

Multiple linear regression

Multiple linear regression

Let us recall the **multiple linear regression** model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

features
predictors
Explanatory
variables
→ error term

where X_j is the j th predictor and β_j quantifies the relationship between that variable and the response.

We interpret β_j as the **average effect** on Y of a one unit increase in X_j , holding all other predictors fixed.

Multiple linear regression

Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p.$$

The parameters are estimated through the ordinary least squares method, OLS, by minimizing

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

↙ actual ↘ prediction


Multiple linear regression: assumptions on error term

We make the following assumptions regarding error terms
($\varepsilon_1, \dots, \varepsilon_N$)

1. errors have mean zero $\rightarrow \bar{\varepsilon} = 0$
2. errors are uncorrelated \rightarrow correlogram can be used to verify
3. errors are uncorrelated with $X_{j,i}$

Multiple linear regression: model fit

The R^2 statistic is given by

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}}$$


In addition to looking at the R^2 , it can be useful to plot the data. Graphical summaries may reveal problems with a model that are not visible from numerical statistics.

Multiple linear regression

In order to test the global significance of the model we

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta_j \neq 0$$

through the F statistic

$$F = \frac{\text{ESS}/p}{\text{RSS}/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Handwritten notes:
An arrow points from p in the numerator to "no of predictors".
An arrow points from $(n-p-1)$ in the denominator to "no of examples".

$F > 4 \rightarrow$ Model Significant

Multiple linear regression

Results may be usefully displayed in an ANOVA table

Source	df	SS	MS	F
Model	p	ESS	MSR	MSR/MSE
Error	n-p-1	RSS	MSE	
Total	n-1	SST		

ESS = SSR
Explained → Regression

Multiple linear regression

After examining the global significance of the model, it is useful to evaluate the significance of parameters. The hypothesis system is

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

and the test is defined as

$$t = \frac{b_j}{\text{se}(b_j)}$$

where b_j is the estimate of the j_{th} coefficient and $\text{se}(b_j)$ is the standard error.

Multiple linear regression: collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another.

Effects of collinearity

- ▶ reduces the accuracy of estimates of the regression coefficients
- ▶ the standard error for β_j grows
- ▶ the t-statistic declines → we may fail to reject $H_0 : \beta_j = 0$

Multiple linear regression: collinearity

how do we detect a problem of collinearity?

- ▶ a simple way to detect collinearity is to look at the **correlation matrix** of the predictors.
- ▶ an element of this matrix that is large in absolute value indicates a pair of highly correlated variables → **collinearity**
- ▶ it is possible for collinearity to exist between three or more variables → **multicollinearity**

Multiple linear regression: collinearity

A better way to assess the multicollinearity is to compute the variance inflation factor, VIF.

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

where R_j^2 is the determination index of the regression of the j_{th} variable on the other $k - 1$ predictors.

- ▶ If $R_j^2 = 0$, then $\text{VIF}_j = 1$.
- ▶ If there is a multicollinearity problem, then $\text{VIF}_j > 1$.

For example, $R_j^2 = 0.9$, $\text{VIF}_j = 10$.

$\text{VIF} > 3.5$ \rightarrow multicollinearity problem exists

Example

$$X_1 = \beta_0 + \beta_2 X_2 + \varepsilon$$

$$X_2 = \beta_0 + \beta_1 X_1 + \varepsilon$$

Let us consider a sample of 10 households and the following variables:

- ▶ Y : yearly amount spent in food (hundreds eur)
- ▶ X_1 : family income (thousands eur)
- ▶ X_2 : number of family members

We first calculate the correlation matrix ...

	Y	X_1	X_2
Y	1	0.884	0.737
X_1		1	0.867
X_2			1

→ highly correlated

Example

We estimate the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

coefficient	estimate	std. error	t-statistic
β_0	3.51865	3.16055	1.1133
β_1	2.27762	0.81261	2.80284
β_2	-0.411406	1.23603	-0.332844

Source	df	SS	MS	F
Model	2	213.422	106.711	12.75
Error	7	58.578	8.3682	
Total	9	272		

$$R^2 = 0.7846$$

How do we interpret these results?

