Søren Johansen

THE ROLE OF ANCILLARITY IN INFERENCE FOR NON-STATIONARY VARIABLES

5 Preprint
May
1994

Institute of Mathematical Statistics
University of Copenhagen

The role of ancillarity in inference for non-stationary variables*

Søren Johansen
Institute of Mathematical Statistics
Universitetsparken 5, 2100 Copenhagen Ø
Denmark

Abstract

Some examples of the regression method are compared with likelihood based inference. It is shown that although the asymptotic theory is distinctly different for ergodic and non-ergodic processes, the likelihood methods lead to the result that asymptotic inference can be conducted in the same way for the two cases by appealing to classical conditioning arguments from statistics using the notion of S-ancillarity or strong exogeneity. It is pointed out that the Fisher information can be considered a measure of the conditional variance of the maximum likelihood estimator given the available information in the sample.

The purpose of this paper is to discuss conditional inference in connection with the usual regression problem in econometrics, and the analysis of the error correction model in the presence of cointegration. The starting point is that inference concerning the cointegrating coefficients is mixed Gaussian, see Phillips (1991), Reinsel and Ahn (1990) or Johansen (1988). Thus the limit distribution itself offers the possibility to make a conditioning argument when deriving the asymptotic distribution of the test statistic for hypotheses on the cointegrating coefficients. It is the intent to investigate to what extent it is possible to argue for the conditioning using ideas of conditioning in the statistical literature. The following quotations by Sir R.A. Fisher are taken from the paper by Efron and Hinkley (1978), who discuss conditioning in the classical case of *i.i.d* measurements.

> Fisher(1934): When these [log likelihood] functions are differentiable successive portions of the [information] loss may be recovered by using as ancillary statistics, in addition to the maximum likelihood estimate the second and higher differential coefficients at the maximum.

*Paper presented as the Frank Paish lecture at the Royal Economic Society Meeting in Exeter 1994.

> Fisher (1925): The function of the ancillary statistic is analogous to providing a true, in place of an approximate, weight for the value of the estimate.

Efron and Hinkley find that in models with parameter ϑ and an ancillary (or approximately ancillary) statistic a, one finds that the inverse information gives an approximation to the conditional variance of $\hat{\vartheta}$ given a with a relative error of T^{-1} rather as was to be expected $T^{-\frac{1}{2}}$. In case the information itself is ancillary we can choose that as the conditioning statistic.

The idea of applying conditioning in regression is of course not new. Bartlett (1939) discusses the concept of conditioning, and notices:

> Consider similarly the test of significance of a regression coefficient. The orthodox theory is to consider the conditional statistic $b|\Sigma_1^n (x_i - \bar{x})^2$, where b is our estimate, and $\Sigma_1^n (x_i - \bar{x})^2$ the sum of squares of deviations of the independent variable x.

The present paper represents an attempt to apply conditioning ideas to the regression and cointegration models for non-stationary variables and contains very little new. Its contribution, if any, is to reinterpret the now standard limit results about mixed Gaussian distributions. We proceed by examples and strive for simplicity to illustrate ideas rather than generality to cover all possible cases.

1 Regression with deterministic regressors

As a first example we consider simple linear regression. This establishes some notation and serves as a reminder of some well known results. We define the process Y_t, $t = 1, ..., T$ by the equations

$$Y_t = \beta' x_t + \epsilon_t, \tag{1}$$

where x_t are deterministic regressors, β an unrestricted m-dimensional parameter and ϵ_t are independent 1-dimensional Gaussian variables with mean zero and variance σ^2, which for simplicity is assumed known. It is well known that ordinary least squares coincides with maximum likelihood estimation in this case and that

$$\hat{\beta} - \beta = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t \epsilon_t \right), \tag{2}$$

which is Gaussian with mean zero and variance $\sigma^2 (\sum_{t=1}^{T} x_t x_t')^{-1}$. The reason that we want the distribution of $\hat{\beta}$ is that we want to be able to conduct inference, that is, test hypotheses about the coefficients of β. If we want to test a simple hypothesis about β then the Wald statistic which is equivalent to the likelihood ratio test is

$$\sigma^{-2}\left(\hat{\beta}-\beta\right)'\sum_{t=1}^{T}x_t x_t'\left(\hat{\beta}-\beta\right), \tag{3}$$

which is distributed as $\chi^2(m)$. A confidence interval or set for β is found from (3) as

$$\{\beta|\sigma^{-2}\left(\hat{\beta}-\beta\right)'\sum_{t=1}^{T}x_t x_t'\left(\hat{\beta}-\beta\right)\leq c\},$$

and for a univariate parameter ($m=1$) we usually communicate

$$\hat{\beta}\pm 2\hat{V}ar\left(\hat{\beta}\right)^{\frac{1}{2}}=\hat{\beta}\pm 2\hat{\sigma}\left(\Sigma_1^T x_t^2\right)^{-\frac{1}{2}}.$$

Thus in this case the distribution of the test statistic and the confidence limit is derived directly from the distribution of the estimated parameter. The distribution theory is standard in the sense that only χ^2 or F- tables are needed.

We now give an analysis of the likelihood function:

$$logL(\beta)=-\frac{1}{2}T\log(2\pi)-\frac{1}{2}T\log\sigma^2-\frac{1}{2}\sigma^{-2}\sum_{t=1}^{T}\left(Y_t-\beta' x_t\right)^2.$$

We find

$$\partial logL(\beta)/\partial\beta=\sigma^{-2}\sum_{t=1}^{T}\left(Y_t-\beta' x_t\right)x_t',$$

$$J_T(\beta)=-\partial^2 logL(\beta)/\partial\beta^2=\sigma^{-2}\sum_{t=1}^{T}x_t x_t'.$$

The negative second derivative is the observed information about β in the whole sample, which in this case is also the expected information $I_T(\beta)=E\left(J_T(\beta)\right)$ since the regressors are deterministic. The Wald test for a simple hypothesis about β can be calculated in three forms which in the present context are identical

$$W_{var}=\left(\hat{\beta}-\beta\right)'\hat{V}ar\left(\hat{\beta}\right)^{-1}\left(\hat{\beta}-\beta\right), \tag{4}$$

$$W_{\exp}=\left(\hat{\beta}-\beta\right)'\hat{I}_T(\beta)\left(\hat{\beta}-\beta\right), \tag{5}$$

$$W_{obs}=\left(\hat{\beta}-\beta\right)'J_T\left(\hat{\beta}\right)\left(\hat{\beta}-\beta\right). \tag{6}$$

Here W_{var} has the estimated variance of $\hat{\beta}$ as the normalizing matrix, whereas $W_{\exp}$ has the estimated expected information as weight matrix. In W_{obs} this is replaced by the observed information, or Hessian matrix, evaluated at the

maximum point of the likelihood function. In the present case all these measures are the same because the regressors are non-stochastic and $J_T(\beta) = I_T(\beta) = Var\left(\hat{\beta}\right)^{-1}$.

