Time Series and Forecasting

Yarema Okhrin
University of Augsburg

Content of the course

- Basics of forecasting and time series analysis
- Forecasting using regression
    - Forecasting cross-sectional data
    - Regression for time series data
    - Forecasting using spline/trigonometric regression
- Time series decomposition
- Exponential smoothing (EWMA, Holt, Holt-Winters, Croston)
- SARIMA modelling
- Special topics in time series and forecasting
    - ARCH/GARCH: models with conditional volatility
    - Outliers in time series
    - Structural breaks in time series
    - Multivariate time series
    - State space models and Kalman filtering
    - Panel data

Useful literature

- John E. Hanke, Dean W. Wichern, 2009, Business Forecasting, Pearson
- Spyros Makridakis, Steven C. Wheelwright, Rob. J. Hyndman, 1998, Forecasting: methods and applications, Wiley
- Max Kuhn, Kjell Johnson, 2013, Applied predictive modeling, Springer
- Philip Hans Frances, 2014, Dick van Dijk and Anne Opschoor, Time Series Models for Business and Economic Forecasting, Cambdridge
- James Hamilton, 1994, Time Series Analysis, Princeton
- ...

Part 1

Objectives, problems and strategies

Objectives

- Forecasts are statements about future unknown quantity of interest. To make these statements we use some available and relevant historical information.
- Forecast provide us with important insights for decision making.
  - *scheduling*
  - *acquiring resources*
  - *determining resources requirements*
  - finance, production, humane resource, sales, marketing, general management, etc.)
- Companies wish to reduce random factors and use more and more complex tools for forecasting.
- Forecasts are always erroneous. The potential deviations - forecast errors - should be analysed carefully.

Categories of forecasting methods

- Quantitative: sufficient information is available
  - **Time series**: predicting the continuation of historical patterns such as the growth in sales or GNP
  - **Explanatory**: understanding how explanatory variables such as prices or ad campaigns affect sales
- Qualitative: little/no quantitative data, but sufficient knowledge
  - Predicting the internet traffic/speed in 2030.
  - Forecasting how a large increase of oil prices will affect economies
- Unpredictable: little or no information is available
  - Predicting the effects of interplanetary travel
  - Predicting the discovery of new forms of energy

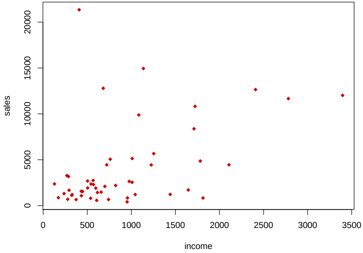
Data and modelling steps I

- Data
  - GIGO principle: *garbage in, garbage out*
  - The data should be reliable and accurate.
  - The data should be relevant.
  - The data should be consistent.
  - The data should be collected for a relevant and correct time period.
- Modeling and evaluation of the model
- Computing the forecasts
- Evaluation of the forecasts

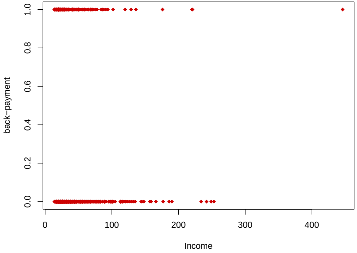
Types of data

- cross sectional data: collected at the same time point or the exact time is irrelevant
- binary data: two possible outcomes (buy vs. not buy, client vs. not client)
- nominal data: several discrete outcomes (choice of a political party, choice of a particular brand)
- ordinal data: several ordered outcomes (quality of products, results of a questionnaire)
- count data: (number or orders, number of insurance claims)
- time series data: collected at successive time periods (monthly sales, monthly unemployment, weekly turnover)
- panel data: several characteristics collected at successive time periods (monthly sales of several branch stores)

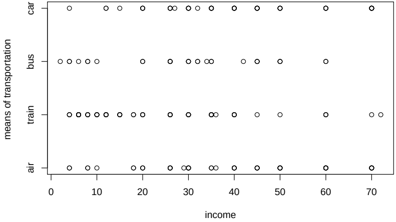
sales vs. CEO income



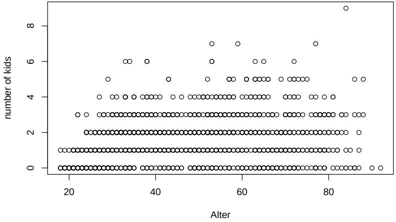
back-payment of a loan

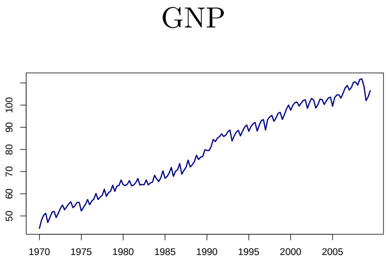


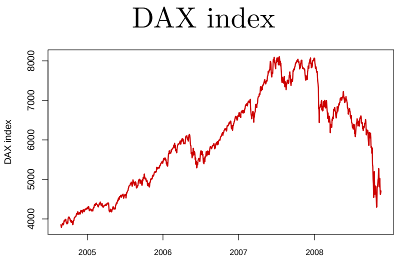
means of transportation

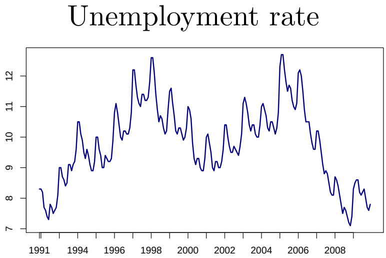


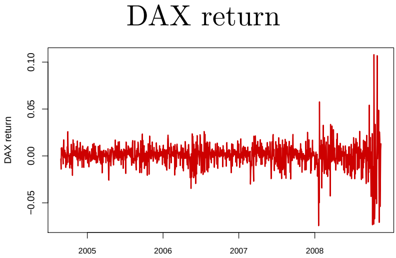
number of kids




GNP
50
60
70
80
90
100
1970
1975
1980
1985
1990
1995
2000
2005


DAX index
DAX index
4000
5000
6000
7000
8000
2005
2006
2007
2008


Unemployment rate
7
8
9
10
11
12
1991
1994
1996
1998
2000
2002
2004
2006
2008


DAX return
DAX return
−0.05
0.00
0.05
0.10
2005
2006
2007
2008

Part 2

Basics of forecasting

Explanatory/cross-sectional forecasting

- Aim: a statement about Y_{new} using information in additional explanatory variables
- The forecast is usually based on a regression-type function

$$\hat{Y}_{new} = \hat{f}(X_{1,new}, ..., X_{K,new})$$

TS forecasting

- Aim: a statement about Y_{t+h} using information at time point t
  - $h = 1$ - one-step-ahead forecast
  - $h > 1$ - multi-step-ahead forecast
- The forecast is usually a function of the historical data $Y_t, Y_{t-1}, Y_{t-2}, ...$ and exploits the specific "memory" of the process using the Box-Jenkins-principle.

$$\hat{Y}_{t+h} = \hat{f}(Y_t, Y_{t-1} ..., Y_{t-p})$$

Note: we consider (almost) exclusively continuous Y's

Judgmental forecasts

- Judgmental forecasts are particularly important for new or rare events.
- Frequently you get a direction of change, but not exact values.
- The forecasts of several experts can be combined using the Delphi method.
- The expert forecasts suffer from *behavioral biases*, e.g. *conservatism, anchoring, wishful thinking, overconfidence, recency, etc.* $\rightsquigarrow$ *behavioral economics*

Types of forecasts I

- Point forecasts: a single value $\hat{Y}_{t+h}$ for the unknown quantity Y_{t+h}.

$$\hat{Y}_{t+h} = \hat{E}(Y_{t+h}|\mathcal{I}_t),$$

where $\mathcal{I}_t$ denotes the information at time point t, e.g. $\mathcal{I}_t = \{Y_t, Y_{t-1}, ...\}$.

Our aim is to find the forecast $\hat{Y}^*_{t+1} = g(\mathcal{I}_t)$, which uses $\mathcal{I}_t$ and minimizes MSE, i.e.

$$MSE(\hat{Y}^*_{t+1}) = E(Y_{t+1} - g(\mathcal{I}_t))^2 \longrightarrow min, \quad \text{w.r.t.} \quad g(\cdot).$$

Types of forecasts II

$$
\begin{aligned}
MSE(g(\mathcal{I}_t)) &= E(Y_{t+1} - g(\mathcal{I}_t))^2 \\
&= E(Y_{t+1} - E(Y_{t+1}|\mathcal{I}_t) + E(Y_{t+1}|\mathcal{I}_t) - g(\mathcal{I}_t))^2 \\
&= E(Y_{t+1} - E(Y_{t+1}|\mathcal{I}_t))^2 + E(E(Y_{t+1}|\mathcal{I}_t) - g(\mathcal{I}_t))^2 \\
&+ 2\,\underbrace{E\big((Y_{t+1} - E(Y_{t+1}|\mathcal{I}_t)) \cdot (E(Y_{t+1}|\mathcal{I}_t) - g(\mathcal{I}_t))\big)}_{=0}.
\end{aligned}
$$

$E(Y_{t+1} - E(Y_{t+1}|\mathcal{I}_t))^2$ does not depend on g and $E(E(Y_{t+1}|\mathcal{I}_t) - g(\mathcal{I}_t))^2$ is minimal for

$$g(\mathcal{I}_t) = E(Y_{t+1}|\mathcal{I}_t).$$

Thus the forecast which minimizes the MSE is the conditional expectation!!!

- Forecast/prediction intervals: we compute the interval $[LB, UB]$, where Y_{t+h} takes a value with some predefined probability. In most of the cases the intervals are built according to the following principle:

$$[LB, UB] = [\hat{Y}_{t+h} + q_{\alpha/2}\sqrt{MSE_h};\ \hat{Y}_{t+h} + q_{1-\alpha/2}\sqrt{MSE_h}],$$

  where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are quantiles of an appropriate distribution. For the true value it holds

$$P(Y_{t+h} \in [LB, UB]) = 1 - \alpha.$$

- Forecast density: we compute the forecast density of Y_{t+h}. In this case we can make statements about

$$P(Y_{t+h} \in (a, b)), \quad P(Y_{t+h} > a), \quad P(Y_{t+h} < b).$$

Problem: for forecast intervals and densities we need assumptions about the distribution of historical data.

Goodness of forecasts I

The goodness of a forecast is measured by the forecast error:

$$\hat{\varepsilon}_{t+h} = Y_{t+h} - \hat{Y}_{t+h}.$$

For a good forecasting procedure the forecast errors should ...

- ... be small $\rightsquigarrow$ loss functions
- ... have no pattern and memory $\rightsquigarrow$ ACF for the forecast errors

Goodness of forecasts II

Loss functions

$$MSE_h = \frac{1}{\tau - h} \sum_{t=1}^{\tau-h} \hat{\varepsilon}^2_{t+h}$$ mean squared error

$$MAE_h = \frac{1}{\tau - h} \sum_{t=1}^{\tau-h} |\hat{\varepsilon}_{t+h}|$$ mean absolute error

$$MAPE_h = \frac{100}{\tau - h} \sum_{t=1}^{\tau-h} \left| \frac{Y_{t+h} - \hat{Y}_{t+h}}{Y_{t+h}} \right|$$ mean absolute % error

R^2 – of LR of $\hat{Y}_{t+h}$ on Y_{t+h} Minzer-Zarnowitz regression

Goodness of forecasts III

$$U_h = \sqrt{\frac{\sum_{t=1}^{\tau-h}\left(\frac{\hat{Y}_{t+h}-Y_{t+h}}{Y_t}\right)^2}{\sum_{t=1}^{\tau-h}\left(\frac{Y_t-Y_{t+h}}{Y_t}\right)^2}} \qquad \text{Theil's } U$$

$U = 1$ – naïve forecast is as good as the one from the model
$U < 1$ – naïve forecast is worse than the one from the model
$U > 1$ – naïve forecast is better than the one from the model

Goodness of forecasts IV

Important:

- Loss functions measure the out-of-sample performance of the underlying model.
- R^2, AIC, BIC, etc. measure the in-sample performance of the underlying model.
- The best in-sample model does not necessarily provide the best forecasts with the smallest loss function and vice versa.

- Very good in-sample models are frequently very complex....
- Very good out-of-sample models are frequently rather simple...

Goodness of forecasts V

There statistical tests to check if one procedure provides significantly better forecasts than other models.

- Equal Predictive Ability
  Diebold, F. X., and Mariano, R. S. (1995), “Comparing Predictive Accuracy”, Journal of Business & Economic Statistics, 13, 253-263.
- Superior Predictive Ability
  Hansen, P.R. (2005), “Test for Superior Predictive Ability”, Journal of Business & Economic Statistics, 23, 365-380.

Goodness of forecasts VI

Equal Predictive Ability (EPA)

Let g be a loss function, e.g. $g(x) = x^2$ or $|x|$, and let $\hat{\varepsilon}^A_{t+h}$ and $\hat{\varepsilon}^B_{t+h}$ be forecast errors from alternative models A and B.

The loss difference is:

$$d_t = g(\hat{\varepsilon}^A_{t+h}) - g(\hat{\varepsilon}^B_{t+h}).$$

H_0 : $E(d) = 0$ - two model provide the equally good forecasts
H_1 : $E(d) \neq 0$ - one model is better

Goodness of forecasts VII

- EPA: sign test with the test statistics

$$S = \frac{2}{\sqrt{\tau - h}} \sum_{t=1}^{\tau-h} (I\{d_t > 0\} - 0.5) \quad \overset{a.}{\sim} N(0,1).$$

  Idea: if H_0 is correct, then half of the d's must be positive. Strong deviations lead to the rejection of H_0.

- EPA: Wilcoxon sign rank test with the test statistics

$$W = \frac{\sum_{t=1}^{\tau-h} I\{d_t > 0\} \cdot rank(|d_t|) \; - \; (\tau - h)(\tau - h + 1)/4}{\sqrt{(\tau - h)(\tau - h + 1)(2(\tau - h) + 1)/24}} \quad \overset{a.}{\sim} N(0,1).$$

  Idea: we take not only the sing into account, but also the ranks.

Goodness of forecasts VIII

- EPA: Diebold-Mariano test
  Idea: we test directly the loss differences

$$DM = \frac{\bar{d}}{\sqrt{\widehat{Var}(\bar{d})}} \overset{a.}{\sim} N(0,1).$$

Rejection area for all three tests:

$$B = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$$

Goodness of forecasts IX

Superior Predictive Ability (SPA)

Let $\hat{\varepsilon}_{t+h}^{B}$ be the benchmark model and $\hat{\varepsilon}_{t+h}^{A_m}$ for $m = 1, ..., M$ the alternative models.

The loss differences with respect to the benchmark model are defined as:

$$d_t^{(m)} = g(\hat{\varepsilon}_{t+h}^{B}) - g(\hat{\varepsilon}_{t+h}^{A_m}).$$

H_0 : $E(d^{(m)}) < 0$ for all $m = 1, ..., M$ - the benchmark model is better
H_1 : $E(d^{(m)}) \geq 0$ for at least one m - at least one model is better than the benchmark

Splitting the data

If forecasting is the main objective of the modelling, then we shall split the data for evaluation purposes.

Approach 1: simple (randomized) splitting

- *Training* data set (70-80%): to fit and to evaluate the model
- *Test* data set: to evaluate the forecasts

Note: different test and training data sets might lead to different conclusions. Thus the measurement of the goodness of the forecasts might be misleading. A robust alternative is *cross-validation*.

