Lecture 06
Prediction Tasks:
Time Series & Regression

Oran Kittithreerapronchai[1]

[1]Department of Industrial Engineering, Chulalongkorn University
Bangkok 10330 THAILAND

last updated: August 7, 2023

Outline

1. Review of Forecasting
2. Time Series Techniques
3. Review of Regression
4. ARIMA = Time Series + Regression
5. Cluster with with R

source: General references [NC20, TSK16, BG19, Pat14]

Recommend Textbook

Hyndman, J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2 .

Types of Forecasting

- **Qualitative/Judgmental:** using subjective inputs
- **Time Series:** using it own past data as inputs
  - Smoothing Technique
  - Trend and Seasonality
  - Classical Decomposition Method
- **Causal/Regression:** using related data/factors as inputs
- **Simulation:** using both Time Series and Causal in computer simulation
  - Imitate consumer choices that give rise to demand
  - Combine time series and causal methods

Component of Observation

$$A = F + E + \epsilon$$

- **Observation:**(A) actual data from history
- **Systematic component:** (F) expected value of demand/ forecasting value

  Level: current de-seasonalized demand
  Tread: growth or decline in demand
  Seasonality: predictable seasonal fluctuation
  Irregular: error or residuals

- **Forecast error:** (E) difference between forecast and actual demand
- **Random component:** (ϵ) part of the forecast that deviates from the systematic component

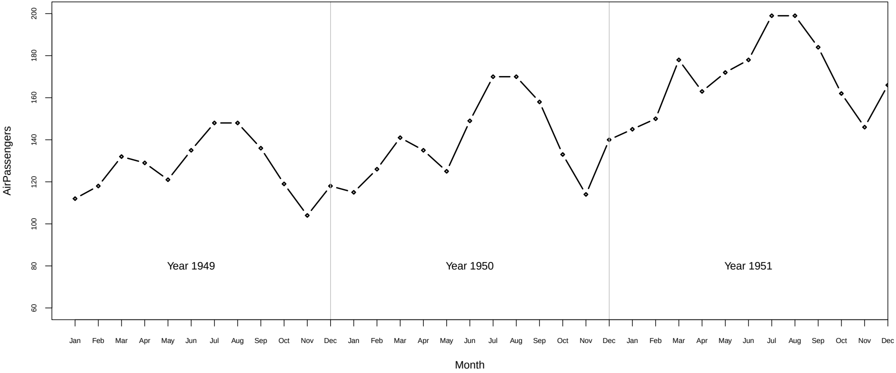
What should we aware before forecast?

- **Description:** story, relationship with other data
- **Time Horizon:** hour, day, week, year
- **Pattern of Data:** seasonal, trend, cycle
- **Forecasting Model:** assumption, data required, parameters, static **VS** dynamic
- **Accuracy:** measuring, how to improve

A good forecaster should:

- be creative & curiosity
- master the 'art' and understand science

Example: US Air Passengers 1949-1951



Facts about Forecasting

- Forecasting is, typically incorrect
- Forecasting is suitable for a group of products
- Forecasting is inaccurate as time horizon increases

source: Chopra and Meindl. 2001. pp. 69

Why do we still need Forecasting?

- Incorrect future is better than knowing nothing
- Incorrect result is manageable

Accuracy of Forecasting

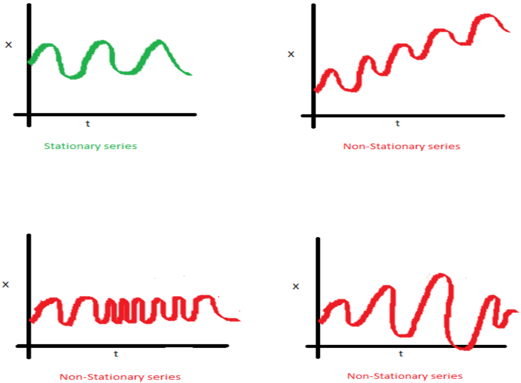
- **Idea:**"average" of $\text{Actual}_t - \text{Forecast}_t$
- **Example:** Mean Error (ME), Mean Absolute Deviation (MAD), Mean Square Error (MSE), Mean Absolute Percentage Error(MAPE), Tracking signal (TS)

