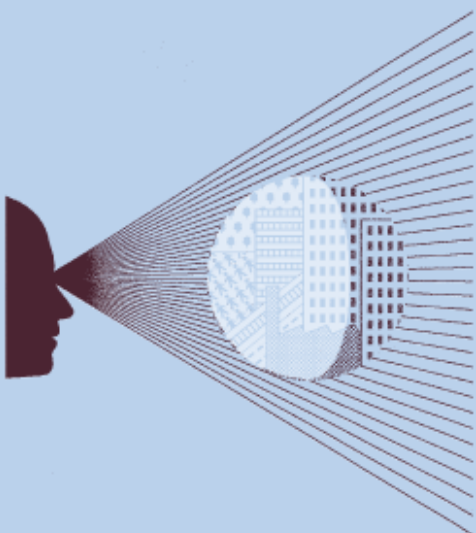


How has the preferred econometric model been derived?

Econometric approach report

**Prepared for
the Department for Transport,
Transport Scotland, and
the Passenger Demand Forecasting Council**

March 2010



Oxera Consulting Ltd is registered in England No. 2589629 and in Belgium No. 0883.432.547. Registered offices at Park Central, 40/41 Park End Street, Oxford, OX1 1JD, UK, and Stephanie Square Centre, Avenue Louise 65, Box 11, 1050 Brussels, Belgium. Although every effort has been made to ensure the accuracy of the material and the integrity of the analysis presented herein, the Company accepts no liability for any actions taken on the basis of its contents.

Oxera Consulting Ltd is not licensed in the conduct of investment business as defined in the Financial Services and Markets Act 2000. Anyone considering a specific investment should consult their own broker or other investment adviser. The Company accepts no liability for any specific investment decision, which must be at the investor's own risk.

© Oxera, 2010. All rights reserved. Except for the quotation of short passages for the purposes of criticism or review, no part may be used or reproduced without permission.

Contents

1	Introduction	1
2	Econometric techniques	3
3	Initial modelling	10
3.1	Dynamics	11
3.2	Choosing between ARDL and ECM	13
3.3	Dynamic panel data models	13
3.4	Diagnostic testing	16
3.5	Comparison of filters	17
3.6	Representation of car ownership variable	18
4	Conclusions	20
A1	Elasticity derivations	21
A1.1	Constant elasticities	21
A1.2	Variable elasticities	21
A1.3	Squared terms	22

List of tables

Table 2.1	Summary of econometric methods	9
Table 3.1	Fixed-effects models for season tickets	12
Table 3.2	Dynamic panel models for season tickets	15
Table 3.3	Static fixed-effects panel models for season tickets	18
Table 3.4	Comparison of different representations of car ownership variable	19

List of figures

Figure 2.1	Approach selection	3
Figure 4.1	Econometric procedure selection	20

1 Introduction

Oxera and Arup have undertaken a study, 'Revisiting the Elasticity-Based Framework', by the Department for Transport (DfT), Transport Scotland and the Passenger Demand Forecasting Council (PDFC). The primary aim of the study is to update and estimate the fares and background growth elasticities contained within the Passenger Demand Forecasting Handbook (PDFH).

The study has a number of secondary objectives, which include:¹

- exploring the use of innovative or alternative econometric techniques;
- re-specifying and extending the core elasticity-based framework;
- improving the underlying data.

The purpose of this report is to describe the econometric techniques that are available for use in the study, together with the preliminary modelling that was undertaken to arrive at the preferred econometric approach.

As part of this study, a number of reports have been produced, detailed below, which form key elements in the formulation of the overall final forecasting framework, and are referenced a number of times here.

Reports prepared by Oxera and Arup for the 'Revisiting the Elasticity-Based Framework' study:

- 'What are the findings from the econometric analysis?' (the *Findings* report)
 - 'Is the data capable of meeting the study objectives?' (the *Data capability* report)
 - 'How has the preferred econometric model been derived?' (the *Econometric approach* report)
 - 'What are the key issue for model specification?' (the *Model specification* report)
 - 'How has the market for rail passenger demand been segmented?' (the *Market segmentation* report)
 - 'Does quality of service affect demand?' (the *Service quality* report)
- 'How should the revised elasticity-based forecasting framework be implemented?' (the *Guidance* report)

The Invitation to Tender for this study requires innovative or alternative econometric techniques to be investigated for use in the study. As this report demonstrates, Oxera has considered a wide range of econometric techniques, some of which are at the forefront of the econometric literature. However, the preferred technique has been derived so that the results of the econometric analysis are as robust as possible. No preference has been given to new techniques over established ones solely because they are new. Oxera has discussed the different econometric approaches with its academic advisers, Professors Anindya Banerjee and Subal Khumbakar, with further input from Professors Manuel Arellano and Stephen Bond.

As a technical report, this document focuses on alternative econometric estimators. Each section begins with a non-technical summary and, where appropriate, a non-technical explanation is given alongside the more technical details.

¹ Department for Transport (2008), 'Rail Passenger Demand Forecasting: Revisiting the Elasticity-Based Framework Request for Proposal and Statement of Requirement', July, pp. 12–13.

The report is structured as follows:

- section 2 discusses possible econometric approaches;
- section 3 details the initial modelling undertaken to arrive at the preferred econometric estimator, together with diagnostic tests and input from Oxera's academic advisers;
- section 4 concludes with a discussion of the preferred econometric approach and the economic models that have subsequently been used to derive the forecasting framework.
- Appendix 1 provides details of how elasticities can be derived from the estimated models.

2 Econometric techniques

The study team has considered a wide range of econometric models and estimators, from the relatively straightforward to the more complex.

Econometric techniques were shortlisted from an initial sift. Some techniques (eg, almost ideal demand systems and structural time series) were quickly discounted for practical or theoretical reasons. Other techniques were considered further. The next section describes the preliminary analysis undertaken to arrive at a preferred model and estimator.