Approach 2: cross-validation

- Make the "training/test" splitting randomly many times
- Note: Cross-validation is not straightforward for time series data!

Leave-one-out cross-validation LOOCV for cross -sectional data:

- The model is estimated n times.
- For the i-th estimation drop the i-th observation, i.e. the validation data set consists of a single observation.
- Determine the out-of-sample forecast $\hat{Y}_i$ and $MSE_i = (\hat{Y}_i - Y_i)^2$.
- LOOCV goodness-of-fit measure is

$$CV = \frac{1}{n}\sum_{i=1}^{n} MSE_i.$$

Cross-validation for TS

- For each time point t estimate the model using the observations
  - $1, \ldots, t-1 \rightsquigarrow$ expanding window
  - $t-\tau, \ldots, t-1 \rightsquigarrow$ moving window
- Compute the forecast for t and $MSE_t = (\hat{Y}_t - Y_t)^2$.
- CVCV goodness-of-fit measure is

$$CV = \frac{1}{n-\tau} \sum_{t=\tau+1}^{n} MSE_t.$$

k-fold cross-validation:

- split the data set k equally large parts.
- Part i is the *validation* data set.
- The model is estimated using the remaining observations and one computes MSE_i for the *validation* data set.
- We repeat it for each validation the data set.
- The final measure is

$$CV = \frac{1}{k} \sum_{i=1}^{k} MSE_i.$$

Note:

- Common values are $k = 5$ or 10.
- *Cross-Validation* can be applied for (almost) any models.
- The application to time series is sometimes more complicated, but works for autoregressive processes.

Forecast combinations

Let $\hat{Y}_{t+1}^{(m)}$ for $m = 1, \ldots, M$ be forecasts from different models, e.g. time series models, smoothing methods, experts.

Note:

- the true model is unknown;
- different models show better or worse performance in different periods.

Idea: weight the forecasts from different models using the current performance measure.

Simple forecast combination

$$\hat{Y}_{t+1} = \frac{1}{\sum_{m=1}^{M} w_t^{(m)}} \left(w_t^{(1)} \hat{Y}_{t+1}^{(1)} + \ldots + w_t^{(M)} \hat{Y}_{t+1}^{(M)} \right),$$

where $w_t^{(m)}$ is the individual weight of a single model.

The weights $w_t^{(m)}$ are updated by taking into account the current performance of each model:

$$w_{t+1}^{(m)} = \lambda w_t^{(m)} + (1-\lambda) g(|\hat{Y}_{t+1}^{(m)} - Y_{t+1}|),$$

with $g(x) = x$, $g(x) = \Phi(x) - 0.5$, etc.

Bayesian model averaging

Idea: weight the forecasts using the Bayes rule.

- $P(\text{model } m)$ - a-priori prob., that the model m is the correct model;
- $P(\mathcal{I}_t|\text{model } m)$ - the conditional probability, that the data comes from model m;
- $P(\text{model } m|\mathcal{I}_t)$ - a-posteriori prob., that for the given data the model m is the right model.
- $E(Y_{t+1}|\text{model } m, \mathcal{I}_t)$ - the optimal forecast, which relies on the given data and assuming that model m is the correct model.

$$E(Y_{t+1}|\mathcal{I}_t) = \sum_{m=1}^{M} E(Y_{t+1}|\text{model } m, \mathcal{I}_t) \cdot P(\text{model } m|\mathcal{I}_t)$$

$$P(\text{model } m|\mathcal{I}_t) = \frac{P(\mathcal{I}_t|\text{model } m)P(\text{model } m)}{\sum_{j=1}^{M} P(\mathcal{I}_t|\text{model } j)P(\text{model } j)}.$$

Characterisation of a TS

Aim: a measure for the strength of the memory of a time series

Statistics: Covariance/Correlation between two variables X und Y

$$\begin{aligned} Cov(X,Y) &= E(X - E(X))(Y - E(Y)) \\ Corr(X,Y) &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \end{aligned}$$

Within time series we examine the relationship between Y_t and Y_{t+h}.

- We write

$$\gamma_h = Cov(Y_{t+h}, Y_t) = E(Y_{t+h} - E(Y_{t+h}))(Y_t - E(Y_t))$$

- γ_h is called autocovariance function at lag h.

Estimator of autocovariance γ_h, $h > 0$:

$$\hat{\gamma}_h = \frac{1}{T} \sum_{t=1}^{T-h} (Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$

($\approx$ sample covariance of (Y_1, Y_{1+h}), ..., (Y_{T-h}, Y_T))

Definition: A time series $\{Y_t : t \in T\}$ is called to be **(weakly) stationary**, if it holds for all $t \in T$ that

1. $E(Y_t)$ does not depend on t (no trend),
2. $Var(Y_t)$ does not depend on t,
3. $Cov(Y_{t+h}, Y_t)$ depends on h, but not on t.

Properties:

It holds $\gamma_0 \geq 0, \gamma_h = \gamma_{-h}$ and $\mid \gamma_h \mid \leq \gamma_0$.

Autocorrelation function (ACF) I

ACF

The autocorrelation ρ_h at lag h measures the strength of linear dependence between Y_t and Y_{t-h}

Assuming stationarity we can write

$$Corr(Y_t, Y_{t+h}) = \frac{Cov(Y_t, Y_{t+h})}{\sqrt{Var(Y_t)Var(Y_{t+h})}} = \gamma_h/\gamma_0 = \rho_h$$

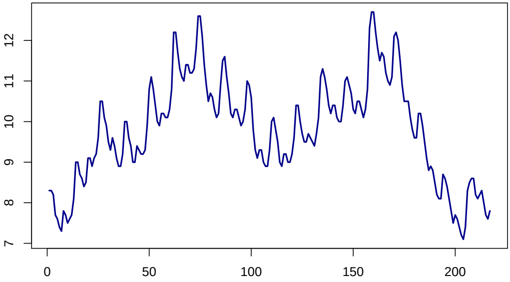
The empirical ACF is then

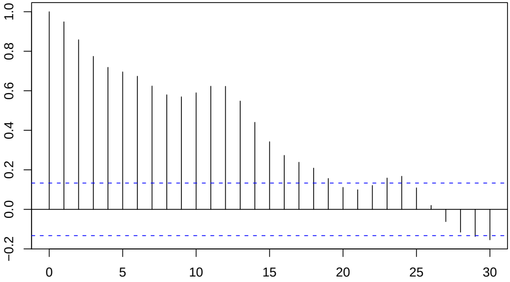
$$\hat{\rho}_h = \frac{\frac{1}{T-h}\sum_{t=1}^{T-h}(Y_t - \bar{Y})(Y_{t+h} - \bar{Y})}{\frac{1}{T}\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

$$|\hat{\rho}_h| \leq 1 \text{ for all } h$$

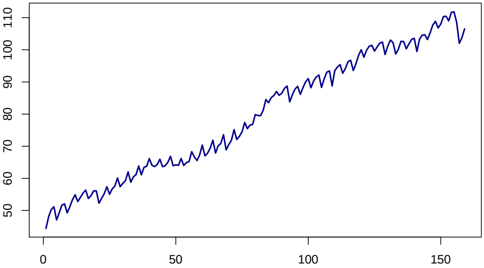
Example: Unemployment rate 12.1991-12.2009 (monthly data)

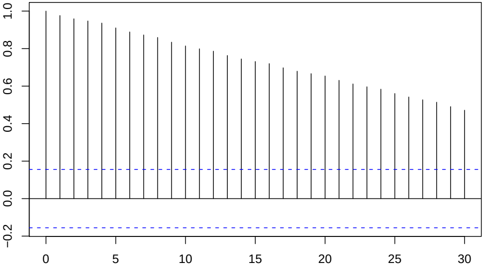
```
 > acf(x, lag.max=20)
          [,1]
 [1,] 1.0000000
 [2,] 0.9489491
 [3,] 0.8581498
 [4,] 0.7741281
 [5,] 0.7186758
 [6,] 0.6955368
 [7,] 0.6736293
 [8,] 0.6240986
 [9,] 0.5797972
[10,] 0.5695241
```


7
8
9
10
11
12
0
50
100
150
200

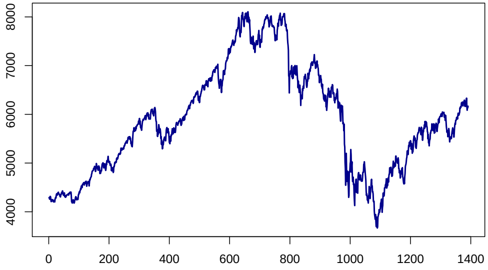

−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

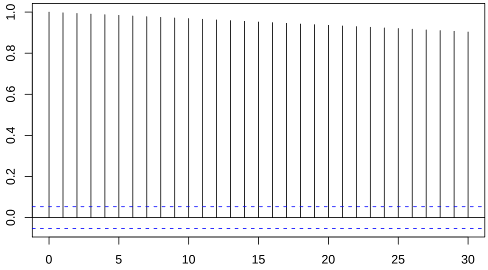
Example: GNP 01.1970-07.2009 (monthly data)



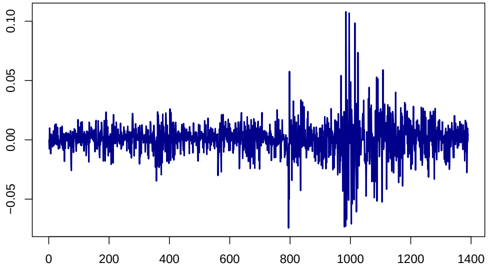


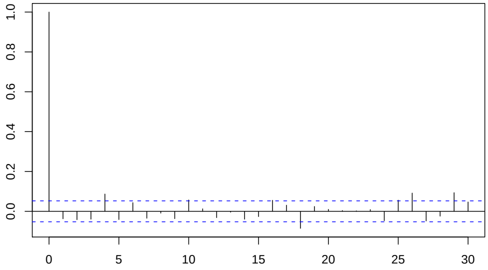
Example: DAX 03.01.2005-03.05.2010 (daily data)



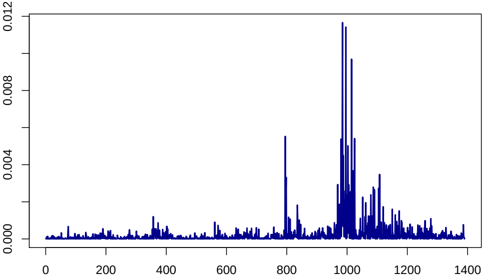


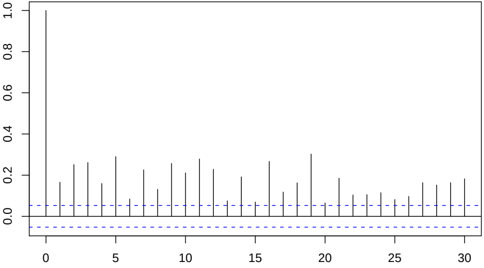
Example: DAX returns 03.01.2005-03.05.2010 (daily data)





Example: Squared returns of DAX 03.01.2005-03.05.2010





Part 3

Forecasting with regression techniques

Objectives of forecasting using regression

Let Y_i be the variable we wish to forecast using predictors $X_{1i}, \ldots, X_{Ji}$, i.e. using liner regression.

Aim: "a statement" about Y_0 using $x_{10}, \ldots, x_{J0}$.

Note: frequently we have data both in cross-section and in time dimension $\rightsquigarrow$ *panel data*

Linear regression

Linear Regression

$$
\begin{aligned}
Y_i &= b_0 + b_1 x_{i1} + \cdots + b_J x_{iJ} + \varepsilon_i, \text{ for } i = 1, \ldots, n \\
E(\varepsilon_i) &= 0 \\
Var(\varepsilon_i) &= \sigma^2 \\
Corr(\varepsilon_i, \varepsilon_j) &= 0 \quad \text{for } i \neq j \\
\varepsilon_i &\sim N(0, \sigma^2)
\end{aligned}
$$

In matrix notation we can write

Linear model

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{b} + \boldsymbol{\varepsilon} \quad \text{withs}$$

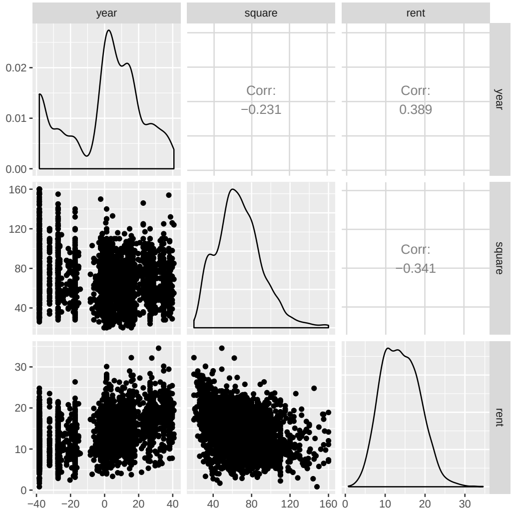
OLS estimation

$$\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \longrightarrow \text{min, wrt. } \boldsymbol{b}$$

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}'\,\boldsymbol{X})^{-1}\,\boldsymbol{X}'\,\boldsymbol{y}$$

$$\widehat{Var(\varepsilon_i)} = \widehat{\sigma}^2 = \frac{1}{n-J-1}\,(\boldsymbol{y} - \boldsymbol{X}\,\widehat{\boldsymbol{b}})'\,(\boldsymbol{y} - \boldsymbol{X}\,\widehat{\boldsymbol{b}})$$

Example: Analyse the impact of the apartment size and the year of construction on the rent (per sqm) for 3082 apartments in Munich.



Y_i	–	rent per sqm.
x_{i1}	–	year of construction (-mean)
x_{i2}	–	(year of construction) ^{2}
x_{i3}	–	1/square

$$rent_i = b_0 + b_1 \text{year}_i + b_2 \text{year}_i^2 + b_3 \frac{1}{\text{square}_i} + \varepsilon_i.$$

2000 observations are used as training data set and the remaining as test data set.

R: `lm`-function

```
  Residuals:
     Min       1Q   Median       3Q      Max
-13.5039  -2.6715  -0.2391   2.6951  16.0871

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.718e+00  2.692e-01   32.39   <2e-16 ***
X2year       8.557e-02  4.305e-03   19.88   <2e-16 ***
X2year2      2.011e-03  1.756e-04   11.45   <2e-16 ***
X2square.inv 2.460e+02  1.358e+01   18.11   <2e-16 ***
---

Residual standard error: 3.978 on 1996 degrees of freedom
Multiple R-squared: 0.2935,     Adjusted R-squared: 0.2925
F-statistic: 276.4 on 3 and 1996 DF,  p-value: < 2.2e-16
```

Note: the estimator depend on the random sample, so we shall threat them as random variables!

