Dynamic Regression Models and Time Series

Jerome Dumortier

17 August 2023

Overview

Packages

Required packages:

- forecast
- Hmisc
- stargazer

Introduction

Dynamic regression and time series topics:

- Trends and seasonality
- Distributed-lag models (including past or lagged independent variables), e.g.,

$$y_t = \alpha + \beta_0 \cdot x_t + \beta_1 \cdot x_{t-1} + \beta_2 \cdot x_{t-2} + \epsilon$$

- Autoregressive model relating present value of a time series to past values and errors (univariate time series), e.g.,

$$y_t = \beta_0 + \beta_1 \cdot y_{t-1} + \epsilon$$

- Forecasting

Trend and Seasonality

Overview

Decomposition of data over time into three components:

- Trend
- Season
- Random component

Trend

- Linear time trend: $y_t = \beta_0 + \beta_1 \cdot t + \epsilon_t$
- Exponential time trend: $\ln(y_t) = \beta_0 + \beta_1 \cdot t + \epsilon_t$
- β_1 in the exponential time trend model is the average annual growth rate (assuming t is in years)

Inclusion of a seasonal component via (quarterly in this case) dummy variables:

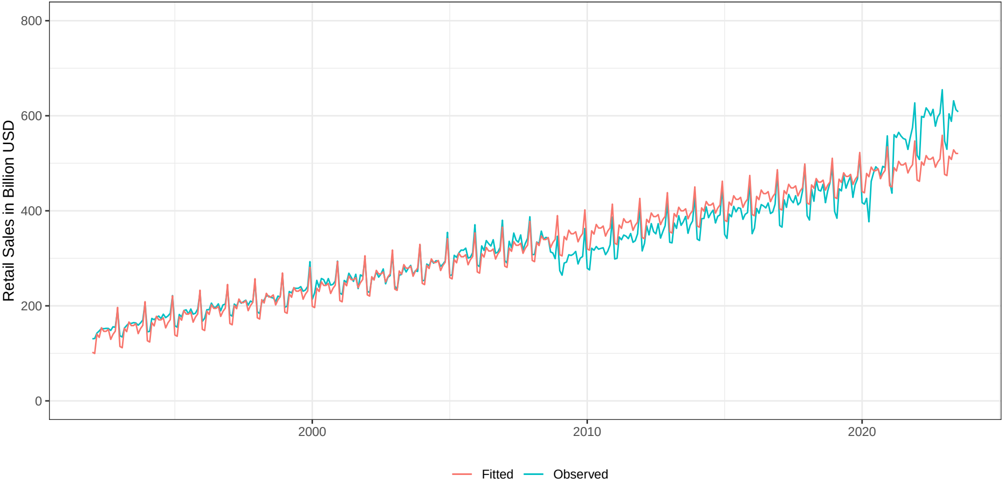
$$y_t = \beta_0 + \delta_1 \cdot Q1_t + \delta_2 \cdot Q2_t + \delta_3 \cdot Q3_t + \beta_1 \cdot x_{1,t} + \cdots + \beta_k \cdot x_{k,t} + \epsilon_t$$

Example using `retail` Data

Implementation to obtain fitted value:

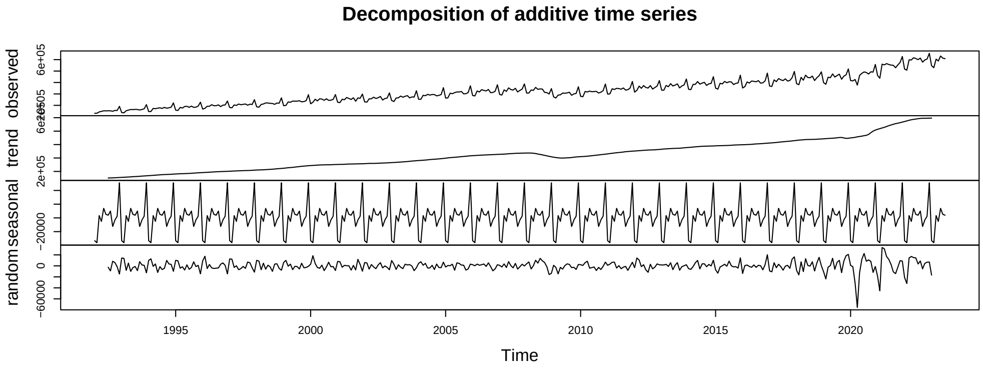
```
retail$date    = as.Date(retail$date,format="%Y-%m-%d")
retail$month   = months(retail$date)
retail$trend   = 1:nrow(retail)
bhat           = lm(retail~factor(month)+trend,data=retail)
retail$fit     = bhat$fitted.values
```

Observed and Fitted `retail` Data



Decomposition of Time Series

```r
retail  = ts(retail$retail,start=c(1992,1),frequency=12)
plot(decompose(retail,type=c("additive")))
```



Distributed-Lag Models

Reasons to Include Lags

Psychological reasons

- Force of habit, e.g., lag in changing consumption habits
- Uncertainty about permanence of change, e.g., getting a new job but with a probationary period.

