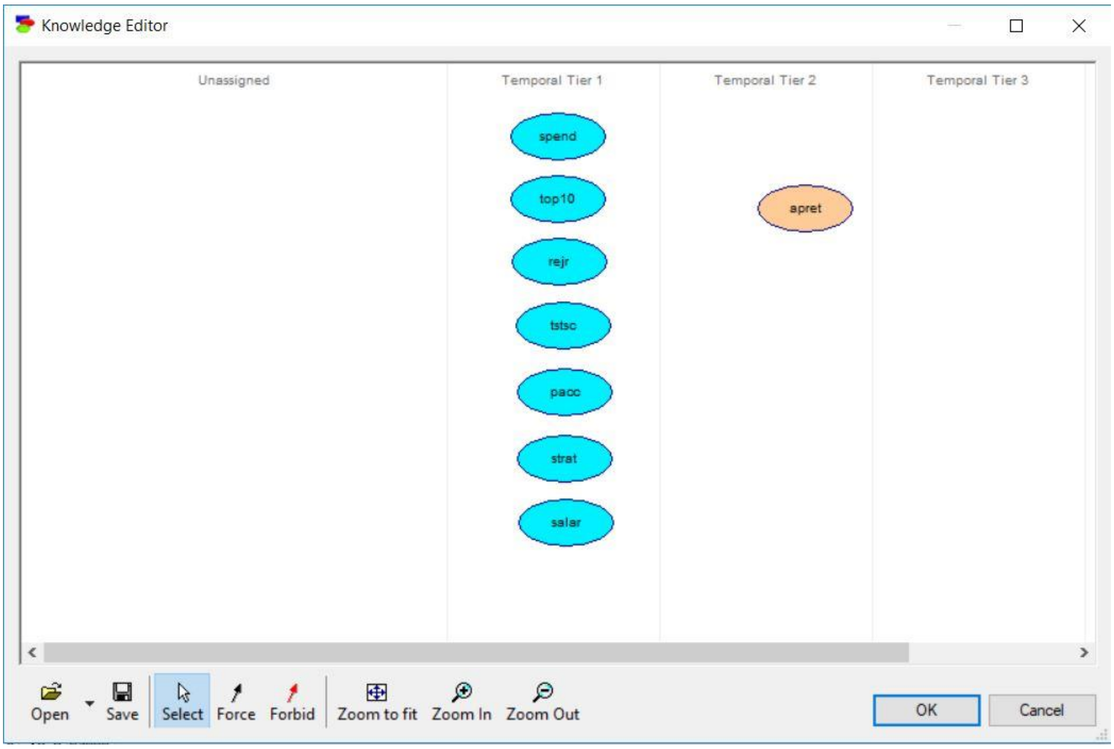
Data Analytics Assignment 6

Jing Pang, Tian Xue, Chuqian Ma, Jiaxiang Leng

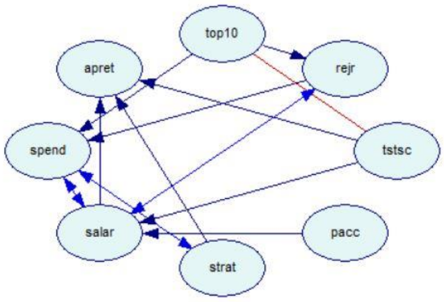
Our task is to verify the conclusion made by Druzdzel & Glymour that student retention rate in U.S. colleges is directly related to the average test scores and high school class standing of the incoming freshmen.