The derivations behind these Wald tests is, apart from some regularity conditions, the following. In a statistical problem with the parameter β we expand the derivative of the log-likelihood function around the maximum likelihood estimator $\hat{\beta}$, see Cox and Hinkley (1974), and find

$$\left[-\partial^2 logL(\beta)/\partial\beta^2\right]\left(\hat{\beta}-\beta\right) \approx \partial logL(\beta)/\partial\beta. \tag{7}$$

An expansion of the likelihood function around $\hat{\beta}$ gives

$$-2log\left(L(\beta)/L\left(\hat{\beta}\right)\right) \approx \left(\hat{\beta}-\beta\right)'\left[-\partial^2 logL(\beta)/\partial\beta^2|_{\beta=\hat{\beta}}\right]\left(\hat{\beta}-\beta\right) = W_{obs}. \tag{8}$$

Under suitable conditions on the observations one can prove that the normed score function $T^{-\frac{1}{2}}\partial logL(\beta)/\partial\beta$ is asymptotically Gaussian, that $T^{-1}J_T(\beta) = -T^{-1}\partial^2 logL(\beta)/\partial\beta^2$ and its expectation converge to a quantity $I(\beta)$, the information per. observation, which is also the variance in the asymptotic distribution of $T^{-\frac{1}{2}}\partial logL(\beta)/\partial\beta$ as well as the inverse of the asymptotic variance of the maximum likelihood estimator $Var\left(\hat{\beta}\right)^{-1}$. The relation (7) implies that

$$T^{\frac{1}{2}}\left(\hat{\beta}-\beta\right) \xrightarrow{w} N[0, I(\beta)^{-1}],$$

and (8) shows that

$$-2log[L(\beta)/L\left(\hat{\beta}\right)] \xrightarrow{w} \chi^2(m).$$

The reason for listing these well known results is that as we relax the conditions on the regressors, some of these results are still valid, while others are valid asymptotically, and still others are incorrect. Note that it is W_{obs} that appears in the expansion of the likelihood ratio test, and that $W_{\exp}$ and W_{var} are approximations to W_{obs}.

2 Regression with ergodic regressors

Consider model (1) but assume now that $\{X_t\}$ is an ergodic and stationary sequence which is independent of the sequence $\{\epsilon_t\}$. Equation (1) has the interpretation as an expression for the conditional distribution of Y_t given $X_t = x_t$ and the past. The regression estimator can be calculated as before and (2) again holds. The exact (marginal) distribution depends on the properties of the sequence X_t, but the conditional distribution given $\{X_t\}$ is the same as before: For fixed values of the sequence $\{X_t\}$ the sequence $\{Y_t\}$ is define by the model (1) with deterministic regressors $x_t = X_t$. Thus for fixed values of $\{X_t\}$ the distribution of $\hat{\beta}$ is Gaussian with conditional mean β and conditional variance

$\sigma^2(\sum_{t=1}^T X_t X_t')^{-1}$. Whatever the distribution of the sequence of regressors we thus find that the assumption of independence between errors and regressors implies that $\hat{\beta}$ conditionally on the sequence $\{X_t\}$ is Gaussian with mean β and variance $\sigma^2(\sum_{t=1}^T X_t X_t')^{-1}$, and hence that the marginal distribution of $\hat{\beta}$ is mixed Gaussian. Note that the conditional distribution of $\hat{\beta}$ given all the $x's$ is the same as the conditional distribution given only the observed information $\sigma^{-2}\Sigma_1^T X_t X_t'$. Thus we call $\Sigma_1^T X_t X_t'$ ancillary if it has exactly the property that Fisher (1925) suggested. It is easily seen that the asymptotic distribution of $T^{\frac{1}{2}}\left(\hat{\beta}-\beta\right)$ is Gaussian with mean zero and variance given by $I(\beta)^{-1}$, where $I(\beta) = \sigma^{-2}E(X_t X_t') = \sigma^{-2}P\lim T^{-1}\sum_{t=1}^T X_t X_t'$.

The Wald statistic derived from the conditional distribution is

$$W_{obs} = \sigma^{-2}\left(\hat{\beta}-\beta\right)'\sum_{t=1}^{T} X_t X_t'\left(\hat{\beta}-\beta\right).$$

This statistic is not only asymptotically χ^2, but the actual distribution is χ^2 since if we condition on the sequence $\{X_t\}$ then the exact conditional Gaussian distribution of Y_t implies that W_{obs} is exactly $\chi^2(m)$ distributed, and since this conditional distribution is the same for any value of the conditioning process $\{X_t\}$, the result also holds unconditionally.

Another way of writing the distributional result is that for any vector ξ it holds that

$$[\xi'(\sum_{t=1}^{T} X_t X_t')^{-1}\xi]^{-\frac{1}{2}}[\xi'\left(\hat{\beta}-\beta\right)] \tag{9}$$

is distributed as $\mathsf{N}(0,\sigma^2)$. This result for ξ equal to a unit vector, e_1 say, gives a way of testing the value of a single coefficient, β_1, by evaluating the deviation between the estimated value $\hat{\beta}_1$and the true value β_1, by a measure of its variation.

Note that $\sigma[\xi'(\sum_{t=1}^T X_t X_t')^{-1}\xi]^{\frac{1}{2}}$ is not the standard deviation of $\hat{\beta}_1$ but rather a consistent estimate of this parameter. It is, however, not really necessary with this asymptotic interpretation, since in this case we have that it is the exact conditional standard deviation. Thus if we could appeal to a "principle of conditionality" we can make exact inference.

The reason that we normalize by $[\xi'(\sum_{t=1}^T X_t X_t')^{-1}\xi]^{\frac{1}{2}}$ is thus not to achieve an asymptotically valid result but because we can exploit the mixed Gaussian distribution in this way. A stronger way of saying this is that we make no use of the actual distribution of the estimator, but rather of its conditional distribution given $\{X_t\}$ or equivalently $\sum_{t=1}^T X_t X_t'$. Thus in the case of regression with ergodic regressors independent of the Gaussian errors we need the conditional distribution given the regressors or equivalently the product moments, not the distribution of the estimator itself. It is of course very difficult to tell the difference because

the implementation of both methods is the same, because of the convergence of the estimated variance to its population value. The problem arises in this simple model, but becomes important only in the next models we consider. See, however, the paper by Efron and Hinkley (1978) for a more careful discussion of the interpretation of the information as an approximate ancillary statistic in the *i.i.d* case.

A likelihood analysis of this model is somewhat more complicated, since we need to specify the joint density of $\{Y_t\}$ and $\{X_t\}$. The assumption of independence between $\{X_t\}$ and $\{\epsilon_t\}$ allows us to calculate the conditional distribution of $\{Y_t\}$ given $\{X_t\}$ by the Gaussian likelihood, and we can then choose any class of distributions we want for the distribution of the process $\{X_t\}$, as long as the process is ergodic. The only other requirement is that the parameter ϑ in the distribution of $\{X_t\}$ is variation independent of the parameter β, that is, they vary without restrictions in a product space. If this is the case then X_t is strongly exogenous for β, see Hendry and Richard (1983) and the whole analysis is as in Section 1. That is, we have

$$L_{Y,X}(\beta,\vartheta) = L_{Y|X}(\beta)\, L_X(\vartheta),$$

which shows that

$$\partial log L_{Y,X}(\beta)/\partial\beta = \partial log L_{Y|X}(\beta)/\partial\beta,$$

$$\partial^2 log L_{Y,X}(\beta)/\partial\beta^2 = \partial^2 log L_{Y|X}(\beta)/\partial\beta^2.$$

Thus the observed information about β is

$$J_T(\beta) = \sigma^{-2}\sum_{t=1}^{T} X_t X_t',$$

and the expected information is

$$I_T(\beta) = \sigma^{-2} T E(X_t X_t') = \sigma^2 T I(\beta).$$

Thus all calculations can be performed in the conditional distribution provided we assume Gaussian distribution and variation independence between the parameters.

