Unit 24: Lagged Regression

Jeffrey Woo

Department of Statistics, University of Virginia

Spring 2016

Readings for Unit 24

Textbook chapter 1.4, 5.7.

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.

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:

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

Cross-Covariance

The cross-covariance function of two jointly stationary processes $\{x_t\}$ and $\{y_t\}$ is

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

Joint Stationarity

Jointly stationary: constant means, autocovariances depending only on lag h, cross-covariance depends only on 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, $\gamma_{xy}(h) = \gamma_{yx}(-h)$.

Cross-Covariance

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

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:

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

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

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

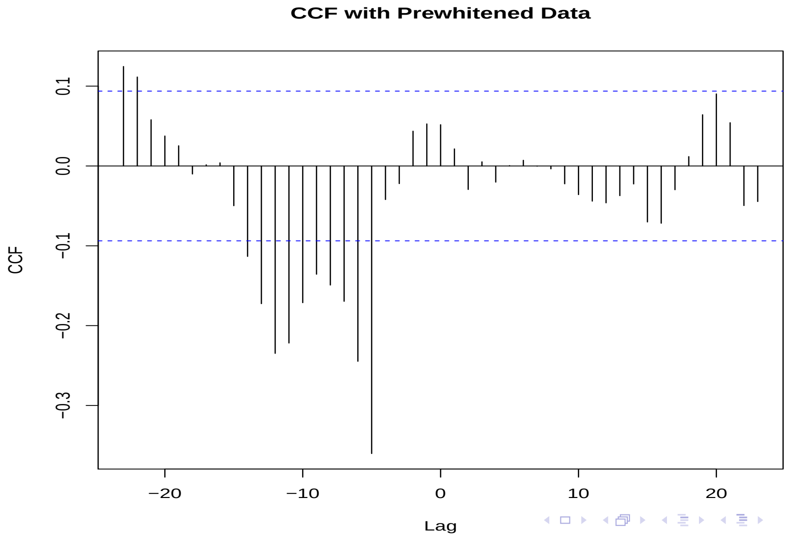
for $h \geq 0$. The sample CCF is

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

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

Sample Cross-Covariance and Sample CCF

Example: CCF of SOI and recruit data.



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.

Lagged Regression Model in Time Domain

We typically consider lagged regression models of the form

$$y_t = \sum_{k=1}^{r} \omega_k y_{t-k} + \sum_{k=0}^{s} \delta_k x_{t-d-k} + u_t. \tag{3}$$

where u_t is a stationary ARMA noise process. So we perform a regression on the lagged versions of both the input and output series to obtain the estimates of $\boldsymbol{\beta} = (\omega_1, \cdots, \omega_r, \delta_0, \delta_1, \cdots, \delta_s)$.

Box-Jenkins Approach

Due to the large number of parameters we are fitting, the following sequential methodology has been developed. **Step 1**: we fit an ARMA model for the input x_t, so we have estimates of $\theta_x(B)$ and $\phi_x(B)$.

Box-Jenkins Approach

Step 2: prewhiten the input and output series by applying the inverse operator $\frac{\phi_x(B)}{\theta_x(B)}$ to the input and output series

Prewhitening

Recall from slide 13 that we need either the input or the output series to be ___________ so we know the theoretical variance of the sample CCF is $1/n$. Thus we prewhiten the input series (and output) so we can study the CCF between the the prewhitened input and output series. Since prewhitening is a linear operation, any linear relationships will be preserved. Note that the operator $\frac{\phi_x(B)}{\theta_x(B)}$ is tailor-made to transform the input to a white noise, not the output.

Box-Jenkins Approach

Step 3: Compute the cross-correlation of $\widetilde{y}_t$, the output series after prewhitening, with w_t, $\gamma_{\widetilde{y}w}(h)$ to estimate the time delay d and suggest a form for (3).

Box-Jenkins Approach

Step 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 (3).

Box-Jenkins Approach

Step 5: Fit an ARMA model for the noise u_t.

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

Worked Example

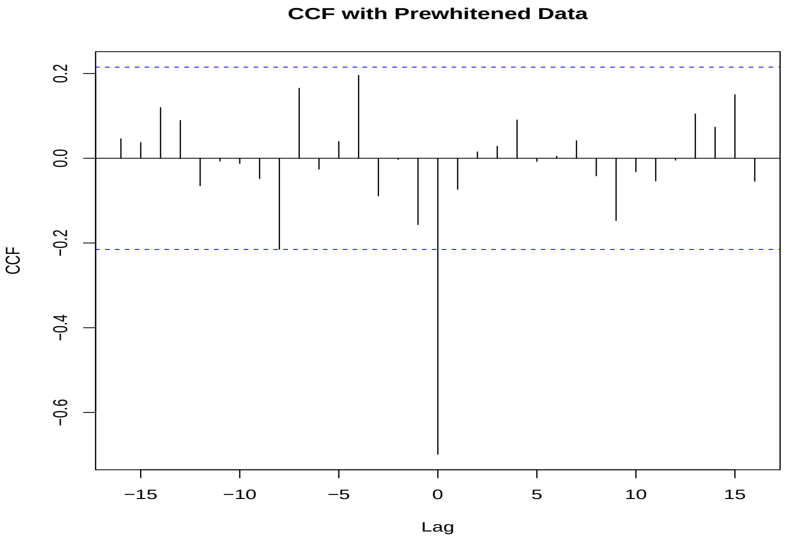
Some of these steps are worked out in some functions in R. What we still need to do is to examine the prewhitened CCF to determine the kind of lagged regression model we should fit, and examine residuals to determine their ARMA structure.