1. Causal Graphs

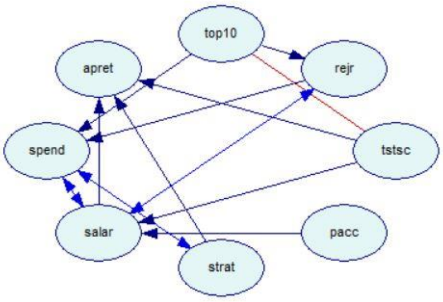
To find out what variables are related to student retention rate, we set the background knowledge as follow:



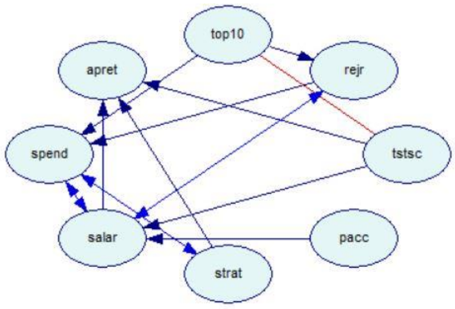
We run PC Learning Algorithm withh the following significance levels: p = 0.2, 0.15, 0.1, 0.05, 0.01, and 0.001, the same as were used in Druzdzel & Glymour 1994. The learned networks are shown as follows:



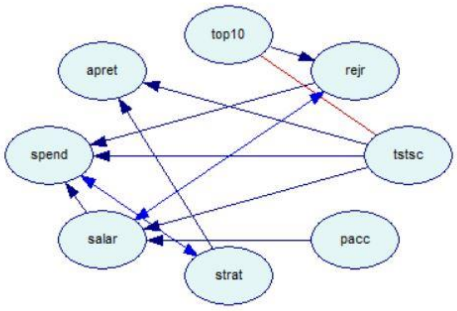
P=0.2



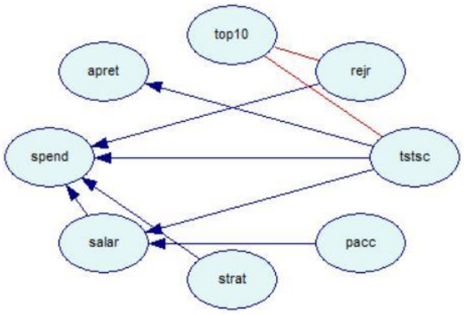
P=0.15



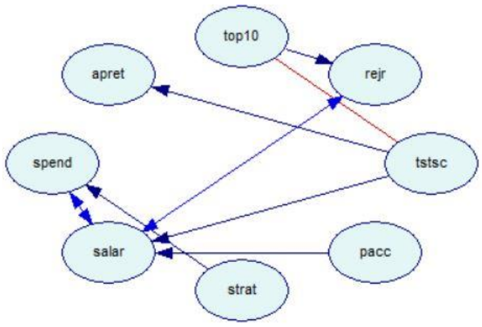
P=0.1



P=0.05



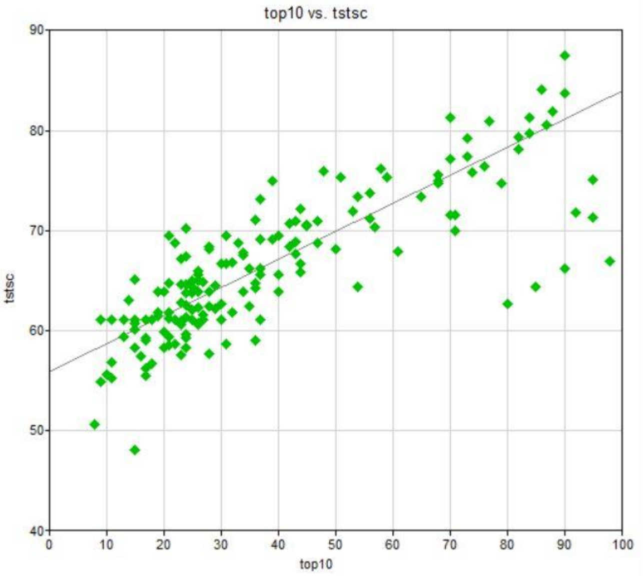
P=0.01



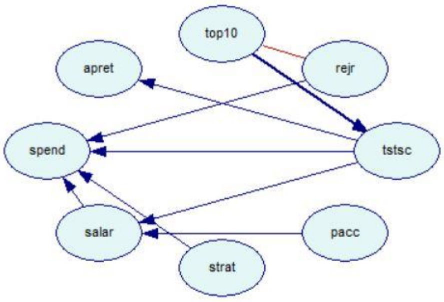
P=0.001

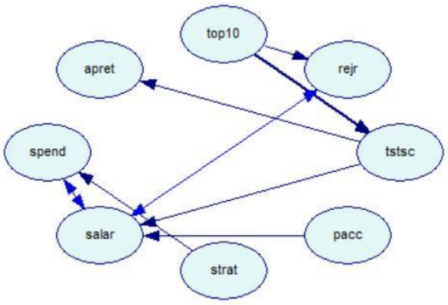
We find out that the only direct causes of the variable apret (average percentage of student retention) are tstsc (average standardized test scores of incoming students), strat (student teacher ratio) and salar (average faculty salary in dollar) when significant levels are set to 0.2, 0.15 and 0.1. However, only strat and tstsc remain when p=0.05. When p=0.01 and p=0.001, tstsc becomes the only variable that is directly related to apret.

To find out latent relationship between tstsc and top10, we perform linear regression between tstsc and top 10. The scatter plot is shown below. Through the plot we find out that top10 and tstsc are directly related.



According to time precedence, it is the percentage of top 10% high school students that influences the scores of incoming students, rather than the score influences the percentage. From this assumption we conclude that top10 directly influences tstsc. The true networks adjusted are shown below.



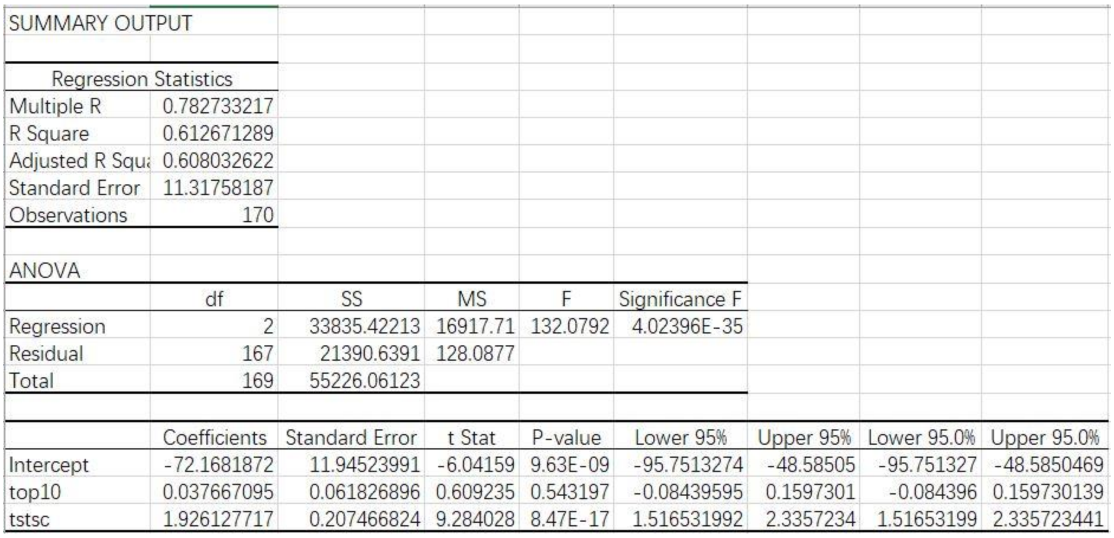


P=0.01 P=0.001

From the graphs above, we conclude that student retention rate is only related to test scores of incoming students and their high school class standings.

