5. Computer exercises

Data: Alcohol consumption

The goal is to build a regression model that explains alcohol consumption expenditures per capita (Q1CPC) with the real price index of alcohol (R1C) and total consumption expenditures per capita (QTOTALPC). The data we use in these demo exercises is real data from Finland and the data consists of yearly time series between 1950 and 1981. The prices of different years have been inflation adjusted to be comparable with the fixed prices of the year 1975.

An elementary model in economics is the log-linear regression model, defined as,

$$\log(\text{Q1CPC}) = \beta_0 + \beta_1 \log(\text{R1C}) + \beta_2 \log(\text{QTOTALPC}) + \varepsilon, \tag{1}$$

where $\log(\cdot)$ is used to denote the natural logarithm.

The model (1) will be estimated in Exercise 5.1 for years 1950-1981. A problem with model (1) is that the alcohol legislation changed in the beginning of the year 1969. This motivates us to add an indicator (or dummy) variable LAW to the model (1). The variable LAW represents the change in the legislation and it is defined as follows.

$$\begin{aligned} \text{LAW} = &\quad 0 \quad \text{for years 1950-1968} \\ \text{LAW} = &\quad 1 \quad \text{for years 1969-1981} \end{aligned}$$

Hence, we obtain the model

$$\log(\text{Q1CPC}) = \beta_0 + \beta_1 \log(\text{R1C}) + \beta_2 \log(\text{QTOTALPC}) + \beta_3 \text{LAW} + \varepsilon. \tag{2}$$

The indicator variable LAW tries to take into account the jump in the level of the alcohol expenditures. Namely, in model (2) the constant term is of the form,

$$\begin{aligned} &\beta_0 && \text{in years 1950-1968,} \\ &\beta_0 + \beta_3 && \text{in years 1969-1981.} \end{aligned}$$

The presumption is that the regression coefficient β_3 is statistically significant and positive.

The model (2) will be estimated in Exercise 5.2 for the years 1950-1981. However, by performing some regression diagnostics, we find that model (2) is not satisfactory when trying to describe the behavior of the variable log(Q1CPC).

The problem with model (2) is that the residuals of the estimated model are heavily autocorrelated. We can sometimes get rid of the autocorrelation issue by utilizing so-called difference models.

Therefore, we try the following difference model to describe alcohol expenditures,

$$D\log(\text{Q1CPC}) = \beta_0 + \beta_1 D\log(R1C) + \beta_2 D\log(\text{QTOTALPC}) + \beta_3 D\text{LAW} + \varepsilon. \tag{3}$$

The difference operation on the dummy variable LAW produces a so called impulse dummy. The model (3) is estimated in Exercise 5.3 for the years 1950-1981. Note that, in the model

(3) we have a different response variable than in the models (1) and (2). Thus, the different models are not directly comparable.
In Exercise 5.4, we modify model (2) by adding dynamic components. We try the following regression model for the alcohol expenditures,

$$\begin{aligned}\log(\mathrm{Q1CPC}_t) =&\beta_0 + \beta_1 \log(\mathrm{Q1CPC}_{t-1})+\\ &\beta_2 \log(\mathrm{R1C}_t) + \beta_3 \log(\mathrm{R1C}_{t-1})+\\ &\beta_4 \log(\mathrm{QTOTALPC}_t) + \beta_5 \log(\mathrm{QTOTALPC}_{t-1})+\\ &\beta_6 \mathrm{LAW}_t + \beta_7 \mathrm{LAW}_{t-1} + \varepsilon_t,\end{aligned} \tag{4}$$

where X_{t-1} is the variable X_t with lag one. Note that, the explanatory variable $\log(\mathrm{Q1CPC}_{t-1})$ is not independent of the error term ε_t. Therefore, the standard assumptions are not satisfied and it is not possible to draw conclusions from the coefficient of determination directly.
In Exercise 5.4, we study the autocorrelation of the residuals by using Breusch–Godfrey test, which is similar to Ljung-Box test. In general, Ljung-Box can be applied to test autocorrelation of the residuals of fitted SARIMA models. However, it is not justified to use Ljung-Box test in regression diagnostics, if the model involves endogenous explanatory variables, that is, variables that are not independent of the residuals. On the other hand, Breusch–Godfrey test is applicable in these situations. In the Breusch–Godfrey test, the null hypothesis is that there is no autocorrelation up to the lag p. The test can be conducted in R with the function `bgtest()`, which is implemented in the package **lmtest**.

