Final Exam

ST565 Winter 2014

Thursday 20th March 2014

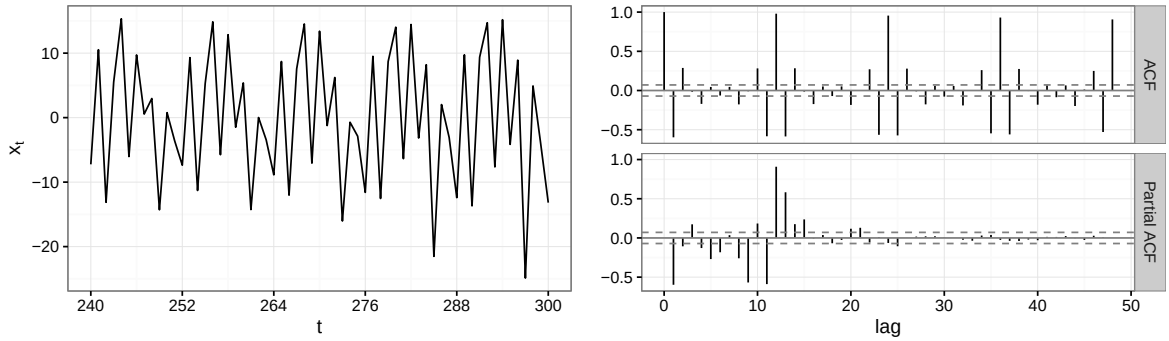
Name: _____

- You have 110 minutes to complete the exam.
- There are 4 questions, answer all of the questions.

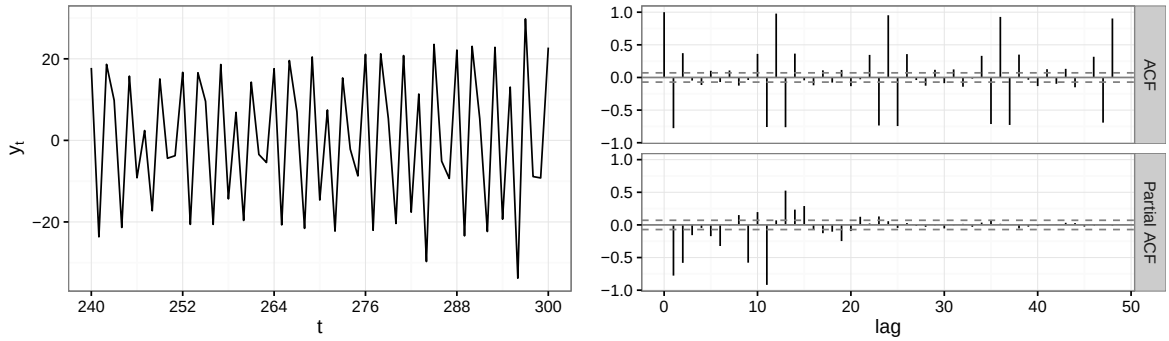
Question	Points	Out of
1		5
2		5
3		5
4		5
Total		20

1. The following plots show time series plots, and ACF/PACF plots for a single **monthly** time series: undifferenced (x_t), differenced once ($y_t = x_t - x_{t-1}$), and differenced seasonally ($v_t = x_t - x_{t-12}$). **Using the plots, suggest a SARIMA model for the data, x_t.** Include your reasoning.

