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*

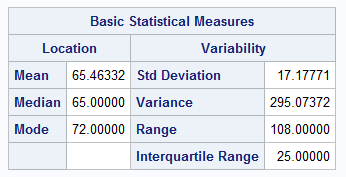
*Expend Instructional expenditure per student*

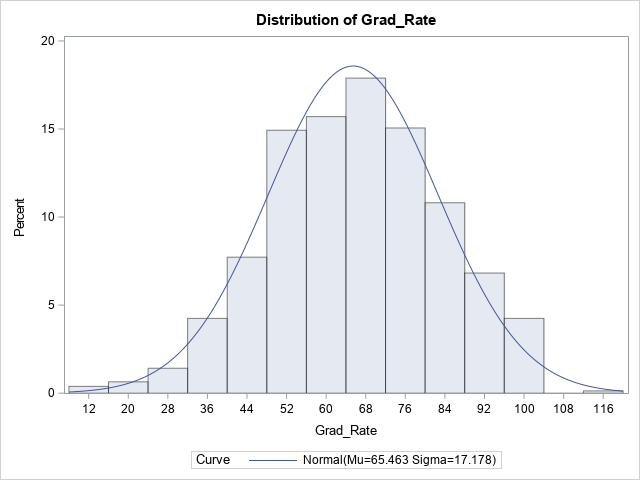
*Grad.Rate Graduation rate in 4 years*

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

**Project steps & questions:**

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).



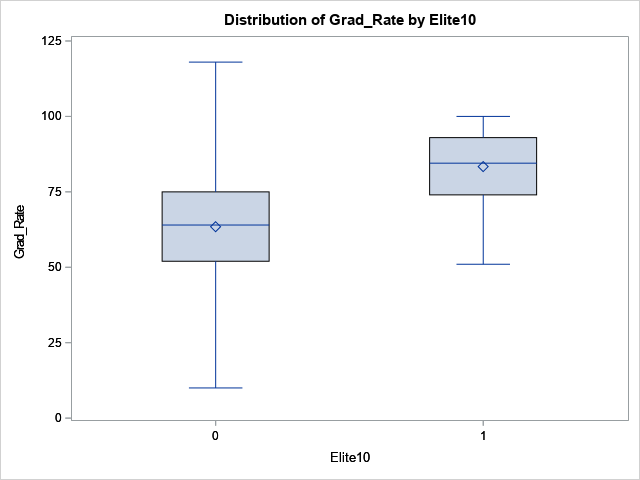
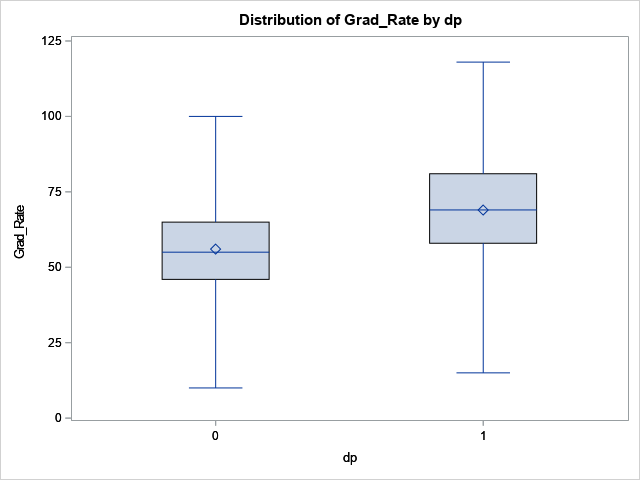


The distribution of grad\_rate is almost symmetric. There is no need to apply and transformation.

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).

