Applied Time Series Analysis
FS 2012 – Week 07

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

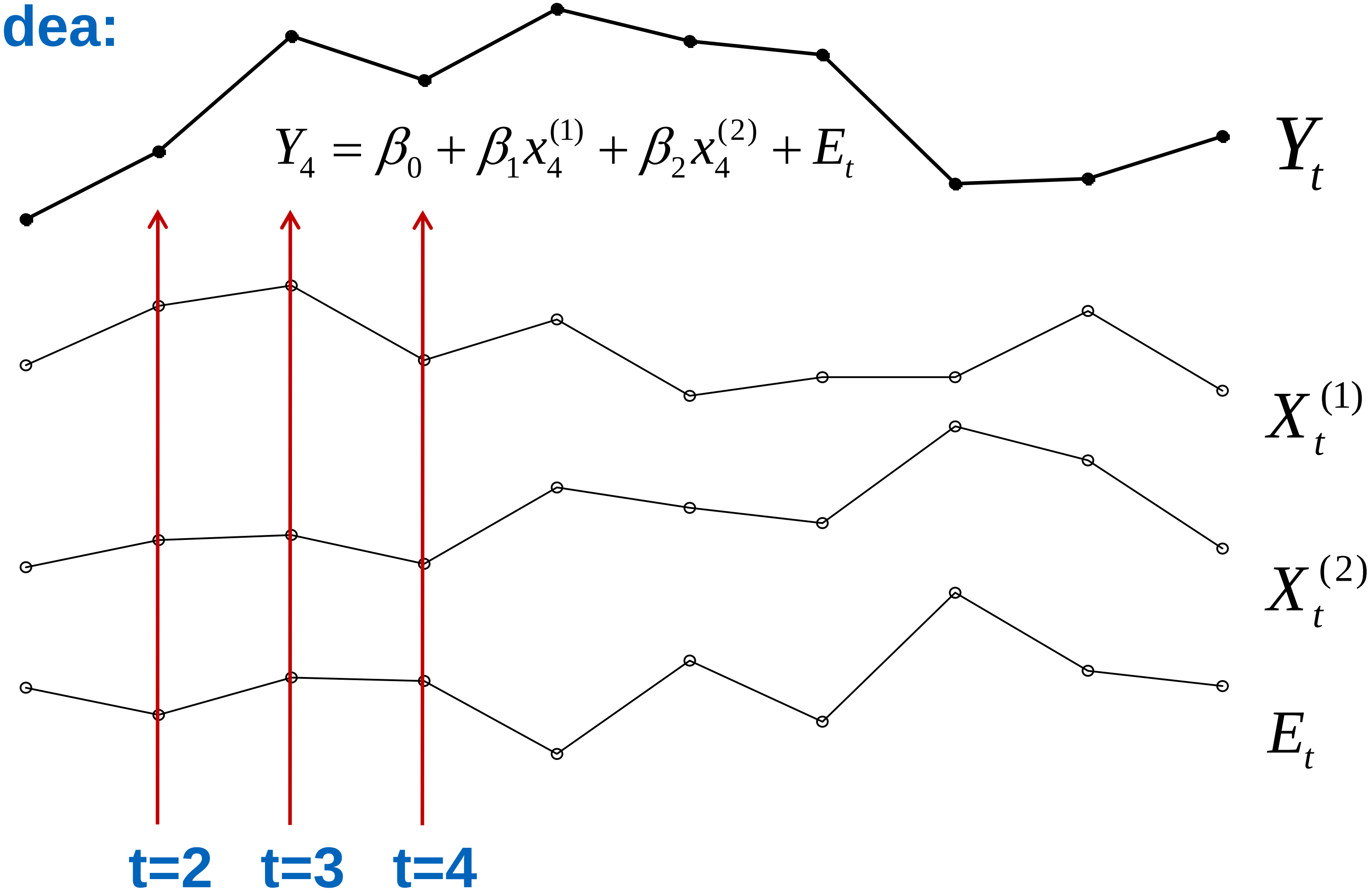
ETH Zürich, April 2, 2012

Applied Time Series Analysis
FS 2012 – Week 07

Time Series Regression

Idea:

$$Y_4 = \beta_0 + \beta_1 x_4^{(1)} + \beta_2 x_4^{(2)} + E_t$$



The Setup

- There is a response time series Y_t
- There is one or several explanatory/descriptive time series $x_t^1,\ldots,x_t^p$
- The goal is to infer the relation between x and Y, i.e. the β_j
- As long as the error series E_t is i.i.d, the usual regression setup with LS-estimates is perfectly fine

→ **Caution and specific procedures are required if the errors are correlated!**

Dealing with Correlated Errors

- In case of time series regression, the error term E_t is usually correlated and not i.i.d.

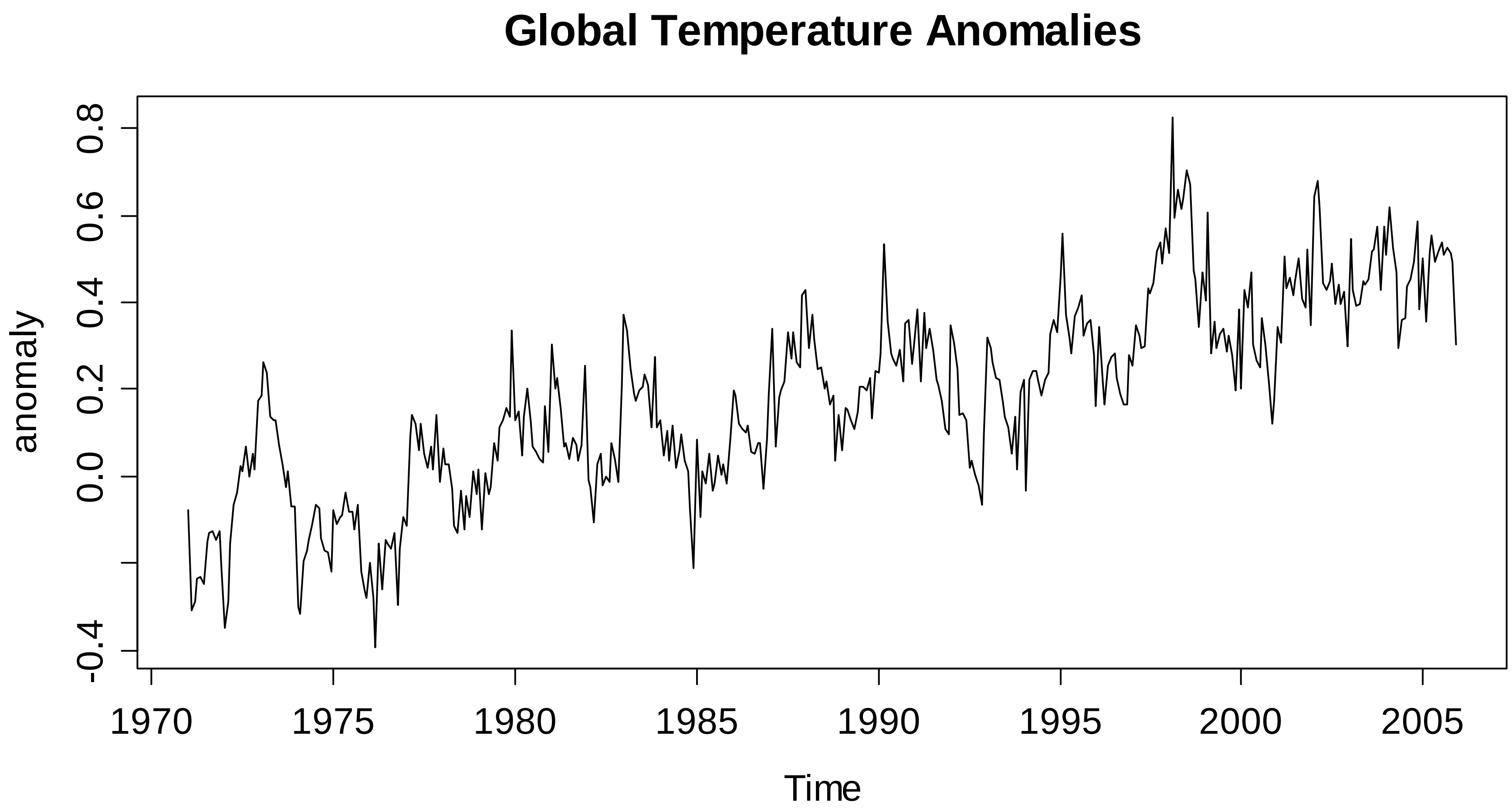
- Then, the estimated β_j are still unbiased, but the usual LS-procedure is no longer efficient and the standard errors can be grossly wrong

- There are procedures that correct for correlated errors:
  - **Cochrane-Orcutt-Method**
  - **Generalized Least Squares**

- **They must be applied in case of correlated errors!**

Applied Time Series Analysis
FS 2012 – Week 07

Example 1: Global Temperature



Applied Time Series Analysis
FS 2012 – Week 07

Example 1: Global Temperature

Temperature = Trend + Seasonality + Remainder

$$Y_t = \beta_0 + \beta_1 t + \beta_2 \cdot 1_{[month="Feb"]} + \ldots + \beta_{12} \cdot 1_{[month="Dec"]} + E_t,$$

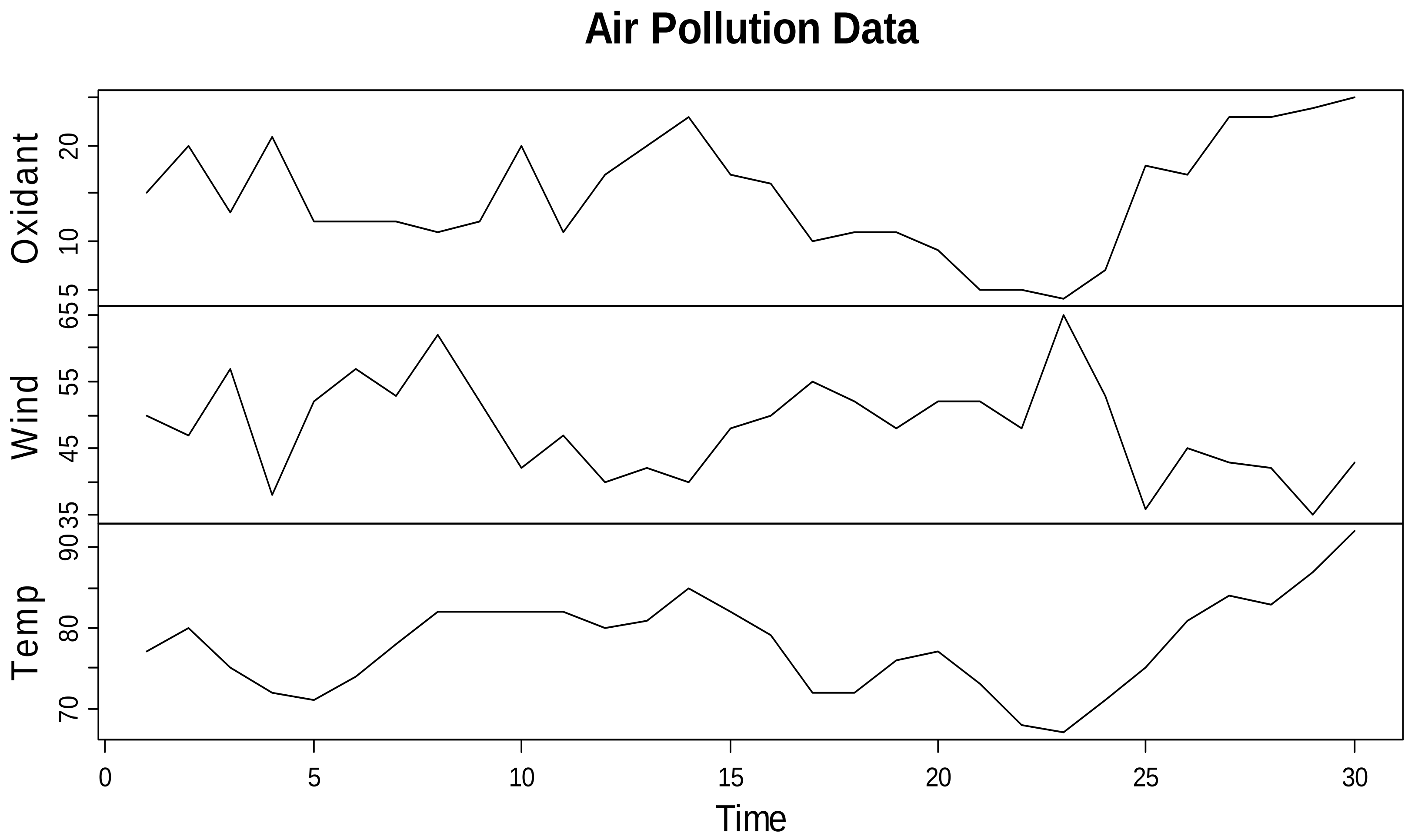
→ Recordings from 1971 to 2005, $n = 420$

→ The remainder term is usually a stationary time series, thus it would not be surprising if the regression model features correlated errors.

