Chapter 9 Solutions

9.2.5 Exercises

Exercise 9.1

The program found next will help confirm the results explained in Section 9.2.4.

The results of the regression model that includes the index variable representing availability of mortgage money is more reliable than the model that includes only population.

starts.pop.model1 in the Program. The estimated (learned) model with just population as independent variable is

$\widehat{starts}_t = -0.0608 + 0.0714 pop_t$

starts.pop.model2 in the program. The model with the population and the mortgage availability index is

$\widehat{starts}_t = -0.010427 + 0.034656 pop_t + 0.760464 index_t$

In both models, the regression slope(s) statistically significant, $p < 0.0000$ and the $R^2 > 0.9$`adjusted`) but the residuals of the model with just population as regressor are autocorrelated, as indicated by the sample residual ACF. The fact that the autocorrelation of the residuals disappears when we add the index variable is indication that the autocorrelation was not because of autocorrelation per se but because of the omission of important variables such as index. When index is added, the residuals of the regression are white noise residuals.

It is interesting for economists to study which independent variable has larger effect on the dependent variable. That requires rewriting the model in standard deviation form. That means that we scale dependent and independent variables to have mean 0 and standard deviation 1. That way, the difference in the slopes indicates which regressor has more impact on the dependent variable. When in standard deviation form for the variables, the estimated model is as follows:

starts.pop.model3 in the program. The model with the population and the mortgage availability index is

$\widehat{starts.st}_t = -0.000 + 0.4668 pop.st_t + 0.5413 index.st_t$

The following R program supports the answers provided.

```
 housing=read.csv("housingchaterjee.csv", header=T)
  head(housing)
  attach(housing)

# metadata: population and starts is in millions.
```

```
## Plot of the three time series

plot.ts(cbind(population,starts,index), type="l",
main="Exercise 9.1.Three variables", ylim=c(0,3)
)


#########
# Fitting a regression for starts and population only.
#########
starts.pop.model1=lm(starts~population)
summary(starts.pop.model1)

sdresiduals1=scale(starts.pop.model1$residuals)

par(mfrow=c(2,1))
plot.ts(sdresiduals1,
main="Modelstarts.pop.model residuals",ylab="standardized residuals")
abline(h=0)

### Include an image of the acf of the residuals with the image above.

acf(sdresiduals1)  #acf shows autocorrelation
dev.off()


####################################
## Run now regression with index and population as regressors
########################################

starts.pop.model2=lm(starts~population+index)
summary(starts.pop.model2)

sdresiduals2=scale(starts.pop.model2$residuals)

par(mfrow=c(2,1))
plot.ts(sdresiduals2,
main="Modelstarts.pop.model2 residuals",ylab="standardized residuals")
abline(h=0)

### Include an image of the acf of the residuals with the image above.That
## is , include three images in one par(mfrow)
acf(sdresiduals2)  #acf shows autocorrelation

dev.off()

### To see which variable affects the dependent variable most, we
 ## can use standardized regression
```

```
pop.st=scale(population)
index.st=scale(index)
starts.st = scale(starts)

starts.pop.model3=lm(starts.st~pop.st+index.st)
summary(starts.pop.model3)
sdresiduals3=scale(starts.pop.model2$residuals)

par(mfrow=c(2,1))
plot.ts(sdresiduals3,
        main="Modelstarts.pop.model2 residuals",ylab="standardized residuals")
     abline(h=0)

### Include an image of the acf of the residuals with the image above.That
## is , include three images in one par(mfrow)
acf(sdresiduals3)  #acf shows autocorrelation

dev.off()
```

□

Exercise 9.2

Program ch9-Exercise9-2.R contains the code on which we base the conclusions.

(a) The simple regression model gives autocorrelated errors. Upon looking at the acf and pacf of the residuals ot the simple regression model, an MA(1) model for the residuals is identified. We then incorporate that information into GLS with that model assumption for the residuals and the residuals of the gls model are the residuals of white noise according to the acf. A Ljung-Box test could also be done.

