Biostatistics — 2nd Set of Exercises

Konstantinos Konstantinidis

February 27, 2025

Answers

1

First we load our data in, in a two dimensional array, with six columns corresponding to the six features measured, and 40 rows corresponding to the participants of the study, using the following code in MATLAB. The data is stored in the variable X.

```
x1 = [
    133; 140; 139; 133; 137; 99; 138; 92; 89; 133; 132; 141; 135; 140; 96; 83; 132;
    100; 101; 80; 83; 97; 135; 139; 91; 141; 85; 103; 77; 130; 133; 144; 03; 90; 83;
    133; 140; 88; 81; 89
];
x2 = [
    132; 150; 123; 129; 132; 90; 136; 90; 93; 114; 129; 150; 129; 120; 100; 71; 132;
    96; 112; 77; 83; 107; 129; 145; 86; 145; 90; 96; 83; 126; 126; 145; 96; 96; 90;
    129; 150; 86; 90; 91
];
x3 = [
    124; 124; 150; 128; 134; 110; 131; 98; 84; 147; 124; 128; 124; 147; 90; 96; 120;
    102; 84; 86; 86; 84; 134; 128; 102; 131; 84; 110; 72; 124; 132; 137; 110; 86; 81;
    128; 124; 94; 74; 89
];
x4 = [
    118; 127; 143; 172; 147; 146; 138; 175; 134; 172; 118; 151; 155; 155; 146; 135;
    127; 178; 136; 180; 175; 186; 122; 132; 114; 171; 140; 187; 106; 159; 127; 191;
    192; 181; 143; 153; 144; 139; 148; 179
];
x5 = [
    64.5; 72.5; 73.3; 68.8; 65.0; 69.0; 64.5; 66.0; 66.3; 68.8; 64.5; 70.0; 69.0;
    70.5; 66.0; 68.0; 68.5; 73.5; 66.3; 70.0; 75.1; 76.5; 62.0; 68.0; 63.0; 72.0;
    68.0; 77.0; 63.0; 66.5; 62.5; 67.0; 75.5; 69.0; 66.5; 66.5; 70.5; 64.5; 74.0;
    75.5
];
x6 = [
    816932; 1001121; 1038437; 965353; 951545; 928799; 991305; 854258; 904858; 955466;
    833868; 1079549; 924059; 856472; 878897; 865363; 852244; 945088; 808020; 889083;
    892420; 905940; 790619; 955003; 831772; 935494; 798612; 1062462; 793549; 866662;
    857782; 949589; 997925; 879987; 834344; 948066; 949395; 893983; 930016; 935863
];
X = [x1 x2 x3 x4 x5 x6];
```

Since we will be performing PCA on the data, we can standardize it first. We can also calculate the mean and standard deviation of each feature and inspect their correlation matrix, to see if there are any strong correlations. To that end we use the following code.

```
means = mean(X);
std_devs = std(X);
fprintf('The mean of each variable are:\n');
disp(means);
fprintf('The standard deviation of each variable are:\n');
disp(std_devs);

X = zscore(X);

corrcoefs = corrcoef(X);
[~, axes] = plotmatrix(X);
title(['Combination of All Initial Variables in pairs and their correlation ', ...
    'coefficients']);
for i = 1:6
    for j = 1:6
        if i == 6
            axes(i, j).XLabel.String = sprintf('x%d', j);
        end
        if j == 1
            axes(i, j).YLabel.String = sprintf('x%d', i);
        end
        if j ~= i
            ax = axes(i, j);
            xpos = min(ax.XLim) + 0.01 * range(ax.XLim);
            ypos = max(ax.YLim) * 0.9;
            text(ax, xpos, ypos, sprintf('%.2f', corrcoefs(i, j)), ...
            'HorizontalAlignment', 'left', ...
            'VerticalAlignment', 'top', ...
            'FontSize', 10, 'FontWeight', 'bold', 'Color', 'red');
        end
    end
end
```

The means and standard deviations of the features are as follows (before standardization):

```
The mean of each variable are:
    1.0e+05 *

    0.0011    0.0011    0.0011    0.0015    0.0007    9.0876

The standard deviation of each variable are:
    1.0e+04 *

    0.0030    0.0024    0.0022    0.0024    0.0004    7.2282
```

The correlation matrix of the features is shown in Figure 1. From the correlation matrix we can see that some features are correlated, such as X_1 and X_2 with a correlation coefficient of 0.83, X_1 and X_3 with a correlation coefficient of 0.76, and X_2 and X_3 with a correlation coefficient of 0.78. Also features X_4 and X_5 are slightly correlated with a coefficient of 0.67 and X_5 and X_6 with a coefficient of 0.57. Hence we expect that with only a few principal components the data set will be adequately described in the transformed feature space.

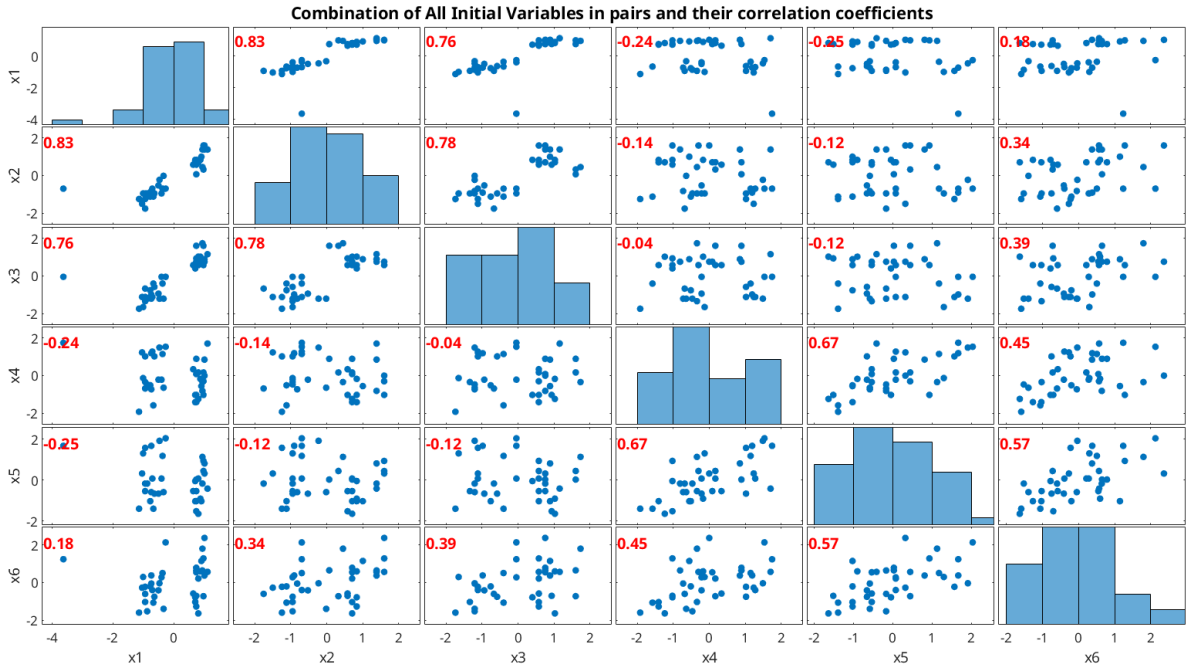

Figure 1: Matrix plot of the six features with their correlation coefficients in red.