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{b}+\boldsymbol{\varepsilon}) = \boldsymbol{b} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$E(\hat{\boldsymbol{b}}) = \boldsymbol{b}$$

$$Var(\widehat{\boldsymbol{b}}) = \begin{pmatrix} Var(\hat{b}_0) & Cov(\hat{b}_0,\hat{b}_1) & \dots & Cov(\hat{b}_0,\hat{b}_J) \\ Cov(\hat{b}_1,\hat{b}_0) & Var(\hat{b}_1) & \dots & Cov(\hat{b}_1,\hat{b}_J) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{b}_J,\hat{b}_0) & Cov(\hat{b}_J,\hat{b}_1) & \dots & Var(\hat{b}_J) \end{pmatrix} = \sigma^2\,(\boldsymbol{X}'\,\boldsymbol{X})^{-1}$$

If the error terms $\boldsymbol{\varepsilon}$ follow normal distribution, it holds:

$$\hat{\boldsymbol{b}} \sim N_{J+1}(\boldsymbol{b},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}), \qquad \hat{b}_j \sim N(b_j,\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{(j,j)}).$$

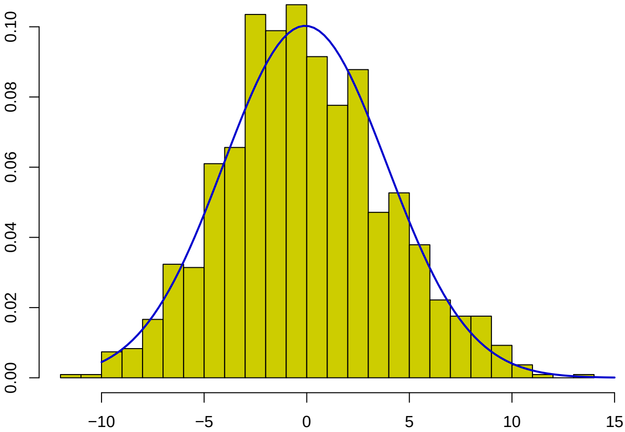
Note: the distribution of error terms is irrelevant for the estimation, but is crucial for tests and forecasts.
The estimated residuals:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \cdots - \hat{b}_J x_{iJ}$$
$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}$$

The distribution of the residuals can be tested using goodness-of-fit tests: χ^2- Test von Pearson, Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, etc.

$$H_0 : \varepsilon_i \sim N(\cdot,\cdot) \quad vs. \quad H_1 : \varepsilon_i \nsim N(\cdot,\cdot)$$

Example: KS-test with $D = 0.0386$, p-value $= 0.0052 \rightsquigarrow$ not normal



Coefficients as forecasts I

Note: b_j is the marginal change in the dependent variable, if X_j changes by one unit.

Thus $c \cdot \hat{b}_j$ is a point forecast of the change in y, if X_j changes for c units.

Since the distribution of $\hat{b}_j$ is known, we can construct prediction intervals

Coefficients as forecasts II

CI for parameters

The unknown parameter lies with probability of $(1-\alpha)\cdot 100\%$ in

$$\left[\hat{b}_j - t_{n-J-1;1-\alpha/2}\cdot\sqrt{\widehat{Var(\hat{b}_j)}};\;\; \hat{b}_j + t_{n-J-1;1-\alpha/2}\cdot\sqrt{\widehat{Var(\hat{b}_j)}}\right]$$

$$\left[\hat{b}_j - t_{n-J-1;1-\alpha/2}\cdot\sqrt{\hat{\sigma}^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{(j,j)}};\;\; \hat{b}_j + t_{n-J-1;1-\alpha/2}\cdot\sqrt{\hat{\sigma}^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{(j,j)}}\right.$$

Ci for coefficients

```
                  2.5 %       97.5 %
(Intercept)   8.190068e+00 9.245759e+00
X2year       7.712932e-02 9.401395e-02
X2year      1.666482e-03 2.355415e-03
X2square.inv 2.193511e+02 2.726137e+02
```

- The interpretation of `X2flaeche.inv` is not feasible .
- If the year B changes by one year (i.e. $B+1$), then the rent changes for $\hat{b}_1 + \hat{b}_2 \cdot (2B+1)$.

$$\hat{b}_1 + \hat{b}_2 \cdot (2B+1) = 8.557164 \cdot 10^{-2} + 2.010948 \cdot 10^{-3} \cdot (2B+1)$$

$$Var(\hat{b}_1 + \hat{b}_2 \cdot (2B+1)) = Var(\hat{b}_1) + (2B+1)^2 Var(\hat{b}_2) + 2(2B+1)Cov(\hat{b}_1, \hat{b}_2)$$

$$\widehat{Var}(\hat{b}_1 + \hat{b}_2 \cdot (2B+1) = 1.853 \cdot 10^{-5} + 3.085 \cdot 10^{-8} \cdot (2B+1) + 2 \cdot (2B+1) \cdot 2.624 \cdot 10^{-7}$$

CI for $b_1 + b_2 \cdot (2B+1)$ is thus

$$\Big[\hat{b}_1 + \hat{b}_2 \cdot (2B+1) - 1.96 \cdot \sqrt{\widehat{Var}(\hat{b}_1 + \hat{b}_2 \cdot (2B+1))};$$
$$\hat{b}_1 + \hat{b}_2 \cdot (2B+1) + 1.96 \cdot \sqrt{\widehat{Var}(\hat{b}_1 + \hat{b}_2 \cdot (2B+1))}\Big]$$

Forecasts I

Let $\boldsymbol{x}_0 = (1, x_{01}, \ldots, x_{0J})'$ be a new vector of observations which WAS NOT used for estimation

- The point forecast:

$$\hat{Y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{b}} = \hat{b}_0 + \hat{b}_1 x_{01} + \cdots + \hat{b}_J x_{0J}$$

- For the true values it holds:

$$Y_0 = b_0 + b_1 x_{01} + \cdots + b_J x_{0J} + \varepsilon_0$$

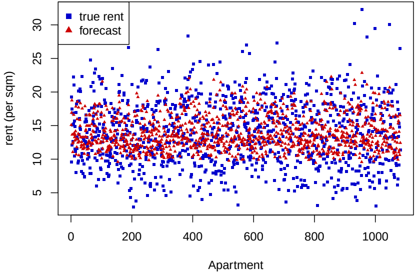
Forecasts II

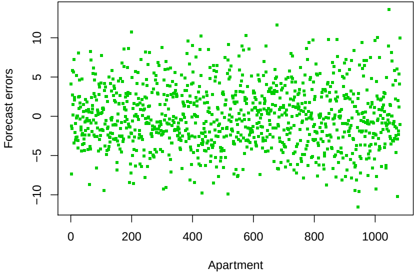
- The forecast error is:

$$\begin{aligned}\hat{\varepsilon}_0 &= Y_0 - \hat{Y}_0 \\ &= \varepsilon_0 + (b_0 - \hat{b}_0) + (b_1 - \hat{b}_1)x_{01} + \cdots + (b_J - \hat{b}_J)x_{0J} \\ &= \varepsilon_0 + \boldsymbol{x}_0'(\boldsymbol{b} - \hat{\boldsymbol{b}})\end{aligned}$$

- The variance of the forecast errors is then:

$$Var(\hat{\varepsilon}_0) = \sigma^2(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)$$





for year=16.69371, year^2=278.6798 and 1/square= 0.01785714 we obaine the forecast $\hat{Y}_0 = 15.09937$ with the forecast error

$$\hat{\varepsilon}_0 = Y_0 - \hat{Y}_0 = 19.2375 - 15.09937 = 4.138126.$$

The variance of the forecast error is:

$$\widehat{Var}(\hat{\varepsilon}_0) = \hat{\sigma}^2(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0) = 15.8364$$

Forecasts for Y_0 in a LR

- point forecast:

$$\hat{Y}_0 = \boldsymbol{x}_0'\hat{b} = \hat{b}_0 + \hat{b}_1 x_{01} + \cdots + \hat{b}_J x_{0J} = \boldsymbol{x}_0'\hat{\boldsymbol{b}}.$$

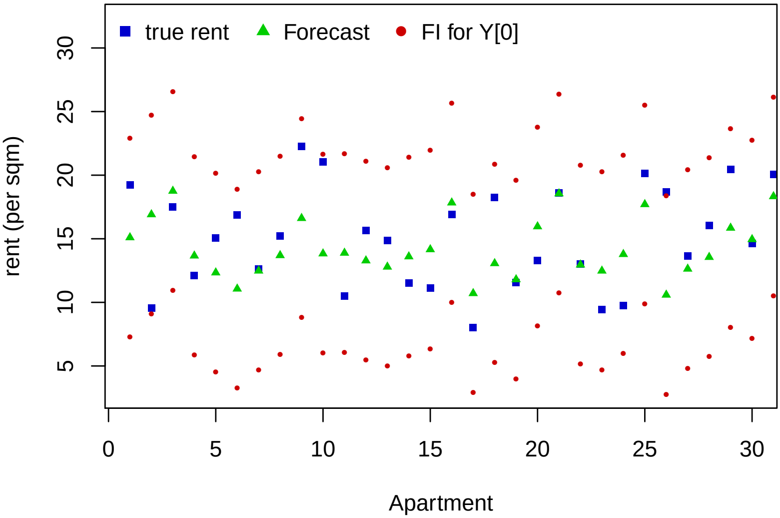
- interval forecast: Y_0 lies with prob. of $(1-\alpha)\cdot 100\%$ in

$$\left[\hat{Y}_0 - t_{n-J-1;1-\alpha}\sqrt{\widehat{Var(\hat{\varepsilon}_0})};\ \hat{Y}_0 + t_{n-J-1;1-\alpha}\sqrt{\widehat{Var(\hat{\varepsilon}_0})}\right]$$

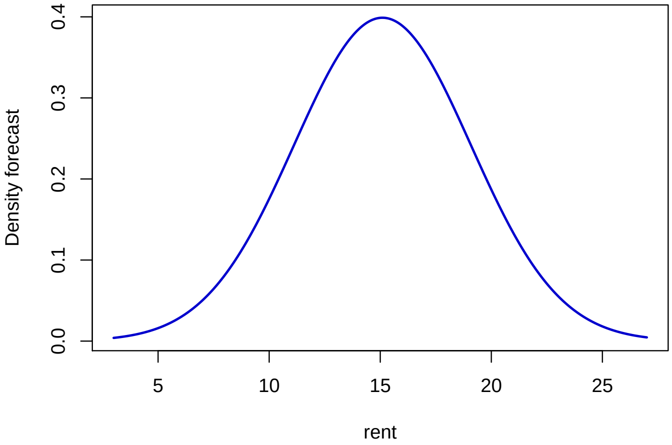
mit $\widehat{Var(\hat{\varepsilon}_0}) = \hat{\sigma}^2(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)$.

- forecast density: for the unknown value Y_0 and the forecast $\hat{Y}_0$ it holds

$$\frac{(Y_0 - \hat{Y}_0)}{\sqrt{\widehat{Var(\hat{\varepsilon}_0})}} \sim t_{n-J-1;1-\alpha}.$$


true rent
Forecast
FI for Y[0]
rent (per sqm)
Apartment
0
5
10
15
20
25
30
5
10
15
20
25
30

The density forecast for an apartment with year=16.69371, year^2=278.6798 and 1/square= 0.01785714. It holds $\hat{Y}_0 = 15.09937$ and $\widehat{Var}(\hat{\varepsilon}_0) = 15.8364$.



Forecasts for $E(Y_0|\boldsymbol{x}_0)$

Note: $\hat{Y}_0$ can be used to estimate not only Y_0, but also $E(Y_0|\boldsymbol{x}_0)$. We are interested NOT in the exact value of Y_0, but in its expected value:

$$E(Y_0|\boldsymbol{x}_0) = b_0 + b_1 x_{01} + \cdots + b_1 x_{0J}$$

$$\begin{aligned}
\hat{\varepsilon}_0^{(e)} &= E(Y_0|\boldsymbol{x}_0) - \hat{Y}_0 \\
&= (b_0 - \hat{b}_0) + (b_1 - \hat{b}_1)x_{01} + \cdots + (b_J - \hat{b}_J)x_{0J} \\
&= \boldsymbol{x}_0'(\boldsymbol{b} - \hat{\boldsymbol{b}})
\end{aligned}$$

with the variance of the forecast error

$$Var(\hat{\varepsilon}_0^{(e)}) = \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$$

Example: the forecast error $\hat{\varepsilon}_0^{(e)}$ cannot be computed, since $E(Y_0|\boldsymbol{x}_0)$ is not observable.

The variance of the forecast error is then:

$$\widehat{Var}(\hat{\varepsilon}_0^{(e)}) = \hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 = 0.1116627$$

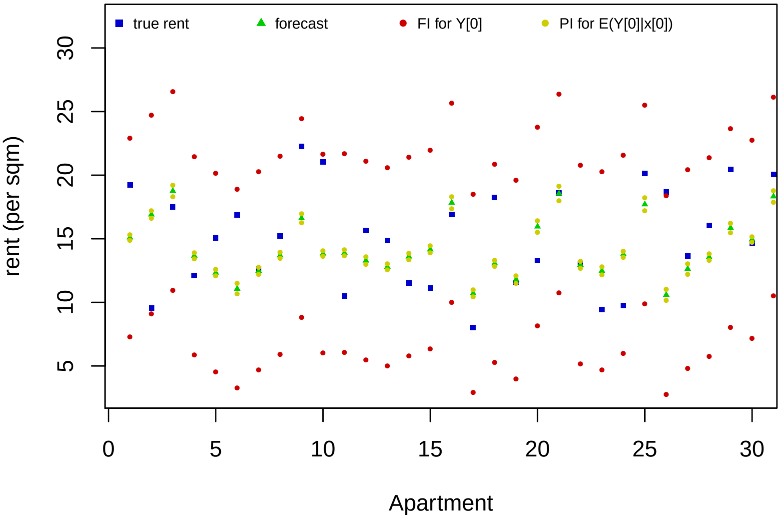
Forecast for $E(Y_0|\boldsymbol{x}_0)$

- Interval forecasts for $E(Y_0|\boldsymbol{x}_0)$: $E(Y_0|\boldsymbol{x}_0)$ lies with prob. of $(1-\alpha)\cdot 100\%$ in

$$\left[\hat{Y}_0 - t_{n-J-1;1-\alpha}\sqrt{\widehat{Var(\hat{\varepsilon}_0^{(e)})}};\ \hat{Y}_0 + t_{n-J-1;1-\alpha}\sqrt{\widehat{Var(\hat{\varepsilon}_0^{(e)})}}\right]$$

mit $\widehat{Var(\hat{\varepsilon}_0^{(e)})} = \hat{\sigma}^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0.$

- Forecast intervals for Y_0 are wider and are called *prediction intervals*.
- Forecast intervals for $E(Y_0|\boldsymbol{x}_0)$ are narrower and are called *confidence intervals*.

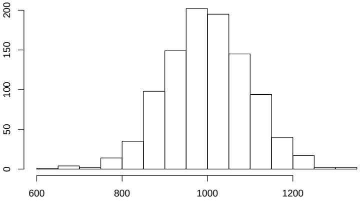

true rent
forecast
FI for Y[0]
PI for E(Y[0]|x[0])
rent (per sqm)
5
10
15
20
25
30
0
5
10
15
20
25
30
Apartment

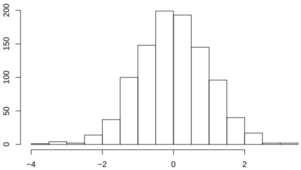
Transformations

The data is frequently transformed. This can improve the stability and the quality of the forecasts.

