**Example Dataset:**

This dataset was collected as part of a student thesis project – and is one of the messiest datasets our lab has worked with ☺. We asked students about their advisors and their stress levels by using lots of scales. I have included the following variables:

* Gender – male / female gender variable.
* Level – factor variable indicating the level of the student (freshman – Ph.D. student)
* Advising – a yes / no variable that indicated if they had an assigned advisor or not, should import as an integer.
* Q1-Q23 – one of the scales we used to assess advising, which was coded 1-7.

**Helpful hint:**

* At some point, you will mess up the dataset. Therefore, I am going to suggest you create new tabs for each step, which means that if you mess up, you can just back up in your code to the last save point. Think about it like a video game with multiple save spots.

**Accuracy**:

* Fix any mislabeled categorical columns.
* Fix any out of range scores.

**Missing data:**

* Check for missing data by participant and eliminate the participants with more than 5%.
* Then check the columns/variables for more than 5% missing.
* Fill in only the columns with less than or equal to 5% missing.

**Outliers**

* Look at z-scores for each column or the total score for scales.
* Take out the outliers if it makes sense.

**Additivity:**

* Create a correlation table, and look for things that are over .90 for problems.

Here’s an example of part of one:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *q1* | *q2* | *q3* |
| q1 | 1 |  |  |
| q2 | 0.620744266 | 1 |  |
| q3 | 0.798862805 | 0.676772681 | 1 |
| q4 | 0.731674231 | 0.625868033 | 0.763395566 |
| q5 | 0.533057156 | 0.491487027 | 0.575065929 |

**The rest of the assumptions set up:**

* For ANOVA, t-tests, correlation: you will use a *fake* regression analyses – it’s considered fake because it’s not the real analysis, just a way to get the information you need to do data screening.
* For regression based tests: you can run the *real* regression analysis to get the same information. The rules are altered slightly, so make sure you make notes in the regression section on what’s different.
* Create a set of random numbers to compare against.
* Run a fake regression.

**Normality:**

* Check out the multivariate histogram. Note, in later videos, I change the way I do this step a little bit to make it easier.

**Linearity:**

* Check out the normal probability plot. You can add a trendline to help you see if it’s on the line.

**Homogeneity/Homoscedasticity:**

* Check out the scatterplot of the residuals.
* For homogeneity:
  + Are the dots roughly centered around zero horizontally and vertically? Use the lines on the graph to help you tell.
  + Don’t go too nuts – one or two random dots does not constitute a problem.
* For homoscedasticity:
  + Are the dots roughly blob shaped? Imagine a line drawn around the dots – it should make blob shape and not a megaphone or UFO.

**Results**

Prior to analysis, holistic, content, structure, stance, sentence fluency, diction and conventions within English papers were examined through Excel for accuracy of data entry, missing values, and fit between their distributions and the assumptions of multivariate analysis. The variables were examined for the 368 participants in the study. All categorical variable labels were examined for typos, and several out of range values for questions on the MBI (1-7) were excluded.

97 participants were excluded for having missing data over 5%, and after excluding these participants, all variables were checked for missing data. Each column had less than 5% missing. These missing values were replaced with mean substitution in Excel. Four outliers were found examining the *z*-scores for total scores on the MBI, using a cut off criterion of *z* = 3.00, *p* < .001. These outliers were deleted, leaving 267 cases. Bivariate correlations were used to check for multicollinearity and singularity, and all values indicated additivity. The multivariate normality plot showed that results were normal. The normal pp plot of regression standardized residuals shows that the variables were linear. The standardized regression scatterplot showed that the results were homogeneic and homoscedastic.