Demo exercises

5.1 Estimate model (1) and study the goodness of fit.

Solution. The file `alcohol.txt` contains observations for the years 1950-1981, where **LR1C**, **LQ1CPC** and **LQTOTALPC** are the logarithms of the variables presented earlier.

```
library(car) # function vif() is contained in the package car
library(forecast)
library(lmtest)
alko<-read.table("alcohol.txt",header=T,sep="\t")
LR1C<-ts(alko$LR1C,start=1950)
LQ1CPC<-ts(alko$LQ1CPC,start=1950)
LQTOTALPC<-ts(alko$LQTOTALPC,start=1950)

model1<-lm(LQ1CPC~LR1C+LQTOTALPC)
summary(model1)

Call:
lm(formula = LQ1CPC ~ LR1C + LQTOTALPC)
Residuals:
      Min        1Q    Median        3Q       Max 
-0.146202 -0.083305 -0.009638  0.082679  0.161945 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.89170    1.98506  -1.457   0.1559    
LR1C        -1.00346    0.37255  -2.693   0.0116 *  
LQTOTALPC    1.46489    0.05904  24.813   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09986 on 29 degrees of freedom
Multiple R-squared:  0.9642,Adjusted R-squared:  0.9617 
F-statistic: 390.1 on 2 and 29 DF,  p-value: < 2.2e-16
```

Comments:

- The regression coefficients corresponding to the price variable LR1C and total expenditures variable LQTOTALPC are statistically significant with 5% level of significance.
- The signs of the regression coefficients of the price and total expenditures variables are as expected: the coefficient of the price variable is negative and the coefficient of the total expenditures variable is positive.
- Interpretations of the regression coefficients as elasticities:
  - If the price goes up by 1 %, then the alcohol expenditures are reduced by 1.003 %.
  - If the total expenditures are increased by 1 %, then the alcohol expenditures are increased by 1.465 %
- The coefficient of determination of the model is 96.42 %.

Next, we study the normality of the residuals, Cook’s distances and compare the fitted model with the original time series.

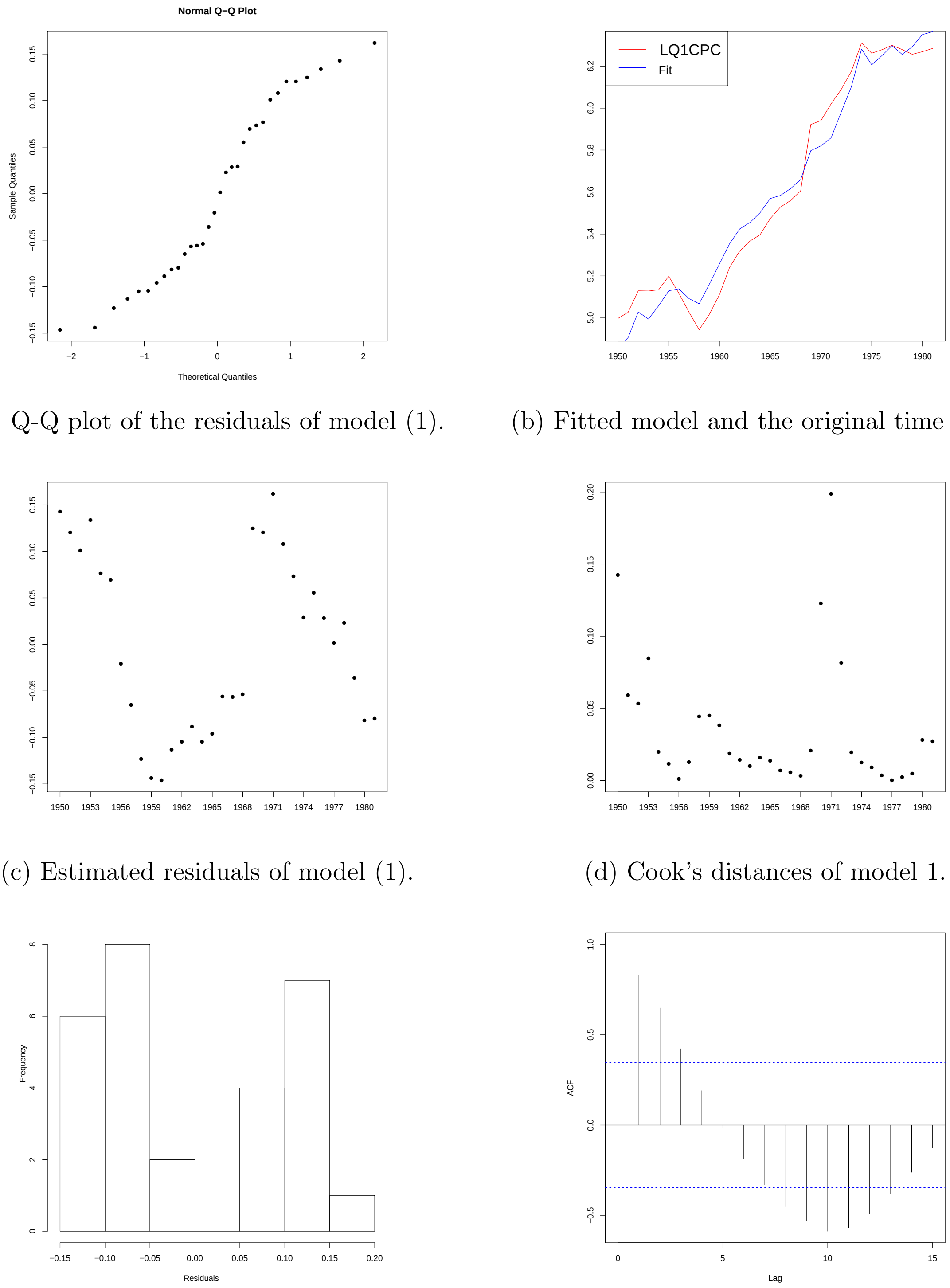
```
qqnorm(model1$residuals)
hist(model1$residuals)
acf(model1$residuals,main="")

plot(model1$residuals,type="p",ylab="Residuals",xlab="Year",pch=16,
     xaxt="n")
axis(1,at=seq(from=1,to=32,by=3),labels=seq(from=1950,to=1981,by=3))

fit <- ts(predict(model1),start=1950)
```

```
#predict() corresponds to model1$fitted.values

plot(LQ1CPC,col="red",xlab="Time",ylab="")
lines(fit,col="blue")
plot(cooks.distance(model1),ylab="Cook’s distances",xlab="Index",
     pch=16)

vif(model1)

LR1C      LQTOTALPC
1.16005   1.16005
```

