Rob J Hyndman

Forecasting using R

11. Dynamic regression

OTexts.com/fpp/9/1/

Outline

Regression with ARIMA errors

Regression models

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

- y_t modeled as function of k explanatory variables $x_{1,t}, \dots, x_{k,t}$.
- Usually, we assume that n_t is WN.
- Now we want to allow n_t to be autocorrelated.

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

and e_t is white noise .

Regression with ARIMA errors

Regression models

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

- y_t modeled as function of k explanatory variables $x_{1,t}, \ldots, x_{k,t}$.
- Usually, we assume that n_t is WN.
- Now we want to allow n_t to be autocorrelated.

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

where $\quad (1 - \phi_1 B)(1 - B) n_t = (1 - \theta_1 B) e_t$

and e_t is white noise .

Regression with ARIMA errors

Regression models

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

- y_t modeled as function of k explanatory variables $x_{1,t}, \ldots, x_{k,t}$.
- Usually, we assume that n_t is WN.
- Now we want to allow n_t to be autocorrelated.

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

where $(1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$

and e_t is white noise .

Regression with ARIMA errors

Regression models

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

- y_t modeled as function of k explanatory variables $x_{1,t}, \ldots, x_{k,t}$.
- Usually, we assume that n_t is WN.
- Now we want to allow n_t to be autocorrelated.

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

where $\quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$

and e_t is white noise .

Regression with ARIMA errors

Regression models

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

- y_t modeled as function of k explanatory variables $x_{1,t}, \ldots, x_{k,t}$.
- Usually, we assume that n_t is WN.
- Now we want to allow n_t to be autocorrelated.

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

and e_t is white noise .

Residuals and errors

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

- Be careful in distinguishing n_t from e_t.
- n_t are the "*errors*" and e_t are the "*residuals*".
- In ordinary regression, n_t is assumed to be white noise and so $n_t = e_t$.

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

Now a regression with ARMA(1,1) error

Residuals and errors

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$
$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

- Be careful in distinguishing n_t from e_t.
- n_t are the "*errors*" and e_t are the "*residuals*".
- In ordinary regression, n_t is assumed to be white noise and so $n_t = e_t$.

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

Now a regression with ARMA(1,1) error

Residuals and errors

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

- Be careful in distinguishing n_t from e_t.
- n_t are the "*errors*" and e_t are the "*residuals*".
- In ordinary regression, n_t is assumed to be white noise and so $n_t = e_t$.

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

Now a regression with ARMA(1,1) error

Residuals and errors

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

- Be careful in distinguishing n_t from e_t.
- n_t are the "*errors*" and e_t are the "*residuals*".
- In ordinary regression, n_t is assumed to be white noise and so $n_t = e_t$.

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

Now a regression with ARMA(1,1) error

Residuals and errors

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$
$$\text{where} \quad (1 - \phi_1 B)(1 - B)n_t = (1 - \theta_1 B)e_t$$

- Be careful in distinguishing n_t from e_t.
- n_t are the "*errors*" and e_t are the "*residuals*".
- In ordinary regression, n_t is assumed to be white noise and so $n_t = e_t$.

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

Now a regression with ARMA(1,1) error

Residuals and errors

Example: n_t = ARIMA(1,1,1)

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$
$$\text{where} \quad (1-\phi_1 B)(1-B)n_t = (1-\theta_1 B)e_t$$

- Be careful in distinguishing n_t from e_t.
- n_t are the "*errors*" and e_t are the "*residuals*".
- In ordinary regression, n_t is assumed to be white noise and so $n_t = e_t$.

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

Now a regression with ARMA(1,1) error

Regression with ARIMA errors

Any regression with an ARIMA error can be rewritten as a regression with an ARMA error by differencing all variables with the same differencing operator as in the ARIMA model.

Original data

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad \phi(B)(1-B)^d n_t = \theta(B) e_t$$

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

$$\text{where} \quad \phi(B) n_t = \theta(B) e_t$$

$$\text{and} \quad y'_t = (1-B)^d y_t, \quad \text{etc.}$$

Regression with ARIMA errors

Any regression with an ARIMA error can be rewritten as a regression with an ARMA error by differencing all variables with the same differencing operator as in the ARIMA model.

Original data

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad \phi(B)(1-B)^d n_t = \theta(B) e_t$$

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

$$\text{where} \quad \phi(B) n_t = \theta(B) e_t$$

$$\text{and} \quad y'_t = (1-B)^d y_t, \quad \text{etc.}$$

Regression with ARIMA errors

Any regression with an ARIMA error can be rewritten as a regression with an ARMA error by differencing all variables with the same differencing operator as in the ARIMA model.

