IDENTIFYING PREDICTORS OF EMERGENCY MEDICAL SERVICE (EMS) DEMAND IN THE CITY OF TORONTO

Parshan BAHRAMI[1], Ernesto DIAZ LOZANO P.[1]

[1]Department of Civil Engineering, University of Toronto

Keywords: Emergency Medical Service, demand modelling, regression modelling

ABSTACT

Predicting Emergency Medical Service (EMS) demand for an urban area is a complex task due to the large number of factors that may drive increased need for ambulance services. Previous studies have modelled EMS demand using various approaches including least-squares regression, ridge regression, and artificial neural networks. However, several of these studies are more than 30 years old, and none have been performed in the City of Toronto. We estimated a multiple least-squares regression model to identify predictor variables that can be used to forecast EMS demand (measured by number of ambulance calls). We found that 6 predictors can be used to estimate EMS demand: number of robberies, number of hazardous incidents, number of seniors living alone, number of premature mortalities, traffic volume, and local employment. Our model can be used by city planners to forecast increase in EMS demand and to assess the placement of ambulance station locations.

BACKGROUND

One of the most important services that a city must provide to its residents is an efficient Emergency Medical Service (EMS). There are several life-threatening events that require urgent medical care within minutes, such as cardiac arrests, strokes, severe trauma, among others. To design an EMS system that can properly address these situations, several cities have defined standards for ambulance response times which are used as targets or metrics. In Toronto, the 2016 response time ranges from 6 minutes for an absolute emergency (e.g. a cardiac arrest), to 25 minutes for a minor injury (Ministry of Health and Long Term Care, 2016). To be able to meet these standards, the city must ensure that it has the right number of ambulances to serve the demand from the different neighbourhoods within the city. However, predicting this demand is not an easy task, as there may be multiple factors that drive the need for EMS in a given neighbourhood. To address this gap, we have prepared a regression model that identifies predictors for EMS demand (measured by number of ambulance calls) in each of the 140 neighbourhoods that constitute the city. We believe this analysis will be useful in evaluating how changes in explanatory variables due to city development will affect the demand for EMS, and thus our model will be a useful tool for service planning.

The development of EMS demand forecasting models started in the 1960's. Early models were extremely inadequate and used simple statistics, for example, the mean time between calls, to predict EMS demand. These models were simply ineffective, as trends and causal factors were ignored.

In 1971, Aldrich et al. were the first to develop a model that predicted per capita EMS demand through causal factors. They used socioeconomic census data from the City of Los Angeles to develop a least-squares linear regression model. Their model used 36 explanatory variables, all of which were determined to be statistically significant. These variables ranged from white population, to total employment, to the number of freeways. The model had an R^2 of 0.927, which represents an extremely good fit.

The next breakthrough came with Siler in 1975. He also used census data to develop a model predicting EMS demand in the City of Los Angeles. Although this sounds very similar to the work of Aldrich, Siler assumed there is no reason to reject the idea of nonlinearity between socioeconomic variables and EMS demand. His final model included four independent variables: ratio of total employment to resident population, ratio of white and married to resident population, ratio of white collar and blue collar employment among female residents, and housing units per area resident. All the explanatory variables were nonlinear, such that their square value was taken, their reciprocal was taken, or their logarithm was taken. The variables were all statistically significant, and the model's R^2 was 0.921, which indicates a great fit.

In 1978, Kvålseth proposed a new method of predicting EMS demand based on socioeconomic, demographic, and health-related predictors in the city of Atlanta, Georgia. He developed both first order and second order models, and interestingly, found that the second order model produced a better fit with only half of the number of variables used in the first order model. The biggest difference in his approach when compared with prior methods was that Kvålseth realized that his independent variables had high multicollinearity, in which using the least-squares regression would be insufficient as it would produce inaccurate results. To handle the complexity of using highly correlated independent variables in his model, Kvålseth used ridge regression instead of the usual least-squares regression method. Ridge regression is a different regression technique, and it partly alleviates the problem of high multicollinearity between the independent variables.

