Modelling the Electric Power Consumption in Germany

Cerasela Măgură
Agricultural Food and Resource Economics (Master students)
Rheinische Friedrich-Wilhelms-Universität Bonn
cerasela.magura@gmail.com

Codruța Elena Duminică
Agricultural Food and Resource Economics (Master students)
Rheinische Friedrich-Wilhelms-Universität Bonn
codruta_e@yahoo.com

Abstract
The main goal of our project is to see how the electric power consumption (kWh) is influenced and fluctuates regarding the total population and the surface area (m^2), in the territory of Germany using tools provided from the econometrics theory, such as regression analysis and time series analysis. The dependent variable is the electric power consumption, while the independent variables are the population and surface area.
The reason behind our choice is the fact that we noticed that once the world gets more in trend with the technology, more electric power is being used; so, we decided to see how the electric power consumption has fluctuated between the years of 1977 and 2010 in Germany, regarding the total population fluctuation and the surface area. The last independent variable is important to indicate and later, maybe in other paper, to compare with other countries, because the concentration of the population depends also on the surface area.

Keywords: electric power consumption, regression analysis.

JEL classifications: C12, C32.

Introduction

Literature review

The liberalization of power markets that started spreading widely during the last decade of the 20th century has significantly changed the energy transaction landscape. Nowadays, electricity is traded in the electricity exchange and in over-the-counter markets. A large part of the traded volume represents energy to be consumed or produced in the future as a forward or future product.

Numerous studies have looked toward energy efficiency and building standards in as an example of the potential for energy savings throughout the world. As we were going through articles to find some research background regarding our theme, we encountered two term with which we have to make the distinction: energy consumption – which refers to the oil power, specially petrol and gas; and electric power consumption – which refers to the electricity consumption.

Previous research on Germany reveals that various factors have to be taken in consideration:

Weather conditions (temperature, humidity, sun light, wind speed), season, economic trends, day of the week, public holidays and vacations, hour of day.

Other models were considering the electric power consumption per capita, related with the GDP per capita and population growth.

So in order to make our model we considered the following variables: electric power consumption, population total and surface area.

Hypothesis

Null hypothesis ("H_0"): the outcome that the researcher does not expect (almost always includes an equality sign).

Alternative hypothesis ("H_A"): the outcome the researcher does expect.

For the β_1 :
H_0: $\beta_1 \leq 0$ => As the time passes and technological progress appears, the electric power consumption remains constant or decreases with respect to the total population.

H_A: $\beta_1 > 0$ => As the time passes and technological progress appears, the electric power consumption increases with respect to the total population.

For the β_2 :
H_0: $\beta_2 \leq 0$ => As the time passes and technological progress appears, the electric power consumption remains constant or decreases with respect to the surface area.

H_A: $\beta_2 > 0$ => As the time passes and technological progress appears, the electric power consumption increases with respect to the surface area.

Methodology

Overall model

The goal of the multiple linear regression model is to estimate the effect on the dependent variable Y of each of the independent variables X_1;……; X_n, or regressors, while holding the others constant. The equation representing the multiple linear regression model is:

$$Y= \beta_0 + \beta_1 * X_1 + \beta_2 * X_{2 + \ldots +} e$$

OLS is the dominant method used in practice for estimating the parameters of linear regression. Some of the theoretical properties of the OLS estimators are highly desirable for the statistical analysis. The most important of which is the fact that they are consistent, efficient and unbiased estimators. In addition, the OLS is successful in estimating the model parameters if the mean of the residuals, E(u), is equal to zero. Furthermore, the variables X1;…; Xn; Y need to be independently and identically distributed random variables. Finally, large outliers should be unlikely, because the coefficients might be sensitive to such disturbances, and there should not be perfect multicollinearity between the regressors since it would lead to a division by zero in the parameter estimation process.

Semi-log Multiple Regression

In this project we used semi-log multiple regression in order to analyse the fluctuation in the electric power consumption in Germany, with respect to the total population and total surface area. Because this consumption depends on various factors and we needed to use a logarithm to make the high values of the energy power consumption more reasonable and fit for our calculus, the use of a left-handed semi-log multiple regression, was compelling.

