Spring 2013 Statistics 153 (Time Series) : Lecture One

Aditya Guntuboyina

22 January 2013

A time series is a set of numerical observations, each one being recorded at a specific time. Time series data arise everywhere. The aim of this course is to teach you how to analyze such data.

We will focus on two approaches to time series analysis: (1) Time Domain Approach, and (2) Frequency Domain Approach (also known as the spectral or Fourier analysis of time series). In the Time domain approach, one works directly with the data while in the Frequency Domain approach, one works with the Discrete Fourier Transform of the data.

Very roughly, 60 percent of the course will be on time domain methods and 40 percent will be on frequency domain methods.

1 Time Series Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample time series data.

We assume that the observed data $x_1, \ldots, x_n$ is a realization of a sequence of random variables $X_1, \ldots, X_n$. The basic strategy of modelling is always to start simple and to build up.

We will study models for time series in this class. Basic Models:

1. White Noise
2. Deterministic trend + white noise
3. Deterministic seasonality + white noise
4. Deterministic trend + deterministic seasonality + white noise
5. Stationary time series models
6. Stationary ARMA models
7. ARIMA models
8. Seasonal ARIMA models
9. Modeling and estimating Spectral Density

1.1 White Noise

$X_1, \ldots, X_n$ are called white noise if they have mean zero, variance σ^2 and are uncorrelated.

An important special case of white noise is Gaussian White Noise where $X_1, \ldots, X_n$ are i.i.d $N(0, \sigma^2)$.

How to check if white noise is a good model for a given dataset? Think in terms of forecasting. For white noise, the given data cannot help in predicting X_{n+1}. The best estimate of X_{n+1} is $E(X_{n+1}) = 0$. In particular, X_1 cannot predict X_2; X_2 cannot predict X_3 and so on. Therefore, the correlation coefficient between $Y = (X_1, \ldots, X_{n-1})$ and $Z = (X_2, \ldots, X_n)$ must be close to zero.

The formula for the correlation between Y and Z is:

$$r := \frac{\sum_{t=1}^{n-1}(X_t - \bar{X}_{(1)})(X_{t+1} - \bar{X}_{(2)})}{\sqrt{\sum_{t=1}^{n-1}(X_t - \bar{X}_{(1)})^2 \sum_{t=1}^{n-1}(X_{t+1} - \bar{X}_{(2)})^2}}$$

where

$$\bar{X}_{(1)} = \frac{\sum_{t=1}^{n-1} X_t}{n-1} \text{ and } \bar{X}_{(2)} = \frac{\sum_{t=1}^{n-1} X_{t+1}}{n-1}.$$

This formula is usually simplified to obtain

$$r_1 = \frac{\sum_{t=1}^{n-1}(X_t - \bar{X})(X_{t+1} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

where $\bar{X} := \sum_{t=1}^{n} X_t/n$. Note that we are calling this correlation r_1 (note the subscript 1). This quantity r_1 is called the sample autocorrelation coefficient (sample ACF) of $X_1, \ldots, X_n$ at lag one. Lag one because this correlation is between X_t and X_{t+1}.

When $X_1, \ldots, X_n$ are obtained from white noise, r_1 is close to zero, particularly when n is large.

One can similarly define sample autocorrelations at other lags:

$$r_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} \qquad \text{for } k = 1, 2, \ldots$$

Here is an important mathematical fact: Under certain additional conditions (which are satisfied for gaussian white noise), if $X_1, \ldots, X_n$ are white noise, then the sample autocorrelations $r_1, r_2, \ldots$ are **independently** distributed according to the normal distribution with mean zero and variance $1/n$.

Note that the variance decreases to zero as n increases and the mean is zero. Thus for large n, the sample autocorrelations should be very close to zero. Also note that the sample autocorrelations for different lags are independent.

Therefore, one way of checking if the white noise model is a good fit to the data is to plot the sample autocorrelations. This plot is known as the **correlogram**. Use the function **acf** in **R** to get the correlogram. The blue bands in the correlogram correspond to levels of $\pm 1.96 n^{-1/2}$.

How to interpret the correlogram? When $X_1, \ldots, X_n$ are white noise, the probability that a fixed r_k lies outside the blue bands equals 0.05. A value of r_k outside the blue bands is **significant** i.e., it gives evidence against pure randomness. However, the overall probability of getting atleast one r_k outside the bands increases with the number of coefficients plotted. For example, if 20 r_ks are plotted, one expects to get one significant value under pure randomness.

Here are a couple rules of thumb for deciding if a correlogram indicates departure from white noise:

- A single r_k just outside the bands may be ignored but two or three values well outside indicate a departure from pure randomness.

- A single significant r_k at a lag which has some physical interpretation such as lag one or a lag corresponding to seasonal variation also indicates evidence of non-randomness.

Spring 2013 Statistics 153 (Time Series) : Lecture Two

Aditya Guntuboyina

24 January 2013

1 Last Class

- Data Examples
- White noise model
- Sample Autocorrelation Function and Correlogram

2 Trend Models

Many time series datasets show an increasing or decreasing trend. A simple model for such datasets is obtained by adding a deterministic trend function of time to white noise:

$$X_t = m_t + Z_t$$

Here m_t is a deterministic trend function and Z_t is white noise. There exist two main techniques for fitting this model to the data.

2.1 Parametric form for m_t and linear regression

Assume a simple parametric form for m_t, say linear or quadratic, and fit it via linear regression.

2.2 Smoothing

Here we estimate m_t without making any parametric assumptions about its form.

The idea is that to get m_t from $X_t = m_t + Z_t$, we need to eliminate Z_t. It is well-known that noise is eliminated by averaging. Consider

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}. \tag{1}$$

If m_t is linear on the interval $[t-q, t+q]$, then check that

$$\frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} = m_t.$$

Thus if m_t is approximately linear over $[t-q, t+q]$, then

$$\hat{m}_t \approx m_t + \frac{1}{2q+1}\sum_{j=-q}^{q} Z_{t+j} \approx m_t.$$

The defining equation for $\hat{m}_t$ will have trouble when calculating averages near end-points. To counter, just define X_t to be X_1 for $t < 1$ and X_n for $t > n$.

$\hat{m}_t$ is also called the Simple Moving Average of X_t.

Key Question: How to choose the smoothing parameter q? Observe that:

$$\hat{m}_t = \frac{1}{2q+1}\sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1}\sum_{j=-q}^{q} Z_{t+j}.$$

If q is very small, then the second term above is not quite small and so the trend estimate will also involve some noise component and therefore $\hat{m}_t$ will be very noisy. On the other hand, if q is large, then the assumption that m_t is linear on $[t-q, t+q]$ may not be quite true and thus, $\hat{m}_t$ may not be close to m_t. This is often referred to as the Bias-Variance tradeoff. Therefore q should be neither too small nor too large.

2.2.1 Parametric Curve Fitting versus Smoothing

Suppose there is good reason to believe that there is an underlying linear or quadratic trend function. In this case, is it still okay to use smoothing?

No, when there is really is an underlying linear trend, fitting a line gives a more precise estimate of the trend. On the other hand, the estimation of trend by smoothing only uses a few observations for each time point and the resulting estimate is not as precise. This is the price one has to pay for giving up the assumption of linearity. Of course, when there is no reason to believe in an underlying linear trend, it might not make sense at all to fit a line. Smoothing is the way to go in such cases.

2.3 More General Filtering

The smoothing estimate (1) of the trend function m_t is a special case of *linear filtering*. A linear filter converts the observed time series X_t into an estimate of the trend $\hat{m}_t$ via the linear operation:

$$\hat{m}_t = \sum_{j=-q}^{s} a_j X_{t+j}.$$

The numbers $a_{-q}, a_{-q+1}, \ldots, a_{-1}, a_0, a_1, \ldots, a_s$ are called the weights of the filter. The Smoothing method is clearly a special instance of filtering with $s = q$ and $a_j = 1/(2q+1)$ for $|j| \leq q$ and 0 otherwise.

One can think of the filter as a (linear) *system* which takes the observed series X_t as input and produces the estimate of trend, $\hat{m}_t$ as output.

In addition to the choice $a_j = 1/(2q+1)$ for $|j| \leq q$, there are other choice of filters that people commonly use.

(1) **Binomial Weights**: Based on the following idea. When we are estimating the value of the trend m_t at t, it makes sense to give a higher weight to X_t compared to $X_{t\pm1}$ and a higher weight to $X_{t\pm1}$ compared to $X_{t\pm2}$ and so on. An example of such weights are:

$$a_j = 2^{-q}\binom{q}{q/2+j} \qquad \text{for } j = -q/2, -q/2+1, \ldots, -1, 0, 1, \ldots, q/2.$$

As in usual smoothing, choice of q is an issue here.

(2) **Spencer's 15 point moving average**: We have seen that simple moving average filter leaves linear functions untouched. Is it possible to design a filter which leaves higher order polynomials untouched? For example, can we come up with a filter which leaves all quadratic polynomials untouched. Yes!

For a filter with weights a_j to leave all quadratic polynomials untouched, we need the following to be satisfied for every quadratic polynomial m_t:

$$\sum_j a_j m_{t+j} = m_t \qquad \text{for all } t$$

In other words, if $m_t = \alpha t^2 + \beta t + \gamma$, we need

$$\sum_j a_j \left(\alpha(t+j)^2 + \beta(t+j) + \gamma\right) = \alpha t^2 + \beta t + \gamma \qquad \text{for all } t.$$

Simplify to get

$$\alpha t^2 + \beta t + \gamma = (\alpha t^2 + \beta t + \gamma)\sum_j a_j + (2\alpha t + \beta)\sum_j j a_j + \alpha \sum_j j^2 a_j \qquad \text{for all } t.$$

This will clearly be satisfied if

$$\sum_j a_j = 1 \qquad \sum_j j a_j = 0 \qquad \sum_j j^2 a_j = 0. \tag{2}$$

An example of such a filter is Spencer's 15 point moving average defined by

$$a_0 = \frac{74}{320}, a_1 = \frac{67}{320}, a_2 = \frac{46}{320}, a_3 = \frac{21}{320}, a_4 = \frac{3}{320}, a_5 = \frac{-5}{320}, a_6 = \frac{-6}{320}, a_7 = \frac{-3}{320}$$

and $a_j = 0$ for $j > 7$. Also the filter is symmetric in the sense that $a_{-1} = a_1, a_{-2} = a_2$ and so on. Check that this filter satisfies the condition (2).

Because this is a symmetric filter, it can be checked that it allows all cubic polynomials to pass unscathed as well.

(3) **Exponential Smoothing**: Quite a popular method of smoothing (wikipedia has a big page on this). It is also used as a forecasting technique.

To obtain $\hat{m}_t$ in this method, one uses only the *previous* observations $X_{t-1}, X_{t-2}, X_{t-3}, \ldots$. The weights assigned to these observations *exponentially decrease* the further one goes back in time. Specifically,

$$\hat{m}_t := \alpha X_{t-1} + \alpha(1-\alpha)X_{t-2} + \alpha(1-\alpha)^2 X_{t-3} + \cdots + \alpha(1-\alpha)^{t-2} X_1 + (1-\alpha)^{t-1} X_0.$$

Check that the weights add up to 1. α is a parameter that determines the amount of smoothing (α here is analogous to q in smoothing). If α is close to 1, there is very little smoothing and vice versa.

Spring 2013 Statistics 153 (Time Series) : Lecture Three

Aditya Guntuboyina

29 January 2013

1 Differencing for Trend Elimination

In the last class, we studied trend models: $X_t = m_t + Z_t$ where m_t is a deterministic trend function and $\{Z_t\}$ is white noise.

The residuals obtained after fitting the trend function m_t in the model $X_t = m_t + Z_t$ are studied to see if they are white noise or have some dependence structure that can be exploited for prediction.

Suppose that the goal is just to produce such detrended residuals. Differencing is a simple technique which produces such de-trended residuals.

One just looks at $Y_t = X_t - X_{t-1}, t = 2, \ldots, n$. If the trend m_t in $X_t = m_t + Z_t$ is linear, then this operation simply removes it because if $m_t = \alpha t + b$, then $m_t - m_{t-1} = \alpha$ so that $Y_t = \alpha + Z_t - Z_{t-1}$.

Suppose that the first differenced series Y_t appears like white noise. What then would be a reasonable forecast for the original series: X_{n+1}? Because Y_t is like white noise, we forecast Y_{n+1} by the sample mean $\bar{Y} := (Y_2 + \cdots + Y_n)/(n-1)$. But since $Y_{n+1} = X_{n+1} - X_n$, this results in the forecast $X_n + \bar{Y}$ for X_{n+1}.

Sometimes, even after differencing, one can notice a trend in the data. In that case, just difference again. It is useful to follow the notation ∇ for differencing:

$$\nabla X_t = X_t - X_{t-1} \qquad \text{for } t = 2, \ldots, n$$

and second differencing corresponds to

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla X_t - \nabla X_{t-1} = X_t - 2X_{t-1} + X_{t-2} \qquad \text{for } t = 3, \ldots, n.$$

It can be shown that quadratic trends simply disappear with the operation ∇^2. Suppose the data $\nabla^2 X_t$ appear like white noise, how would you obtain a forecast for X_{n+1}?

Differencing is a quick and easy way to produce detrended residuals and is a key component in the ARIMA forecasting models (later). A problem however is that it does not result in any estimate for the trend function m_t.

2 Random Walk with Drift Model

Consider the following model for X_t:

$$R_t = \delta + R_{t-1} + W_t$$

for $t = 1, 2, \ldots$, with initial condition $R_0 = 0$ and W_t being white noise. This model can also be written as:

$$R_t = \delta t + \sum_{j=1}^{t} W_j$$

When $\delta = 0$, this model is called the Random Walk model. This is used often to model trend. This would be a stochastic model for trend as opposed to the previous ones which are deterministic models.

Consider the model $X_t = m_t + Z_t$ where Z_t is white noise and m_t is a random walk with drift: $m_t = \delta + m_{t-1} + W_t$. This is an example of a Dynamic Linear Model (DLM). W_t is called evolution error and Z_t is called observational error.

The differenced series for X_t is:

$$\nabla X_t = X_t - X_{t-1} = m_t - m_{t-1} + Z_t - Z_{t-1} = \delta + W_t + Z_t - Z_{t-1}.$$

Therefore, ∇X_t is a detrended series. Ways for modelling detrended series through stationary models will be studied later.

3 Models for Seasonality

Many time series datasets exhibit seasonality. Simplest way to model this is: $X_t = s_t + Z_t$ where s_t is a periodic function of a known period d i.e., $s_{t+d} = s_t$ for all t. Such a function s models seasonality. These models are appropriate, for example, to monthly, quarterly or weekly data sets that have a seasonal pattern to them.

This model, however, will not be applicable for datasets having both trend and seasonality which is the more realistic situation. These will be studied a little later.

Just like the trend case, there are three different approaches to dealing with seasonality: fitting parametric functions, smoothing and differencing.

3.0.1 Fitting a parametric seasonality function

The simplest periodic functions of period d are: $a\cos(2\pi ft/d)$ and $a\sin(2\pi ft/d)$. Here f is a positive integer. The quantity a is called *Amplitude* and f/d is called *frequency* and its inverse, d/f is called *period*. The higher f is, the more rapid the oscillations in the function are.

More generally,

$$s_t = a_0 + \sum_{f=1}^{k} (a_j \cos(2\pi ft/d) + b_j \sin(2\pi ft/d)) \tag{1}$$

is a periodic function. Choose a value of k (not too large) and fit this to the data.

For $d = 12$, there is no need to consider values of k that are more than 6. With $k = 6$, every periodic function with period 12 can be written in the form (1). More on this when we study the frequency domain analysis of time series.

3.0.2 Smoothing

Because of periodicity, the function s_t only depends on the d values $s_1, s_2, \ldots, s_d$. Clearly s_1 can be estimated by the average of $X_1, X_{1+d}, X_{1+2d}, \ldots$. For example, for monthly data, this corresponds to

estimating the mean term for January by averaging all January observations. Thus

$$\hat{s}_i := \text{average of } X_i, X_{i+d}, X_{i+2d}, \ldots$$

Note that here, we are fitting 12 parameters (one each for $s_1, \ldots, s_d$) from n observations. If n is not that big, fitting 12 parameters might lead to overfitting.

3.0.3 Differencing

How can we obtain residuals adjusted for seasonality from the data without explicitly fitting a seasonality function? Recall that a function s is a periodic function of period d if $s_{t+d} = s_t$ for all t. The model that we have in mind here is: $X_t = s_t + Z_t$.

Clearly $X_t - X_{t-d} = s_t - s_{t-d} + Z_t - Z_{t-d} = Z_t - Z_{t-d}$. Therefore, the lag-$d$ differenced data $X_t - X_{t-d}$ do not display any seasonality. This method of producing deseasonalized residuals is called *Seasonal Differencing*.

4 Data Transformations

Suppose that the time series data set has a trend and that the variability increases along with the trend function. An example is the UKgas dataset in R. In such a situation, transform the data using the logarithm or a square root so that the resulting data look reasonably homoscedastic (having the same variance throughout).

Why log or square root? It helps to know a little bit about variance stabilizing transformations. Suppose X is a random variable having mean m. A *very heuristic* calculation gives an *approximate* answer for the variance of a function $f(X)$ of the random variable X? Expand $f(X)$ in its Taylor series up to *first order* around m:

$$f(X) \approx f(m) + f'(m)(X - m)$$

As a result,

$$\text{var}(f(X)) \approx \text{var}\left(f(m) + f'(m)(X - m)\right) = (f'(m))^2 \text{var}(X).$$

Thus if

1. $\text{var}(X) = Cm$ and $f(x) = \sqrt{x}$, we would get $\text{var}(X) \approx C/4$.
2. $\text{var}(X) = Cm^2$ and $f(x) = \log x$, we would get $\text{var}(X) \approx C$.

The key is to note that in both the above cases, the approximate variance of $f(X)$ does not depend on m anymore.

The above rough calculation suggests the following insight into time series data analysis. A model of the form $X_t = m_t + W_t$ where m_t is a *deterministic* function and W_t is purely random or stationary (next week) assumes that the variance of X_t does not vary with t. Suppose however that the time plot of the data shows that the variance of X_t increases with its mean m_t, say $\text{var}(X_t) \propto m_t$. Then the rough calculation suggests that $\text{var}(\sqrt{X_t})$ should be approximately constant (does not depend on t) and hence the model $m_t + W_t$ should be fit to the transformed data $\sqrt{X_t}$ instead of the original data X_t. Similarly, if $\text{var}(X_t) \propto m_t^2$, then $\text{var}(\log X_t)$ should be approximately constant.

Thus, if the data show increased variablity with a trend, then apply a transformation such as log or square root depending on whether the variability in the *resulting* data set is constant across time.

By the way, *count* data are usually modelled via Poisson random variables and the variance of a Poisson equals its mean. So one typically works with square roots while dealing with count (Poisson) data.

If one uses the model $X_t = m_t + W_t$ with a non-deterministic (stochastic) trend function m_t, this automatically allows for X_t to have a variance that changes with t. In that case, we may not need to use transformations on the data. These models can be seen as special cases of State Space Models that we will briefly look at later.

Box-Cox transformations: The square-root and the logarithm are special cases of the Box-Cox Transformations given by:

$$\begin{aligned} Y_t &= \frac{X_t^\lambda - 1}{\lambda} \qquad \text{if } \lambda \neq 0 \\ &= \log X_t \qquad \text{if } \lambda = 0. \end{aligned} \tag{2}$$

Square root essentially corresponds to $\lambda = 1/2$.

Fall 2013 Statistics 151 (Linear Models) : Lecture Four

Aditya Guntuboyina

10 September 2013

1 Recap

1.1 The Regression Problem

There is a response variable y and p explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between y and $x_1, \ldots, x_p$.

There are n subjects and data is collected on the variables from these subjects.

Data on the response variable is $y_1, \ldots, y_n$ and is represented by the column vector $Y = (y_1, \ldots, y_n)^T$ (the T here stands for transpose).

Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$. This data is represented by the $n \times p$ matrix X whose (i,j)th entry is x_{ij}. In other words, the ith row of X has data collected from the ith subject and the jth column of X has data for the jth variable.

1.2 The Linear Model

1. $y_1, \ldots, y_n$ are assumed to be random variables but x_{ij} are assumed to be non-random.
2. It is assumed that

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \qquad \text{for } i = 1, \ldots, n$$

where $e_1, \ldots, e_n$ are uncorrelated random variables with mean zero and variance σ^2.

In matrix notation, the second assumption above can be written as

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n$$

where I_n denotes the $n \times n$ identity matrix. $\beta = (\beta_1, \ldots, \beta_p)^T$ and $e = (e_1, \ldots, e_n)^T$.

If $Z_1, \ldots, Z_n$ are random variables with $Z = (Z_1, \ldots, Z_n)^T$, then $Cov(Z)$ denotes the $n \times n$ matrix whose (i,j)th entry denotes the covariance between Z_i and Z_j. In particular, the ith diagonal entry of $Cov(Z)$ would denote the variance of the random variable Z_i. Therefore, $Cov(e) = \sigma^2 I_n$ is a succinct way of saying that the covariance between e_i and e_j would equal 0 when $i \neq j$ and σ^2 when $i = j$.

2 The Intercept Term

Among other things, the linear model stipulates that

$$\mathbb{E}y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

for $i = 1, \ldots, n$. This implies that when the values of the explanatory variables $x_{i1}, \ldots, x_{ip}$ are all equal to 0, then $\mathbb{E}y_i = 0$. This is of course not always a reasonable assumption. One therefore modifies the linear model slightly by stipulating that

$$\mathbb{E}y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{1}$$

Now $\mathbb{E}y_i$ does not have to be zero when all the explanatory variables take on the value zero. The term β_0 above is known as the *intercept* term. Usually, in linear models, one **always** includes the intercept term by default.

If we let x_0 denote the "variable" which always takes the value 1, then (1) can be written as

$$\mathbb{E}y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Therefore this model with the intercept term is just the same as the previous linear model with this additional variable along with the p explanatory variables.

With the intercept term, one can write the linear model in matrix form as

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n$$

where X denotes the $n \times (p+1)$ matrix whose first column consists of all ones and the rest of the columns correspond to the values of the p explanatory variables and $\beta = (\beta_0, \ldots, \beta_p)^T$.

When $p = 1$ i.e., when there is only one explanatory variable, this linear model (with the intercept term) is called the *simple linear regression model*: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

From now on, we will always consider linear models with the intercept term which means that the first column of X (which is an $n \times (p+1)$ matrix) is always the vector of ones and β is a vector of length $p + 1$.

3 Estimation in the Linear Model

The quantities $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\sigma^2 > 0$ are *parameters* in the linear model. These need to be estimated from the data. The process of estimating the parameters is also referred to as fitting the linear model to data.

Let us first focus on the estimation of β.

The idea behind the linear model is that one tries to explain the response value y_i via the linear combination $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. It makes sense, therefore, to estimate β by the **minimizer** of the sum of squares

$$S(\beta) := \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}\right)^2.$$

Using matrix notation, this can be written as

$$S(\beta) = ||Y - X\beta||^2.$$

The norm $||x||$ of a vector $x = (x_1, \ldots, x_n)^T$ is defined as $||x|| := \sqrt{x_1^2 + \cdots + x_n^2}$. Note the equality $||x||^2 = x^T x$. Using this, we can write

$$S(\beta) = (Y - X\beta)^T (Y - X\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta.$$

This can be minimized via calculus. Take partial derivatives with respect to β_i for $i = 0, 1, \ldots, p$ and equate them to 0. It is easy to check that

$$\nabla S(\beta) = 2X^T X\beta - 2X^T Y.$$

where

$$\nabla S(\beta) = \left(\frac{\partial S(\beta)}{\partial \beta_1}, \ldots, \frac{\partial S(\beta)}{\partial \beta_p} \right).$$

denotes the gradient of $S(\beta)$ with respect to $\beta = (\beta_1, \ldots, \beta_p)^T$. It follows therefore that the minimizer of $S(\beta)$ satisfies the equality

$$X^T X\beta = X^T Y. \tag{2}$$

This gives p linear equations for the p unknowns $\beta_1, \ldots, \beta_p$. This important set of equations are called *normal equations*. Their solution, denoted by $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ gives an estimate of β called the least squares estimate. If the values of the p explanatory variables for a subject are $\lambda_1, \ldots, \lambda_p$, then the estimate of his mean response is given by $\hat{\beta}_0 + \lambda_1 \hat{\beta}_1 + \cdots + \lambda_p \hat{\beta}_p$.