Original data

$$y_t = b_0 + b_1 x_{1,t} + \cdots + b_k x_{k,t} + n_t$$

$$\text{where} \quad \phi(B)(1-B)^d n_t = \theta(B) e_t$$

After differencing all variables

$$y'_t = b_1 x'_{1,t} + \cdots + b_k x'_{k,t} + n'_t.$$

$$\text{where} \quad \phi(B) n_t = \theta(B) e_t$$

$$\text{and} \quad y'_t = (1-B)^d y_t, \quad \text{etc.}$$

Modeling procedure

Problems with OLS and autocorrelated errors

1. OLS no longer the best way to compute coefficients as it does not take account of time-relationships in data.
2. Standard errors of coefficients are incorrect — most likely too small. This invalidates tests and prediction intervals.

- Second problem more serious because it can lead to misleading results.

Modeling procedure

Problems with OLS and autocorrelated errors

1. OLS no longer the best way to compute coefficients as it does not take account of time-relationships in data.
2. Standard errors of coefficients are incorrect — most likely too small. This invalidates tests and prediction intervals.

- Second problem more serious because it can lead to misleading results.
- If standard errors obtained using OLS are used, some explanatory variables may appear to be significant when, in fact, they are not. This is known as "spurious regression."

Modeling procedure

Problems with OLS and autocorrelated errors

1. OLS no longer the best way to compute coefficients as it does not take account of time-relationships in data.
2. Standard errors of coefficients are incorrect — most likely too small. This invalidates tests and prediction intervals.

- Second problem more serious because it can lead to misleading results.
- If standard errors obtained using OLS too small, some explanatory variables may appear to be significant when, in fact, they are not. This is known as “spurious regression.”

Modeling procedure

Problems with OLS and autocorrelated errors

1. OLS no longer the best way to compute coefficients as it does not take account of time-relationships in data.
2. Standard errors of coefficients are incorrect — most likely too small. This invalidates tests and prediction intervals.

- Second problem more serious because it can lead to misleading results.
- If standard errors obtained using OLS too small, some explanatory variables may appear to be significant when, in fact, they are not. This is known as “spurious regression.”

Modeling procedure

Problems with OLS and autocorrelated errors

1. OLS no longer the best way to compute coefficients as it does not take account of time-relationships in data.
2. Standard errors of coefficients are incorrect — most likely too small. This invalidates tests and prediction intervals.

- Second problem more serious because it can lead to misleading results.
- If standard errors obtained using OLS too small, some explanatory variables may appear to be significant when, in fact, they are not. This is known as “spurious regression.”

Modeling procedure

- Estimation only works when all predictor variables are deterministic or stationary and the errors are stationary.
- So difference stochastic variables as required until all variables appear stationary. Then fit model with ARMA errors.
- `auto.arima()` will handle order selection and differencing (but only checks that errors are stationary).

Modeling procedure

- Estimation only works when all predictor variables are deterministic or stationary and the errors are stationary.
- So difference stochastic variables as required until all variables appear stationary. Then fit model with ARMA errors.
- `auto.arima()` will handle order selection and differencing (but only checks that errors are stationary).