Comments for the model (1):

- By Figures 1a and 1e, the residuals do not look normally distributed.
- By Figures 1c and 1f, the residuals seem to be heavily correlated.
- By the variance inflation factor (VIF), multicollinearity is not a problem here.
- The reason for the correlatedness of the residuals can be seen from Figure 1b, where the fitted curve stays above and below the response variable LQ1CPC for long time periods.
- The model does not take account of the change in the legislation (the beginning of the year 1969). This is also visible in the Cook’s distances (Figure 1d).

Thus, the model (1) cannot be considered to be sufficient. **Note that, in the context of linear regression, it is not allowed to extrapolate outside the interval of observations without a valid justification!**



(a) Q-Q plot of the residuals of model (1).

(b) Fitted model and the original time series.

(c) Estimated residuals of model (1).

(d) Cook's distances of model 1.

(e) Estimated residuals of model (1).

(f) ACF of the estimated residuals of model 1.

Figure 1: Graphical regression diagnostics for model 1.

5.2 Estimate and study model (2).

Solution. Estimate the model (2).

```
zeros <- rep(0,19)
ones <- rep(1,13)
LAW=ts(c(zeros,ones),start=1950)
model2=lm(LQ1CPC~LR1C+LQTOTALPC+LAW)

summary(model2)

Call:
lm(formula = LQ1CPC ~ LR1C + LQTOTALPC + LAW)

Residuals:
      Min        1Q    Median        3Q       Max
-0.144576 -0.031179  0.008463  0.048923  0.086176

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.21798    1.40182   0.155  0.87754
LR1C        -0.88570    0.24650  -3.593  0.00124 **
LQTOTALPC    1.04355    0.07818  13.349 1.16e-13 ***
LAW          0.31738    0.05106   6.216 1.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06588 on 28 degrees of freedom
Multiple R-squared:  0.9849,Adjusted R-squared:  0.9833
F-statistic: 610.5 on 3 and 28 DF,  p-value: < 2.2e-16
```

Comments:

- The regression coefficients corresponding to the price variable LR1C and the total expenditures variable LQTOTALPC are statistically significant with 5% level of significance.
- The estimates for the regression coefficients differ from the estimates of the model (1).
- The signs of the regression coefficients for the price and total expenditures are as expected: the coefficient of the price variable is negative and the coefficient of the total expenditures variable is positive.
- Interpretations of the regression coefficients as elasticities:

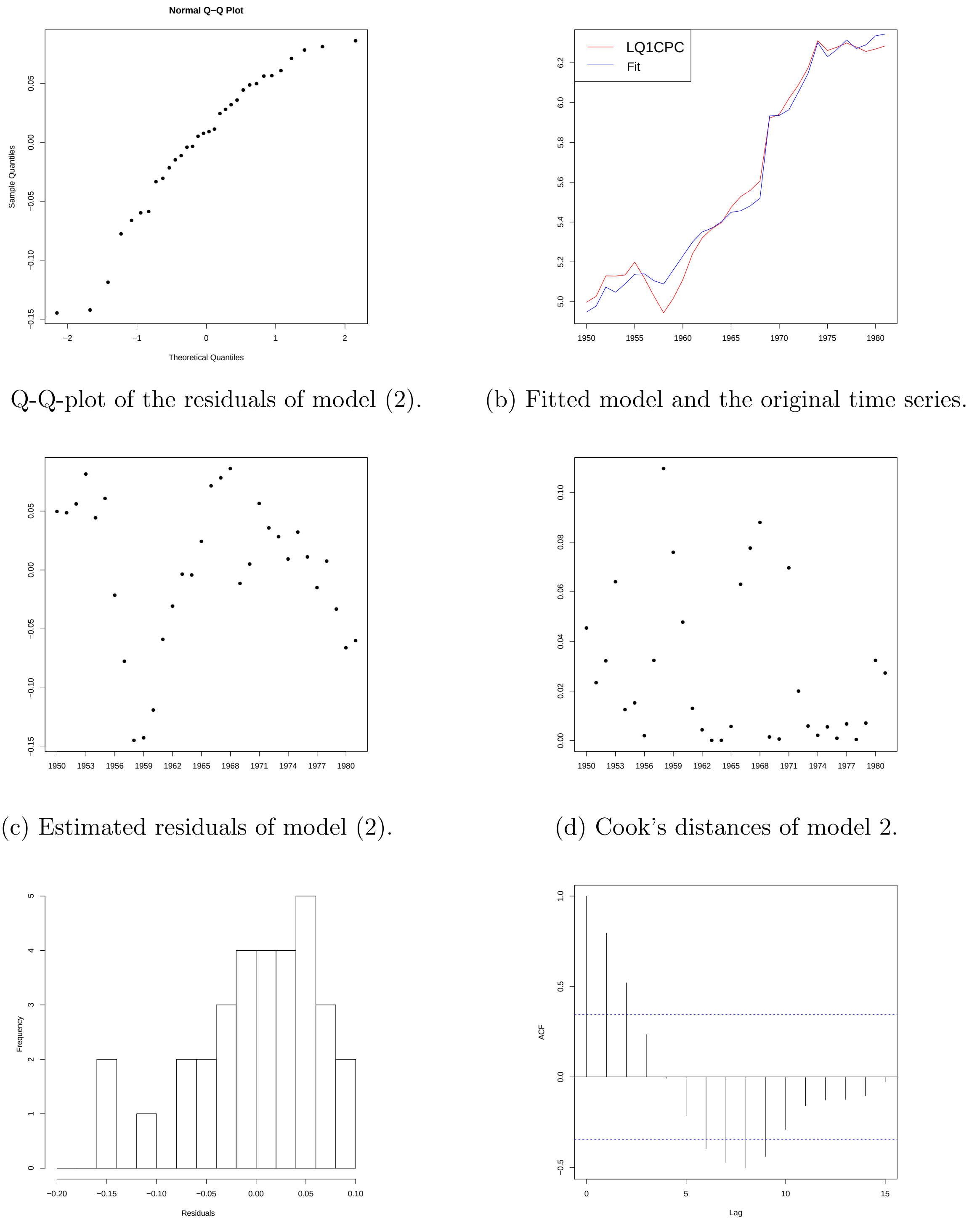
  - If the price goes up by 1 %, then the alcohol expenditures are reduced by 0.89 %.
  - If the total expenditures are increased by 1 %, then the alcohol expenditures are increased by 1.04 %.
- The regression coefficient 0.318 corresponding to LAW is statistically significant with 5% level of significance.
- The sign of the regression coefficient of LAW is as expected.
- The coefficient of determination has increased to 98.49 %.

Study the goodness of the fit:

```
hist(model2$residuals,breaks=seq(from=-0.2,to=0.1,by=0.02),
     xlab="Residuals",ylab="Frequency",main=" ")
acf(model2$residuals,main="")
qqnorm(model2$residuals,pch=16)
plot(model2$residuals,type="p",ylab="Residuals",xlab="Year",pch=16,
     xaxt="n")
axis(1,at=seq(from=1,to=32,by=3),labels=seq(from=1950,to=1981,by=3))
fit2 <- ts(predict(model2),start=1950)
plot(LQ1CPC,col="red",xlab="Time",ylab="")
lines(fit2,col="blue")
legend("topleft", legend=c("LQ1CPC", "Fit"),
       col=c("red","blue"),lty=c(1,1),cex=1.8)

plot(cooks.distance(model2),ylab="Cook's distance",xlab="Index",
     pch=16)
vif(model2)
LR1C        LQTOTALPC     LAW
1.166943    4.674154      4.636612
```

Comments related to the model (2):

- By Figure 2a, the distribution of the residuals does not seem to be normal. The histogram of the residuals is skewed, which is evidence against normality (Figure 2e).
- By Figures 2c and 2f, the residuals are strongly correlated.
- By VIF, there is no problem with multicollinearity.
- The reason for the correlation of the residuals can be seen from Figure 2b. The fitted curve stays long time periods above and below the values of the response variable LQ1CPC.
- The model takes into account the change in the legislation (beginning of 1969).
- By the regression diagnostics, this model is not satisfactory in explaining the alcohol expenditures.



