Problem Sheet 2

1. Suppose that $\{Y_t\}$ is a stationary process with constant mean $\mu_t = \mathrm{E}(Y_t) = \mu$ and autocorrelation function ρ_k. Define

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

to be the sample mean of $Y_1, Y_2, \ldots, Y_n$.

   (a) Show that $\bar{Y}$ is an unbiased estimator of μ; that is, show that $\mathrm{E}(\bar{Y}) = \mu$. Note that this result holds regardless of the values of ρ_k.

   (b) find the expression for $\mathrm{var}(\bar{Y})$ as a function of λ_0, n, ρ_k, where $\lambda_0 = \mathrm{var}(Y_t)$.

   (c) If Y_t is a white noise process, what does $\mathrm{var}(\bar{Y})$ reduce to?

   (d) Suppose that Y_t is a MA(1) process; i.e.,$Y_t = e_t - \theta e_{t-1}$,where θ is a fixed parameter and e_t is a zero mean white noise process with $\mathrm{var}(e_t) = \sigma_e^2$. Use the result in part (b) to find an expression for $\mathrm{var}(\bar{Y})$.

2. Let $\{X_t\}$ be a zero mean stationary Gaussian process. we define the process $\{Y_t\}$:

$$Y_t = X_t X_{t-1}.$$

   (a) Determine the autocovariance function $\gamma_{k,Y}$ of the process Y_t in terms of $\gamma_{k,X}$ (the autocovariance fonction of X_t).

   we give the following formula:

$$\mathrm{E}(X_j X_k X_l X_m) = \mathrm{E}(X_j X_k)\mathrm{E}(X_l X_m) + \mathrm{E}(X_j X_l)\mathrm{E}(X_k X_m) + \mathrm{E}(X_j X_m)\mathrm{E}(X_k X_l)$$

   (b) If $\{X_t\}$ is an MA(1) process:

$$X_t = e_t - \theta_1 e_{t-1},$$

   give the form of $\gamma_{k,Y}$ in terms of θ_1 and σ_e^2.

3. Suppose that X_t is a stationary AR(1) process with parameter $\phi \neq 0$, that is,

$$X_t = \phi X_{t-1} + e_t,$$

where e_t is zero mean white noise with $\mathrm{var}(e_t) = \sigma_e^2$.

Define the two processes Y_t and W_t by

$$Y_t = \beta_0 + \beta_1 t + X_t,$$

$$W_t = Y_t - Y_{t-1}.$$

where β_0 and β_1 are fixed constants.

   (a) Find the mean and variance of Y_t. Is the Y_t process stationary?

   (b) Find the mean and variance of W_t. Is the W_t process stationary?

4. Consider the process $Y_t = \mu + X_t$, where $X_t = e_t - \theta e_{t-1}$ and $e_t \sim \mathrm{iid}N(0,1)$. Let $n = 100$.

   (a) Show that $\mathrm{E}(X_t) = 0$ and $\mathrm{var}(X_t) = \lambda_0 = 1 + \theta^2$.

   (b) Find the sampling distribution of Y .

   (c) Suppose that $\theta = -0.5$. Use the R code

```
> Y.t = 10 + arima.sim(list(order = c(0,0,1), ma = 0.5), n = 100)
```

to generate a realization of this process when $\mu = 10$. Compute a 99 percent confidence interval for μ (which in this problem is known to be 10). Does your interval include 10? You can use the `mean(Y.t)` command to compute the sample mean of $Y_1, Y_2, \dots, Y_{100}$.

(d) Repeat part (c) with $\theta = 0.5$. You will use the same command to simulate a new realization, except use `ma = -0.5`. R uses the convention of negating the parameter θ.

(e) Compare the confidence intervals in parts (c) and (d) in terms of interval length. Explain why one interval should be shorter.

5. The TSA library contains the data set `co2`, which lists monthly carbon dioxide levels in northern Canada from 1/1994 to 12/2004. To load the data in R, remember that you need to first type:

```
> library(TSA)
> data(co2)
```

(a) Construct a time series plot for these data. You can do this using the following R command:

```
> plot(co2, ylab="CO2 levels", xlab="Year", type="o")
```

Describe all systematic patterns you see in the plot.