- Standardization: makes the interpretation difficult, but simplifies the inference and precision

$$x_i^* = \frac{x_i - \bar{x}}{s_x}.$$





- **Reduction of asymmetry**: many methods work only with symmetric data

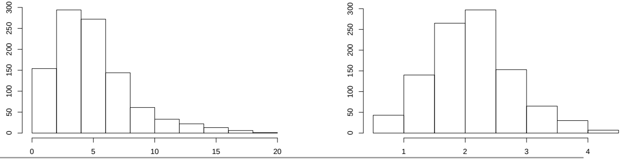
$$\text{Skewness} = \frac{(x_i - \bar{x})^3}{(n-1)s_x^{3/2}}.$$

If skewness≥ 0, then the distribution is right-skewed, else it is left-skewed.

$$x_i^* = ln(x_i), \qquad \sqrt{x_i}, \qquad \frac{1}{x_i}, \qquad \frac{x^\lambda - 1}{\lambda}, \;\; \lambda \neq 0$$

Or the Box-Cox-transformation with an estimated parameter λ

$$x_i^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ ln(x_i) & \text{for } \lambda = 0 \end{cases}.$$



Transformation of Y

$$ln(y_i) = z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iJ} + u_i.$$

Using LS approach we estimate the parameters and obtain for $(x_{01}, \ldots, x_{0J})$ the forecasts $\hat{z}_0$.

But: it is in general wrong to forecast y_0 by $\hat{y}_0 = e^{\hat{z}_0}$! It holds

$$E(\hat{Z}_0 | x_{01}, \ldots, x_{0J}) = z_0$$

but

$$E(e^{\hat{z}_0} | x_{01}, \ldots, x_{0J}) \neq e^{z_0} = y_0$$

Thus the forecasts are biased.

If $Z \sim N(\mu, \sigma^2)$, then

$$E(e^Z) = e^{\mu + \frac{1}{2}\sigma^2}.$$

Thus if the residuals are Gaussian, then the following forecasts are optimal:

$$\hat{y}_0^{(opt)} = e^{\hat{z}_0 + \frac{1}{2}\widehat{Var(\hat{z}_0)}} = e^{\hat{z}_0 + \frac{1}{2}\widehat{Var(\hat{\varepsilon}_0)}}$$

Note:

- Compute both forecasts and choose the method with a better fit.
- The optimal forecasts depend on the type of transformation.

Example: in-sample vs. out-of-sample fit?

$$\text{Model 1}: rent_i = b_0 + b_1 \text{year}_i + b_2 \text{year}_i^2 + b_3 \frac{1}{\text{square}_i} + \varepsilon_i$$

$$\text{Model 2}: rent_i = b_0 + b_1 \text{year}_i + b_2 \text{square}_i + \varepsilon_i$$

$$R^2_{M1} = 0.2925, \qquad R^2_{M2} = 0.2081$$

For the *out-of-sample* forecasts we obtain:

$$MSE_{M1} = \frac{1}{1082} \sum_{i=1}^{1082} (\hat{Y}_i^{(M1)} - Y_i)^2 = 15.81969,$$

$$MSE_{M2} = \frac{1}{1082} \sum_{i=1}^{1082} (\hat{Y}_i^{(M2)} - Y_i)^2 = 17.57824,$$

$$MAE_{M1} = \frac{1}{1082} \sum_{i=1}^{1082} |\hat{Y}_i^{(M1)} - Y_i| = 3.375251,$$

$$MAE_{M2} = \frac{1}{1082} \sum_{i=1}^{1082} |\hat{Y}_i^{(M2)} - Y_i| = 3.1876$$

$$d_i^{(MSE)} = (\hat{Y}_i^{(M1)} - Y_i)^2 - (\hat{Y}_i^{(M2)} - Y_i)^2$$
$$d_i^{(MAE)} = |\hat{Y}_i^{(M1)} - Y_i| - |\hat{Y}_i^{(M2)} - Y_i|$$

- **Sign-test**: is the median of d_i equal 0, so is for a half of the sample model 1 a better choice, and the other half the model 2.

```
  SIGN.test(loss1mae-loss2mae, md=0)
        One-sample Sign-Test

data:  loss1mae - loss2mae
s = 614, p-value = 1.013e-05
alternative hypothesis:
    true median is not equal to 0
sample estimates:
median of x
  0.3046949
```

```
> SIGN.test(loss1-loss2, md=0)

        One-sample Sign-Test

data:  loss1 - loss2
s = 614, p-value = 1.013e-05
alternative hypothesis:
    true median is not equal to 0
95 percent confidence interval:
 0.5176113 1.2298277
sample estimates:
median of x
  0.8558778
```

- **Wilcoxon-sign-rank-test**: is the median of d_i equal 0

```
> wilcox.test(loss1mae-loss2mae)

data:  loss1mae - loss2mae
V = 346940.5, p-value = 1.512e-07
```

```
> wilcox.test(loss1-loss2)

data:  loss1 - loss2
V = 349199, p-value = 4.481e-08
```

- **Diebold-Mariano-t-test**: ist the expectation of d_i equal 0

```
> t.test(loss1mae-loss2mae)
        One Sample t-test

data:  loss1mae - loss2mae
t = 5.2616, df = 1081, p-value = 1.722e-07
alternative hypothesis:
    true mean is not equal to 0
sample estimates:
mean of x
0.1876511
```

```
> t.test(loss1-loss2)
        One Sample t-test

data:  loss1 - loss2
t = 5.0525, df = 1081, p-value = 5.117e-07
alternative hypothesis:
    true mean is not equal to 0
sample estimates:
mean of x
  1.75855
```

$\rightsquigarrow$ Model 1 is significantly better, if MSE is used as a criteria. For MAE the model 2 is better.

Example: Models 1 and 2 (s. above)

```
> library("cvTools")
> Z = lm(Y~year+year2+square.inv, data=cv.data);
> cvFit(Z, data=cv.data, y=cv.data$Y, K=n, foldType="random")
Leave-one-out CV results:
      CV
3.979901
> cvFit(Z, data=cv.data, y=cv.data$Y, K=5, foldType="random")
5-fold CV results:
      CV
3.983318


> Z2 = lm(Y~year+square, data=cv.data2);
> cvFit(Z, data=cv.data2, y=cv.data$Y, K=n, foldType="random")
Leave-one-out CV results:
      CV
4.205233
> cvFit(Z, data=cv.data2, y=cv.data$Y, K=5, foldType="random")
5-fold CV results:
      CV
4.205515
```

Linear regression with time series data

Now: the dependent and independent variables are time series

Problems:

- The error terms are correlated $\rightsquigarrow$ the assumption $Cov(\varepsilon_i, \varepsilon_j)$ is not fulfilled ;
- there might be trends and seasonalities in the data.

Autocorrelated error terms

If the errors are correlated, then

$$Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Omega} = \begin{pmatrix} \sigma^2 & \sigma_{12} & \dots & \sigma_{1\tau} \\ \sigma_{21} & \sigma^2 & \dots & \sigma_{2\tau} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\tau 1} & \sigma_{\tau 2} & \dots & \sigma^2 \end{pmatrix}$$

Then it holds

$$Var(\hat{\boldsymbol{b}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \text{matrix},$$

and the true variance might be either under- or overestimated.

Visual analysis of residuals

$$
\begin{aligned}
\hat{\varepsilon}_t &= Y_t - \hat{Y}_t \\
\hat{\rho}_{\varepsilon,h} &= \widehat{Corr}(\hat{\varepsilon}_t, \hat{\varepsilon}_{t+h}) \\
&= \frac{\sum_{t=1}^{\tau-h}(\hat{\varepsilon}_t - \bar{\hat{\varepsilon}}_t)(\hat{\varepsilon}_{t+h} - \bar{\hat{\varepsilon}}_{t+h})}{\sqrt{\sum_{t=1}^{\tau-h}(\hat{\varepsilon}_t - \bar{\hat{\varepsilon}}_t)^2 \sum_{t=1}^{\tau-h}(\hat{\varepsilon}_{t+h} - \bar{\hat{\varepsilon}}_{t+h})^2}}
\end{aligned}
$$

Note:

Y_t – real investititions

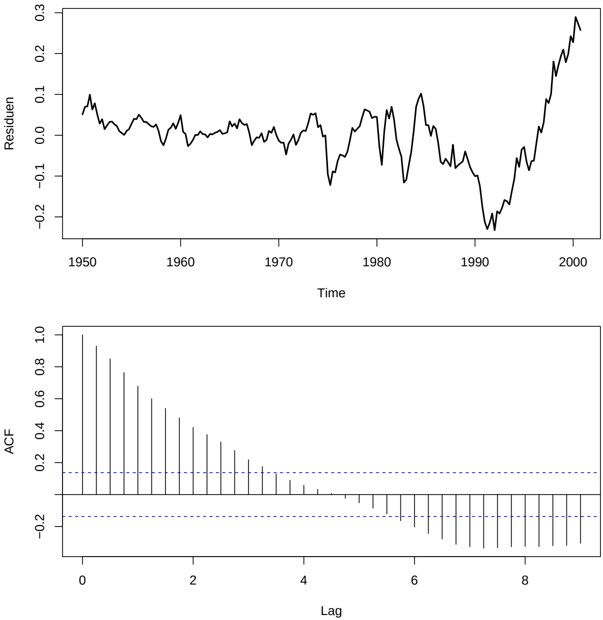
X_1 – GNP

X_2 – inflation

X_3 – interest rates

```
 Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.151802   0.015706  -9.665  < 2e-16 ***
realgdp      0.185548   0.003183  58.299  < 2e-16 ***
infl        -0.007832   0.002347  -3.337  0.00101 **
realint     -0.009045   0.002901  -3.118  0.00209 **
---

Residual standard error: 0.08518 on 200 degrees of freedom
Multiple R-squared: 0.9534,     Adjusted R-squared: 0.9527
F-statistic:  1364 on 3 and 200 DF,  p-value: < 2.2e-16
```


Residuen
0.3
0.2
0.1
0.0
−0.1
−0.2
1950
1960
1970
1980
1990
2000
Time
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0
2
4
6
8
Lag

A statistical tool to check for autocorrelation is the Durbin-Watson test.

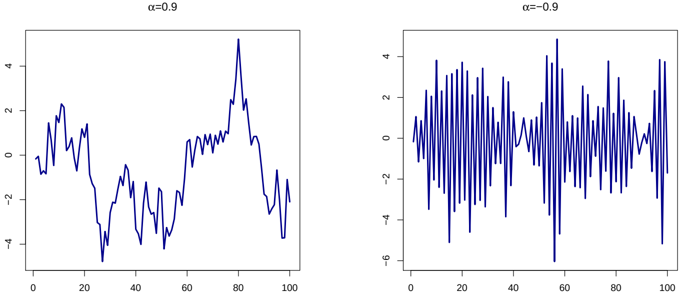
Idea: check the strength of the correlation between two subsequent residuals.

$$H_0 \quad : \quad Corr(\varepsilon_t, \varepsilon_{t+1}) = 0$$
$$H_1 \quad : \quad Corr(\varepsilon_t, \varepsilon_{t+1}) > (<) 0$$

- The test statistics:

$$d = \frac{\sum_{t=1}^{\tau-1}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t+1})^2}{\sum_{t=1}^{\tau}\hat{\varepsilon}_t^2}$$

- It holds $d \in [0;4]$. If d is close to 4, then we suspect negative autocorrelation. Is d close to 0, the we suspect positive autocorrelation.



- Note: d does not follow any standard distribution $\rightsquigarrow$ check p-values.

Example Durbin/Watson-test

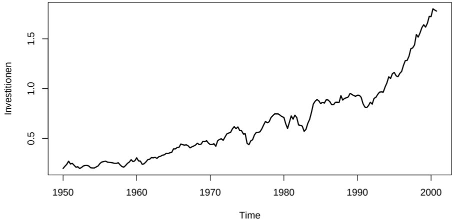
```
data:  Z
DW = 0.0922, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The residuals are strongly autocorrelated and the results might be misleading.

Note: taking autocorrelation into account is not trivial (GLS, 2SLS).

Time variables

- It the dependent variable has a clear trend, one uses the time as explanatory variable.




```
       Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.567e+03  1.964e+02  -7.979 1.16e-13 ***
X.time       1.641e+00  2.019e-01   8.128 4.67e-14 ***
X.time2     -4.294e-04  5.188e-05  -8.277 1.85e-14 ***
realgdp      5.664e-01  2.511e-02  22.556  < 2e-16 ***
infl         4.762e-03  1.881e-03   2.532 0.012105 *
realint      7.974e-03  2.119e-03   3.764 0.000221 ***
---

Residual standard error: 0.05036 on 198 degrees of freedom
Multiple R-squared: 0.9839,     Adjusted R-squared: 0.9835
F-statistic:  2417 on 5 and 198 DF,  p-value: < 2.2e-16
```

- Seasonal dummies

$$\begin{aligned} D_1 = 1 \quad &- \quad \text{for the 1st quarter, else 0;} \\ D_2 = 1 \quad &- \quad \text{for the 2nd quarter, else 0;} \\ D_3 = 1 \quad &- \quad \text{for the 3rd quarter, else 0.} \end{aligned}$$

$$\begin{aligned} D_1 = 1 \quad &- \quad \text{for January, else 0;} \\ D_2 = 1 \quad &- \quad \text{for February, else 0;} \\ &\vdots \\ D_{11} = 1 \quad &- \quad \text{for November, else 0.} \end{aligned}$$

Example:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.574e+03  1.979e+02  -7.953 1.45e-13 ***
X.time       1.648e+00  2.034e-01   8.100 5.87e-14 ***
X.time2     -4.313e-04  5.228e-05  -8.248 2.35e-14 ***
realgdp      5.674e-01  2.532e-02  22.414  < 2e-16 ***
infl         4.733e-03  1.893e-03   2.500 0.013259 *
realint      8.176e-03  2.152e-03   3.799 0.000194 ***
D1TRUE       2.062e-03  1.004e-02   0.205 0.837436
D2TRUE       3.186e-03  1.005e-02   0.317 0.751647
D3TRUE      -3.620e-03  1.020e-02  -0.355 0.722937
---

Residual standard error: 0.05068 on 195 degrees of freedom
Multiple R-squared: 0.9839,     Adjusted R-squared: 0.9833
F-statistic:  1492 on 8 and 195 DF,  p-value: < 2.2e-16
```

- Trading days: the foreacast of monthly sales depends heavily on the number of the individual weekdays, e.g. number of saturdays

$$
\begin{array}{lcl}
T_1 & - & \text{number of Mondays ;} \\
T_2 & - & \text{number of Tuesdays;} \\
\vdots & & \\
T_7 & - & \text{number of Saturdays ;}
\end{array}
$$

- Special effects

  $Z = 1$ – if it is a month before Easter or Christmas, else 0;

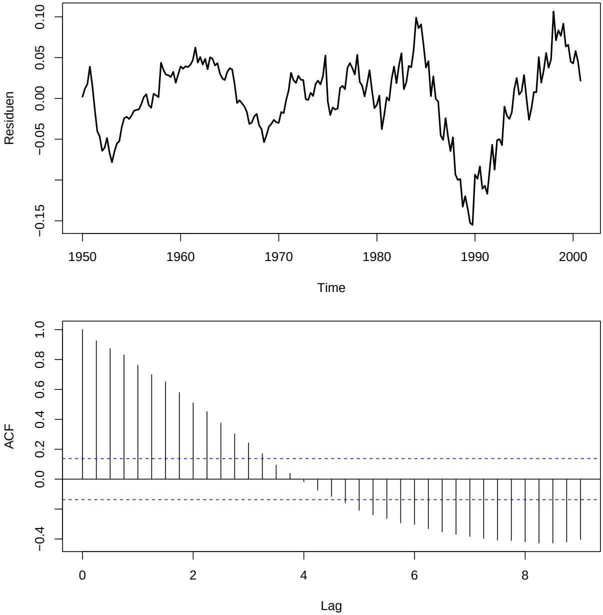
  $Z = 1$ – 1 after a technical improvement and 0 before;

  $Z = 1$ – in one economic phase and 0 in another phase;

Example: $J = 0$ before 1992 and $J = 1$ after 1992.

