## Main Instructions

1. Open "caschools.sav" in SPSS and reacquaint yourself with the meaning of each variable. Inspect frequencies and plots for each variable.
2. Our research question is to see what school characteristics predict academic performance. "read" and "math" represent academic performance. All other variables besides district, school, and county will be used as potential predictors. However, many of the predictors and the outcome variable should be modified. In particular, (a) convert "grades" to a numeric variable; (b) create a new variable that is teachers per student; (c) create a new variable that is computers per student; (d) Apply a transformation to student numbers to create a more normally distributed version of student numbers (log might be a good choice); (e) combine reading and mathematics performance into a single measure of academic performance.
3. Produce a correlation matrix of all the predictor variables (where you've created a transformed variable, use that instead of the original) and the dependent variable (i.e., academic performance). Tidy up the correlation matrix as per these instructions: <http://jeromyanglim.blogspot.com.au/2009/02/formatting-correlation-matrices-in.html> . Interpret the pattern of results.
4. Run a multiple regression predicting academic performance from the predictors (i.e., where you've created a transformed variable, use that instead of the original).
5. Check features of the model's relation to the data: (a) homoscedasticity, (b) normality of residuals; (c) absence of outliers in the residuals; (d) linearity of relationship between predictors and dependent variable; (e) multicollinearity of the predictors. Interpret what this means and consider whether you should do anything as a result (e.g., exclude predictor, transform predictor, remove certain cases, etc.)
6. Re-run multiple regression if necessary and interpret substantive results: (a) interpret overall model fit and significance (b) interpret individual coefficients (unstandardised beta, standardised beta, semi-partial r), confidence intervals, and statistical significance.
7. Provide an overall interpretation of the results.

## Additional Exercises

1. Answer the question: to what extent do school features predict mathematics performance more than reading performance?
2. Answer the question: "grades", calworks, lunch, income, english, are all baseline measures predicting academic performance; Which of following three variables "teachers per student", "computers per student" and "expenditure per student" leads to the biggest incremental prediction in academic performance over and above the baseline measures? Run three separate hierarchical regressions where based on model 1 you add expenditure

# Answers

## 1. Open "caschools.sav" in SPSS and reacquaint yourself with the meaning of each variable. Inspect frequencies and plots for each variable.

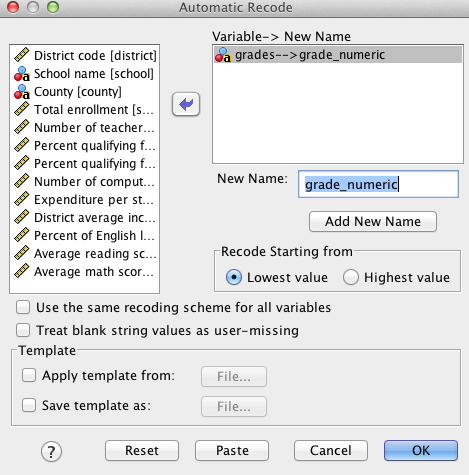
Frequencies are available through "analyze - descriptives - frequencies" (put in all variables). You can also request a histogram which would be an appropriate summary for the numeric variables. The aim of this process is to reacquaint you with the meaning of each variable. Think about what scale each variable is on. Think about the general descriptive statistics for each variable. Recognise that some variables are just identifiers (like district, and school) whereas others are characterising important properties of the schools (like how many students there are).

## 2. Our research question is to see what school characteristics predict academic performance. "read" and "math" represent academic performance. All other variables besides district, school, and county will be used as potential predictors. However, many of the predictors and the outcome variable should be modified. In particular, (a) convert "grades" to a numeric variable; (b) create a new variable that is teachers per student; (c) create a new variable that is computers per student; (d) Apply a transformation to student numbers to create a more normally distributed version of student numbers (log might be a good choice); (e) combine reading and mathematics performance into a single measure of academic performance.

It is a common requirement that before substantive analyses can be performed many transformations need to be applied to the raw data.

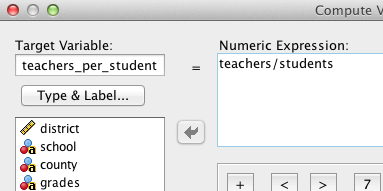
(a) Grades is by default stored as string variable and thus, not suited to inclusion in a multiple regression that expects numeric input. You can use "transform - automatic recode" to quickly convert a categorical variable into numeric variable with value labels corresponding to the relevant numbers. However, it doesn't give you quite as much control as "Transform - recode into different variable" in terms of how numbers are assigned to categories.

We'll use automatic recode here:



(b) Teachers per student is more relevant predictor than raw numbers of teachers.

Use "Transform compute" and do the simple arithmetic



If you paste the SPSS syntax is also quite simple, and you should get in the habit of saving and transitioning to using syntax for these simple transformations.

