Ramadan Gannud
Assignment-5

Problem 1 [5 pts] – to be answered by everyone

You will continue the prediction, confidence interval and prediction interval for the **banking** dataset that was analyzed in Assignment 4. Since you would have altered the dataset to exclude outliers/influential points and/or multicollinearity, use the dataset and the code that was used to generate your final model. Note: Make sure you rerun the whole banking code from assignment 4, before you do this last part.

a) **Use the fitted regression model from Assignment 4 to predict the average bank balance for a specific zip code area where there is a plan to open a new branch. Census data in that area show the following values: median age is 34 years, median education is 13 years, median income is $89,000, median home value is $160,000, median wealth is 140,000. Using SAS, compute the predicted average bank balance, 95% confidence interval and prediction interval for your estimate. Make sure to use SAS coding to determine the values. Include all relevant outputs. Discuss your findings.**

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	28930	497.3768	27942	29918	25721	32139	.
2	38517	40407	411.7951	39589	41225	37246	43567	-1890
3	40618	40078	343.6694	39396	40761	36950	43207	539.6379
4	35206	34267	242.6509	33785	34749	31176	37358	938.9733

The predicted average bank balance is **$28930** with 95% C.I equal to ($27942, $29918) and P.I ($25721, $32139).

b) **Copy and paste your FULL SAS code into the word document along with your answers.**

```
*compute predictions;
data pred;
input balance age education homeVal wealth;
datalines;
. 34 13 160000 140000
;

*join datasets;
data predict;
set pred Bankingfull_new1;
run;

proc print;
run;

proc reg;
model balance= age education homeVal wealth/p clm cli;
run;
```

PROBLEM 2 [20 pts] – to be answered by everyone

This problem asks you to build a model for the college dataset (college.csv) that contains the following variables:

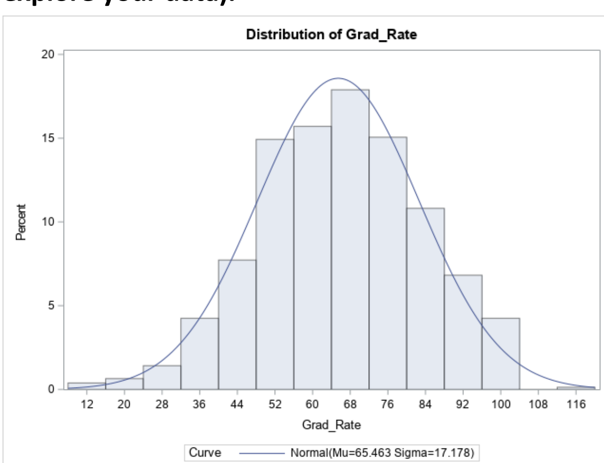
School	*School name*
Private	*public/private indicator. YES if university is private, NO if university is public.*
Accept.pct	*percentage of applicants accepted*
Elite10	*Elite schools with majority of students from the top 10% of their high school class (0- Not Elite, 1-Elite)*
F.Undergrad	*number of full-time undergraduate students*
P.Undergrad	*number of part-time undergraduate students*
Outstate	*Out-of-state tuition*
Room.Board	*room and board costs*
Books	*estimated book costs*
Personal	*Estimated personal spending*
PhD	*Percent of faculty with PhD*
Terminal	*Faculty with terminal degrees (terminal degree is a university degree that is either highest on the academic track or highest on the professional track in a given field of study)*
S.F.Ratio	*Student/faculty ratio*
perc.alumni	*Percent of alumni who donate*
Expend	*Instructional expenditure per student*
Grad.Rate	*Graduation rate in 4 years*

Apply regression analysis techniques to analyze the relationship among the observed variables and build a model to predict Graduation Rates (Grad.Rate).

Note: Before you start, open the college.csv file, and examine the data.

Answer the following questions.

a) Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation (This is the data exploration stage, therefore use the appropriate statics to explore your data).



Moments			
N	777	Sum Weights	777
Mean	65.4633205	Sum Observations	50865
Std Deviation	17.1777099	Variance	295.073717
Skewness	-0.1137773	Kurtosis	-0.2052265
Uncorrected SS	3558769	Corrected SS	228977.205
Coeff Variation	26.2402056	Std Error Mean	0.61624691

Quantiles (Definition 5)	
Level	Quantile
100% Max	118
99%	100
95%	95
90%	89
75% Q3	78
50% Median	65
25% Q1	53
10%	44
5%	37
1%	22
0% Min	10

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10	586	100	378
15	385	100	512
15	5	100	523
18	67	100	669
21	395	118	96

The distribution is symmetric since the Median is almost equal to the Mean. Also, the histogram gives us a symmetric shape with normal distribution. The skewness is -0.11 which is very small and makes it a normal distribution.

b) **Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission).**

Since the Private variable has qualitative values, we should define a dummy variable.

	Z1 numprivate
YES	1
NO	0

numprivate = 1 if Private = YES

numprivate = 0 if Private = NO

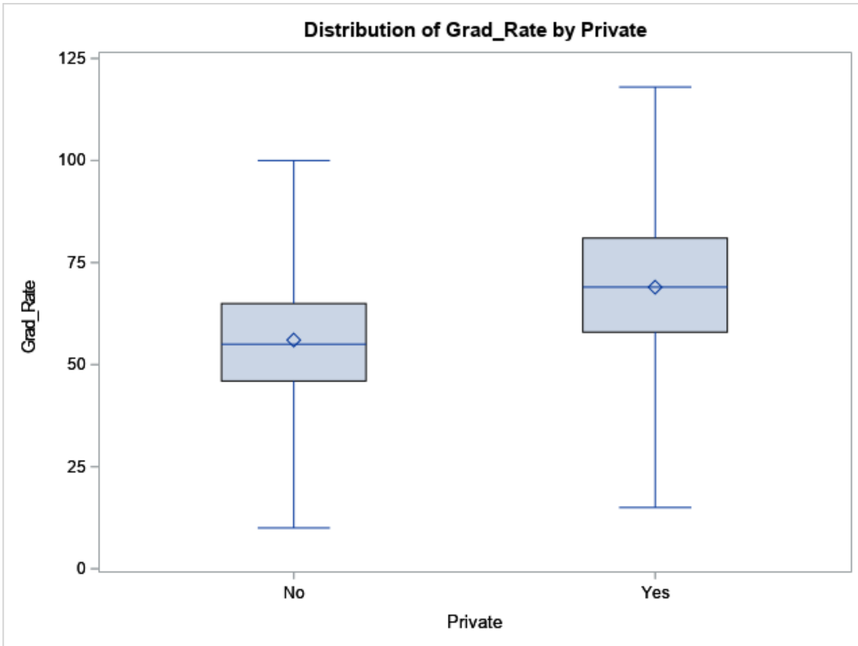
In numprivate and Elite10 variables, the dots are scattered around 0 and 1 because they are dummy variables. Thus, they won't show any linear relationship with Grad_Rate. There is no significant correlation between Grade_Rate and any of the independent variable in the dataset. All independent X-variables show no signification linear association between them and the Grade_Rate variable. Furthermore, Outstate shows a very small positive linear relationship with a correlation value of 0.57.

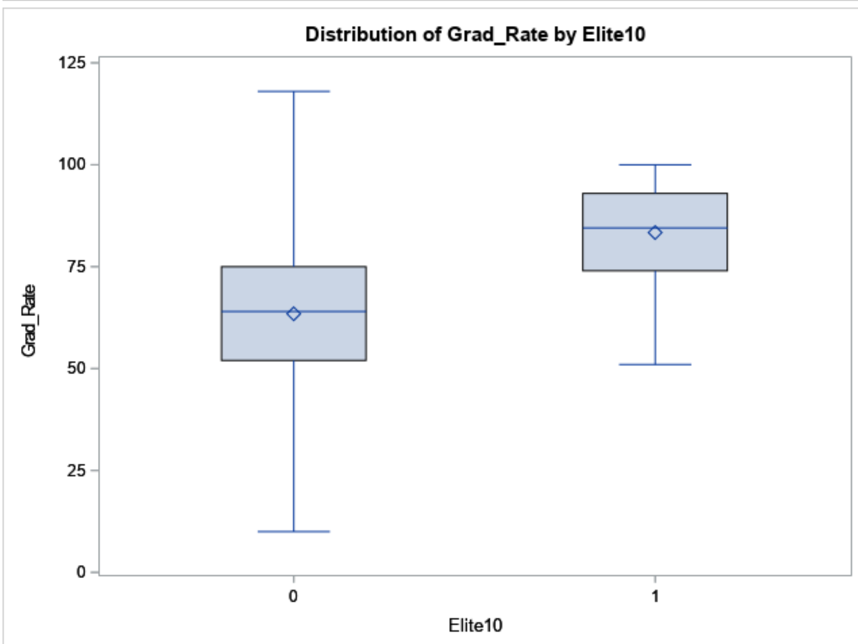
c) **Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Include the boxplots and discuss your findings. (See SAS Procedures section on D2L if you need the code to generate a boxplot).**

According to the boxplots below, the graduation rates vary by university type and the university status (Elite or not). The mean and the median of graduation rates in private schools are higher than they are in public schools. The range of graduation rates is bigger in private schools.

