DOI: 10.1111/j.1752-5209.2007.0015.x

Modelling and forecasting football attendances

J. James Reade

St. Cross College, University of Oxford

Abstract. Estimating demand functions for football match attendance is increasingly popular, with focus usually on top divisions of national leagues and on groups of teams. Here, demand for one lower division team, Oldham Athletic, is analyzed. General-to-specific model selection is employed to find a model to explain attendances, which is then used to forecast attendances. The model explains and predicts attendances well, and additionally designed forecast combinations provide competitive forecasts due to the possibility of structural breaks.

'Somebody's got to watch Oldham, somebody's got to watch Wigan . . . and somebody's got to watch Coventry'

Richard Keys

1 Introduction

The desire to know what will happen in the future is almost certainly the most pressing question facing anyone in any sphere of life. As such, economists are continually asked to forecast, despite boasting a track record worse than an England cricket team in Australia. The cause of poor forecasts is that the world keeps changing, also the reason why people still ask economists to forecast. This 'non-stationarity' means that data series stretching over any period of time will not have constant means or variances, a consideration which, unmodelled, can cause problems for inference and forecasting. Hendry and Clements (2001) have provided a framework for forecasting in this 'non-stationary' world: better economic models will not forecast better because the world on which they are based has changed; instead forecasting methods that are flexible to these changes are required. Naive methods are proposed and shown to have good properties in the presence of structural breaks. In this note, many of these ideas will be discussed in the context of forecasting attendances at Oldham Athletic football matches. A model of demand for attendance will be constructed in Section 2, and its forecast performance will be assessed in Section 3. Section 4 concludes.

2 Modelling demand for football matches

Simmons (1996) and Garcia and Rodriguez (2001) investigate demand for attendance at top division English and Spanish league matches respectively, while Simmons and Forrest (2005) consider a panel of lower division English football matches. Different considerations affect lower division demand from top division demand such as level of brand loyalty, and here one dimension of the lower division panel is considered, as the author, having followed the team for nearly twenty years, is acquainted with systematic and idiosyncratic factors that have affected demand.

As a primary focus of any demand analysis is estimating demand elasticities, a log-linear regression model will be used, where y_t is log of attendance, X_t are the K (possibly logged) independent variables yet to be determined, and β is a $K \times 1$ vector of coefficients:

$$y_t = \beta X_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2). \tag{1}$$

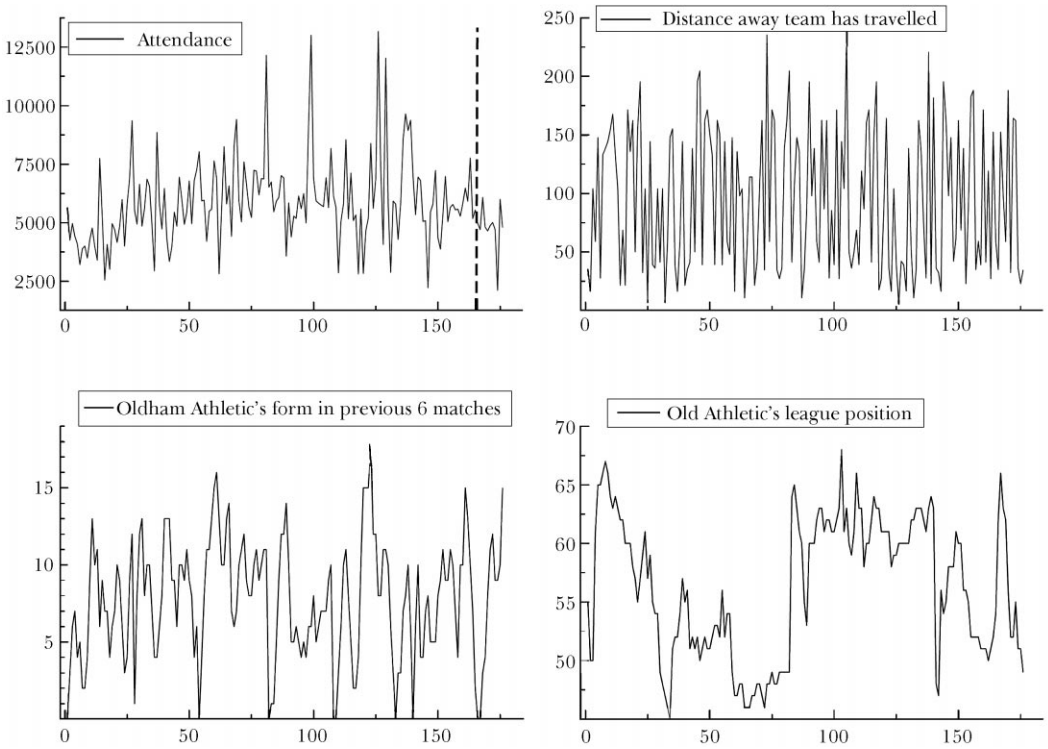
With just four near capacity crowds in nearly two hundred matches, there is no truncated data problem.[1] Thus OLS is used for estimation, providing the estimate $\hat{\beta} = \sum_{t=1}^{T} X_t y_t / \sum_{t=1}^{T} X_t^2$ for β in (1).

