COMPG0118: Computational Modeling in Biomedical Imaging - Coursework 2

Due: March 9th, 2021, 16:00

February 25, 2021

Submission Guideline: Electronic submission through Moodle in **PDF format**. Do **not** submit any other document formats, such as Microsoft Word or Open Office. The submission consists of two parts.

1. A written report; maximum 3-page long (excluding figures); font arial; minimum font size 10.
2. A code listing: no page limit.

Answer Guideline: Please give clear and concise **explanation** for your solutions and include the results from the key **intermediate** steps.

Topic: General Linear Model (GLM) and Permutation Testing (15 pts)

Part 1: GLM unifies all forms of classical hypothesis testing. This exercise demonstrates this capability with two standard examples.

1. Two-sample t-test for testing the hypothesis that two groups have different means:

   (a) Simulate sampled data from two groups with different means but the same variance. Set the sample size for both groups to 20, the respective means to 1.5 and 2.0, the stochastic (random) component to a normal distribution with 0 mean and 0.2 standard deviation (Use the matlab command `rng` to fix the random seed to your **student ID** and then use `randn` to simulate the desired noise). Compute the mean and standard deviation of each sample. Verify that the values are as expected.

   Note: By simulating the stochastic component e, we have the **ground truth** of the deviation of the sampled data from the respective means. You will be asked to relate this to the error estimated with GLM later.

   For the remainder of this part (1.1), please re-use the same simulated data from (a).

   (b) Compute the two-sample t-statistic of the two samples with matlab's built-in function `ttest2`.

   Note: Make sure that you have carefully studied the usage of this matlab function and understood the meaning of its output(s).

   (c) Compute the t-statistic with the GLM model:

$$Y = X_1\beta_1 + X_2\beta_2 + e\,, \tag{1}$$

   where X_1 and X_2 are the dummy variables indicating the membership for the respective groups, β_1 and β_2 are the corresponding coefficients, and e is the stochastic component. X_1 and X_2 are two column vectors (i.e., just a single column) with the number of rows equal to the total number of observations (i.e., the total number of the data points). Writing in terms of the design matrix $X = [X_1\ X_2]$, $Y = X\beta + e$ with $\beta = [\beta_1\ \beta_2]'$. The entry X_{ij} is either 1 if the observation Y_i belongs to the group j or 0 otherwise.

   i. Derive the design matrix X and determine the dimension of its column space $C(X)$, which we will denote as $\dim(X)$.

   ii. Derive the general formula of the perpendicular projection operator P_X corresponding to any $C(X)$ and demonstrate that P_X satisfies the key properties of a perpendicular projection operator; use your formula to determine the P_X for $C(X)$ in question; compute the trace of P_X and explain what the result means.

iii. Use P_X to determine $\hat{Y}$, the projection of Y into $C(X)$; explain what $\hat{Y}$ means and why $C(X)$ is also known as the estimation space.

iv. Compute $R_X = I - P_X$ (I denotes the identity matrix of the same dimensions as P_X) and demonstrate that R_X is also a perpendicular projection operator.

v. Use R_X to determine $\hat{e}$, the projection of Y into the error space $C(X)^\perp$, which provides an estimate of the error vector; determine the dimension of $C(X)^\perp$.

vi. Determine the angle between $\hat{e}$ and $\hat{Y}$ and explain if this is what one should expect.

vii. Derive the general formula for estimating the model parameters of any GLM and explain why this is known as a least squares estimate; use your formula to determine $\hat{\beta}$, the estimate to the model parameters of the GLM considered here.

viii. Estimate the variance of the stochastic component $\hat{e}$ with

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n - \dim(X)} \tag{2}$$

and explain why this is also known as the mean squared errors (MSE).

ix. Estimate the covariance matrix of the estimated model parameters $\hat{\beta}$ with

$$S_{\hat{\beta}} = \hat{\sigma}^2 (X'X)^{-1}. \tag{3}$$

Use the estimated covariance matrix to determine the standard deviation of the model parameters.

x. Derive the appropriate contrast vector λ for comparing the group differences in the means, then the reduced model X_0 corresponding to the null hypothesis H_0: $\lambda'\beta = 0$.

xi. Use the reduced model X_0 to compute the (additional) error as a result of placing the constraint $\lambda'\beta = 0$, then estimate the F-statistic of comparing the reduced model X_0 to the full model X. What are the degrees of freedoms of the F-statistic in question?

xii. The F-statistic can only be used to determine whether the two groups have different means. To test whether one group has a higher means than another, determine the t-statistic with

$$t = \frac{\lambda'\hat{\beta}}{\sqrt{\lambda' S_{\hat{\beta}} \lambda}}. \tag{4}$$

(Is your answer identical to that in b)?)

xiii. Explain the meaning of the model parameters. (See (a)) What should be their ground truth values?

xiv. Compute the projection of the ground truth deviation e into $C(X)$. Explain how does this relate to $\hat{\beta}$ and the ground truth β.

xv. Compute the projection of the ground truth deviation e into $C(X)^\perp$. How does this relate to $\hat{e}$?

