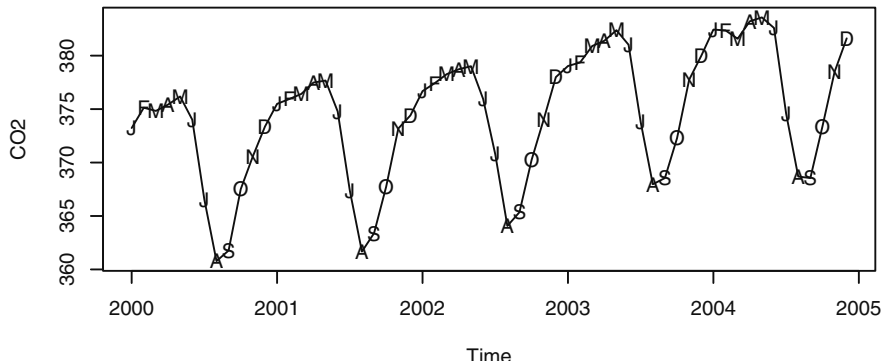
# Ch 10+: Model building

## Motivation

- We discussed the large class of seasonal ARIMA$(p, d, q)$x$(P, D, Q)_s$ models that can be used for forecasting time series.
- In Ch 10 (Ch. 10.4 in the book), we discussed how to select a model from this large class of models to forecast the data series CO2 (below).
- This approach was somewhat informal because it was based on visual inspection of plots.
- While such approaches are valuable for exploratory time series analysis, more formal/automated model selection approaches can help to make model selection less subjective and easier to carry out.

**Exhibit 10.2 Carbon Dioxide Levels with Monthly Symbols**



# Time series model building

- Let's discuss a more formal and automated approach to model building.
- These slides contain a combination of material from the book and additional reference material (so they are a bit more word-y than usual):
  - We will follow some of the recommendations from Hyndman and Khandakar (2008): Automatic Time Series Forecasting: The forecast package for R. Journal of Statistical Software, Vol 27-3.
  - This paper is uploaded on IVLE and here referred to as H&K.
  - All R-code is in "modelbuilding.R".

# Time series model building: review of models I

- $Y_t$ is a multiplicative ARMA$(p, q)$x$(P, Q)_s$ process with
  - constant term $\theta_0$,
  - seasonal period $s$,
  - AR characteristic polynomial $\phi(x)\Phi(x)$ with

$$\begin{aligned}\phi(x) &= 1 - \phi_1 x - \phi_2 x^2 - \ldots - \phi_p x^p, \\ \Phi(x) &= 1 - \Phi_1 x^s - \Phi_2 x^{2\cdot s} - \ldots - \Phi_P x^{P\cdot s},\end{aligned}$$

  - MA characteristic polynomial $\theta(x)\Theta(x)$ with

$$\begin{aligned}\theta(x) &= 1 - \theta_1 x - \theta_2 x^2 - \ldots - \theta_q x^q, \\ \Theta(x) &= 1 - \Theta_1 x^s - \Theta_2 x^{2\cdot s} - \ldots - \Theta_Q x^{Q\cdot s},\end{aligned}$$

if $Y_t$ is defined as follows:

$$\phi(B)\Phi(B)Y_t = \theta_0 + \theta(B)\Theta(B)e_t.$$

# Time series model building: review of models II

- When including non-seasonal and/or seasonal differencing, we obtain multiplicative ARIMA$(p,d,q)$x$(P,D,Q)_s$ processes.
- A process $Y_t$ is a multiplicative ARIMA$(p,d,q)$x$(P,D,Q)_s$ process with
  - constant term $\theta_0$,
  - seasonal period $s$,
  - non-seasonal orders $p, q$ and seasonal orders $P, Q$ and characteristic functions as described on the previous slide

  if $Y_t$ is defined as follows:

$$\phi(B)\Phi(B)(1-B^s)^D(1-B)^d Y_t = \theta_0 + \theta(B)\Theta(B)e_t.$$

  or equivalently

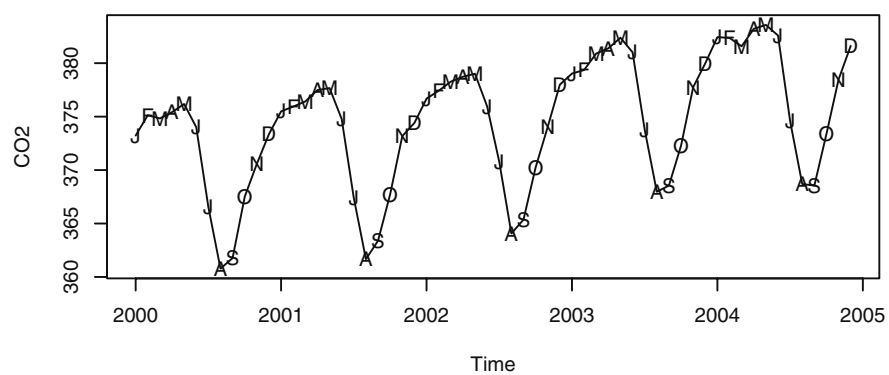
$$\phi(B)\Phi(B)W_t = \theta_0 + \theta(B)\Theta(B)e_t,$$

  for $W_t = \nabla_s^D \nabla^d Y_t = (1-B^s)^D(1-B)^d Y_t$.