2.1 The data

Data for football matches, such as attendance and score, are very easily available back to 2000, giving a sample of 176 matches. Demand factors might be categorised into: own price, price/existence of alternatives, quality of product and income.

1. The near-capacity crowds are explained by one-off factors (a promotion play-off match, a free-entry offer, and two cup matches against large local rivals), and are dummied out.

Figure 1 Plots of four data series; attendance, distance away team travelled, Oldham's form measured by points won in previous six matches, and Oldham's league position (lower is higher in the table).



Own price Own price is not a particularly informative variable, as ticket prices are held fixed for entire seasons (with occasional discounts), with only four increases, from £14 to £20, over the sample.

Price/existence of alternative goods Oldham Athletic play in Greater Manchester, which has the highest concentration of professional football teams in the world, and hence alternative products are not hard to imagine. Additionally shopping, cinema, television and computer games all provide enticing alternatives for possible attendees. Over the sample period considered, it's unlikely that these factors have changed much as almost all the major entertainment developments in Oldham were complete in 2000. The timing of a match may matter: midweek evening matches may coincide with other social events as opposed to the traditional 'football' time of 3pm on a Saturday. Televised football matches will also likely affect midweek attendance.[2]

The quality of the product The quality of the two teams determines the attraction of a particular match, although measuring quality is not easy. Match results, the form of the team, will be one clear measure of quality. Changes in the composition of the teams (primarily Oldham) will matter, such as player sales or acquisitions, managerial changes and boardroom takeovers. Two time horizons appear relevant. Some supporters buy tickets for the entire season, and having made sunk cost, are less affected by short-term form. Season ticket holder numbers accounted for around two-thirds of the average attendance in the last four seasons. Short-term factors that affect attendance of non-season ticket holders may include the opposing team: local rival teams provide excitement, and also a higher number of visiting supporters, while larger visiting teams also would be expected to increase demand for a match. The competition the team is playing in may affect demand, as Oldham compete in a league competition and three cup competitions.

Income and supply Over the sample period considered here, economic stability means that it is unlikely income would play a factor in attendance determination. There is also no identification problem between demand and supply: stadium capacity has remained fixed at 13,700 over the entire sample.

Figure 1 plots four of the variables considered above; the attendance, the distance the visiting team has travelled to

2. Television companies are not allowed to show matches beginning at 3pm Saturday.

Table 1 Selected model: dependent variable is the logarithm of match attendance, $\log(Att)_t$.

	Coefficient (t-value)		Coefficient (t-value)		Coefficient (t-value)
Constant	7.244	*Midweek*	−0.186	*CompBal*	−0.002
	(18.2)		(−6.68)		(−2.28)
$\log(Att)_{t-1}$	0.079	*AprMay*	0.097	*BoxingDay*	0.234
	(2.18)		(2.87)		(2.97)
$\log(Att)_{t-2}$	0.083	*Player*Δ	−0.025	*LastHGame*	−0.003
	(2.42)		(−2.62)		(−1.34)
Form	0.017	*Manager*	0.121	*D*0102	0.092
	(5.35)		(2.33)		(3.04)
LeagueCup	−0.421	*Grimsby*	0.842	*D*0203	0.140
	(−5.83)		(6.57)		(4.36)
LDVVans	−0.617	*Torquay*	0.276	*D*0304	0.146
	(−13.5)		(2.1)		(4.58)
$\log(Distance)$	−0.073	*Playoffs*	0.480	*D*0405	0.190
	(−5.23)		(3.73)		(6.26)
ClubSize	0.115	*OnTV*	0.258	*Rivals*	0.111
	(7.73)		(3.13)		(3.0)
Wed	0.075				
	(1.22)				
$\hat{\sigma}$	0.12	AR F(2, 147)	1.41 (0.25)	Hetero F(32, 116)	1.72 (0.02)*
R^2	0.86	ARCH F(1, 147)	0.06 (0.82)	RESET F(1, 148)	3.62 (0.06)
F(24, 149)	39.4 (0.0)**	Normality $\chi^2(2)$	1.22 (0.5)		

visit Oldham, the form Oldham have displayed in recent matches, and Oldham's league position over the sample.

