NAME:
Student Number:

STAT 485/685
Midterm Examination

Instructor: Richard Lockhart 19 October 2017

Instructions: This is a closed book exam but you are permitted two sheets of paper with notes – written (or printed) on both sides. You are permitted a calculator but not one with wifi capability; you should not really need a calculator. The exam is out of **30**. You should have 21 pages including two pages for extra work and one page for the grades. **DON'T PANIC.**

General Comments: Many of the answers which I marked and which needed to be in sentences had sentences that conveyed no meaning to me. I took marks off for saying things that were untrue, sloppy, meaningless, or incomprehensible. I tried not to be satisfied with imagining I could guess what you meant but had not written. So in marking questions 2, 4, 5, 6, and 7 you lost marks for saying things like that but then gained marks for things I thought were good ideas. I also have made a list of slightly more specific complaints which might apply to several questions:

- *Use of 'it' and other pronouns like 'this' and 'that'.* The sentence 'It is not a good model' cannot be the first sentence of an answer unless you are relying on the presence of the word 'model' near the end of the question. The word 'it' and other pronouns stand for something and it must be clear what thing they stand for; if not you should have lost marks.

- People are not going to take you seriously as a data analyst if all you ever talk about is p-values and 'it is not significant' and so on. You need to say something *about the data*. Interpretations of statistical analyses should refer to the context of the data; otherwise what would be the point in telling you about the context?

- Lots of people believe they can use the acf to detect non-stationarity. It is true that a series with a linear trend will produce a slowly decaying acf and a series with a periodic mean will produce a periodic component in the acf which does not die away. But there are stationary processes with slowly decaying acfs and even (look at the cosine process in the text) with periodic correlation structures.

- More challengingly: it is possible for a stationary series to have a slowly damped oscillating periodogram; the sunspot data in this exam might be an example.

- Suppose I have two variables, say X and Y, which happen to be independent and to have normal distributions. If I get a sample of pairs X_i, Y_i and regress the Ys on the X values I will get a small slope and a low R^2. But the model will not be 'bad' and you should not say the 'fit' is bad just because you accept the null hypothesis that the slope is 0. In the situation I am describing the model is *good* because it is right. It is just that X is not useful as a predictor of Y.

- People also seemed to believe that the presence of autocorrelation in our time series indicates a model failure; on the contrary we use that correlation to help with forecasting. We are trying to find / estimate / characterize that autocorrelation for use in forecasting and for use in computing realistic standard errors.

- People used the idea that for residuals which are symmetric about 0 the median should be close to 0 and the first and third quartiles should be about the same size but with opposite sign. But they made judgments about how different the median was from 0 without looking at the scale of the data. It is good to compare the differences between the median and 0 and between Q3 and −Q1 to the standard deviation of the residuals; this sd is produced by lm under the label ‘Residual standard error’.

- People often seem not to have read to discover whether the data were annual, monthly, weekly or what.

- If you have $R^2 = 0.5$ you are not ok to say the model “explains 50% of the data”. In my view that is just wrong. It explains, in a certain sense, 50% of the variability in the data.

1. Suppose ϵ_1, ϵ_2, and so on are independent noise with mean 0 and standard deviation 3. Let $Y_0 = 0$ and

$$Y_1 = \epsilon_1, \quad Y_2 = 2Y_1 + \epsilon_2, \quad Y_3 = 2Y_2 + \epsilon_3$$

and so on.

(a) Compute the mean and and standard deviation of Y_3. [3 marks]

There are many ways to do this problem. In addition to the method below you could use $Y_1 = \epsilon_1$, $Y_2 = 2\epsilon_1 + \epsilon_2$ and $Y_3 = 4\epsilon_1 + 2\epsilon_2 + \epsilon_3$ to compute means, variances and covariances.

$$\begin{aligned}
\mathrm{E}(Y_3) &= \mathrm{E}(2Y_2 + \epsilon_3) \\
&= 2\mathrm{E}(Y_2) + \mathrm{E}(\epsilon_3) \\
&= 2\mathrm{E}(2Y_1 + \epsilon_2) \\
&= 4\mathrm{E}(Y_1) \\
&= 4\mathrm{E}(\epsilon_1) \\
&= 0
\end{aligned}$$

Similarly

$$\begin{aligned}
\mathrm{Var}(Y_3) &= \mathrm{Var}(2Y_2 + \epsilon_3) \\
&= 4\mathrm{Var}(Y_2) + \sigma^2 \\
&= 4\mathrm{Var}(2Y_1 + \epsilon_2) + \sigma^2 \\
&= 16\mathrm{Var}(Y_1) + 4\sigma^2 + \sigma^2 \\
&= 16\mathrm{Var}(\epsilon_1) + 5\sigma^2 \\
&= 21\sigma^2 \\
&= 21 \cdot 3^2 = 189.
\end{aligned}$$

So the SD of Y_3 is $\sqrt{189}$; I don't need that converted to a decimal or simplified.

(b) Compute $\gamma_{2,3}$ and $\gamma_{3,2}$. [3 marks]

By definition

$$\begin{aligned}\gamma_{2,3} &= \mathrm{Cov}(Y_2, Y_3) \\ &= 2\mathrm{Cov}(Y_2, Y_2) + \mathrm{Cov}(Y_2, \epsilon_3) \\ &= 2\mathrm{Var}(2Y_1 + \epsilon_2) + 0 \\ &= 8\mathrm{Var}(Y_1) + 2\mathrm{Var}(\epsilon_2) \\ &= 8\mathrm{Var}(\epsilon_1) + \mathrm{Var}(\epsilon_2) \\ &= 10 \cdot 3^2 \\ &= 90\end{aligned}$$

You are also asked for $\gamma_{3,2} = \mathrm{Cov}(Y_3, Y_2)$. This is the same as $\mathrm{Cov}(Y_2, Y_3) = 90$.

