

Introduction to Time Series

Lecture 2: Time series models, trend and autocovariance

Matthew F. Dixon

January 19th, 2022

Outline

- 1 General notation for time series data and models
- 2 Trend
- 3 Autocovariance and autocorrelation
- 4 Least squares estimation of a trend
- 5 Autocorrelated residuals with trend estimation
- 6 Generalized least squares for trend estimation

Background Reading (Brockwell and Davis)

- Section 1.3.2 on estimating trends
- Section 1.4 on Autocorrelation
- See pages 212, 386 generalized least squares (GLS) estimation

Definition: Time series data and time series models

- A time series is a sequence of numbers, called **data**. In general, we will suppose that there are N numbers, $y_1^*, y_2^*, \dots, y_N^*$, collected at an increasing sequence of times, t_1, t_2, \dots, t_N .
- We write $1:N$ for the sequence $\{1, 2, \dots, N\}$ and we write the collection of numbers $\{y_n^*, n = 1, \dots, N\}$ as $y_{1:N}^*$.
- A time series model is a collection of jointly defined random variables, Y_1, Y_2, \dots, Y_N .
- We write this collection of random variables as $Y_{1:N}$.
- Like all jointly defined random variables, the distribution of $Y_{1:N}$ is defined by a joint density function, which we write as

$$f_{Y_{1:N}}(y_1, \dots, y_N; \theta). \quad (1)$$

- Here, θ is a vector of parameters.
- The density is defined for arbitrary outcomes $y_{1:N}$. Its value at the data, $y_{1:N}^*$, is called the **likelihood**.

- Our notation for densities generalizes. We write $f_Y(y)$ for the density of a random variable Y evaluated at y , and $f_{YZ}(y, z)$ for the joint density of the pair of random variables (Y, Z) evaluated at (y, z) . We can also write $f_{Y|Z}(y | z)$ for the conditional density of Y given Z .
- For discrete data, such as count data, our model may also be discrete and we interpret the density function as a probability mass function. Expectations and probabilities are integrals for continuous models, and sums for discrete models. Otherwise, everything remains the same. We will write formulas only for the continuous case. You can swap integrals for sums when working with discrete models.
- Scientifically, we postulate that $y_{1:N}^*$ is a realization of $Y_{1:N}$ for some unknown value of θ .
- Our notation distinguishes between the model, $Y_{1:N}$; an arbitrary realization of the model, $y_{1:N}$; the specific sequence of numbers that we observed as data, $y_{1:N}^*$.
- Time series texts commonly ignore these distinctions. For example, ? call all these quantities y_t .

Review: Random variables

Question 2.1. What is a random variable?

Question 2.2. What is a collection of jointly defined random variables?

Question 2.3. What is a probability density function? What is a joint density function? What is a conditional density function?

Question 2.4. What does it mean to say that “ θ is a vector of parameters?”

There are different answers to these questions, but you should be able to write down an answer that you are satisfied with.

Review: Expectation

Random variables usually have an expected value, and in this course they always do. We write $\mathbb{E}[X]$ for the expected value of a random variable X .

Question 2.5. Review question: What is the expected value? How is it defined? How can it fail to exist for a properly defined random variable?

Definition: The mean function, or trend

- We define the **mean function**, for $n \in 1:N$, by

$$\mu_n = \mathbb{E}[Y_n] = \int_{-\infty}^{\infty} y_n f_{Y_n}(y_n) dy_n$$

- We use the words “mean function” and “trend” interchangeably.
- We say “function” since we are thinking of μ_n as a function of n .
- Sometimes, it makes sense to think of time as continuous. Then, we write $\mu(t)$ for the expected value of an observation at time t . We only make observations at the discrete collection of times $t_{1:N}$ and so we require $\mu(t_n) = \mu_n$.
- A time series may have measurements evenly spaced in time, but our notation does not insist on this. In practice, time series data may contain missing values or unequally spaced observations.