$$\text{ME} = \frac{1}{N}\sum_{t=1}^{N} A_t - F_t \qquad \text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(A_t - F_t)^2$$

$$\text{MAD} = \frac{1}{N}\sum_{t=1}^{N} |A_t - F_t| \qquad \text{MAPE} = \frac{1}{N}\sum_{t=1}^{N} \frac{100\,|A_t - F_t|}{A_t}$$

$$\text{bias} = \sum_{t=1}^{N} A_t - F_t \qquad \text{TS} = \frac{\sum_{t=1}^{N} A_t - F_t}{\sum_{t=1}^{N} |A_t - F_t|}$$

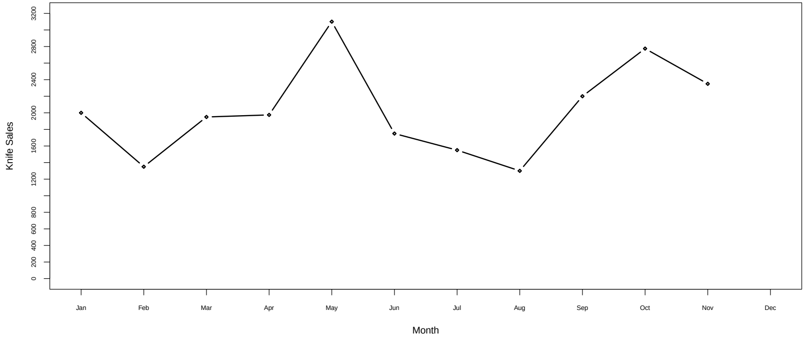
Stationary Process



Data are eventually repeated with the same process

Smoothing: Simple Forecasting Methods

- **Assumption:** recent past ≈ future
- **Time Horizon:** short period
- **Data Pattern:** nearly constant
- **Benefit:** remove randomness, reduce sizes of data
- **Example:** Moving Average, Exponential Smoothing



Moving Average: MA(q)

- using average value of q pervious periods as forecast

$$F_t = \frac{1}{q}\sum_{i=1}^{q} A_{t-i}$$

$$\begin{aligned} F_t &= \text{Smoothing value at time } t \\ A_t &= \text{Actual value at time } t \\ q &= \text{Numbers of interested period} \end{aligned}$$

Example of Moving Average

Month	Knife Demands	MA(3)	MA(5)
Jan	2000	-	-
Feb	1350	-	-
Mar	1950	-	-
Apr	1975	1767	-
May	3100	1758	-
Jun	1750	2342	2075
Jul	1550	2275	2025
Aug	1300	2133	2065
Sep	2200	1533	1935
Oct	2770	1683	1980
Nov	2350	2092	1915
Dec	-	2440	2034

source: Singkarlsiri C., 1997. pp.10-25

Exponential Smoothing Model

- using a previous value and previous error as forecast

$$
\begin{aligned}
F_t &= F_{t-1} + \alpha\,(A_{t-1} - F_{t-1}) \\
&= \alpha\,A_{t-1} + (1-\alpha)F_{t-1}
\end{aligned}
$$

$$
\begin{aligned}
F_t &= \text{Smoothing value at time } t \\
A_t &= \text{Actual value at time } t \\
\alpha &= \text{Exponential factor}, \alpha \in [0,1]
\end{aligned}
$$

- **Idea:** Forecast $= \alpha$ Actual $+ (1-\alpha)$ Old Forecast

Why do we call "Exponential Smoothing"?

$$
\begin{aligned}
F_t &= \alpha\ A_{t-1} + (1-\alpha)\mathbf{F_{t-1}} \\
&= \alpha\ A_{t-1} + (1-\alpha)\left[\alpha\ A_{t-2} + (1-\alpha)F_{t-2}\right] \\
&= \alpha\ A_{t-1} + \alpha(1-\alpha)A_{t-2} + (1-\alpha)^2\mathbf{F_{t-2}}
\end{aligned}
$$

What does it mean?

