Search this site...

STAT 501 Regression Methods

Home / 14 / 14.1

14.1 - Autoregressive Models

A **time series** is a sequence of measurements of the same variable(s) made over time. Usually the measurements are made at evenly spaced times - for example, monthly or yearly. Let us first consider the problem in which we have a *y*-variable measured as a time series. As an example, we might have *y* a measure of global temperature, with measurements observed each year. To emphasize that we have measured values over time, we use "*t*" as a subscript rather than the usual "*i*," i.e., y_t means y measured in time period t.

An **autoregressive model** is when a value from a time series is regressed on previous values from that same time series. for example, y_t on y_{t-1}:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t.$$

In this regression model, the response variable in the previous time period has become the predictor and the errors have our usual assumptions about errors in a simple linear regression model. The **order** of an autoregression is the number of immediately preceding values in the series that are used to predict the value at the present time. So, the preceding model is a first-order autoregression, written as AR(1).

If we want to predict y this year (y_t) using measurements of global temperature in the previous two years (y_{t-1}, y_{t-2}), then the autoregressive model for doing so would be:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t.$$

This model is a second-order autoregression, written as AR(2), since the value at time t is predicted from the values at times $t-1$ and $t-2$. More generally, a k^{th}-order autoregression, written as AR(*k*), is a multiple linear regression in which the value of the series at any time *t* is a (linear) function of the values at times $t-1, t-2, \ldots, t-k$.

Autocorrelation and Partial Autocorrelation

The coefficient of correlation between two values in a time series is called the **autocorrelation function** (**ACF**) For example the ACF for a time series y_t is given by:

$$\mathrm{Corr}(y_t, y_{t-k}).$$

This value of *k* is the time gap being considered and is called the **lag**. A **lag 1** autocorrelation (i.e., *k* = 1 in the above) is the correlation between values that are one time period apart. More generally, a **lag *k*** autocorrelation is the correlation between values that are *k* time periods apart.

The ACF is a way to measure the linear relationship between an observation at time *t* and the observations at previous times. If we assume an AR(*k*) model, then we may wish to only measure the association between y_t and y_{t-k} and filter out the linear influence of the random variables that lie in between (i.e., $y_{t-1}, y_{t-2}, \ldots, y_{t-(k-1)}$), which requires a transformation on the time series. Then by calculating the correlation of the transformed time series we obtain the **partial autocorrelation function** (**PACF**).