→ The applied question which is of importance here is whether there is a significant trend, and a significant seasonal variation

Applied Time Series Analysis
FS 2012 – Week 07

Example 2: Air Pollution



Applied Time Series Analysis
FS 2012 – Week 07

Example 2: Air Pollution

Oxidant = Wind + Temperature + Error

$$Y_t = \beta_0 + \beta_1 x_t^1 + \beta_2 x_t^2 + E_t$$

→ Recordings from 30 consecutive days, $n = 30$

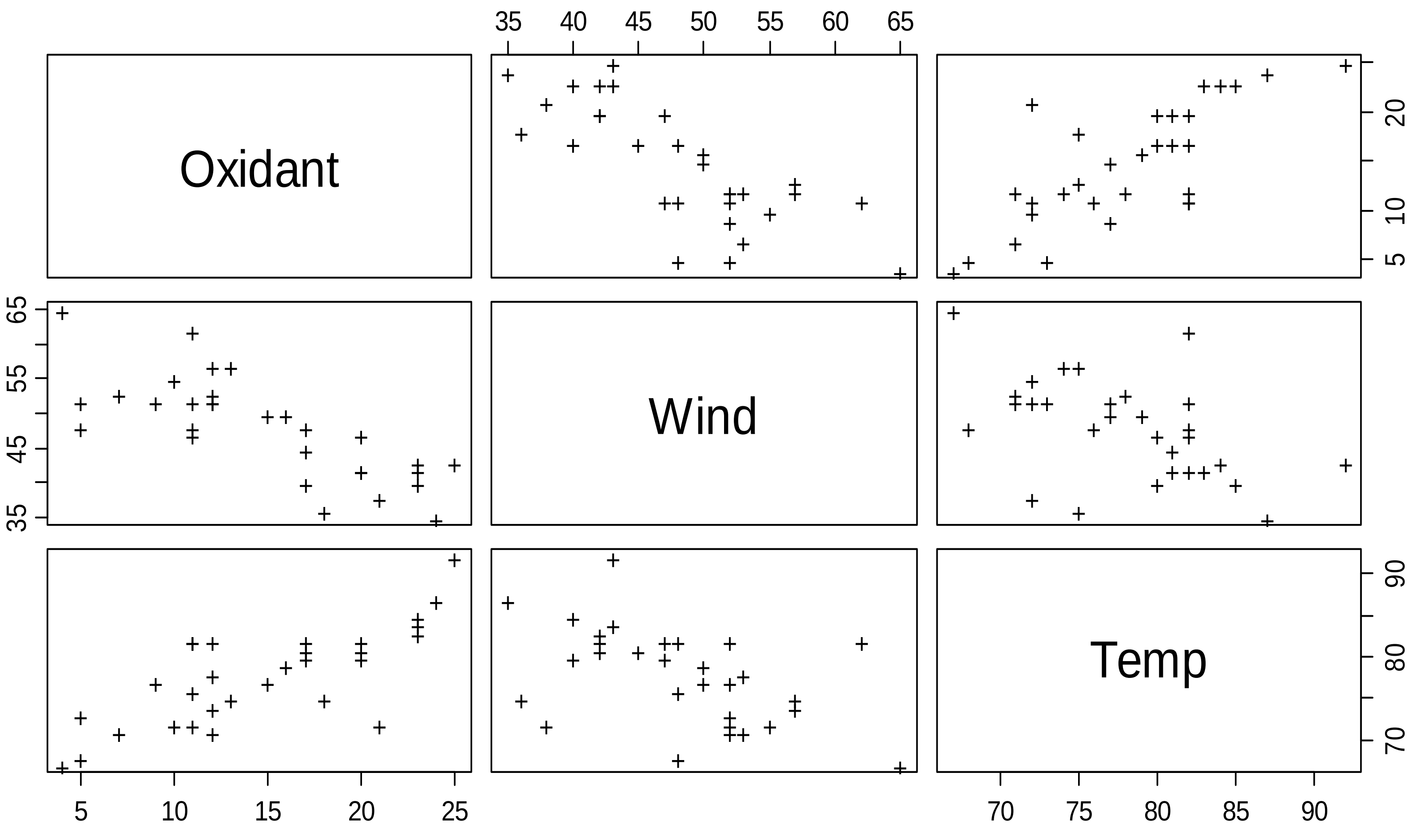
→ The data are from the Los Angeles basin, USA

→ The pollutant level is influence by both wind and temperature, plus some more, unobserved variables.

→ It is well conceivable that there is "day-to-day memory" in the pollutant levels, i.e. there are correlated errros.

Applied Time Series Analysis
FS 2012 – Week 07

Example 2: Air Pollution



Time Series Regression Model

$$Y_t = \beta_0 + \beta_1 x_t^{(1)} + \ldots + \beta_q x_t^{(q)} + E_t$$

- $t = 1, \ldots, N$
- no feedback from Y_t onto the predictors (i.e. input series)
- E_t are independent from $x_s^{(j)}$ for all j and all s, t
- E_t (generally) are dependent (e.g. an ARMA(p,q)-process)

Facts When Using Least Squares

In case of correlated errors, the effect on point estimates is:

- the estimated coefficients $\beta_1, ..., \beta_q$ are unbiased
- the estimates are no longer optimal: $Var\left(\hat{\beta}_j\right) > \min_* Var\left(\hat{\beta}_j^*\right)$

Important is the effect on the standard errors of the estimates:

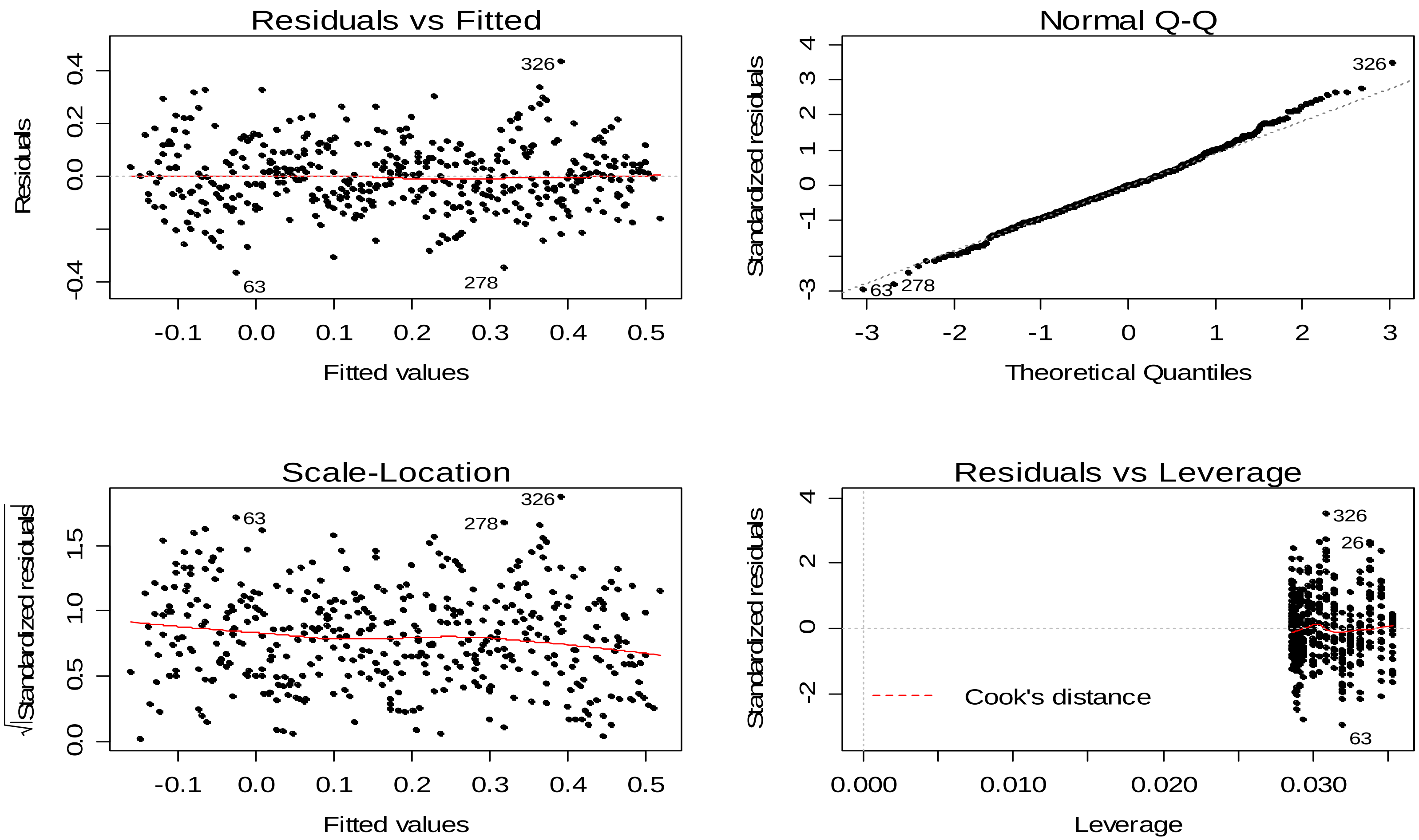
- $V\hat{a}r\left(\hat{\beta}_j\right)$ can be grossly wrong!
- often, the standard errors are underestimated
- too small CIs & spuriously significant results

