

# Chapter 5

## Linear Stochastic Models

### 5.1 Least Squares

Suppose that we observe some dependent variable (e.g. number of red blood cells) as a function of some independent variable (e.g. dosage of a drug), based on  $n$  experiments. The empirical data consists of  $(x_1, y_1), \dots, (x_n, y_n)$ , where  $x_i$  is the value of the independent variable used for the  $i$ 'th experiment, and  $y_i$  is the corresponding value of the dependent variable.

We wish to find a simple mathematical model that summarizes the relationship of the dependent variable to the independent variable. The simplest such model is a linear model of the form

$$y(x) = ax + b$$

If  $n = 2$  with  $x_1 \neq x_2$ , we can find  $a$  and  $b$  by solving a linear system of two equations in two unknowns. If  $n > 2$ , the system is overdetermined, and we can apply the method of “least squares”.

Specifically, let  $\vec{y} = (y_1, \dots, y_n)^T$ ,  $\vec{x} = (x_1, \dots, x_n)^T$ ,  $\vec{e} = (1, 1, \dots, 1)^T$ . A reasonable means of finding the “best” values of  $a$  and  $b$  is to select  $a$  and  $b$  as to minimize some measure of distance between  $\vec{y}$  and  $a\vec{x} + b\vec{e}$ . One notion of distance that leads to a particularly nice system of determining equations for  $a$  and  $b$  is to measure the distance between  $\vec{z}_1, \vec{z}_2 \in \mathbb{R}^n$  via

$$\|\vec{z}_1 - \vec{z}_2\|,$$

where

$$\|\vec{w}\|^2 = \vec{w}^T \Lambda \vec{w}.$$

Here,  $\Lambda$  is a given  $n \times n$  symmetric positive definite matrix. The minimizers  $a, b$  of the sum of squares

$$\min_{a,b} \|\vec{y} - a\vec{x} - b\vec{e}\|^2$$

satisfy the linear system

$$\begin{pmatrix} \vec{x}^T \Lambda \vec{x} & \vec{e}^T \Lambda \vec{x} \\ \vec{x}^T \Lambda \vec{e} & \vec{e}^T \Lambda \vec{e} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \vec{y}^T \Lambda \vec{x} \\ \vec{y}^T \Lambda \vec{e} \end{pmatrix}.$$

Our choice of a quadratic form as our (squared) notion of distance is what leads to a linear system for the minimizer. Other notions of distance would lead to more complex optimization problem. The case in which  $\Lambda = I$  is called “ordinary least squares”, whereas  $\Lambda \neq I$  is called “weighted least squares”.

This approach to fitting a linear model to observed data leaves open several questions:

1. How should one choose the matrix  $\Lambda$ ?
2. How can one assign “error bars” to our slope and intercept values  $a$  and  $b$ ?

3. Is there any way to objectively test the linear model against the even simpler model in which  $a = 0$  (the “constant” model)?

A statistical formulation of this linear modeling problem will permit us to address these issues.

## 5.2 Linear Regression Models with Gaussian Residuals

We turn to the statistical view of how to build a linear model. We now view the dependent data values as random variables. In particular, we assume that  $Y_i$  is a rv corresponding to the measured response of the dependent variable (e.g. blood pressure) as a function of the independent variable (e.g. drug dosage). Specifically, the “linear regression” model assumes that

$$Y_i = a^* x_i + b^* + \varepsilon_i$$

where  $a^*$  and  $b^*$  are the “true” slope and intercept values, and  $\varepsilon_i$  is a rv describing the “residual error” in the linear model corresponding to observation  $Y_i$ . The great majority of the literature on linear regression presumes that  $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$  is a Gaussian random vector with mean  $\vec{0}$  and covariance matrix  $\sigma^2 C$ , where  $\sigma^2$  is unknown (and  $C$  is specified by the statistician and is therefore known). We follow the literature here, and make the assumption that the residuals have this Gaussian structure.

This statistical model has three unknown parameters, namely  $a^*$ ,  $b^*$  and  $\sigma^2$ . The principle of maximum likelihood asserts that  $a^*$ ,  $b^*$  and  $\sigma^2$  should be estimated as the maximizer  $(\hat{a}, \hat{b}, \hat{\sigma}^2)$  of the likelihood.

$$\frac{1}{(2\pi\sigma^2)^{n/2} |\det C|^{1/2}} \exp\left(-\frac{1}{2}(\vec{Y} - a\vec{x} + b\vec{e})^T C^{-1}(\vec{Y} - a\vec{x} + b\vec{e})\right).$$

Here,  $\vec{Y} = (Y_1, \dots, Y_n)^T$ . (This likelihood presumes that  $C$  has been specified as a positive definite matrix. Any reasonable model for  $C$  will have this property.)

