2012 South Central SAS Users Group Conference
Houston, Texas -- November 5-6, 2012

Predicting Student Enrollments Using SAS

Harjanto Djunaidi[1]

Abstract

Data-driven strategic decisions have been utilized and used many years in other areas than in education. For example, most financial or manufacturing companies have used simulation, mathematical programming and statistical approaches to improve their ability to make sound strategic investment and resource allocation decisions. Recent dynamic changes in the US and global economy, federal government financial assistance regulations as well as in the competitive environments where higher education institutions are operating have forced colleges and universities in the US to find new operational paradigms which can be applied to optimize their resources. The *Institutional Research Intelligence* (IRI) is the next needed concept which offers future tools to provide vital information to the administrators. One possible application of the IRI concepts is to use statistical analyses such as logistic regression model to predict student yield. This new approach will max out the enrollment outcome more efficiently. IRI provides the Office of Students Recruitment and Enrollment the ability to predict how many students who were offered admissions will turn them down. The logistic regression model generates the probability which can be used in the decision making process. Therefore, administrators might be able to affect the enrollment outcomes, allocate resources more efficiently and make better strategic decisions on financial award, class room and course management as well as resident hall assignments. IRI provides vital information in the process of making admission decision in a real time, right after the applicants transmitted their credentials electronically. IRI helps decision makers at higher education institutions to outsmart their competitors while delivering the best quality of services to their students.

[1] Heidelberg University—Institutional and Market Research Office, Tiffin, Ohio 44883.

Introduction

Perhaps, only a handful people have expected the 2008 financial crises will ever happen in the US and the world. When the bankruptcy news of Lehman Brothers' broke out the financial market communities from Shanghai to Frankfurt were stunned. The chain reactions followed by the collapse of other big banks have forced the US government to intervene in order to prevent further chaos in the global financial market. Will the same crises happen in the education industry?

With the students' loan default on the rise, fewer jobs are available to absorb the college graduates, less availability of taxpayers' money to fund higher education, increasing excess supply of higher education services and the newly introduced federal funding regulations known as SAP criteria are the ingredients for a perfect storm for those who operate under BAU mindset. As shown in Table 1 during the period of 2007 to 2011, the average enrollment growth in the US has increased by 7.25%. Though the average growth in each state in the US has shown a favorable number, the magnitude is different from one state to the other. Twelve states have experienced a double digit enrollment growth, while growth in 38 states was below the national average. A more meaningful picture can be generated by subtracting the average state number with the national average growth (7.25%) as shown in the last column of the same table. Using this comparison, one notices that the positive growth only occurred in 22 states and others experienced an unfavorable growth. Comparing the state and national growth data though is a more conservative approach, it potentially provides additional information. For example, a college administrator may want to know if his or her institutional performed above or below the national average which can be used as the base of comparison in the decision making process. The aggregate information shown in Table 1 only paints a bigger picture and it may have less value for an individual college. It does not provide strategic information that can be used in the decision making process. Given the fact that the growth in 38 states was below the national average, it may be the interest of policy and colleges' decision makers to find out the reasons behind the trends. For example, one may have the interest to know which players have contributed to the positive growth in a particular state. Or what type of services that have been offered by the successful institutions to drive the favorable number? As one can see, Connecticut has experienced a tremendous growth in 2007 (46.71%). The published data showed that one institution which offered courses and degrees online contributed to a tremendous growth in that year. The data showed that its enrollments have increased from 3 to 111 students resulted to almost 3700% growth increase. If this outlier is removed, it certainly will show different picture[2]. Readers need to be cautious when making any inference based on the information.

To increase student enrollments always be the interest of the administrators. If student enrollments keep increasing overtime, professionals in the industry may not need to concern too much. Recent supply increase and possible decrease in the funding availability for college students are the two most important structural changes that have occurred recently. Using a comparative static equilibrium analysis, one

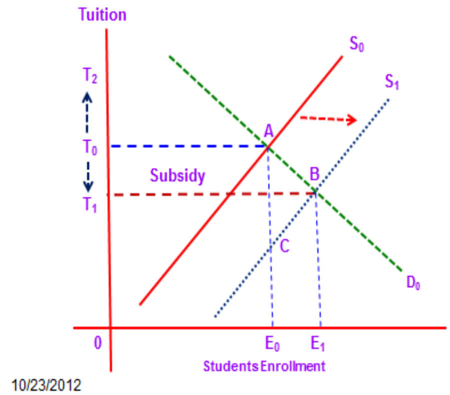
[2] Complete data are available at http://www.IRIntelligence.org.

might be able to analyze the potential impacts of such increase in the supply as shown in Figure 1. Theoretically, increasing supply for the services will have negative effects on the price for acquiring the college education. It supposed to decline from T_0 to T_1. However, this never happened in the past, at least for now. In fact, annual college tuition has increase by $764.21 per year (*Source: Harjanto Djunaidi: Institutional Research Intelligence: Go beyond Reporting, upcoming book, summer, 2013*[3]).