Two important questions arise are: (1) **Does there exist a solution to the normal equations** and (2) **If yes, then is the solution unique?**

The answer to the first question is **yes**. The normal equations always have a solution. The reason is the following: $X^T Y$ lies in the column space of X^T. Further, the column spaces of X^T and $X^T X$ are identical and thus $X^T Y$ can always be written as $X^T X u$ for some vector u.

The answer to the second question is **yes if $X^T X$ is invertible** and **no if $X^T X$ is not invertible**.

Do the normal equations (2) admit a *unique* solution? Answer: **No in general. Yes if $X^T X$ is invertible**.

If $X^T X$ is invertible, then the solution to the normal equations is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$. The estimate of the linear function $\lambda^T \beta$ for a vector $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ is then given by $\lambda^T (X^T X)^{-1} X^T Y$.

If $X^T X$ is not invertible, then the normal equations have many solutions. In this case, how does one estimate β? Here, it actually turns out that the vector β **cannot** be estimated. This is explained next.

4 When $X^T X$ is not necessarily invertible

The vector β cannot be estimated in this case.

Observe first that $X^T X$ being invertible is equivalent to the rank of X being equal to $p + 1$. Thus when $X^T X$ is not invertible, the rank of X is strictly smaller than $p + 1$. In other words, some column of X is a linear combination of the rest of the columns of X i.e., at least one of explanatory variables is redundant in the sense that it can be written as a linear combination of the other explanatory variables.

Let us consider an example here. Suppose $p = 2$ and that the two explanatory variables x_1 and x_2 are actually the same i.e., $x_{i1} = x_{i2}$ for each $i = 1, \ldots, n$. It should be clear then that the rank of X is at most 2. The linear model can then be written as

$$y_i = \beta_0 + (\beta_1 + \beta_2)\, x_{i1} + \epsilon_i$$

for $i = 1, \ldots, n$. It should be clear that from these observations, the parameters β_1 and β_2 **cannot** be estimated. On the other hand, $\beta_1 + \beta_2$ can be estimated.

Thus when X^TX is not invertible, the parameter vector β cannot be estimated while certain special linear combinations can be estimated.

It can be shown that a linear combination $\lambda^T\beta$ can be estimated if and only if λ lies in the column space of X^T. This is equivalent to saying that λ lies in the column space of X^TX because the column spaces of X^T and X^TX are always equal.

In the example just discussed, the vector $(0,1,1)^T$ is in the column space of X^T which implies that $\beta_1+\beta_2$ is estimable. On the other hand, the vector $(0,1,0)^T$ is not in the column space of X^T which implies that β_1 is not estimable.

When X^TX is invertible, then the column space of X^T contains all $(p+1)$ dimensional vectors and then every linear combination of β is estimable.

5 Least Squares Estimates

Consider the normal equations $X^TX\beta = X^TY$. Let $\hat{\beta}_{ls}$ denote any solution (it is unique only if X^TX is invertible).

Let $\lambda^T\beta$ be estimable (i.e., λ lies in the column space of X^T or equivalently the column space of X^TX). Then estimate $\lambda^T\beta$ by $\lambda^T\hat{\beta}_{ls}$. This is called the least squares estimate of $\lambda^T\beta$.

Result 5.1. *If $\lambda^T\beta$ is estimable, then $\lambda^T\hat{\beta}_{ls}$ is the same for every solution $\hat{\beta}_{ls}$ of the normal equations. In other words, the least squares estimate of $\lambda^T\beta$ is unique.*

Proof. Since $\lambda^T\beta$ is estimable, the vector λ lies in the column space of X^TX and hence $\lambda = X^TXu$ for some vector u. Therefore,

$$\lambda^T\hat{\beta}_{ls} = u^TX^TX\hat{\beta}_{ls} = u^TX^TY$$

where the last equality follows from the fact that $\hat{\beta}_{ls}$ satisfies the normal equations. Since u only depends on λ, this proves that $\lambda^T\hat{\beta}_{ls}$ does not depend on the particular choice of the solution $\hat{\beta}_{ls}$ of the normal equations. □

Thus when $\lambda^T\beta$ is estimable, it is estimated by the least squares estimate $\lambda^T\hat{\beta}_{ls}$ (which is uniquely defined). When $\lambda^T\beta$ is not estimable, it of course does not make sense to try to estimate it.

Fall 2013 Statistics 151 (Linear Models) : Lecture Five

Aditya Guntuboyina

12 September 2013

1 Least Squares Estimate of β in the linear model

The linear model is

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n$$

where Y is $n \times 1$ vector containing all the values of the response, X is $n \times (p+1)$ matrix containing all the values of the explanatory variables (the first column of X is all ones) and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ (β_0 is the intercept).

As we have seen last time, β is estimated by minimizing $S(\beta) = ||Y - X\beta||^2$. Taking derivatives with respect to β and equating to zero, one obtains the normal equations

$$X^T X\beta = X^T Y.$$

If $X^T X$ is invertible (this is equivalent to the rank of X being equal to $p+1$), then the solution to the normal equations is unique and is given by

$$\hat{\beta} := (X^T X)^{-1} X^T Y$$

This is the least squares estimate of β.

2 Special Case: Simple Linear Regression

Suppose there is only one explanatory variable x. The matrix X would then of size $n \times 2$ where the first column of X consists of all ones and the second column of X equals the values of the explanatory variable $x_1, \ldots, x_n$. Therefore

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Check that

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

where $\bar{x} = \sum_i x_i/n$. Also let $\bar{y} = \sum_i y_i/n$. Because

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

we get

$$(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}.$$

Also

$$X^TY = \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Therefore

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

Simplify to obtain

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{pmatrix}.$$

Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \frac{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{y} - \hat{\beta}_1 \bar{x}$$

If we get a new subject whose explanatory variable value is x, our prediction for its response is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{1}$$

If the predictions given by the above are plotted on a graph (with x plotted on the x-axis), then one gets a line called the **Regression Line**.

The Regression Line has a much nicer expression than (1). To see this, note that

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \bar{x}\hat{\beta}_1 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

This can be written as

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} (x - \bar{x}) \tag{2}$$

Using the notation

$$r := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad s_x := \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}, \quad s_y := \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2},$$

we can rewrite the prediction equation (2) as

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}. \tag{3}$$

r is the correlation between x and y which is always between -1 and 1.

As an implication, note that if $(x - \bar{x})/s_x = 1$ i.e., if the explanatory variable value of the subject is one standard deviation above the sample mean, then its response variable is predicted to be only r standard deviations above its mean. Francis Galton termed this **regression to mediocrity** which is where the name regression comes from.

3 Basic Mean and Covariance Formulae for Random Vectors

We next want to explore properties of $\hat{\beta} = (X^T X)^{-1} X^T Y$ as an estimator of β in the linear model. For this we need a few facts about means and covariances.

Let $Z = (Z_1, \ldots, Z_k)^T$ be a random vector. Its expectation $\mathbb{E}Z$ is defined as a vector whose ith entry is the expectation of Z_i i.e., $\mathbb{E}Z = (\mathbb{E}Z_1, \mathbb{E}Z_2, \ldots, \mathbb{E}Z_k)^T$.

The covariance matrix of Z, denoted by $Cov(Z)$, is a $k \times k$ matrix whose (i, j)th entry is the covariance between Z_i and Z_j.

If $W = (W_1, \ldots, W_m)^T$ is another random vector, the covariance matrix between Z and W, denoted by $Cov(Z, W)$, is a $k \times m$ matrix whose (i, j)th entry is the covariance between Z_i and W_j. Note then that, $Cov(Z, Z) = Cov(Z)$.

The following formulae are very important:

1. $\mathbb{E}(AZ + c) = A\mathbb{E}(Z) + c$ for any constant matrix A and any constant vector c.
2. $Cov(AZ + c) = ACov(Z)A^T$ for any constant matrix A and any constant vector c.
3. $Cov(AZ + c, BW + d) = ACov(Z, W)B$ for any pair of constant matrices A and B and any pair of constant vectors c and d.

The linear model is

$$Y = X\beta + e \qquad \text{with } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 I_n.$$

Because of the above formulae (remember that X and β are fixed),

$$\mathbb{E}Y = X\beta \quad \text{and} \quad Cov(Y) = \sigma^2 I_n.$$

4 Properties of the Least Squares Estimator

Assume that $X^T X$ is invertible (equivalently, that X has rank $p + 1$) and consider the least squares estimator

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

What properties does $\hat{\beta}$ have as an estimator of β?

4.1 Linearity

An estimator of β is said to be linear if it can be written as AY for some matrix A. Clearly $\hat{\beta} = (X^T X)^{-1} X^T Y$ is of this form and hence it is a linear estimator of β.

4.2 Unbiasedness

An estimator for a parameter is said to be unbiased if its expectation equals the parameter (for all values of the parameter).

The expectation of the least squares estimator is (using the formula for expectation: $\mathbb{E}AZ = A\mathbb{E}Z$)

$$\mathbb{E}\hat{\beta} = \mathbb{E}((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T \mathbb{E}Y = (X^T X)^{-1} X^T X\beta = \beta$$

In particular, this means that $\mathbb{E}\hat{\beta}_i = \beta_i$ for each i which implies that each $\hat{\beta}_i$ is an unbiased estimator of β_i. More generally, for every vector λ, the quantity $\lambda^T\hat{\beta}$ is an unbiased estimator of $\lambda^T\beta$.

4.3 Covariance Matrix

The Covariance matrix of the estimator $\hat{\beta}$ can be easily calculated using the formula: $Cov(AZ) = ACov(Z)A^T$:

$$Cov(\hat{\beta}) = Cov((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^TCov(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

In particular, the variance of $\hat{\beta}_i$ equals σ^2 multiplied by the ith diagonal element of $(X^TX)^{-1}$. Once we learn how to estimate σ, we can use this to obtain standard errors for $\hat{\beta}_i$.

4.4 Optimality - The Gauss-Markov Theorem

The Gauss-Markov Theorem states that $\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator). This means that $\hat{\beta}$ is the "best" estimator among all **linear and unbiased** estimators of β. Here, "best" is in terms of variance. This implies that $\hat{\beta}_i$ has the **smallest variance** among all linear and unbiased estimators of β_i.

Fall 2013 Statistics 151 (Linear Models) : Lecture Six

Aditya Guntuboyina

17 September 2013

We again consider $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I_n$. β is estimated by solving the normal equations $X^T X\beta = X^T Y$.

1 The Regression Plane

If we get a new subject whose explanatory variable values are $x_1, \ldots, x_p$, then our prediction for its response variable value is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_p x_p$$

This equation represents a plane which we call the regression plane.

2 Fitted Values

These are the values predicted by the linear model for the n subjects.

The values of the explanatory variables are $x_{i1}, \ldots, x_{ip}$ for the ith subject. Thus the linear model prediction for the ith subject is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

Because the value of the response variable for the ith subject is y_i, it makes sense to call the above prediction $\hat{y}_i$. Thus

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \qquad \text{for } i = 1, \ldots, n.$$

These values $\hat{y}_1, \ldots, \hat{y}_n$ are called fitted values and the vector $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ is called the vector of fitted values. This vector can be written succinctly as $\hat{Y} = X\hat{\beta}$. Because $\hat{\beta} = (X^T X)^{-1} X^T Y$, we can write

$$\hat{Y} = X(X^T X)^{-1} X^T Y.$$

The vector of fitted values $\hat{Y}$ is the (orthogonal) projection of Y onto the column space of X.

Let $H = X(X^T X)^{-1} X^T$ so that $\hat{Y} = HY$. Because multiplication by H changes Y into $\hat{Y}$, the matrix H is called the **Hat Matrix**. It is very important in linear regression. It has the following three easily verifiable properties:

1. It is a symmetric $n \times n$ matrix.
2. It is idempotent i.e., $H^2 = H$.
3. $HX = X$.

4. The ranks of H and X are the same.

These can be easily derived from the definition $H = X(X^TX)^{-1}X^T$. Because of these, we get

$$\mathbb{E}\hat{Y} = \mathbb{E}(HY) = H(\mathbb{E}Y) = HX\beta = X\beta = \mathbb{E}Y. \tag{1}$$

Thus $\hat{Y}$ and Y have the same expectation. Also

$$Cov(\hat{Y}) = Cov(HY) = HCov(Y)H^T = H(\sigma^2 I)H = \sigma^2 H.$$

3 Residuals

The difference between y_i and $\hat{y}_i$ is called the residual for the ith subject. $\hat{e}_i := y_i - \hat{y}_i$. The vector $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)^T$ is called the vector of residuals. Clearly

$$\hat{e} = Y - \hat{Y} = (I - H)Y.$$

The vector of residuals $\hat{e}$ acts as a proxy for the **unobserved error vector** e.

The most important fact about the residuals in the linear model is that they are orthogonal to the column space of X. This happens because $HX = X$ so that

$$\hat{e}^TX = ((I - H)Y)^TX = Y^T(I - H)X = Y^T(X - HX) = 0.$$

As a result $\hat{e}^TXu = 0$ for every vector u which means that $\hat{e}$ is orthogonal to the column space of X.

The first column of X consists of ones. Because $\hat{e}$ is orthogonal to everything in the column space of X, it must therefore be orthogonal to the vector of ones which means that

$$\sum_{i=1}^{n} \hat{e}_i = 0.$$

$\hat{e}$ is also orthogonal to every column of X:

$$\sum_{i=1}^{n} \hat{e}_i x_{ij} = 0 \qquad \text{for every } j$$

The vector of fitted values belongs to the column space of X because $\hat{Y} = X\hat{\beta}$. Thus, $\hat{e}$ is also orthogonal to $\hat{Y}$.

Because $X^T\hat{e} = 0$, the residuals satisfy $rank(X) = p + 1$ linear equalities. Hence, although there are n of them, they are effectively $n - p - 1$ of them. The number $n - p - 1$ is therefore referred to as the *degrees of freedom* of the residuals $\hat{e}_1, \ldots, \hat{e}_n$.

The expectation of $\hat{e}$ is

$$\mathbb{E}\hat{e} = \mathbb{E}\left((I - H)Y\right) = (I - H)(\mathbb{E}Y) = (I - H)X\beta = (X - HX)\beta = 0.$$

Alternatively $\mathbb{E}\hat{e} = \mathbb{E}(Y - \hat{Y}) = \mathbb{E}Y - \mathbb{E}\hat{Y} = 0$ by (1).

The Covariance matrix of $\hat{e}$ is

$$Cov(\hat{e}) = Cov((I - H)Y) = (I - H)Cov(Y)(I - H) = \sigma^2(I - H). \tag{2}$$

Note that the residuals have different variances.

4 The Residual Sum of Squares

The sum of squares of the residuals is called RSS:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}^T \hat{e} = Y^T (I - H) Y = e^T (I - H) e.$$

What is the residual sum of squares where there are no explanatory variables in the model (the model in this case only contains the intercept term)? Ans: $\sum_{i=1}^{n} (y_i - \bar{y})^2$ where $\bar{y} = (y_1 + \cdots + y_n)/n$. This quantity is called the TSS (Total Sum of Squares). The vector $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$ has $n - 1$ degrees of freedom (because this is a vector of size n and it satisfies the linear constraint that sum is zero).

What is the residual sum of squares in simple linear regression (when there is exactly one explanatory variable)? Check that in simple linear regression:

$$RSS = \left(1 - r^2\right) \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where r is the sample correlation between $(y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$. Because $1 - r^2 \leq 1$, the RSS in simple linear regression is smaller than the RSS in the linear model with no explanatory variables.

In general, RSS decreases (or remains the same) as we add more explanatory variables to the model.

5 The Coefficient of Determination

This is more commonly referred to as R-squared. It is defined as

$$R^2 := 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

What is the point of this definition? One of the goals of regression is to predict the value of the response variable for future subjects. For this purpose, we are given data $y_1, \ldots, y_n$ and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Suppose we are told to predict the response of a future subject **without** using any of the data on the explanatory variables i.e., we are only supposed to use $y_1, \ldots, y_n$. In this case, it is obvious that our prediction for the next subject would be $\bar{y}$. The error of this method of prediction on the ith subject is $y_i - \bar{y}$ and the total error is therefore:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

If, on the other hand, we are allowed to use data on the explanatory variables, then the prediction will be given by

$$\hat{\beta}_0 + x_1 \hat{\beta}_1 + \cdots + x_p \hat{\beta}_p.$$

The error of this prediction on the ith subject is the residual $\hat{e}_i$ and the total error is the Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Because using the explanatory variables is always better than not using them, RSS is always smaller than or equal to TSS (**this fact is crucially reliant on the fact that there is an intercept in our model**).

If RSS is very small compared to TSS, it means that the explanatory variables are really useful in predicting the response. On the other hand, if RSS is only a little bit smaller than TSS, it means that we are not really gaining much by using the explanatory variables. The quantity R^2 tries to quantify how useful the explanatory variables are in predicting the response. It always lies between 0 and 1

1. **If R^2 is high, it means that RSS is much smaller compared to TSS and hence the explanatory variables are really useful in predicting the response.**
2. **If R^2 is low, it means that RSS is only a little bit smaller than TSS and hence the explanatory variables are not useful in predicting the response.**

It must be noted that R^2 is an *in-sample* measure of prediction accuracy. In other words, the predictions are checked on the subjects already present in the sample (as opposed to checking them on new subjects). In particular, these are the same subjects on whom the model is fitted (or trained), so R^2 can be made to look very good by fitting models with lots of parameters.

Because RSS decreases when more parameters are added to the model, R^2 increases when more parameters are added to the model.

Fall 2013 Statistics 151 (Linear Models) : Lecture Seven

Aditya Guntuboyina

19 September 2013

1 Last Class

We looked at

1. **Fitted Values**: $\hat{Y} = X\hat{\beta} = HY$ where $H = X(X^TX)^{-1}X^TY$. $\hat{Y}$ is the projection of Y onto the column space of X.
2. **Residuals**: $\hat{e} = Y - \hat{Y} = (I - H)Y$. $\hat{e}$ is orthogonal to **every vector** in the column space of X. The degrees of freedom of the residuals is $n - p - 1$.
3. **Residual Sum of Squares**: $RSS = \sum_{i=1}^n \hat{e}_i^2 = \hat{e}_i^T\hat{e}_i Y^T(I - H)Y$. RSS decreases when more explanatory variables are added to the model.
4. **Total Sum of Squares**: $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Can be thought of the RSS in a linear model with no explanatory variables (only the intercept term).
5. **Coefficient of Determination or Multiple** R^2: Defined as $1-(RSS/TSS)$. Always lies between 0 and 1. High value means that the explanatory variables are useful in explaining the response and low value means that the explanatory variables are not useful in explaining the response. R^2 increases when more explanatory variables are added to the model.

2 Expected Value of the RSS

What is the expected value of RSS?

$$\mathbb{E}(RSS) = \mathbb{E}e^T(I - H)e = \mathbb{E}\left(\sum_{i,j}(I - H)(i,j)e_ie_j\right) = \sum_{i,j}(I - H)(i,j)(\mathbb{E}e_ie_j)$$

Because $\mathbb{E}(e_ie_j)$ equals 0 when $i \neq j$ and σ^2 otherwise, we get

$$\mathbb{E}(RSS) = \sigma^2\sum_{i=1}^n(I - H)(i,i) = \sigma^2\left(n - \sum_{i=1}^n H(i,i)\right)$$

The sum of the diagonal entries of a square matrix is called its trace i.e, $tr(A) = \sum_i a_{ii}$. We can therefore write

$$\mathbb{E}(RSS) = \sigma^2\left(n - tr(H)\right).$$

A very important fact about trace is $tr(AB) = tr(BA)$. Thus

$$tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_{p+1}) = p + 1.$$

We proved

$$\mathbb{E}(RSS) = \sigma^2(n-p-1).$$

An *unbiased* estimator of σ^2 is therefore given by

$$\hat{\sigma^2} := \frac{RSS}{n-p-1}.$$

And σ is estimated by

$$\hat{\sigma} := \sqrt{\frac{RSS}{n-p-1}}.$$

This $\hat{\sigma}$ is called the **Residual Standard Error**.

3 Standard Errors of $\hat{\beta}$

We have seen that $\mathbb{E}\hat{\beta} = \beta$ and that $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. The standard error of $\hat{\beta}_i$ is therefore defined as $\hat{\sigma}$ multiplied by the square root of the ith diagonal entry of $(X^TX)^{-1}$. The standard error gives an idea of the accuracy of $\hat{\beta}_i$ as an estimator of β_i. These standard errors are part of the R output for the summary of the linear model.

4 Standardized or Studentized Residuals

The residuals $\hat{e}_1, \ldots, \hat{e}_n$ have different variances. Indeed, because $Cov(\hat{e}) = \sigma^2(I-H)$, we have

$$var(\hat{e}_i) = \sigma^2(1-h_{ii})$$

where h_{ii} denotes the ith diagonal entry of H. Because h_{ii} can be different for different i, the residuals have different variances.

The variance can be standardized to 1 if we divide the residuals by $\sigma\sqrt{1-h_{ii}}$. But because σ is unknown, one divides by $\hat{\sigma}\sqrt{1-h_{ii}}$ and we call the resulting quantities **Standardized Residuals** or **Studentized Residuals**:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

The standardized residuals $r_1, \ldots, r_n$ are very important in regression diagnostics. Various assumptions on the unobserved errors $e_1, \ldots, e_n$ can be checked through them.

5 Normality of the Errors

Everything that we did so far was only under the assumption that the errors $e_1, \ldots, e_n$ were uncorrelated, had mean zero and variance σ^2. But if we want to test hypotheses about or if we want confidence intervals for linear combinations of β, we need distributional assumptions on the errors.

For example, consider the problem of testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. If H_0 were true, this would mean that the first explanatory variable has no role (in the presence of the other explanatory variables) in determining the expected value of the response. An obvious way to test this hypothesis is to look at the value of $\hat{\beta}_1$ and then to reject H_0 if $|\hat{\beta}_1|$ is large. But how large? To answer this question, we need to understand how $\hat{\beta}_1$ is distributed under the null hypothesis H_0. Such a study requires some distributional assumptions on the errors $e_1, \ldots, e_n$.

The mose standard assumption on the errors is that $e_1, \ldots, e_n$ are independently distributed according to the normal distribution with mean zero and variance σ^2. This is written in multivariate normal notation as $e \sim N(0, \sigma^2 I_n)$.

6 The Multivariate Normal Distribution

A random vector $U = (U_1, \ldots, U_p)^T$ is said to have the multivariate normal distribution with parameters μ and Σ if the joint density of $U_1, \ldots, U_p$ is given by

$$(2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(\frac{-1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu)\right) \qquad \text{for } u \in \mathbb{R}^d.$$

Here $|\Sigma|$ denotes the determinant of Σ.

We use the notation $U \sim N_p(\mu, \Sigma)$ to express that U is multivariate normal with parameters μ and Σ.

Example 6.1. *An important example of the multivariate normal distribution occurs when $U_1, \ldots, U_p$ are independently distributed according to the normal distribution with mean 0 and variance σ^2. In this case, it is easy to show $U = (U_1, \ldots, U_p)^T \sim N_p(0, \sigma^2 I_p)$.*

The most important properties of the multivariate normal distribution are summarized below:

1. When $p = 1$, this is just the usual normal distribution.
2. **Mean and Variance-Covariance Matrix**: $\mathbb{E}U = \mu$ and $Cov(U) = \Sigma$.
3. **Independence of linear functions can be checked by multiplying matrices**: Two linear functions AU and BU are independent if and only if $A\Sigma B^T = 0$. In particular, this means that U_i and U_j are independent if and only if the (i, j)th entry of Σ equals 0.
4. Every linear function is also multivariate normal: $a + AU \sim N(a + A\mu, A\Sigma A^T)$.
5. Suppose $U \sim N_p(\mu, I)$ and A is a $p \times p$ symmetric and idempotent (symmetric means $A^T = A$ and idempotent means $A^2 = A$) matrix. Then $(U-\mu)^T A(U-\mu)$ has the chi-squared distribution with degrees of freedom equal to the rank of A. This is written as $(U-\mu)^T A(U-\mu) \sim \chi^2_{rank(A)}$.

7 Normal Regression Theory

We assume that $e \sim N_n(0, \sigma^2 I_n)$. Equivalently, $e_1, \ldots, e_n$ are independent normals with mean 0 and variance σ^2.

Under this assumption, we can calculate the distributions of many of the quantities studied so far.

7.1 Distribution of Y

Since $Y = X\beta + e$, we have $Y \sim N_n(X\beta, \sigma^2 I_n)$.

