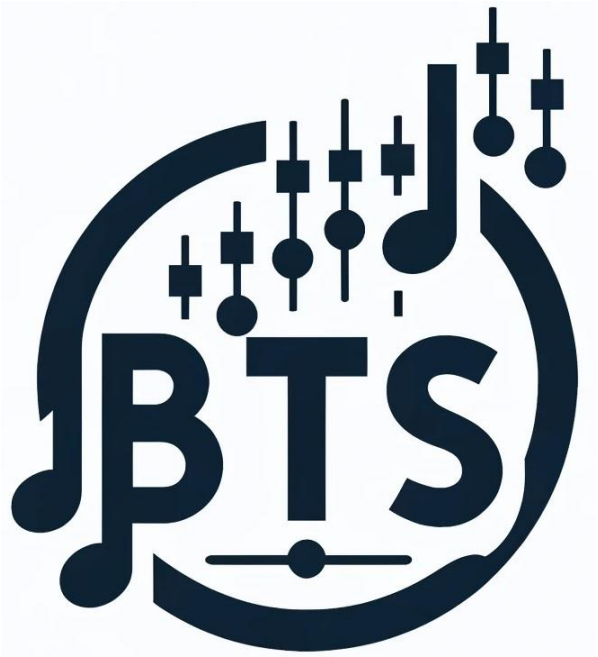
Basic Testing System (BTS) Version 1.0 User Tutorial



Hunter Dlugas, M.S.

Janaka Liyanage, Ph.D.

Seongho Kim, Ph.D.

Note

This tutorial describes how to use the statistical R/Shiny package **BTS** (Basic Testing System) **Version 1.0**, focusing mainly on technical guidelines. This manual assumes that users have completed at least a college-level statistics course. For those needing statistical knowledge, please refer to any introductory college-level statistics book.

Installation

The *R/Shiny* package **BTS** is designed to run in a web browser using the *Shiny* package (https://shiny.posit.co/) under *RStudio* (https://posit.co/download/rstudio-desktop/). **BTS** is freely available on *GitHub* (https://github.com/hdlugas/BTS).

To run **BTS**, first install *R* (version 4.4.0 or higher) and *RStudio* (version 2024.04.0 or higher). Then, download **BTS** from *GitHub* (https://github.com/hdlugas/BTS) and extract it in the current directory. Next, open the *BTSv1.R* file in RStudio. Finally, click the **'Run App'** button located in the top right corner of the *BTSv1.R* file.

Contents

About BTS

By default, the **About BTS** tab will be displayed with a brief introduction upon launching the BTS app (**Figure 1**).

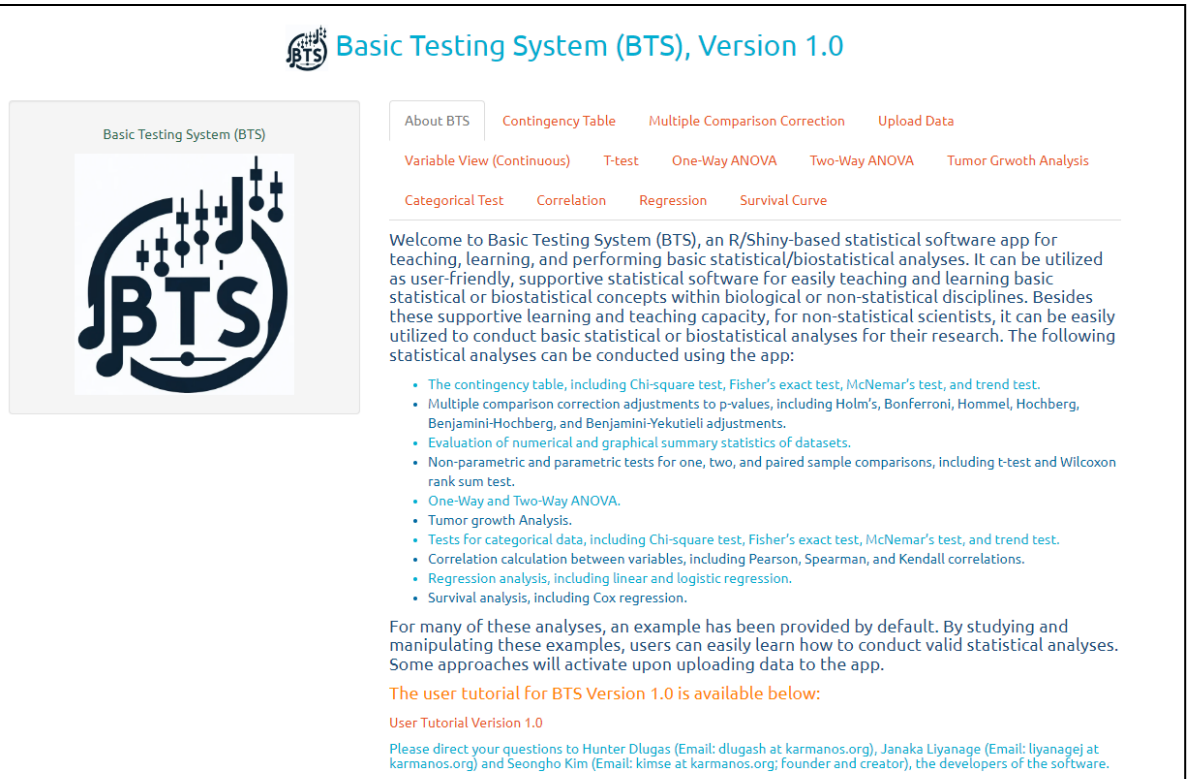
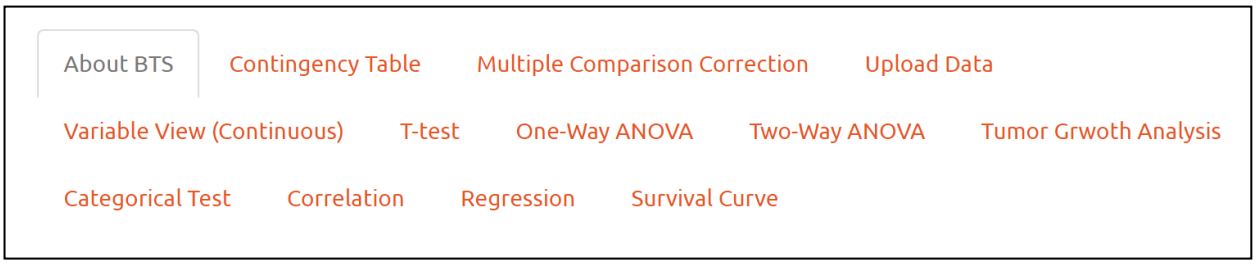


Figure 1. The **About BTS** tab.

The app has thirteen tabs on the top of the right panel, including the **About BTS** tab, aligned next to each other:



Statistical analyses can be conducted on these tabs by choosing the corresponding tab for each analysis. Each tab primarily consists of left and right panels. The left panel is the input panel that facilitates data upload and various parameter selections for performing each statistical procedure. The corresponding outputs are displayed on the right output panel.

Contingency Table

This section explains how to perform a contingency table analysis. To perform this analysis, a user must use the **Contingency Table** tab (**Figure 2**). The **Contingency Table** tab performs a similar function to the **Categorical Test** tab, but the main difference lies in the input data. While the **Contingency Table** tab analyzes data directly entered by the user into the subpanel, the **Categorical Test** tab analyzes data uploaded through the **Upload Data** tab. Please refer Categorical Test section for more detail.

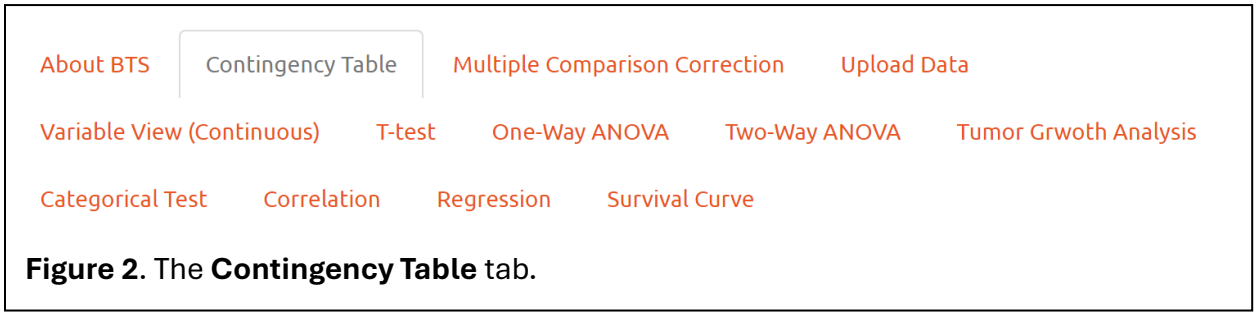


Figure 2. The **Contingency Table** tab.

Parameter Selection

A user needs to select a type of test and enter the data into the app as shown in **Figure 3**. The data must be entered in a contingency table format, arranged in rows and columns, through the **'Data'** multi-line text input. Each data point in each row must be separated by only one space, as the space functions as the data separator. If two spaces are entered consecutively, the app treats it as a missing data point separated by left and right spaces.

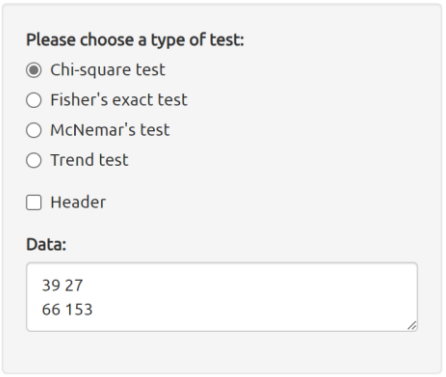


Figure 3. The parameter selection panel for the **Contingency Table** tab.

Note that the Chi-square test can also be performed as a goodness-of-fit test for one categorical variable via the **Contingency Table** tab. If a goodness-of-fit test for one categorical variable is performed with the Chi-square test, data should be entered as one row through the **'Data'** multi-line text input, separating each number by a space.

There are four options to perform a categorical test: Chi-square test (Chi-square), Fisher's exact test (Fisher's exact), McNemar's test (McNemar's), and Trend test (Trend). The Chi-square test and Fisher's exact test are used to investigate the association between two categorical variables. When the sample size is small, especially if the expected frequencies in any of the cells of a contingency table are less than 5, the Chi-square test's approximation can be inaccurate. In such cases, Fisher's exact test provides an exact p-value. Generally, Fisher's exact test is preferred for small samples, while the Chi-square test is more efficient for larger samples and larger tables. The McNemar's test is used to assess paired data. The trend test is used to evaluate the trend between a categorical variable and an ordinal variable.

Output

The output panel (right panel) is composed of three subpanels:

- Data
- Expected value (for a Chi-squared test only)
- Hypothesis test

The **Data** subpanel shows the input data in the format of a contingency table, which must be used to check if the data in the contingency table are entered correctly into the app. The **Expected value** subpanel provides the expected cell counts, assisting in the decision-making process to switch to a Fisher's exact test if needed. This subpanel is available only for a Chi-square test. The **Hypothesis test** subpanel reports the results of the hypothesis testing.

Example

The example data is a toy example and a hypothetical data provided by the BTS. It contains the treatment response pre- and post-treatment for 250 patients by drugs (drug_A and drug_B) and is composed of the following four variables:

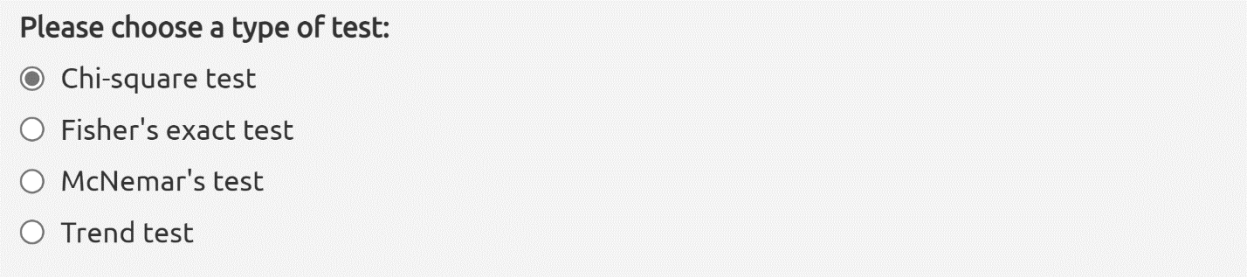
- **Treatment**: a factor with groups: drug_A and drug_B.
- **Drug_amount**: an ordinal factor with the drug level: Level_1 to Level_5.
- **Pre_Response**: a factor with the treatment response from a standard of care: Response and No_response.
- **Post_Response**: a factor with the treatment response from new treatments: Response and No_response.

The hypothesis to investigate is whether a tumor growth is associated with treatment.

Step by Step

A. Chi-square test

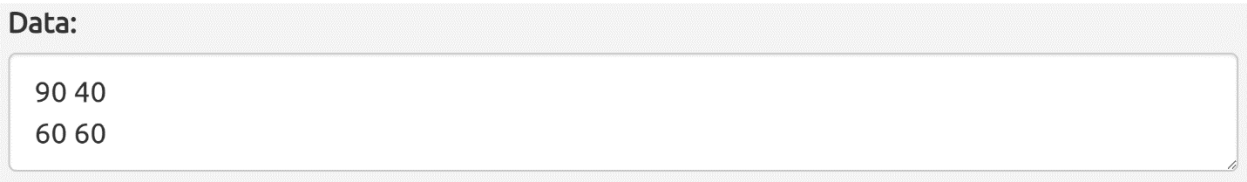
Choose the '**Chi-square test**':



The contingency table contains data for '**Pre_Response**' with two levels ('**No_response**' and '**Response**') and '**Treatment**' with two levels ('**drug_A**' and '**drug_B**'):

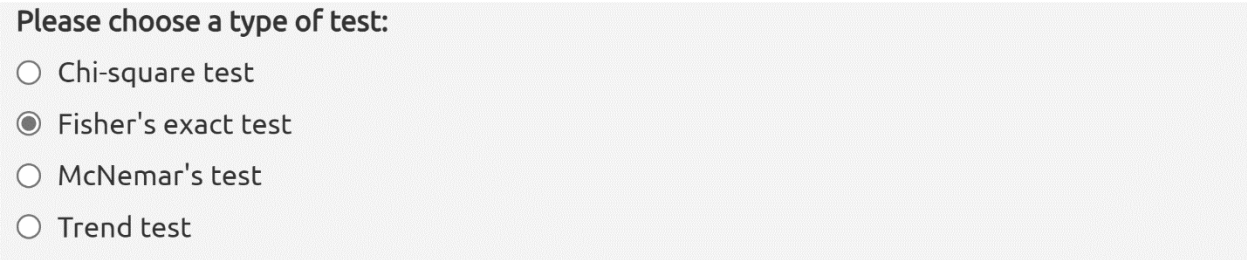
	drug_A	drug_B
No_response	90	40
Response	60	60

Enter the contingency table into the app via '**Data**' multi-line text input. Each data point in each row must be separated by a space, and a space must not be entered after the last data point in each row:

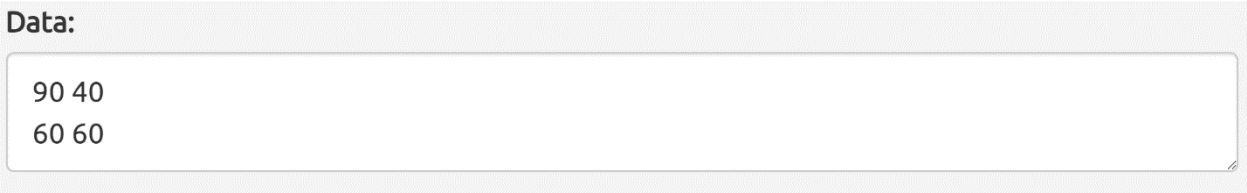


B. Fisher's exact test

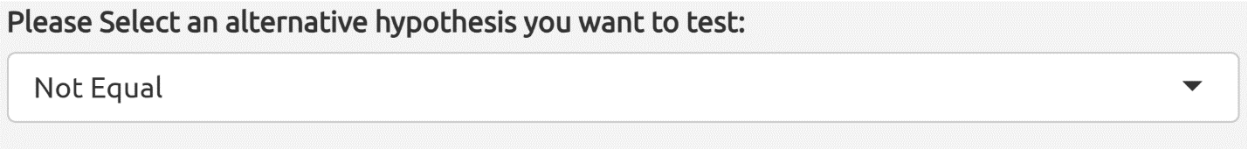
Choose the '**Fisher's exact test**':



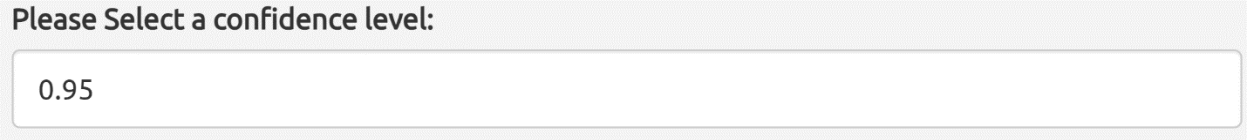
Enter the contingency table to the app through '**Data**' multi-line text input. Each data point in each row must be separated by a space, and a space must not be entered after the last data point in each row:



Choose the alternative hypothesis. Depending on the alternative hypothesis of interest, a two-tailed, left-tailed, or right-tailed test can be selected by choosing the options of '**Not Equal**', '**Less'**, or '**Greater'**. In this example, the alternative hypothesis is that tumor growth is not associated with treatment. Accordingly, choose '**Not Equal'**:



Select 95% confidence level:

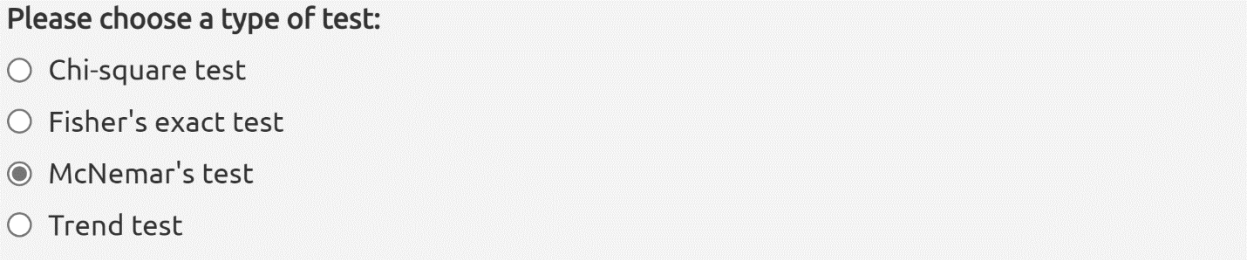


C. McNemar's test

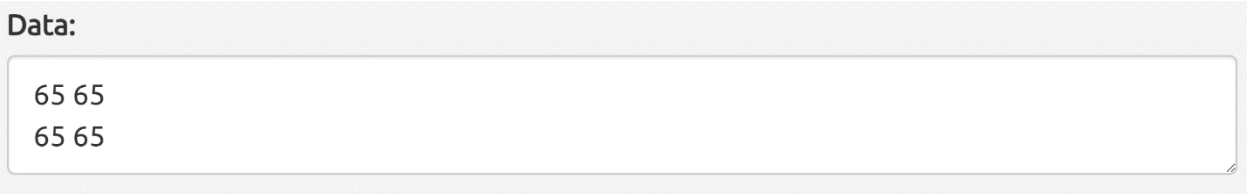
The contingency table contains the paired data of '**Pre_response**' and '**Post_response**' with two levels ('**No_response**' and '**Response**'):

	No_response	Response
No_response	65	65
Response	65	65

Choose the '**McNemar's test**':



Enter the table into the app via '**Data'** multi-line text input. Each data point in each row must be separated by a space, and a space must not be entered after the last data point in each row:

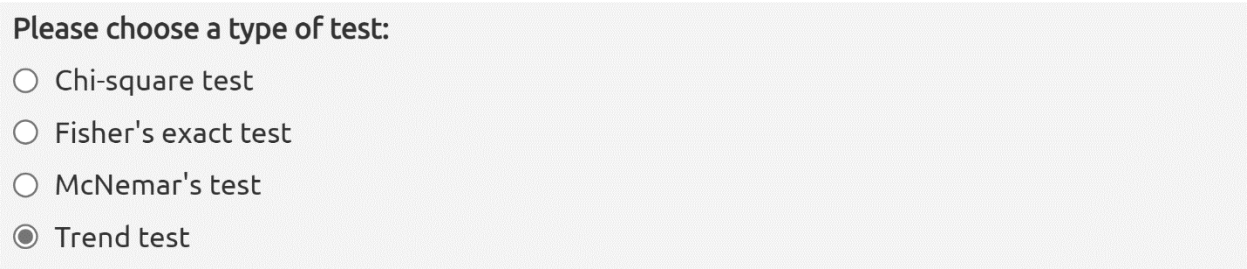


D. Trend test

The contingency table contains the trend data of '**Post_response**' with two levels ('**No_response**' and '**Response**') and '**Drug_amount** with five levels ('**Level_1**',..., '**Level_5**'):

	Level_1	Level_2	Level_3	Level_4	Level_5
No_response	26	26	26	26	26
Response	24	24	24	24	24

Choose the '**Trend test**':



Enter the contingency table into the app via the 'Data' multi-line text input. Each data point in each row must be separated by a space, and a space must not be entered after the last data point in each row:

Data:

```
26 26 26 26 26
24 24 24 24 24
```

Results and Interpretation

A. Chi-square test

Data

This subpanel provides the data read by the app in the form of a contingency table. Since each variable has two levels, there is a 2 × 2 contingency table. The data has been entered correctly into the app:

Data

```
     V1 V2
[1,] 90 40
[2,] 60 60
```

Expected values

Expected values

```
     V1 V2
[1,] 78 52
[2,] 72 48
```

This subpanel shows the expected counts for each cell, which will be used to decide whether a Fisher's exact test should be used instead. In this example, all expected counts are greater than 5, so a Chi-square test is still valid.

