ASSIGNMENT 2

Chunyao Wang(chw132),
Wenxing Li(wel85),
Jie Rong(jir18)

1. Overview

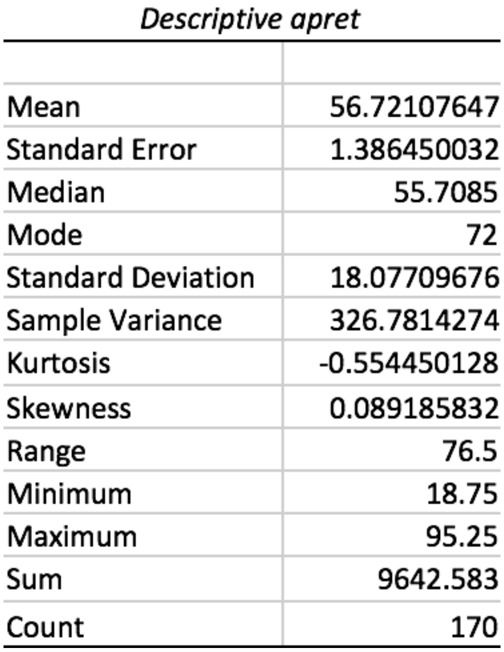
This project analyzes the data set of 170 tuples. Each tuple represents a college described by 8 attributes (their meanings are listed below).

spend - average spending per student (in dollars)
apret - average retention rate (i.e., percentage of students making it through the studies)
top10 - percentage of incoming freshmen who were among the top 10% students in their high schools
rejr - school's rejection rate (percentage of applicants denied admission)
tstsc - average test scores of incoming freshmen
pacc - percent of admitted applicants who accept university's offer
strat - student-teacher ratio
salar - average faculty salary (in dollars)

First we use descriptive analysis to describe the basic features of data and figure out the law hidden behind them. Then we try to find the relations among apret, tstsc and salar by linear regression. We choose Microsoft Excel as the data analysis tool and import data into it form retention.txt.

2. Descriptive statistics and histograms for apret, tstsc, and salar

2.1 apret



Descriptive apret

Mean	56.72107647
Standard Error	1.386450032
Median	55.7085
Mode	72
Standard Deviation	18.07709676
Sample Variance	326.7814274
Kurtosis	-0.554450128
Skewness	0.089185832
Range	76.5
Minimum	18.75
Maximum	95.25
Sum	9642.583
Count	170

Chart2.1 descriptive statistics for apret

The chart shows descriptive statistics for percentage of students making it through the studies. According to the chart, almost half of the students do struggle with their studies because mean and median are very close and the minimum is as low as 18.75%. As it can be seen, majority of the students to perform acceptably due to mode 72% and maximum 95.25%.

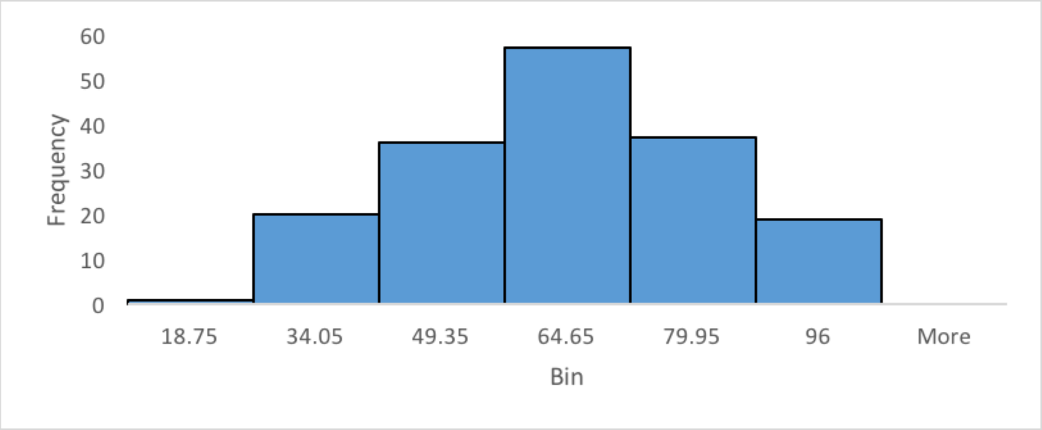


Fig2.1 histogram for apret

According to fig2.1, there are obvious differences among colleges' average rate of retention by combining that difference with the variance and the deviation.

