CASE PROBLEM: ALUMNI GIVING

1. Use methods of descriptive statistics to summarize the data.

	Graduation Rate	% of Classes Under 20	Student-Faculty Ratio	Alumni Giving Rate
Mean	83.04	55.73	11.54	29.27
Standard Error	1.24	1.90	0.70	1.94
Median	83.85	59.5	10.50	29
Mode	92	65	13	13
Standard Deviation	8.61	13.19	4.85	13.44
Minimum	66	29	3	7
Maximum	97	77	23	67

	Graduation Rat	*% of Classes Under 20*	*Student-Faculty Ratio*	*Alumni Giving Rate*
Graduation Rate	1			
% of Classes Under 20	0.58278843	1		
Student-Faculty Ratio	-0.6049379	-0.785559252	1	
Alumni Giving Rate	0.75594359	0.645650419	-0.742397463	1

The correlation matrix among all the variables is shown above. The highest negative correlation between Student-Faculty Ratio and percentage of Classes Under 20 indicates that the high percentage of classes under 20 is, the lower student-faculty ratio is. We can also see that Alumni Giving Rate is positively correlated with Graduation Rate, suggesting that schools with higher graduation rate have higher alumni giving rate. On the contrary, Student-Faculty Ratio has a negative correlation with Alumni giving rate, which shows that schools with higher student-faculty ratio have lower alumni giving rate.

2. Develop an estimated simple linear regression model that can be used to predict the alumni giving rate, given the graduation rate. Discuss your findings.

(1) Graduation Rate and Alumni Giving Rate

Regression Statistics								
Multiple R	0.75594359							
R Square	0.57145071							
Adjusted R Square	0.56213442							
Standard Error	8.89432811							
Observations	48							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	4852.46183	4852.46183	61.3388789	5.2382E-10			
Residual	46	3639.01734	79.1090726					
Total	47	8491.47917						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-68.761183	12.5826557	-5.4647591	1.821E-06	-94.088755	-43.43361	-94.088755	-43.43361
Graduation Rate	1.180516	0.15073148	7.83191413	5.2382E-10	0.87710927	1.48392273	0.87710927	1.48392273

In the Regression Statistics section, $R^2 = 0.5715$, which means that approximately 57% of the variation in Alumni Giving Rate is explained by Graduation Rate. The significance F is nearer to zero and the P-value is much less than 0.05, we deduce the following formula:

$$Y = 1.18X_1 - 68.76$$

(2) % of Classes Under 20 and Alumni Giving Rate

Regression Statistics								
Multiple R	0.64565042							
R Square	0.41686446							
Adjusted R Square	0.4041876							
Standard Error	10.3752247							
Observations	48							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	3539.79591	3539.79591	32.8838909	7.2281E-07			
Residual	46	4951.68326	107.645288					
Total	47	8491.47917						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-7.3860676	6.56547229	-1.1249865	0.26643066	-20.601678	5.82954265	-20.601678	5.82954265
% of Classes Under 20	0.65776869	0.1147048	5.73444774	7.2281E-07	0.4268799	0.88865748	0.4268799	0.88865748

In the Regression Statistics section, $R^2 = 0.4169$, which means that approximately 42% of the variation in Alumni Giving Rate is explained by Percentage of Classes Under 20. The significance

F is nearer to zero and the P-value is much less than 0.05, we come up with the following formula:

$$Y = 0.66X_2 - 7.39$$

(3) Student-Faculty Ratio and Alumni Giving Rate

Regression Statistics								
Multiple R	0.74239746							
R Square	0.55115399							
Adjusted R Square	0.54139647							
Standard Error	9.10251579							
Observations	48							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	4680.11265	4680.11265	56.4850379	1.5442E-09			
Residual	46	3811.36651	82.8557938					
Total	47	8491.47917						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	53.0138271	3.42145043	15.4945478	7.0588E-20	46.1268046	59.9008497	46.1268046	59.9008497
Student-Faculty Ratio	-2.0571547	0.27371603	-7.5156529	1.5442E-09	-2.6081165	-1.5061929	-2.6081165	-1.5061929

In the Regression Statistics section, $R^2 = 0.5512$, which means that approximately 55% of the variation in Alumni Giving Rate is explained by Student-Faculty Ratio. The significance F is nearer to zero and the P-value is much less than 0.05, we come up with the following formula:

$$Y = -2.06X_3 + 53.01$$

3. Develop an estimated multiple linear regression model that could be used to predict the alumni giving rate using Graduation Rate, % of Classes Under 20, and Student-Faculty Ratio as independent variables. Discuss your findings.