Hypothesis test

This subpanel prints out the results from a hypothesis test, which is a Chi-square test. The result shows that there is a significant association between '**Pre_Response**' and '**Treatment'** at the 5% level:

Hypothesis test

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  ttab2
X-squared = 8.8308, df = 1, p-value = 0.002962
```

B. Fisher's exact test

Data

This subpanel provides the data read by the app in the form of a contingency table. Since each variable has two levels, there is a 2 × 2 contingency table. The data has been entered correctly into the app:

Data

```
     V1 V2
[1,] 90 40
[2,] 60 60
```

Hypothesis test

This subpanel prints out the results from a hypothesis test, which is a Fisher's exact test. The result shows that there is a significant association between '**Pre_Response**' and '**Treatment**' at the 5% level:

Hypothesis test

```
        Fisher's Exact Test for Count Data

data:  ttab2
p-value = 0.002874
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.300417 3.900725
sample estimates:
odds ratio
  2.242518
```

C. McNemar's test

Data

This subpanel provides the data read by the app in the form of a contingency table. Since each variable has two levels, there is a 2 × 2 contingency table. The data has been entered correctly into the app:

Data

```
     V1 V2
[1,] 65 65
[2,] 65 55
```

Hypothesis test

This subpanel prints out the results from a hypothesis testing, which is a McNemar's test. The result shows that there is no association between '**Pre_Response**' and '**Post_response**' at the 5% level:

Hypothesis test

```
        McNemar's Chi-squared test

data:  ttab2
McNemar's chi-squared = 0, df = 1, p-value = 1
```

D. Trend test

Contingency table

This subpanel provides the data read by the app in the form of a contingency table. Since each variable has two levels, there is a 2 × 5 contingency table. The data has been entered correctly into the app:

Data

```
     V1 V2 V3 V4 V5
[1,] 26 26 26 26 26
[2,] 24 24 24 24 24
```

Hypothesis test

This subpanel prints out the results from a hypothesis testing, which is a Trend test. The result shows that there is no linear trend between '**Post_Response**' and '**Drug_amount**' at the 5% level:

Hypothesis test

```
        Chi-squared Test for Trend in Proportions

data:  ttab2[1, ] out of apply(ttab2, 2, sum) ,
 using scores: 1 2 3 4 5
X-squared = 6.3109e-30, df = 1, p-value = 1
```

Multiple Comparison Correction

This section explains adjusting p-values for controlling the type I error rates for multiple hypothesis testing. The Holm, Bonferroni, Hommel, Hochberg, Benjamini-Hochberg, and Benjamini-Yekutieli adjustments are commonly used to adjust p-values. These adjusted p-values can be obtained using the Multiple Comparison Correction tab on the BTS app. To perform this analysis, a user must use the Multiple **Comparison Correction** tab (**Figure 4**).

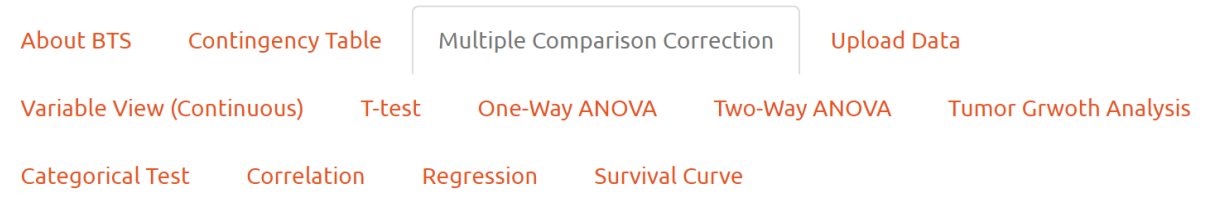


Figure 4. The **Multiple Comparison Correction** tab.

Parameter Selection

A user needs to select one or several preferred adjustment m ethods from the Holm, Bonferroni, Hommel, Hochberg, Benjamini-Hochberg, and Benjamini-Yekutieli adjustments, as shown in **Figure 5**. Multiple p-values that need to be adjusted must be entered into the app through the text input located at the bottom of the panel.

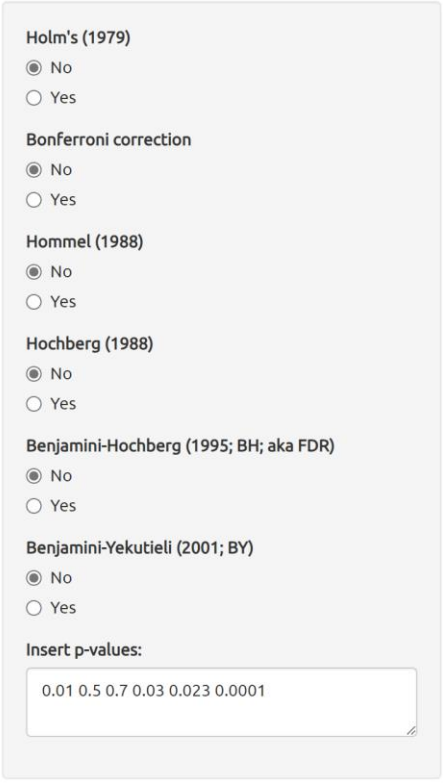


Figure 5. The parameter selection panel for the **Multiple Comparison Correction** tab.

Output

The output panel (right panel) is composed of three subpanels:

- p-values
- Corrected p-values
- Details from R help

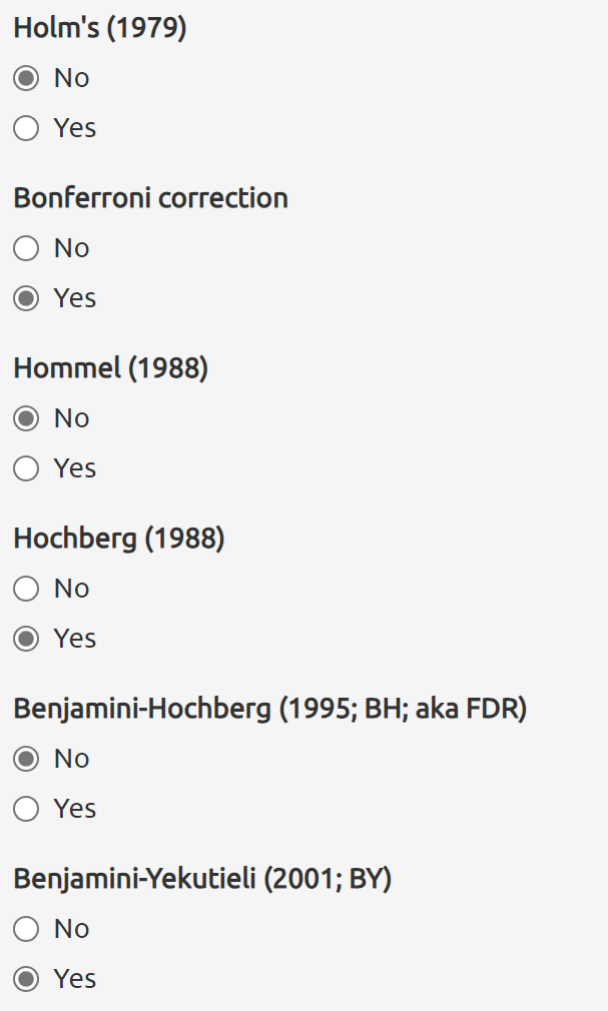
The first subpanel displays the p-values that are read by the app. The unadjusted p-values that need to be corrected must be entered through the text input located at the bottom of the panel.

Example

Consider the unadjusted p-values of 0.01, 0.06, 0.009, 0.081, 0.003, and 0.98 obtained through six multiple comparison tests. Suppose we are interested in applying Bonferroni, Hochberg, and Benjamini-Yekutieli adjustments to these p-values for controlling the type I error of performing six multiple tests.

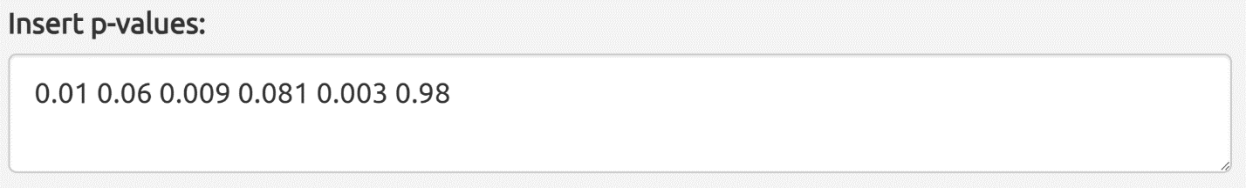
Step by Step

Choose '**Yes'** for Bonferroni correction, Hochberg (1988), and Benjamini-Yekutieli (2001;BY), while keeping '**No**' for the others.



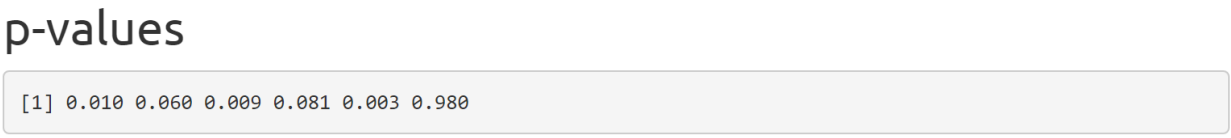
Enter the unadjusted p-values into the app through '**Insert p-values**' text input. Each p-value must be separated by only one space. An space must not be entered at the end of the

last p-vale. Entering more than one space between two numbers result in missing values during data reading:



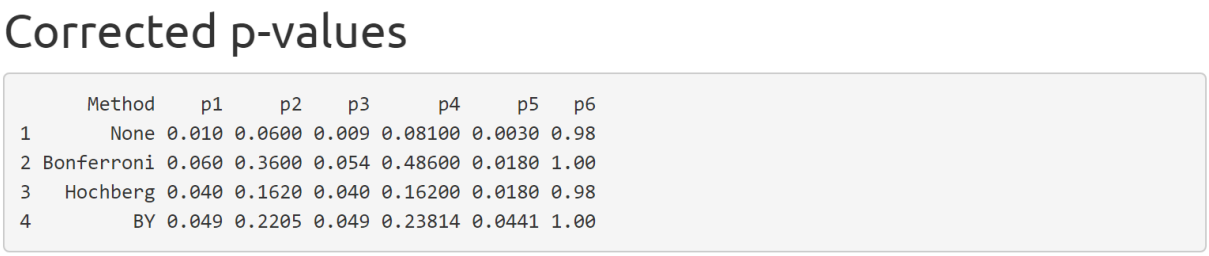
Results and Interpretation

p-values read by the app



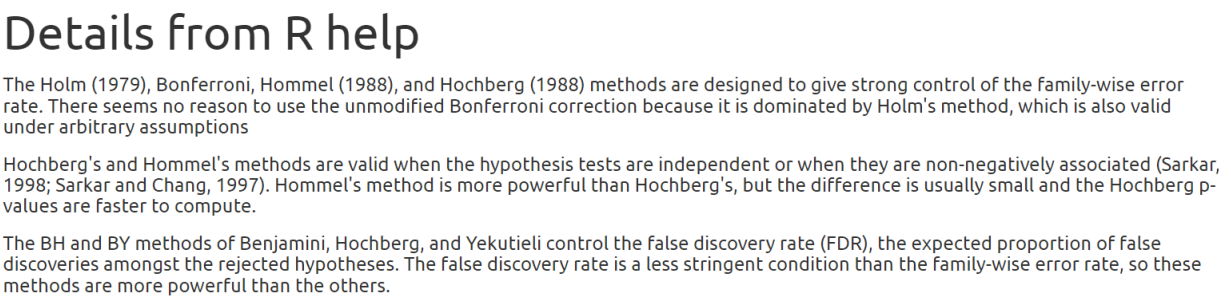
The app has read the p-values correctly as the entered and read values are consistent.

Corrected p-values.



The adjusted p-values under each selected method together with the unadjusted p-values are displayed. Among the three selected methods, Bonferroni is the most conservative procedure, while Hochberg is the least conservative adjustment.

Details extracted from R help



The app extracts and displays the R help information of the selected methods.

Upload Data

This section explains how to upload data to the app for executing statistical analyses. The external data file must be a CSV or text file. To upload data, a user must use the **Upload Data** tab (**Figure 6**).

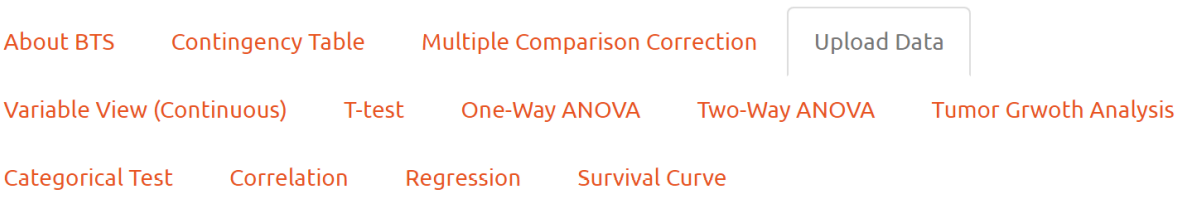


Figure 6. The **Upload Data** tab.

Parameter Selection

A user needs to select an external CSV or text data file for uploading. An option for exporting data from R packages is also available. These data upload options are shown in **Figure 7**. If an external data file is a text file, then the user needs to:

1. Select how the data points are separated from one another in the file.
2. Select how the character variables are indicated in the file.

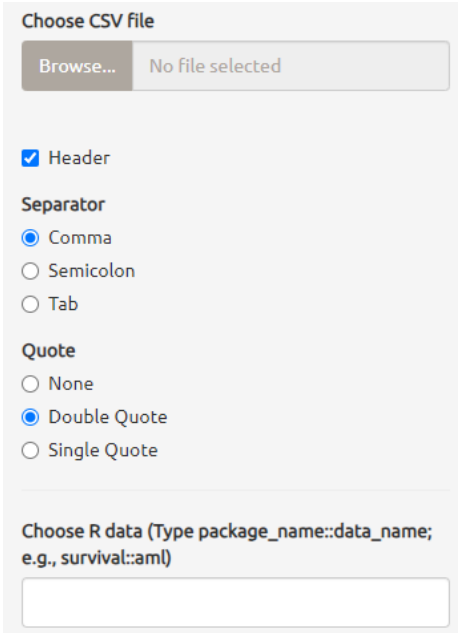


Figure 7. The parameter selection panel for the **Upload Data** tab.

By default, the app recognizes the data separation and how the categorical variable stored in a CSV data file. As a result, the choice of the separator and quote is not needed when uploading external CSV data file.

If the user wants to access a data set stored in an R package, enter the names of the data set together with the package to the bottom text field. All above mentioned steps of external data upload are not needed for uploading R data. Note that the R package needs to be installed before exporting data from the package.

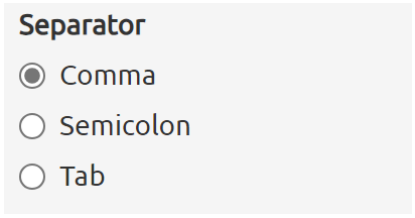


Figure 8. The option for a data separator for the **Upload Data** tab.

There are three options for the type of data separator in an external text data file: Comma, Semicolon, and Tab (**Figure 8**). Accordingly, it is essential to specify which data separation method is used in the external text data file when reading it into R. This can be achieved by choosing the correct separation among '**Comma'**, '**Semicolon'**, or '**Tab'** options. Based on the selection, the app correctly separates the data points or data columns when reading the data. If the external data file is a CSV file, or if a user wants to upload a data set from an R package, then these steps are not necessary.

In **Figure 9**, another parameter allows users to specify how the character variables are stored in a data file. These levels, along with the variable names, can be enclosed within single quotes, double quotes, or without any quotes. Users must select the corresponding option from the available choices: '**None'**, '**Double Quote'**, and '**Single Quote**'. If the external data file is a CSV file, or if a user wants to upload a data set from an R package, then this step can be ignored.

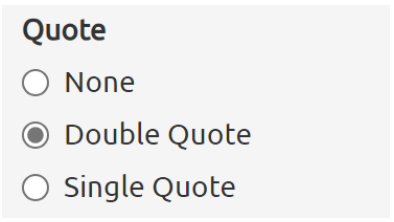


Figure 9. The option for a data separator for the **Upload Data** tab.

Output

The output panel (right panel) is composed of four subtabs (**Figure 10**):

- Data
- Summary
- Description
- R Data list

The first subtab displays the uploaded data, the second subtab displays the summary statistics of the uploaded data, the third subtab displays the path to the temporary folder

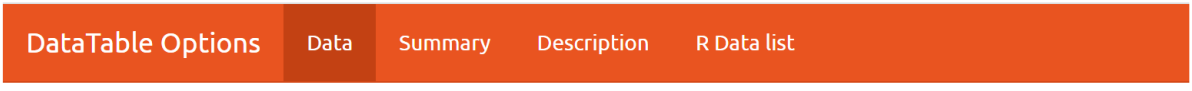


Figure 10. The subtabs of the output panel of the **Upload Data** tab.

that stored the uploaded data, or the data description available in the R package if the data are uploaded from an R package. The final subtab lists the datasets available in installed R libraries.

Example