Technological or economic reasons

- Difficulty to change practices due to high cost

Institutional reasons

- Contractual obligations that cannot be modified in the short-run

Relationship between income and consumption

Assume the following relationship between income and consumption:

$$C_t = \alpha + \beta_0 \cdot I_t + \beta_1 \cdot I_{t-1} + \beta_2 \cdot I_{t-2}$$

Example: Increase in income from \$4,000 to \$5,000

- Assume that $\alpha_0 = 100$, $\beta_0 = 0.4$, $\beta_1 = 0.3$, and $\beta_2 = 0.2$.
- What is the long-run consumption with \$4,000?
- How does the consumption change over the time when receiving the increase of \$1,000

Note that $\sum_{i=0}^{2} \beta_i = 0.9$

Long-Run Multiplier

Distributed-lag models (including pasted or lagged independent variables):

$$y_t = \alpha + \beta_0 \cdot x_t + \beta_1 \cdot x_{t-1} + \beta_2 \cdot x_{t-2} + \cdots + \beta_k \cdot x_{t-k} + \epsilon$$

Long-run multiplier (or long-run propensity):

$$\sum_{i=1}^{k} \beta_i = \beta_0 + \beta_1 + \beta_2 + \cdots + \beta_k = \beta$$

Koyck Method for Distributed-Lag Models

Assumption: All β_k are of the same sign, then $\beta_k = \beta_0 \cdot \lambda^k$ for $k = 0, 1, 2, \dots, \infty$. Characteristics of this assumption:

- $\lambda < 1$ gives less weight to distant βs
- Long-run multiplier is finite, i.e.,

$$\sum_{k=0}^{\infty} \beta_k = \beta_0 \left(\frac{1}{1 - \lambda} \right)$$

Regression model equation

$$y_t = \alpha + \beta_0 \cdot x_t + \beta_0 \cdot \lambda \cdot x_{t-1} + \beta_0 \cdot \lambda^2 \cdot x_{t-2} + \cdots + \epsilon_t$$

Reformulated equation: $y_t = \alpha \cdot (1 - \lambda) + \beta_0 \cdot x_t + \lambda \cdot y_{t-1} + \upsilon$

Koyck Method for Distributed-Lag Models

```
##
## Call:
## lm(formula = consumption ~ income + Lag(consumption), data = usdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4165.9   -67.7    12.9    83.0  2677.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.52067   47.28993  -0.011   0.9912
## income             0.05931    0.01851   3.204   0.0015 **
## Lag(consumption)   0.93770    0.02046  45.837   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322.4 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 1.524e+05 on 2 and 302 DF,  p-value: < 2.2e-16
```

Time Series

Overview

Stochastic process:

- Collection of random variables ordered in time

Stationary process: If the time series is not stationary then the analysis cannot be generalized to other time periods.

- Constant mean: $E(y_t) = \mu$
- Constant variance: $Var(y_t) = \sigma^2$
- Constant covariance depending on h but not t: $\gamma_h = Cov(y_t, y_{t-h})$

White noise:

- Purely random stochastic process with mean zero and constant variance.

Autoregressive Model of Order 1: AR(1)

AR(1) Model

$$y_t = \alpha + \phi_1 \cdot y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. Properties of an AR(1) process:

- Mean of x_t: $\mu = \frac{\alpha}{1-\phi_1}$
- Variance: $Var(x_t) = \frac{\sigma_\epsilon^2}{1-\phi_1^2}$
- Correlation: $\rho_h = \phi_1^h$ where h represents the number of periods separating the observations

Requirement for stationary AR(1) is that $|\phi_1| < 1$. Trending time series are usually not stationnary.