Significant Model
since $F > 4$
but there might be the chance of VIF

Example

Let us compute the Variance Inflation Factor.

This may be easily computed for X_1 e X_2 considering that $R^2 = (r_{X_1X_2})^2 = (0.867)^2 = 0.75$ so that

$$\text{VIF}_{X_1} = 1/(1 - 0.75) = 4$$

$$\text{VIF}_{X_2} = 1/(1 - 0.75) = 4$$

There is a multicollinearity problem: solution \rightarrow remove X_2 from the model and estimate a simple regression with X_1 .

Multiple linear regression with time series

Many business and economic problems involve the use of time series data.

The linear regression model may be usefully employed to model monthly, quarterly or yearly data.

► A linear trend may be easily included through a predictor $X_{1,t} = t$.

► Seasonality may modeled with seasonal dummy variables.

As a general rule, we use $s - 1$ dummy variables to describe s periods (to avoid perfect multicollinearity).

monthly seasonality requires $\rightarrow 12 - 1 = 11$ dummy variables

Month
Jan \rightarrow
Feb
Dec

Jan Feb . . . Dec

1 0 . . . 0

0 Jan = 1 + 0 Feb + 0 Mar + . . . Dec + ϵ

Multiple linear regression with time series

For instance, a model for quarterly data with trend and seasonality may be

$$Y_t = \beta_0 + \beta_1 t + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \varepsilon_t$$

Trend and seasonality are modelled as a series of straight lines with different intercept and same slope. The first quarter is described with the model $Y_t = \beta_0 + \beta_1 t$.

Parameters $\beta_2, \beta_3, \beta_4$ describe the variation with respect to β_0 due to seasonality.

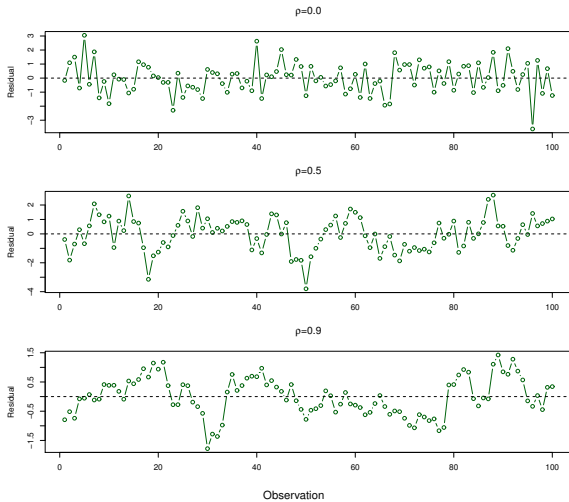
$$\begin{aligned} \text{For } S_1 &\rightarrow Y_t = \beta_0 + \beta_1 t + \varepsilon_t \\ \text{For } S_2 &\rightarrow Y_t = \boxed{\beta_0 + \beta_2} + \beta_1 t + \varepsilon_t \end{aligned}$$

Multiple linear regression with time series

- ▶ Time series data tend to be autocorrelated
- ▶ Autocorrelation occurs when a the effect of a variable is spread over time. For example, a change in prices may have an effect on both current and future sales
- ▶ Autocorrelation may be detected through a graphical inspection of residuals
- ▶ Specific tests on residuals

.

Autocorrelated residuals



Autocorrelated residuals

A typical example of autocorrelation is defined as

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

with

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

where ρ is the correlation between sequential errors and ν_t is an erratic component with mean zero and constant variance.

If $\rho = 0$ allora $\varepsilon_t = \nu_t$.