# Time series model building: overview

- What are the tasks involved in selecting candidate ARIMA$(p, d, q)$x$(P, D, Q)_s$ models, e.g. for the CO2 series?
  - Step 1: Select the order $d$ of non-seasonal and $D$ of seasonal differencing.
  - Step 2:
    - Select the orders $p, q, P, Q$ and decide whether or not the constant term $\theta_0$ should be included,
    - Decide whether or not "in-between" predictors (lagged $Y$'s or past white noise terms) should be removed.
- Steps 1 and 2 may result in a set of candidate models. If so, we need
  - Step 3: Which model to use for forecasting?

**Exhibit 10.2 Carbon Dioxide Levels with Monthly Symbols**



# Step 1: Selecting the values for $d$ and $D$.

- We follow the approach suggested by H&K, who recommend using *unit root tests* to determine $d$ and $D$, a commonly used approach.
- Unit-root tests try to answer the question whether a time series is stationary or not, by investigating whether a unit root exists in the AR-characteristic equation of an ARMA specification of the series (hence the name).
  - E.g. if $Y_t = Y_{t-1} + e_t$, the AR-characteristic equation is $\phi(x) = 1 - x = 0$, which has a unit root $x = 1$.
  - For this example, a unit root test for $Y_t$ is likely to suggest that there is evidence of a unit root and thus that the series is not stationary, so we should difference the series.

# Differences in unit-root tests

- Most unit-root tests, such as the augmented Dickey-Fuller Unit-Root test (p.128 in the book) are based on the null hypothesis ($H_0$) that the series is non-stationary, e.g. that a unit-root exists.
  - Based on such tests, we would difference a series if we cannot reject $H_0$.
  - E.g. if $Y_t = Y_{t-1} + e_t$, the augmented Dickey-Fuller Unit-Root test is likely to suggest that cannot reject $H_0$, so we would difference the series.
- However, H&K point out that such tests bias results towards over-differencing. They propose to use unit-root tests that are based on a null hypothesis of no unit-root.
  - Based on such tests, we would difference a series if we reject $H_0$.

# A recommended unit-root test

- For non-seasonal data, H&K recommend the "KPSS test", based on the $H_0$ that there is no unit-root (no differencing is needed).
- Details on this test are outside the class material, but can be found here: *Kwiatkowski D, Phillips PCB, Schmidt P and Shin Y (1992): Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. Journal of Econometrics 54:159-178.*
- We will use the test as follows to select $d$:
  - If the test results is significant (if the p-value for the test statistic is very small), the null hypothesis of no unit-root is rejected, so we will difference the time series.
  - After differencing the time series, the test is repeated until the first insignificant test result is obtained.
- R-function: ndiffs, which (conveniently) outputs the number of times the series should be differenced

```
> data(co2)
> ndiffs(co2,)
[1] 1
```

# Seasonal unit-root tests

- Seasonal unit-root tests are similar to non-seasonal unit-root tests:
  - For seasonal data, unit-root tests refer to the unit roots of the seasonal AR characteristic equation.
  - If seasonal roots exist, seasonal differencing is needed to obtain a stationary series.
- H&K propose to use the Osborn-Chui-Smith-Birchenhall (1988) test, with the null hypothesis that a seasonal unit root exists.
  - Approach: if the test result is not significant, the time series is "seasonally differenced".
- H&K recommend carrying out the test procedure for seasonal differencing **before** the test procedure for non-seasonal differencing.
  - E.g., if seasonal differencing is necessary, we would apply the non-seasonal unit-root test to the seasonal differenced series $\nabla_s^D Y_t$.

# Unit-root tests: CO2 data example

- R-function for the seasonal unit-root test: nsdiffs
  This function outputs the number of times the series should be seasonally differenced.
- For the CO2 data:

  ```
  > nsdiffs(co2,12)
  [1] 1
  ```

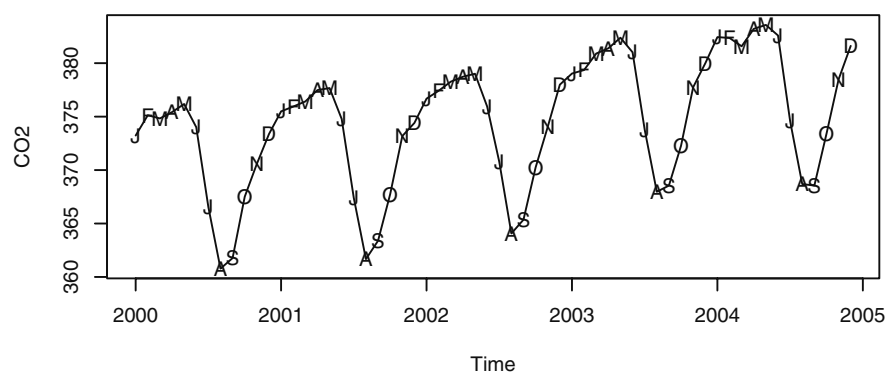
  thus we would use $W_t = Y_t - Y_{t-12}$.
- Do we still need non-seasonal differencing?

  ```
  > ndiffs(diff(co2,12))
  [1] 0
  ```
- Note that seasonal-nonseasonal testing sequence approach differs from the informal approach used in the book and results in seasonal differencing only!

