Modelli Statistici Avanzato a.a. 2017/18
Homework n. 2

The dataset `hsb.dta` is derived from the High School and Beyond Survey. The data are discussed and analyzed in Raudenbush, Bryk, Cheong and Congdon (2004) *HLM6: Hierarchical Linear and Nonlinear Modeling*. The variables used for the homework are

- `schoolid`: school identifier
- `mathach`: a measure of mathematics achievement
- `ses`: socioeconomic status based on parental education, occupation, and income
- `minority`: dummy variable for student being nonwhite

1. Keep only data on the five schools with the lowest values of `schoolid` (1224, 1288, 1296, 1308, and 1317). Also drop the variables not listed above.
2. Obtain the means and standard deviations for the continuous variables and frequency tables for the categorical variables. Also obtain the mean and standard deviation of the continuous variables for each of the five schools (using the `table` or `tabstat` command).
3. Produce a histogram and a boxplot of `mathach`.
4. Produce a scatterplot of `mathach` versus `ses`. Also produce a scatterplot for each school (using the `by()` option).
5. Treating `mathach` as the response variable y and `ses` as an explanatory variable x, consider the linear regression of y on x:
$$y_i \ = \ \beta_1 + \beta_2 x_i + \epsilon_i, \quad \epsilon_i | x_i \sim N(0, \sigma^2)$$
   a. Fil the model.
   b. Report and interpret the estimates of the three parameters of this model.
   c. Interpret the confidence interval and *p*-value for the coefficient of `ses`.

6. Using the predict command, create a new variable `yhat` equal to the predicted values of `mathach`:
$$\widehat{y_i} \ = \ \widehat{\beta}_1 + \widehat{\beta}_2 x_i$$

7. Produce a scatterplot of `mathach` versus `ses` with the regression line (`yhat` versus `ses`) superimposed. Produce the same scatterplot by school. Does it appear as if schools differ in their mean math achievement after controlling for ses?
8. Extend the regression model by including dummy variables for the schools.
   a. Fit the model with and without school dummy variables.
   b. Interpret the coefficients of the dummy variables.
   c. Test the null hypothesis that the population coefficients of all dummy variables are zero (in Stata use `testparm`).
9. Add interactions between the school dummies and `ses` and interpret the estimated coefficients (in Stata use the syntax with #)