Paper AA-16-2013

Leading and Lagging Indicators in SAS®

David J. Corliss

Magnify Analytic Solutions, Detroit, MI

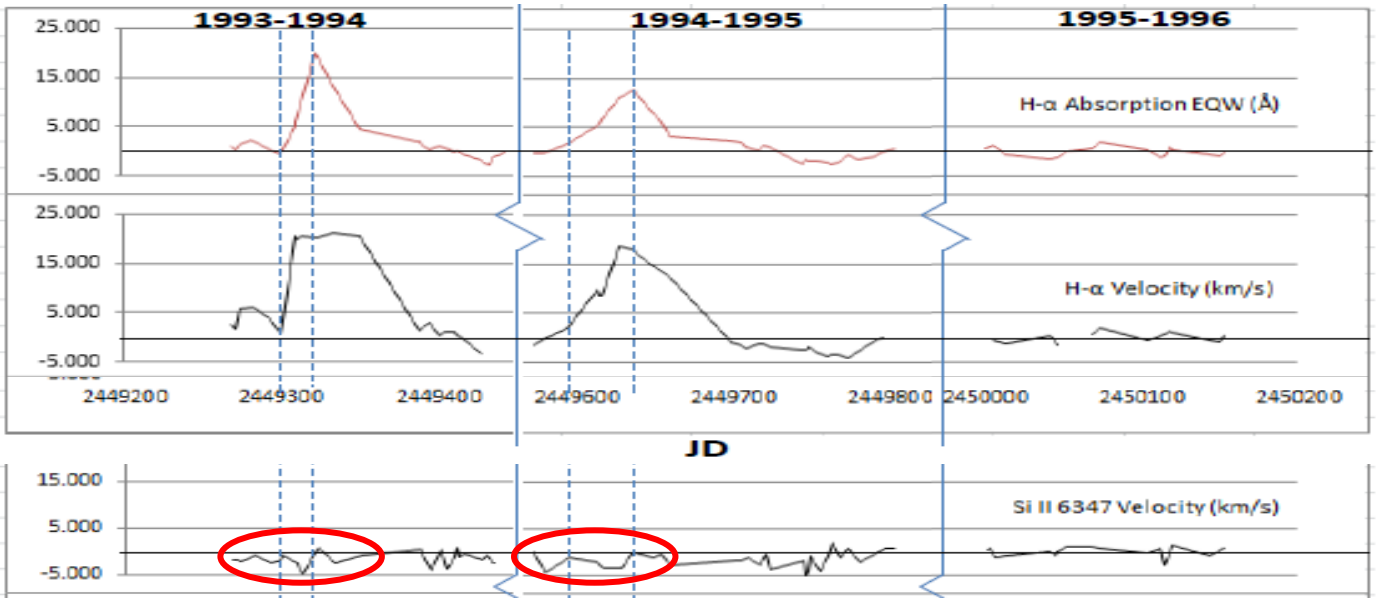
ABSTRACT

Leading indicators are familiar to us from economics as factors whose changes are correlated with future events, while lagging indicators are correlated with previous activity. However, leading and lagging indicators appear in many areas on interest: changes in a person's diet, such as gaining or losing weight, correspond to risk group changes at a later date. An increase in poverty over time corresponds to an increasing prison population later on. Lagging indicators may be used to infer the past, unobserved history of a physical systems from automobile repairs to stellar explosions. This paper demonstrates code in Base SAS® for identifying leading and lagging indicators and measuring the difference in time between two linked behaviors. A facility is included for addressing the presence of missing data by suppressing points in time where the data are insufficient to support accurate results. Examples are given in biostatistics, social sciences and astrophysics as well as econometrics to demonstrate how leading and lagging indicators may be used to better understand correlated past and future behaviors.

Keywords: Leading, Lagging, Time Series

INTRODUCTION

Mathematically, leading and lagging indicators are correlations where the maximum correlation is observed with a temporal offset between the two observables being compared. In this technique, the offset is created by adding a constant value to the time of one set of observations. The correlation coefficient between the two sets of observations is then calculated using the observed times for one set and the offset times for the other set. To find leading and lagging indicators, a series of incremental offsets may be used, with the correlation coefficient calculated for each offset amount. Where a high level of correlation or anti-correlation is found, the offset with the highest absolute correlation coefficient may indicate the presence of a time lag between two related events.



