# A comparative assessment of aggregate car ownership model estimation methodologies

Sambracos, Evangelos and Paravantis, John

University of Piraeus

19 May 2006

# A COMPARATIVE ASSESSMENT OF AGGREGATE CAR OWNERSHIP MODEL ESTIMATION METHODOLOGIES

**J.A. Paravantis and E. Sambracos**

University of Piraeus

## Abstract

This work examines the implications of advances in time series analysis on car ownership modeling in Greece. Variables include adults' population ratio, GDP per capita, car occupancy, bus kilometers, inflation and unemployment. We developed and compared (a) a classical regression model estimated on raw levels, (b) an econometric model estimated on data stationarized using graphical and unit root tests and (c) an "atheoretical" ARIMA model. Although significant methodological implications were noted, all models forecast 48 to 49 private cars per 100 inhabitants by the year 2010, a development of momentous energy and environmental implications.

Keywords: car ownership; aggregate models; regression; time series analysis; forecasting.

## 1. INTRODUCTION

As Kennedy aptly points out (2001), in the recent past practitioners more or less ignored the effects of nonstationarity in variables involved in time series regression; this has certainly been the case with most car ownership studies (including many referenced in this work). Yet, advances in unit root research indicate that nonstationarity must be taken into account, not by merely resorting to an alternate estimation method (such as generalized least squares) but by rethinking model specification. To avoid nonsensical results from spurious regression, trend stationary variables must be detrended and difference stationary (i.e. integrated) variables must be differenced before entering a regression equation either as response or predictors. In this work we investigate the impact of these important developments in time series analysis on aggregate car ownership modeling by considering Greece as a case study and developing a "naïve", an "econometric" and an "atheoretical" car ownership model with the objective of providing forecasts to the year 2010.

## 2. LITERATURE REVIEW

Scholl, Schipper and Kiang (1996) discuss how the demand for passenger transport is affected by lifestyles, income, fuel prices, labor structure, travel time and cost as well as urban development and point out that population growth magnifies the impact of these effects. Developing an aggregate model of car ownership in Asian countries, Prevedouros and An (1998) identified population, income and unemployment rate as important car ownership factors and warned that income, car prices and fuel prices usually present multicollinearity problems. Lam and Tam (2002) chose population, population density, annual gross domestic product, first registration tax, annual license fee, gasoline price, annual passenger trips on public transport and annual railway passenger kilometers in order to estimate an aggregate car ownership model for Hong Kong.

Cars and population are strongly related although population is also likely to be correlated to economic measures such as Gross Domestic Product (GDP). Therefore, to avoid collinearity problems, the car ownership ratio (e.g. number of cars per 100 persons) is often utilized as a better choice of a dependent variable. A related population metric of interest, the adults ratio, has been shown by Gately (1990) to be important in explaining the rapid growth in the number of drivers from the mid 1960s to the mid 1970s. Goodbody Economic Consultants (2000) suggest that there is a close correlation between car ownership and economic growth while the London Borough of Croydon supports a correlation between gross annual household income and household car ownership (1995). Higher income is regarded as the driving force behind the increase in car ownership (Baldwin Hess and Ong, 2001) that leads to a decline in the relative importance of bus and local/intercity train travel (Scholl, Schipper and Kiang, 1996). The relationship between income and car ownership is likely to be nonlinear because car ownership grows with income level but the impact of income declines as a certain saturation level is approached: the logistic, the quadratic and the Gompertz function have been employed to model this effect (Button, Fowkes and Pearman, 1980; Dargay and Gately, 1997; Dargay and Gately, 1999). On the cost side, Dargay and Gately (1999) examined fixed (e.g. insurance, road tax, vehicle licensing fees and garaging fees) and variable (e.g. fuel costs, maintenance and repairs, oil, parking fees, tolls) vehicle costs. Contrary to Hong Kong where first registration tax and first license fee are high in an effort to control the number of private cars (Lam and Tam, 2002), in Greece they are quite low and do not function as disincentives in owning a car; car prices on the other hand, include heavy taxation and may influence the decision to own a car but they are collinear with population, GDP and other independent variables. Paravantis and Prevedouros (2001) found both gasoline price and

