Appendix 3

REPORT ON HEVERT REGRESSION ANALYSIS OF IMPACT OF PRICE AND INCOME ON UNCOLLECTIBLES

By Edward Bodmer

Consultant to Commission Staff

A. Interpretation of R-Squared

The R-squared statistic in a regression equation measures the total variation in the dependent variable – uncollectible expenses – divided by the amount of variation that is explained by all of the variables in the regression equations. If the regression equation that includes price, income and other variables would perfectly predict the historic actual level of uncollectible expense, then the R-squared would be 100%. On the other hand, if the regression equation predicts nothing, then the R-squared is zero. The more nuanced issue in this case is whether the R-squared can be interpreted as the amount of variation beyond the control of management when variables such as lagged uncollectible accounts expense and seasonal adjustment dummy variables for the moratorium period are included in the regression equation.

It is helpful in interpreting R-squared results to consider a regression equation where only income and price are included in the analysis. In this case, one could use the R-squared statistic to conclude how much of the change in uncollectible expense over time is due to changes in price and income which are presumably out of the control of management. For example, if there is an increase in write-offs of $100, and the R-squared is 69%, then one could say that $69 of the $100 increase is due to price and income.

CMP's regression analysis includes two lagged variables associated with the uncollectible expenses themselves. These variables capture trends in the data and seasonal fluctuations. If lagged dependent variables are used in a regression, then attempting to interpret the R-squared as representing things outside the control of management becomes a much murkier issue.

To illustrate this point, assume that a hypothetical company is evaluating its management of operating costs and is attempting to discover whether it could have been more efficient in controlling costs. In making the analysis, pretend that the company uses a lag of the expense variable as did CMP. Assume that the company finds that expenses have increased, but it also finds that almost all of the increase in expenses can be explained with a regression equation that forecasts the operating expense as a function of lagged expenses in the prior year. Say that this company which uses a regression equation to evaluate its efficiency, instituted a policy of allowing employees to have memberships in a fancy golf club three years ago. The next year, the golf memberships were retained, but the company also allowed each employee to use first class tickets on planes whenever they travel. In the third year, the company went even further and paid the cost of cigarettes for each employee who wanted them. Because of the gradual movement in expenditures, lagged variables may suggest that most of variation is due to the lagged dependent variable. The measured R-squared, results from gradual movements in efficiency, and not factors beyond the control of management. In truth, all of the trend in expenses is due to management behavior and nothing is out of the control of management. In the hypothetical example, interpreting the R-squared as representing things outside of the control of management does not work anymore. As explained below, in these circumstances where the lagged dependent variable seems to explain much of the variation, it is much better to make an adjustment for autocorrelation, as CMP regularly does in its regressions of energy sales.

In addition to including lagged dependent variables, CMP adds a variable that adjusts for seasonality in the data. This variable, called moratorium for shut-offs, is a dummy variable that has a value of one from December to March (even though the moratorium period begins in November and extends to April). When the seasonal adjustment variable is included directly in the regression equation, the R-squared may appear high, but this is due to smoothing of seasonal swings and should not be attributed to factors out of the control of management. As with the lagged dependent variable discussed above, inclusion of a seasonal adjustment variable renders the R-squared much more difficult to interpret.

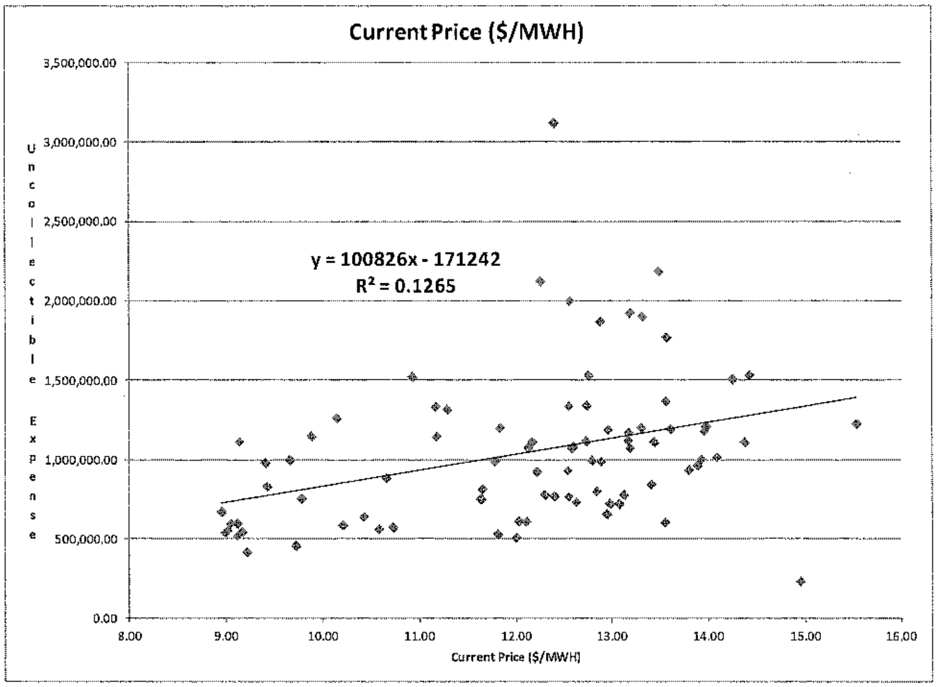
To illustrate distortions that arise from using R-squared results without accounting for seasonal adjustments in an appropriate way, we have created another hypothetical example. This time, pretend that the board of directors of a company is trying to use a regression analysis to measure its effectiveness in setting the salary of the chief executive officer of the company. The board assumes that if the R-squared is 70%, then 70% of changes in the compensation of the CEO are related to factors out of its control and should not be considered in the evaluation. Now pretend that for the first four years of the analysis, the CEO receives only $20 million in salary and an additional $20 million as a bonus payment each September. If a seasonal dummy variable is included in the regression, it would look like the regression equation explains much of the year to year variation even though there is no variation in the annual data. Next, assume that in year five, the CEO manipulates the compensation formula and is able to increase both his salary and September bonus each to $50 million. Assume that the entire increase to $100 million (salary and bonus) from $40 million is due to problems with the compensation formula. However, if a seasonal dummy variable is used in the formula, the regression equation would have a very high R-squared and it would appear that much of the variation from month to month is due to the September bonus. Here, as in the case with the lagged dependent variable, the R-squared cannot be used to make conclusions about what is in control of management and what is out of the control of management. In the hypothetical example, the entire increase in salary and bonus of $60 million in the fifth year is due to problems with the CEO. If the R-squared is used to interpret responsibilities of the board of directors, then an incorrect conclusion would be made that much of the compensation increase is due to factors out of control of the board.

