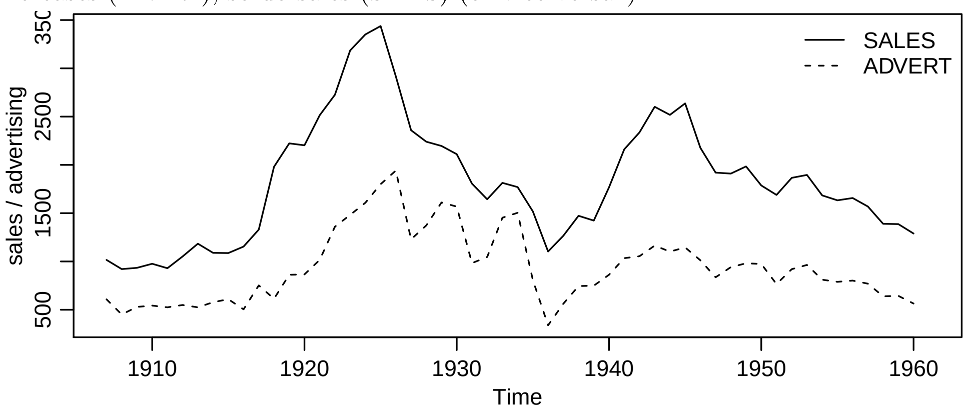
Solution to Series 5

1. a) In the time series plot, the dependence of the two series is evident. When advertising expenditure increases (`ADVERT`), so do sales (`SALES`) (or vice versa?).

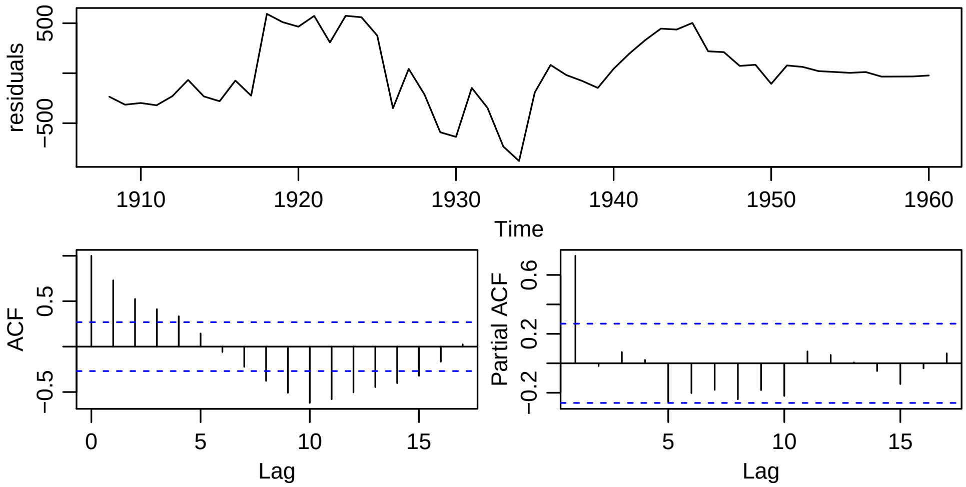


b) We regard the model

$$\texttt{SALES}_t = \beta_0 + \beta_1\,\texttt{ADVERT}_t + \beta_2\,\texttt{ADVERT}_{t-1} + E_t\,.$$

R commands and output:

```
> lm.advert1 <- lm(SALES ~ ADVERT + ADVERT1, data = ts.advert)
> summary(lm.advert1)
Call:
lm(formula = SALES ~ ADVERT + ADVERT1, data = ts.advert)

Residuals:
   Min     1Q Median     3Q    Max
-877.9 -224.4  -18.1  211.1  593.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 496.68768  135.76609   3.658  0.00061 ***
ADVERT        1.35243    0.22704   5.957 2.55e-07 ***
ADVERT1       0.08066    0.22753   0.355  0.72445
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 346.3 on 50 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.7081,       Adjusted R-squared: 0.6965
F-statistic: 60.66 on 2 and 50 DF,  p-value: 4.259e-14
> res.advert1 <- ts(resid(lm.advert1), start = 1908)
> layout(matrix(c(1, 1, 2, 3), 2, 2, byrow = TRUE))
> plot(res.advert1, ylab = "residuals")
> acf(res.advert1, plot = TRUE)
> acf(res.advert1, type = "partial", plot = TRUE)
```



The time series plot of residuals, and the corresponding correlograms, show that the errors are correlated and behave as an AR(1) process.

