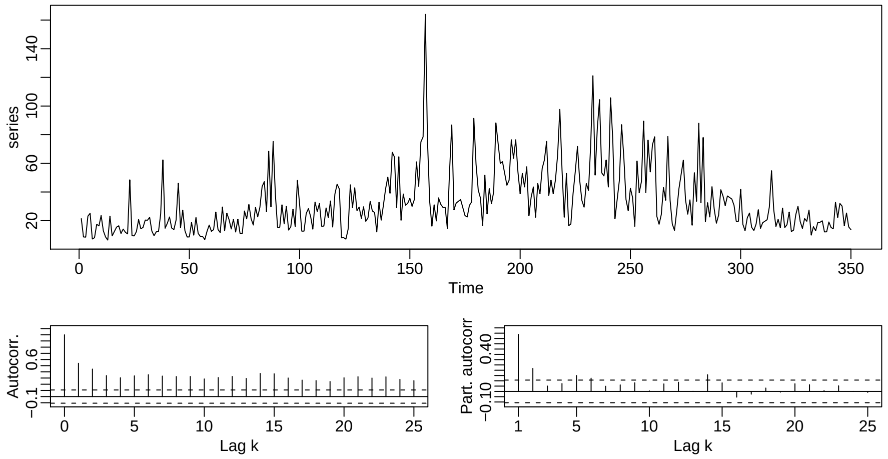
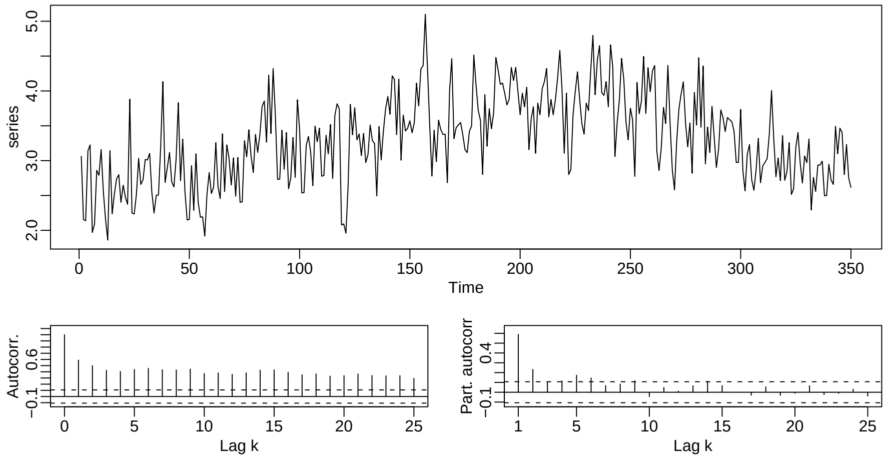
Solution to Series 5

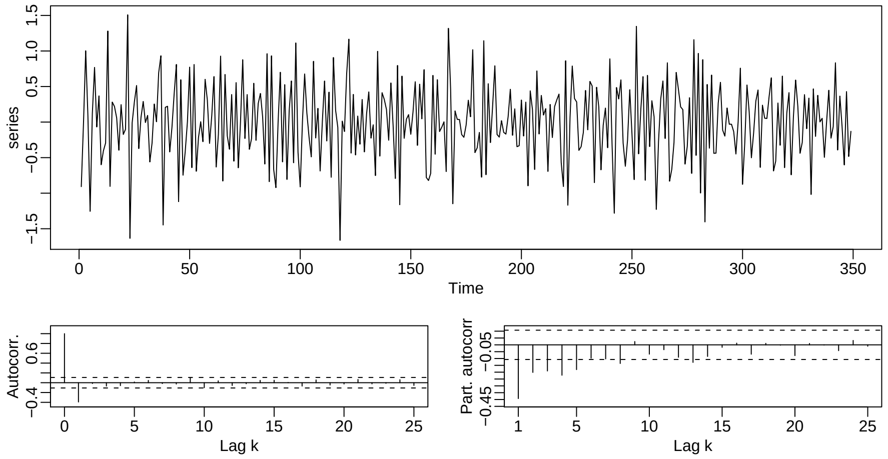
1. a) In some areas the variance is much smaller than in others. The "peak" in the middle indicates that a logarithmic transformation must first be applied to the data. If we look at the correlogram, we notice that the ordinary autocorrelations decay far too slowly. Even for large lags, they still lie outside the confidence band:



