TimeSeries for Economics and Finance

From Data to Signals

Florian Ielpo
f.ielpo@lombardodier.com

Lombard Odier Investment Management

February 1, 2023

Class 1:
Linear Univariate Models

Why is estimation so important in economics and finance?

- Economists and financial engineers produce many models with sets of unknown parameters.
- These models need to be made as "realistic" as possible.
- We therefore need to squeeze seem so that they can resemble reality.
- How do we do that? $\Rightarrow$ we need to "calibrate" their parameters in a way that make them the closest they can get to reality.
- Many technics: OLS, GMM, Maximum Likelihood, Simulated Method of Moments, Indirect Inference, Instrumental Variables... When use what?
- Many models: CAPM, Gordon and Shapire, Taylor rule, Growth models, Vasicek models... they all incorporate some time series dimensions that needs to be dealt with.
- Another way of thinking of timeseries analysis: exploring linear relationships between economic and financial variables.

The Sims Criticism

Two ways to do time series analysis:

- Starting from a theoretical model, and trying to estimate its parameters given a time period.

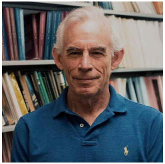
  $\Rightarrow$ sometimes refered to as *calibration*.

- Starting from a time series, exploring its salient feature and creating a model around them.

  $\Rightarrow$ Usually called the "Sims Criticism", after its article “Macroeconomics and Reality.” Econometrica 1980, 48: 1 (January): 1-48.

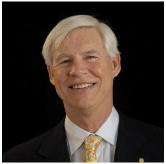
Some of the Econometricians We Are Going To Talk About

Christophe Sims



Clive Granger

Robert Engle



Monika Piazzesi

Andrew Patton

Yacine Aït Sahalia

Yacine Aït Sahalia

Christian Gourieroux

A first toy example

$$r_t = \alpha + \beta_1 x_t^{(1)} + \beta_2 x_t^{(2)} + ... + \beta_p x_t^{(p)} + \sigma \epsilon_t$$

defines the linear model with:

- α the constant term, aka the intercept
- r_t is known as the 'endogenous' variable: it depends on other factor
- $x_t^{(j)}$ are the 'exogenous' variables: they *explain* r_t without anything else explaining them
- ϵ_t is what the exogenous variables cannot explain: model error, observation errors...
- the β_j turn $x_t^{(j)}$ into r_i

$\Rightarrow x_t^{(j)}$ are "factors" in the financial literature.

The Fama-French model

$$r_t = \alpha + \beta_1 f_t^{(Market)} + \beta_2 f_t^{(SMB)} + \beta_3 f_t^{(HML)} + \sigma\epsilon_t$$

This model is widely used to explain the performance of individual stocks across three market factors:

- r_t is the return on a given individual stock.
- Market is the market factor (what globally happens across equity indices)
- SMB stands for Small Minus Big: the return of the largest companies vs. the return on the smallest (size factor)
- HML stands for High Minus Low: the returns of the most expensive vs. the cheapest companies (value factor, high book to market value vs. low book to market value)
- ϵ_t is now whatever the model does not explain and/or uncertainty sources arising from individual stocks (idiosyncratic risk).

$\Rightarrow$ We still need to estimate the model, i.e. assigining the most realistic values to its parameters (the β and α).

Estimation Technics

There exists broadly three estimation technics that you have heard of:

- Ordinary Least Squares: assigning to parameters values such that the errors between reality and model are as small as possible.
- Maximum Likelihood: assigning to parameters values such that the model has the greatest achievable probability of being right.
- Generalized Method of Moments: assigning to parameters values that make sure that the moments of the model are as close as possible to the moment of the phenomenon we are trying to describe.

In this class: mainly Maximum Likelihood as for the model class we will be dealing with, it is usually the strongest method.

The Least Square Estimation to the Linear Model

OLS (Ordinary Least Square) estimators are easily obtained when presented in the form of matrices.
We can rewrite the linear model in a matrix form:

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ 2 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_1^{Market} \\ x_2^{Market} \\ \vdots \\ x_n^{Market} \end{pmatrix} + \beta_2 \begin{pmatrix} x_1^{SMB} \\ x_2^{SMB} \\ \vdots \\ x_n^{SMB} \end{pmatrix} + \beta_3 \begin{pmatrix} x_1^{HML} \\ x_2^{HML} \\ \vdots \\ x_n^{HML} \end{pmatrix} + \sigma \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^{Market} & x_1^{SMB} & x_1^{HML} \\ 1 & x_2^{Market} & x_2^{SMB} & x_2^{HML} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{Market} & x_n^{SMB} & x_n^{HML} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \sigma \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\Leftrightarrow R = X\beta^\top + \sigma\epsilon$$

