General Linear Model and Permutation Testing

1. General Linear Model

1.1. Two Sample T-test

We start by creating simulated data for two groups of 20 as indicated in the instructions. Setting random seed to student id, a normally distributed stochastic component scaled by 0.2 standard deviation (std for brevity) is added to the groups with 1.5 and 2.0 mean values. Computing the mean of samples results in 1.5648 and 2.0445 for groups 1 and 2 respectively, with std of 0.2558 and 0.2213, values within expectation.

A two-sample t-statistics (t-stat for brevity) using matlab `ttest2` function results in -6.3424, with 38 degrees of freedom (df), and p-value of $1.9326e^{-7}$. This results in rejection of the null hypothesis at the default 5% significance level. Confidence interval (CI) of -0.6328 and -0.3266 doesn't contain 0, indicating significant difference between the two groups.

1.2. GLM Model $Y = X_1\beta_1 + X_2\beta_2 + e$

The design matrix X consists of 40 rows and 2 columns, with the first 20 rows of column 1 and rows 21 to 40 of column 2 filled with ones, and the rest zero, indicating membership of the two groups. The dimension of column space $C(X)$, $dim(X)$, is 2, equal to its rank since two columns are linearly independent.

The projection operator P_x to any $C(X)$ is calculated as $X \cdot \text{pinv}(X' \cdot X) \cdot X'$ where X' is transpose of X and $\text{pinv}(X)$ is pseudoinverse of X, resulting in a symmetric and idempotent matrix. Inverse could be used if the matrix is invertible. Symmetric property can be proven with derivation 1 whereas idempotent with derivation 2.

$$\begin{aligned} P_x' &= (X \cdot \text{pinv}(X' \cdot X) \cdot X')' = (X')' \cdot \text{pinv}(X' \cdot X)' \cdot X' \\ &= X \cdot \text{pinv}(X' \cdot X) \cdot X' = P_x \end{aligned} \tag{1}$$

$$\begin{aligned} P_x \cdot P_x &= X \cdot \text{pinv}(X' \cdot X) \cdot X' \cdot X \cdot \text{pinv}(X' \cdot X) \cdot X' \\ &= X \cdot \text{pinv}(X' \cdot X) \cdot X' = P_x \end{aligned} \tag{2}$$

We calculate P_x of our design matrix X that is a 40×40 matrix with 2 20×20 matrix of 0.05 in their diagonals and rest zero, it has trace 2 which is the sum of its eigenvalues, it is also $dim(X)$, the dimension of the subspace spanned by X.

We stack both groups' observation to create Y, and calculate $\hat{Y} = P_x \cdot Y$, which is the projection of Y onto $C(X)$. $\hat{Y}$ is the best predicted value of Y living in $C(X)$, turns out to be a length 40 vector with first 20 entries 1.5648 and latter 20 entries of 2.0445, which are also the empirical means of the two groups. $C(X)$ is known as the "estimation space" for this reason, because it is the full enclosed space that any estimation of Y can be made.

Residual R_x is calculated as $I - P_x$, it is also a perpendicular projection operator with symmetry property: $R_x' = (I - P_x)' = I' - P_x' = I - P_x = R_x$, and idempotent property: $R_x \cdot R_x' = R_x \cdot R_x = (I - P_x) \cdot (I - P_x) = I - P_x - P_x + P_x \cdot P_x = I - P_x$. This results in a 40×40 matrix, where two 20×20 blocks are at top left and bottom right corners. Values in the blocks are -0.05 except for the diagonal which are 0.95, zero elsewhere.

We compute $\hat{e}$ using $R_x \cdot Y$, which effectively gives an estimate of the error vector. Orthogonal complement $C(X)^\perp$ is the error space that Y is projected onto using R_x. Its dimension is $n - \text{rank}(X)$, which is 38 in given n is 40 and $\text{rank}(X) = 2$, verifiable with trace of R_x.

