0.1 ARIMA: ARIMA Models for Time Series Data

Use auto-regressive, integrated, moving-average (ARIMA) models for time series data. A time series is a set of observations ordered according to the time they were observed. Because the value observed at time t may depend on values observed at previous time points, time series data may violate independence assumptions. An ARIMA(p, d, q) model can account for temporal dependence in several ways. First, the time series is differenced to render it stationary, by taking d differences. Second, the time dependence of the stationary process is modeled by including p auto-regressive and q moving-average terms, in addition to any time-varying covariates. For a cyclical time series, these steps can be repeated according to the period of the cycle, whether quarterly or monthly or another time interval. ARIMA models are extremely flexible for continuous data. Common formulations include, ARIMA(0, 0, 0) for least squares regression (see Section **??**), ARIMA(1, 0, 0), for an AR1 model, and ARIMA(0, 0, 1) for an MA1 model. For a more comprehensive review of ARIMA models, see Enders (2004).

Syntax

```
> z.out <- zelig(Diff(Y, d, ds=NULL, per=NULL) ~ lag.y(p, ps=NULL)
                 + lag.eps(q, qs=NULL) + X1 + X2,
                 model="arima", data=mydata, ...)
> x.out <- setx(z.out, X1 = list(time, value), cond = FALSE)
> s.out <- sim(z.out, x=x.out, x1=NULL)
```

Inputs

In addition to independent variables, `zelig()` accepts the following arguments to specify the `ARIMA` model:

- `Diff(Y, d, ds, per)` for a dependent variable Y sets the number of non-seasonal differences (`d`), the number of seasonal differences (`ds`), and the period of the season (`per`).
- `lag.y(p, ps)` sets the number of lagged observations of the dependent variable for non-seasonal (`p`) and seasonal (`ps`) components.
- `lag.eps(q, qs)` sets the number of lagged innovations, or differences between the observed value of the time series and the expected value of the time series for non-seasonal (`q`) and seasonal (`qs`) components.

In addition the user can control the estimation of the time series with the following terms:

- . . .: Additional inputs. See `help(arima)` in the stats library for further information.

Stationarity

A stationary time series has finite variance, correlations between observations that are not time-dependent, and a constant expected value for all components of the time series (Brockwell and Davis 1991, p. 12). Users should ensure that the time series being analyzed is stationary before specifying a model. The following commands provide diagnostics to determine if a time series `Y` is stationary.

- `pp.test(Y)`: Tests the null hypothesis that the time series is non-stationary.
- `kpss.test(Y)`: Tests the null hypothesis that the time series model is stationary.

The following commands provide graphical means of diagnosing whether a given time series is stationary.

- `ts.plot(Y)`: Plots the observed time series.
- `acf(Y)`: Provides the sample auto-correlation function (correlogram) for the time series.
- `pacf(Y)`: Provides the sample partial auto-correlation function (PACF) for the time series.

These latter two plots are also useful in determining the p autoregressive terms and the q lagged error terms. See Enders (2004) for a complete description of how to utilize ACF and PACF plots to determine the order of an ARIMA model.

Examples

1. No covariates

   Estimate the ARIMA model, and summarize the results.

```
> data(approval)

> z.out1 <- zelig(Diff(approve, 1) ~ lag.eps(2) + lag.y(2), data = approval,
+     model = "arima")
> summary(z.out1)
```

   Set the number of time periods (ahead) for the prediction to run. for which you would like the prediction to run:

```
> x.out1 <- setx(z.out1, pred.ahead = 10)
```

