Stability of parameters estimated on Cross-sectional data

Jens Erik Nielsen, Rikke J. Broegård and Mogens Fosgerau
Danmarks TransportForskning

Abstract

This paper discusses the stability of parameters used in models estimated on cross-sectional data and thus the appropriateness of using such models in forecasting. The problem of parameter instability is demonstrated by using a simple logit model for car availability in the Danish households, which is estimated on data from the Danish travel diary data between 1995 and 1999. The combination of using a large and highly reliable data set, together with a simple model makes a good case for examining the inherent stability of the parameters. The effects of wealth on car availability receive special attention because these are expected to change over time. Cohort effects are also addressed.

Introduction:

Traffic models are increasingly used to make forecasts for improving the understanding of future trends in travel demand. They frequently form the basis upon which decisions are made about future actions, such as large infrastructure investments. There are several Danish examples of traffic models that are based on cross-sectional data. These include the main parts of the fixed link projects in Denmark (Great Belt (Storebælt 1991), Oresund (Øresundskonsortiet 1999) and Femern (Trafikministeriet 1999)) and national model systems like PETRA (Transportrådet 1999) and ALTRANS (Christensen et. al. 2001), as well as various models for the Copenhagen region e.g. Orestad Traffic model (Jovicic and Hansen 2001). Many more examples can be found in other countries.

For disaggregate modelling, the data are mostly cross-sectional, that is, they represent observational units (variables) at a single point in time (or a short period). It is clear that the use of such models is necessarily based on the assumption that any relationships that are estimated from cross-sectional data can reasonably be extended into the future. This assumption is the subject of the current article.

Despite the important consequences and costs of decisions that are being taken based on these models there are surprisingly few studies, which have investigated whether parameters estimated from cross-sectional data are sufficiently stable over time. Indeed, the investment in developing the models contrasts sharply with the investment in testing the assumptions upon which they are based. For example one possible problem related to using models that are estimated on cross-sectional data is that the models may not necessarily account for income, wealth and cohort effects (Jansson 1988, 1989).

This article utilises a short time series of cross-sectional data to investigate the stability of parameters estimated on such data. This allows us to focus on the issue at hand, namely the stability of the relationship between exogenous and endogenous variables, while avoiding the need to forecast exogenous variables. The household choice between having none, one or more cars available is used as a test case and a logit model is used to model this choice. The paper will mainly focus on the impact of wealth on parameter stability, and it examines the question of whether wealth affect the stability of the parameters.

The model and the data:

A logit model was constructed for the number of cars available to Danish household in order to test the stability of the selected parameters. The model predicts whether a given household has zero, one or more than one car available. The model variables are shown in table 1.

Table 1 Model variables

Variable	Variable description	Variable	Variable description
ASC	Constant	u_large	Home in large city (over 70000 inh.)
inc	Natural logarithm of after tax income	u_suburb	Home in suburbs of Copenhagen
time_pub	Distance to public transport	u_small	Home in small town (2000 to 10000 inh.)
lic	Drivers licence holding variable	u_country	Home in the country (less than 2000 inh.)
lic_f	Drivers licence holding variable if a women	ac_det	Living in a detached house
age	Age	ac_ter	Living on a semi-detached house
age2	Age squared	ac_farm	Living on a farm
u_copen	Home in Copenhagen		

Data from the Danish travel diary between 1995 and 1999 provide the input to the model[1]. The data contain information concerning an interviewee (IP), as well as some variables describing the household of the IP. In order to keep the sample homogeneous, observations were only included for households containing couples with children[2]. Income has been discounted, allowing comparison between different years[3].

[1] The data contains 18530 observations with 2404 observations in 1995, 3958 observations in 1996, 3948 observations in 1997, 4157 observations in 1998 and 4063 observations in 1999.

[2] Consequently the model does not represent all types of households. However, as the focus of the article is on the stability of parameters over time, this is less problematic.

[3] Efforts were made to ensure consistency over time concerning the population segments reached by the travel diary, for example through the exclusion of people younger than 16 and older than 74. All observations where the IP is not a part of the “head of household couple” have been excluded in order to avoid young people who is living with their parents, thus having high car availability rates and low personal income.

