Input missing values and outliers detection

Juan Camilo Rivera. [1] Hugo Andres Dorado. [1]

[1]Big data and site-specific agriculture
Decision and Policy Analysis
Centro Internacional de Agricultura Tropical

Data analysis course, 2018

Outline

1. Input missing values
   - Vector Autoregression (VAR)
   - RMWAGEN package
2. Outliers detection
   - Univariate approach
   - Cook's distance
   - Welsch-Kuh distance

Vector Autoregression (VAR)

Definition VAR

Vector autoregression is a stocastic model to identify the linear relationship between time series. This is a generalization of AR (autoregressive model)

Definition AR

$AR(p)$ autoregressive model of order p is defined as:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t$$

where $\varphi_1, ..., \varphi_p$ are parameters model, c is constant and ϵ_t is noise.

Vector Autoregression (VAR)

- There is a package, parstm, in R that contains functions to construct a model of autogression.

Example:
Real gross domestic product data in Germany (1960.1-1990.4)

Criterio	1	2	3	4
AIC	-661.60	-680.89	-669.84	-661.54
BIC	-636.30	-644.44	-622.31	-603.0
$F(\phi_{p+1,s} = 0)$	8.54	0.80	1.35	2.91
$p - value$	0.00	0.53	0.26	0.03

Vector Autoregression (VAR)

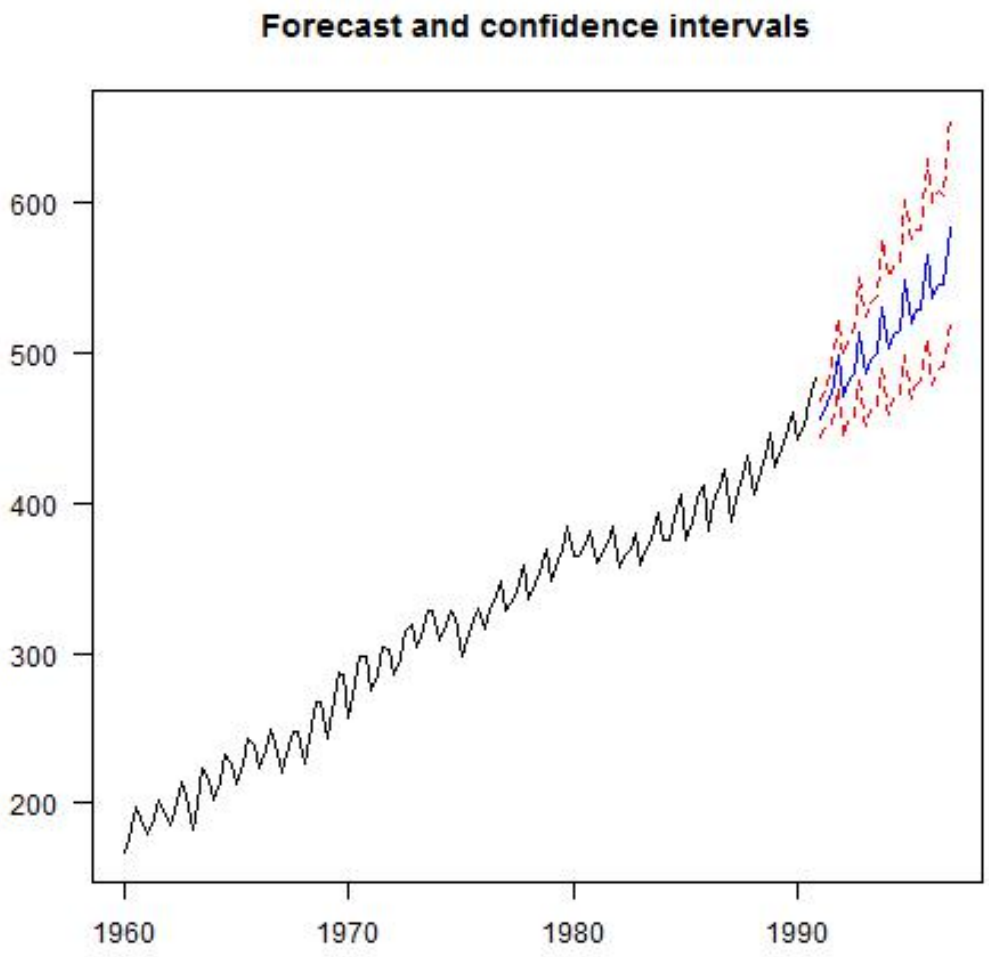
A $F - test$ is used for checking the periodicity of model.
The Fpar.test function computes $p - values$ for F statistc.