2.2 tstsc

Descriptive tstsc

Mean	66.1641647
Standard Error	0.53498157
Median	64.7815
Mode	61.111
Standard Deviation	6.97530626
Sample Variance	48.6548974
Kurtosis	0.19642638
Skewness	0.57321757
Range	39.375
Minimum	48.125
Maximum	87.5
Sum	11247.908
Count	170

Chart2.1 descriptive statistics for tstsc

The chart shows descriptive statistics for percentage of students making it through the studies. According to the chart, median and mode are within the range of sixties. Minimum 48.13 and maximum 87.5 is another proof that majority of the students are stuck between this range.

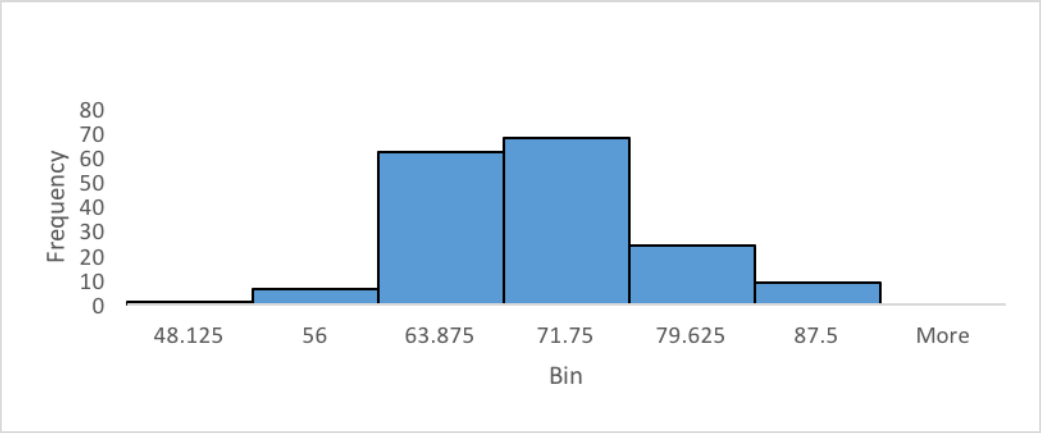


Fig2.1 histogram for tstsc

As the histogram also shows, the max value of average test scores of incoming freshmen is 87.5 and the minimum value of it is 48.13. Whole average rate is medium. The difference of the max and minimum score is 39.37, and observing the variance and deviation informs that colleges' average test scores of incoming freshmen are not so different with each other compared with the average retention rate that is showed above.

2.3 salar

Descriptive salar

Mean	61357.6471
Standard Error	751.839401
Median	61150
Mode	48000
Standard Deviation	9802.78646
Sample Variance	96094622.3
Kurtosis	-0.2310967
Skewness	0.25787668
Range	49260
Minimum	38640
Maximum	87900
Sum	10430800
Count	170

Chart2.1 descriptive statistics for salar

The chart shows descriptive statistics for average faculty salary (in dollars). Central tendency informs that average value and the exact middle value of the set are almost identical however the sample variance and standard deviation shows how different the salaries can vary among the schools. Mode or the most frequent value however is 48 000 dollars per year which is significantly closer to the minimum than maximum.

