Chapter 3: Regression Methods for Trends

- Time series exhibiting *trends* over time have a mean function that is some simple function (not necessarily constant) of time.
- The example random walk graph from Chapter 2 showed an upward trend, but we know that a random walk process has constant mean zero.
- That upward trend was simply a characteristic of that one random realization of the random walk.
- If we generated other realizations of the random walk process, they would exhibit different "trends".
- Such "trends" could be called *stochastic trends*, since they are just random and not fundamental to the underlying process.

Deterministic Trends

- The more important type of trend is a *deterministic trend*, which is related to the real nature of the process.
- Example: The plot of Dubuque temperature over time shows a periodic *seasonal* trend that reflects how the location is oriented to the sun across the seasons.
- In other examples, the trend in the time series might be *linear* or *quadratic* or some other function.
- Often a time series process consists of some specified trend, plus a random component.
- We commonly express such time series models using the form $Y_t = \mu_t + X_t$, where μ_t is a trend and X_t is a random process with mean zero for all t.

Estimation of a Constant Mean

- Consider a time series with a constant mean function, i.e., $Y_t = \mu + X_t$.
- The usual estimate of μ is the sample mean

$$\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$$

- This is an unbiased estimator of μ and it has variance

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

- If X_t is white noise ($\rho_k = 0$, i.e., zero autocorrelation), then $var(\bar{Y})$ simply equals γ_0/n, the familiar "population variance divided by the sample size."

Precision of Sample Mean

- In general, negative autocorrelation implies the sample mean will have smaller variance (greater precision).
- Positive autocorrelation implies the sample mean will have larger variance (worse precision).
- In other cases, some ρ_k's are positive and some are negative.
- In many stationary processes, the autocorrelation decays toward zero quickly as the lag increases so that

$$\sum_{k=0}^{\infty} |\rho_k| < \infty.$$

More on the Precision of the Sample Mean

- If $\rho_k = \phi^{|k|}$ for some $-1 < \phi < 1$, then by summing a geometric series and using a large-sample result:

$$var(\bar{Y}) \approx \left(\frac{1+\phi}{1-\phi} \right) \frac{\gamma_0}{n}$$

- For nonstationary random processes, the variance of $\bar{Y}$ can be undesirable.
- For example, if X_t is a random walk process, then $var(\bar{Y})$ *increases* as n increases, which is not good! We would need to estimate μ in some other way in this case.

Regression Methods

- Now we consider several common nonconstant mean trend models: linear, quadratic, seasonal means, and cosine trends.
- A linear trend is expressed as:

$$\mu_t = \beta_0 + \beta_1 t$$

- The least squares method chooses the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the least squares criterion:

$$Q(\beta_0, \beta_1) = \sum_{t=1}^{n} [Y_t - (\beta_0 + \beta_1 t)]^2$$

- The resulting estimates have the familiar formulas in equation (3.3.2) on page 30, and can be found easily using software (see R examples on simulated random walk data).

Other Regression Methods

- A quadratic trend is expressed as:

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

- The least squares method minimizes the least squares criterion:

$$Q(\beta_0, \beta_1, \beta_2) = \sum_{t=1}^{n} [Y_t - (\beta_0 + \beta_1 t + \beta_2 t^2)]^2$$

- Before fitting a linear (or quadratic) model, it is important to ensure that this trend truly represents the deterministic nature of the time series process and is not simply an artifact of the randomness of that realization of the process (random walk example).

Cyclical or Seasonal Trends

- The *seasonal means* approach represents the mean function with a different parameter for each level.
- For example, suppose each measured time is a different month, and we have observed data over a period of several years.
- The seasonal means model might specify a different mean response for each of the 12 months:

$$\mu_t = \begin{cases} \beta_1 \text{ for } t = 1, 13, 25, \ldots \\ \beta_2 \text{ for } t = 2, 14, 26, \ldots \\ \vdots \\ \beta_{12} \text{ for } t = 12, 24, 36, \ldots \end{cases}$$

- This is similar to an ANOVA model in which the parameters are the mean response values for each factor level.

Details of Seasonal Means Model

- The model as presented on the previous slide does not contain an intercept, and that fact needs to be specified in the fitting software.
- See R example for the model fit on the Dubuque temperature data.
- An alternative formulation does include an intercept and omits one of the β's in the previous model.
- The parameters are interpreted differently in that model formulation (see Dubuque temperature example).

Harmonic Regression for Cosine Trends

- The seasonal means model makes no assumption about the shape of the mean response function over time.
- A more specific model might assume that the mean response varies over time in some regular manner.
- For example, a model for temperature data might assume mean temperatures across time rise and fall in a periodic pattern, such as:

$$\mu_t = \beta \cos(2\pi f t + \phi),$$

where β is the *amplitude*, f the frequency, and ϕ the phase.

Fitting a Harmonic Regression Model

- To fit the model, it is useful to consider a transformation of the mean response function:

$$\mu_t = \beta_1 \cos(2\pi ft) + \beta_2 \sin(2\pi ft),$$

where $\beta = \sqrt{\beta_1^2 + \beta_2^2}$ and $\phi = \arctan(-\beta_2/\beta_1)$.

