Forecasting with linear regression

Yannig Goude– yannig.goude@edf.fr

AGENDA

1. FORECASTING: PRINCIPLES AND METHODOLOGY
2. LINEAR REGRESSION
3. APPLICATION TO ELECTRICITY LOAD FORECASTING

FORECASTING: principles and methodology

FORECASTING

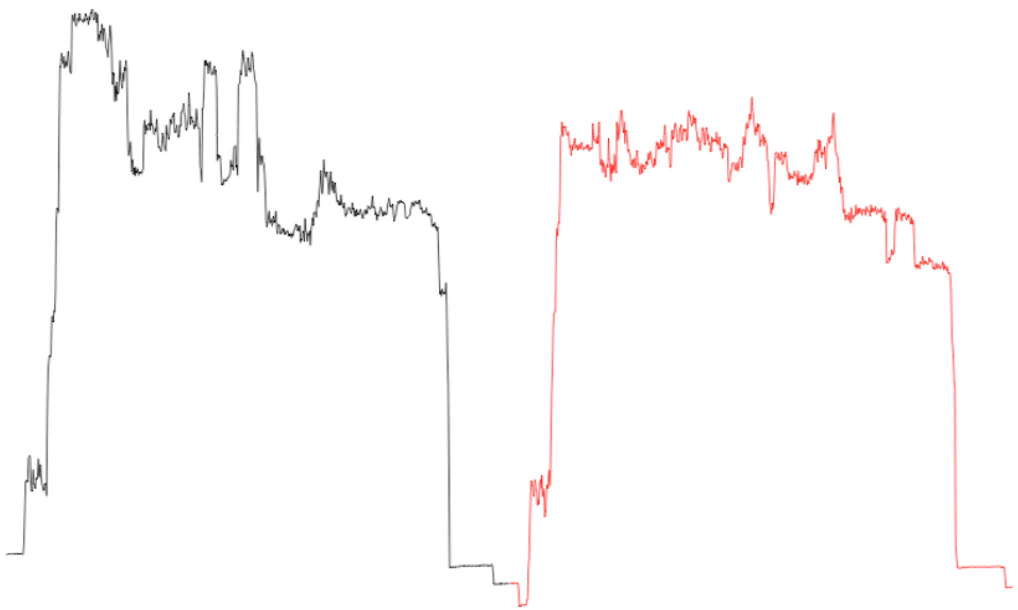
Can take many forms:

- Expert advice
- Prospective (science fiction, anticipation)
- Scenario generation
- What if scenarios
- Physical modeling

We will focus here on forecasting from **data**

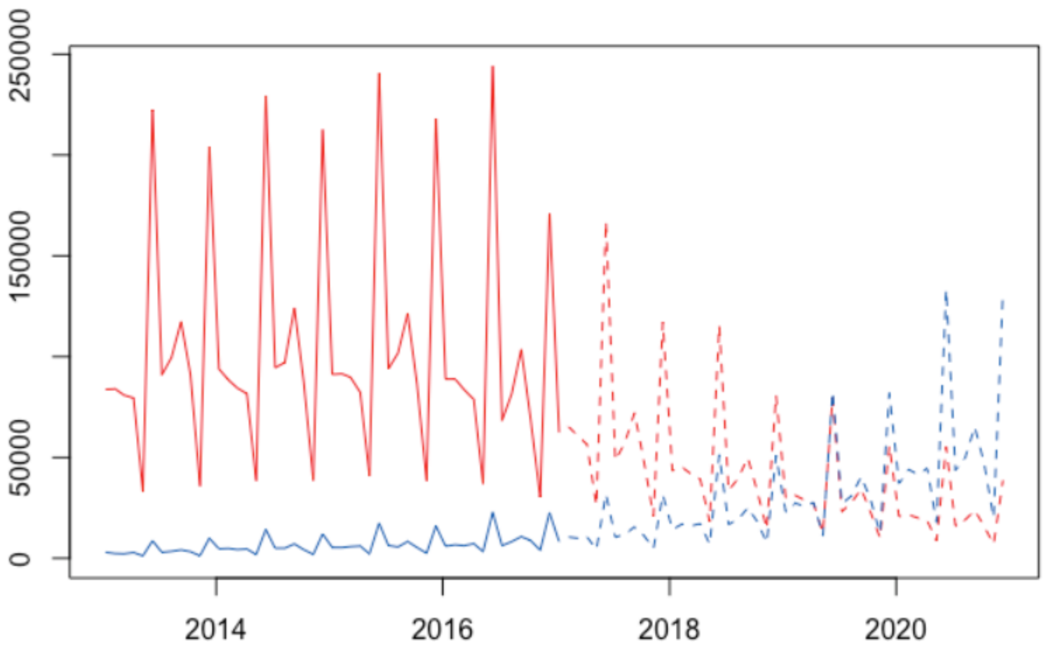
FORECASTING FROM DATA

Forecasting electricity load (industrial consumption at 5 min resolution) at a one day horizon:



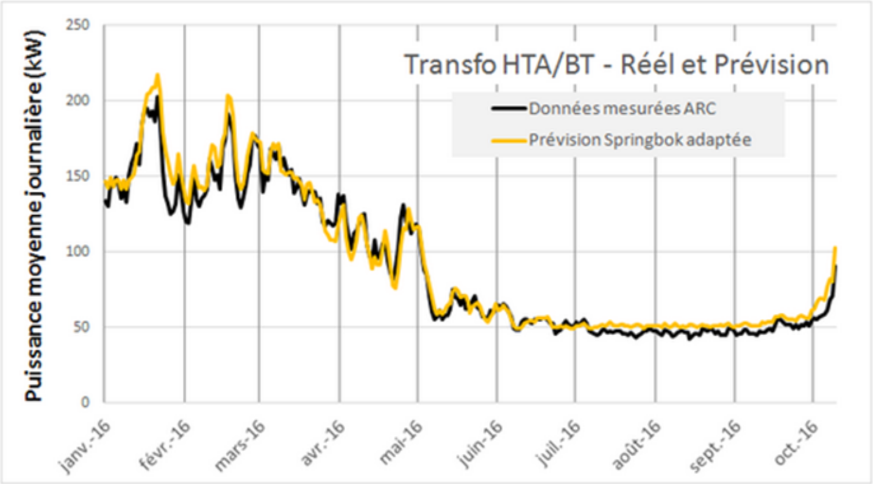
FORECASTING FROM DATA

Forecasting car sales (UK) at a 2 years horizon:



FORECASTING FROM DATA

Forecasting low voltage stations load :



FORECASTING FROM DATA

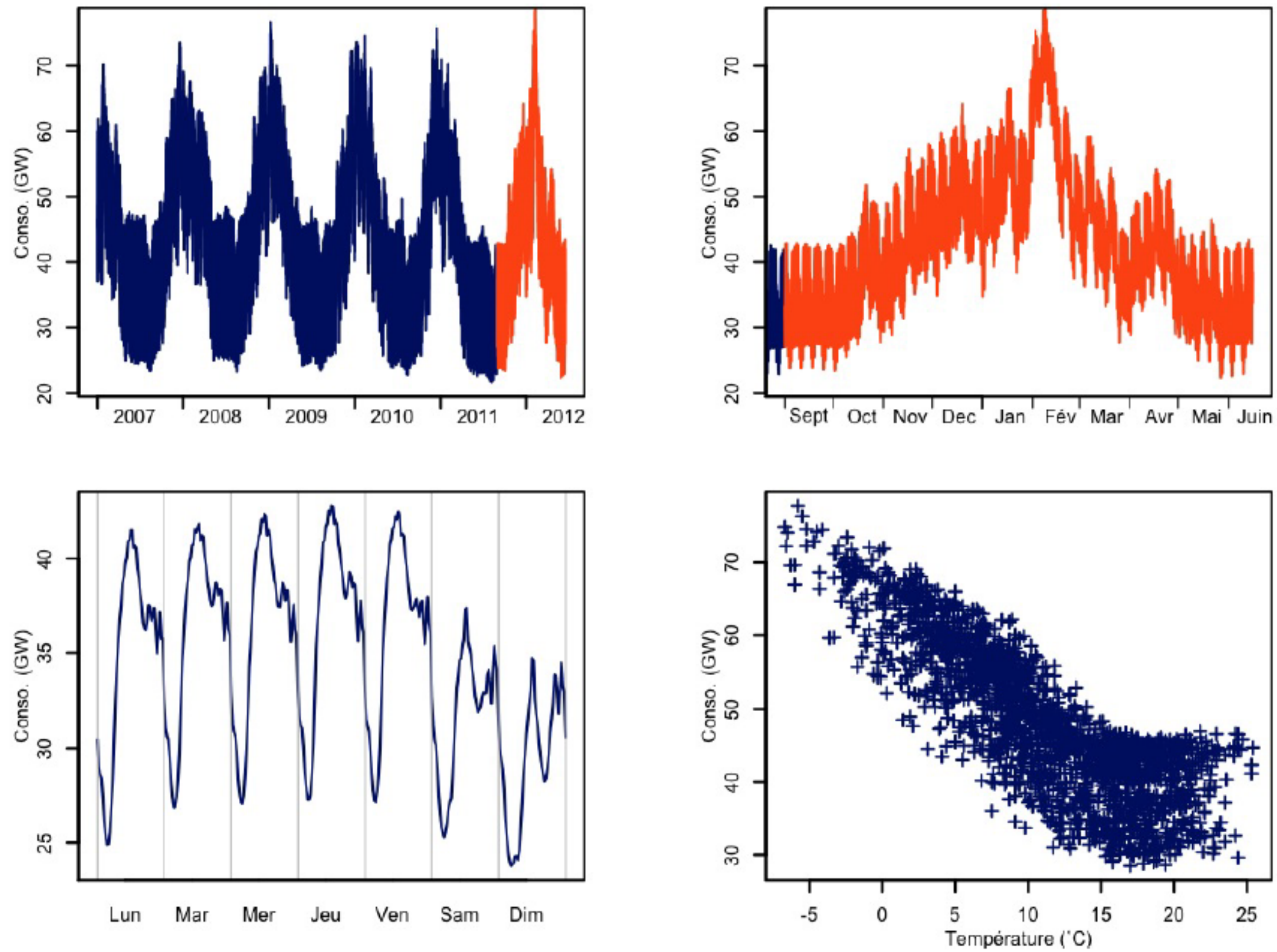
- Characterize statistical properties of the data:
  - Endogenous (time dependancy)
  - Exogenous (dependancy with other covariates)
- Correlation/causality
- Stationnarity
- Mean forecast/ distribution forecast/ quantile forecast

We restrict here to **statistical models** which are a good compromise between quality of the forecast and interpretability of the models.