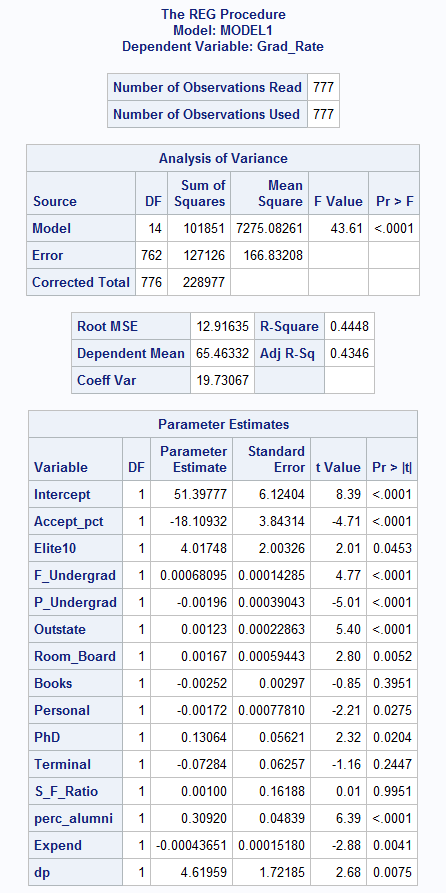
From the scatterplots, Grad.Rate shows weak positive linear association with Outstate, Room.Board, PhD, and perc.alumni. Grad.Rate shows weak negative linear association with S.F.Ratio. Grad.Rate shows no linear association with Accept.pct, F.Undergrad, P.Undergrad, Books, Personal, Terminal, and Expend. We cannot say linear association between Grad.Rate and dummy variables (Elite10 and dp).

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).



Considering the mean graduation rate, the graduation rate in private universities is higher than in public universities, and elite universities is higher than not elite universities. The range of graduation rate in private universities is slightly wider than in public universities. The range of graduation rate in not elite universities is much wider than elite universities.

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.



H0: βi = 0

Ha: At least one coefficient βj ≠ 0

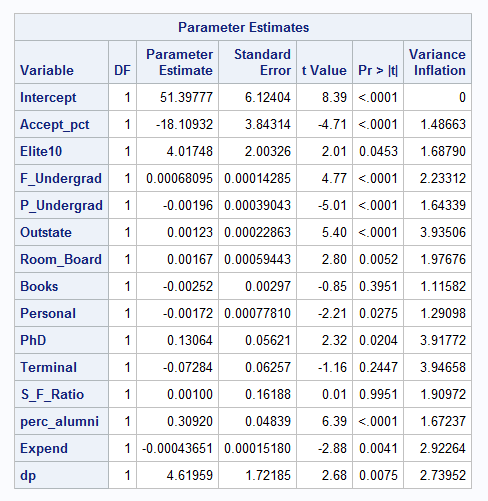
F-value = 43.61

P-value < 0.001

Conclusion: The H0 is rejected, and there is at least one x-variable that has a significant effect on Y. The F-test gives strong support to the fitted model.

According to parameter estimate, Accept\_pct has the most effect on Grad.Rate. The Adj R2 is 0.4346 which is not good enough. S\_F\_Ratio is the most not significant variable. Books and Terminal are also not significant.

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.

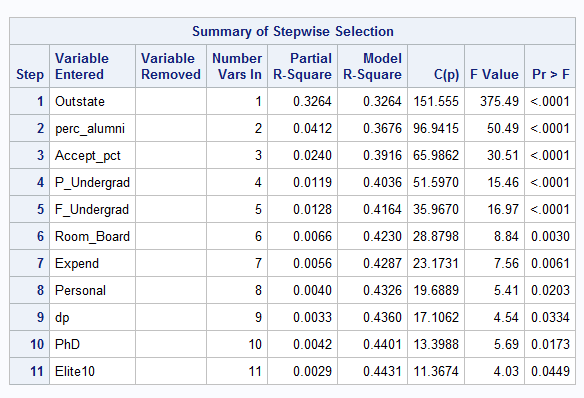


There is no multi-collinear problem because the variance inflation of all variables is less than 10.