# Time series model building: back to the overview

- What are the tasks involved in selecting candidate ARIMA$(p, d, q)$x$(P, D, Q)_s$ models, e.g. for the CO2 series?
  - Step 1: Select the order $d$ of non-seasonal and $D$ of seasonal differencing.
  - Step 2:
    - Select the orders $p, q, P, Q$ and decide whether or not the constant term $\theta_0$ should be included,
    - Decide whether or not "in-between" predictors (lagged $Y$'s or past white noise terms) should be removed.
- Steps 1 and 2 may result in a set of candidate models. If so, we need
  - Step 3: Which model to use for forecasting?

**Exhibit 10.2 Carbon Dioxide Levels with Monthly Symbols**



# Step 2: Determining $p, q, P, Q$ and inclusion of model terms

- Question to answer: What ARMA model to use for the (differenced) series?
- It is NOT recommended to simply choose orders arbitrarily large: you may obtain a good fit in-sample (small estimate for the white noise variance) but when the fitted model is then used for forecasting, you're more likely to make greater forecast errors.
- To overcome the problem of overfitting, model selection criteria are commonly used for selecting orders.
  - Not just in time series analysis, e.g. also in regression analysis!
- These criteria are based on a negative measure of model fit combined with a positive penalty term for the number of parameters in the model.
  - Derivation of criteria is outside class material; we focus on how to use them for model selection.
  - Question 1: Are models with lower OR higher values for the criteria preferred?

# Model selection using criteria

- Model selection based on model selection criteria corresponds to finding the model or models with the lowest or close-to-lowest criteria.
- A commonly used criterion in time series modeling is the AICc, the bias-corrected version of the AIC (the Akaike Information Criterion) proposed by Hurvich and Tsai (1989) (book p.131):

$$AICc = -2 \cdot \log(L) + 2k + \frac{2(k+1)(k+2)}{n-k-2},$$

where

- $L$ is the maximized likelihood function (plugging in the MLE estimates into the likelihood function),
- $n$ is the number of observations in the time series,
- $k$ is the number of parameters in the model, which is $p+q+P+Q$ if no constant term $\theta_0$ was included, $p+q+P+Q+1$ otherwise.

# How to use the AICc for model selection

- The "best" candidate model is the model with the lowest value for the criterion.
- Other simpler models are also considered as candidate models if differences between their criterion and that of the best model are small.
- Generally, if the "best" model (with the smallest value for the criteria) has $AICc = A$, models with $AICc < A + 2$ also considered as candidate models.

# Example

- Comparison of two models for CO2 data below (the first model is the one we selected informally in Ch 10.4).
- What information does the AICc provide us here?

```
> moda <- Arima(co2, order = c(0,1,1),
 seasonal = list(order = c(0,1,1), period = 12),
 method = "ML", include.drift = FALSE)
> summary(moda)
...
AIC=285.08   AICc=285.29   BIC=293.41

> modb <- Arima(co2, order = c(0,1,2),
 seasonal = list(order = c(0,1,1), period = 12),
 method = "ML", include.drift = FALSE)
> summary(modb)
...
AIC=287.05   AICc=287.4   BIC=298.16
```

# Model selection using criteria: BIC

- Another commonly used criterion is the Bayesian information criterion (BIC):

$$BIC = -2 \cdot \log(L) + k \log(n),$$

  with $L$, $k$ and $n$ as explained for the AICc.
- BIC is used in the same way as the AICc for selecting candidate models.
- What criterion should we use?

# Model selection using criteria: AICc or BIC?

- Unfortunately, there is no clear recommendation that works well for all data sets.
- The choice of what criterion to use (e.g., AICc or BIC) depends on the time series. E.g.
  - BIC tends to perform well in large samples w.r.t. getting the correct order if the true process indeed follows an ARMA model,
  - while the AICc enjoys the property that it will lead to an optimal model that is closest to the true process among the class of models under study (where closeness is measured in terms of the Kullback-Leibler divergence).
- The comparison of the two criteria is outside the material of this class.
- We will either choose one criterion for a given data set or apply both and consider all candidate models as indicated by both criteria.

