Journal of Air Transport Management 62 (2017) 90–98

Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: www.elsevier.com/locate/jairtraman

Modelling the asymmetric probabilistic delay of aircraft arrival

J.V. Pérez–Rodríguez [a, *], J.M. Pérez–Sánchez [b], E. Gómez–Déniz [c]

[a] Department of Quantitative Methods for Economics & Business, University of Las Palmas de Gran Canaria, Spain
[b] Department of Quantitative Methods for Economics & Business, University of Granada, Spain
[c] Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, Spain

ARTICLE INFO

Article history:
Received 31 October 2016
Received in revised form
6 March 2017
Accepted 7 March 2017

Keywords:
Airports
Asymmetric link
Delay
Logit models
Bayesian estimation

ABSTRACT

The main purpose of this paper is to present an asymmetric logit probability model to estimate and predict the daily probabilities of delay in aircraft arrivals. The proposed model takes into account statistical regularity, noting that more arrivals are on time than delayed, thus reflecting an asymmetric pattern of behaviour. The data analysed were obtained from the BTS and IATA databases for December 2014, corresponding to delays within the US airspace system for each carrier, measured at various US airports. The model was evaluated by analysing both in–sample and out–of–sample data, for main and control samples. The performance of the proposed asymmetric Bayesian logit model was compared with that of two others: frequentist logit and symmetric Bayesian logit. The main conclusion drawn is that the model we propose obtains the best fit, according to the statistics considered, and identifies a novel delaying factor, namely distance, which is not identified by the other models analysed.

1. Introduction

Most studies addressing dichotomous outcomes, such as success vs. failure, use classical logit and probit models, and therefore assume that the responses are symmetric. Nevertheless, in practice, the real proportion of (for example) successes and failures may not be symmetric. If this is the case, application of these classical models can led to model misspecification and a misinterpretation of the marginal effects and unidentified predictors, the consequences of which could be very significant. In the present study, we examine data corresponding to aircraft arrival and departure delays, which often present just this kind of asymmetry.

Arrival and departure delays in the airspace system are important variables because they cause significant losses to airlines and create problems for passengers, airports and staff. Delays can be categorised into gate delay, taxi–out delay, en–route delay, terminal delay and taxi–in delay (see Mueller and Chatterji, 2002). Since traffic management decisions are influenced by the predicted demand, better demand forecasting is always desirable. Departure time uncertainty is the major cause of demand prediction error; therefore, increased departure time reliability will directly increase the accuracy of demand prediction (Mueller and Chatterji, 2002). In consequence, scheduling and policy decision makers should seek to minimise the risk of delay and thus improve the forecasting accuracy of departure times when a probabilistic delay time model is used. Accordingly, it is important to determine the causes of delays in the airspace system, such as factors related to aircraft, airline operations, changes of procedure and traffic volume.

Several approaches can be taken to analyse this issue. On the one hand, we can attempt to estimate the actual duration of the delay. For example, Allan et al. (2001) analysed several determinants of flight delay at one US airport (Newark International Airport) and showed that adverse weather conditions influenced flight delays. On the other hand, Mueller and Chatterji (2002) modelled delay assuming it to be a random variable that follows a statistical distribution. Their study, seeking to improve delay prediction, analysed the departure, en–route and arrival delays of aircraft that operated out of one of ten major U.S. hub airports. Kwan and Hansen (2011) analysed causal factors including airport congestion, total traffic and en–route weather. The estimation results obtained suggested that airport congestion, measured by arrival queuing delay, was a major contributor to average delay (about 32%). Nevertheless, these authors concluded that a model with a single explanatory variable is inadequate to describe the reality of a system. Wong and Tsai (2012) analysed flight delay propagation employing a survival method (the Cox proportional hazard model). These authors developed departure and arrival

* Corresponding author.
E-mail addresses: jv.perez-rodriguez@ulpgc.es (J.V. Pérez–Rodríguez), josemag@ugr.es (J.M. Pérez–Sánchez), emilio.gomez-deniz@ulpgc.es (E. Gómez–Déniz).

http://dx.doi.org/10.1016/j.jairtraman.2017.03.001
0969-6997/

delay models that showed how flight delay propagation can be formulated through repeated chain effects in aircraft rotations performed by a Taiwanese domestic airline. Other papers have also analysed delay propagation using other econometric methods; see, for example, Xu et al. (2005, 2007), Liu and Ma (2008) and Cao and Fang (2012), among others, who used a Bayesian network approach, Pyrgiotis et al. (2013), who analysed a network of airports using a queuing model, and Derudder et al. (2010) and Diana (2011), who analysed the prediction of arrival delays using spatial analysis.

Another possibility is to analyse the probability of delay. To our knowledge, only a few studies have taken this approach. Among them, Abdel-Aty et al. (2007) identified the periodic patterns of arrival delay for non-stop domestic flights at the Orlando International Airport during 2002–2003. Using logistic regression, their results showed that time of day, day of week, season, flight distance, precipitation at Orlando International Airport and scheduled time intervals between successive flights were significantly correlated with arrival delay. In this field, too, Tu et al. (2008) attempted to model flight departure delay probability distributions, but did not assess all of the relevant aviation and meteorological parameters. In another study, Wesonga et al. (2012) analysed the probability of arrival and departure delay at Entebbe airport (Uganda), using a multiple parametric approach to determine the probability of aircraft delay. In this study, a robust approach was used to include the apparently significant meteorological and aviation parameters while computing the exact probabilities of delay.

