Agribusiness Analysis and Forecasting

Autoregressive Process, Part I

Henry Bryant

Texas A&M University
Autoregressive Process (AR)

An autoregressive (AR) time series model amounts to forecasting a variable using only its own past values.

We are going to focus on the application and less on the estimation calculations because AR models can be simply estimated using OLS.

Simetar estimates AR models easily with a menu and provides forecasts of the time series model.
AR Process

- AR is a forecasting methodology ideal for variables without clear relationships to other variables in the sense of a structural model.

- An AR process in the simplest form is a regression model such as:

\[ Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots) \]

- Notice there are no structural variables, just lags of the variable itself.
AR Process

General steps for applying an autoregressive model are:

1. Graph the data series to see what patterns are present.
2. Test data for stationarity with Dickey-Fuller (D-F) tests.
   - If original series is not stationary then difference it until it is.
   - Number of Differences (p) to make a series stationary is determined using the D-F Test.
3. Use the stationary (differenced) data series to determine the number of Lags that best forecasts the historical period.
   - Use the Schwarz Criteria (SIC), autocorrelation table, or partial-autocorrelation table to determine the best number of lags (q) to include when estimating the model.
4. Estimate the $AR(p, q)$ Model with OLS and make recursive forecasts.
Stationarity

A series is *covariance stationary* if the mean and variability is constant, i.e., the same for the future as for the past, in other words.

- $E(Y_t) = E(Y_{t-1}) = \mu$
- $\sigma^2_{T+i} = \sigma^2_{Historical} < \infty$
- $\text{Cov}(Y_t, Y_{t-k}) = \gamma_k$ and does not depend on time.
- This is a crucial assumption because if $\sigma^2$ depends on $t$, then forecast variance will explode over time.
Step to Insure the Data are Stationary

- Take differences of the data to make it stationary.
- The first difference of the raw data in $Y$ is
  \[ D_{1,t} = Y_t - Y_{t-1} \]
- Calculate the second difference of $Y$ using the first difference $(D_{1,t})$ or
  \[ D_{2,t} = D_{1,t} - D_{1,t-1} \]
- stop differencing data when series is stationary.
Make Data Series Stationary

Example Difference table for a time series data set

<table>
<thead>
<tr>
<th>t</th>
<th>Y</th>
<th>D₁</th>
<th>D₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>71.47</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>70.06</td>
<td>-1.41</td>
<td>-1.82</td>
</tr>
<tr>
<td>4</td>
<td>70.31</td>
<td>0.25</td>
<td>1.86</td>
</tr>
</tbody>
</table>
Test for Stationarity

Dickey-Fuller Test for stationarity

First D-F test: Are original data stationary?

\[ D_{1,t} = \alpha + \beta Y_{t-1} \]

- \( H_0 \): the data are non-stationary
- Parameters can be estimated using OLS
- D-F Test statistic is the \( t \) statistic on \( \beta \).
- If \( t \) is less than the critical value of -2.9 (more negative), reject \( H_0 \) at the 5% level.
- For instance, if you get a D-F statistic of -3.2, which is more negative than -2.9, then independent series are stationary.
Next Level of Testing for Stationarity

- Second D-F Test: Testing for stationarity of the $D_{1,t}$ series with the Second D-F Test.
- Here we are testing if the $Y$ series will be stationary after only one differencing
  - So we are asking if the $D_{1,t}$ series is stationary.
- Estimate regression for

\[ D_{2,t} = \alpha + \beta D_{1,t-1} \]

- $t$ statistic on slope $\beta$ is the second D-F test statistic.
- Check if the $t$ statistic is more negative than -2.90.
Test for Stationarity

- Estimate regression for: \( D_{1,t} = \alpha + \beta Y_{t-1} \).
- D-F is -1.868. You can see it is the \( t \) statistic for the \( \beta \) on the original series.
Test for Stationarity

- Estimated regression for $D_{2,t} = \alpha + \beta D_{1,t-1}$.
- D-F is -12.948, which is the $t$ ratio on the slope parameter $\beta$.
- See the residuals oscillate about a mean of zero, no trend in either series.
- Intercept is 0.121 or about zero, so the mean is more likely to be constant.
DF Stationarity Test in Simetar

Dickey-Fuller (DF) function in *Simetar*

= DF ( Data Series, Trend, No. of Lags, No. of Diff to Test)

where:

- **Data Series** is the location of the data.
- **Trend** is "False" for the test described in the previous slides.
- **No. of Lags** is zero for the test described in the previous slides.
- **No. of Diff** is the number of differences to test.

