Postgraduate Program: "Data Science and Information Technologies"
Course: Biostatistics

Postgraduate student: Eleni Panagiotopoulou
Registration number: 7115152300011

Exercise 1

A psychiatrist studied the effect of a combination of two drugs on treating a disease. Patients in the study received the drug A at one of two dose levels (low dose and high dose) and the drug B at one of three dose levels (no drug, low dose and high dose). Patients in the study were divided into groups so that each dose level combination of drug A and drug B corresponded to three patients.

After two weeks of drugs reception, their impact was measured by subjecting the patients of the study to a specially designed psychometric test. The performance to the test was measured on a scale of 0 to 50 and follows the normal distribution with the same variances for all cases. The measurements are listed in the table below:

		Drug B		
		No drug	**Low dose**	**High dose**
Drug A	**Low dose**	38 32 30	37 35 40	36 39 43
	High dose	40 45 36	39 42 46	39 48 47

Test at the 5% significance level, the validity of the following:

(a) The test performance is the same for both doses of drug A.

(b) The test performance is the same for all three doses of drug B.

(c) The change of dose of drug A does not alter the effect of drug B on psychometric test performance and vice versa, that is, the interaction between drug A and drug B is zero.

Answer

In this analysis, the combination of the of two drugs is examined in its potency to treat a disease. In this case, we need to simultaneously examine the effects of two independent variables (drug A and drug B) on a dependent variable (impact on treating the disease). The method of the two-way ANOVA will be employed in order to determine the impact of each drug and their combination to the treatment. During this experiment, drug A was administered in low dose and high dose in combination with drug B in low dose, high dose or total absence. Three patients were tested for each combination of the drugs A and B. Since the first variable (drug A) has two levels (low and high dose) and the second variable (drug B) has three levels (zero, low and high dose), the present analysis will be approached with *2x3 two-way ANOVA*.

For N-way ANOVA application, the assumption of independence must be satisfied which in our case it is given that there is no relationship either between the observations among the groups or between the observations of the same group. Also, normality and equality of variances are fulfilled. Three separate F-ratios will be calculated in order to determine the part of the variability of the dependent variable that can be attributed to either drug A, drug B or their combination. First, we insert our data into MATLAB.

X = [38 32 30 40 45 36; 37 35 40 39 42 46; 36 39 43 39 48 47]';

Since for each combination, 3 patients are examined, our repetitions equal three. The 2x3 two-way ANOVA in MATLAB:

[p,table, stats] = anova2(X, 3);

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	80.778	2	40.389	2.61	0.1149
Rows	150.222	1	150.222	9.69	0.009
Interaction	3.444	2	1.722	0.11	0.8957
Error	186	12	15.5		
Total	420.444	17			

Figure 1. ANOVA table generated by the above command.

(a) For determining if the test performance is the same for both doses of drug A, the mean values of the observations grouped by drug A must be equal. In this case, we should examine the p-value of the rows in the significance level of 0.05.

H_0: The mean values of the two different levels of drug A are equal ($\mu_{low} = \mu_{high}$).
H_1: The mean values are not equal. The test performance is not the same for both doses of drug A.

The p-value in this case equals 0.009 which is extremely lower than the significance level 0.05. The null hypothesis can be rejected in favor of the alternative one indicating that ***the test performance is not the same for both doses of drug A***.

(b) Now in the case of drug B, we will approach it with the same manner. In order for its test performance to be the same for all three administration levels, we examine the p-value of the columns.

H_0: The mean values of the three different levels of drug B are equal ($\mu_{zero} = \mu_{low} = \mu_{high}$).
H_1: The mean values are not equal. The test performance is not the same for both doses of drug B.

The p-value for the analysis is 0.1149 which is higher than the significance level of 0.05. The null hypothesis cannot be rejected; thus ***the test performance is the same for all three cases of drug B.***

(c) Now, we will investigate if the change of dose of drug A does not alter the effect of drug B on psychometric test performance and vice versa, that is, the interaction between drug A and drug B is zero. The p-value of the interaction equals to 0.8957 >> 0.05 which is the significance level. This suggests that the evidence does not suffice to reject the null hypothesis, that there is no interaction between the two factors. Essentially, ***the change of dose of drug A does not alter the effect of drug B on psychometric test performance and vice versa.***

Exercise 2

Construct the Life Table of a population, which is characterized by the following:

a. The initial number of people (the root of the Life Table) is equal to 60,000.

b. The percentage p_x of individuals aged x who survive in their (x + 1)-th birthdays is provided by the following table:

X	p_x
0	0.9906
1	0.9908
2	0.9915
3	0.9922
4	0.9930
5	0.9940
6	0.9960
7	0.9980
8	0.9984
9	0.9988
10	0.9992

Construct the Life Table for one-year steps.

Answer

The life table records the mortality pattern in respect to age for a population and, consequently, it provides a foundation for calculating life expectancy at different ages. First, the definitions of every symbol that will be included in the Life Table will be provided as well as their way of calculation.

Table 1. Definition of the symbols used in the Life Table and how they can be calculated.

Symbol	*Definition*	*Calculation*
p_x	Percentage of individuals aged x that survive in their (x+1)-th birthdays	[given]
q_x	Percentage of individuals aged x that do not survive in their (x+1)-th birthdays	$q_x = 1 - p_x$
l_x	Number of individuals aged x that survive in their x-th birthday	$l_x = l_{x-1} \bullet p_{x-1}$
d_x	Number of individuals that die between their x-th and their (x+1)-th birthdays	$d_x = q_x \bullet l_x$
L_x	Number of individuals' years lived between their x and (x+1) birthdays	$L_x = l_x - ½\, d_x$
T_x	Number of individuals' years lived after the x-th birthday	$T_x = L_x + L_{x+1} + \ldots + L_n$
e_x	Number of additional years that on average live those who survived on their x-th birthday (life expectancy left on birthday x).	$e_x = T_x / l_x$

Considering the table above, it is now feasible to fill in the Life Table for the percentage given since it is provided that the radix is equal to 60000 people.