- μ_n may depend on θ , the parameter vector. We can write $\mu_n(\theta)$ to make this explicit.
- We write $\hat{\mu}_n(y_{1:N})$ to be some estimator of μ_n , i.e., a map which is applied to data to give an estimate of μ_n . An appropriate choice of $\hat{\mu}_n$ will depend on the data and the model.
- Usually, applied statistics courses do not distinguish between estimators (functions that can be applied to any dataset) and estimates (an estimator evaluated on our data). For thinking about model specification and diagnosing model misspecification it is helpful to bear the difference in mind.
- The **estimated mean function** or **estimated trend** is the value of the estimator when applied to our data. Here, we write this as

$$\hat{\mu}_n = \hat{\mu}_n(y_{1:N}^*). \quad (2)$$

- We are using a standard abuse of notation where $\hat{\mu}_n$ denotes both the estimator function and its value evaluated at $y_{1:N}^*$.

- For example, suppose a model with $\mu_n = \mu$, so the mean is assumed constant. In this case, the model is called **mean stationary**. Then, we might estimate μ using the mean estimator,

$$\hat{\mu}(y_{1:N}) = \frac{1}{N} \sum_{n=1}^N y_n. \quad (3)$$

- $\hat{\mu} = \hat{\mu}(y_{1:N}^*)$ denotes the sample mean.
- We can compute the sample mean, $\hat{\mu}$, for any dataset. It is only a reasonable estimator of the mean function when a mean stationary model is appropriate.
- Notice that trend is a property of the model, and the estimated trend is a function of the data.
- Similarly, data cannot be mean stationary. A model can be mean stationary.

Question 2.6. Properties of models vs properties of data.

Consider these two statements. Does it matter which one we use?

“The data looks mean stationary.”

“A mean stationary model looks appropriate for these data.”

Definition: The autocovariance function

- We will assume that variances and covariances exist for the random variables $Y_{1:N}$. We write

$$\gamma_{m,n} = \mathbb{E}[(Y_m - \mu_m)(Y_n - \mu_n)]. \quad (4)$$

This is the **autocovariance** function, a function of m and n .

- We write Γ for the $N \times N$ matrix whose (m, n) entry is $\gamma_{m,n}$.
- If the covariance between two observations depends only on their time difference, the time series model is **covariance stationary**. For observations equally spaced in time, the autocovariance function is then a function of a lag, h ,

$$\gamma_h = \gamma_{n,n+h}. \quad (5)$$

- For a covariance stationary model, and some mean estimate $\hat{\mu}_n$, an estimate for γ_h is the **sample autocovariance function**,

$$\hat{\gamma}_h = \frac{1}{N} \sum_{n=1}^{N-h} (y_n^* - \hat{\mu}_n) (y_{n+h}^* - \hat{\mu}_{n+h}). \quad (6)$$

Definition: The autocorrelation function

- Dividing the autocovariance by the variance gives the **autocorrelation function** ρ_h given by

$$\rho_h = \frac{\gamma_h}{\gamma_0}.$$

We can analogously construct the standard autocorrelation estimator,

$$\hat{\rho}_h(y_{1:N}) = \frac{\hat{\gamma}_h(y_{1:N})}{\hat{\gamma}_0(y_{1:N})},$$

which leads to an estimate known as the **sample autocorrelation**,

$$\hat{\rho}_h = \hat{\rho}_h(y_{1:N}^*) = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}.$$

- It is common to use ACF as an acronym for any or all of the autocorrelation function, sample autocorrelation function, autocovariance function, and sample autocovariance function. **If you use the acronym ACF, you are expected to define it, to remove the ambiguity.**

Sample statistics exist without a model

- The sample autocorrelation and sample autocovariance functions are statistics computed from the data. They exist, and can be computed, even when the data are not well modeled as covariance stationary. However, in that case, it does not make sense to view them as estimators of the autocorrelation and autocovariance functions (which exist as functions of a lag h only for covariance stationary models).
- Formally, we should not talk about the correlation or covariance of data. These are properties of models. We can talk about the sample autocorrelation or sample autocovariance of data.