Furthermore, the mean and the median of graduation rates in elite schools is higher than they are in not elite ones. The difference is obvious although the range is smaller in elite schools.

According to the second graph, 75% of the records of graduation rates in elite schools are bigger than 75% of the records of graduation rates in not elite ones.





d) Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output.

Using the absolute value of standardized estimate to determine the predictors with significant effect on graduation rates. The strongest predictor is Outstate since the standardized estimate is the highest 0.289. When performing t-test on individual parameters, books, Terminal, S_F_Ratio have p-values that are higher than 0.05 which make them insignificant X-variables. The other variables' p-values are less than 0.05 which make them significant X-variables.
The coefficient value of the parameter of X measures the predicted change in Y for any unit increase in X while the other independent variables stay constant. For example, if school changed from public to private, graduation rate will increase by4.62%. Also, if acceptance percentage increases by 1%, graduation rate increases by 18.1%.
Adj-R^2 0.4346 does not show a good and a higher Adj-R^2 will give a better model.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	101851	7275.08261	43.61	<.0001
Error	762	127126	166.83208		
Corrected Total	776	228977			

Root MSE	12.91635	R-Square	0.4448
Dependent Mean	65.46332	Adj R-Sq	0.4346
Coeff Var	19.73067		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	51.39777	6.12404	8.39	<.0001	0
numprivate	1	4.61959	1.72185	2.68	0.0075	0.11986
Accept_pct	1	-18.10932	3.84314	-4.71	<.0001	-0.15508
Elite10	1	4.01748	2.00326	2.01	0.0453	0.07033
F_Undergrad	1	0.00068095	0.00014285	4.77	<.0001	0.19228
P_Undergrad	1	-0.00196	0.00039043	-5.01	<.0001	-0.17333
Outstate	1	0.00123	0.00022863	5.40	<.0001	0.28918
Room_Board	1	0.00167	0.00059443	2.80	0.0052	0.10644
Books	1	-0.00252	0.00297	-0.85	0.3951	-0.02426
Personal	1	-0.00172	0.00077810	-2.21	0.0275	-0.06772
PhD	1	0.13064	0.05621	2.32	0.0204	0.12418
Terminal	1	-0.07284	0.06257	-1.16	0.2447	-0.06243
S_F_Ratio	1	0.00100	0.16188	0.01	0.9951	0.00023113
perc_alumni	1	0.30920	0.04839	6.39	<.0001	0.22306
Expend	1	-0.00043651	0.00015180	-2.88	0.0041	-0.13269

Fitted Regression line Expression:

Graduation Rate = 51.39 + 4.62*Private – 18.11*Accept_pct + 4.017*Elite + 0.00068*F_Undergrad – 0.002*P_Undergrad + 0.0012*Outstate + 0.0016*Room_Board – 0.00252*Books – 0.0017*Personal + 0.13*PhD – 0.072*Terminal + 0.001*S_F_Ratio + 0.31*perc_alumni – 0.00043*Expend

Elite = 1 for Elite schools, Elite = 0 for not Elite schools

Private = 1 when school is private, Private = 0 when school is public

e) Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. Include the relevant output and discuss your answer.

Diagnosing Multicollinearity:

1. Scatterplot matrix and Pearson correlation matrix for each pair of x variables:
   There is a collinearity problem between PhD and Terminal because the correlation value between them is 0.85. There is small collinearity between Expand and Outstate with correlation value of

0.67. Also, there is small collinearity between Outstate and Room_board because the correlation value between them is 0.65.

2. Compute Variance Inflation Factor (VIF):
   All VIF are less than 10. The highest ones are Terminal, Outstate, and PhD.
   These VIFs suggest no collinearity.
3. Compute Tolerance value (TOL):
   All TOLs are bigger than 0.1. The smallest ones are Terminal, Outstate, and PhD.
   These TOLs suggest no collinearity.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	51.39777	6.12404	8.39	<.0001	.	0
numprivate	1	4.61959	1.72185	2.68	0.0075	0.36503	2.73952
Accept_pct	1	-18.10932	3.84314	-4.71	<.0001	0.67266	1.48663
Elite10	1	4.01748	2.00326	2.01	0.0453	0.59245	1.68790
F_Undergrad	1	0.00068095	0.00014285	4.77	<.0001	0.44780	2.23312
P_Undergrad	1	-0.00196	0.00039043	-5.01	<.0001	0.60850	1.64339
Outstate	1	0.00123	0.00022863	5.40	<.0001	0.25413	3.93506
Room_Board	1	0.00167	0.00059443	2.80	0.0052	0.50588	1.97676
Books	1	-0.00252	0.00297	-0.85	0.3951	0.89620	1.11582
Personal	1	-0.00172	0.00077810	-2.21	0.0275	0.77460	1.29098
PhD	1	0.13064	0.05621	2.32	0.0204	0.25525	3.91772
Terminal	1	-0.07284	0.06257	-1.16	0.2447	0.25338	3.94658
S_F_Ratio	1	0.00100	0.16188	0.01	0.9951	0.52364	1.90972
perc_alumni	1	0.30920	0.04839	6.39	<.0001	0.59796	1.67237
Expend	1	-0.00043651	0.00015180	-2.88	0.0041	0.34216	2.92264

f) Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate. You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R^2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, include the outputs.

The first selection is the adj-R^2 method:

Number in Model	Adjusted R-Square	R-Square	Variables in Model
12	0.4356	0.4443	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal perc_alumni Expend
13	0.4353	0.4448	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD Terminal perc_alumni Expend
12	0.4351	0.4438	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD perc_alumni Expend
11	0.4351	0.4431	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD perc_alumni Expend
13	0.4348	0.4443	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal S_F_Ratio perc_alumni Expend
14	0.4346	0.4448	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend
13	0.4343	0.4438	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD S_F_Ratio perc_alumni Expend
12	0.4343	0.4431	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD S_F_Ratio perc_alumni Expend

The best model has the highest Adj R2 = 0.4356 and the number of variables in the model is 12.

The second selection is cp:

Number in Model	C(p)	R-Square	Variables in Model
11	11.3674	0.4431	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD perc_alumni Expend
12	11.7240	0.4443	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal perc_alumni Expend
12	12.3565	0.4438	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD perc_alumni Expend
13	13.0000	0.4448	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD Terminal perc_alumni Expend
12	13.3667	0.4431	numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD S_F_Ratio perc_alumni Expend
10	13.3988	0.4401	numprivate Accept_pct F_Undergrad P_Undergrad Outstate Room_Board Personal PhD perc_alumni Expend
11	13.7091	0.4414	numprivate Accept_pct F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal perc_alumni Expend

The best model is the first one with cp = 11.367 and 11 X-variables. Cp is approximately equal to k+1.

g) Fit a final regression model M1 for Grad.Rate based on the results in f) – i.e. optimal model. Explain your choice. Write down the expression of the estimated model M1.

Root MSE	12.91101	R-Square	0.4431
Dependent Mean	65.46332	Adj R-Sq	0.4351
Coeff Var	19.72251		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	48.40380	4.62103	10.47	<.0001
numprivate	1	4.77018	1.68907	2.82	0.0049
Accept_pct	1	-17.78222	3.79718	-4.68	<.0001
Elite10	1	4.02179	2.00221	2.01	0.0449
F_Undergrad	1	0.00066311	0.00014112	4.70	<.0001
P_Undergrad	1	-0.00196	0.00039013	-5.03	<.0001
Outstate	1	0.00121	0.00022699	5.35	<.0001
Room_Board	1	0.00153	0.00058784	2.61	0.0092
Personal	1	-0.00182	0.00076376	-2.38	0.0174
PhD	1	0.08424	0.03706	2.27	0.0233
perc_alumni	1	0.30598	0.04806	6.37	<.0001
Expend	1	-0.00044650	0.00013904	-3.21	0.0014

Difference = 0.4443 - 0.4351 = 0.0092 = 0.92%

I would choose the first selection with 12 X-variables because Terminal adds 0.92% to the Adj R2.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50.15666	4.81689	10.41	<.0001
numprivate	1	4.53671	1.69814	2.67	0.0077
Accept_pct	1	-17.71061	3.79599	-4.67	<.0001
Elite10	1	3.99877	2.00145	2.00	0.0461
F_Undergrad	1	0.00067321	0.00014128	4.77	<.0001
P_Undergrad	1	-0.00195	0.00039005	-5.00	<.0001
Outstate	1	0.00124	0.00022775	5.44	<.0001
Room_Board	1	0.00162	0.00059173	2.74	0.0062
Personal	1	-0.00183	0.00076352	-2.40	0.0166
PhD	1	0.13733	0.05554	2.47	0.0136
Terminal	1	-0.07955	0.06200	-1.28	0.1999
perc_alumni	1	0.31078	0.04818	6.45	<.0001
Expend	1	-0.00044063	0.00013906	-3.17	0.0016

