1. **PROBLEM 1 [25 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:

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g., $ or $ 99 million, %, etc.) otherwise points will be deducted.***

page1image28233280

*School Private Accept.pct Elite10 F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate*

Apply regression model to predict

Apply regression analysis techniques to analyze the relationship among the observed variables and build a Graduation Rates (Grad.Rate). **Note: Depending on how you import your data (INFILE or when you use a proc print.**

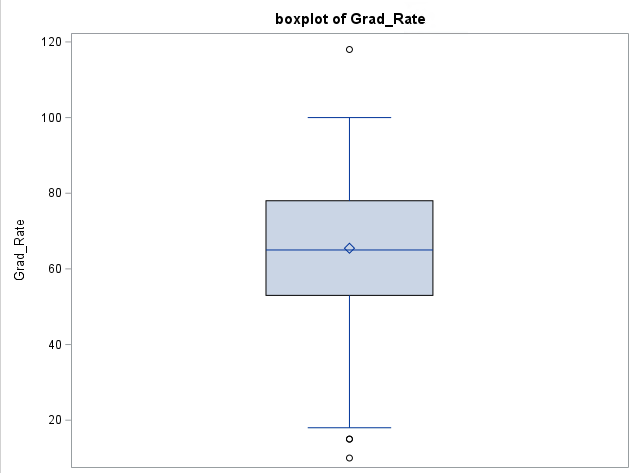
A screenshot of a computer

Description automatically generated with medium confidence

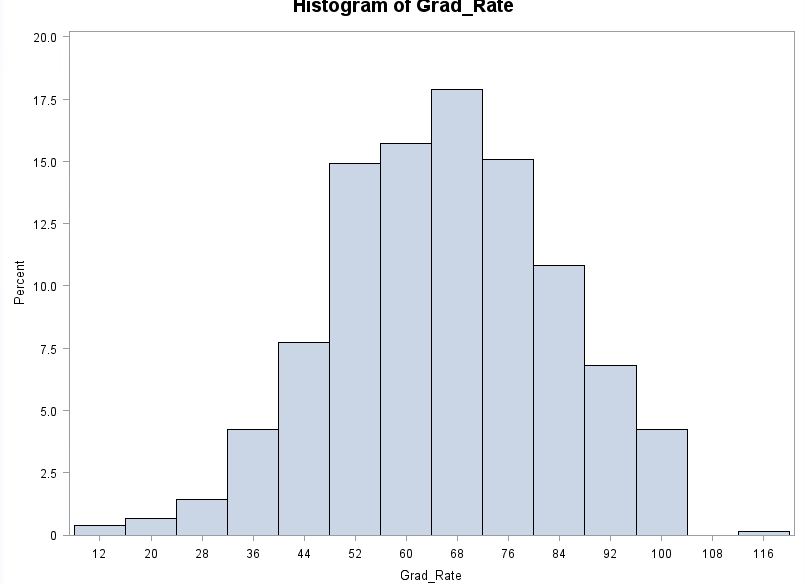
***Note: Before you start, open the college.csv file, and examine the data.***

Answer the following questions.

1. 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).
   1. **Data Exploration:**
      1. **Visualize (Boxplots, histograms)**
         1. **Boxplot:**

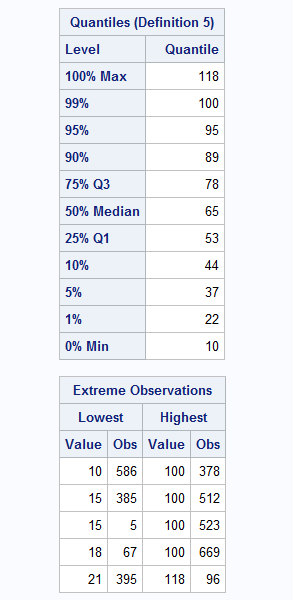
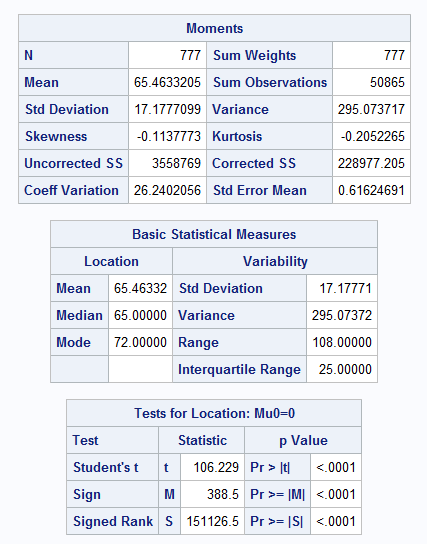


* + - 1. **Histogram:**

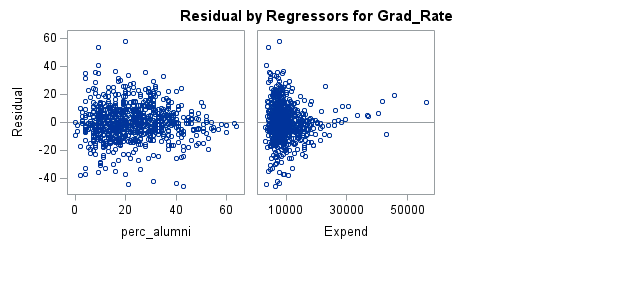


**Discussion:** If we review the histogram and boxplot, we will notice that grade rate has at least one observation above 100%. It is statistically impossible to reach a 118% graduation rate and is wrongly manipulating the dataset. Therefore, this outlier should be removed for data mesentery. There also exist two outliers (as seen in the boxplot) these should be considered for removal after further analysis since they can disrupt the predictive model. We can also determine the central tendency of our data via our visualization's above, which appears somewhat left skewed (this may change to a normal curve if we remove the data entry error).