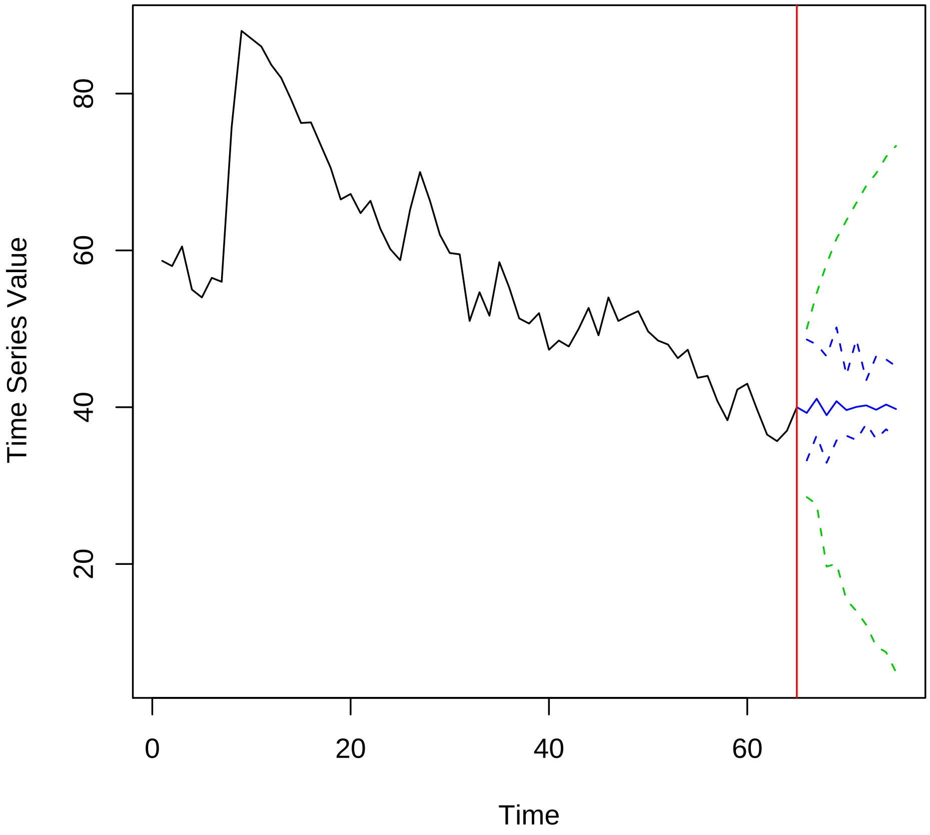
   Simulate the predicted quantities of interest:

```
> s.out1 <- sim(z.out1, x = x.out1)
```

Summarize and plot the results:

```
> summary(s.out1)
```

```
> plot(s.out1, lty.set = 2)
```



2. Calculating a treatment effect

Estimate an ARIMA model with exogenous regressors, in addition to lagged errors and lagged values of the dependent variable.

```
> z.out2 <- zelig(Diff(approve, 1) ~ iraq.war + sept.oct.2001 +
+      avg.price + lag.eps(1) + lag.y(2), data = approval, model = "arima")
```

To calculate a treatment effect, provide one counterfactual value for one time period for one of the exogenous regressors (this is the counterfactual treatment).

```
> x.out2 <- setx(z.out2, sept.oct.2001 = list(time = 45, value = 0),
+     cond = T)
```

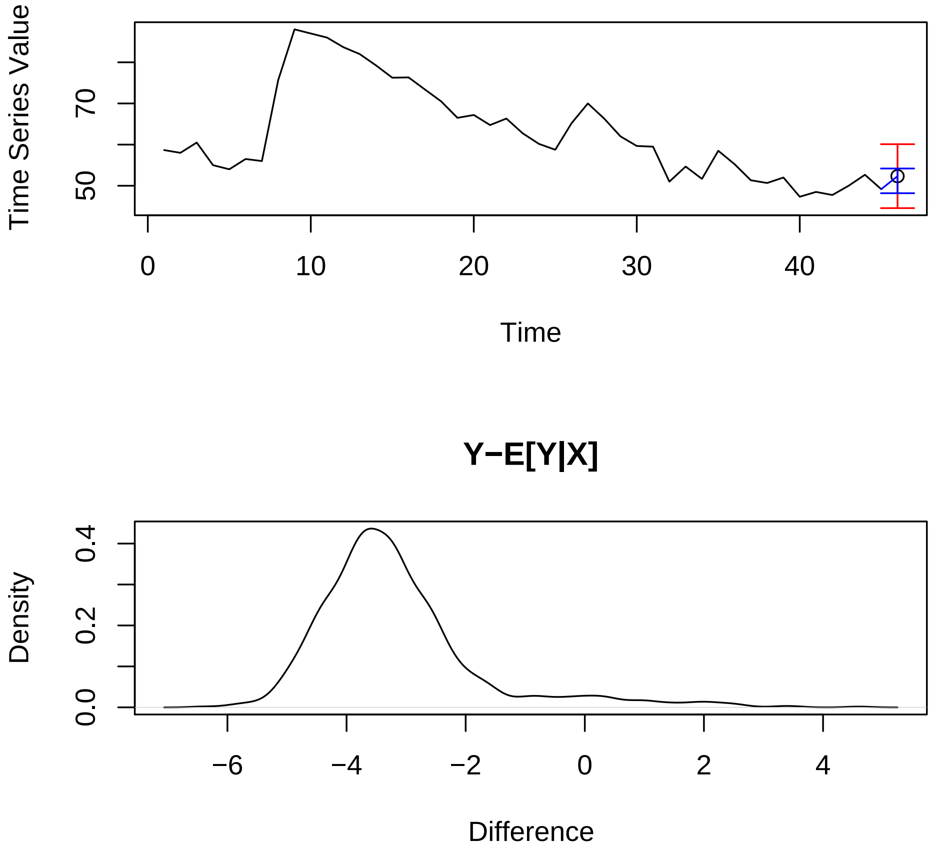
Simulate the quantities of interes

```
> s.out2 <- sim(z.out2, x = x.out2)
```

