Statistical Time Series Analysis

Clayton W. Schupp, Galvanize

Spring 2015

What Constitutes a Time Series

Time Series Analysis: the analysis of experimental data that have been observed at different points in time (usually equally spaced time points)

The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of conventional statistical methods which are based on independent and identically distributed data

The first eight weeks or so of this course will cover statistical techniques to forecast future observations based on past observations

Basic Methodology

1. a “pattern” is first attained from the data at hand
2. the “pattern” is then extrapolated into the future to prepare a forecast

Assumptions? The “pattern” we’ve observed will continue...

Some applications:

- predicting/forecasting future gas prices
- global warming? Predicting future global temperatures
- population growth

Components of a Time Series

- Trend: the long-run upward or downward movement of the series
- Cycle: the upward or downward movement of the series around the trend (think of a wave)
- Seasonal Variations: patterns in the data that follow yearly patterns (think of seasonal temperatures)
- Irregular Variations: the remaining erratic movements in the series that cannot be accounted for

Our goal is to estimate the trend, cycle, and seasonal components of a time series so that all that is left is irregular fluctuations (often referred to as white noise)

Time Series Regression

Most useful when the parameters describing the time series to be forecast remain constant over time.

$$y_t = TR_t + \epsilon_t \qquad where \qquad \epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

1. Linear Trend

$$TR_t = \beta_0 + \beta_1 t$$

2. Quadratic Trend

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

3. Pth Order Polynomial Trend

$$TR_t = \beta_0 + \beta_1 t + \cdots + \beta_P t^p$$

Adding a Seasonality Factor

$$y_t = TR_t + SN_t + \epsilon_t \qquad \textit{where} \qquad \epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

1. Main assumption: constant seasonal variation for basic time series regression
2. We model the seasonal factor SN_t using indicator variables (aka "dummy" variables)
   - Need to model L seasons in SN_t using $L-1$ indicator variables:

$$SN_t = \beta_{S_1} \mathbb{1}\{S_1\} + \beta_{S_2} \mathbb{1}\{S_2\} + \cdots + \beta_{S_{L-1}} \mathbb{1}\{S_{L-1}\}$$

where

$$\mathbb{1}\{S_i\} = \begin{cases} 1 & \text{If } t \text{ is season } i \\ 0 & \text{Otherwise} \end{cases}$$

Additive Decomposition

Model:

$$y_t = TR_t + SN_t + CL_t + IR_t$$

where TR_t, SN_t, CL_t, IR_t are defined to be the trend, seasonal, cyclical, and irregular factors

Our goal is to estimate the above factors with point estimates tr_t, sn_t, cl_t, ir_t and use these estimates for forecasting.

Method

1. Estimate SN_T by grouping seasons and computing the average and normalizing
2. Estimate TR_T by deseasonalizing the observations and fitting a standard regression
3. Estimate CL_T by removing the season and trend and performing a moving average
4. Estimate IR_T by looking at the residual after removing the main components

Forecasting

Since our goal is prediction, we assume that the pattern continues in the future and that there is no pattern in the irregular component and so we predict IR_t to be zero

Thus the point estimate of the forecast at time t is

$$\hat{y}_t = tr_t + sn_t + cl_t$$

An appropriate prediction interval for y_t is

$$\hat{y}_t \pm B_{t,\alpha}$$

where $B_{t,\alpha}$ is the error bound in a $100(1-\alpha)\%$ prediction interval for d_t

Main Types

1. Simple Exponential Smoothing: used when the time series has no significant linear trend (i.e. slope) but the mean is changing over time
2. Holt's Trend Corrected Exponential Smoothing: used when a series has a linear trend and a slope that is changing over time
3. Holt-Winter's Method: extension of Holt's trend method for a series that has seasonality

Simple Exponential Smoothing

- If no trend exists and the mean remains constant, then our standard linear model

  $$y_t = \beta_0 + \varepsilon \qquad \textit{where} \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

  can be used giving us the simple estimate:

  $$\hat{y} = b_0 = \frac{1}{n} \sum y_t$$

  giving equal weight to all data points
- If instead, the mean might be changing slowly over time then it might be better to give more recent observations greater weight than older observations

Simple Exponential Smoothing gives the most recent observations the greatest weight and allows the forecaster to detect changes in the mean level and incorporate them into the model.

Method

1. Begin with an estimate of the mean at time point zero by averaging a subset (n_S) of the observations.

$$l_0 = \frac{1}{n_S} \sum_{t=1}^{n_S} y_t$$

2. Compute updated estimates using the following smoothing equation:

$$l_t = \alpha y_t + (1 - \alpha) l_{t-1}$$

where $0 < \alpha < 1$ is chosen to be an arbitrary value to be improved during the optimization process.