The above figure presents a time series of three observed physical quantities of a variable star designated HR 1040. The amplitude of the wavelets (red circles in the bottom plot) observed in the silicon data acts as a leading indicator, correlated with variation in the two Hydrogen features in the top and center plots. The hydrogen features follow the silicon data with a time lag of 12 days. Conversely, the simultaneous large variation in the two hydrogen features constitutes a lagging indicator for the silicon wavelet. The presence of such an event in hydrogen is strongly predictive of wavelet in silicon 12 days earlier, even if the wavelet is not directly observed.

While two factors may appear uncorrelated when no time lag is used, a correlation may become visible when a time-varying lag between the two time series is used. Without a time lag, the quarterly data below appear to be unrelated, with a correlation coefficient of 0.399 (left). However, if a time lag is introduced, a correlation of 0.747 is found if the second leads the first by two quarters

Correlations of X with varying lead or lag interval of Y

Lead / Lag	(None)		-3 Weeks		-2 Weeks		-1 Week		Even		+1 Week		+2 Weeks	
Correlation	0.399		0.718		0.747		0.526		0.399		0.193		-0.355	
Week	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
				1.00										
				1.12		1.00								
				0.89		1.12		1.00						
Week 1	0.22%	1.00	0.22%	1.06	0.22%	0.89	0.22%	1.12	0.22%	1.00	0.22%		0.22%	
Week 2	0.526	1.12	0.80%	1.12	0.80%	1.06	0.80%	0.89	0.526	1.12	0.80%	1.00	0.80%	
Week 3	0.56%	0.89	0.56%	1.05	0.56%	1.12	0.56%	1.06	0.56%	0.89	0.56%	1.12	0.56%	1.00
Week 4	-0.50%	1.06	-0.50%	0.69	-0.50%	1.05	-0.50%	1.12	-0.50%	1.06	-0.50%	0.89	-0.50%	1.12
Week 5	-0.18%	1.12	-0.18%	0.74	-0.18%	0.69	-0.18%	1.05	-0.18%	1.12	-0.18%	1.06	-0.18%	0.89
Week 6	0.15%	1.05	0.15%	0.53	0.15%	0.74	0.15%	0.69	0.15%	1.05	0.15%	1.12	0.15%	1.06
Week 7	-1.01%	0.69	-1.01%	0.60	-1.01%	0.53	-1.01%	0.74	-1.01%	0.69	-1.01%	1.05	-1.01%	1.12
Week 8	-1.74%	0.74	-1.74%	0.52	-1.74%	0.60	-1.74%	0.53	-1.74%	0.74	-1.74%	0.69	-1.74%	1.05
Week 9	-1.24%	0.53	-1.24%	0.81	-1.24%	0.52	-1.24%	0.60	-1.24%	0.53	-1.24%	0.74	-1.24%	0.69
Week 10	-0.18%	0.60	-0.18%	0.79	-0.18%	0.81	-0.18%	0.52	-0.18%	0.60	-0.18%	0.53	-0.18%	0.74
Week 11	0.40%	0.52	0.40%	0.97	0.40%	0.79	0.40%	0.81	0.40%	0.52	0.40%	0.60	0.40%	0.53
Week 12	0.10%	0.81	0.10%		0.10%	0.97	0.10%	0.79	0.10%	0.81	0.10%	0.52	0.10%	0.60
Week 13	0.05%	0.79	0.05%		0.05%		0.05%	0.97	0.05%	0.79	0.05%	0.81	0.05%	0.52
Week 14	-0.20%	0.97	-0.20%		-0.20%		-0.20%		-0.20%	0.97	-0.20%	0.79	-0.20%	0.81
												0.97		0.79
														0.97

CALCULATING THE CORRELATION COEFFICIENT FOR A TIME SERIES

The following process implements a moving offset in a time series and calculates the correlation coefficient at each offset value. The code supports testing multiple fields for the presence of a correlation with a time lag.

```
%MACRO OFFSET(OFFSET);
DATA WORK.ORIG_DATA;
   SET DATALIB.ORIG_DATA;
   OFFSET_DATE = DATE;
   KEEP DATE VAR_1 OFFSET_DATE;
RUN;
```

```
PROC SORT DATA=WORK.ORIG_DATA;
   BY OFFSET_DATE;
RUN;

DATA WORK.OFFSET_DATA;
   SET WORK.ORIG_DATA;
   BY OFFSET_DATE;
   OFFSET_DATE = DATE + &OFFSET.;
   KEEP VAR_2 VAR_3 VAR_4 VAR_5 OFFSET_DATE;
RUN;

DATA WORK.CORR_DATA;
   MERGE WORK.ORIG WORK.OFFSET_DATA;
   BY OFFSET_DATE;
RUN;
```