The Least Square Estimation to the Linear Model

The OLS estimator solves the following least square criterion:

$$\beta_{OLS} = \min_{\beta} \left(R - X\beta^\top\right)^\top \left(R - X\beta^\top\right)$$

Key matrix operations:

- let $y = Ax$ be a matrix product, x being a $p \times 1$ matrix of parameters and A a $n \times p$ matrix. Then:$\partial_x y = A$.
- let $y = x^\top Ax$ be a matrix product, x being a $p \times 1$ matrix of parameters and A a $n \times p$ matrix. Then:$\partial_x y = x^\top(A^\top + A)$. If A is symmetrical, then $\partial_x y = 2x^\top A$.

The least square estimate solves:

$$\min_{\beta} R^\top R - R^\top X\beta^\top - \beta X^\top R + \beta X^\top X\beta^\top$$

Exercise: writes down the FOC and solves them for β.

The Least Square Estimation to the Linear Model

Opening brackets yields:

$$\left(R - X\beta^\top\right)^\top \left(R - X\beta^\top\right) = (R - \beta X^\top)(R - X\beta^\top)$$

$$= R^\top R - R^\top X\beta^\top - \beta X^\top R + \beta X^\top X\beta^\top$$

Differentiating with respect to β yields:

$$\partial_\beta R^\top R - R^\top X\beta^\top - \beta X^\top R + \beta X^\top X\beta^\top = -X^\top R - X^\top R + 2\beta X^\top X$$

Finally solving yields

$$\beta = (X^\top X)^{-1}(X^\top R)$$

The Least Square Estimation to the Linear Model

The OLS estimator is a cocktail of random variable, which means it is also a random variable and its distribution is usually Gaussian either as:

- conditionally upon the knowledge of the x_ts, it is Gaussian because of the Central Limit Theorem (asymptotic normality).
- or because the conditional distribution of ϵ is Gaussian (finite distance normality).

$$\begin{aligned}\hat{\beta}^\top &= (X^\top X)^{-1}(X^\top R) = (X^\top X)^{-1}(X^\top(X\beta^\top + \sigma\epsilon)) \\ &= (X^\top X)^{-1}X^\top X\beta^\top + \sigma(X^\top X)^{-1}X^\top\epsilon\end{aligned}$$

This implies:

$$E[\hat{\beta}^\top] = \beta^\top$$

$$V[\hat{\beta}^\top] = \sigma^2 V[(X^\top X)^{-1}X^\top\epsilon] = \sigma^2(X^\top X)^{-1}$$

The Student Test

A test to check whether a parameter's value is equal to an assumed value.

$$H_0 : \beta_i = x$$

$$H_1 : \beta_i \neq x$$

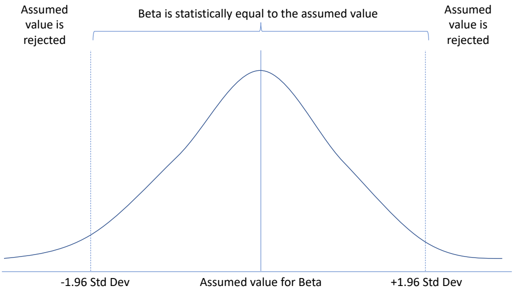
Under H_0,

$$\beta_i \sim N(0, \sigma_{\beta_i})$$

Therefore, 95% of the time, if β_i is really equal to x as assumed under H_0, the following statement should be true:

$$\frac{\hat{\beta}_i - x}{\sigma_{\beta_i}} \in [-1.96 + 1.96]$$

The Student Test



The Least Square Estimation to the Linear Model

The R2

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.
- 0% indicates that the model explains none of the variability of the response data around its mean.

$$R2 = 1 - \frac{(R - \hat{R})^\top (R - \hat{R})}{V(R)}$$

Fama French Model Estimated

Fama French Model:

$$r_t = \alpha + \beta_1 f_t^{(Market)} + \beta_2 f_t^{(SMB)} + \beta_3 f_t^{(HML)} + \sigma \epsilon_t$$

Estimated using daily data (taken from Fama-French) over the 2011-2020 period.

Output Table:

	β_3	β_2	β_1	α	R2
Apple	-0,55	-0,30	1,13	0,00	0,48
Netflix	-1,02	0,32	1,10	0,00	0,18
Pfizer	-0,11	-0,30	0,76	0,00	0,42

Over that period:

- Apple and Netflix have been aggressive stocks, Pfizer more defensive.
- 18 to 48% of daily returns are explained by Fama-French, Netflix bearing the highest idiosyncratic risk.
- Pfizer is defensive, rather big and anti-value (more growth). Netflix has been on the Small side.

