1-1-1986

Economic forecasting

Peter D. Stevens

Peter D. Stevens

May 10,1986

Economic Forecasting

Look at table 1. Is it possible from this data to predict the future? This data is a time series; a list of the values of a variable over time. Using statistical techniques, we can manipulate this data to forecast future values. There are two basic approaches to forecasting, model building and time data analysis. Model building, the more complex of the two, takes basic economic relationships and develops a simple model for the data. The simple income and consumption equations, $Y=C+I$ and $C=\alpha+\beta Y$, X, are a good example of a model. The regression analysis as a method of forecasting shall examined under the model building approach. The second method, data based analysis, derives its forecasts on the past values of the forecasted variable. Under this approach, we shall look at exponential smoothing and the ARIMA techniques. These three techniques, regression, exponential smoothing, and ARIMA modeling shall be explained and then applied to an empirical example.

Model building joins economic theory, statistics, and mathematics to develop models or equations on economic variables. If we can develop equations that simulate the economy, then we can simulate what will happen in the future. Regression is based on the principle that one variable depends on another, $Y=f(X)$. For example, demand for a good depends on its price.

From this basic relationship, regression analysis tries to uncover any linear relationship that exists. The simplest linear relationship is the line itself. To better understand regression, lets use start the equation, Y=mX+b, except in forecasting we use

(1) $Y=\alpha+\beta X + \varepsilon$

where α and β are the unknown parameters of the population upon which regression estimates are based, and ε is error term. The error term which is caused by measurement error, randomness in human behavior, and unobservable values. When we estimate (1), the error term will be replaced by a residual term - a "fit" of the true unknown error.

If we plot our given data we can see a basic linear relationship. What we want is a single line drawn in between the data points or as close to them as possible. The idea behind regression is that there is a line which best fits the data. The line that best fits the data must have the sum of the distances of the data points away from the line is at a minimum, so we use the formula:

(1a) $\text{Min} \sum_{i=1}^{n} (\hat{Y}_i - Y)^2$

where $\hat{Y}_i$ = the fitted or predicted value of Y_i. The equation is squared because we want to deal with only positive values. The line itself is simply an estimate of the relation. To solve the equation, we must find estimate for α and β . The standard practice for indicating estimates, as opposed to population notation, is to use '^' notation, pronounced 'hat'. Our parameters becomes $\hat{\beta}$ (beta hat) and $\hat{\alpha}$ (alpha hat). To find $\hat{\alpha}$ and $\hat{\beta}$, we must take the derivatives of (1a) with respect to α and β. We then get the following:

(2) $\hat{\beta} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{VAR(X)}$

(3) $\hat{\alpha} = \hat{\beta}\,\bar{Y} - \bar{Y}$

and (1) now becomes

(4) $Y = \hat{\alpha} + \hat{\beta} X + e$

What happens if we introduce more variables into the model; for example, the demand for goods is depends on the price of the good, income, and even seasonal factors as we shall see in our empirical evidence. Our initial model then must be expanded to the following:

(5) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

where X_1, X_2, X_3 are different independent variables and where $\beta_0, \beta_1, \beta_2, \beta_3$, are the coefficients of those variables. The general model is of the form

(6) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$

Like the simple regression model, estimates must be made for the β's. This accomplished by more complex matrix techniques which are described in appendix I.

Once we have built these models, we can forecast Y by filling in the estimated values for β's and for the conditional values of the X's. This will allow us to predict what the values of Y conditional on given X's.

The second approach is the data based statistical analysis. This method deals only with the past values of the variable instead of building a model with additional variables as regression does. To derive forecasts, data based methods find relationships between the current time period and its past. We shall look at two techniques, exponential smoothing and ARIMA, an acronym for integrated autoregressive moving average.

To better understand exponential smoothing, let us first look at the most simple method in smoothing, the moving average. To find a moving average all one needs to do is pick a interval period for your data, say 3 months, and take the average or mean for that period of time. This process eliminates the peaks and valleys in the data or any random error, and shows the basic cycle. The large the interval of time you use the flatter your time series becomes. If you choose 7 or 12 months, your time series becomes flatter, and if you pick n to be the interval you would simply have a straight line or just the mean.