Season	$p - value$
Intercepts	0.00
Trends	0.00

The table above shows that periodicity is not rejected.

Vector Autoregression (VAR)

The function predictpiar makes prediction based on VAR model.
The figure shows 24 ahead forecasts in PIAR(2) model.



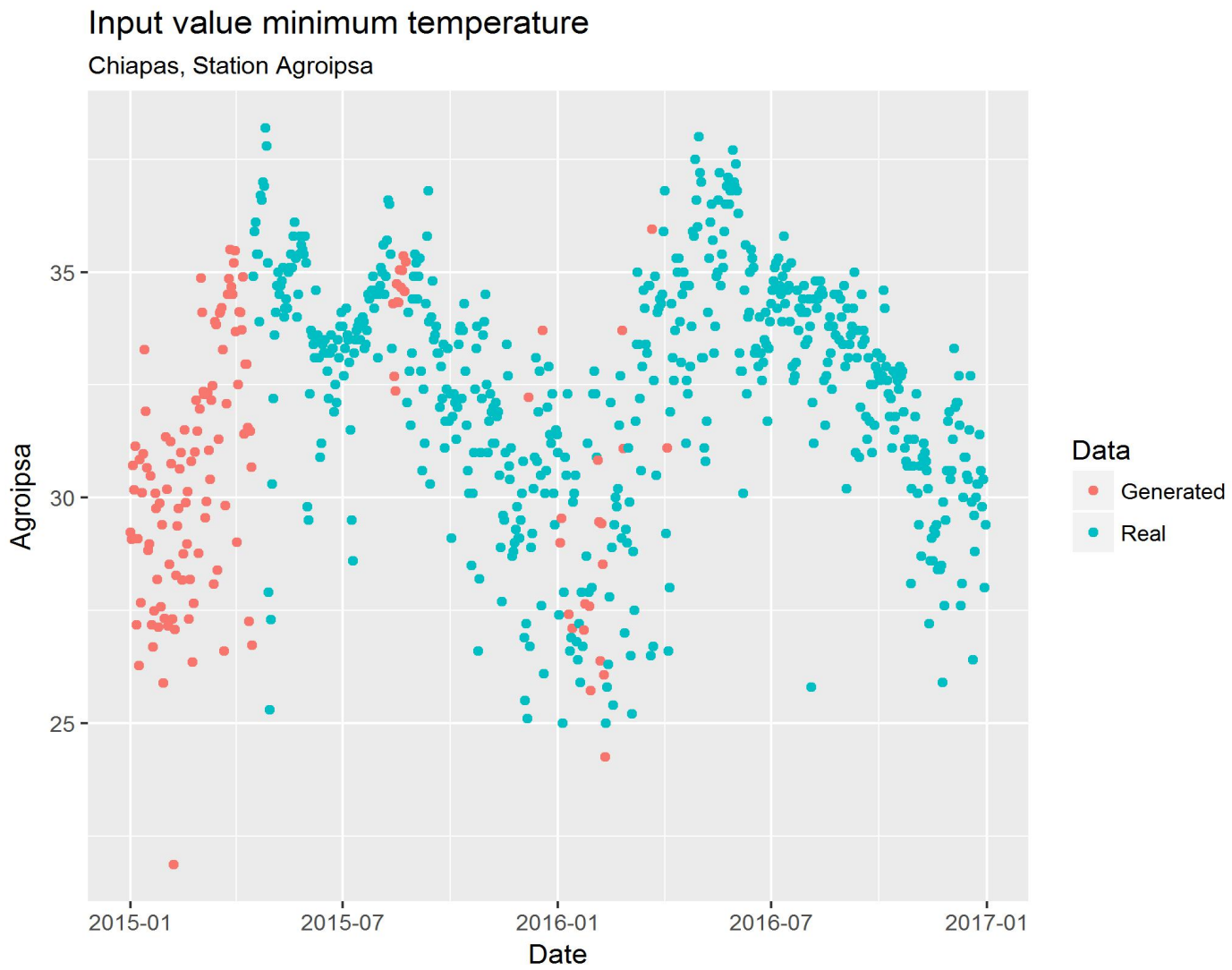
