Multivariate Regression

Setting and Notation

Univariate modeling

In previous units, a realization of length n (say on the variable "monthly sales") was denoted by:

$$X_t = X_1, X_2, \ldots, X_n$$

Our analyses in previous units have shown

- How to use an observed realization of length n to model the correlation structure within the time series on sales
  - Using AR, ARMA, ARIMA, seasonal, etc. models
- How to predict future sales: X_{n+1}, X_{n+2}, ...

This is called a ***univariate (one-variable)*** approach

Multivariate modeling

In this unit we will consider ***multivariate*** approaches that use ***more than one time-series variable*** to better understand data, make decisions (forecasts, etc.)

Multiple Regression with Correlated Errors

The first technique we will discuss is a direct extension of the multiple regression model. In the time series setting, each of the variables (independent and dependent) depend on time and occur as realizations of the same length.

- In the case in which each of the dependent and independent variables depend on time, it is common for the errors to be auto-correlated
  - This was true when testing for trend in the temperature data
    - In which case the only independent variables was time
  - We used the Cochrane-Orcutt procedure
  - Another option is an MLE approach that we will illustrate in the next example

Multiple Regression with Correlated Errors

Notation

We will denote the multiple time series regression model by

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_m X_{tm}, + Z_t$$

where Z_t may satisfy an AR(p) process

- Each realization is of length *n.* In the equation above

  $Y_t = Y_1,\ Y_2,\ \ldots\ ,\ Y_n$ and $Z_t = Z_1, Z_2, \cdots\ ,\ Z_n$

- We refer to the corresponding *m* independent variables

  $X_{t1} = X_{11},\ X_{21},\ \ldots\ ,\ X_{n1}$

  $X_{t2} = X_{12},\ X_{22},\ \ldots\ ,\ X_{n2}$

  …

  $X_{tm} = X_{1m},\ X_{2m},\ \ldots\ ,\ X_{nm}$

Note that for the independent variables:

- The first subscript is time $(1, 2, \ldots, n)$
- The second subscript indicates which variable it is

DataScience@SMU

Multivariate Regression

R-Based Analysis

Multiple Regression with Correlated Errors

R-based analysis strategy

In the following R code, **y** denotes the realization Y_t, **z** denotes Z_t, and **x1**, … , **xm** denote $X_{t1}, \ldots, X_{tm}$.

Step 1: Perform a regression analysis and model the residuals. The following is example code (with **m**=3):

```
ksfit=lm(y~x1+x2+x3)
phi=aic.wge(ksfit$residuals,p=0:8)
```

Step 2: Use function ARIMA to perform the MLE analysis which estimates the coefficients in the multiple regression while simultaneously modeling Z_t as an AR(phi$p)

```
fit=arima(sales,order=c(phi$p,0,0)xreg=cbind(x1,x2,x3))
```

fit$coef contains the AR coefficients, the constant, and the coefficients on **x1**, **x2**, and **x3**

Multiple Regression with Correlated Errors

R-based analysis strategy

The command

`fit`

produces the following example "dummy" output (assuming `phi$p`=2)

```
Coefficients:
      ar1       ar2   intercept      x1      x2      x3
      1.6      -0.8      10.5        2.3     3.1     0.3
s.e.  0.6       0.3       2.2        0.9     1.2     0.2
sigma2 estimated as 1.4: log likelihood=-16,aic=20.1
```

- Recall that we don't usually look at the SE's for AR and MA coefficients (the factor table gives more information).
- Function `arima` doesn't give p-values, but in general if the absolute value of the coefficient is over two times the SE, this is evidence at the .05 level that the variable is useful.
- The final model residuals are given in `fit$resid`, and they should be white
  - Check with residual plots and/or Ljung-Box Test.
- Compare competing models with AIC/AIC/BIC etc.
- Use the model to forecast and/or answer any additional QOIs.

DataScience@SMU

Multivariate Regression: Example

Sales Data

Multiple Regression with Correlated Errors

Example

We are interested identifying variables that impact sales (Y) (variable sales in the R code on the next slide).

Variables we consider are (data in file BusinessSales.csv)

- TV advertising expenditures (X_1) (variable ad_tv)
- Online advertising expenditures (X_2) (ad_online)
- Discount on product (X_3) (variable discount)

We have data for the past 100 weeks. That is, for each variable, we have a time series realization of length $n = 100$ (weeks):

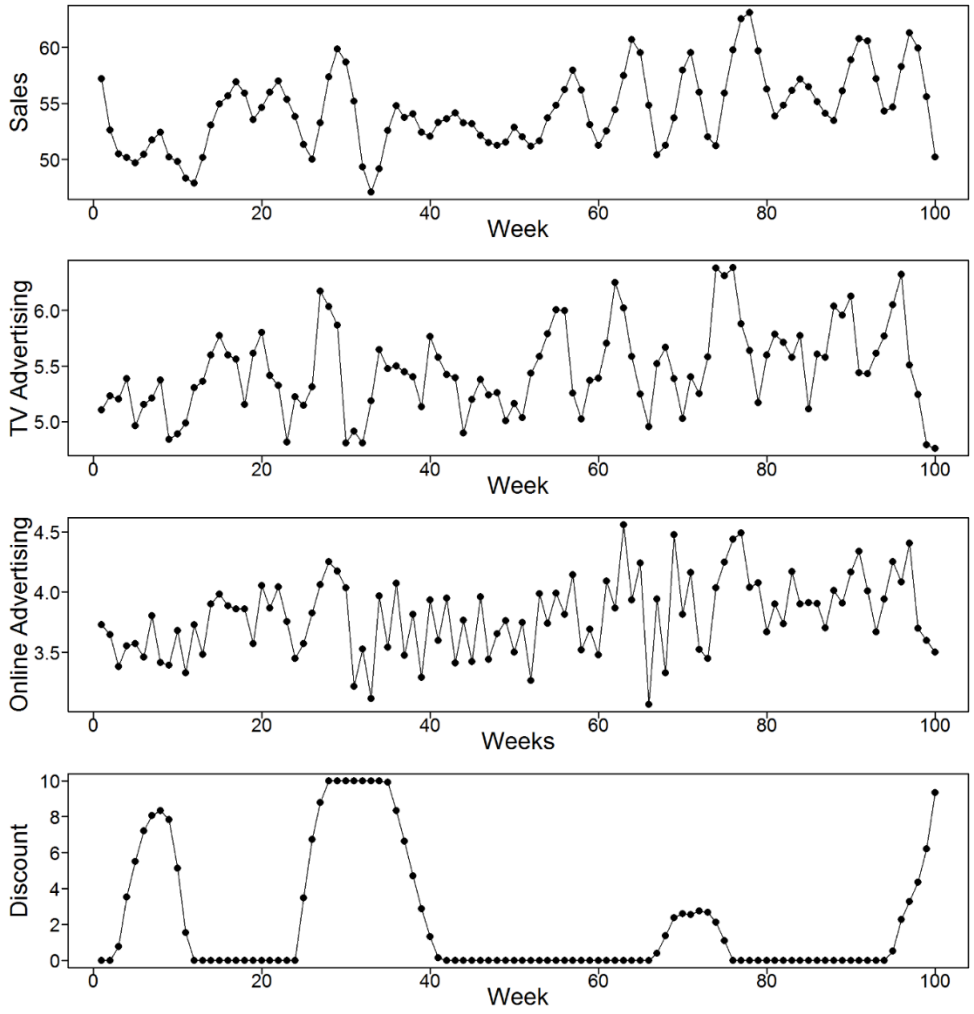
sales: $Y_t = Y_1, Y_2, \ldots, Y_{100}$ (x \$10,000)

ad_tv: $X_{t1} = X_{11}, X_{21}, \ldots, X_{100,1}$ (x \$10,000)

ad_online: $X_{t2} = X_{12}, X_{22}, \ldots, X_{100,2}$ (x \$10,000)

discount: $X_{t3} = X_{13}, X_{23}, \ldots, X_{100,3}$ (% discount)

Hypothetical Sales Data



Multiple Regression with Correlated Errors

R code

```
#Assuming a data.frame exists with corresponding names below.
ksfit=lm(sales~ad_tv+ad_online+discount, data = BSales)
aic.wge(ksfit$residuals,p=0:8, q=0)  # AIC picks p=7
fit=arima(Bsales$sales,order=c(7,0,0),xreg=BSales[,3:5])
fit
```

Output time series multiple regression analysis

```
Coefficients:
      ar1      ar2     ar3     ar4     ar5     ar6      ar7
    1.4734  -0.8921  0.0749  0.0919  0.0438  0.1865  -0.1287
s.e.0.1107   0.1963  0.2167  0.2057  0.2049  0.1852   0.1080

intercept    ad_tv   ad_online   discount
54.5513     0.0703    -0.0934    -0.1514
 2.2040     0.3434     0.2075     0.1315

sigma^2 estimated as 1.411: log likelihood=-161.1, aic=346.21
```

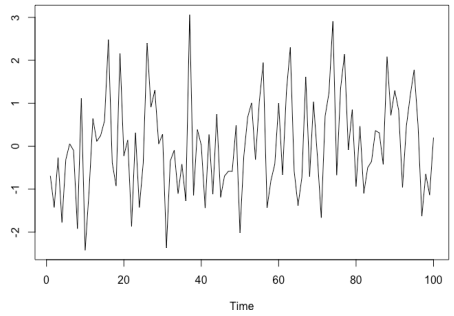
Multiple regression equation

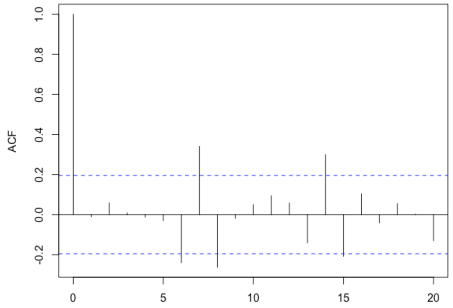
$$sales = 54.55 + .07(ad_{tv}) - .09(online_{ad}) - .15(discount)$$

Comments about Regression Equation

Results:

- The errors in the standard multiple regression satisfy an AR(7) and are correlated
- None of the variables ad_tv, ad_online, or discount are significantly different from zero
- The final model residuals are not white (fit$resid)





acf(fit$resid)

ltest = ljung.wge(fit$resid)

ltest$pval

```
> ltest$pval
[1] 7.445469e-06
```

There is strong evidence that the residuals are serially correlated.

