

### III. Forecasting and Regression Models

For additional reading on this topic, please see Chapter 20 of the Course Reader.

#### 1 Forecasting

Forecasting is closely related to prediction. In these notes, we:

- describe the mathematical foundations of “prediction theory”
- describe the mathematical models that underline the forecasting methods described in the text.

#### Best Mean Square Prediction

Suppose, on the basis of historical information, that we know that the demand  $D$  for a product follows the density function  $f_D(x)$ . What should we use as our prediction  $\hat{D}$  for  $D$ ?

Answer: Choose  $\hat{D} = a$ , where  $a$  minimizes the “mean square prediction error”, namely choose  $a$  to minimize

$$E[(D - a)^2] = E[D^2] - 2aE[D] + a^2.$$

The minimizing value of  $a$  is  $a^* = E[D]$ . So, our prediction is just the expected value of  $D$ !

Suppose now that we observe some quantity  $X$  (e.g. GDP growth) that has historically been correlated with  $\hat{D}$ . Assuming that  $X = 1.7\%$  (say), the residual uncertainty in  $D$  is described by the conditional distribution. The best mean square predictor of  $D$ , given that  $X = 1.7$ , is then the mean of this conditional distribution, namely

$$E[D|X = 1.7].$$

In general, if we observe the values  $X_1 = x_1, \dots, X_d = x_d$ , the best mean square predictor of  $D$  will be the conditional expectation

$$E[D|X_1 = x_1, \dots, X_d = x_d].$$

#### Constant-Level Forecasting

Given a time series  $(Y_j : -\infty < j < \infty)$ , a “constant-level forecast” for  $Y_{n+1}$ , based on observing the past values  $Y_j, j \leq n$ , of the time series is

$$\hat{Y}_{n+1} = \mu.$$

Underlying Statistical Model: Suppose the  $Y_j$ 's are iid. Then,

$$\begin{aligned}\hat{Y}_{n+1} &= E[Y_{n+1}|Y_n = y_n, Y_{n-1} = y_{n-1}, \dots] \\ &= E[Y_{n+1}],\end{aligned}$$

so we obtain a “constant-level forecast” with  $\mu = E[Y_n]$ .

### Averaging Forecasting Method

Given a time series  $(Y_j : j \geq 1)$ , an “averaging forecast” for  $Y_{n+1}$  takes the form

$$\hat{Y}_{n+1} = \frac{1}{n}(Y_1 + \cdots + Y_n)$$

Underlying Statistical Model: Consider the constant-level forecast described above. If we have an enormous amount of statistical data available, we essentially will know the mean  $\mu$ . However, if our only knowledge of  $\mu$  comes from the observed time series values already collected, then we will need to estimate  $\mu$  from the observed data. The standard estimator for the mean of a population is the “sample mean”. So, if we have observed  $(Y_j : 1 \leq j \leq n)$ , then the sample mean is

$$\frac{1}{n}(Y_1 + \cdots + Y_n).$$

So, in this setting, the implemented version of constant level forecasting takes the form

$$\hat{Y}_{n+1} = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

### Last-Value Forecasting

Given a time series  $(Y_j : -\infty < j < \infty)$ , a “last-value forecast” for  $Y_{n+1}$ , based on  $Y_j, j \leq n$ , is

$$\hat{Y}_{n+1} = Y_n.$$

Underlying Statistical Model: Suppose

$$Y_{n+1} = Y_n + Z_{n+1}, \tag{3.1.1}$$

where the  $Z_i$ 's are iid with  $E[Z_i] = 0$ . Then,

$$\begin{aligned} \hat{Y}_{n+1} &= E[Y_n + Z_{n+1} | Y_n = y_n, Y_{n-1} = y_{n-1}, \dots] \\ &= y_n \end{aligned}$$

It is common in financial asset pricing theory to assume that the log-prices satisfy (3.1.1). If one believes that such a model governs log-prices, then “last-value forecasting” is appropriate.

### Moving Average Forecasting Method

Here, the “moving average forecast” takes the form

$$\hat{Y}_{n+1} = \frac{1}{d}(Y_n + Y_{n-1} + \cdots + Y_{n-d+1})$$

Underlying Statistical Model: This is the best mean square predictor if the time series obeys

$$Y_{n+1} = \frac{1}{d}(Y_n + Y_{n-1} + \cdots + Y_{n-d+1}) + Z_{n+1}$$

where the  $Z_i$ 's are iid with  $E[Z_i] = 0$ .

### Exponential Smoothing Forecasts

“Exponential smoothing forecasts” for a time series  $(Y_j : -\infty < j < \infty)$  follow the recursion

$$\hat{Y}_{n+1} = \alpha Y_n + (1 - \alpha)\hat{Y}_n,$$

for some constant  $\alpha \in (0, 1)$ .

Underlying Statistical Model: This is the best mean square predictor if the time series obeys

$$Y_{n+1} - Y_n = Z_{n+1} - (1 - \alpha)Z_n, \quad (3.1.2)$$

where the  $Z_n$ 's are iid with  $E[Z_n] = 0$ .

To see this, note that if  $\Delta Y_{n+1} = Y_{n+1} - Y_n$ , then we can re-write (3.1.2) as

$$\begin{aligned} Z_{n+1} &= (1 - \alpha)Z_n + \Delta Y_{n+1} \\ &= (1 - \alpha)[(1 - \alpha)Z_{n-1} + \Delta Y_n] + \Delta Y_{n+1} \\ &= \Delta Y_{n+1} + (1 - \alpha)\Delta Y_n + (1 - \alpha)^2 Z_{n-1} \\ &= \sum_{j=0}^{\infty} (1 - \alpha)^j \Delta Y_{n+1-j} \\ &= \sum_{j=0}^{\infty} (1 - \alpha)^j [Y_{n+1-j} - Y_{n-j}] \\ &= Y_{n+1} - \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j Y_{n-j} \end{aligned}$$

so

$$\begin{aligned} 0 &= E[Z_{n+1} | Y_n = y_n, Y_{n-1} = y_{n-1}, \dots] \\ &= \hat{Y}_{n+1} - \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j y_{n-j}. \end{aligned}$$

Hence,

$$\begin{aligned} \hat{Y}_{n+1} &= \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j Y_{n-j} \\ &= \alpha Y_n + (1 - \alpha)[\alpha Y_{n-1} + \alpha(1 - \alpha)Y_{n-2} + \dots] \\ &= \alpha Y_n + (1 - \alpha)\hat{Y}_n. \end{aligned}$$

This is a time series model that is widely applied by statisticians as a fit to real data. This, plus the simplicity of the forecast update rule, explains its popularity.

## 2 Statistical Modeling of Time Series

In predicting future values of a time series, there are three basic approaches that are followed :

1. forecasting method : These are easy to use, and only loosely supported by statistical principles. They are suitable for “quick and dirty” implementation.
2. statistical methods : These are based on sound statistical theory, and are widely used in applications for which lots of historical data is available.
3. stochastic modeling methods : This approach is also based on sound scientific principles, but relies on stochastic models (typically Markovian) to make predictions. These methods can be applied in settings where there is limited (or no) historical data available.