1.A

Perform a Principal Component Analysis (PCA) to the above 6 variables in order to determine:

i. The smallest number of Principal Components that describe at least 80% 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 the question *i.*

It is given that the assumptions required for PCA application are valid.

The following code performs PCA on the data and answers the questions. The first part performs PCA on the data and calculates the cumulative explained variance of the principal components. We then find the smallest number of principal components that explain at least 80% of the variability. The second part calculates the coordinates of each principal component with respect to the initial variables and the third part calculates the coordinates of the data set in the coordinates system defined by the principal components. The analysis is shown in the code below.

```
[coeff, score, latent, ~, explained] = pca(X);

%%%%%%%%%%
% A. i. %
%%%%%%%%%%
cumulative_explained = cumsum(explained);
num_components = find(cumulative_explained >= 80, 1);
fprintf(['The smallest number of principal components that explain at least ', ...
    '80%% of the variability is %d.\n'], num_components);
figure;
plot(cumulative_explained, 'LineWidth', 2);
hold on;
scatter(1:6, cumulative_explained, 'filled');
xlabel('Number of Principal Components');
ylabel('Cumulative Explained Variance (%)');
title('Cumulative Explained Variance by Number of Principal Components');
grid on;
hold off;

%%%%%%%%%%
% A. ii. %
%%%%%%%%%%
fprintf(['The coordinates of each principal component with respect to the ', ...
    'initial variables are:\n']);
disp(coeff(:, 1:num_components));

%%%%%%%%%%%
% A. iii. %
%%%%%%%%%%%
fprintf(['The coordinates of the data set in the coordinates system defined ', ...
    'by the principal components are:\n']);
disp(score(:, 1:num_components));
```

The text output of the first part of the script is shown below. From that, we can see that the smallest number of principal components that explain at least 80% of the variability is 2. The plot of the cumulative explained variance over the number of principal components is shown in Figure 2.

```
The smallest number of principal components that explain at least 80% of the
variability is 2.
```

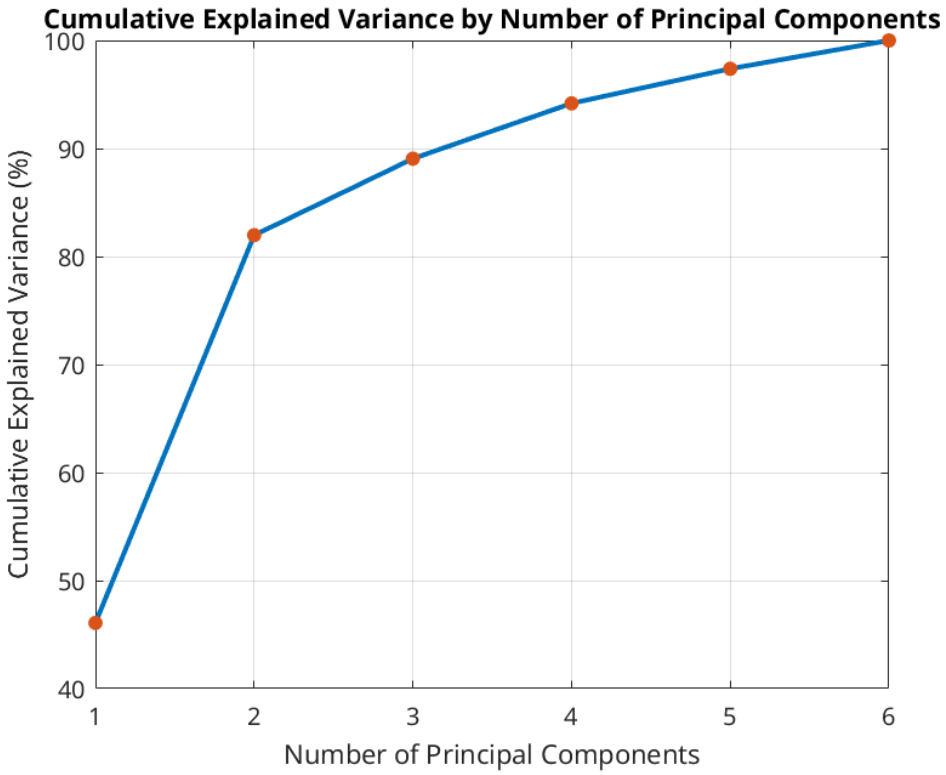

Figure 2: Plot showing the cummulative explained variance over the number of principal components.

The text output of the second part is provided below. The coordinates of each principal component with respect to the initial variables are shown. The first column corresponds to the first principal component and the second column to the second principal component.

```
The coordinates of each principal component with respect to the initial variables
are:
    0.5567   -0.0624
    0.5605    0.0491
    0.5447    0.0953
   -0.1341    0.5648
   -0.1380    0.5994
    0.2056    0.5535
```

The text output of the third part is shown below. The coordinates of the data set in the coordinates system defined by the principal components are shown. The first column corresponds to the first principal component and the second column to the second principal component. Each row corresponds to a participant of the study.

```
The coordinates of the data set in the coordinates system defined by the principal
components are:
    1.2630   -2.0636
    2.0231    0.7617
    1.9815    1.6057
    1.2573    1.0133
    1.7808   -0.2285
   -0.7038    0.0520
    2.0031   -0.2202
   -1.4017   -0.2991
   -1.3583   -0.8995
    1.3337    0.9869
    1.2213   -1.9380
    2.3095    1.5864
    1.1706    0.2968
    1.3651    0.0680
   -1.0478   -0.8288
   -1.8775   -0.9104
    1.0611   -1.0038
   -1.0251    1.5835
   -0.9695   -1.5787
   -2.2906    0.6229
   -2.2266    1.2841
   -1.5150    1.8698
    1.4586   -2.5045
    1.9752   -0.1241
   -1.0328   -2.3673
    1.6721    1.2601
   -1.8983   -1.3169
   -0.6110    3.2406
   -2.1564   -2.9562
    0.9043   -0.4098
    1.4470   -1.8070
    1.9689    1.1328
   -2.6450    2.8561
   -1.6502    0.4488
   -1.8733   -1.2002
    1.3944    0.0863
    1.8467    0.4801
   -1.2993   -1.0973
   -2.0906    0.7295
   -1.7645    1.7885
```

1.B

Produce the Y-index calculated using the relation:

$$Y = 3(X_1 + X_2 + X_3) - 4X_4 + 5X_5 + 3X_6$$

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.

The following code calculates the Y-index using the given relation and then finds the coefficients of the multiple linear regression model of the variable Y, taking as independent variables the principal components. The assumptions required for the application of the multiple linear regression model are valid. The code is shown below.

