Lecture 14

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Reading: Cryer and Chen (2008): Chapter 3.3-3.4; Chapter 6.4; Chapter 11.3-11.4

MATH 4070: Regression and Time-Series Analysis

Whitney Huang
Clemson University

Notes

Agenda

1. Time Series Regression Models
2. Generalized Least Squares Regression
3. Unit Root Tests in Time Series Analysis
4. Spurious Correlation and Prewhitening

Notes

Time Series Regression

Suppose we have the following time series model for $\{Y_t\}$:

$$Y_t = m_t + \eta_t,$$

where

- m_t captures the mean of $\{Y_t\}$, i.e., $\mathbb{E}(Y_t) = m_t$
- $\{\eta_t\}$ is a zero mean stationary process with ACVF $\gamma_\eta(\cdot)$

The component $\{m_t\}$ may depend on time t, or possibly on other explanatory series

Notes

Example Models for m_t: Trends and Seasonality

- Constant trend model: For each t let $m_t = \beta_0$ for some unknown parameter β_0
- Simple linear regression: For unknown parameters β_0 and β_1,
$$m_t = \beta_0 + \beta_1 x_t,$$
where $\{x_t\}$ is some explanatory variable indexed in time (may just be a function of time or could be other series)
- Harmonic regression: For each t let
$$m_t = A\cos(2\pi\omega t + \phi),$$
where $A > 0$ is the amplitude (an unknown parameter), $\omega > 0$ is the frequency of the sinusoid (usually known), and $\phi \in (-\pi, \pi]$ is the phase (usually unknown). We can rewrite this model as
$$m_t = \beta_0 x_{1,t} + \beta_1 x_{2,t},$$
where $x_{1,t} = \cos(2\pi\omega t)$ and $x_{2,t} = \sin(2\pi\omega t)$

Notes

Multiple Linear Regression Model

Suppose there are p explanatory series $\{x_{j,t}\}_{j=1}^{p}$, the time series model for $\{Y_t\}$ is
$$Y_t = m_t + \eta_t,$$
where
$$m_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,t},$$
and $\{\eta_t\}$ is a mean zero stationary process with ACVF $\gamma_\eta(\cdot)$
We can write the linear model in matrix notation:
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta},$$
where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ is the observation vector, the coefficient vector is $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$, $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)^T$ is the error vector, and the design matrix is
$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{bmatrix}$$

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

Model Estimates & Distribution for i.i.d. Errors

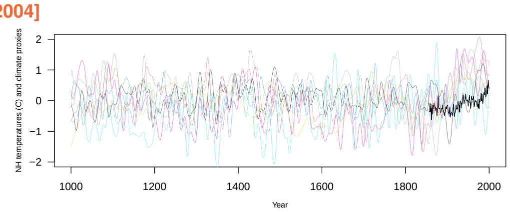
Suppose $\{\eta_t\}$ is i.i.d. $\mathrm{N}(0, \sigma^2)$. Then the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is
$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$
with
$$\hat{\sigma}^2 = \frac{\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}\right)}{n - (p+1)}$$

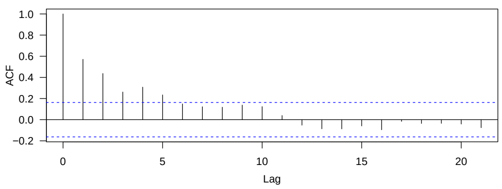
- Gauss-Markov theorem: $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$
- We have
$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1})$$
is independent of
$$\frac{(n-(p+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

Temperatures and Tree Ring Proxies [Jones & Mann, 2004]





Residuals from a linear regression fit are correlated in time $\Rightarrow$ OLS is not appropriate here ☹

Notes

Generalized Least Squares Regression

When dealing with time series the errors $\{\eta_t\}$ are typically correlated in time

- Assuming the errors $\{\eta_t\}$ are a stationary Gaussian process, consider the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta},$$

  where $\boldsymbol{\eta}$ has a multivariate normal distribution, i.e., $\boldsymbol{\eta} \sim \mathrm{N}(\boldsymbol{0}, \Sigma)$

