Homework 2
UMN STAT 5511
Charles R. Doss
Solution

The usual formatting rules:

- Your homework (HW) should be formatted to be easily readable by the grader.
- You may use knitr or Sweave in general to produce the code portions of the HW. However, the output from knitr/Sweave that you include should *be only what is necessary to answer the question*, rather than just any automatic output that R produces. (You may thus need to avoid using default R functions if they output too much unnecessary material, and/or should make use of `invisible()` or `capture.output()`.)
  - For example: for output from regression, the main things we would want to see are the estimates for each coefficient (with appropriate labels of course) together with the computed OLS/linear regression standard errors and p-values. If other output is not needed to answer the question, it should be suppressed!
- Code snippets that directly answer the questions can be included in your main homework document; ideally these should be preceded by comments or text at least explaining what question they are answering. Extra code can be placed in an appendix.
- All plots produced in R should have appropriate labels on the axes as well as titles. Any plot should have explanation of what is being plotted given clearly in the accompanying text.
- Plots and figures should be *appropriately sized,* meaning they should not be too large, so that the page length is not too long. (The arguments fig.height and fig.width to knitr chunks can achieve this.)
- **Directions for "by-hand" problems:** In general, credit is given for (correct) shown work, not for final answers; so show **all** work for each problem and explain your answer fully.

Questions:

1. (Prediction using the cross-correlation function) Assume that $Y_t = aX_{t-\ell} + W_t$ for some number a. The series X_t leads Y_t if $\ell > 0$ and is said to lag Y_t if $\ell < 0$. Assume that $E(X_t) = E(Y_t) = 0$, that $\{X_t\}$ is stationary and that $W_t \sim \mathrm{WN}(0, \sigma^2)$ is uncorrelated with the whole series X_t. Let γ_x denote the autocovariance function of $\{X_t\}$.

   (a) Is Y_t stationary?

   (b) Compute the cross covariance function between Y_t and X_s, for any s and t. (Your answer will depend on γ_x, the autocovariance function of X_t.)

   (c) Compute the cross correlation function between Y_t and X_s, for any s and t. (Your answer will depend on γ_x, the autocovariance function of X_t.)

   Solution:

   (a) We have $EY_t = aEX_{t-\ell}$. We have $\mathrm{Var}(Y_t) = a^2 \mathrm{Var}(X_{t-\ell}) + \sigma_w^2 = a^2\gamma_x(0) + \sigma_w^2$ by independence of $X_{t-\ell}$ and W_t, and stationarity of X_t. For $h > 0$, $\mathrm{Cov}(Y_t, Y_{t-h}) = \mathrm{Cov}(aX_{t-\ell}, aX_{t-\ell-h}) = a^2\gamma_x(h)$.

   Thus, since X_t is stationary, Y_t has constant mean, constant variance and its autocovariance is a function of the time difference only. We conclude Y_t is indeed (weakly) stationary.

   (b) $\mathrm{Cov}(Y_t, X_s) = \mathrm{Cov}(aX_{t-\ell} + W_t, X_s) = a\,\mathrm{Cov}(X_{t-\ell}, X_s) + \mathrm{Cov}(W_t, X_s) = a\gamma_x(|t - \ell - s|)$, where we have used that W_t is independent of X_s for all (t, s) and that X_t is stationary by assumption.

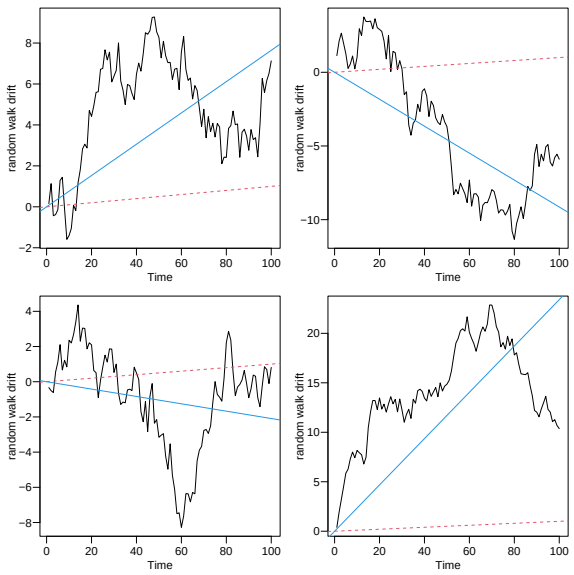
   (c) Using the calculation of $\mathrm{Var}(Y_t)$ from above, the cross-correlation is