A model design has been chosen for which 1999 was the year of reference and the parameters for 1995 to 1998 are relative to the parameters estimated for 1999. This parameterisation allows direct examination of the differences in parameters relative to the parameters for 1999.

Estimation and analysis:

Initially the model is based upon data for the entire time period in order to allow parameters to differ among years. Subsequently, restrictions are introduced by requiring the parameter values to be identical for all of the years 1995-1999 (i.e. assuming stability over time). The statistics and estimation of results for these two models are shown in table 2[4].

Table 2 Estimation results for the full model and the restricted model

	Different parameters for each year	Parameters identical between years
Number of observations	18530	18530
Log likelihood, zero coefficients	-20357.3	-20357.2
Log likelihood, constants only (degrees of freedom)	-14834.2 (10)	-14834.3 (2)
Log likelihood, final model (degrees of freedom)	-12228.2 (150)	-12334.2 (30)
Rho-squared, zero coefficients	0.3993	0.3941
Rho-squared, constants only	0.1757	0.1685
χ^2-value for the restriction	0.00000	

With the χ^2-value being very low, the chi-square test strongly rejects the restriction imposed on the parameters and we therefore conclude that the parameters cannot be assumed to be constant over time.

The model is now estimated using 1999 data only, in order to explore the cause of instability. The statistics for the resulting model are shown in 3.

Table 3 Estimation results the model estimated on 1999 data only

	Results of estimation on 1999 data only
Number of observations	4063
Log likelihood, zero coefficients	-4463.7
Log likelihood, constants only (degrees of freedom)	-3303.8 (2)
Log likelihood, final model (degrees of freedom)	-2728.9 (30)
Rho-squared, zero coefficients	0.3886
Rho-squared, constants only	0.1740

The resulting model is used to make a backcast for the years 1995 to 1998. The actual and predicted numbers of cars per household are shown in figure 1. The figure shows that the model overestimates the number of cars in earlier years. The error is very small in 1997 and 1998, but it becomes considerable for 1995 and 1996.

[4] All estimation results including parameter estimates for this article are available from the authors on request.

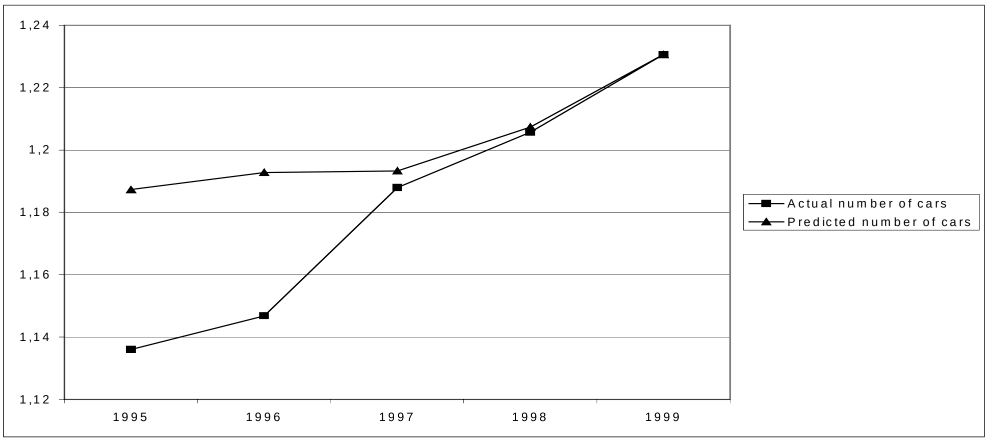

Figure 1 Average number of cars in a household (actual and predicted)

In the figure below the estimated distribution of households according to their car availability is compared with the observed distribution. It shows that the model overestimates the number of households with two or more cars available. It also shows that the model underestimates the number of households with only one car available. However, the difference between the predicted and actual number of households with no car availability is quite small, indicating the model predicts a shift from having one car available to having two or more cars available. Such a shift would increase the average number of cars in the households, which in turn would explain the error observed in figure 1.