- The generalized least squares (GLS) estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{Y},$$

  with

$$\hat{\sigma}^2 = \frac{\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)}{n - (p+1)}$$

Notes

Distributional Properties of Estimators

Gauss-Markov theorem: $\boldsymbol{\beta}_{\mathrm{GLS}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$

- We have

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}\right)^T)$$

- The variance of linear combinations of $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is less than or equal to the variance of linear combinations of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, that is:

$$\mathrm{Var}\left(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right) \le \mathrm{Var}\left(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)$$

Notes

Applying GLS in Practice

The main problem in applying GLS in practice is that Σ depends on ϕ, $\boldsymbol{\theta}$, and σ^2 and we have to estimate these

- A two-step procedure
  1. Estimate β by OLS, calculating the residuals $\hat{\boldsymbol{\eta}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, and fit an ARMA to $\hat{\eta}$ to get Σ
  2. Re-estimate β using GLS
- Alternatively, we can consider one-shot maximum likelihood methods

Notes

Likelihood-Based Regression Methods

Model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta},$$

where $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \Sigma)$

$$\Rightarrow \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \Sigma)$$

We maximum the Gaussian likelihood

$$L_n(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right]$$

with respect to the regression parameters β and ARMA parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, σ^2 simultaneously

Notes

Comparison of Two-Step and One-Step Estimation Procedures

Let's conduct a Monte Carlo simulation with the following data-generating mechanism:

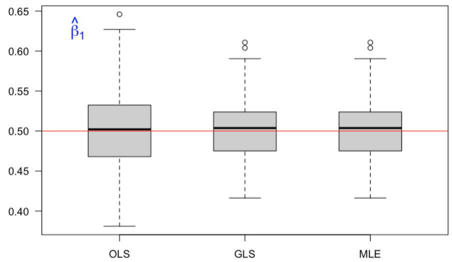
$$Y_t = 3 + 0.5x_t + \eta_t,$$

where $\eta_t = 0.8\eta_{t-1} + Z_t - 0.4Z_{t-1}$, $Z_t \sim \mathrm{N}(0, 1)$.

1. Simulate 500 replications, each with 200 data points
2. Apply the two-step procedure: fit OLS, extract residuals, estimate ARMA model for $\hat{\Sigma}$, then refit using GLS.
3. Apply the one-step procedure to jointly estimate regression and ARMA parameters
4. Compare the estimation performance

Notes

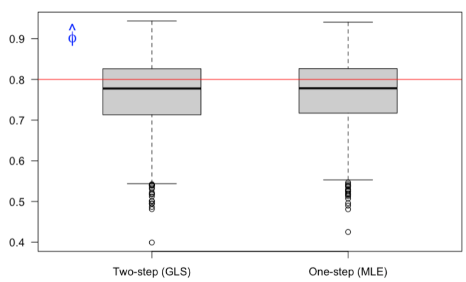
Comparing Regression Slope Estimates



Method	OLS	GLS	MLE
Bias	-4e-4	9e-4	9e-4
Sd	0.046	0.035	0.035
CI coverage	90.8%	93.6%	93.6%
CI width	0.162	0.129	0.129

Notes

Comparing ARMA Estimates



Method	GLS	MLE
Bias	-0.038	-0.036
Sd	0.090	0.089
CI coverage	96.6%	96.2%
CI width	0.330	0.328

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

An Example: Lake Huron Levels

Model:

$$Y_t = m_t + \eta_t$$

where

$m_t = \beta_0 + \beta_1 t$

$\{\eta_t\}$ is some ARMA(p, q) process

- Scientific Question: Is there evidence that the lake level has changed linearly over the years 1875-1972?
- Statistical Hypothesis:

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

Fitting Result form the Two-Step Procedure

1. OLS:

```
lm(formula = LakeHuron ~ years)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50997 -0.72726  0.00083  0.74402  2.53565

Coefficients:
               Estimate Std. Error t value
(Intercept) 625.554918   7.764293  80.568
years        -0.024201   0.004036  -5.996
```

