MAST90104: A First Course in Statistical Learning

Week 8 Practical and Workshop

1 Workshop questions

1. An industrial psychologist is investigating absenteeism among production-line workers, based on different types of work hours: (1) 4-day week with a 10-hour day, (2) 5-day week with a flexible 8-hour day, and (3) 5-day week with a structured 8-hour day. A study is conducted and the following data obtained of the average number of days missed:

	Work plan		
	1	2	3
Mean	9	6.2	10.1
Number	100	85	90

They also find $s^2 = 110.15$.

(a) Test the hypothesis that the work plan has no effect on the absenteeism.
Solution: For this data, we can fit the less than full rank model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

This can be reparameterised as

$$y_{ij} = \gamma_i + \epsilon_{ij}$$

using

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Easy to see that

$$\widetilde{X}^T\widetilde{X} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix}, \quad (\widetilde{X}^T\widetilde{X})^{-1} = \begin{bmatrix} 1/n_1 & 0 & 0 \\ 0 & 1/n_2 & 0 \\ 0 & 0 & 1/n_3 \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}$$

.
From the notes, we know that $\widehat{\boldsymbol{\gamma}}$ would be the sample mean of each group, $\widehat{\boldsymbol{\gamma}} = (9, 6.2, 10.1)^T$
The hypothesis that the work plan has no effect on the absenteeism is $H_0 : \tau_1 = \tau_2 = \tau_3$, which is true if and only if

$$\tau_1 - \tau_2 = 0, \text{ and } \tau_2 - \tau_3 = 0$$

. Since $\gamma_i = \mu + \tau_i$, this is equivalent to testing $\gamma_1 - \gamma_2 = 0$ and $\gamma_2 - \gamma_3 = 0$.
In this model, the hypothesis is $H_0 : L\beta = \mathbf{0}$ where

$$L = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

.
Note that

$$\widetilde{L} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

.
The test statistics is

$$F^* = \frac{(L\widehat{\beta})^T[\widetilde{L}(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{L}^T]^{-1}L\widehat{\beta}/2}{s^2}$$

We can compute

$$[\widetilde{L}(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{L}^T]^{-1} = \frac{n_1n_2^2n_3}{n_1n_2+n_2^2+n_2n_3}\begin{bmatrix} 1/n_2+1/n_3 & 1/n_2 \\ 1/n_2 & 1/n_1+1/n_2 \end{bmatrix}$$

And $L\widehat{\beta} = \widetilde{L}\widehat{\gamma} = \begin{bmatrix} 2.8 & -3.9 \end{bmatrix}^T$. Substitute this $n_1 = 100$, $n_2 = 85$ and $n_3 = 90$ to the formula gives $F^* = 3.200371$. The test statistic follows an F distribution with degrees of freedom 2 and $(n_1+n_2+n_3)-3$

```
> gammahat <- c(9,6.2,10.1)
> s2 <- 110.15
> n <- c(100,85,90)
> r <- 3
> XtildetXtildeinv <- diag(c(1/n))
> (Ltilde <- matrix(c(1,-1,0,0,1,-1),2,3,byrow=T))
[,1] [,2] [,3]
[1,]    1   -1    0
[2,]    0    1   -1
> (Fstat <- t(Ltilde%*%gammahat)%*%solve(Ltilde%*%XtildetXtildeinv%*%t(Ltilde))
%*%Ltilde%*%gammahat/2/s2)
[,1]
[1,] 3.200371
> pf(Fstat,2,sum(n)-r,lower=F)
[,1]
[1,] 0.04228613
```

Therefore we reject the null hypothesis at a 5% level: work plan has an effect on absenteeism.

(b) Test the hypothesis that work plans 1 and 3 have the same rate of absenteeism.

Solution The test is $H_0 : \tau_1 = \tau_3$ or $L\beta = 0$ where $L = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix}$. The corresponding $\widetilde{L} = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$ Using the formula of the F-test above, we can compute test statistic $F^* = 0.5203431$, it has an F distribution with degrees of freedom 1 and $(n_1+n_2+n_3)-3$

```
> Ltilde <- matrix(c(1,0,-1),1,3,byrow=T)
[,1] [,2] [,3]
[1,]    1    0   -1
> Fstat <- t(Ltilde%*%gammahat)%*%solve(Ltilde%*%XtildetXtildeinv%*%t(Ltilde))
%*%Ltilde%*%gammahat/1/s2
[,1]
[1,] 0.5203431
> pf(Fstat,1,sum(n)-r,lower=F)
[,1]
[1,] 0.471315
```

We cannot reject the null hypothesis.

2. Prove Theorem 6.2 using the following steps.

(a) Show that under the conditions of Theorem 6.1, the column space of XC is the same as the column space of X.

Solution: Every column of XC is a linear combination of columns in X so it is in the column space of X.

Since, under the conditions of the Theorem, XC is full rank, the columns of XC are thus a basis for the column space of X. Hence every element of the column space of X can be expressed as a linear combination of the columns of XC. That is every element of the column space of X is in the column space of XC, showing that the two column spaces are the same.