Table 2. Presented below is the Life Table featuring an initial population of 60,000 individuals and corresponding initial percentages as provided from the beginning. The table was constructed using Excel.

X	l_x	d_x	p_x	q_x	L_x	T_x	e_x
0	**60000**	564	0.9906	0.0094	59718	634896	10.58
1	59436	547	0.9908	0.0092	59163	575178	9.68
2	58889	501	0.9915	0.0085	58639	516015	8.76
3	58388	455	0.9922	0.0078	58161	457376	7.83
4	57933	406	0.993	0.007	57730	399215	6.89
5	57527	345	0.994	0.006	57355	341485	5.94
6	57182	229	0.996	0.004	57068	284130	4.97
7	56953	114	0.998	0.002	56896	227062	3.99
8	56839	91	0.9984	0.0016	56794	170166	2.99
9	56748	68	0.9988	0.0012	56714	113372	2
10	56680	45	0.9992	0.0008	56658	56658	1

Exercise 3

A study involved 74 patients who were in one of the three stages of the development of a specific disease. The study included the measurement of six clinical parameters denoted by X_1, X_2, ..., X_6. The results of the measurements are recorded in the following table:

α/α	Stage of disease	X_1	X_2	X_3	X_4	X_5	X_6
1	1	22	3	2	2.5	11	3.58
2	1	17	3	1	3.0	11	2.53
3	1	22	0	0	3.0	12	3.08
4	1	17	5	2	3.0	15	3.20
5	3	23	3	3	2.5	11	3.70
6	3	25	4	4	2.5	12	3.64
7	1	20	3	3	4.5	16	2.93
8	1	15	4	4	4.0	20	2.41
9	1	18	3	4	4.0	21	2.73
10	1	26	0	0	3.0	10	2.87
11	1	20	3	3	2.0	16	2.93
12	1	16	3	4	3.5	17	2.93
13	1	19	3	3	3.5	13	3.08
14	1	14	3	3	4.0	20	2.28
15	1	14	2	2	3.5	16	2.19
16	1	21	3	3	3.0	13	2.24
17	1	29	3	3	2.5	9	2.93
18	1	16	4	4	4.0	20	2.56
19	1	22	3	3	3.5	17	2.73
20	1	22	2	3	2.0	16	2.73
21	1	24	2	2	2.0	7	2.73
22	1	19	3	3	3.5	13	2.56
23	2	23	4	3	1.5	6	3.89
24	2	35	5	5	2.0	8	3.70
25	2	24	4	4	2.5	8	3.54
26	2	21	4	4	2.5	8	3.55
27	2	30	5	4	2.0	8	3.54
28	1	18	2	2	4.0	17	2.47
29	1	16	2	2	3.5	16	2.71
30	1	17	2	2	4.5	21	2.94
31	3	21	3	1	2.5	16	3.37
32	1	28	4	0	1.5	9	3.15
33	1	21	3	3	2.0	10	3.08
34	2	25	5	5	3.0	10	3.05
35	2	28	4	4	2.5	5	3.30
36	1	12	3	4	3.5	22	2.47
37	1	12	3	4	2.5	18	2.47

38	1	14	3	3	3.5	15	2.47
39	1	30	4	4	3.5	11	3.73
40	1	22	4	3	3.0	9	2.73
41	1	14	4	3	3.5	16	2.75
42	1	14	4	4	3.0	16	2.75
43	1	15	3	2	3.5	23	2.26
44	1	18	3	0	3.0	15	2.43
45	1	20	3	3	3.5	17	3.08
46	1	21	4	4	4.0	20	2.41
47	1	19	3	3	4.5	16	2.93
48	1	19	3	4	2.0	16	2.93
49	1	18	4	4	4.0	20	2.73
50	1	19	3	3	4.5	14	3.08
51	1	24	1	1	2.0	10	2.73
52	1	16	3	3	3.5	17	2.41
53	3	14	0	0	3.5	14	3.58
54	1	28	3	3	2.0	11	3.05
55	1	34	5	4	2.5	11	2.97
56	1	25	3	0	4.0	17	3.37
57	1	26	0	0	1.5	8	3.54
58	1	18	2	2	5.0	16	3.23
59	1	18	4	4	4.0	20	2.73
60	1	18	1	2	1.5	7	3.08
61	1	19	3	3	2.0	16	2.93
62	1	19	3	3	3.5	17	2.93
63	1	19	0	0	3.5	13	3.08
64	1	24	2	2	2.0	7	2.73
65	3	26	3	3	3.0	10	3.72
66	2	35	5	4	2.5	11	3.81
67	2	18	5	5	2.5	14	3.06
68	2	31	5	5	3.0	9	3.21
69	2	18	5	5	2.0	11	3.05
70	3	25	4	3	3.0	15	3.78
71	3	41	5	4	3.0	15	3.78
72	3	25	4	3	2.0	16	3.78
73	3	23	4	3	2.5	12	3.74
74	3	17	5	3	2.5	14	2.98

(a) Perform a Principal Component Analysis (PCA) using variables X_1, X_2 ,..., X_6, to determine:

i. The smallest number of Principal Components that describe at least 85% of the total variability.

ii. The coordinates of each of the Principal Components of the question i. with respect to the initial variables.

iii. The coordinates of the data in the above table in the coordinates system defined by the Principal Components of question i.

It is given that the assumptions required to perform PCA are valid.

(b) Derive the dependent variable Y from the values of the variables in the above table using the relation:

$$Y= 18 + X_1 – 3*X_2 +2 *X_3 – 5*X_4 +4*X_5 – 7*X_6$$

Next, find the coefficients of the multiple linear regression model of the variable Y taking as independent variables the Principal Components of the question A.i. It is given that the assumptions required for the application of the multiple linear regression model are valid.