2. AR:

```
arima(x = lm$residuals, order = c(2, 0, 0), include.mean = FALSE)

Coefficients:
         ar1      ar2
      1.0050  -0.2925
s.e.  0.0976   0.1002
```

3. Refit GLS
   Will leave it to you as an exercise

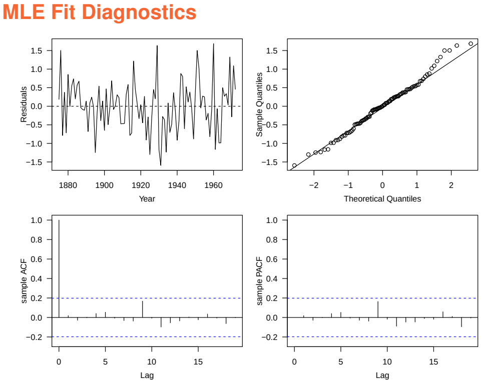
Notes

Fitting Result from One-Step MLE

```
> mle <- arima(LakeHuron, order = c(2, 0, 0),
+            xreg = cbind(rep(1,length(LakeHuron)), years),
+            include.mean = FALSE)
> mle

Call:
arima(x = LakeHuron, order = c(2, 0, 0), xreg = cbind(rep(1, length(LakeHuron)),
    years), include.mean = FALSE)

Coefficients:
         ar1      ar2  rep(1, length(LakeHuron))
      1.0048  -0.2913                   620.5115
s.e.  0.0976   0.1004                    15.5771
        years
      -0.0216
s.e.   0.0081

sigma^2 estimated as 0.4566:  log likelihood = -101.2,  aic = 212.4
```


Notes

MLE Fit Diagnostics




```
> plot.residuals(years, resid(mle), xlab = "Year", ylab = "Residuals")

        Box-Ljung test

data:  y
X-squared = 6.2088, df = 19, p-value = 0.9974
```


Notes

Comparing Confidence Intervals

Regression Slope β_1:

Method	2.5%	Point Est.	97.5%
OLS	-0.0322	-0.0242	-0.0162
MLE	-0.0374	-0.0216	-0.0057

AR ϕ_1:

Method	2.5%	Point Est.	97.5%
GLS	0.813	1.005	1.196
MLE	0.813	1.005	1.196

AR ϕ_2:

Method	2.5%	Point Est.	97.5%
GLS	-0.489	-0.293	-0.096
MLE	-0.488	-0.291	-0.095

Notes

Unit Root Tests: Tests for Non-Stationarity

Suppose we have $X_1, \cdots, X_n$ that follow the model

$$(X_t - \mu) = \phi(X_{t-1} - \mu) + Z_t,$$

where $\{Z_t\}$ is a $\mathrm{WN}(0, \sigma^2)$ process

- A unit root test considers the following hypotheses:

$$H_0 : \phi = 1 \text{ versus } H_a : |\phi| < 1$$

- Note that where $|\phi| < 1$ the process is stationary (and causal) while $\phi = 1$ leads to a nonstationary process
- Exercise: Letting $Y_t = \nabla X_t = X_t - X_{t-1}$, show that

$$\begin{aligned} Y_t &= (1-\phi)\mu + (\phi - 1)X_{t-1} + Z_t \\ &= \phi_0^* + \phi_1^* X_{t-1} + Z_t, \end{aligned}$$

where $\phi_0^* = (1-\phi)\mu$ and $\phi_1^* = (\phi - 1)$

Notes

Unit Root Tests via Ordinary Least Squares Argument

- We can estimate ϕ_0^* and ϕ_1^* using ordinary least squares
- Using the estimate of ϕ_1^*, $\hat{\phi}_1^*$, and its standard error, $\hat{\mathrm{SE}}(\hat{\phi}_1^*)$, the Dickey-Fuller statistics is

$$T = \frac{\hat{\phi}_1^*}{\hat{\mathrm{SE}}(\hat{\phi}_1^*)}$$

- Under H_0 this statistic follows a Dickey-Fuller distribution. For a level α test we reject if the observed test statistic is smaller than a critical value C_α