Since $\hat{e}' \cdot \hat{Y} = 0$, the angle between them is 90 degrees, which makes sense because $\hat{Y}$ lives in $C(X)$ and $\hat{e}$ in $C(X)^\perp$ minimising the error.

To derive the formula of parameter estimation for general linear model, we start with $\hat{e}' \cdot X \cdot \alpha = 0$; substituting $\hat{e}$ with $Y - X \cdot \hat{\beta}$, setting $Y' \cdot X - \hat{\beta}' \cdot X' \cdot X = 0$, finally estimated parameters $\hat{\beta} = (X' \cdot X)^{-1} \cdot X' \cdot Y$. This is the general formula for GLM using least squares estimate, named so because it also minimises the sum of squared residuals. We use this approach to estimate $\hat{\beta}$, resulting again in 1.5648 and 2.0445, the empirical means of the two groups.

The estimated noise $\hat{\sigma}^2$ is calculated as $\frac{\hat{e}' \cdot \hat{e}}{n - dim(X)}$, resulting in 0.0572. This is also known as the mean squared error (MSE), because it is the result after $\hat{e}' \cdot \hat{e}$ multiplication and averaged in unbiased form.

The covariance matrix $S_{\hat{\beta}}$ of $\hat{\beta}$ is calculated as $\hat{\sigma}^2 \cdot (X' \cdot X)^{-1}$, resulting in a diagonal matrix of size 2×2

with elements 0.0029. Then squaring it results in the std of $\hat{\beta}$, 0.0535 for both parameters.

Now, the appropriate contrast vector λ to compare group difference in means is $[1, -1]$, since the null hypothesis is that the two groups have the same mean, $H_0 = \lambda' \cdot \beta$. Thus, the reduced model X_0 would be $X \cdot N(\lambda')$, where N is the null space of λ. Conceptually it would be summing two columns of X, leaving only one column in X_0. In practice we see X_0 being a 40×1 matrix with all 0.7071 entries.

The error $\hat{e}_{H_0}$ is calculated based on the reduced model, obtaining $\text{SSR}(X_0) = 4.4749$ and $\text{SSR}(X) = 2.1738$. With df $v_1 = 1$ and $v_2 = 38$, the f-statistics using $\frac{(\text{SSR}(X_0) - \text{SSR}(X))/v_1}{\text{SSR}(X)/v_2}$ gives 40.2256 and p-value $1.9326e^{-7}$ calculated using $1 - \text{fcdf}(F, v_1, v_2)$, which is the same as the two-sample t-test.

We compute t-stat using the formula $\frac{\lambda' \cdot \hat{\beta}}{\sqrt{\lambda' \cdot S_{\hat{\beta}} \cdot \lambda}}$, resulting in -6.3424 which is the same as the two-sample t-test we conducted earlier.

The meaning of model parameters lies in their contribution to the observation for being part of group 1 or group 2 respectively. The parameters are the empirical means of the two groups matching to the means calculated in subsection 1.1 Nonetheless, the ground truth values are 1.5 and 2.0 respectively.

The projection of ground truth deviation e into $C(X)$, which we calculate first with $e = Y - X \cdot \beta$, and then $e_{C(X)} = P_x \cdot e$, results in 0.0648 the first 20 and 0.0445 the last 20, in 40-element vector. This relates to $\hat{\beta}$ and β in that $\hat{\beta} = \text{inv}(X' \cdot X) \cdot X' \cdot Y$ and $Y = X \cdot \beta + e$, thus $\hat{\beta} = \beta + \text{inv}(X' \cdot X) \cdot X' \cdot e$. And it turns out that $X' \cdot P_x = X'$, then $X' \cdot e_{C(X)} = X' \cdot e$. Therefore, $e_{C(X)}$ represents the part of ground truth deviation e that can be explained within $C(X)$ and contributes to the estimation of $\hat{\beta}$ by causing deviations in the estimated parameters.

