

# Chapter 1

## An overview of Statistical Methods

### 1.1 Introduction

Technological invention, trade globalization and internet revolution have brought us into **new era of financial markets**:

- Many new financial products have been introduced.
- Switch from management based on accrual accounting to that

based on daily marking-to-market.

- Important milestones in 1973:
  - options exchange opened in Chicago
  - Black-Scholes option pricing formula
  - Merton's general equilibrium model for security pricing.
- Derivative markets have experienced extraordinary growth.

based on daily marking-to-market.

- Important milestones in 1973:
  - options exchange opened in Chicago
  - Black-Scholes option pricing formula
  - Merton's general equilibrium model for security pricing.
- Derivative markets have experienced extraordinary growth.

Professionals in finance now routinely use **sophisticated statistical techniques** and modern computation power in

♠ portfolio management; ♠ securities regulation; ♠ proprietary trading  
♣ asset pricing; ♣ financial consulting; ♣ risk management.

based on daily marking-to-market.

- Important milestones in 1973:
  - options exchange opened in Chicago
  - Black-Scholes option pricing formula
  - Merton's general equilibrium model for security pricing.
- Derivative markets have experienced extraordinary growth.

Professionals in finance now routinely use **sophisticated statistical techniques** and modern computation power in

♠ portfolio management; ♠ securities regulation; ♠ proprietary trading  
♣ asset pricing; ♣ financial consulting; ♣ risk management.

**Financial econometrics** studies econometric and statistical aspects of finance. It is an active field of **integration** of

- finance (problems; products);
- economics (theory);
- statistics and mathematical sciences (quantitative tools).

**Financial econometrics** studies econometric and statistical aspects of finance. It is an active field of **integration** of

- finance (problems; products);
- economics (theory);
- statistics and mathematical sciences (quantitative tools).

**Example:**

- Complex products pose new challenges on their valuation and hedging.
- Sophisticated stochastic models are introduced to capture salient features of underlying economic variables.

— Statistical tools are used to identify parameters of stochastic models; and to simulate complex financial systems; to empirically test stochastic models and systems; to price assets directly from empirical data; and to manage market risks.

Financial Econometrics deals with **cross-sectional** and **longitudinal** data.

## 1.2 An overview of statistical methods

Sampling data:  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ .

Objectives: Estimate the population parameter  $\boldsymbol{\theta}$ , attach associated estimation errors, and make inferences about  $\boldsymbol{\theta}$ .

Modeling: Think  $\mathbf{x}$  as realizations from the random variable  $\mathbf{X}$ , having joint density  $f(\mathbf{x}, \boldsymbol{\theta})$ .

## 1.2 An overview of statistical methods

Sampling data:  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ .

Objectives: Estimate the population parameter  $\boldsymbol{\theta}$ , attach associated estimation errors, and make inferences about  $\boldsymbol{\theta}$ .

Modeling: Think  $\mathbf{x}$  as realizations from the random variable  $\mathbf{X}$ , having joint density  $f(\mathbf{x}, \boldsymbol{\theta})$ .

Example 1.1: (logistic regression) In horse-race wagering, to predict the probability of a horse wins a race, the following logistic regression model is frequently used. Let  $Y = 1$  (or  $= 0$ ) be the indicator that a particular horse wins (or loses) a race and  $\mathbf{X}$  be its associated characteristics or attributes. (e.g.  $X_1$  is the public odds on winning 3

minutes before the race;  $X_2$  is the weight a horse carries;  $X_3$  track number;  $X_4$  is the jockey; and so on.).

minutes before the race;  $X_2$  is the weight a horse carries;  $X_3$  track number;  $X_4$  is the jockey; and so on.).

It is frequently assumed that

$$\begin{aligned} P(Y = 1|\mathbf{X}) &= \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} \\ &= \frac{\exp(\mathbf{X}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}^T \boldsymbol{\beta})}. \end{aligned}$$

minutes before the race;  $X_2$  is the weight a horse carries;  $X_3$  track number;  $X_4$  is the jockey; and so on.).

It is frequently assumed that

$$\begin{aligned} P(Y = 1|\mathbf{X}) &= \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} \\ &= \frac{\exp(\mathbf{X}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}^T \boldsymbol{\beta})}. \end{aligned}$$

Suppose that we have a random sample  $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$  from the above logistic regression model. In the above notation,  $\mathbf{x} = \{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$  and  $\mathbf{X}$  is the population under which the data were drawn. To predict the probability of winning, we need

- to estimate the unknown parameters  $\boldsymbol{\beta}$  and their associated standard errors;

- to select important variables; testing some coefficients are zero;
- to construct confidence intervals for winning probability

$$p^* = \frac{\exp(\mathbf{x}^{*T} \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^{*T} \boldsymbol{\beta})},$$

given the characteristic  $\mathbf{x}^*$ .

- to select important variables; testing some coefficients are zero;
- to construct confidence intervals for winning probability

$$p^* = \frac{\exp(\mathbf{x}^{*T} \boldsymbol{\beta})}{1 + \exp(\mathbf{x}^{*T} \boldsymbol{\beta})},$$

given the characteristic  $\mathbf{x}^*$ .

**Example 1.2.** The Fifth National Bank of Springfield was charged in court with that its female employees received substantially smaller salaries than its male employees. For each of its 208 employees, the data in 1995 include

♠ Salary   ♠ Job Grade (1–6);   ♠ Year Hired ;   ♠ Prior Experience  
♣ Year Born;   ♣ Gender;   ♣ Education level;   ♣ Indicator of PC Job

One possible model is

$$\begin{aligned} \text{Salary} = & \beta_0 + \beta_1 \text{JobGrade} + \beta_2 \text{YrExp} + \beta_3 \text{YrEdu} \\ & + \beta_4 \text{age} + \beta_5 \text{Gender} + \varepsilon \end{aligned}$$

$$\begin{aligned} \text{Salary} = & \beta_0 + \beta_1 \text{JobGrade} + \beta_2 \text{YrExp} + \beta_3 \text{YrEdu} \\ & + \beta_4 \text{age} + \beta_5 \text{Gender} + \varepsilon \end{aligned}$$

This is a continuous type of response variable.

- The interest is if  $\beta_5 = 0$  or  $\beta_5 > 0$ , which is the salary difference between males and females controlling other variables.
- Should we revise the model? e.g. if the age enters the equation linearly? if the salary and JobGrade are linearly related?
- Is there any discrimination in terms of promotion, even if there is no gender difference after adjusting for job grade. This is a categorical type of response.

### 1.2.1 Maximum Likelihood Estimation

Joint Density:  $f(\mathbf{x}_1, \dots, \mathbf{x}_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta})$  for independent data.

Log-likelihood:  $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{x}_i; \boldsymbol{\theta})$

Maximum likelihood estimator: Find  $\hat{\boldsymbol{\theta}}$  to maximize  $\ell(\boldsymbol{\theta})$  or to solve the likelihood equation

$$\ell'(\boldsymbol{\theta}) = 0,$$

where  $\ell'(\boldsymbol{\theta})$  is the gradient vector of  $\ell(\boldsymbol{\theta})$ .

### 1.2.1 Maximum Likelihood Estimation

Joint Density:  $f(\mathbf{x}_1, \dots, \mathbf{x}_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta})$  for independent data.

Log-likelihood:  $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{x}_i; \boldsymbol{\theta})$

Maximum likelihood estimator: Find  $\hat{\boldsymbol{\theta}}$  to maximize  $\ell(\boldsymbol{\theta})$  or to solve the likelihood equation

$$\ell'(\boldsymbol{\theta}) = 0,$$

where  $\ell'(\boldsymbol{\theta})$  is the gradient vector of  $\ell(\boldsymbol{\theta})$ .

Example 1.1 (Cont.). For this model,  $\tilde{\mathbf{x}}_i = (\mathbf{x}_i, y_i)$ . The joint density is

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n)P(Y_1 = y_1, \dots, Y_n = y_n | \mathbf{x}_1, \dots, \mathbf{x}_n).$$

The first term does not involve  $\boldsymbol{\beta}$ . Thus, the likelihood function is equivalent to the conditional likelihood

$$\prod_{i=1}^n P(Y_i = y_i | \mathbf{x}_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1 - y_i},$$

where  $p_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$ . Hence,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}.$$

The first term does not involve  $\boldsymbol{\beta}$ . Thus, the likelihood function is equivalent to the conditional likelihood

$$\prod_{i=1}^n P(Y_i = y_i | \mathbf{x}_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1 - y_i},$$

where  $p_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$ . Hence,

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}.$$

**Newton-Raphson algorithm**: In general, the MLE can not be explicitly found. One can find the MLE through the Newton-Raphson algorithm. For a given initial estimate  $\hat{\boldsymbol{\theta}}_0$ , by Taylor's expansion,

$$0 = \ell'(\hat{\boldsymbol{\theta}}) \approx \ell'(\hat{\boldsymbol{\theta}}_0) + \ell''(\hat{\boldsymbol{\theta}}_0)(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_0).$$

Thus, we update the estimator as

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0 - \ell''(\hat{\boldsymbol{\theta}}_0)^{-1} \ell'(\hat{\boldsymbol{\theta}}_0).$$

One can iterate the above steps until convergence.

**One-step estimator**: When an initial estimator  $\hat{\boldsymbol{\theta}}_0$  is root-n consistent, then one-step iteration suffices to obtain the same asymptotic efficiency as the fully iterated one.

Thus, we update the estimator as

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0 - \ell''(\hat{\boldsymbol{\theta}}_0)^{-1} \ell'(\hat{\boldsymbol{\theta}}_0).$$

One can iterate the above steps until convergence.

**One-step estimator**: When an initial estimator  $\hat{\boldsymbol{\theta}}_0$  is root-n consistent, then one-step iteration suffices to obtain the same asymptotic efficiency as the fully iterated one.

**Initial estimator**: Use some simple methods such as the method of moments or approximation scheme.

**Asymptotic normality**: Under some regularity conditions,  $\hat{\boldsymbol{\theta}}$  is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{L} N(0, I^{-1}(\boldsymbol{\theta})),$$

where  $I(\boldsymbol{\theta}) = -\lim_{n \rightarrow \infty} n^{-1} E\ell''(\boldsymbol{\theta})$  is the Fisher information matrix.

**Estimate of asymptotic covariance matrix:** The asymptotic covariance matrix  $I^{-1}(\boldsymbol{\theta})/n = [nI(\boldsymbol{\theta})]^{-1}$  is usually estimated by  $\hat{\boldsymbol{\Sigma}} = [-\ell''(\hat{\boldsymbol{\theta}})]^{-1}$ , the inverse of Hessian matrix at the MLE and the standard error of each estimator is given by its diagonal element.

where  $I(\boldsymbol{\theta}) = -\lim_{n \rightarrow \infty} n^{-1} E \ell''(\boldsymbol{\theta})$  is the Fisher information matrix.

**Estimate of asymptotic covariance matrix:** The asymptotic covariance matrix  $I^{-1}(\boldsymbol{\theta})/n = [nI(\boldsymbol{\theta})]^{-1}$  is usually estimated by  $\hat{\boldsymbol{\Sigma}} = [-\ell''(\hat{\boldsymbol{\theta}})]^{-1}$ , the inverse of Hessian matrix at the MLE and the standard error of each estimator is given by its diagonal element.

**Example 1.1** (Cont.). It can easily be calculated that

$$\ell''(\boldsymbol{\beta}) = -\sum_{i=1}^n \hat{p}_i(1 - \hat{p}_i) \mathbf{x}_i \mathbf{x}_i^T.$$

Hence, the estimated variance-covariance for  $\hat{\beta}$  is

$$\hat{\Sigma} = \left( \sum_{i=1}^n \hat{p}_i(1 - \hat{p}_i) \mathbf{x}_i \mathbf{x}_i^T \right)^{-1}.$$

In particular, the standard error of  $\hat{\beta}_i$  is the square root of the  $i$ -th element and the asymptotic covariance between the  $\hat{\beta}_i$  and  $\hat{\beta}_j$  is the  $(i, j)$  element of  $\hat{\Sigma}$ .

### Example 1.3. Credit risk / Health risk

The “burn data” were collected by the General Hospital Burn Center at the University of Southern California, concerning about the survival status of patients suffering from various types of burns along with associated covariates.

### Example 1.3. Credit risk / Health risk

The “burn data” were collected by the General Hospital Burn Center at the University of Southern California, concerning about the survival status of patients suffering from various types of burns along with associated covariates. We take

- the response variable as the survival status of burn victims and
- covariates as ethnicity (white and non-white),
- age (A – under 40, B – Between 40 and 60, C — Over 60),
- ratio of third-degree burn areas (continuous),
- prior respiratory disease (No or Yes).

To account possible nonlinear effect, we introduce the variable dratio, which discretizes the variable ratio (A —0%, B – between 0% and 15%, C – between 15% and 70%, D — over 70%). After deleting some missing cases, we have 975 cases remaining.

To account possible nonlinear effect, we introduce the variable dratio, which discretizes the variable ratio (A — 0%, B – between 0% and 15%, C – between 15% and 70%, D — over 70%). After deleting some missing cases, we have 975 cases remaining.

As the first attempt, the generalized linear models are fitted with the following command in Splus.

```
options(contrasts=c("contr.treatment", "contr.treatment"))
burn.glm <- glm(formula = surv ~ ratio+ Age + dratio
+ race + resp, family = binomial, data = burn.sub)
summary(burn.glm)
```

	Coefficients	Std. Error	t value
(Intercept)	4.286370156	0.4304437	9.95802794
ratio	-2.034206415	0.9673477	-2.10286997
AgeB	-1.025769716	0.2555389	-4.01414325

AgeC	-2.730384558	0.2746412	-9.94164112
dratioB	-0.982073606	0.4894461	-2.00650017
dratioC	-0.736154133	0.5341817	-1.37809683
dratioD	-0.643689332	0.9556140	-0.67358715
race	-0.238447516	0.2313350	-1.03074558
resp	-0.003632491	0.3037939	-0.01195709

e.g. SE for  $\hat{\beta}_8$  and  $\hat{\beta}_9$  are .2313 and .3038, which are the square-root 8-th and 9-th diagonal elements of  $\hat{\Sigma}$ . Further,  $\text{cov}(\beta_8, \beta_9) = 0.0035$ , which is the (8,9)-element of  $\hat{\Sigma}$ .

AgeC	-2.730384558	0.2746412	-9.94164112
dratioB	-0.982073606	0.4894461	-2.00650017
dratioC	-0.736154133	0.5341817	-1.37809683
dratioD	-0.643689332	0.9556140	-0.67358715
race	-0.238447516	0.2313350	-1.03074558
resp	-0.003632491	0.3037939	-0.01195709

e.g. SE for  $\hat{\beta}_8$  and  $\hat{\beta}_9$  are .2313 and .3038, which are the square-root 8-th and 9-th diagonal elements of  $\hat{\Sigma}$ . Further,  $\text{cov}(\beta_8, \beta_9) = 0.0035$ , which is the (8,9)-element of  $\hat{\Sigma}$ .

### 1.2.2 The delta method

**Question:** What is the standard error for the survival probability

$$\hat{p}^* = \frac{\exp(\mathbf{x}^{*T} \hat{\beta})}{1 + \exp(\mathbf{x}^{*T} \hat{\beta})}?$$

**The delta method**: Suppose that  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{L} N(0, V(\boldsymbol{\theta}))$ .

Then,

$$\begin{aligned} \sqrt{n}\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\} &\approx g'(\boldsymbol{\theta})^T \sqrt{n}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\} \\ &\xrightarrow{L} N\{0, g'(\boldsymbol{\theta})^T V(\boldsymbol{\theta}) g'(\boldsymbol{\theta})\}, \end{aligned}$$

**The delta method**: Suppose that  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{L} N(0, V(\boldsymbol{\theta}))$ .

Then,

$$\begin{aligned} \sqrt{n}\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\} &\approx g'(\boldsymbol{\theta})^T \sqrt{n}\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\} \\ &\xrightarrow{L} N\{0, g'(\boldsymbol{\theta})^T V(\boldsymbol{\theta})g'(\boldsymbol{\theta})\}, \end{aligned}$$

where  $g'(\boldsymbol{\theta})$  is the gradient vector of  $g(\boldsymbol{\theta})$ , namely

$$g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}) \sim N\{0, n^{-1}g'(\boldsymbol{\theta})^T V(\boldsymbol{\theta})g'(\boldsymbol{\theta})\}.$$

**Example 1.1**(Cont.). Let us compute the standard error of the

estimator  $\hat{p}^* = \frac{\exp(\mathbf{x}^{*T}\hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}^{*T}\hat{\boldsymbol{\beta}})}$ . It can easily be computed that

$$\frac{\partial p^*}{\partial \boldsymbol{\beta}} = p^*(1 - p^*)\mathbf{x}^*.$$

Thus, the asymptotic variance is

$$\{\hat{p}^*(1 - \hat{p}^*)\}^2 \mathbf{x}^{*T} \hat{\Sigma} \mathbf{x}^*.$$

Thus, the asymptotic variance is

$$\{\hat{p}^*(1 - \hat{p}^*)\}^2 \mathbf{x}^{*T} \hat{\Sigma} \mathbf{x}^*.$$

**Example 1.3** (cont). As an illustration, let us fit a simpler model based on the previous analysis.

```
burn1.glm <- glm(formula = surv ~ ratio+ Age, family = binomial, data = burn.sub)
```

	Coefficients	Std. Error	t value
(Intercept)	3.621051	0.2291687	15.800809
ratio	-2.391849	0.2693361	-8.880536
AgeB	-1.045473	0.2509960	-4.165297
AgeC	-2.779915	0.2595052	-10.712368

Correlation of Coefficients:

	(Intercept)	ratio	AgeB
ratio	-0.7444388		
AgeB	-0.4803412	0.0984625	
AgeC	-0.5203382	0.1701190	0.3762085

For a white in age group B with 30% burn ratio, estimate his survival probability.

For a white in age group B with 30% burn ratio, estimate his survival probability.

$$\hat{p} = \frac{\exp(3.6211 - 2.3918 * 0.3 - 1.0454)}{1 + \exp(3.6211 - 2.3918 * 0.3 - 1.0454)} = 0.8651.$$

Its standard error can be computed by using the above formula:

$$SE(\hat{p}) = (.865 * .135)^2 \mathbf{x}^{*T} \hat{\Sigma} \mathbf{x}^*,$$

For a white in age group B with 30% burn ratio, estimate his survival probability.

$$\hat{p} = \frac{\exp(3.6211 - 2.3918 * 0.3 - 1.0454)}{1 + \exp(3.6211 - 2.3918 * 0.3 - 1.0454)} = 0.8651.$$

Its standard error can be computed by using the above formula:

$$SE(\hat{p}) = (.865 * .135)^2 \mathbf{x}^{*T} \hat{\Sigma} \mathbf{x}^*,$$

where  $\mathbf{x}^* = (1, .3, 1, 0)$  and

$$\hat{\Sigma} = \begin{pmatrix} 0.229^2 & -.229 * .269 * .744 & -.229 * .251 * .480 & -.229 * .260 * .520 \\ -.269 * .229 * .744 & .269^2 & .269 * .251 * .098 & .269 * .260 * .170 \\ \dots & & & \\ \dots & & & \end{pmatrix}.$$

### 1.2.3 Hypothesis testing and confidence intervals

Testing of hypothesis: Suppose that  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$  and we wish to test

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}.$$

Let  $\boldsymbol{\Sigma}^{11}$  be the first  $d \times d$  submatrix of  $\widehat{\boldsymbol{\Sigma}} = [-\ell''(\widehat{\boldsymbol{\theta}})]^{-1}$ .

### 1.2.3 Hypothesis testing and confidence intervals

Testing of hypothesis: Suppose that  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$  and we wish to test

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}.$$

Let  $\boldsymbol{\Sigma}^{11}$  be the first  $d \times d$  submatrix of  $\widehat{\boldsymbol{\Sigma}} = [-\ell''(\widehat{\boldsymbol{\theta}})]^{-1}$ .

— Wilks test (MLR):  $T_1 = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_0)\}$ , where  $\widehat{\boldsymbol{\theta}}_0$  is the MLE under  $H_0$ , which is  $(\boldsymbol{\theta}_{10}^T, \widehat{\boldsymbol{\theta}}_{20}^T)^T$ .

— Wald test:  $T_2 = (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})^T (\boldsymbol{\Sigma}^{11})^{-1} (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})$ ;

— Rao test:  $T_3 = \ell'(\widehat{\boldsymbol{\theta}}_0)^T (\boldsymbol{\Sigma}^{11})^{-1} \ell'(\widehat{\boldsymbol{\theta}}_0)$ .

**Null distribution:** These three tests are asymptotically equivalent and have null distribution  $\chi_d^2$ , where  $d$  is the number of restrictions under  $H_0$ . There is a duality between testing and confidence region.

