# Analysis of Time Series, L8

Rolf Larsson

Uppsala University

8 april 2025

# Today

- 3.7: Integrated models
- 3.8: Building ARIMA models

# Integrated models

Recall:

**Definition (2.4)**

The *backshift operator* is defined by

$$Bx_t = x_{t-1}.$$

For $k = 1, 2, \ldots$, $B^k x_t = x_{t-k}$.

**Definition (2.5)**

*Differences* of order $d$ are defined by

$$\nabla^d x_t = (1 - B)^d x_t.$$

Special cases:

- $\nabla^1 x_t = \nabla x_t = (1 - B)x_t = x_t - x_{t-1}$
- $\nabla^2 x_t = (1 - B)^2 x_t = x_t - 2x_{t-1} + x_{t-2}$

# Integrated models

- Two types of trends:
  1. Deterministic (example: $x_t = \beta_0 + \beta_1 t + w_t$)
  2. Stochastic (example: $x_t = x_{t-1} + w_t$)
- May be removed by differencing.
- Watch out for overdifferencing!
- Example: $x_t = x_{t-1} + w_t \Rightarrow \nabla x_t = w_t \Rightarrow \nabla^2 x_t = \nabla w_t = w_t - w_{t-1}$. $\mathrm{corr}(\nabla x_{t+1}, \nabla x_t) = 0$, $\mathrm{corr}(\nabla^2 x_{t+1}, \nabla^2 x_t) = -1/2 \neq 0$.
- Too much differencing may introduce extra autocorrelations and non invertibility!

# Integrated models

Example:

- Let
$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + y_t,$$
where $\beta_2 \neq 0$ and $y_t$ is stationary.
- How many differences are required to make $x_t$ stationary?

# Integrated models

## Definition (3.11)

A process $\{x_t\}$ is said to be ARIMA$(p, d, q)$ if

$$\nabla^d x_t = (1 - B)^d x_t$$

is ARMA$(p, q)$. We may write the model as

$$\phi(B)(1 - B)^d x_t = \theta(B) w_t.$$

If $E(\nabla^d x_t) = \mu$, we write the model as

$$\phi(B)(1 - B)^d x_t = \delta + \theta(B) w_t,$$

where

$$\delta = \mu(1 - \phi_1 - \ldots - \phi_p).$$

Why?

# Integrated models

Prediction (based on infinite past):

- Write the process on AR form to obtain the forecast.
- Write the process on MA form to obtain the prediction error.
- Example: Forecasting IMA(1,1)

$$\nabla x_t = w_t - \lambda w_{t-1}.$$

Leads to the Holt and Winter method:

$$\tilde{x}_{n+1} = (1-\lambda)x_n + \lambda \tilde{x}_n.$$

Why?

# Building ARIMA models

1. Check if the series is stationary. If not:
    - Remove trends by differencing.
    - Make the variance constant by transformations.
2. Identify an ARMA model by looking at
    - ACF
    - PACF
    - Information criteria (AIC, BIC...)
3. Estimate the model.
4. Check the model fit by performing residual diagnostics.
5. If the model does not fit well, start over from 1.

# Building ARIMA models

- Make the variance constant!
- Box-Cox transformation (chap.2, p.59)

$$y_t = \begin{cases} \frac{x_t^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log x_t, & \lambda = 0. \end{cases}$$

- Modified (not in book):

$$\tilde{y}_t = \mathrm{gm}(x)^{1-\lambda} y_t,$$

where $\mathrm{gm}(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$.

- By Taylor expansion (why?),

$$\frac{x_t^\lambda - 1}{\lambda} = \log x_t + O(\lambda).$$

# Building ARIMA models

Model identification via ACF and PACF:

| | AR($p$) | MA($q$) | ARMA($p, q$) |
|---|---|---|---|
| ACF | Tails off | Cuts off after lag $q$ | Tails off |
| PACF | Cuts off after lag $p$ | Tails off | Tails off |

# Building ARIMA models

Let $\hat{\sigma}_k^2 = SSE_k/n$, where $SSE_k$ is the residual sum of squares and $k$ is the number of parameters in the model.

**Definition (Akaike Information Criterion)**

$$\text{AIC} = \log(\hat{\sigma}_k^2) + \frac{2k}{n}.$$

In R: $\text{AIC} = -2\log(L) + 2k$.