Answer

Principal Component Analysis is a common practice used to reduce the dimensionality of the data. In this case, we are given 6 different variables for 74 patients that are in different stages of the same disease. PCA is a tool that discovers the redundancy in the multivariate data and is able to extract the relationships between the variables achieving dimension reduction without significant loss of information. The method primarily aims to identify the directions in the data with the highest variability. Subsequently, it projects the data onto these directions and seeks to minimize the sum of squared distances between the original points and their projections onto the principal components. Naturally, in order to perform this technique, certain assumptions considering the data, must be met. These include that the variables involved must follow the multivariate normal distribution while at the same time the relationship between each pair of variables must be linear. In addition, the ratio of the number of the observations to the number of variables must be large. Ideally, outliers are excluded from the analysis. Our data fulfill all the previous assumptions.

First, we ensure that the scale differences are minimized between the variables by standardizing them. This is possible by dividing each column by its standard deviation. Our data are represented by the *X* variable. The code used for standardizing follows.

```
data_stand = [X(:,1)./std(X(:,1)), X(:,2)./std(X(:,2)), X(:,3)./std(X(:,3)), X(:,4)./std(X(:,4)),
              X(:,5)./std(X(:,5)), X(:,6)./std(X(:,6))];
```

The standardized data will be the input for the PCA analysis. The MATLAB *correlation()* and *eig()* functions will be employed.

```
correlation=corr(data_stand);
[eigenvectors, eigenvalues] = eig(correlation);
```

```
eigenvectors =

   -0.5109   -0.0045   -0.3646   -0.6550    0.2620    0.3291
   -0.2357   -0.6473    0.0613    0.0270   -0.6664    0.2773
   -0.1384   -0.6798    0.2007    0.0207    0.5596   -0.4060
    0.4406   -0.2060   -0.6674   -0.2779   -0.2125   -0.4423
    0.4955   -0.2721   -0.2387    0.2259    0.3544    0.6684
   -0.4739    0.0494   -0.5662    0.6645    0.0586   -0.0864
```

Figure 2. The eigenvectors of the data.

```
eigenvalues =

    2.6537         0         0         0         0         0
         0    1.7333         0         0         0         0
         0         0    0.7037         0         0         0
         0         0         0    0.3877         0         0
         0         0         0         0    0.2270         0
         0         0         0         0         0    0.2947
```

Figure 3. Data's eigenvalues.

```
correlation =

    1.0000    0.2893    0.1301   -0.4094   -0.5798    0.6137
    0.2893    1.0000    0.7403   -0.0802   -0.0116    0.2077
    0.1301    0.7403    1.0000    0.0103    0.0716    0.0590
   -0.4094   -0.0802    0.0103    1.0000    0.6601   -0.3690
   -0.5798   -0.0116    0.0716    0.6601    1.0000   -0.5055
    0.6137    0.2077    0.0590   -0.3690   -0.5055    1.0000
```

Figure 4. Correlation matrix between the six variables. Each strong correlation between the variables is represented by a different color.

At a first glance, the correlation matrix suggests five strong correlations. We focus on the absolute values that are also greater than 0.5.

Generally, the sum of eigenvalues is equal to the number of variables, which is 6 in this case. This relationship is expressed mathematically as the sum of the diagonal elements of the covariance matrix being equal to the number of variables. The same results of the eigenvalues can be obtained and summarized by the *pca()* function, stored in a variable named 'latent'.

```
[coeff, score, latent] = pca(data_stand);
```

```
latent =

    2.6537
    1.7333
    0.7037
    0.3877
    0.2947
    0.2270
```

```
coeff =

    0.5109    0.0045    0.3646   -0.6550    0.3291   -0.2620
    0.2357    0.6473   -0.0613    0.0270    0.2773    0.6664
    0.1384    0.6798   -0.2007    0.0207   -0.4060   -0.5596
   -0.4406    0.2060    0.6674   -0.2779   -0.4423    0.2125
   -0.4955    0.2721    0.2387    0.2259    0.6684   -0.3544
    0.4739   -0.0494    0.5662    0.6645   -0.0864   -0.0586
```

Figure 5. The variables 'latent' and the 'coefficient' of the aforementioned command. The latent variable includes the eigenvalues sorted from the biggest to the smallest quantity.

(a) (i)

The variable 'latent' stores the eigenvalues associated with each principal component, aligning with the respective columns of the 'coeff' matrix. According to the results, if only the first component is kept, we have retained the 2.6537/6 = **44.22%** of the initial variability. In case we also include the second component, we gather the (2.6537+1.7333)/6 = **73.12%** of the initial variability. If the third component is also added, we end up retaining (2.6537+1.7333+0.7037)/6 = **84.84%**. In order to be precise, we won't round the percentage here since the query demands at least 85%. By adding one more component, (2.6537+1.7333+0.7037+0.3877)/6 = **91.31%** of the total variability is explained. In this case, the least number of components that describe at least 85% of the total variability is 4. By using the following piece of code, we can calculate in MATLAB the percentage of total variability explained by each element.

```
lat = latent';
percentage = 100*lat/(sum(lat));
```

The output is:

```
percentage =

   44.2275   28.8875   11.7290    6.4618    4.9116    3.7825
```

Figure 6. A way to corroborate our previous calculations. By adding the variability of the principal components (while it is organized in descending order) we are able to show that the first 4 components explain 91.31% of the total variability (44.2275+28.8875+11.7290+6.4618=91.3058%)

(ii)

The coefficients of all components are summarized in the 'coeff' table, where every column corresponds to a principal component. The principal components are constructed as linear combinations of the original four variables, resulting in the creation of new variables. Therefore, for each Principal Component mentioned in sub-question i, there exist specific linear combinations with respect to the initial variables that define them.