   $$\mathrm{Cov}(Y_t, X_s)/\sqrt{Var(Y_t)Var(X_s)} = a\gamma_x(|t - \ell - s|)/\sqrt{(\sigma_w^2 + a^2\gamma_x(0))\gamma_x(0)}.$$

2. Question 2.3, Shumway and Stoffer, 4th edition (Note: The question is somewhat different than in previous editions).

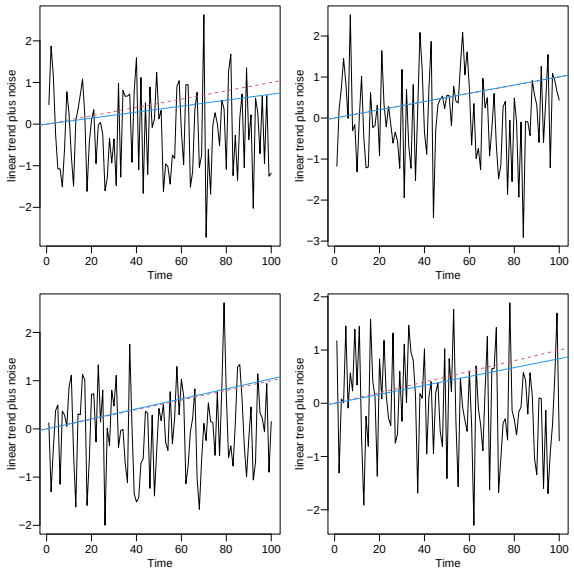
   Solution:

(a)
```
par(mfrow=c(2,2),mar=c(2.5,2.5,0,0)+0.5,mgp=c(1.6,0.6,0))
for(i in c(1:4)){
  x<-ts(cumsum(rnorm(100,0.01,1)))
  model<-lm(x~time(x)+0,na.action = NULL)
  plot(x,ylab='random walk drift',las=1)
  abline(a=0,b=0.01,col=2,lty=2)
  abline(model,col=4)
}
```



The dashed line is the true mean function and the solid line is the fitted one.

(b)
```
par(mfrow=c(2,2),mar=c(2.5,2.5,0,0)+0.5,mgp=c(1.6,0.6,0))
for(i in c(1:4)){
  x<-ts(rnorm(100))
  y<-0.01*time(x)+x
  model<-lm(y~time(x)+0,na.action = NULL)
  plot(x,ylab='linear trend plus noise',las=1)
  abline(a=0,b=0.01,col=2,lty=2)
  abline(model,col=4)
}
```



The dashed line is the true mean function and the solid one is the fitted one.

(c) This question explores two very different models or data generating mechanisms. The estimated line based on the linear trend model does quite well ("is consistent", we say) whereas based on the random walk it does poorly.

We saw in class the theoretical property that random walks are nonstationary because the variance of a random walk accumulates over time. This simulation shows what it means that the "trend" that we (think we) see in a random walk is actually variance rather than a true trend. (The four different instantiations of the random walk had four different "trends", whereas the four different simulations in (b) had very similar trends.)