Some recent development have been done in the field of **AI** methods (e.g. deep learning) but these approach are still **black boxes.**

FORECASTING FROM DATA

Exogenous/endogenous dependance: the exemple of french electricity load

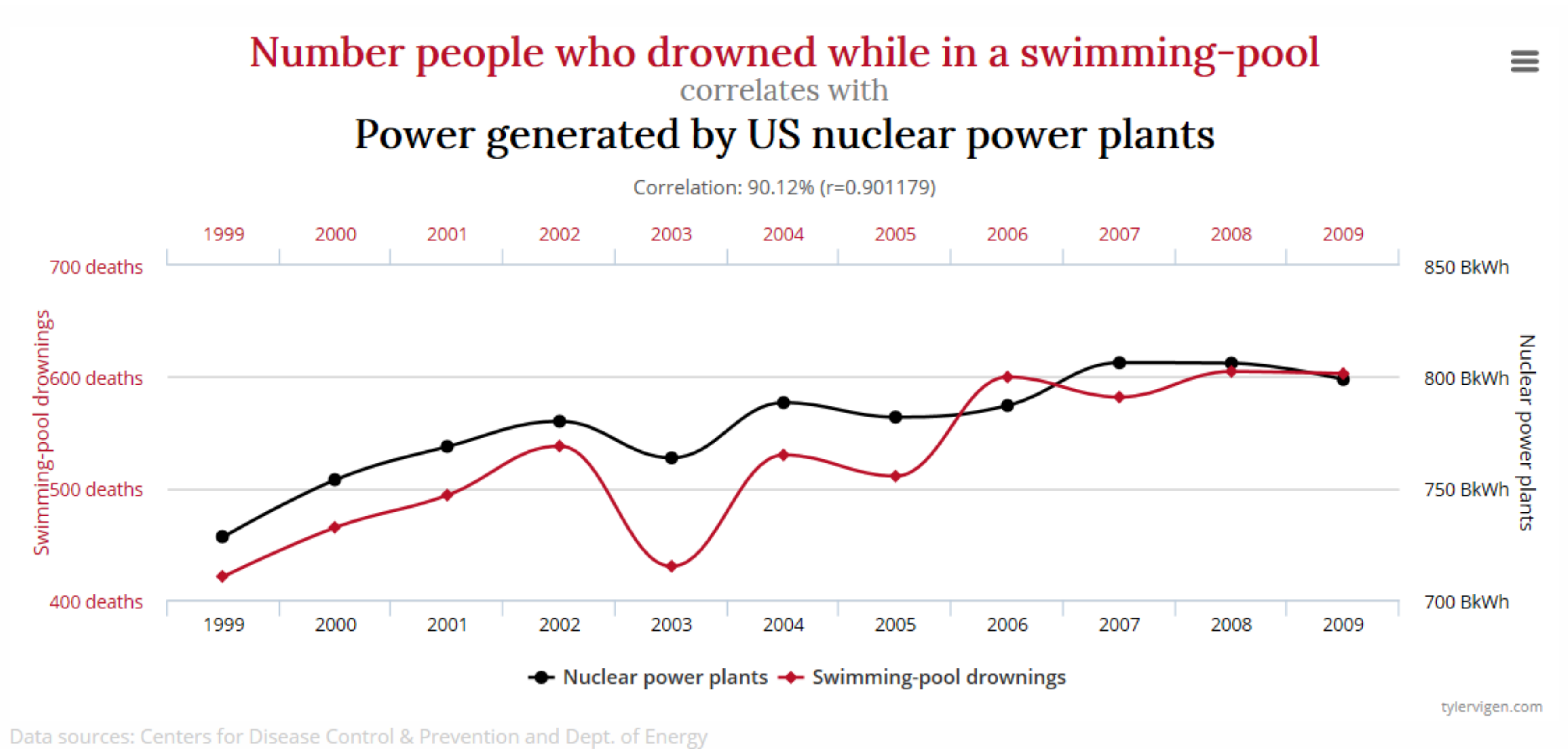


Weekly cycle

Temperature dependancy