**Definition (Bayesian Information Criterion)**

$$\text{BIC} = \log(\hat{\sigma}_k^2) + \frac{k\log(n)}{n}.$$

- BIC is recommended for large samples, AIC for small samples.
- It is *not recommended* to compare AIC, BIC between ARIMA models with different $d$, or for different transformations.

# Building ARIMA models

- Test $H_0$: The residuals are white noise vs $H_1$: $\neg H_0$.
- Ljung-Box-Pierce Q statistic (p.139)

$$Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}_e^2(h)}{n-h},$$

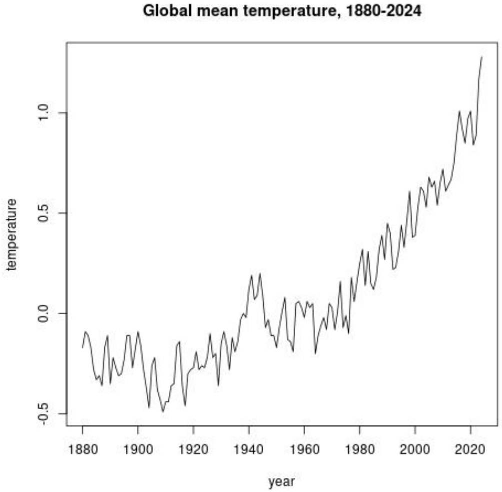
where $\hat{\rho}_e(h)$ are the estimated autocorrelations of the residuals.

- Asymptotically, $Q \sim \chi^2_{H-p-q}$.

# Building ARIMA models

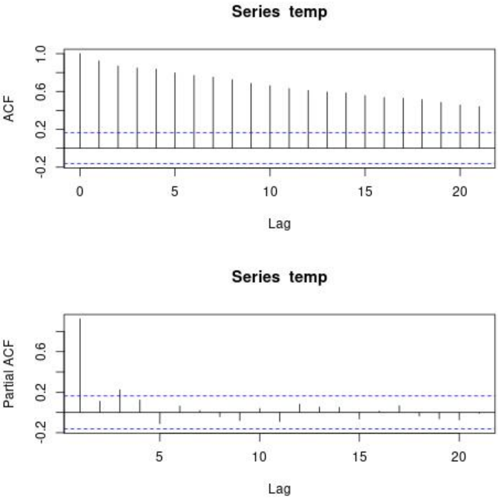
1. Check if the series is stationary. If not:
    - Remove trends by differencing.
    - Make the variance constant by transformations.
2. Identify an ARMA model by looking at
    - ACF
    - PACF
    - Information criteria (AIC, BIC...)
3. Estimate the model.
4. Check the model fit by performing residual diagnostics.
5. If the model does not fit well, start over from 1.

# Building ARIMA models

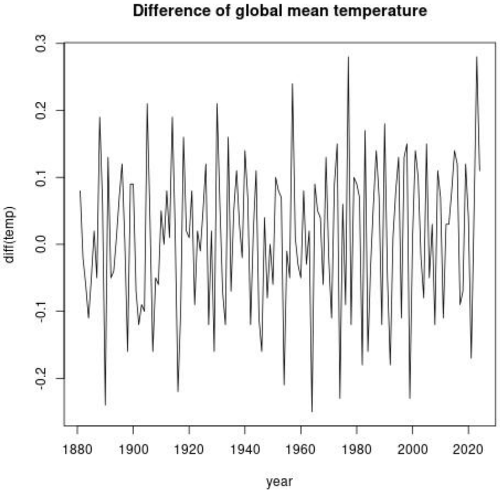


# Building ARIMA models

Global mean temperature, ACF and PACF (typical signs of a trend):

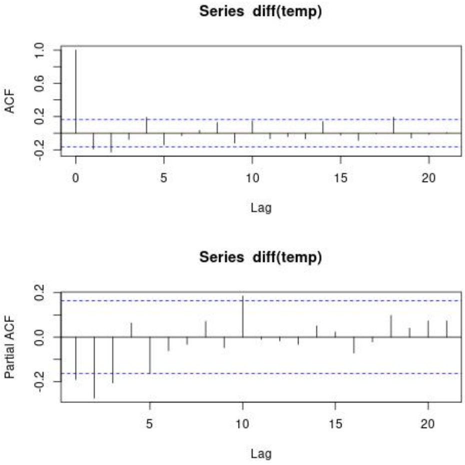


# Building ARIMA models



# Building ARIMA models

Difference of global mean temperature, ACF (cuts off after lag 2) and PACF (tails off?):