Zurich University
of Applied Sciences

Applied Time Series Analysis
FS 2012 – Week 07

Finding Correlated Errors

1) Start by fitting an OLS regression and analyze residuals

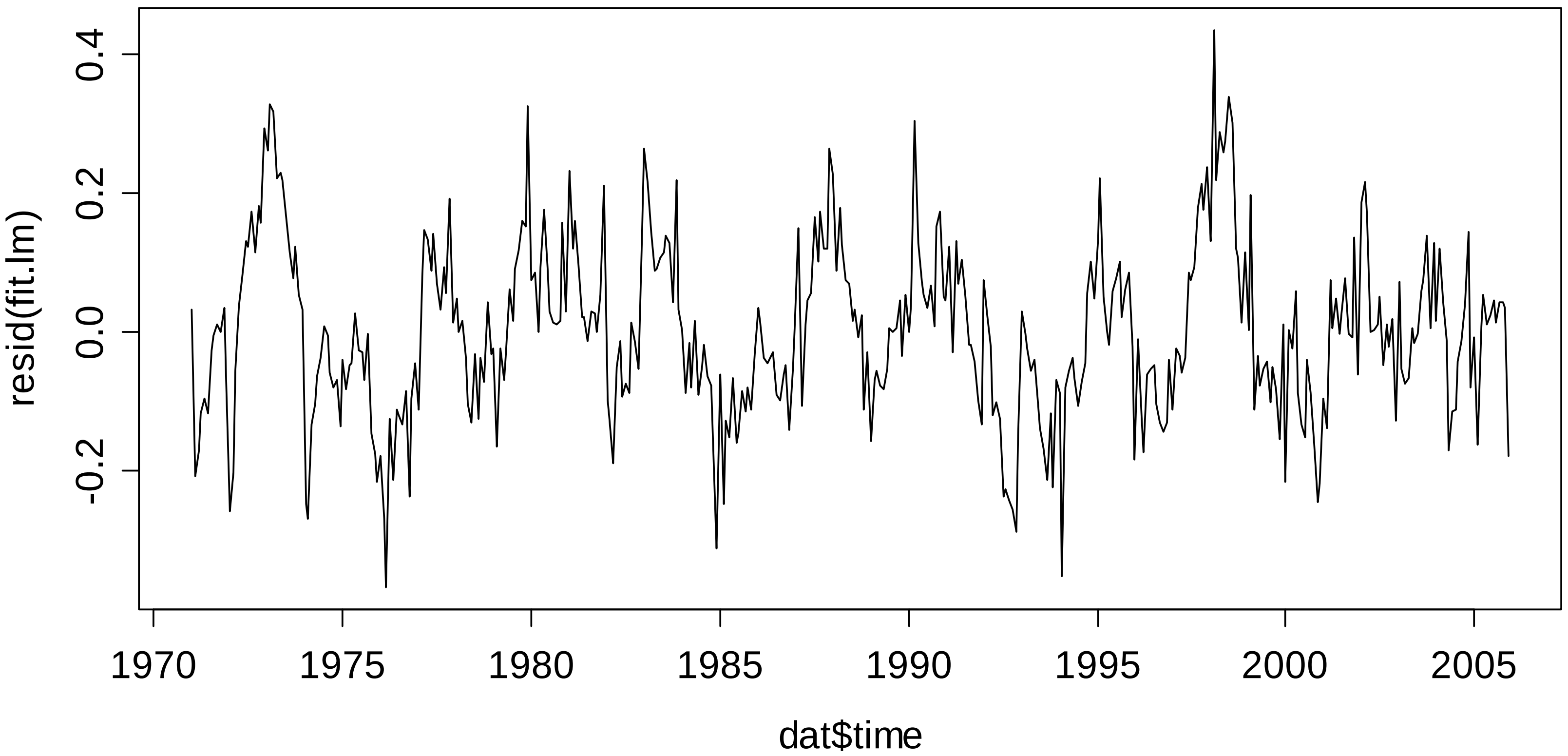


Applied Time Series Analysis
FS 2012 – Week 07

Finding Correlated Errors

2) Continue with a time series plot of OLS residuals

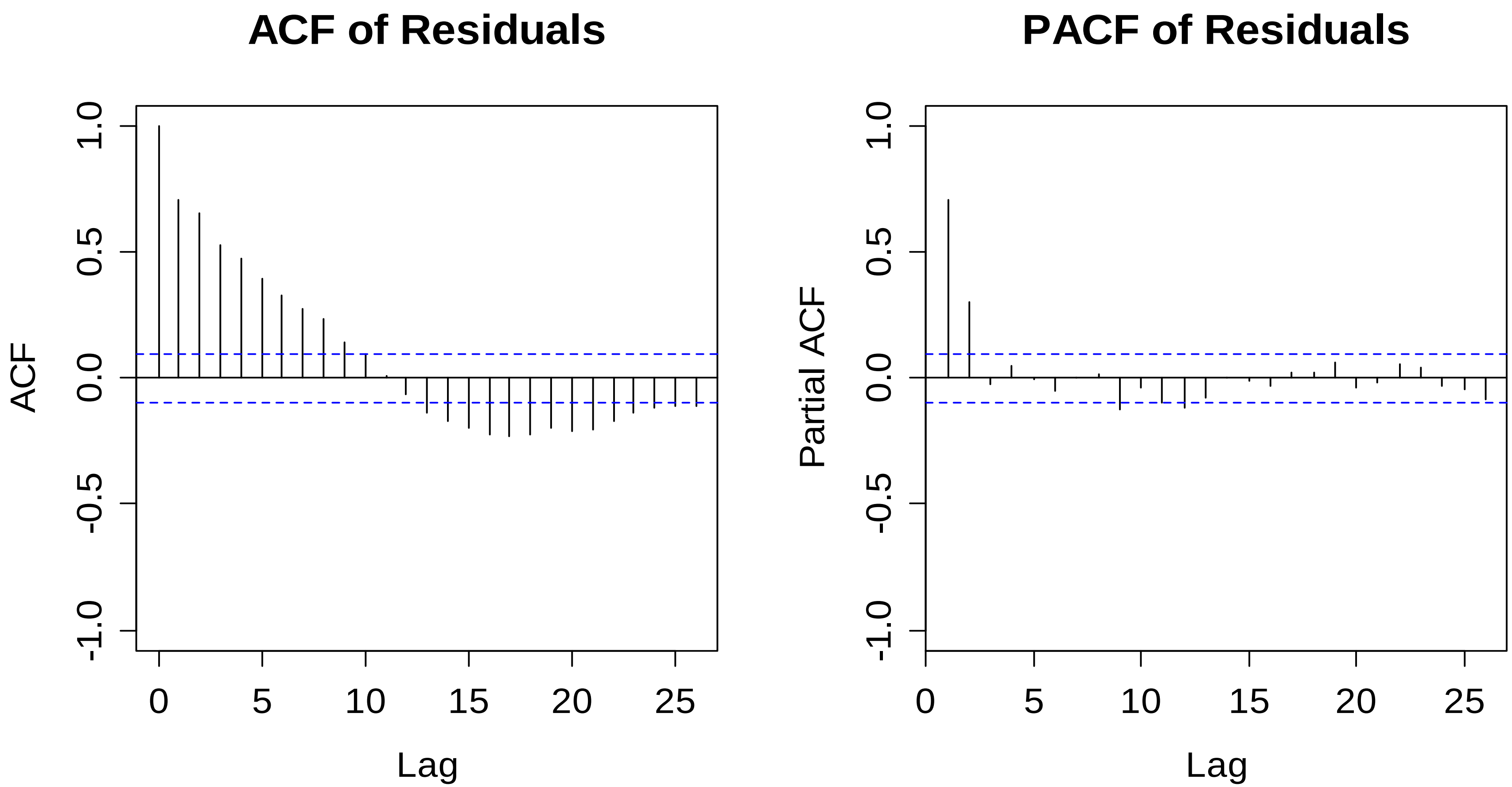
Residuals of the lm() Function



Applied Time Series Analysis
FS 2012 – Week 07

Finding Correlated Errors

3) Also analyze ACF and PACF of OLS residuals



Model for Correlated Errors

→ It seems as if an AR(2) model provides an adequate model for the correlation structure observed in the residuals of the OLS regression model.

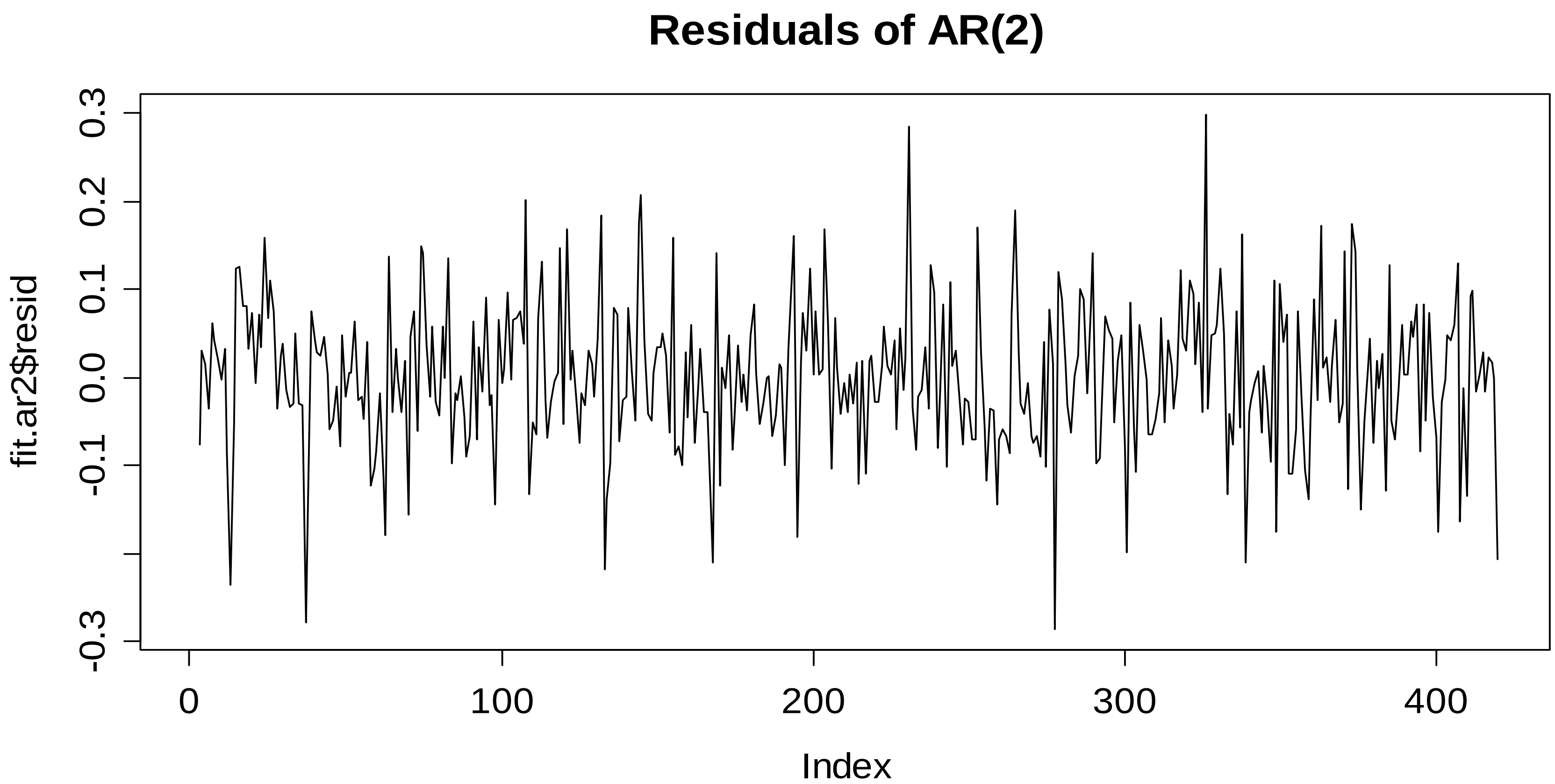
```
> fit.ar2 <- ar.burg(resid(fit.lm)); fit.ar2

Call: ar.burg.default(x = resid(fit.lm))

Coefficients:
     1       2
0.4945  0.3036

Order selected 2  sigma^2 estimated as  0.00693
```

→ Residuals of this AR(2) model must look like white noise!

Applied Time Series Analysis
FS 2012 – Week 07

Does the Model Fit?

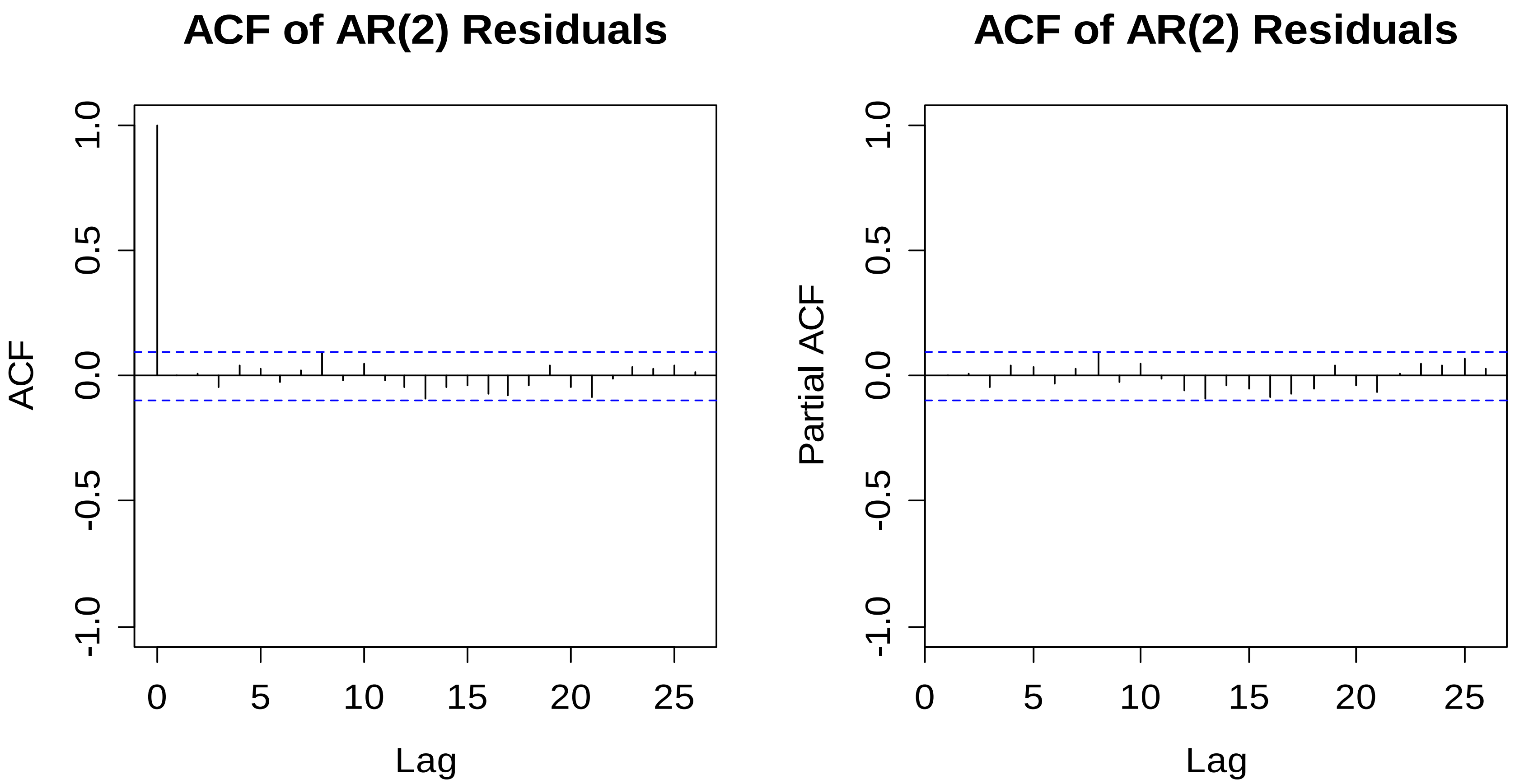
5) Visualize a time series plot of the AR(2) residuals



Applied Time Series Analysis
FS 2012 – Week 07

Does the Model Fit?