7.2 Distribution of $\hat{\beta}$

Because $\hat{\beta} = (X^T X)^{-1} X^T Y$ is a linear function of Y, it has a multivariate normal distribution. We already saw that $\mathbb{E}\hat{\beta} = \beta$ and $Cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$. Thus $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^T X)^{-1})$.

7.3 Distribution of Fitted Values

$\hat{Y} = HY$. Thus $\mathbb{E}\hat{Y} = H\mathbb{E}(Y) = HX\beta = X\beta$. Also $Cov(\hat{Y}) = Cov(HY) = \sigma^2 H$. Therefore $\hat{Y} \sim N_n(X\beta, \sigma^2 H)$.

7.4 Distribution of Residuals

$\hat{e} = (I - H)Y$. We saw that $\mathbb{E}\hat{e} = 0$ and $Cov(\hat{e}) = \sigma^2(I - H)$. Therefore $\hat{e} \sim N_n(0, \sigma^2(I - H))$.

7.5 Independence of residuals and $\hat{\beta}$

Recall that if $U \sim N_p(\mu, \Sigma)$, then AU and BU are independent if and only if $A\Sigma B^T$.

This can be used to verify that $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{e} = (I - H)Y$ are independent. To see this, observe that both are linear functions of $Y \sim N_n(X\beta, \sigma^2 I)$. Thus if $A = (X^T X)^{-1} X^T Y$, $B = (I - H)$ and $\Sigma = \sigma^2 I$, then

$$A\Sigma B^T = \sigma^2 (X^T X)^{-1} X^T (I - H) = \sigma^2 (X^T X)^{-1} (X^T - X^T H)$$

Because $X^T H = (HX)^T = X^T$, we conclude that $\hat{\beta}$ and $\hat{e}$ are independent.

Also check that $\hat{Y}$ and $\hat{e}$ are independent.

7.6 Distribution of RSS

Recall

$$RSS = \hat{e}^T \hat{e} = Y^T (I - H) Y = e^T (I - H) e.$$

So

$$\frac{RSS}{\sigma^2} = \left(\frac{e}{\sigma}\right)^T (I - H) \left(\frac{e}{\sigma}\right).$$

Because $e/\sigma \sim N_n(0, I)$ and $I - H$ is symmetric and idempotent with rank $n - p - 1$, we have

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Fall 2013 Statistics 151 (Linear Models) : Lecture Eight

Aditya Guntuboyina

24 September 2013

1 Normal Regression Theory

We assume that $e \sim N_n(0, \sigma^2 I_n)$. Equivalently, $e_1, \ldots, e_n$ are independent normals with mean 0 and variance σ^2. As a result of this assumption, we can calculate the following:

1. **Distribution of** Y: Since $Y = X\beta + e$, we have $Y \sim N_n(X\beta, \sigma^2 I_n)$.
2. **Distribution of** $\hat{\beta}$: $\hat{\beta} = (X^T X)^{-1} X^T Y \sim N_{p+1}(\beta, \sigma^2 (X^T X)^{-1})$.
3. **Distribution of Residuals**: $\hat{e} = (I - H)Y$. We saw that $\mathbb{E}\hat{e} = 0$ and $Cov(\hat{e}) = \sigma^2(I - H)$. Therefore $\hat{e} \sim N_n(0, \sigma^2(I - H))$.
4. **Independence of residuals and** $\hat{\beta}$: Recall that if $U \sim N_p(\mu, \Sigma)$, then AU and BU are independent if and only if $A\Sigma B^T$.

   This can be used to verify that $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{e} = (I - H)Y$ are independent. To see this, observe that both are linear functions of $Y \sim N_n(X\beta, \sigma^2 I)$. Thus if $A = (X^T X)^{-1} X^T Y$, $B = (I - H)$ and $\Sigma = \sigma^2 I$, then

   $$A\Sigma B^T = \sigma^2 (X^T X)^{-1} X^T (I - H) = \sigma^2 (X^T X)^{-1} (X^T - X^T H)$$

   Because $X^T H = (HX)^T = X^T$, we conclude that $\hat{\beta}$ and $\hat{e}$ are independent.

   Also check that $\hat{Y}$ and $\hat{e}$ are independent.
5. **Distribution of RSS**: $RSS = \hat{e}^T \hat{e} = Y^T (I - H) Y = e^T (I - H) e$. So

   $$\frac{RSS}{\sigma^2} = \left(\frac{e}{\sigma}\right)^T (I - H) \left(\frac{e}{\sigma}\right).$$

   Because $e/\sigma \sim N_n(0, I)$ and $I - H$ is symmetric and idempotent with rank $n - p - 1$, we have

   $$\frac{RSS}{\sigma^2} \sim \chi^2_{n-p-1}.$$

2 How to test $H_0 : \beta_j = 0$

The are two equivalent ways of testing this hypothesis.

2.1 First Test: t-test

It is natural to base the test on the value of $\hat{\beta}_j$ i.e., reject if $|\hat{\beta}_j|$ is large. How large? To answer this, we need to look at the distribution of $\hat{\beta}_j$ under H_0 (called the null distribution). Under normality of the errors, we have seen that $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^TX)^{-1})$. In other words,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 v_j)$$

where v_j is the jth diagonal entry of $(X^TX)^{-1}$. Under the null hypothesis, when $\beta_j = 0$, we thus have

$$\frac{\hat{\beta}_j}{\sigma\sqrt{v_j}} \sim N(0,1).$$

This can be used to construct a test but the problem is that σ is unknown. One therefore replaces it by the estimate $\hat{\sigma}$ to construct the test statistic:

$$\frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} = \frac{\hat{\beta}_j/\sigma\sqrt{v_j}}{\hat{\sigma}/\sigma} = \frac{\hat{\beta}_j/\sigma\sqrt{v_j}}{\sqrt{RSS/(n-p-1)\sigma^2}}$$

Now the numerator here is $N(0,1)$. The denominator is $\sqrt{\chi^2_{n-p-1}/(n-p-1)}$. Moreover, the numerator and the denominator are independent. Therefore, we get

$$\frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \sim t_{n-p-1}$$

where t_{n-p-1} denotes the t-distribution with $n-p-1$ degrees of freedom.

p-value for testing $H_0 : \beta_j = 0$ can be got by

$$\mathbb{P}\left(|t_{n-p-1}| > \left|\frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)}\right|\right).$$

Note that when $n-p-1$ is large, the t-distribution is almost the same as a standard normal distribution.

2.2 Second Test: F-test

We have just seen how to test the hypothesis $H_0 : \beta_j = 0$ using the statistic $\hat{\beta}_j/s.e(\hat{\beta}_j)$ and the t-distribution.

Here is another natural test for this problem. The null hypothesis H_0 says that the explanatory variable x_j can be dropped from the linear model. Let us call this reduced model m.

Also, let us call the original model M (this is the full model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$).

The following presents another natural test for H_0. Let the Residual Sum of Squares in the model m be denoted by $RSS(m)$ and let the RSS in the full model be $RSS(M)$. It is always true that $RSS(M) \leq RSS(m)$. Now if $RSS(M)$ is *much smaller* than $RSS(m)$, it means that the explanatory variable x_j contributes a lot to the regression and hence cannot be dropped i.e., we reject the null hypothesis H_0. On the other hand, if $RSS(M)$ is *only a little smaller* than $RSS(m)$, then x_j does not really contribute a lot in predicting y and hence can be dropped i.e., we do not reject H_0.

Therefore one can test H_0 via the test statistic:

$$RSS(m) - RSS(M)$$

We would reject the null hypothesis if this is large. How large? To answer this, we need to look at the **null distribution** of $RSS(m) - RSS(M)$. We show (in the next class) that

$$\frac{RSS(m) - RSS(M)}{\sigma^2} \sim \chi^2_1$$

under the null hypothesis. Since we do not know σ^2, we estimate it by

$$\hat{\sigma}^2 = \frac{RSS(M)}{n-p-1},$$

to obtain the test statistic:

$$\frac{RSS(m) - RSS(M)}{RSS(M)/(n-p-1)}$$

The numerator and the denominator are independent (to be shown in the next class). This independende will not hold if the denominator were $RSS(m)/(n-p)$. Thus under the null hypothesis

$$\frac{RSS(m) - RSS(M)}{RSS(M)/(n-p-1)} \sim F_{1,n-p-1}.$$

p-value can therefore be got by

$$\mathbb{P}\left(F_{1,n-p-1} > \frac{RSS(m) - RSS(M)}{RSS(M)/(n-p-1)}\right).$$

2.3 Equivalence of These Two Tests

It turns out that these two tests for testing $H_0 : \beta_j = 0$ are equivalent in the sense that they give the same p-value. This is because

$$\left(\frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)}\right)^2 = \frac{RSS(m) - RSS(M)}{RSS(M)/(n-p-1)}$$

This is not very difficult to prove but we shall skip its proof.

Fall 2013 Statistics 151 (Linear Models) : Lecture Nine

Aditya Guntuboyina

26 September 2013

1 Hypothesis Tests to Compare Models

Let M denote the full regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$$

which has p explanatory variables.

Let m denote a sub-model of M that is obtained by a linear constraint on the parameter $\beta = (\beta_0, \ldots, \beta_p)$ of M. Examples:

1. For the constraint $\beta_1 = 0$, the model m becomes: $y_i = \beta_0 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$.
2. For $\beta_1 = \beta_2 = 0$, the model m becomes $y_i = \beta_0 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + e_i$.
3. For $\beta_1 = \cdots = \beta_p = 0$, the model m becomes $y_i = \beta_0 + e_i$.
4. For $\beta_1 = \beta_2$, the model m becomes $y_i = \beta_0 + \beta_1(x_{i1} + x_{i2}) + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + e_i$.
5. For $\beta_1 = 3$, the model m becomes $y_i = \beta_0 + 3x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$.

How do we test the hypothesis $H_0 : m$ against $H_1 : M$? Let q be the number of explanatory variables in m. This test can be carried out by noting that if $RSS(m) - RSS(M)$ is large, then m is not a good model for the data and we therefore reject H_0. On the other hand, if $RSS(m) - RSS(M)$ is small, we do not reject H_0.

The test is therefore based on $RSS(m) - RSS(M)$. What is the distribution of this quantity under the null bypothesis? We will show that

$$\frac{RSS(m) - RSS(M)}{\sigma^2} \sim \chi^2_{p-q}$$

where q is the number of explanatory variables in m and p is the number of explanatory variables in M. Let us now prove this fact. Let the hat matrices in the two models be denoted by $H(m)$ and $H(M)$ respectively. Write

$$RSS(m) = Y^T(I - H(m))Y \quad \text{and} \quad RSS(M) = Y^T(I - H(M))Y$$

so that

$$RSS(m) - RSS(M) = Y^T(H(M) - H(m))Y.$$

We need the null distribution of $RSS(m) - RSS(M)$. So we shall assume that $Y = X(m)\beta(m) + e$ (where $X(m)$ is the X-matrix in the model m). It is important to realize that $H(m)X(m) = X(m)$ and also $H(M)X(m) = X(m)$. So

$$RSS(m) - RSS(M) = e^T(H(M) - H(m))e.$$

$H(M) - H(m)$ is a symmetric $n \times n$ matrix of rank $p - q$. It is also idempotent because

$$(H(M) - H(m))^2 = H(M) + H(m) - 2H(M)H(m). \tag{1}$$

Now $H(M)H(m) = H(m)$. To see this, it is enough to show that $H(M)H(m)v = H(m)v$ for every vector v. Now recall that $H(m)v$ denotes the projection of v onto the column space of $X(m)$. And $H(M)H(m)v$ projects $H(m)v$ projects onto the column space of $X(M)$ (which equals the original X matrix). But because the column space of $X(m)$ is contained in the column space of $X(M)$, it follows that $H(m)v$ is already contained in the column space of $X(M)$. Thus its projection onto the column space of $X(M)$ equals itself. So $H(M)H(m)v = H(m)v$.

Because $H(M)H(m) = H(m)$, it follows from (1) that

$$(H(M) - H(m))^2 = H(M) - H(m).$$

Therefore, under the null hypothesis

$$\frac{RSS(m) - RSS(M)}{\sigma^2} = \left(\frac{e}{\sigma}\right)^T (H(M) - H(m))\frac{e}{\sigma} \sim \chi^2_{p-q}.$$

Since we do not know σ^2, we estimate it by

$$\hat{\sigma}^2 = \frac{RSS(M)}{n - p - 1},$$

to obtain the test statistic:

$$\frac{RSS(m) - RSS(M)}{RSS(M)/(n - p - 1)}$$

The numerator and the denominator are independent because $RSS(m) - RSS(M) = Y^T(H(M) - H(m))Y$ and $RSS(M) = Y^T(I - H(M))Y$ and the product of the matrices

$$(H(M) - H(m))\,(I - H(M)) = H(M) - H(M)^2 - H(m) + H(m)H(M) = H(M) - H(M) - H(m) + H(m) = 0.$$

Thus under the null hypothesis

$$\frac{(RSS(m) - RSS(M))/(p - q)}{RSS(M)/(n - p - 1)} \sim F_{p-q,n-p-1}.$$

p-value can therefore be got by

$$\mathbb{P}\left(F_{p-q,n-p-1} > \frac{(RSS(m) - RSS(M))/(p - q)}{RSS(M)/(n - p - 1)}\right).$$

If the null hypothesis can be written in terms of a single linear function of β, such as $H_0 : \beta_1 + 5\beta_3 = 5$. Then it can also be tested via the t-test; using the statistic:

$$\frac{\hat{\beta}_1 + 5\hat{\beta}_3 - 5}{s.e(\hat{\beta}_1 + 5\hat{\beta}_3)}$$

which has the t-distribution with $n - p - 1$ degrees of freedom under H_0. This test and the corresponding F-test will have the same p-value.

1.1 Testing for all explanatory variables

How do we test $H_0 : \beta_1 = \cdots = \beta_p = 0$ against its complement? Just take m to be the model $y_i = \beta_0 + e_i$. In this case, $RSS(m) = TSS$ and $q = 0$ and $RSS(M) = RSS$. Thus the p-value is

$$\mathbb{P}\left\{F_{p,n-p-1} > \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}\right\}.$$

Spring 2013 Statistics 153 (Time Series) : Lecture Ten

Aditya Guntuboyina

21 February 2013

1 Best Linear Prediction

Suppose that Y and $W_1, \ldots, W_m$ are random variables with zero means and finite variances. Let $\text{cov}(Y, W_i) = \zeta_i, i = 1, \ldots, m$ and

$$\text{cov}(W_i, W_j) = \Delta(i, j) \qquad \text{for } i, j = 1, \ldots, m.$$

What is the best **linear predictor** of Y in terms of $W_1, \ldots, W_m$?

The best linear predictor $a_1, \ldots, a_m$ is characterized by the property that $Y - a_1 W_1 - \cdots - a_m W_m$ is uncorrelated with $W_1, \ldots, W_m$. In other words:

$$\text{cov}(Y - a_1 W_1 - \cdots - a_m W_m, W_i) = 0 \qquad \text{for } i = 1, \ldots, m.$$

Note that this gives m equations in the m unknowns $a_1, \ldots, a_m$. The ith equation can be rewritten as

$$\zeta_i - \Delta(i, 1) a_1 - \cdots - \Delta(i, m) a_m = 0.$$

In other words, this means that ζ_i equals the ith row of Δ multiplied by the vector $a = (a_1, \ldots, a_m)^T$ which is same as the ith element of the vector Δa. Thus these m equations can be written in one line as $\Delta a = \zeta$.

Another way to get this defining equation for the coefficients of the best linear predictor is to find values of $a_1, \ldots, a_m$ that minimize

$$\begin{aligned} F(\mathbf{a}) &:= \mathbb{E}\left(Y - a_1 W_1 - \cdots - a_m W_m\right)^2 \\ &= \mathbb{E}\left(Y - a^T W\right)^2 \\ &= \mathbb{E} Y^2 - 2\mathbb{E}((a^T W) Y) + \mathbb{E}(a^T W W^T a) \\ &= \mathbb{E} Y^2 - 2 a^T \zeta + a^T \Delta a. \end{aligned}$$

Differentiate with respect to a and set equal to zero to get

$$-2\zeta + 2\Delta a = 0$$

or $a = \Delta^{-1}\zeta$. Therefore the best linear predictor of Y in terms of $W_1, \ldots, W_m$ equals $\zeta^T \Delta^{-1} W$.

The special case of this for $m = 1$ (when there is only one predictor W_1) may be more familiar. When $m = 1$, we have $\zeta_1 = \text{cov}(Y, W_1)$ and $\Delta(1, 1) = \text{var}(W_1)$. Thus, the best predictor or Y in terms of W_1 is

$$\frac{\text{cov}(Y, W_1)}{\text{var}(W_1)} W_1.$$

Now consider a stationary mean zero time series $\{X_t\}$. Using the above with $Y = X_n$ and $W_1 = X_{n-1}$, we get that the best predictor of X_n in terms of X_{n-1} is

$$\frac{\text{cov}(X_n, X_{n-1})}{\text{var}(X_{n-1})} X_{n-1} = \frac{\gamma_X(1)}{\gamma_X(0)} X_{n-1} = \rho_X(1) X_{n-1}$$

What is the best predictor for X_n in terms of $X_{n-1}, X_{n-2}, \ldots, X_{n-k}$? Here we take $Y = X_n$ and $W_i = X_{n-i}$ for $i = 1, \ldots, k$. Therefore

$$\Delta(i,j) = \text{cov}(W_i, W_j) = \text{cov}(X_{n-i}, X_{n-j}) = \gamma_X(i-j)$$

and

$$\zeta_i = \text{cov}(Y, W_i) = \text{cov}(X_n, X_{n-i}) = \gamma_X(i).$$

With these Δ and ζ, solve for $\Delta a = \zeta$ to obtain the coefficients of $X_{n-1}, \ldots, X_{n-k}$ in the best linear predictor of X_n.

Consider the special case of the AR(p) model: $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$. Directly from the defining equation and causality, it follows that $X_n - \phi_1 X_{n-1} - \cdots - \phi_p X_{n-p}$ is uncorrelated with $X_{n-1}, X_{n-2}, \ldots$. We thus deduce that the best linear predictor of X_n in terms of $X_{n-1}, X_{n-2}, \ldots$ equals $\phi_1 X_{n-1} + \phi_2 X_{n-2} + \cdots + \phi_p X_{n-p}$.

2 The Partial Autocorrelation Function (pacf)

2.1 First Definition

Let $\{X_t\}$ be a mean zero stationary process. The Partial Autocorrelation at lag h, denoted by $pacf(h)$ is defined as the coefficient of X_{t-h} in the best linear predictor for X_t in terms of $X_{t-1}, \ldots, X_{t-h}$.

Check that $pacf(1)$ is the same as the autocorrelation at lag one, $\rho(1)$. But $pacf(h)$ for $h > 1$ can be quite different from $\rho(h)$.

For the AR(p) model: $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$, check that $pacf(p) = \phi_p$ and that $pacf(h) = 0$ for $h > p$.

2.2 Second Definition

From the first definition, it is not quite clear why this is called a correlation. This will be apparent from the second definition.

The pacf at lag h is defined as the correlation between X_t and X_{t-h} **with the effect of the intervening variables $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$ removed**. Let $\beta_1 X_{t-1} + \cdots + \beta_{h-1} X_{t-h+1}$ denote the best linear predictor of X_t in terms of $X_{t-1}, \ldots, X_{t-h+1}$. By stationarity, the two sequences

$$X_t, X_{t-1}, \ldots, X_{t-h+1}$$

and

$$X_{t-h}, X_{t-h+1}, \ldots, X_{t-1}$$

have the same covariance matrix. Indeed, if $W_i = X_{t-i+1}$ and $\tilde{W}_i = X_{t-h+i-1}$ for $i = 1, \ldots, h$, then the covariance between W_i and W_j equals $\gamma_X(i-j)$ which is the same as the covariance between $\tilde{W}_i$ and $\tilde{W}_j$.

Therefore, the best linear prediction of X_{t-h} in terms of $X_{t-h+1}, \ldots, X_{t-1}$ equals $\beta_1 X_{t-h+1} + \cdots + \beta_{h-1} X_{t-1}$.

The pacf at lag h is defined as

$$pacf(h) = \text{corr}\left(X_t - \beta_1 X_{t-1} - \cdots - \beta_{h-1} X_{t-h+1}, X_{t-h} - \beta_1 X_{t-h+1} - \cdots - \beta_{h-1} X_{t-1}\right).$$

In other words, $pacf(h)$ is the correlation between the **errors in the best linear predictions** of X_t and X_{t-h} in terms of the intervening variables $X_{t-1}, \ldots, X_{t-h+1}$.

The key fact is that for an AR(p) model, $pacf(h)$ equals zero for lags $h > p$. To see this: note that for $h > p$, the best linear predictor for X_t in terms of $X_{t-1}, \ldots, X_{t-h+1}$ equals $\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}$. In other words, $\beta_1 = \phi_1, \ldots, \beta_p = \phi_p$ and $\beta_i = 0$ for $i > p$.

Therefore for $h > p$, we have

$$\begin{aligned} pacf(h) &= \text{corr}\left(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) \\ &= \text{corr}\left(Z_t, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) = 0, \end{aligned}$$

by causality.

The equivalence between the two definitions of $pacf(h)$ can be proved by linear algebra. We will skip this derivation.

3 Estimating *pacf* from Data

How does one estimate $pacf(h)$ from data for different lags h? The coefficients $a_1, \ldots, a_h$ of $X_{t-1}, \ldots, X_{t-h}$ in the best linear predictor of X_t are obtained by solving an equation of the form $\Delta a = \zeta$.

Now all the elements of Δ and ζ are of the form $\gamma_X(i-j)$ for some i and j. Therefore, a natural method of estimating $pacf(h)$ is to estimate the entries in Δ and ζ by the respective sample autocorrelations to obtain $\hat{\Delta}$ and $\hat{\zeta}$ and then to solve the equation $\hat{\Delta}\hat{a} = \hat{\zeta}$ for $\hat{a}$. Note that $pacf(h)$ is precisely a_h.

It has been shown that when the data come from an AR(p) model, the sample partial autocorrelations at lags greater than p are approximately **independently normally** distributed with zero means and variances $1/n$. Thus for $h > p$, bands at $\pm 1.96 n^{-1/2}$ can be used for checking if an AR(p) model is appropriate.

4 Summary

For an $MA(q)$ model, the autocorrelation function $\rho_X(h)$ equals zero for $h > q$. Also for $h > q$, the sample autocorrelation functions r_h are approximately normal with mean 0 and variance w_{hh}/n where $w_{hh} := 1 + 2\rho^2(1) + \cdots + 2\rho^2(q)$.

For an $AR(p)$ model, the partial autocorrelation function $pacf(h)$ equals zero for $h > p$. Also for $h > p$, the sample autocorrelation functions r_h are approximately normal with mean 0 and variance $1/n$.

If the sample acf for a data set cuts off at some lag, we use an MA model. If the sample pacf cuts off at some lag, we use an AR model.

What if neither of the above happens? How do we then choose an appropriate ARMA model? Here is a general strategy:

1. Try ARMA(p, q) for various choices of p and q.
2. For a fixed p and q, fit the ARMA(p, q) model to the data (we will soon learn how to do this).

3. See how good the fit is. Select p and q so that the fit is good **while making sure there is no overfitting**.

How to check if a model is fits the data well but does not overfit? This is a problem of model selection. Often automatic criteria like AIC, FPE, BIC are used. One should also use judgement.

Our plan is as follows:

1. How to fit an ARMA model to data?
2. How to assess goodness of fit?
3. Choosing p and q by an automatic Model selection technique.

Fall 2013 Statistics 151 (Linear Models) : Lecture Eleven

Aditya Guntuboyina

03 October 2013

1 One Way Analysis of Variance

Consider the model

$$y_{ij} = \mu_i + e_{ij} \qquad \text{for } i = 1, \ldots, t \text{ and } j = 1, \ldots, n_i$$

where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.

This model is used for the following kinds of situations:

1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments. $y_{i1}, \ldots, y_{in_i}$ denote the scores of the subjects that received the ith treatment.
2. We are looking at some performance of n subjects who can naturally be divided into t groups. We would like to see if the performance difference between the subjects can be explained by the fact that there in these different groups. $y_{i1}, \ldots, y_{in_i}$ denote the performance of the subjects in the ith group.