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

mdl = fitlm(score(:, 1:num_components), Y);

fprintf(['The coefficients of the multiple linear regression model of the ', ...
    'variable Y, taking as independent variables the principal components ', ...
    'are:\n']);
disp(mdl.Coefficients.Estimate);
```

The text output of the script is shown below. The coefficients of the multiple linear regression model of the variable Y, taking as independent variables the principal components, are shown. The first coefficient corresponds to the intercept and the next coefficients correspond to the coefficients of the two principal components.

```
The coefficients of the multiple linear regression model of the variable Y, taking
as independent variables the principal components are:
    0.0000
    5.4487
    2.6440
```

2

First we load the data in two arrays, one for each type of the three types of instruments, and one for each time of the day. The data is stored in the variables `types` and `times` respectively. We also store the names of the types and times in the variables `types_names` and `times_names`. The code is shown below.

```
type1 = [6.3; 7.1; 5.5; 5.9; 5.3; 6.6; 6.8; 7.2; 8.2; 9.1; 6.4; 7.5];
tyep2 = [6.1; 3.9; 4.3; 4.8; 5.3; 3.9; 4.2; 4.1; 4.3; 5.8; 4.1; 5.2];
type3 = [3.2; 4.2; 4.8; 5.3; 5.1; 3.7; 4.8; 4.7; 4.9; 5.5; 4.8; 5.7];
types = [type1, tyep2, type3];
types_names = ["type1", "type2", "type3"];

morning = [6.3; 7.1; 5.5; 5.9; 6.1; 3.9; 4.3; 4.8; 3.2; 4.2; 4.8; 5.3];
noon = [5.3; 6.6; 6.8; 7.2; 5.3; 3.9; 4.2; 4.1; 5.1; 3.7; 4.8; 4.7];
afternoon = [8.2; 9.1; 6.4; 7.5; 4.3; 5.8; 4.1; 5.2; 4.9; 5.5; 4.8; 5.7];
times = [morning, noon, afternoon];
times_names = ["morning", "noon", "afternoon"];
```

Test, at a significane level of 5%, the validity of the following hypotheses:

a. The type of analyzers does not affect sales.

b. The time of day does not affect sales.

c. There is no interaction between the type of analyzers and the time of day.

Since we will be performing ANOVA on our data, we should first check if the assumptions for ANOVA are valid. The first assumption states that the samples should be independent. We can claim that this assumption holds because of the nature of the measurements. The second and third assumptions state that the populations follow the normal distribution and the population variances are equal, respectively. To that end we utilize the Lilliefors test and a Q-Q plot for each of the measurements and the vartestn (both for the `types` and the `times` array).

The Lilliefors test check whether a population follows the normal distribution and the Q-Q plot visualizes that. MATLAB's vartest uses the Bartlett test with the null hypothesis that the columns of data vector x come from normal distributions with the same variance and returns a summary table and a boxplot.

The MATLAB code provided below achieves those tasks, first for the `types` and then for the `times` arrays.

```
disp("Lilliefors normality test");
for i = 1:3
    [h, p, ksstat, cv] = lillietest(types(:, i), 'alpha', 0.05);
    disp([types_names(i), "result", h, "p-value", p]);
end
for i = 1:3
    figure;
    hold on;
    normplot(types(:, i));
    title(sprintf("Normality plot for analyzer %s", types_names(i)));
    hold off;
end
disp("Vartest");
[p, stats] = vartestn(types);
p

disp("Lilliefors normality test");
for i = 1:3
    [h, p, ksstat, cv] = lillietest(types(:, i), 'alpha', 0.05);
    disp([times_names(i), "result", h, "p-value", p]);
end
for i = 1:3
    figure;
    hold on;
    normplot(types(:, i));
    title(sprintf("Normality plot for time of day %s", times_names(i)));
    hold off;
end
disp("Vartest");
[p, stats] = vartestn(times);
p
```

The text output of the first part concerning the `types` array is provided below. We can see that the measurements corresponding to the type 2 equipment **do not follow a normal distribution**, however the standard deviations of the populations appear equal.

```
Lilliefors normality test
Warning: P is greater than the largest tabulated value, returning 0.5.
> In lillietest (line 207)
    "type1"    "result"    "0"    "p-value"    "0.5"
    "type2"    "result"    "1"    "p-value"    "0.016809"
    "type3"    "result"    "0"    "p-value"    "0.061578"
Vartest
p =
    0.3132
```

This deviation from the normal distribution can be seen in the QQ plot of the second type of instruments (Figure 3).

Note: For the sake of continuing with the question, I decided to ignore this finding, but keep in mind that normally the analysis of this data set would end here.

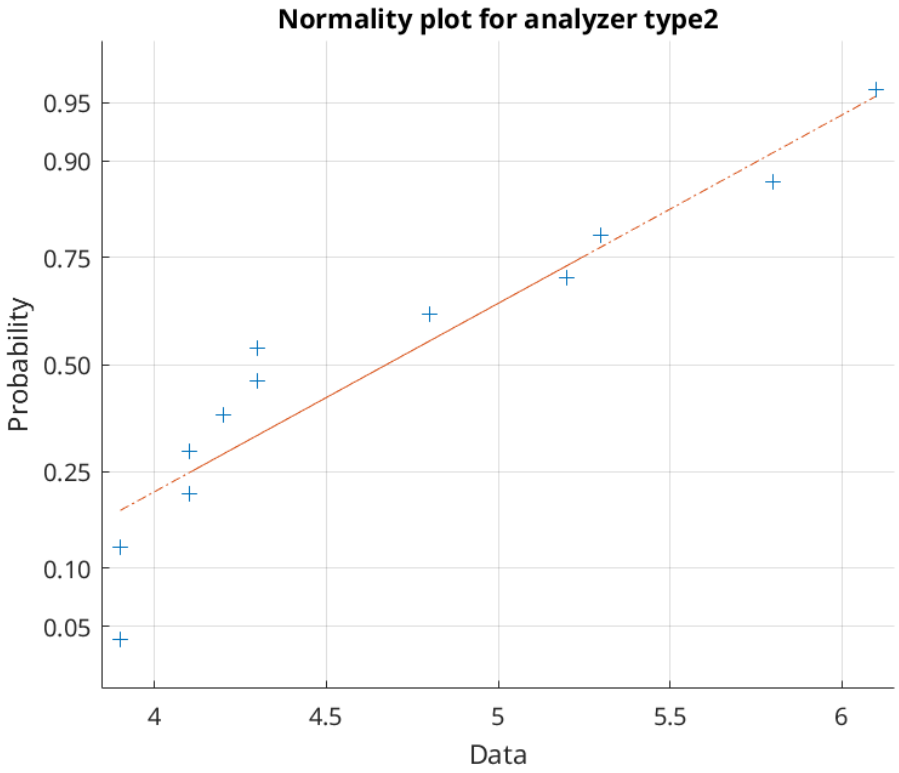

Figure 3: QQ plot of the measurements of the type 2 analyzer.

Next, we inspect the text output of the second part of the script, concerning the `times` array. Like before, we see that the measurements corresponding to noon **do not follow the normal distribution**, while the standard deviations appear equal.