Summary:

- These results are not very enlightening (or encouraging)
- Let's try adding a trend term to the model

Multiple Regression with Correlated Errors

Add a trend term (t=week) to the model

```
t=1:100
ksfit=lm(sales~t+ad_tv+ad_online+discount, data = BSales)
aic.wge(ksfit$residuals,p=0:8,q=0:0)   # AIC picks p=6
fit=arima(BSales$sales,order=c(6,0,0),xreg=cbind(t,BSales[,3:5]))
fit
```

Output time series multiple regression analysis

```
Coefficients:
         ar1      ar2     ar3     ar4     ar5      ar6
      1.4090  -0.8778  0.0180  0.0562  0.1676  -0.0364
s.e.  0.1131   0.1978  0.2182  0.2057  0.1785   0.1050

     intercept       t     ad_tv   ad_online  discount
       51.9224  0.0465  0.1123     -0.0508   -0.1701
        2.2242  0.0148  0.3549      0.1939    0.1052

sigma^2 estimated as 1.363:  log likelihood = -159.14,  aic = 342.29
```

time (week) is significant, other variables are not

Multiple regression equation

$$sales = 51.92 + .05(week) + .11(ad_{tv}) - .05(online_{ad}) - .17(discount)$$

Comments about Regression Equation with Trend Included

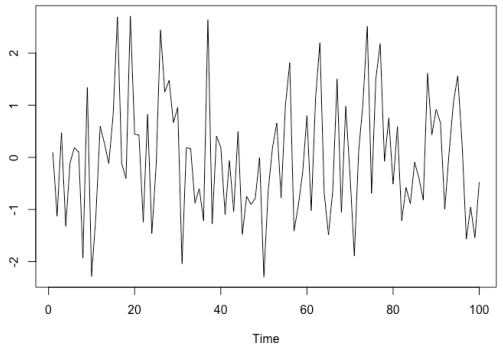
Results:

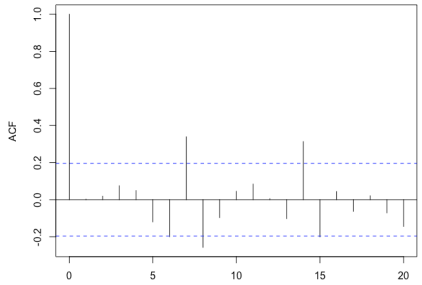
- Again the errors in the standard multiple regression satisfy an AR(6) and are correlated
- Time was significant (with positive slope) so it seems that sales are increasing with time
- None of the variables ad_tv, ad_online, or discount are significantly different from zero
- The final model residuals are not white

```
ltest = ljung.wge(fit$resid)
ltest$pval
```

```
> ltest$pval
[1] 1.66855e-05
```

There is strong evidence that the residuals are serially correlated.





acf(fit$resid)

Summary:

- Let's keep working on this

DataScience@SMU

Multivariate Regression

Lagged Variables

How About Using Lagged Variables?

Advertising for the week didn't seem to predict sales for that week

- It could be that advertising has a "lag" effect
- That is, advertising for the current week may have more effect on sales next week than this week

We next considered a multiple regression in which the "advertising variables at time *t*" are actually costs for the previous week

- These will be denoted ad_tv1 and ad_online1
- The new variables will be

    ad_tv1[t]=ad_tv[t-1]

    ad_online1[t]=ad_online[t-1]

Creating Lagged Variables in R

```
#Example:
#With dplyr lag function
df = data.frame(Y = c(1,1,2,3,4,4,5,8),X1 = c(5,6,6,7,7,8,8,9))
df$X1_L1 = dplyr::lag(df$X1,1)
df$X1_L2 = dplyr::lag(df$X1,2)
df
```

```
> df
  Y X1 X1_L1 X1_L2
1 1  5    NA    NA
2 1  6     5    NA
3 2  6     6     5
4 3  7     6     6
5 4  7     7     6
6 4  8     7     7
7 5  8     8     7
8 8  9     8     8
```

DataScience@SMU

Multivariate Regression: Example Part 2

Using Lagged Variables

Multiple Regression with Correlated Errors

```
ad_tv1 = dplyr::lag(BSales$ad_tv,1)
ad_online1 = dplyr::lag(BSales$ad_online,1)
discount = BSales$discount
BSales$ad_tv1= ad_tv1
BSales$ad_online1 = ad_online1
ksfit=lm(sales~ad_tv1+ad_online1+discount, data = BSales)
aic.wge(ksfit$residuals,p=0:8,q=0:0)  # AIC picks p=7
fit=arima(BSales$sales,order=c(7,0,0),xreg=cbind(ad_tv1, ad_online1, discount))
fit
```

With Lagged "ad" variables

$Sales_t = \beta_0 + \beta_1 ad\_tv_{t-1} + \beta_2 ad\_online_{t-1} + \beta_3 discount_{t,} + Z_t$ Z_t is AR(7)

```
Coefficients:
          ar1     ar2      ar3      ar4      ar5      ar6     ar7
      -0.5197  0.1348  -0.0787  -0.1155  -0.1954  -0.0004  0.0485
s.e.   0.1839  0.1689   0.1233   0.1256   0.1309   0.1884  0.1574

      intercept  ad_tv1  ad_online1  discount
         4.8382  3.4341      8.1152   -0.0573
s.e.     2.8270  0.6166      1.2447    0.0281

sigma^2 estimated as 1.642:  log likelihood = -165.43,  aic = 354.86
```

both "ad" variables are highly significant

Multiple Regression with Correlated Errors

```
t=1:100
ad_tv1 = dplyr::lag(BSales$ad_tv,1)
ad_online1 = dplyr::lag(BSales$ad_online,1)
BSales$ad_tv1= ad_tv1
BSales$ad_online1 = ad_online1

ksfit=lm(sales~t + ad_tv1+ad_online1+discount, data = BSales)
aic.wge(ksfit$residuals,p=0:8,q=0:0)  # AIC picks p=7
fit=arima(Bsales$sales,order=c(7,0,0),xreg=cbind(t, ad_tv1, ad_online1, discount))
fit
```

Lagged "ad" variables and trend

$Sales_t = \beta_0 + \beta_1 t + \beta_2 ad\_tv_{t-1} + \beta_3 ad\_online_{t-1} + \beta_4 discount_{t,} + Z_t$ Z_t is AR(7)

```
Coefficients:
          ar1     ar2      ar3      ar4      ar5     ar6     ar7
      -0.5062  0.1265  -0.1053  -0.1388  -0.1985  0.0109  0.0493
s.e.   0.1883  0.1537   0.1138   0.1160   0.1206  0.1805  0.1535

      intercept       t  ad_tv1  ad_online1  discount
         6.2215  0.0065  3.3180      7.8248   -0.0453
s.e.     2.7820  0.0038  0.6288      1.3020    0.0276

sigma^2 estimated as 1.577:  log likelihood = -163.43,  aic = 352.87
```

both "ad" variables are highly significant

Multiple Regression with Correlated Errors

Lagged "ad" variables

```
Coefficients:
          ar1     ar2      ar3      ar4      ar5      ar6     ar7
      -0.5197  0.1348  -0.0787  -0.1155  -0.1954  -0.0004  0.0485
s.e.   0.1839  0.1689   0.1233   0.1256   0.1309   0.1884  0.1574

      intercept  ad_tv1  ad_online1  discount
         4.8382  3.4341      8.1152   -0.0573
s.e.     2.8270  0.6166      1.2447    0.0281

sigma^2 estimated as 1.642:  log likelihood = -165.43,  aic = 354.86
```

both "ad" variables are highly significant

Lagged "ad" variables and trend

```
Coefficients:
          ar1     ar2      ar3      ar4      ar5     ar6     ar7
      -0.5062  0.1265  -0.1053  -0.1388  -0.1985  0.0109  0.0493
s.e.   0.1883  0.1537   0.1138   0.1160   0.1206  0.1805  0.1535

      intercept       t  ad_tv1  ad_online1  discount
         6.2215  0.0065  3.3180      7.8248   -0.0453
s.e.     2.7820  0.0038  0.6288      1.3020    0.0276

sigma^2 estimated as 1.577:  log likelihood = -163.43,  aic = 352.87
```

AIC favors model with trend

Comments about Results Using Lagged Variables

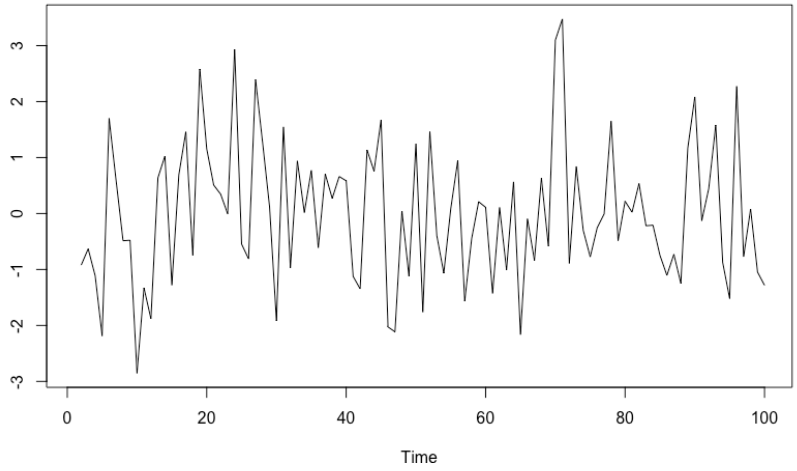
Results:

- Again, the errors in the standard multiple regression are correlated.
- Discount was not significant with or without trend in the model.
- Lagged variables ad_tv1 and ad_online1 were highly significant.
- In the presence of the lagged variables, time (trend) was not significant (although favored by AIC).
- The final model residuals (with trend) are white:

```
ltest = ljung.wge(fit$resid)
ltest$pval
```

```
> ltest$pval
[1] 0.6758827
```

There are not enough evidence to suggest that the residuals are serially correlated.



Summary:

- Advertising effects seem to be delayed

Lagged Variables in Multiple Regression with Correlated Errors

The preceding example makes an important point regarding lagged variables in multiple regression.

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_m X_{tm} + Z_t$$

Note from the model:

- All independent variables and the dependent variable are evaluated at the same time t.
  - This "restriction" seems to eliminate the use of lagged relationships in the multiple regression model.
- However, using the technique employed in the sales example, you can let, for example, X_{t1}, actually denote the TV-ad expenditures at time $t - 1$.
- This technique allows for identifying "lagged" relationships although the model itself seems to not allow them.

A Note about Lagged Variables in Multiple Time Series Analysis

When predicting gross domestic product (GDP) at time t, economists use leading economic indicators that have been observed at time $t-k$ such as

- Unemployment rate
- Income and wages
- Corporate profits
- Etc.

Point to remember:

Don't forget to consider lagged information.

DataScience@SMU

Multivariate Regression

Cross-Correlation

Cross-Correlation between 2 Time Series

A useful tool for detecting the existence of "lagged relationships" in multivariate time series analysis is the cross-correlation function.

The "cross-correlation" between variables X_{t1} and X_{t2} at lag k is the ***correlation between X_{t1} and $X_{t+k,2}$***.

