Unit 24: Lagged Regression

Taylor R. Brown

Department of Statistics, University of Virginia

Fall 2020

Readings for Unit 24

Textbook chapter 1.4 (page 23 to 25), 5.5.

Last Unit

1. Linear Regression with AR errors.

Motivation

We'll explore the lagged regression model: used to identify a relationship between two time series with a lagged effect.

1. Bivariate Processes

2. Lagged Regression Model

3. Worked Example

Bivariate Processes

Consider the bivariate time series $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$.
Define the following:

- $\mathrm{E}(x_t) = \mu_x, \mathrm{E}(y_t) = \mu_y$.
- $\gamma_x(h) = \mathrm{Cov}(x_t, x_{t+h}), \gamma_y(h) = \mathrm{Cov}(y_t, y_{t+h})$.

Cross-Covariance

The cross-covariance function, $\gamma_{xy}(h)$, measures the strength of the linear relationship between two variables at a certain lag. If $\{x_t\}$ and $\{y_t\}$ are jointly stationary processes, then

$$\gamma_{xy}(h) = \mathrm{E}\left[(x_{t+h} - \mu_x)(y_t - \mu_y)\right]. \tag{1}$$

Cross-Covariance

- $\gamma_{xy}(h)$: y_t is leading x_t.
- $\gamma_{xy}(-h)$: x_t is leading y_t.

Toy example: Consider x_t being the gas input and y_t the CO2 output of a furnace. The fluctuations of y_t is delayed with respect to the fluctuations of x_t due to chemical reaction time for gas to produce CO2.

Cross-Correlation

The cross-correlation function of jointly stationary $\{x_t\}$ and $\{y_t\}$ is

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}. \tag{2}$$

Properties:

- $\rho_{xy}(h) \neq \rho_{yx}(-h)$.
- $\rho_{xy}(h) = \rho_{yx}(-h)$.
- $|\rho_{xy}(h)| \leq 1$.

Joint Stationarity

Jointly stationary: constant means, autocovariances depending only on lag h, cross-covariance depends only on lag h.

Worked Example

Consider the following processes: $x_t = w_t + w_{t-1}$, $y_t = x_t - x_{t-1}$. Derive the cross-covariance function, cross-correlation function, and show that $\{x_t\}$ and $\{y_t\}$ are jointly stationary.

Sample Cross-Covariance and Sample CCF

Sample cross-covariance

$$\hat{\gamma}_{xy}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y})$$

for $h \geq 0$. The sample CCF is

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}$$

If **either** $\{x_t\}$ or $\{y_t\}$ is **white noise**, then $\hat{\rho}_{xy}(h) \sim N(0, 1/n)$. This is a special case of **Bartlett's Theorem** (Theorem A.8 in the appendix).

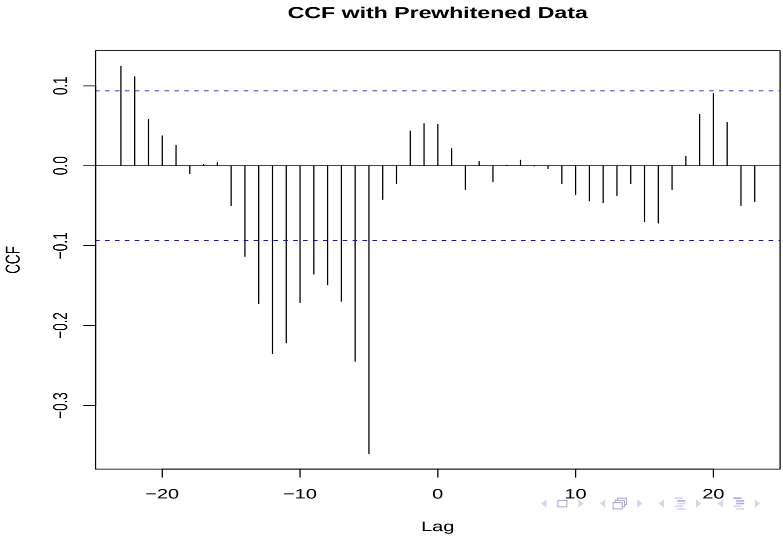
Prewhitening

If neither $\{x_t\}$ nor $\{y_t\}$ is white noise, then the sampling distribution of cross-correlation estimates are problematic, and so hypothesis tests to detect significant CCFs are unreliable. Thus, we typically **prewhiten** the data, then produce the CCF plot of the data.

