, main = "ACF of Durations")
```

Code

```
data_new5 <- diurnalAdj(dur = dur, method = "cubicSpline", aggregation = "none", nodes = c(seq(570,960,5),960))
ts.plot(data_new5$adjDur, ylab = "Durations", main = "Time Plot of Durations")
acf(data_new5$adjDur, main = "ACF of Durations")

data_new7 <- diurnalAdj(dur = dur, method = "smoothSpline", aggregation = "none", nodes = c(seq(570,960,5),960), spar = 0)
ts.plot(data_new7$adjDur, ylab = "Durations", main = "Time Plot of Durations")
acf(data_new7$adjDur, main = "ACF of Durations")

data_new8 <- diurnalAdj(dur = dur, method = "FFF", aggregation = "none", Q = 50)
ts.plot(data_new8$adjDur, ylab = "Durations", main = "Time Plot of Durations")
acf(data_new8$adjDur, main = "ACF of Durations")

data_new = data_new7

# Fit ACD models
m1 <- acdFit(durations = data_new, model = "ACD", dist = "exponential")
acf(m1$residuals, main = "EACD(1,1) ACF of Residuals")
acf(m1$residuals^2, main = "EACD(1,1) ACF of Residuals^2")
Box.test(m1$residuals, lag = 10, "Ljung")
Box.test(m1$residuals^2, lag = 10, "Ljung")
Box.test(m1$residuals, lag = 20, "Ljung")
Box.test(m1$residuals^2, lag = 20, "Ljung")
```

Code

```
m2 <- acdFit(durations = data_new, model = "ACD", dist = "weibull")
acf(m2$residuals, main = "WACD(1,1) ACF of Residuals")
acf(m2$residuals^2, main = "WACD(1,1) ACF of Residuals^2")
Box.test(m2$residuals, lag = 10, "Ljung")
Box.test(m2$residuals^2, lag = 10, "Ljung")
Box.test(m2$residuals, lag = 20, "Ljung")
Box.test(m2$residuals^2, lag = 20, "Ljung")

# Best model
m1$goodnessOfFit
m2$goodnessOfFit
```

Thank you