A less popular but alternative and emerging approach to forecasting demand is the use of artificial neural networks, or ANNs. An ANN is a computational model based on the structure and functions of biological neural networks of the human brain. Although nonlinear relationships and multicollinearity can partly be addressed through different regression techniques previously stated, ANNs are a much superior alternative method as a whole. The main trade off is complexity: although ANNs can provide for a more accurate forecasting model, ANNs are more complex and as a result less used in the academic community. A large deal of work and research into ANNs for EMS demand forecasting was done by Setzler (2007).

Based on the existing literature, we believe that using a least-squares regression model is an appropriate way to forecast EMS demand that provides a model that is simple to estimate and use. Therefore, our main objective is to identify which explanatory variables, from those available through the City of Toronto's Open Data Catalogue, can be used to model EMS demand in the different neighbourhoods of the city. Since our model will be estimated based on the 2008 data that is available online, we will use the model to predict the EMS demand for the 2011 year, which can be used for validation once the data for 2011 is made public.

METHODOLOGIES

Having a large set of accurate data is vital in developing a multiple linear regression model. The City of Toronto's *Open Data Catalogue* was used to obtain all the required data. The 2008 EMS demand for each of Toronto's 140 different neighbourhoods was determined as the dependent variable. Possible predictors were collected by obtaining demographic, economic, environmental, health, transportation, and safety data for each of Toronto's 140 different neighbourhoods. After compiling all possible predictors into one data set, there were 127 predictors present. To reduce this numbers, each dataset was examined to determine its eligibility to be used as a predictor for EMS demand. In other words, all possible predictors that had no logical correlation to EMS demand (for example, "Business Licensing"), were removed. After this "sense check" step, only 44 predictors remained.

Stepwise backward elimination procedure was used to take the initial set of 44 predictors and narrow it down to the final model. The final model was estimated using multiple linear least-squares regression. Starting with the initial set of 44 predictors, two variables were removed at a time until the final model was developed. After each step of the stepwise backward elimination, the newest model was evaluated using the various techniques presented below.

Significance test of independent variables

A 95% one-tailed t-statistic test was used to test the significance of each independent variable. A one-tailed test was performed instead of a two-tailed test because the signs of all independent variables was logically reasoned to be positive. At each step in the backward elimination process, up to two variables that failed the t-test were removed. Additionally, variables that had a negative coefficient (and thus negative t-statistics) were also removed due to the stated assumption of positive signage.

R^2 check

At each step, the *adjusted* R^2 of each succeeding model was analyzed and compared to the adjusted R^2 of the preceding model to verify the model's goodness of fit. The adjusted R^2 simply "adjusts" for the number of predictors in the model, providing for a more accurate result.

ANOVA

ANOVA was performed on each succeeding model to calculate the F statistic to determine if the variation in Y (EMS demand) was in fact being explained by the variation in X (the predictors). A 95% confidence level was used for this procedure.

Residuals Analysis

The residuals plots of each successive model were graphed and analyzed. The goal was to obtain normally distributed residuals with constant variance over the range of EMS demand.

Partial F-test

Partial F-tests were performed between *pairs* of consecutive models. This was done to ensure that each of the variables removed for each successive model were in fact *not* usefully contributing to the model. A 95% confidence level was used for this procedure.

Multicollinearity check

After the final model was developed, the correlation matrix between all of the independent variables was obtained and analyzed. A correlation rate of 0.6 or higher was determined to classify high correlation between two independent variables. Variables that had high correlation with other variables were considered for removal, as way to improve the final model and reduce collinearity between independent variables. One variable, number of break and enters, had a significantly high correlation coefficient (0.78) two other variables, hazardous incidents and robberies. Since hazardous incidents and robberies seemed to have a more logical causation relationship with EMS demand, and these two variables did not have high correlation with any other dependant variables, number of breaks and enters was removed to avoid multicollinearity.

RESULTS

The final model was estimated using 6 regressors as explanatory variables, which are summarized in Table 1. We believe that most of the explanatory variables selected for the model are reasonable predictors for ambulance service, either because they directly indicate events that may require EMS (such as hazardous incidents), or because they may be related with increased accidents (road volume).