inflation to bear a significant effect in their autoregressive railway passenger models, possibly reflecting the fact that cheaper gasoline prices make traveling by car more affordable while inflation tends to impact motoring costs such as maintenance, insurance and toll fees. To this end, unemployment may also be a variable of interest. Finally, vehicle occupancy (i.e. the vehicle loading factor) should be associated with car ownership. Car pooling is oftentimes encouraged (although not in Greece) with measures such as entry into fast moving lanes; car pooling may also indicate a societal change to a more environmentally friendly position possibly associated with more usage of mass transport. We decided to investigate the effect of this important parameter including recent original estimates of vehicle occupancy values for the case of Greece (Danos, 2004).

Based on our literature review and considering data availability, we propose the following car ownership formulation (with expected signs preceding variable names):

$$\text{CARS100} = f(+\text{PCTADULT}, +\text{GDPPC}, -\text{INFL}, -\text{UNEMPL}, -\text{CAROCC}, -\text{BUSKM}) \quad (1)$$

where PCTADULT represents the percent of population above 17 years of age, GDPPC is per capita GDP, INFL stands for inflation, UNEMPL represents unemployment, CAROCC is car occupancy and BUSKM equals annual vehicle kilometres of a single mass transit bus.

## 3. METHODOLOGY

We develop three alternative car ownership models, taking Greece as a case study and using car ownership and socioeconomic data that span the period 1970 to 2003 with the objective of predicting car ownership to the year 2010:

1. At first, we develop a "naïve" car ownership model by carrying out Ordinary Least Squares (OLS) regression while largely disregarding the issue of nonstationarity. We forgo formal unit root testing, a choice often faced by researchers in the case of small sample sizes (such as annual data) that make rejecting the hypothesis of a unit root fairly difficult. We try to avoid seemingly good fits that are the result of spurious regression not by stationarizing variables beforehand but by being watchful, for instance, for formulations that render a very high coefficient of determination ($R^2$) combined with a very low Durbin-Watson statistic (an almost sure sign of spurious regression). Among these models that produce residuals with correlograms that appear to originate from white noise, we select the best formulation by utilizing appropriate fit measures as well as other more advanced (albeit equally easy to use) criteria (i.e. AIC and BIC).
2. Next, we develop an "econometric" car ownership model, taking into consideration any nonstationarity in both dependent and independent variables. Using graphical means and formal testing, we examine variables in order to determine whether they are stationary, trend stationary or difference stationary (i.e. integrated) processes and we transform nonstationary variables by either detrending (in the case of trend stationary) or differencing (in the case of integrated processes). We select our best formulation utilizing appropriate fit and information criteria as before.
3. Finally, we develop an "atheoretical" car ownership model using the ARIMA (autoregressive integrated moving average) technique that is oftentimes quoted as performing at least equally well (if not better) than traditional regression techniques (Makridakis, Wheelwright and Hyndman, 1998).

The Gretl (version 1.5.0) econometric package (Baiocchi and Distaso, 2003) and Statgraphics (version 5.1) were used for graphing and statistical analysis.

## 4. RESULTS

### *4.1. "Naïve" regression approach*

Initially, we examine Pearson correlation coefficients (R) among the dependent and independent variables. CARS100, the dependent variable, was found to be strongly correlated to PCTADULT and GDPPC ($R \geq 0.9$). UNEMPL showed an extremely strong positive (possibly spurious) association with CARS100 (R=0.89), contrary to prior expectations. The correlation coefficients of CAROCC (–0.881) and BUKM (–0.828) also showed strong negative association with CARS100, as expected. Finally, INFL was negatively associated with CARS100. Although multicollinearity does not affect the ability of a model to predict (Makridakis, Wheelwright and Hyndman, 1998), we also looked at correlation coefficients among independent variables (some of which may signify spurious association among time series) in order to avoid including highly correlated variables in the same model. Luckily, multicollinearity may be conveniently assessed by Variance Inflation Factors (VIFs) that are easy to compute and interpret (Studenmund, 1992).

