Time Series I:
Trend, Seasonality, and Autoregressive Models

Alan B. Gelder
06E:071, The University of Iowa[1]

1 Time Series Patterns

Time series data: Time series is the term for data that is measured over time. For instance, we could track a retail store's monthly sales over several years. We could follow the price of a particular stock throughout a trading day or even over the entire trading history of a company. We can also chart population or GDP or corn yield in Iowa over time. These are all examples of time series data.

Trend: In time series, the word *trend* is used to denote a long term pattern that appears in the data. An increasing or positive trend indicates long term growth, while a decreasing or negative trend represents a decline.

Seasonality: Retail sales typically peaks in November and December as people shop for the holidays. Construction jobs are more abundant in the summer than during the winter months—especially in Iowa. School supplies, such as notebooks, folders, and lined paper, are in strong demand in August and September as students and parents do back-to-school shopping. *Seasonality* in time series data can be described as a regularly repeating pattern that happens on fixed intervals (such as the calendar year).

Seasonality does not include patterns that repeat on an irregular schedule. For instance, GDP decreases during recessions and then increases after recessions. However, since recessions do not happen on regularly scheduled intervals, then the decreasing and increasing pattern of GDP in connection with recessions is not considered seasonal.

Time series data may only have a trend, or it may only have seasonality, or it can have both or neither. **(Draw examples of time series data with these different patterns.)**

Other patterns: Time series data can have an exponential shape. For example,

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

the processing capacity of computers over time would likely have an exponential shape. A logarithmic shape may appear in animal populations. Initially animal populations grow rapidly, then population growth tapers off as the total population approaches the carrying capacity of the environment. It is also possible that no apparent pattern can be seen. For instance, the stock market typically has an increasing trend. However, the stock market can jump all over the place in the short run so that no clear pattern arises.

2 Trend and Seasonal Models

2.1 The Trend Only Model

Time is always an x variable: In time series models, time is *always* a predictor variable. In practice, there are a multiple ways to enter time as a predictor variable.

- The time variable can be a period counter. For instance, if you have daily data, then a period counter would equal 1 on the first day, 2 on the second day, 3 on the third, etc. This method works for hourly, daily, weekly, monthly, quarterly, and yearly data provided that data is available for each period (so that you do not jump from April to August suddenly).
- The time variable can be the year (if your data is yearly). This works since years count up by 1 and never repeat. Months, on the other hand, repeat each year, so this method would not work.
- If you have have quarterly data, then the year and quarter could be combined as a single time variable (e.g. 2002.00, 2002.25, 2002.50, 2002.75, 2003.00, 2003.25, 2003.50, 2003.75, 2004.00, etc.). Similar decimal approaches could be used to account for monthly or weekly data. However, this decimal method will change the interpretation of the slope for the time variable. **(How?)**

The trend only model: The trend in a time series model is measured by the predictor variable *time*. The population regression line for the trend only model is as follows:

$$y = \beta_0 + \beta_1 \times \text{Time} + \varepsilon$$

The sample regression line is

$$\hat{y} = b_0 + b_1 \text{Time}$$

Interpretation of the trend: The trend is estimated by b_1, the slope coefficient for time. We can interpret b_1 as follows:

For every additional period, y increases by b_1 units on average. (Make sure to specify the units and the time period.)

2.2 Autocorrelation in Time Series

A common problem: Time series data is notorious for having self-correlated residuals, or autocorrelation. **(What are residuals?)** There are several examples that can be used to help give intuition for why this is the case. For instance, if unemployment is high today, it is very likely that unemployment will still be high tomorrow. It may change somewhat, but today's unemployment level is going to be correlated with the unemployment level tomorrow. Using a sports example, if a college football team has a lot of all-star juniors in their program, it is likely that the team will also do very well when those athletes are seniors. The strength of the football team in one season is correlated with the strength of the team in the next season.

Check the residuals! Autocorrelation can be detected by seeing if there are any *clear* patterns in the residuals (ideally there would be no pattern, which would indicate that there is no autocorrelation). Recall that the presence of autocorrelation violates one of our underlying mathematical assumptions for regression. Hence, if we do have autocorrelation, then we should take measures to correct it and account for it.

What is valid and invalid when there is autocorrelation?
Autocorrelation causes the estimates of the standard errors in regression to be *invalid* (such as SE_{b_i} and $SE(\hat{y})$). Hence, we *cannot* do the following:

- Confidence intervals for β_i
- Confidence intervals for $\hat{y}$
- Prediction intervals for $\hat{y}$

However, even if we do have autocorrelation, we *can* still

- Interpret b_i
- Calculate $\hat{y}$ to obtain a simple prediction

2.3 Including Seasonality

Season is a categorical variable. So if the data have a clear seasonal pattern, then we can incorporate that into our time series model by using seasonal indicator (binary) variables.[2] If our data is quarterly, then we have four possible categories that can represent the season. If our data is monthly, then each month is a possible category for describing the season.

