

מודלים סטטיסטיים ויישומיהם 52518 תשע"ח – פתרון תרגיל 1

1. כתבו קוד R לביצוע מבחן F של גרסיה. עבור הנתונים בקובץ המצורף `targil1.xlsx`, השתמשו בתוכנית שלכם כדי לבדוק את השערת האפס שהמתקדמים של המשתנים x_8, \dots, x_{13} שווים כולם לאפס.

קוד R:

```
FTest <- function(y,X,Hcols){  
  X0 <- X[,-Hcols]  
  reg.full <- summary(lm(y ~ X))  
  reg.0 <- summary(lm(y ~ X0))  
  SSE <- reg.full$df[2] * (reg.full$sigma)^2  
  SSE0 <- reg.0$df[2] * (reg.0$sigma)^2  
  Fstat <- (((SSE0 - SSE) / (length(Hcols))) / ((SSE) / (reg.full$df[2])))  
  return(pf(Fstat, length(Hcols), reg.full$df[2], lower.tail = F))  
}
```

```
library(readxl)  
D <- read_excel('targil1.xlsx')  
X <- as.matrix(D[,2:14])  
y <- D[, "Y"]
```

```
FTest(y,as.matrix(X),c(8:13))
```

הפלט המתקבל הוא $3.679 \cdot 10^{-46}$, לכן בהנתן השערת אלטרנטיבה פשוטה נדחה את H_0 (במעט) בכל רמת מובהקות.

2. בקובץ המצורף `fertilizers.csv` מצויים נתוני תנובות של שדות (בעלי אותו גודל) שקיבלו זני דשן שונים. על מנת להשוות בין סוגי הדשן, מבקשים לערוך מבחן F.

א. הגדירו מהי השערת האפס ומה השערת האלטרנטיבה במקרה זה.

ב. חשבו את ממוצעי הקבוצות השונות, הגדירו את המטריצה W וחשבו את סטטיסטי המבחן F.

ג. בהתאם להשערות אותן הגדרתם בסעיף א' ולסטטיסטי המבחן אותו קיבלתם בסעיף ב', הכריעו האם לדחות את השערת האפס או לא. הסבירו את צעדיכם.

א. השערת האפס היא כי לסוג הדשן אין השפעה על התנובה הממוצעת, כלומר $\bar{y}_1 = \bar{y}_2 = \dots = \bar{y}_I$. בהתאם, השערת האלטרנטיבה תהיה כי לסוג הדשן ישנה השפעה על התנובה הממוצעת.

ב.

$$I = 4, N = 24, n_1 = n_2 = n_3 = n_4 = 6, \\ \bar{y}_1 = 21, \bar{y}_2 = 28.6, \bar{y}_3 = 25.867, \bar{y}_4 = 29.2, \bar{y}_\cdot = 26.167$$

$$W = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & & \\ 1 & 0 & & \\ 0 & 1 & & \\ & \vdots & & \vdots \\ & 1 & 0 & \\ & 0 & 1 & \\ \vdots & & \vdots & \\ & & 1 & 0 \\ & & 0 & 1 \\ & & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\|\hat{y} - \hat{y}^{(0)}\|^2 = \sum_{i=1}^I n_i (\bar{y}_i - \bar{y}_\cdot)^2 = 6(26.694 + 5.921 + 0.09 + 9.201) = 251.44$$

$$\begin{aligned}\|y - \hat{y}\|^2 &= \sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_{i.})^2 = 61.033 \\ F &= \frac{\frac{\|\hat{y} - \hat{y}^{(0)}\|^2}{I-1}}{\frac{\|y - \hat{y}\|^2}{N-I}} = \frac{\frac{251.44}{3}}{\frac{61.033}{20}} = 27.465\end{aligned}$$

ג. המכנים של סטטיסטי המבחן הם 3 ו-20 בהתאמה לכן נבדוק מול התפלגות $\mathcal{F}_{3,20}$. ה-p-value המתקבל עבור סטטיסטי זה הוא $2.712 \cdot 10^{-7}$ לכן בהנתן אלטרנטיבה פשוטה, נדחה את השערת האפס כמעט בכל רמת מובהקות.

3. הראו כי בניתוח שונות חד-כווני עם $I = 2$, הסטטיסטי F שקול לריבוע הסטטיסטי t (כאשר מדובר במבחן t להפרש תוחלות עבור שתי קבוצות בלתי-תלויות).