The estimators satisfy

$$\begin{pmatrix} \vec{x}^T C^{-1} \vec{x} & \vec{e}^T C^{-1} \vec{x} \\ \vec{x}^T C^{-1} \vec{e} & \vec{e}^T C^{-1} \vec{e} \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \vec{Y}^T C^{-1} \vec{x} \\ \vec{Y}^T C^{-1} \vec{e} \end{pmatrix}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} (\vec{Y} - \hat{a}\vec{x} - \hat{b}\vec{e})^T C^{-1} (\vec{Y} - \hat{a}\vec{x} - \hat{b}\vec{e}).$$

It is common to choose  $C = I$  in the linear regression model. This corresponds to an assumption of *iid* residual errors. However, one need not choose  $C = I$ . For example, suppose that one assumes that the variability of  $Y_i$  scales with the magnitude of  $x_i$ . In this case, one would set  $C = \text{diag}(x_1^2, x_2^2, \dots, x_n^2)$ . This leads to a “weighted least squares” problem. Note that the statistical formulation helps suggest plausible forms for  $C$ .

This statistical framework also permits us to develop “error bars” for our estimates of  $a^*$  and  $b^*$ . Let

$$\bar{x} = \frac{1}{n} \sum_1^n x_i,$$

$$s_{xx} = \sum_1^n (x_i - \bar{x})^2,$$

$$\text{SSE} = \text{sum of squares of estimated residuals} = \sum_1^n (Y_i - \hat{a}x_i - \hat{b})^2.$$

It can be shown that when  $C = I$ ,

$$\frac{\hat{a} - a^*}{\sqrt{\frac{\text{SSE}/(n-2)}{s_{xx}}}} \stackrel{\mathcal{D}}{=} t_{n-2},$$

$$\frac{\hat{b} - b^*}{\sqrt{\frac{\text{SSE}}{(n-2)} \left( \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)}} \stackrel{\mathcal{D}}{=} t_{n-2},$$

where  $t_{n-2}$  is a so-called Student-t rv with  $n - 2$  degrees of freedom (and is a “tabulated distribution”). It follows that if one selects  $z$  so that  $\text{P} \{-z \leq t_{n-2} \leq z\} = 1 - \delta$ , then

$$\left[ \hat{a} - z \sqrt{\frac{\text{SSE}/(n-2)}{s_{xx}}}, \hat{a} + z \sqrt{\frac{\text{SSE}/(n-2)}{s_{xx}}} \right],$$

$$\left[ \hat{b} - z \sqrt{\frac{\text{SSE}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)}, \hat{b} + z \sqrt{\frac{\text{SSE}}{n-2} \left( \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)} \right]$$

are *exact*  $100(1 - \delta)\%$  confidence interval for  $a^*$  and  $b^*$ , respectively<sup>1</sup>

We turn next to the issue of testing the linear model ( $a^* \neq 0$ ) versus the constant model ( $a^* = 0$ ) when  $C = I$ . Under the hypothesis that  $a^* = 0$ ,

$$\frac{\hat{a}^2 s_{xx}}{\text{SSE}/(n-2)} \stackrel{\mathcal{D}}{=} F_{1,n-2}$$

where  $F_{1,n-2}$  is a rv having the  $F$  distribution with 1 and  $n - 2$  degrees of freedom. If we choose  $z$  so that  $\text{P} \{F_{1,n-2} > z\} = \gamma$  (with  $\gamma$  small), then it is rare that the statistic  $\frac{\hat{a}^2 s_{xx}}{\text{SSE}/(n-2)}$  exceeds  $z$  when  $a^* = 0$ . (A common value of  $\gamma$  is 0.05.) Hence, if  $\frac{\hat{a}^2 s_{xx}}{\text{SSE}/(n-2)} \leq z$ , we view the data as being consistent with  $a^* = 0$  (i.e. the “constant” model), whereas if the statistic is larger than  $z$ , we reject the hypothesis that  $a^* = 0$ .

### 5.3 Linear Regression Model with non-Gaussian Residuals

In some applications settings, one may prefer to avoid assuming that the residuals are Gaussian. Here, we illustrate the use of the bootstrap in this context.

We assume that for  $1 \leq i \leq n$ ,

$$Y_i = a^* x_i + b^* + \varepsilon_i$$

where  $a^*$  and  $b^*$  are the “true” shape and intercept values, and  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  are *iid* rv’s with  $\text{E}[\varepsilon_i] = 0$  and  $\text{var}(\varepsilon_i) < \infty$ . Because this is a model in which  $C$  is implicitly assumed to be the identity, we estimate  $a^*$  and  $b^*$  as in ordinary least squares:

$$\hat{a} = \frac{\sum_1^n (x_i Y_i - \bar{x} \bar{Y})}{\sum_1^n (x_i - \bar{x})^2}$$

$$\hat{b} = \bar{Y} - \hat{a} \bar{x}$$

where  $\bar{Y} = \frac{1}{n} \sum_1^n Y_i$ . Under modest assumptions on the  $x_i$ ’s, it can be shown that

$$\hat{a} \xrightarrow{p} a^*$$

$$\hat{b} \xrightarrow{p} b^*$$

<sup>1</sup>See Chapter 12 of *Probability and Statistics for the Engineering, Computing, and Physical Sciences* by E.R. Dougherty. Prentice Hall (1990), for details.

as  $n \rightarrow \infty$ . This guarantees that when  $n$  is large, each of the estimated residuals

$$\hat{\varepsilon}_i = Y_i - \hat{a}x_i - \hat{b}$$

is close to its corresponding true residual.