2. Linear Regression

Same as the approach used in [Druzdzel & Glymour 1994], we apply linear regression to the relation between the indicators of the quality of incoming freshmen: tstsc(average test scores) and top10(class rating) and apret(freshmen retention rate) to obtain a quantitative measure of these interactions. The output using excel is shown below.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.782733217
R Square	0.612671289
Adjusted R Squa	0.608032622
Standard Error	11.31758187
Observations	170

ANOVA

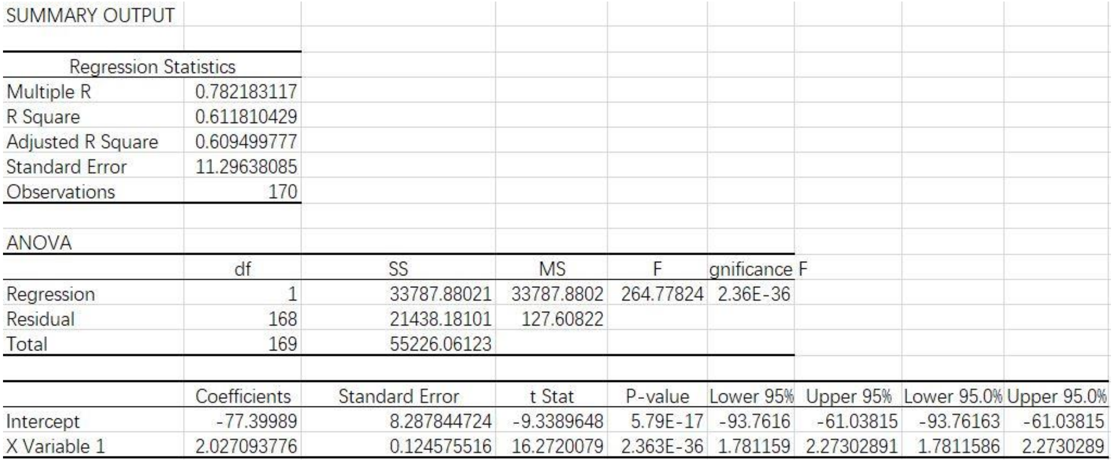
	df	SS	MS	F	Significance F
Regression	2	33835.42213	16917.71	132.0792	4.02396E-35
Residual	167	21390.6391	128.0877		
Total	169	55226.06123			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-72.1681872	11.94523991	-6.04159	9.63E-09	-95.7513274	-48.58505	-95.751327	-48.5850469
top10	0.037667095	0.061826896	0.609235	0.543197	-0.08439595	0.1597301	-0.084396	0.159730139
tstsc	1.926127717	0.207466824	9.284028	8.47E-17	1.516531992	2.3357234	1.51653199	2.335723441

The equation is:

apret = -72.1682 + 0.0377 top10 + 1.9261 tstsc, R-sq(adj)=60.8%

As the indicator of top10 is too small compared with that of tstsc, we repeat the linear regression with only tstsc. The output using excel is shown below.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.782183117
R Square	0.611810429
Adjusted R Square	0.609499777
Standard Error	11.29638085
Observations	170

ANOVA

	df	SS	MS	F	gnificance F
Regression	1	33787.88021	33787.8802	264.77824	2.36E-36
Residual	168	21438.18101	127.60822		
Total	169	55226.06123			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-77.39989	8.287844724	-9.3389648	5.79E-17	-93.7616	-61.03815	-93.76163	-61.03815
X Variable 1	2.027093776	0.124575516	16.2720079	2.363E-36	1.781159	2.27302891	1.7811586	2.2730289

The equation is:
apret = -77.3999 + 2.0271 tstsc, R-sq(adj)=61.2%

3. Conclusion

From the network learned by PC algorithm, test scores and class standing are the only two variables related to retention rate. Factors such as student faculty ratio, faculty salary, and university's educational expenses per student are all independent to retention rates, and, therefore, do not seem to directly influence student retention.

From the regression above, we find out that average standardized test scores of incoming students and the percentage of incoming freshmen who were among the top 10% students in their high schools explain nearly 62% of the variance in retention rates. This finding proves the conclusion made by Druzdzel & Glymour in 1994 that "student retention is directly related to the average test scores and high school class standing of the incoming freshmen". In their research, "test scores and class standing explain 52.6%of the variance in freshmen retention rate and62.5%of the variance in graduation rate".