As shown in Figure 1, after the supply has increased and in order to maintain the level of enrollment and the tuition at the original level (E_0 and T_0, instead of at T_1), the industry may need to get a subsidy from the government. The different between T_0 and T_1 shows the amount of subsidy needed to enroll ***one*** additional student. Most of the private schools will try to fill the gap by asking more generous support from the alumni or to take commercial loan to cover the deficit and to keep up with increasing cost of operation. The 2008 economic crises have caused less donations were collected and it has a direct impact on the ability of a school to maintain its level of operation. Higher education institutions which are managed based on the old paradigm may not be able to survive in the new competitive environments. A new mindset, a new way of thinking are needed to cope with the new reality. IRI offers the almost ideal solution for that purpose.

[3] The electronic version of the book is planned to be launched in the end of November, 2012 and it can be accessed through the following website: http://www.IRIntelligence.org

Figure 1 – Increasing Supply: Impact on Competition Tuition and State Subsidy



Situations are getting bleaker when one considers the reality that the rate of supply change for education services has increased tremendously in the past years because of the technological change. When regulators permit for-profit higher education institutions enter the industry, it further pushed the supply curve to the right/outward from S_0 to S_1. If one adds the impact of declining student's loan or 2011 policy changes on Title IV financial assistance known as SAP[4] into the analyses, one might be able to see the compounding effects on demand for higher education. When funding for higher education institutions (HEIs) decreases, then demand for education theoretically will decrease further. For example, if less student loan is available then the impact is pretty obvious. A portion of the college applicants might not be able to

[4] The government has introduced the SAP (Satisfactory Academic Progress) criteria concerning Title IV financial assistance. Section § 668.34 stated that "an institution must establish a reasonable satisfactory academic progress policy for determining whether an otherwise eligible student is making satisfactory academic progress in his or her educational program and may receive assistance under the title IV, HEA programs". This new regulation was in effect starting July, 1, 2011,

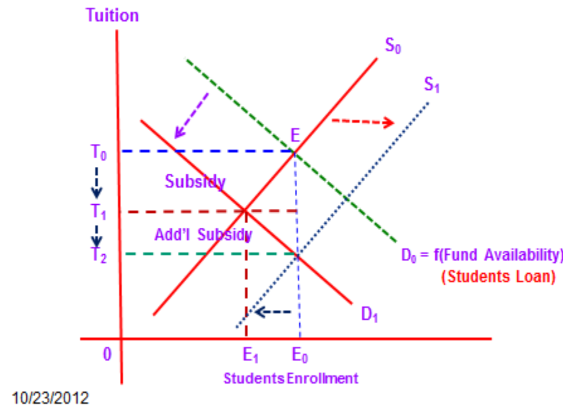
take classes without some sort of government's assistance. This situation causes demand curve to shift further down from D_0 to D_1 as shown in Figure 2. If the policy makers or administrators want to keep the level of student enrollment at E_0 after the change has occurred, then *additional* per unit subsidy is needed as measured by $(T_1 – T_2)$. This simple comparative static analysis shows that the US colleges and universities need to work extra hard to keep the ship from sinking. There is no other choice for both private and public universities /colleges to survive without making drastic strategic changes. Given the new facts, there are strong interests from some of the players to find new and better ways to manage their institution. There is a clear evidence that shows the top private universities and colleges are pretty aware of the situation while smaller one are still operating as usual (BAU—business as usual).

Objective

The purpose of this paper is to show how HEIs can operate more efficiently by applying the IRI concepts and other more rigorous tools such as statistical analyses to cope with recent phenomenal changes in the competitive environments.