Fama French Model Estimated

Output of the Fama-French Regression with Netflix (2010-2020)

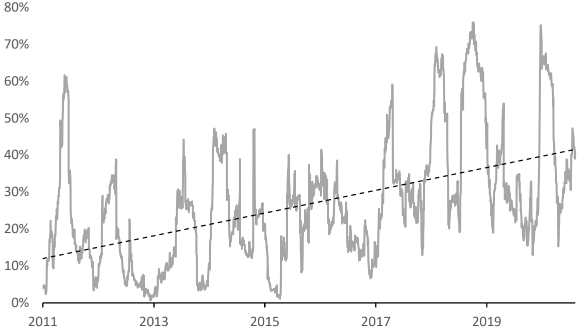
	β_3	β_2	β_1	α
Estimate	-1,02	0,32	1,10	0,00
Standard Deviation	0,09	0,11	0,05	0,00
Estimate/Standard Deviation Ratio (Student)	-11,43	2,95	20,17	1,59
R2	0,18			

Conclusions:

- Netflix has been over the period a growth small stock rather aggressive.
- All parameters are statistically significant but the α.
- The 3-factor model explains 18% of Netflix's returns.

Fama French Model Estimated

Rolling estimation of Netflix's R2



Maximum likelihood

Very important methodology:
Probability of observing a draw $x_1, x_2, ..., x_n$ given a parametric model:

$$P(x_1, x_2, ..., x_n|\theta) = P(\theta|x_1, x_2, ..., x_n)$$

ML method means picking θ such that this probability is max, i.e. so that it can be as likely as possible that this sample has been drawn from this parametric model.
Problem: for most model impossible to compute this probability given that:

- for a continuous distribution this probability is equal to 0.
- for most models, the joint distribution cannot be derived (need to find a trick there).

$\Rightarrow$ yet ML makes best use of all the information about the distribution.

Maximum likelihood

Solution: probability is proportional to density

$$P(x_1|\theta) = \int_{x_1-\epsilon}^{x_1+\epsilon} f(x)dx$$

Instead of joint probability, use joint density:

$$f(x_1, x_2, ..., x_n|\theta)$$

With iid data:

$$f(x_1, x_2, ..., x_n|\theta) = \prod f(x_t|\theta)$$

Maximizing a product is numerically complex, better to maximize the log of it

$$\max \log \prod f(x_t|\theta)$$

Which then becomes:

$$\max \sum \log f(x_t|\theta)$$

For most distribution, density uses exponentials $\Rightarrow$ tractable expression.

Maximum likelihood

ML estimates solve for i.i.d. observations:

$$\max_{\theta} \sum_{i=1}^{n} \log f(x_t|\theta)$$

or equivalently:

$$\sum_{i=1}^{n} \partial_\theta \log f(x_t|\theta) = 0$$

Several imporant hypotheses:

- Identification: $\forall \theta \neq \theta^*$, $\sum_{i=1}^{n} \log f(x_t|\theta) \neq \sum_{i=1}^{n} \log f(x_t|\theta^*)$. One set of parameter, one likelihood value. Can be violated for several finance models such as Vaiscek model, or in probit models...
- first three derivatives of $\log f(x_t|\theta)$ are continuous and finite $\forall \theta$
- $\mathbb{E}[\partial_\theta \log f(x_t|\theta)] < \infty$ and $\mathbb{E}[\partial_{\theta^2} \log f(x_t|\theta)] < \infty$
- $|\partial_{\theta^3} \log f(x_t|\theta)| < h$, with $\mathbb{E}[h] < \infty$.

Last three conditions mean we deal with regular densities.

Maximum likelihood

Why love maximum likelihood? Because of the asymptotic efficiency of its estimates.
ML estimates are:

- Consistency :$\mathbb{E}[\hat{\theta}_{ML}] = \theta_0$ or plim$\hat{\theta} = \theta_0$
- Asymptotic normality: $\hat{\theta} \rightarrow N\left(\theta_0, \left[-\mathbb{E}\left[\partial_{\theta^2} \log L\right]\right]^{-1}\right)$
- Asymptotic efficiency: $\hat{\theta}$ reaches the FDCR lower bound.
- Invariance: the ML estimate of $\gamma_0 = g(\theta_0)$ is $\hat{\gamma}_0 = g(\theta_0)$

Third item means that when the ML conditions are granted, then ML estimates are the most efficient estimates of the world.
$I_\theta = \mathbb{E}\left[\partial_{\theta^2} \log L\right]$ is called the information matrix, as it informs its user on the variance of the estimates around the optimal value.
Also: $-\mathbb{E}\left[\partial_{\theta^2} \log L\right] = \mathbb{E}\left[\partial_\theta \log L \times \partial_\theta \log L\right]$ for a single parameter – or its matrix equivalent.