(c) Is the time series Y_t stationary? Explain. [1 mark]

No. It is easy to check for instance that $\gamma_{1,2} = 8\mathrm{Var}(Y_1) = 72$ which is not the same as $\gamma_{2,3}$. For a stationary process $\gamma_{1,2} = \gamma_{2,3}$. There are lots of ways to arrive at the same conclusion.

2. If I kept track of the number of calories I eat each day I would get a time series.

(a) Would you expect the time series to be stationary over a short period of time like say a few months? Explain your answer. [2 marks]

Answers will be evaluated on the quality of their reason so it is possible that I will accept answers like "No, I expect calories eaten to be higher on weekends so the mean depends on the day of the week." I would also accept "Yes, over such a short time period there is not likely to be much change in the behaviour of this series."

Some people would say "stationary" and then give a reason for it not being stationary and vice versa. Those people got low marks.

(b) Would you expect the time series to be stationary over a long period of time like say 10 years? Explain your answer. [2 marks]

As in part a answers will be evaluated on the quality of their reason so it is possible that I will accept either "Yes" or "No" with a good reason. I think the best answer is 'No' since over a long period there is a good chance my behaviour will change.

If you said 'non-stationary' to part a) then you *need* to say the same for b). A series which is not stationary over the short term is not stationary. A number of people talked about the need for lots of data; this has nothing to do with whether or not the series is actually stationary though lots of data might help you judge whether or not the series really is stationary. Also see my comments about acf and stationarity in Q4 below.

3. Suppose that $\ldots, \epsilon_1, \epsilon_0, \epsilon_1, \ldots$ is white noise; you may assume as I said in my notes that the various ϵ_t are independent. Use σ^2 for the variance of ϵ_t. Define $Y_t = \epsilon_t \epsilon_{t-1}$.

   (a) Compute $\mathrm{E}(Y_t)$. [1 mark]

   $$\mathrm{E}(Y_t) = \mathrm{E}(\epsilon_t \epsilon_{t-1}) = \mathrm{E}(\epsilon_t)\mathrm{E}(\epsilon_{t-1}) = 0 \cdot 0 = 0.$$

   (b) Compute $\mathrm{Var}(Y_t)$. [1 mark]

   Since Y_t has mean 0 we find

   $$\begin{aligned} \mathrm{Var}(Y_t) &= \mathrm{E}(Y_t^2) \\ &= \mathrm{E}(\epsilon_t^2 \epsilon_{t-1}^2) \\ &= \mathrm{E}(\epsilon_t^2)\mathrm{E}(\epsilon_{t-1}^2) \\ &= \sigma^4 \end{aligned}$$

   (c) Compute $\mathrm{E}(Y_1 Y_2)$. [1 mark]

   We have

   $$\begin{aligned} \mathrm{E}(Y_1 Y_2) &= \mathrm{E}(\epsilon_0 \epsilon_1^2 \epsilon_2) \\ &= \mathrm{E}(\epsilon_0)\mathrm{E}(\epsilon_1^2)\mathrm{E}(\epsilon_2) \\ &= 0 \end{aligned}$$

   (d) Compute $\mathrm{Cov}(Y_1, Y_2)$. [1 mark]

   Since the Y_i have mean 0 the answer is the same as in (c).

   (e) Compute $\mathrm{Cov}(Y_t, Y_s)$ for general t and s. [1 mark]

   If s and t are different then just is in (c) you get 0.

   (f) What other process has the same mean and autocorrelation? [1 mark]

   White noise.

   We also accepted a few answers which were examples of white noise rather than the phrase “white noise” itself.

4. In the next few questions I ask you to comment on graphs and `R` output. The first dataset is from the `fpp` package in `R`. It is part of a package to accompany *Forecasting: principles and practice* by Rob J. Hyndman and George Athanasopoulos.