**Null distribution:** These three tests are asymptotically equivalent and have null distribution  $\chi_d^2$ , where  $d$  is the number of restrictions under  $H_0$ . There is a duality between testing and confidence region.

**Example 1.1** (Cont.). To test if  $\beta_i = 0$ , the Wald test is to reject  $H_0$  when

$$T_1 = \frac{\widehat{\beta}_i^2}{\widehat{\sigma}_i^2} \geq \chi_1^2(1 - \alpha) \iff |z_i| \geq z_{1-\alpha/2},$$

where  $z_i = \widehat{\beta}_i / \widehat{\sigma}_i$  with  $\widehat{\sigma}_i^2$  being the  $i$ -th diagonal element of  $\widehat{\Sigma}$ . To test  $H_0 : \beta_1 = \beta_2 = 0$ , the likelihood ratio test is

$$T_1 = 2\{\ell(\widehat{\beta}) - \ell(0, 0, \widehat{\beta}_{20})\} \sim \chi_2^2,$$

where  $\widehat{\beta}_{20}$  is the MLE under the null hypothesis, maximizing  $\ell(0, 0, \beta_2)$ .

**Example 1.3.** (Cont.) Let us back to the larger model (9 variables).

Suppose that we wish to test

$$H_0 : \beta_8 = \beta_9 = 0.$$

The Wald test statistic is

$$T_2 \approx \left( \frac{\hat{\beta}_8}{\widehat{\text{SE}}(\beta_8)} \right)^2 + \left( \frac{\hat{\beta}_9}{\widehat{\text{SE}}(\beta_9)} \right)^2 = 1.0307^2 + (-0.0196)^2 = 1.063,$$

since  $\hat{\beta}_8$  and  $\hat{\beta}_9$  are weakly correlated. P-value =  $P(\chi_2^2 > 1.063) = 0.5877$ .

**Example 1.3.** (Cont.) Let us back to the larger model (9 variables).

Suppose that we wish to test

$$H_0 : \beta_8 = \beta_9 = 0.$$

The Wald test statistic is

$$T_2 \approx \left( \frac{\hat{\beta}_8}{\widehat{\text{SE}}(\beta_8)} \right)^2 + \left( \frac{\hat{\beta}_9}{\widehat{\text{SE}}(\beta_9)} \right)^2 = 1.0307^2 + (-0.0196)^2 = 1.063,$$

since  $\hat{\beta}_8$  and  $\hat{\beta}_9$  are weakly correlated. P-value =  $P(\chi_2^2 > 1.063) = 0.5877$ . No strong enough evidence against  $H_0$ . The 95% confidence for  $\beta_8$  and  $\beta_9$  is approximately

$$\{(\beta_8, \beta_9) : (\beta_8 - .238)^2 / .231^2 + (\beta_9 + .004)^2 / .3037^2 \leq 5.99\}.$$

Similarly, if we wish to test

$$H_0 : \beta_5 = \cdots = \beta_9 = 0.$$

The Wilks statistic is

$$T_1 = 2\{\ell(\text{all}) - \ell(\text{age, ratio})\} = 614.12 - 607.62 = 6.49.$$

The degree of freedom  $d = 5$ . P-value =  $P(\chi_5^2 > 6.49) = 0.2613$ .

Accept the null hypothesis.

Similarly, if we wish to test

$$H_0 : \beta_5 = \cdots = \beta_9 = 0.$$

The Wilks statistic is

$$T_1 = 2\{\ell(\text{all}) - \ell(\text{age, ratio})\} = 614.12 - 607.62 = 6.49.$$

The degree of freedom  $d = 5$ . P-value =  $P(\chi_5^2 > 6.49) = 0.2613$ .

Accept the null hypothesis.

**Confidence regions**: With confidence  $1 - \alpha$ , the unknown parameter falls in the confidence region

$$\left\{ \boldsymbol{\theta}_1 : (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)^T (\boldsymbol{\Sigma}^{11})^{-1} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \leq \chi_d^2(1 - \alpha) \right\},$$

which is an ellipsoid. The likelihood ratio based confidence region is

$$\left\{ \boldsymbol{\theta} : 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta})] \leq \chi_p^2(1 - \alpha) \right\},$$

where  $p$  is the number of parameters in  $\boldsymbol{\theta}$ .

## Chapter 2

### Asset returns and Linear Time Series

Tsay: Chap 2, FY: §2.1, §2.2, Chap 3, GJ: Chap 2.

#### 2.1 Returns

Much of financial econometrics concerns about dynamics of the returns and their associated volatilities of assets. Let us first clarify some of their meanings. Let  $S_t$  be the price of an asset (portfolio) at

time  $t$ .

One-period simple return:  $R_t = (S_t - S_{t-1})/S_{t-1}$ .

$k$ -period simple return:  $R_t[k] = S_t/S_{t-k} - 1$ . Note that

$$R_t[k] = \prod_{j=0}^{k-1} (1 + R_{t-j}) - 1 \approx R_t + \cdots + R_{t-k+1}.$$

time  $t$ .

One-period simple return:  $R_t = (S_t - S_{t-1})/S_{t-1}$ .

$k$ -period simple return:  $R_t[k] = S_t/S_{t-k} - 1$ . Note that

$$R_t[k] = \prod_{j=0}^{k-1} (1 + R_{t-j}) - 1 \approx R_t + \cdots + R_{t-k+1}.$$

The **annualized return** is defined by

$$\left\{ \prod_{j=0}^{k-1} (1 + R_{t-j}) \right\}^{1/k} - 1 \approx k^{-1} \sum_{j=0}^{k-1} R_{t-j}.$$

Continuously compounding:

$$A_t = C \left(1 + \frac{r}{m}\right)^{mt} \xrightarrow{m \rightarrow \infty} C \exp(rt).$$

log-return:  $r_t = \log(S_t/S_{t-1}) = \log S_t - \log S_{t-1}$ . It is easy to see

$$r_t = \log(1 + R_t) \approx R_t.$$

The  $k$ -period return is defined as  $r_t[k] = \log(S_t/S_{t-k})$ , which satisfies

$$r_t[k] = r_t + r_{t-1} + \cdots + r_{t-k+1}.$$

Most of the returns in the class refer to the log-returns.

log-return:  $r_t = \log(S_t/S_{t-1}) = \log S_t - \log S_{t-1}$ . It is easy to see

$$r_t = \log(1 + R_t) \approx R_t.$$

The  $k$ -period return is defined as  $r_t[k] = \log(S_t/S_{t-k})$ , which satisfies

$$r_t[k] = r_t + r_{t-1} + \cdots + r_{t-k+1}.$$

Most of the returns in the class refer to the log-returns.

Returns with dividend payments:

$$R_t = \frac{S_t + D_t}{S_{t-1}} - 1; \quad r_t = \log \frac{S_t + D_t}{S_{t-1}},$$

where  $D_t$  is the dividend payment during the period.

Excess returns: difference between the asset's return and the return on some reference assets. Reference assets include

— **riskless**: 3-month US Treasury bills as a proxy;

— **market portfolio**: SP500 index or CRSP (center for research in security prices) value-weighted index as a proxy.

— **market portfolio**: SP500 index or CRSP (center for research in security prices) value-weighted index as a proxy.

## 2.2 Stationarity

Linear time series techniques are frequently employed to forecast the returns of financial assets. An important concept is the stationarity, which is about **structural invariability** across time so that historical relationship can be aggregated.

**Definition:**  $\{X_t : t = 0, \pm 1, \pm 2, \dots\}$  is (weak) stationary if

(i)  $EX_t = \mu,$

(ii)  $\text{Cov}(X_t, X_{t+k}) = \gamma(k),$  independent of  $t,$  — called ACF.

Thus,  $\gamma(0) = \text{var}(X_t).$

The correlation between  $X_t$  and  $X_{t+k}$  **does not change** over time, in addition to having constant mean and variance.

**Definition:**  $\{X_t : t = 0, \pm 1, \pm 2, \dots\}$  is (weak) stationary if

(i)  $EX_t = \mu,$

(ii)  $\text{Cov}(X_t, X_{t+k}) = \gamma(k),$  independent of  $t,$  — called ACF.

Thus,  $\gamma(0) = \text{var}(X_t).$

The correlation between  $X_t$  and  $X_{t+k}$  **does not change** over time, in addition to having constant mean and variance.

**Definition:**  $\{X_t : t = 0, \pm 1, \pm 2, \dots\}$  is strict (strong) stationary if the distribution of  $(X_1, \dots, X_k)$  and  $(X_{t+1}, \dots, X_{t+k})$  are the same for any  $k$  and  $t.$

## Use of stationarity: **Strong stationarity** $\implies$

— Weak stationarity (if moments exist)

—  $E(X_{t+h} | X_t, \dots, X_{t-k}) = g(X_t, \dots, X_{t-k})$

$$E(X_{1+h} | X_1, \dots, X_{1-k}) = g(X_1, \dots, X_{1-k})$$

♠ **nonlinear relationship remains stable.**

—  $\text{var}(X_{t+h} | X_t, \dots, X_{t-k}) = \sigma^2(X_t, \dots, X_{t-k})$

$$\text{var}(X_{1+h} | X_1, \dots, X_{1-k}) = \sigma^2(X_1, \dots, X_{1-k}).$$

## Use of stationarity: **Strong stationarity** $\implies$

- Weak stationarity (if moments exist)
- $E(X_{t+h}|X_t, \dots, X_{t-k}) = g(X_t, \dots, X_{t-k})$   
 $E(X_{1+h}|X_1, \dots, X_{1-k}) = g(X_1, \dots, X_{1-k})$
- ♠ **nonlinear relationship remains stable.**
- $\text{var}(X_{t+h}|X_t, \dots, X_{t-k}) = \sigma^2(X_t, \dots, X_{t-k})$   
 $\text{var}(X_{1+h}|X_1, \dots, X_{1-k}) = \sigma^2(X_1, \dots, X_{1-k})$ .

The functions  $g(\cdot)$  and  $\sigma^2(\cdot)$  are independent of  $t$ .

**Weak stationarity**  $\implies$  **linear relationship remains stable**:

$$E(X_{t+h} - \beta_0 - \beta_1 X_t - \dots - \beta_{k+1} X_{t-k})^2 = \text{MSE of linear prediction.}$$

- MSE is independent of  $t$  and depends only ACF;

- best linear prediction rule independent of  $t$ ;
- first two-moments are sufficient for linear time series.

### 2.3 Autocorrelation Function

Autocorrelation Function is defined as

$$\begin{aligned}\rho(k) &= \text{Corr}(X_{t+k}, X_t) = \frac{\text{Cov}(X_{t+k}, X_t)}{\sqrt{\text{Var}(X_{t+k})} \sqrt{\text{Var}(X_t)}} \\ &= \frac{\gamma(k)}{\gamma(0)} \quad \text{for stationary time series}\end{aligned}$$

- best linear prediction rule independent of  $t$ ;
- first two-moments are sufficient for linear time series.

### 2.3 Autocorrelation Function

Autocorrelation Function is defined as

$$\begin{aligned}\rho(k) &= \text{Corr}(X_{t+k}, X_t) = \frac{\text{Cov}(X_{t+k}, X_t)}{\sqrt{\text{Var}(X_{t+k})} \sqrt{\text{Var}(X_t)}} \\ &= \frac{\gamma(k)}{\gamma(0)} \quad \text{for stationary time series}\end{aligned}$$

- ♠ measures linear strength of  $X_t$  and  $X_{t+k}$ ;
- ♠  $-1 \leq \rho(k) \leq 1$  and is even  $\rho(k) = \rho(-k)$ ;

♠ is a semi-positive definite function: for any  $a_i$

$$\sum_{i=1}^k \sum_{j=1}^k \rho(i-j) a_i a_j = \gamma(0)^{-1} \text{var} \left( \sum_{i=1}^k a_i X_{t-i} \right) \geq 0.$$

### Estimate of ACVF:

$$\begin{aligned} \hat{\gamma}(k) &= \frac{1}{T} \sum_{t=1}^{T-k} (X_t - \bar{X})(X_{t+k} - \bar{X}), \quad k = 0, 1, \dots, T-1, \\ &\cong \text{sample COV of } \{(X_t, X_{t+k}), t = 1, \dots, T-k\}, \end{aligned}$$

where  $\bar{X} = \sum_{t=1}^T X_t / T$ .

♠ is a semi-positive definite function: for any  $a_i$

$$\sum_{i=1}^k \sum_{j=1}^k \rho(i-j) a_i a_j = \gamma(0)^{-1} \text{var} \left( \sum_{i=1}^k a_i X_{t-i} \right) \geq 0.$$

### Estimate of ACVF:

$$\begin{aligned} \hat{\gamma}(k) &= \frac{1}{T} \sum_{t=1}^{T-k} (X_t - \bar{X})(X_{t+k} - \bar{X}), \quad k = 0, 1, \dots, T-1, \\ &\cong \text{sample COV of } \{(X_t, X_{t+k}), t = 1, \dots, T-k\}, \end{aligned}$$

where  $\bar{X} = \sum_{t=1}^T X_t / T$ .

**Estimated ACF:**  $\hat{\rho}(k) = \hat{\gamma}(k) / \hat{\gamma}(0)$ .

- impossible to estimate  $\hat{\rho}(k)$  or  $\hat{\gamma}(k)$  for  $k \geq T$ ;
- $\hat{\rho}(k)$  and  $\hat{\gamma}(k)$  can not be estimated accurately for large  $k$ ;
- rule of thumb:  $T \geq 50$  and  $k \leq T/4$ .

**MA( $q$ )-model:**  $X_t = \mu + \sum_{j=1}^q a_j \varepsilon_{t-j} + \varepsilon_t, \{\varepsilon_t\} \sim \text{IID}(0, \sigma^2).$

ACF plays an active role in model identification:

e.g. MA(2) model  $X_t = \mu + \varepsilon_t + a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2}$ . Then

$$X_{t-3} = \mu + \varepsilon_{t-3} + a_1 \varepsilon_{t-4} + a_2 \varepsilon_{t-5}$$

Hence,  $\rho(3) = \rho(4) = \dots = 0$ . But the sample ACF will not be exactly zero. What is the **confidence limit**?

**MA( $q$ )-model:**  $X_t = \mu + \sum_{j=1}^q a_j \varepsilon_{t-j} + \varepsilon_t, \{\varepsilon_t\} \sim \text{IID}(0, \sigma^2).$

ACF plays an active role in model identification:

e.g. MA(2) model  $X_t = \mu + \varepsilon_t + a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2}$ . Then

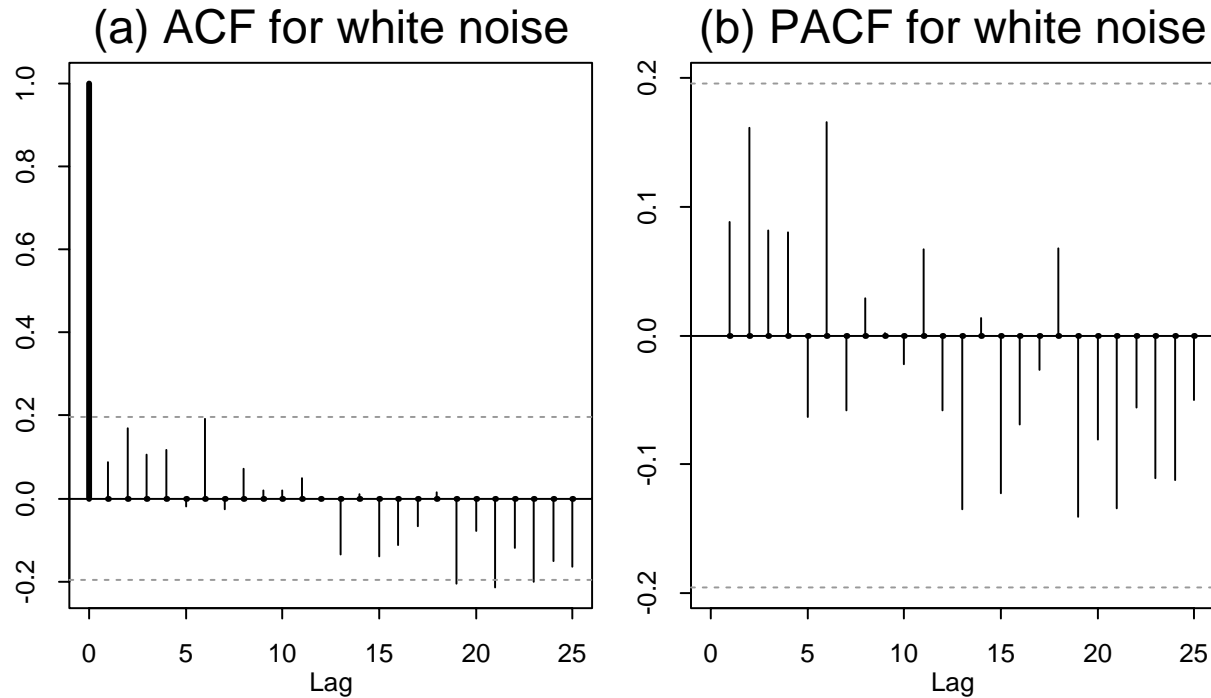
$$X_{t-3} = \mu + \varepsilon_{t-3} + a_1 \varepsilon_{t-4} + a_2 \varepsilon_{t-5}$$

Hence,  $\rho(3) = \rho(4) = \dots = 0$ . But the sample ACF will not be exactly zero. What is the **confidence limit**?

**Theorem 1** *Let  $X_t$  follow an MA( $q$ ) model. Then*

$$\sqrt{T} \hat{\rho}(j) \xrightarrow{D} \mathcal{N}\left(0, 1 + 2 \sum_{i=1}^q \rho^2(i)\right), \quad j > q$$

(see Theorem 2.8 of Fan and Yao).

Figure 2.1: ACF and PACV for a simulated white noise series with  $T=100$ .

Thus, for  $j > q$ , we expect that 95% of sample correlations  $\hat{\rho}(j)$  fall in the interval

$$\pm \frac{1.96}{\sqrt{T}} \left\{ 1 + 2 \sum_{i=1}^q \hat{\rho}^2(i) \right\}^{1/2}.$$

## 2.4 Predicability of asset returns

Predicability of asset returns is fundamental and stated in several forms, depending on the mathematical requirement. One of the forms is the **I.I.D. random walk hypothesis** for the asset returns  $\{X_t\}$ :

$$\{X_t\} \sim \text{IID}(0, \sigma^2).$$

## 2.4 Predicability of asset returns

Predicability of asset returns is fundamental and stated in several forms, depending on the mathematical requirement. One of the forms is the **I.I.D. random walk hypothesis** for the asset returns  $\{X_t\}$ :

$$\{X_t\} \sim \text{IID}(0, \sigma^2).$$

Another is that the **returns are uncorrelated**. In both cases,

$$\sqrt{T}\hat{\rho}(j) \xrightarrow{D} \mathcal{N}(0, 1), \quad \text{for } j \neq 0.$$

$\implies$  There is 95% chance that  $\hat{\rho}(j)$  falls in  $1.96T^{-1/2}$ .

Tests for (uncorrelated) random walk hypothesis:

$$H_0 : \rho(j) = 0, \quad \forall j \neq 0.$$

**Portmanteau statistic** (Box and Pierce 1970):  $Q^*(m) = T \sum_{j=1}^m \hat{\rho}^2(j)$ .

**Ljung and Box (1978)**:  $Q(m) = T(T + 2) \sum_{j=1}^m \frac{\hat{\rho}^2(j)}{T - j}$ .

- Choice of  $m$ :  $m \approx \log T$ .
- See §7.4 of Fan and Yao (2003) for more modern tests.
- Under the white noise hypothesis, both test statistics follow asymptotically  $\chi_m^2$ -distribution.

**Portmanteau statistic** (Box and Pierce 1970):  $Q^*(m) = T \sum_{j=1}^m \hat{\rho}^2(j)$ .

**Ljung and Box (1978)**:  $Q(m) = T(T + 2) \sum_{j=1}^m \frac{\hat{\rho}^2(j)}{T - j}$ .

- Choice of  $m$ :  $m \approx \log T$ .
- See §7.4 of Fan and Yao (2003) for more modern tests.
- Under the white noise hypothesis, both test statistics follow asymptotically  $\chi_m^2$ -distribution.