Effect of Data Offset

Original Data

Week	Y_1	Y_2	Y_3
1	0.22%	1.00	0.806
2	0.526	1.12	0.627
3	0.56%	0.89	0.059
4	-0.50%	1.06	0.873
5	-0.18%	1.12	0.504
6	0.15%	1.05	0.447
7	-1.01%	0.69	0.503
8	-1.74%	0.74	0.657
9	-1.24%	0.53	0.186
10	-0.18%	0.60	0.221
11	0.40%	0.52	0.507
12	0.10%	0.81	0.887
13	0.05%	0.79	0.651
14	-0.20%	0.97	0.806

Offset = -3

Week	Y_1	Y_2	Y_3
-2		1.00	0.806
-1		1.12	0.627
0		0.89	0.059
1	0.22%	1.06	0.873
2	0.526	1.12	0.504
3	0.56%	1.05	0.447
4	-0.50%	0.69	0.503
5	-0.18%	0.74	0.657
6	0.15%	0.53	0.186
7	-1.01%	0.60	0.221
8	-1.74%	0.52	0.507
9	-1.24%	0.81	0.887
10	-0.18%	0.79	0.651
11	0.40%	0.97	0.806
12	0.10%		
13	0.05%		
14	-0.20%		

Offset = 0

Week	Y_1	Y_2	Y_3
1	0.22%	1.00	0.806
2	0.526	1.12	0.627
3	0.56%	0.89	0.059
4	-0.50%	1.06	0.873
5	-0.18%	1.12	0.504
6	0.15%	1.05	0.447
7	-1.01%	0.69	0.503
8	-1.74%	0.74	0.657
9	-1.24%	0.53	0.186
10	-0.18%	0.60	0.221
11	0.40%	0.52	0.507
12	0.10%	0.81	0.887
13	0.05%	0.79	0.651
14	-0.20%	0.97	0.806

Offset = 3

Week	Y_1	Y_2	Y_3
1	0.22%		
2	0.526		
3	0.56%		
4	-0.50%	1.00	0.806
5	-0.18%	1.12	0.627
6	0.15%	0.89	0.059
7	-1.01%	1.06	0.873
8	-1.74%	1.12	0.504
9	-1.24%	1.05	0.447
10	-0.18%	0.69	0.503
11	0.40%	0.74	0.657
12	0.10%	0.53	0.186
13	0.05%	0.60	0.221
14	-0.20%	0.52	0.507
15		0.81	0.887
16		0.79	0.651
17		0.97	0.806

Once the offset has been implemented, the correlation coefficient can be calculated:

```
PROC CORR DATA=WORK.CORR_DATA OUT=WORK.TEMP;
   ODS OUTPUT KendallCorr=KendallCorr;
RUN;
```

```
DATA WORK.TEMP;
   SET WORK.TEMP;
   OFFSET = &OFFSET.;
RUN;

DATA HR_1040.CORR_OFFSET;
   SET HR_1040.CORR_OFFSET WORK.TEMP;
   IF _NAME_ = 'VAR_1';
   KEEP _NAME_ _TYPE_ OFFSET VAR_1_1 VAR_2 VAR_3 VAR_4 VAR_5;
RUN;
```

PROC CORR Output

Week		Y_1	Y_2
-2	Y_1	1.00000	0.80609
-2	Y_2	0.80609	1.00000
-1	Y_1	1.00000	0.87292
-1	Y_2	0.87292	1.00000
0	Y_1	1.00000	0.0000
0	Y_2	0.0000	1.00000
1	Y_1	1.00000	0.50297
1	Y_2	0.50297	1.00000
2	Y_1	1.00000	0.18609
2	Y_2	0.18609	1.00000
3	Y_1	1.00000	0.88662
3	Y_2	0.88662	1.00000
4	Y_1	1.00000	0.65077
4	Y_2	0.65077	1.00000
5	Y_1	1.00000	0.47486
5	Y_2	0.47486	1.00000
.	.	.	.

