Foundations of Medical Statistics

Analytical Techniques & Regression Assignment

Instructions for submission

You must hand in your assignment by **2pm** on **Tuesday 13th December 2022**. You should submit your assignment electronically via the submission point on the Foundations of Medical Statistics page on Moodle.

Background

Huntington's disease (HD) is a neurodegenerative disease caused by a genetic factor, specifically a CAG repeat expansion in *HTT*, the gene that encodes huntingtin, which is on chromosome 4. HD is most common in Europe, North American, and Australia, where the prevalence is approximately 6 per 100,000. HD is characterised by progressive motor dysfunction, cognitive decline and psychiatric disturbance. The mean age of clinical onset is around 45, and most people with the CAG repeat expansion seem healthy until adulthood. The length of the CAG repeat accounts for around 60% of the variability in age at clinical onset, with individuals with longer repeat lengths typically having earlier onset than those with shorter repeat lengths.

In this assignment you will explore differences in performance on a cognitive test between controls without HD, and people who have a CAG repeat expansion in *HTT* but who, as yet, do not have a diagnosis of clinical onset of the disease. This latter group are termed "pre-manifest HD gene carriers".

The cognitive test considered here is the symbol digit modalities test (SDMT). The test involves pairing abstract symbols with specific numbers. The test requires elements of attention, visuoperceptual processing, working memory, and psychomotor speed. The higher the score the better the performance, with maximum possible score of 110.

Data

The data here are from 120 pre-manifest HD gene carriers and 123 controls without HD. The Stata file **ATREG_assign2022.dta** (available from the Foundations Moodle page) contains the following variables:

Variable	*Description and coding*
id	Participant identifier
group	Disease status (2 = pre-manifest HD gene carrier, 1 = control)
age	Age (years)
cag	CAG repeat length (not recorded in controls)
sdmt	SDMT score

Statistical Analysis

Use Stata or R to help carry out the following tasks.

1) Exploratory analysis

a) Use appropriate numerical summaries and graphical techniques to explore (separately in each group) the distributions of SDMT score, age and CAG repeat length (HD gene carriers only).

b) Construct a table reporting appropriate key summary statistics for each group.

2) Comparison of SDMT scores between groups

a) Use appropriate statistical techniques to compare the mean and the standard deviation of SDMT scores between the pre-manifest HD gene carriers and controls. Interpret your findings.

b) Examine the correlation between age and SDMT scores in each of the two groups. By considering the distributions of age in the two groups and these correlations, explain why the comparison of the means in 2a) might be biased. Predict how you might expect the estimated difference in mean SDMT scores between the groups to change were you to adjust for age.

c) Use an appropriate linear regression (ANCOVA) model to compare SDMT scores in the two groups adjusting for age. Interpret your findings. Are the results in agreement with your prediction in 2b)?

d) Construct appropriate plots and carry out any other statistical procedures that you think are appropriate to investigate whether the assumptions underlying the model in 2c) are violated and whether the results can be considered robust.

3) Dependency of SDMT scores on age and CAG repeat length in pre-manifest HD gene carriers

a) In the HD gene carriers alone fit linear regression models relating SDMT scores to (i) age alone, (ii) CAG repeat length alone, and (iii) age and CAG repeat length.

b) Use appropriate statistical techniques to explore the association between age and CAG repeat length in the HD gene carriers. Use these results to help inform the interpretation of the results obtained in 3a).

c) For the each of the models in 3a) construct appropriate plots and carry out any other statistical procedures that you think are appropriate to investigate whether the assumptions underlying the models are violated and whether the results can be considered robust.

4) Comparison of SDMT scores between subgroups

a) Create a new variable that takes the values: 1 for healthy controls; 2 for HD gene carrier with a CAG repeat length of ≤ 42; 3 for HD gene carrier with a CAG repeat length of ≥ 43. Explore the distributions of age, SDMT score and CAG repeat length in the three subgroups.

b) Use an appropriate linear regression (one-way ANOVA) model to compare mean SDMT scores in the three groups formed in 4a).

c) Use an appropriate linear regression (ANCOVA) model to compare mean SDMT scores in the three groups adjusting for age.

d) Construct appropriate plots and carry out any other statistical procedures that you think are appropriate to investigate whether the assumptions underlying the models in 4b) & 4c) are violated and whether the results can be considered robust.

e) Construct an appropriate graph to show how the fitted value for SDMT score from the model in 4c) relates to age and subgroup. On this graph also show the raw data on SDMT scores by subgroup.

Report

The report should be **no more than 4 pages long** including tables and figures, but not including the 1-page appendix detailed below. Reports longer than 4 pages will be penalized by 1 grade. The font size for the text of the report should not be smaller than Times New Roman 11 point (you do not have to use Times New Roman) and normal spaced margins should be used. Font size for tables and figures can be smaller but these should be appropriately sized and clearly readable.

The report should be in three sections:

1) **Statistical Methods:** This should be a brief description of the statistical techniques that you have used for this assignment. The description should be concise, but include enough detail for a competent data analyst to repeat your analysis given the data. Unless you have used any non-standard techniques a description of the techniques in words, without algebra, is sufficient. The suggested length is around half of a page. **[15% of marks]**

2) **Results:** The results section should include the summary table referred to in 1b) and one or more table(s) presenting the **results** obtained from 2), 3) and 4) above. You should also present the figure referred to in 4e). You do not need to report in detail all the analyses that you have carried out, but you should describe the relevant results from parts 1-4. Be precise in your reporting, especially in the headings and labels of tables and figures. In your description of the results there is no need to define or explain the general meaning of statistical terms such as '*p*-value'. Do not include any non-graphical Stata output i.e. results copied directly from the Stata screen. The suggested length is around 3 pages. **[65% of marks]**

3) **Discussion:** Summarise your findings and discuss any limitations of this study and your analysis. Do not include any review of the literature. The suggested length is around half of a page. **[10% of marks]**

Appendix

On a separate page write down the algebraic expression for the model specified in 4c). Carefully define each term in the expression and give the interpretation of each of the population parameters (i.e. α, $\beta 1$, $\beta 2$, etc.). This should be included as an appendix to the report on a 5^{th} page i.e. please submit the whole assignment as a single document. The algebraic expression can be written by hand or using an equation editor. **[10% of marks]**