To resolve this problem, a two stage analysis can be used. In the first stage, the data should be seasonally adjusted and then in the second stage, the analysis should be made on the seasonally adjusted data. In our hypothetical example, the salary and bonus could be totaled over the whole year. Then, all of the $60 million increase would be in the unexplained portion and the R-squared would be zero.

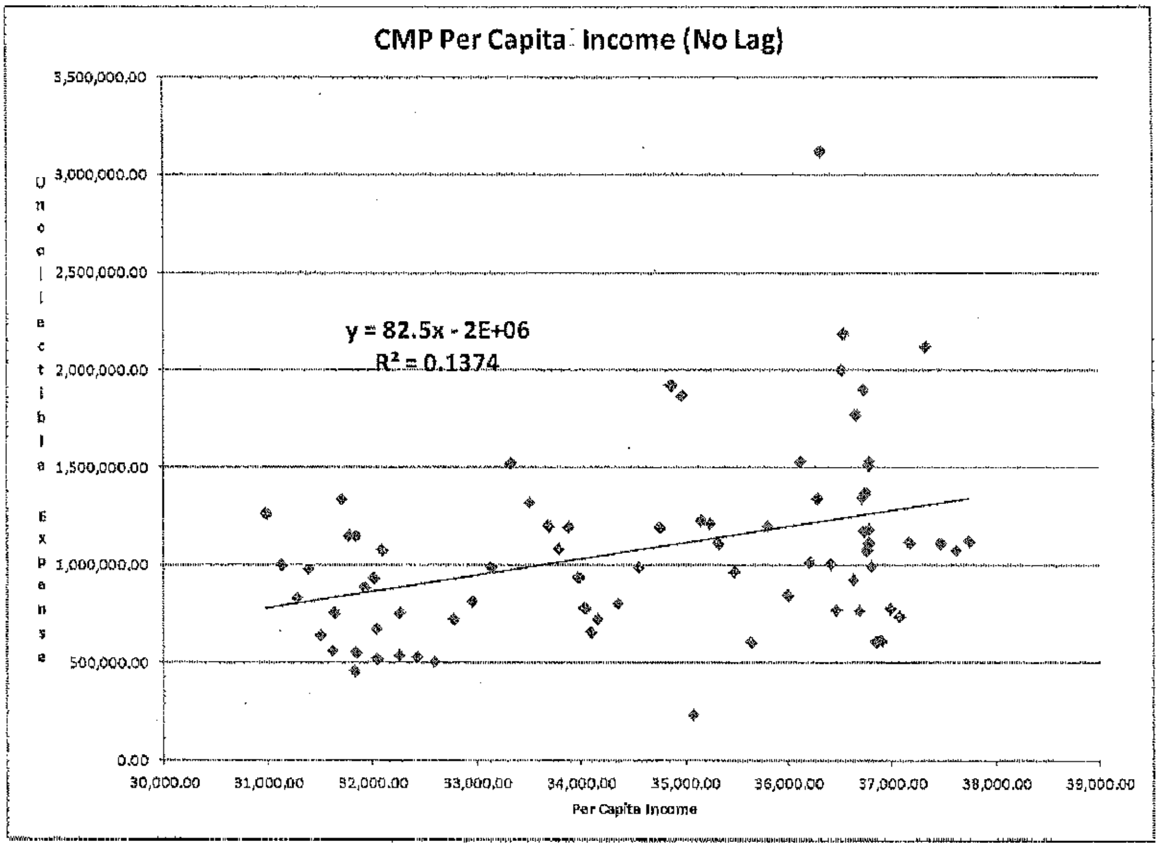
B. CMP Data Used in Regression Analysis of Uncollectible Expenses

Before working through the details of our adjusted regression equations, some of the data used by CMP in its analysis is reviewed in this section. We present scatter plots of price versus uncollectible expense to demonstrate that much of the change in uncollectible expense is driven by price. The scatter plot of current price (with no lags) against uncollectible expenses demonstrates that the CMP data does not have a very strong relationship. Other scatter plots of lagged prices (lags of one to twelve) against uncollectible expense show even weaker relationships.