Even the time series of logarithmic data cannot yet be regarded as stationary, since it exhibits clear trends (first increasing, then decreasing), which can however be eliminated by taking first differences:



The plot of first differences for the transformed series shows that stationarity can now be assumed:



R commands:

```
> f.acf(d.varve)
> f.acf(log(d.varve))
> f.acf(diff(log(d.varve)))
```

b) The correlogram plotted in part a) indicates an ARIMA(1,1,1) process (or perhaps an ARIMA(0,1,1) process). Fitting these two models, we see that the ARIMA(1,1,1) model is very good at describing the logarithmic data. In both fitted models, the algorithm converges; of the two models, ARIMA(1,1,1) has a smaller AIC.

The estimated coefficients are $\widehat{\beta}_1 = -0.84$ for the fitted ARIMA(0,1,1) model and $\widehat{\alpha}_1 = 0.25$, $\widehat{\beta}_1 = -0.91$ for the fitted ARIMA(1,1,1) model. For both models, the estimated mean is $\widehat{\mu} = -0.00127$, which leads us to assume the data do not need correcting by their mean. Furthermore, the estimated error variances are 0.224 (for the ARIMA(0,1,1) model) and 0.2138 (for the ARIMA(1,1,1) model).

Thus the ARIMA(0,1,1) model looks as follows:

$$Y_t = X_t - X_{t-1}$$

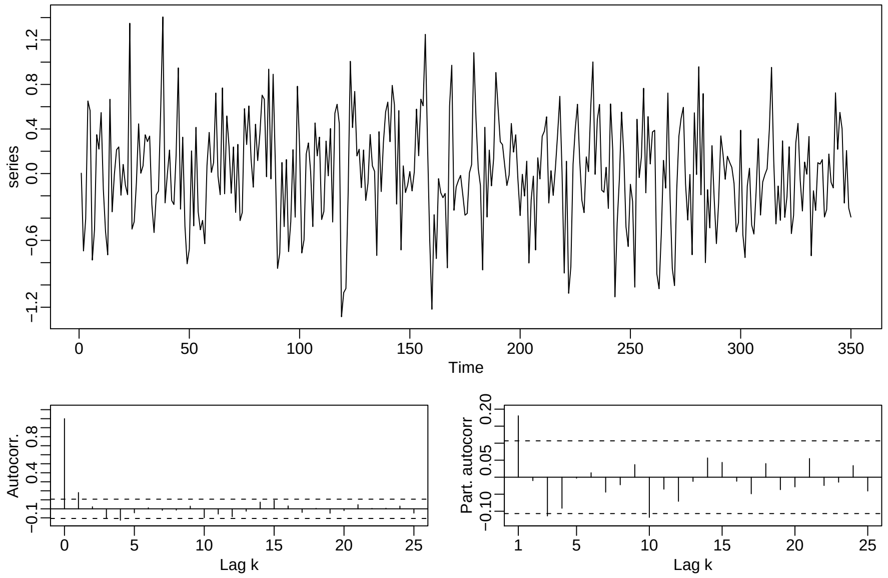
$$Y_t = E_t - 0.84 \cdot E_{t-1}; \quad \sigma^2_{E_t} = 0.224$$

