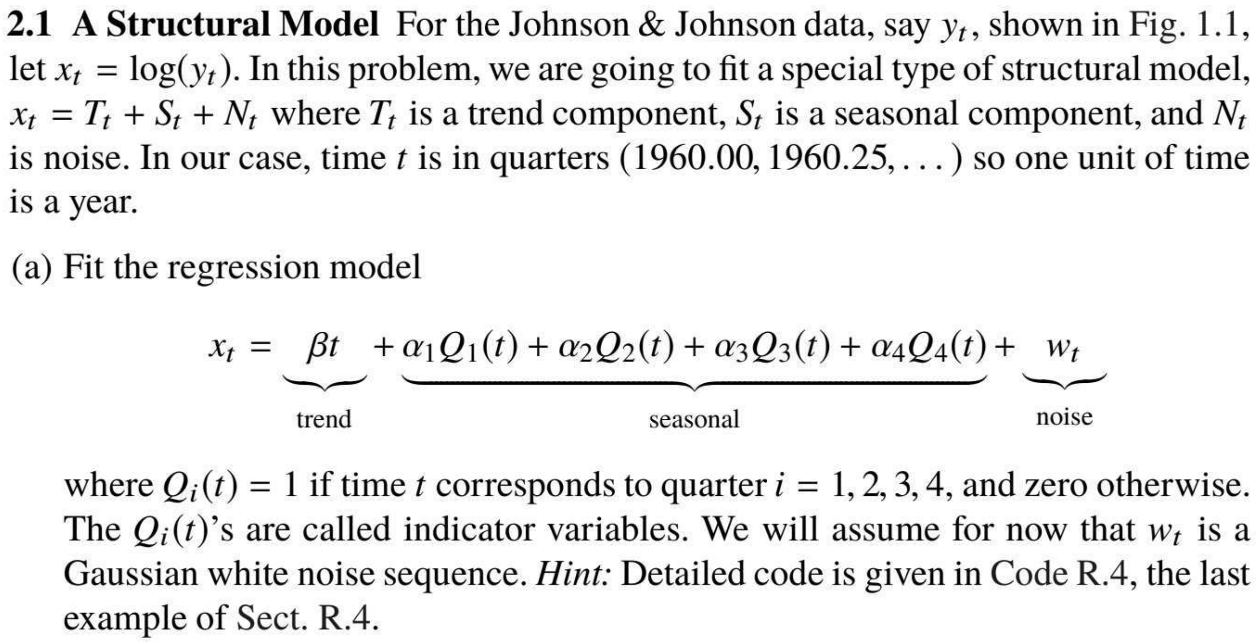
Chapter 2

2.1



2.1 A Structural Model For the Johnson & Johnson data, say y_t, shown in Fig. 1.1, let $x_t = \log(y_t)$. In this problem, we are going to fit a special type of structural model, $x_t = T_t + S_t + N_t$ where T_t is a trend component, S_t is a seasonal component, and N_t is noise. In our case, time t is in quarters (1960.00, 1960.25, . . .) so one unit of time is a year.

(a) Fit the regression model

$$x_t = \underbrace{\beta t}_{\text{trend}} + \underbrace{\alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t)}_{\text{seasonal}} + \underbrace{w_t}_{\text{noise}}$$

where $Q_i(t) = 1$ if time t corresponds to quarter $i = 1, 2, 3, 4$, and zero otherwise. The $Q_i(t)$'s are called indicator variables. We will assume for now that w_t is a Gaussian white noise sequence. *Hint:* Detailed code is given in Code R.4, the last example of Sect. R.4.

```
y = jj; x = log(y)
trend = time(x) - 1970 # Centres time
Q = factor(cycle(x))
fit1 <- lm(x~0+trend+Q, na.action=NULL) # No intercept
summary(fit1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29318 -0.09062 -0.01180  0.08460  0.27644

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
trend 0.167172   0.002259   74.00   <2e-16 ***
Q1    1.052793   0.027359   38.48   <2e-16 ***
Q2    1.080916   0.027365   39.50   <2e-16 ***
Q3    1.151024   0.027383   42.03   <2e-16 ***
Q4    0.882266   0.027412   32.19   <2e-16 ***
```

(b) If the model is correct, what is the estimated average annual increase in the logged earnings per share?

The estimated annual increase is $\hat{\beta}$ = .167.

(c) If the model is correct, does the average logged earnings rate increase or decrease from the third quarter to the fourth quarter? And, by what percentage does it increase or decrease?

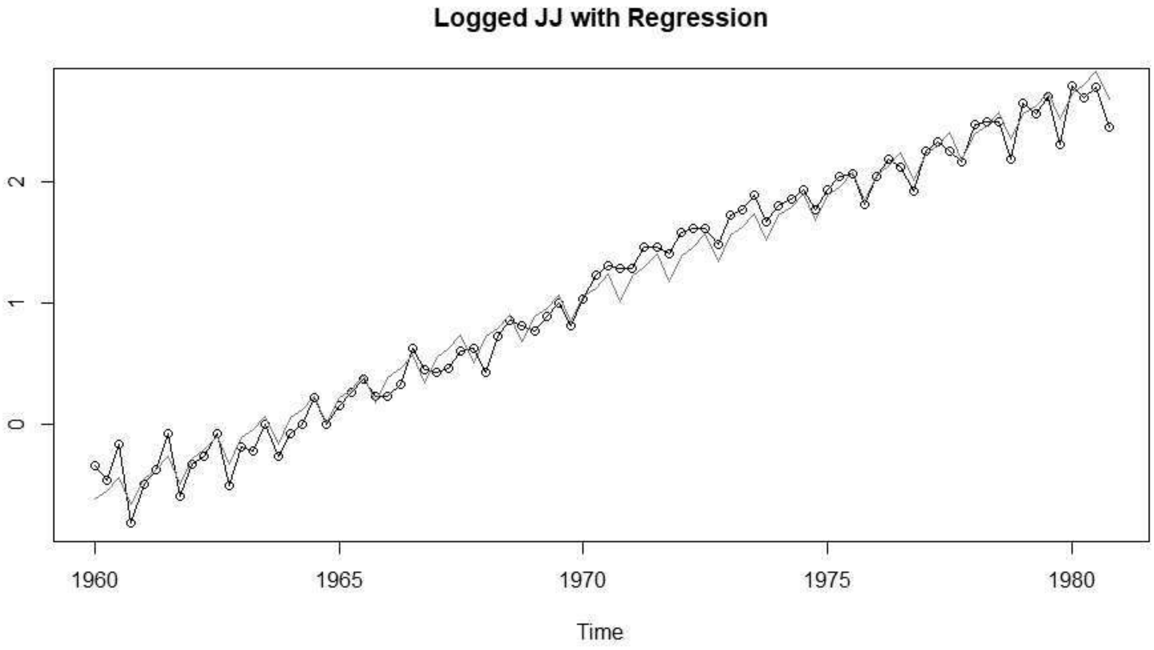
The average logged earnings rate decreases by 0.27 from the third to fourth quarter, a roughly 23% decrease.

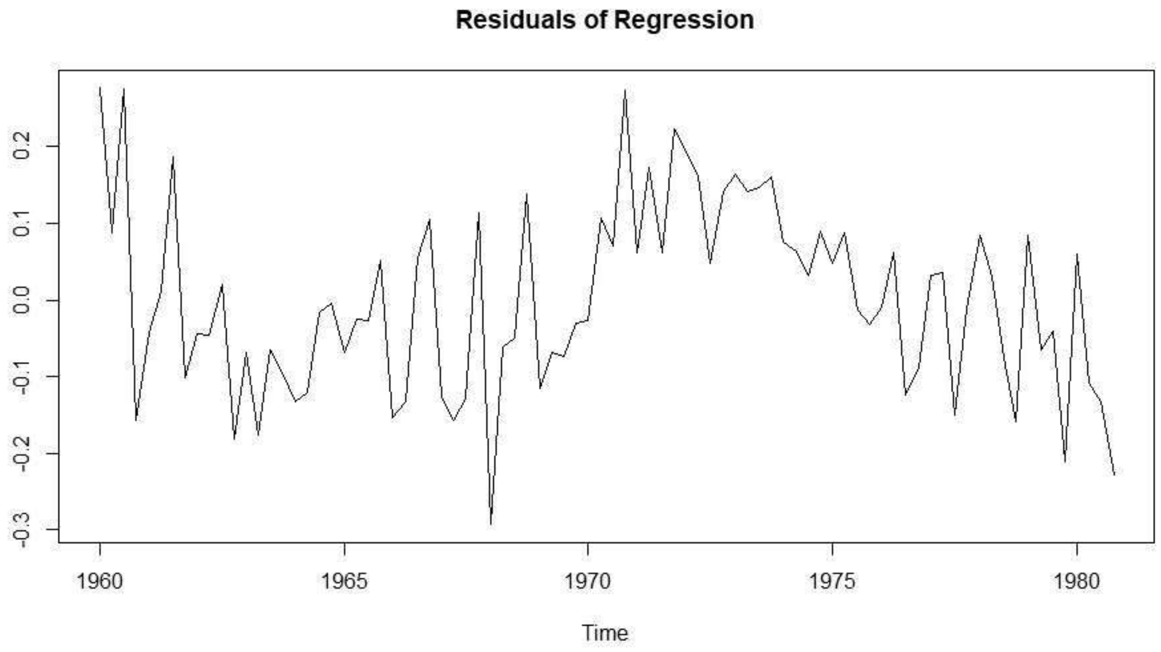
(d) What happens if you include an intercept term in the model in (a)? Explain why there was a problem.

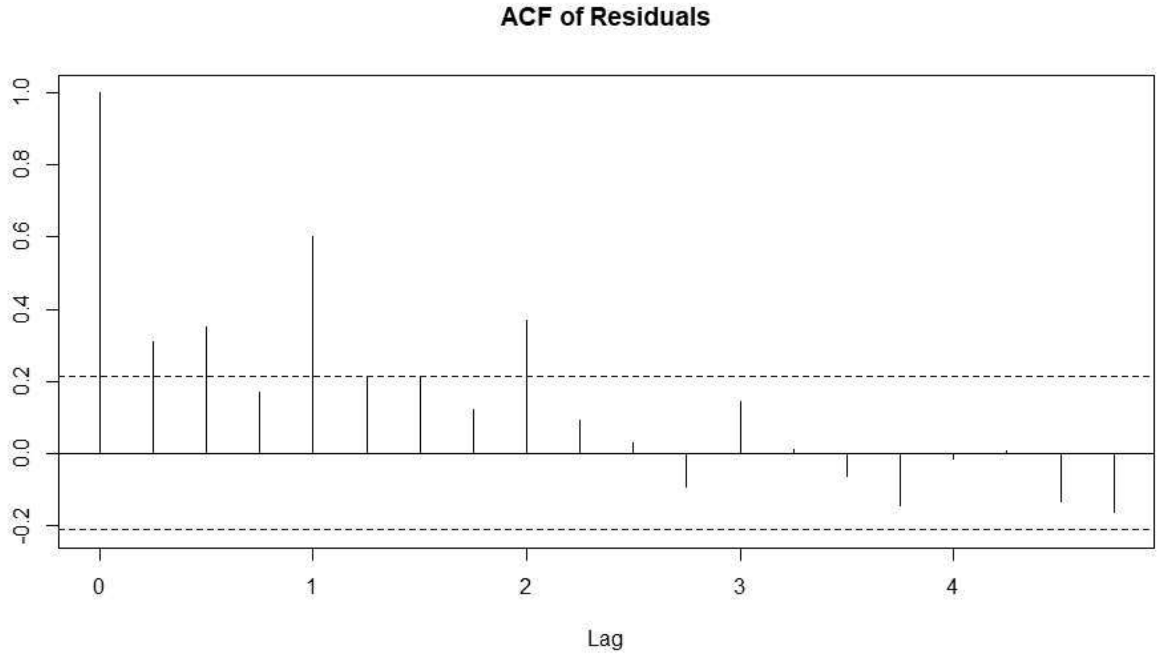
If the model is fitted with an intercept, R will remove Q1, as it will be overparameterised.

(e) Graph the data, x_t, and superimpose the fitted values, say $\hat{x}_t$, on the graph. Examine the residuals, $x_t - \hat{x}_t$, and state your conclusions. Does it appear that the model fits the data well (do the residuals look white)?

```
plot(x, type="o", main='Logged JJ with Regression', ylab='')
lines(fitted(fit1), col=2)
plot.ts(resid(fit1), main="Residuals of Regression", ylab='')
acf(resid(fit1), main="ACF of Residuals", ylab='')
```

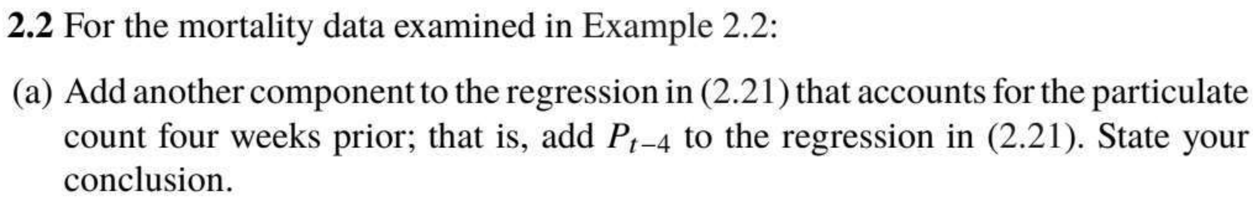






We can still see substantial correlation in the residuals, particularly at the yearly cycle. The residuals do not look white.