The example text data file, named **carData_SLID.txt**, contains data from the 1994 wages of the Canadian Survey of Labor and Income Dynamics for the province of Ontario. This dataset is also available under the name "SLID" in the **carData** R package. The variables are as follows:

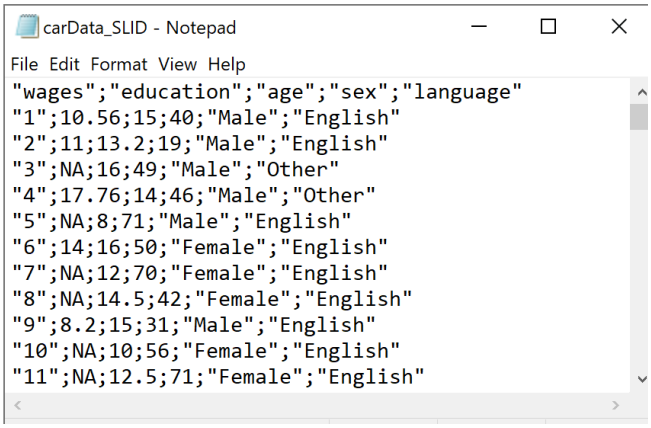


Figure 11. The first eleven rows of the **carData_SLID.txt**.

- **wages:** Composite hourly wage rate from all jobs.
- **education:** Number of years of schooling.
- **age:** in years.
- **sex:** a factor with levels: Female, Male.
- **language: a** factor with levels: English, French, Other

Figure 11 displays the data for the first eleven individuals. Note that the variable names, row numbers, and levels of the factors "**sex**" and "**language**" are stored within double quotes. The data in each line is separated by semicolons. This dataset should be properly uploaded into the app before performing any statistical analysis by correctly specifying this data separator and double quotes.

Step by Step

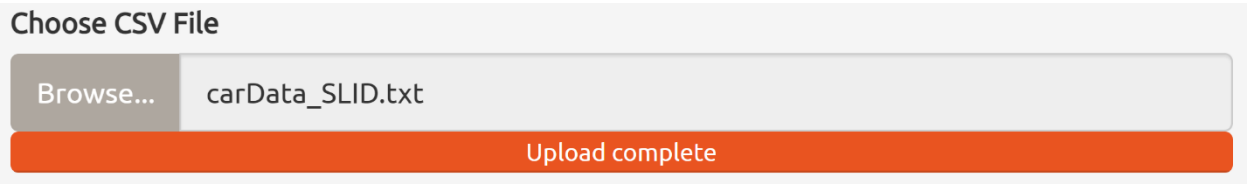
To upload data, go to the **Upload Data** tab.

If a user wants to upload SLID data from the **carData** R package, they should use the '**Choose R data (Type package_name::data_name; e.g., survival::aml)**' text input at the bottom of the right panel and enter the names of the data and package. Ignore all other

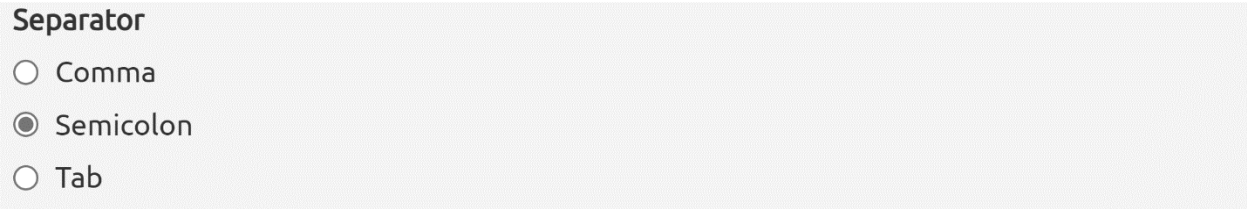
options available on the input panel, as they are required only for external data upload. Make sure that the **carData** package is installed in R before uploading data.

To upload an external CSV or text data file, click the "**Browse...**" button under the "**Choose CSV File**" section. A window will pop up, allowing you to select the data file from your computer. Locate and choose the desired data file through the popup window. For example, in this case, the **carData_SLID.txt** file is selected and uploaded.

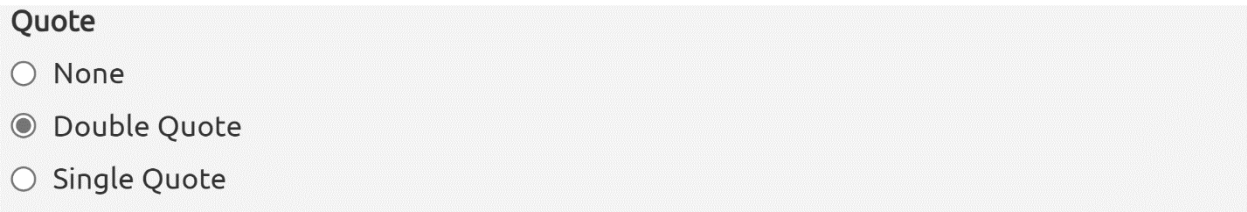
After a successful data upload, the name of the data file will be displayed in the text box next to the browse button, and the message "**Upload complete**" will be displayed underneath the browse button:



Choose "**Semicolon"** as the data separator. Ignore this step If the external data file is a CSV file:



Choose "**Double Quote"** as the quote. Ignore this step if the external data file is a CSV file:



Results and Interpretation

Data subtab

DataTable Options | Data | Summary | Description | R Data list

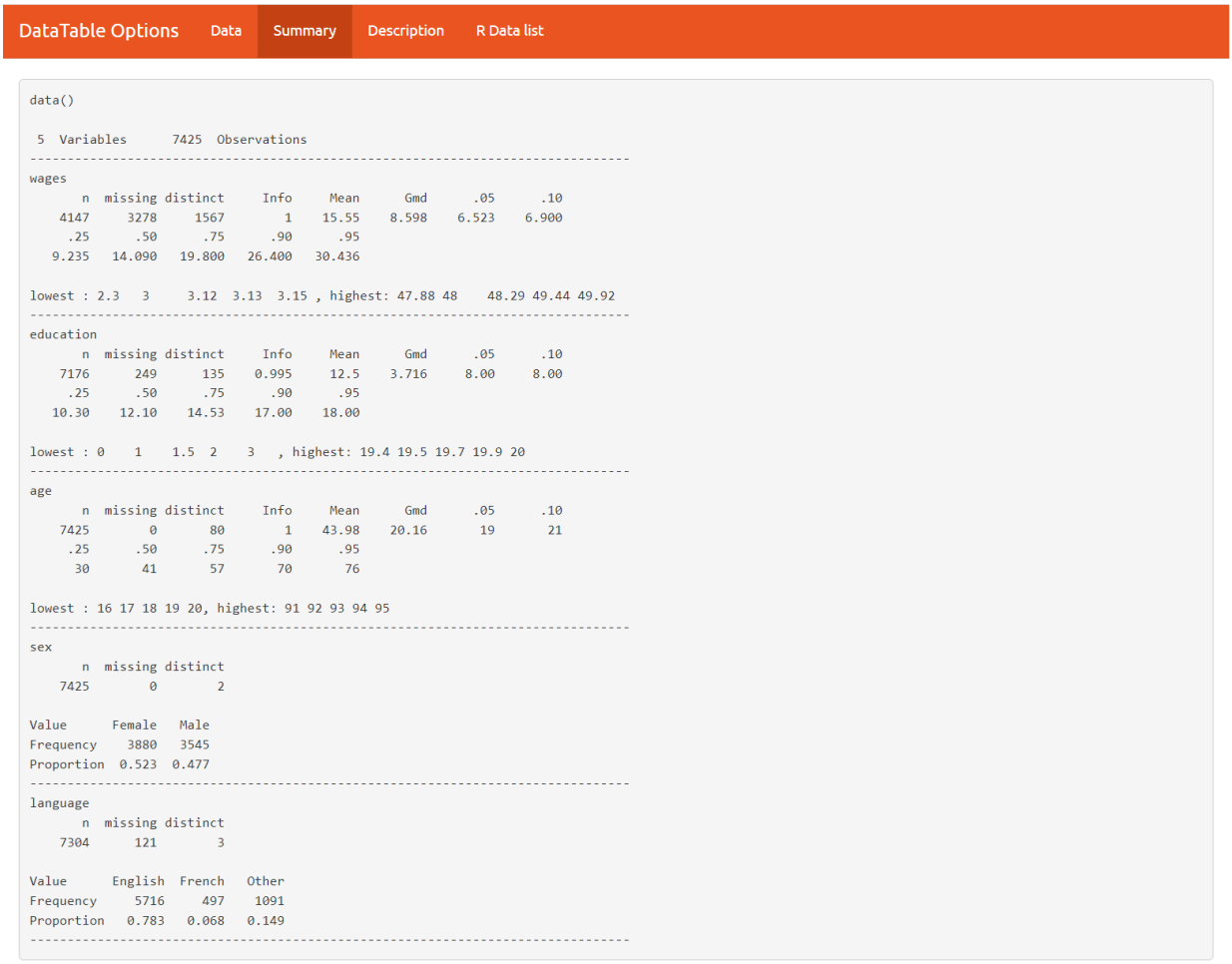
Show 10 entries Search:

	wages	education	age	sex	language
1	10.56	15	40	Male	English
2	11	13.2	19	Male	English
3		16	49	Male	Other
4	17.76	14	46	Male	Other
5		8	71	Male	English
6	14	16	50	Female	English
7		12	70	Female	English
8		14.5	42	Female	English
9	8.2	15	31	Male	English
10		10	56	Female	English

Showing 1 to 10 of 7,425 entries Previous 1 2 3 4 5 … 743 Next

The data columns are separated correctly, and the data is correctly arranged under each variable. By default, data for ten individuals appears per page, which can be changed to 25, 50, and 100 through the Show entries dropdown menu on the top left. The rest of the data can be viewed by navigating through each page, clicking on the page number at the bottom right. If a user wants to view data on a particular row of the data set, enter the row required data row number to '**Search**' text input in the top right. All the data staring from the entered row will be displayed on this panel.

Summary subtab

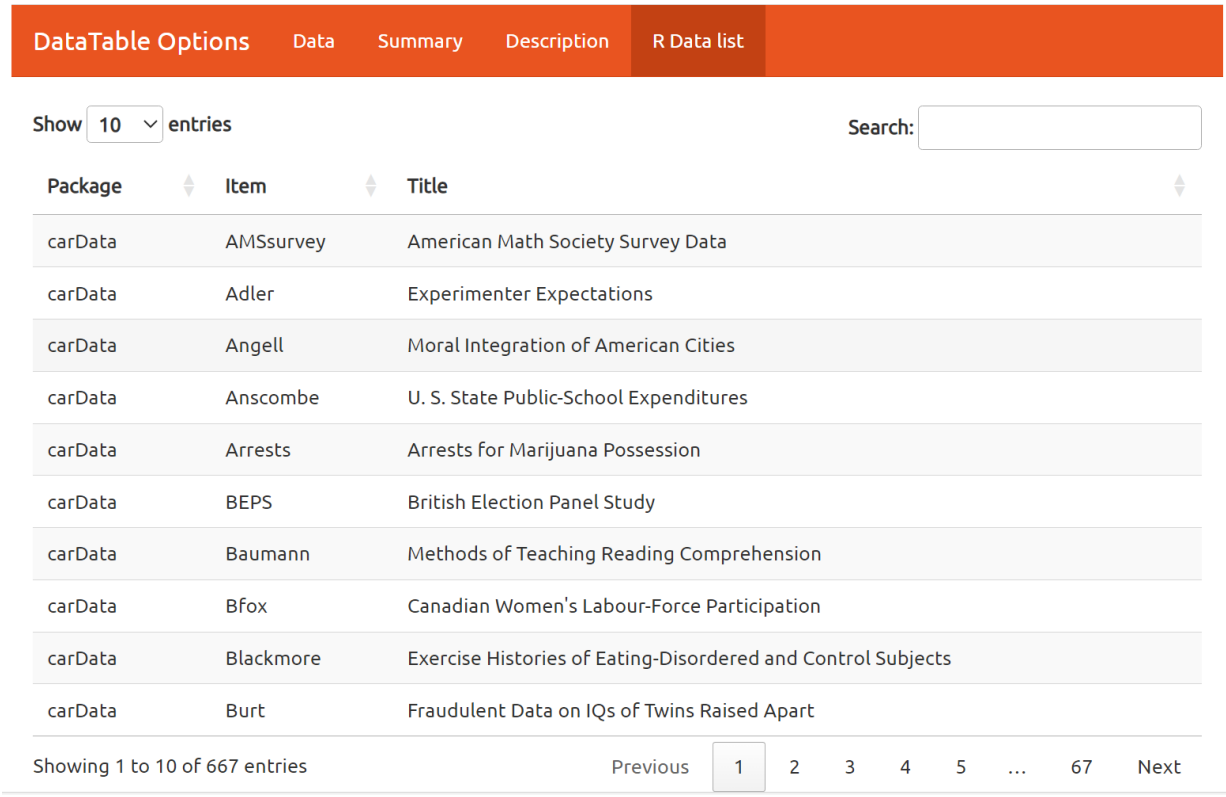


The numerical summaries of each data column are displayed under the summary subtab. This includes information such as the number of observations, number of missing values, etc., for each column.

Description

If an external data file is uploaded, the path to the temporary folder containing the uploaded data on the computer is displayed. If data are uploaded from an R package, the data description from the R package will be displayed.

R Data list



A list of available data from R packages is displayed. Please note that only the installed R packages will be listed. If a particular dataset from a specific package is not listed, then install the package and check the list again.

Variable View (Continuous)

This section explains numerical and graphical inspection techniques for continuous data to investigate if any outliers are present and if the data are normally or symmetrically distributed. Normal distribution and the absences of outliers are commonly assumed in most statistical methods. It is crucial to assess these conditions before performing any statistical analysis using continuous data. To conduct these investigations, users must utilize the **Variable View (Continuous)** tab (**Figure 12**).

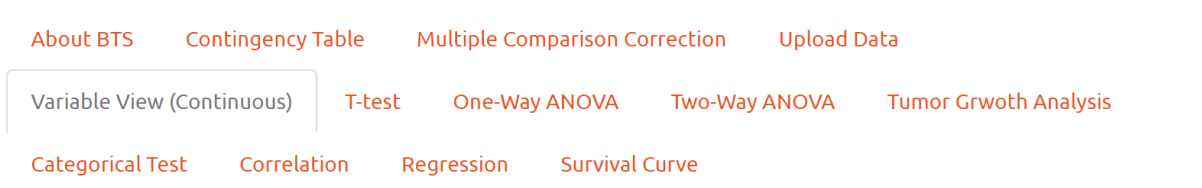


Figure 12. The **Variable View (Continuous)** tab.

Parameter Selection

A user needs to select a total of three parameters, as shown in **Figure 13**, to investigate continuous variables. The following three parameters must be decided by the user:

1. Select the desired number of bins for the histogram. The default is 2.
2. Choose a type of transformation for the continuous variable. The default is '**No transformation**'.
3. If data need to view by a group, select the group variable.

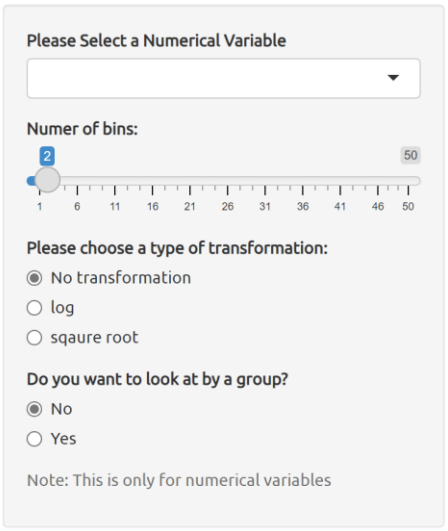


Figure 13. The parameter selection panel for the **Variable View (Continuous)** tab.

Output

The output panel (right panel) is composed of six subpanels:

- Histogram
- Density plot

- Boxplot
- Q-Q plot
- Shapiro-Wilk's normality test
- Summary statistics

The first and second subpanels display the histogram and the density plot for the continuous variable, respectively. The third subpanel shows the boxplot. The fourth and fifth subpanels present the Q-Q plot and Shapiro-Wilk's normality test results, respectively. The last subpanel provides an overall summary of the continuous variable.

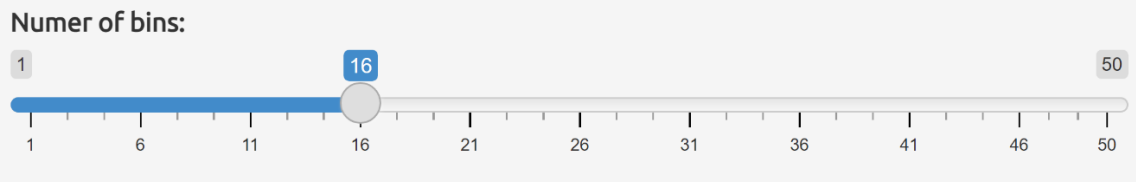
Example

The example uses the same dataset, **carData_SLID.txt**, discussed above. It aims to investigate the continuous variable of wages by grouping the variable **'sex'**.

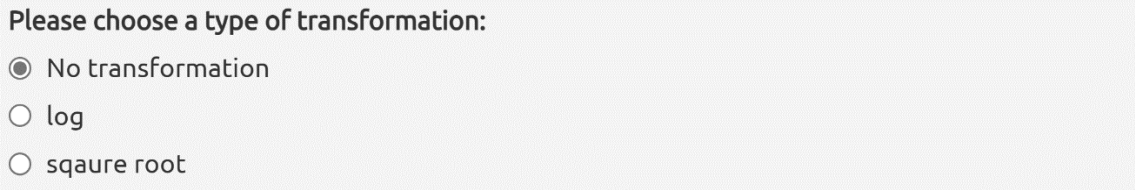
Step by Step

To upload the dataset to the app, follow the steps outlined in the "**Upload Data**" section above. Once the data is uploaded, navigate to the Variable View (Continuous) tab. Select the "**wages**" variable:

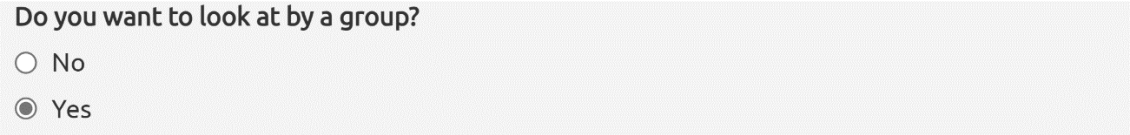
Choose the number of bins to be 16. The number 16 is chosen arbitrarily, and users can adjust this number as needed to obtain a better plot of the histogram:



Select the type of transformation for the continuous variable. Default is '**No transformation'**. If users are unsure about which transformation to apply, they should select 'No transformation' and assess the Q-Q plot, boxplot, and Shapiro-Wilk's normality test. Then, users might choose a type of transformation and reassess the Q-Q plot, boxplot, Shapiro-Wilk's normality test. This process should be repeated until a satisfactory transformation is found.



Select '**Yes**' for conducting the investigation by group. If the interest is to investigate all wages data without grouping by sex, then choose '**No**' option.



Then select the group variable:

Results and Interpretation

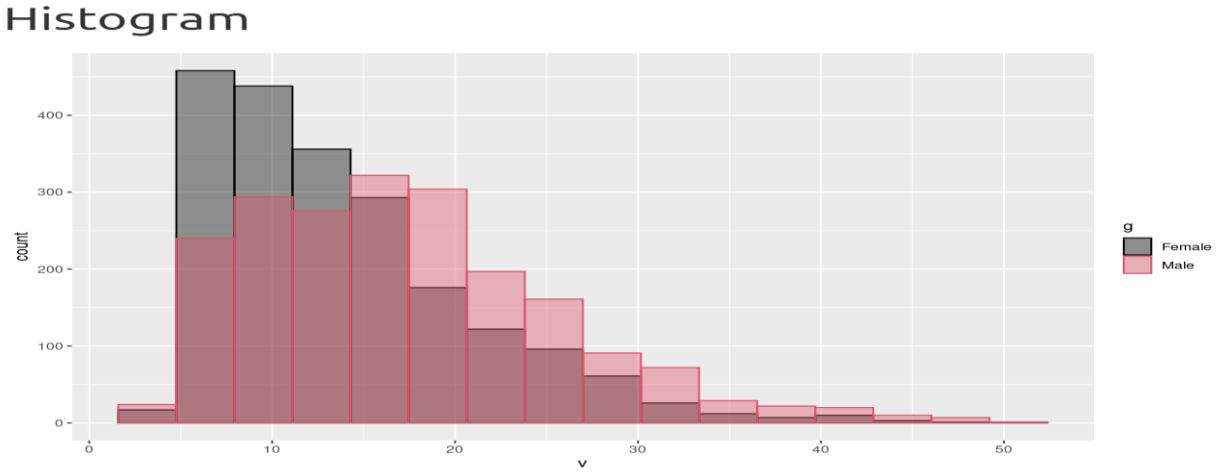
When no group investigation is performed, there is one plot and one output available within each subplot for the entire continuous data.

When an investigation is conducted by grouping, there are multiple plots and multiple numerical results available within each subplot, depending on the number of levels of the grouping variable.

Histogram

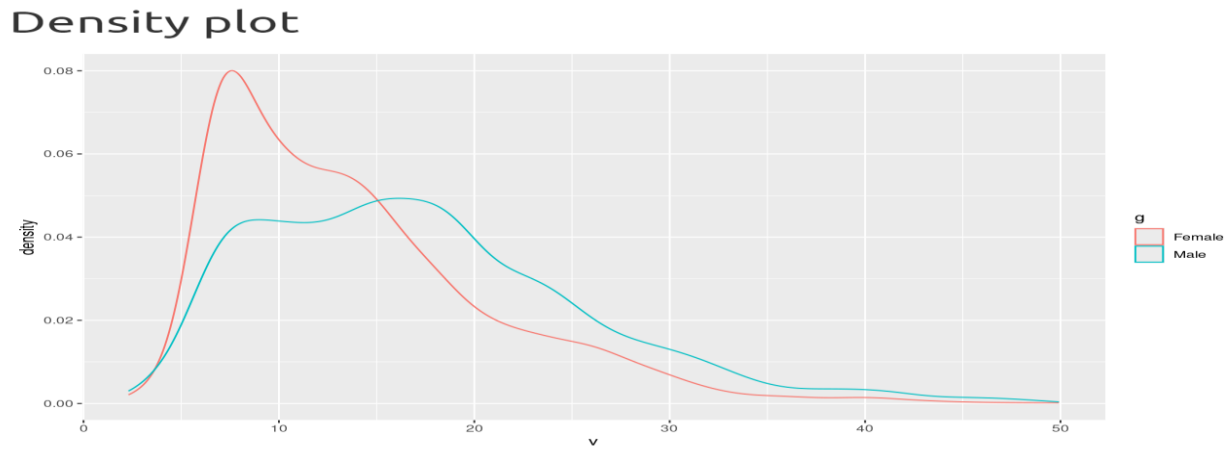
Two histograms for wages data grouping by the sex variable are overlaid in the histogram plot. The red histogram represents males, and the grey histogram represents females. Both

histograms appear to be skewed.



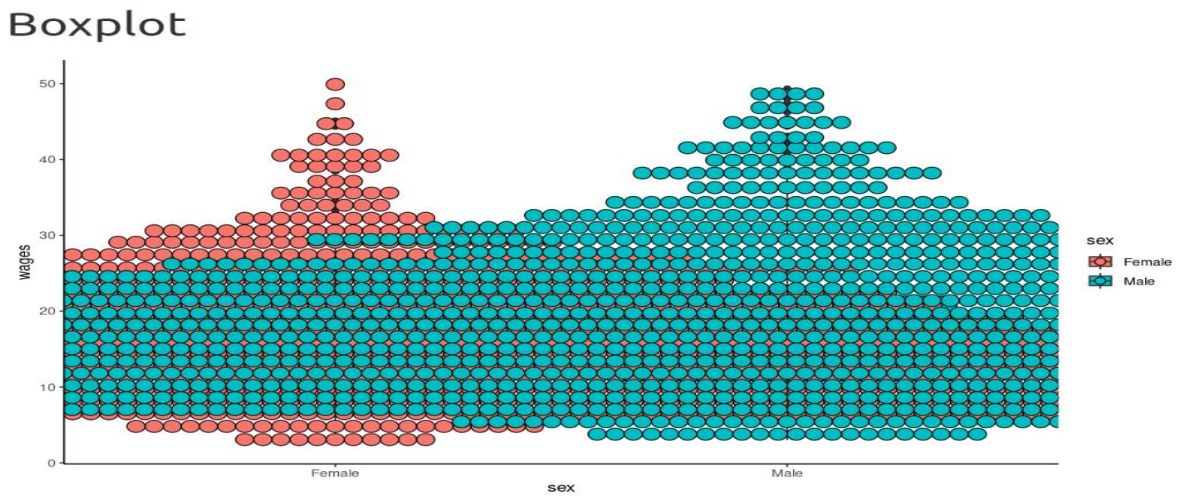
Density plot

Two density plots for wages data corresponding to males and females are overlaid in the density plot, represented in red and green. Both densities appear to be skewed.



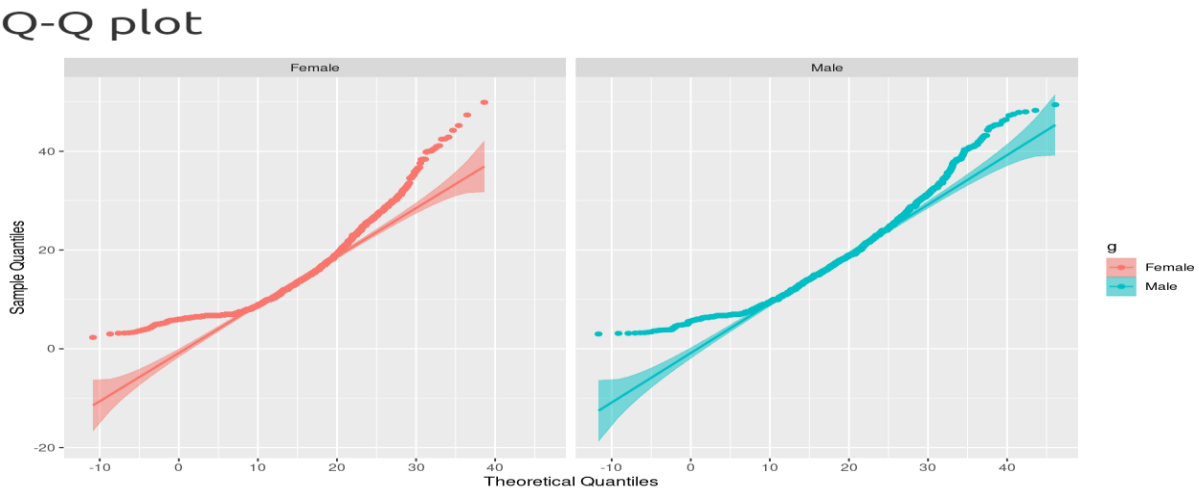
Boxplot

Two boxplots for the wages data relevant to males and females are overlaid in the boxplot.



Q-Q plot

Q-Q plots of the wages data for males and females are displayed in the Q-Q plot. Both plots largely deviate from the line and the confidence band, indicating the skewness of the distributions.



Shapiro-Wilk's normality test

The Shapiro-Wilk's normality test results for both male and female wages indicate significant deviation from the normal distribution, as both p-values are less than 0.001.

Sharpiro-Wilk's normality test

