Control Function Approach for Addressing Endogeneity on the London-Amsterdam Bimodal Corridor

Thomas E. Guerrero B.[a], Nicolò Avogadro[b], and Raul Ramos[c]

[a]Department of Civil Engineering, Universidad Francisco de Paula Santander Seccional Ocaña, Sede El Algodonal Ocaña, Colombia.
[b]Department of Management, Information and Production Engineering, University of Bergamo, Italy.
[c]Departamento de Ingeniería de Transporte y Logística, Pontificia Universidad Católica de Chile, Santiago, Chile.

Short summary

This research examines the competition between High-Speed Rail and air travel in the London-Amsterdam bimodal corridor. Using Discrete Choice Models, we revealed endogeneity in the fare. To address this issue, we employed the Control Function approach, which allowed us to identify and rectify inaccuracies in fare estimates. Our findings indicate that the corrected model significantly outperforms the original endogenous model in terms of accuracy and reliability, underscoring the critical importance of addressing these methodological challenges.
Keywords: Endogeneity; Discrete Choice Model; Control function; High-Speed Rail; Air transport.

1 Introduction

High-Speed Rail (HSR) has become an affordable and sustainable alternative to air travel for domestic and cross-border trips worldwide (see, e.g., Zhang et al., 2019). Following this trend, researchers are increasingly investigating drivers underpinning transport demand in markets where HSR competes with air carriers. The standard approach to perform these analyzes is the use of linear econometric models or adopt gravity formulation approaches where the dependent variable is the aggregated demand or the market share of the single alternative (see, e.g., Avogadro et al., 2023). However, at an individual level, transport demand is not continuous, but discrete, as single passengers choose the best alternative to make their trip. Thus, empirical studies that use individual data extensively rely on Discrete Choice Models (DCM) to investigate passenger preferences.

Most studies investigating HSR and air transport demand using passenger choices leveraged Stated Preferences (SP) surveys alone or combined with Revealed Preferences (RP) data (see, e.g., Hong & Najmi, 2022; Bergantino & Madio, 2020). Not surprisingly, from a methodological perspective, the majority of previous works estimate Multinomial (MNL) or Nested Logit (NL) models, which are the most common DCM applied in the transport literature since they have closed-form expressions for the choice probability, making them computationally easy to estimate. Regarding modal attributes, fare, frequency, and travel time are among the most investigated explanatory variables. However, other variables such as the access and egress time as well as comfort and reliability also demonstrated to affect passenger choices.

Notably, all previous studies leveraging DCM systematically disregard possible correlations between the error term and the explanatory variables, which may lead to biased and inconsistent estimates of the model parameters (C. A. Guevara & Ben-Akiva, 2012). This so-called endogeneity problem has been extensively investigated in the case of linear models. For example, it is well-known that the fare may be endogenous because higher consumption allows transport providers with market power to charge higher prices (Birolini et al., 2020). Moreover, demand may also affect the frequency, so even this variable could introduce endogeneity problems (Birolini et al., 2020). Another source of endogeneity is the correlation of ticket price with unobservable quality characteristics that can affect demand, an effect known as omitted variable bias (Li et al., 2020).

Based on this gap, the contribution of the current paper is twofold: (i) evaluate potential sources of endogeneity in DCM analyzing a multimodal corridor featuring the competition between HSR and air transport; (ii) apply the Control Function (CF) approach to rectify endogeneity issues proposing the fare observed in similar markets and the price of the single alternative's fuel (i.e. oil or electricity) as instruments.

2 Modeling framework

Let us assume a DCM with endogeneity due to omitting a certain variable, q, correlated with an observable variable X. Let V_{in} denote the utility obtained by individual n when choosing alternative i:

$$V_{in} = ASC_i + \beta_y Y_{in} + \beta_x X_{in} + \underbrace{\beta_q q_{in} + e_{in}}_{\varepsilon_{in}}, \tag{1}$$

where ASC_i is an Alternative Specific Constant, Y_{in} is a set of known (measurable) and exogenous attributes, X_{in} is the possible endogenous variable, β_y, β_x and β_q are parameters to be estimated, and e_{in} is an exogenous error term. Then, assuming that q_{in} is an unobservable variable, the specification proposed by the modeler is:

$$V_{in} = ASC_i + \beta_y Y_{in} + \beta_x X_{in} + \varepsilon_{in}, \tag{2}$$

where the new error term, ε_{in}, contains both e_{in} and q_{in}. The variable X_{in} is endogenous because it is correlated with ε_{in} through q_{in} in Eq. (2). The correlation arises because in Eq. (3), the variable X_{in} depends on q_{in} as follows:

$$X_{in} = \gamma_0 + \gamma_{z_1} z_{1,in} + \gamma_{z_2} z_{2,in} + \gamma_y Y_{in} + \underbrace{\gamma_q q_{in} + \phi_{in}}_{\delta_{in}}, \tag{3}$$

where $z_{1,in}$ and $z_{2,in}$ are the so-called Instrumental Variables (IVs or instruments), and ϕ_{in} is an exogenous error term. For illustration, we assume that the endogenous variable X_{in} is correlated with the exogenous variables Y_{in}, and the error term δ_{in} contains both ϕ_{in} and q_{in}. Then, it can be shown that the DCM corrected by endogeneity has the following functional form (C. A. Guevara & Ben-Akiva, 2012):

$$V_{in} = \widetilde{ASC}_i + \widetilde{\beta}_y Y_{in} + \widetilde{\beta}_x X_{in} + \beta_\delta \widetilde{\delta}_{in} + \widetilde{e}_{in} \tag{4}$$

where $\widetilde{\delta}_{in}$ is a proper estimator of δ_{in}. The intuition is that $\widetilde{\delta}_{in}$ captures the part of the endogenous variable X_{in} correlated with the error term ε_{in}. Therefore, if the instruments $z_{1,in}$ and $z_{2,in}$ are truly exogenous and correlated with the endogenous variable, $\widetilde{\delta}_{in}$ should be added to the utility V_{in} to control the endogeneity problem. The practical implementation of the CF approach follows two main stages. First, find an estimator for δ_{in}, which can be computed as the residual of the Ordinary Least Squares (OLS) regression of X_{in} on the instruments ($z_{1,in}$, $z_{2,in}$) and the exogenous variables (Y_{in}). Second, estimate the DCM considering $\widetilde{\delta}_{in}$, X_{in} and Y_{in} as explanatory variables to obtain consistent estimators $\widetilde{\beta}_x$ for the parameter β_x (C. A. Guevara & Ben-Akiva, 2012).

We apply this method to analyze one of the latest additions to the European HSR network: the Eurostar connection between Amsterdam and London. The HSR alternative entered this highly dense route in 2018, gaining a market share of about 10% against a broad set of air alternatives, including traditional and low-cost carriers, connecting Amsterdam with various airports within the London airport system. To analyze passenger preferences in this market, we leveraged a set of revealed preferences data from the International Passenger Survey (IPS), conducted by the UK Office for National Statistics (ONS), interviewing passengers traveling to and from the UK by air or the channel tunnel.

3 Results

Firstly, while the possibility of endogeneity in frequency is somewhat argued in transportation literature, for the London-Amsterdam market we did not find any proof of endogeneity concerning this variable. This is likely due to the peculiarities of the specific market. Contrariwise, we proved endogeneity in the fare when estimating the MNL. To correct this issue, we propose two instruments: (i) the average fares observed for trips that occurred during the same month and year but on a similar market (i.e., London-Paris) by mode (z_{fare}), and (ii) the average price of each alternative's power source (i.e. oil or electricity) for the month and year when the trip occurred (z_{power}).

The two IVs we propose are correlated with the endogenous variable fare and do not confound with market share. In other words, they can affect the endogenous variable through aggregate travel demand, but not the individual passenger's travel utility and associated unobservable service attributes. Fuel cost has been considered a valid instrument as it is correlated with ticket price and not confounded with market share (Birolini et al., 2020). Similarly, Lurkin et al. (2017) use average prices for other markets as an effective instrument to control for the effects of endogeneity in modeling airline price-demand elasticity.

Table 1 reports the endogenous and corrected DCM estimates. In addition to relevant supply attributes that affects passenger choices, both models were estimated including Alternative Specific Constants (ASCs), where each alternative is a single travel option. The left-hand side model on Table 1 considers on-board travel time, access/egress time, the logarithm of the weekly frequency, the expected delay, and the fare, which are the variables that usually contribute to the choice of consumers in this context. Frequency is included in logarithmic form for two reasons. First, to account for the expected decreasing marginal utility of frequency. Second, since a route alternative is considered as an aggregation of individual flights/trains, the logarithmic form is the most suitable for a characteristic that captures the size of an aggregated alternative. Except for the one related to the fare, the endogenous model parameters show correct signs and are statistically significant at 99%. However, the positive sign of the fare parameter is clear evidence of the endogeneity effects of this model, because it is not in line with the basic economic theory and typical passenger behavior.