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	101730	8477.53112	50.90	<.0001
Error	764	127247	166.55344		
Corrected Total	776	228977			

Root MSE	12.90556	R-Square	0.4443
Dependent Mean	65.46332	Adj R-Sq	0.4356
Coeff Var	19.71418		

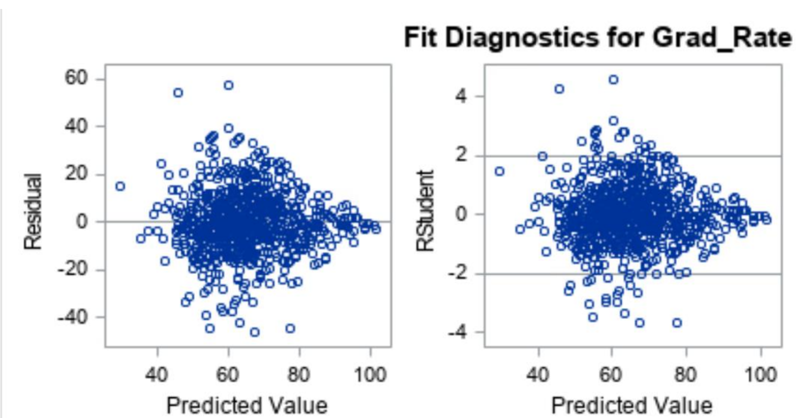
Final Regression Model M1:

Graduation Rate = 50.156 + 4.5367*Private – 17.71*Accept_pct + 3.998*Elite + 0.00067*F_Undergrad – 0.00195*P_Undergrad + 0.00124*Outstate + 0.00162*Room_Board – 0.00183*Personal + 0.137*PhD – 0.07955*Terminal + 0.31*perc_alumni – 0.00044*Expend

Elite = 1 for Elite schools, Elite = 0 for not Elite schools

Private = 1 when school is private, Private = 0 when school is public

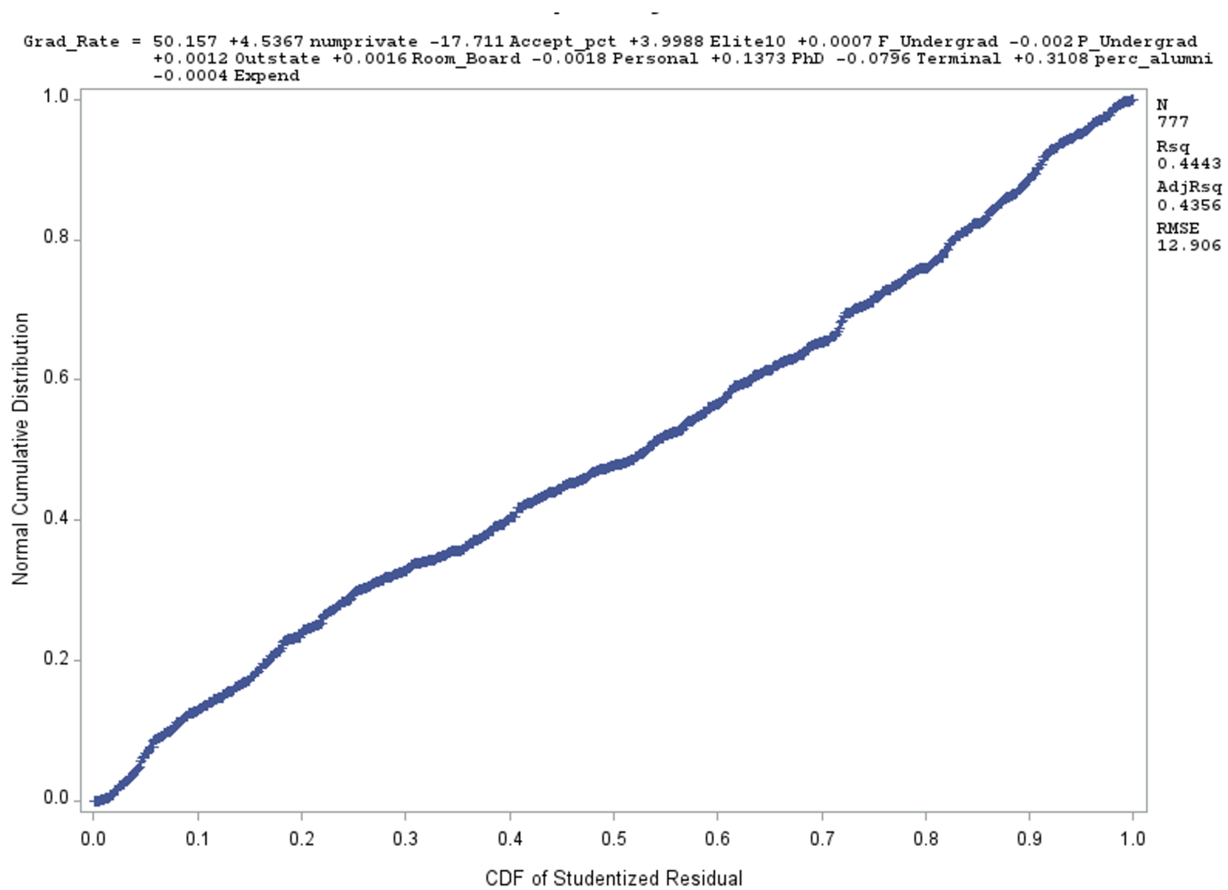
h) Draw a plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis? Include the outputs and explain.



Points are randomly scattered, and residual analysis show no concern for the model fit.

i) Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? Include the outputs and explain.

Plot the normal probability plot of the residuals. It’s almost straight which means it’s normal.



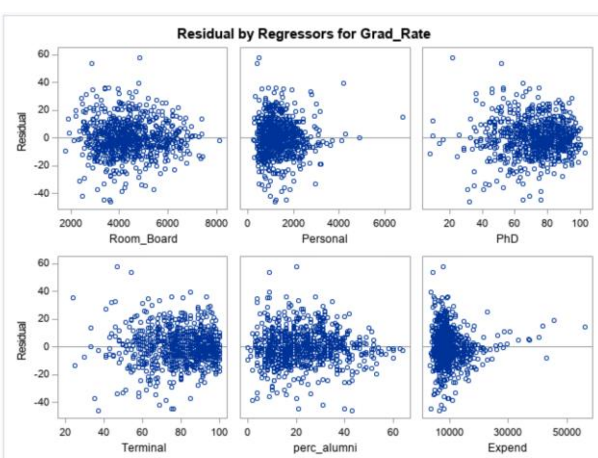
j) Are there any outliers or Influential Points? Compute appropriate statistics. Include the outputs. Take any action you think is necessary and explain why/why not you took these actions?

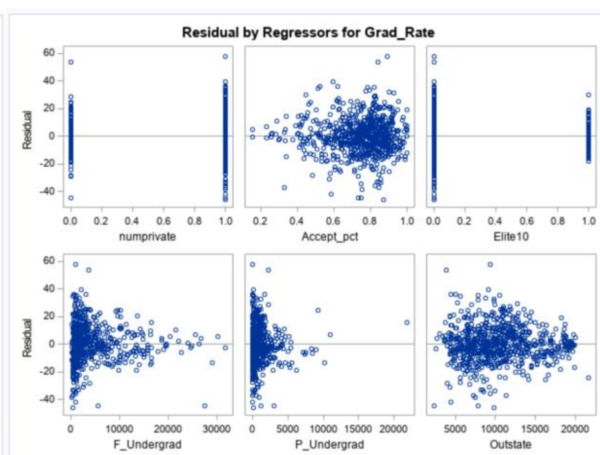
Yes, there are multiple outliers and influential points as they are shown in the figures below.

