Gaussian processes

The class of Gaussian processes is one of the most widely used families of stochastic processes for modeling dependent data observed over time, or space, or time and space. The popularity of such processes stems primarily from two essential properties. First, a Gaussian process is completely determined by its mean and covariance functions. This property facilitates model fitting as only the first- and second-order moments of the process require specification. Second, solving the prediction problem is relatively straightforward. The best predictor of a Gaussian process at an unobserved location is a linear function of the *observed* values and, in many cases, these functions can be computed rather quickly using recursive formulas.

The fundamental characterization, as described below, of a Gaussian process is that all the finite-dimensional distributions have a multivariate normal (or Gaussian) distribution. In particular the distribution of each observation must be normally distributed. There are many applications, however, where this assumption is not appropriate. For example, consider observations $x_1, \ldots, x_n$, where x_t denotes a 1 or 0, depending on whether or not the air pollution on the tth day at a certain site exceeds a government standard. A model for there data should only allow the values of 0 and 1 for each daily observation thereby precluding the normality assumption imposed by a Gaussian model. Nevertheless, Gaussian processes can still be used as building blocks to construct more complex models that are appropriate for non-Gaussian data. See [3–5] for more on modeling non-Gaussian data.

Basic Properties

A real-valued stochastic process $\{X_t, t \in T\}$, where T is an index set, is a Gaussian process if all the finite-dimensional distributions have a multivariate normal distribution. That is, for any choice of distinct values $t_1, \ldots, t_k \in T$, the random vector $\mathbf{X} = (X_{t_1}, \ldots, X_{t_k})'$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu} = \mathrm{E}\mathbf{X}$ and covariance matrix $\Sigma = \mathrm{cov}(\mathbf{X}, \mathbf{X})$, which will be denoted by

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$$

Provided the covariance matrix Σ is nonsingular, the random vector $\mathbf{X}$ has a Gaussian probability density function given by

$$f_{\mathbf{X}}(x) = (2\pi)^{-n/2}(\det\ \Sigma)^{-1/2} \times \exp(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})) \quad (1)$$

In environmental applications, the subscript t will typically denote a point in time, or space, or space and time. For simplicity, we shall restrict attention to the case of time series for which t represents time. In such cases, the index set T is usually $[0, \infty)$ for time series recorded continuously or $\{0, 1, \ldots, \}$ for time series recorded at equally spaced time units.

The *mean* and *covariance* functions of a Gaussian process are defined by

$$\mu(t) = \mathrm{E}X_t \quad (2)$$

and

$$\gamma(s, t) = \mathrm{cov}(X_s, X_t) \quad (3)$$

respectively. While Gaussian processes depend only on these two quantities, modeling can be difficult without introducing further simplifications on the form of the mean and covariance functions. The assumption of stationarity frequently provides the proper level of simplification without sacrificing much generalization. Moreover, after applying elementary transformations to the data, the assumption of stationarity of the transformed data is often quite plausible.

A Gaussian time series $\{X_t\}$ is said to be stationary if

1. $m(t) = \mathrm{E}X_t = \mu$ is independent of t, and
2. $\gamma(t + h, t) = \mathrm{cov}(X_{t+h}, X_t)$ is independent of t for all h.

For stationary processes, it is conventional to express the covariance function γ as a function on T instead of on $T \times T$. That is, we define $\gamma(h) = \mathrm{cov}(X_{t+h}, X_t)$ and call it the autocovariance function of the process. For stationary Gaussian processes $\{X_t\}$, we have

3. $X_t \sim N(\mu, \gamma(0))$ for all t, and
4. $(X_{t+h}, X_t)'$ has a bivariate normal distribution with covariance matrix

$$\begin{bmatrix} \gamma(0) & \gamma(h) \\ \gamma(h) & \gamma(0) \end{bmatrix}$$

for all t and h.