The projection of e into $C(X)^\perp$ instead, is $e_{C(X)^\perp} = R_x \cdot e$. Using the similar logic of $P_x \cdot X = X$, we actually find out that this is same as $\hat{e}$, $\hat{e} = R_x \cdot Y = (I - P_x) \cdot (X \cdot \beta + e) = (X \cdot \beta - P_x \cdot X \cdot \beta) + (I - P_x) \cdot e = (I - P_x) \cdot e = R_x \cdot e$. This makes sense since that the error $\hat{e}$ of the model are the same as the projection of the ground truth error e to the residual space, space orthogonal to $C(X)$.

1.3. GLM Model $Y = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + e$

This design matrix X still with 40 rows will have 3 columns, the first one all ones and the other two the same as the previous model. The dimension is 2, because the three columns are no longer linearly independent, for example the first column is the sum of the other two. We can check this with the rank as well.

P_x results in a very similar matrix to the previous model, with only minor numbers in the 0 squares, most probably due to numerical precision and using pseudo inverse. The corresponding estimation space is conceptually same as before, since the new intercept column lives in the same space formed by the original two columns.

The contrast vector λ in this case should be $[0, 1, -1]$, as we still want to test the null hypothesis that β_1 and β_2 have equal effects, and the reduced model X_0 should be $X \cdot N(\lambda')$ with this new λ. X_0 results to be a 40×2 matrix where first column is -0.2071 and second 1.2071.

We estimate parameters $\hat{\beta}$ and its covariance matrix as usual, and get a t-stat of -6.3424, which is consistent to all previous results.

The parameters are respectively $\hat{\beta} = [1.2031, 0.3617, 0.8414]$. The first parameter $\hat{\beta}_0$ indicates the prevalence of observation irrespective of membership, equivalent to the baseline. And the $\hat{\beta}_1$ and $\hat{\beta}_2$ are the deviations of empirical means from the intercept. Also, we observe that $\hat{\beta}_1 + \hat{\beta}_2 = \hat{\beta}_0$, consistent with the contrast vector and linear dependency of the columns.

1.4. GLM Model $Y = X_0\beta_0 + X_1\beta_1 + e$

X in this model is simply dropping the third column of the previous model, resulting in a 40×2 matrix with first column full of ones and second column rows 1 to 20 filled with 1s and 21 to 40 with 0s. The dimension is 2, as the two columns are linearly independent, verified by the rank.

The contrast vector λ is set to $[0, 1]$, because we want to test the null hypothesis that the group 1 has no effect. The reduced model X_0 is $X \cdot N(\lambda')$, resulting in a 40×1 matrix with all -1.

Resulting t-stat is still -6.3424, consistent with the previous models. The interpretation of the parameters $\hat{\beta}$ are that $\hat{\beta}_0$ is the empirical mean of group 2 (serving as the baseline intercept). And $\hat{\beta}_1$ is the deviation for being member of group 1, matching with the difference between the empirical means of the two groups.

If the model is $Y = X_0\beta_0 + e$, we cannot test null hypothesis as before, because β_0 cannot differentiate between the two groups.

1.5. Paired T-test

If we now consider repeated observations of same subjects, conducting paired t-test using `ttest` function, we obtain t-stat of -5.9258, with 19 df, p-value increased to $1.0515e^{-5}$, and wider CI of $[-0.6491, -0.3103]$. Despite less extreme t-stat, the p-value remains significant, indicating group difference

in repeated observations.

Turning to the GLM model proposed in the instructions, we set up the design matrix X as a 40×22 matrix, with first column full of 1s, second column half 0s (rows 1 to 20) and half 1s (rows 21 to 40), and the rest columns as a vertical stack of identity matrix, representing subjects at each time point. Due to one linear dependency of sum of dummy columns equating to the intercept column, $\mathrm{rank}(X) = 21$.

