**Krissy Groom**

**CSC423**

**8/5/18**

**HWK 3**

**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Summer Assignment-3** | **Total Points: 25 pts for CSC 423/25 pts for CSC 324**

**Due Date: 08/07/2018 by 11:59 pm**

***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.***

**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 2, before you do this last part.

***Answer****: The feedback I received from Homework 2 was that I should have removed Income vs. Wealth so I went back and redid my model using Income. I checked for multicollinearity, outliers and influential points, checked model assumptions and goodness of fit. I will add the output for updated final model and will continue to answer the questions for this part of Assignment 3 moving forward with this model:*

***Code:***

*\*Final model with Income removed and outliers and influential points removed*

*TITLE 'M2 Regression Analysis with removed points';*

*PROC REG data=bankingfull\_new5;*

*MODEL Balance=Age Education HomeVal Wealth /influence r stb vif;*

*\*residual plots;*

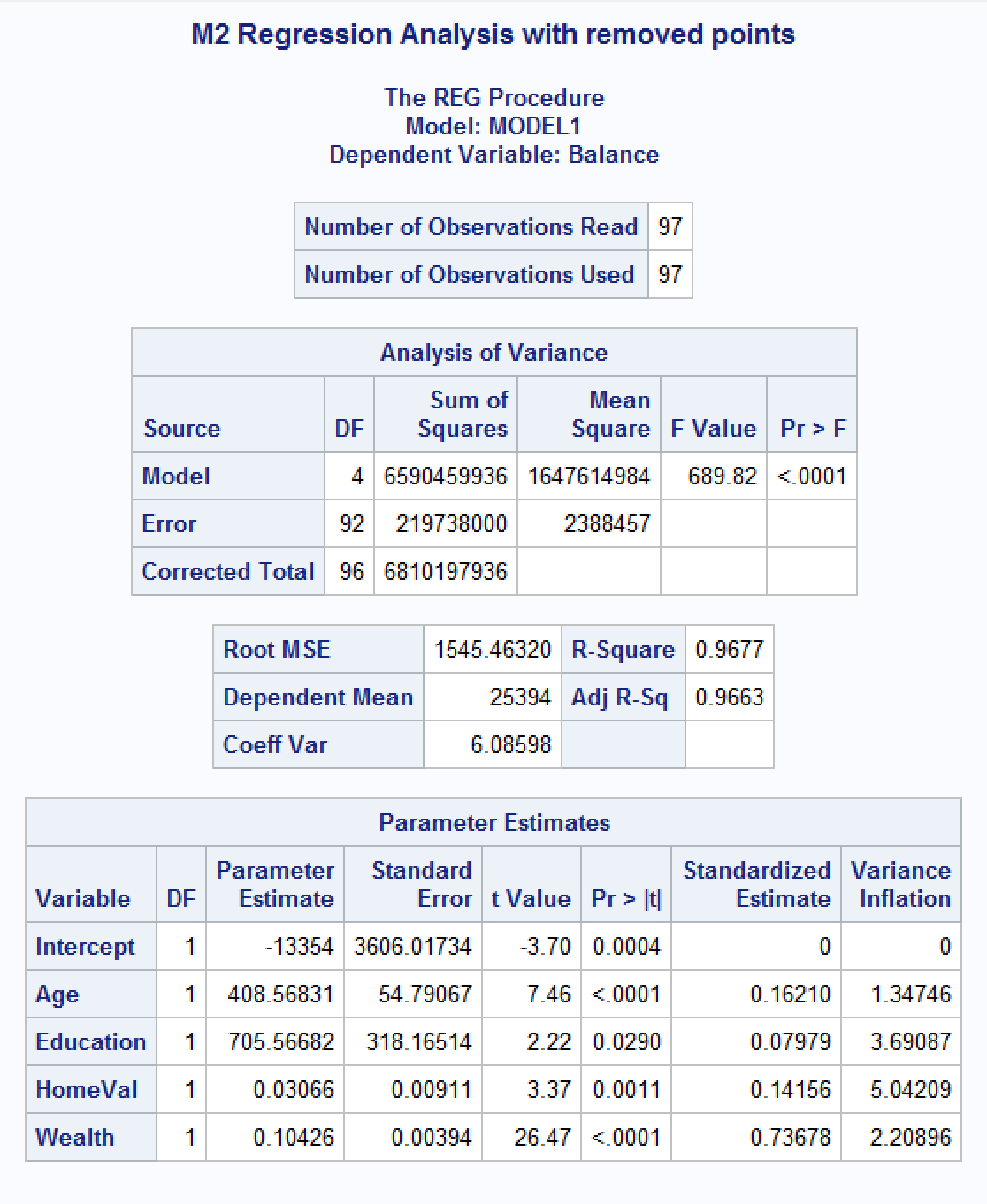
*PLOT student.\*predicted.;*

*PLOT student.\*(Age Education HomeVal Wealth);*

*PLOT npp.\*student.;*

*run;*

***Output:***

******

1. Use the fitted regression model from Assignment 2 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.

***Answer****: The final fitted model equation:*

*Average bank balance = -13354 + 408.56831(Age) + 705.56682(Education) + 0.03066(HomeVal) + 0.10426(Wealth)*

***Code****:*

*\*\*compute predictions on new value;*

*title "Compute Predictions";*

*\*\*creates dataset with new value;*

*data pred;*

*input Age Education Income HomeVal Wealth;*

*datalines;*

*34 13 89000 160000 140000*

*;*

*run;*

*proc print;*

*run;*

*\*join new dataset with final bankingfull dataset;*

*data prediction;*

*set pred bankingfull\_new5;*

*run;*

*proc print;*

*run;*

*\*compute regression analysis and confidence interval for average estimate;*

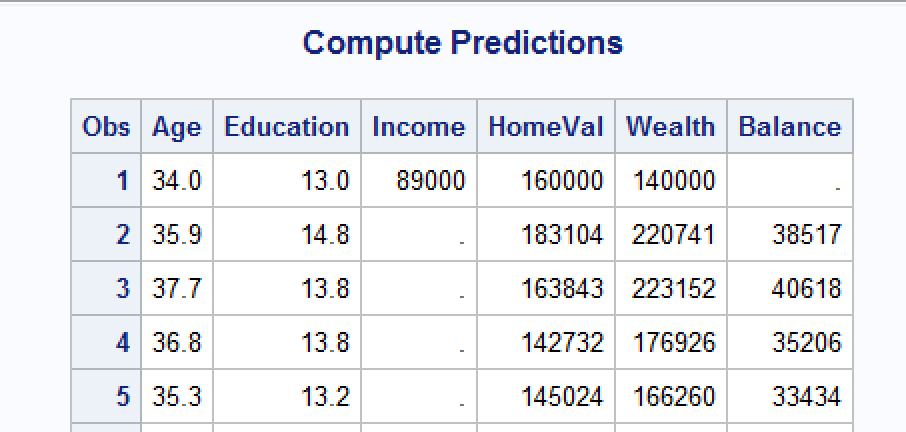
*proc reg;*

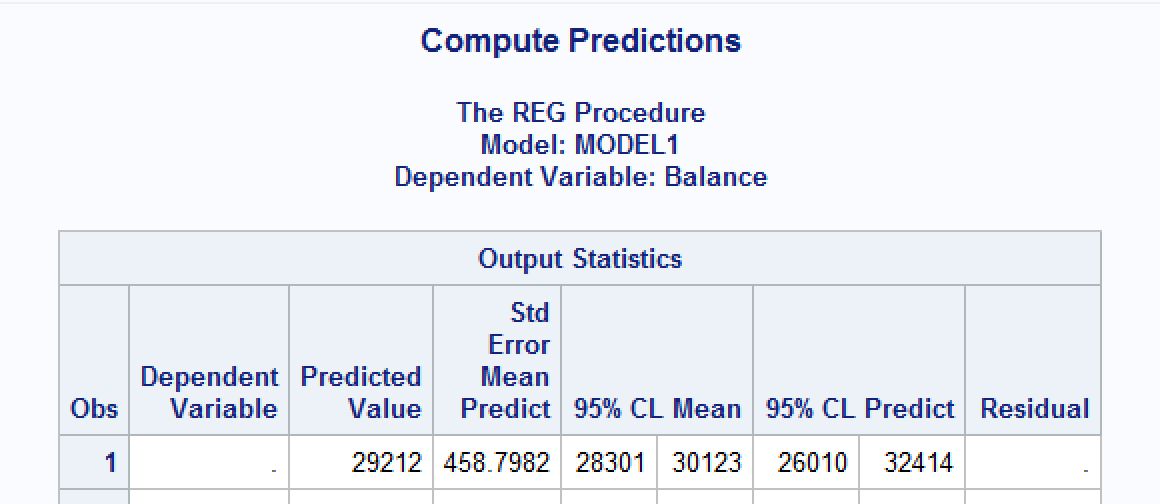
*model Balance=Age Education HomeVal Wealth/ p clm cli;*

*run;*

***Output:***







***Findings:***

*The predicted value, or that on average the bank can expect for the average bank balance in the specific zip code to be $28,212 with 95% CI ($28,301, $30,123) and 95% PI ($26,010, $32,414).*

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

*See above Answers*

**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: Depending on how you import you data (INFILE or IMPORT) the SAS may relabel the column names. Make sure to use the variable names that appear when you use a proc print.**

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

Answer the following questions.

***Answer:*** *Getting data into dataset with dummy vars:*

*\*\*Problem 2;*

*\*\*a);*

*\* Get college data into dataset;*

*TITLE 'College data with dummy vars';*

*data grad;*

*INFILE "S:\CSC423\HW3\College.csv" delimiter = ',' MISSOVER firstobs=2;*

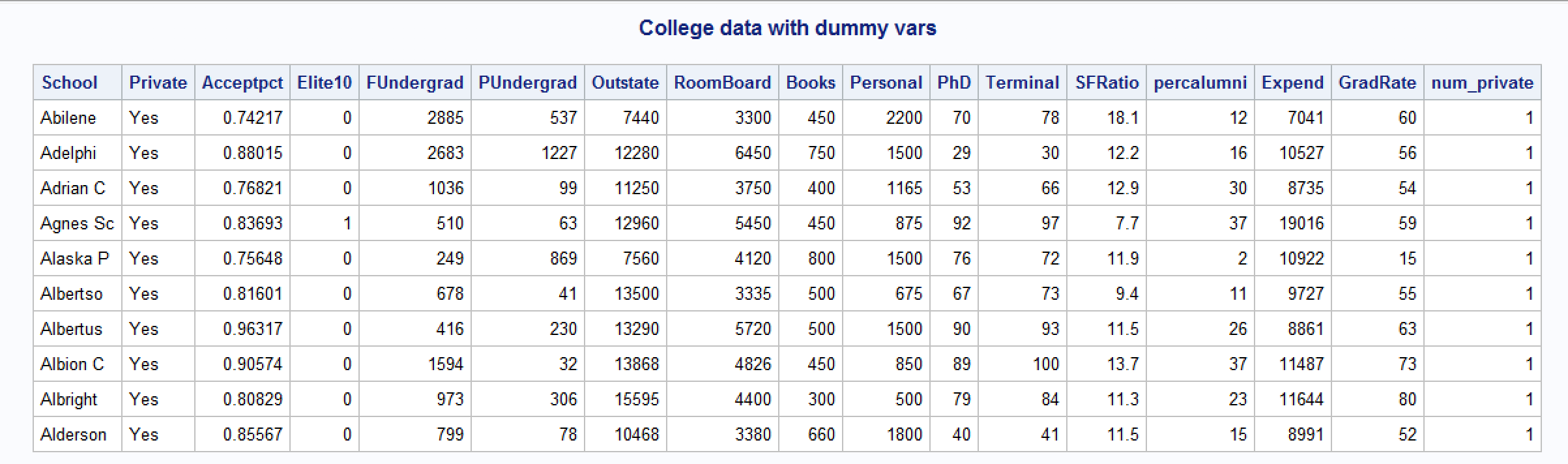
*INPUT School $ Private $ Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio percalumni Expend GradRate;*

*num\_private=(private="Yes");*

*run;*

*proc print data=grad(obs=10) noobs;*

*run;*

**

1. 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).