```
Lilliefors normality test
Warning: P is greater than the largest tabulated value, returning 0.5.
> In lillietest (line 207)
    "morning"    "result"    "0"    "p-value"    "0.5"
    "noon"    "result"    "1"    "p-value"    "0.016809"
    "afternoon"    "result"    "0"    "p-value"    "0.061578"
Vartest
p =
    0.4839
```

This deviation from the normal distribution can be seen in the QQ plot of the noon measurements (Figure 3).

Note: Like before, for the sake of continuing with the question, I decided to ignore this finding, but keep in mind that normally the analysis of this data set would end here.

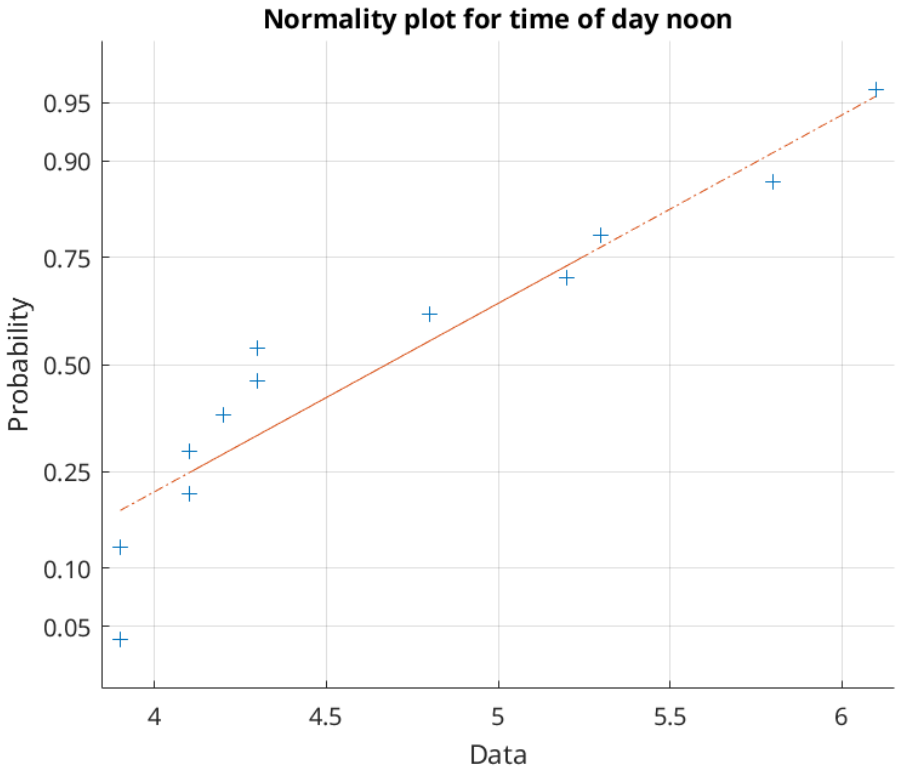

Figure 4: QQ plot of the measurements at noon.

2.a.

To check if the type of analyzer affects sales we can use the ANOVA under the null hypothesis that all population means are equal and an alternative hypothesis that at least one pair of means is not equal:

$$H_0 : \mu_{C_i} = \mu_{C_j} \ \forall i, j,$$
$$H_1 : \exists\ i, j \text{ with } i \neq j \text{ such that } \mu_{C_i} \neq \mu_{C_j}.$$

We can use the one-way ANOVA with the following MATLAB code:

```
[p, ~, stats] = anova1(types, types_names);
disp(["p-value", p]);
multcompare(stats);
```

The table output is provided below. The p-value is nearly zero, so at a 5% level of confidence, the null hypothesis is rejected for the alternative, meaning there is a statistically significant difference between the means of the analyzer types.

Source	SS	df	MS	F	Prob>F
Columns	36.2872	2	18.1436	23.71	4.13825e-07
Error	25.2517	33	0.7652		
Total	61.5389	35			

Using the multcompare function we can perform a multiple comparison test on the results. From the plot (Figure 5) we can see that the the type 1 analyzers have a greater mean than the pther 2, i.e. they sell more.

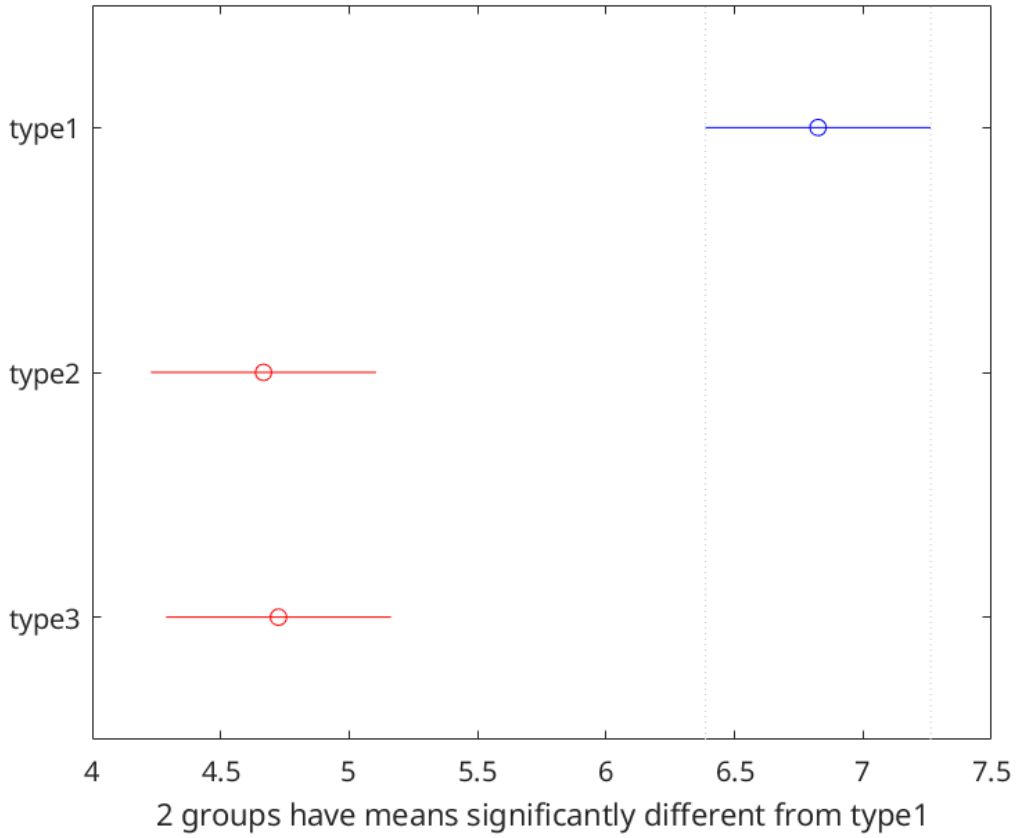

Figure 5: QQ plot of the measurements at noon.

2.b.

To check if the time of day affects sales we can use the ANOVA under the null hypothesis that all population means are equal and an alternative hypothesis that at least one pair of means is not equal

We can use the one-way ANOVA with the following MATLAB code:

```
[p, ~, stats] = anova1(times, times_names);
disp(["p-value", p]);
multcompare(stats);
```

The table output is provided below. We can see that the p-value is 0.213, so at a significance level of 5% the null hypothesis can not be rejected, meaning the time of day does not affect the number of sales of analyzers. There is no need to continue with the multcompare function.