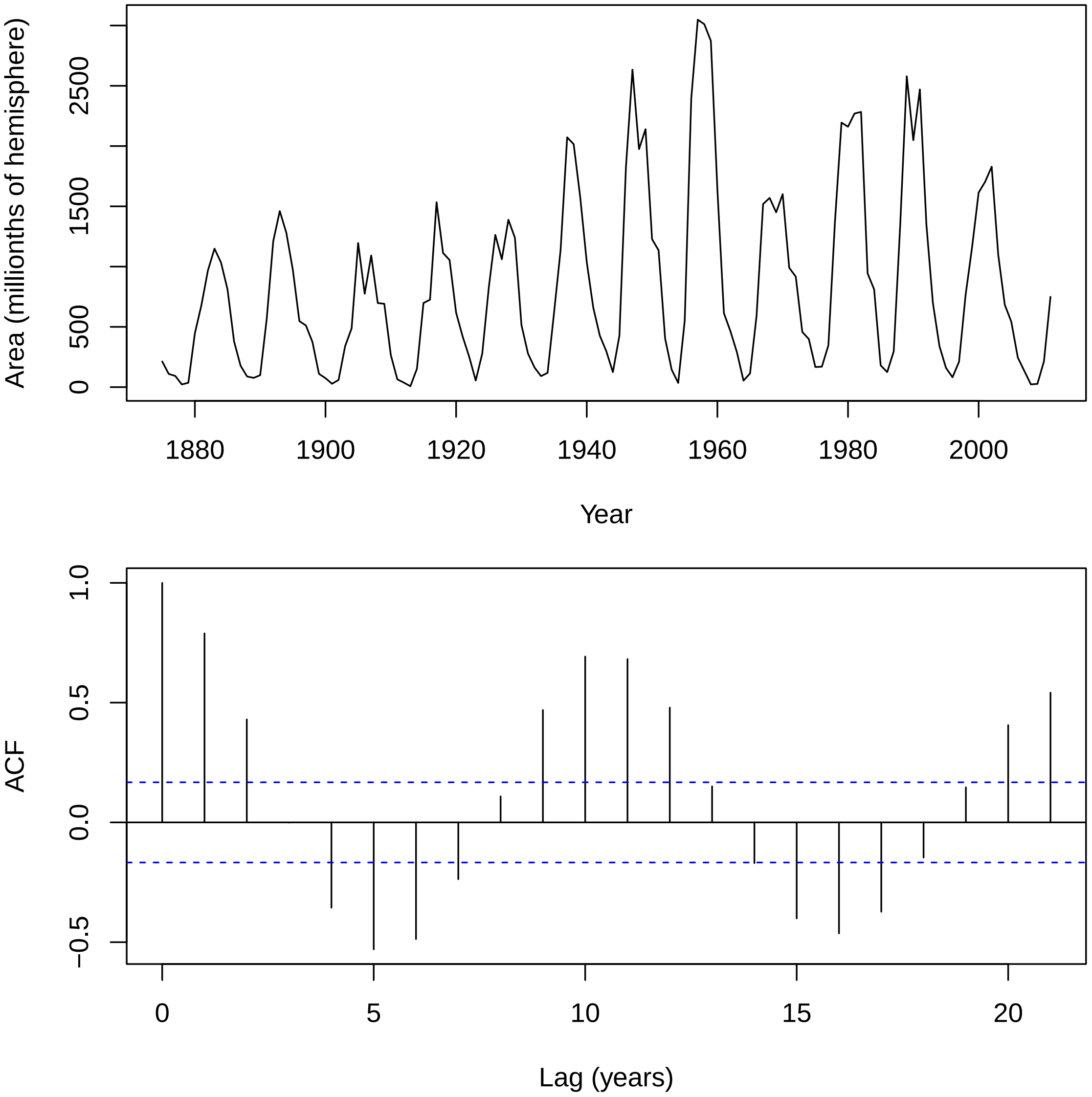
   > Annual averages of the daily sunspot areas (in units of millionths of a hemisphere) for the full sun. Sunspots are magnetic regions that appear as dark spots on the surface of the sun. The Royal Greenwich Observatory compiled daily sunspot observations from May 1874 to 1976. Later data are from the US Air Force and the US National Oceanic and Atmospheric Administration. The data have been callibrated to be consistent across the whole history of observations. More information is available at http://solarscience.msfc.nasa.gov/greenwch.shtml.

I ran the code

```
data("sunspotarea")
pdf("sunspots.pdf")
par(mfrow=c(2,1),mar=c(4,4,1,1)+0.1)
plot(sunspotarea,main=’’,xlab="Year",ylab="Area (millionths of hemisphere)")
acf(sunspotarea,main=’’, xlab=’Lag (years)’)
dev.off()
```

The graph produced is on the next page.

Does the series look stationary? Explain (on the next page). [2 marks]



Answer for sunspot data:

Actually I don't see much change in the behaviour across the time so my prefered answer is 'Yes'. But I admit that the peaks are higher in the middle of the time period and so I will take 'No' if you make out some case involving perhaps higher variability in the middle of the time period or lower at the beginning say.

In the end I took a lot of 'no' answers where people argued that they could see a change in the mean or in the variability (and said what that change was). I also took answers which argued that there was some sort of harmonic mean. Some people talked about ”seasonal” trend; this is annual data and I am not sure what those people meant by “seasonal”. In class I will take up the fact that it is hard to tell apart a non-stationary perfect cosine wave and some AR processes.

A lot of people believe that you can use the acf to detect non-stationarity. The acf is plotted based on an assumption of stationarity so this is not generally true. I did say in class that for series with a trend you get an acf which decays to 0 very slowly. It is also true that a perfect cosine wave component to a time series shows up in the acf as a perfect cosine wave which does *not* decay to 0.

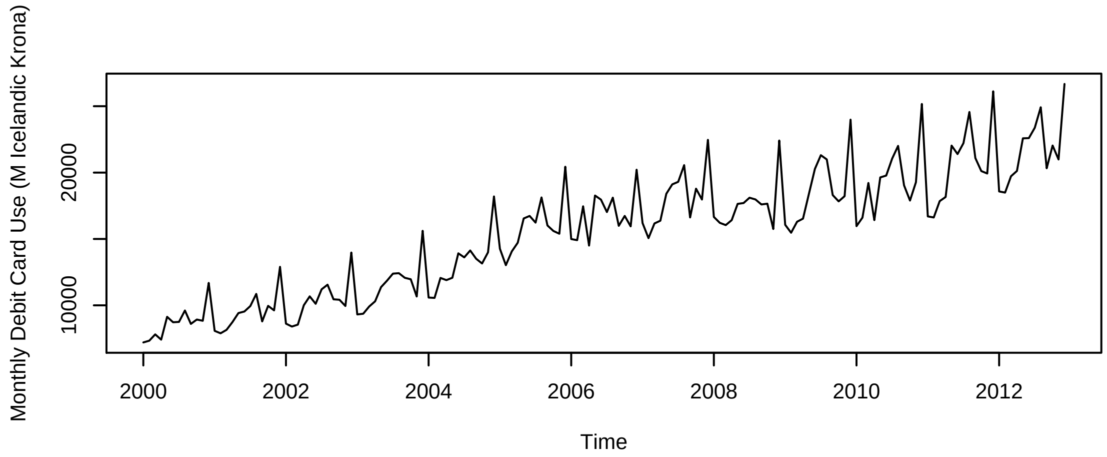
5. This data set is from the `fpp` package in `R`. It is part of a package to accompany *Forecasting: principles and practice* by Rob J. Hyndman and George Athanasopoulos. It records for each month, starting in January 2000 and ending December 2012, usage of debit cards in Iceland, in Millions of Icelandic Krona.