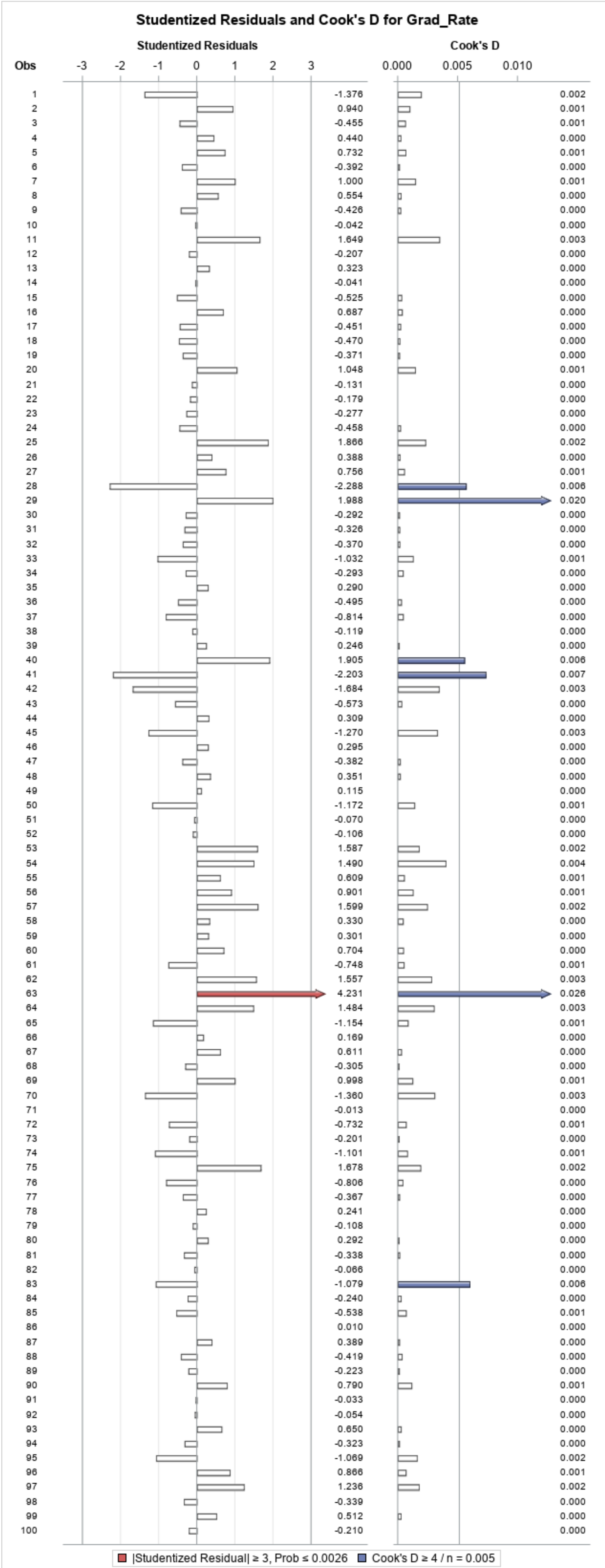
I would exclude all the outliers and influential points and rerun the model again.

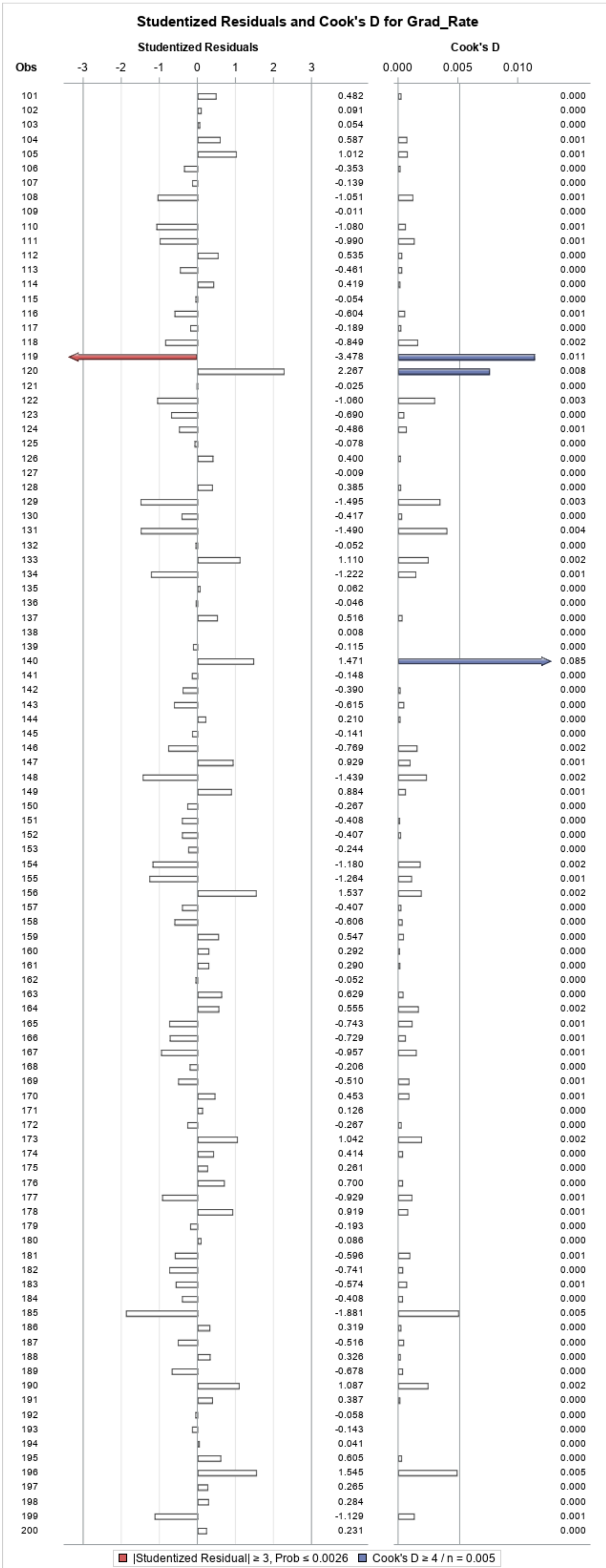
Outliers are the observation numbers (63, 119, 256, 274, 287, 437, 492)

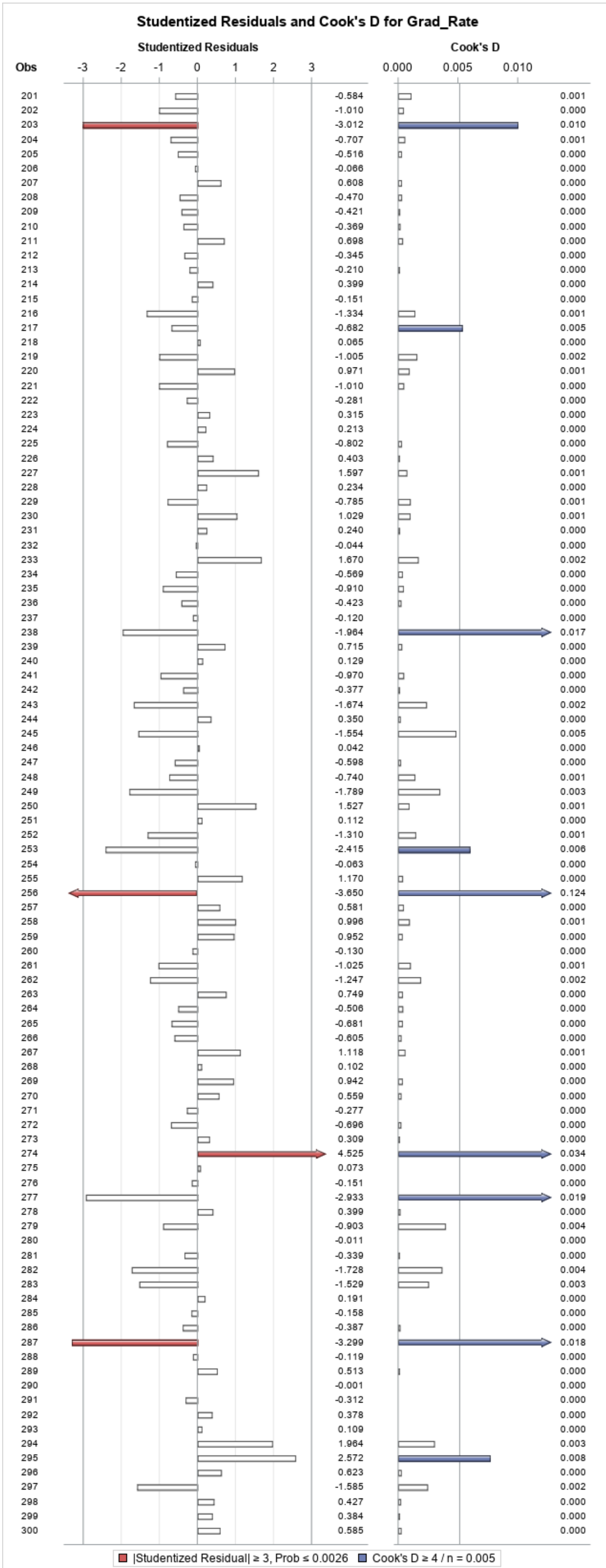
Influential points are the observation numbers (29, 63, 119, 140, 256, 238, 274, 277, 287, 437, 439, 492, 559, 568, 738, 743, 771)

