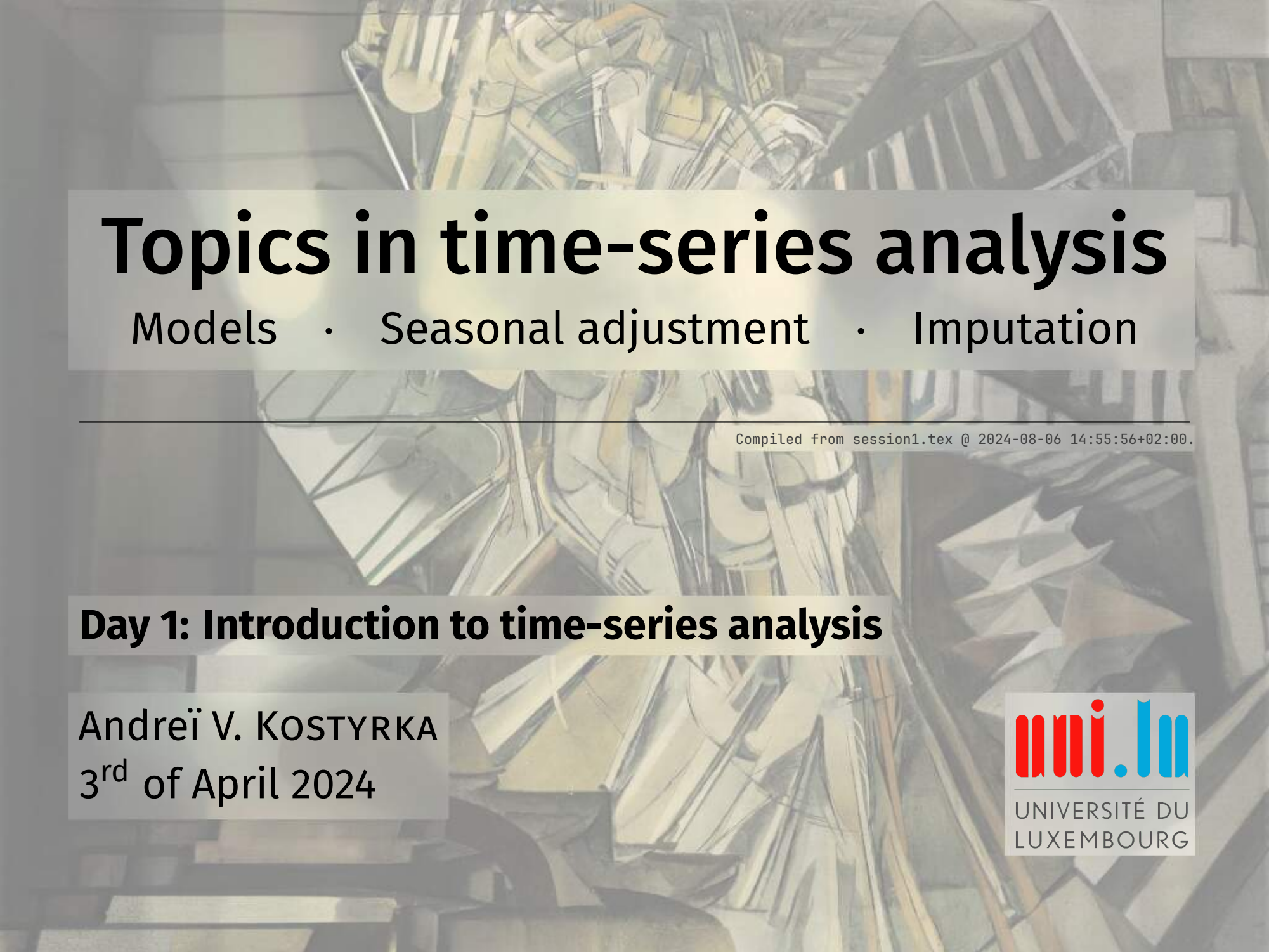


Topics in time-series analysis

Models · Seasonal adjustment · Imputation

Compiled from session1.tex @ 2024-08-06 14:55:56+02:00.

Day 1: Introduction to time-series analysis

Andreï V. KOSTYRKA

3rd of April 2024

UNIVERSITÉ DU LUXEMBOURG

Presentation structure

1. Administrative formalities
2. Stationary and non-stationary processes
3. Statistical theory for TSA
4. Linear time-series model estimation and inference

Administrative formalities

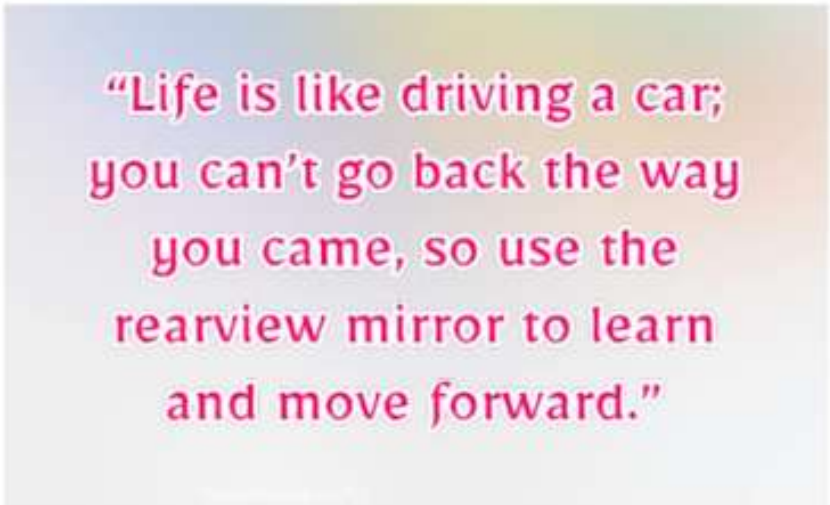
Why this course exists

- The demand for time-series analysis is going up (cross-country analysis, short-term forecasting in hundreds of economy sectors)
  - Yet, data collection of many indicators started only recently
  - For the crucial macro-economic variables, monthly or quarterly data is the best that one can get
- There are many doctoral courses dedicated to cross-sectional data analysis, but few on time series
  - No empirical time-series modelling course ⇒ fills the gap
- Promotion of transparent and harmonised data analysis (required by most EU institutions) and reproducible research at Uni.LU
  - Sometimes, the data are bad, and one has to do something at gunpoint and be honest about the limitations

Uncertainty is scary

- Humans hate uncertainty; it is quite scary when everything is uncertain, when anything can present danger
- Economists want to predict the future as accurately as possible
- When working with time series, one has to use only the past to make decisions at the present moment about the future

Motivation for time-series analysis



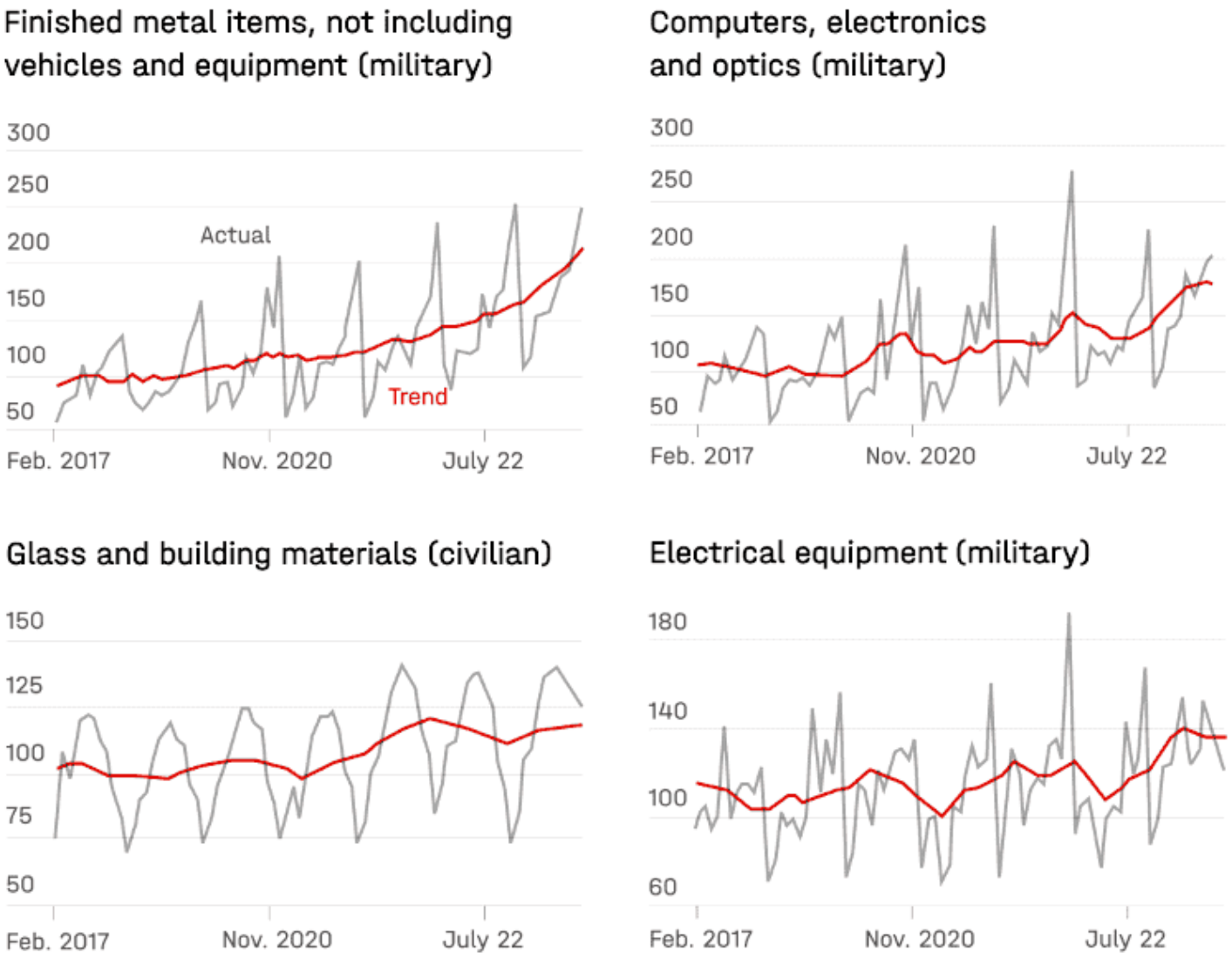
How this course was shaped

- I was working at STATEC / Hendyplan in 2022–2023
  - Seasonal adjustment, imputation, and forecasting were my daily bread
- The SA part of this course is an extension of what I had taught at workshops
- There is a high demand for harmonised time series / panels in the EU
  - Eurostat-endorsed software for processing of multi-country time series is much more feature-rich and user-friendly than the U. S. Census software

Course goals

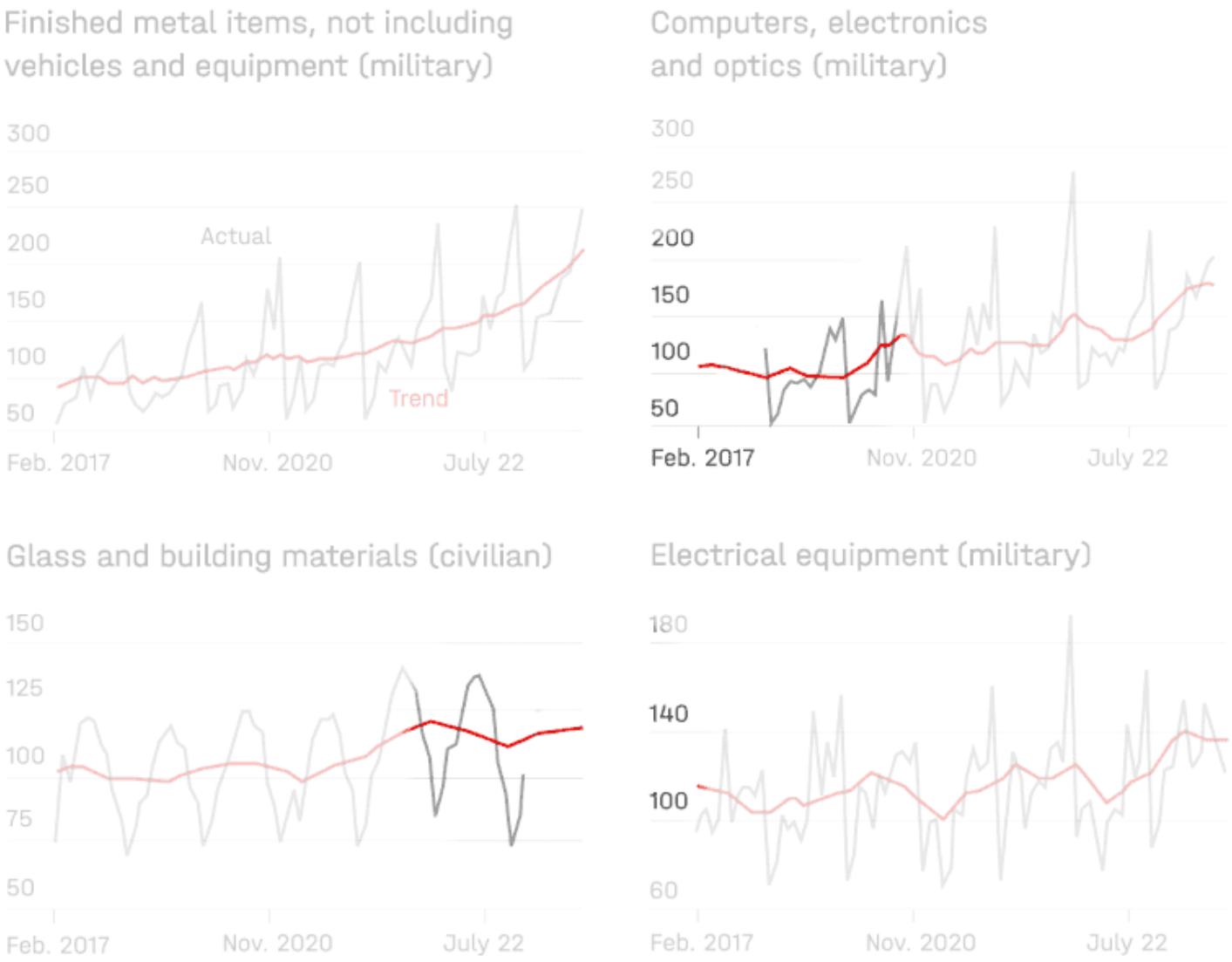
- Estimate several common time-series model specifications, including seasonal ones
  - Produce, benchmark, and back-test multiple forecasts
- Remove the seasonal component via semi-parametric and parametric methods
  - Diagnose the adequacy of the model and adjustment
- Depending on the time / your preferences:
  1. State-space / dynamic factor models and dimensionality reduction
  2. Reconstruction and imputation of multiple times series

Example: you will be able to do this



Credit: Centre for Macroeconomic Analysis and Short-Term Forecasting.