Summarizing and plotting the quantities of interest.

```
> summary(s.out2)
```

```
> plot(s.out2)
```



3. Calculating first differences

Continuing the example from above, calculate first differences by selecting several counterfactual values for one of the exogenous regressors.

```
> x.out3 <- setx(z.out2, sept.oct.2001 = list(time = 45:50, value = 0))
> x1.out3 <- setx(z.out2, sept.oct.2001 = list(time = 45:50, value = 1))
```

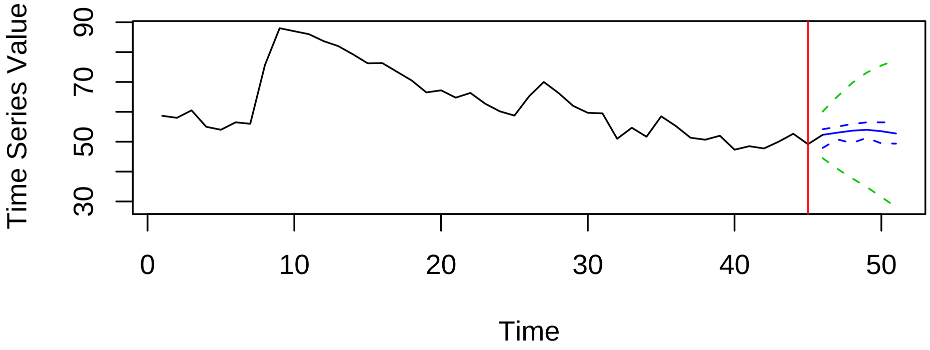
Simulating the quantities of interest

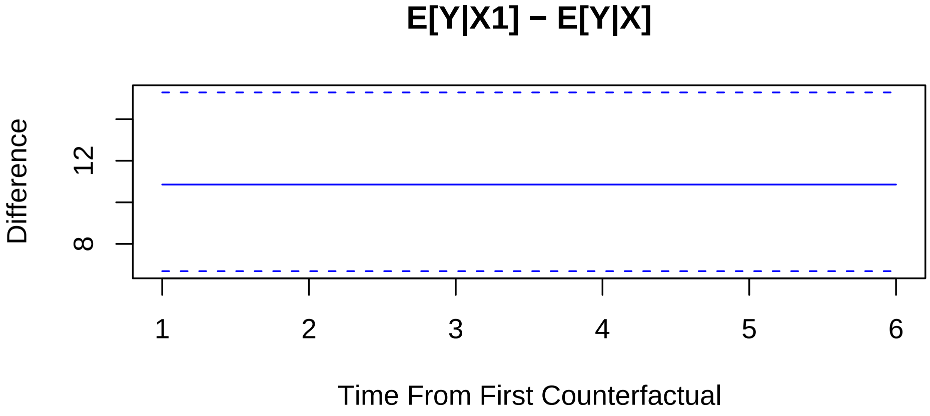
```
> s.out3 <- sim(z.out2, x = x.out3, x1 = x1.out3)
```

Summarizing and plotting the quantities of interest. Choosing `pred.se = TRUE` only displays the uncertainty resulting from parameter estimation.

```
> summary(s.out3)
```

```
> plot(s.out3, pred.se = TRUE)
```





Model

Suppose we observe a time series Y, with observations Y_i where i denotes the time at which the observation was recorded. The first step in the ARIMA procedure is to ensure that this series is stationary. If initial diagnostics indicate non-stationarity, then we take additional differences until the diagnostics indicate stationarity. Formally, define the difference operator, ∇^d, as follows. When $d = 1$, $\nabla^1 Y = Y_i - Y_{i-1}$, for all observations in the series. When $d = 2$, $\nabla^2 Y = (Y_i - Y_{i-1}) - (Y_{i-1} - Y_{i-2})$. This is analogous to a polynomial expansion, $Y_i - 2Y_{i-1} + Y_{i-2}$. Higher orders of differencing ($d > 2$) following the same function. Let Y^* represent the stationary time series derived from the initial time series by differencing Y d times. In the second step, lagged values of Y^* and errors $\mu - Y_i^*$ are used to model the time series. ARIMA utilizes a state space representation of the ARIMA model to assemble the likelihood and then utilizes maximum likelihood to estimate the parameters of the model. See Brockwell and Davis (1991) Chapter 12 for further details.

- A stationary time series Y_i^* that has been differenced d times has *stochastic component*:

  $$Y_i^* \sim \text{Normal}(\mu_i, \sigma^2),$$

  where μ_i and σ^2 are the mean and variance of the Normal distribution, respectively.