In the case where some data is missing for certain points in time, the number of data points used to calculate the correlation coefficient will vary with the value of the offset. This can result in having sufficient data to accurately calculate the correlation coefficient for some offset values but not others. This is addressed by capturing n along with the correlation coefficient for each offset value, allowing for suppression of correlation values based on insufficient data:

```
DATA WORK.CORR_N;
   SET WORK.CORR_N WORK.TEMP;
   IF _TYPE_ = 'N';
   N = VAR_1;
   KEEP _TYPE_ OFFSET N;
RUN;
%MEND OFFSET;
```

The following code creates a scatter plot, with the offset value on the horizontal axis and the correlation coefficient at each offset on the vertical axis. The WHERE statement is critical, as it suppresses plotting any correlation coefficient not supported by a sufficient number of observations.

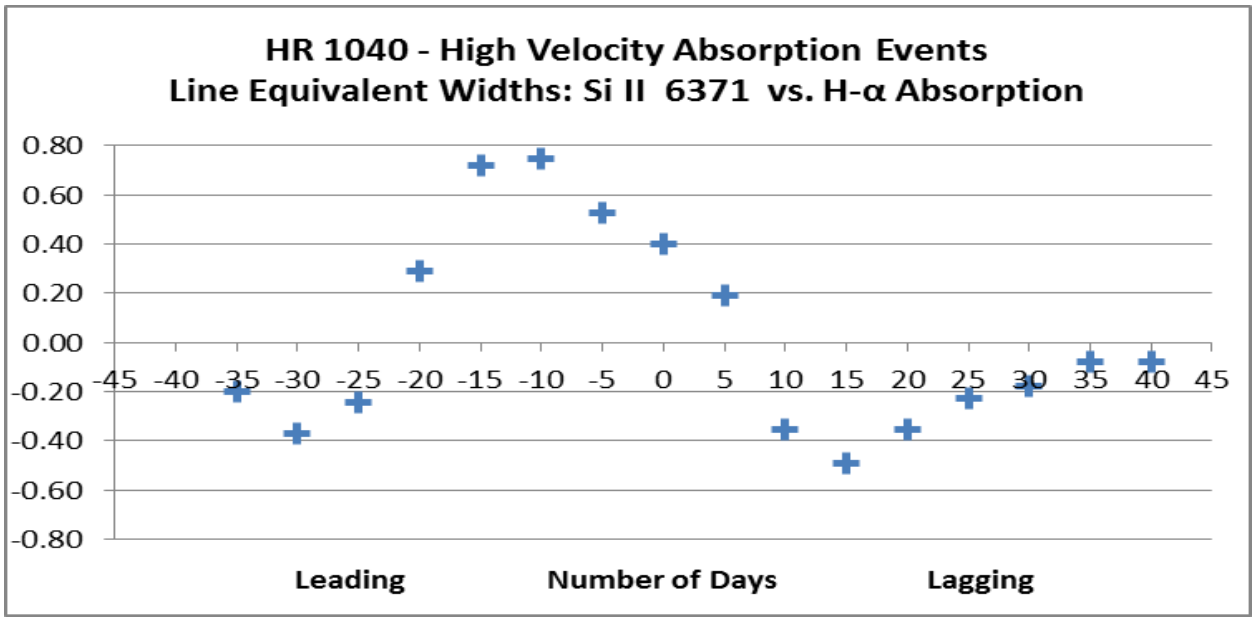
```
axis1 order=(-30 to 30 by 5) label=('OFFSET (DAYS)');
axis2 order=(-1 to 1 by 0.2) label=('CORRELATION COEFFICIENT');
legend1;

PROC GPLOT DATA=HR_1040.CORR_OFFSET;
   PLOT H_ALPHA_AB * OFFSET
   H_ALPHA_LEFT_EM * OFFSET
   H_ALPHA_RIGHT_EM * OFFSET
   SI_II_6347 * OFFSET
   / HAXIS = AXIS1 VAXIS = AXIS2 LEGEND = LEGEND1;

   WHERE CORR_N GE 35;

RUN;
QUIT;
```