RMWAGEN package

A Daily Weather Generator generates weather time series with same statiscal patterns of the observed ones.
Generates daily of following weather variables:

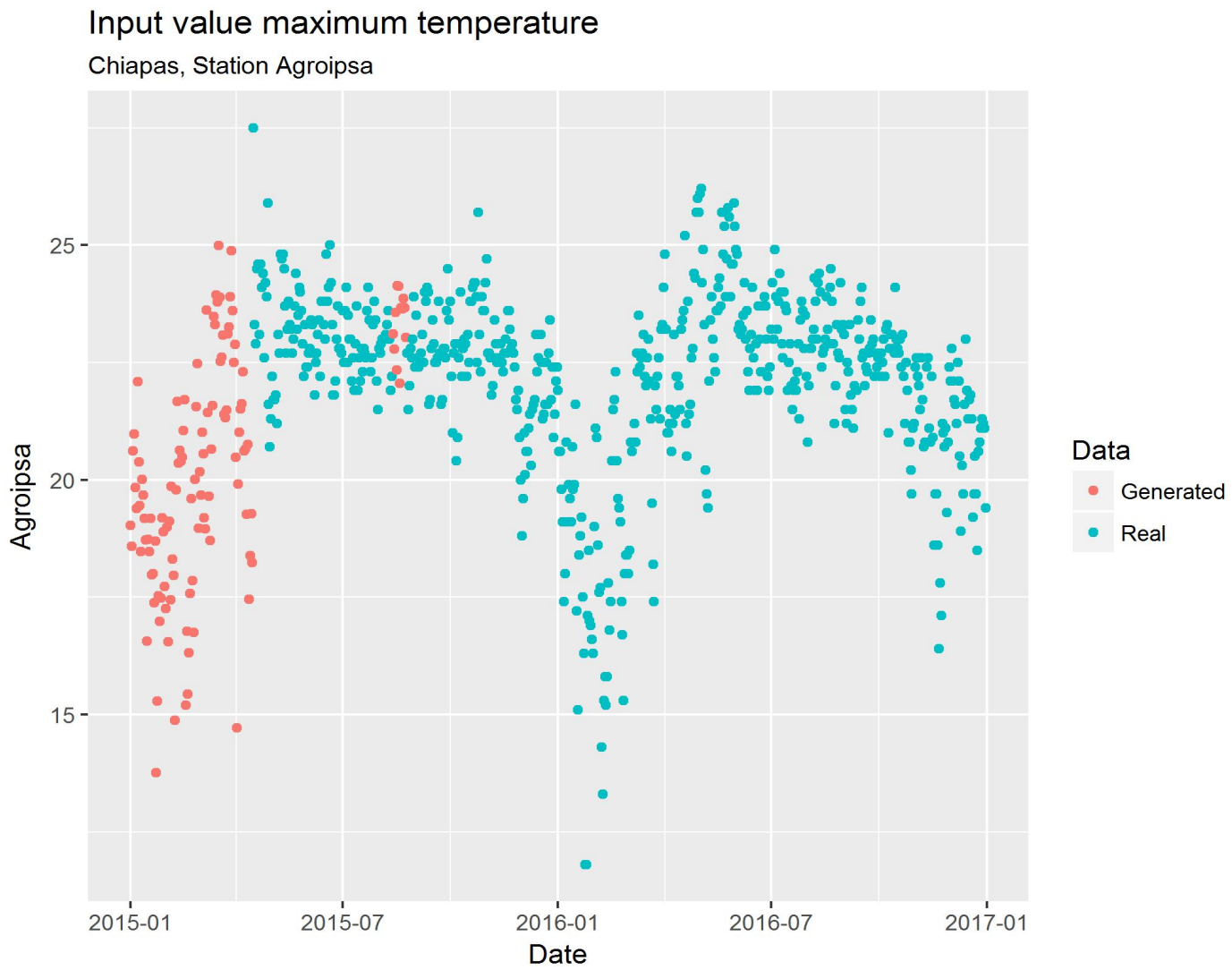
- Maximum temperature
- Minimum temperature
- Precipitation