$$pc_1 = 0.5109 * X_1 + 0.2357 * X_2 + 0.1384 * X_3 - 0.4406 * X_4 - 0.4955 * X_5 + 0.4739 * X_6$$

$$pc_2 = 0.0045 * X_1 + 0.6473 * X_2 + 0.6798 * X_3 + 0.2060 * X_4 + 0.2721 * X_5 - 0.0494 * X_6$$

$$pc_3 = 0.3646 * X_1 - 0.0613 * X_2 - 0.2007 * X_3 + 0.6674 * X_4 + 0.2387 * X_5 + 0.5662 * X_6$$

$$pc_4 = -0.6550 * X_1 + 0.0270 * X_2 - 0.0207 * X_3 - 0.2779 * X_4 + 0.2259 * X_5 + 0.6645 * X_6$$

(iii)

The variable 'score' contains the data in the new coordinate system defined by the principal components and thus, it constitutes the representation of our data in the space of the main components. Essentially, it has the same size as the initial data table. This means that the rows of the variable 'score' correspond to the observations, while the columns correspond to the

components. So in order to determine the task given, we only need to take a look into the 'score' variable.

X =

22.0000	3.0000	2.0000	2.5000	11.0000	3.5800
17.0000	3.0000	1.0000	3.0000	11.0000	2.5300
22.0000	0	0	3.0000	12.0000	3.0800
17.0000	5.0000	2.0000	3.0000	15.0000	3.2000
23.0000	3.0000	3.0000	2.5000	11.0000	3.7000
25.0000	4.0000	4.0000	2.5000	12.0000	3.6400

score =

1.1039	-0.8715	0.3410	0.7433	-0.0226	0.2616
-0.8003	-1.1412	-0.7400	-0.4108	-0.0699	1.1643
-0.5521	-3.1465	0.6096	-0.1959	-0.0802	-0.3569
-0.0950	0.5503	0.0770	0.8405	0.5597	1.3680
1.4201	-0.3808	0.4054	0.8214	-0.2890	-0.2132
1.7036	0.6973	0.3163	0.5962	-0.0919	-0.2748

Figure 7. The truncated outputs of the 'X' (initial data) and 'score' variables. Since the observations are 74, this snapshot shows the first 6 observations in order to save space. We can focus on the first observation of both matrices so that we can see how it was initially recorded and how it is represented after principal component analysis.

(b) In this part we will derive the dependent variable Y from the values of the variables in the above table using the relation:

$$Y= 18 + X_1 – 3*X_2 +2 *X_3 – 5*X_4 +4*X_5 – 7*X_6$$

Hence, the analysis entails multiple linear regression, constructing a model that incorporates multiple independent variables. This model comprises a set of partial regression coefficients. In order to calculate Y with this manner, the following line was used (*X* is our initial data):

```
Y = 18 + X(:,1) - 3*X(:,2) + 2*X(:,3) - 5*X(:,4) + 4*X(:,5) - 7*X(:,6);
```

The Y values of the output are shown here:

[41.4400 39.2900 51.4400 46.6000 43.6000 49.0200 55.9900 72.1300 79.8900 48.9100

68.4900 62.9900 46.9400 73.0400 61.1700 57.3200 46.9900 72.0800 68.3900 74.8900

38.8900 50.5800 24.2700 44.1000 32.7200 29.6500 38.2200 64.7100 59.5300 73.9200

59.9100 40.4500 44.4400 41.6500 26.4000 82.2100 71.2100 54.2100 44.3900 35.8900

53.2500 57.7500 86.6800 54.9900 63.9400 78.1300 54.9900 69.4900 72.8900 45.9400

51.8900 64.6300 45.4400 55.6500 55.7100 58.4100 43.7200 50.3900 72.8900 35.9400

67.4900 63.9900 49.9400 38.8900 39.9600 50.8300 53.0800 42.5300 43.6500 55.5400

70.5400 64.5400 44.3200 48.6400];

Now, the coefficients of the multiple linear regression model will be calculated of the variable Y taking as independent variables the Principal Components of the question (a)(i). It is given that the assumptions required for the application of the multiple linear regression model are valid.

In the initial phase of multiple linear regression analysis, the first step involves creating graphs depicting the dependent variable against each independent variable. Subsequently, principal components (pc1, pc2, pc3, pc4) are calculated as linear combinations of the initial variables X1, X2, X3, X4, X5 and X6). The MATLAB *plotmatrix()* function is then employed to visually assess whether a linear relationship exists between each independent variable and the dependent variable. The observed relationships appear to be linear, in alignment with the expectations from the exercise. Consequently, multiple regression analysis is deemed suitable for application. This is executed using the regress function in MATLAB, which provides the least squares line of the dependent variable Y based on the multidimensional space of independent variables. It's noteworthy that the fixed term was previously incorporated in the table of independent variables. The regress function output includes, among other elements, a column vector 'c' containing the coefficients of the regression line. Additionally, the 'stats' variable holds information such as the R^2 value, along with F and p-values from the analysis of variance (ANOVA).

```
pc1 = 0.5109*X(:,1) + 0.2357 * X(:,2) + 0.1384 * X(:,3) - 0.4406 * X(:,4) - 0.4955 * X(:,5) + 0.4739 *
X(:,6);
pc2 = 0.0045 * X(:,1) + 0.6473 * X(:,2) + 0.6798 * X(:,3) + 0.2060 * X(:,4) + 0.2721* X(:,5) - 0.0494 *
X(:,6);
pc3 = 0.3646 * X(:,1) - 0.0613 * X(:,2) - 0.2007 * X(:,3) + 0.6674 * X(:,4) + 0.2387 * X(:,5) + 0.5662 *
X(:,6);
pc4 = -0.6550 * X(:,1) + 0.0270 * X(:,2) - 0.0207* X(:,3) - 0.2779 * X(:,4) + 0.2259 * X(:,5) + 0.6645 *
X(:,6);
```

```
PC=[pc1 pc2 pc3 pc4];

dt=[ones(size(PC(:,1))) PC];
[c, cint, r, rint, stats] = regress(Y, dt);
```