Pseudo/Quasi Maximum likelihood

What happens if the model is misspecified?

$\Rightarrow$ pseudo maximum likelihood: under some circumstances, the ML estimates of wrongly specified model remains consistent! Globally, $f(.)$ must below to the exponential family.

$$\sqrt{n}(\hat{\theta}_{PML} - \theta_0) \rightarrow N\left(0, H^{-1}\Phi H^{-1}\right)$$

with

$$\Phi = Cov\left(\frac{\partial \log f(x_t)}{\partial \theta}\right)$$

Usually: Φ estimated via $\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial \log f(x_t|\theta_0)}{\partial \theta_0}\right]\left[\frac{\partial \log f(x_t|\theta_0)}{\partial \theta_0}\right]$

When $\Phi = -H(\theta_0)$ then PML and ML estimates are the same.

Conditional Densities

In the case of timeseries, serial dependencies are common things: $f(x_t)$ are *not* i.i.d.
Solution: conditioning on past information.

$$f(x_1, x_2, ..., x_n) = f(x_1) \times f(x_2|x_1) \times \ldots f(x_n|x_1, \ldots, x_{n-1})$$

The loglikelhihood is then the sum of the conditional log-likelihoods:

$$LogL = \sum_{t=1}^{n} \log f(x_t|x_1, \ldots, x_{t-1})$$

Usually denoted:

$$LogL = \sum_{t=1}^{n} \log f(x_t|\underline{x_{t-1}})$$

The Change in Variable Theorem

The Change in Variable Theorem is essential to compute the conditional distribution for a lot of models.

Theorem Suppose X is continuous with probability density function $f_X(x)$. Let $y = h(x)$ with h a strictly increasing continuously differentiable function with inverse $x = g(y)$. Then $Y = h(X)$ is continuous with probability density function $f_Y(y)$ given by

$$f_Y(y) = f_X(g(y))g'(y)$$

Example: Let $\epsilon_t \sim N(0, 1)$. Find the distribution of $Y_t = a + \sigma\epsilon_t$.
Solution:

$$f_Y(y) = f_\epsilon\left(\frac{y-a}{\sigma}\right) \times \left(\frac{y-a}{\sigma}\right)'$$

Therefore

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{1}\left(\frac{y-a}{\sigma}\right)^2\right)$$

this implies $Y \sim N(a, \sigma)$

Maximum Likelihood Estimation

Considering the following model:

$$r_i = \alpha + \beta x_t + \sigma \epsilon_i$$

with ϵ_i following a Gaussian distribution $N(0, 1)$.
Step 1: using the change of variable theorem, find the distribution of r_i.

$$f_{r_i}(r_i) = f_{\epsilon_i}\left(\frac{r_i - \mu}{\sigma}\right) \times \frac{\partial \frac{r_i - \mu}{\sigma}}{\partial r_i}$$

From that result, $r_i \sim N(\alpha + \beta x_t, \sigma)$.
Step 1: write the log-likelihood of the model and derive the FOC:

$$\log L(r_1, r_2, ..., r_n|\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \sum_{i=1}^{n}\frac{1}{2}\left(\frac{r_i - \alpha - \beta x_t}{\sigma}\right)^2$$

Maximum Likelihood Estimation

FOC:

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} x_t \left(\frac{r_i - \alpha - \beta x_t}{\sigma} \right) = 0$$

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} \left(\frac{r_i - \alpha - \beta x_t}{\sigma} \right) = 0$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(r_i - \alpha - \beta x_t)^2}{\sigma^3} = 0$$

Which yields the following estimates:

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} (r_i - \beta_i x_t), \hat{\beta} = \frac{\sum_{i=1}^{n} (r_i - \hat{\alpha} - \beta_i x_t)}{\sum_{i=1}^{n} x_t^2}$$

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (r_i - \hat{\alpha} - \hat{\beta}_i x_t)^2$$

How to assess the quality of a model?

Two different measures:

1. In sample: the R2 ("R-square") looks into the quality of the fit:

$$R^2 = 1 - \frac{\mathbb{V}[\hat{\epsilon}_i]}{\mathbb{V}[r_i]}$$

   Ranges from 0 to 1, close to one when the residuals have almost no variability left.

2. Out of sample:
   - Root Mean Squared Errors$= \sqrt{\frac{1}{n}\sum_i (r_i - \hat{r}_i)^2}$, with $\hat{r}_i$ the forecast variables r_i, with n forecasts.
   - Mean Absolute Errors$= \frac{1}{n}\sum_i |r_i - \hat{r}_i|$.