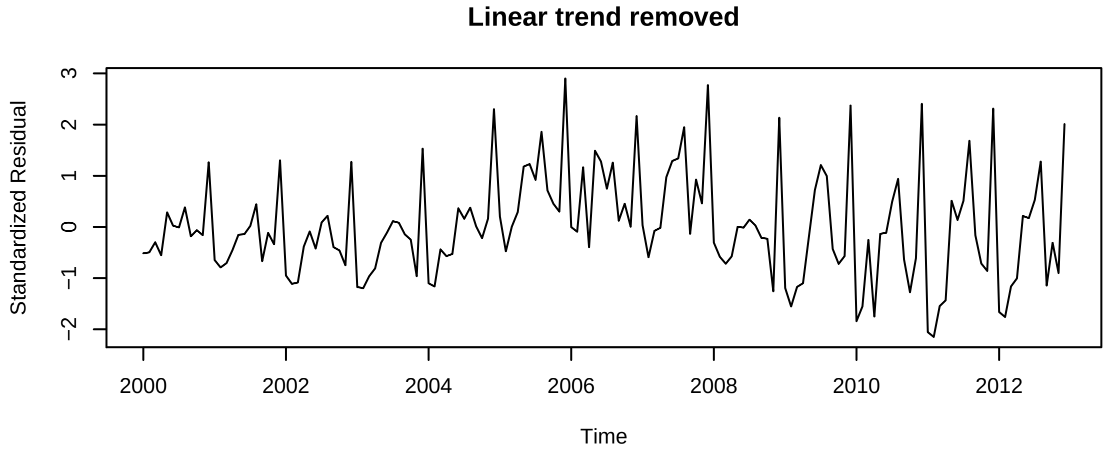
```
data(debitcards)
plot(debitcards,ylab="Monthly Debit Card Use (M Icelandic Krona)")
time = time(debitcards)
s.=season(debitcards)
fit = lm(debitcards~time)
summary(fit)
debit.res = rstudent(fit)
attributes(debit.res)=attributes(debitcards)
plot(debit.res,ylab="Standardized Residual",main="Linear trend removed")
fit2 = lm(debitcards~time+s.)
summary(fit2)
debit.res2 = rstudent(fit2)
attributes(debit.res2)=attributes(debitcards)
plot(debit.res2,ylab="Standardized Residual")
plot(debitcards,ylab="Monthly Debit Card Use (M Icelandic Krona)")
plot(debit.res,ylab="Standardized Residual",main="Linear trend removed")
plot(debit.res2,ylab="Standardized Residual",main="Linear plus seasonal")
```

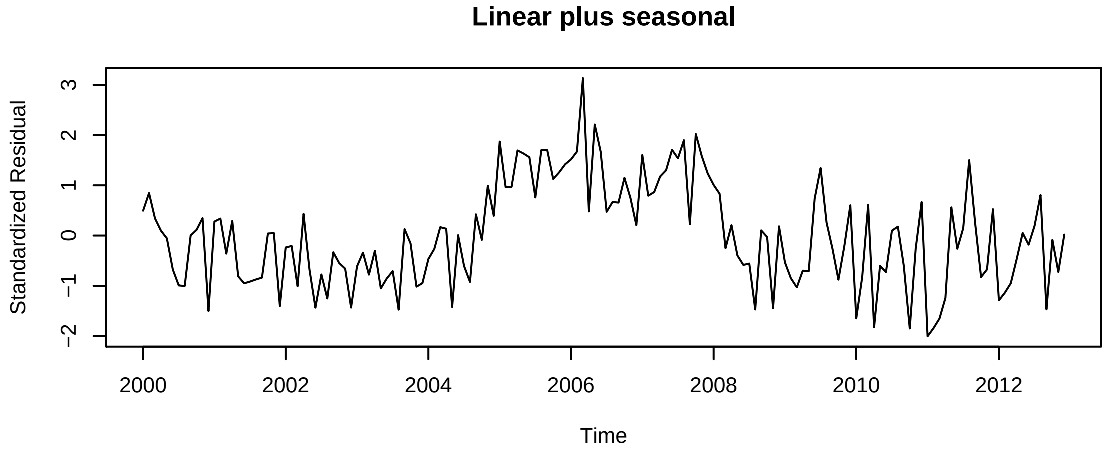
Here comes: output, then a page of graphs and the question after that!

```
> summary(fit)

Call:
```

```
lm(formula = debitcards ~ time)

Residuals:
    Min      1Q  Median      3Q     Max
-4130.2 -1202.3  -223.3   892.8  5537.8

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.256e+06  8.397e+04  -26.86   <2e-16 ***
time         1.132e+03  4.185e+01   27.05   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1961 on 154 degrees of freedom
Multiple R-squared: 0.8261,Adjusted R-squared: 0.825
F-statistic: 731.6 on 1 and 154 DF,  p-value: < 2.2e-16


> summary(fit2)

Call:
lm(formula = debitcards ~ time + s.)

Residuals:
    Min      1Q  Median      3Q     Max
-2178.7  -854.5   -14.6   736.2  3358.8

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2214591.9    49345.6 -44.879  < 2e-16 ***
time            1110.6       24.6  45.150  < 2e-16 ***
s.February      -344.7      450.9  -0.764  0.44584
s.March          590.5      450.9   1.310  0.19244
s.April          367.0      450.9   0.814  0.41704
s.May           2166.6      451.0   4.804 3.88e-06 ***
s.June          2343.2      451.0   5.195 6.90e-07 ***
s.July          2621.8      451.1   5.813 3.85e-08 ***
s.August        3399.4      451.1   7.535 5.08e-12 ***
s.September     1197.5      451.2   2.654  0.00885 **
s.October       1310.5      451.3   2.904  0.00427 **
s.November       868.6      451.4   1.924  0.05628 .
s.December      5646.6      451.5  12.507  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1150 on 143 degrees of freedom
Multiple R-squared: 0.9445,Adjusted R-squared: 0.9399
F-statistic: 202.9 on 12 and 143 DF,  p-value: < 2.2e-16
```