```
$Female

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.89726, p-value < 2.2e-16


$Male

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.95009, p-value < 2.2e-16
```

Summary

The wages data are available for 2077 females with a mean of 13.89 and a standard deviation of 7.09. For males, wages are available for 2070 individuals, with a mean of 17.22 and a standard deviation of 8.28. Additionally, other numerical summaries such as median, minimum, maximum, range, etc., are also available for the wages of both groups:

Summary

```
$Female
   vars    n  mean   sd median trimmed  mad min   max range skew kurtosis   se
X1    1 2077 13.89 7.09  12.35   12.93 6.63 2.3 49.92 47.62 1.26     1.76 0.16

$Male
   vars    n  mean   sd median trimmed  mad min   max range skew kurtosis   se
X1    1 2070 17.22 8.28  16.09   16.46 8.21   3 49.44 46.44 0.86     0.72 0.18
```

T-test

This section explains T-tests, including one-sample, two-sample, and paired-sample t-tests, along with the equivalent nonparametric tests of Wilcoxon signed-rank, two-sample Wilcoxon, and matched-pairs Wilcoxon tests. For parametric t-tests, data are assumed to be normally distributed, while the nonparametric Wilcoxon tests are free of distributional assumptions. To perform all these tests, users must utilize the **T-test** tab (**Figure 14**).

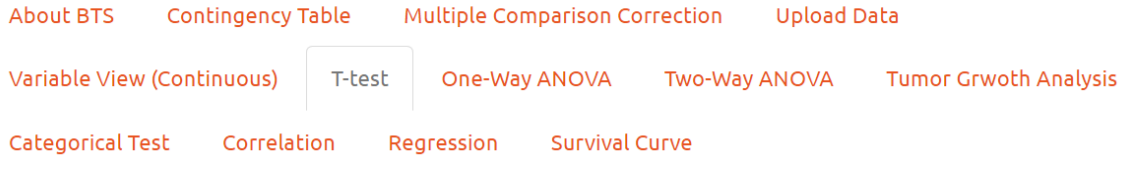


Figure 14. The **T-test** tab.

Parameter Selection

A user needs to select a total of seven parameters, as shown in **Figure 15** under the default setting. To perform a t-tests or equivalent Wilcoxon tests, the following five parameters must be decided by the user:

1. Is the test nonparametric? The default is 'No'.
2. Select a type of transformation for the data. The default is 'No transformation'.
3. Select one, two, or paired sample. The default is one sample. Note that the appearance of this panel changes upon selection of either paired or two-sample, which will be discussed further in the following sections.
4. Select the numerical variable (X1 or X).
5. Select an alternative hypothesis. The default is 'Not equal (X1 != X2)'.
6. Choose the assumed population mean under the null hypothesis. The default is 0.
7. Choose the confidence level. The default is 95%.

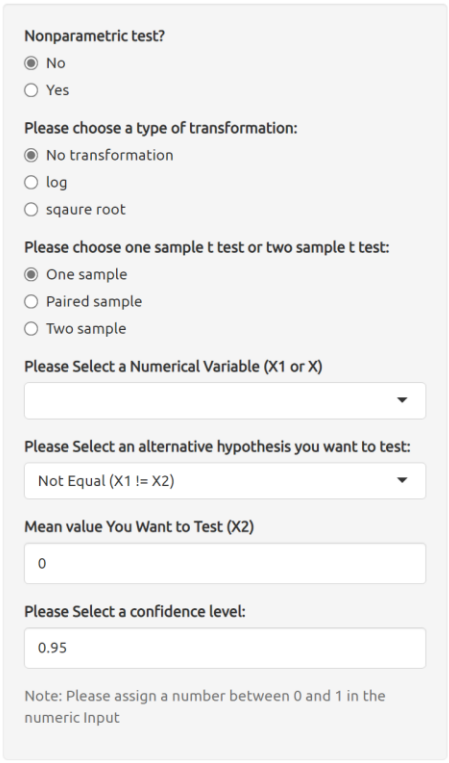


Figure 15. The parameter selection panel for the **T-test** tab.

There is an option for choosing a parametric t-test or nonparametric Wilcoxon test (**Figure 16**). A t-test is selected by default by choosing the '**No**' option for '**Nonparametric test?**'. A Wilcoxon test can be executed by choosing the '**Yes**' option. There are three other options for the type of transformation for the data (**Figure 17**): **No transformation**, **log transformation (log)**, and **square root transformation (square root)**. Based on the selection, the data will be transformed before performing a t or Wilcoxon test. If users are unsure about which transformation to apply, they should select 'No transformation' and assess the Q-Q plot, boxplot, and Shapiro-Wilk's normality test. Then, users might choose a type of transformation and reassess the Q-Q plot, boxplot, Shapiro-Wilk's normality test. This process should be repeated until a satisfactory transformation is found. Another three options for selecting one, two, or paired test, either a t or a Wilcoxon, are also available (**Figure 18**). The default setting is for one sample.

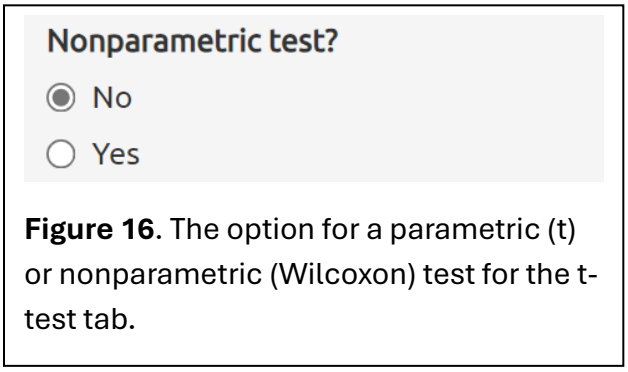


Figure 16. The option for a parametric (t) or nonparametric (Wilcoxon) test for the t-test tab.

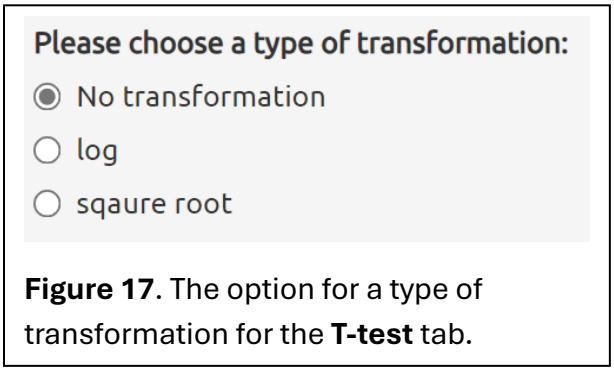


Figure 17. The option for a type of transformation for the **T-test** tab.

Output

The output panel (right panel) is composed of five subpanels for one and paired sample t-tests:

- Boxplot
- Q-Q plot
- Sharpiro-Wilk's normality test
- Summary statistics
- Hypothesis testing

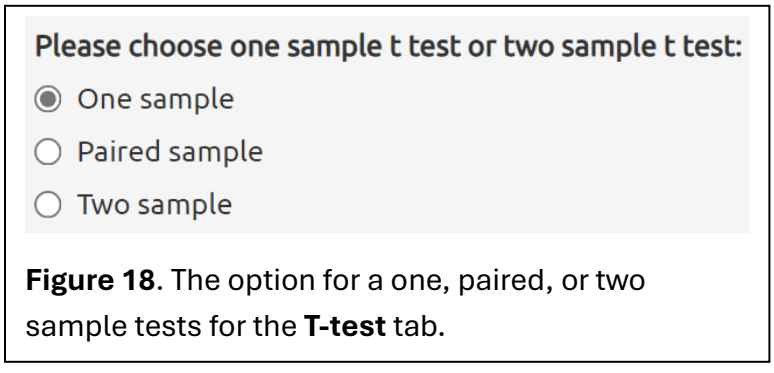


Figure 18. The option for a one, paired, or two sample tests for the **T-test** tab.

An additional subpanel of Levene's homogeneity of variance test is available for the two-sample t-tests. This subpanel displays the hypothesis test results of homogeneity of the variances of the data of the two samples. For Wilcoxon tests, the output panel consists of the following:

- Boxplot: Summarizes the data graphically.
- Summary statistics: Shows the numerical summaries of the data.
- Hypothesis testing: Displays the outcomes of hypotheses tests.

In addition, for both t-tests and Wilcoxon tests, the Q-Q plot and Shapiro-Wilk's normality test subpanels show the graphical and formal test results for evaluating the normality condition of the data.

A. Example for one-sample t-test and Wilcoxon signed-rank test

The example data, referred to as **intake data**, is accessible in the R package **ISwR**. It comprises paired values of premenstrual and postmenstrual energy intake for 11 women and consists of the following two variables:

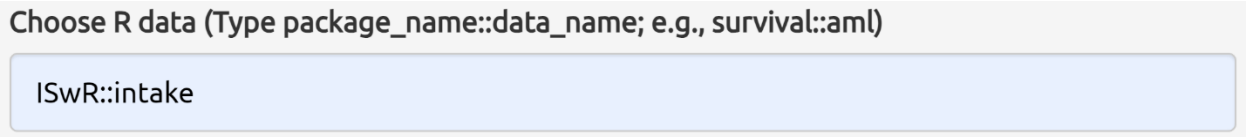
- **pre**: a numeric vector representing premenstrual intake (in kilojoules, kJ).
- **post**: a numeric vector representing postmenstrual intake (in kilojoules, kJ).

For both the one-sample t-test and the Wilcoxon signed-rank test, only the data from the "**pre**" column is utilized. However, both the "**pre**" and "**post**" paired data will be employed to demonstrate paired tests later in this section. The hypothesis is:

- premenstrual energy intake distribution might have a mean of 7725 kJ.

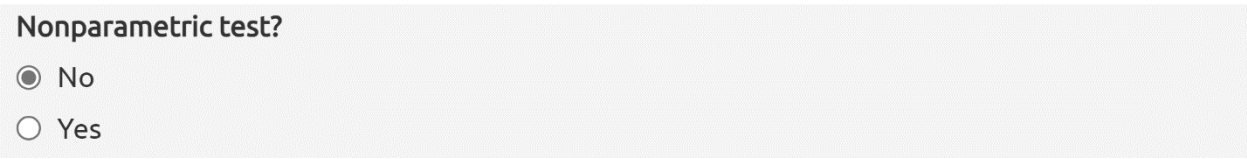
Step by Step

To upload data, go to the **Upload Data** tab and type "**ISwR::intake**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:



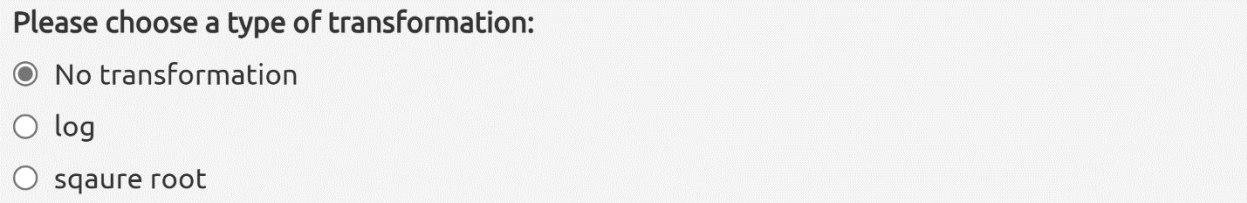
Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

Once the data is uploaded, go to the **T-test** tab. For performing t-test under the normality assumption, select '**No**' for the Nonparametric test:

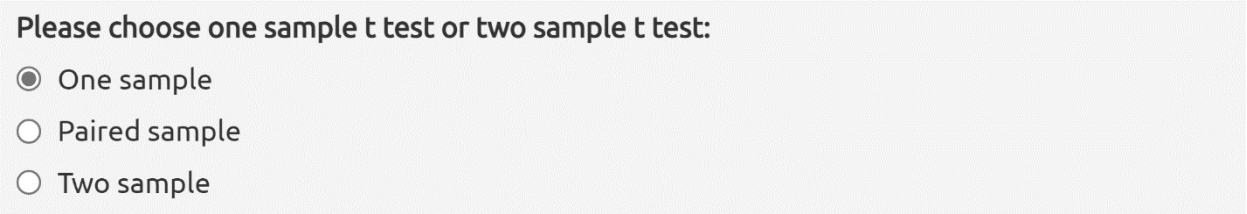


For performing distributional free nonparametric Wilcoxon signed-rank test, choose '**Yes**' for the Nonparametric test instead of '**No**' option.

Choose '**No transformation**' for the data. A user may revisit this option and select a different one if needed after assessing the normality of data:



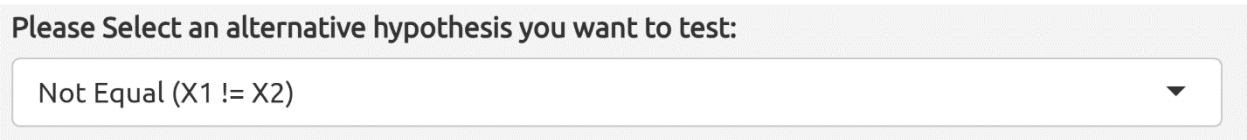
Choose '**One sample**' among One, Paired, and Two samples:



Selecting either the **'Paired sample**' or **'Two sample'** options enables the performance of paired and two-sample tests, respectively. The appearance of the parameter selection panel on the right side, as depicted in **Figure 15**, will adjust accordingly based on the chosen option, as elaborated in the subsequent example sections.

Select the **pre** data column:

The options **'Not equal (X1!=X2)**', **'Less (X1<X2)**', and **'Greater (X1>X2)**' facilitate the performance of two-tailed, left-tailed, and right-tailed tests, respectively, depending on the research question. In this example, where the alternative hypothesis is that the mean of the premenstrual energy intake is not equal to 7725, corresponding to a two-tailed test, select the default setting of **'Not equal (X1!=X2)**':



Enter the null hypothesis mean of 7725:

Mean value You Want to Test (X2)

7725

Finally, enter the 95% confidence level:

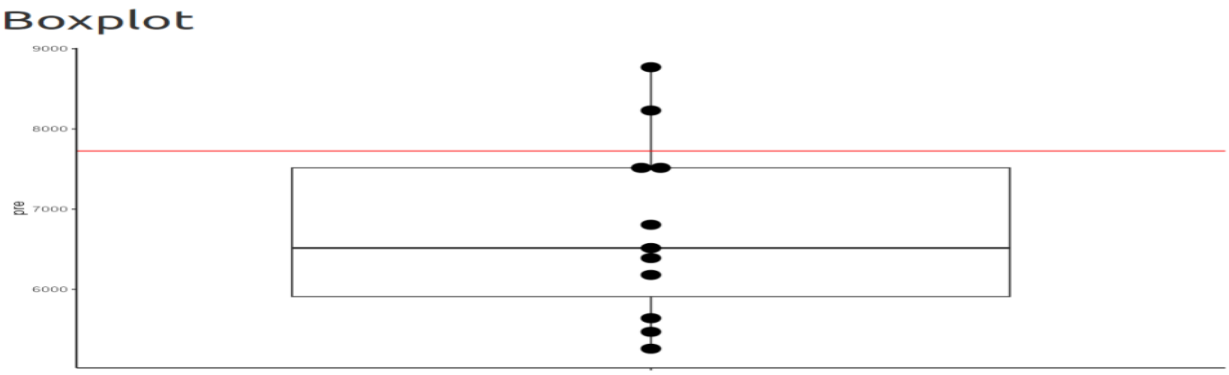
Please Select a confidence level:

0.95

Note: Please assign a number between 0 and 1 in the numeric Input

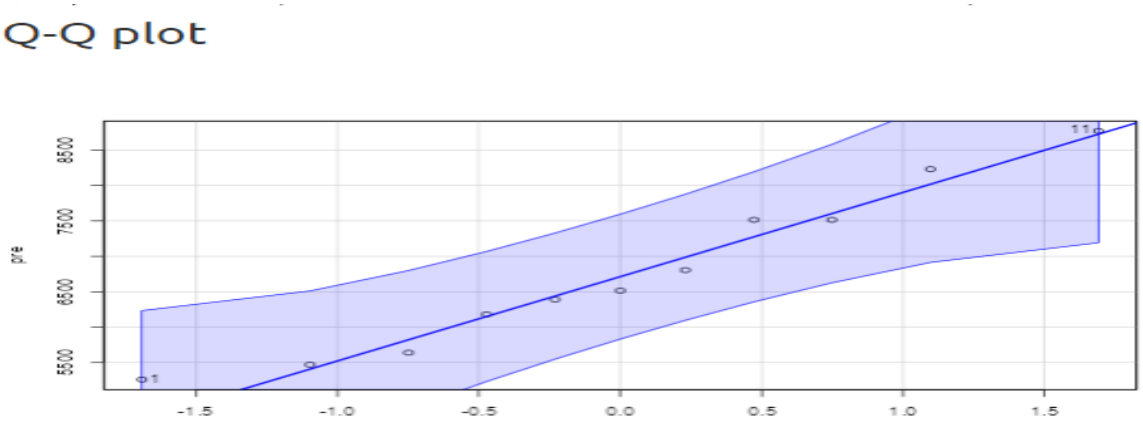
Results and Interpretation

Boxplot



The data exhibit an approximately symmetric distribution without any outliers.

Q-Q and density plots



All data points are distributed around the line, on average, and within the 95% confidence band.

Sharpiro-Wilk's normality test

Sharpiro-Wilk's normality test

```
        Shapiro-Wilk normality test

data:  wdata$val
W = 0.95237, p-value = 0.6743
```

The p-value is greater than 0.05, leading to the failure to reject the null hypothesis that the data follows a normal distribution at a two-sided 5% level.

Summary statistics

Summary statistics

```
   vars  n    mean      sd median trimmed    mad  min  max range skew kurtosis
X1    1 11 6753.64 1142.12   6515 6695.56 1482.6 5260 8770  3510 0.32    -1.33
       se
X1 344.36
```

One sample t-test results

Hypothesis testing

```
        One Sample t-test

data:  new.var1
t = -2.8208, df = 10, p-value = 0.01814
alternative hypothesis: true mean is not equal to 7725
95 percent confidence interval:
 5986.348 7520.925
sample estimates:
mean of x
 6753.636
```

The p-value obtained from the statistical test is 0.018, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis. This result indicates that there is enough evidence to conclude that the mean of the women's premenstrual energy intake is significantly different from 7725 kJ at a two-sided 5% significance level.

One sample Wilcoxon signed-rank test results

Suppose that the distribution of the data is unknown, and the Wilcoxon signed-rank test is performed by relaxing the distribution assumption. Accordingly, the '**Yes'** option is chosen for the '**Nonparametric test'**. The output consists only of a boxplot, summary statistics (which are the same as above), and the hypothesis testing results:

Hypothesis testing

```
        Wilcoxon signed rank test with continuity correction

data:  new.var1
V = 8, p-value = 0.0293
alternative hypothesis: true location is not equal to 7725
```

The same conclusion holds as the one-sample t-test, even though the p-values are different.

B. Example for paired-sample t-test and matched-pairs Wilcoxon test

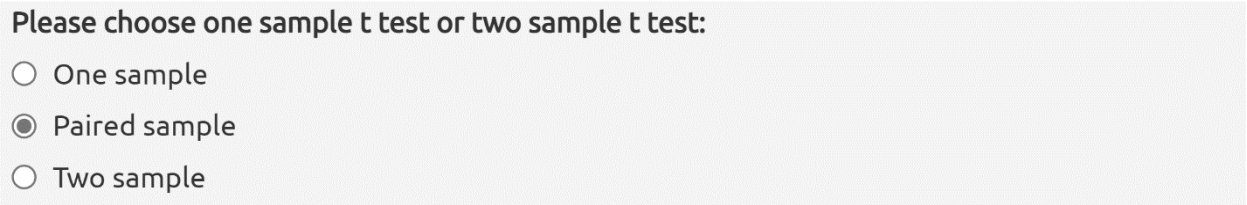
The example data is the same intake data discussed in the above section. Both premenstrual and postmenstrual paired data will be used in this section. The objective is to

test if the premenstrual and postmenstrual energy intake distributions have the same mean.

Step by Step

The steps for data upload, the choice of parametric or nonparametric tests, and data transformation are similar to those used in the above one-sample t-test and Wilcoxon signed-rank test.

Choose 'Paired sample' from among the options: One, Paired, and Two samples:



Note that the appearance of the right parameter selection panel in **Figure 15** will change to that shown in **Figure 19** upon selecting the '**Paired sample**' option. The default option for choosing the assumed mean for the null hypothesis will disappear from the parameter selection tab. An option for choosing the second data set of the paired sample will appear.

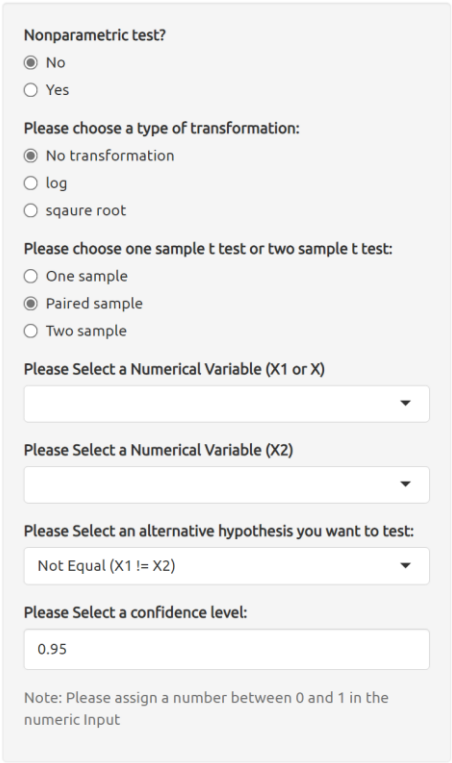
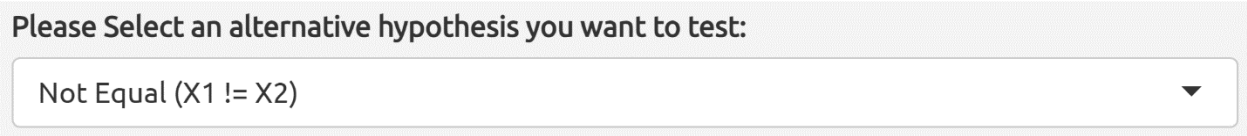


Figure 19. The parameter selection panel for the **T-test** tab, when the '**Paired sample'** option is chosen.

Select the '**pre**' data column of the paired sample:

Select the '**post**' data column of the paired sample:

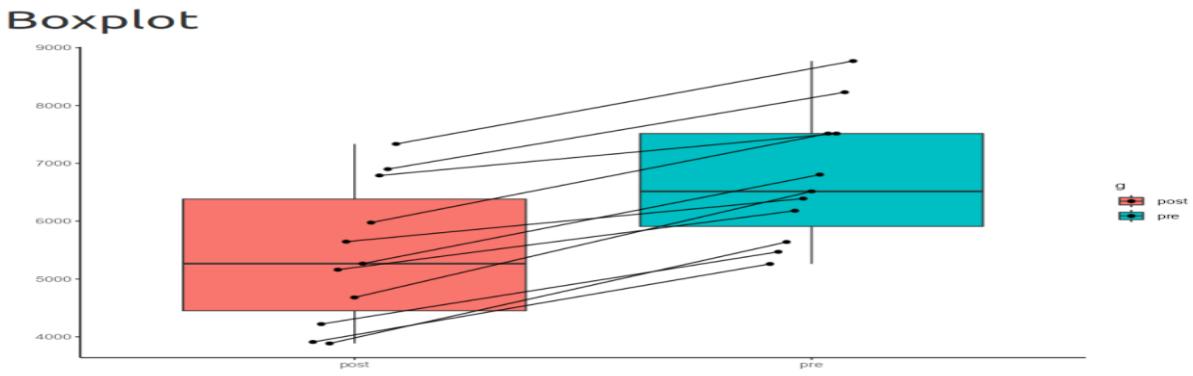
For this example, the alternative hypothesis is that the mean of the premenstrual energy intake is not equal to the mean of postmenstrual energy intake, corresponding to a two-tailed test. Accordingly, choose the default setting of '**Not equal (X1 != X2)**':



To complete the analysis, enter the 95% confidence level, similar to the one-sample t-test as shown previously.

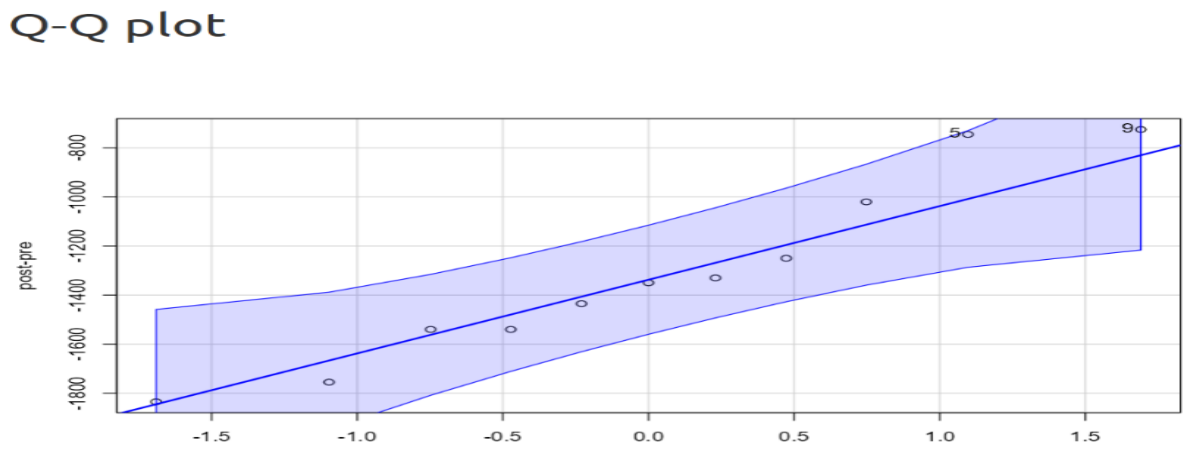
Results and Interpretation

Boxplot



Both premenstrual and postmenstrual energy intake data exhibit approximately symmetric distributions. However, upon visual examination, it appears that the premenstrual energy intake levels tend to be larger than those of the postmenstrual levels.

Q-Q and density plots



The data points representing the difference between postmenstrual and premenstrual energy intake are, on average, distributed around the central line and within the 95% confidence band.

Sharpiro-Wilk's normality test

Sharpiro-Wilk's normality test

```
        Shapiro-Wilk normality test

data:  diff.wdata
W = 0.93737, p-value = 0.4901
```

The p-value is larger than 0.05, indicating a failure to reject the null hypothesis that the difference between postmenstrual and premenstrual energy intake data follows a normal distribution at a two-sided 5% significance level.

Summary statistics

Summary statistics

```
$post
   vars  n    mean      sd median trimmed     mad  min  max range skew kurtosis
X1    1 11 5433.18 1216.83   5265 5393.89 1549.32 3885 7335  3450 0.16    -1.56
       se
X1 366.89

$pre
   vars  n    mean      sd median trimmed    mad  min  max range skew kurtosis
X1    1 11 6753.64 1142.12   6515 6695.56 1482.6 5260 8770  3510 0.32    -1.33
       se
X1 344.36
```

Paired- sample t-test results

Hypothesis testing

```
        Paired t-test

data:  new.var1$v1 and new.var1$v2
t = 11.941, df = 10, p-value = 3.059e-07
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean difference
       1320.455
```

The p-value is less than 0.001, leading to rejection of the null hypothesis. The data provide strong evidence to conclude that the mean of women's premenstrual energy intake differs from that of postmenstrual energy intake at a two-sided 5% significance level.

Matched-pairs Wilcoxon test results

By assuming the distribution of the data is unknown, the test can be performed nonparametrically using the matched-pairs Wilcoxon test. Accordingly, the **'Yes'** option is chosen for the **'Nonparametric test?'** The output includes boxplots, summary statistics (which remain the same as above), and the hypothesis testing results:

Hypothesis testing