Our emphasis here will be on the second approach, namely that of statistical model-building. Three different types of data sets present themselves in applied work :

1. data sets in which some deterministic “trend” is present (but in which the “noise” around that trend is iid)
2. data sets in which no deterministic trend is present (but the time series is strongly autocorrelated)
3. data sets in which some deterministic trend is present and in which the “noise” around that trend is autocorrelated.

We will focus our discussion here on the first two cases.

### Time Series with Deterministic Trends : the Linear Case

Suppose that  $y(t)$  is the amount of carbon dioxide per cubic meter at time  $t$ . A simple model postulates a linear relationship :

$$y(t) = a^*t + b^*,$$

so that  $CO_2$  content increases linearly. Of course, if  $Y_i$  is the carbon dioxide content measured at time  $t_i$ ,  $Y_i$  does not follow a perfect straight line (perhaps because of measurement error or other complexities not reflected in the model). We model  $(Y_i : 1 \leq i \leq n)$  as

$$Y_i = a^*t_i + b^* + \epsilon_i, \tag{3.2.1}$$

where the  $\epsilon_i$ 's are iid  $N(0, \sigma_*^2)$  rv's.

**Question:** How do we estimate  $a^*$ ,  $b^*$ , and  $\sigma_*^2$  from the observed data?

**Remark:** If you are familiar with “linear regression”, you will recognize the model (3.2.1) as being a linear regression model.

To estimate  $a^*$ ,  $b^*$ , and  $\sigma_*^2$ , we use the method of maximum likelihood. The likelihood based on observing the time series value  $Y_i$  at time  $t_i$  (for  $1 \leq i \leq n$ ) is

$$L(a, b, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - at_i - b)^2}{2\sigma^2}\right)$$

so the log-likelihood is

$$\mathcal{L}(a, b, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n \frac{(Y_i - at_i - b)^2}{2\sigma^2}.$$

At a maximizer  $(\hat{a}, \hat{b}, \hat{\sigma}^2)$ ,  $\hat{a}$  and  $\hat{b}$  must solve the pair of linear equations

$$\begin{aligned} \sum_{i=1}^n (Y_i - \hat{a}t_i - \hat{b})t_i &= 0 \\ \sum_{i=1}^n (Y_i - \hat{a}t_i - \hat{b}) &= 0 \end{aligned}$$

The solution of this pair of linear equations is

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^n t_i Y_i - n(\sum_{j=1}^n Y_j/n)(\sum_{i=1}^n t_i/n)}{\sum_{i=1}^n t_i^2 - n(\sum_{j=1}^n t_j/n)(\sum_{i=1}^n t_i/n)} \\ \hat{b} &= \frac{1}{n} \sum_{i=1}^n Y_i - \hat{a} \frac{1}{n} \sum_{i=1}^n t_i. \end{aligned}$$

The standard estimator of  $\sigma_*^2$  is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}t_i - \hat{b})^2.$$

**Remark:** We divide by  $n-2$  in the definition of the estimator  $\hat{\sigma}^2$  to guarantee that  $\hat{\sigma}^2$  is unbiased as an estimator of  $\sigma^2$  (i.e.  $E(\hat{\sigma}^2) = \sigma^2$ ). Note also that if  $n = 2$ , we can always pass a straight line through the two points, so the variance of the “noise” appears to be zero when the sample size  $n \leq 2$ . Since we intuitively expect variability in the noise to be present, dividing by  $n-2$  (rather than, say,  $n-1$ ) renders the variance estimator undefined for  $n \leq 2$  (as is appropriate).

As usual, our best (mean square) predictor at some future time  $t$  for our time series is the conditional expectation of the distribution of  $Y$  at time  $t$ , namely  $a^*t + b^*$ . Substituting our estimators  $\hat{a}$  and  $\hat{b}$ , we obtain the predictor  $\hat{a}t + \hat{b}$ .

### **A. Prediction Interval at Time $t$**

Given our linear model (3.2.1), the value of  $Y$  at time  $t$  has a distribution that is normal with mean  $a^*t + b^*$  and variance  $\sigma_*^2$ . So, we expect the value of  $Y$  at time  $t$  to exhibit some “noise” around the trend value  $a^*t + b^*$ . We would like to construct a “prediction interval” for the time series value  $Y$  at time  $t$  that reflects the presence of this noise and that takes into account that because we do not know  $a^*$  and  $b^*$ , they must be estimated from the data via  $\hat{a}$  and  $\hat{b}$ .

To construct a prediction interval with confidence level  $100(1 - \delta)\%$ , we do the following :

1. We select a value  $z$  so that

$$P(-z \leq t_{n-2} \leq z) = 1 - \delta,$$

where  $t_{n-2}$  is a Student-t rv with  $n - 2$  degrees of freedom. (The Student-t rv has a tabulated distribution that can be found in many statistics books and is built into many spreadsheets.)

2. Compute

$$\hat{\gamma} = \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(t - \bar{t})^2}{\sum_{i=1}^n t_i^2 - n(\bar{t})^2} \right)}$$

where  $\bar{t} = n^{-1} \sum_{j=1}^n t_j$ .

3. The  $100(1 - \delta)\%$  prediction interval is

$$[\hat{a}t + \hat{b} - z\hat{\gamma}, \hat{a}t + \hat{b} + z\hat{\gamma}].$$

The interpretation of the prediction interval is that we are  $100(1 - \delta)\%$  confident that the time series value at time  $t$  will fall in the interval.

### **B. Confidence Intervals for the Slope and Intercept**

To construct a  $100(1 - \delta)\%$  confidence interval for the slope  $a^*$ , we :

1. Select a value of  $z$  so that

$$P(-z \leq t_{n-2} \leq z) = 1 - \delta,$$

where  $t_{n-2}$  is a Student-t rv with  $n - 2$  degrees of freedom.

2. The  $100(1 - \delta)\%$  confidence interval is :

$$\left[ \hat{a} - z \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n t_i^2 - n(\bar{t})^2}}, \hat{a} + z \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n t_i^2 - n(\bar{t})^2}} \right]$$

To construct a  $100(1 - \delta)\%$  confidence interval for the intercept  $b^*$ , we :

1. Select a value of  $z$  so that

$$P(-z \leq t_{n-2} \leq z) = 1 - \delta,$$

where  $t_{n-2}$  is a Student-t rv with  $n - 2$  degrees of freedom.

2. The  $100(1 - \delta)\%$  confidence interval for  $b^*$  is :

$$\left[ \hat{b} - z \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{\bar{t}^2}{\sum_{i=1}^n t_i^2 - n(\bar{t})^2} \right)}, \hat{b} + z \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{\bar{t}^2}{\sum_{i=1}^n t_i^2 - n(\bar{t})^2} \right)} \right]$$

**C. Testing the Hypothesis of a Linear Relationship**

Intuition tells us that we should always adopt the simplest model that is consistent with the data. A simpler model than our linear model (3.2.1) is to assume that  $a^* = 0$ , so that

$$Y_i = b^* + \epsilon_i$$

for  $1 \leq i \leq n$  (i.e. the  $Y_i$ 's are iid normal rv's with mean  $b^*$  and variance  $\sigma_*^2$ ).