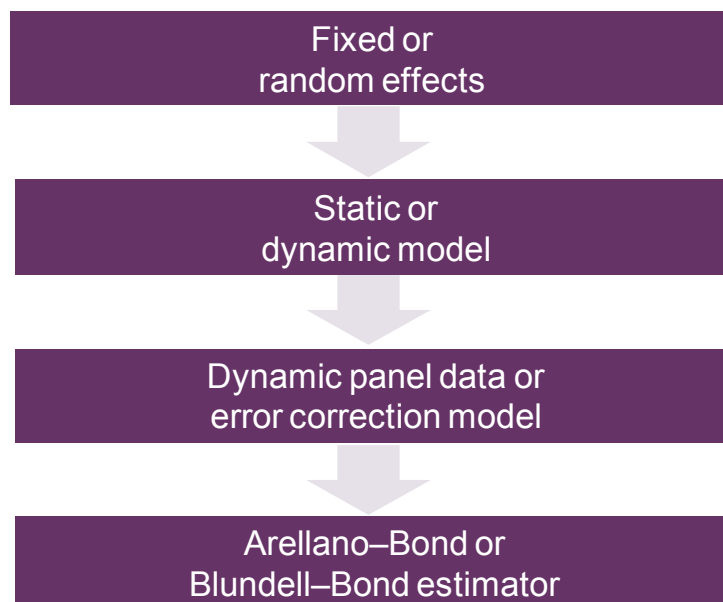
There are a number of possible approaches which the study team has analysed. These range from straightforward (static pooled ordinary least squares, OLS) to more complex dynamic panel models.

A number of decisions have been taken in order to arrive at the preferred econometric estimator:

- whether to use a pooled estimator,² or effects-based models which are designed to take into account cross-sectional variation/heterogeneity;
- whether the preferred model is dynamic (ie, whether it includes lagged adjustments);
- if the preferred model is dynamic, what the preferred estimator is;
- whether the estimates can be used in the forecasting framework;
- whether the data is capable of estimating such a model.

Figure 2.1 illustrates the decision process used to arrive at the preferred estimator.

Figure 2.1 Approach selection



Source: Oxera.

Two main, separate, decisions have had to be made: which static/dynamic estimation method to use; and whether to use an effects-based model. These decisions are discussed in more detail in section 3.

² For more details, see the *Model specification* report.

The section below describes the models and estimators which are available for use in this study, together with their advantages and disadvantages. The suitable methodologies are then considered further in section 3.

A non-technical description of each method is given in italics at the start of each sub-section.

2.1.1 Auto-regressive distributed lag models

Description of method

In an auto-regressive distributed lag (ARDL) model, the variable of interest is assumed to be a function of the past values of itself (auto-regressive) and the current and past values of other variables (distributed lag). They can relatively easily be extended to incorporate panel data.

ARDL models can accommodate a variety of lag structures and include well-known models such as static regressions as special cases.

The general form is:

$$y_t = \mu + \sum_{k=1}^p \rho_k y_{t-k} + \sum_{j=0}^q \beta_j x_{t-j} + \varepsilon_t$$

where y is the dependent variable and x is an explanatory variable. The static model is the case where $p = 0$ and $q = 0$.

Advantages

This type of model can accommodate very general lag structures and can easily be extended to incorporate panel data.

Disadvantages

Models of this type are likely to have difficulties in successfully identifying the 'correct' relationships between the variables in data which contain a unit root (see Box 2.1 below), as issues of spurious correlation may arise. As also noted in Box 2.1, economic time series often display evidence of a unit root and therefore it is important to consider this issue. However, the existence, or otherwise, of a unit root needs to be analysed on a case-by-case basis. Should a unit root be found in the data, the issue can be resolved by modelling in differences (ie, using the difference between two years of data as the dependent variable in the regression).

One criticism that has been levelled at ARDL models is that, if there is a stochastic (random) trend present in the data, the dynamics in an ARDL model will be approximating this trend rather than modelling 'real' dynamics.³ However, if there is not a stochastic trend in the data, this criticism is not valid. The presence, or otherwise, of such a stochastic trend is an empirical issue and is difficult to identify given the small number of time-series observations in the dataset available for this study. (There are a maximum of 18 time-series observations available in the dataset.)

³ See, for example, MVA Consultancy (2008), 'Robust Foundations: Econometric Analysis of Long Time Series Rail Passenger Demand Aggregates', November, Appendix C.

Box 2.1 Unit roots

A unit root is defined as where the value of a variable is equal to its previous level plus or minus a random error. Unit roots are often present in data which trends over time. More technically:

$$y_t = y_{t-1} + \varepsilon_t$$

Where y_t is the variable of interest at time t and ε_t is a random error term.

This can be a problem: if there are two variables which are generated by such a process, they can appear to be closely related (without an actual relationship existing between them); this is known as spurious correlation.⁴

The problems associated with unit roots can be avoided by:

- using cointegration techniques (such as the error correction model, ECM—see below);
- modelling the variable in differences, which would result in the loss of one year of data for each cross-sectional observation.

Many approaches have been developed to test for unit roots in time-series data—for example, the Dickey–Fuller, KPSS, and Phillip–Perron tests.⁵ However, testing for unit roots in panel data is less developed.⁶

The ARDL models discussed in this section implicitly pool the data (note that the intercept does not vary across cross-sectional units). Pooling the data constrains all the parameters in a model (including the intercept) to be the same across all cross-sectional units. (For more discussion on the assumptions and limitations of pooling data, and when this approach is valid, see the *Model specification* report).

Data requirements

This type of model requires no data beyond the standard requirements for model estimation. The data requirements for an econometric model are set out in the *Model specification* report.

Feasibility

This type of model is straightforward to estimate, and standard diagnostic tests can be used to identify errors in specification. However, the model may need to be enhanced to an ECM (see below), or estimated in differences to adequately incorporate data which may contain a unit root.

2.1.2 Error correction models

Description of method

When data series move together over time (as is common in economic variables, such as demand, income, etc), standard statistical techniques such as OLS may find a spurious relationship between the variables. To counteract this, an ECM identifies a long-run relationship between the variables, while allowing for short-run deviations from this relationship. As with the ARDL model above, ECMs can be extended in a relatively straightforward way to allow for panel data.

Time-series data often contains a unit root (see Box 2.1). An ECM allows for this by identifying a long-run relationship between variables such as demand and income, often based on economic theory, but allowing for short-term deviations from this long-run relationship.

⁴ This is the subject of a developed econometric literature. See, for example, Nelson, C.R. and Plosser, C.I. (1982), 'Trends and Random Walks in Macroeconomic Time Series', *Journal of Monetary Economics*, **10**, pp. 139–62.

⁵ Greene, W. (2003), *Econometric Analysis*, New Jersey: Prentice Hall.

⁶ See, for example, Baltagi, B. (2005), *Econometric Analysis of Panel Data*, Chichester: John Wiley & Sons, Ltd, pp. 237–47.