5) ACF and PACF plots of AR(2) residuals



Global Temperature: Conclusions

- The residuals from OLS regression are visibly correlated.

- An AR(2) model seems appropriate for this dependency.

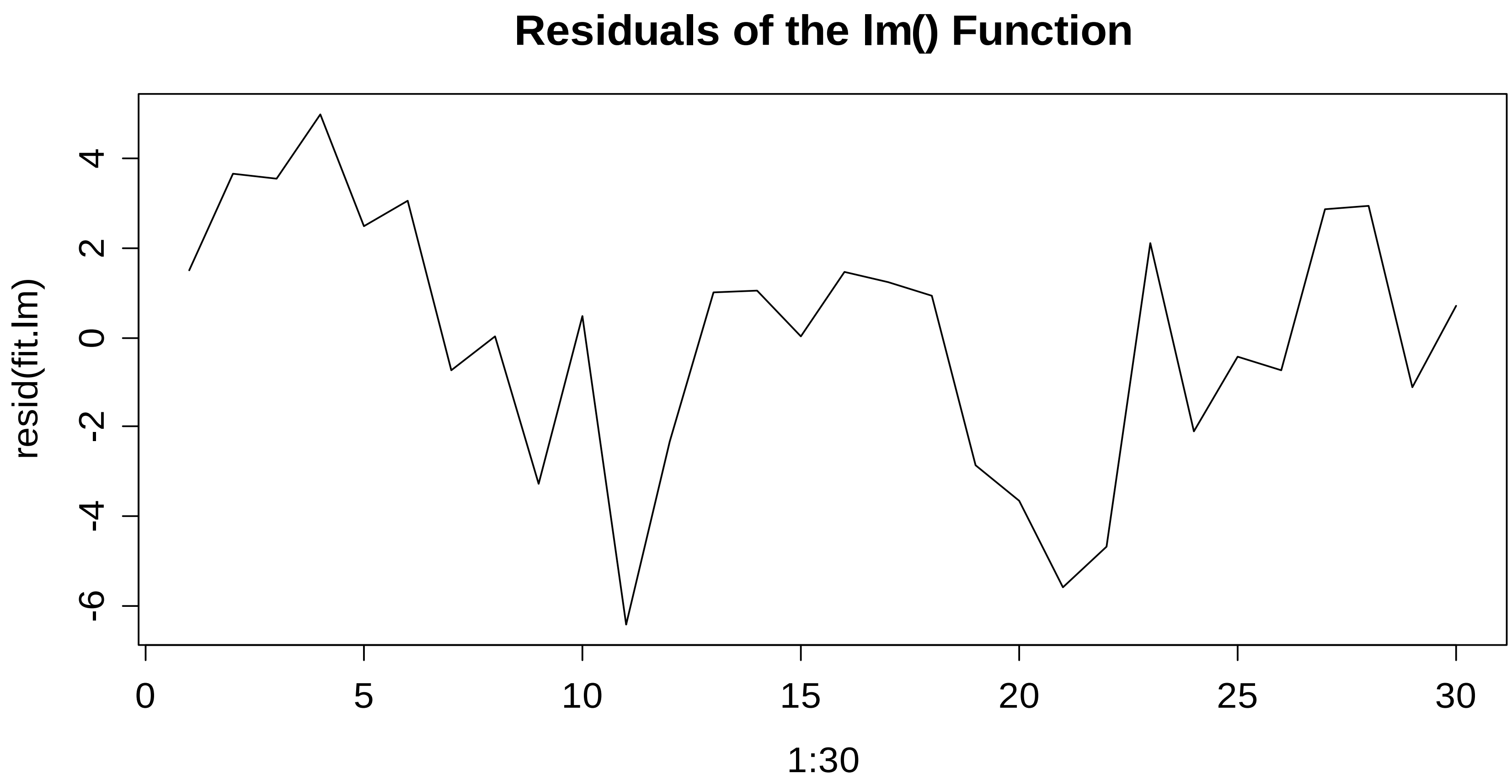
- The AR(2) yields a good fit, because its residuals have white noise properties. We have thus understood the dependency of the regression model errros.

→ We need to account for the correlated errors, else the coefficient estimates will be unbiased but inefficient, and the standard errors are wrong, preventing successful inference for trend and seasonality

Applied Time Series Analysis
FS 2012 – Week 07

Air Pollution: OLS Residuals

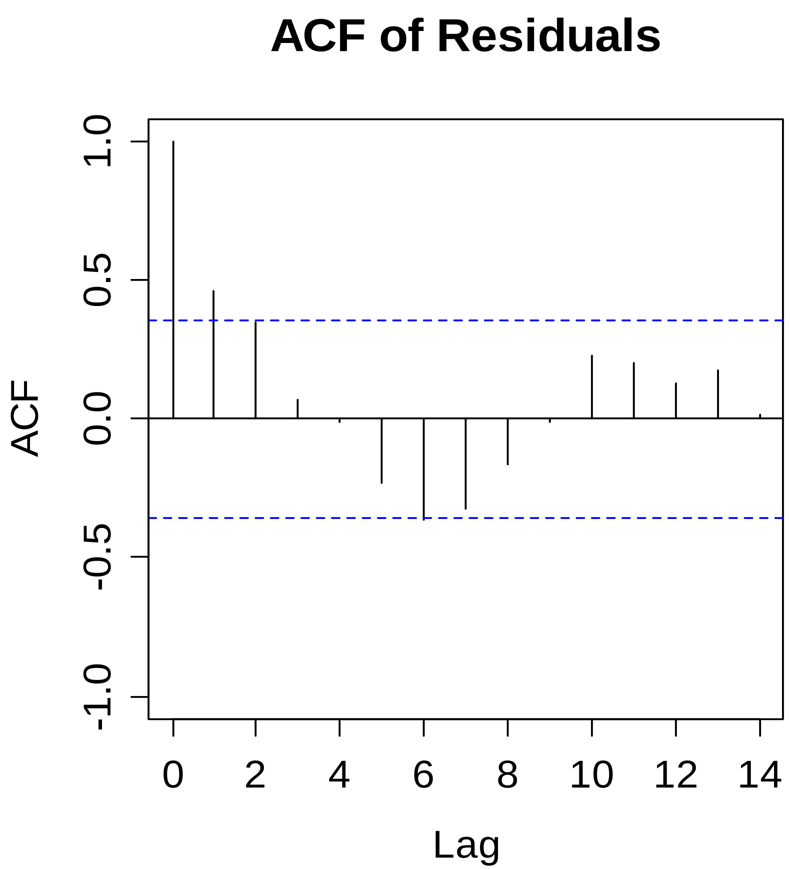
Time series plot: dependence present or not?

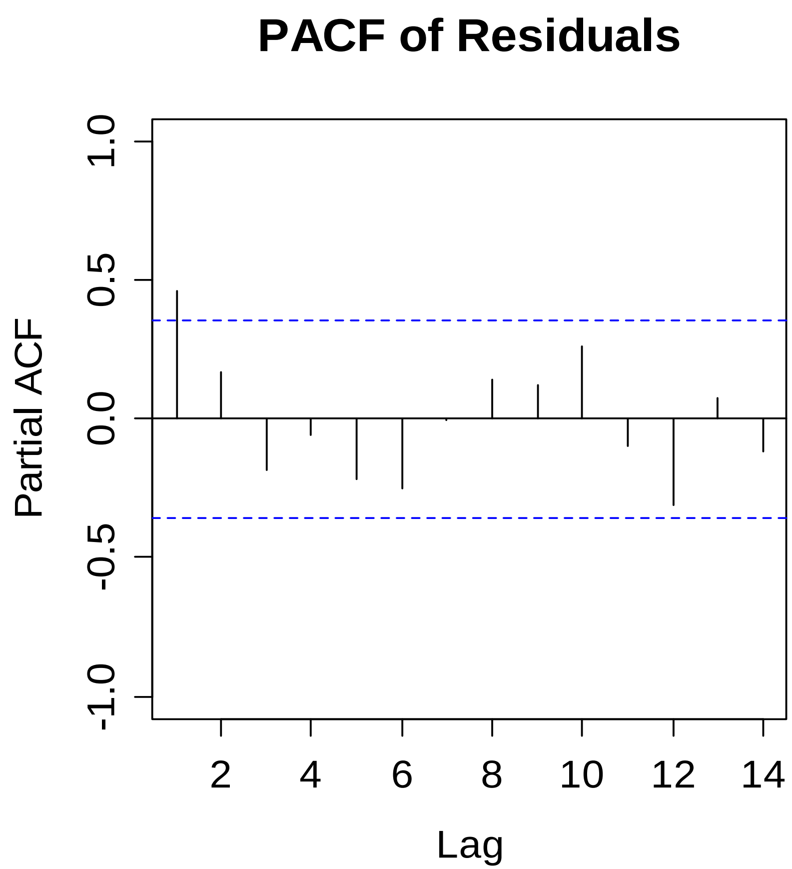


Applied Time Series Analysis
FS 2012 – Week 07

Air Pollution: OLS Residuals

ACF and PACF suggest: *there is AR(1) dependence*





Pollutant Example

```
> summary(erg.poll,corr=F)
Call: lm(formula = Oxidant ~ Wind + Temp, data = pollute)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.20334   11.11810  -0.468    0.644
Wind        -0.42706    0.08645  -4.940 3.58e-05 ***
Temp         0.52035    0.10813   4.812 5.05e-05 ***

Residual standard error: 2.95 on 27 degrees of freedom
Multiple R-squared: 0.7773, Adjusted R-squared: 0.7608
F-statistic: 47.12 on 2 and 27 DF,  p-value: 1.563e-09
```

Pollutant Example

```
> summary(erg.poll,corr=F)
Call: lm(formula = Oxidant ~ Wind + Temp, data = pollute)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.20334   11.11810  -0.468    0.644
Wind        -0.42706    0.08645  -4.940 3.58e-05 ***
Temp         0.52035    0.10813   4.812 5.05e-05 ***

Residual standard error: 2.95 on 27 degrees of freedom
Multiple R-squared: 0.7773, Adjusted R-squared: 0.7608
F-statistic: 47.12 on 2 and 27 DF,  p-value: 1.563e-09
```

Durbin-Watson Test

also see the blackboard...

- The Durbin-Watson approach is a dull test for (auto)-correlated errors in regression modeling
- Many statistics software packages automagically yield a decision or p-value for this test
- A rejection of its null hypothesis should always be taken as a serious hint for correlated errors
- **A non-rejection doesn‘t mean much!**

→ **Better to check ACF/PACF of residuals!**

Durbin-Watson Test

Example 1: Global Temperature

```
> library(lmtest)
> dwtest(fit.lm)
data:  fit.lm
DW = 0.5785, p-value < 2.2e-16
alt. hypothesis: true autocorrelation is greater than 0
```

Example 2: Air Pollution

```
> dwtest(fit.lm)
data:  fit.lm
DW = 1.0619, p-value = 0.001675
alt. hypothesis: true autocorrelation is greater than 0
```

Applied Time Series Analysis
FS 2012 – Week 07

Generalized Least Squares

→ **See the blackboard for full explanation**

- OLS regression assumes a diagonal error covariance matrix, but there is a generalization to $Var(E) = \sigma^2 \Sigma$.