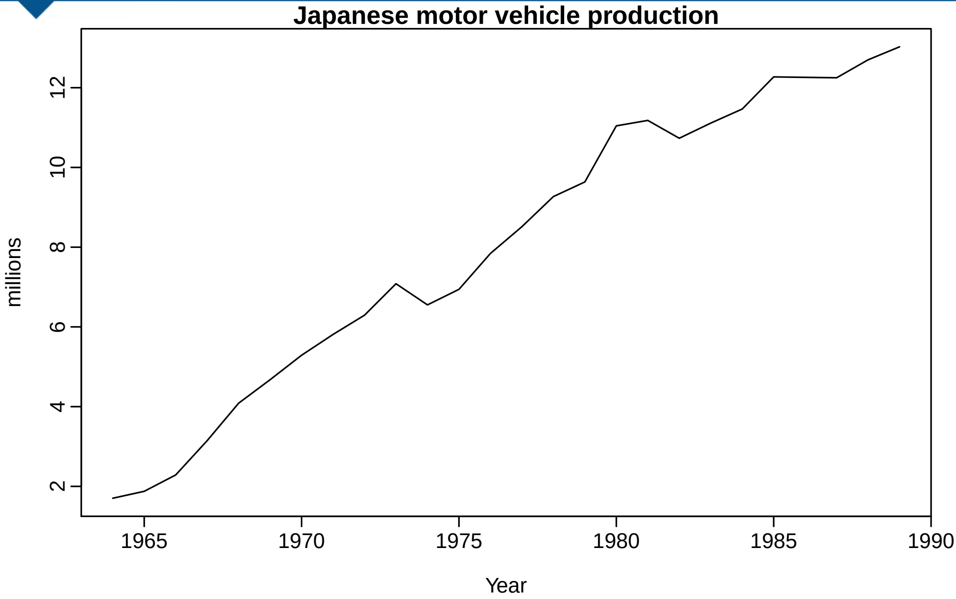
Modeling procedure

- Estimation only works when all predictor variables are deterministic or stationary and the errors are stationary.
- So difference stochastic variables as required until all variables appear stationary. Then fit model with ARMA errors.
- `auto.arima()` will handle order selection and differencing (but only checks that errors are stationary).

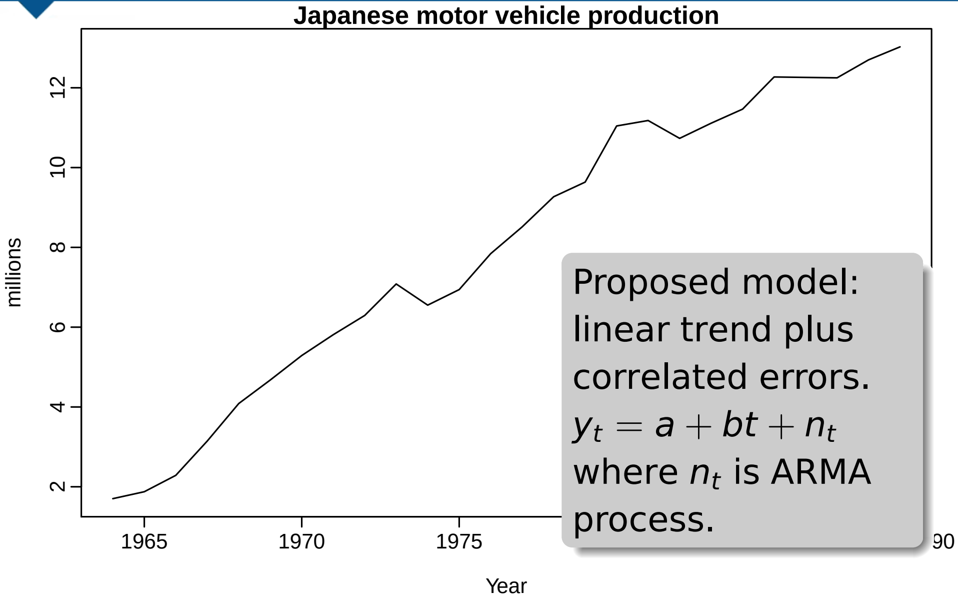
Outline

1. Regression with ARIMA errors
2. **Example: Japanese cars**
3. Using Fourier terms for seasonality
4. Example: Sales of petroleum & coal products

Example: Japanese cars



Example: Japanese cars

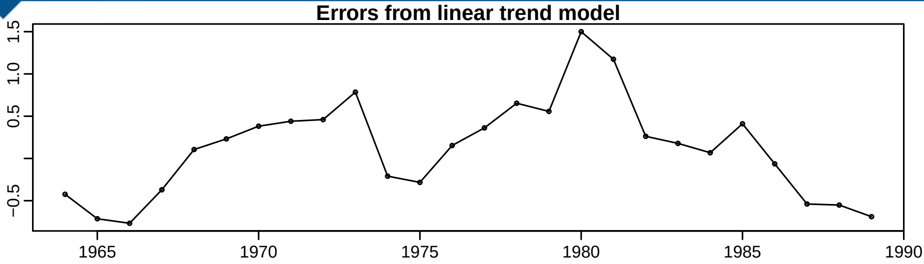


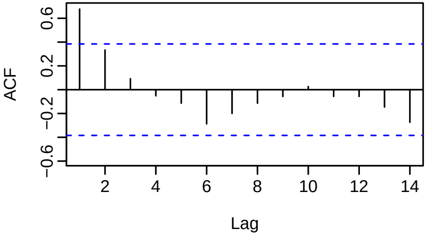
Proposed model: linear trend plus correlated errors.

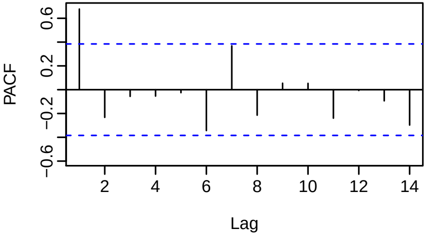
$y_t = a + bt + n_t$

where n_t is ARMA process.

