Capstone 1: Statistical Data Analysis

Significant variables

Some of the variables that were first assumed to be significant either were found to be not, or were too sparsely represented to be adequately tested. The most significant variables are age group (child, adolescent, or adult), known jaundice at birth (this is a little problematic, because it is not always known, but the default assumption is 'no'), the relationship of the assessor to the assessee, and whether or not there is a family history of autism.

Significant differences between subgroups in your data

Although the assessment data was originally divided into three age categories of people (child, adolescent, and adult), I further divided the data into much smaller age groups and saw that there was a lot of variation of results within the 'adult' category.

Jaundice displays a gender-dependent relationship with assessment results, so these two subsets were teased apart for individual analysis.

Correlations between pairs of variables

Strong correlations exist between the assessment result (dependent) and many independent variables. The largest correlation exists between age and result - adults having a significantly lower result than under 18s. Other independent variables that show a correlation to result are: whether the assessor was the parent or the person themselves, family history of autism, known jaundice at birth, and between adults and adolescents, as well as adults and children (although not a large difference between children and adolescents). No correlation was found between result and gender, or any meaningful correlation for ethnicity or country of residence.

Although there was quite a positive correlation between self-assessment and assessment by a healthcare professional, the number of data available from healthcare professionals was too small to be statistically confident of the correlation.

Inferential statistical tests

To analyze and look for potential statistical relationships between the dependent variable and multiple independent variables, I used an independent (aka two sample) t-test of the population means, without the assumption of equal variance. This inferential test is appropriate because it requires one continuous dependent variable (in this case the resultant score from the assessment) and one two-level, categorical, independent variable (e.g. gender, age group, jaundice at birth). The null assumption of an independent t-test is that the means of both populations are equal. Usually, this test is used with an aim to reject the null hypothesis and be able to say that the two populations do not share the same mean.