Exponential smoothing is similar to the moving average. The basic assumption of this type of smoothing is that present values depend upon past value, but the more you go into the past the less effect they have. To show this you give decreasing weights to these past values,

(7) $F_t = \theta Y_t + (1-\theta) F_{t-1}$

where F_t is the forecast of the current period, F_{t-1} is the forecast of period $t-1$, θ is the weighing factor and Y_t is the variable of the current period. This Equation only shows two periods. If it is factored out we can better see how the past effects the future to a lesser degree.

(8) $F_t = \theta Y + \theta(1-\theta) Y_{t-1} + \theta(1-\theta) Y_{t-2} + \theta(1-\theta) Y_{t-3} + \ldots$

where F_t is the forecast, $Y_t, Y_{t-1}, Y_{t-2}, Y_{t-3}$ are the values of Y at periods t,t-1,t-2,t-3, and θ is the weighing factor.

Estimates for θ are ad hoc. There is no rigorous method of finding θ. Trial and error seems to be the best method. Business often use this method because it is easy to compute, and once you find a θ that gives you a good prediction, you have a forecast.

The second method of data based analysis is somewhat a combination of the first two but much more rigorous than exponential smoothing. Before we discuss ARIMA methods, the similar ARMA shall be explained. ARMA method is made up of two equation parts, autoregressive (AR) and moving average (MA). The equation for AR is similar (6); however, instead of regressing on separate individual variables, AR regresses on its self, auto. The general equation for AR (p) is of the form:

(9) $Z_t = \phi Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \ldots + \phi_p Z_{t-p} + \varepsilon_t$

where $Z_t, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-p}$ are values of Z in time t,t-1,t-2,t-3 and t-p, and ε_t is the error term

The other part of the ARMA model is moving average. The MA is similar to the simple moving average discussed earlier only on a theoretical basis. In this process, the MA is not a interval average, but is stochastic in nature, based on moving averages of current and lagged white noise error terms. The basic assumption behind MA is that current cycles fit models based on current and past errors or white noises. You must remember that we are only trying to predict Y; we are not attempting to set up a model that will explain Y. Some data just happens to fit best in time series based on linear combinations of current and past errors.

The equation for MA (q) is similar to the AR, but now the error term, , goes into the past.

(10) $Z_t = \varepsilon_t - \psi_1\varepsilon_{t-1} - \psi_2\varepsilon_{t-2} - \psi_3\varepsilon_{t-3} - \ldots - \psi_q\varepsilon_{t-q}$

where $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}$, ε_{t-q} , are white noise and ψ_1 , ψ_2 , ψ_3, ψ_q, are weighing factors.

When you combine the autoregression and the moving average, you get are third possibility the ARMA (p,q).

(11) $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \varepsilon_t - \psi_1\varepsilon_{t-1} - \psi_2\varepsilon_{t-2} - \ldots \psi_q\varepsilon_{t-q}.$

The basic assumption here is that the best fit line for Y is based upon both on its past values and on its current and past errors.

We can further modify the above equations by constructing an ARIMA, an integrated autoregressive/moving average. This method, along with ARMA modeling, is known as the Box/Jenkins approach, due to Box and Jenkins (1975)). Data often contains long term or seasonal trends. If the data has any non stationary component, trend, then it must be removed so we can properly forecast.(Box and Jenkins (1975)) This stationary transformation can be accomplished by the process of differencing. Differencing is accomplished simply by adding a (1-B) or (1-B ^{S}) term which represents the nonstationarity and seasonal nonstationarity, respectively.

The Box-Jenkins method as mentioned earlier is a methodology of determining which of the models and parameters to use. This model has basically three steps: 1) identify the model, 2) estimate the model, 3) diagnostic checking of the model. Current research (Tsay and Tiao, 1984) has shown optimal methods to identify what type of ARMA model best fits the data. One method is based on the simple auto correlation function. Autocorrelation, like autoregression, is self correlation of the same time series. By lagging the time series, we can produce additional time series which can be correlated upon each other. This process shows us the dependence of the same values but at different time periods (Wheelwright and Makridakis(1985)).

Another identification tool is the partial autocorrelation function (PACF). This process is very complicated and will only be mentioned here. (for further information see Nelson(1973)). Another complicated tool is the ´corner´ method developed by Tsay and Tiao (Tsay and Tiao(1984)). This method used to find mixed ARMA (p,q) models where neither p or q equal zero.