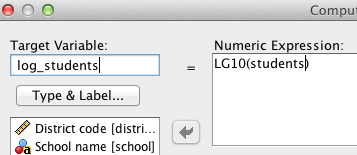
(c) Thus, including computers per student we have the following syntax

COMPUTE teachers\_per\_student=teachers/students.

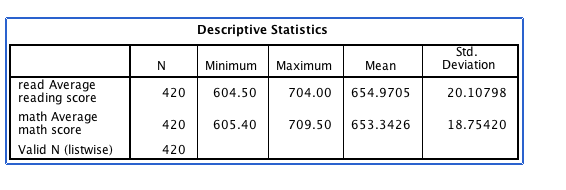
COMPUTE computers\_per\_student=computer/students.

EXECUTE.

(d) There are several different transformations to normalise positively skewed count variable. A log transformation is a common one. Under transform compute, you will see the function group "arithmetic" and under that you will see that "lg10" applies a log base 10 transformation.



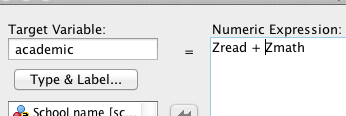
(e) When combining two performance measures like reading and mathematics we generally want to make sure that they are weighted equally. Thus, we should check that they are on the same scale (e.g., analyze - descriptives -descriptives)



They do have a similar standard deviation so we could probably just take the average of the two variables or the general solution would be to convert both to z-scores and average the two z-scores.

I'll show the z-score approach here:

"Analyze-descriptives-descriptives" and select "save as standardised variables" and put in math and read. This will create zmath and zread in the datafile which are z-score versions of the original variables (i.e., mean of zero and sd of 1; you may wish to check this). Then use "transform - compute" to sum the two zscores



## 3. Produce a correlation matrix of all the predictor variables (where you've created a transformed variable, use that instead of the original) and the dependent variable (i.e., academic performance). Tidy up the correlation matrix as per these instructions: http://jeromyanglim.blogspot.com.au/2009/02/formatting-correlation-matrices-in.html . Interpret the pattern of results.

"Analyze - correlate - bivariate"

Thus, if you paste the set of variables should probably look something like this:

CORRELATIONS

/VARIABLES=calworks lunch expenditure income english grade\_numeric

teachers\_per\_student computers\_per\_student log\_students academic

/PRINT=TWOTAIL SIG /MISSING=PAIRWISE.

The website link describes a process of making the correlations more presentable. Follow those steps:

The correlations provide a rich set of information. In particular, socioeconomic factors are strongly correlated with school academic performance. Resource factors such as expenditure, teachers and computers per student and show positive but much weaker relationships with academic performance. Resource factors don't seem to be strongly correlated with socioeconomic factors.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | Percent qualifying for CalWorks (income assistance) |  |  |  |  |  |  |  |  |  |
| 2 | Percent qualifying for reduced-price lunch | .74 |  |  |  |  |  |  |  |  |
| 3 | Expenditure per student | .07 | -.06 |  |  |  |  |  |  |  |
| 4 | District average income (in USD 1,000) | -.51 | -.68 | .31 |  |  |  |  |  |  |
| 5 | Percent of English learners | .32 | .65 | -.07 | -.31 |  |  |  |  |  |
| 6 | grade span of district | .07 | .11 | -.17 | -.16 | -.05 |  |  |  |  |
| 7 | teachers\_per\_student | -.02 | -.13 | .63 | .23 | -.19 | -.09 |  |  |  |
| 8 | computers\_per\_student | -.15 | -.20 | .29 | .20 | -.25 | -.02 | .31 |  |  |
| 9 | log\_students | .08 | .09 | -.16 | .12 | .38 | -.03 | -.35 | -.34 |  |
| 10 | academic | -.63 | -.87 | .19 | .71 | -.64 | -.16 | .23 | .27 | -.12 |

## 4. Run a multiple regression predicting academic performance from the predictors (i.e., where you've created a transformed variable, use that instead of the original).