A general stochastic process $\{X_t\}$ satisfying conditions 1 and 2 is said to be weakly or second-order stationary. The first- and second-order moments of weakly stationary processes are invariant with respect to time translations. A stochastic process $\{X_t\}$ is strictly stationary if the distribution of $(X_{t_1}, \ldots, X_{t_n})$ is the same as $(X_{t_{1+s}}, \ldots, X_{t_{n+s}})$ for any s. In other words, the distributional properties of the time series are the same under any time translation. For Gaussian time series, the concepts of weak and strict stationarity coalesce. This result follows immediately from the fact that for weakly stationary processes, $(X_{t_1}, \ldots, X_{t_n})$ and $(X_{t_{1+s}}, \ldots, X_{t_{n+s}})$ have the same mean vector and covariance matrix. Since each of the two vectors has a multivariate normal distribution, they must be identically distributed.

Properties of the Autocovariance Function

An autocovariance function $\gamma(\cdot)$ has the properties:

1. $\gamma(0) \geq 0$,
2. $|\gamma(h)| \leq \gamma(0)$ for all h,
3. $\gamma(h) = \gamma(-h)$, i.e. $\gamma(\cdot)$ is an even function.

Autocovariances have another fundamental property, namely that of non-negative definiteness,

$$\sum_{i,j=1}^{n} a_i \gamma(t_i - t_j) a_j \geq 0 \tag{4}$$

for all positive integers n, real numbers $a_1, \ldots, a_n$, and $t_1, \ldots, t_n \in T$. Note that the expression on the left of (4) is merely the variance of $a_1 X_{t_1} + \cdots + a_n X_{t_n}$ and hence must be non-negative. Conversely, if a function $\gamma(\cdot)$ is non-negative definite and even, then it must be an autocovariance function of some stationary Gaussian process.

Gaussian Linear Processes

If $\{X_t, t = 0, \pm 1, \pm 2, \ldots, \}$ is a stationary Gaussian process with mean 0, then the Wold decomposition allows X_t to be expressed as a sum of two independent components,

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t \tag{5}$$

where $\{Z_t\}$ is a sequence of independent and identically distributed (iid) normal random variables with mean 0 and variance σ^2, $\{\psi_j\}$ is a sequence of square summable coefficients with $\psi_0 = 1$, and $\{V_t\}$ is a *deterministic* process that is independent of $\{Z_t\}$. The Z_t are referred to as *innovations* and are defined by $Z_t = X_t - \mathrm{E}(X_t | X_{t-1}, X_{t-2}, \ldots)$. A process $\{V_t\}$ is deterministic if V_t is completely determined by its past history $\{V_s, s < t\}$. An example of such a processes is the random sinusoid, $V_t = A\cos(\nu t + \Theta)$, where A and Θ are independent random variables with $A \geq 0$ and Θ distributed uniformly on $[0, 2\pi)$. In this case, V_2 is completely determined by the values of V_0 and V_1. In most time series modeling applications, the deterministic component of a time series is either not present or easily removed.

Purely nondeterministic Gaussian processes do not possess a deterministic component and can be represented as a Gaussian linear processes,

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \tag{6}$$

The autocovariance of $\{X_t\}$ has the form

$$\gamma(h) = \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \tag{7}$$

The class of autoregressive (AR) processes, and its extensions, autoregressive moving-average (ARMA) processes, are dense in the class of Gaussian linear processes. A Gaussian AR(p) process satisfies the recursions

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t \tag{8}$$

where $\{Z_t\}$ is an iid sequence of $\mathrm{N}(0, \sigma^2)$ random variables, and the polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ has no zeros inside or on the unit circle. The AR(p) process has a linear representation (6) where the coefficients are found as functions of the ϕ_j (see [2]). Now for any Gaussian linear process, there exists an AR(p) process such that the difference in the two autocovariance functions can be made arbitrarily small for all lags. In fact, the autocovariances can be matched up perfectly for the first p lags.