***Answer:***

***Code:***

*\*Histogram and normal curve for GradRate;*

*TITLE 'Histogram + normal curve for GradRate';*

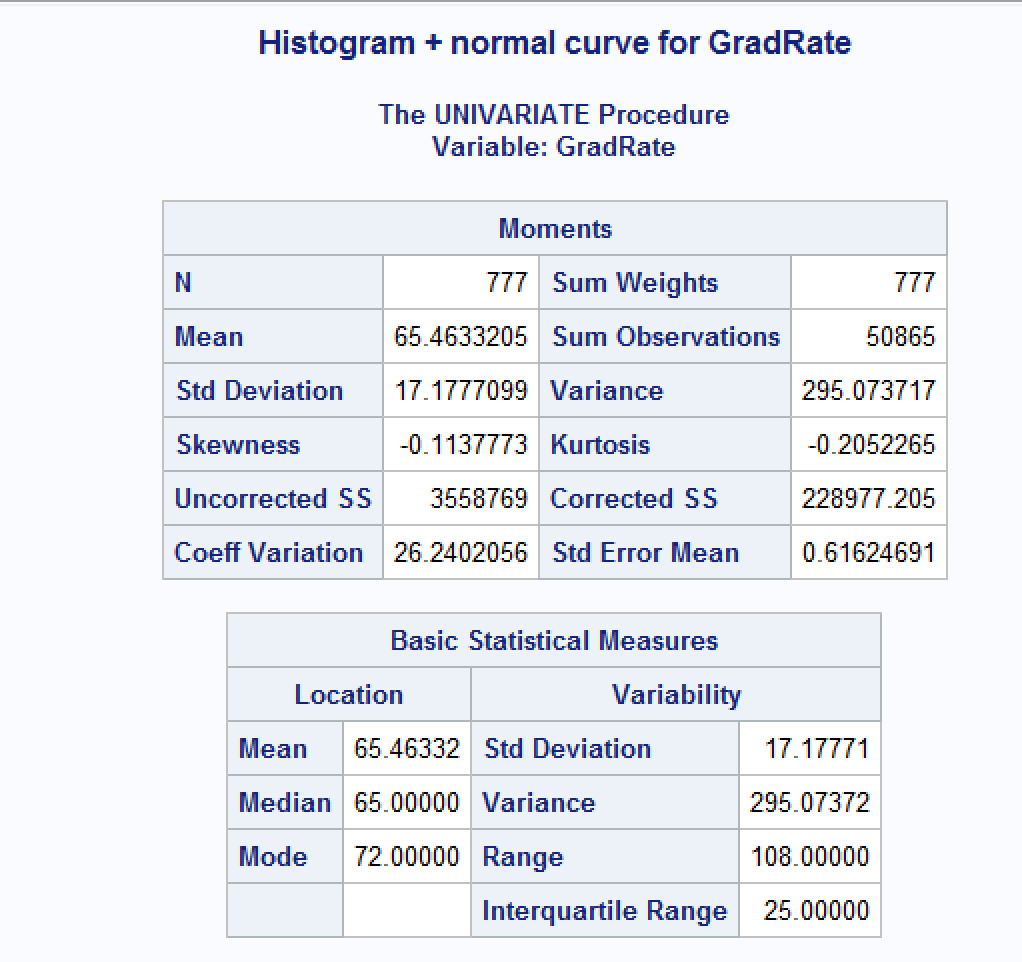
*proc univariate normal;*

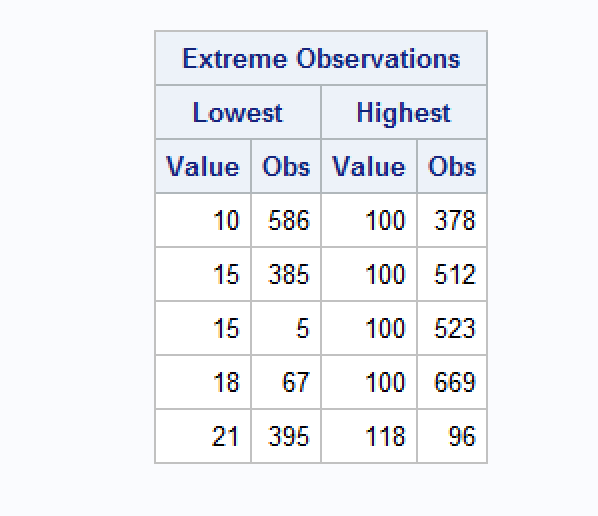
*var GradRate;*

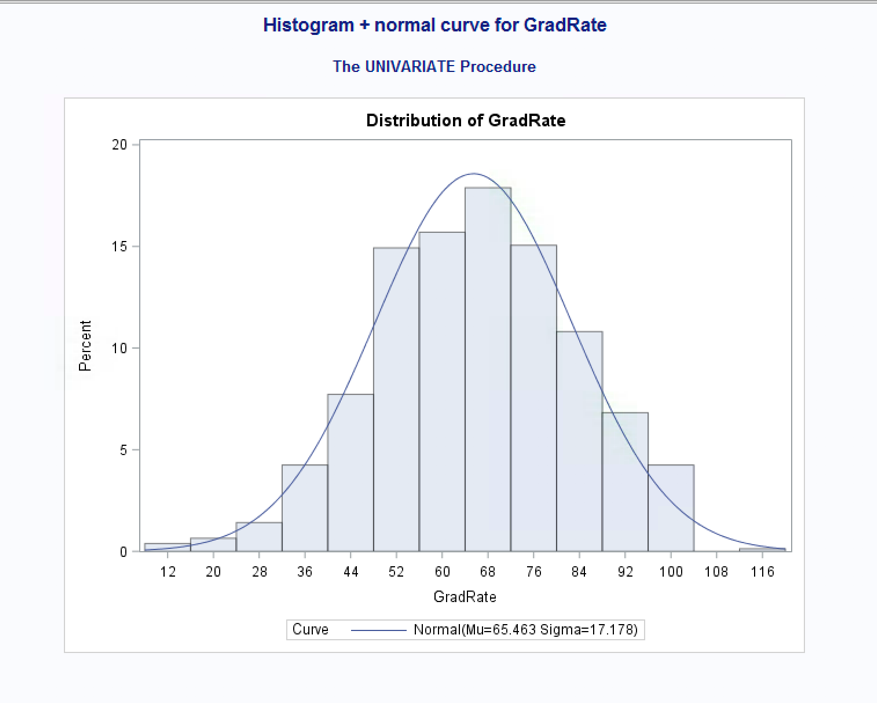
*histogram / normal (mu = est sigma = est);*

*run;*

***Output:***

******

******

******

***Findings:***

*I can see that the dependent variable of GradRate – graduation rate in 4 years, is pretty normally distributed, which I see from the histogram and the normal probability curve output. This is confirmed for me when I look at the mean and median which are almost exactly the same. In looking at the Extreme Observations output, I see that there is likely one outlier, point 96 with a value of 118, which doesn’t make sense as a grad rate so I’m guessing this could be a data input error - I will explore this further throughout my analysis.*

1. 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).

***Answer:***

***Code:***

*\*Gplot for GradRate and x vars;*

*TITLE 'Gplot GradRate\*(x vars)';*

*PROC GPLOT;*

*PLOT GradRate\*(Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio percalumni Expend);*