Based on VAR models.

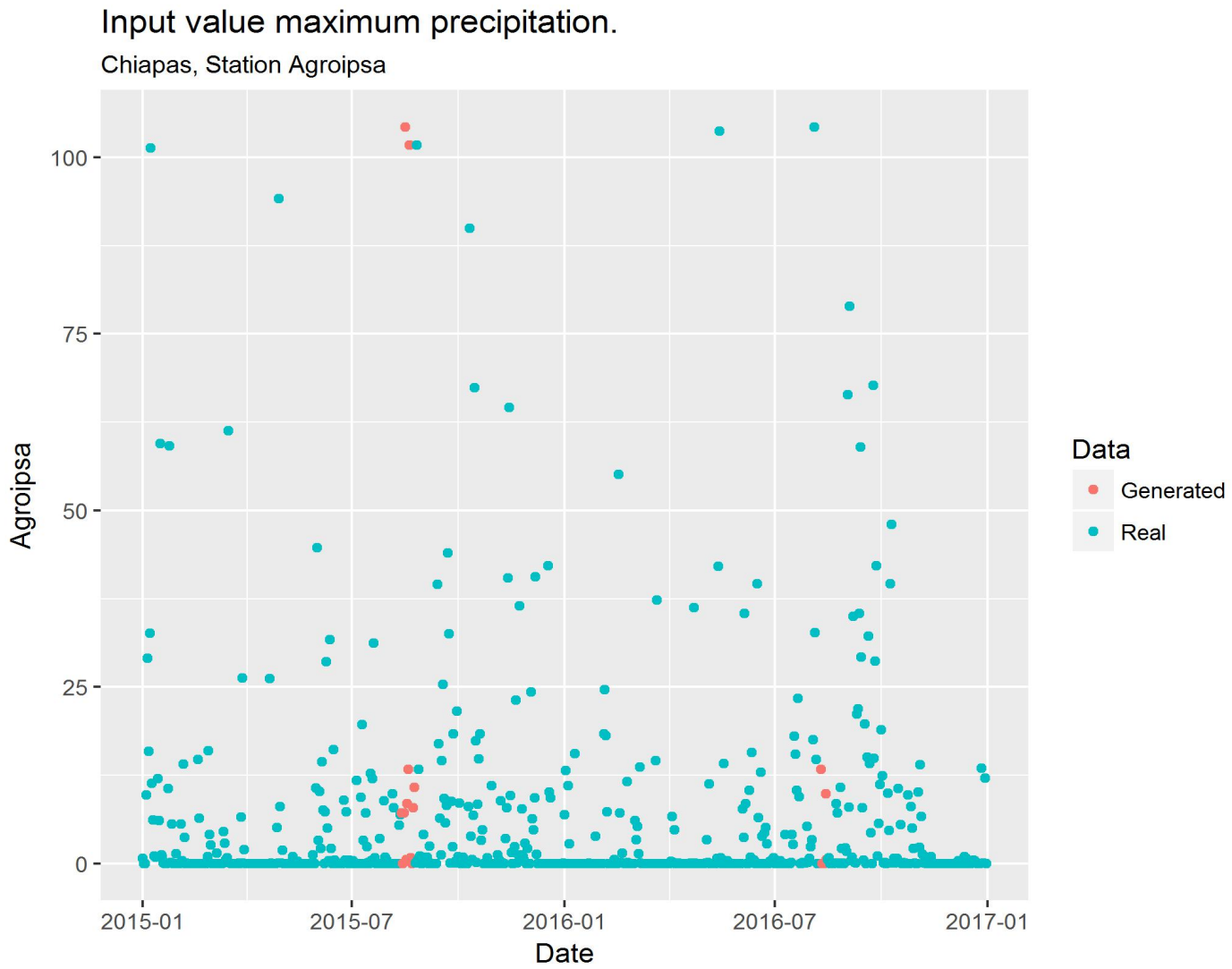
Minimum temperature



Maximum temperature



Precipitation



Outliers detection

Univariate approach

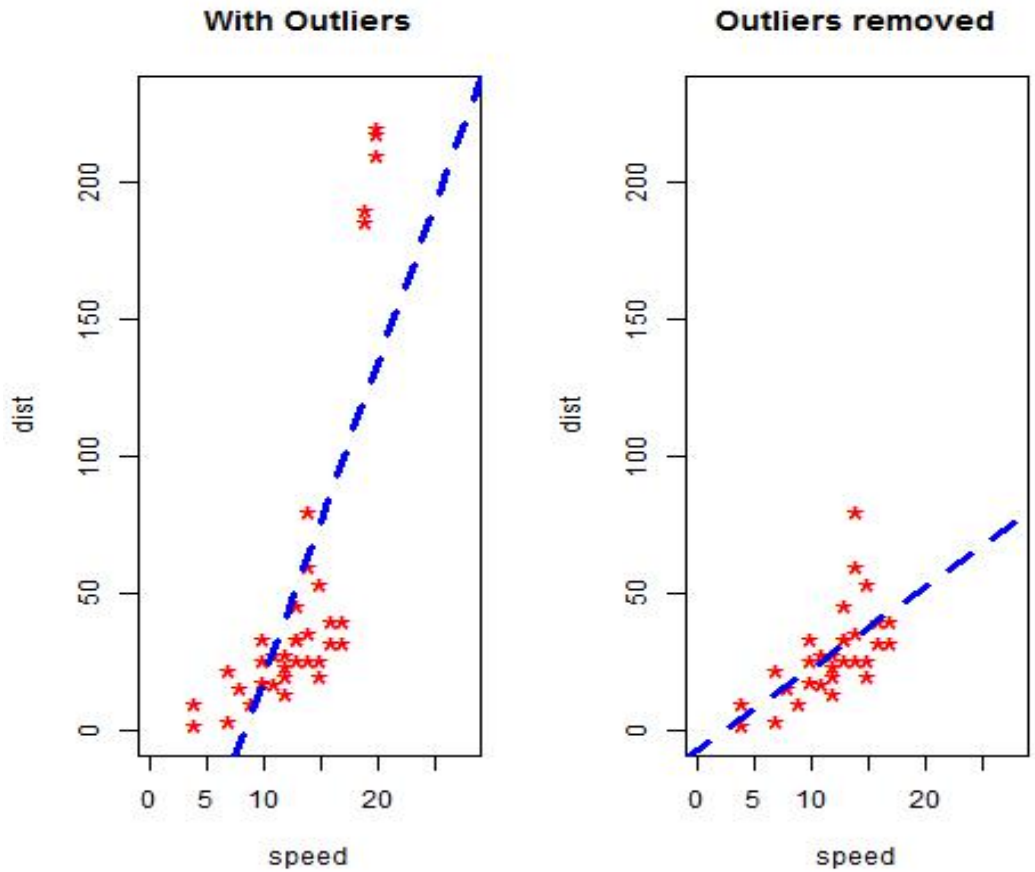
For continuous variable. The outliers are those observations that lie outside:

$$1.5 * IQR \tag{1}$$

where IQR, `Inter Quartile Range` is the difference between 75th and 25th quartiles.

