

# Unit 24: Lagged Regression

Jeffrey Woo

Department of Statistics, University of Virginia

Spring 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  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .  
Define the following:

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

# 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

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

# 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

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

Properties:

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

# Joint Stationarity

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

Recall that the autocovariance function is symmetric. The cross-covariance function,  $\gamma_{xy}(h)$ , is not symmetric, i.e.  $\gamma_{xy}(h) \neq \gamma_{xy}(-h)$ . However,  $\gamma_{xy}(h) = \gamma_{yx}(-h)$ .

## 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  $\{x_t\}$  or  $\{y_t\}$  is \_\_\_\_\_, then  $\hat{\rho}_{xy}(h) \sim N(0, 1/n)$ .

# Prewhitening

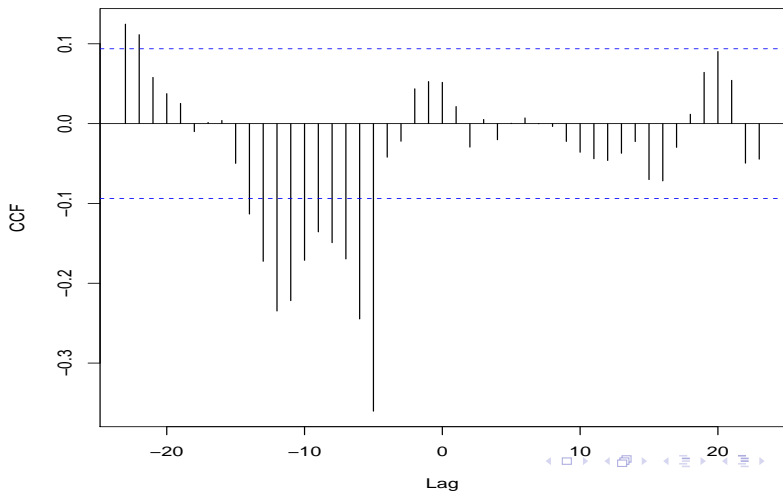
If neither  $\{x_t\}$  nor  $\{y_t\}$  is white noise, then hypothesis tests to detect significant CCFs are unreliable. Thus, we typically \_\_\_\_\_ the data, then produce the CCF plot of the data.

- 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.

**CCF with Prewhitened Data**



# Sample Cross-Covariance and Sample CCF

Peak appears at  $h = -5$ , this indicates that SOI at time  $t - 5$  has \_\_\_\_\_ with recruitment at time  $t$ . SOI \_\_\_\_\_ 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

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

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

# 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.

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

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

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

# Box-Jenkins Approach

Expanding the backshifts in  $\delta(B)$  and  $\omega(B)$  in (5), we obtain

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

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

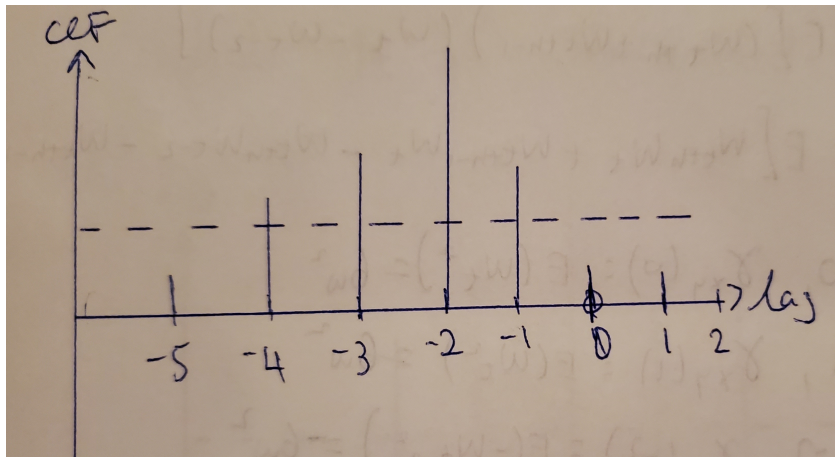
# Box-Jenkins Approach

- 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  $\beta = (\omega_1, \dots, \omega_r, \delta_0, \delta_1, \dots, \delta_s)$ .
- We normally just consider  $u_t$  to be an ARMA process, and use the methods discussed in Unit 23 to estimate  $u_t$ .