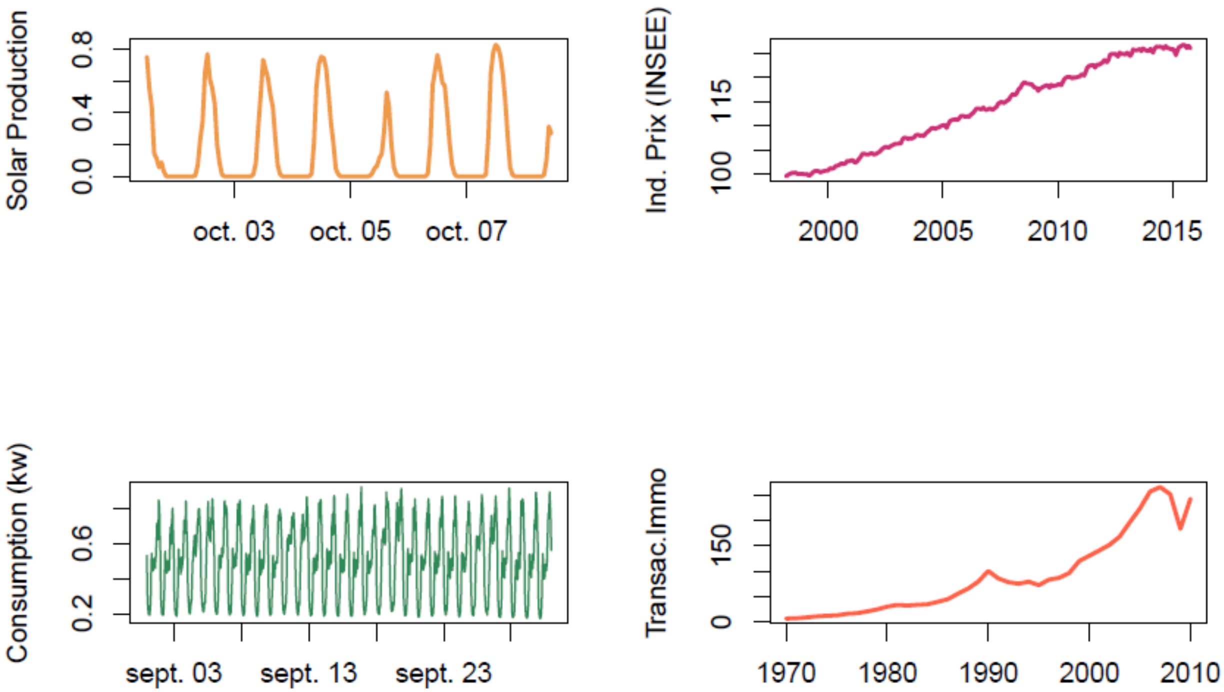
FORECASTING FROM DATA

Correlation/causality



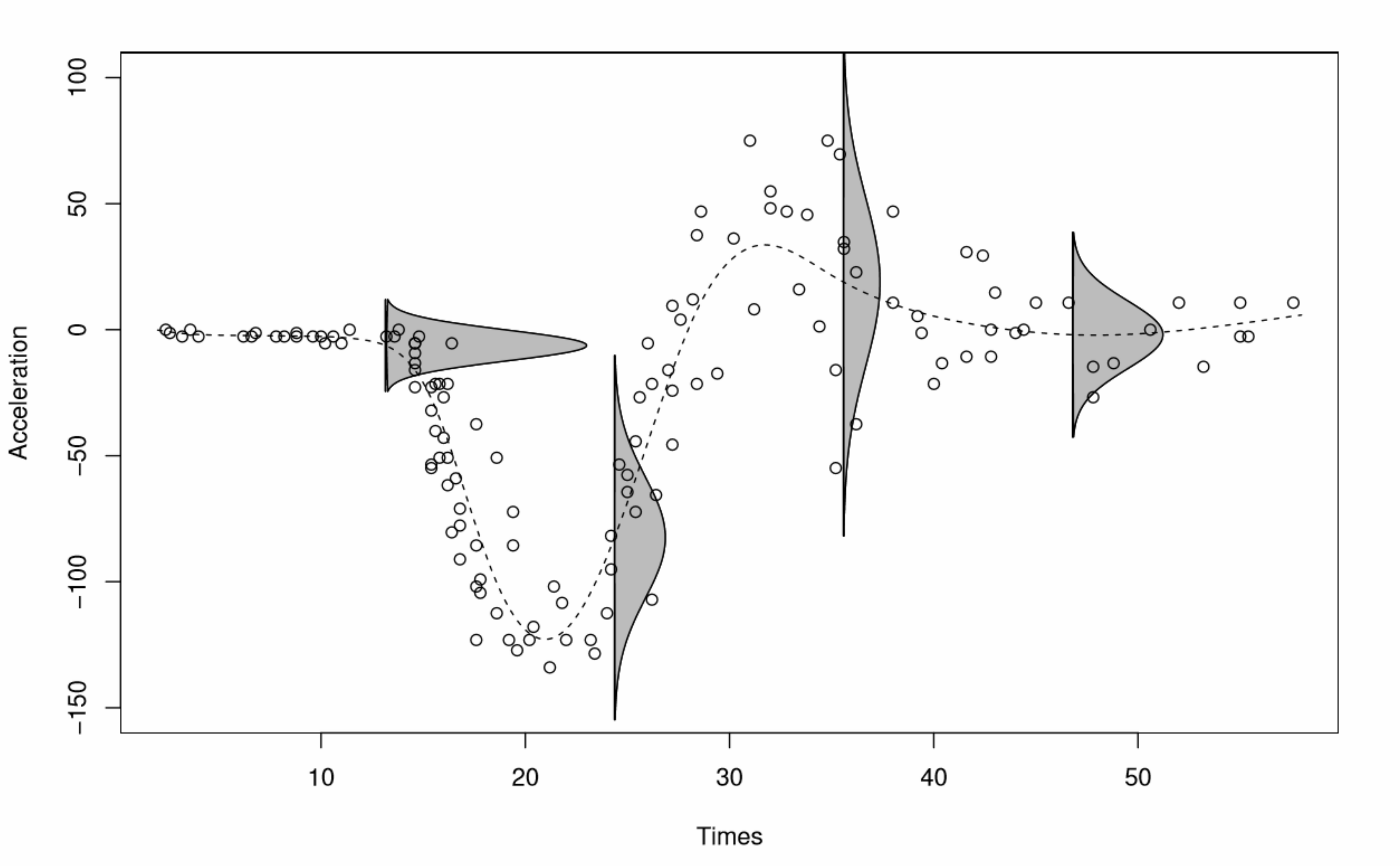
FORECASTING FROM DATA

Stationnarity: is the law of the process stable with time?



FORECASTING FROM DATA

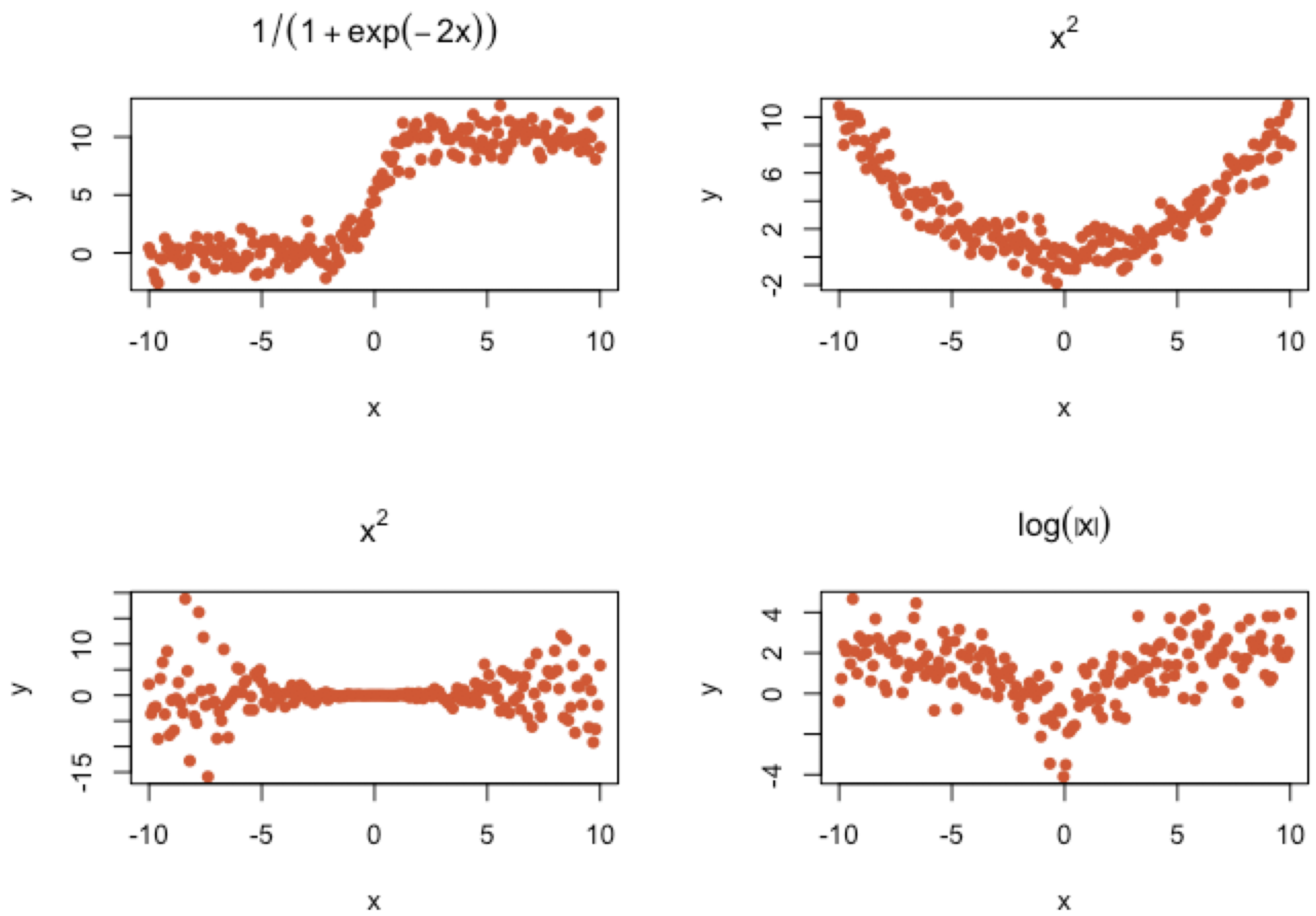
Mean forecast, density forecast, quantile forecast