We find the cross-correlation at lag k using the following ordered pairs

X_{t1}	$X_{t+k,2}$
X_{11}	$X_{1+k,2}$
X_{21}	$X_{2+k,2}$
...	
$X_{n-k,1}$	X_{n2}

Cross-Correlation between 2 Time Series

To calculate the cross-correlation between X_{t1} and $X_{t,2}$ at lag k, i.e. the correlation between X_{t1} and $X_{t+k,2}$ where $k > 0$, we use the formula

$$\hat{}_{ij}(k) = \frac{\sum_{t=1}^{n-k} (X_{ti} - \bar{X}_i)(X_{t+k,j} - \bar{X}_j)}{\sqrt{\sum_{t=1}^{n} (X_{ti} - \bar{X}_i)^2} \sqrt{\sum_{t=1}^{n} (X_{tj} - \bar{X}_j)^2}}$$

Cross-Correlation between 2 Time Series

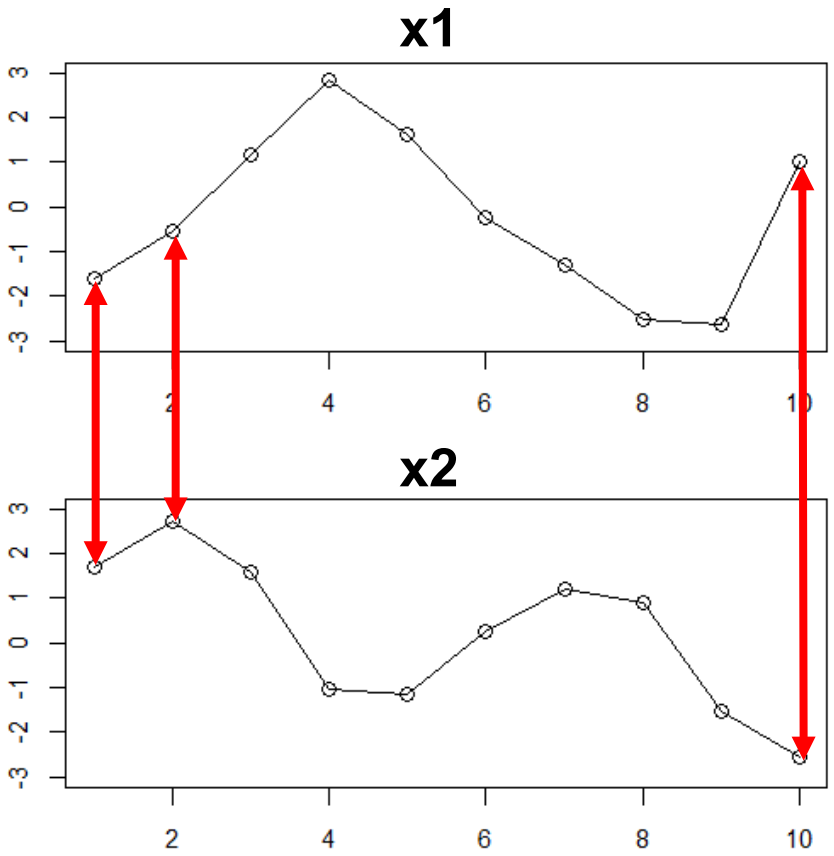
As mentioned, $\hat{\rho}_{ij}(k)$ essentially calculates the correlation based on the ordered pairs

$\hat{\rho}_{12}(k)$

X_{t1}	$X_{t+k,2}$
X_{11}	$X_{1+k,2}$
X_{21}	$X_{2+k,2}$
...	
$X_{n-k,1}$	X_{n2}

$\hat{\rho}_{12}(0)$

X_{t1}	$X_{t,2}$
X_{11}	X_{12}
X_{21}	X_{22}
...	
$X_{10,1}$	$X_{10,2}$



Cross-Correlation between 2 Time Series

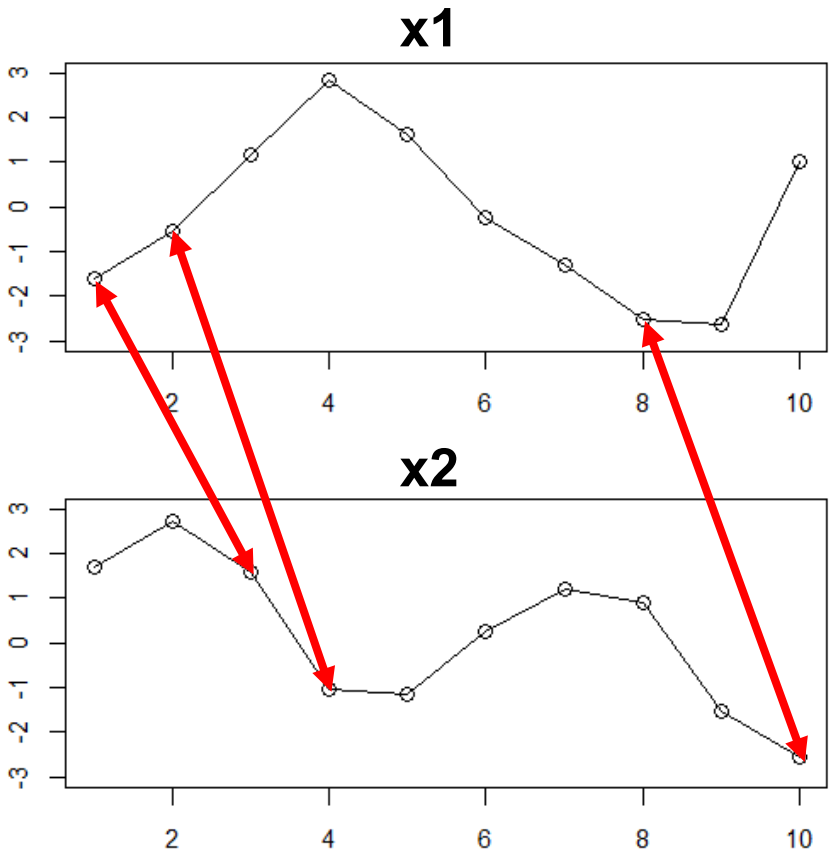
As mentioned, $\hat{\rho}_{ij}(k)$ essentially calculates the correlation based on the ordered pairs

$\hat{\rho}_{12}(k)$

X_{t1}	$X_{t+k,2}$
X_{11}	$X_{1+k,2}$
X_{21}	$X_{2+k,2}$
...	
$X_{n-k,1}$	X_{n2}

$\hat{\rho}_{12}(2)$

X_{t1}	$X_{t+2,,2}$
X_{11}	X_{32}
X_{21}	X_{42}
...	
$X_{8,1}$	$X_{10,2}$



Cross-Correlation between 2 Time Series

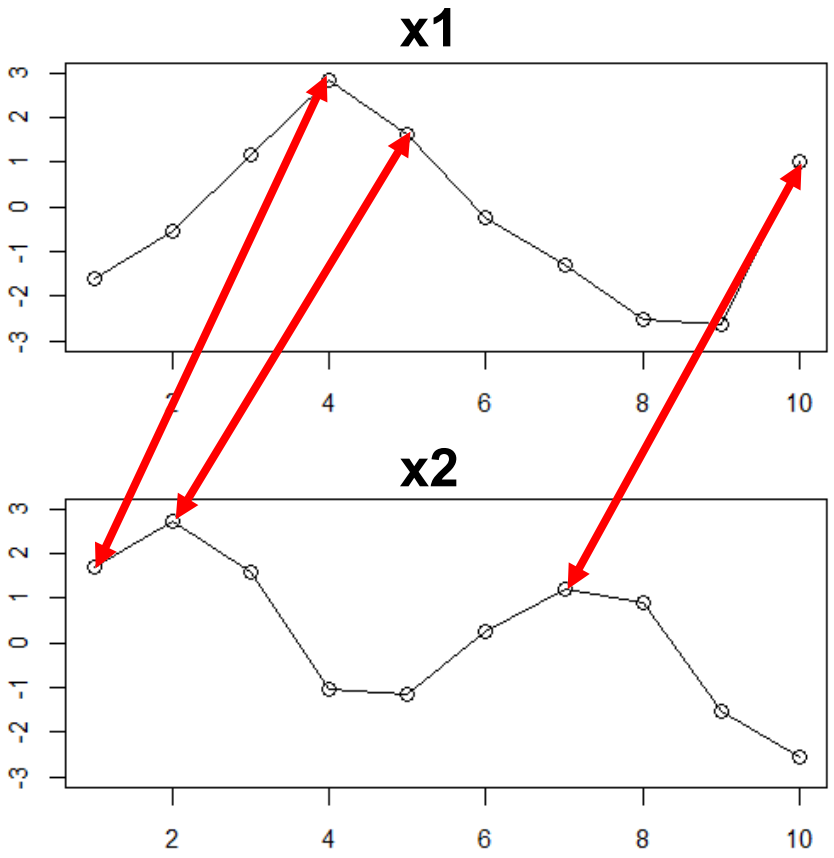
As mentioned, $\hat{\rho}_{ij}(k)$ essentially calculates the correlation based on the ordered pairs

$\hat{\rho}_{12}(k)$

X_{t1}	$X_{t+k,2}$
X_{11}	$X_{1+k,2}$
X_{21}	$X_{2+k,2}$
...	
$X_{n-k,1}$	X_{n2}

$\hat{\rho}_{12}(-3)$

X_{t1}	$X_{t+2,,2}$
X_{41}	X_{12}
X_{51}	X_{22}
...	
$X_{10,1}$	$X_{7,2}$



Graphical Presentation of Cross-Correlations

It is typical to plot cross-correlations for variables X_{ti} and X_{tj} using vertical bars (similar to autocorrelations) for a range of k values

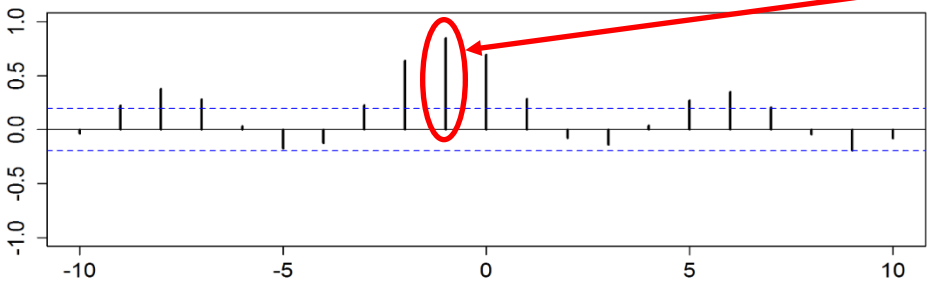
- I.e. we plot $\hat{\rho}_k,\ M_1 \leq k \leq M_2$.