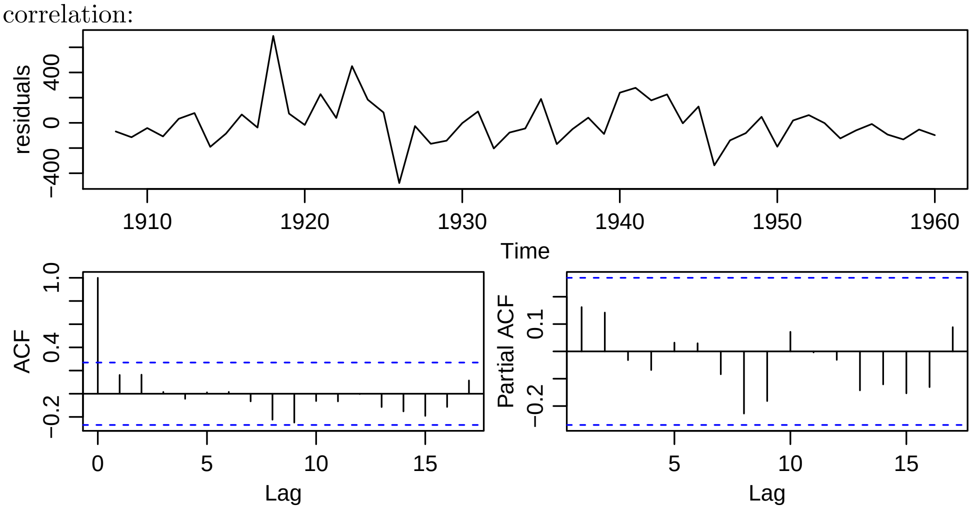
Consequences:

Correlation of residuals means that subsequently, the confidence intervals for coefficients β_0, β_1 and β_2 are inaccurate, which has an adverse effect on predictions and their precision. Since the setup of this exercise means that prediction is our main interest, this model really should be improved first.

c) We extend the model from part b) by introducing the variable $\texttt{SALES}_{t-1} = \texttt{SALES1}$:

$$\texttt{SALES}_t = \beta_0 + \beta_1\,\texttt{ADVERT}_t + \beta_2\,\texttt{ADVERT}_{t-1} + \beta_3\,\texttt{SALES}_{t-1} + E_t\,.$$

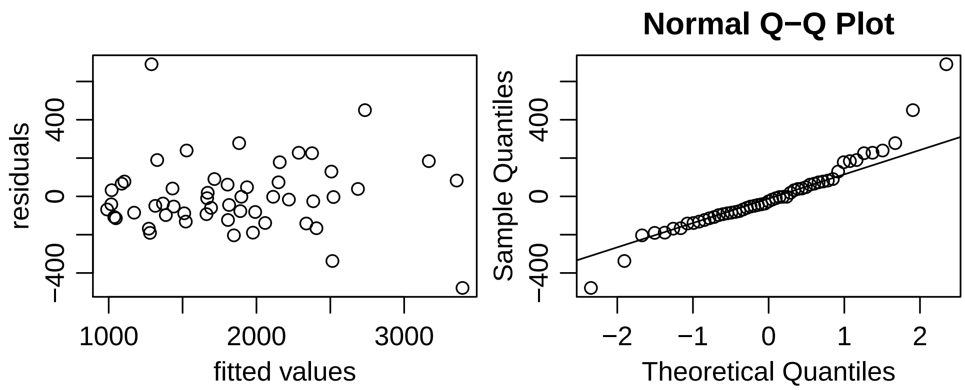
Note that the variable `SALES` serves both as a target and as an explanatory variable.
The summary of the `lm` object (not printed here) shows that all three explanatory variables are significant. The plot of residuals—and moreover, the correlograms—no longer exhibit unwanted correlation:



By including the additional variable $\texttt{SALES}_{t-1}$, we have succeeded in eliminating the autocorrelation of residuals from the model in part b).