So, we would like to test :

$$\begin{array}{ll} H_0(\text{ null hypothesis } ) & : \quad a^* = 0 \\ & \text{versus} \\ H_1(\text{ alternative hypothesis } ) & : \quad a^* \neq 0 \end{array}$$

Here is the test :

1. Choose a significance level  $\alpha$  for the test. (This is the probability that the test will reject the null hypothesis when the null hypothesis is correct.) Typical values for  $\alpha$  are  $\alpha = 0.01$  or  $\alpha = 0.05$ .
2. Find the value  $f$  so that

$$P(F_{1,n-2} \geq f) = \alpha,$$

where  $F_{1,n-2}$  is an F distribution with 1 and  $n - 2$  degrees of freedom. (This is a tabulated distribution.)

3. If

$$\frac{\sum_{i=1}^n (\hat{a}t_i + \hat{b} - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \hat{a}t_i - \hat{b})^2 / (n - 2)} \geq f,$$

reject the null hypothesis ; otherwise accept the null hypothesis.

Note that if  $a^* = 0$  (as assumed under the null hypothesis), then we expect the estimator  $\hat{a}$  to be close to zero, so each term appearing in the numerator of the test statistic should then be close to  $\hat{b} - \bar{Y}$ . But  $\bar{Y}$  is an estimator for  $b^*$  when  $a^*$  equals zero, so  $\hat{b} - \bar{Y}$  should be close to zero, and hence the test statistic should be small under the null hypothesis. This offers intuition as to why the hypothesis test takes the above form.

**Time Series with Deterministic Trends : the Nonlinear Case**

Because we wish to take advantage of the extensive set of tools available for linear regression, the standard approach for dealing with non-linear trends is to try to “transform the data” so that the trend becomes linear after the transformation.

**Exponential Trends:** Suppose that we believe that the basic form of the trend is exponential growth in time, so that the trend  $y(t)$  takes the form

$$y(t) = c^* \exp(a^*t).$$

Then,  $\log(y(t)) = \log(c^*) + a^*t = a^*t + b^*$ , where  $b^* = \log(c^*)$ . So,  $\log(y(t))$  has a linear trend.

If we believe that our time series  $(Y_i : 1 \leq i \leq n)$  has an exponential trend, then we should analyze instead  $(\tilde{Y}_i : 1 \leq i \leq n)$ , where  $\tilde{Y}_i = \log(Y_i)$ . Fit our linear regression model to the  $(\tilde{Y}_i, \tilde{t}_i)$ 's, thereby obtaining  $\hat{a}$  and  $\hat{b}$ . Of course, in the end, we need to exponentiate in order to transform back to our original scale. For example, the  $100(1 - \delta)\%$  prediction interval then takes the form

$$[\exp(\hat{a}t + \hat{b} - z\hat{\gamma}), \exp(\hat{a}t + \hat{b} + z\hat{\gamma})]$$

**Power Law Trends:** Suppose that we instead believe that the basic form of the trend is as a power of time, so that the trend takes the form

$$y(t) = c^* t^{a^*}.$$

Then,  $\log(y(t)) = \log(c^*) + a^* \log(t) = a^* \log(t) + b^*$ , where  $b^* = \log(c^*)$ . So,  $\log(y(t))$  is linear in  $\log(t)$ .

If we believe that our time series  $(Y_i : 1 \leq i \leq n)$  has such a trend, then we should fit a linear regression model to the  $(\tilde{Y}_i, \tilde{t}_i)$ 's, where  $\tilde{Y}_i = \log(Y_i)$  and  $\tilde{t}_i = \log(t_i)$ , thereby obtaining  $\hat{a}$  and  $\hat{b}$ . As in the exponential case, we need to transform back again to get back to our original scale. For example, the  $100(1 - \delta)\%$  prediction interval at time  $t$  takes the form

$$[\exp(\hat{a} \log(t) + \hat{b} - z\hat{\gamma}), \exp(\hat{a} \log(t) + \hat{b} + z\hat{\gamma})].$$

**Reference:** A good reference (with worked out numerical examples) on linear regression modeling is : “Probability and Statistics for the Engineering, Computing, and Physical Sciences” by Edward R. Dougherty, Prentice Hall, Englewood Cliffs, NT (1990).

### Time Series without Trends : The Autoregressive Process of Order 1

We now extend the linear regression model to the case in which we regress not on an exogenous variable (e.g. time) but on an endogenous variable (in the current setting, the prior values of the time series). Such models are called “autoregressive time models” (because we are regressing on prior values of the time series itself).

A first order autoregressive sequence  $(Y_n : n \geq 0)$  is one satisfying

$$Y_{n+1} = a^* Y_n + b^* + \epsilon_{n+1}$$

where  $(\epsilon_n : n \geq 1)$  is a sequence of iid  $N(0, \sigma_*^2)$  rv's.

**Question:** How do we estimate  $a^*$ ,  $b^*$ , and  $\sigma_*^2$  from the observed data?

As usual, we apply the method of maximum likelihood. Note that conditional on  $Y_0$ , the likelihood of the observed time series  $(Y_i : 1 \leq i \leq n)$  is

$$L(a, b, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - aY_{i-1} - b)^2}{2\sigma^2}\right)$$

so the log-likelihood is

$$\mathcal{L}(a, b, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n \frac{(Y_i - aY_{i-1} - b)^2}{2\sigma^2}.$$

At a maximizer  $(\hat{a}, \hat{b}, \hat{\sigma}^2)$ ,  $\hat{a}$  and  $\hat{b}$  must solve the pair of linear equations

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{a}Y_{i-1} - \hat{b})Y_{i-1} &= 0 \\ \sum_{i=1}^n (Y_i - \hat{a}Y_{i-1} - \hat{b}) &= 0.\end{aligned}$$

The solution of this pair of linear equations is

$$\begin{aligned}\hat{a} &= \frac{\sum_{i=1}^n Y_i Y_{i-1} - n(\sum_{j=1}^n Y_j/n)(\sum_{i=0}^{n-1} Y_i/n)}{\sum_{i=0}^{n-1} Y_i^2 - n(\sum_{j=1}^n Y_j/n)(\sum_{i=0}^{n-1} Y_i/n)} \\ \hat{b} &= \frac{1}{n} \sum_{i=1}^n Y_i - \hat{a} \frac{1}{n} \sum_{i=0}^{n-1} Y_i.\end{aligned}$$

The standard estimator for  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}Y_{i-1} - \hat{b})^2$$

### Predicting Future Values of the Autoregressive Process of Order 1

Note that

$$Y_{n+m} = \sum_{j=0}^{m-1} (a^*)^j (\epsilon_j + b^*) + (a^*)^m Y_n$$

and hence the “best prediction” of  $Y_{n+m}$  given  $Y_0 = y_0, \dots, Y_n = y_n$  is

$$\begin{aligned}E[Y_{n+m} | Y_0 = y_0, \dots, Y_n = y_n] &= \sum_{j=0}^{m-1} (a^*)^j b^* + (a^*)^m y_n \\ &= \begin{cases} (a^*)^m y_n + \frac{(1-(a^*)^m)}{1-a^*} b^* & , a^* \neq 1 \\ y_n + m b^* & , a^* = 1. \end{cases}\end{aligned}$$

This means of forecasting the future value at time  $n + m$  works when  $a^*$  and  $b^*$  are known. But, typically  $a^*$  and  $b^*$  must be estimated by  $\hat{a}$  and  $\hat{b}$ , suggesting the forecast at time  $n + m$  (based on observing  $Y_0 = y_0, \dots, Y_n = y_n$ )

$$\sum_{j=0}^{m-1} (\hat{a})^j \hat{b} + (\hat{a})^m y_n.$$

To construct a prediction interval for  $Y_{n+m}$  (based on observing  $Y_0, \dots, Y_n$ ) with confidence level  $100(1 - \delta)\%$ , we do the following :

1. We select a value  $z$  so that

$$P(-z \leq N(0, 1) \leq z) = 1 - \delta.$$

2. The  $100(1 - \delta)\%$  prediction interval is

$$\left[ \sum_{j=0}^{m-1} (\hat{a})^j \hat{b} + (\hat{a})^m y_n - z\hat{\sigma}, \sum_{j=0}^{m-1} (\hat{a})^j \hat{b} + (\hat{a})^m y_n + z\hat{\sigma} \right].$$

When  $n$  is large, this interval  $I$  has the property that

$$P(Y_{n+m} \in I | Y_0 = y_0, \dots, Y_n = y_n) \approx 1 - \delta$$

as required.