The contrast vector λ is set to $[0, 1, 0, 0, \ldots, 0]$, to test the hypothesis that time point does not have an effect. Calculated t-stat is 5.9258, which is the same value but opposite sign as the paired t-test. As the t-distribution is symmetric, we perform two-tail test using `abs(t_stat)`, resulting in p-value of $1.0515e^{-5}$, verifying findings on both ways.

2. Permutation Testing

2.1. Single Permutation Test

Simulating a dataset with same means and std but with sizes $n_1 = 6$ and $n_2 = 8$ results in a t-stat of -4.3788 and p-value of $8.9814e^{-4}$ using `ttest2` function, indicating significant difference between groups.

D is obtained concatenating the observations of two groups. Then we find all possible permutations of the indices to fit in group 1 from the indices range, using `nchoosek` setting k to n_1. For each permutation, we insert D with indices shown in the permutation to group 1, and the rest to group 2 using `setdiff`. We calculate the t-stat for each permutation and store them to create its empirical distribution.

Strictly following the instructions, we find p-value is 1, meaning 100% of the values are greater or equal to the original t-stat. This suggests that we fail to reject this orientated one-tailed hypothesis. However, this p-value is extreme, so if we conduct a two-tailed test (or change direction of testing) we would obtain significant p-values that lead to rejection of the hypothesis. For instance, using absolute t-stat which is two-tailed t-test, the p-value is $6.66e^{-4}$, corresponding to $2/3003$ cases of t-stats same or more extreme than the original at two ends. Compared to the split significance level of 2.5%, we would still reject the null hypothesis.

To use difference between means as the test statistic, we calculate this difference between means of two groups instead of t-stat each permutation. It results in the same p-value as the previous part, on both versions. Indicating that the two groups are significantly different on both test statistics.

2.2. Approximate Permutation Test

Reminding that we set up random seed for replication, we compute approximate permutation-based p-value by sampling from all possible permutations. We use 1000 permutations where each iteration `randperm` is used to generate permuted index, assign labels by order, and then calculate t-stat. We check if any of them is the original label-position permutation and repeat the process if this is not the case. It took 2 loops in this case. The one-tailed p-value resulted in 0.9980, strictly following the greater or equal to rule, while the two-tailed p-value is 0.0020.

This number is less extreme than the previous two, and one of the reasons could be appearance of duplicates. Here by duplicate we mean label-position combination duplicates, meaning their grouping is the same, rather than exact sequence duplicates. The latter is highly improbable with $14!$ possibilities. The former, however, appears 41 times in our 1000 permutations. The effects of these duplicates are that they bias the distribution by overweighting certain combinations, under-representing other possible partitions. This results in inflating the p-value, because the duplicates introduce redundant points in the denominator not introducing new unique cases, effectively reducing the statistical significance.

2.3. Single Threshold Test for Multiple Comparisons Correction

We firstly load and collect imaging maps and mask as instructed, for 2 groups, each with 8 subjects. Then iterate over all voxels of the mask, skipping the zero ones. At each iteration, we extract the subject data from each map in both groups, and calculate the t-stat using GLM with model of Section 1.2. After collecting all t-stat, the maximum t-stat identified within 19908 voxels is 6.5294.

To build the permutation version of the test, we create all possible combinations of 8 out of 16 subjects using `nchoosek`. This results in 12870 combinations, for each of which we permute the group memberships and calculate the t-stat same as paragraph before. Due to the large size of permutations and voxels, the process took around 5100 seconds.

The multiple-comparisons-corrected p-value is computed by comparing the proportion of max t-stats in the permutation distribution that is greater or equal to the original max t-stat, resulting in 0.00918. The maximum t-stat threshold for p-value of 5%, on the other hand, is the 5% quantile (rounding down for strictness) of reverse-sorted max t-stats, resulting in 6.9383. This controls Familiy-Wise Error Rate (FWER).