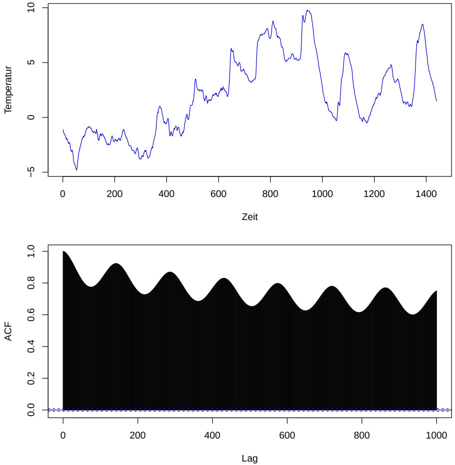
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.255e+03  2.117e+02  -5.927 1.36e-08 ***
X.time       1.321e+00  2.174e-01   6.074 6.34e-09 ***
X.time2     -3.475e-04  5.585e-05  -6.221 2.90e-09 ***
realgdp      5.515e-01  2.483e-02  22.210  < 2e-16 ***
infl         2.191e-03  1.978e-03   1.108 0.269322
realint      4.642e-03  2.279e-03   2.037 0.043008 *
J           -6.876e-02  1.999e-02  -3.439 0.000712 ***
---

Residual standard error: 0.04903 on 197 degrees of freedom
Multiple R-squared: 0.9848,     Adjusted R-squared: 0.9843
F-statistic:  2126 on 6 and 197 DF,  p-value: < 2.2e-16
```


Residuen
0.10
0.05
0.00
−0.05
−0.15
1950
1960
1970
1980
1990
2000
Time
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.4
0
2
4
6
8
Lag

- If we have very strong seasonalities the trigonometric function may be helpful.

Example: 10-minutes temperature for 2008, 52400 observations

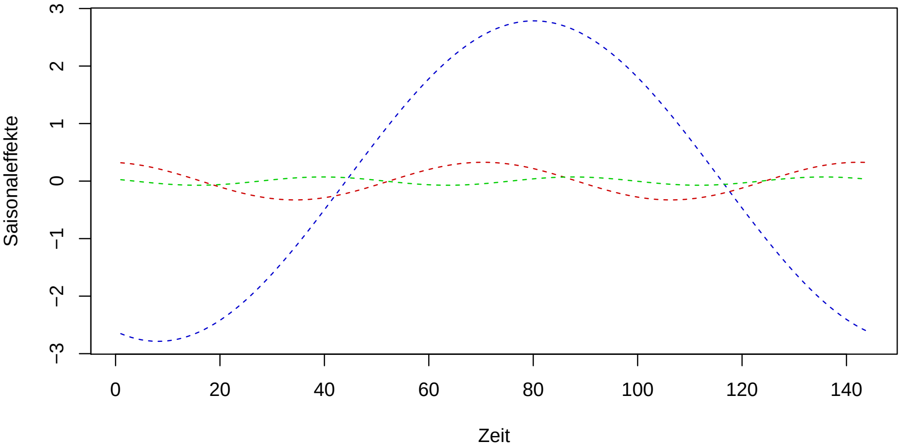


$$T_t = a_0 + a_1 t + a_2 t^2 + \sum_{p=1}^{3} \left[b_p cos\left(2p\pi \frac{X_t}{144}\right) + d_p sin\left(2p\pi \frac{X_t}{144}\right) \right] + \varepsilon_t$$

where $X_t = 1, ..., 144$ is an intraday period.

```
  Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -4.069e+00  5.650e-02  -72.017   <2e-16 ***
XT           1.580e-03  4.980e-06  317.327   <2e-16 ***
XT^2        -2.963e-08  9.202e-11 -322.023   <2e-16 ***
Xcos1       -2.613e+00  2.664e-02  -98.079   <2e-16 ***
Xsin1       -9.671e-01  2.663e-02  -36.316   <2e-16 ***
Xcos2        3.239e-01  2.664e-02   12.160   <2e-16 ***
Xsin2       -4.686e-02  2.663e-02   -1.759   0.0785 .
Xcos3        3.295e-02  2.663e-02    1.237   0.2160
Xsin3       -6.393e-02  2.664e-02   -2.400   0.0164 *
---