### Time Series without Trends: Autoregressive Processes of Order $p$

A more flexible family of autoregressive models is given by  $p$ 'th order autoregressions. A  $p$ 'th order autoregressive sequence ( $Y_n : n \geq 0$ ) is one satisfying

$$Y_{n+1} = \sum_{j=1}^p a_j^* Y_{n+1-j} + b^* + \epsilon_{n+1}$$

for  $n \geq p - 1$ , where  $(\epsilon_n : n \geq 1)$  is a sequence of iid  $N(0, \sigma_*^2)$  rv's.

**Question:** How do we estimate  $a_1^*, \dots, a_p^*, b^*$  and  $\sigma_*^2$  from the observed data?

Conditional on  $Y_0, \dots, Y_{p-1}$ , the likelihood of the observed time series ( $Y_i : p \leq i \leq n$ ) is

$$\begin{aligned} L(a_1, \dots, a_p, b, \sigma^2) \\ = \prod_{i=p}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(Y_i - a_1 Y_{i-1} - \dots - a_p Y_{i-p} - b)^2}{2\sigma^2} \right) \end{aligned}$$

so the log-likelihood is

$$\begin{aligned} \mathcal{L}(a_1, \dots, a_p, b, \sigma^2) \\ = -\frac{(n-p+1)}{2} \log(2\pi\sigma^2) - \sum_{i=p}^n \frac{(Y_i - a_1 Y_{i-1} - \dots - a_p Y_{i-p} - b)^2}{2\sigma^2}. \end{aligned}$$

At a maximizer  $(\hat{a}_1, \dots, \hat{a}_p, \hat{b}, \hat{\sigma}^2)$ ,  $\hat{a}_1, \dots, \hat{a}_p, \hat{b}$  must solve the linear system of  $(p+1)$  equations in  $(p+1)$  unknowns, namely

$$\begin{aligned} \sum_{i=p}^n (Y_i - \hat{a}_1 Y_{i-1} - \dots - \hat{a}_p Y_{i-p} - \hat{b}) Y_{i-j} &= 0, \quad 1 \leq j \leq p, \\ \sum_{i=p}^n (Y_i - \hat{a}_1 Y_{i-1} - \dots - \hat{a}_p Y_{i-p} - \hat{b}) &= 0. \end{aligned}$$

Once  $\hat{a}_1, \dots, \hat{a}_p, \hat{b}$  have been computed from the above linear system of equations,  $\sigma_*^2$  can be estimated by

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

### Predicting Future Values of the Autoregressive Process of Order $p$

One means of obtaining such predictions is to use the power of matrix theory. We start by writing a  $p$ 'th order autoregression using matrix/vector notation. Put

$$\vec{Y}_n = (Y_n, Y_{n-1}, \dots, Y_{n-p+1})^T.$$

Note that

$$\begin{pmatrix} Y_{n+1} \\ Y_n \\ \vdots \\ Y_{n-p+2} \end{pmatrix} = \begin{pmatrix} a_1^* & a_2^* & \cdots & \cdots & a_p^* \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} Y_n \\ Y_{n-1} \\ \vdots \\ Y_{n-p+1} \end{pmatrix} + \begin{pmatrix} b^* \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \epsilon_{n+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

so we can write this as

$$\vec{Y}_{n+1} = F^* \vec{Y}_n + \vec{b}^* + \vec{\epsilon}_{n+1}$$

where

$$\begin{aligned} \vec{b}^* &= (b^*, 0, \dots, 0)^T, \\ \vec{\epsilon}_n &= (\epsilon_n, 0, \dots, 0)^T, \end{aligned}$$

and

$$F^* = \begin{pmatrix} a_1^* & a_2^* & \cdots & \cdots & a_p^* \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

As for a first order (scalar) autoregression, we can write  $\vec{Y}_{n+m}$  as

$$\vec{Y}_{n+m} = \sum_{j=0}^{m-1} (F^*)^j (\vec{b}^* + \vec{\epsilon}_{n+m-j}) + (F^*)^m \vec{Y}_n.$$

So,

$$E[\vec{Y}_{n+m} | \vec{Y}_0 = \vec{y}_0, \dots, \vec{Y}_n = \vec{y}_n] = \sum_{j=0}^{m-1} (F^*)^j \vec{b}^* + (F^*)^m \vec{y}_n.$$

But  $Y_n = e^T \vec{Y}_n$ , where

$$e = (1, 0, \dots, 0)^T$$

So, our best prediction of  $Y_{n+m}$ , based on observing  $Y_0 = y_0, \dots, Y_n = y_n$ , is

$$\begin{aligned} \hat{Y}_{n+m} &= E[Y_{n+m} | Y_0 = y_0, \dots, Y_n = y_n] \\ &= e^T \sum_{j=0}^{m-1} (F^*)^j \vec{b}^* + e^T (F^*)^m \vec{y}_n. \end{aligned}$$

where  $\vec{y}_n = (y_n, y_{n-1}, \dots, y_{n-p+1})^T$ .

**Example 3.2.1:** If  $(Y_n : n \geq 0)$  is a second order autoregression satisfying

$$Y_{n+1} = (1/2)Y_n + (1/8)Y_{n-1} + 3 + \epsilon_{n+1}$$

where the  $\epsilon_n$ 's are iid  $N(0, 9)$  rv's with  $Y_0 = 1, Y_1 = 1/5, Y_2 = 7, Y_3 = 4$ , then our best forecast of  $Y_7$  is

$$\widehat{Y}_7 = (1, 0) \sum_{j=0}^3 \begin{pmatrix} 1/2 & 1/8 \\ 1 & 0 \end{pmatrix}^j \begin{pmatrix} 3 \\ 0 \end{pmatrix} + (1, 0) \begin{pmatrix} 1/2 & 1/8 \\ 1 & 0 \end{pmatrix}^4 \begin{pmatrix} 4 \\ 7 \end{pmatrix}.$$

As in the case of a first order (scalar) autoregressive process, we typically must estimate  $a_1^*, \dots, a_p^*$ , and  $b^*$  from the observed data. Let  $\widehat{a}_1, \dots, \widehat{a}_p$ , and  $\widehat{b}$  be the estimators discussed earlier (obtained by solving a system of linear equations). Put

$$\widehat{F} = \begin{pmatrix} \widehat{a}_1 & \widehat{a}_2 & \cdots & \widehat{a}_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$\widehat{\vec{b}} = (\widehat{b}, 0, \dots, 0)^T.$$