- Prewhitening transforms both variables in a way that the one of the variables becomes white noise after transformation.
- We then produce a CCF plot of the data, and can reliably interpret the plot.

Sample Cross-Covariance and Sample CCF

Example: CCF of SOI and recruit data with prewhitening.



Sample Cross-Covariance and Sample CCF

Peak appears at $h = -5$, this indicates that SOI at time $t - 5$ has **strongest correlation** with recruitment at time t. SOI **leads** recruitment by 5 months. The CCF is negative, which tells us that the two time series move in opposite directions: increase in SOI is associated with a decrease in recruitment.

1. Bivariate Processes
2. Lagged Regression Model
3. Worked Example

Lagged Regression Model in Time Domain

We typically consider lagged regression models of the form

$$y_t = \sum_{j=0}^{\infty} \alpha_j x_{t-j} + \eta_t = \alpha(B) x_t + \eta_t \tag{3}$$

where $\alpha(B) = \sum_{j=0}^{\infty} \alpha_j B^j$ is the **transfer function** and η_t is a stationary ARMA noise process.

Lagged Regression Model versus Regression with ARMA errors

Unlike the model from the previous chapter, we assume both x_t and η_t each have their own process:

$$\phi(B)x_t = \theta(B)w_t$$

$$\phi_\eta(B)\eta_t = \theta_\eta(B)z_t$$

where w_t and z_t are white noise processes.

Box-Jenkins Approach

Box & Jenkins have proposed that $\alpha(B)$ in (3) can often be expressed as a ratio of polynomials involving a smaller number of coefficients, along with a specific delay, d, i.e.

$$\alpha(B) = \frac{\delta(B)B^d}{\omega(B)}, \tag{4}$$

where

- $\delta(B) = \delta_0 + \delta_1 B + \cdots + \delta_s B^s$ and
- $\omega(B) = 1 - \omega_1 B - \cdots - \omega_r B^r$.

Box-Jenkins Approach

Subbing (4) into (3), we obtain

$$y_t = \frac{\delta(B)B^d}{\omega(B)} x_t + \eta_t \tag{5}$$

Multiplying both sides by $\omega(B)$ we get

$$y_t = \sum_{k=1}^{r} \omega_k y_{t-k} + \sum_{k=0}^{s} \delta_k x_{t-d-k} + u_t. \tag{6}$$

where $u_t = \omega(B)\eta_t$.

This is a regression model with ARMA errors again!

Box-Jenkins Approach

We will fit the model via MLE, but to identify **which** model we should fit, here's a helpful derivation:

$$y_t = \alpha(B)x_t + \eta_t$$

After pre-whitening, because $\phi(B)x_t = \theta(B)w_t$

$$\begin{aligned}
\tilde{y}_t &= \frac{\phi(B)}{\theta(B)} y_t \\
&= \frac{\phi(B)}{\theta(B)} \left[\alpha(B)x_t + \eta_t\right] \\
&= \alpha(B)\frac{\phi(B)}{\theta(B)} x_t + \frac{\phi(B)}{\theta(B)}\eta_t \\
&= \alpha(B)w_t + \frac{\phi(B)}{\theta(B)}\eta_t
\end{aligned}$$

Box-Jenkins Approach (continued)

$$\tilde{y}_t = \alpha(B)w_t + \frac{\phi(B)}{\theta(B)}\eta_t$$

`TSA::prewhiten` gives us

$$\begin{aligned}
\gamma_{w,\tilde{y}}(-h) &= \operatorname{Cov}\left(w_{t-h}, \tilde{y}_t\right) \\
&= \operatorname{Cov}\left(w_{t-h}, \alpha(B)w_t + \frac{\phi(B)}{\theta(B)}\eta_t\right) \\
&= \operatorname{Cov}\left(w_{t-h}, \sum_{j=0}^{\infty} \alpha_j w_{t-j}\right) \\
&= \alpha_h \sigma_w^2
\end{aligned}$$

Sample Cross-Covariance and Sample CCF

Example: CCF of SOI and recruit data with prewhitening.