In the equation below, there is a long-run relationship between the variables y and w , which both contain a unit root (by assumption), but the short-run relationship is affected by w and another variable, x , which does not contain a unit root.⁷

$$\Delta y_t = \sum_{k=1}^p \rho_k \Delta y_{t-k} + \sum_{j=0}^q \beta_j x_{t-j} + \sum_{l=0}^r \gamma_l \Delta w_{t-l} + \lambda(y_{t-1} - \theta w_{t-1}) + \varepsilon_t$$

where ρ , β , γ , and λ are parameters to be estimated and ε_t is a random error term.

Advantages

The model is flexible and provides both short- and long-run elasticities, in addition to being consistent in dealing with data which contains a unit root.

Disadvantages

These models can be relatively complex to specify due to the two-stage conceptual framework which requires the long-run relationship to be identified before the short-run dynamics are determined.

Data requirements

This type of model requires no data beyond the standard requirements for model estimation, although a longer time series may be required to identify the lag structure.

Feasibility

This type of model has similar data requirements to that of dynamic panel effects-based models, such as those set out below. However, using an ECM (and hence modelling in differences) results in the loss of one degree of freedom per flow.

2.1.3 Dynamic panel effects-based models

Description of method

Panel data often contains data with a substantial degree of heterogeneity in both levels and responses to changes in explanatory variables. For example, the data in this study has data on demand between London and Manchester, which had approximately 1.1m journeys in 2007, and Hillfoot and Partick (two stations in Scotland), which had approximately 11,000 journeys in 2007. An effects-based model compensates for the differences in the levels of variable by allowing each flow to have a separate intercept. Effects-based models allow for some heterogeneity in levels and omitted factors,⁸ and the market segmentation accounts for differences in responses. Hence, the elasticities are assumed to be the same across all flows within each model.

There are two main types of panel effects models: fixed and random effects. Both models contain an intercept which varies between cross-sectional units (flows), but are constant over time. However, fixed-effects models contain a term which is fixed but differs across cross-sectional units—ie, flows (μ_i)—while, for a random-effects model, the intercept is assumed to be the same in expectation, but varies according to a random variable (u_i). The key distinction between the two types of model is whether or not the unobserved effects (which the effects, fixed or random, are allowing for) are correlated with the observed effects.⁹ For example, if the cost of bus and coach travel (which is unobserved) is correlated with the variables included in the model (fare, car cost, etc) then fixed effects may be more appropriate than random effects.

⁷ This example is given with two explanatory variables, but the principle generalises to any number of explanatory variables.

⁸ It is not possible to determine whether the heterogeneity is due to omitted factors or levels since they are, to a large extent, indistinguishable. For example, it may be that past omitted factors have caused the difference in levels.

⁹ For more details, see Greene, W. (2003), *Econometric Analysis*, New Jersey: Prentice Hall, p. 285.

The estimated elasticities are constrained to be the same across all flows within a market segment. This constraint is imposed because the market segmentation aims to group together flows with similar behavioural responses.

Fixed effects:

$$y_{it} = \sum_{k=1}^p \rho_k y_{i,t-k} + \sum_{j=0}^q \beta_j x_{i,t-j} + \mu_i + \varepsilon_t$$

Random effects:

$$y_{it} = \sum_{k=1}^p \rho_k y_{i,t-k} + \sum_{j=0}^q \beta_j x_{i,t-j} + (\mu + u_i) + \varepsilon_t$$

where ρ and β are parameters to be estimated and ε_t is a random error term. It is possible to use a Hausman test to determine whether the data supports fixed or random effects (see Box 3.1). Note that the Hausman test may not be defined in small samples.

Advantages

These models allow for heterogeneity in the data, which provides for more accurate identification of the elasticities of interest. Models of this type can also estimate a wide range of lag structures.

Disadvantages

These models (dynamic panel models specified as an ARDL model) may not be adequate to deal with variables which contain unit roots, as is often the case in time-series and panel data models (see Box 2.1). Where this is the case, these models will need to be enhanced to an ECM or other approach (such as modelling the differences in the variables).

Data requirements

These types of models require the dependent variable to be in a panel data format, as per the data for this study. The structure of the explanatory variables is less important since they can usually be linked to the dependent variable by making assumptions about how the variable differs across the panel, although explanatory variables which match the level of aggregation of the dependent variable are preferred. The assumptions required are discussed in the *Model specification* report.

Feasibility

The issue with this class of models is not whether it is feasible to estimate them (given the data available for this study), but whether they are the most appropriate means of estimating the elasticities to be used in a forecasting framework (ie, whether, given the dataset available for use in this study, this class of models is optimal). This is an empirical question, and is considered in more detail in section 3.¹⁰

2.1.4

Others

A number of other types of econometric models were discounted after an initial sift, for a number of reasons, including practicality and the ability to produce meaningful long-term forecasts.

The two models considered in this sub-section are structural time series and almost ideal demand systems models.

¹⁰ It is worth noting that Wardman and Dargay (2007) recommended the use of a fixed-effects panel specification in preference to the ratio approach currently used in the PDFH. See Wardman, M. and Dargay, J. (2007), 'External Factors Data Extension and Modelling', Institute for Transport Studies, Leeds University, March, p. 7.

Structural time series

Structural time series models are very flexible and allow for elasticities, together with a trend, which can change over time.

More technically, these models allow for varying coefficients and a stochastic slope, and as such provide a very flexible way to model time-series data. This methodology requires a long time dimension to the data in order to identify the dynamics. The time dimension of the dataset available for this study is relatively short, and for this reason the study team has not used this methodology.

Almost ideal demand systems

The rationale underlying this approach is that consumers allocate expenditure first among different areas of expenditure (ie, food, leisure, transport, etc), subsequently to mode of transport (ie, car, bus, rail, etc), and finally to ticket type (ie, standard class full, standard class reduced, first class full, etc). This sequential allocation enables restrictions on choices from consumer theory to be imposed on the model in a way which is difficult to do for more standard approaches.

This is a flexible methodology used to estimate elasticities. The model allocates consumer expenditure among different baskets of goods, where a consumer's expenditure share for a particular good or service (i) is defined as:

$$w_i = \frac{p_i q_i}{m}$$

where p_i is the price paid for good i, q_i is the quantity of good i purchased, and m is the total expenditure on all goods in the demand system.

$$w_i = a_i + \sum_{j=1}^k \gamma_{ij} p_j + \beta_i \ln(m/P)$$

Here, the expenditure shares of each product (or categories of product) are related to prices and total real expenditure on that product category in a log-linear way, where p_j is the price of the j th item in the category, m is total household expenditure allocated to that product category, and P is the aggregate price index of the product category.

