• Terms to verify that you know from previous classes (we won’t be covering unless you ask)
  • Random Variable, Mean, Expected Value, Variance, Standard Deviation
  • Population
  • Sample, sample mean, sample variance, sample standard deviation
  • Normal, T, Chi-square, F
  • Correlation, regression, residuals, SSE (sum of squared error = sum of squared residuals)
  • MSE (SE of regression)

• Terms that you may or may not be familiar with from previous classes (which we will be covering as we need them)
  • Covariance
  • Conditional Distribution, Conditional mean, conditional variance
  • Likelihood, log-likelihood, maximum likelihood
  • AIC, BIC (SIC)

• Homework assignment (due Tue, Jan 12)

I. Some conceptual/theoretical problems

1. Provide brief definitions or explanations for each term in the above “verify that you know” list.
2. Provide a brief, plain English, explanation of at least two of the terms on the “terms that you may or may not be familiar with” list (an explanation of what the thing is or how it is used is sufficient).

Not to be graded ... but possibly credit/no-credit for doing or not doing this.

II. Some conceptual/theoretical problems

1. If \( y_t = \beta_0 + \beta_1 t + \varepsilon_t \) and the \( \varepsilon_t \) are independent of each other mean zero random variables, explain why a first difference (i.e. \( y_t - y_{t-1} \)) would have no trend. [Hint: \( E(\varepsilon_t) = 0 \).]

Here we see that:

\[
y_t - y_{t-1} = (\beta_0 + \beta_1 t + \varepsilon_t) - (\beta_0 + \beta_1(t-1) + \varepsilon_{t-1}) \\
= \beta_1 + \varepsilon_t - \varepsilon_{t-1}.
\]

Thus \( E(y_t - y_{t-1}) = \beta_1 \). Interestingly enough, the mean of the first difference is the slope in the original trend.

2. If we have a linear trend plus seasonality model: \( y_t = \beta_0 + \beta_1 t + \beta_{Jan} I_{Jan,t} + \beta_{Feb} I_{Feb,t} + \cdots \beta_{Nov} I_{Nov,t} + \varepsilon_t \) and the \( \varepsilon_t \) are independent of each other mean zero random variables and \( I_{Month,t} \) is a month indicator, show that a first difference would still have seasonality.

Consider a time \( t \) which corresponds to a February. For such a time,

\[
y_t - y_{t-1} = (\beta_0 + \beta_1 t + \beta_{Feb} + \varepsilon_t) - (\beta_0 + \beta_1(t-1) + \beta_{Jan} + \varepsilon_{t-1}) \\
= \beta_1 + \beta_{Feb} - \beta_{Jan} + \varepsilon_t - \varepsilon_{t-1},
\]

and even in expectation, we have

\[
E(y_t - y_{t-1}) = \beta_1 + \beta_{Feb} - \beta_{Jan}.
\]
Similarly across the whole year, unless all seasonal effects (the monthly betas) are identical and zero, there would then be seasonality that cannot be eliminated by a 1st difference.

3. With the linear trend plus seasonality model, show that a seasonal difference (i.e. $y_t - y_{t-12}$) would have neither trend nor seasonality.

Again, consider a time $t$ which corresponds to a February. For such a time,

$$y_t - y_{t-12} = (\beta_0 + \beta_1 t + \beta_{Feb} + \epsilon_t) - (\beta_0 + \beta_1 (t - 12) + \beta_{Feb} + \epsilon_{t-1})$$

$$= 12\beta_1 + \epsilon_t - \epsilon_{t-12}$$

where in expectation, we have

$$E(y_t - y_{t-12}) = 12\beta_1.$$

Thus, via a seasonal difference, we can eliminate both a linear trend and seasonality. The seasonal difference also is related to the linear trend; the annual slope (the average change for a 1 year increase in time) is that seasonal difference.

4. In the formulation of the model in part (2), what is $\beta_0$? Use plain English.

Literally, the intercept is the average value of $y$ when all $x$'s are zero. In this case it would be the typical value for the December before we started collecting data. [Note: it is convenient that $t = 0$ happened to correspond to a December in our theoretical model.]

5. In the formulation of the model in part (2), what is $\beta_1$? Use plain English.

The slope is the average change in $y$ for a 1-unit increase in $t$, having controlled for the other $x$-variables being held as fixed. The problem here, is that from month-to-month there are predictable seasonal changes because the other $x$-variable will NOT be unchanged when there is a one month increase in time. This suggests we should go with one of:

A – the slope is the average monthly increase in the seasonally adjusted $y$ values.

Or

B – the slope is $1/12$ of the annual change in the $y$ values.

6. In the formulation of the model in part (2), what is $\beta_0 + \beta_1 t$? Use plain English.

As the monthly indicator that was excluded from the model was the December indicator, we would interpret $\beta_0 + \beta_1 t$ as the December trendline.
III. Some data oriented problems

1. For Beer production (R code to read this in is provided):
   
a. Plot time series and describe the dataset.

   This series has an increasing trend until about 1975 or 1980 and is relatively level since then. There is also a definite seasonal pattern. We also see that the amplitude of the seasonality (and possibly the irregularity) is increasing with the mean.

   b. Create ACF and describe.