Once the θ_i´s and Φ_i´s have been estimated, obtaining $\hat{\Theta}_i$ and $\hat{\Phi}_i$, forecasted equations can be constructed. For example, the equation for an AR(1) would be $\hat{X}_{n+1}=\hat{\Phi}_1 X_n$ for 1 period ahead, $\hat{X}_{n+2}=\hat{\Phi}_1 X_{n+1}$ for two periods ahead, etc..

II. An Empirical Example

Table 1 shows us the data that we will be using in this example. I forecasted income from this data for a one year period, periods 157 to 168. Table 2 shows the forecasted values derived from the three methods. As we can see from figure 1, regression analysis has done the best at predicting the actual values of Y.

This brings up an important concern about forecasting. As you remember regression is a model building method. In model building two errors can make data meaningless, data collection error and model construction error. The relationship I used was $Y = f(P, Q)$ where P is price of alcohol and Q is the quantity of alcohol. This model is ridiculous; price and quantity of alcohol have no baring on income. It was just luck that caused the predicted values of Y to be so good. The R square coefficient is .9379 which is significant. So, when one forecasts he or she must be vary careful to use logical models and good data.

A better model is $Q = f(P, Y)$. If we forecast the same 12 months for quantity of alcohol using regression. As one could imagine, demand for is very seasonal, so to set up a good model we must take into account this seasonality by creating dummy variables for the months. Our regression equation becomes

(12) $Y = \beta_0 + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + \ldots + \beta_{12} D12 + \beta_{13} P + \beta_{14} Q + e$

where $D1, D2, D3, \ldots, D_{12}$ are the variables for the months, P is price and Y equals income. Solving this equation, we get the forecast shown in table 3. We can see that alcohol in june and in december because of the holiday seasons. We see though that the forecasts are not that good. I think this is due seasonal variables we had to include. The next step would be to try other models and methods in a hope to develop a good forecast.

III. Summary

We have looked at three methods of forecasting and a basic empirical example. Many consider forecasting as an art from; hopefully, this paper has eliminated the mysticism of basic forecasting and has replaced it with a basic understanding of currently used forecasting techniques.

Appendix I.

The purpose of this appendix is to give a basic understanding of the advanced process of solving a multiple regression. This process depends on matrix algebra to solve for the betas. The general equation for regression is

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\dots+\beta_k X_k$$

When using a large amount of data this can be stated in Matrix form.

$$\begin{bmatrix} Y \\ Y \\ Y \\ \cdot \\ Y \end{bmatrix} = \begin{bmatrix} 1 & X_{(1,1)} & X_{(1,2)} & X_{(1,3)} & \cdots & X_{(1,k)} \\ 1 & X_{(2,1)} & X_{(2,2)} & \cdot & & \cdots \\ 1 & X_{(3,1)} & \cdot & \cdot & & \cdots \\ \cdot & \cdot & \cdot & \cdot & & \cdots \\ 1 & X_{(n,1)} & \cdot & \cdot & \cdots & X_{(n,k)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

The matrix when reduced to shorthand matrix notation looks similar to the simple regression model,

$$\underset{n\times 1}{\underline{Y}} = \underset{n\times k}{X}\,\underset{k\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{\varepsilon}}$$

If the matrix is of full column rank(i.e. there are k+1 linear columns) then the same calculus equation mentioned under regression can be applied here.

min $\sum(\hat{Y}_i - Y_i)^2$ where $Y_i=\hat{\beta}_0+\hat{\beta}_1 X_1 + \dots \hat{\beta}_k X_k$ and yields the solution vector, $\underline{b} = (X'X)^{-1}X'Y$ where "$'$" denotes transpose and "-1" denotes matrix inverse. This fitted value $\hat{Y}_i$ can also be thought of as the predicted Y value. If we can give reasonable forecasts of $X_1, X_2, \ldots, X_k$ say in period n+1, then we can predict Y_i in period n+1 by using the formula given for $\hat{Y}_{n+1}$. The variance of this forecast can also be determined. $\hat{Y}_{n+1}$ can be expressed in the following matrix form. Let $\underline{X}_{n+1} = (1, X_{1,n+1}, X_{2,n+1}, \ldots, X_{k,n+1})$. Then $\hat{Y}_{n+1} = \underline{X}'_{n+1} b$. The forecast error can be determined using the formula $\mathrm{Var}(\hat{Y}_{n+1}) = \underline{X}_{n+1}' \mathrm{Var}(\underline{b})\underline{X}_{n+1}$ where $\mathrm{Var}(b) = s^2(X'X)^{-1}$ is the variance of the estimated beta vector, b. This way a confidence interval for the forecasts can be calculated.