Although pure autocorrelation may be expected of monthly or quarterly data, it is less likely in annual data (Studenmund, 1992) where any autocorrelation present is more likely impure i.e. a sign of errors in specification rather than a violation of technical assumptions (Hendry and Mizon, 1978). We employed the Durbin-Watson statistic (D) and the Lagrange Multiplier (LM) test to test autocorrelation. Noticing that models tended to systematically underpredict actual CARS100 values at the end of the time series, we run a exploratory piecewise linear regression that gave an optimum inflection point between years 1995 and 1996 (corresponding to an apparent slope change of GDPPC), a finding consistent with a state-sponsored non-catalytic vehicle retirement program (that was in effect for most of the 1990s) in tandem with the start of a boom of the Greek Stock Exchange that lasted until the start of the 2000s and turned many stock holders into millionaires overnight. In order to capture the effect of this shock event, we decided to introduce into our model a slope dummy that must include both a dummy intercept (D95) and a dummy slope term (GDPPC95), both of which were set to zero prior to 1996 (Studenmund, 1992). The resulting model (M1) was superior to all previous models: variable coefficients had the expected sign and were statistically significant and the standard error of the regression line was smaller. Although the Durbin-Watson test was inconclusive, an LM test did not reject the hypothesis of no autocorrelation up to order one. On the issue of multicollinearity, we now got extremely high VIF values but this was clearly an artifact due to the introduction of D95 and GDPPC95 (VIFs do not work well in the case of binary data). Leverage plots confirmed that collinearity was not a problem and since all regression coefficients were statistically significant, we decided that the model is valid and declared it to be our best "naïve" model formulation:

$$\begin{gathered}\text{CARS100} = -88.167 + 1.559\ \text{PCTADULT} + 0.119\ \text{GDPPC} - 23.782\ \text{D85} \\ + 0.371\ \text{GDPPC95} - 0.281\ \text{BUSKM} - 0.524\ \text{CAROCC} \\ \textit{t=24.037 (p=0.000); t=5.667 (p=0.000); t=–10.908 (p=0.000);} \\ \textit{t=11.615 (p=0.000); t=–5.912 (p=0.000); t=–1.572 (p=0.128)} \end{gathered} \quad (2)$$

T-statistic and p values (in italics) refer to variable coefficients (the significance of the constant term is of no concern). We note that INFL and UNEMPL are not included in M1 but we retained CAROCC although it is marginally not significant due to its theoretical appeal. Model M1 provides an almost perfect fit to historical data (Figure 1) as indicated by an impressively high $R^2$ value (0.999, which should be interpreted cautiously in the case of time

series data).

*4.2. "Econometric" approach*

In order to develop a better ("econometric") model that takes into account latest findings in time series analysis, we have to examine the stationarity (or lack thereof) of dependent and independent variables. In general, a proper investigation into the stationarity of a time series includes (a) graphical analysis, (b) a look at correlograms and (c) unit root testing. To transform a nonstationary times series into stationary, we first deflate it (if applicable) and/or attempt a logarithm (or other appropriate) transformation. If no transformation renders the series stationary, we proceed with formal unit root testing by the augmented Dickey-Fuller test (DF) that tests the null hypothesis of a unit root and helps us separate trend stationary from difference stationary processes. Dickey-Fuller tests have low power, i.e. tend to accept the unit root hypothesis more frequently than warranted and thus find a unit root even when none exists (Gujarati, 2003). Power is higher with larger data sets although it is also related to the time span of the data (e.g. a 30-year data set may be better than a data set covering 100 weeks). To identify the optimal order of differencing (almost always equal to 1, rarely above 2), we use two empirical rules: (a) keep differencing until the autocorrelation of the first lag becomes zero or negative and (b) select the differencing that renders the smallest standard deviation of the transformed series (Nau, 2006).