(a) Q-Q-plot of the residuals of model (2).

(b) Fitted model and the original time series.

(c) Estimated residuals of model (2).

(d) Cook's distances of model 2.

(e) Estimated residuals of model (2).

(f) ACF of the estimated residuals of model 2.

Figure 2: Graphical regression diagnostics for model 2.

5.3 Estimate and study model (3).

Solution. Compute the differenced variables, estimate the model and study the goodness of the fit similarly as in the previous exercises.

```
DLQ1CPC <- diff(LQ1CPC)
DLR1C <- diff(LR1C)
DLQTOTALPC <- diff(LQTOTALPC)
DLAW <- diff(LAW)

model3<-lm(DLQ1CPC~DLR1C+DLQTOTALPC+DLAW)
summary(model3)

Call:
lm(formula = DLQ1CPC ~ DLR1C + DLQTOTALPC + DLAW)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08272 -0.01250  0.00000  0.02027  0.05541

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.010343   0.009273  -1.115    0.275
DLR1C       -0.816697   0.133193  -6.132 1.50e-06 ***
DLQTOTALPC   1.372390   0.225936   6.074 1.74e-06 ***
DLAW         0.196386   0.037765   5.200 1.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03383 on 27 degrees of freedom
Multiple R-squared:  0.8264,Adjusted R-squared:  0.8071
F-statistic: 42.84 on 3 and 27 DF,  p-value: 2.113e-10


vif(model3)
DLR1C        DLQTOTALPC       DLAW
1.061327     1.273221         1.206251
```

Comments:

- All the coefficients of the difference model (3) are statistically significant (with the exception of the constant term) with 5% level of significance.
- The estimates for the model (3) are clearly different when compared to the estimates of the model (2).

- The signs of the regression coefficients of the price and total expenditures variables are as expected: the coefficient of the price variable is negative and the coefficient of the total expenditures variable is positive.
- Interpretations of the regression coefficients as elasticities:
  - If the price goes up by 1%, then the alcohol expenditures are reduced by 0.817 %.
  - If the total expenditures are increased by 1 %, then the alcohol expenditures are increased by 1.37 %.
- The coefficient of the instant effect of the dummy variable LAW is 0.196.
- The coefficient of determination is 82.6 %.