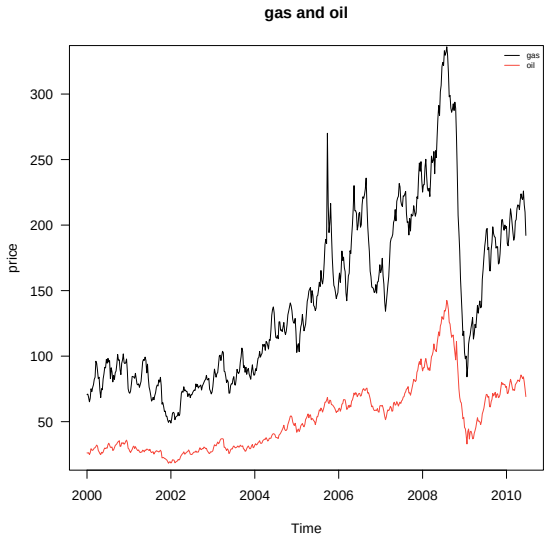
One way to think about this is to think about prediction: predicting future values based on the estimates in part (b) will tend to do well, but in part (a) the estimated line will be useless for prediction.

Another thing to notice is that the series as a whole is much more variable in (a) than in (b). For instance, the last observation (X_{100}) goes from around -10 to +4 in (a) whereas in (b) it is always between -2 and 2.

3. Question 2.10, Shumway and Stoffer. For (f)(iii), you can do both analysis of the residuals as you would in a non-time series context (e.g., a QQ-plot) and analysis of the correlation of the residuals (using the ACF).

Solution:

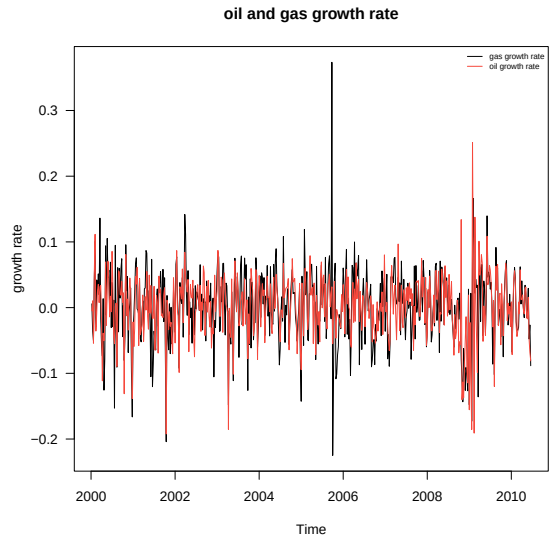
(a)
```
library(lattice)
library(astsa)
par(mfrow=c(1,1))
plot.ts(gas,ylab="price",main="gas and oil",ylim=c(25,325),col='1',las=1)

lines(oil, col='2')
legend("topright", legend = c("gas","oil"),lty = 1:1, col = 1:2, bty = 'n',
       cex=0.6)
```



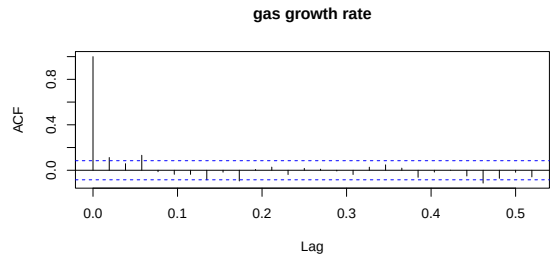
The series look like random walks, perhaps with drift. So, it is not stationary. There is one visible very large jump present. Excluding that one, there are still several other quite large jumps present. This suggests there are periods of heavy volatility or heavy tailed behavior. Ignoring those, the random walk (with drift) model seems reasonable.

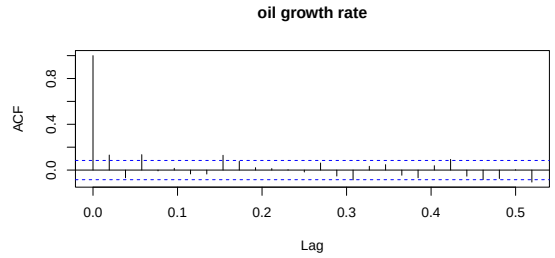
(b) If $X_{t+1} = (1+r)X_t$ then $\log(X_{t+1}/X_t) = \log(1+r)$ which is approximately r if r is close to 0. so,$\nabla \log(x_t)$ is a good approximation.

