Journal of Mathematical Sciences: Advances and Applications
Volume 32, 2015, Pages 57-71

FORECASTING THE NUMBER OF STUDENTS IN GENERAL EDUCATION IN UNIVERSITY COLLEGE USING MATHEMATICAL MODELLING

IBRAKHIMJON RAKHIMOV and MANISHA M. KANKAREJ

Department of Mathematics and Statistics
University College
Zayed University
Dubai
UAE
e-mail: Ibrahim.Rahimov@zu.ac.ae
Manisha.Kankarej@zu.ac.ae

Abstract

In this paper, a quantitative forecasting model has been created for studying the dynamics of the student population entering University College in Zayed University, UAE. Admission, curriculum, pre registration, registration, grading, and record management are all supported by the mathematical model in one or the other way. This model explores the projection of the number of students enrolled in General Education (GE). A statistical modelling method is adapted in developing the model. As a case study, we have used the data available on the website and also the data provided by the office of Institutional Research, Zayed University. The model predicts that the number of students in GE has the tendency to increase linearly with the slope of 71.227 and 114.95 for Dubai and Abu Dhabi campuses, respectively, and the limits for 95% confidence interval are (44.01, 98.44) and (87.42, 142.48).

2010 Mathematics Subject Classification: 62J12, 62J05, 62Q99.
Keywords and phrases: student population, enrollment forecasting, linear model, standard error, confidence interval, posterior probability.
Received March 9, 2015

1. Introduction

Forecasting is the process of making statements about the events whose actual outcomes have not been observed. We can also explain it as the estimation of some variable of interest at some specified future date. Prediction is similar but of more a general term. Both refer to formal statistical methods to make the judgments for future. Quantitative forecasting models are used to forecast future data as a function of the past data.

Different modelling methods have been studied by different researchers regarding the forecasting of data according to the availability of the data and the situations fitted by these data. Some researchers used descriptive and explanatory model to forecast the number of students [2]. A study on the enrollment forecasting for an upper division general education using regression analysis is done in Grand Valley State University [4]. One more study is conducted on enrollment forecasting for Community College [5]. A study on Markov chain model was explored to forecast enrollment [6], [7]. Simple moving average, single exponential smoothing and double exponential smoothing, double exponential smoothing are also studied [12].

Predicting the number of students is important for estimating the distributed budget into academic institution, it may contribute the action plan and may be used as information for giving long-term policy. This model also helps the administrator to understand the enrollment pattern and the factors that influences the number of students expected to enroll. Understanding the inflow and the outflow of the students in the department or college is necessary to the university management including the performance measures.

Mathematical modelling is a useful tool to study population dynamics. In the present context, the model of the number of students enrolled in GE in University College can be used for predicting and

understanding the key factor that influences the changes, e.g., past enrollment data, curriculum or the program offered, prerequisite, etc. This model is built from using real data and then determining a mathematical formula with parameters that fit to available data. In particular, using regression analysis one can obtain the values of such parameters. In this study, we mathematically model the number of students in University College for understanding the underlying mechanism about the inflow of the students.

2. Data

The accurate data plays a major role in forecasting and prediction in order to avoid the uncertainty in the forecasts. That is why it is most important that the data has to be accurate for the forecasts.

Regarding our research there are many factors affecting the number of students entering GE program in Zayed University. There are some students who are

- New direct entry;
- New students from Academic Bridge Program (ABP); and
- Continuing GE program.

Though our study reflects the number of these students together, also we found that the number of the students continuing GE is much more than the number of students who are direct entry and coming from ABP every semester. The data collected for this study is the number of students enrolled in Dubai and Abu Dhabi campus every year starting from 2004 till 2014 for GE program. We created a linear trend line for the given data together (means from 2004 till 2014) but the percentage error obtained for this case is determined to be very large as compared to the trend line when the data was considered separately for two independent sets by dividing it into two parts. The causes of this phenomenon can be seen in charts on Figure 1 below which shows that the rate of change is

essentially different before and after 2009. Therefore, we divided the data in two parts and created two separate trend line, which reduced the percentage error to a very large extent.