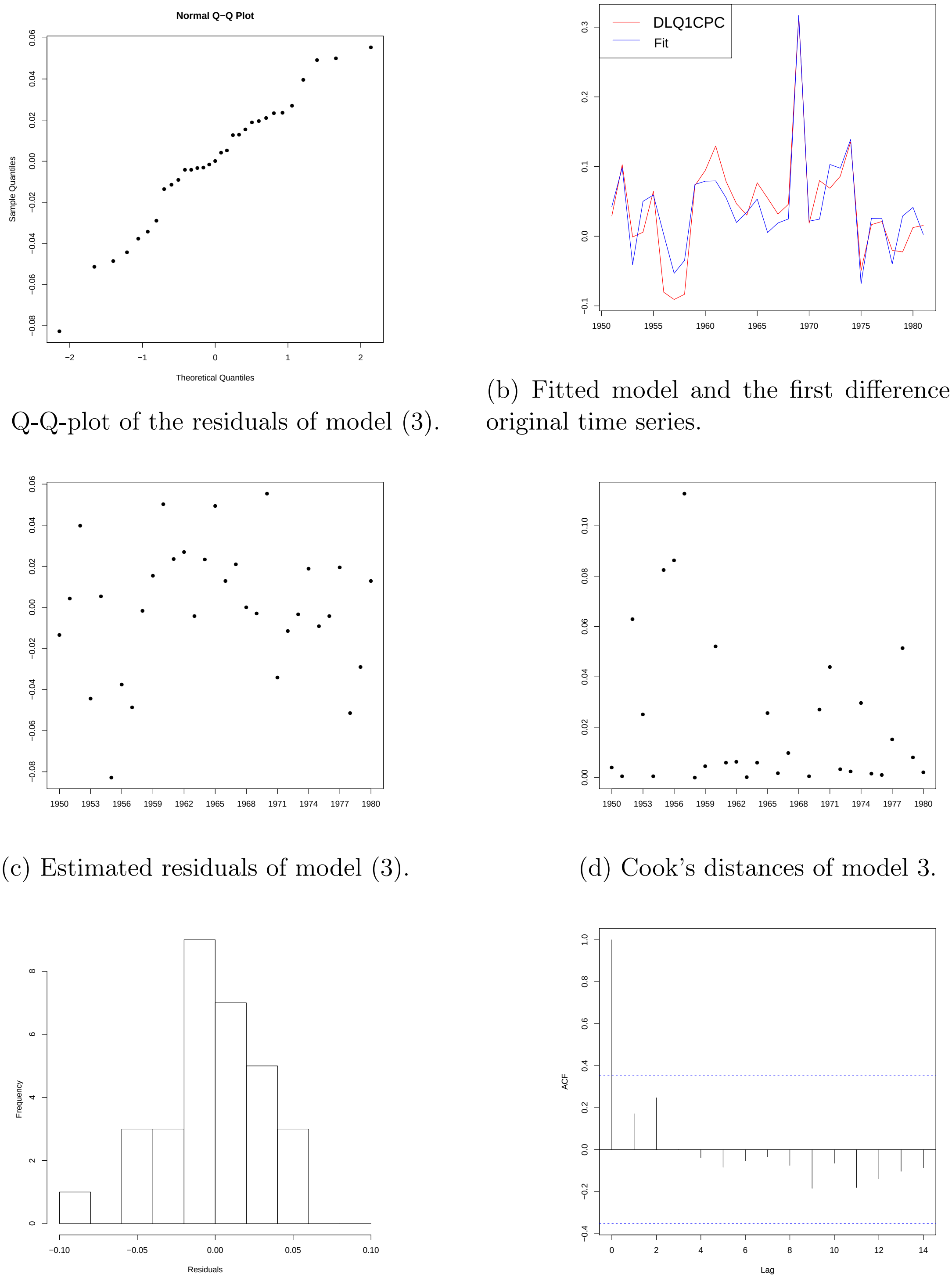
Remark: The coefficient of determination of the difference model (3) is not comparable with the coefficients of determinations corresponding to models (1) and (2), since the response variable is not the same.

Comments related to the model (3):

- By Figures 3a and 3e, the residuals could be normally distributed.
- By Figures 3c and 3f, the residuals are not correlated
- By the Breusch-Godfrey test, the residuals are not correlated: the null hypothesis is accepted with 5% level of significance for all lags. The test can be performed as follows.

```
install.packages("lmtest")
library(lmtest)

model3_bg <- rep(NA,27)

# Breusch-Godfrey can be performed up to order:
# (sample size) - (number of estimated parameters) = 31-4 = 27

for (i in 1:27)
{
  model3_bg[i]= bgtest(model3, order=i)$p.value
}

which(model3_bg > 0.05)
# Null hypothesis of no autocorrelation accepted with all lags
```

- By VIF, multicollinearity is not a problem.
- By plotting the residuals against the fitted values (or observed values, or time), there does not seem to be evidence of heteroscedasticity.
- By the regression diagnostics, the model is satisfactory.



(a) Q-Q-plot of the residuals of model (3).

(b) Fitted model and the first difference of the original time series.

(c) Estimated residuals of model (3).

(d) Cook's distances of model 3.

(e) Estimated residuals of model (3).

(f) ACF of the estimated residuals of model 3.

Figure 3: Graphical regression diagnostics for model 3.

5.4 Estimate and study model (4).

Solution. Estimate the model (4). In the R-print below, we have the regression coefficients in order $\beta_0, \beta_1, \ldots, \beta_7$. Note that, when variables of the form X_{t-1} and X_t are considered, the last and the first observation are omitted, respectively, when the model (4) is estimated.

```
n <- nrow(alko)

model4 <- lm(LQ1CPC[-1]~ LQ1CPC[-n] + LR1C[-1] +LR1C[-n]+
               LQTOTALPC[-1] + LQTOTALPC[-n]+ LAW[-1] +LAW[-n])

summary(model4)

Call:
lm(formula = LQ1CPC[-1] ~ LQ1CPC[-n] + LR1C[-1] + LR1C[-n] +
    LQTOTALPC[-1] + LQTOTALPC[-n] + LAW[-1] + LAW[-n])

Residuals:
      Min        1Q    Median        3Q       Max
-0.073939 -0.013803  0.002322  0.013555  0.071924

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.36137    0.86190  -0.419 0.678911
LQ1CPC[-n]     0.91197    0.10461   8.718 9.52e-09 ***
LR1C[-1]      -0.82927    0.16234  -5.108 3.57e-05 ***
LR1C[-n]       0.71352    0.17579   4.059 0.000486 ***
LQTOTALPC[-1]  1.46946    0.24887   5.905 5.10e-06 ***
LQTOTALPC[-n] -1.31522    0.27235  -4.829 7.13e-05 ***
LAW[-1]        0.17751    0.04468   3.973 0.000602 ***
LAW[-n]       -0.19188    0.04707  -4.077 0.000465 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03442 on 23 degrees of freedom
Multiple R-squared:  0.9964,Adjusted R-squared:  0.9954
F-statistic: 922.3 on 7 and 23 DF,  p-value: < 2.2e-16
```