**Example 1.** Consider the monthly log-returns for IBM stocks from January 1926 to December 1997. It can be calculated that

$$Q(5) = 5.8, \quad d.f. = 5, \quad \text{p-value} = 0.33;$$

$$Q(10) = 13.7, \quad d.f. = 10, \quad \text{p-value} = 0.19 .$$

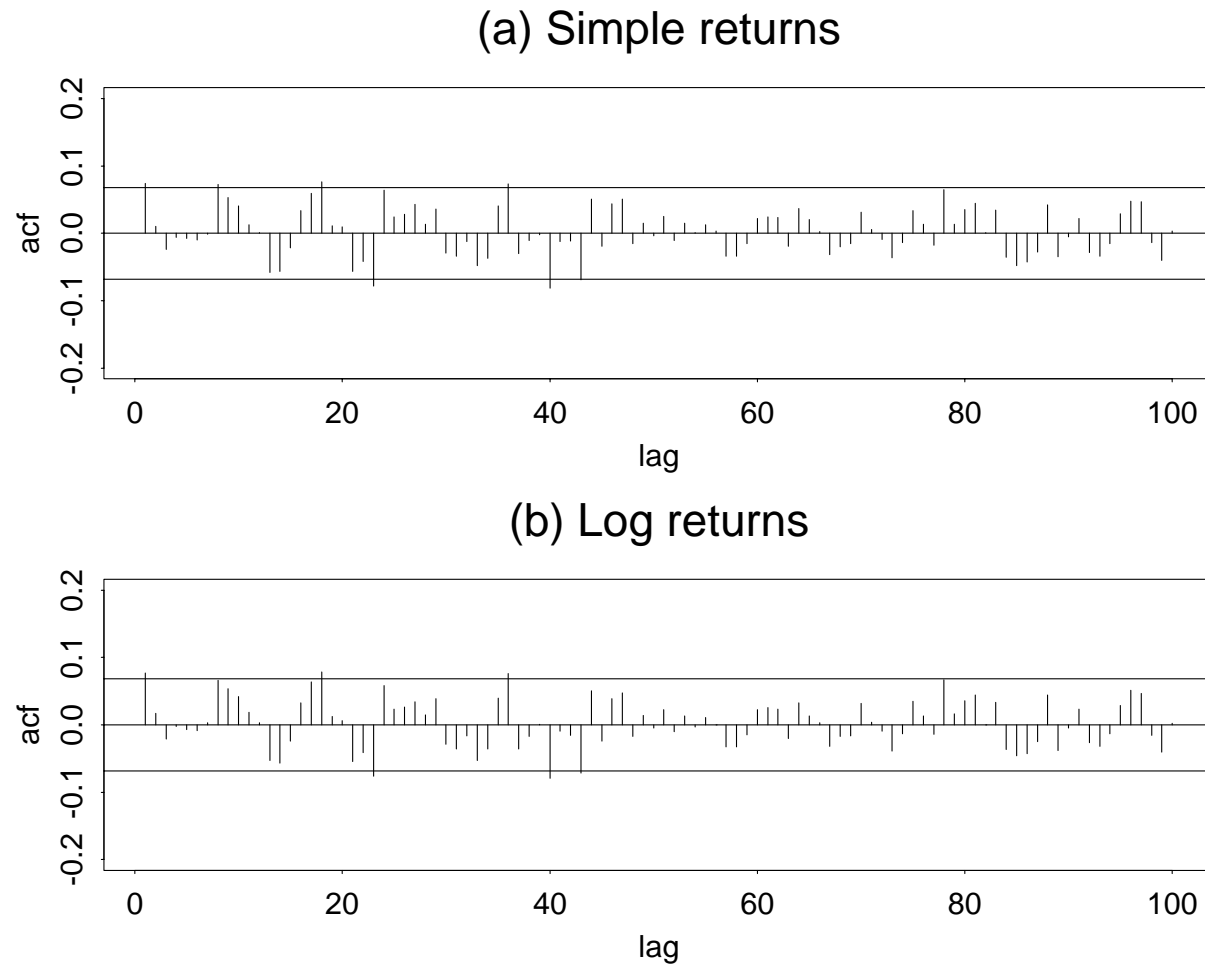


Figure 2.2: ACFs for the monthly simple- and log-returns of IBM stocks.

No evidence against the null hypothesis: random walk.

For the monthly log-return of CRSP value-weighted index in the same period, we have

$$Q(5) = 26.9, \quad d.f. = 5, \quad \text{p-value} \leq 0.03\%,$$

$$Q(10) = 32.7, \quad d.f. = 10, \quad \text{p-value} \leq 0.03\%.$$

Strong evidence against the random walk hypothesis.

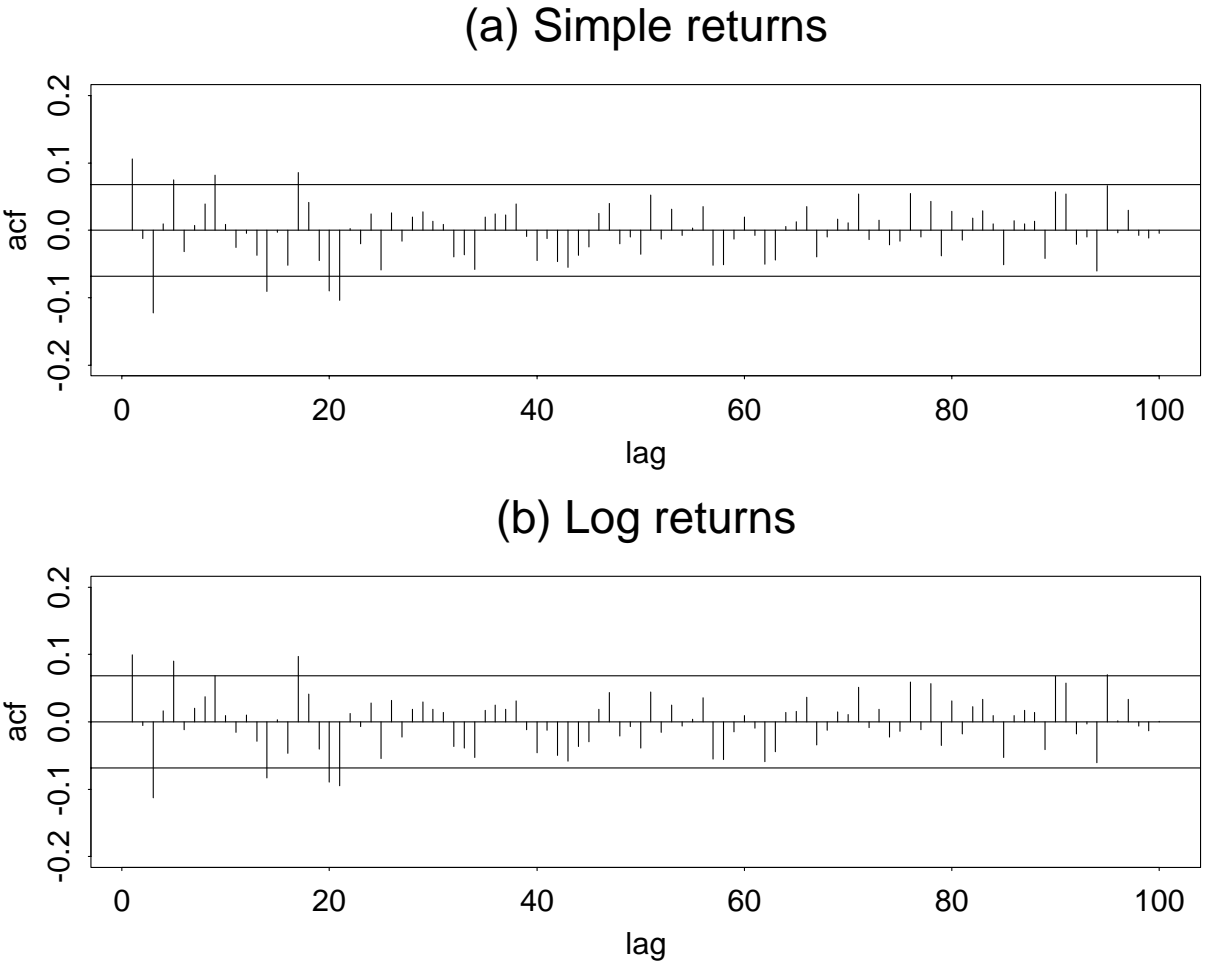


Figure 2.3: ACFs for the monthly simple- and log-returns of CRSP value-weighted index.

## 2.5 Autoregressive model

— is a simple and useful class of models for forecasting the returns of assets.

## 2.5 Autoregressive model

— is a simple and useful class of models for forecasting the returns of assets.

$$\begin{aligned}\text{AR}(p)\text{-model: } X_t &= b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t \\ &= b_0 + (b_1 B + \cdots + b_p B^p) X_t + \varepsilon_t,\end{aligned}$$

where  $B$  is

$$\text{Backshift operator: } B^k X_t = X_{t-k}, k = \pm 1, \pm 2, \cdots.$$

## 2.5 Autoregressive model

— is a simple and useful class of models for forecasting the returns of assets.

$$\begin{aligned} \text{AR}(p)\text{-model: } X_t &= b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t \\ &= b_0 + (b_1 B + \cdots + b_p B^p) X_t + \varepsilon_t, \end{aligned}$$

where  $B$  is

$$\text{Backshift operator: } B^k X_t = X_{t-k}, k = \pm 1, \pm 2, \dots$$

Suppose that  $X_t$  is stationary with mean  $\mu$ . Then

$$\mu = b_0 + (b_1 + \cdots + b_p)\mu \quad \implies \quad \mu = \frac{b_0}{1 - b_1 - \cdots - b_p}.$$

Stationarity: Let  $b(z) = 1 - b_1 z - \cdots - b_p z^p$  be the characteristic

function. Then

$$b(B)X_t = b_0 + \varepsilon_t \quad \Longrightarrow \quad X_t = b(B)^{-1}(b_0 + \varepsilon_t).$$

The invertibility requires some technical conditions:

**$b(z) = 0$  has roots outside the unit circle,**

i.e.  $|b(z)| > 0$  for  $|z| \leq 1$ . If so,

$$b(z)^{-1} = \sum_{j=0}^{\infty} c_j z^j$$

function. Then

$$b(B)X_t = b_0 + \varepsilon_t \quad \Longrightarrow \quad X_t = b(B)^{-1}(b_0 + \varepsilon_t).$$

The invertibility requires some technical conditions:

**$b(z) = 0$  has roots outside the unit circle,**

i.e.  $|b(z)| > 0$  for  $|z| \leq 1$ . If so,

$$b(z)^{-1} = \sum_{j=0}^{\infty} c_j z^j$$

and

$$\begin{aligned} X_t &= \sum_{j=0}^{\infty} c_j B^j (b_0 + \varepsilon_t) \\ &= \sum_{j=0}^{\infty} c_j (b_0 + \varepsilon_{t-j}). \quad \longleftarrow \text{Strong stationarity} \\ &\quad \longleftarrow \text{MA of infinity order} \end{aligned}$$

**Theorem 2** (FY Theorem 2.1): An AR( $p$ ) process is stationary if

$$\inf_{|z| \leq 1} |b(z)| > 0.$$

**Example 2.** For the AR(1)-model:  $X_t = bX_{t-1} + \varepsilon_t$ , the characteristic function is  $b(z) = 1 - bz$  and has a root  $1/b$ . Thus, when  $|b| < 1$ , the series is stationary. Now consider the following AR(3) model:

$$X_t = 1.8X_{t-1} - 1.05X_{t-2} + 0.2X_{t-3} + \varepsilon_t.$$

**Theorem 2** (FY Theorem 2.1): An AR( $p$ ) process is stationary if

$$\inf_{|z| \leq 1} |b(z)| > 0.$$

**Example 2.** For the AR(1)-model:  $X_t = bX_{t-1} + \varepsilon_t$ , the characteristic function is  $b(z) = 1 - bz$  and has a root  $1/b$ . Thus, when  $|b| < 1$ , the series is stationary. Now consider the following AR(3) model:

$$X_t = 1.8X_{t-1} - 1.05X_{t-2} + 0.2X_{t-3} + \varepsilon_t.$$

The characteristic function

$$\begin{aligned} b(z) &= 1 - 1.8z + 1.05z^2 - 0.2z^3 \\ &= (1 - 0.8z)(1 - 0.5z)^2, \end{aligned}$$

which has a root outside the unit circle. Thus, it is stationary.

Autocorrelation of AR( $p$ ): Since  $X_{t-k}$  depends only on  $\varepsilon_{t-k}$  and its past,

$$\text{Cov}(b(B)X_t, X_{t-k}) = \text{Cov}(b_0 + \varepsilon_t, X_{t-k}) = 0, \quad \text{for } k > 0$$

$$\implies \gamma(k) - b_1\gamma(k-1) - \dots - b_p\gamma(k-p) = 0 \quad \forall k > 0.$$

This can be written as

$$b(B)\gamma(k) = (B - z_1) \cdots (B - z_p)\gamma(k) = 0, \quad \text{for } k > 0,$$

where  $z_1, \dots, z_p$  are the roots of  $b(z)$ .

**Autocorrelation of AR( $p$ ):** Since  $X_{t-k}$  depends only on  $\varepsilon_{t-k}$  and its past,

$$\text{Cov}(b(B)X_t, X_{t-k}) = \text{Cov}(b_0 + \varepsilon_t, X_{t-k}) = 0, \quad \text{for } k > 0$$

$$\implies \gamma(k) - b_1\gamma(k-1) - \dots - b_p\gamma(k-p) = 0 \quad \forall k > 0.$$

This can be written as

$$b(B)\gamma(k) = (B - z_1) \cdots (B - z_p)\gamma(k) = 0, \quad \text{for } k > 0,$$

where  $z_1, \dots, z_p$  are the roots of  $b(z)$ .

**Solution:** The above difference equation is similar to that of AR( $p$ )

model and the solution is of form:

$$\gamma(k) = \alpha_1 z_1^{-k} + \dots + \alpha_p z_p^{-k}.$$

Hence,  $\rho(k) \rightarrow 0$  exponentially fast. (FY: Prop 2.2). For example, for AR(1):  $X_t = bX_{t-1} + \varepsilon_t$ ,  $\rho(k) = b^{|k|}$ .

Hence,  $\rho(k) \rightarrow 0$  exponentially fast. (FY: Prop 2.2). For example, for AR(1):  $X_t = bX_{t-1} + \varepsilon_t$ ,  $\rho(k) = b^{|k|}$ .

ACF helps us to identify a model. Figure 2.4 shows ACF for AR(1) with  $b = 0.7$  or  $-0.7$ . Figure 2.5 shows ACF for

- AR(4):  $X_t = 0.5X_{t-1} + 0.3X_{t-2} - 0.7X_{t-3} + 0.2X_{t-4} + \varepsilon_t$ ,
- MA(4):  $X_t = \varepsilon_t + 0.6\varepsilon_{t-1} + 0.6\varepsilon_{t-2} + 0.3\varepsilon_{t-3} + 0.7\varepsilon_{t-4}$ ,
- ARMA(2,2):  $X_t = 0.8X_{t-1} - 0.6X_{t-2} + \varepsilon_t + 0.7\varepsilon_{t-1} + 0.4\varepsilon_{t-2}$ .

Hence,  $\rho(k) \rightarrow 0$  exponentially fast. (FY: Prop 2.2). For example, for AR(1):  $X_t = bX_{t-1} + \varepsilon_t$ ,  $\rho(k) = b^{|k|}$ .

ACF helps us to identify a model. Figure 2.4 shows ACF for AR(1) with  $b = 0.7$  or  $-0.7$ . Figure 2.5 shows ACF for

- AR(4):  $X_t = 0.5X_{t-1} + 0.3X_{t-2} - 0.7X_{t-3} + 0.2X_{t-4} + \varepsilon_t$ ,
- MA(4):  $X_t = \varepsilon_t + 0.6\varepsilon_{t-1} + 0.6\varepsilon_{t-2} + 0.3\varepsilon_{t-3} + 0.7\varepsilon_{t-4}$ ,
- ARMA(2,2):  $X_t = 0.8X_{t-1} - 0.6X_{t-2} + \varepsilon_t + 0.7\varepsilon_{t-1} + 0.4\varepsilon_{t-2}$ .

## Model identification and partial autocorrelation

MA( $q$ ) model has a distinct rubric:  $\rho(k) = 0, \forall k > q$ . Do we have a similar rubric? This is accomplished by the partial autocovariance function  $\pi(\cdot)$  defined below.

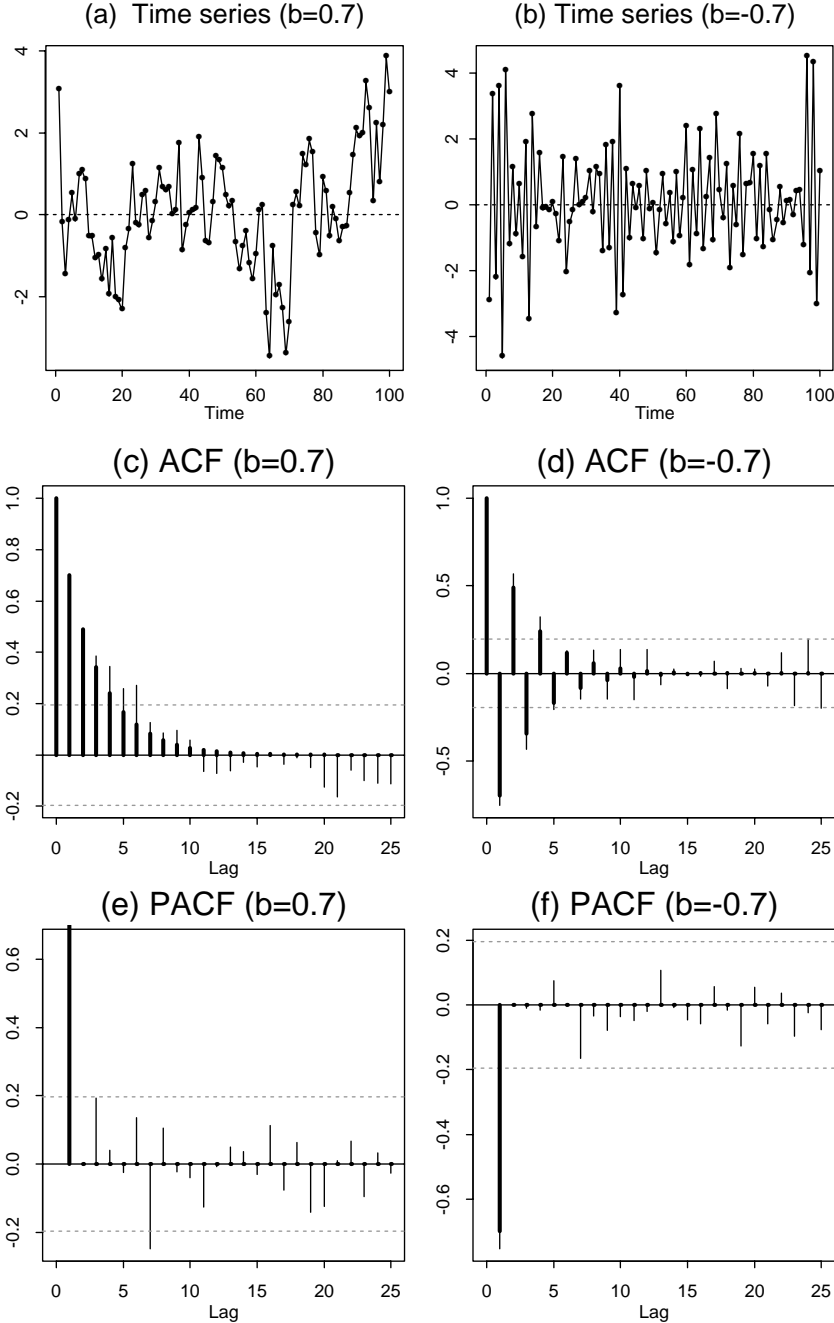


Figure 2.4: Simulated time series and their ACFs and PACFs, T= 100.