* + 1. **Univariate for extreme observations** 
       1. **Univariate Procedure**



**Discussion:** Considering the univariate procedure’s extreme observations table we notice 10 outliers including the data entry error (value 118, obs 96), first noted in the above graphs. In our univariate procedure we can now note the specific observations which represent our outliers. We should remove observation 96 and consider removing our other extreme observations to create the most representative predictive model.

* + 1. **Transformation: Should we transform our model? Answer: Yes**
       1. **Residual Plot**
          1. 
          2. 
          3. 

**Discussion:** To detect problems and anomalies we can refer to our residual plots and check assumptions of independence and variance (excluding our qualitative variable d\_private and Elite10).

**Residual Plots Descriptions per X-Variable:**

**Accept\_pct:** Appears funnel-like with a greater proportion of observations lying beyond the 0.5 Accept\_pct. Overall problematic, cannot conclude independence and constant variance due to funnel-like behavior.

**F\_Undergrad:** Appears funnel-like with a greater proportion of observations lying below 10,000 full time undergraduate students. Overall problematic, we cannot conclude independence and constant variance due to funnel-like behavior.

**P\_Undergrad:** Appears funnel-like with a greater proportion of observations lying below 5,000 full time undergraduate students. Overall problematic, we cannot conclude independence and constant variance due to funnel-like behavior.

**OutState:** Appears to clear all assumptions. Maintains constant variance, independence, and points are randomly scattered across the zero-line. Overall, unproblematic.

**Room\_Board:** Appears to clear all assumptions. Maintains constant variance, independence, and points are randomly scattered across the zero-line. Overall, unproblematic.

**Books:** Although it appears unbalanced towards the left, the residual plot clears all assumptions. Maintains constant variance, independence, and linearity across the zero-line. Overall, unproblematic.

**Personal:** Although it appears unbalanced towards the left, the residual plot clears all assumptions. Maintains constant variance, independence, and linearity across the zero-line. Overall, unproblematic.

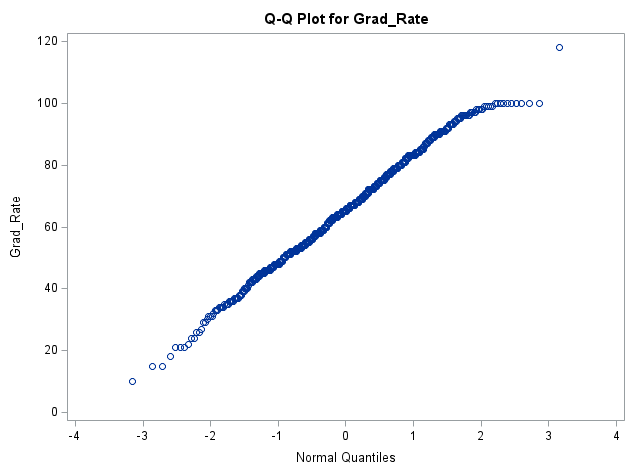
**PhD:** Potentially problematic, appears to be funneling. Overall problematic, we cannot conclude independence and constant variance due to funnel-like behavior.

**Terminal:** Potentially problematic, appears to be funneling. Overall problematic, we cannot conclude independence and constant variance due to funnel-like behavior.

**S\_F\_Ratio:** Appears to clear all assumptions. Maintains constant variance, independence, and points are randomly scattered across the zero-line. Overall, unproblematic.

**perc\_alumni:** Potentially problematic, appears to be funneling. Overall problematic, we cannot conclude independence and constant variance due to funnel-like behavior.

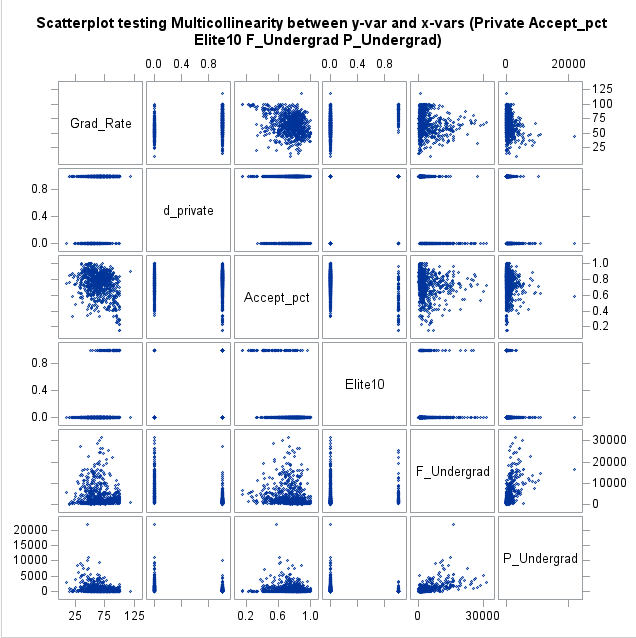
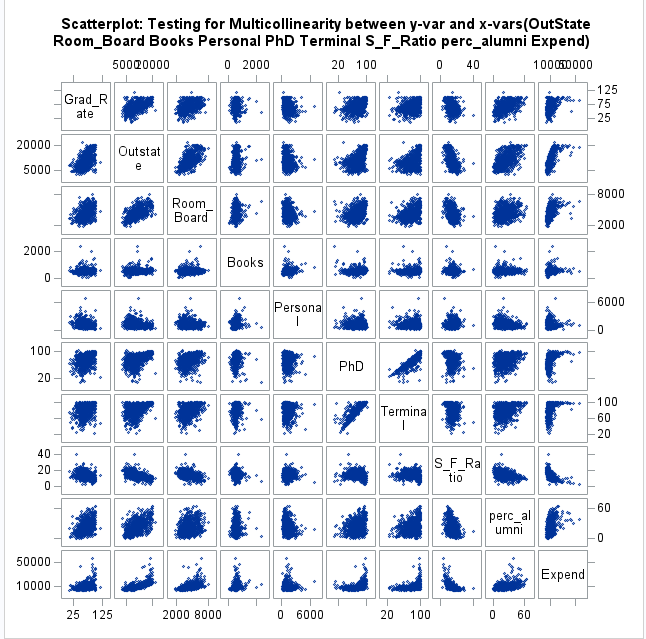
**Expand:** Appears funnel-like with a greater proportion of observations lying below 30,000 instructional expenditure per student. Overall problematic, we cannot conclude independence and constant variance due to funnel-like behavior.