Regression Statistics								
Multiple R	0.83662453							
R Square	0.69994061							
Adjusted R Square	0.67948201							
Standard Error	7.60972478							
Observations	48							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	5943.53107	1981.17702	34.2125451	1.4323E-11			
Residual	44	2547.94809	57.9079112					
Total	47	8491.47917						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-20.720134	17.521365	-1.1825639	0.24333306	-56.032125	14.5918566	-56.032125	14.5918566
Graduation Rate	0.7481828	0.16595996	4.50821272	4.799E-05	0.41371248	1.08265312	0.41371248	1.08265312
% of Classes Under 20	0.02904065	0.13932132	0.20844367	0.83584449	-0.251743	0.30982432	-0.251743	0.30982432
Student-Faculty Ratio	-1.1920107	0.3867231	-3.0823364	0.0035384	-1.9713999	-0.4126215	-1.9713999	-0.4126215

Looking at the p-value for the independent variables in the last section, we can see that p-value of intercept and % of Classes Under 20 are less than 0.05, therefore, we ignore the independent variable percentage of Classes Under 20 and test again and gain the following tables.

Regression Statistics								
Multiple R	0.83644743							
R Square	0.69964431							
Adjusted R S	0.68629516							
Standard Err	7.52841155							
Observations	48							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	5941.01505	2970.50752	52.4111817	1.7653E-12			
Residual	45	2550.46412	56.6769805					
Total	47	8491.47917						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-19.106313	15.5500587	-1.2286972	0.22557243	-50.425739	12.2131131	-50.425739	12.2131131
Graduation F	0.75573539	0.16022577	4.71669052	2.3478E-05	0.43302412	1.07844667	0.43302412	1.07844667
Student-Facu	-1.2459535	0.28430229	-4.3824955	6.9542E-05	-1.8185677	-0.6733393	-1.8185677	-0.6733393

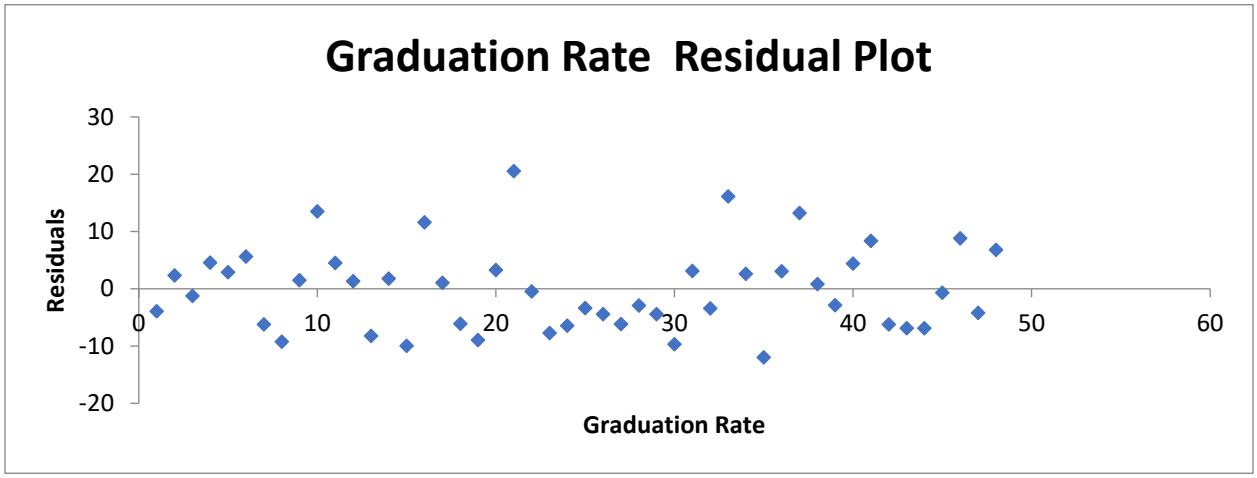
The value of R^2 indicates that 70% of the variation in the dependent variable is explained by these independent variables.