(c)
```
gas_gr <- diff(log(gas))
oil_gr <- diff(log(oil))
plot.ts(gas_gr,main="oil and gas growth rate",ylab = 'growth rate',col='1',las=1)
lines(oil_gr, col='2')
legend("topright", legend = c("gas growth rate","oil growth rate"),
       lty = 1:1, col = 1:2, bty = 'n', cex=0.6)
```



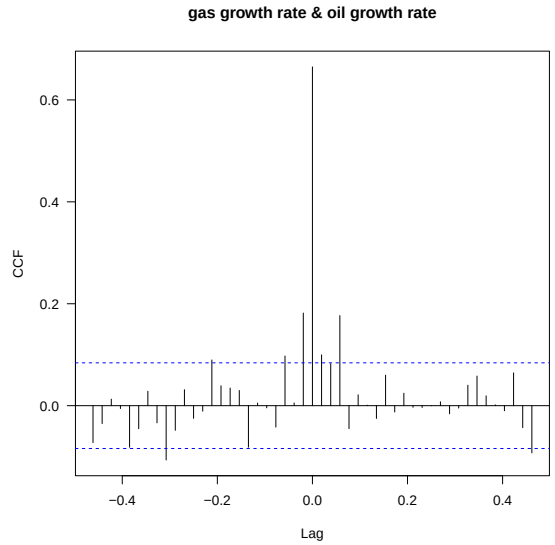

```
par(mfrow=c(2,1))
acf(gas_gr,main = 'gas growth rate')
acf(oil_gr,main= 'oil growth rate')
```





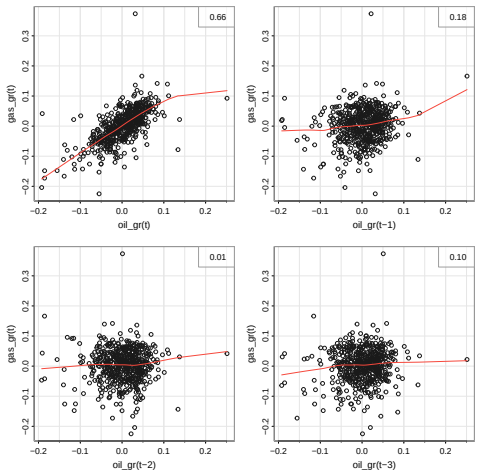
We can see that the transformed data looks fairly stationary since most of the ACF (excluding lag 0) lies within the 95% confidence interval.

(d)
```
par(mfrow=c(1,1))
ccf(oil_gr, gas_gr, main = 'gas growth rate & oil growth rate',
    ylab = 'CCF',las=1)
```



The plot here is of $\gamma_{\text{oil,gas}}(h) = \text{Cov}(O_{t+h}, G_t)$. The strongest correlation on the plot is at $h = 0$; the two series are strongly contemporaneously correlated. Significant CCF values in this plot with lag > 0 indicate that gas leads oil; significant values when the lag is < 0 indicate that oil leads gas. We know that oil is used to create gas and so we would expect, a priori, that oil would lead gas. That would indicate we would see significant values with lag ≤ 0. We do indeed see that at a one week lag ($h = -1$) oil significantly leads gas. (It is debatable whether oil leads gas at 3 weeks, $h = -3$.) From this plot at lag $h = 3$ we also see that gas seems to lead oil by three weeks, and maybe also at weeks $h = 1, 2$. As the textbook mentions, this might be considered to be a feedback loop (e.g., where the price of gas is high and so oil sellers decide/realize they could increase the price of oil and gas sellers would still pay for it).

(e)
```
lag2.plot(oil_gr,gas_gr,3,corr=T,smooth=T)
```



We can see that nearly each plot has outliers, one at the top and one at the right. If we ignore the outliers, all of the four plots show a linear relationship. However, we can see there is a strong linear relationship between the current oil growth rate and the current gas growth rate. For the one to three weeks lead time of oil price, the linear relationships are not as strong as it the contemporaneous one is.