Example: Japanese cars







Example: Japanese cars

1. We will fit a linear trend model:

$$y_t = a + bx_t + n_t$$

where $x_t = t - 1963$.

2. `auto.arima` chooses an AR(1) model for n_t.
3. Full model is

$$y_t = a + bx_t + n_t \qquad \text{where} \quad n_t = \phi_1 n_{t-1} + e_t$$

and e_t is a white noise series.

Example: Japanese cars

1. We will fit a linear trend model:

$$y_t = a + bx_t + n_t$$

   where $x_t = t - 1963$.

2. `auto.arima` chooses an AR(1) model for n_t.
3. Full model is

$$y_t = a + bx_t + n_t \quad \text{where} \quad n_t = \phi_1 n_{t-1} + e_t$$

   and e_t is a white noise series.

Example: Japanese cars

1. We will fit a linear trend model:

   $$y_t = a + bx_t + n_t$$

   where $x_t = t - 1963$.
2. `auto.arima` chooses an AR(1) model for n_t.
3. Full model is

   $$y_t = a + bx_t + n_t \qquad \text{where} \quad n_t = \phi_1 n_{t-1} + e_t$$

   and e_t is a white noise series.

Example: Japanese cars

Parameter		Estimate	Standard Error
AR(1)	ϕ_1	0.736	0.152
Intercept	a	1662.	504.1
Slope	b	463.6	32.31

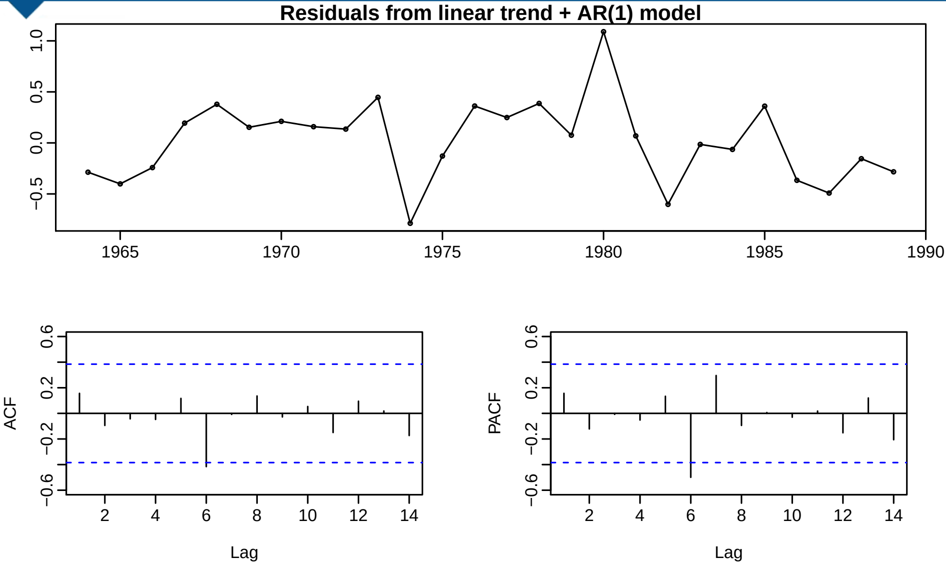
- We need to check that the residuals (e_t) look like white noise.

Example: Japanese cars

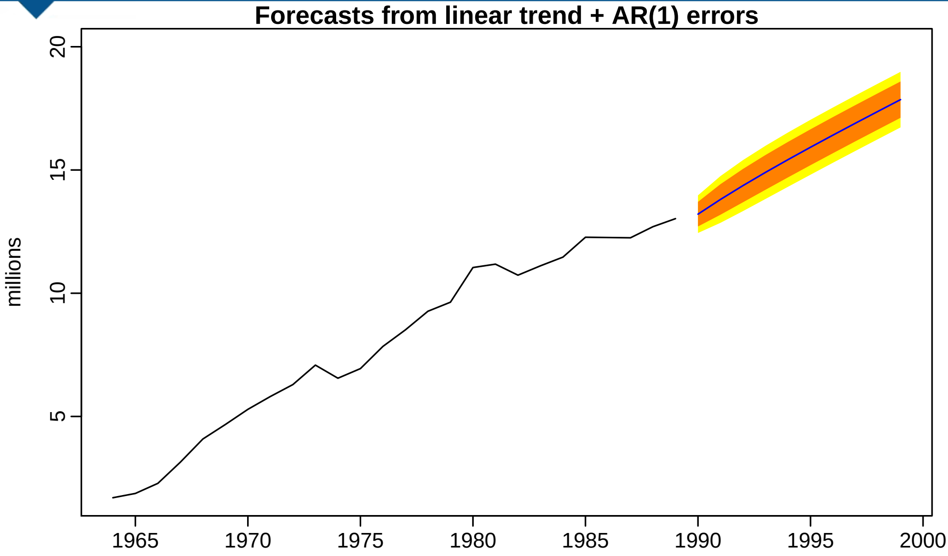
Parameter		Estimate	Standard Error
AR(1)	ϕ_1	0.736	0.152
Intercept	a	1662.	504.1
Slope	b	463.6	32.31