- Then β_1 and β_2 enter the regression equation linearly and we can estimate them using linear regression with $\cos(2\pi ft)$ and $\sin(2\pi ft)$ as predictors, along with an intercept term:

$$\mu_t = \beta_0 + \beta_1 \cos(2\pi ft) + \beta_2 \sin(2\pi ft),$$

Meanings of the Parameters

- The *amplitude* β is the height of the cosine curve from its midpoint to its top.
- The *frequency* f is the reciprocal of the *period*, which measures how often the curve's pattern repeats itself.
- For monthly data having $t = 1, 2, \ldots, 12, \ldots$, the period is 12 and the frequency $f = 1/12$.
- If the data are monthly but time is measured in years, e.g., $t = 2016, 2016.0833, 2016.1667$, etc., then the period is 1 and f would be 1 in this case.
- See R example on the Dubuque temperature data.
- Note that we can add more harmonic terms to the regression equation to create a more complicated mean response function, which will have more parameters, but which may improve the fit.

Properties of the Least Squares Estimators

- Section 3.4 gives some theoretical results about the variance of the least squares estimators.
- When the random component $\{X_t\}$ is not white noise, the least squares estimators are not the Best Linear Unbiased Estimators (BLUEs).
- However, under some general conditions on the random component $\{X_t\}$, it is known that if the trend over time is either:
  - polynomial
  - a trigonometric polynomial
  - seasonal means
  - any linear combination of these

  then the least squares estimates of the trend coefficients have approximately the same variance as the BLUEs, for large sample sizes.

Regression Output from Software

- The previously stated theoretical results imply some trust in the least squares method, at least for some types of trend, but the standard errors we obtain from software may still be incorrect (for moderate sample sizes) if the random component is not white noise.
- The R output gives us least squares estimates of the unknown parameters.
- It also gives the *residual standard deviation*, which estimates $\sqrt{\gamma_0}$, the standard deviation of $\{X_t\}$:

$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2},$$

where $\hat{\mu}_t$ is the trend with the parameter estimates plugged in, and p is the number of parameters estimated in μ_t.

- The smaller s, the better the fit, so the value of s can be used to compare alternative trend models.

More Regression Output from Software

- The *coefficient of determination* R^2 is the proportion of variation in the time series that is explained by the estimated trend.
- An adjusted R^2 measure is similar, but penalizes models with more parameters.
- Like s, the adjusted R^2 can be used to compare different trend models; a higher adjusted R^2 indicates a better model.
- R also produces estimated standard errors of the coefficient estimates, but these are only valid when the random component is white noise.

More Model Selection Criteria

- The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are other model selection criteria (similar to adjusted R^2).
- They are smaller for better fitting models, but they also penalize models with more parameters.
- See R examples for displaying AIC and BIC.
- One model selection strategy is to pick the model with the smallest AIC (or smallest BIC).

Caution About Model Selection Criteria

- The AIC and BIC are **likelihood-based** model selection criteria.
- It only makes sense to compare two models via AIC or BIC if the dependent variable is exactly the same for both models.
- For example, you should not use AIC or BIC to compare a model whose response variable is price to another model whose response variable is the logarithm of price (apples and oranges; completely different likelihood functions).
- We will see approaches to determine whether the response variable should enter the model in a transformed manner, but this should **not** be determined using AIC or BIC.

Residual Analysis

- For any observation at time t, the residual $\hat{X}_t = Y_t - \hat{\mu}_t$ serves as a prediction of the unobserved stochastic component.
- It can be used to assess the true nature of X_t, for example, does X_t behave like white noise?
- We can plot the residuals (often standardized) against time and look for any patterns that might deviate from white noise.
- We can also plot the residuals against the fitted trend values $\hat{\mu}_t$ and look for patterns.
- For example, does the variation in residuals change as the fitted trend values change?
- If we see notable patterns in the plots, we may rethink whether our model assumptions are appropriate.
- See examples with Dubuque temperature data.

More Residual Analysis

- We can examine whether the residuals appear normally distributed using a histogram, or, even better, a normal Q-Q plot of the (standardized) residuals.
- A roughly straight-line pattern in the Q-Q plot is evidence that the assumption of a normally distributed stochastic component is reasonable.
- The Shapiro-Wilk test is a formal test for normality.
- The null hypothesis is that the stochastic component is normal. We would doubt this normality assumption only if the Shapiro-Wilk p-value were small, say less than 0.05.

Assessing Independence

- The *runs* test on the residuals is one way to test whether the stochastic component is independent across time.
- A *run* is a sequence of one or more residuals that are each above (or below) the overall median residual.
- If there are many runs, this indicates excessive alternation back and forth across the median, a sign of negative autocorrelation.
- Very few runs indicates large residuals tend to be followed by large residuals, and negative residuals by negative – a sign of positive autocorrelation.
- The null hypothesis of the runs test is that there is independence – a small p-value would show evidence against this independence (either based on a very small or very large number of runs).
- See example of runs test with Dubuque data.