...and hopefully (if we have time) this



Credit: Centre for Macroeconomic Analysis and Short-Term Forecasting.

Intended schedule

Hard must:

1. Basic linear time-series models: theory and practice in R
2. Parametric and non-parametric seasonal adjustment and quality assessment in JDemetra+
3. Model-free and model-based multiple imputation techniques for stationary data in R

Stretch goals:

- Dimensionality reduction: dynamic factor models, principal component analysis
- Combining forecasts and imputations

Open-source packages for TSA

- JDemetra+
  - Java-based, extensible, recommended by Eurostat for SA
- R
  - Amazing capabilities for academic research
  - ML = maximum likelihood (mostly)
- Python
  - Huge choice of packages for data miners and big-data analysts in the industry
  - ML = machine learning (although there is no machine learning when the sample size is $T = 25$ for Luxembourg!)
- Julia
  - Very fast, but not so many packages for custom models

We shall be using R

Commercial statistical packages for TSA

- Matlab
  - Good for matrix operations and certain parametric models, inconvenient debugging, inflexible in terms of data structures
- Stata
  - Good for cross-sectional and panel analysis, poor time-series capabilities, very limited programming features
- EViews
  - OK for specifying state-space models, irredeemably outdated, feature-poor, lacks adequate numerical methods, no active community, failing statistical methods
- SPSS, SAS, Gauss, Minitab
  - *De mortuis nil nisi bonum*

What this course is not

- Not econometrics
  - But we shall revise the relevant basics and essential statistics today
- Not quantitative finance, volatility forecasting, portfolio optimisation
  - Yet provides certain relevant knowledge
- Not general signal processing
- Not modelling extra-complicated time series
  - We discuss how to reversibly transform ‘irregular’ TS into ‘regular’ ones that are safe to analyse

What this course is

- Methods that are used at EU statistical bureaus
- Linear and generalised linear methods for TS model estimation
- Linear filters and additive / multiplicative decomposition techniques
- Convenient asymptotic approximations even when '$T \to \infty$' is questionable

During the course

- 5 days = 5 sessions
  - 10-minute break in the middle (grab a coffee)
  - Study at home, ask questions about unclear concepts during the sessions
- Having a laptop is completely optional (you can follow the screen), but carrying out empirical analysis at least once in your spare time is a must
  - The earlier you spot a problem (e. g. broken Java installation), the earlier I shall be able to help you

Syllabus

- Contains the intended agenda
- Contains links to openly published learning resources (books, online tutorials etc.)
  - Extra material will be provided on Moodle
  - Your suggestions for learning resources are welcome
- Contains homework descriptions and full final project proposals

Bibliography

- Looks scary because it is long
  - Everything is on Moodle now
- Contains sources in the order of relevance
- The bare minimum is:
  1. Any graduate-level treatment of linear time-series models (The legendary Box & Jenkins book (5^{th} ed.) is appropriate)
  2. Eurostat (2018) 'Handbook on Seasonal Adjustment'
  3. Little & Rubin (2019) 'Statistical analysis with missing data'

Grading

- 5 · 2% attendance + 30% small assignment + 60% project
- The assignment is short: carry out seasonal adjustment of 2 time series, comment on its adequacy (in a GUI-based programme with a mouse, no coding required!)
  - Each participant gets their own data set on Moodle
- Final project: **choose the task that is the most relevant for your research** or the one that uses the data set that you know well
  - 7 conceptually different tasks to choose from
  - You may reuse your existing material in the assignment

Technical requirements for all assignments

1. Submission: written report + script / source code applying the empirical methods and reproducing these results
   - I should be able to reproduce these results – attach the data if they come from elsewhere
   - Optional: plots or videos – ZIP everything together
2. Use **open-source** statistical packages: R and/or JDemetra+, Python, Julia (or other open-source ones)
3. Write the report in plain text / Markdown / LaTeX / Jupyter / knitr / Sweave
   - `.doc[x]` and `.odt` are **not accepted**

If something is missing from this tutorial

- ESS Guidelines on SA (2015 edition)
- European Statistical Training Programme
  - Introduction to SA and JDemetra+ – 2024
  - 11–13th of June, 2024, Cologne, Germany
  - Application deadline: 15th of April, 2024
- ESTP 2021 PowerPoint slides on SA
- Eurostat 2018 handbook on SA (very deep)

Notational conventions

- Capital italic Latin: scalar and vector random variables (X, Y); subscript in brackets = coordinate ($X^{(2)}$)
- Capital Roman Latin: non-random matrices (A, V)
- Greek: parameters and numeric constants (β, θ)
  - With subscript: true values (β_3, θ_0)
  - With diacritics: estimators ($\hat{\beta}_{LS}, \tilde{\theta}$), which are random variables, not constants
- Lowercase Latin: functions ($f_{X,Z}(u, v)$, $g(X, Y, \theta)$) or their arguments ($f(t)$), or unimportant constants ($c = 12$)
- A vector X is always a column vector, X' is a row ($'$ means transposition)
- Operators: $\mathbb{E}X$, $\operatorname{Var} X$, $\mathbb{E}(Y \mid X)$

Stationary and non-stationary processes

Time-series samples

The properties of random variables are studied and the models are estimated in **finite samples**, i. e. realisations of random variables collected in periods $t = 1, \dots, T$ (we consider **discrete time** only).

- $X_1, X_2, ..., X_t$ is a sequence of non-independent non-identically-distributed random variables
- The sequence $\{X_t\}_{t=1}^{T}$ can be treated as a multivariate random variable
- From the economic perspective, we call $X_1, X_2, \dots, X_t$ a **time series** (or process), where t is a discrete (at most countable) time indicator

Weak stationarity

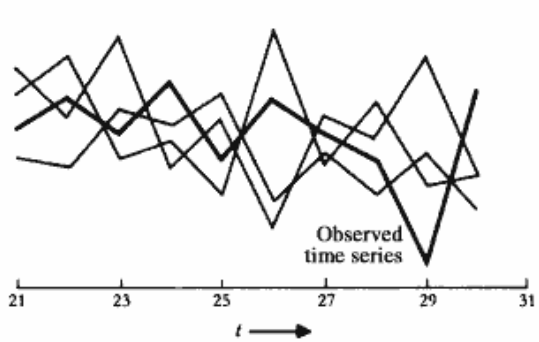
Statistical analysis is the easiest when certain aspects of these random variables are homogeneous: inference requires somewhat identically (or similarly) distributed RVs.

A weakly second-order-stationary process Y_t:

- $\forall t, s: \mathbb{E}Y_t = \mathbb{E}Y_s = \mu$
- $\forall t, s: \operatorname{Var} Y_t = \operatorname{Var} Y_s = \sigma^2$
- $\forall t, s: \operatorname{Cov}(Y_t, Y_s) = f(|t - s|)$, where f is some function

Ergodicity

Problems: only one realisation of a TS process (cf. cross-sections where one can collect more observations).

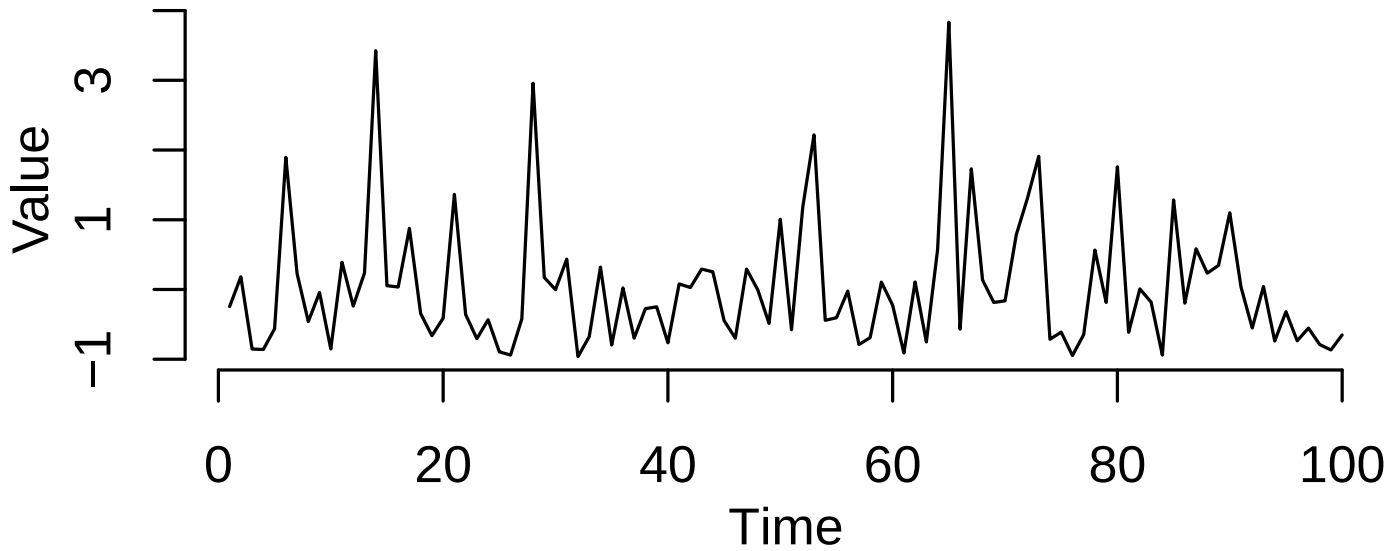


Credit: Box, Jenkins, Reinsel, Ljung (2015).

- WLLN: $n^{-1} \sum_{i=1}^{n} X_i \xrightarrow[n \to \infty]{\mathbb{P}} \mathbb{E}X$ for IID X_i
- In TSA, $\{X_1, \dots, X_T\} = \{X_t\}_{t=1}^{T}$ are one single realisation of the RV X, and (X_i, X_j) are not independent
  - In a parallel universe, many different realisations of $\{X_t\}_{t=1}^{T}$ could be collected
- Ergodicity: observing only one sequence of non-IID $\{X_t\}_{t=1}^{T}$ is sufficient for recovering the properties of X
- Sufficient conditions: $\mathrm{Cov}(X_t, X_{t-h}) \xrightarrow{h \to \infty} 0$ (mean), $\sum_{h=0}^{\infty} \left| \mathrm{Cov}(X_t, X_{t-h}) \right| < \infty$ (variance)

Example: white noise

Centred exponential distribution: $f_X(t) = \lambda \exp(-\lambda t - 1)$ for $t > -1/\lambda$. One realisation with with $\lambda = 1$:



$\mathbb{E}X_t = 0$, $\operatorname{Var} X_t = 1$, $\operatorname{Cov}(X_t, X_s) = 0$ for $t \neq s$.

Is X_t stationary? Is X_t symmetric around the mean?

Wold decomposition

Y_t is a weakly stationary process *if and only if* it can be represented in the following form:

$$Y_t = \mu_0 + \sum_{i=0}^{\infty} \psi_i U_{t-i} + V_t,$$

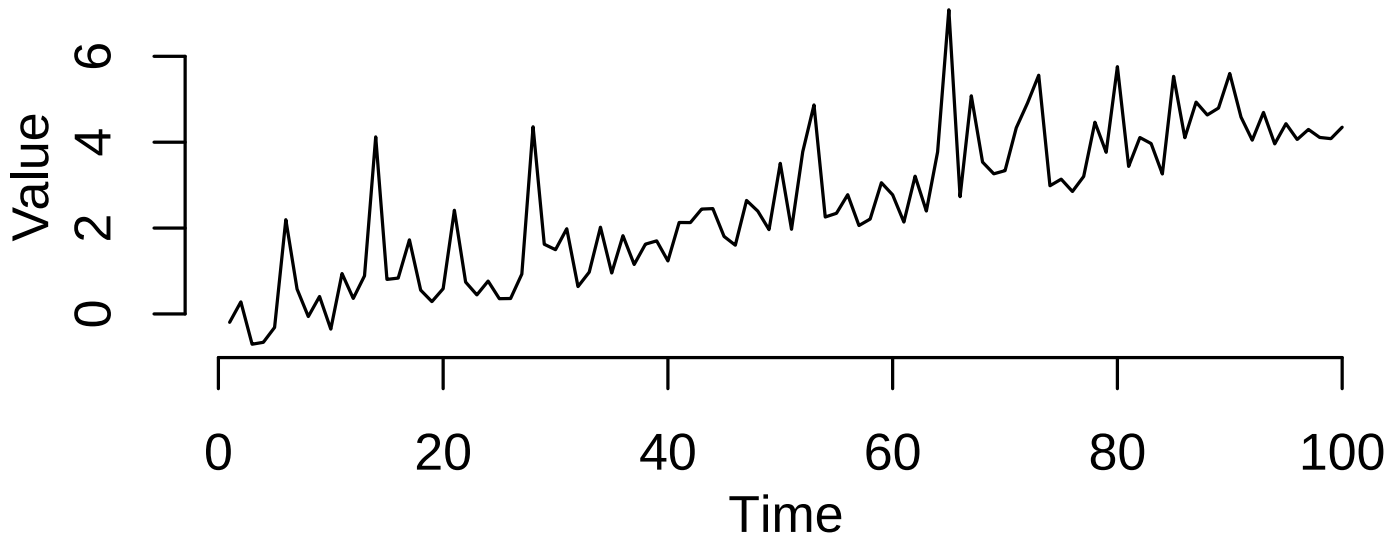
where U_{t-i} is the white-noise error, $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$, and V_t is deterministic and uncorrelated with U_s for all s (and some other technical conditions).

The error U_t is the linear forecast error based on all available information: $U_t \coloneqq Y_t - \text{BLP}(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots)$.

