Statistics 203: Introduction to Regression and Analysis of Variance

Time Series: Brief Introduction

Jonathan Taylor

Today's class

- Models for time-correlated noise.
- Stationary time series.
- ARMA models.
- Autocovariance, power spectrum.
- Diagnostics.

Modelling correlation

- In the mixed effects model

$$Y = X\beta + Z\gamma + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2 I)$ and $\gamma \sim N(0, D)$ we were essentially saying

$$Y \sim N(X\beta, ZDZ^t + \sigma^2 I)$$

- We then estimated D from the data (more precisely, R does this for us).
- We can impose structure on D if necessary. For example, in two-way random effects ANOVA, we assumed that $\alpha_i, \beta_j, (\alpha\beta)_{ij}$ were independent mean zero normal random variables.
- In summary, a mixed effect model can be thought of as modelling the correlation in the errors of Y coming from "sampling from a population."

Other models of correlation

- Not all correlations come from sampling.
- Another common source is correlation in time.
- Example: imagine modelling monthly temperature in a given location over many years.
  - $Y_t = \mu_{t\%12} + \varepsilon_t, 1 \leq t \leq T$
  - Clearly, μ will vary smoothly as a function of t, but there will also be correlation in ε_t due to "weather systems" that last more than one day.
  - To estimate μ "optimally" and (especially to) make inferences about μ we should take these correlations into account.
- Time series models are models of such (auto)correlation. Good references: *Priestley*, "Spectral Theory and Time Series"; *Brockwell and Davis*, "Introduction to Time Series and Forecasting."
- Nottingham temperature example.
- Today we will just talk about time series in general.

Autoregressive models

- Simplest stationary "auto"correlation

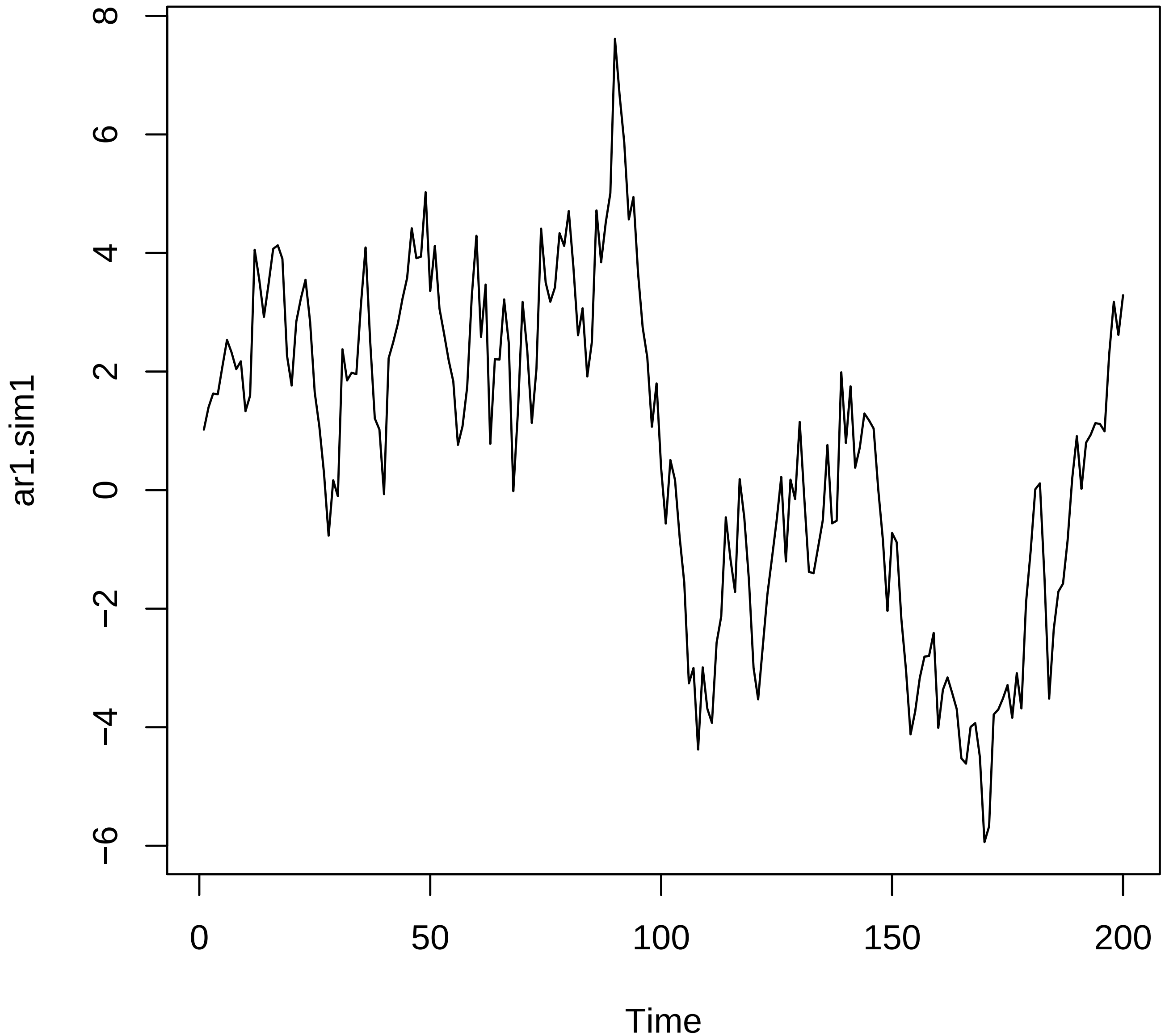
$$\varepsilon_t = \alpha \cdot \varepsilon_{t-1} + \eta_t$$