AR(1) Model: jcars

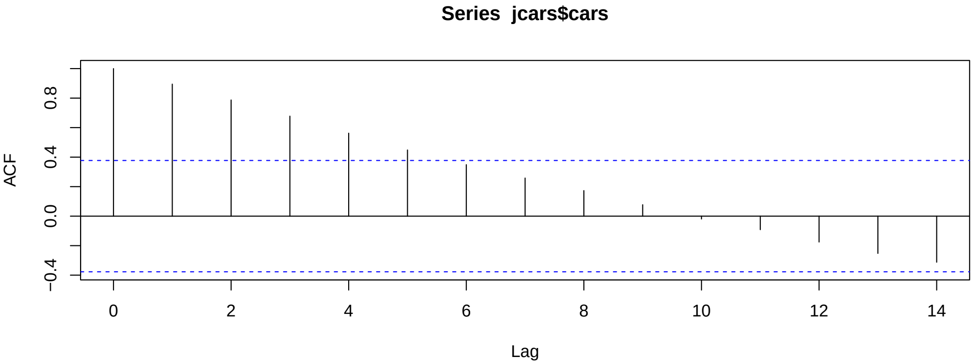
```
jcars = subset(jcars,year>1962)
bhat1 = lm(cars~Lag(cars),data=jcars)
summary(bhat1)
```

```
##
## Call:
## lm(formula = cars ~ Lag(cars), data = jcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -996.98 -210.64   11.38  250.13 1013.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 667.3499   186.1964   3.584   0.0015 **
## Lag(cars)     0.9716     0.0221  43.972   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 412.9 on 24 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9877, Adjusted R-squared:  0.9872
## F-statistic:  1934 on 1 and 24 DF,  p-value: < 2.2e-16
```

Autocorrelation Function (ACF)

ACF: Correlation between x_t and x_{t-1}, x_{t-2}, x_{t-3}, and so on.

- Identification of possible time series structure using function `acf()`
- Use on time series and regression residuals



Moving Average Models

A moving average term in a time series model is a past error (multiplied by a coefficient), e.g., MA(1):

$$x_t = \mu + w_t + \theta_1 \cdot w_{t-1}$$

where $w_t \sim N(0, \sigma_w^2)$. The MA(2) model is written as:

$$x_t = \mu + w_t + \theta_1 \cdot w_{t-1} + \theta_2 \cdot w_{t-2}$$

Properties of an MA(1) model:

- $E(x_t) = \mu$
- $Var(x_t) = \sigma_w^2 \cdot (1 + \theta_1^2)$
- ACF is $\rho_1 = \theta_1/(1 + \theta_1^2)$ and $\rho_h = 0$ for $h \geq 2$

Random Walk

Let ϵ_t be white noise then the random walk without drift is

$$y_t = y_{t-1} + \epsilon_t$$

Example:

$$y_1 = y_0 + \epsilon_1$$

$$y_2 = y_1 + \epsilon_2 = y_0 + \epsilon_1 + \epsilon_2$$

This is not a stationary process and it can be shown that $E(y_t) = y_0$ and $Var(y_t) = t \cdot \sigma^2$. However

$$y_t - y_{t-1} = \Delta y_t = \epsilon_t$$

Random Walk and Autoregressive Models

Let ϵ_t be white noise then the random walk with drift is

$$y_t = \alpha + y_{t-1} + \epsilon_t$$

where α is the drift parameter. It can be shown that $E(y_t) = y_0 + \alpha \cdot t$ and $Var(y_t) = t \cdot \sigma^2$. An autoregressive model AR(p) can be written as

$$y_t = \alpha + \sum_{i=1}^{p} \phi_p \cdot y_{t-p} + \epsilon_t$$

Forecasting

Overview

AR(p) with δ being the mean:

$$y_t - \delta = \alpha_1 \cdot (y_{t-1} - \delta) + \alpha_2 \cdot (y_{t-2} - \delta) + \cdots + \alpha_p \cdot (y_{t-p} - \delta) + \epsilon_t$$

Moving average (MA) process: MA(q)

$$y_t = \mu + \beta_0 \cdot \epsilon_t + \beta_1 \cdot \epsilon_{t-1} + \beta_2 \cdot \epsilon_{t-2} + \cdots + \beta_q \cdot \epsilon_{t-q}$$