```
Source      SS      df      MS       F     Prob>F
-----------------------------------------------
Columns    5.5039    2   2.75194   1.62   0.2131
Error     56.035    33   1.69803
Total     61.5389   35
```

2.c.

To check if there is an interraction between the type of analyzers and the time of day we can use the two-way ANOVA with the null hypothesis that there is no interaction between the two factors and an alternative hypothesis that there is an interaction between the two factors.

```
[p, ~, stats] = anova2(types, 4, "on");
disp(["p-value", p]);
multcompare(stats);
```

The table is provided below. We can see that for the interaction, the p-value is 0.4467, meaning the null hypothesis can not be rejected at a 5% level of significance, meaning there is no interaction between the type of analyzer and the time of day.

```
Source          SS      df      MS        F     Prob>F
-------------------------------------------------------
Columns       36.2872    2   18.1436   28.32   0
Rows           5.5039    2    2.7519    4.3    0.024
Interaction    2.4528    4    0.6132    0.96   0.4467
Error         17.295    27    0.6406
Total         61.5389   35
```

Note: we could have also used the two-way ANOVA to check for all questions (a, b, and c). The p-value (`PROB>F` columns) for the columns corresponds to the one-way ANOVA test on the analyzer types, the p-value for the rows corresponds to the one-way ANOVA test on the time of day.

3

We load the data in two column vectors, one for the `X` and one for the `Y` measurements, using the following code.

```
X = [4; 7; 1; 5; 8; 5; 2; 4; 3];
Y = [80; 92; 52; 76; 106; 100; 69; 71; 65];
```

Before proceeding with the questions, we must first check that the two sets of measurements follow the normal distribution. To that end, we can use the Lilliefors test and a QQ plot using the following code.

```
disp("Lillefors normality tests");

[h, p] = lillietest(X, 0.05);
disp("X");
disp(["result: ", h, "p-value", p]);

[h, p] = lillietest(Y, 0.05);
disp("Y");
disp(["result: ", h, "p-value", p]);

figure;
hold on;
normplot(X);
title("Q-Q plot for the X variable");
hold off;

figure;
hold on;
normplot(Y);
title("Q-Q plot for the Y variable");
hold off;
```

The text output is provided below. For both cases the null hypothesis is not rejected, indicating that both `X` and `Y` follow normal distributions.

```
Lillefors normality tests
X
    "result: "    "0"    "p-value"    "0.5"
Y
    "result: "    "0"    "p-value"    "0.5"
```

The QQ plots for X and Y are shown below (Figure 6 and Figure 7). The plots seem to aggree with the Lilliefors test.

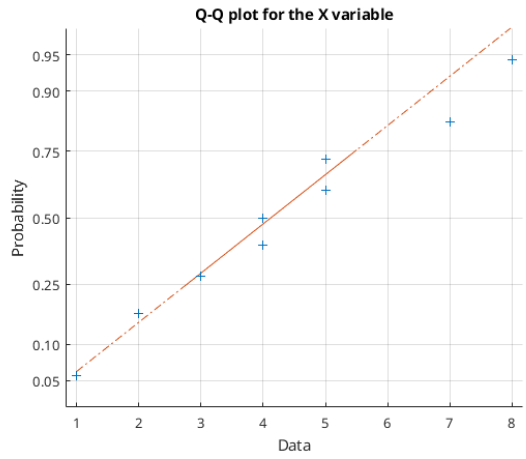

Figure 6: QQ plot of the X measurements.

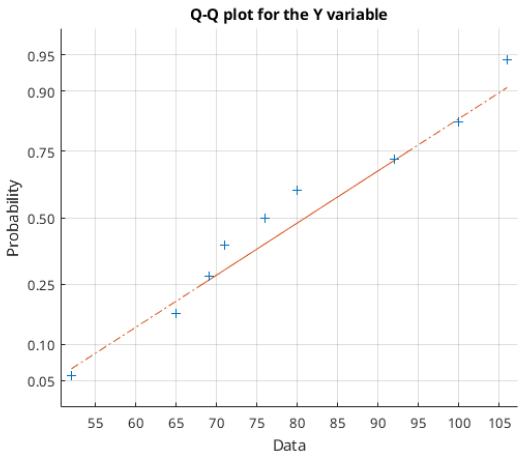

Figure 7: QQ plot of the Y measurements.

Next:

a. Plot the scatter diagram of the pairs of measurements.

b. Can you infer from the visual observation of the above diagram that x and y are correlated?

c. Calculate the appropriate correlation coefficient.

d. Can we reject the hypothesis that the correlation coefficient for the population is zero at the 5% significance level?

3.a.

To visualize the scatter plot of the `X` and `Y` measurement pairs we can use the `scatter` function in MATLAB using the following.

```
figure;
hold on;
scatter(X, Y, 75, "filled");
title("Scatter Plot of Measurements");
xlabel("X");
ylabel("Y");
xlim([min(X) - 0.05*range(X), max(X) + 0.05*range(X)]);
ylim([min(Y) - 0.05*range(Y), max(Y) + 0.05*range(Y)]);
hold off;
```

The generated scatter plot is shown in the Figure 8 below.

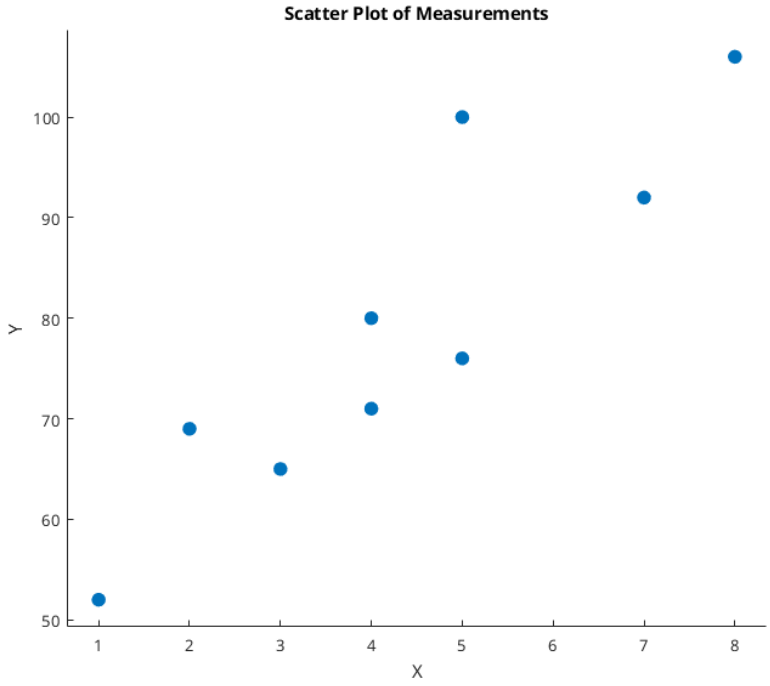

Figure 8: Scatter plot of the X and Y measurement pairs.

3.b.

The scatter plot (Figure 8) of the measurements shows that the `Y` values increase somewhat linearly as the `X` values increase, indicating we –at least visually– can claim that they have a positive correlation.

3.c.

The Pearson's correlation coefficient can be calculated using the following:

```
[rho, pval] = corr(X, Y);
disp(["Correlation coefficient: ", rho]);
```

The text output is provided below. As we can see, the `X` and `Y` measurements have a 0.89663 correlation coefficient.