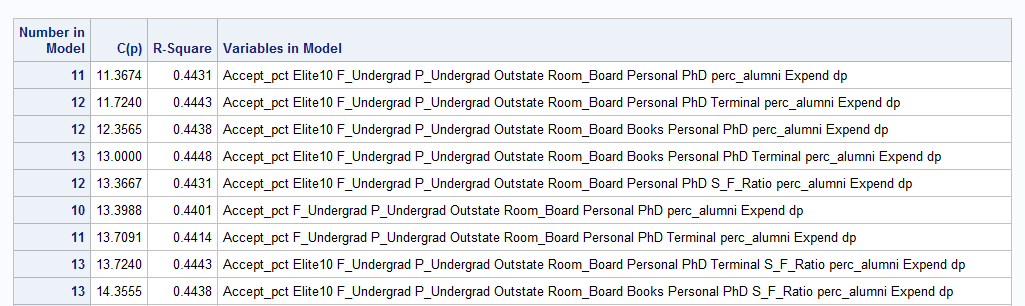
1. Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate*.* You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, include the outputs.

Stepwise:





Cp:

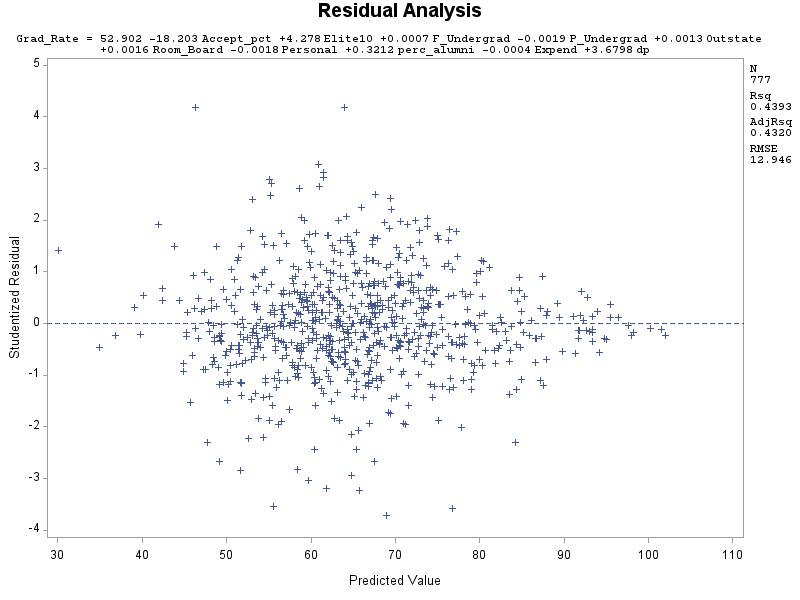


1. Fit a final regression model **M1** for Grad.Rate based on the results in f) – i.e. optimal model. Explain your choice. Write down the expression of the estimated model **M1**.

Stepwise method and Cp method give the same result. In Cp method, Cp should < or ≈ K+1. The R2 of the first one is slightly less than the second one, however the first one has fewer variables. Therefore the model M1 is:

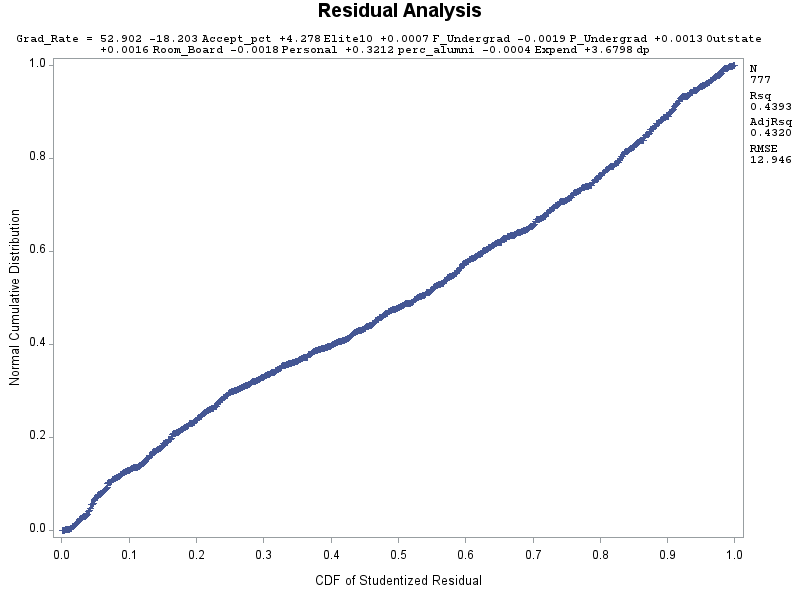
Grad\_Rate = 48.404 + -17.782\*Accept\_pct + 4.022\*Elite10 + 0.0007\*F\_Undergrad – 0.002\*P\_Undergrad + 0.0012\*Outstate + 0.00153\*Room\_Board – 0.0018\*Personal + 0.084PhD + 0.306\*perc\_alumni – 0.0004Expend + 4.77\*dp

