Relationships between Two Time Series

- Most of our attention in previous chapters has been on modeling a single time series, and on using that model to forecast future values of the series.
- Sometimes the time series of interest is related to one (or more) other time series, which we could call *covariate time series*.
- We have seen the fish population levels over time are related to sea temperature values over time.
- Pasture production in Africa over time is related to certain climate variables that are measured over time.
- We can use tools such as *cross-correlation* and *time series regression* to explore how one time series may predict or explain another.

Notation for Working with Two Time Series

- Let $\{Y_t\}$ represent the main time series of interest (the response), and let $\{X_t\}$ represent the explanatory time series (the covariate series).
- Define the *cross-covariance function* as $\gamma_{t,s}(X, Y) = cov(X_t, Y_s)$ for each pair of integers t and s.
- Two time series $\{X_t\}$ and $\{Y_t\}$ are called jointly (weakly) stationary if their mean functions are constant and their cross-covariance $\gamma_{t,s}(X, Y)$ depends only on the time difference $t - s$.

The Cross-Correlation Function of Two Time Series

- If $\{X_t\}$ and $\{Y_t\}$ are jointly stationary, then their *cross-correlation function* (CCF) is $\rho_k(X, Y) = corr(X_t, Y_{t-k}) = corr(X_{t+k}, Y_t)$.
- Note that $\rho_0(X, Y)$ measures the *contemporaneous* (same-time) linear association between X and Y.
- And $\rho_k(X, Y)$ measures the lag-k cross-correlation, i.e, the linear association between X_t and Y_{t-k}.
- Note that $corr(X_t, Y_{t-k})$ need not equal $corr(X_t, Y_{t+k})$.

More on Cross-Correlation

- Sometimes the effect of the X variable on Y only manifests itself after a delay of a few time units.
- For example, suppose (monthly) pasture production Y is affected by the rainfall level X two months previously.
- Suppose the variables X_t and Y_t follow the regression model

$$Y_t = \beta_0 + \beta_1 X_{t-d} + e_t,$$

  where the X's are iid and the e_t are white noise, independent of the X's.
- The true CCF $\rho_k(X, Y)$ is zero at all lags except at lag $k = -d$, at which lag it has the same sign as β_1.
- We say here that X is "leading" Y by d time units.
- The value of X will take effect on Y at a time d units into the future.

Sample Cross-Correlation Function

- The sample cross-correlation function between the paired samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ is:

$$r_k(X, Y) = \frac{\sum(X_t - \bar{X})(Y_{t-k} - \bar{Y})}{[\sum(X_t - \bar{X})^2]^{1/2}[\sum(Y_t - \bar{Y})^2]^{1/2}}$$

  where the summation is over all indices that make sense.
- Note that if $X = Y$, this reduces to the formula for the sample ACF.
- Under the white-noise error model, any sample cross-correlation values that are greater than $1.96/\sqrt{n}$ can be considered significantly different from zero.

A Sample CCF on Some Simulated Data

- See the R code for an example of simulated $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ (with white-noise errors) with X leading Y by $d = 2$ time units.
- The theoretical CCF should be zero everywhere except lag -2.
- We see the sample CCF for these simulated data is significant at lag -2 and marginally significant at lag 3, but having at least one "false positive" (falsely significant) value is likely when we look at 33 CCF values.

A Model with Autocorrelated Errors

- The regression model we just looked at assumed the error terms were white noise.
- In reality, with time series data, it is common that the errors would display some autocorrelation.
- Suppose the variables X_t and Y_t follow the more general regression model

$$Y_t = \beta_0 + \beta_1 X_{t-d} + Z_t,$$

  where the X's are iid and the Z_t follow some $ARIMA(p, d, q)$ model, independent of the X's.
- Again here, X is "leading" Y by d time units, so that the value of X will take effect on Y at a time d units into the future.

Issues with the Sample CCF when the Errors are Autocorrelated

- When the errors are autocorrelated, using the $1.96/\sqrt{n}$ rule to judge at which lags the CCF values are significant can lead to *many* false positives.
- When X and Y are independent and the errors are white noise, we would expect the $1.96/\sqrt{n}$ rule to produce about 5% false positives.
- But when X and Y are both $AR(1)$ processes each with $\phi = 0.75$, the false positive rate is 30%.
- When X and Y are both $AR(1)$ processes each with $\phi = 0.9$, the false positive rate is 53%!
- When the series are nonstationary, the problem is even worse.
- This phenomenon leads to the diagnosis of "spurious correlations" (apparent correlations that are not really present).

