Please read ALL INSTRUCTIONS very carefully:

1. Deadline for submission: 3/14/2016, at start of class (multiple choice part will be taken that night also).

2. The take home portion of the midterm will be worth 70% of the total. The remaining 30% will be from a multiple choice exam to be taken on 3/14/2016. The worth of each individual question in this portion of the exam is indicated on the question. Together, the take-home and multiple choice parts of the exam are worth 40% of your overall grade for the class.

3. Complete the exam on your own.
   - Remember this is a formal assessment, just like an exam you are completing in class in a controlled time. As such, it should be an assessment of your capabilities alone – remember what the penalty for academic dishonesty is – 0 on this assignment, an inevitable fail in the class, and possible dismissal from the program/university. Please do not force me to even consider such punishment. In statistics, collusions are especially easy to spot. Please notice the absence of smiley face here.

4. Submission guidelines
   a. Please word process your answers, and double space throughout. Any tables, graphs or other output required to supplement your answers should be cut and pasted from SPSS directly into the appropriate place in your word document (just take screenshots). If you have trouble with this, ask for help.

5. The data file to use for this midterm is on the course web site on the midterm page

The data file is named and labeled, but here is a full description of the variables in the file:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>Name of state (in quotation marks)</td>
</tr>
<tr>
<td>Spend</td>
<td>Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)</td>
</tr>
<tr>
<td>ptratio</td>
<td>Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994</td>
</tr>
<tr>
<td>tsalary</td>
<td>Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)</td>
</tr>
<tr>
<td>persat</td>
<td>Percentage of all eligible students taking the SAT, 1994-95</td>
</tr>
<tr>
<td>satverb</td>
<td>Average verbal SAT score, 1994-95</td>
</tr>
<tr>
<td>satmath</td>
<td>Average math SAT score, 1994-95</td>
</tr>
<tr>
<td>sattot</td>
<td>Average total score on the SAT, 1994-95</td>
</tr>
</tbody>
</table>

And now, the questions (cue drum roll):

1. **(15 points total) Data manipulation, descriptive stats, z-scores**
   a. *Use the [Transform_Compute] command to do this*
      Create a new variable, called “caring” which is the product of “spend+tsalary-ptratio” (in other words the sum of the first two variable minus the third). Now create another variable, “zcaring” which is formed using the z-scores of spend, tsalary and ptratio, combined in the same way (sum of the first two, minus the third). Which state comes top for caring, and which for zcaring? Explain why different states might come top for the “caring” and “zcaring” variables, using what you know of z scores for the explanation. (5 points)
   b. For each of the 3 variables used to comprise the variable “caring,” find the state with the maximum value. Which of these three maximum values is the least likely to occur, statistically
speak (think z-scores)? List any states where that maximum value would be expected to occur less than 5% of the time, based on the distribution of scores across the states (again, think z-scores). (10 pts)

2. (10 points total) Use SPSS to generate correlations and scattergrams among all possible pairs of the following variables: spend, ptratio, tsalary, sattot. (I won’t need to see the scattergrams, but they may help you to visualize the kind of relationship present in each case)
   a. Report each correlation coefficient in the manner shown in Cronk (i.e. using appropriate APA style) (5 points).
   b. Interpret the relationship among SAT score and the other three variables – speculate as to what might be the reasons for the level and direction of association denoted by the correlation coefficient in each of the 3 cases? (5 points)

3. (15 points total) Simple Regression.
   a. Construct a model to predict total SAT score from state expenditure. Interpret the model’s significance. Also, state the regression equation, and state the expected total SAT score for a state spending $10,000 per pupil. (10 points)
   b. Test the assumptions of the regression model (normality, linearity, homoscedasticity). Explain whether any assumptions were violated, and if they were state whether these would be likely to result in under- or over-estimating the strength of the relationship between the 2 variables. (5 points)

4. (15 points) Multiple regression and model reduction.
   Create a regression model using all other 6 variables to predict total SAT score. Clearly, there is at least one problem with it. Use what you know of model reduction to reduce the model to make it as good a model as you can. Remember, the purpose of the model is to be able to predict student’s performance in SAT exams before they take them. Use your common sense (or maybe not common – think carefully about what each variable is, and what role it could/should play in predicting SAT), and all principles explained relevant to model reduction, to arrive at your best choice of regression model to predict total SAT score. Your grade on this question will be determined by your explanation of your choices at each stage of model reduction, not by the final model you end up with. So make sure you include a good argument for your choices at each stage of the process. Also include the R^2, ANOVA, and coefficient tables from the output at each stage of the analysis. Once you have arrived at your final model, fully report this model, using the format laid out in class and in Cronk.
   Finally, report the residual for Kansas according to this model.

5. (15 points) Theory. For the z-score formula (assume normal distributions in all cases):
   a. Explain, with reference to the formula, why converting two variables to z-scores allows direct comparison between them (5 points).
   b. Let’s say you missed this mid-term, and there were no make-ups, but you had a legitimate excuse. What would be a better (fairer) procedure for me to take, and why? (5 points)
      i. Take your mean score on the final, and give you the same score on the mid-term.
      ii. Find your z-score for the final, and use your z-score on the final, together with the mean and standard deviation of the mid-term to turn that z-score onto a score on the mid-term. That’s your mid-term score. Thus:

\[
Z_{\text{final}} = \frac{\text{Score}_{\text{final}} - \bar{X}_{\text{final}}}{SD_{\text{final}}}, \text{ so } (Z_{\text{final}} \times SD_{\text{midterm}}) + \bar{X}_{\text{midterm}} = \text{Score}_{\text{midterm}}
\]
iii. Find your z-score for the final, and use the mean and standard deviation of the final to generate a score on the mid-term. That’s your mid-term score. Thus:

\[ Z_{\text{final}} = \frac{\text{Score}_{\text{final}} - \overline{X}_{\text{final}}}{\text{SD}_{\text{final}}}, \text{ so } (Z_{\text{final}} \times \text{SD}_{\text{final}}) + \overline{X}_{\text{final}} = \text{Score}_{\text{ midterm}} \]

iv. Take the mean score of the entire class on the mid-term, and give you that instead.

c. Refer to the top and bottom of the correlation equation to explain why the added data point in panel (1) below might affect the overall correlation of the data set differently than the added data point in panel (2). (5 points)