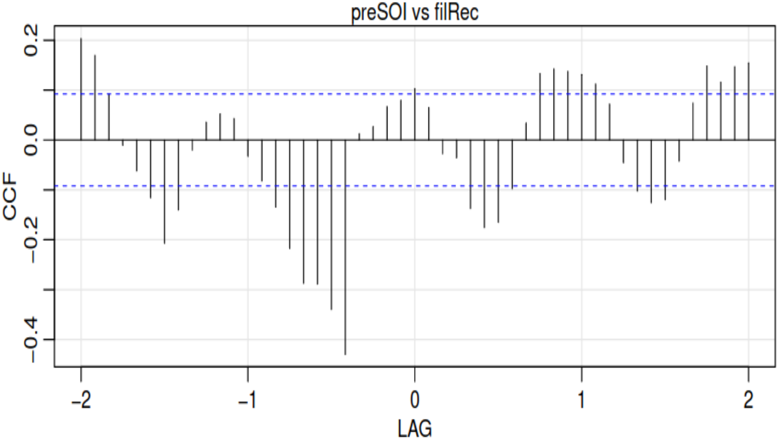


Fig. 5.11. Sample CCF of the prewhitened, detrended SOI and the similarly transformed Recruitment series; negative lags indicate that SOI leads Recruitment.

Box-Jenkins Approach

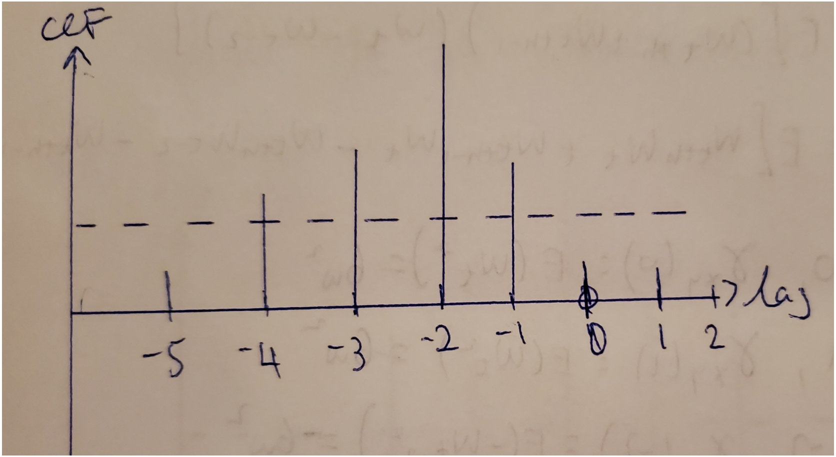
- So we perform a regression of y_t on the lagged versions of both y_t and x_t series to obtain the estimates of $\boldsymbol{\beta} = (\omega_1, \cdots, \omega_r, \delta_0, \delta_1, \cdots, \delta_s)$.
- We normally just consider u_t to be an ARMA process, and use the methods discussed in Unit 23 to estimate u_t.

Box-Jenkins Methodology for Lagged Regression

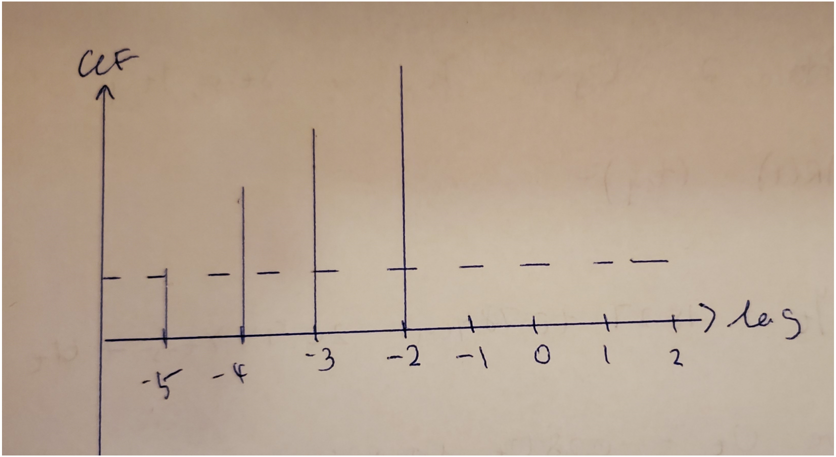
1. Fit an ARMA model for x_t, so we have estimates of $\theta_x(B)$ and $\phi_x(B)$.
2. Prewhiten the variables by applying the operator $\frac{\phi_x(B)}{\theta_x(B)}$ to both variables.
3. Compute the cross-correlation of the variables (after prewhitening) to estimate the time delay d and suggest a form for (6).
4. Obtain $\hat{\beta} = (\hat{\omega}_1, \cdots, \hat{\omega}_r, \hat{\delta}_0, \hat{\delta}_1, \cdots, \hat{\delta}_s)$ using a regression of the form in (6). Store the residuals from this regression.
5. Fit an ARMA model for the noise u_t using the residuals from the previous step and using the techniques mentioned in Unit 23.
6. Estimate the overall model using MLE (`sarima`)

Steps 1-3 are handled by `prewhiten`. (4) is handled by `lm`.