FORECASTING: descriptive statistics

FORECASTING: DESCRIPTIVE ANALYTICS

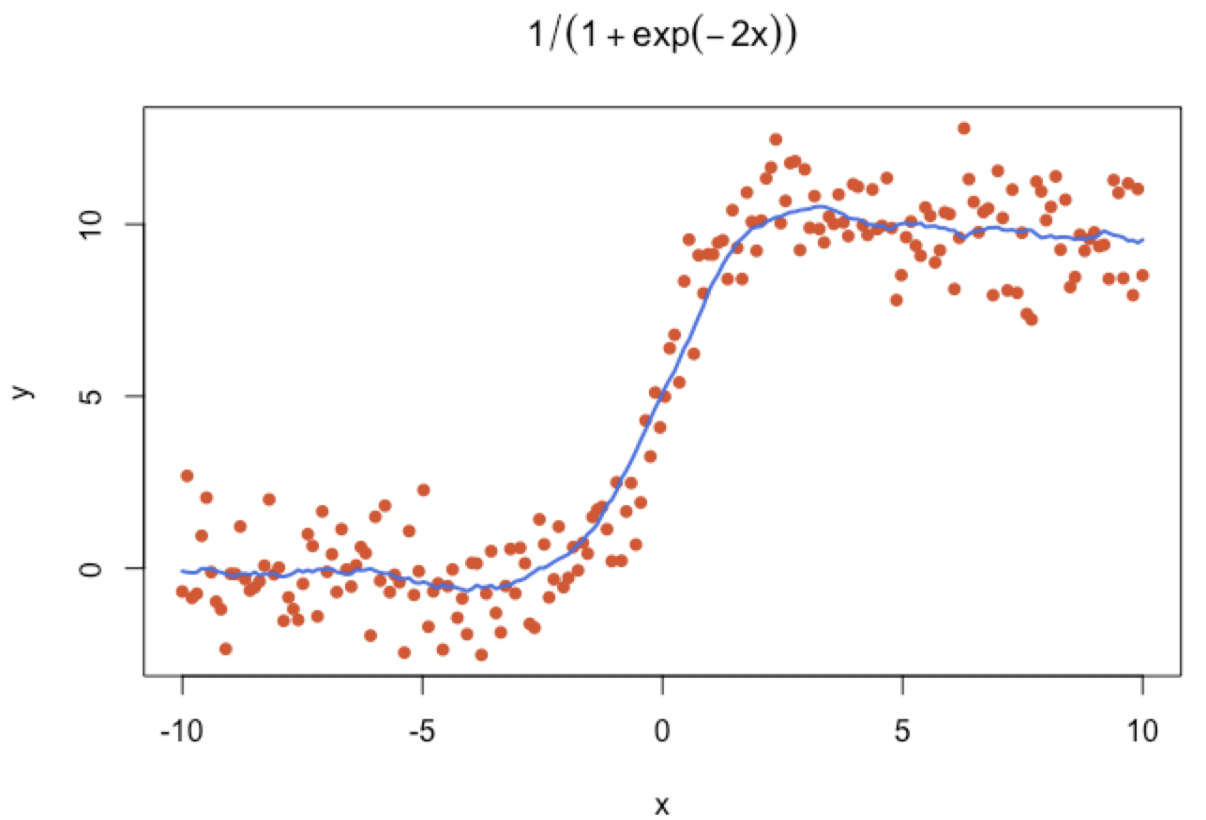
Scatter plots



FORECASTING: DESCRIPTIVE ANALYTICS

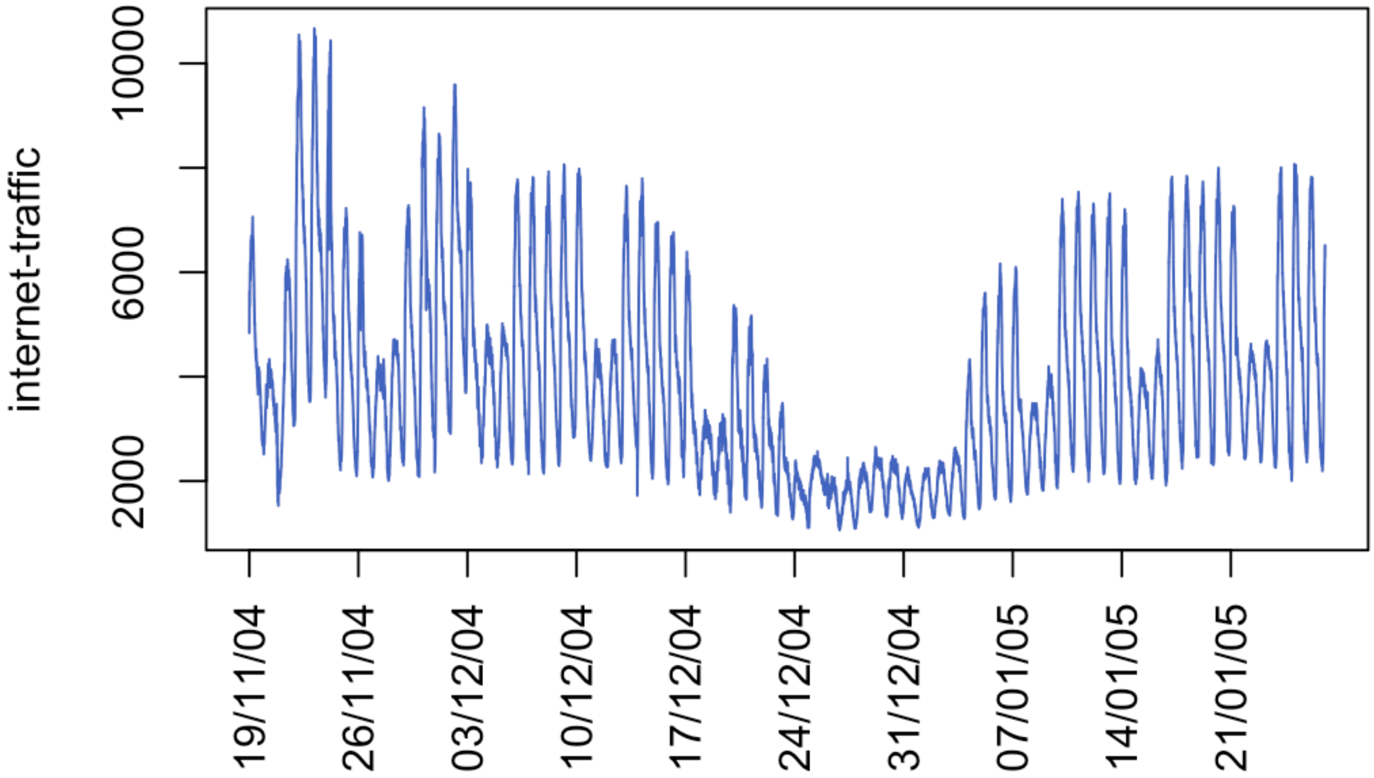
Scatter plots / kernel smoothing

$$\widehat{f}_h(x) = \frac{\sum_{t=1}^{n} y_t K\left(\frac{x-t}{h}\right)}{\sum_{t=1}^{n} K\left(\frac{x-t}{h}\right)}$$



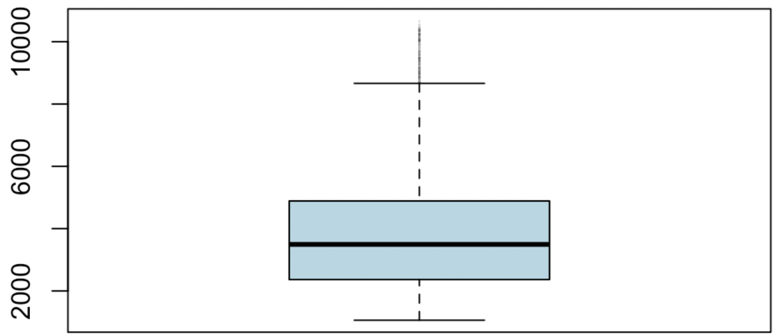
FORECASTING: DESCRIPTIVE ANALYTICS

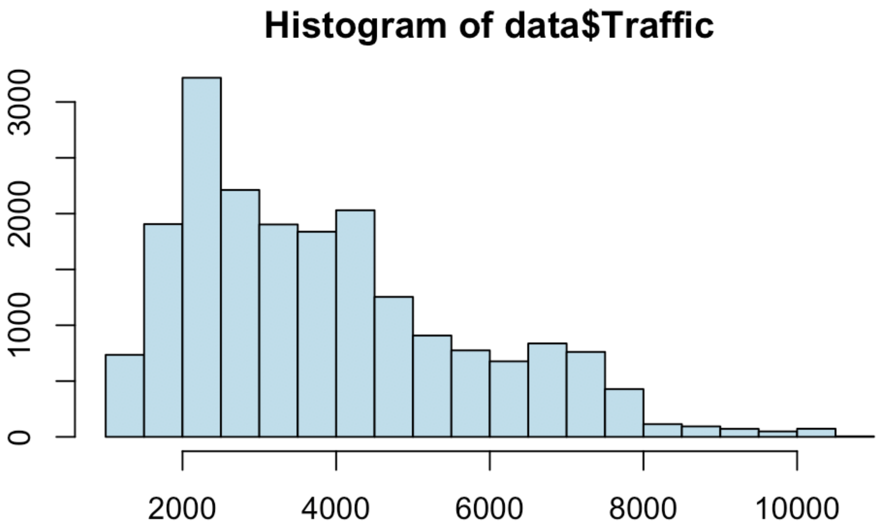
…or time plots



FORECASTING: DESCRIPTIVE ANALYTICS

Histograms/ boxplots





Autocorrelation

More specifically for time series, these statistics are useful:

- the **empirical mean** of a time series $(y_t)_{1 \leq t \leq n}$: $\bar{y}_n = \frac{1}{n} \sum_{t=1}^{n} y_t$
- the empirical standart deviation to estimate its **dispersion** : $\sigma_n = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \bar{y}_n)^2}$
- empirical auto-covariance γ or autocorrelation ρ indicate the temporal (linear) **dependancy**:

$$\gamma_n(h) = \frac{1}{n} \sum_{t=1}^{n-h} (y_t - \bar{y}_n)(y_{t+h} - \bar{y}_n)$$

$$\rho_n(h) = \frac{\gamma_n(h)}{\gamma(0)}$$

remark that $\gamma_n(0) = \sigma_n^2$.

thus $\rho_n(h)$ is the estimate of the correlation between y_t and y_{t+h}, supposing this correlation exists and is stable in function of time (stationnarity)

Partial Autocorrelation

- partial autocorrelation (PACF) : dependancy between two instant t and $t+k$ conditionnally to what happened at times $t+1,...,t+k-1$

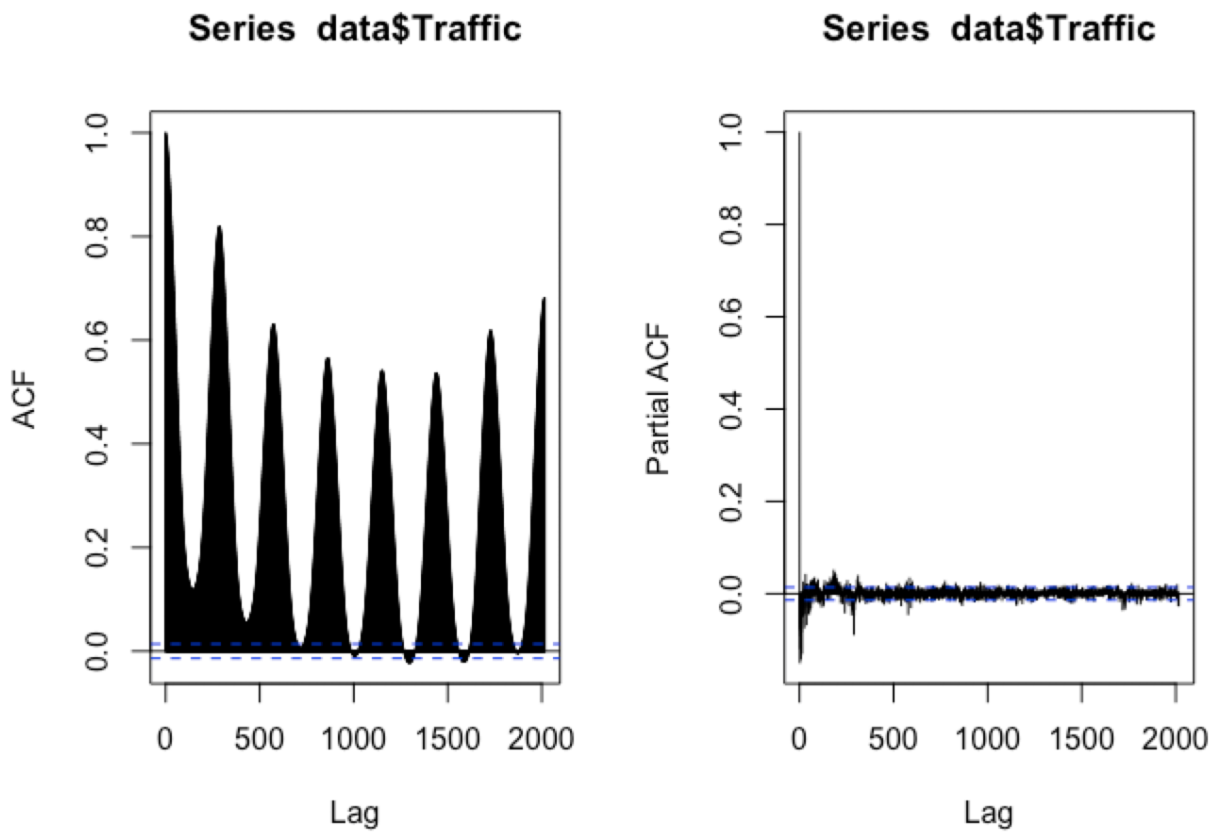
To obtain order h PACF one have to consider the following linear model:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_h y_{t-h} + \varepsilon_t$$

the order h PACF is defined as α_h and can be estimated solving the OLS problem.