Table 1 Descriptive statistics for model explanatory variables

Variable	N	Mean	St. Dev.	Variable description
Y	140	1,536	982	Total calls for ambulance service
X_1	140	118	63	Hazardous incidents reported by Toronto Fire Services
X_2	140	29	23	No. of robberies registered in the ECRIME and CIPS databases
X_3	140	4,372	2,033	Average 24 hour collector road volume
X_4	140	234	71	Number of deaths that occur before 75 years of age
X_5	140	9,349	18,491	Total number of jobs for persons that are 15+ years old
X_6	140	638	361	Number of citizens 65+ years living alone

As can be seen in Table 2, there is low correlation between most of the dependent variables, which reduces the probability that issues with multicollinearity may arise. This was achieved because some potential dependent variables in the dataset were eliminated from previous iterations of the model due to high correlation. This will be further discussed in the limitations section.

Table 2 Correlation Matrix for the regression variables

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1						
X_1	0.840	1					
X_2	0.780	0.608	1				
X_3	0.108	0.097	-0.006	1			
X_4	0.255	0.152	0.262	-0.125	1		
X_5	0.723	0.581	0.551	0.097	0.041	1	
X_6	0.484	0.526	0.236	0.063	-0.113	0.161	1

Table 3 summarizes the regression model parameters, as well as all the relevant statistical and goodness of fit information about the model. All the parameters are statistically significant under a 95% level of confidence (assuming a one tailed test, which is appropriate since we expect all these variables to have a positive sign). Further, the goodness of fit of the model is excellent: the adjusted R^2 indicates that over 88% of the variation observed in the dependent variable can be explained by the model. Lastly, based on the P-value of the F statistic, we can say with a confidence level of 95% that the explanatory variables have an effect on the dependent variable.

Table 3 Linear regression model parameters, t and p statistics, and confidence intervals (95% level of confidence)

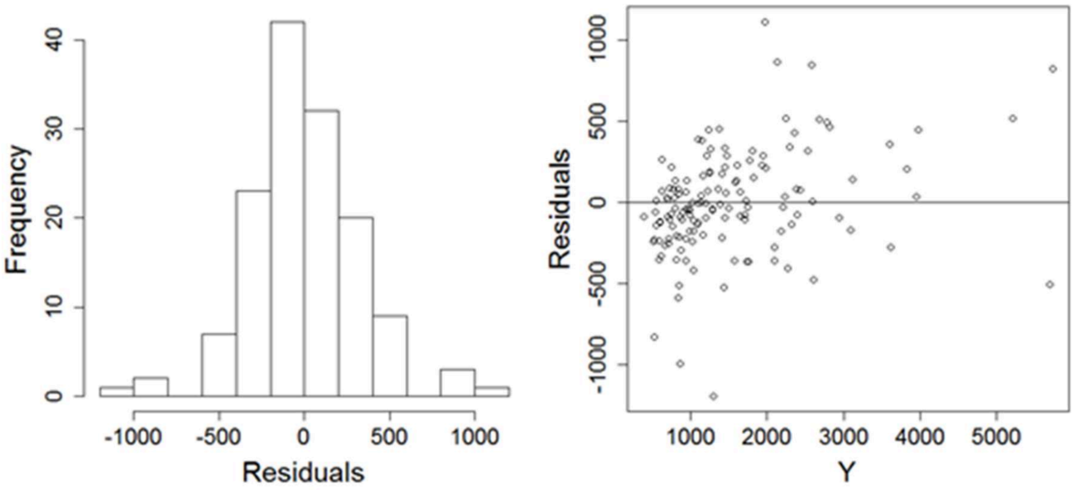
	Parameter (s.e.)	t	p	2.5 %	97.5 %
Constant	-512.259 (133.581)	-3.835	0.0002	-776.477	-248.041
X_1	5.238 (0.713)	7.348	$1.83\text{x}10^{-11}$	3.828	6.648
X_2	14.185 (1.694)	8.372	$6.91\text{x}10^{-14}$	10.833	17.536
X_3	0.025 (0.014)	1.781	0.078	-0.003	0.053
X_4	1.839 (0.428)	4.295	$3.34\text{x}10^{-5}$	0.992	2.686
X_5	0.016 (0.002)	7.978	$6.06\text{x}10^{-13}$	0.012	0.020
X_6	0.521 (0.096)	5.410	$2.84\text{x}10^{-17}$	0.331	0.711
Observations	140				
R^2	0.892				
Adjusted R^2	0.887				
Residual Std. Error	330.387 (df = 133)				
F Statistic	182.371 (df = 6; 133) (p = $2.2\text{x}10^{-16}$)				