Checking the assumption on the distribution of residuals:

```
> par(mfrow = c(1, 2), mar = c(3, 3, 2, 0.1))
> plot(fitted(lm.advert2), resid(lm.advert2), xlab = "fitted values",
+   ylab = "residuals")
> qqnorm(resid(lm.advert2))
> qqline(resid(lm.advert2))
```



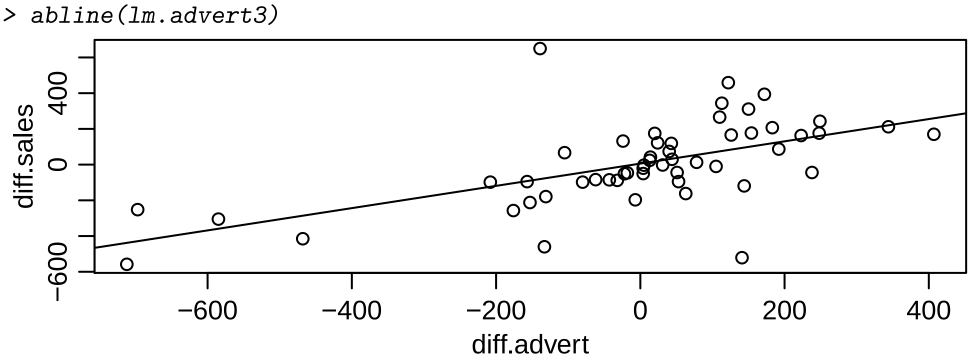
In the time series plot of residuals, and in the normal and Tukey-Anscombe plots, however, 2 outliers are visible. These observations should be looked at more closely. Simply omitting them is not an option, since this obviously causes problems for a time series. (Simply omitting outliers is a bad habit anyway…)

d) We regard the model

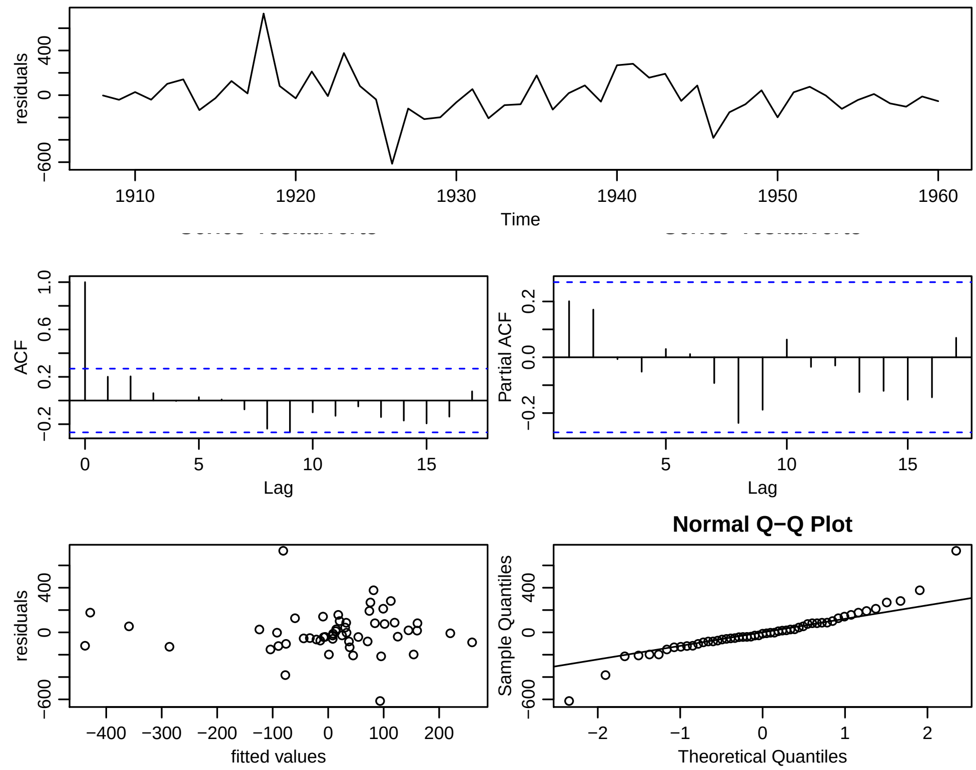
$$\texttt{D\_SALES}_t = \beta_0 + \beta_1\, \texttt{D\_ADVERT}_t + E_t\,,$$

where $\texttt{D\_SALES}_t = \texttt{SALES}_t - \texttt{SALES}_{t-1}$ and $\texttt{D\_ADVERT}_t = \texttt{ADVERT}_t - \texttt{ADVERT}_{t-1}$ are the first-order differences. The fitted line is shown in the following plot:

```
> diff.sales <- ts.advert[, "SALES"] - ts.advert[, "SALES1"]
> diff.advert <- ts.advert[, "ADVERT"] - ts.advert[, "ADVERT1"]
> lm.advert3 <- lm(diff.sales ~ diff.advert)
> plot(diff.advert, diff.sales, type = "p", xy.labels = FALSE, xy.lines = FALSE)
> abline(lm.advert3)
```