- The regression model can be rewritten as:

$$\begin{aligned} y &= X\beta + E \\ S^{-1}y &= S^{-1}X\beta + S^{-1}E \\ y' &= X'\beta + E' \qquad \text{with } Var(E') = \sigma^2 I \end{aligned}$$

- One obtains the generalized least square estimates:

$$\hat{\beta} = (X^T \Sigma^{-1} X) X^T \Sigma^{-1} y \quad \text{with} \quad Var(\hat{\beta}) = (X^T \Sigma^{-1} X)^{-1} \sigma^2$$

Applied Time Series Analysis
FS 2012 – Week 07

Generalized Least Squares

For using the GLS approach, i.e. for correcting the dependent errors, we need an estimate of the error covariance matrix Σ.

The two major options for obtaining it are:

1) **Cochrane-Orcutt (for AR(p) correlation structure only)**
iterative approach: i) β, ii) α, iii) β

2) **GLS (Generalized Least Squares, for ARMA(p,q))**
simultaneous estimation of β and α

→ **Full explanation of the two different approaches is provided on the blackboard!**

GLS: Syntax

Package `nlme` has function `gls()`. It does only work if the correlation structure of the errors is provided. This has to be deterimined from the residuals of an OLS regression first.

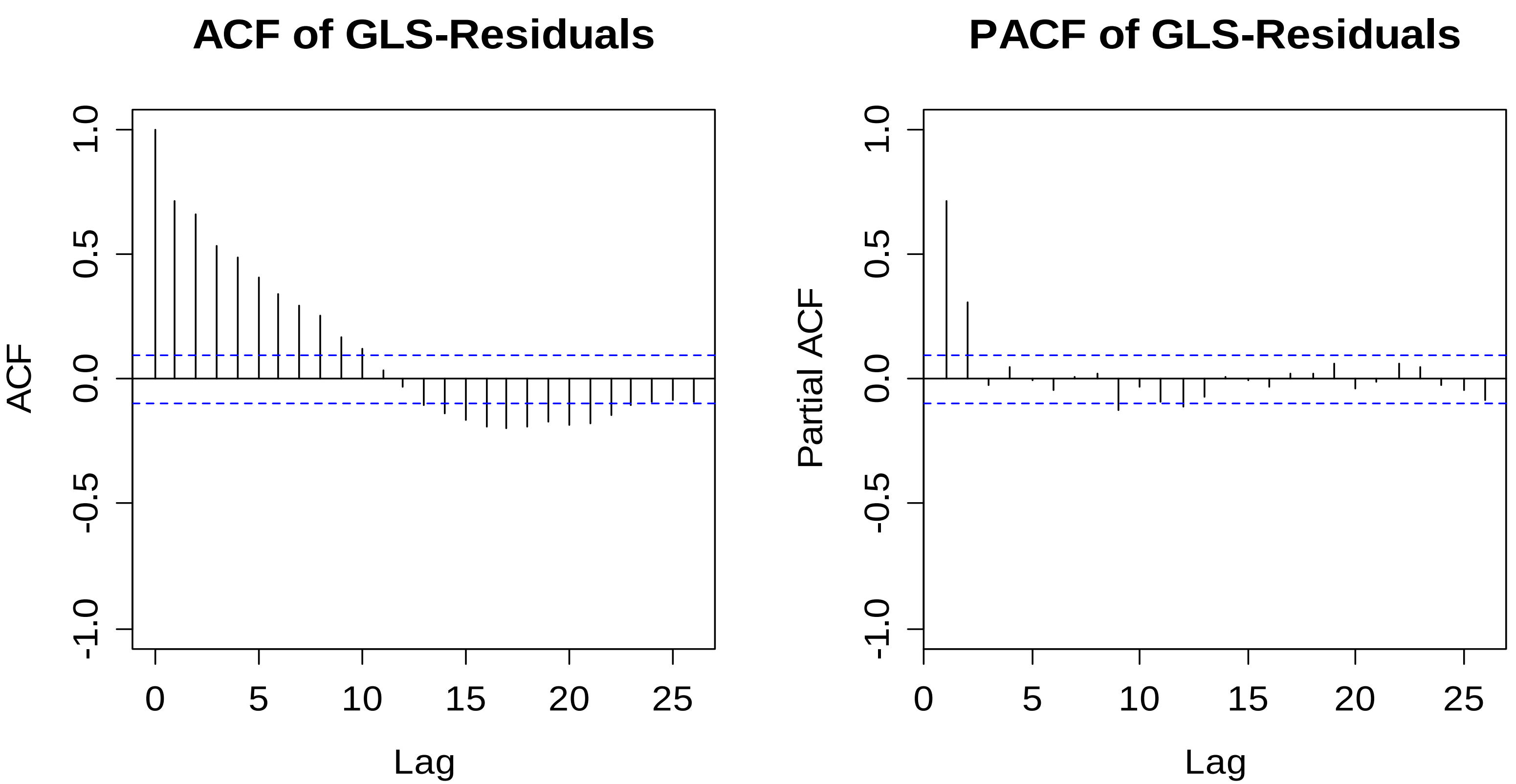
```
> library(nlme)
> corStruct <- corARMA(form=~time, p=2)
> fit.gls <- gls(temp~time+season, data=dat,
                    correlation=corStruct)
```

The output contains the **regression coefficients** and their **standard errors**, as well as the **AR-coefficients** plus some further information about the model (Log-Likeli, AIC, ...).

Applied Time Series Analysis
FS 2012 – Week 07

GLS: Residual Analysis

The residuals from a GLS must look like coming from a time series process with the respective structure:



GLS/OLS: Comparison of Results

→ **The trend in the global temperature is significant!**

OLS

```
> coef(fit.lm)["time"]
      time
0.01822374
> confint(fit.lm, "time")
          2.5 %    97.5 %
time 0.01702668 0.0194208
```

GLS

```
> coef(fit.gls)["time"]
      time
0.02017553
> confint(fit.gls, "time")
          2.5 %     97.5 %
time 0.01562994 0.02472112
```

Applied Time Series Analysis
FS 2012 – Week 07

GLS/OLS: Comparison of Results

→ The seasonal effect is not significant!

OLS

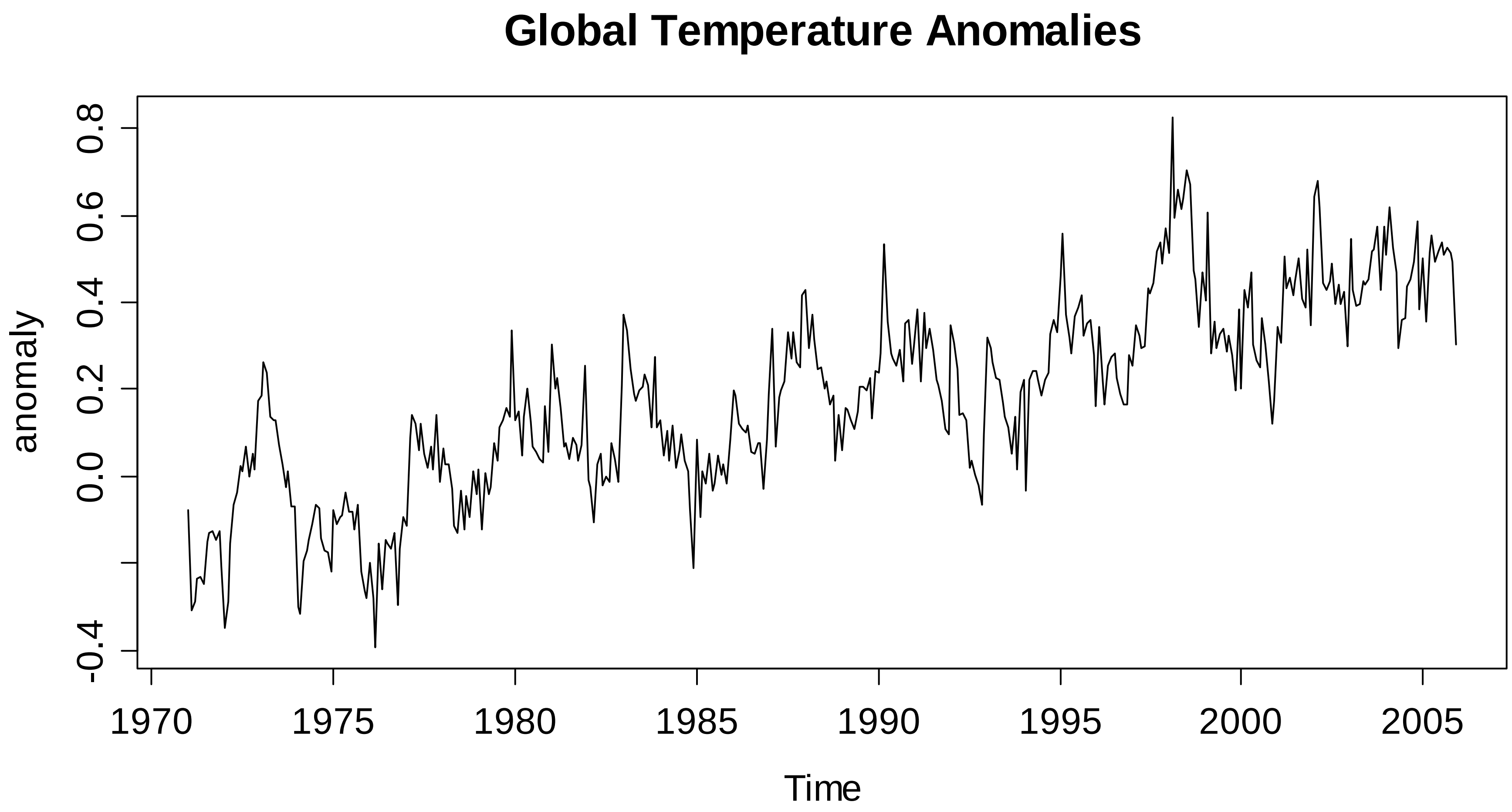
```
> drop1(fit.lm, test="F")
 temp ~ time + season
       Df Sum of Sq      RSS     AIC  F value  Pr(F)
<none>                6.4654 -1727.0
time    1   14.2274 20.6928 -1240.4 895.6210 <2e-16 ***
season 11    0.1744  6.6398 -1737.8   0.9982 0.4472
```

GLS

```
> anova(fit.gls)
Denom. DF: 407
            numDF  F-value p-value
(Intercept)     1 78.40801  <.0001
time            1 76.48005  <.0001
season         11  0.64371  0.7912
```

Applied Time Series Analysis
FS 2012 – Week 07

Example 1: Global Temperature



Applied Time Series Analysis
FS 2012 – Week 07

Air Pollution: Results

Both predictors are significant with both approaches...

OLS

```
> confint(fit.lm, c("Wind", "Temp"))
          2.5 %     97.5 %
Wind -0.6044311 -0.2496841
Temp  0.2984794  0.7422260
```

GLS

```
> confint(fit.gls, c("Wind", "Temp"))
          2.5 %     97.5 %
Wind -0.5447329 -0.2701709
Temp  0.2420436  0.7382426
```

→ But still, it is important to use GLS with correlated errors!