α	0.01	0.05	0.10
C_α	-3.43	-2.86	-2.57

- We can extend to other processes (AR(p), ARMA(p,q), and MA(q))–see Brockwell and Davis [2016, Section 6.3] for further details

Notes

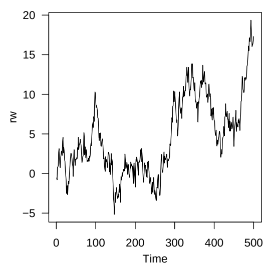
Unit Root Test: Simulated Examples

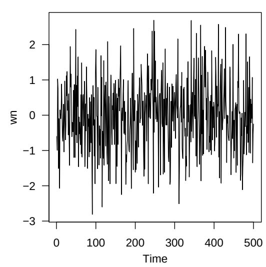
Recall

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + Z_t,$$

where $\phi_0^* = (1-\phi)\mu$ and $\phi_1^* = (\phi - 1)$

Let's demonstrate the test with a simulated random walk ($\phi = 1$) and a simulated white noise ($\phi = 0$)





Notes

Unit Root Test: Simulated Examples Cont'd

```
> diff.rw <- diff(rw); n <- length(rw)
> ys <- diff.rw; xs <- rw[1:(n-1)]
> ols.rw <- lm(ys ~ xs); summary(ols.rw)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10125    0.05973   1.695   0.0906 .
xs          -0.01438    0.00899  -1.600   0.1102


> diff.wn <- diff(wn)
> ys <- diff.wn; xs <- wn[1:(n-1)]
> ols.wn <- lm(ys ~ xs); summary(ols.wn)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001138   0.045329  -0.025     0.98
xs          -1.002420   0.044843 -22.354   <2e-16
```

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

Augmented Dickey-Fuller Test in R

Augmented Dickey-Fuller (ADF) Test: to check for the presence of a unit root in a time series and determine if the series is stationary

H_0 : The time series has a unit root (non-stationary)

H_1 : The time series is stationary

If p-value $<$ significance level (e.g., 0.05), reject $H_0 \Rightarrow$ stationary

```
> library(tseries)
> adf.test(rw)

        Augmented Dickey-Fuller Test

data:  rw
Dickey-Fuller = -1.9203, Lag order = 7, p-value =
0.612
alternative hypothesis: stationary
```

```
> adf.test(wn)
Warning in adf.test(wn) : p-value smaller than printec

        Augmented Dickey-Fuller Test

data:  wn
Dickey-Fuller = -7.8953, Lag order = 7, p-value =
0.01
alternative hypothesis: stationary
```

Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

Lagged Regression and Cross-Covariances

Consider the lagged regression model:

$$Y_t = \beta_0 + \beta_1 X_{t-d} + \varepsilon_t,$$

where X's are iid random variables with variance σ_X^2 and the ε's are also white noise with variance σ_ε^2 and are independent of the X's

The cross-covariance function of $\{Y_t\}$ and $\{X_t\}$ is

$$\gamma_{XY}(h) = \mathbb{E}\left[(X_{t+h} - \mu_X)(Y_t - \mu_Y)\right],$$

and the cross-correlation function (CCF) is

$$\rho_{XY}(h) = \frac{\gamma_{XY}(h)}{\sqrt{\gamma_X(0)\gamma_Y(0)}}.$$

If $d > 0$, we say X_t leads Y_t, and we have CCF is identically zero except for lag $h = -d$, where CCF is $\frac{\beta_1 \sigma_X}{\sqrt{\beta_1^2 \sigma_X^2 + \sigma_\varepsilon^2}}$