For quarterly data, we can construct binary variables representing each quarter as follows:

Period	Year	Quarter	Q1	Q2	Q3	Q4
1	2007	1	1	0	0	0
2	2007	2	0	1	0	0
3	2007	3	0	0	1	0
4	2007	4	0	0	0	1
5	2008	1	1	0	0	0
6	2008	2	0	1	0	0
7	2008	3	0	0	1	0
8	2008	4	0	0	0	1
9	2009	1	1	0	0	0
10	2009	2	0	1	0	0
11	2009	3	0	0	1	0
12	2009	4	0	0	0	1

The population regression equation for a time series model that incorporates both trend and seasonality for quarterly data is

$$y = \beta_0 + \beta_1 \text{Time} + \gamma_1 Q1 + \gamma_2 Q2 + \gamma_3 Q3 + \varepsilon$$

Note that we did not include the variables for all four quarters into the model. However, the model still allows us to distinguish the fourth quarter. **(How?)**.

While the model above does not include the variable for Q4, we could also have a model that includes Q2, Q3, and Q4, but not Q1. The key is that we need to leave

[2]It may be helpful to review the categorical predictor variables section in the Multiple Regression II notes.

one of the seasonal variables out of the model.

Interpreting the seasonal coefficients: The seasonal coefficients are always interpreted in relation to the seasonal variable that is excluded from the model. For instance, in the model above, $\hat{\gamma}_1$ is the estimated average seasonal effect of the first quarter relative to the fourth quarter.

2.4 Extrapolation and Forecasting

Recall that extrapolation is the term for using a regression model to make predictions outside the range of the data that the regression model is based on. Regression models work best when we stay within the range of our data. However, there are times when we do want to make predictions beyond the range of our data. For instance, business forecasts use current and past market information to predict market information in the future (which is outside the range of the data). Likewise, meteorology uses information on past and present weather conditions to predict weather conditions in the future.

Statistics is not a crystal ball, and so we cannot predict the future perfectly (especially when natural disasters occur, or wars break out, or the stock market crashes, etc.) Forecasting entails using statistics to make educated predictions about the future. Here are a couple of guiding principles for forecasting:

- It is easier to predict the near future than it is to predict the far future.
- Time series predictions about the future are always based on the idea that past trends continue into the future (e.g. no sudden wars, natural disasters, etc.).

2.5 Trade Employment

[This example is adapted from Dr. Whitten's notes]

The trade employment data set includes the number of people (in thousands) who worked in the trade industry in Wisconsin for each month from January 1970 until December 1974.

- Make a time plot of the trade employment data. Describe any trend and seasonal patterns.

– Stat > Time Series > Time Series Plot > Series: Trade Employment > Time/Scale > Time Scale: Stamp > Stamp Columns: Date > OK > OK

- Find the trend only model for predicting employment. Check and store the residuals.

- Estimate employment in July 1975.

- Find the trend and seasonal model for predicting employment.

- Again, check and store the residuals. Do we still have autocorrelation?

- Interpret the February seasonal effect.

- What is the estimated average difference in employment between March and September?

3 Autoregressive Models

In this section, we will study how to improve our forecasting ability and construct confidence intervals for our forecasts when we have autocorrelation.

Recall that the residual is defined as

$$e = y - \hat{y}$$

Simply put, the residual is the portion of the data that cannot be explained by the regression equation. It is the information that is left over. In the absence of autocorrelation, the residuals are randomly distributed. However, when we do have autocorrelation, then the residuals are not random. This means that there is still some information in the residuals that we can harness in order to improve our forecasting ability.

Autocorrelation implies that a residual at time t is correlated with the residuals in the past. So the way to harness the information that the residuals still contain is

to use the residuals from the past as the predictor variables for the residual at time t. This idea is called an autoregressive (AR) model.

An AR(1) model predicts the error at time t based on one previous period:

$$e_t = \beta_1 e_{t-1} + \upsilon_t$$

An AR(2) model predicts the error at time t based on two previous periods:

$$e_t = \beta_1 e_{t-1} + \beta_2 e_{t-2} + \upsilon_t$$

In general, an AR(p) model predicts the error at time t based on p previous periods:

$$e_t = \beta_1 e_{t-1} + \beta_2 e_{t-2} + \ldots + \beta_p e_{t-p} + \upsilon_t$$

Choosing the right number of periods: If we do not include enough previous periods in our AR model, then we may not be extracting all of the information that is contained in the residuals. On the other hand, if we include too many previous periods in our AR model, then the variance of the coefficients (β_i) will increase, and our model will not be as accurate as it could be. We need a method to choose how many previous periods we should include in our AR model.