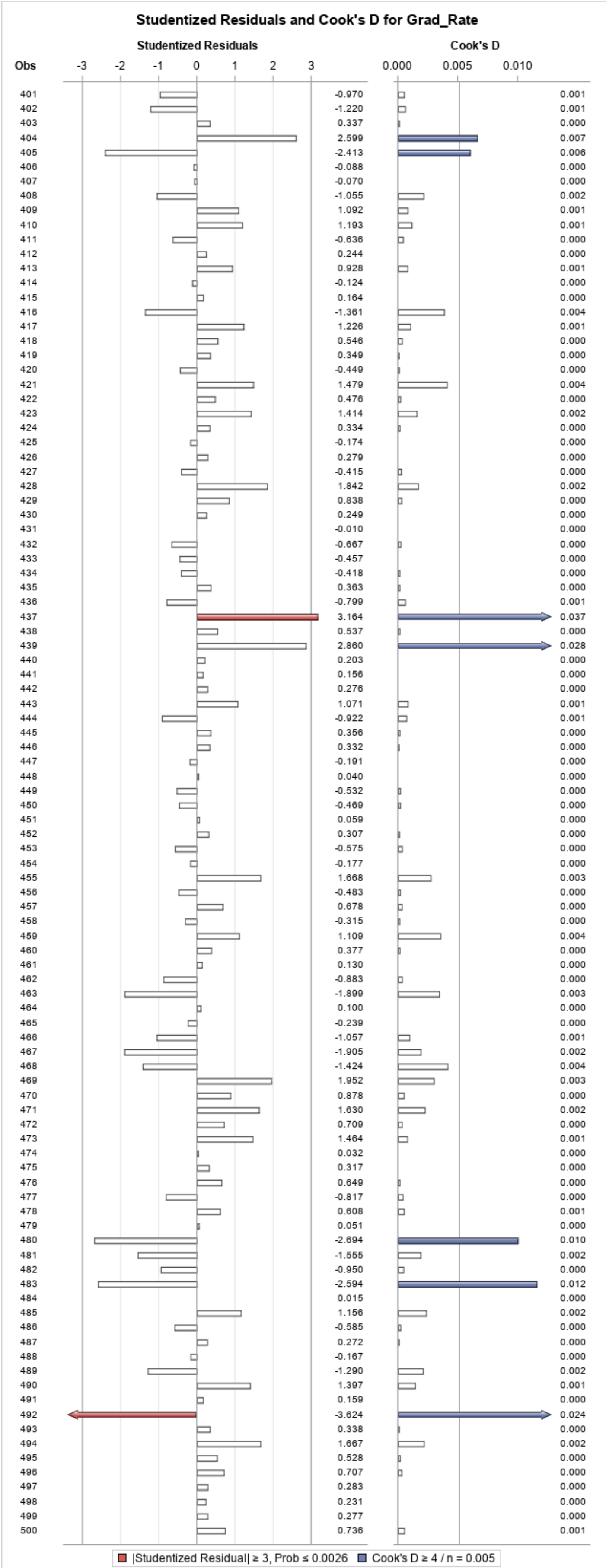




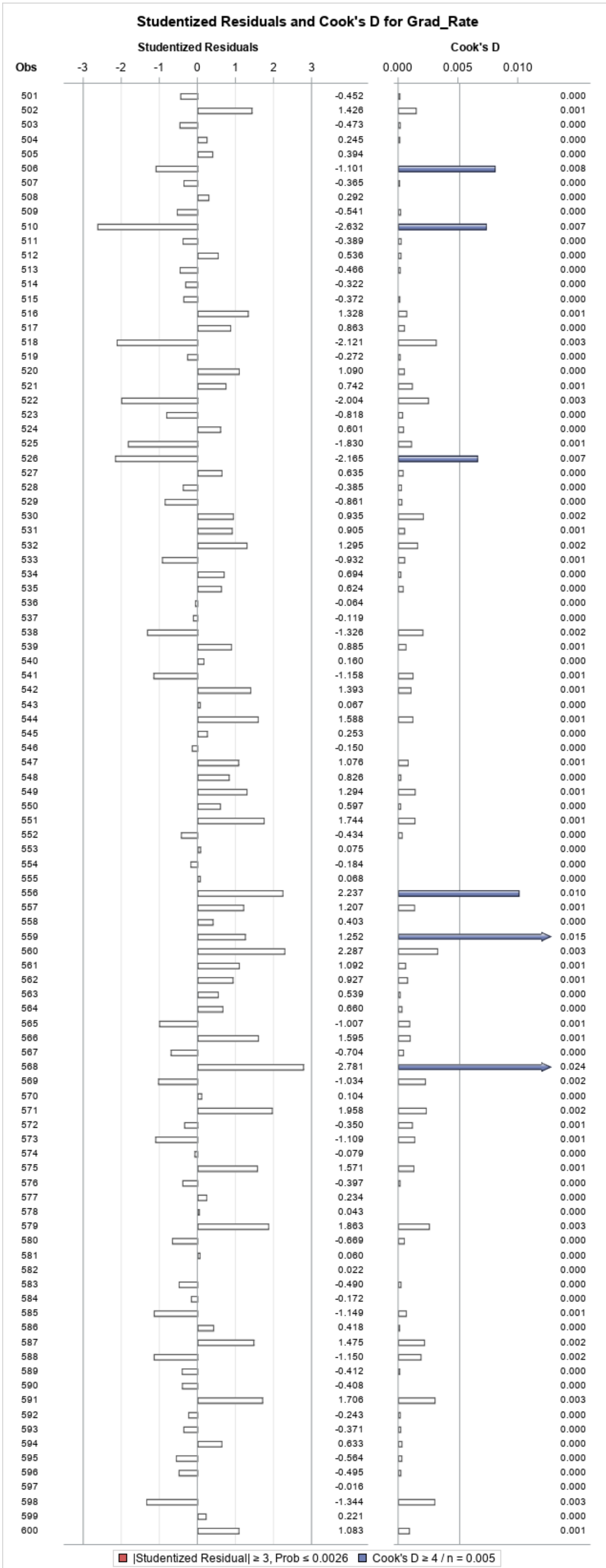

Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
1 -1.376 0.002
2 0.940 0.001
3 -0.455 0.001
4 0.440 0.000
5 0.732 0.001
6 -0.392 0.000
7 1.000 0.001
8 0.554 0.000
9 -0.426 0.000
10 -0.042 0.000
11 1.649 0.003
12 -0.207 0.000
13 0.323 0.000
14 -0.041 0.000
15 -0.525 0.000
16 0.687 0.000
17 -0.451 0.000
18 -0.470 0.000
19 -0.371 0.000
20 1.048 0.001
21 -0.131 0.000
22 -0.179 0.000
23 -0.277 0.000
24 -0.458 0.000
25 1.866 0.002
26 0.388 0.000
27 0.756 0.001
28 -2.288 0.006
29 1.988 0.020
30 -0.292 0.000
31 -0.326 0.000
32 -0.370 0.000
33 -1.032 0.001
34 -0.293 0.000
35 0.290 0.000
36 -0.495 0.000
37 -0.814 0.000
38 -0.119 0.000
39 0.246 0.000
40 1.905 0.006
41 -2.203 0.007
42 -1.684 0.003
43 -0.573 0.000
44 0.309 0.000
45 -1.270 0.003
46 0.295 0.000
47 -0.382 0.000
48 0.351 0.000
49 0.115 0.000
50 -1.172 0.001
51 -0.070 0.000
52 -0.106 0.000
53 1.587 0.002
54 1.490 0.004
55 0.609 0.001
56 0.901 0.001
57 1.599 0.002
58 0.330 0.000
59 0.301 0.000
60 0.704 0.000
61 -0.748 0.001
62 1.557 0.003
63 4.231 0.026
64 1.484 0.003
65 -1.154 0.001
66 0.169 0.000
67 0.611 0.000
68 -0.305 0.000
69 0.998 0.001
70 -1.360 0.003
71 -0.013 0.000
72 -0.732 0.001
73 -0.201 0.000
74 -1.101 0.001
75 1.678 0.002
76 -0.806 0.000
77 -0.367 0.000
78 0.241 0.000
79 -0.108 0.000
80 0.292 0.000
81 -0.338 0.000
82 -0.066 0.000
83 -1.079 0.006
84 -0.240 0.000
85 -0.538 0.001
86 0.010 0.000
87 0.389 0.000
88 -0.419 0.000
89 -0.223 0.000
90 0.790 0.001
91 -0.033 0.000
92 -0.054 0.000
93 0.650 0.000
94 -0.323 0.000
95 -1.069 0.002
96 0.866 0.001
97 1.236 0.002
98 -0.339 0.000
99 0.512 0.000
100 -0.210 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005