(b) Show that if two full-rank linear model design matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same column space, then the eigenvectors of their hat matrices are the same.

Solution: Note that since $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same column space, then every column of $\mathbf{X}_1$ can be expressed as a linear combination of columns of $\mathbf{X}_2$. Hence, $\mathbf{X}_1 = \mathbf{X}_2\mathbf{A}$. Moreover, every column of $\mathbf{X}_2$ can be expressed as a linear combination of columns of $\mathbf{X}_1$. Hence, $\mathbf{X}_2 = \mathbf{X}_1\mathbf{B}$.

Now consider their respective hat matrices H_1, H_2.

Then by Spectral theorem, each H_1 and H_2 has n real eigenvalues and corresponding eigenvectors. Moreover, by theorem 2.2, their eigenvalues are all either 0 or 1 since they are idempotent, symmetric matrices.

Take an eigenvector, $\mathbf{x}$, for H_1 which has eigenvalue 1.

Then $\mathbf{x}$ is also an eigenvector with eigenvalue 1 for H_2 since

$$H_2\mathbf{x} = H_2H_1\mathbf{x} = H_1\mathbf{x} = \mathbf{x},$$

the third step follows from $H_2H_1 = X_2(X_2^TX_2)^{-1}X_2^TX_2A(X_1^TX_1)^{-1}X_1 = H_1$.

Next, take an eigenvector $\mathbf{z}$ for H_1 which has eigenvalue 0.

Then $\mathbf{z}$ is also an eigenvector with eigenvalue 0 for H_2 since

$$H_2\mathbf{z} = X_2(X_2^TX_2)^{-1}B^TX_1^T\mathbf{z} = X_2(X_2^TX_2)^{-1}B^TX_1^T\underbrace{X_1(X_1^TX_1)^{-1}X_1^T\mathbf{z}}_{=\mathbf{0}} = \mathbf{0},$$

(c) Hence show that if the column space for two linear models is the same, the fitted values are the same.

Solution: Since H_1 is a symmetric and idempotent matrix, there exists a P such that $P^TH_1P = D$, where D is a diagonal matrix with diagonal entries of eigenvalues of H_1 which are equal 0 or 1. Since H_1 and H_2 have same eigenvalue-eigenvector pairs, $P^TH_2P = D$. Hence, $H_1 = PDP^T = H_2$. Therefore, the fitted value under model 1 equals $H_1\mathbf{y} = H_2\mathbf{y}$ which is equals to the fitted values under model 2.

(d) Prove Theorem 6.2.

Solution:

Consider two choices of p by r matrices C_1 and C_2 such that $\widetilde{X}_1 = XC_1$ and $\widetilde{X}_2 = XC_2$ are full rank. By part (a), $\widetilde{X}_1$ and $\widetilde{X}_2$ have the same column space. Hence, by part (b), $\widetilde{H}_1 = \widetilde{X}_1(\widetilde{X}_1^T\widetilde{X}_1)^{-1}\widetilde{X}_1^T$ and $\widetilde{H}_2 = \widetilde{X}_2(\widetilde{X}_2^T\widetilde{X}_2)^{-1}\widetilde{X}_2^T$ have the same eigenvalue-eigenvector pairs. In fact, by part (c), the fitted values under both models are equal, i.e., $\widetilde{H}_1\mathbf{y} = \widetilde{H}_2\mathbf{y}$ for all n-dimensional vector $\mathbf{y}$. Since SS_{res}, SS_{reg}, residuals, and mean sqaured-errors are function of the hat matrix, then these quantities are equal under both models.

3. Consider question 2 in practical class this week, where we study the effect of various breeds and diets on the milk yield of cows. A study is conducted on 9 cows and the following data obtained:

	Diet		
Breed	1	2	3
1	18.8	16.7	19.8
	21.2		23.9
2	22.3	15.9	21.8
		19.2	

(a) Express this as a two-factor model with no interaction in matrix form.

Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 18.8 \\ 21.2 \\ 16.7 \\ 19.8 \\ 23.9 \\ 22.3 \\ 15.9 \\ 19.2 \\ 21.8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

and $\boldsymbol{\varepsilon}$ is as expected.

(b) Express this as a two-factor model with interaction in matrix form.

Solution: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 18.8 \\ 21.2 \\ 16.7 \\ 19.8 \\ 23.9 \\ 22.3 \\ 15.9 \\ 19.2 \\ 21.8 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \xi_{11} \\ \xi_{12} \\ \xi_{13} \\ \xi_{21} \\ \xi_{22} \\ \xi_{23} \end{bmatrix}$$

and $\boldsymbol{\varepsilon}$ is as expected.

(c) Express the hypothesis that there is no interaction in terms of your parameters. Eliminate any redundancies.

Solution: We know that we require $(I-1)(J-1) = 2$ hypotheses, so we take the obviously non-redundant hypotheses

$$\begin{aligned} (\xi_{11} - \xi_{12}) - (\xi_{21} - \xi_{22}) &= 0 \\ (\xi_{11} - \xi_{13}) - (\xi_{21} - \xi_{23}) &= 0. \end{aligned}$$