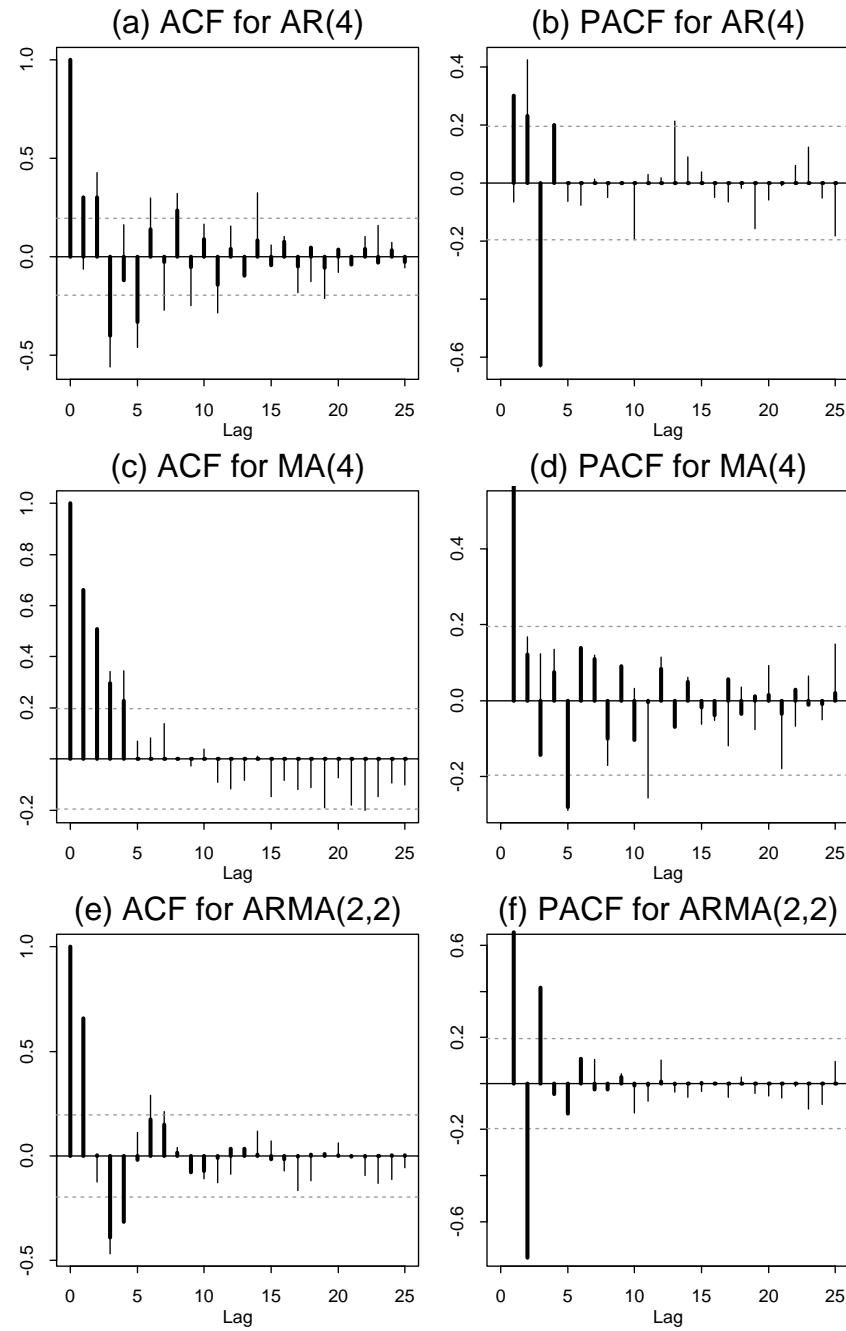


Figure 2.5: Simulated time series and their ACFs and PACFs,  $T=100$ .

**PACF**:  $\pi(1) = \text{Corr}(X_1, X_2) = \rho(1)$ ; and for  $k \geq 2$ ,

$$\pi(k) = \text{Corr}(R_{1|2,\dots,k}, R_{k+1|2,\dots,k}),$$

where  $R_{j|2,\dots,k}$  = residual of  $X_j$  on  $(X_2, \dots, X_k)$ .

**Proposition** (*FY Thm. 2.9*):  $\pi(k) = b_{kk}$  where  $(b_{k1}, \dots, b_{kk})$  minimizes

$$E(X_t - \beta_0 - \beta_1 X_{t-1} - \dots - \beta_k X_{t-k})^2.$$

**PACF**:  $\pi(1) = \text{Corr}(X_1, X_2) = \rho(1)$ ; and for  $k \geq 2$ ,

$$\pi(k) = \text{Corr}(R_{1|2,\dots,k}, R_{k+1|2,\dots,k}),$$

where  $R_{j|2,\dots,k}$  = residual of  $X_j$  on  $(X_2, \dots, X_k)$ .

**Proposition** (*FY Thm. 2.9*):  $\pi(k) = b_{kk}$  where  $(b_{k1}, \dots, b_{kk})$  minimizes

$$E(X_t - \beta_0 - \beta_1 X_{t-1} - \dots - \beta_k X_{t-k})^2.$$

Now, for an AR(p) model

$$X_t = b_0 + b_1 X_{t-1} + \dots + b_p X_{t-p} + \varepsilon_t,$$

It is clear for  $k > p$ ,

$$b_{k1} = b_1, \dots, b_{kp} = b_p, \quad b_{k(p+1)} = \dots = b_{kk} = 0.$$

Hence,

$$\pi(k) = 0 \quad \forall k > p.$$

**Proposition** (*FY Prop. 3.1*): If  $\pi(k)$  is estimated by the least-squares method below and if the series  $\{X_t\}$  follows an AR( $p$ ) model with i.i.d. white noise, then

$$T^{1/2}\hat{\pi}(k) \xrightarrow{D} N(0, 1), \text{ for } k > p.$$

Hence,

$$\pi(k) = 0 \quad \forall k > p.$$

**Proposition** (*FY Prop. 3.1*): If  $\pi(k)$  is estimated by the least-squares method below and if the series  $\{X_t\}$  follows an AR( $p$ ) model with i.i.d. white noise, then

$$T^{1/2}\hat{\pi}(k) \xrightarrow{D} N(0, 1), \text{ for } k > p.$$

For the monthly returns of the CRSP value-weighted index, it can be computed that

p	1	2	3	4	5	6	7	8	9	10
PACF	0.11	-0.02	-0.12	0.04	0.07	-0.06	0.02	0.06	0.06	-0.01
AIC	-5.807	-5.805	-5.817	-5.816	-5.819	-5.821	-5.819	-5.820	-5.821	-5.818

$T = 864$  (Jan., 1926 — Dec. 1997). The standard error under AR( $p$ ) model is  $SE = \frac{1}{\sqrt{T}} = 0.034$ ,  $2SE = 0.068$ ,  $\hat{p} = 3$  or  $5$ .

---

p	1	2	3	4	5	6	7	8	9	10
PACF	0.11	-0.02	-0.12	0.04	0.07	-0.06	0.02	0.06	0.06	-0.01
AIC	-5.807	-5.805	-5.817	-5.816	-5.819	-5.821	-5.819	-5.820	-5.821	-5.818

---

$T = 864$  (Jan., 1926 — Dec. 1997). The standard error under AR( $p$ ) model is  $SE = \frac{1}{\sqrt{T}} = 0.034$ ,  $2SE = 0.068$ ,  $\hat{p} = 3$  or 5.

## Parameter estimation

The AR coefficient is estimated by the least-squares (pseudo MLE) or the maximum likelihood method:

$$\sum_{t=p+1}^T (X_t - b_0 - b_1 X_{t-1} - \cdots - b_p X_{t-p})^2.$$

---

p	1	2	3	4	5	6	7	8	9	10
PACF	0.11	-0.02	-0.12	0.04	0.07	-0.06	0.02	0.06	0.06	-0.01
AIC	-5.807	-5.805	-5.817	-5.816	-5.819	-5.821	-5.819	-5.820	-5.821	-5.818

---

$T = 864$  (Jan., 1926 — Dec. 1997). The standard error under AR( $p$ ) model is  $SE = \frac{1}{\sqrt{T}} = 0.034$ ,  $2SE = 0.068$ ,  $\hat{p} = 3$  or  $5$ .

## Parameter estimation

The AR coefficient is estimated by the least-squares (pseudo MLE) or the maximum likelihood method:

$$\sum_{t=p+1}^T (X_t - b_0 - b_1 X_{t-1} - \cdots - b_p X_{t-p})^2.$$

This is the same as the linear regression model

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_p + \varepsilon,$$

where at time  $t - 1$ ,  $Y = X_t$ ,  $X_1 = X_{t-1}, \dots, X_p = X_{t-p}$ .

where at time  $t - 1$ ,  $Y = X_t, X_1 = X_{t-1}, \dots, X_p = X_{t-p}$ .

The data are of form

$Y$	$X_1$	$\dots$	$X_p$
$X_{p+1}$	$X_p$	$\dots$	$X_1$
$X_{p+2}$	$X_{p+1}$	$\dots$	$X_2$
	$\dots$	$\dots$	
$X_T$	$X_{T-1}$	$\dots$	$X_{T-p}$

**Residuals**:  $\hat{\varepsilon}_t = X_t - \hat{b}_0 - \hat{b}_1 X_{t-1} - \dots - \hat{b}_p X_{t-p}$

♣ provide raw materials for model checking; ♣ ideal fit if the residual series behaves like white noise series.

**Residual variance**:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{T - 2p - 1} = \frac{\sum_{t=p+1}^T \hat{\varepsilon}_t^2}{T - 2p - 1}$$

For the CRSP value-weight index, fitting an AR(3) results in

$$r_t = 0.0103 + 0.104r_{t-1} - 0.010r_{t-2} - 0.120r_{t-3} + \varepsilon_t$$

SE	0.002	0.034	0.034	0.034
----	-------	-------	-------	-------

For the CRSP value-weight index, fitting an AR(3) results in

$$\begin{array}{rcccc} r_t & = & 0.0103 & + & 0.104r_{t-1} & - & 0.010r_{t-2} & - & 0.120r_{t-3} & + & \varepsilon_t \\ \text{SE} & & 0.002 & & 0.034 & & 0.034 & & 0.034 & & \end{array}$$

For example, for testing whether the intercept  $b_0$  (or the mean return) is zero, we compute the t-statistic  $0.0103/0.002 = 5$  and hence its associated P-value is nearly 0. In other words, we have very strong evidence against the hypothesis  $H_0 : b_0 = 0$ , namely, the monthly returns are statistically positive.

### Remarks:

— Series are weakly dependent (small  $b$ 's).

—  $b_0$  is significantly positive. The expect return is

$$\hat{\mu} = \frac{\hat{b}_0}{1 - \hat{b}_1 - \hat{b}_2 - \hat{b}_3} = 0.01 \quad (\text{monthly return})$$

Annualized expected return =  $(1 + 0.01)^{12} - 1 \approx 12.6\%$ .

—  $b_0$  is significantly positive. The expect return is

$$\hat{\mu} = \frac{\hat{b}_0}{1 - \hat{b}_1 - \hat{b}_2 - \hat{b}_3} = 0.01 \quad (\text{monthly return})$$

Annualized expected return =  $(1 + 0.01)^{12} - 1 \approx 12.6\%$ .

Actual annualized return (1/1926—12/1997)

$$\left[ \prod_{t=1}^{864} (1 + R_t) \right]^{12/864} - 1 \approx 10.53\%.$$

— For checking the random walk of the residual series,

$$Q(10) = 15.8, \quad d.f. = 10 - 3 = 7, \quad \text{p-value} = 2.7\%.$$

Note that a small p-value does not imply the deviation is large. In fact, with  $T = 864$ , I consider the deviation is reasonable.

—  $b_0$  is significantly positive. The expect return is

$$\hat{\mu} = \frac{\hat{b}_0}{1 - \hat{b}_1 - \hat{b}_2 - \hat{b}_3} = 0.01 \quad (\text{monthly return})$$

Annualized expected return =  $(1 + 0.01)^{12} - 1 \approx 12.6\%$ .

Actual annualized return (1/1926—12/1997)

$$\left[ \prod_{t=1}^{864} (1 + R_t) \right]^{12/864} - 1 \approx 10.53\%.$$

— For checking the random walk of the residual series,

$$Q(10) = 15.8, \quad d.f. = 10 - 3 = 7, \quad \text{p-value} = 2.7\%.$$

Note that a small p-value does not imply the deviation is large. In fact, with  $T = 864$ , I consider the deviation is reasonable.

[Order Selection](#) ( §3.4, FY)

**Aim**: To select an order of  $AR(p)$  model to minimize an estimated PE. There are a few criteria for assessing prediction errors. See §1.5 and §3.4 of FY for further discussion.

**Aim**: To select an order of AR( $p$ ) model to minimize an estimated PE. There are a few criteria for assessing prediction errors. See §1.5 and §3.4 of FY for further discussion.

**Akaike information Criterion**:

- $AIC(p) = -2 (\max \log\text{-likelihood}) + 2(\text{No. parameters})$   
 $= (T - L) \log(\hat{\sigma}_p^2) + 2(p + 1) + \text{constant}, \quad \text{for } p = 0, 1, \dots, L,$   
 where  $\hat{\sigma}_p^2$  is the estimated residual variance. The first part measures the lack of fit and the second part **penalizes** the complexity of the model.
- $AICC(p) = (T - L) \log(\hat{\sigma}_p^2) + \frac{2(p+1)(T-L)}{T-L-p-2}$ , a correction of AIC.

**Aim**: To select an order of AR( $p$ ) model to minimize an estimated PE. There are a few criteria for assessing prediction errors. See §1.5 and §3.4 of FY for further discussion.

### **Akaike information Criterion**:

- $AIC(p) = -2 (\max \log\text{-likelihood}) + 2(\text{No. parameters})$   
 $= (T - L) \log(\hat{\sigma}_p^2) + 2(p + 1) + \text{constant}, \quad \text{for } p = 0, 1, \dots, L,$   
 where  $\hat{\sigma}_p^2$  is the estimated residual variance. The first part measures the lack of fit and the second part **penalizes** the complexity of the model.
- $AICC(p) = (T - L) \log(\hat{\sigma}_p^2) + \frac{2(p+1)(T-L)}{T-L-p-2}$ , a correction of AIC.

### **Bayesian information criterion**:

$$\begin{aligned} \text{BIC}(p) &= -2(\max \log\text{-likelihood}) + \log(T - L) \text{ (No. parameters)} \\ &= (T - L) \log(\hat{\sigma}_p^2) + (p + 1) \log(T - L) \quad \text{for } p = 0, 1, \dots, L. \end{aligned}$$

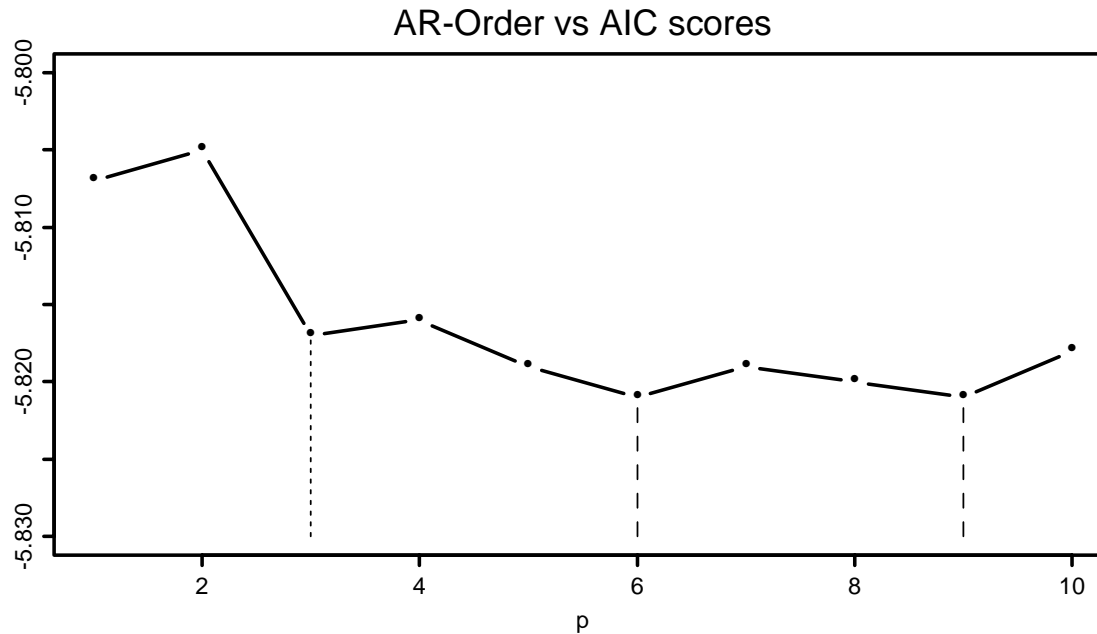


Figure 2.6: order versus AIC for the CRSP value weighted data

For CRSP value-weighted index,

$\hat{p} = 3$  — the point where AIC stops decreasing dramatically;

$\hat{p} = 6$  or  $9$  — minimum.

## Remarks:

- AIC usually selects too many parameters;
- BIC can select too few parameters;
- BIC and AICC penalize more heavily on **complexity** than AIC;

## 2.6 Prediction

Forecast future  $X_{T+m}$ ,  $T$  — forecast origin,  $m$  — forecast horizon.

Prediction error:  $= \mathbb{E}(X_{T+m} - f(X_T, X_{T-1}, \dots))^2$ .

**Best prediction rule**: find  $f$  to minimize the prediction error:

$$X_T(m) = \arg \inf_f \mathbb{E}(X_{T+m} - f)^2 = \mathbb{E}(X_{T+m} | X_1, \dots, X_T).$$

**Best prediction rule**: find  $f$  to minimize the prediction error:

$$X_T(m) = \arg \inf_f \mathbb{E}(X_{T+m} - f)^2 = \mathbb{E}(X_{T+m} | X_1, \dots, X_T).$$

The prediction is more precise, but the prediction rule is hard to estimate from the data. Hence, one can also limit the class of the prediction.

**Best linear prediction**:

$$X_T^L(m) = \hat{\beta}_{m,0} + \hat{\beta}_{m,1}X_T + \dots + \hat{\beta}_{m,p}X_{T+1-p},$$

where  $\hat{\beta}_{m,0}, \dots, \hat{\beta}_{m,p}$  minimize the expected prediction error

$$\mathbb{E}(X_{T+m} - \beta_0 - \beta_1X_T - \dots - \beta_pX_{T+1-p})^2.$$

This can be found by the least-squares method and the method depends only on the **weak stationarity**, not AR( $p$ ) assumption.

**Best prediction rule**: find  $f$  to minimize the prediction error:

$$X_T(m) = \arg \inf_f \mathbb{E}(X_{T+m} - f)^2 = \mathbb{E}(X_{T+m} | X_1, \dots, X_T).$$

The prediction is more precise, but the prediction rule is hard to estimate from the data. Hence, one can also limit the class of the prediction.

**Best linear prediction**:

$$X_T^L(m) = \hat{\beta}_{m,0} + \hat{\beta}_{m,1}X_T + \dots + \hat{\beta}_{m,p}X_{T+1-p},$$

where  $\hat{\beta}_{m,0}, \dots, \hat{\beta}_{m,p}$  minimize the expected prediction error

$$\mathbb{E}(X_{T+m} - \beta_0 - \beta_1X_T - \dots - \beta_pX_{T+1-p})^2.$$

This can be found by the least-squares method and the method depends only on the **weak stationarity**, not AR( $p$ ) assumption.

For the AR(p) model:

$$X_t = b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t,$$

the best one-step predictor

$$\begin{aligned} X_T(1) &= E(b_0 + b_1 X_T + \cdots + b_p X_{T+1-p} + \varepsilon_{t+1} | X_1, \dots, X_T) \\ &= b_0 + b_1 X_T + \cdots + b_p X_{T+1-p}, \end{aligned}$$

which coincides with the best linear predictor. In practice, one step-forecasting is

$$\hat{X}_T(1) = \hat{b}_0 + \hat{b}_1 X_T + \cdots + \hat{b}_p X_{T+1-p}.$$

**Size of prediction error**: Ignoring error in estimating coefficient, it is  $\sigma$ . There are statistical methods to account for that.

For the AR(p) model:

$$X_t = b_0 + b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t,$$

the best one-step predictor

$$\begin{aligned} X_T(1) &= E(b_0 + b_1 X_T + \cdots + b_p X_{T+1-p} + \varepsilon_{t+1} | X_1, \dots, X_T) \\ &= b_0 + b_1 X_T + \cdots + b_p X_{T+1-p}, \end{aligned}$$

which coincides with the best linear predictor. In practice, one step-forecasting is

$$\hat{X}_T(1) = \hat{b}_0 + \hat{b}_1 X_T + \cdots + \hat{b}_p X_{T+1-p}.$$

**Size of prediction error**: Ignoring error in estimating coefficient, it is  $\sigma$ . There are statistical methods to account for that.

2-step ahead prediction: Note that

$$X_{T+2} = b_0 + b_1 X_{T+1} + \cdots + b_p X_{T+2-p} + \varepsilon_{T+2}.$$

Hence, the two-step prediction is

$$X_T(2) = b_0 + b_1 X_T(1) + \cdots + b_p X_{T+2-p},$$

which is linear predictor and hence coincides with the best linear prediction. The 2-step prediction error is

$$e_T(2) = X_{T+2} - X_T(2) = \varepsilon_{T+2} + b_1 \varepsilon_{T+1}.$$

The size of the two-step prediction error is

$$\sqrt{\text{Var}(e_T(2))} = \sqrt{(1 + b_1^2)\sigma^2},$$

which is larger than one-step.

