36-350: Data Mining **Handout 26** **November 24, 2003**

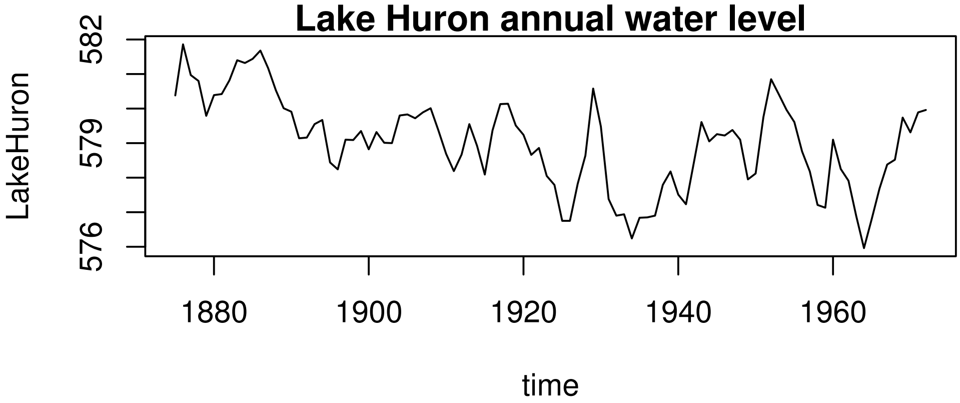
Modeling time-series data

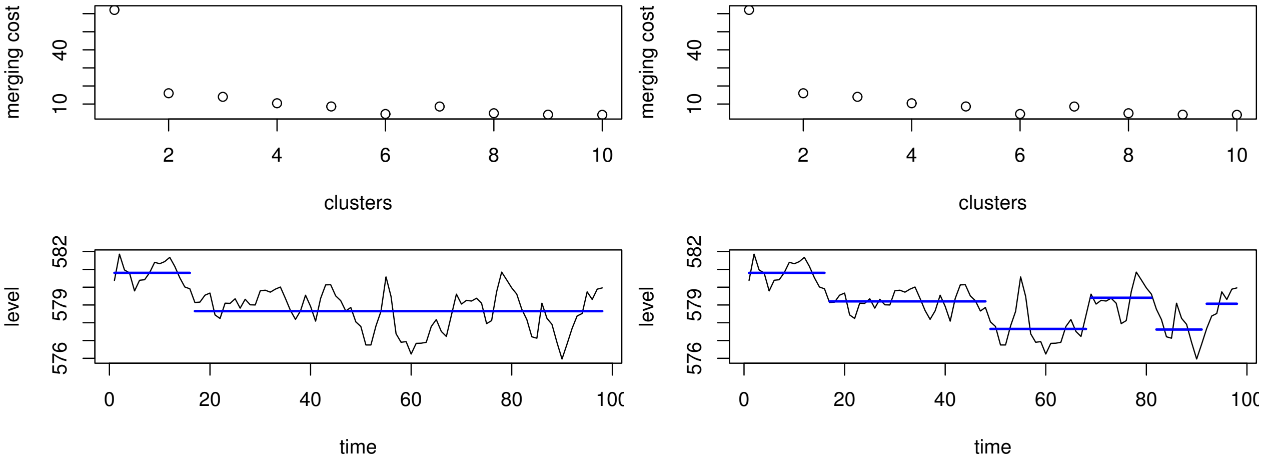
- Predict the value as a function of time (change-point model)
- Predict the value as a function of previous values (auto-regression)

It is common for a time series to hover around one value, then abruptly change to another value, and so on. These change-points can be found by clustering or by building a regression tree. In the clustering method, we repeatedly merge neighboring points, as in Ward's method. After determining the right number of clusters, the cuts that remain are the change-points.