FORECASTING: DESCRIPTIVE ANALYTICS

ACF and PACF



FORECASTING: linear model

THE LINEAR REGRESSION MODEL

We consider a target Y which is a real random variable that we want to forecast according to:

- paste value of Y
- and/or covariates X_1, ..., X_p

We assume that the data are generated according to the following (linear regression) model:

$$y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,p}\beta_p + \varepsilon_i$$

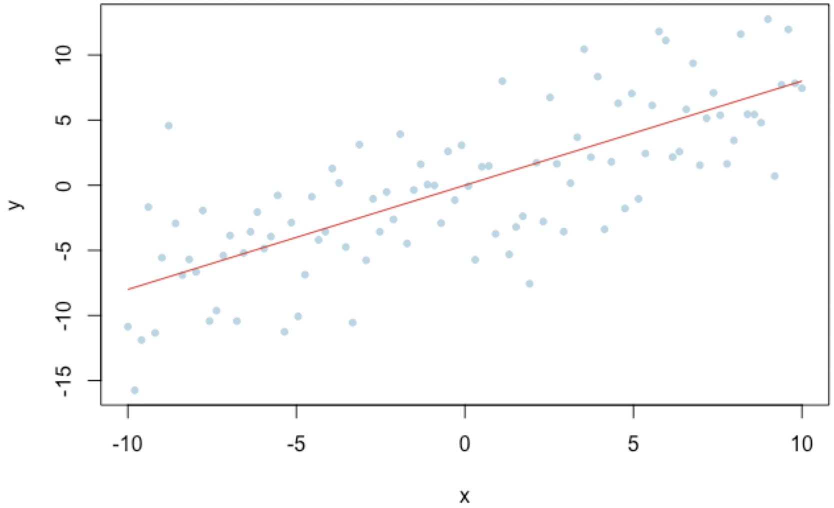
for $i = 1, \ldots, n$ observations. Our objective is to estimate the unknown parameters $\beta = (\beta_1, \ldots, \beta_p)$

We will also suppose:

- ε_i are independant and indentically distributed (iid)
- mean 0 and constant variance: $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$

THE LINEAR REGRESSION MODEL

This can be rewritten with matrix notations:



$$Y = X\beta + \varepsilon$$

where

- X has n rows and p columns and is of rank p
- $Y \in \mathcal{R}^p$, $\beta \in \mathcal{R}^p$
- $E(\varepsilon) = 0$, $V(\varepsilon_i) = \sigma^2$

suppose that we measure our performance with the quadratic loss, we can solve the well known ordinary least square problem to infer β from the data:

$$\min_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} (y_i - x_i\beta)^2 = \min_{\beta \in \mathcal{R}^p} ||Y - X\beta||^2$$

This can be seen as a convex optimisation problem where we minimise the fonction $g : \beta \to \sum(y_i - x_i\beta)^2$.

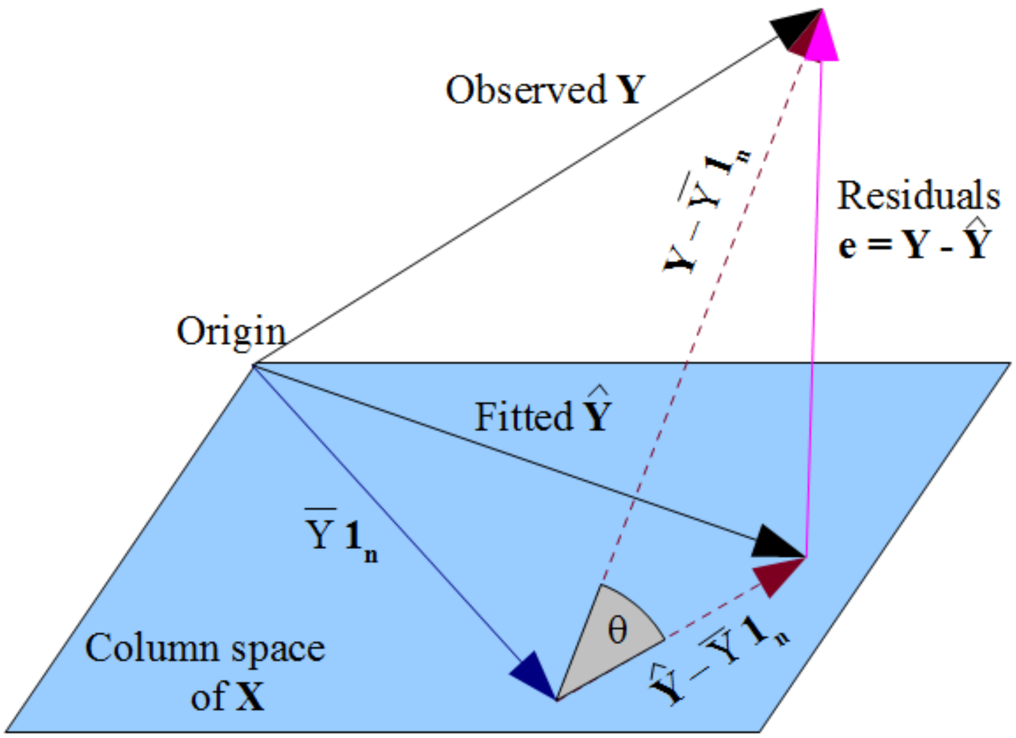
So, $\widehat{\beta}$ satisfying $\frac{dg}{d\beta}(\widehat{\beta}) = 0$, ie:

$$\frac{dg}{d\beta} = -2X^TY + 2X^TX\beta = 2X^T(X\beta - Y)$$

thus, $X\widehat{\beta} = Y$ and $X^TX\widehat{\beta} = X^TY$ and, as X is rank p:

$$\widehat{\beta} = (X^TX)^{-1}X^TY$$

Geometric interpretation



THE LINEAR REGRESSION MODEL

Useful statistics:

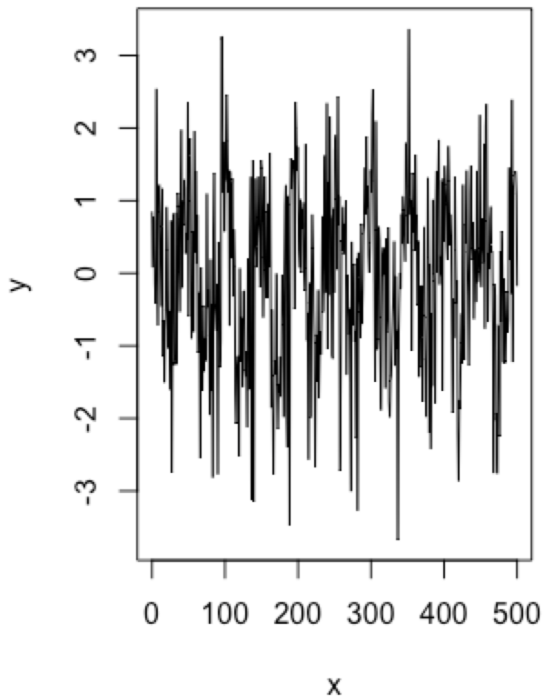
- $R^2 = \cos^2(\theta) = \frac{||\widehat{Y} - \bar{Y}||^2}{||Y - \bar{Y}||^2} = 1 - \frac{||Y - \widehat{Y}||^2}{||Y - \bar{Y}||^2}$ the proportion of variance explained by our model. It indicates wether the regression is close to the observations (including the noise variance).
- don't work if we compare different nature of models (e.g. multiplicative vs additive)
- can induce overfitting as R^2 increases as p increases, there is an adjusted version to take dimension p into account: $R_a^2 = 1 - \frac{n}{n-p} \frac{||Y - \widehat{Y}||^2}{||Y - \bar{Y}||^2}$

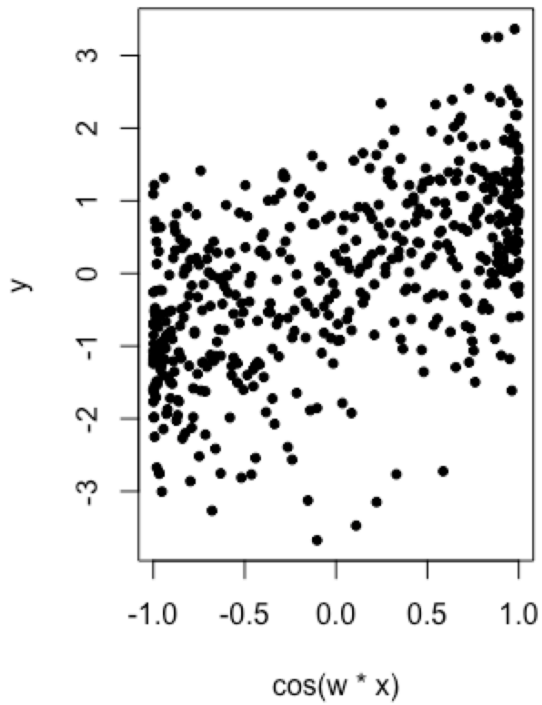
FEATURE ENGENEERING

Linear models are a powerful tool as, conditional to the good transformations of the data $X \rightarrow Xnew$, we can often express Y as a linear combination of $Xnew$. Here are a few examples.