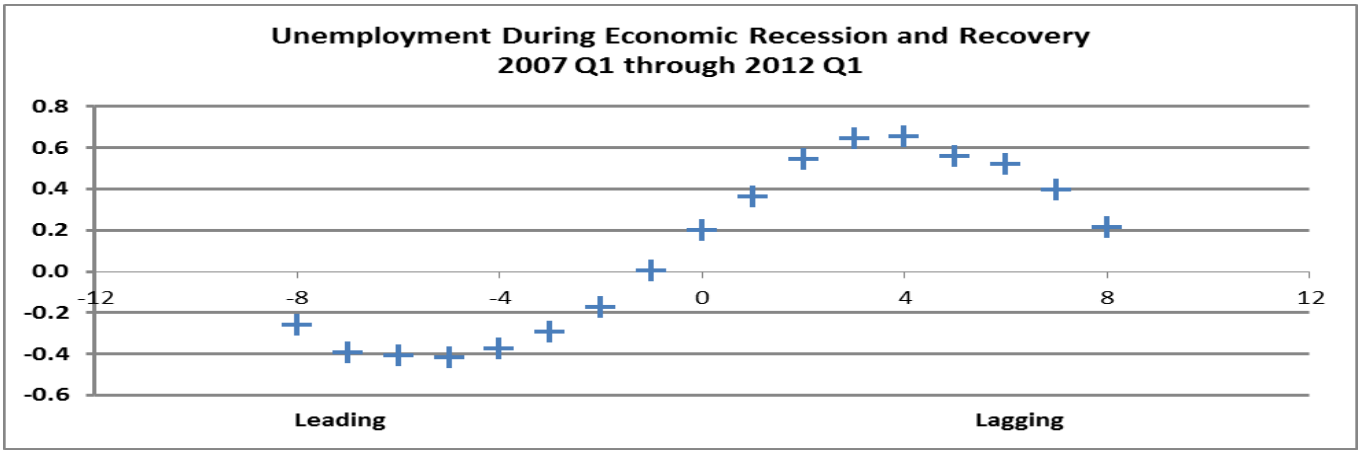
INTERPRETING THE RESULTS



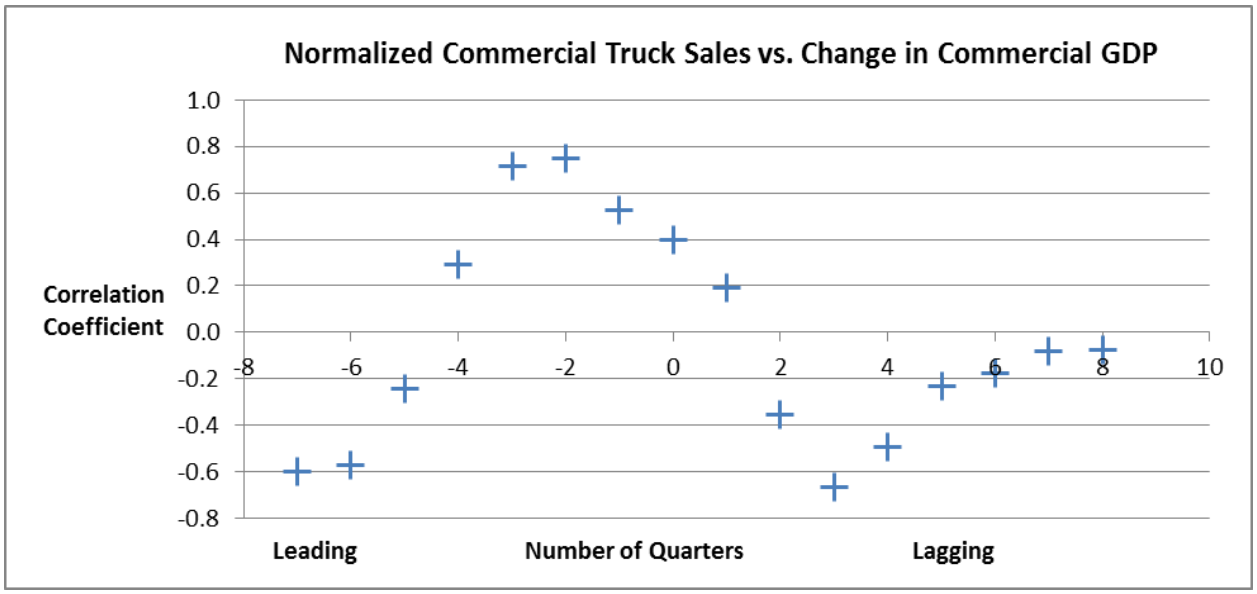
In the above plot, the Silicon II 6371 Å absorption feature leads the H-α feature by 10 to 15 days; interpolation gives a final estimated value of 12 days.