We can replace the observed information $\sigma^{-2}\sum_{t=1}^{T} X_t X_t'$ in W_{obs} by the expected information $I_T(\beta)$ to get $W_{\exp}$, but this expectation should then be calculated in the distribution of X_t. If we estimate it by the natural estimate namely $J_T(\beta)$ then we get W_{obs} again. One could in principle calculate

$$Var\left(\hat{\beta}\right) = E[Var(\hat{\beta}|\{X_t\})] = \sigma^2 E[(\sum_{t=1}^{T} X_t X_t')^{-1}],$$

but again a natural estimator is $J_T(\beta)^{-1}$ and W_{var} is then also equal to W_{obs}.

Thus the difference between W_{obs}, W_{var}, $W_{\exp}$ is small and not very important asymptotically.

3 Ancillarity and exogeneity

At this point it becomes important to remind about the conditionality arguments that have been discussed in statistics. In a statistical model given by the densities $f(z,\vartheta)$, $\vartheta \in \Theta$ for the random variable Z we call the statistic $t(Z)$ ancillary if the density of Z factorizes into a product of the conditional density given $t(Z)$ and the marginal density of $t(Z)$ which does not depend on the parameter:

$$f(z,\vartheta) = g[t(z)]h[z|t(z),\vartheta], \vartheta \in \Theta.$$

That is, the marginal distribution does not depend on the parameter and in this sense the observation of $t(Z)$ does not contain any information about the parameter ϑ. This is a precise definition and an interpretation which tries to capture the meaning of the word ancillary that Fisher alluded to in the above quotations. In econometrics we rarely meet this concept because the models considered are very complicated.

Another way of approaching the topic is the notation of S-ancillarity Barndorff-Nielsen (1978) or strong exogeneity Hendry and Richard (1983). Let $f(z,\varphi,\lambda)$, $(\varphi,\lambda) \in \Theta$ define a statistical model. The statistic $t(Z)$ is called S-ancillary for $\tau(\varphi)$ if

$$f(z,\varphi,\lambda) = g[t(z),\lambda]h[z|t(z),\varphi], (\varphi,\lambda) \in A \times B.$$

Thus the marginal distribution of $t(Z)$ does contain parameters, but they have "nothing to do with" the parameters of interest. Note that a consequence of S-ancillarity is that the maximum likelihood estimator for φ can be derived solely from the conditional distribution given $t(Z)$. This is just a consequence of the decomposition of the likelihood function. The principle of conditionality, on the other hand, asserts that since the distribution of $t(Z)$ contains no information about the parameter of interest $\tau(\varphi)$ the variation in the data due to the variation of $t(Z)$ is irrelevant for inference concerning $\tau(\varphi)$, and hence the estimator of τ should be evaluated in the distribution conditional on $t(Z)$. Thus confidence intervals for τ should be based upon the conditional variance of $\hat{\tau}$ given $t(Z)$, not the marginal variance of $\hat{\tau}$.

With this terminology we see that in the situation with ergodic regressors independent of the errors ϵ_t, the process X_t is strongly exogenous or S- ancillary and inference can and should be conducted conditionally on the process $\{X_t\}$.

It is an important aspect of strong ancillarity that it requires the correct parameterization. That is, sometimes the strong exogeneity is only valid after the model has been reparameterized into variation independent parameters, and

conversely if $t(Z)$ is strongly exogenous for some parameter $\tau(\varphi)$ in a model with parameter (φ, λ) it need not be if we reparametrize into (φ, γ) where $\gamma = \gamma(\varphi, \lambda)$.

4 Regression with non-ergodic regressors

In the next example we consider equation (1) and let the process $\{X_t\}$ to be non-ergodic and independent of the error $\{\epsilon_t\}$. In this case the regression estimator again satisfies (2), and the conditional distribution given the sequence $\{X_t\}$ is the same as before. Hence again tests on β can be conducted in the conditional distribution using the χ^2 distribution, since $\{X_t\}$ is strongly exogenous.

The likelihood formulation in this case is exactly as in section 2 in that the model so far only specifies the distribution of $\{Y_t\}$ given $\{X_t\}$. If we choose as before a class of distributions for $\{X_t\}$ parameterized by ϑ which is variation independent of β, then the observed information is as before:

$$J_T(\beta) = \sigma^{-2} \sum_{t=1}^{T} X_t X_t',$$

and the expected information becomes

$$I_T(\beta) = \sigma^{-2} E\left(\sum_{t=1}^{T} X_t X_t'\right).$$

The variance of the estimator for β is calculated as

$$Var\left(\hat{\beta}\right) = E[Var\left(\hat{\beta}|\{X_t\}\right)] = \sigma^2 E[(\sum_{t=1}^{T} X_t X_t')^{-1}].$$

We investigate the Wald test statistics (4), (5) and (6) in order to see how they are related in the non-ergodic case. Consider for simplicity that X_t is a random walk so that the model becomes

$$Y_t = \beta' X_t + \epsilon_{1t},$$

$$\Delta X_t = \epsilon_{2t},$$

where ϵ_t, $t = 1, ..., T$ are independent Gaussian in $1 + m$ dimensions with mean zero and variance matrix

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sum \end{pmatrix}.$$

In this case the asymptotics is a bit more tricky. It holds that

$$T^{-\frac{1}{2}} \sum_{t=1}^{[Tu]} \epsilon_t \xrightarrow{w} B(u),$$

where $B = (B_1, B_2')'$ is an $1+m$ dimensional Brownian motion such that B_1 and B_2 are independent. It follows, see Phillips and Durlauf (1986), that

$$T^{-2}\sum_{t=1}^{T} X_t X_t' \xrightarrow{w} \int_0^1 B_2(u) B_2(u)' du, \tag{10}$$

$$T^{-1}\sum_{t=1}^{T} X_t \epsilon_t \xrightarrow{w} \int_0^1 B_2(dB_1). \tag{11}$$

The first result involves an ordinary integral of the continuous Brownian motion and the second integral involves a stochastic integral. It is not important to understand the exact definition of a stochastic integral for the present presentation. It suffices to think of it as a limit of the sums

$$\sum_{k=1}^{N} B_2(t_k)[B_1(t_{k+1}) - B_1(t_k)],$$

where $0 < t_1 < ... < t_N < t_{N+1} = 1$ is a fine partition of the unit interval. Thus the stochastic integral mimics the definition of the sums on the left hand side of (11).

From the results (10) and (11) we find the asymptotic distribution

$$T\left(\hat{\beta} - \beta\right) \xrightarrow{w} \left[\int_0^1 B_2(u) B_2(u)' du\right]^{-1} \int_0^1 B_2(dB_1).$$

This distribution is also mixed Gaussian. The reason for this is that if we condition on B_2, then $\int_0^1 B_2(u) B_2(u)' du$ is a constant and $\int_0^1 B_2(dB_1)$ is Gaussian with mean zero and variance $\sigma^2 \int_0^1 B_2(u) B_2(u)' du$, since B_1 and B_2 are independent. To see this consider $\Sigma_1^N B_2(t_k)[B_1(t_{k+1}) - B_1(t_k)]$ which is Gaussian with mean zero and conditional variance $\Sigma_1^N B_2(t_k) B_2(t_k)' \sigma^2 (t_{k+1} - t_k)$. This, however, is approximately $\sigma^2 \int_0^1 B_2(u) B_2(u)' du$. Hence the limit distribution is mixed Gaussian with mixing parameter $\sigma^2 \int_0^1 B_2(u) B_2(u)' du$.