Table 1. The number of students enrolled in GE

Year	S. No.	Semester	Students enrolled in Dubai	Students enrolled in Abu Dhabi
2004	1	Spring 2004	309	325
	2	Fall 2004	440	384
2005	3	Spring 2005	412	326
	4	Fall 2005	492	437
2006	5	Spring 2006	467	418
	6	Fall 2006	548	496
2007	7	Spring 2007	483	441
	8	Fall 2007	574	476
2008	9	Spring 2008	557	455
	10	Fall 2008	695	517
2009	11	Spring 2009	611	519
	12	Fall 2009	867	683
2010	13	Spring 2010	860	790
	14	Fall 2010	1153	1052
2011	15	Spring 2011	1035	1142
	16	Fall 2011	1351	1407
2012	17	Spring 2012	1247	1387
	18	Fall 2012	1542	1664
2013	19	Spring 2013	1342	1670
	20	Fall 2013	1578	1809
2014	21	Spring 2014	1385	1664
	22	Fall 2014	1598	1796

3. Statistical Model

In this section, we have formulated the number of students in the GE for Dubai and Abu Dhabi, using quantitative forecasting model as it also support a user to achieve a greater efficiency and effectiveness of decision-making. Based on the data collected for the number of students in GE from Spring 2004 till Fall 2014 following line chart have been created.

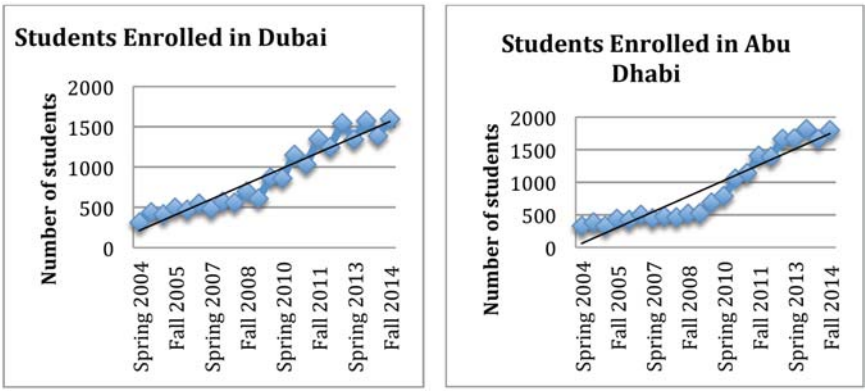

Figure 1. The student population chart for Dubai and Abu Dhabi based on Table 1.

The charts given above shows that between 2009 and 2010 there is a major change in the inflow of the students, that is why the rate of change from 2004 till 2009 is different than the rate of change of 2010-2014. The same change has been noticed in fall semester also.

Figure 1 suggests using simple linear model to describe the growth of the student population in the University College.

$$y = a + bx, \tag{1}$$

where x = semester number and y = number of students.

The parameters of the model a and b are subject to estimate based on the data available.

Here data starts from spring 2004, so spring 2004 corresponds to $x = 1$ and so on. Further the data is divided in part 1 (Spring 2004 to Spring 2009) and part 2 (Fall 2009 to Fall 2014). Similar partitions are done for Abu Dhabi as well. Later on in Section 5, we demonstrate that there is a strong evidence to support the hypothesis that slopes for part 1 and 2 are essentially different based for the data available. The standard least squares technique gives the following equations. We also provide the coefficient of determination for each regression line.

- Dubai Part 1 : $y = 28.591x + 336.45$, $R^2 = 0.8179$.
- Dubai Part 2 : $y = 71.227x + 841.55$, $R^2 = 0.79571$.
- Abu Dhabi Part 1 : $y = 18.091x + 327.27$, $R^2 = 0.77445$.
- Abu Dhabi Part 2 : $y = 114.95x + 679.73$, $R^2 = 0.90839$. (2)

As all the slopes are positive, which shows that, the number of students is growing linearly. The coefficient of determination R^2 (it is a number that indicates how well data fit a statistical model) shows the portion of the total variation in the number of students that is explained by its relationship with semesters or years. Later we provide the value of the standard error (which is the square root of sum of squares of the errors divided by $n - 2$, where n is the number of observations). We denote A_i = *Actual value*, F_i = *Fitted value*. Then we can write Error = $A_i - F_i$, percentage error = $|A_i - F_i| / A_i * 100$.

The critical value of t-distribution with $n - 2 = 9$ degrees of freedom for 95% confidence interval and is $t_{\frac{\alpha}{2}} = 2.262$ for each part, what concerns the standard error, we calculate it as

$$S_{\epsilon} = \sqrt{\frac{\sum_{i=1}^{n} (A_i - F_i)^2}{n - 2}}.$$

The limits of the 95% confidence interval of y for given particular x is

$$\hat{y} \pm t_{\frac{\alpha}{2}} S_{\in} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Based on Equations (2), we also calculated the predicted number of students for the coming semesters. The obtained results are provided in Tables 2 and 3 for Dubai and Abu Dhabi campuses. In the given tables, last two columns indicate the 95% confidence interval limits for the predicted number of students corresponding to each semester.