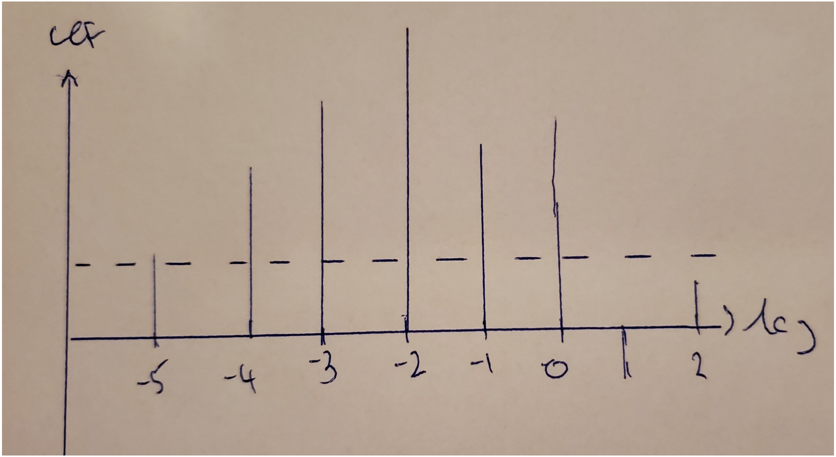
Common Patterns in CCF Plot



Common Patterns in CCF Plot



Common Patterns in CCF Plot



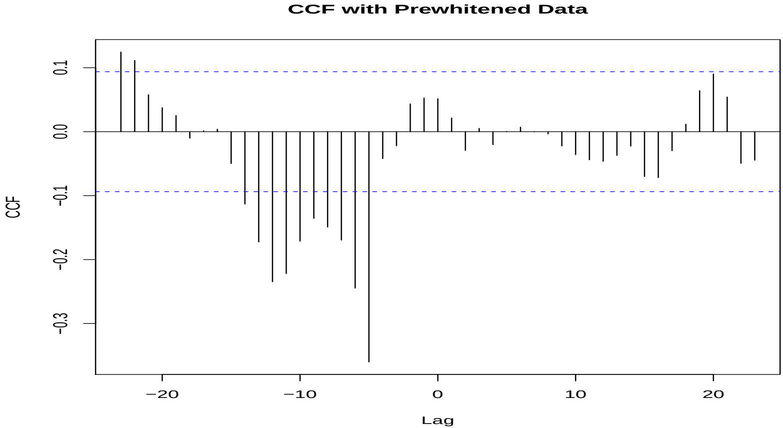
Worked Example

Some of these steps are worked out in some functions in R. What we still need to do is to examine the prewhitened CCF to determine the kind of lagged regression model we should fit (steps 1-3), and examine residuals to determine their ARMA structure (step 5).

Worked Example

We will use the Southern Oscillation Index and recruitment datasets, which contain monthly data on the changes in air pressure and estimated number of new fish in the central Pacific Ocean from 1950 to 1987. We wish to fit a lagged regression model (6) for the number of new fish against lagged versions of number of new fish and the change in air pressure in the Central Pacific Ocean.

Worked Example

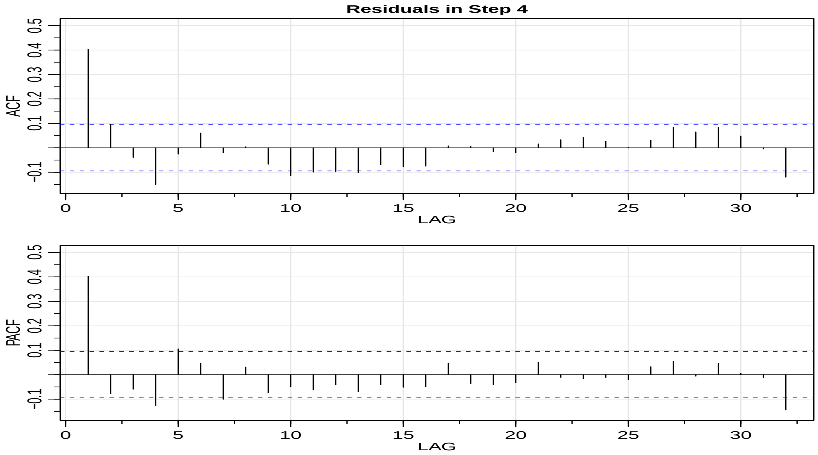


What form should (6) take?