The difference between the ergodic case and the non-ergodic case is that in the ergodic case the asymptotic distribution is mixed Gaussian with a degenerate limiting mixing parameter, whereas in the non-ergodic case of a random walk the limit distribution is mixed Gaussian with a non-degenerate mixing distribution. Thus the asymptotic distribution of the estimator is mixed Gaussian and not well suited for making inference about β, see for instance Phillips (1994). The Wald statistic W_{var} derived from the marginal distribution of $\hat{\beta}$ is

$$W_{var} = (\hat{\beta} - \beta)' Var(\hat{\beta})^{-1} (\hat{\beta} - \beta) = \sigma^{-2} (\hat{\beta} - \beta)' E(\Sigma_{t=1}^T X_t X_t')^{-1} (\hat{\beta} - \beta).$$

The asymptotic distribution of this can be derived by the above results but is clearly rather complex. There is, however, also no reason to conduct this test,

since we have available the more natural Wald test statistic W_{obs} derived from the likelihood function, for which we can find the limit distribution.

We find

$$\begin{aligned} W_{obs} = & \quad \sigma^{-2}\left(\hat{\beta}-\beta\right)' \textstyle\sum_{t=1}^{T} X_t X_t' \left(\hat{\beta}-\beta\right) \\ = & \quad \sigma^{-2} T\left(\hat{\beta}-\beta\right)' [T^{-2}\textstyle\sum_{t=1}^{T} X_t X_t'] T\left(\hat{\beta}-\beta\right) \\ \xrightarrow{w} & \sigma^{-2} \textstyle\int_0^1 (dB_1) B_2' \left[\int_0^1 B_2(u) B_2(u)' du\right]^{-1} \int_0^1 B_2 (dB_1). \end{aligned}$$

For given B_2 this has the form $Z'Var(Z)^{-1}Z$, where Z is Gaussian with mean zero, and hence distributed as $\chi^2(m)$. This is not surprising since the exact distribution is also χ^2. Conditionally on $\Sigma_1^T X_t X_t'$ the statistic W_{obs} is $\chi^2(m)$ and hence also unconditionally.

Finally we consider

$$\begin{aligned} W_{\exp} &= \sigma^{-2}\left(\hat{\beta}-\beta\right)' E\left[\textstyle\sum_{t=1}^{T} X_t X_t'\right]\left(\hat{\beta}-\beta\right) \\ &= \tfrac{1}{2} T(T+1)\sigma^{-2}\left(\hat{\beta}-\beta\right)' \Sigma\left(\hat{\beta}-\beta\right). \end{aligned}$$

Again the limit distribution can be derived but is non-standard and depends on nuisance parameters. Note that the asymptotic properties of the three different forms of the Wald statistics are entirely different in the non-ergodic case. The only manageable one is W_{obs}, whereas W_{var} and $W_{\exp}$ have very difficult limit distributions.

In the previous cases the very strong assumption about independence between the difference $Y_t - \beta' X_t$ and the regressor X_t makes the results of limited use in practice and they are of course only given as an excuse for the discussion of the mixed Gaussian distribution. It is known that even in the ergodic case the lack of independence implies a bias in the regression estimator, and this carries in some sense over to the non-ergodic case, see Phillips (1991). What happens is that the limit distribution of the regression estimate becomes rather complicated and it is not so obvious how one should modify the regression estimator in order to avoid the bad properties, see Phillips and Hansen (1990) and Park (1992). We therefore turn to the likelihood method that has the advantage that it almost automatically compensates for complications in the dependence structure in the model by suggesting a new estimator.

In the non-ergodic ($I(1)$) case the observed information $J_T(\beta)$ grows like T^2, as does the expected information $I_T(\beta)$, but $T^{-2}J_T(\beta)$ does not converge to the same limit as $T^{-2}I_T(\beta)$, but to a stochastic limit. Thus even in the limit the observed information about β is random. This means that in the classical case of inference for stationary processes the information per observation in a long series of observations, is roughly the same for every series, whereas for the case of inference for non-stationary processes, the information normalized by T^2 is random even in the limit. This shows that there are sample paths or

series of realized values that sometimes contain very little information about the parameter. We can see why. A random walk usually exhibits a trending behavior which shows that the information, as measured by the cumulated sum of squares, is very large, but clearly a random walk can by accident in a given sample fluctuate around the value zero, in which case the information never builds up. Thus we should be aware that for some realizations there is little information about β in the sample, whereas for others there is a lot of variation. This is reflected in the choice of W_{obs} based upon the observed information, where deviations of $\hat{\beta}$ from β are measured by the cumulated information in the actual sample rather than by the expected information.

It is therefore argued that for non-stationary variables the proper choice of Wald statistic is not the usual W_{var} based on a measure of the variance of the estimator, but rather W_{obs} which takes into account the actual information in the sample that one has obtained. Another way of saying this is that the usual approximation that leads from W_{obs} to W_{var} is not valid for non-ergodic processes. A consequence is that for the univariate case ($m = 1$) a confidence set is not given by $\hat{\beta} \pm 2\hat{V}ar\left(\hat{\beta}\right)^{\frac{1}{2}}$ which would correspond to choosing W_{var} as the test statistic, but rather to $\hat{\beta} \pm 2J_T\left(\beta\right)^{-\frac{1}{2}}$ corresponding to W_{obs} as the test statistic.

The conditions for applying the likelihood methods is that one needs to specify a full model for all the variables in the model and we thus have to be more precise in the formulation and checking of the model. A class of models that have proved useful in the analysis of macro data is the class of vector autoregressive models and the next two examples will deal with such models and investigate to what extent the problem of correlation between error and regressor can be formulated and solved within this framework.

5 Non-ergodic regressors which are correlated with the error

The first example of the type of problem that can be solved by analyzing the likelihood function and which leads to a modified estimator is

$$Y_t = \beta' X_t + \epsilon_{1t} \tag{12}$$

$$\Delta X_t = \epsilon_{2t}. \tag{13}$$

We assume that $\epsilon_t = (\epsilon_{1t}, \epsilon'_{2t})'$ are independent $1 + m$-dimensional Gaussian with mean zero and variance

$$\begin{pmatrix} \sigma^2 & \omega'\Sigma \\ \Sigma\omega & \Sigma \end{pmatrix}.$$

Note that correlation between ϵ_{1t} and ϵ_{2t} implies a correlation between the regressor X_t and the error ϵ_{1t}. The parameters are (β, ω, Σ) which vary freely.

This model was discussed in detail by Phillips (1991) and is useful as an example of the problems that can arise. A property of (13) is that X_t is a random walk and (12) then shows that Y_t is non-stationary even though the combination $Y_t - \beta' X_t$ is stationary. This is an instance of cointegration between Y_t and X_t, which are called integrated processes, see Engle and Granger (1987). The expectation of Y_t given X_t and the past is given by

$$E\left(Y_t | X_t, past\right) = \beta' X_t + \omega' \Delta X_t.$$

Thus we can replace equation (12) by the regression equation

$$Y_t = \beta' X_t + \omega \Delta X_t + \epsilon_{1.2t}, \tag{14}$$

where $\epsilon_{1.2t} = \epsilon_{1t} - \omega \epsilon_{2t}$ is independent of ϵ_{2t} and has variance $\sigma^2_{1.2} = \sigma^2 - \omega' \Sigma \omega$. It is seen that a regression of Y_t on X_t and ΔX_t will yield consistent estimates of β and ω. Since the distribution of X_t only depends on Σ, all information about β has been extracted by the above regression which also gives the maximum likelihood estimator. Thus the modification given by an analysis of the model, that is, the maximum likelihood estimator, is to include ΔX_t in the regression. Note that $\{X_t\}$ is still strongly exogenous if we reparametrize the model by $(\beta, \omega, \sigma^2_{1.2})$ and Σ. This serves as a justification for conditioning on the process $\{X_t\}$ when making inference about β. Due to the strong exogeneity we can condition on the process $\{X_t\}$ and then (14) just expresses a regression equation for Y_t which shows that the distribution of $\hat{\beta}$ is Gaussian if we condition on $\{X_t\}$. Thus the conditional Gaussian distribution that we meet here is again a result of a structure whereby $\hat{\beta}$ is Gaussian conditionally on the strongly exogenous or S-ancillary statistic $\{X_t\}$. It is not enough here to condition on S_{xx}, but we need also $S_{\Delta\Delta}$ and $S_{x\Delta}$ corresponding to the information in the conditional model. Note that the information in the conditional experiment depends only on the conditioning variable, and that the information is therefore deterministic in the conditional distribution rather than stochastic.