To construct error bars in this setting, we apply the bootstrap. If we knew  $a^*$ ,  $b^*$ , and the exact distribution of the residuals, we could use Monte Carlo simulation to numerically compute the distribution of (for example)

$$\frac{\hat{a} - a^*}{\sqrt{\frac{\text{SSE}/(n-2)}{s_{xx}}}}.$$

(Of course, in the Gaussian setting with  $C = I$ , this is known to be a  $t_{n-2}$  rv. In the non-Gaussian setting, this rv has a complicated and unknown distribution.) Specifically, we could sample the distribution of the  $\varepsilon_i$ 's  $n$  iid times, yielding  $\varepsilon_{11}, \dots, \varepsilon_{1n}$ . Set  $Y_{1i} = a^*x_i + b^* + \varepsilon_{1i}$ , for  $1 \leq i \leq n$ , and compute the ordinary least squares estimates  $\hat{a}_1$  and  $\hat{b}_1$  corresponding to the data set  $(x_1, Y_{11}), \dots, (x_n, Y_{1n})$ . If we repeat the process  $m$  independent times (for a total of  $mn$  samples from the distribution of the  $\varepsilon_i$ 's), thereby yielding  $\hat{a}_1, \hat{b}_1, \dots, \hat{a}_m, \hat{b}_m$ , we could estimate the required distribution via

$$\frac{1}{m} \sum_{i=1}^m I\left(\frac{\hat{a}_i - a^*}{\sqrt{\frac{\sum_1^n (Y_{ij} - \hat{a}_i x_j - \hat{b}_i)^2 / (n-2)}{s_{xx}}}}\right) \leq \cdot).$$

Of course, we generally don't have the ability to cheaply obtain  $mn$  such samples from the distribution of the  $\varepsilon_i$ 's.

The bootstrap philosophy replaces  $a^*$  by  $\hat{a}$ ,  $b^*$  by  $\hat{b}$ , and the distribution of the  $\varepsilon_i$ 's by the  $\hat{\varepsilon}_i$ 's. Sample the  $\hat{\varepsilon}_i$ 's  $n$  iid times (with replacement), thereby yielding  $\varepsilon_{11}^*, \dots, \varepsilon_{1n}^*$ , and compute the ordinary least squares estimator,  $\hat{a}_1^*$  and  $\hat{b}_1^*$ , corresponding to the data set  $(x_1, Y_{11}^*), \dots, (x_n, Y_{1n}^*)$ , where  $Y_{1j}^* = \hat{a}x_j + \hat{b} + \varepsilon_{1j}^*$ . We now repeat this process  $m$  independent times, thereby yielding  $m$  bootstrap estimates  $\hat{a}_1^*, \hat{b}_1^*, \dots, \hat{a}_m^*, \hat{b}_m^*$ . The required distribution can be estimated via

$$\frac{1}{m} \sum_{i=1}^m I\left(\frac{\hat{a}_i^* - \hat{a}}{\sqrt{\frac{\sum_1^n (Y_{ij}^* - \hat{a}_i^* x_j - \hat{b}_i^*)^2 / (n-2)}{s_{xx}}}}\right) \leq \cdot).$$

The above estimated distribution is then used to construct a confidence interval for  $a^*$  in the usual way.

A similar bootstrap method can be used to produce confidence intervals for  $b^*$  (that are asymptotically valid as  $m, n \rightarrow \infty$ ) or to produce asymptotically valid hypothesis testing regions.

**Remark 5.1:** Suppose that we assume the  $\varepsilon_i$ 's have a covariance matrix that is known up to an (unknown) factor  $\sigma^2$ , so that

$$Y_i = a^*x_i + b^* + \varepsilon_i,$$

where  $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$  is assumed to have a positive definite covariance matrix  $\sigma^2 C$ , where  $C$  is known and  $\sigma^2$  is assumed unknown. In this case, we can use the fact that  $C$  is known to compute the Cholesky factorization  $C = LL^T$ . Note that

$$L^{-1}\vec{Y} = a^*L^{-1}\vec{x} + b^*L^{-1}\vec{e} + L^{-1}\vec{\varepsilon}.$$

Hence, if we set  $\vec{Z} = L^{-1}\vec{Y}$ ,  $\vec{w} = L^{-1}\vec{x}$ ,  $\vec{v} = L^{-1}\vec{e}$ , and  $\vec{\nu} = L^{-1}\vec{\varepsilon}$ , we arrive at the model

$$\vec{Z} = a^*\vec{w} + b^*\vec{v} + \vec{\nu},$$

where  $\vec{\nu}$  has mean zero and  $\sigma^2 I$  as its covariance matrix. If we now additionally assume that the  $\nu_i$ 's are iid, the bootstrap can be applied to this transformed model (and hence to the original model with covariance matrix  $\sigma^2 C$ ).

## 5.4 Data Transformations

In many applied settings, one expects that a non-linear model might offer a better explanation of the data. For example, one might postulate basic trends of the form:

$$y(x) = c \exp(ax) \quad (\text{exponential trend})$$

or

$$y(x) = cx^a \quad (\text{power law trend}).$$

Simple data transformations reduce these models to a linear model. In the presence of an exponential trend, fit the linear regression model to  $(x_1, \log Y_1), \dots, (x_n, \log Y_n)$ , while one fits a linear model to  $(\log x_1, \log Y_1), \dots, (\log x_n, \log Y_n)$  in the presence of a power law trend.