   This ACF shows both seasonality and trend. We can see the seasonality because at a period of 1 and 2 (years ... 12 and 24 months) we see peaks in the ACF. The trend component is visible in that the decay in the ACF is very very slow.

   c. For each series explain why a linear trend estimate from a regression would be appropriate way to “detrend” the series.

   If we were to use a linear trend estimate via a linear regression, it would be better than nothing. However, it would not fully capture the trend in this series where the trend is nonlinear. Graphs of the Trend estimate, residuals (the detrended version of the series) and ACF of the residuals at right.

   We can see in these residuals and the ACF that they are clearly still seasonal but less trend.
d. If there is evident seasonality or trend, remove it with an appropriate difference then create ACF and describe.

At right we have the original series, seasonal difference \((y_t - y_{t-12})\) and the ACF of the seasonal difference. Here we see that the seasonal difference doesn’t appear to have a trend at all. The ACF also shows that the nature of any seasonal behavior is vastly reduced. There is no longer a strong positive autocorrelation. In fact, now it seems a bit negative: that if \(y_t - y_{t-12}\) is positive the lag 12 value, \(y_{t-12} - y_{t-24}\) would likely be negative. Perhaps odd, but that seems to be what the data are saying here.

e. Plot instead the \(\log_{10}\) (whether natural log or log base 10 doesn’t matter, but log base 10 makes more sense with such economic data as this) of the Beer production. Comment on whether the log transformation would be more appropriate for analysis. [Hint: the \(\log10()\) function in R can be used.]

Using a log-scale (see bottom graph at right) we can see that the series is a bit better than in the original scale (see graph immediately at the right). While there still is a trend, we can see that the seasonal behavior is now about the same size in magnitude. As our conceptual model for time series required the seasonal component to be identical across the cycles, the log-scale data would be more appropriate for analysis.
2. Repeat a-d for the Global Temperature dataset.
   a. Plot time series and describe the dataset.

   This series has an overall upwards trend. Before 1900, the series seems as if the mean isn’t increasing. Since then, it has been … and much more so after about 1970.

   b. Create ACF and describe.

   The ACF (again, at right) shows that there is a trend. It might be hard to spot, but at lags of 1 (year) and especially at 2 (years), we see a slight increase in the ACF values relative to the other, nearby ACF values.

   c. For each series explain why a linear trend estimate from a regression would be appropriate way to “detrend” the series.

   The linear trend (shown at right) along with residuals and their ACF show pretty clearly that the trend is lessened by estimating the trend with a linear fit. However, there is still quite clearly trend in the residual series as seen in the time series plot of the residuals and in the ACF of the residuals.
d. If there is evident seasonality or trend, remove it with an appropriate difference then create ACF and describe.

The seasonally differentiated doesn't appear to have a trend in the time-series plot. The ACF of the seasonally differenced data, due to the somewhat exponential (not linear) decay seems to not really be exhibiting trend.

e. Plot instead the log_{10} (whether natural log or log base 10 doesn't matter, but log base 10 makes more sense with such economic data as this) of the Beer production. Comment on whether the log transformation would be more appropriate for analysis. [Hint: the log10() function in R can be used.]

At the right we have the log transformation (note: because our Global temperatures are in departures from mean, some of the values are negative ... and when we need a log transform of values which are sometimes negative, we need to first add a constant that causes every value to be positive ... in this case, I added a 2).

On the log-scale we do NOT see the series as any better than the original series, in terms of the amount of variability.

A power transformation (in this case, cubing the Global temps + 2), seems to maybe make the series better for analysis.
3. Sketch time series of hourly temperature data for four days in SLO in February during a warming trend (heading from below average temps to above average temps). Sketch ACF for such a series.

The student could (or maybe should) draw this by hand. I faked a time series via:

```
Temp = 10*cos(2*pi*hour/24+.8*pi)+55 +20*hour/(4*24)+.75*rnorm(96)
```

For this (or any such series), the time series should clearly show an increase in the temperatures with a diurnal cycle. (In my case, my index, "hour" is the number of hours past midnight.)

The ACF shows both a seasonal pattern with the sinusoidal pattern (lags of 24 and 48 seem to exhibit higher autocorrelation than other nearby lags, indicating the 24 hour period) as well as trend as evidenced by the linear trend in the ACF that is above and beyond the sinusoidal seasonality.
IV. Yet another conceptual problem.

1. If \( \text{SE}_{ACF(k)} \) is approximately \( n^{1/2} \), for each of the ACF values in parts 1b and 2b above, add a margin of error around the sample ACFs you created to provide an approximate 95% CI for the underlying ACF value for the process being observed.

For this question, the student should develop a 95% CI for the process ACF by adding to and subtracting from the sample ACF a value equal to \( 2/\sqrt{n} \). A student could do so with pencil or by creating pictures sort of like:

The key here is that the CI for the process ACF, at any lag, is 95% sure to be found inside the sample ACF +/- \( 2/\sqrt{n} \) range.

[Note: I am not presenting here the comparable graph for the Global Temperatures dataset.]

2. Explain why such an approximate 95% CI would not include zero only if the sample ACF is outside the blue range provided by R’s acf() function.

This is essentially the duality between confidence intervals and hypothesis testing. Like how a CI for the difference in population means includes zero if and only if the hypothesis test for equality of means would not be rejected ... the CI for the process autocorrelation would include zero if and only if the sample autocorrelation is not statistically significantly different from zero.