30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Regression Forecasting of Patient Admission Data

Justin Boyle, Marianne Wallis, Melanie Jessup, Julia Crilly, James Lind, Peter Miller, Gerard Fitzgerald

Abstract— **Forecasting is an important aid in many areas of hospital management, including elective surgery scheduling, bed management, and staff resourcing. This paper describes our work in analyzing patient admission data and forecasting this data using regression techniques. Five years of Emergency Department admissions data were obtained from two hospitals with different demographic techniques. Forecasts made from regression models were compared with observed admission data over a 6-month horizon. The best method was linear regression using 11 dummy variables to model monthly variation (MAPE=1.79%). Similar performance was achieved with a 2-year average, supporting further investigation at finer time scales.**

I. Introduction

THE majority of acute admissions in hospital arise from the Emergency Department (ED). With experience, hospital bed managers can identify which days of the week admissions from ED may reach saturation level (i.e. access block) because of the pre-assigned inpatient beds for elective surgery. With the current retrospective, reactive bed management approach, a common scenario experienced is the occurrence of a full ED on a Monday morning and by midday, the hospital has been put on ambulance bypass. A typical resolution to this situation is cancellation of elective surgery patients at the last minute. This action is very inefficient for physical and human resource utilization and has left many patients dissatisfied with the health system.

With hospital occupancy rates approaching 100% on a regular basis, more efficient management of inpatient beds to reduce access block is becoming essential [1], [2]. Access block has been shown to impair the function of emergency departments, and lead to less favourable outcomes for patients [3], [4]. Access block can be reduced by managing inpatient beds prospectively, thereby booking elective surgery on specific times and days of the year where there is lower demand from ED. We wished to test the assumption by others that emergency admissions are difficult to predict [5].

Manuscript received April 15, 2008.

J. Boyle is with the Australian E-Health Research Centre, CSIRO ICT Centre, PO.Box 10842, Adelaide St, Brisbane, 4000, Australia (phone: +617 30241606; fax: +617 30241690; e-mail: justin.boyle@csiro.au).

M. Wallis and M. Jessup are with Griffith University, Qld, (e-mail m.wallis@griffith.edu.au, m.jessup@griffith.edu.au)

J. Lind, J. Crilly and P. Miller are with the Gold Coast and Toowoomba Hospitals, Queensland Health (e-mail: James_Lind@health.qld.gov.au, Julia_Crilly@health.qld.gov.au, Peter_Miller@health.qld.gov.au)

G. Fitzgerald is with Queensland University of Technology (e-mail: gj.fitzgerald@qut.edu.au).

Several attempts have been made to develop mathematical models to predict the number of likely emergency admissions. Bagust *et al.* [5] model the dynamics of the hospital system using discrete-event stochastic simulation using a MS Excel spreadsheet. The model was based on simulated emergency admissions for a hypothetical hospital, and they conclude that spare bed capacity is essential for the effective management of emergency admissions. Champion *et al.* [6] used the SPSS Trends package to automatically identify optimal models to forecast monthly emergency department presentations. The authors report that a simple seasonal exponential smoothing model provided optimal forecasting performance, and forecasts for the first five months of 2006 compared well with observed attendance data. Reis and Mandl [7] used the SAS package to fit ARIMA models to nearly a decade of ED presentation data, and report a Mean Absolute Percentage Error of 9.37% when validated against the final 2 years of the dataset. Similar work was performed by Pereiras [8] who used the SPSS Trends package and other commercial forecasting software to predict monthly blood transfusion demand. The results of these studies indicate that univariate time series methods can be useful to forecast monthly health data. Our work differs from these studies in that we wish to predict admissions as distinct from all presentations, as this represents demands made on hospital beds. Also, when compared to other regression attempts which only report the degree of fit (R^2) of the forecasting model to the data [9],[10] our validation is based on a "hold-out" set of data not included in developing the models.

II. Data Access

Following ethics and State Government Privacy legislation approvals, we obtained five years of ED presentation and admission data (1/7/02 – 30/6/07) for two hospitals in Queensland, Australia. The two hospitals were chosen for their different demographic characteristics. Toowoomba reflects an entire regional population (~90,000) served by one ED with a fairly stable population, unlike the Gold Coast, which has one of the busiest EDs in the state, a large itinerant population and numerous other EDs serving the area. Patient statistics for the two EDs are presented in Table I. Although the data supplied for the project was deidentified, it was considered confidential as patients had not given their explicit consent for their records to be used for this project. Additional applications to government privacy bodies were worthwhile to enable the modelling to be based on actual patient records rather than simulated data.