Estimating a trend by least squares

Let's analyze a time series of global mean annual temperature downloaded from `climate.nasa.gov/system/internal_resources/details/original/647_Global_Temperature_Data_File.txt`. These data are in degrees Celsius measured as an anomaly from a 1951-1980 base. This is climatology jargon for saying that the sample mean of the temperature over the interval 1951-1980 was subtracted from all time points.

```
global_temp <- read.csv("../data/Global_Temperature.csv", header = TRUE)
str(global_temp)
```

Mean global temperature anomaly, degrees C

```
plot(Annual~Year,data=global_temp,ty="l")
```


- These data should make all of us pause for thought about the future of our planet.
- Understanding climate change involves understanding the complex systems of physical, chemical and biological processes driving climate.
- It is hard to know if gigantic models that attempt to capture all important parts of the global climate processes are in fact a reasonable description of what is happening.
- There is value in relatively simple statistical analysis, which can at least help to tell us what evidence there is for how things are, or are not, changing.
- Here is a quote from *Science* (18 December 2015, volume 350, page 1461; I've added some emphasis).

“Scientists are still debating whether—and, if so, how—warming in the Arctic and dwindling sea ice influences extreme weather events at midlatitudes. **Model limitations, scarce data** on the warming Arctic, and the **inherent variability** of the systems make answers elusive.”

Fitting a least squares model with a quadratic trend

Perhaps the simplest trend model that makes sense looking at these data is a quadratic trend,

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2. \quad (7)$$

To write the least squares estimate of β_0 , β_1 and β_2 , we set up matrix notation. Write

$$\mu = (\mu_1, \mu_2, \dots, \mu_N)^t \quad (8)$$

for the column vector describing the mean function, and similarly,

$$\beta = (\beta_0, \beta_1, \beta_2)^t. \quad (9)$$

Then, defining

$$Z = \begin{pmatrix} 1 & 1880 & 1880^2 \\ 1 & 1881 & 1881^2 \\ 1 & 1882 & 1882^2 \\ \vdots & \vdots & \vdots \end{pmatrix}, \quad (10)$$

we can write

$$\mu = Z\beta.$$

We write $y_{1:N}^*$ as a column vector,

$$y = (y_1, y_2, \dots, y_N)^t. \quad (11)$$

The ordinary least squares (OLS) estimator of β is

$$\hat{\beta}_{OLS}(y_{1:N}) = (Z^t Z)^{-1} Z^t y, \quad (12)$$

with corresponding OLS estimate

$$\hat{\beta}_{OLS} = \hat{\beta}_{OLS}(y_{1:N}^*) = (Z^t Z)^{-1} Z^t y^*.$$

We can carry out this computation in R by

```
lm_fit <- lm(Annual~Year+I(Year^2),data=global_temp)
```

where `I()` is a function that tells R to construct `Year^2` as a variable, and inhibits interpretation in the R model formula notation.

```
summary(lm_fit)
```

We can check visually how well this model fits the data.

```
yr <- 1880:2026
Z <- cbind(1,yr,yr^2)
beta <- coef(lm_fit)
prediction <- Z%*%beta
plot(Annual~Year,data=global_temp,ty="l",xlim=range(yr),
     ylim=range(c(global_temp$Annual,prediction),na.rm=TRUE),
     lty="dashed")
lines(x=yr,y=prediction,col="red")
```

- The overall estimated trend seems a reasonable fit for the data.
- If we want to attach uncertainty to our parameter estimates, and consequently to our forecast, we need a time series model $Y_{1:N}$, which we write in column vector form as

$$Y = (Y_1, Y_2, \dots, Y_N)^t.$$

- The usual model behind OLS is the independent error model, known in time series analysis as the **white noise** error model:

$$[L1] \quad Y = Z\beta + \epsilon,$$

where $\epsilon = \epsilon_{1:N}$ is a vector of independent identically distributed (iid) random variables with mean zero and constant variance, $\mathbb{E}[\epsilon_n] = 0$ and $\text{Var}(\epsilon_n) = \sigma^2$. Standard linear model software, such as `lm` in R, provides confidence intervals based on this model.

- Under model L1, the estimator $\hat{\beta}_{OLS}(y_{1:N})$ is unbiased. This can be checked:

$$\begin{aligned} \mathbb{E}[\hat{\beta}_{OLS}(Y_{1:N})] &= \mathbb{E}[(Z^t Z)^{-1} Z^t Y] \\ &= \mathbb{E}[(Z^t Z)^{-1} Z^t \{Z\beta + \epsilon\}] \\ &= (Z^t Z)^{-1} Z^t \{Z\beta + \mathbb{E}[\epsilon]\} \\ &= (Z^t Z)^{-1} (Z^t Z) \beta \\ &= \beta \end{aligned}$$

- A result for linear models is that $\hat{\beta}_{OLS}(y_{1:N})$ is the minimum variance unbiased estimator for model L1.