## 5.5 Multiple Linear Regression

Of course, it is common in many applications to try to explain a dependent variable (e.g. blood pressure) as a function of multiple explanatory variables (e.g. dosage of a drug, body weight). Here, we have an  $\mathbb{R}^d$ -valued vector  $x_i$  that represents the levels of the  $d$  explanatory variables associated with experimental value  $i$ , and  $Y_i$  is the corresponding value of the dependent variable. We assume that

$$Y_i = a^{*T} x_i + b^* + \varepsilon_i$$

for  $1 \leq i \leq n$ , where  $a^* \in \mathbb{R}^d$  and  $b^*$  are the “true” parameters, and  $(\varepsilon_1, \dots, \varepsilon_n)^T$  is an  $n$  dimensional Gaussian rv with mean 0 and covariance matrix  $\sigma^2 C$  with unknown  $\sigma^2$  (but known  $C$ ).

Here, the likelihood is given by

$$\frac{1}{(2\pi\sigma^2)^{n/2} |\det C|^{1/2}} \exp\left(-\frac{1}{2}(\vec{Y} - xa - b\vec{e})^T C^{-1}(\vec{Y} - xa - b\vec{e})\right)$$

where  $\vec{Y} = (Y_1, \dots, Y_n)^T$  and  $x$  is the  $n \times d$  matrix in which the  $i$ 'th row is  $x_i$ . The maximum likelihood estimators  $\hat{a}$ ,  $\hat{b}$  and  $\hat{\sigma}^2$ , satisfy

$$\begin{pmatrix} \vec{x}^T C^{-1} \vec{x} & \vec{x}^T C^{-1} \vec{e} \\ \vec{e}^T C^{-1} \vec{x} & \vec{e}^T C^{-1} \vec{e} \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \vec{x}^T C^{-1} \vec{Y} \\ \vec{e}^T C^{-1} \vec{Y} \end{pmatrix}$$

$$\hat{\sigma}^2 = \frac{1}{n} (\vec{Y} - x\hat{a} - \hat{b}\vec{e})^T C^{-1} (\vec{Y} - x\hat{a} - \hat{b}\vec{e})$$

All the ideas described in the context of (simple) linear regression models with  $d = 1$  generalize in a suitable way to the multiple linear regression context: confidence region procedures for  $a^*$ , hypothesis testing, bootstrap procedures for non-Gaussian residuals, etc.

## 5.6 The Correlation Model

Here, we adapt the linear regression model slightly, so that the  $x_i$ 's are now modeled themselves as random variables (so, in particular, the experimentalist does not control the  $x_i$ 's at which measurements are gathered). For example, we might collect  $n$  specimens of (say) a fish, and study the relationship between the (random) weight  $X_i$  of the  $i$ th fish, and the amount  $Y_i$  of pollutant stored in the tissues of the fish (for  $1 \leq i \leq n$ ).

The precise statistical specification of this so-called “correlation” model assumes that

$$Y_i = a^{*T} X_i + b^* + \varepsilon_i$$

for some “true”  $a^* \in \mathbb{R}^d$  and  $b^* \in \mathbb{R}$ , where  $((X_i, \varepsilon_i) : 1 \leq i \leq n)$  is a set of  $n$  iid pairs with  $E[\varepsilon_i] = 0$ , and  $\text{var}(\varepsilon_i) < \infty$ . (We permit  $X_i \in \mathbb{R}^d$  to be vector valued, so as to permit  $Y_i$  to depend on multiple characteristics of each specimen.) Put  $\tilde{X}_i = X_i - E[X_i]$ , and  $\tilde{Y}_i = Y_i - E[Y_i]$ , for  $1 \leq i \leq n$ . Note that the best affine predictor of  $Y_i$  given  $X_i$  must be  $a^{*T}X_i + b^*$ , and hence

$$a^* = (E[\tilde{X}_1 \tilde{X}_1^T])^{-1} E[\tilde{X}_1 \tilde{Y}_1]$$

$$b^* = E[Y_1] - a^{*T} E[X_1]$$

(We assume here, and throughout, that the covariance matrix,  $E[\tilde{X}_1 \tilde{X}_1^T]$ , is non-singular.)

We describe now the bootstrap procedure that would be used to deal with such a correlation model (in the presence of non-Gaussian residual errors).

**Exercise 5.1:** Suppose that  $E[\|X_1\|^2] < \infty$  and  $E[\varepsilon_1^2] < \infty$ . Put

$$\hat{a}_n = \left( \frac{1}{n} \sum_1^n (X_i - \bar{X}_n)^T (X_i - \bar{X}_n) \right)^{-1} \cdot \left( \frac{1}{n} \sum_1^n (X_i - \bar{X}_n)^T (Y_i - \bar{Y}_n) \right),$$

$$\hat{b}_n = \bar{Y}_n - \hat{a}_n \bar{X}_n$$

where

$$\bar{X}_n = \frac{1}{n} \sum_1^n X_i,$$

$$\bar{Y}_n = \frac{1}{n} \sum_1^n Y_i.$$

Prove that

$$\hat{a}_n \rightarrow \hat{a} \quad \text{a.s.},$$

$$\hat{b}_n \rightarrow \hat{b} \quad \text{a.s.}$$

as  $n \rightarrow \infty$ .