Table 1

SAS

OBS	I	Y	P	Q
1	1	709.5	101.7	23988100
2	2	710.5	101.8	24157600
3	3	711.3	101.5	28862000
4	4	710.3	101.9	28872800
5	5	718.0	101.9	30688400
6	6	722.0	102.1	30457100
7	7	727.4	102.4	28583300
8	8	730.6	102.7	28408900
9	9	731.4	103.0	27791700
10	10	732.7	103.8	33030700
11	11	731.8	103.9	33708800
12	12	734.0	104.0	42843100
13	13	734.8	103.7	28736400
14	14	736.8	103.8	25099400
15	15	740.3	104.1	30994000
16	16	755.9	104.3	28922000
17	17	751.5	104.8	28395100
18	18	749.2	105.1	29304100
19	19	757.7	105.5	28008100
20	20	760.7	105.7	27141600
21	21	762.1	105.4	30223800
22	22	755.3	105.5	32493800
23	23	754.5	105.7	35167400
24	24	757.8	105.8	47713300
25	25	769.5	104.8	24602400
26	26	770.2	105.8	24988800
27	27	773.4	105.8	31487100
28	28	773.6	105.9	28755500
29	29	775.0	106.0	29223900
30	30	791.1	106.2	33791000
31	31	776.3	106.3	28975900
32	32	781.5	107.0	30651500
33	33	782.2	107.0	30375800
34	34	782.2	106.8	31372500
35	35	783.4	108.9	38640800
36	36	790.1	107.9	47289400
37	37	788.1	107.4	28084200
38	38	793.1	108.5	28822300
39	39	794.9	108.5	33369700
40	40	798.2	108.8	24731700
41	41	803.7	108.5	32481200
42	42	784.1	108.5	35187800
43	43	806.2	108.0	27824900
44	44	814.2	108.8	28346000
45	45	814.8	108.8	30710300
46	46	828.8	108.9	33731800
47	47	839.8	108.8	39528200
48	48	844.5	109.0	48343100
49	49	845.9	108.1	28263100
50	50	857.9	108.1	28756400
51	51	861.8	109.2	33869800
52	52	862.5	109.1	30485800
53	53	861.1	108.9	33661800
54	54	868.2	108.8	33700000
55	55	868.4	109.0	28690100
56	56	864.1	109.0	32358100
57	57	871.3	109.5	29493500
58	58	873.8	109.5	35109200
59	59	874.6	109.4	41071900
60	60	871.3	109.5	47131800
61	61	865.5	109.6	29781100
62	62	868.6	109.7	28333700
63	63	855.6	109.8	34106000
64	64	857.6	109.8	33534200
65	65	860.2	109.7	34227000
66	66	861.4	109.7	34032800
67	67	864.6	110.5	32858300
68	68	859.1	111.2	32470800
69	69	854.5	112.0	30324000
70	70	855.0	112.5	35788200
71	71	849.1	112.8	38928200
72	72	878.5	113.0	48888700
73	73	843.4	113.1	31184700
74	74	845.7	113.5	28902600
75	75	846.1	113.5	32681900
76	76	855.2	113.4	33151800
77	77	922.4	113.2	28838000
78	78	886.7	113.7	35712800
79	79	873.0	113.7	33362300
80	80	878.8	114.4	32055100
81	81	882.8	114.8	32377700
82	82	888.0	115.0	38867200
83	83	885.8	114.8	38472200
84	84	888.4	114.9	52134700
85	85	893.0	115.6	30218300
86	86	901.7	115.4	28100200
87	87	903.8	115.8	38860100