2.2 Selecting the best model

Selecting a particular model from a host of possible explanatory variables is not simple. Relevant variables may be omitted, and irrelevant variables kept. The econometrics literature is littered with model selection techniques, and critiques of them. Two automated model selection programs, PcGets (Hendry and Krolzig, 2001), and autometrics (Doornik, 2006), are based on the 'General-to-Specific' reduction theory of Hendry (1995, Chapter 9), and address many criticisms of model selection. From a well specified econometric model (passing standard mis-specification tests), all feasible search paths are taken, thus avoiding path dependency. These path searches involve omitting statistically insignificant variables only if the resulting model passes all mis-specification tests. A path is terminated when no further variables can be omitted, and the resulting model is an approximation to the true data generating mechanism, and simulation studies (Krolzig and Hendry, 2001) suggest that models selected using PcGets are very close to the true data generation mechanism, or the linear approximation thereof.

2.3 The model

Table 1 contains the final model. Autometrics chose 24 explanatory variables for attendance. The selected model passes all mis-specification tests at a 1% level, although there is a slight heteroskedasticity concern. 86.4% of the variation in attendance is captured by the model, and the standard deviation of the model is 12%. Notable features of the model are that price does not enter; it is insignificant, as price shows little variation over the sample compared to attendance. Improved quality of product has a positive effect on attendance: the attendance increases if Oldham are playing better (Form), if the visiting team is a club of larger stature (ClubSize), if the two teams are more closely matched in terms of league position (CompBal, difference in league position before match), and if the match is against a rival team.[3] Attendance decreases if the match is in a particular competition (the LDV Vans Trophy reduces attendances by 60%, the League Cup by 40%), and if the visiting team has travelled further (a 10% increase in distance decreases the attendance by 0.7%). The importance of a given match has an effect: in April and May, as the season ends and matches have great importance for promotion and relegation, attendances increase (AprMay). Changes in the quality of the product have an effect: a new manager increases the gate by 12.1%, while (surprisingly)

3. Rival teams are defined as Bradford City, Huddersfield Town, Wigan Athletic, Blackpool and Stockport County.

new players decrease the attendance by 2.5%.[4] The existence of alternatives perhaps suggests why midweek games have almost 20% lower crowds.[5] LastHGame, the number of days since the previous home game, appears to have the wrong sign; one would expect that with a smaller number of days between matches, fewer supporters attend owing to the cost involved.[6] There are seasonal effects: four dummy variables taking unity for a particular season are significant, and show that the mean in the middle three seasons of the sample was considerably higher than at the start or end, something noticeable in Figure 1. The Boxing Day fixture, traditionally one of the biggest of the year, attracts 23.4% more attendees. Finally there is a number of one-off effects; Grimsby relates to a match where entry was free, Torquay to a league match with substantially reduced entry prices, and Playoffs to a promotion play-off match which attracted a capacity crowd.

3 Forecasting Attendance

Forecasts are carried out based on the model: estimation is over the period $t = 1, \ldots, T$, and one-step ahead forecasts are made to $T + 1$, as supporters make and revise decisions based on the most recent few matches and events. The forecast is:

$$\hat{y}_{T+1} = \hat{\beta} X_{T+1}. \tag{2}$$

Forecasts can be assessed ex post by their error, which at $T + 1$ is:

$$\hat{e}_{T+1} = y_{T+1} - \hat{y}_{T+1}. \tag{3}$$

A conventional method of assessing forecast performance is shortening the estimation sample and comparing forecasts to known outcomes over a 'training period'. A model that forecasts well on this training period does not necessarily predict more accurately future values of y_t, because structural breaks can happen at any moment and render a good model bad, and vice versa. Furthermore, the information set was different in previous periods, hence a different model might have been selected.

The training period here is the current football season. From Figure 1, attendances this season (after the dashed vertical line) are distinctly lower than they were in previous seasons: in the previous season two attendances fell below 5000; already this season eight have been below 5000. Within this are three notable observations: Match 3 (attendance 6080), the visit of a team that, following two consecutive promotions, brought an unexpectedly large number of supporters; Match 9 (2118), an LDV Vans Trophy match, which, at about 40% of 5000, is not an outlier from Table 1; Match 10 (6001), against local rivals Bradford, so also not an outlier. This apparent downward shift, allied with clear systematic patterns, presents a challenge for forecasting.