CMP Price vs. Uncollectibles

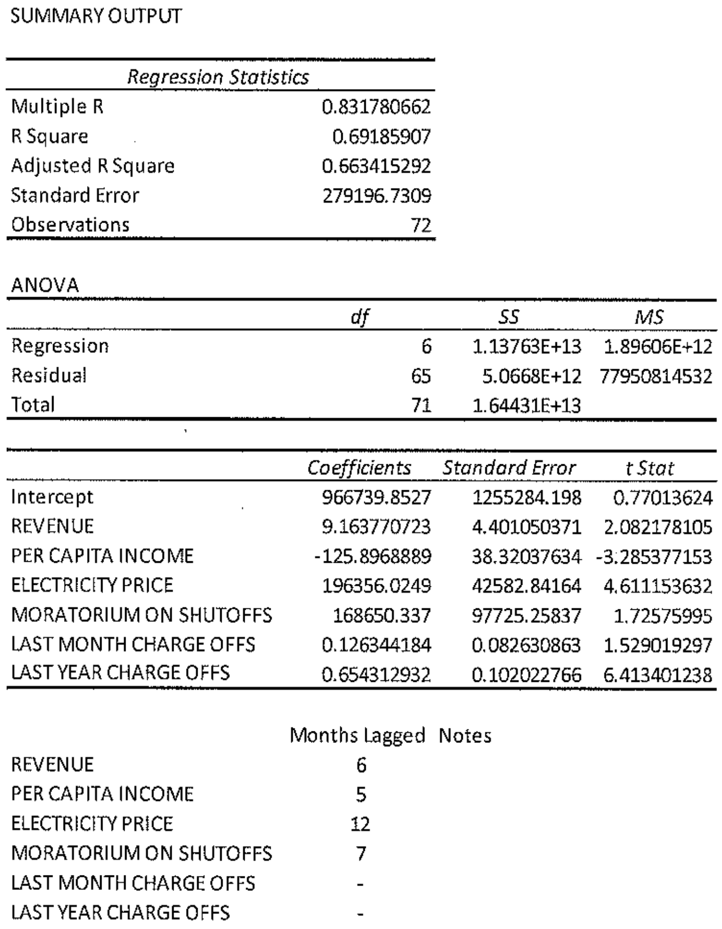


In addition to the examining the relationship between uncollectible expenses and price, a scatter plot of income relative to uncollectible expenses is shown in the graph below (the income has no lags as with the price variable above). As with the price variable for CMP, there appears to be some positive relationship between uncollectible expenses and income. However, this means that the relationship is backwards and goes in the wrong direction. Increases in income should cause the level of uncollectible to increase and not the other way around.



The CMP regression equation is summarized in the table below. The table shows that the R-squared is 69% which seems to imply that at most 30% of the variation in uncollectible expense from period to period is due to factors within the control of management. The t-statistic on the lagged dependent variable is 6.41 in the CMP regression which is larger than the t-statistic for the other variables.[1]

[1] A lower t-statistic implies less possibility that the data can decipher a true underlying relationship.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831780662
R Square	0.69185907
Adjusted R Square	0.663415292
Standard Error	279196.7309
Observations	72