Example: mean non-stationarity (1/2)

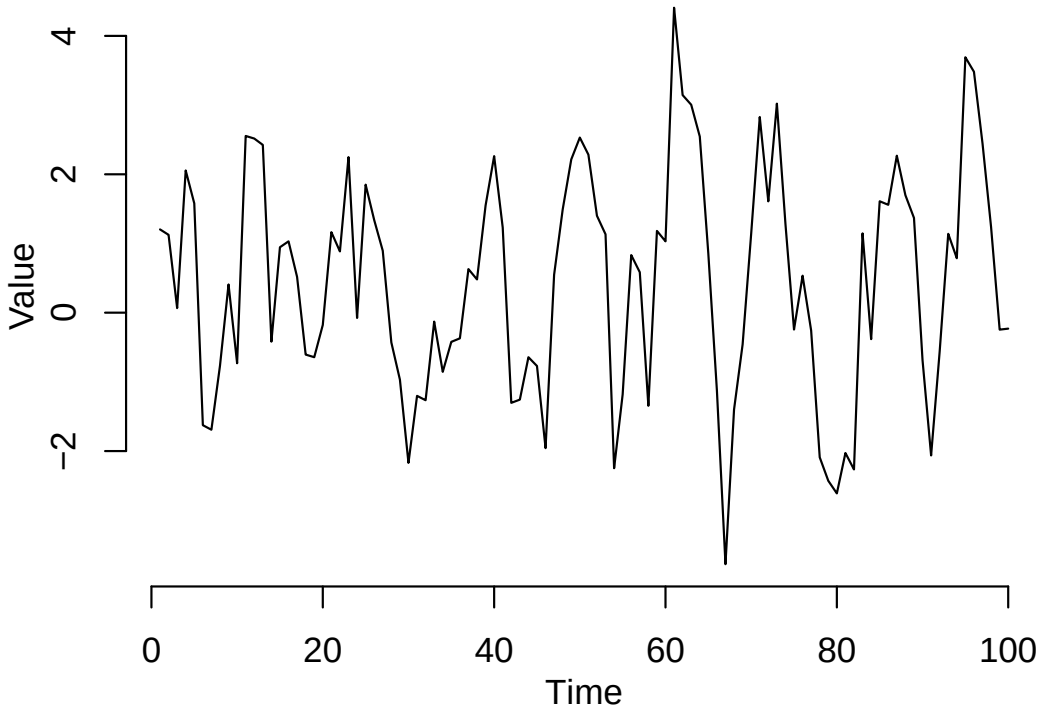
Suppose that Y_t is growing linearly in time:

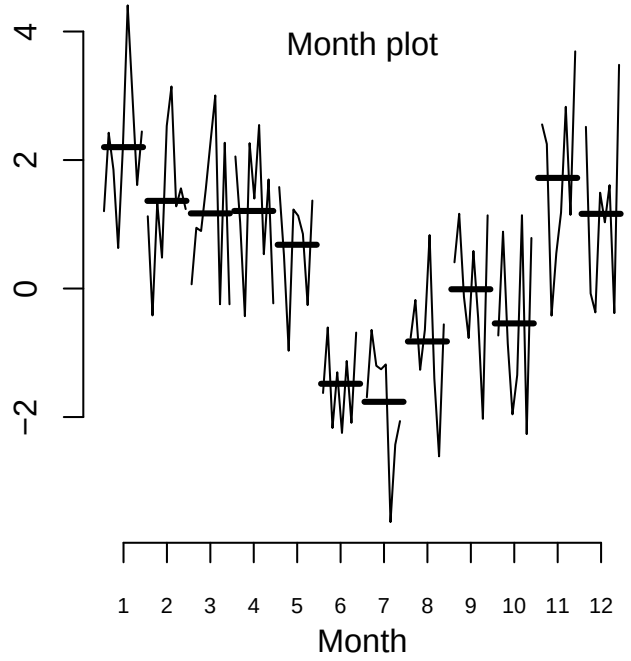
$$Y_t = 0.05t + U_t, \quad \mathbb{E}U_t = 0, \quad \operatorname{Var} U_t = \sigma^2$$



One should de-trend (= de-mean) this **trend-stationary** process (or model the mean tendency explicitly).

Example: mean non-stationarity (2/2)

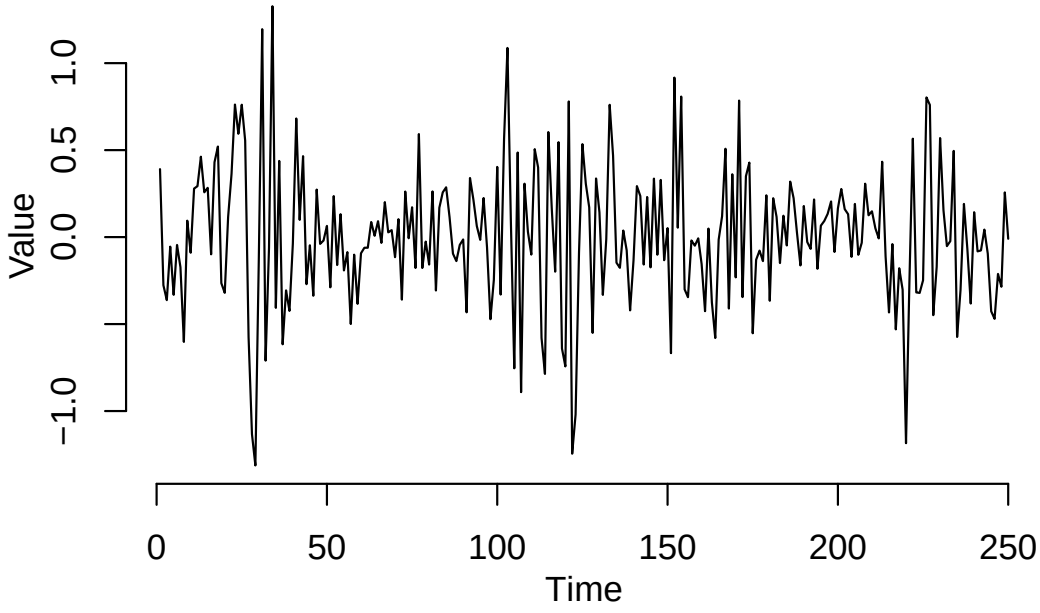


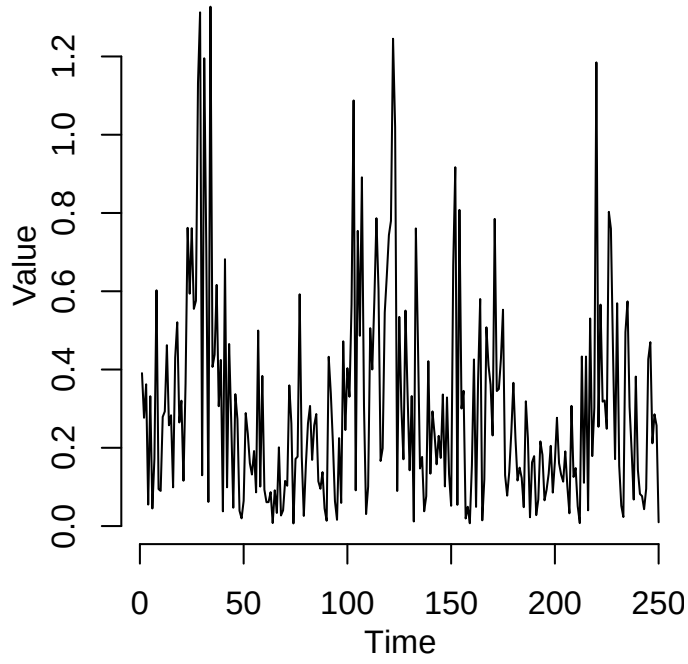


One should de-seasonalise (= seasonally adjust) this **seasonal** process (or model the seasonality explicitly).

Example above: electricity consumption monthly effects.

Example: variance non-stationarity



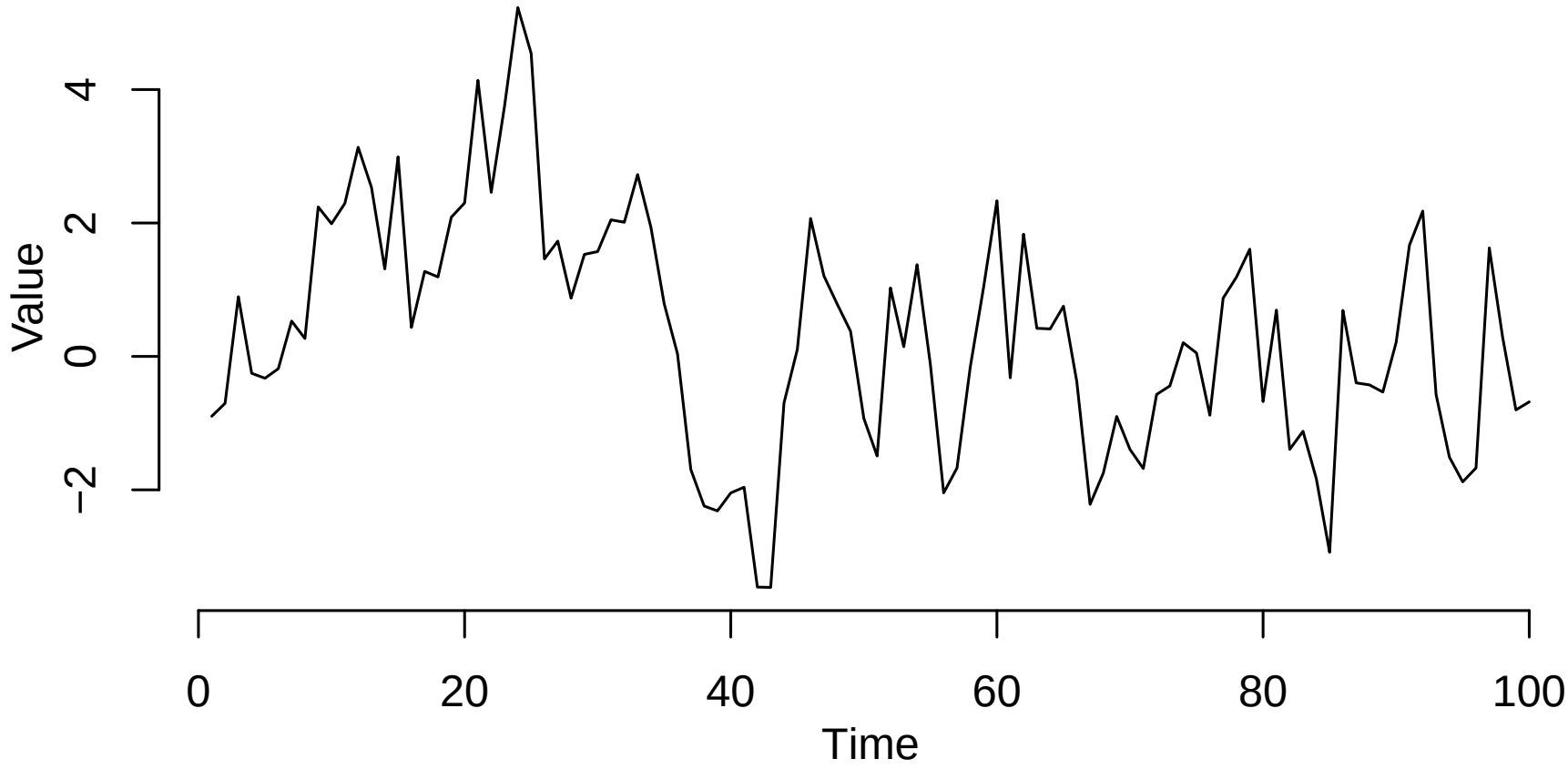


Suppose that Y_t behaves like market returns and exhibits volatility clustering:

$$Y_t = \sqrt{h_t} U_t, \quad h_t = 0.01 + 0.7 h_{t-1} + 0.25 Y_{t-1}$$

Conditional heteroskedasticity should be modelled.

Example: covariance non-stationarity



$\mathrm{Cov}(X_t, X_{t-h})$ must be a function of h only (not of t).

Eyeballing stationarity

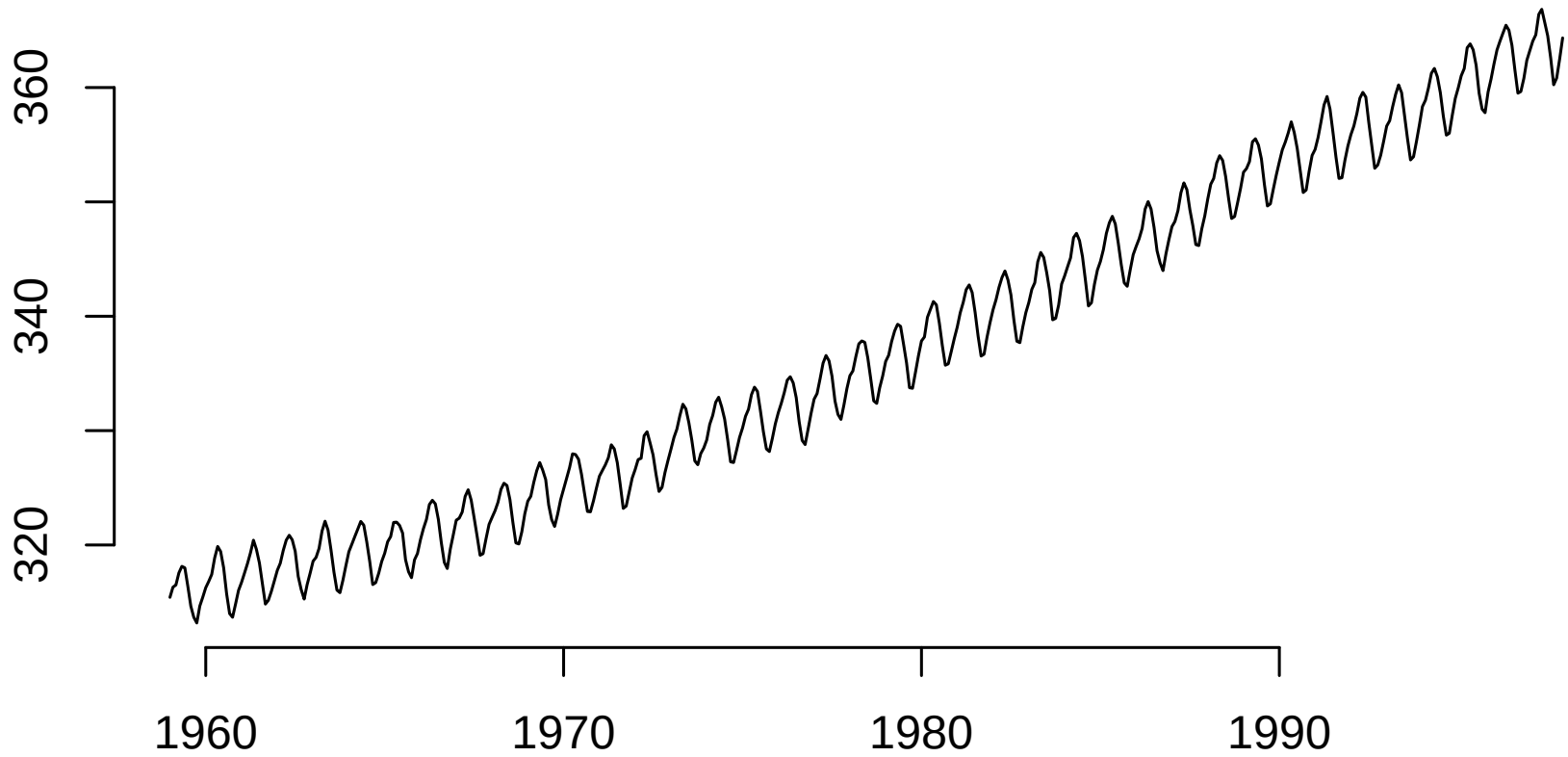
Always look at the input series before applying statistical techniques developed for *(put the type here)* processes!