*RUN;*

*\*scatterplot of gradrate vs percalumni by num\_private;*

*symbol1 value=dot color=red;*

*symbol2 value=plus color=blue;*

*proc gplot data=grad;*

*plot GradRate\*percalumni=num\_private;*

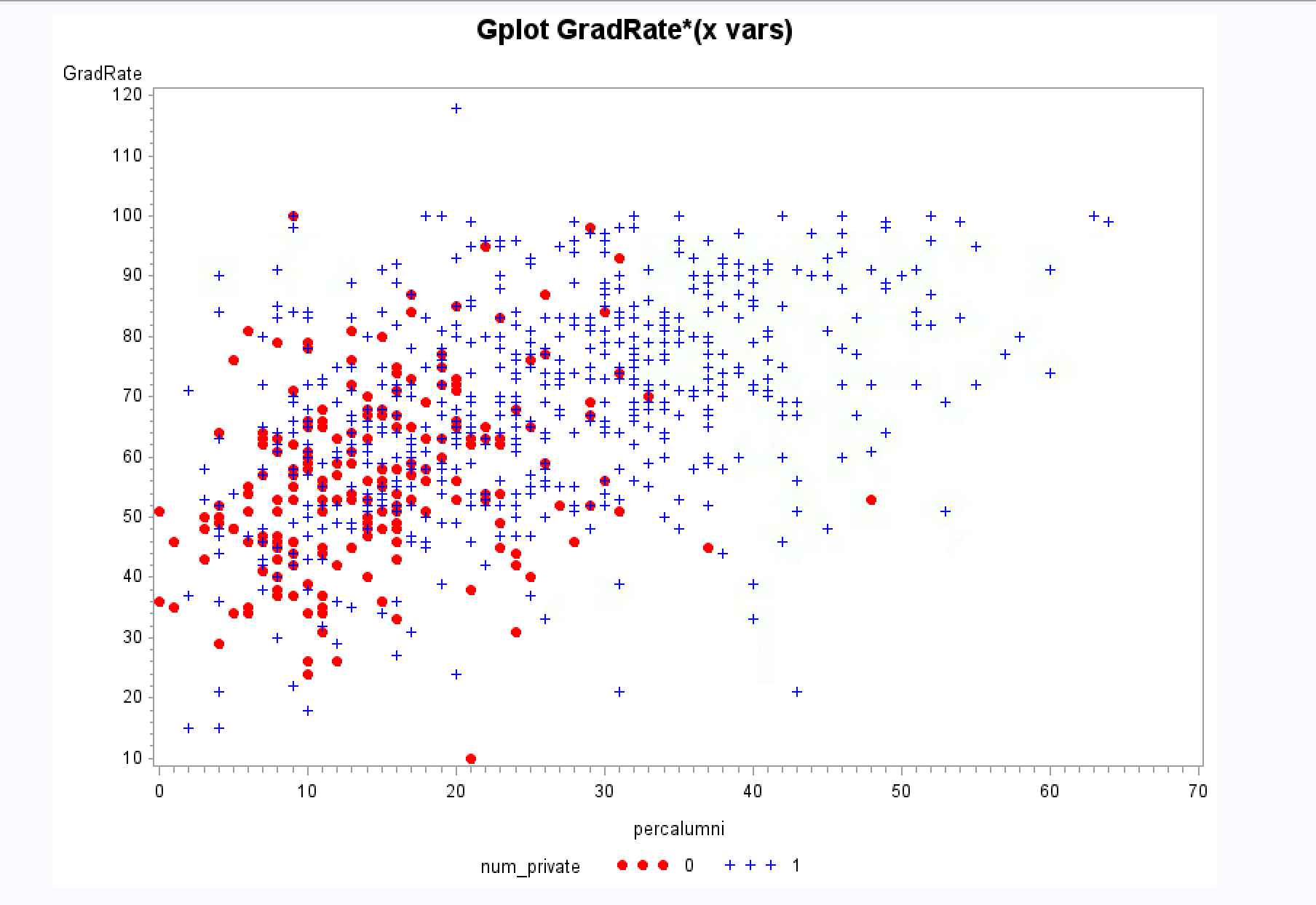
*plot GradRate\*percalumni=Elite10;*

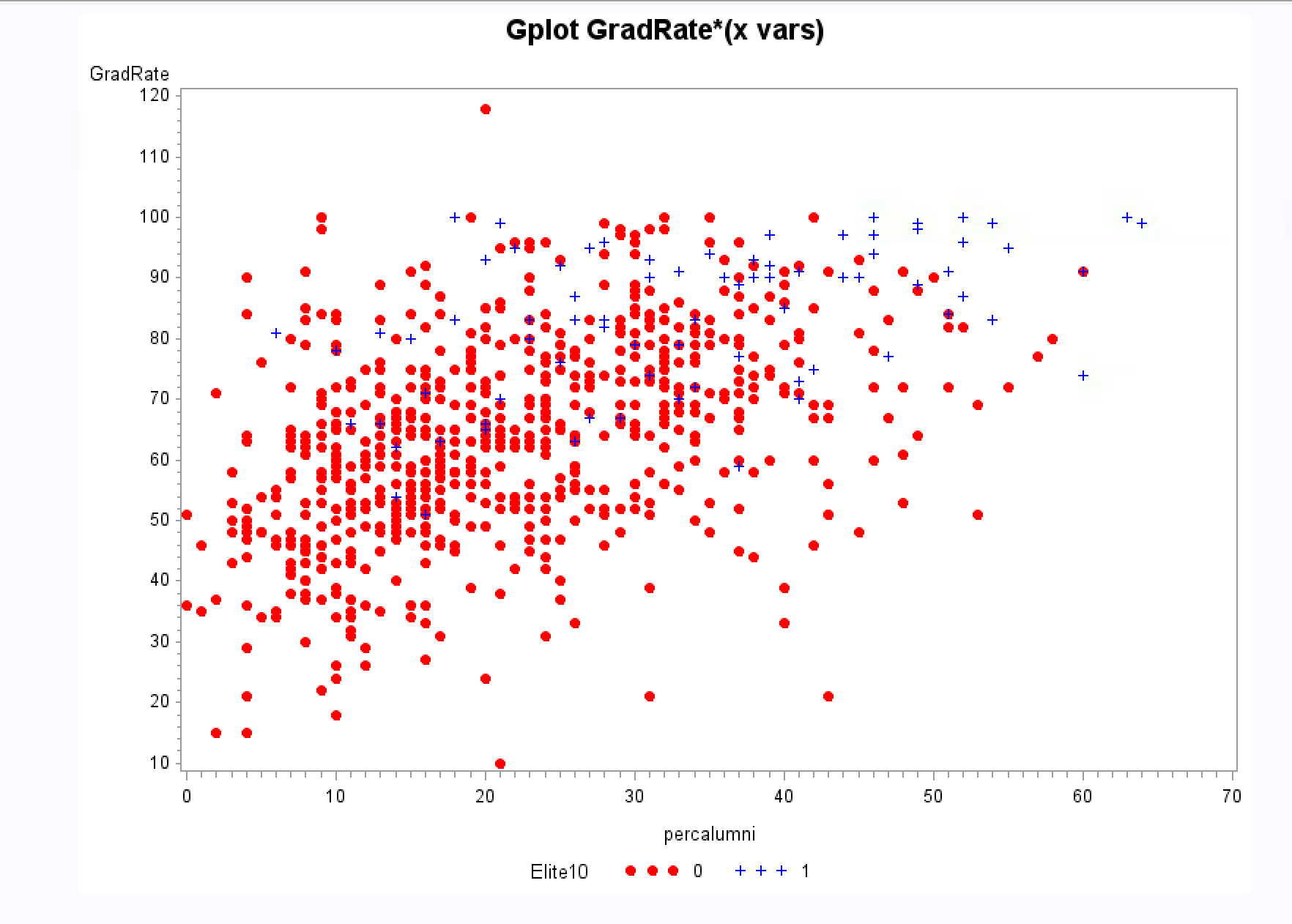
*run;*

***Output:***

*(See above – no need to include scatterplots)*

*Adding the plots for the categorical variables:*

**

**

***Findings:***

*I noticed the following observations between GradRate and the independent variables:*

*GradRate and:*

* *Acceptpct – shows a linear relationship*
* *Elite10 – does not tell me anything about the relationship b/c is dummy var*
* *FUndergrad – appears to be linear relationship with possible outliers or influential points*
* *PUndergrad – I think that this shows a linear relationship, although the data distribution for part-time undergrads looks very skewed right*
* *Outstate – shows a linear relationship*
* *Roomboard – shows a linear relationship with possible outlier*
* *Books – another plot that I think is linear although skewed*
* *Personal – shows a linear relationship*
* *Phd – shows linear relationship*
* *Terminal – shows linear relationship*
* *SFRatio – shows linear relationship*
* *Percalumni – shows linear relationship*
* *Expend – shows possible linear relationship that is being pulled by an outlier and or influential point*
* *Num\_private – categorical variable so does not tell me anything from this plot*

*I looked at GradRate vs percalumni with regards to num\_private, and Elite10. I can see in the plot with num\_private variable that it appears that there are a higher number of alumni who donate along with a higher 4 year graduation rate for private schools vs. non-private schools. The plot for Elite10 looks like there are more schools considered non-elite than there are that are considered elite but of those that are elite, it appears that in general there is a higher number of alumni who donate.*

*I can see that all the plots are being influenced by at least one outlier – the point that shows almost 120% for grad rate. I will explore that when I look at outliers and influential points.*

1. 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).

***Answer:***

***Code:***

*\* boxplot of gradrate by num\_private;*

*\* need to sort data by num\_private first;*

*proc sort data=grad;*

*by num\_private;*

*proc boxplot data=grad;*

*plot GradRate\*num\_private;*

*run;*

*\* boxplot of gradrate by num\_private;*

*\* need to sort data by num\_private first;*

*proc sort data=grad;*

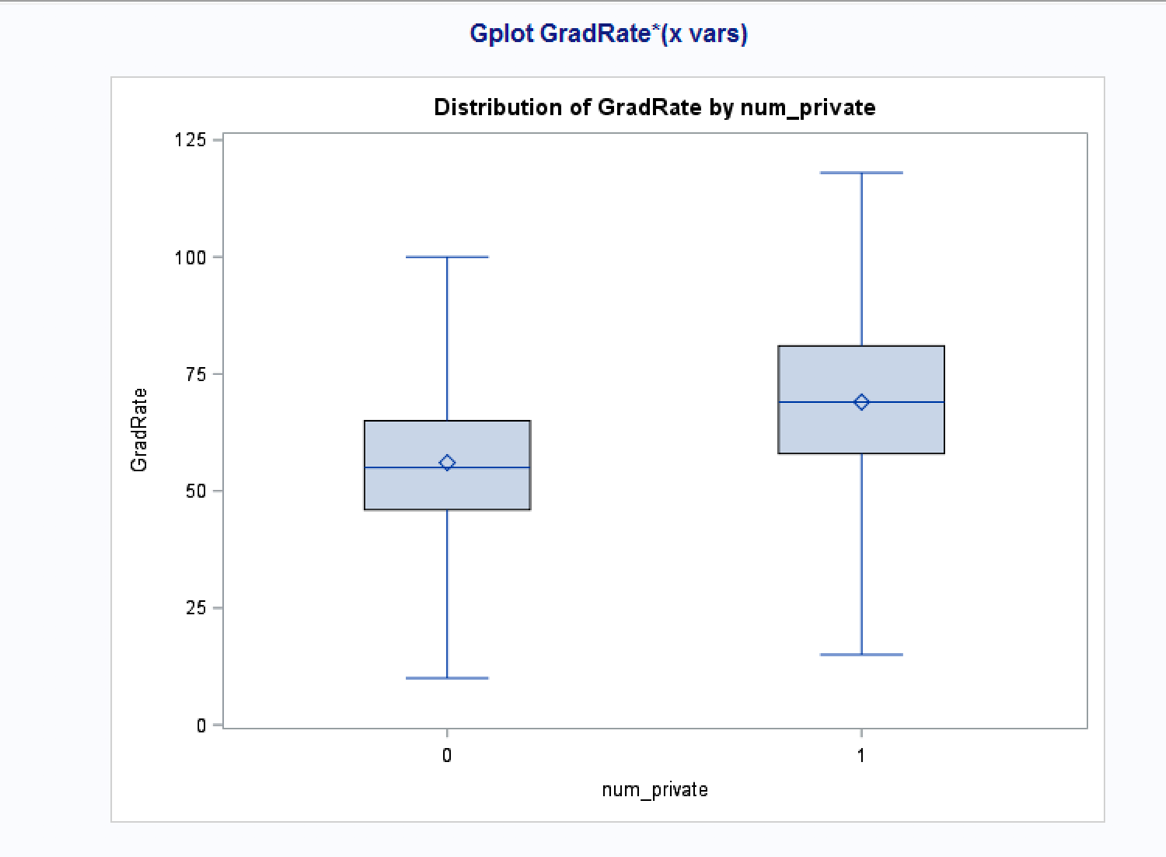
*by Elite10;*

*proc boxplot data=grad;*

*plot GradRate\*Elite10;*

*run;*

***Output:***

******

******

***Findings:***

*The above boxplots tell me that the 4 year graduation for private vs non-private schools is on average higher and that the 4 year graduation rate for elite schools vs non-elite schools is on average higher with less of a range among max and min grad rates.*

1. 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.

***Answer:***

***Code:***

*\*fit the full model with all predictors for GradRate;*

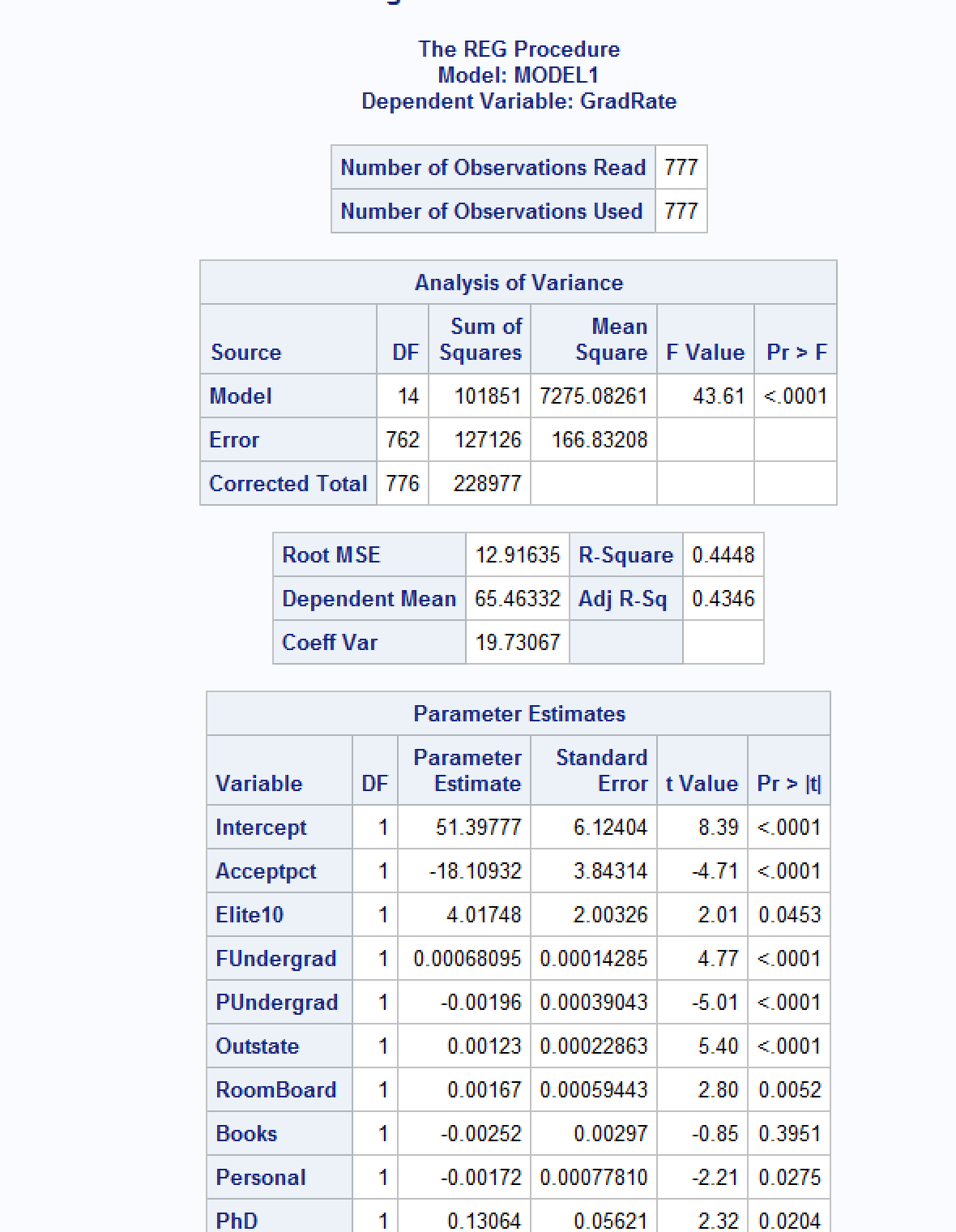
*title "Full Regression Model for GradRate";*