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.



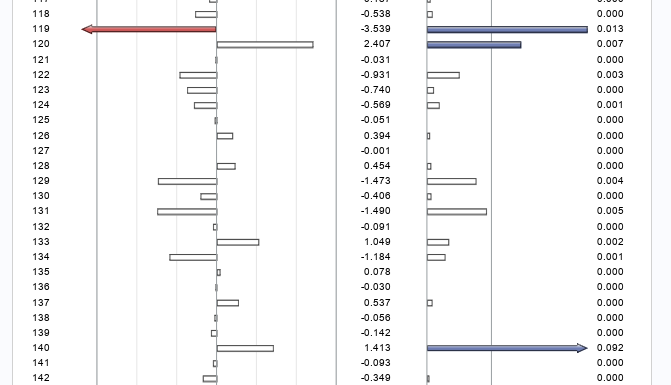
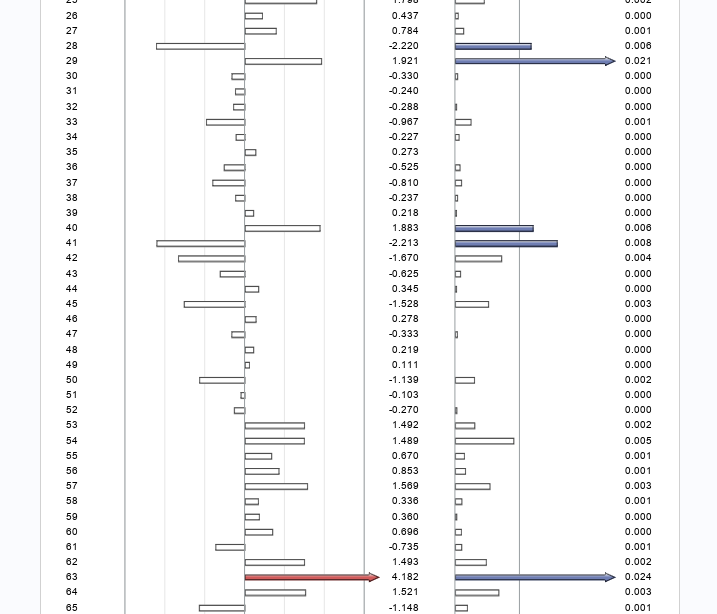
Plot looks OK for the most part. It seems like constant variance and independence maybe very slightly violated.