TABLE I
EMERGENCY DEPARTMENT PATIENT STATISTICS 2002-2007

	Toowoomba Hospital Mean ± 1 Standard Dev.	Gold Coast Hospital Mean ± 1 Standard Dev.
Daily Presentations	113 ± 14	152 ± 20
Daily Admissions	22 ± 5	50 ± 8
Admission Rate	20% ± 4%	33% ± 5%

III. TIME AND DATE ANALYSIS

The data includes date and time of presentation and admission which provides useful information on peak admission times experienced within the EDs, and the days of the week that represent higher ED workloads and hospital bed demands. Figure 1 shows the mean and 95% Confidence Interval band for the days of the week (left) and months of analysis (right) for the arrival time of all presentations. The left plots indicate that at both hospitals, the busiest days for presentations are over the weekend and Mondays. The right plots shows that presentations at Toowoomba were fairly stable, whilst the Gold Coast experienced an overall increase in the number of patients presenting over the five years (approximately 40% increase). Population growth over the study period was 1.3% (Toowoomba) and 3.3% (Gold Coast), which highlights the effect of the large itinerant population at the Gold Coast.

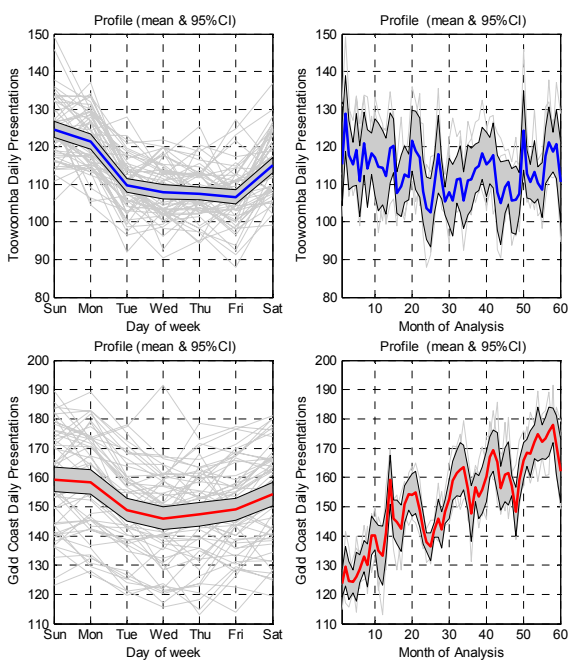

Fig. 1. Day-of-week and Monthly trends for patient presentations; Top: Toowoomba Hospital; Bottom: Gold Coast Hospital

If we consider the time that patients that require admission leave the ED and are admitted to a bed, we can see that Mondays (and Tuesdays at the Gold Coast) are busiest (Figure 2). Interestingly the trend over the months of analysis shows a plateau effect at the Gold Coast, which could be attributed to bed capacity being reached.

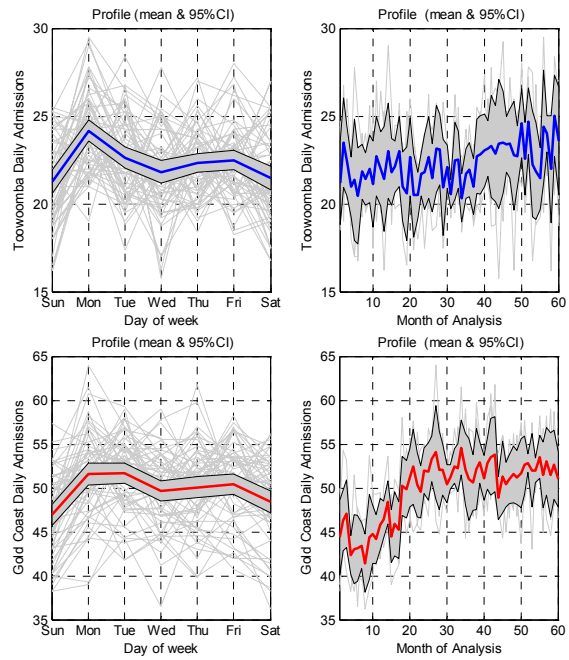

Fig. 2. Day-of-week and Monthly trends for patient admissions; Top: Gold Coast Hospital; Bottom: Toowoomba Hospital

IV. REGRESSION MODELS

The time that Admitted patients leave the ED is of most interest to bed managers as it represents when patients require beds. Thus this is the data that we wish to forecast. The general regression model followed was:

$$y = X\beta + \varepsilon \quad , \tag{1}$$

where y = response, X=design matrix, β=parameters, and ε = random disturbances. Seasonality was modeled by including dummy variables in the design matrix (Table II), where each variable had only two allowable values, 0 or 1, depending on the month.