Prediction

Recall that if two random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ have a joint normal distribution, i.e.

$$\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

and Σ_{22} is nonsingular, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ has a multivariate normal distribution with mean

$$\boldsymbol{\mu}_{X_1|X_2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \qquad (9)$$

and covariance matrix

$$\Sigma_{X_1|X_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad (10)$$

The key observation here is that the *best mean square error predictor* of $\mathbf{X}_1$ in terms of $\mathbf{X}_2$ (i.e. the multivariate function $\mathbf{g}(\mathbf{X}_2)$ that minimizes $\mathrm{E}\|\mathbf{X}_1 - \mathbf{g}(\mathbf{X}_2)\|^2$, where $\|\cdot\|$ is Euclidean distance) is $\mathrm{E}(\mathbf{X}_1|\mathbf{X}_2) = \boldsymbol{\mu}_{X_1|X_2}$ which is a *linear* function of $\mathbf{X}_2$. Also, the covariance matrix of prediction error, $\Sigma_{X_1|X_2}$, does not depend on the value of $\mathbf{X}_2$. These results extend directly to the prediction problem for Gaussian processes.

Suppose $\{X_t, t = 1, 2, \ldots\}$ is a stationary Gaussian process with mean μ and autocovariance function $\gamma(\cdot)$ and that based on the random vector consisting of the first n observations, $\mathbf{X}_n = (X_1, \ldots, X_n)'$, we wish to predict the next observation X_{n+1}. Prediction for other lead times is analogous to this special case. Applying the formula in (9), the best one-step-ahead predictor of X_{n+1} is given by

$$\widehat{X}_{n+1} := \mathrm{E}(X_{n+1}|X_1, \ldots, X_n) = \mu + \phi_{n1}(X_n - \mu) + \cdots + \phi_{nn}(X_1 - \mu) \qquad (11)$$

where

$$(\phi_{n1}, \ldots, \phi_{nn})' = \Sigma_n^{-1}\gamma_n \qquad (12)$$

$\Sigma_n = \mathrm{cov}(\mathbf{X}_n, \mathbf{X}_n)$, and $\gamma_n = \mathrm{cov}(\mathbf{X}_{n+1}, \mathbf{X}_n) = (\gamma(1), \ldots, \gamma(n))'$. The mean square error of prediction is given by

$$v_n = \gamma(0) - \gamma_n'\Sigma_n^{-1}\gamma_n \qquad (13)$$

These formulas assume that Σ_n is nonsingular. If Σ_n is singular, then there is a linear relationship among $X_1, \ldots, X_n$ and the prediction problem can then be recast by choosing a generating prediction subset consisting of *linear independent* variables. The covariance matrix of this prediction subset will be nonsingular. A mild and easily verifiable condition for ensuring nonsingularity of Σ_n for all n is that $\gamma(h) \to 0$ as $h \to \infty$ with $\gamma(0) > 0$ (see [1]).

While (12) and (13) completely solve the prediction problem, these equations require the inversion of an $n \times n$ covariance matrix which may be difficult and time consuming for large n. The Durbin–Levinson algorithm (see [1]) allows one to compute the coefficient vector $\phi_n = (\phi_{n1}, \ldots, \phi_{nn})'$ and the one-step prediction errors v_n recursively from ϕ_{n-1}, v_{n-1}, and the autocovariance function.

The Durbin–Levinson Algorithm

The coefficients ϕ_n in the calculation of the one-step prediction error (11) and the mean square error of prediction (13) can be computed recursively from the equations

$$\phi_{nn} = \left(\gamma(n) - \sum_{j=1}^{n}\phi_{n-1,j}\gamma(n-1)\right)^{-1} v_{n-1}^{-1}$$

$$\begin{bmatrix}\phi_{n,1}\\ \vdots\\ \phi_{n,n-1}\end{bmatrix} = \begin{bmatrix}\phi_{n-1,1}\\ \vdots\\ \phi_{n-1,n-1}\end{bmatrix} - \phi_{nn}\begin{bmatrix}\phi_{n-1,n-1}\\ \vdots\\ \phi_{n-1,1}\end{bmatrix}$$

$$v_n = v_{n-1}(1 - \phi_{nn}^2) \qquad (14)$$

where $\phi_{11} = \gamma(1)/\gamma(0)$ and $v_0 = \gamma(0)$.

If $\{X_t\}$ follows that AR(p) process in (8), then the recursions simplify a great deal. In particular, for $n > p$, the coefficients $\phi_{nj} = \phi_j$ for $j = 1, \ldots, p$ and $\phi_{nj} = 0$ for $j > p$ giving

$$\widehat{X}_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n-p} \qquad (15)$$

with $v_n = \sigma^2$.