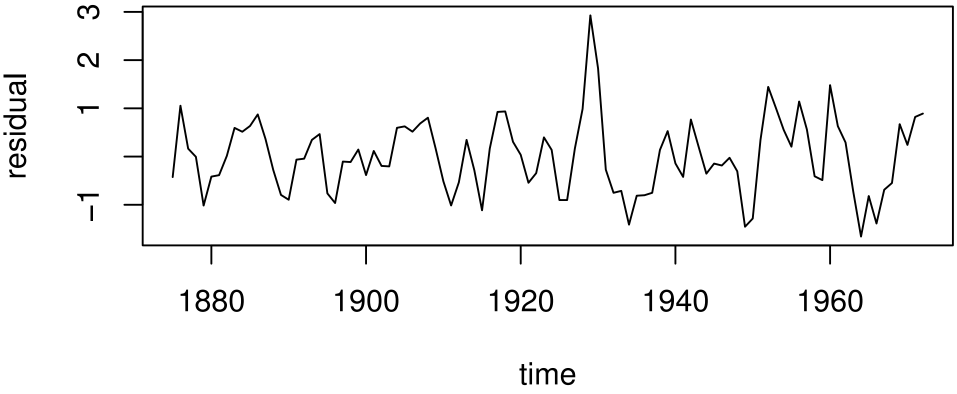
In the regression tree method, we simply predict the value as a function of time. After pruning the tree to determine the right number of splits, those that remain are the change-points.

Annual measurements of the level, in feet, of Lake Huron 1875–1972.