Measures for Autocorrelation

14.4 - Examples of Applying Cochrane-Orcutt Procedure

14.5 - Advanced Methods

Software Help 14

Lesson 15: Logistic, Poisson & Nonlinear Regression

Resource Menu

Notation Used in this Course

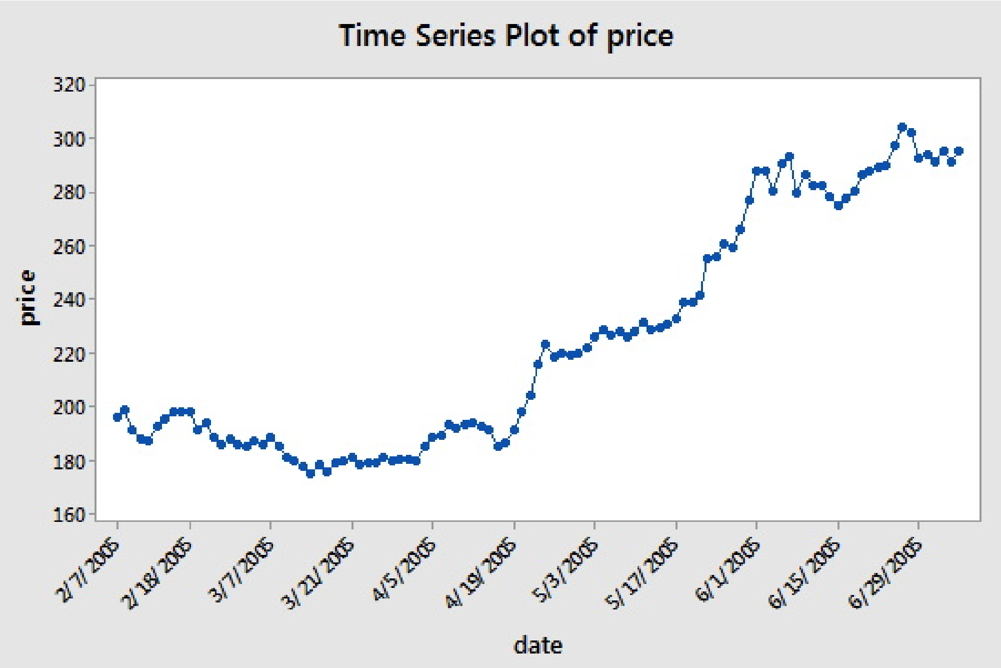
Common Procedures in Minitab

The PACF is most useful for identifying the order of an autoregressive model. Specifically, sample partial autocorrelations that are significantly different from 0 indicate lagged terms of y that are useful predictors of y_t. To help differentiate between ACF and PACF, think of them as analogues to R^2 and partial R^2 values as discussed previously.

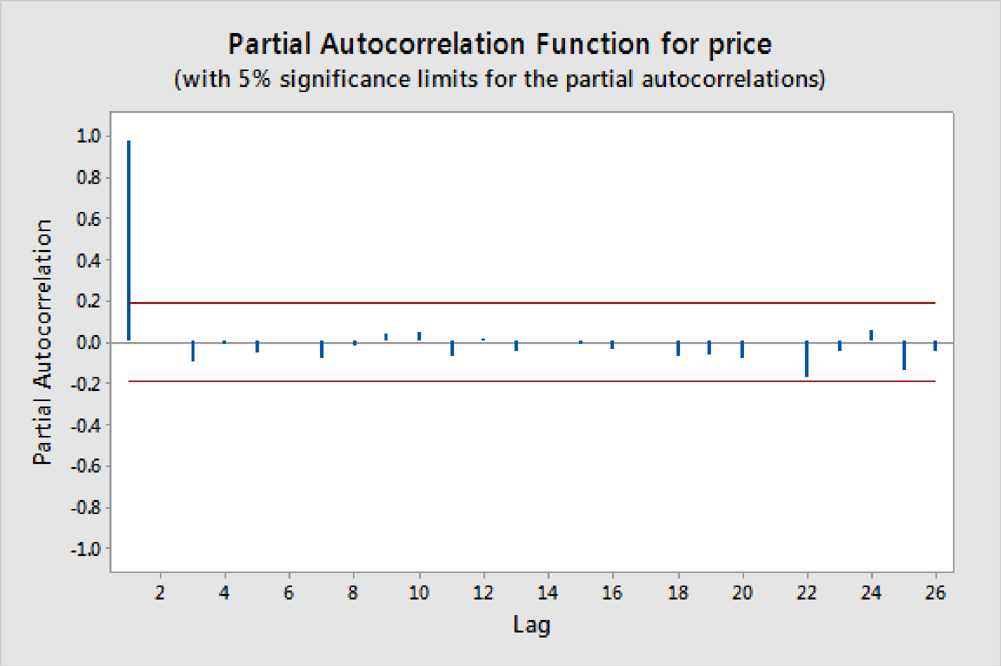
Graphical approaches to assessing the lag of an autoregressive model include looking at the ACF and PACF values versus the lag. In a plot of ACF versus the lag, if you see large ACF values and a non-random pattern, then likely the values are serially correlated. In a plot of PACF versus the lag, the pattern will usually appear random, but large PACF values at a given lag indicate this value as a possible choice for the order of an autoregressive model. It is important that the choice of the order makes sense. For example, suppose you have blood pressure readings for every day over the past two years. You may find that an AR(1) or AR(2) model is appropriate for modeling blood pressure. However, the PACF may indicate a large partial autocorrelation value at a lag of 17, but such a large order for an autoregressive model likely does not make much sense.

Example 14-1: Google Data

The **Google Stock dataset** consists of *n* = 105 values which are the closing stock price of a share of Google stock during 2-7-2005 to 7-7-2005. We will analyze the dataset to identify the order of an autoregressive model. A plot of the stock prices versus time is presented in the figure below (Minitab: Stat > Time Series > Time Series Plot, select "price" for the Series, click the Time/Scale button, click "Stamp" under "Time Scale" and select "date" to be a Stamp column):