Example of Spurious Correlation

- See the R example analyzing the monthly U.S. milk production and monthly U.S. electricity production from January 1994 to December 2005.
- Both series appear nonstationary (upward mean trend) and definitely seasonal.
- The sample CCF shows significant cross-correlations at many lags.
- But it is likely that these are spurious correlations.
- Section 11.4 discusses *prewhitening*, a method for disentangling the linear association between X and Y from their autocorrelation.
- If we take first differences and seasonal differences of the two time series, and use the `prewhiten` function, we see the sample CCF of the prewhitened data shows that the two time series are basically uncorrelated, which is more sensible.

Another Example of CCF with Prewhitened Data

- See the R example analyzing the weekly (logged) potato-chip sales and weekly average price from September 1998 to September 2000.
- Both series may be nonstationary, so we work with the first differences.
- The sample CCF of the prewhitened differenced data shows significant cross-correlations at only lag zero, and this sample correlation is strongly negative.
- It appears there is contemporaneous negative correlation between the first differences of price and sales.
- When the price from one week to the next goes down, the sales from one week to the next tends to go up.

Regression with Time Series

- If we want to use the covariate time series $\{X_t\}$ to predict the response time series $\{Y_t\}$, then time series regression is a useful tool.
- If both $\{Y_t\}$ and $\{X_t\}$ have white-noise errors, then ordinary least-squares methods can be used to regress $Y_1, \ldots, Y_n$ on $X_1, \ldots, X_n$.
- Ordinary least squares (OLS) regression can be implemented with the `lm` function in R.
- But often, the errors of the regression model are autocorrelated.

Regression with Time Series with Autocorrelated Errors

- Consider the regression model

$$Y_t = \beta_0 + \beta_1 X_{t-d} + Z_t,$$

  where Z_t is a noise process with some autocovariance function $\gamma_z(s, t)$.
- More generally, we could have several explanatory time series in the regression model.
- If Z_t follows a particular stationary ARMA model, we could identify an operator $\pi(B)$ that will transform Z_t into white noise: $\pi(B)Z_t = e_t$.
- In other word, this operator will *whiten* Z_t.
- Then the operator can be applied to the entire regression model and weighted least squares can be run on the computer to estimate the β's.

Details about Time Series Regression with Autocorrelated Errors

- The problem is that we do not know in advance the process that Z_t follows.
- In practice, we can guess this process based on the residuals from an OLS fit and proceed as follows:
  1. Run an OLS regression of Y_t on X_t (or on the several covariate series, if there is more than one) and retain the residuals from this OLS fit.
  2. Specify some ARMA-type model for the residuals, using our usual specification techniques.
  3. Run weighted least squares (WLS) with the noise process specified to follow that ARMA model.
  4. Check to see whether the residuals from this WLS fit resemble white noise, and repeat (2)-(4) using another ARMA model if they do not.

Implementing Time Series Regression with Autocorrelated Errors

- This approach can be implemented in R with the `arima` function in the `TSA` package or the `sarima` function in the `astsa` package.
- With each function, the covariate time series are named in the `xreg` argument.
- See the R example of the regression of logged sales on price in the Bluebird potato chip data.
- See the R example of the regression of cardiovascular mortality on time, (centered) temperature, squared temperature, and particulate level.
- We will examine a public transportation time series regression example after discussing outlier detection.

A Time Series Regression with Lagged Variables and a SARIMA Noise Process

- In the previous examples, we set $d = 0$; that is, X did not "lead" Y in time.
- When appropriate, we could use X_{t-d} instead of X_t in the regression model.
- When predicting the recruitment (fish population) value Y based on the SOI level X, there is evidence that X leads Y by about 6 months (see lagged scatterplot matrix in R).
- We fit the model

$$Y_t = \beta_0 + \beta_1 X_{t-6} + \beta_2 D_{t-6} + \beta_3 D_{t-6} \times X_{t-6} + Z_t,$$

where D_t is a dummy variable that is 1 if SOI > 0 and 0 otherwise.

- An examination of the residuals after a WLS fit shows seasonality, so we include a seasonal AR component in our model for the noise process (see R example).