OTHER EXAMPLES

A classic example of a lagging indicator is the unemployment rate, very commonly described as lagging behind changes in GDP. Using data from the United States Bureau of Economic Analysis, this hypothesis can be tested and measured. During the recent cycle of recession and recovery, overall US unemployment is found to lag behind GDP changes by 11 months. A trace of anti-correlation at 5 ½ months with a coefficient of -0.42; too weak to accept but too strong to dismiss. This feature indicates the presence of a weak autocorrelation resulting from the cyclical nature of the economic cycle.

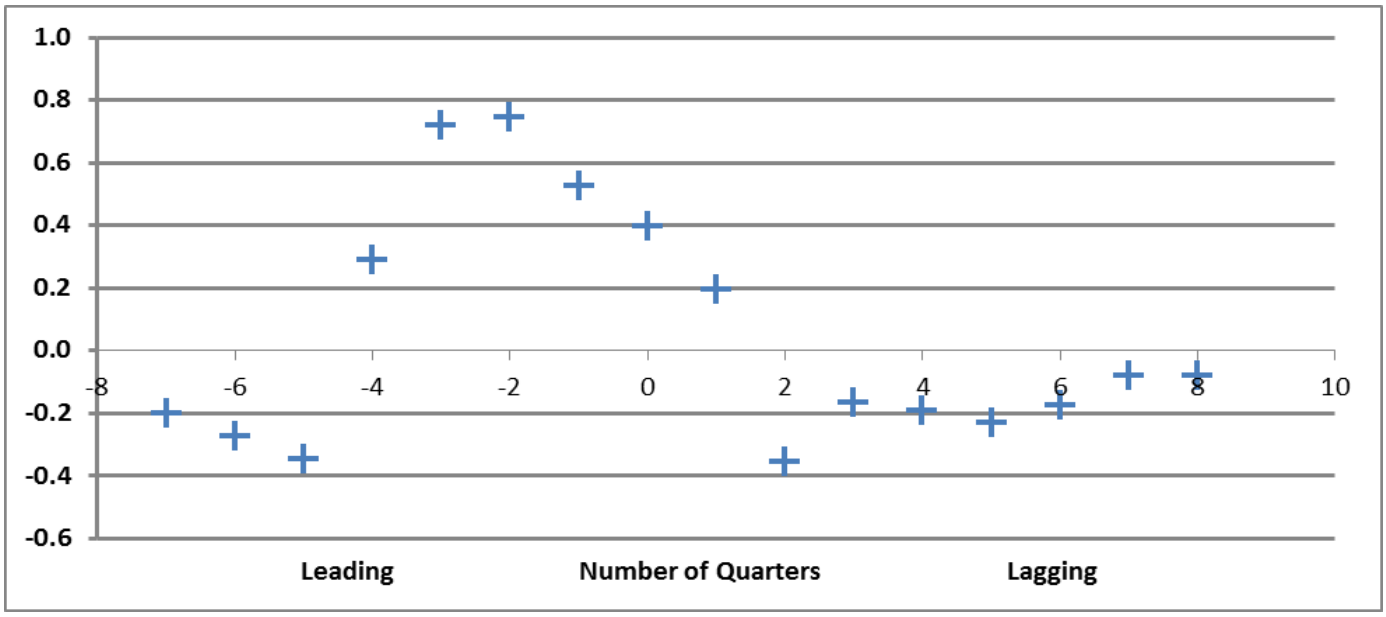


While some factors such as unemployment lag behind changes in GDP, leading indicators also exist. Sales of commercial trucks, for example, while following a highly cyclical pattern reflected alternation between correlation and anti-correlation, show a clear peak in the time varying correlation coefficient of 0.75 at -2 months. Thus, while businesses may delay both layoffs during a recession and re-hiring during a recovery, durable goods purchases are driven by work contracted but not yet paid. Durable goods were first described as a leading indicator by Geoffrey Moore in 1958.