For the ARIMA(1,1,1) model, we similarly have

$$Y_t = X_t - X_{t-1}$$

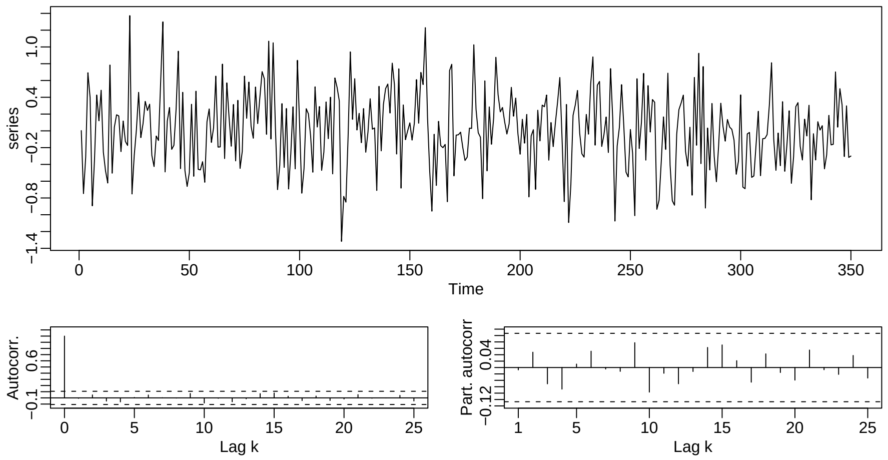
$$Y_t = 0.25 \cdot Y_{t-1} + E_t - 0.91 \cdot E_{t-1}; \quad \sigma^2_{E_t} = 0.214$$

Residuals of the fitted ARIMA(0,1,1) process:



The first ordinary (=first partial) autocorrelation clearly lies outside the confidence band. Thus the residuals cannot be considered independent.

Residuals of the fitted ARIMA(1,1,1) process:



These residuals no longer exhibit any undesired structure.

R commands:

ARIMA(0,1,1) model:

```
> mean(diff(log(d.varve)))                                  # -0.001271813
> r.varve.m1 <- arima(log(d.varve), order=c(0,1,1))
> r.varve.m1$code                                 # Code = 0, i.e. convergence
> r.varve.m1
```

Result:

```
Call:
arima(x = log(d.varve), order = c(0, 1, 1))

Coefficients:
          ma1
      -0.8421
s.e.   0.0411

sigma^2 estimated as 0.224:  log likelihood = -234.77,  aic = 473.53
```

```
> f.acf(resid(r.varve.m1))
```

ARIMA(1,1,1) model:

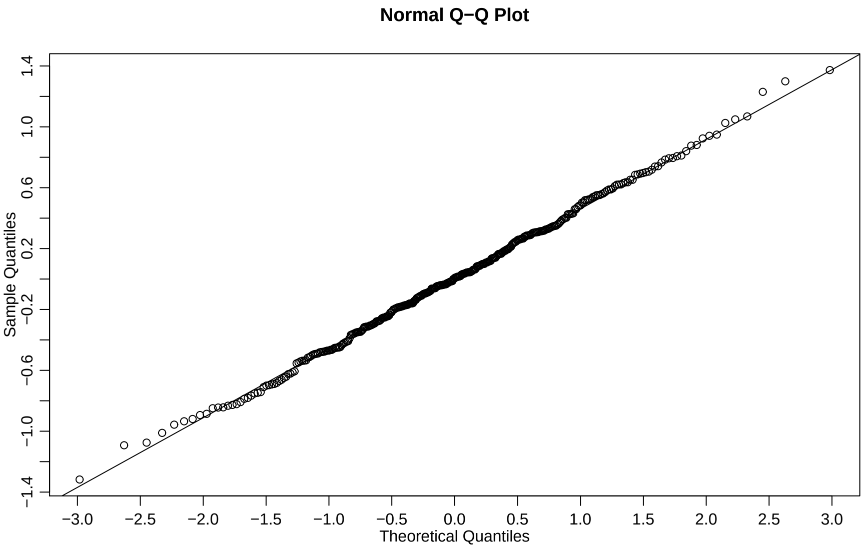
```
> r.varve.m2 <- arima(log(d.varve),order=c(1,1,1))
> r.varve.m2$code                                 # Code = 0, i.e. convergence
> r.varve.m2
```