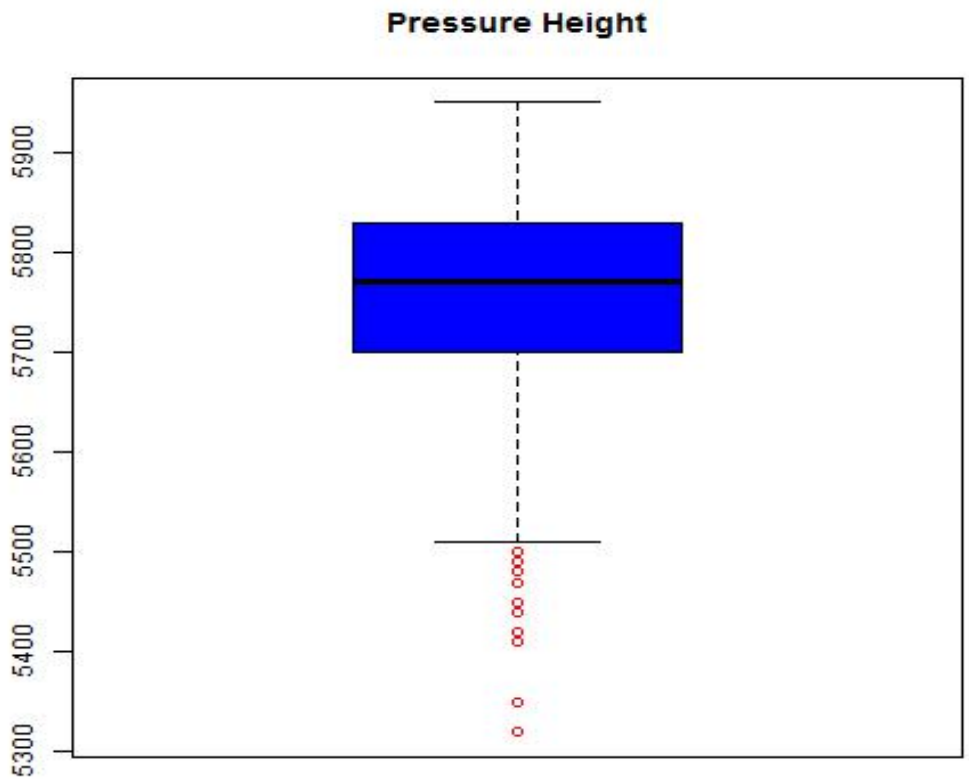
Univariate approach

Example:
Outliers affects linear models fit. The cars is a dataset that is composed by speed and velocity of cars in 1920.



Univariate approach

Example:
Detection of outliers using boxplot. Ozone is a dataset of weather variables of 2015 in U.S.



Cook's distance

It is used to measure the influence of observation i on the estimation of the regression parameters.

$$C_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}^{-i}\right)}{\sigma^2(p+1)} \tag{2}$$

where $\hat{y}_j^{-i}$ is the prediction at point x_j and p amount points. A large value of C_i indicates that ith observation is influential.

The function **cooks.distance ()** computes cooks distance.

Cook's distance

Example:
The weight at birth data is about risks associated with low weight at birth. Using the following linear model for weight at birth (BWT).

