Effects of multicollinearity on electricity consumption forecasting using partial least squares regression

Gulder Kemalbay a,*, Ozlem Berak Korkmazoglu a

Yildiz Technical University, Davutpasa Campus, Faculty of Art&Science, Department of Statistics, 34220, Esenler, Istanbul, Turkey

Abstract

The electricity forecasting is important for accurate investment planning of energy production/generation; however energy is also essential input for economical and industrial development. In this study, we aim to discuss the impact of economic growth on annual electricity consumption. It is true that economic variables usually have correlations with each with variable degrees. Partial least squares regression (PLSR) is one of the effective ways of dealing with the multicollinearity problem especially arises in several econometric models. In situations when there are a large number of highly correlated explanatory variables, decomposition by PLSR can be used to select a small number of linear combinations of the original variables which explain the great amount of covariance between explanatory and response variables. Thus, the aim of this study is to perform the partial least squares regression method to forecast annual electricity consumption using historical data for Turkey.

Keywords: Partial least squares, multicollinearity, economic growth, electricity consumption;

1. Introduction

Electricity consumption has increased over the years and so the topic was usually examined by researchers. Several electricity consumption studies have been published on aggregate macro data at the country, sub-national or state level. In recent years, some of these studies have been devoted to investigating the relationship between electricity consumption and economic growth and from these empirical findings it can be said that the direction of the relationship between electricity consumption and economic growth could be differ by countries and their government policies. Several of electricity forecasting models have been developed using economic, social, geographic and demographic factors. This paper proposes a model for electricity consumption in Turkey and by following the literature, economic variables are used as explanatory variables which are highly correlated. Partial least squares regression (PLSR) is one of the effective ways of dealing with the multicollinearity problem especially arises in several econometric models. Thus, the aim of this study is to perform the partial least squares regression method to forecast annual electricity consumption using historical data for Turkey.

* Corresponding Author name. Tel.: +90-0212-383-4429
E-mail address: kemalbay@yildiz.edu.tr

© 2012 Published by Elsevier Ltd. Selection and/or peer review under responsibility of Prof. Dr. Hüseyin Arasli

Keywords: Partial least squares, multicollinearity, economic growth, electricity consumption;
2. Partial Least Squares Regression

PLSR easily handle multicollinearity problem in situations when there are a large number of highly correlated explanatory variables and singularity of explanatory matrix $X$ when the number of predictors is large compared to the number of observations. Thus it is sometimes referred as soft modeling because of ordinary least squares regression makes hard assumptions including lack of multicollinearity among explanatory variables. There are several approaches to cope with this problem like eliminating some predictors using stepwise regression or principal component regression (PCR). By contrast, PLSR method is initialized with decomposition of both $X$ and $Y$, thus these components include information about the correlation between $X$ and $Y$ whereas PCR only concentrates on the variance of $X$. As in the PCR, it is followed by a regression step where the components are used to predict $Y$.

2.1. A Decomposition Method of both Predictors and Response Variable

For simplicity, we are interested in predicting the single-response model in matrix algebra term given as $Y = X\beta + \varepsilon$ where $Y$ is nx1 vector consisting of n observation on one response variable, $X$ is nxp matrix consisting of n observation on p explanatory variables $\beta$ is px1 vector of unknown regression coefficients and $\varepsilon$ is nx1 vector of errors whose rows are independently and identically distributed. The aim of PLSR is to obtain information about both of $X$ and $Y$ by constructing new explanatory variables, often called factors, latent variables, or components, where each component is a linear combination of $X_1, X_2, \ldots, X_p$. The matrix of explanatory variables $X$ is decomposed as $X=TP^T$ with $TT^T=I$, where $I$ is identity matrix, $T$ is called score matrix and $P$ is called loading matrix. The columns of $T$ are latent vectors which could be any set of orthogonal vectors spanning the column space of $X$. However, in order to specify $T$ we need to find two sets $w$ and $c$ which represent the weights of $X$ and $Y$, respectively, to obtain a linear combination of the columns of $X$ and $Y$ which make their covariance maximum. Then, Y is estimated as $\hat{Y} = TBC^T$ where $B$ is diagonal matrix of regression weights and $C$ is weight matrix of $Y$.

