Statistical Data Analysis Problem sheet 3

1. Exercise 1

[37]
```
#importing all the libraries
import pandas as pd
import numpy as np
```

[38]
```
#getting the X values as a dataframe
dfx = pd.read_csv('/content/drive/MyDrive/1-DS/X.txt', sep=",", header=None)
```

[39]
```
#getting the Y values as a dataframe
dfy = pd.read_csv('/content/drive/MyDrive/1-DS/Y.txt', header= None)
```

Now we need to convert them into matrix

[40]
```
#converting to numpy array to get matrix
x = dfx.to_numpy()
y = dfy.to_numpy()
```

[41]
```
#adding an identity column to the x to equalize the full rank p + 1
x = np.insert(x, 0, 1.0, axis=1)
```

As the LS estimator is equivalent to the ML estimator based on the maximization of the log likelihood, we can estimate the beta hat from it.

[42]
```
#beta hat estimating
xt = np.transpose(x)
xtx = xt.dot(x)

beta = np.linalg.inv(xtx).dot(xt).dot(y)
beta
```

[42]

```
#beta hat estimating
xt = np.transpose(x)
xtx = xt.dot(x)

beta = np.linalg.inv(xtx).dot(xt).dot(y)
beta
```

```
array([[-0.00800698],
       [ 0.88161162],
       [-2.45938171],
       [-0.97715699]])
```

- *Here we get B0, B1, B2 and B3 for the multiple linear regression model*
- As the mean is 0 and the constant variance is 1 which gives us that N is a standard normal distrbution.

```
#now we can get the sigma hat square in the multiple linear regression model
sgms =  (np.transpose(y-x.dot(beta))*(y-x.dot(beta))) / 201
sgms
```

```
array([[ 1.43756944e-04, -2.60179376e-04,  4.88288715e-05, ...,
        -8.27929177e-04, -5.93215518e-04, -6.80968604e-04],
       [-2.60179376e-04,  4.70887219e-04, -8.83732288e-05, ...,
         1.49843263e-03,  1.07363470e-03,  1.23245515e-03],
       [ 4.88288715e-05, -8.83732288e-05,  1.65853462e-05, ...,
        -2.81216658e-04, -2.01493183e-04, -2.31299633e-04],
       ...,
       [-8.27929177e-04,  1.49843263e-03, -2.81216658e-04, ...,
         4.76823381e-03,  3.41646407e-03,  3.92185420e-03],
       [-5.93215518e-04,  1.07363470e-03, -2.01493183e-04, ...,
         3.41646407e-03,  2.44791410e-03,  2.81002872e-03],
       [-6.80968604e-04,  1.23245515e-03, -2.31299633e-04, ...,
         3.92185420e-03,  2.81002872e-03,  3.22571018e-03]])
```

[44]

```
#now we can also get the adjusted estimator of the variance hat
sgad = ((np.transpose(y).dot(y))- (np.transpose(beta).dot(xt).dot(y))) / (201 - 3 - 1)
sgad
```

```
array([[0.97422819]])
```

Statistical Data Analysis

Problem Sheet 3

sat sun mon tue wed thu fri Date: / /

(2)

(i) Here,

Least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T y$

and REML estimator $\hat{\sigma}^2_{ad} = \frac{1}{n-p-1} \hat{\epsilon}^T \hat{\epsilon}$

Now, $Cov(\hat{\beta}) = cov((X^T X)^{-1} X^T y)$ [Plugging $\hat{\beta}$ value] ...... (i)

We know, $Cov(Xy) = E[(Xy - E[Xy])(Xy - E[Xy])^T]$

$= E[X(y - E[y])(y - E[y])^T X^T]$

$= X E[(y - E[y])(y - E[y])^T] X^T$

$= X\, cov(y) X^T$ [putting value of cov(y)] .... (ii)

By using (ii) in (i), we get,

$cov(\hat{\beta}) = (X^T X)^{-1} X^T cov(y) [((X^T X)^{-1}) X^T]^T$

$= (X^T X)^{-1} X^T cov(y) X [(X^T X)^{-1}]^T$

$= (X^T X)^{-1} X^T X [(X^T X)^T]^{-1} Cov(y)$

$= (X^T X)^{-1} (X^T X) (X^T X)^{-1} cov(y)$

$= (X^T X)^{-1} Cov(y)$

$\therefore Cov(\hat{\beta}) = \sigma^2 (X^T X^{-1})$ [$Cov(y) = \sigma^2 I$]

(showed)

HEARTS

sat sun mon tue wed thu fri Date: / /

(ii)

Here, $\hat{\epsilon} = y - X\hat{\beta}$ [with respect to ϵ]

$$= y - X(X^T X)^{-1} X^T y$$

$$= y(I_n - X(X^T X)^{-1} X^T)$$

$$= (X\beta + \epsilon)(I_n - X(X^T X)^{-1} X^T)$$

$$= X\beta - X(X^T X)^{-1} X^T X\beta + (I_n - X(X^T X)^{-1} X^T)\epsilon$$

Now, $((X^T X)^{-1} X X^T) = I_{p+1}$

$$= X\beta - X\beta + (I_n - X(X^T X) X^T)\epsilon$$

$$= (I_n - X(X^T X) X^T)\epsilon \quad \text{[linear properties of } \epsilon\text{]}$$

$$= M\epsilon$$

Here $M = (I_n - X(X^T X)^{-1} X^T)$ is a symmetric and idempotent matrix.