- **Mean stationarity:** if the horizontal level of the series systematically drifts away vertically, the series is not mean-stationary
- **Variance stationarity:** if the 'tube width' around the series is variable (looks like pulsations), the series is not variance-stationary
- **Covariance stationarity:** if the frequency of mean crossings changes substantially in time, the series is not covariance-stationary

Rejections are more informative than non-rejections.

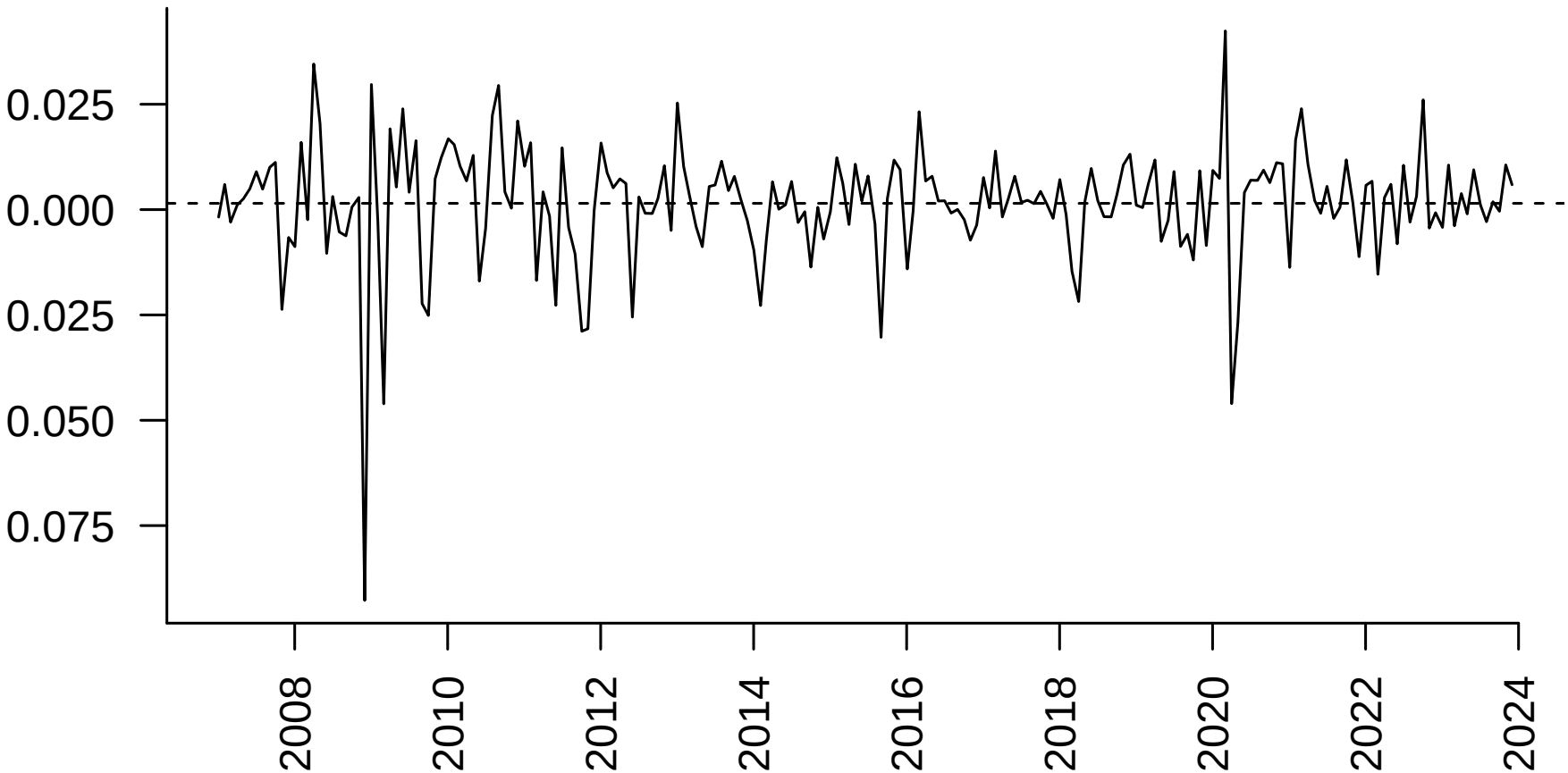
Example: CO_2 concentration (parts per mln)

Data set in R: `datasets::co2`



Example: S&P500 monthly returns

Data set in R: `tidyquant::tq_get("SPY")`



Statistical theory for TSA

Chaotic vs. stochastic models

Can someone explain the difference?

- Chaotic = deterministic (surprise!), complex, irregular, non-linear, exhibiting patterns (attractors)
  - Butterfly effect: sensitive to initial conditions
  - No randomness (Laplace's demon)
- Stochastic = random, linear or non-linear, simple or complex, regular or irregular, without any guarantee of patterns
  - The influence of the initial condition fades away

Parametric vs. semi-parametric models

- Fully parametric
- Fully non-parametric
- Partially parametric, partially non-parametric (semi-parametric)

Fully parametric model example:

$$Y = \alpha + X'\beta + U, \quad U \mid X \sim \mathcal{N}(0, \sigma^2)$$

Fully non-parametric model example:

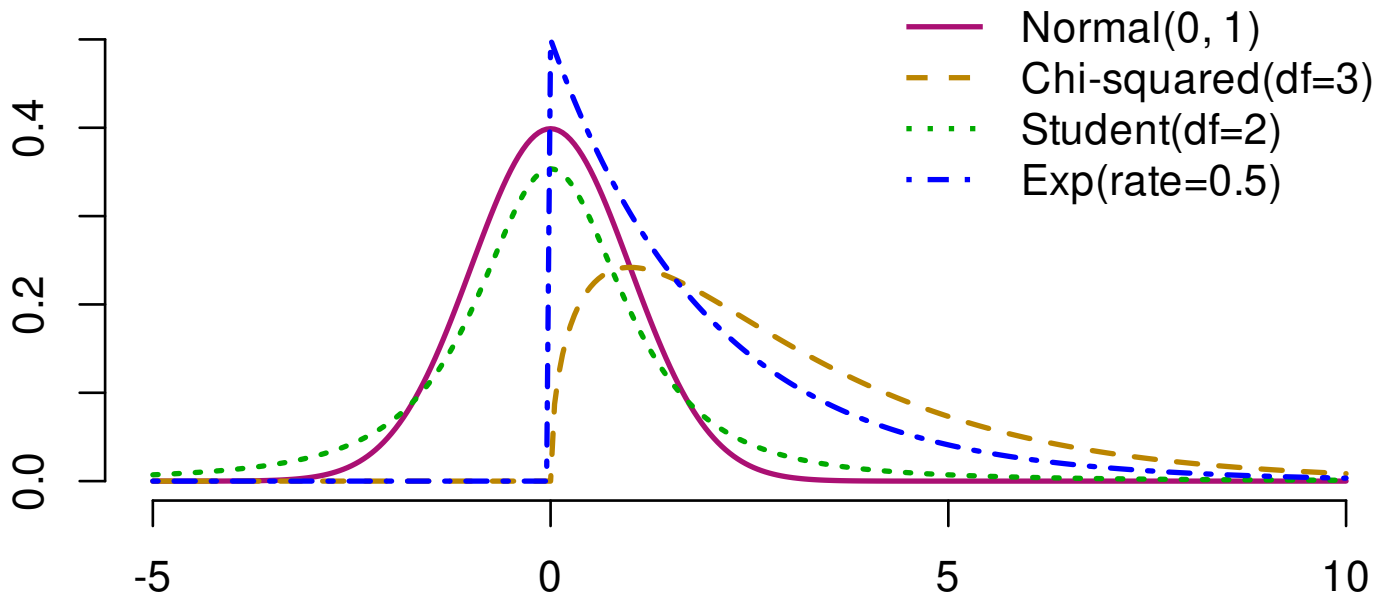
$$Y = f(X) + U, \quad \mathbb{E}(U \mid X) = 0, \quad f \text{ is unknown}$$

Semi-parametric model example:

$$Y = \alpha + X'\beta + U, \quad \mathbb{E}(U \mid X) = 0$$

Parametric density

We may specify the conditional density of the model error given the observables (Y, X, θ): 'Everything that the model does not capture is, e. g., Gaussian!'



- Simplifies estimation and inference
- A huge leap of faith (but sometimes, the necessary evil)

Density function

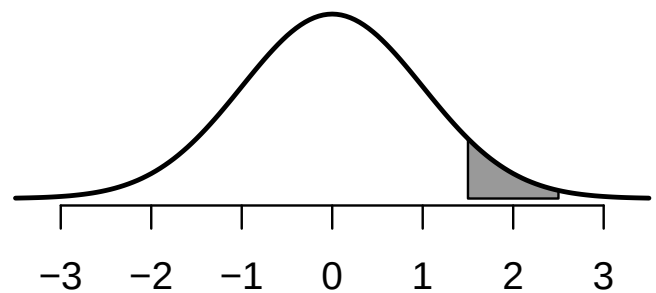
Probability density function (PDF) of RV X: such a non-negative function $f_X(t)$ that its integral over any area A yields the probability that $X \in A$.

Consequence: $\int_{\mathbb{R}^{\dim X}} f_X(t)\,\mathrm{d}t = 1$.

Univariate Gaussian (normal) density:

$$f_{\mathcal{N}(\mu,\sigma^2)}(t) \coloneqq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

$\mathbb{P}(\text{standard Gaussian} \in [1.5, 2.5])$:
$\int_{1.5}^{2.5} f_{\mathcal{N}(0,1)}(t) \approx 6.06\%$



Moments of random variables

m^{th} moment = the mean of the m^{th} power of X

$$\mu_m \coloneqq \mathbb{E}X^m = \int_{\mathbb{R}} t^m f_X(t)\,\mathrm{d}t$$

- $\mu_1 = \mathbb{E}X$, mean, is the first moment
- $\mu_2 = \mathbb{E}X^2$
  - $\mathbb{E}X^2 - \mu_1^2 = \operatorname{Var} X = \sigma^2 = \mathbb{E}(X - \mu_1)^2$, i. e. variance is the second central moment
  - Cross-moments: $\operatorname{Cov}(X, Y) \coloneqq \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$
- Skewness: $\mathbb{E}\left(\frac{X-\mu_1}{\sigma}\right)^3$, kurtosis: $\mathbb{E}\left(\frac{X-\mu_1}{\sigma}\right)^4$

Moments can be infinite:
for a Student(ν) RV, $\mathbb{E}|X|^k < \infty$ for $\nu > k$.

Auto-covariance function

For weakly stationary processes,

$$\gamma(h) \coloneqq \mathrm{Cov}(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)]$$

is the auto-covariance function.

- $\gamma(h) = \gamma(-h)$
- $\gamma(0) \geq \gamma(h)$

Auto-correlation function

Pearson's linear correlation:

$$\rho(X, Y) \coloneqq \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}\, X \cdot \mathrm{Var}\, Y}}$$

Auto-correlation:

$$\rho(h) \coloneqq \mathrm{Cor}(X_t, X_{t+h}) = \frac{\mathrm{Cov}(X_t, X_{t+h})}{\sqrt{\mathrm{Var}\, X_t \cdot \mathrm{Var}\, X_{t+h}}} = \frac{\gamma(h)}{\gamma(0)}$$

Remember: $-1 \leq \rho(h) \leq 1$.

Method of moments (MM)

Under certain regularity conditions (necessary for the WLLN and CLT), population parameters can be estimated by sample averages.

Replace integrals in expectations with sums:

- $\mathbb{E}X = \mu \iff \int_{\mathbb{R}} (t-\mu) f_X(t)\,\mathrm{d}t = 0$ becomes
  $\frac{1}{T}\sum_{t=1}^{T}(X_t - \mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t$
- $\operatorname{Var} X = \sigma^2 \iff \int_{\mathbb{R}} (t-\mu)^2 f_X(t)\,\mathrm{d}t = \sigma^2$ becomes
  $\frac{1}{T}\sum_{t=1}^{T}(X_t - \mu)^2 = \sigma^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(X_t - \hat{\mu})^2$

Mean estimation

We are interested in the first moment (theoretical / population mean) of X, $\mathbb{E}X$.

We collected 5 observations $\{X_t\}_{t=1}^{T} = \{15, 22, 13, 11, 24\}$.

Then, $\hat{\mu} = \frac{1}{5}\sum_{t=1}^{5} X_t = \frac{85}{5} = 17$ is the MM estimate of the mean.

This estimator can be plugged in wherever the explicit expression for the theoretical mean appears.

Variance estimation

$\{X_t\}_{t=1}^T = \{15, 22, 13, 11, 24\}$, $\hat{\mu} = 17$.