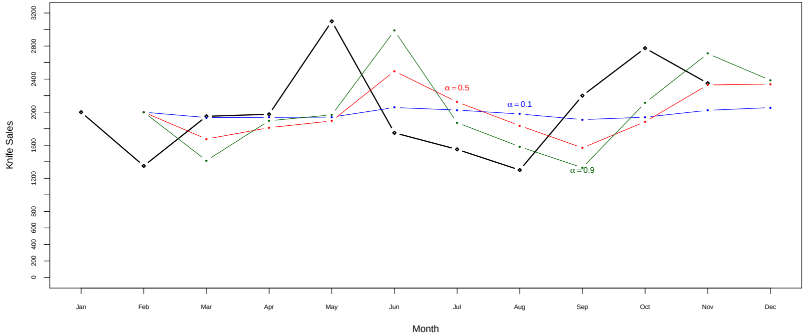
- Effects of actual value and error exponentially decay
- α controls the decay rate; F_1 is initial forecast value
- if $\alpha = 0$, no effect of actual value
- if $\alpha = 1$, no effect of forecast value

How to choose F_1 and α?

- **Good News**: effect of F_0 will decay; typically $F_1 = A_1$
- **Bad News**: select 'right' α is difficult $\rightarrow$ try out and error

Month	Knife Demands	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
Jan	2000	-	-	-
Feb	1350	2000	2000	2000
Mar	1950	1935	1675	1415
Apr	1975	1937	1813	1897
May	3100	1940	1894	1967
Jun	1750	2056	2497	2987
Jul	1550	2026	2123	1874
Aug	1300	1978	1837	1582
Sep	2200	1910	1568	1328
Oct	2770	1939	1884	2113
Nov	2350	2023	2330	2709
Dec	-	2056	2340	2386

Excerise:



Knife Sales
3200
2800
2400
2000
1600
1200
800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
α = 0.5
α = 0.1
α = 0.9

Basic Time Series in R

- **ts:**

```
set.seed(937)
myTS <- ts(10*rnorm(24)+30,start=c(2008,1),frequency = 12)
summary(myTS) ; tsp(myTS) ; frequency(myTS) ; deltat(myTS)
cycle(myTS); time(myTS)

summary(AirPassengers)
decompose(AirPassengers,type="multiplicative")
plot(decompose(AirPassengers))
require(ggplot2)  ;  autoplot(AirPassengers)

air.trin  <- window(AirPassengers,start=1948.0,end=1958.917)
air.test  <- window(AirPassengers,start=1959.0,end=1959.917)
```

- forecast

```
require(forecast)
tsdisplay(air.trin)         ;  tsoutliers(air.trin)     ;  tsclean(air.trin)
ggseasonplot(air.trin)      ;  ggmonthplot(air.trin)    ;  ggtsdisplay(air.trin)
##-- transform using moving avg or box-cox
lambda.opt <- BoxCox.lambda(air.trin)  ##
autoplot(BoxCox(air.trin,lambda = lambda.opt))
```

- **Simple**

```
naive(air.trin,h=12)$mean     ; meanf(air.trin,h=12)$upper
snaive(air.trin,h=12)$fitted  ; rwf(air.trin,h=12,drift=T)$residuals ## rand walk
autoplot(air.trin) +
  autolayer(meanf(air.trin, h=12),series="Mean", PI=F) +
  autolayer(naive(air.trin, h=12),series="ïNave", PI=F) +
  autolayer(snaive(air.trin, h=12),series="S.ïnave", PI=T)
```