- periodic data: Fourier basis regression

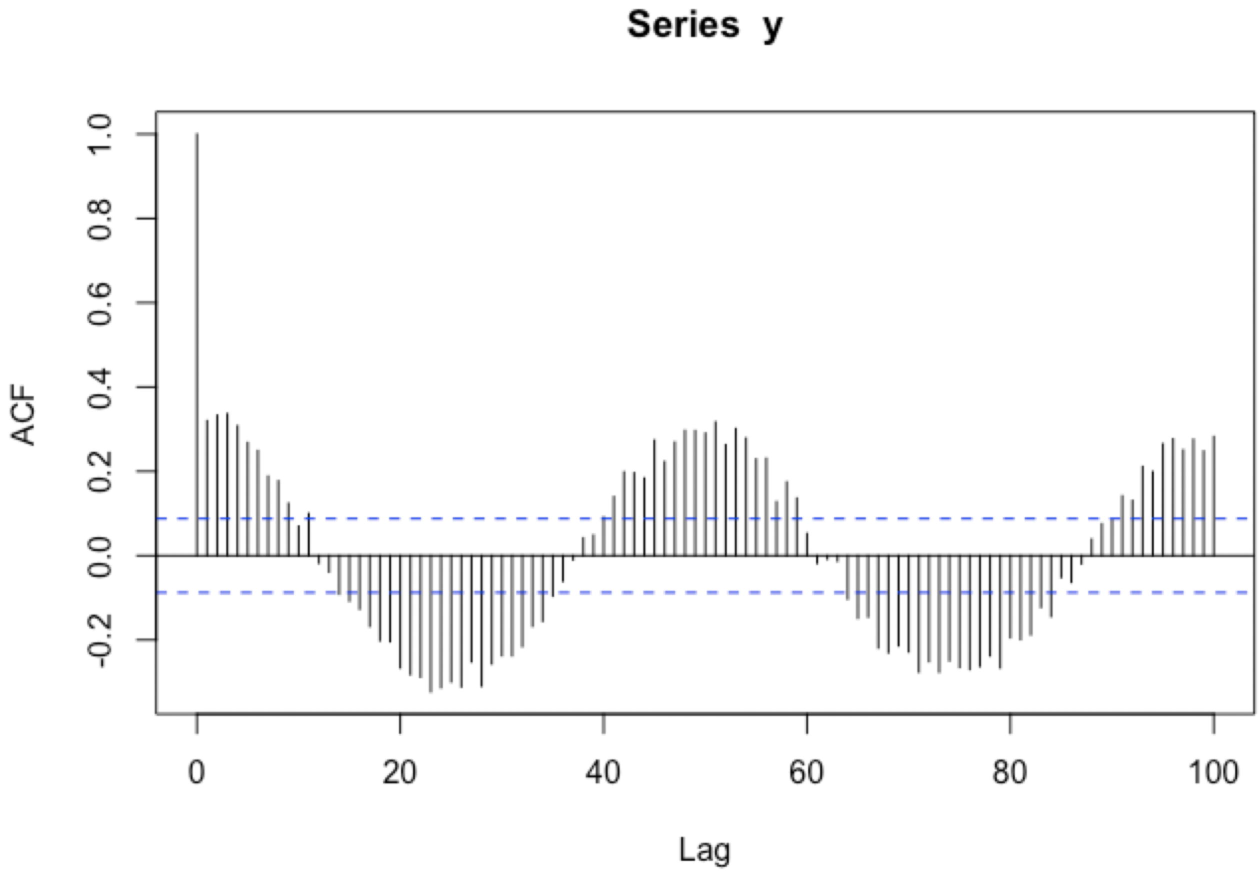
For a well chosen $\omega = 2 * \pi / T$ and an harmonic k: $Xnew = \cos(k * \omega x)$





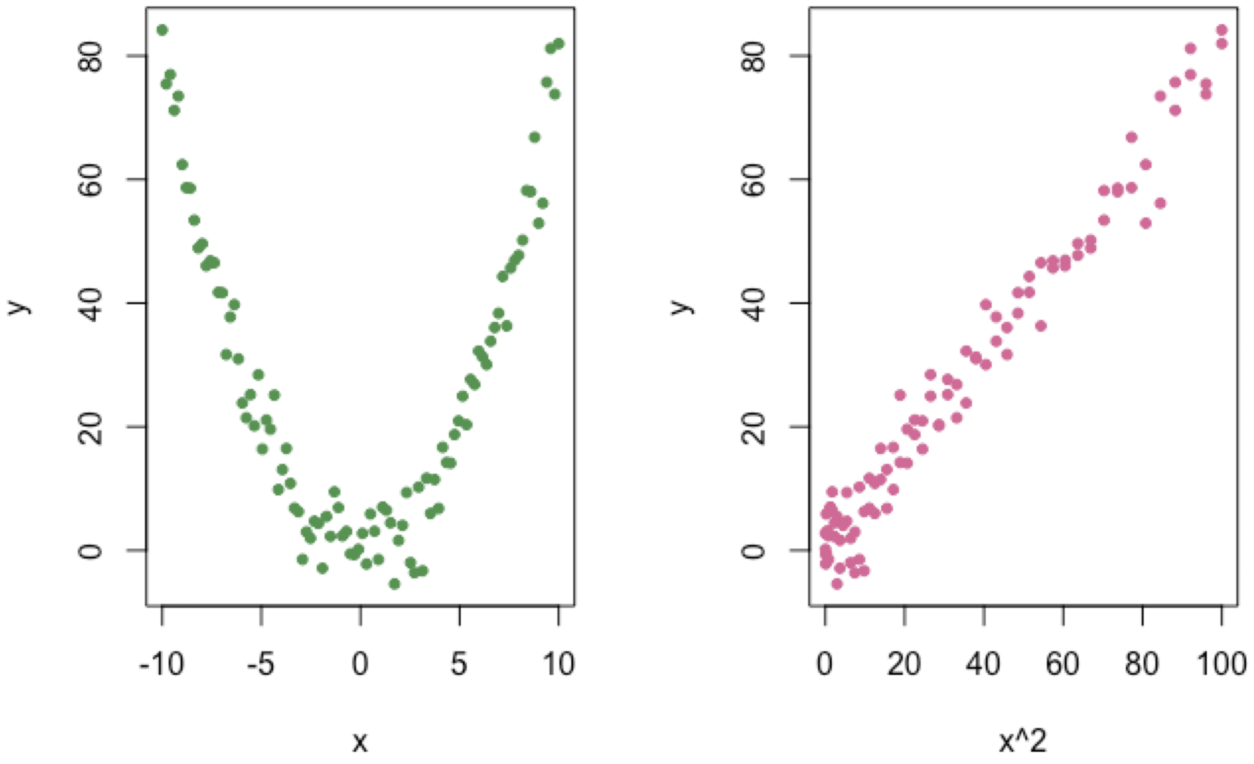
FEATURE ENGENEERING

ω can be chosen according to a frequency analysis of the signal:

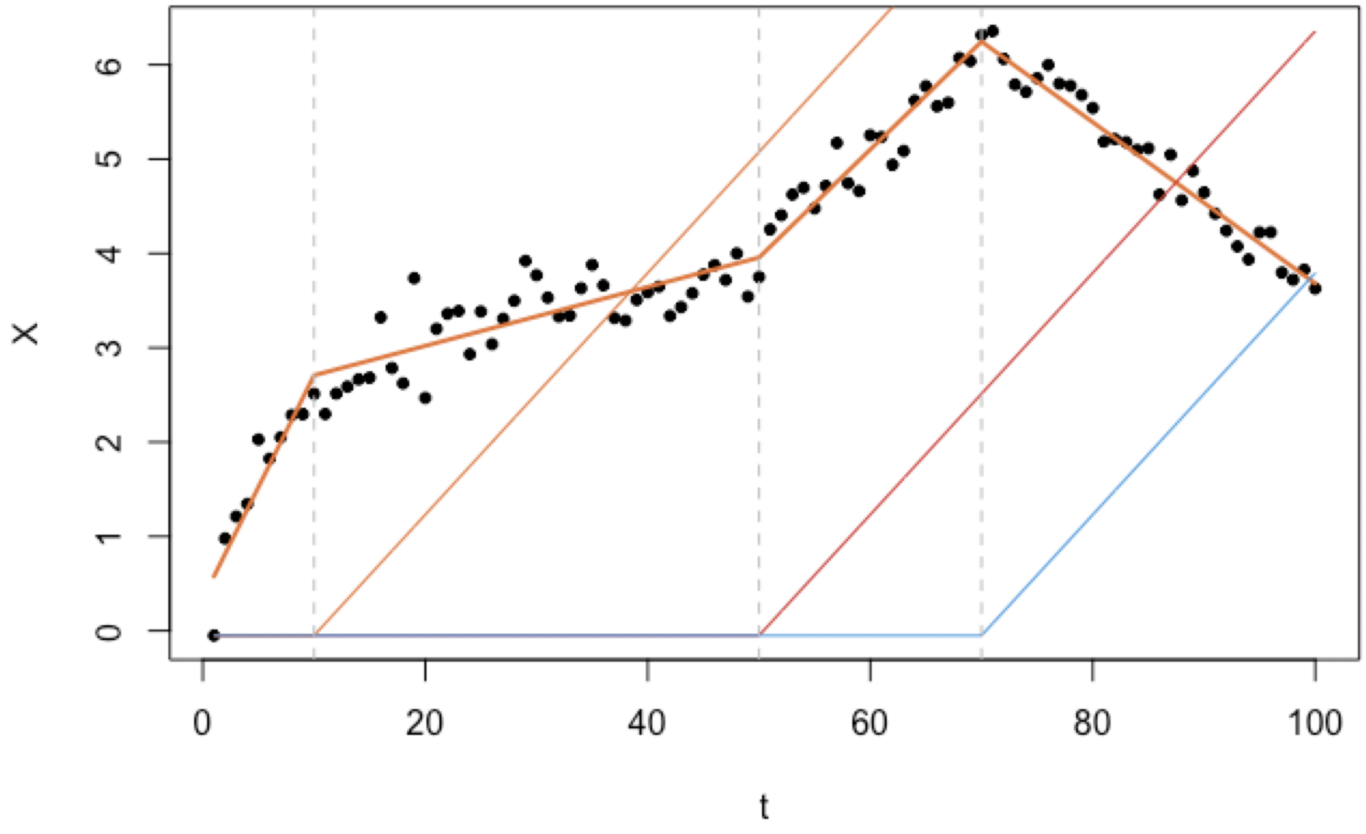


FEATURE ENGENEERING

- polynomial transformation:



- spline basis decomposition: e.g. truncated power functions $1, x, (x-k)_+$



MODEL VALIDATION

The forecaster needs an objective criteria to:

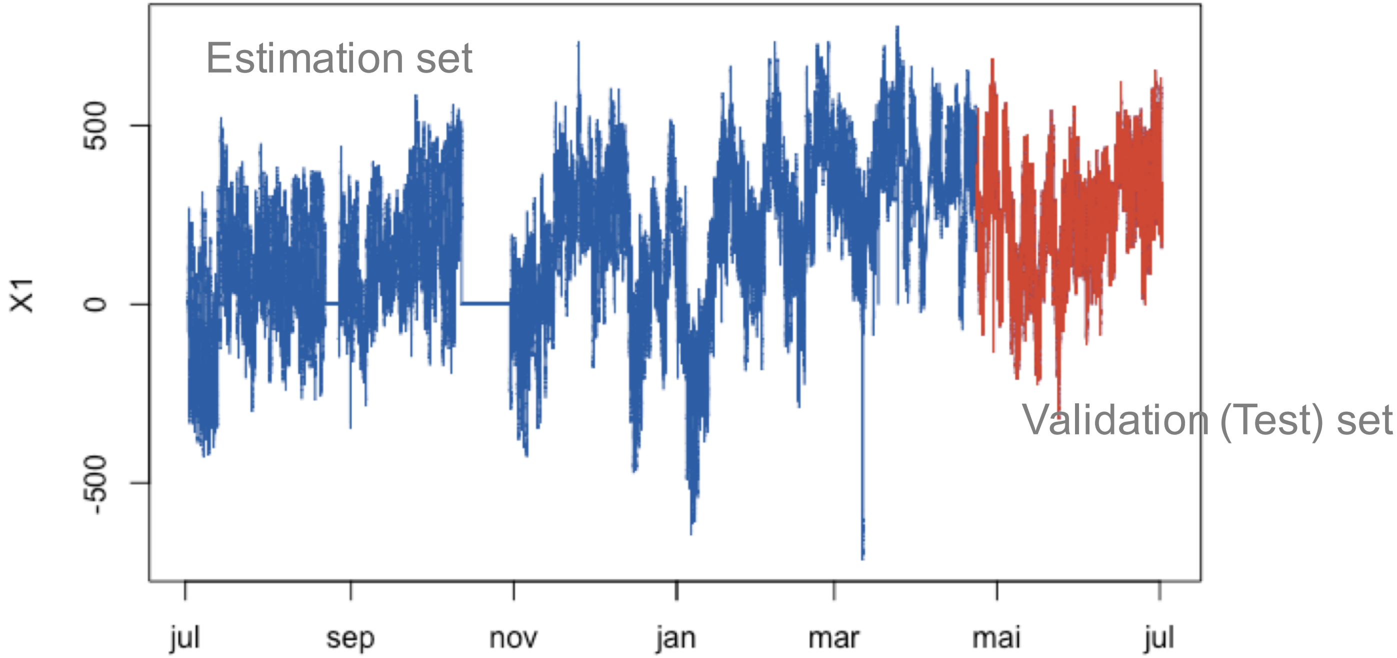
- Select the set of covariates to include into the model
- Find the good transformation of the covariates
- Calibrate the model
- Have a good estimate of its forecasting performances

Many pitfalls:

- Overfitting
- Extrapolation problem (trends)
- Most of the time data are not iid

MODEL VALIDATION

Test set



Work only if:

- the data have the same generation process in the 2 sets
- we have enough data to split it

Try to choose the test set in accordance with your final purpose/ the characteristic of the data

Cross validation

leave one out:

- choose randomly $i \in \{1, \ldots, n\}$
- fit your model on all the data except i ie $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)$, denotes this model $\widehat{\phi}_{-i}$
- estimate a forecast error $(y_i - \widehat{\phi}_{-i}(X_i))^2$
- repeat that N times and compute an estimate of the forecast error of your model ϕ

$$R_N(\phi) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{\phi}_{-i}(X_i))^2$$

MODEL VALIDATION

Cross validation

K-fold:

For $k \in \{1, \ldots, K\}$,

- choose randomly $I_k = (i_1, \ldots, i_Q) \in \{1, \ldots, n\}^Q$, where $K * Q = n$
- fit your model on all the data except I_k, denotes this model $\widehat{\phi}_{-I_k}$
- estimate a forecast error $R_{I_k} = \frac{1}{Q} \sum_{k=1}^{Q} (y_{i_k} - \widehat{\phi}_{-I_k}(x_{i_k}))^2$

then compute an estimate of the forecast error of your model ϕ

$$R_K(\phi) = \frac{1}{K} \sum_{i=1}^{K} R_{I_k}$$

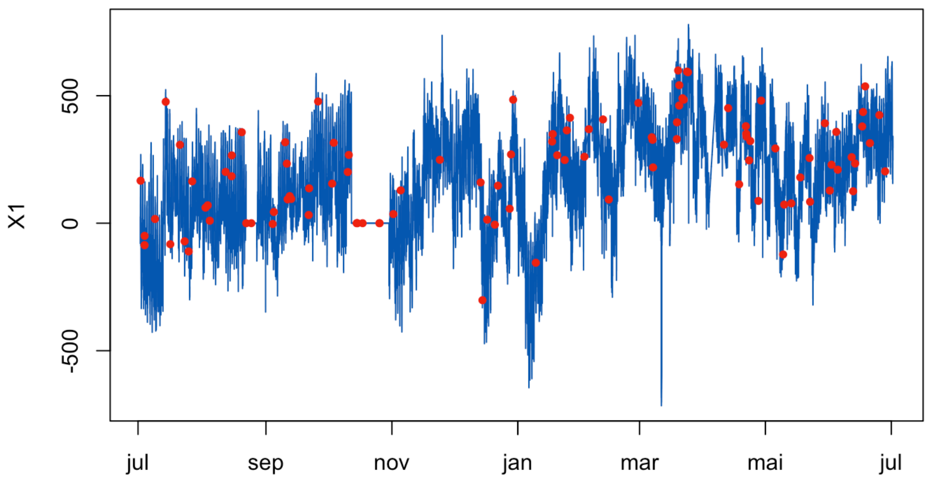
Remarks:

- I_k is here randomly sample but it could be blocks of consecutives observations (blockwise K-fold CV) so that I_1,..., I_K is a partition of $1, \ldots, n$. This is particularly relevant in the time series context.

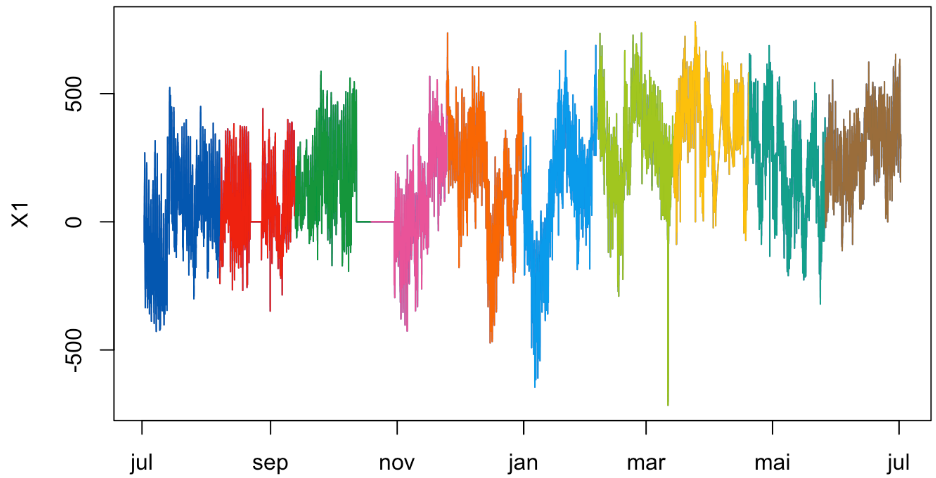
- in addition to the error $R_K(\phi)$, if K is sufficiently large, we can also compute a measure of the uncertainty of this estimate (variance, quantile…)

MODEL VALIDATION

Cross validation



Blockwise10-folds CV



MODEL VALIDATION

Sequential testing

For $t \in \{n_0, \ldots, n\}$:

- fit a model $\widehat{\phi}_t$ on the data $(x_1, y_1), \ldots, (x_t, y_t)$
- forecast y_{t+1} as $\widehat{\phi}_t(x_{t+1})$

then compute an estimate of the forecast error of your model ϕ

$$R_{n_0}(\phi) = \frac{1}{n - n_0} \sum_{t=n_0+1}^{n} (y_t - \widehat{\phi}_t(x_t))^2$$

MODEL VALIDATION

Well chosen blockwise test set

Yearly seasonnal time series

MODEL VALIDATION

For linear model we have this convenient property:

$$CV = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{y}_i)^2}{(1 - H_{i,i})^2} \qquad \widehat{\varepsilon_i^{-i}} = \widehat{\varepsilon}_i/(1 - H_{i,i})$$

Preuve:

- lemme d'inversion matriciel:
  Soit M une matrice symétrique inversible $p \times p$, u et v deux vecteurs de taille p. Alors:

$$(M + uv')^{-1} = M^{-1} - \frac{M^{-1}uv'M^{-1}}{1 + u'M^{-1}v}$$

- $X'X = X'_{-i}X_{-i} + x_i x'_i$
- $X'Y = X'_{-i}Y_{-i} + x_i y_i$
- $h_{i,i} = x'_i(X'X)^{-1}x_i$