# Automated model selection in R

- The function "auto.arima" in the "forecast" package provides an automated procedure.
- Based on default settings:
  - It selects $d$ and $D$ using the unit-root tests that we discussed.
  - It uses the AICc to select values for $p, q, P, Q$ (with maxima for $p$ and $q$ set at 5, and maxima for $P$ and $Q$ set at 2) and to decide whether to include or exclude $\theta_0$ if $d + D \leq 1$.
  - A "stepwise" procedure is used for finding the "best" model (see H&K reference), but for smaller datasets, we can carry out an exhaustive search.

# Example: CO2 data, automated selection using AICc

```
> auto.arima(co2,
+            stepwise = FALSE,
+            approximation= FALSE, # do NOT use an approximat
+            ic="aicc")
Series: co2
ARIMA(1,0,1)(0,1,1)[12] with drift

Coefficients:
         ar1      ma1     sma1   drift
      0.8349  -0.4630  -0.8487  0.1520
s.e.  0.0864   0.1313   0.1343  0.0055

sigma^2 estimated as 0.5537:  log likelihood=-122.48
AIC=254.96   AICc=255.48   BIC=268.9
```

- Set 'approximation' to FALSE to get exact calculations of the AICc.
- Resulting model: ARIMA(1,0,1)x$(1,1,1)_{12}$ with drift.

# Example: CO2 data, automated selection using BIC

```
> auto.arima(co2,
           stepwise = FALSE,  approximation= FALSE,
           ic="bic")
Series: co2
ARIMA(1,0,1)(0,1,1)[12] with drift

Coefficients:
         ar1      ma1     sma1   drift
      0.8349  -0.4630  -0.8487  0.1520
s.e.  0.0820   0.1246   0.1274  0.0052

sigma^2 estimated as 0.4983:  log likelihood=-136.09
AIC=282.18   AICc=282.7   BIC=296.11
```

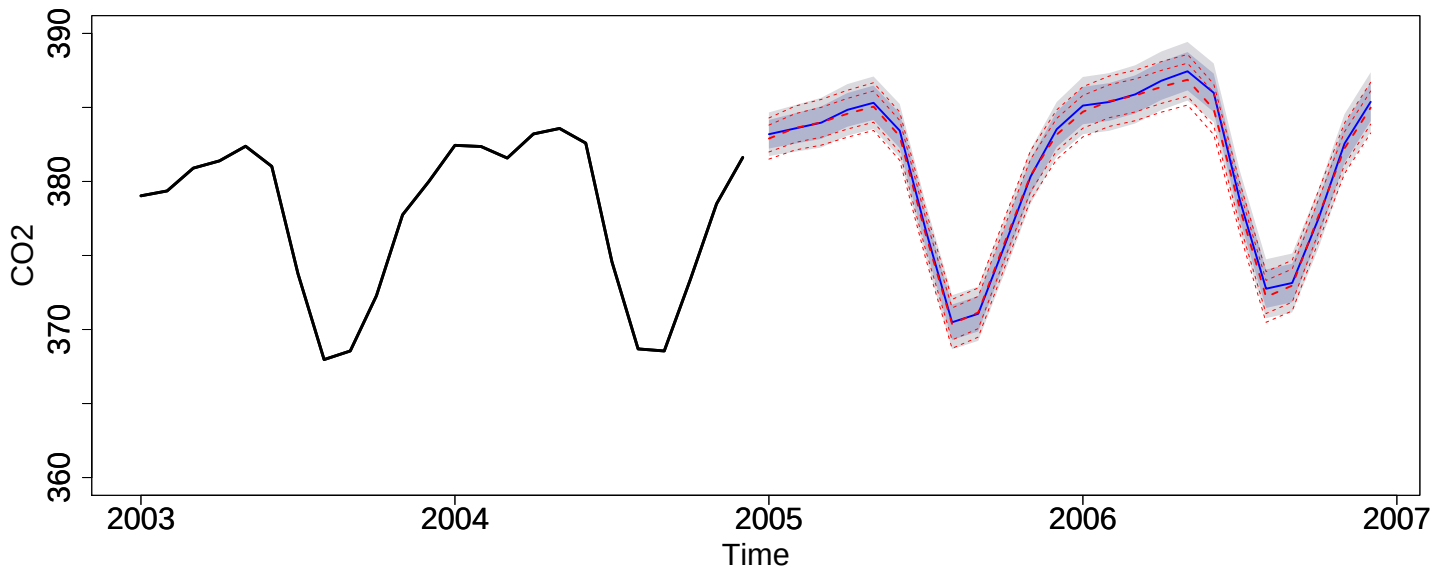
- Resulting model using BIC: ARIMA(1,0,1)x$(1,1,1)_{12}$ with drift, same as AICc here!
- What if we would have found two different models?

