1

Statistical Preliminary

This chapter covers the basic statistical methods that are mostly used in univariate voxel-level approaches. However, these basic methods are equally useful in brain network analysis as well. Most of network modeling techniques are based on the voxel-level methods. Readers familiar with univariate statistical methods can skip this chapter.

1.1 General Linear Models

General linear models (GLM) have been widely used in brain imaging and network studies. The GLM is a very flexible and general statistical framework encompassing a wide variety of fixed-effect models such as multiple regressions, the analysis of variance (ANOVA), the multivariate analysis of variance (MANOVA), the analysis of covariance (ANCOVA), and the multivariate analysis of covariance (MANCOVA) (Timm and Mieczkowski, 1997). More complex multilevel or hierarchical models such as the mixed-effects models and structural equation models (SEM) are also viewed as special cases of general linear models.

GLM provides a framework for testing various associations and hypotheses while accounting for nuisance covariates in the model in a straightforward fashion. The effect of age, sex, brain size, and possibly IQ may have severe confounding effects on the final outcome of many brain network studies. Older populations' reduced functional activation could be the consequence of age-related atrophy of neural systems (Mather et al., 2004). Brain volumes are significantly larger for children with autism 12 years old and younger compared with normally developing children (Aylward et al., 1999). Therefore, it is desirable to account for various confounding factors such as age and sex. This can be done using GLM automatically. The parameters of GLM are

mainly estimated by the least squares estimation and have been implemented in many statistical packages such as R[1] (Pinehiro and Bates, 2002), statistical parametric mapping (SPM)[2] and fMRI-STAT.[3]

We assume there are n subjects. Let y_i be the response variable at a node or edge, which is mainly coming from images and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$ to be the variables of interest and $\mathbf{z}_i = (z_{i1}, \cdots, z_{ik})$ to be nuisance variables corresponding to the ith subject. Then we have GLM

$$y_i = \mathbf{z}_i \boldsymbol{\lambda} + \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)^\top$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ are unknown parameter vectors to be estimated. We assume ϵ to be the usual zero mean Gaussian noise.

The significance of the variable of interests $\mathbf{x}_i$ is determined by testing the null hypothesis

$$H_0 : \boldsymbol{\beta} = 0 \text{ vs. } H_1 : \boldsymbol{\beta} \neq 0.$$

The fit of the reduced model corresponding to $\beta = 0$, i.e.,

$$y_i = \mathbf{z}_i \boldsymbol{\lambda}, \tag{1.1}$$

is measured by the sum of the squared errors (SSE):

$$\text{SSE}_0 = \sum_{i=1}^{n} (y_i - \mathbf{z}_i \widehat{\boldsymbol{\lambda}}_0)^2,$$

where $\widehat{\boldsymbol{\lambda}}_0$ is the least squares estimation obtained from the reduced model. The reduced model (1.1) can be written in a matrix form

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} z_{11} & \cdots & z_{1k} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nk} \end{pmatrix}}_{\mathbf{Z}} \underbrace{\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}}_{\boldsymbol{\lambda}}.$$

By multiplying $\mathbf{Z}^\top$ on the both sides, we obtain

$$\mathbf{Z}^\top \mathbf{y} = \mathbf{Z}^\top \mathbf{Z} \boldsymbol{\lambda}.$$

Now the matrix $\mathbf{Z}^\top \mathbf{Z}$ is a full rank and can be invertible if $n \geq k$, i.e., there are more subjects than the number of parameters. The matrix equation then can be solved by performing a matrix inversion

$$\widehat{\boldsymbol{\lambda}}_0 = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}.$$

[1] www.r-project.org
[2] www.fil.ion.ucl.ac.uk/spm
[3] www.math.mcgill.ca/keith/fmristat

Similarly the fit of the full model corresponding to $\boldsymbol{\beta} \neq 0$, i.e.,

$$y_i = \mathbf{z}_i \boldsymbol{\lambda} + \mathbf{x}_i \boldsymbol{\beta}$$

is measured by

$$\text{SSE}_1 = \sum_{i=1}^{n} (y_i - \mathbf{z}_i \widehat{\boldsymbol{\lambda}}_1 - \mathbf{x}_i \widehat{\boldsymbol{\beta}}_1)^2,$$

where $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\beta}}_1$ are the least squares estimation from the full model. The full model can be written in a matrix form by concatenating the row vectors $\mathbf{z}_i$ and $\mathbf{x}_i$ into a larger row vector $(\mathbf{z}_i, \mathbf{x}_i)$, and the column vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ into a larger column vector $(\boldsymbol{\lambda}^\top, \boldsymbol{\beta}^\top)^\top$, i.e.,

$$y_i = (\mathbf{z}_i, \mathbf{x}_i) \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\beta} \end{pmatrix}.$$