```
        Wilcoxon signed rank test with continuity correction

data:  new.var1$v1 and new.var1$v2
V = 66, p-value = 0.00384
alternative hypothesis: true location shift is not equal to 0
```

While the p-value differs from that of the paired-sample t-test due to the use of different tests, the conclusion remains the same, as the p-value is less than 0.05.

C. Example for two-sample t-test and two-sample Wilcoxon test

The example data is sourced from the **ISwR** R package and pertains to energy expenditure data. It comprises 22 rows and 2 columns. The dataset contains information on the energy expenditure of groups consisting of lean and obese women. It includes the following two variables:

- **expend**: a numeric vector representing 24-hour energy expenditure (MJ).
- **stature**: a factor variable with levels 'lean' and 'obese'.

Under the normality assumption for the data, a two-sample t-test is implemented. By relaxing the distribution assumption, a two-sample Wilcoxon test is also implemented for illustration purposes. The objective is to test if the means of the 24-hour energy expenditure of lean and obese women are the same.

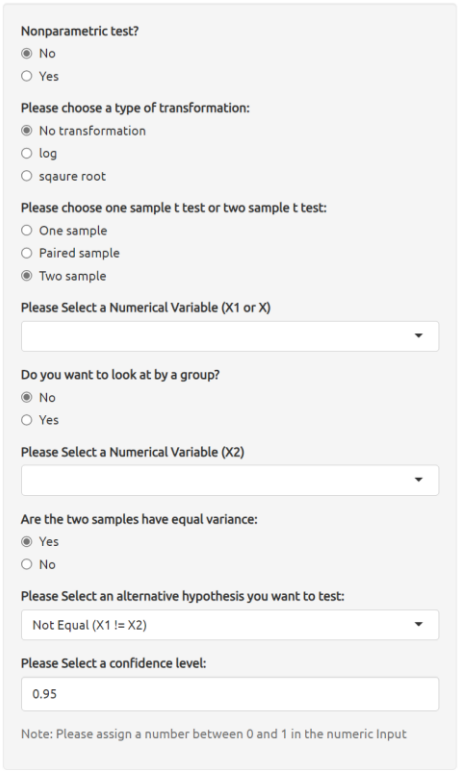
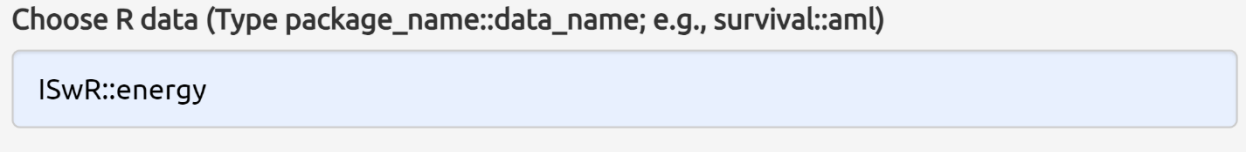


Figure 20. The parameter selection panel for the **T-test** tab, when the '**Two sample'** option is chosen.

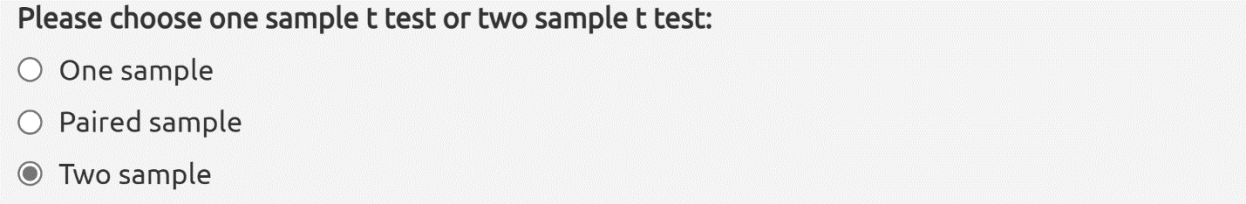
Step by Step

To upload data, go to the **Upload Data** tab and type "**ISwR::intake**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:



The choice of parametric or nonparametric tests and the transformation of data steps are similar to those used in the one-sample t-test and Wilcoxon signed-rank test. Refer to that section for further details.

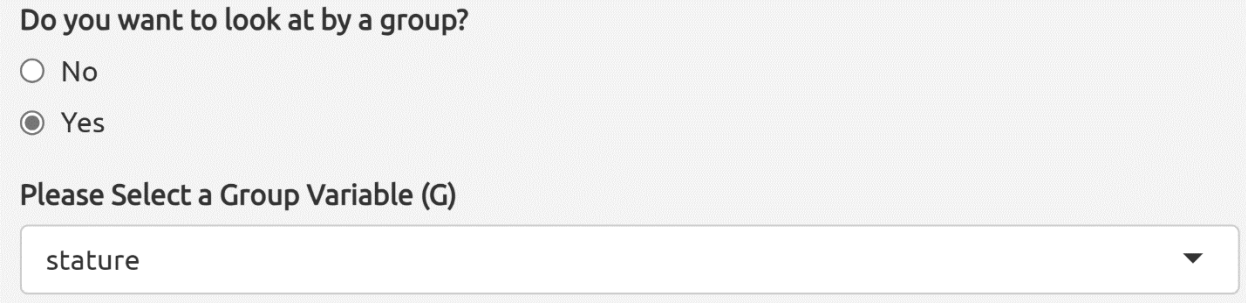
Choose '**Two sample'** among One, Paired, and Two samples:



Note that the appearance of the right parameter selection panel changes to match **Figure 20** upon selecting the '**Two-sample**' option. This is different from both **Figure 15** and **Figure 19**.

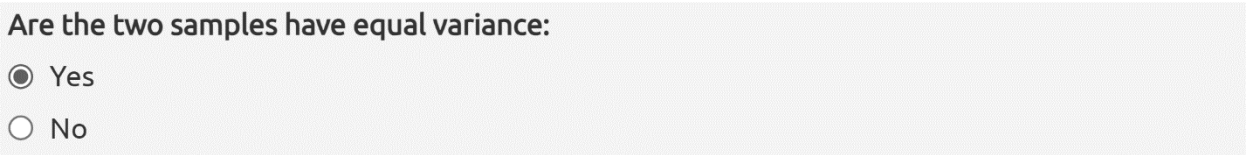
Select the **expend** data column as the sample:

Select '**Yes'** option for conducting the analysis by group and select **stature** for the grouping variable:



Note that the choice of the '**Yes'** or '**No'** option for the analysis by group depends on how the data is stored in the dataset. The expenditure values for both groups of lean and obese women have been stored in the single **expend** column of the energy data. The data can be split into two groups using the two levels (**lean** and **obese**) of the **stature** variable, which is the grouping variable. If the data is stored in this format (numerical data of both groups in one column and the grouping factor in another column), the grouping option is used by choosing the '**Yes'** option as demonstrated in this step. On the other hand, if the numerical data of the two groups are stored in two separate columns, then the '**No'** option should be chosen as the data has already been grouped, and the numerical values of the second group should be specified. If the '**No'** option is chosen, the prompt '**Please select a group variable (G)**' changes to '**Please select a numerical variable (X2)'** for specifying the data for the second group:

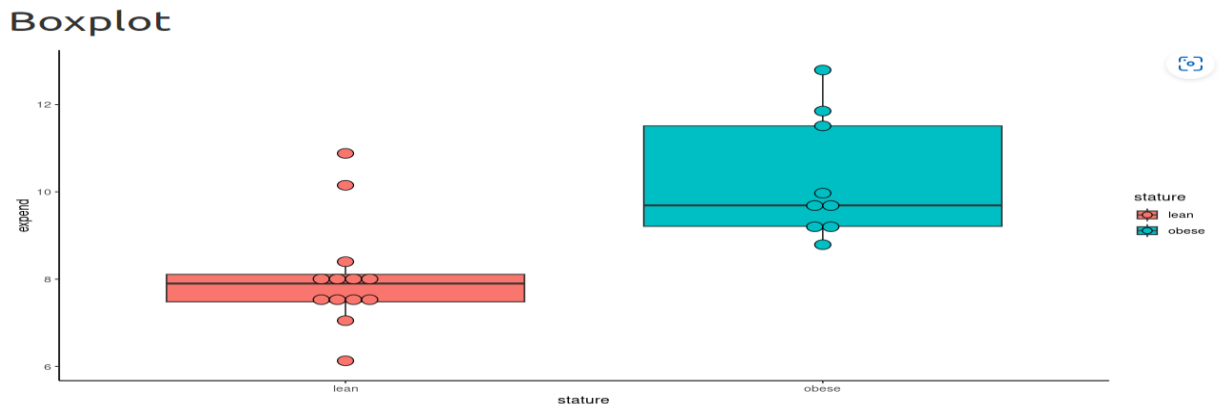
Choose '**Yes'** for homogeneous variance.



If a user is unsure about which option to choose, they should select '**Yes'** and assess the Levene's homogeneity of variance test results. If the result indicates homogeneity of variances across the two groups, the '**Yes'** option and the corresponding rest of the analysis are valid. Under this option, the app uses the pooled variance for standard error calculation, assuming equal variances of the two groups. If the Levene's homogeneity of variance test result indicates non-homogeneous variances, then the rest of the analysis should be executed by choosing the '**No**' option. Correspondingly, the app uses the unpooled variance for standard error calculation. For the alternative hypothesis, choose the '**Not equal (X1 != X2)**' option, and choose a 95% confidence level, similar to the above one-sample t-test.

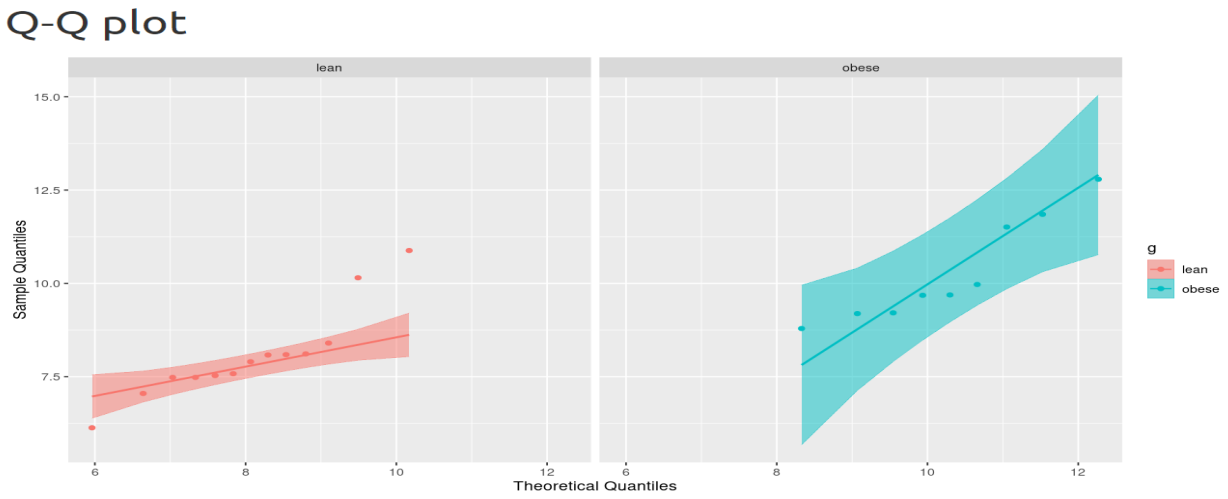
Results and Interpretation

Boxplot



Both lean and obese data appear to have slightly asymmetric distributions. The mean energy expenditure of obese women appears to be relatively larger than that of lean women.

Q-Q and density plots



The data points for energy consumption among obese women are generally distributed around the average line and within the 95% confidence band. Among lean women, most of the data also appear to be distributed around the average line and within the 95% confidence band, with the exception of three outliers lying beyond the confidence band and away from the average line.

Sharpiro-Wilk's normality test

Sharpiro-Wilk's normality test

```
$lean

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.86733, p-value = 0.04818


$obese

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87603, p-value = 0.1426
```

The p-value for the lean group is marginal at a two-sided 5% level. However, for the obese group, the p-value is larger than 0.05, indicating that there is not enough evidence to reject the null hypothesis. This suggests that the data may follow a normal distribution.

Levene's homogeneity of variance test

Levene's homogeneity of variance test

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.2677 0.6105
      20
```

The p-value is larger than 0.05, indicating that there is not enough evidence to reject the null hypothesis of homogeneous variance across the two groups.

Summary statistics

Summary statistics

```
$lean
   vars  n mean   sd median trimmed  mad  min   max range skew kurtosis   se
X1    1 13 8.07 1.24    7.9    7.99 0.62 6.13 10.88  4.75 0.91     0.18 0.34

$obese
   vars n mean  sd median trimmed  mad  min   max range skew kurtosis   se
X1    1 9 10.3 1.4   9.69    10.3 0.74 8.79 12.79     4 0.59     -1.4 0.47
```

Two- sample t-test results

Hypothesis testing

```
        Two Sample t-test

data:  v by g
t = -3.9456, df = 20, p-value = 0.000799
alternative hypothesis: true difference in means between group lean and group obese is not equal to 0
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

The p-value is less than 0.001. Therefore, the null hypothesis is rejected. The data provide enough evidence to conclude that the mean of the lean women's energy consumption is different from that of the obese women's energy consumption at a two-sided 0.1% level.

Matched-pairs Wilcoxon test results

By assuming the distribution of the data is unknown, the test can be performed nonparametrically using the two-sample Wilcoxon test. Accordingly, choose the 'Yes' option for the 'Nonparametric test?' The rest of the choices are the same as the above two-sample test options. The output only consists of boxplots, summary statistics, which are the same as above, and the hypothesis testing result:

Hypothesis testing

```
        Wilcoxon rank sum test with continuity correction

data:  v by g
W = 12, p-value = 0.002122
alternative hypothesis: true location shift is not equal to 0
```

The p-value is different from the p-value of the paired-sample t-test since the tests are different. However, the conclusion is the same as the two-sample t-test, as the p-value (0.002) is less than 0.05.

One-Way ANOVA

This section explains a one-way (or one-factor) analysis of variance (ANOVA). A one-way ANOVA is a method used to examine the effect of one independent variable (always categorical) on a single dependent variable (always continuous). To perform this analysis, you must use the **One-Way ANOVA** tab (**Figure 21**).

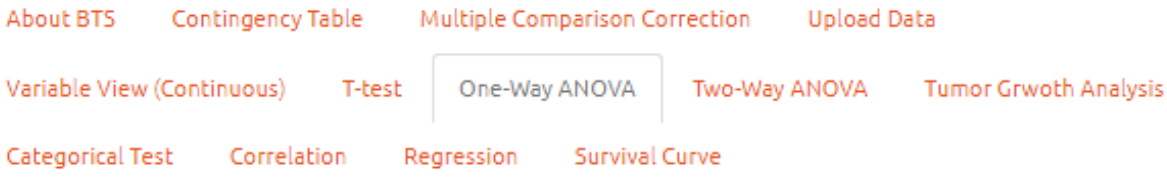


Figure21. The **One-Way ANOVA** tab.

Parameter Selection

A user needs to select a total of six parameters as shown in **Figure 22**. To perform a one-way ANOVA, the following five parameters must be decided by the user:

8. Nonparametric test? The default is 'No'.
9. Select an outcome (X1) variable.
10. Select a type of transformation for the outcome variable (X1). The default is 'No transformation'.
11. Select the factor (G).
12. Select whether the sample variance is homogeneous or not. The default is 'Yes'.
13. Select a correction method for pairwise post-hoc p-values. The default is 'Holm's'.

There are three options for the type of transformation for the outcome variable (X1): No transformation, log transformation (log), and square root transformation (square root) (**Figure 23**). Based on the selection, the outcome variable (X1) will be transformed before performing a one-way ANOVA. If a user is unsure

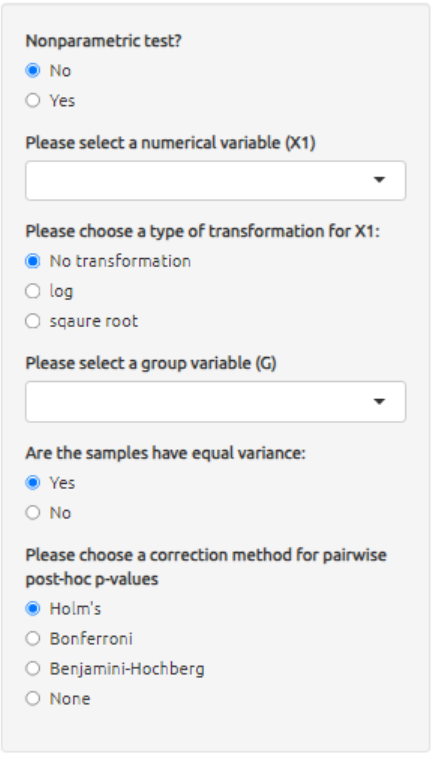


Figure 22. The parameter selection panel for the **One-Way ANOVA** tab.

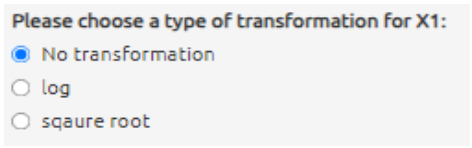


Figure 23. The option for a type of transformation for the **Two-Way ANOVA** tab.

about which transformation to apply, they should select 'No transformation' and assess the Q-Q plot and residual plot. Then, the user might choose a type of transformation and reassess the Q-Q plot, Shapiro-Wilk's normality test, and residual plot. This process should be repeated until a satisfactory transformation is found.

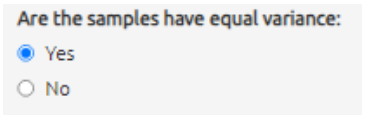


Figure 24. The option for a type of transformation for the **Two-Way ANOVA** tab.

There is an option for the homogeneity of variance condition for the outcome variable (X1) across the groups determined by the factor variable (**Figure 24**). If users are unsure about whether the variance of the outcome is homogeneous or nonhomogeneous, they should select 'Yes' and assess Bartlett's test of homogeneity of variances results. If the test indicates nonhomogeneous variances, The analysis should be conducted using the choice of 'No' for the equal variance condition.

Another parameter is for the post-hoc procedure to correct multiplicity (**Figure 25**). Four methods are available: Holm's procedure (Holm's), Bonferroni correction (Bonferroni), Benjamini-Hochberg correction (Benjamini-Hochberg), and no post-hoc procedure (None). The Bonferroni correction is the most conservative approach. The Benjamini-Hochberg correction is used to control the false discovery rate (FDR) and is typically used for omics data.

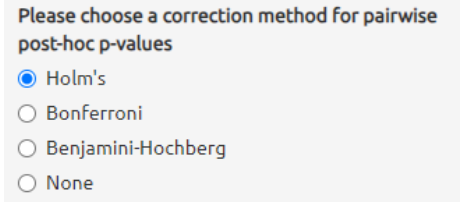


Figure 25. The option for a post-hoc procedure for the **Two-Way ANOVA** tab.

Output

The output panel (right panel) is composed of nine subpanels:

- Boxplot
- Mean plot
- Q-Q plot

- Sharpiro-Wilk's normality test
- Residual plot
- Bartlett's homogeneity test
- Summary statistics
- One-way ANOVA
- Post-hoc pairwise comparisons

The first two subpanels show the overall graphical summary of the data on the outcome variable and the mean differences among the groups. The next four describe and evaluate the distribution of residuals. The next subpanel describes the numerical summaries of the outcome among the groups. The last two subpanels display the outcomes of hypotheses tests.

Example

The example data, called **red.cell.folate data** is available in the R package **ISwR**. It contains data on red cell folate levels in patients receiving three different methods of ventilation during anesthesia and is composed of the following two variables:

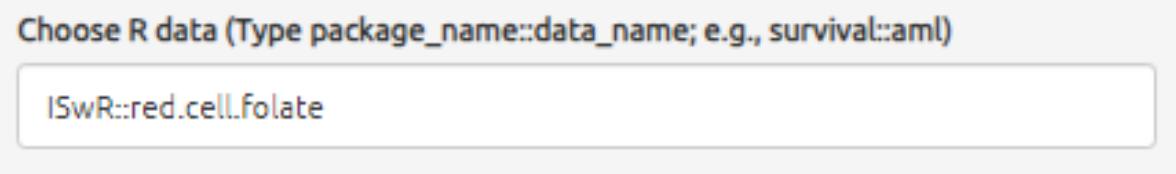
- **ventilation**: a factor with levels N2O+O2,24h: 50% nitrous oxide and 50% oxygen, continuously for 24 hours; N2O+O2, op: 50% nitrous oxide and 50% oxygen, only during operation; O2,24h: no nitrous oxide but 35%–50% oxygen for 24 hours.
- **folate**: a numeric vector, folate concentration (µg/l).

The hypothesis to investigate is:

- If the ventilation is associated with the folate concentration.

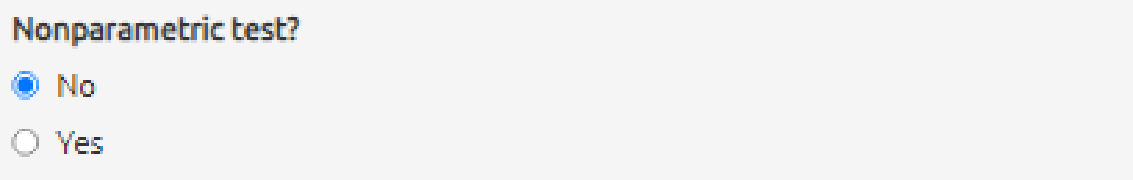
Step by Step

To upload data, go to the **Upload Data** tab and type "**ISwR:: red.cell.folate**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:



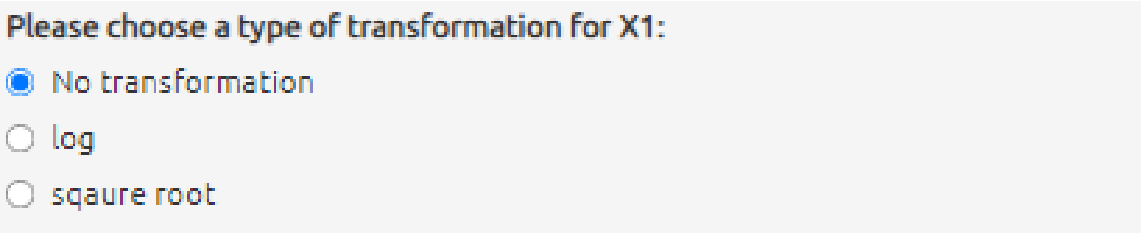
Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

Once the data is uploaded, go to the **One-Way ANOVA** tab. Then choose '**No**' for the nonparametric test. The users may select '**Yes**' option if they want to perform nonparametric equivalent Kruskal–Wallis test:

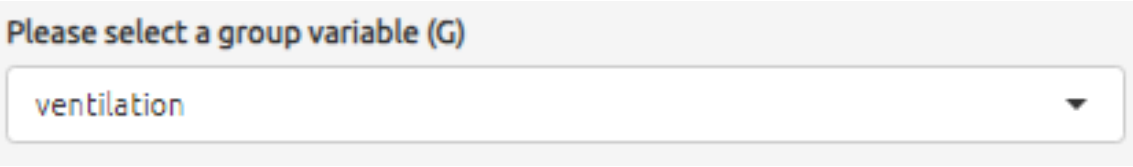


Select '**folate**' as the outcome variable (X1):

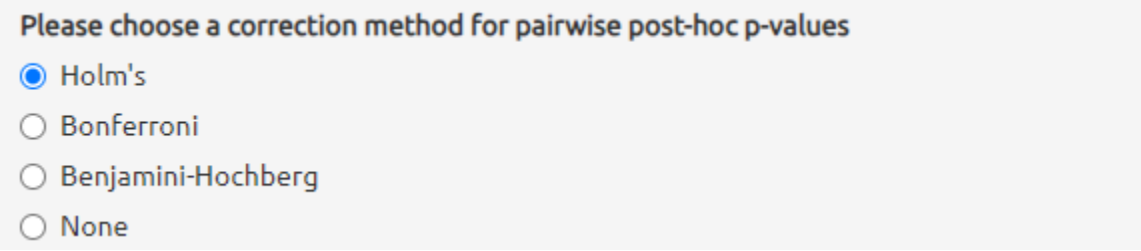
Choose '**No transformation**' for the outcome variable. A user may revisit this option and select a different one if needed after assessing the distribution of residuals:



Select '**ventilation**' as grouping/factor variable (G):



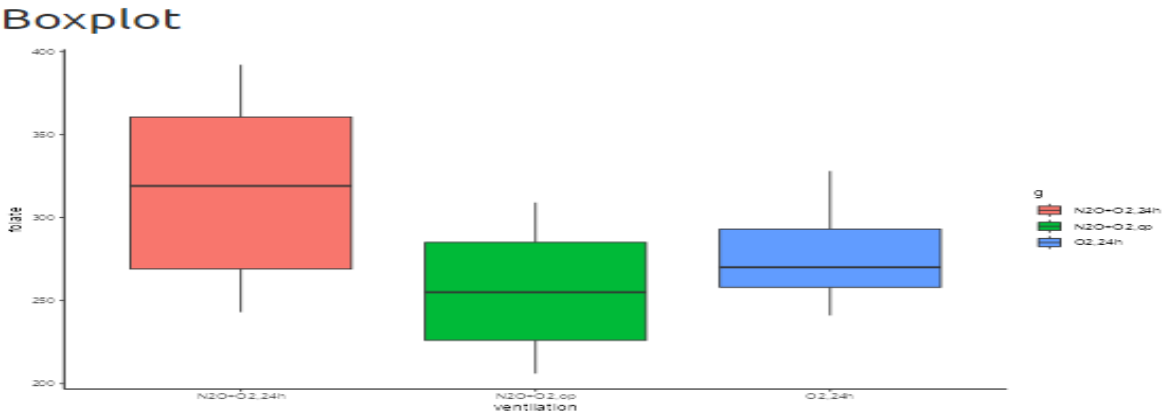
Finally, select choose '**Holm's**' as the correction method:



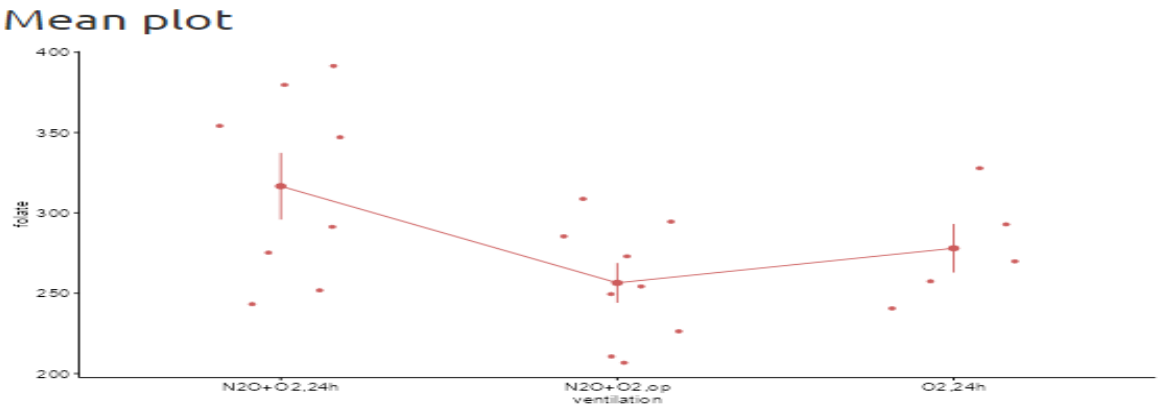
Results and Interpretation

Boxplot

The folate data are symmetrically distributed within each group of ventilation without any outliers. The average folate concentration is slightly higher under the N2O+O2,24h group among the three ventilation groups. The boxplot is given below:



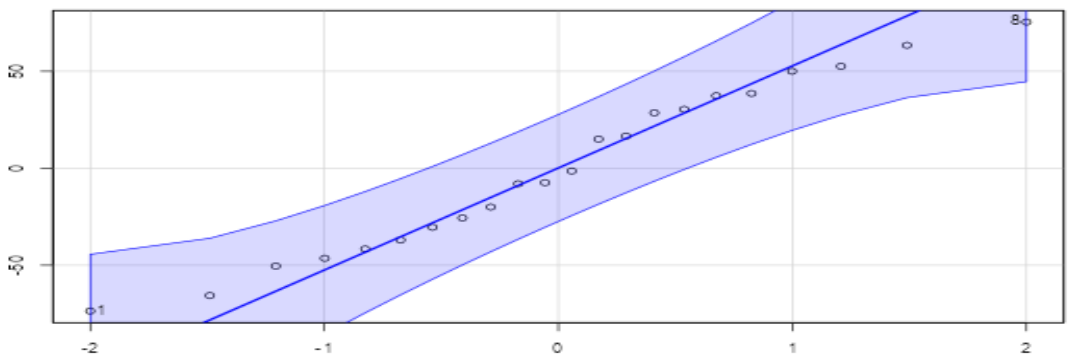
Mean plot



Like the boxplot, the average folate concentration is slightly higher under the N2O+O2,24h group among the three ventilation groups.

Q-Q plot

Q-Q plot



The data points appear to be distributed around the line and within the 95% confidence band on average, indicating approximate normality.

Sharpiro-Wilk's normality test

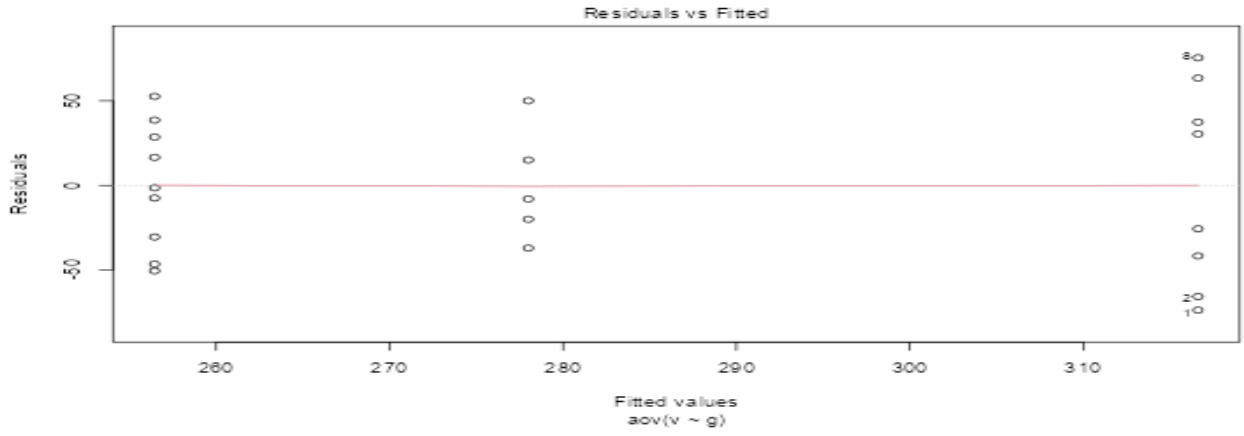
Sharpiro-Wilk's normality test