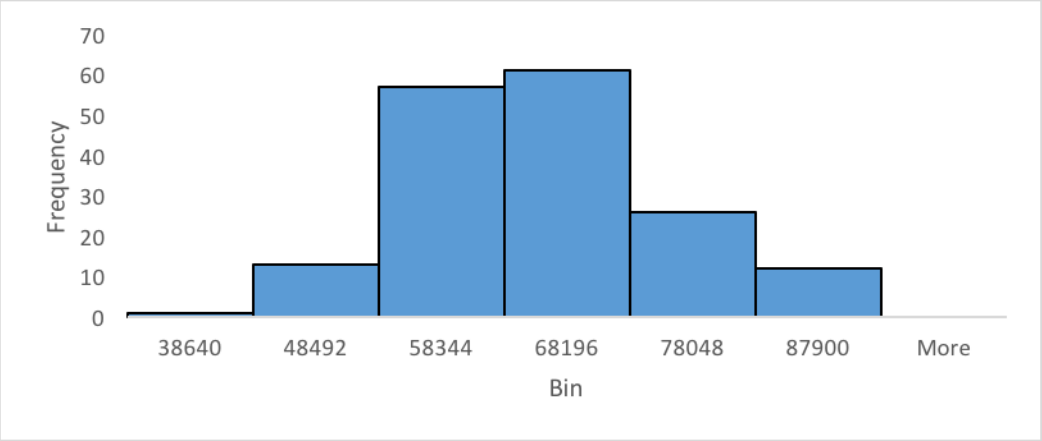
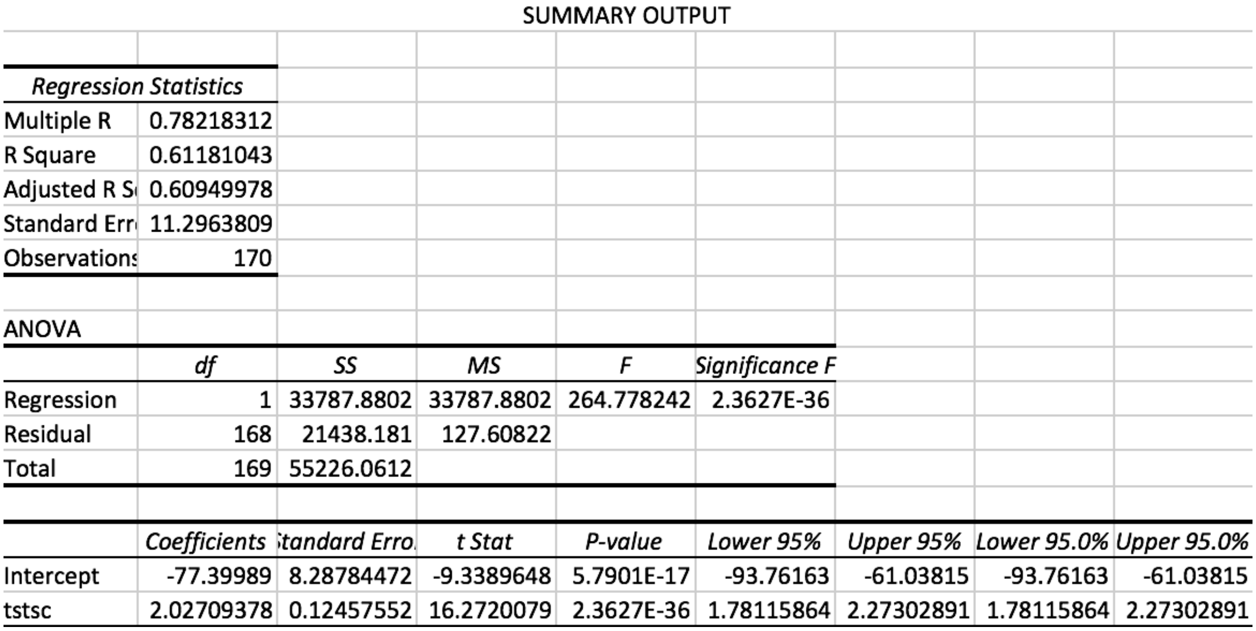


Fig2.1 histogram for salar

Fig2.1 shows the max value of average faculty salary is 87900 and the minimum value of it is 38640. The mean of average faculty salary is 61357.6, and the variance is 9.61 and the deviation is 9802.79. Whole average rate is normal. The difference of the max and minimum score is 49260.

