Miniproject#1 ~ Energy Consumption Forecasting

Paul J. Kellenberger

EC413-990

Introduction

Energy is a critical component of all industrialized countries. Consumption, production, and expenditures on energy are important drivers of economies all over the world. Energy shortages can cost countries further expansion, create stagnation, and implode any booming economy. From an economical standpoint, it would be useful to be able to identify, quantify, and predict the variables that determine energy consumption – which is the purpose of this project.

Based on data of energy demand from the United Kingdom from the years 1967 to 1997, I will create a model that can be used in forecasting the demand for energy (*Ew*) in this country using linear regression techniques associated with the AR(1) regression model. Mean Square Error and Root Mean Square Error, of the holdout period 1998-2001, will be used to determine error of the model(s) and the F-test will be used to determine the significance of variables used in the model(s).

The data set variables are:

Ew = Energy consumption by final user (thousand tons of oil equivalent)
Xw = Expenditure on energy by final user (millions of current dollars)
GDP95 = Gross domestic product (millions of 1995 dollars)
GDPcurr = Gross domestic product (millions of current dollars)
POP = UK population in millions
Tave = Annual average temperature in Great Britain (Celsius)
Tjan - Average January temperature in Great Britain (Celsius)

Most of these variables are self explanatory; however, *Tave* and *Tjan* are important because there is a relationship between the temperature and energy consumption. As it gets colder or warmer more energy is used to heat and cool homes and businesses, increasing the demand for energy (*Ew*).

In economics, it is sometimes desirable to express terms such as *Ew*, *Xw*, *GDP95*, and *GDPcurr* in per capita terms. In other words, divide the preceding variables by population (*POP*). A new data set was created with these four variables.

Ew_pc = Per capita energy consumption by final user in tons
Xw_pc = Per capita expenditure on energy by final user in current dollars
GDP95_pc = Gross domestic product per capita in 1995 dollars
GDPcurr_pc = Gross domestic product per capita in current dollars

Tjan and *Tave* are not divided by population (*POP*) because there is no relationship between these two variables.

Model Selection

During the process of model selection, I found that there were eight (8) models of considerable interest. The four models below were adjusted for per capita analysis.

$$\hat{Y} = \alpha + \beta_1 GDP95\_pc + \beta_2 RPenergy\_pc + \beta_3 Tave + \beta_4 Year + \beta_5 Tjan + \beta_6 Ew\_pc1 + \varepsilon$$
$$\hat{Y} = \alpha + \beta_1 GDP95\_pc + \beta_2 RPenergy\_pc + \beta_3 Tave + \beta_5 Tjan + \beta_6 Ew\_pc1 + \varepsilon$$
$$\hat{Y} = \alpha + \beta_1 GDP95\_pc + \beta_2 RPenergy\_pc + \beta_3 Tave + \beta_5 Tjan + \varepsilon$$
$$\hat{Y} = \alpha + \beta_1 GDP95\_pc + \beta_2 RPenergy\_pc + \beta_3 Tave + \beta_4 Year + \beta_5 Tjan + \varepsilon$$

The other four models consist of taking the Log of *GDP95*, *RPenergy*, and *Ew1*. These models are based on the four above. The log-linear model took into consideration constant prices and elasticity of demand characteristics; plus, the numbers were rather big and the log function made the numbers easier to handle. Two new variables were created in order to obtain the real price of energy, *RPenergy* and *Ew1*.

$$RPenergy = \frac{NPenergy}{GDPdef}$$

where,

$$NPenergy = \frac{Xw}{Ew}$$

&

$$GDPdef = \frac{GDPcurr}{GDP95}$$

$$Ew1 = Ew_{t-1}$$

These relationships proved invaluable. The real price of energy, *RPenergy*, is equal to the nominal price of energy, *NPenergy*, divided by the GDP deflator, *GDPdef*. The GDP deflator is obtained by dividing GDP in current dollars by GDP in 1995 dollars. For the per capita model, the same process is done except the variables listed above were in per capita terms. In the log-linear model, these variables are put into log form. The previous year's (Ew_{t-1}) energy consumption is the variable *Ew1*; therefore, *Ew1* is a lag variable. This creation of the variable is important because, often times in forecasting, last year's consumption is often a determinate of next year's consumption. In other words, it is a boundary point; energy usage usually does not decline in industrialized countries as years pass. The variable, *Year*, is a trend variable; this variable is usually important because energy consumption typically has a positive upward trend (assuming all other variables are held constant). *Year* is unaffected by log and per capita adjustments.

I performed the F-test on the eight (8) models. There were two unrestricted models, which are listed below:

$$Log(\hat{Y}) = \alpha + \beta_1 LGDP95 + \beta_2 LRPenergy + \beta_3 Tave + \beta_4 Year + \beta_5 Tjan + \beta_6 Lenergy1$$

$$\hat{Y} = \alpha + \beta_1 GDP95\_pc + \beta_2 RPenergy\_pc + \beta_3 Tave + \beta_4 Year + \beta_5 Tjan + \beta_6 Ew\_pc1$$

The log-linear restricted models were based on the top unrestricted model. The per capita restricted models were based upon the bottom unrestricted model. Restricted model 1 had the trend variable, *Year*, excluded. Restricted model 2 had the lag variable, *Ew1*, excluded. Restricted model 3 had both the trend variable, *Year*, and the lag variable, *Ew1*, excluded. The restrictions were the null hypothesis, with the alternative hypothesis not being the null hypothesis. Below are the results of the F-test and their associated critical values.