To be able to run the regression, we used E-Views software and for the preparation of the data, we used Microsoft Office Excel. The general model used, as mentioned above, is a semi-log left-handed functional form:

$$\log Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_n * X_n + e$$

The particular model used to run the estimation, in accordance with our data choice is:

$$\log\_energy = \beta_0 + \beta_1 * population\_total + \beta_2 * surface\_area + e$$

It is called left-handed semi-log functional form because the logarithm is on the left side of the equation and because only one part of it contains it.

We also decided upon this form because we expect that an increase in X_1 (population_total), will cause Y (electric power consumption) to increase at an increasing rate.

Time series analysis

A time series is an ordered sequence of observations, which in our case are 34. The ordering is usually made through time as implied by its name, but also spatial ordering is possible. Examples of time series can be found in economics, engineering, natural and social sciences. Time series can be either continuous or discrete, however, discrete values can be managed more easily in terms of computational complexity. The main objective of time series analysis is the better understanding and description of a mechanism, the forecast of future values and the improvement in the control a system.

These are the main reasons why we chose to run a time series analysis because it offers us a vision of the electric consumption in time and not just at a certain moment. Of course, that knowing the fluctuations in the past and now, having the greater picture, noticing also the trend, we could make a forecast regarding this consumption.

Data

In order to create a concrete model, reliable data are necessary. Since electric power consumption is a complex process that depends on several different factors, electric power consumption alone is not enough to perform an accurate forecast model. Therefore, in order to analyse the effects of these factors and better understand the nature of electric power consumption, electric power consumption data were used along with the total population, surface area and calendar data.

The most important information necessary for the formulation of such a model is the electric power consumption. Historic data of electric power consumption for Germany were available from The World Bank. The electric power consumption can be better understood by studying its moving average for different time windows.

Results

The raw data required obtaining the estimates of the regression coefficients, their standard errors, etc; from these calculations the following data was obtained (Table 1):

Table 1: E-views regression output

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LOG_ENERGY
Method: Least Squares
Date: 01/17/13 Time: 08:51
Sample: 1977 2010
Included observations: 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-241.5294	115.0592	-2.099175	0.0440
POPULATION_TOTAL	2.22E-08	8.67E-09	2.560180	0.0156
SURFACE_AREA	0.000747	0.000324	2.306950	0.0279

R-squared	0.657285	Mean dependent var	26.97793
Adjusted R-squared	0.635175	S.D. dependent var	0.095398
S.E. of regression	0.057621	Akaike info criterion	-2.785755
Sum squared resid	0.102926	Schwarz criterion	-2.651076
Log likelihood	50.35784	Hannan-Quinn criter.	-2.739826
F-statistic	29.72714	Durbin-Watson stat	0.231995
Prob(F-statistic)	0.000000		

$\hat{\beta}_0$ = -241.5294 se ($\hat{\beta}_1$)=115.0592

$\hat{\beta}_1$ = 2.22E-08 se ($\hat{\beta}_1$)=8.67E-09

$\hat{\beta}_2$ = 0.000747 se ($\hat{\beta}_2$)=0.000324

R^2 = 0.0657285 and the adjusted R^2 = 0.635175

The estimated regression line, therefore is:

$\hat{Y}$= -241.5294+2.22E-08*population_total+0.000747*surface_area

Overall model fit

One of the most important methods to see how well the regression fits the data is to check the value of the R^2, or coefficient of determination. The coefficient of determination represents how much of the variance in Y is explained by all Xi's together. The R^2 ranges between 0 and 1. If the coefficient of determination has a value as close to 1 as possible, that means that the fit, between our data and our estimations, is a very close and good one and that our variables do explain the dependent variable Y.

In our case, the R^2 has a value of 0.657285, which indicates that 65% fluctuation in Y is explained by our two variables jointly, the rest of the percentage could be explained by other variables or residuals, which are the variables not included in this model. Therefore, because the R^2 is bigger that 60% we can affirm that our model is nicely fitted.

Statistical analysis

The statistical analysis starts by checking if the coefficients are significant, therefore we look at the probabilities column, which in our case we can see that all the values there are smaller than 0.05, which means that we have a very small chance that our coefficients are zero, so they are significant in our model.