Analysis of residuals:



The correlograms do not exhibit any undesired correlation. All the ordinary and partial autocorrelations lie inside the confidence band.
However, the time series plot of residuals and the normal and Tukey-Anscombe plots again contain 2 outliers. The fitted model is

$$\texttt{D\_SALES}_t = 5.668 + 0.623 \cdot \texttt{D\_ADVERT}_t + E_t\,.$$

The intercept $\widehat{\beta}_0 = 5.668$ is not significant and could possibly be removed from the model.

e) Comparison of both models:

c) $\texttt{SALES}_t = \beta_0 + \beta_1\,\texttt{ADVERT}_t + \beta_2\,\texttt{ADVERT}_{t-1} + \beta_3\,\texttt{SALES}_{t-1} + E_t$

d) $\texttt{D\_SALES}_t = \beta_0 + \beta_1\,\texttt{D\_ADVERT}_t + E_t$;
corresponds to the model
$\texttt{SALES}_t = \beta_0 + \beta_1\,\texttt{ADVERT}_t - \beta_1\,\texttt{ADVERT}_{t-1} + \texttt{SALES}_{t-1} + E_t$

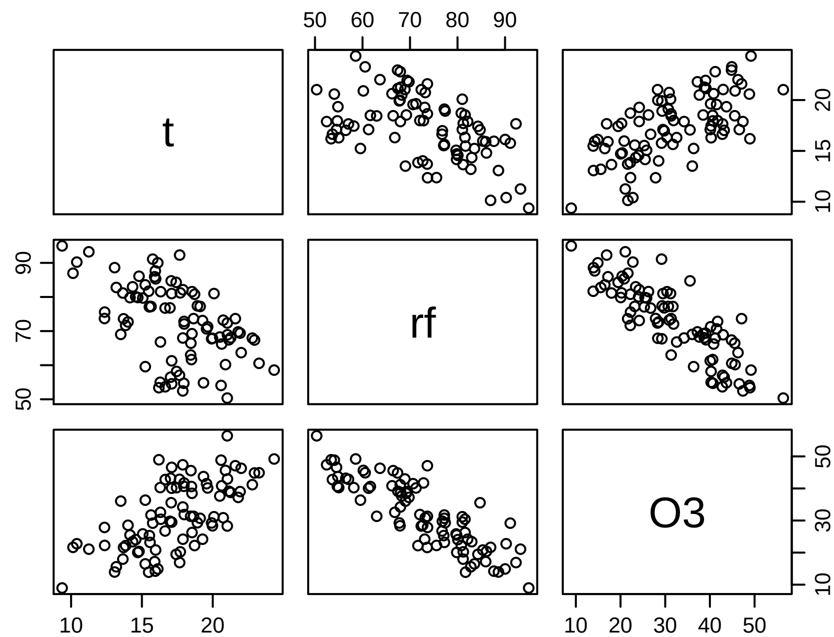
- In both models the errors satisfy the assumption of independence. However, both models breach the assumption on their distribution, and there are outliers.
- Both models contain the same explanatory variables, but the model in part d) contains restrictions on the regression coefficients (only 2 coefficients are estimated here!).
- The second model is somewhat simpler to interpret than the first one. However, model d) does not fit as well as model c): its R^2 is only 0.344 compared to 0.915 in model c).

Notes — Outlook:
In this example it is difficult to determine which series influences the other one. The theory distinguishes two settings:

- Both series influence each other. Such models are called **bivariate autoregressive models**.
- Only one of the series (y_t) depends on the other one (x_t). Such models are termed **transfer function models**.

The connection between the two time series can be investigated using so-called **cross-correlations**, which you will encounter later. In both cases, however, both y_t and x_t must be assumed to be **stationary** time series.