The **Durbin-Watson test** is typically used to diagnose this kind of autocorrelation The system of hypothesis is

$$H_0 : \rho = 0 \quad H_1 : \rho > 0$$

Durbin-Watson test

The Durbin-Watson test is defined as

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

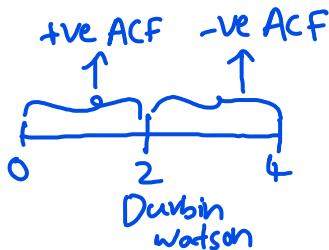
The values of DW range between 0 and 4 with a central value of 2.

For large samples, the following holds

$\rightarrow DW = 2(1 - r_1(e))$

where $r_1(e)$ is the residual autocorrelation at lag 1.

Since $-1 < r_1(e) < 1$, then $0 < DW < 4$.



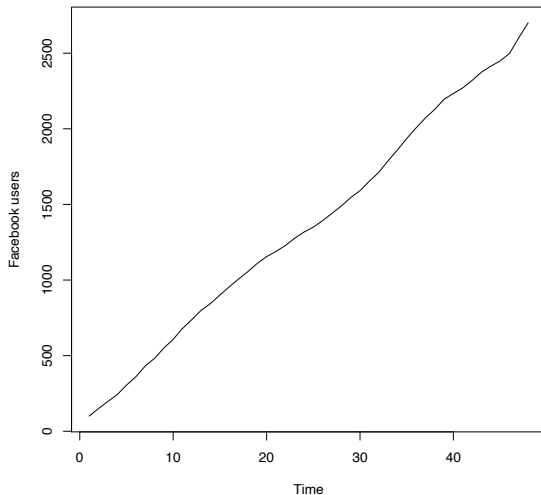
Autocorrelation: solutions

To solve the problem of autocorrelation we need to examine the model:

- ▶ is the functional form correct?
- ▶ are there any omitted variables?