By displaying the components of the vector ‘c’, we are able to construct the multiple linear regression model:

vector c: [14.2605 -5.4494 1.9004 0.5353 -4.7615]

The model that describes our data taking into account Principal Component Analysis:

Y = 14.2605 – 5.4494*pc1 + 1.9004*pc2 + 0.5353*pc3 – 4.7615*pc4

Exercise 4

The measurements in a sample of 20 healthy men in order to find the relationship between age and triglycerides concentration are shown in the following table (independent measurements):

Age (in years)	Triglycerides concentration (mgr/lt)	Age (in years)	Triglycerides concentration (mgr/lt)
12	28	**43**	83
14	53	**46**	95
18	107	**49**	112
24	88	**52**	126
26	91	**58**	84
29	61	**62**	120
32	98	**63**	195
35	80	**66**	166
38	130	**68**	153
40	53	**69**	92

(a) Plot the graph of the data points (triglycerides concentration with respect to age).

(b) On the above graph, plot, based on the criterion of least squares, the line that "fits" best to the data points. What is the slope of this line and at what point does it intersect the vertical axis?

(c) How much of the triglycerides concentration variability is explained by the above line?

(d) Use the above least squares line to calculate the triglycerides concentration predictions for the following ages: 5, 35, 65, 95.

(e) Can we reject the hypothesis that the R^2 for the population from which the above sample comes is zero, at a significance level of 5%?

Hint: Present analytically the assumptions needed to apply the linear regression and test their validity.

Answer

The goal of the analysis is to assess the relationship between the two quantitative variables of age and triglyceride concentration. This indicates that this investigation is based on theoretical knowledge, and it does not concern a correlation but a regression problem. Commencing, the visualization of the data is imperative in order to evaluate if a linear relationship is possible between our two variables in question. The problem will be approached using MATLAB. The data are introduced:

```
age = [12, 14, 18, 24, 26, 29, 32, 35, 38, 40, 43, 46, 49, 52, 58, 62, 63, 66, 68, 69]';

trigl_conc = [28, 53, 107, 88, 91, 61, 98, 80, 130, 53, 83, 95, 112, 126, 84, 120, 195, 166, 153, 92]';
```

(a) A graph of the two variables is plotted, the independent variable corresponds to the age of the subject while the dependent variable corresponds to its triglyceride concentration.

scatter(age, trigl_conc, 'blue', 'filled');

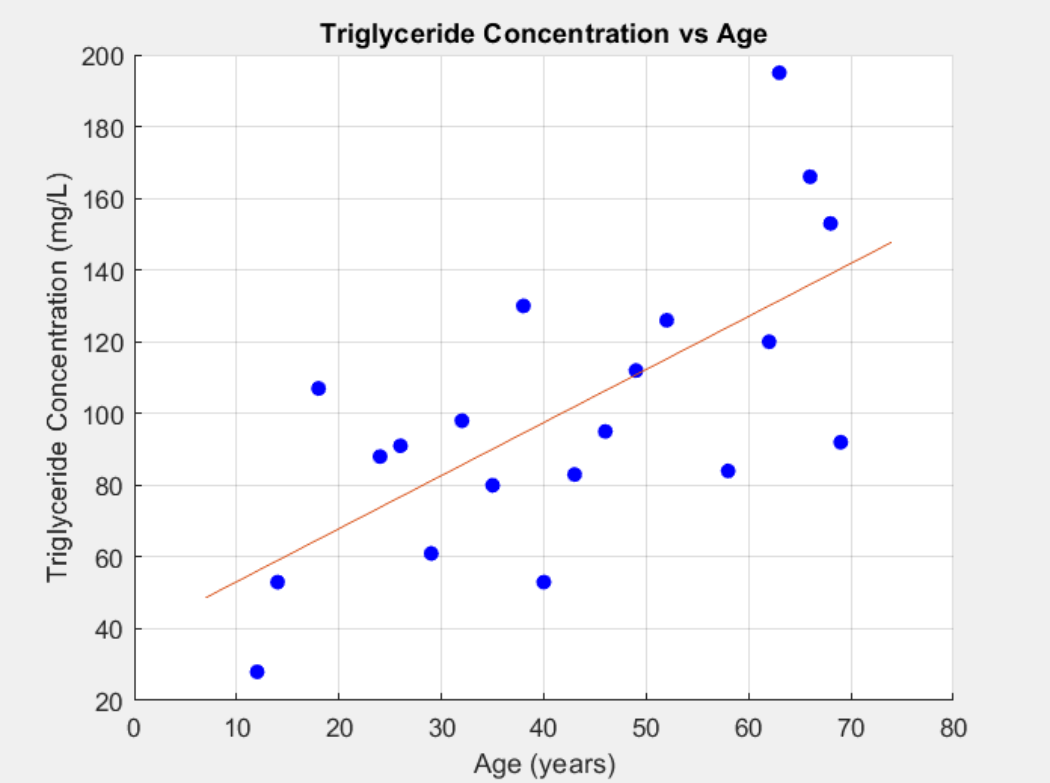

Figure 8. A scatterplot of the variables. We can deduce that a linear relationship is possible, however it appears to be somewhat weak. The following analysis will help us reach a well-funded conclusion.

(b) The next step suggests finding the line that best fits the data, based on the least-squares method. For this purpose, the MATLAB function *regress* will be used.

```
X = [ones(size(age)) age];
C = regress(trigl_conc, X);
```

The output of this command is C = [38.3293 1.4792]; indicating that the regression line is the following.

$$y = 38.33 + 1.48x$$

The slope of the line is 1.48 while and it intersects the vertical axis at 38.33.

The application of the least-squares method, however, presupposes the facts that the dependent variable follows a normal distribution and that the variance of the distribution of the dependent variable is the same for all values of the independent variable. The assumptions that all observations are independent and that the relationship between the variables is linear are fulfilled.