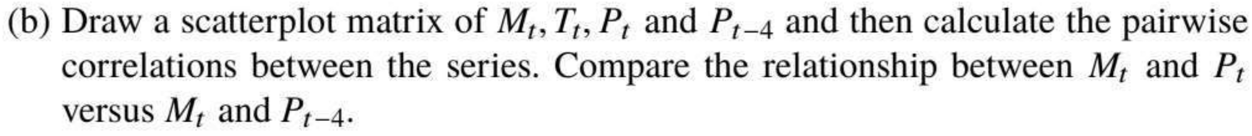
2.2

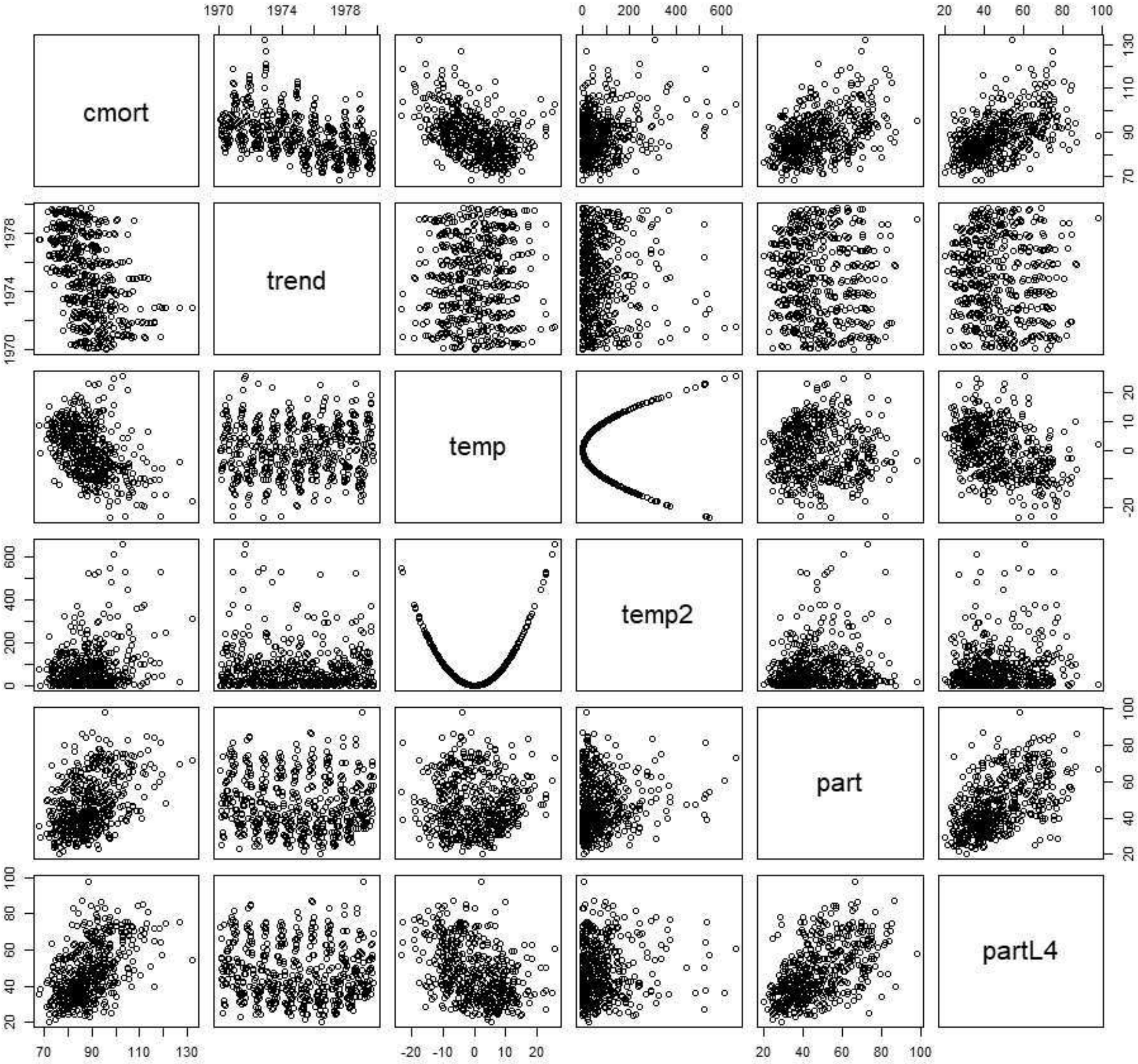

2.2 For the mortality data examined in Example 2.2:

(a) Add another component to the regression in (2.21) that accounts for the particulate count four weeks prior; that is, add P_{t-4} to the regression in (2.21). State your conclusion.

```
ded = ts.intersect(cmort, trend=time(cmort), temp, temp2=temp^2, part,
partL4=lag(part,-4))
fit <- lm(cmort~trend + temp + temp2 + part + partL4, data=ded)
summary(fit)

Residuals:
    Min      1Q  Median      3Q     Max
-18.228  -4.314  -0.614   3.713  27.800

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.808e+03  1.989e+02  14.123  < 2e-16 ***
trend       -1.385e+00  1.006e-01 -13.765  < 2e-16 ***
temp        -4.058e-01  3.528e-02 -11.503  < 2e-16 ***
temp2        2.155e-02  2.803e-03   7.688 8.02e-14 ***
part         2.029e-01  2.266e-02   8.954  < 2e-16 ***
partL4       1.030e-01  2.485e-02   4.147 3.96e-05 ***
```

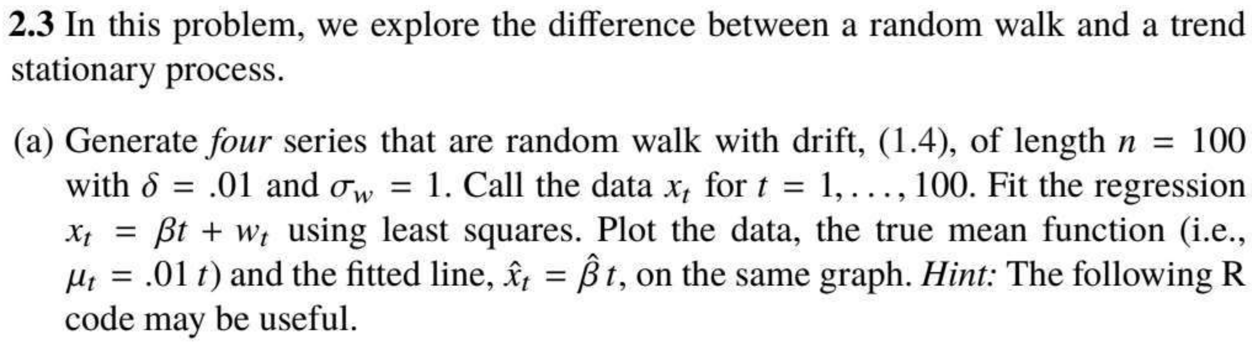
We can see that the partL4 parameter is statistically significant. The AIC prefers the updated model (4.692916 v 4.773793), however, the BIC prefers the original (4.751563 v 4.824156).


(b) Draw a scatterplot matrix of M_t, T_t, P_t and P_{t-4} and then calculate the pairwise correlations between the series. Compare the relationship between M_t and P_t versus M_t and P_{t-4}.



The relationships are similar between P_t and M_t, and P_{t-4} and M_t, with a slightly elongated cluster for both, starting from the bottom left moving towards the top right. Running cor(ded) will show that the correlation of P_{t-4} with M_t is 0.52. pairs(ded) gives the scatterplots.

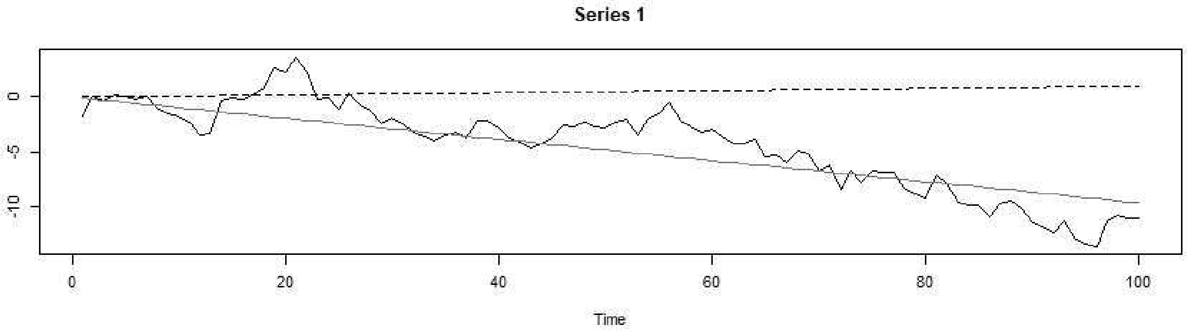
2.3

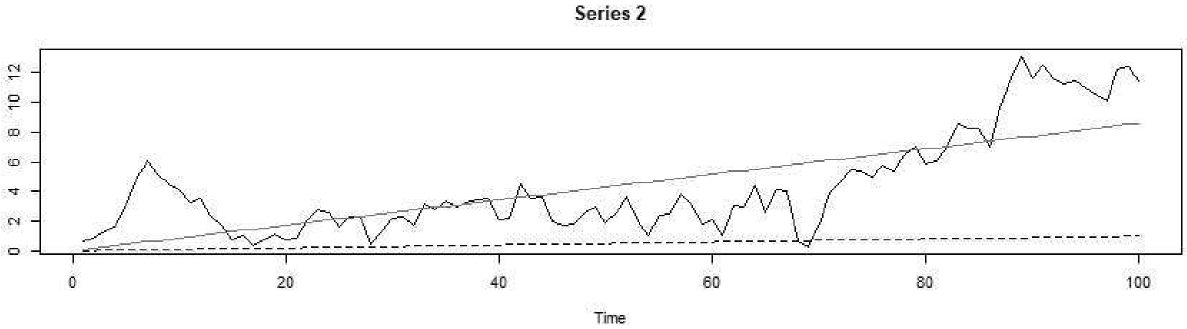


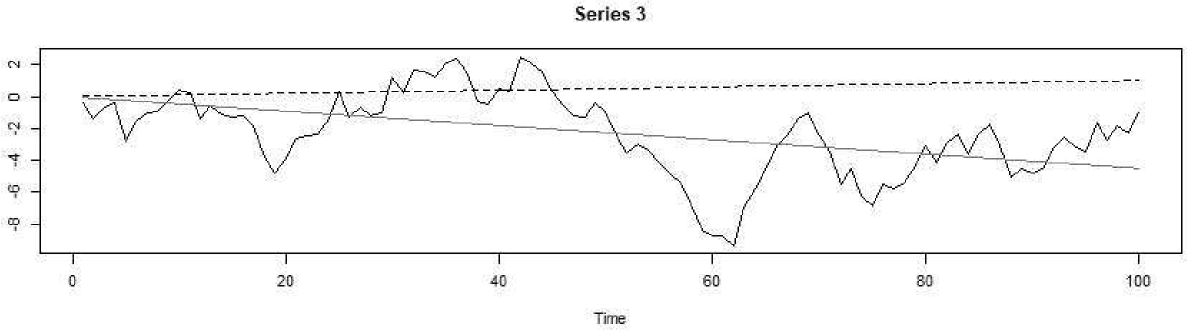
2.3 In this problem, we explore the difference between a random walk and a trend stationary process.

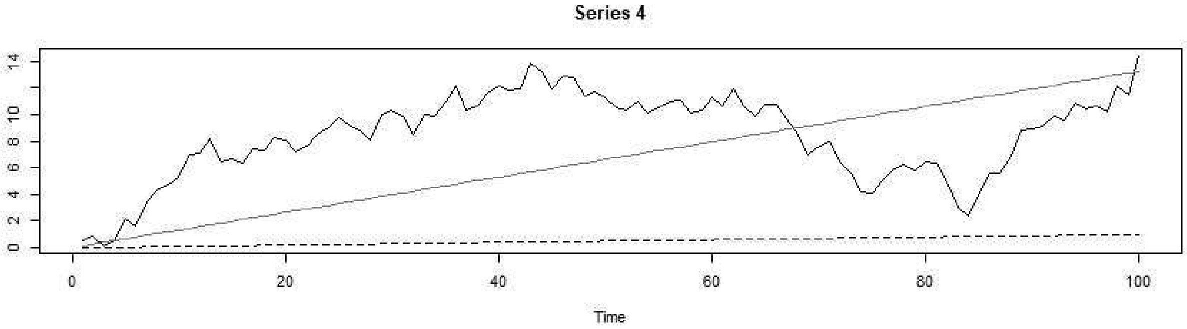
(a) Generate *four* series that are random walk with drift, (1.4), of length $n = 100$ with $\delta = .01$ and $\sigma_w = 1$. Call the data x_t for $t = 1, \dots, 100$. Fit the regression $x_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .01\, t$) and the fitted line, $\hat{x}_t = \hat{\beta}\, t$, on the same graph. *Hint:* The following R code may be useful.

```
par(mfrow=c(4,1))

drift_series = function(num, n=100, drift=0.01) {
  w = rnorm(n)
  x = rep(0, n+1)
  for (i in 2:(n+1)) {
    x[i] = drift + x[i-1] + w[i-1]
  }
  fit <- lm(x[-1] ~ 0+time(x[-1])) # No intercept
  plot.ts(x[-1], main=sprintf("Series %s", num), ylab='')
  lines(fitted(fit), col=4)
  lines(0.01 * time(x[-1]), lty=2)
}

for (i in 1:4) {
  drift_series(i)
}
```