According to Problem Exercise 5.1,  $\hat{a}_n$  and  $\hat{b}_n$  are (for large sample sizes  $n$ ) close to  $a^*$  and  $b^*$ . Suppose that we sample  $(X_{11}^*, Y_{11}^*), \dots, (X_{1n}^*, Y_{1n}^*)$ , from the collection of observations  $(X_1, Y_1), \dots, (X_n, Y_n)$ , independently (and with replacement). Put

$$\bar{X}_1^* = \frac{1}{n} \sum_1^n X_i$$

$$\bar{Y}_1^* = \frac{1}{n} \sum_1^n Y_i$$

$$\hat{a}_1^* = \left( \frac{1}{n} \sum_{i=1}^n (X_{1i}^* - \bar{X}_1^*)^T (X_{1i}^* - \bar{X}_1^*) \right)^{-1} \cdot \left( \frac{1}{n} \sum_{i=1}^n (Y_{1i}^* - \bar{Y}_1^*)^T (X_{1i}^* - \bar{X}_1^*) \right)$$

$$\hat{b}_1^* = \bar{Y}_1^* - \hat{a}_1^* \bar{X}_1^*$$

If we independently generate  $m$  such bootstrap samples from  $(X_1, Y_1), \dots, (X_n, Y_n)$ , we obtain  $m$  pairs  $(\hat{a}_1^*, \hat{b}_1^*), \dots, (\hat{a}_m^*, \hat{b}_m^*)$ . If  $m, n$  are both large, the distribution of

$$\frac{1}{n} \sum_{i=1}^n I((X_i - \bar{X}_n)^T (X_i - \bar{X}_n))^{-\frac{1}{2}} (\hat{a}_n - a^*)$$

can be approximated by

$$\frac{1}{m} \sum_{i=1}^m I\left(\left(\frac{1}{n} \sum_{j=1}^n (X_{ij}^* - \bar{X}_i^*)^T (X_{ij}^* - \bar{X}_i^*)\right)^{-\frac{1}{2}} (\hat{a}_i^* - \hat{a}_n)\right) \leq \cdot$$

This bootstrap procedure can be used to construct confidence regions for  $a^*$  and  $b^*$  for the correlation model, as well as hypothesis testing regions<sup>2</sup>.

## 5.7 Modeling Deterministic Dynamical Systems via Differential Equations

Let  $y = (y(t) : t \geq 0)$  be a deterministic dynamical system described by a  $p^{\text{th}}$  order differential equation, namely

$$y^{(p)} = f(y(t), y^{(1)}(t), \dots, y^{(p-1)}(t))$$

for some given function  $f : \mathbb{R}^p \mapsto \mathbb{R}$ . Of course, such a  $p^{\text{th}}$  order equation can always be reduced to a first order equation by introducing a suitable state variable, namely

$$x(t) = (y(t), y^{(1)}(t), \dots, y^{(p-1)}(t))^T.$$

Then we have  $\dot{x}(t) = g(x(t))$ , where  $g : \mathbb{R}^p \mapsto \mathbb{R}^p$  is given by

$$g(x_1, \dots, x_p) = (x_2, \dots, x_p, f(x_1, \dots, x_{p-1}))^T.$$

An especially important case is that of a  $p^{\text{th}}$  order linear differential equation of the form

$$y^{(p)} = \sum_{j=1}^p \beta_j y^{(p-j)} + c \tag{5.1}$$

in which case we have

$$\frac{d}{dt} \begin{pmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(p-1)}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ \beta_p & \beta_{p-1} & \beta_{p-2} & \dots & \beta_1 \end{pmatrix} \begin{pmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(p-1)}(t) \end{pmatrix} \tag{5.2}$$

## 5.8 Linear Difference Equation of $p^{\text{th}}$ Order

**Definition 5.1:** Given a sequence  $(y_n : n \geq 0)$ , define the  $p^{\text{th}}$  difference by

$$\Delta^1 y_n = y_{n+1} - y_n$$

for  $k = 1$  and for  $k > 1$ ,

$$\Delta^k y_n = \Delta^{k-1} y_{n+1} - \Delta^{k-1} y_n$$

Then the discrete-time analog to (5.1) is

$$\Delta^p y_n = \sum_{j=1}^p \beta_j \Delta^{p-j} y_n + c \tag{5.3}$$

---

<sup>2</sup>See Chapter 4 of The Bootstrap and Edgeworth Expansion by Peter Hall, Springer-Verlag (1992) for details.