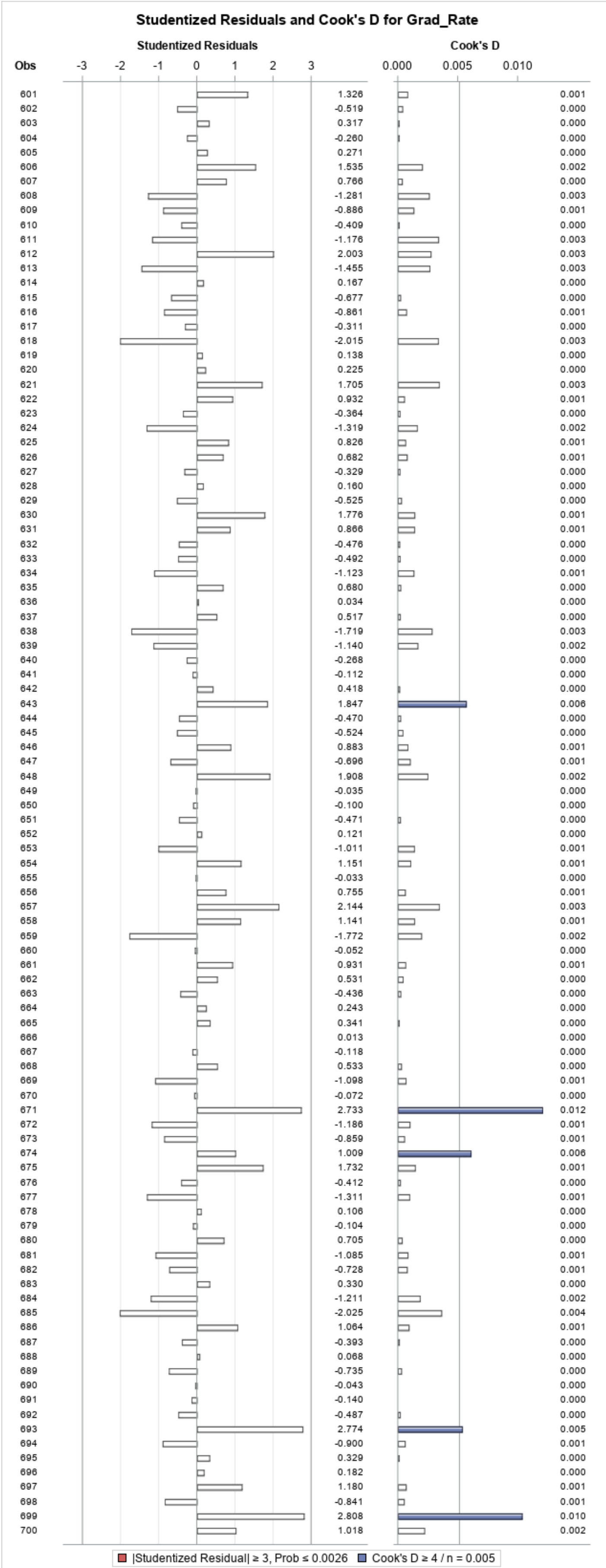

Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
101 0.482 0.000
102 0.091 0.000
103 0.054 0.000
104 0.587 0.001
105 1.012 0.001
106 -0.353 0.000
107 -0.139 0.000
108 -1.051 0.001
109 -0.011 0.000
110 -1.080 0.001
111 -0.990 0.001
112 0.535 0.000
113 -0.461 0.000
114 0.419 0.000
115 -0.054 0.000
116 -0.604 0.001
117 -0.189 0.000
118 -0.849 0.002
119 -3.478 0.011
120 2.267 0.008
121 -0.025 0.000
122 -1.060 0.003
123 -0.690 0.000
124 -0.486 0.001
125 -0.078 0.000
126 0.400 0.000
127 -0.009 0.000
128 0.385 0.000
129 -1.495 0.003
130 -0.417 0.000
131 -1.490 0.004
132 -0.052 0.000
133 1.110 0.002
134 -1.222 0.001
135 0.062 0.000
136 -0.046 0.000
137 0.516 0.000
138 0.008 0.000
139 -0.115 0.000
140 1.471 0.085
141 -0.148 0.000
142 -0.390 0.000
143 -0.615 0.000
144 0.210 0.000
145 -0.141 0.000
146 -0.769 0.002
147 0.929 0.001
148 -1.439 0.002
149 0.884 0.001
150 -0.267 0.000
151 -0.408 0.000
152 -0.407 0.000
153 -0.244 0.000
154 -1.180 0.002
155 -1.264 0.001
156 1.537 0.002
157 -0.407 0.000
158 -0.606 0.000
159 0.547 0.000
160 0.292 0.000
161 0.290 0.000
162 -0.052 0.000
163 0.629 0.000
164 0.555 0.002
165 -0.743 0.001
166 -0.729 0.001
167 -0.957 0.001
168 -0.206 0.000
169 -0.510 0.001
170 0.453 0.001
171 0.126 0.000
172 -0.267 0.000
173 1.042 0.002
174 0.414 0.000
175 0.261 0.000
176 0.700 0.000
177 -0.929 0.001
178 0.919 0.001
179 -0.193 0.000
180 0.086 0.000
181 -0.596 0.001
182 -0.741 0.000
183 -0.574 0.001
184 -0.408 0.000
185 -1.881 0.005
186 0.319 0.000
187 -0.516 0.000
188 0.326 0.000
189 -0.678 0.000
190 1.087 0.002
191 0.387 0.000
192 -0.058 0.000
193 -0.143 0.000
194 0.041 0.000
195 0.605 0.000
196 1.545 0.005
197 0.265 0.000
198 0.284 0.000
199 -1.129 0.001
200 0.231 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005