The validity of the assumptions will be tested using residual analysis. Essentially, the distribution of residuals and their relationship to the variables is examined in order to deduce

if there are any deviations from the assumptions that must hold in order to apply linear regression. The residuals represent the differences between the observations of the dependent variable and the values y_{est} calculated using the regression model from above. If they follow the normal distribution and they have the same variance for all values of the independent variable, the assumptions we examine hold. We will use the standardized residuals in order to check for normality and the studentized residuals to check for the data's variance. The calculation of both can be done using the MATLAB function *regstats*.

```
mystats = regstats(trigl_conc, age, 'linear', {'standres', 'studres'});
```

A qqplot of the standardized residuals will give away whether they follow a normal distribution or not.

```
qqplot(mystats.standres);
```

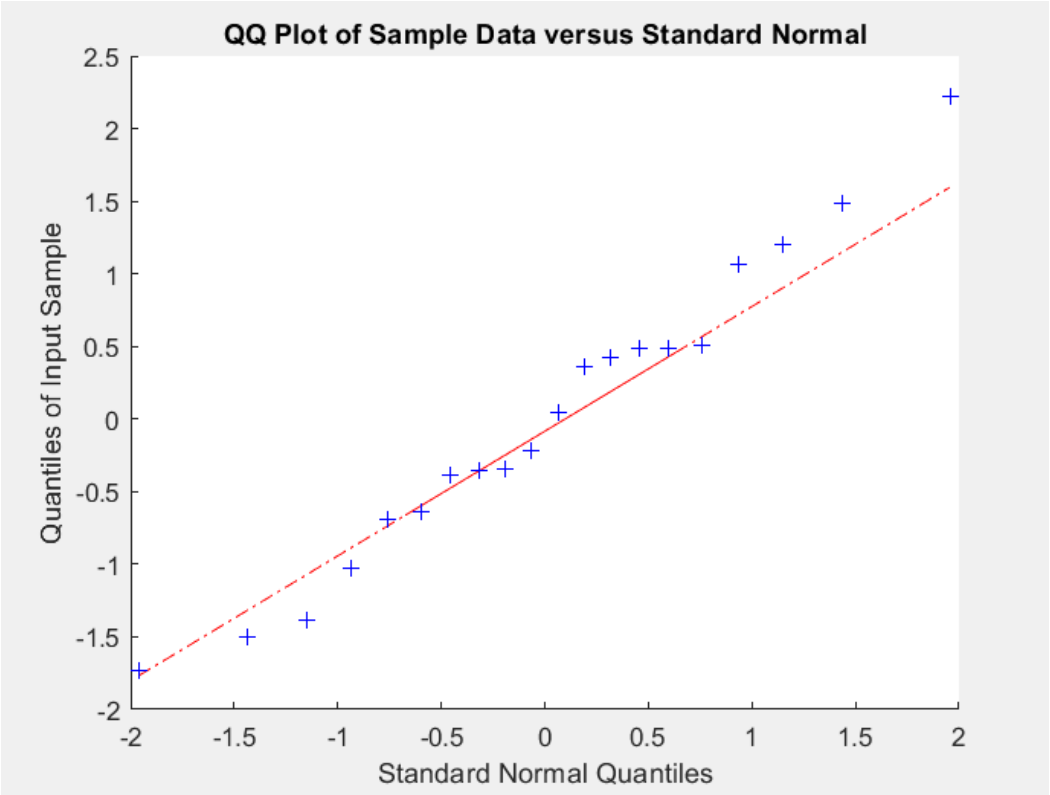

Figure 9. A qqplot of standardized errors. It is obvious that the points are almost on the same line, indicating that a normal distribution is followed.

The assumption of normality seems to be satisfied. We should also assess that data's variance complies with the initial assumption. In this case the estimated values are calculated by multiplying the predictor matrix X to the vector C.

```
yest = X * C;
```

Now, we will create a plot of the studentized residuals against the estimated values of y (yest). This will essentially help us understand whether the variance of the dependent variable is the same for all values of the independent one.

```
scatter(yest, mystats.studres);
```