ANOVA

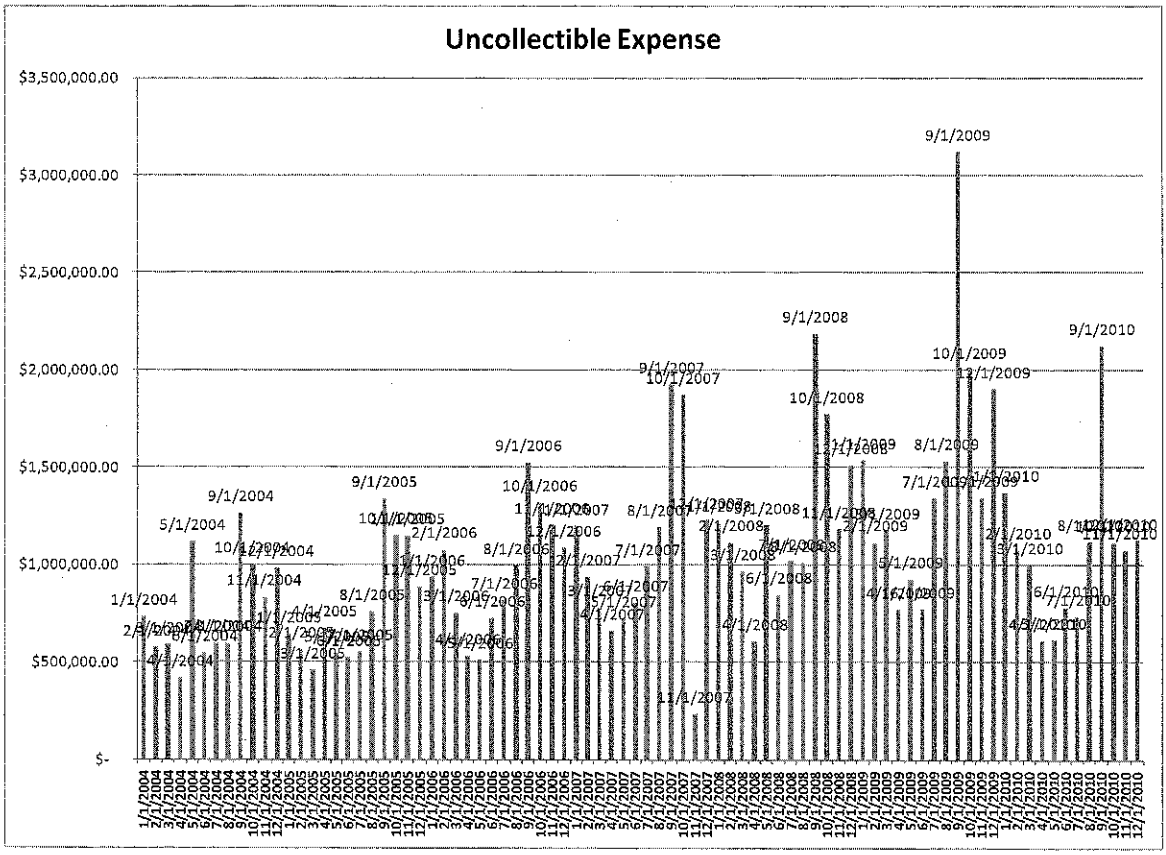
	df	*SS*	*MS*
Regression	6	1.13763E+13	1.89606E+12
Residual	65	5.0668E+12	77950814532
Total	71	1.64431E+13	

	Coefficients	*Standard Error*	*t Stat*
Intercept	966739.8527	1255284.198	0.77013624
REVENUE	9.163770723	4.401050371	2.082178105
PER CAPITA INCOME	-125.8968889	38.32037634	-3.285377153
ELECTRICITY PRICE	196356.0249	42582.84164	4.611153632
MORATORIUM ON SHUTOFFS	168650.337	97725.25837	1.72575995
LAST MONTH CHARGE OFFS	0.126344184	0.082630863	1.529019297
LAST YEAR CHARGE OFFS	0.654312932	0.102022766	6.413401238

	Months Lagged	Notes
REVENUE	6	
PER CAPITA INCOME	5	
ELECTRICITY PRICE	12	
MORATORIUM ON SHUTOFFS	7	
LAST MONTH CHARGE OFFS	-	
LAST YEAR CHARGE OFFS	-	

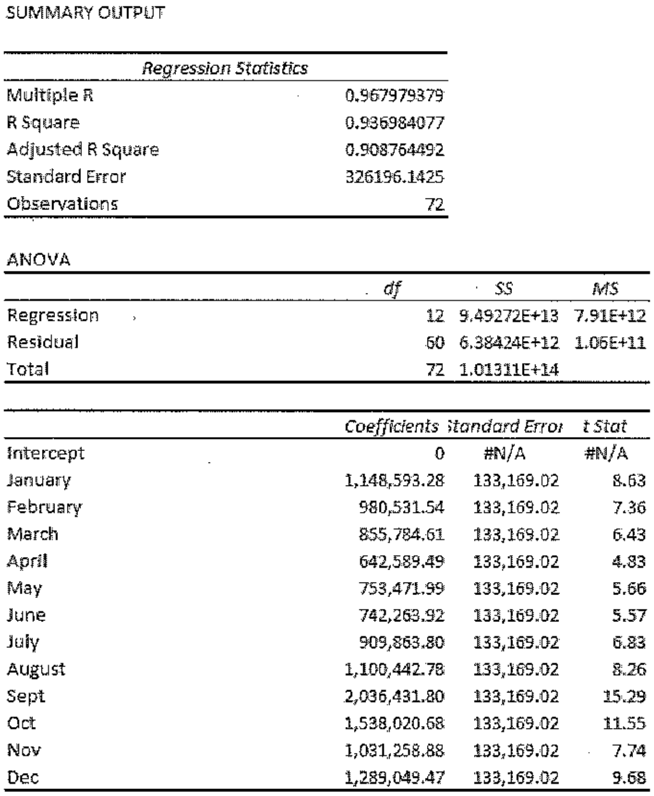
C. Seasonal Adjustment

When CMP presented regression equations for the number of complaints last year as part of Docket No. 2009-217, the Company used annual data. In this case CMP developed the regression equations using monthly data. Use of monthly data means that some variation may arise from seasonal patterns. To review the issue, we first present a graph of month by month uncollectible expenses below. Dates highlighted on the top of each bar show that uncollectible expenses are far higher in the month of September than in any other month and this phenomenon occurs year after year. The graph also shows that expenses in October are relatively high while expenses in April and March are relatively low.