```
        Shapiro-Wilk normality test

data:  res
W = 0.966, p-value = 0.6188
```

The p-value is greater than 0.05, failing to reject the null hypothesis that the residual distribution follows a normal distribution at a two-sided 5% level.

Residual plot

Residual plot



The residuals appear to be randomly scattered around the zero line. The spread of the residuals seems fairly consistent across the range of fitted values, suggesting homoscedasticity.

Bartlett's homogeneity test

Bartlett's homogeneity test

```
        Bartlett test of homogeneity of variances

data:  v by g
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

The p-value is greater than 0.05, failing to reject the null hypothesis that the residual variance is homogeneous at a two-sided 5% level.

Summary statistics

Summary statistics

```
$`N2O+O2,24h`
   vars n   mean    sd median trimmed   mad min max range skew kurtosis    se
X1    1 8 316.62 58.72    319  316.62 77.84 243 392   149    0    -1.92 20.76

$`N2O+O2,op`
   vars n   mean    sd median trimmed   mad min max range  skew kurtosis    se
X1    1 9 256.44 37.12    255  256.44 44.48 206 309   103 -0.06    -1.68 12.37

$`O2,24h`
   vars n mean    sd median trimmed  mad min max range skew kurtosis   se
X1    1 5  278 33.76    270     278 34.1 241 328    87 0.36    -1.72 15.1
```

The sample sizes among the groups are slightly different. The average folate concentration, its standard deviation, median, and range are slightly higher for the N2O+O2,24h group compared to the other two groups. Skewness is fairly low for all groups.

One-Way ANOVA results

One-way ANOVA

```
        One-way analysis of means (not assuming equal variances)

data:  v and g
F = 2.9704, num df = 2.000, denom df = 11.065, p-value = 0.09277
```

Since the p-value is larger than 0.05, the null hypothesis of there is no association between the ventilation and folate concentration is not rejected at a two-sided 5% level. As a result, there is not enough evidence to conclude that folate concentration is associated with ventilation.

Post-hoc pairwise comparisons

Post-hoc pairwise comparisons

```
        Pairwise comparisons using t tests with non-pooled SD

data:  wdata$v and wdata$g

           N2O+O2,24h N2O+O2,op
N2O+O2,op 0.087      -
O2,24h    0.321      0.321

P value adjustment method: holm
```

This subpanel shows the pairwise comparisons among the groups. The post-hoc test will be executed only when the overall ANOVA test is significant to identify the different group. In this example, the post-hoc test has no validity as the ANOVA test is not significant. Consequently, the output is meaningless.

Two-Way ANOVA

This section explains a two-way (or two-factor) analysis of variance (ANOVA). A two-way ANOVA is a method used to examine the effect of two independent variables (always categorical) on a single dependent variable (always continuous). To perform this analysis, a user must use the **Two-Way ANOVA** tab (**Figure 26**), as the **One-Way ANOVA** tab is limited to analyses involving only one independent variable.

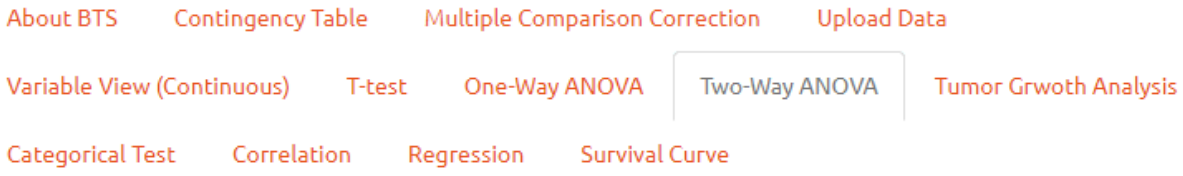


Figure 26. The **Two-Way ANOVA** tab.

The hypotheses that a two-way ANOVA considers are:

- The population means of the first factor are equal.
- The population means of the second factor are equal.
- There is no interaction between the two factors.

The assumptions for a two-way ANOVA that a user needs to consider are:

- The population (residual) is (approximately) normally distributed.
- The samples are independent.
- The variances are equal.

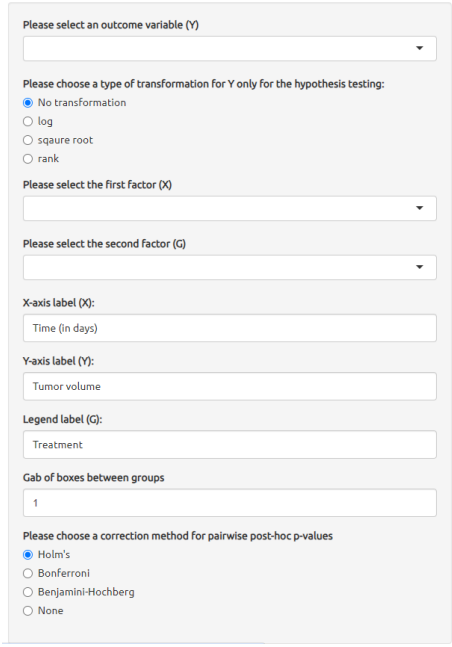


Figure 27. The parameter selection panel for the **Two-Way ANOVA** tab.

Parameter Selection

A user needs to select a total of nine parameters as shown in **Figure 27**. To perform a two-way ANOVA, the following five parameters must be decided by the user:

3. Select an outcome variable (Y)
4. Select a type of transformation for the outcome variable (Y). The default is 'No transformation'.
5. Select the first factor (X)

6. Select the second factor (G)
7. Select a correction method for pairwise post-hoc p-values. The default is 'Holm's'.

The remaining parameters are related to the plot generated by a two-way ANOVA:

8. Add the x-axis label (X). The default is 'Time (in days)'.
9. Add the y-axis label (Y). The default is 'Tumor volume'.
10. Add the legend label (G). The default is 'Treatment'.
11. Select the gap of boxes between groups. The default is 1.

There are four options for the type of transformation for the outcome variable (Y): No transformation, log transformation (log), square root transformation (square root), and rank transformation (rank) (**Figure 28**). Based on the selection, the outcome variable (Y) will be transformed before performing a two-way ANOVA. If a user is unsure about which transformation to apply, they should select 'No transformation' and assess the Q-Q plot, density plot, and residual plot. Then, the user might choose a type of transformation and reassess the Q-Q plot, density plot, Shapiro-Wilk's normality test, and residual plot. This process should be repeated until a satisfactory transformation is found.

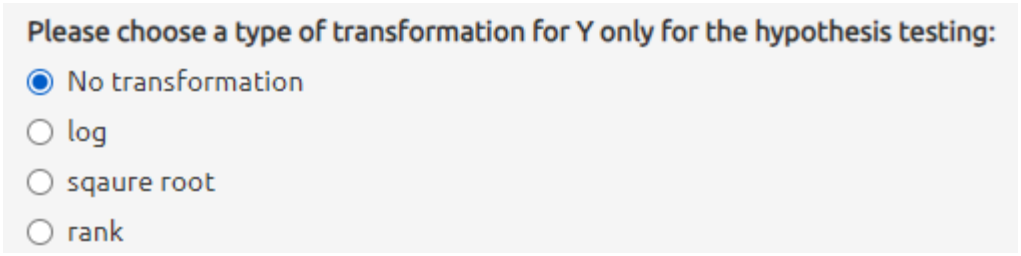


Figure 28. The option for a type of transformation for the **Two-Way ANOVA** tab.

Another parameter is for the post-hoc procedure to correct multiplicity (**Figure 29**). Four methods are available: Holm's procedure (Holm's), Bonferroni correction (Bonferroni), Benjamini-Hochberg correction (Benjamini-Hochberg), and no post-hoc procedure (None). The Bonferroni correction is the most conservative approach. The Benjamini-Hochberg correction is used to control the false discovery rate (FDR) and is typically used for omics data.

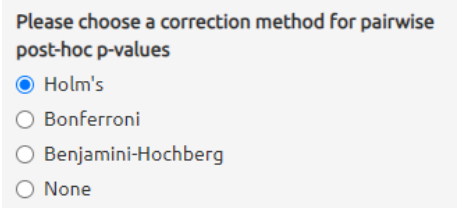


Figure 29. The option for a post-hoc procedure for the **Two-Way ANOVA** tab.

Output

The output panel (right panel) is composed of seven subpanels:

- Q-Q and density plot
- Sharpiro-Wilk's normality test
- Residual plot
- Interaction plot
- Box plot
- Summary
- Post-hoc pairwise comparisons

The first three subpanels describe and evaluate the distribution of residuals. The next two subpanels show the overall graphical summary of the data and the interaction between two factors on the outcome variable. The last two subpanels display the outcomes of hypotheses tests.

Example

The example data, called **Coking** data, is available in the R package **ISwR**. It contains the time to coking in an experiment where the oven width and temperature were varied and is composed of the following three variables:

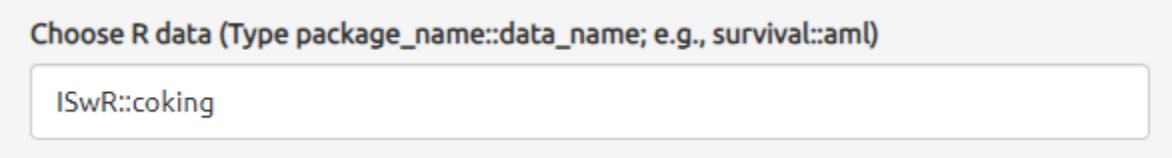
- **width**: a factor with levels 4, 8, and 12, giving the oven width in inches.
- **temp**: a factor with levels 1600 and 1900, giving the temperature in Fahrenheit.
- **time**: a numeric vector, time to coking.

The hypotheses to investigate are:

- If the time to coking is associated with the temperature (Main effect).
- If the time to coking is associated with the oven width (Main effect).
- If there is an interaction between the temperature and the oven width on the time to coking (Interaction effect).

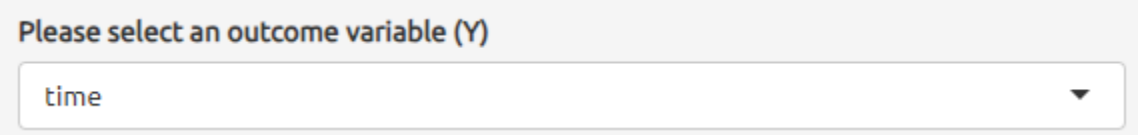
Step by Step

To upload data, go to the **Upload Data** tab and type "**ISwR::coking**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:

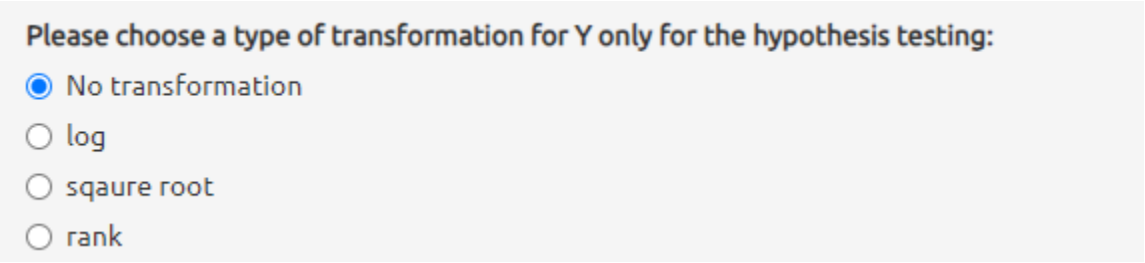


Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data. In case the coking data is not in the **R Data list**, please ensure that the R package *ISwR* has been installed.

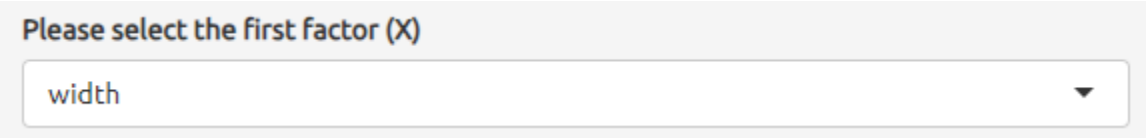
Once the data is uploaded, go to the **Two-Way ANOVA** tab. Then select '**time**' as the outcome variable:

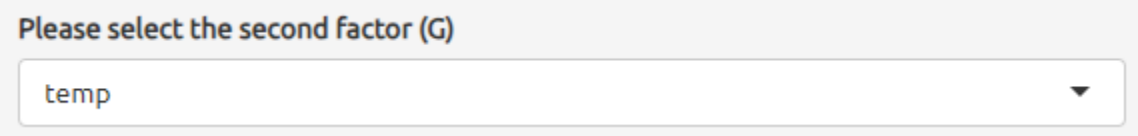


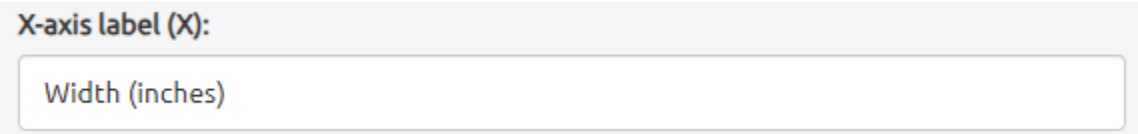
Choose '**No transformation**' for the outcome variable. A user may revisit this option and select a different one if needed after assessing the distribution of residuals:

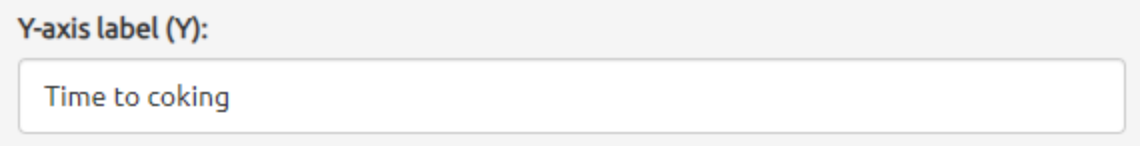


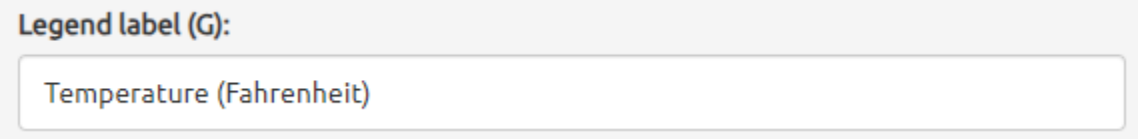
Select the first factor (X) and the second factor (Y), along with typing the labels and the gap size:



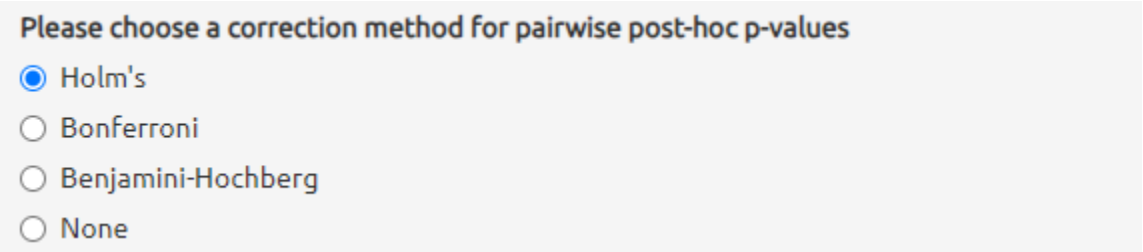






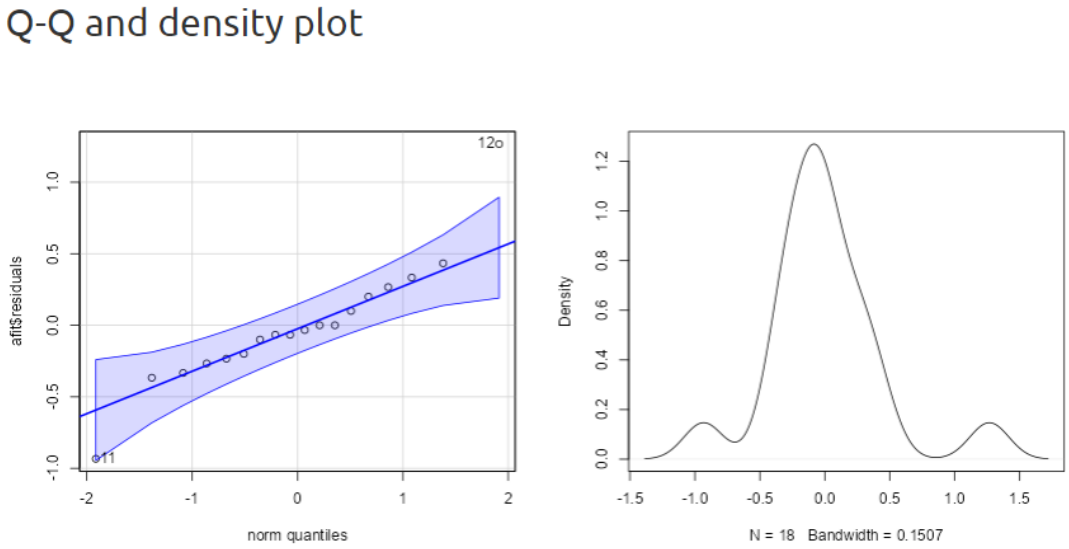


Finally, select choose '**Holm's**' as the correction method:



Results and Interpretation

Q-Q and density plots



One data point (labelled 12) is outside the 95% confidence band, but the density plot appears to be fairly symmetric.