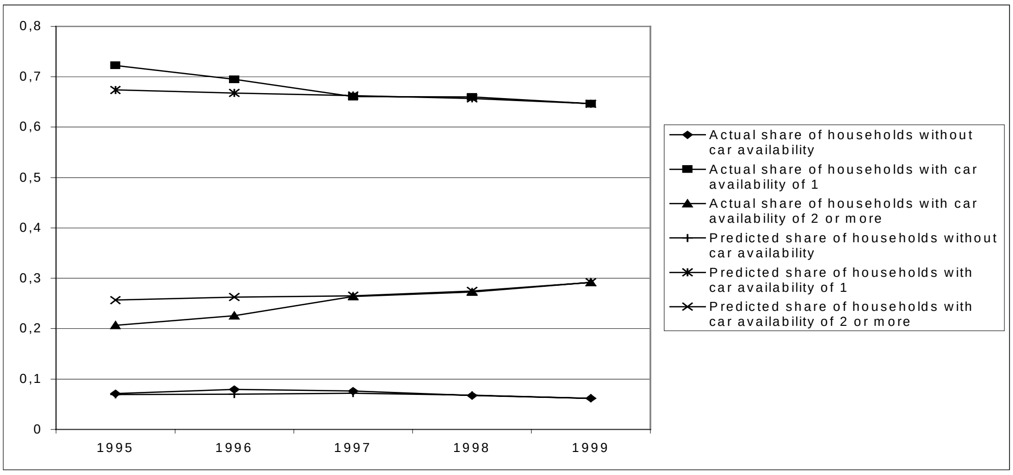

Figure 2 Share of households grouped by number of cars available (actual and predicted)

It is evident from figure 1 and figure 2 that the instability of data mainly shows up in 1995 and 1996. In order to examine this in more detail, the model is now used to compare predicted and actual car availability in different kinds of households in 1995 and 1996, according to their type of accommodation. These results are shown in table 4.

Table 4 Car availability in different accommodations (1995 and 1996)

No car availability (number of households)								
Accommodation type	Actual (1995)	Predicted (1995)	Difference	Standard deviation	Actual (1996)	Predicted (1996)	Difference	Standard deviation
Detached house	70	55,8	-14,2	6,9	103	83,9	-19,1	8,6
Semi-detached house	31	43,4	12,4	5,4	60	69,7	9,7	6,8
Apartment	67	64,8	-2,2	6,1	147	117,2	-29,8	7,9
Farm	2	2,6	0,6	1,6	3	5,1	2,1	2,1

One car available (number of households)								
Accommodation type	Actual (1995)	Predicted (1995)	Difference	Standard deviation	Actual (1996)	Predicted (1996)	Difference	Standard deviation
Detached house	1214	1139,9	-74,1	18,1	1947	1864,9	-82,1	23,2
Semi-detached house	204	188,5	-15,5	7,0	316	305,7	-10,3	9,0
Apartment	182	179,2	-2,8	7,1	261	280,4	19,4	9,1
Farm	137	113,0	-24,0	7,3	227	192,1	-34,9	9,7

Two or more cars available (number of households)								
Accommodation type	Actual (1995)	Predicted (1995)	Difference	Standard deviation	Actual (1996)	Predicted (1996)	Difference	Standard deviation
Detached house	355	443,3	88,3	17,2	629	730,2	101,2	22,1
Semi-detached house	28	31,1	3,1	5,1	51	51,6	0,6	6,5
Apartment	16	21,0	5,0	4,2	22	32,4	10,4	5,2
Farm	98	121,4	23,4	7,2	192	224,8	32,8	9,6

The table shows that the largest shift in car availability occurred for the households living on farms and in detached houses. The model underestimates the number of households living on farms having one car available with more than 3,3 and 3,6 times the standard deviation and with more than 4,0 and 3,5 times the standard deviation for detached houses in 1995 and 1996 respectively. At the same time it overestimates the number of households living on farms having two or more cars available by more than 3,2 and 3,4 times the standard deviation and with more than 5,1 and 4,6 times the standard deviation for households living in detached houses. With a high proportion of the households living in detached houses, the error for this segment has a high impact on the overall results shown in figure 1. We will therefore focus on this segment in the following.