Simulation Study: Model

We want to study the effect of correlated errors on the quality of estimates when using the least squares approach:

$$x_t = t/50$$

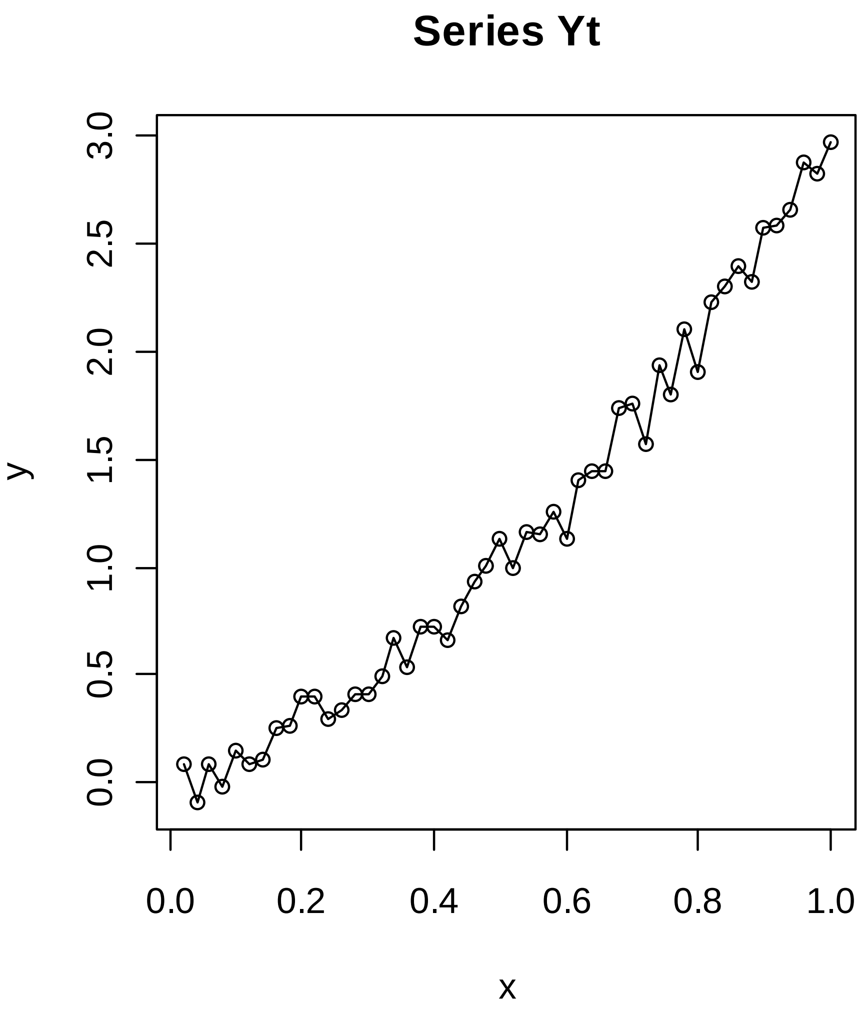
$$y_t = x_t + 2x_t^2 + E_t$$

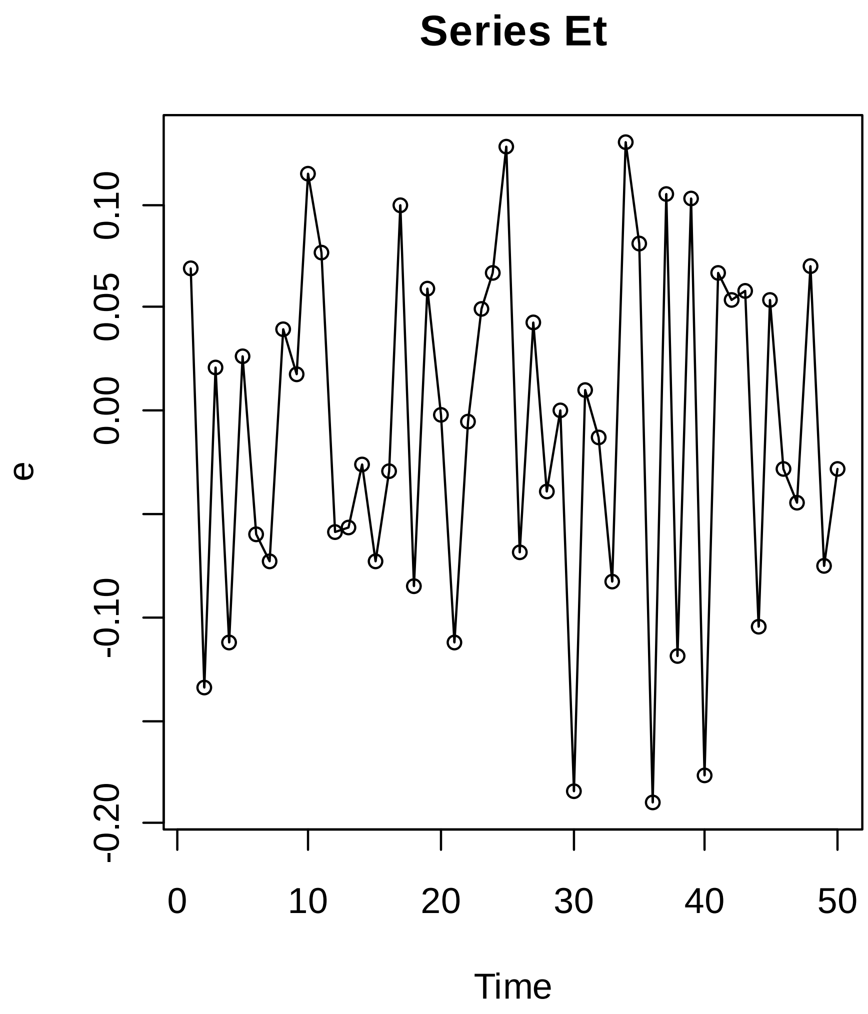
where E_t is from an AR(1)-process with $\alpha = -0.65$ and $\sigma = 0.1$.

We generate 100 realizations from this model and estimate the regression coefficient and its standard error by:

1) LS
2) GLS

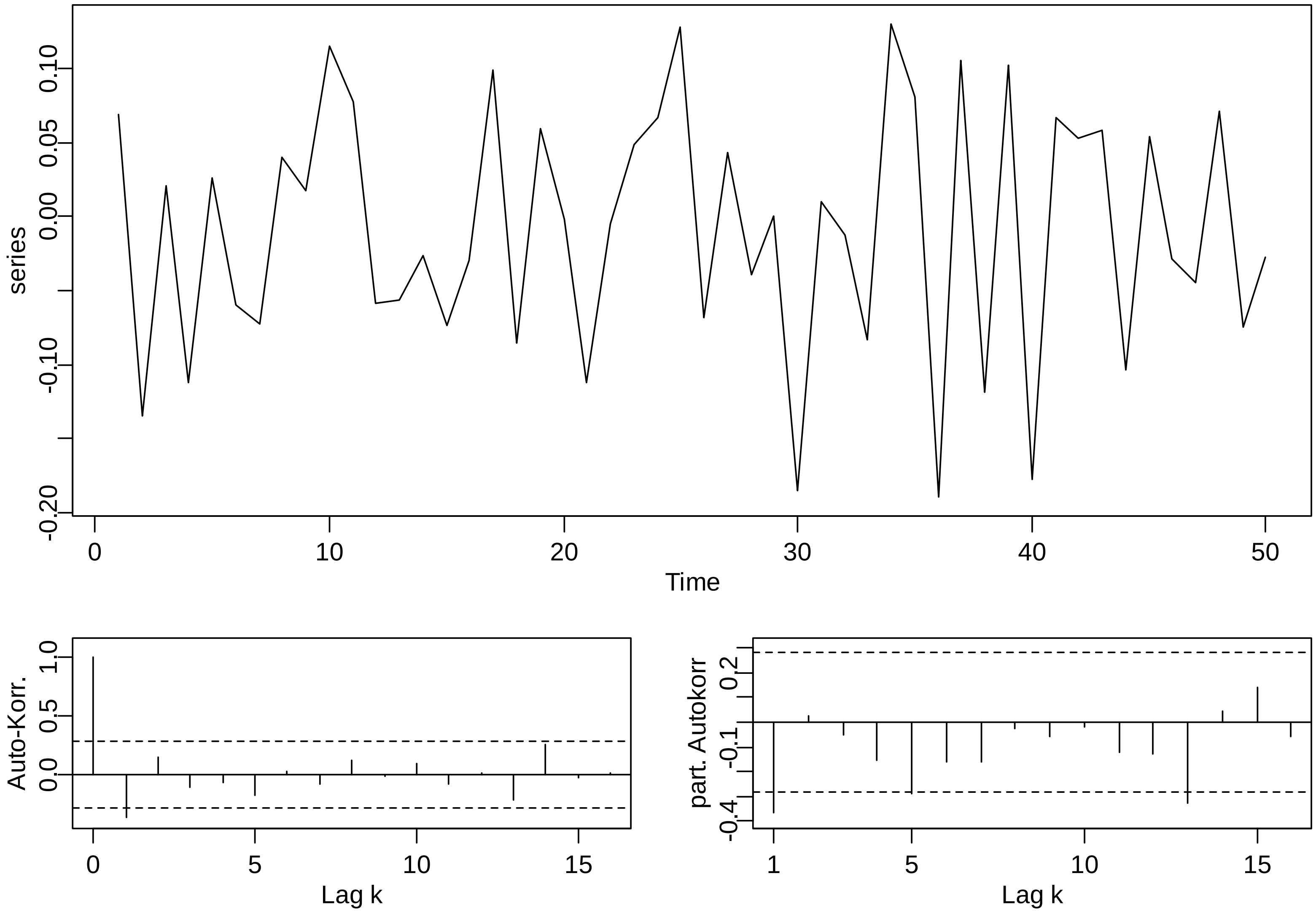
Simulation Study: Series





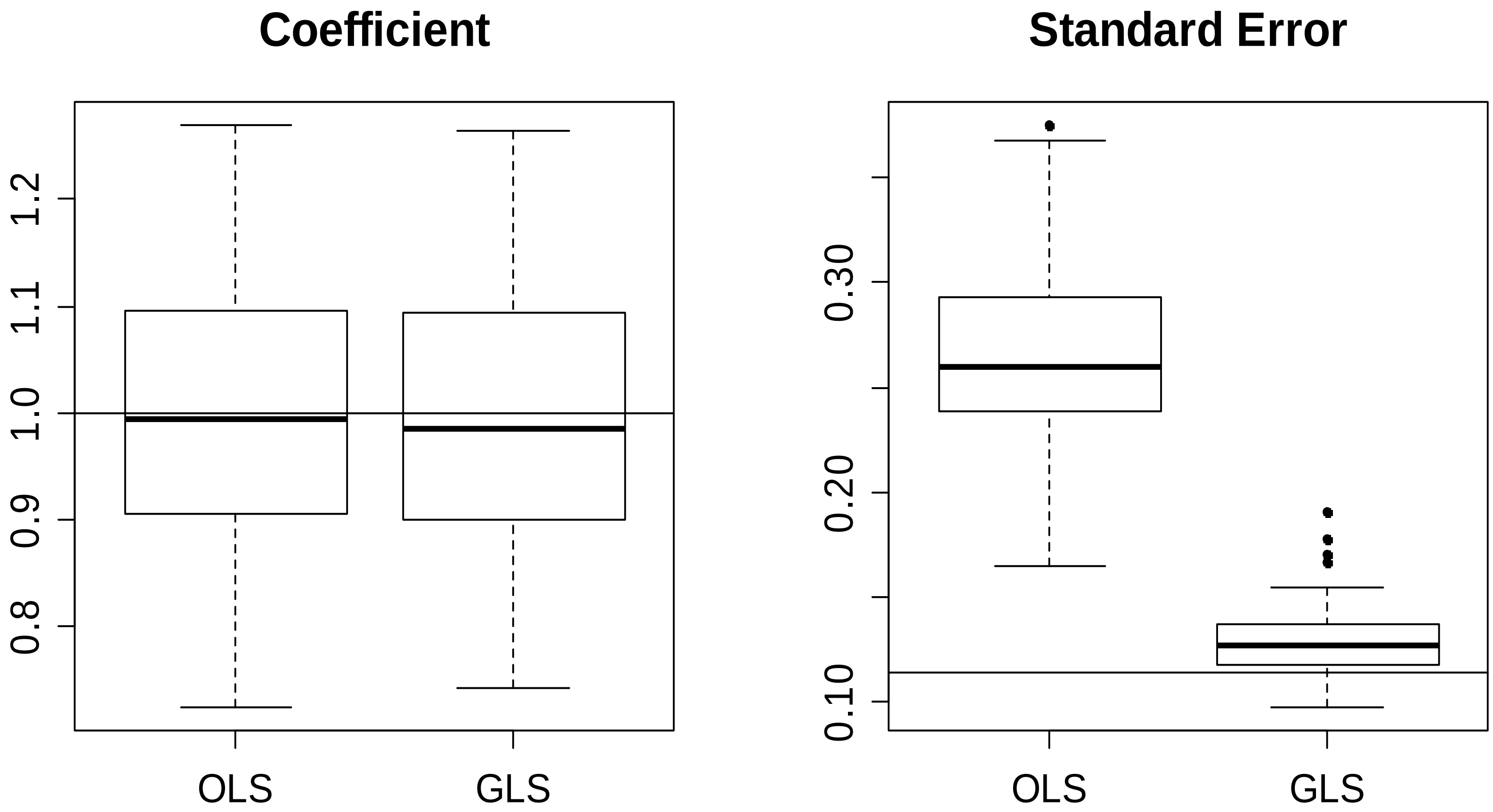
Applied Time Series Analysis
FS 2012 – Week 07

Simulation Study: ACF of the Error Term



Applied Time Series Analysis
FS 2012 – Week 07

Simulation Study: Results



Missing Input Variables

- (Auto-)correlated errors are often caused by the non-presence of crucial input variables.