The natural thing to do here is to substitute  $\widehat{F}$  and  $\widehat{\vec{b}}$  for  $F$  and  $\vec{b}$  in our forecasting formula, namely to forecast  $Y_{n+m}$  via

$$e^T \sum_{j=0}^{m-1} (\widehat{F})^j \widehat{\vec{b}} + e^T (\widehat{F})^m \vec{y}_n.$$

We can also obtain a prediction interval for  $Y_{n+m}$  (based on observing  $Y_0 = y_0, \dots, Y_n = y_n$ ) with confidence level  $100(1 - \delta)\%$ , we do the following:

1. We select a value  $z$  so that

$$P(-z \leq N(0, 1) \leq z) = 1 - \delta.$$

2. The  $100(1 - \delta)\%$  prediction interval is

$$\left[ e^T \sum_{j=0}^{m-1} (\widehat{F})^j \widehat{\vec{b}} + e^T (\widehat{F})^m \vec{y}_n - z\widehat{\sigma}, e^T \sum_{j=0}^{m-1} (\widehat{F})^j \widehat{\vec{b}} + e^T (\widehat{F})^m \vec{y}_n + z\widehat{\sigma} \right].$$

When  $n$  is large, this interval  $I$  has the property that

$$P(Y_{n+m} \in I | Y_0 = y_0, \dots, Y_n = y_n) \approx 1 - \delta$$

as required.

### Testing Forecasting Methods

Suppose that we wish to forecast the value of a time series at time 105, based on observing the time series at time epochs  $1, 2, \dots, 100$ . The most common approach to doing this is to split the

observed time series into two pieces: the “in-sample” piece (e.g. the observed time series at times  $1, 2, \dots, 50$ ) and the “out-of-sample” piece (e.g. the observed time series at times  $51, 52, \dots, 100$ ).

We usually have several different forecasting methods that we believe may work well. (e.g. autoregressive models of several different orders, exponential smoothing with several different smoothing parameters, etc.) We will use the in-sample data to fit any parameters required by the forecasting method (e.g. the autoregressive parameters  $a_1^*, \dots, a_p^*, b^*$ ) and the out-of-sample data to compare the quality of the different competing forecasting methods.

For example, suppose we wish to compare the forecasting power of :

- an autoregressive model of order 1
- an autoregressive model of order 2
- exponential smoothing with  $\alpha = 0.7$ .

We use the in-sample data (i.e. the time series at times  $1, 2, \dots, 50$ ) to estimate the parameters for the two autoregressive models (using the methods discussed earlier). Once we have fit the parameters from the observed in-sample data, compare the quality of the “5-step” predictions (recall that we are trying to predict the time series at time 105, based on observing the time series to time 100) based on the out-of-sample data. For example, if we want to use the forecasting method that minimizes mean square prediction error, compute

$$\frac{1}{50} \sum_{i=51}^{100} (\hat{Y}_i - y_i)^2 \tag{1}$$

for each of the forecasting methods you are considering, where  $\hat{Y}_i$  is the prediction of  $Y_i$  based on observing  $Y_0 = y_0, \dots, Y_{i-5} = y_{i-5}$  (and using the parameters estimated from the in-sample data). Then, use the forecasting method that minimizes (1) to forecast the value at time 105.

The reason that the sample is split into an “in-sample” and “out-of-sample” piece is to avoid the following problem. Suppose that we fit the parameters for our two autoregressive models from the full data set of 100 observations and then use the same 100 observations to test the quality of the predictions. Note that the order 1 model is “nested” inside the order 2 model (because an order 1 model is just an order 2 model with  $a_2^*=0$ ). Under the normal distribution assumption, the principle of maximum likelihood amounts to finding a value of the autoregressive parameters that minimizes the sum of the mean square prediction errors for the data set. Because the models are nested, it is certain that one can make the errors smaller with the model having more parameters. So, if we fit the models and compare the models using exactly the same data, we will always choose the model having the most parameters. But we intuitively know that if the data set is of limited size, estimating all the parameters accurately may be impossible, so using the model with the most parameters may result in “overfitting” the data. (Think of fitting an autoregressive model of order 97 to 100 data points!). By testing the quality of the predictions on a “fresh” data set, this allows the model comparison to be done more fairly and allows us to take into account possible overfit issues. In particular, it is often the case in practice that a lower order model predicts better on the out-of-sample data than does the high order model.

### 3 Model Selection

Suppose that you observe a time series  $\{Y_t\}$ ,  $t = 1, 2, \dots, n$ , and you want to develop a method for predicting the next value  $Y_{n+1}$ . Very often plotting the values  $\{Y_t\}$  against time  $t$  provides ideas on whether the time series has a deterministic trend (linear, for example), or whether there is no trend in the observations, and other models (autoregression, for example) should be used. However, in some cases visual data analysis can not provide insights on what kind of model to choose. In this case, special model selection procedures are used to decide on what kind of model to rely on. Sometimes the model selection procedure first fits the data to a more general model that includes particular suspicious models as special cases, and then decides on a particular model to use for forecasting by analyzing the input of each of the underlying sub-models to the data fit.

Suppose that you are not sure if you should rely on a linear trend to describe your time series:

$$Y_t = a \cdot t + b + \epsilon_t, \quad t = 1, 2, \dots, n \quad (3.3.1)$$

where  $\{\epsilon_t\}$  are independent identically distributed random errors with normal probability distribution with zero mean and some finite variance, or if you should use an autoregressive model of order one:

$$Y_t = a \cdot Y_{t-1} + b + \epsilon_t, \quad t = 2, 3, \dots, n \quad (3.3.2)$$

where  $\{\epsilon_t\}$  are again random errors with the same assumptions as before. To resolve the issue, you can assume the following model for your time series:

$$Y_t = a_1 \cdot t + a_2 \cdot Y_{t-1} + b + \epsilon_t, \quad t = 2, 3, \dots, n \quad (3.3.3)$$

which includes both models (3.3.1) and (3.3.2) as special cases. Indeed, if we set  $a_2 = 0$  in (3.3.3) then we get the linear regression model (3.3.1), and if we set  $a_1 = 0$  in (3.3.3) then we get the autoregression model (3.3.2).

Model (3.3.3) can be viewed as a multiple linear regression model, and you can use the Matlab *regress* function to estimate coefficients  $\{a_1, a_2, b\}$  and to obtain the confidence intervals for them (please refer to the Linear Regression with Matlab section for the description of how to do it). After you have obtained the estimates and confidence intervals, you should analyze the confidence intervals to see whether  $a_1$  and  $a_2$  are significantly different from zero from the statistical standpoint. The rule is as follows: if a confidence interval for a coefficient includes zero, then this coefficient is not statistically significant. So:

- if your confidence interval for  $a_1$  includes zero, then the input of the linear trend to your model is not statistically significant, and you should rely on the model (3.3.2);
- if your confidence interval for  $a_2$  includes zero, then there is no statistically significant linear relationship on previous observations, and you should rely on the model (3.3.1);

Note that you can also end up with both coefficients being significant, and then you should use the model (3.3.3) for forecasting, or both of them being insignificant, and then your model reduces to simply  $Y_t = b + \epsilon_t$ .