- We need to check that the residuals (e_t) look like white noise.

Example: Japanese cars



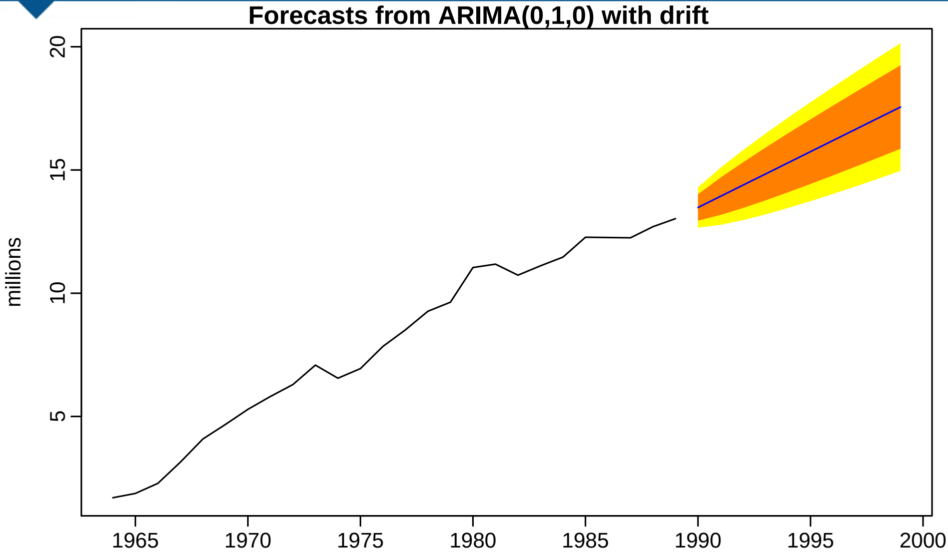
Example: Japanese cars



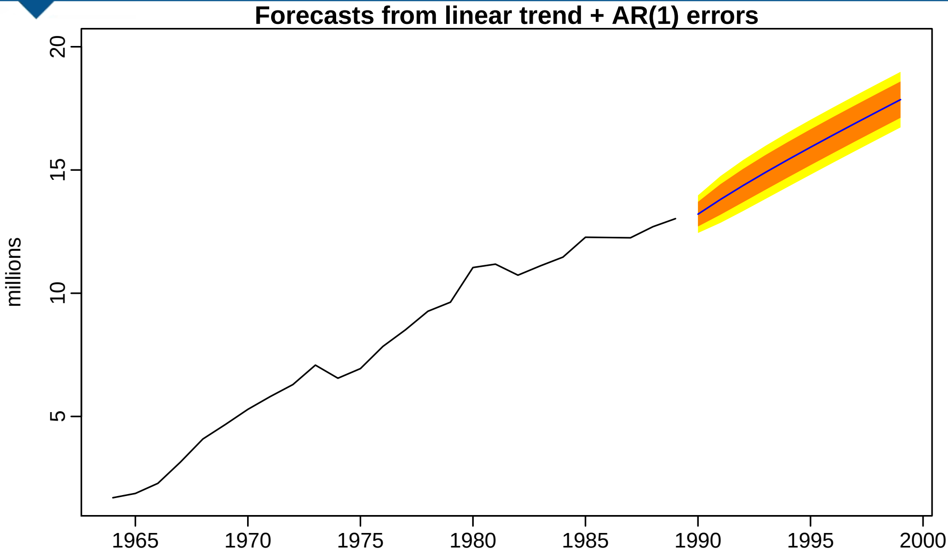
Example: Japanese cars

```
> fit2 <- auto.arima(x)
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.4530
s.e.  0.0836

sigma^2 estimated as 0.1749:
log likelihood = -13.68
AIC = 31.36   AICc = 31.9   BIC = 33.8
```

Example: Japanese cars



Example: Japanese cars



Outline

1. Regression with ARIMA errors
2. Example: Japanese cars
3. **Using Fourier terms for seasonality**
4. Example: Sales of petroleum & coal products

Fourier series

Any periodic function of period m can be approximated using

$$\sum_{k=1}^{K} \left\{ \sin\left(\frac{2\pi kt}{m}\right) + \cos\left(\frac{2\pi kt}{m}\right) \right\}$$

- The approximation can be made exact with large K.
- Each k represents a harmonic.
- `fourier` generates a matrix of Fourier terms.