The training period forecasts are banded into two groups – models based on the model selected above, denoted 'structural models', and models constructed without any reference to the economic processes underlying attendance determination, such as random walk models and combinations of forecasts, denoted 'naive models' – and plotted in Figure 2.

Considering the structural-based models, the Model plot suggests that the model forecasts reasonably well, in particular in recent matches, picking up the large variation in Matches 9 and 10. The Model* plot is the forecasts produced using the model selected by autometrics on the data up to the forecast. This model does badly, in part because the Match 3 outlier hinders any data-based attempt to identify the downward shift. Clements and Hendry (2004) discuss the benefits of adding an 'intercept correction' term, a dummy variable taking unity for the last t_b observations of estimation, and extended into the forecast period. It corrects the level of the process y_t for any break that might have happened, so forecasts are made around the new level. Intercept corrections in the form of the seasonal dummies ($D0607$, etc) are included in the final selected model, but these are not selected earlier in the training period. Once added the model begins to forecast better, except the inexplicable woeful forecast of Match 7. The efficacy of these 'intercept corrections' can be seen in the plot where they are not included in the model, where the model generally over-predicts.

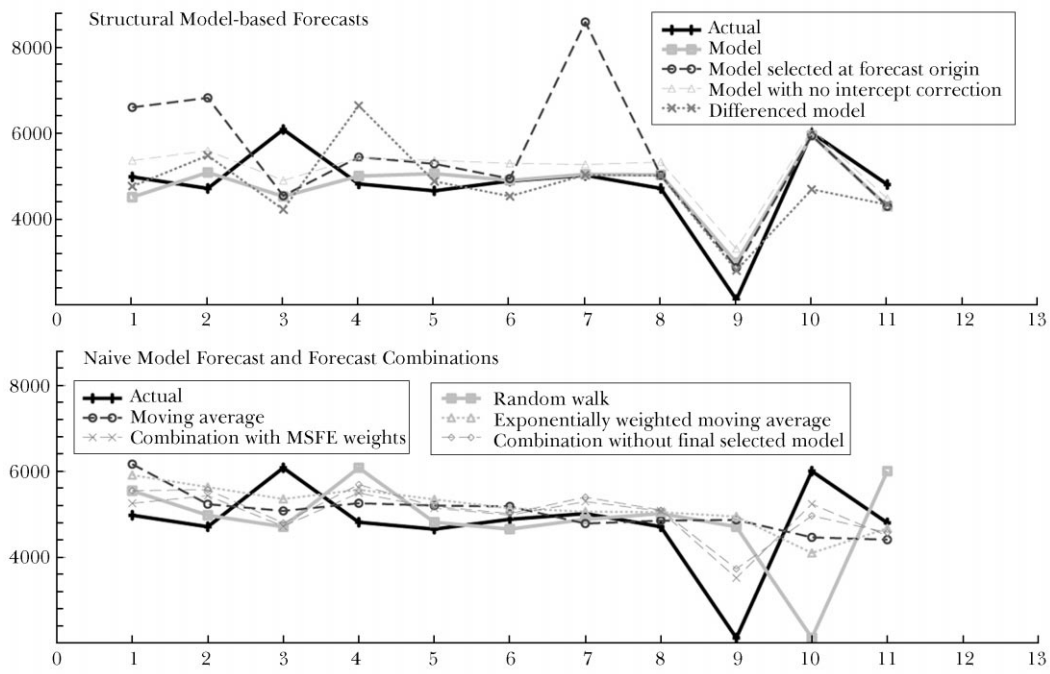
Another forecast model proposed by Clements and Hendry (2004) is a differenced version of (1):

$$\hat{\Delta} y_{T+1} = \beta \Delta X_{T+1} \Rightarrow y_{T+1} = \beta \Delta X_{T+1} + y_T \tag{4}$$

This model copes with level shifts by using the level of the process only at time T, and is plotted in Figure 2. The aforementioned 'outliers' in Matches 3 and 10 cause forecast failure as the model erroneously assumes the previous observation is the new level of the process; otherwise the model forecasts competitively.