Some others tests people could ask you about

1. Fisher test: a significant test for all parameters in the meantime. H_0 is all parameters are equal to 0. Under H_0 the following test statistic follows a Fisher distribution:

$$F_{\text{test}} = \frac{R^2}{1-R^2}\frac{n-p-1}{p} \sim F(p, n-p-1),$$

   with n the number of observations and p the number of parameters.

2. Student test: testing for only one parameter's significance. Significance means "can this parameter be set to 0?". $H_0 : \theta = 0$, with θ a given parameter. Under H_0,

$$\frac{\hat{\theta}}{\sigma(\theta)} \sim N(0,1),$$

   provided that you have enough information.

3. Durbin and Watson test: an unusual test to check that errors are not *autocorrelated*. The test statistics:

$$d = \frac{\sum_i (\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_i \hat{\epsilon}_i^2}.$$

   d oscillates between 4 and 0. 2 means no autocorrelation.

Forecasting

Once the relationship is estimated, we can make forecasts: $r_j = \beta x_j + \sigma \epsilon_j$ (simplified version with no intercept).
The 'forecast' error: $e_j = y_j - \hat{y}_j = x_j(\beta - \hat{\beta}) + \sigma \epsilon_j$, with the variance:

$$V(e_j) = x_j^2 V(\hat{\beta}) + \sigma^2 = x_j^2 \frac{\sigma^2}{\sum_{i=1}^n x_t^2} + \sigma^2$$

Forecast errors are function of

- the estimation errors' variance.
- the forecast x_j's values
- the volatility of the x_t as

$$V(e_j) = \sigma^2 \left(1 + \frac{x_j^2}{\sum_{i=1}^n x_t^2} \right)$$

A toy example: the Taylor rule

Hypothesized relationship between the Fed's decision rate and economic variables:

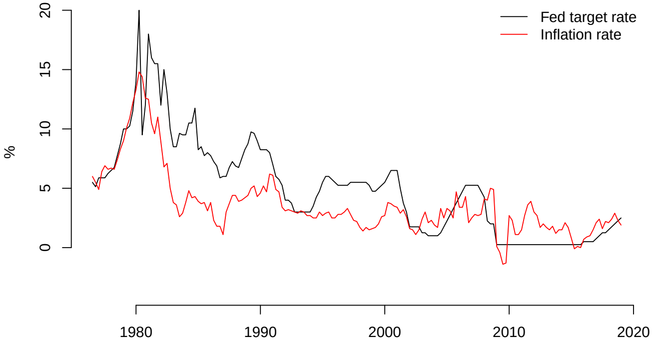
$$R_t = \alpha + \beta_\pi \pi_t + \beta_g g_t + \epsilon_t$$

with

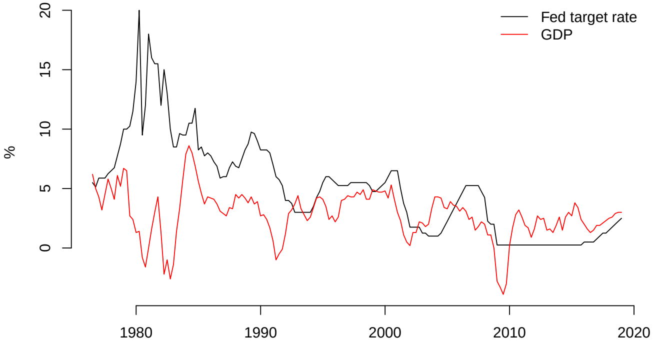
1. π_t the inflation rate at time t
2. g_t the growth rate of GDP at time t
3. R_t the Fed's decision rate.

The model assumes that ϵ_t is iid and that $Cov(\epsilon_t, \epsilon_{t-1}) = 0$: monetary policy shocks are not persistent.