Sharpiro-Wilk's normality test

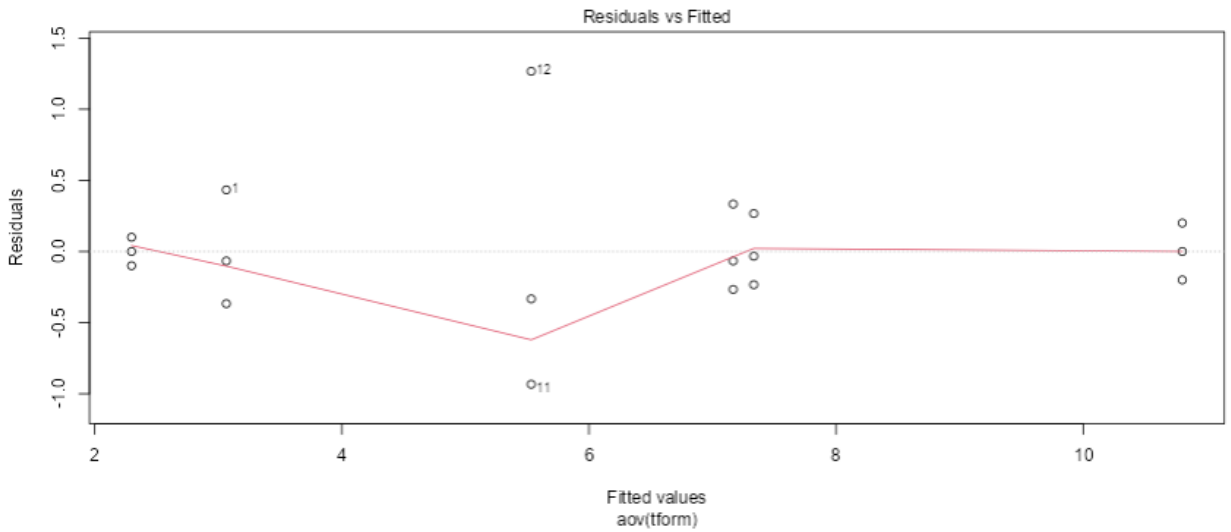
Sharpiro-Wilk's normality test

```
        Shapiro-Wilk normality test

data:  res
W = 0.89984, p-value = 0.05708
```

The p-value is marginal, failing to reject the null hypothesis that the residual distribution follows a normal distribution at a two-sided 5% level.

Residual plot

Residual plot



The residuals appear to be randomly scattered around the zero line. Although there are two data points (labeled 11 and 12) that are somewhat distant from the zero line, the spread of the residuals seems fairly consistent across the range of fitted values, suggesting homoscedasticity.

Interaction and box plots

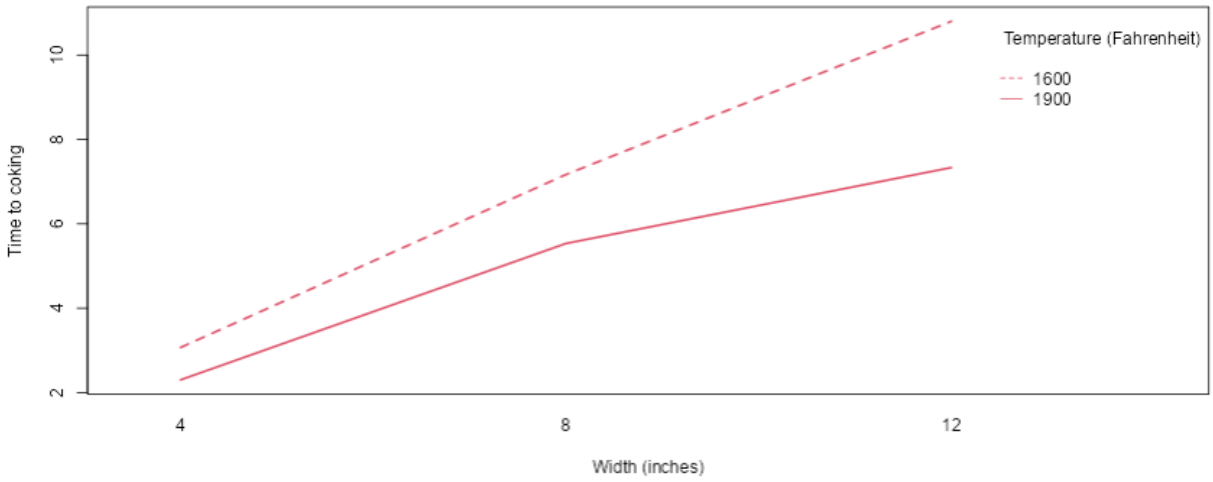
The interaction and box plots can be summarized as follow:

- Main Effect of Oven Width: As oven width increases, the time to coking increases.
- Main Effect of Temperature: At a lower temperature (1600°F), the time to coking is longer compared to a higher temperature (1900°F).

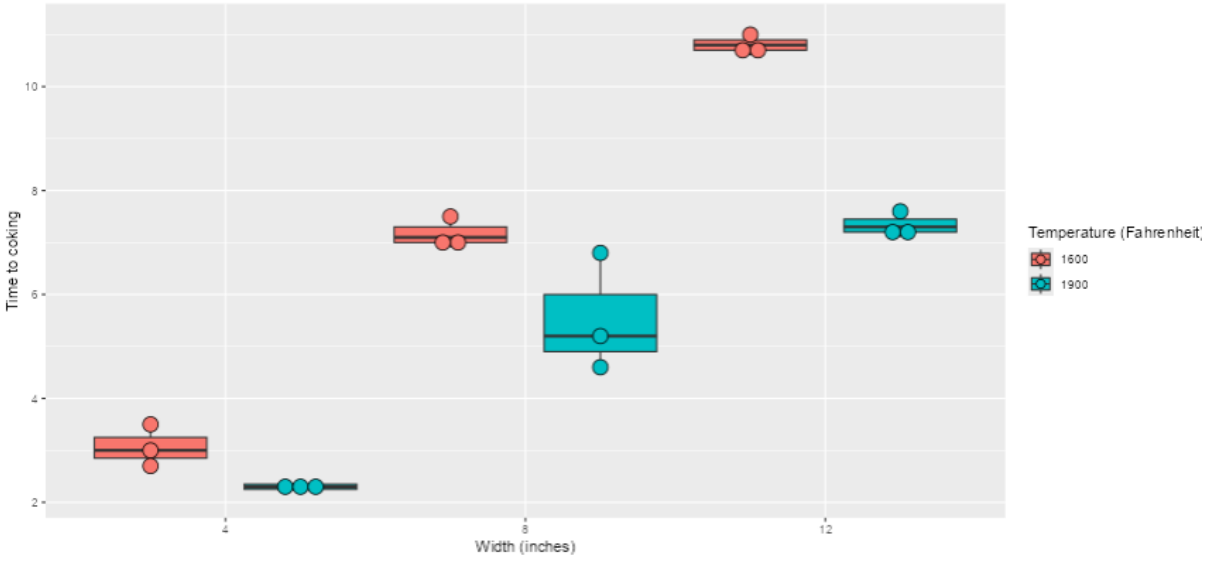
- Interaction Effect: The non-parallel lines suggest that the effect of oven width on time to coking depends on the temperature, indicating an interaction between these two factors.

In conclusion, both oven width and temperature affect the time to coking, and their effects are not independent of each other, indicating an interaction. The interaction and box plots are below:

Interaction plot



Box plot



Two-Way ANOVA results

Summary

```
            Df Sum Sq Mean Sq F value   Pr(>F)
width        2 123.14   61.57  222.10 3.31e-10 ***
temp         1  17.21   17.21   62.08 4.39e-06 ***
width:temp   2   5.70    2.85   10.28   0.0025 **
Residuals   12   3.33    0.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results can be summarized as follows:

- **width**: The oven width has a highly significant effect on the time to coking. The very low p-value indicates strong evidence against the null hypothesis, suggesting that changes in oven width significantly affect the time to coking.
- **temp**: The temperature has a highly significant effect on the time to coking. The very low p-value indicates strong evidence against the null hypothesis, suggesting that temperature significantly affects the time to coking.
- **width:temp**: There is a significant interaction effect between oven width and temperature on the time to coking. The p-value is low, indicating that the effect of one factor depends on the level of the other factor.

Overall, in terms of main effects, both oven width and temperature have highly significant effects on the time to coking. For the interaction effect, there is a significant interaction between oven width and temperature, indicating that the effect of oven width on time to coking depends on the temperature level.

Post-hoc pairwise comparisons

Post-hoc pairwise comparisons

```
  width temp1 temp2 unadjusted.p adjusted.p
1     4  1900  1600       0.0332     0.0665
2     8  1900  1600       0.0742     0.0742
3    12  1900  1600       0.0000     0.0001
```

This subpanel shows the unadjusted and adjusted p-values for the pairwise comparisons between temperatures given a specific width.

Tumor Growth Analysis

This section explains a tumor growth analysis, which is a special example of a two-way ANOVA with a time variable as a factor. To perform this analysis, a user must use the **Tumor Growth Analysis** tab (**Figure 30**).

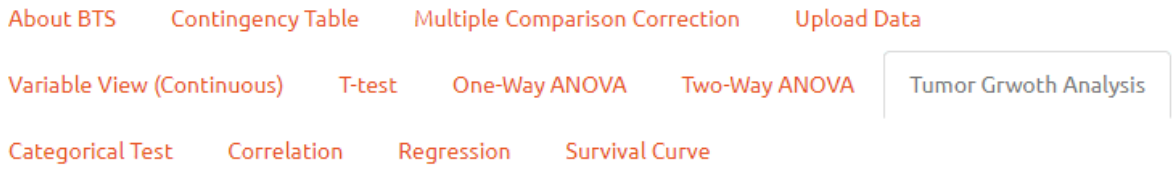


Figure 30. The **Tumor Growth Analysis** tab.

The hypothesis that a tumor growth analysis considers is whether a tumor growth is associated with groups, as well as whether tumor size/volume is associated with groups. The assumptions for a tumor growth analysis are the same as those for a two-way ANOVA, as follows:

- The population (residual) is (approximately) normally distributed.
- The samples are independent.
- The variances are equal.

Parameter Selection

A user needs to select a total of nine parameters as shown in **Figure 31**. To perform a tumor growth analysis, the following six parameters must be decided by the user:

1. Select a variable for individual IDs (I)
2. Select an outcome variable (Y).
3. Select a time variable (T)
4. Select a group variable (G)
5. Select a type of error bar in the mean plot. The default is 'mean+-SD'.
6. Select the type of transformation for the outcome variable (Y). The default is 'No transformation'.

The remaining parameters are related to the labels for the plots generated by a tumor growth analysis:

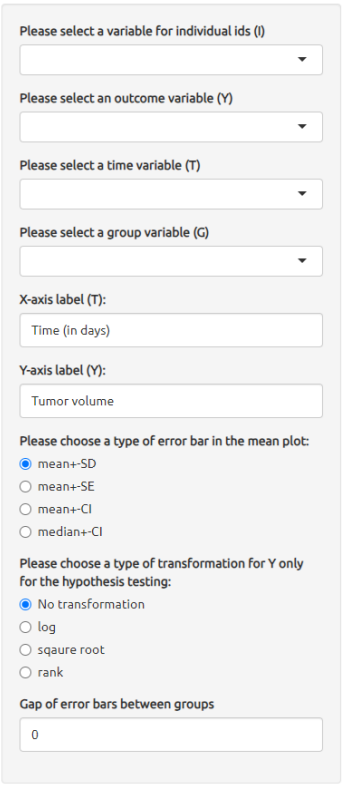


Figure 31. The parameter selection panel for the **Tumor Growth Analysis** tab.

7. Add the x-axis label (T). The default is 'Time (in days)'.
8. Add the y-axis label (Y). The default is 'Tumor volume'.
9. Select the gap of boxes between groups. The default is 0.

Similar to a two-way ANOVA, there are four options for the type of transformation for the outcome variable (Y): No transformation, log transformation (log), square root transformation (square root), and rank transformation (rank) (**Figure 31**). Based on the selection, the outcome variable (Y) will be transformed before performing a tumor growth analysis. If a user is unsure about which transformation to apply, they should select 'No transformation' and assess the Q-Q plot, density plot, and residual plot. Then, the user might choose a type of transformation and reassess the Q-Q plot, density plot, Shapiro-Wilk's normality test, and residual plot. This process should be repeated until a satisfactory transformation is found.

Output

The output panel (right panel) is composed of seven subpanels:

- Spider plot
- Spider plot by group
- Mean growth plot
- Linear mixed-effects model summary
- Q-Q and density plot
- Sharpiro-Wilk's normality test
- Residual plot

The first three subpanels provide the overall graphical summary of the data and the growth rate. The next subpanel shows the results for a tumor growth analysis using a linear mixed-effects analysis. The last three subpanels describe and evaluate the distribution of residuals.

Example

The example data is a toy example and a hypothetical data provided by the BTS. It contains the tumor size at each week for each mouse by group and is composed of the following four variables:

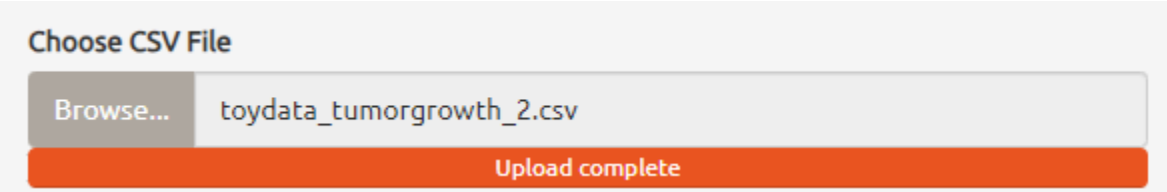
- **treatment**: a factor with groups NoTx (No treatment) and Tx (Treatment).
- **week**: a numeric factor, time points to measure the tumor size in week.

- **tumor_size**: a numeric vector, tumor size measured at a certain week.
- **mouse**: a numeric factor, a mouse ID

The hypothesis to investigate is a tumor growth is associated with a treatment.

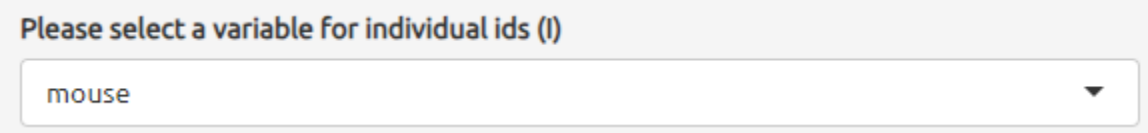
Step by Step

To upload data, go to the **Upload Data** tab and choose the csv file "**toydata_tumorgrowth_2.csv**":

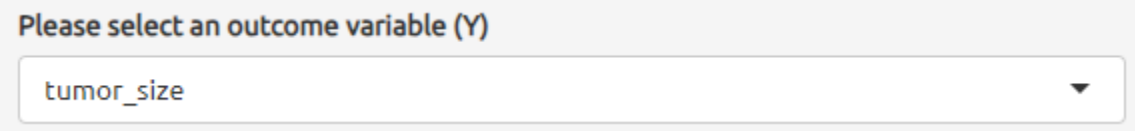


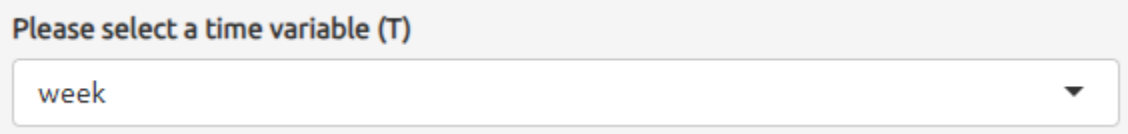
Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

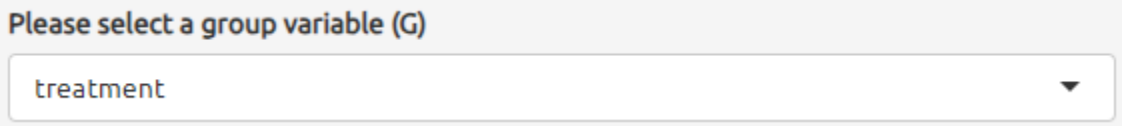
Once the data is uploaded, go to the **Tumor Growth Analysis** tab. Then select '**mouse**' as the individual IDs:

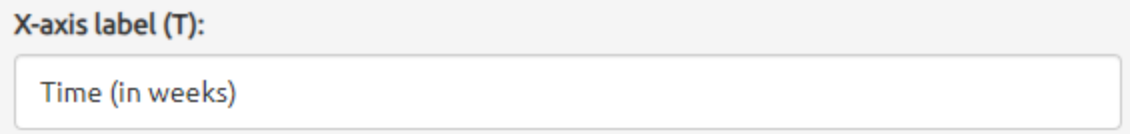


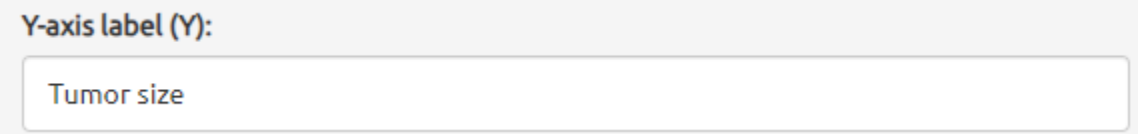
Select the outcome variable (Y), a time variable (T), and a group variable (G), along with typing the labels:



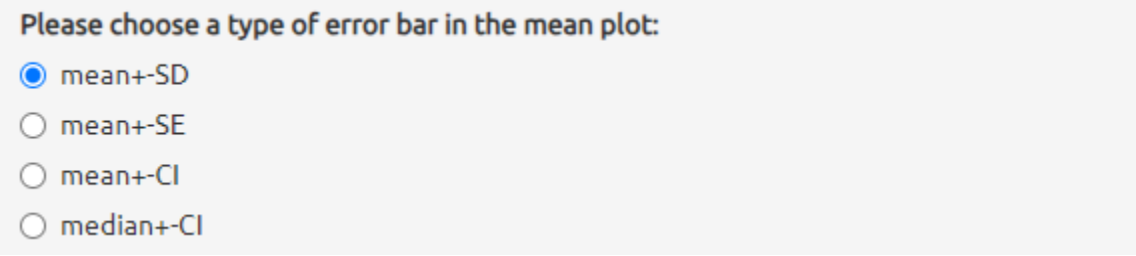




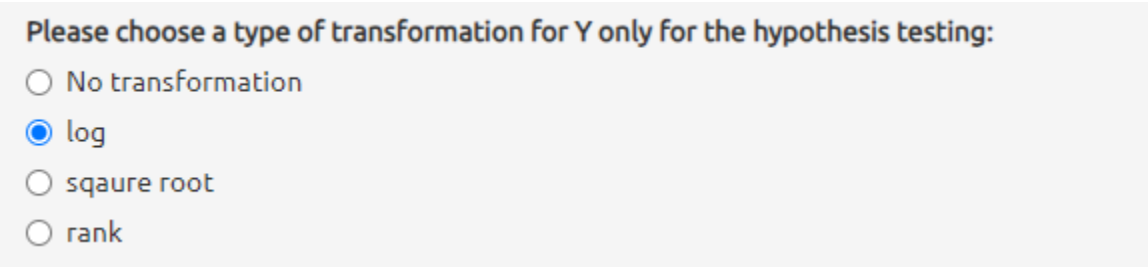




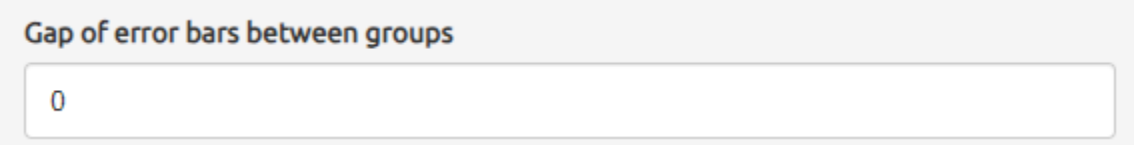
Then choose the type of error in the mean plot:



Choose '**log**' for the outcome variable. A user may revisit this option and select a different one if needed after assessing the distribution of residuals:



Finally, select the gap size:

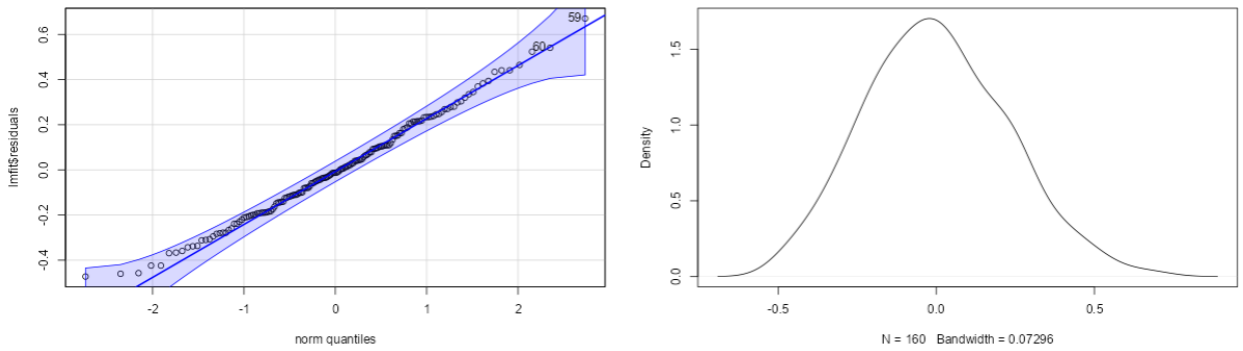


Results and Interpretation

Q-Q and density plots

Although these are not in the first subpanel, it is highly recommended to check the distribution of residuals before interpreting results. No data point is outside the 95% confidence band, and the density plot appears to be fairly symmetric.

Q-Q and density plot



Sharpiro-Wilk's normality test

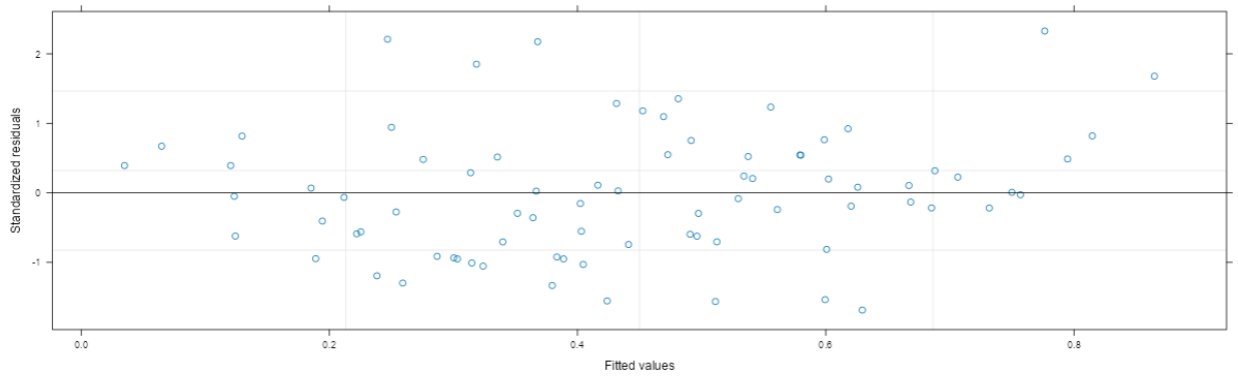
Sharpiro-Wilk's normality test