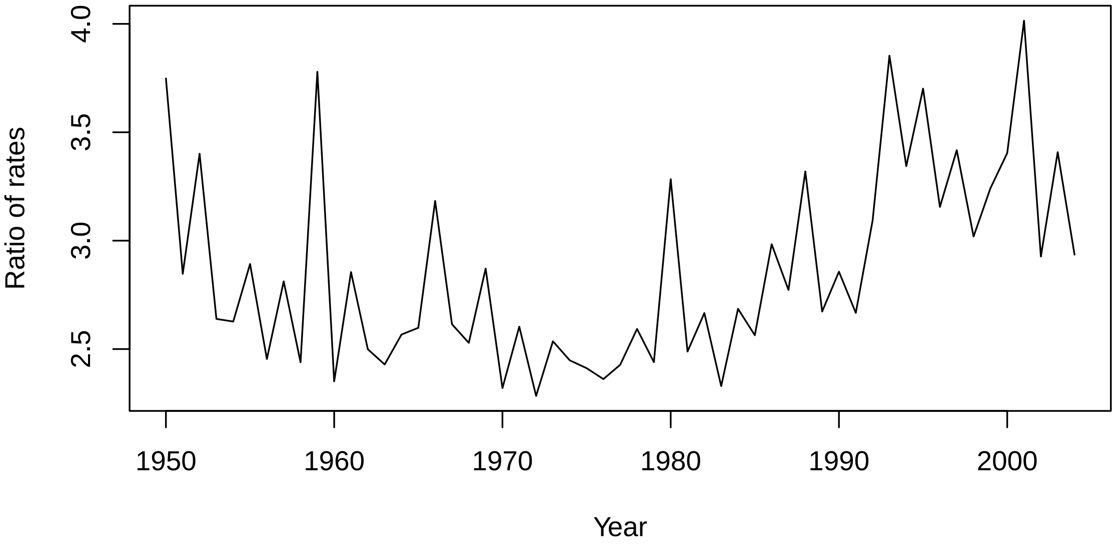
Question: Interpret the output and the graphs and comment on what else you would like to plot to examine the quality of the fits. [4 marks]

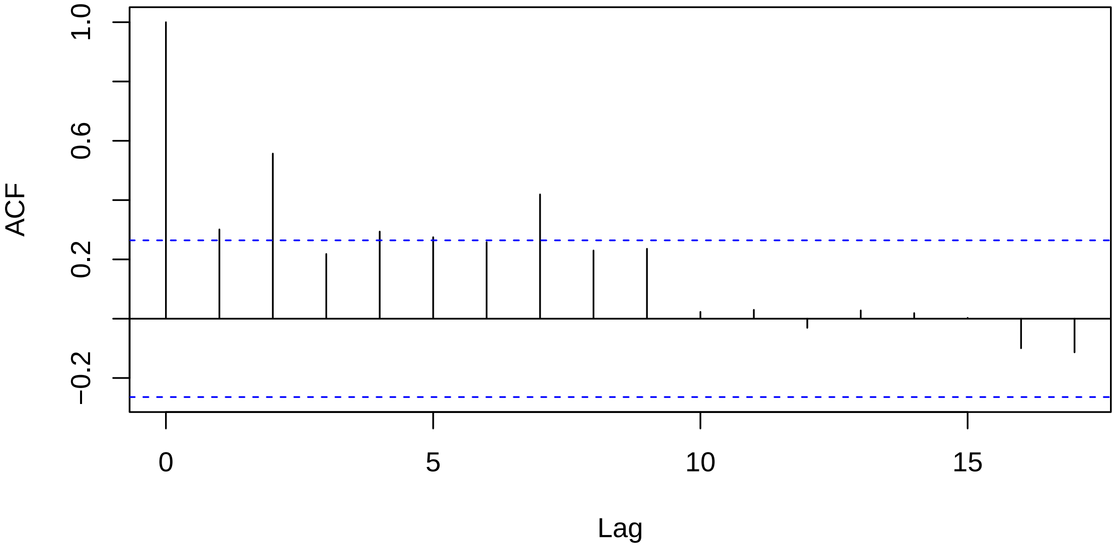
Here are some points you could make:

- There is a clear increasing trend: see the graph and the lm output. In the latter the rate of increase is about 1,000 to 1,100 Million Krona per year (or more precisely from the output.
- There is a very clear annual cycle with December being particularly different.
- The residuals from the linear plus seasonal fit show a surprising high spot in the range 2007 to 2008. (Not part of the course but that ends with the economic crash of 2008; Iceland was heavily affected because their economy had become enormously dependent on their financial system; high residuals before the crash may be produced by the boom before the bust.)
- I would like to see a month plot of the residuals. I think December might be a problem as I suggested it was for a time series we looked at in class. I might assess the normality of the residuals too. (But when I actually did this after marking the exam I saw not too much.)

