Project 0

August 26, 2024

You are working with an investigator who is interested in modelling the association between a set of predictors and a set of outcomes. Specifically, the investigator obtained measures of resting state functional connectivity (the predictors) which they wish to correlate with a set of measures of physical function using a set of (separate) linear regression models (no regularization). Suppose that the investigator collected P measures of functional connectivity and M measures of physical function on $i = 1, \dots, N$ participants, with $N > P > M$.

The investigator is interested in identifying which measures of resting state functional connectivity are associated with individual measures of physical function, after adjusting for the other resting state connectivity values. Further, the investigator is concerned both with family-wise type-I error rate (the probability of making at least one false discovery) and power (the probability of rejecting the null hypothesis when the alternative is true). Let $\boldsymbol{y}^m = [y_1^m, \dots, y_N^m]^t \in \mathbb{R}^N$ denote the response vector for the m^{th} measure of physical function, $\boldsymbol{x}^p = [x_1^p, \dots, x_N^p]^t \in \mathbb{R}^N$ denote the vector containing the p^{th} functional connectivity measure, and $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_P]$ be the $N \times (P+1)$ design matrix associated with all functional connectivity measures (note the column $\mathbf{1}$ corresponding to the "intercept" term in the design matrix). You may assume that $\boldsymbol{X}$ is full column rank. You may assume that the residuals from linear regression are normally distributed and that residuals across persons are uncorrelated. The investigator's data are contained in the data file *fmri_phys_func.csv*.

(a) How many regression models is the investigator asking you to fit? Write out the models using general notation. State all relevant assumptions.

(b) How many parameters are you estimating in total? Briefly explain your answer.

(c) How many parameters is the investigator interested in? Briefly explain your answer.

(d) Fit the regressions requested by the investigator to the data. Report estimates, confidence intervals, and p-values in figure/table format. **Concisely** summarize the findings relevant to the investigator. You may use more than one figure/table to present findings, but it should be easy for readers to connect your summary in words to the supporting tables/figures.

(e) The investigator wants to test each of the regression coefficients individually to identify measures of physical function while preserving a family-wise type-I error rate of $\alpha = 0.05$. The Bonferroni correction (`https://en.wikipedia.org/wiki/Bonferroni_correction`) is one such approach. Report the Bonferroni corrected quantities in a table format. Comment on difference from what you observed in (c).

(f) The Bonferroni correction is, in general, overly conservative. Given the investigator's interest in both preserving type-I error rate and maximal power, they ask you to come up with an alternative approach. Here we will attempt to find a more powerful approach.

   i. Using the model assumptions, identify the distribution of the regression coefficients for each of the separate models, $\hat{\boldsymbol{\beta}}^m$.

   ii. One approach to addressing this problem is to identify the distribution of the random vector $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}^{1^t}, \dots, \hat{\boldsymbol{\beta}}^{M^t}]^t$ (the joint distribution of the estimated regression coefficients). Is $\hat{\boldsymbol{\beta}}$ multivariate normal? If so, explain why. If not, identify one or more assumptions that would result in $\hat{\boldsymbol{\beta}}$ being multivariate normal. Hint: Think about what quantities in linear regression have distributional assumptions.

   iii. Using the assumptions from iii., find the parameters that define the multivariate normal distribution for $\hat{\boldsymbol{\beta}}$.