The sequence of coefficients $\{\phi_{jj}, j \geq 1\}$ is called the partial autocorrelation function and is a useful tool for model identification. The partial autocorrelation at lag j is interpreted as the correlation between X_1 and X_{j+1} after correcting for the intervening observations $X_2, \ldots, X_j$. Specifically, ϕ_{jj} is the correlation of the two residuals obtained by regression of X_1 and X_{j+1} on the intermediate observations $X_2, \ldots, X_j$. Of particular interest is the relationship between ϕ_{nn} and the reduction in the one-step mean square error as the number of predictors is increased from $n - 1$

to n. The one-step prediction error has the following decomposition in terms of the partial autocorrelation function:

$$v_n = \gamma(0)(1 - \phi_{11}^2) \cdots (1 - \phi_{nn}^2) \tag{16}$$

For a Gaussian process, $X_{n+1} - \widehat{X}_{n+1}$ is normally distributed with mean 0 and variance v_n. Thus,

$$\widehat{X}_{n+1} \pm z_{1-\alpha/2} v_n^{-1/2}$$

constitute $(1 - \alpha)$ 100% prediction bounds for the observation X_{n+1}, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. In other words, X_{n+1} lies between the bounds $\widehat{X}_{n+1} \pm z_{1-\alpha/2} v_n^{-1/2}$ with probability $1 - \alpha$.

Estimation for Gaussian Processes

One of the advantages of Gaussian models is that an explicit and closed form of the likelihood is readily available. Suppose that $\{X_t, t = 1, 2, \ldots, \}$ is a stationary Gaussian time series with mean μ and autocovariance function $\gamma(\cdot)$. Denote the data vector by $\mathbf{X}_n = (X_1, \ldots, X_n)$ and the vector of one-step predictors by $\widehat{\mathbf{X}}_n = (\widehat{X}_1, \ldots, \widehat{X}_n)'$, where $\widehat{X}_1 = \mu$ and $\widehat{X}_j = \mathrm{E}(X_j | X_1, \ldots, X_{j-1})$ for $j \geq 2$. If Σ_n denotes the covariance matrix of $\mathbf{X}_n$, which we assume is nonsingular, then the likelihood of $\mathbf{X}_n$ is

$$L(\Sigma_n, \mu) = (2\pi)^{-n/2} (\det \Sigma_n)^{-1/2} \times \exp\left(-\tfrac{1}{2}(\mathbf{X}_n - \mu\mathbf{1})' \Sigma_n^{-1} (\mathbf{X}_n - \mu\mathbf{1})\right) \tag{17}$$

where $\mathbf{1} = (1, \ldots, 1)'$. Typically, Σ_n will be expressible in terms of a finite number of unknown parameters, $\beta_1, \ldots, \beta_r$, so that the maximum likelihood estimator of these parameters and μ are those values that maximize L for the given dataset. Under mild regularity assumptions, the resulting maximum likelihood estimators are approximately normally distributed with covariance matrix given by the inverse of the Fisher information.

In most settings, direct-closed-form maximization of L with respect to the parameter set is not achievable. In order to maximize L using numerical methods, either derivatives or repeated calculation of the function are required. For moderate to large sample sizes n, calculation of both the determinant of Σ_n and the quadratic form in the exponential of L can be difficult and time consuming. On the other hand, there is a useful representation of the likelihood in terms of the one-step prediction errors and their mean square errors. By the form of $\widehat{\mathbf{X}}_n$, we can write

$$\mathbf{X}_n - \widehat{\mathbf{X}}_n = A_n \mathbf{X}_n \tag{18}$$

where A_n is a lower triangular square matrix with ones on the diagonal. Inverting this expression, we have

$$\mathbf{X}_n = C_n(\mathbf{X}_n - \widehat{\mathbf{X}}_n) \tag{19}$$

where C_n is also lower triangular with ones on the diagonal. Since $X_j - \mathrm{E}(X_j | X_1, \ldots, X_{j-1})$ is uncorrelated with $X_1, \ldots, X_{j-1}$, it follows that the vector $\mathbf{X}_n - \widehat{\mathbf{X}}_n$ consists of uncorrelated, and hence independent, normal random variables with mean 0 and variance v_{j-1}, $j = 1, \ldots, n$. Taking covariances on both sides of (19) and setting $D_n = \mathrm{diag}\{v_0, \ldots, v_{n-1}\}$, we find that