In parallel to the above comparison, table 5 compares actual and predicted car availability between different age groups. The model generally underestimates the number of households having one car available and overestimates the number of households with two or more cars available. The table shows that the error is concentrated in the age group representing people between 35 and 44. In this group the underestimation of households with one car available is more than 3 times the standard deviation and the overestimation of households with two or more cars available is more than 4 times the standard deviation. More general it can be said that the model underestimates the number of households with middle-aged people having one car available and overestimates the number of households with middle-aged people having two or more cars available.

Table 5 Car availability in different age groups (1995 and 1996)

No car availability (number of households)								
Age-group	Actual (1995)	Predicted (1995)	Difference	Standard deviation	Actual (1996)	Predicted (1996)	Difference	Standard deviation
15-24	10	12.7	2,7	2,5	8	5,8	-2,2	1,8
25-34	128	110.1	-17,9	8,3	72	66,6	-5,4	6,7
35-44	111	93.4	-17,6	8,3	70	61,9	-8,4	6,7
45-54	59	50.5	-8,5	6,1	19	26,1	7,1	4,5
55-64	5	8.2	3,2	2,3	1	5,8	4,8	2,0
65-74	0	1.0	1,0	0,9	0	0,4	0,4	0,6
One car available (number of households)								
Age-group	Actual (1995)	Predicted (1995)	Difference	Standard deviation	Actual (1996)	Predicted (1996)	Difference	Standard deviation
15-24	48	39,8	-8,2	3,3	23	21,7	-1,3	2,4
25-34	858	840,1	-17,9	15,4	561	524,5	-36,5	12,1
35-44	1172	1106,4	-65,6	18,1	754	708,1	-45,9	14,5
45-54	603	589,3	-13,7	13,9	344	322,0	-22,0	10,2
55-64	65	61,8	-3,2	4,7	50	41,4	-8,6	3,7
65-74	5	5,6	0,6	1,5	5	2,9	-2,1	1,1
Two or more cars available (number of households)								
Age-group	Actual (1995)	Predicted (1995)	Difference	Standard deviation	Actual (1996)	Predicted (1996)	Difference	Standard deviation
15-24	2	7,6	5,6	2,4	1	4,5	3,5	1,7
25-34	234	269,7	35,7	13,4	120	161,9	41,9	10,5
35-44	357	440,1	83,1	16,7	221	275,0	54,0	13,3
45-54	263	285,1	22,1	12,9	139	153,9	14,9	9,6
55-64	33	33,0	0,0	4,3	15	18,8	3,8	3,3
65-74	5	3,4	-1,6	1,3	1	2,7	1,7	1,0

In order to determine whether the previously observed instability can be explained by the parameters of age, income and type of accommodation we estimate the model with restrictions on all parameters but these. The statistics for this partial restriction are shown in table 2.

Table 6 Estimation results for the full model and the partially restricted model

	Full model	All parameters restricted but income, age and type of accommodation
Number of observations	18530	18530
Log likelihood, zero coefficients	-20357.3	-20357,3
Log likelihood, constants only (degrees of freedom)	-14834.2 (10)	-14834,2 (2)
Log likelihood, final model (degrees of freedom)	-12228.2 (150)	-12285,2 (72)
Rho-squared, zero coefficients	0,3993	0,3965
Rho-squared, constants only	0,1757	0,1718
χ^2-value for the reduction	0,005	

Once again the χ^2-value is very low and the chi-square test rejects the proposed restriction. Therefore, there must be other causes besides age, income and type of accommodation, which contribute to the observed instability. However, although these parameters do not provide a satisfying explanation, they might be important contributors to the instability.

The large error in the backcast occurring between 1995 and 1997 indicates the existence of significant effects taking place in this time-period not captured by the model. One possible explanation of the change could be related to the housing market. Housing prices have risen steadily in the years from 1993 and onwards, as shown in figure 3.