Since $\hat{\mu}$ has already been estimated independently through the first moment equation in the MM approach,

$$\mathbb{E}(X - \mathbb{E}X)^2 = \sigma^2$$

becomes in finite samples

$$\hat{\sigma}^2 = \frac{1}{5}\sum_t (X_t - \hat{\mu})^2 = \frac{130}{5} = 26$$

Properties of (co-)variances

- For vector random variables,
  $\operatorname{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)'$
- $\operatorname{Cov}(X, X) = \operatorname{Var} X$
- $\operatorname{Var}(X \pm Y) = \operatorname{Var} X + \operatorname{Var} Y \pm \operatorname{Cov}(X, Y) \pm \operatorname{Cov}(Y, X)$
- $\operatorname{Var} \alpha X = \alpha^2 \operatorname{Var} X$
- $\operatorname{Var} AX = A(\operatorname{Var} X)A'$

Variance of the sample average

$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t$, the sample average is an estimator of the mean.

$$\operatorname{Var}\hat{\mu} = \operatorname{Var}\frac{1}{T}\sum_{t=1}^{T} X_t = \frac{1}{T^2}\sum_{t=1}^{T}\operatorname{Var} X_t + \frac{2}{T^2}\sum_{s<t}\operatorname{Cov}(X_t, X_s)$$

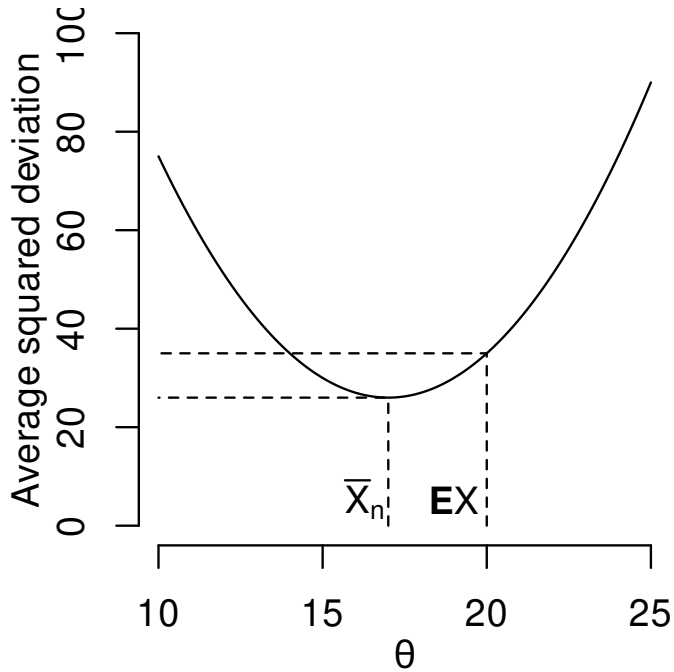
- IID: $\operatorname{Var} X_t = \sigma^2$, $\operatorname{Cov}(X_t, X_s) = 0$ for $t \neq s \Rightarrow \widehat{\operatorname{Var}}\,\hat{\mu} = \frac{\hat{\sigma}^2}{T}$
- INID: $\operatorname{Var} X_t = \sigma_t^2$, hence, one estimator of $\operatorname{Var}\hat{\mu}$ is $\frac{1}{T^2}\sum_{t=1}^{T}(X_t - \hat{\mu})^2 = \frac{1}{T}\widehat{\operatorname{Var}}_{\text{MM}} X_t = \frac{\hat{\sigma}^2}{t}$ (NB: σ^2 is purely cerebral)
- NINID: cleverly estimate the non-zero covariances by imposing a structure
  - Cross-sectional ⇒ spatial correlations with block matrices
  - Temporal ⇒ auto-correlations with spectral methods

This forms the basis of **consistent** VCOV estimation!

Higher moment estimation pitfalls

Estimation of 2nd and higher moments is **tricky**, and MM / ML can yield consistent yet **biased** estimators that perform poorly in small samples.

Example. The MM variance estimator, the average squared deviation from a constant, $\frac{1}{T}\sum_t (X_t - \theta)^2$, is always minimised (w. r. t. θ) at $\hat{\theta} = \bar{X}_T$.
The true Var X, i. e. the expected squared deviation from the true $\mathbb{E}X$, is higher than the MM estimate with probability 1!



Finite-sample calibration

Uncertainty is always under-estimated, typically by some factor of order T^{-1} $\Rightarrow$ some form of finite-sample calibration is recommended.

- Use **conservative** variance estimators that are not biased downwards
  - Use $\frac{T}{T-1} \cdot \hat{\sigma}^2_{\text{MM}} = \frac{1}{T-1} \sum_t (X_t - \hat{\mu})^2$ for $\widehat{\text{Var}}\, X$
- Use more conservative critical values for inference
  - In the past, researchers would use Student-t or Fisher's F critical values (distributions with fatter tails)
  - These days, researchers prefer Bartlett correction and/or bootstrap calibration and/or empirical-likelihood-based inference

Covariance estimation

$$\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \quad \Rightarrow \quad \text{use } \frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X}_T)(Y_t - \bar{Y}_T)$$

Auto-covariance $\mathrm{Cov}(X_t, X_{t-h})$ MM estimator:

$$\hat{\gamma}(h) = \frac{1}{T-h}\sum_{t=h+1}^{T}(X_t - \bar{X}_T)(X_{t-h} - \bar{X}_T)$$

Auto-correlation $\mathrm{Cor}(X_t, X_{t-h})$ MM estimator:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{(T-h)^{-1}\sum_{t=h+1}^{T}(X_t - \bar{X}_T)(X_{t-h} - \bar{X}_T)}{T^{-1}\sum_{t=1}^{T}(X_t - \bar{X}_T)^2}$$

NB. Multiple finite-sample alternative estimators are possible, e. g., the MM estimator of $\mathbb{E}X_{t-h}$ may use only the last $T-h$ points, or the $\frac{T}{T-1}$ factor can be added to $\hat{\gamma}(0)$.

Linear models

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + U$$

Compact notation:

$$Y = \tilde{X}'\theta + U, \quad \tilde{X} \coloneqq (1 \; X_1 \; \ldots \; X_k)', \quad \theta \coloneqq (\alpha \; \beta_1 \; \ldots \; \beta_k)'$$

Exogeneity assumption: $\mathbb{E}(U \mid X) = 0$.

No parametric assumption about the conditional distribution $f_{Y|X,\theta} \Rightarrow$ this model is **semi-parametric**.

If we specify that $Y \mid X \sim \mathcal{N}(\tilde{X}'\theta, v(X))$ for a specified $v(\cdot) > 0$, or $\frac{Y - \tilde{X}'\theta}{\sigma} \mid X \sim t_5$, then, this specification becomes **parametric**.

Causal interpretation

$$FUELSALES = \beta_0 + \beta_1 P_{Lux} + \beta_2 P_{abroad} + \beta_3 COMMUTERS + \beta_4 COVID + U$$

- $\frac{\partial}{\partial P_{abroad}} FUELSALES = \beta_2 + \frac{\partial}{\partial P_{abroad}} U$
- Causal interpretation: one variable changes, all others (including the error!) remain constant
  - If the foreign fuel price changes by 1 Euro, fuel sales will change by β_2 units *ceteris paribus*
  - $\frac{\partial}{\partial P_{abroad}} U = 0$ is our exogeneity assumption

Conditional density

Suppose that in a model, we *know* the joint distribution of (Y, X). For simplicity, assume continuous distributions $f_{Y,X}(y, x)$.

Conditional density – as if X were not random, i. e. took a specific value:

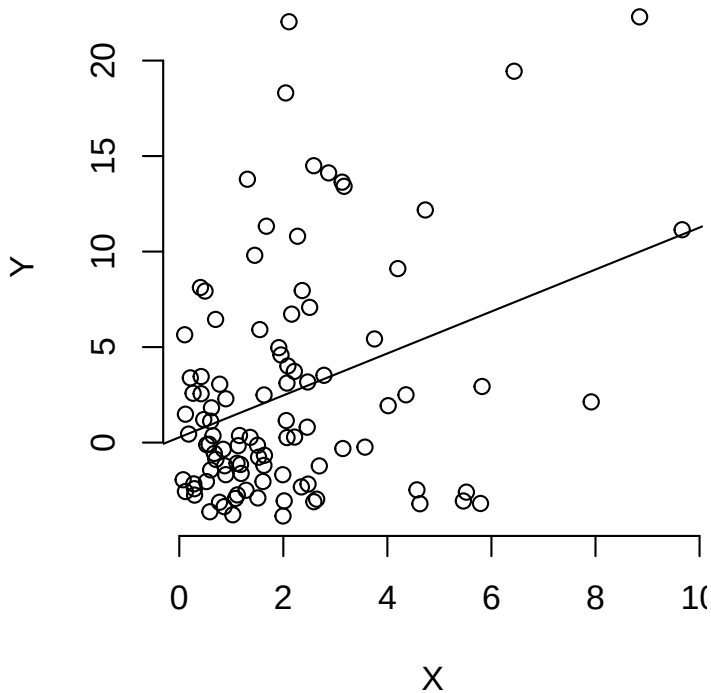
$$f_{Y|X=x}(y) = \frac{f_{Y,X}(y, x)}{f_X(x)}$$

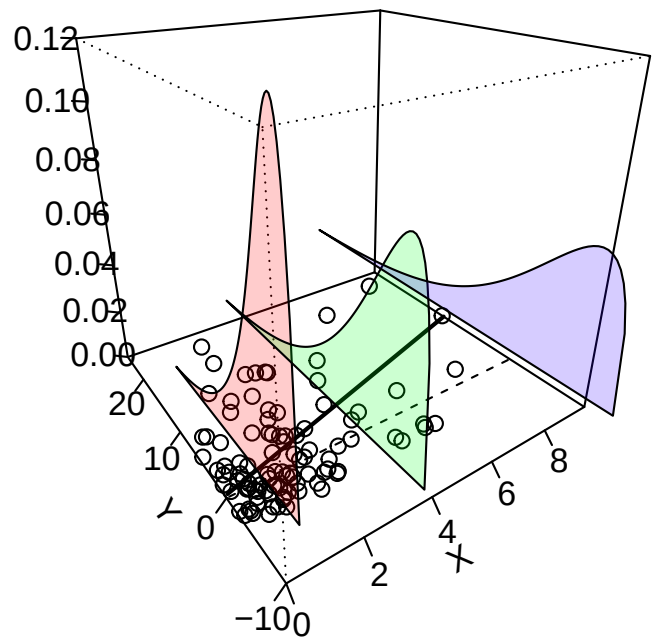
Sometimes, we may not know $f_{Y,X}$ at all – but we can assume a specific conditional distribution $f_{Y|X=x}(y)$ directly.

Conditional density example

Linear model with zero-mean heteroskedastic error:
$\mathbb{E}(U \mid X) = 0$, $\mathbb{E}U = 0$, $\mathrm{Var}(U \mid X) = 4(X+1)^2$:

$$Y = 1 + 1 \cdot X + U, \quad X \sim \exp(0.5), \quad \frac{U}{X+1} + 3 \sim \chi^2_3$$





Conditional distribution: $Y \mid X, \theta \sim \frac{Y-\theta^{(1)}-\theta^{(2)}X}{X+1} + 3 \sim \chi^2_3 \big|_{\theta=\theta_0}$.

Does the error distribution matter?

Consider a general non-linear model with additive errors:

$$Y = h(X, \theta_0) + U$$

- In some models, $\mathbb{E}(U \mid X) = 0$ alone is sufficient to allow the estimation of parameters
- In some models (especially non-linear ones with a limited dependent variable), knowledge of the conditional density $\text{PDF}_{Y|X}$ is required
  - It is unknown, but the researcher can assume $f_{Y|X,\theta}$
  - The marginal distribution of X is uninformative about θ
- Sometimes, the results coincide; sometimes, they do not

Linear time-series model estimation and inference

Estimation paradigms

1. Maximise some goodness-of-fit measure
2. Minimise some discrepancy measure

However, all of them can be characterised as minimum-distance methods:

- OLS: minimise the Euclidean norm of the residual vector
- GMM: minimise the distance between the average moment function and zero
- ML: minimise the Kullback–Leibler divergence

Approaches may turn out to be equivalent: in a linear model, minimising the sum of squared residuals = maximising the Gaussian likelihood = maximising R^2.

Estimators used in applied economics

Since everything is a minimum-distance estimator, consider the problem of 'solving' some useful economic model by 'finding' some 'optimal' parameters θ using data $\{Z_i\}_{i=1}^n$:

$$\hat{\theta} \coloneqq \arg\min_{\theta} s\Big(\ell(Z_1, \theta), \dots, \ell(Z_n, \theta)\Big)$$

- ℓ: loss function (x^2 for OLS, $|x|$ for LAD, minus likelihood for ML, high-breakdown-point losses...)
- s: aggregating statistic (average, trimmed average, weighted average, median, quantile ...)

OLS: minimising the sum = minimising the average because $\arg\min_\theta \sum_i (Y_i - X_i'\theta)^2 = \arg\min_\theta \frac{1}{n}\sum (Y_i - X_i'\theta)^2$.

OLS estimator of linear model parameters

Suppose that the researcher collected a random sample of cross-sectional data, i. e. IID observations $\{(Y_i, X_i)\}_{i=1}^n$.

$$Y_i = \tilde{X}_i' \theta_0 + U_i, \quad i = 1, \dots, n$$

Then, the ordinary least-squares (OLS) estimator is:

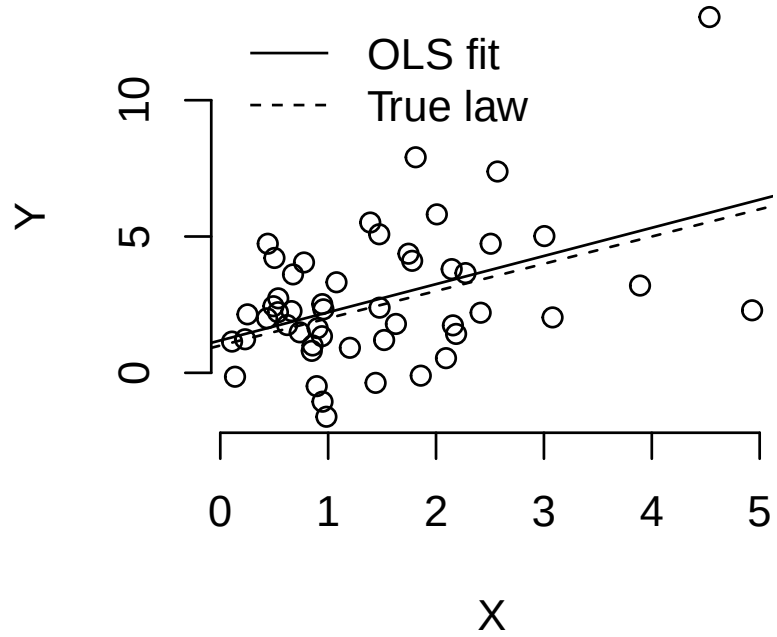
$$\hat{\theta}_{\text{LS}} \coloneqq \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{X}_i'\right)^{-1}}_{\text{MM estimator of } \mathbb{E}\tilde{X}\tilde{X}'} \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i Y\right)}_{\text{MM estimator of } \mathbb{E}\tilde{X}Y},$$

where $\theta_0 \coloneqq (\mathbb{E}\tilde{X}\tilde{X}')^{-1}\mathbb{E}\tilde{X}Y$.

OLS intuition

$$\hat{\theta}_{\text{LS}} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \tilde{X}_i' \theta)^2$$

$$\hat{\theta}_{\text{LS}} = \arg\min_{\theta} \widehat{\text{Var}} (Y_i - \tilde{X}_i' \theta)$$



- The OLS estimator is the minimiser of the unconditional error variance
  - The variance of OLS residuals, $\hat{U} \coloneqq Y - \tilde{X}' \hat{\theta}_{\text{LS}}$, around the fitted hyper-plane is the smallest
- $\tilde{X}' \hat{\theta}_{\text{LS}}$ is the best linear predictor (BLP) of Y given X
  - $\tilde{X}' \hat{\theta}_{\text{LS}}$ is the projection of Y onto the linear space of X

OLS estimator variance (HC)

Randomness in $\hat{\theta}_{\text{LS}}$ = discrepancy between the means and sample averages: $\hat{\theta}_{\text{LS}} = \theta_0 + \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i U_i\right)$.