* + - 1. **Normal Plot**
         1. 

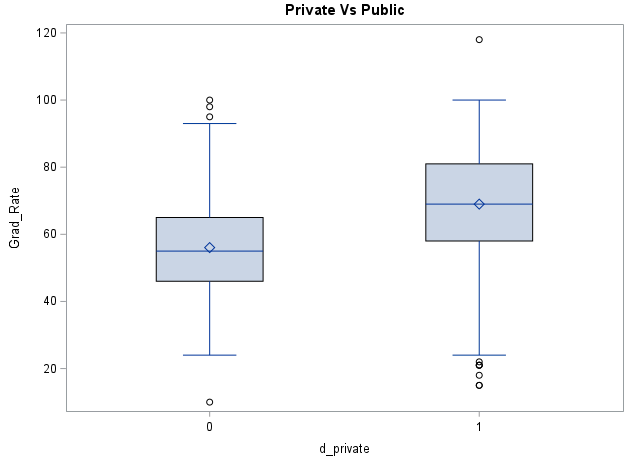
**Discussion:** Normal plot appears overall linear with minor curving towards quantile 3.

* + 1. **Flagging for multicollinearity (R, scatterplot)**
       1. **VIF & Tol**
       2. **(Note: Tol <= 0.10 suggests collinearity. VIF >= 10.0 suggests collinearity)**
          1. 

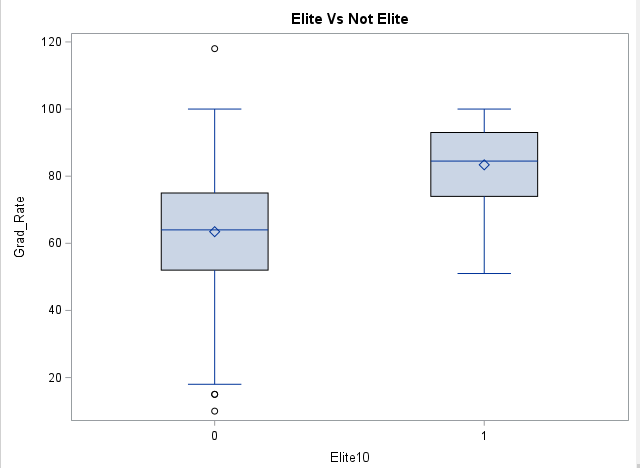
**Discussion:** As noted above, Tol <= 0.10 suggests collinearity, VIF >= 10.0 suggests collinearity. In terms of multicollinearity, the parameter estimates suggest that no x-variable are collinear. However, p-values propose that there does exist insignificant values to the model, such as Books, Terminal, and S\_F\_Ratio. We will conduct further analysis to argue their significance.

* + - 1. **Matrix 1:**
         1. 
      2. **Matrix 2:**
         1. 

1. b) Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationship between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission).
   1. **Discussion:** We can conclude that there exists no multicollinearity (no collinear pairs) and whether there exists a linear relationship (further emphasized by R/Adj-R2/F-value).
   2. **Linearity: Description per x-variable**
      1. **Accept\_Pct: Does not appear linear**
         1. **R: 0.2871** **R2: 0.0824** **Adj-R2:** **0.0812** **F-value: 69.55**
         2. **Suggests weak linearity**
         3. **F-value suggests Accept\_Pct is a good fit for Grad\_Rate.**
      2. **F\_Undergrad: Does not appear linear**
         1. **R: 0.0787** **R2: 0.0062** **Adj-R2: 0.0049** **F-value:4.84**
         2. **Supports lack of linearity**
         3. **F-value suggests F\_Undergrad is not a very significant variable regarding Grad\_Rate**
      3. **P\_Undergrad: Does not appear linear**
         1. **R: 0.2569** **R2: 0.0660** **Adj-R2:** **0.0648** **F-value: 54.81**
         2. **Suggests weak linearity**
         3. **F-value suggests P\_Undergrad is a good fit for Grad\_Rate.**
      4. **OutState: Appears linear**
         1. **R: 0.5713** **R2: 0.3264** **Adj-R2:** **0.3255** **F-value:375.49**
         2. **Suggests moderate linearity**
         3. **F-value suggests OutState is a good fit for Grad\_Rate.**
      5. **RoomBoard: Appears linear**
         1. **R: 0.4249** **R2: 0.1806** **Adj-R2: 0.1795** **F-value:170.79**
         2. **Suggests moderate linearity**
         3. **F-value suggests RoomBoard is a good fit for Grad\_Rate.**
      6. **Books: Does not appear linear**
         1. **R: 0** **R2: 0** **Adj-R2:** **-0.0013**  **F-value: 0**
         2. **Does not suggest linearity**
         3. **F-value suggests Books is not a very significant variable regarding Grad\_Rate**
      7. **Personal: Does not appear linear**
         1. **R: 0.2692** **R2: 0.0725** **Adj-R2:** **0.0713** **F-value: 60.62**
         2. **Suggests weak linearity**
         3. **F-value suggests Personal is a good fit for Grad\_Rate.**
      8. **PhD: Appears linear**
         1. **R: 0.3049** **R2: 0.0930** **Adj-R2: 0.0919** **F-value:79.51**
         2. **Suggests weak linearity**
         3. **F-value suggests PhD is a good fit for Grad\_Rate.**
      9. **Terminal: Appears linear**
         1. **R: 0.2894** **R2: 0.0838** **Adj-R2: 0.0826** **F-value: 70.91**
         2. **Suggests weak linearity**
         3. **F-value suggests Terminal is a good fit for Grad\_Rate.**
      10. **S\_F\_Ratio: Appears linear**
          1. **R: 0.3067** **R2: 0.0941** **Adj-R2: 0.0929** **F-value: 80.49**
          2. **Suggests weak linearity**
          3. **F-value suggests S\_F\_Ratio is a good fit for Grad\_Rate.**
      11. **Perc\_alumni: Appears linear**
          1. **R: 0.4909** **R2: 0.2410** **Adj-R2:** **0.2400**  **F-value:246.05**
          2. **Suggests moderate linearity**
          3. **F-value suggests perc\_alumni is a good fit for Grad\_Rate.**
      12. **Expend: Does not appear linear**
          1. **R: 0.3903** **R2: 0.1524** **Adj-R2:** **0.1513**  **F-value:139.31**
          2. **Suggests weak-moderate linearity**
          3. **F-value suggests Expend is a good fit for Grad\_Rate.**
2. 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).
   1. **Boxplot: private vs public (private = 1)**