MODEL VALIDATION

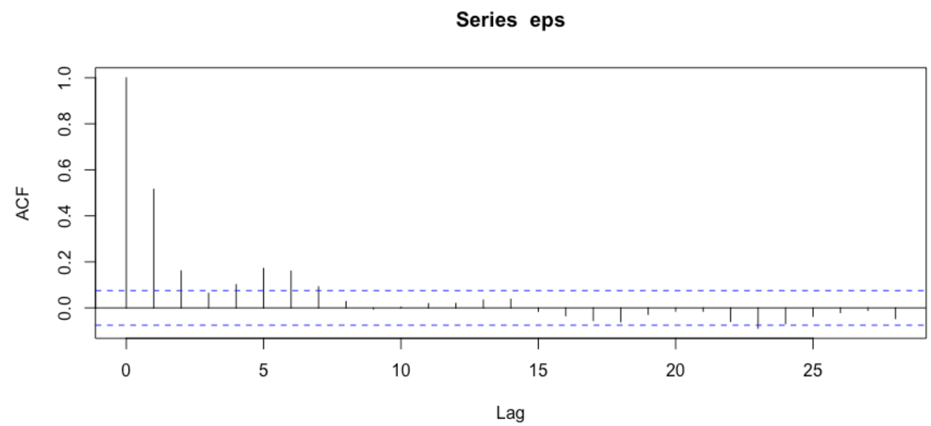
Résiduals checks

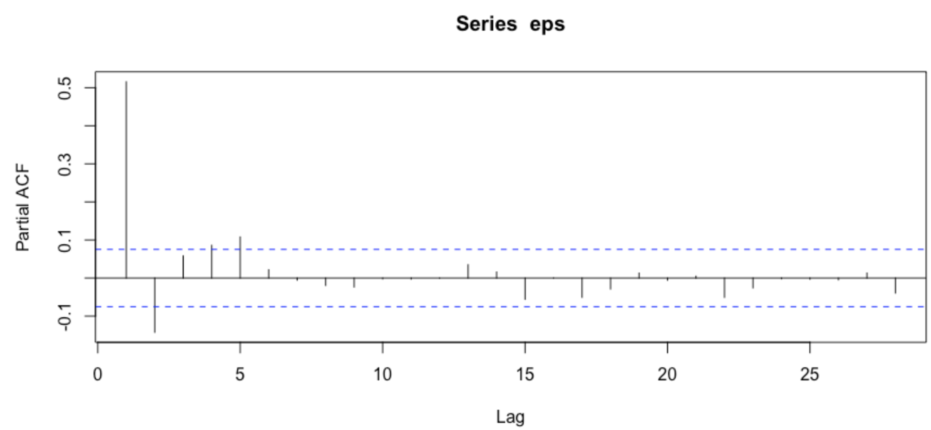
- Independance: acf, pacf
- Test de Box-Pierce

$$H_0(h) : \rho_\varepsilon(1) = \rho_\varepsilon(2) = ... = \rho_\varepsilon(h) = 0$$

$$H_1(h) : \exists k \in (1, ..., h) \text{ t.q } \rho_\varepsilon(k) \neq 0$$

$$Q_{BP}(h) = n \sum_{j=1}^{h} \widehat{\rho}_\varepsilon(j)^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2(h-k)$$

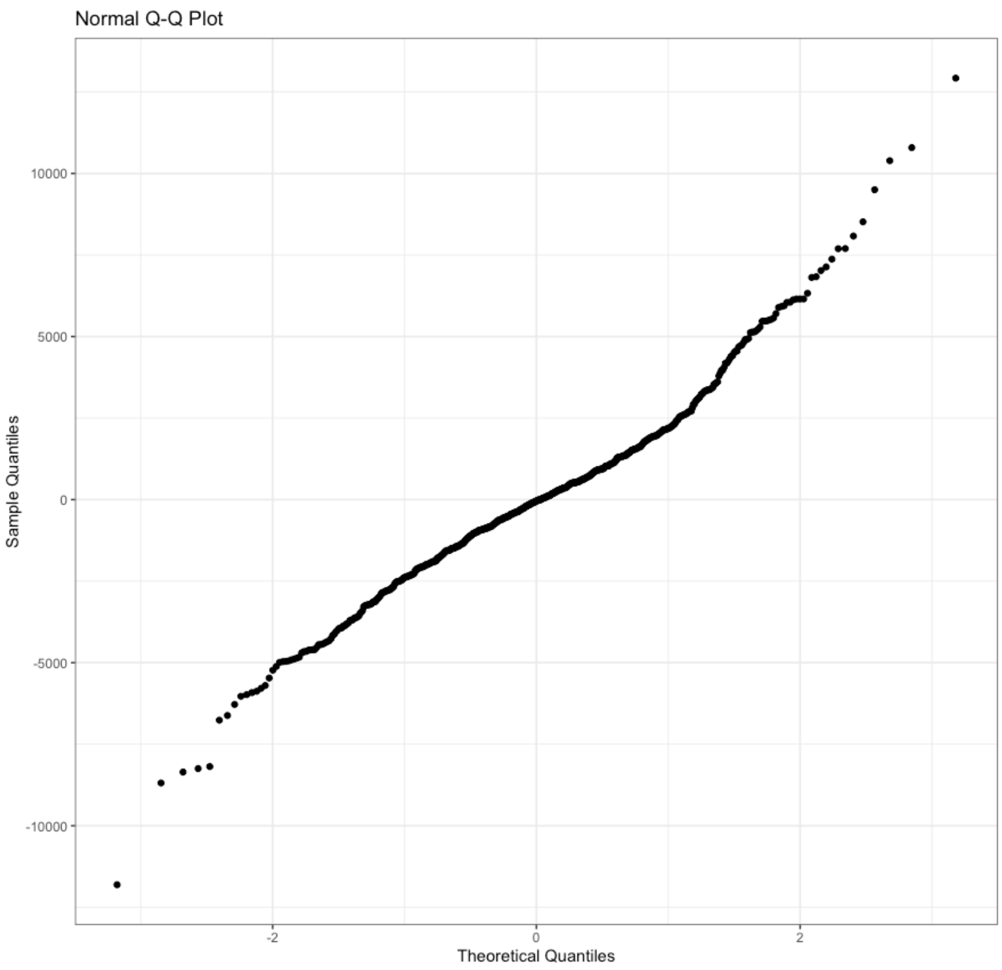




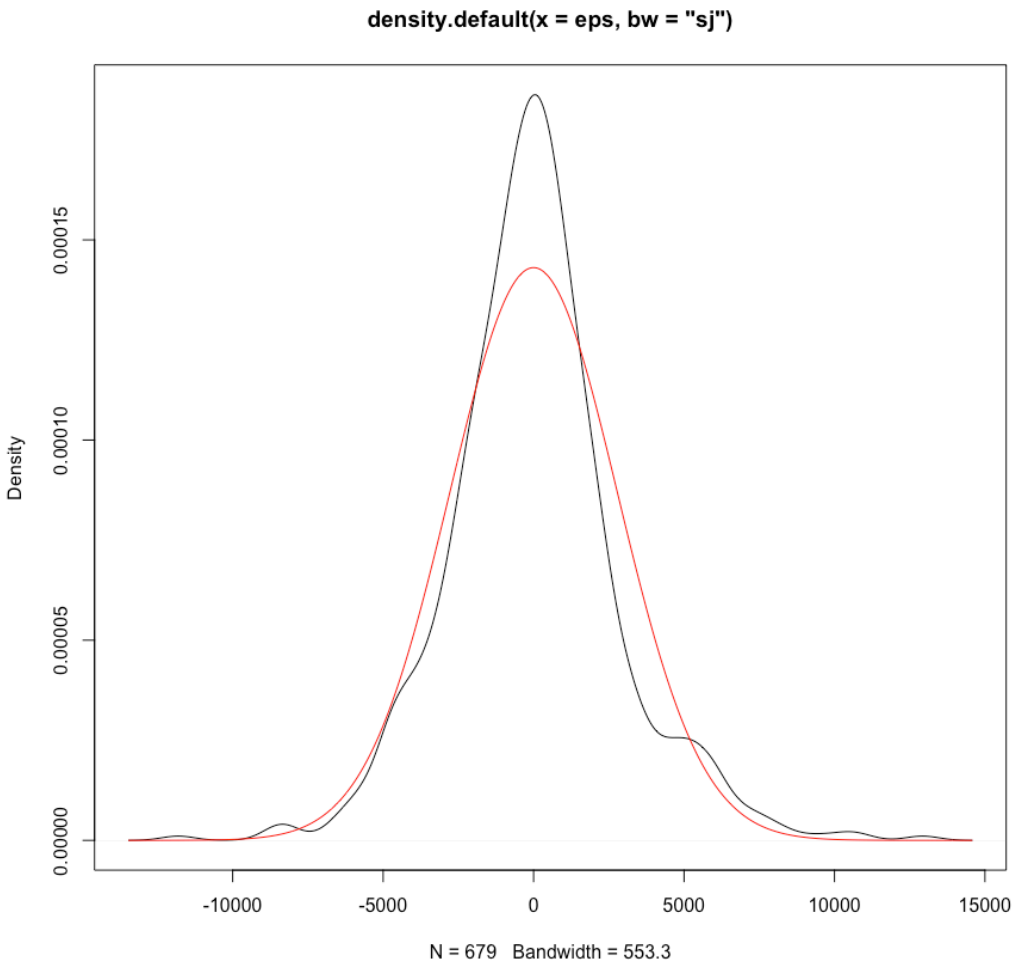
MODEL VALIDATION

Adequation to a given distribution:

- Qqplot
- Density estimation
- Tests: chi2, kolmogorov-Smirnov



MODEL VALIDATION



MODEL VALIDATION

Now, linear modeling of french electricity consumption….