Data Exploration and Analysis

Analyze a data set containing 13 attributes of mother-baby pairs where the child has been diagnosed with the PKU inherited disorder. (Information on PKU is below). Explore the data for possible relationships between the predictor variables and the cognitive development of the 1000 babies at age one. There are 1000 observations of mother-baby pairs in this dataset.

Analyze the data set and write a report to summarize your conclusions about the relationship of the predictor variables to the response variables. (The data set variables are described below). Your report should include plots that convey much of the information. The text you write for the report should be brief - let your plots convey much of your findings. You can paste relevant outputs (dataset and model summaries) into an appendix and refer to them so support your conclusions. You may turn in your report in Word, pdf, or html format. R permits you to generate these formats using R markdown to create an .Rmd file. If you wish you may generate your Word, pdf, or html file using R markdown. In any case please turn in both your executable script and your report.

Your analysis should include these approaches:

1. Read the data into an R data frame and detect whether or not the data needs pre-processing to deal with missing attributes. Check that the data types are reasonable.
2. Plot a histogram or density distribution of each variable.
3. Check the variables for correlations by using **cor(df)** where **df** is the name of your data frame.
4. Use **lm** to fit a regression model to calculate the response variable **MDI** as a function of the predictor variable **phe_avg.** Perform a summary of the model by entering **summary(fit1)** where **fit1** is the linear model you fit. In your report, discuss the summary of the linear model: for the significant coefficients, state what one unit increase in each predictor variable will result in for the response variable. Also discuss the **adjusted R^2**. Execute **plot(fit1,1)** as part of your script. In class we will discuss the summary of a fitted model and the use of the plot command
5. Use **lm** to fit a regression model to calculate the response variable **PDI** as a function of the predictor variable **phe_avg**. Follow the steps in #4 above and include them in your report.
6. Fit two other regression models that will model a response variable (see below) as a function of predictor variables. Try to achieve a high **adjusted R^2**, indicating goodness of fit of your model to the data. Follow the steps in #4 above and include them in your report.

PKU

PKU (phenylketonuria) is a well-documented genetic disorder where the body is unable to metabolize the amino acid phenylalanine due to the lack of the enzyme phenylalanine hydroxylase, coded for by the PAH gene. If not metabolized, phenylalanine can build up in the blood and brain, affecting brain development and function. It can lead to cognitive disability and seizures.

A person with PKU may prevent cognitive damage by strictly avoiding foods containing phenylalanine. It is therefore vitally important to detect the disorder before any damage is done. For this reason, newborn screening programs are used to detect PKU very early so the child can follow a strict diet and avoid the cognitive effects of the disease. PKU is recessive, so it takes 2 mutated alleles of the gene (one from each parent) in order for the individual to be affected. People who have only one mutated allele are carriers, but do not themselves have the disease. Most carriers are unaware that they carry the mutated gene and can pass it on to their offspring.

The PKU mother-baby data set:

Predictor variables:

- Mother's age (y)
- Mother's Verbal IQ
- Mother's Performance IQ
- Mother's Full-scale IQ
- Weight gain (lb)
- Variability (SD) of Phe (umol/L)
- Average Phe exposure (umol/L)
- Gestation at term(weeks)
- Birth length (cm)
- Birth head circumference (cm)

Response variables:

- Birth weight (g)
- Bayley (Mental Development Index) **MDI** at age 1
- Bayley (Psychomotor Development Index) **PDI** at age 1