  where $\eta \sim N(0, \sigma^2 I)$ are i.i.d. Normal random variables, $|\alpha| < 1$.
- This is called an *auto-regressive* process: "auto" because ε is like a regression of ε on its past.
- It is called $AR(1)$ because it only goes 1 time point into the past.
- Covariance / correlation function

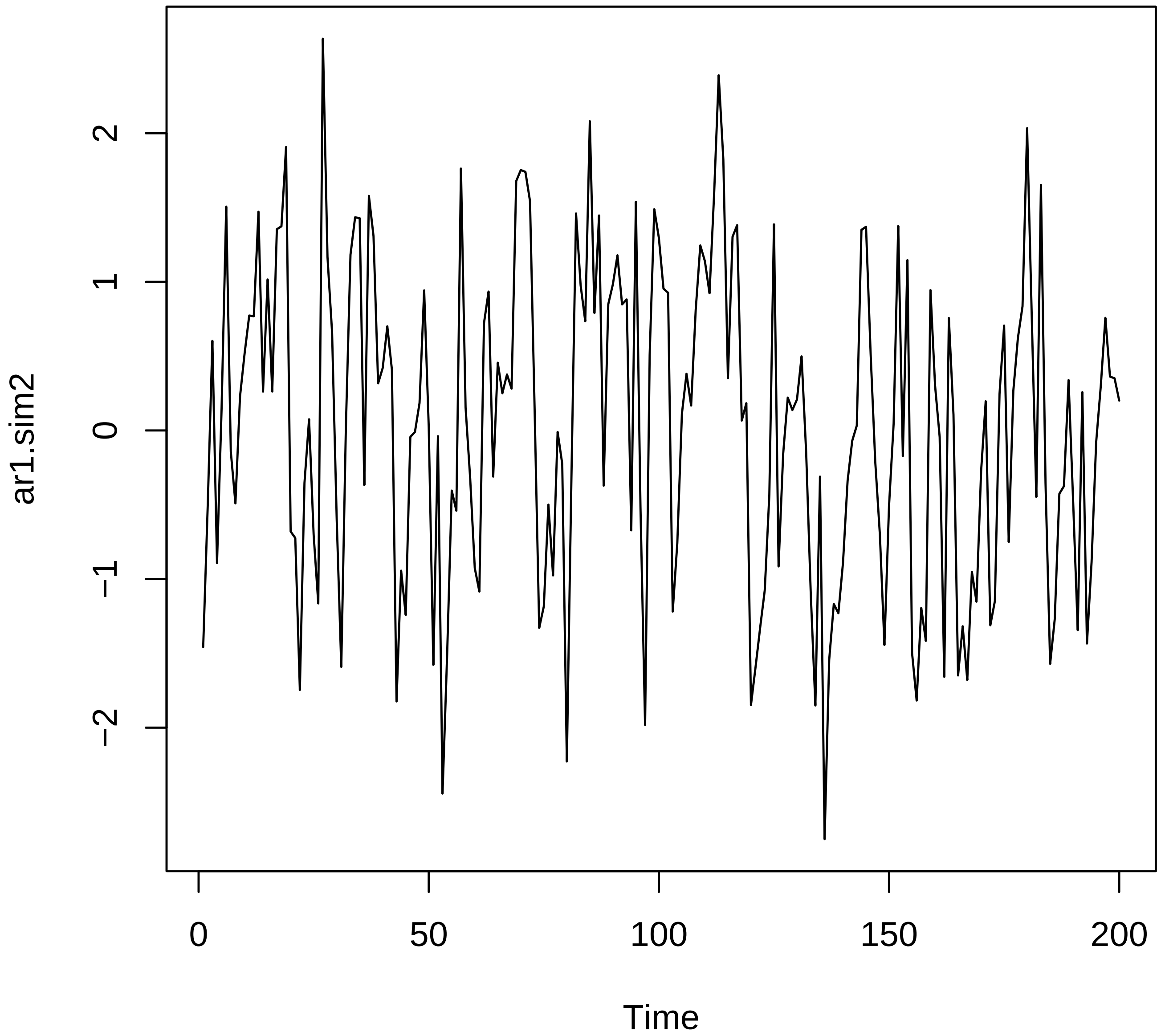
$$\text{Cov}(\varepsilon_t, \varepsilon_{t+j}) = \frac{\sigma^2 \alpha^{|j|}}{1 - \alpha^2}, \qquad \text{C}or(\varepsilon_t, \varepsilon_{t+j}) = \alpha^{|j|}$$