- Some people made much of the fact that coefficients of some months are not statistically significant. This means only that the amount of money spent via debit cards is not much different in those months than in January (after adjusting for a linear trend in time). I would not, myself call these seasonal coefficients *slopes*. They are slopes in a certain strict formal sense but it is better to think of them as the difference between two straight lines: one line for January and the other a line for the month in question.
- I took marks off for saying things that are wrong even if you said lots that was right.
- The linear plus seasonal fit is far better than the linear only fit even though the residual plot for the linear only fit seems at first glance a bit better. The trouble is that there is clearly a strong seasonal pattern with very high summer spending and very very high December spending.
- Some people attributed the slope to population growth or inflation. The change from left to right is a growth of well over 100 % in not that many years to population growth is totally unlikely as an explanation. But it appears that Icelandic inflation has was quite high over this time period and that would account for most or even all of the trend, I think. The inflation rate grew rather sharply in 2008; the economic crisis that year had a very big impact on Iceland because its banks largely became bankrupt.
- If you suggested adding a quadratic trend term I was happy with that but I really want clear precise sentences. Lots of people lost a lot of marks for vague or even meaningless remarks.

6. This data set is from Gapminder. I started with a time series of annual homicides of female victims per 100,000 females in the United States and the same series for Canada. Then I computed the ratio US divided by Canada. The plots below apply to this ratio. Data start in 1950 and end with 2004.





Question for homicide data: Discuss the plots and describe what you might do in `R` to investigate this series: what would you plot, what tests and other procedures might you apply? [3 marks]

I think this looks reasonably stationary. I might consider a quadratic trend, too. If I removed such a trend I would plot the residuals, the residual acf, and consider the normality of the standardized residuals. Overall, however, I don't see a lot to worry about. Again I will look for good reasons in whatever answer you give.

So in marking most people lost marks for saying extra things I don't agree with or for just giving a list of plot names which amounted to all the ones we have studied. If you talk about doing something with residuals you have to tell me clearly what residuals – how would you run `lm`?

I believe quite strongly that it is useful to observe that the numbers range from 2.5 to 4. This is the most interesting feature of the data and a statistician should remember that there has to be some point in the analysis. I do not know yet why there is any autocorrelation and particularly why there is some at lag 2 and apparently lag 7 (for lag 2 I wonder about differences in reporting practices). Naturally one should study the joint behaviour of the two series, Canada and the US, rather than just the ratio summary; any statistician ought to say so.

The total number of Canadian female homicide victims has ranged from about 150 to a high of around around 270 – back in 1991. They are now down at the low end. Under a Poisson model for homicide occurence the standard deviation of the number of deaths in a year where the mean is 200 is about 14. So year over year changes of up to 30 of more were credible ($2\sqrt{200+200} = 40$ is the SD of the change year to year if the means were both 200 and the actual variability is a bit higher because of the small number of multiple homicides like Marc Lepine in 1989). The annual Canadian rate has variability on the order of $sqrt200/200 = 7\%$; you apply that 7% to the very tiny actual rate – about 0.5 women per 100,000.

7. This data set is in `astsa` in `R`. It is drawn from the data `lap` which gives weekly data described as “LA Pollution-Mortality Study (1970-1979, weekly data).” The original data set contained weekly mortality (number of deaths that week in LA from any cause) and respiratory mortality (number of those deaths due to breathing or lung issues). The data set plotted below is the ratio: the fraction of weekly deaths which have a respiratory cause.

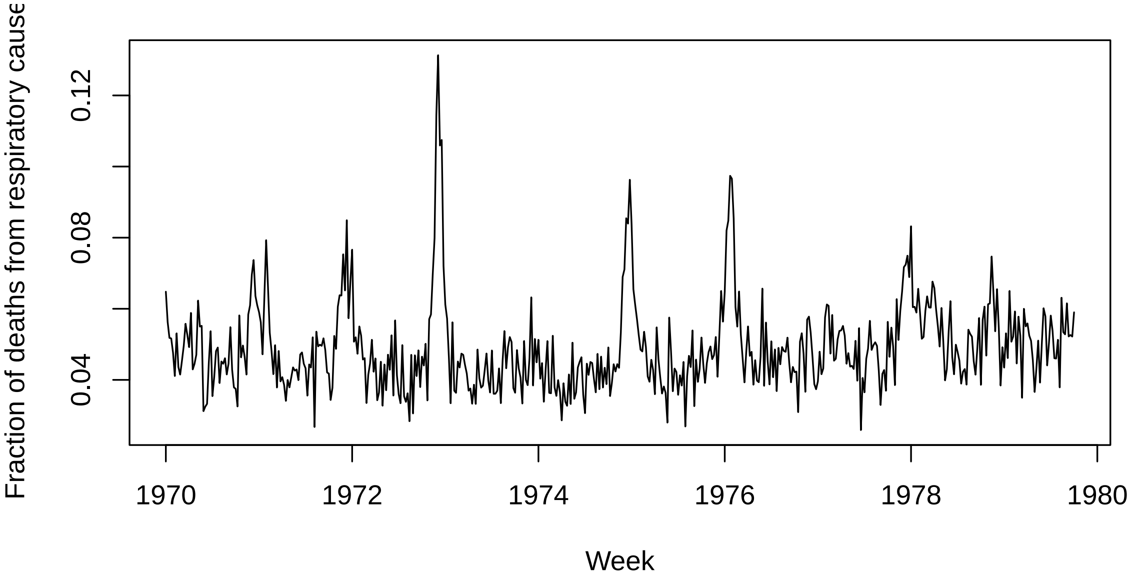
```
library(astsa)
data(lap)
tmort=lap[,1]
rmort=lap[,2]
plot(rmort/tmort,xlab="Week",ylab="Fraction of deaths from respiratory causes")
acf(rmort/tmort,xlab="Week",main='')
ratiomort=rmort/tmort
attributes(ratiomort)=attributes(tmort)
mortality.fit.linear = lm(ratiomort~time(ratiomort))
summary(mortality.fit.linear)
mortality.res=rstudent(mortality.fit.linear)
attributes(mortality.res)=attributes(tmort)
postscript("MortalityDiagnostics.ps",horizontal=F,height=8,width=6)
par(mfrow=c(3,1))
plot(mortality.res)
hist(mortality.res)
qqnorm(mortality.res)
dev.off()
```