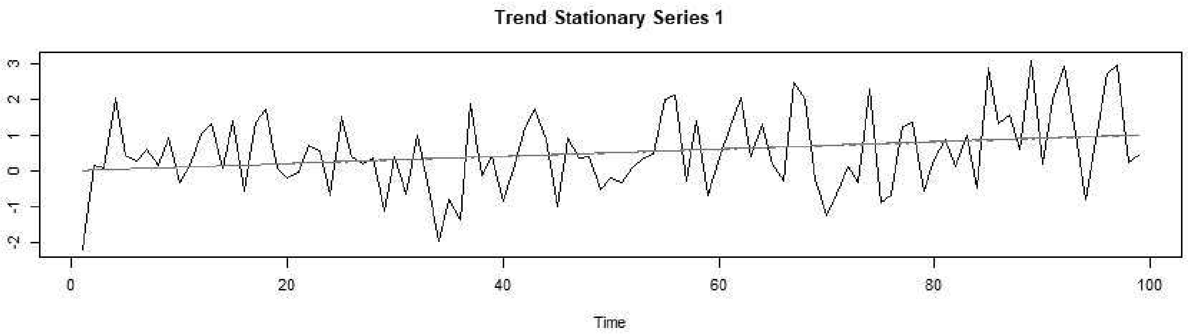


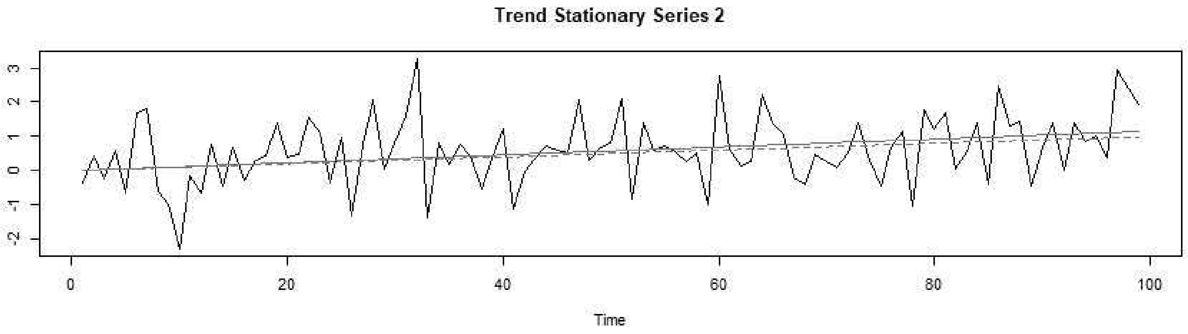


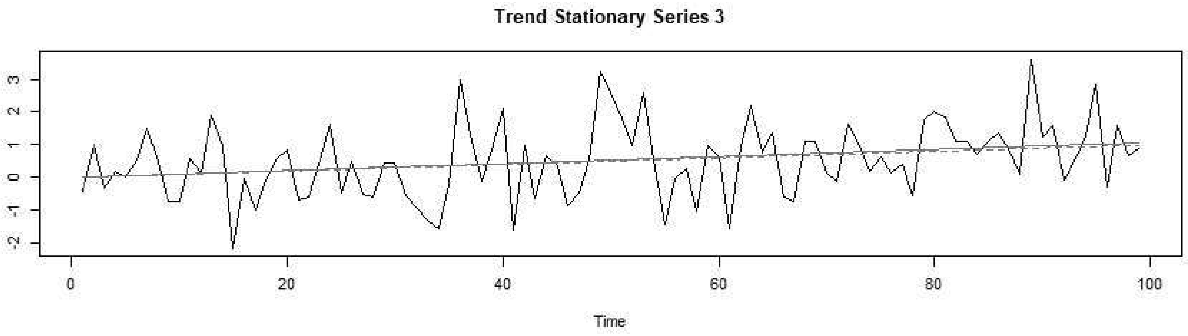


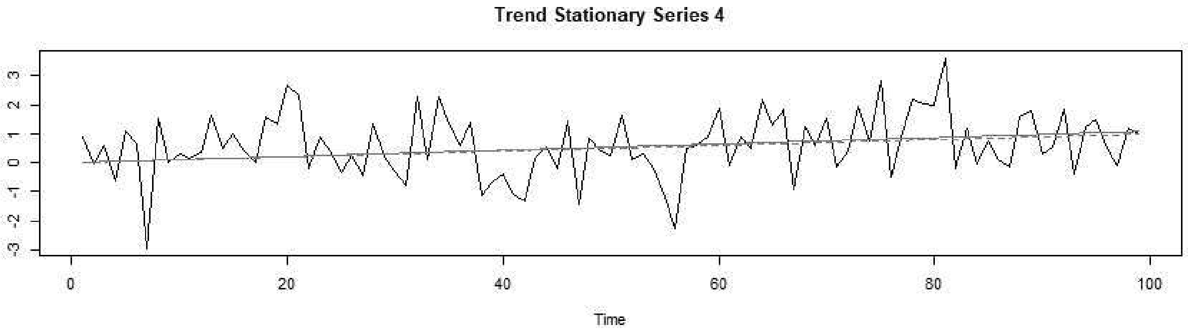
(b) Generate *four* series of length $n = 100$ that are linear trend plus noise, say $y_t = .01\,t + w_t$, where t and w_t are as in part (a). Fit the regression $y_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .01\,t$) and the fitted line, $\hat{y}_t = \hat{\beta}\,t$, on the same graph.

```
trend_station = function(num, n=100, drift=0.01) {
  w = rnorm(n)
  x = rep(0, n)
  for (i in 1:n) {
    x[i] = drift*i + w[i]
  }
  fit <- lm(x[-1] ~ 0+time(x[-1]))
  plot.ts(x[-1], main=sprintf("Trend Stationary Series %s", num), ylab='')
  lines(fitted(fit), col=4)
  lines(0.01 * time(x[-1]), lty=2, col=2)
}

for (i in 1:4) {
  trend_station(i)
}
```





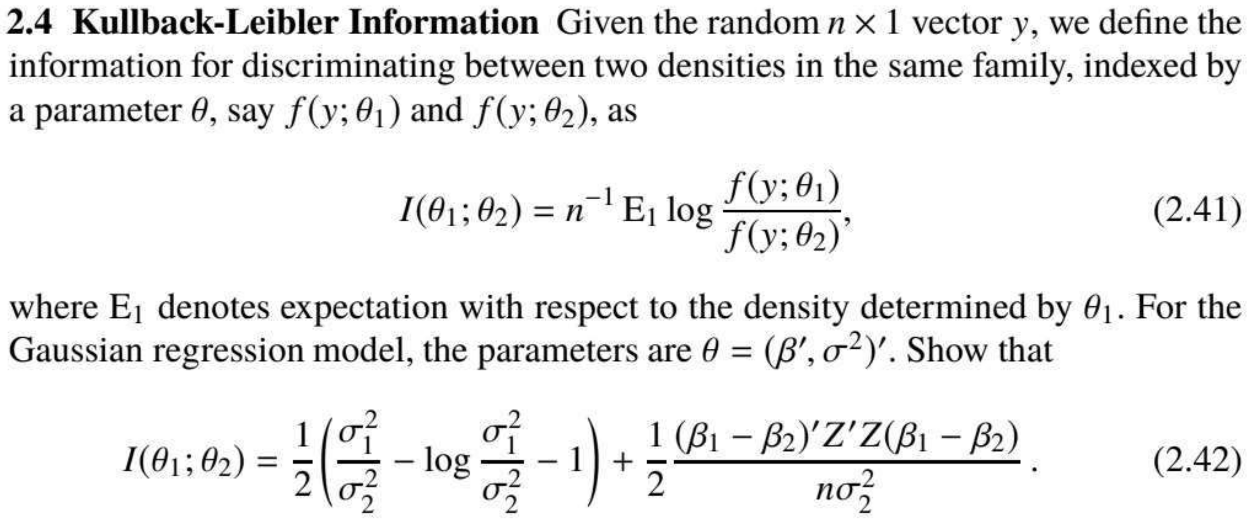




(c) Comment (what did you learn from this assignment).

The regression line and the true mean function more closely match for the trend stationary data, whereas for the random walk with drift data, although it is possible for them to closely match, the mean function and regression fit are likely to diverge by a considerable degree. The data points for the trend stationary series tend to fluctuate around the mean, and not diverge too far; while the random walk with drift data points curve out their own path, which may end up in the opposite direction to the mean function.

2.4



2.4 Kullback-Leibler Information Given the random $n \times 1$ vector y, we define the information for discriminating between two densities in the same family, indexed by a parameter θ, say $f(y;\theta_1)$ and $f(y;\theta_2)$, as

$$I(\theta_1;\theta_2) = n^{-1}\,\mathrm{E}_1 \log \frac{f(y;\theta_1)}{f(y;\theta_2)}, \tag{2.41}$$

where E_1 denotes expectation with respect to the density determined by θ_1. For the Gaussian regression model, the parameters are $\theta = (\beta', \sigma^2)'$. Show that

$$I(\theta_1;\theta_2) = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{1}{2}\frac{(\beta_1-\beta_2)'Z'Z(\beta_1-\beta_2)}{n\sigma_2^2}. \tag{2.42}$$

The PDF for the entire set of data is given by

$P(\tilde{Y} = \tilde{y}|\theta) = P(Y_1 = y_1 Y_n = y_n|\theta)$

If they are independent variables, then

$P(\tilde{Y} = \tilde{y}|\theta) = P(Y_1 = y_1)P(Y_n = y_n|\theta)$

This implies that when taking the log of a PDF and you have a sample of observed values:

$l_x(\mu,\sigma^2) = \sum_{i=1}^{n} l_{x_i}(\mu,\sigma^2)$

Hence,

$$log\frac{f(y;\theta_1)}{f(y;\theta_2)} = log\frac{\frac{1}{\sigma_1\sqrt{2\pi}}exp(-\frac{(y-\beta_1'z_t)^2}{2\sigma_1^2})}{\frac{1}{\sigma_2\sqrt{2\pi}}exp(-\frac{(y-\beta_2'z_t)^2)}{2\sigma_2^2})}$$

$$= log\frac{1}{\sigma_2}exp(-\frac{(y-\beta_1'z_t)^2}{2\sigma_1^2}) - log\frac{1}{\sigma_1}exp(-\frac{(y-\beta_2'z_t)^2)}{2\sigma_2^2})$$

$$= -log\sum_{t=1}^{n}\sigma_1 + log\sum_{t=1}^{n}\sigma_2 + \sum_{t=1}^{n} -\frac{(y_t-\beta_1'z_t)^2}{2\sigma_1^2} - \sum_{t=1}^{n} -\frac{(y_t-\beta_2'z_t)^2}{2\sigma_2^2}$$

$$= -\frac{n}{2}log\sigma_1^2 + \frac{n}{2}log\sigma_2^2 - \frac{1}{2\sigma_1^2}\sum_{t=1}^{n}(y_t-\beta_1'z_t)^2 + \frac{1}{2\sigma_2^2}\sum_{t=1}^{n}(y_t-\beta_2'z_t)^2$$

Now,

$$E_1[(y_t-\beta_2')^2] = E_1[y_ty_t - y_t\beta_2'z_t - y_t\beta_2'z_t + \beta_2'z_t\beta_2'z_t]$$

$$= E_1[y_ty_t] - \beta_1'z_t\beta_1'z_t + \beta_1'z_t\beta_1'z_t - \beta_1'z_t\beta_2'z_t - \beta_1'z_t\beta_2'z_t + \beta_2'z_t\beta_2'z_t$$

$$= \sigma_1^2 + (\beta_1-\beta_2)'z_tz_t'(\beta_1-\beta_2)$$

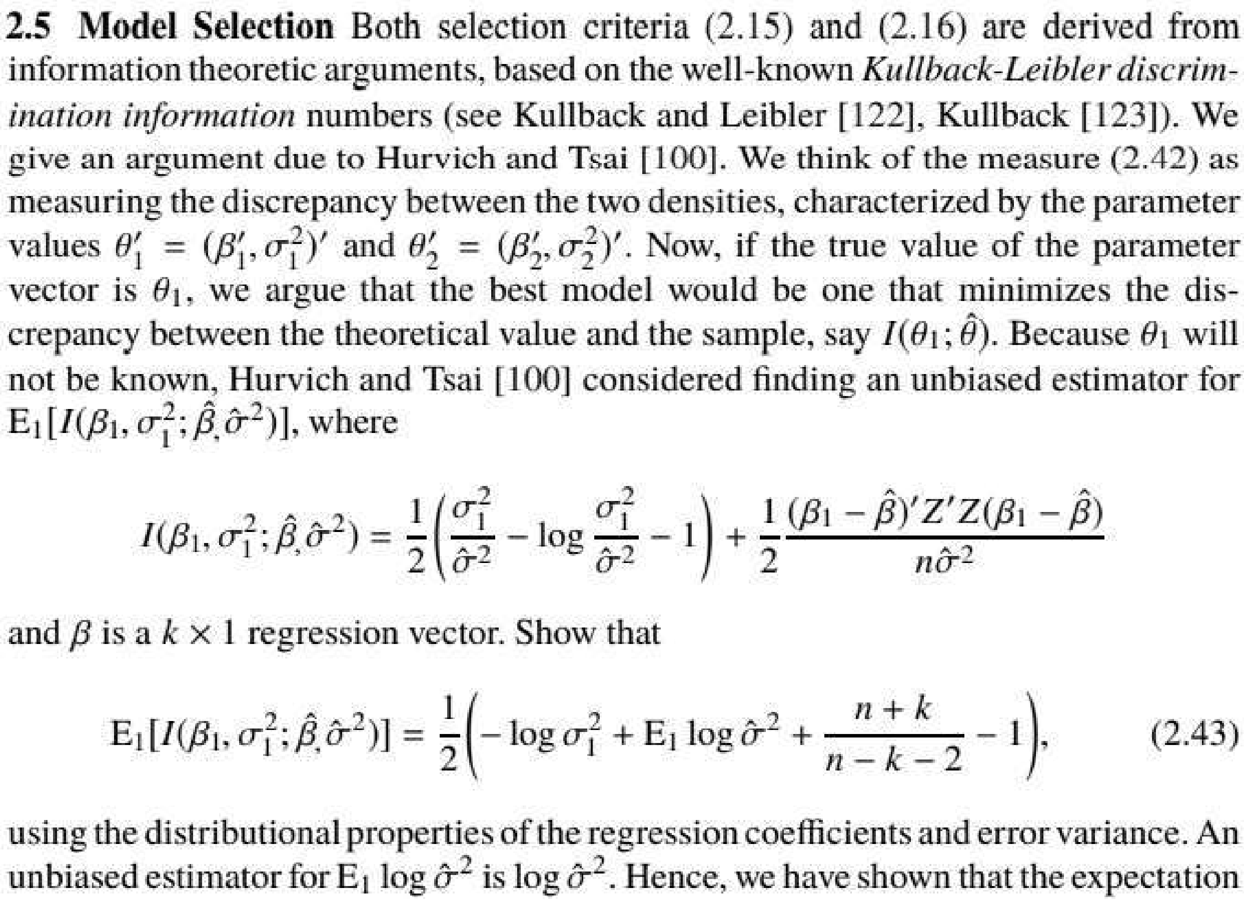
And,

$$E_1[(y_t - \beta_1' z_t)^2] = E_1[y_t y_t] - \beta_1' z_t \beta_1' z_t = \sigma_1^2$$

$$\sum_{t=1}^{n} (\beta_1 - \beta_2)' z_t z_t' (\beta_1 - \beta_2) = (\beta_1 - \beta_2)' Z' Z (\beta_1 - \beta_2)$$

Filling in the gaps, we arrive at the result.