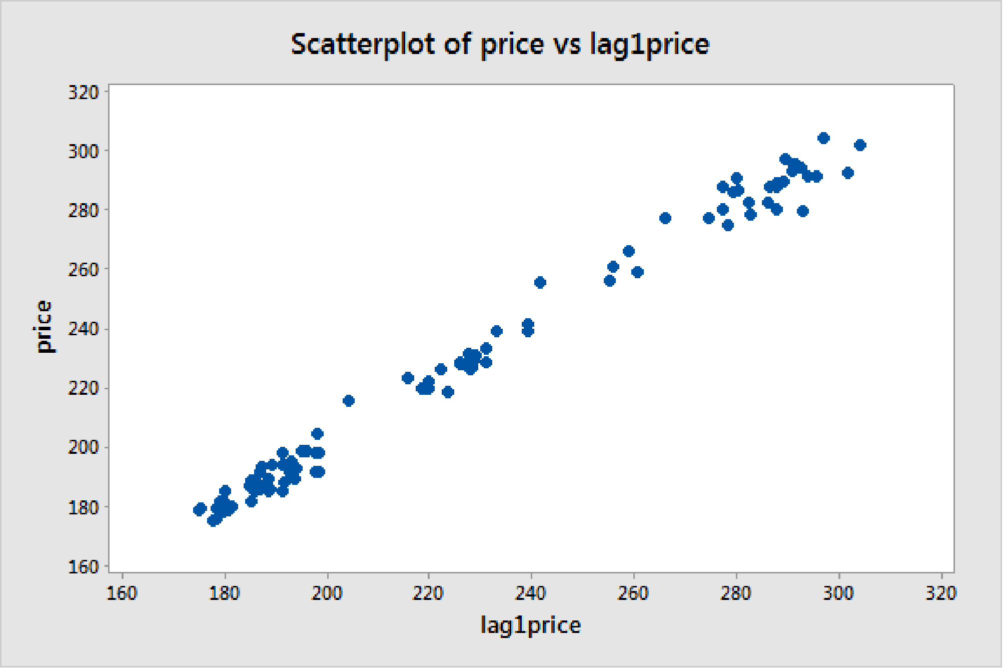
Consecutive values appear to follow one another fairly closely, suggesting an autoregression model could be appropriate. We next look at a plot of partial autocorrelations for the data:



To obtain this in Minitab select Stat > Time Series > Partial Autocorrelation. Here we notice that there is a significant spike at a lag of 1 and much lower spikes for the subsequent lags. Thus, an AR(1) model would likely be feasible for this data set.

Approximate bounds can also be constructed (as given by the red lines in the plot above) for this plot to aid in determining large values. Approximate $(1-\alpha)\times 100\%$ significance bounds are given by $\pm z_{1-\alpha/2}/\sqrt{n}$. Values lying outside of either of these bounds are indicative of an autoregressive process.

We can next create a lag-1 price variable and consider a scatterplot of price versus this lag-1 variable:



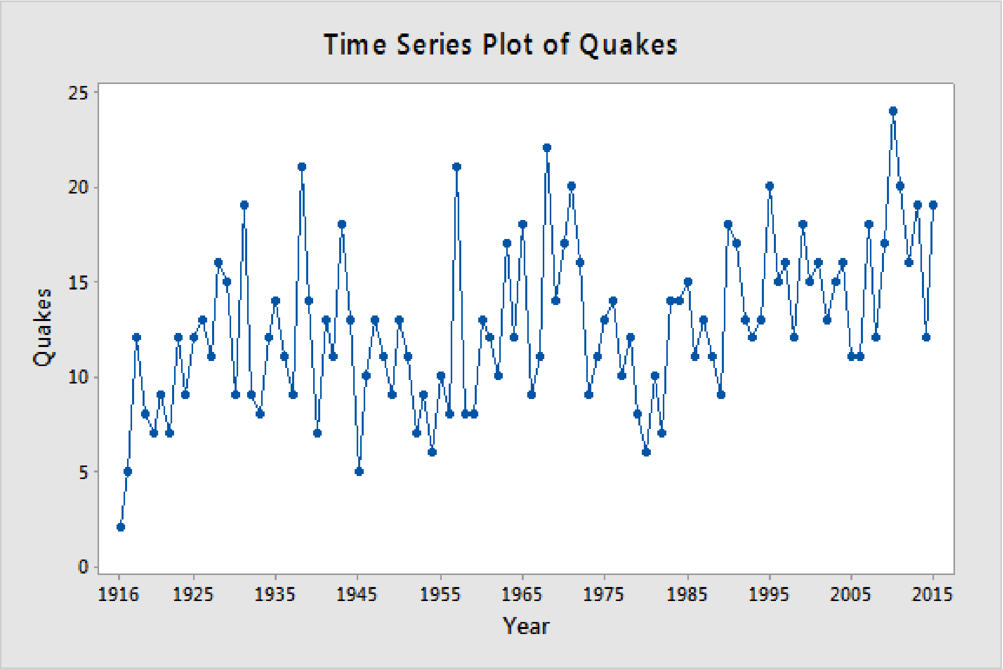
There appears to be a moderate linear pattern, suggesting that the first-order autoregression model