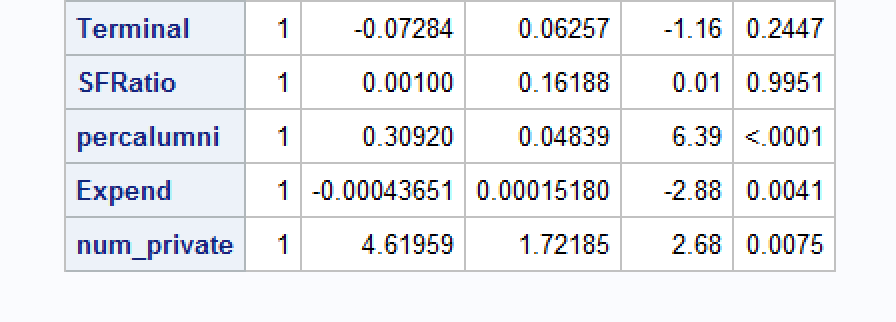
*proc reg data=grad;*

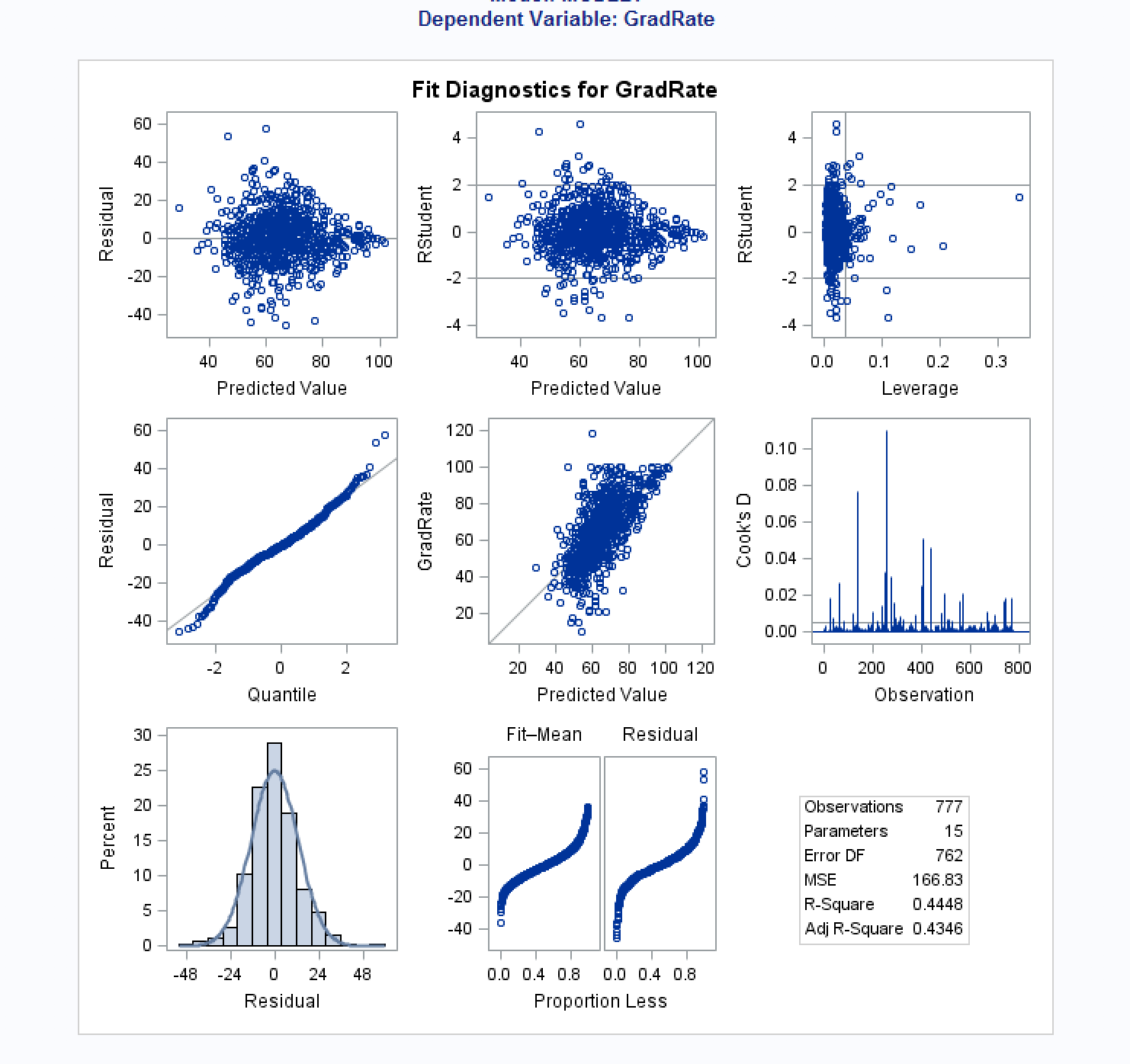
*model GradRate = Acceptpct Elite10 FUndergrad PUndergrad Outstate RoomBoard Books Personal PhD Terminal SFRatio percalumni Expend num\_private;*