Comments

- It is not possible to draw direct conclusions regarding the significance of the regression coefficients based on the t-tests. However, the results give some general direction, and the results indicate that all regression coefficients would be statistically significant (with the exception of the constant).

- The coefficient of the variable LQ1CPC with lag 1 is 0.91, which implies that the adjustment to changes in prices and total expenditures is rather fast.
- The signs of the coefficients of the price and total expenditures variables with lag 0 are as expected: the coefficient -0.83 of the price variable is negative and the coefficient +1.47 of the total expenditures variable is positive. These coefficients describe the instant effects of changes in prices and total expenditures.
- The signs of the coefficients of the price and total expenditures variables with lag 1 are also as expected.
- Long term elasticities are:

$$\begin{aligned} &\text{Price:} && \frac{\beta_2+\beta_3}{1-\beta_1} \approx -1.31, \\ &\text{Total expenditures:} && \frac{\beta_4+\beta_5}{1-\beta_1} \approx 1.75. \end{aligned}$$

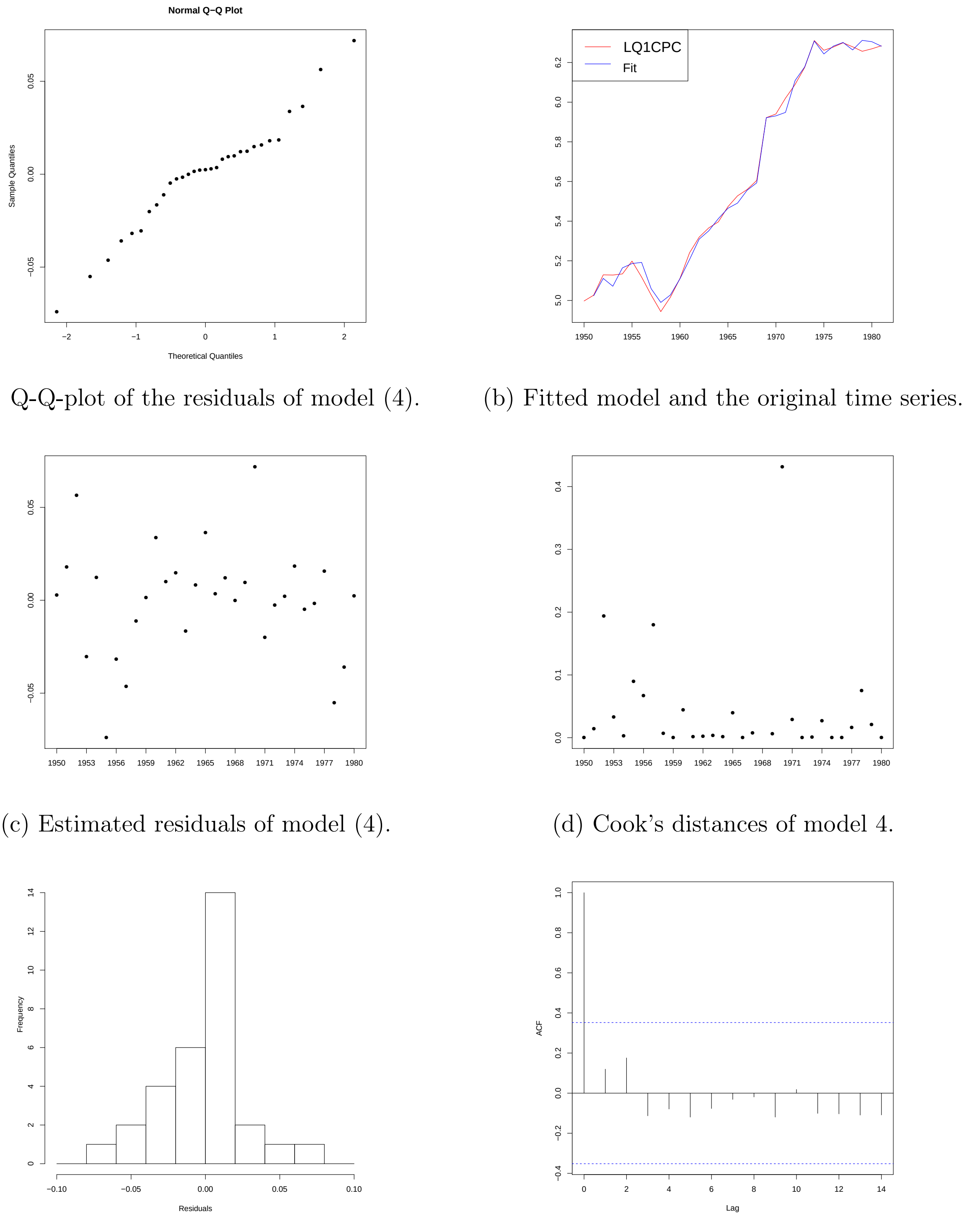
- Interpretations of the regression coefficients of price and total expenditures variables with lag 0:
  - If the price goes up by 1%, then the alcohol expenditures are instantly reduced by (without a lag) 0.83%.
  - If the total expenditures are increased by 1%, then the alcohol expenditures are increased by 1.47%.
- Interpretations of the long term elasticities of price and total expenditures variables:
  - If the price goes up by 1%, then the alcohol expenditures are reduced by 1.31% in the long term.
  - If the total expenditures are increased by 1%, then the alcohol expenditures are increased by 1.75% in the long term.
- The coefficient of the instant effect of the dummy variable LAW is 0.177 and the long term coefficient is small ($(\beta_6+\beta_7)/(1-\beta_1) \approx -0.16$). Hence, the change in the legislation has rather minor effect on the behaviour of the consumers in a long term, which seems plausible.
- Additionally, it is not possible to draw conclusions from the coefficient of the determination.