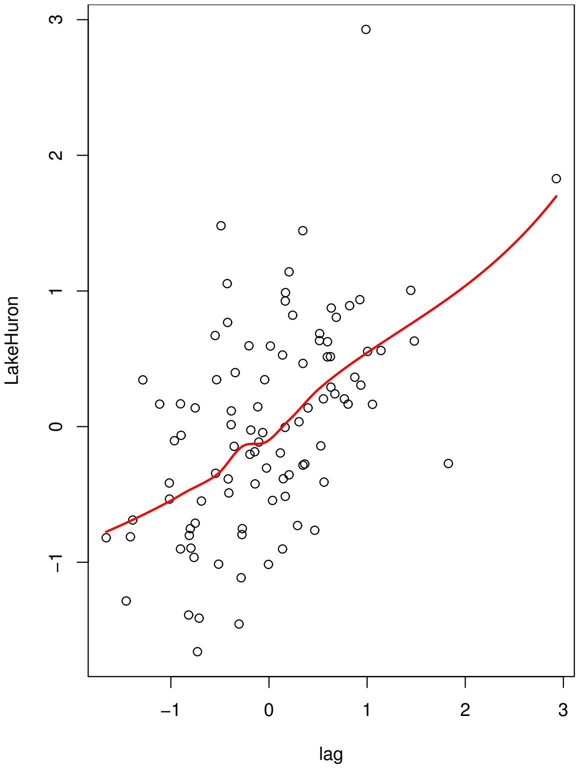




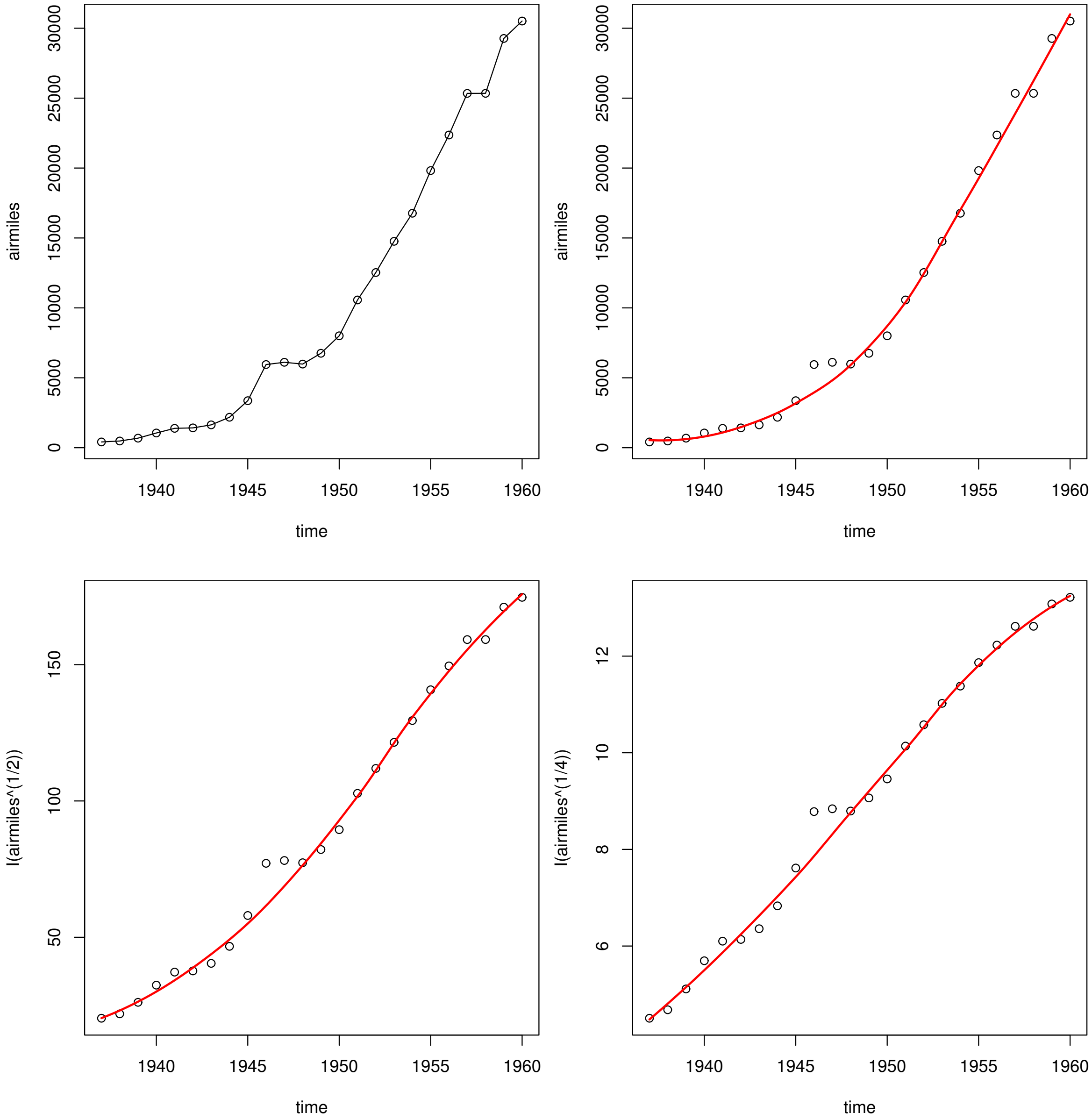
This locally-constant model can be treated as a regression model. Use the residuals from that model to find unusual values.