A DF test carried out on CARS100 does not reject the null hypothesis of a unit root (i.e. nonstationarity) when the test regression is estimated with only a constant (t=2.3515; p=1), neither with a constant and a linear trend (t=1.0198; p=0.9999) nor with a constant and a quadratic trend (t=–0.8076; p=0.9926). We conclude that, since CARS100 is nor stationary neither trend stationary, it must be difference stationary and using our two empirical rules, we decide that CARS100 is a second-order integrated process. In a similar fashion, we conclude that PCTADULT may be modeled as stationary around a quadratic trend but, following the empirical rules, we decide to model it as an integrated process of order three (an unusually high degree of differencing but this approach renders the smallest standard deviation); GDPPC and BUSKM are first order integrated processes; CAROCC could be modeled as (linear) trend stationary but is best represented as a second order integrated process (lower standard deviation); INFL could be considered stationary, most likely around a quadratic trend but is best modeled as a first order integrated process; finally, UNEMPL could be regarded as a (linear) trend stationary process but we prefer second differences (lower standard deviation).

Having stationarized all variables, we now proceed with model estimation based on D2CARS100 (second differences of CARS100), D3PCTADULT (second differences of PCTADULT), D1GDPPC (first differences of GDPPC), D1BUSKM (first differences of BUSKM), D2CAROCC (second differences of CAROCC), D1INFL (first differences of INFL) and D2UNEMPL (second differences of UNEMPL). As expected, correlations among the dependent and independent variables are much weaker since spurious associations have been excluded, e.g. D2CARS100 is now only weakly associated (and with the expected sign) to both D3PCTADULT (0.229) and D1GDPPC (0.109) that in the previous model were major influences. It becomes obvious that regression models ran on stationary rather than level data are much less likely to suffer from multicollinearity. This time, our best model (M2) resulted from a backwards stepwise procedure, i.e. we included all independent variables and then omitted those with the wrong sign and/or insignificant, testing the hypothesis that the coefficients of omitted variables were zero:

$$\text{D2CARS100} = 0.0488 - 1.803\ \text{D2CAROCC} - 0.0837\ \text{D2UNEMPL}$$
$$t=-3.242\ (p=0.003);\ t=-1.047\ (p=0.305) \qquad (3)$$

Compared to M1, not only is the formulation different but the coefficient of determination is also much smaller ($R^2$=0.338).

### *4.3. "Atheoretical" approach*

In this final modeling attempt, we estimate an ARMA model on the second differences of CARS100 that were previously found to be stationary, i.e. we estimate an ARIMA(i,2,j) model. Because this approach relies only on autoregressive and moving average terms of CARS100, it is regarded as an "atheoretical" approach that estimates forecasts in a mechanistic fashion. In estimating an "optimal" ARMA model, we keep in mind that mild underdifferencing may be compensated by adding more autoregressive (AR) terms in the final model while mild overdifferencing may be compensated with more moving average (MA) terms in the final model (Nau, 2006). Guided by plots of the autocorrelation (ACF) and partial autocorrelation (PACF) function, we chose the model that rendered residuals that looked like white noise. It turned out that the second differences of CARS100 constitute mild overdifferencing and a MA(1) term was added which, although not significant, was kept because it resulted in narrower prediction intervals. Taking into account that an ARMA model estimated on second differences should not have a constant term, our best ARMA model is:

$$\text{D2CARS100} = 0.197\ \text{MA(1)}$$
$$t=1.094\ (p=0.283) \qquad (4)$$

which is an ARIMA(0,2,1) model (and is equivalent to linear exponential smoothing).

Incidentally, all ARMA models up to AR(2) and MA(2) render almost identical forecasts (e.g. 48.93 to 49.03 cars per 100 people for 2010) although in mixed models (i.e. models that include both AR and MA terms) AR and MA terms may be canceling each other out, so simpler configurations with fewer terms are usually preferred. These similar forecasts imply that attributing overly attention to proper ARMA model selection, although a good practice (Box, Jenkins and Reinsel, 1994) may not carry significant practical implications; in our opinion, it is more important to obtain the correct degree of differencing (i.e. order of integration) prior to estimating alternative ARMA models (i.e. "I" is the most important part of ARIMA).