## 5.9 Stochastic Linear Difference Equations of $p^{\text{th}}$ Order

The stochastic analog to a constant sequence  $z_n = c$  is an iid sequence  $(V_n : n \geq 0)$ . Hence, the natural stochastic analog to (5.3) is a stochastic sequence  $(Y_n : n \geq 0)$  satisfying

$$\Delta^p Y_n = \sum_{j=1}^p \beta_j \Delta^{p-j} Y_n + V_n. \quad (5.4)$$

Now observe that

$$\Delta^k Y_n = \sum_{j=1}^k \binom{k}{j} (-1)^{k-j} Y_{n+j}.$$

As a consequence, we may write (5.4) in the form

$$Y_n = \sum_{j=1}^p a_j Y_{n-j} + V_n \quad (5.5)$$

for  $n \geq p$  (for suitably chosen  $a_j$ 's).

Note that  $Y_n$  is expressed as a linear combination of the  $p$  previous values of the  $Y$ -sequence, namely  $Y_{n-1}, \dots, Y_{n-p}$ . In other words,  $Y_n$  is “regressed” on the  $p$  previous values of the same  $Y$ -sequence, and hence it is “autoregressed”.

### Definition 5.2:

A sequence  $Y = (Y_n : n \geq 0)$  satisfying (5.5) with  $(V_n : n \geq 0)$  iid is called a  $p^{\text{th}}$  order **autoregressive** sequence.

The autoregressive sequence is said to be **Gaussian** if the  $V_n$ 's are Gaussian.

Any  $p^{\text{th}}$  order (scalar) autoregression can be expressed as a first order (vector) autoregression, by following the same idea as that leading to (5.2). Put

$$X_n = (Y_{n-p+1}, \dots, Y_n)^T$$

and note that

$$X_{n+1} = F X_n + Z_{n+1} \quad (5.6)$$

where

$$F = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ a_p & a_{p-1} & a_{p-2} & \dots & a_1 \end{pmatrix}$$

$$Z_{n+1} = (0, 0, \dots, 0, V_{n+1})^T$$

## 5.10 Stability Properties of the Autoregressive Sequences

If  $V_n = c$  in (5.5), the deterministic sequence governed by (5.5) remain bounded if and only if the spectral radius of  $F$  (i.e. the maximum of the moduli of the eigenvalues of  $F$ ) is less than 1. This turns out to be the right condition to guarantee “stability” of an autoregressive sequence.

**Exercise 5.2:**

Let  $X = (X_n : n \geq 0)$  satisfy (5.6) for  $n \geq 0$  with  $E\|Z_n\| < \infty$  and  $(Z_n : n \geq 0)$  iid.

1. Show that  $X_n = F^n X_0 + \sum_{j=0}^{n-1} F^j Z_{n-j}$ .
2. Prove that  $X_n \stackrel{\mathcal{D}}{=} F^n X_0 + \sum_{j=0}^{n-1} F^j Z_j$ .
3. Prove that if the spectral radius of  $F$  is less than one, then  $X_n \implies X_\infty$  as  $n \rightarrow \infty$ , where

$$X_\infty \stackrel{\mathcal{D}}{=} \sum_{j=0}^{\infty} F^j Z_j.$$

4. If the  $Z_n$ 's are Gaussian with covariance  $C$ , show that  $X_\infty$  is Gaussian with mean  $(I - F)^{-1} E Z_1$  and covariance matrix  $\Lambda$  satisfying

$$\Lambda = F \Lambda F^T + C.$$

5. Prove that  $\Lambda$  can be computed via the recursion

$$\Lambda_{n+1} = F \Lambda_n F^T + C$$

for  $n \geq 0$ , subject to  $\Lambda_0 = 0$ .

Requiring the eigenvalues of  $F$  to have moduli less than one is equivalent to requiring that the  $p$  roots  $z_1, \dots, z_p$  of the degree  $p$  polynomial

$$z^p - \sum_{j=1}^p a_j z^{p-j} \tag{5.7}$$

all have modulus less than one.

## 5.11 Stationary Version of a Stable Autoregressive Sequence

Suppose that the  $p$  roots of (5.7) are all less than one in modulus, and  $E\|Z_1\| < \infty$ . If we then initialize  $X$  at time  $-r$  at 0, then

$$X_k = \sum_{j=0}^{k+r-1} F^j Z_{k-j}$$

where  $(Z_n : n \geq 0)$  is a sequence of iid copies of the random variable  $Z_1$ . To indicate the dependence of  $X_k$  on  $r$ , we write it as  $X_{k,-r}$ . Observe that as  $r \rightarrow \infty$ ,

$$X_{k,-r} \rightarrow X_k^* \quad \text{a.s.}$$

for each  $k \in \mathbb{Z}$ , where

$$X_k^* = \sum_{j=0}^{\infty} F^j Z_{k-j} \stackrel{\mathcal{D}}{=} X_\infty$$

Note that  $X^* = (X_k^* : k \in \mathbb{Z})$  satisfies the recursion

$$X_{k+1}^* = F X_k^* + Z_{k+1}$$

and is stationary in the sense that  $(X_{m+k}^* : k \in \mathbb{Z}) \stackrel{\mathcal{D}}{=} (X_k^* : k \in \mathbb{Z})$

**Definition 5.3:**

The sequence  $X^*$  is said to be the **stationary version** of  $X$ .

We interpret a stationary process as representing a system that was initialized at time  $-\infty$  and is in stochastic equilibrium at every finite  $t$ .