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

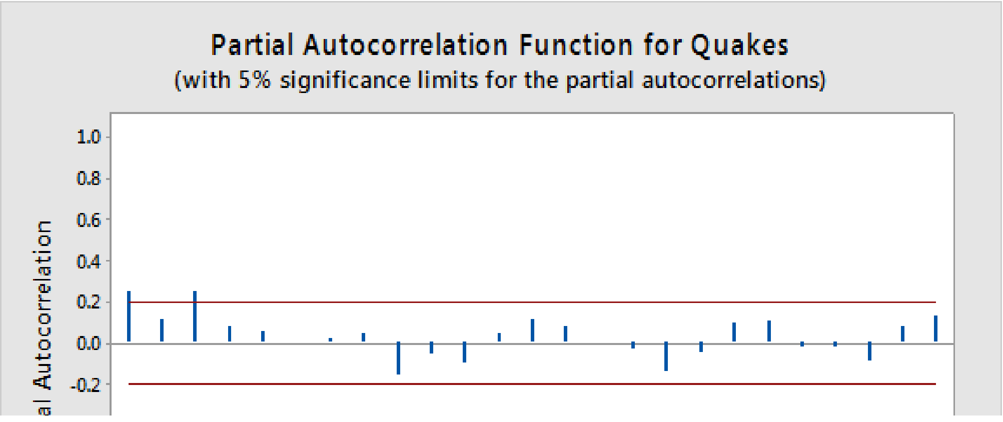
could be useful.

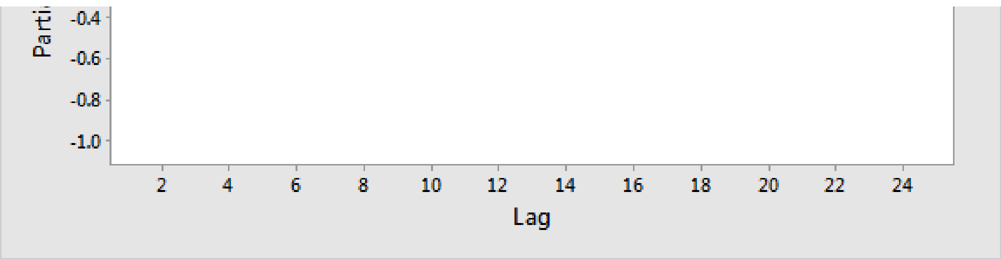
Example 14-2: Quake Data

Let y_t = the annual number of worldwide earthquakes with magnitude greater than 7 on the Richter scale for *n* = 100 years (Earthquakes data obtained from USGS Website). The plot below gives a time series plot for this dataset.



The plot below gives a plot of the PACF (partial autocorrelation function), which can be interpreted to mean that a third-order autoregression may be warranted since there are notable partial autocorrelations for lags 1 and 3.





The next step is to do a multiple linear regression with the number of quakes as the response variable and lag-1, lag-2, and lag-3 quakes as the predictor variables. (In Minitab, we used *Stat* >> *Time Series* >> *Lag* to create the lag variables.) In the results below we see that the lag-3 predictor is significant at the 0.05 level (and the lag-1 predictor p-value is also relatively small).

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.85068	13.88%	11.10%	6.47%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.45	1.79	3.61	0.000	
lag1Quakes	0.164	0.101	1.63	0.106	1.07
lag2Quakes	0.171	0.101	0.70	0.484	1.12
lag3Quakes	0.2693	0.0978	2.75	0.007	1.09

Regression Equation

Quakes = 6.45 + 0.164 lag1Quakes + 0.071 lag2Quakes + 0.2693 lag3Quakes

« Previous Next »