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>AA</th>
<th>AB</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dickey-Fuller Test assuming no trend and 0 lags</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>No. Diff</td>
<td>Trend</td>
<td>Lags</td>
<td>DF Test Statistic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>FALSE</td>
<td>0</td>
<td>-1.868 =DF($C$9:$C$212,W3,X3,V3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>FALSE</td>
<td>0</td>
<td>-12.948 =DF($C$9:$C$212,W4,X4,V4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>FALSE</td>
<td>0</td>
<td>-24.967 =DF($C$9:$C$212,W5,X5,V5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>TRUE</td>
<td>0</td>
<td>-1.952 =DF($C$9:$C$212,W7,X7,V7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>TRUE</td>
<td>0</td>
<td>-12.916 =DF($C$9:$C$212,W8,X8,V8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>TRUE</td>
<td>0</td>
<td>-24.903 =DF($C$9:$C$212,W9,X9,V9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Summarize Stationarity**

- $Y_t$ is the original data series.
- $D_{i,t}$ is the $i^{th}$ difference of the $Y_t$ series.
- We difference the data to make it stationary to guarantee the assumption that both mean and variance are constant.
- Dickey-Fuller test to determine the no. of differences needed to make series stationary.  
  $=$DF(Data range, False, 0, No. of Differences)
- Test as many differences as necessary with and without trend and zero lags using $=$DF().
- Select the lowest number of differences with a DF test statistic more negative than -2.90 for the purpose of estimating the AR model (described next).
**NEXT:** Determine the Number of Lags in the AR model

- Number of Lags, $q$, is the number of lagged values on the right-hand-side of the OLS equation.
- If the series is stationary with 1 difference, estimate the OLS model
  \[ D_{1,t} = \alpha + \beta_1 D_{1,t-1} + \beta_2 D_{1,t-2} + \ldots \]
- The only question that remains is how many lags ($q$) of $D_{1,t}$ will we need to forecast the series.
- To determine the number of lags we use several tests.
Determining No. of Lags (Method #1)

- Build a Sample Autocorrelation Table (SAC)  
  \[ \text{AUTOCORR(Data Series, No. Lags, No. Diff)} \]

- Pick best no. of lags based on the last lag with a statistically significant t value.

<table>
<thead>
<tr>
<th>Y</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>AA</th>
<th>AB</th>
<th>AC</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Sample Autocorrelation Coefficient Table to test for the best number of Lags</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>No. Diff</td>
<td>No. Lags</td>
<td>Auto Corr Coef</td>
<td>t Statistic</td>
<td>SE. Est.</td>
<td>Formula for Autocorr() Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0.08781</td>
<td>1.2511223</td>
<td>0.07019</td>
<td>(=\text{AUTOCORR(C9:C212, W10, V10)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>2</td>
<td>0.11502</td>
<td>1.6263217</td>
<td>0.07073</td>
<td>(=\text{AUTOCORR(C9:C212, W11, V11)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>3</td>
<td>0.00875</td>
<td>0.1221616</td>
<td>0.07164</td>
<td>(=\text{AUTOCORR(C9:C212, W12, V12)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>4</td>
<td>-0.10255</td>
<td>-1.431334</td>
<td>0.07165</td>
<td>(=\text{AUTOCORR(C9:C212, W13, V13)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>5</td>
<td>-0.08893</td>
<td>-1.228838</td>
<td>0.07237</td>
<td>(=\text{AUTOCORR(C9:C212, W14, V14)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>6</td>
<td>-0.14881</td>
<td>-2.041179</td>
<td>0.0729</td>
<td>(=\text{AUTOCORR(C9:C212, W15, V15)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>7</td>
<td>-0.09578</td>
<td>-1.287613</td>
<td>0.07438</td>
<td>(=\text{AUTOCORR(C9:C212, W16, V16)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>8</td>
<td>-0.09946</td>
<td>-1.326378</td>
<td>0.07499</td>
<td>(=\text{AUTOCORR(C9:C212, W17, V17)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>9</td>
<td>0.09946</td>
<td>1.3149739</td>
<td>0.07564</td>
<td>(=\text{AUTOCORR(C9:C212, W18, V18)})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>10</td>
<td>0.02987</td>
<td>0.3915373</td>
<td>0.07628</td>
<td>(=\text{AUTOCORR(C9:C212, W20, V20)})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bar chart of autocorrelation coefficients in Sample \textit{AUTOCORR()} Table.