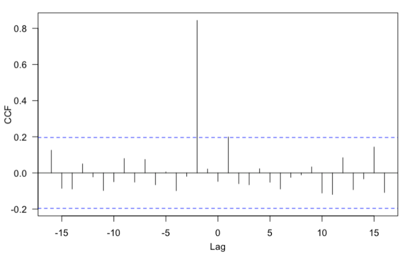
Notes

Lagged Regression and Its CCF

Consider the following reggression model:

$$Y_t = X_{t-2} + \varepsilon_t,$$

where $X_t \overset{i.i.d}{\sim} \mathrm{N}(0,1)$, $\varepsilon_t \overset{i.i.d}{\sim} \mathrm{N}(0, 0.25)$, and X's and ε's are independent to each other. The CCF is $\frac{1}{\sqrt{1+0.25}} = 0.8944$ when $h = -2$, and 0 otherwise

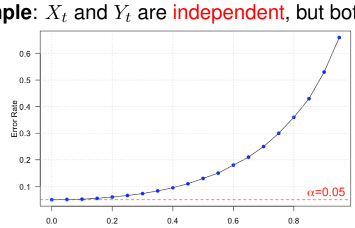


Notes

Spurious Correlations

- The lagged regression discussed earlier may be too restrictive, as X_t, Y_t, and ε_t could be temporally correlated
- Temporal dependence makes the horizon blue dashed lines $(\pm 1.96/\sqrt{n})$ unreliable
- This can lead to spurious correlations

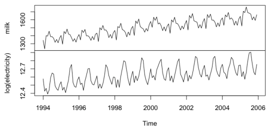
Example: X_t and Y_t are independent, but both follow an

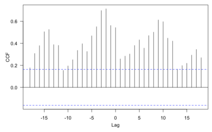


AR(1)

Notes

Spurious Correlations: An Example with Milk and Electricity Data





- **Observed Correlation**: Milk production and electricity usage show a high correlation due to shared seasonal patterns
- **Temporal Dependence**: Both series exhibit seasonality and autocorrelation, making raw correlations misleading
- **Key Takeaway**: Spurious correlations highlight the need for detrending and deseasonalizing in time series analysis

Notes

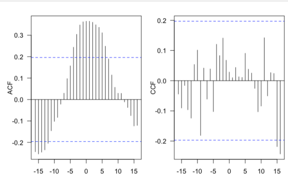
Understanding Prewhitening

Prewhitening: A technique to remove autocorrelation in a time series before analyzing cross-correlations

Steps in Prewhitening:

- Fit a time series model (e.g., ARMA) to $\{X_t\}$ and filter it to obtain residuals
- Apply the same model to $\{Y_t\}$ for consistent filtering
- Compute the cross-correlation of the residuals

```
x <- arima.sim(n = 100, list(ar = 0.9))
y <- arima.sim(n = 100, list(ar = 0.9))
par(las = 1, mgp = c(2.2, 1, 0), mar = c(3.6, 3.6, 0.8, 0.6), mfrow = c(1, 2))
ccf(x, y)
prewhiten(x, y)
```



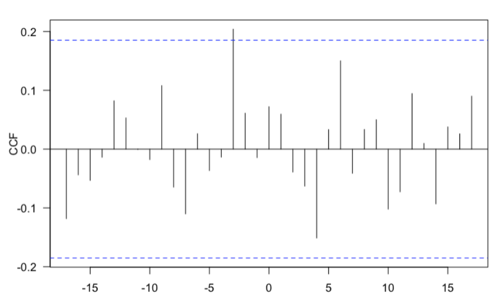
Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes

Applying Prewhitening to the Milk and Electricity Data Example

```
> me.dif = ts.intersect(diff(diff(milk, 12)),
+ diff(diff(log(electricity), 12)))
> prewhiten(as.vector(me.dif[, 1]), as.vector(me.dif[, 2]), ylab = 'CCF')
> par(las = 1, mgp = c(2.2, 1, 0), mar = c(3.6, 3.6, 0.8, 0.6))
> prewhiten(as.vector(me.dif[, 1]), as.vector(me.dif[, 2]), ylab = 'CCF')
```



Regression with Time Series Errors, Unit Root Tests, Spurious Correlations, and Prewhitening

Time Series Regression Models
Generalized Least Squares Regression
Unit Root Tests in Time Series Analysis
Spurious Correlation and Prewhitening

Notes