Without making any distributional assumptions, analyse $\operatorname{Var}\hat{\theta}_{\text{LS}} = \operatorname{Var}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i U_i\right)\right]$.

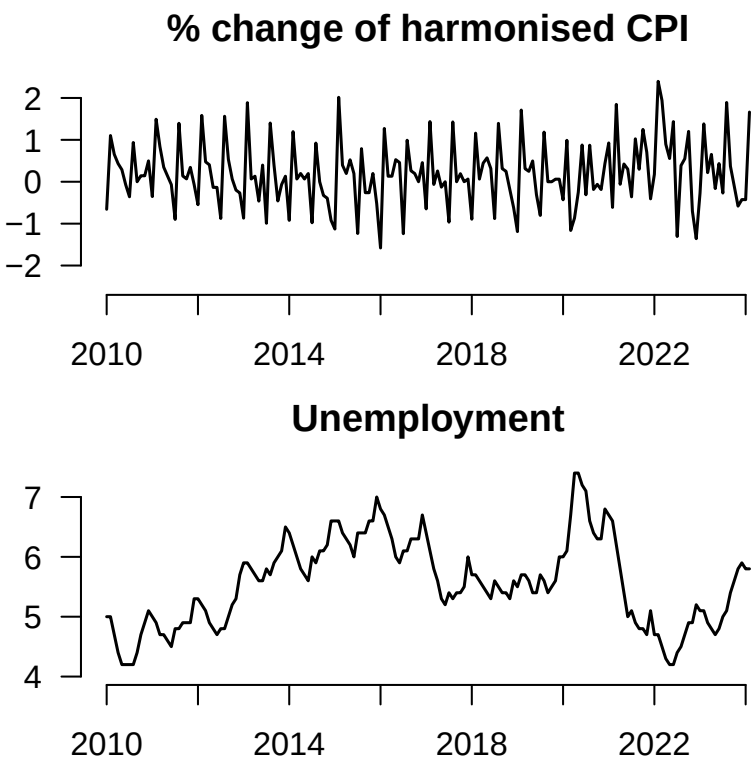
Applying $\operatorname{Var} AX = A(\operatorname{Var} X)A'$, using the IID assumption (covariances are zero $\Rightarrow \operatorname{Var}\sum = \sum \operatorname{Var}$) and $\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)'$:

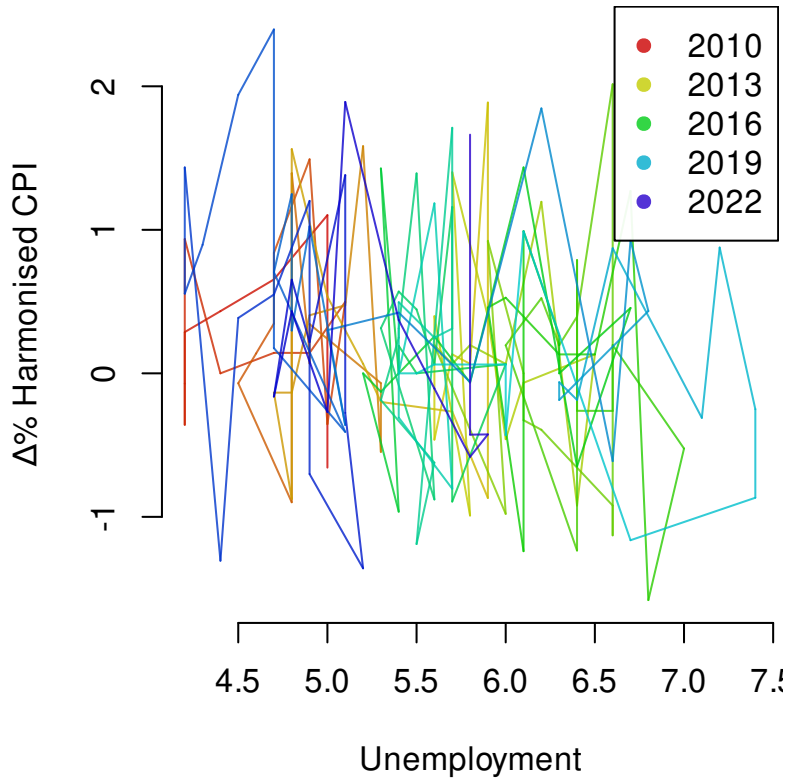
$$\operatorname{Var}\hat{\theta}_{\text{LS}} = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i' U_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\right)^{-1}$$

Replacing U_i with $\hat{U}_i$ yields $\widehat{\operatorname{Var}}\,\hat{\theta}_{\text{LS}}$, the famous White's heteroskedasticity-consistent VCOV estimator (HC_0).

Regression with time-series data

Task: estimate the average slope of the Phillips curve in Luxembourg for 2010–2023. Can we still do it via OLS?





Hint: does anything change in the algebra?

OLS with time-series data

There are observations $\{(Y_t, X_t)\}_{t=1}^T$.

$$Y_t = \tilde{X}_t'\theta + U_t, \quad t = 1, \dots, T$$

Then, the ordinary least-squares (OLS) estimator is:

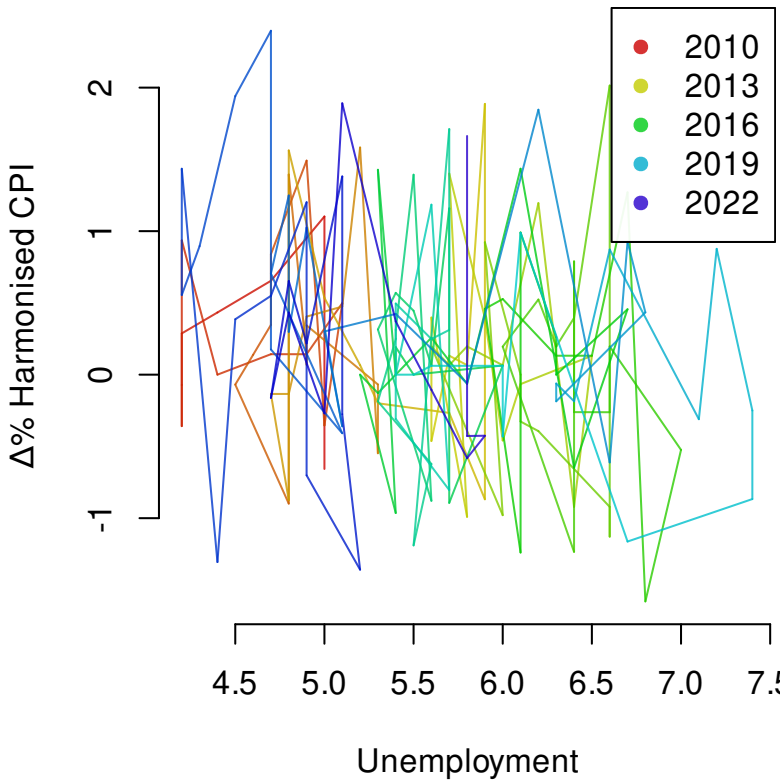
$$\hat{\theta}_{\text{LS}} := \left(\frac{1}{T}\sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \tilde{X}_t Y_t\right)$$

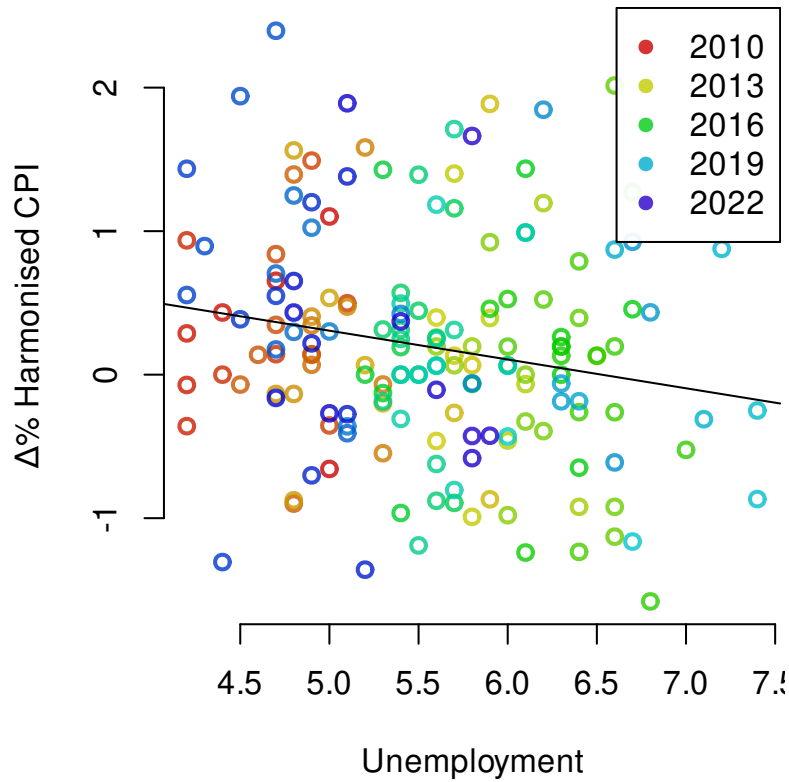
Is it consistent, though?

- As long as the model is correctly specified, $\mathbb{E}(U \mid X) = 0$, or *at least* $\mathbb{E}U = \mathbb{E}UX = 0$ (this is the first-order condition)

Example: Phillips curve

Task: estimate the average slope of the Phillips curve in Luxembourg for 2010–2023.





Same $\hat{\theta}_{LS}$ as if there were no time dependence.

OLS estimator variance complication

$$\operatorname{Var}\hat{\theta}_{\mathrm{LS}} = \operatorname{Var}\Big[\Big(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t'\Big)^{-1}\Big(\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\tilde{X}_t U_t\Big)\Big].$$

The complications for inference are due to the fact that $\operatorname{Var}\big(\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_t U_t\big)$ depends on **all** the covariances $(X_t U_t, X_s U_s) \quad \forall t, s = 1, \dots, T$!

Solution 1 (bad): since for stationary processes, the auto-covariance depends only on the lag, estimate all $\operatorname{Cov}(X_t\hat{U}_t, X_s\hat{U}_s)$ empirically and plug those covariances.

Why it is bad: in finite samples, it may not be positive semi-definite. In addition, it is very noisy, and there are T^2 unknowns!

OLS estimator variance (HAC)

Recall that for weakly stationary processes,
(1) $\operatorname{Cov}(A_t, A_{t+h}) = f(|h|)$ and (2) $f(|h|) \xrightarrow{h\to\infty} 0$.
Use (1) to reduce the number of estimands from T^2 to T and (2) to forcibly downweight / zero out long covariances.

Ideas (Newey & West, 1987):

1. $\operatorname{Var} \frac{1}{T}\sum_{t=1}^{T} A_t = \frac{1}{T^2}\sum_{t=1}^{T} \operatorname{Var} A_t + \frac{2}{T^2}\sum_{h=1}^{T-1}(T-h)\operatorname{Cov}(A_t, A_{t+h})$
2. Estimate only a fraction ~ $\sqrt[3]{T}$ of covariances
3. Add $w(|h|) \xrightarrow{h\to\infty} 0$

$$\widehat{\operatorname{Var}}\,\hat{\theta}_{\text{LS}} = \frac{1}{T}\Big(\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t'\Big)^{-1}\Big(\frac{1}{T}\sum_{t,s=1}^{T} w(|t-s|)\tilde{X}_t\tilde{X}_s'\hat{U}_t\hat{U}_s\Big)\Big(\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_t\tilde{X}_t'\Big)^{-1}$$

Pitfalls of HAC estimation

Popular kernel functions $w(|h|)$: Bartlett (triangular), quadratic spectral ($\sim \sin(h)/h^2$), truncated ($\mathbb{I}(h \leq h^*)$). Procedures for auto-choice of the scaling bandwidth are available in software (Andrews 1991, Newey & West 1994).

But! These estimators are much less reliable because they depend on many tuning parameters.

Discordant standard errors are to be expected even with 'nice' stationary input series.

Resampling (e. g. bootstrap) easily yields HC variances, but with dependent observations, HAC-consistent bootstrap also depends on tuning parameters.

Distributed lags

Can we add lags of X_t as explanatory variables?

$$Y_t = \beta_0 + \beta_1' X_t + \beta_2' X_{t-1} + U_t, \quad \mathbb{E}(U \mid X_t, X_{t-1}) = 0$$

Example: reversal effect if $\beta_1 = 0.7$, $\beta_2 = -0.2$ (the more one consumes today, the less they consume tomorrow).

Problem: if there is strong persistence in X_t, then, it is hard to distinguish the separate effects of X_t and X_{t-1}, which is called poor identification.