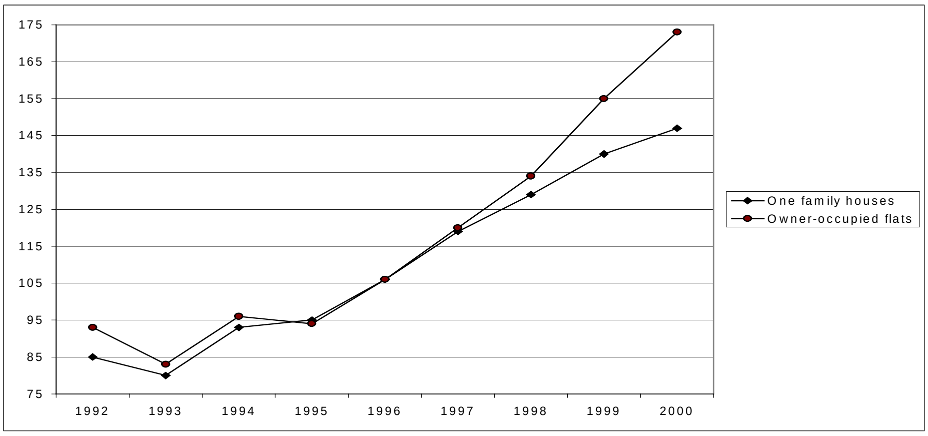

Figure 3 Price index for real estate (1980=100. Data from StatBank Denmark)

This housing price increase was followed by a wave of real estate loan consolidations and homeowners were the main beneficiaries of a capital gain. The development in lending activities for owner-occupied dwellings presented in figure 4 shows that there has been a large increase in the number of new loans, especially between 1993 and 1998.

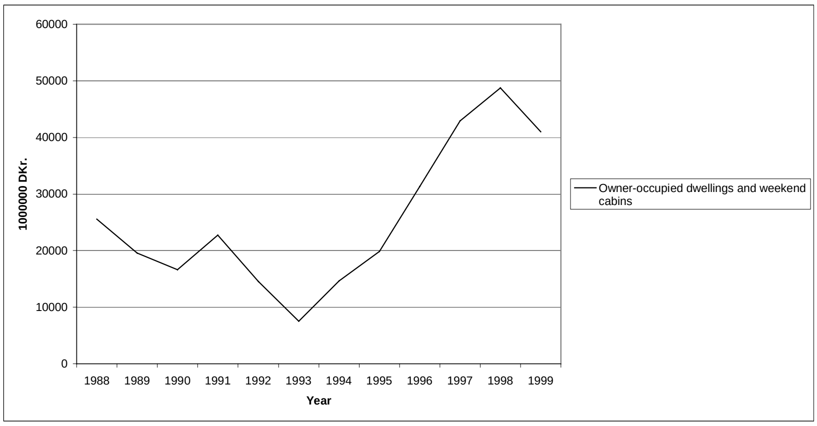

Figure 4 Lending activities of mortgage credit associations (Data from StatBank Denmark, Danmarks Statistik 1993 and Danmarks Statistik 1991)

These developments might explain the estimation differences observed in table 3. The hypothesis would be that the households realising the highest capital gain are homeowners and that they use this gain to purchase a second car. The initial model does not capture this wealth effect. The model estimated only on 1999 data applies this trend as the norm for the entire period and therefor overestimates the number of cars for the earlier part of the period, before the effect of the loan consolidation took place.

Another possible explanation for the estimation differences could be a cohort effect. Figure 5 and 6 show that a larger proportion of young people have a driver's license relative to older generations (see also Jansson 1989, 1990). In parallel, the desired availability of cars in those households with younger people may be greater than in other households. Additionally, young couples have higher need for a second car, due to the fact that younger women tend to work more than do older women (Danmarks Statistik 2000).