2.2. A PLSR Algorithm

Partial least squares regression is defined as the projection onto an orthogonal space of the column or row-centered and column or row-normalized data. Without centering, both the mean variable value and the variation around that mean are involved in selecting factors. The PLS regression is based on nonlinear iterative partial least squares (NIPALS) algorithm with main steps as follows:

**Step1.** Transform matrices $X$ and $Y$ into $E$ and $F$, respectively, which are row-centered and row-normalized data to have mean 0 and standart deviation 1 as follows: $E = [X^*_y]_{n \times p}$, $F = [Y^*_i]_{n \times 1}$, $X^*_y = X_y - \overline{X}_y / S_{X_y}$, $Y^*_i = Y_i - \overline{Y}_i / S_{Y_i}$, $i = 1, \ldots, n$; $j = 1, \ldots, p$.

**Step2.** Before starting the iteration, initialize the vector $u=F$.

**Step3.** Compute weights for $X$ as $w = E^TF / \|E^TF\|$, $\|w\| = 1$ and factor scores for $X$ as $t = Ew / \|Ew\|$, $\|t\| = 1$.

**Step4.** Compute weights for $Y$ as $c = F^Tt / \|F^Tt\|$, $\|c\| = 1$ and scores for $Y$ as $u = Fc$.

**Step5.** If $t$ has not converged then go to Step 1; otherwise compute $b = t^Tu$, where $b$ is diagonal element of regression weight matrix $B$ and compute factor loadings for $X$ as $p = E^Tb$.

**Step6.** Substract the effect of $t$ from $E$ and $F$ as $E = E - tp^T$ and $F = F - btc^T$.

**Step7.** The sum of squares of $X$ ($Y$) explained by a latent vector which is denoted by $p^Tp$ ($b^T$) and dividing the explained sum of squares by the corresponding total sum of squares, the proportion of variance explained is obtained.

**Step8.** If $E$ is null matrix, then all latent vectors is found, otherwise re-iterate the process from initial Step 1.

Briefly, the main goal is to find vectors $t = Ew$ and $u = Fc$ with constraints $w^Tw = t^Tt = 1$ such that $t^Tu$ be maximal.
3. Empirical Results: Electricity Forecasting based on PLSR

To examine the ability of economic variables to explain the electricity forecasting, we used gross national product of primary industry ($X_1$), gross national product of secondary industry ($X_2$), gross national product of tertiary industry ($X_3$), installed capacity ($X_4$), electricity production ($X_5$), and government expenditure ($X_6$). The historical data has been selected from 1981 to 2010 for Turkey and has been made stationary by taking the logarithm. The presence of multicollinearity between the explanatory variables in the model is shown in Table 1.

| Variables | $X_1$  | $X_2$  | $X_3$  | $X_4$  | $X_5$  | $X_6$ |
|-----------|--------|--------|--------|--------|--------|--------|
| $X_1$     | 1.000  | 0.999  | 0.997  | 0.975  | 0.994  | 0.989  |
| $X_2$     | 0.999  | 1.000  | 0.999  | 0.974  | 0.992  | 0.984  |
| $X_3$     | 0.997  | 0.999  | 1.000  | 0.976  | 0.991  | 0.981  |
| $X_4$     | 0.975  | 0.974  | 0.976  | 1.000  | 0.998  | 0.980  |
| $X_5$     | 0.994  | 0.972  | 0.991  | 0.998  | 1.000  | 0.992  |
| $X_6$     | 0.989  | 0.984  | 0.981  | 0.980  | 0.992  | 1.000  |

From Table 1, it is obvious that the explanatory variables are highly correlated. Under such a scenario the OLS estimates are still BLUE; however, they may be unstable which means that it will be harder to produce significant coefficients. The problem with evaluating such a model is that it may demonstrate adequate prediction capability on the training data. The quality of prediction for a random model does not always increase with the number of components used in the model. Thus, it is critical to determine the optimal number of components to keep for building a model. A straightforward approach is cross validation method which is generally used for determining the optimal number of components to take into account. The most used technique is leave-one-out (LOO) cross validation is given following table which shows how much predictor and response variation is explained by each component:

| Validation: RMSEP | Intercept | 1 Components | 2 Components |
|-------------------|-----------|--------------|--------------|
| Cross Validation  | 0.6476    | 0.07846      | 0.04295      |
| % Variance Explained (X) | 99.83    | 99.93        |
| % Variance Explained (Y) | 98.64    | 99.65        |