3. linear regression

3.1 linear regression of apret on tstsc



SUMMARY OUTPUT

Regression Statistics								
Multiple R	0.78218312							
R Square	0.61181043							
Adjusted R S	0.60949978							
Standard Err	11.2963809							
Observations	170							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	33787.8802	33787.8802	264.778242	2.3627E-36			
Residual	168	21438.181	127.60822					
Total	169	55226.0612						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-77.39989	8.28784472	-9.3389648	5.7901E-17	-93.76163	-61.03815	-93.76163	-61.03815
tstsc	2.02709378	0.12457552	16.2720079	2.3627E-36	1.78115864	2.27302891	1.78115864	2.27302891

Chart3.1 regression of apret on tstsc

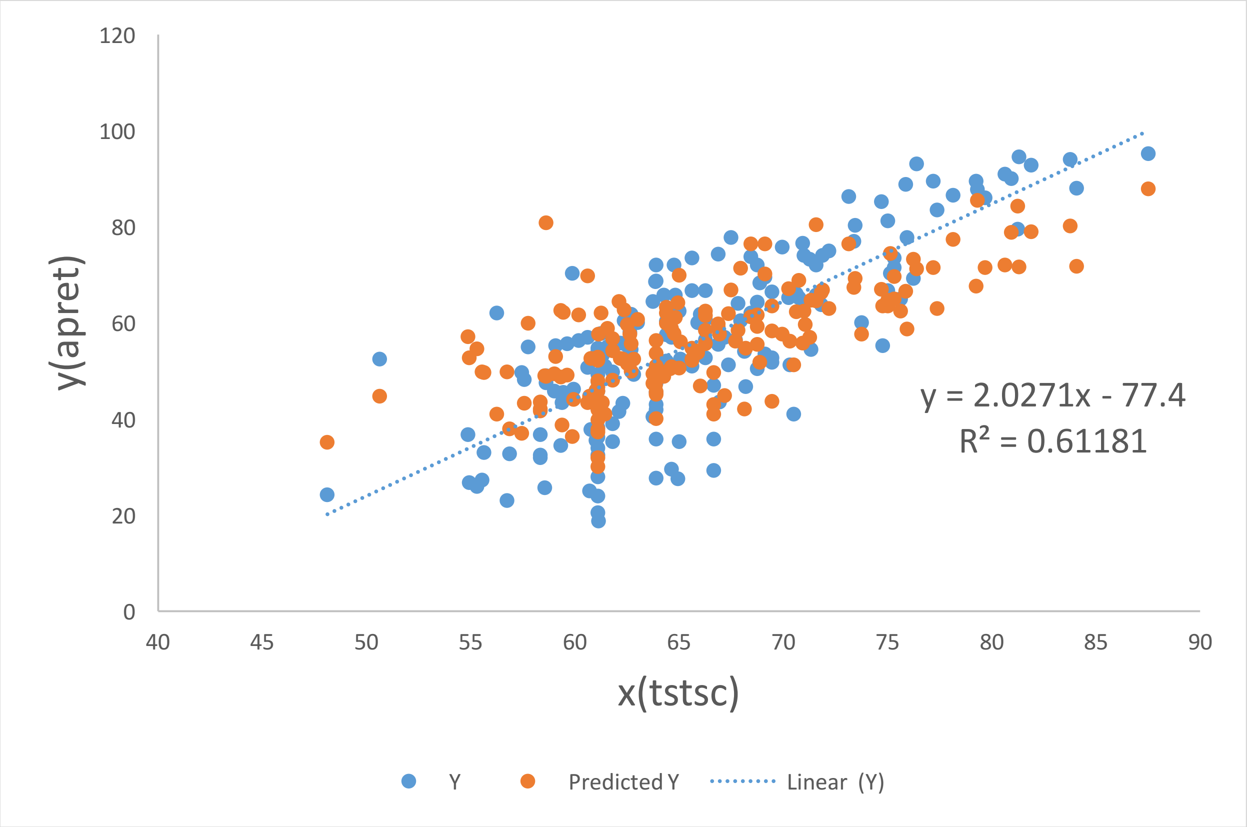
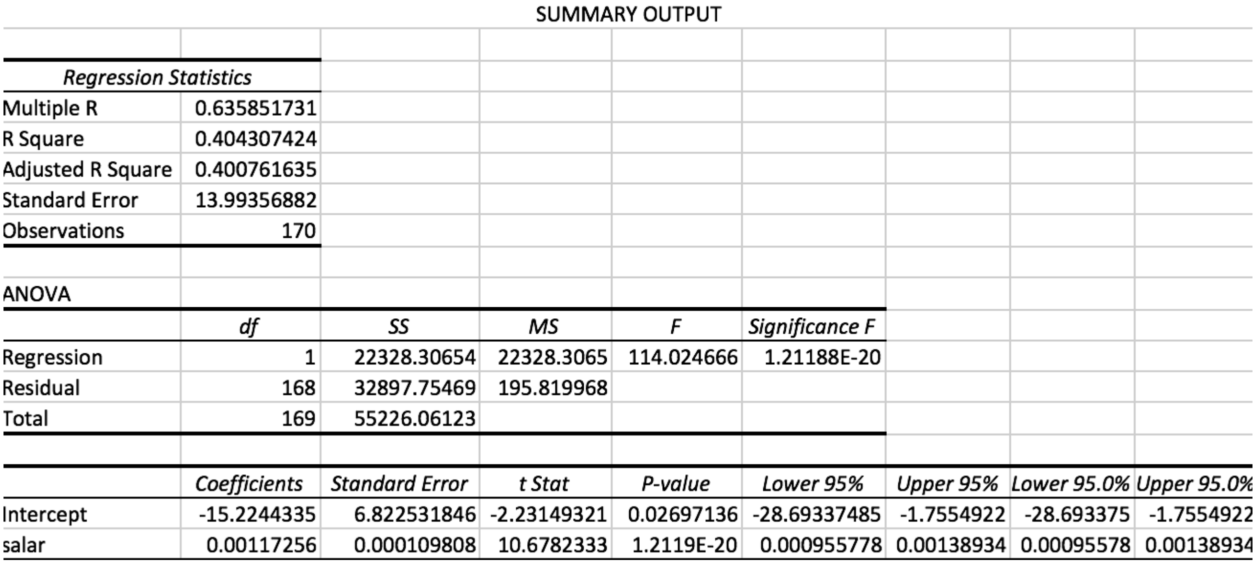