Often this model is also written as

$$y_{ij} = \mu + \tau_i + e_{ij} \qquad \text{for } i = 1, \ldots, t \text{ and } j = 1, \ldots, n_i \tag{1}$$

where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score. In this model, μ and the individual τ_is are not estimable. It is easy to show that here a parameter $\lambda\mu + \sum_{i=1}^{t} \lambda_i \tau_i$ is estimable if and only if $\lambda = \sum_{i=1}^{t} \lambda_i$. Because of this lack of estimability, people often impose the condition $\sum_{i=1}^{t} \tau_i = 0$. This condition ensures that all parameters μ and $\tau_1, \ldots, \tau_t$ are estimable. Moreover, it provides a nice interpretation. μ denotes the baseline response value and τ_i is the value by which the response value needs to be adjusted from the baseline μ for the group i. Because $\sum_i \tau_i = 0$, some adjustments will be positive and some negative but the overall adjustment averaged across all groups is zero.

How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model? This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m). What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$. Write

$$\begin{aligned}
\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \mu_i)^2 \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2 \sum_{i=1}^{t} (\bar{y}_i - \mu_i) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) + \sum_{i=1}^{t} n_i (\bar{y}_i - \mu_i)^2 \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{t} n_i (\bar{y}_i - \mu_i)^2 .
\end{aligned}$$

Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$. If we write μ_i as $\mu + \tau_i$ with $\sum_i \tau_i = 0$, then the least squares estimate of μ is $\bar{y}$ and the least squares estimate of τ_i is $\bar{y}_i - \bar{y}$.

The RSS in the full model is

$$RSS(M) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 .$$

Check that the RSS in the reduced model is

$$RSS(m) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{t} n_i (\bar{y}_i - \bar{y})^2 .$$

Thus the F-statistic for testing $H_0 : \mu_1 = \cdots = \mu_t$ is

$$T = \frac{\sum_{i=1}^{t} n_i (\bar{y}_i - \bar{y})^2/(t-1)}{\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2/(n-t)}$$

which has the F-distribution with $t-1$ and $n-t$ degrees of freedom under H_0.

2 Permutation Tests

We have studied hypothesis testing in the linear model via the F-test so far. Suppose we want to test a linear hypothesis about $\beta = (\beta_0, \ldots, \beta_p)^T$ in the full linear model (denoted by M). We first construct a reduced model which incorporates the hypothesis in the full model M. Call this reduced model m. We then look at the quantity:

$$T := \frac{(RSS(m) - RSS(M))/(p-q)}{RSS(M)/(n-p-1)}.$$

It makes sense to reject the null hypothesis if T is large. To answer the question: *how large is large?*, we rely on the assumption of normality of the errors i.e., $e \sim N(0, \sigma^2 I)$ to assert that $T \sim F_{p-q,n-p-1}$ under H_0. As a result, a p-value can be obtained as $\mathbb{P}\{F_{p-q,n-p-1} > T\}$.

Suppose we do not want to assume normality of errors. Is there any way to obtain a p-value? This is possible in some cases via permutation tests. We provide two examples below.

2.1 Testing for all explanatory variables

We want to test the null hypothesis that all explanatory variables can be thrown away without assuming that $e \sim N(0, \sigma^2 I)$. Under the null hypothesis, we assume that if the response variable y has no relation to the explanatory variables. Therefore, it is plausible to assume that under the null hypothesis, the values of the response variable $y_1, \ldots, y_n$ are randomly distributed between the n subjects without relation to the predictors. This motivates the following test:

1. Randomly permute the response values: $y_1, \ldots, y_n$.
2. Calculate the quantity
$$\frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}.$$
with the response values being the permuted values in the pervious step.
3. Repeat the above pair of steps a large number of times.

4. This results in a large number of values of the test statistic (one for each permutation of the response values). Let us call them $T_1, \ldots, T_N$. The p-value is calculated as the proportion of $T_1, \ldots, T_N$ that exceed the original test statistic value T (T is calculated with the actual unpermuted response values $y_1, \ldots, y_n$).

The idea behind this test is as follows: From the given data, we calcuate the value of

$$\frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}.$$

We need to know how extreme this value is under the null hypothesis. Under the assumption of normality, we can assess this by the F-distribution. But we need to do this without assuming normality. For this, we try to generate values of this quantity under the null hypothesis. The idea is to do this by calculating the statistic after permuting the response values. Because once the response values are permuted, all association between the response and explanatory variables breaks down so that the values of

$$\frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}.$$

for the permuted response values resembles values generated under the null hypothesis. The p-value is then calculated as the proportion of these values larger than the observed value.

2.2 Testing for a single explanatory variable

How do we test if, say, the first explanatory variable is useful? We calcuate the t-statistic:

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$$

and calculate p-value by comparing it with the t_{n-p-1} distribution (which requires normality). How to do this without normality?

We can follow the permutation test by permuting the values of x_1. For each permutation, we calculate the t-statistic and the p-value is the proportion of these t-values which are larger than the observed t-value in absolute value.

Spring 2013 Statistics 153 (Time Series) : Lecture Twelve

Aditya Guntuboyina

05 March 2013

1 Plan

So Far:

1. Trend and Seasonality
2. Stationarity
3. ARMA models

Still to come in Time Domain Techniques:

1. How to fit ARMA models to data.
2. ARIMA models
3. SARIMA models
4. Forecasting
5. Model Diagnostics and Selection

2 Recap: Fitting AR models to data

Assuming that the order p is known. Carried out by invoking the function $ar()$ in R.

1. **Yule Walker or Method of Moments**: Finds the AR(p) model whose acvf equals the sample autocorrelation function at lags $0, 1, \ldots, p$. Use yw for method in R.
2. **Conditional Least Squares**: Minimizes the conditional sum of squares: $\sum_{i=p+1}^{n}(x_i - \mu - \phi_1(x_{i-1} - \mu) - \cdots - \phi_p(x_{i-p} - \mu))^2$ over μ and $\phi_1, \ldots, \phi_p$. And σ^2 is achieved by the average of the squared residuals. Use *ols* for method in R. In this method, given data $x_1, \ldots, x_n$, R fits a model of the form $x_t - \bar{x} = intercept + \phi(x_{t-1} - \bar{x}) + residual$ to the data. The fitted value of intercept can be obtained by calling $\$x.intercept$. One can convert this to a model of the form $x_t = intercept + \phi x_{t-1} + residual$. Check the help page for the R function ar.ols.
3. **Maximum Likelihood**: Maximizes the likelihood function (which is relatively straightforward to write down but which requires an optimization routine to maximize). Use *mle* for method in R. This method is complicated.

It is usually the case that all these three methods yield similar answers. The default method in R is Yule-Walker.

3 Asymptotic Distribution of the Yule-Walker Estimates for AR models

For n large, the approximate distribution of $\sqrt{n}\left(\hat{\phi}-\phi\right)$ is normal with mean 0 and variance covariance matrix $\sigma_Z^2\Gamma_p^{-1}$ where Γ_p is the $p\times p$ matrix whose (i,j)th entry is $\gamma_X(i-j)$.

For example, in the AR(1) case:

$$\Gamma_p=\Gamma_1=\gamma_X(0)=\sigma_Z^2/(1-\phi^2).$$

Thus $\hat{\phi}$ is approximately normal with mean ϕ and variance $(1-\phi^2)/n$.

For AR(2), using

$$\gamma_X(0)=\frac{1-\phi_2}{1+\phi_2}\frac{\sigma_Z^2}{(1-\phi_2)^2-\phi_1^2}\qquad\text{and}\qquad\rho_X(1)=\frac{\phi_1}{1-\phi_2},$$

we can show that $(\hat{\phi}_1,\hat{\phi}_2)$ is approximately normal with mean (ϕ_1,ϕ_2) and variance-covariance matrix is $1/n$ times

$$\begin{pmatrix}1-\phi_2^2 & -\phi_1(1+\phi_2)\\ -\phi_1(1+\phi_2) & 1-\phi_2^2\end{pmatrix}$$

Spring 2013 Statistics 153 (Time Series) : Lecture Thirteen

Aditya Guntuboyina

07 March 2013

1 Asymptotic Distribution of the Estimates for AR models

The following holds for each of the Yule-Walker, Conditional Least Squares and ML estimates:

For n large, the approximate distribution of $\sqrt{n}\left(\hat{\phi}-\phi\right)$ is normal with mean 0 and variance covariance matrix $\sigma_Z^2\Gamma_p^{-1}$ where Γ_p is the $p\times p$ matrix whose (i,j)th entry is $\gamma_X(i-j)$.

1.1 Proof Sketch

Assume $\mu=0$ for simplicity. It is easiest to work with the conditional least squares estimates. The AR(p) model is:

$$X_t=\phi_1X_{t-1}+\cdots+\phi_pX_{t-p}+Z_t.$$

We may write this model in matrix notation as:

$$X_t=\mathbb{X}_{t-1}^T\phi+Z_t$$

where $\mathbb{X}_{t-1}$ is the $p\times 1$ vector $\mathbb{X}_{t-1}=(X_{t-1},X_{t-2},\ldots,X_{t-p})^T$ and ϕ is the $p\times 1$ vector $(\phi_1,\ldots,\phi_p)^T$. The conditional least squares method minimizes the sum of squares:

$$\sum_{t=p+1}^{n}\left(X_t-\phi^T\mathbb{X}_{t-1}\right)^2$$

with respect to ϕ. The solution is:

$$\hat{\phi}=\left(\sum_{t=p+1}^{n}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^T\right)^{-1}\left(\sum_{t=p+1}^{n}\mathbb{X}_{t-1}X_t\right).$$

Writing $X_t=\mathbb{X}_{t-1}^T\phi+Z_t$, we get

$$\hat{\phi}=\phi+\left(\sum_{t=p+1}^{n}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^T\right)^{-1}\left(\sum_{t=p+1}^{n}\mathbb{X}_{t-1}Z_t\right).$$

As a result,

$$\sqrt{n}\left(\hat{\phi}-\phi\right)=\left(\frac{1}{n}\sum_{t=p+1}^{n}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p+1}^{n}\mathbb{X}_{t-1}Z_t\right). \tag{1}$$

The following assertions are intuitive (note that $\mathbb{X}_{t-1}$ and Z_t are uncorrelated and hence independent under the gaussian assumption) and can be proved rigorously:

$$\frac{1}{n}\sum_{t=p+1}^{n}\mathbb{X}_{t-1}\mathbb{X}_{t-1}^T \to \Gamma_p \qquad \text{as } n\to\infty \text{ in probability}$$

and

$$\frac{1}{\sqrt{n}}\sum_{t=p+1}^{n}\mathbb{X}_{t-1}Z_t \to N\left(0,\sigma_Z^2\Gamma_p\right) \qquad \text{as } n\to\infty \text{ in distribution.}$$

These results can be combined with the expression (1) to prove that $\sqrt{n}\left(\hat{\phi}-\phi\right)$ converges in distribution to a normal distribution with mean 0 and variance covariance matrix $\sigma_Z^2\Gamma_p^{-1}$.

1.2 Special Instances

In the AR(1) case:

$$\Gamma_p = \Gamma_1 = \gamma_X(0) = \sigma_Z^2/(1-\phi^2).$$

Thus $\hat{\phi}$ is approximately normal with mean ϕ and variance $(1-\phi^2)/n$.

For AR(2), using

$$\gamma_X(0) = \frac{1-\phi_2}{1+\phi_2}\frac{\sigma_Z^2}{(1-\phi_2)^2-\phi_1^2} \qquad \text{and} \qquad \rho_X(1) = \frac{\phi_1}{1-\phi_2},$$

we can show that $(\hat{\phi}_1,\hat{\phi}_2)$ is approximately normal with mean (ϕ_1,ϕ_2) and variance-covariance matrix is $1/n$ times

$$\begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}$$

Note that the approximate variances of both $\hat{\phi}_1$ and $\hat{\phi}_2$ are the same. Observe that if we fit AR(2) model to a dataset that comes from AR(1), then the estimate of $\hat{\phi}_1$ might not change much but the standard error will be higher. We lose precision. See Example 3.34 in the book.

2 More General ARMA model fitting

2.1 Invertibility

Consider the case of the MA(1) model whose acvf is given by $\gamma_X(0) = \sigma_Z^2(1+\theta^2)$ and $\gamma_X(1) = \theta\sigma_Z^2$ and $\gamma_X(h) = 0$ for all $h \geq 2$. It is easy to see that for $\theta = 5, \sigma_Z^2 = 1$, we get the same acvf as for $\theta = 1/5, \sigma_Z^2 = 25$. In other words, there exist different parameter values that give the same acvf. More generally, the parameter pairs (θ,σ_Z^2) and $(1/\theta,\theta^2\sigma_Z^2)$ correspond to the same acvf.

This implies that one **can not uniquely** estimate the parameters of an MA(1) model from data. A natural fix is to consider only those MA(1) for which $|\theta| < 1$. This condition is called *invertibility*. The condition $|\theta| < 1$ for the MA(1) model is equivalent to stating that the moving average polynomial $\theta(z)$ has all roots of magnitude strictly larger than one. This gives the general definition of invertibility for an ARMA process.

An ARMA model $\phi(B)(X_t-\mu) = \theta(B)Z_t$ is said to be invertible if all roots of the moving average polynomial $\theta(z)$ have magnitude strictly larger than one. It can be shown (in analogy with causality)

that this condition is equivalent to Z_t being written as a linear combination of the present and past values of X_t alone.

From now on, we shall only consider stationary, causal and invertible ARMA models i.e., we shall assume that both the polynomials $\phi(z)$ and $\theta(z)$ do not have any roots in the unit disk.

Let us now study the problem of fitting a stationary, causal and invertible ARMA model to data assuming that the orders p and q are known.

Each of the three methods for the AR model fitting carry over (with additional complications) to the general ARMA case. It is easiest to start by learning the relevant R function. The function to use is *arima*(). We will see later that ARIMA is a more general class of models that include the ARMA models as a special case (actually, ARIMA is just differencing + ARMA). This function *arima*() can be used to fit ARMA models to data. It also has a method argument that has three values: CSS-ML, ML and CSS, the default being CSS-ML.

2.2 Yule-Walker or Method of moments

This proceeds, in principle, by solving some subset of the following set of equations for the unknown parameters $\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p$ and σ_Z^2 (and μ is estimated by the sample mean)

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = (\psi_0\theta_k + \psi_1\theta_{k+1} + \cdots + \psi_{q-k}\theta_q)\sigma_Z^2$$

for $0 \leq k \leq q$ and

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = 0 \qquad \text{for } k > q.$$

Note that ψ_j above are functions of $\theta_1, \ldots, \theta_q$ and $\phi_1, \ldots, \phi_p$.

This method of estimation has the following problems:

1. It is cumbersome (unless we are in the pure AR case): Solutions might not always exist to these equations (for example, in the MA(1), this method entails solving $r_1 = \theta/(1+\theta^2)$ which of course does not have a solution when $r_1 \notin [-0.5, 0.5]$). The parameters are estimated in an arbitrary fashion when these equations do not have a solution.
2. The estimators obtained are *inefficient*. The other techniques below give much better estimates (smaller standard errors).

Because of these problems, no one uses method of moments for estimating the parameters of a general ARMA model. R does not even have a function for doing this. Note, however, that both of these problems disappear for the case of the pure AR model.

2.3 Conditional Least Squares

Let us first consider the special case of the MA(1) model: $X_t - \mu = Z_t + \theta Z_{t-1}$. We want to fit this model to data $x_1, \ldots, x_n$. If the data were indeed generated from this model, then

$$Z_1 = x_1 - \mu - \theta Z_0; Z_2 = x_2 - \mu - \theta Z_1; \ldots; Z_n = x_n - \mu - \theta Z_{n-1}.$$

If we set Z_0 to its mean 0, then for **every fixed values** of θ and μ, we can recursively calculate $Z_1, \ldots, Z_n$. We can then compute the sum of squares $\sum_{i=1}^n Z_i^2$. This value would change for different values of θ. We would then choose the value of θ for which it is small (this is accomplished by an optimization procedure).

This is also called conditional least squares because this minimzation is obtained when one tries to maximize the conditional likelihood of the data conditioning on $z_0 = 0$.

Note that conditional likelihood works differently in the AR(1) case compared to the MA(1) case. It works in a yet another different way in the ARMA(1, 1) case for example. Here the model is $X_t - \mu - \phi(X_{t-1} - \mu) = Z_t + \theta Z_{t-1}$. Here it is convenient to set Z_1 to be zero. Then we can write

$$Z_2 = x_2 - \mu - \phi(x_1 - \mu); Z_3 = x_3 - \mu - \phi(x_2 - \mu) - \theta Z_2; \ldots; Z_n = x_n - \mu - \phi(x_{n-1} - \mu) - \theta Z_{n-1}.$$

After this, one forms the sum of squares $\sum_{i=2}^{n} Z_i^2$ which can be computed for every fixed values of θ, ϕ and μ. One then minimizes these resulting sum of squares over different values of the unknown parameters.

For a general ARMA(p, q) model:

$$X_t - \mu - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

we set $Z_t = 0$ for $t \leq p$ and calculate recursively

$$Z_t = X_t - \mu - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q}$$

for $t = p + 1, \ldots, n$. This is equivalent to writing the likelihood coniditioning on $X_1, \ldots, X_p$ and $Z_t = 0$ for $t \leq p$. If $q = 0$ (AR models), minimizing the sum of squares is equivalent to linear regression and no iterative technique is needed. If $q > 0$, the problem becomes nonlinear regression and numerical optimization routines need to be used.

In R, this method is performed by calling the function *arima*() with the method argument set to *CSS* (CSS stands for conditional sum of squares).

2.4 Maximum Likelihood

This method is simple in principle. Assume that the errors $\{Z_t\}$ are gaussian. Write down the likelihood of the observed data $x_1, \ldots, x_n$ in terms of the unknown parameter values $\mu, \theta_1, \ldots, \theta_q$, $\phi_1, \ldots, \phi_p$ and σ_Z^2. Maximize over these unknown parameter values.

It is achieved in R by calling the function *arima*() with the method argument set to *ML* or *CSS-ML*. ML stands of course for Maximum Likelihood. R uses an optimization routine to maximize the likelihood. This routine is iterative and needs suitable initial values of the parameters to start. In CSS-ML, R selects these starting values by CSS. I do not quite know how the starting values are selected in ML. The default method for the arima function in R is CSS-ML. The R output for the methods CSS-ML and ML seems to be identical.

Spring 2013 Statistics 153 (Time Series) : Lecture Fourteen

Aditya Guntuboyina

12 March 2013

1 Asymptotic Distribution of ARMA ML Estimates

Let $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$. The distribution of $\hat{\beta}$ is approximately normal with mean β and variance covariance matrix $\sigma_Z^2 \Gamma_{p,q}^{-1}/n$ where $\Gamma_{p,q}$ is a $(p+q) \times (p+q)$ matrix of the form:

$$\begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}$$

The (i,j)th entry of the $p \times p$ matrix $\Gamma_{\phi\phi}$ equals $\gamma_A(i-j)$ where $\{A_t\}$ is the AR(p) process: $\phi(B)A_t = Z_t$. Similarly, the $q \times q$ matrix $\Gamma_{\theta\theta}$ has (i,j)th entry equalling $\gamma_B(i-j)$ for the AR(q) process $\theta(B)B_t = Z_t$. The (i,j)th entry of $\Gamma_{\phi\theta}$ equals the covariance between A_i and B_j.

1.1 Special Cases

The result in particular states that the variance-covariance matrix for AR and MA models will be very similar (the only difference is in signs). In the MA(1) case:

$$\Gamma_\theta = \Gamma_1 = \sigma_Z^2/(1-\theta^2).$$

Thus $\hat{\theta}$ is approximately normal with mean θ and variance $(1-\theta^2)/n$.

For MA(2), $(\hat{\theta}_1, \hat{\theta}_2)$ is approximately normal with mean (θ_1, θ_2) and variance-covariance matrix is $1/n$ times

$$\begin{pmatrix} 1-\theta_2^2 & \theta_1(1-\theta_2) \\ \theta_1(1-\theta_2) & 1-\theta_2^2 \end{pmatrix}$$

.

For ARMA(1, 1), to calculate $\Gamma_{\phi\theta}$, we must find the covariance between A_1 and B_1 where $A_1 - \phi A_0 = Z_t$ and $B_1 + \theta B_0 = Z_t$. Write

$$\Gamma_{\phi\theta} = \mathrm{cov}(A_1, B_1) = \mathrm{cov}(\phi A_0 + Z_1, -\theta B_0 + Z_t) = -\phi\theta\Gamma_{\phi\theta} + \sigma_Z^2$$

which gives $\Gamma_{\phi\theta} = \sigma_Z^2/(1+\phi\theta)$. This gives that $(\hat{\phi}, \hat{\theta})$ is approximately normal with mean (ϕ, θ) and variance-covariance matrix is $1/n$ times the **inverse** of

$$\begin{pmatrix} (1-\phi^2)^{-1} & (1+\phi\theta)^{-1} \\ (1+\phi\theta)^{-1} & (1-\theta^2)^{-1} \end{pmatrix}$$

.

2 ARIMA Models

ARIMA is essentially differencing plus ARMA. We have seen previously that differencing is commonly used on time series data to remove trends and seasonality.

For example, differencing can be used for

1. Removing polynomial trends: Suppose the data come from the model $Y_t = \mu_t + X_t$ where μ_t is a polynomial of order k and X_t is stationary, then differencing of order k: $\nabla^k Y_t = (I - B)^k Y_t$ results in stationary data to which an ARMA model can be fit.
2. Random walk models: Suppose that the data come from the random walk model: $Y_t = Y_{t-1} + X_t$ where X_t is stationary. Then clearly $\nabla Y_t = X_t$ is stationary and an ARMA model can be fit to this difference data.

Such models, which after appropriate differencing, reduce to ARMA models are called ARIMA models.

Definition 2.1 (ARIMA)**.** *A process Y_t is said to be ARIMA(p, d, q) with mean μ if $X_t = (I - B)^d Y_t$ is ARMA(p, q) with mean μ. In other words:*

$$\phi(B)(X_t - \mu) = \theta(B)Z_t,$$

where $\{Z_t\}$ is white noise.

3 Fitting ARIMA models

Just use the function *arima(dataset, order = c(p, d, q))*. I suggest you always use this function. If you know that you want to fit a pure AR model, you might consider the *ar()* function.

The arima function will give you the estimates of μ (under the name intercept), $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$. It will also give you the estimated standard errors. An estimate of σ^2 is also provided.

Spring 2013 Statistics 153 (Time Series) : Lecture Fifteen

Aditya Guntuboyina

14 March 2013

1 ARIMA Forecasting

Forecasting for a future observation, x_{n+m}, is done using the best linear predictor of X_{n+m} in terms of $X_1, \ldots, X_n$. The coefficients of the best linear predictor involve the parameters of the ARIMA model used for $x_1, \ldots, x_n$. These paramters are estimated from data.

We have already seen how the best linear predictor of a random variable Y in terms of $W_1, \ldots, W_m$ is calculated.

Suppose that all the random variables $Y, W_1, \ldots, W_m$ have mean zero. Then the best linear predictor is $a_1 W_1 + \cdots + a_m W_m$ where $a_0, \ldots, a_m$ are characterized by the set of equations:

$$\operatorname{cov}(Y - a_1 W_1 - \cdots - a_m W_m, W_i) = 0 \qquad \text{for } i = 1, \ldots, m.$$

The above gives m equations in the m unknowns $a_1, \ldots, a_m$. The equations can be written in a compact form as $\Delta a = \zeta$ where $\Delta(i,j) = \operatorname{cov}(W_i, W_j)$ and $\zeta_i = \operatorname{cov}(Y, W_i)$.

If the random variables $Y, W_1, \ldots, W_m$ have different means: $\mathbb{E}Y = \mu_Y$ and $\mathbb{E}W_i = \mu_i$, then the best linear predictor of $Y - \mu_Y$ in terms of $W_1 - \mu_1, \ldots, W_m - \mu_m$ is given by $a_1(W_1 - \mu_1) + \cdots + a_m(W_m - \mu_m)$ where $a_1, \ldots, a_m$ are given by the same equation $\Delta a = \zeta$. Thus, in these non-zero mean case, the best linear predictor of Y in terms of $W_1, \ldots, W_m$ is

$$\mu_Y + a_1(W_1 - \mu_1) + \cdots + a_m(W_m - \mu_m).$$

The prediction error is measured by

$$\mathbb{E}\left(Y - \mu_Y - a_1(W_1 - \mu_1) - \cdots - a_m(W_m - \mu_m)\right)^2.$$

For ARMA models, there exist iterative algorithms for quickly calculating the best linear predictors of X_{n+m} based on $X_1, \ldots, X_n$ and the corresponding prediction errors recursively over n and m e.g., Durbin-Levinson and Innovations. These do not explicit inversion of the matrix Δ.