The distribution of the estimators follows from the relation

$$\begin{pmatrix} \hat{\beta} - \beta \\ \hat{\omega} - \omega \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{x\Delta} \\ S_{\Delta x} & S_{\Delta\Delta} \end{pmatrix}^{-1} \begin{pmatrix} S_{x\epsilon} \\ S_{\Delta\epsilon} \end{pmatrix},$$

which gives

$$\hat{\beta} - \beta = \left(S_{xx} - S_{x\Delta} S^{-1}_{\Delta\Delta} S_{\Delta x}\right)^{-1} \left(S_{x\epsilon_{1.2}} - S_{x\Delta} S^{-1}_{\Delta\Delta} S_{\Delta\epsilon_{1.2}}\right).$$

Here $T^{-1} S_{xx} = T^{-2} \Sigma_1^T X_t X_t' \xrightarrow{w} \int_0^1 B_2(u) B_2(u)' du$, and $S_{x\Delta} = T^{-1} \Sigma_1^T X_t \Delta X_t'$ and $S_{\Delta\Delta} = T^{-1} \Sigma_1^T \Delta X_t \Delta X_t'$ are of the order of magnitude of a constant, whereas $S_{\Delta\epsilon_{1.2}} = T^{-1} \Sigma_1^T \Delta X_t \epsilon_{1.2t}$ tends to zero, since $\Delta X_t = \epsilon_{2t}$ is independent of $\epsilon_{1.2t}$. Finally $S_{x\epsilon_{1.2}} \xrightarrow{w} \int_0^1 B_2(dB_{1.2})$. Thus we find that

$$T\left(\hat{\beta}-\beta\right)=\left(T^{-1}S_{xx}\right)^{-1}S_{x\epsilon_{1.2}}+o_P\left(1\right) \tag{15}$$

$$\xrightarrow{w}\left[\int_0^1 B_2\left(u\right)B_2\left(u\right)'du\right]^{-1}\int_0^1 B_2\left(u\right)\left(dB_{1.2}\right).$$

Here $B=\left(B_1,B_2'\right)'$ is a Brownian motion generated by ϵ_t and $B_{1.2}=B_1-\omega B_2$ with variance $\sigma_{1.2}^2=\sigma^2-\omega'\Sigma\omega$ is independent of B_2. We note that again the limit distribution is mixed Gaussian, because of the independence of the Brownian motions B_2 and $B_{1.2}$.

As a comparison consider what would happen if the regression is carried out in equation (12) without taking into account the correlation between the errors.

We would then find

$$\beta_{ols}=S_{xx}^{-1}S_{xy},$$

which is different from $\hat{\beta}$, and that

$$T\left(\beta_{ols}-\beta\right)=\left(T^{-1}S_{xx}\right)^{-1}S_{x\epsilon_1}\xrightarrow{w}\left[\int_0^1 B_2\left(u\right)B_2\left(u\right)'du\right]^{-1}\int_0^1 B_2\left(dB_1\right)$$

where B_1 is generated from ϵ_{1t}. Now in general B_1 and B_2 are dependent and hence the expectation of the limit distribution given B_2 is different from zero. This implies that the natural Wald statistic given by

$$W=\sigma^{-2}\left(\beta_{ols}-\beta\right)\sum_{t=1}^{T}X_tX_t'\left(\beta_{ols}-\beta\right),$$

will not follow a χ^2 distribution but a mixed non-central χ^2 distribution.

What has been achieved by the analysis of the model, rather than the straight forward regression, is that the estimator is modified so that the limit distribution of the estimator is mixed Gaussian and hence usual χ^2 inference is possible.

Above we have analyzed the model by finding the conditional expectation and variance of Y_t given X_t and ΔX_t in order to split up the likelihood function. A direct analysis of the likelihood function would yield

$$\begin{gathered}logL\left(\beta,\omega,\Sigma\right)=-\tfrac{1}{2}T\log\sigma_{1.2}^2-\tfrac{1}{2}T\log\left|\Sigma\right| \\ -\tfrac{1}{2}\textstyle\sum_{t=1}^{T}\left(Y_t-\beta X_t-\omega\Delta X_t\right)^2\sigma_{1.2}^{-2}-\tfrac{1}{2}\sum_{t=1}^{T}\Delta X_t'\Sigma^{-1}\Delta X_t\end{gathered}$$

where $\sigma_{1.2}^2=\sigma^2-\omega'\Sigma\omega$ with derivatives

$$\begin{array}{ll}\partial logL\left(\beta,\omega\right)/\partial\beta & =\sum_{t=1}^{T}\left(Y_t-\beta'X_t-\omega'\Delta X_t\right)X_t'\sigma_{1.2}^{-2}\,, \\ \partial logL\left(\beta,\omega\right)/\partial\omega & =\sum_{t=1}^{T}\left(Y_t-\beta'X_t-\omega'\Delta X_t\right)\Delta X_t'\sigma_{1.2}^{-2}\,, \\ -\partial^2logL\left(\beta,\omega\right)/\partial\beta^2 & =\sum_{t=1}^{T}X_tX_t'\sigma_{1.2}^{-2}\,, \\ -\partial^2logL\left(\beta,\omega\right)/\partial\omega^2 & =\sum_{t=1}^{T}\Delta X_t\Delta X_t'\sigma_{1.2}^{-2}, \\ -\partial^2logL\left(\beta,\omega\right)/\partial\beta\partial\omega & =\sum_{t=1}^{T}X_t\Delta X_t'\sigma_{1.2}^{-2}.\end{array}$$

Thus if $\sigma_{1.2}^{-2}$ and Σ were known

$$J_T(\beta,\omega) = T\begin{pmatrix} S_{xx} & S_{x\Delta} \\ S_{\Delta x} & S_{\Delta\Delta} \end{pmatrix}\sigma_{1.2}^{-2},$$

and the information in the likelihood profile or the "marginal" information about β is

$$J_T(\beta) = T\left(S_{xx} - S_{x\Delta}S_{\Delta\Delta}^{-1}S_{\Delta x}\right)\sigma_{1.2}^{-2} = TS_{xx.\Delta}\sigma_{1.2}^{-2}.$$

We can then interpret the result (14) as saying that

$$J_T(\beta)^{\frac{1}{2}}\left(\hat{\beta}-\beta\right) \xrightarrow{w} N(0,I). \tag{16}$$

Hence even in the non-ergodic case the information matrix is the proper normalization of the deviation between $\hat{\beta}$ and β. A similar formulation of the limit result can also be found in Krämer (1986), even though the general framework there does not allow the calculation of information matrices. Note that $J_T(\beta)^{-1}$ is not an estimator of $Var\left(\hat{\beta}\right)$. It is the conditional variance of $\hat{\beta}$ given the strongly exogenous variables or equivalently the information in the sample, and in fact (16) is an exact rather than a limit result since $\{Y_t\}$ given $\{X_t\}$ is Gaussian. The Wald statistic W_{obs} is calculated as

$$W_{obs} = T\sigma_{1.2}^{-1}\left(\hat{\beta}-\beta\right)' S_{xx.\Delta}\left(\hat{\beta}-\beta\right)$$

which by (15) converges to

$$\sigma_{1.2}^{-1}\int_0^1 (dB_{1.2})\,B_2'\left[\int_0^1 B_2(u)B_2'(u)du\right]^{-1}\int_0^1 B_2\,(dB_{1.2})'$$

which by the conditioning argument is χ^2 distributed, since B_2 and $B_{1.2}$ are independent. Again (16) is exactly χ^2 since it can be interpreted a the Wald statistic in the conditional model for $\{Y_t\}$ given $\{X_t\}$.