Variance/covariance for the white noise error model

- The variance/covariance matrix of $\hat{\beta}_{OLS}(Y_{1:N})$ under this model (?) is

$$\text{Cov}[\hat{\beta}_{OLS}(Y_{1:N})] = \sigma^2 (Z^t Z)^{-1},$$

which is estimated using an estimate for σ of

$$\hat{\sigma}_{OLS} = \sqrt{\frac{1}{N-d} (y^* - Z\hat{\beta}_{OLS})^t (y^* - Z\hat{\beta}_{OLS})},$$

where d is the number of covariates, i.e., the number of columns of Z .

Let's look at the residuals to assess how appropriate this model is here.

```
acf(resid(lm_fit))
```

Investigating the ACF plot provided by R

- The horizontal dashed lines on the graph of the sample autocorrelation function (ACF) give a measure of chance variation under the null hypothesis that the residuals are iid.
- At each lag h , the chance that the estimated ACF falls within this band is approximately 95%, under the null hypothesis.
- Thus, under the null hypothesis, one expects a fraction of $1/20$ of the lags of the sample ACF to fall outside this band.
- Here, the sample ACF confirms what we can probably see from the plot of the fitted model: the variation around the fitted model is clustered in time, so the sample ACF of the residuals is not consistent with a model having independent error terms.

Question 2.7. How does R construct these horizontal dashed lines?

Figuring out what R does

- How would you check what R actually does when it constructs these dashed lines? What approximation is being made? When is that approximation appropriate?
- Hint: If you type 'acf' in R, you get the source code for the acf function. You'll see that the plotting is done by a service function 'plot.acf'. This service function is part of the package, and is not immediately accessible to you. Nevertheless, you can check the source code as follows
- Notice, either from the help documentation '?acf' or the last line of the source code 'acf' that this function resides in the package 'stats'.
- Now, you can access this namespace directly, to list the source code, by
stats:::plot.acf

- To relate this source code to the task of testing for lack of correlation, a standard topic in undergrad introductory statistics courses, the critical line of code seems to be

```
clim0 <- if (with.ci) qnorm((1 + ci)/2)/sqrt(x$n.used)
```

This appears to use a normal distribution approximation for the sample autocorrelation estimator, with mean zero and standard deviation $1/\sqrt{N}$.

Generalized least squares

- Suppose for the moment that we knew the covariance matrix, Γ , for a model with dependent errors,

$$[L2] \quad Y = Z\beta + \zeta, \quad \zeta \sim N[0, \Gamma].$$

We read “ $\zeta \sim N[0, \Gamma]$ ” as “ ζ follows a multivariate normal distribution with mean zero and covariance matrix Γ .”

- The minimum variance unbiased estimator of β for model L2 is the generalized least square (GLS) estimator

$$\hat{\beta}_{GLS}(y_{1:N}) = (Z^t \Gamma^{-1} Z)^{-1} Z^t \Gamma^{-1} y.$$

- The OLS estimator remains unbiased for L2 (you can check this as an exercise). In this sense it remains a reasonable estimator. It is often a practical solution to use the OLS estimator, especially for preliminary data analysis. We don't know Γ so can't necessarily make a good estimator based on the GLS model. It might be easier to get an estimate of Γ once we have a reasonable estimate of the trend.

- For model L2, the variance of the OLS estimator is

$$\text{Var} \left[\hat{\beta}_{OLS}(Y_{1:N}) \right] = (Z^t Z)^{-1} Z^t \Gamma Z (Z^t Z)^{-1}. \quad (13)$$

This is different from the variance under model L1.

- **CONCLUSION.** It is okay to do ordinary linear regression for data which are not well modeled with uncorrelated errors. However, if we do so, we should not trust the error estimates coming from L1.
- This is an example of a situation where some parts of the output from statistical software are reasonable (here, the parameter estimates from `lm`) and other parts are unreasonable (the corresponding standard errors and any tests based on them). The theory helps us decide which bits of computer output to use and which to ignore.