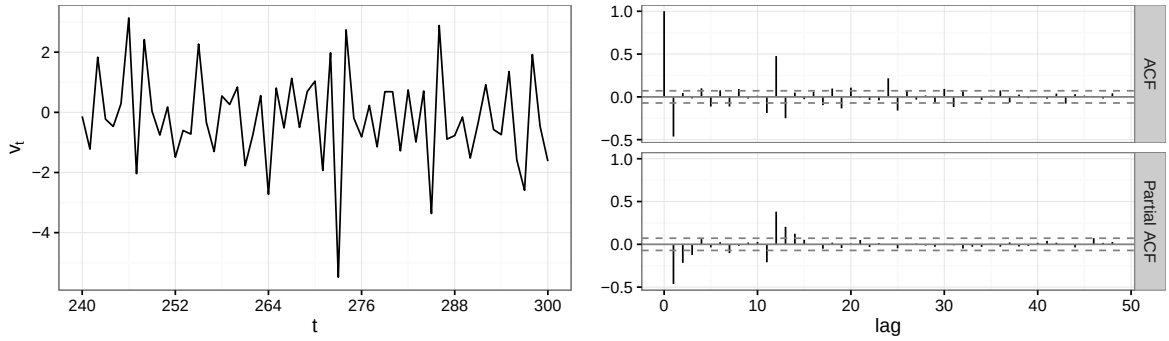
   Undifferenced, x_t



   Differenced once, $y_t = x_t - x_{t-1}$



   Differenced seasonally, $v_t = x_t - x_{t-12}$



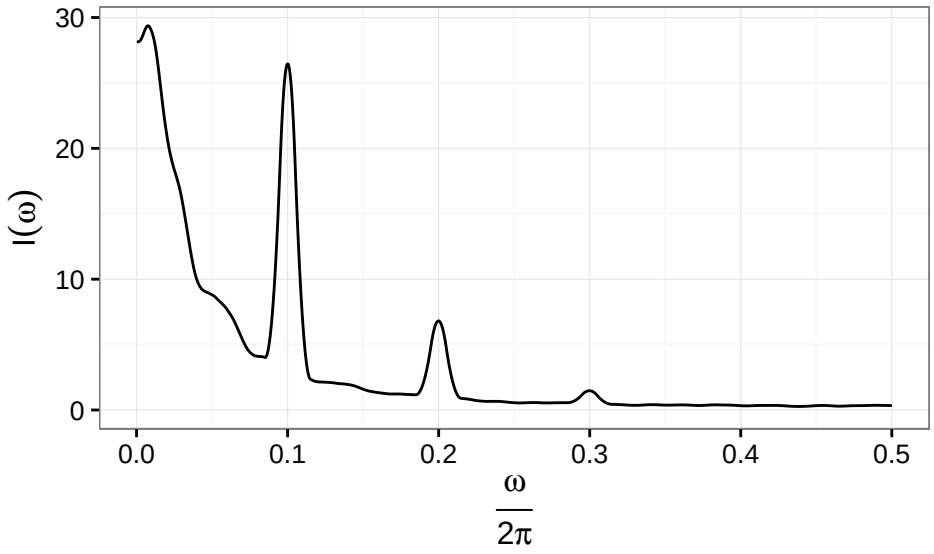
Solution:

- Differencing once seasonally appears to give a stationary series, therefore D = 1, d = 0.

- Correlation at seasonal lags (12, 24, 36) seems to be decaying slowly in the ACF and cutting off after one seasonal lag in the PACF, suggesting P = 1, Q = 0
- Correlation at non-seasonal lags (1, 2, ...) seems to be decaying slowly in the PACF and cutting off after lag 1 in the ACF, suggesting p = 0, q = 1

SARIMA$(0, 0, 1) \times (1, 1, 0)_{12}$

2. The following is a periodogram from an observed time series x_t, $t = 1, \ldots, 5000$. **Suggest a model for x_t, and describe what features you might see in a time series plot of x_t.**



Solution: The periodogram shows distinct peaks at 0.1, 0.2 and 0.3. This suggests a dominant periodic component of frequency 0.1 and two harmonics.

Ignoring the peaks, the periodogram is higher at lower frequencies than at higher frequencies, suggesting some positive autocorrelation.

A tentative model might be:

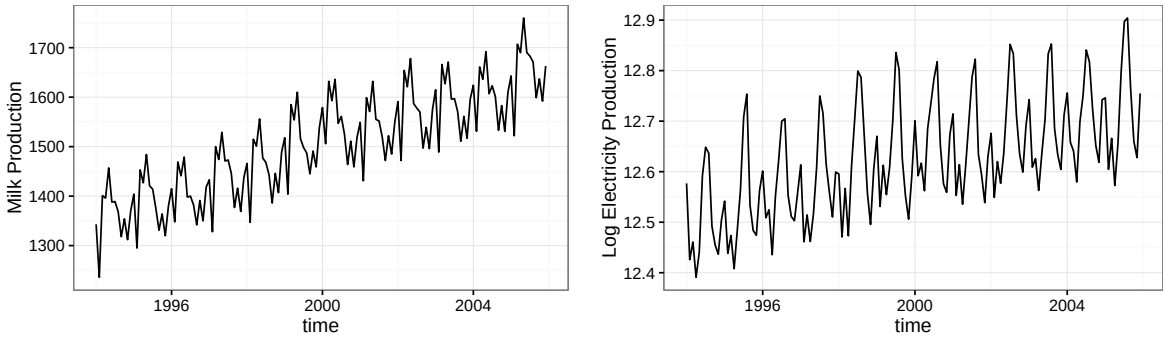
$$\begin{aligned} x_t = & a_1 cos(0.2\pi t) + b_1 sin(0.2\pi t) + \\ & a_2 cos(0.4\pi t) + b_2 sin(0.4\pi t) + \\ & a_3 cos(0.6\pi t) + b_3 sin(0.6\pi t) + z_t \end{aligned}$$

where z_t is a stationary time series process with **positive autocorrelation**, maybe an AR(1) with parameter $0 < \alpha < 1$.

Examining the time series plot of x_t should reveal a periodic component the cycles once every 10 (1/0.1) time steps. The presence of harmonics indicates the periodic component is not exactly sinusoidal, so this may be an asymmetric pattern, the fact there are only three harmonics suggests it is a relatively smooth pattern.