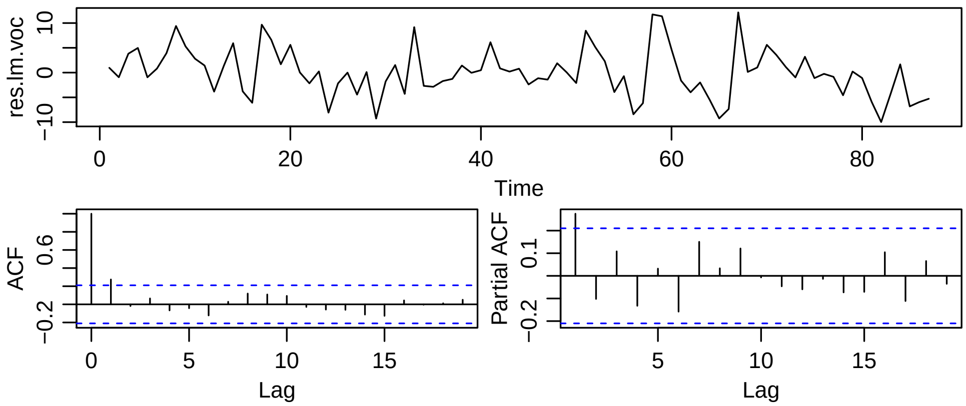
2. a)



b) Both explanatory variables are significant:

```
> lm.voc <- lm(O3 ~ ., data = ts.voc)
> summary(lm.voc)

Call:
lm(formula = O3 ~ ., data = ts.voc)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9833 -2.7762 -0.0024  2.0879 12.1565

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.60570    6.54426   9.719 2.12e-15 ***
t            0.99033    0.19733   5.019 2.87e-06 ***
rf          -0.67404    0.05577 -12.085  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.917 on 84 degrees of freedom
Multiple R-squared: 0.7905,       Adjusted R-squared: 0.7855
F-statistic: 158.5 on 2 and 84 DF,  p-value: < 2.2e-16
```

However, the residuals behave badly: they are not uncorrelated:



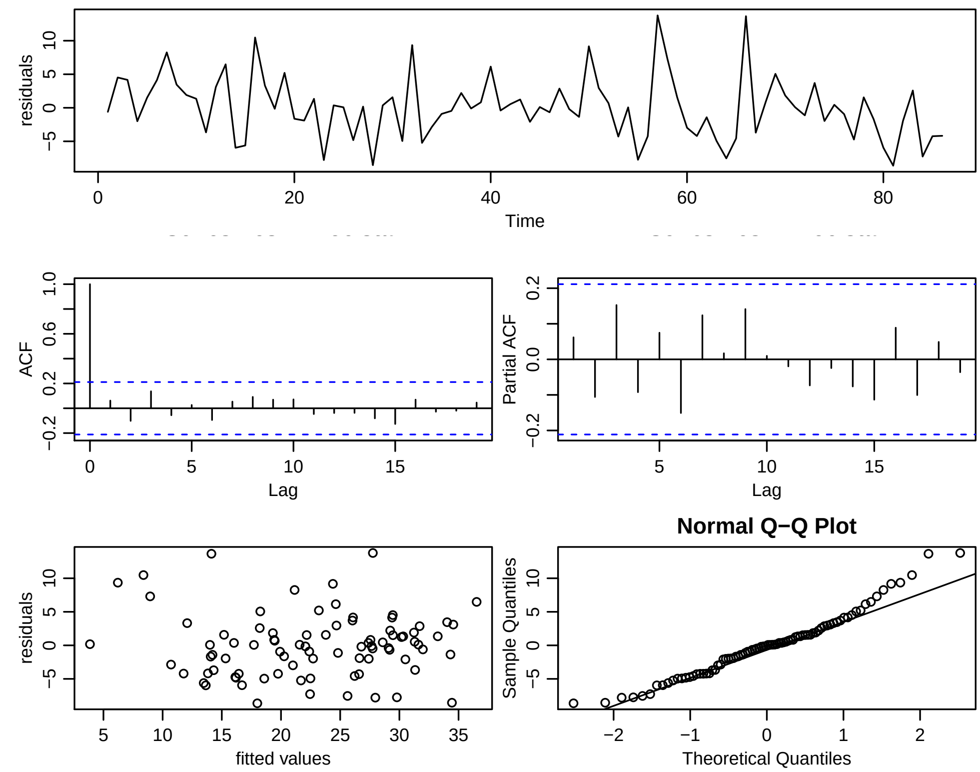
c) The PACF plot of the residuals in part b) indicates that the residuals should be modeled as an AR(1) process. We fit its parameter and transform the model accordingly:

```
> alpha <- ar(res.lm.voc, method = "yw", order.max = 1)$ar
> ts.voc.star <- ts.voc - alpha[1] * lag(ts.voc, -1)
> colnames(ts.voc.star) <- colnames(ts.voc)
```