The reason for the very high expenses in September is likely due to CMP "putting the press on" ratepayers before the moratorium period after which consumers cannot be disconnected. This extreme seasonality in the data can cause serious problems with interpretation of a regression equation if it is not dealt with in an appropriate manner. For example, if the income lag is fiddled with enough, one may able to array a decline in income with the month of September. Alternatively, the jump in standard offer prices and the income decrease after the Lehman collapse could be matched to the month of September through changing the lag structure of the price and income variable. In this case, it would appear that the price and income are causing the movement in uncollectible expense, when in fact it is simply CMP's September press that is driving the results.

To resolve issues associated with the high level of seasonality in the data, we have developed a regression equation that computes the effect of each month on the uncollectible expense. This equation uses a series of dummy variables for each month. Results of the seasonal adjustment process demonstrate that most of the movement in the month to month uncollectible expense is simply due to seasonal variation as demonstrated by the R-squared statistic of 94% shown in the table below. The seasonal variables also explain why the lagged dependent variable has such a high level of significance in the regression. Use of the lagged dependent variable however is a very crude way to account for the seasonality and it causes the R-squared to simply pick-up seasonal movements in the data.



SUMMARY OUTPUT

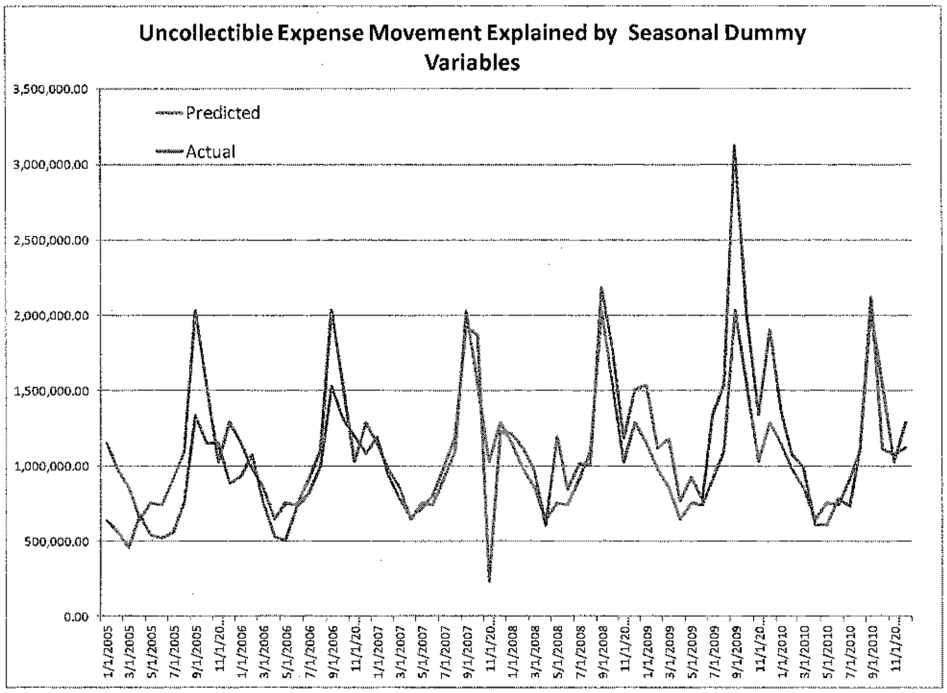
Regression Statistics	
Multiple R	0.967979379
R Square	0.936984077
Adjusted R Square	0.908764492
Standard Error	326196.1425
Observations	72

ANOVA

	df	*SS*	*MS*
Regression	12	9.49272E+13	7.91E+12
Residual	60	6.38424E+12	1.06E+11
Total	72	1.01311E+14	

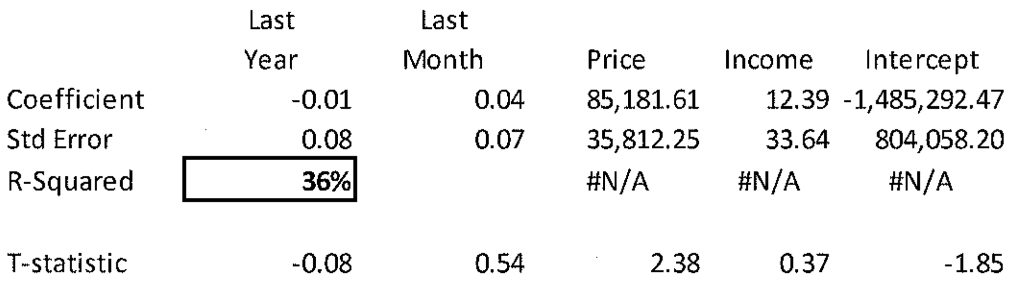
	Coefficients	*Standard Error*	*t Stat*
Intercept	0	#N/A	#N/A
January	1,148,593.28	133,169.02	8.63
February	980,531.54	133,169.02	7.36
March	855,784.61	133,169.02	6.43
April	642,589.49	133,169.02	4.83
May	753,471.99	133,169.02	5.66
June	742,263.92	133,169.02	5.57
July	909,863.80	133,169.02	6.83
August	1,100,442.78	133,169.02	8.26
Sept	2,036,431.80	133,169.02	15.29
Oct	1,538,020.68	133,169.02	11.55
Nov	1,031,258.88	133,169.02	7.74
Dec	1,289,049.47	133,169.02	9.68