3.1 Partial Autocorrelation:

One tool that we can use to help determine the right number of previous periods (or "lags") for our AR model is called *partial autocorrelation.*

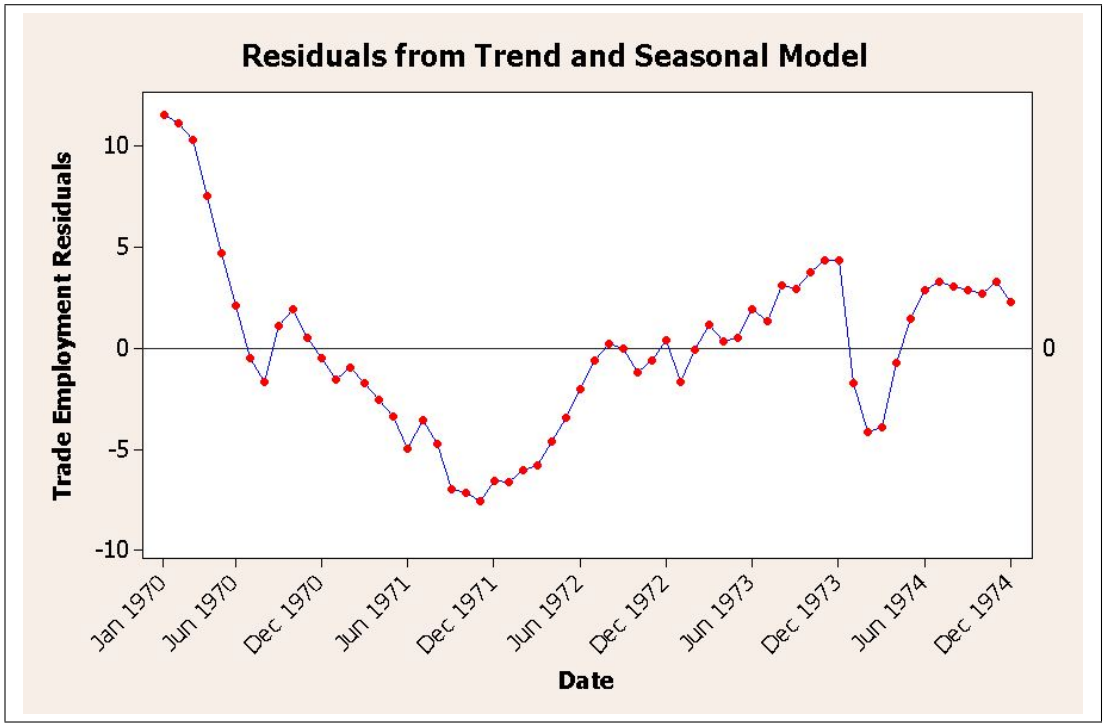
To do partial autocorrelation in Minitab, you need to first store the residuals from a regression:

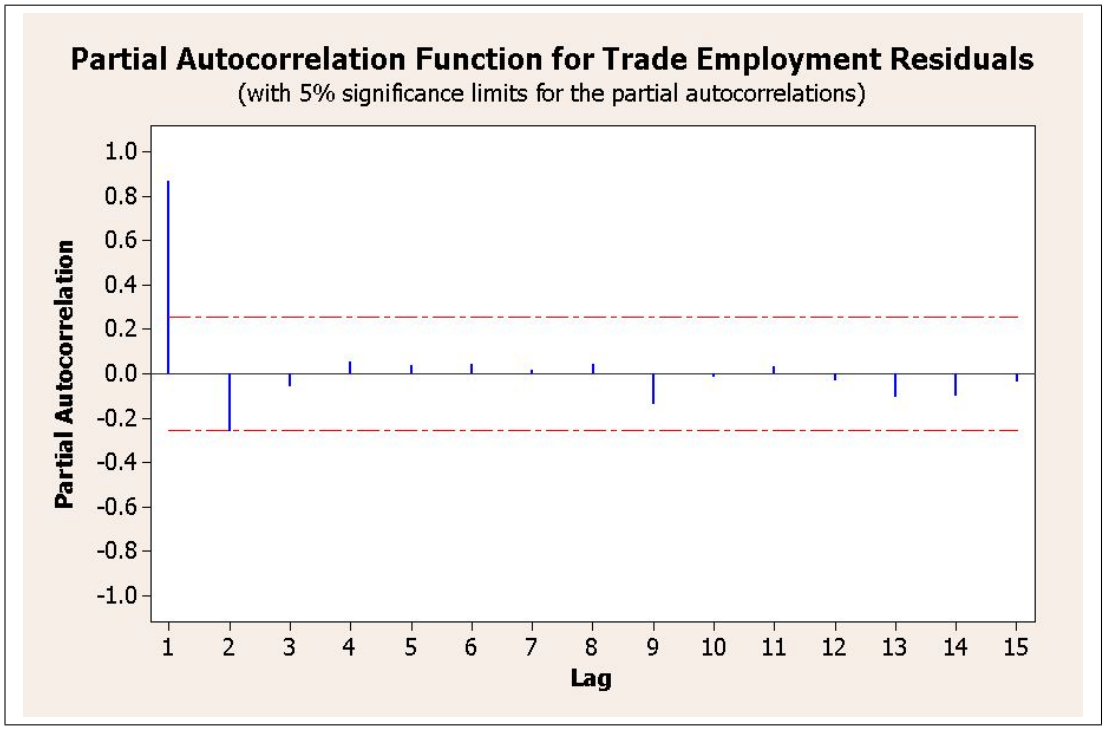
- Stat > Regression > Regression > (Enter response and predictor variables) > Storage > Diagnostic Measures: Residuals > OK > OK

Partial autocorrelation can then be done as follows:

- Stat > Time Series > Partial Autocorrelation > Series: (Enter residuals) > OK

Trade employment continued: It will be easiest to explain how to use partial autocorrelation through an example, so let's return to the trade employment data. The following two figures show the residuals from the trend and seasonal model for predicting trade employment and the partial autocorrelation plot for those residuals.


Residuals from Trend and Seasonal Model
Trade Employment Residuals
10
5
0
-5
-10
0
Jan 1970
Jun 1970
Dec 1970
Jun 1971
Dec 1971
Jun 1972
Dec 1972
Jun 1973
Dec 1973
Jun 1974
Dec 1974
Date


Partial Autocorrelation Function for Trade Employment Residuals
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Lag

In the first figure, note that the residuals in the most recent months of the data are consistently positive. This means that our model is underpredicting the true employment in the most recent months ($e = y - \hat{y} > 0$ or rather $y > \hat{y}$). Since we are consistently underpredicting employment, we can improve our forecast for January 1975 by increasing $\hat{y}$ slightly. We are going to make an autoregressive model which will specify how much we should increase our forecast.

The second figure has a really long vertical line at one lag, a shorter vertical line at two lags, and even shorter vertical lines for higher numbers of lags. There are also dotted horizontal lines at 2.5 and -2.5. The partial autocorrelation plot uses the horizontal lines as a cut-off for the number of previous residuals (or lags) that are significant for predicting the current residual. If the vertical line for a lag extends beyond the horizontal lines at 2.5 or -2.5, then the residuals from that lagged number of periods are significant.

For this data, the residuals at $t - 1$ appear to be significant for predicting the residuals at time t. The residuals at time $t - 2$ also appear to be significant, but we will need to check this numerically in order to verify that they are significant. The partial residuals plot is designed to give us a quick idea of how many lags are significant. However, we should always verify the results numerically.

3.2 AR Models:

We will verify the number of lags that are significant by constructing autoregressive models and checking the p-values. We want *all* of the lags in our autoregressive model to be significant.

The Minitab commands for building an AR model are as follows:

- Stat > Time Series > ARIMA > Series: (Enter the residuals) > Do *not* include a constant term in the model (*unselect* the box) > Autoregressive: (Enter the number of lags) > OK

In our example, the first two lags appear to be significant from the partial autocorrelation plot. However, just to be conservative, begin by building an AR model with three lags. (It is always a good idea to check one additional lag from what the partial autocorrelation plot suggests is significant.)

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	1.4055	0.1336	10.52	0.000
AR	2	-0.5150	0.2109	-2.44	0.018
AR	3	0.0470	0.1287	0.36	0.717

As you can see in this Minitab output, the first two lags are significant, but the third lag is not (use $\alpha = 0.05$). Now we will check the AR model with two lags.

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	1.3828	0.1201	11.51	0.000
AR	2	-0.4530	0.1160	-3.91	0.000

Here both of the lags are significant, so this becomes our chosen model. It is written as follows:

$$e_t = 1.3828e_{t-1} - 0.4530e_{t-2}$$

This is called an AR(2) model since the residuals at time t are predicted by the residuals from the two previous periods.