- Model is "stationary" because $\text{Cov}(\varepsilon_t, \varepsilon_{t+j})$ depends only on $|j|$.

$AR(1), \alpha = 0.95$



$AR(1), \alpha = 0.5$



$AR(k)$ models

- The $AR(1)$ model can be easily generalized to the $AR(p)$ model:

$$\varepsilon_t = \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j} + \eta_t$$

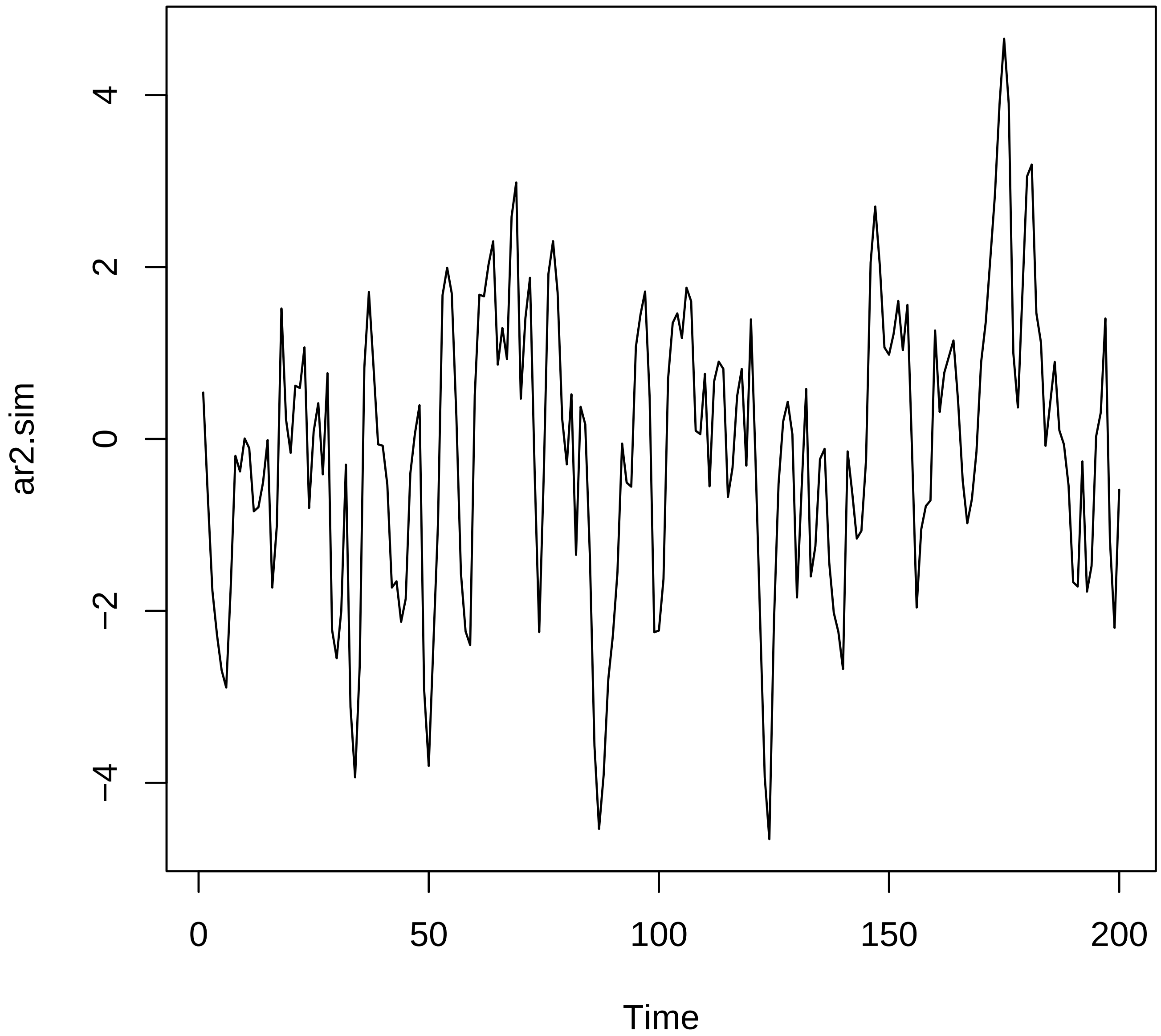
where $\eta \sim N(0, \sigma^2 I)$ are i.i.d. Normal random variables.