88	88	904.4	115.8	33018000
89	89	904.5	115.8	32722200
90	90	903.3	116.8	38450100
91	91	906.3	116.2	33048100
92	92	908.1	116.2	31798000
93	93	909.3	116.5	33248300
94	94	909.0	116.5	34368200
95	95	915.8	116.0	41809400
96	96	918.4	115.7	53408800
97	97	919.2	115.8	28867000
98	98	910.8	116.8	28162000
99	99	928.5	116.7	38873000
100	100	930.7	116.6	33284000
101	101	932.8	117.2	32759000
102	102	938.4	118.8	38824000
103	103	948.2	117.8	31013000
104	104	950.8	118.0	33255000
105	105	958.4	118.3	34327000
106	106	982.3	118.8	34832000
107	107	985.2	118.8	43838000
108	108	970.3	118.8	54828000
109	109	980.5	120.2	30548000
110	110	971.8	120.8	30189000
111	111	981.7	120.7	38415000
112	112	988.0	121.2	33418000
113	113	981.2	122.5	54380000
114	114	981.3	122.6	38752000
115	115	982.3	122.7	32057000
116	116	984.9	122.7	35770000
117	117	995.4	122.9	34208000
118	118	999.7	123.4	37283000
119	119	1002.8	123.9	44518000
120	120	1011.7	124.1	62928000
121	121	1017.5	124.4	32034000
122	122	1571.1	124.5	30375000
123	123	1588.1	124.8	37426000
124	124	1584.8	125.4	34457000
125	125	1601.5	126.5	38800000
126	126	1609.0	126.8	38983000
127	127	1631.8	127.4	32435000
128	128	1640.4	128.4	35038000
129	129	1647.8	129.0	35555000
130	130	1663.1	128.4	38547000
131	131	1678.4	130.7	44247000
132	132	1695.0	130.8	53748000
133	133	1758.9	131.6	33122000
134	134	1763.3	132.6	32018000
135	135	1775.1	132.8	34684000
136	136	1775.8	133.4	35428000
137	137	1783.8	133.9	34698000
138	138	1793.0	134.7	37281000
139	139	1824.9	135.2	36349000
140	140	1837.7	135.9	34478000
141	141	1859.2	137.6	34682000
142	142	1881.2	137.7	38249000
143	143	1900.1	138.9	42828000
144	144	1918.0	138.9	54398000
145	145	1931.4	140.0	32841000
146	146	1946.6	140.0	30877000
147	147	1965.4	141.6	38878000
148	148	1978.6	141.3	37032000
149	149	1984.8	142.5	34418000
150	150	1996.3	143.0	37718000
151	151	2025.5	143.7	35909000
152	152	2043.2	144.7	33207000
153	153	2057.3	145.4	34747000
154	154	2078.7	144.0	39071000
155	155	2088.8	144.8	41708000
156	156	2091.8	146.0	54088000
157	157	2108.2	146.9	30898000
158	158	2177.5	146.8	30220000
159	159	2124.4	147.0	35886000
160	160	2148.3	147.2	35127000
161	161	2152.5	148.2	33288000
162	162	2155.6	148.3	37202000
163	163	2194.8	148.9	33479000
164	164	2196.7	149.6	32745000
165	165	2202.7	149.8	34931000
166	166	2213.5	150.0	38327000
167	167	2221.8	150.7	43130000
168	168	2235.1	150.2	51878000