# Step 3: what candidate model to use for forecasting?

- Some things to consider when comparing candidate models are
  - Which has the "best" diagnostics?
    Candidate models for which diagnostic checking reveals potential issues (e.g., outliers may be present) are less preferable than candidate models where no problems are detected.
  - Do they make similar predictions?
- If selected models with acceptable diagnostics give very similar forecasts, and forecasting is your goal, any of the candidate models will do (e.g., choose the simplest one).
- However, if selected models make very different predictions, there is simply a lot of uncertainty about which model to use.
  - In this situation, more advanced techniques can be used, that deal with combining the forecasts of several models (outside class material).
  - If you are working with a long time series, cross validation exercises, whereby part of the series is left out to select a model based on the partial series, can be considered, to compare forecast errors for the left-out series across selected models.

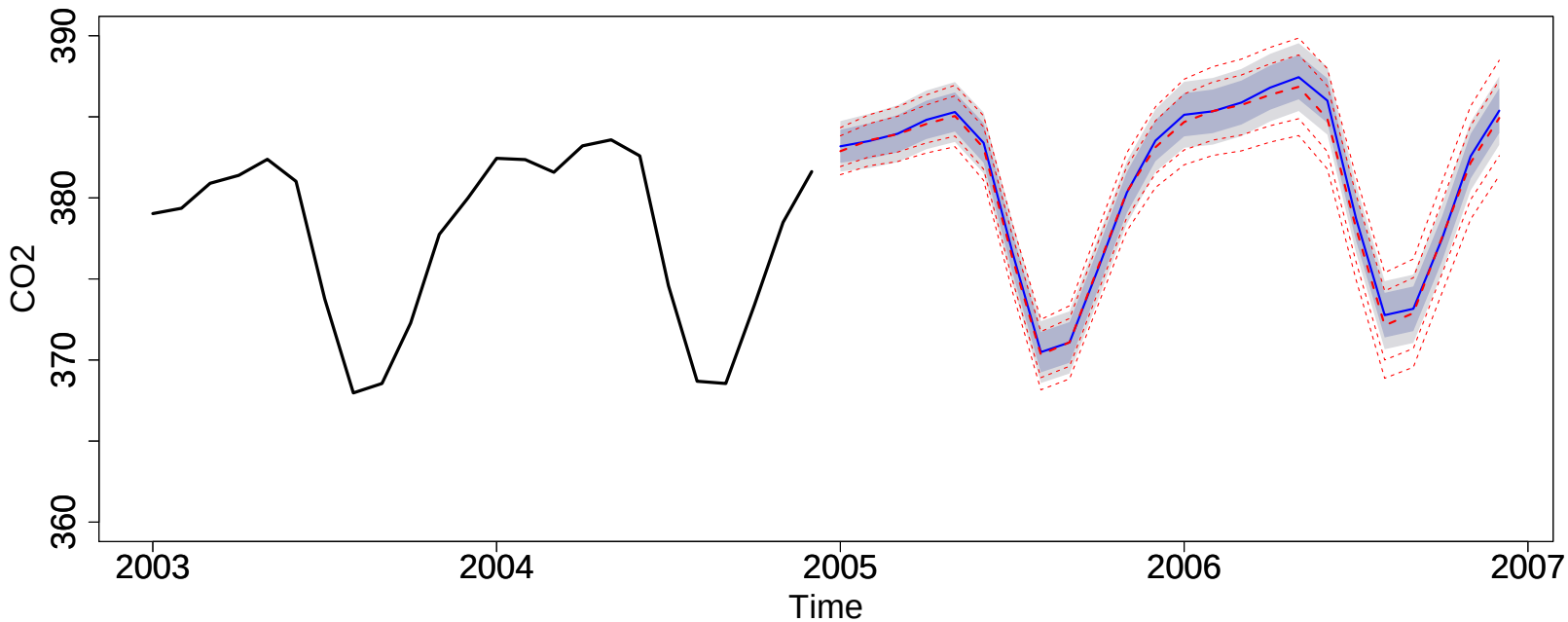
# CO2 data: comparison of forecasts of different models

- Based on using an approximation for the AICc, a different model was identified: ARIMA(3,0,0)x$(2,1,1)_{12}$ with drift. How to compare both models?
- Yet to do: checking of model diagnostics.
- Comparison of the forecasts of the two models in plot below, where: Blue = appr. AICc model, Red = BICc model.
- Conclusion?



# CO2 data: comparison of forecasts

- Note that the AICc and BIC models were different from the ARIMA(0,1,1)x$(0,1,1)_{12}$ model we used in Ch 10.4 through informal selection.
- Comparison of AICc (blue) and "informal" model (red) below:
- Point forecasts are very similar, the informally selected model just gives wider PIs.