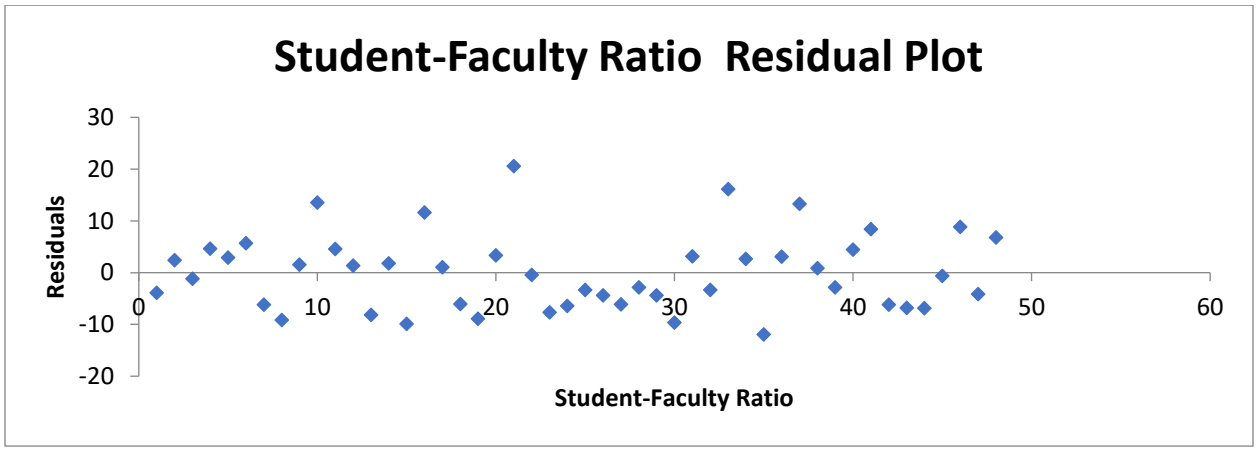
From the ANOVA section, we may test for a significance of regression. At a 5% significance level, we reject the null hypothesis because Significance F s essentially zero.

Looking at the p-value for the independent variables in the last section, we see that the P-value of Graduation Rate and Student-Faculty Rate are smaller than 0.05, however, P-value of Intercept is larger than 0.05, but statistically speaking, we do not pay much attention to intercept, therefore we keep using the data. We reject the null hypothesis that each partial regression coefficient is zero and conclude that each of them is statistically significant.

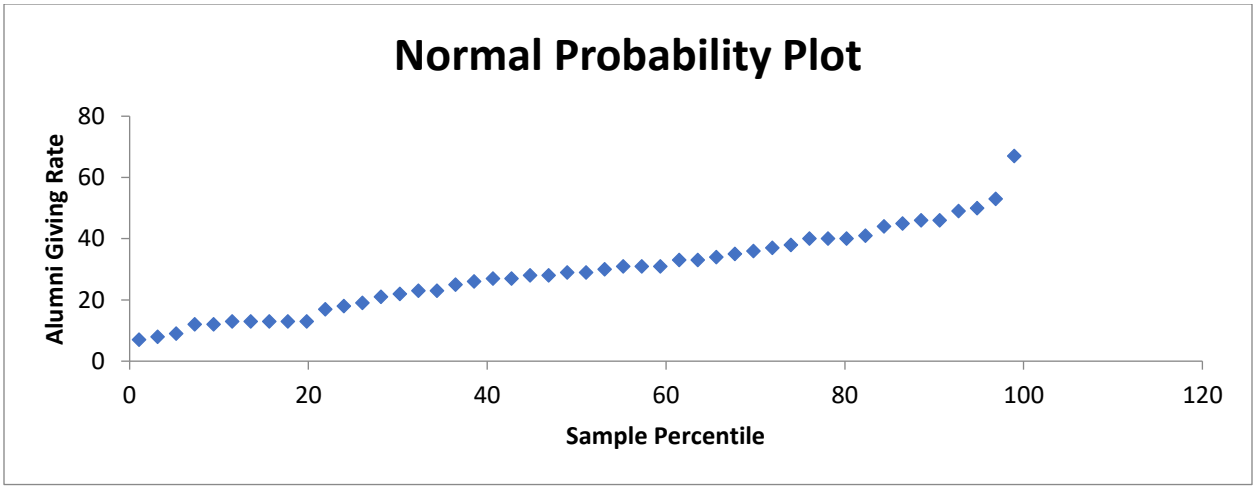
As a consequence, we can give the following formula.

$$Y = 0.76X_1 - 1.25X_2 - 19.11$$





The above residual plots validate these assumptions.



The normal probability plot does not suggest any serious departures from normality.