2.5



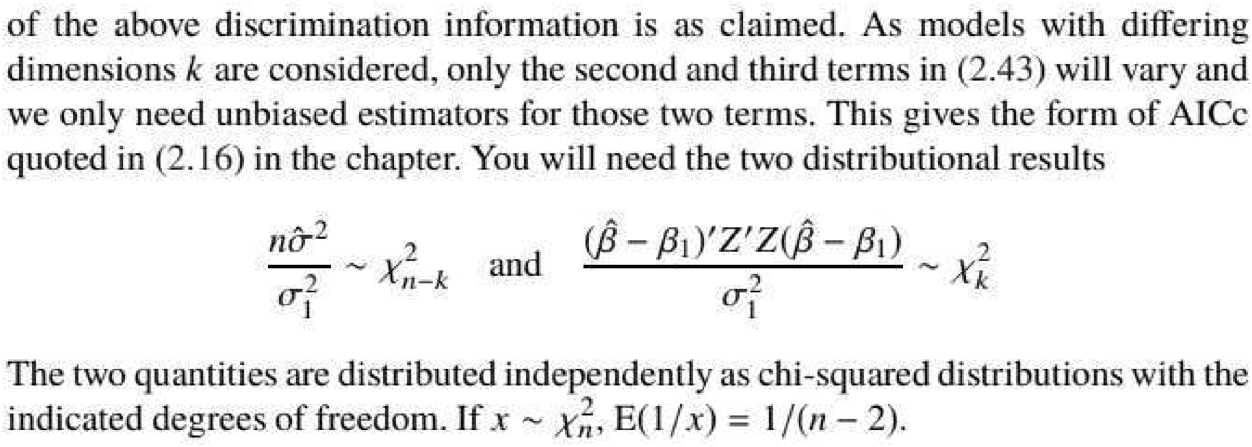
2.5 Model Selection Both selection criteria (2.15) and (2.16) are derived from information theoretic arguments, based on the well-known *Kullback-Leibler discrimination information* numbers (see Kullback and Leibler [122], Kullback [123]). We give an argument due to Hurvich and Tsai [100]. We think of the measure (2.42) as measuring the discrepancy between the two densities, characterized by the parameter values $\theta_1' = (\beta_1', \sigma_1^2)'$ and $\theta_2' = (\beta_2', \sigma_2^2)'$. Now, if the true value of the parameter vector is θ_1, we argue that the best model would be one that minimizes the discrepancy between the theoretical value and the sample, say $I(\theta_1; \hat{\theta})$. Because θ_1 will not be known, Hurvich and Tsai [100] considered finding an unbiased estimator for $\mathrm{E}_1[I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2)]$, where

$$I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2) = \frac{1}{2}\left(\frac{\sigma_1^2}{\hat{\sigma}^2} - \log\frac{\sigma_1^2}{\hat{\sigma}^2} - 1\right) + \frac{1}{2}\frac{(\beta_1 - \hat{\beta})' Z' Z (\beta_1 - \hat{\beta})}{n\hat{\sigma}^2}$$

and β is a $k \times 1$ regression vector. Show that

$$\mathrm{E}_1[I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2)] = \frac{1}{2}\left(-\log\sigma_1^2 + \mathrm{E}_1 \log\hat{\sigma}^2 + \frac{n+k}{n-k-2} - 1\right), \qquad (2.43)$$

using the distributional properties of the regression coefficients and error variance. An unbiased estimator for $\mathrm{E}_1 \log\hat{\sigma}^2$ is $\log\hat{\sigma}^2$. Hence, we have shown that the expectation



of the above discrimination information is as claimed. As models with differing dimensions k are considered, only the second and third terms in (2.43) will vary and we only need unbiased estimators for those two terms. This gives the form of AICc quoted in (2.16) in the chapter. You will need the two distributional results

$$\frac{n\hat{\sigma}^2}{\sigma_1^2} \sim \chi^2_{n-k} \quad \text{and} \quad \frac{(\hat{\beta}-\beta_1)'Z'Z(\hat{\beta}-\beta_1)}{\sigma_1^2} \sim \chi^2_k$$

The two quantities are distributed independently as chi-squared distributions with the indicated degrees of freedom. If $x \sim \chi^2_n$, E$(1/x) = 1/(n-2)$.

$$Define : \beta Z = (\hat{\beta}-\beta_1)'Z'Z(\hat{\beta}-\beta_1)$$

$$E_1[I(\beta_1,\sigma_1^2;\hat{\beta},\hat{\sigma^2})] = \frac{1}{2}E_1[\frac{\sigma_1^2}{\hat{\sigma^2}}] - \frac{1}{2}E_1[log\frac{\sigma^2}{\hat{\sigma^2}}] - \frac{1}{2} + \frac{1}{2}E_1[\frac{\beta Z}{n\hat{\sigma^2}}]$$

Given the Chi distributions, we have

$$E_1[\frac{\sigma_1^2}{n\hat{\sigma^2}}] = \frac{1}{n-k-2}$$

$$E_1[\frac{\beta Z}{\sigma_1^2}] = k$$

Also,

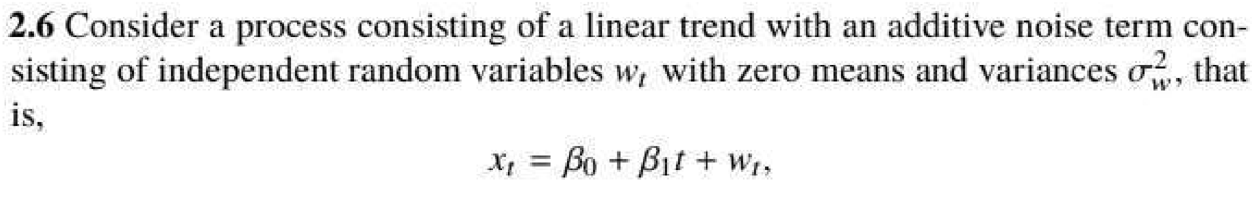
$$E_1[\frac{\beta Z}{n\hat{\sigma^2}}] = E_1[\frac{\beta Z}{\sigma_1^2}\frac{\sigma_1^2}{n\hat{\sigma^2}}] = E_1[\frac{\beta Z}{\sigma_1^2}]E_1[\frac{\sigma_1^2}{n\hat{\sigma^2}}] = \frac{k}{n-k-2}$$

Hence,

$$E_1[I(\beta_1,\sigma_1^2;\hat{\beta},\hat{\sigma^2})] = \frac{n}{2}\frac{1}{n-2} - log\sigma_1^2 + log\hat{\sigma^2} - \frac{1}{2} + \frac{1}{2}k\frac{1}{n-k-2}$$

Which simplifies to the result.

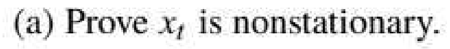
2.6



2.6 Consider a process consisting of a linear trend with an additive noise term consisting of independent random variables w_t with zero means and variances σ_w^2, that is,

$$x_t = \beta_0 + \beta_1 t + w_t,$$

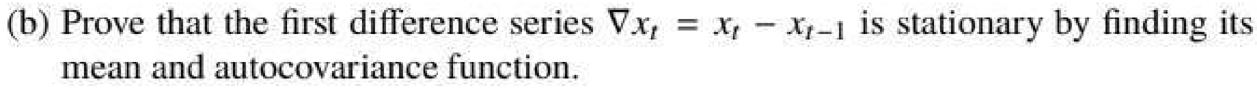
where β_0, β_1 are fixed constants.



(a) Prove x_t is nonstationary.

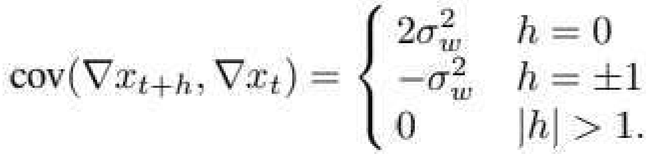
$$\mathbb{E}[x_t] = \beta_0 + \beta_1 t$$

Hence, the mean is non-constant and depends on t. Hence, the series is non-stationary.

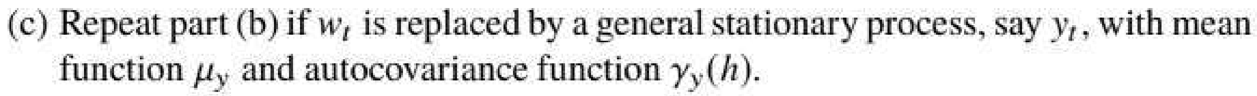

(b) Prove that the first difference series $\nabla x_t = x_t - x_{t-1}$ is stationary by finding its mean and autocovariance function.

$$\nabla x_t = \beta_1 + w_t - w_{t-1}$$
$$\mathbb{E}[\nabla x_t] = \beta_1$$
$$\gamma(h) = \mathbb{E}[w_t w_{t+h} - w_t w_{t-1+h} - w_{t-1} w_{t+h} + w_{t-1} w_{t-1+h}]$$

Hence,

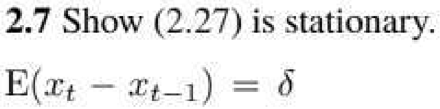

$$\operatorname{cov}(\nabla x_{t+h}, \nabla x_t) = \begin{cases} 2\sigma_w^2 & h = 0 \\ -\sigma_w^2 & h = \pm 1 \\ 0 & |h| > 1. \end{cases}$$

The mean is constant. The mean and autocovariance function both do not depend on time.

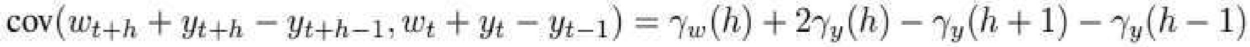

(c) Repeat part (b) if w_t is replaced by a general stationary process, say y_t, with mean function μ_y and autocovariance function $\gamma_y(h)$.

$$\gamma(h) = \mathbb{E}[y_t y_{t+h} - y_t y_{t-1+h} - y_{t-1} y_{t+h} + y_{t-1} y_{t-1+h}]$$
$$= \gamma_y(h) - \gamma_y(h-1) - \gamma_y(h+1) + \gamma_y(h)$$
$$= 2\gamma_y(h) - \gamma_y(h-1) - \gamma_y(h+1)$$

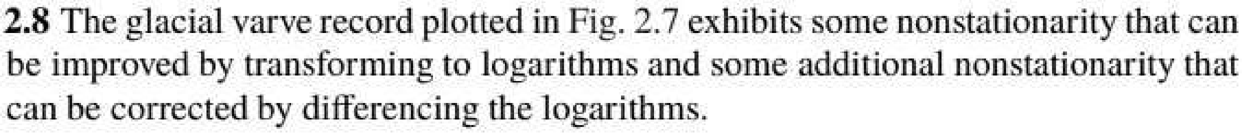
2.7

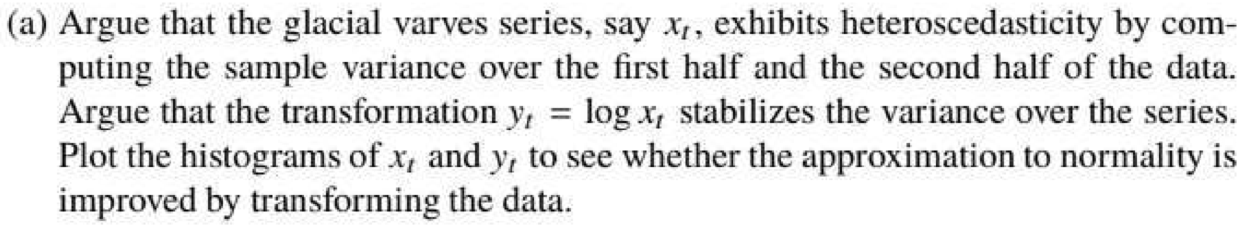

2.7 Show (2.27) is stationary.

$$\mathrm{E}(x_t - x_{t-1}) = \delta$$


$$\operatorname{cov}(w_{t+h} + y_{t+h} - y_{t+h-1}, w_t + y_t - y_{t-1}) = \gamma_w(h) + 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1)$$

2.8


2.8 The glacial varve record plotted in Fig. 2.7 exhibits some nonstationarity that can be improved by transforming to logarithms and some additional nonstationarity that can be corrected by differencing the logarithms.


(a) Argue that the glacial varves series, say x_t, exhibits heteroscedasticity by computing the sample variance over the first half and the second half of the data. Argue that the transformation $y_t = \log x_t$ stabilizes the variance over the series. Plot the histograms of x_t and y_t to see whether the approximation to normality is improved by transforming the data.

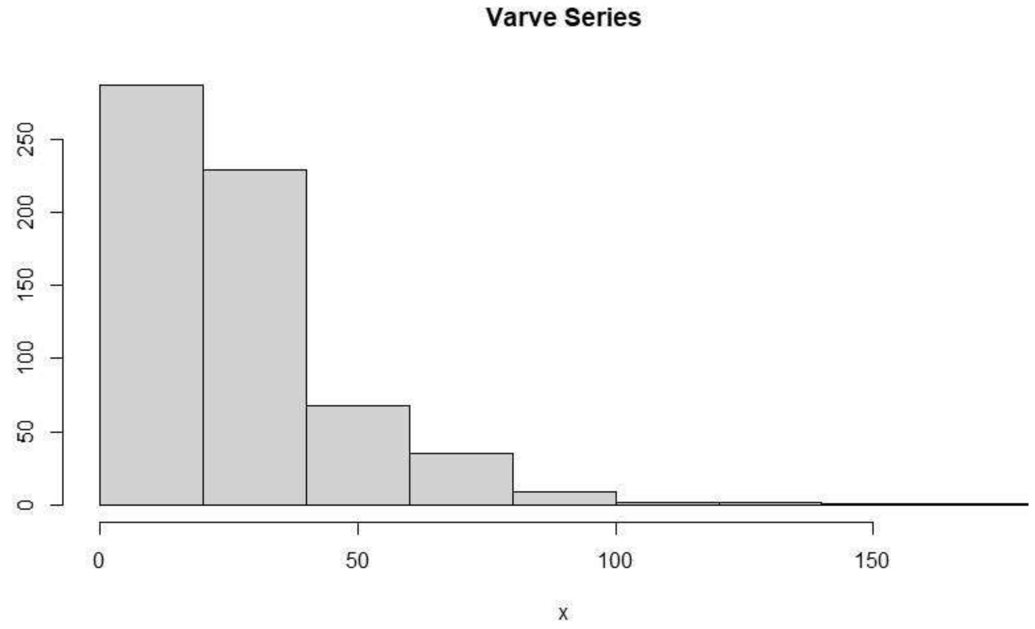
```
library(astsa)
x = varve
n = length(x)
x1 = x[1:(n/2)]; x2 = x[((n/2)+1):n]
sd1 = sd(x1) # 11.55238
sd2 = sd(x2) # 24.38217
```

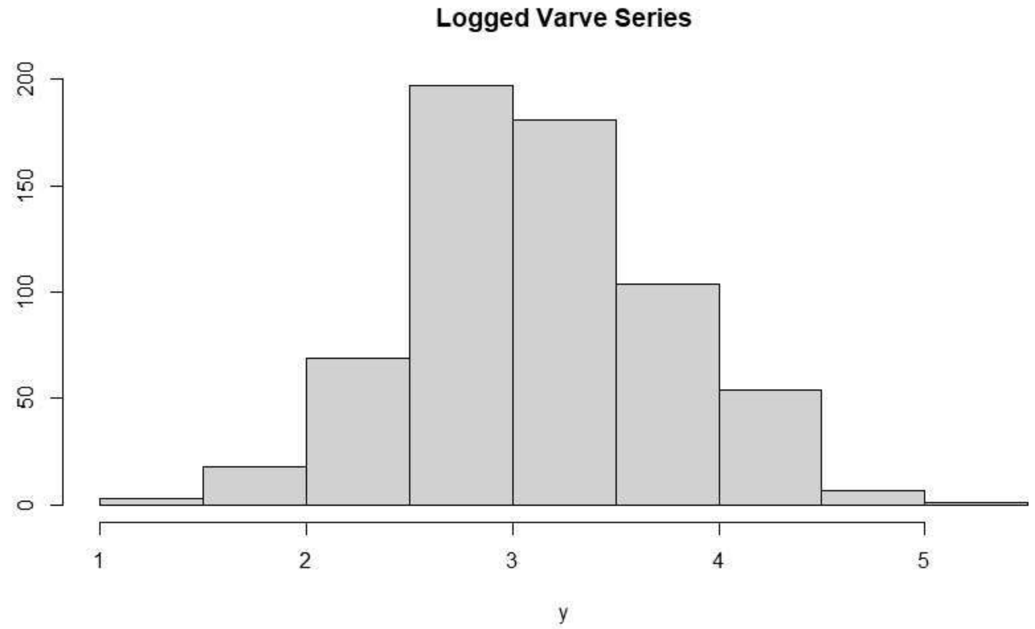
The standard deviation of the second half of the data is significantly greater (over 2x) than that of the first half, implying a non-constant standard deviation.

```
y = log(x)
y1 = y[1:(n/2)]; y2 = y[((n/2)+1):n]
sdy1 = sd(y1) # 0.5203092
sdy2 = sd(y2) # 0.6718415
```