The study team has not estimated an almost ideal demand systems model for this study in view of the prohibitive data requirements for estimating such a model at the required disaggregate level.

2.1.5

Summary

This section has reviewed several possible econometric techniques for use in this study. Table 2.1 summarises the methods reviewed and their theoretical suitability and feasibility for use within this study.

Table 2.1 Summary of econometric methods

	Suitable for use in the study	Lag structure	Variable elasticities	Generalised cost/time	Monetised/ non-monetised variables	Income variable interaction
ARDL	✓	✓	✓	✓	✓	✓
ECM	✓	✓	✓	✓	✓	✓
Panel effects	✓	✓	✓	✓	✓	✓
Almost ideal demand systems	x	x	✓	✓	✓	✓
Structural time series	x	✓	✓	✓	✓	✓

Source: Oxera.

As can be seen from the table, a number of techniques may have been useful for estimating the models in this study. Almost ideal demand systems and structural time series were not investigated further because of the practical considerations set out above. The remaining three suitable techniques all have strengths and weaknesses, and the preferred option has been determined both by the data as well as theoretical considerations. It is possible to use combinations of these techniques; for example, ARDL and effects-based models, or effects-based and an ECM. These combinations have also been considered.

The next section considers the initial modelling undertaken in order to decide which option, or combination of options (ARDL, ECM, or panel effects), was preferred.

3 Initial modelling

Following the identification of the possible econometric models/estimators in section 2, the preliminary modelling undertaken to arrive at a preferred econometric model and estimator is described in this section. This modelling also investigated how the car ownership variable should be specified and whether the data cleaning process makes an undue difference to the estimated elasticities.

Preliminary modelling was undertaken at a national level for season and non-season tickets.

The use of effects-based models is important to allow for the heterogeneity in the dataset. The choice between fixed and random effects was based on extensive statistical analysis, which suggested that fixed effects were preferable.

The preliminary analysis revealed that static models result in unreasonably large parameter estimates and very high t-statistics, and hence a dynamic specification is preferred. There are two possible alternatives for dynamic models: an ECM or an ARDL model. The difference lies in whether the long-run relationships are modelled explicitly (in an ECM) or implicitly (in an ARDL model).

Following advice from Oxera's academic advisers, an ARDL model was chosen for use in this study due to the relatively small number of time-series observations available in the dataset, which is likely to make identification of explicit long-run elasticities challenging.

This leads to the final decision required in order to arrive at the preferred econometric estimator: which of the two common dynamic panel data estimators is preferred? On the basis of its theoretical advantages, the Blundell–Bond estimator has been chosen in favour of the Arellano–Bond estimator.

A number of specifications for the car ownership variable were tested; however, due to the theoretical properties and empirical performance, the proportion of households without access to a car has been retained as the preferred specification for this study. This is the same form as is currently in the PDFH.¹¹

The preliminary modelling also suggests that the data cleaning process has not had an unduly large effect on the estimated elasticities.

This section discusses the initial modelling undertaken for the study, based on the diagnostics available for use with the preferred model, and the input received by Oxera from a number of prominent academics.

The aims of the initial modelling were to investigate:

- the optimal formulation of the car ownership variable;
- whether the data cleaning process applied to the data makes an undue difference to the estimated elasticities;¹²
- the preferred estimator.

During the preliminary modelling, the dataset was split into season and non-season tickets and national models were estimated. However, it is important to note that the final modelling has been conducted on full fare, reduced fare and season tickets. Therefore, the results presented in this section are *for descriptive purposes only and should not be interpreted as robust models*.

The next section discusses static and dynamic models, and the impact of different filters using national season ticket data.

¹¹ The elasticity for this variable will alter depending on the value of the proportion of households without access to a car, as it enters the model in a log-linear formulation.

¹² The data cleaning process is detailed in the *Data capability* report.

3.1 Dynamics

One of the most important issues investigated during the preliminary modelling phase is that of dynamics—ie, whether a model is required that allows for adjustments to changes in the explanatory variables after the current period.

A static model assumes that all effects on the dependent variable from changes in the explanatory variables are completed within the same period as the change occurs (ie, within one year for this dataset).

Dynamic models include lagged values of the dependent variable as an explanatory variable in the regression equation, together with lagged values of the explanatory variables.

3.1.1 Initial model results

The choice of explanatory variables was based on existing industry practice (such as the PDFH) and insights from the 'Rail Trends Report'.¹³ Season tickets are typically purchased by commuters and therefore employment was included as a key explanatory variable, while for non-season tickets, income is included. In line with previous studies, demographic variables, car journey time (CJT), car ownership, cost of other means of transport and generalised journey time (GJT) were included as drivers of rail demand. In addition, an index of service quality has been included as an explanatory variable (see the *Service quality* report). This report focuses on elasticities with respect to fares, socio-economic/demographic factors, and other modes.

As explained in more detail in section 2.1.3, an important consideration in the estimation of panel data models is whether to use fixed or random effects, or no panel effects. Comparisons were made between random and fixed-effects models and, for most of the models, random effects were rejected in favour of fixed effects, based on the Hausman test statistic (see Box 3.1 for more details). This is in addition to the theoretical rationale for using fixed effects in preference to random effects; namely that the unobserved variables are likely to be correlated with the observed variables and, therefore, that the random-effects models are likely to be inconsistent (ie, they will not produce accurate estimates of the 'true' elasticity). Accordingly, the modelling proceeded with fixed-effects models, some of which are reported in Table 3.1 below.¹⁴

Box 3.1 Hausman test

The Hausman test can be used to test whether fixed or random effects are preferred. In essence, it tests whether it is valid to assume that the unobserved effects are uncorrelated with the observed variables.

The basis of the test is that, if the unobserved effects are correlated with the observed effects, the random-effects estimator is inconsistent, but the fixed-effects estimator is not. However, if the unobserved effects are not correlated with the observed effects then the fixed-effects estimator is still consistent, while the random-effects model is both consistent and efficient.

The calculation of the test statistic is described in a number of textbooks, and so is not repeated here.¹⁵

¹³ Ove Arup & Partners Ltd (2009), 'Rail Trends Report', March 30th.