TABLE II
DESIGN MATRIX WITH DUMMY VARIABLES TO HANDLE SEASONALITY

Time (Months)	J	F	M	A	M	J	J	A	S	O	N	D
Jan'03	1	0	0	0	0	0	0	0	0	0	0	0
Feb'03	0	1	0	0	0	0	0	0	0	0	0	0
Mar'03	0	0	1	0	0	0	0	0	0	0	0	0
Apr'03	0	0	0	1	0	0	0	0	0	0	0	0
:												
Dec'06	0	0	0	0	0	0	0	0	0	0	0	1

There were several variations of this model:

1. Linear model with 12 monthly dummy variables;
2. Linear model with 11 monthly dummy variables, to assess the potential for multicollinearity; In multiple regression, computational problems arise if explanatory variables are highly correlated with one another and the regression coefficients associated with those explanatory variables can be unstable.

3. Quadratic model - containing dummy variables, time and time-squared;
4. Normalised model - where data was normalised by the number of days in each month to reduce the effect of shorter months;
5. Autoregressive model for the error with two stages:
   (i) regression model for original data:
   $$y_t = (X_t * b) + r_t \quad (2)$$
   (ii) autoregressive model for residuals:
   $$r_t = (\rho * r_{t-1}) + u_t \quad (3)$$

Initially, four years of admission data were aggregated into monthly totals and used to generate regression fits. An example is shown in Figure 3:

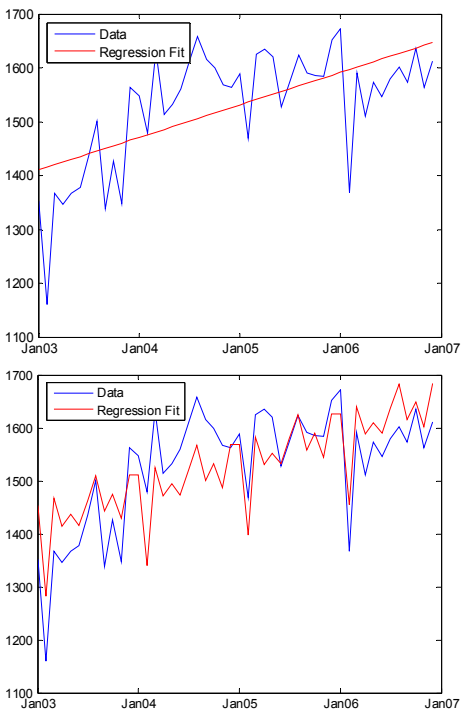

Fig. 3. Linear regression fit; Left: Without dummy variables R^2=0.41 Right: With dummy variables R^2=0.62

There was autocorrelation evident between adjacent residuals (measured by plotting the residuals and testing formally with the Durbin-Watson statistic) in all models except the autoregressive model, and forecasts for the next 4 years with this model are shown in Figure 4.

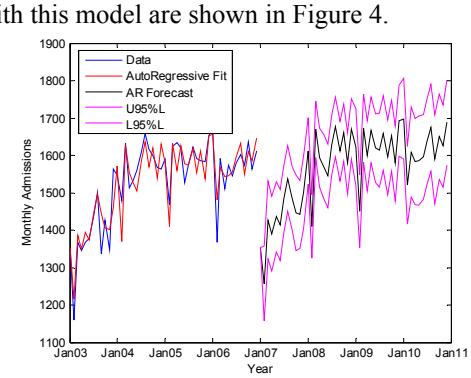

Fig. 4. Autoregressive forecast over a 4-year horizon

It is obvious that this forecast horizon is not useful, as the model exhibits the admissions behaviour of the earlier years before the plateau effect. Thus it was desired to assess forecast accuracy using two years instead of four. With this time frame, residuals were not correlated for the models. Example forecasts for the six months Jan'07-Jun'07 are shown in Figure 5. Actual admission data are plotted in the right hand forecast plots as crosses (+) and it can be seen that observed data mostly fall within the 95% Prediction band.