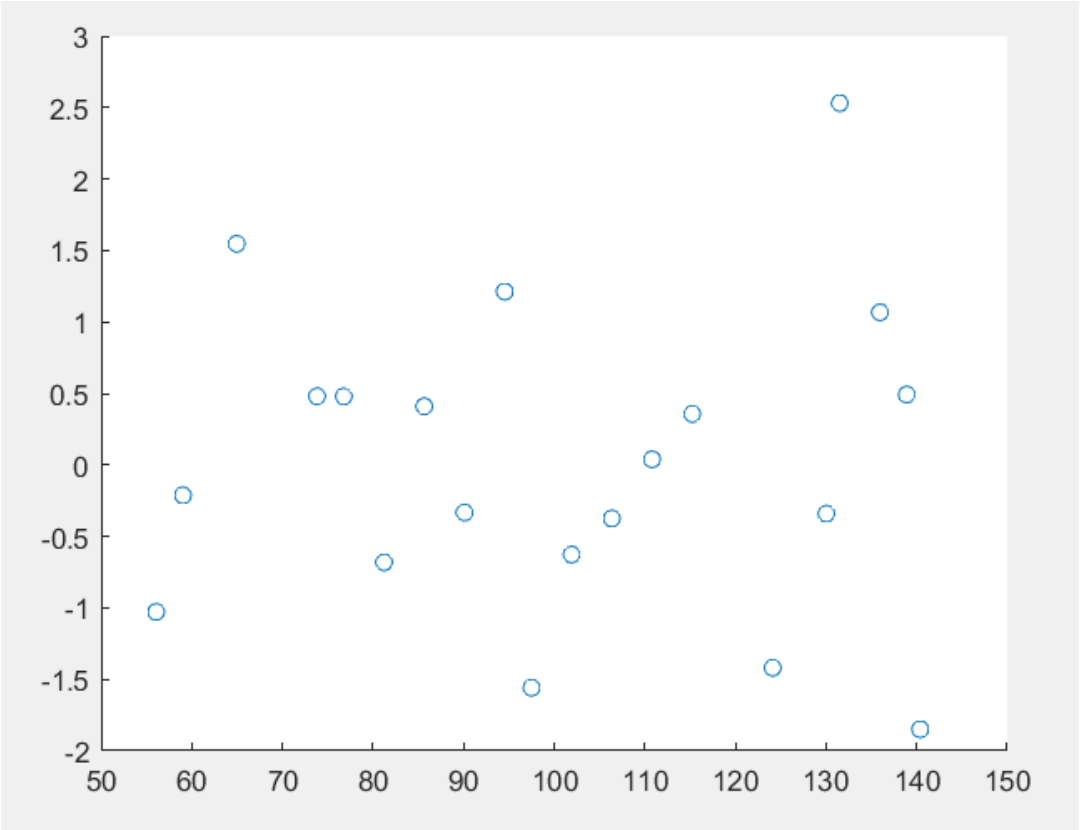

Figure 10. The scatterplot. The output of the command appears to support the equality of variances of the dependent variables.

In the scatterplot it is observed that the variances do not deviate from the value 0, while their values are very close. It can be concluded that the assumption for equalities of variance is also satisfied.

(c) A statistic metric that shows the percentage of the variability of the dependent variable explained by the regression model is R^2. By using the following command, the R^2 will correspond to the first value of the 'stats' variable.

```
[c, cint, r, rint, stats] = regress(trigl_conc, X);
```

The analysis shows that R^2=0.4571, so 45.71% of the variability of the dependent variable is justified by the regression model. This value is highly inadequate as it is implied that the 54,29% of the variability of the triglyceride's concentration is still inexplicable.

(d) In this query, our regression model will be used for predictions in order to determine the expected triglyceride concentration (mg/L) of the ages 5, 35, 65, 95.

```
Xtest = [5, 35, 65, 95];
```

By using the following command, essentially inserting each value we want to test to our regression model, the result matrix will give the predicted value for each one of the wanted ages.

```
yest_test = [ones(size(Xtest)) Xtest] * C;
```

Table 3. Table summarizing the results.

Age (years)	Predicted value of triglyceride concentration (mg/L)
5	45.7
35	90.1
65	134.5
95	178.8

We should note at this point that the age values 5 and 95 are out of limits for our estimated model since the max age value of our dataset is 69 while its minimum value is 12. In this case, our model is not trustworthy and precise outside of bounds and it is possible to yield inaccurate results.

(e) ANOVA is used in order to test the null hypothesis.

H_0: The R^2 for the population of which the sample comes is zero, at a significance level of 5% ($b = 0$).
H_1: The R^2 is not zero ($b \neq 0$).

The p-value of this is already given by the *regress* function in MATLAB and it corresponds to the third value of the vector 'stats'. The results indicate that the p-value is equal to 0.0011 that is less than 0.05, so we can reject the null hypothesis in favor of the alternative. The R^2 of our regression model is not zero which means that our model has some explanatory power in describing the relationship between the age and the triglycerides concentration (mg/L) of men.

Exercise 5

The following was recently published on a news website:

"A vaccine for Alzheimer's disease gives encouraging results to mice"

Researchers from Laval University in Quebec in collaboration with scientists from a well-known pharmaceutical company have found a way to activate the brain's natural defense mechanisms against Alzheimer's disease. This major breakthrough presented in the journal "Proceedings of the National Academy of Science" paves the way for the development of effective therapies, even a preventive vaccine for this neurodegenerative disease that is thought to be the epidemic of the 21st century due to the aging of the world population.

The "guilty" beta-amyloid

One of the main features of Alzheimer's disease is the production in the brain of a toxic molecule known as beta-amyloid. B-amyloid is deposited in plaques on the patient's brain. The microglial cells, the "defense" line of the nervous system, are unable to eliminate it.

Now the Laval University research team, led by Serge Rivest, Professor of Medical Science, in collaboration with specialists from a well-known pharmaceutical company, has identified a molecule that activates microglial cells. The molecule called MPL (monophosphoryl lipid A) has long been used as an adjuvant in vaccines and has been shown to be safe.

High efficiency of the molecule

Experiments in mice with Alzheimer's disease showed that weekly injections of MPL over a 12-week period "eliminated" up to 80% of the beta-amyloid plaques from the animal's brain. In addition, specific tests measuring the learning ability of experimental animals (by performing new tasks) showed a significant improvement in their cognitive ability after receiving treatment.

Therapeutic or preventive vaccine

The researchers believe that the MPL molecule could have two potential uses: It could be used first by intramuscular injection to Alzheimer's patients with the aim of "braking" the progression of their disease. Secondly, it could be one of the "components" of a vaccine designed to trigger the production of antibodies against beta-amyloid.

"The vaccine could be given in therapeutic form - to people already suffering from Alzheimer's disease to boost their physical immunity," Professor Rivest said, adding that "it could also be given preventively to people with increased risk factors for Alzheimer's disease."

Professor concluded that "when our team started working on Alzheimer's disease ten years ago our goal was to develop better treatments for Alzheimer's patients. The new discovery shows us that we are rather close to our goal."

Describe the way you think the above study was organized and carried out. In particular, analyze the following:

a. Which hypotheses were tested by the researchers?

b. What data were collected and used to test each one of the above hypotheses (question a.)?

c. Which methods were used to test each one of the above hypotheses? Justify your answer.

d. Which were the results of the application of the above methods according to the aforementioned summary?