4. Based on the results in part 2 and 3, do you believe another regression model may be more appropriate? Estimate this model, and discuss your result.

By considering part 2 and 3, win order to estimate a better model, when we construct a model with all available independent variables, we find out that the independent variable Percentage of Classes Under 20 is not significant by looking at its P-value.

Therefore, we remove the independent variable Percentage of Classes Under 20 and construct a new model. We find out that all the rest independent variables are significant. Based on this, we construct the following model.

$$Y = 0.76X_1 - 1.25X_2 - 19.11$$

Regression Statistics	
Multiple R	0.83644743
R Square	0.69964431
Adjusted R S	0.68629516
Standard Err	7.52841155
Observations	48

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5941.01505	2970.50752	52.4111817	1.7653E-12
Residual	45	2550.46412	56.6769805		
Total	47	8491.47917			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-19.106313	15.5500587	-1.2286972	0.22557243	-50.425739	12.2131131	-50.425739	12.2131131
Graduation F	0.75573539	0.16022577	4.71669052	2.3478E-05	0.43302412	1.07844667	0.43302412	1.07844667
Student-Facu	-1.2459535	0.28430229	-4.3824955	6.9542E-05	-1.8185677	-0.6733393	-1.8185677	-0.6733393

5. What conclusions and recommendations can you derive from your analysis? What universities are achieving a substantially higher alumni giving rate than would be expected, given their Graduation Rate, % of Classes Under 20, and Student/Faculty Ratio? What universities are achieving a substantially lower alumni giving rate than would be expected, given their Graduation Rate, % of Classes Under 20, and Student/Faculty Ratio? What other independent variables could be included in this model?

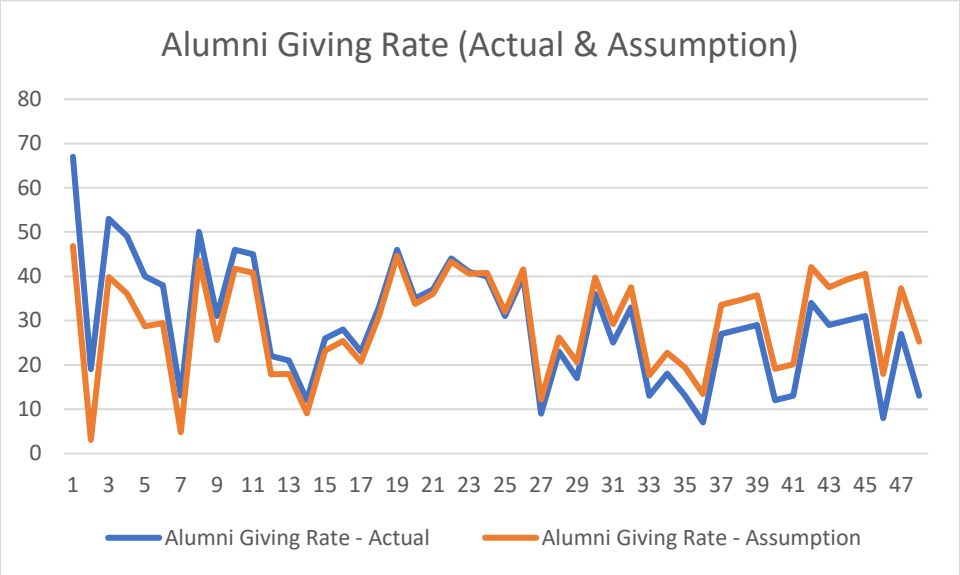
(1) My conclusions:

A. Generally speaking, building a multiple linear regression is better than building a simple linear regression because the dependent variable is usually affected by a couple of independent variables. In this case, R^2 of multiple linear regression is higher than R^2 of simple linear regression, which suggests that the dependent variable can be explained more in the multiple linear regression.

B. After constructing a model, we should take the test results into consideration because sometimes the independent variables are not significance.

(2) By applying the formula $Y = 0.76X_1 – 1.25X_2 – 19.11$, we come up with the estimated figures and draw the following chart, the trend of the two lines are similar, some universities witness higher substantial alumni giving rate while the others have lower alumni giving rate.

According to the actual funding and our model, Princeton University, U. of Florida, Dartmouth College have higher alumni giving rate, and U. of California-San Diego, Johns Hopkins University and U. of Michigan-Ann Arbor have lower alumni giving rate.



(3) Other independent variables: Public/Private; Job employment rate; enterprise faculty rate;