## 5.12 Prediction for Autoregressive Sequences

Suppose that we wish to compute the best mean square predictor of  $X_{n+m}$ , given the past “history”  $X_j, j \leq n$ . If  $E\|Z_1\|^2 < \infty$ , this is just

$$E[X_{n+m}|X_j : j \leq n].$$

This, of course, is equal to

$$F^m X_n + \sum_{j=0}^{m-1} F^j E Z_1. \quad (5.8)$$

Hence, we can use this formula to predict  $Y_{n+m}$  based on  $(Y_n, Y_{n-1}, \dots, Y_{n-p+1})^T$  (equal to  $X_n^T$ ).

## 5.13 Parameter Estimation for Gaussian Autoregressive Sequences

In order to use an autoregressive model (in a real-world setting), we must first estimate the parameters from observed data. If we assume that the  $V_n$ 's are iid Gaussian with (unknown) mean  $\mu^*$  and (unknown) variance  $\sigma^{*2}$ , then the  $p^{\text{th}}$  order autoregressive model contains  $p+2$  unknown parameters, namely  $\alpha_1^*, \dots, \alpha_p^*, \mu^*$  and  $\sigma^{*2}$ . (Here  $\alpha_1^*, \dots, \alpha_p^*$  are the “true” autoregressive coefficients). Here, the “partial likelihood” (referred to as partial because it is a likelihood that conditions on  $Y_0, \dots, Y_{p-1}$  and does not take full advantage of the information that may be present in this initialization) based on observing  $(Y_j : 0 \leq j < n+p)$  is given by

$$(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=0}^{n-1} (Y_{i+p} - a_1 Y_{i+p-1} - \dots - a_p Y_i - \mu)^2\right).$$

The maximum likelihood estimators  $\hat{a}_1, \dots, \hat{a}_p, \hat{\mu}$  and  $\hat{\sigma}^2$  solve the linear system

$$\begin{pmatrix} \sum_{i=0}^{n-1} Y_{i+p-1}^2 & \sum_{i=0}^{n-1} Y_{i+p-1} Y_{i+p-2} & \dots & \sum_{i=0}^{n-1} Y_{i+p-1} \\ \sum_{i=0}^{n-1} Y_{i+p-1} Y_{i+p-2} & \sum_{i=0}^{n-1} Y_{i+p-2}^2 & \dots & \sum_{i=0}^{n-1} Y_{i+p-2} \\ \dots & \dots & \dots & \dots \\ \sum_{i=0}^{n-1} Y_{i+p-1} Y_i & \sum_{i=0}^{n-1} Y_{i+p-2} Y_i & \dots & \sum_{i=0}^{n-1} Y_i \\ \sum_{i=0}^{n-1} Y_{i+p-1} & \sum_{i=0}^{n-1} Y_{i+p-2} & \dots & n \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \\ \hat{\mu} \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n-1} Y_{i+p} Y_{i+p-1} \\ \sum_{i=0}^{n-1} Y_{i+p} Y_{i+p-2} \\ \vdots \\ \sum_{i=0}^{n-1} Y_{i+p} Y_i \\ \sum_{i=0}^{n-1} Y_{i+p} \end{pmatrix} \quad (5.9)$$

with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=0}^{n-1} (Y_{i+p} - \hat{a}_1 Y_{i+p-1} - \dots - \hat{a}_p Y_i - \hat{\mu})^2$$

As in the settings of conventional regression models, exact confidence regions and hypothesis testing have been developed in this Gaussian setting. Details can be found in the enormous literature on so-called “time series” models.

## 5.14 Parameter Estimation for Autoregressive Sequences with non-Gaussian Residuals

We now turn to the issue of how to deal with an autoregressive sequence  $Y = (Y_n : n \geq 0)$  for which

$$Y_n = a_1^* Y_{n-1} + \dots + a_p^* Y_{n-p} + \mu^* + \varepsilon_n$$

where  $(\varepsilon_n : n \geq 0)$  is a iid (possibly non-Gaussian) sequence with  $E[\varepsilon_0] = 0$  and  $\text{var}(\varepsilon_0) < \infty$ .

We first deal with the prediction problem in the presence of known parameters  $a_1^*, \dots, a_p^*, \mu^*$  and a known distribution for the  $\varepsilon_i$ 's. Conditional on  $(X_j : j \leq n)$ ,  $X_{m+n}$  has conditional mean

$$F^m X_n + \sum_{j=0}^{m-1} F^j E Z_1;$$

see (5.8) above. If the  $Z_n$ 's are Gaussian, the conditional distribution of  $X_{n+m}$  is

$$N\left(F^m X_n + \sum_{j=0}^{m-1} F^j E Z_1, \Lambda_m\right) \quad (5.10)$$

where  $\Lambda_m = F \Lambda_{m-1} F^T + E(Z_1 - E Z_1)(Z_1 - E Z_1)^T$  with  $\Lambda_0 = 0$ . This conditional distribution can be used to make predictions such as