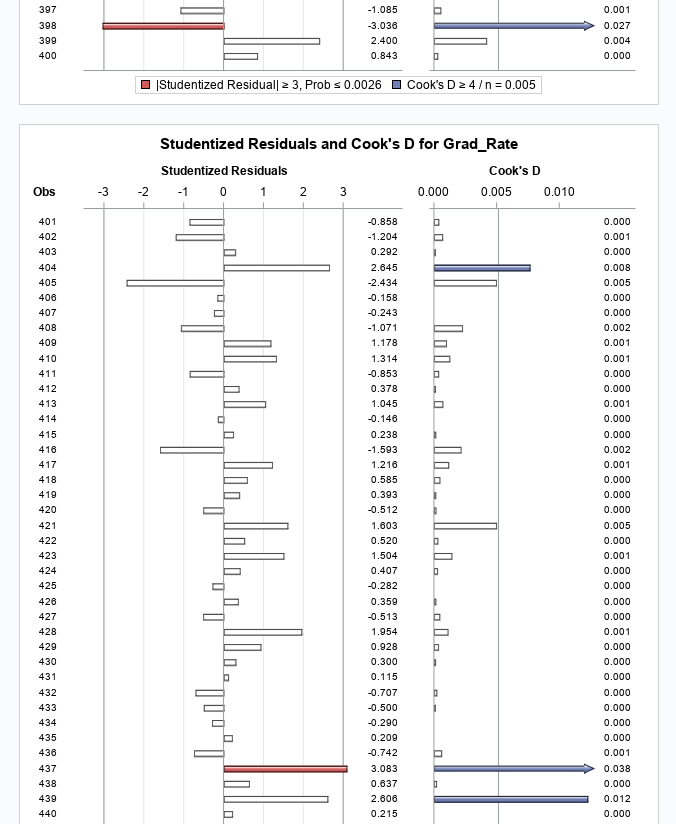
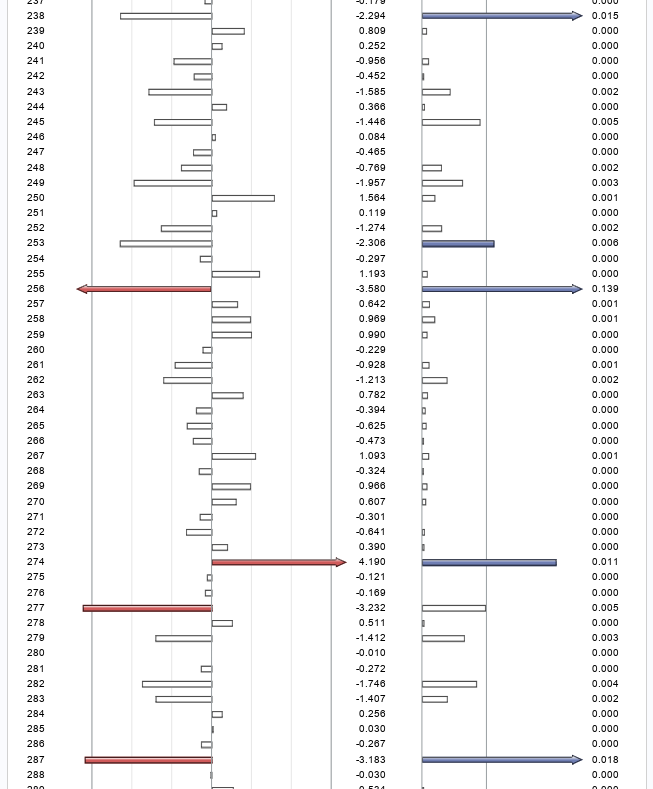
1. Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? Include the outputs and explain.

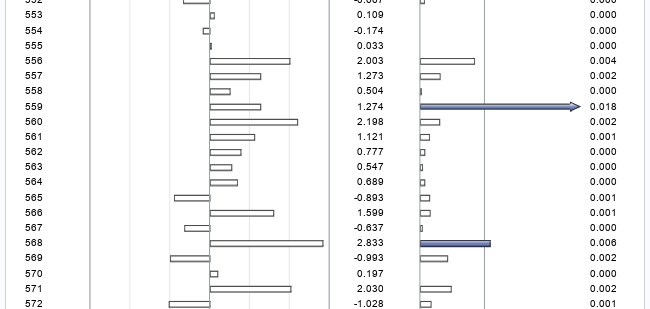
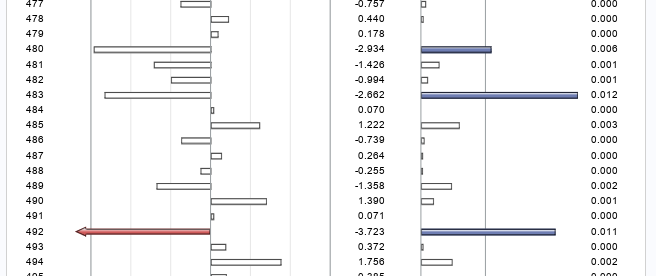