2-step ahead prediction: Note that

$$X_{T+2} = b_0 + b_1 X_{T+1} + \cdots + b_p X_{T+2-p} + \varepsilon_{T+2}.$$

Hence, the two-step prediction is

$$X_T(2) = b_0 + b_1 X_T(1) + \cdots + b_p X_{T+2-p},$$

which is linear predictor and hence coincides with the best linear prediction. The 2-step prediction error is

$$e_T(2) = X_{T+2} - X_T(2) = \varepsilon_{T+2} + b_1 \varepsilon_{T+1}.$$

The size of the two-step prediction error is

$$\sqrt{\text{Var}(e_T(2))} = \sqrt{(1 + b_1^2)\sigma^2},$$

which is larger than one-step.

## Multiple-step forecast: (FY, Prop 3.3)

$$X_T(m) = b_0 + b_1 X_T(m-1) + \cdots + b_p X_T(m-p)$$

with  $X_T(-k) = X_{T-k}$ ,  $k = 0, 1, \dots, p$ . According to Prop. 3.4, FY,

$$\sigma^2(m) \xrightarrow{m \rightarrow \infty} \text{Var}(X_t)$$

♠ long-term forecast approaches to its unconditional mean and the error is the same as marginal one.

♠ referred to as the mean reversion.

Example 3: Consider the CRSP value-weighted index, the following

AR(5) model is used to predict the monthly log-return

$$r_t = 0.0075 + 0.103r_{t-1} + 0.002r_{t-2} - 0.114r_{t-3} + 0.032r_{t-4} + 0.084r_{t-5} + \varepsilon_t$$

with  $\sigma = 0.054$ ,  $T = 858$ .

with  $\sigma = 0.054$ ,  $T = 858$ .

**Results**: The actual and forecasted one are summarized in the following table and Fig. 2.7.

Step	1	2	3	4	5	6
Forecast	0.0071	-0.0008	0.0086	0.0154	0.0141	0.0100
SE	0.0541	0.0545	0.0545	0.0549	0.0549	0.0550
Actual	0.0762	-0.0365	0.0580	-0.0341	0.0311	0.0183

**Rolling and Cross-validation\***:

- Does AR(3) predict better than MA(5)? (**How to compare?**)
- It is hard to believe that the price dynamic remains constant over a long time horizon. So one might decide to use more recent data (e.g. post-war data).
- The former requires more accurately assessing PE and the latter requires to fit a model within

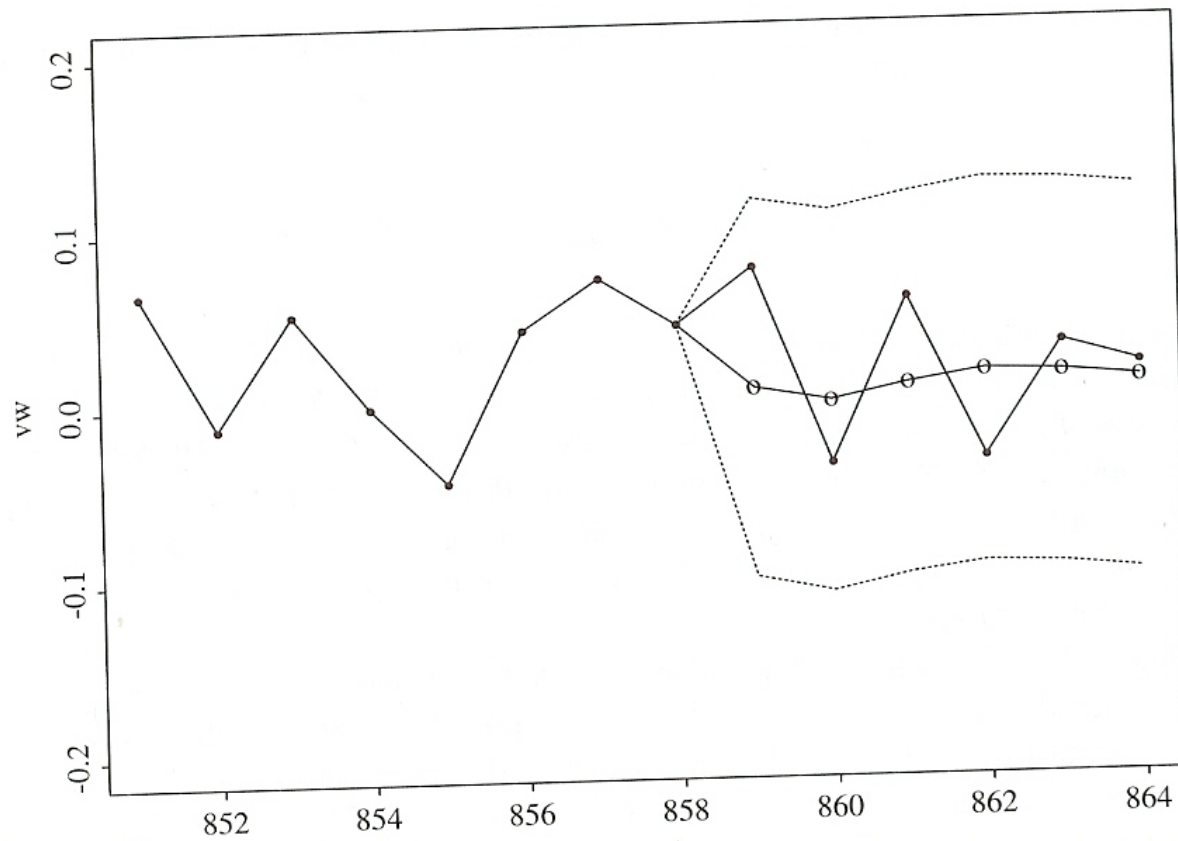


Figure 2.7: Out sample forecast

a smaller window (**How to choose window size?**).

— Clearly, there are needs for assessing prediction error.

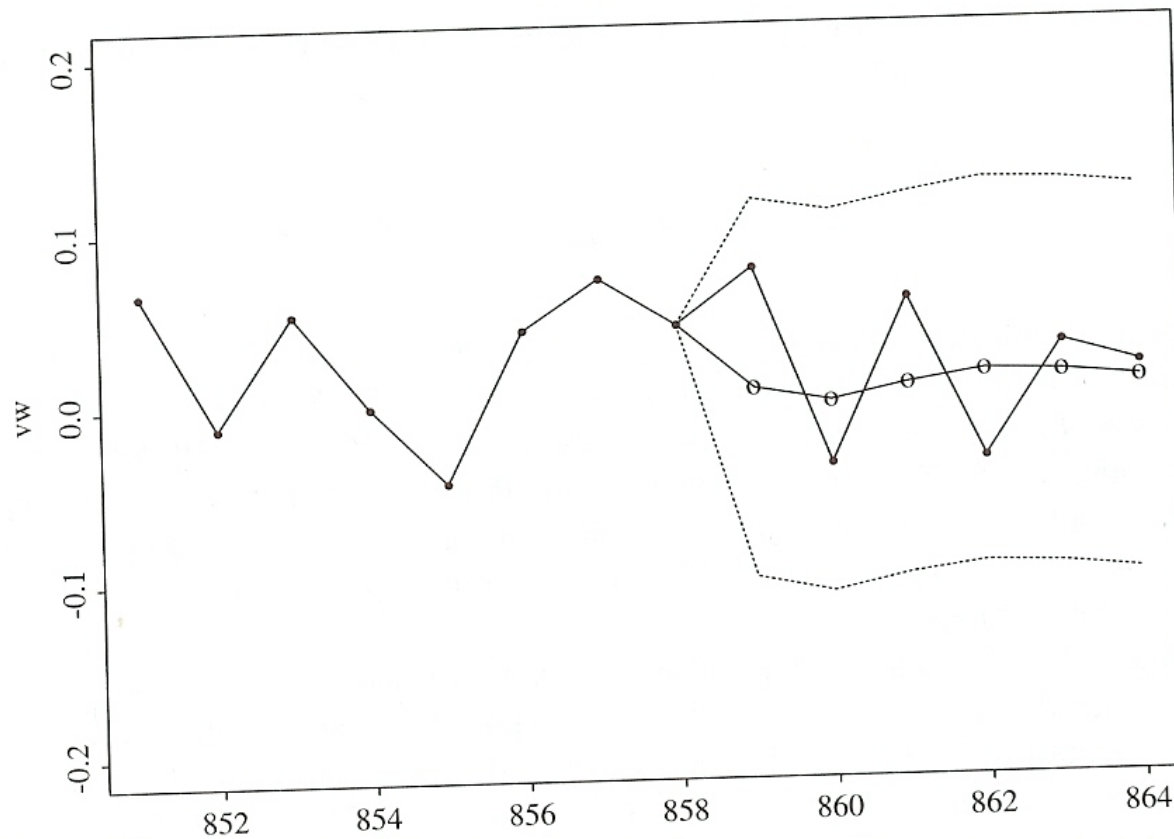


Figure 2.7: Out sample forecast

a smaller window (**How to choose window size?**).

— Clearly, there are needs for assessing prediction error.

**Assessing Prediction Errors\***: The simplest method is to use the first  $n - m$  data points to learn about the model (estimating parameters) and use the last  $m$  data to compute the prediction

error. Here, both **static** (parameters are not updated) or **rolling technique** (parameters are re-estimated as prediction along) can be used.

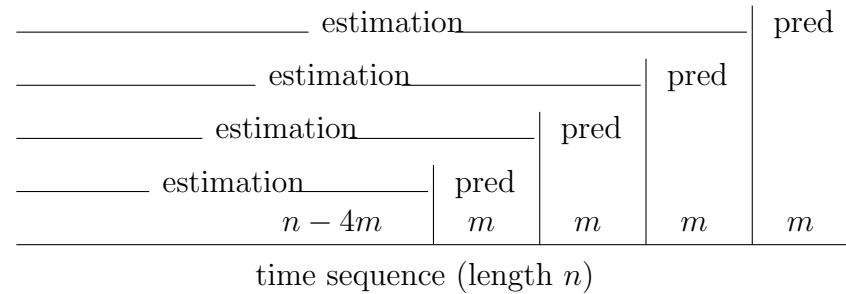


Figure 2.8: Schematic illustration of assessing prediction errors

The PE can be more **reliably assessed** by the following extension of cross-validation idea, which is schemely illustrated above.

error. Here, both **static** (parameters are not updated) or **rolling technique** (parameters are re-estimated as prediction along) can be used.

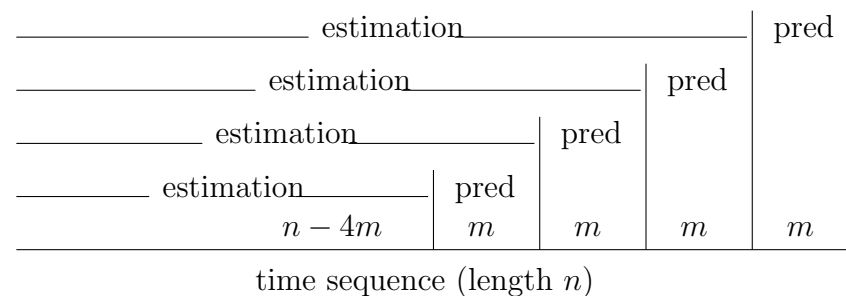


Figure 2.8: Schematic illustration of assessing prediction errors

The PE can be more **reliably assessed** by the following extension of cross-validation idea, which is schemely illustrated above.

- For each prediction period, we can compute PE based on future data in the prediction period.
- The division makes it easier to compute (saving computing resource).
- Aggregate  $m$  PEs to get the overall PE.

The idea is closely related to the **cross-validation** for cross-section data, which is summarized

as follows.

- choose  $m$  at random from the  $n$  data points (testing data);
- use  $(n - m)$  data points to estimate parameter (learning data);
- validate the model by computing PE for  $m$  out sample data ;
- repeat the above process a number of times.

as follows.

- choose  $m$  at random from the  $n$  data points (testing data);
- use  $(n - m)$  data points to estimate parameter (learning data);
- validate the model by computing PE for  $m$  out sample data ;
- repeat the above process a number of times.

## 2.7 Sampling Frequency\*

Time units of financial data: minute, daily, monthly, ...

For an illustration, let  $\{X_t\}$  be an hourly series and  $Y_t = X_{mt}$  be a daily series. Assume that

$$X_t = \rho X_{t-1} + \varepsilon_t.$$

Then,

$$\begin{aligned} Y_t &= \rho X_{mt-1} + \varepsilon_{mt} \\ &= \rho(\rho X_{mt-2} + \varepsilon_{mt-1}) + \varepsilon_{mt} \\ &= \rho^m Y_{t-1} + \tilde{\varepsilon}_t, \end{aligned}$$

as follows.

- choose  $m$  at random from the  $n$  data points (testing data);
- use  $(n - m)$  data points to estimate parameter (learning data);
- validate the model by computing PE for  $m$  out sample data ;
- repeat the above process a number of times.

## 2.7 Sampling Frequency\*

Time units of financial data: minute, daily, monthly, ...

For an illustration, let  $\{X_t\}$  be an hourly series and  $Y_t = X_{mt}$  be a daily series. Assume that

$$X_t = \rho X_{t-1} + \varepsilon_t.$$

Then,

$$\begin{aligned} Y_t &= \rho X_{mt-1} + \varepsilon_{mt} \\ &= \rho(\rho X_{mt-2} + \varepsilon_{mt-1}) + \varepsilon_{mt} \\ &= \rho^m Y_{t-1} + \tilde{\varepsilon}_t, \end{aligned}$$

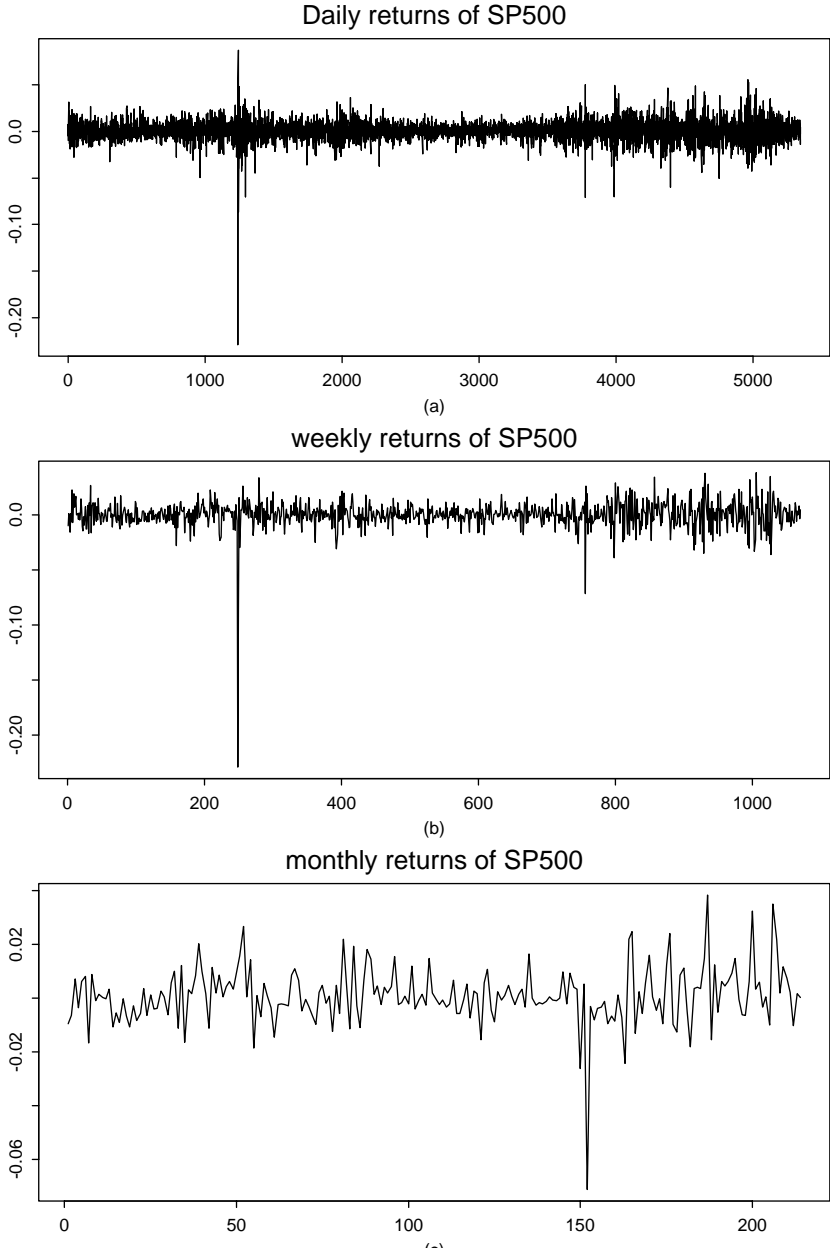


Figure 2.9: Log-return for the S & P500 index with different sampling frequencies in the period 11/22/1982 — 1/30/2004.

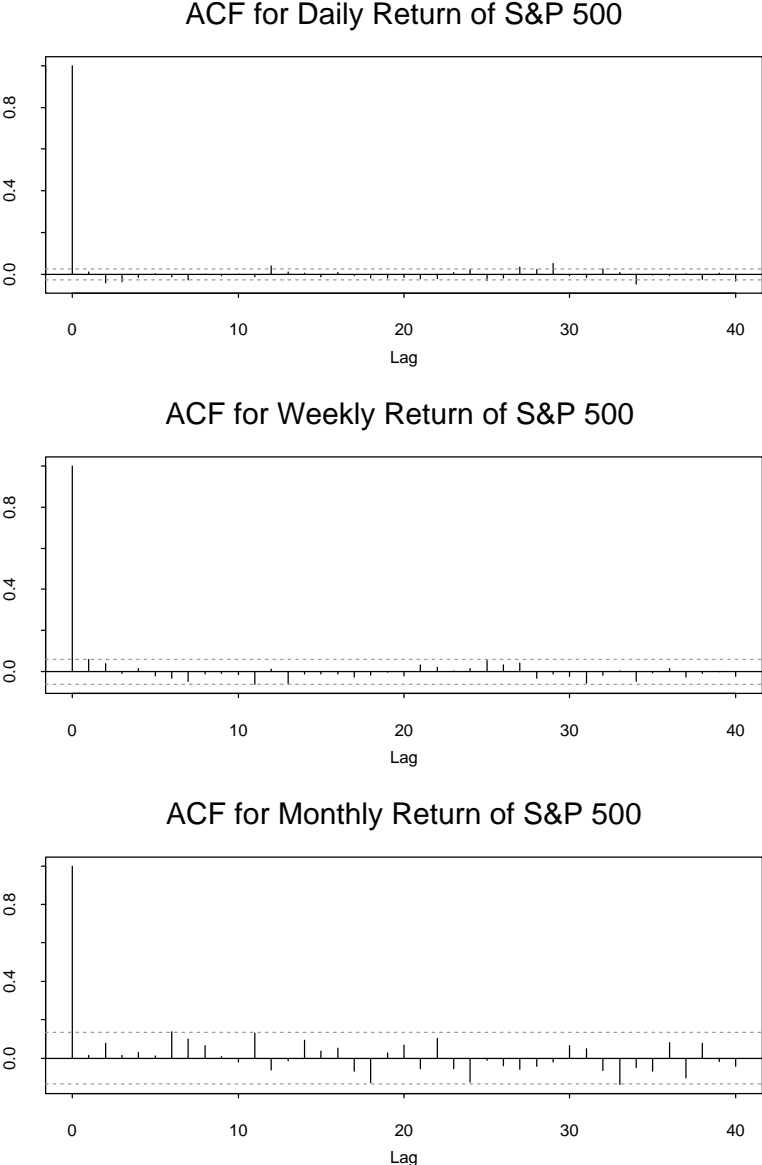


Figure 2.10: ACFs for SP500 returns with different frequency.

where  $\tilde{\varepsilon}_t = \varepsilon_{mt} + \rho\varepsilon_{m(t-1)} + \cdots + \rho^{m-1}\varepsilon_{m(t-1)+1}$ . The series  $\{\tilde{\varepsilon}_t\}$  is uncorrelated white noise with

$$\text{Var}(\tilde{\varepsilon}_t) = \sigma^2(1 + \rho^2 + \cdots + \rho^{2(m-1)}).$$

Hence,  $Y_t$  is also an AR(1) process.

When the time unit is small, the series converge to a continuous time process. Consider

**Ornstein-Uhlenbeck process**:  $dX_t = -\lambda X_t dt + \sigma dW_t$ ,

where  $\{W_t\}$  is the **Brownian motion** (Wiener process), satisfying

- $W(0) = 0$ ;
- (independent increment)  $W_s$  and  $W_t - W_s$  are independent;
- (Normality):  $W_t - W_s \sim \mathcal{N}(0, t - s)$ .

where  $\tilde{\varepsilon}_t = \varepsilon_{mt} + \rho\varepsilon_{m(t-1)} + \dots + \rho^{m-1}\varepsilon_{m(t-1)+1}$ . The series  $\{\tilde{\varepsilon}_t\}$  is uncorrelated white noise with

$$\text{Var}(\tilde{\varepsilon}_t) = \sigma^2(1 + \rho^2 + \dots + \rho^{2(m-1)}).$$

Hence,  $Y_t$  is also an AR(1) process.

