Unintended biases of an electricity demand forecast based on a double-log regression

Woo, Chi Keung; Zarnikau, J.; Cao, Kang Hua

Published in:
Electricity Journal

DOI:
10.1016/j.tej.2020.106866

Published: 01/12/2020

Document Version:
Peer reviewed version

Unintended biases of an electricity demand forecast based on a double-log regression

C.K. Woo[a], J. Zarnikau[b,*], K.H. Cao[c]

[a] Department of Asian and Policy Studies, Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong

[b] Department of Economics, University of Texas at Austin, 2225 Speedway, Austin, TX 78712, USA

[c] Department of Economics, Hong Kong Baptist University, 34 Renfrew Road, Kowloon Tong, Hong Kong

E-mail addresses: chiwoo@eduhk.hk (C.K. Woo), jayz@utexas.edu (J. Zarnikau), kanghuacao@hkbu.edu.hk (K.H. Cao)

[*] Corresponding author at: Department of Economics, University of Texas at Austin, 2225 Speedway, Austin, TX 78712, USA.

Keywords: Double-log regression, electricity demand forecast, forecast bias, forecast precision

Abstract

This paper identifies the unintended biases occasionally not recognized when using a double-log regression to make an electricity demand forecast. It shows that ignoring the stochastic nature of forecasts of income, price and weather can vastly overstate the forecast's precision, potentially causing inadequate resource procurement for reliable service at least cost. Fortunately, the overstated precision is readily avoidable because its correction uses information available when making the forecast.

1. Introduction

An accurate and precise electricity demand forecast is critical for policy modeling (Manne et al., 1979) and resource planning (Chao, 1983; Hobbs, 1995; Wilkerson et al., 2014). The goal of this paper is to illustrate the forecast based on a double-log regression can appear to have high precision, potentially causing erroneous decisions on resource procurement.

The double-log specification is popular for analysing electricity demand (Espey and Espey, 2004; Labandeira et al., 2017).[1] This is because it has slope coefficients that are income and price elasticities used by governments and electric utilities for quantifying the effects of price and income changes on electricity consumption (Orans, 2008; EIA, 2017).[2] Its results are less influenced by the presence of outliers than a linear specification's. Finally, it pre-empts the odd forecast outcome of negative MWh numbers.

The example in the next section identifies the unintended biases occasionally not recognized when using a double-log regression for MWh forecasting, which requires forecasts of income, price and weather that are inevitably stochastic. Such biases, however, are readily avoidable because their correction uses information available when making a MWh forecast. Hence, resource procurement based on a double-log MWh forecast should account for the biases identified by this paper.

2. Example

To illustrate the extent of unintended forecast biases in connection to a double-log MWh regression, we use an example from the second author's course of Markets for Electricity at UT Austin. For mathematical details of these biases, see Appendix 1.

[1] Woo (1994) shows how to use the likelihood ratio (LLR) test to determine whether a linear or double-log specification should be used for a demand analysis. Treating the Box-Cox specification as the null hypothesis, we presume that the LLR test has not rejected the double-log specification at the 5% significance level.
[2] For the case of residential customers, the double-log specification is theoretically sound, consistent with consumer utility-maximizing behaviour under various settings (Hausman, 1987; Woo, 1994; Woo et al., 2012).

This example uses monthly electricity sales for Jan-2008 to Oct-2015 in the U.S. state of Michigan. Sales are modeled as monthly MWh usage per customer in a double-log regression, whose regressors are the log of the average price of electricity (real $ per MWh), the log of the state's per-capita income, and two weather variables: cooling degree days (CDD) and heating degree days that can have zero values.[3] Data were obtained from websites maintained by the U.S. Department of Energy's Energy Information Agency (EIA) and the U.S. Bureau of Labor Statistics.

Table 1 reports the ordinary least-squares (OLS) regression results.[4] The adjusted R^2 of 0.81 suggests a reasonably good fit and all coefficient estimates have correct signs.

Using the regression results in Table 1, the forecast assumptions in Table 2 and the formulas in Appendix 1, we perform the following MWh calculations for the forecast month of September 2016:[5]

- The understated forecast level denoted by MWh_1 that ignores a ln(MWh) forecast's standard error (SE).
- The understated forecast level denoted by MWh_2 that is based on the biased SE of a ln(MWh) forecast.
- The unbiased forecast level denoted by MWh_3 that is based on the unbiased SE of a ln(MWh) forecast.