```
"Correlation coefficient: "    "0.89663"
```

3.d.

To check whether the correlation coefficient is zero at 5% significance level we can utilize the p-value we got from the `corr` function (from above). The null hypothesis states that the two measurements are not correlated (i.e. their correlation coefficient is zero), while the alternative hypothesis states the opposite.

```
disp(["p-value: ", pval]);
alpha = 0.05;
if pval <= alpha
    disp("Null hypothesis rejected. X and Y are correlated");
else
    disp("Null hypothesis not rejected. X and Y are not correlated");
end
```

The text output of the MATLAB script is provided below. The p-value is calculated at 0.001, meaning that at a 5% significance level, the null hypothesis is rejected, the measurements are correlated.

```
"p-value: "    "0.0010556"
Null hypothesis rejected. X and Y are correlated
```

4

We load the data in two vectors, one named `years` and one named `salary`, corresponding to the experience and salary of the nurses respectively, using the following:

```
years = [18; 10; 4; 5; 6; 3; 16; 8; 14];
salary = [57; 50; 25; 28; 33; 19; 50; 45; 52];
```

a. Provide the scatter plot of the data using experience as an independent variable and monthly salary as a dependent variable.

b. On the above graph, plot the least squares line that fits the data points.

c. What percentage of the variability of the dependent variable is explained by the above line?

d. Provide the interpretation of the coefficients of the above line.

e. What is the forecasted monthly salary of a nurse when her experience is equal to:

   i. 9 years.

   ii. 15 years.

   iii. 21 years

f. Can we reject the hypothesis that there is no linear relationship between the monthly salary and the experience for nurses in country X, at a 5% significance level?

4.a.

The scatter plot of the data can be easily generated by using the following:

```
hold on;
scatter(years, salary, 75, "filled");
title("Experience vs Salary");
xlabel("Experience (years)");
ylabel("Montlhy salary (hundreds of Euros)");
xlim([min(years) - 0.05*range(years), max(years) + 0.05*range(years)]);
ylim([min(salary) - 0.05*range(salary), max(salary) + 0.05*range(salary)]);
hold off;
```

The scatter plot of salary over experience is provided below (Figure 9).

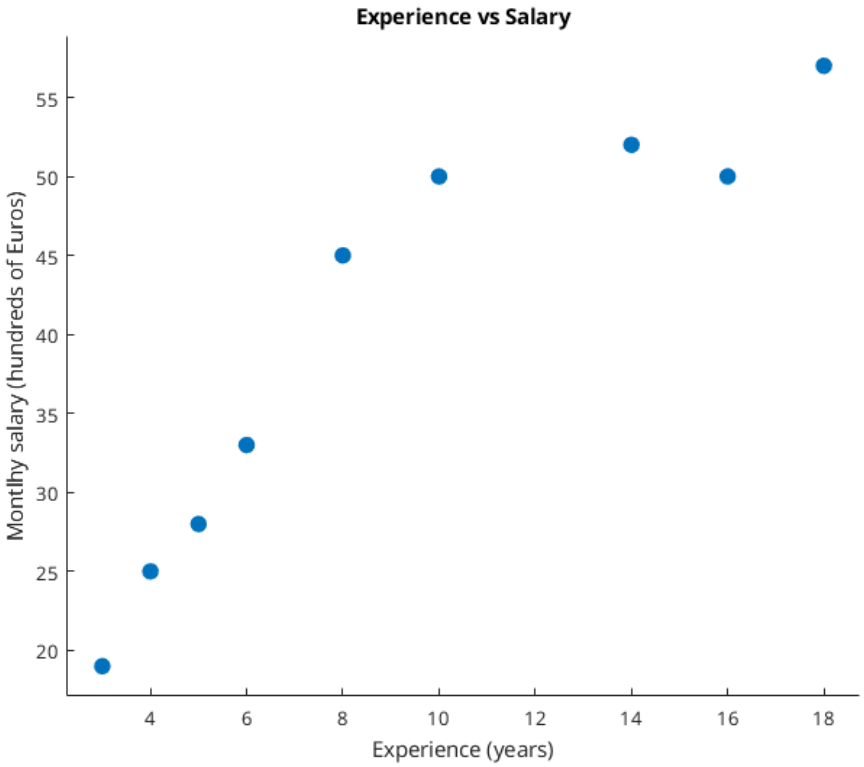

Figure 9: Scatter plot of the data set.

4.b.

The regression line can be displayed over the scatter plot with the following code:

```
x = [ones(size(years)) years];
C = regress(salary, x);
figure;
hold on;
scatter(years, salary, 75, "filled");
title("Experience vs Salary");
xlabel("Experience (years)");
ylabel("Montlhy salary (hundreds of Euros)");
xlim([min(years) - 0.05*range(years), max(years) + 0.05*range(years)]);
ylim([min(salary) - 0.05*range(salary), max(salary) + 0.05*range(salary)]);
x_fit = linspace(min(years), max(years), 100);
y_fit = C(1) + C(2) * x_fit;
plot(x_fit, y_fit, 'r-', 'LineWidth', 2);
legend('Data points', 'Least squares regression line');
hold off;
```

The plot is shown below (Figure 10).

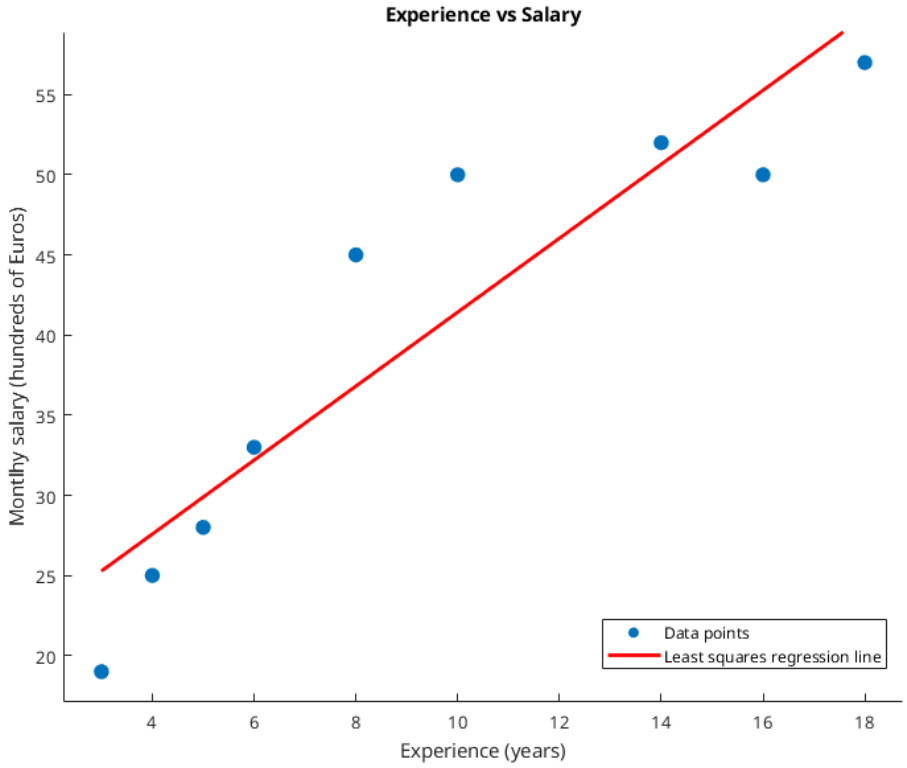

Figure 10: Scatter plot of the data set with a linear regression line in red.

In order to accept a least squares fit, we must first test that our assumptions are correct. The following assumptions must hold:

i. The dependent variable must follow the normal distribution.

ii. The variance of the dependent variable is the same for all values of the independent variable (constant variance).

iii. The relationship between the dependent and the independent variable is linear for the population.

iv. All observations are independent.

Using the following MATLAB code we can check whether these assumptions hold. The code below performs residual analysis on the linear regression model. The Lilliefors test is performed on the standardized residuals. Similarly, the QQ plot is used for the residuals. The studentized residuals are plotted against the estimates of the dependent variable to inspect for constant variance. For the linearity test there is nothing more needed, as we have shown the scatter plot above and we can see that they do seem linearly correlated. Finally, the observations are independent by the definition of the experiment.