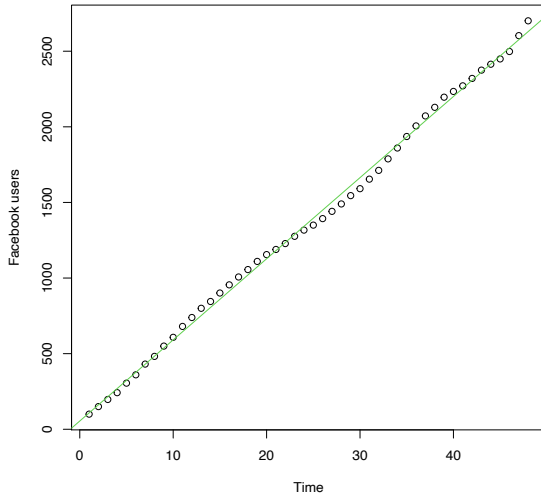
Example

Facebook users: quarterly data 2008-2020



Example

Facebook users: simple linear regression



Example

Facebook users: simple linear regression

```
lm(formula = fb ~ time)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	54.5363	10.9917	4.962	1e-05 ***
time	53.6507	0.3905	137.378	<2e-16 ***

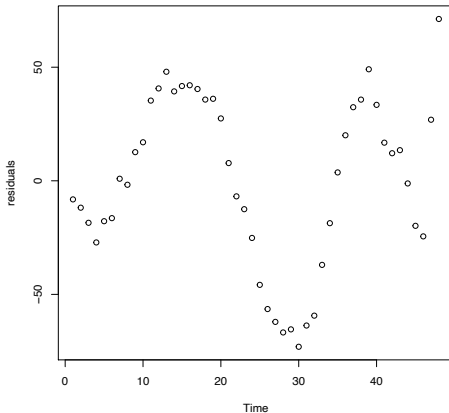
Residual standard error: 37.48 on 46 degrees of freedom

Multiple R-squared: 0.9976, Adjusted R-squared: 0.9975

F-statistic: 1.887e+04 on 1 and 46 DF, p-value: < 2.2e-16

Example

Facebook users: residuals

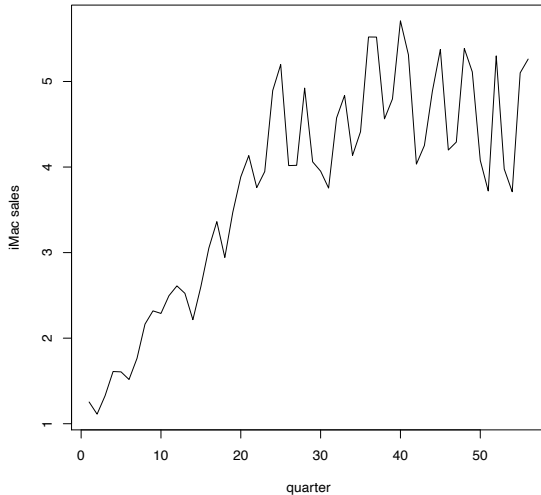


Durbin-Watson test: $DW = 0.16378$ p-value $< 2.2e-16$

Positive autocorrelation in residuals

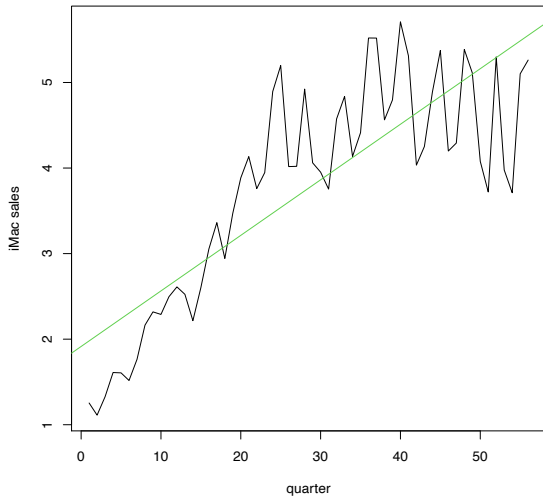
Example

iMac sales: quarterly data 2006-2019



Example

iMac sales: simple linear regression



Example

iMac sales: linear regression with trend and seasonality

Call:

```
tslm(formula = mac.ts ~ trend + season)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.60158	-0.42293	-0.00687	0.54972	1.42797

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.155255	0.236078	9.129	2.62e-12	***
trend	0.064591	0.005613	11.507	8.68e-16	***
season2	-0.640448	0.256052	-2.501	0.0156	*
season3	-0.460039	0.256237	-1.795	0.0785	.
season4	0.176727	0.256544	0.689	0.4940	

Residual standard error: 0.6773 on 51 degrees of freedom

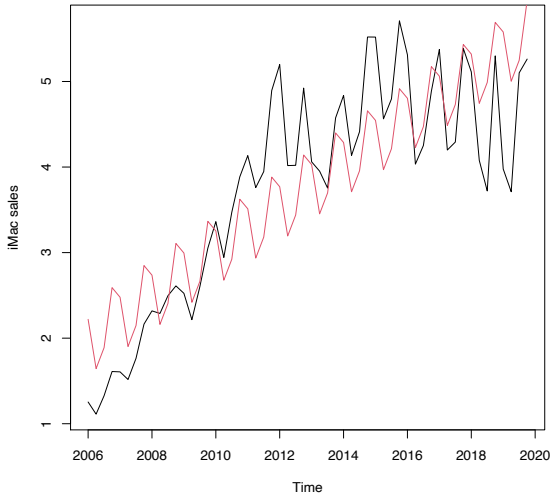
Multiple R-squared: 0.7436, Adjusted R-squared: 0.7235

F-statistic: 36.97 on 4 and 51 DF, p-value: 1.695e-14

Example

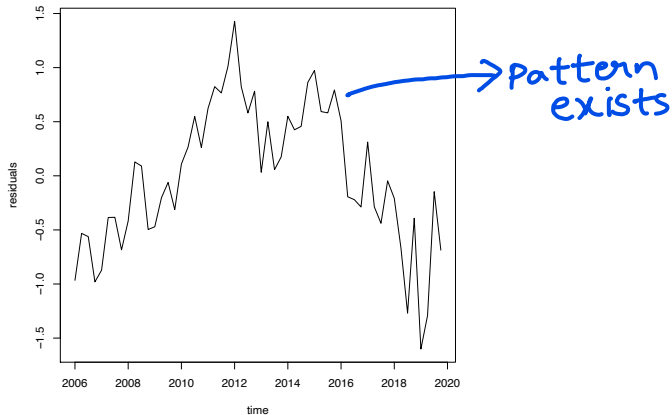
iMac sales: linear regression with trend and seasonality

$$\text{Sales} = \beta_0 + \beta_1 t + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \varepsilon_t$$