We see that the Wald statistic given by W_{obs} is the statistic that makes inference easy, in the sense that we can apply the usual χ^2 tables, whereas W_{var} and $W_{\exp}$ which are interpreted without appeal to the conditionality argument are difficult to handle.

6 The cointegration model

As the final example we consider the simple cointegration model for a $p=(1+m)$ -dimensional process $Z_t = (Y_t, X_t')'$

$$\Delta Z_t = \alpha\beta' Z_{t-1} + \epsilon_{1t} \tag{17}$$

where again $\epsilon_1, ..., \epsilon_T$ are independent $N_{m+1}(0,\Omega)$ and the (α,β) are parameters. We assume for simplicity that Ω is known and that there is only one cointegrating

relation. If we solve for Y_t, that is, let $\beta' = (1, -B')$ then the cointegrating relation is

$$Y_t = B'X_t + U_t, \tag{18}$$

where U_t is a stationary process with properties derived from the above equations, that is

$$U_t = \sum_{i=0}^{\infty} (1 + \beta'\alpha)^i \beta'\epsilon_{t-i}, \tag{19}$$

provided as will be assumed $|1+\beta'\alpha| < 1$. Thus if we consider (18) as a regression equation the regressor is correlated with the error U_t.

The likelihood analysis of the model leads to a reduced rank regression as first derived by Anderson (1951). This procedure is treated in detail by many authors, see Johansen (1988), Ahn and Reinsel (1988), Reinsel and Ahn (1990) and will not be reported here. Instead we discuss the likelihood equations and indicate how the limit distribution for $\hat{\beta}$ can be derived from the likelihood equations.

We find the derivative with respect to β to be

$$\partial logL\left(\alpha, \beta\right) / \partial\beta = \alpha'\Omega^{-1} \sum_{t=1}^{T} \left(\Delta Z_t - \alpha\beta' Z_{t-1}\right) Z'_{t-1}$$

which shows that the maximum likelihood estimator satisfies

$$\hat{\alpha}'\Omega^{-1} \sum_{t=1}^{T} \left(\Delta Z_t - \hat{\alpha}\hat{\beta}' Z_{t-1}\right) Z'_{t-1} = 0.$$

Inserting the expression for ΔZ_t from the equations (17) we find the relation

$$\hat{\alpha}'\Omega^{-1} \sum_{t=1}^{T} [(\alpha\beta' - \hat{\alpha}\hat{\beta}')Z_{t-1} + \epsilon_t] Z'_{t-1} = 0. \tag{20}$$

The above model only identifies α and β up to a constant factor. Any choice of maximum likelihood estimator $\hat{\beta}$ can be decomposed as

$$\hat{\beta} = \beta b + \alpha_\perp c,$$

with $b = (\alpha'\beta)^{-1}\alpha'\hat{\beta}$ so that we define a normalized maximum likelihood estimator

$$\tilde{\beta} = \hat{\beta} b^{-1} = \alpha_\perp c b^{-1}.$$

with the property that

$$\tilde{\beta} - \beta \in sp(\alpha_\perp).$$

The proper normalization of $\tilde{\beta} - \beta$ is by T and not $T^{\frac{1}{2}}$ as is usually the case, and if we let

$$\tilde{\beta} - \beta = \alpha_\perp T^{-1} B_T,$$

then B_T converges weakly. We shall find the limit distribution of B_T and hence that of $\tilde{\beta}$ and from this the distribution of $\hat{\beta}$ for any other normalization.

The estimator $\tilde{\alpha} = \hat{\alpha}\hat{\beta}'\alpha(\beta'\alpha)^{-1}$ is also consistent and $\tilde{\alpha} = \alpha + T^{-\frac{1}{2}}A_T$ where A_T is weakly convergent. Hence we find that

$$\alpha\beta' - \tilde{\alpha}\tilde{\beta}' = -(\tilde{\alpha} - \alpha)\beta' - \alpha(\tilde{\beta} - \beta)' - (\tilde{\alpha} - \alpha)(\tilde{\beta} - \beta)',$$

which inserted into (20) gives

$$\tilde{\alpha}'\Omega^{-1}T^{-1}\Sigma_{t=1}^{T}\epsilon_t Z_{t-1}'\alpha_\perp = (\alpha'\Omega^{-1}\alpha)B_T'(T^{-2}\Sigma_{t=1}^{T}\alpha_\perp' Z_{t-1}Z_{t-1}'\alpha_\perp) + O_P\left(T^{-\frac{1}{2}}\right).$$

From Granger's representation theorem we find

$$Z_t = \beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\alpha_\perp'\Sigma_{i=1}^{t}\epsilon_i + stationary\ process,$$

so that

$$\alpha_\perp' T^{-\frac{1}{2}} Z_{[Tu]} = \alpha_\perp' T^{-\frac{1}{2}}\Sigma_{t=1}^{[Tu]}\epsilon_i + O_P(T^{-\frac{1}{2}}) \xrightarrow{w} \alpha_\perp' W(u) = F(u),$$

and

$$\alpha'\Omega^{-1}T^{-\frac{1}{2}}\Sigma_{t=1}^{[Tu]}\epsilon_i \xrightarrow{w} \alpha'\Omega^{-1}W(u) = G(u),$$

say. This implies that in the limit we have

$$T^{-2}\Sigma_{t=1}^{T}\alpha_\perp' Z_{t-1}Z_{t-1}'\alpha_\perp \xrightarrow{w} \int_0^1 \alpha_\perp' W(u)W'(u)\alpha_\perp du = \int_0^1 F(u)F(u)'du,$$

and

$$\alpha'\Omega^{-1}T^{-1}\Sigma_{t=1}^{T}\epsilon_t Z_{t-1}'\alpha_\perp \xrightarrow{w} \alpha'\Omega^{-1}\int_0^1 (dW)W'\alpha_\perp = \int_0^1 (dG)F',$$

and hence

$$T\left(\tilde{\beta} - \beta\right) = \alpha_\perp B_T \xrightarrow{w} \alpha_\perp[\int_0^1 F(u)F(u)'du]^{-1}\int_0^1 F(dG)'(\alpha'\Omega^{-1}\alpha)^{-1}. \qquad (21)$$

This shows that the limit distribution is constructed as a mixed Gaussian distribution with the permanent shocks $F(u) = \alpha_\perp' W(u)$ as mixing distribution, and the transitory shocks $G(u) = \alpha'\Omega^{-1}W(u)$ describing the stochastic variation in the conditional limit distribution.