The IRI approach helps HEIs to increase effectiveness of their student enrollment management and to improve other students' metrics such as retention and graduation rate. Managing student enrollments efficiently have played a pivotal role in the past. However, its role is even more important today than ever for reasons which have been described above. This paper shows and demonstrates such applications of the IRI[5] new concept. The benefit of using this approach will also be presented and discussed.

Figure 2 – Decreasing Funding and Increasing Supply: Impact on Government Subsidy



Data

There are 100 observations used to derive and show the example of using logistic model to manage student enrollments more efficiently[6]. This is a

---

[5] The IRI concept was first developed by *Harjanto Djunaidi* in 2010. The concept was first introduced and presented at 2012 North Carolina Community College System Annual Conference in Raleigh, NC on October 8, 2012 and at 2012 South Central SAS Users Group Annual Conference in Houston, TX on November 5-6, 2012. Applications of IRI to cope and adapt with recent dynamic changes in the industry and to increase the efficiency and effectiveness of students' recruitment, persistent and graduation rate are among many chapters discussed in his upcoming book—*Institutional Research Intelligence: Go beyond Reporting*. The electronic version of the book will be available on the following website: http://www.IRIntelligence.org. Author's email: harry.djunaidi@IRIntelligence.org

[6] Ideally, more observations are needed to estimate a credible model.

created data set and it does not represent any institution. There are 14 variables originally, but only 6 explanatory variables used in the final estimation. These variables are:

X3= Students' AP English test scores.
X6= Students' AP Chemistry test scores.
X7= Students' ACT Science sub-scores.
X10= Students' ACT Mathematic sub-scores.
FAward = Offered financial award package.
FC=Family contribution.

The Model

The logistic regression is applied to estimate the effects of a set or a group of explanatory variables (in the final model, this study has 6 independent variables (IVs) as shown above) on the DV which has two possible values (either 0 or 1). The target group (students who ***turn down*** admissions offer) will be coded as "1". The reference group (students who ***accept*** the admissions offer) is coded as "0". The probability of applicants to turn the offer down equals to the odds of the exponential function of the linear regression equation (α + βx) as shown in equation (1). Equation (2) is known as the Logistic/Logit function. It represents the log odds or natural logarithm of the odds and it serves as a link function between the odds and the linear regression equation. Equation (3) shows the odds that students may turn down the admission offer equals to the exponential function of the linear regression equation.

$$(1)\ \pi(x) = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}$$

$$(2) Ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$$

and

$$(3)\ \frac{\pi(x)}{1-\pi(x)} = e^{(\alpha+\beta x)}$$

π (x) is the probability of an event x (“the odds”) has a value equals to 1. In this study x represents applicants’ final decision to turn down admissions offer (x=1). The parameter α and β are intercept and regression coefficient, respectively. Ln and e, respectively are natural log and exponential function.

Knowing each applicant’s probability [π (x=1)] to turn down the admission offer ***in advance*** gives the Students Enrollment Office (SEO) or Admission Office (AO) the opportunity to sway the applicants’ unsettled decision toward more incline to accept the offer through different instruments. For example, the office may offer a more attractive financial package to the promising applicants. On the other hand, the school may not need to commit more resources to recruit and to bring a weaker group of students who have higher probability not to accept the offer.

A successful students’ recruitment has a direct impact on the school’s financial situation, the ability to predict the probability that certain group of applicants will or will not accept admissions offer are becoming more important than ever for the strategic planning purposes. The common practices that most schools have been doing in the past were to offer admission to as many qualified students as they can with the expectation that certain portion of these admitted students will go somewhere else. This strategy may work in the past, but it is not a cost effective approach. With recent declining in the recruitment budget and other resources, the AO needs to work smarter and more efficiently. IRI provides the additional ability to the AO or SEO to affect the outcome toward a more favorable for the school.

Result and Discussion

The results of applying the logistic regression model to predict student enrollment are presented in Table 2. Most of the IVs are significant either at a one or five percent confident level. The concordance or C-statistic measures how well the model is able to discriminate between observations at different levels of the outcome. Hosmer and Lemeshow consider c values of 0.7 to 0.8 to show acceptable discrimination. C values of 0.8 to 0.9 indicate excellent discrimination while any values greater or equal to 0.90 shows outstanding discrimination. The C statistics value is 0.899 which falls under the threshold for excellent or outstanding discrimination power. It means the model is capable to differentiate the student population into two groups—those who might turn down or accept admission offer.