(b) We used the arima() function to estimate the MA(1) model of the residuals. The model is

$\widehat{\epsilon_t} = -0.3570$
$\widehat{\epsilon_{t-1}} + 0.0002$ The model is estimated using the arima() function and the residuals of this arima model are white noise.

(c) Since the residuals of the simple regression model are autocorrelated we could suspect that there could be another variable explaining. The scatter plot showing the relation between eruptions and waiting indicates that there are two clusters of points. Those corresponding to large previous eruptions and those corresponding to smaller previous eruptions. Perhaps including a factor variable indicating the cluster might help. The data are bimodal, and hence the Normality assumption for residuals that is assumed when interpreting inference results does not hold. We thus reestimate the simple regression model with that information. However, it makes no difference in the residuals. They are still correlated. Perhaps other geological variables might explain. After all geysers are complex natural phenomena.

□

Exercise 9.3

Run the program for this Exercise and pay attention to the comments.

□

9.3.1 Exercises

Exercise 9.4

(a) Nothing to write.

(b) The time plot reveals that amplitudes of the seasonals are increasing over time, thus we take log to stabilize the variance across time. The log also helps give a histogram for logged data a little more symmetric than the skewed-right histogram of the raw data.

(c) There is a problem with the residuals of the model requested. The residuals have cyclical patterns (as indicated by the time plot) and increase over time (as indicated by the time plot). According to the acf they are nonstationary. The increasing variability seen in the time plot means that they are not variance stationary. The variance has not been stabilized enough with the log transformation of the time series.Explain what you are doing. The equation of the fitted model with the coefficients is given below.

$\hat{y}_t = 7.27 + 0.00796t - 0.0000t^2 - 0.00199D2 + 0.06598eD3 + 0.03288eD4 + 0.1462D5 + ... + 0.0235D12$

This model with just quadratic trend is not good enough. So we try a 4th degree polynomial for the trend, which gives an ACF of an AR borderline nonstationary and perhaps also MA...

(d) The correlogram of the residual of the model with dummies and fourth degree polynomial could be, as said earlier the correlogram of an AR model close to nonstationary, perhaps due to the strong seasonal. We tried different models for the residuals, but the only one that leaves us with the residuals of the residual model being those of white noise is an AR(12), a rather large model. This suggests that we should do a better job at fitting the seasonal in the original regression model. Dummy variables do not work so well with real data.

(e) We fitted an AR(12) to the residuals.

See Program, part (d)

(f) The residuals are white noise, according to the ACF.

(g) See program, part (g)

(h) A forecast using the glm capabilities, like all forecasts involving multiple regression, require the dataframe with the values of the independent variables for the future. We create that in the program.

We did not create a training and a test set with the data. Thus, we can not calculate the RMSE to see how good is a forecast. The plot of the forecast suggests that electricity production will continue increasing exponentially in the future.

Note: The reader could try different model specifications for the trend function or perhaps different pre-differencing transformations and see the effect that this has on the residuals and the model required for the residuals.

□

9.6 Problems

Problem 9.1

Program *seatbelt.R* and datasets *ksi.txt* and *ksi-2.txt* are needed to answer this question.

(a) The time plot of the seasonally adjusted time series with the trend estimate superimposed indicates that the trend changes slope after 1983 and overall the trend remains at an average level lower than before 1983.

(b) A regression model of accidents against time and seatbelt factor results in residuals that are not white noise, according to the acf. Upon identification of a model for the residuals using the ACF and the PACF, we tentatively fit an AR(3) to the residuals, and using gls include that in the gls regression specification.