Example

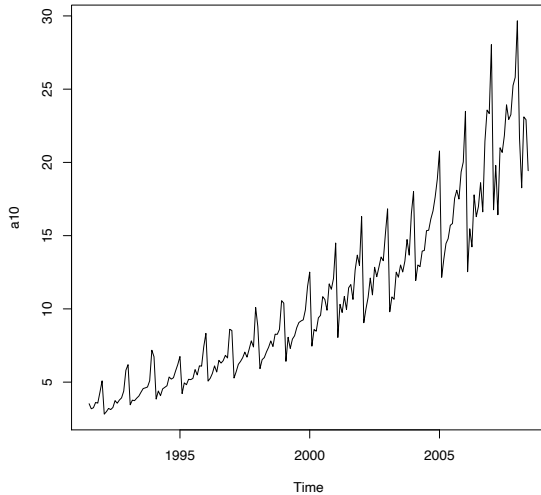
iMac sales: residuals



Residuals clearly show a nonlinear behaviour

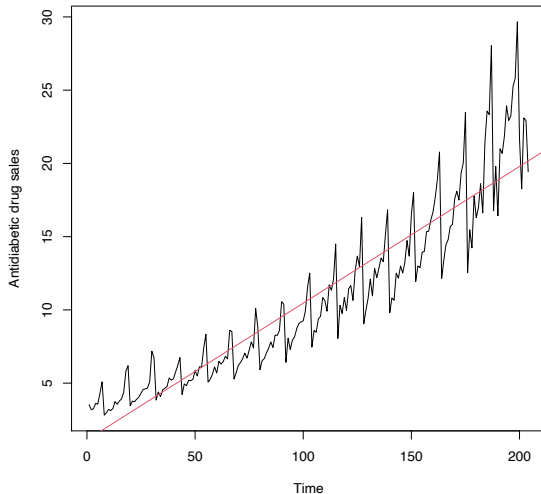
Example

Monthly sales of a drug



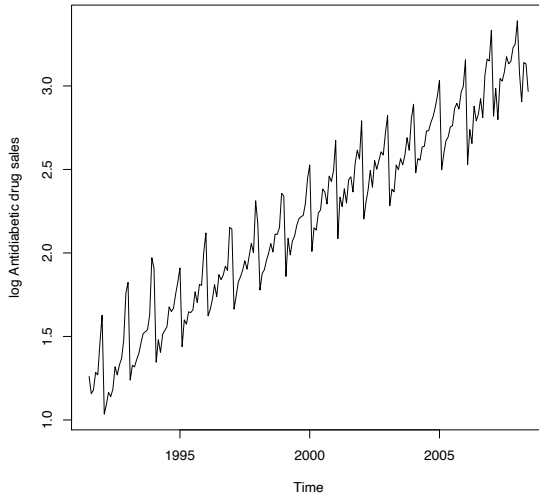
Example

Monthly sales of a drug: simple linear regression



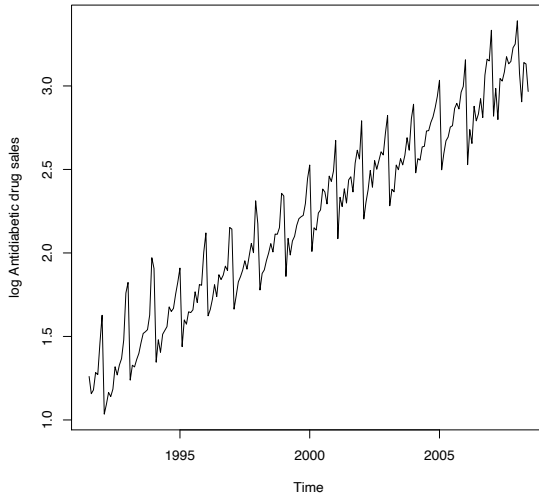
Example

Monthly sales of a drug: log transformation



Example

Monthly sales of a drug: log transformation



Example

Monthly sales of a drug: simple linear regression with log transformation

Call:

```
lm(formula = la10 ~ t)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.36954	-0.09621	-0.00889	0.07139	0.43395