Both the financial awards (F_Award) and family contribution (FC1) have a correct negative sign. The negativity of financial award shows that offered admission applicants may have expected a higher award than what have been offered. Had the AO knew that a certain group of applicants weight financial award heavier than any other factors, then AO office might be able to change the outcome (intervention) by offering more scholarship money or at least match other institutions' offer of financial aids. *Without the application of IRI, the AO may have a little clue or does not have any clue at all* of what strategic action needs to be applied (at the individual applicant level) in order to keep the most talented and strongest applicants to accept the offer.

As shown in Table 2, each of the ACT sub-scores in the logarithm form is significant and has a positive sign. This means the number of students who decline the offer will increase as their ACT sub-score get higher. The most talented and strongest applicants have the perception that competitors' offers are more attractive ***and*** that they may have better reputations. Therefore, the applicants more prefer to accept the other offers. If the applicants' reason to turn down the admission is because of a less attractive financial package, then this becomes a short-run enrollment problem. It is easier to be solved and fixed. However, the institution is facing more serious problems if rejection is due to non-financial factors such as the school's reputation or other non-price/tuition/money related matters.

Table 2 - Logistic Model for Students Enrollment

Variable	Estimated	Chi-square	Pr>Chisq	Explanations
Intercept	-27.5709	11.133	0.0008	A constant
LX3	4.3959	10.968	0.0009	Log Students' AP English test score
LX10	5.6274	9.333	0.0023	Log students' ACT mathematics test sub-score
LFCI	-1.4177	7.056	0.0079	Log of family contribution
LX6	2.9697	5.702	0.0170	Log Students' AP Chemistry test score
LX7	1.6420	5.257	0.0219	Log students' ACT science test sub-score
F_AWARD	-0.0002	3.698	0.0545	Financial aids award
C-Statistics = 0.899				

The IRI surely helps both the SEO and AO to operate smarter and more efficiently in their effort to increase students' enrollment. Applying students' recruitment strategy using IRI potentially helps the institution to increase students' retention and graduation rate as well. As previously discussed, IRI helps the school's score card to improve over time. It surely is better than any other schools that operate under the BAU mindset. The potential benefit of applying IRI is shown visually by the ROC graph.

ROC Graph

The ROC (Receiving Operating Characteristic) is a visual presentation of the gain or the benefit of using of IRI's concepts in decision making process. It visually shows the effects of implementing the model on the ability of the concerned parties to identify applicants who may not accept the admissions offer ahead of the game. The horizontal axis is labeled as (1- Specificity) and the vertical axis is called Sensitivity. Both (1 –Specificity) and Sensitivity have a value between 0 and 1.

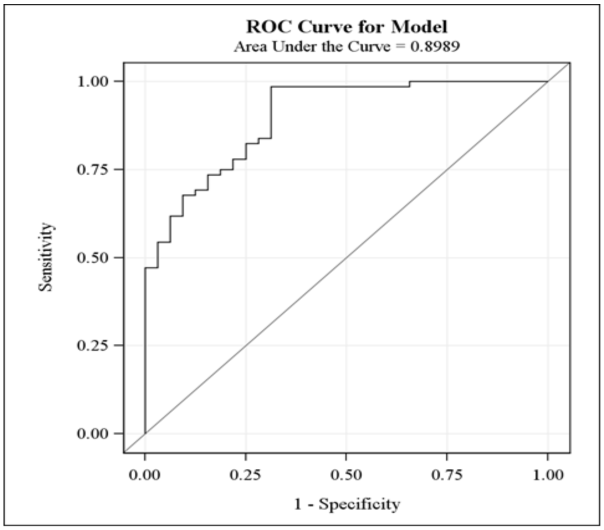
As previously discussed, the dependent variable in the logistic modeling has two values. This value equals to "1" if students decline an offer [turn down (TD)]. Otherwise, it equals to "0" i.e., if students accept the offer (ACO). In such a case, there are four possible combinations of events. If the model prediction is TD and the actual outcome is also TD, then one might call it CTD (correct TD). However, if the actual outcome is ACO or accept admission offer, but the prediction is TD then it is said false TD (FTD). Likewise, if the predicted and actual outcome matches in the case of ACO, then it is called correct ACO (CACO). Otherwise, it is called false ACO (FACO). One can explain this concept more easily using a 2by2 contingency table as shown in Table 3 below.