Using the coefficients in the above table, the predicted level of uncollectible expense due only to seasonal factors can be compared to the actual level of expenses. The graph below shows that, without any price or income variable, the fit is very close.



Once the seasonal regression is complete, the remaining amount of uncollectible expense not explained by seasonal factors can be computed. This is simply the difference between the two lines in the graph above. This difference is used as the dependent variable in a regression equation to evaluate how much of the variation not due to simple seasonal patterns is caused by price and income levels. Results of this second stage equation are shown in the table below.

Results of the regression equation with seasonally adjusted data show that the R-squared drops to 36%. Perhaps more importantly, the income variable is not significant and has the incorrect sign. This equation with corrected seasonal adjustment shows that the data cannot be used to make conclusions about whether income really affects the level of uncollectible expenses. The income variable and the price variable shown below have the same lags as those assumed by CMP (the price has a lag of 12 months and the income variable has a lag of 5 months). Finally, note that when the seasonal adjustment is made, the lagged dependent variables for uncollectible expenses from one year and one month earlier are not significant (they have a t-statistic of below 1.0 as is the case of the income variable.)



	Last Year	Last Month	Price	Income	Intercept
Coefficient	-0.01	0.04	85,181.61	12.39	-1,485,292.47
Std Error	0.08	0.07	35,812.25	33.64	804,058.20
R-Squared	**36%**		#N/A	#N/A	#N/A
T-statistic	-0.08	0.54	2.38	0.37	-1.85

The R-squared statistic in the above equation can be interpreted to represent the variation in the annual level of uncollectible expense that is not explained by price and income. This equation implies that most of the variation in uncollectible expense on an annual basis is explained by factors other than price and income.

D. Autocorrelation

When CMP develops its regression equations for purposes of making its sales forecasts, the data has a similar structure to the uncollectible expense analysis. In particular, the sales analysis uses a dependent variable that has seasonality and develops an equation that is a function of both income and price. When CMP makes its sales forecast, the Company does not use lagged dependent variables but instead makes an adjustment for autocorrelation in the data. The adjustment for autocorrelation and the exclusion of lagged dependent variables is consistent with econometric theory that suggests among other things that the R-squared statistic is distorted unless an autocorrelation adjustment is made. The autocorrelation adjustment is required because if the lagged dependent variable is not included in the regression, trends in the unexplained values follow trends and are not independent from one period to the next. For statisticians, this is a very bad thing.

To create the regression with an autocorrelation adjustment, a three stage process can be used that is somewhat analogous to the two stage process for computing seasonal adjustments described above. First, the regression equation is run without the lagged independent variable. Next, unexplained residuals from the first stage are regressed against lagged values in order to determine the trends in the data (that should not be present if the regression is to produce unbiased results). Finally, the trend coefficient from the second step regression, called the autocorrelation coefficient, is used to create new variables that apply the formula:

Transformed Value = Original Value – Autocorrelation Coefficient x Lagged Value

When applying this process to the CMP data, the regression equation from the final step of the process is shown below. One notable aspect of this regression is that the R-squared statistic of 31% is similar to the regression using the seasonal adjustment approach. However, in the case with the autocorrelation adjustment shown below, the income variable has the correct sign and a t-statistic above 2.0 and the t-statistics for both the price and income variables are below the t-statistics in the CMP regression equation. The lower t-statistics imply that there is less possibility that the data can decipher a true underlying relationship. As the R-squared is 31% in the regression with autocorrelation adjustment, the analysis demonstrates that most of the variation in uncollectible expenses cannot be explained by the price and income variables. The regression equation shown below uses the lag structure assumed by CMP.

	Lagged Moratorium 7	**Lagged Price Lag 12**	**Per Capita Lagged Lag 5**	**Revenue**	**Intercept**
Coefficient	267,339.03	163,656.40	-234.70	-3.97	2,735,176.82
Std Error	84,003.43	46,793.74	95.66	7.41	1,462,144.59
R-squared	29%	298541.4217	#N/A	#N/A	#N/A
t-statistic	3.18	3.50	-2.45	-0.54	1.87

The seasonally adjusted regression discussed in the last section also exhibits some minor autocorrelation because there is a gradual movement in uncollectible expenses from month to month. When this regression equation is adjusted for autocorrelation, the resulting regression is shown in the table below. This regression that includes both the seasonality and the autocorrelation adjustments is the most appropriate equation.