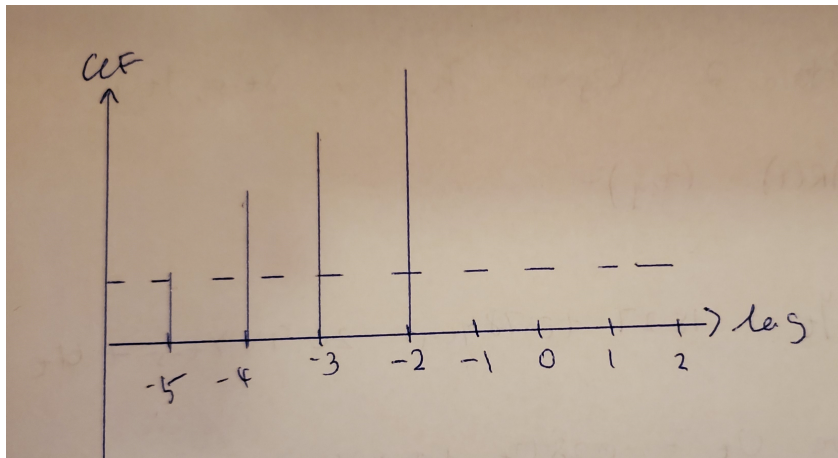
# Box-Jenkins Methodology for Lagged Regression

- 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  $\hat{\beta} = (\hat{\omega}_1, \dots, \hat{\omega}_r, \hat{\delta}_0, \hat{\delta}_1, \dots, \hat{\delta}_s)$  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.

# Common Patterns in CCF Plot

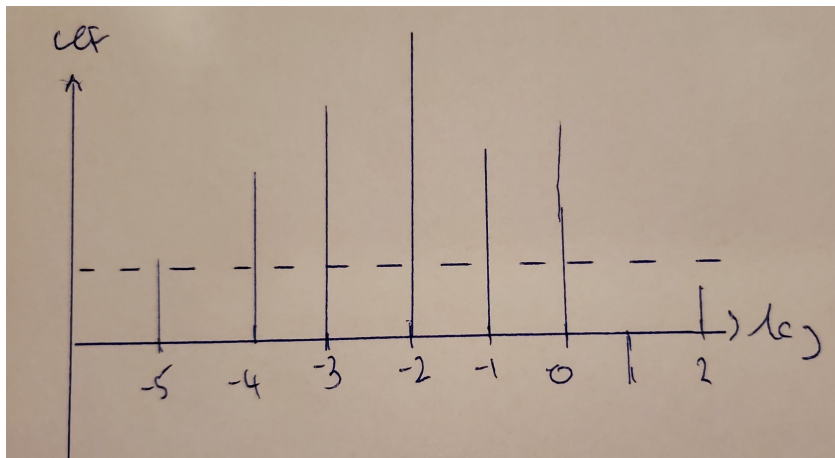


# Common Patterns in CCF Plot





# Common Patterns in CCF Plot



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

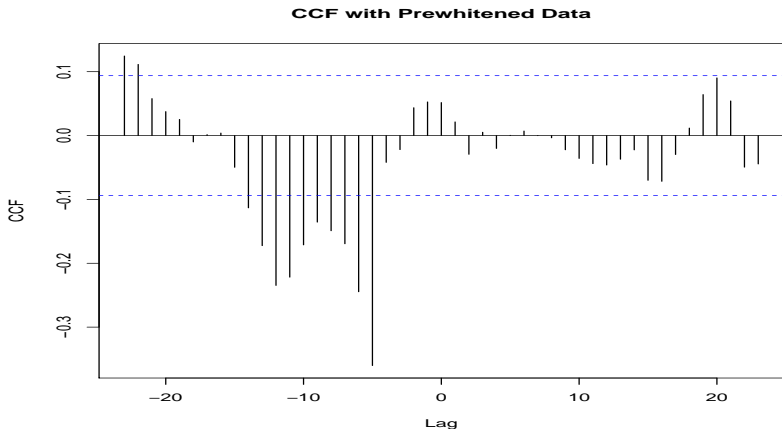
## Worked Example

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 (step 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