Adv Time Series in R

- **Overall:**

```
air.holt <- holt(air.trin,h=12) ; autoplot(air.holt)
forecast(air.holt,h=12)         ; predict(air.holt,n.ahead=12)

checkresidual(air.holt)         ;    accuracy(air.holt,air.test)

autoplot(ses(air.trin,h=12))    ;    autoplot(holt(air.trin,h=12))
autoplot(hw}(air.trin,h=12))    ;
autoplot(est(air.trin,alpha=0.5))

rbind(accuracy(ses(air.trin,h=12,alpha=0.5,initial="simple")),
  accuracy(ses(air.trin,h=12))
)
ets(air.trin)         ## multi-purpose and optimal tool for expo series
```

- **Decomp:**

```
ma(air.trin, order=11)          ;   autoplot(decompose(air.trin))
air.trin.stl <- stl(air.trin,t.window=13, s.window="periodic")
autoplot(air.trin.stl)          ;   remainder(air.trin.stl)
seasonal(air.trin.stl)          ; seasadj(air.trin.stl)
trendcycle(air.trin.stl)
```

- **ARIMA:**

```
arima(air.trin,order=c(0,0,0),seasonal=list(order=c(0,0,0)))
acf(stlf(air.trin,h=12)$residuals)          ## check acf of arima
pacf(arima(air.trin,c(1,0,1))$residuals)    ## check partial acf of arima
```

Regression Re-Cap

- **Regression**: a function of ind. variables (predictors, x_i) to predict a dep variable (response; y)
- **Linear Regression**: predicting y with a linear function of x_i as follows:

$$y = \beta_0 + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1n}x_n$$

- **Assumptions**:
  - **L**inear relationship between predictors and responses → plot
  - **I**ndependent predictors → VIF ≤ 4
  - multivariate **N**ormal of all variables → Q-Q plot, `ks.test()`
  - equal **E**rror terms in regression, a.k.a Homoscedasticity
  - little or no **AUTO**correlation → DW ≈ 2

Linear Regression as Matrix

$$
\begin{aligned}
Y_i &= \beta_0 + \boldsymbol{\beta_1} \mathbf{X}_i + \epsilon_i, \\
&\quad \text{where } \epsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2) \\
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} \beta_0 + \boldsymbol{\beta}_1 \mathbf{X}_1 \\ \beta_0 + \boldsymbol{\beta}_1 \mathbf{X}_2 \\ \vdots \\ \beta_0 + \boldsymbol{\beta}_1 \mathbf{X}_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
\mathbf{Y} &= \begin{bmatrix} \mathbf{1} & \mathbf{X}_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{bmatrix} + \boldsymbol{\epsilon} \\
&= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}
\end{aligned}
$$

Solve for $\boldsymbol{\beta}$ to min $\boldsymbol{\epsilon}^2$

$$
\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \equiv \boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= [\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}]'\,[\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}] \\
& \quad \text{apply FOC on } \boldsymbol{\beta} \\
\mathbf{0} &= -2\mathbf{X}'\,[\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}] \\
(\mathbf{X}'\mathbf{X})\,\boldsymbol{\beta} &= \mathbf{X}'\mathbf{Y} \\
\boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}
\end{aligned}
$$

Simple Linear Regression in R

- **Building:**

```
data("Davis", package="carData")    ;     self <- Davis
View(self)                          ;     edit(self)
summary(self)                       ## any error?
lm(weight~.,data=self)              ## linear?
self.lm <-  lm(weight~I(height), data=self )
fit.lm  <-  step(lm(weight~., data=self))
summary(fit.lm)                     ;     plot(fit.lm)
```

- **Predict:**

```
testData <- data.frame(sex=factor(c('M','M','F','F'))
                       ,weight=c(80,77,80,77),height=c(182,161,182,161)
                       ,repwt=c(78,78,78,78),repht=c(180,170,180,170))
predict(fit.lm,newdata = testData)

confint(fit.lm)                     ;     residuals(fit.lm)
fitted(fit.lm)                      ;     anova(fit.lm)
```