¹⁴ See Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: The MIT Press, pp. 288–291.

¹⁵ Ibid.

Table 3.1 Fixed-effects models for season tickets

Dependent variable: log journey	Model 1 (static)	Model 2 (static)	Model 3 (dynamic)	Model 4 (dynamic)
Log fare	-1.122 (-91.43)	-1.161 (-109.34)	-1.122 (-91.43)	-0.725 (-53.55)
Log total jobs at destination	0.682 (6.95)	0.851 (7.87)	-0.0555 (-0.38)	0.272 (4.16)
Log working age population at origin	6.261 (22.05)	5.989 (31.00)	0.909 (2.57)	1.296 (5.66)
Proportion of households without car	-0.0447 (-10.53)	0.0544 (21.43)	0.0221 (2.84)	0.00526 (1.12)
Car journey time	0.900 (1.22)	9.429 (9.56)	-1.458 (-2.29)	-1.658 (-2.17)
Log fuel price	1.141 (11.61)	3.435 (48.81)	0.366 (4.96)	3.435 (48.81)
Log GJT	-0.821 (-13.66)	-0.283 (-4.46)	0.0582 (0.81)	-0.0420 (-0.68)
Service quality index	0.441 (9.68)		0.152 (4.01)	0.0607 (1.99)
Strike		-0.613 (-42.08)	0.485 (17.64)	0.601 (88.51)
Hatfield derailment	-0.0780 (-5.72)	-0.360 (-21.39)	0.00963 (0.98)	-0.360 (-21.39)
Constant	-54.77 (-22.46)	-75.49 (-40.91)	-4.086 (-1.38)	-10.99 (-5.30)
Number of observations	91,651	170,178	76,816	91,694
Adjusted R ²	0.026	0.089		

Note: t-statistics in parentheses. The dynamic models are ARDL specifications.
Source: Oxera.

The signs of most of the parameter estimates are in line with economic theory, and for the dynamic models, tests for auto-correlation revealed that the basic conditions to apply the estimators were satisfied (see section 3.4).

The results of the preliminary estimation revealed the following issues which required further investigation:

- high t-statistics in the static models may suggest a spurious regression (see Box 2.1);
- some of the estimated elasticities are very high in absolute value for the static models (eg, the working age population elasticity in Models 1 and 2);
- the aggregation of the data into season and non-season tickets has combined a large volume of data and there could be several segments in the market. This issue is discussed in the *Market segmentation* report. The results reported in Table 3.1 are from the preliminary modelling and should not be relied upon (see the *Findings* report for the final results);
- the correlation between fare, GJT, CJT and car cost introduces multicollinearity into the model. This is likely to be because all of these variables are positively related to

distance. However, in the final models, multicollinearity is less of a problem due to the large amount of cross-sectional variation.

The greatest concerns prompted by these conclusions are the potentially spurious regression and the very high elasticity estimates which occur in the static models. Therefore, a dynamic model specification is preferred to a static specification. A number of possible dynamic panel data models have been considered, and these are addressed in the next section.

3.2 Choosing between ARDL and ECM

There are two main classes of dynamic panel data estimators: ARDL models, or ECMs. The relatively short time series of the dataset available for use in this study (a maximum of 18 annual observations) may prevent a single pooled model being estimated for each market segment. Panel data can improve the difficulties associated with the short time series, but introduces other complexities into the analysis. Both ARDL and ECMs can be constructed as time-series or panel data approaches, and therefore the choices between ARDL and ECM, and effects-based or not, are separate, as indicated by Figure 4.1 in section 4.

Following consultation with Professor Banerjee, Oxera has determined that the time-series element of the dataset available for this study may pose problems for identifying the explicit long-run dynamics required to estimate an ECM. In light of this, ARDL-type dynamic panel data models, which have been developed specifically for datasets with a small time dimension, have been used.

A dynamic model can be achieved by using a lagged dependent variable as an explanatory variable. However, since the lagged dependent variable is endogenous, it is correlated with the unobserved panel effects by construction, which makes standard fixed- or random-effects estimators inconsistent. Therefore, more advanced dynamic panel data estimators are required to estimate consistent elasticities.

The issues highlighted in sections 2.1.1 and 2.1.3 relating to unit roots have been considered carefully, and the analysis has included testing for these issues (see section 3.4 for more detail).

Two main estimators are available to use in the case of large N , small T datasets (such as that currently available for use in this study). The two techniques are known as Arellano & Bond, and Blundell & Bond, following the papers published by these authors.¹⁶ The next section details the analysis that has been undertaken to determine which of these estimation techniques is preferred.

3.3 Dynamic panel data models

The models set out below have been developed specifically for datasets with a large cross-sectional component (N) and a small time-series component (T). The difference between the two lies in the assumptions which need to be made to use the estimators. The Blundell–Bond estimator requires more assumptions than the Arellano–Bond estimator, but is more efficient and is still consistent when the data is highly dependent on previous data.

3.3.1 Process

Initial models were computed using the estimation techniques outlined by Arellano & Bond and Blundell & Bond. Arellano & Bond suggest an estimator constructed to deal with the

¹⁶ Arellano, M. and Bond, S. (1991), 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies*, **58**, pp. 277–297. See also Blundell, R. and Bond, S. (1998), 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics*, **87**, pp. 115–143.

endogenous nature of the lagged dependent variable.¹⁷ This approach is broadly similar to standard instrumental variables (IV) estimators, with different variables being used as instruments (or proxies) for the lagged dependent variable.¹⁸ The main requirement for applying these estimators is that there should be no serial correlation in the residuals (see section 3.4 for more detail); otherwise the instrumentation is not valid and hence the estimator is inconsistent. The Arellano–Bond estimator can perform poorly if the coefficient on the lagged dependent variable is large—ie, if the time series is close to containing a unit root.

Blundell & Bond propose a system of equations/GMM estimator that accounts for this problem. The additional assumption which needs to be made to use the Blundell–Bond estimator is that the initial observations must have a very small impact on the observed observations. Given the length of time since railways were first introduced in Great Britain, this assumption would appear to hold for the dataset available for this study. Note that there are currently no formal tests of this assumption available in the literature.

3.3.2 Initial results

Both Model 1 and Model 2 have similar specifications: the instruments include all possible lags of the dependent variable and one lag of the other explanatory variables. The difference lies in the estimation technique. Model 1 has been estimated using Arellano–Bond estimation techniques and Model 2 has been computed using Blundell–Bond. The results differ between the two models, indicating the importance of selecting the most appropriate estimation technique.