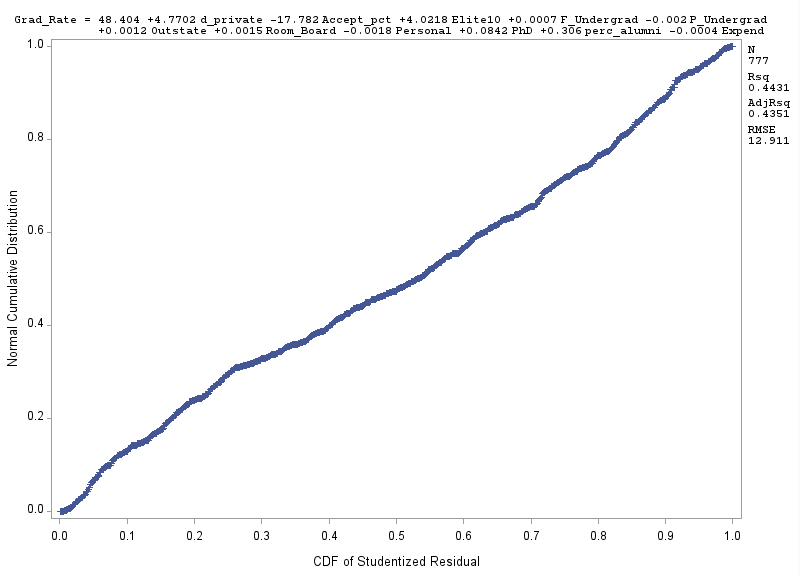


* + 1. **Discussion:** In comparing the graduation rates of individuals in public schools to those of private schools, we learn that the interquartile range of private schools depicts higher graduation rates than those of public schools. However there do exist data entry errors and outliers that could be manipulating the data to some degree. Even so, from the boxplots we can note that the mean graduation rate for private schools is approximately 70% while the mean graduation rate for public schools is approximately 55%. In addition, private schools have an overall larger quartile range, ranging from approximately 22% to 100% graduation rate versus public school's quartile range from approximately 22% to 95%.
  1. **Boxplot: elite vs not elite (elite = 1)**



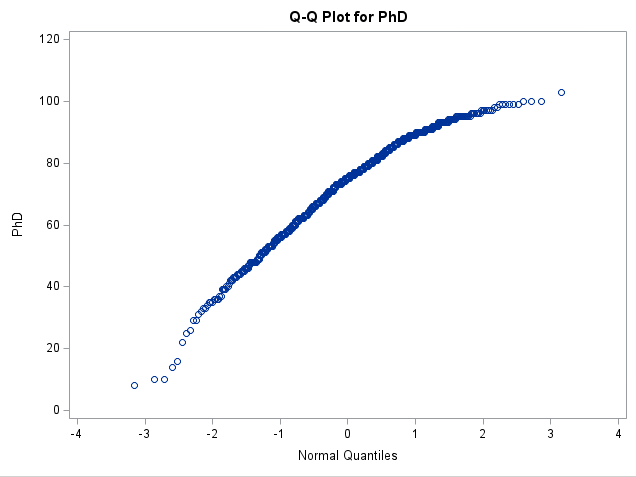
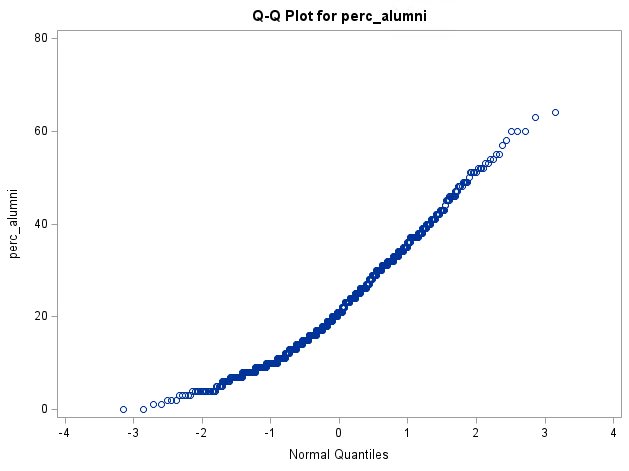
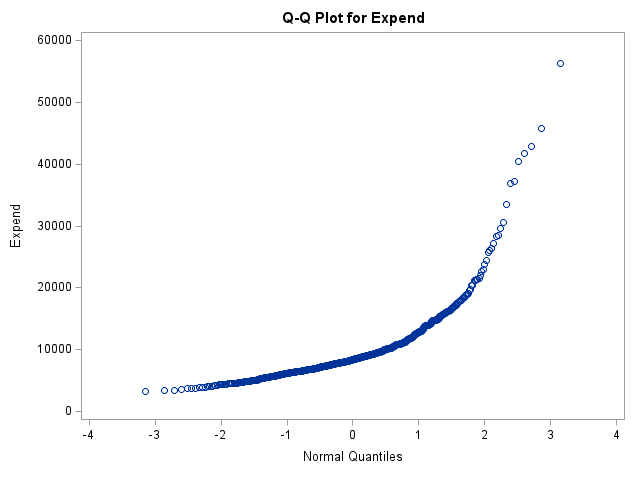
* + 1. **Discussion:** In comparing the elite schools to non-elite schools, the mean graduation rate from elite schools is larger at approximately 90% compared to non-elite schools mean at approximately 65%. In addition, the range for elite schools is smaller, ranging from 55% to 100% graduation rate compared to non-elite schools, ranging from approximately 18% to 100% (not including outliers). Overall, the boxplots suggest that elite schools have a higher, more concentrated graduation rate, while non-elite schools having larger range of graduation rates with a smaller mean graduation rate.