In R, try MA(2) for differences (ARIMA(0,1,2)):

```
> arima(temp,order=c(0,1,2))

Call:
arima(x = temp, order = c(0, 1, 2))

Coefficients:
          ma1      ma2
      -0.3010  -0.2118
s.e.   0.0783   0.0702

sigma^2 estimated as 0.01106:  log likelihood = 119.84,
aic = -233.68
```

Among many models, ARIMA(3,1,1) gave the smallest AIC.

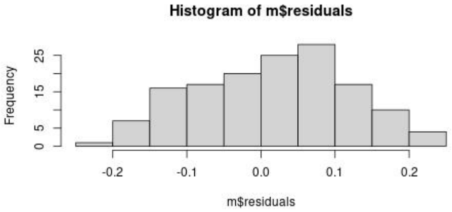
```
> arima(temp,order=c(3,1,1))

Call:
arima(x = temp, order = c(3, 1, 1))

Coefficients:
          ar1      ar2      ar3     ma1
      -1.0247  -0.5030  -0.3750  0.7904
s.e.   0.1566   0.1125   0.0795  0.1607

sigma^2 estimated as 0.01053:  log likelihood = 123.27,
aic = -236.55
```

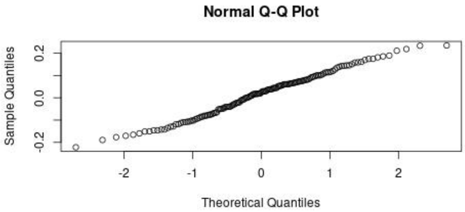
Check: all coefficients are significant, i.e. outside the two times s.e. bound.

Residual diagnostics, ARIMA(3,1,1):

```
> m=arima(temp,order=c(3,1,1))
> par(mfrow=c(2,1))
> hist(m$residuals)
> qqnorm(m$residuals)
> dev.off()
> tsdiag(m)
```

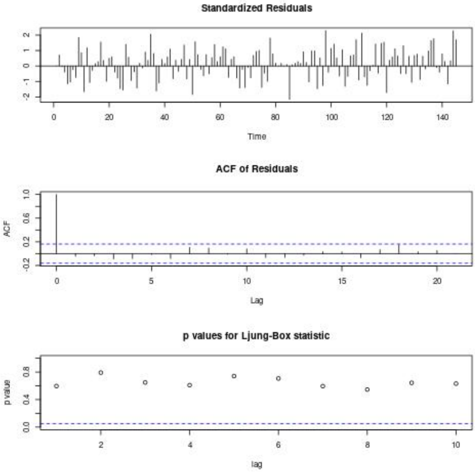
# Building ARIMA models



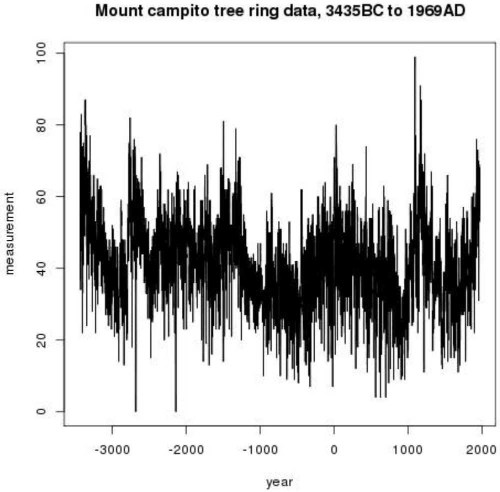

Normal distribution ok (?)

# Building ARIMA models



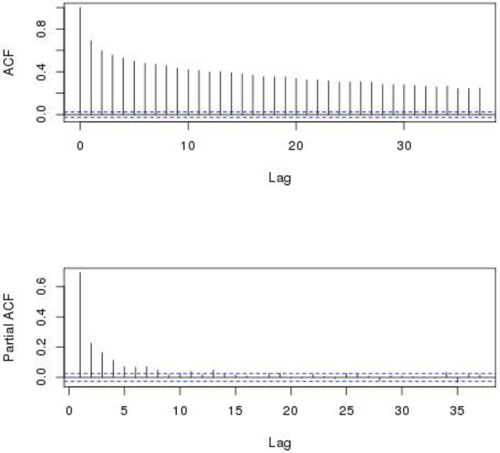

No sign of autocorrelations in the residuals!