FIGURE [illegible]

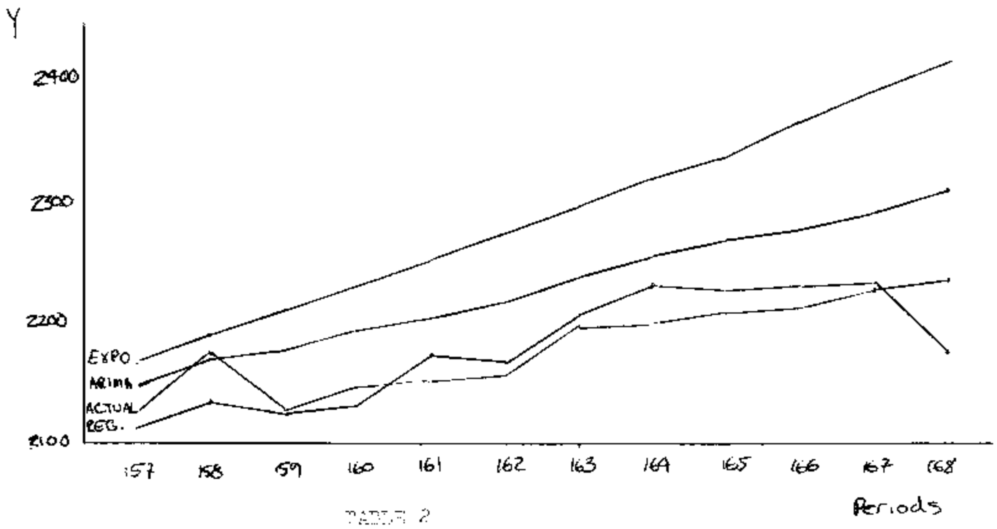

TABLE 2

	ARIMA	Exponential Smoothing	Regression	Actual
157	2150	2162	2105	2[illegible]
158	2[illegible]65	2[illegible]	2[illegible]5[illegible]	217[illegible]
159	2[illegible]78	2206	2[illegible]25	2[illegible]2[illegible]
160	2[illegible]2	2226	2128	2[illegible]46
161	2203	22[illegible]	2[illegible]	2[illegible]52
162	22[illegible]7	2272	2[illegible]3	2[illegible]5[illegible]
163	2235	2294	2[illegible]98	2[illegible]
164	2248	2315	2226	2[illegible]97
165	2261	2338	2228	22[illegible]
166	2276	235[illegible]	2226	22[illegible]
167	229[illegible]	2365	2223	2222
168	2306	2405	2[illegible]72	2235

FIGURE 2

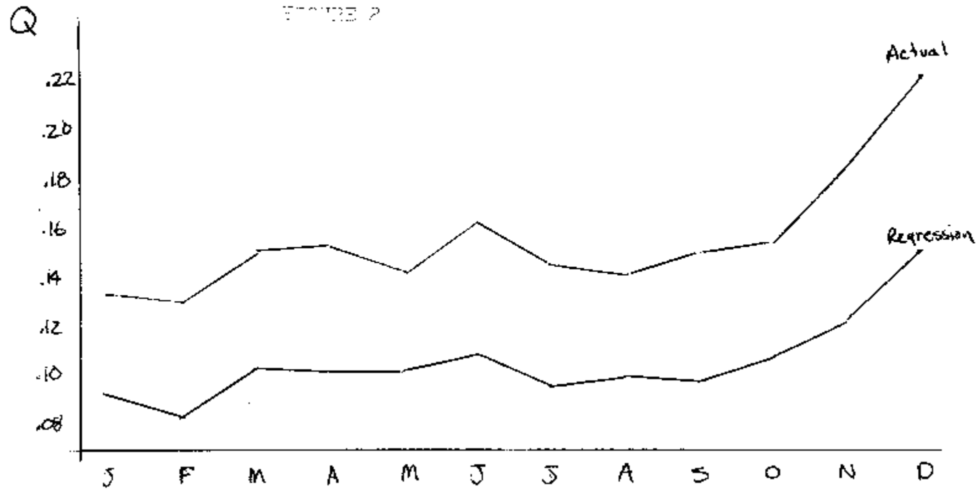

Table 3

	Regression	Actual
Jan	.092	.132
Feb	.087	.130
March	.106	.156
April	.102	.150
May	.102	.145
June	.109	*.16[illegible]
July	.0[illegible]	.144
Aug	.09[illegible]	.14[illegible]
Sept	.09[illegible]	.152
Oct	.1[illegible]	.157
Nov	.123	.186
Dec	.155	.223

References

Box, G.E.P., and Jenkins, G.M. (1975). Time Series Analysis: Forercasting and Control, San Francisco: Holden Day.

Nelson, Charles R. (1973). Applied Time Series Analysis for managerial Forecasting, San Francisco: Holdenday.

Tsay, R.S. and Tiao, G.C. (1984). Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Non-stationary ARMA models. Journal of the American Statistical Association 79 84-96.

Wheelwright, Steven C. and Makridakis, Spyros (1985). Forecasting Methods for Management, New York: Wiley-Interscience