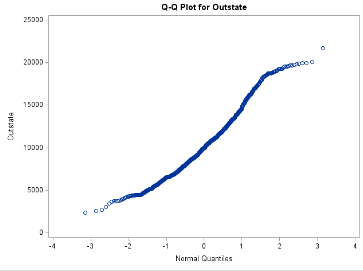
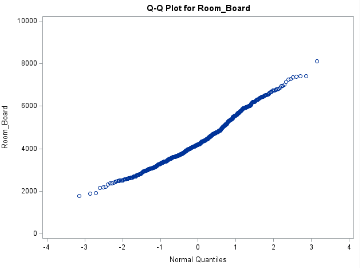
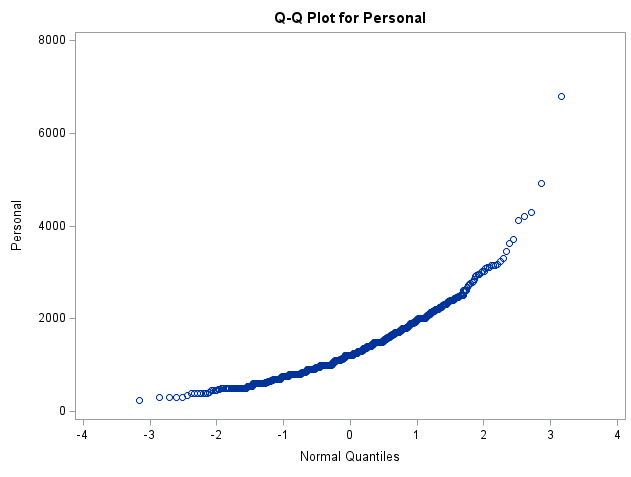
1. d) Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output.
   1. Refer to **Linearity: Description per X-variables** in Question 2 to get detailed parameter estimates, significance, goodness-of-fit and AdjR2 values for all x-variables.
   2. Full model:
      1. Grad\_Rate = 4.61959**d\_private** -18.10932**Accept\_pct +** 4.01748**Elite10 +** 0.00068095**F\_Undergrad** –0.00196**P\_Undergrad +** 0.00123**OutState +** 0.00167**Room\_Board** -0.00252**Books** -0.00172 **Personal** + 0.13064**PhD +** -0.07284**Terminal +** 0.00100**S\_F\_Ratio +** 0.30920**perc\_alumni +** -0.00043651**Expend**
   3. 
   4. Full model significance, goodness-of-fit, AdjR2 :
      1. **R: 0.6669** **R2: 0.4448** **Adj-R2:** **0.4346**  **F-value:43.61**
   5. **Discussion:** In review of the full model, we can note that the model has moderate linearity, but weak goodness-of-fit measurements. The goodness-of-fit measurements signify that only 43.46% of variance is explained by the full model with a moderate F-value of 43.61 that passes the p-value test at <0.0001. It is recommended to fit the full-model to procure a more significant model.
2. e) Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF (Variance Inflation Factor) statistics. Include the relevant output and discuss your answer.
   1. Refer to **Flagging for Multicollinearity** above in Question 1. As written in the discussion, there were no x-variables that showed signs of multicollinearity. Therefore, multicollinearity is not a problem.
3. 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-R2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, including the outputs.
   1. **Variable Selection Method: Backward**
      1. **Fitted predictors:**
         1. **Accept\_pct, d\_private, Elite10, F-Undergrad, P\_Undergrad, OutState, Room\_Board, Personal, PhD, perc\_alumni, Expend**
         2. **R: 0.6656**  **R2: 0.4431**  **AdjR2: 0.4351** **F-value:55.33**
         3. **Error:** **166.69412** **Cp: 11.3674**
      2. **Fitted Predictors with Transformed x-variables:**
         1. **Sq\_Accept\_pct, d\_private, Elite10, log\_F\_Undergrad, sqrt\_OutState, log\_Room\_Board, log\_Personal, sqrt\_perc\_alumni, inv\_Expend**
         2. **R:0.7293 R2:0.5319 AdjR2:0.5261 F-value:91.30**
         3. **Error:117.65529 Cp:10.3690**
   2. **Variable Selection Method: cp**
      1. **Fitted predictors:**
         1. **Accept\_pct, d\_private, Elite10, F\_Undergrad, P\_Undergrad, Outstate, Room\_Board, Personal, PhD, perc\_alumni, Expend**
         2. **R: 0.6656**  **R2: 0.4431**  **AdjR2: 0.4351** **F-value:55.33**
         3. **Error:** **166.69412** **Cp: 11.3674**
      2. **Fitted Predictors with Transformed x-variables:**
         1. **Sq\_Accept\_pct, d\_private, Elite10, log\_F\_Undergrad, sqrt\_OutState, log\_Room\_Board, log\_Personal, sqrt\_perc\_alumni, inv\_Expend**
         2. **R:0.7293 R2:0.5319 AdjR2:0.5261 F-value:91.30**
         3. **Error:117.65529 Cp:10.3690**
4. 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**.
   1. Final model M1: Grad\_Rate = -25.83333-sq(10.25946) **Sq\_Accept\_pct t +** 7.26646**d\_private +** 3.64145**Elite10 + log(**3.99909) **log\_F\_Undergrad + sqrt(**0.24940) **sqrt\_OutState + log(**4.42951) **log\_Room\_Board** –log(3.12611) **log\_Personal + sqrt(**3.78752) **sqrt\_perc\_alumni +inv(**31296**) inv\_Expend**
   2. **Discussion:** This M1 model is the optimal model since the variable selection procedure methods, backwards and cp, offered these variables as the variables that contribute to the best fit. The backward method selects variables based on a variable’s statistical significance, removing the variable if it does not fit the model appropriately (p-value/AdjR2). The cp method selects variables based on multiple models to assess which model contains the best fit in consideration to cp value and AdjR2.
5. 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.
   1. **Plot:**