# Building ARIMA models

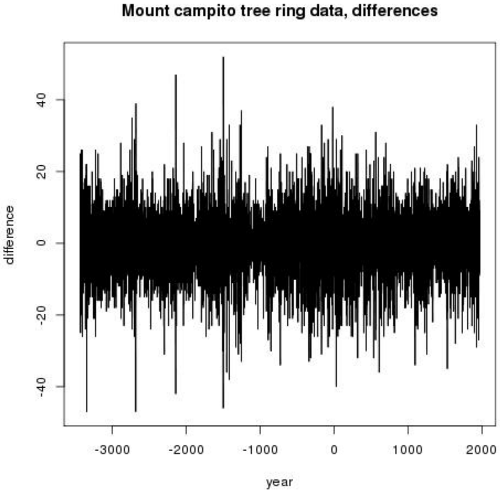


# Building ARIMA models

Mount Campito, ACF and PACF. Maybe not stationary?

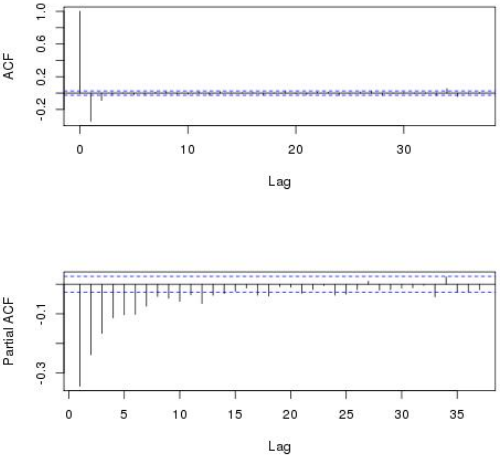


# Building ARIMA models



# Building ARIMA models

Mount Campito differences, ACF and PACF. MA(2)?



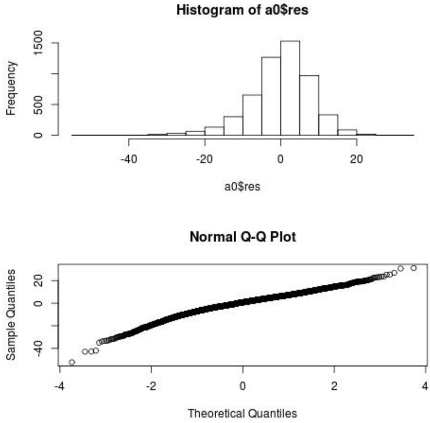
MA(2) for differences, estimation in R (intercept not significant):

```
> dy=diff(y)
> a=arima(dy,order=c(0,0,2));a

Call:
arima(x = dy, order = c(0, 0, 2))

Coefficients:
          ma1      ma2  intercept
      -0.5449  -0.1921     0.0009
s.e.   0.0130   0.0140     0.0289

sigma^2 estimated as 65.34:  log likelihood = -18961.64,
aic = 37931.27
```

MA(2) without constant (note: AIC two units lower):

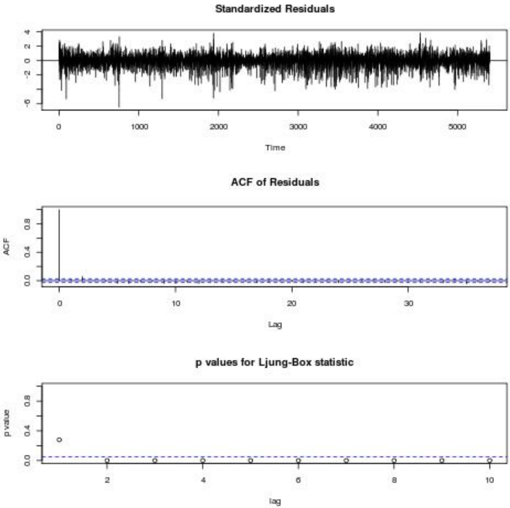
```
> a0=arima(dy,order=c(0,0,2),include.mean=FALSE);a0

Call:
arima(x = dy, order = c(0, 0, 2), include.mean = FALSE)

Coefficients:
          ma1      ma2
      -0.5449  -0.1921
s.e.   0.0130   0.0140

sigma^2 estimated as 65.34:  log likelihood = -18961.64,
aic = 37929.27
```

```
> par(mfrow=c(2,1))
> hist(a0$res)
> qqnorm(a0$res)
```




Skew distribution!?

```
> tsdiag(a0)
```




Small p values, i.e. significant autocorrelations.