(b) To enhance the usefulness of the plot, add monthly plotting symbols using the following R commands:

```
> plot(co2, ylab="CO2 levels", xlab="Year", type='l')
> points(y=co2, x=time(co2), pch=as.vector(season(co2)), cex=0.75)
```

Which months are consistently associated with the highest CO2 levels? The lowest?

(c) Consider fitting the simple straight line regression model (using the method of least squares):

$$Y_t = \beta_0 + \beta_1 t + e_t$$

to the data, where Y_t denotes the carbon dioxide level at time t. This model says that the CO2 level is a linear function of time. You can do this using the R commands:

```
> model = lm(co2 ~ time(co2))
> summary(model)
```

What are the least squares estimates of β_0 and β_1? Write out the equation of the least squares regression line. In classical linear regression, what are the "usual" assumptions for the error terms e_t?

(d) Now construct a plot which superimposes the least squares fit over the time series plot from part (b). You can do this using the R commands:

```
> plot(co2, ylab="CO2 levels", xlab="Year", type='l')
> points(y=co2, x=time(co2), pch=as.vector(season(co2)), cex=0.75)
> abline(model)
```

(e) As with any regression analysis, we can use the residuals to assess the quality of the model. Use the following R commands to create a plot of the residuals from the least squares fit versus time:

```
> plot(y=rstudent(model), x=as.vector(time(co2)), xlab="Year", ylab="Standardised residuals",
+ type='l')
> points(y=rstudent(model), x=as.vector(time(co2)), pch=as.vector(season(co2)), cex=0.75)
```

Comment on the adequacy of the straight line regression model in part (c). What other types of models might do a better job in capturing the systematic part of this time series?

(f) We will now fit the model

$$C_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi f t) + \beta_3 \sin(2\pi f t) + X_t,$$

where $\mathrm{E}(X_t) = 0$. This deterministic part $\mu_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi f t) + \beta_3 \sin(2\pi f t)$ contains both linear and trigonometric trend components. Note that there are 12 observations per year, so we take the frequency f = 1.

To fit the model, we can use the R commands:

```
> har. <- harmonic(co2,1)
> fit <- lm(co2~har. + time(co2))
> summary(fit)
```

What are the least squares estimates? Write out an equation for the fitted model. Is all of the R output relevant? Explain.

(g) Construct a graph which plots the points along with the fitted regression model superimposed. You can use the R commands:

```
> plot(ts(fitted(fit),freq=12,start=c(1994,1)),ylab="Monthly CO2 levels",
                          type='l',ylim=range(c(fitted(fit),co2)))
> points(co2)
```

Is this model appear to fit the data better than the straight line model? How would yourate the fit overall?

(h) Use the remaining code to perform the model diagnostics we discussed in class (Section3.5 in the notes).

```
> plot(rstudent(fit),ylab="Std.residuals",xlab="Year",type="o")
> abline(h=0)> hist(rstudent(fit),main="Hist. of std.residuals",xlab="Std.residuals")
> qqnorm(rstudent(fit),main="QQ plot of std.residuals")
> shapiro.test(rstudent(fit))> runs(rstudent(fit))
> acf(rstudent(fit),main="Sample ACF for std.residuals")
```

Interpret everything and give an overall assessment of the model we have fit in this problem. Do the residuals look to resemble a stationary white noise process? Or, is there still noticeable structure left in them?

6. Consider the ventilation data. Then, read the data into R.

(a) Fit a straight-line regression model to the data for detrending purposes.

(b) Do a thorough residual analysis of the detrended (residual) process using the techniques we discussed in class. Do the residuals from your straight-line model fit resemble a zero mean white noise process?

7. The data set prescrip, lists monthly prescription costs for the months August 1986 to March 1992. These data are from the State of New Jersey Prescription Drug Program and are the cost per prescription claim during this time period.

(a) Construct a time series plot for the data. Describe the appearance of the series.

(b) Use the R command

```
diff.1 <- diff(prescrip)
```

to calculate the first differences $\nabla Y_t = Y_t - Y_{t-1}$ and plot the first differences. Describe the appearance of this plot and how it compares with the plot of the original series.

(c) Use all of the model diagnostic checks on the difference process ∇Y_t. Do the data differences resemble a normal zero mean white noise process?