On the next slide we plot cross-correlations between sales, and the three explanatory variables

- ad_tv
- ad_online
- discount

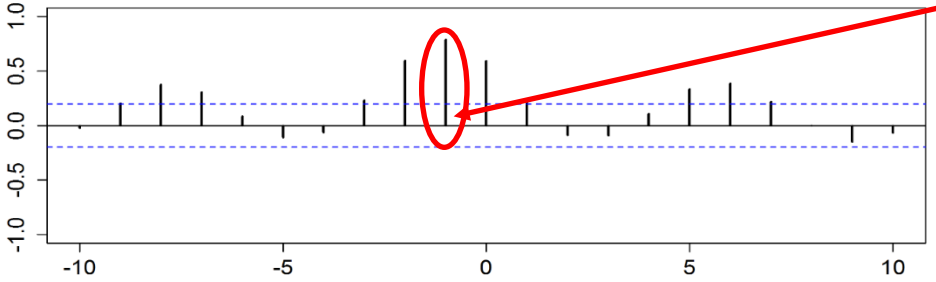
Cross-correlations

X_1 =sales, X_2= ad_tv



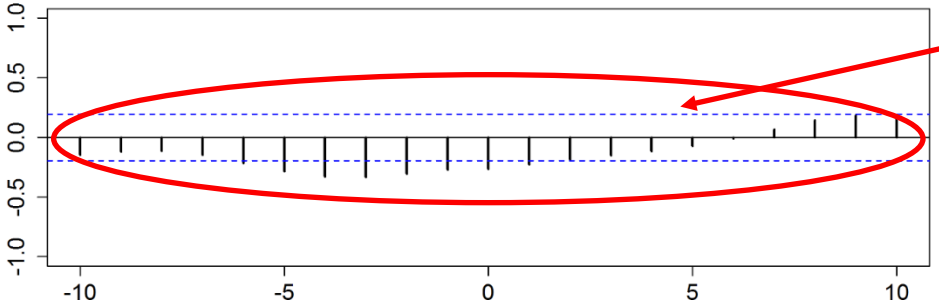
$X_{t1}, X_{t-1,2}$

X_1 =sales, X_2 = ad_online



$X_{t1}, X_{t-1,2}$

X_1 =sales, X_2 = discount



No strong cross-correlations

Be careful:

To be consistent with Woodward, et al., 2017, we have defined the cross-correlation between X_{t1} and X_{t2} at lag k to be

- The correlation between X_{t1} and $X_{t+k,2}$
- I.e. the second variable has the "time shift"

However, it is just as common to define the cross-correlation between X_{t1} and X_{t2} at lag k to be

- The correlation between $X_{t+k,1}$ and X_{t2}
- I.e. the first variable has the "time shift"

Please note:

Base R function ccf uses the ***second definition***!

In order to find the cross-correlation between **x1** and **x2** using our definition use the statement **`ccf(x2,x1)`**

DataScience@SMU

VAR Models | Definition and Example

Vector AR (VAR) Models

Simultaneously modeling several time series

In the multiple regressions with correlated errors we did not take into account the possible correlation structure within and among the independent variables.

In this setting there is no distinction between dependent and independent variables.

Our goal is to see how the interrelationships among the variables help with such things as forecasting one or more of the variables..

The variables are denoted as before:

$$X_{t1} = X_{11},\ X_{21},\ \ldots\ ,\ X_{n1}$$
$$X_{t2} = X_{12},\ X_{22},\ \ldots\ ,\ X_{n2}$$
$$\ldots$$
$$X_{tm} = X_{1m},\ X_{2m},\ \ldots\ ,\ X_{nm}$$

- First subscript is time
- Second subscript is variable number

Example: Bivariate VAR(1) Process

Recall AR(1): $X_t = (1 - \varphi_1)\mu + \varphi_1 X_{t-1} + a_t$

- Univariate: One variable, X_t
- The value of X_t involves the lag 1 value of X_t, i.e. X_{t-1}

Bivariate VAR(1):

$$X_{t1} = (1 - \varphi_{11})\mu_1 - \varphi_{12}\mu_2 + \varphi_{11}X_{t-1,1} + \varphi_{12}X_{t-1,2} + a_{t1}$$

$$X_{t2} = -\varphi_{21}\mu_1 + (1 - \varphi_{22})\mu_2 + \varphi_{21}X_{t-1,1} + \varphi_{22}X_{t-1,2} + a_{t2}$$

- Two variables: X_{t1} and X_{t2}
- The value of X_{t1} and X_{t2} involve lag one values $X_{t-1,1}$ and $X_{t-1,2}$

Matrix Notation

It is convenient to write VAR models in matrix notation.

Bivariate VAR(1) model in matrix notation

For example, the bivariate VAR(1) model (with zero mean) can be written as

$$\begin{pmatrix} X_{t1} \\ X_{t2} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} a_{t2} \\ a_{t1} \end{pmatrix}$$

where

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} (1-\varphi_{11})\mu_1 - \varphi_{12}\mu_2 \\ -\varphi_{21}\mu_1 + (1-\varphi_{22})\mu_2 \end{pmatrix}$$

Matrix Notation

Before we write the formula for a bivariate AR(2) model, note

- There will be two matrices of coefficients φ_{ij}.
- We will modify the notation, and for example, φ_{11} will be denoted $\varphi_{11(1)}$
- The vector $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$ is a (messy) extension of the VAR(1) case and in both cases $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$ is simply a vector of linear combinations of μ_1 and μ_2

Bivariate VAR(2) model

$$\begin{pmatrix}X_{t1}\\ X_{t2}\end{pmatrix} = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \begin{pmatrix}\varphi_{11(1)} & \varphi_{12(1)}\\ \varphi_{21(1)} & \varphi_{22(1)}\end{pmatrix}\begin{pmatrix}X_{t-1,1}\\ X_{t-1,2}\end{pmatrix} + \begin{pmatrix}\varphi_{11(2)} & \varphi_{12(2)}\\ \varphi_{21(2)} & \varphi_{22(2)}\end{pmatrix}\begin{pmatrix}X_{t-2,1}\\ X_{t-2,2}\end{pmatrix} + \begin{pmatrix}a_{t2}\\ a_{t1}\end{pmatrix}$$

Bivariate VAR(2)

Expanding the VAR(2) matrix equation on the previous slide, we get

$$X_{t1} = \beta_1 + \varphi_{11(1)}X_{t-1,1} + \varphi_{12(1)}X_{t-1,2} + \varphi_{11(2)}X_{t-2,1} + \varphi_{12(2)}X_{t-2,2} + a_{t1}$$

$$X_{t2} = \beta_2 + \varphi_{21(1)}X_{t-1,1} + \varphi_{22(1)}X_{t-1,2} + \varphi_{21(2)}X_{t-2,1} + \varphi_{22(2)}X_{t-2,2} + a_{t1}$$

Note that:

- X_{t1} depends on lagged $t-1$ and $t-2$ versions of both X_{t1} and X_{t2}
- Clearly, writing the equations in expanded form will get very cumbersome very quickly for higher order models and more than two variables
- We use R functions!
  - CRAN package **`vars`** is particularly useful

DataScience@SMU

Forecasting with VAR Models

Forecasting with VAR(p) Models

Forecasting with VAR(p) models is simply an extension of forecasting with AR(p) models

Quick review: forecasting with AR(p) models

Forecasts for an AR(p) model are based on the underlying AR(p) equation and are given by

$$\hat{X}_{t_0}(\ell) = \bar{X}(1-\varphi_1-\cdots-\varphi_p)+\varphi_1\hat{X}_{t_0}(\ell-1)+\cdots+\varphi_p\hat{X}_{t_0}(\ell-p)$$

Specifically, the ℓ-step ahead forecast, $\hat{X}_{t_0}(\ell)$, for the univariate variable X_t depends on $\ell-1$ through $\ell-p$ step ahead forecasts, i.e. $\hat{X}_{t_0}(\ell-1),\ldots,\hat{X}_{t_0}(\ell-p)$ (some of which may be actual observed values)

Forecasting with VAR(p) Models

For simplicity we will show forecasts for a bivariate VAR(1) model

$$X_{t1} = (1-\varphi_{11})\mu_1 - \varphi_{12}\mu_2 + \varphi_{11}X_{t-1,1} + \varphi_{12}X_{t-1,2} + a_{t1}$$
$$X_{t2} = -\varphi_{21}\mu_1 + (1-\varphi_{22})\mu_2 + \varphi_{21}X_{t-1,1} + \varphi_{22}X_{t-1,2} + a_{t2}$$

The forecasts are given by

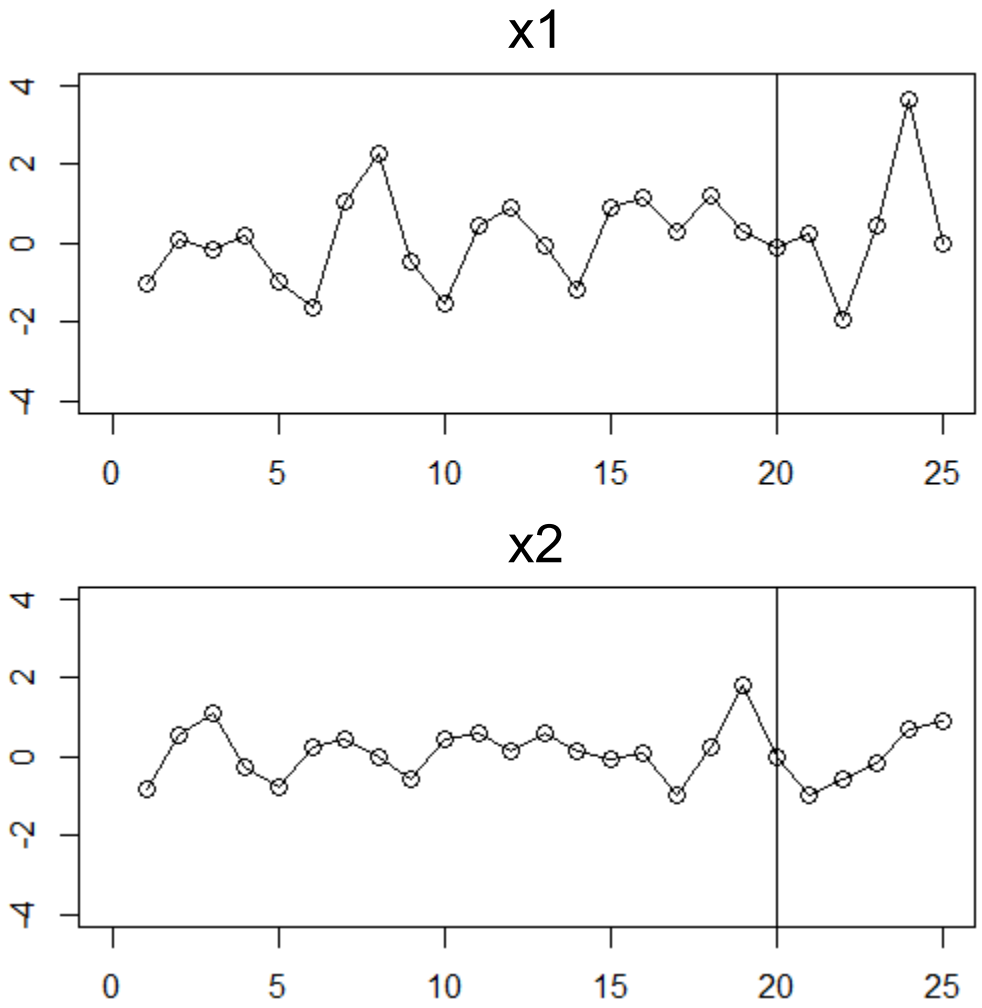
$$\hat{X}_{t_0 1}(\ell) = (1-\varphi_{11})\bar{X}_1 - \varphi_{12}\bar{X}_2 + \varphi_{11}\hat{X}_{t_0 1}(\ell-1) + \varphi_{12}\hat{X}_{t_0 2}(\ell-1)$$
$$\hat{X}_{t_0 2}(\ell) = -\varphi_{21}\bar{X}_1 + (1-\varphi_{22})\bar{X}_2 + \varphi_{21}\hat{X}_{t_0 1}(\ell-1) + \varphi_{22}\hat{X}_{t_0 2}(\ell-1)$$

Specifically, the ℓ-step ahead forecasts, $\hat{X}_{t_0 1}(\ell)$, for the variable X_{t1} depend on $\ell-1$ step ahead forecasts for both variables X_{t1} and X_{t2}. Forecasts, $\hat{X}_{t_0 2}(\ell)$, have a similar form.