2 Time Series Data Analysis

1. Exploratory analysis.
2. Decide if it makes sense to transform the data (either for better interpretation or for stabilizing the variance).
3. Deal with trend or seasonality. Either by fitting deterministic models or by smoothing or differencing.

4. Fit an ARMA model to the residuals obtained after trend and seasonality are removed.

5. Check if the fitted ARMA model is adequate (Model Diagnostics).

6. Forecast.

3 Model Diagnostics

After fitting an ARIMA model to data, one can form the residuals: $x_i - \hat{x}_i^{i-1}$ by looking at the difference between the ith observation and the best linear prediction of the ith observation based on the previous observations $x_1, \ldots, x_{i-1}$. One usually standardizes this residual by dividing by the square-root of the corresponding prediction error.

If the model fits well, the standardized residuals should behave as an iid sequence with mean zero and variance one. One can check this by looking at the plot of the residuals and their correlogram. Departures from gaussianity also need to be assessed (this is done by looking at the Q-Q plot).

Let $r_e(h)$ denote the sample acf of the residuals from an ARMA fit. For the fit to be good, the residuals have to be iid with mean zero and variance one which implies that $r_e(h)$ for $h = 1, 2, \ldots$ have to be i.i.d with mean 0 and variance $1/n$.

In addition to plotting $r_e(h)$, there is a formal test that takes into account the magnitudes of $r_e(h)$ together. This is the Ljung-Box-Pierce test that is based on the so-called Q-statistic:

$$Q := n(n+2) \sum_{h=1}^{H} \frac{r_e^2(h)}{n-h}.$$

Under the null hypothesis of model adequacy, the distribution of Q is asymptotically χ^2 with degrees of freedom $H - p - q$. The maximum lag H is chosen arbitrarily (typically 20). Thus, one would reject the null at level α if the observed value of Q exceeds the $(1-\alpha)$ quantile of the χ^2 distribution with $H-p-q$ degrees of freedom.

4 Seasonal ARIMA Models

These provide models that have non-zero autocorrelations for small lags (say, 0 and 1) and also at some seasonal lag (say, 12) and zero autocorrelation for all other lags.

Consider the MA model: $X_t = Z_t + \Theta Z_{t-12} = (1 + \Theta B^{12}) Z_t$. This can be thought of as an MA(12) model with $\theta_1 = \cdots = \theta_{11} = 0$ and $\theta_{12} = \Theta$. This is a stationary model whose autocovariance function is non-zero only at lags 0 and 12. It is therefore called a seasonal MA(1) model with seasonal period 12.

Generalizing, a seasonal MA(Q) model with seasonal period s is defined by

$$X_t = Z_t + \Theta_1 Z_{t-s} + \Theta_2 Z_{t-2s} + \cdots + \Theta_Q Z_{t-Qs}.$$

This stationary model has autocorrelation that is non-zero only at lags $0, s, 2s, \ldots, Qs$. Note that this is just a MA(Qs) model with the MA polynomial $1 + \Theta_1 z^s + \Theta_2 z^{2s} + \cdots + \Theta_Q z^{Qs}$.

For the co2 dataset, we need a stationary model with non-zero autocorrelations at lags 1, 11, 12 and 13 (and zero autocorrelation at all other lags). An example of such a model is given by:

$$X_t = Z_t + \theta Z_{t-1} + \Theta Z_{t-12} + \theta\Theta Z_{t-13}.$$

More compactly, this model can be written as

$$X_t = (1 + \theta B + \Theta B^{12} + \theta\Theta B^{13})Z_t = (1 + \theta B)(1 + \Theta B^{12})Z_t.$$

This is just a MA(12) model with the MA polynomial $(1 + \theta z)(1 + \Theta z^{12})$.

It is easy to check that for this model:

$$\gamma_X(h) = (1 + \theta^2)(1 + \Theta^2)\sigma_Z^2,$$

$$\rho_x(1) = \frac{\theta}{1 + \theta^2} \quad \text{and} \quad \rho_X(12) = \frac{\Theta}{1 + \Theta^2}$$

and

$$\rho_X(11) = \rho_X(13) = \frac{\theta\Theta}{(1 + \theta^2)(1 + \Theta^2)}.$$

At every other lag, the autocorrelation $\rho_X(h)$ equals zero.

More generally, we can consider ARMA models with AR polynomial $\phi(z)\Phi(z)$ and MA polynomial $\theta(z)\Theta(z)$ where

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p,$$
$$\Phi(z) = 1 - \Phi_1 z^s - \Phi_2 z^{2s} - \cdots - \Phi_p z^{Ps}$$

and

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q,$$
$$\Theta(z) = 1 + \Theta_1 z^s + \Theta_2 z^{2s} + \cdots + \Theta_Q z^{Qs}.$$

This is called the **multiplicative seasonal ARMA**$(p, q) \times (P, Q)_s$ **model with seasonal period** s.

In the co2 example, we wanted to use such a model to the first and seasonal differenced data. Specifically, we want to use the model ARMA$(0, 1) \times (0, 1)_{12}$ to the seasonal and first differenced data: $\nabla\nabla_{12}X_t$. A sequence $\{Y_t\}$ is said to be a **multiplicative seasonal ARIMA model** with nonseasonal orders p, d, q, seasonal orders P, D, Q and seasonal period s if the differenced series $\nabla^d\nabla_s^d Y_t$ satisfies an ARMA$(p, q) \times (P, Q)_s$ model with seasonal period s.

Therefore, we want to fit the multiplicative seasonal ARIMA model with nonseasonal orders $0, 1, 1$ and seasonal orders $0, 1, 1$ with seasonal period 12 to the co2 dataset. This model can be fit to the data using the function *arima*() with the *seasonal* argument.

5 Overfitting as a Diagnostic Tool

After fitting an adequate model to the data, fit a slightly more general model. For example, if an AR(2) model seems appropriate, overfit with an AR(3) model. The original AR(2) model can be confirmed if while fitting the AR(3) model:

1. The estimate of the additional ϕ_3 parameter is not significantly different from zero.
2. The estimates of the common parameters, ϕ_1 and ϕ_2, do not change significantly from their original estimates.

How does one choose this general model to overfit? While fitting a more general model, one should not increase the order of both the AR and MA models. Because it leads to lack of identifiability issues. For example: consider the MA(1) model: $X_t = (1+\theta B)Z_t$. Then by multiplying by the polynomial $1-\phi z$ on both sides: we see that X_t also satisfies the ARMA(1, 2) model: $X_t - \phi X_{t-1} = Z_t + (\theta - \phi)Z_{t-1} + \phi\theta Z_{t-2}$.

But note that the parameter ϕ is not unique and thus if we fit an ARMA(1, 2) model to a dataset that is from MA(1), we might just get an arbitrary estimate for ϕ.

In general, it is a good idea to find the general overfitting model based on the analysis of the residuals. For example, if after fitting an MA(1) model, a not too small correlation remains at lag 2 in the residuals, then overfit with an MA(2) and not ARMA(1, 1) model.

Spring 2013 Statistics 153 (Time Series) : Lecture Sixteen

Aditya Guntuboyina

19 March 2013

1 Seasonal ARMA Models

The doubly infinite sequence $\{X_t\}$ is said to be a seasonal ARMA(P, Q) process with period s if it is stationary and if it satisfies the difference equation $\Phi(B^s)X_t = \Theta(B^s)Z_t$ where $\{Z_t\}$ is white noise and

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

and

$$\Theta(B^s) = 1 + \Theta_1 B^2 + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}.$$

Note that these can also be viewed as ARMA(Ps, Qs) models. However note that these models have $P + Q + 1$ (the 1 is for σ^2) parameters while a general ARMA(Ps, Qs) model will have $Ps + Qs + 1$ parameters. So these are much sparser models.

Unique Stationary solution exists to $\Phi(B^s)X_t = \Theta(B^s)Z_t$ if and only if every root of $\Phi(z^s)$ has magnitude different from one. Causal stationary solution exists if and only if every root of $\Phi(z^s)$ has magnitude strictly larger than one. Invertible stationary solution exists if and only if every root of $\Theta(z^s)$ has magnitude strictly larger than one.

The ACF and PACF of these models are **non-zero** only at the seasonal lags $h = 0, s, 2s, 3s, \ldots$. At these seasonal lags, the ACF and PACF of these models behave just as the case of the unseasonal ARMA model: $\Phi(B)X_t = \Theta(B)Z_t$.

2 Multiplicative Seasonal ARMA Models

In the co2 dataset, for the first and seasonal differenced data, we needed to fit a stationary model with non-zero autocorrelations at lags 1, 11, 12 and 13 (and zero autocorrelation at all other lags). We can use a MA(13) model to this data but that will have 14 parameters and therefore will likely overfit the data. We can get a much more parsimonious model for this dataset by *combining* the MA(1) model with a seasonal MA(1) model of period 12. Specifically, consider the model

$$X_t = (1 + \Theta B^{12})(1 + \theta B)Z_t.$$

This model has the autocorrelation function:

$$\rho_x(1) = \frac{\theta}{1 + \theta^2} \quad \text{and} \quad \rho_X(12) = \frac{\Theta}{1 + \Theta^2}$$

and

$$\rho_X(11) = \rho_X(13) = \frac{\theta\Theta}{(1 + \theta^2)(1 + \Theta^2)}.$$

At every other lag, the autocorrelation $\rho_X(h)$ equals zero. This is therefore a suitable model for the first and seasonal differenced data in the co2 dataset.

More generally, we can combine, by multiplication, ARMA and seasonal ARMA models to obtain models which have special autocorrelation properties with respect to seasonal lags:

The **Multiplicative Seasonal Autoregressive Moving Average Model** ARMA$(p, q) \times (P, Q)_s$ is defined as the stationary solution to the difference equation:

$$\Phi(B^s)\phi(B)X_t = \Theta(B^s)\theta(B)Z_t.$$

The model we looked at above for the co2 dataset is ARMA$(0, 1) \times (0, 1)_{12}$.

Another example of a multiplicative seasonal ARMA model is the ARMA$(0, 1) \times (1, 0)_{12}$ model:

$$X_t - \Phi X_{t-12} = Z_t + \theta Z_{t-1}.$$

The autocorrelation function of this model can be checked to be $\rho_X(h) = \Phi^h$ for $h \geq 0$ and

$$\rho_X(12h-1) = \rho_X(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h \qquad \text{for } h = 0, 1, 2, \ldots$$

and $\rho_X(h) = 0$ at all other lags.

When you get a stationary dataset whose correlogram shows interesting patterns at seasonal lags, consider using a multiplicative seasonal ARMA model. You may use the R function *ARMAacf* to understand the autocorrelation and partial autocorrelation functions of these models.

3 SARIMA Models

These models are obtained by combining differencing with multiplicative seasonal ARMA models. These models are denoted by ARIMA$(p, d, q) \times (P, D, Q)_s$. This means that after differencing d times and seasonal differencing D times, we get a multiplicative seasonal ARMA model. In other words, $\{Y_t\}$ is ARIMA$(p, d, q) \times (P, D, Q)_s$ if it satisfies the difference equation:

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d Y_t = \delta + \Theta(B^s)\theta(B)Z_t.$$

Recall that $\nabla_s^d = (1 - B^s)^d$ and $\nabla^d = (1 - B)^d$ denote the differencing operators.

In the co2 example, we wanted to use the model ARMA$(0, 1) \times (0, 1)_{12}$ to the seasonal and first differenced data: $\nabla\nabla_{12}X_t$. In other words, we want to fit the SARIMA model with nonseasonal orders $0, 1, 1$ and seasonal orders $0, 1, 1$ with seasonal period 12 to the co2 dataset. This model can be fit to the data using the function *arima*() with the *seasonal* argument.

4 AIC

AIC stands for Akaike's Information Criterion. It is a model selection criterion that recommends choosing a model for which:

$$AIC = -2\log(\text{maximum likelihood}) + 2k$$

is the smallest. Here k denotes the number of parameters in the model. For example, in the case of an ARMA(p, q) model with a non-zero mean μ, we have $k = p + q + 2$.

The first term in the definition of AIC measures the fit of the model i.e., the model performance on the given data set. The term $2k$ serves as a penalty function which penalizes models with too many parameters.

While comparing a bunch of models for a given dataset, you may use the AIC. There are other criteria as well. For example, the BIC (Bayesian Information Criterion) looks at:

$$BIC = -2\log(\text{maximum likelihood}) + k\log n.$$

Note that the penalty above is larger than that of AIC. Consequently, BIC selects more parsimonious models compared to AIC.

5 Time Series Cross Validation

Read the two articles on Rob Hyndman's blog: `http://robjhyndman.com/hyndsight/crossvalidation/` for a simple introduction to cross validation in general and `http://robjhyndman.com/hyndsight/tscvexample/` for cross validation specific to time series.

There are many ways to do cross validation for time series. Suppose we have monthly data for m years $x_1, \ldots, x_n$ where $n = 12m$ and the objective is to predict the data for the next year (This is similar to the midterm problem which has weekly data instead of monthly). Suppose we have ℓ competing models $M_1, \ldots, M_\ell$ for the dataset. We can use cross-validation in order to pick one of these models in the following way:

1. Fix a model M_i. Fix $k < m$.
2. Fit the model M_i to the data from the first k years.
3. Using the fitted model, predict the data for the $(k+1)$st year.
4. Calculate the sum of squares of errors of prediction for the $(k+1)$st year.
5. Repeat these steps for $k = k_0, \ldots, m-1$ where k_0 is an arbitrary value of your choice.
6. Average the sum of squares of errors of prediction over $k = k_0, \ldots, m-1$. Denote this value by CV_i and call it the Cross Validation score of model M_i.
7. Calculate CV_i for each $i = 1, \ldots, \ell$ and choose the model with the smallest Cross-Validation score.

Spring 2013 Statistics 153 (Time Series) : Lecture Seventeen

Aditya Guntuboyina

21 March 2013

1 Overfitting as a Diagnostic Tool

After fitting an adequate model to the data, fit a slightly more general model. For example, if an AR(2) model seems appropriate, overfit with an AR(3) model. The original AR(2) model can be confirmed if while fitting the AR(3) model:

1. The estimate of the additional ϕ_3 parameter is not significantly different from zero.
2. The estimates of the common parameters, ϕ_1 and ϕ_2, do not change significantly from their original estimates.

How does one choose this general model to overfit? While fitting a more general model, one should not increase the order of both the AR and MA models. Because it leads to lack of identifiability issues. For example: consider the MA(1) model: $X_t = (1+\theta B)Z_t$. Then by multiplying by the polynomial $1-\phi z$ on both sides: we see that X_t also satisfies the ARMA(1, 2) model: $X_t - \phi X_{t-1} = Z_t + (\theta - \phi)Z_{t-1} + \phi\theta Z_{t-2}$. But note that the parameter ϕ is not unique and thus if we fit an ARMA(1, 2) model to a dataset that is from MA(1), we might just get an arbitrary estimate for ϕ.

In general, it is a good idea to find the general overfitting model based on the analysis of the residuals. For example, if after fitting an MA(1) model, a not too small correlation remains at lag 2 in the residuals, then overfit with an MA(2) and not ARMA(1, 1) model.

2 Sines and Cosines

Frequency domain techniques: Using sines and cosines to study time series data.

Sinusoid: $R\cos(2\pi ft + \Phi)$. The following terminology is standard. R is called the *amplitude*, f is called the *frequency* and Φ is called the *phase*. The quantity $1/f$ is called the *period* and $2\pi f$ is termed the *angular frequency*. Note that three parameters R, f and Φ are involved in the definition of the sinusoid.

The function can also be written as $A\cos 2\pi ft + B\sin 2\pi ft$ where $A = R\cos\Phi$ and $B = R\sin\Phi$. This parametrization also has three parameters: f, A and B.

Spring 2013 Statistics 153 (Time Series) : Lecture Eighteen

Aditya Guntuboyina

2 April 2013

1 The Sinusoid

Sinusoid: $R\cos(2\pi ft + \Phi)$. The following terminology is standard. R is called the *amplitude*, f is called the *frequency* and Φ is called the *phase*. The quantity $1/f$ is called the *period* and $2\pi f$ is termed the *angular frequency*. Note that three parameters R, f and Φ are involved in the definition of the sinusoid.

The function can also be written as $A\cos 2\pi ft + B\sin 2\pi ft$ where $A = R\cos\Phi$ and $B = R\sin\Phi$. This parametrization also has three parameters: f, A and B.

Yet another way of representing the sinusoid is to use complex exponentials:

$$\exp(2\pi ift) = \cos(2\pi ft) + i\sin(2\pi ft).$$

Therefore

$$\cos(2\pi ft) = \frac{\exp(2\pi ift) + \exp(-2\pi ift)}{2} \text{ and } \sin(2\pi ft) = \frac{\exp(2\pi ift) - \exp(-2\pi ift)}{2i}.$$

Thus $A\cos 2\pi ft + B\sin 2\pi ft$ can also be written as a linear combination of $\exp(2\pi ift)$ and $\exp(-2\pi ift)$.

Sinusoids at certain special frequencies have nice orthogonality properties. Consider the vector:

$$u = (1, \exp(2\pi i/n), \exp(2\pi i2/n), \ldots, \exp(2\pi i(n-1)/n)).$$

This is the sinusoid $\exp(2\pi ift)$ at frequency $f = 1/n$ evaluated at the time points $t = 0, 1, 2, \ldots, (n-1)$.

Also define the vector corresponding to the sinusoid $\exp(2\pi ift)$ at frequency $f = j/n$ evaluated at $t = 0, 1, \ldots, (n-1)$ by

$$u^j = (1, \exp(2\pi ij/n), \exp(2\pi i2j/n), \ldots, \exp(2\pi i(n-1)j/n)).$$

These vectors $u^0, u^1, u^2, \ldots, u^{n-1}$ are orthogonal. In other words, the dot product between u^k and u^l is zero if $k \neq l$. Therefore, every data vector $x := (x_1, \ldots, x_n)$ can be written as a linear combination of $u^0, u^1, \ldots, u^{n-1}$.

2 The Discrete Fourier Transform

Let $x_0, \ldots, x_{n-1}$ denote data (real numbers) of length n.

The DFT of $\{x_t\}$ is given by $b_j, j = 0, 1, \ldots, n-1$, where

$$b_j = \sum_{t=0}^{n-1} x_t \exp\left(-\frac{2\pi ijt}{n}\right) \qquad \text{for } j = 0, \ldots, n-1.$$

Therefore, $b_0 = \sum x_t$. Also for $1 \le j \le n-1$,

$$b_{n-j} = \sum_t x_t \exp\left(-\frac{2\pi i(n-j)t}{n}\right) = \sum_t x_t \exp\left(\frac{2\pi ijt}{n}\right) \exp\left(-2\pi it\right) = \bar{b}_j.$$

Note that this relation only holds if $x_0, \ldots, x_{n-1}$ are real numbers.

Thus for $n = 11$, the DFT can be written as:

$$b_0, b_1, b_2, b_3, b_4, b_5, \bar{b}_5, \bar{b}_4, \bar{b}_3, \bar{b}_2, \bar{b}_1.$$

and for $n = 12$, it is

$$b_0, b_1, b_2, b_3, b_4, b_5, b_6 = \bar{b}_6, \bar{b}_5, \bar{b}_4, \bar{b}_3, \bar{b}_2, \bar{b}_1.$$

Note that b_6 is necessarily real because $b_6 = \bar{b}_6$.

The DFT is calculated by the R function `fft()`.

The original data $x_0, \ldots, x_{n-1}$ can be recovered from the DFT using:

$$x_t = \frac{1}{n} \sum_{j=0}^{n-1} b_j \exp\left(\frac{2\pi ijt}{n}\right) \qquad \text{for } t = 0, \ldots, n-1.$$

Thus for $n = 11$, one can think of the data as the 11 real numbers $x_0, x_1, \ldots, x_{10}$ or, equivalently, as one real number b_0 along with 5 complex numbers $b_1, \ldots, b_5$.

For $n = 12$, one can think of the data as the 12 real numbers $x_0, \ldots, x_{11}$ or, equivalenlty, as two real numbers, b_0 and b_6, along with the 5 complex numbers $b_1, \ldots, b_5$.

The following identity always holds:

$$n \sum_t x_t^2 = \sum_{j=0}^{n-1} |b_j|^2 .$$

Because $b_{n-j} = \bar{b}_j$, their absolute values are equal and we can write the above sum of squares identity in the following way. For $n = 11$,

$$n \sum_t x_t^2 = b_0^2 + 2|b_1|^2 + 2|b_2|^2 + 2|b_3|^2 + 2|b_4|^2 + 2|b_5|^2$$

and for $n = 12$,

$$n \sum_t x_t^2 = b_0^2 + 2|b_1|^2 + 2|b_2|^2 + 2|b_3|^2 + 2|b_4|^2 + 2|b_5|^2 + b_6^2.$$

Note that there is no need to put an absolute value on b_6 because it is real.

Because $b_0 = \sum_t x_t = n\bar{x}$, we have

$$n \sum_t x_t^2 - b_0^2 = n \sum_t x_t^2 - n^2 \bar{x}^2 = n \sum_t (x_t - \bar{x})^2.$$

Thus the sum of squares identity can be written for n odd, say $n = 11$, as

$$\sum_t (x_t - \bar{x})^2 = \frac{2}{n}|b_1|^2 + \frac{2}{n}|b_2|^2 + \frac{2}{n}|b_3|^2 + \frac{2}{n}|b_4|^2 + \frac{2}{n}|b_5|^2$$

and, for n even, say $n = 12$, as

$$\sum_t (x_t - \bar{x})^2 = \frac{2}{n}|b_1|^2 + \frac{2}{n}|b_2|^2 + \frac{2}{n}|b_3|^2 + \frac{2}{n}|b_4|^2 + \frac{2}{n}|b_5|^2 + \frac{1}{n}b_6^2.$$

Spring 2013 Statistics 153 (Time Series) : Lecture Nineteen

Aditya Guntuboyina

4 April 2013

1 DFT Recap

Given data $x_0, \ldots, x_{n-1}$, their DFT is given by $b_0, b_1, \ldots, b_{n-1}$ where

$$b_j := \sum_{t=0}^{n-1} x_t \exp\left(-\frac{2\pi i j t}{n}\right) \qquad \text{for } j = 0, 1, \ldots, n-1.$$

Two key things to remember are:

1. $b_0 = x_0 + \cdots + x_{n-1}$
2. $b_{n-j} = \bar{b}_j$ for $1 \leq j \leq n-1$

For odd values of n, say $n = 11$, the DFT is comprised of the real number b_0 and the $(n-1)/2$ complex numbers $b_1, \ldots, b_{(n-1)/2}$.

For even values of n, say $n = 12$, the DFT consists of two real numbers b_0 and $b_{n/2}$ and the $(n-2)/2$ complex numbers $b_1, \ldots, b_{(n-2)/2}$.

The original data $x_0, x_1, \ldots, x_{n-1}$ can be recovered from the DFT by the formula:

$$x_t = \frac{1}{n} \sum_{j=0}^{n-1} b_j \exp\left(\frac{2\pi i j t}{n}\right) \qquad \text{for } t = 0, 1, \ldots, n-1.$$

This formula can be written succintly as:

$$x = \frac{1}{n} \sum_{j=0}^{n-1} b_j u^j$$

where $x = (x_0, \ldots, x_{n-1})$ denotes the data vector and u^j denotes the vector obtained by evaluating the sinusoid $\exp(2\pi i j t/n)$ at times $t = 0, 1, \ldots, n-1$. We have seen in the last class that $u^0, u^1, \ldots, u^{n-1}$ are orthogonal with $u^j \cdot u^j = n$ for each j.

As a result, we have the sum of squares identity:

$$n \sum_t x_t^2 = \sum_{j=0}^{n-1} |b_j|^2 .$$

The absolute values of b_j and b_{n-j} are equal because $b_{n-j} = \bar{b}_j$ and hence we can write the above sum of squares identity in the following way. For $n = 11$,

$$n \sum_t x_t^2 = b_0^2 + 2|b_1|^2 + 2|b_2|^2 + 2|b_3|^2 + 2|b_4|^2 + 2|b_5|^2$$

and for $n = 12$,

$$n\sum_t x_t^2 = b_0^2 + 2|b_1|^2 + 2|b_2|^2 + 2|b_3|^2 + 2|b_4|^2 + 2|b_5|^2 + b_6^2.$$

Note that there is no need to put an absolute value on b_6 because it is real.