Appendix 1 mathematically shows $MWh_1 < MWh_2 < MWh_3$, supporting our first claim that a biased MWh forecast like MWh_1 or MWh_2 is less than the unbiased MWh_3.

[3] We define CDD = monthly sum of max(daily maximum temperature – 65°F, 0) and HDD = monthly sum of max(65°F - daily minimum temperature, 0). These variables capture the effect of weather on air conditioning and space heating.

[4] The double-log regression could have been estimated using generalized least squares (GLS) to account for the presence of autocorrelation or heteroskedasticity. Appendix 1 shows that the biases continue to exist because the calculations of the MWh forecast's level and SE does not qualitatively depend on the choice of estimation method.

[5] September 2016 is chosen because it is a likely month of scheduled maintenance of peaking generation units after their heavy summer utilization.

We also calculate MWh_2's SE denoted by d_2 and MWh_3's SE denoted by d_3. Appendix 1 mathematically shows $d_2 < d_3$, supporting our second claim that a wrongly prepared MWh forecast can appear to be more precise than the unbiased MWh forecast.

To provide the empirics of biases, Table 3 reports the results of the above calculations, yielding the following findings. First, MWh_1, MWh_2 and MWh_3 are virtually identical. Second, MWh_2's SE is about half of MWh_3's SE, falsely conveying MWh_2's precision. The implication of this erroneous precision is potential inadequacy in least-cost resource procurement for providing reliable service.

3. Conclusion

In the context of a double-log MWh regression, ignoring the stochastic nature of forecasts for income, price and weather can understate a MWh forecast's standard error by a large amount. Correcting the understatement is straightforward, using information available when making a MWh forecast. Hence, when using a double-log regression for MWh forecasting, least-cost resource procurement should avoid the biases identified by this paper.

Acknowledgements

C.K. Woo's research is funded by research grants (#4388 and #4400) from the Education University of Hong Kong. J. Zarnikau's contribution to this paper is part his ongoing research on electricity economics at UT Austin.

Appendix 1. Theory and calculations

Suppose the double-log MWh regression based on N monthly observations is:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad (1)$$

where $\boldsymbol{Y}$ = (N × 1) column vector of the natural log of MWh sales per customer; and $\boldsymbol{X}$ = (N × 5) matrix that contains a constant equal to one, natural-log values of per capita real income ($ per month) and real electricity price ($ per MWh), and two weather variables: cooling degree days (CDD) and heating degree days (HDD). The (5 × 1) column vector $\boldsymbol{\beta}$ contains β_1 = intercept, β_2 = income elasticity > 0, β_3 = price elasticity < 0, and β_4 = CDD's sales effect > 0, and β_5 = HDD's sales effect > 0.[6] The (N × 1) column vector $\boldsymbol{\varepsilon}$ contains normally and independently distributed (NID) errors, each of which has zero mean and variance σ^2.[7]

Suppose $\boldsymbol{b}$ is the OLS estimate of $\boldsymbol{\beta}$, whose estimated variance is $\boldsymbol{W} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} s^2$ where s^2 = mean squared error of the ln(MWh) regression. Further suppose $\boldsymbol{x}_f$ for forecast month f is a row vector that contains a constant equal to one and month f's forecasts for the remaining regressors. Let $\boldsymbol{U}_f$ denote $\boldsymbol{x}_f$'s estimated variance matrix. The first row and first column of $\boldsymbol{U}_f$ contain zeroes. The remaining elements in $\boldsymbol{U}_f$ are $\boldsymbol{x}_f$'s variance and covariance estimates.

An unbiased MWh forecast based on equation (1) is:

$$z_f = \exp(\boldsymbol{x}_f \boldsymbol{b} + 0.5\, v_f^2); \quad (2)$$

where v_f = standard error (SE) given by equation (3) below (Johnson and Kotz, 1970). Hence, ignoring v_f tends to understates z_f.

We now state the unintended bias in a ln(MWh) forecast's SE. Feldstein (1971) shows that v_f^2 is the sum of three strictly positive terms:

[6] Including additional regressors (e.g., customer demographics) complicates the appendix's algebraic exposition without the benefit of improved insights.