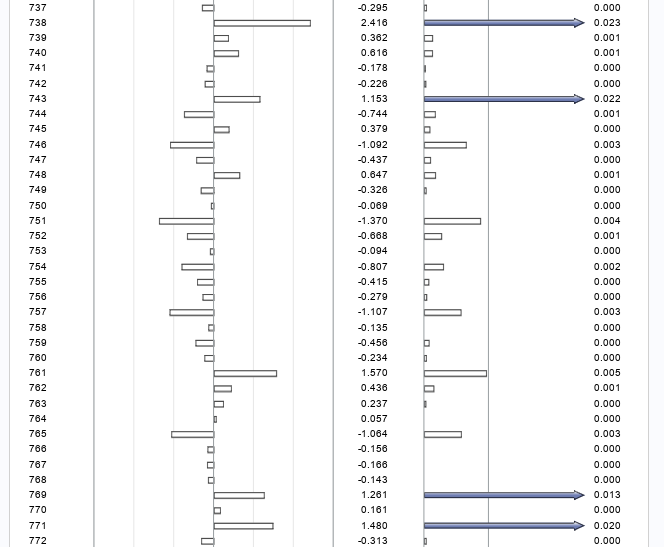
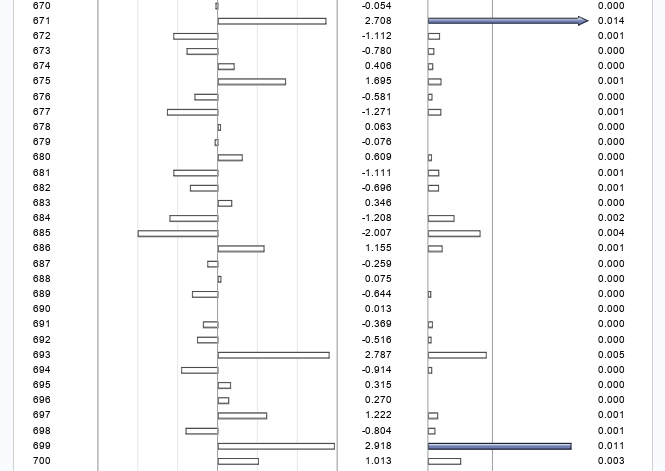
Plot shows almost 45degree straight line. Thus, it is linear and normally distributed.

1. 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?









There are 9 outliers and 14 influential points.

Remove the outliers and influential points, rerun the model, Adj-R2 increase from 43% to 48%



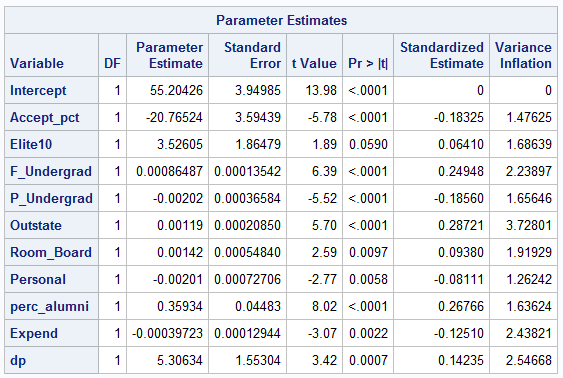
There are 4 more outliers after rerun the model, if we remove the new outliers the Adj-R2 only increase 1%. Therefore, it is not necessary to remove the new ouliers.

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



Adj-R2=0.4773. Therefore, 47.73% of the variation in Grad\_Rates is explained by its relationship with Accept\_pct, Elite10, F\_Undergrad, P\_Undergrad, Outstate, Room\_Board, Personal, perc\_alumni Expend, and dp.

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.



The standardized estimate of Outstate is 0.287, therefore it is the most important predictors.

If the universities type change from public to private, the graduation rates will increase 5.3%. Thus, it shows a significant different.

If the universities change from not elite to elite, the graduation rates will increase 3.5%. Therefore, elite universities do have higher graduation rates