The resulting gls estimated model has white noise residuals. Based on this model, the trend model for the date before the seatbelt law and the model after are:

Before:

$\widehat{accidents} = 1425.14 + 14.6761 * time$

After:

$\widehat{accidents} = 1425.14 - 625.35 + 14.6761 * time$

Thus, there is an average reduction in the trend of approximately 625 accidents after the law was passed. It appears that the law had a significant effect in the number of accidents by shifting the trend down.

There is statistically significant difference in the trends because the coefficient of the dummy variable for seatbelts has p-value almost 0.

However, we have some reservation with this model. Presumably other variables affect the number of accidents, so it is no surprise that in the simple regression model fit before the gls one, the R-square was less than 50%.

(c) With the differenced data, we do not need gls. The residuals of the regression model are white noise. The dummy variable representing the seatbelt policy has a statistically significant coefficient of -236 approximately. Thus the trend in the change variable gets reduced significantly reduced down by an average of that amount.

□

Problem 9.3

(a) The time plot of the data reveals the following:

Lumber production is fluctuating around a constant mean with no specific pattern. For that reason, the first thing that comes to mind is just to estimate the mean. The variable time would be used as a trend only if we thought there was a trend. We did the autocorrelation of the lumber variable and noticed that the data is just white noise. The ACF has no significant autocorrelation at all.

(b) The regression line is flat, an estimate of the mean of the data only.

$$\widehat{y} = 35652$$

$$\widehat{\beta_o} = 35652 = \widehat{\mu}$$

$$se(\widehat{\beta_0}) = 372; \quad t = 95.85; \quad p - value = 0.000000$$

Based on these results of the regression, the mean of lumber production is significantly different from 0 (p-value =0.0000), and is at 35652 million board feet.

(c) The Durbin Watson test, is not rejecting the null hypothesis of 0 autocorrelation in any of its versions, one sided or two sided. The p-value is larger than 0.05 in all the tests' versions. Thus there is no statistical significant evidence that the residuals are autocorrelated. The plot of the standardized residuals confirms that, as there are no runs of positives or negative residuals. and there is no pattern of zig zag positive to negative repeatedly. The ACF of the standardized residuals confirms that there is no autocorrelation. The model we fitted did not create any patterns that introduce autocorrelation then.

The runs test confirms the conclusion of the DW test and the residual plot. There is no statistically significant evidence to reject the null hypothesis of no autocorrelation.

Note: notice how in the R program we obtain the prediction intervals for the true values of the data in the forecast period. For example,

The prediction interval for 1977 is

$$35651.87 \pm 2.0452(2037\sqrt{(1+\frac{1}{30})}) = (31416.93, 39886.81)$$

We are 95% confident that in 1977, the volume of lumber will be in that interval.