Solutions:

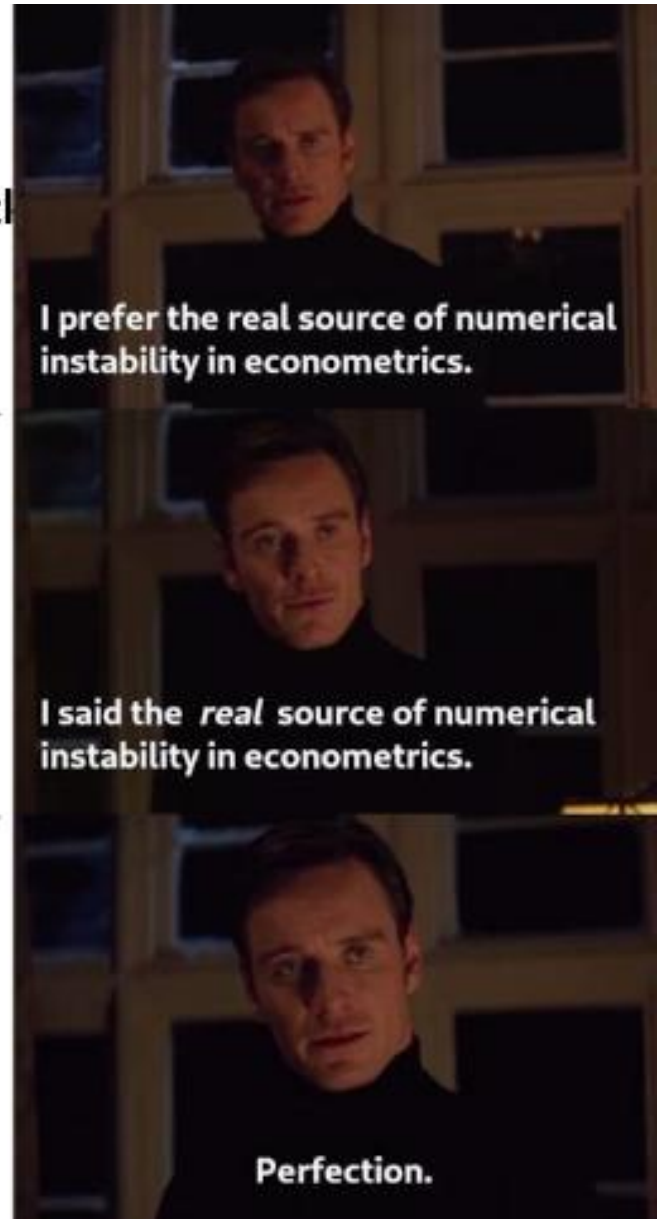
- Re-parametrise the model, carry out a linear transformation, use X_t and $X_t - X_{t-1}$ as regressors
- Put economic constraints (e. g. $|\beta_1| \geq |\beta_2|$, or $3 \cdot \beta_2 = \beta_1$)

Specification re-thinking

Quasi-multi-collinearfhtrtl

Weak instruments

Lack of identification

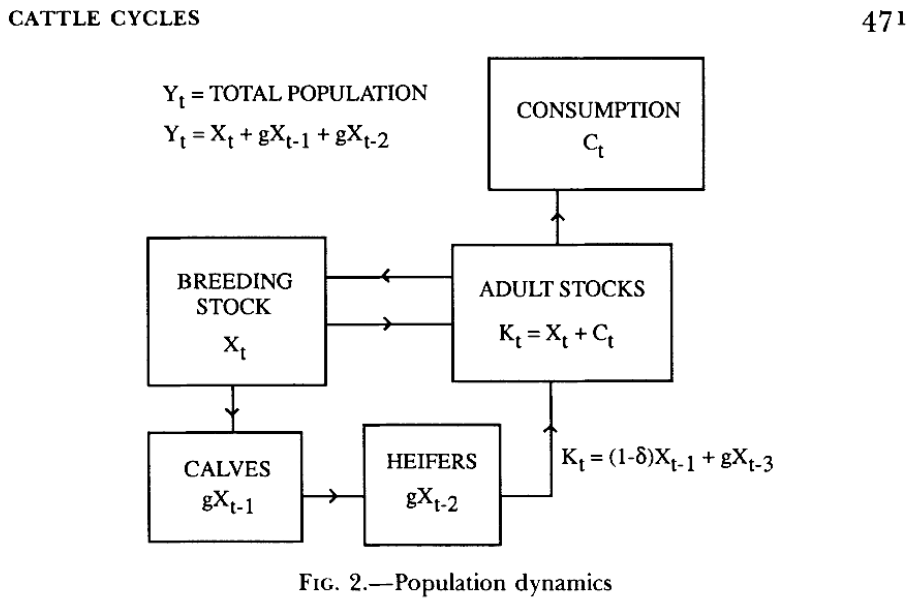


Multiple distributed lags

What if there are effects that are more distant in time?

$$Y_t = \beta_0 + \beta_1' X_t + \beta_3' X_{t-3} + \beta_4' X_{t-4} + U_t, \quad \mathbb{E}(U \mid X_t, \dots, X_{t-4}) = 0$$

Example: multi-year cattle cycles; 3–4 years for pork, 8–12 years for beef (Rosen, Murphy, Scheinkman, 1994, JPE). Same OLS algebra (assuming correct specification).



CATTLE CYCLES 471

FIG. 2.—Population dynamics

cattle are homogeneous and undifferentiated, independent of age or prior fertility; the slaughter age is exogenously fixed at 2 years;[3] and there are no interactions between trends and cycles.

Auto-regression

Can we add Y_t's own lags as explanatory variables?

AR(1) process:

$$Y_t = \mu_0 + \varphi_1 Y_{t-1} + U_t, \quad \mathbb{E}(U \mid Y_{t-1}) = 0$$

AR(p) process:

$$Y_t = \mu_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + U_t,$$
$$\mathbb{E}(U \mid Y_{t-1}, \ldots, Y_{t-p}) = 0$$

Stationary auto-regressive processes usually have **extremely short memory**. For AR(1),
$\mathrm{Cor}(Y_t, Y_{t-1}) = \mathrm{Cor}(\mu_0 + \varphi_1 Y_{t-1} + U_t, Y_{t-1}) = \varphi_1 \in (-1, 1)$
(exponential decay).

Auto-regression and distributed lag (ARDL)

We can combine two specifications: Y_t can be explained by its past and other contemporaneous and lagged variables.

$$Y_t = \mu_0 + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=0}^{q} X'_{t-j}\beta_j + U_t$$

$$\mathbb{E}(U \mid Y_{t-1,\dots,p}, X_{t-0,\dots,q}) = 0$$

How can such models be estimated? OLS (all these variables are observed). Define

$$\tilde{X} \coloneqq (1, Y_{t-1}, \dots, Y_{t-p}, X_t, X_{t-1}, \dots, X_{t-q})'$$

and apply standard OLS matrix algebra.

Moving-average processes

Recall the Wold decomposition for stationary processes:

$$Y_t = \mu_0 + \sum_{i=0}^{\infty} \psi_i U_{t-i} + V_t$$

Assume that only the immediate past is of interest – consider a **moving-average** process, MA(q):

$$Y_t = \mu_0 + U_t + \theta_1 U_{t-1} + \ldots + \theta_q U_{t-q} + \varepsilon_t$$

How can one estimate θ? **Not by OLS**, since U_t are unobserved! (We handle this case soon.)

Auto-regression and moving average (ARMA)

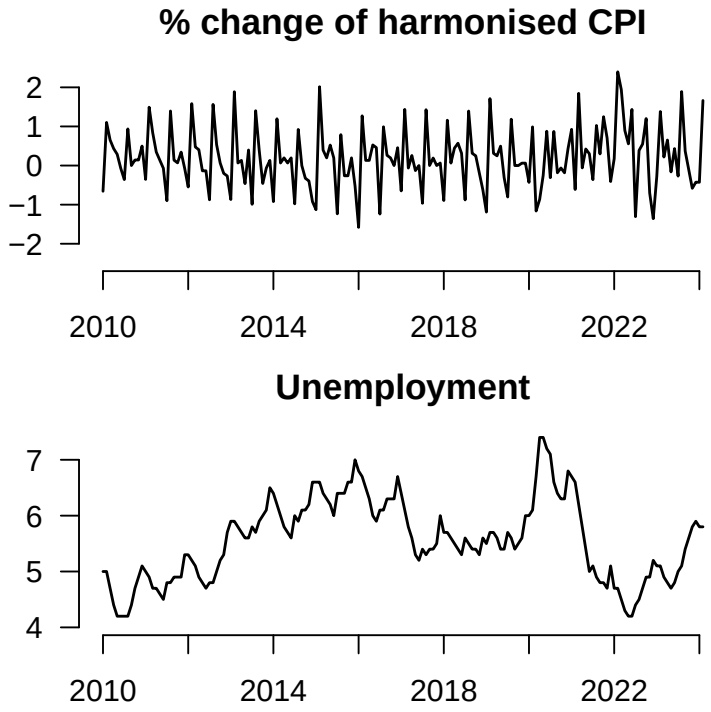
We can combine two specifications: Y_t can be explained by its past, and the error term can exhibit persistence.

$$Y_t = \mu_0 + U_t + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j U_{t-j}$$

- Promoted by Box and Jenkins (1970)
- No external regressors are needed (assuming that the past is informative about the future)
- Often out-performs complex structural models

Seasonal AR models

Some phenomena exhibit quasi-cyclical behaviour, usually related to the Earth rotation or cobweb effect.
Cycle length: $c = 4$ for quarterly, $c = 12$ for monthly data.



Cyclicalities can be (at least partially) captured by

$$Y_t = \mu_0 + \varphi_1 Y_{t-1} + \varphi_c Y_{t-c} + U_t$$

Estimable by OLS.

Seasonal ARMA models

We can add seasonal lags to ARMA models.

$$Y_t = \mu_0 + U_t + \underbrace{\sum_{\substack{i=1 \\ i \bmod c \neq 0}}^{p} \varphi_p Y_{t-i} + \sum_{\substack{j=1 \\ j \bmod c \neq 0}}^{q} \theta_j U_{t-j}}_{\text{non-seasonal part}} + \underbrace{\sum_{i=1}^{p_s} \varphi_{i \cdot c} Y_{t-ic} + \sum_{j=1}^{q_s} \theta_{j \cdot c} U_{t-jc}}_{\text{seasonal part}}$$

c is the cycle length.

- Rarely anything with more than two lags is used
- Notation: SARMA$(p, q)(p_s, q_s)$

What is a good TS model?

- ARMA processes do not exist in real life and are used merely as convenient approximations
- Modelling assumption: U_t is white noise ⇒ 'residuals look like white noise' is good enough
- Plato's cave: the world cannot be learned using linear models with imprecise data; we are like the ER doctors, slapping on a plaster and calling it a day
- Occam's razor: explanations that posit fewer entities are to be preferred

Maximum likelihood

With the help of providence, assume some convenient and flexible parametric density $f_{Y|X;\theta}$. It is known up to θ – suppose that exists θ_0 such that $f_{Y|X;\theta_0}$ is the true law (and a couple of technical assumptions).

Then, θ_0 can be estimated by maximising the expected logarithm of the conditional density:

$$\theta_0 = \arg\max_{\theta} \mathbb{E} \log f_{Y|X;\theta}(Y)$$

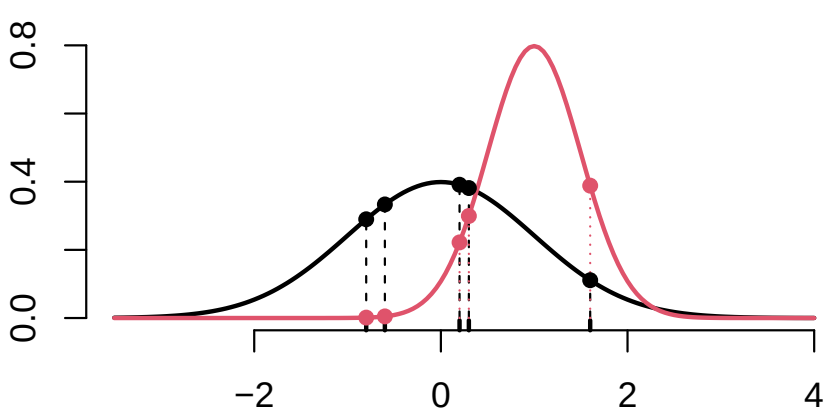
Given data $\{(Y_t, X_t)\}_{t=1}^T$:

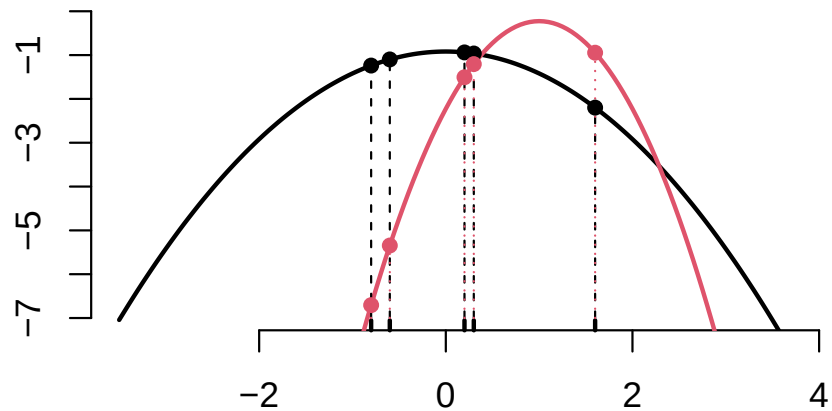
$$\hat{\theta}_{\text{ML}} \coloneqq \arg\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} \log f_{Y_t|X_t;\theta}(Y_t) = \arg\max_{\theta} \prod_{t=1}^{T} f_{Y_t|X_t;\theta}(Y_t)$$

Maximum likelihood intuition

Assuming that unknown θ defines a data-generating process (model), which value of θ most likely generated our data?

Example: distribution fitting. Find such (μ, σ^2) that would generate a Gaussian curve with the highest average log-likelihood for IID $\{Y_t\}_{t=1}^5 = \{-0.8, -0.6, 0.2, 0.3, 1.6\}$.





$\frac{1}{5}\sum_{t=1}^{5} \log f_{\mathcal{N}(0,1)}(Y_t) = -1.3 > \frac{1}{5}\sum_{t=1}^{5} \log f_{\mathcal{N}(1,0.5)}(Y_t) = -3.1$

Maximum likelihood fitting

- Assume a conditional density function $t \mapsto f_{Y|X;\theta}(t)$ that fully describes the model
- Re-interpret it as a likelihood function $\theta \mapsto f_{Y|X;\theta}(Y)$

$$\hat{\theta}_{\mathrm{ML}} \coloneqq \arg\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} \log f_{Y_t|X_t;\theta}(Y_t) \coloneqq \arg\max_{\theta} \frac{1}{T} \mathcal{L}_T(\theta)$$

- Find $\hat{\theta}_{\mathrm{ML}}$ as the average log-likelihood maximiser
  - Solve the FOC $\frac{1}{T} \sum_{t=1}^{T} \nabla_\theta \mathcal{L}_T(\theta)$ whilst praying that this problem is well-behaved (the global maximum of $\mathcal{L}_T(\theta)$ exists and is unique; $\mathcal{L}_T$ is smooth in θ)
  - Use any reasonable numerical optimisation technique (recommendation: stochastic, then gradient-based)

Maximum likelihood testing

1. Estimate a model without restrictions ⇒ get $\mathcal{L}(\hat{\theta}_{\text{UR}})$
2. Fix k hypothesised parameter values / impose k constraints and estimate a restricted model w. r. t. remaining parameters ⇒ get $\mathcal{L}(\hat{\theta}_{\text{R}})$

Then, if $\mathcal{H}_0$: 'the constraints hold' is true,

$$2[\mathcal{L}(\hat{\theta}_{\text{UR}}) - \mathcal{L}(\hat{\theta}_{\text{R}})] \xrightarrow[T\to\infty]{d} \chi^2_k$$

The LR test requires two estimations (unrestricted and restricted), but is universally most powerful (UMP, i. e. detects deviations from the null) for point hypotheses.

We skip the Wald and Lagrange-multiplier tests.