Sample autocorrelation function

- The sample autocorrelation function r_k is an estimate of the autocorrelation ρ_k between values separated by lag of size k:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{1}$$

- A *correlogram* is a plot of r_k against k which can be used to assess dependence over time.
- The `acf` function in R produces the sample autocorrelation function for each lag k, along with dashed horizontal lines at plus/minus two standard errors ($\pm 2/\sqrt{n}$).
- If the process follows white noise, we expect the autocorrelations at each lag to remain within the dotted lines.
- See R examples.

Regression with Lagged Variables

- Sometimes the response variable at the present time is related to some predictor variable's value several time periods in the past.

$$Y_t = \beta_0 + \beta_1 X_{t-k} + \epsilon_t$$

- This can be done fairly easily in R; see example with recruitment and SOI data.

Dealing with Nonstationary Processes

- We often see processes which can be represented by the form $Y_t = \mu_t + X_t$, where μ_t is a *nonconstant* trend and X_t is a stationary noise process with mean zero.
- The overall process Y_t is a nonstationary process, since its mean function is not constant.
- *Detrending* is a method of estimating the trend, then subtracting the fitted values from Y_t to get the residuals.
- The detrended residual process is then stationary.
- Another way of transforming nonstationary processes into stationary processes is by *differencing*.
- This amounts to obtaining the *differences* of successive Y-values: $\nabla Y_t = Y_t - Y_{t-1}$.
- Then the *differenced* time series, $\{\nabla Y_t\}$, is stationary.

Detrending or Differencing?

- An advantage of detrending is that we obtain an estimate of the trend, which may be useful.
- An advantage of differencing is that we don't need to estimate any parameters (or even assume any model for the trend).
- If the trend estimate is not needed, then converting the process to stationarity is more easily done by differencing.

The Backshift Operator

- Define the *backshift* operator B as

$$BY_t = Y_{t-1}$$

- Similarly, $B^2 Y_t = Y_{t-2}$, and in general,

$$B^k Y_t = Y_{t-k}$$

- The *inverse* of the backshift operator is the *forward-shift operator* B^{-1}, defined so that $B^{-1} Y_{t-1} = Y_t$, and $B^{-1} B Y_t = Y_t$.

The Backshift Operator and Differencing

- We see that the first difference operator ∇ can be expressed as

$$\nabla Y_t = (1 - B) Y_t$$

- The second difference is simply

$$\nabla^2 Y_t = (1 - B)^2 Y_t = (1 - 2B + B^2) Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$$

which is equal to

$$\begin{aligned} \nabla(\nabla Y_t) &= \nabla(Y_t - Y_{t-1}) \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}. \end{aligned}$$

- In general,

$$\nabla^d Y_t = (1 - B)^d Y_t$$

- See examples of differencing of chicken price and global temperature data.

Lagged Scatterplot Matrix

- Relations between two time series (possibly lagged) may not be linear.
- One way to investigate the form of associations (including lags) between time series is a lagged *scatterplot matrix*.
- A scatterplot matrix could show associations between Y_t and Y_{t-k} for $k = 1, 2, \ldots$. (investigating lags in a single time series).
- Or it could show associations between Y_t and X_{t-k} for $k = 0, 1, 2, \ldots$. (investigating associations and lags between two separate time series).
- See examples with recruitment and SOI data.

Smoothing Time Series

- *Smoothing* is a general way to discover long-term (possibly nonlinear) trends in time series.
- The *moving average* smoother represents the mean at time t by the average of the observed values around t:

$$m_t = \sum_{j=-k}^{k} a_j Y_{t-j},$$

  where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$.
- If the nonzero a_j's are all equal, then this is a simple moving average. Otherwise, it is a weighted moving average.

Kernel Smoothing of Time Series

- A *kernel* smoother determines the a_j weights according to a *kernel* function, which is a symmetric density function (often chosen to be a normal density) centered at zero.
- When calculating m_t, values close to t are given more weight by the kernel function.
- The *bandwidth* of the kernel function controls the width of the density function used.
- The larger the bandwidth, the smoother the overall m_t curve appears.
- See examples in R using the SOI series.

Lowess Smoothing

- *Lowess* smoothing is a weighted regression method that is similar to kernel smoothing.
- The *span* is the fraction of the values in the time series that are used at each calculation of m_t.
- The span plays a similar role as the bandwidth.
- The larger the span, the smoother the overall m_t curve appears.
- Lowess is also a good way to present a nonlinear relationship between two time series.
- See R examples on SOI data (one time series), and the temperature and mortality data (two time series).

Classical Structural Modeling

- Classical structural modeling decomposes the time series into several distinct components, for example:

$$Y_t = T_t + S_t + \epsilon_t,$$

  where T_t is a trend component, S_t is a seasonal component, and ϵ_t is a noise component.
- The R function `stl` estimates and plots each of these components.
- The decomposition is not unique.
- This does not work well for every time series – some series may involve another cyclic component C_t, for example, a business cycle (beyond seasonal) in sales data.
- Sometimes it is unclear whether some pattern should be considered trend or a business cycle – see R example with Hawaiian hotel occupancy data.