Example

```
fit <- auto.arima(y, seasonal=FALSE,
  xreg=fourier(y, K=3))
```

Fourier series

Any periodic function of period m can be approximated using

$$\sum_{k=1}^{K} \left\{ \sin\left(\frac{2\pi kt}{m}\right) + \cos\left(\frac{2\pi kt}{m}\right) \right\}$$

- The approximation can be made exact with large K.
- Each k represents a harmonic.
- `fourier` generates a matrix of Fourier terms.

Example

```
fit <- auto.arima(y, seasonal=FALSE,
  xreg=fourier(y, K=3))
```

Fourier series

Any periodic function of period m can be approximated using

$$\sum_{k=1}^{K} \left\{ \sin\left(\frac{2\pi kt}{m}\right) + \cos\left(\frac{2\pi kt}{m}\right) \right\}$$

- The approximation can be made exact with large K.
- Each k represents a harmonic.
- `fourier` generates a matrix of Fourier terms.

Example

```
fit <- auto.arima(y, seasonal=FALSE,
  xreg=fourier(y, K=3))
```

Fourier series

Any periodic function of period m can be approximated using

$$\sum_{k=1}^{K} \left\{ \sin\left(\frac{2\pi kt}{m}\right) + \cos\left(\frac{2\pi kt}{m}\right) \right\}$$

- The approximation can be made exact with large K.
- Each k represents a harmonic.
- `fourier` generates a matrix of Fourier terms.

Example

```
fit <- auto.arima(y, seasonal=FALSE,
  xreg=fourier(y, K=3))
```

Fourier series

Any periodic function of period m can be approximated using

$$\sum_{k=1}^{K} \left\{ \sin\left(\frac{2\pi k t}{m}\right) + \cos\left(\frac{2\pi k t}{m}\right) \right\}$$

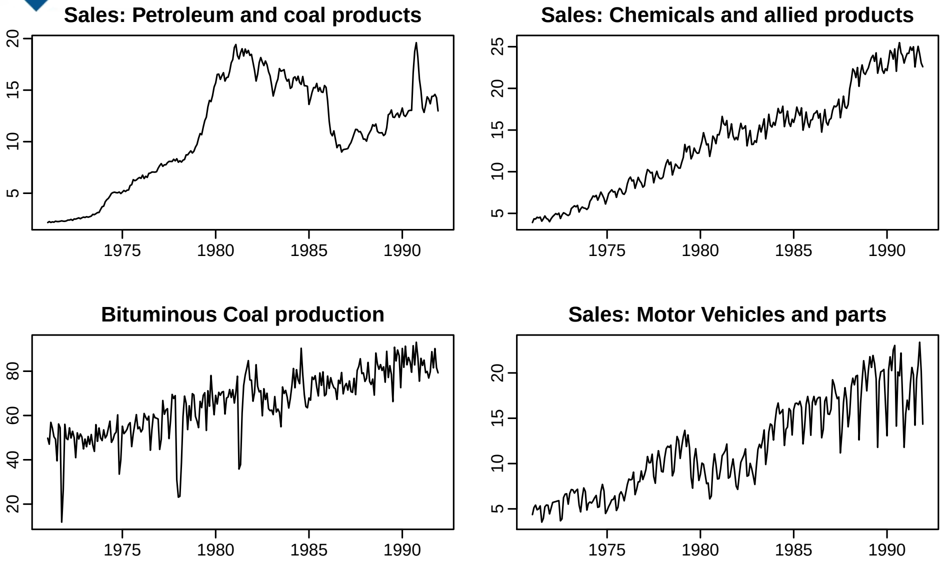
- The approximation can be made exact with large K.
- Each k represents a harmonic.
- `fourier` generates a matrix of Fourier terms.