Then the parameters of the full model can be estimated in the least squares fashion. Note that

$$\begin{aligned} \text{SSE}_1 &= \min_{\boldsymbol{\lambda}_1, \boldsymbol{\beta}_1} \sum_{i=1}^{n} (y_i - \mathbf{z}_i \boldsymbol{\lambda}_1 - \mathbf{x}_i \boldsymbol{\beta}_1)^2 \\ &\leq \min_{\boldsymbol{\lambda}_0} \sum_{i=1}^{n} (y_i - \mathbf{z}_i \boldsymbol{\lambda}_0)^2 = \text{SSE}_0. \end{aligned}$$

So the larger the value of $\text{SSE}_0 - \text{SSE}_1$, more significant the contribution of the coefficients $\boldsymbol{\beta}$ is. Under the assumption of the null hypothesis H_0, the test statistic is the ratio

$$F = \frac{(\text{SSE}_0 - \text{SSE}_1)/p}{\text{SSE}_0/(n-p-k)} \sim F_{p,n-p-k}. \tag{1.2}$$

The larger the F value, it is more unlikely to accept H_0.

1.1.1 *T*-Statistic

When $p = 1$, the test statistic F is distributed as $F_{1,n-1-k}$, which is the square of the student t-distribution with $n-1-k$ degrees of freedom, i.e., t^2_{n-1-k}. In this case, it is better to use t-statistic. The advantage of using the t-statistic is that the test statistic can provide the direction of the group difference that the F-statistic cannot provide.

Let

$$c = (\underbrace{0, \cdots, 0}_{k}, 1, \underbrace{0, \cdots, 0}_{p-1})^\top$$

be the contrast vector of size $k + p$. The incorporation of the contrast vector makes the algebraic derivation straightforward. Consider testing the significance of $H_0 : \beta_1 = 0$. The least squares estimation of β_1 can be written as

$$\widehat{\beta}_1 = c \begin{pmatrix} \widehat{\boldsymbol{\lambda}} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix}.$$

Under the assumption $\epsilon_i \sim N(0, \sigma^2)$,

$$\mathbb{E}\widehat{\beta}_1 = \beta_1.$$

Further, the variance

$$\mathbb{V}\widehat{\beta}_1 = c\mathbb{V}\begin{pmatrix} \widehat{\boldsymbol{\lambda}} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} c^\top = \sigma^2 c^\top \left([\mathbf{ZX}]^\top \mathbf{ZX}\right)^{-1} c.$$

Thus, the unbiased estimator of σ^2 is given by

$$\mathrm{SSE}_1/(n-1-k).$$

We plug this estimator into σ^2. Then the test statistic under the null hypothesis is

$$T = \frac{\widehat{\beta}_1}{\sqrt{\mathbb{V}\widehat{\beta}_1}} \sim t_{n-1-k}.$$

1.1.2 R-Square

The R-square of a model explains the proportion of variability in measurement that is accounted by the model. Sometime R-square is called the coefficient of determination and it is given as the square of a correlation coefficient for a very simple model. For a linear model involving the response variable y_i, the total sum of squares (SST) measures total total variation in response y_i and is defined as

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

where $\bar{y}$ is the sample mean of y_i.

On the other hand, SSE measures the amount of variability in y_i that is not explained by the model. Note that SSE is the minimum of the sum of squared residual of any linear model, SSE is always smaller than SST. Therefore, the amount of variability explained by the model is SST-SSE. The proportion of variability explained by the model is then

$$R^2 = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}},$$

which is the coefficient of determination. The R-square ranges between 0 and 1 and the value larger than 0.5 is usually considered significant.