ראינו בכתה כי בניתוח שונות חד-כווני מתקיים $F = \frac{\frac{\|\hat{y} - \hat{y}^{(0)}\|^2}{I-1}}{\frac{\|y - \hat{y}\|^2}{N-I}}$. עבור $I = 2$ נקבל:

$$F = \frac{\frac{\|\hat{y} - \hat{y}^{(0)}\|^2}{I-1}}{\frac{\|y - \hat{y}\|^2}{N-I}} = \frac{\frac{\|\hat{y} - \hat{y}^{(0)}\|^2}{1}}{\frac{\|y - \hat{y}\|^2}{N-2}} = \frac{\frac{\|\hat{y} - \hat{y}^{(0)}\|^2}{\sigma^2}}{\frac{\|y - \hat{y}\|^2}{\sigma^2(N-2)}} = \frac{\frac{\|\hat{y} - \hat{y}^{(0)}\|^2}{\sigma^2}}{\frac{\|y - \hat{y}\|^2(N-2)}{\sigma^2(N-2)^2}} =$$

המונה מתפלג χ_1^2 והמכנה מתפלג χ_{N-2}^2 כלומר סה"כ נקבל כי הביטוי מתפלג $\frac{\chi_1^2}{\chi_{N-2}^2/(N-2)^2}$, זהו ריבוע של $\frac{\sim N(0,1)}{\sqrt{\frac{\chi_{N-2}^2}{(N-2)^2}}}$

שזו בדיוק התפלגות t_{N-2} .

4. נתייחס לניתוח שונות חד-כיווני. בכיתה הוצג המודל בצורה $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, כאשר:

$$\mu = \frac{1}{I} \sum_{i=1}^I \mu_i, \quad \alpha_i = \mu_i - \mu$$

בספרים מסוימים מוגדר μ באופן הבא:

$$\mu = \frac{1}{N} \sum_{i=1}^I n_i \mu_i$$

נניח כי $I = 3, n_1 = 9, n_2 = 7, n_3 = 5$. עם ההגדרה החדשה של μ נגדיר גם

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \eta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

א. מצאו את המטריצה X המקיימת $E[Y] = X\eta$.

ב. מצאו את המטריצה W המקיימת $E[Y] = W\beta$.

ג. מצאו ביטויים לכל רכיבי הוקטור η במונחי רכיבי הוקטור β ואת המטריצה H המקיימת $\eta = H\beta$.

ד. מצאו ביטויים לכל רכיבי הוקטור β במונחי רכיבי הוקטור η ואת המטריצה G המקיימת $\beta = G\eta$.

ה. וודאו ע"י חישוב ב- R כי מתקיים $W = XH$ וכן $X = WG$.

.N

$$E[y] = X \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 \\ \vdots (8 \text{ more}) \\ \mu + \alpha_2 \\ \vdots (6 \text{ more}) \\ \mu + \alpha_3 \\ \vdots (4 \text{ more}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots (8 \text{ more}) \\ 1 & 0 & 1 \\ \vdots (6 \text{ more}) \\ 1 & -1.8 & -1.4 \\ \vdots (4 \text{ more}) \end{bmatrix} \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}}_{\mathbf{x}}$$

.ב

$$E[y] = W \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots (8 \text{ more}) \\ \mu_2 \\ \vdots (6 \text{ more}) \\ \mu_3 \\ \vdots (4 \text{ more}) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \vdots (8 \text{ more}) \\ 0 & 1 & 0 \\ \vdots (6 \text{ more}) \\ 0 & 0 & 1 \\ \vdots (4 \text{ more}) \end{bmatrix}}_W \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}}_{\beta}$$

.ג

$$\eta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{3}{7}\mu_1 + \frac{1}{3}\mu_2 + \frac{5}{21}\mu_3 \\ \frac{4}{7}\mu_1 - \frac{1}{3}\mu_2 - \frac{5}{21}\mu_3 \\ -\frac{3}{7}\mu_1 + \frac{2}{3}\mu_2 - \frac{5}{21}\mu_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 3/7 & 1/3 & 5/21 \\ 4/7 & -1/3 & -5/21 \\ -3/7 & 2/3 & -5/21 \end{bmatrix}}_H \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}}_{\beta}$$

.ד

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu - 1.8\alpha_1 - 1.4\alpha_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1.8 & -1.4 \end{bmatrix}}_G \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}}_{\eta}$$

ה. קוד R:

```
X <- matrix(ncol = 3, data = c(rep(c(1,1,0),9),rep(c(1,0,1),7),rep(c(1,-1.8,-1.4),5)), byrow = T)
```

```
W <- matrix(ncol = 3, data = c(rep(c(1,0,0),9),rep(c(0,1,0),7),rep(c(0,0,1),5)), byrow = T)
```

```
H <- matrix(ncol = 3, data = c(3/7,1/3,5/21,4/7,-1/3,-5/21,-3/7,2/3,-5/21), byrow = T)
```

```
G <- matrix(ncol = 3, data = c(1,1,0,1,0,1,1,-1.8,-1.4), byrow = T)
```

```
all.equal(W %*% G, X) && all.equal(X %*% H, W)
```