Econometric analysis

The econometric analysis begins with an evaluation of the coefficients, which in our case, the first one (population_total) has a value of 2.22E-08, which has a positive sign, as we expected, because the total population has a positive influence on the electric power consumption. So, we could say that, one unit increase in total population, leads to an increase in the electric power consumption by 0.00000222% a year in the case of ceteris paribus.

Regarding the second coefficient, the surface area, with a value of 0.000747, as we suspected, it is also positive and greater that zero. Therefore, we could say that for an increase of surface area with one unit, the electric power consumption will increase by 0.0747% a year, in the situation of ceteris paribus.

Hypothesis

Regarding the acceptance or rejection of the hypothesis, in order to reject H_0 we need to check is the p-value < level of significance (0.05) and has the sign of H_A. Therefore, in our case, we can reject the null hypothesis because for all variables respect the fact that p-value < level of significance:

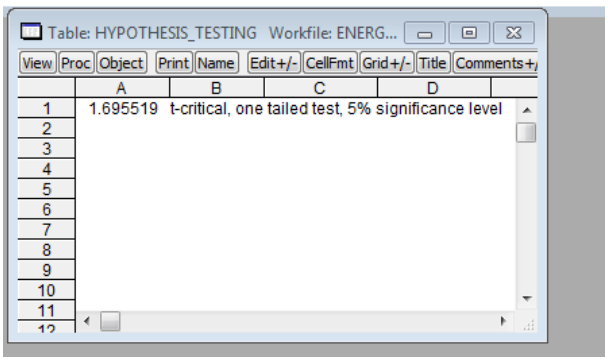
β_1 p-value: 0.0156 < 0.05
β_2 p-value: 0.0279 < 0.05

T-test

With regard to the t-test, we have considered the various costs involved in making Type I and Type II Errors and have chosen 5% as the level of significance. We have 34 observations, therefore, the calculated degree of freedom is 34-2=32. At a 5% level of significance, the critical t-value, t_c, was calculated in E-Views and is illustrated in Table 2, to be 1.695.

Table 2: T-critical E-view's output

```
table hypothesis_testing
hypothesis_testing (1,1)=@qtdist(.95,(regression.@regobs-regression.@ncoef))
hypothesis_testing (1,2)="t-critical, one tailed test, 5% significance level"
```



As we know from theory, the decision rule for the t-test is to reject H_0 if $|t_k| > t_c$ and if t_k also has the sign implied by H_A. In our case, this amounts to the following two conditions:

For β_1: Reject H_0 if $|2.56| > 1.695$ and if 2.56 is positive.
For β_2: Reject H_0 if $|2.30| > 1.695$ and if 2.30 is positive.

Multicollinearity

Perfect multicollinearity violates Classical Assumption VI, which specifies that no explanatory variable is a perfect linear function of any other explanatory variables.

When testing for multicollinearity, we discovered that our regression has imperfect multicollinearity, because our explanatory variables are imperfectly linearly related. Therefore, first, we tested for multicollinearity with the help of the simple correlation coefficient which in our case is 0.80 as it can be observed in Table 3.

Table 3: Simple correlation coefficient E-view's output

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Correlation

	POPULATIO...	SURFACE_...
POPULATIO...	1.000000	0.800720
SURFACE_...	0.800720	1.000000

Secondly, we tested with the variance inflation factors by doing two steps: first we ran a OLS regression that had our X_i (surface_area) as a function of all the other explanatory variables, which in our case was just one, the population_total, so the equation was like it follows:

$$Surface\_area = \beta_0` + \beta_1` * population\_total + e$$

After doing this first step and observing the regression output which is present in Table 4, we calculated the VIF (variance inflation factor) by this formula:

$$VIF(\hat{\beta}_i) = \frac{1}{(1-R_i^2)}$$

Table 4: Multicollinearity regression E-view's output

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SURFACE_AREA
Method: Least Squares
Date: 01/21/13 Time: 17:42
Sample: 1977 2010
Included observations: 34

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	355290.0	228.0799	1557.743	0.0000
POPULATION_TOTAL	2.14E-05	2.84E-06	7.561385	0.0000

R-squared	0.641153	Mean dependent var	357014.1
Adjusted R-squared	0.629939	S.D. dependent var	51.70532
S.E. of regression	31.45370	Akaike info criterion	9.791933
Sum squared resid	31658.73	Schwarz criterion	9.881719
Log likelihood	-164.4629	Hannan-Quinn criter.	9.822552
F-statistic	57.17455	Durbin-Watson stat	0.113425
Prob(F-statistic)	0.000000		

Which in our case is:

$$VIF(\hat{\beta}_i) = \frac{1}{(1-0.641153)} = 2.78668$$

Analysing this result we must know that the higher the value of VIF the more severe are the effects of the multicollinearity. Because there is no table of critical VIF values, a common rule of thumb is that if a given VIF is greater than 5 the multicollinearity is severe. But, in our case, our VIF value is 2.78, so we can conclude that we do have multicollinearity but not at a severe level.