1.1.3 Sum of *T*-Statistics

Often there is a situation such as a meta-analysis, where we have to sum the t-statistic images or networks (Chung et al., 2017b). Note that a t-statistic for large degrees of freedom (above 30) is very close to standard normal, i.e., $N(0,1)$. For n identically distributed possibly dependent t-statistics $t^1, \cdots, t^n$, the variance of sum $\sum_{j=1}^n t^j$ is approximately given by (Billingsley, 1995)

$$\mathbb{V}\left(\sum_{j=1}^{n} t^j\right) \approx n + \sum_{i \neq j} \mathbb{E}(t^i t^j),$$

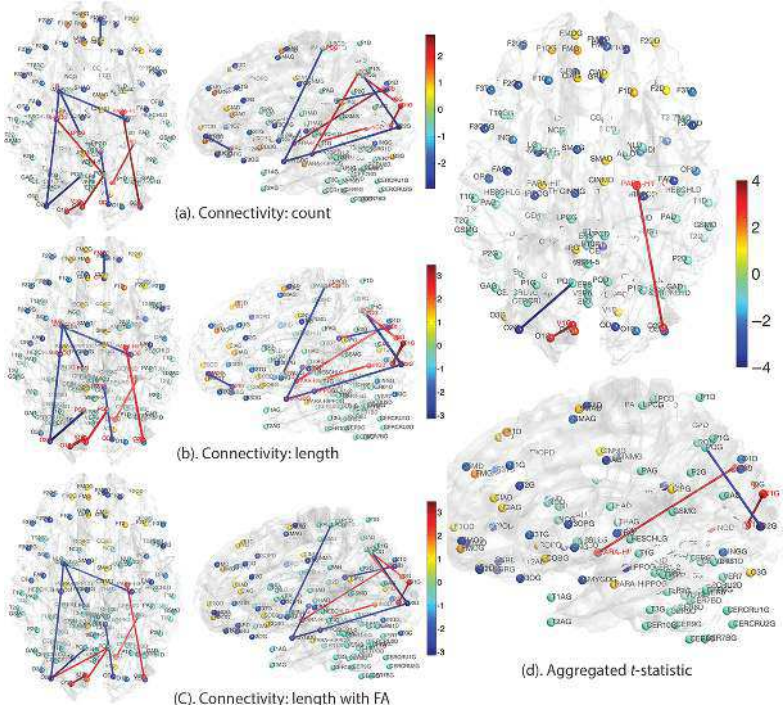

Figure 1.1 (a)–(c) t-statistic results of group difference between maltreated children and normal controls for three different connectivity methods (Chung et al., 2017b). Only the connections at the p-value less than 0.01 (uncorrected) are shown. (d) The three t-statistic maps are aggregated to form a single t-statistic.

where $\mathbb{E}(t^i t^j)$ is the correlation between t^i and t^j. We used the fact $\mathbb{E}t^j = 0$. Then, we have the *aggregated t-statistic* given by

$$T = \frac{\sum_{j=1}^{n} t^j}{\sqrt{n + \sum_{i \neq j} \mathbb{E}(t^i t^j)}} \sim N(0,1).$$

If the statistics t^j are all independent, since t^j are close to standard normal, $\mathbb{E}(t^i t^j) \approx 0$. The dependency increases the variance estimate and reduces the aggregated t-statistic value. Unfortunately, it is difficult to estimate the correlations directly since only one t-statistic map is available for each t^j. $\mathbb{E}(t^i t^j)$ can be empirically estimated by computing correlations over the entries of t-statistic maps t^i and t^j (see Figure 1.1).

1.2 Logistic Regression

Logistic regression is useful for setting up a probabilistic model on the strength of connectivity and performing classification (Subasi and Ercelebi, 2005). Suppose k regressors $X_1, \cdots, X_k$ are given. These are both imaging and nonimaging biomarkers such as gender, age, education level, and memory test score. Let $x_{i1}, \cdots, x_{ik}$ denote the measurements for the ith subject. Let the response variable Y_i be the probability of connection modeled as a Bernoulli random variable with parameter π_i, i.e.,

$$Y_i \sim \text{Bernoulli}(\pi_i).$$

$Y_i = 0, 1$ indicates the edge connected (assigned number 1) or disconnected (assigned number 0) respectively. π_i is then the likelihood (probability) of the edge connected, i.e., $\pi_i = P(Y_i = 1)$.