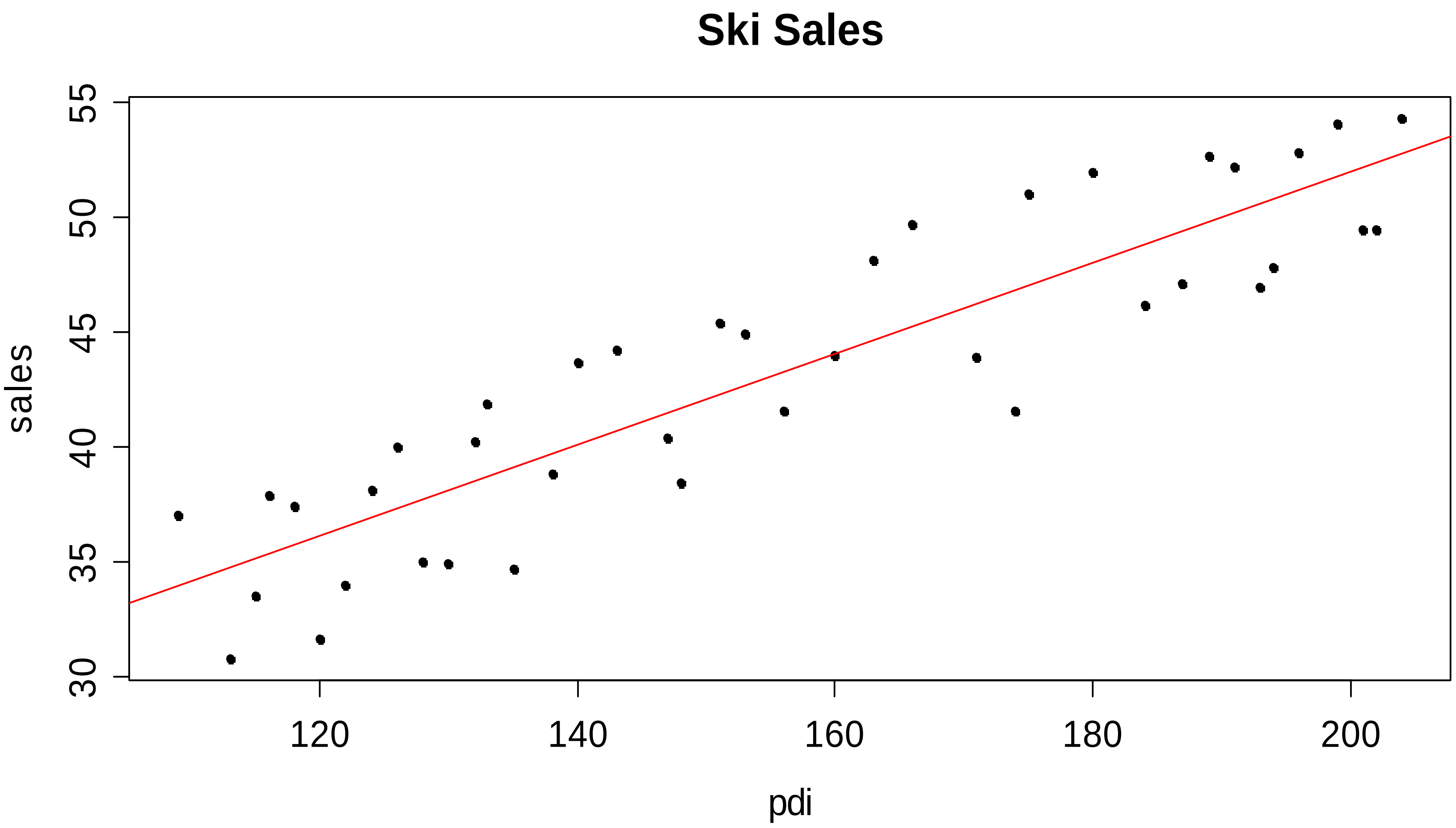
- In this case, it is much better to identify the not-yet-present -variables and include them in the analysis.

- However, this isn‘t always possible.

→ **regression with correlated errors can be seen as a sort of emergency kit for the case where the non-present variables cannot be added.**

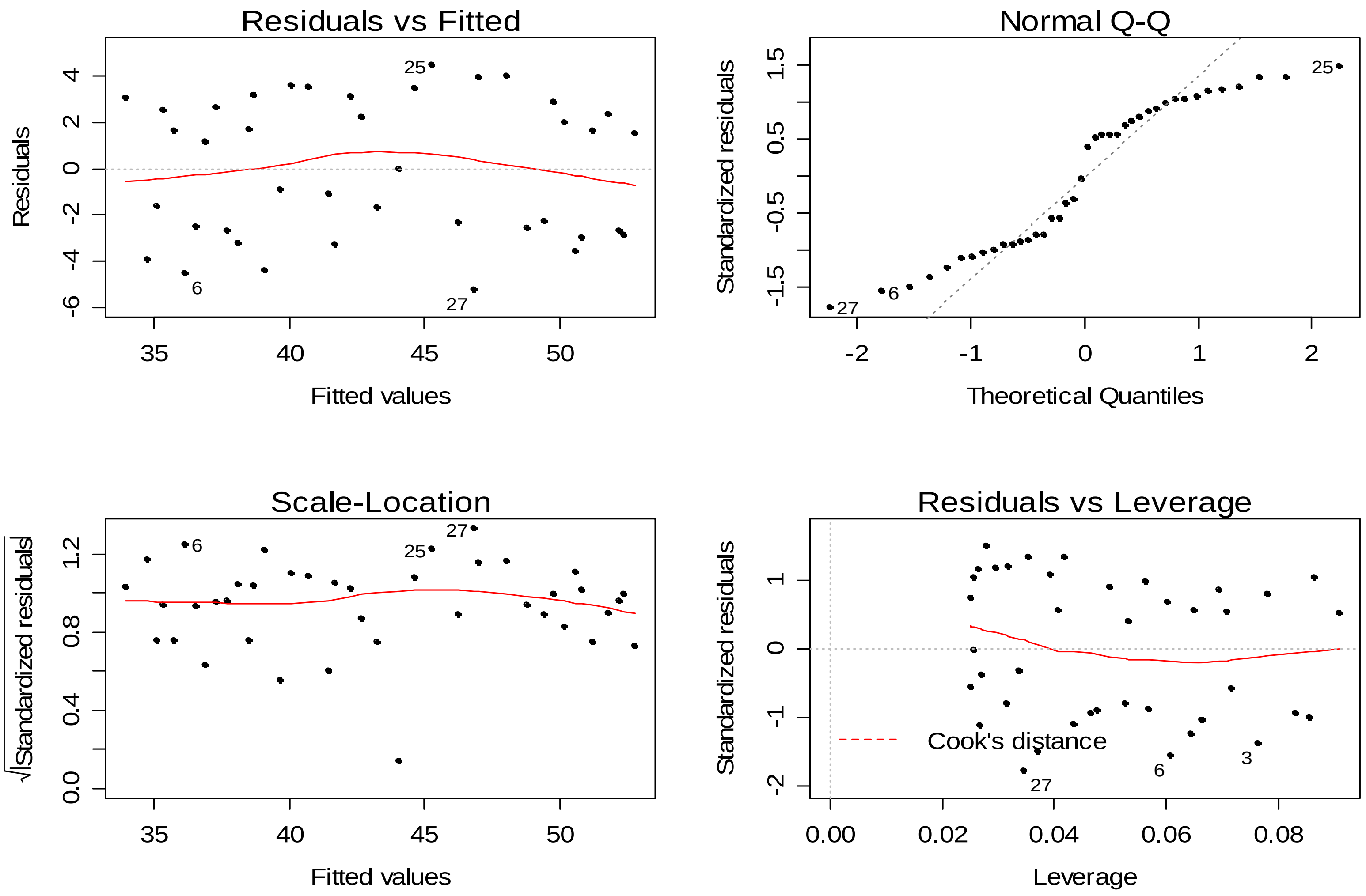
Applied Time Series Analysis
FS 2012 – Week 07

Example: Ski Sales



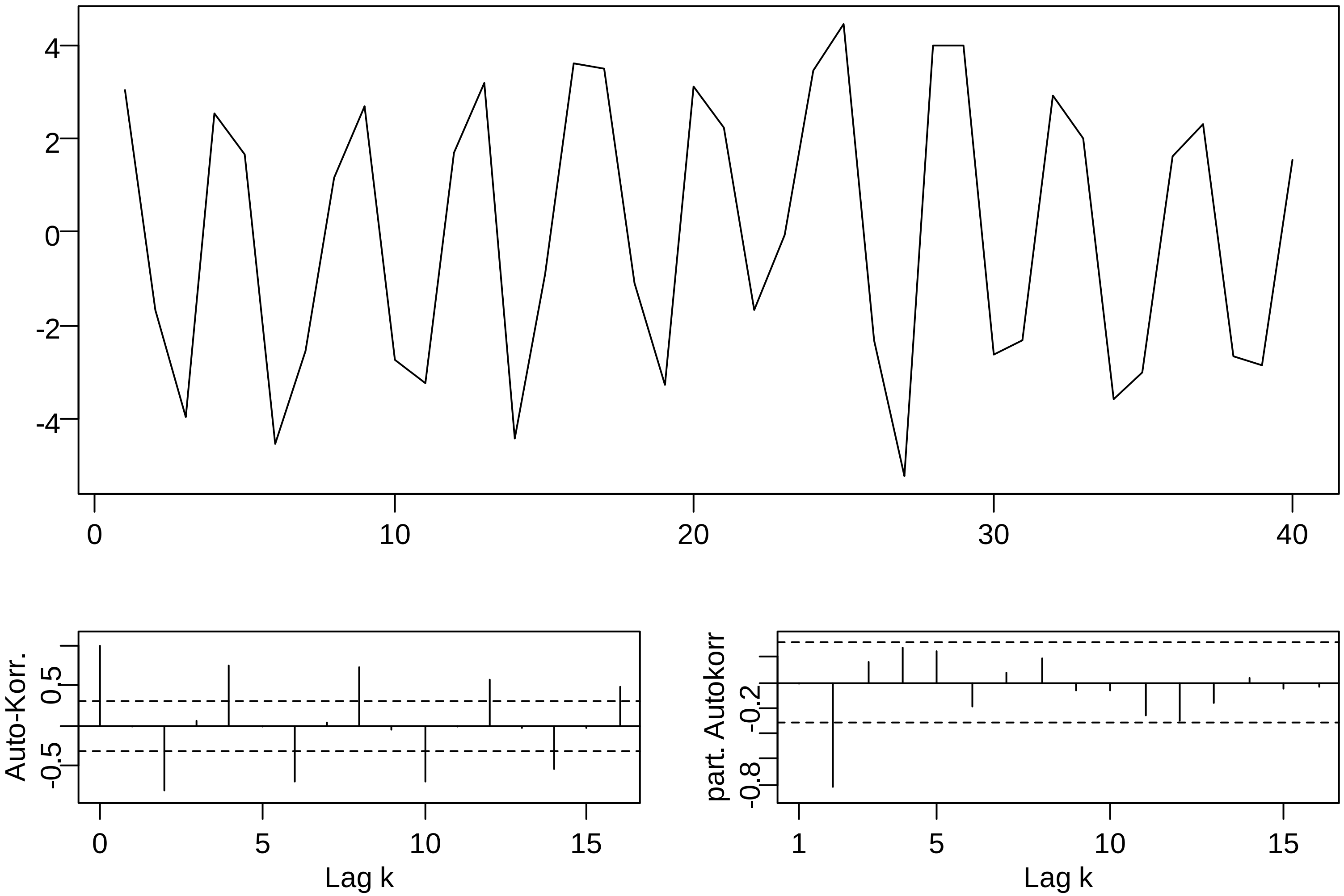
Applied Time Series Analysis
FS 2012 – Week 07