# Building ARIMA models

MountCampito: Find the ARMA($p, q$) model for differences without constant with the smallest AIC:

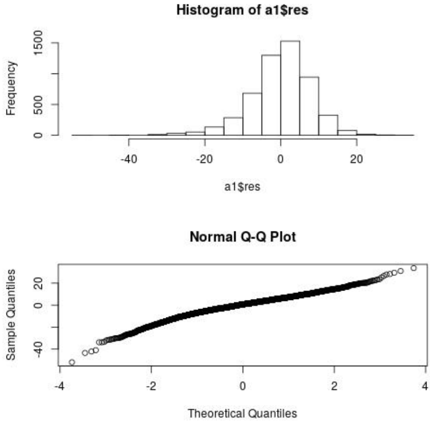
| $p$ | $q$ | AIC |
|---|---|---|
| 0 | 0 | 39339.2 |
| 0 | 1 | 38107.1 |
| 1 | 0 | 38657.1 |
| 0 | 2 | 37929.3 |
| 1 | 1 | 37885.3 |
| 2 | 0 | 38341.7 |
| 0 | 3 | 37901.2 |
| 1 | 2 | 37854.4 |
| 2 | 1 | 37872.9 |
| 3 | 0 | 38191.7 |
| 0 | 4 | 37887.8 |
| 1 | 3 | 37818.9 |
| 2 | 2 | **37815.9** |
| 3 | 1 | 37851.8 |
| 4 | 0 | 38123.2 |
| 2 | 3 | 37826.7 |
| 3 | 2 | 37818.0 |
| 3 | 3 | 37819.8 |

Try ARMA$(2,2)$:

```
> a1=arima(dy,order=c(2,0,2),include.mean=FALSE);a1

Call:
arima(x = dy, order = c(2, 0, 2), include.mean = FALSE)

Coefficients:
         ar1      ar2      ma1     ma2
      1.1512  -0.2216  -1.7007  0.7059
s.e.  0.0367   0.0260   0.0322  0.0313

sigma^2 estimated as 63.93:  log likelihood = -18902.95,
aic = 37815.9
```

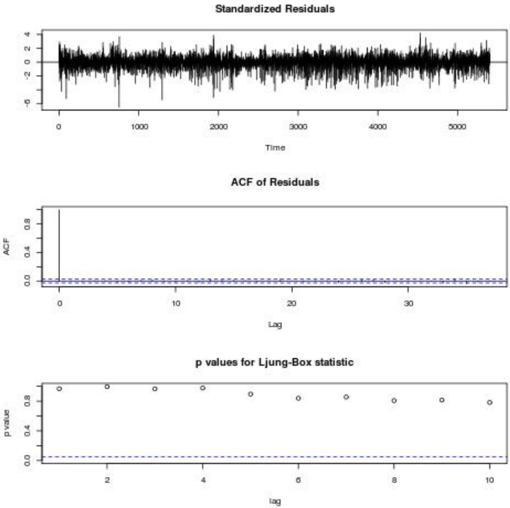
Observe: The invertibility condition $\theta_2 - \theta_1 < 1$ is not satisfied!

```
> par(mfrow=c(2,1))
> hist(a1$res)
> qqnorm(a1$res)
```



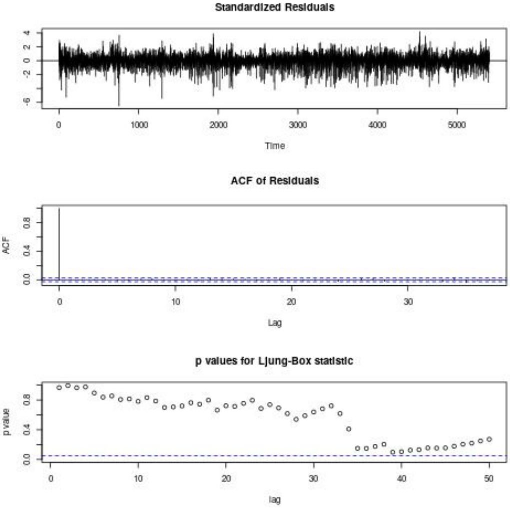

Still a bit skew!

```
> tsdiag(a1)
```




Better on autocorrelations, but try more lags for Ljung-Box!