Historical evolution



Historical evolution

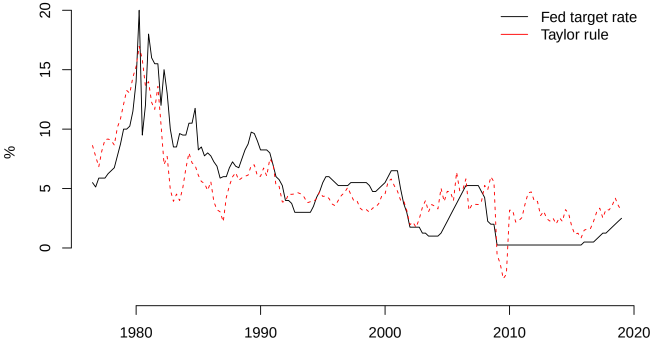


Estimation results: Taylor rule

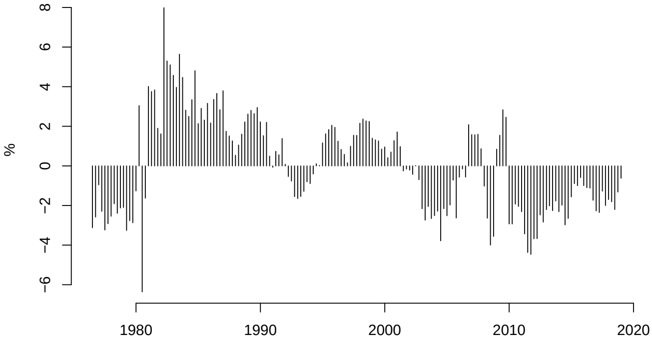
```
Call:
lm(formula = y ~ as.matrix(x[, c(1, 3)]))

Residuals:
    Min      1Q  Median      3Q     Max
-6.3599 -2.0724 -0.1582  1.7291  7.9847

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                           0.12618    0.39696   0.318  0.75099
as.matrix(x[, c(1, 3)])CPI.YOY.Index  1.10910    0.06675  16.615  < 2e-16 ***
as.matrix(x[, c(1, 3)])GDP.CYOY.Index 0.29671    0.09219   3.218  0.00155 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.449 on 168 degrees of freedom
Multiple R-squared:  0.6337,    Adjusted R-squared:  0.6293
F-statistic: 145.3 on 2 and 168 DF,  p-value: < 2.2e-16
```

```
>
> d=sum((res[-1]-res[-length(res)])^2)/sum(res^2)
>
> d
[1] 0.3673082
```

Historical fit



Residual's behaviour



What is the impact of an MA noise?

Time series effect, i.e. the disturbances are not what you think but rather $\epsilon_t = \rho\eta_{t-1} + \eta_t$. In such as case, the true model is

$$r_t = \beta r_{m,t} + \rho\eta_{t-1} + \eta_t$$

. This leads to

$$\hat{\beta} = \frac{\sum_t r_{m,t}(\beta r_{m,t} + \rho\eta_{t-1} + \eta_t)}{\sum_t r_{m,t}^2} = \beta + \rho\frac{\sum_t r_{m,t}\eta_{t-1}}{\sum_t r_{m,t}^2} + \frac{\sum_t r_{m,t}\eta_t}{\sum_t r_{m,t}^2}$$

.
Taking expectations yields $\hat{\beta} = \beta + \rho\frac{Cov(\eta_{t-1}r_{m,t})}{V[r_{m,t}]}$.
$\Rightarrow$ The last model is a *time series* model and it incorporates an MA(1) disturbance.

Time series analysis

Founding element: white noises.
ϵ_t follows a white noise process if

- i.i.d.
- Gaussian
- Expectation 0 and variance σ_ϵ.

Time series models are combinations of white noise with path dependent components.
In financial modeling, white noises are very important: Black-Scholes model is a white noise model augmented with a drift.

Stationarity

Two different definitions:

1. Strict stationarity: Let X_t be a timeseries process. $\forall h$ X_t and X_{t+h} have the *same* distribution. Hard to prove, hard to test.
2. Second order stationarity: Let X_t be a timeseries process. Then if
   - $\mathbb{E}[X_t] = \mu < \infty$
   - $\mathbb{V}[X_t] = \sigma^2 < \infty$
   - $Cov(X_t, X_{t+h}) = f(h) < \infty$

   X_t is said to be second order stationary.

And then the Wold theorem: any second order stationary process X_t can always be represented as an infinite sum of past and present shocks:

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} + \kappa_t,$$

where $\psi_i \in \mathbb{R}$, $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ and ϵ_t is a white noise. κ_t is a function of t and is not stochastic.

Classic timeseries models

Three classic timeseries models:

- Autoregressive models (AR(p)) models:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

- Moving Average models (MA(q)) models:

$$X_t = \sum_{i=1}^{q} \psi_i \epsilon_{t-i} + \epsilon_t$$

- Autoregressive models (ARMA(p,q)) models:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \psi_i \epsilon_{t-i} + \epsilon_t$$

Identification

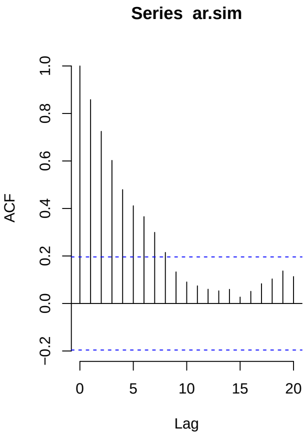
How to identify an MA model from an AR model?
Before performing an estimation, their "autocorrelogram" are informative: autocorrelation between "lags" is obtained from the following formula:

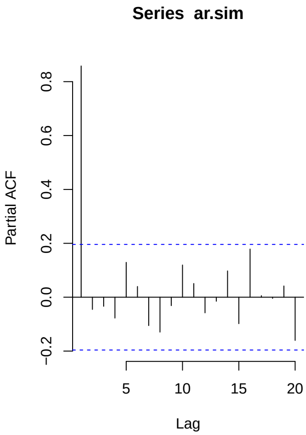
$$\gamma_h = \frac{cov(X_t, X_{t+h})}{V(X_t)}$$

- In the case of an MA(q) process, the autocorrelation function is different from 0 up to $h = q$ and 0 afterward.
- In the case of an AR(p) process, the autocorrelation function always different from zero and decays slowly to zero as h increases.

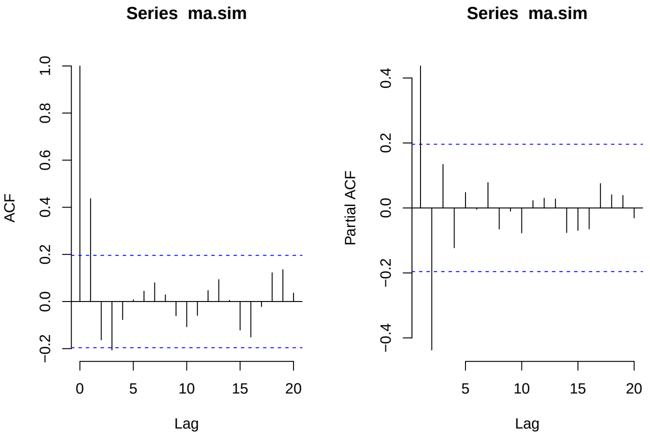
This result can be easily explained with the Wold representation of an AR process.
Partial autocorrelation can be computed from a linear regression of X_t on its lags and yields exactly the opposite image.

ACF/PACF AR(1) process





ACF/PACF MA(1) process



Estimation

Both timeseries models are no longer iid. How to deal with that? By using conditional densities, we can estimate their parameters by maximum likelihood:

$$f(X_1, X_2, X_3, ..., X_n) = f(X_1)f(X_2|X_1)f(X_3|X_2, X_1) \times ... \times f(X_n|X_1, ..., X_{n-1})$$

Taking the log of this expression yields the following log likelihood:

$$\log L = \log f(X_1) + \log f(X_2|X_1) + \log f(X_3|X_2, X_1) + ... + \log f(X_n|X_1, ..., X_{n-1})$$

The loglikelihood estimates are obtained by maximizing this expression.
The conditional densities are obtained by using the usual change in variable theorem.
Important remark: the MA process is path dependent and the ϵ_t are unobservable. We can only know its likelihood function for a given set of parameters. It therefore needs to be maximized numerically. Once estimated it yields a timeseries of individual shocks as a by-product.
The case of an AR process is much simpler given it is a linear model in observed variables (past realizations of X_t): OLS estimates can be used an coincide with ML estimates.

How to chose p and q?

Three information criterions can be used:

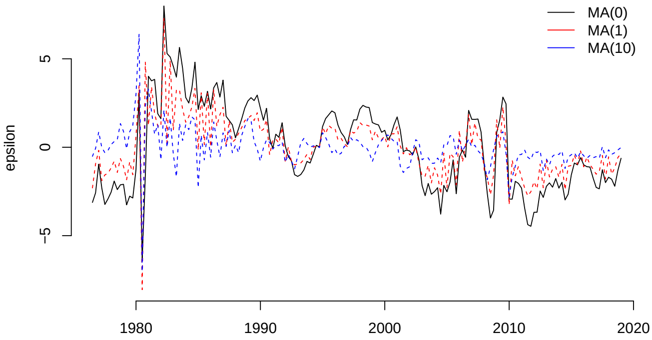
- Akkaike criterion: $AIC(p,q) = 2(p+q) - 2\log L$
- Schwartz criterion: $BIC(p,q) = 2(p+q)\log n - 2\log L$
- Hanan and Quinn criterion: $HQ(p,q) = 2(p+q)\log\log n - 2\log L$,

with n the number of observations. The right p and q should minimize one of these three information criterion.