Output for mortality ratio

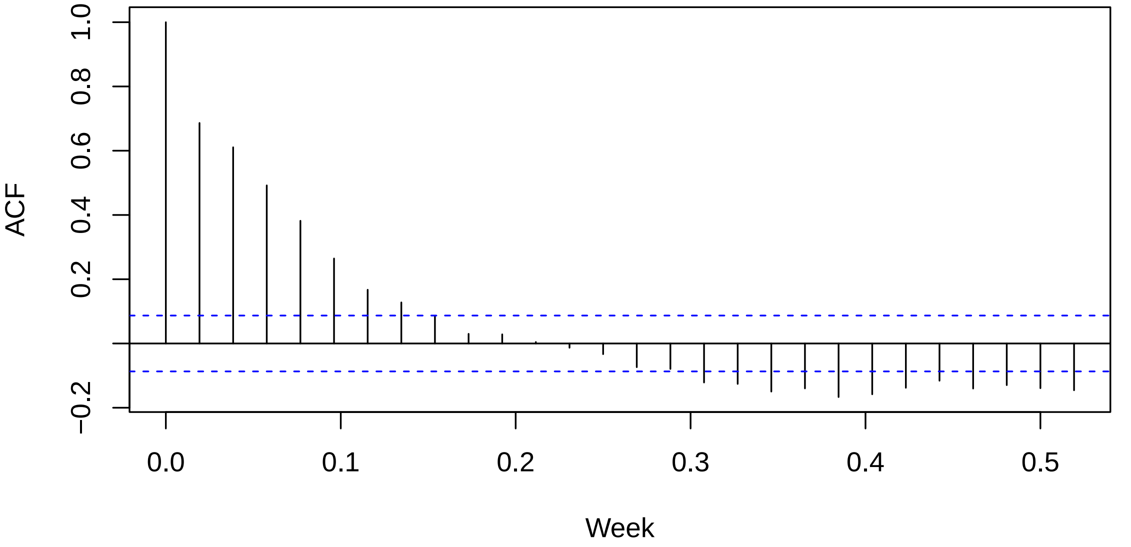
```
> summary(mortality.fit.linear)
```

```
Call:
lm(formula = ratiomort ~ time(ratiomort))

Residuals:
      Min        1Q    Median        3Q       Max
-0.024401 -0.008228 -0.001846  0.004841  0.083254

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.9199137  0.3919707  -2.347   0.0193 *
time(ratiomort)  0.0004906  0.0001985   2.472   0.0138 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01262 on 506 degrees of freedom
Multiple R-squared:  0.01193,Adjusted R-squared:  0.009979
F-statistic: 6.111 on 1 and 506 DF,  p-value: 0.01377
```

Graphs for mortality ratio: first the series and its acf.