- Condition on α's: all roots of the (complex) polynomial

$$\phi_\alpha(z) = 1 - \sum_{j=1}^{p} \alpha_j z^j$$

are within the unit disc in the complex plane.

$AR(2), \alpha_1 = 0.9, \alpha_2 = -0.2$



Moving average & $ARMA(p,q)$ models

- $MA(q)$ is another stationary model:

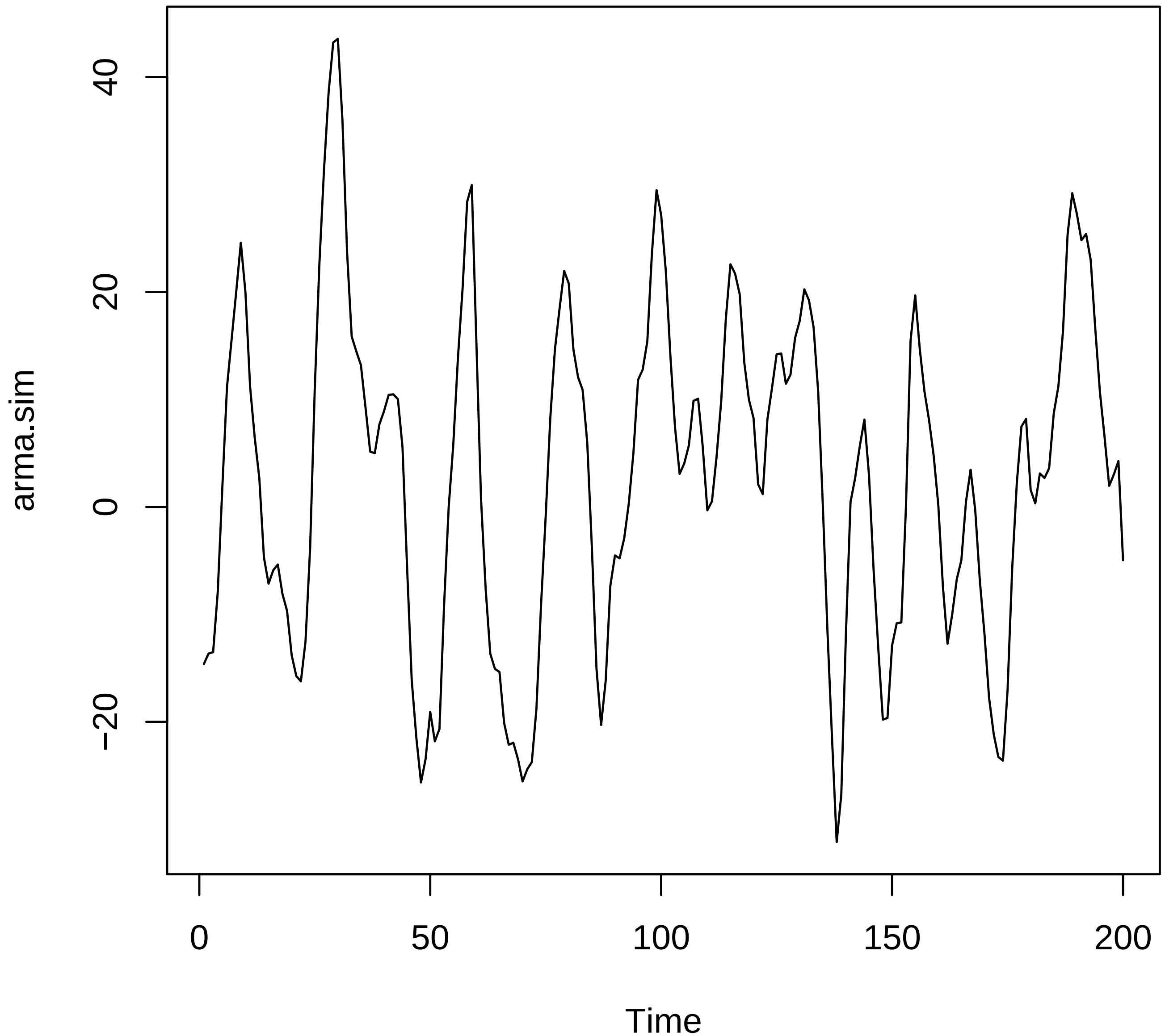
$$\varepsilon_t = \sum_{j=0}^{q} \beta_j \eta_{t-q}$$