- The *systematic component*, μ_i is modeled as

  $$\mu_i = x_i\beta + \alpha_1 Y_{i-1}^* + \ldots + \alpha_p Y_{i-p}^* + \gamma_1 \epsilon_{i-1} + \ldots + \gamma_q \epsilon_{i-q}$$

  where x_i are the explanatory variables with associated parameter vector β; Y^* the lag-p observations from the stationary time series with associated parameter vector α; and ϵ_i the lagged errors or innovations of order q, with associated parameter vector γ.

Quantities of Interest

- The expected value (`qi$ev`) is the mean of simulations from the stochastic component,

  $$\text{E}(Y_\text{i}) = \mu_i = x_i\beta + \alpha_1 Y_{i-1}^* + \ldots + \alpha_p Y_{i-p}^* + \gamma_1 \epsilon_{i-1} + \ldots + \gamma_q \epsilon_{i-q}$$

  given draws of β, α, and γ from their estimated distribution.

- The first difference (`qi$fd`) is:

  $$\text{FD}_i = E(Y|x_{1i}) - E(Y|x_i)$$

- The treatment effect (`qi$t.eff`), obtained with `setx(..., cond = TRUE)`, represents the difference between the observed time series and the expected value of a time series with counterfactual values of the external regressors. Formally,

  $$\text{t.eff}_\text{i} = Y_i - E[Y_i|x_i]$$

  Zelig will not estimate both first differences and treatment effects.

Output Values

The output of each Zelig command contains useful information which the user may view. For example, if the user runs `z.out <- zelig(Diff(Y,1) + lag.y(1) + lag.eps(1) + X1, model = "arima", data)` then the user may examine the available information in `z.out` by using `names(z.out)`, see the coefficients by using `z.out$coef` and a default summary of information through `summary(z.out)`. `tsdiag(z.out)` returns a plot of the residuals, the ACF of the residuals, and a plot displaying the *p*-values for the Ljung-Box statistic. Other elements, available through the $ operator are listed below.

- From the `zelig()` output object `z.out`, you may extract:
  - `coef`: parameter estimates for the explanatory variables, lagged observations of the time series, and lagged innovations.
  - `sigma2`: maximum likelihood estimate of the variance of the stationary time series.
  - `var.coef`: variance-covariance matrix for the parameters.
  - `loglik`: maximized log-likelihood.
  - `aic`: Akaike Information Criterion (AIC) for the maximized log-likelihood.
  - `residuals`: Residuals from the fitted model.
  - `arma`: A vector with seven elements corresponding to the AR and MA, the seasonal AR and MA, the period of the seasonal component, and the number of non-seasonal and seasonal differences of the dependent variable.
  - `data`: the name of the input data frame.
- From the `sim()` output object `s.out` you may extract quantities of interest arranged as matrices, with the rows indicating the number of the simulations, and the columns representing the simulated value of the dependent variable for the counterfactual value at that time period. `summary(s.out)` provides a summary of the simulated values, while `plot(s.out)` provides a graphical representation of the simulations. Available quantities are:
  - `qi$ev`: the simulated expected probabilities for the specified values of `x`.
  - `qi$fd` : the simulated first difference for the values that are specified in `x` and `x1`.
  - `qi$t.eff`: the simulated treatment effect, difference between the observed `y` and the expected values given the counterfactual values specified in `x`.

How to Cite

To cite the ARIMA Zelig module:

Justin Grimmer. 2007. "ARIMA: Models for Time Series Data," in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software," `http://gking.harvard.edu/zelig`.

To cite Zelig as a whole, please reference these two sources:

> Kosuke Imai, Gary King, and Olivia Lau. 2007. "Zelig: Everyone's Statistical Software," `http://GKing.harvard.edu/zelig`.

> Kosuke Imai, Gary King, and Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development," *Journal of Computational and Graphical Statistics*, forthcoming, `http://gking.harvard.edu/files/abs/z-abs.shtml`.

See also

The ARIMA function is part of the stats package (Venables and Ripley 2002) For an accessible introduction to identifying the order of an ARIMA model consult Enders (2004) In addition, advanced users may wish to become more familiar with the state-space representation of an ARIMA process (Brockwell and Davis 1991) Additional options for ARIMA models may be found using `help(arima)`.

Bibliography

Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, Springer-Verlag, 2nd ed.

Enders, W. (2004), *Applied Econometric Time Series*, Wiley, 2nd ed.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer-Verlag, 4th ed.

Bibliography

Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, Springer-Verlag, 2nd ed.

Enders, W. (2004), *Applied Econometric Time Series*, Wiley, 2nd ed.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer-Verlag, 4th ed.