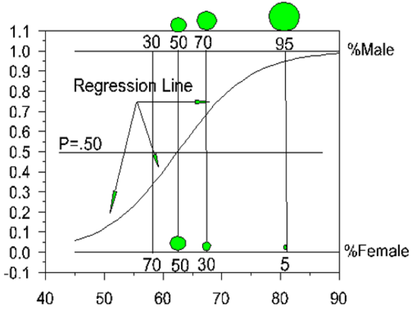
- **Verify:**

```
require(olsrr)
ols_plot_cooksd_chart(fit.lm)   ;     ols_vif_tol(fit.lm)
ols_test_normality(fit.lm)      ;     ols_test_correlation(fit.lm)
ols_test_outlier(fit.lm)        ;     ols_plot_added_variable(fit.lm)
```

General Regression Model

- **What**: an extensions of regression that allows predictors to be any functions and response from other distributions (e.g., Poisson & Binomial) or other complex function
- **Types**:
  - Logistic Regression response or predictor is binary, e.g. forecasting probability
  - Poisson Regression response or predictor is integer, e.g. forecasting number of awards
  - Non-Linear Regression estimation function consists of non-linear terms
  - Non-Parametric Regression estimation function is not predetermined, but based on data (**not cover here**)

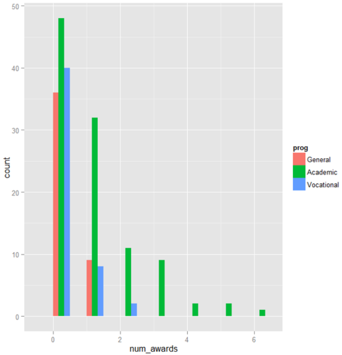
Logistic Regression



- **Concept:** T/F = probability $[0, 1]$
  - **response:** $F(\mathbf{X}) = \frac{1}{1+e^{-(\beta_0+\boldsymbol{\beta_1}\mathbf{X})}}$
  - **linear regression:** $\log\left(\frac{F(\mathbf{X})}{1-F(\mathbf{X})}\right) = \beta_0 + \boldsymbol{\beta_1}\mathbf{X}$
- **R Command**

```
height.glm <-   glm(Gender~Height,data=height,family = "binomial")
prob <-   data.frame(Gender=factor(rep("M",21)),Height=60:80)
prob$prob <   predict(height.glm,newdata= prob,type = "response")
```

Poisson Regression



- **R Command:**

```
glm(num_awards~prog + math, family="poisson", data=award)
predict(award.pois,type="response")
plot(jitter(award$math),jitter(award$num_awards)
  ,col=award$prog,pch=16,cex=0.5,xlab="Math",ylab="# awards")
```

What is ARIMA?

combination of linear regression and traditional time series, i.e.,

$$y_t = \beta_0 + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1n}x_n$$

and

$$x_n = f(y_{t-n})$$

COMPONENTS OF ARIMA: $ARIMA(p, d, q)$

- Autocorrelation (AR): linear regression of previous actual

$$F_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i F_{t-i} + E_t$$

- Integrated (I): previous/lagged value,

$$F_t = \sum_{j=1}^{d} F_{t-d} + E_t$$

- Moving Average (MA): linear regression of previous error

$$F_t = \theta_0 + \sum^{q} \theta_k E_{t-k} E_t$$

Special Cases

Constant = ARIMA(0,0,0)

$$F_t = C + E_t$$

Random Walk = ARIMA(0,1,0) no constant

$$F_t = 0 + F_{t-1} + E_t$$

Simple Expo Smoothing = ARIMA(0,1,1)

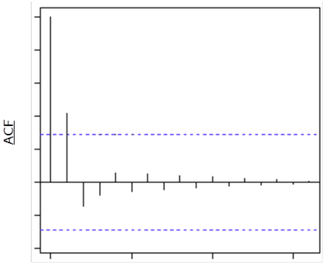
$$\begin{aligned} F_t &= F_{t-1} + \theta_1 E_{t-1} + E_t \\ &= F_{t-1} + \theta_1 (A_{t-1} - F_{t-1}) + E_t \\ &= (1-\theta_1) F_{t-1} + \theta_1 A_{t-1} + E_t \end{aligned}$$