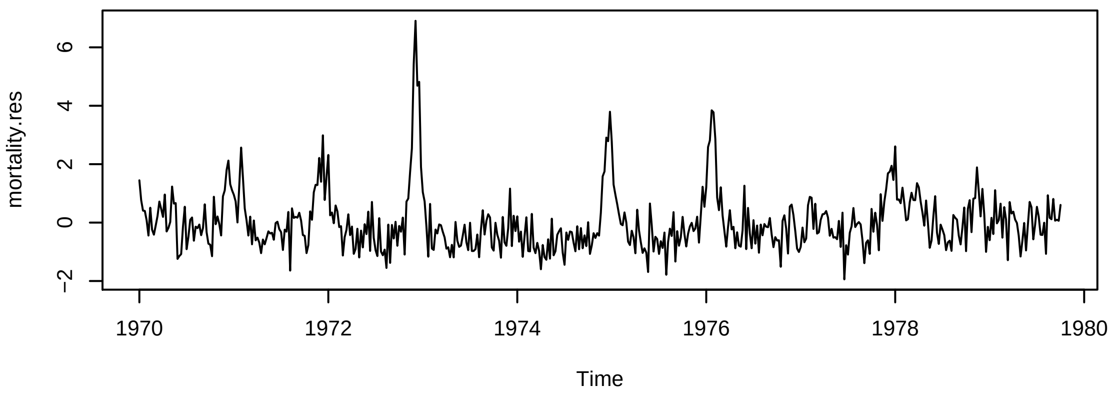


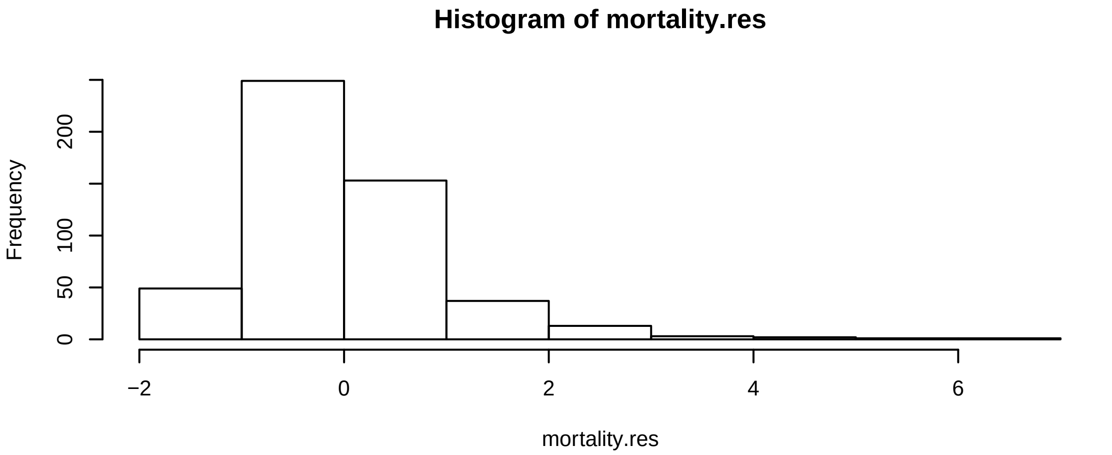
Fraction of deaths from respiratory cause
0.04
0.08
0.12
1970
1972
1974
1976
1978
1980
Week

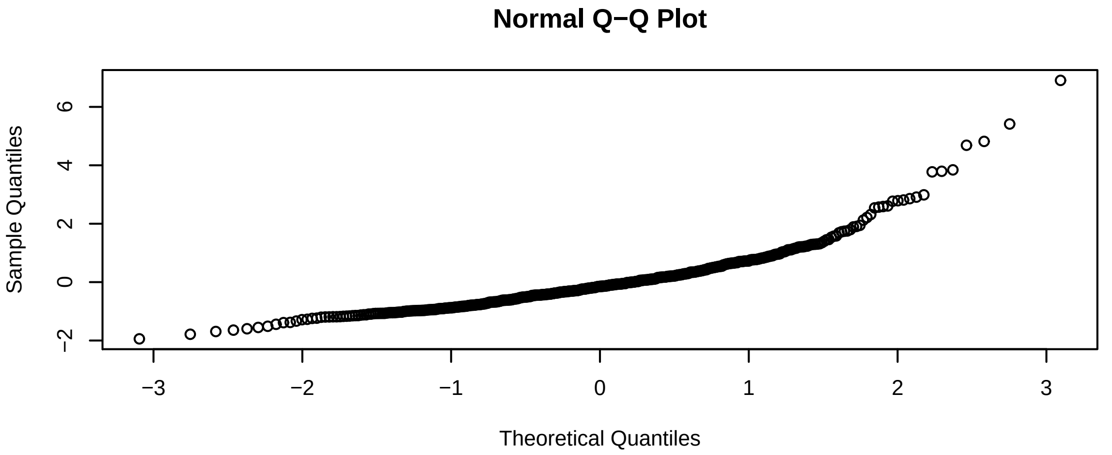


ACF
−0.2
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
Week

Graphs for mortality ratio: last 3 graphs produced by `R` code above.







Question for mortality ratio data: Interpret the output and the graphs and comment on problems revealed by those. [4 marks]

A first point: the acf graph has an error. I could not see that anyone noticed. The data are weekly but the time period is not ”Week” but ”Year”. Thus the negative correlations at 0.5 are correlations between Y values some 6 months apart. This mistake is visible in label ”Week” underneath the x-axis in the plot of the original series.

Some thoughts I will look for in your answers:

- Notice that the multiple R^2 (this is the square of the usual correlation coefficient when you just regress on time) values are tiny.
- The slope for time is very small (though it is just statistically significant – remember that the statistical significance can be substantially wrong if there is autocorrelation).
- The residuals are clearly not normal: the qq plot has a curve, the histogram is skewed to the right and there is a clear outlier (or short patch of outliers) around 1973.
- There is a standardized residual of 6! None of the negative residuals is even as small as -2.
- It looks to me like we need to check for an annual cycle in this weekly data.

- The spike in 1973 needs to be understood. Some people talked about the effect of pollution on respiratory deaths and that question is certainly the reason for collecting this data. But the data I gave you have *no* pollution data at all; so naturally you can't tell what caused this spike.
- Lots of people said low R^2 meant a bad fit. It doesn't; it means that the linear trend will do little to help with forecasting.
- Lots of people looked only an R^2 and ignored the direction and size of the trend. I think it is worth saying that the slope is fairly small in the sense that the change over the 10 year period is only 0.005 compared to a mean for the series of 0.05 (10 times larger) and a standard deviation for the series of 0.01. Not negligible but not huge – half an SD.
- It turns out that the mortality ratio is much higher in the winter than in the summer.
- Lots of people showed me they don't know the difference between left and right skewed.
- The $q - q$ plot shows a heavy right hand tail not both tails. A heavy left tail would go below the line on the left of the picture.