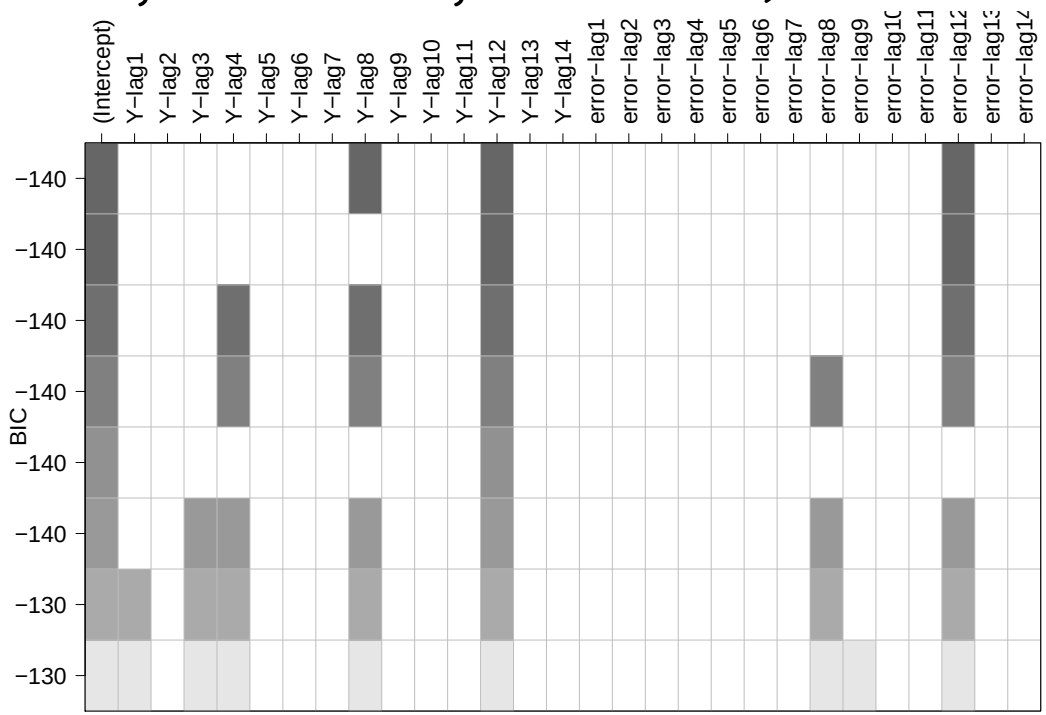
# Further model selection options

- So far, we discussed using AICc or BIC for selecting orders $p, q, P, Q$ and deciding whether to include or exclude $\theta_0$ in an (seasonal) ARIMA model.
- Model selection criteria can also be used to find the subset of non-zero coefficients for the candidate variables (lagged $Y_t$'s and past white noise terms).
- One automated approach is implemented for ARMA$(p, q)$ models in the armasubset function in the TSA library. A nice overview plot is produced when using this function.
- Two examples follow (one for a simulated data set and one for a real data set).
  - Example for CO2 in code.
- Note however that the function seems to always include an intercept into the model, so it is not helpful when exploring models for which you would like to fix the mean at 0.

# Example of subset selection (p132-133 book)

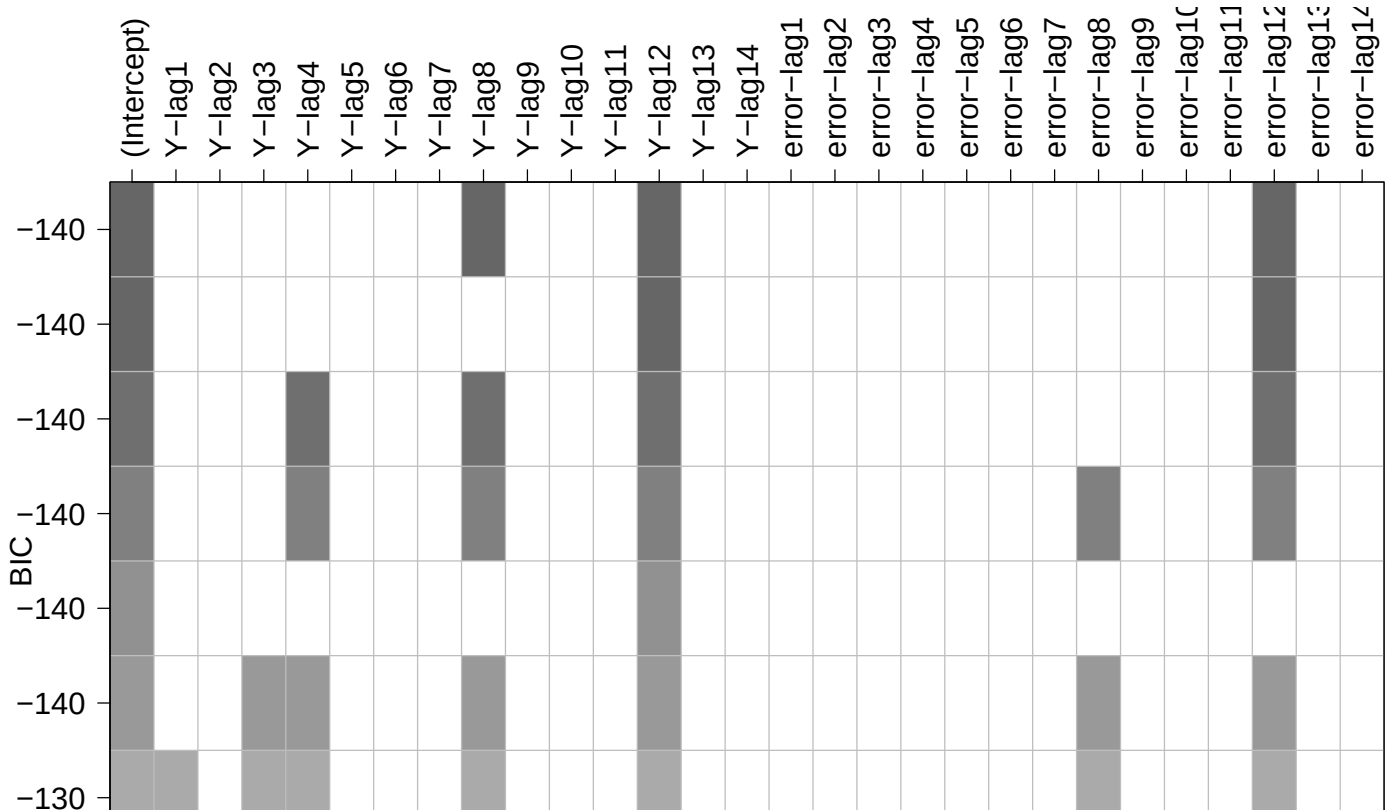
- Time series "test" considered is a simulated ARMA(12,12) model with $\phi_i = \theta_j = 0$ for $i, j \neq 12$: $Y_t = \phi_{12} Y_{t-12} + e_t - \theta_{12} e_{t-12}$.
- Each row in the overview plot corresponds to a subset ARMA model where the cells of the variables selected for the model are shaded.
- The models are sorted according to their BIC, with better models (lower BIC) placed in higher rows and with darker shades.

```
armasubsets(test, nar = 14, nma = 14)
```