$$\hat{\epsilon}^T \hat{\epsilon} = \epsilon^T M^T M \epsilon = \epsilon^T M \epsilon$$

Here we get a scalar. Using trace operator 'tr' we get,

$$E[\hat{\epsilon}^T \hat{\epsilon}] = E[\epsilon^T M \epsilon]$$

$$= E[tr(\epsilon^T M \epsilon)]$$

$$= E[tr(M \epsilon \epsilon^T)]$$

$$= tr(M E[\epsilon \epsilon^T])$$

$$= tr(M \sigma^2 I_n)$$

$$= \sigma^2 tr(M)$$

HEART'S

Now, plugging the value of M,

$$\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = \sigma^2 \, tr\left(I_n - X(X^TX)^{-1}X^T\right)$$

$$= \sigma^2 \left[tr(I_n) - tr\left(X(X^TX)^{-1}X^T\right)\right]$$

$$= \sigma^2 (n-p-1) \quad [\text{using } tr(A+B)] \quad \ldots \text{(i)}$$

So, $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{1}{n-p-1}\hat{\epsilon}^T\hat{\epsilon}\right]$

$$= \mathbb{E}\left[\frac{1}{n-p-1}\cdot\sigma^2(n-p-1)\right] \quad [\text{using (i)}]$$

$$= \sigma^2 \quad \text{(showed)}$$

(3) The corresponding statistical model is the Hardy - Weinberg model. It is used to compare gene frequencies in each population over a period of time. According to it's principle, at equilibrium, the gene or the allele frequencies within a population will be same in ~~each~~ all generations. A population of alleles must meet five rules to be considered "in equilibrium":

(1) No occurances of allele changes
(2) No migration in or out of the population.
(3) Random mating must occur.

sat sun mon tue wed thu fri Date: / /

(4) No genetic drift

(5) No natural selection.

environmental change may occur in this rules.

To solve the equilibrium question we need,

$$p+q=1$$

$$p^2+2pq+q^2=1$$

Here, p is the frequency of the dominant allele.

q is the frequency of the recessive allele.

p^2 is the frequency of individuals with the homozygous dominant genotype.

$2pq$ is the frequency of individuals with the heterozygous genotype.

q^2 is the frequency of individuals with the homozygous recessive genotype.

Now, Given, alleles are AA, Aa, aa.

Here, allele frequencies are equal to the products of the corresponding allele frequencies.

Thus, $P_{AA} = P_x = p^2$

$$P_{Aa} = P_y = 2P(1-P)$$

$$P_{aa} = P_z = (1-P)^2$$

HEART'S

sat sun mon tue wed thu fri Date: / /

Now, the likelihood of P is:

$$L(P) = \frac{n!}{x!\,y!\,z!} (p^2)^x [2p(1-p)]^y . [(1-p)^2]^z$$

The natural log likelihood of ~~L(P)~~ p is

$$l(P) = \ln(n!) - \ln(x!) - \ln(y!) - \ln(z!) + \ln(p^{2x})$$
$$+ \ln 2p^y + \ln(1-p)^y + \ln(1-p)^{2z}$$
$$= \ln(n!) - \ln(x!) - \ln(y!) - \ln(z!) + 2x\ln(p)$$
$$+ \ln(2) + y\ln(p) + y\ln(1-p) + 2z\ln(1-p)$$
$$\ldots\ldots \text{(i)}$$

Now, we will take a derivative respect to P.

$$\therefore \frac{d}{dp} l(P) = \frac{d}{dp}[2x\ln(p)] + \frac{d}{dp}[y\ln(p)] + \frac{d}{dp}[y\ln(1-p)]$$
$$+ \frac{d}{dp}[2z\ln(1-p)]$$
$$= 2x/p + y/p - y/(1-p) - 2z/(1-p)$$
$$= \frac{2x+y}{p} - \frac{2z+y}{1-p} \ldots\ldots \text{(ii)}$$

To estimate p, the derivative $= 0$.

$$\therefore \frac{2x+y}{p} - \frac{2z+y}{1-p} = 0$$

$$\text{or, } \frac{2x+y}{p} = \frac{2z+y}{1-p}$$

$$\text{or, } (1-p)(2x+y) = 2zp + py$$

$$\text{or, } 2x+y-2xp-py-py = 2zp$$

$$\text{or, } 2p(x+y+z) = 2x+y$$

$$\text{or, } p = \frac{2x+y}{2(x+y+z)} \quad [\because x+y+z=n]$$

$\therefore$ Maximum likelihood estimator for $p = \frac{2x+y}{2n}$.