From Table 2, it is seen that %99.93 of the variance of $X$ and % 99.65 of variance of $Y$ are explained by two components. One may select a number of components after which the cross validation error, generally used root mean squared error of prediction (RMSEP), does not show a significant decrease. It is often simpler to judge the RMSEP s by plotting them against the number of components as follows.
We decide optimal number of components is two since the cross validation error RMSEP (0.04295) does not show a significant decrease after two components. After selecting the number of components, the quality of cross-validated predictions can be performed by plotting them versus measured values as given in Figure 2.

![Figure 2. Cross-validated Predictions versus Measured Values](image)

There is not any irregularity in the predictors, such as distinct grouping patterns or outliers as shown in Figure 2. In PLSR, the predictor and response variables are considered as a block of variables. Then, PLSR extracts the score vectors which serve as a new predictor representation and regresses the response variables on these new predictors. The scores for two components obtained by the NIPALS algorithm are given in Table 3 as follows

| Obs. | Scores for X (T) | Scores for Y (U) | Obs. | Scores for X (T) | Scores for Y (U) |
|------|-----------------|-----------------|------|-----------------|-----------------|
|      | t_1             | t_2             | u_1  | u_2             | t_1             | t_2             | u_1  | u_2             |
| 1    | -5.307          | 0.015           | -5.525| -0.063          | 16              | 0.693           | -0.078| 0.540           | -0.044|
| 2    | -4.997          | 0.000           | -5.184| -0.054          | 17              | 1.165           | -0.079| 1.035           | -0.037|
| 3    | -4.820          | 0.066           | -5.001| -0.052          | 18              | 1.558           | -0.074| 1.378           | -0.052|
| 4    | -4.416          | 0.042           | -4.392| 0.007           | 19              | 1.914           | -0.080| 1.573           | -0.098|
| 5    | -3.978          | -0.065          | -4.031| -0.015          | 20              | 2.281           | -0.079| 1.948           | -0.096|
| 6    | -3.662          | -0.007          | -3.627| 0.010           | 21              | 2.650           | -0.145| 1.885           | -0.220|
| 7    | -3.271          | 0.059           | -2.975| 0.085           | 22              | 2.960           | -0.162| 2.179           | -0.225|
| 8    | -2.835          | 0.036           | -2.579| 0.073           | 23              | 3.132           | -0.108| 2.589           | -0.156|
| 9    | -2.412          | 0.093           | -2.169| 0.070           | 24              | 3.279           | -0.046| 2.992           | -0.083|
| 10   | -2.064          | 0.046           | -1.758| 0.088           | 25              | 3.411           | 0.022| 3.355           | -0.016|
| 11   | -1.684          | 0.017           | -1.502| 0.052           | 26              | 3.534           | 0.060| 3.823           | 0.083 |
| 12   | -1.255          | -0.031          | -1.047| 0.060           | 27              | 3.614           | 0.145| 4.228           | 0.176 |
| 13   | -0.866          | -0.060          | -0.583| 0.082           | 28              | 3.760           | 0.155| 4.442           | 0.196 |
| 14   | -0.266          | -0.056          | -0.403| -0.039          | 29              | 3.789           | 0.169| 4.284           | 0.142 |
| 15   | 0.199           | -0.051          | 0.062 | -0.039          | 30              | 3.890           | 0.197| 4.462           | 0.164 |

Score matrix for X, i.e T is composed of linear combination of weight matrix W and X as expressed in equation T=XW. The columns of T is called latent vectors. Scores for Y is used in modelling response variable as given Y=UC.

| Obs. | Weights for T | Loadings for X |
|------|---------------|----------------|
|      | w_1           | p_1            | p_2            |
| X_1  | 0.548         | 0.546          | 0.698          |
| X_2  | 0.559         | -0.123         | 0.56           |
| X_3  | 0.607         | -0.500         | 0.608          | -0.645         |
| X_4  | -             | 0.483          | -              | 0.279          |
| X_5  | -             | 0.397          | -              | 0.254          |
| X_6  | -             | 0.286          | -              | 0.256          |
Weights for $T$ show that which predictors are most represented in each factor. Those predictors with small weights are less important than those with large weights in absolute value. The $X$-weights $W$ represent the correlation between the $X$-variables and the $Y$-scores $U$. The $X$-loadings and $X$-weights are usually very similar to each other. The $X$-loadings give the combination of predictors that comprise each PLSR component.