We can now perform a least squares regression with the transformed time series:

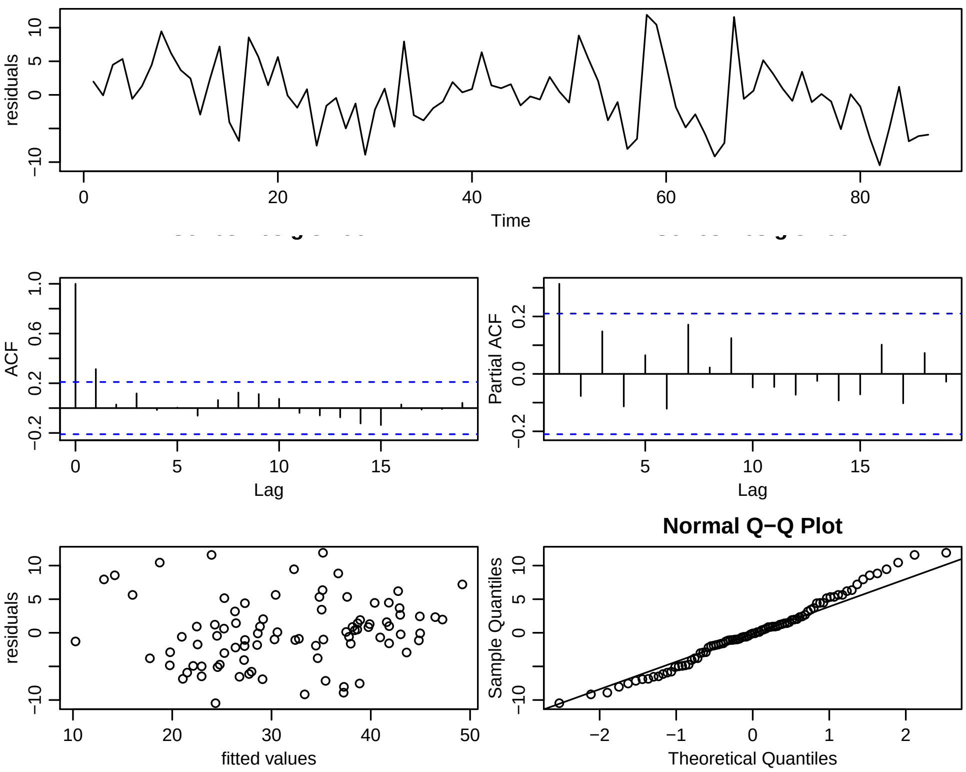
```
> lm.voc.star <- lm(O3 ~ ., data = ts.voc.star)
> summary(lm.voc.star)

Call:
lm(formula = O3 ~ ., data = ts.voc.star)

Residuals:
   Min     1Q Median     3Q    Max 
-8.638 -3.495 -0.018  2.131 13.774 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 44.68763    5.58291   8.004 6.38e-12 ***
t            0.93329    0.23743   3.931 0.000175 ***
rf          -0.63378    0.06446  -9.833 1.42e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.729 on 83 degrees of freedom
Multiple R-squared: 0.7127,       Adjusted R-squared: 0.7057 
F-statistic: 102.9 on 2 and 83 DF,  p-value: < 2.2e-16
```

Both explanatory variables are still significant. The residual analysis gives good results now—up to few outliers:



d)
```
> library(nlme)
> times <- time(ts.voc)
> gls.voc <- gls(O3 ~ ., data = ts.voc, method = "ML",
+   correlation = corARMA(form = ~ times, p = 1, q = 0))
> summary(gls.voc)
Generalized least squares fit by maximum likelihood
  Model: O3 ~ .
  Data: ts.voc
      AIC      BIC   logLik
  523.246 535.5756 -256.623

Correlation Structure: AR(1)
 Formula: ~times
 Parameter estimate(s):
      Phi
0.3162185

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 60.84755  7.822679  7.778352   0e+00
t            0.94651  0.236042  4.009938   1e-04
rf          -0.62598  0.065669 -9.532313   0e+00

 Correlation:
   (Intr) t
t  -0.851
rf -0.891  0.534

Standardized residuals:
       Min          Q1         Med          Q3         Max
```

```
-2.15109265 -0.61024167 -0.01873279  0.52524855  2.44588076

Residual standard error: 4.86868
Degrees of freedom: 87 total; 84 residual
```



e)