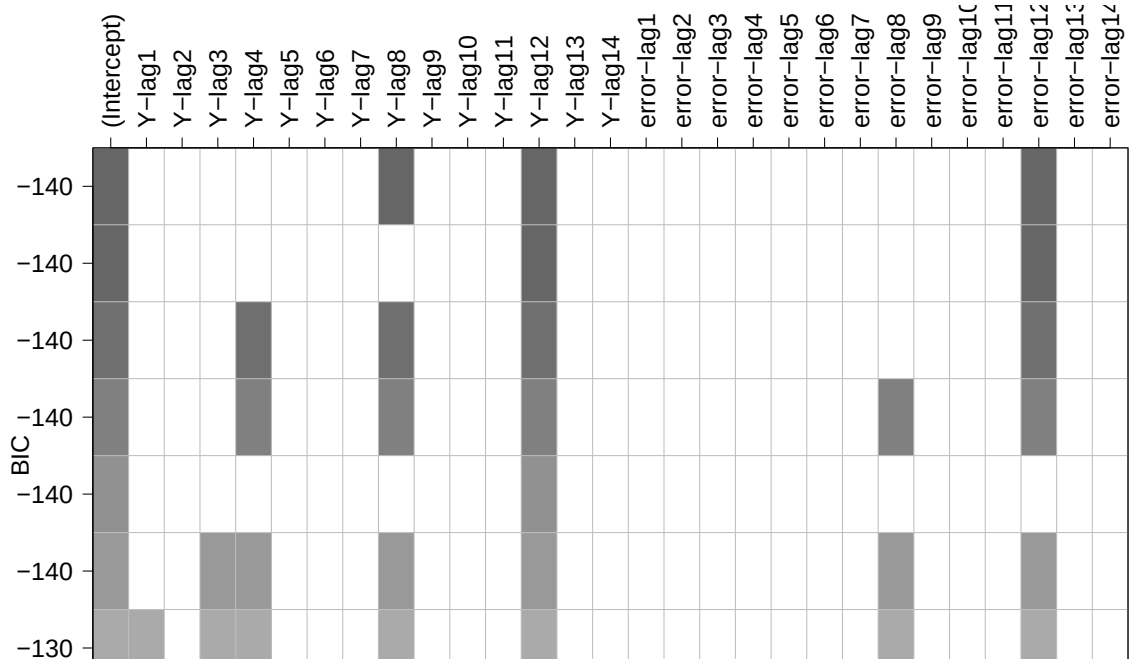
# Subset selection: interpreting the display

- The top row tells us that the subset ARMA(14,14) model with the smallest BIC contains only lags 8 and 12 of the observed time series and lag 12 of the error process.
- The next best model contains lag 12 of the time series and lag 12 of the errors (the true model), while the third best model contains lags 4, 8, and 12 of the time series and lag 12 of the errors.