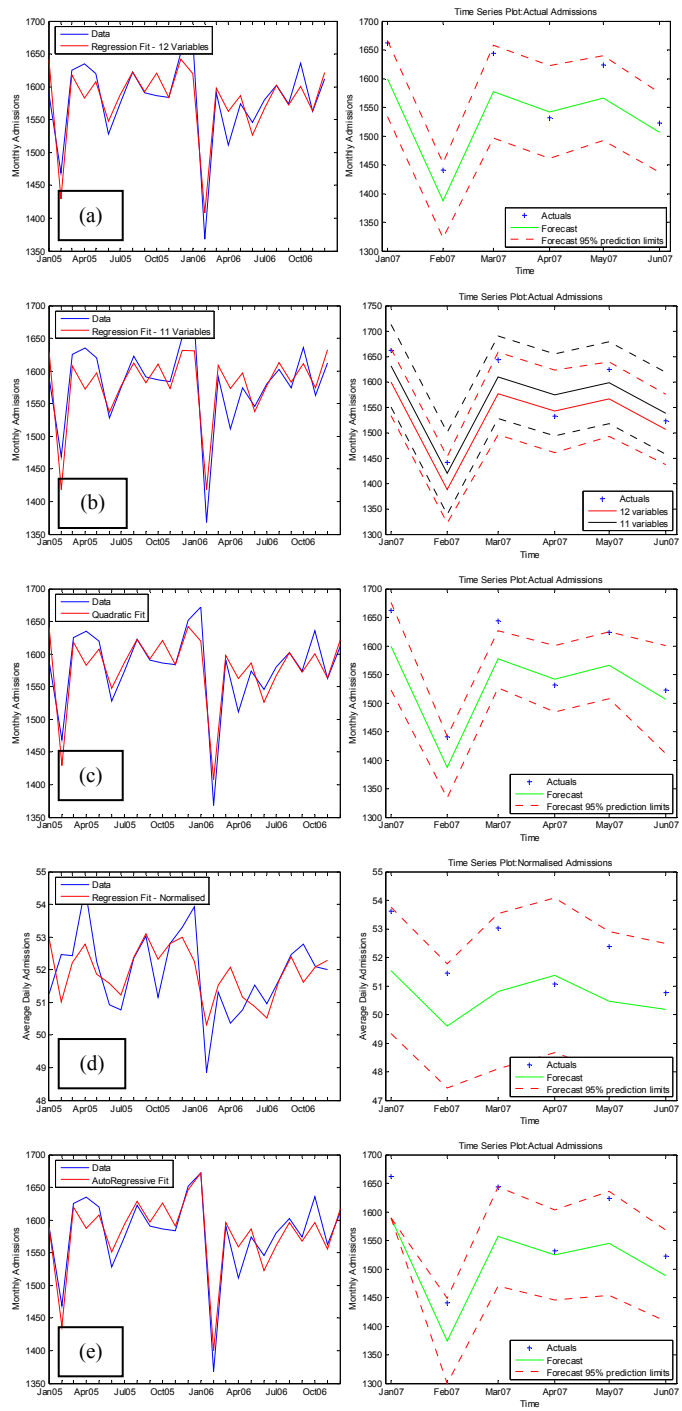

Fig. 5. Models (Left) and Forecasts (Right): a) Linear 12 variables R^2=0.81 B) Linear 11-variables R^2=0.78 c) Quadratic R^2=0.81 d) Normalised R^2=0.41 e) Autoregressive R^2=0.87.

The accuracy of each forecast method is summarised in Table III and Figure 6 using the mean absolute percentage error defined as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|PE_t\right| , \qquad (4)$$

where PE_t is the Percentage Error of forecasts, defined as:

$$PE_t = \left(\frac{Y_t - F_t}{Y_t}\right) \times 100. \qquad (5)$$

TABLE III
ACCURACY OF REGRESSION FORECASTS JAN'07-JUN'07

Model	Absolute Percentage Error $\lvert PE_t \rvert$ (%)						MAPE
	Jan	Feb	Mar	Apr	May	Jun	
Linear-12 variables	3.8	3.7	4.1	0.7	3.6	1.1	2.81%
Linear-11 variables	1.8	1.5	2.1	2.8	1.6	1.0	1.79%
Quadratic	3.8	3.7	4.1	0.7	3.6	1.1	2.81%
Autoregressive	4.4	4.6	5.3	0.5	4.9	2.3	3.66%
Normalised	3.9	3.6	4.2	0.6	3.7	1.2	2.85%
Extrapolated	7.3	1.2	3.2	5.2	3.3	8.2	4.73%
2-year average	1.9	1.6	2.2	2.7	1.7	0.9	1.82%