Comments related to the model (4):

- By Figures 4a and 4e, the residuals could be normally distributed.
- By Figures 4c and 4f, the residuals are not correlated.
- By the Breusch-Godfrey test, the residuals are not correlated. The null hypothesis is accepted with 5% level of significance for all lags:

```
model4_bg <- rep(NA,23)

# Breusch-Godfrey can be performed up to order:
# (sample size) - (number of estimated parameters) = 31 - 8 = 23

for (i in 1:23)
{
  model4_bg[i]= bgtest(model4, order=i)$p.value
}

which(model4_bg > 0.05)
# Null hypothesis of no autocorrelation accepted with all lags
```

- By VIF, there is strong multicollinearity in the model. This is unsurprising, as the model involves same variables with different lags.
- By the residual diagrams, there is no evidence of heteroscedasticity.
- The model takes into account the change in the legislation.
- By Figure 4b, the fitted model coincides better with the original time series than the fits of models (1) and (2).
- We consider this model to be sufficient in explaining the alcohol expenditures.



(a) Q-Q-plot of the residuals of model (4).

(b) Fitted model and the original time series.

(c) Estimated residuals of model (4).

(d) Cook's distances of model 4.

(e) Estimated residuals of model (4).

(f) ACF of the Estimated residuals of model 4.

Figure 4: Graphical regression diagnostics for model 4.

Homework

5.5 The file `t38.txt` contains three quarterly time series. The time series start from the first quarter of the year 1953 and the corresponding time series are,

CONS	= total consumption (billions)
INC	= income (billions)
INFLAT	= inflation (%)

The time series **CONS** and **INC** represent the observed total consumption and income in an imaginary country. The time series **INFLAT** represents inflation. The goal is to estimate a so-called consumption function that explains the time series **CONS** with the time series **INC** and **INFLAT**.

The conventional linear regression model for the response variable CONS is

$$\mathrm{CONS}_t = \beta_0 + \beta_1 \mathrm{INC}_t + \beta_2 \mathrm{INFLAT}_t + \varepsilon_t. \tag{5}$$

Assignments:

a) Estimate model (5) and study the goodness of fit.

b) Estimate the difference model corresponding to (5) and study the goodness of fit.

c) Estimate dynamic regression model:

$$\begin{aligned}\mathrm{CONS}_t =& \beta_0 + \beta_1 \mathrm{CONS}_{t-1} + \beta_2 \mathrm{INC}_t + \beta_3 \mathrm{INC}_{t-1} \\ &+ \beta_4 \mathrm{INFLAT}_t + \beta_5 \mathrm{INFLAT}_{t-1} + \varepsilon_t,\end{aligned} \tag{6}$$

and study the goodness of fit.

d) Which of the previous models are sufficient in explaining the behavior of the response variable CONS?