As a result, your workflow for model selection looks as follows:

1. Perform estimation for model (3) and obtain estimates and confidence intervals for coefficients.
2. Analyze statistical significance of coefficients and exclude insignificant terms.
3. Perform parameter estimation for the resulting model and use it for forecasting.

## 4 Numerical Example

Below are yields of 8 consecutive batches of a chemical process.<sup>1</sup>

89.7 81.4 84.5 84.8 87.3 79.7 85.1 81.7

We wish to fit an appropriate model to the yields and predict the yield for time period 9 (2 time periods into the future, given that the eight data points represent the observations collected in time periods 0,1,2,...,7). Suppose we wish to compare two different models:

- an autoregressive model of order 1
- exponential smoothing with  $\alpha = 0.7$ .

First, we use data in the first 4 time periods as in-sample data.

In the autoregressive model of order 1,  $Y_n$ , the batch of the  $n$ th time period, can be modeled as

$$Y_{n+1} = a^*Y_n + b^* + \epsilon_{n+1}$$

where  $(\epsilon_n : n \geq 1)$  is a sequence of iid  $N(0, \sigma_*^2)$  rv's. Then the estimators  $\hat{a}$  and  $\hat{b}$  for  $a^*$  and  $b^*$  are

$$\hat{a} = \frac{\sum_{i=1}^n Y_i Y_{i-1} - n(\sum_{j=1}^n Y_j/n)(\sum_{i=0}^{n-1} Y_i/n)}{\sum_{i=0}^{n-1} Y_i^2 - n(\sum_{j=1}^n Y_j/n)(\sum_{i=0}^{n-1} Y_i/n)}$$

$$\hat{b} = \frac{1}{n} \sum_{i=1}^n Y_i - \hat{a} \frac{1}{n} \sum_{i=0}^{n-1} Y_i.$$

In this example,  $n = 3$ ,  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ ,  $Y_3 = 84.8$ , so we get

$$\hat{a} = \frac{81.4 \cdot 89.7 + 84.5 \cdot 81.4 + 84.8 \cdot 84.5 - 3\left(\frac{81.4+84.5+84.8}{3}\right)\left(\frac{89.7+81.4+84.5}{3}\right)}{(89.7^2 + 81.4^2 + 84.5^2) - 3\left(\frac{81.4+84.5+84.8}{3}\right)\left(\frac{89.7+81.4+84.5}{3}\right)} = -0.0018$$

$$\hat{b} = \frac{1}{3}(81.4 + 84.5 + 84.8) - (-0.0018)\frac{1}{3}(89.7 + 81.4 + 84.5) = 83.72.$$

Hence, the best prediction of  $Y_{n+2}$  can be predicted from  $Y_0 = y_0, \dots, Y_n = y_n$  by

$$\hat{b} + \hat{a}\hat{b} + \hat{a}^2 y_n.$$

For our 8 period observed time series, we decided above to reserve the first 4 periods as our “in sample” data set and the last 4 time periods as the “out of sample” data set. So, we must use our fitted model to predict the values of the time series in each of the 4 “out of sample” time

<sup>1</sup>Thanks to Hyndman, R.J. (n.d.) *Time Series Data Library*,

<http://www-personal.buseco.monash.edu.au/hyndman/TSDL/>

for supplying the above data.

points. In other words, recalling that our ultimate goal is to forecast 2 time periods in the future, we predict  $Y_4$  based on  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ , and  $Y_2 = 84.5$ . The value  $Y_5$  is predicted on the basis of  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ , and  $Y_3 = 84.8$ . The value  $Y_6$  is predicted on the basis of  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ ,  $Y_3 = 84.8$ , and  $Y_4 = 87.3$ . Finally, the value  $Y_7$  is predicted on the basis of  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ ,  $Y_3 = 84.8$ ,  $Y_4 = 87.3$ , and  $Y_5 = 79.7$ . This yields the following forecasts:

$$\begin{aligned}\widehat{Y}_4 &= \widehat{b} + \widehat{a}\widehat{b} + \widehat{a}^2(84.5) \\ &= 83.5696 \\ \widehat{Y}_5 &= \widehat{b} + \widehat{a}\widehat{b} + \widehat{a}^2(84.8) \\ &= 83.5696 \\ \widehat{Y}_6 &= \widehat{b} + \widehat{a}\widehat{b} + \widehat{a}^2(87.3) \\ &= 83.5696 \\ \widehat{Y}_7 &= \widehat{b} + \widehat{a}\widehat{b} + \widehat{a}^2(79.5) \\ &= 83.5696\end{aligned}$$

We can now compare the four forecasts above to the actual observed “out of sample” time series values at times 4, 5, 6, and 7, namely to the values 87.3, 79.7, 85.1, and 81.7. If we use mean square prediction error as our selection criterion, we then get

$$\begin{aligned}\frac{1}{4} \sum_{i=4}^7 (\widehat{Y}_i - y_i)^2 &= \frac{1}{4} \{(83.5696 - 87.3)^2 + (83.5696 - 79.7)^2 + (83.5696 - 85.1)^2 + (83.5696 - 81.7)^2\} \\ &= 8.6818\end{aligned}$$

as the mean square prediction error associated with our first order autoregressive model.

We now turn to our exponential smoothing forecast with  $\alpha = 0.7$ . Here we have no parameters to estimate from the “in sample” data. One complication here is that we need to adapt our exponential smoothing method to predict 2 units ahead in time (rather than the more standard one time step forecast). There is a brief discussion of this in Section 20.6 of the Course Reader. It suggests that the “two step ahead exponential smoothing forecast” should be equal to the “one step ahead exponential smoothing forecast”. The reason for this has to do with the statistical model that underlies exponential smoothing (See the handout on Forecasting). Note that  $Y_{n+2} - Y_{n+1} = Z_{n+2} - (1 - \alpha)Z_{n+1}$ . If we compute the conditional expectation of each side of the equality given  $Y_0 = y_0, \dots, Y_n = y_n$  (and use the fact that both  $Z_{n+1}$  and  $Z_{n+2}$  are independent of  $Y_0, \dots, Y_n$ ), we see that  $E(Y_{n+2}|Y_0 = y_0, \dots, Y_n = y_n) = E(Y_{n+1}|Y_0 = y_0, \dots, Y_n = y_n)$ . In other words, the “two step ahead exponential smoothing forecast” should be equal to the “one step ahead exponential smoothing forecast”. So, all we need to do here is compute our standard “one step ahead exponential smoothing forecast” and use that in place of the “two step ahead exponential smoothing forecast”.

As in the first order autoregressive modeling context, we predict  $Y_4$  based on  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ , and  $Y_2 = 84.5$ . Our forecast for  $Y_4$  should equal our forecast for  $Y_3$ , based on the observations through period 2. The value  $Y_5$  is predicted on the basis of  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ , and  $Y_3 = 84.8$ . (Our forecast for  $Y_5$  should equal our forecast for  $Y_4$ ). The value  $Y_6$  is predicted on the basis of  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ ,  $Y_3 = 84.8$ , and  $Y_4 = 87.3$ . (Again, our forecast for  $Y_6$  should equal our forecast for  $Y_5$  based on the observations through period 4.) Finally, the value  $Y_7$  is predicted on the basis of  $Y_0 = 89.7$ ,  $Y_1 = 81.4$ ,  $Y_2 = 84.5$ ,  $Y_3 = 84.8$ ,  $Y_4 = 87.3$ , and  $Y_5 = 79.7$ . (Our forecast for  $Y_7$  should equal our forecast for  $Y_6$ , based on our observations of the time series through period 5).