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2577135	0.0216920	57.98	<2e-16 ***
t	0.0093211	0.0001835	50.80	<2e-16 ***

Residual standard error: 0.1543 on 202 degrees of freedom

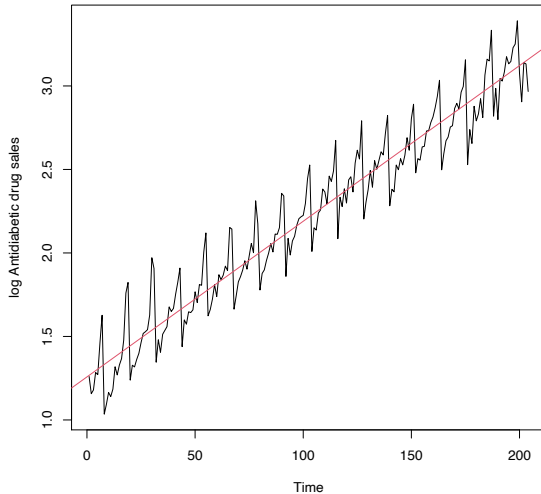
Multiple R-squared: 0.9274, Adjusted R-squared: 0.927

F-statistic: 2580 on 1 and 202 DF, p-value: < 2.2e-16

*better
result*

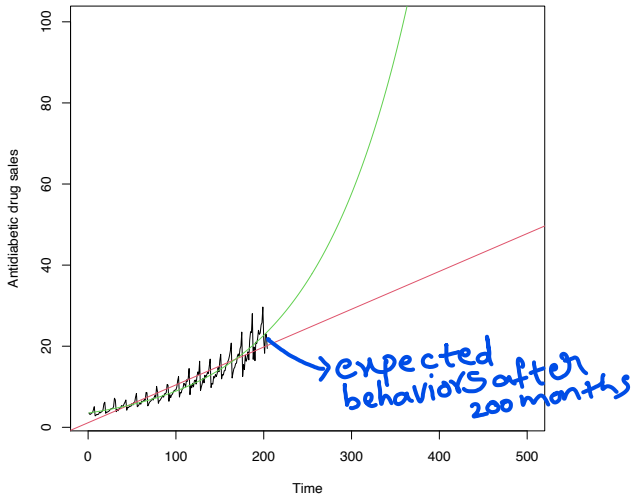
Example

Monthly sales of a drug: log transformation



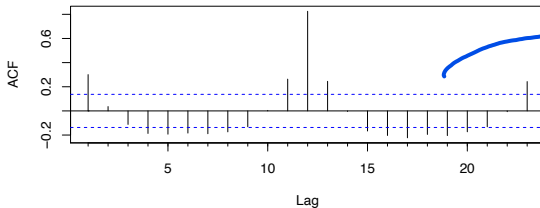
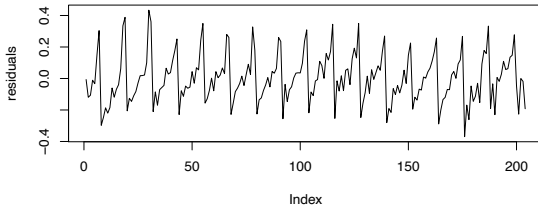
Example

Monthly sales of a drug: model comparison



Example

Monthly sales of a drug: residuals



Auto correlation exists