Ski Sales: Residual Diagnostics

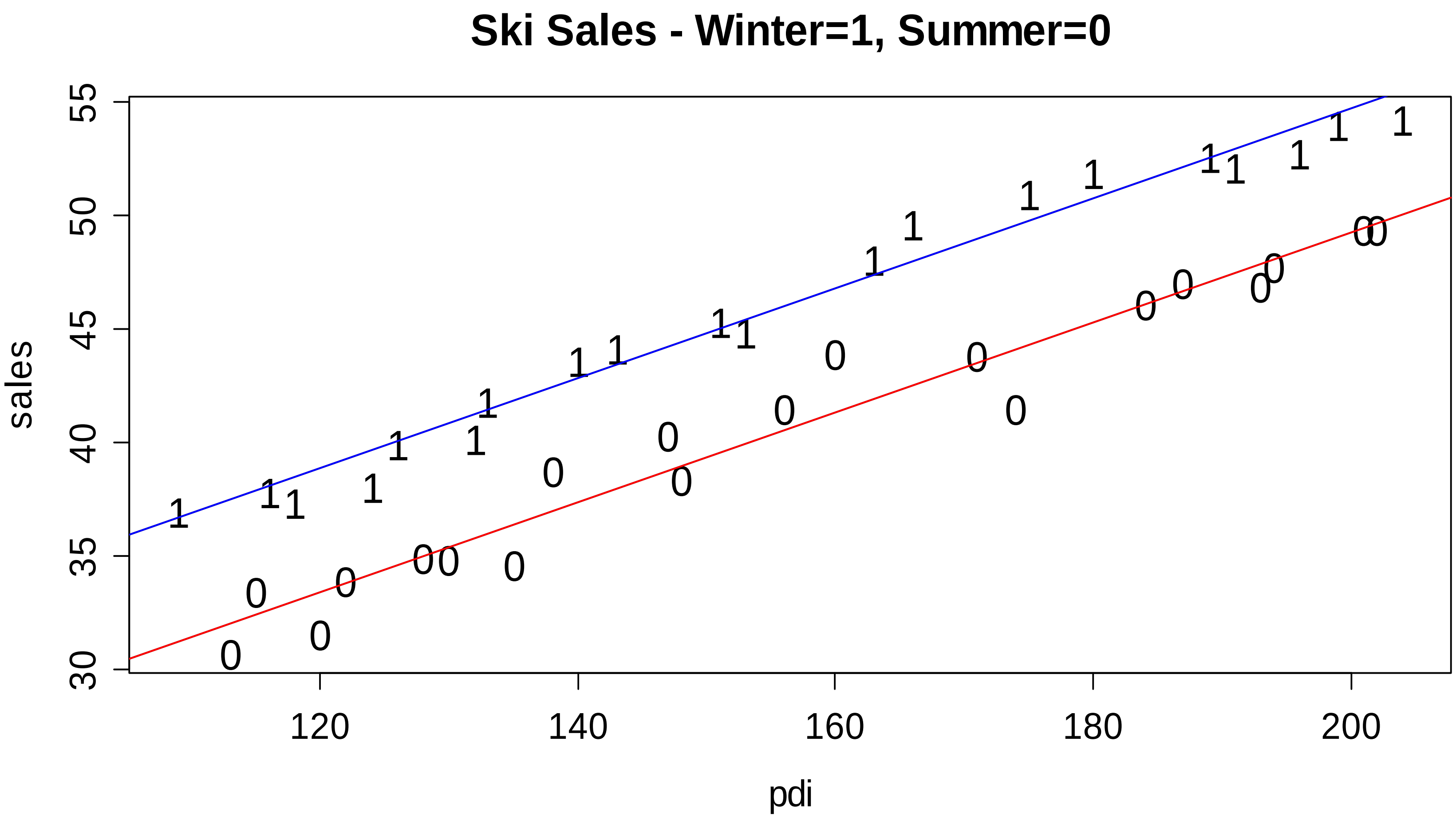


Applied Time Series Analysis
FS 2012 – Week 07

Ski Sales: ACF/PACF of Residuals

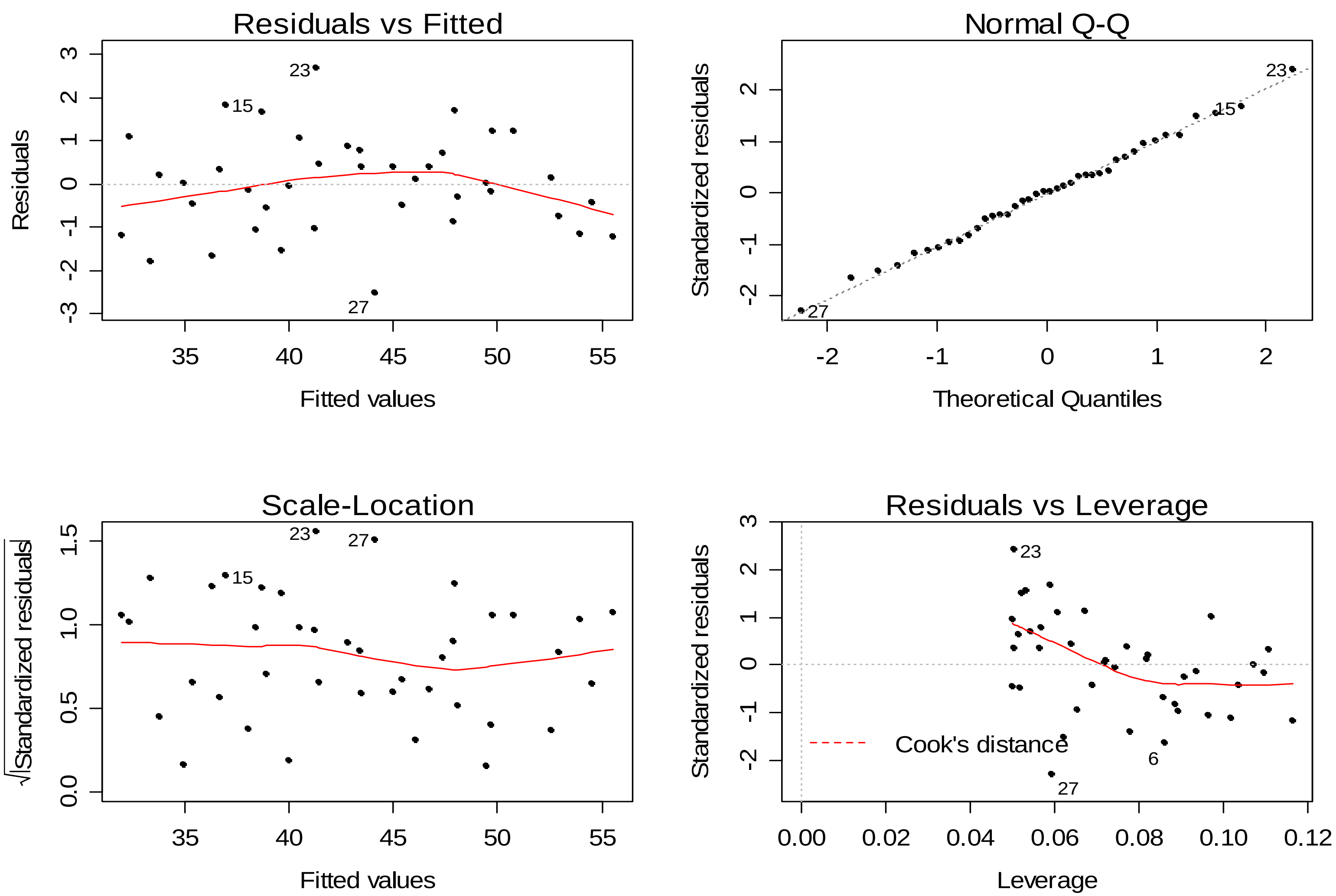


Ski Sales: Model with Seasonal Factor

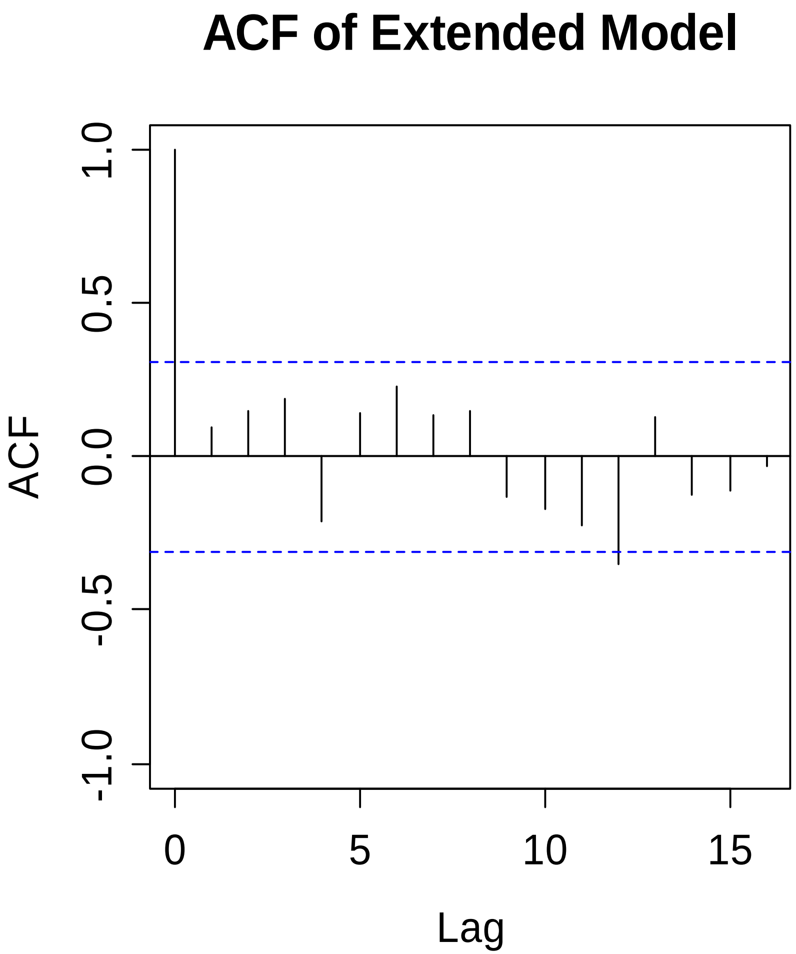


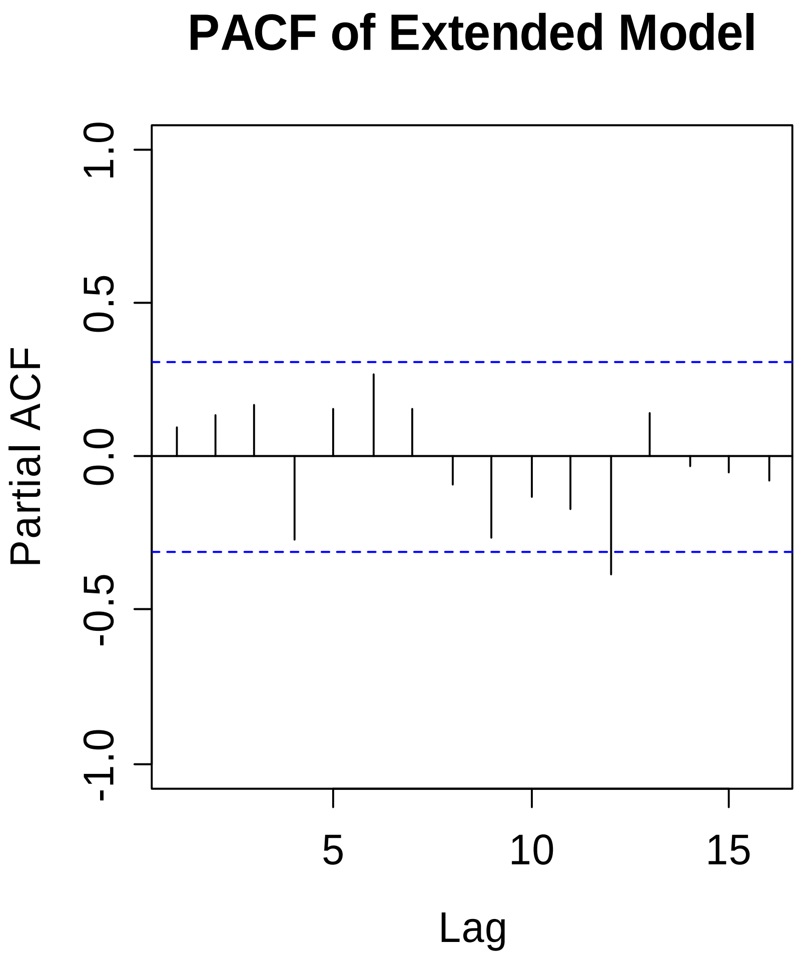
Applied Time Series Analysis
FS 2012 – Week 07

Residuals from Seasonal Factor Model



Residuals from Seasonal Factor Model





Ski Sales: Summary

- the first model (sales vs. PDI) showed correlated errors
- the Durbin-Watson test failed to indicate this correlation
- this apparent correlation is caused by ommitting the season
- adding the season removes all error correlation!

→ ***the emergency kit „time series regression“ is, after careful modeling, not even necessary in this example. This is quite often the case!***