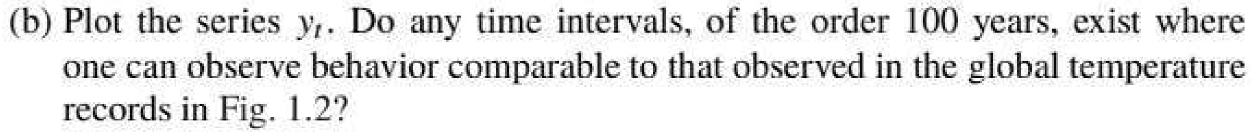
After logging data, we still have a difference in the standard deviations of the first and second halves of the data. However, the difference has been reduced significantly. The second standard deviation is only 1.2x greater.

```
par(mfrow=c(2,1))
hist(x, main="Varve Series", ylab='')
hist(y, main="Logged Varve Series", ylab='')
```



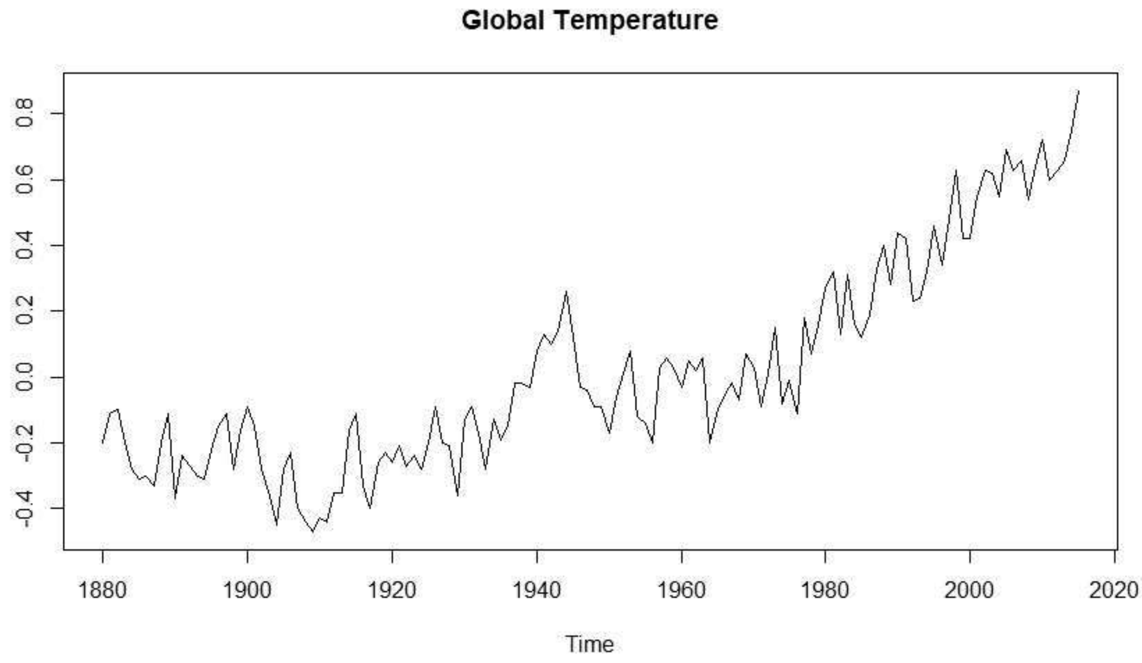


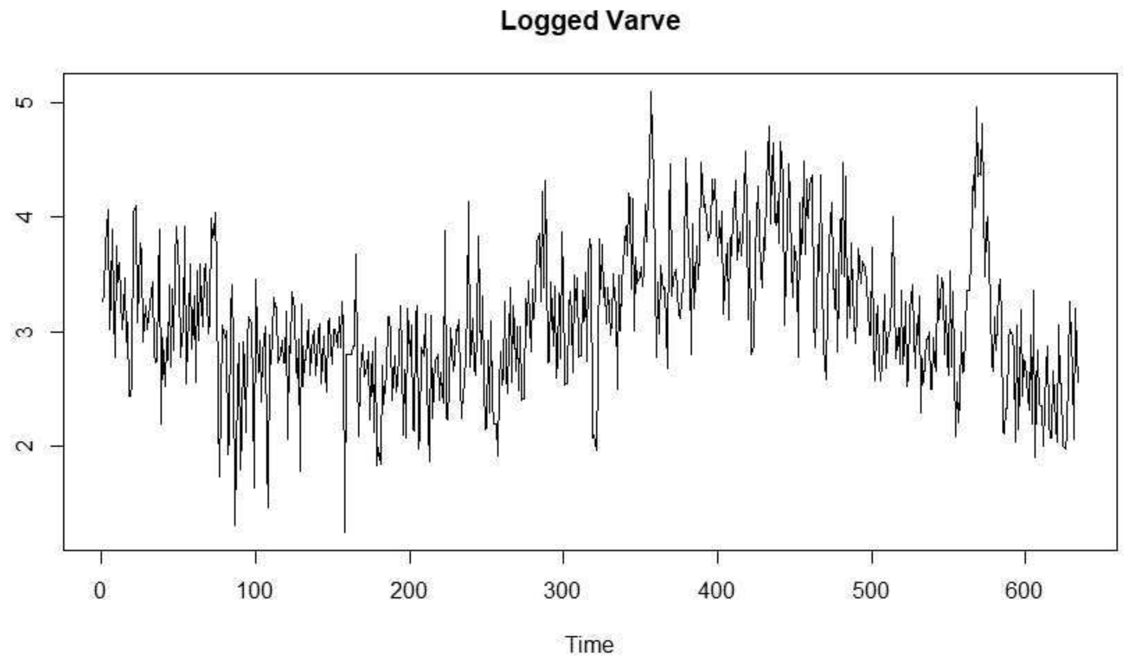
After transforming the data, the peak of the histogram becomes more centralised, like a normal distribution.



(b) Plot the series y_t. Do any time intervals, of the order 100 years, exist where one can observe behavior comparable to that observed in the global temperature records in Fig. 1.2?

```
g = globtemp
plot.ts(g, main="Global Temperature", ylab='')
plot.ts(y, main="Logged Varve", ylab='')
```

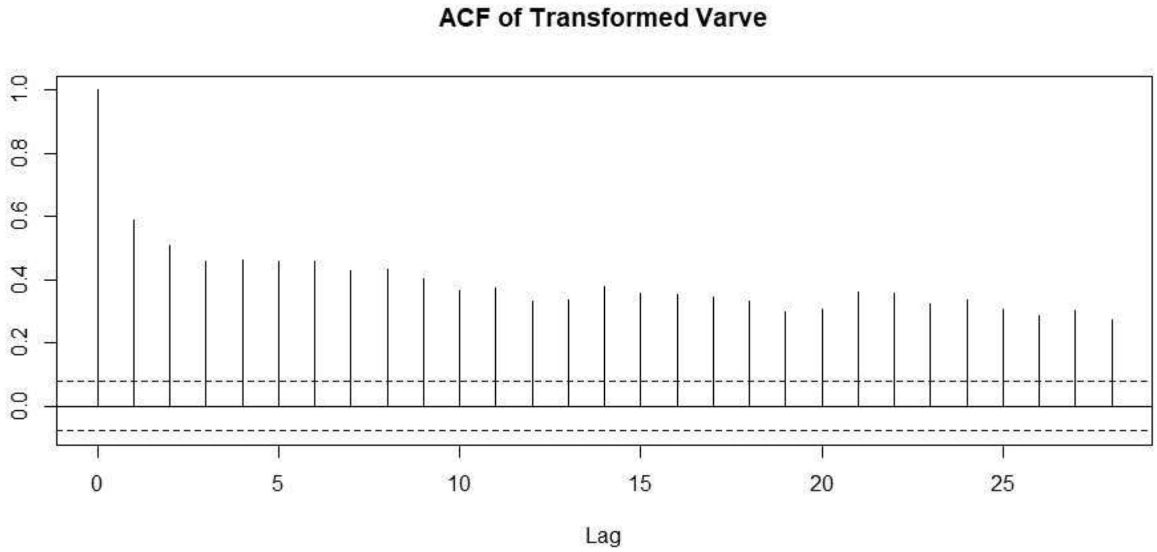


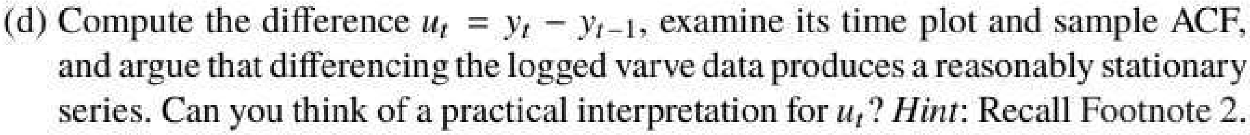


The data between 300 and 450 show a positive trend that is similar to the global temperature data.

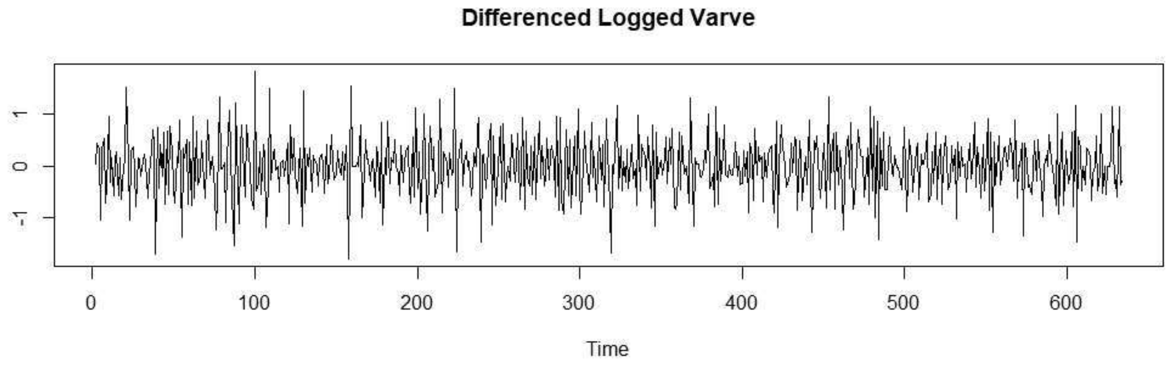
(c) Examine the sample ACF of y_t and comment.

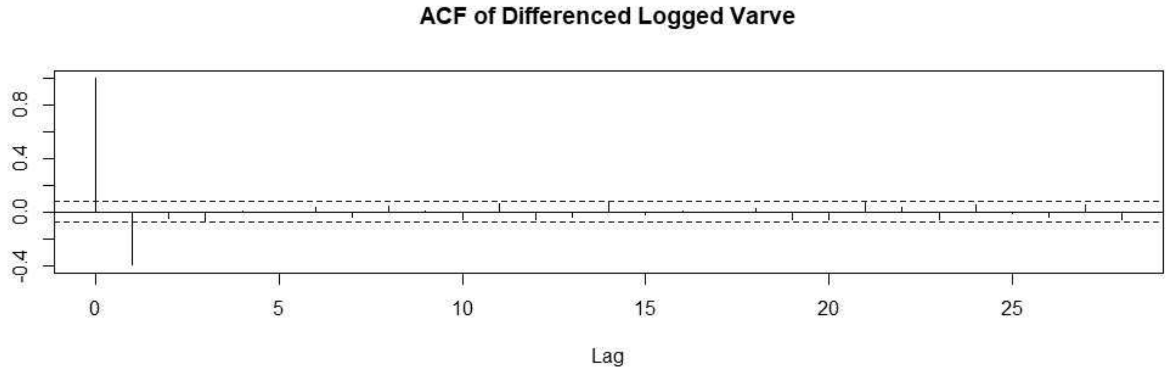
The ACF of the transformed data is showing significant correlation at every lag shown, however, it is declining slowly in a linear way.



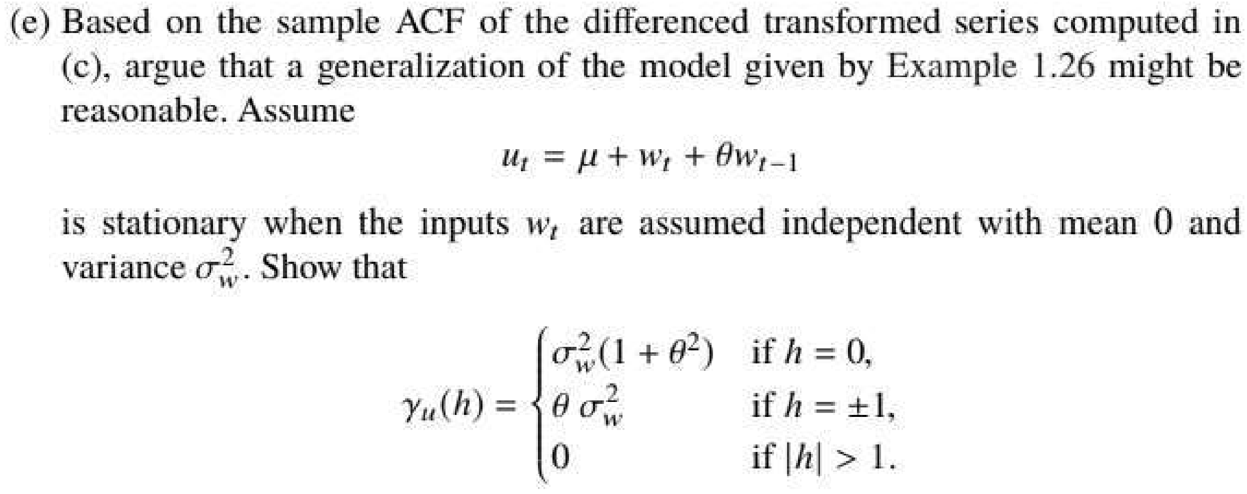

(d) Compute the difference $u_t = y_t - y_{t-1}$, examine its time plot and sample ACF, and argue that differencing the logged varve data produces a reasonably stationary series. Can you think of a practical interpretation for u_t? *Hint*: Recall Footnote 2.

```
d = diff(y)
plot.ts(d, main="Differenced Logged Varve", ylab='')
acf(d, main="ACF of Differenced Logged Varve", ylab='')
```





We see that the fully transformed is showing a relatively stable variance across the plot. The data points are also hovering around 0, which suggests that it has a constant mean. The ACF is only showing significant correlation at lag 1.