Because $b_0 = \sum_t x_t = n\bar{x}$, we have

$$n\sum_t x_t^2 - b_0^2 = n\sum_t x_t^2 - n^2\bar{x}^2 = n\sum_t (x_t - \bar{x})^2.$$

Thus the sum of squares identity can be written for n odd, say $n = 11$, as

$$\sum_t (x_t - \bar{x})^2 = \frac{2}{n}|b_1|^2 + \frac{2}{n}|b_2|^2 + \frac{2}{n}|b_3|^2 + \frac{2}{n}|b_4|^2 + \frac{2}{n}|b_5|^2$$

and, for n even, say $n = 12$, as

$$\sum_t (x_t - \bar{x})^2 = \frac{2}{n}|b_1|^2 + \frac{2}{n}|b_2|^2 + \frac{2}{n}|b_3|^2 + \frac{2}{n}|b_4|^2 + \frac{2}{n}|b_5|^2 + \frac{1}{n}b_6^2.$$

2 DFT of the Cosine Wave

Let $x_t = R\cos(2\pi f_0 t + \phi)$ for $t = 0, \ldots, n-1$. We have seen in R that when f_0 is a Fourier frequency (i.e., of the form k/n for some k), the DFT has exactly one spike but when f_0 is not a Fourier frequency, there is leakage. We prove this here.

We can, without loss of generality, assume that $0 \le f_0 \le 1/2$ because:

1. If $f_0 < 0$, then we can write $\cos(2\pi f_0 t + \phi) = \cos(2\pi(-f_0)t - \phi)$. Clearly, $-f_0 \ge 0$.
2. If $f_0 \ge 1$, then we write

   $$\cos(2\pi f_0 t + \phi) = \cos(2\pi[f_0]t + 2\pi(f - [f_0])t + \phi) = \cos(2\pi(f - [f_0])t + \phi),$$

   because $\cos(\cdot)$ is periodic with period 2π. Clearly $0 \le f - [f_0] < 1$.
3. If $f_0 \in [1/2, 1)$, then

   $$\cos(2\pi f_0 t + \phi) = \cos(2\pi t - 2\pi(1 - f_0)t + \phi) = \cos(2\pi(1 - f_0)t - \phi)$$

   because $\cos(2\pi t - x) = \cos x$ for all integers t. Clearly $0 < 1 - f_0 \le 1/2$.

Thus given a cosine wave $R\cos(2\pi f t + \phi)$, one can always write it as $R\cos(2\pi f_0 t + \phi')$ with $0 \le f_0 \le 1/2$ and a phase ϕ' that is possibly different from ϕ. This frequency f_0 is said to be an **alias** of f. From now on, whenever we speak of the cosine wave $R\cos(2\pi f_0 t + \phi)$, we assume that $0 \le f_0 \le 1/2$.

If $\phi = 0$, then we have $x_t = R\cos(2\pi f_0 t)$. When $f_0 = 0$, then $x_t = R$ and so there is no oscillation in the data at all. When $f_0 = 1/2$, then $x_t = R\cos(\pi t) = R(-1)^t$ and so $f_0 = 1/2$ corresponds to the maximum possible oscillation.

What is the DFT of $x_t = R\cos(2\pi f_0 t + \phi)$ for $0 \le f_0 \le 1/2$? The formula is

$$b_j := \sum_{t=0}^{n-1} x_t \exp\left(-\frac{2\pi i j t}{n}\right).$$

Suppose $f = j/n$ and we shall calculate

$$b(f) = \sum_{t=0}^{n-1} x_t \exp\left(-2\pi i f t\right).$$

The easiest way to calculate this DFT is to write the cosine wave in terms of complex exponentials:

$$x_t = \frac{R}{2}\left(e^{2\pi i f_0 t} e^{i\phi} + e^{-2\pi i f_0 t} e^{-i\phi}\right).$$

It is therefore convenient to first calculate the DFT of the complex exponential $e^{2\pi i f_0 t}$.

2.1 DFT of $y_t = e^{2\pi i f_0 t}$

The DFT of $y_t = e^{2\pi i f_0 t}$ is given by

$$\sum_{t=0}^{n-1} y_t e^{-2\pi i f t} = \sum_{t=0}^{n-1} e^{2\pi i (f_0 - f) t}$$

where $f = j/n$. Let us denote this by $S_n(f_0 - f)$ i.e.,

$$S_n(f_0 - f) = \sum_{t=0}^{n-1} e^{2\pi i (f_0 - f) t}. \tag{1}$$

This can clearly be evaluated using the geometric series formula to be

$$S_n(f_0 - f) = \frac{e^{2\pi i (f_0 - f) n} - 1}{e^{2\pi i (f_0 - f)} - 1}$$

It is easy to check that

$$e^{i\theta} - 1 = \cos\theta + i \sin\theta - 1 = 2e^{i\theta/2} \sin\theta/2.$$

As a result

$$S_n(f_0 - f) = \frac{\sin \pi n (f_0 - f)}{\sin \pi (f_0 - f)} e^{i\pi (f_0 - f)(n-1)}$$

Thus the absolute value of the DFT of $y_t = e^{2\pi i f_0 t}$ is given by

$$|S_n(f - f_0)| = \left|\frac{\sin \pi n (f_0 - f)}{\sin \pi (f_0 - f))}\right| \qquad \text{where } f = j/n$$

This expression becomes meaningless when $f_0 = f$. But when $f_0 = f$, the value of $S_n(f_0 - f)$ can be directly be calculated from (1) to be equal to n.

The behavior of $|S_n(f - f_0)|$ can be best understood by plotting the function $g \mapsto (\sin \pi n g)/(\sin \pi g)$. This explains leakage.

The behavior of the DFT of the cosine wave can be studied by writing it in terms of the DFT of the complex exponential.

If f_0 is not of the form k/n for any j, then the term $S_n(f - f_0)$ is non-zero for all f of the form j/n. This situation where one observes a non-zero DFT term b_j because of the presence of a sinusoid at a frequency f_0 different from j/n is referred to as **Leakage**.

Leakage due to a sinusoid with frequency f_0 not of the form k/n is present in all DFT terms b_j but the magnitude of the presence decays as j/n gets far from f_0. This is because of the form of the function $S_n(f - f_0)$.

There are two problems with Leakage:

1. Fourier analysis is typically used to separate out the effects due to different frequencies; so leakage is an undesirable phenomenon.

2. Leakage at j/n due to a sinusoid at frequency f_0 can mask the presence of a true sinusoid at frequency j/n.

How to get rid of leakage? The easy answer is to choose n appropriately (ideally, n should be a multiple of the periods of all oscillations). For example, if it is monthly data, then it is better to have whole year's worth of data. But this is not always possible. We will study a leakage-reducing technique later.

3 DFT of a Periodic Series

Suppose that the data $x_0, x_1, \ldots, x_{n-1}$ is periodic with period h i.e., $x_{t+hu} = x_t$ for all integers t and u. Let n be an integer multiple of h i.e., $n = kh$. For example, suppose we have monthly data collected over 10 years in which case: $h = 12$, $k = 10$ and $n = 120$.

Suppose that DFT of the data $x_0, \ldots, x_{n-1}$ is $b_0, b_1, \ldots, b_{n-1}$. Suppose also that the DFT of the data in the first cycle: $x_0, x_1, \ldots, x_{h-1}$ is $\beta_0, \beta_1, \ldots, \beta_{h-1}$.

We shall express b_j in terms of $\beta_0, \ldots, \beta_{h-1}$. Let $f = j/n$ for simplicity.

By definition

$$b_j = \sum_{t=0}^{n-1} x_t \exp\left(-2\pi i t f\right).$$

Break up the sum into

$$\sum_{t=0}^{h-1} + \sum_{t=h}^{2h-1} + \cdots + \sum_{t=(k-1)h}^{kh-1}$$

The lth term above can be evaluated as:

$$\begin{aligned}\sum_{t=(l-1)h}^{lh-1} x_t \exp\left(-2\pi i f t\right) &= \sum_{s=0}^{h-1} x_s \exp\left(-2\pi i f (s + (l-1)h)\right) \\ &= \exp\left(-2\pi i f (l-1)h\right) \sum_{s=0}^{h-1} x_s \exp(-2\pi i f s).\end{aligned}$$

Therefore

$$\begin{aligned} b_j &= \sum_{l=1}^{k} \exp\left(-2\pi i f(l-1)h\right) \sum_{s=0}^{h-1} x_s \exp(-2\pi i f s) \\ &= \sum_{s=0}^{h-1} x_s \exp(-2\pi i f s) \sum_{l=1}^{k} \exp\left(-2\pi i f(l-1)h\right) \\ &= S_k(fh) \sum_{s=0}^{h-1} x_s \exp(-2\pi i f s) \\ &= S_k(jh/n) \sum_{s=0}^{h-1} x_s \exp(-2\pi i j s/n) \\ &= S_k(j/k) \sum_{s=0}^{h-1} x_s \exp(-2\pi i (j/k) s/h) \end{aligned}$$

Thus $b_j = 0$ if j is not a multiple of k and when j is a multiple of k, then $|b_j| = k|\beta_{j/k}|$.

Thus the original DFT terms $\beta_0, \beta_1, \ldots, \beta_{h-1}$ now appear as $b_0 = k\beta_0, b_k = k\beta_1, b_{2k} = k\beta_2$ etc. until $b_{(h-1)k} = k\beta_{h-1}$. All other b_js are zero.

4 DFT and Sample Autocovariance Function

We show below that

$$\frac{|b_j|^2}{n} = \sum_{|h|<n} \hat{\gamma}(h) \exp\left(-\frac{2\pi i j h}{n}\right) \qquad \text{for } j = 1, \ldots, n-1$$

where $\hat{\gamma}(h)$ is the sample autocovariance function. This gives an important connection between the dft and the sample autocovariance function.

To see this, observe first, by the formula for the sum of a geometric series, that

$$\sum_{t=0}^{n-1} \exp\left(-\frac{2\pi i j t}{n}\right) = 0 \qquad \text{for } j = 1, \ldots, n-1.$$

In other words, if the data is constant i.e., $x_0 = \cdots = x_{n-1}$, then b_0 equals nx_0 and b_j equals 0 for all other j. Because of this, we can write:

$$b_j = \sum_{t=0}^{n-1} (x_t - \bar{x}) \exp\left(-\frac{2\pi i j t}{n}\right) \qquad \text{for } j = 1, \ldots, n-1.$$

Therefore, for $j = 1, \ldots, n-1$, we write

$$\begin{aligned}
|b_j|^2 = b_j \bar{b}_j &= \sum_{t=0}^{n-1} \sum_{s=0}^{n-1} (x_t - \bar{x})(x_s - \bar{x}) \exp\left(-\frac{2\pi i j t}{n}\right) \exp\left(\frac{2\pi i j s}{n}\right) \\
&= \sum_{t=0}^{n-1} \sum_{s=0}^{n-1} (x_t - \bar{x})(x_s - \bar{x}) \exp\left(-\frac{2\pi i j (t-s)}{n}\right) \\
&= \sum_{h=-(n-1)}^{n-1} \sum_{t,s:t-s=h} (x_t - \bar{x})(x_{t-h} - \bar{x}) \exp\left(-\frac{2\pi i j h}{n}\right) \\
&= n \sum_{|h|<n} \hat{\gamma}(h) \exp\left(-\frac{2\pi i j h}{n}\right).
\end{aligned}$$

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty

Aditya Guntuboyina

9 April 2013

1 DFT Again

Data is denoted by $x_0, x_1, \ldots, x_{n-1}$.

DFT is denoted by $b_0, b_1, \ldots, b_{n-1}$.

The DFT is calculated from data by

$$b_j := \sum_{t=0}^{n-1} x_t \exp\left(-\frac{2\pi i j t}{n}\right) \qquad \text{for } j = 0, 1, \ldots, n-1. \tag{1}$$

The data is calculated from the DFT by

$$x_t = \frac{1}{n} \sum_{j=0}^{n-1} b_j \exp\left(\frac{2\pi i j t}{n}\right) \qquad \text{for } t = 0, 1, \ldots, n-1.$$

Remember that $b_0 = x_0 + \cdots + x_{n-1}$ and $b_{n-j} = \bar{b}_j$ for $1 \le j \le n-1$. In the textbook, the formula for the DFT is given by (1) with an extra factor of $n^{-1/2}$. I have dropped this factor to make the definition compatible with the R function *fft*.

For odd values of n, the DFT is comprised of the real number b_0 and the $(n-1)/2$ complex numbers $b_1, \ldots, b_{(n-1)/2}$.

For even values of n, the DFT consists of two real numbers b_0 and $b_{n/2}$ and the $(n-2)/2$ complex numbers $b_1, \ldots, b_{(n-2)/2}$.

2 What does the DFT do?

Suppose $x_t = R\cos(2\pi f_0 t + \Phi)$ for $t = 0, 1, \ldots, n-1$. We have seen in the last class that we only have to consider frequencies in the range $0 \le f_0 \le 1/2$ (because every other frequency has an *alias* in the interval $[0, 1/2]$).

Assume first that f_0 is of the form k/n for some k where $0 \le k/n \le 1/2$. Then the DFT is given by

$$\begin{aligned} b_j &= \sum_{t=0}^{n-1} R\cos(2\pi(k/n)t + \Phi)\exp\left(-2\pi i(j/n)t\right) \\ &= \frac{Re^{i\Phi}}{2} \sum_{t=0}^{n-1} \exp\left(2\pi i t \frac{j-k}{n}\right) + \frac{Re^{-i\Phi}}{2} \sum_{t=0}^{n-1} \exp\left(-2\pi i t \frac{j+k}{n}\right). \end{aligned}$$

Note that we do not need to consider the DFT b_j for $j/n > 1/2$. So we assume that $0 \leq j/n \leq 1/2$. Because the original cosine wave was assumed to have frequency in the range $[0, 1/2]$, we have $0 \leq k/n \leq 1/2$. Check that $0 < (k+j)/n < 1$ when $j \neq k$. Because of all this, we get that the second term above is always zero and the first term equals zero when $j \neq k$ and equals $Re^{i\Phi}n/2$ when $j = k$. Therefore the DFT of the cosine wave with a frequency k/n for $0 \leq k/n \leq 1/2$ is $b_k = nRe^{i\Phi}/2$ for and $b_j = 0$ for $j \neq k$ and $0 \leq j/n \leq 1/2$.

Now consider data that is linear combination of multiple frequencies:

$$x_t = \sum_{l=1}^{m} R_l \cos(2\pi t(k_l/n) + \Phi_l) \tag{2}$$

where each k_l is an integer satisfying $0 \leq k_l/n \leq 1/2$. Because the definition of the DFT is linear in the data $\{x_t\}$, it follows that the DFT of (2) is given by

$$b_j = \begin{cases} nR_l e^{i\Phi_l}/2 & \text{if } j = k_l \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq j/n \leq 1/2$.

This shows that the DFT picks out the frequencies present in the data. The strength (absolute value) of the DFT at a frequency is proportional to the amplitude (R_l) of the cosine wave at that frequency.

3 Interpreting the DFT

The DFT writes the given data in terms of sinusoids with frequencies of the form k/n. Frequencies of the form k/n are called Fourier frequencies.

Suppose that we are given a dataset $x_0, \ldots, x_{n-1}$. We have calculated its DFT: $b_0, b_1, \ldots, b_{n-1}$ and we have plotted $|b_j|$ for $j = 1, \ldots, (n-1)/2$ for odd n and for $j = 1, \ldots, n/2$ for even n.

If we see a single spike in this plot, say at b_k, we are sure that the data is a sinusoid with frequency k/n.

If we get two spikes, say at b_{k_1} and b_{k_2}, then the data is slightly more complicated: it is a linear combination of two sinusoids at frequencies k_1/n and k_2/n with the strengths of these sinusoids depending on the size of the spikes.

Multiple spikes indicate that the data is made up of many sinusoids at Fourier frequencies and, in general, this means that the data is more complicated.

However, sometimes one can see multiple spikes in the DFT even when the structure of the data is not very complicated. A typical example is leakage due to the presence of a sinusoid at a non-Fourier frequency.

The DFT of a sinusoid at a non-Fourier frequency is calculated in the following way: Consider the signal $x_t = e^{2\pi f_0 t}$ where $f_0 \in [0, 1/2]$ is not necessarily of the form k/n for any k. Its DFT is given by

$$b_j := \sum_{t=0}^{n-1} x_t e^{-2\pi i t(j/n)} = \sum_{t=0}^{n-1} e^{2\pi i(f_0 - (j/n))t}.$$

If we denote the function

$$S_n(g) := \sum_{t=0}^{n-1} e^{2\pi i g t} \tag{3}$$

then we can write

$$b_j = S_n(f_0 - (j/n)).$$

The function $S_n(g)$ can clearly be evaluated using the geometric series formula to be

$$S_n(g) = \frac{e^{2\pi ign} - 1}{e^{2\pi ig} - 1}$$

Because

$$e^{i\theta} - 1 = \cos\theta + i\sin\theta - 1 = 2e^{i\theta/2}\sin\theta/2,$$

we get

$$S_n(g) = \frac{\sin \pi ng}{\sin \pi g} e^{i\pi g(n-1)}$$

Thus the absolute value of the DFT of $y_t = e^{2\pi i f_0 t}$ is given by

$$|b_j| = |S_n(f_0 - (j/n))| = \left| \frac{\sin \pi n(f_0 - (j/n))}{\sin \pi(f_0 - (j/n)))} \right|$$

This expression becomes meaningless when $f_0 = j/n$. But when $f_0 = f$, the value of $S_n(f_0 - j/n)$ can be directly be calculated from (3) to be equal to n.

The behavior of this DFT can be best understood by plotting the function $g \mapsto (\sin \pi ng)/(\sin \pi g)$.

4 Leakage Reduction by Hanning

Hanning is a technique to reduce leakage which says: Multiply the data by the *window* or *fader*:

$$w_t = 1 - \cos(2\pi t/n) \qquad \text{for } t = 0, 1, \ldots, n-1$$

and then take the DFT.

Why does it work? The following is the DFT of $y_t = w_t e^{2\pi i f_0 t}$ (below f stands for j/n)

$$\begin{aligned}
b_y(f) &:= \sum_{t=0}^{n-1} w_t e^{2\pi i f_0 t} e^{-2\pi i f t} \\
&= \sum_{t=0}^{n-1} (1 - \cos(2\pi t/n))\, e^{2\pi i (f_0 - f)t} \\
&= \sum_{t=0}^{n-1} e^{2\pi i (f_0 - f)t} - \frac{1}{2}\sum_{t=0}^{n-1} e^{2\pi i (f_0 - f + 1/n)t} - \frac{1}{2}\sum_{t=0}^{n-1} e^{2\pi i (f_0 - f - 1/n)t} \\
&= S_n(f_0 - f) - \frac{1}{2}S_n(f_0 - f + 1/n) - \frac{1}{2}S_n(f_0 - f - 1/n).
\end{aligned}$$

Clearly $b_y(f_0) = n = b(f_0)$ because $S_n(1/n) = 0$. Suppose $g = f_0 - f$. To eliminate leakage, we need to make sure that $b_y(f)$ is close to zero when f is not equal to f_0. This is not quite possible but what we shall heuristically show is that when $|f - f_0|$ is reasonably large compared to $1/n$, then $b_f(f)$ is close to zero.

Let g denote $f - f_0$ so that

$$b_y(f) = S_n(g) - \frac{1}{2}S_n(g - 1/n) - \frac{1}{2}S_n(g + 1/n). \tag{4}$$

We derived in the last section that

$$S_n(g) = \frac{\sin \pi ng}{\sin \pi g} e^{\pi ig(n-1)}.$$

Therefore,

$$S_n(g-1/n) = \frac{\sin \pi n(g-1/n)}{\sin \pi(g-1/n)} e^{\pi i(g-1/n)(n-1)}$$

Now $\sin \pi n(g-1/n) = -\sin \pi ng$ and if $1/n$ is small compared to g, then $\sin \pi(g-1/n) \approx \sin \pi g$. Also when $1/n$ is small compared to g, we have

$$e^{\pi i(g-1/n)(n-1)} = e^{\pi ig(n-1)} e^{-i\pi(n-1)/n} \approx e^{\pi ig(n-1)} e^{-i\pi} = -e^{\pi ig(n-1)}.$$

Therefore, if $1/n$ is small compared to g, we have

$$S_n(g-1/n) \approx S_n(g)$$

and similarly $S_n(g+1/n) \approx S_n(g)$. Therefore, from (4), we get $b_y(f) \approx 0$ provided $f - f_0$ is not too small compared to $1/n$. Also $b_y(f_0) = b(f_0)$. Thus leakage is reduced.

It is usually the case however that for f close to f_0, $b_y(f)$ is much larger than $b(f)$. Thus the price that is paid for the reduction in leakage is that the peaks are slightly rounder at the top compared to the peaks without hanning.

5 DFT and Sample Autocovariance Function

We show below that

$$\frac{|b_j|^2}{n} = \sum_{|h|<n} \hat{\gamma}(h) \exp\left(-\frac{2\pi ijh}{n}\right) \qquad \text{for } j = 1, \ldots, n-1$$

where $\hat{\gamma}(h)$ is the sample autocovariance function. This gives an important connection between the dft and the sample autocovariance function.

To see this, observe first, by the formula for the sum of a geometric series, that

$$\sum_{t=0}^{n-1} \exp\left(-\frac{2\pi ijt}{n}\right) = 0 \qquad \text{for } j = 1, \ldots, n-1.$$

In other words, if the data is constant i.e., $x_0 = \cdots = x_{n-1}$, then b_0 equals nx_0 and b_j equals 0 for all other j. Because of this, we can write:

$$b_j = \sum_{t=0}^{n-1} (x_t - \bar{x}) \exp\left(-\frac{2\pi ijt}{n}\right) \qquad \text{for } j = 1, \ldots, n-1.$$

Therefore, for $j = 1, \ldots, n-1$, we write

$$\begin{aligned}
|b_j|^2 = b_j \bar{b}_j &= \sum_{t=0}^{n-1} \sum_{s=0}^{n-1} (x_t - \bar{x})(x_s - \bar{x}) \exp\left(-\frac{2\pi ijt}{n}\right) \exp\left(\frac{2\pi ijs}{n}\right) \\
&= \sum_{t=0}^{n-1} \sum_{s=0}^{n-1} (x_t - \bar{x})(x_s - \bar{x}) \exp\left(-\frac{2\pi ij(t-s)}{n}\right) \\
&= \sum_{h=-(n-1)}^{n-1} \sum_{t,s:t-s=h} (x_t - \bar{x})(x_{t-h} - \bar{x}) \exp\left(-\frac{2\pi ijh}{n}\right) \\
&= n \sum_{|h|<n} \hat{\gamma}(h) \exp\left(-\frac{2\pi ijh}{n}\right).
\end{aligned}$$

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty One

Aditya Guntuboyina

11 April 2013

1 The Periodogram

In the last class, we saw the following connection between the DFT and the sample autocovariance function:

$$\frac{|b_j|^2}{n} = \sum_{h:|h|<n} \hat{\gamma}(h) \exp\left(-\frac{2\pi i j h}{n}\right) \qquad \text{for } j = 1, \ldots, [n/2].$$

The function

$$I(j/n) := \frac{|b_j|^2}{n} = \sum_{h:|h|<n} \hat{\gamma}(h) \exp\left(-\frac{2\pi i j h}{n}\right) \qquad \text{for } j = 1, \ldots, [n/2] \tag{1}$$

is called the *periodogram* of the data $x_0, x_1, \ldots, x_{n-1}$. **The periodogram gives the strengths of sinusoids at various frequencies in the data**.

2 The Spectral Density

Suppose $\{X_t\}$ is a doubly infinite sequence of random variables that is stationary. Let $\{\gamma(h)\}$ denote their autocovariance function. In analogy with the definition (1) of the Periodogram, we define

$$f(\lambda) := \sum_{h=-\infty}^{\infty} \gamma(h) \exp\left(-2\pi i \lambda h\right) \qquad \text{for } -1/2 \leq \lambda \leq 1/2 \tag{2}$$

and call this quantity the *Spectral Density* of the stationary sequence of random variables, $\{X_t\}$. Because the complex exponentials $e^{-2\pi i \lambda h}$ are all periodic in λ with period 1, we only need to define f on an interval of length 1 and, by convention, we focus on the interval $[-1/2, 1/2]$. In fact, note that f is symmetric and we really only need to worry about $[0, 1/2]$.

In analogy with the periodogram, the spectral density will give the strengths of sinusoids at various frequencies in the data.

We have defined the spectral density in terms of the autocovariance function. It turns that the autocovariance function can also be obtained from the spectral density: To see this, just multiply both sides of (2) by $e^{2\pi i \lambda k}$ for a fixed k and integrate from $\lambda = -1/2$ to $\lambda = 1/2$ to get:

$$\gamma(k) = \int_{1/2}^{1/2} e^{2\pi i \lambda k} f(\lambda) d\lambda \tag{3}$$

In other words, the autocovariance function and the spectral density provide equivalent information about the stationary process $\{X_t\}$.