In addition the variation about this periodic component would appear correlated.

3. Below are time series plots of milk production per cow and log electricity production in the United States from 1994 to 2005.



A lactose intolerant researcher is interested in the relationship between the two and fits a regression of log electricity production on milk production per cow and gets the following output from R:

```
summary(lm(log_elec ~ milk))

##
## Call:
## lm(formula = log_elec ~ milk)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17386 -0.06666 -0.01961  0.06003  0.21276
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.175e+01  1.136e-01 103.441  < 2e-16 ***
## milk        5.812e-04  7.534e-05   7.714 1.95e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09481 on 142 degrees of freedom
## Multiple R-squared:  0.2953,Adjusted R-squared:  0.2904
## F-statistic: 59.51 on 1 and 142 DF,  p-value: 1.949e-12

exp( 100 * 5.812e-04) * 100

## [1] 105.9842
```

They conclude:

> There is extremely convincing evidence that milk production is associated with electricity production. A one hundred unit increase in milk production per cow is associated with a 6% increase in median electricity production. To reduce electricity needs we should just drink less milk!

What is wrong with the researcher's approach? What would you suggest they do?

Solution: Both log energy production and milk production are non-stationary (both have clear seasonality, and increasing trend). So, this is probably a case of "spurious correlation": when the response and explanatory variable are non-stationary time series in a regression, Type I error is greatly inflated, and we tend to declare a significant correlation too often when none really exists.

The series appear correlated, but that is really just picking up the fact that both are highly seasonal and trending in the same direction.

The researcher should transform both series to be stationary first, here differencing once seasonally and once non-seasonally should work, then perform the regression. Spurious correlation can also occur when the series are auto-correlated so he should also check the residuals of the regression for correlation by examining their ACF and PACF plots. If there does appear to be correlation in the residuals we should use a correlated errors model.

After differencing and modelling any correlation in the errors, the researcher can be more confident a significant p-value is a true correlation, not a spurious one. Although they still need to beware making causal conclusions based on observational data.

4. You have been put in charge of collecting air samples at a remote forest site in the Cascades. Each day you drive two hours from your accommodation to the site and take samples at noon. You can save a lot of time calibrating the sampling equipment if you know the temperature at the site before you leave your accommodation, however the weather station at the site isn't wireless, so your only option is to guess what the temperature is given what you know before you leave.

   At the time you leave your accommodation, you know the temperature at your accommodation today, and the temperature of the site yesterday as well as past values for both. **Discuss how you might go about building a model to forecast**:

   - today's temperature at the site,
   - tomorrow's temperature at the site, and
   - the temperature at the site on Jan 27 2015

   Mention what ideas you might explore and how you would explore them, as well as how your approach (or forecasts) might differ between the three time horizons.

Solution: This is a very open ended question! There are many "right" answers.

Considerations you probably should address:

- temperature is highly seasonal, knowing tomorrow is "March 20" should give me some information on what temperature to predict
- temperature tends to be autocorrelated, today's deviation from the usual March 20 temperature should help predict tomorrow's deviation from usual March 21 temperature
- my two locations aren't that far away, so there should be some correlation in their temperature. (Although if the weather tends move W to E, and my site is W of my accommodation, knowing the temperature at my accommodation might not help predict the temperature at the site)

Let x_t be today's temperature at my accommodation and y_t be today's temperature at the site.

I would start by exploring the seasonal pattern in both x_t and y_t with the aim of modelling and temporarily removing it from both series. This serves two purposes, I have a decent model to use to make long term forecasts (i.e. y_t on Jan 27 2015) and I can explore the correlation structure without the distraction of the seasonal patterns. Let $s_{x,t}$ and $s_{y,t}$ be the seasonality at time t in series x and y respectively.

I would then examine auto and cross correlation plots of both x_t and y_t with seasonality removed. I would try to determine whether short term forecasts (y_t or y_{t-1}) are going to be better made by using y_{t-1} or x_t, e.g. is the correlation stronger between the

temperature today at my accommodation and the site, or between yesterday's temperature at the site and today's temperature at the site. Essentially considering models of the form:

$$y_t = s_{y,t} + \alpha(y_{t-1} - s_{y,t-1}) + \beta(x_t - s_{x,t}) + z_t$$

where z_t is possibly a time series process. Bigger lags might be indicated in the correlation plots, non-linearity might be suggested in scatter plots of $x_t - s_{x,t}$ against $y_t - s_{y,t}$ etc.

I should have a way to pick the "best" forecasting model (or at least convince myself my model does better than just saying $\hat{y}_t = y_{t-1}$). One way might be to use say last year's data to fit the model, then see how well they do predicting this year's temperatures.

There might be trend (i.e. climate change) to worry about.