Heteroskedasticity

Pure heteroskedasticity occurs when Classical Assumption V, which assumes constant variance of the error term, is violated, of course, when talking about a correctly specified equation.

To test for heteroskedasticity we made a White test in E-views and the output can be observed in Table 5.

Table 5: White test- Heteroskedasticity E-view's output

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: White

F-statistic	4.545860	Prob. F(3,30)	0.0097
Obs*R-squared	10.62565	Prob. Chi-Square(3)	0.0139
Scaled explained SS	9.104424	Prob. Chi-Square(3)	0.0279

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 01/21/13 Time: 20:48
Sample: 1977 2010
Included observations: 34
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.285184	2.810711	-0.101463	0.9199
POPULATION_TOTAL	-5.48E-09	9.98E-08	-0.054952	0.9565
POPULATION_TOTAL^2	-6.17E-17	4.40E-16	-0.140268	0.8894
POPULATION_TOTAL*SURFACE_AREA	3.93E-14	2.83E-13	0.138867	0.8905

R-squared	0.312519	Mean dependent var	0.003027
Adjusted R-squared	0.243771	S.D. dependent var	0.004412
S.E. of regression	0.003837	Akaike info criterion	-8.178372
Sum squared resid	0.000442	Schwarz criterion	-7.998801
Log likelihood	143.0323	Hannan-Quinn criter.	-8.117133
F-statistic	4.545860	Durbin-Watson stat	0.653438
Prob(F-statistic)	0.009656		

Regarding the interpretation of the White test, we know that the variable denoted with Obs*R^2 is the White test statistic. It is computed as the number of observations times R^2 from the test regression. This test statistic has a chi-square distribution with degrees of freedom equal to the number of slope parameters.

Therefore, to make a decision about the fact of having or not heteroskedasticity we need to know that if Obs*R^2 is larger than the critical chi-square value found in Statistical Table B-8 found in Appendix, then we can reject the null hypothesis that there is no heteroskedasticity; and if Obs*R^2 is less than the critical chi-square value, then we cannot reject the null hypothesis of no heteroskedasticity.

In our case, the value of Obs*R^2 is 10.62 which is smaller than the critical chi-square value which is 43.77. Therefore, we cannot reject the null hypothesis of no heteroskedasticity.

Conclusions

Theoretically as we can see from our empirical model, the trend in the electric power consumption is increasing every year, but this variable cannot only depend on the number of population and surface area, we have to take into consideration other dependent variable as GDP per capita, energy consumption (oil, petrol, bio-fuel, gas), and others variables.

Regarding the limitations of our model we can state that because of more omitted variables and not being able to drop some of our variables, our model has multicollinearity and cannot be reliable for good future forecasts.

In conclusion our empirical model tries to estimate an overall consumption of electric power trend over a time series.

References

Boug, P. (2000), "Modelling energy demand in Germany – A Cointegration Approach", *Statistic Research Norway Department*

Sotiropoulos, E. (2012), Master Thesis "Modelling of German Electricity Load for pricing forward contracts", *EEH – Power Systems Laboratory, Swiss Federal Institute of Technology (ETH) Zurich*

Kandel, A. Sheridan, M. MCauliffe, P. (2008), "A comparison of per capita electricity consumption in the United States and California", *California Energy Commission Presented at: 2008 ACEEE Summer Study on Energy Efficiency in Buildings Asilomar Conference Center Pacific Grove, California August 17-22, 2008*

Jeffrey M. Wooldridge, (2011), "Introductory Econometrics – A modern approach", *International Student Edition*

The World Bank, http://data.worldbank.org/indicator

http://www.statisticsmentor.com/tables/table_t.htm