When the time unit is small, the series converge to a continuous time process. Consider

**Ornstein-Uhlenbeck process:**  $dX_t = -\lambda X_t dt + \sigma dW_t$ ,

where  $\{W_t\}$  is the **Brownian motion** (Wiener process), satisfying

- $W(0) = 0$ ;
- (independent increment)  $W_s$  and  $W_t - W_s$  are independent;
- (Normality):  $W_t - W_s \sim \mathcal{N}(0, t - s)$ .

Then

$$X_{t+\Delta} - X_t \approx -\lambda X_t \Delta + \sigma \Delta^{\frac{1}{2}} \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1),$$

or

$$\begin{aligned} X_{t+\Delta} &\cong (1 - \lambda\Delta)X_t + \sigma\Delta^{\frac{1}{2}}\varepsilon_t \\ &\cong e^{-\lambda\Delta}X_t + \sigma\Delta^{\frac{1}{2}}\varepsilon_t. \end{aligned}$$

The time series  $Y = X_{t\Delta}$  follows approximately the AR(1) model

$$Y_{t+1} = e^{-\lambda\Delta}Y_t + \sigma\Delta^{\frac{1}{2}}\varepsilon_t,$$

and low-frequency data

$$X_{tm\Delta} = e^{-\lambda\Delta m}X_{(t-1)m\Delta} + \sigma(m\Delta)^{\frac{1}{2}}\varepsilon_t.$$

In other words, the AR(1) model can be regarded as an approximation to the  $OU$  process. Indeed, the approximation can be made exactly (see the Vasicek model in §9.2.2).

## 2.8 Unit root and Martingale Hypothesis

Fundamental question in Financial Econometric: Whether asset prices are predictable.

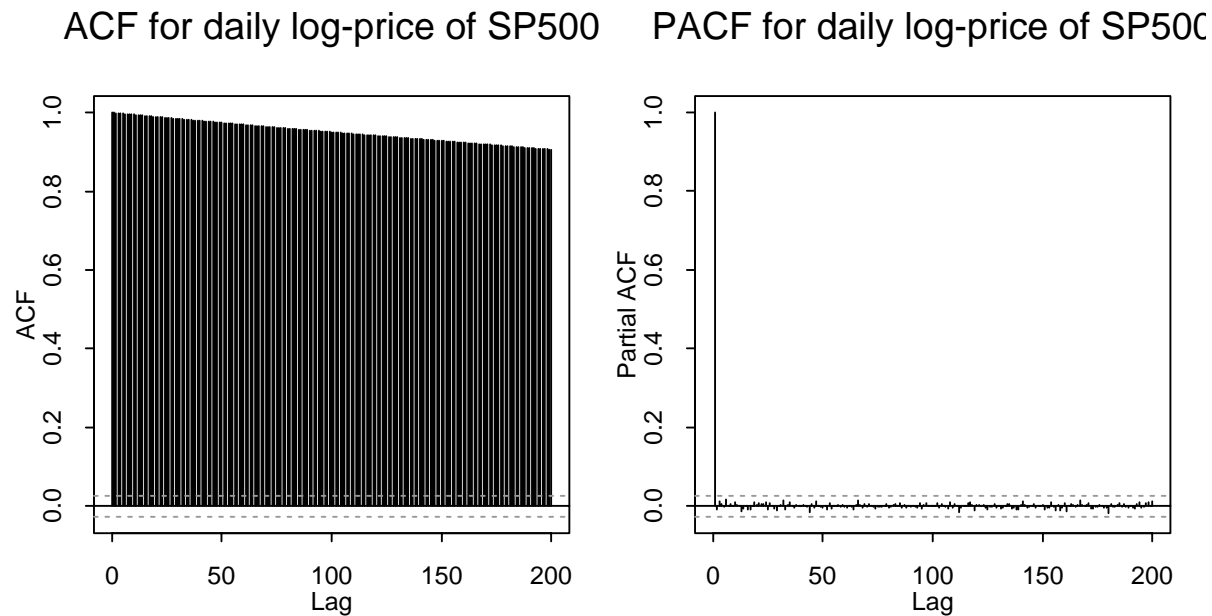


Figure 2.11: ACF and PACF for daily log-price of SP500 index.

Consider the log-price of SP500:  $X_t = \log(S_t)$ , where  $S_t =$  index

at  $t$ . The autocorrelation of  $\{X_t\}$  features strong persistence. From PACF plot, it is reasonable to assume that

$$X_t = X_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \text{ white noise.}$$

It is an AR(1) model whose characteristic equation  $1 - x$  has unit root.

at  $t$ . The autocorrelation of  $\{X_t\}$  features strong persistence. From PACF plot, it is reasonable to assume that

$$X_t = X_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \text{ white noise.}$$

It is an AR(1) model whose characteristic equation  $1 - x$  has unit root.

Note that  $X_t = X_0 + \sum_{i=1}^t \varepsilon_i$ , which is

- nonstationary;
- a random walk without drift.

The random walk with a drift is defined as

$$\begin{aligned} X_t &= \mu + X_{t-1} + \varepsilon_t \\ &= \mu t + X_0 + \sum_{i=1}^t \varepsilon_i. \end{aligned}$$

The covariance structure is given

$$\text{Cov}(X_s, X_t) = \text{Var}(X_0) + s\text{Var}(\varepsilon_i), \quad \text{for } s < t,$$

which is **similar to that of Brownian motion.**

The random walk with a drift is defined as

$$\begin{aligned} X_t &= \mu + X_{t-1} + \varepsilon_t \\ &= \mu t + X_0 + \sum_{i=1}^t \varepsilon_i. \end{aligned}$$

The covariance structure is given

$$\text{Cov}(X_s, X_t) = \text{Var}(X_0) + s\text{Var}(\varepsilon_i), \quad \text{for } s < t,$$

which is **similar to that of Brownian motion**.

These non-stationary processes can be transformed into stationary via differencing:  $Y_t = X_t - X_{t-1}$ , which is stationary.

**Unit root test**: It is natural to embed the random walk into the

AR(1) model:

$$X_t = \rho X_{t-1} + \varepsilon_t \quad \text{without drift;}$$

$$X_t = \mu + \rho X_{t-1} + \varepsilon_t \quad \text{with drift.}$$

The random walk hypothesis:  $H_0 : \rho = 1$ . This corresponds to the parameter at the boundary of the parameter space. Conventional theory does not apply.

AR(1) model:

$$X_t = \rho X_{t-1} + \varepsilon_t \quad \text{without drift;}$$

$$X_t = \mu + \rho X_{t-1} + \varepsilon_t \quad \text{with drift.}$$

The random walk hypothesis:  $H_0 : \rho = 1$ . This corresponds to the parameter at the boundary of the parameter space. Conventional theory does not apply.

Estimation of  $\rho$  by the least squares-method:

- **Model without drift**

$$\hat{\rho} = \frac{\sum_{t=2}^T X_t X_{t-1}}{\sum_{t=2}^T X_{t-1}^2}.$$

- **Model with drift**

$$\begin{aligned}\tilde{\mu} &= \bar{X}_2 - \tilde{\rho}\bar{X}_1 \\ \tilde{\rho} &= \frac{\sum_{t=2}^T (X_t - \bar{X}_2)(X_{t-1} - \bar{X}_1)}{\sum_{t=2}^T (X_{t-1} - \bar{X}_1)^2},\end{aligned}$$

where  $\bar{X}_2 = (T-1)^{-1} \sum_{t=2}^T X_t$  and  $\bar{X}_1 = (T-1)^{-1} \sum_{t=2}^T X_{t-1}$ .

- **Model with drift**

$$\begin{aligned}\tilde{\mu} &= \bar{X}_2 - \tilde{\rho}\bar{X}_1 \\ \tilde{\rho} &= \frac{\sum_{t=2}^T (X_t - \bar{X}_2)(X_{t-1} - \bar{X}_1)}{\sum_{t=2}^T (X_{t-1} - \bar{X}_1)^2},\end{aligned}$$

where  $\bar{X}_2 = (T-1)^{-1} \sum_{t=2}^T X_t$  and  $\bar{X}_1 = (T-1)^{-1} \sum_{t=2}^T X_{t-1}$ .

- Under the null hypothesis  $H_0 : \rho = 1$ , both estimators  $\hat{\rho}$  and  $\tilde{\rho}$  converge to 1 at **rate**  $O(T^{-1})$ .

Test rule (Dickey-Fuller tests, JASA, 79, 427-431): Reject  $H_0$  at

$\alpha = 5\%$ , when

$$T(1 - \hat{\rho}) > 14.1 \text{ without drift;}$$

$$T(1 - \tilde{\rho}) > 21.8 \text{ with drift.}$$

**Example 4**. For the S&P 500 daily log-prices,  $T = 5348$ . For testing against random-walk without a drift, it can be computed that

$$\hat{\rho} = 1.0006, \quad T(1 - \hat{\rho}) = -3.2088.$$

We have weak evidence against the random walk hypothesis, as the test statistic is smaller than the critical value.

**Example 4**. For the S&P 500 daily log-prices,  $T = 5348$ . For testing against random-walk without a drift, it can be computed that

$$\hat{\rho} = 1.0006, \quad T(1 - \hat{\rho}) = -3.2088.$$

We have weak evidence against the random walk hypothesis, as the test statistic is smaller than the critical value.

Similarly, for random walk hypothesis with a drift,

$$\tilde{\rho} = 0.9997106, \quad T = 5348, \quad T(1 - \tilde{\rho}) = 1.5479,$$

which is consistent with the random walk hypothesis.

For monthly log-prices,  $T = 214$ . It can be computed that

$$\hat{\rho} = 1.0050, \quad T(1 - \hat{\rho}) = -0.321,$$

and

$$\tilde{\rho} = 0.99339, \quad T(1 - \tilde{\rho}) = 1.4135.$$

Both do not provide strong evidence against  $H_0$ .

and

$$\tilde{\rho} = 0.99339, \quad T(1 - \tilde{\rho}) = 1.4135.$$

Both do not provide strong evidence against  $H_0$ .

### Remarks:

- This does not imply that the null hypothesis must be true.
- The Dickey-Fuller is a parametric test on the price process  $\{X_t\}$ .  
The basic assumption is that  $\{X_t\}$  follows an AR(1) model, which might be wrong to begin with.
- The Ljung-Box is a NP test for uncorrelatedness based on the return process. They have different assumptions to begin with.

Example 5: Suppose that unknown to us, the data follow the fol-

lowing nonlinear auto-regression model

$$X_t = 0.8 \frac{X_{t-1}^2}{1 + X_{t-1}^2} + \varepsilon_t,$$

but we make the assumption

$$X_t = b_0 + b_1 X_{t-1} + \varepsilon_t$$

and test  $H_0 : b_1 = 0$ . Even if  $H_0$  is accepted at  $\alpha = 5\%$ , we cannot conclude with certainty that  $\{X_t\}$  is a white noise series.

lowing nonlinear auto-regression model

$$X_t = 0.8 \frac{X_{t-1}^2}{1 + X_{t-1}^2} + \varepsilon_t,$$

but we make the assumption

$$X_t = b_0 + b_1 X_{t-1} + \varepsilon_t$$

and test  $H_0 : b_1 = 0$ . Even if  $H_0$  is accepted at  $\alpha = 5\%$ , we cannot conclude with certainty that  $\{X_t\}$  is a white noise series.

**Martingale**: Let  $E_t$  be the conditional expectation given the information up to time  $t$ . A sequence  $\{Y_t\}$  is called **martingale** if it is observable at time  $t$  and

$$E_{t-1} Y_t = Y_{t-1}, \quad \forall t.$$

Hence,  $\varepsilon_t = Y_t - Y_{t-1}$  is called a martingale difference, satisfying

$$E_t \varepsilon_{t+1} = 0.$$

Hence,  $\varepsilon_t = Y_t - Y_{t-1}$  is called a martingale difference, satisfying  $E_t \varepsilon_{t+1} = 0$ .

**Motivation**: The returns of assets are uncorrelated, but not independent. This is evidenced by the following correlograms.

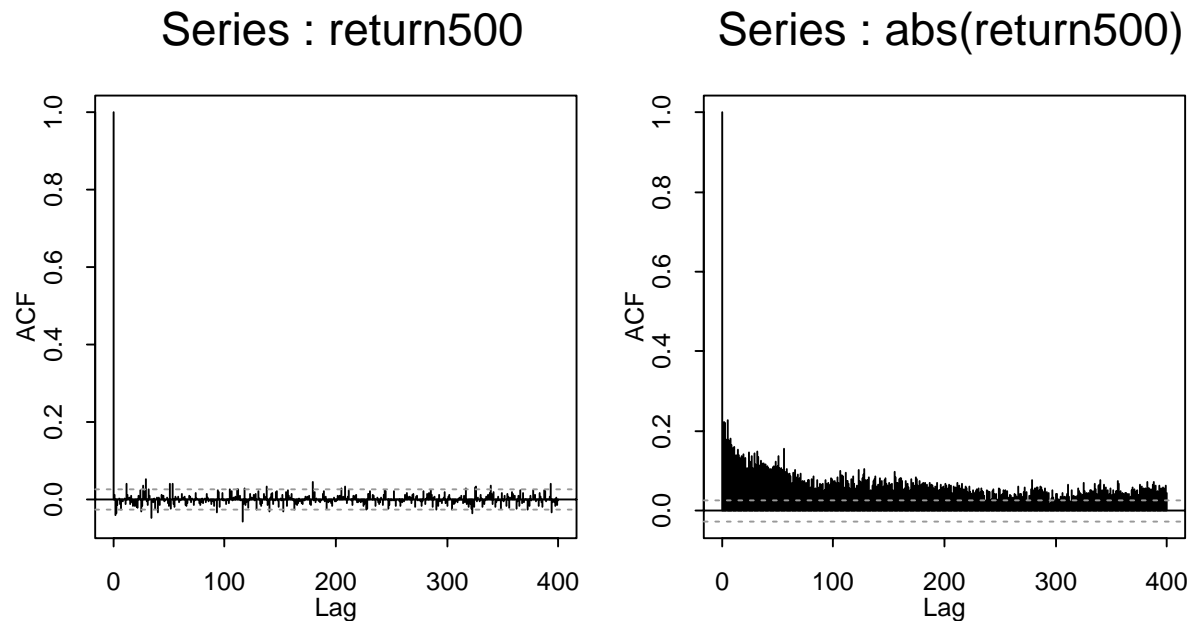


Figure 2.12: ACFs of the log-returns and absolute log-returns of the S&P 500 index. (should be replaced by another data, using squared return series)

## Properties of martingale difference:

(i)  $E\varepsilon_t = E\{E_{t-1}\varepsilon_t\} = 0$  (double expectation formula);

(ii)  $E_t\varepsilon_{t+m} = E_t\{E_{t+m-1}\varepsilon_{t+m}\} = 0, \quad \forall m \geq 1$  (unpredictable conditional mean);

(iii)  $E_t\varepsilon_t\varepsilon_{t+m} = E\{E_t(\varepsilon_t\varepsilon_{t+m})\} = 0, \quad \forall m > 0$  (uncorrelated).

Relationship:  $\{\text{i.i.d.}\} \implies \{\text{martingale difference}\} \implies \{\text{uncorrelated}\}.$

## Properties of martingale difference:

- (i)  $E\varepsilon_t = E\{E_{t-1}\varepsilon_t\} = 0$  (double expectation formula);
- (ii)  $E_t\varepsilon_{t+m} = E_t\{E_{t+m-1}\varepsilon_{t+m}\} = 0, \quad \forall m \geq 1$  (unpredictable conditional mean);
- (iii)  $E_t\varepsilon_t\varepsilon_{t+m} = E\{E_t(\varepsilon_t\varepsilon_{t+m})\} = 0, \quad \forall m > 0$  (uncorrelated).

Relationship:  $\{\text{i.i.d.}\} \implies \{\text{martingale difference}\} \implies \{\text{uncorrelated}\}.$

Martingale hypothesis: The asset price is a random walk  $X_t = X_{t-1} + \varepsilon_t$  with martingale difference  $\varepsilon_t$ . Hence

$$E_{t-1}X_t = X_{t-1} + E_{t-1}\varepsilon_t = X_{t-1}$$

is a martingale.

**Remarks:** The in-predicability of asset returns  $r_t$  is the same as the random walk of (logarithm) asset prices. The meaning of in-predicability takes three mathematical forms (weakest to strongest):

- the series  $\{r_t\}$  is uncorrelated (the future expected return can not be predicted by **linear** rules);

**Remarks:** The in-predicability of asset returns  $r_t$  is the same as the random walk of (logarithm) asset prices. The meaning of in-predicability takes three mathematical forms (weakest to strongest):

- the series  $\{r_t\}$  is uncorrelated (the future expected return can not be predicted by **linear** rules);
- the series  $\{r_t\}$  is a martingale difference (the future expected return can not be predicted even by **nonlinear** rules);
- the series  $\{r_t\}$  is an i.i.d. sequence (**nothing** can be predicted, including volatility).

**Remarks:** The in-predicability of asset returns  $r_t$  is the same as the random walk of (logarithm) asset prices. The meaning of in-predicability takes three mathematical forms (weakest to strongest):

- the series  $\{r_t\}$  is uncorrelated (the future expected return can not be predicted by **linear** rules);
- the series  $\{r_t\}$  is a martingale difference (the future expected return can not be predicted even by **nonlinear** rules);
- the series  $\{r_t\}$  is an i.i.d. sequence (**nothing** can be predicted, including volatility).

The martingale hypothesis gives a more **reasonable hypothesis** on the price dynamics (returns are not predictable, but volatility can

possibly be). However, the Ljung and Box and the Dickey-Fuller tests are only testing uncorrelated white noise.

Implications of martingale hypothesis: is a form of **efficient market hypothesis**. Suppose that we have initial endowment  $C_0$ , invested in a riskless asset with price 1 and risky asset with price  $S_t$ .

possibly be). However, the Ljung and Box and the Dickey-Fuller tests are only testing uncorrelated white noise.

Implications of martingale hypothesis: is a form of **efficient market hypothesis**. Suppose that we have initial endowment  $C_0$ , invested in a riskless asset with price 1 and risky asset with price  $S_t$ .

At date  $t$ :  $\alpha_t$  dollars of cash and  $\beta_t$  shares of stock, with value  $C_t = \alpha_t + \beta_t S_t$ .

At time  $t + 1$ : value is  $\alpha_t + \beta_t S_{t+1}$ . Suppose a new strategy invests  $\alpha_{t+1}$  in cash and  $\beta_{t+1}$  in stock.

Self-financing:  $C_{t+1} = \alpha_{t+1} + \beta_{t+1} S_{t+1} = \alpha_t + \beta_t S_{t+1}$  ( $\alpha_{t+1}$  and

$\beta_{t+1}$  allow to be negative, “short positions”). Hence,

$$C_{t+1} - C_t = \beta_t(S_{t+1} - S_t).$$

If an asset price is a martingale, then

$$E_t(C_{t+1} - C_t) = 0.$$

$\beta_{t+1}$  allow to be negative, “short positions”). Hence,

$$C_{t+1} - C_t = \beta_t(S_{t+1} - S_t).$$

If an asset price is a martingale, then

$$E_t(C_{t+1} - C_t) = 0.$$

— No matter how complex the strategy of portfolio allocations is, the expected return at future is zero.

— Impossible to have probability 1 such that

$$C_t > C_0,$$

namely **no arbitrage** opportunity exists (Recall “double-strategy” in gambling).

- The market is efficient that even a skilled investor has no sure advantage.
- Efficient market hypothesis or martingale hypothesis can be tested using the Ljung-Box and Dickey-Fuller tests (in fact tested only uncorrelatedness). We do not have any strong empirical evidence against the efficient market hypothesis based on the Ljung-Box and Dickey-Fuller tests.

## 2.9 ARMA Processes

Are the AR models rich enough? Why ARMA ?

**Wold Theorem:** Any stationary process can be written as

$$X_t = \mu + \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_k\varepsilon_{t-k} + \cdots ,$$

where  $\{\varepsilon_t\}$  is a white noise,  $\sum a_i^2 < \infty$ . This is a moving average of infinity order.