$$\Sigma_n = C_n D_n C_n' \tag{20}$$

and

$$(\mathbf{X}_n - \mu\mathbf{1})' \Sigma_n^{-1} (\mathbf{X}_n - \mu\mathbf{1}) = (\mathbf{X}_n - \widehat{\mathbf{X}}_n)' D_n^{-1} (\mathbf{X}_n - \widehat{\mathbf{X}}_n) = \sum_{j=1}^{n} \frac{(\mathbf{X}_j - \widehat{\mathbf{X}}_j)^2}{v_{j-1}} \tag{21}$$

It follows that $\det \Sigma_n = v_0 v_1 \ldots v_{n-1}$ so that the likelihood reduces to

$$L(\Sigma_n, \mu) = (2\pi)^{-n/2} (v_0 v_1 \ldots v_{n-1})^{-1/2} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \frac{(\mathbf{X}_j - \widehat{\mathbf{X}}_j)^2}{v_{j-1}}\right) \tag{22}$$

The calculation of the one-step prediction errors and their mean square errors required in the computation of L based on (22) can be simplified further for a variety of time series models such as ARMA processes. We illustrate this for an AR process.

Gaussian Likelihood for an AR(p) Process

If $\{X_t\}$ is the AR(p) process specified in (8) with mean μ, then one can take advantage of the simple form for the one-step predictors and associated mean square errors. The likelihood becomes

$$L(\phi_1, \ldots, \phi_p, \mu, \sigma^2) = (2\pi)^{-(n-p)/2} \sigma^{-(n-p)}$$

$$\times \exp\left(-\frac{1}{2}\sum_{j=p+1}^{n}\frac{(\mathbf{X}_j-\widehat{\mathbf{X}}_j)^2}{\sigma^2}\right)(2\pi)^{-p/2}$$

$$\times (v_0v_1\dots v_{p-1})^{-1/2}\exp\left(-\frac{1}{2}\sum_{j=1}^{p}\frac{(\mathbf{X}_j-\widehat{\mathbf{X}}_j)^2}{v_{j-1}}\right) \qquad (23)$$

where, for $j > p$, $\widehat{X}_j = \mu + \phi_1(X_{j-1} - \mu) + \cdots + \phi_p(X_{j-p-1} - \mu)$ are the one-step predictors. The likelihood is a product of two terms, the conditional density of $\mathbf{X}_n$ given $\mathbf{X}_p$ and the density of $\mathbf{X}_p$. Often, just the conditional maximum likelihood estimator is computed which is found by maximizing the first term. For the AR process, the conditional maximum

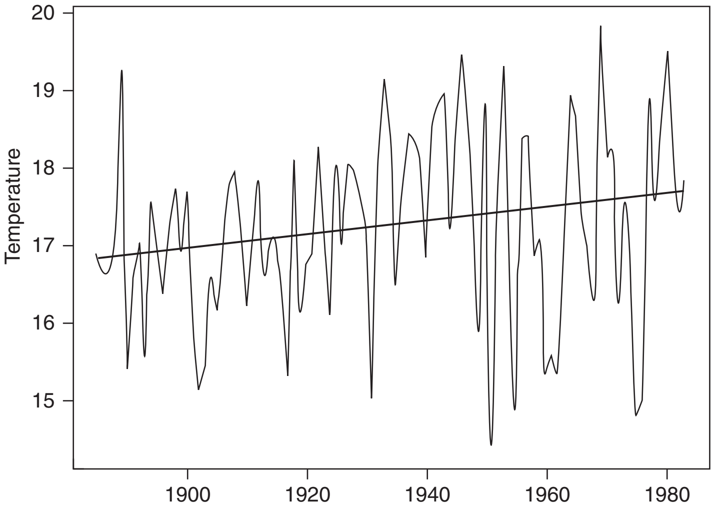

Figure 1 Average maximum temperature, 1885–1993. Regression line is $16.83 + 0.008\,45t$