¹⁷ The approach is a generalised method of moments (GMM) estimator, which uses estimates of the sample mean, variance, etc, to estimate the parameters of interest.

¹⁸ An instrument is a variable which is correlated with the endogenous variable but is not correlated with the error term.

Table 3.2 Dynamic panel models for season tickets

Dependent variable: log journey	Model 1	Model 2
Log fare	-1.122 (-91.43)	-0.725 (-53.55)
Log total jobs at destination	-0.0555 (-0.38)	0.272 (4.16)
Log working age population at origin	0.909 (2.57)	1.296 (5.66)
Proportion of households without car	0.0221 (2.84)	0.00526 (1.12)
Car journey time	-1.458 (-2.29)	-1.658 (-2.17)
Log fuel price	0.366 (4.96)	3.435 (48.81)
Log GJT	0.0582 (0.81)	-0.0420 (-0.68)
Service quality index	0.152 (4.01)	0.0607 (1.99)
Lag of journey	0.485 (17.64)	0.601 (88.51)
Hatfield derailment	0.00963 (0.98)	-0.360 (-21.39)
Constant	-4.086 (-1.38)	-10.99 (-5.30)
Number of observations	76,816	91,694

Note: t-statistics in parentheses.
Source: Oxera.

Tests for auto-correlation revealed that the basic conditions to apply the estimators were satisfied.

As expected, the lag of journey is positive and significant. The Blundell–Bond model (Model 2) is more in keeping with the predictions of economic theory—eg, with a positive (and statistically significant) coefficient on total jobs at destination.

The Blundell–Bond estimator has a number of theoretical advantages over the Arellano–Bond estimator. It was developed to estimate models using data where the autoregressive coefficient is close to one (ie, the data is strongly determined by the value in the previous period).¹⁹ Moreover, the Blundell–Bond estimator is also more efficient.²⁰ **Therefore, the Blundell–Bond estimator is the preferred econometric technique for use in this study.**

The next section considers the diagnostic tests available to test the assumptions and conditions underlying the chosen modelling technique.

¹⁹ Baltagi, B.H. (2005), *Econometric Analysis of Panel Data*, Chichester: John Wiley & Sons Ltd, p. 148.

²⁰ *ibid.*

3.4 Diagnostic testing

Diagnostic testing in panel data is not as advanced as diagnostic testing in cross-sectional or time-series analysis. However, there are a number of diagnostic tests which can be used, including tests for:

- autocorrelation;
- cointegration;
- instrument validity.

Box 3.1 Academic advice

The study team has corresponded with Professors Arellano & Bond, in addition to the Oxera study team advisers, Professors Banerjee and Khumbakar.

Input from Professors Bond and Arellano focused on diagnostic testing of the models—in particular, whether there were diagnostic tests available beyond the autocorrelation and moment validity tests (detailed below), and how to express the fit of the model to the data.

The academics' input suggests that there are very few tests available, other than those below, possibly because when the cross-sectional element of the data is large (as is the case for this study), standard errors can be estimated which are robust (asymptotically) to heteroscedasticity and non-normality (ie, despite the presence of potentially non-normal and heteroscedastic errors, it is possible to estimate standard errors that are not affected by these factors).

There does not appear to be a consensus on the optimal measure of model fit to report. It was suggested that the squared correlation coefficient between actual and fitted results be used since this is bounded by zero and one, with a result closer to one corresponding to a better 'fit'.

The sections below discuss these points further.

3.4.1 Autocorrelation

A test for autocorrelation in dynamic panel data was developed by Arellano & Bond. Both the Arellano–Bond and the Blundell–Bond estimators include an assumption that there is no residual autocorrelation in the model.

The Arellano–Bond test for autocorrelation is a statistical test for correlation in the first-differenced errors. There will be autocorrelation in the first differences because of the construction of the model, and hence it is to be expected that the null hypothesis of no autocorrelation will be rejected for the first differences. However, rejection of the null hypothesis of no autocorrelation at higher orders (eg, at second or third differences) implies that the moment conditions are not valid and hence that the estimator is not valid.

3.4.2 Cointegration

There is a developed literature on testing for cointegration in panel data.²¹ However, one of the challenges with testing the stationarity of residuals in panel data is the interpretation of a rejection of the null hypothesis that 'all cross-sectional units are stationary'. This implies that at least one of the cross-sectional units is non-stationary. If the cross-sectional units in question are the residuals from a regression, this might imply that at least one of the cross-sectional units has a spurious regression with the others. In practice, this means that the testing for spurious relationships between variables in a panel dataset is not straightforward.

It is important to note that the power of these tests is also expected to be weak in this current study due to the low number of time-series observations.

²¹ See, for example, the survey in Mátyás, L. and Sevestre, P. (2008), *The Econometrics of Panel Data: Fundamentals and Recent Development in Theory and Practice*, Berlin: Springer, pp. 279–322.

3.4.3 Moment validity

The estimators mentioned above are only valid if the moment conditions are valid—ie, the instruments used in the estimation fulfil the assumptions of being uncorrelated with the error term and correlated with the lagged dependent variable. However, there are no direct tests for this assumption. A second-best alternative is to test whether the instruments are valid using the Sargan test.²² As stated above, the validity of the instruments is important for consistent estimates of the parameters of interest. However, this test can be rejected for a number of reasons, and rejection of the null hypothesis does not necessarily imply that the model is misspecified. If the error term is not homoscedastic, the test statistic does not have a standard asymptotic distribution—ie, critical values cannot be tabulated. This does not mean that the model is misspecified, but that the Sargan test will be failed, and hence, under these circumstances, the test is not a useful test of model specification.

3.4.4 Model fit

The fit of the model to the data is an important consideration for all models, but particularly so when using models for forecasting. However, it is important to avoid ‘tweaking’ the models solely to improve the fit to the available data since this may result in models that provide a good fit to the dataset, but do not forecast well.

The approach to measuring model fit in panel data models is still an area of active research, as the R^2 measure from linear models is not appropriate. Oxera has received advice from academic advisers on the optimal way to report the model fit, but there is no consensus on this (see Box 3.1).

On the advice of Oxera’s academic advisers, the squared correlation coefficient between actual and fitted values has been used as the measure of model fit for this study. This provides a number—for each model which is bounded by zero and one—where a result closer to one implies a better match between the actual and fitted results.