Worked Example

After deciding the appropriate (lagged) regression, fit the model, and examine the ACF and PACF of the residuals to decide their ARMA structure.

Worked Example

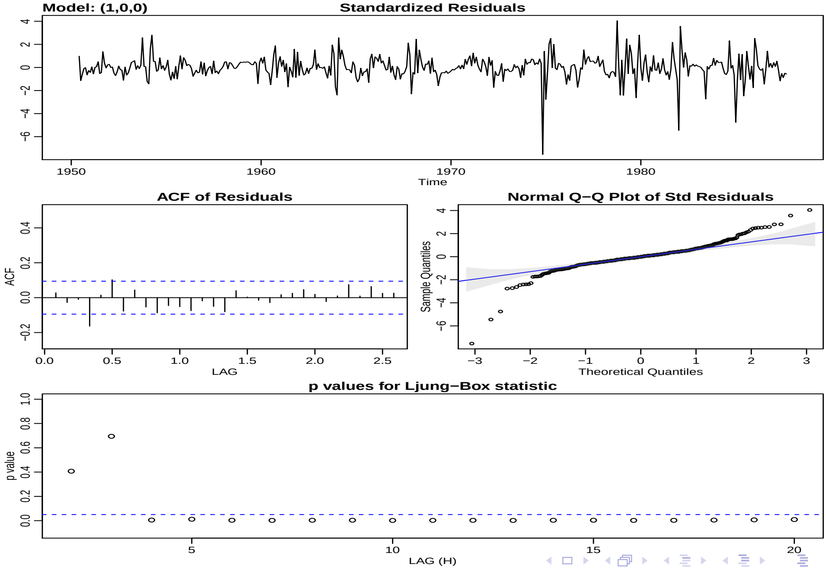


Possible structure?

Worked Example

Fit the (lagged) regression model and specify the ARMA structure of the residuals.

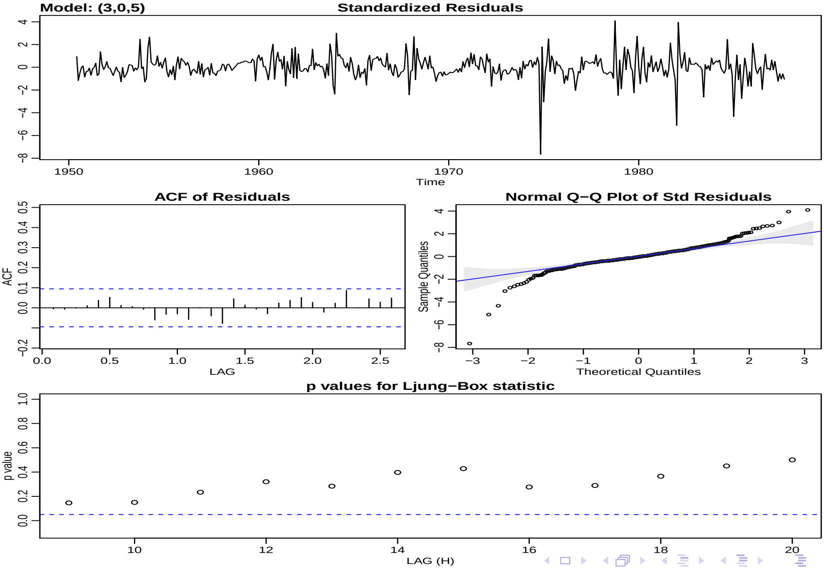
Worked Example



Worked Example

When we think we want to choose a model, make sure to examine the residuals to ensure they appear to be white. Ljung-Box statistics should be insignificant.

Worked Example



Worked Example

```
sigma^2 estimated as 46.31:  log likelihood = -1

$degrees_of_freedom
[1] 437

$ttable
              Estimate     SE  t.value p.value
ar1            -0.2826 0.1826  -1.5478  0.1224
ar2             0.5776 0.1523   3.7925  0.0002
ar3             0.4309 0.1284   3.3572  0.0009
ma1             0.7642 0.1771   4.3152  0.0000
ma2            -0.2591 0.1239  -2.0921  0.0370
ma3            -0.6097 0.1219  -5.0007  0.0000
ma4            -0.4753 0.0779  -6.0998  0.0000
ma5            -0.2947 0.0538  -5.4720  0.0000
intercept      15.2659 1.2889  11.8437  0.0000
lag(rec, -1)    0.7826 0.0199  39.3027  0.0000
lag(soi, -5)  -20.9372 1.0621 -19.7132  0.0000
```