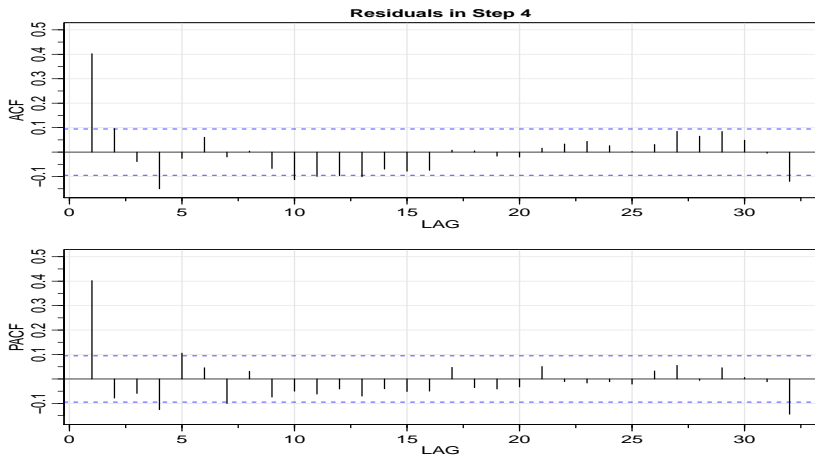


What form should (6) take?

## 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



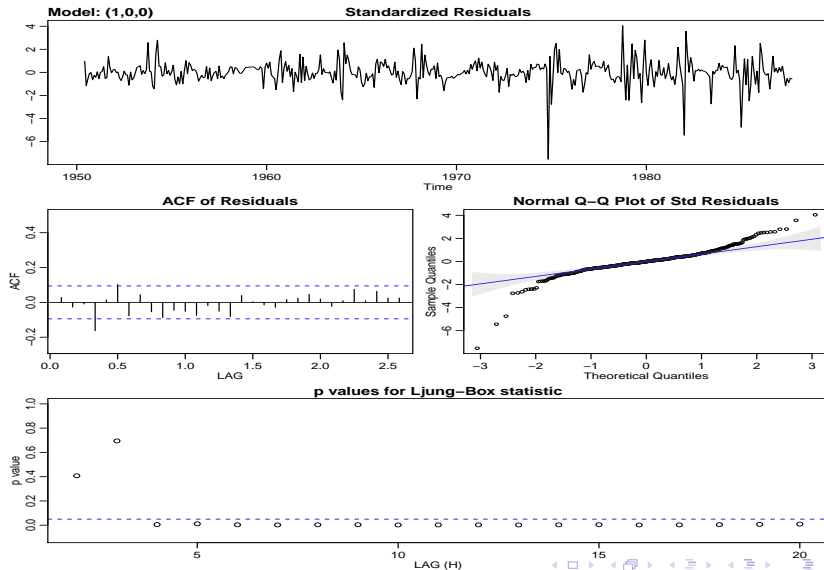
Possible structure?

# Worked Example

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



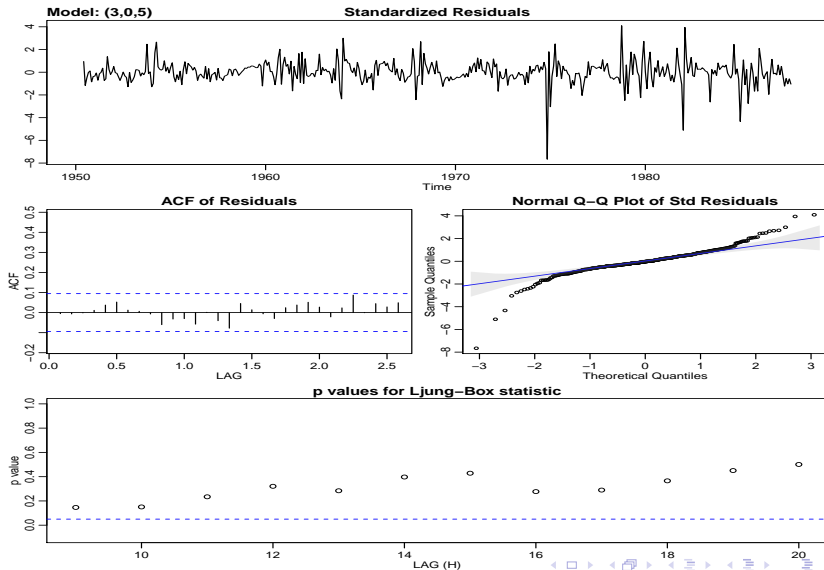
# Worked Example



## 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 = -14
$degrees_of_freedom
[1] 437

$tttable
```

	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