Motivated by the desire to improve the accuracy of demand prediction, both en–route and at airports, using probabilistic delay forecasting, we analyse departure and arrival data for U.S. airports with different volumes of traffic and significant delays. Following both Abdel-Aty et al. (2007) and Wesonga et al. (2012), we also use logistic regressions. Specifically, we analyse not only how the airport factor and the airline factor can influence delays, but also the distance between airports, departure delay and daily patterns. However, unlike the latter studies, our paper is conducted using an asymmetric logit model. This choice was made because it has been observed that the proportion of on–time flights (leaving/arriving within 15 min of the scheduled time) is generally higher than that of delayed ones.

This study examines data corresponding to air traffic delay statistics compiled in the United States, and uses an asymmetric logit model in the belief that this provides better results than the standard fitted logit model.[1]

The rest of this paper is organised as follows. Section 2 presents the methodology, including a brief description of the three probabilistic logit models analysed, with respect to aircraft delay: the classical logit model, the symmetric Bayesian logit model and the asymmetric Bayesian logit model. Section 3 presents the data included in the analysis and the sampling procedure. Section 4 presents the estimation performed, the main results obtained and their discussion. Finally, in Section 5 we summarise the main conclusions drawn.

[1] This approach has been successfully used in other studies; for example, Chen et al. (1999) applied a Bayesian approach and an asymmetric link in analysing binary response data, when one response is much more frequent than the other. Similarly, Bermúdez et al. (2008) applied asymmetric logistic regression to model fraudulent behaviour, using a Spanish insurance database. In the area of health care, Sáez–Castillo et al. (2010) used an asymmetric logistic link to predict infection rates in a General and Digestive Surgery hospital department. More recently, Pérez–Sánchez et al. (2014) analysed the risk factors of automobile insurance claims, considering an asymmetric link in the logistic regression.

2. Methodology: logistic models

2.1. Classical models

Logistic regression has long been the standard method for studying the relationship between a binary response variable and one or more predictors or explanatory variables, using a cumulative density function (cdf), termed Ψ. Let x be a vector of explanatory variables and y the response variable taking values in $\{0,1\}$. This can be expressed as $\Pr(y=1|x,\beta)=\Psi(x'\beta)$. Then, by taking Ψ as the cdf of the logistic distribution, we obtain the logistic regression. In this case, the probability density function is symmetric about zero. Thus, the cdf approaches 1 at the same rate as it approaches 0. However, in many practical situations this is not a reasonable outcome, because data are often positively or negatively skewed and contain a substantial proportion of zeros (non–zeros) with respect to the proportion of non–zeros (zeros). This asymmetry may arise in diverse practical situations and then the logit model is not really appropriate. Since the pioneering study by Prentice (1976), various new models of dichotomous choice have been proposed to overcome this problem (under logit and probit assumptions), greatly assisted by advances in computer technology and software development. See, for example, Stukel (1988, 1990), Chen et al. (1999), Fernández and Steel (1998), Fletcher et al. (2005), Bazán et al. (2006, 2010) and Kumar and Manju (2015).

As observed by Stukel (1988) and Chen et al. (1999), the use of an asymmetric link function is recommended for binary response data when one response is much more frequent than the other.

The classical logit model is based on the following ideas. Let $y=(y_1,y_2,\ldots,y_n)'$ denote an $n\times 1$ vector of a dependent dichotomous variable, and let $x_i=(x_{i1},\ldots,x_{ik})'$ denote the $k\times 1$ vector of covariates for the set i. Here, x_{i1} may be $\mathbf{1}$, which corresponds to an intercept. A fit regression model is used to estimate the probability of belonging to a group included in y_i. In this study of flight delays, if $y_i=1$ the ith flight lands late, and $y_i=0$ otherwise. We assume that $y_i=1$ with probability p_i and $y_i=0$ with probability $1-p_i$. The regression model is given by $p_i=F(x'_i\beta)$, where F is the inverse of the standard logistic cumulative function (link function), and $\beta=(\beta_1,\ldots,\beta_k)'$ is a $k\times 1$ vector of regression coefficients, which represents the effect of each variable x_i on the model. Thus, the likelihood function, denoted as $l(y|x,\beta)$, is given by

$$l(y|x,\beta)=\prod_{i=1}^{n}\left[F(x'_i\beta)\right]^{y_i}\left[1-F(x'_i\beta)\right]^{1-y_i}, \tag{1}$$

where $F(s)=1/(1+e^{-s})$, $-\infty<s<\infty$, is a symmetric function with respect to zero. Regression coefficients are usually estimated by numerical evaluation of the likelihood function. In the present case, thus, the model provides the probability of each flight landing with delay. The next step is to consider a cutoff in this probability in order to determine whether a flight will land on time or not. The logit model was evaluated using STATA econometric software.

2.2. Bayesian models allowing symmetry and asymmetry

The regression logit model outlined above is too simple to be used for any serious empirical work when the sample data present asymmetry between the two values of the binary response variable. In this context, the Bayesian approach is a powerful tool providing more flexible models in regression analysis.

The main idea of the Bayesian regression model (Zellner, 1971; Koop, 2003) is to consider that the regression coefficients are random and fit a distribution function (the prior distribution). We propose two alternative Bayesian estimations of the logit model.