The previous section discussed the modelling undertaken to determine the preferred econometric approach and this section has discussed how the models can be tested for possible errors in specification, or breaches of the underlying model assumptions. The next two sections consider whether the data cleaning process is likely to have had an unduly large effect on the parameter estimates, and hence the estimated elasticities, and how the car ownership variable could enter into the analysis.

3.5 Comparison of filters

The sensitivity of the estimated parameters to the data cleaning process was analysed to compare the results of models estimated on a dataset where observations which changed by more than 500% (year on year) were removed and where observations which changed by more than 100% (year on year) were removed. The filtering process is discussed in the *Data capability* report. This analysis also included a small general-to-specific modelling exercise, and the parameter estimates for the preferred models are reported in Table 3.3.

²² See StataCorp (2007), *Longitudinal/Panel Data: Release 10*, College Station Texas, Stata Press, pp. 89–91.

Table 3.3 Static fixed-effects panel models for season tickets

Dependent variable: log journey	Model 1 (500% cut-off)	Model 1A (100% cut-off)	Model 2 (500% cut-off)	Model 2A (100% cut-off)
Hatfield derailment	-0.0659 (-5.35)	-0.0780 (-5.72)	-0.267 (-18.37)	-0.360 (-21.39)
Log fare	-0.869 (-79.26)	-1.122 (-91.43)	-1.004 (-105.52)	-1.161 (-109.34)
Log GJT	-0.928 (-17.11)	-0.821 (-13.66)	-0.534 (-9.22)	-0.283 (-4.46)
Car journey time	1.415 (2.13)	0.900 (1.22)	9.597 (11.28)	9.429 (9.56)
Log fuel price	1.140 (12.75)	1.141 (11.61)	2.610 (41.06)	3.435 (48.81)
Service quality index	0.416 (10.04)	0.441 (9.68)		
Strike			-0.536 (-39.56)	-0.613 (-42.08)
Proportion of households without car	-0.0436 (-11.42)	-0.0447 (-10.53)	0.0371 (15.41)	0.0544 (21.43)
Log total jobs at destination	0.548 (6.23)	0.682 (6.95)	0.636 (6.46)	0.851 (7.87)
Log working age population at origin	7.430 (28.87)	6.261 (22.05)	7.294 (41.97)	5.989 (31.00)
Constant	-63.25 (-28.68)	-54.77 (-22.46)	-79.10 (-48.87)	-75.49 (-40.91)
Number of observations	101,451	91,651	171,646	170,178
Adjusted R ²	0.027	0.026	0.089	0.089

Note: t-statistics in parentheses.
Source: Oxera.

The preferred model specification did not change across the two sensitivities. Models 1A and 2A reported in Table 3.3 correspond to the static models reported in Table 3.1, repeated here to aid comparison of the results. A comparison of the models in Table 3.3 reveals that many of the parameter estimates have similar magnitudes regardless of the data filter applied.²³ This analysis suggests that, for season tickets, the data cleaning process does not unduly alter the parameter estimates (and hence the elasticities that form the focus of this study).

3.6 Representation of car ownership variable

Different functional forms of the car ownership variable were tested to determine which best captured the behavioural response in rail demand to changes in car ownership. In the initial models, natural logs of the car ownership variable (households without a car) were tried. While the estimates from these models are similar to the preferred model, this functional form is not in line with the requirements of the mathematical framework specified in Chapter B1 of

²³ As noted above, these numbers should be considered illustrative since these models are not estimated in the same way as the final models.

the PDFH v5 (p. 7). According to Chapter B1.4 of PDFH v5, car ownership should be included as the proportion of households *without* a car. This specification is desirable because it results in a decreasing marginal effect of an increase in car ownership.

Several different functional forms were tested by using the number of households without cars, as both an absolute and a proportional metric. The coefficient estimates favoured the proportional specification. Table 3.4 illustrates this by comparing the two functional forms for the proposed model. Model 1 includes car ownership as a proportion of total households, while Model 2 includes car ownership in levels. The sign of the estimated coefficient changes between the two models, with Model 2 (estimated using the Blundell–Bond estimator, in levels) producing the expected sign coefficient.

Table 3.4 Comparison of different representations of car ownership variable

Dependent variable: log journey	Model 1	Model 2
Log fare	-1.126 (-91.69)	-1.169 (-109.90)
Employment at destination	1.585 (8.99)	2.239 (13.61)
Log population at origin	8.301 (22.01)	6.200 (21.19)
Proportion of households without car at origin	0.0496 (19.37)	
Number of households without car at origin		-0.0000325 (-9.49)
Car journey time	1.693 (2.30)	10.68 (10.79)
Log fuel price	0.864 (8.59)	3.012 (38.01)
Log GJT	-0.813 (-13.52)	-0.291 (-4.58)
Service quality index	0.426 (9.44)	
Strike		-0.558 (-35.86)
Hatfield derailment	-0.0656 (-4.79)	-0.346 (-20.57)
Constant	-54.77 (-22.46)	-75.49 (-40.91)
Number of observations	91651	170178
Adj. R ²	0.015	0.088

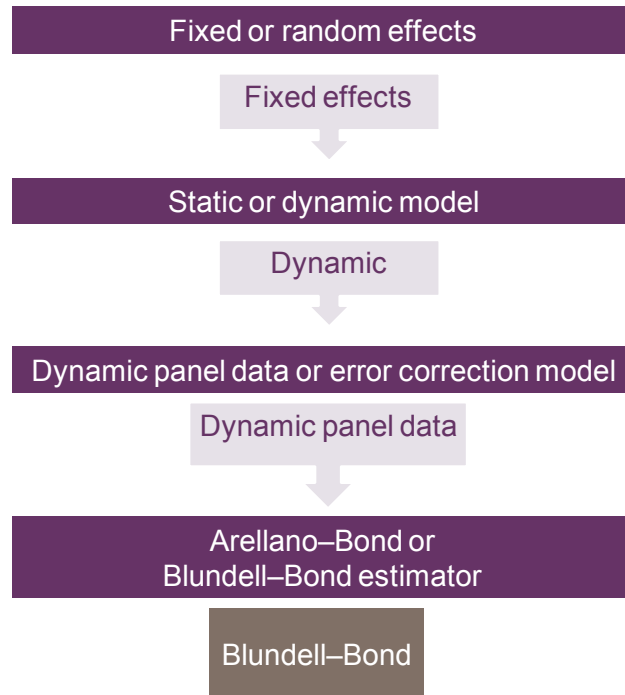
Note: t-statistics in parentheses.
Source: Oxera.