We use the following R program to reach those conclusions.

```
 lumberdata=read.csv("lumber.csv",header=T)
lumber=lumberdata$lumber
lumber=ts(lumber,start=1947,frequency=1)
acf(lumber) # lumber is white noise.
####
# (a)
####
plot.ts(lumber,main="lumber annual  data 1947-1976")


#####
# (b)
#####

lumberols=lm(lumber~1)
abline(lumberols)


####
# (c)
####
## We first check for autocorrelation of the residuals informally with a
## time plot and formally with the ACF of the standardized residuals.

summary(lumberols)
sresiduals=scale(residuals)  #standardize the residuals

sresiduals.ts= ts(sresiduals,start=1947,freq=1)

par(mfrow=c(2,1))
```

```
plot.ts(sresiduals.ts,main="plot of standardized residuals, 1947-1976")
abline(h=0)
acf(sresiduals.ts)
dev.off()

## we can see that the residuals are white noise (from ACF and also lack
## of pattern in the time plot of the standardized residuals.

install.packages("lmtest")
library(lmtest)
dwtest(lumberols)
dwtest(lumberols,alternative="greater")
dwtest(lumberols,alternative="less")
dwtest(lumberols,alternative="two.sided")

#####################################################
#Extra: runs test -not asked in the problem

#### (d)runs test (d)
install.packages("tseries")
library(tseries)
test1=runs.test(factor(sign(sresiduals)),
alternative="less")
test1
test2=runs.test(factor(sign(sresiduals)),
alternative="greater")
test2

test3=runs.test(factor(sign(sresiduals)),
alternative="greater")
test3



################################################

#### Extra, not asked in the problem
# focus on the fit and the forecast. Notice how
## the out of sample intervals are prediction intervals.

#obtain list with fitted values,prediction interval for y,error of fit
pred.int.insample=predict(lumberols, lumber, interval=c("predict"),se.fit=T)
pred.int.insample
#obtain matrix with fitted values and confidence intervals for mu
pred.ci.insample=predict(lumberols,lumber,interval=c("confidence"))
pred.ci.insample

#extract things out
#extract prediction points
```

```
predicted=pred.int.insample$fit[,1]
#extract standard errors of predictions
se.predict=pred.int.insample$se.fit
se.predict
#extract prediction interval bounds for y_t
lower95pred= pred.int.insample$fit[,2]
upper95pred=pred.int.insample$fit[,3]
#extract prediction interval bounds for mu
lower95ci=pred.ci.insample[,2]
upper95ci=pred.ci.insample[,3]

#put together all in one data frame
all=data.frame(lumber, predicted,se.predict,lower95pred, upper95pred,
lower95ci, upper95ci, residuals)
all

all.ts=ts(all, start=1947, freq =1 )
all.ts

plot(all.ts[,1],ylim=c(30000, 40000))
lines(all.ts[,2],col="red")
lines(all.ts[,4],col="blue")
lines(all.ts[,5], col="blue")
lines(all.ts[,6], col="brown")
lines(all.ts[,7], col="brown")

## the outofsample newdata in this case consists of the observation number
newdata=data.frame(c(31,32,33,24))
forecast=predict(lumberols,newdata, interval="prediction")
forecast
summary(forecast)



forecast.ts=ts(forecast,start=1977, freq=1)
forecast.ts

y=ts(c(lumber,forecast[,1]),start=1947,freq=1)
y

plot.ts(y,ylim=c(30000, 40000) )

abline(v=1977)
lines(all.ts[,2],col="red")
lines(all.ts[,4],col="blue")
lines(all.ts[,5], col="blue")
lines(all.ts[,6], col="brown")
lines(all.ts[,7], col="brown")
lines(forecast.ts[,2],col="green")
```

```
lines(forecast.ts[,3],col="green")
   content...
```

□

Problem 9.6

Notice the missing paragraph:

If initially both the DW and the residual plot indicated autocorrelation we would be inclined to suggest that the pattern of dependence in the residuals is AR(1). When the pattern of dependence is other than first order, the line plot of residuals against time will still be informative. However, the DW statistics may not yield useful information.

□

Problem 9.7

(a) The histogram of the calculator sales looks close to normal, so the assumption of normality of the residuals of the model we use will makes sense. For the rest of the problem, use the program in *calculator.R*.

(b) The plot produced by the R program indicates that the number of calculators used has increased over time, with no apparent seasonality.

(c) A regression model with calculator as dependent variable as a function of a linear tred was fitted to the data.

$\widehat{calculator}_t = -189702.56 + 96.89time$

The residuals of this model are white noise, and the coefficients are statistically significant. The $R^2 = 0.76$.

(d) No matter what the alternative hypothesis is for the Durbin-Watson test, there is no statistically significant AR(1) autocorrelation in the residuals of the regression model. Notice that the D-W test only tests for AR(1) of the residuals.

□

9.7 Quiz

Question 9.1

$3.3602 - (1.09 + 0.46) = 1.8102$

$-3.1769 - (1.09 + 0.46 * 2 - 4.77) = -0.4169$

Question 9.3

(a) and (b)

□

Question 9.5

White noise

□

Question 9.7

Finding too many parameters that are not significant.