Example

```
fit <- auto.arima(y, seasonal=FALSE,
  xreg=fourier(y, K=3))
```

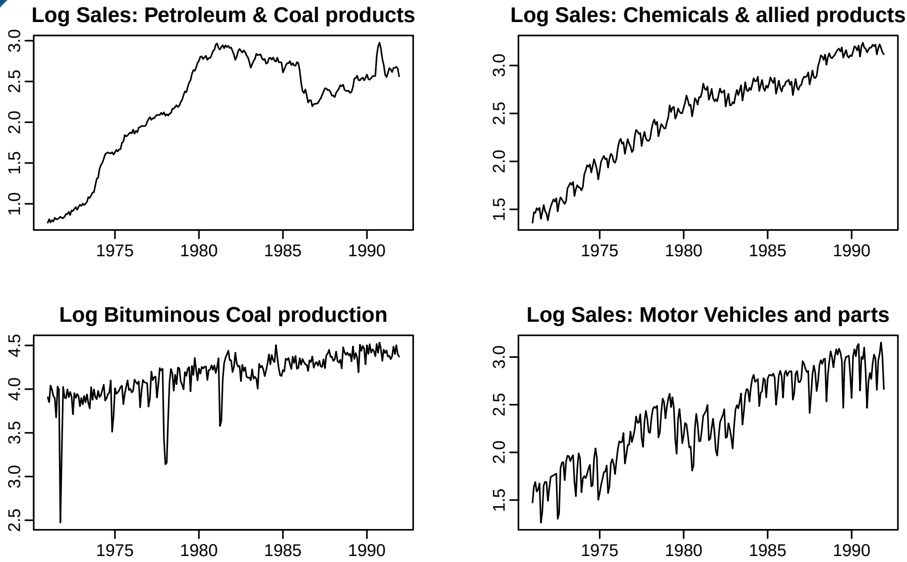
Outline

Example: Sales of petroleum & coal products

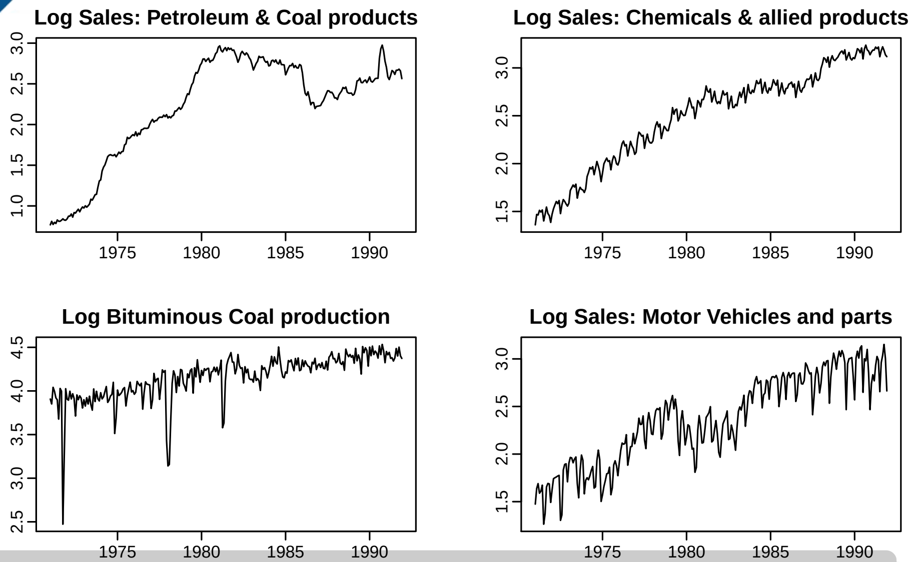


U.S. monthly sales: Jan 1971 – Dec 1991.

Example: Sales of petroleum & coal products



Example: Sales of petroleum & coal products



Series clearly non-stationary, so difference.

Example: Sales of petroleum & coal products

$$y'_t = b_1 x'_{1,t} + b_2 x'_{2,t} + b_3 x'_{3,t} + n_t$$

- $y_t =$ log petroleum sales
- $x_{1,t} =$ log chemical sales
- $x_{2,t} =$ log coal production
- $x_{3,t} =$ log motor vehicle and parts sales.

- `auto.arima` selects an ARIMA(0,0,1)(1,0,2)$_{12}$ model for n_t.
- Full model is $y_t = b_1 x_{1,t} + b_2 x_{2,t} + b_3 x_{3,t} + n_t$

$$(1-\Phi_1 B^{12})(1-B)n_t = (1+\theta_1 B)(1+\Theta_1 B^{12}+\Theta_2 B^{24})e_t.$$

Example: Sales of petroleum & coal products

$$y'_t = b_1 x'_{1,t} + b_2 x'_{2,t} + b_3 x'_{3,t} + n_t$$

- y_t = log petroleum sales
- $x_{1,t}$ = log chemical sales
- $x_{2,t}$ = log coal production
- $x_{3,t}$ = log motor vehicle and parts sales.