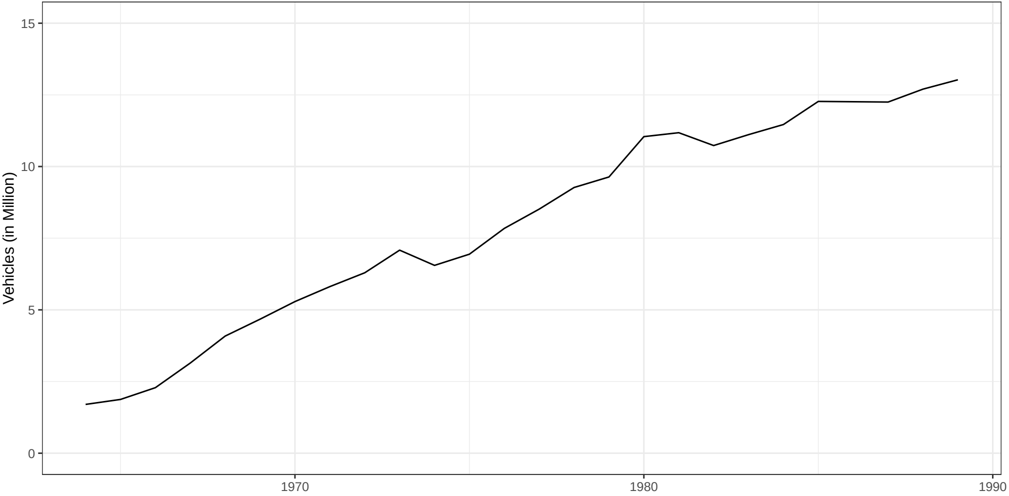
Autoregressive and moving average (ARMA) process: ARMA(p,q)

$$y_t = \theta + \alpha_1 \cdot y_{t-1} + \beta_0 \cdot \epsilon_t + \beta_1 \cdot \epsilon_{t-1}$$

Autoregressive Integrated Moving Average (ARIMA) Model: ARIMA(p,d,q)

- Correction for non-stationary time series

Japanese Car Production 1964-1989



Forecasting Procedures

Model 1: Regular OLS Model

$$y_t = \beta_0 + \beta_1 \cdot t + \epsilon_t$$

Model 2: Autoregressive Model

$$y_t = \beta_0 + \beta_1 \cdot t + n_t \quad \text{where} \quad n_t = \phi_1 \cdot n_{t-1} + \epsilon_t$$

Note: Production volume after 1963

```r
bhatols    = lm(cars~year,data=jcars)
bhatarima = Arima(jcars$cars,order=c(1,0,0),
                  include.constant=TRUE,include.drift=TRUE)
```

Model 1: Regular OLS

Implementation to obtain fitted value:

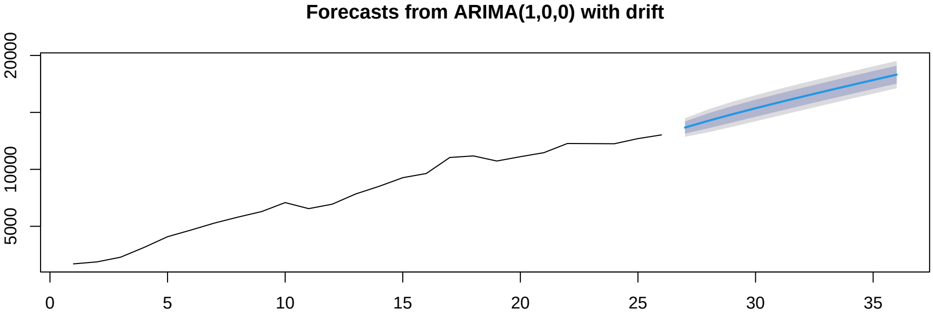
```
##
## Call:
## lm(formula = cars ~ year, data = jcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.62 -406.49   47.09  353.35 1351.64
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -924484.82   30143.45  -30.67   <2e-16 ***
## year            471.81      15.25   30.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.2 on 24 degrees of freedom
## Multiple R-squared:  0.9755, Adjusted R-squared:  0.9745
## F-statistic: 957.1 on 1 and 24 DF,  p-value: < 2.2e-16
```

Model 2: Autoregressive Model

```
## Series: jcars$cars
## ARIMA(1,0,0) with drift
##
## Coefficients:
##          ar1  intercept     drift
##       0.7363  1662.4148  463.5637
## s.e.  0.1347   471.7223   29.2265
##
## sigma^2 = 171700:  log likelihood = -192.38
## AIC=392.77   AICc=394.67   BIC=397.8
##
## Training set error measures:
##                    ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
## Training set 17.28081 389.7285 311.2957 -0.7522648 5.354775 0.5840007 0.1564618
```

Plot

```r
plot(forecast(bhatarima))
```


Forecasts from ARIMA(1,0,0) with drift