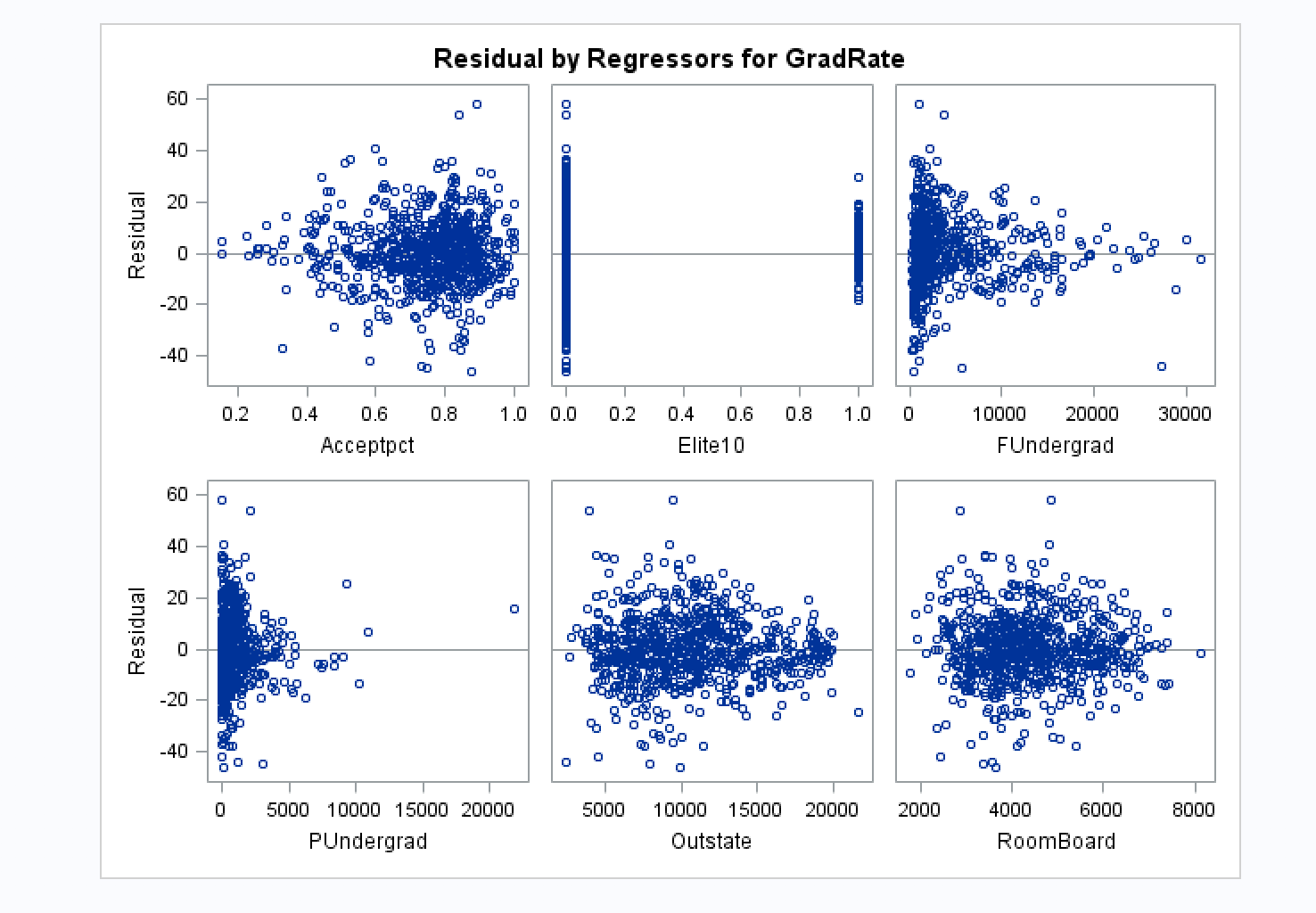
*run;*

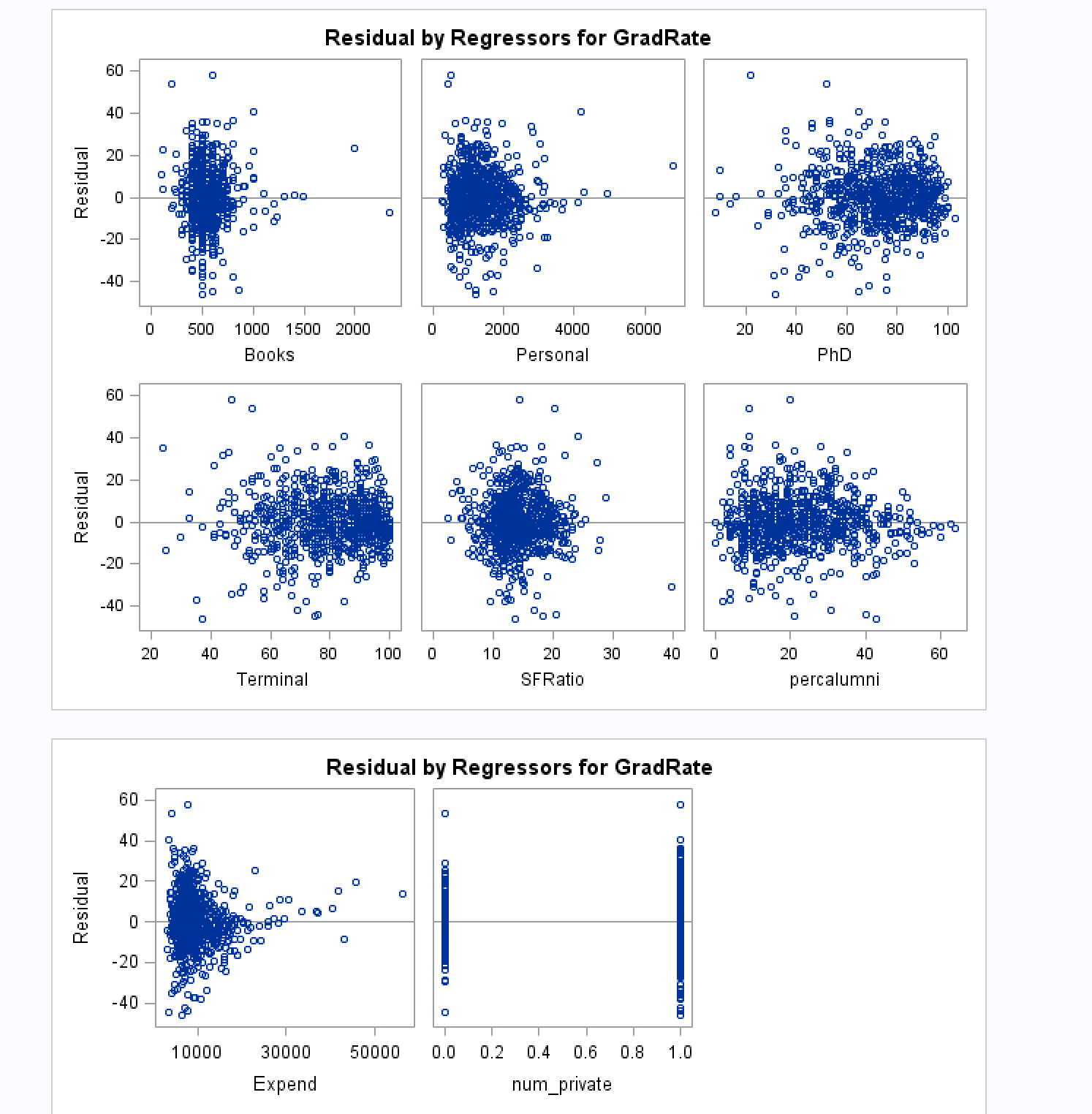
***Output:***

******

******

******

******

******

***Findings:***

*I can see from the parameter estimates table that the majority of the predictor variables are significate except for Books(p-val = 0.3951), Terminal(p-val = 0.2447), and SF ratio(p\_val = 0.9951), all of which are > alpha = 0.05.*

*I can see that the Adj R2 is 0.4346 which is not very high therefore I am assuming that the model can be improved.*

*As for goodness-of-fit, I see the the F-value is not very high at 43.61 although the p-value associated with it is very low at p < .0001, which tells me that at least one of the predictors is significant in predicting graduation rate and we can therefore reject the null hypothesis that states that no predictors are significant.*

*In looking at the residual plots, I am seeing some constant variance and independence violation, along with a curve in the normal probability plot. I see the violations particularly strong in the residual plots for FUndergrad, PUndergrad, Books, Acceptpct, Expend. This suggest a need for transformation of the variables. I tried transforming the y variable but did not notice a difference with the constant variance and independence violations by viewing the residual plots.*

*So I then transformed the x-variables that have plots that appear to violate model assumptions the most. I will post my code and outputs below:*

*\* transforming x vars to view residuals;*

*title "Transform x Variables";*

*data t\_grad;*

*set grad;*

*lnexpend=log(Expend);*

*lnfunder=log(FUndergrad);*

*lnpunder=log(PUndergrad);*

*lnbooks=log(Books);*

*lnpersonal=log(Personal);*

*run;*

*\*fit the full model with all predictors for GradRate;*

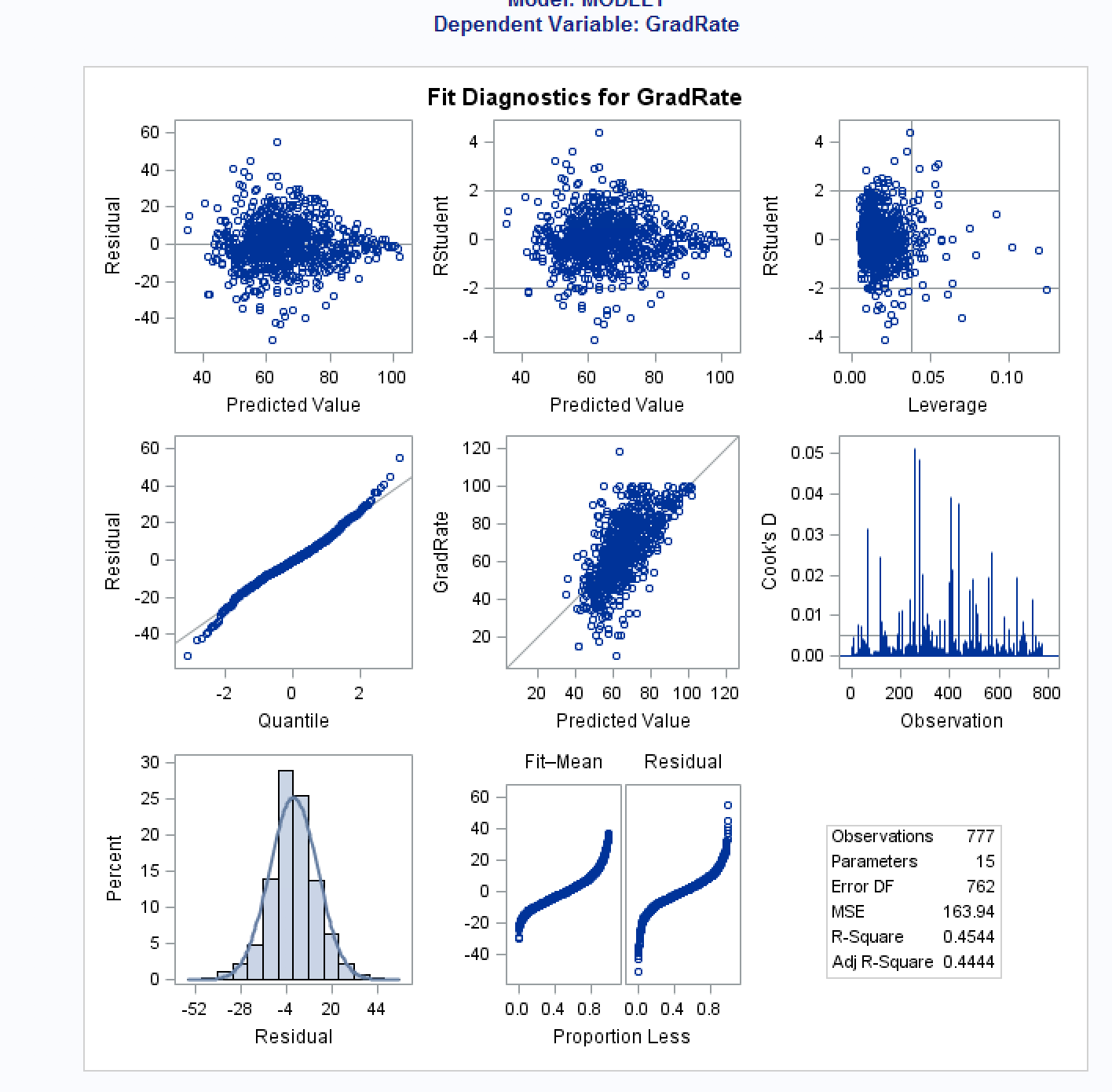
*title "Full Regression Model for GradRate";*