Analyse - Regression - linear

DEPENDENT academic

IVS: works lunch expenditure income english grade\_numeric

teachers\_per\_student computers\_per\_student log\_students

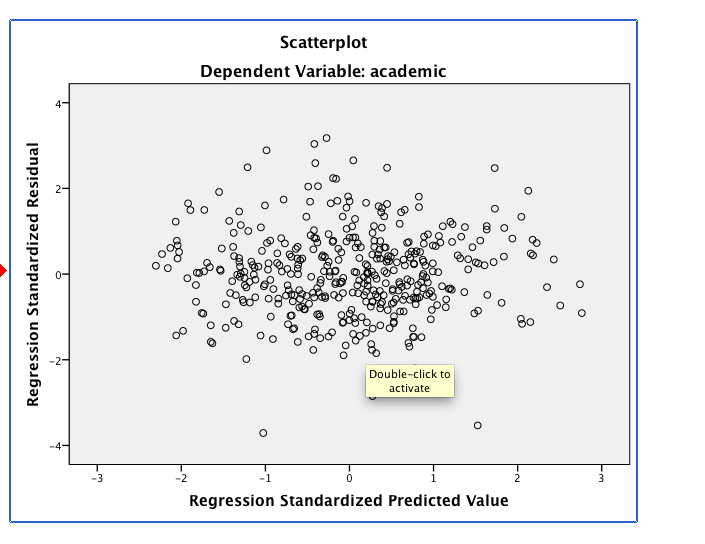
Under statistics add: confidence intervals, descriptives, part and partial correlations and collinearity diagnositcs and casewise diagnostics

Under plots at x: zpred by y: zresid

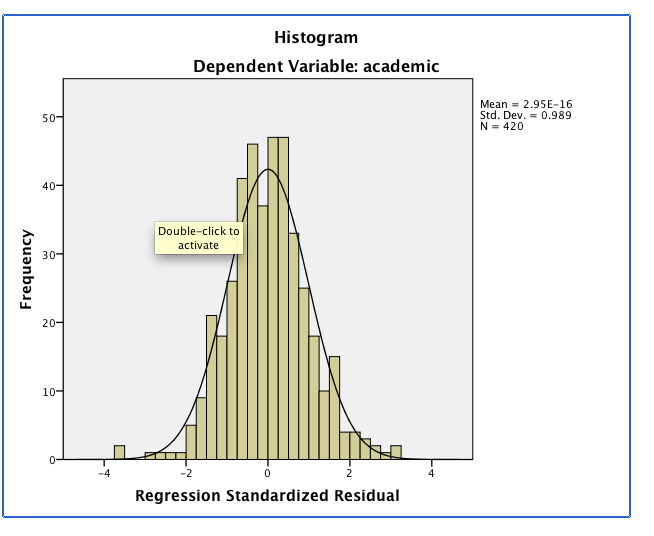
Under plots add histogram and produce all partial plots.

## 5. Check features of the model's relation to the data: (a) homoscedasticity, (b) normality of residuals; (c) absence of outliers in the residuals; (d) linearity of relationship between predictors and dependent variable; (e) multicollinearity of the predictors. Interpret what this means and consider whether you should do anything as a result (e.g., exclude predictor, transform predictor, remove certain cases, etc.)

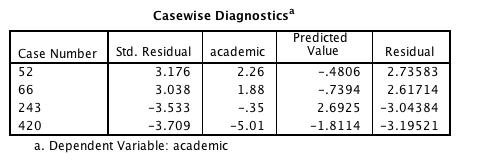
(a) homoscdasticity: The plot of predicted and residuals suggests that the degree of spread in the residuals is fairly constant across predicted values (thus, homoscedasticity)



(b) normality of residuals: The residuals look very normal; there are a couple of possible outliers with low negative residuals.

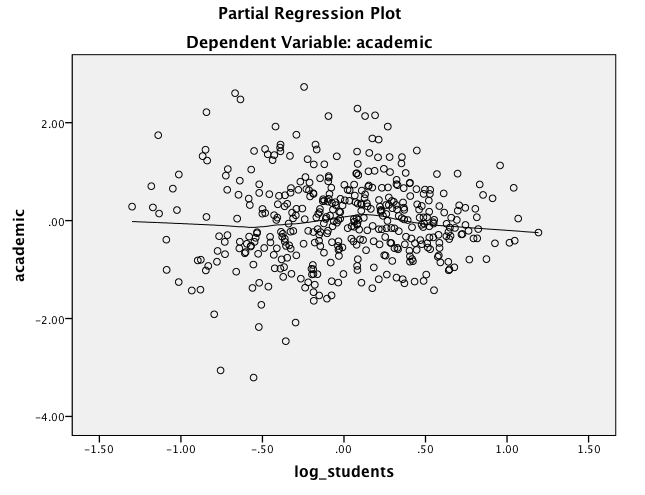


(c) outliers in residuals: There were a few outlier cases that weren't modelled well by the data, but, given the number of cases this probably isn't too big a deal.



For example, inspection of case 420 showed that it had the worse academic performance of any school. These cases could be further inspected to understand what might be going on.

(d) linearity: examination of the partial regression plots can give an indication of linearity of the relationship where there is a relationsip. Double clicking and selecting fit line - loess can help you see what if any relationship is present. Here's an example where there really is not a relationship at all. The absence of a relationship isn't a problem as much as a non-linear relationship is. If the relationship is non-linear we would want to model that non-linear effect. In this case the predictor is just weak to non-existant.



