Problem 1

Academic abilities [nels]

Problem

Find the factors that affect academic abilities and acquired knowledge of American high school students from the *Northeast* of the U.S. in selected subjects. The factors that could potentially affect academic abilities are: *family background*, *school characteristics*, *student attitudes*.

Specifications

1. The measures of academic abilities and acquired knowledge are either scores or percentiles (select one) from standardized tests in mathematics, English, and natural sciences.

   In this assignment, consider only academic abilities and acquired knowledge in *mathematics*.

2. Choose whether the response variable will be modeled on the original scale or log-transformed scale.

3. Initially, consider only students from the *urban* locations (`f2.sch.loc = Urban`) and build a suitable linear regression model. Make sure that the final regression model has a straightforward interpretation. As you build the model, pay attention to the signs and magnitudes of the regression coefficients.

4. Summarize the results of the final regression model, interpret its coefficients, run a test and calculate a confidence interval for the effect of each covariate included in the final model.

5. Which covariates contribute most to explaining the academic abilities of students?

6. Check if the assumption of equal variance is satisfied by the final model.

7. In the final model, replace the usual variance estimator by the sandwich (White) estimator, which does not assume homoscedasticity. Repeat the tests and recalculate the confidence intervals. Have the results changed? Would you expect them to change given your previous assessment of the homoscedasticity assumption?

8. Assume that the final model is suitable for all three locations (*urban*, *suburban*, *rural* as recorded in the variable `f2.sch.loc`) and evaluate (by proper statistical tests) whether the effect of each of the included covariates changes with the location of the school.

Requirements

Write a report (prepared by LaTeX, LibreOffice, MS Word, ...) summarizing your solution to the problem. Include data manipulation statements, and definitions of new variables (those used by you). Include the code for fitting the final model, calculation of tests and confidence intervals. Formulate specific answers to each of the questions asked in the Specifications paragraph.

Mail the report in the pdf format (file named as `Surname_Firstname_1.pdf`) and related R script (file named as `Surname_Firstname_1.R`) to komarek@karlin.mff.cuni.cz.

Deadline: ***Monday February 28, 2022 [06:59 CET]***.

Population

The data were collected in a large national survey of the U.S. high school students called "National Education Longitudinal Study" conducted by the U.S. Department of Education. For this assignment, only data from the high schools in the *Northeast* of the U.S. are available.

The study enrolled students who were in the 8th grade (last year of the middle school) in 1988. The data include information collected at baseline (in 1988) and during the second follow-up that was performed four years later (typically during the 2nd year of the high school).

Dataset

The dataset can be downloaded from
https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2021_22/nmst432/Problem_1/AdvRegr_1_nelsNE.RData

The dataframe is called `nelsNE`. It contains 2 312 rows (students) and 52 variables. Variable labels are included in the string vector `varlabels`.

Variable list: See Table 1. Additional information can be found in the codebook available from
https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2021_22/nmst432/Problem_1/AdvRegr_1_codebook_nels.pdf

Hints

- Consider which variables will be considered for inclusion in the regression model. Some of them are clearly useless. Even a highly significant variable may be unsuitable for inclusion in the model. Variables that contain many missing values may be problematic.
- Be careful about missing values. Some of them may be missing because the question is not applicable to the particular person. Exclude the rows containing missing values before you start building the model but check and report how many observations have been lost.
- Do not spend too much time by regression diagnostics (except those requested by the Specifications) and consideration of interactions.
- Recode factor covariates with too many levels. You can use the function `Recode()` in package `Epi` to do this efficiently.

Table 1: Variable coding table

Variable Name	Variable Label	When Collected*	On whom Collected
id	Student ID	U	Student
fam.sz	Family size	B	Family
fam.comp	Family composition	B	Family
fam.educ	Highest parent education	B	Family
fam.mar	Marrital status of parents	B	Family
f2.yng.bro	Number of younger brothers	F2	Family
f2.yng.sis	Number of younger sisters	F2	Family
f2.mo.wrk	Mother working	F2	Family
f2.mo.occup	Mother's occupation	F2	Family
f2.fa.wrk	Father working	F2	Family
f2.fa.occup	Father's occupation	F2	Family
f2.old.bro	Number of older brothers	F2	Family
f2.old.sis	Number of older sisters	F2	Family
ses.perc	Socio-economic status percentile	F2	Family
f2.sch.reg	School region (f2)	F2	School
f2.sch.loc	School location (f2)	F2	School
f2.perc.soc	Social science percentile	F2	Student
f2.sco.soc	Social science score	F2	Student
f2.perc.math	Math percentile	F2	Student
f2.sco.math	Math score	F2	Student
f2.perc.read	Reading percentile	F2	Student
f2.sco.read	Reading score	F2	Student
f2.perc.sci	Science percentile	F2	Student
f2.sco.sci	Science score	F2	Student
b.mo	Month of birth	F2	Student
b.yr	Year of birth	F2	Student
race	Race	F2	Student
f2.mo.left	Month left school	F2	Student
f2.yr.left	Year left school	F2	Student
f2.gpa	Cumulative grade average last year	F2	Student
f2.gr.comp	Average grade Comput. sci.	F2	Student
f2.gr.eng	Average grade English	F2	Student
f2.gr.lang	Average grade Foreign language	F2	Student
f2.gr.math	Average grade Mathematics	F2	Student
f2.gr.sci	Average grade science	F2	Student
f2.gr.soc	Average grade Social studies	F2	Student
f2.reas.lft	Reason left school	F2	Student
f2.moved	Times moved since 1988	F2	Student
f2.othsch	Times changed school since 1988	F2	Student
f2.menrol	Math enrollment past 2 years	F2	Student
f2.games	Hours of video games play	F2	Student
f2.tv	Hours of TV	F2	Student
f2.grade	Current grade	F2	Student
f2.teach	The teaching is good at my school	F2	Student
f2.disc	Discipline is fair in my school	F2	Student
f2.cig	Cigarettes per day	F2	Student
f2.alco	Times alcohol last month	F2	Student
f2.drunk	Times drunk last 2 weeks	F2	Student
f2.weed	Times marihuana last month	F2	Student
f2.skip	Times skip classes	F2	Student
f2.arrest	Times arrested	F2	Student
sex	Gender	F2	Student

* U = universal, B = baseline, F2 = 2nd follow-up