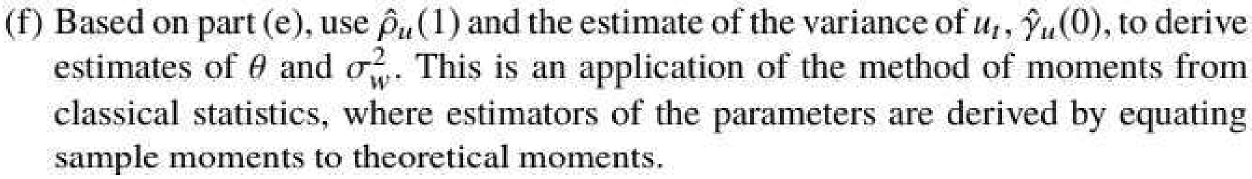
(e) Based on the sample ACF of the differenced transformed series computed in (c), argue that a generalization of the model given by Example 1.26 might be reasonable. Assume

$$u_t = \mu + w_t + \theta w_{t-1}$$

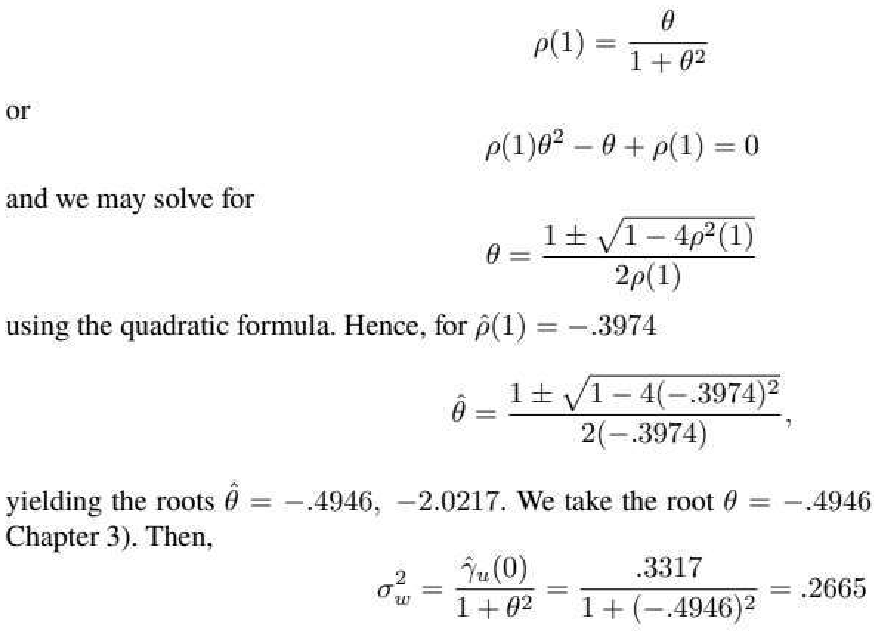
is stationary when the inputs w_t are assumed independent with mean 0 and variance σ_w^2. Show that

$$\gamma_u(h) = \begin{cases} \sigma_w^2(1+\theta^2) & \text{if } h = 0, \\ \theta\, \sigma_w^2 & \text{if } h = \pm 1, \\ 0 & \text{if } |h| > 1. \end{cases}$$

The ACF of the generalised model only has a significant lag at h=1, which is what we observe for the sample ACF.



(f) Based on part (e), use $\hat{\rho}_u(1)$ and the estimate of the variance of u_t, $\hat{\gamma}_u(0)$, to derive estimates of θ and σ_w^2. This is an application of the method of moments from classical statistics, where estimators of the parameters are derived by equating sample moments to theoretical moments.



$$\rho(1) = \frac{\theta}{1+\theta^2}$$

or

$$\rho(1)\theta^2 - \theta + \rho(1) = 0$$

and we may solve for

$$\theta = \frac{1 \pm \sqrt{1-4\rho^2(1)}}{2\rho(1)}$$

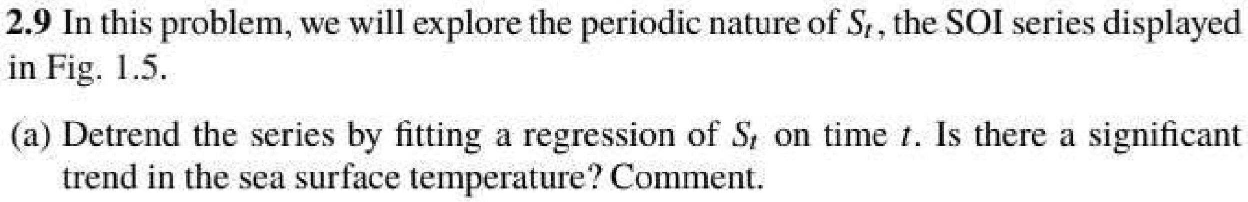
using the quadratic formula. Hence, for $\hat{\rho}(1) = -.3974$

$$\hat{\theta} = \frac{1 \pm \sqrt{1-4(-.3974)^2}}{2(-.3974)},$$

yielding the roots $\hat{\theta} = -.4946, \ -2.0217$. We take the root $\theta = -.4946$ (this is the invertible root, see Chapter 3). Then,

$$\sigma_w^2 = \frac{\hat{\gamma}_u(0)}{1+\theta^2} = \frac{.3317}{1+(-.4946)^2} = .2665$$

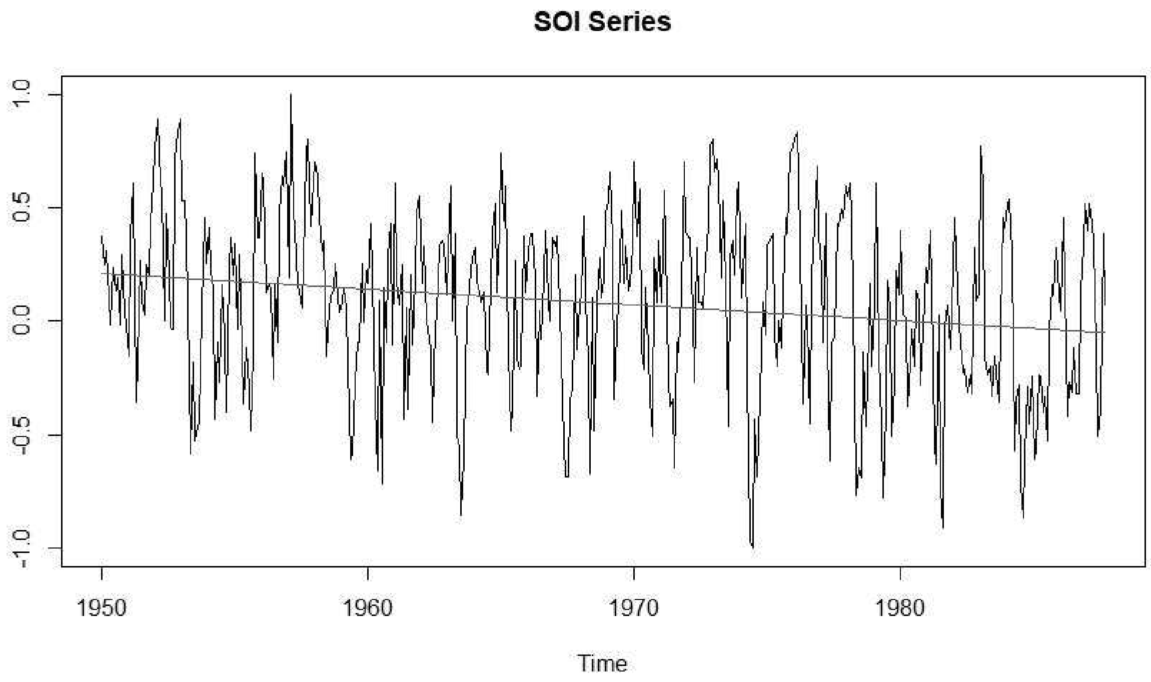
2.9

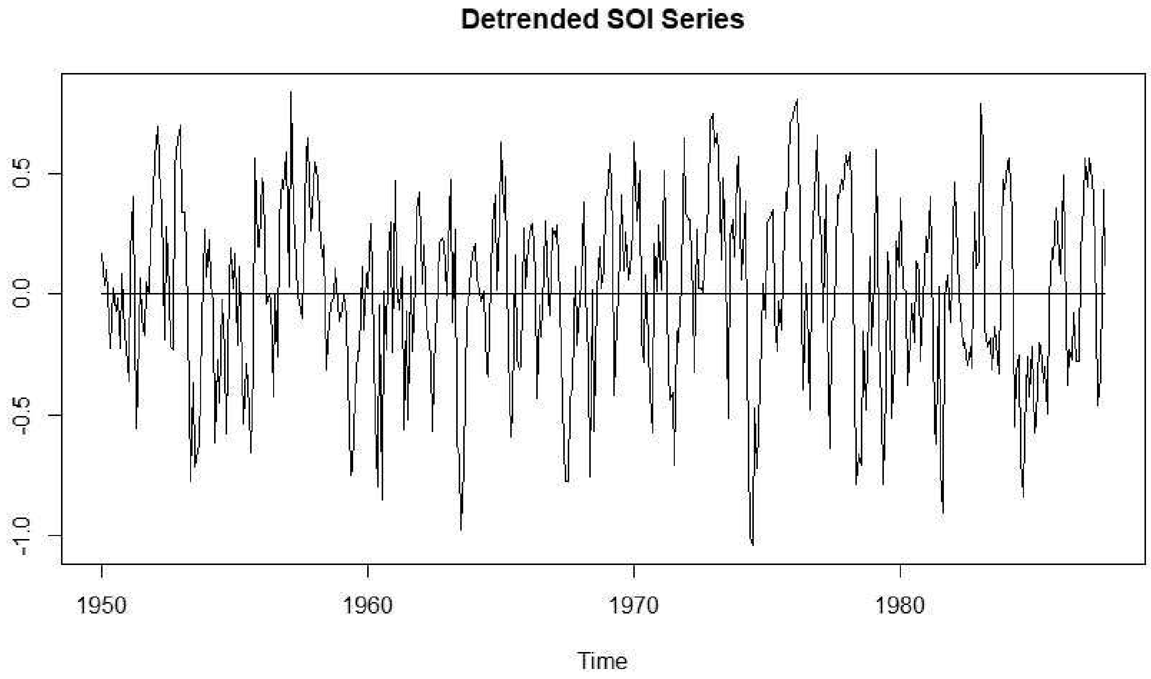


2.9 In this problem, we will explore the periodic nature of S_t, the SOI series displayed in Fig. 1.5.

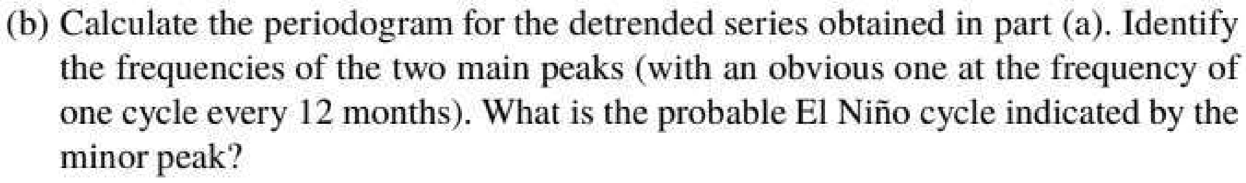
(a) Detrend the series by fitting a regression of S_t on time t. Is there a significant trend in the sea surface temperature? Comment.

```
par(mfrow=c(2,1))
plot.ts(soi, main="SOI Series", ylab='')
model <- lm(soi ~ time(soi), na.action=NULL)
lines(fitted(model), col=4)
soi.d = resid(model) # detrended = soi - fitted(model) also works
plot.ts(soi.d, main="Detrended SOI Series", ylab='')
model.2 <- lm(soi.d ~ time(soi.d), na.action=NULL)
lines(fitted(model.2, col=4))
```





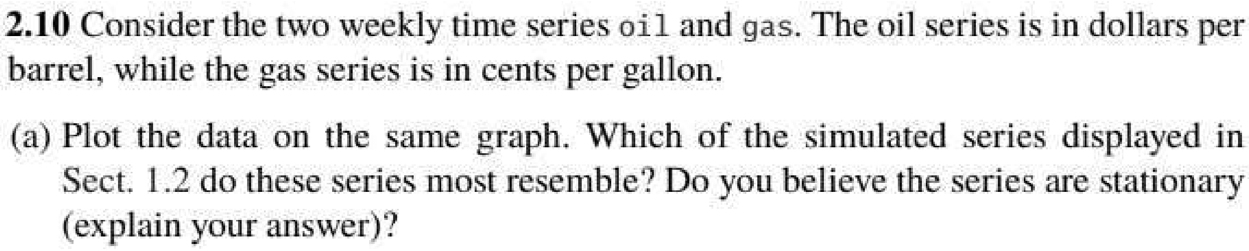
We see that the mean of the detrended series is constant over, compared to the original series, where we see that the mean is decreasing and non-constant.



(b) Calculate the periodogram for the detrended series obtained in part (a). Identify the frequencies of the two main peaks (with an obvious one at the frequency of one cycle every 12 months). What is the probable El Niño cycle indicated by the minor peak?

See solutions manual.