```
% Assumptions %
% Residual analysis:
stats = regstats(salary, years, 'linear', {'standres','studres' });
stats.standres
stats.studres
% 1. Normality test
[h, p] = lillietest(stats.standres, 0.05);
disp("Standardized residuals:");
disp(["result: ", h, "p-value", p]);
figure;
hold on;
normplot(stats.standres);
title("Q-Q plot for the standardized residuals");
hold off;
% 2. Constant variance test
yest = x * C;
figure;
hold on;
title("Constant variance test");
xlabel("Y_{est}");
ylabel("Studentized residuals");
scatter(yest, stats.studres);
ylim([-10, 10]);
hold off;
```

The text output of the first part testing for the normality of the standardized residuals is provided below. As we can see, the null hypothesis is not rejected, meaning the values do follow a normal distribution.

```
Standardized residuals:
    "result: "    "0"    "p-value"    "0.5"
```

The QQ plot of the standardized residuals is shown below (Figure 11) and as we can see, it agrees with the conclusion made by the Lilliefors test.

The scatter plot to check for the constant variance is also shown below (Figure 12). Visually, the variance seems to be constant, so we can accept the least squares regression line.

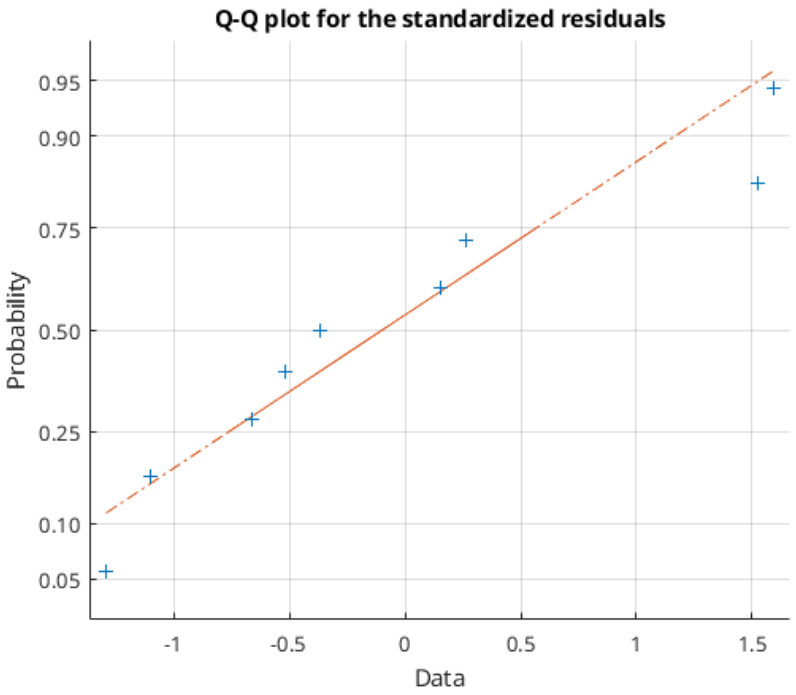

Figure 11: QQ plot of the standardized residuals.

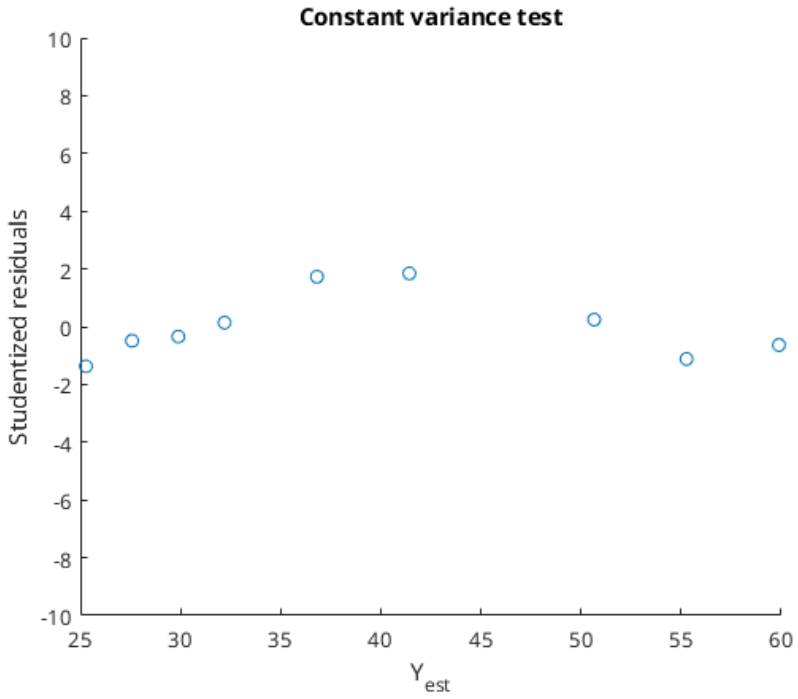

Figure 12: Constant variance test on the studentized residuals.

4.c.

The percentage of the variability explained by the regression line can be quantified with the square of the Pearson's correlation coefficient, as shown below.

```
R = corr(years, salary);
Rsq = R^2;
Rsq_perc = Rsq * 100;
fprintf('Percentage of variability explained: %.2f%%\n', Rsq_perc);
```

The text output is shown below, 85% of the variability of the dependent variable is explained by the regression line.