Now consider linear model

$$Y_i = \mathbf{x}_i^\top \beta + \epsilon_i, \tag{1.3}$$

where $\mathbf{x}_i^\top = (1, x_{i1}, \cdots, x_{ik})$ and $\beta^\top = (\beta_0, \cdots, \beta_k)$. We may assume

$$\mathbb{E}\epsilon_i = 0, \quad \mathbb{V}\epsilon_j = \sigma^2.$$

However, linear model (1.3) is no longer appropriate since

$$\mathbb{E}Y_j = \pi_i = \mathbf{x}_i^\top \beta$$

but $\mathbf{x}_i^\top \beta$ may not be in the range $[0, 1]$. The inconsistency is caused by trying to match continuous variables x_{ij} to categorical variable Y_i directly. To address this problem, we introduce the *logistic regression function* g:

$$\pi_i = g(x_i) = \frac{\exp(\mathbf{x}_i^\top \beta_i)}{1 + \exp(\mathbf{x}_i^\top \beta_i)}. \tag{1.4}$$

Using the *logit function*, we can write (1.4) as

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \mathbf{x}_i^\top \beta_i.$$

1.2.1 Maximum Likelihood Estimation

The unknown parameters β are estimated via the maximum likelihood estimation (MLE) over n subjects at each edge. The likelihood function is

$$\begin{aligned} L(\beta|y_1, \cdots, y_n) &= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^\top \beta_i)}{1 + \exp(\mathbf{x}_i^\top \beta_i)} \right]^{y_i} \prod_{i=1}^{n} \left[\frac{1}{1 + \exp(\mathbf{x}_i^\top \beta)} \right]^{1-y_i}. \end{aligned}$$

The loglikelihood function is given by

$$\begin{aligned} \log L(\beta) &= \text{const.} + \sum_{i=1}^{n} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \\ &= \text{const.} + \sum_{i=1}^{n} y_i \mathbf{x}_i^\top \beta + \log(1 - \pi_i) \end{aligned}$$

and its maximum is obtained when

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \pi_i) = 0.$$

In simplifying the expression, we used the following identities

$$\frac{\partial \pi_i}{\partial \beta_0} = \pi_i (1 - \pi_i)$$

and

$$\frac{\partial \pi_i}{\partial \beta_1} = x_i \pi_i (1 - \pi_i).$$

Since the logistic regression function π is in complicated form, the maximum is obtained numerically. Define the *information matrix* $I(\beta)$ to be

$$I(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta' \partial \beta} - \sum_{i=1}^{n} \pi_i (1 - \pi_i) \mathbf{x}_i \mathbf{x}_i^\top.$$

Then the Newton–Raphson algorithm is used to find the MLE in an iterative fashion. Starting with an arbitrary initial vector β^0, we estimate iteratively

$$\beta^{j+1} = \beta^j + I(\beta^j)^{-1} \frac{\partial \log L(\beta)}{\partial \beta}(\beta^j).$$

Many computational packages such as R and MATLAB have the logistic regression model fitting procedure.

Although we do not have the explicit formulas for the MLE, using the asymptotic normality of the MLE, the distributions of the estimators can be approximately determined. For large sample size n, the distribution of $\widehat{\beta}$ is approximately multivariate normal with means β with the covariance matrix $I(\widehat{\beta})^{-1}$.

1.2.2 Best Model Selection

Consider following full model:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Let $\beta^{(1)} = (\beta_0, \cdots, \beta_q)^\top$ and $\beta^{(2)} = (\beta_{q+1}, \cdots, \beta_p)^\top$. The parameter $\beta^{(1)}$ corresponds to the parameters of the *reduced model*. Then we are interested in testing

$$H_0 : \beta^{(2)} = 0.$$

Define the *deviance* D of a model as $D = -2 \log L(\widehat{\pi})$, which is distributed asymptotically as χ^2_{n-p-1}. Let $\widehat{\pi}^{(p)}$ and $\widehat{\pi}^{(q)}$ be the estimated success probabilities for the full and reduced models, and let D_p and D_q be the associated deviances. Then the log-likelihood ratio statistic for testing $\beta^{(2)} = 0$ is

$$2[\log L(\widehat{\pi}^{(p)}) - \log L(\widehat{\pi}^{(q)})] = D_q - D_p \sim \chi^2_{p-q}.$$

1.2.3 Logistic Discriminant Analysis

Discriminant analysis resulting from the estimated logistic model is called the *logistic discrimination*. We classify the ith subject according to a *classification rule*. The simplest rule is to assign the ith subject as group 1:

$$P(Y_i = 1) > P(Y_i = 0).$$

This statement is equivalent to $\pi_i > 1/2$. Depending on the bias and the error of the estimation, the value $1/2$ can be adjusted. For the fitted logistic model, we classify the ith subject as group 1 if $\mathbf{x}_i^\top \beta_i > 0$ and as 0 if $\mathbf{x}_i^\top \beta_i < 0$. The plane $\mathbf{x}_i^\top \beta = 0$ is the *classification boundary* that separates two groups.