The final model equation is as follows:

$$Y = -512.259 + 5.238X_1 + 14.185X_2 + 0.025X_3 + 1.839X_4 + 0.016X_5 + 0.521X_6 \qquad (1)$$

The quality of the model can further be validated by looking at analysis of residuals. A good model should have residuals that are normally distributed and with an equal variance. Figure 1a and Figure 1b show the histogram of the residual frequency as well as a plot of the residuals versus the dependent variable Y. From the histogram, it can be observed that the residuals seem to follow a normal distribution. Further, from the residual plot, it can be observed that overall the residuals seem to have a relatively constant variance. There are a few points that skew the plot, but there is no significant increase or decrease in residual variance as a function of Y. Therefore, the residuals of the model appear to meet the criteria of being normally distributed and having a constant variance.

Figure 1 Residual analysis figures a) histogram of residual frequency and b) residual plot versus dependent variable Y



DISCUSSION

Overall, we feel that our model incorporates logical predictors for EMS demands, and has high statistical significance and goodness of fit. The variables incorporate a diverse pool of predictors: including health, demographics, safety, and transportation metrics. Further, our variables also incorporate several mechanisms for driving EMS demand: from direct-effect mechanisms (such as a hazardous incident occurring that requires an ambulance) to indirecteffect mechanisms (such as increased road volume which in turn increases the probability of a severe accident that requires an ambulance).

To properly assess the effect of each predictor on the EMS demand, we can evaluate their coefficients, and also assess their correlation with the defendant variable. With regards to the coefficients, it is important to look at them in the context of the magnitude of its related variable. For example, X_2 has the largest coefficient (14.185), but its mean value is fairly small (29). In comparison, X_5 has the smallest coefficient (0.016), but its mean value is the largest among the independent variables (9,349). With regards to the correlation between the predictors and the dependent variable, we can see that the only variables with low correlation are X_3 and X_4. This indicates that, although the variables X_3 and X_4 have a statistically significant effect on Y, their relationship with Y may be nonlinear. The remaining dependent variables have high correlation to EMS demand, so a linear relationship is more evident.

Our model has an apparently large constant (-512.259). We acknowledge that this implies that the model is capturing a large percentage of the error in the constant term which is not ideal. However, that does not mean it is a bad model, especially when considering the high goodness of fit of 89%.

There were several variables in the source datasets that initially seemed liked reasonable predictors (such as pedestrian collisions, assaults, murders, fire and fire alarms, and traffic collisions) but which did not show high statistical significance, and thus were excluded in the final model. We believe that this may be due to several factors. Firstly, it is possible that the order in which the variables were eliminated affected the t-test at each stage. To mitigate this, we only removed up to 2 variables at each iteration and attempted to keep the variables that seemed the most logical predictors of EMS demand until the last iterations. Secondly, issues with multicollinearity between these variables may have affected the t-statistics calculated at each iteration of the model. Multicollinearity can cause the standard error of the affected coefficients to be high, which in turn leads to a lower t-statistic. For example, high correlation between traffic volume and pedestrian collisions could cause one of these variables to have a lower t- statistic.

Model comparison

Comparing our model to those in the literature reviewed is challenging due to the large differences in context between the cities and timeframes of these models. The most similar model in the literature is the one developed by Aldrich et al. (1971) who also utilized least-square regression. Although their model includes 36 explanatory variables as opposed to 6, a few of our variables have a similar counterpart in their model (seniors living alone, employment, and road volume). Further, our unadjusted R^2 of 0.892 is comparable to the R^2 of 0.927 in the model developed by Aldrich et al. (1971).