Equivalence: ML and OLS

Consider a linear regression model:

$$Y_t = \tilde{X}_t'\theta + U_t, \quad \mathbb{E}(U \mid X) = 0$$

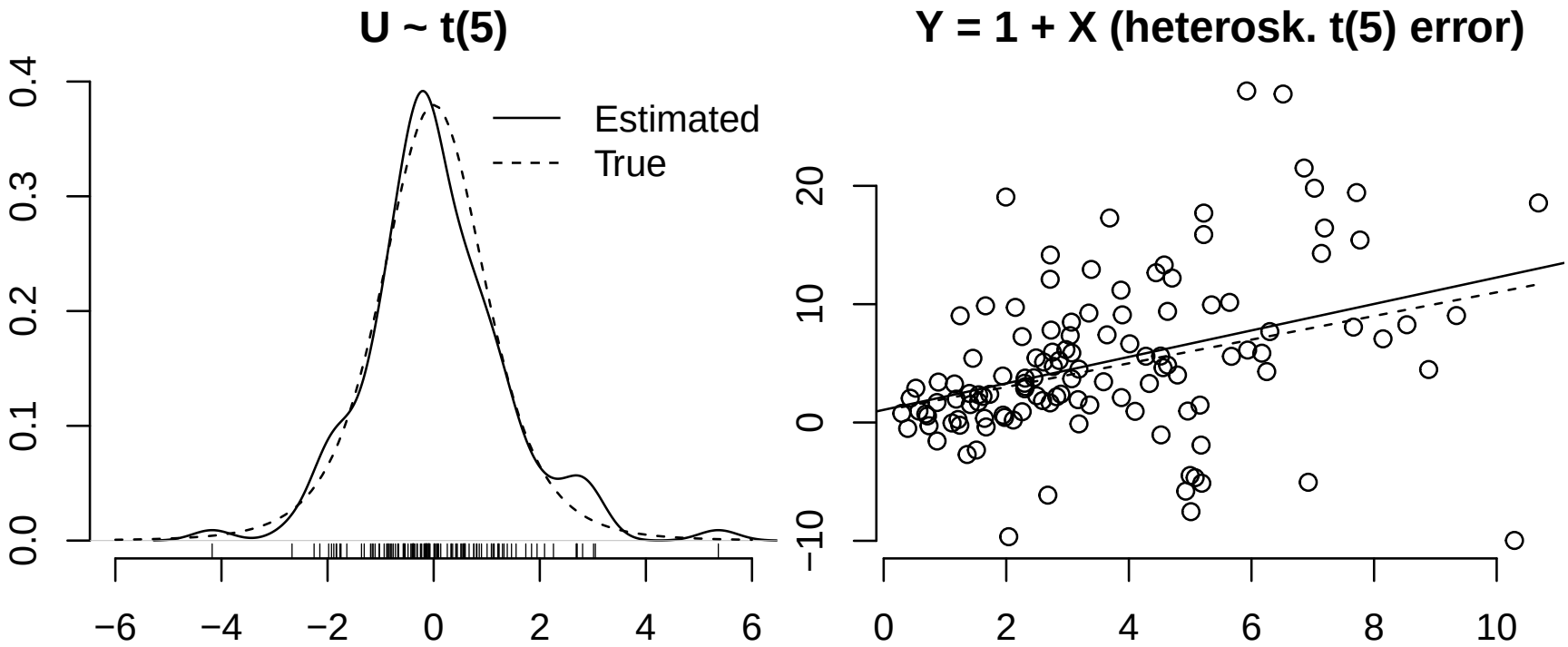
$$\hat{\theta}_{\text{LS}} \coloneqq \arg\min_{\theta} \frac{1}{T} \sum_{t=1}^{n} (Y_t - \tilde{X}_t'\theta)^2$$

Add the normality assumption: $U \mid X \sim \mathcal{N}(0, \sigma^2)$:

$$(\hat{\theta}_{\text{ML}}, \hat{\sigma}^2_{\text{ML}}) \coloneqq \arg\max_{\theta, \sigma^2} \frac{1}{T} \sum_{t=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_t - \tilde{X}_t'\theta)^2}{2\sigma^2} \right) \right)$$

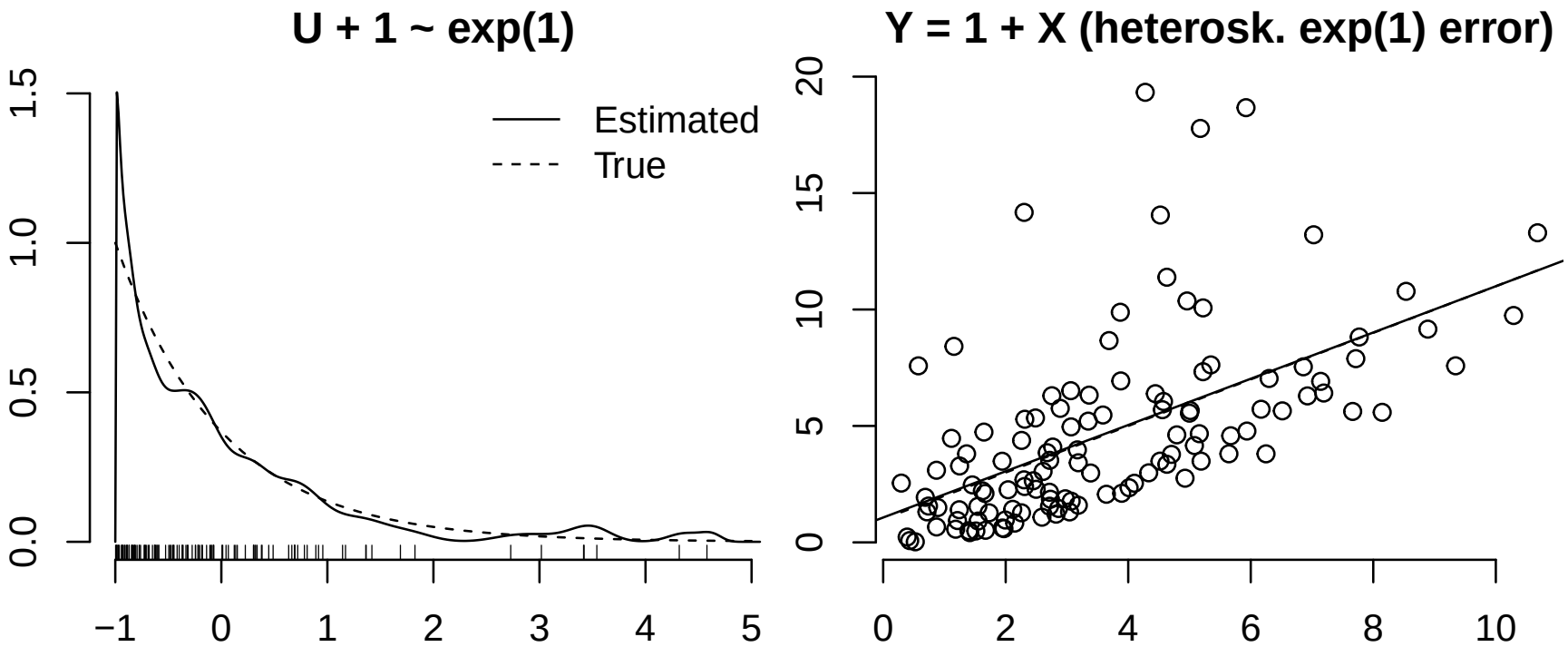
Claim. $\hat{\theta}_{\text{LS}} \equiv \hat{\theta}_{\text{ML}}$ in this case.

Avoid unnecessary assumptions (1/2)



OLS works well (is consistent) for correctly specified models ($\mathbb{E}(U \mid X) = 0$) even if $U \mid X$ is non-Gaussian (e. g. has heavier tails).

Avoid unnecessary assumptions (2/2)



OLS works well (is consistent) for correctly specified models ($\mathbb{E}(U \mid X) = 0$) even if $U \mid X$ is asymmetric.

ML under the wrong model

Suppose that the true conditional density is $f_{Y_t|X_t,\theta_0}$ but instead, one assumes $g_{Y_t|Z_t,\theta_0}$.

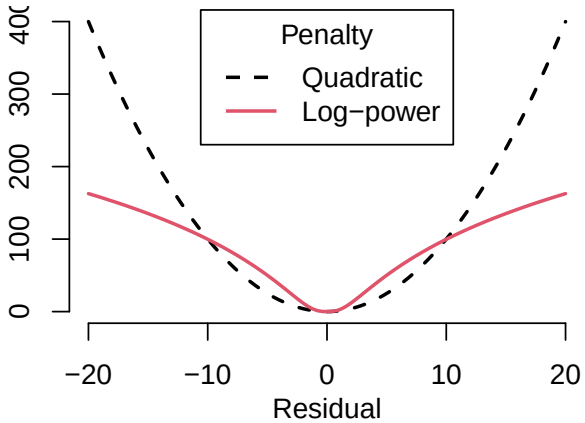
Example: assuming normality where it is clearly violated.

Can θ_0 be estimated consistently by maximising the wrong likelihood?

- In the general case, $\hat{\theta}_{\text{ML}}$ converges to the parameter of the closest model (i. e. the wrong one) $\Rightarrow$ bias, inconsistency, inefficiency
- Some estimates are robust to mis-specification, and the estimator variance can be consistently estimated using the sandwich formula due to White (1982)

Equivalence: ML and robust regression

Researchers in finance often assume 'heavy tails' and non-zero probabilities of extreme events and work with data containing influential observations.



Assuming the Student-t error distribution and maximising the Student-t log-likelihood is the same as minimising the robust penalty $\frac{1}{T}\sum_{t=1}^{T}\log\left(1+\frac{U(Y_t,X_t,\theta)^2}{\delta_1}\right)^{\delta_2}$ for some chosen $\delta_1,\delta_2>0$.

ML estimation is more popular in risk modelling than minimising Huber-like penalties that grow slower than the quadratic OLS penalty.

Arbitrary parametric densities

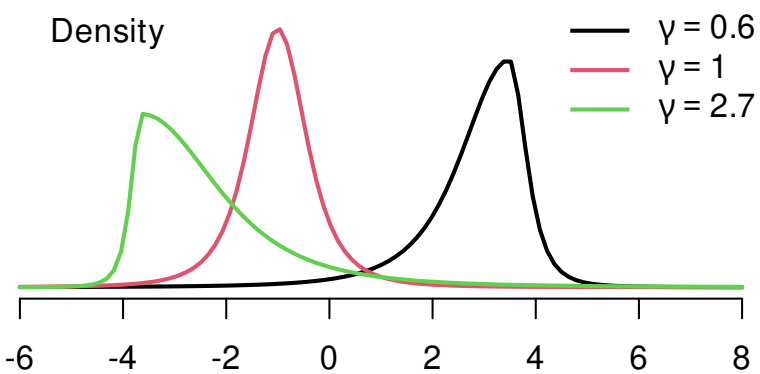
Feel free to assume any parametric conditional density / joint distribution that is common in the field.

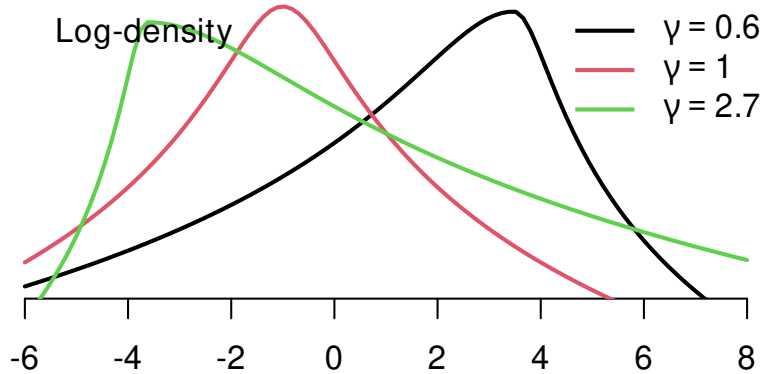
At least **try** something more flexible than the Gaussian distribution.

Skew-Student distribution

Fernandez & Steel (1998) proposed generalising any distribution to account for asymmetry and heavy tails.

- Take any unimodal density $f_X(t)$ symmetric around 0
- Rescale its left part by $\gamma > 0$ and the right by $1/\gamma$:
  $f_{X^*}(t) := \frac{2}{\gamma+1/\gamma}[f_X(t/\gamma)\mathbb{I}(t \geq 0) + f_X(\gamma t)\mathbb{I}(t < 0)]$
  - Re-centre and re-normalise if necessary to match the moments of the original distribution
- Drawback: less numerically stable (solution: use fixed γ)





ML estimation of AR(p) models

$$Y_t = \mu_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + U_t$$

Assuming $U_t/\sigma \sim \mathcal{N}(0,1)$, denoting ϕ the PDF of $\mathcal{N}(0,1)$:

$$U_t(\theta_0) = Y_t - \mu_0 - \sum_{i=1}^{p} \varphi_i Y_{t-i}$$

$$\log f_{Y_t|Y_{t-1},\ldots;\theta} = \log \phi(U_t(\theta)/\sigma)/\sigma$$

Maximise $\frac{1}{T}\sum_{t=p+1}^{T} \log \frac{1}{\sigma}\phi(U_t(\theta)/\sigma)$: for any value $(\tilde{\theta}, \tilde{\sigma}^2)$, compute the residuals $U_t(\tilde{\theta})$, evaluate the log-densities, add them up to get $\mathcal{L}(\tilde{\theta}, \tilde{\sigma}^2)$, find the direction of its growth w. r. t. θ, choose a better guess of (θ, σ^2) until convergence.

ML estimation of ARMA(p, q) models

$$Y_t = \mu_0 + U_t + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \varphi_p Y_{t-p} + \sum_{j=1}^{q} \theta_q U_{t-q}$$

$$U_t(\cdot) = Y_t - \mu_0 - \sum_{i=1}^{p} \varphi_i Y_{t-i} - \sum_{j=1}^{q} \theta_q U_{t-q}$$

Problem: unlike AR(p), we cannot generate $\{U_t\}_{t=q+1}^{T}$.
Solution: make some assumptions about p extra values $\underline{Y} \coloneqq Y_0, Y_{-1}, \ldots$ and q extra values $\underline{U} \coloneqq U_0, U_{-1}, \ldots$.

If $(\underline{Y}, \underline{U})$ are known, $U_1(\theta_0, \underline{Y}, \underline{U})$ can be *conditionally* computed, and $\{U_t\}_{t=1}^{T}$ reconstructed for plugging into $\phi(\cdot)$.

Recommendation: trust the built-in ARMA functions (they make smart guesses about $\underline{Y}$ and $\underline{U}$)!

Thank you for your attention!