DataScience@SMU

VAR Models | Example

Bivariate VAR Time Series Example



We know data out to t=25

We will perform analysis based on data only up to t=20

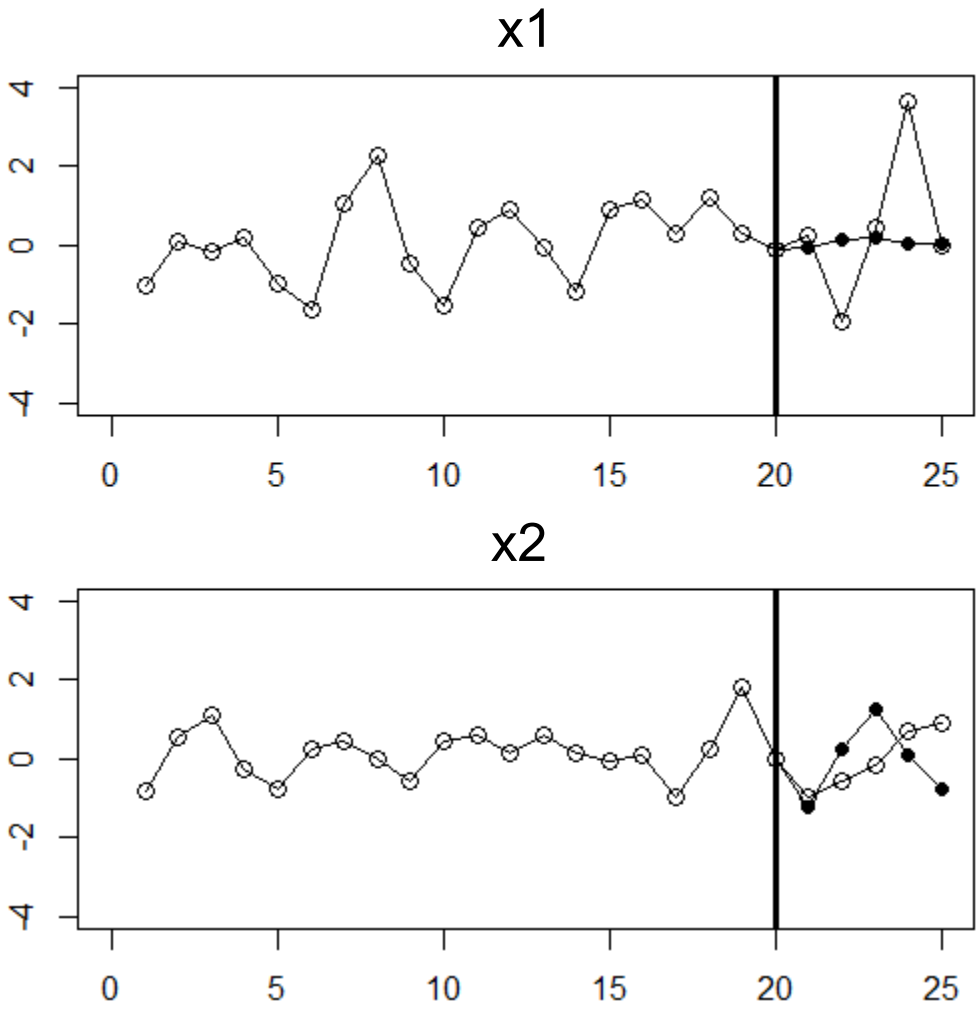
R Code using tswge and vars to Analyze Bivariate Data (x1,x2)

```
x1.25=c( -1.03,  0.11, -0.18, 0.20, -0.99, -1.63, 1.07,  2.26, -0.49, -1.54,  0.45,  0.92,
       -0.05, -1.18,  0.90,  1.17,  0.31,  1.19,  0.27, -0.09,  0.23, -1.91,  0.46,  3.61, -0.03)
x2.25=c( -0.82,  0.54,  1.13, -0.24, -0.77,  0.22,  0.46, -0.03, -0.59,  0.45,  0.59,  0.15,
       0.60,  0.13, -0.04,  0.12, -0.96,  0.23,  1.81, -0.01, -0.95, -0.55, -0.15,  0.71,  0.90)
x1=x1.25[1:20]
x2=x2.25[1:20]
p1=aic.wge(x1,p=0:8,q=0:0)
# aic picks p=2
x1.est=est.ar.wge(x1,p=p1$p)
fore.arma.wge(x1,phi=x1.est$phi,n.ahead=5,lastn=FALSE,limits=FALSE)
p2=aic.wge(x2,p=0:8,q=0:0)
# aic picks p=2
x2.est=est.ar.wge(x1,p=p2$p)
fore.arma.wge(x2,phi=x2.est$phi,n.ahead=5,lastn=FALSE,limits=FALSE)
#
# VAR and VARselect are from CRAN package vars
X=cbind(x1,x2)
VARselect(X, lag.max = 6, type = "const",season = NULL, exogen = NULL)
#VARselect picks p=5 (using AIC)
lsfit=VAR(X,p=5,type="const")
preds=predict(lsfit,n.ahead=5)
# preds$fcst$x1[1,1] - [5,1] are the VAR forecasts for x1.  Similar for x2
library(RColorBrewer)
fanchart(preds, colors = brewer.pal(n = 8, name = "Blues")) # Change color pallet to make distinguishable.
```

DataScience@SMU

VAR Models | Example

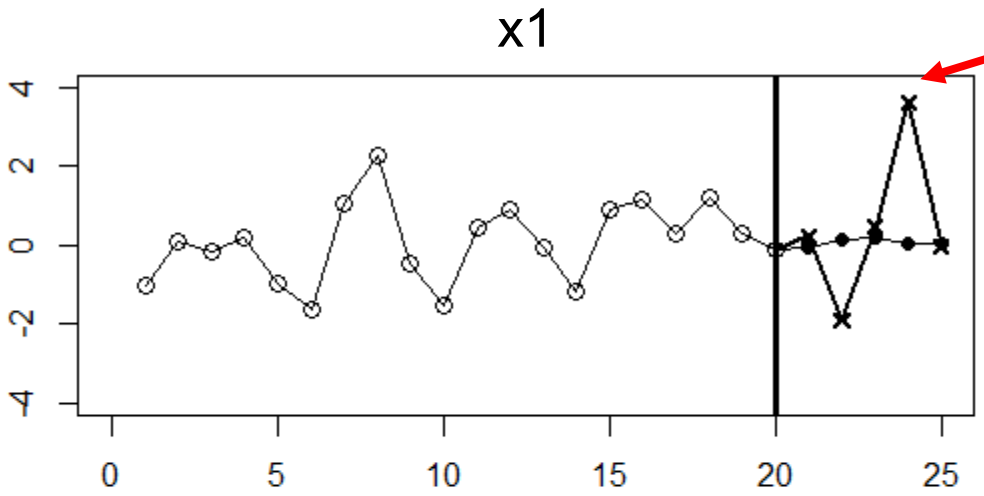
Univariate Forecasts

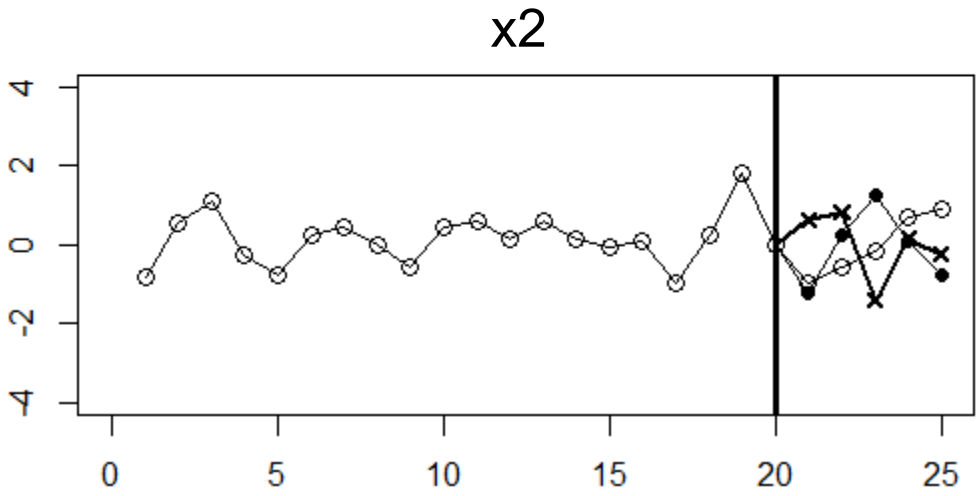


These are each realizations from AR(2) models, and the solid dots are the "univariate" forecasts for each realization based on data to time t=20

- The forecasts (especially for x1) are not very informative

Forecasting with Bivariate VAR Models





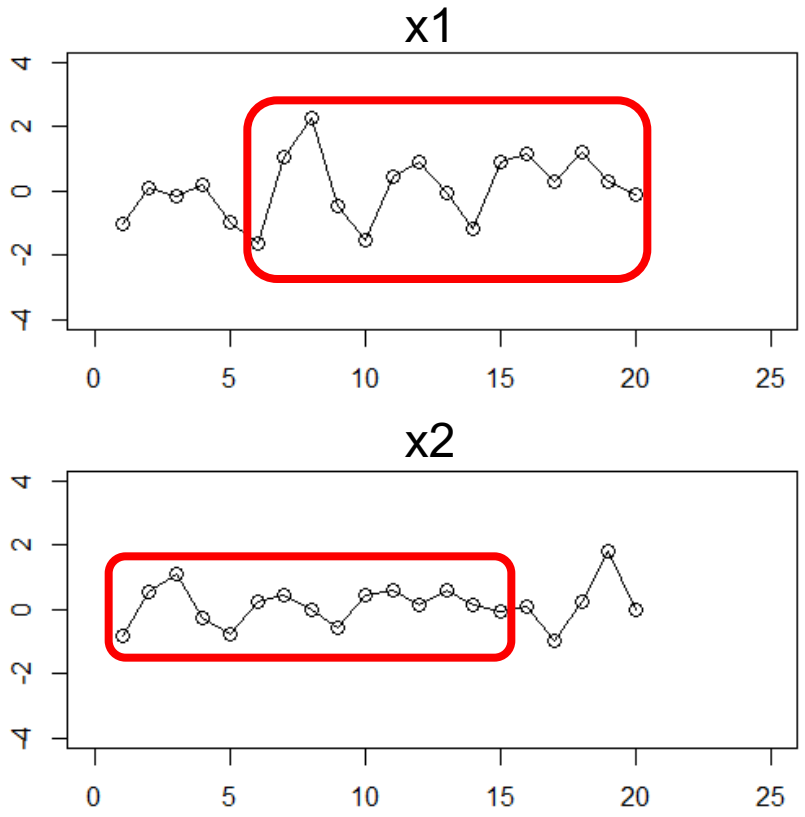
wow!!