Table 2. Predicted student population and 95% confidence interval limits for Dubai part 2

S. No.	Students enrolled in Dubai	Predicted value	Percentage error	Lower limit	Upper limit
1	867	912.78	5%	536.83	1197.17
2	860	984.00	14%	540.08	1179.92
3	1153	1055.23	8%	841.29	1464.71
4	1035	1126.46	9%	729.28	1340.72
5	1351	1197.69	11%	1048.94	1653.06
6	1247	1268.91	2%	946.17	1547.83
7	1542	1340.14	13%	1239.94	1844.06
8	1342	1411.37	5%	1036.28	1647.72
9	1578	1482.59	6%	1266.29	1889.71
10	1385	1553.82	12%	1065.08	1704.92
11	1598	1625.05	2%	1267.83	1928.17

Table 3. Predicted student population and 95% confidence interval limits for Abu Dhabi part 2

S. No.	Students enrolled in Abu Dhabi	Predicted value	Percentage error	Lower limit	Upper limit
1	683	794.68	16%	349.01	1016.99
2	790	909.63	15%	466.38	1113.62
3	1052	1024.58	3%	736.68	1367.32
4	1142	1139.53	0%	832.74	1451.26
5	1407	1254.48	11%	1101.44	1712.56
6	1387	1369.43	1%	1082.68	1691.32
7	1664	1484.38	11%	1358.44	1969.56
8	1670	1599.33	4%	1360.74	1979.26
9	1809	1714.28	5%	1493.68	2124.32
10	1664	1829.23	10%	1340.38	1987.62
11	1796	1944.18	8%	1462.01	2129.99

It is natural that we can give the forecasted number of students for each of the data sets using model (1). Following table gives forecasted values for Dubai and Abu Dhabi in Spring and Fall 2015.

Table 4. Forecasted number of students for Dubai and Abu Dhabi in Spring and Fall 2015

Semester	Dubai	Abu Dhabi
Spring 2015	1696	2059
Fall 2015	1767	2174

4. Confidence Intervals for the Parameters of the Model

Here we construct 95% confidence intervals for the slope and intercept of the model (1). For this, we first find the standard errors of $\hat{a}$ and $\hat{b}$ by using the following formulas:

$$S_a = S_\epsilon \sqrt{\frac{\sum x_i^2}{n\sum (x_i - \overline{x})^{2'}}}, \quad S_b = \frac{S_\epsilon}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}}.$$

Next, we used formulas

$$\hat{a} \pm t_{\frac{\alpha}{2}} S_a, \quad \hat{b} \pm t_{\frac{\alpha}{2}} S_b.$$

To construct the confidence intervals for a and b, respectively.

As it was mentioned before that $t_{\frac{\alpha}{2}}$ is the critical value of the t-distribution with $n-2$ degrees of freedom and for 95% of confidence interval with $n-2=9$ degrees of freedom, $t_{\frac{\alpha}{2}} = 2.262$. The standard calculations results the following Tables 5 and 6. Columns 2 and 3 correspond to part 1 and part 2 of Dubai, whereas columns 4 and 5 correspond to part 1 and part 2 of Abu Dhabi.

Table 5. Confidence interval for a

	Dubai		Abu Dhabi	
$\hat{a}$	336.45	841.55	327.27	114.95
S_ε	47.16	126.17	34.13	127.63
S_a	30.50	81.59	22.07	82.53
Lower limit	267.46	656.99	277.34	493.04
Upper limit	405.44	1026.11	377.20	866.42

Table 6. Confidence interval for b

	Dubai		Abu Dhabi	
$\hat{b}$	28.591	71.227	18.09	114.95
S_b	4.50	12.03	3.25	12.17
Lower limit	18.42	44.01	10.73	87.42
Upper limit	38.76	98.44	25.45	142.48

5. Assessing the Fit and the Significance of the Relationship

The primary measure of the fit of a regression model is the coefficient of determination R^2. It shows how much variation in y the regression line explains on x. We can see from Equations (2) that the percentage of the variation in the number of students y explained by the change of semester x varies from 77% to 90% (Abu Dhabi part 2 data). This allows us to estimate the fit of the model as good.

