STAT 520, Fall 2015: Homework 3

1. Let $\{Y_t\}$ be a stationary process with constant mean $E(Y_t) = \mu$ and autocorrelation function ρ_k. Define $\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ to be the sample mean of $Y_1, Y_2, \ldots, Y_n$. Recall that

$$\text{var}(\overline{Y}) = \frac{\gamma_0}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right],$$

where $\gamma_0 = \text{var}(Y_t)$.

   (a) Find $\text{var}(\overline{Y})$ when $\{Y_t\}$ is white noise, i.e. independent and identically distributed.

   (b) Find $\text{var}(\overline{Y})$ when $\{Y_t\}$ is MA(1) with parameter θ. Hint: see p. 82 in Tebbs' notes.

   (c) Find $\text{var}(\overline{Y})$ when $\{Y_t\}$ is AR(1) with parameter ϕ. Hint: see p. 90 in Tebbs' notes. You do not need to simplify this.

   For parts (b) and (c) determine whether $\text{var}(\overline{Y})$ is smaller or larger (i.e. better or worse) than the white noise case for MA(1) with $\theta = -0.5$ and $\theta = 0.5$, and AR(1) with $\phi = 0.5$ and $\phi = -0.5$; use $n = 5$ for both. Show your calculations.

2. Consider the process $Y_t = \mu + X_t$, where $X_t = e_t - \theta e_{t-1}$, i.e. $\{X_t\}$ is MA(1), and $e_t \sim$ iid $\mathcal{N}(0, 1)$. Let $n = 100$.

   (a) Show that $E(X_t) = 0$ and $\text{var}(X_t) = \gamma_0 = 1 + \theta^2$. Because linear combinations of normal random variables are normally distributed, it follows immediately that $X_t \sim \mathcal{N}(0, 1 + \theta^2)$ and $Y_t \sim \mathcal{N}(\mu, 1 + \theta^2)$.

   (b) Find the sampling distribution of $\overline{Y}$.

   (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? If it does not, you probably did something wrong. You can use the `mean(Y.t)` command to compute the sample mean of $Y_1, Y_2, \ldots, Y_{100}$.

   (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, as we will see in Chapter 4.

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

3. The TSA library contains the data set `co2`, which lists monthly carbon dioxide levels in northern Canada from 1/1994 to 12/2004. In this problem, we will use R to fit the model

$$\text{CO2}_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi f t) + \beta_3 \sin(2\pi f t) + X_t,$$

where $E(X_t) = 0$. The deterministic part of the model

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

contains both linear and trigonometric trend components. Note that there are 12 observations per year, so we take the frequency $f = 1$.

(a) 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.

(b) 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)
```

How would you rate the fit overall?

(c) Use the remaining code to perform the model diagnostics we discussed in class (Section 3.5 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?

4. The TSA library contains the data set `prescrip`, which 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 (Section 3.5 in the notes) on the difference process $\{\nabla Y_t\}$. Do the data differences resemble a normal zero mean white noise process?

5. Tuberculosis, commonly known as TB, is a bacterial infection that can spread through the lymph nodes and bloodstream to any organ in your body (it is most often found in the lungs). Most people who are exposed to TB never develop symptoms, because the bacteria can live in an inactive form in the body. But if the immune system weakens, such as in people with HIV or in elderly adults, TB bacteria can become active and fatal if untreated. The numbers of TB cases (per month) in the United States from January 2000 to December 2009 are catalogued in the data file `tb`, read into R via

```
tb=ts(read.table("http://people.stat.sc.edu/hansont/stat520/tb.txt"))
```

(a) Use the regression methods in Chapter 3 to "detrend" the data, that is, fit a deterministic trend model of the form

$$Y_t = \mu_t + X_t,$$

where μ_t is a deterministic trend function and $E(X_t) = 0$. Obviously, you are to examine the data and identify a plausible trend function. Produce a plot that displays the data with your fitted model superimposed over them (like in the notes). Hint: there may be more than one type of trend, e.g. both seasonal and polynomial. Recall that we discussed two types of seasonal trends.

(b) Examine the standardized residuals $\{\widehat{X}_t^*\}$ from your fitted model for normality and independence. What are your conclusions? Do the standardized residuals look to resemble a normal, zero mean white noise process?

(c) Display the sample ACF for the standardized residuals in part

(d) What type of stationary model would be useful in modeling the standardized residual process?

Cryer and Chan problems: 4.1, 4.2 (use R function `ARMAacf`; see R code for Chapter 4), 4.5abc. For 4.5abc only obtain a plot of the autocorrelation function in R; don't worry about the roots, etc.