Consider a simple MA(1) model

$$X_t = \varepsilon_t - b\varepsilon_{t-1} = (1 - bB) \cdot \varepsilon_t$$

$$\implies (1 - bB)^{-1}X_t = \varepsilon_t$$

$$\implies (1 + bB + b^2B^2 + \dots)X_t = \varepsilon_t$$

$$\implies X_t = -bX_{t-1} - b^2X_{t-2} - \dots + \varepsilon_t.$$

Consider a simple MA(1) model

$$X_t = \varepsilon_t - b\varepsilon_{t-1} = (1 - bB) \cdot \varepsilon_t$$

$$\implies (1 - bB)^{-1}X_t = \varepsilon_t$$

$$\implies (1 + bB + b^2B^2 + \dots)X_t = \varepsilon_t$$

$$\implies X_t = -bX_{t-1} - b^2X_{t-2} - \dots + \varepsilon_t.$$

♠ Even a simple MA model, it takes a large AR model to approximate it. (not a parsimonious model)

♠ Likewise, a simple AR(1) model will take a large order of an MA process to approximate it.

Assume that the series  $\{X_t\}$  has zero mean (mean has been removed).

## ARMA model:

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q},$$

where  $\{\varepsilon_t\} \sim \mathcal{WN}(0, \sigma^2)$ , denoted by  $\{X_t\} \sim \text{ARMA}(p, q)$ . It can be written as

$$b(B)X_t = a(B)\varepsilon_t,$$

where  $b(z) = 1 - b_1 z - \cdots - b_p z^p$ ,  $a(z) = 1 + a_1 z + \cdots + a_q z^q$ .

## ARMA model:

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q},$$

where  $\{\varepsilon_t\} \sim \mathcal{WN}(0, \sigma^2)$ , denoted by  $\{X_t\} \sim \text{ARMA}(p, q)$ . It can be written as

$$b(B)X_t = a(B)\varepsilon_t,$$

where  $b(z) = 1 - b_1 z - \cdots - b_p z^p$ ,  $a(z) = 1 + a_1 z + \cdots + a_q z^q$ .

**Identification:** The order as defined above is not identifiable:

$$(1 - 0.7B)b(B)X_t = (1 - 0.7B)a(B)\varepsilon_t$$

is  $\text{ARMA}(p + 1, q + 1)$ . Thus, we assume  $a(z)$  and  $b(z)$  do not share the same root.

Stationarity: (FY: Theorem 2.1) When

$$\inf_{|z|<1} |b(z)| > 0,$$

the ARMA model can be written as

$$X_t = b(B)^{-1}a(B)\varepsilon_t = d_0\varepsilon_t + d_1\varepsilon_{t-1} + \cdots ,$$

an infinite order of MA and hence is stationary.

**Stationarity:** (FY: Theorem 2.1) When

$$\inf_{|z|<1} |b(z)| > 0,$$

the ARMA model can be written as

$$X_t = b(B)^{-1}a(B)\varepsilon_t = d_0\varepsilon_t + d_1\varepsilon_{t-1} + \cdots,$$

an infinite order of MA and hence is stationary.

**Yule-Walker equation:** Note that

$$\text{Cov}\{b(B)X_t, X_{t-k}\} = \text{Cov}\{a(B)\varepsilon_t, X_{t-k}\} = 0, \quad \text{if } k > q.$$

It follows

$$\gamma(k) - b_1\gamma(k-1) - \cdots - b_p\gamma(k-p) = 0, \text{ for } k > q$$

or

$$b(B)\gamma(k) = (B - z_1) \cdots (B - z_p)\gamma(k) = 0, \text{ for } k > q,$$

where  $z_1, \cdots, z_p$  are the roots of  $b(z)$ .

or

$$b(B)\gamma(k) = (B - z_1) \cdots (B - z_p)\gamma(k) = 0, \text{ for } k > q,$$

where  $z_1, \dots, z_p$  are the roots of  $b(z)$ .

**Solution:** The above difference equation is similar to that of AR( $p$ ) model and the solution is of form:

$$\gamma(k) = \alpha_1 z_1^{-k} + \cdots + \alpha_p z_p^{-k}.$$

Hence,  $\rho(k) \rightarrow 0$  exponentially fast.

or

$$b(B)\gamma(k) = (B - z_1) \cdots (B - z_p)\gamma(k) = 0, \text{ for } k > q,$$

where  $z_1, \dots, z_p$  are the roots of  $b(z)$ .

**Solution:** The above difference equation is similar to that of AR( $p$ ) model and the solution is of form:

$$\gamma(k) = \alpha_1 z_1^{-k} + \cdots + \alpha_p z_p^{-k}.$$

Hence,  $\rho(k) \rightarrow 0$  exponentially fast.

**Example 6.** Consider the ARMA(1,1) model. For  $k > 1$ , we have

$$\gamma(k) = b_1 \gamma(k-1) = \cdots = b_1^{k-1} \gamma(1),$$

which decays exponentially. Using  $X_{t-1} = b_1 X_{t-2} + \varepsilon_{t-1} + a_1 \varepsilon_{t-2}$ ,

we have

$$\gamma(1) = \text{Cov}(X_t, X_{t-1}) = \text{Cov}(b_1 X_{t-1} + a_1 \varepsilon_{t-1}, X_{t-1}) = b_1 \gamma(0) + a_1 \sigma^2.$$

Note that

$$\gamma(0) = \text{var}(X_t) = b_1^2 \gamma(0) + (1 + a_1^2) \sigma^2 + 2a_1 b_1 \sigma^2.$$

Hence,

$$\gamma(0) = \frac{1 + a_1^2 + 2a_1 b_1}{1 - b_1^2} \sigma^2.$$

and

$$\gamma(1) = \frac{a_1 + b_1(1 + a_1^2 + a_1 b_1)}{1 - b_1^2} \sigma^2, \quad \gamma(k) = \gamma(1) b_1^k, \text{ for } k \geq 1$$

**Estimation:** For given orders  $p$  and  $q$ , the parameters can be estimated by the **Quasi Maximum Likelihood** (QML) method under

the assumption that  $\{\varepsilon_t\} \sim i.i.d \mathcal{N}(0, \sigma^2)$ . See Chapter 3 of FY for details.

Let  $\mathbf{X}_T$  be the observed series and  $\mathbf{\Sigma}$  be its covariance matrix, which depends on  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\sigma$ . Under the Gaussian assumption,

$$\mathbf{X}_T \equiv (X_1, \dots, X_T)' \sim \mathcal{N}(0, \mathbf{\Sigma}).$$

Hence, the likelihood function (the density function of  $\mathbf{X}_T$ ) can be written as

$$L(\mathbf{a}, \mathbf{b}, \sigma^2) \propto |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \mathbf{X}_T' \mathbf{\Sigma}^{-1} \mathbf{X}_T \right\}.$$

the assumption that  $\{\varepsilon_t\} \sim i.i.d \mathcal{N}(0, \sigma^2)$ . See Chapter 3 of FY for details.

Let  $\mathbf{X}_T$  be the observed series and  $\mathbf{\Sigma}$  be its covariance matrix, which depends on  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\sigma$ . Under the Gaussian assumption,

$$\mathbf{X}_T \equiv (X_1, \dots, X_T)' \sim \mathcal{N}(0, \mathbf{\Sigma}).$$

Hence, the likelihood function (the density function of  $\mathbf{X}_T$ ) can be written as

$$L(\mathbf{a}, \mathbf{b}, \sigma^2) \propto |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \mathbf{X}_T' \mathbf{\Sigma}^{-1} \mathbf{X}_T \right\}.$$

—  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\sigma^2$  are estimated by maximizing  $L(\mathbf{a}, \mathbf{b}, \sigma^2)$ ; the likelihood principle furnishes estimated parameters and its associated standard error. The **estimated covariance matrix** is the neg-

ative Hessian matrix of the log-likelihood.

- direct computation involves inverting a  $T \times T$  matrix and new algorithm is needed.
- matrix inversion is avoided by a **prewhitening** technique:
  - find the best linear predictor  $\hat{X}_t$  for  $X_t$  recursively;
  - the residual series  $\{X_t - \hat{X}_t\}$  are uncorrelated ;
  - express the series  $X_t$  as linear combination of  $\{X_t - \hat{X}_t\}$ .

ative Hessian matrix of the log-likelihood.

- direct computation involves inverting a  $T \times T$  matrix and new algorithm is needed.
- matrix inversion is avoided by a **prewhitening** technique:
  - find the best linear predictor  $\hat{X}_t$  for  $X_t$  recursively;
  - the residual series  $\{X_t - \hat{X}_t\}$  are uncorrelated ;
  - express the series  $X_t$  as linear combination of  $\{X_t - \hat{X}_t\}$ .

**Order Selection:** The idea for selecting AR models continues to

apply. Let  $\hat{\ell}(p, q) = \max_{\mathbf{a}, \mathbf{b}, \sigma} \log L(\mathbf{a}, \mathbf{b}, \sigma)$ . Then

$$\text{AIC}(p, q) = -2\hat{\ell}(p, q) + 2(p + q + 1);$$

$$\text{AICC}(p, q) = -2\hat{\ell}(p, q) + \frac{2(p + q + 1)T}{T - p - q - 2};$$

$$\text{BIC}(p, q) = -2\hat{\ell}(p, q) + (p + q + 1) \log(T - L);$$

and the rolling and PE idea continues to apply.

**Prediction:** Let  $X_T(m) = E_T(X_{T+m})$ . Then, from

$$X_{T+m} = \sum_{j=1}^p b_j X_{T+m-j} + \varepsilon_{T+m} + \sum_{j=1}^q a_j \varepsilon_{T+m-j},$$

we have

$$X_T(m) = \sum_{j=1}^p b_j X_T(m-j) + \sum_{j=1}^q a_j \varepsilon_T(m-j),$$

apply. Let  $\hat{\ell}(p, q) = \max_{\mathbf{a}, \mathbf{b}, \sigma} \log L(\mathbf{a}, \mathbf{b}, \sigma)$ . Then

$$\text{AIC}(p, q) = -2\hat{\ell}(p, q) + 2(p + q + 1);$$

$$\text{AICC}(p, q) = -2\hat{\ell}(p, q) + \frac{2(p + q + 1)T}{T - p - q - 2};$$

$$\text{BIC}(p, q) = -2\hat{\ell}(p, q) + (p + q + 1) \log(T - L);$$

and the rolling and PE idea continues to apply.

**Prediction:** Let  $X_T(m) = E_T(X_{T+m})$ . Then, from

$$X_{T+m} = \sum_{j=1}^p b_j X_{T+m-j} + \varepsilon_{T+m} + \sum_{j=1}^q a_j \varepsilon_{T+m-j},$$

we have

$$X_T(m) = \sum_{j=1}^p b_j X_T(m-j) + \sum_{j=1}^q a_j \varepsilon_T(m-j),$$

where

$$\varepsilon_T(i) = \mathbb{E}_T\{\varepsilon_{T+i}\} = \begin{cases} 0 & \text{if } i > 0 \\ \varepsilon_{T+i} & \text{if } i \leq 0 \end{cases}.$$

Thus, starting from the one-step ahead prediction, we compute the two-step ahead prediction and so on. An alternative approach is to represent ARMA( $p, q$ ) in an AR( $\infty$ ) process. When  $\sup_{|z| \leq 1} |a(z)| > 0$ ,

$$a(B)^{-1}b(B)X_t = \varepsilon_t,$$

$$X_t - \sum_{j=1}^{\infty} c_j X_{t-j} = \varepsilon_t$$

with  $c_k = b_k - \sum_{j=1}^{k-1} c_j a_{k-j}$  and convention that  $b_{p+j} = a_{q+j} =$

$$0, \quad \forall j \geq 1.$$

0,  $\forall j \geq 1$ . Thus

$$X_T(m) = \sum_{j=1}^{\infty} c_j X_T(m-j).$$

**Example 7**. Consider an ARMA(1,1) model

$$X_t - bX_{t-1} = \varepsilon_t - a\varepsilon_{t-1}.$$

For  $m \geq 2$ , it is easy to see

$$X_T(m) - bX_T(m-1) = 0 \implies \hat{X}_{T+m} \equiv X_T(m) = b^{m-1}X_T(1).$$

The long-term forecasting is nearly impossible.

Let us now consider the one-step forecasting.

Let us now consider the one-step forecasting. Note that

$$\begin{aligned}
 \varepsilon_t &= (1 - aB)^{-1}(X_t - bX_{t-1}) \\
 &= \sum_{j=0}^{\infty} a^j B^j (X_t - bX_{t-1}) \\
 &= X_t + a \sum_{j=1}^{\infty} a^{j-1} X_{t-j} - b \sum_{j=1}^{\infty} a^{j-1} X_{t-j} \\
 &= X_t - (b - a) \sum_{j=1}^{\infty} a^{j-1} X_{t-j}.
 \end{aligned}$$

Hence,

$$X_T(1) = (b - a) \{X_T + aX_{T-1} + a^2X_{T-2} + \cdots\}.$$

In particular, when  $b = 1$ , we have

$$\widehat{X}_{T+1} = (1 - a)(X_T + aX_{T-1} + a^2X_{T-2} + \cdots)$$

— exponential smoothing.

$$= (1 - a)X_T + a\widehat{X}_T$$

(possibly non-stationary time series, based on the continuity of the series)

In particular, when  $b = 1$ , we have

$$\widehat{X}_{T+1} = (1 - a)(X_T + aX_{T-1} + a^2X_{T-2} + \cdots)$$

— exponential smoothing.

$$= (1 - a)X_T + a\widehat{X}_T$$

(possibly non-stationary time series, based on the continuity of the series)

**Example 7**: Consider the daily return of SP500 index. Fitting the ARMA(1,1) model results in

$$(r_t - 0.0399) = -0.3379(r_{t-1} - 0.0399) + \varepsilon_t - 0.359\varepsilon_{t-1}.$$

```

return500 <- 100*diff(sp500)
mean(return500)
return500 <- return500 - mean(return500)
return500 <- as.ts(return500)
#acf(return500)
#acf(return500, lag.max=100, type="partial")
return500.arima <- arima.mle(return500, list(order=c(1,0,1)))
arima.diag(return500.arima)
fore500 <- arima.forecast(return500, return500.arima$model, 10)
tsplot(fore500$mean, fore500$mean+2*fore500$std.err,
fore500$mean-2*fore500$std.err)

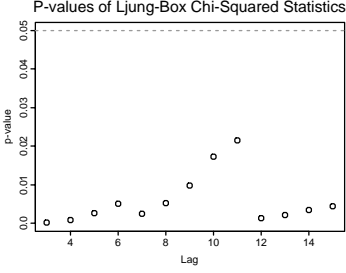
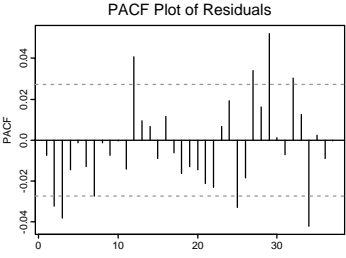
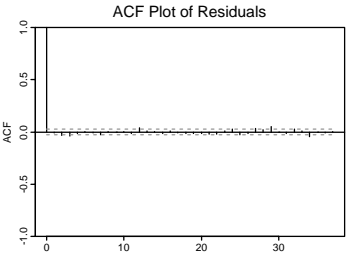
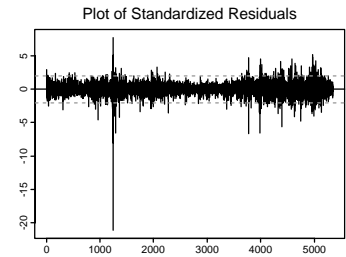
```

	ARMA(1,1)	AR(1)	AR(2)	AR(3)	MA(1)	MA(2)	MA(3)
AIC	15962	15961	15952	15945	15964	15957	15952
$\sigma$	1.159	1.159	1.157	1.156	1.159	1.157	1.156

## 2.10 ARIMA model

When a financial time series appears to have a slow-varying **time**

MA Model Diagnostics: return



ARIMA(1 0 1) Model with Mean 0

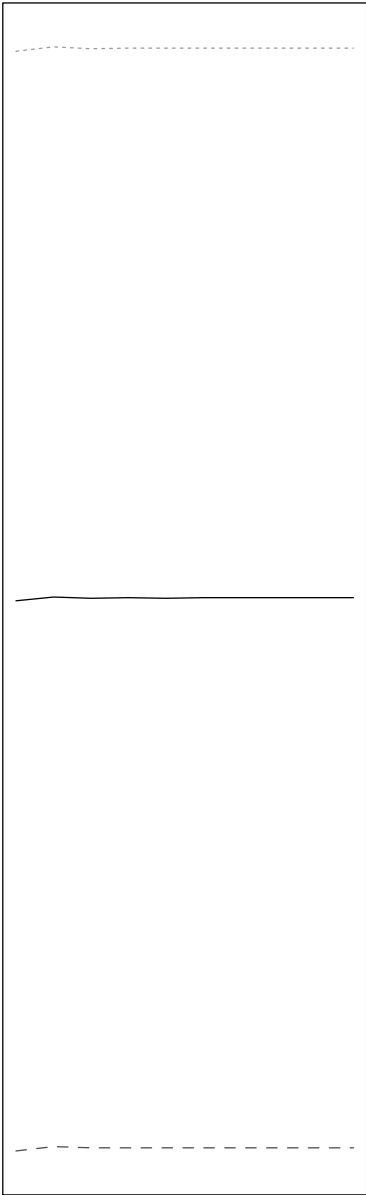


Figure 2.13: ARMA fit to S&P500 returns.

**trend**, a common practice is to remove the trend by differencing.

e.g. Let  $X_t$  be the log-price of SP500 index, and

$$r_t = X_t - X_{t-1} = (1 - B)X_t$$

be the log-return. We hope to model  $r_t$  by an ARMA( $p, q$ ) model.

Since  $X_t$  is the integration of  $r_t$ , the model is called an autoregressive integrated moving average (ARIMA) process. Note that the ARIMA model for ( $d > 0$ ) is non-stationary.

**trend**, a common practice is to remove the trend by differencing.

e.g. Let  $X_t$  be the log-price of SP500 index, and

$$r_t = X_t - X_{t-1} = (1 - B)X_t$$

be the log-return. We hope to model  $r_t$  by an ARMA( $p, q$ ) model.

Since  $X_t$  is the integration of  $r_t$ , the model is called an autoregressive integrated moving average (ARIMA) process. Note that the ARIMA model for ( $d > 0$ ) is non-stationary.

**ARIMA**: Let  $Y_t = (1 - B)^d X_t$ . If

$$b(B)Y_t = a(B)\varepsilon_t \iff b(B)(1 - B)^d X_t = a(B)\varepsilon_t,$$

$\{X_t\}$  is called ARIMA model with order  $p, d$  and  $q$ , denoted by

$\{X_t\} \sim \text{ARIMA}(p, d, q)$ .

The techniques for ARMA model can readily be extended.

## 2.11 Persistence and Long Memory Processes\*

FY: §2.5, RT: §2.10 .

For ARMA models,  $|\rho(k)| \leq Cr^k$ ,  $r < 1$ ,  $k = 0, 1, 2, \dots$ . There also exists a class models with

$$\rho(k) \sim Ck^{2d-1}, \quad \text{as } k \rightarrow \infty, \quad d < 0.5.$$

For  $d \in (0, 0.5)$ ,  $\sum |\rho(k)| = \infty$ . Such a process is called a **long memory process**.

e.g. ACF of log-price of S&P500 appears to decay slowly;

e.g. ACF of absolute log-return of S&P500 appears persistent.

The techniques for ARMA model can readily be extended.

## 2.11 Persistence and Long Memory Processes\*

FY: §2.5, RT: §2.10 .

For ARMA models,  $|\rho(k)| \leq Cr^k$ ,  $r < 1$ ,  $k = 0, 1, 2, \dots$ . There also exists a class models with

$$\rho(k) \sim Ck^{2d-1}, \quad \text{as } k \rightarrow \infty, \quad d < 0.5.$$

For  $d \in (0, 0.5)$ ,  $\sum |\rho(k)| = \infty$ . Such a process is called a **long memory process**.

e.g. ACF of log-price of S&P500 appears to decay slowly;

e.g. ACF of absolute log-return of S&P500 appears persistent.