We now conduct a hypothesis test to determine whether there is a significant linear relationship between independent variable x and a dependent variable y. If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables. We test $H_0 : b = 0$ against alternative $H_a : b \neq 0$ with the level of significance $\alpha = 0.01$. For this, we first find the standard error of the slope from

$$S_b = \frac{S_{\in}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

The test statistic given by $t = \dfrac{\hat{b}}{S_b}$ has a t distribution with $n - 2$ degree of freedom.

The following table gives the values of these variables for 2 parts of the data.

Table 7. The values of S_b and test statistic

	Dubai		Abu Dhabi	
$\hat{b}$	28.59	71.22	18.09	114.95
S_b	4.50	12.03	3.25	12.17
$t = \dfrac{\hat{b}}{S_b}$	6.36	5.92	5.56	9.50

For $\frac{\alpha}{2} = 0.005$, we have the rejection region $(-\infty, -3.25) \cup (3.25, \infty)$. We see that in all cases the value of the test statistic falls into rejection region. Hence, we reject the hypotheses $H_o : b = 0$ with only 1% type 1 error which shows the slope in our model is significant in all cases.

We now conduct a hypothesis test to check that the slopes related to data before and after 2009 are significantly different. For this, we denote by b_1 and b_2. The slopes related to part 1 and part 2 data. First we consider the data from Dubai. We test $H_0 : b_1 = b_2$ against $H_a : b_1 \neq b_2$ with the level of significance $\alpha = 0.01$. To conduct our test, we use the following test statistic (see [3]):

$$z = \frac{(\widehat{b_1} - \widehat{b_2})}{\sqrt{(\widehat{SEb_1}^2 + \widehat{SEb_2}^2)}}. \tag{3}$$

We find for the Dubai data that $z = 3.28$ and $Z_{0.005} = 2.576$ which shows we reject the null hypotheses. Thus, we have a strong evidence to conclude the slopes related to the data before and after 2009 are significantly different. The same calculations for the data from Abu Dhabi show that $z = 7.7$ and again we reject the null hypotheses. This justifies considering the data before and after 2009 separately.

6. Some Applications of the Model

Table 4 gives the forecasted number of students in Spring and Fall 2015 for Dubai and Abu Dhabi campuses. Of course one can construct confidence intervals for the predicted number of students or for the mean of predicted number of students based on the data available. Here we discuss about posterior predictive distribution of the number of students. We use the simple fact that the posterior predictive distribution of $\widehat{Y}(12)$ is t-distribution with $n-2$ degrees of freedom, mean is given by the predicted number of students and the standard deviation is calculated by

$$SE(\widehat{Y}(12)) = S_{\epsilon}\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \quad (4)$$

We can easily see from Table 4 that for Spring 2015

$$E\left(\widehat{Y}(12)\right) = 1696, \quad E\left(\widehat{Y}(12)\right) = 2059,$$

for Dubai and Abu Dhabi, respectively.

First we consider posterior predictive distribution for Dubai. To calculate the standard error of $\widehat{Y}(12)$ based on formula (4), we use the fact that for the second part of the data $S_{\epsilon} = 126.1724$ (see Table 5) and calculate the value of the square root as 1.0215. This gives us that $SE(\widehat{Y}(12)) = 128.8857$. Thus we can conclude that

$$P\{\widehat{Y}(12) \le x \mid data\} = t_9(1696,\ 128.8857;\ x), \quad (5)$$

where $t_n(c,\ d;\ x)$ denotes t-distribution with degrees of freedom n, mean c and the standard deviation d. By similar analysis, we find that for Abu-Dhabi data

$$P\{\widehat{Y}(12) \le x \mid data\} = t_9(2059,\ 130.375;\ x). \quad (6)$$

We now consider some numerical examples. From the data Table 1 for Dubai, we can find the percentage growth of the number of students from Spring to Fall 2014 is

$$\frac{1598-1385}{1385} * 100 = 15.379\%.$$

Say approximately 15% growth. To find the probability that the growth will be more than 15% from Fall 2014 to Spring 2015, given the Dubai data, we can use formula (5) and the table for the t-distribution. For this, we first estimated the number of students when 15% grows as $1598 + 240 = 1838$. Hence, we have

$$P\{\widehat{Y}(12) > 1838 \mid data\} = 1 - t_9(0,\ 1;\ 1.102) = 0.1495.$$

This allows us to suggest that there is about 15% chance that the number of students will increase more that 15% from the Fall 2014 to Spring 2015 in Dubai.