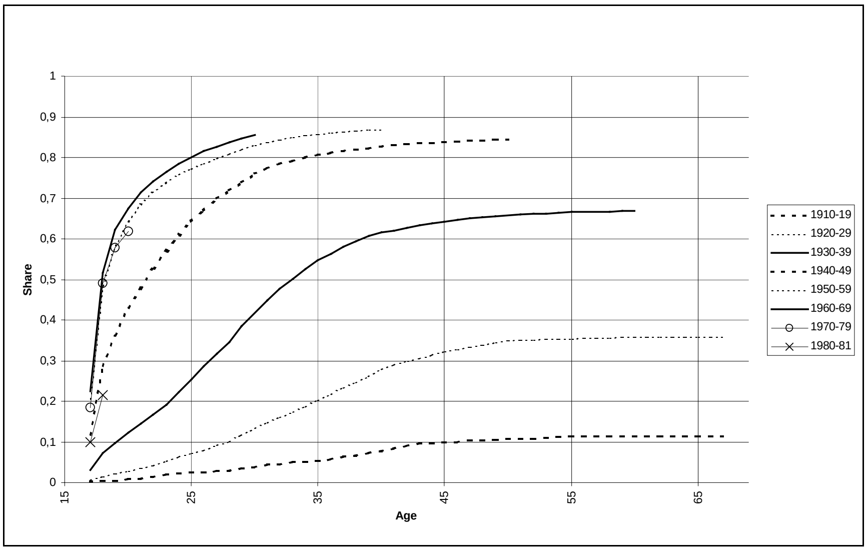

Figure 5 Driver license-holding rates for men

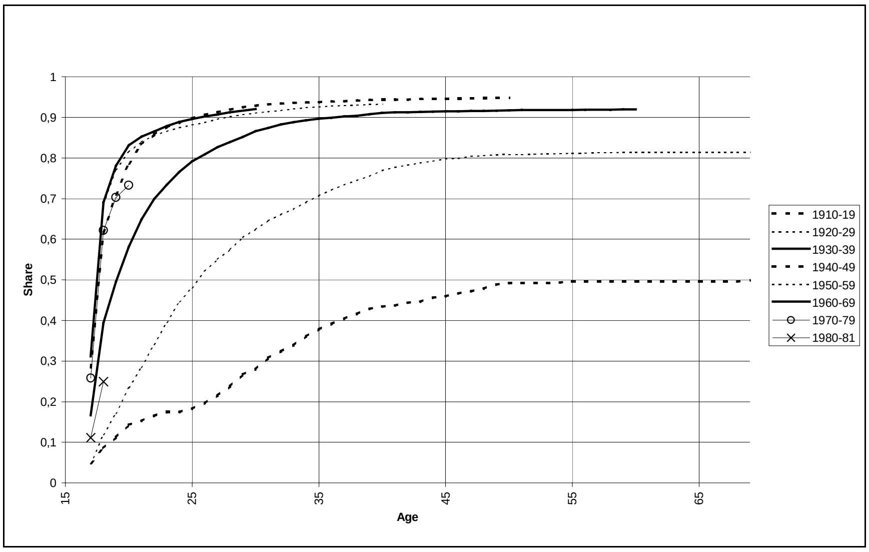

Figure 6 Driver license-holding rates for women

This cohort effect influences the car availability in different age groups. Using a parameter for license holding we are able to capture some of this effect. However, there might still be some cohort effect, which is not captured by the model.

The model is now used to calculate income elasticities for car availability for 1995 to 1999. The results are shown in figure 7.

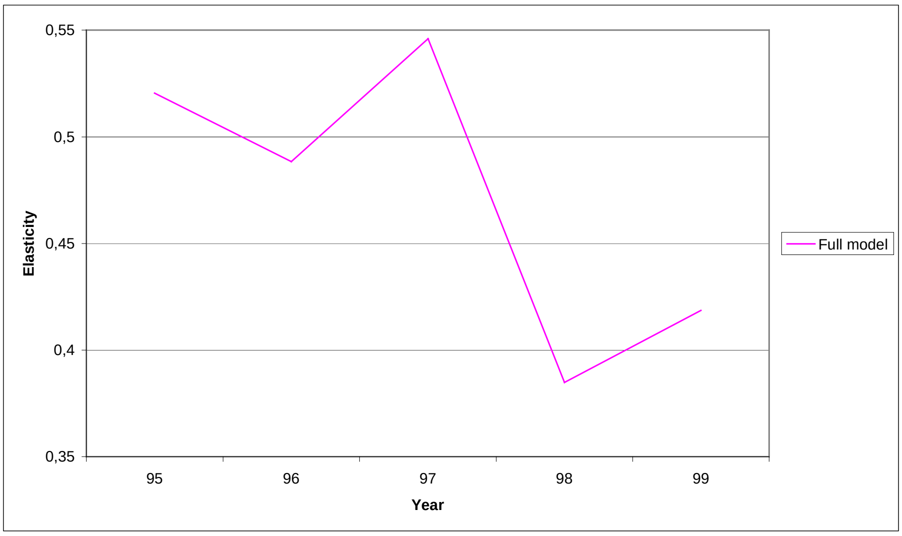

Figure 7 Income elasticities for car availability in the households.