Residual standard error: 4.311 on 52391 degrees of freedom
Multiple R-squared: 0.6875,     Adjusted R-squared: 0.6874
F-statistic: 1.441e+04 on 8 and 52391 DF,  p-value: < 2.2e-16
```


Saisonaleffekte
Zeit
3
2
1
0
−1
−2
−3
0
20
40
60
80
100
120
140

B-spline regression

The time-series is modelled by a set of polynomials.

$$Y_t = b_0 + \sum_{i=1}^{k} b_i B_i^{(q)}(t) + \varepsilon_t,$$

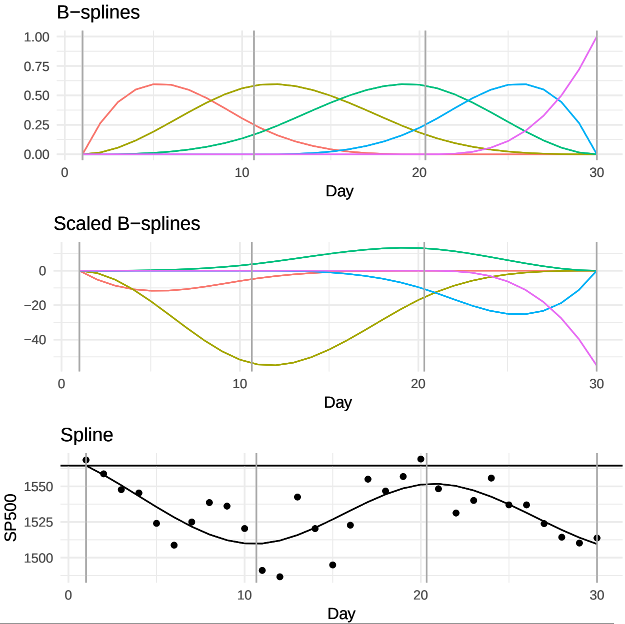
where

- B-splines $B_i^{(q)}$ are k polynomials of order q.
- Equally spaced grid $t_0, \ldots, t_{k+1}$
- B-spline of order q in the subinterval i is recursively defined as

$$B_i^{(q)}(t) = \alpha_{i,q}(t) B_i^{(q-1)}(t) + (1 - \alpha_{i+1,q}(t)) B_{i+1}^{(q-1)}(t), \quad \text{with}$$

$$\alpha_{i,q}(t) = \frac{t - t_i}{t_{i+q-1} - t_i} \qquad \text{and} \qquad B_i^{(0)}(t) = \mathbb{1}_{[t_i,\, t_{i+1})}(t).$$

Setup: $q = 5$, $k = 3$, $d = 30$



Part 4

Time series decomposition

Time series decomposition

Let Y_t be a time series with time index $t = 1, \ldots, \tau$. Y_t is a RV for each t.

A time series can be decomposed into four components:

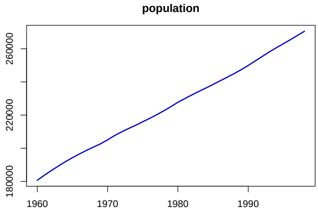
- long term trend component T;
- repeating seasonal component S;
- a component which repeats over several periods, cyclic component C;
- irregular residual component I.

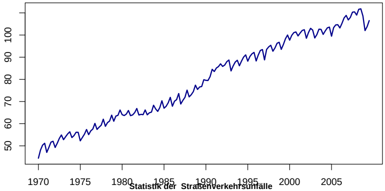
Additive time series model:

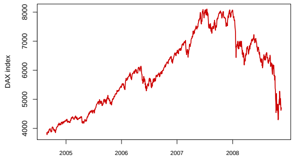
$$Y_t = T_t + S_t + C_t + I_t$$

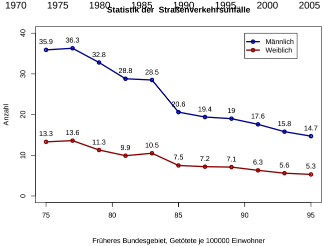
Time series with trends

Trend is a long term component, which explains increases or decreases of the time series over many periods.





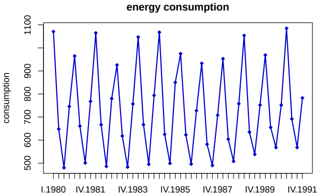


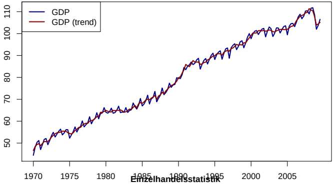


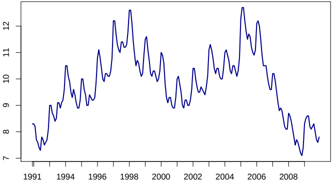
Modeling: exponential smoothing, simple regression (deterministic trend, e.g. linear, exp-trend), ARIMA-models (stochastic trends)

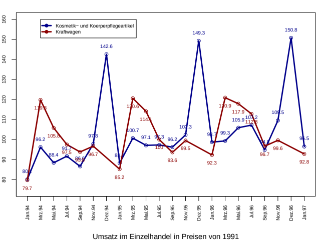
Time series with seasonal components

Seasonal component explains wave-type fluctuations around the trend, which repeat over fixed time periods.





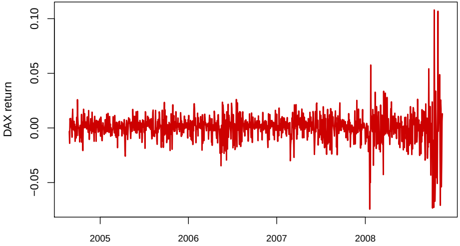




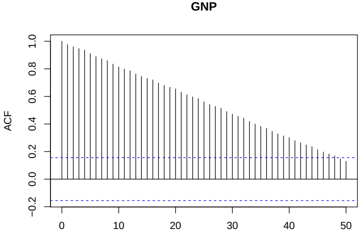
Modeling: classical decomposition, seasonal exponential smoothing, multiple regression, seasonal ARIMA-models

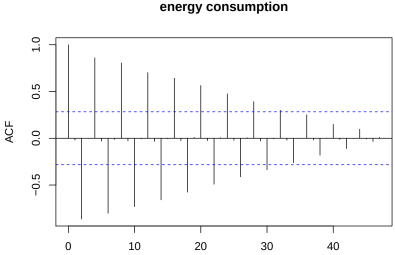
Residual component

The residual component shows no specific pattern, but just random behavior.

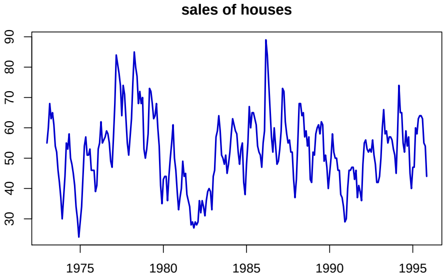


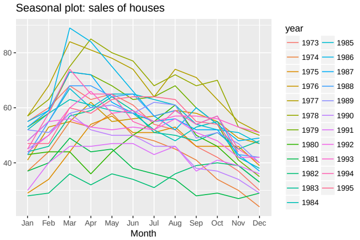
Modeling: simple smoothing, ARMA-models

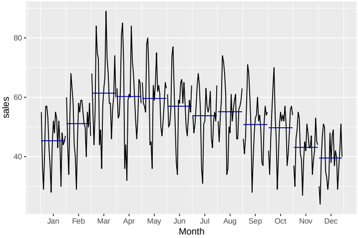




- A very slowly falling ACF indicates a stochastic or a deterministic trend.
- A quickly (exponentially) falling ACF indicates a stationary process.
- Regular spikes in the ACF indicate a seasonal component.


sales of houses
90
80
70
60
50
40
30
1975
1980
1985
1990
1995


Seasonal plot: sales of houses
80
60
40
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month
year
1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984
1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995


sales
80
60
40
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month

Trend: simple moving average I

Aim: extract the trend component of a time series.

Idea:

- the observations which are close in time are similar
- the average eliminates the irregular component and reduces the impact of seasonalities
- thus the average contains the trend

Question: how many observations should be averaged?

Trend: simple moving average II

Simple k-MA moving average (k is odd)

Let k be the odd order of the moving average and $m = (k-1)/2$. Then the trend component at time point t equals

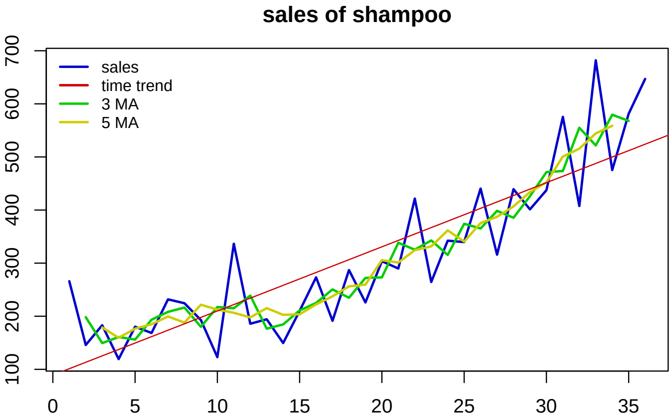
$$T_t = \frac{1}{k} \sum_{j=-m}^{m} Y_{t+j}.$$

For $k = 3$ it holds $m = 1$ and

$$T_t = \frac{1}{3} \cdot (Y_{t-1} + Y_t + Y_{t+1}).$$

Note: the trend cannot be computed for the first and for the last m observations. Particularly important are the last!

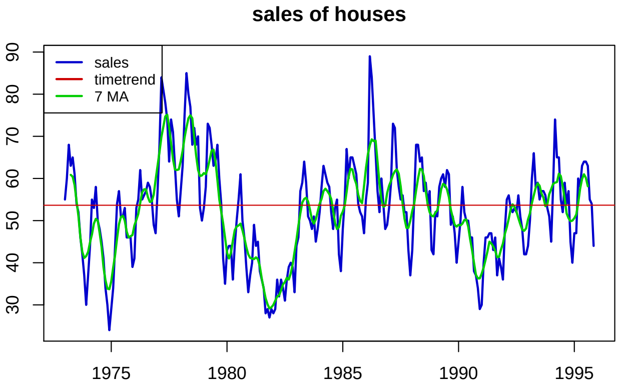
Trend: simple moving average III



Function `ma(...., order=k)` in R.

Trend: simple moving average IV

Problem: for complicated time series the k-MA method is useless.



Trend: centered moving average I

Problem: What to do with even k?

Let $k = 4$.

$$\begin{aligned}
T_{2.5} &= \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4) \\
T_{3.5} &= \frac{1}{4}(Y_2 + Y_3 + Y_4 + Y_5) \\
T_3'' &= \frac{1}{2}(T_{2.5} + T_{3.5}) \\
&= \frac{1}{4}(0.5 \cdot Y_1 + Y_2 + Y_3 + Y_4 + 0.5 \cdot Y_5)
\end{aligned}$$

Trend: centered moving average II

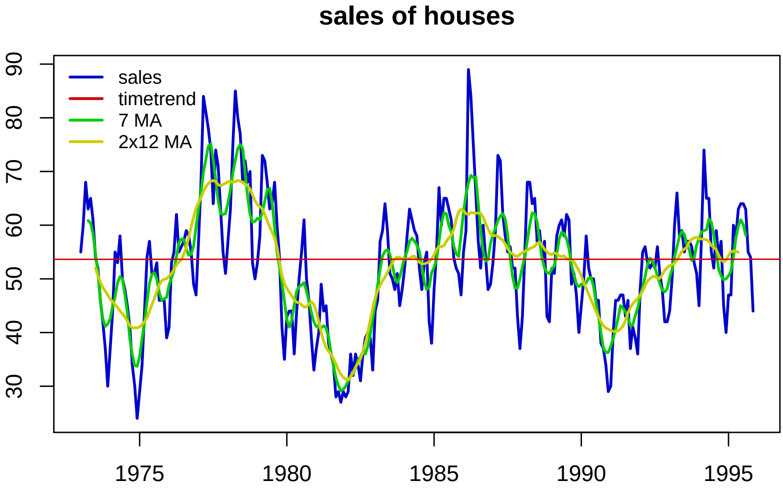
Centered $2 \times k$-MA moving average (even k)

Let k be an even order and $m = k/2$. Then the trend at time point t is

$$T_t = \frac{1}{k}(0.5 \cdot Y_{t-m} + \sum_{j=-m+1}^{m-1} Y_{t+j} + 0.5 \cdot Y_{t+m}).$$

- 2×4-MA for quarterly seasonality and monthly data
- 2×12-MA for annual seasonality and monthly data

Trend: centered moving average III



function `ma(...., order=k, centre=T)` in R.

Trend: double moving average I

Let $k = 3$. Then 3×3-MA is defined as:

$$
\begin{aligned}
T_2 &= \frac{1}{2}(Y_1 + Y_2 + Y_3) \\
T_3 &= \frac{1}{2}(Y_2 + Y_3 + Y_4) \\
T_4 &= \frac{1}{2}(Y_3 + Y_4 + Y_5) \\
T_3'' &= \frac{1}{3}(T_2 + T_3 + T_4) \\
&= \frac{1}{9}(Y_1 + 2 \cdot Y_2 + 3 \cdot Y_3 + 2 \cdot Y_4 + Y_5)
\end{aligned}
$$

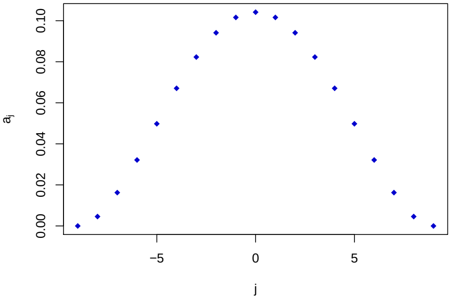
Trend: weighted moving averages I

In general

$$T_t = \sum_{j=-m}^{m} a_j Y_{t+j}, \text{ with } a_{-j} = a_j.$$

- Spencer's weights
- Henderson's weights
- $$a_j = \frac{Q(j,m)}{\sum_{i=-m}^{m} Q(i,m)}, \text{ with } Q(i,m) = \begin{cases} (1-(i/m)^2)^2, & \text{for } -m \le i \le m \\ 0, & \text{else} \end{cases}$$

Trend: weighted moving averages II



Trend: local polynomial regr. (LOESS) I

Problem: The classical regression assumes the same regression for all observations.

$$\sum_{t=1}^{\tau}(Y_t - b_0 - b_1 \cdot t)^2 \longrightarrow min, \text{ w.r.t. } b_0, b_1.$$

Aim: the regression is fitted just to a small fraction of data. To estimate the function in t_0 we solve

$$\sum_{t=1}^{\tau} w_t(t_0)(Y_t - b_0 - b_1 \cdot (t - t_0) - \frac{1}{2}b_2(t - t_0)^2)^2 \longrightarrow min, \text{ w.r.t. } b_0, b_1, b_2,$$

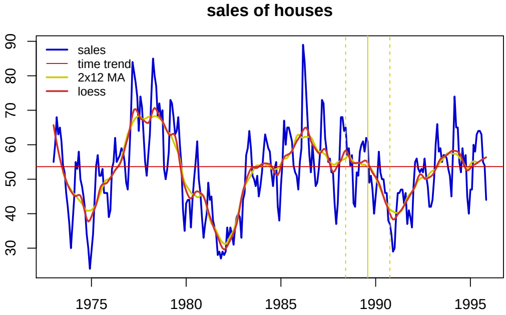
where

$$w_t(t_0) = W\left(\frac{t_i - t_0}{h}\right); \qquad W(u) = \begin{cases} (1 - |u|^3)^3, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

Trend: local polynomial regr. (LOESS) II

- h is the span (bandwidth) parameter which controls the smoothness
- here: 2nd order local polynomial, but other values are possible

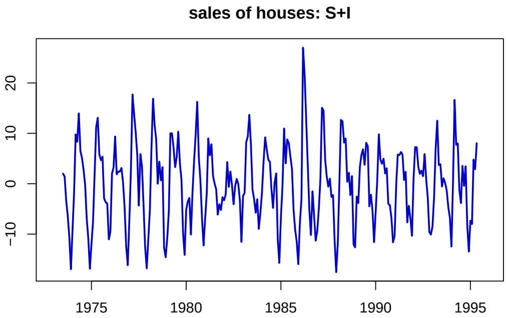
loess-function in R:



Seasonal component I

If the trend T_t is already extracted, then the seasonal and the irregular components are obtained from:

$$S_t + I_t = Y_t - T_t$$



Seasonal component II

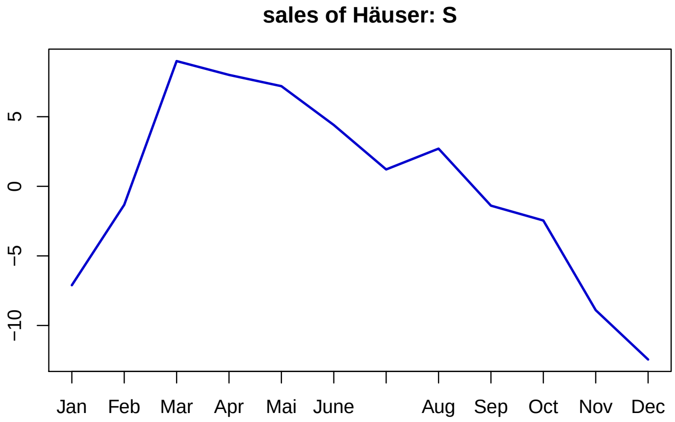
Idea: The seasonal component is constant from period to period, but the irregular component should be on average zero.

Let M^* be the number of full seasonal periods (number of years for annual seasonality) and m^* is the length of a seasonal period (e.g. 12). Then the seasonal component for month i is

$$\frac{1}{M^*}\sum_{j=1}^{M^*}(S_{(j-1)\cdot m^*+i} + I_{(j-1)\cdot m^*+i}),$$

i.e. the seasonal component for January is the average of all January values of $S_t + I_t$.

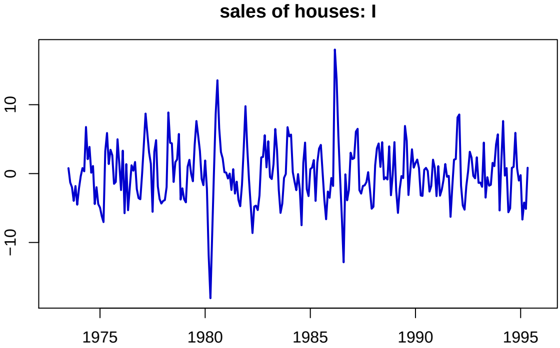
Seasonal component III

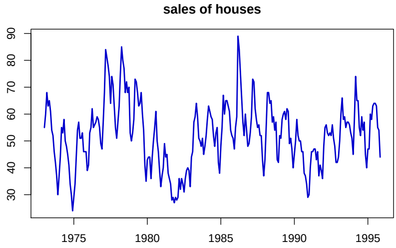


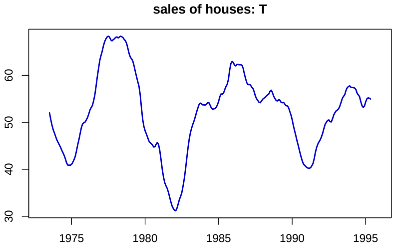
Irregular component

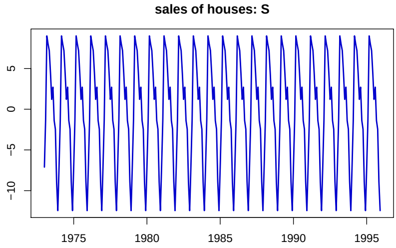
If T_t and S_t are extracted, then the irregular component equals:

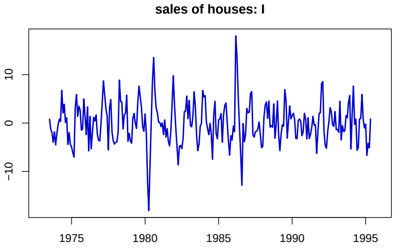
$$I_t = Y_t - T_t - S_t$$




sales of houses
30
40
50
60
70
80
90
1975
1980
1985
1990
1995


sales of houses: T
30
40
50
60
1975
1980
1985
1990
1995


sales of houses: S
−10
−5
0
5
1975
1980
1985
1990
1995


sales of houses: I
−10
0
10
1975
1980
1985
1990
1995

Part 5

Exponential smoothing