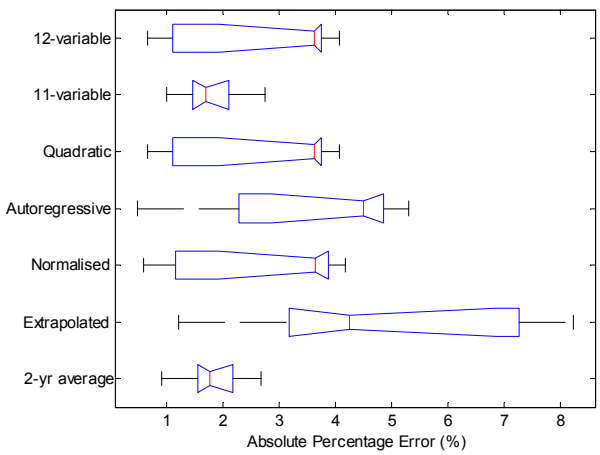

Fig. 6. Comparison of Forecast Techniques; Only the means of the 11-variable and the Extrapolated methods are significantly different.

V. DISCUSSION

For interest we compared forecasts made with the regression models with two simpler methods. Abdel-Aal and Mangoud [11] describe the estimation of patient volume at month k of year 2007, $\hat{A}_k(2007)$, from the corresponding known value for year 2006, $A_k(2006)$, using

$$\hat{A}_k(2007) = A_k(2006)\frac{\overline{A}_e(2007)}{\overline{A}(2006)}; \quad k=1, 2, \ldots.12 \qquad (6)$$

where $\bar{A}(2006)$ is the actual annual mean for year 2006 and $\bar{A}_e(2007)$ is the extrapolated annual mean for year 2007 as determined from a polynomial fit to data for all available years. The forecasted values using a polynomial of order 1 are not as good as the regression models (MAPE=4.73%). Also included in Table III is the MAPE for a simple average of the monthly values for the preceding 2 years (ie. Jan'07 = (Jan'05 + Jan'06)/2. It can be seen that this method performs very well at the monthly level. Overall the linear regression model with 11 dummy variables for monthly variation had the lowest mean absolute percentage error. When performing multiple comparison testing on the means of each technique, only the means of the 11-variable technique and the Extrapolated method are significantly different ie. simple methods, such as the two-year average method should not be dismissed. Future work involves assessing accuracy across finer time periods (daily and hourly) and implementing additional forecasting methods (exponential smoothing and Box-Jenkins Autoregressive Integrated Moving Average methods).

VI. CONCLUSION

The highest accuracy was linear regression with monthly variation modeled with 11 dummy variables. This method had the lowest mean absolute percentage error for forecasts, and it is noted that this does not necessarily correspond to the model with the highest R^2 value when fitted to training data. A simple averaging technique shows comparable forecasting performance to regression analysis and will be assessed along with other methods across finer time periods.

ACKNOWLEDGMENT

The assistance of Staff within Decision Support Services, Queensland Health who assisted with the extraction and deidentification of data is acknowledged.

REFERENCES

[1] B. Asplin, K. Rhodes, L. Crain *et al.* Measuring emergency department crowding and hospital capacity, Academic Emergency Medicine, vol. 9(5), pp. 366-367, 2002

[2] P. Cameron, P. Scown, D. Campbell, Managing access block, Australian Health Review, vol. 25 (4), pp. 59-68, 2002

[3] D. Brown, Marked increase in patients who leave the ED without treatment, Academic Emergency Medicine, vol. 9(5), pp. 510, 2002

[4] P. Cameron, Increase in patient mortality at 10 days associated with emergency department overcrowding, MJA vol. 184(5), pp. 203-4, 2006

[5] A. Bagust, M. Place, J. Posnett, Dynamics of bed use in accommodating emergency admissions: stochastic simulation model, BMJ vol. 319 (7203), pp. 155-158, 1999

[6] R. Champion, L. Kinsman, G. Lee, *et al.*, Forecasting emergency department presentations, Aust Health Rev vol. 31(1), pp. 83-90, 2007

[7] B. Reis B, K. Mandl, Time series modeling for syndromic surveillance, BMC Med Inform Decis Mak. vol. 23;3:2. Epub 2003, Jan 2003

[8] A Pereira, Performance of time-series methods in forecasting the demand for red blood cell transfusion, Transfusion vol.44(5), pp. 739-46, 2004

[9] D. Holleman, R. Bowling, C. Gathy, Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data, J Gen Intern Med vol. 11(4), pp. 237-239, 1996

[10] A. Diehl, M. Morris, S. Mannis, Use of calendar and weather data to predict walk-in attendance, South Med J, vol. 74(6), pp.709-712, 1981

[11] R. Abdel-Aal, A. Mangoud, Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis, Computer Methods and Programs in Biomedicine, vol. 56(3), pp. 235-247, 1998