Answer

In the present study, researchers found that MPL (monophosphoryl lipid A) is a substance able to activate the phagocytotic ability of microglial cells against the non-desirable beta-amyloids that form plaques on Alzheimer's patients' brains. It should be noted that MPL is a chemically detoxified lipid A moiety derived from *Salmonella minnesota* R595 LPS. This is important because LPS is a major component present in the outer surface membrane to almost all Gram-negative bacteria (like *Salmonella sp.)*. Given its origin, it's a known strong stimulator of innate or natural immunity for the human immune system.

Researchers were based on previous tested information that beta-amyloids deposited in plaques are responsible for memory deterioration and patient decline. Aiming for their destruction essentially suggests an improvement of life quality of patients. Also, since the MPL molecule has been used in adjuvants, the practice to administrate it to patients is known to be safe.

(a) There are five hypotheses that were tested by the researchers:

(i) The MPL molecule has the ability to activate microglial cells.

As the article mentions:

"Now the Laval University research team, led by Serge Rivest, Professor of Medical Science, in collaboration with specialists from a well-known pharmaceutical company, **has identified a molecule that activates microglial cells**. The molecule called MPL (monophosphoryl lipid A) has long been used as an adjuvant in vaccines and has been shown to be safe."

(ii) MPL can eliminate beta-amyloid plaques from mice's brain.

"**Experiments in mice with Alzheimer's disease showed that weekly injections of MPL over a 12-week period "eliminated" up to 80% of the beta-amyloid plaques from the animal's brain**."

(iii) MLP improves the cognitive ability of experimental animals (mice) after receiving treatment.

"In addition, **specific tests measuring the learning ability of experimental animals (by performing new tasks) showed a significant improvement in their cognitive ability after receiving treatment.**"

(iv) MLP can be given as vaccine to boost their physical immunity.

"The researchers believe that the MPL molecule could have two potential uses: **It could be used first by intramuscular injection to Alzheimer's patients with the aim of "braking" the progression of their disease**. Secondly, it could be one of the "components" of a vaccine designed to trigger the production of antibodies against beta-amyloid."

"**The vaccine could be given in therapeutic form - to people already suffering from Alzheimer's disease to boost their physical immunity**," Professor Rivest said, adding that "it could also be given preventively to people with increased risk factors for Alzheimer's disease."

Now each of these hypotheses will be analyzed in respect to (b) what data are collected and used in order to test them, (c) which methods are used to test them and (d) what are the results of the above methods' application according to the summary.

Hypothesis i: "The MPL molecule has the ability to activate microglial cells."

In this context we would need data from microglial cells and we should measure the activation of specific factors that take part in immune responses. There are two groups of cells. The control group (A) and the group that MPL is administered to (B). Then the factors are measured for each cell group. The statistical analysis that could be used is the ttest for mean difference between the two groups.

H_0: There is no significant difference between the activation levels of the factors between A and B.

H_1: There is significant difference between the activation levels of the factors in the groups A and B.

It seems that the null hypothesis can be rejected in a significance level of 0.05 (or 0.1) in favor of the alternative hypothesis showing that MPL has the ability to activate microglial cells.

Hypothesis ii: "MPL can eliminate beta-amyloid plaques from mice's brain."

After the first hypothesis that concerned the activation of specific cells, now it should be tested in vivo in order to assess its effectiveness inside an organism. In this experiment, there is the need of one group of mice with Alzheimer's disease. In this case, we measure the quantity of beta-amyloids in the initial stage, before administering MPL. The statistical method that could be used is the ttest again. After 12-weeks of MPL injection the same group is measured for the concentration of beta-amyloids in its brain.

H_0: There is no significant difference in the concentration of beta-amyloids in the brains of mice with Alzheimer's before and after 12 weeks of MPL injection.

H_1: There is significant difference in the concentration of beta-amyloids in the brains of mice with Alzheimer's before and after 12 weeks of MPL injection.

After rejecting the null hypothesis in a significance level of 0.05 or 0.1, the percentage reduction could be calculated since the difference is now significant, via the following formula :

([initial concentration] – [final concentration] / [initial concentration]) * 100

Hypothesis iii: "MLP improves the cognitive ability of experimental animals (mice) after receiving treatment."

In the same manner as before, one group of mice with Alzheimer's is needed. Now cognitive ability is measured and "quantified" via specific tests before the treatment. After the treatment the cognitive ability is again measured. The statistical method to be used in this context is the ttest. The two hypotheses concerning the ttest of comparison between two dependent groups:

H_0: There is no significant difference in the cognitive abilities of mice with Alzheimer's disease before and after the MPL treatment.

H_1: There is significant difference in the cognitive abilities of mice with Alzheimer's disease before and after the MPL treatment.

The null hypothesis would be rejected since it is known to us that MPL increases the cognitive ability after treatment.

Hypothesis iv: "MLP can be given as vaccine to boost their physical immunity."

This hypothesis is based on hypothesis ii. It has been shown that the injections of MPL eliminate the β-amyloid plaques in mice during a 12-month period. With the same method as ii, and after every prerequisite is fulfilled, it could move to clinical trials involving humans. The same experiment could be conducted but in this case, a group of people with Alzheimer's would have the concentration of beta-amyloids in their brain measured. Then, after the vaccine, the same measurement would take place. The two hypotheses for the ttest, comparison between the means of two dependent groups, follow:

H_0: There is no significant difference in the concentration of beta-amyloids in the brains of humans with Alzheimer's before and after MPL treatment.

H_1: There is significant difference in the concentration of beta-amyloids in the brains of humans with Alzheimer's before and after MPL treatment.

In case that H_0 is rejected in a significance level of 0.05 or 0.1, we can conclude that beta-amyloids are gradually destroyed and thus, the vaccine boosts tha patients' physical immunity.

Before these statistical analyses we should, of course, make sure that our data fulfill the assumptions in order for ttest to be used:

- Normality of the data should be assured.
- For independent samples, homogeneity of variances should be assessed.
- Independence between groups (in case of hypothesis i).
- For the paired samples t-test (like hypothesis ii,iii,iv), the dependent variable should be measured on an interval or ratio scale.