Fig3.1 regression of apret on tstsc

3.2 linear regression of apret on salar



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.635851731
R Square	0.404307424
Adjusted R Square	0.400761635
Standard Error	13.99356882
Observations	170

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	22328.30654	22328.3065	114.024666	1.21188E-20
Residual	168	32897.75469	195.819968		
Total	169	55226.06123			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-15.2244335	6.822531846	-2.23149321	0.02697136	-28.69337485	-1.7554922	-28.693375	-1.7554922
salar	0.00117256	0.000109808	10.6782333	1.2119E-20	0.000955778	0.00138934	0.00095578	0.00138934

Chart3.2 regression of apret on salar

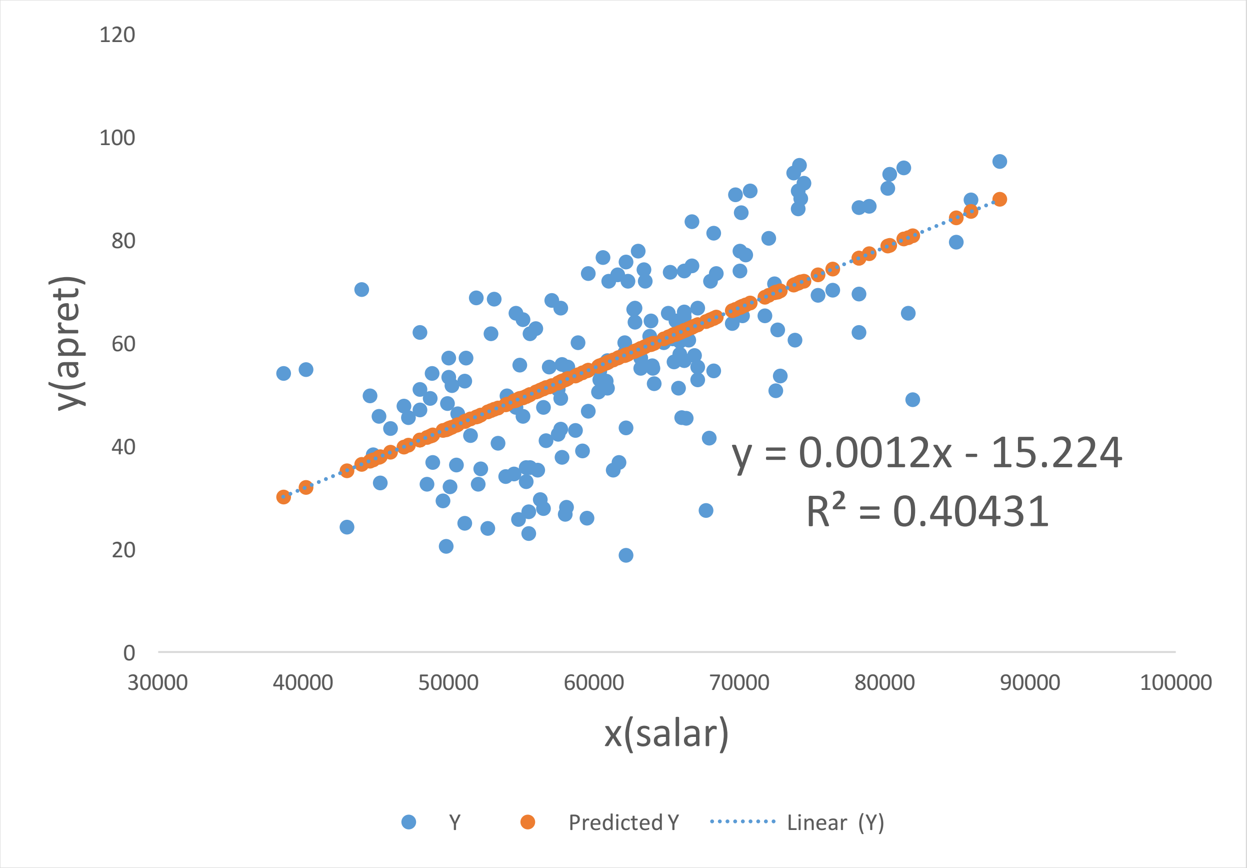


Fig3.2 regression of apret on salar

3.3 linear regression of apret on both tstsc and salar

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.78975521
R Square	0.62371329
Adjusted R Square	0.61920686
Standard Error	11.1550941
Observations	170

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	34445.22832	17222.6142	138.405259	3.59606E-36
Residual	167	20780.83291	124.436125		
Total	169	55226.06123			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-75.911107	8.209780049	-9.246424	1.0684E-16	-92.11943723	-59.702777	-92.119437	-59.702777
tstsc	1.7375403	0.176080788	9.86785849	2.208E-18	1.389909111	2.08517148	1.38990911	2.08517148
salar	0.00028797	0.000125293	2.29839395	0.02277903	4.06102E-05	0.00053533	4.061E-05	0.00053533

Chart3.3 regression of apret on both tstsc and salar

Chart3.3 shows the relationship of apret on both tstsc and salar by linear regression. According to the chart, apret is more linear dependent on salar with a coefficients as 1.738 than tatac with a much lower coefficients as 0.00029. R square 0.624 means 62.4% of variation of apret y bar is explained by repressor tstsc and salar. The standard error 11.2 estimates the standard deviation of the error, this value is not expected according to the coefficient number, we generate:

apret = -75.9111+1.7375*tstsc+0.0003*salar