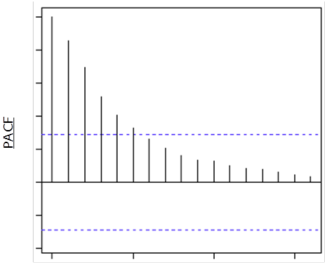
Double Expo Smoothing = ARIMA(0,2,2)

Choosing ARIMA model

- **Autocorrelation Function (ACF)** correlative of series compared to itself (lag-h)
- **Partial Autocorrelation Function (PACF)** ACF after removing effect of previous term

Spike in value of ACF lag-1 to lag-h indicates MA(h), whereas Spike in value of PACF lag-1 to lag-h indicates AR(h)





Case Study: Regression of SG Load

Background

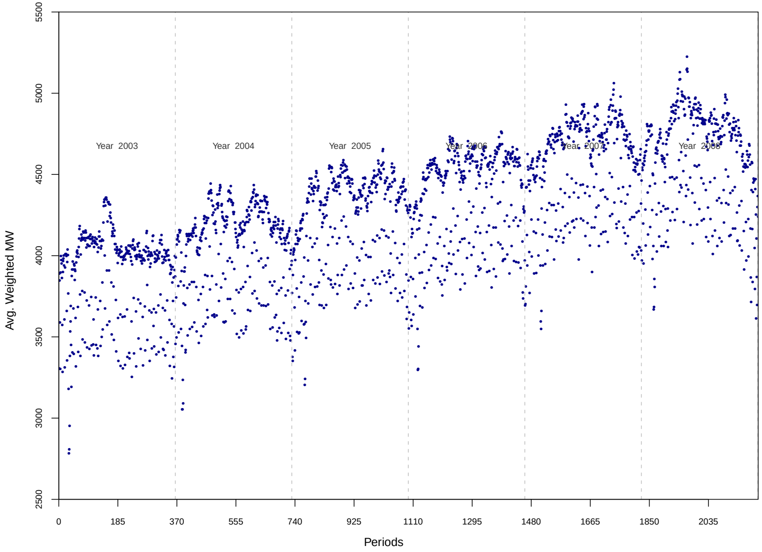
- Deregulated market
- 80% is gas-fired generation plant
- Several disruptions in 2006
- LNG Terminal

Questions

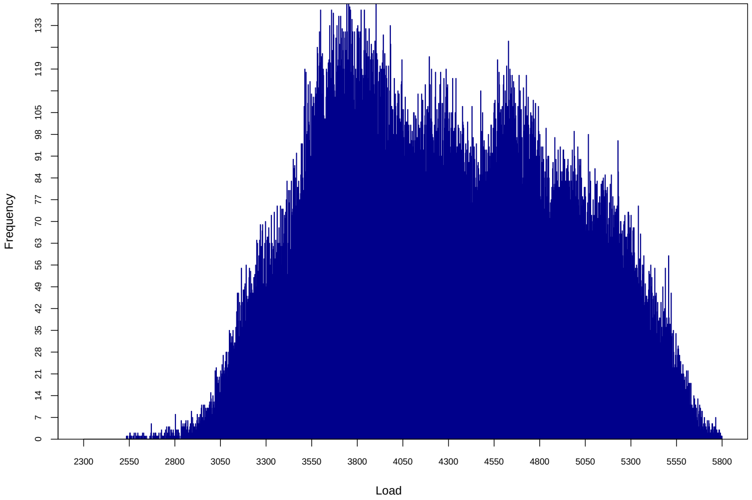
- Can LNG Terminal reduce price volatility?
- When should SG burn LNG and at which portion?

$\rightarrow$ **What is pattern of electricity loads?**

Historical loads: half-hourlydaily



Electricity load histogram



Load Histogram

What should be factors of load?

