Model selection – “hands on” exercise

Question: What state has the smartest high school students, on average?

Data: State average SAT score, in SAT.csv

Results, with no adjustment: IA has the highest average. Looks great, but wait, IA has among the lowest percent of HS students who take the SAT.

Where is IA after adjusting for relevant variables associated with SAT score? The variables in the data set are:

 Takers: % of HS seniors who took the SAT

 income: median family income of test takers

 years: # years of formal education in social sciences, natural sciences, and humanities, summed

 public: % of test takers attending public schools

 expend: per pupil state expenditure on secondary schools

 rank: median class ranking (as a percentile) of test takers in their class (100% = top student).

1) Which variables should be included in a multiple regression model that predicts SAT scores?

2) Any issues or concerns with this model?

3) After fixing the major issue, which variables should be included in a multiple regression model that predicts SAT scores?

4) What state is the most above what is expected given the variables in the multiple regression model?

Hint: Look at the usual residuals. A large positive residual for a state tells you it has students who did unusually well on the SAT.

5) Is Iowa still exceptional, after adjusting for relevant variables?