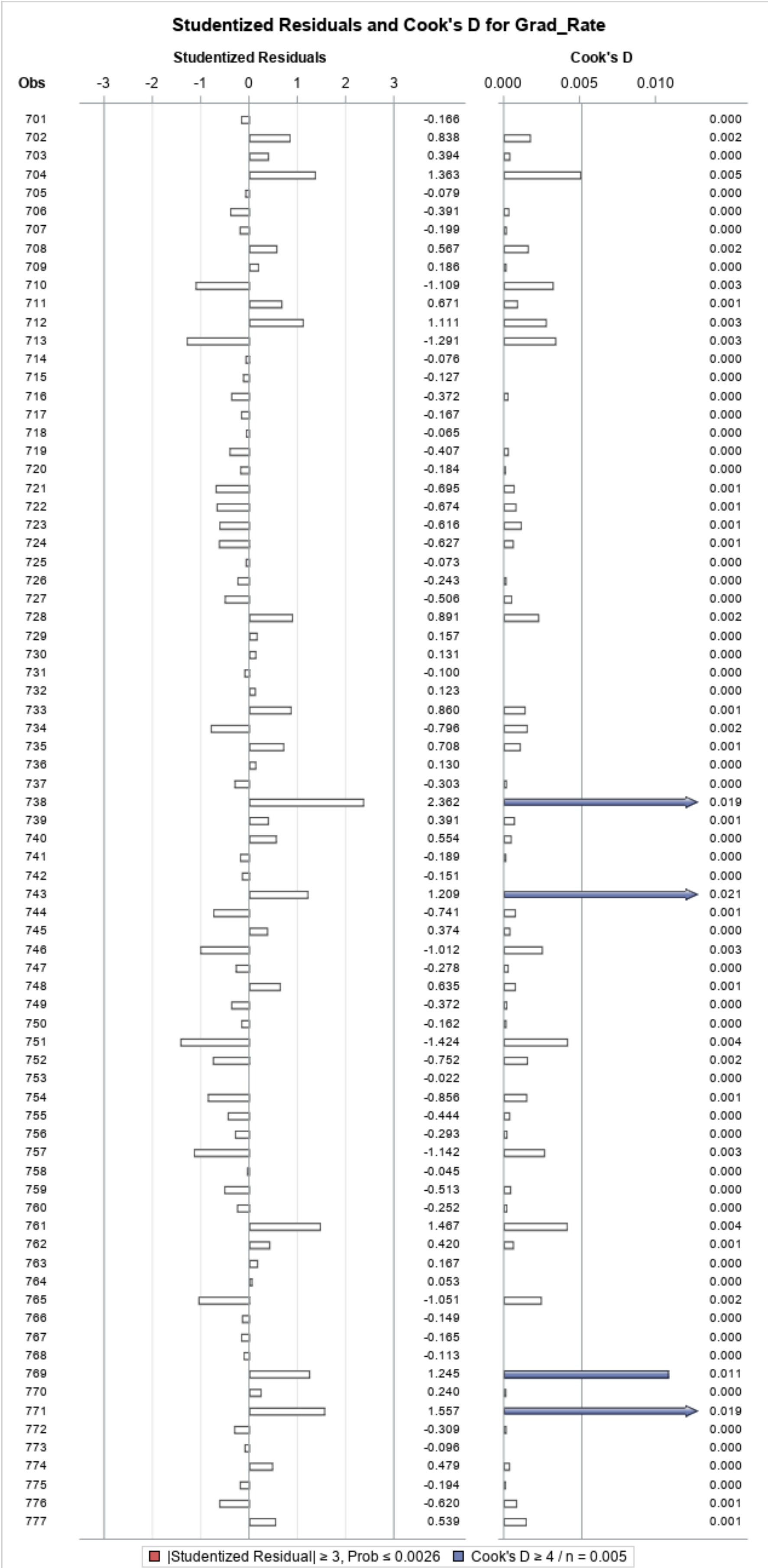

Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
201 -0.584 0.001
202 -1.010 0.000
203 -3.012 0.010
204 -0.707 0.001
205 -0.516 0.000
206 -0.066 0.000
207 0.608 0.000
208 -0.470 0.000
209 -0.421 0.000
210 -0.369 0.000
211 0.698 0.000
212 -0.345 0.000
213 -0.210 0.000
214 0.399 0.000
215 -0.151 0.000
216 -1.334 0.001
217 -0.682 0.005
218 0.065 0.000
219 -1.005 0.002
220 0.971 0.001
221 -1.010 0.000
222 -0.281 0.000
223 0.315 0.000
224 0.213 0.000
225 -0.802 0.000
226 0.403 0.000
227 1.597 0.001
228 0.234 0.000
229 -0.785 0.001
230 1.029 0.001
231 0.240 0.000
232 -0.044 0.000
233 1.670 0.002
234 -0.569 0.000
235 -0.910 0.000
236 -0.423 0.000
237 -0.120 0.000
238 -1.964 0.017
239 0.715 0.000
240 0.129 0.000
241 -0.970 0.000
242 -0.377 0.000
243 -1.674 0.002
244 0.350 0.000
245 -1.554 0.005
246 0.042 0.000
247 -0.598 0.000
248 -0.740 0.001
249 -1.789 0.003
250 1.527 0.001
251 0.112 0.000
252 -1.310 0.001
253 -2.415 0.006
254 -0.063 0.000
255 1.170 0.000
256 -3.650 0.124
257 0.581 0.000
258 0.996 0.001
259 0.952 0.000
260 -0.130 0.000
261 -1.025 0.001
262 -1.247 0.002
263 0.749 0.000
264 -0.506 0.000
265 -0.681 0.000
266 -0.605 0.000
267 1.118 0.001
268 0.102 0.000
269 0.942 0.000
270 0.559 0.000
271 -0.277 0.000
272 -0.696 0.000
273 0.309 0.000
274 4.525 0.034
275 0.073 0.000
276 -0.151 0.000
277 -2.933 0.019
278 0.399 0.000
279 -0.903 0.004
280 -0.011 0.000
281 -0.339 0.000
282 -1.728 0.004
283 -1.529 0.003
284 0.191 0.000
285 -0.158 0.000
286 -0.387 0.000
287 -3.299 0.018
288 -0.119 0.000
289 0.513 0.000
290 -0.001 0.000
291 -0.312 0.000
292 0.378 0.000
293 0.109 0.000
294 1.964 0.003
295 2.572 0.008
296 0.623 0.000
297 -1.585 0.002
298 0.427 0.000
299 0.384 0.000
300 0.585 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005



Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
401 -0.970 0.001
402 -1.220 0.001
403 0.337 0.000
404 2.599 0.007
405 -2.413 0.006
406 -0.088 0.000
407 -0.070 0.000
408 -1.055 0.002
409 1.092 0.001
410 1.193 0.001
411 -0.636 0.000
412 0.244 0.000
413 0.928 0.001
414 -0.124 0.000
415 0.164 0.000
416 -1.361 0.004
417 1.226 0.001
418 0.546 0.000
419 0.349 0.000
420 -0.449 0.000
421 1.479 0.004
422 0.476 0.000
423 1.414 0.002
424 0.334 0.000
425 -0.174 0.000
426 0.279 0.000
427 -0.415 0.000
428 1.842 0.002
429 0.838 0.000
430 0.249 0.000
431 -0.010 0.000
432 -0.667 0.000
433 -0.457 0.000
434 -0.418 0.000
435 0.363 0.000
436 -0.799 0.001
437 3.164 0.037
438 0.537 0.000
439 2.860 0.028
440 0.203 0.000
441 0.156 0.000
442 0.276 0.000
443 1.071 0.001
444 -0.922 0.001
445 0.356 0.000
446 0.332 0.000
447 -0.191 0.000
448 0.040 0.000
449 -0.532 0.000
450 -0.469 0.000
451 0.059 0.000
452 0.307 0.000
453 -0.575 0.000
454 -0.177 0.000
455 1.668 0.003
456 -0.483 0.000
457 0.678 0.000
458 -0.315 0.000
459 1.109 0.004
460 0.377 0.000
461 0.130 0.000
462 -0.883 0.000
463 -1.899 0.003
464 0.100 0.000
465 -0.239 0.000
466 -1.057 0.001
467 -1.905 0.002
468 -1.424 0.004
469 1.952 0.003
470 0.878 0.000
471 1.630 0.002
472 0.709 0.000
473 1.464 0.001
474 0.032 0.000
475 0.317 0.000
476 0.649 0.000
477 -0.817 0.000
478 0.608 0.001
479 0.051 0.000
480 -2.694 0.010
481 -1.555 0.002
482 -0.950 0.000
483 -2.594 0.012
484 0.015 0.000
485 1.156 0.002
486 -0.585 0.000
487 0.272 0.000
488 -0.167 0.000
489 -1.290 0.002
490 1.397 0.001
491 0.159 0.000
492 -3.624 0.024
493 0.338 0.000
494 1.667 0.002
495 0.528 0.000
496 0.707 0.000
497 0.283 0.000
498 0.231 0.000
499 0.277 0.000
500 0.736 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005


Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
501 -0.452 0.000
502 1.426 0.001
503 -0.473 0.000
504 0.245 0.000
505 0.394 0.000
506 -1.101 0.008
507 -0.365 0.000
508 0.292 0.000
509 -0.541 0.000
510 -2.632 0.007
511 -0.389 0.000
512 0.536 0.000
513 -0.466 0.000
514 -0.322 0.000
515 -0.372 0.000
516 1.328 0.001
517 0.863 0.000
518 -2.121 0.003
519 -0.272 0.000
520 1.090 0.000
521 0.742 0.001
522 -2.004 0.003
523 -0.818 0.000
524 0.601 0.000
525 -1.830 0.001
526 -2.165 0.007
527 0.635 0.000
528 -0.385 0.000
529 -0.861 0.000
530 0.935 0.002
531 0.905 0.001
532 1.295 0.002
533 -0.932 0.001
534 0.694 0.000
535 0.624 0.000
536 -0.064 0.000
537 -0.119 0.000
538 -1.326 0.002
539 0.885 0.001
540 0.160 0.000
541 -1.158 0.001
542 1.393 0.001
543 0.067 0.000
544 1.588 0.001
545 0.253 0.000
546 -0.150 0.000
547 1.076 0.001
548 0.826 0.000
549 1.294 0.001
550 0.597 0.000
551 1.744 0.001
552 -0.434 0.000
553 0.075 0.000
554 -0.184 0.000
555 0.068 0.000
556 2.237 0.010
557 1.207 0.001
558 0.403 0.000
559 1.252 0.015
560 2.287 0.003
561 1.092 0.001
562 0.927 0.001
563 0.539 0.000
564 0.660 0.000
565 -1.007 0.001
566 1.595 0.001
567 -0.704 0.000
568 2.781 0.024
569 -1.034 0.002
570 0.104 0.000
571 1.958 0.002
572 -0.350 0.001
573 -1.109 0.001
574 -0.079 0.000
575 1.571 0.001
576 -0.397 0.000
577 0.234 0.000
578 0.043 0.000
579 1.863 0.003
580 -0.669 0.000
581 0.060 0.000
582 0.022 0.000
583 -0.490 0.000
584 -0.172 0.000
585 -1.149 0.001
586 0.418 0.000
587 1.475 0.002
588 -1.150 0.002
589 -0.412 0.000
590 -0.408 0.000
591 1.706 0.003
592 -0.243 0.000
593 -0.371 0.000
594 0.633 0.000
595 -0.564 0.000
596 -0.495 0.000
597 -0.016 0.000
598 -1.344 0.003
599 0.221 0.000
600 1.083 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005


Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
601 1.326 0.001
602 -0.519 0.000
603 0.317 0.000
604 -0.260 0.000
605 0.271 0.000
606 1.535 0.002
607 0.766 0.000
608 -1.281 0.003
609 -0.886 0.001
610 -0.409 0.000
611 -1.176 0.003
612 2.003 0.003
613 -1.455 0.003
614 0.167 0.000
615 -0.677 0.000
616 -0.861 0.001
617 -0.311 0.000
618 -2.015 0.003
619 0.138 0.000
620 0.225 0.000
621 1.705 0.003
622 0.932 0.001
623 -0.364 0.000
624 -1.319 0.002
625 0.826 0.001
626 0.682 0.001
627 -0.329 0.000
628 0.160 0.000
629 -0.525 0.000
630 1.776 0.001
631 0.866 0.001
632 -0.476 0.000
633 -0.492 0.000
634 -1.123 0.001
635 0.680 0.000
636 0.034 0.000
637 0.517 0.000
638 -1.719 0.003
639 -1.140 0.002
640 -0.268 0.000
641 -0.112 0.000
642 0.418 0.000
643 1.847 0.006
644 -0.470 0.000
645 -0.524 0.000
646 0.883 0.001
647 -0.696 0.001
648 1.908 0.002
649 -0.035 0.000
650 -0.100 0.000
651 -0.471 0.000
652 0.121 0.000
653 -1.011 0.001
654 1.151 0.001
655 -0.033 0.000
656 0.755 0.001
657 2.144 0.003
658 1.141 0.001
659 -1.772 0.002
660 -0.052 0.000
661 0.931 0.001
662 0.531 0.000
663 -0.436 0.000
664 0.243 0.000
665 0.341 0.000
666 0.013 0.000
667 -0.118 0.000
668 0.533 0.000
669 -1.098 0.001
670 -0.072 0.000
671 2.733 0.012
672 -1.186 0.001
673 -0.859 0.001
674 1.009 0.006
675 1.732 0.001
676 -0.412 0.000
677 -1.311 0.001
678 0.106 0.000
679 -0.104 0.000
680 0.705 0.000
681 -1.085 0.001
682 -0.728 0.001
683 0.330 0.000
684 -1.211 0.002
685 -2.025 0.004
686 1.064 0.001
687 -0.393 0.000
688 0.068 0.000
689 -0.735 0.000
690 -0.043 0.000
691 -0.140 0.000
692 -0.487 0.000
693 2.774 0.005
694 -0.900 0.001
695 0.329 0.000
696 0.182 0.000
697 1.180 0.001
698 -0.841 0.001
699 2.808 0.010
700 1.018 0.002
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005


