Multiple Regression Homework Assignment

The file `States.txt` on the course website presents education data for the U. S. states and the District of Columbia (Washington D.C.).

1.   Using these data, perform a multiple linear regression of the average SAT math score (y) for the states on the percentage of students taking the SAT exam (x_1) and teachers' pay (x_2). Interpret the regression coefficients (a, b_1, and b_2). What is R^2 for the regression? How is it to be interpreted?

It is not feasible to do this problem by hand, so you will need to use the *R Commander*. In addition to the state names (as two-character abbreviations), the data file includes the following variables:

> `region`, coded as follows:
>> `ENC` (East North Central)
>> `ESC` (East South Central)
>> `MA` (Middle Atlantic)
>> `MTN` (Mountain)
>> `NE` (Northeast)
>> `PAC` (Pacific)
>> `SA` (South Atlantic)
>> `WNC` (West North Central)
>> `WSC` (West South Central)
>
> `population` of the state in 1000s
> `satVerbal`, average verbal score in the state of high school students taking
>> the SAT university-entry exam
>
> `satMath`, average math score on the SAT exam
> `percentTaking`, percentage of graduating high school students who take the SAT
> `percentNoHS`, percentage of adults in the state without a high school diploma
> `teacherPay`, average teachers' pay in the state in $1000s

Read the data in R via the *Data ►Import data ► from text file, clipboard, or URL...* menu in the *R Commander*.

Performing a multiple regression with the *R Commander* is almost the same as performing a simple regression, except that you need to select more than one explanatory variable in the *Linear Regression* dialog: just control-click on more than one variable in the *Explanatory variables* list.

2.   Regress the average SAT math scores on teachers' pay (i.e., ignoring the percentage of the students taking the exam). Is the coefficient for teachers' pay in this simple regression substantially different from the coefficient for the same variable in the multiple regression? If so, how do you explain the difference?

3.   To check that the partial relationships between SAT math scores and the two explanatory variables are linear, plot residuals from the regression model against each explanatory variable:

Make the multiple regression (presumably the first regression that you fit) the "active" model in the *R Commander* by clicking on the *Model* button (near the top of the *R Commander* window) and selecting `RegModel.1` in the resulting dialog box.

Add residuals and fitted values to the data set via the *Models ► Add observation statistics to data...* menu. (You can uncheck all of the boxes except the ones for residuals and fitted values.)

Use *Graphs ► Scatterplot...* to construct scatterplots of the residuals against each of `percentTaking` and `teacherPay` and against the fitted values. Assuming that the multiple regression is the first regression equation that you fit, the residuals should be named `RegModel.1.residuals` and the fitted values `RegModel.1.fitted`. In drawing these scatterplot, I suggest that you check the boxes for *Smooth line* and *Show spread* in the *Options* tab, and set the span for the smoother to a larger value, such as 90.

Do the residual plots reveal any problems for the multiple linear regression?

Submit the R Commander output and graphs along with the written portion of the homework.