There is one problem however with the definition of the spectral density. The infinite sum in (2) need not always make sense. Indeed, the complex exponentials $\exp(-2\pi i\lambda h)$ always have a magnitude of 1 and so the sum (2) only makes sense when $\{\gamma(h)\}$ decay sufficiently quickly. A sufficient (but not necessary) condition for (2) to make sense is $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$.

It is not too hard to find examples where the sum on the right hand side in (2) does not make sense. For example, consider the process where $X_t = A\cos 2\pi\lambda_1 t + B\sin 2\pi\lambda_1 t$ where A and B are uncorrelated random variables both with mean 0 and variance σ^2 and $0 < \lambda_1 < 1/2$ is a fixed (non-random) frequency. This process is clearly stationary and its autocovariance function equals $\gamma(h) = \sigma^2\cos 2\pi\lambda_1 h$. Clearly, this does not decay fast enough and $\sum_h \gamma(h)\exp(-2\pi i\lambda h)$ does not make sense for any λ.

It turns out that one may not be able to define a spectral density for every stationary process $\{X_t\}$. But one can always define a Spectral Distribution Function. The analogy is to random variables (Not all random variables have densities but they all have distribution functions).

Before defining the spectral distribution function, let us briefly discuss elementary expectations.

2.1 Review of Expectations

Let X be a random variable. The distribution function of X is defined as $F(x) = \mathbb{P}\{X \le x\}$. The function F is non-negative, right-continuous, non-decreasing and satisfies:

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = 1.$$

The expectation of a function $g(X)$ of X is sometimes denoted by:

$$\mathbb{E}g(X) = \int g(x)dF(x).$$

The computation of this expectation is rather easy in the following two cases:

1. X is a discrete random variable taking values $x_1 < \cdots < x_k$ with probabilities $p_1, \ldots, p_k$. In this case, F has a jump of size p_i at x_i and is constant between x_i and x_{i+1}. And, $\int g(x)dF(x) = \sum_i g(x_i)p_i$.
2. X has a density f. In this case, $F(x) = \int_{-\infty}^{x} f(x)dx$ and $\mathbb{E}g(X) = \int g(x)f(x)dx$.

The quantity $\int g(x)dF(x)$ can also be defined for F which are non-negative, right-continuous, non-decreasing and satisfy:

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} F(x) = \sigma^2$$

for some $\sigma^2 > 0$. In this case F/σ^2 is a distribution function and thus the integral $\int g(x)dF(x)$ can be defined as

$$\int g(x)dF(x) = \sigma^2 \int g(x)d\tilde{F}(x) \qquad \text{where } \tilde{F}(x) = \frac{F(x)}{\sigma^2}.$$

2.2 Spectral Distribution Function and Spectral Density

Let $\{X_t\}$ be a stationary sequence of random variables and let $\gamma_X(h) = \mathrm{cov}(X_t, X_{t+h})$ denote the autocovariance function.

A theorem due to Herglotz (sometimes attributed to Bochner) states that **every** autocovariance function γ_X can be written as:

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{2\pi i h\lambda} dF(\lambda),$$

where $F(\cdot)$ is a non-negative, right-continuous, non-decreasing function on $[-1/2, 1/2]$ with $F(-1/2) = 0$ and $F(1/2) = \gamma_X(0)$. Moreover, F is uniquely determined by γ_X.

If F has a density f, then f is called the *Spectral Density* of $\{X_t\}$.

A sufficient condition (but not necessary) for the existence of the spectral density is the condition $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$. And in this case, the spectral density exists and is given by the formula (2).

2.3 Discrete Spectrum Example

Suppose

$$X_t = \sum_{j=1}^{m} \left(A_j \cos(2\pi\lambda_j t) + B_j \sin(2\pi\lambda_j t)\right) \qquad \text{for } t = \ldots, -2, -1, 0, 1, 2, \ldots,$$

where the frequencies $0 < \lambda_1 < \cdots < \lambda_m < 1/2$ are fixed and $A_1, B_1, A_2, B_2, \ldots, A_m, B_m$ are *uncorrelated* random variables with common mean 0 and $\text{var}(A_j) = \sigma_j^2 = \text{var}(B_j)$. The covariance between X_t and X_{t+h} equals:

$$\sum_j \sigma_j^2 \left(\cos(2\pi\lambda_j t)\cos(2\pi\lambda_j(t+h)) + \sin(2\pi\lambda_j t)\sin(2\pi\lambda_j(t+h))\right) = \sum_j \sigma_j^2 \cos(2\pi\lambda_j h).$$

Because this covariance does not depend on t, the process $\{X_t\}$ is stationary with autocovariance function $\gamma_X(h) = \sum_{j=1}^{m} \sigma_j^2 \cos(2\pi\lambda_j h)$. This autocovariance function $\gamma_X(h)$ can be written as:

$$\gamma_X(h) = \sum_{j=1}^{m} \sigma_j^2 \left(\frac{e^{2\pi i\lambda_j h} + e^{-2\pi i\lambda_j h}}{2}\right)$$

Thus $\gamma_X(h)$ equals $\int_{-1/2}^{1/2} e^{2\pi i h\lambda} dF(\lambda)$ where F corresponds to the discrete distribution which takes values

$$-\lambda_m < \cdots < -\lambda_1 < \lambda_1 < \cdots < \lambda_m$$

with weights

$$\frac{\sigma_m^2}{2}, \ldots, \frac{\sigma_1^2}{2}, \frac{\sigma_1^2}{2}, \ldots, \frac{\sigma_m^2}{2}.$$

Note that this is a symmetric distribution. Thus the spectral distribution function puts mass only at the frequencies that are present in $\{X_t\}$. Moreover, the mass at a particular frequency λ_j is proportional to the variance σ_j^2 at that frequency. The total mass of the spectral distribution is:

$$\frac{\sigma_m^2}{2} + \cdots + \frac{\sigma_1^2}{2} + \frac{\sigma_1^2}{2} + \cdots + \frac{\sigma_m^2}{2} = \sigma_1^2 + \cdots + \sigma_m^2 = \gamma_X(0).$$

2.4 White Noise

For white noise $\gamma_X(h) = 0$ for $h \neq 0$ and $\gamma_X(0) = \sigma^2$. Thus $\sum_h |\gamma_X(h)| < \infty$ and the spectral density is given by:

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_X(h) \exp\left(-2\pi i\lambda h\right) = \gamma_X(0) = \sigma^2 \qquad \text{for all } -1/2 \leq \lambda \leq 1/2.$$

The idea is that all frequencies are present in white noise in equal amounts.

2.5 Spectral Density for ARMA processes

Theorem 2.1. *Let $\{Y_t\}$ be a mean-zero, stationary process with Spectral Distribution Function F_Y. Define $\{X_t\}$ by*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} \qquad \text{where } \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

Then $\{X_t\}$ is stationary with Spectral Distribution Function:

$$F_X(\lambda) = \int_{-1/2}^{\lambda} \left| \sum_j \psi_j e^{2\pi i j \lambda} \right|^2 dF_Y(\lambda). \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

Proof. The autocovariance of X_t is (note that X_t has mean zero because $\{Y_t\}$ was assumed to have zero mean):

$$\gamma_X(h) = \mathbb{E}X_t X_{t+h} = \mathbb{E}(\sum_j \psi_j Y_{t-j})(\sum_k \psi_k Y_{t+h-k}) = \sum_{j,k} \psi_j \psi_k \gamma_Y(h-k+j).$$

By the definition of the spectral distribution function, we can write:

$$\gamma_Y(h-k+j) = \int_{-1/2}^{1/2} e^{2\pi i (h-k+j)\lambda} dF_Y(\lambda).$$

Therefore,

$$\begin{aligned}
\gamma_X(h) &= \sum_{j,k} \psi_j \psi_k \int_{-1/2}^{1/2} e^{2\pi i (h-k+j)\lambda} dF_Y(\lambda) \\
&= \int_{-1/2}^{1/2} e^{2\pi i h \lambda} \sum_{j,k} \psi_j \psi_k e^{-2\pi i k \lambda} e^{2\pi i j \lambda} dF_Y(\lambda) \\
&= \int_{-1/2}^{1/2} e^{2\pi i h \lambda} \left(\sum_j \psi_j e^{2\pi i j \lambda} \right) \left(\sum_k \psi_k e^{-2\pi i k \lambda} \right) dF_Y(\lambda) \\
&= \int_{-1/2}^{1/2} e^{2\pi i h \lambda} \left| \sum_j \psi_j e^{2\pi i j \lambda} \right|^2 dF_Y(\lambda)
\end{aligned}$$

This is of the form:

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{2\pi i h \lambda} dF_X(\lambda)$$

with

$$dF_X(\lambda) = \left| \sum_j \psi_j e^{2\pi i j \lambda} \right|^2 dF_Y(\lambda).$$

The proof is complete. □

It follows from the above theorem that if $\{Y_t\}$ has a spectral density f_Y, then X_t also has a spectral density that is given by

$$f_X(\lambda) = \left| \sum_{j=-\infty}^{\infty} \psi_j e^{2\pi i j \lambda} \right|^2 f_Y(\lambda) \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

If we use the notation $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$, then the spectral density $f_X(\lambda)$ can be written as:

$$f_X(\lambda) = |\psi(e^{2\pi i\lambda})|^2 f_Y(\lambda)$$

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty Two

Aditya Guntuboyina

16 April 2013

1 Spectral Distribution Function

Let $\{X_t\}$ be a stationary sequence of random variables and let $\gamma_X(h) = \text{cov}(X_t, X_{t+h})$ denote the autocovariance function.

A theorem due to Herglotz (sometimes attributed to Bochner) states that **every** autocovariance function γ_X can be written as:

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{2\pi i h \lambda} dF(\lambda),$$

where $F(\cdot)$ is a non-negative, right-continuous, non-decreasing function on $[-1/2, 1/2]$ with $F(-1/2) = 0$ and $F(1/2) = \gamma_X(0)$. Moreover, F is uniquely determined by γ_X.

This function F is called the *Spectral Distribution Function* of $\{X_t\}$. If F has a density f i.e., if F can be written as

$$F(x) := \int_{-1/2}^{x} f(t)dt$$

then f is called the *Spectral Density* of $\{X_t\}$.

A sufficient condition (but not necessary) for the existence of the spectral density is the condition $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$. And in this case, the spectral density exists and is given by the formula:

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-2\pi i \lambda h) \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

The spectral distribution function is an as important quantity for a stationary process as the autocovariance function.

2 Linear Time-Invariant Filters

A linear time-invariant filter uses a set of specified coefficients $\{a_j\}$ for $j = \ldots, -2, -1, 0, 1, 2, 3, \ldots$ to transform an input time series $\{X_t\}$ into an output time series $\{Y_t\}$ according to the formula:

$$Y_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}.$$

The filter is determined by the coefficients $\{a_j\}$ which are often assumed to satisfy $\sum_{j=-\infty}^{\infty} |a_j| < \infty$.

Suppose that the input series $\{X_t\}$ is given by

$$X_t = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$$

Such an $\{X_t\}$ is often called an *impulse function*. The output of the filter $\{Y_t\}$ can then be easily seen to be $Y_t = a_t$. For this reason the filter coefficients $\{a_j\}$ are often collectively known as the *impulse response function*.

The two main examples of linear time-invariant filters that we have seen so far are (1) the moving average filter which has the impulse response function: $a_j = 1/(2q+1)$ for $|j| \leq q$ and $a_j = 0$ otherwise; and (2) Differencing which corresponds to the filter $a_0 = 1$ and $a_1 = -1$ and all other a_js equal zero. We have seen that these two filters act very differently; one estimates trend while the other eliminates it.

Suppose that the input time series $\{X_t\}$ is stationary with autocovariance function γ_X. What is the autocovariance function of $\{Y_t\}$? Observe that

$$\gamma_Y(h) := \text{cov}\left(\sum_j a_j X_{t-j}, \sum_k a_k X_{t+h-k}\right) = \sum_{j,k} a_j a_k \text{cov}(X_{t-j}, X_{t+h-k}) = \sum_{j,k} a_j a_k \gamma_X(h-k+j). \quad (1)$$

Note that the above calculation shows also that $\{Y_t\}$ is stationary.

Suppose now that the spectral density of the input stationary series $\{X_t\}$ is f_X. What then is the spectral density f_Y of the output $\{Y_t\}$?

Because the spectral density of $\{X_t\}$ equals f_X, we have

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{2\pi i h\lambda} f_X(\lambda) d\lambda.$$

We thus have from (1) that

$$\gamma_Y(h) = \sum_j \sum_k a_j a_k \int e^{2\pi i(h-k+j)\lambda} f_X(\lambda) d\lambda = \int e^{2\pi i h\lambda} f_X(\lambda) \left(\sum_j \sum_k a_j a_k e^{-2\pi i k\lambda} e^{2\pi i j\lambda}\right) d\lambda \quad (2)$$

Let us now define the function

$$A(\lambda) := \sum_j a_j e^{-2\pi i j\lambda} \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

Note that this function only depends on the filter coefficients $\{a_j\}$. From (2) it clearly follows that

$$\gamma_Y(h) = \int e^{2\pi i\lambda h} f_X(\lambda) A(\lambda) \overline{A(\lambda)} d\lambda,$$

where, of course, $\overline{A(\lambda)}$ denotes the complex conjugate of $A(\lambda)$. As a result, we have

$$\gamma_Y(h) = \int e^{2\pi i\lambda h} f_X(\lambda) \left|A(\lambda)\right|^2 d\lambda.$$

This is clearly of the form $\gamma_Y(h) = \int e^{2\pi i\lambda h} f_Y(\lambda) d\lambda$. We therefore have

$$f_Y(\lambda) = f_X(\lambda) \left|A(\lambda)\right|^2 \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \quad (3)$$

In other words, the action of the filter on the spectrum of the input is very easy to explain. It modifies the spectrum by multiplying it with the function $|A(\lambda)|^2$. Depending on the value of $|A(\lambda)|^2$, some frequencies may be enhanced in the output while other frequencies will be diminished.

This function $\lambda \mapsto |A(\lambda)|^2$ is called the *power transfer function* of the filter. The function $\lambda \mapsto A(\lambda)$ is called the *transfer function* or the *frequency response function* of the filter.

The spectral density is very useful while studying the properties of a filter. While the autocovariance function of the output series γ_Y depends in a complicated way on that of the input series γ_X, the dependence between the two spectral densities is very simple.

Example 2.1 (Power Transfer Function of the Differencing Filter)**.** *Consider the Lag s differencing filter: $Y_t = X_t - X_{t-s}$ which corresponds to the weights $a_0 = 1$ and $a_s = -1$ and $a_j = 0$ for all other j. Then the transfer function is clearly given by*

$$A(\lambda) = \sum_j a_j e^{-2\pi i j\lambda} = 1 - e^{-2\pi i s\lambda} = 2i\sin(\pi s\lambda)e^{-\pi i s\lambda},$$

where, for the last equality, the formula $1 - e^{i\theta} = -2i\sin(\theta/2)e^{i\theta/2}$ is used. Therefore the power transfer function equals

$$|A(\lambda)|^2 = 4\sin^2(\pi s\lambda) \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

To understand this function, we only need to consider the interval $[0, 1/2]$ because it is symmetric on $[-1/2, 1/2]$.

When $s = 1$, the function $\lambda \mapsto |A(\lambda)|^2$ is increasing on $[0, 1/2]$. This means that first order differencing enhances the higher frequencies in the data and diminishes the lower frequencies. Therefore, it will make the data more wiggly.

For higher values of s, the function $A(\lambda)$ goes up and down and takes the value zero for $\lambda = 0, 1/s, 2/s, \ldots$. In other words, it eliminates all components of period s.

Example 2.2. *Now consider the moving average filter which corresponds to the coefficients $a_j = 1/(2q+1)$ for $|j| \le q$. The transfer function is*

$$\frac{1}{2q+1}\sum_{j=-q}^{q} e^{-2\pi i j\lambda} = \frac{S_{q+1}(\lambda) + S_{q+1}(-\lambda) - 1}{2q+1},$$

where it may be recalled (Lecture 19) that

$$S_n(g) := \sum_{t=0}^{n-1} \exp(2\pi i g t) = \frac{\sin(\pi n g)}{\sin(\pi g)} e^{i\pi g(n-1)}$$

. Thus

$$S_n(g) + S_n(-g) = 2\frac{\sin(\pi n g)}{\sin(\pi g)}\cos(\pi g(n-1)),$$

which implies that the transfer function is given by

$$A(\lambda) = \frac{1}{2q+1}\left(2\frac{\sin(\pi(q+1)\lambda)}{\sin(\pi\lambda)}\cos(\pi q\lambda) - 1\right),$$

This function only depends on q and can be plotted for various values of q. For q large, it drops to zero very quickly. The interpretation is that the filter kills the high frequency components in the input process.

3 Spectral Densities of ARMA Processes

Suppose $\{X_t\}$ is a stationary ARMA process: $\phi(B)X_t = \theta(B)Z_t$ where the polynomials ϕ and θ have no common zeroes on the unit circle. Because of stationarity, the polynomial ϕ has no roots on the unit circle.

Let $U_t = \phi(B)X_t = \theta(B)Z_t$. Let us first write down the spectral density of $U_t = \phi(B)X_t$ in terms of that of $\{X_t\}$. Clearly, U_t can be viewed as the output of a filter applied to X_t. The filter is given by $a_0 = 1$ and $a_j = -\phi_j$ for $1 \le j \le p$ and $a_j = 0$ for all other j. Let $A_\phi(\lambda)$ denote the transfer function of this filter. Then we have

$$f_U(\lambda) = |A_\phi(\lambda)|^2 f_X(\lambda). \tag{4}$$

Similarly, using the fact that $U_t = \theta(B)Z_t$, we can write

$$f_U(\lambda) = |A_\theta(\lambda)|^2 f_Z(\lambda) = \sigma_Z^2 |A_\theta(\lambda)|^2 \tag{5}$$

where $A_\theta(\lambda)$ is the tranfer function of the filter with coefficients $a_0 = 1$ and $a_j = \theta_j$ for $1 \le j \le q$ and $a_j = 0$ for all other j. Equating (4) and (5), we obtain

$$f_X(\lambda) = \frac{|A_\theta(\lambda)|^2}{|A_\phi(\lambda)|^2}\sigma_Z^2 \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

Now

$$A_\phi(\lambda) = 1 - \phi_1 e^{-2\pi i\lambda} - \phi_2 e^{-2\pi i(2\lambda)} - \cdots - \phi_p e^{-2\pi i(p\lambda)} = \phi(e^{-2\pi i\lambda}).$$

Similarly $A_\theta(\lambda) = \theta(e^{-2\pi i\lambda})$. As a result, we have

$$f_X(\lambda) = \sigma_Z^2 \frac{|\theta(e^{-2\pi i\lambda})|^2}{|\phi(e^{-2\pi i\lambda})|^2} \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

Note that the denominator on the right hand side above is non-zero for all λ because of stationarity.

Example 3.1 (MA(1))**.** *For the MA(1) process: $X_t = Z_t + \theta Z_{t-1}$, we have $\phi(z) = 1$ and $\theta(z) = 1 + \theta z$. Therefore*

$$\begin{aligned} f_X(\lambda) &= \sigma_Z^2 \left|1 + \theta e^{2\pi i\lambda}\right|^2 \\ &= \sigma_Z^2 \left|1 + \theta\cos 2\pi\lambda + i\theta \sin 2\pi\lambda\right|^2 \\ &= \sigma_Z^2 \left[(1 + \theta\cos 2\pi\lambda)^2 + \theta^2 \sin^2 2\pi\lambda\right] \\ &= \sigma_Z^2 \left[1 + \theta^2 + 2\theta\cos 2\pi\lambda\right] \qquad \textit{for } -1/2 \le \lambda \le 1/2. \end{aligned}$$

Check that for $\theta = -1$, the quantity $1 + \theta^2 + 2\theta\cos(2\pi\lambda)$ equals the power transfer function of the first differencing filter.

Example 3.2 (AR(1))**.** *For AR(1): $X_t - \phi X_{t-1} = Z_t$, we have $\phi(z) = 1 - \phi z$ and $\theta(z) = 1$. Thus*

$$f_X(\lambda) = \sigma_Z^2 \frac{1}{|1 - \phi e^{2\pi i\lambda}|^2} = \frac{\sigma_Z^2}{1 + \phi^2 - 2\phi\cos 2\pi\lambda} \qquad \textit{for } -1/2 \le \lambda \le 1/2.$$

Example 3.3 (AR(2))**.** *For the AR(2) model: $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t$, we have $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ and $\theta(z) = 1$. Here it can be shown that*

$$f_X(\lambda) = \frac{\sigma_Z^2}{1 + \phi_1^2 + \phi_2^2 - 2\phi_1(1 - \phi_2)\cos 2\pi\lambda - 2\phi_2 \cos 4\pi\lambda} \qquad \textit{for } -1/2 \le \lambda \le 1/2.$$

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty Three

Aditya Guntuboyina

18 April 2013

1 Nonparametric Estimation of the Spectral Density

Let $\{X_t\}$ be a stationary process with $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$. We have then seen that $\{X_t\}$ has a spectral density that is given by

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_X(h) e^{-2\pi i \lambda h} \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \tag{1}$$

Suppose now that we are given data $x_1, \ldots, x_n$ from the process $\{X_t\}$. How then would we estimate $f(\lambda)$ without making any parametric assumptions about the underlying process? This is our next topic.

Why would we want to estimate the spectral density nonparametrically?

When we were fitting ARMA models to the data, we first looked at the sample autocovariance or autocorrelation function and we then tried to find the ARMA model whose theoretical acf matched with the sample acf. Now the sample autocovariance function is a nonparametric estimate of the theoretical autocovariance function of the process. In other words, we first estimated $\gamma(h)$ nonparametrically by $\hat{\gamma}(h)$ and then found an ARMA model whose $\gamma_{ARMA}(h)$ is close to $\hat{\gamma}(h)$.

If we can estimate the spectral density nonparametrically, we can similarly use the estimate for choosing a parametric model. We simply choose the ARMA model whose spectral density is closest to the non-parametric estimate.

Another reason for estimating the spectral density comes from the problem of estimating filter coefficients. Suppose that we know that two processes $\{X_t\}$ and $\{Y_t\}$ are related to each other through a linear time-invariant filter. In other words, $\{Y_t\}$ is the output when $\{X_t\}$ is the input to a filter. Suppose, that we do not know the filter coefficients however but we are given observations from both the input and the output process. The goal is to estimate the filter. In this case, a natural strategy is to estimate the spectral densities of f_X and f_Y from data and then to use $f_Y(\lambda) = f_X(\lambda)|A(\lambda)|^2$ to obtain an estimate of the power transfer function of the filter (to obtain an estimate of the transfer function itself, one needs to use cross-spectra). This is one of the applications of spectral analysis. We might not always be able to make parametric assumptions about $\{X_t\}$ and $\{Y_t\}$ so it makes sense to estimate the spectral densities nonparametrically.

Nonparametric estimation of the spectral density is more complicated than the nonparamtric estimation of the autocovariance function. The main reason is that the natural estimator does not work well.

Because of the formula (1) for the spectral density in terms of the autocovariance function $\gamma_X(h)$, a natural idea to estimate $f(\lambda)$ is to replace $\gamma_X(h)$ by its estimator $\hat{\gamma}(h)$ for $|h| < n$ (it is not possible to

estimate $\gamma(h)$ for $|h| > n$). This would result in the estimator:

$$I(\lambda) = \sum_{h:|h|<n} \hat{\gamma}(h) e^{-2\pi i \lambda h} \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

When $\lambda = j/n \in (0, 1/2]$, the above quantity is just the periodogram:

$$I(j/n) = \frac{|b_j|^2}{n} \qquad \text{where } b_j = \sum_t x_t \exp\left(-\frac{2\pi i j t}{n}\right)$$

Unfortunately, $I(\lambda)$ is not a good estimator of f_X. This can be easily seen by simulations. Just generate data from white noise and observe that the periodogram is very wiggly while the true spectral density is constant. The fact that $I(\lambda)$ is a bad estimator can also be verified mathematically in the following way.