The ROC graph as shown in Figure 3 can be drawn by plotting the relationships between FTDR and CTDR as x and y axes, respectively. This graph represents the relative trade-offs between these two rates. The y (Sensitivity) axis represents CTDR and the x (1 -Specificity) axis depicts the FTDR. The (0, 1) coordinate of the ROC graph shows a perfect correct prediction of applicants who turn down offer and at the same time there is no FTD prediction i.e., a perfect model.

		Table 3 - Applicants Yield Contigency Table	
		Actual Outcome	
		Turn Down Offer (TD)	Accept Offer (ACO)
Prediction	(TD)'	Correct TD =40	False TD =30
Outcome	(ACO)'	False ACO =20	Corrcet ACO =30
		Total =60	Total ACO=60
CTDR	66.67%	CTDR=Correct Turn Down Rate	
FTDR	50.00%	FTDR=False Turn Down Rate	

This point is also known as perfect classification. The diagonal or tangent 45 degree line (which is equal to 1) is also known as no-discrimination line represents a random guess of which applicants will turn down admission offer.

Figure 3 – Students Enrollment ROC Graph



The area above the diagonal line represents the ability of the model to predict the number of TDs; which certainly is better than the random guess. Perhaps, a better way to interpret the tangent 45 degree line in this case as the ability of the administrators to predict randomly how many students will turn down the admission offer (***random guess***). An example of a random guess is to send out as many offers possible to the strongest candidates on the list with the expectation that some of them will turn them down. IRI works differently. It provides new insights and information more accurately. It helps SEO or AO the ability to affect the enrollment outcomes more favorably toward the institution. This new approach adds the ability to predict and manage the admissions outcome at the individual applicant level instead of at the general applicants' pool. IRI provides the ability and strategic information before the offer of admissions and award package are sent out. The areas under the ROC graph indicate the probability of the model to correctly rank the pair of (CTD, FTD). This example demonstrated the clear benefit of applying the approach in making strategic decisions such that the institutions might be able to increase their operational efficiency. IRI helps HEIs to outsmart their competitors while delivering the best quality of services to their students.

References

Harjanto Djunaidi., *Institutional Research Intelligence: Go Beyond Reporting*. Upcoming Book. Summer 2013.

Harjanto Djunaidi., Predicting Students Enrollment Using SAS. *2012 South Central SAS Users Group Conference*. Houston, Texas, November 5-6, 2012.

Harjanto Djunaidi., Data-driven Decision: Managing Higher Education Institutions more Effectively. *2012 North Carolina Community College System Conference*. Raleigh, October 7-9, 2012.

Harjanto Djunaidi., Institutional Research Intelligence: Go Beyond Reporting. *2012 South Central SAS Users Group Conference*. Houston, Texas, November 5-6, 2012.

Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. New York, NY: John Wiley & Sons, Inc; 2000.

Logistic Regression Internet Access: http://en.wikipedia.org/wiki/Logistic_regression