- `auto.arima` selects an ARIMA$(0,0,1)(1,0,2)_{12}$ model for n_t.
- Full model is $y_t = b_1 x_{1,t} + b_2 x_{2,t} + b_3 x_{3,t} + n_t$

  $(1-\Phi_1 B^{12})(1-B)n_t = (1+\theta_1 B)(1+\Theta_1 B^{12}+\Theta_2 B^{24})e_t.$

Example: Sales of petroleum & coal products

$$y'_t = b_1 x'_{1,t} + b_2 x'_{2,t} + b_3 x'_{3,t} + n_t$$

- $y_t =$ log petroleum sales
- $x_{1,t} =$ log chemical sales
- $x_{2,t} =$ log coal production
- $x_{3,t} =$ log motor vehicle and parts sales.

- `auto.arima` selects an ARIMA(0,0,1)(1,0,2)$_{12}$ model for n_t.
- Full model is $y_t = b_1 x_{1,t} + b_2 x_{2,t} + b_3 x_{3,t} + n_t$

$$(1-\Phi_1 B^{12})(1-B)n_t = (1+\theta_1 B)(1+\Theta_1 B^{12}+\Theta_2 B^{24})e_t.$$

Example: Sales of petroleum & coal products

Parameter		Estimate	S.E
MA(1)	θ_1	0.331	0.058
Seasonal AR(1)	Φ_1	0.916	0.053
Seasonal MA(1)	Θ_1	-0.661	0.086
Seasonal MA(2)	Θ_2	-0.116	0.068
Log Chemicals	b_1	0.296	0.065
Log Coal	b_2	-0.029	0.013
Log Vehicles	b_3	-0.014	0.023

- AIC $= -913.8$
- Consider dropping motor vehicles and parts variable: AIC$= -915.6$.

Example: Sales of petroleum & coal products

Parameter		Estimate	S.E
MA(1)	θ_1	0.331	0.058
Seasonal AR(1)	Φ_1	0.916	0.053
Seasonal MA(1)	Θ_1	-0.661	0.086
Seasonal MA(2)	Θ_2	-0.116	0.068
Log Chemicals	b_1	0.296	0.065
Log Coal	b_2	-0.029	0.013
Log Vehicles	b_3	-0.014	0.023

- AIC $= -913.8$
- Consider dropping motor vehicles and parts variable: AIC$= -915.6$.

Forecasting

- To forecast a regression model with ARIMA errors, we need to forecast the regression part of the model and the ARIMA part of the model and combine the results.
- When future predictors are unknown, they need to be forecast first.
- Forecasts of macroeconomic variables may be obtained from the national statistical offices.
- Separate forecasting models may be needed for other explanatory variables.
- Some explanatory variable are known into the future (e.g., time, dummies).

Forecasting

- To forecast a regression model with ARIMA errors, we need to forecast the regression part of the model and the ARIMA part of the model and combine the results.
- When future predictors are unknown, they need to be forecast first.
- Forecasts of macroeconomic variables may be obtained from the national statistical offices.
- Separate forecasting models may be needed for other explanatory variables.
- Some explanatory variable are known into the future (e.g., time, dummies).

Forecasting

- To forecast a regression model with ARIMA errors, we need to forecast the regression part of the model and the ARIMA part of the model and combine the results.
- When future predictors are unknown, they need to be forecast first.
- Forecasts of macroeconomic variables may be obtained from the national statistical offices.
- Separate forecasting models may be needed for other explanatory variables.
- Some explanatory variable are known into the future (e.g., time, dummies).

Forecasting

- To forecast a regression model with ARIMA errors, we need to forecast the regression part of the model and the ARIMA part of the model and combine the results.
- When future predictors are unknown, they need to be forecast first.
- Forecasts of macroeconomic variables may be obtained from the national statistical offices.
- Separate forecasting models may be needed for other explanatory variables.
- Some explanatory variable are known into the future (e.g., time, dummies).

Forecasting

- To forecast a regression model with ARIMA errors, we need to forecast the regression part of the model and the ARIMA part of the model and combine the results.
- When future predictors are unknown, they need to be forecast first.
- Forecasts of macroeconomic variables may be obtained from the national statistical offices.
- Separate forecasting models may be needed for other explanatory variables.
- Some explanatory variable are known into the future (e.g., time, dummies).