Suppose that the data x_t are generated from gaussian white noise with variance σ^2 (their mean is zero because they are white noise). What is the distribution of $|b_j|^2/n$ for $j/n \in [0, 1/2]$? Write

$$\begin{aligned}
\frac{|b_j|^2}{n} &= \frac{1}{n}\left|\sum_{t=0}^{n-1} x_t \exp\left(-\frac{2\pi i j t}{n}\right)\right|^2 \\
&= \frac{1}{n}\left|\sum_t x_t \cos(2\pi j t/n) - i \sum x_t \sin(2\pi j t/n)\right|^2 \\
&= \frac{1}{n}\left(A_j^2 + B_j^2\right),
\end{aligned}$$

where

$$A_j = \sum_t x_t \cos(2\pi j t/n) \text{ and } B_j = \sum_t x_t \sin(2\pi j t/n).$$

If we also assume normality of $x_1, \ldots, x_n$, then (A_j, B_j) are jointly normal with

$$\mathrm{var} A_j = \sigma^2 \sum_{t=0}^{n-1} \cos^2(2\pi j t/n) \text{ and } \mathrm{var} B_j = \sigma^2 \sum_{t=0}^{n-1} \sin^2(2\pi j t/n).$$

Also

$$\mathrm{cov}(A_j, B_j) = \sigma^2 \sum_{t=0}^{n-1} \cos(2\pi j t/n) \sin(2\pi j t/n).$$

It can be checked that

$$\begin{aligned}
\sum_{t=0}^{n-1} \cos^2(2\pi j t/n) &= n \qquad \text{when } j \text{ is either 0 or } n/2 \\
&= n/2 \qquad \text{when } j \text{ is neither 0 nor } n/2.
\end{aligned}$$

and

$$\begin{aligned}
\sum_{t=0}^{n-1} \sin^2(2\pi j t/n) &= 0 \qquad \text{when } j \text{ is either 0 or } n/2 \\
&= n/2 \qquad \text{when } j \text{ is neither 0 nor } n/2.
\end{aligned}$$

and

$$\sum_{t=0}^{n-1} \cos(2\pi j t/n) \sin(2\pi j t/n) = 0.$$

Thus when j is neither 0 nor $n/2$ (recall that $0 \le j/n \le 1/2$), we have

$$\frac{\sqrt{2} A_j}{\sigma\sqrt{n}} \sim N(0,1) \quad \text{and} \quad \frac{\sqrt{2} B_j}{\sigma\sqrt{n}} \sim N(0,1)$$

which implies that

$$\frac{2}{n\sigma^2}A_j^2 \sim \chi_1^2 \quad \text{and} \quad \frac{2}{n\sigma^2}B_j^2 \sim \chi_1^2.$$

Also because they are independent, we have for $j/n \in (0, 1/2)$

$$\frac{2}{\sigma^2}I(j/n) = \frac{2|b_j|^2}{n\sigma^2} = \frac{2}{n\sigma^2}A_j^2 + \frac{2}{n\sigma^2}B_j^2 \sim \chi_2^2$$

or $I(j/n) \sim (\sigma^2/2)\chi_2^2$.

For $j = 0$ or $n/2$, we have $B_j = 0$ and $A_j \sim N(0, \sigma^2 n)$ which implies that $|b_j|^2/n \sim \sigma^2\chi_1^2$.

It is important to notice that the distribution of $I(j/n)$ does not depend on n. One can also check that (A_j, B_j) is independent of $(A_{j'}, B_{j'})$ for $j \neq j'$.

Therefore, when the data $x_1, \ldots, x_n$ are generated from the Gaussian White Noise model, the periodogram ordinates $I(j/n)$ for $0 < j \leq n/2$ are independent random variables having the distribution $(\sigma^2/2)\chi_2^2$ for $0 < j < n/2$ and $\sigma^2\chi_1^2$ for $j = n/2$. Because of this independence and the fact that the distribution does not depend on n, it should be clear that $I(\lambda)$ is not a good estimate of $f(\lambda)$.

We have done the above calculations for data from the gaussian white noise. For general ARMA processes, under some regularity conditions, it can be shown that when n is large, the random variables:

$$\frac{2I(j/n)}{f(j/n)}, \qquad \text{for } 0 < j < n/2$$

are approximately independently distributed according to the χ_2^2 distribution.

Note that because the χ_2^2 distribution has mean 2, the expected value of $I(j/n)$ is approximately $f(j/n)$. In other words, the periodogram is approximately unbiased. On the other hand, the variance of $I(j/n)$ is approximately $f^2(j/n)$. So, in the gaussian white noise case, for example, the variance of the periodogram ordinates is σ^4 which does not decrease with n. This and the approximate independence of the neighboring periodogram ordinates makes the periodogram very noisy and a bad estimator of the true spectral density.

2 Modifying the Periodogram for good estimates of the spectral density

2.1 Method One

The approximate distribution result allows us to write:

$$\frac{2I(j/n)}{f(j/n)} \approx 2 + 2U_j \qquad \text{for } 0 < j < n/2,$$

where $U_1, U_2, \ldots$ are independent, have mean zero and variance 1. In other words $\{U_j\}$ is white noise. Thus

$$I(j/n) = f(j/n) + U_j f(j/n) \qquad \text{for } 0 < j < n/2.$$

Therefore, we can think of $I(j/n)$ as an uncorrelated time series with a trend $f(j/n)$ that we wish to estimate. Our previous experience with trend estimation suggests that we do this by smoothing $I(j/n)$ with a filter, say the simple moving average filter:

$$\frac{1}{2m+1}\sum_{k=-m}^{m} I\left(\frac{j+k}{n}\right).$$

More generally, we can consider using unequal weights as well to yield estimators of the form:

$$\hat{f}(j/n) = \sum_{k=-m_n}^{m_n} W_n(k) I\left(\frac{j+k}{n}\right).$$

Note that if we take $m_n = 0$, we get back the periodogram. We can extend this definition of $\hat{f}$ to the entire interval $[0, 1/2]$ in the following way: For each $\lambda \in [0, 1/2]$, let $g(\lambda, n)$ denote the multiple of $1/n$ that is closest to λ. Define

$$\hat{f}(\lambda) = \hat{f}(g(\lambda, n)).$$

It can be shown that this estimator is consistent (i.e., it gets closer and closer to $f(j/n)$ as n becomes larger) provided:

$$m_n \to \infty \quad \text{and} \quad \frac{m_n}{n} \to 0 \tag{2}$$

as $n \to \infty$. One also needs the weights $W_n(k)$ to be symmetric: $W_n(k) = W_n(-k)$, nonnegative $W_n(k) \geq 0$, add up to 1: $\sum_{k=-m_n}^{m_n} W_n(k) = 1$ and their sum of squares to go to zero: $\sum_{k=-m_n}^{m_n} W_n^2(k) \to 0$ as $n \to \infty$. Note that all these conditions are satisfied for the simple moving average filter with m chosen as in (2).

If the above conditions are met, then, for $0 < \lambda < 1/2$, we have

$$\mathbb{E}\hat{f}(\lambda) \approx f(\lambda) \quad \text{and} \quad \text{var}(\hat{f}(\lambda)) \approx \left(\sum_{k=-m_n}^{m_n} W_n^2(k)\right) f^2(\lambda).$$

When λ equals 0 or $1/2$, the variance is twice the one given by the equation above. The expectation is still the same.

Also the covariance between $\hat{f}(\lambda_1)$ and $\hat{f}(\lambda_2)$ is approximately zero.

2.2 Method Two

Here is a slightly different way of coming up with estimators for the spectral density that are different from the periodogram. The periodogram is defined by:

$$I(\lambda) = \sum_{h:|h|<n} \hat{\gamma}(h) \exp(-2\pi i \lambda h) \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \tag{3}$$

Note that the above formula involves the estimates of all the autocovariances $\gamma_X(h)$ for $|h| < n$. Now we know that from a sample of size n, it is impossible to come up with good estimates of $\gamma_X(h)$ for h close to n. This is often cited as a reason why the periodogram is not a good estimator. In light of this reason, a reasonable way to obtain better estimators is to truncate the sum on the right hand side of (3) by omitting $\hat{\gamma}(h)$ for h near n. In other words, we consider

$$\tilde{f}(\lambda) = \sum_{h:|h|\leq r} \hat{\gamma}(h) \exp(-2\pi i \lambda h).$$

If we assume that $r = r_n$ is a function of n such that $r_n \to \infty$ and $r_n/n \to 0$ as $n \to \infty$, then $\tilde{f}$ is a sum of $(2r+1)$ terms, each with a variance of about $1/n$. In this case, under regularity conditions, it can be shown that $\tilde{f}$ is a consistent estimator of f.

More generally, we can take

$$\tilde{f}(\lambda) = \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) \hat{\gamma}(h) \exp(-2\pi i \lambda h),$$

where $w(x)$ is a symmetric $w(x) = w(-x)$ function satisfying $w(0) = 1$, $|w(x)| \leq 1$ and $w(x) = 0$ for $|x| > 1$. This is sometimes called a *lag window spectral density estimator.*

2.3 Equivalence of the Two Methods

We shall now show that these two ways of improving the periodogram: by smoothing it and the lag window spectral density estimator are essentially the same. To see this, we first need an *inverse* relationship between $I(\lambda)$ and $\hat{\gamma}(h)$. We have defined $I(\lambda)$ as

$$I(\lambda) := \sum_{h:|h|<n} \hat{\gamma}(h) e^{-2\pi i \lambda h} \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

It is possible to invert this formula to write $\hat{\gamma}(k)$ in terms of $I(\lambda)$. Fix an integer k with $|k| < n$ and multiply both sides of the above formula by $e^{2\pi i \lambda k}$. Integrating the resulting expression with respect to λ from $-1/2$ to $1/2$, we get

$$\int_{-1/2}^{1/2} e^{2\pi i \lambda k} I(\lambda) d\lambda = \sum_{h:|h|<n} \hat{\gamma}(h) \int_{-1/2}^{1/2} e^{2\pi i \lambda (k-h)} d\lambda = \hat{\gamma}(k).$$

This therefore implies

$$\hat{\gamma}(k) = \int_{-1/2}^{1/2} e^{2\pi i \lambda k} I(\lambda) d\lambda. \tag{4}$$

In other words, the function $I(\lambda)$ is the spectral density corresponding to the sample autocorrelation function. Using the formula (4), we can write the lag window spectral density estimator as

$$\begin{aligned}
\tilde{f}(\lambda) &= \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) \hat{\gamma}(h) e^{-2\pi i \lambda h} \\
&= \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) \int_{-1/2}^{1/2} e^{2\pi i \rho h} I(\rho) d\rho \; e^{-2\pi i \lambda h} \\
&= \int_{-1/2}^{1/2} I(\rho) \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) e^{2\pi i (\rho - \lambda) h} d\rho.
\end{aligned}$$

By the change of variable $\rho = \lambda + u$, we get

$$\tilde{f}(\lambda) = \int I(\lambda + u) \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) e^{2\pi i u h} du.$$

Letting

$$W(u) = \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) e^{2\pi i u h},$$

we get that

$$\tilde{f}(\lambda) = \int I(\lambda + u) W(u) du.$$

Thus the lag window spectral density estimator $\tilde{f}$ can also be thought of as obtained by smoothing the periodogram.

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty Four

Aditya Guntuboyina

23 April 2013

In the last class, we discussed the problem of nonparametrically estimating a spectral density. The natural estimator is:

$$I(\lambda) := \sum_{h:|h|<n} \hat{\gamma}(h) e^{-2\pi i \lambda h} \qquad \text{for } -1/2 \leq \lambda \leq 1/2$$

which for $\lambda = j/n \in (0, 1/2]$ coincides with the periodogram:

$$I(j/n) = \frac{|b_j|^2}{n} \qquad \text{where } b_j = \sum_t x_t \exp\left(-\frac{2\pi i j t}{n}\right)$$

The key result about the periodogram is that under some regularity conditions which hold for all ARMA processes under the gaussian noise, it can be shown that when n is large, the random variables:

$$\frac{2I(j/n)}{f(j/n)} \qquad \text{for } 0 < j < n/2$$

are approximately independently distributed according to the χ^2_2 distribution. As a result, $I(\lambda)$ is not a good estimator of $f(\lambda)$.

We studied two modifications of the periodogram:

1. **Moving average smoothing**: Choose an integer $m \geq 1$ and estimate $f(j/n)$ by

$$\hat{f}(j/n) := \frac{1}{2m+1} \sum_{k=-m}^{m} I\left(\frac{j+k}{n}\right)$$

   or more generally

$$\hat{f}(j/n) := \sum_{k=-m}^{m} W(k) I\left(\frac{j+k}{n}\right)$$

   where $W(k)$ are nonnegative weights summing to one. This estimator is based on the approximate representation $I(j/n) \approx f(j/n) + U_j f(j/n)$ for $0 < j < n/2$ where $\{U_j\}$ is white noise.

2. **Lag Window Spectral Density Estimator**: Choose an integer $r \geq 1$ and estimate $f(j/n)$ by

$$\hat{f}(j/n) := \sum_{h:|h|\leq r} \hat{\gamma}(h) \exp\left(-2\pi i \lambda h\right)$$

   or more generally

$$\hat{f}(j/n) := \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) \hat{\gamma}(h) \exp\left(-2\pi i \lambda h\right)$$

where $w(x)$ is a symmetric i.e., $w(x) = w(-x)$ function satisfying $w(0) = 1$, $|w(x)| \leq 1$ and $w(x) = 0$ for $|x| > 1$. This estimator is based on the formula:

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \lambda h}$$

and the idea that $\gamma(h)$ for large h become close to zero (because $\sum_h |\gamma(h)| < \infty$) and they are also difficult to estimate from the data.

1 Equivalence of these Two Estimators

We shall now show that these two ways of improving the periodogram: by smoothing it and the lag window spectral density estimator are essentially the same. To see this, we first need an *inverse* relationship between $I(\lambda)$ and $\hat{\gamma}(h)$. We have defined $I(\lambda)$ as

$$I(\lambda) := \sum_{h:|h|<n} \hat{\gamma}(h) e^{-2\pi i \lambda h} \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

It is possible to invert this formula to write $\hat{\gamma}(k)$ in terms of $I(\lambda)$. Fix an integer k with $|k| < n$ and multiply both sides of the above formula by $e^{2\pi i \lambda k}$. Integrating the resulting expression with respect to λ from $-1/2$ to $1/2$, we get

$$\int_{-1/2}^{1/2} e^{2\pi i \lambda k} I(\lambda) d\lambda = \sum_{h:|h|<n} \hat{\gamma}(h) \int_{-1/2}^{1/2} e^{2\pi i \lambda (k-h)} d\lambda = \hat{\gamma}(k).$$

This therefore implies

$$\hat{\gamma}(k) = \int_{-1/2}^{1/2} e^{2\pi i \lambda k} I(\lambda) d\lambda. \tag{1}$$

In other words, the function $I(\lambda)$ is the spectral density corresponding to the sample autocorrelation function. Using the formula (1), we can write the lag window spectral density estimator as

$$\begin{aligned}
\tilde{f}(\lambda) &= \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) \hat{\gamma}(h) e^{-2\pi i \lambda h} \\
&= \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) \int_{-1/2}^{1/2} e^{2\pi i \rho h} I(\rho) d\rho \; e^{-2\pi i \lambda h} \\
&= \int_{-1/2}^{1/2} I(\rho) \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) e^{2\pi i (\rho - \lambda) h} d\rho.
\end{aligned}$$

By the change of variable $\rho = \lambda + u$, we get

$$\tilde{f}(\lambda) = \int I(\lambda + u) \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) e^{2\pi i u h} du.$$

Letting

$$W(u) = \sum_{h:|h|\leq r} w\left(\frac{h}{r}\right) e^{2\pi i u h},$$

we get that

$$\tilde{f}(\lambda) = \int I(\lambda + u) W(u) du.$$

Thus the lag window spectral density estimator $\tilde{f}$ can also be thought of as obtained by smoothing the periodogram.

2 Approximate Confidence Intervals for $f(j/n)$

Recall that the random variables

$$\frac{2I(j/n)}{f(j/n)} \qquad \text{for } 0 < j < n/2$$

are approximately independently distributed according to the χ^2_2 distribution.

Therefore, approximately

$$\hat{f}(j/n) = \frac{1}{2m+1} \sum_{k=-m}^{m} I\left(\frac{j+k}{n}\right) \approx \frac{f(j/n)}{2(2m+1)} \sum_{k=-m}^{m} \frac{2I((j+k)/n)}{f((j+k)/n)}.$$

This would allow us to approximate the distribution of $\hat{f}(j/n)$ in the following way:

$$2(2m+1)\frac{\hat{f}(j/n)}{f(j/n)} \sim \chi^2_{2(2m+1)}.$$

If $\chi^2_{2(2m+1)}(\alpha/2)$ and $\chi^2_{2(2m+1)}(1-\alpha/2)$ satisfy

$$\mathbb{P}\left\{\chi^2_{2(2m+1)}(\alpha/2) \le \chi^2_{2(2m+1)} \le \chi^2_{2(2m+1)}(1-\alpha/2)\right\} = 1-\alpha,$$

then we conclude that approximately

$$\mathbb{P}\left\{\chi^2_{2(2m+1)}(\alpha/2) \le 2(2m+1)\frac{\hat{f}(j/n)}{f(j/n)} \le \chi^2_{2(2m+1)}(1-\alpha/2)\right\} \approx 1-\alpha.$$

This would lead to the following confidence interval for $f(j/n)$ of level approximately $1-\alpha$:

$$2(2m+1)\frac{\hat{f}(j/n)}{\chi^2_{2(2m+1)}(1-\alpha/2)} \le f(j/n) \le 2(2m+1)\frac{\hat{f}(j/n)}{\chi^2_{2(2m+1)}(\alpha/2)}.$$

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty Five

Aditya Guntuboyina

25 April 2013

Suppose $x_1, \ldots, x_n$ are data that we assume come from a stationary process $\{X_t\}$ with mean zero and variance σ^2.

Previously, we studied the stationary ARMA models for modelliing such data. Today we ask if there are any other natural ways of modelling such data. In particular, we investigate if it makes sense to use sines and cosines to model $x_1, \ldots, x_n$.

The simplest stationary model using sines and cosines is

$$X_t = A\cos(2\pi\lambda t) + B\sin(2\pi\lambda t)$$

where $0 \leq \lambda \leq 1/2$ is a fixed constant and A and B are uncorrelated random variables with mean 0 and variance σ^2. We have seen many times in the past that $\{X_t\}$ is stationary with mean 0 and variance σ^2.

More complicated stationary models with sines and cosines can be constructed by taking linear combinations of the form

$$X_t = \sum_{j=1}^{m} \left(A_j \cos(2\pi\lambda_j t) + B_j \sin(2\pi\lambda_j t)\right) \tag{1}$$

where $0 \leq \lambda \leq 1/2$ is a fixed constant and $A_1, B_1, A_2, B_2, \ldots, A_m, B_m$ are all uncorrelated random variables with mean zero and

$$\mathrm{var}(A_j) = \mathrm{var}(B_j) = \sigma_j^2.$$

Let $\sum_{j=1}^{m} \sigma_j^2 = \sigma^2$ so that the variance of the process $\{X_t\}$ equals σ^2.

It turns out that the model (1) can approximate any stationary model provided m is large enough and $\lambda_1, \ldots, \lambda_m$ and $\sigma_1^2, \ldots, \sigma_m^2$ are chosen appropriately. For example, the choices

$$\lambda_j = \frac{j}{2m} \text{ and } \sigma_j^2 = \frac{\sigma^2}{m} \qquad \text{for } j = 1, \ldots, m$$

for m large lead to a very good approximation of the white noise model.

To approximate a general stationary process using sines and cosines, we can use its spectral density. We know that the spectral density $f(\lambda)$ satisfies

$$\int_{-1/2}^{1/2} f(\lambda) d\lambda = \sigma^2.$$

For real valued stationary processes, the spectral density is symmetric around zero. Therefore,

$$\int_{0}^{1/2} f(\lambda) d\lambda = \frac{\sigma^2}{2}.$$

Approximating the integral on the left hand side above by a Riemann sum (assuming such an approximation is valid), we obtain

$$\frac{\sigma^2}{2} = \sum_{j=1}^{m} \int_{\lambda_{j-1}}^{\lambda_j} f(\lambda) d\lambda \approx \frac{1}{2m} \sum_{j=1}^{m} f(\lambda_j) \qquad \text{where } \lambda_j = \frac{j}{2m}.$$

If we now take

$$\lambda_j = \frac{j}{2m} \text{ and } \sigma_j^2 = \frac{f(\lambda_j)}{m}$$

Then, for large m, the process (1) will approximate the stationary process with spectral density f. This method can be used to, for example, simulate ARMA processes without using the *arima.sim* function in R.

When m equals ∞, the process (1) can be defined to make sense so that it has exactly the same spectral density f. But this requires stochastic integration and is beyond the scope of this class.

The above analysis conveys the key role of the spectral density for the study of stationary processes. It essentially tells us all that there is to know about the stationary process.

Spring 2013 Statistics 153 (Time Series) : Lecture Twenty Six

Aditya Guntuboyina

30 April 2013

We will continue discussion on the estimation of the spectral density. Estimators are given by:

$$\hat{f}(j/n) := \sum_{k=-m}^{m} W_m(k) I\left(\frac{j+k}{n}\right)$$

The set of weights $\{W_m(k)\}$ is often referred to as a kernel or a spectral window.

Simplest choice of $W_m(k)$ is

$$W_m(k) = \frac{1}{2m+1} \qquad \text{for } -m \leq k \leq m.$$

This window is called the *Daniell Spectral Window*.

One can get these estimates directly in R by using the function *spec.pgram* and *kernel*.

The bandwidth of a spectral window is defined as the standard deviation of the weighting distribution. It is actually this standard deviation that controls the bias of the estimator. This can be justified by a second order Taylor expansion as follows. The expected value of $\hat{f}(j/n)$ is

$$\mathbb{E}\hat{f}(j/n) = \sum_{k=-m}^{m} W_m(k) f\left(\frac{j+k}{n}\right)$$

Let $\lambda = j/n$ for ease of notation. Then by a second order Taylor expansion around λ, we get

$$\mathbb{E}\hat{f}(\lambda) = \sum_{k=-m}^{m} W_m(k)\left(f(\lambda) + \frac{k}{n} f'(\lambda) + \frac{k^2}{2n^2} f''(\lambda)\right)$$

If the weights are such that $\sum_k W_m(k) = 1$ and $\sum_k k W_m(k) = 0$ (satisfied for the Daniell kernel for example), then

$$\mathbb{E}\hat{f}(\lambda) - f(\lambda) = \frac{f''(\lambda)}{2} \sum_{k=-m}^{m} \left(\frac{k}{n}\right)^2 W_m(k)$$

The bandwidth of the kernel is given by

$$\sqrt{\sum_{k=-m}^{m} \left(\frac{k}{n}\right)^2 W_m(k)}.$$

For the Daniell kernel, the bandwidth is given by the standard deviation of the uniform distribution on $\{-m/n, -(m-1)/n, \ldots, (m-1)/n, m/n\}$ which is very close to the standard deviation of the continuous uniform distribution on $[-m/n, m/n]$ which equals:

$$\sqrt{\frac{(2m)^2}{12n^2}} \approx \sqrt{\frac{L^2}{12n^2}} = \frac{L}{n\sqrt{12}}$$

Repeated use of the Daniell kernel yields non-uniform weights. For example, the Daniell kernel for $m = 1$ corresponds to the three weights $(1/3, 1/3, 1/3)$. Applying it to a sequence of numbers $\{u_t\}$ leads to the smoother:

$$\hat{u}_t = \frac{u_{t-1} + u_t + u_{t+1}}{3}.$$

Applying the Daniell kernel again to $\hat{u}_t$ gives

$$\hat{\hat{u}}_t := \frac{\hat{u}_{t-1} + \hat{u}_t + \hat{u}_{t+1}}{3} = \frac{1}{9}u_{t-2} + \frac{2}{9}u_{t-1} + \frac{3}{9}u_t + \frac{2}{9}u_{t+1} + \frac{1}{9}u_{t+2}.$$

Thus application of the Daniell kernel is equivalent to applying the kernel $(1/9, 2/9, 3/9, 2/9, 1/9)$ to the data. This is a non-uniform kernel with a higher bandwidth. Note also that these weights equal the convolution of the Daniell kernel. In other words, if X_1 and X_2 both have the pmfs $(1/3, 1/3, 1/3)$, then $X_1 + X_2$ has the pmf $(1/9, 2/9, 3/9, 2/9, 1/9)$. If we keep applying the Daniell kernel repeatedly, we get spectral windows that look very much like a gaussian pdf.

Another common kernel choice is the modified Daniell kernel which puts half-weights at the end-points. The book also talks about the Dirichlet kernel and the Fejer kernel.