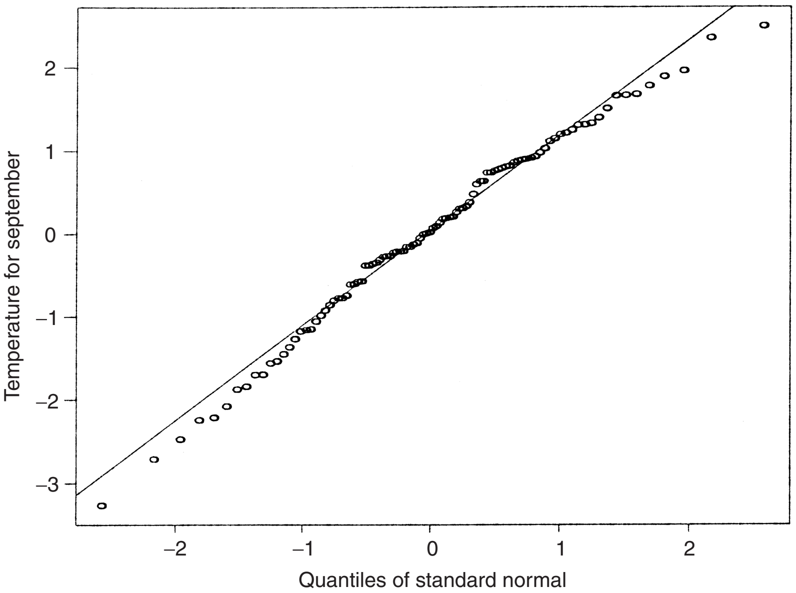

Figure 2 *QQ* plot for normality of the innovations

likelihood estimator can be computed in closed form.

Example This example consists of the average maximum temperature over the month of September for the years 1895–1993 in an area of the US whose vegetation is characterized as tundra. The time series $x_1, \ldots, x_{99}$ is plotted in Figure 1. Here we investigate the possibility of the data exhibiting a slight linear trend. After inspecting the residuals from fitting a least squares regression line to the data, we entertain a time series model of the form

$$X_t = \beta_0 + \beta_1 t + W_t \tag{24}$$

where $\{W_t\}$ is the Gaussian AR(1),

$$W_t = \phi_1 W_{t-1} + Z_t \tag{25}$$

and $\{Z_t\}$ is a sequence of iid $\mathrm{N}(0, \sigma^2)$ random variables. After maximizing the Gaussian likelihood over the parameters β_0, β_1, ϕ_1, and σ^2, we find that the maximum likelihood estimate of the mean function is $16.83 + 0.008\,45t$. The maximum likelihood parameters of ϕ_1 and σ^2 are estimated by 0.1536 and 1.3061, respectively. The maximum likelihood estimates of β_0 and β_1 can be viewed as generalized least squares estimates assuming that the residual process follows the estimated AR(1) model. The resulting standard errors of these estimates are 0.277 81 and 0.004 82, respectively, which provides some doubt about the significance of a nonzero slope of the line. Without modeling the dependence in the residuals, the slope would have been deemed significant using classical inference procedures. By modeling the dependence in the residuals, the evidence in favor of a nonzero slope has diminished somewhat. The *QQ* plot of the estimated innovations is displayed in Figure 2. This plot shows that the AR(1) model is not far from being Gaussian. Further details about inference procedures for regression models with time series errors can be found in [2, Chapter 6].

References

[1] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer-Verlag, New York.

[2] Brockwell, P.J. & Davis, R.A. (1996). *Introduction to Time Series and Forecasting*, Springer-Verlag, New York.

[3] Diggle, Peter J., Liang, Kung-Yee & Zeger, Scott L. (1996). *Analysis of Longitudinal Data*, Clarendon Press, Oxford.

[4] Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer-Verlag, New York.

[5] Rosenblatt, M. (2000). *Gaussian and Non-Gaussian Linear Time Series and Random Fields*, Springer-Verlag, New York.

RICHARD A. DAVIS