[7] Replacing NID errors with autoregressive or heteroskedastic errors leads to GLS estimation that does not qualitatively alter our subsequent discussion of a MWh forecast's level and standard error. Further, we avoid overstating the OLS coefficient estimates' precision by using robust SE that are autocorrelation-heteroskedasticity-consistent (Greene, 2012).

$$v_f^2 \quad = \quad A_f + B_f + C_f. \tag{3}$$

The first term is $A_f = (\boldsymbol{x}_f (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \boldsymbol{x}_f^{\mathrm{T}} + 1)\, s^2 > 0$, the forecast variance under the assumption that $\boldsymbol{x}_f$ is non-stochastic (Greene, 2002). For clarity, we use $v_{Af} = A_f^{1/2}$ to denote the understated forecast SE noted in Section 2. Because v_{Af} excludes B_f and C_f, it is less than v_f.

The second term is $B_f = \boldsymbol{b}^{\mathrm{T}} \boldsymbol{U}_f \boldsymbol{b} > 0$ is $\boldsymbol{U}_f$'s effect magnified by $\boldsymbol{b}$ on v_f^2. To compute B_f, we first express $B_f = \Sigma_j \Sigma_k\, b_j\, b_k\, r_{jk}\, u_{fj}\, u_{fk}$ for j (or k) = 2 for ln(income), 3 for ln(price), 4 for CDD, and 5 for HDD; where b_j = estimate of β_j; r_{jk} = correlation between forecasts j and k; and u_{fj} = SE of forecast j. We then find r_{jk} based on the historic correlations among the double-log regression's non-intercept regressors. As a result, B_f's calculation uses the information available when making a MWh forecast.

The third term is C_f = trace of $\boldsymbol{W} \otimes \boldsymbol{U}_f = \Sigma_j\, w_j^2\, u_{fjj}^2 > 0$, where w_j^2 = variance of b_j. Based on the information available when making a MWh forecast, C_f is the combined effect of $\boldsymbol{W}$ and $\boldsymbol{U}_f$ on v_f^2.

We now state the three possible calculations of the MWh forecast level:

- $\mathrm{MWh}_{1f} = \exp(\boldsymbol{x}_f \boldsymbol{b})$ is the understated level that ignores a ln(MWh) forecast's SE.
- $\mathrm{MWh}_{2f} = \exp(\boldsymbol{x}_f \boldsymbol{b} + 0.5\, v_{Af}^2)$ is the understated level that is based on v_{Af} = understated SE of a ln(MWh) forecast.
- $\mathrm{MWh}_{3f} = \exp(\boldsymbol{x}_f \boldsymbol{b} + 0.5\, v_f^2)$ is the unbiased level that is based on v_f = unbiased SE of a ln(MWh) forecast.

A comparison of the three MWh levels indicates $\mathrm{MWh}_{1f} < \mathrm{MWh}_{2f} < \mathrm{MWh}_{3f}$, thus supporting our first claim that a biased forecast like MWh_{1f} or MWh_{2f} has a lower level than the unbiased forecast MWh_{3f}.

Finally, we make two calculations of the MWh forecast's SE (Johnson and Kotz, 1970):

$$d_{2f} \quad = \quad \mathrm{MWh}_2\, [\exp(v_{Af}^2) - 1]^{1/2}; \tag{4.1}$$

$$d_{3f} \quad = \quad \mathrm{MWh}_3\, [\exp(v_f^2) - 1]^{1/2}. \tag{4.2}$$

Equations (4.1) and (4.2) indicate $d_{2f} < d_{3f}$ because $MWh_2 < MWh_3$ and $v_{4f} < v_f$. This finding supports our second claim that understating a ln(MWh) forecast's SE can cause a biased MWh forecast to appear more precise than the unbiased MWh forecast.

References

Chao, H.P., 1983. Peak load pricing and capacity planning with demand and supply uncertainty. Bell Journal of Economics14(1), 179-90.

EIA, 2017. Assumptions to the Annual Energy Outlook 2017. Washington D.C.: U.S. Energy Information Administration. Available at https://www.eia.gov/outlooks/aeo/assumptions/pdf/0554(2017).pdf (accessed on 20 April 2020).