Table 1: Endogenous and corrected DCM for the multimodal London-Amsterdam market.

Variable	Alternative	Endogenous model		Corrected model	
		Coefficient*	Std. Error**	Coefficient*	Std. Error**
Alternative Specific Constant (ASC)	AIR 1	-11.230	0.8463	-8.2560	0.8675
	AIR 2	-10.960	0.8202	-8.0765	0.8396
	AIR 3	-8.8570	0.7765	-6.3023	0.7930
	AIR 4	-8.1550	0.7815	-5.6479	0.7986
	AIR 5	-9.2730	0.7876	-6.3260	0.8118
	AIR 6	-9.1810	0.7974	-6.1407	0.8228
	AIR 7	-8.6980	0.8252	-6.2176	0.8399
	AIR 8	-9.1370	0.7994	-6.6753	0.8135
	AIR 9	-9.3210	0.8356	-6.5675	0.8528
	HSR	(base)	-	(base)	-
On-board travel time	All	-0.0638	0.0053	-0.0457	0.0054
Access/egress time	All	-0.0529	0.0006	-0.0532	0.0007
Log (weakly frequency)	All	0.6787	0.0562	0.6541	0.0591
Expected delay	All	-0.0140	0.0033	-0.0114	0.0033
Fare	All	**0.0026**	0.0003	**-0.0101**	0.0010
Fare's residual	All	-	-	0.0147	0.0010
Sample size		*5199*		*5199*	
Log-likelihood		*-8921.164*		*-8892.846*	

*All parameters are significant at the 99% level.

**Standard errors determined using Bootstrap.

The right-hand side of Table 1 reports the model corrected for endogeneity using the CF approach. Since the model is estimated in two stages, the standard errors cannot be directly inferred from the Fisher-information matrix. Therefore, the variance-covariance matrix must be determined using a non-parametric method, in this case, we use bootstrapping (Petrin & Train, 2003). The corrected model provides one additional estimate compared to the endogenous model: the one corresponding to the fare residuals derived from the first stage of the CF method.

Before discussing the estimated parameters, let us verify the hypothesis of fare endogeneity in the uncorrected model and the compliance of the proposed IVs to exogeneity and relevance requirements. Considering the former, the endogeneity of the fare can be proved following the Rivers & Vuong (1988) method by evaluating the significance of the residuals of the endogenous variables in the second stage of the CF approach. In our case, the fare's residuals are significant in the second stage of the CF approach, thus there is evidence that the model exhibits endogeneity due to the fare variable. Concerning the relevance condition, we consider the recommendation of C. Guevara & Navarro (2015) to check that the F test is greater than 10 for the first stage regression in the CF method. On the other hand, the instruments' exogeneity condition for DCM was confirmed through S_{mREF} test, which is the most recommended due to its larger power, smaller size distortion, and more robustness compared to others analyzed by C. A. Guevara (2018). Finally, to evaluate the goodness of the two models (endogenous and corrected), we used the likelihood ratio (LR) test. Since the LR value exceeds the critical value, we conclude that the corrected model outperforms the restricted one. Compliance with the relevance and exogeneity conditions, as well as the results of the goodness of fit test, are summarized in Table 2.

Table 2: Relevance, exogeneity, and goodness tests for the Instrumental Variables

Property	Test	Value	Threshold	
Relevance	F-test (fare)	11.85	> 10	✓
Exogeneity	S_{mREF}	0.204	< 3.84	✓
Goodness	LR	56.64	> 3.84	✓

Considering the coefficients of the corrected model, we observe that the CF method changes the sign of the fare coefficient, which confirms the bias in the model where endogeneity is overlooked. Overall, the model confirms the sensitivities toward the different travel characteristics reported in previous transport mode choice studies. Namely, increases in travel time, access time, expected delay, and fare negatively influence passenger utility. Concurrently, higher frequency increases passenger attitudes toward the single travel alternative.