The elasticity differs between 0.385 and 0.546. The difference between the starting and endpoint is approximately 0.1. A forecast estimating the effect on the number of cars, in response to a 10% increase in income would differ by approximately 17000, depending on which of these elasticities are used[5]. However, in comparison with other recent transport studies (e.g. Birkeland et. al. 2000), it seems that the elasticities found here for 1995, 1996 and 1997 might be too high[6]. As seen, overestimated income elasticities are transformed into considerable errors when used in forecasting.

Conclusions:

The stability of parameters estimated on cross-sectional data has been addressed using a simple logit model for car availability. The model was estimated using a short time series and it was shown that the hypothesis that parameters are stable over time might lead to errors when such a model is used in forecasting. In other words, some parameters show instability over time. When the model was used to make a backcast, it overestimated car availability in households. These erroneous estimates mainly involved middle-aged households living in detached houses. On this basis, it is likely that a wealth effect may have changed these households' accessibility to cars, thereby allowing them to aquire a second car. This could

[5] In 2000 there roughly 1.7 million cars in the Danish households (Danmarks Statistik, 2000)

[6] Birkeland et. al. (2000) uses both a cross-sectional analysis and a pseudo-panel analysis to estimate income elasticities. The cross-sectional analysis finds income elasticities ranging from 0.28 to 0.48 and the pseudo-panel analysis finds the income elasticity to be 0.19. They also has references to a number of other studies, all but one finding income elasticity to be lower than 0.45.

partly explain the error in the backcast, while cohort effects could offer another part of the explanation. By using the model to calculate income elasticities it was shown that the lack of stability might also give rise to erroneous forecasting. The inclusion of variables normally assumed to be exogenous might offer a (partial) solution to the problem. This is due to the fact that their inclusion may capture some of the effects causing instability. Still, the results presented in this article underscores of the importance of critically evaluating the assumptions upon which models are based, and scrutinising the results in the light of the restrictions that are imposed by these assumptions.

References:

Ben-Akiva, M. and S. R. Lerman (1995) *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press, Cambridge, Massachusetts.

Birkeland, M. E., Brems, C. R. & Kabelmann, T. (2000) Analyser af personers transportarbejde, 1975-1998, *Trafikdage På Aalborg Universitet 2000,* Konferencerapport, 549-558.

Christensen, L., Kveiborg, O. & Rich, J. H. (2001) *ALTRANS, En Model for Persontrafik, En oversigt over metoder og resultater,* Faglig rapport fra DMU, Afdeling for Systemanalyse, forthcoming.

Danmarks Statistik (2000) *Statistisk Årbog 2000*, Danmarks Statistik, November, 2000.

Danmarks Statistik (1993) *Statistisk Årbog 1993*, Danmarks Statistik, August, 1993.

Danmarks Statistik (1991) *Statistisk Årbog 1991*, Danmarks Statistik, September, 1991.

Jansson, J. O. (1990) Car Ownership Entry and Exit Propensities of Different Generations – A Key Factor for the Development of the Total Car Fleet, *Oxford Conference on Travel and Transportation*, July, 417-435.

Jansson, J. O. (1989) Car Demand Modelling and Forecast, A New Approach, *Journal of Transport Economics and Policy*, **13(2)**, pp. 125-140.

Jovicic, G. & Hansen, C. O. (2001), The Orestad Traffic Passenger Demand Model, *Trafikdage på Aalborg Universitet 2001*, forthcoming.

StatBank Denmark, *www.statistikbanken.dk.*

Storebælt (1991) *Øst-vesttrafikmodellen, Prognoser for trafikken mellem Øst- og Vestdanmark*, Storebælt, February 1991.

Trafikministeriet (1999) *Femer Bælt-Forbindelsen – Forundersøgelser – Resumerapport,* Trafikministeriet, Marts 1999.

Transportrådet (1999) *PETRA – analysemodel for persontransport*, Transportrådet, Notat 99-06, Oktober 1999.

Øresundskonsortiet (1999) *Traffic Forecast Model, The Fixed Link across Øresund*, Øresundskonsortiet, June 1999.