For this worked example, we examine the (log-transformed) sales and price of a certain potato chip from Bluebird Foods. The first step would be to transform the time series to obtain stationarity, and then examine the CCF for the prewhitened data. For this dataset, we take the first difference of both time series to obtain stationarity, and examine the CCF.

Worked Example

Some common patterns of CCF to look out for.

Worked Example

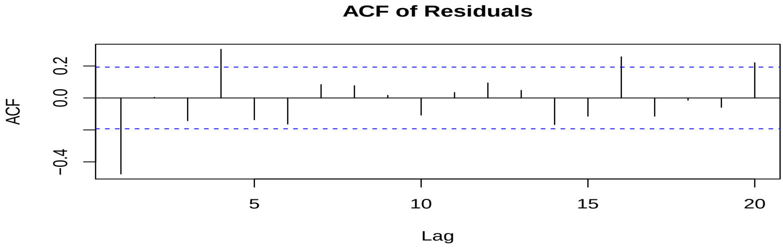


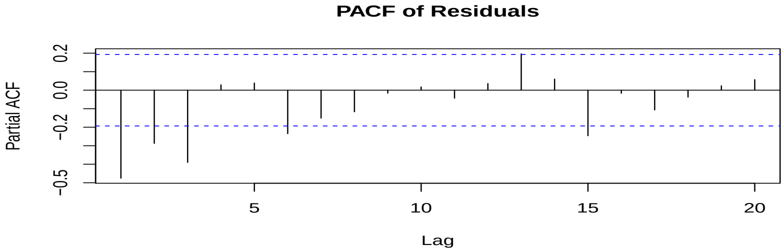
What should we regress on?

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

```
Call:
arima(x = dy, order = c(3, 0, 4), xreg = data.frame(dx))

Coefficients:
          ar1      ar2      ar3     ma1      ma2      ma3     ma4  intercept
      -1.0465  -0.7252  -0.0315  0.2559  -0.0707  -0.7453  0.2096    -0.0009
s.e.   0.3617   0.4148   0.2862  0.3519   0.1730   0.0933  0.2977     0.0037
           dx
      -2.5797
s.e.   0.1215

sigma^2 estimated as 0.02502:  log likelihood = 40.93,  aic = -63.85
```

Any comments?

Worked Example

```
Call:
arima(x = dy, order = c(2, 0, 3), xreg = data.frame(dx))

Coefficients:
          ar1      ar2      ma1     ma2      ma3  intercept       dx
      -0.0125  -0.9565  -0.7926  0.8786  -0.6680    -0.0010  -2.4473
s.e.   0.0879   0.0687   0.1665  0.1144   0.1138     0.0036   0.1293

sigma^2 estimated as 0.02693:  log likelihood = 39.12,  aic = -64.25
```

Any comments?

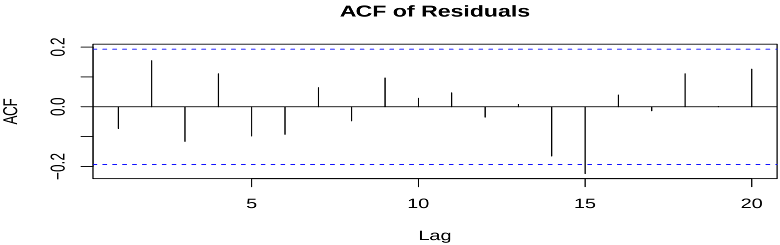
Worked Example

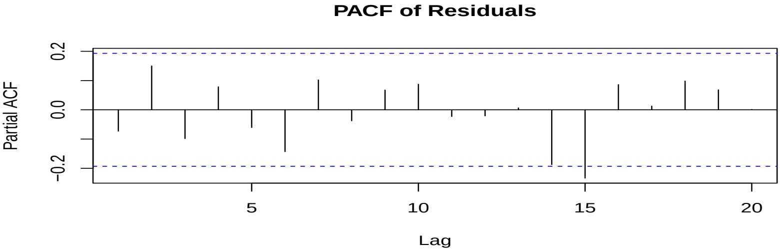
```
Call:
arima(x = dy, order = c(2, 0, 3), xreg = data.frame(dx), fixed = c(0, NA, NA,
    NA, NA, 0, NA))

Coefficients:
      ar1      ar2      ma1     ma2      ma3  intercept       dx
        0  -0.9488  -0.8129  0.8901  -0.6578          0  -2.4510
s.e.    0   0.0641   0.0855  0.0776   0.1109          0   0.1282

sigma^2 estimated as 0.02697:  log likelihood = 39.07,  aic = -68.15
```

Worked Example

When we think we want to choose a model, make sure to examine the residuals to ensure they appear to be white.

Worked Example





Any comments?