We now consider the similar problem for Abu-Dhabi data. Again from the data Table 1, we can see that in Abu-Dhabi the percentage growth of the number of students from Spring to Fall 2014 is

$$\frac{1796 - 1664}{1664} * 100 = 7.9327.$$

What is the probability that growth will be more than 8% from Fall 2014 to Spring 2015, given the Abu Dhabi data. Using formula (6) this time, we found that

$$P\{\hat{Y}(12) > 1796 + 144 \mid data\} = 1 - t_9(0, 1; -0.9127) = 0.8074.$$

Thus, we can suggest that there is about 80% chance that the number of students will increase more that 8% next semester in Abu-Dhabi. In conclusion, we note that similar posterior predictive distributions can be derived for the slope $\hat{b}$ and intercept $\hat{a}$ of the model and calculate various posterior probabilities related to the parameters.

7. Conclusion

Enrollment forecasting is both an art and a science. Statistical techniques have improved significantly but still hard to come by and many events can influence actual path of prediction. The most effective models emerge when changes in trends can be explained by the changes in the surrounding competitive higher education environment or changes within the institution. Nevertheless, as we pointed out earlier, the model gives some useful information on how to forecast the number of students in the future. The model gives linear relationship between the number of students and calendar years. Qualitatively, this model indicates increasing of the number of students. This shows that the data are likely to be characterized by linear curve.

Using these models, we can predict the student populations for next calendar year, say 2015 and further. Predicting the number of such students for next calendar year is useful for education strategic management and planning of the department such as preparing enough teachers for students coming up next year. In addition, we can use such predictions for course schedule managements. For example, the department can make a decision about the sections or classrooms for the students who will be arriving. Being known about the growth, the institution becomes better informed about the patterns of the demand and it facilitates the formulation of the investment decisions and strategies at all levels in the system.

Finally, the model modification along with an alternative method can be considered for future study. For example, instead of using linear fit, one can assume logistic regression or other nonlinear functional forms. Hence, the transition from year to year could be considered. Overall, it is important to appreciate that enrollment forecasting is an iterative and collaborative process.

References

[1] G. G. Woodworth, Biostatistics: A Bayesian Introduction, John Wiley & Sons, Inc., Hoboken, 2004.

[2] Nichaphat Patanarapeelert and Klot Patanarapeelert, Forecasting number of students in University department: Modeling approach, Open Journal of Applied Sciences 3 (2013), 293-297.

[3] R. Paternoster, Robert Brame, Paul Mazerolle and Alex Piquero, Using the correct statistical test, for the equality of regression coefficients, Criminology 36(4) (1998), 859-866.

[4] S. Choudhuri, C. R. Standridge, C. Griffin and W. Wenner, Enrollment Forecasting for an Upper Division General Education Component, Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference, Milwaukee, 10-13 October 2007, pp. T3E-25-T3E-28.

[5] S. Guo, Three Enrollment Forecasting Models: Issues in Enrollment Projection for Community Colleges, Presented at the 40th RP Conference Asilomar Conference Grounds Pacific Grove, California, May 1-3, 2002.

[6] J. M. Fraser, S. Djumin and J. J. Mager, The University as Educational Lab, in Proc. 1999 American Society for Engineering Education Annual Conference and Exposition, 1999.

[7] J. C. Segura-Ramirez and W. Chang, Using Markov Chain and Nearest neighbor Criteria in an Experience Based Study Planning System with Linear Time Search Scalability, in Proc. 2006 IEEE International Conference, Waikoloa Village, HI, (2006), 395-403.

[8] D. Y. Young and L. J. Redlinger, Modeling Student Flows through the University's Pipelines, Proceedings of the 41st Forum of the Association for Institutional Research, Long Beach, 5 June 2001, pp. 1-13.

[9] J. D. Logan and W. R. Wolesensky, Mathematical Methods in Biology, John Wiley & Sons, Hoboken, 2009.

[10] M. F. Triola, Elementary Statistics, Pearson Education, Inc., Boston, 2004.

[11] R. Peck, C. Olsen and J. L. Devore, Introduction to Statistics and Data Analysis, Brooks/Cole Cengage Learning, Boston, 2012.

[12] R. Q. Lavilles and M. J. B. Arcilla, Enrollment forecasting for school management system, International Journal of Modeling and Optimization 2(5) (2012), 563-566. doi:10.7763/IJMO.2012.V2.183

■