A family of models with  $\rho(k) \sim Ck^{2d-1}$  is given by the **fractional difference**:

$$(1 - B)^d X_t = \varepsilon_t, \quad -0.5 < d < 0.5.$$

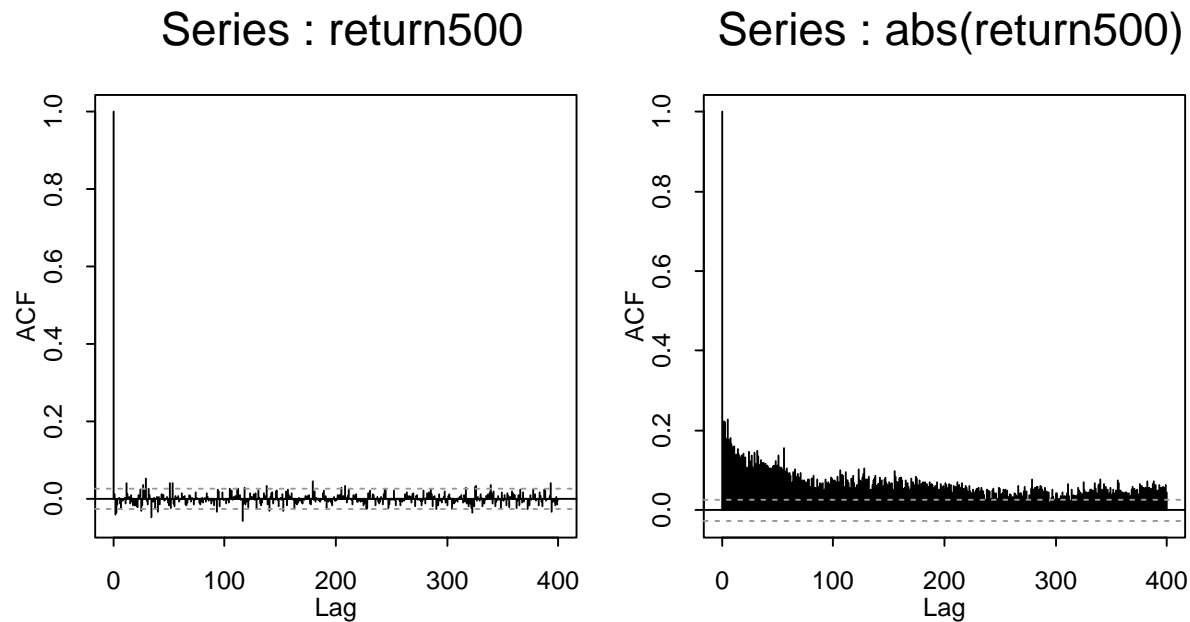


Figure 2.14: ACFs of the log-returns and absolute log-returns of the S&P 500 index.

Here,

$$(1 - B)^d = 1 - dB + \frac{d(1-d)}{2!}B^2 - \dots - (-1)^k \frac{d(1-d)\cdots(k-1-d)}{k!}B^k - \dots$$

Thus,  $X_t$  admits an  $AR(\infty)$  representation

$$X_t - dX_{t-1} + \dots + (-1)^k \frac{d(1-d)\cdots(k-1-d)}{k!}X_{t-k} - \dots = \varepsilon_t.$$

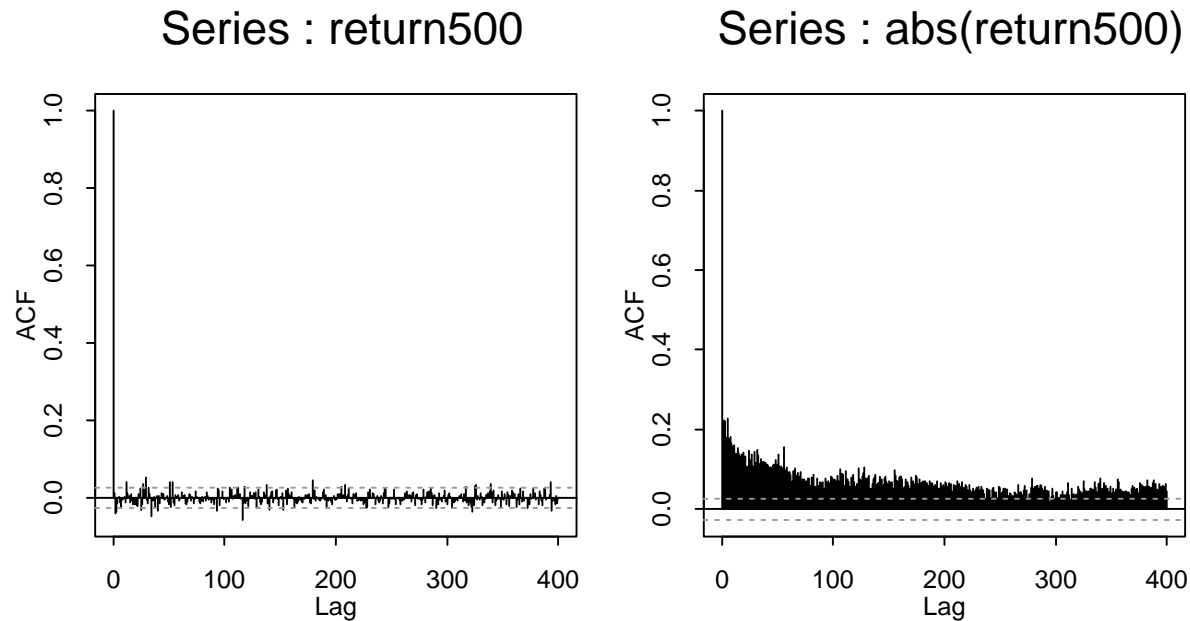


Figure 2.14: ACFs of the log-returns and absolute log-returns of the S&P 500 index.

Here,

$$(1 - B)^d = 1 - dB + \frac{d(1-d)}{2!}B^2 - \dots - (-1)^k \frac{d(1-d) \cdots (k-1-d)}{k!}B^k - \dots$$

Thus,  $X_t$  admits an  $AR(\infty)$  representation

$$X_t - dX_{t-1} + \dots + (-1)^k \frac{d(1-d) \cdots (k-1-d)}{k!}X_{t-k} - \dots = \varepsilon_t.$$

**Properties.** For the process defined above,

$$(i) \rho(k) = \frac{d(1+d) \cdots (k-1+d)}{(1-d)(2-d) \cdots (k-d)} \sim Ck^{2d-1}, \quad k \rightarrow \infty.$$

$$(ii) \text{ PACF: } \pi(k) = d/(k-d).$$

(iii) The Fourier transform of the ACF admits

$$f(\omega) \sim \omega^{-2d}, \quad \text{as } \omega \rightarrow 0.$$

— the spectral density function

—  $d$  can be estimated from log-periodogram with  $\omega$  small.

**FARIMA:** In general, if  $\{X_t\}$  satisfies

$$b(B)(1-B)^d X_t = a(B)\varepsilon_t,$$

then it is called an FARIMA( $p, d, q$ )  $\iff$  Fractional ARIMA model.

$$(i) \rho(k) = \frac{d(1+d)\cdots(k-1+d)}{(1-d)(2-d)\cdots(k-d)} \sim Ck^{2d-1}, \quad k \rightarrow \infty.$$

$$(ii) \text{ PACF: } \pi(k) = d/(k-d).$$

(iii) The Fourier transform of the ACF admits

$$f(\omega) \sim \omega^{-2d}, \quad \text{as } \omega \rightarrow 0.$$

— the spectral density function

—  $d$  can be estimated from log-periodogram with  $\omega$  small.

**FARIMA:** In general, if  $\{X_t\}$  satisfies

$$b(B)(1-B)^d X_t = a(B)\varepsilon_t,$$

then it is called an FARIMA( $p, d, q$ )  $\iff$  Fractional ARIMA model.

Note that

$$a(B)^{-1}b(B)(1-B)^d X_t = \varepsilon_t \iff a(B)^{-1}b(B)X_t = \eta_t,$$

where  $\eta_t = (1 - B)^{-d}\varepsilon_t$  is a long-memory process. Thus, FARIMA can be viewed as an ARMA model driven by a long memory noise.

## 2.12 Seasonal Models

Some financial time series exhibits periodic behavior. e.g. quarterly earning per share of Johnson and Johnson 1960-1980. The features of earning data include serial correlation and seasonality.

### Differencing:

— Serial correlation is weakened by a differencing:  $\Delta X_t = X_t - X_{t-1}$ .

The ACF is given in Figure 2.15, labeled (dx).

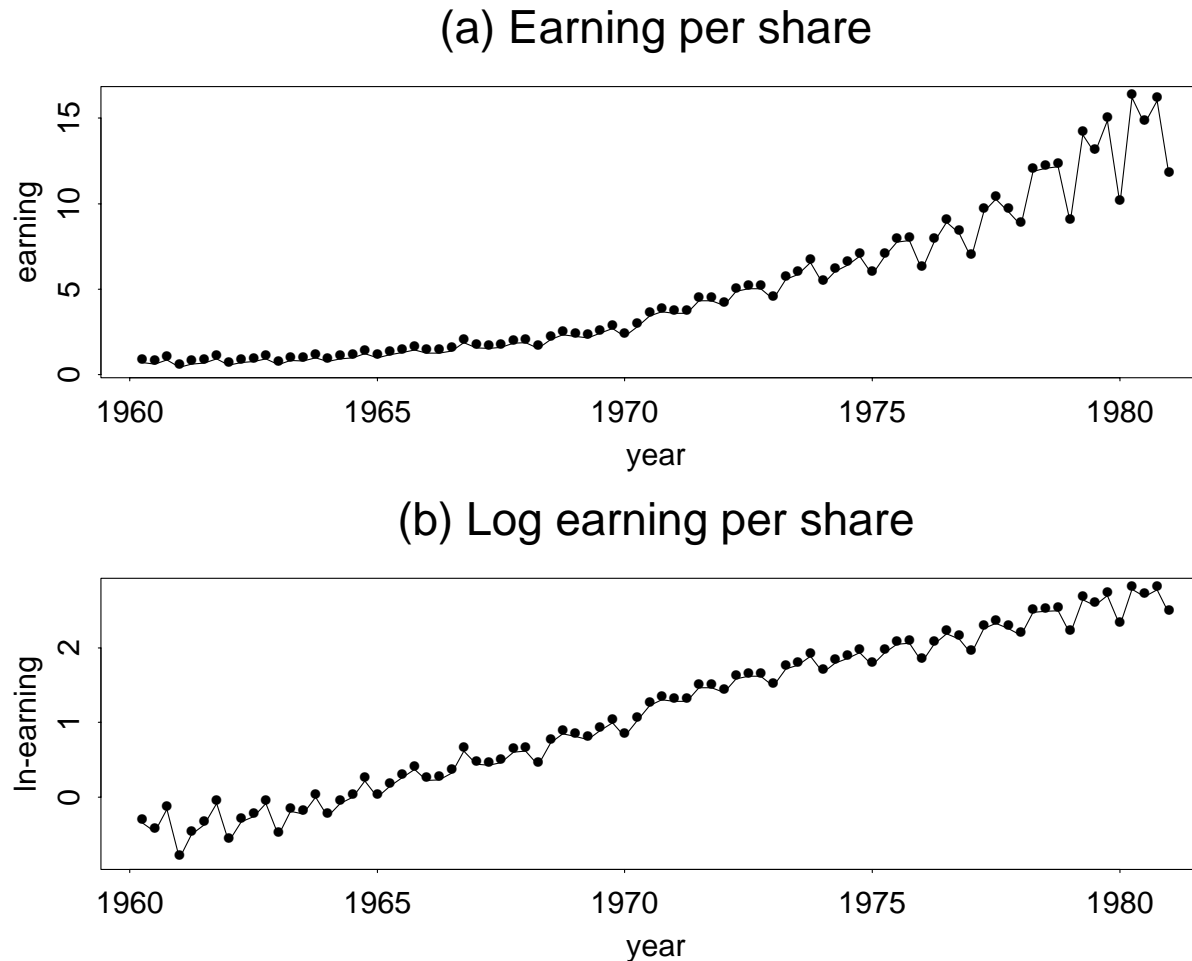


Figure 2.15: Quarterly earnings of Johnson and Johnson 1960 – 1980.

— Seasonality is handled by a seasonal difference:  $\Delta_4 = (1 - B^4)$ .  
 The ACF is given in Figure 2.15, labeled (ds).

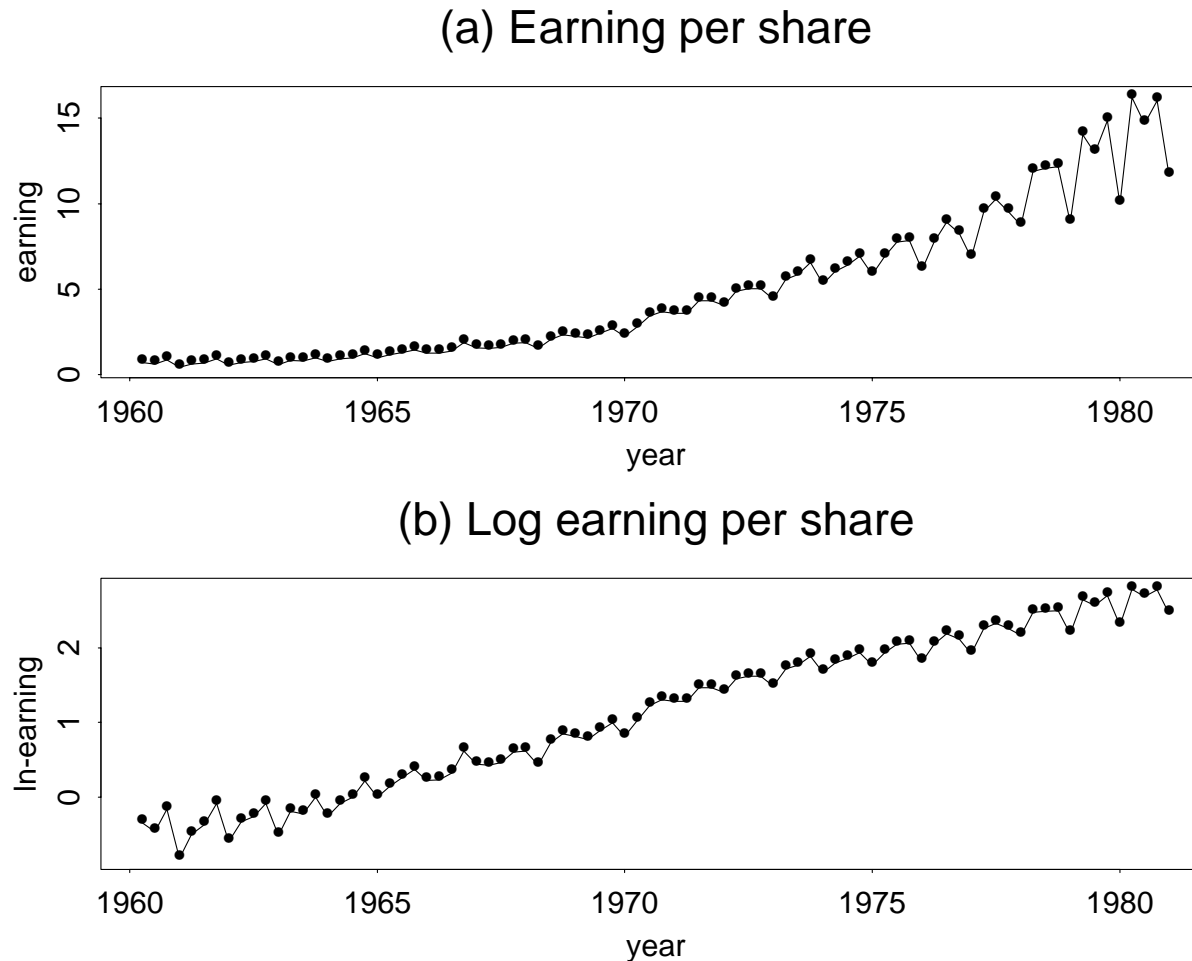


Figure 2.15: Quarterly earnings of Johnson and Johnson 1960 – 1980.

— Seasonality is handled by a seasonal difference:  $\Delta_4 = (1 - B^4)$ .  
 The ACF is given in Figure 2.15, labeled (ds).

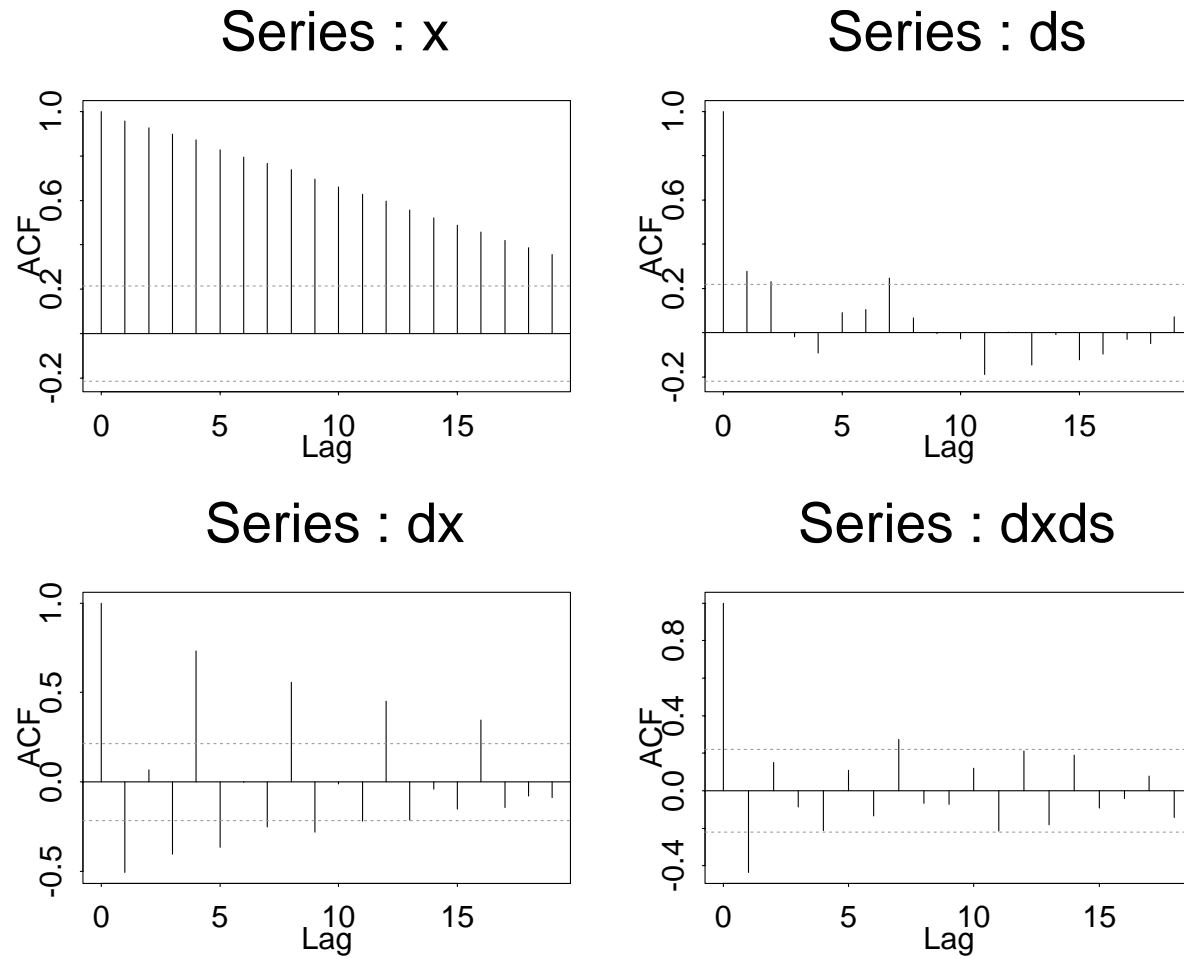


Figure 2.16: ACFS of earning series based on differences.

— seasonal difference with periodicity  $m$  means  $\Delta_m = (1 - B^m)$ .

— Serial correlation and seasonality are handled by

$$(1 - B^4)(1 - B)X_t = X_t - X_{t-1} - X_{t-4} + X_{t+5}.$$

The resulting series has ACF given in Figure 2.11 (d).

After the above preprocessing, we can construct a periodic seasonal model such as

### Airline model:

$$(1 - B^m)(1 - B)X_t = (1 - \theta B)(1 - \eta B^m)\varepsilon_t, \quad |\theta| < 1, |\eta| < 1.$$

— Serial correlation and seasonality are handled by

$$(1 - B^4)(1 - B)X_t = X_t - X_{t-1} - X_{t-4} + X_{t+5}.$$

The resulting series has ACF given in Figure 2.11 (d).

After the above preprocessing, we can construct a periodic seasonal model such as

### Airline model:

$$(1 - B^m)(1 - B)X_t = (1 - \theta B)(1 - \eta B^m)\varepsilon_t, \quad |\theta| < 1, |\eta| < 1.$$

Letting  $Y_t = (1 - B^m)(1 - B)X_t$ . This is a specific MA model for  $Y_t$ :

$$Y_t = \varepsilon_t - \theta\varepsilon_{t-1} - \eta\varepsilon_{t-m} + \theta\eta\varepsilon_{t-m-1}.$$

For the quarterly earning data, based on the MLE, we have

$$(1 - B)(1 - B^4)X_t = (1 - 0.678B)(1 - 0.314B^4)\varepsilon_t, \quad \sigma = 0.089.$$

(0.080)                      (0.101)

The estimation was based on 84 data points. The Ljung-Box statistic of residuals show  $Q(12) = 10.0$  with p-value 0.44. Based on first 76 quarters data, the 8-quarter predictions are shown below.



Figure 2.17: Forecasted and actual earnings