# Subset selection: selecting candidate models

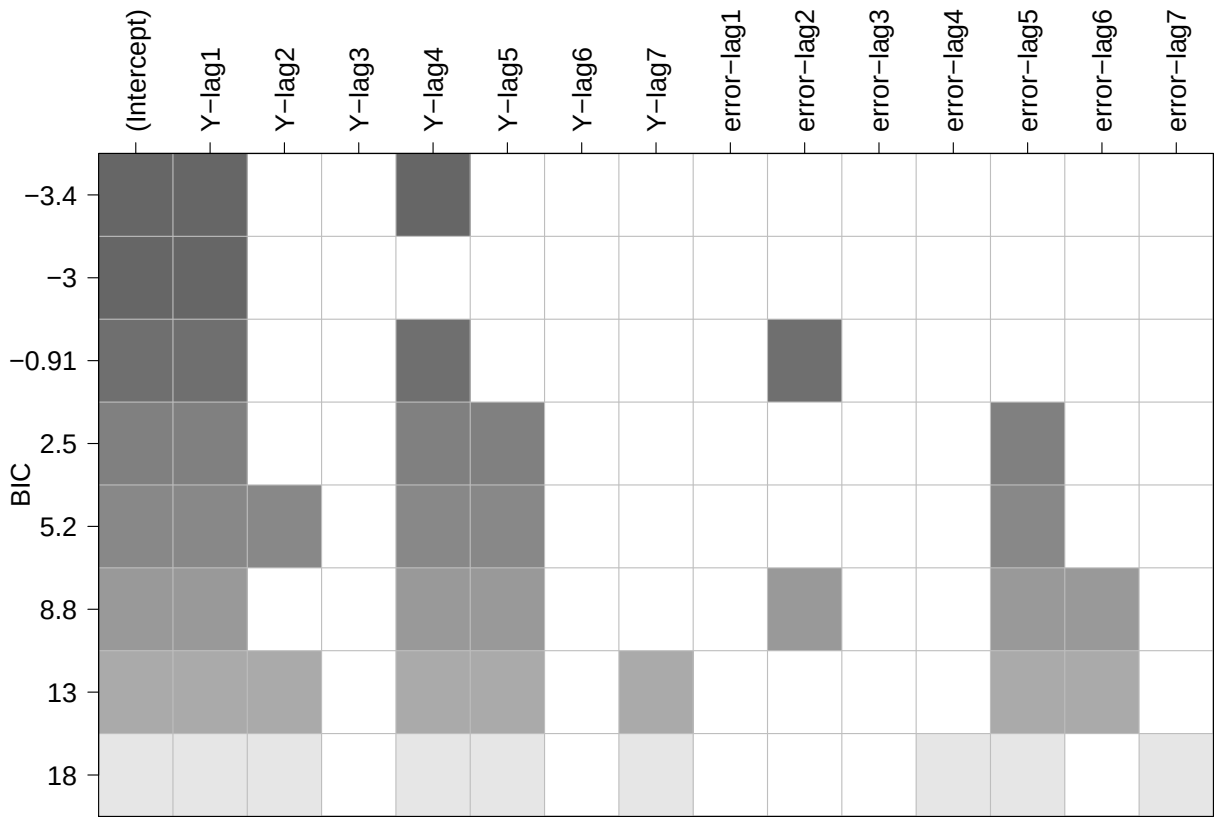
- If this were a display for a real data series, how to continue?
- The BIC values for the first 6 models are all very similar, so would be worthy of further study.
- Also note that $Y_{t-12}$ and $e_{t-12}$ are the two variables most frequently found in the various subset models. Such a finding suggests that they may be more important variables (as they indeed are for this simulation!) so make sure to explore candidate model(s) that include those predictors.



# Another example: oil price data (p.139)

- We discussed in Ch.5 that taking a log-transform and (non-seasonal) differencing seemed appropriate for the oil price series.
- What variables should be included?

```
armasubsets(diff(log(oil.price)), nar = 7, nma = 7)
```



# How to fit an ARMA model with coefficients that are fixed at 0?

- The argument "fixed" can be used, with 0s added for the parameters that are 0, NA otherwise.
- Note that the parameters are ordered as follows: (AR(p), MA(q) AR(P), AR(Q), drift/intercept)
- Example for oil price series:

```
mod <- Arima(log(oil.price), order = c(4,1,0),
             fixed=c(NA,0,0,NA,NA),
             transform.pars = FALSE,
             include.drift = TRUE)
```

```
> summary(mod)
ARIMA(4,1,0) with drift
         ar1  ar2  ar3      ar4   drift
      0.2335    0    0  -0.0863  0.0043
s.e.  0.0659    0    0   0.0672  0.0062
```

# Summary of model selection approaches

- When selecting candidate ARIMA$(p, d, q)$x$(P, D, Q)_s$ models:
  - Unit-root tests can be used to determine whether (seasonal) differencing is necessary.
  - Model selection criteria, such as AICc and BIC can be used to select seasonal ARIMA models, or to select subsets of variables for ARMA models.
- As a default approach in R, we can
  - Use the auto.arima function to obtain a candidate model via AICc/BIC.
  - Use armasubsets to look into other candidate models.
- When working with real time series data, most likely, this may lead to several candidate model. If so, consider if models differ w.r.t.
  - model diagnostics,
  - forecasts.