### *4.4. Comparing model forecasts*

In order to forecast the dependent variable to 2010 will all three models, we must first develop predictions for the independent variables. In the case of PCTADULT, we used official UN projections (medium variant population growth scenario); for GDPPC, we employed the OECD scenario that predicts a 3.2% growth to the year 2005 and a 3.6% growth thereafter. In the case of variables for which there exist no reliable third party forecasts, we used ARIMA and selected the most parsimonious model specification for which residuals appear to be pure noise (Box, Jenkins and Reinsel, 1994). In the case of BUSKM we excluded bus privatization years (that were quite atypical compared to the rest of the series) and forecast a further decline in vehicle kilometers of a single bus; finally, we predicted a further decline in CAROCC, a result we also feel comfortable with.

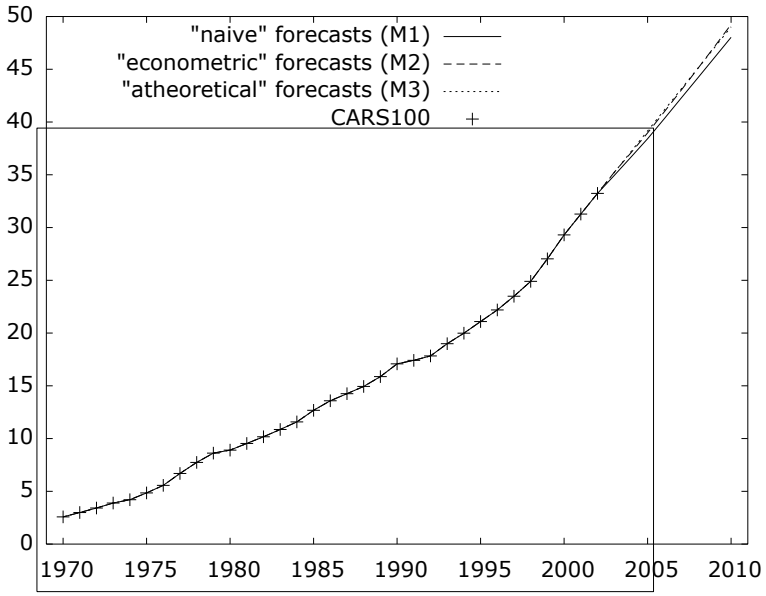


Figure 1. Forecasts of alternative modeling approaches.

CARS100 forecasts by the "naïve" (M1), the "econometric" (M2) and the "atheoretical" method (M3) are shown in Figure 1. Considering forecasts confidence intervals (not shown), the predictions are not significantly different and all three demonstrate a further considerable increase in passenger car ownership between 2005 and 2010. Based on these very similar results from the three alternate approaches, we feel confident in predicting that ownership levels in Greece will, more or less, reach one car per two inhabitants by 2010, a development of momentous energy and environmental implications.

Although holding out data for validation is a very important practice in time series analysis, it is difficult to implement with small data sets: holding out a good portion of the sample may decrease the size of the estimation period to such an extend that it would be too small to be a reliable indicator of future performance. Although no details are shown, holding out the period 1996 to 2002 had a dramatic impact on forecasting: alternative models invariably failed to recognize the slope increase in CARS100 (noted between 1995 and 1996) and significantly underpredicted 1996 to 2002 historical data. This demonstrates both the potential impact of significant intervention events on time series as well as the limitations of statistical forecasting.