The solid dots are the univariate forecasts from the previous slide

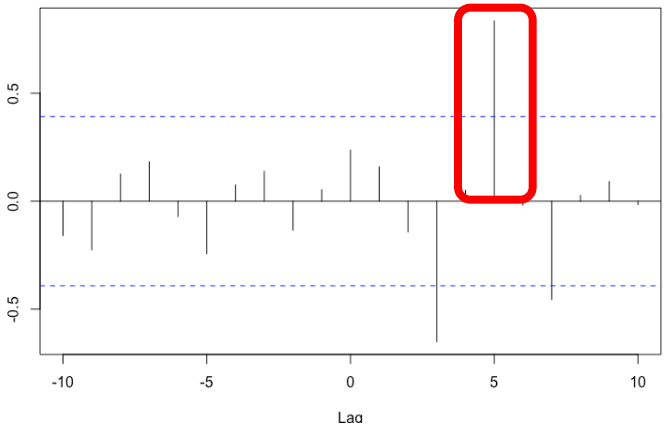
The “x”’s denote the forecasts based on the VAR model fit to the data

What Happened?

- Why are the forecasts for x1 so good?
- Let's look at the data again.



x1(*t*) is essentially a scaled version of 2*x2(*t*-5)



a = ccf(x2.25,x1.25)

```
Autocorrelations of series 'X', by lag

   -10     -9     -8     -7     -6     -5     -4     -3     -2     -1      0
-0.016  0.091  0.026 -0.455 -0.019  0.835  0.050 -0.651 -0.143  0.159  0.236
     1      2      3      4      5      6      7      8      9     10
 0.053 -0.134  0.139  0.075 -0.243 -0.071  0.183  0.125 -0.225 -0.159
```

Bivariate VAR(5) Model

$$X_{t1} = \varphi_{11(1)}X_{t-1,1} + \varphi_{12(1)}X_{t-1,2} + \varphi_{11(2)}X_{t-2,1} + \varphi_{12(2)}X_{t-2,2} +$$
$$\varphi_{11(3)}X_{t-3,1} + \varphi_{12(3)}X_{t-3,2} + \varphi_{11(4)}X_{t-4,1} + \varphi_{12(4)}X_{t-4,2} +$$
$$\varphi_{11(5)}X_{t-5,1} + \varphi_{12(5)}X_{t-5,2} + \beta_1 + a_{t1}$$

X_{t2} is defined analogously

After running the R code for VAR modeling, if you type lsfit you obtain the following output:

Estimated coefficients for equation x1: (eg. x1.l3 is the coefficient on $X_{t-3,1}$)

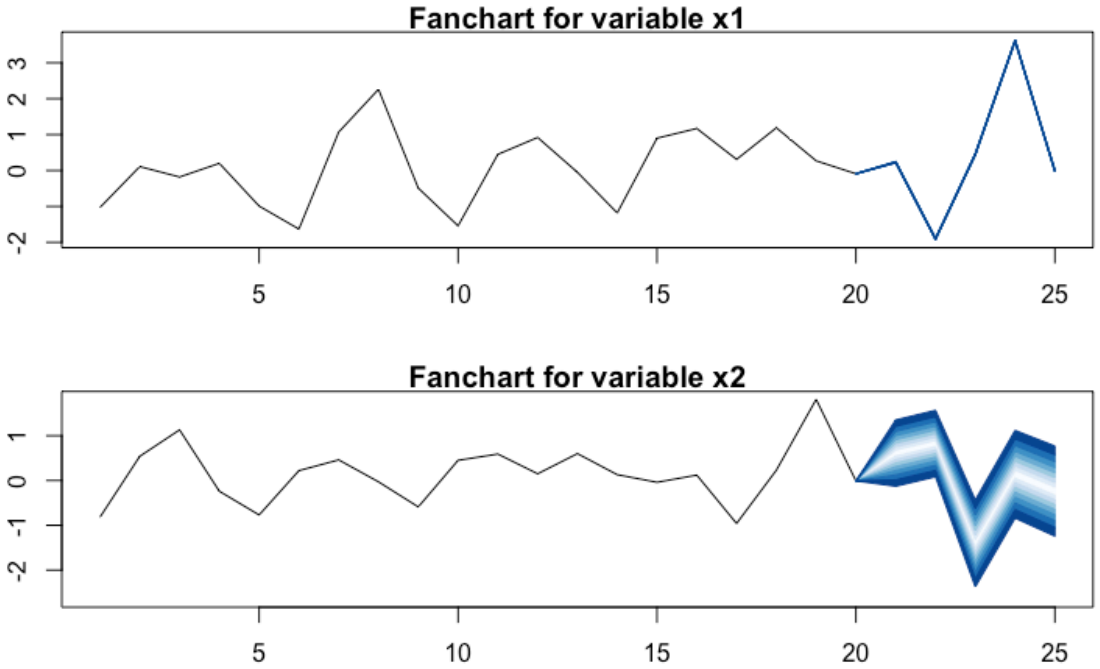
==========

x1 = x1.l1 + x2.l1 + x1.l2 + x2.l2 + x1.l3 + x2.l3 + x1.l4 + x2.l4 + x1.l5 + x2.l5 + const

x1.l1	x2.l1	x1.l2	x2.l2	x1.l3	x2.l3	x1.l4	x2.l4	x1.l5	x2.l5	const
-0.0043	0.0025	0.0041	-0.0078	-0.0048	0.0097	0.0045	-0.0079	-0.0022	2.0046	-0.0002

Note: In the formula for X_{t1}, the coefficient of $X_{t-5,2}$ is about two while all other coefficients are approximately zero.

Bivariate VAR(5) Model: Fan Chart and Prediction Intervals




```
library(RColorBrewer)
fanchart(preds, colors = brewer.pal(n = 8, name = "Blues")) # Change color pallet to make distinguishable
```

Note: Note that the margin or error for the x1 variable is so small that the forecasts and the upper and lower limits are nearly overlapped. We used the RColorBrewer package to change the colors to be more distinguishable.

Comments about the Bivariate VAR(5) Example

- This example was constructed to show how one variable in a VAR model can be a leading indicator for another variable.
- Because variable $X_{t1} = 2X_{t-5,2}$, we would expect the VAR forecasts to somehow take advantage of this fact.
- Notice that
  - Both X_{t1} and X_{t2} were generated as AR(2) processes.
  - AIC applied to each series separately picked an AR(2).
  - Function **`VARselect`** identified the VAR model as a VAR(5).
    - It was necessary that *p* was at least 5 so that lag 5 would be in the fitted model
    - AIC detected this "by itself"
    - The fitted VAR(5) model gave essentially perfect forecasts for X_{t1} for steps ahead up to 5
    - The forecasts for X_{t2} showed no detectable improvement over the univariate forecasts.

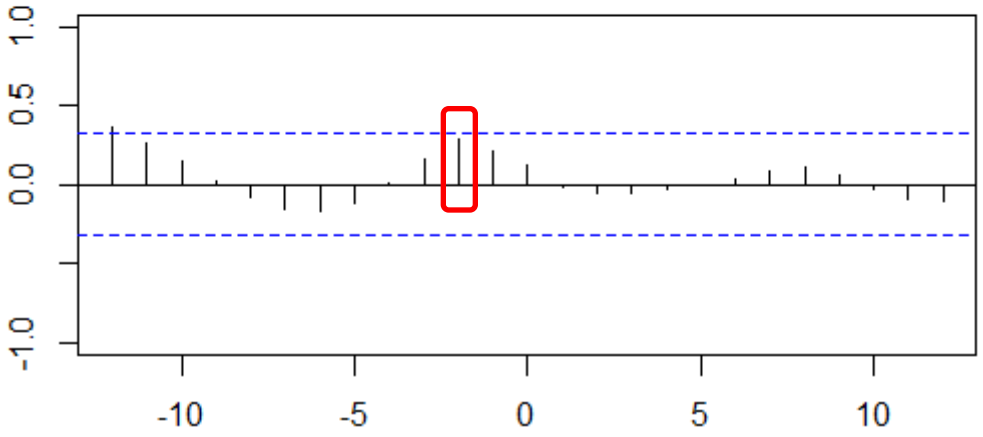
DataScience@SMU

Melanoma and Sunspots Example

Melanoma Sunspot Example
Annual Data: 1936-1972

This data set has caused interest because of the fact that there is some evidence that the melanoma incidence at year *t* is related to the sunspot number at year *t*-2.