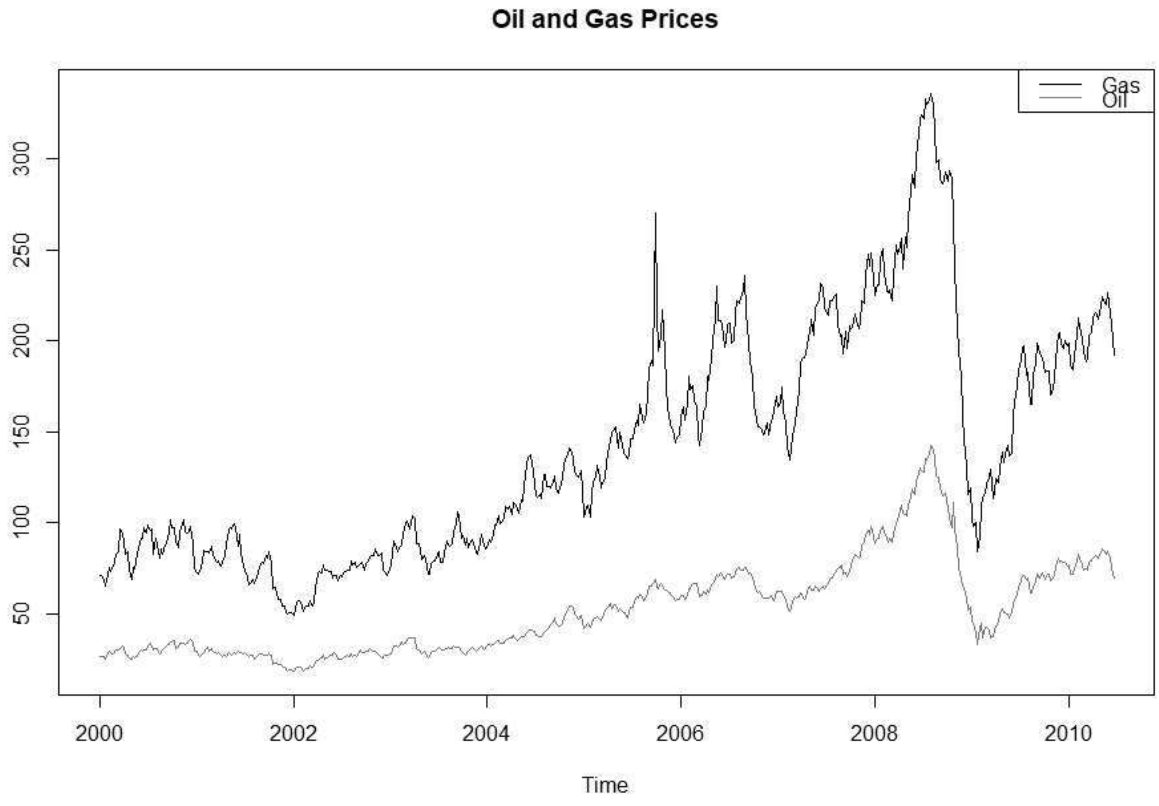
2.10



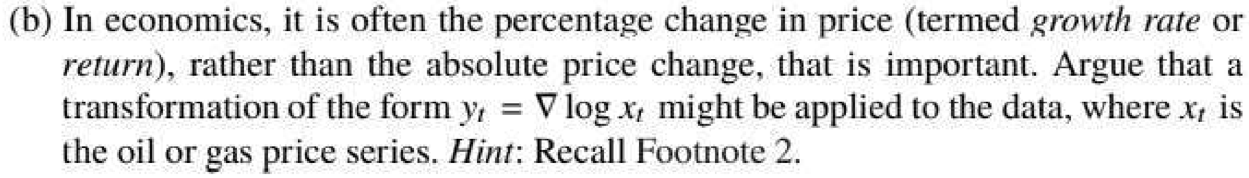
2.10 Consider the two weekly time series oil and gas. The oil series is in dollars per barrel, while the gas series is in cents per gallon.

(a) Plot the data on the same graph. Which of the simulated series displayed in Sect. 1.2 do these series most resemble? Do you believe the series are stationary (explain your answer)?

```
ts.plot(gas, oil, main="Oil and Gas Prices", col=1:2)
legend(x="topright", legend=c("Gas", "Oil"), lty=c(1,1), col=c(1,2))
```

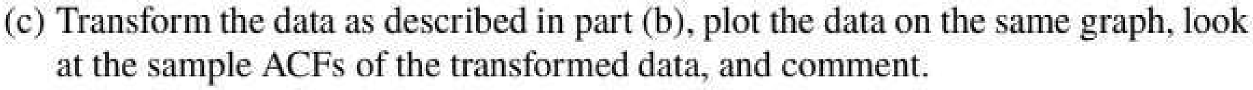


We see that the mean of both series is increasing, which implies non-stationarity for both. They look like the random walk with drift models and random walks are non-stationary.



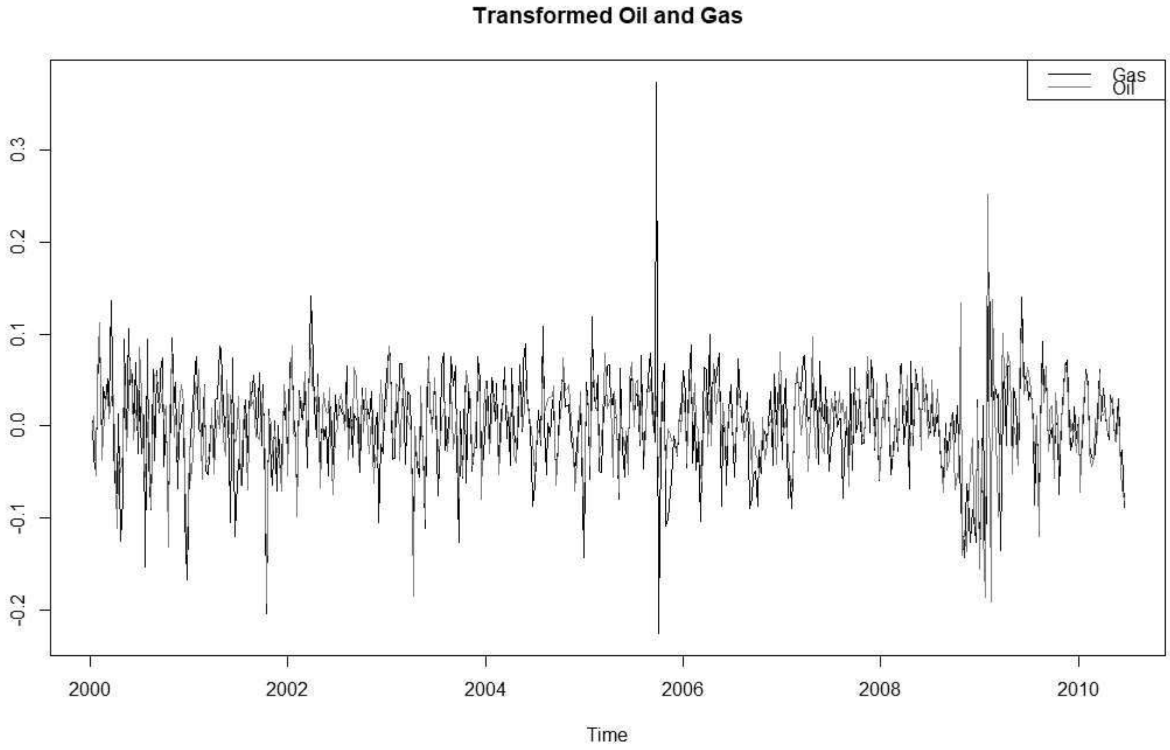
(b) In economics, it is often the percentage change in price (termed *growth rate* or *return*), rather than the absolute price change, that is important. Argue that a transformation of the form $y_t = \nabla \log x_t$ might be applied to the data, where x_t is the oil or gas price series. *Hint*: Recall Footnote 2.

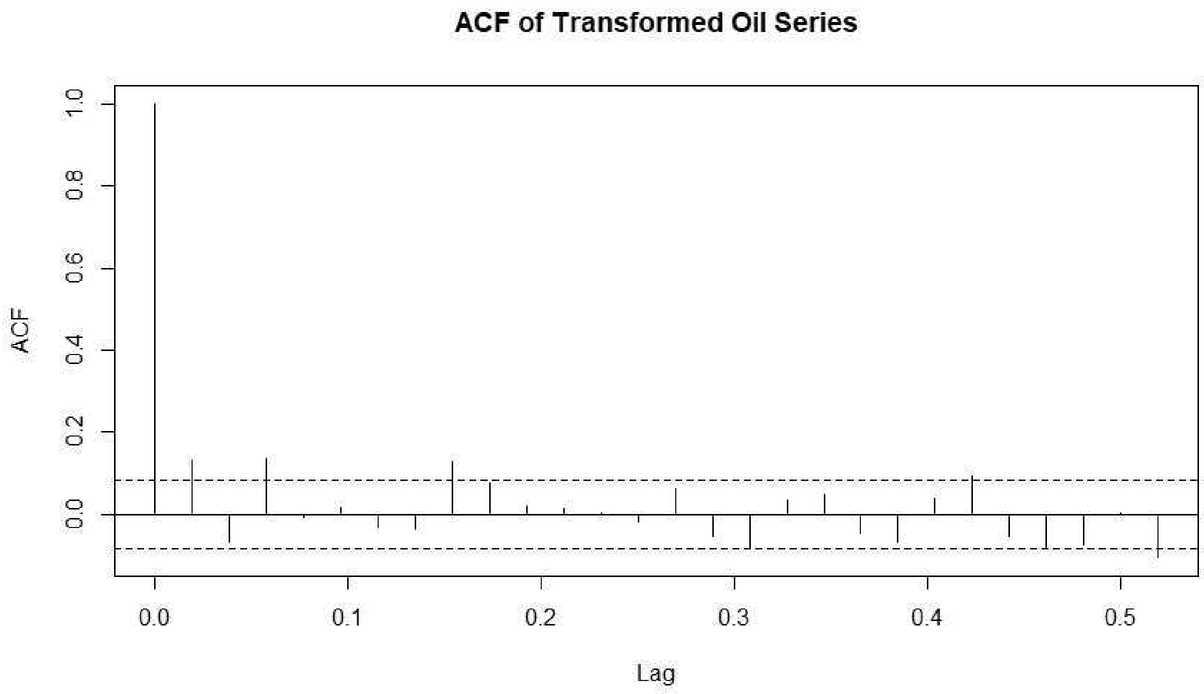
The percentage change in price can be approximated by a log of the quotient between the two price points. So, applying a delta log transformation would generate a series of returns.

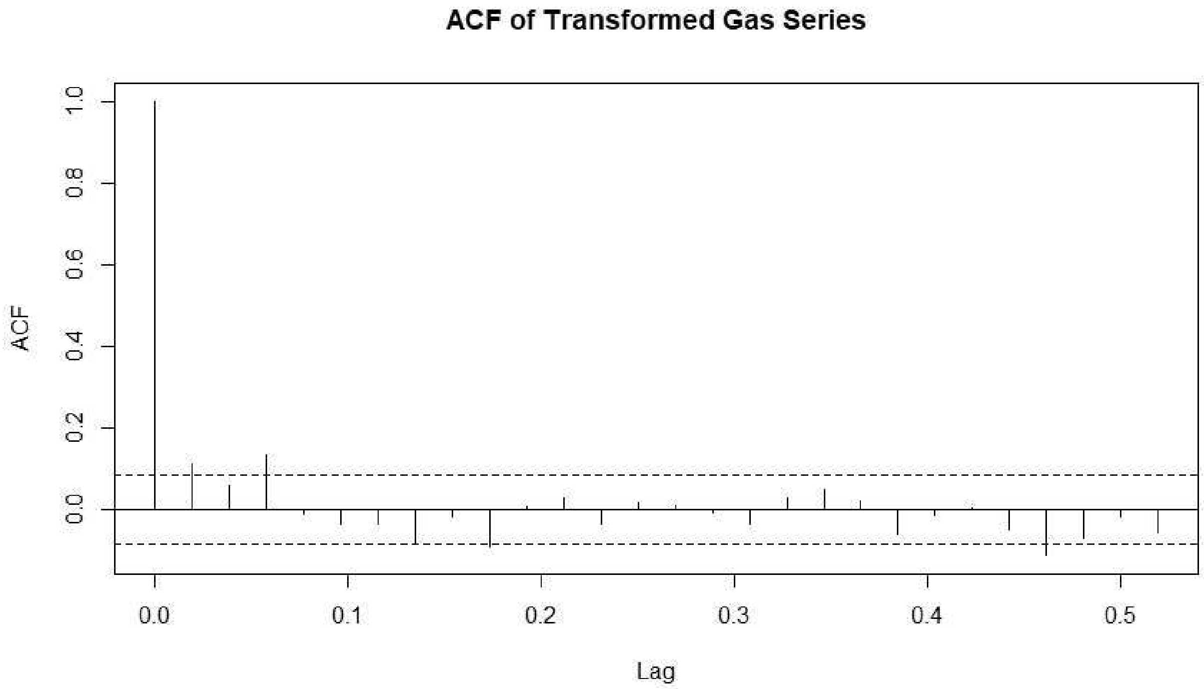


(c) Transform the data as described in part (b), plot the data on the same graph, look at the sample ACFs of the transformed data, and comment.

```
o.l = diff(log(oil)); g.l = diff(log(gas))
ts.plot(g.l, o.l, main="Transformed Oil and Gas", col=1:2)
legend(x="topright", legend=c("Gas", "Oil"), lty=c(1,1), col=c(1,2))
acf(o.l, main="ACF of Transformed Oil Series")
acf(g.l, main="ACF of Transformed Gas Series")
```



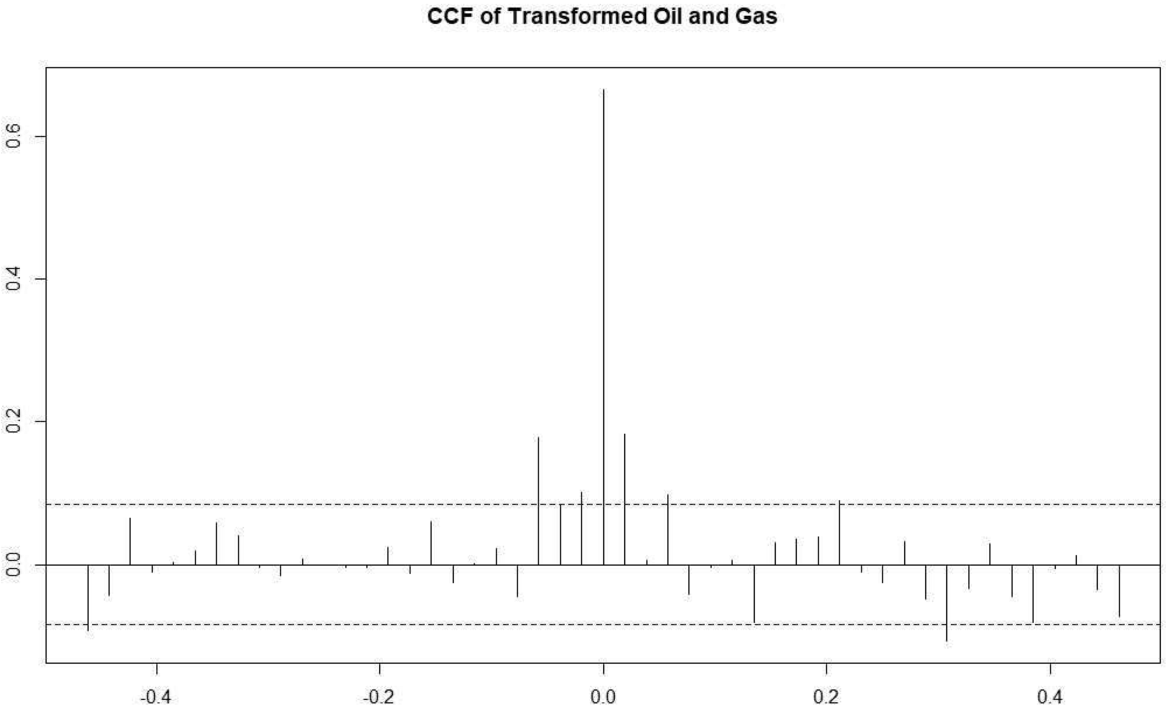




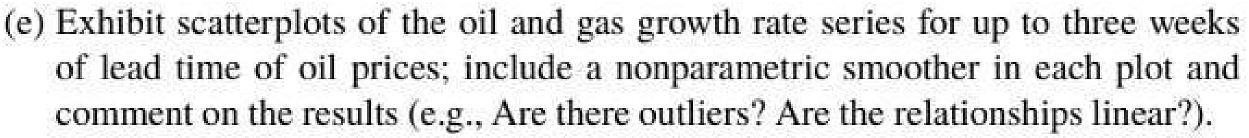
We see that the transformed data series display a constant mean and the respective ACFs only show significant correlation at a few lags.