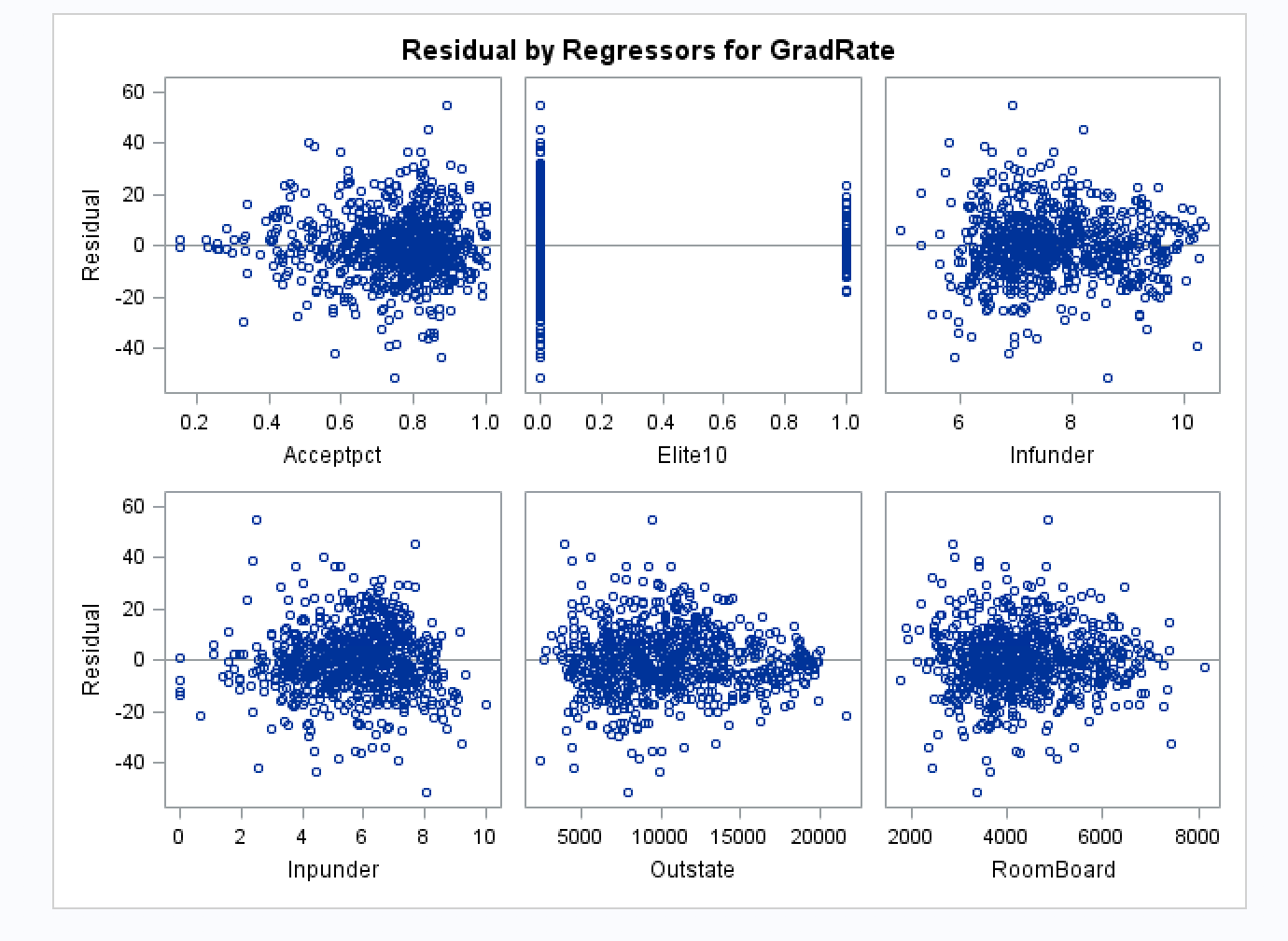
*proc reg data=t\_grad;*

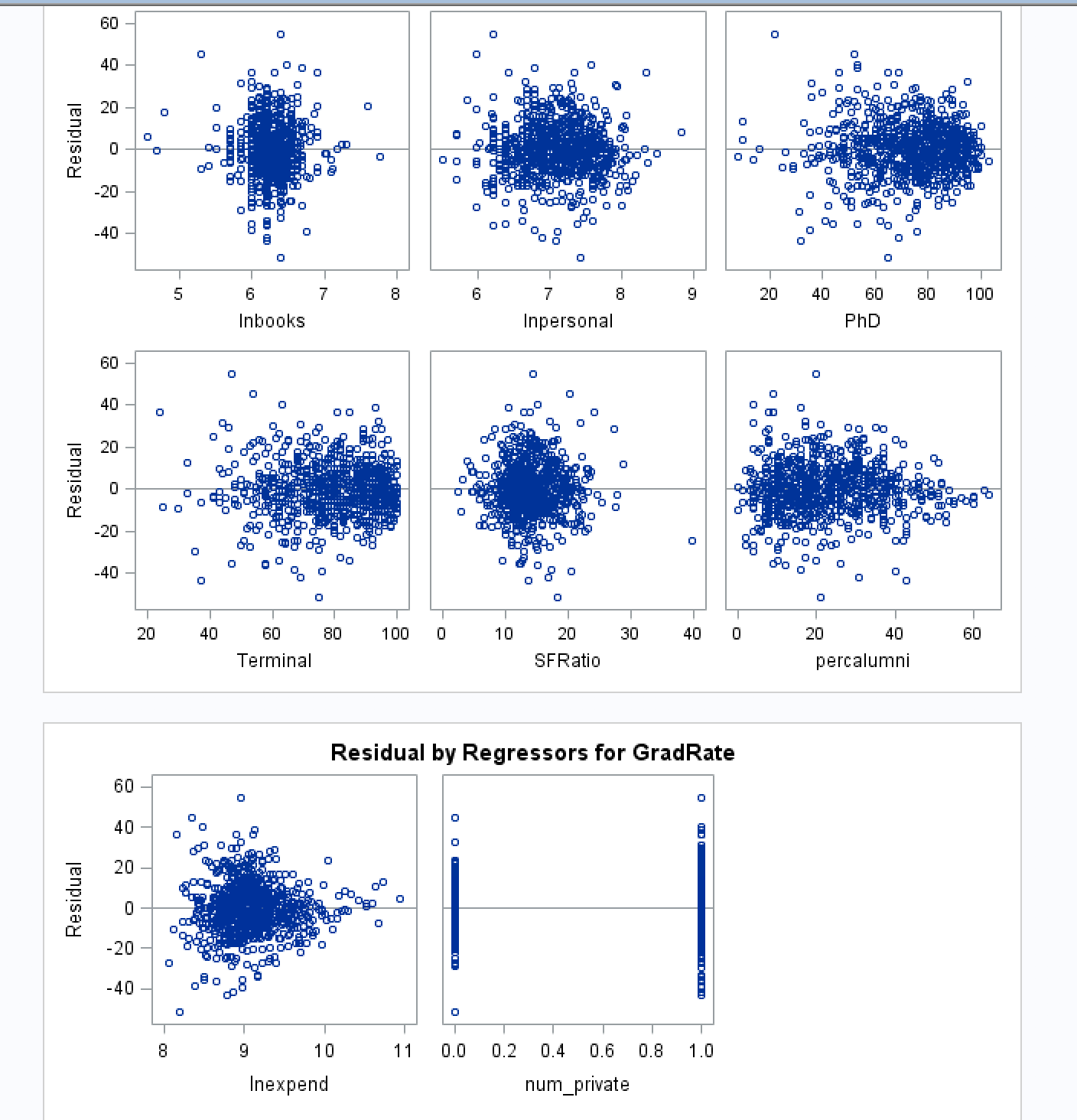
*model GradRate= Acceptpct Elite10 lnfunder lnpunder Outstate RoomBoard lnbooks lnpersonal PhD Terminal SFRatio percalumni lnexpend num\_private;*

*run;*

**

**

**



*After the transformations, the plots appear be more evenly scattered around the midline, the Adj R2 value went up very slightly, and the F-value went up slightly. I will use this final model with transformations to continue with the analysis.*

1. 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.

***Answer:***

***Code:***

*\* compute correlation values;*

*title "Pearson Correlation Matrix";*

*proc corr;*

*var GradRate Acceptpct Elite10 lnfunder lnpunder Outstate RoomBoard lnbooks lnpersonal PhD Terminal SFRatio percalumni lnexpend num\_private;*

*run;*

*\*fit the full model with all predictors for GradRate;*

*title "Full Regression Model for GradRate";*

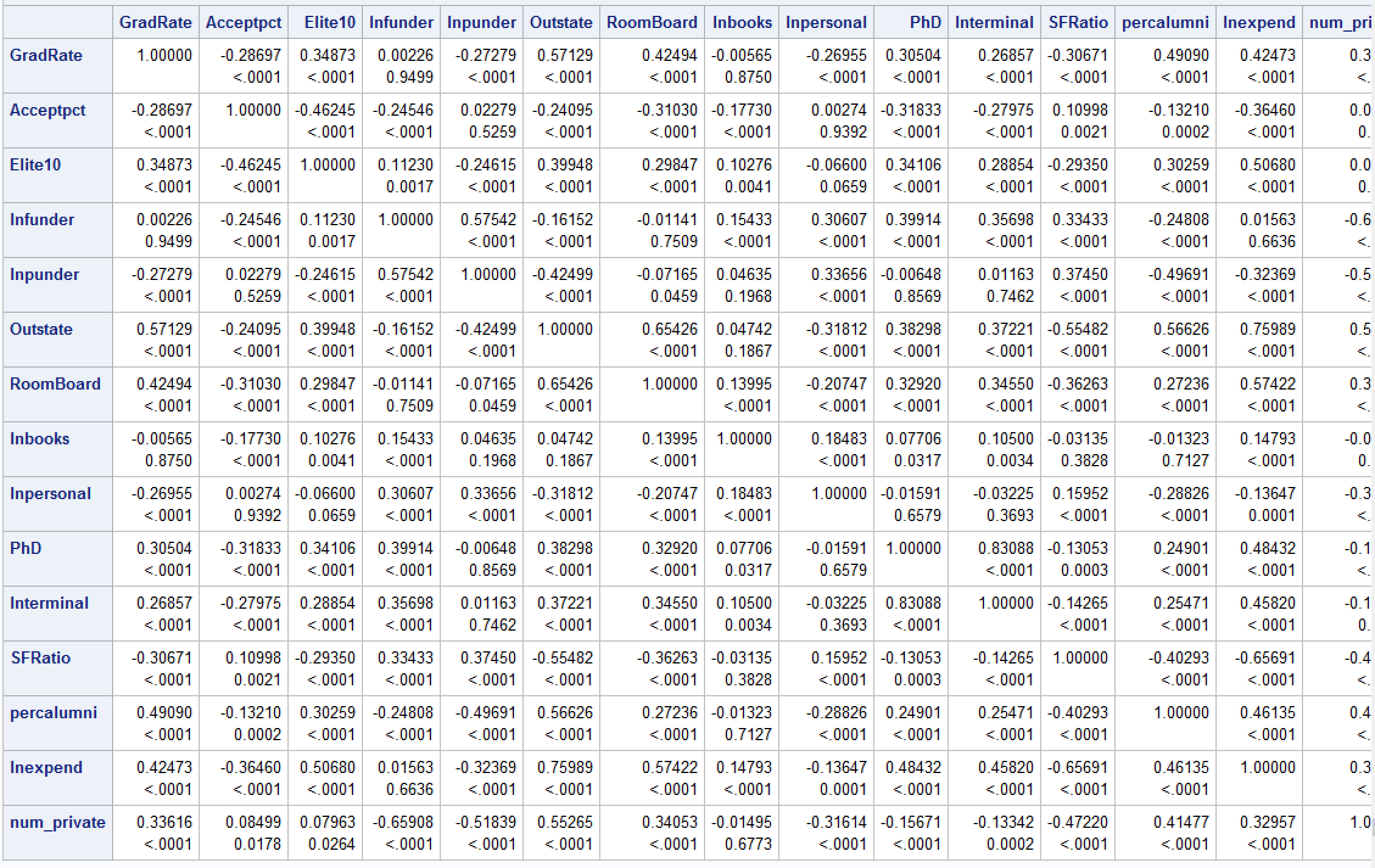
*proc reg data=t\_grad;*

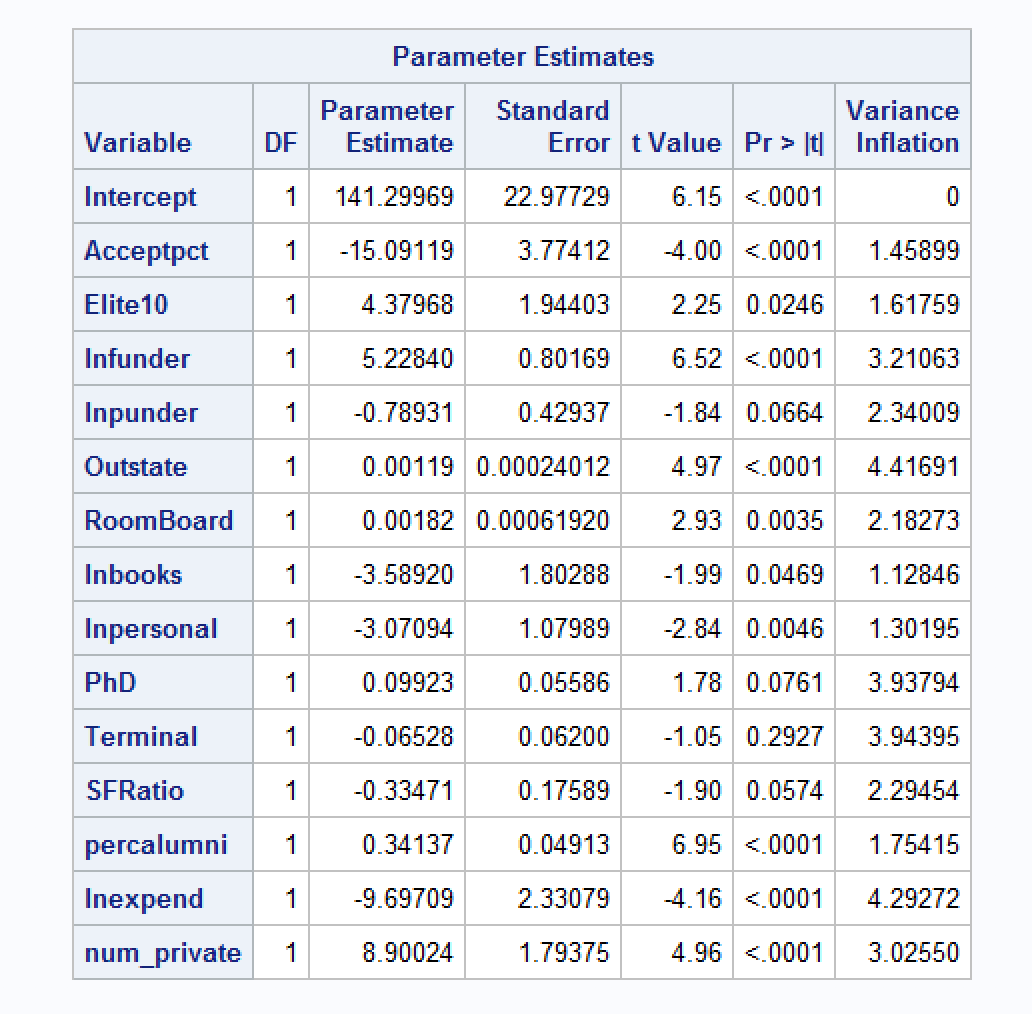
*model GradRate= Acceptpct Elite10 lnfunder lnpunder*

*Outstate RoomBoard lnbooks lnpersonal PhD Terminal SFRatio percalumni lnexpend num\_private /vif;*

*run;*

***Output:***

******

******

***Findings:*** *I do not see any multi-collinearity among the variables from looking at the correlation table and also from looking at VIF output. I see that all of the values are < 10, which tells me that we do not have multi-collinearity in the model.*

1. 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, include the outputs.

title "Selection Method-1: Adj-R2 Selection Method";