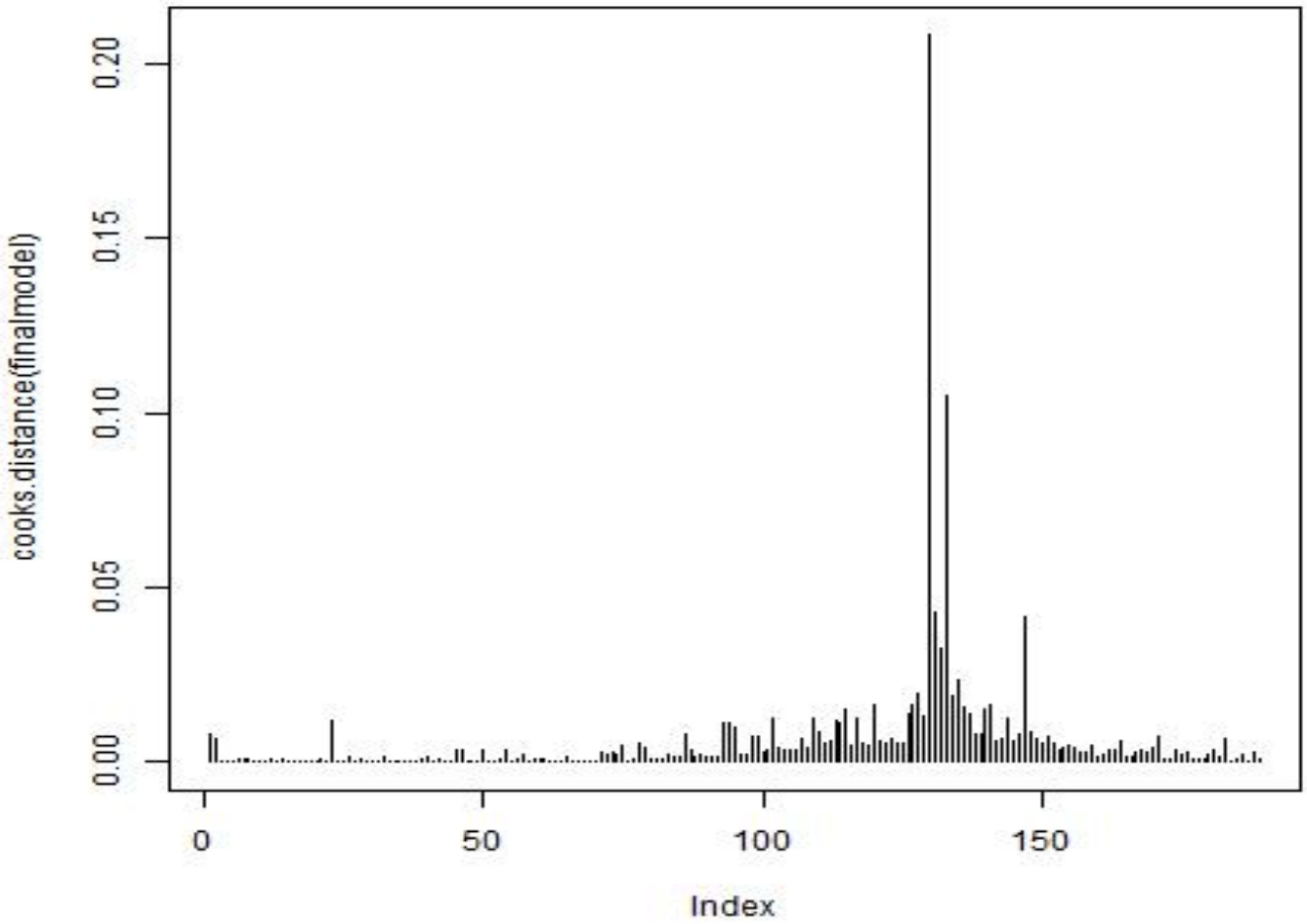
$$BWT = \beta_0 + \beta_1 SMOKE + \beta_2 AGE + \beta_3 LWT \tag{3}$$

where

- SMOKE: smoke during pregnancy, yes = 1 no =0
- AGE: age of mother
- LWT: weight of mother at last menstrual period

Cook's distance

```
finalmodel<-lm(BWT  SMOKE + AGE + LWT, data=birth.weight)
plot (cooks.distance(finalmodel), type ="h")
```