4 Conclusions

We researched sources of endogeneity in the multimodal corridor between Amsterdam and London. It is shown that the frequency is not endogenous, because of certain market features, confirming that numerous particular market configurations may limit the occurrence of this problem, adds to the literature on frequency. Our correction of the endogeneity of prices was completed successfully with two IVs. Where it concerns values and awareness of bias stemming from these endogenous specifications.

Dealing with endogeneity adds a layer of complexity to our efforts in modeling the London-Amsterdam multimodal (HSR and air transportation) market. We identified the fare variable as a potential troublemaker due to the simultaneous estimation and the omission of specific factors that usually affect this variable. The results from correction using the CF approach highlight that the corrected model performs better than the original model (endogenous). This finding was tested according to LR tests.

Endogeneity concerns for the case study on the London-Amsterdam route, where two modes compete, were adequately addressed using the CF framework. The CF method proves to be a well-suited methodology for this application. During the research, an important challenge was to find a candidate set of instruments that show relevance and exogeneity, supporting their use in correcting endogeneity in the endogenous variable (fare). The selected instrumental variables were: (i) the average fares observed for trips that occurred during the same month and year but on a similar market (i.e., London-Paris) by mode, and (ii) the average price of each alternative's power source (i.e., oil or electricity) for the month and year when the trip occurred. The identification of these variables represents a significant contribution, affirming their validity. Future researchers are encouraged to adopt the CF approach and highlight instrumental variables as a reliable guide for addressing potential endogeneity issues in mode choice models.

References

Avogadro, N., Pels, E., & Redondi, R. (2023, June). Policy impacts on the propensity to travel by HSR in the Amsterdam – London market. *Socio-Economic Planning Sciences*, *87*, 101585. (Publisher: Pergamon) doi: 10.1016/J.SEPS.2023.101585

Bergantino, A. S., & Madio, L. (2020). Intermodal competition and substitution. HSR versus air transport: Understanding the socio-economic determinants of modal choice. doi: 10.1016/j.retrec.2020.100823

Birolini, S., Cattaneo, M., Malighetti, P., & Morlotti, C. (2020). Integrated origin-based demand modeling for air transportation. *Transportation Research Part E: Logistics and Transportation Review*, *142*, 102050. doi: 10.1016/j.tre.2020.102050

Guevara, C., & Navarro, P. (2015). Detection of weak instruments when correcting for endogeneity in binary logit models. In *Proceedings 14th international conference on travel behaviour research (iatbr), windsor, uk.*

Guevara, C. A. (2018). Overidentification tests for the exogeneity of instruments in discrete choice models. *Transportation Research Part B: Methodological*, *114*, 241–253. doi: 10.1016/j.trb.2018.05.020

Guevara, C. A., & Ben-Akiva, M. E. (2012). Change of scale and forecasting with the control-function method in logit models. *Transportation Science*, *46*(3), 425–437. doi: 10.1287/trsc.1110.0404

Hong, S.-J., & Najmi, H. (2022, July). Impact of High-speed rail on air travel demand between Dallas and Houston applying Monte Carlo simulation. *Journal of Air Transport Management*, *102*, 102222. Retrieved 2023-10-27, from `https://www.sciencedirect.com/science/article/pii/S0969699722000436` doi: 10.1016/j.jairtraman.2022.102222

Li, H., Wang, K., Yu, K., & Zhang, A. (2020, September). Are conventional train passengers underserved after entry of high-speed rail?-Evidence from Chinese intercity markets. *Transport Policy*, *95*, 1–9. doi: 10.1016/j.tranpol.2020.05.017

Lurkin, V., Garrow, L. A., Higgins, M. J., Newman, J. P., & Schyns, M. (2017, June). Accounting for price endogeneity in airline itinerary choice models: An application to Continental U.S. markets. *Transportation Research Part A: Policy and Practice*, *100*, 228–246. doi: 10.1016/j.tra.2017.04.007

Petrin, A., & Train, K. (2003, January). *Omitted product attributes in discrete choice models* (Working Paper No. 9452). National Bureau of Economic Research. doi: 10.3386/w9452

Rivers, D., & Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, *39*(3), 347–366. doi: 10.1016/0304-4076(88)90063-2

Zhang, A., Wan, Y., & Yang, H. (2019, September). Impacts of high-speed rail on airlines, airports and regional economies: A survey of recent research. *Transport Policy*, *81*, A1–A19. (Publisher: Pergamon) doi: 10.1016/J.TRANPOL.2019.06.010