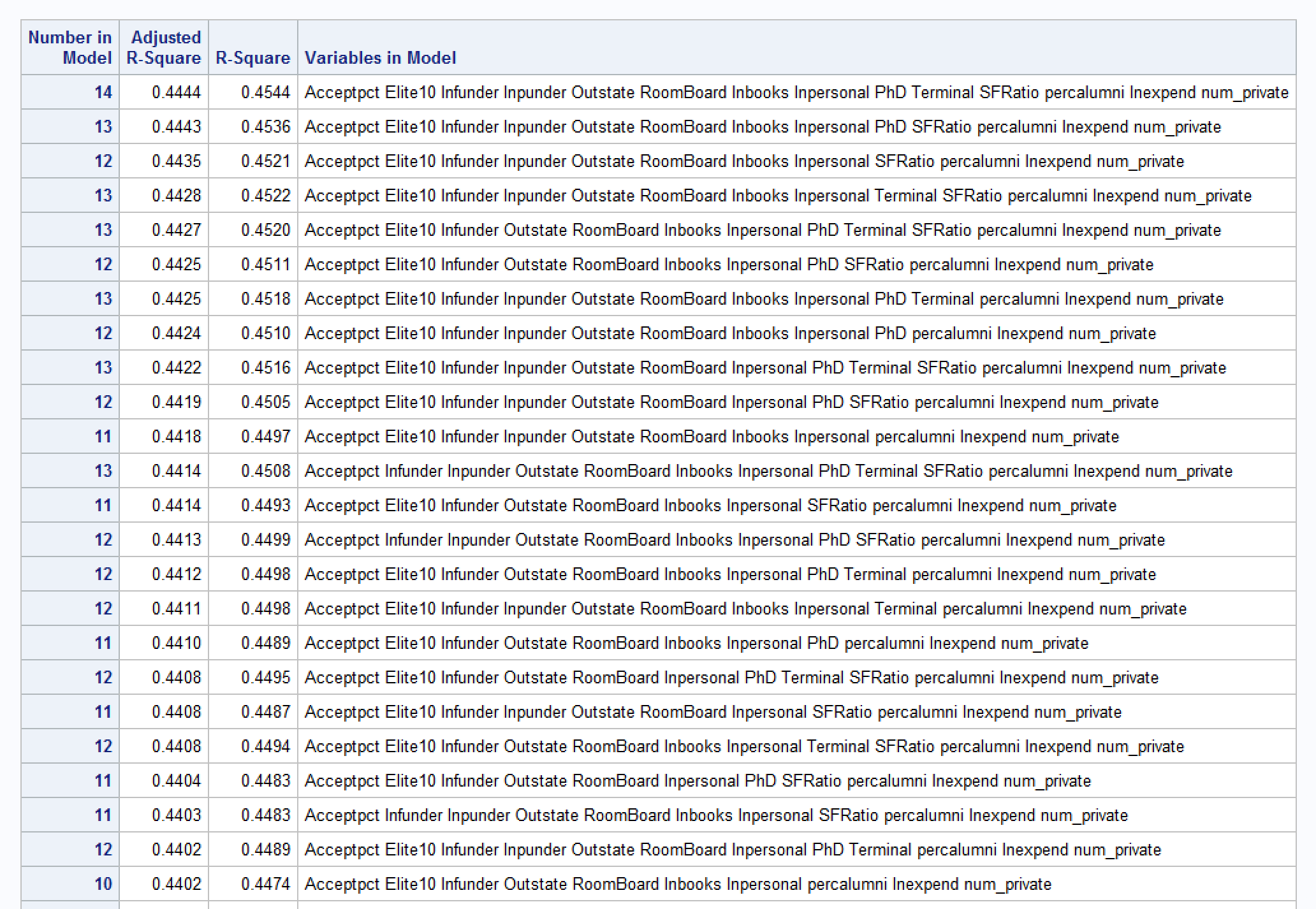
\* Selection Method-1: Adj-R2 Selection Method;

proc reg data=t\_grad;

model GradRate= Acceptpct Elite10 lnfunder lnpunder

Outstate RoomBoard lnbooks lnpersonal PhD Terminal SFRatio percalumni lnexpend num\_private /selection=adjrsq;

run;



title "Selection Method-2: forward Selection Method";

\* Selection Method-1: Adj-R2 Selection Method;

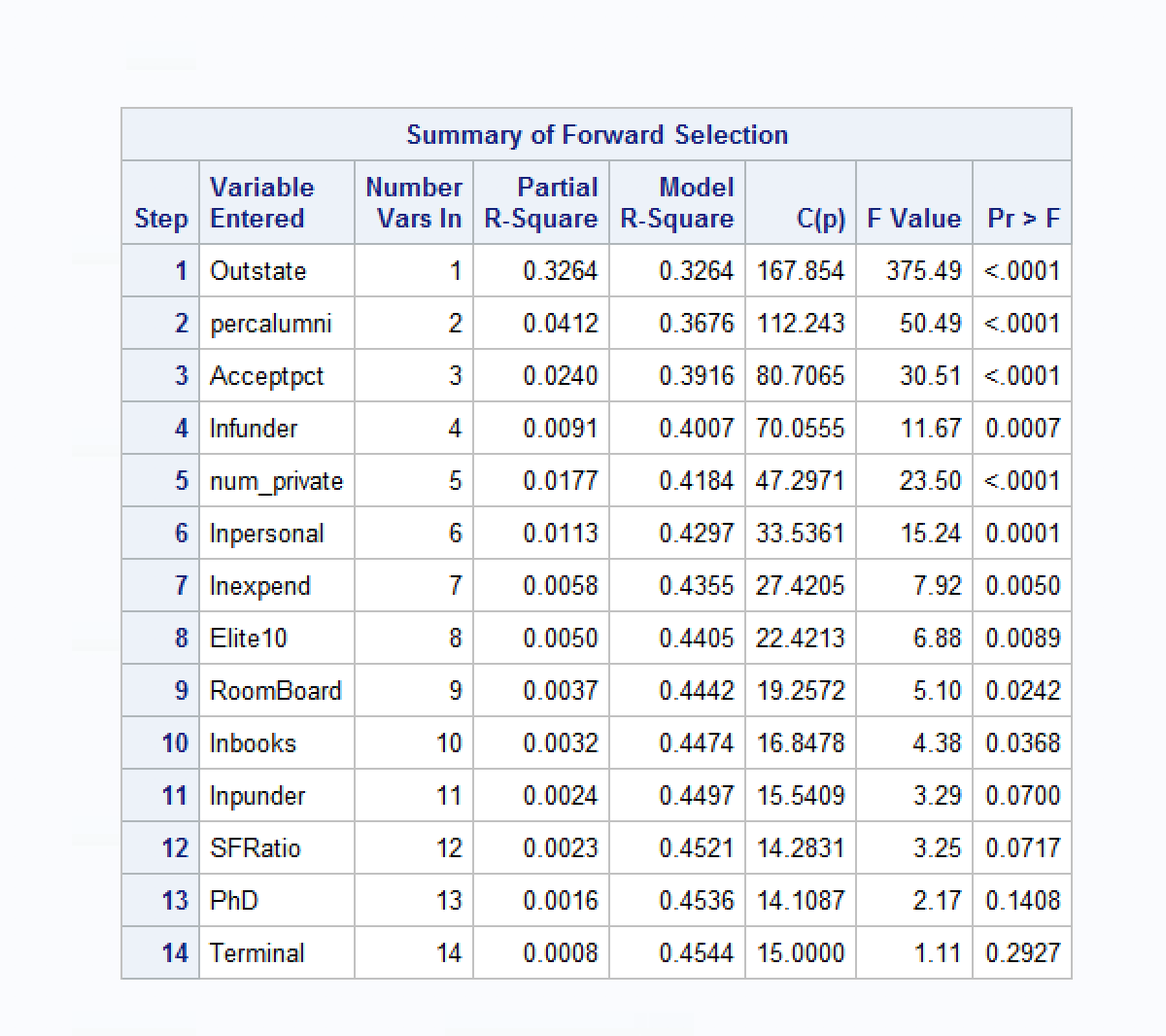
proc reg data=t\_grad;

model GradRate= Acceptpct Elite10 lnfunder lnpunder

Outstate RoomBoard lnbooks lnpersonal PhD Terminal SFRatio percalumni lnexpend num\_private /selection=forward;

run;





1. 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**.

***Answer:***

*Based on the model selection process and by looking at significance of variables, I can see that the Adj R2 does not increase very much at all when I choose the 12 variable model at Adj R2 = 0.4435 Also looking at forward selection, I see that the final two variables, PhD and Terminal can be removed without changing partial R2significantly. In fact, I could probably eliminate more variables as well that do not add very much to the Adj R2 value, and when fit the full model, I can see from the standard estimates that SFRatio is insignificant as well, with p-value = 0.0717. When I removed SFRatio, I found that lnpunder was also insignificant at p-val = 0.0700. The final model with PhD, Terminal, SFRatio, and lnpunder all removed gives me a model with all predictors significant, less predictors to avoid overfitting, a slightly higher F-value and only a very small insignificant change in Adj R2 = at 0.4402. See code and output below:*

*\*fit the full model after model selection – with lnpunder;*

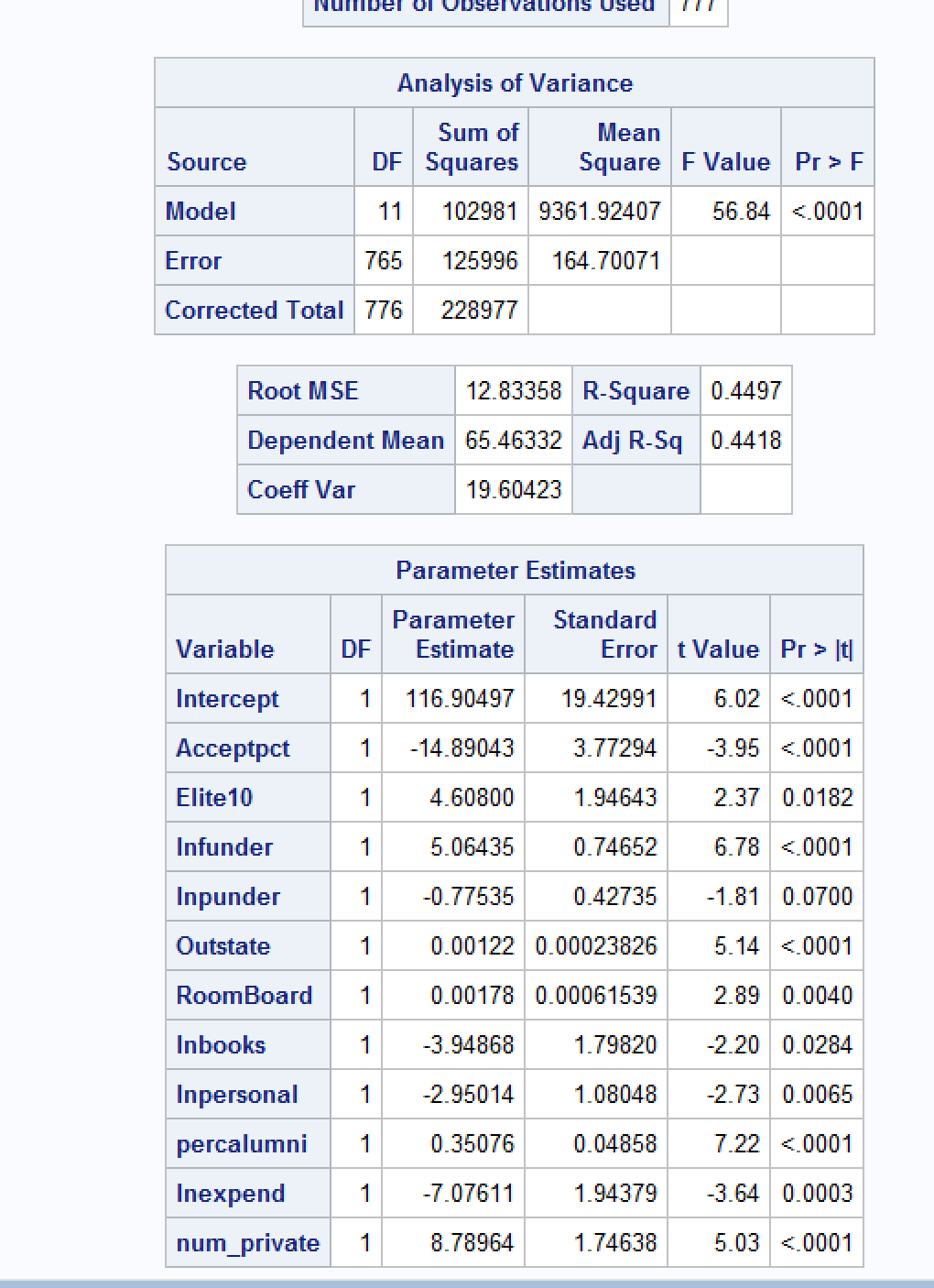
*title "Full Regression Model for GradRate";*

*proc reg data=t\_grad;*

*model GradRate= Acceptpct Elite10 lnfunder lnpunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend num\_private;*

*run;*

**

*\*fit the full model after model selection;*

*title "Full Regression Model for GradRate";*

*proc reg data=t\_grad;*

*model GradRate= Acceptpct Elite10 lnfunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend num\_private;*

*PLOT student.\*predicted.;*

*PLOT student.\*(Acceptpct lnfunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend);*

*PLOT npp.\*student.;*

*run;*

**

*The final fitted model with all significant predictors is:*

*4 year Graduation Rate = 114.45730 – 15.16039(Acceptpct) + 5.23926(Elite10) + 4.45228(lnfunder) + 0.00130(Outstate) + 0.00144(RoomBoard) – 3.76103(lnbooks) – 3.20521(lnpersonal) + 0.36963(percalumni) – 6.66241(lnexpend) + 8.80755(num\_private)*

*where num\_private=1 when Private=’Yes’ , otherwise 0*

1. 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.