Welsch - Kuh distance

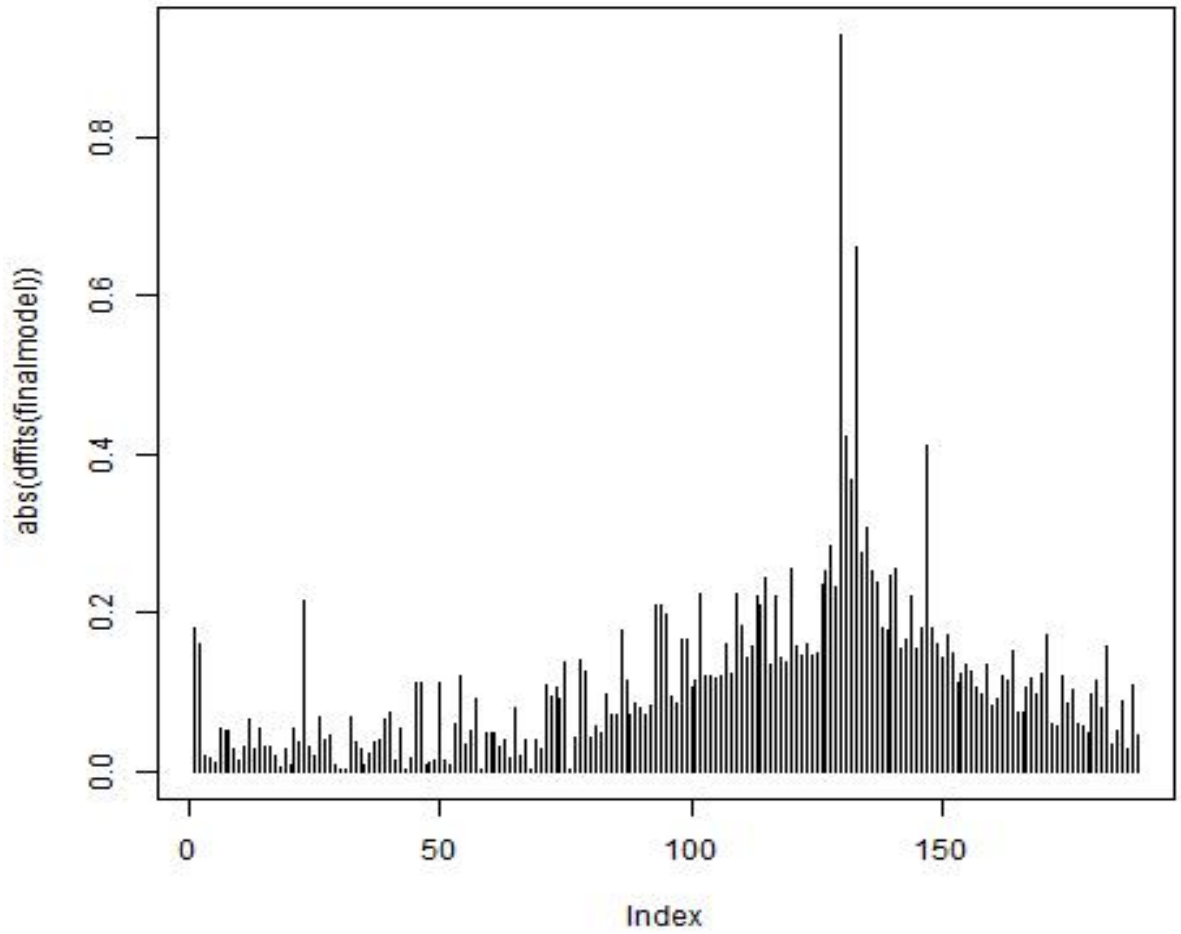
$$Dffts_i = \frac{\hat{y}_i - \hat{y}_i^{-i}}{\sigma_{-i}\sqrt{h_{ii}}} \tag{4}$$

Large $|Dffts_i|$ indicates that observation i has influence on the estimate $\hat{y}_i$.

The observation is considered influential if $|Dffts_i| \geq 2\sqrt{\frac{p+1}{n}}$.

Welsch - Kuh distance

```
threshold.fitt <- 2*sqrt((8+1)/(189))
birth.weight$ID[abs(dffits(finalmodel)) >= threshold.fitt]
```



For Further Reading I

Pierre Lafaye de Micheaux.
The R Software,
Fundamentals of programming and statistical analysis
Springer, 2013.