3.3 Forecasting

Now that we have selected the appropriate AR model, we can use it make predictions about the residuals beyond the range of our data. This will enable us to account for the systematic overprediction or underprediction that occurs when there is autocorrelation.

In Minitab, we need to return to the ARIMA window (Stat > Time Series > ARIMA; make sure that Minitab still has the correct information entered to produce your chosen AR model). Then we need to add the following commands:

- Forecasts > Lead: (Enter the number of future periods that you want forecasts for) > OK

In our example, we can make forecast predictions of the residuals for the first six months of 1975 by entering 6 in the "Lead" box.

```
Forecasts from period 60

                          95% Limits
Period  Forecast     Lower    Upper  Actual
    61   1.67024  -1.07141  4.41188
    62   1.27539  -3.40326  5.95404
    63   1.00709  -5.14883  7.16301
    64   0.81493  -6.42727  8.05713
    65   0.67074  -7.35697  8.69844
    66   0.55838  -8.03316  9.14992
```

The Minitab output includes the forecast for $\hat{e}_{61}$ through $\hat{e}_{66}$ (the 61st through 66th months from the start of the data in January 1970). Minitab also provides a 95% confidence interval for these forecasts of $\hat{e}$.

The forecasts for $\hat{e}$, however, are only one part of our final forecast for these months. We also need the predictions for $\hat{y}$, which can be obtained using the trend and seasonal model for trade employment.

We can either obtain $\hat{y}_{61}$ by plugging 61 into month and 1 in for January in the trend and seasonal model. Or we can have Minitab do the calculation:

- Stat > Regression > Regression > Response: Trade Employment > Predictors: Month S1-S11 > Options > Prediction intervals for new observations: 61 1 0 0 0 0 0 0 0 0 0 0 > OK > OK

Since there are eleven seasonal predictor variables, we need to enter a one for January and zeros for February through November. (**Why isn't there a variable for December?**) Part of the Minitab output is shown below:

```
Predicted Values for New Observations

New Obs      Fit  SE Fit        95% CI              95% PI
      1  375.775   2.499  (370.748, 380.802)  (364.951, 386.599)
```

Our final forecast will be the sum of $\hat{y}$ and $\hat{e}$. That is,

$$F_t = \hat{y}_t + \hat{e}_t$$

Hence, the forecast for trade employment in January 1975 (month 61) is

$$F_{61} = \hat{y}_{61} + \hat{e}_{61} = 375.775 + 1.67024 = 377.44524$$

Since employment is expressed in thousands, then our forecast for trade employment in Wisconsin in January 1975 is 377,445 employees.

Confidence intervals for forecasts: We can construct 95% confidence intervals for a forecast by adding the upper and lower bounds of the confidence interval for $\hat{e}_t$ to $\hat{y}_t$. For example, the confidence interval for $\hat{e}_{61}$ is (-1.07141, 4.41188). Therefore, a 95% confidence interval for the forecast of F_{61} is

$$(375.775 - 1.07141, \quad 375.775 + 4.41188) = (374.70359, \quad 380.18688)$$

Interpretation of a forecast: As was previously stated, forecasting beyond the range of our data is a form of extrapolation and must be done with caution. We are assuming that the same things that happened in the past will continue to happen in the future. Hence, interpretations of forecasts should indicate that we are assuming that previous trends continue.

The 95% confidence interval for the forecast of January 1975 trade employment can be interpreted as follows:
We are 95% confident that trade employment in Wisconsin in January 1975 will be between 374,704 and 380,187, *assuming* previous employment trends continue.

More practice:

- Find a simple forecast for employment in April 1975
- Find a 95% confidence interval for employment in April 1975
- Interpret the confidence interval
- Compare the width of the confidence interval for F_{61} with the prediction interval given in the Minitab output for $\hat{y}_{61}$.
- Compare the width of the confidence interval for January 1975 and April 1975. What is the intuition for this?
- Quickly review the steps for improving forecasts starting from the beginning of the trade employment example

3.4 Summary

The five steps for improving forecasts when there is autocorrelation are as follows:

1. Store the residuals from the time series model
2. Make a partial autocorrelation plot for the residuals
3. Construct an AR model where all of the lagged predictors are significant
4. Predict e_t using the AR model; predict y_t using the time series model
5. Make a forecast for period t: $F_t = \hat{y}_t + \hat{e}_t$

We can also make confidence intervals for our forecasts. This can be done by making confidence intervals for $\hat{e}_t$, and then adding the lower and upper bound of the confidence interval for $\hat{e}_t$ to $\hat{y}_t$.