## 5. CONCLUSIONS

Our analyses brought out important differences in methodology but rather insignificant differences in results, certainly nowhere as striking as one would expect from the advice of workers who strongly admonish against using OLS with nonstationary data. Considering our findings in view of related literature, Dargay and Gately (1997) estimated a simple aggregate car ownership nonlinear model that related car ownership to per capita income on a sample of 26 countries with data spanning the period from 1973 to 1992: in the case of Greece they assumed an annual car usage of 15000 km and forecast a car ownership ratio of a mere 33 cars per 100 people for the year 2015, a number that, in fact, was exceeded in 2002. In a more detailed work (Dargay and Gately, 1999), the authors employed a similar model of car ownership with data from 26 countries over the time period 1960 to 1992 and projected, in the case of Greece, a value of 35 cars per 100 people for the year 2015. The results of our modeling work show that their projections may in fact underestimate significantly the

increase in car ownership expected by 2010, underlining the ever increasing impact of private automobiles on overall increases of both fuel consumption and $CO_2$ emissions from road passenger transport in medium and low income countries.

## REFERENCES


Baiocchi, G. and W. Distaso (2003). "GRETL: Econometric Software for the GNU Generation", *Journal of Applied Econometrics*, 18, pp. 105–10.

Baldwin Hess, D. and P.M. Ong (2001). "Traditional Neighborhoods and Auto Ownership", Working Paper #37, *Working Paper Series*, The Ralph and Goldy Lewis Center for Regional Policy Studies, School of Public Policy and Social Research, University of California at Los Angeles (UCLA).

Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*, 3rd edition, Prentice-Hall International.

Button, K.J, A.S. Fowkes and A.D. Pearman (1980). "Disaggregate and Aggregate Car Ownership Forecasting in Great Britain", *Transportation Research Part A*, Vol.14A, pp.263-273.

Danos, S. (2004). *A Comparative Evaluation of Aggregate Car Ownership Models in Order to Forecast Fuel Consumption and $CO_2$ Emissions in Passenger Transport: the Case of Greece*, MSc Thesis, University of Piraeus [in Greek].

Dargay, J. and D. Gately (1997). "Vehicle Ownership to 2015: Implications for Energy Use and Emissions", *Energy Policy*, Vol.25, Nos 14-15, pp.1121-1127.

Dargay, J. and D. Gately (1999). "Income's Effect on Car and Vehicle Ownership", *Transportation Research Part A*, Vol.33, pp.101-138.

Gately, D. (1990). "The US Demand for Highway Travel and Motor Fuel", *Energy Journal*, 11(3), pp.59-72.

Goodbody Economic Consultants (2000). *Travel Demand*, Dublin, Ireland, November.

Gujarati, D. (2003). *Basic Econometrics*. 4th Edition, McGraw-Hill.

Hendry, D.F. and G.E. Mizon (1978). "Serial Correlation as a Convenient Simplification, Not a Nuisance: A Comment on a Study for Money by the Bank of England", *The Economic Journal*, 88, pp.549-563, September.

Kennedy, P. (2001). *A Guide to Econometrics*, 4th Edition, The MIT Press, Cambridge, Massachusetts.

Lam, W.H.K. and M.L. Tam (2002). "Reliability of Territory-Wide Car Ownership Estimates in Hong Kong", *Journal of Transport Geography*, 10 (2002), pp.51-60.

London Borough of Croydon (1995). *Transport*, Croydon Environment Audit, No.7, London.

Makridakis, S., S.C. Wheelwright and R.J. Hyndman (1998). *Forecasting: Methods and Applications*, 3rd edition, John Wiley and Sons.

Nau, B. (2006). "Identifying the Order of Differencing", Online Lecture Notes, *Forecasting* (Decision 411), Fuqua School of Business, Duke University, (http://www.duke.edu/%7Ernau/411arim2.htm, accessed January 2006).

Paravantis, J.A. and P.D. Prevedouros (2001). "Railroads in Greece: History, Characteristics and Forecasts", *Transportation Research Record*, No.1742, pp.34-44.

Prevedouros, P.D. and P. An, (1998). "Automobile Ownership in Asian Countries: Historical Trends and Forecasts", *ITE Journal*, 68, pp.24-29.

Scholl, L., L. Schipper and N. Kiang (1996). "$CO_2$ Emissions from Passenger Transport: A Comparison of International Trends from 1973 to 1992", *Energy Policy*, Vol.24, No.1, pp.17-30.

Studenmund, A.H. (1992). *Using Econometrics*, 2nd edition, Harper Collins.