이대영 (2022311154)

2022-04-14

HW6 Report

1. The test statistic used for testing $H_0: C\underset{\sim}{\mu}_1 = C\underset{\sim}{\mu}_2$ vs. $H_a: C\underset{\sim}{\mu}_1 \neq C\underset{\sim}{\mu}_2$ is

$$T^2 = (\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' C' \left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right) C S_p C'\right]^{-1} C(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2) \overset{H_0}{\sim} \frac{(n_1+n_2-2)}{n_1+n_2-q-1} F_{q,\, n_1+n_2-q-1}$$

where C is $q \times p$ matrix ($q \leq p$) and $\text{rank}(C) = q$

After defining functions for calculating the mean vector and covariance matrix, we define a function 'multi_prof_analysis' which conducts such a hypothesis test and calculates the p-value.

In [4]:
```
#Multivariate profile analysis assuming equal covariances
def multi_prof_analysis(df1,df2,C):
    n1 = len(df1.index)
    n2 = len(df2.index)
    q = len(C.index)
    x1_bar = mean_vector(df1)
    x2_bar = mean_vector(df2)
    S1 = cov_matrix(df1)
    S2 = cov_matrix(df2)
    Sp = ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2) #pooled covariance matrix
    t2 = ((x1_bar-x2_bar).dot(C.T).dot(np.linalg.inv((1/n1+1/n2)*C.dot(Sp).dot(C.T))).dot(C).dot((x1_bar-x2_bar).T)).iloc[0,0]
    fvalue = t2*(n1+n2-q-1)/((n1+n2-2)*q)
    pvalue = f.sf(fvalue, q, n1+n2-q-1)
    return {'t2': t2, 'f-value': fvalue, 'p-value': pvalue}
```

2. (a) First, we save 'turtle.dat' data as 'turtle'. We then divide 'turtle' into two dataframes, 'female' and 'male' based on gender. We create a matrix 'C_par' to use for parallel profile testing. We use the 'multi_prof_analysis' function defined in the code in #1 to test the hypothesis

$H_0: \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$ vs H_a: not H_0

This can also be expressed as

$H_0: C\underset{\sim}{\mu}_1 = C\underset{\sim}{\mu}_2$ vs. H_a: not H_0 where $C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$

In [8]:
```
#(a) Are the profiles parallel?
C_par=pd.DataFrame([[1,-1,0],[0,1,-1]])
C_par
```

Out[8]:

	0	1	2
0	1	-1	0
1	0	1	-1

In [9]:
```
multi_prof_analysis(female,male,C_par)
#Reject null hypothesis.
#Profiles are not parallel.
```

Out[9]:
```
{'t2': 15.423169058017201,
 'f-value': 7.543941387073631,
 'p-value': 0.0014947765346784245}
```

p-value = 0.00149 < 0.05. At significance level 0.05, we reject the null hypothesis. The profiles are not parallel.

(b) We create a matrix 'C_coin' to use for coincident profile testing. We use the 'multi_prof_analysis' function again to test the hypothesis

$H_0: \mu_{1i} = \mu_{2i}$, $i = 1, 2, 3$ vs. H_a: not H_0

This can also be expressed as

$H_0: 1'\underset{\sim}{\mu}_1 = 1'\underset{\sim}{\mu}_2$ vs. H_a: not H_0 where $1' = [1, 1, 1]$

'C_coin' matrix is equivalent to $1'$ in the hypothesis.

In [10]:
```
#(b) Are the profiles coincident?
C_coin=pd.DataFrame([np.ones(3)])
C_coin
```

Out[10]:

	0	1	2
0	1.0	1.0	1.0

In [11]:
```
multi_prof_analysis(female,male,C_coin)
#Reject null hypothesis.
#If the profiles are not parallel, they cannot be coincident.
#As expected, the profiles are not coincident.
```

Out[11]:
```
{'t2': 24.964840464171786,
 'f-value': 24.96484046417179,
 'p-value': 8.894702339275906e-06}
```

p-value = $8.8947 \times 10^{-6} < 0.05$. At significance level 0.05, we reject the null hypothesis. The profiles are not coincident. If profiles are not parallel, they cannot be coincident. We know from #2(a) that the profiles are not parallel, so the result is as expected.

(c) First, we preprocess the data so that we can use it in a function from a Python package.