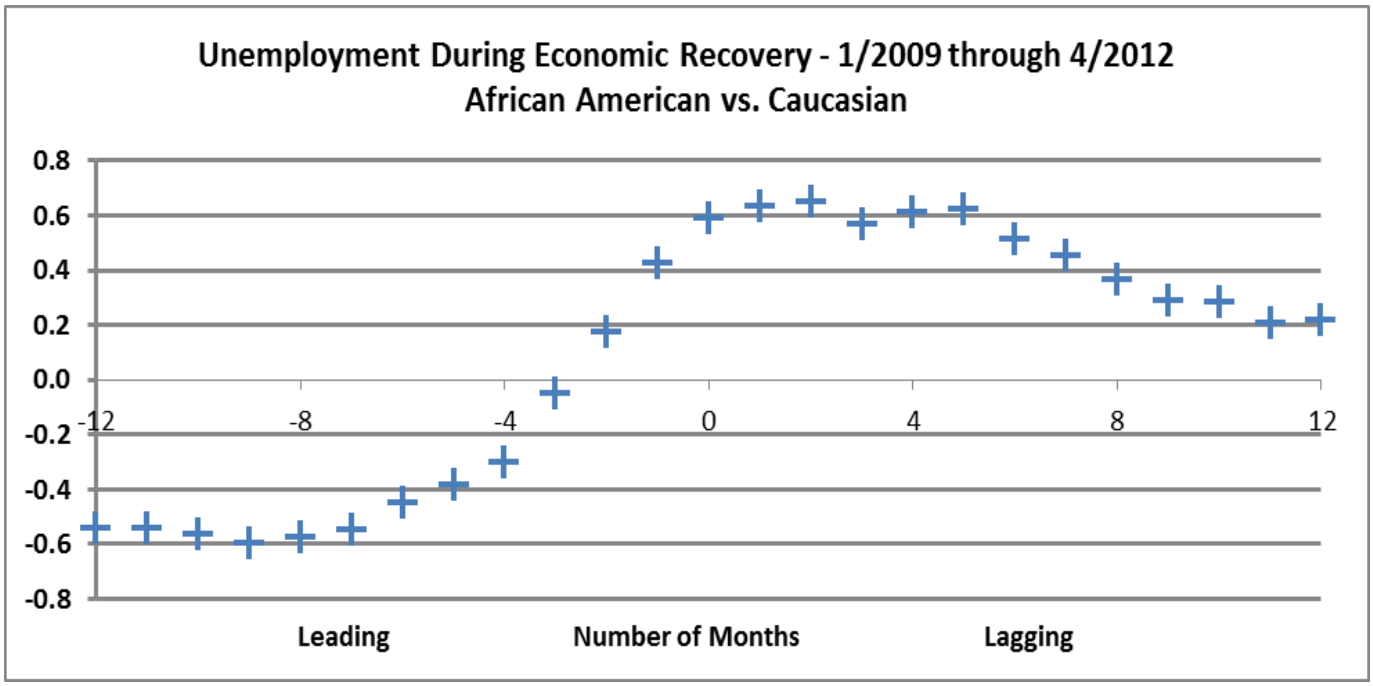


In the case of risk losses, an economic downturn and create a workload crisis as companies seek to reduce staff with increasing losses. If reductions are applied to risk management staff even as workload increases, additional losses can result. A workload forecast model can deliver adequate staffing to reduce losses.

US Unemployment and Mortgage Delinquency
2007 Q1 through 2012 Q1



Where economic factors intersect with public policy and programs, leading and lagging indicators can help to tell a powerful story. In addition to overall US unemployment, the United States Bureau of Economic Analysis publishes breakdown by major ethnic group. Analysis of these breakdowns reveals that African Americans who lost their job during the recent economic down turn were unemployed, on average, for three to six months longer than Caucasian unemployed.



CONCLUSION

Factors in a time series analysis can be tested for leading / lagging behavior by calculating the correlation coefficient for a range of time lags. The amount of time lag between two indicators can be measured by finding the time difference at the maximum correlation coefficient. Leading / lagging indicators find application in many areas beyond economics, including business intelligence / decision science, natural sciences, medical applications and public policy.

REFERENCES

Moore, Geoffrey "Forecasting Industrial Production: A Comment" *Journal of Political Economy*, February, 1958

ACKNOWLEDGMENTS

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David J Corliss
Magnify Analytic Solutions
1 Kennedy Square, Suite 500
Detroit, MI 48224
Phone: 313.202.6323
Email: dcorliss@magnifyas.com