Back on the Taylor rule estimation

```
> arima(y,order=c(0,0,0),,x[,c(1,3)])

Call:
arima(x = y, order = c(0, 0, 0), xreg = x[, c(1, 3)])

Coefficients:
      intercept  CPI.YOY.Index  GDP.CYOY.Index
         0.1262         1.1091          0.2967
s.e.     0.3935         0.0662          0.0914

sigma^2 estimated as 5.892:  log likelihood = -394.28,  aic = 796.55
> arima(y,order=c(0,0,1),,x[,c(1,3)])

Call:
arima(x = y, order = c(0, 0, 1), xreg = x[, c(1, 3)])

Coefficients:
         ma1  intercept  CPI.YOY.Index  GDP.CYOY.Index
      0.7190     0.4989         1.0245          0.2719
s.e.  0.0415     0.4510         0.0777          0.1015

sigma^2 estimated as 2.984:  log likelihood = -336.48,  aic = 682.96
```

Back on the Taylor rule estimation



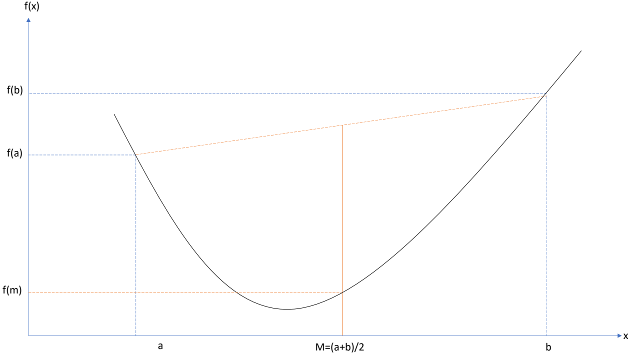
No analytical solutions?

Most of the time, you will NOT find an analytical solution to the ML max program. How can we deal with it?
Numerical optimization: when things get complicated.
Assume you need to minimize a function $h(x)$ of one variable, with $x \in [a, b]$ but can't do it through a closed form formula. Two solutions:

- Grid search: slide and dice $[a, b]$ and compute the function for each node. Retain x that yielded the max value.
- Dichotomy method: if the function has a unique max:
  - compute $m = (a + b)/2$
  - if $f(b) > f(a) > f(m)$ then the solution is between m and a and b become m
  - if $f(a) > f(b) > f(m)$ then the solution is between m and b and a become m
  - Continue until $abs(a - b) < \epsilon$.

  Works only if $f''(x) \leq 0$

An Illustration of the Dichotomy



No analytical solutions?

What happens when more than one parameter needs to be estimated? numerical optimization.

Two solutions:

- be lazy, use 'optim' in R. Eventually, you'll chose this solution.
- be a good student and understand the basics of the Newton Raphson optimization method.

A crash introduction to Newton Raphson

All gradient based optimization methods are based on the same idea: create a sequence of estimators that converges towards the 'right' solution.
Let θ be the sequence of parameters we need to estimate and θ_i the i^{th} step of the optimization.
We want to create a sequence such that

$$\theta_{i+1} = \theta_i + \lambda_i \Delta_i$$

with

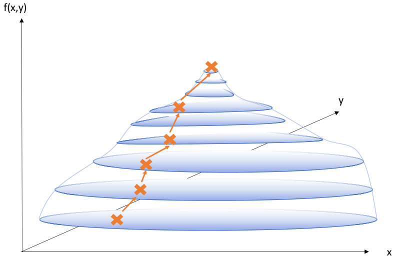
- λ_i is the size of the step to be done
- Δ_i the direction of the step.

Gradient-based methods use $\Delta_i = W_i G_i$, where G_i is the gradient matrix

$$G_i = \left[\frac{\partial LL(\theta_i)}{\partial \theta_i} \right]$$

and W_i a definitive positive matrix.

An Illustration of the Newton Raphson Method



A crash introduction to Newton Raphson

Newton-Raphson:

Assume you want to maximize $f : \mathbb{R}^k \rightarrow \mathbb{R}$, a function that can be differenciated twice and whose derivatives are continuous.
A Taylor approximation of $f(.)$ yields:

$$f(x+h) = f(x) + G^\top h + \frac{1}{2} h^\top H h$$

, with H the Hessian matrix of f.
Differenciating this expression and solving it for zero yield the optimal h:

$$\partial_h f(x+h) = G + Hh = 0 \Rightarrow h = -H^{-1} G$$

.
This means: around θ_i, the best steop is $-H_i^{-1}$ to be combined with $\Delta_i = G_i$

A crash introduction to Newton Raphson

Newton-Raphson's programming:

1. Start from a given θ_0
2. Compute G_0 and H_0
3. Compute $\lambda_0 = 1$ and $\Delta_1 = -H_0^{-1} G_0$
4. obtain θ_1
5. test that $||G|| < \epsilon$, if not start again with θ_1 as a starting point.

The computation of H can be tedious, even numerically. Trick of the day: use the BHHH approximation:

$$-H = GG^\top$$

, as

$$-\mathbb{E}\left[\partial_\theta^2 LL()\right] =$$

The sequence becomes

$$\theta_{i+1} = \theta_i + (GG^\top)^{-1} G$$