- Year: regular VS recession; 2003 . . . 2009
- Month: Quarter; January . . . December
- Week: weekday VS weekend; Monday . . . Sunday
- Day: peak VS off-peak; peak VS semi peak VS off-peak (exact time)

Regression model of electrical load

Proposed Model

$$
\begin{aligned}
f(\text{load}_i) &= g(\text{time}_i) \\
&+ a^m B^m(\text{month}_i) + b^w 1_i^w + b^{p\text{-}w} 1_i^{p\text{-}w} + \text{constant} + \epsilon_t
\end{aligned}
$$

, where

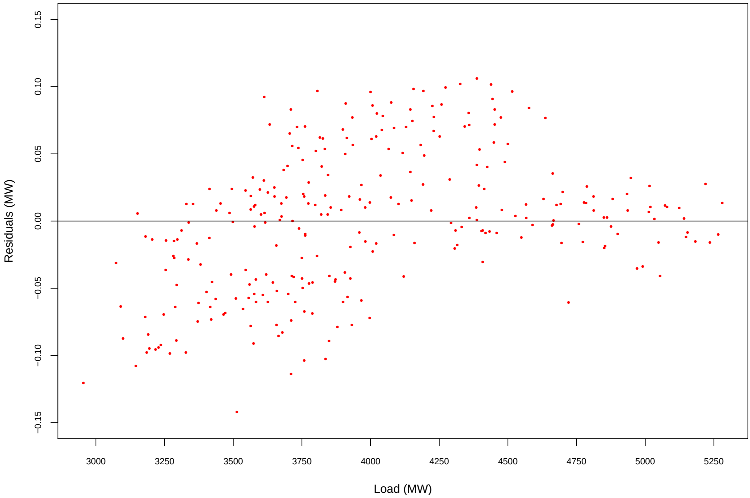
$$
\begin{aligned}
B^m(\text{month}) &= \text{integer for monthly seasonality effect} \\
1^w &= \text{binary for weekday effect} \\
1^{p\text{-}w} &= \text{binary for peak-weekday effect}
\end{aligned}
$$

Selection of Regression model

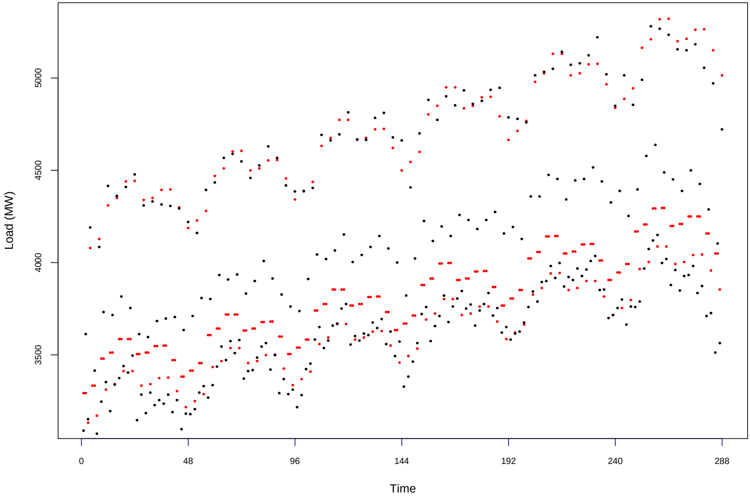
AIC: Fitness of regression model

$g(\cdot)\backslash f(\cdot)$	load	$\ln(\text{load})$	$\sqrt{\text{load}}$	load^2
time	3898.881	-854.503	1120.327	9074.690
$\ln(\text{time})$	3959.229	-806.069	1174.416	9142.393
$\sqrt{\text{time}}$	3902.624	-855.422	1121.700	9083.050
time^2	3947.126	-804.609	1169.673	9118.886

Residuals of ln(load) and time



Actual values and Estimation values



Actual and Estimation values

Reference

[BG19] Brad Boehmke and Brandon M Greenwell.
Hands-on machine learning with R.
CRC press, 2019.

[NC20] Fred Nwanganga and Mike Chapple.
Practical machine learning in R.
John Wiley & Sons, 2020.

[Pat14] Manas A Pathak.
Beginning data science with R.
Springer, 2014.

[TSK16] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar.
Introduction to data mining.
Pearson Education India, 2016.