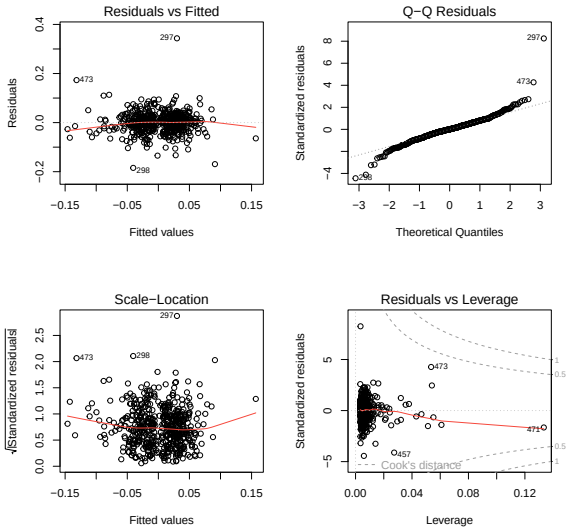
(f) i.
```
poil<-diff(log(oil))
pgas<-diff(log(gas))
indi<-ifelse(poil<0,0,1)
new<-ts.intersect(pgas,poil,poilL=lag(poil,-1),indi)
fit<-lm(pgas~poil+poilL+indi,data=new)
summary(fit)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -0.00644490 0.003464303 -1.860374 6.337704e-02
## poil         0.68312729 0.058369065 11.703584 2.379825e-28
## poilL        0.11192714 0.038554377  2.903098 3.846226e-03
## indi         0.01236821 0.005515705  2.242362 2.534376e-02
```

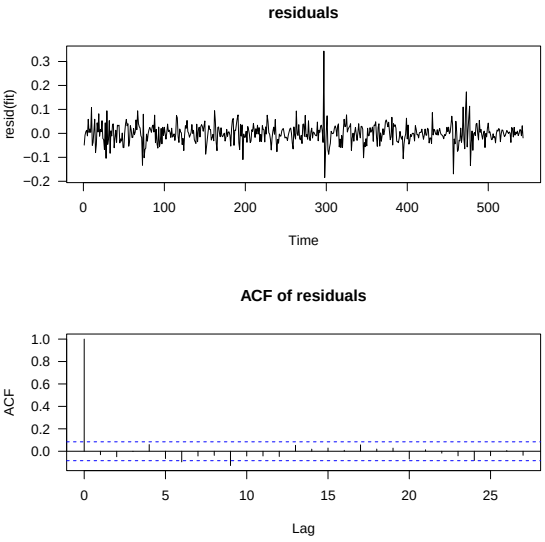
Our analysis here ignores the autocorrelated errors (i.e., the time series structure) and pretends we can apply standard linear regression techniques. Since we are ignoring the autocorrelated errors, our analysis is very preliminary, and not statistically reliable (but is the best we can do at this point in the course).
If we imagine/pretend that the p-values were "correct", then we can say that all of the coefficients except the intercept are significant at the significance level 0.05. The increasement of O_t, O_{t-1} and I_t will increase G_t, since all of their coefficients are positive. The magnitude of the coefficient of O_t is the largest, indicating that the increasement of O_t will bring the largest change in G_t. The low F test p value and medium R^2 show that this model is reasonable. (Again, none of the "significance" can be truly relied on, but we get some preliminary sense from this analysis.)

ii. When the model has a negative growth, it means indi=0, so the model is
$G_t = -0.006445 + 0.683127O_t + 0.111927O_{t-1}$
when there is a positive growth, the indi is 1, so the intercept is: $-0.006445 + 0.012368 = 0.00592331$
so the model is: $G_t = 0.00592331 + 0.683127O_t + 0.111927O_{t-1}$
If we assume the p-values are reliable, then the results support the asymmetry hypothesis, since the coefficient of I_t is statistically significant.