Espey, J.A., Espey, M., 2004. Turning on the lights: a meta-analysis of residential electricity demand elasticities. Journal of Agricultural and Applied Economics 36, 65-81.

Feldstein, M.S.,1971. The error of forecast in econometric models when the forecast-period exogenous variables are stochastic. Econometrica 39, 55–60.

Greene, W.H., 2012. Econometric Analysis. Pearson Education.

Hausman, J.A., 1987. Exact consumer surplus and deadweight loss. American Economic Review 74(1), 662-676.

Hobbs, B.F., 1995. Optimization methods for electric utility resource planning. European Journal of Operational Research 83, 1-20.

Johnson, N.L., Kotz, S., 1970. Distribution in Statistics: Continuous Univariate Distributions. Houghton Miffin.

Labandeira, X., Labeaga, J.M., López-Otero, X., 2017. A meta-analysis on the price elasticity of energy demand. Energy Policy 102, 549-68.

Manne, A.S., Richels, R.G., Weyant, J.P., 1979. Feature article – energy policy modeling: a survey. Operations Research 27(1), 1-36.

Orans R., 2008. Direct Testimony, 2008 Long Term Acquisition Plan, Appendix E. Vancouver, B.C. Hydro. Available at https://www.bcuc.com/Documents/Proceedings/2008/BCH_LTAP_B-1-1_APPENDICES/Appendix%20E.pdf (accessed on 20 April 2020).

Wilkerson, J., Larsen, P., Barbose, G., 2014. Survey of Western U.S. electric utility resource plans. Energy Policy 66, 90-103.

Woo, C.K., 1994. Managing water supply shortage: interruption vs. pricing. Journal of Public Economics 54, 145-160.

Woo, C.K., Zarnikau, J., Kollman, E., 2012. Exact welfare measurement for double-log demand with partial adjustment. Empirical Economics 42,171–180.

Table 1. Double-log regression results for Michigan; sample period: Jan-2008 to Oct-2015 that contains 94 monthly observations for ln(MWh sales per customer)

Variable: coefficient	Estimate	Standard error	*p*-value
Adjusted R^2	0.8148		
Mean squared error: s^2	0.0011		
Intercept: b_1	0.2316	1.2563	0.8540
ln(real income per capita): b_2	0.0219	0.0797	0.7840
ln(real electricity price): b_3	-0.0033	0.0799	0.9670
Cooling degree days (CDD): b_4	0.0010	0.0001	0.0000
Heating degree days (HDD): b_5	0.0001	0.0000	0.0000

Note: The *p*-values for the coefficient estimates are based on robust standard errors that are autocorrelation-heteroscedasticity-consistent (Greene, 2012).

Table 2. Forecasts of the non-intercept regressors in Table 1 for September 2016

Variable	Value	Standard error
ln(real income per capita)	8.77	0.02
ln(real electricity price)	-9.92	0.04
Cooling degree days (CDD)	105.33	69.78
Heating degree days (HDD)	162.33	284.52

Note: The forecast of for the four variables are obtained through PROC FROECAST of SAS (2004) that uses time series modeling to automatically produce the numbers required by the calculations listed in Section 2.

Table 3: Results based on the example described by Tables 1 and 2

MWh forecast's level					MWh forecast's standard error *d*		
MWh_1	MWh_2	MWh_3	$E_1 = MWh_1 / MWh_3$	$E_2 = MWh_2 / MWh_3$	d_2	d_3	$E_d = d_2 / d_3$
1.7940	1.7956	1.7983	0.9976	0.9985	0.0756	0.1244	0.6077

Notes: (1) E_1 and E_2 measure the extent of understatement of biased MWh forecast levels. Appendix 1 mathematically demonstrates $MWh_1 < MWh_2 < MWh_3$, where MWh_1 = MWh forecast that ignores the ln(MWh) forecast's SE; MWh_2 = MWh forecast based on the ln(MWh) forecast's understated SE; and MWh_3 = MWh forecast based on the ln(MWh) forecast's unbiased SE.
(2) E_d measures the extent of understatement of a MWh forecast's SE. Appendix 1 mathematically demonstrate $d_2 < d_3$.