  where $\eta \sim N(0, \sigma^2 I)$ are i.i.d. Normal random variables.
- No conditions on β's – this is always stationary.
- $ARMA(p,q)$ model:

$$\varepsilon_t = \sum_{l=1}^{p} \alpha_l X_{t-l} + \sum_{j=0}^{q} \beta_j \eta_{t-q}$$

  where $\eta \sim N(0, \sigma^2 I)$ are i.i.d. Normal random variables.

$ARMA(2,4)$



Stationary time series

- In general a (Normally distributed) time series (ε_t) is stationary if

$$\text{Cov}(\varepsilon_t, \varepsilon_{t+j}) = R(|j|)$$

  for some "covariance" function R.
- If errors are not normally distributed then the process is called weakly stationary, or stationary in mean-square.
- The function $R(t)$ can generally be expressed as the Fourier transform of a spectral density

$$R(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it\omega} f_R(\omega)\, d\omega$$

  where f is called the "spectral" density of the process.
- Conversely

$$f_R(t) = \sum_t e^{-it\omega} R(t)$$

- The function f_R is sometimes called the "power spectrum" of ε.

Estimating autocovariance / correlation

- Natural estimate of covariance function for $t \geq 0$ based on observing $(\varepsilon_1, \dots, \varepsilon_n)$

$$\widehat{R}(t) = \frac{1}{n} \sum_{j=1}^{n-t} (\varepsilon_{j+t} - \overline{\varepsilon})(\varepsilon_j - \overline{\varepsilon}).$$

- Estimate of correlation function

$$\widehat{\text{Cor}}(t) = \frac{\widehat{R}(t)}{\widehat{R}(0)}.$$

Estimating power spectrum

- Estimate of power spectrum based on observing $(\varepsilon_1, \dots, \varepsilon_n)$ is called the *periodogram*, the discrete Fourier transform of ε

$$I(\omega) = \frac{\left|\sum_{t=1}^{n} e^{-i\omega t}\varepsilon_t\right|^2}{n}$$

- In fact

$$I(\omega) = \sum_t \widehat{R}(t) e^{-i\omega t}$$

  i.e. it is the Fourier transform of $\widehat{R}(t)$.

- It is customary to use a smoothed periodogram as an estimate of f_R

$$\widehat{f}_R(\omega) = \int K_h((\lambda - \omega)/h) I(\lambda)\, d\lambda.$$

  for some kernel K_h.

- If ε's are i.i.d. (hence stationary), then

$$\mathbb{E}(I(\omega)) = \text{Var}(\varepsilon).$$

Diagnostics

- Suppose we fit an $ARMA(p, q)$ model to observations $(\varepsilon_1, \dots, \varepsilon_n)$: how can we tell if the fit is "good"?
- How do we do this? By residuals of course. In an $AR(p)$ model, for instance, define

$$\widehat{\eta}_t = X_t - \sum_{j=1}^{p} \widehat{\alpha}_j X_{t-j}.$$

- These should look like an i.i.d. sequence, at least roughly.
- Can plot residuals themselves, autocorrelation function of residuals, and cumulative periodogram.