iii.
```
par(mfrow=c(2,2))
plot(fit)
```




```
par(mfrow=c(2,1))
plot.ts(resid(fit), main = 'residuals',las=1)
acf(resid(fit), main = 'ACF of residuals',las=1)
```



We see that the residuals plot and scale-location plot don't have a obvious pattern except there are several outliers, so we can claim that the equal variance is satisfied. The QQ plot shows a violation about normal distribution (heavy tailed), part of the reason is that there are outliers. Hence, if we remove the outliers and refit the model, the model diagonistic result will be much better.
Also, the ACF of residuals show that there is not significant correlation between residuals, which supports that our model is reasonable.

4. Consider the setup of the previous question (Question 2.10, Shumway and Stoffer), and let us focus on just the oil series. One model we might consider for the (untransformed) oil series is the random walk with drift model, $X_t = \delta_1 + X_{t-1} + W_t$ where $W_t \sim \text{WN}(0, \sigma^2)$. If we let $X_0 = \delta_0$ be a constant "intercept" term, then we have checked in class that we can write $X_t = \delta_0 + \delta_1 t + \sum_{s=1}^{t} W_s$. The mean of X_t is thus $\delta_0 + \delta_1 t$. We might be interested in estimating this (linear) regression function.

(a) Use `lm()` to regress the (untransformed) oil series on time. Print the `summary()` of the results and plot the data with the regression line. Comment briefly on the statistical significance of the coefficients.

(b) Compute the F-statistic for testing $H_0 : \delta_1 = 0$ against $H_A : \delta_1 \neq 0$ (i.e., for testing whether there is a drift) and the corresponding p-value; you can do this by using `lm()` and `summary()`.

(c) Now we return to the random walk with drift model and the results of 4b. We want to assess whether the p-values that we computed actually mean anything. We will run a simulation study to assess this, as follows.

Simulate a random walk with no drift X_t, for $t = 1, \ldots, 545$. (You may want to use the `cumsum` function.) Assume $W_t \overset{\text{iid}}{\sim} N(0, 1)$ (you may also take $\delta_0 = 0$ although the value of δ_0 will not matter here). Run the regression we ran previously in 4a of X_t on time, $E(X_t) = \delta_0 + \delta_1 t$. Get the p-value for testing $H_0 : \delta_1 = 0$. (Note: you can get p-values from `summary()$coefficients`.)

Do this procedure $M = 1000$ times. Report the proportion of p-values that are smaller than .05. Provide a comment explaining what this means for the p-value reported in 4b.

Solution:

(a)
```
library(astsa)
oilreg <- lm(oil~time(oil))
summary(oilreg)$coef
```

```
##                   Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) -13475.162986 452.5800215 -29.77410 3.187894e-116
## time(oil)        6.745965   0.2256995  29.88915 8.670476e-117
```

```
plot(oil,main = 'oil price over time',las=1)
abline(oilreg)
```

All of the coefficients are significant.

(b)
```
nn <- length(oil)
avg <- mean(oil)
SSE0 <- sum((oil-avg)^2)
SSEfull <- sum((resid(oilreg))^2)
avg <- mean(oil)
```

```
Fstat <- ((SSE0-SSEfull)/ 1) /
      (SSEfull/(nn-2))
Fstat

## [1] 893.3612

pval <- pf(q=Fstat, df1=1, df2=nn-2, lower.tail=F)
pval

## [1] 8.670476e-117
```

Using summary() gives the same result.

(c)
```
set.seed(1)
MM <- 1000
nn <- 545
pvals <- numeric(length=MM)
for (ii in 1:MM){
ww <- rnorm(nn)
xx <- ts(cumsum(ww))
reg <- lm(xx~time(xx))
pvals[ii] <- summary(reg)$coeff[2,4]
}
mean(pvals<.05)

## [1] 0.943
```

We see that the type I error rate exceeds 90% while we would expect it to be around 5%. This implies that the result in 4b is not reliable.