In [12]:
```
#(c) Repeat (a) and (b) using Python packages.
#Repeat (a)
#Data preprocessing
diff = turtle.drop('gender',axis=1)
diff = diff.dot(C_par.T)
diff.columns = ['x12','x23']
para = diff.join(pd.DataFrame(turtle['gender']))
para
```

Out[12]:

	x12	x23	gender
0	17	43	female
1	19	46	female
2	17	44	female
3	19	44	female
4	21	44	female
5	31	42	female
6	28	49	female
7	34	48	female
8	31	51	female
9	31	51	female
10	34	52	female

We use the 'MANOVA' function from the 'statsmodels' Python package to repeat parallel profile testing as in #2(a).

In [13]:
```
#Test for parallel profile
para_test = MANOVA.from_formula('x12 + x23 ~ gender', data = para)
print(para_test.mv_test())
#p-value=0.0015
#Reject null hypothesis
#Same result as #2(a)
```

```
                    Multivariate linear model
================================================================

----------------------------------------------------------------
       Intercept         Value  Num DF  Den DF  F Value  Pr > F
----------------------------------------------------------------
          Wilks' lambda  0.0175 2.0000 45.0000 1260.1630 0.0000
          Pillai's trace  0.9825 2.0000 45.0000 1260.1630 0.0000
  Hotelling-Lawley trace 56.0072 2.0000 45.0000 1260.1630 0.0000
     Roy's greatest root 56.0072 2.0000 45.0000 1260.1630 0.0000
----------------------------------------------------------------

----------------------------------------------------------------
            gender         Value  Num DF  Den DF F Value Pr > F
----------------------------------------------------------------
            Wilks' lambda 0.7489 2.0000 45.0000  7.5439 0.0015
            Pillai's trace 0.2511 2.0000 45.0000  7.5439 0.0015
    Hotelling-Lawley trace 0.3353 2.0000 45.0000  7.5439 0.0015
       Roy's greatest root 0.3353 2.0000 45.0000  7.5439 0.0015
================================================================
```

p-value = 0.0015. The result is the same as in #2(a), so we reject the null hypothesis.

For repeating #2(b), we create univariate x1+x2+x3 vectors and save them as 'female' and 'male' for each group. Assuming equal covariances, we use the 'ttest_ind' function from the 'scipy' Python package to repeat #2(b).

```
In [14]: #Repeat (b)
         #univariate x1+x2+x3
         female = turtle[turtle['gender'] == 'female'].iloc[:,:3].sum(axis=1)
         male = turtle[turtle['gender'] == 'male'].iloc[:,:3].sum(axis=1)

In [15]: #Test for coincident profile
         ttest_ind(female, male, equal_var = True) #We assumed equal covariances
         #p-value=8.8947e-06
         #Reject null hypothesis
         #Same result as #2(b)

Out[15]: Ttest_indResult(statistic=4.996482809354179, pvalue=8.894702339275784e-06)
```

p-value = 8.8947×10^{-6}. The result is the same as in #2(b), so we reject the null hypothesis.

(d) We use the 'MANOVA' function again to apply oneway MANOVA approach. Our model is

$$\underset{\sim}{X}_{ij} = \underset{\sim}{\mu}_i + \underset{\sim}{\varepsilon}_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \cdots, 24 \end{cases}$$

and the hypothesis is

$H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ vs. $H_a: \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$

For oneway MANOVA, we must assume that:

1. The random samples from different populations are independent.
2. All populations have a common covariance matrix
3. Each population is multivariate normal.

```
In [18]: man = MANOVA.from_formula('x1+x2+x3 ~ gender', data = turtle)
         print(man.mv_test())
         #p-value=0.0000 < 0.05
         #Reject null hypothesis
         #At significance level 0.05, the two population means are not equal
```

```
------------------------------------------------------------
       Intercept         Value  Num DF  Den DF  F Value  Pr > F
------------------------------------------------------------
         Wilks' lambda  0.0144 3.0000 44.0000 1001.2534 0.0000
        Pillai's trace  0.9856 3.0000 44.0000 1001.2534 0.0000
 Hotelling-Lawley trace 68.2673 3.0000 44.0000 1001.2534 0.0000
    Roy's greatest root 68.2673 3.0000 44.0000 1001.2534 0.0000
------------------------------------------------------------

------------------------------------------------------------
           gender         Value  Num DF  Den DF F Value Pr > F
------------------------------------------------------------
            Wilks' lambda 0.3886 3.0000 44.0000 23.0782 0.0000
           Pillai's trace 0.6114 3.0000 44.0000 23.0782 0.0000
   Hotelling-Lawley trace 1.5735 3.0000 44.0000 23.0782 0.0000
      Roy's greatest root 1.5735 3.0000 44.0000 23.0782 0.0000
============================================================
```

p-value = 0.0000 < 0.05. Thus, we reject the null hypothesis. At significance level 0.05, the two population means are not equal.