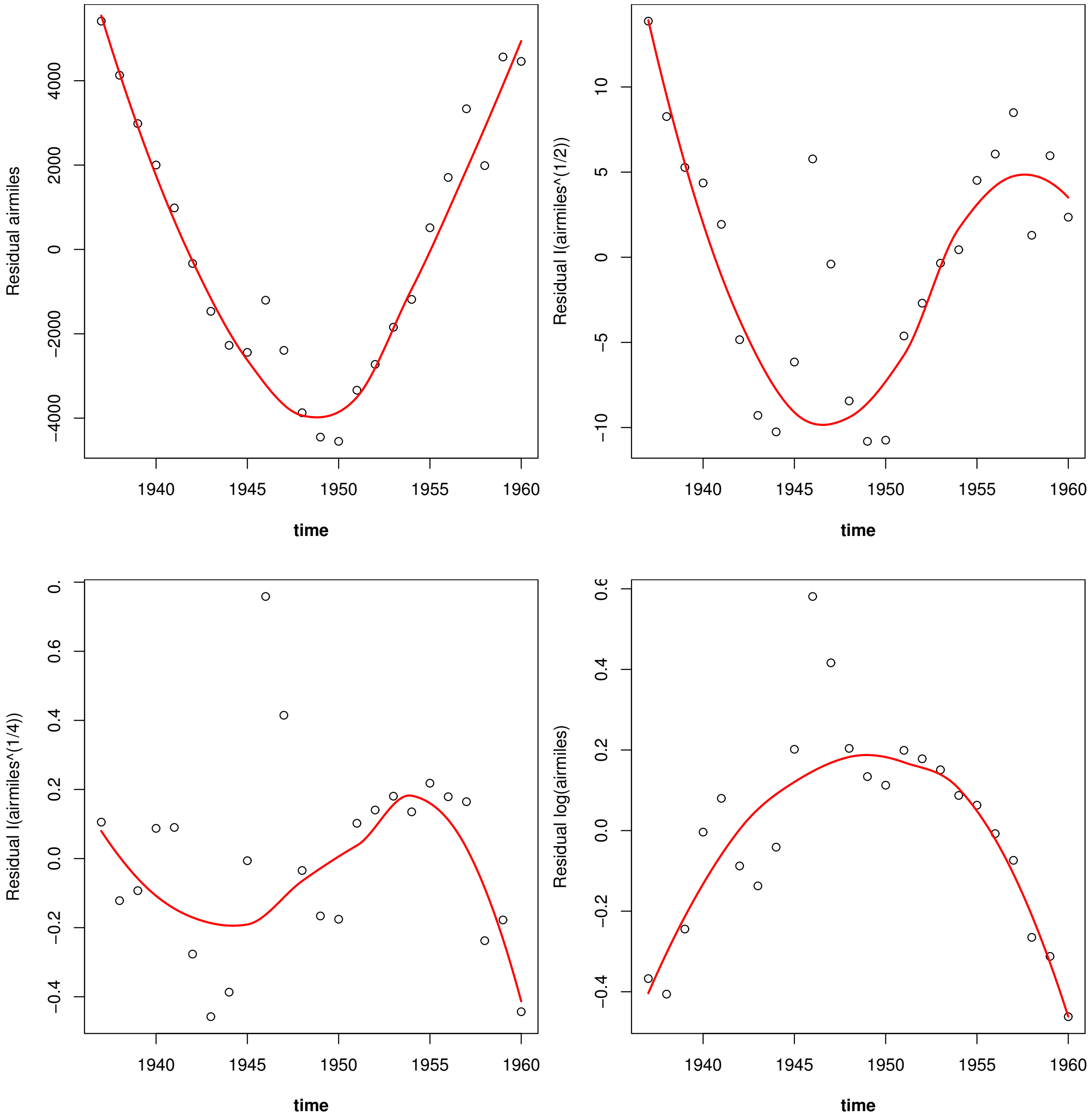


After removing the change-points, we can apply other modeling techniques. For example, here is an autoregression:



The revenue passenger miles flown by commercial airlines in the United States for each year from 1937 to 1960. What is the best model, and where are the outliers?

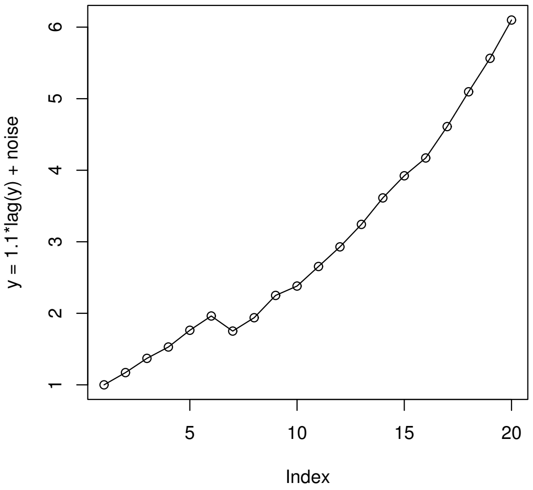


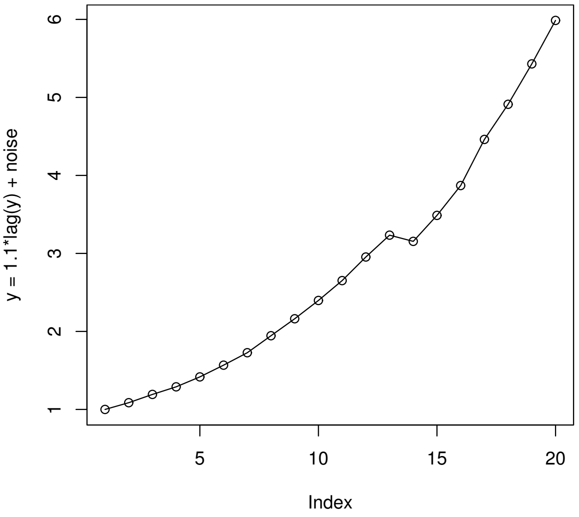


Instead of guessing different models and fitting each one, we use the residuals to guide us to the right model. The initial curvature suggests a transformation, eventually leading us to a fourth-root model. A linear model on the untransformed response gets $R^2 = 0.91$, while a linear model on the fourth-root gets $R^2 = 0.99$.