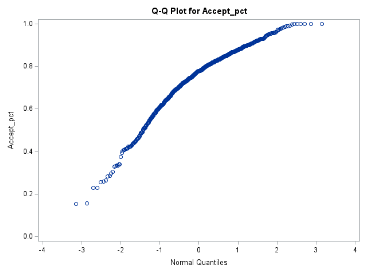
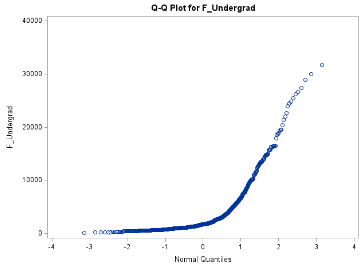
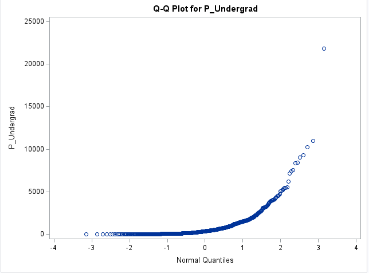


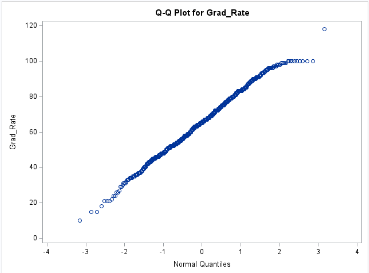
**Discussion:** The plot does not indicate any striking problems. There does exist some curvature to the plot that isn’t indicative of linearity, however, it doesn’t appear too problematic.

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









**Discussion:** By utilizing the normal plot of the residual for our variables, we can notice that normality is not satisfied. If normality was satisfied all the above plots would be linear. However, plots concerning x-variables Expand, perc\_alumni, PhD, Personal, OutState, P\_Undergrad, F\_Undergrad and Accept\_pct do not appear linear. It is recommended to transform the model.

**Transforming the model:**

**Plots:**

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated**

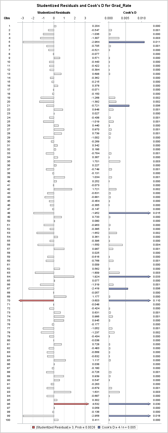
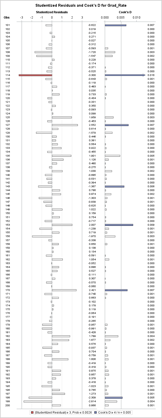
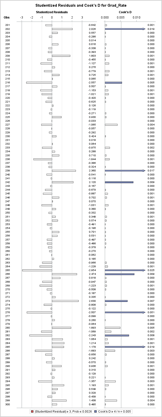
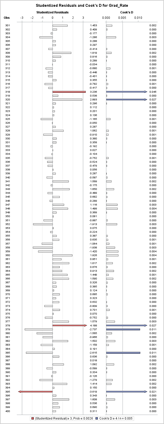
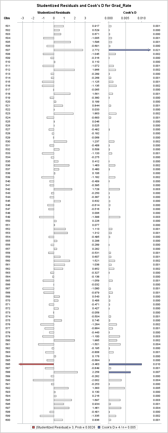
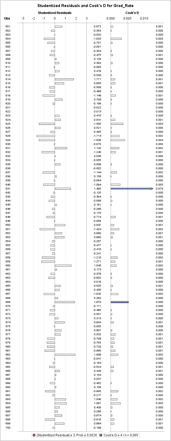
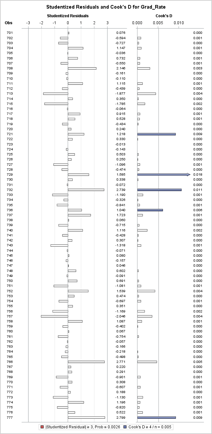
**Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated**

**Discussion:** The variables appear linear after the transformations of the x-variables, such that the x-variables are now sq\_Accept\_pct, log\_F\_Undergrad, sqrt\_OutState, log\_Room\_Board, log\_Personal , sqrt\_perc\_alumni, inv\_Expend (excluding d\_private and Elite10).

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



**Discussion:** All red values are outliers, and all blue values are influential points. These observations are

**outliers:** 70, 96, 99, 114, 318, 378, 395, 586

**Influential points:** 5, 21, 48, 67, 70, 96, 99, 114, 127, 143, 153, 170, 198, 202, 239, 242, 265, 266, 275, 280, 282, 285, 318, 320, 378, 379, 385, 395, 419, 427, 446, 452, 498, 507, 586, 588, 592, 641, 669, 721, 729, 732, 752, 777

To improve our grad\_rate efficiency I will remove the outliers and influential points.