This still leaves the question of how to initialize our exponential smoothing forecast. To initialize our exponential smoothing forecast at time zero, let’s just start it at zero. So,  $\widehat{Y}_0 = 0$ . We can now

recursively compute  $\widehat{Y}_1, \widehat{Y}_2, \dots, \widehat{Y}_6$  using our observed time series values.

$$\begin{aligned}\widehat{Y}_1 &= (0.7)y_0 + (0.3)\widehat{Y}_0 \\ &= (0.7)(89.7) \\ &= 62.7900 \\ \widehat{Y}_2 &= (0.7)y_1 + (0.3)\widehat{Y}_1 \\ &= (0.7)(81.4) + (0.3)(62.7900) \\ &= 75.8170 \\ \widehat{Y}_3 &= (0.7)y_2 + (0.3)\widehat{Y}_2 \\ &= (0.7)(84.5) + (0.3)(75.8170) \\ &= 81.8951 \\ \widehat{Y}_4 &= (0.7)y_3 + (0.3)\widehat{Y}_3 \\ &= (0.7)(84.8) + (0.3)(81.8951) \\ &= 83.9285 \\ \widehat{Y}_5 &= (0.7)y_4 + (0.3)\widehat{Y}_4 \\ &= (0.7)(87.3) + (0.3)(83.9285) \\ &= 86.2886 \\ \widehat{Y}_6 &= (0.7)y_5 + (0.3)\widehat{Y}_5 \\ &= (0.7)(79.7) + (0.3)(86.2886) \\ &= 81.6766\end{aligned}$$

Our discussion above justifies using  $\widehat{Y}_3 = 81.8951$  as our forecast for period 4 (based on observing the time series up to time 2),  $\widehat{Y}_4 = 83.9285$  as our forecast for period 5 (based on observing the time series up to time 3),  $\widehat{Y}_5 = 86.2886$  as our forecast for period 6 (based on observing the time series up to time 4), and  $\widehat{Y}_6 = 81.6766$  as our forecast for period 7 (based on observing the time series up to time 5).

We can now again compare the four forecasts above to the actual observed “out of sample” time series values at times 4, 5, 6, and 7, namely to the values 87.3, 79.7, 85.1, and 81.7. Using mean square prediction error as our selection criterion, we then get

$$\begin{aligned}\frac{1}{4} \sum_{i=4}^7 (\widehat{Y}_i - y_i)^2 &= \frac{1}{4} \{(81.8951 - 87.3)^2 + (83.9285 - 79.7)^2 + (86.2886 - 85.1)^2 + (81.6766 - 81.7)^2\} \\ &= 12.1266.\end{aligned}$$

as the mean square prediction error associated with our exponential smoothing method with  $\alpha = 0.7$ . On the basis of comparing our two mean square prediction errors on the “out of sample” data, we see that the autoregressive model yields lower “two step ahead” forecast errors.

We therefore should use the autoregressive method to make our forecast in period 9, based on the observed data up to period 7. The forecast for period 9 is  $\widehat{b} + \widehat{a}\widehat{b} + \widehat{a}^2 y_7 = \widehat{b} + \widehat{a}\widehat{b} + \widehat{a}^2(81.7) = 83.5696$ .

## 5 Linear Regression with Matlab

Matlab function *regress* addresses the multiple linear regression problem based on the least squares approach and provides the automatic means for estimating the model parameters and performing the regression model diagnostics. Here is the description of the *regress* function in Matlab help:

REGRESS Multiple linear regression using least squares.

`b = REGRESS(y,X)` returns the vector of regression coefficients, `b`, in the linear model  $y = Xb$ , (`X` is an `n`x`p` matrix, `y` is the `n`x1 vector of observations).

`[B,BINT,R,RINT,STATS] = REGRESS(y,X,alpha)` uses the input, `ALPHA` to calculate  $100(1 - \text{ALPHA})$  confidence intervals for `B` and the residual vector, `R`, in `BINT` and `RINT` respectively. The vector `STATS` contains the R-square statistic along with the `F` and `p` values for the regression.

The `X` matrix should include a column of ones so that the model contains a constant term. The `F` and `p` values are computed under the assumption that the model contains a constant term, and they are not correct for models without a constant. The R-square value is the ratio of the regression sum of squares to the total sum of squares.

The use of this function is illustrated below using a simple example. Suppose that we are given the following independent observations for variable `Y` at time `t`:

<i>t</i>	1	2	3	4	5	6	7	8	9	10
<i>Y</i>	2	0	5.5	7.9	12	12.5	14	15	19	19.9

We want to check if there is a linear relationship between `Y` and `t`, and if it exists, to predict `Y` when `t` is 11 and 12. Plotting `Y` against `t` suggests that our assumption about the linear relationship is probably true (Fig. 1):

```
>> t = [ 1 2 3 4 5 6 7 8 9 10 ]
```

```
t =
```

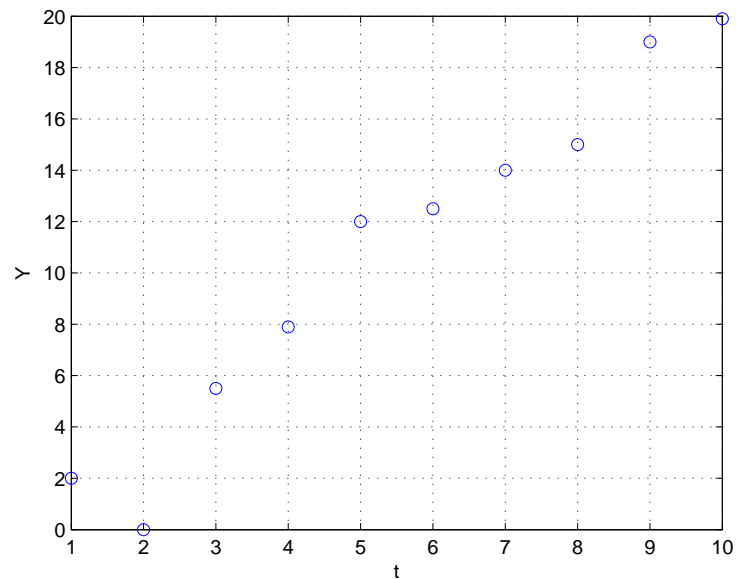
```
    1    2    3    4    5    6    7    8    9   10
```

```
>> Y = [ 2 0 5.5 7.9 12 12.5 14 15 19 19.9 ]
```

```
Y =
```

```
    2.0000    0    5.5000    7.9000   12.0000   12.5000   14.0000   15.0000   19.0000   19.9000
```

```
>> plot( t, Y, 'o' ); grid; xlabel( 't' ); ylabel( 'Y' );
```