The performance of classification technique is measured by the *error rate* γ, the overall probability of misclassification. The *cross-validation* is used to estimate the error rate. This is done by randomly partitioning the data into the training and the testing sets. In the *leave-one-out* scheme, the training set consists of $n-1$ subjects, while the testing set consists of one subject. Suppose the ith subject is taken as the test set. Then using the training set, we determine the logistic model. Using the predicted model, we test if the ith subject is correctly classified. The error rate obtained in this fashion is denoted as e_{-i}. Note that $e_{-i} = 0$ if the subject is classified correctly while $e_{-i} = 1$ if the subject is misclassified. The *leave-one-out error rate* is then given by

$$\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} e_{-i}.$$

To formally test the statistical significance of the discriminant power, we use Press's Q statistic (Hair et al., 1998), which is given by

$$n(2\gamma - 1)^2 \sim \chi_1^2.$$

Press's Q statistic is asymptotically distributed as χ^2 with one degree of freedom.

1.3 Random Fields

At the voxel level, it is often necessary to model measurements at each voxel as a random field. For instance, the deformation field of warping a brain to another brain is often modeled as a continuous random field (Chung et al., 2001b). The generalization of a continuous stochastic process defined in $\mathbb{R}$ to a higher dimensional abstract space is called a *random field*. For an introduction to random fields, see (Yaglom, 1987; Dougherty, 1999; Adler and Taylor, 2007). In the random field theory as introduced in (Worsley, 1994; Worsley et al., 1996b), measurement Y at voxel position $x \in \mathcal{M}$ is modeled as

$$Y(x) = \mu(x) + \epsilon(x),$$

where μ is the unknown functional signal to be estimated and ϵ is the measurement error, which is modeled as a random variable at each fixed x. Then the collection of random variables $\{\epsilon(x) : x \in \mathcal{M}\}$ is called a *stochastic process* or *random field*. The more precise measure-theoretic definition can be found in (Adler and Taylor, 2007). Random field modeling can be done beyond the usual Euclidean space to curved cortical and subcortical manifolds (Joshi, 1998; Chung et al., 2003a). Most of concepts in random fields are the continuous generalization of random vectors.

Definition 1.1 *Given a probability space, a random field $T(x)$ defined in $\mathbb{R}^n$ is a function such that for every fixed $x \in \mathbb{R}^n$, $T(x)$ is a random variable on the probability space.*

Definition 1.2 *The covariance function $R(x, y)$ of a random field T is defined as*

$$R(x, y) = \mathbb{E}\big[T(x) - \mathbb{E}T(x)\big]\big[T(y) - \mathbb{E}T(y)\big].$$

If the joint distribution of T at points $x_1, \cdots, x_m$

$$P\Big(T(x_1) \le z_1, \cdots, T(x_m) \le z_m\Big)$$

is invariant under the translation

$$(x_1, \cdots, x_m) \to (x_1 + \tau, \cdots, x_m + \tau),$$

T is said to be stationary or homogeneous.

For a stationary random field T, its covariance function is

$$R(x, y) = f(x - y)$$

for some function f. A special case of stationary fields is an *isotropic* field, which requires the covariance function to be rotation invariant, i.e.,

$$R(x, y) = f(|x - y|)$$

for some function f (Yaglom, 1987).

1.3.1 Gaussian Fields

The most important class of random fields is Gaussian fields. A more rigorous treatment can be found in Adler and Taylor (2007). Let us start defining a multivariate normal distribution from a Gaussian random variable.

Definition 1.3 *A random vector $T = (T_1, \cdots, T_m)$ is multivariate normal if $\sum_{i=1}^m c_i T_i$ is Gaussian for every possible $c_i \in \mathbb{R}$.*

Then a Gaussian random field can be defined from a multivariate normal distribution.

Definition 1.4 *A random field T is a* Gaussian random field *if $T(x_1), \cdots, T(x_m)$ are multivariate normal for every $(x_1, \cdots, x_m) \in \mathbb{R}^m$.*

An equivalent definition to Definition 1.4 is as follows. T is a Gaussian random field if the finite joint distribution

$$P(T(x_1) \le z_1, \cdots, T(x_m) \le z_m)$$

is a multivariate normal for every $(x_1, \cdots, x_m)$.