The information is found from

$$J_T(\beta) = -\partial^2 logL(\alpha,\beta)/\partial\beta^2 = \alpha'\Omega^{-1}\alpha\sum_{t=1}^{T} Z_{t-1}Z_{t-1}' \in O_P\left(T^2\right),$$

whereas it is seen that

$$J_T(\alpha) = -\partial^2 logL(\alpha,\beta)/\partial\alpha^2 = \Omega^{-1}\beta' \sum_{t=1}^{T} Z_{t-1}Z'_{t-1}\beta \in O_P(T),$$

$$J_T(\beta,\alpha) = -\partial^2 logL(\alpha,\beta)/\partial\beta\partial\alpha \in O_P(T).$$

This shows that the marginal information about β, which can be derived from the concentrated likelihood function, is given by

$$J_T(\beta) - J_T(\beta,\alpha) J_T(\alpha)^{-1} J_T(\alpha,\beta)$$

$$\approx J_T(\beta) = \alpha'\Omega^{-1}\alpha \sum_{t=1}^{T} Z_{t-1}Z'_{t-1}.$$

Thus the Wald statistic W_{obs} which appears as an approximation to the likelihood ratio test is approximately equal to

$$W_{obs} \approx \left(\tilde{\beta}-\beta\right)' \sum_{t=1}^{T} Z_{t-1}Z'_{t-1} \left(\tilde{\beta}-\beta\right)\left(\alpha'\Omega^{-1}\alpha\right).$$

By (21) we find that

$$W_{obs} \xrightarrow{w} \int_0^1 (dG)F'[\int_0^1 F(u)F(u)'du]^{-1} \int_0^1 F(dG)'(\alpha'\Omega^{-1}\alpha)^{-1} \qquad (22)$$

For fixed $F(u) = \alpha'_\perp W(u)$ this is just a $\chi^2(m)$ distributed since $G(u) = \alpha'\Omega^{-1}W(u)$ is independent of $F(u)$, hence also unconditionally the limit of W_{obs} is $\chi^2(m)$. Thus inference in the cointegration model concerning β involves the same conditioning argument as in the regression model with non-ergodic regressors.

Note that $T^{-2}J_T(\beta) \approx \alpha'\Omega^{-1}\alpha T^{-2}\Sigma_1^T Z_{t-1}Z'_{t-1}$ is convergent and that

$$T^{-2}EJ_T(\beta) \approx (\alpha'\Omega^{-1}\alpha)T^{-2}E\left(\Sigma_1^T Z_{t-1}Z'_{t-1}\right)$$

is convergent but not to the same value. The first converges to a random variable and the second to a constant. Note also that the asymptotic variance of $\hat{\beta}$ is not given by the inverse limit of $T^{-2}EJ_T(\beta)$, the normalized expected information. Thus again we find that in the non-ergodic case it holds that W_{obs}, W_{var} and $W_{\exp}$ behave rather differently.

This has implication for the simulation studies that are performed to study the small sample behavior of the estimator for β, see Bewley, Orden, Yang and Fisher (1993). In a given simulated set of data generated form equations (17) and (20) one should calculate $\hat{\beta}$ and $\hat{\beta}-\beta$ but also the information in the sample given by $\sum_{t=1}^{T} Z_{t-1}Z'_{t-1}$. Some samples will have a lot of information about β and others very little, thus a histogram of the calculated $\hat{\beta}$ values will be a histogram of many stochastic quantities with a varying precision. This aspect

is lost if one calculates say $\bar{\beta} = N^{-1}\sum_{i=1}^{N}\hat{\beta}_i$ and $N^{-1}\sum_{i=1}^{N}\left(\hat{\beta}_i - \bar{\beta}\right)\left(\hat{\beta}_i - \bar{\beta}\right)'$ on the basis of many simulated values $\hat{\beta}_1, ..., \hat{\beta}_N$. Instead if one is interested in a linear combination $\xi'\beta$ one should calculate the quantities

$$[\xi'(\sum_{t=1}^{T} Z_{t-1}Z'_{t-1})^{-1}\xi]^{-\frac{1}{2}}\xi'\left(\hat{\beta}_i - \beta\right),$$

which will be asymptotically Gaussian.

It is seen that again the conditioning argument in the limit distribution involves conditioning on the (continuous analogue) of the common trends $\alpha'_\perp\Sigma_{i=1}^{t}\epsilon_i$.

This can be interpreted by saying that inference on the variation around the attractor set $sp(\beta_\perp)$ as measured by $\beta'Z_t$ should be conducted conditionally on the common trends that move the process along the attractor set.

It is the purpose of this paper to investigate to what extent this idea can be made precise in the cointegration model using the notion of strong exogeneity.

One possible solution to the problem is to consider a different model where α is known, then only β is unknown and the equations take the form

$$\begin{aligned}\bar{\alpha}'\Delta Z_t &= \beta' Z_{t-1} + \bar{\alpha}'\epsilon_t,\\ \alpha'_\perp\Delta Z_t &= \alpha'_\perp\epsilon_t.\end{aligned}$$

We see that now $\alpha'_\perp\Delta Z_t = \alpha'_\perp\epsilon_t$ or $\alpha'_\perp\Sigma_{i=1}^{t}\epsilon_i$ is strongly exogenous and that β can be determined by regression of $\bar{\alpha}'\Delta Z_t$ on Z_{t-1} and $\alpha'_\perp\Delta Z_t$ like in Section 2.

If α is unknown such a precise result does not hold. Instead we shall make the following approximate argument which also works in the general cointegration model

$$\Delta Z_t = \alpha\beta' Z_{t-1} + \Sigma_{i=1}^{k-1}\Gamma_i\Delta Z_{t-i} + \epsilon_t.$$

We let $\Omega_{\alpha\alpha} = \alpha'\Omega\alpha$ and let $\omega = \alpha'\Omega\alpha_\perp(\alpha'_\perp\Omega\alpha_\perp)^{-1}$,and $\Omega_{\alpha\alpha.\alpha_\perp} = \alpha'\Omega\alpha - \alpha'\Omega\alpha_\perp(\alpha'_\perp\Omega\alpha_\perp)^{-1}\alpha'_\perp\Omega\alpha$. We denote R_{0t} and R_{1t} the residuals after regressing ΔZ_t and Z_{t-1} on $U_t = (\Delta Z_{t-1}, ..., \Delta Z_{t-k+1})$. The residuals satisfy

$$R_{0t} = \hat{\alpha}\hat{\beta}' R_{1t} + \hat{\epsilon}_t,$$

so that

$$\hat{\alpha}'_\perp R_{0t} = \hat{\alpha}'_\perp\hat{\epsilon}_t.$$

Theorem 1 *In the general cointegration model the process $\hat{\alpha}'_\perp\Sigma_{i=1}^{t}R_{0i}$ is approximately strongly exogenous in the sense that*

1.

$$T^{-\frac{1}{2}}\hat{\alpha}'_\perp\Sigma_{i=1}^{[Tu]}R_{0i} - T^{-\frac{1}{2}}\alpha'_\perp\Sigma_{i=1}^{[Tu]}\epsilon_i \xrightarrow{P} 0.$$

2. The distribution of $T^{-\frac{1}{2}}\alpha'_\perp\Sigma_{i=1}^{[Tu]}\epsilon_i$ depends only on $\Omega_{\alpha_\perp\alpha_\perp}$.

3. The distribution of $\{Z_t\}$ depends on the parameters $(\alpha, \beta, \Gamma_1, ..., \Gamma_{k-1}, \omega, \Omega_{\alpha\alpha.\alpha_\perp})$ which vary independently of the marginal variance $\alpha'_\perp\Omega\alpha_\perp$.