	Price	Income	Intercept
Coefficient	88,615.52	7.25	-1,113,747.20
Std Error	36,748.26	33.39	697,853.77
R-Squared	27.7%	239448.3	#N/A
	13.04823	68	#N/A
t-statistic	2.41	0.22	-1.60

The results of this regression show even lower R-squared and t-statistic values than in the prior table. This supports, even more strongly, that most of the variation in uncollectible expenses are not explained by price and income variables.

E. Time Period Selected for Lagged Variables

In explaining how CMP derived the lags in variables, CMP stated: "The number of months that each variable was lagged was based on determining the best fit for the model, while taking into account the reasonableness of the assumption." (CMP response to EX-05-07.) Using a sensitivity analysis to determine the variables to include and how to lag them suggests that CMP may have been searching for supportive data and implies that the whole process may simply boil down to searching around and finding an equation that looks best – with a high R-squared and correct signs on the variables. This process implies that the regression equations do not really mean anything other than looking around in some random data.

If one applies lags to the variables (including a lagged dependent variable) and alternative autocorrelation adjustments one undoubtedly can find an equation that fits the historic data and also meets the objectives of the person developing the regression equation. However, the fact that the predicted data provides a close fit does not mean the equation is a true reflection of how uncollectible expenses are affected by price and income.

The issue of using variables that are not logical is explained by Peter Kennedy as follows: "If economic theory cannot defend the use of a variable as an explanatory variable, it should not be included in the set of potential independent variables. Such theorizing should take place before any empirical testing of the appropriateness of potential independent variables; this guards against the adoption of an independent variable just because it happened to "explain" a significant portion of the variation in the dependent variable in the particular sample at hand."

To illustrate the effect of different time lags, the tables below compute regression equations with different time lags. The sensitivity analysis shows the R-squared, the coefficient on the income variable and the t-statistic for the income variable using different price and income lags. The lag on the moratorium variable is assumed to have a lag of zero since there is no economic rationale for assuming a zero lag for this variable. The three tables below show the sensitivity analysis for the CMP equation, the autocorrelation adjustment, and the regression without the seasonal adjustment.

CMP Regression Maximum 68.34% Minimum 58.91%

		Lagged Price												
	68%	0	1	2	3	4	5	6	7	8	9	10	11	12
Lagged Income	0	59.31%	60.41%	59.11%	58.91%	59.13%	62.01%	62.43%	62.96%	63.38%	63.03%	63.90%	65.49%	68.34%
	1	59.28%	60.18%	59.11%	59.00%	59.13%	61.73%	62.19%	62.72%	63.12%	62.73%	63.60%	65.14%	68.01%
	2	59.32%	59.97%	59.21%	59.21%	59.21%	61.37%	61.85%	62.41%	62.79%	62.41%	63.18%	64.60%	67.50%
	3	59.36%	59.92%	59.27%	59.33%	59.28%	61.25%	61.77%	62.34%	62.74%	62.33%	63.07%	64.38%	67.41%
	4	59.36%	59.93%	59.26%	59.32%	59.26%	61.29%	61.82%	62.46%	62.87%	62.43%	63.14%	64.46%	67.51%
	5	59.34%	59.94%	59.22%	59.24%	59.22%	61.39%	61.91%	62.58%	63.09%	62.62%	63.34%	64.69%	67.86%
	6	59.37%	59.89%	59.27%	59.31%	59.27%	61.25%	61.78%	62.43%	62.94%	62.55%	63.28%	64.63%	67.74%
	7	59.46%	59.86%	59.39%	59.47%	59.39%	61.01%	61.54%	62.16%	62.61%	62.27%	63.06%	64.36%	67.46%
	8	59.54%	59.86%	59.49%	59.58%	59.49%	60.86%	61.36%	61.97%	62.39%	62.07%	62.83%	64.11%	67.23%
	9	59.67%	59.91%	59.66%	59.77%	59.65%	60.75%	61.18%	61.74%	62.14%	61.87%	62.59%	63.71%	66.76%
	10	59.84%	60.00%	59.87%	60.02%	59.86%	60.71%	61.07%	61.56%	61.93%	61.71%	62.39%	63.34%	66.07%
	11	60.01%	60.12%	60.08%	60.26%	60.07%	60.72%	61.04%	61.47%	61.79%	61.62%	62.27%	63.09%	65.47%
	12	60.09%	60.19%	60.16%	60.34%	60.16%	60.74%	61.05%	61.45%	61.74%	61.58%	62.21%	62.96%	65.17%