1. k) Analyze the AdjR2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.
   1. Final model with transformation: **AdjR2: 0.5261**
   2. **Discussion:** At 0.5264 or 52.64%, the AdjR2 identifies the percentage of variance is explained by the model.
2. 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.
   1. **Fitted model without outliers and influential points and transformation:**
      1. Grad\_Rate = -25.83333-sq(10.25946) **Sq\_Accept\_pct t +** 7.26646**d\_private +** 3.64145**Elite10 + log(**3.99909) **log\_F\_Undergrad + sqrt(**0.24940) **sqrt\_OutState + log(**4.42951) **log\_Room\_Board** –log(3.12611) **log\_Personal + sqrt(**3.78752) **sqrt\_perc\_alumni +inv(**31296**) inv\_Expend**
   2. **Most Important Predictors:**
      1. **Table

         Description automatically generated**
      2. **Predictors ordered by most important to least important**
         1. **Sqrt\_perc\_alumni**
         2. **Sqrt\_OutState**
         3. **Log\_F\_Undergrad**
         4. **D\_private**
         5. **Sq\_Accept\_Pct**
         6. **Inv\_Expend**
         7. **Log\_Personal**
         8. **Log\_Room\_Board**
         9. **Elite10**
   3. **Difference between private vs public university graduation rates:**
      1. **Note:** Considering the variable d\_private singularly we know that there would be a graduation rate difference of 7.2664%, signifying that if the university is private there would be an increase in our predictive graduation rate by nearly 8% which is a considerably significant difference.
      2. **Graduation rate for private schools:** Graduation rate increases by 7.2664%.
      3. **Graduation rate for public schools:** Graduation rate would be less than private schools by 7.2664%.
   4. **Difference between elite vs non-elite university graduation rates:**
      1. **Note:** Considering the variable Elite10 singularly we know that there would be a graduation rate difference of 3.64145%, signifying that if the university is made of the top 10% students in high school or Elite10, there would be a 3.64145% in graduation rate in comparison to non-elite schools. This can be considered a moderately significant difference.
      2. **Graduation rate for elite schools:** Graduation rate increases by 3.64145%
      3. **Graduation rate for non-elite schools:** Graduation rate would be less than elite schools by 3.64145%
3. m) Use the final regression model to predict the graduation rate for the following values. Using SAS, compute the predicted graduation rate, 95% confidence interval and prediction interval for your estimate. Make sure to use your findings.

Table

Description automatically generated

**Predicted Graduation Rate: 64.32%**

Grad\_Rate = -25.83333-sq(10.25946) **(0.87) +** 7.26646**(1)+** 3.64145**(0) + log(**3.99909)**(3000) + sqrt(**0.24940) **(6500)+ log(**4.42951) **(3300)** –log(3.12611) **(1350)+ sqrt(**3.78752)**(13) +inv(**31296**) (5201)**