National Center for Education Statistics. http://nces.ed.gov/ipeds/datacenter/

SAS/STAT® 9.3

34 CFR 668.34-Satisfactory academic progress. Internet access: http://www.ifap.ed.gov/qahome/qaassessments/sap.html

Table 1 - Undergraduate Enrollment (Unduplicated Head Count) Growth by State							
	Enrollment Growth Year					Overall	Compare Growth
State Name	2007	2008	2009	2010	2011	Average	with US Average
Connecticut	46.71%	6.75%	6.50%	7.55%	3.50%	14.20%	6.96%
Kentucky	21.59%	12.70%	12.95%	18.98%	2.46%	13.74%	6.49%
Louisiana	25.75%	5.54%	7.16%	18.54%	6.67%	12.73%	5.48%
Utah	26.14%	2.08%	11.45%	12.43%	11.54%	12.73%	5.48%
New Jersey	10.87%	9.09%	13.16%	15.90%	14.12%	12.63%	5.38%
Nevada	5.38%	6.10%	13.85%	24.03%	7.87%	11.45%	4.20%
Florida	12.77%	9.05%	14.56%	15.78%	3.64%	11.16%	3.91%
Delaware	12.02%	8.89%	4.84%	14.27%	12.50%	10.50%	3.26%
Indiana	8.43%	1.42%	24.22%	15.22%	2.34%	10.32%	3.08%
Arizona	4.88%	12.66%	12.88%	17.87%	2.98%	10.26%	3.01%
West Virginia	1.50%	-2.86%	12.77%	27.68%	10.18%	9.86%	2.61%
South Carolina	10.26%	6.82%	17.74%	12.06%	0.94%	9.57%	2.32%
Maryland	6.71%	7.29%	10.59%	16.55%	6.64%	9.56%	2.31%
California	7.65%	8.78%	15.22%	11.63%	4.51%	9.56%	2.31%
Oklahoma	16.74%	5.38%	2.35%	12.07%	11.03%	9.52%	2.27%
Hawaii	-1.86%	5.81%	5.27%	16.03%	21.36%	9.32%	2.08%
Georgia	9.30%	5.79%	11.88%	13.90%	2.51%	8.68%	1.43%
Ohio	8.82%	6.19%	14.26%	11.78%	1.96%	8.60%	1.36%
Missouri	6.81%	3.37%	7.36%	22.19%	1.34%	8.21%	0.97%
Minnesota	9.32%	5.96%	7.97%	12.98%	1.13%	7.47%	0.23%
Texas	5.26%	7.02%	9.71%	10.55%	4.71%	7.45%	0.20%
Iowa	3.38%	1.32%	4.92%	6.89%	19.90%	7.28%	0.04%
Colorado	-1.34%	8.41%	6.55%	17.61%	2.89%	6.83%	-0.42%
New Mexico	4.18%	3.30%	10.98%	8.82%	6.79%	6.81%	-0.43%
Oregon	2.15%	8.52%	9.49%	11.94%	1.43%	6.71%	-0.54%
Tennessee	8.85%	3.00%	8.41%	10.52%	2.21%	6.60%	-0.65%
Mississippi	1.84%	3.04%	8.48%	14.49%	4.70%	6.51%	-0.74%
Illinois	15.54%	4.70%	4.37%	6.12%	1.01%	6.35%	-0.90%
North Carolina	4.00%	3.16%	9.18%	9.91%	4.75%	6.20%	-1.05%
New York	8.98%	4.78%	6.14%	7.74%	3.36%	6.20%	-1.05%
Arkansas	4.97%	10.38%	5.87%	6.46%	2.47%	6.03%	-1.22%
Kansas	2.21%	3.64%	10.07%	9.02%	4.98%	5.98%	-1.26%
Michigan	10.17%	6.96%	5.24%	7.19%	-0.42%	5.83%	-1.42%
Idaho	7.96%	-0.76%	10.83%	10.19%	0.77%	5.80%	-1.45%
Pennsylvania	4.56%	3.19%	6.53%	10.49%	2.99%	5.55%	-1.69%
North Dakota	5.43%	0.63%	1.39%	16.56%	3.40%	5.48%	-1.76%
Washington	0.38%	3.50%	6.07%	12.98%	3.49%	5.28%	-1.96%
Wisconsin	8.53%	1.56%	3.21%	10.12%	2.38%	5.16%	-2.09%
Wyoming	0.47%	-1.66%	5.86%	13.92%	4.95%	4.71%	-2.54%
Masschusetts	5.38%	6.95%	2.93%	5.63%	2.42%	4.66%	-2.58%
Alabama	3.60%	5.89%	7.65%	6.66%	-0.75%	4.61%	-2.64%
District of Columbia	2.02%	2.60%	6.39%	-0.48%	12.15%	4.54%	-2.71%
South Dakota	-0.55%	0.42%	2.60%	8.50%	5.37%	3.27%	-3.98%
Nebraska	5.11%	-0.78%	-1.16%	8.48%	4.23%	3.18%	-4.07%
Maine	2.52%	3.94%	2.67%	3.08%	2.88%	3.02%	-4.23%
Vermont	8.27%	1.88%	0.21%	3.57%	1.01%	2.99%	-4.26%
Montana	0.33%	-0.73%	6.54%	10.69%	-2.91%	2.78%	-4.46%
Rhode Island	3.40%	2.18%	-0.30%	2.07%	5.12%	2.50%	-4.75%
Alaska	-11.18%	4.51%	1.83%	12.16%	4.24%	2.31%	-4.93%
New Hampsire	0.10%	-0.62%	2.92%	4.19%	1.53%	1.62%	-5.62%
US Growth	**7.326%**	**4.554%**	**7.851%**	**11.671%**	**4.826%**	**7.246%**	N/A
Source: Derived from NCES (Internet access, October. 2012)							