Naive forecasts I

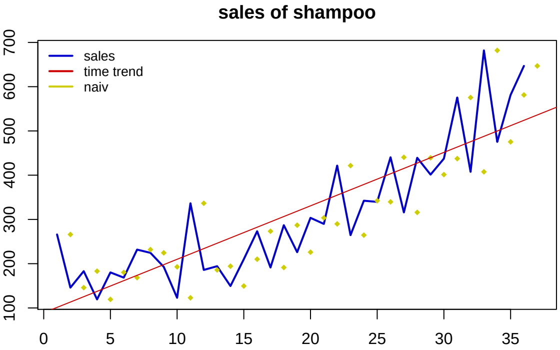
- Naive forecasts without/with taking the seasonality into account:

$$\hat{Y}_{t+1} = Y_t \qquad \text{without seasonality}$$
$$\hat{Y}_{t+1} = Y_{t+1-s} \qquad \text{with seasonality,}$$

where $s = 12$ indicates annual seasonality.

Naive forecasts II

Example: sales of shampoo



The red line corresponds to the time trend from a linear regression

$$\text{shampoo}_t = b_0 + b_1 \cdot t + \varepsilon_t.$$

Naive forecasts III

- Naive forecasts with absolute trend (*same-change* principle)

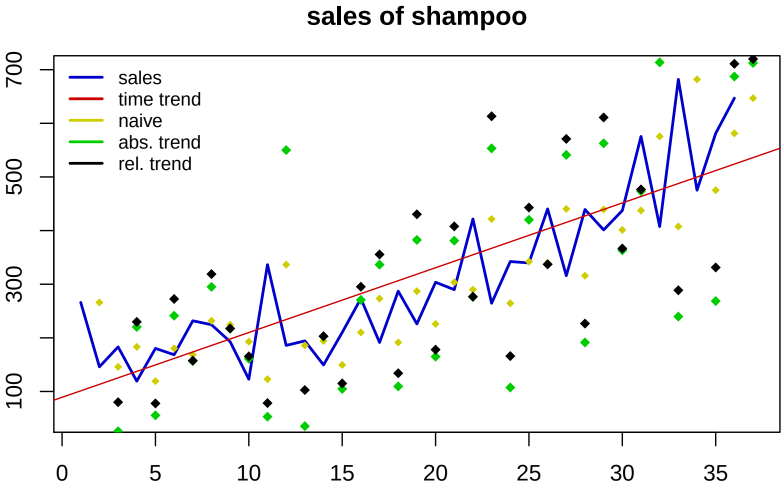
$$\hat{Y}_{t+1} = Y_t + (Y_t - Y_{t-1}).$$

- Naive forecasts with relative trend (*same-change* principle)

$$\hat{Y}_{t+1} = Y_t \cdot \frac{Y_t}{Y_{t-1}}.$$

- Naive forecasts with seasonality and absolute trend

$$\hat{Y}_{t+1} = Y_{t+1-s} + (Y_{t+1-s} - Y_{t+1-2s}).$$



	naive	abs. trend	rel. trend
MSE_1	11715.388	40484.661	57703.342
MAE_1	88.220	164.326	180.089
U_1	1.000	2.826	3.067

Forecasting with smoothing I

Now: Forecasting with smoothed historical observations as an alternative to the ARMA modelling.

Average as a forecast

$$\hat{Y}_{t+1} = \frac{1}{t}\sum_{i=1}^{t} Y_i$$

Forecasting with smoothing II

Recursive computation of forecasts:

$$\hat{Y}_{t+2} = \frac{1}{t+1} \sum_{i=1}^{t+1} Y_i = \frac{1}{t+1} (t\hat{Y}_{t+1} + Y_{t+1})$$

Note: average can be used for forecasting if the data has

- no trend and
- no seasonality.

Forecasting with moving averages

MAF(k) - *moving average forecast*

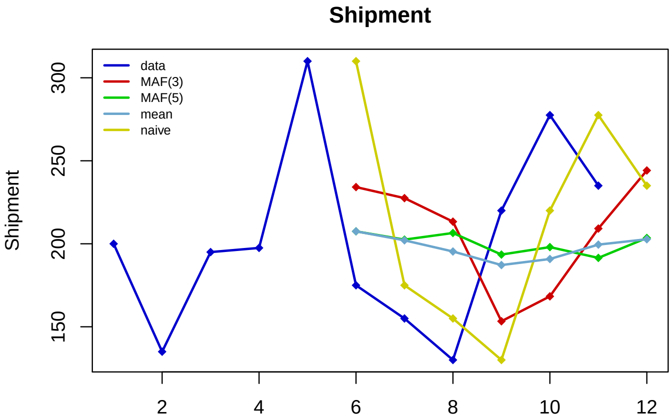
$$\hat{Y}_{t+1} = \frac{1}{k} \sum_{j=t-k+1}^{t} Y_j$$

$$\hat{Y}_{t+2} = \frac{1}{k} \sum_{j=t-k+2}^{t+1} Y_j = \hat{Y}_{t+1} + \frac{1}{k}(Y_{t+1} - Y_{t-k+1})$$

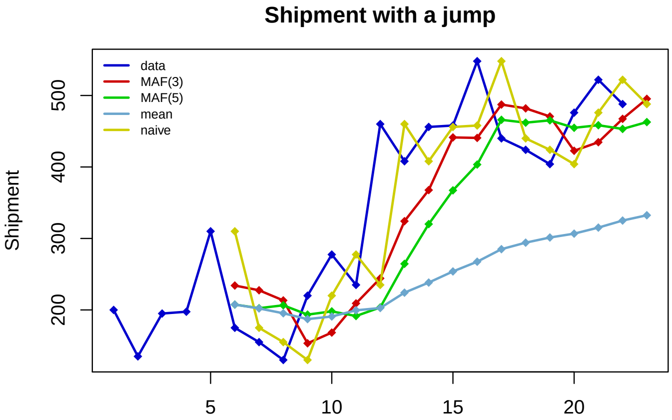
Advantages:

- we take only the recent k observations into account;
- the information set is constant.

Disadvantage: seasonality



	method			
	MAF(3)	MAF(5)	average	naive
MSE	5455.093	3013.250	2898.778	5410.417
MAE	69.444	51.000	49.987	61.667
Theil's U	1.127	0.810	0.788	1.000



	method			
	MAF(3)	MAF(5)	average	naive
MSE	7464.186	9625.844	25561.437	6880.912
MAE	74.235	77.765	139.409	63.706
Theil's U	1.059	1.065	1.349	1.000

Forecasting with exponential smoothing

Note: moving averages weight the historical obervations equally (with $1/k$)

Idea: the impact of past values should decrease.

Assumption: no trend and no seasonality.

EWMA(α)- *exponentially weighted moving average*

$$\hat{Y}_{t+1} = \alpha(Y_t - \hat{Y}_t) + \hat{Y}_t = \alpha Y_t + (1-\alpha)\hat{Y}_t$$

with $\alpha \in (0, 1]$.

$$\hat{Y}_{t+h} = \hat{Y}_{t+1} \text{ for } h > 1.$$

Note:

- EMWA forecasts have statistical optimality properties (Muth, JASA, 1960)
- The forecasts are easy to implement and are frequently used in practice; RiskMetrics approach to volatility forecasting in risk management

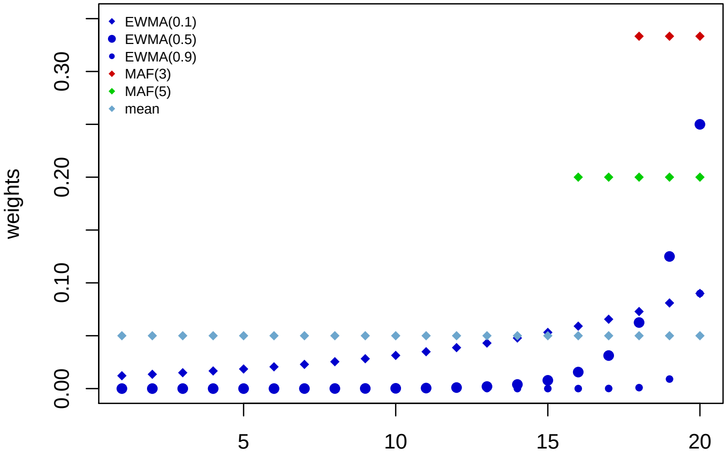
$$\begin{aligned}\hat{Y}_{t+1} &= \alpha Y_t + (1-\alpha)[\alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1}] \\ &\vdots \\ &= \alpha[Y_t + (1-\alpha)Y_{t-1} + \cdots + (1-\alpha)^{t-2}Y_2] + (1-\alpha)^{t-1}\hat{Y}_2,\end{aligned}$$

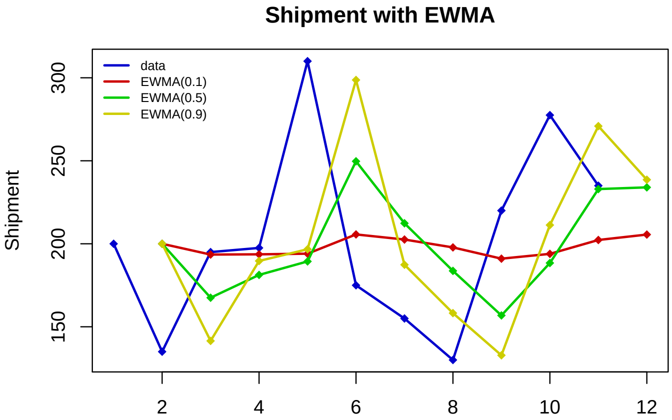
mit $\hat{Y}_2 = Y_1$ (initialisation).

Note: α plays the role of memory parameter:

- α close to zero $\rightsquigarrow$ historical values have strong impact on the forecast;
- α close to one $\rightsquigarrow$ current values have strong impact on the forecast;
- $\alpha = 1 \rightsquigarrow$ naive forecast.

The weight of Y_s in the EWMA(α) forecast for Y_{t+1} is: $\alpha(1-\alpha)^{t-s}$





	method			
	EWMA(0.1)	EWMA(0.5)	EWMA(0.9)	naive
MSE	3438.332	4347.237	5039.368	5295.000
MAE	47.758	56.937	61.318	61.000
Theil's U	0.809	0.922	0.982	1.000

Adaptive EWMA-forecasts

The EWMA-forecast can be optimized w.r.t. α-parameter, i.e. choose α with the smallest MSE-value.

Problem: in phases with strong/little changes or with/without trend different parameters may be optimal.

Solution: adaptive EWMA-forecasts, i.e. the parameter α can be changed adaptively.

Adaptive EWMA-forecast

$$
\begin{aligned}
\hat{Y}_{t+1} &= \alpha_t(Y_t - \hat{Y}_t) + \hat{Y}_t, \\
\alpha_{t+1} &= \left|\frac{A_t}{M_t}\right|, \\
A_t &= \beta(Y_t - \hat{Y}_t) + (1-\beta)A_{t-1} \\
M_t &= \beta|Y_t - \hat{Y}_t| + (1-\beta)M_{t-1}
\end{aligned}
$$

Idea:

- A_t is a smoothed forecast of the forecast error and M_t serves as normalizing factor
- If A_t is large, this implies that the last forecats were bad and the recent value should get more weight.
- If A_t is small, this implies that the forecasts are good and the historical values get more weight.
- Frequently there is a delay in the computation of α to avoid the impact of outliers.

Example: Initialisation:

$$\hat{Y}_2 = Y_1, \quad \alpha_2 = \alpha_3 = \alpha_4 = \beta = 0.2, \quad A_1 = M_1 = 0$$

Periode	Y_t	$\hat{Y}_t$	$Y_t - \hat{Y}_t$	A_t	M_t	α_t
1	200.0000					
2	135.0000	200.0000	-65.0000	-13.0000	13.0000	0.2000
3	195.0000	187.0000	8.0000	-8.8000	12.0000	0.2000
4	197.5000	188.6000	8.9000	-5.2600	11.3800	0.2000
5	310.0000	190.3800	119.6200	19.7160	33.0280	0.4622
6	175.0000	245.6701	-70.6701	1.6388	40.5564	0.5969
7	155.0000	203.4837	-48.4837	-8.3857	42.1419	0.0404
8	130.0000	201.5246	-71.5246	-21.0135	48.0184	0.1990
9	220.0000	187.2921	32.7079	-10.2692	44.9563	0.4376
10	277.5000	201.6055	75.8945	6.9635	51.1440	0.2284
11	235.0000	218.9418	16.0582	8.7825	44.1268	0.1362
12		221.1282				0.1990

Holt forecasts

Problem: MAF forecasts cannot be used for data with a trend.

Aim: a forecasting method, which uses a exponential smoothing and captures trends.

Idea: introduce a trend component, which reacts to the changes in the level of the observations.

Holt forecasts

$$
\begin{aligned}
L_t &= \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1}) && \text{Level} \\
T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} && \text{Trend} \\
\hat{Y}_{t+h} &= L_t + T_t h
\end{aligned}
$$

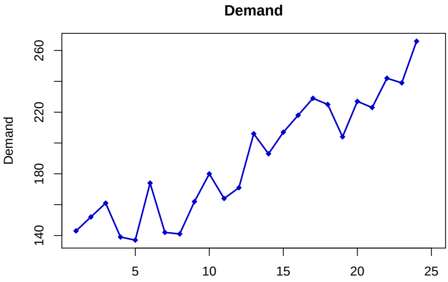
- Set $T_t = 0$ to obtain MAF forecasts.
- If the level changes, then there is a trend: the trend T_t consists of the smoothed changes in the level.
- The forecast consists of the level and the h-step-ahead forecast of the trend.

The smoothing parameters α and β can be determined by optimization:

$$MSE(\alpha, \beta) \longrightarrow min, \quad \text{w.r.t.} \quad \alpha, \beta.$$

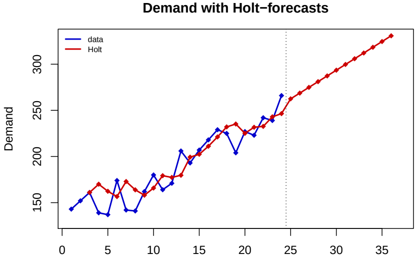
R: function `HoltWinters(....., gamma=F)`

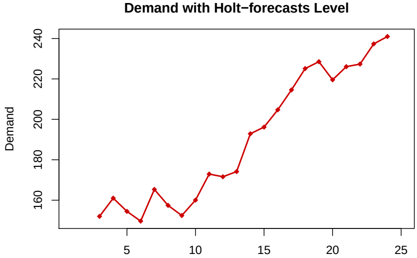
Example: monthly demand for a particular product.

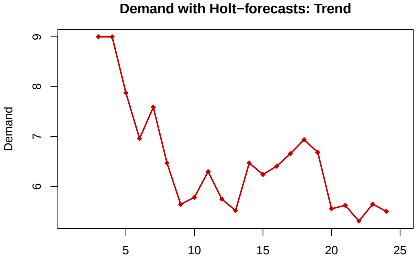


Optimal parameters: $\hat{\alpha} = 0.5011$, $\hat{\beta} = 0.0723$ with $MSE = 287.3911$

The parameters are used to determine the level, the trend and the forecasts.







$$\text{MSE} = 287.3911,$$
$$\text{MSE(EWMA)} = 311.7059$$

Holt-Winters forecasts

Problem: Holt method does not work for data with seasonality.

Aim: a forecasting method, which uses a exponential smoothing, captures trends and seasonality.

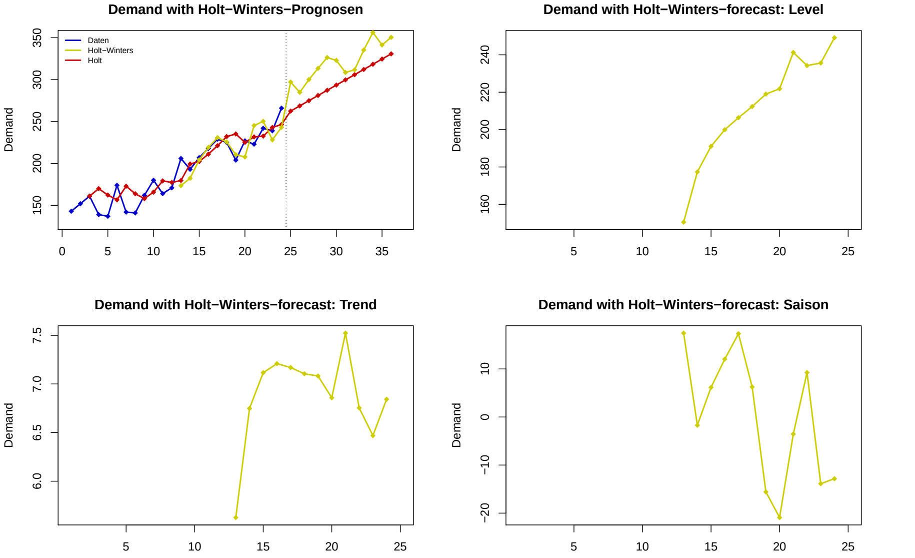
Idea: add a seasonal component

Holt-Winters forecasts

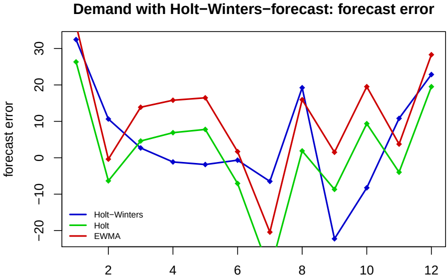
$$
\begin{aligned}
L_t &= \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1}) && \text{Level} \\
T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} && \text{Trend} \\
S_t &= \gamma(Y_t - L_t) + (1-\gamma)S_{t-s} && \text{Season} \\
\hat{Y}_{t+h} &= L_t + T_t h + S_{t-s+h}
\end{aligned}
$$

- The seasonal component is determined as a smoothed deviation of Y_t from the level L_t.
- The optimal parameters can be found by minimizing the MSE.

Example: $\alpha = 0.6541$, $\beta = 0.0528$, $\gamma = 0.1$



Comparison



	Holt-Winters	Holt	EWMA
MSE	232.7140	204.0641	326.5109
MAE	11.6099	11.1333	14.5121
Theil's U	0.7100	0.6226	0.9962192

Aim: comparison of EWMA and Holt forecasts.

$$d_t = (Y_t - \hat{Y}_t^{EWMA})^2 - (Y_t - \hat{Y}_t^{HW})^2.$$

H_0 : both models are equivalent.
H_1 : one model is better.