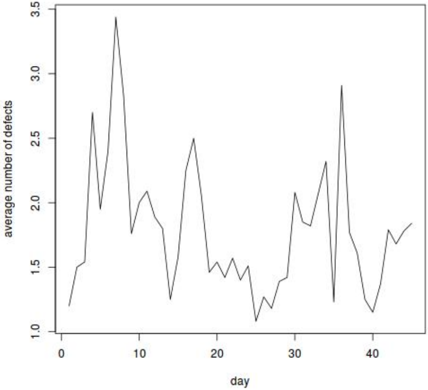
```
> tsdiag(a1,50)
```




Ljung-Box a bit suspicious here (p values for high lags close to 0.05)...

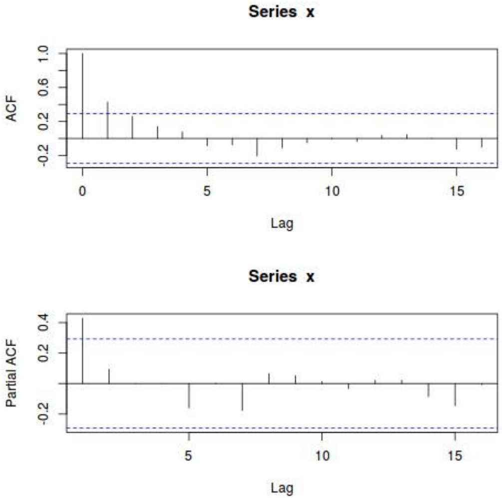
# Building ARIMA models

Daily average number of defects per truck found in the final inspection at the end of the assembly line of a truck manufacturing plant.
(Looks stationary.)



# Building ARIMA models

Trucks, ACF (tails off) and PACF (cuts off after lag 1). AR(1)?

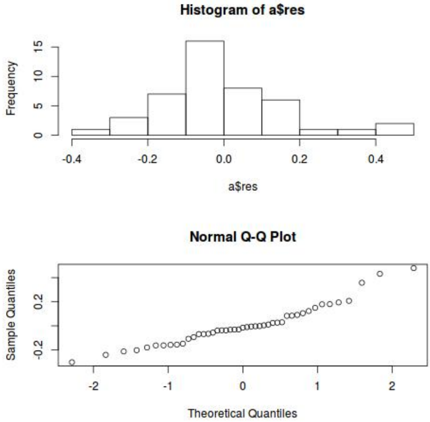


Try AR(1):

```
> a=arima(x,order=c(1,0,0));a

Call:
arima(x = x, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.4322     1.7799
s.e.  0.1340     0.1189

sigma^2 estimated as 0.2118:  log likelihood = -29.04,
aic = 64.07
```

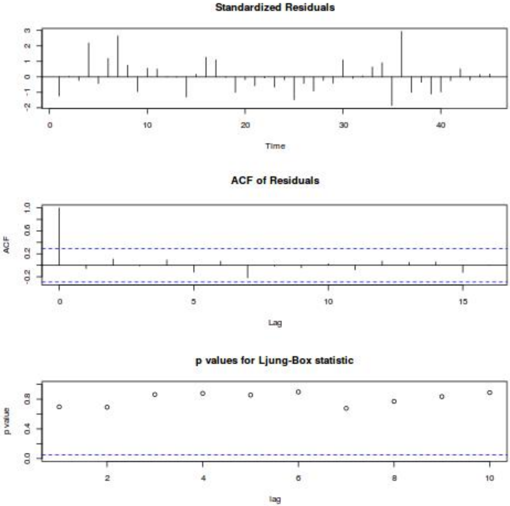
Significant coefficients!

```
> par(mfrow=c(2,1))
> hist(a$res)
> qqnorm(a$res)
```