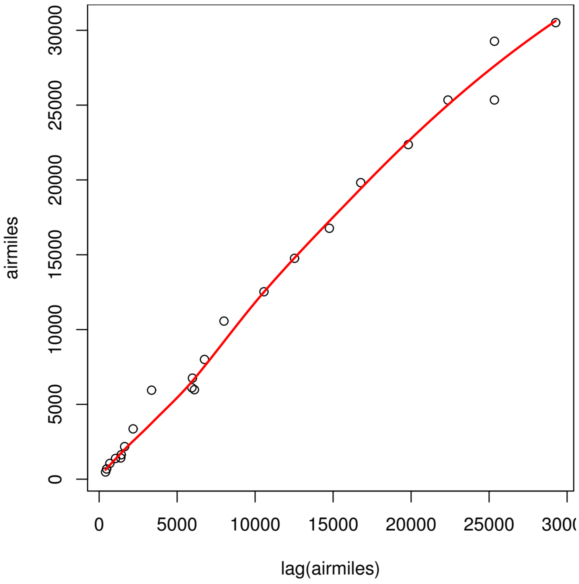
One interpretation of this model is that people look at the year before deciding to fly. A more sensible explanation is that the airlines were growing each year. But that suggests a different model, where the number of miles is determined primarily by last year's total, not by the absolute year.

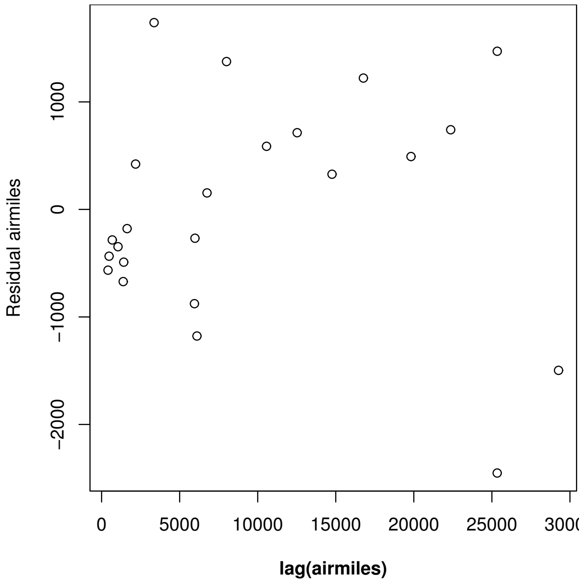
Markov model—The current value is a function of previous values, instead of time. These models are fit by **auto-regression** (regressing the time series on lagged versions of itself). Some examples of Markov data:

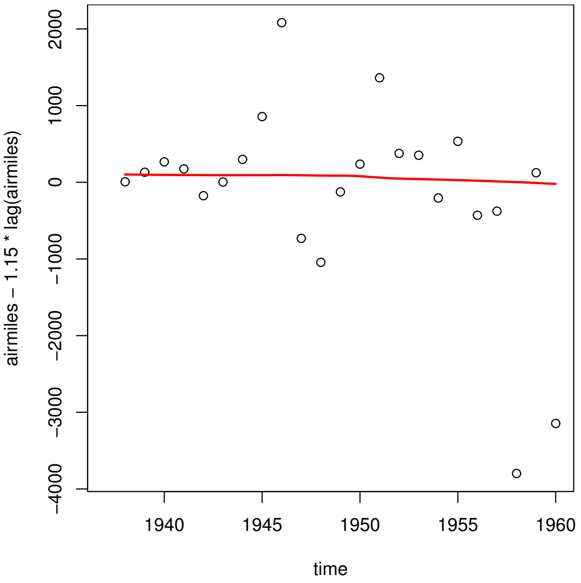


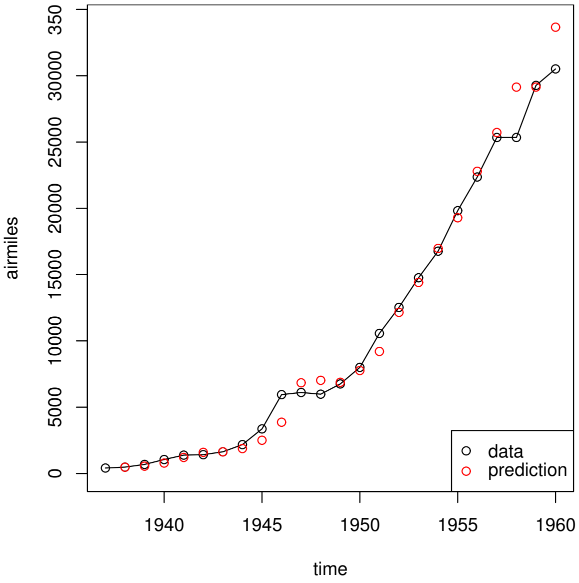


Notice how the Markov model behaves when there is a sudden change. It starts a new trend, instead of returning to the old one (which would happen if it were a function of time alone). The airmiles data does seem to follow this pattern. To see for sure, we do an auto-regression:









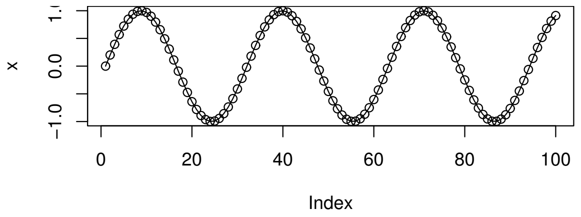
The residuals from least-squares still show a trend, because of two outliers. The residuals become flat with a coefficient of 1.15, so the correct model is `airmiles = 1.15*lag(airmiles) + noise`. In other words, airmiles increase by (an average of) 15% each year. This model has $R^2 = 0.99$ and makes sensible predictions after a sudden change (above right).

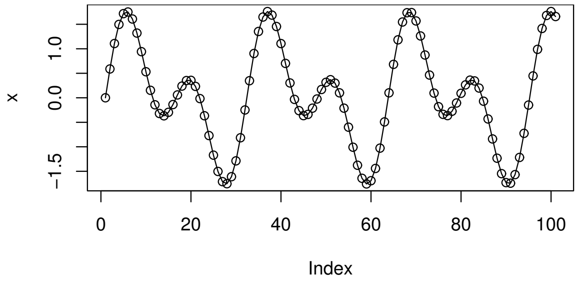
As an example of how powerful this technique is, consider a sinewave (left). As a function of time, it is nonlinear, but as a function of the past, a simple linear model suffices:

$$x_t = 1.96x_{t-1} - x_{t-2}$$

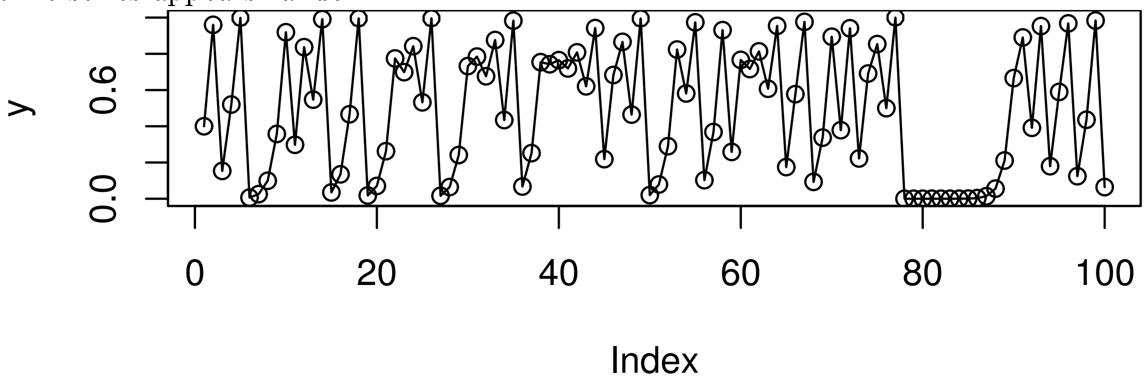
By using multiple lags, more complex repeating patterns can be modeled (right):

$$x_t = 3.95x_{t-1} - 5.9x_{t-2} + 3.95x_{t-3} - x_{t-4}$$

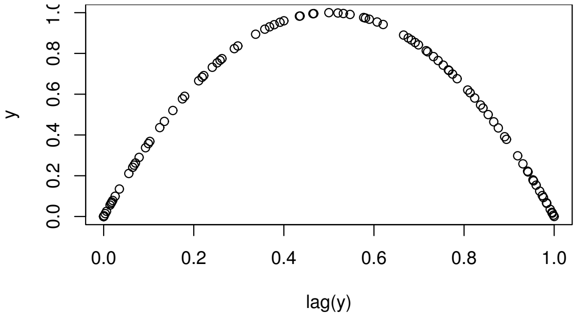




This time-series appears random:



but actually it follows a simple autoregression:



$$y_t = 4y_{t-1}(1 - y_{t-1})$$