	OLS		C & O		GLS	
	estimate	std. err.	estimate	std. err.	estimate	std. err.
$\hat{\beta}_0$	63.61	6.54	61.61	7.7	60.85	7.82
$\hat{\beta}_1$	0.99	0.2	0.93	0.24	0.95	0.24
$\hat{\beta}_2$	−0.67	0.06	−0.63	0.06	−0.63	0.07

3. a) i) $X_t = t + E_t$ is **not stationary** since the expectation value $E[X_t] = E[t + E_t] = t$ is not stationary.

ii) We have

$$Y_t = X_t - X_{t-1} = t + E_t - (t - 1 + E_{t-1}) = 1 + E_t - E_{t-1}\,,$$

hence Y_t is a **stationary** MA(1) process with $\mu = 1$ and $\beta_1 = -1$.

iii) The time series $Z_t = X_t - t$ is **stationary**: $Z_t = X_t - t = t + E_t - t = E_t$.

b) • **Series** Y_t has autocovariances

$$\begin{aligned}
\gamma_{11}(k) &= \mathrm{Cov}(Y_t, Y_{t+k}) = \mathrm{Cov}(1 + E_t - E_{t-1}, 1 + E_{t+k} - E_{t+k-1}) \\
&= \mathrm{Cov}(E_t, E_{t+k}) - \mathrm{Cov}(E_t, E_{t+k-1}) - \mathrm{Cov}(E_{t-1}, E_{t+k}) + \mathrm{Cov}(E_{t-1}, E_{t+k-1}) \\
&= \begin{cases} 2\sigma^2 & k = 0 \\ -\sigma^2 & k = \pm 1 \\ 0 & |k| > 1 \end{cases}
\end{aligned}$$

Therefore we find the autocorrelations

$$\begin{aligned}
\rho_{11}(0) &= 1\,, \\
\rho_{11}(\pm 1) &= \frac{\gamma_{11}(1)}{\gamma_{11}(0)} = -\frac{1}{2}\,, \\
\rho_{11}(k) &= 0\,, \quad \text{für } |k| > 1\,.
\end{aligned}$$

- **Series Z_t:** Since $Z_t = E_t$ is white noise, we have, trivially, $\gamma_{22}(0) = \sigma^2$ and $\gamma_{22}(k) = 0$ for $|k| \geq 1$. Hence we have $\rho_{22}(0) = 1$ and $\rho_{22}(k) = 0$ for $|k| \geq 1$.
- **Cross-correlations between Y_t and Z_t:** The cross-covariances are

$$\begin{aligned}
\gamma_{12}(k) &= \mathrm{Cov}(Y_{t+k}, Z_t) = \mathrm{Cov}(1 + E_{t+k} - E_{t+k-1}, E_t) \\
&= \mathrm{Cov}(E_{t+k}, E_t) - \mathrm{Cov}(E_{t+k-1}, E_t) \\
&= \begin{cases} \sigma^2 & k = 0 \\ -\sigma^2 & k = 1 \\ 0 & \text{sonst} \end{cases}
\end{aligned}$$

Hence we find the cross-correlations

$$\rho_{12}(k) = \frac{\gamma_{12}(k)}{\sqrt{\gamma_{11}(0)\gamma_{22}(0)}} = \begin{cases} 1/\sqrt{2} = 0.71 & k = 0 \\ -1/\sqrt{2} = -0.71 & k = 1 \\ 0 & \text{otherwise} \end{cases}$$

In this example, the cross-correlation $\rho_{12}(k)$ describes the connection between Y_{t+k} (MA(1) model) and E_t (white noise). By construction of the series, the cross-correlation always vanishes, except for lags 0 and 1.

c) The simulated processes Y_t and Z_t behave as we expect from theory.
Simulation with R:

```
> ts.E <- ts(rnorm(201))
> ts.X <- (1:201) + ts.E
> ts.Y <- diff(ts.X)
> ts.Z <- ts.E
```

Plot of the auto- and crosscorrelations:

```
> acf(ts.intersect(ts.Y, ts.Z), ylim = c(-1, 1), plot = TRUE)
```