To determine which factors to eliminate from the analysis, we look at the Variable Importance for the Projection (VIP) which is defined as to analysis the explanatory power of selected factors to the response. If a predictor has a relatively small coefficient (in absolute value) and a small value of VIP (usually less than 0.8 to be small), then it is a prime candidate for deletion. The following Table lists the VIP for all predictor variables:

| X   | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|-----|-------|-------|-------|-------|-------|-------|
| Comp.1-VIP | 1.002139 | 0.999512 | 0.997213 | 0.993456 | 1.006222 | 1.001411 |
| Comp.2-VIP | 1.001411 | 0.999348 | 0.998214 | 0.992757 | 1.006201 | 1.002020 |

The bigger VIP, the more importance of $X$. Then, we can say that electricity production ($X_5$) has most explanatory power; however gross national product of primary industry ($X_1$) and government expenditure ($X_6$) almost have the same power with ($X_5$) while ($X_2$), ($X_3$) and ($X_4$) have also similar power of factors to annual electricity forecasting.

4. Conclusions

A forecasting model is proposed for electricity consumption based on PLSR method. By analyzing a real data, it is confirmed that this method is effective forecasting in situation multicollinearity. The results show that the electricity production has most affected the electricity consumption.

References

Abdi, H. (2003). PLS-Regression; Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks Sage.

Ahnay, T. (1996). A simulation study on comparison of prediction models when only a few components are relevant. Computational Statistics and Data Analysis, 21:87–107, 1996.

Altunay, G., & Karagöl, E. (2005). Electricity Consumption and Economic Growth: Evidence from Turkey. Energy Economics, 27, 849-856.

Barker, M., & Rayens, W.S. (2003). Partial least squares for discrimination. Journal of Chemometrics, 17:166–173.

Chen, S. T., H. Kuo. & Chen, C.C. (2007). The Relationship between GDP and Electricity Consumption in 10 Asian Countries. Energy Policy, 35, 2611-2621.

Egelioglu, F., Mohamad, A.A., & Guven, H. (2001). Economic variables and electricity consumption in Northern Cyprus. Energy, 26:355–62.

Frank, I.E., & Friedman, J.H. (1993). A Statistical View of Some Chemometrics Regression Tools. Technometrics, 35:109–147.

Geladi, P., & Kowalski B. (1986). Partial least square regression: A tutorial. Analytica Chimica Acta, 35:1–17.

Goutis, C. (1996). Partial least squares yields shrinkage estimators. The Annals of Statistics, 24:816–824.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning, Springer.

Harris, J.L., & Liu, L. (1993). Dynamic structural analysis and forecasting of residential electricity consumption. International Journal of Forecast, 9:437–55.

Holland I.S. (1990). PLS regression and statistical models. Scandinavian Journal of Statistics, 17, 97–114.

Höskuldson, A. (1988). PLS regression methods. Journal of Chemometrics, 2, 211-228.

Hulland, J.S. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. Strategic Management Journal, 20:195–204.

Kraft, J., & A. Kraft. (1978). On the Relationship Between Energy and GNP. Journal of Energy Development, 3, 401-403.

Meng, M., & Shang, W. (2008). Research on Annual Electric Power Consumption Forecasting Based on Partial Least-squares Regression. International Seminar on Business and Information Management.

Meng, M., & Niu, D. (2011). Annual Electricity Consumption Analysis and Forecasting of China Based on Few Observations Methods. Energy Conversion and Management, 52, 953-957.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiah, editor, Multivariate Analysis, 391-420. Academic Press, NewYork.