Based on the theoretical advantages of the proportional formulation, the car ownership variable in the further analysis will be the proportion of households without access to a car.

4 Conclusions

This report has outlined how the preferred econometric estimator has been selected. Figure 4.1 illustrates this process, and the decisions made.

Figure 4.1 Econometric procedure selection



Source: Oxera.

The preferred econometric estimator is the Blundell–Bond estimator, which has been derived for the type of modelling that has been undertaken for this study—ie, for datasets with a large cross-sectional dimension and small time dimension.

This has been selected after considering a substantial range of econometric estimators on theoretical grounds, which led to the rejection of some possibilities as being unsuitable for use in this study (eg, structural time series and almost ideal demand systems).

Following this theoretical consideration, preliminary modelling was undertaken to determine which of a number of techniques were most appropriate (ie, whether the preferred model was static or dynamic, etc). This process was detailed in section 3.

The decision to use a dynamic estimator on a dataset with a relatively small time-series component has resulted in the use of an estimator that has been derived for just such ‘large N, small T’ datasets.

The justification and selection of the preferred econometric estimator is a key step in estimating a new elasticities-based forecasting framework. The chosen econometric approach will be used to estimate a number of functional forms for each market segment. The economic model, set out in the *Model specification* report, and the results of the analysis, are presented out in the *Findings* report.

Appendix 1 shows how elasticities can be derived from the dynamic panel data models for the functional forms set out above.

A1 Elasticity derivations

This appendix provides details of how to derive elasticities from the estimated model parameters.

A1.1 Constant elasticities

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 x_{t-2} + \varepsilon_t$$

Where y and x are both in logs and therefore the elasticity of y with respect to x is $\frac{dy_t}{dx_t}$

$$\text{At time } t: \quad \frac{dy_t}{dx_t} = \beta_2$$

$$\begin{aligned} \text{At time } t+1: \quad \frac{dy_{t+1}}{dx_t} &= \beta_1 \frac{dy_t}{dx_t} + \beta_3 \\ &= \beta_1 \beta_2 + \beta_3 \end{aligned}$$

$$\begin{aligned} \text{At time } t+2: \quad \frac{dy_{t+2}}{dx_t} &= \beta_1 \frac{dy_{t+1}}{dx_t} + \beta_4 \\ &= \beta_1^2 \beta_2 + \beta_1 \beta_3 + \beta_4 \end{aligned}$$

and so on.

Therefore, one-year elasticity: $\vartheta_1 = \beta_2$.

The three-year elasticity is equal to the sum of the elasticities at time t , $t+1$ and $t+2$:

$$\begin{aligned} \vartheta_3 &= \beta_2 + \beta_1 \beta_2 + \beta_1^2 \beta_2 + \beta_1 + \beta_1 \beta_3 + \beta_4 \\ &= \beta_2 (1 + \beta_1 + \beta_1^2) + \beta_3 (1 + \beta_1) + \beta_4 \end{aligned}$$

$$T \text{ year elasticity (T very large): } \vartheta_{1,r} = \frac{(\beta_2 + \beta_3 + \beta_4)}{1 - \beta_1}$$

A1.2 Variable elasticities

$$\log(y_t) = \beta_0 + \beta_1 \log(y_{t-1}) + \beta_2 x_t + \beta_3 x_{t-1} + \beta_4 x_{t-2} + \varepsilon_t$$

Where y_t and x_t are levels (ie, are not logged).

$$\begin{aligned} \text{Elasticity} &= \frac{dy_t}{dx_t} \cdot \frac{x_t}{y_t} \\ &= \left(\frac{dy_t}{y_t} \right) / \left(\frac{dx_t}{x_t} \right) \\ &= d\log y_t / \left(\frac{dx_t}{x_t} \right) \\ &= \frac{d\log y_t}{dx_t} \cdot x_t \end{aligned}$$

$$\frac{d \log y_t}{dx_t} = \beta_2$$

Therefore, the one-year elasticity: $\vartheta_1 = \beta_2 x_t$. From this point, the derivation is similar to the constant elasticities:

$$\vartheta_3 = [\beta_2(1 + \beta_1 + \beta_1^2) + \beta_3(1 + \beta_1) + \beta_4] x_t$$

$$\vartheta_{1,r} = \frac{(\beta_2 + \beta_3 + \beta_4) x_t}{(1 - \beta_1)}$$

A1.3 Squared terms

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_t^2 + \beta_4 x_{t-1} + \beta_5 x_{t-2} + \varepsilon_t$$

where y_t and x_t are both in logs, and therefore the elasticity of y with respect to x is $\frac{dy_t}{dx_t}$.

At time t : $\frac{dy_t}{dx_t} = \beta_2 + 2\beta_3 x_t$

At time $t+1$: $\frac{dy_{t+1}}{dx_t} = \beta_1 \frac{dy_t}{dx_t} + \beta_4$

$$= \beta_1 (\beta_2 + 2\beta_3 x_t) + \beta_4$$

At time $t+2$: $\frac{dy_{t2}}{dx_t} = \beta_1 \frac{dy_{t+1}}{dx_t} + \beta_5$

$$= \beta_1 [\beta_1 (\beta_2 + 2\beta_3 x_t) + \beta_4] + \beta_5$$

Therefore, one-year elasticity: $\vartheta_1 = \beta_2 + 2\beta_3 x_t$

$$\vartheta_3 = (\beta_2 + 2\beta_3 x_t)(1 + \beta_1 + \beta_1^2) + \beta_4(1 + \beta_1) + \beta_5$$

$$\vartheta_{1,r} = \frac{\beta_2 + 2\beta_3 x_t + \beta_4 + \beta_5}{(1 - \beta_1)}$$

Park Central
40/41 Park End Street
Oxford OX1 1JD
United Kingdom

Tel: +44 (0) 1865 253 000
Fax: +44 (0) 1865 251 172

Stephanie Square Centre
Avenue Louise 65, Box 11
1050 Brussels
Belgium

Tel: +32 (0) 2 535 7878
Fax: +32 (0) 2 535 7770

Thavies Inn House
7th Floor
3/4 Holborn Circus
London EC1N 2HA
United Kingdom

Tel: +44 (0) 20 7822 2650
Fax: +44 (0) 20 7822 2651