Predicting demand

We utilized our model to estimate 2011 EMS demand as a form of model validation. Unfortunately the 2011 EMS demand has not yet been released; however, data for all 6 explanatory variables has been released for 2011. Using the developed model and 2011 datasets for the explanatory variables, the EMS demand is estimated to have been 289,182. The *actual* 2008 EMS demand was 215,064 – thus the developed model estimates a 35% increase between 2008 and 2011. Once the 2011 EMS demand data becomes available, the model's accuracy will be validated.

Limitations

The regression technique utilized in this paper was ordinary least-squares (OLS) linear regression. Multicollinearity is when two or more variables in a regression model are highly correlated. For an OLS regression model, the case of multicollinearity is not desirable and should be avoided. In our situation, the final model had originally consisted of seven independent variables that were all statistically significant. However, one of those variables (number of breaks and enters) had very high correlation with two other variables. As a result, that variable was removed from the model, and while the adjusted R^2 dropped slightly, this was done to avoid the stated issue of multicollinearity. One way to keep correlated values and maintain a valid model is to use different techniques, like those discussed in the Background section. These can include ridge regression and artificial neural networks.

The residuals of the developed model were normally distributed with constant variance (without any visible patterns or distortions), and the adjusted R^2 was 89%, an extremely good fit. While the residual analysis and adjusted R^2 indicate that a linear model is simply "sufficient", there is no reason to believe that a nonlinear model should not be tested. Nonlinear and higher-order linear models can be tested to see if they prove to be more accurate.

A major limitation is the usage of an old dataset. As previously stated, the dataset used is from 2008, which is latest EMS demand data available. As a result, although the 140 neighbourhoods provide for an observation of 140 instances for the dependent and independent variables, the model was still based on a single year. Additionally, this single year is eight years past, and thus the relations of the predictors could have changed significantly.

CONCLUSIONS

We have estimated a simple multiple linear regression model to predict EMS demand for the 140 neighbourhoods that constitute the City of Toronto, based on the 2008 data provided by the city. We have found that there are six variables that can be used to predict the EMS demand with a high goodness of fit, and that these variables vary significantly in nature. This has a few implications for the city. Firstly, if newer data is made available and our model can be validated, it can be a useful tool for city planners to easily forecast increase in ambulance demand based on expected changes in the predictors. Secondly, the application of our model can be expanded to evaluate whether current EMS dispatch stations are placed close to the neighbourhoods that will have the highest predicted demand.

ACKNOWLEDGEMENTS

We would like to thank Sami Hasnine for his advice and assistance in developing this paper. Further, we would like to acknowledge Hlavac (2015) for developing the stargazer R Package that was used for creating most of the tables in this paper.

REFERENCES

Aldrich, C. A., Hisserich, J. C., & Lave, L. B. (1971). An analysis of the demand for emergency ambulance service in an urban area. *American Journal of Public Health*, *61*(6), 1156–1169.

Hlavac, M. (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables. (Version R package version 5.2.). Retrieved from http://CRAN.R-project.org/package=stargazer

Kvålseth, T. O. (1978). First-and second-order ridge regression models of the rate of demand for emergency medical services. *Applied Mathematical Modelling*, *2*(3), 209–215. https://doi.org/10.1016/0307-904X(78)90010-0

Ministry of Health and Long Term Care. (2016). MOHLTC - EHS - Land Ambulance Response Time Standard. Retrieved December 6, 2016, from http://www.health.gov.on.ca/english/public/Program/ehs/land/responsetime.html

Setzler, H. (2007). *Developing an Accurate Forecasting Model for Temporal and Spatial Ambulance Demand Via Artificial Neural Networks: A Comparative Study of Existing Forecasting Techniques Vs. an Artificial Neural Network*. University of North Carolina. Retrieved from https://books-google-ca.myaccess.library.utoronto.ca/books/about/Developing_an_Accurate_Forecasting_Model.html?hl=es&id=frbi1ssr0qUC

Siler, K. F. (1975). Predicting demand for publicly dispatched ambulances in a metropolitan area. *Health Services Research*, *10*(3), 254.