Studentized Residuals and Cook's D for Grad_Rate
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010
701 -0.166 0.000
702 0.838 0.002
703 0.394 0.000
704 1.363 0.005
705 -0.079 0.000
706 -0.391 0.000
707 -0.199 0.000
708 0.567 0.002
709 0.186 0.000
710 -1.109 0.003
711 0.671 0.001
712 1.111 0.003
713 -1.291 0.003
714 -0.076 0.000
715 -0.127 0.000
716 -0.372 0.000
717 -0.167 0.000
718 -0.065 0.000
719 -0.407 0.000
720 -0.184 0.000
721 -0.695 0.001
722 -0.674 0.001
723 -0.616 0.001
724 -0.627 0.001
725 -0.073 0.000
726 -0.243 0.000
727 -0.506 0.000
728 0.891 0.002
729 0.157 0.000
730 0.131 0.000
731 -0.100 0.000
732 0.123 0.000
733 0.860 0.001
734 -0.796 0.002
735 0.708 0.001
736 0.130 0.000
737 -0.303 0.000
738 2.362 0.019
739 0.391 0.001
740 0.554 0.000
741 -0.189 0.000
742 -0.151 0.000
743 1.209 0.021
744 -0.741 0.001
745 0.374 0.000
746 -1.012 0.003
747 -0.278 0.000
748 0.635 0.001
749 -0.372 0.000
750 -0.162 0.000
751 -1.424 0.004
752 -0.752 0.002
753 -0.022 0.000
754 -0.856 0.001
755 -0.444 0.000
756 -0.293 0.000
757 -1.142 0.003
758 -0.045 0.000
759 -0.513 0.000
760 -0.252 0.000
761 1.467 0.004
762 0.420 0.001
763 0.167 0.000
764 0.053 0.000
765 -1.051 0.002
766 -0.149 0.000
767 -0.165 0.000
768 -0.113 0.000
769 1.245 0.011
770 0.240 0.000
771 1.557 0.019
772 -0.309 0.000
773 -0.096 0.000
774 0.479 0.000
775 -0.194 0.000
776 -0.620 0.001
777 0.539 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.005

After removing all these 17 observations and rerun the model, there will be no significant outliers in the new dataset.

k) Analyze the AdjR^2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.

Number of Observations Read	760
Number of Observations Used	760

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	105928	8827.36271	63.77	<.0001
Error	747	103409	138.43237		
Corrected Total	759	209337			

Root MSE	11.76573	R-Square	0.5060
Dependent Mean	65.44079	Adj R-Sq	0.4981
Coeff Var	17.97920		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50.11302	4.55976	10.99	<.0001
numprivate	1	4.92773	1.61092	3.06	0.0023
Accept_pct	1	-18.35460	3.53220	-5.20	<.0001
Elite10	1	2.91244	1.85202	1.57	0.1162
F_Undergrad	1	0.00091273	0.00013843	6.59	<.0001
P_Undergrad	1	-0.00269	0.00045261	-5.95	<.0001
Outstate	1	0.00128	0.00021934	5.84	<.0001
Room_Board	1	0.00170	0.00054863	3.09	0.0021
Personal	1	-0.00215	0.00075699	-2.83	0.0047
PhD	1	0.14337	0.05253	2.73	0.0065
Terminal	1	-0.08179	0.05917	-1.38	0.1673
perc_alumni	1	0.35439	0.04505	7.87	<.0001
Expend	1	-0.00061098	0.00014575	-4.19	<.0001

The Adj R2 = 0.4981 which is the highest so far. This model has p-value <.0001 that is almost 0.

l) Draw conclusions on graduation rates based on your regression analysis. What are the most important predictors in your model? Does your model show a significant difference in graduation rates between private and public universities? Do "elite" universities have higher graduation rates? Explain.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	50.11302	4.55976	10.99	<.0001	0
numprivate	1	4.92773	1.61092	3.06	0.0023	0.13238
Accept_pct	1	-18.35460	3.53220	-5.20	<.0001	-0.16223
Elite10	1	2.91244	1.85202	1.57	0.1162	0.05234
F_Undergrad	1	0.00091273	0.00013843	6.59	<.0001	0.26324
P_Undergrad	1	-0.00269	0.00045261	-5.95	<.0001	-0.21004
Outstate	1	0.00128	0.00021934	5.84	<.0001	0.30726
Room_Board	1	0.00170	0.00054863	3.09	0.0021	0.11179
Personal	1	-0.00215	0.00075699	-2.83	0.0047	-0.08224
PhD	1	0.14337	0.05253	2.73	0.0065	0.13914
Terminal	1	-0.08179	0.05917	-1.38	0.1673	-0.07134
perc_alumni	1	0.35439	0.04505	7.87	<.0001	0.26506
Expend	1	-0.00061098	0.00014575	-4.19	<.0001	-0.17596

The most important predictors are the one with the highest absolute value of standardized estimate which are Outstate, then (pwec_alumni, F_Undergrad, P_Undergrad) respectively. They also have the smallest p-values which are almost zero. They have a significant effect on graduation rates.

My model shows a significant difference between private and public schools. If a school changes from public to private, the graduation rates according to my model will increase by 4.927%.

Elite universities show 2.91% higher graduation rates than not elite universities.

My Last Model:

Graduation Rate = 50.113 + 4.9277*Private – 18.354*Accept_pct + 2.912*Elite + 0.000912*F_Undergrad – 0.00269*P_Undergrad + 0.00128*Outstate + 0.0017*Room_Board – 0.00215*Personal + 0.1433*PhD – 0.08179*Terminal + 0.35439*perc_alumni – 0.00061*Expend

Elite = 1 for Elite schools, Elite = 0 for not Elite schools

Private = 1 when school is private, Private = 0 when school is public

m) Copy and paste your FULL SAS code into the word document along with your answers.

```
*a;
*import data from file;
proc import datafile="S:\HW5\College.csv" out=myd replace;
delimiter=',';
getnames=yes;
run;

*Create dummy variable for Private;
data college;
set myd;
numprivate = 1;
if Private = 'No' then numprivate = 0;
run;

Proc print data = college (obs= 20);
run;

title "HISTOGRAM of Grad_Rate";
proc univariate normal;
var Grad_Rate;
histogram / normal (mu = est sigma = est);
run;

*b;
proc sgscatter;
matrix Grad_Rate numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni
Expend;
run;

proc gplot;
plot Grad_Rate*(numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni
Expend);
run;

proc corr;
var Grad_Rate numprivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend;
run;

*c;
*Boxplot - by Private;
proc sort;
```

```
by Private;
RUN;

PROC BOXPLOT;
PLOT Grad_Rate*Private;
RUN;

proc sort;
by Elite10;
RUN;

*Boxplot - by Elite;
PROC BOXPLOT;
PLOT Grad_Rate*Elite10;
RUN;

*d;
*Model 1- full model with all predictors;
proc reg;
model Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend
/stb;
run;

*e;
proc reg;
model Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend
/vif tol;
run;


*f;
PROC REG data=college ;
*Backward selection method;
MODEL Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni
Expend/SELECTION = adjrsq;
run;

PROC REG data=college ;
*CP selection method;
MODEL Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni
Expend/SELECTION = cp;
run;


*g;
PROC REG data=college ;
MODEL Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD perc_alumni Expend;
run;

*First Model;
PROC REG data=college ;
MODEL Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend;
run;
```

```
*h;
plot student.*(numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend);
* Residual plot: residuals vs pred. values;
plot student.*predicted.;

* Normal probability plot or QQ plot;
plot npp.*student.;
run;

*i;
proc reg data = college;
model Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend/influence r;
plot student.*(numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend predicted.);
plot npp.*student.;
run;

*j;
*deleting Multiple observations;
data college_new;
set college;
if _n_ in (29 63 119 140 256 238 274 277 287 437 439 492 559 568 738 743
771) then delete;
run;

*rerunning the model without outlier using the new dataset;
proc reg data = college_new;
model Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend/influence r;
plot student.*(numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend predicted.);
plot npp.*student.;
run;

*k;
proc reg data = college_new;
model Grad_Rate=numprivate Accept_pct Elite10 F_Undergrad P_Undergrad
Outstate Room_Board Personal PhD Terminal perc_alumni Expend/stb;
run;
```