Cross-covariance of melanoma at time *t* and sunspot at time *t+k*

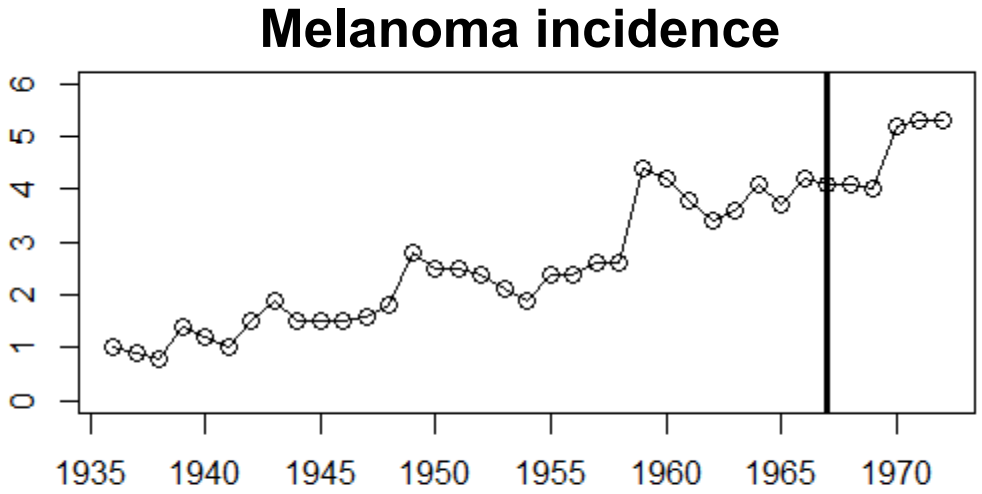


The "peak" at -2 suggests that sunspot number at time *t*-2 is related to melanoma at time *t*

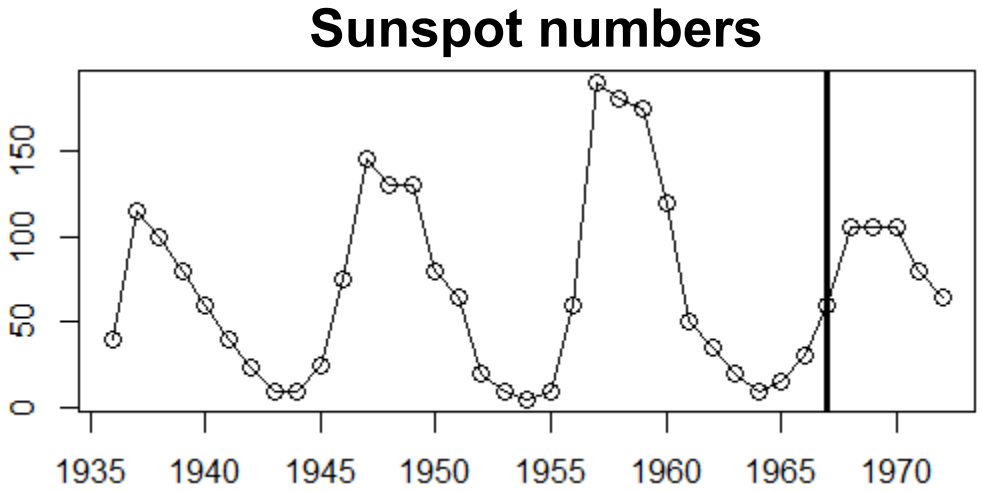
But, note that the "peak" is within the error bars for zero cross-correlations

- Relationship may not be strong

Melanoma Sunspot Example
Annual Data: 1936-1972



We know data out to t=1972



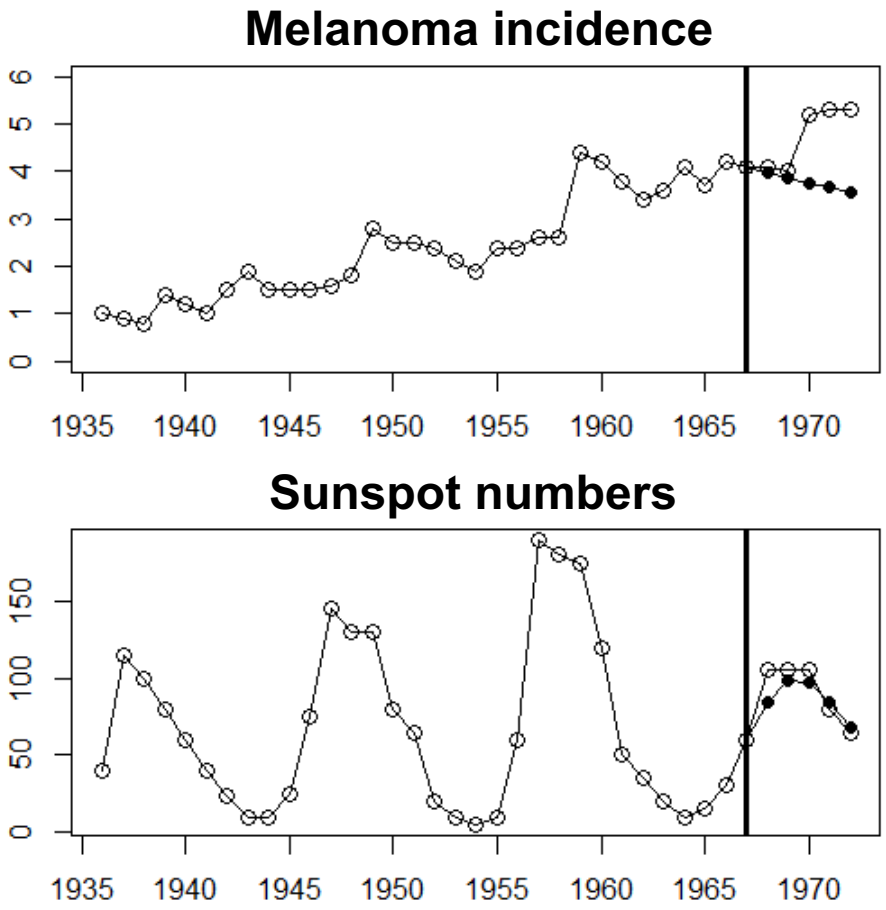
We will perform analysis based on data only up to t=1967

R Code using tswge and vars to Analyze Melanoma-Sunspot Data

```
# melanoma incidence and sunspot numbers 1936-1972
 melanoma=c(1.0, 0.9, 0.8, 1.4, 1.2, 1.0, 1.5, 1.9, 1.5, 1.5, 1.5, 1.6, 1.8, 2.8, 2.5, 2.5, 2.4, 2.1, 1.9, 2.4, 2.4, 2.6, 2.6,
4.4, 4.2, 3.8, 3.4, 3.6, 4.1, 3.7, 4.2, 4.1, 4.1, 4.0, 5.2, 5.3, 5.3)
sunspot=c(40, 115, 100,  80,  60,  40,  23,  10,  10,  25,  75, 145, 130, 130,  80,  65,  20,  10,   5,  10, 60, 190, 180, 175,
120,  50,  35,  20,  10,  15,  30,  60, 105, 105, 105,  80,  65)
mel.67=melanoma[1:32]
sun.67=sunspot[1:32]
p.mel=aic.wge(mel.67,p=0:8,q=0:0)
p.mel$p
mel.est=est.ar.wge(mel.67,p=p.mel$p)
fore.arma.wge(mel.67,phi=mel.est$phi,n.ahead=5,lastn=FALSE,limits=FALSE)
p.sun=aic.wge(sun.67,p=0:8,q=0:0)
p.sun$p
sun.est=est.ar.wge(sun.67,p=p.sun$p)
fore.arma.wge(sun.67,phi=sun.est$phi,n.ahead=5,lastn=FALSE,limits=FALSE)
#
# VAR and VARselect are from CRAN package vars
X=cbind(mel.67,sun.67)
VARselect(X, lag.max = 6, type = "const",season = NULL, exogen = NULL) #AIC = 5.04
#VARselect picks p=4 (using AIC)
lsfit=VAR(X,p=4,type='const')
preds=predict(lsfit,n.ahead=5)
#preds$fcst$mel[1,1]-[5,1] are the VAR forecasts formenlanoma.  Similar for sunspot.
plot(seq(1,37,1),melanoma, type = "b", ylim = c(0,6))
points(seq(33,37,1),preds$fcst$mel.67[1:5,1],type = "b", pch = 15)
fanchart(preds)
```

Melanoma Sunspot Example

Univariate forecasts

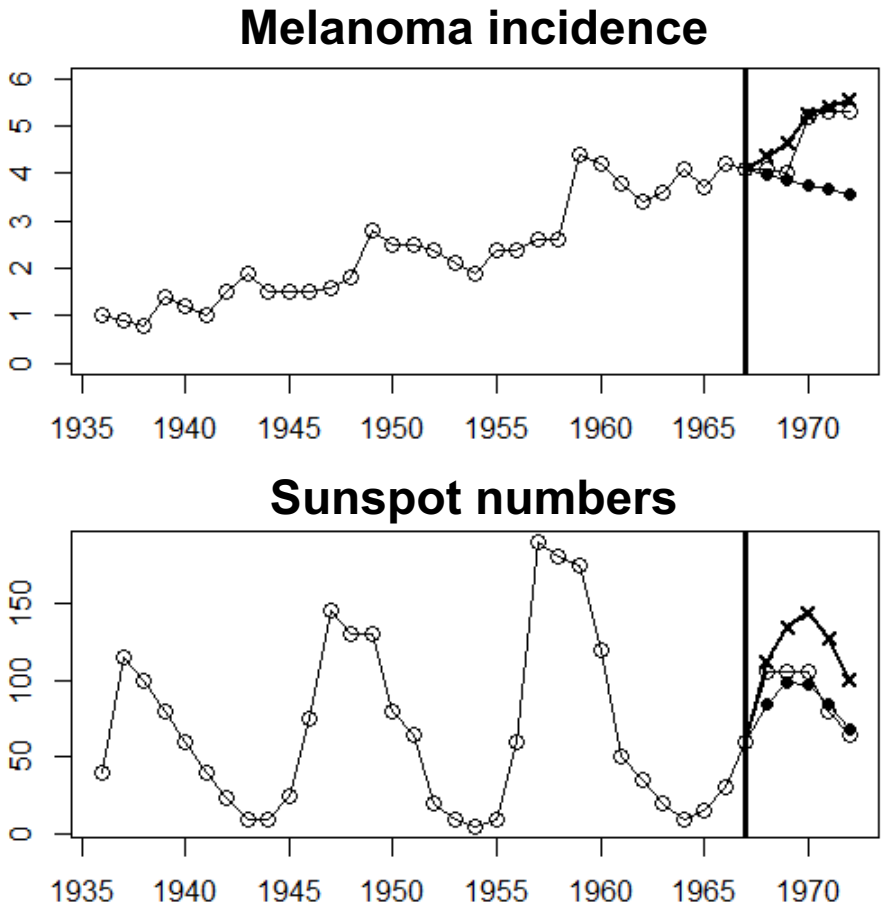


AIC selects AR(1) and AR(2) for the melanoma and sunspot univariate models, respectively.