Result:

```
Call:
arima(x = log(d.varve), order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.2461  -0.9140
s.e.  0.0590   0.0234

sigma^2 estimated as 0.2138:  log likelihood = -226.65,  aic = 459.3
```

```
> f.acf(resid(r.varve.m2))
```

c) The correlation structure of the residuals has already been examined in part b). The residuals of the ARIMA(1,1,1) process do not look normally distributed:

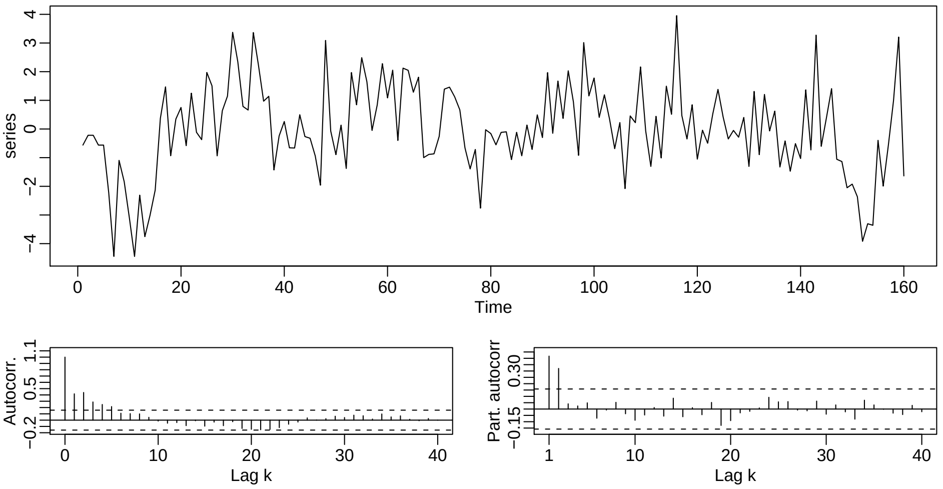


R commands:

```
> qqnorm(r.varve.m2$resid)
> qqline(r.varve.m2$resid)
```

2. a) The correlogram of the residuals shows that significant correlation is present. Consequently, all confidence intervals and tests in the output of lm can be wildly inaccurate. It is thus impossible for zoologists to conclude which explanatory variables are needed in the model.

```
> r.bel.lm <- lm(NURSING   ., data=beluga)
> summary(r.bel.lm)
Call:
lm(formula = NURSING ~ ., data = d.beluga)

Residuals:
     Min       1Q   Median       3Q      Max
-4.44568 -0.90180 -0.08505  1.09525  3.95477

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5602842  0.5502170   1.018  0.31012
PERIOD       0.0001998  0.0031937   0.063  0.95020
BOUTS        0.8784967  0.3336237   2.633  0.00932 **
LOCKONS      2.3903512  0.2035042  11.746  < 2e-16 ***
DAYNIGHT    -0.3416237  0.2510156  -1.361  0.17550
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.582 on 155 degrees of freedom
Multiple R-Squared: 0.842,     Adjusted R-squared: 0.8379
F-statistic: 206.5 on 4 and 155 DF,  p-value: < 2.2e-16


> d.resid <- ts(resid(r.bel.lm))
> f.acf(d.resid)
```



b) Due to the partial autocorrelations present, an AR(2) model for the residuals makes sense. Note that the ordinary autocorrelations make up a dampened sine curve, a property typical of AR processes. We can use the Burg algorithm to estimate both AR parameters.

Executing

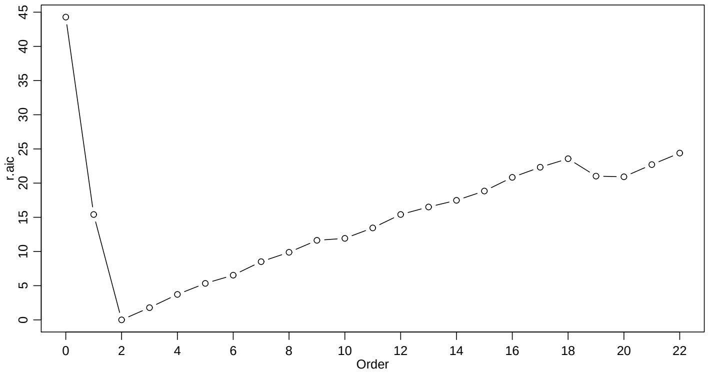
```
> r.burg <- ar(d.resid, method="burg", order.max=2, aic=F)
> str(r.burg)
```

in R, we obtain:

$\alpha_1 = 0.284, \alpha_2 = 0.321.$

Note: We can also use the AIC plot to determine the order of the process:

```
> r.aic <- ar(d.resid, method="burg")$aic
> plot(0:(length(r.aic)-1), r.aic, xlab="Order", type="b")
```



It seems that $p = 2$ is a good order to take.

c) We have

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 \cdot t + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \beta_4 X_{4,t} + E_t \qquad\qquad (t = 1, \ldots, 160) \\ \text{with} \quad & E_t = \alpha_1 E_{t-1} + \alpha_2 E_{t-2} + U_t \qquad U_t \text{ i.i.d. }, \; E[U_t] = 0, \; \mathrm{Var}[U_t] = \sigma^2 \;, \end{aligned}$$

where $Y_t =$ `NURSING`, $X_{1,t} = t =$ `PERIOD`, $X_{2,t} =$ `BOUTS`, $X_{3,t} =$ `LOCKONS` and $X_{4,t} =$ `DAYNIGHT`. Computing $Y_t^* = Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2}$:

$$\begin{aligned} Y_t^* &= Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} \\ &= \beta_0 + \beta_1 \cdot t + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \beta_4 X_{4,t} + E_t \\ &\quad -\alpha_1 \big(\beta_0 + \beta_1 \cdot (t-1) + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + \beta_4 X_{4,t-1} + E_{t-1}\big) \\ &\quad -\alpha_2 \big(\beta_0 + \beta_1 \cdot (t-2) + \beta_2 X_{2,t-2} + \beta_3 X_{3,t-2} + \beta_4 X_{4,t-2} + E_{t-2}\big) \\ &= \beta_0(1 - \alpha_1 - \alpha_2) + \beta_1(t - \alpha_1(t-1) - \alpha_2(t-2)) \\ &\quad +\beta_2(X_{2,t} - \alpha_1 X_{2,t-1} - \alpha_2 X_{2,t-2}) + \ldots + E_t - \alpha_1 E_{t-1} - \alpha_2 E_{t-2} \\ &= \beta_o^* + \beta_1 X_{1,t}^* + \beta_2 X_{2,t}^* + \beta_3 X_{3,t}^* + \beta_4 X_{4,t}^* + U_t \end{aligned}$$

The explanatory variables and the target must all be transformed as follows:

$$x_t^* = x_t - \widehat{\alpha}_1 x_{t-1} - \widehat{\alpha}_2 x_{t-2} = x_t - 0.284 \cdot x_{t-1} - 0.321 \cdot x_{t-2}$$

* This transformation, and the subsequent normal regression, can be performed in R using the following code. Note that the residuals now no longer exhibit correlation.

```
> t.ar <- r.burg$ar
> ## Transform the entire multivariate time series
> d.beluga.tr <- d.beluga - t.ar[1]*lag(d.beluga,-1) - t.ar[2]*lag(d.beluga,-2)
> ## Set new (meaningful) colnames
> colnames(d.beluga.tr) <- paste(colnames(d.beluga),".tr",sep="")
[1] "PERIOD.tr"   "BOUTS.tr"    "NURSING.tr"  "LOCKONS.tr"  "DAYNIGHT.tr"
> t.intercept <- rep((1-t.ar[1]-t.ar[2]),nrow(d.beluga.tr))
> r.lm.tr <- lm(NURSING.tr ~ -1 + t.intercept + PERIOD.tr + BOUTS.tr +
+               LOCKONS.tr + DAYNIGHT.tr, data=d.beluga.tr)
> f.acf(r.lm.tr$resid)
```