3. Next we want to find the best α by minimizing the Sum of Squared Error (SSE)

Error Correction Form of Smoothing Equation

$$\begin{aligned} l_t &= \alpha y_t + (1-\alpha) l_{t-1} \\ &= \alpha y_t + l_{t-1} - \alpha l_{t-1} \\ &= l_{t-1} + \alpha (y_t - l_{t-1}) \end{aligned}$$

- Remember that l_{t-1} is the forecast of y_t so $(y_t - l_{t-1}$ is the forecast error
- Therefore l_t is the estimate of y_t plus a fraction of the forecast error of y_t

Forecasting

- We can now forecast into the future, $y_{T+\tau}$. Our estimates will become less accurate as we move further from our last observation, y_t, and the prediction intervals will get wider.
- All point predictions will be based on l_T which is the forecast for y_{T+1}, the first time point after our final observation.
- The prediction interval for any time point, τ, in the future will be:

$$l_T \pm z_{\alpha/2} s \sqrt{1 + (\tau - 1)\alpha^2} \qquad where \qquad s = \sqrt{\frac{SSE}{T-1}}$$

Holt's Trend Corrected Exponential Smoothing

- If a time series is increasing or decreasing at a fixed rate, use linear regression model

$$y_t = \beta_0 + \beta_1 t + \varepsilon \qquad \textit{where} \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

  where the increase between time, $t - 1$, and time, t, is simply the growth rate, β_1
- What if both the mean level and the growth rate are changing with time? We need a model to describe these changing levels

Method

1. Let l_{t-1} estimate the mean at time period, $t-1$, and b_{t-1} estimate the growth rate. Now our estimate of y_t is $l_{t-1} + b_{t-1}$
2. Begin with an estimate of the mean and growth rate at time point zero. Use standard linear regression on a subset of the data.
3. Compute updated estimates using the following smoothing equations:

$$\begin{aligned} l_t &= \alpha y_t + (1-\alpha)[l_{t-1} + b_{t-1}] \\ b_t &= \gamma(l_t - l_{t-1}) + (1-\gamma)b_{t-1} \end{aligned}$$

   where $0 < \alpha, \gamma < 1$ are chosen to be arbitrary values to be improved during the optimization process.
4. Next we want to find the best α, γ by minimizing the Sum of Squared Error (SSE)

Forecasting

- We can now forecast into the future, $y_{T+\tau}$. Our estimates will become less accurate as we move further from our last observation, y_t, and the prediction intervals will get wider.
- All point predictions will be based on l_T, b_T and time τ:

$$\hat{y}_{T+\tau} = l_T + \tau b_T$$

- The prediction interval for any time point, τ, in the future will be:

$$l_T \pm z_{\alpha/2} s \sqrt{1 + \sum_{j=1}^{\tau-1} \alpha^2 (1 + j\gamma)^2} \quad where \quad s = \sqrt{\frac{SSE}{T-2}}$$

Holt-Winter's Exponential Smoothing

- If we had a fixed trend and fixed seasonal factor, our previous model was:

$$y_t = TR_t + SN_t + \varepsilon \qquad \textit{where} \qquad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

- If linear trend is changing with time and has constant seasonal variation we can use the additive Holt-Winter's method to extend Holt's method

Method

1. Let sn_{t-L} denote the most recent estimate of the seasonal factor for the season in time t where L is the number of seasons.
2. Begin with Holt's trend corrected method on a subset of the data
3. Our seasonal baseline factors are found as follows:
    1. Detrend the data by computing: $y_t - \hat{y}_t$
    2. Compute the baseline seasonal factors by averaging the detrended values over the years for each season
4. Compute updated estimates using the following smoothing equations:

$$
\begin{aligned}
l_t &= \alpha(y_t - sn_{t-L}) + (1-\alpha)[l_{t-1} + b_{t-1}] \\
b_t &= \gamma(l_t - l_{t-1}) + (1-\gamma)b_{t-1} \\
sn_t &= \delta(y_t - l_t) + (1-\delta)sn_{t-L}
\end{aligned}
$$

Forecasting

- Point predictions: $\hat{y}_{T+\tau} = l_T + \tau b_T + sn_{T+\tau-L}$
- The prediction interval for any time point, τ, in the future will be:

$$\hat{y}_{T+\tau} \pm z_{\alpha/2} s \sqrt{c_\tau} \quad \text{where} \quad s = \sqrt{\frac{SSE}{T-3}}$$

and

$$\begin{aligned} c_\tau &= 1 && \text{if } \tau = 1 \\ &= 1 + \sum_{j=1}^{\tau-1} \alpha^2 (1 + j\gamma)^2 && \text{if } 2 \le \tau \le L \\ &= 1 + \sum_{j=1}^{\tau-1} [\alpha(1 + j\gamma) + d_{j,L}(1-\alpha)\delta]^2 && \text{if } \tau > L \\ d_{j,L} &= 1 \text{ if } j \text{ is a multiple of L} \\ &= 0 \text{ otherwise} \end{aligned}$$

ARIMA Models

Afternoon Lecture

ARIMA Models

The Box-Jenkins Methodology applies autoregressive moving average models to find the best fit of a time series based on past values.