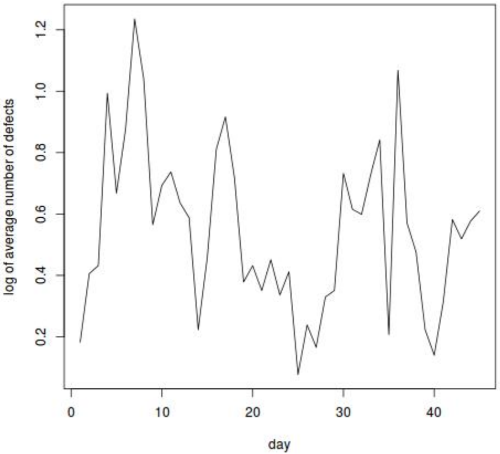

Maybe some outliers to the right?

```
> tsdiag(a)
```


Standardized Residuals
Time
ACF of Residuals
ACF
Lag
p values for Ljung-Box statistic
p value
lag

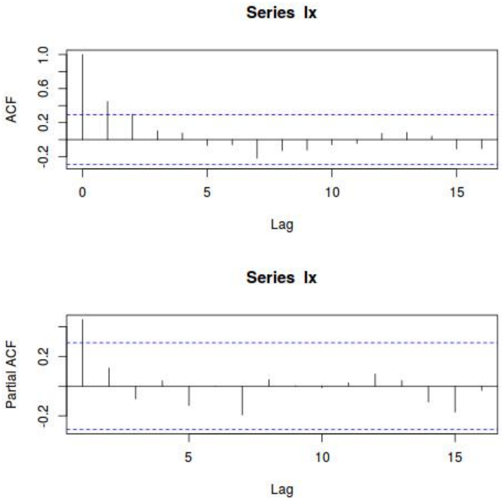
# Building ARIMA models

Take logarithms (to get a smaller effect of outliers):



# Building ARIMA models

## Trucks in logs, ACF and PACF

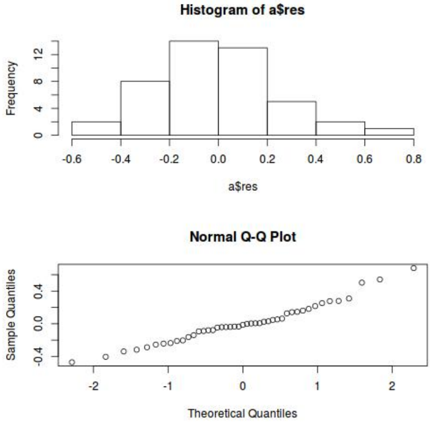


Try AR(1):

```
> lx=log(x)
> a=arima(lx,order=c(1,0,0));a

Call:
arima(x = lx, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.4582     0.5391
s.e.  0.1330     0.0641

sigma^2 estimated as 0.05625:  log likelihood = 0.78,
aic = 4.43
```

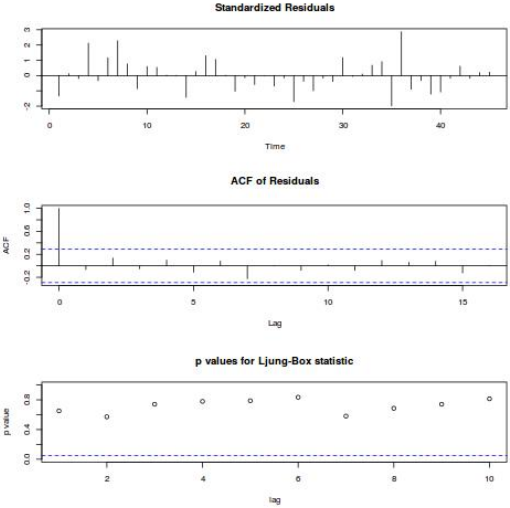
Do not compare AIC to the previous value!

```
> par(mfrow=c(2,1))
> hist(a$res)
> qqnorm(a$res)
```




Less pronounced outliers now.

```
> tsdiag(a)
```


Standardized Residuals
Time
ACF of Residuals
ACF
Lag
p values for Ljung-Box statistic
p value
lag

# News of today

- Removing trends by differencing
- Model building:
    - Transformation
    - Identification
    - Estimation
    - Diagnostics