(d) Plot the CCF of the transformed data and comment The small, but significant values when `gas` leads `oil` might be considered as feedback.

```
ccf(o.l, g.l, main="CCF of Transformed Oil and Gas", ylab='', xlab='')
```

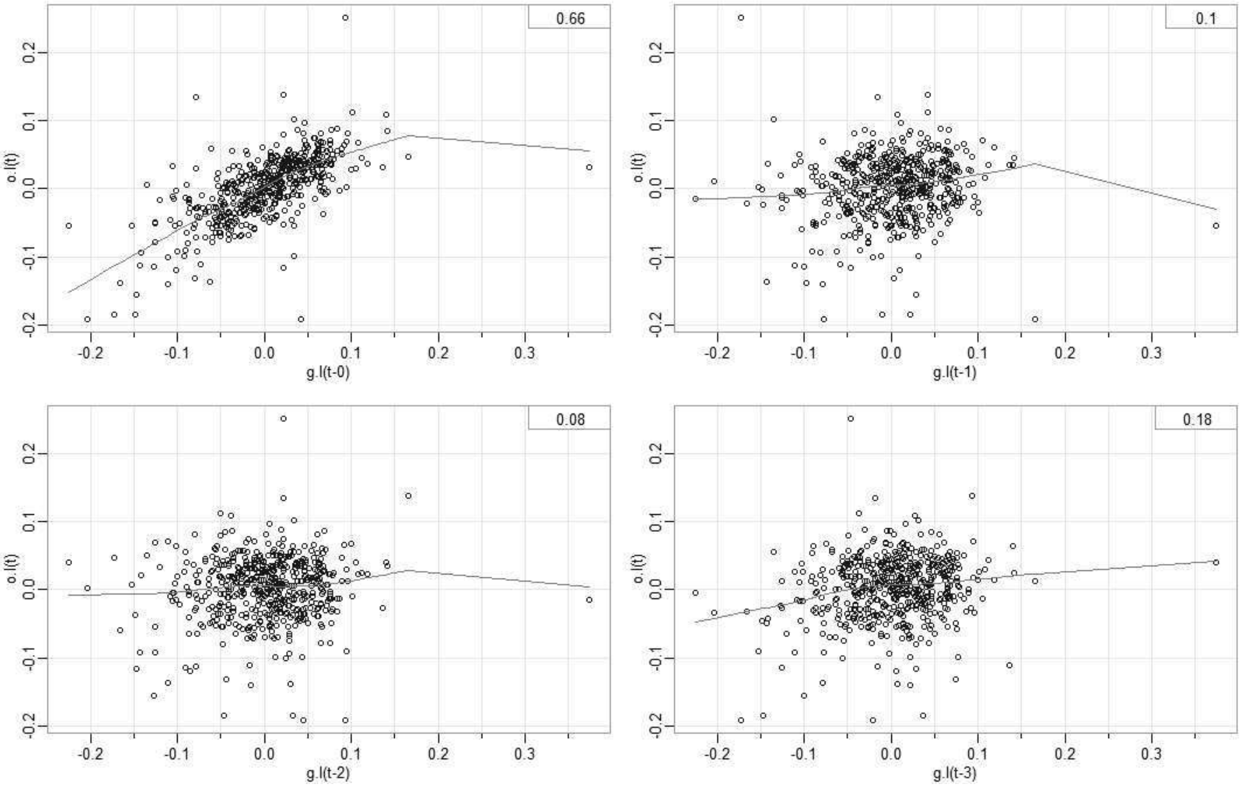


There is strong cross-correlation at lag 0, also oil one week ahead and feedback for gas three weeks ahead.

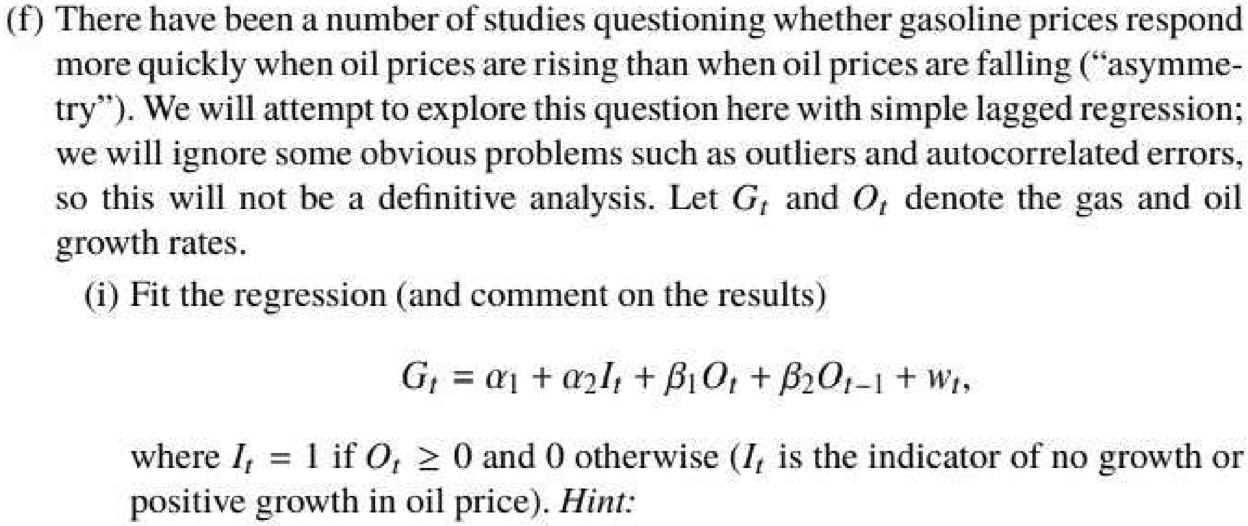


(e) Exhibit scatterplots of the oil and gas growth rate series for up to three weeks of lead time of oil prices; include a nonparametric smoother in each plot and comment on the results (e.g., Are there outliers? Are the relationships linear?).

```
lag2.plot(g.l, o.l, 3)
```



There are a few outliers, however, besides these, the data shows fairly linear relationships.



(f) There have been a number of studies questioning whether gasoline prices respond more quickly when oil prices are rising than when oil prices are falling ("asymmetry"). We will attempt to explore this question here with simple lagged regression; we will ignore some obvious problems such as outliers and autocorrelated errors, so this will not be a definitive analysis. Let G_t and O_t denote the gas and oil growth rates.

(i) Fit the regression (and comment on the results)

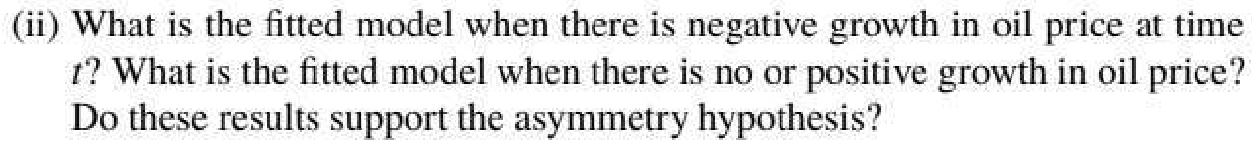
$$G_t = \alpha_1 + \alpha_2 I_t + \beta_1 O_t + \beta_2 O_{t-1} + w_t,$$

where $I_t = 1$ if $O_t \geq 0$ and 0 otherwise (I_t is the indicator of no growth or positive growth in oil price). *Hint:*

```
ind = ts(as.numeric(o.l > 0), start=c(2000,2), end=c(2010, 5), frequency=52)
# Can also use
# ind = ifelse(o.l < 0, 0, 1)
fish = ts.intersect(gas=g.l, oil=o.l, oilL=lag(o.l, -1), indic=ind, dframe=TRUE)
model <- lm(gas~indic+oil+oilL, data=fish)
```

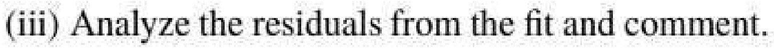
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.18447 -0.02318 -0.00033  0.02204  0.34352 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.006472   0.003582  -1.807  0.07136 .  
indic        0.012272   0.005704   2.151  0.03191 *  
oil          0.683728   0.059855  11.423  < 2e-16 ***
oilL         0.112463   0.039534   2.845  0.00462 ** 
```

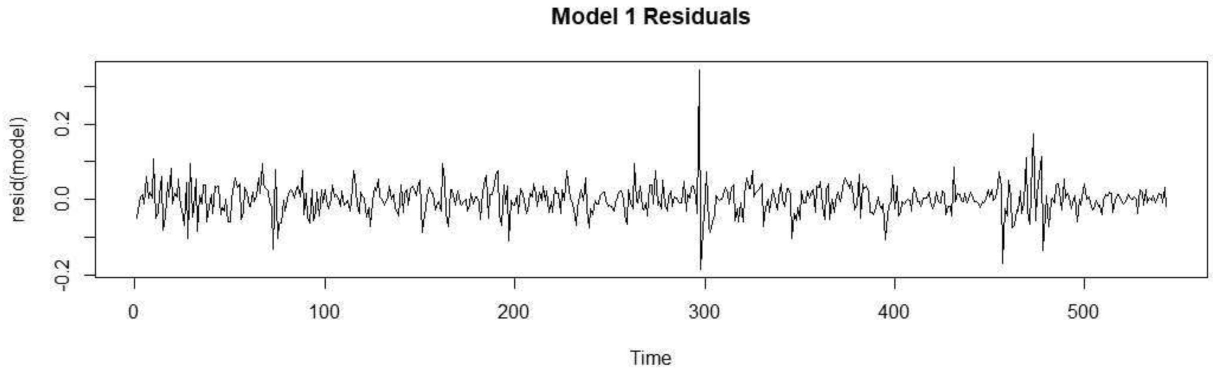
The indicator and lag parameters aren’t as significant as oil. In any case, we only see a slight increase in response from gas prices given increasing oil prices.

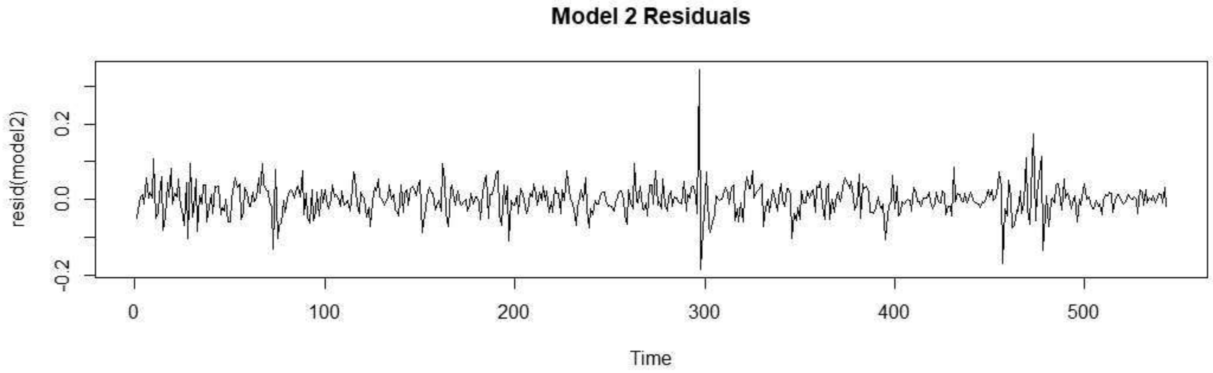

(ii) What is the fitted model when there is negative growth in oil price at time *t*? What is the fitted model when there is no or positive growth in oil price? Do these results support the asymmetry hypothesis?

```
ind2 = ifelse(o.l > 0, 0, 1)
fish2 = ts.intersect(gas=g.l, oil=o.l, oilL=lag(o.l, -1), indic2=ind2, dframe=TRUE)
model2 <- lm(gas~indic2+oil+oilL, data=fish2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18460 -0.02167 -0.00030  0.02176  0.34352 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.005675   0.003125   1.816  0.06994 .  
indic2      -0.011785   0.005514  -2.137  0.03303 *  
oil          0.687749   0.058380  11.781  < 2e-16 ***
oilL         0.112152   0.038570   2.908  0.00379 ** 
```

We see similar levels of significance for negative growth in oil prices, and a similar increase in responsiveness. Hence, we do not see any signs of asymmetry from the models.

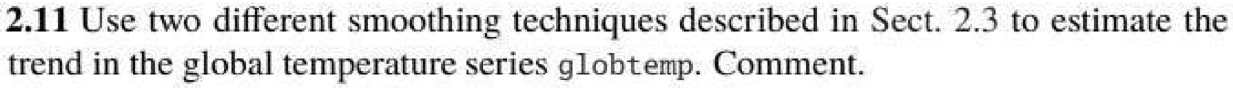

(iii) Analyze the residuals from the fit and comment.





From the residuals, we see that the models are pretty good, except for a few outliers.

2.11



2.11 Use two different smoothing techniques described in Sect. 2.3 to estimate the trend in the global temperature series `globtemp`. Comment.

```
plot.ts(globtemp, main="Global Temperature and Filters", ylab='')
weights = c(0.5, rep(1,11), 0.5)/12
ma = filter(globtemp, sides=2, filter=weights)
ks = ksmooth(time(globtemp), globtemp, "normal", bandwidth=5)
lines(ma, col=4, lwd=2)
lines(ks, lty=2, col=2, lwd=2)
legend(x="topright", legend=c("MA", "KS"), lty=1:2, lwd=rep(2,2), col=c(4,2))
```