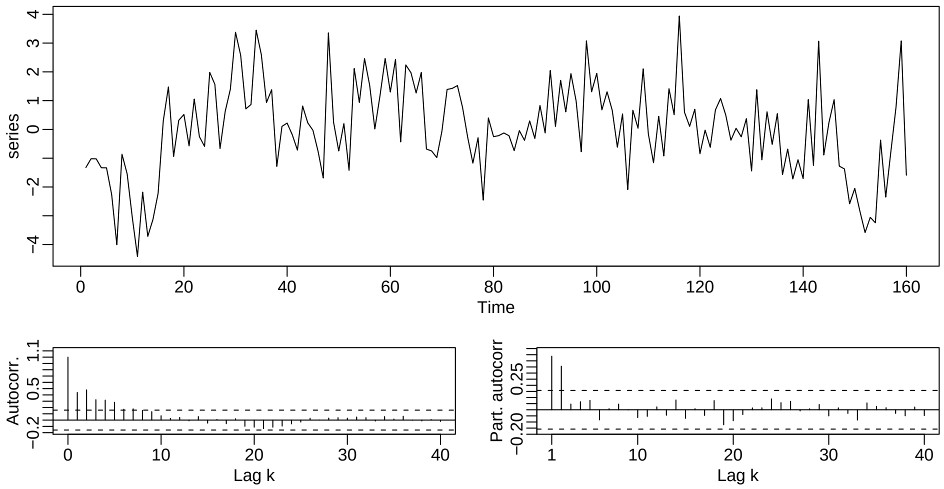
d) The procedure `gls()` can be used for much more general models than those you have already seen. The argument `correlation` can be used for specifying the correlation structure of the residuals. In principle an AR(p) model is merely a special case of the so-called ARMA(p, q) model taking $q = 0$ (cf. chap. 9). This explains the overly complex expression `corARMA(value=c(...,...), p=2, q=0, fixed=F)`. The AR coefficients computed in Part b) can be used as starting values by specifying them in the argument `value`. Errors in different time periods can be specified as being correlated by means of the argument `form=` $\sim$ `PERIOD` of `corARMA`. This is necessary, as the entries in the data matrix can be arranged in any way.

R output from `summary(r.bel.gls)`:

```
Generalized least squares fit by maximum likelihood
  Model: NURSING ~ BOUTS + LOCKONS + DAYNIGHT + PERIOD
  Data: d.beluga
      AIC      BIC   logLik
  560.396 584.9974 -272.198

Correlation Structure: ARMA(2,0)
 Formula: ~PERIOD
 Parameter estimate(s):
     Phi1      Phi2
0.2739964 0.3653668

Coefficients:
                Value Std.Error   t-value p-value
(Intercept)  1.3218871 0.7678364  1.721574  0.0871
BOUTS        0.2961684 0.3370588  0.878685  0.3809
LOCKONS      2.5681923 0.1964012 13.076257  <.0001
DAYNIGHT    -0.3080293 0.1549160 -1.988363  0.0485
PERIOD       0.0024982 0.0062754  0.398090  0.6911

 Correlation:
         (Intr) BOUTS  LOCKON DAYNIG
BOUTS    -0.303
LOCKONS  -0.101 -0.811
DAYNIGHT -0.014 -0.135  0.067
PERIOD   -0.607 -0.233  0.251  0.024

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.80055625 -0.58763749  0.01738824  0.65602061  2.49854120

Residual standard error: 1.577031
Degrees of freedom: 160 total; 155 residual
```

These coefficient estimates differ markedly from those in Part a). We obtain $\alpha_1 = 0.274$ and $\alpha_2 = 0.365$, which can be found in the above R output at *Parameter estimate(s)* (here labelled as *Phi1* and *Phi2*). In particular note that the standard errors of the explanatory variables sometimes differ greatly from those in the regression model.

Residual analysis:



There are only small differences to the model using ordinary regression. This is because residuals denote the difference between observations and model-derived fitted values – and the least squares estimates of coefficients do make sense here. It is merely the standard errors of the least squares method that are wrong. The residuals form an AR(2) process; thus the chosen correlation structure is correct.