Fig. 1. Plot of variable  $Y$  vs. time  $t$ 

Now we want to use the *regress* function to estimate the model parameters and perform regression diagnostics. The *regress* function requires that the matrix of independent variables includes a column of ones so that the model contains a constant term, so we create a matrix  $x$  as:

```
>> x = [ ones(10,1) t' ]
```

```
x =
```

```
1    1
1    2
1    3
1    4
1    5
1    6
1    7
1    8
1    9
1   10
```

Now let us apply the *regress* function to our data (note that we transposed  $t$  and  $Y$  according to the function requirements):

```
>> [b,bint,r,rint,stats] = REGRESS( Y', x )
```

```
b =
```

```
-1.2333
```

```
2.1842

bint =

-3.6760    1.2094
 1.7906    2.5779

r =

 1.0491
-3.1352
 0.1806
 0.3964
 2.3121
 0.6279
-0.0564
-1.2406
 0.5752
-0.7091

rint =

-1.9053    4.0035
-5.0099   -1.2604
-3.2863    3.6475
-3.1580    3.9507
-0.6955    5.3198
-2.9509    4.2066
-3.6271    3.5144
-4.5383    2.0571
-2.7006    3.8509
-3.7394    2.3212

stats =

 0.9534   163.6974   0.0000
```

The output vector  $b$  is the vector of regression coefficients such that  $Y_t = b_1 + t \cdot b_2 + \epsilon_t$ , where  $\{\epsilon_t\}$  are independent identically normally distributed random errors with zero mean and finite variance. The output variable  $bint$  provides the confidence intervals (CI) for the corresponding coefficients (default is 95% confidence level). These confidence intervals can be used to assess the significance of the regression coefficients: if an interval for a coefficient contains zero, then this coefficient is not significantly different from zero from the statistical standpoint, and thus can be set to zero in the resulting regression model. In our case we get:

- 95% CI for  $b_1$  is  $[-3.6760; 1.2094]$ , it contains zero so the constant term is not statistically significant and can be excluded from further consideration.
- 95% CI for  $b_2$  is  $[1.7906; 2.5779]$ , it does not contain zero and we conclude that there is a

statistically significant evidence of linear relationship between  $Y$  and  $t$ , with the least squares approach leading to the estimates of slope  $b_2 = 2.1842$ .

Since the constant term  $b_1$  is not statistically significant, you now want to estimate the regression slope  $b_2$  assuming that  $b_1 = 0$ . Although the `regress` function requires the model to have constant term to correctly calculate the diagnostic values in variable `stat`, you can still obtain the rest of the output correctly even when there is no constant term in the model. In particular, you can get an estimate of  $b_2$  assuming that  $b_1 = 0$  as follows:

```
>> b = regress( Y', t' )
```

```
b =  
    2.0081
```

Note that you estimate of the regression slope is different from the one obtained before. As a result, our model that can be used for prediction of  $Y$  given  $t$  looks as follows:

$$Y = 2.0081 \cdot t,$$

and plugging in  $t = 11$ ,  $t = 12$  leads to  $Y = 22.0886$  and  $Y = 24.0966$  respectively. Fig. 2 provides a graphical illustration of the obtained model fit.

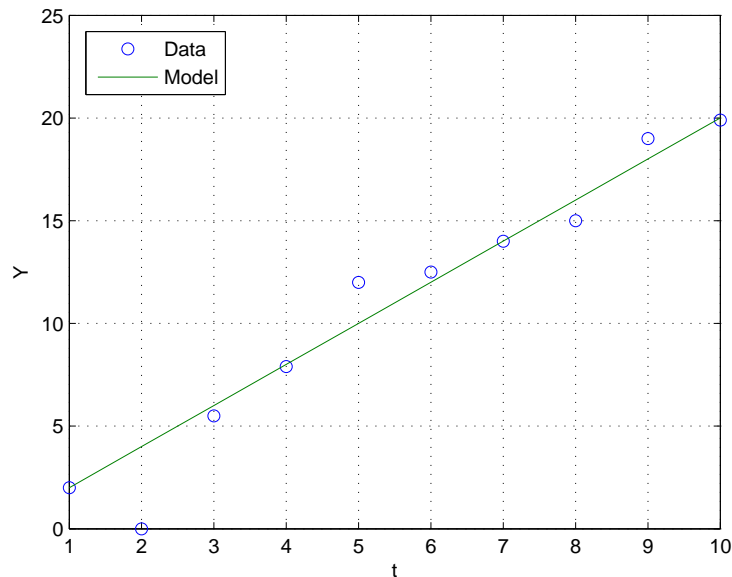


Fig. 2. Original data and estimated linear regression model

Note that the `regress` function also returns variables `r`, `rint` and `stats`. The output vector `r` contains the residuals for the estimated model, which in our case are calculated as  $r_t = Y_t C[1; t] \cdot b$ . The values of the residuals show the accuracy of the approximation achieved by our model: the smaller the residual is, the better we fit a particular observation. The variable `rint` contains confidence intervals for the corresponding residuals; if a confidence interval for a particular residual does not contain zero, then the corresponding observation is suspicious for being an outlier. Note that the vector `r` can be used to calculate the achieved mean square error:

```
>> mse = mean( r.^2 )
```

```
mse =
    1.9236
```

The vector *stats* contains more advanced statistical diagnostics for our model. *stats*<sub>1</sub> is the value of the coefficient of determination, the closer this value is to one, the better our regression model explains the data. *stats*<sub>3</sub> is the *p*-value for the F-test, and is used to assess the overall model significance in case of multiple independent variables. Usually, if this value is less than 0.05 then the chances of all regression coefficients being equal to zero (or equivalently being insignificant) are very low.

Suppose that now you want to check if introducing an autoregressive component to your model improves your understanding of the data. In other words, you want to check if the following model that includes an autoregressive component of order one is better than the simple regression model that you just created:

$$Y_t = b_1 + t \cdot b_2 + Y_{t-1} \cdot b_3 + \epsilon_t, \quad t = 2, 3, \dots, 10, \quad (3.5.1)$$

where  $\{\epsilon_t\}$  are independent identically distributed random errors with normal probability distribution with zero mean and some finite variance. To do that, you first create a new matrix of independent variables to be used in *regress* function:

```
>> x_ar = [ ones(9,1) t(2:10)' Y(1:9)' ]
```

```
x_ar =
    1.0000    2.0000    2.0000
    1.0000    3.0000         0
    1.0000    4.0000    5.5000
    1.0000    5.0000    7.9000
    1.0000    6.0000   12.0000
    1.0000    7.0000   12.5000
    1.0000    8.0000   14.0000
    1.0000    9.0000   15.0000
    1.0000   10.0000   19.0000
```

Then the *regress* function returns the following:

```
>> [b,bint] = REGRESS( Y(2:10)', x_ar )
```

```
b =
   -2.4204
    2.6033
   -0.1478
```

```
bint =
   -7.4015    2.5608
```

0.3786	4.8279
-1.1105	0.8149

Analyzing the confidence intervals in *bint* you conclude that only the coefficient  $b_2$  in the model (3.5.1) is statistically significant, and since  $b_3$  is not significantly different from zero, the autoregressive component that you tried to introduce to the model does not help to explain the nature of the observed data. In other words, observation  $Y_t$  does not depend on the previous observation  $Y_{t-1}$ .