$$P(Y_{n+m} > z | Y_j : 0 \leq j \leq n) \quad (5.11)$$

If the  $Z_n$ 's are non-Gaussian, computing (5.11) is non-trivial and must generally be implemented via Monte-Carlo. In particular, to compute the conditional distribution of  $X_{n+m}$  (conditional on  $X_j, j \leq n$ ), we generate  $mr$  independent copies of  $Z_1$ , call them  $Z_{1,1}, \dots, Z_{r,m}$  and use the Monte-Carlo estimator (based on  $r$  independent simulations of the history of  $X$  over  $[n, n+m]$ )

$$\frac{1}{r} \sum_{i=1}^r I\left(F^m X_n + \sum_{j=0}^{m-1} F^j Z_{i,j+1} \in \cdot\right)$$

to compute

$$P(X_{n+m} \in \cdot | X_j, j \leq n)$$

We now turn to the question of parameter estimation in the setting of non-Gaussian residuals. Note that

$$\langle \varepsilon_n, Y_{n-i} \rangle = 0$$

for  $i \geq 1$ , so that

$$\langle Y_n - a_1^* Y_{n-1} - \dots - a_p^* Y_{n-p} - \mu^*, Y_{n-i} \rangle = 0$$

for  $i \geq 1$ . In other words, the "true" parameters  $a_1^*, \dots, a_p^*$  and  $\mu^*$  satisfy the linear system

$$a_1^* E Y_{n-1} Y_{n-i} + \dots + a_p^* E Y_{n-p} Y_{n-i} + \mu^* E Y_{n-i} = E Y_n Y_{n-i}$$

for  $i \geq 1$ . A square linear system of  $p+1$  equations is obtained by taking the first  $p+1$  such equations (i.e.  $1 \leq i \leq p+1$ ).

Given observations  $(Y_j : 0 \leq j \leq n+p)$ , we can estimate  $E Y_{l-k} Y_{l-i}$  via  $Y_{l-k} Y_{l-i}$ , suggesting that we consider the linear system

$$\hat{a}_1 \frac{1}{n} \sum_{l=p+1}^{p+n} Y_{l-1} Y_{l-i} + \dots + \hat{a}_p \frac{1}{n} \sum_{l=p+1}^{p+n} Y_{l-p} Y_{l-i} + \hat{\mu} \frac{1}{n} \sum_{l=p+1}^{p+n} Y_{l-i} = \frac{1}{n} \sum_{l=p+1}^{p+n} Y_l Y_{l-i} \quad (5.12)$$

for  $1 \leq i \leq p+1$ . (Note the similarity of (5.12) to (5.9). (What explains the similarity?)

**Exercise 5.3:**

Suppose that the roots of (5.7) are all less than one in modulus, and assume that  $E\varepsilon_1^4 < \infty$ . Prove that  $\hat{a}_i \xrightarrow{p} a_i^*$ ,  $1 \leq i \leq p$  and that  $\hat{\mu} \xrightarrow{p} \mu^*$  as  $n \rightarrow \infty$ .

To produce confidence regions for  $a_1^*$ ,  $\dots$ ,  $a_p^*$  and  $\mu^*$ , we can apply the bootstrap idea. For  $p \leq i \leq n+p$ , let

$$\hat{\varepsilon}_i = Y_i - \hat{a}_1 Y_{i-1} - \dots - \hat{a}_p Y_{i-p} - \hat{\mu}$$

be the  $i^{\text{th}}$  estimated residual. To create a bootstrap sample of the autoregressive sequence, sample  $\varepsilon_{1,p}^*$ ,  $\dots$ ,  $\varepsilon_{1,n+p}^*$   $n+1$  independent times from the set of estimated residuals  $\{\hat{\varepsilon}_p, \dots, \hat{\varepsilon}_{n+p}\}$ . For  $p \leq i \leq n+p$ , compute

$$Y_{1,i}^* = \hat{a}_1 Y_{1,i-1}^* + \dots + \hat{a}_p Y_{1,i-p}^* + \hat{\mu} + \varepsilon_{1,i}^*$$

subject to  $Y_{1,j}^* = Y_j$  for  $0 \leq j < p$ .

From the bootstrapped autoregressive sequence  $(Y_{1,j}^* : 0 \leq j \leq n+p)$ , solve the linear system corresponding to (5.12) for  $\hat{a}_{1,1}^*$ ,  $\dots$ ,  $\hat{a}_{1,p}^*$ ,  $\hat{\mu}^*$ . If we repeat this bootstrap procedure  $m$  independent times, then

$$\frac{1}{m} \sum_{i=1}^m I(\hat{a}_{i,j}^* - \hat{a}_j \in \cdot)$$

will (for large  $m$  and  $n$ ) be close to

$$P(\hat{a}_j - a_j^* \in \cdot)$$

from which a large-sample confidence interval for  $a_j^*$  can be obtained. In a similar way, we can obtain a large-sample bootstrap confidence interval for  $\mu^*$ .

**Exercise 5.4:**

Extend the bootstrap procedure to produce prediction regions for  $Y_{n+m}$ , based on observing  $Y_j$ ,  $0 \leq j \leq n$ , that take into account parameter uncertainty in estimating  $a_1^*$ ,  $\dots$ ,  $a_p^*$  and  $\mu^*$  from the observed data.