***Answer:***

*\*fit the full model after model selection;*

*title "Full Regression Model for GradRate";*

*proc reg data=t\_grad;*

*model GradRate= Acceptpct Elite10 lnfunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend num\_private;*

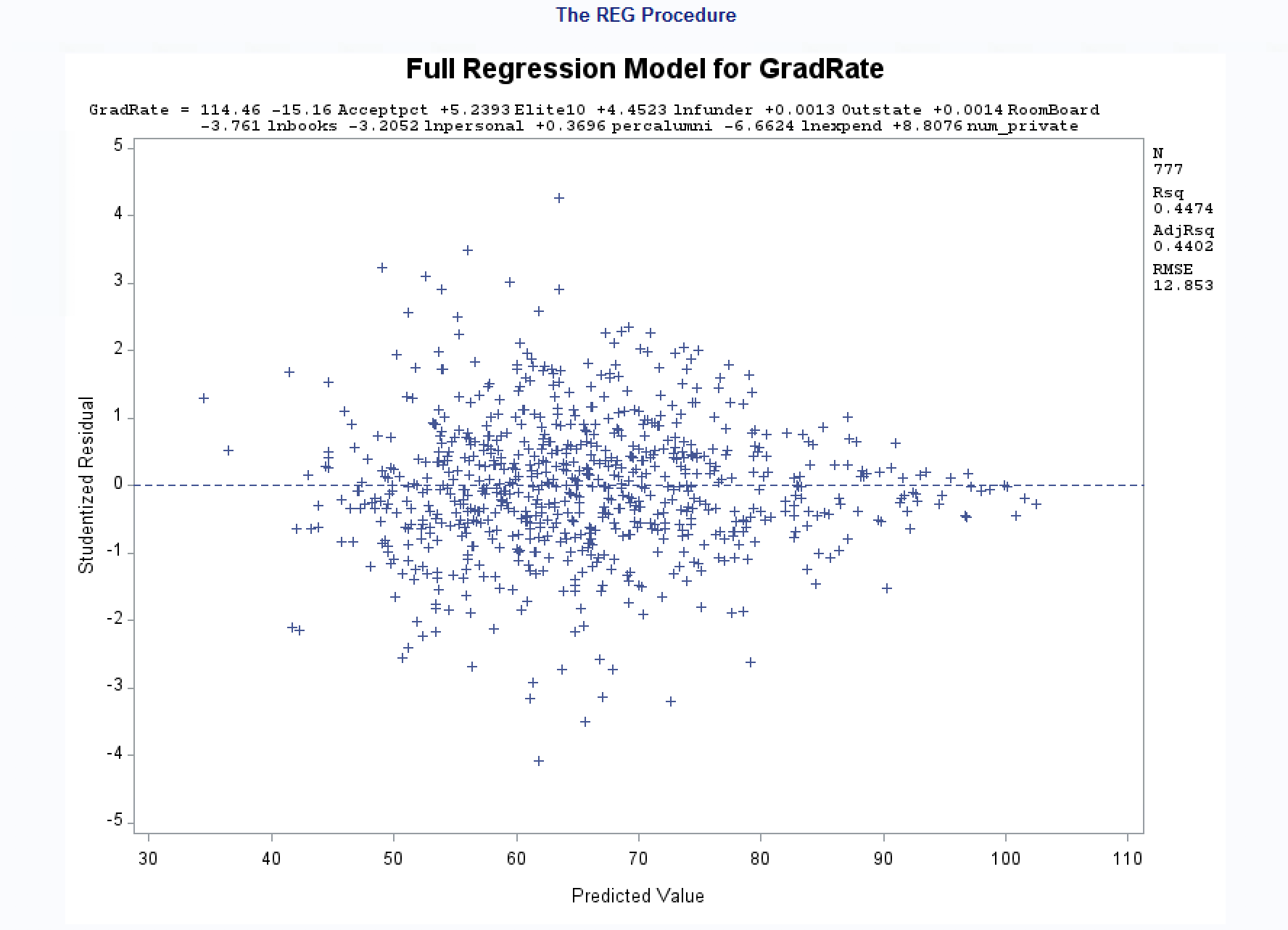
*PLOT student.\*predicted.;*

*PLOT student.\*(Acceptpct lnfunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend);*

*PLOT npp.\*student.;*

*run;*

**

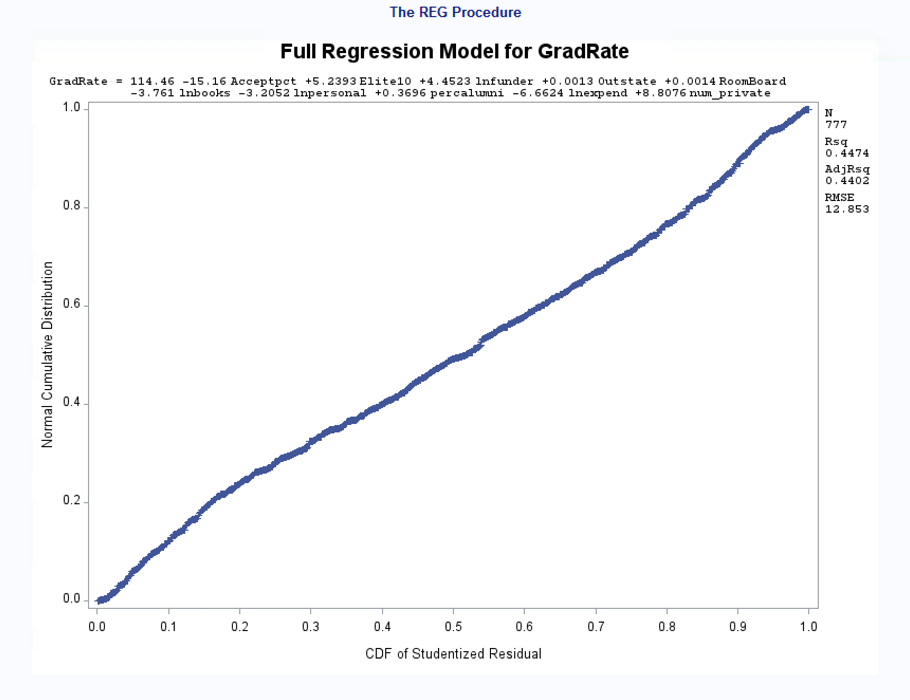
*I do see some outliers for sure and possible influential points. Otherwise, the residual plots shows a fairly random scattering of plots around the midline.*

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

***Answer:***

*The normal probability plot of residuals shows a slight s-shape to it, likely from the possible outliers and influential points. When these are addressed, my assumption is that it will likely become more aligned with the normal probability line.*

**

**

1. Are there any outliers or Influential Points? Compute appropriate statistics. Include the outputs and explain why/why not?

***Answer:***

***Code:***

*\*fit the full model after model selection;*

*title "Full Regression Model for GradRate";*

*proc reg data=t\_grad;*

*model GradRate= Acceptpct Elite10 lnfunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend num\_private/ stb vif influence r;*

*run;*

*\*dropping outliers and influential points - multiple ones;*

*TITLE 'M1 with outlier and influential point removed';*

*data t\_grad10;*

*set t\_grad9;*

*if \_n\_ = 147 then delete;;*

*run;*

*\*with point removed;*

*title "Full Regression Model for GradRate with points removed";*

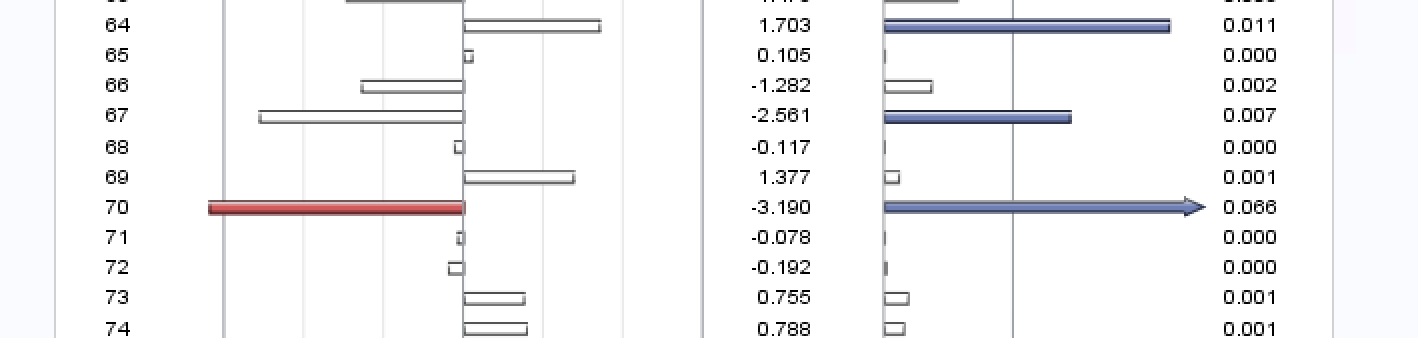
*proc reg data=t\_grad10;*

*model GradRate= Acceptpct Elite10 lnfunder*

*Outstate RoomBoard lnbooks lnpersonal percalumni lnexpend num\_private/ stb vif influence r;*

*run;*

***Output:***

******

******

***Findings:***

*Yes, the above output shows one example that I found of over 10 outlier/influential points so I removed one at a time until the* AdjR2 did not show any more improvement. After removing outliers and influential points, the above table shows my final output showing the full fitted model without outliers and influential points.

1. Analyze the AdjR2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.

***Answer:***

*The final fitted model explains approximately 54% of the variation in graduation rates among universities which is not very good. This tells me that there may be other variables that we have not accounted for in our dataset that account for graduation rates of universities.*

1. 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.

***Answer:***

*My conclusion of this final model is that it is an ok model for understanding graduation rates among universities. The most significant predictors of graduation rate are the percent of alumni who donate, the out-of-state tuition, and the number of full-time undergraduate students. Given the final fitted model:*

*4 year Graduation Rate = 114.45730 – 15.16039(Acceptpct) + 5.23926(Elite10) + 4.45228(lnfunder) + 0.00130(Outstate) + 0.00144(RoomBoard) – 3.76103(lnbooks) – 3.20521(lnpersonal) + 0.36963(percalumni) – 6.66241(lnexpend) + 8.80755(num\_private)*

*where num\_private=1 when Private=’Yes’ , otherwise 0*

*I can see graduation rates increase by 8.8 percentage points when we compare a non-private school to a private school and that there is approximately a 5.24 point increase in graduation rate from non-elite school to elite school. So there seem to be higher graduation rates for private and elite schools.*

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

**See all code above**