	F-test		**Critical Values**	
	Log-linear Model	**Per Capita Model**	**Log-linear Model**	**Per Capita Model**
Restricted Model 1	**2.376**	**1.62**	**4.26**	**4.26**
Restricted Model 2	**8.977**	**9.71**	**4.26**	**4.26**
Restricted Model 3	**5.828**	**5.21**	**3.4**	**3.4**

If the F-test for the respective model was greater than the critical value, than the null hypothesis was rejected and the unrestricted model was chosen. If the F-test was less than the critical value, the restricted model was chosen. After the F-test, only two models were left:

$$\hat{Y}_A = \alpha + \beta_1 GDP95\_pc + \beta_2 RPenergy\_pc + \beta_3 Tave + \beta_5 Tjan + \beta_6 Ew\_pc1$$
$$Log(\hat{Y}_B) = \alpha + \beta_1 LGDP95 + \beta_2 LRPenergy + \beta_3 Tave + \beta_5 Tjan + \beta_6 Lenergy1$$

These models were then used to determine the holdout period (1998-2001) energy consumption (*Ew*). The MSE and root MSE of the two models were compared; whatever model had the lowest root MSE was the better model. Below are the results:

	Root MSE
Model A	1813.15
Model B	1804.20

Therefore, the final model is Model B. To double check the F-test, I also calculated the holdout period MSE on the two unrestricted models as well. My F-test calculations were correct, because the MSE on the two unrestricted models were higher than the two models above.

The next step was to see how accurate Model B was to the simple naïve model. The simple naïve model assumed that this year's energy consumption (Ew_t) would be the same as last year's consumption (Ew_{t-1}). In other words:

$$Ew_t = Ew_{t-1}$$ Or the lag variable, *Ew1*, from the previous models.

$$\hat{Y} = Ew_{t-1} + \varepsilon$$

The root MSE for the simple naïve model was 155318.83. Compared to Model B (or any other model for that matter) this was not a very good model. Therefore, all of the models tried above were an improvement over the simple naïve model.

The final model with coefficients was:

$$Log(\hat{Y}) = 5.67 + 0.1445 * LGDP95 - 0.197 * LRPenergy - 0.017 * Tave - 0.003 * Tjan + 0.36 * Lenergy1$$

Interpretation of the Final Model

In an economy with zero GDP (and an implicit zero energy expenditure) and an average temperature in January and throughout the year of zero degrees Celsius, energy consumption would be *e* raised to the 5.67 power, or 290.03, thousand tons of oil equivalent. For every 1 unit increase in *LGDP95*, log(*Ew*) increased by 14.45%. For every 1 unit increase in *LRPEnergy*, log(*Ew*) decreased by 19.7%. For every 1 unit increase in *Tave*, log(*Ew*) decreased by 1.7%. For every 1 unit increase in *Yjan*, log(*Ew*) decreased by 0.3%. For every 1 unit increase in *Lenergy1*, log(*Ew*) increased by 36%. In other words, there is a negative relationship between energy consumption (*Ew*) average yearly temperature (Celsius), average January temperature (Celsius), and the price for energy. Furthermore, as income rose, consumers consumed more energy. It was intuitive that the model showed that last year's energy consumption would shape this year's consumption in some form or another.

Conclusion

It is interesting to note that the error between the two restricted models were so close. The two models were most definitely an improvement over the unrestricted models used in the F-test. However, the error in Model B was still too high at 1804.2 thousand tons of energy consumption during the holdout period. Nonetheless, it was better than the simple naïve MSE of 155318.83. The results were logical and straightforward to interpret; however, more sophisticated econometric techniques should be used in the future to obtain a more accurate model.

Excel Spreadsheet of AR(1) and Naïve Model:

AR(1) Model:

	Actual *Ew*	Forecast Y	Error	Error^2
1998	155784	155725.8	58.19032832	3386.114
1999	157130	154100.4	3029.556894	9178215
2000	158684	156735.9	1948.060532	3794940
2001	160789	160998.7	-209.7155449	43980.61
			MSE	**3255130**
			SQRT_MSE	**1804.2**

Naïve Model:

	Actual *Ew*	Forecast Y	Error	Error^2
1998	155784	153902	-152918.6525	2.34E+10
1999	157130	155784	-154689.2605	2.39E+10
2000	158684	157130	-156026.6689	2.43E+10
2001	160789	158684	-157602.5236	2.48E+10
			MSE	**2.4E+10**
			SQRT_MSE	**155319**

The SAS Printout of the Final Model, Model B:

The REG Procedure
Model: MODEL1
Dependent Variable: Lenergy

Number of Observations Read	31
Number of Observations Used	30
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.03910	0.00782	24.90	<.0001
Error	24	0.00754	0.00031406		
Corrected Total	29	0.04664			

Root MSE	0.01772	R-Square	0.8384
Dependent Mean	11.89291	Adj R-Sq	0.8047
Coeff Var	0.14901		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.66960	1.12065	5.06	<.0001
LGDP95		1	0.14482	0.02591	5.59	<.0001
LRPenergy		1	-0.19746	0.03093	-6.38	<.0001
Tave	Tave	1	-0.01736	0.00888	-1.96	0.0622
Tjan	Tjan	1	-0.00276	0.00254	-1.08	0.2894
Lenergy1		1	0.36040	0.10938	3.29	0.0031