**95% CI:**

**Table

Description automatically generated with medium confidence**

**95% PI:**

Table

Description automatically generated

1. Copy and paste your FULL SAS code into the word document along with your answers.
2. **proc** **import** file = "C:\Users\DVALDE12\Desktop\A5\College.csv"
3. out = college
4. dbms = csv
5. replace;
6. **run**;
7. **proc** **print** data = college;
8. **run**;
9. \*creating dummy variable for x-var Private;
10. **data** college;
11. set college;
12. d\_private = (Private = 'Yes');
13. sq\_Accept\_pct = (Accept\_pct)\*\***2**;
14. log\_F\_Undergrad = log(F\_Undergrad);
15. log\_P\_Undergrad = log(P\_Undergrad);
16. sqrt\_OutState = sqrt(OutState);
17. log\_Room\_Board = log(Room\_Board);
18. log\_Personal = log(Personal);
19. sq\_PhD = PhD\*\***2**;
20. sqrt\_perc\_alumni = sqrt(perc\_alumni);
21. inv\_Expend = **1**/Expend;
22. **run**;
23. \*boxplot;
24. **proc** **sgplot** data = college;
25. vbox Grad\_Rate;
26. **run**;
27. \*histogram;
28. **proc** **univariate** data = college;
29. title "Histogram: Grad\_Rate";
30. var Grad\_Rate;
31. Histogram;
32. **run**;
33. \*univariate proc;
34. **proc** **univariate** data = college;
35. title "Univariate Grad\_Rate";
36. var Grad\_Rate;
37. **run**;
38. \*transformation;
39. \*flagging for multicollinearity;
40. **proc** **sgscatter**;
41. title "Matrix: Grad\_Rate & x-var (d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad)";
42. matrix Grad\_Rate d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad;
43. **run**;
44. **proc** **sgscatter**;
45. title "Matrix: Grad\_Rate & x-var ( OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend)";
46. matrix Grad\_Rate OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend;
47. **run**;
48. \*vif & tol;
49. **proc** **reg** data = college;
50. title "VIF & TOl Between Grad\_Rate and All X-var";
51. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend / vif tol;
52. **run**;
53. \*normal plot;
54. **proc** **univariate** data = college;
55. var Grad\_Rate;
56. qqplot;
57. **run**;
58. \*finding outliers and influential pts;
59. **proc** **reg** data =college;
60. title "outliers & influential obs Grad\_Rate";
61. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend / influence r;
62. plot student.\*(d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend predicted.);
63. plot npp.\*student.;
64. **run**;
65. \*individually checking x-var r/adj-r2/f-value;
66. \*proc reg data = college;
67. \*model Grad\_Rate = Accept\_Pct;
68. \*run;
69. \*proc reg data = college;
70. \*model Grad\_Rate = F\_Undergrad;
71. \*run;
72. \*proc reg data = college;
73. \*model Grad\_Rate = P\_Undergrad;
74. \*run;
75. \*proc reg data = college;
76. \*model Grad\_Rate = OutState;
77. \*run;
78. \*proc reg data = college;
79. \*model Grad\_Rate = Room\_Board;
80. \*run;
81. \*proc reg data = college;
82. \*model Grad\_Rate = Books;
83. \*run;
84. \*proc reg data = college;
85. \*model Grad\_Rate = Personal;
86. \*run;
87. \*proc reg data = college;
88. \*model Grad\_Rate = PhD;
89. \*run;
90. \*proc reg data = college;
91. \*model Grad\_Rate = Terminal;
92. \*run;
93. \*proc reg data = college;
94. \*model Grad\_Rate = S\_F\_Ratio;
95. \*run;
96. \*proc reg data = college;
97. \*model Grad\_Rate = perc\_alumni;
98. \*run;
99. **proc** **reg** data = college;
100. model Grad\_Rate = Expend;
101. **run**;
102. \*boxplot private/public;
103. **proc** **sgplot** data = college;
104. title "Private Vs Public";
105. vbox Grad\_Rate / category = d\_private;
106. **run**;
107. **proc** **sgplot** data = college;
108. title "Elite Vs Not Elite";
109. vbox Grad\_Rate / category = Elite10;
110. **run**;
111. \*Backward;
112. **proc** **reg** data = college;
113. title "Variable Selection: Backward";
114. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend / selection=backward;
115. **run**;
116. \*fitted-model Backward;
117. **proc** **reg** data = college;
118. title "Variable Selection: Fitted Model Backward";
119. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend;
120. **run**;
121. \*Cp not fitted;
122. **proc** **reg** data = college;
123. title "Variable Selection: cp";
124. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend / selection=cp;
125. **run**;
126. \*Plotting studentized;
127. **proc** **reg** data =college;
128. title "M1 Studentized";
129. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend;
130. plot student.\*(d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend predicted.);
131. plot npp.\*student.;
132. **run**;
133. \*Plotting predicted;
134. **proc** **reg** data = college;
135. title "M1 Predicted";
136. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend;
137. plot student.\*(d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend);
138. plot student.\*predicted.;
139. plot npp.\*student.;
140. **run**;
141. \*plotting normal;
142. **proc** **univariate** data = college;
143. title 'normal';
144. var Grad\_Rate d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend;
145. qqplot;
146. **run**;
147. \*finding outliers and influential pts;
148. **proc** **reg** data =college;
149. title "outliers & influential";
150. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend / influence r;
151. plot student.\*(d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend predicted.);
152. plot npp.\*student.;
153. **run**;
154. \*removing outliers and influential pts;
155. **data** college\_new;
156. set college;
157. if \_n\_ in (**70**, **96**, **99**, **114**, **318**, **378**, **395**, **586**, **5**, **21**, **48**, **67**, **127**, **143**, **153**, **170**, **198**, **202**, **239**, **242**, **265**, **266**, **273**, **280**, **282**, **285**, **320**, **379**, **385**, **419**, **427**, **446**, **452**, **498**, **507**, **588**, **592**, **641**, **669**, **721**, **729**, **732**, **752**, **777**) then delete;
158. **run**;
159. **proc** **reg** data = college\_new;
160. title 'new\_model w/o outliers';
161. model Grad\_Rate = d\_private Accept\_pct Elite10 F\_Undergrad P\_Undergrad OutState Room\_Board Personal PhD perc\_alumni Expend;
162. **run**;
163. \*transforming:all x-var using sqrt;
164. **proc** **univariate** data = college\_new;
165. title 'normal: transform of x-vars';
166. var Grad\_Rate d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad log\_P\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sq\_PhD sqrt\_perc\_alumni inv\_Expend;
167. qqplot;
168. **run**;
169. \*new model with transforms;
170. **proc** **reg** data = college\_new;
171. title 'new\_model w/o outliers';
172. model Grad\_Rate = d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad log\_P\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sq\_PhD sqrt\_perc\_alumni inv\_Expend;
173. **run**;
174. \*backward selection on transformed model;
175. **proc** **reg** data = college\_new;
176. title "Variable Selection: Backward + Transformation";
177. model Grad\_Rate = d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad log\_P\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sq\_PhD sqrt\_perc\_alumni inv\_Expend/ selection= backward;
178. **run**;
179. \*fitted transform;
180. **proc** **reg** data = college\_new;
181. title "Variable Selection: fit";
182. model Grad\_Rate = d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sqrt\_perc\_alumni inv\_Expend;
183. **run**;
184. \*cp fitted with transformed x-var;
185. **proc** **reg** data = college\_new;
186. title "Variable Selection: Cp + Transformation";
187. model Grad\_Rate = d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad log\_P\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sq\_PhD sqrt\_perc\_alumni inv\_Expend/ selection= cp;
188. **run**;
189. \*most important x-var;
190. **proc** **reg** data = college\_new;
191. title "important x-var";
192. model Grad\_Rate = d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sqrt\_perc\_alumni inv\_Expend/ stb;
193. **run**;
194. \*CI;
195. title 'CI';
196. **proc** **means** data = college\_new alpha = **0.05** mean std stderr clm min p25 p50 p75 max ;
197. var Grad\_Rate d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sqrt\_perc\_alumni inv\_Expend;
198. **run**;
199. \*PI;
200. **proc** **reg** data = college\_new;
201. title 'PI';
202. model Grad\_Rate = d\_private sq\_Accept\_pct Elite10 log\_F\_Undergrad sqrt\_OutState log\_Room\_Board log\_Personal sqrt\_perc\_alumni inv\_Expend/cli clm;
203. **run**;