4. Player moves is defined as acquisitions minus sales. Most variation in this variable is between seasons, and hence perhaps picks up the phenomenon that early season attendances are usually lower than later in the season. Also, as the club produces youth players, this variable is predominantly negative as these players eventually move on.
5. The anomaly with Wednesday is explained by the fact Oldham very rarely play midweek games on a Wednesday and hence the Wed variable simply picked up two outlying observations.
6. This perhaps reflects that the match after a long break has generally been an uninspiring match; the effect is only significant at a 10% level.

Figure 2 Forecasts from various models; upper panel structural based models, lower panel naive forecasts and combinations.



Moving to the bottom panel of Figure 2, the first naive device is a random walk model:

$$\hat{y}_{T+1} = y_T \tag{5}$$

The forecast is simply the previous attendance, and it reacts well to the initial shift, providing the best forecast of Match 2. Two moving average models are considered. A moving average of the previous three matches is plotted, along with an exponentially weighted moving average, with a smoothing parameter of $\alpha = 0.3$ in:

$$s_t = \alpha y_T + (1 - \alpha)s_{t-1}, \quad \hat{y}_{T+1} = s_T. \tag{6}$$

One imagines these series will move quickly enough to capture the effect of a break, and indeed around the middle of the sample these models forecast well. However, Matches 9 and 10 display the advantage of structural models; none of the naive models picks up the variation that is explained by cup competition and local rivals in the structural model.

Finally, a simple forecast combination is attempted. Bates and Granger (1969) combined two forecast models using a number of different weights based on the mean squared forecast error, while Clements and Hendry (2004) showed that combination may be effective if models respond differently to shocks. A combination of the other seven forecasts is plotted, and provides a forecast that is better than all the naive devices.[7] There is no reason why a more intelligent combination might not forecast better; the inverse of the mean absolute error was used here. Weights could vary for each prediction given the nature of the match; e.g. the LDV Vans Trophy attracts very few spectators, hence structural models may be given stronger weight. In league matches, averaging might help bring structural forecasts back towards the average, reducing forecast error.

This is not an exhaustive array of forecast models, but it illustrates the problems for forecasting attendances, due to large systematic variation nested within varying averages. Forecast combination might prove beneficial, and this is a subject of future research.

4 Conclusions

In this note, general-to-specific methods of model selection have been applied to demand for attendance at Oldham Athletic football matches. The resulting model explains attendance well, and is shown to forecast quite well. Naive models cannot outperform it, despite a structural break near the forecast origin. This reflects that the data do not satisfy textbook definitions of stationarity or non-stationarity, and shows that the forecaster needs to be pragmatic in his approach to forecasting.

7. In real-time, such a forecast would not be possible. Combining the six models known at each point provides a similar forecast, shown in Figure 2.

Acknowledgements

As a Christian, I firstly thank God. I would also like to thank David Hendry, Jennie Castle and Nick Fawcett.

References

Bates, J. and Granger, C.W.J. (1969) 'The combination of forecasts'. *Operations Research Quarterly*, 20, 451–468.

Doornik, J.A. (2006) 'Autometrics: further applications of automatic model selection'. Mimeo, Nuffield College.

Garcia, J. and Rodríguez, P. (2001) 'The determinants of football match attendance revisited: Empirical evidence from the Spanish football league', Economics Working Papers 555, Department of Economics and Business, Universitat Pompeu Fabra.

Hendry, D.F. (1995) *Dynamic Econometrics*, Oxford, Oxford University Press.

Hendry, D.F. and Clements, M.P. (2001) *Forecasting Non-stationary Economic Time Series*, Cambridge, MA, MIT Press.

Hendry, D.F. and Clements, M.P. (2004) 'Pooling of forecasts', *Econometrics Journal*, 7, 1–31.

Hendry, D.F. and Krolzig, H.-M. (2001) *Automatic Econometric Model Selection Using PcGets*, Timberlake Consultants Ltd, London.

Krolzig, H.-M. and Hendry, D.F. (2001) 'Computer automation of general-to-specific model section procedures', *Journal of Economic Dynamics and Control*, 25 (6–7), 831–866.

Simmons, R. (1996) 'The demand for English league football: A club-level analysis', *Applied Economics*, 28 (2), 139–155.

Simmons, R. and Forrest, D. (2005) 'New issues in attendance demand: The case of the English football league', Lancaster University Management School Working Paper 2005/004.

James Reade is a third year DPhil candidate (Economics) at St. Cross College having completed the MPhil in Economics at Oxford. He studied undergraduate economics at Durham University. He was recently part of the victorious Oxford team at the Econometric Game in Amsterdam.

