Brain.sas: Explanation of code

Goals of code:
Focus on diagnostics for linear regression models using the brain dataset in SAS. The first few lines read the data and compute the log transform of each variable.

Remember: if you use proc import to read a .xlsx or .csv file, you need a separate data step, with a set command, to compute new variables.

Regression diagnostics: `\r vif` and `output`
If you want to see the actual numbers for regression diagnostics, adding the `/r` option to the proc reg prints out diagnostics for each observation.

The VIF (variance inflation factor) diagnostic is characteristic of each variable (not each observation). To print that out, add `/vif` to the proc reg statement.

Note: If you have two options, e.g. `/r` and `/vif`, you only need one `/` to separate the model piece from the options. This is illustrated in the example code in the last proc reg.

If you want to store the diagnostics in a data set, which can then be printed or plotted, use `output` with the appropriate keyword. The example code gives keywords for the diagnostics we'll use in class and a few more that I'll briefly mention.

Many diagnostics, e.g., Cook's Distance, are best plotted against the observation number. This gives you a quick indication of the largest value (i.e., the answer to the question, do I need to care about influential observations?) and help you identify which point is the one (or more) with large D values.

Storing the observation number in a data set: `i = _n_;`
The easiest way to produce a full page plot using the observation number is to create a new variable (I use i) containing the observation number. This is done in the data step that creates brain. Inside a data step, SAS creates various "internal" variables. These have values that are defined inside the data step but are not saved, so they are not accessible to procs. When SAS creates a variable, it usually has `_` before and after, so the created variable is not confused with a variable name you want. One of those "internal" variables is `_N_`. This contains the observation number. To save that information, copy it to a variable you name (e.g. `i`), which will then be stored and can used in subsequent procs.

Ways for diagnostics:

- Extracting Predicted Values and Residuals: `output out=resids p=predicted r=residual`
- Requesting Standardized Residuals: `student=isresid rstudent=esresid`
- Calculating Cook's D: `cookd=d`

- Calculating Variance Inflation Factor (VIF): `/vif`

Identifying interesting observations:

- Outliers: data points that significantly differ from the majority of the data.

  Identification Methods:

  Look at the standardized residuals

- Influential points: observations that, if removed, would significantly change the regression model's results.

  Identification Methods:

  Cook's Distance: Measures the influence of each observation. Points with Cook's distance greater than a certain threshold (like 4/n or 1) are considered highly influential.

- Residual Analysis: to check the assumption of homoscedasticity (equal variances) and to identify outliers or unusual patterns.

  Methods:

  Standardized Residuals: Observations with usually large or unusually small standardized residuals should be investigated. If the model fits, 95% of the standardized residuals should be between -2 and 2 and very very few should be less than -4 or larger than 4.

- Multicollinearity Assessment: to identify predictor variables that are highly correlated with each other.

  Methods:

  Variance Inflation Factor (VIF): A VIF greater than 5-10 suggests significant multicollinearity.