- The solid dots are the "univariate" forecasts for each realization based on data to time t=1967
- Melanoma forecasts (from a simple AR(1) model) are quite poor
- Sunspot forecasts for short steps ahead are good

Melanoma Sunspot Example

Bivariate VAR(4) forecasts

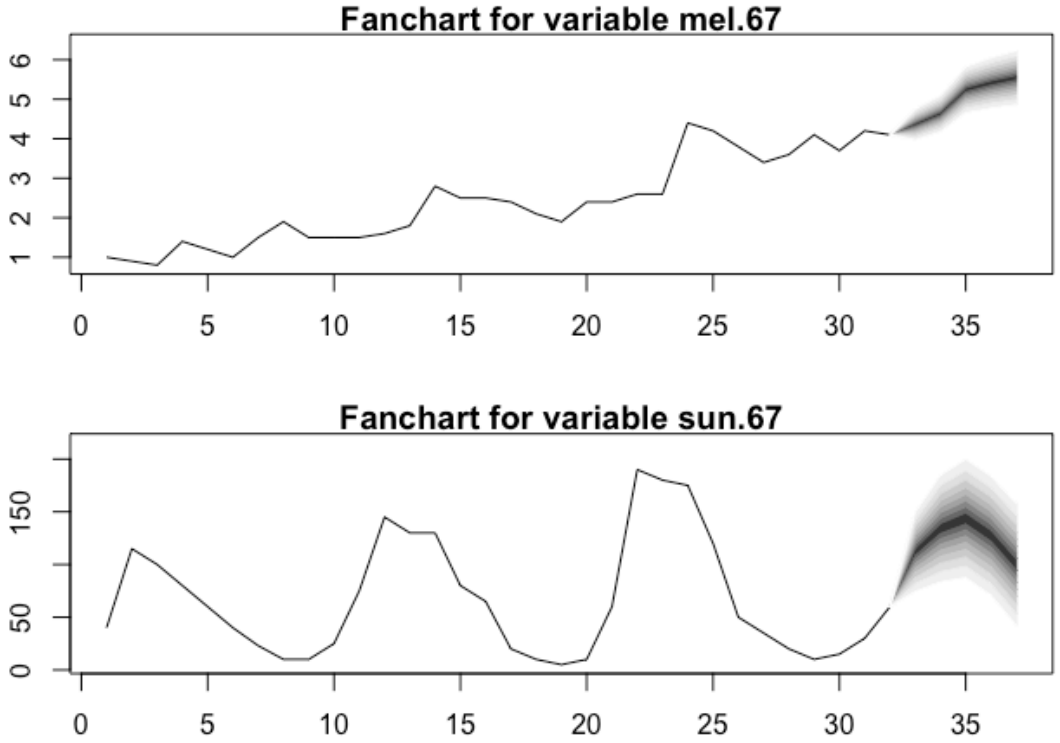


The VAR forecasts are shown by "***x***"

- Forecasts for melanoma have greatly improved
- Probably based on the fact that melanoma incidence at time t is related to the sunspot number at time *t*-2
- The small cross-correlation made more difference in forecasts than might be expected
- Short term sunspot forecasts are not improved with the melanoma information

Melanoma Sunspot Example

Plots with prediction intervals



Again, judging by the wider confidence intervals, we can see that sunspots provide more information in predicting melanoma than the other way around.

Multiple Regression Analysis of Melanoma-Sunspot Data

Before leaving this example, we examine the melanoma-sunspot data using a multiple regression with correlated errors.

- Since melanoma tends to trend upward we include t and sunspot as independent variables for predicting melanoma.

```
t=1:37
ksfit=lm(melanoma~sunspot+t)
phi=aic.wge(ksfit$residuals,p=0:8,q=0:0) # AIC picks p=2
fit=arima(melanoma,order=c(phi$p,0,0),xreg=cbind(t,sunspot))
fit
```

Output:

Coefficients:

	ar1	ar2	intercept	t	sunspot
	0.4795	-0.2461	0.3957	0.1157	0.0021
s.e.	0.1627	0.1713	0.1951	0.0077	0.0017

- AIC picked an AR(2) for the residuals
- There is a significant trend but sunspot is not a significant predictor

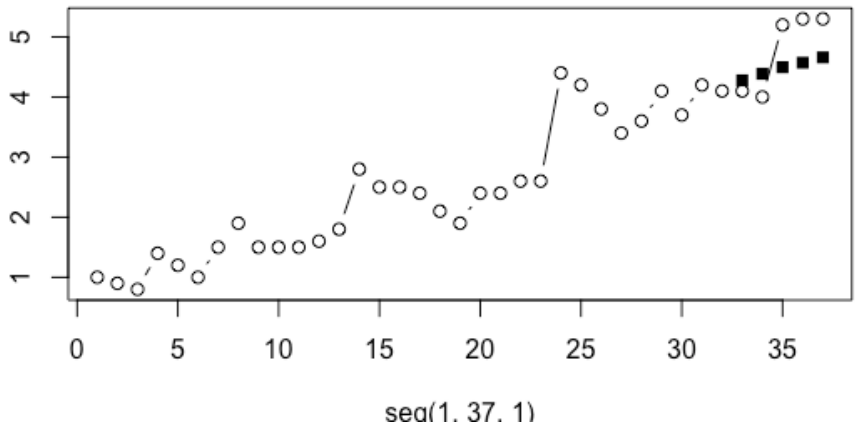
aic = 45.17

Multiple Regression Analysis of Melanoma-Sunspot Data

Before leaving this example, we examine the melanoma-sunspot data using a multiple regression with correlated errors.

- Since melanoma tends to trend upward we include t and sunspot as independent variables for predicting melanoma.

```
t=1:32
ksfit=lm(melanoma[1:32]~sunspot[1:32]+t)
phi=aic.wge(ksfit$residuals,p=0:8,q=0:0) # AIC picks p=2
fit=arima(melanoma[1:32],order=c(phi$p,0,0),xreg=cbind(t,sunspot[1:32]))
preds = predict(fit,newxreg = data.frame(t = c(33,34,35,36,37), sunspot = sunspot[33:37]))
plot(seq(1,37,1),melanoma, type = "b")
points(seq(33,37,1),preds$pred,type = "b", pch = 15)
```



aic = 45.17

ase = ?

Multiple Regression Analysis of Melanoma-Sunspot Data

Remembering that melanoma incidence seemed to be related to sunspot activity ***two years earlier***, we examine a lagged effect (along with time)

- We compute variables sun2 and mel2 below

```
t=1:35
sun2=sunspot[1:35]
for(i in 1:35){mel2[i]=melanoma[i+2]}
ksfit=lm(mel2~t+sun2)
phi=aic.wge(ksfit$residuals,p=0:8,q=0:0)  # AIC picks p=1
fit=arima(mel2,order=c(phi$p,0,0),xreg=cbind(t,sun2))
fit
```

Output:

Coefficients:

	ar1	intercept	t	sun2
	0.3124	0.2746	0.1177	0.0064
s.e.	0.1732	0.1608	0.0068	0.0013

- AIC picked an AR(1) for the residuals
- The coefficients for trend ***and*** sunspots are significant

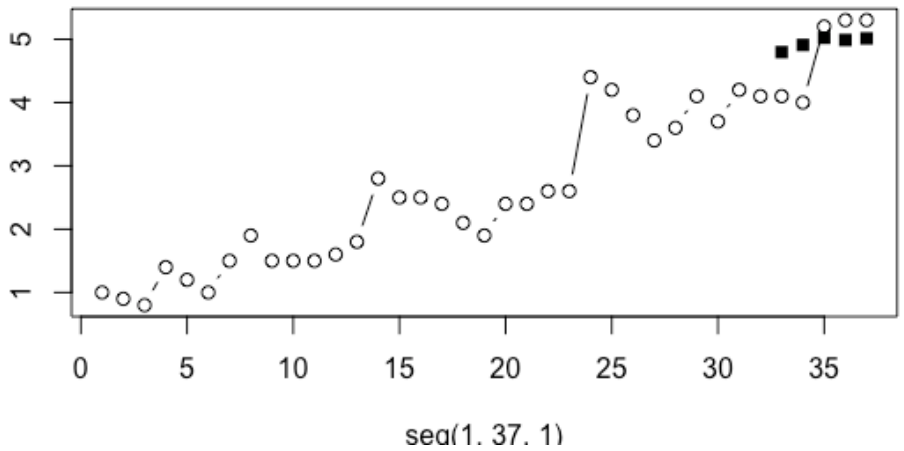
aic = 22.94

Multiple Regression Analysis of Melanoma-Sunspot Data

Remembering that melanoma incidence seemed to be related to sunspot activity ***two years earlier***, we examine a lagged effect (along with time)

- We compute variables sun2 and mel2 below

```
t=1:30
sun2=sunspot[1:30]
mel2 = c()
for(i in 1:30){mel2[i]=melanoma[i+2]}
ksfit=lm(mel2~t+sun2)
phi=aic.wge(ksfit$residuals,p=0:8,q=0:0)  # AIC picks p=1
fit=arima(mel2,order=c(phi$p,0,0),xreg=cbind(t,sun2))
preds = predict(fit,newxreg = data.frame(t = c(33,34,35,36,37), sunspot = sunspot[33:37]))
plot(seq(1,37,1),melanoma, type = "b")
points(seq(33,37,1),preds$pred,type = "b", pch = 15)
```



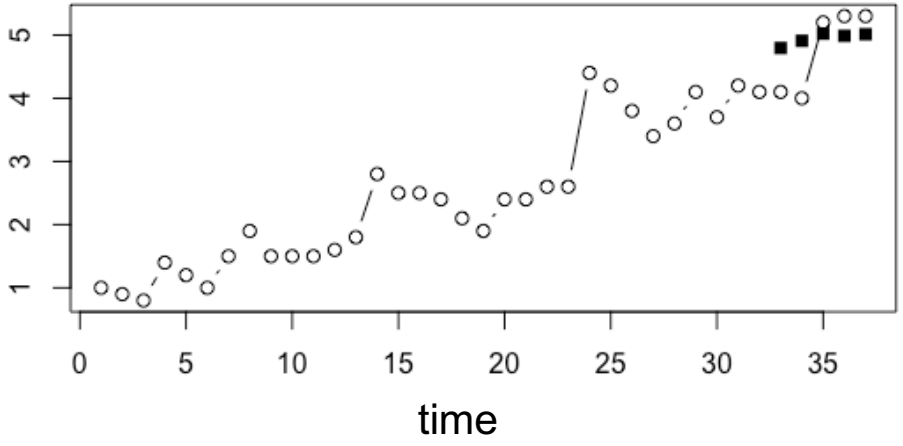
aic = 22.94

ase = ?

Comparison!!!

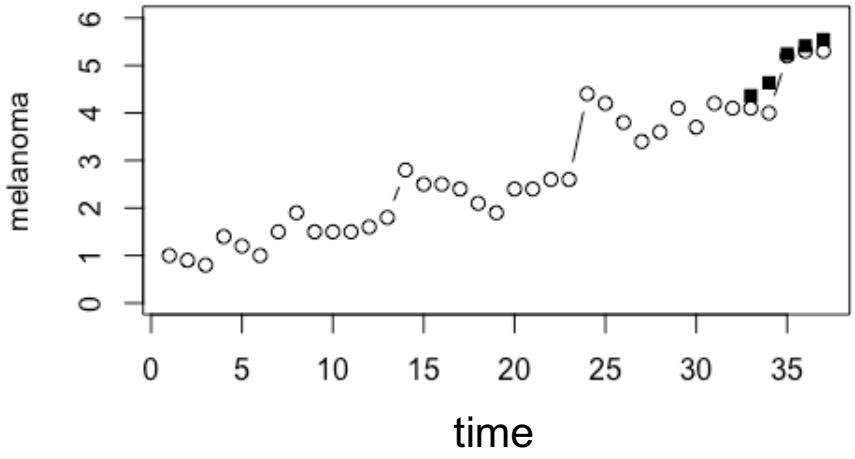
Melanoma incidence

Multiple Regression Lag = 2, AR(1)



aic = 22.94

Melanoma incidence VAR(4)



aic = 5.04

DataScience@SMU