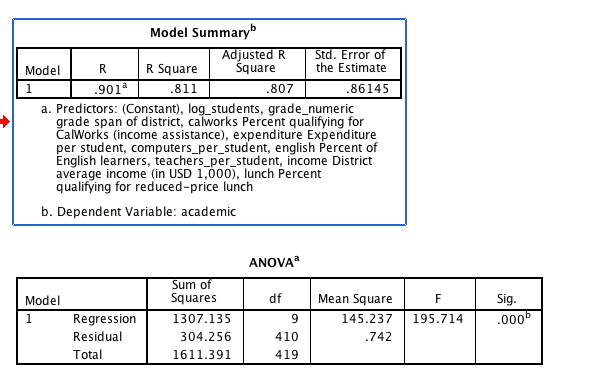
(e) collinearity diagnostics

There is clearly some correlation between predictors which may influence the interpretation of the results. In particular the lunch variable has a fairly low tolerance (i.e., .171). Thus, most of that variable is explained by other predictors. Specifically, only 17.1% of the lunch variable is unique, presumably because it overlaps with variables capturing aspects of socio-economic status. You could consider removing.

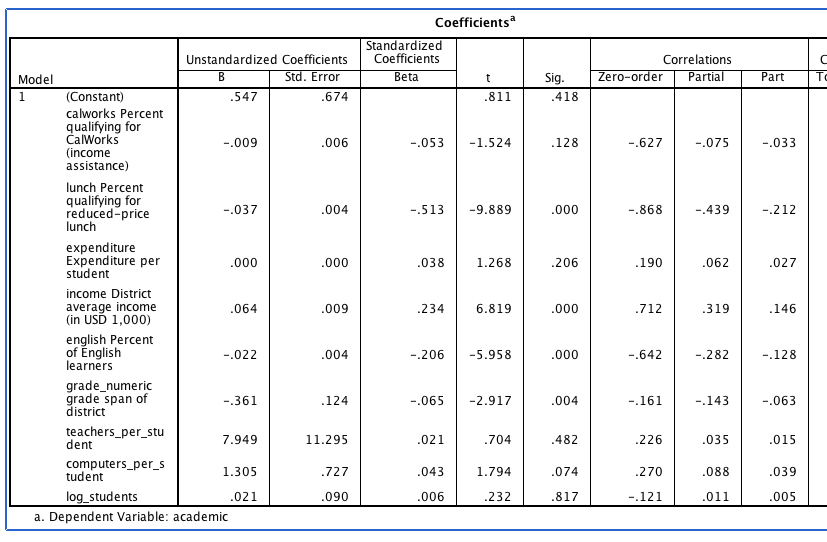
## 6. Re-run multiple regression if necessary and interpret substantive results: (a) interpret overall model fit and significance (b) interpret individual coefficients (unstandardised beta, standardised beta, semi-partial r), confidence intervals, and statistical significance.

I'll go the simple approach of retaining all predictors

(a) Overall model :The predictors explained 81.1% of variance in academic performance. This was statistically significant, F (9, 410) = 195.7, p<.001.



(b) coefficients:



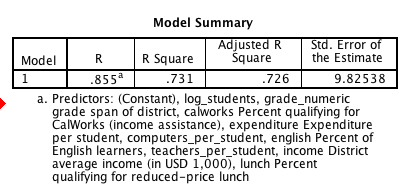
## 7. Provide an overall interpretation of the results.

In general, the best predictors of school academic performance were the various indicators of socioeconomic status. In particular, percentage qualifying for reduce price lunch was the best predictor, but higher district income, and lower proportions of English learners were also predictive. Resource variables (e.g., teacher-student ratios) were not statistically significant.

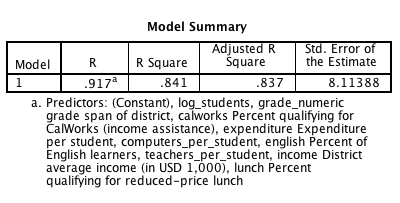
## 8. Answer the question: to what extent do school features predict mathematics performance more than reading performance?

Run the regression above with math and then reading as dependent variables

With math as DV:



with read as DV



While we have not explicitly explored how to test this for statistical significance. It is pretty clear that reading (r-square = .84) is better predicted than mathematics (r-square = .73).

It would be interesting to explore the coefficients and think theoretically about why this might be the case

## 9. Answer the question: "grades", calworks, lunch, income, english, are all baseline measures predicting academic performance; Which of following three variables "teachers per student", "computers per student" and "expenditure per student" leads to the biggest incremental prediction in academic performance over and above the baseline measures?

Re-run three regressions removing the three predictors, the use the block functionality to put for example "expenditure in the second block", you may also want to click "statistics - r-squared change" and repeat for each of the three predictors of interest

In this case each of the variables only adds about .2% of variance.