```
Percentage of variability explained: 85%
```

4.d.

The coefficients stored in the `C` variable are the intercept and the slope of the regression line. We can inspect them using the following:

```
disp(["C(1)", C(1)]);
disp(["C(2)", C(2)]);
```

The text output is shown below. The first line corresponds to the Y intercept, meaning the line crosses the Y axis at 18.3554. The second corresponds to the slope of the regression line, meaning the salary increases by 2307.2 Euros as experience increases by one year.

```
"C(1)"    "18.3554"
"C(2)"    "2.3072"
```

4.e.

The coefficients stored in the `C` variable can be used to form predictions on a nurses salary based on their experience. We can perform those predictions with the following code:

```
%   i. 9 years
forecast_9 = [1 9] * C;
disp("Expected salary in hundreds of Euros for a nurse with 9 years experience:")
disp(forecast_9);

%   ii. 15 years
forecast_15 = [1 15] * C;
disp("Expected salary in hundreds of Euros for a nurse with 15 years experience:")
disp(forecast_15);

%   iii. 21 years
forecast_21 = [1 21] * C;
disp("Expected salary in hundreds of Euros for a nurse with 21 years experience:")
disp(forecast_21);
```

The text output is shown below. Keep in mind that for more than 18 years of experience we can not form any predictions (since we have measurements for up to 18 years of experience), so the last one should be discarded.

```
Expected salary in hundreds of Euros for a nurse with 9 years experience:
   39.1198
Expected salary in hundreds of Euros for a nurse with 15 years experience:
   52.9628
Expected salary in hundreds of Euros for a nurse with 21 years experience:
   66.8058
```

4.f.

To check if there is no linear relationship between the monthly salary and the experience for nurses in country X, at a 5% significance level we can use the ANOVA with three equivalent null hypotheses:

1. There is no linear relationship in the population between the dependent and the independent variable.
2. The regression coefficient b in the population is equal to zero.
3. The R square for the population is equal to zero.

If the p-value is less than the significance level, the null hypothesis is rejected. The test is performed by using the following:

```
[p, ~, stats] = anova1(salary, years);
alpha = 0.05;
if p <= alpha
    fprintf("Null hypothesis rejected, there is linear relationship, (p-value =
    ↪  %d)", p);
else
    fprintf("Null hypothesis not rejected, there is no alinear relationship, (p-value
    ↪  = %d)", p);
end
```

The text output is shown below. We can see that the p-value is extremely small, hence there is a linear relationship between the values.

```
Null hypothesis rejected, there is linear relationship, (p-value = 0)
```

5

The researchers claim that a hypertension pill containing the drug propranolol could potentially reduce or completely erase bad memories related to fear. They decided to test their claims by first "teaching" volunteers to fear spiders by subjecting them to harmless but painful electrical shocks when looking at photos of spiders, while volunteers in the control group did not receive a shock. On the second day some volunteers were given propranolol, while others took a placebo. On the final day all volunteers looked at photos of spiders again, while a device measured their anxiety levels by looking at sublte contractions of the muscles around their eyes.

First, when reading about the structure of the study, we can see that the research team correctly split their volunteers in two groups, where one group was shocked to "teach" fear of spiders, while the second was not, and then some of them were given the drug while others were given a placebo, example given in Table 1. Also, if the grouping was made with a random method, then the groups are independent from each other, which is extremely important for the hypothesis tests that have been performed. We know that there were 60 volunteers in the study but since we do not have any information on how many volunteers were in each group we can assume that they were split somewhat evenly.

	Shocked	**Not shocked**
Propranolol	group 1	group 2
Placebo	group 3	group 4

Table 1: Example table of how the volunteers were grouped in the experiment.

One thing to mention here, is that the researchers measured how much the volunteers feared spiders, by inducing that fear themselves, however there is no mention of filtering out volunteers that already had a fear of spiders, which is an important distinction. This could confound the results, as their anxiety levels might not be solely due to the shocks.

The research team tested the hypothesis that propranolol reduces fear. In other words the mean of the measured stress levels of group 1 being less than that of group 3, while the means of group 2 and group 4 being equal (to rule out the possibility that propranolol reduces stress, regardless of "taught" fear). Also, by comparing the means of group 1 and group 4, the team can estimate whether the drug reduces stress to the levels of individuals who did not "learn" to fear spiders. More formally:

$$
\begin{aligned}
&H_1: \ \mu_1 < \mu_3 \\
&H_2: \ \mu_2 = \mu_4 \\
&H_3: \ \mu_1 = \mu_4 \\
&\text{where } \mu_i \text{ is the mean of the } i\text{'th group}
\end{aligned}
$$

All of the above can be carried out by perfofming a two-way ANOVA test. This test is appropriate because it allows for examining the main effects of the shock condition and the drug condition, as well as any interaction effects between the two. The report should include the significance level α or the p-value, but it does not, so we can not comment on that.

Before performing a two-way ANOVA test, several key assumptions must be met to ensure the validity of the results. First, the data should satisfy the assumption of independence, meaning that the observations in each group are independent of each other. This

is typically achieved through random assignment of participants to groups, as mentioned above. Second, the dependent variable (e.g., anxiety levels) should be normally distributed within each group, which can be checked using statistical tests like the Lilliefors test or visual methods like Q-Q plots. Third, the assumption of homogeneity of variances must hold, meaning that the variance of the dependent variable is approximately equal across all groups. This can be tested Bartlett's test. If these assumptions are violated, the results of the ANOVA may be unreliable, and alternative non-parametric tests (e.g., Kruskal-Wallis test) or transformations of the data may be necessary. Additionally, the dependent variable should be measured on a continuous scale (e.g., anxiety levels measured by muscle contractions), and the independent variables (e.g., shock condition and drug condition) should be categorical. Ensuring these assumptions are met is critical for the accuracy and interpretability of the ANOVA results.

The article concludes that propranolol fights the phobic reaction by weakening the painful memory, so their first and second hypotheses must have been accepted, but we do not have any information about the third one, however it should have been also accepted.

One thing to note is that the researches claim that the mechanism invlovled in the suppression of these fearful memories is a process called "reconsolidation". Reconsolidation is a process where memories are modified as they are recalled, so for propranolol to be responsible for the suppression of the memories related to fear of spiders, it should be in the volunteers system, in its active form, during the time when the memory is recalled.

The authors do not include any information on that, however one thing we can deduce is that propronalol must be helping in reconsolidating memories in very small concentrations. The volunteers took propronalol in the second day, but their memories of spiders were recalled the next day, on day three; and since propronalol's plasma half life is 3 to 6 hours[1], then by the time of memory recall, propronalol's concentration in the volunteers' plasma would be ranging from 0.3% to 6.3% of its initial concentration[2], which is an interesting finding!

[1]From the FDA's list of approved drugs:
`https://www.accessdata.fda.gov/drugsatfda_docs/label/2021/205410s006lbl.pdf`

[2]Equation for calculating the concentration of a drug based on its half-life: $C = \left(\frac{1}{2}\right)^{t/\lambda_{1/2}}$ where: t is the total time elapsed and $\lambda_{1/2}$ is the half-life of the drug.