HW Question Week 6

STA2101F 2021

Due October 28 2021 11.59 pm

Homework to be submitted through Quercus

This question is based on the article "The impact of a lack of mathematical education on brain development and future attainment" by Zacharopoulos, et al.. The article and supplementary appendix are posted on the course web page. The authors ran two experiments (see *Materials and Methods* on p.6, 1st paragraph), but we will focus on the first experiment only, which the authors also call "the A-level cohort".

(a) The *Materials and Methods* section describes the authors' dependent variable, let's call it y: what is this and how was it coded? How many students were included in Experiment 1? How many had $y = 1$ and how many had $y = 0$?

(b) On p.2 we read "Based on the existing literature, we hypothesized that the lack of mathematical education would be associated with reduced GABA and/or increased glutamate." I think both GABA and glutamate were measured in two different brain regions, MFG and IPS, so there were four potential explanatory variables of interest. Figure 2D shows the fitted values for a model that used MFG-GABA as the explanatory variable. Write out an equation and R pseudo-code for the model that was used to obtain these fitted values. (It's described in the second paragraph of the Results section.)

(c) Figures 2A and 2B compare the scores on "a numerical operation attainment test", and a "mathematical reasoning attainment test" in the "math" and "non-math" groups. In the text we read (Results par.1), for Figure 2A, $t(84) = -5.27, p < 0.001$. How was $t(84)$ computed? What do the error bars on the boxplots indicate? How are these error bars related to the comparison between the two groups?

(d) In the "Statistical Analyses" subsection of the "Materials and Methods" section, (p.7), the authors mention Levene's test. What hypotheses were they testing, and why?

(e) In the "Results" section on p.3 (left), they discuss potential confounding. What confounding variables did they consider? Write out an equation and R pseudo-code for one of the models described in this paragraph.

(f) There are other subsections in the Results section – "Dissociating ..." refers to Experiment 2, and "MFG GABA Predicts..." refers to a follow-up analysis that they

did to relate scores on the mathematical reasoning test taken "19 months later" to MFG-GABA and some other covariates. In this paragraph the degrees of freedom for the t statistics seem to be 33, instead of 84 as above. Why is it so much smaller?

(g) *Bonus/PhD SM, Exercise 10.4.1*: Suppose $y_1, \ldots, y_n$ follow a binary logistic model in which y_i takes value 1 with probability $p_i = p_i(\beta) = \exp(x_i^T\beta)/\{1 + \exp(x_i^T\beta)\}$ and value 0 otherwise, for $i = 1, \ldots, n$. (i) Show that the maximum likelihood estimate of β, $\hat{\beta}$, satisfies $X^T y = X^T\hat{p}$, where $\hat{p}_i = p_i(\hat{\beta})$. (ii) The *deviance* is defined to be twice the difference in the maximized log-likelihood functions computed under the model $p_i = p_i(\beta) = \exp(x_i^T\beta)/\{1 + \exp(x_i^T\beta)\}$, i.e. $\ell(\hat{\beta})$, and the maximized log-likelihood function in a so-called *saturated* model in which p_i is not linked to x_i. In this saturated model the maximum likelihood estimate is $\tilde{p}_i = y_i$. Show that this means that the deviance depends entirely on the vector $p(\hat{\beta})$. (iii) If $p_1 = \cdots = p_n = p$, then show that $\hat{p} = \bar{y}$, and verify that the deviance is $-2n\{\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})\}$.