The explanatory power of the distant lags is not large enough to warrant including in the model, based on their t stats, so do not include them.
Autocorrelation Charts of Sample Autocorrelation Coefficients (SAC)

- **Lag k**
  - A SAC that cuts off
  - Dying down in a dampened exponential fashion – no oscillation
  - Dying down extremely slowly
  - Dying down quickly
  - Dying down in a dampened exponential fashion – oscillation
Determining the Number of Lags (Method #2)

- Use Schwarz Information Criterion (SIC) for an information-theoretic determination of the best number of lags.
- Find the number of lags which minimizes the SIC.
- In Simetar use the `ARLAG()` function which returns the optimal number of lags based on SIC test.
  
  \[ =ARLAG(\text{Data Series}, \text{Constant}, \text{No. of Differences}) \]
Number of Lags for AR(p,q) (Method #3)

- Partial autocorrelation coefficients used to estimate number of lags for \( D_{i,t} \) in model.
- If \( D_{1,t} \) is stationary then, define \( D_{1,t}^* = D_{1,t} - \bar{D}_{1} \):
- Test for one lag use \( \beta_1 \) from OLS regression model

\[
D_{1,t}^* = \beta_1 D_{1,t-1}^* + e_t
\]

- Test for two lags use \( \beta_2 \) from OLS regression model

\[
D_{1,t}^* = \beta_1 D_{1,t-1}^* + \beta_2 D_{1,t-2}^* + e_t
\]

- Test for three lags use \( \beta_3 \) from OLS regression model

\[
D_{1,t}^* = \beta_1 D_{1,t-1}^* + \beta_2 D_{1,t-2}^* + \beta_3 D_{1,t-3}^* + e_t
\]

After each regression we only use the beta \( (\beta_i) \) for the last lagged term, i.e., the bold ones above. Use the \( t \) test on the last \( \beta_i \) to determine contribution of the last lag to explaining \( D_{1,t}^* \).
Note: Partial vs. Sample Autocorrelation

- Partial autocorrelation coefficients (PAC) show the contribution of adding one more lag (PAUTOCORR).
  - It takes into consideration the impacts of lower order lags.
  - A $\beta$ for the 3rd lag shows the contribution of 3rd lag after having lags 1-2 in place.

$$D_{1,t}^* = \beta_1 D_{1,t-1}^* + \beta_2 D_{1,t-2}^* + \beta_3 D_{1,t-3}^* + e_t$$

- Sample autocorrelation coefficients (SAC) show contribution of adding a particular lag (AUTOCORR).
  - A SAC for 3 lags shows the contribution of just the 3rd lag.

$$D_{1,t}^* = \beta D_{1,t-3}^* + e_t$$

- Thus the SAC does not equal the PAC.
Number of Lags for Time Series Model

- Some authors suggest using SAC to determine the number of differences to achieve stationarity.

- If the SAC cuts off or dies down rapidly it is an indicator that the series is stationary.

- If the SAC dies down very slowly, the series is not stationary.

- This is a good check of the DF test, but we will rely on the DF test for stationarity.
Summarize Stationarity/Lag Determination

- Make the data series stationary by differencing the data.
  - Use the Dickey-Fuller Test ($DF < -2.90$) to find how many differences necessary to make the data stationary ($p$).
  - Use the $=DF()$ function in Simetar.
- Use the sample autocorrelation coefficients (SACs) to determine how many lags ($q$) to include in the AR model.
  
  $$=\text{AUTOCORR()} \text{ function in Simetar}$$
  
  (array formula!)
- Or... minimize the Schwarz Information Criterion to determine the number of lags ($q$) to include.
  
  $$=\text{ARLAG()} \text{ or } =\text{ARSCHWARZ()} \text{ functions in Simetar}$$