```
 360.27138   73.11708  -13.91563  -46.12191  -57.09005  -49.33493
-931.59911  366.44564  419.47051  -19.78895  100.99468  141.35047

-931.59911  -57.09005  -49.33493  -46.12191  -19.78895  -13.91563
  73.11708  100.99468  141.35047  360.27138  366.44564  419.47051
```

Sign test: R: `SIGN.test` from BSDA package.

$$T = \frac{2}{12} \cdot \sum_{i=1}^{12} (I(d_t > 0) - 0.5) = 0.$$

$\rightsquigarrow$ H_0 is not rejected.

Wilcoxon sign rank test: R: `wilcox.test`

The p-value of the test is $0.3804 > 0.05$. Thus H_0 cannot be rejected.

$\rightsquigarrow$ Both models are equally good!

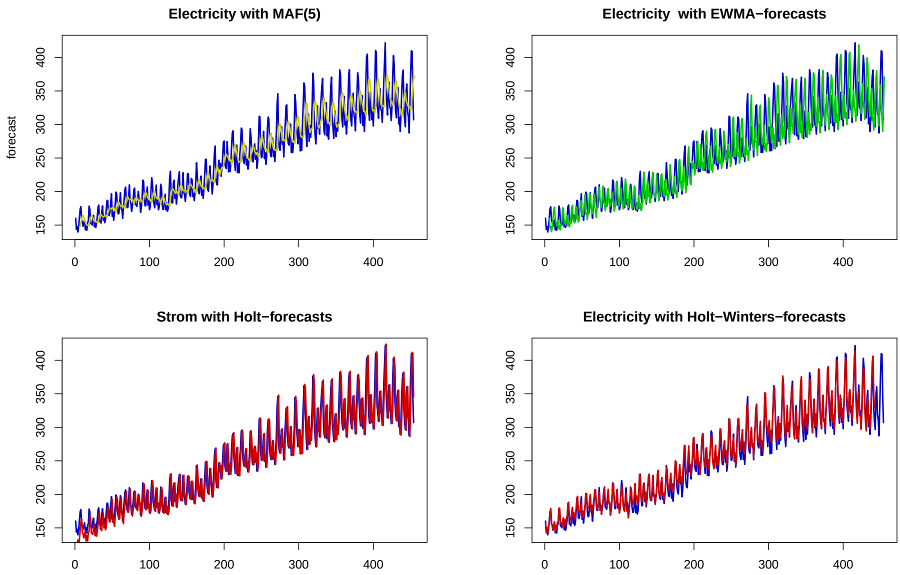
Note: small samples and the asymptotics is not reliable!

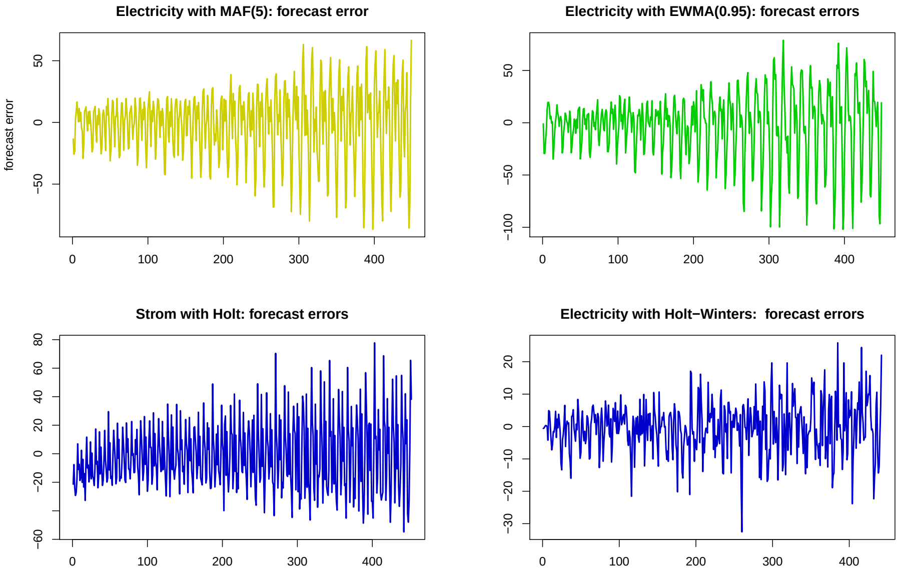
Example: electricity production, USA, monthly data, 01.1973-10.2010.

One-step ahead forecasts with

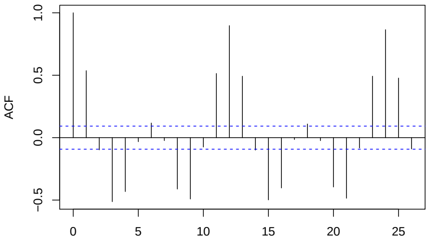
- MAF(5)
- EWMA with $\alpha = 0.95$
- Holt method with α=1 and $\beta = 0.02471944$
- Holt-Winters- method with α=0.2843972, $\beta = 0.006568855$ and $\gamma = 0.4586164$

	MAF(5)	EWMA	Holt	Holt-Winters
MSE	782.768	1044.776	562.967	63.361
MAE	21.341	23.458	19.501	6.045
Theil's U	1.173	1.312	1.055	0.356

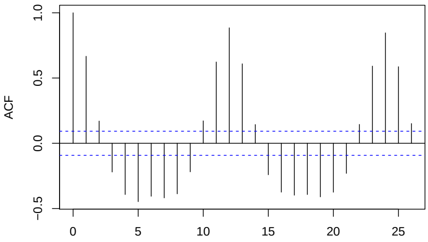

Electricity with MAF(5)
forecast
Electricity with EWMA−forecasts
Strom with Holt−forecasts
Electricity with Holt−Winters−forecasts
150
200
250
300
350
400
0
100
200
300
400


Electricity with MAF(5): forecast error
forecast error
Electricity with EWMA(0.95): forecast errors
Strom with Holt: forecast errors
Electricity with Holt−Winters: forecast errors

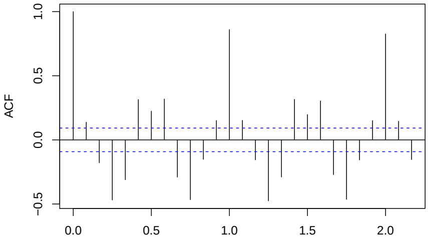
Electricity with MAF(5): ACF of the forecast errors



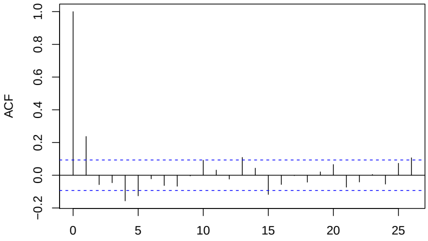
Electricity with EWMA(0.95): ACF of the forecast errors



Strom with Holt: ACF of the forecast errors



Electricity with Holt−Winters: ACF of the forecast errors

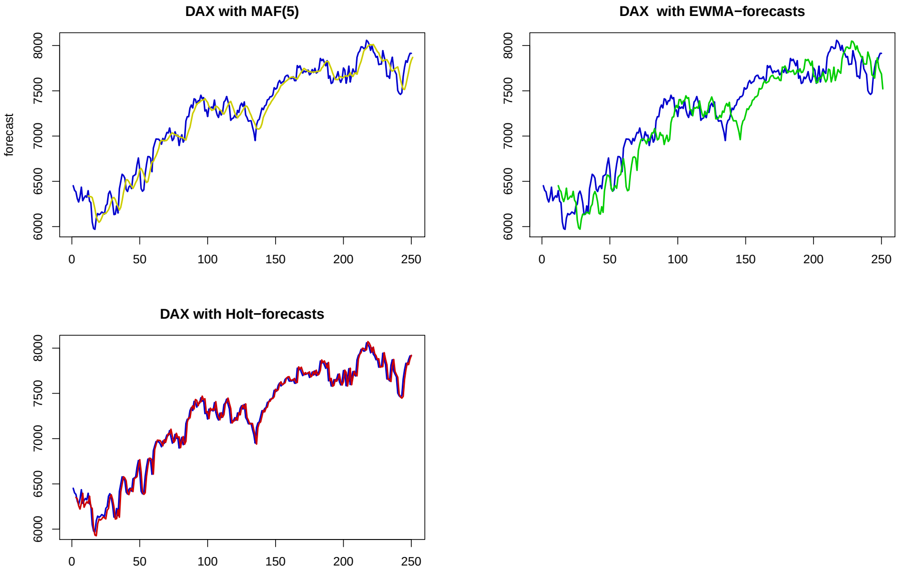


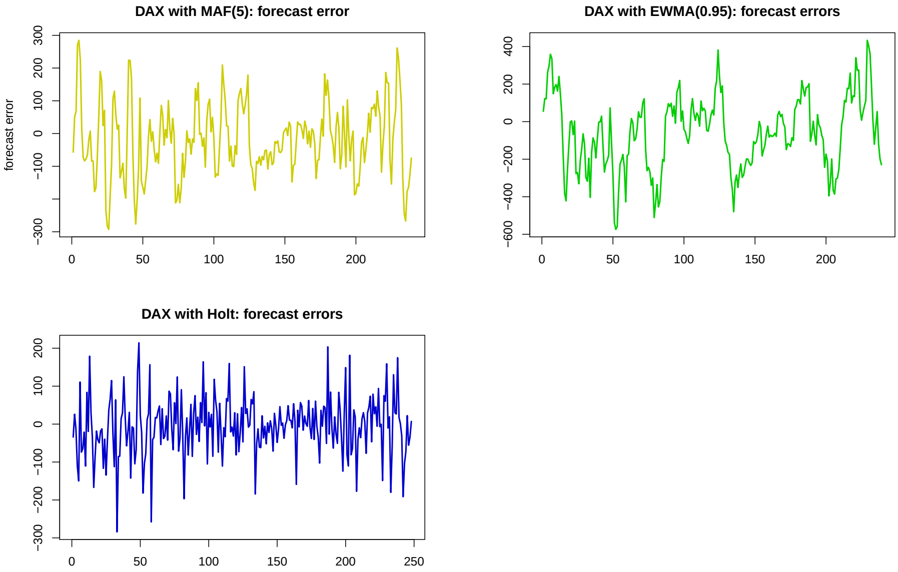
Example: DAX index, daily data, 01.05.2012-01.05.2013.

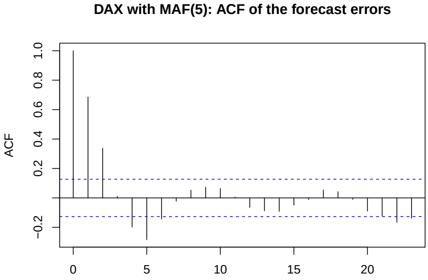
One-step ahead forecasts using

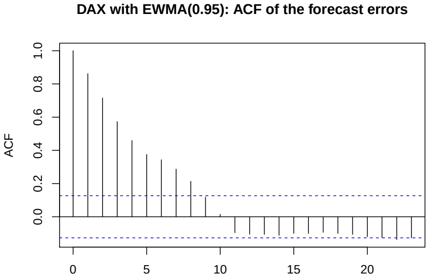
- MAF(5)
- EWMA with $\alpha = 0.95$
- Holt method with α=1 and $\beta = 0.04207012$

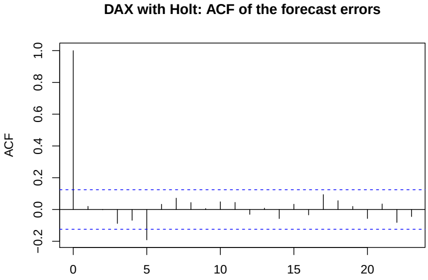
	MAF(5)	EWMA	Holt
MSE	12186.073	26397.249	5987.952
MAE	86.766	128.441	57.627
Theil's U	1.458	2.135	1.027


DAX with MAF(5)
forecast
6000
6500
7000
7500
8000
0
50
100
150
200
250
DAX with EWMA−forecasts
DAX with Holt−forecasts


DAX with MAF(5): forecast error
forecast error
−300
−100
0
100
200
300
0
50
100
150
200
DAX with EWMA(0.95): forecast errors
−600
−400
−200
0
200
400
0
50
100
150
200
DAX with Holt: forecast errors
−300
−200
−100
0
100
200
0
50
100
150
200
250


DAX with MAF(5): ACF of the forecast errors
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0
5
10
15
20


DAX with EWMA(0.95): ACF of the forecast errors
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20


DAX with Holt: ACF of the forecast errors
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20

Most popular generalizations

- STL decomposition: a seasonal-trend decomposition based on loess (Cleveland et al. 1990, Journal of Official Statistics)
- X-12-ARIMA: approach of the *U.S. Bureau of the Census*
- ETS: exponential smoothing state space model (Hyndman et al. 2002, International Journal of Forecasting)

STL decomposition

Advantages: highly resistant to outliers; any seasonal period; works even with missing values

- Inner loop

Step 1 Subtract the trend: $Y_t - T_t$

Step 2 de-trended observations for each month are smoothed by loess and glued for a complete seasonal TS

Step 3 $3 \times 12 \times 12$-MA and loess are applied to the preliminary S_t from Step 2

Step 4 The final S_t is estimated as the difference between the seasonal components in Step 3 and Step 2

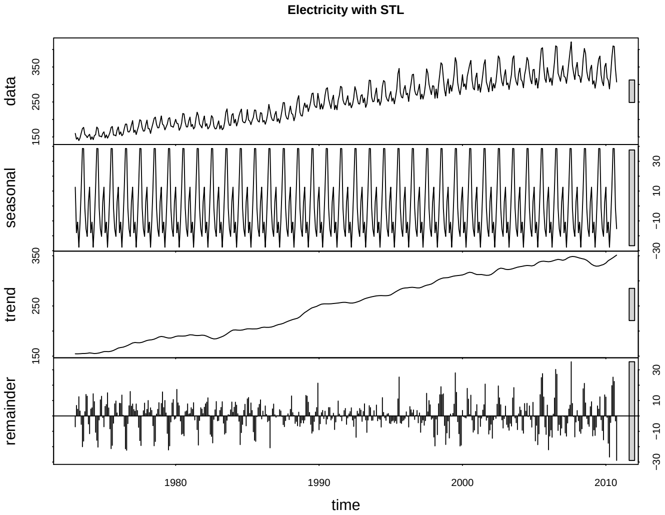
Step 5 Compute seasonally adjusted time series as $Y_t - S_t = T_t + I_t$

Step 6 Apply loess to $Y_t - S_t$ to obtain T_t

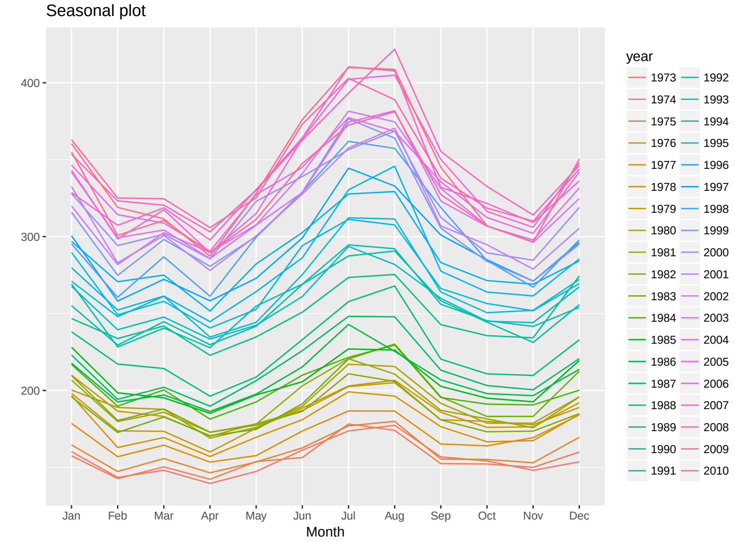
- Outer loop: repeat the inner loop by using the final trend component in Step 1
- Parameter estimation: two loess smoothing parameters in Steps 2 and 6
  - The 1st controls the variation of the season
  - The 2nd controls the variation of the trend

Example:

```
> elec.stl = stl(usmelec, s.window="periodic", robust=FALSE)
> plot(elec.stl)
```

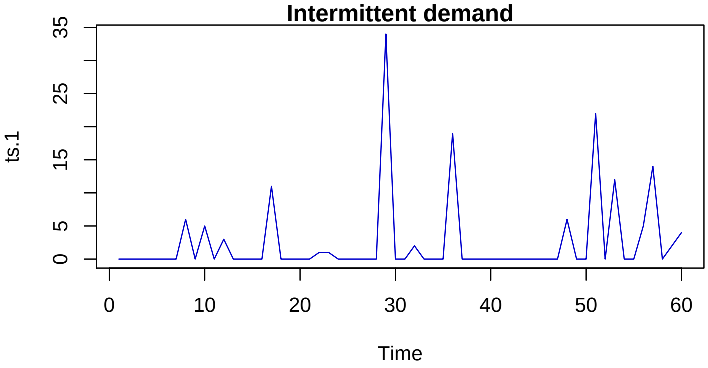


```
> ggseasonplot(usmelec, main="Seasonal plot")
```



Intermittent (sporadic) demand

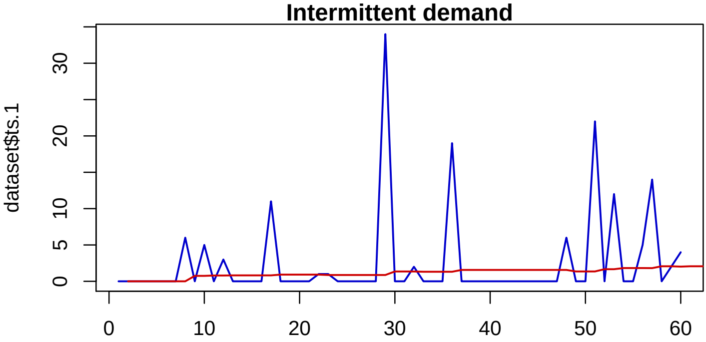
Problem: frequently the data is not systematic, but contains longer periods of zeros.



Croston's method: model the level of the non-zero TS and the waiting time till the next non-zero value separately.

Let q be the current number of consecutive zero-demand periods.

$$\begin{aligned} Z_{t+1} &= \alpha Y_t + (1-\alpha) Z_t \\ V_{t+1} &= \alpha q + (1-\alpha) V_t \\ \hat{Y}_{t+1} &= Z_{t+1}/V_{t+1} \end{aligned}$$



Notes:

- The exponential smoothing is easy to implement.
- It does not require any statistical model for the data.
- The optimal parameters can be found by minimizing loss functions.
- Only point forecasts are possible.