Proof: Let $U_t = (\Delta Z'_{t-1}, ..., \Delta Z'_{t-k+1})'$ then

$$R_{0t} = \Delta Z_t - \Sigma_{t=1}^T \Delta Z_t U'_t [\Sigma_{t=1}^T U_t U'_t]^{-1} U_t$$

and

$$T^{-\frac{1}{2}} \hat{\alpha}'_\perp \Sigma_{t=1}^{[Tu]} R_{0t} = T^{-\frac{1}{2}} \hat{\alpha}'_\perp \Sigma_{t=1}^{[Tu]} \hat{\epsilon}_t$$

$$= \hat{\alpha}'_\perp (T^{-\frac{1}{2}} \Sigma_{t=1}^{[Tu]} \epsilon_t - [T^{-1} \Sigma_{t=1}^T \epsilon_t U'_t][T^{-1} \Sigma_{t=1}^T U_t U'_t]^{-1} T^{-\frac{1}{2}} \Sigma_{t=1}^{[Tu]} U_t$$

Now $\hat{\alpha}_\perp \xrightarrow{P} \alpha_\perp$ and $T^{-\frac{1}{2}} \Sigma_{t=1}^{[Tu]} U_t$ and $T^{-1} \Sigma_{t=1}^T U_t U'_t$ are bounded in probability, whereas $T^{-1} \Sigma_{t=1}^T \epsilon_t U'_t \xrightarrow{P} 0$, so that the first statement is proved.

The distribution of $T^{-1} \Sigma_{t=1}^{[Tu]} \epsilon_t$ depends only on the covariance as indicated, and the conditional distribution has the parameters as described. Finally it is a well known result that for the Gaussian distribution the marginal variance is variation independent of the regression coefficient and the conditional variance.

This result is not terribly satisfactory since the relation between weak convergence and conditioning is not so clear. The result indicates that one can consider the common trends approximately weakly exogenous for inference on the cointegrating relations.

This means that the conditioning argument made in the asymptotic distribution in order to prove the asymptotic χ^2 distribution of W_{obs} can in some sense be considered a consequence of strong exogeneity or S-ancillarity. Note that asymptotically we only need condition on the variable $\alpha'_\perp \int_0^1 W(u) W(u)' du \alpha_\perp$ which is the weak limit of the stochastic part of the information concerning β.

Thus in the asymptotic sense described above we can say that inference concerning β should be conditional on the available information on β which is measured by the cumulated variation of the common trends.

7 Conclusion

By a few examples we have illustrated some results from inference for ergodic and non-ergodic processes. It is argued that the classical result that the inverse information measures the variance of the maximum likelihood estimator is not the correct formulation in the non-ergodic case. What holds here is that the information measures the conditional variance of the maximum likelihood estimator given the available information in the sample. Thus it is argued that an analysis of the likelihood function suggests that the information should be considered an ancillary quantity in the sense of Fisher (1934). Hence inference should be conducted conditional on the information. Thus the proper basis for inference on β is not the distribution of the estimator but the conditional distribution given the information.

8 References

Ahn, S.K. and Reinsel, G.C. (1988). 'Nested reduced-rank autoregressive models for multiple time series.' *Journal of the American Statistical Association,* vol. 83, pp. 849-856.

Anderson, T. W. (1951). 'Estimating linear restrictions on regression coefficients for multivariate normal distributions.' *Annals of Mathematical Statistics,* vol. 22, pp. 327-51.

Bartlett, M.S. (1939). 'A note on the interpretation of quasi-sufficiency.' *Biometrika,* vol. 31, pp. 391-2.

Barndorff-Nielsen, O. E. (1978). '*Information and Exponential Families in Statistical Theory.*' New York: John Wiley and Sons.

Bewley, R., Orden, D., Yang, M. and Fisher, L.A. (1993). 'Comparison of Box-Tiao and Johansen Canonical Estimator of cointegrating Vectors in VEC(1) Models.' *Journal of Econometrics* vol ? pp. ??.

Cox, R.D. and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Efron, B. and Hinkley, D.V. (1978). 'Assessing the accuracy of the maximum likelihood estimator. Observed versus expected Fisher information.' *Biometrika,* vol. 65, pp. 475-87.

Engle, R.F. and Granger, C.W.J. (1987). 'Co-integration and error correction: representation, estimation and testing.' *Econometrica,* vol. 55, pp. 251-76.

Fisher, R.A. (1925). 'Theory of statistical estimation.' *Proceedings of the Cambridge Philosophical Society,* vol. 22, pp. 700-25.

Fisher, R.A. (1934). 'Two new properties of mathematical likelihood.' *Proceedings of the Royal Society,* Series A, vol. 144, pp. 285-307.

Fisher, R.A. (1934). *Statistical Methods for Scientific Inference.* 3rd. edition, Edinburgh: Oliver and Boyd.

Hendry, D.F and Richard, J.-F. (1983). 'The econometric analysis of economic time series.' *International Statistical Review,* vol. 51, pp. 111-63.

Johansen, S. (1988). 'Statistical analysis of cointegration vectors.' *Journal of Economic Dynamics and Control,* vol.12, pp. 231-54.

Johansen, S. (1991). 'Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models.' *Econometrica,* vol. 59, pp. 1551-80.

Krämer, W. (1986). 'Least squares regression when the independent variable follows an ARIMA process.' *Journal of the American Statistical Association,* vol. 81, pp. 150-54.

Park, J.Y. (1992). 'Canonical cointegrating regressions.' *Econometrica,* vol. 60, pp. 119-43.

Phillips, P.C.B. (1991). 'Optimal inference in cointegrated systems.' *Econometrica,* vol. 59, pp. 283-306.

Phillips, P.C.B. (1994). 'Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models.' *Econometrica,* vol. 62, pp.73- 93.

Phillips, P.C.B. and Durlauf, S.N. (1986). 'Multiple Time Series Regression with Integrated Processes.' *Review of Economic Studies,* vol. 53, pp. 473-95.

Phillips, P.C.B. and Hansen, B.E. (1990). 'Statistical inference on instrumental variables regression with $I(1)$ processes.' *Review of Economic Studies,* vol. 57, pp. 99-124.

Reinsel, G.C. and Ahn, S.K. (1990).'Vector autoregressive models with unit roots and reduced rank structure, estimation, likelihood ratio test, and forecasting.' *Journal of Time Series,* vol. 13, pp. 283-95.

Preprints 1993

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK. TELEPHONE +45 35 32 08 99.

No. 1 Hansen, Henrik and Johansen, Søren: Recursive Estimation in Cointegration VAR-Models.

No. 2 Stockmarr, A. and Jacobsen, M.: Gaussian Diffusions and Autoregressive Processes: Weak Convergence and Statistical Inference.

No. 3 Nishio, Atsushi: Testing for a Unit Root against Local Alternatives

No. 4 Tjur, Tue: StatUnit - An Alternative to Statistical Packages?

No. 5 Johansen, Søren: Likelihood Based Inference for Cointegration of Non-Stationary Time Series.

Preprints 1994

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK. TELEPHONE 45 35 32 08 99, FAX 45 35 32 07 72.

No. 1 Jacobsen, Martin: Weak Convergence of Autoregressive Processes.

No. 2 Larsson, Rolf: Bartlett Corrections for Unit Root Test Statistics.

No. 3 Anthony W.F. Edwards, Anders Hald, George A. Barnard: Three Contributions to the History of Statistics.

No. 4 Tjur, Tue: Some Paradoxes Related to Sequential Situations.

No. 5 Johansen, Søren: The Role of Ancillarity in Inference for Non-Stationary Variables.