Autocorrelation Maximum 24.73% Minimum 0.23%

		Lagged Price												
	19%	0	1	2	3	4	5	6	7	8	9	10	11	12
Lagged Income	0	9.19%	12.12%	8.10%	6.95%	7.13%	10.41%	13.58%	15.63%	12.98%	9.91%	13.55%	15.18%	24.73%
	1	6.79%	9.05%	5.22%	4.21%	4.79%	7.59%	10.80%	13.04%	10.73%	7.60%	11.34%	12.20%	22.13%
	2	4.65%	5.58%	2.55%	1.88%	3.07%	4.99%	7.72%	10.04%	8.21%	5.25%	8.55%	8.97%	19.51%
	3	3.74%	3.22%	0.93%	0.60%	2.46%	3.08%	5.40%	8.03%	6.64%	3.77%	6.92%	6.69%	17.78%
	4	3.69%	3.12%	0.96%	0.58%	2.42%	3.19%	5.65%	8.44%	6.97%	3.87%	7.17%	6.74%	17.84%
	5	3.74%	3.26%	1.18%	0.73%	2.37%	3.67%	6.00%	8.80%	7.43%	4.18%	7.80%	7.32%	18.96%
	6	3.83%	3.23%	1.49%	1.05%	2.31%	4.19%	6.52%	8.54%	7.38%	4.37%	8.39%	7.96%	20.26%
	7	3.22%	1.49%	0.54%	0.44%	1.76%	3.19%	4.59%	5.85%	4.94%	2.76%	6.40%	6.05%	18.57%
	8	3.15%	0.80%	0.25%	0.23%	1.44%	2.93%	3.74%	4.22%	3.66%	1.86%	4.92%	4.55%	16.96%
	9	3.86%	0.95%	0.61%	0.43%	1.61%	3.08%	3.58%	3.47%	2.91%	1.44%	3.82%	3.28%	14.58%
	10	6.02%	3.04%	2.62%	1.97%	3.07%	4.77%	5.36%	4.89%	3.63%	2.12%	4.18%	3.91%	13.33%
	11	8.30%	5.86%	5.03%	4.00%	4.96%	7.26%	8.17%	7.60%	5.60%	3.69%	5.73%	5.93%	14.30%
	12	9.90%	8.35%	7.01%	5.78%	6.37%	9.51%	10.94%	10.50%	7.75%	5.39%	7.57%	8.10%	16.04%

Seasonal Maximum 24.73% Minimum 0.23%

		Lagged Price												
	19%	0	1	2	3	4	5	6	7	8	9	10	11	12
Lagged Income	0	9.19%	12.12%	8.10%	6.95%	7.13%	10.41%	13.58%	15.63%	12.98%	9.91%	13.55%	15.18%	24.73%
	1	6.79%	9.05%	5.22%	4.21%	4.79%	7.59%	10.80%	13.04%	10.73%	7.60%	11.34%	12.20%	22.13%
	2	4.65%	5.58%	2.55%	1.88%	3.07%	4.99%	7.72%	10.04%	8.21%	5.25%	8.55%	8.97%	19.51%
	3	3.74%	3.22%	0.93%	0.60%	2.46%	3.08%	5.40%	8.03%	6.64%	3.77%	6.92%	6.69%	17.78%
	4	3.69%	3.12%	0.96%	0.58%	2.42%	3.19%	5.65%	8.44%	6.97%	3.87%	7.17%	6.74%	17.84%
	5	3.74%	3.26%	1.18%	0.73%	2.37%	3.67%	6.00%	8.80%	7.43%	4.18%	7.80%	7.32%	18.96%
	6	3.83%	3.23%	1.49%	1.05%	2.31%	4.19%	6.52%	8.54%	7.38%	4.37%	8.39%	7.96%	20.26%
	7	3.22%	1.49%	0.54%	0.44%	1.76%	3.19%	4.59%	5.85%	4.94%	2.76%	6.40%	6.05%	18.57%
	8	3.15%	0.80%	0.25%	0.23%	1.44%	2.93%	3.74%	4.22%	3.66%	1.86%	4.92%	4.55%	16.96%
	9	3.86%	0.95%	0.61%	0.43%	1.61%	3.08%	3.58%	3.47%	2.91%	1.44%	3.82%	3.28%	14.58%
	10	6.02%	3.04%	2.62%	1.97%	3.07%	4.77%	5.36%	4.89%	3.63%	2.12%	4.18%	3.91%	13.33%
	11	8.30%	5.86%	5.03%	4.00%	4.96%	7.26%	8.17%	7.60%	5.60%	3.69%	5.73%	5.93%	14.30%
	12	9.90%	8.35%	7.01%	5.78%	6.37%	9.51%	10.94%	10.50%	7.75%	5.39%	7.57%	8.10%	16.04%

The tables demonstrate that R-squared statistic is very sensitive to the lags which are chosen. This demonstrates that making any conclusions from the regression is simply not reliable.

F. Conclusion

The general conclusion of this analysis is that CMP's regression equations do not demonstrate that most of the variation in uncollectible expense over time has been due to prices and income, factors outside the control of management. The statistical analysis prepared by CMP is problematic because:

1. The analysis does not account for the very high seasonality in the data. Most of the variation from month to month is due to seasonal factors and ignoring these factors creates a meaningless regression.

2. The CMP equation does not account for autocorrelation that is present in the data. Without adjusting for autocorrelation, the regression is biased and the R-squared statistic is significantly overstated.

Once the seasonality and the autocorrelation adjustments are made, the R-squared of the regression is only 28%. This analysis shows that a regression equation cannot be used to conclude that most of the variation in uncollectible expenses is due to factors outside of the control of management.