(d) Compute the t-statistic with a different model:

$$Y = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + e\,, \tag{5}$$

which differs from the model in Eqn. (1) only in the constant variable X_0 (the intercept) to explain the common effect between the groups, such that all its entries are equal to 1.

i. Write down the design matrix X and determine the dimension of its column space $\dim(X)$.

ii. Compute the corresponding P_X and how does it compare to the one in (c)? How does the corresponding estimation space compare to the one in (c)?

iii. Determine the appropriate contrast vector λ and the reduced model corresponding to the constraint $\lambda'\beta = 0$.

iv. Determine the t-statistic. Here the variance matrix of the estimated parameters need to be computed differently because $(X'X)^{-1}$ is no longer defined when X is not full rank. The inverse should be replaced by the pseudo-inverse (matlab built-in function for this is `pinv`).

v. Explain the meaning of the model parameters.

(e) Compute the t-statistic with yet another model:

$$Y = X_0\beta_0 + X_1\beta_1 + e\,, \tag{6}$$

i.e., we simply drop the dummy variable X_2.

i. Write down the design matrix X and determine the dimension of its column space $\dim(X)$.

ii. Determine the appropriate contrast vector λ and the reduced model corresponding to the constraint $\lambda'\beta = 0$.

iii. Determine the t-statistic.

iv. Explain the meaning of the model parameters.

(f) Can we test the same hypothesis with an even simpler model:

$$Y = X_0\beta_0 + e\,? \tag{7}$$

2. Paired t-test for comparing the means from repeat measurements from the same group of subjects.

I am assuming there are 40 subjects here

(a) Now treat the simulated data from above as the repeat observations from the same group of subjects, i.e., treating the two groups now as the repeat measurements of the same set of individuals at two different points in time.

I'm assuming we're doing a paired t-test for the two repeats and comparing against the two sample t-test

i. Compute the paired t-statistic using matlab's built-in function `ttest`. How does this statistic compare to the one estimated with two-sample t-test?

For the remainder of this part (1.2), please re-use the same simulated data from (a).

(b) Compute the same statistic with the GLM model:

$$Y = X_0\beta_0 + X_1\beta_1 + \sum_{i=1}^{N} S_i s_i + e\,, \tag{8}$$

where X_0 is the constant variable, X_1 is the explanatory variable for indicating different time points, N is the number of subjects, S_i is the dummy variable for indicating if an observation is made on the subject i, s_i is the parameter for S_i.

i. Determine the design matrix X and its rank.

ii. Write down the appropriate contrast vector λ.

iii. Compute the t-statistic. (Is your answer identical to that in (a)?)

Part 2: Permutation testing makes few distributional assumptions and therefore can often provide more robust inference than the parametric approaches using GLM. This exercise demonstrates this both for individual testing as well as for multiple comparisons correction.

1. Single permutation test for testing the hypothesis that two groups have different means:

(a) Simulate a similar dataset as in Part 1.1 but with much smaller sample sizes of $n_1 = 6$ and $n_2 = 8$ respectively, then determine the t-statistic and p-value using `ttest2`.

(b) Compute the exact permutation-based p-value.

i. First construct an one-dimensional array D of size $n = n_1 + n_2$ to store the observations with the first n_1 entries for group 1 and the rest for group 2.

ii. Then construct all the valid permutations of D such that the sample size of each group is maintained. You can do this with matlab's built-in function `combnk`. (Consider what the input arguments should be.)

iii. Next, compute the t-statistics for all the membership permutations to construct the empirical distribution of the t-statistic.

iv. Finally, determine the p-value by finding the percentage of the permutations with a t-statistic greater than and equal to that of the original labeling.

How does this value compare to the one from (a)?

(c) Repeat (b) but rather than using the t-statistic, use the difference between the means as the test statistic.

(d) Compute the approximate permutation-based p-value. Considering all possible permutations is only feasible for very small sample sizes. For larger sample sizes, we would need a way to determine a random subset from all the valid permutations. One way to do this is to use matlab's built-in command `randperm` to generate a random set of permutations of the integers from 1 to $n_1 + n_2$. Since these integers are precisely the indices for the data array D, each permutation of these indices provide a permutation of the data array.

i. Use this approach for estimating p-value with t-statistic and 1000 permutations only. Note that one must always include the original labeling.

ii. How does the estimated p-value compare to that of (b) and (c)?

iii. Check if there ere any duplicates in these 1000 permutations. Explain how the existence of dupliates may affect the p-value estimation.

2. Single threshold test for multiple comparisons correction:
This problem uses the provided fractional anisotropy (FA) maps for two groups of 8 subjects each. The FA maps for group 1 have their filenames starting with CPA; the ones for group 2 with PPA.
Below is an example code snippet for loading the maps into the matlab:

```
fid = fopen('CPA4_diffeo_fa.img', 'r', 'l') % little-endian
data = fread(fid, 'float') % 16-bit floating point
data = reshape(data, [40 40 40]) % dimension 40x40x40
```

An additional binary volume, `wm_mask.img`, is provided, which defines the region-of-interest (ROI) for statistical analysis. You only need to analyze the voxels with a non-zero value in the mask.

(a) Compute, for each voxel in the ROI, the two-sample t-statistic between the groups using a GLM of choice (not with `ttest2`), then determine the maximum t-statistic among all the voxels.

(b) Use the same strategy as in 1-(b) to determine all the possible permutations of group labels, then repeat (a) for each permutation to construct the empirical distribution of the maximum t-statistic. This is for the 1_b in part 2

(c) Determine the multiple-comparisons-corrected p-value by finding the percentage of the permutations with a maximum t-statistic greater than that of the original labeling.

(d) Determine the maximum t-statistic threshold corresponding to p-value of 5%.