3-stage approach:

1. Model Identification
   - making sure data is stationary
   - identifying seasonality
   - Using plots of autocorrelation (ACF) and partial autocorrelation (PACF) to decide which autoregressive and/or moving average components to include
2. Parameter Estimation via MLE or non-linear least squares
3. Model Checking by testing the model residuals

If estimation is inadequate, we return to step 1 and iterate.

Autocorrelation Function

Correlation at adjacent points of the same series is measured by the autocorrelation function (ACF):

$$\rho_y(h) = \frac{\gamma_y(h)}{\gamma_y(0)} \qquad\qquad -1 < \rho_y(h) < 1$$

Since the true ACF is unknown, we utilize a sample version of the ACF

$$\hat{\rho}_y(h) = \frac{\hat{\gamma}_y(h)}{\hat{\gamma}_y(0)}$$

where

$$\hat{\gamma}_y(h) = \frac{1}{n}\sum_{t=1}^{n-h}(y_{t+h} - \bar{y})(y_t - \bar{y}) \quad \textit{with} \quad \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

Autocorrelation Function

Under the assumption that the underlying process y_t is gaussian white noise (after we have taken the appropriate steps), the approximate standard error of the ACF is

$$\sigma_{\hat{\rho}} = \frac{1}{\sqrt{n}}$$

This implies that asymptotically

$$\hat{\rho}_y(h) \sim N\left(0, \frac{1}{n}\right)$$

From normal theory, values within $\pm 1.96\sigma_{\hat{\rho}}$ might be reasonable for $\alpha = 0.05$.

Partial Autocorrelation Function

On can think of the PACF as the simple correlation between two points separated by a lag h, say y_t and y_{t-h}, with the effect of the intervening points$y_{t-1}, y_{t-2}, \dots, y_{t-h+1}$ conditioned out.

Suppose we want to predict y_t from $y_{t-1}, \dots, y_{t-h}$ using some linear function of these past values. Consider minimizing the mean square prediction error

$$MSE = E[(y_t - \hat{y}_t)^2]$$

using the predictor

$$\hat{y}_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_h y_{t-h}$$

over the possible values of the weighting coeffcients $a_1, \dots, a_h$ where we assume, for convenience that y_t has been adjusted to have zero mean.

Partial Autocorrelation Function

The partial autocorrelation function (PACF) is defined as the value of the last coeffcient

$$\phi_{hh} = a_h$$

In practice, we minimize the sample error sum of squares

$$SSE = \sum_{t=h+1}^{n} \left[(y_t - \bar{y}) - \sum_{k=1}^{h} a_k (y_{t-k} - \bar{y}) \right]^2$$

with the estimated partial correlation defined as $\hat{\phi}_{hh} = \hat{a}_h$ and

$$\hat{\phi}_{hh} \sim N\left(0, \frac{1}{n}\right)$$

Stationary Time Series

A stationary time series is one for which the statistical behavior of a set of observations $y_{t_1}, y_{t_2}, \ldots y_{t_k}$ is identical to that of the shifted set of observations $y_{t_{1+h}}, y_{t_{2+h}}, \ldots y_{t_{k+h}}$ for any collection of time points $t_1, t_2, \ldots t_k$ and for any shift h (lag)

This is the definition of *strong stationarity* and is too strong for most applications

Weak Stationarity

There is a relaxed definition, referred to as *weak stationarity* which requires only that the first and second moments satisfy the following constraints:

1. $$E(y_t) = \mu \qquad \forall t$$
2. $$\gamma_y(h) = E[(y_{t+h} - \mu)(y_t - \mu)]$$

where E is the usual expectation over the population density, h is the time shift (lag), and $\gamma_y(h)$ is called the autocovariance function and we additionally assert that

$$\gamma_y(h) = \gamma_y(-h)$$

Nonstationarity

Most time series are not stationary to begin with, so need to modify the series to improve the approximation of stationarity by

- detrending $\longrightarrow$ Constraint 1
- differencing $\longrightarrow$ Constraint 1
- transformations $\longrightarrow$ Constraint 2

Detrending

The general version of a nonstationary time series is to assume a general linear trend of the form

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

The natural thing to do is to consider the residual

$$\hat{e}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 t$$

as a plausible stationary series where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated intercept and slope coefficients based on least squares estimation.

Differencing

A common first step for achieving stationity is with the first difference

$$\bigtriangledown y_t = y_t - y_{t-1}$$

$\bigtriangledown y_t$ is useful for series with trend.

Higher order differences are defined as successive applications of the operator $\bigtriangledown$. For example, the second difference is

$$\begin{aligned} \bigtriangledown^2 y_t &= \bigtriangledown[\bigtriangledown y_t] \\ &= \bigtriangledown[y_t - y_{t-1}] \\ &= [y_t - y_{t-1}] - [y_{t-1} - y_{t-2}] \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

If the model also contains a quadratic trend term, $\bigtriangledown^2 y_t$ usually reduces the model to a stationary form.

Higher Order Differences

In order to find higher order differences, we must first define the **backshift operator (B)**:

$$By_t = y_{t-1}$$

and we can extend it to higher powers:

$$B^2 y_t = B(By_t) = By_{t-1} = y_{t-2} \longrightarrow B^k y_t = y_{t-k}$$

Higher order differences are defined as

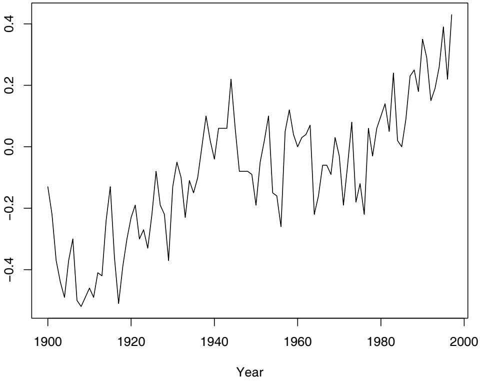
$$\triangledown^d y_t = (1-B)^d y_t$$

Lets find the fourth difference $\triangledown^4 y_t$:

$$\begin{aligned} \triangledown^4 y_t &= (1-B)^4 y_t \\ &= (1-4B+6B^2-4B^3+B^4) y_t \\ &= y_t - 4y_{t-1} + 6y_{t-2} - 4y_{t-3} + y_{t-4} \end{aligned}$$

Example: Global Temperatures

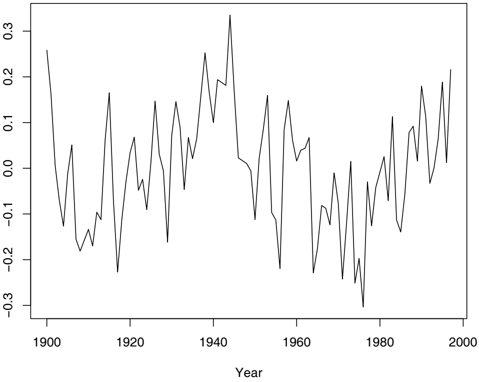
Consider a global temperatures series: the data are a combination of land-air average temperature anomalies for the years 1900-1997.



Example: Global Temperatures Continued

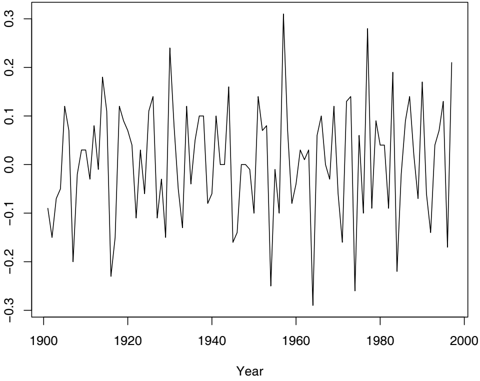
We first look at detrending the time series and examine the residuals from the model

$$\widehat{Temp} = -12.2 + 0.006 * Years$$



Example: Global Temperatures Continued

Next, look at differencing the time series with the first difference



Transformations

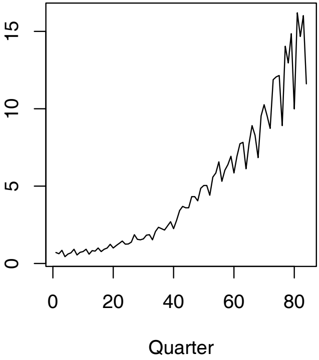
A transformation that cuts down the values of larger peaks of a time series and emphasizes the lower values may be effective in reducing nonstationary behavior due to changing variance. Examples:

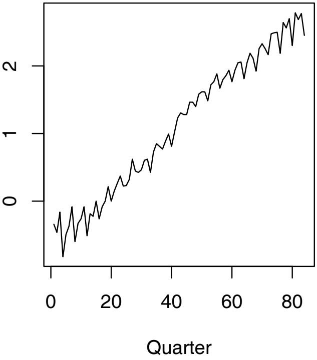
- The logarithmic transformation, $y_t = log(y_t)$ usually exponential-base (natural log)
- The square root transformation, $y_t = \sqrt{y_t}$, useful for count data
- More general transformations fall within the Box-Cox family

$$y_t = \begin{cases} \frac{1}{\lambda}(y_t^\lambda - 1) & \text{if } \lambda \neq 0 \\ \ln(y_t) & \text{if } \lambda = 0 \end{cases}$$

Example: Johnson & Johnson Data

The following figures shows quarterly earnings per share for Johnson&Johnson from 1960 to 1980 before and after a log transformation.





Assessing Stationarity

Box-Jenkins models describe **stationary** time series, so we must determine if the series is stationary and if not, transform it to attain stationarity.

We can use the ACF to determine if the time series is stationary

- If the ACF dies down quickly (relatively few significant lags) we can consider the series stationary
- If it dies down slowly then we need to try some transformations to either remove the trend or stabilize the variance
- We will then use the new transformed series as the "working" time series for the remainder of the analysis.

General Steps

Once we have achieved stationarity, we can then use the ACF and PACF to identify a potential model:

- If there is a spike ("highly significant") at lag k where all lags afterward are 'not significant' we can focus on the relationship between y_T and y_{T-k}
- We say the ACF or PACF "cuts off after lag k" if all the lags beyond are 'not significant'
- Need to realize that both the ACF and PACF might not "cut off" at the same lag. For instance the PACF might not cut off at all while the ACF has cut off

Autoregressive (AR) Models

The idea behind AR(p) models is that the present value of the series can be explained as a function of p past values

$$AR(p) : y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

We can also use the backshift operator to write the model more succinctly:

$$\Phi(B) y_t = \varepsilon_t$$

where $\Phi(B) = 1 - \phi_1(B) - \cdots - \phi_p(B^p)$ and $B^k y_t = y_{t-k}$

The general pattern of an AR(p) model will have an ACF that dies down in a steady fashion dominated by damped exponential decay and the PACF will have a significant peak at lag p and then cut off.

Moving Average (MA) Models

As an alternative to the AR(p) representation in which y_t is a linear combination of previous observations, the MA(q) model assumes that white noise is combined linearly to form the observed data

$$MA(q) : y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

We can also use the backshift operator to write the model more succinctly:

$$y_t = \Theta(B)\varepsilon_t$$

where $\Theta(B) = 1 + \theta_1(B) + \cdots + \theta_q(B^q)$ and $B^k y_t = y_{t-k}$

The general pattern of an MA(q) model will have an PACF that dies down in a steady fashion dominated by damped exponential decay and the ACF will have a significant peak at lag q and then cut off.

Autoregressive Moving Average (ARMA) Models

ARMA(p,q) models are naturally a combination of AR(p) and MA(q) models such that the model is written:

$$\Phi(B)y_t = \Theta(B)\varepsilon_t$$

The general pattern of an ARMA(p,q) model will have an PACF than has a significant peak at laq *p* and then cut off while the ACF will have a significant peak at lag *q* and then cut off.

ARIMA(p,d,q) models are ARMA(p,q) models that use the difference (d) to achieve stationarity.

Box-Jenkins Method

General Steps:

- Identify the steps needed to achieve stationarity (detrending, differencing, and/or transformations)
- Utilize ACF to see if stationarity is achieved. Lags should die down quickly.
- Use ACF/PACF to identify potential models by where the lags cut off.
    1. PACF identifies AR(p)
    2. ACF identifies MA(q)
- Fit model and examine residuals
    1. If white noise, you're done
    2. Else, identify additional components and update model

Seasonal ARIMA

Seasonal ARIMA (or SARIMA(p,d,q)x(P,D,Q)xL) are an extension of ARIMA(p,d,q) models to address seasonality.

Methods used are identical to those above except we focus on the seasonal lags (L) including taking a seasonal difference $\bigtriangledown^D$ which is the difference in the seasonal lags.

For Example:

1. If quarterly data, we would focus on lags: 4, 8, 12, etc
2. If monthly data, we would focus on lags: 12, 24, 36, etc
3. If daily data, we might focus on lags: 7, 14, 21, etc
4. If hourly data, we might focus on lags: 24, 48, 72, etc