```
        Shapiro-Wilk normality test

data:  res
W = 0.99221, p-value = 0.5392
```

The p-value is large, failing to reject the null hypothesis that the residual distribution follows a normal distribution at a two-sided 5% level.

Residual plot

Residual plot

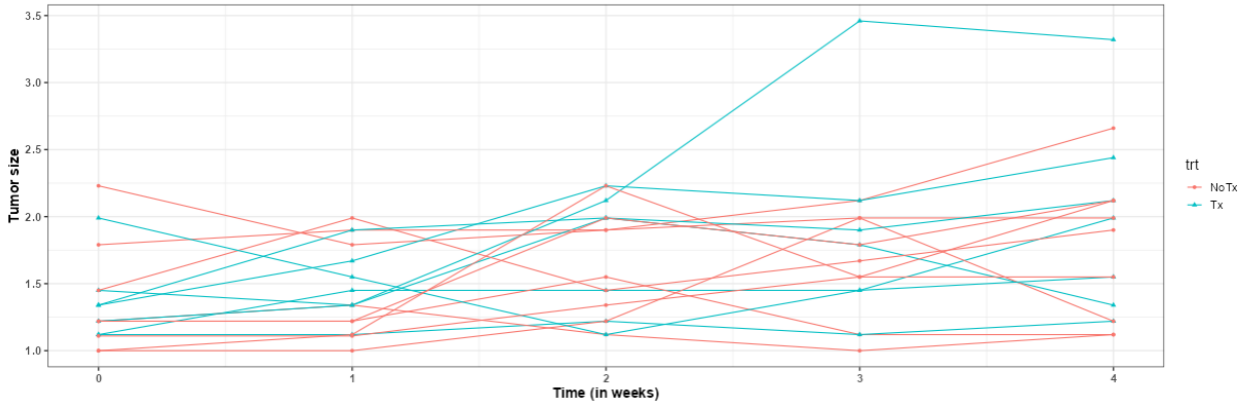


The residuals are randomly scattered with no clear pattern, suggesting that the model's assumptions are reasonably met. The spread of residuals is consistent, indicating homoscedasticity. There are no extreme outliers, although a few points deviate more than others. Overall, this residual plot suggests that the regression model provides a good fit for the data and that the key assumptions for a valid regression analysis are satisfied.

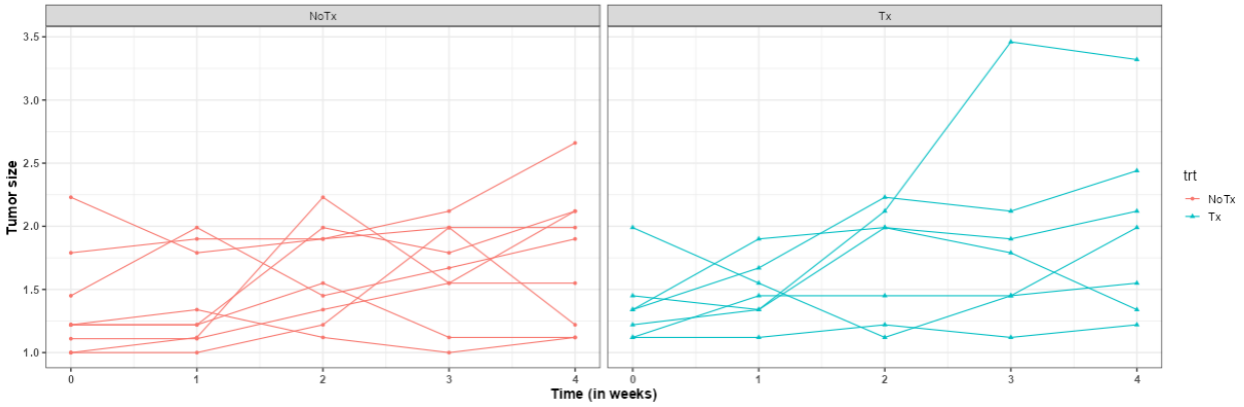
Spider plot and spider plot by group

Both plots show that there is not that much difference between two treatment conditions, implying that there is no treatment effect on tumor size. It appears that some mice in treatment group have a higher growth rate and a higher overall tumor size. These plots are shown below:

Spider plot

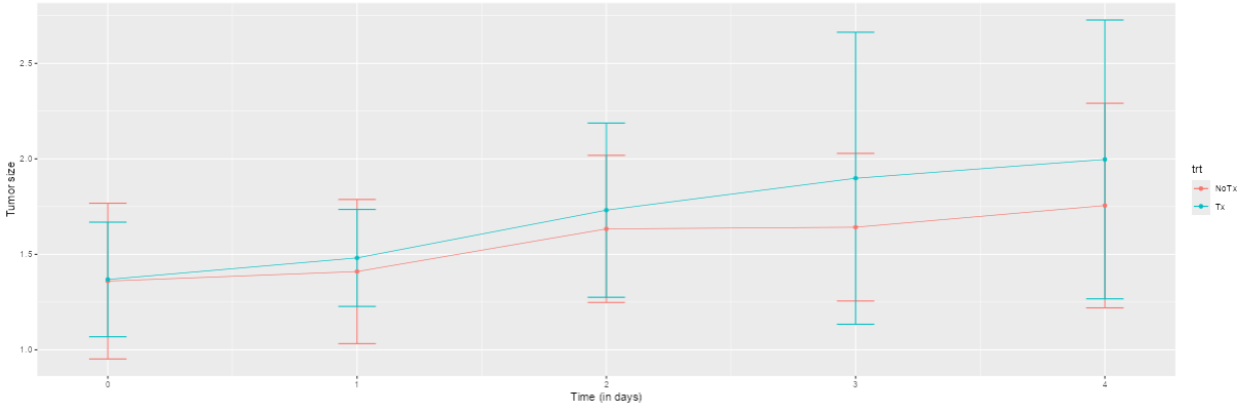


Spider plot by group



Mean growth plot

Mean growth plot



The mean growth plot shows the similar trend that is observed in the spider plots.

Linear mixed-effects results

Linear mixed-effects model

Summary

	numDF	denDF	F-value	p-value
(Intercept)	1	62	74.76608	<.0001
week	1	62	22.65184	<.0001
treatment	1	14	0.52630	0.4801
week:treatment	1	62	0.55415	0.4594

Based on this model, while tumor growth rate changes significantly over time, there is no evidence to suggest that the treatment has a significant effect on tumor growth rate compared to no treatment. Additionally, the interaction between week and treatment is not significant, indicating that the treatment does not significantly alter the tumor growth rate trajectory over time compared to the no treatment group.

Categorical Test

This section explains how to perform a categorical test. To perform this analysis, a user must use the **Categorical Test** tab (**Figure 32**). The **Categorical Test** tab performs a similar function to the **Contingency Table** tab, but the main difference lies in the input data. While the **Contingency Table** tab analyzes data directly entered by the user into the subpanel, the **Categorical Test** tab analyzes data uploaded through the **Upload Data** tab.

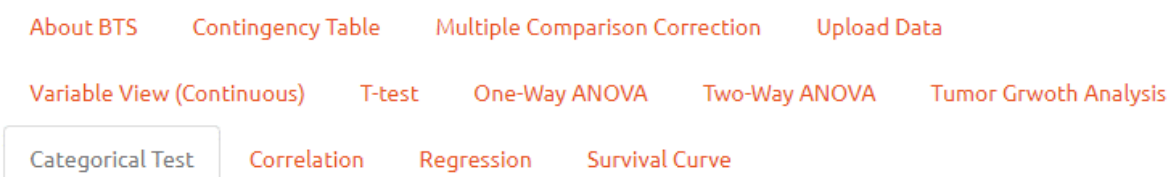


Figure 32. The **Categorical Test** tab.

Parameter Selection

A user needs to select a total of three parameters as shown in **Figure 33**. To perform a categorical test, the following three parameters must be decided by the user:

1. Select a type of test.
2. Select the first categorical variable (X1).
3. Select the second categorical variable (X2)

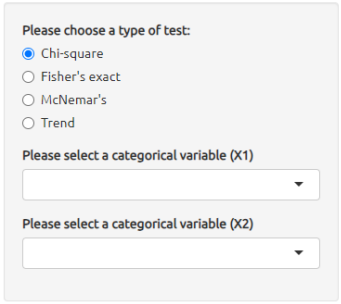


Figure 33. The parameter selection panel for the **Categorical Test** tab.

There are four options to perform a categorical test: Chi-square test (Chi-square), Fisher's exact test (Fisher's exact), McNemar's test (McNemar's), and Trend test (Trend). As described in the section for the **Contingency Table** tab, the Chi-square test and Fisher's exact test are used to investigate the association between two categorical variables. It is worthwhile to note that the Chi-square test can also be used for a goodness-of-fit test for one categorical variable, but this analysis is available only under the **Contingency Table** tab. When the sample size is small, especially if the expected frequencies in any of the cells of a contingency table are less than 5, the Chi-square test's approximation can be inaccurate. In such cases, Fisher's exact test provides an exact p-value. Generally, Fisher's exact test is preferred for small samples, while the Chi-square test is more efficient for larger samples and larger tables. The McNemar's test is used to assess paired data. The trend test is used to evaluate the trend between a categorical variable and an ordinal variable.

Output

The output panel (right panel) is composed of three or four subpanels:

- Levels
- Contingency table
- Expected contingency table (for a Chi-square test only)
- Hypothesis testing

The first subpanel '**Levels**' displays the levels of each categorical variable. The **Contingency table** subpanel shows the input data in the format of a contingency table. The **Expected contingency table** subpanel provides the expected cell counts, assisting in the decision-making process to switch to a Fisher's exact test if needed. This subpanel is available only for a Chi-square test. The last subpanel reports the results of the hypothesis testing.

Example

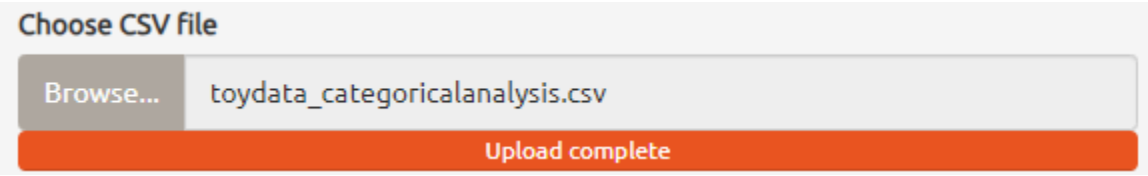
The example data is a toy example and a hypothetical data provided by the BTS. It contains the treatment response pre- and post-treatment for 250 patients by drugs (drug_A and drug_B) and is composed of the following four variables:

- **Treatment**: a factor with groups: drug_A and drug_B.
- **Drug_amount**: an ordinal factor with the drug level: Level_1 to Level_5.
- **Pre_Response**: a factor with the treatment response from a standard of care: Response and No_response.
- **Post_Response**: a factor with the treatment response from new treatments: Response and No_response.

The hypothesis to investigate is whether a tumor growth is associated with treatment.

Step by Step

To upload data, go to the **Upload Data** tab and choose the csv file "**toydata_categoricalanalysis.csv**":

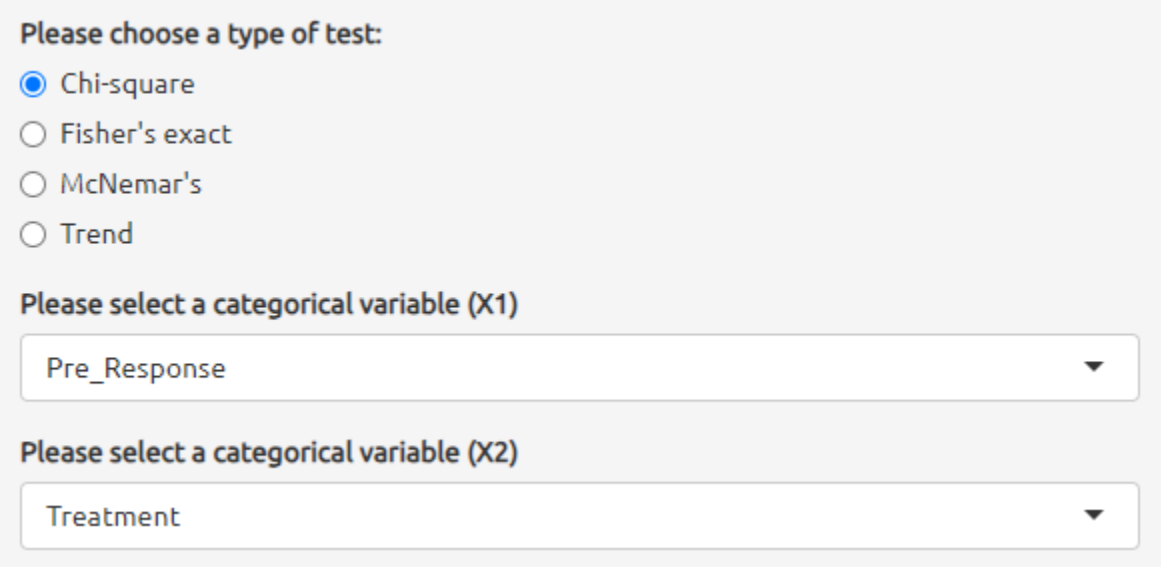


Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

Once the data is uploaded, go to the **Categorical Test** tab. The following outlines the step-by-step procedures for each categorical test separately.

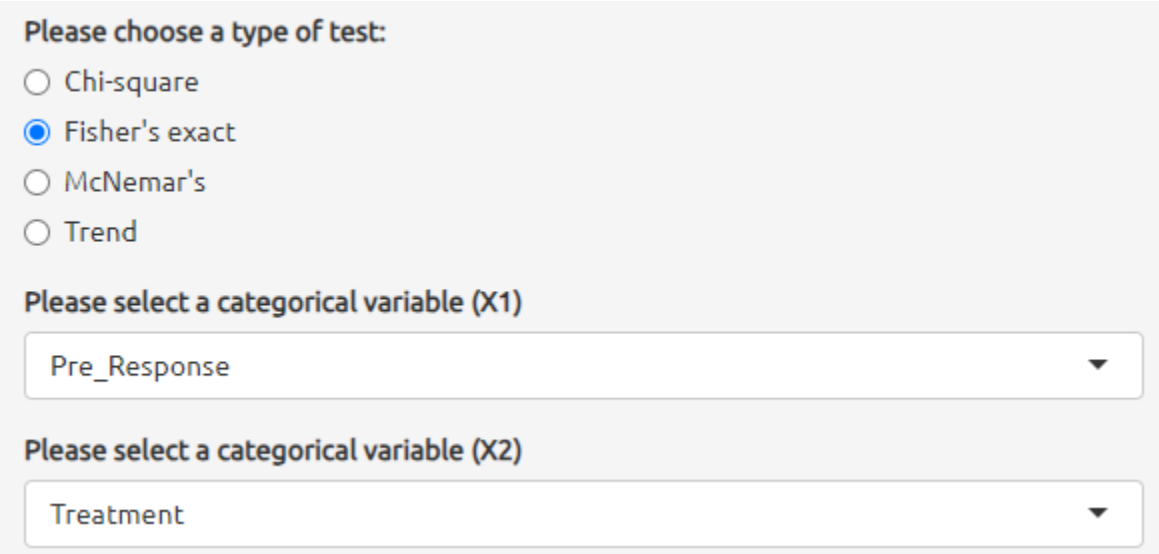
A. Chi-square test

Choose the '**Chi-square**' and select '**Pre_Response**' and '**Treatment**' for categorical variables (X1 and X2):



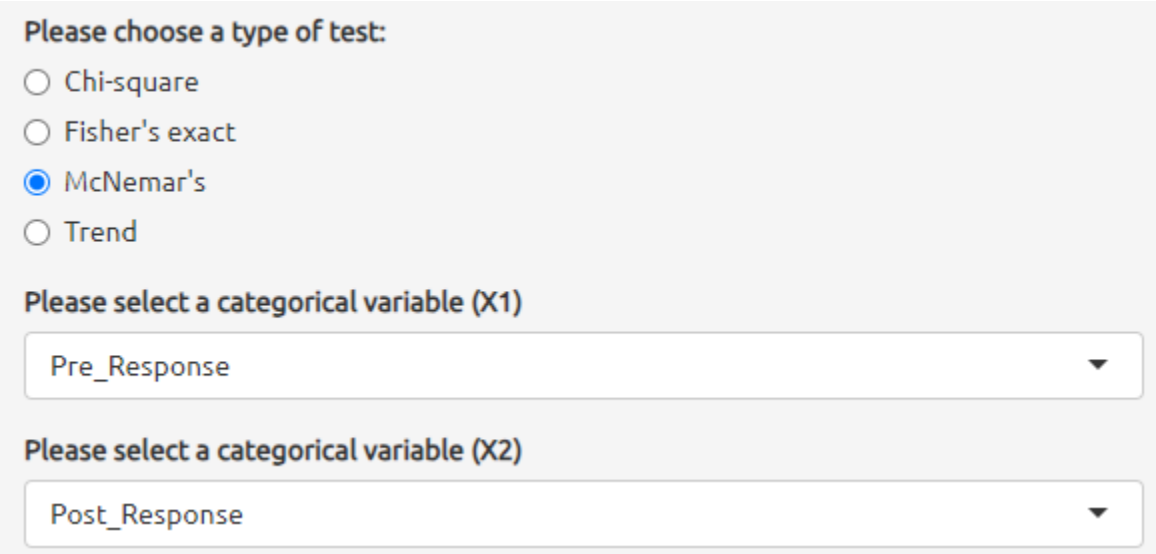
B. Fisher's exact test

Choose the '**Fisher's exact**' and select '**Pre_Response**' and '**Treatment**' for categorical variables (X1 and X2):



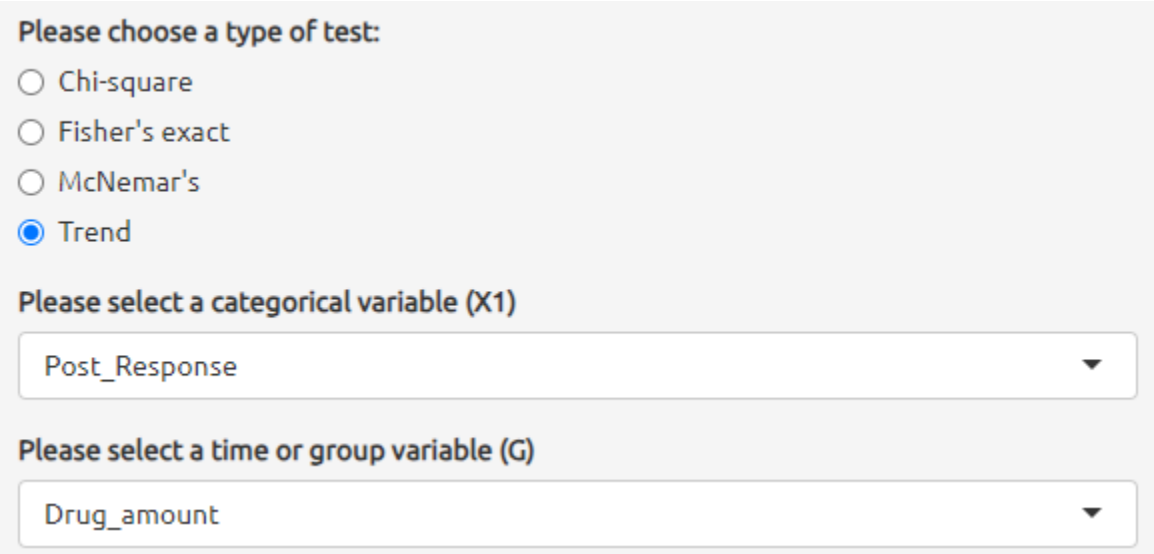
C. McNemar's test

Choose the '**McNemar's**' and select '**Pre_Response**' and '**Post_Response**' for categorical variables (X1 and X2):



D. Trend test

Choose the '**Trend**' and select '**Post_Response**' and '**Drug_amount**' for the categorical variables (X1 and X2):

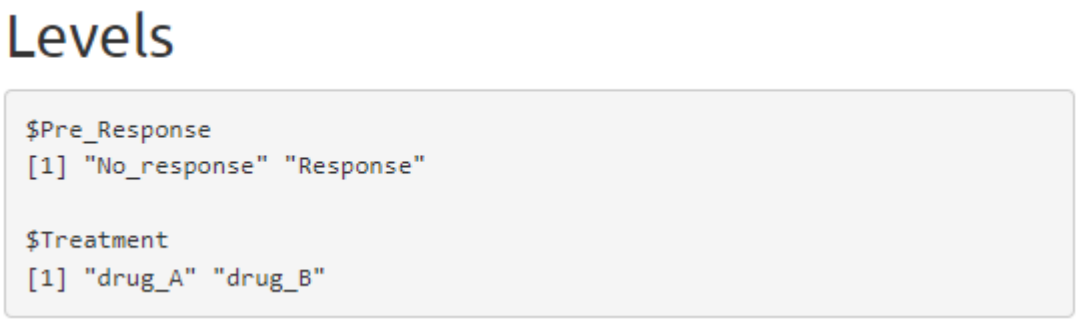


Results and Interpretation

A. Chi-square test

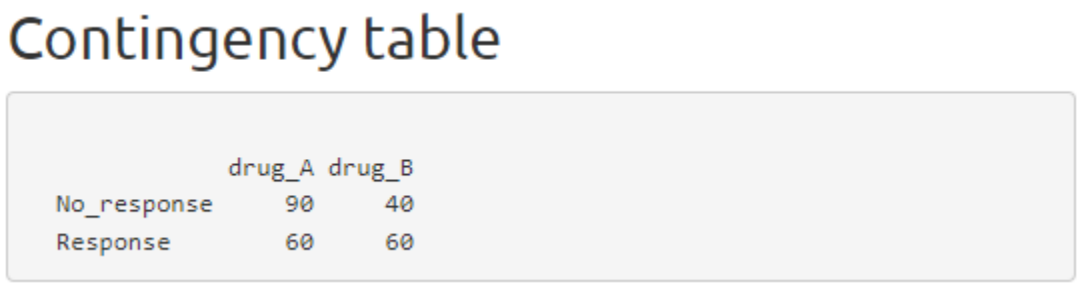
Levels

This subpanel provides information about the categorical variables of interest, such as the number of levels and the values of each level. It shows that '**Pre_Response**' has two levels ('**No_response**' and '**Response**') and '**Treatment**' has two levels ('**drug_A**' and '**drug_B**'):

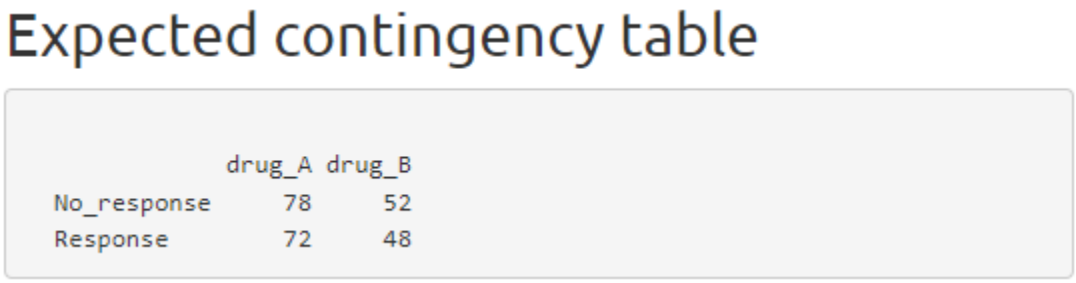


Contingency table

This subpanel provides the data in the form of contingency table. Since each variable has two levels, there is a 2 × 2 contingency table:



Expected contingency table



This subpanel shows the expected counts for each cell, which will be used to decide whether a Fisher's exact test should be used instead. In this example, all expected counts are greater than 5, a Chi-square test is still valid.

Hypothesis testing

This subpanel prints out the results from a hypothesis testing, which is a Chi-square test. The result shows that there is a significant association between '**Pre_Response**' and '**Treatment**' at the 5% level:

Hypothesis testing

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  ttab
X-squared = 8.8308, df = 1, p-value = 0.002962
```

B. Fisher's exact test

Levels

This subpanel provides information about the categorical variables of interest, such as the number of levels and the values of each level. It shows that '**Pre_Response**' has two levels ('**No_response**' and '**Response**') and '**Treatment**' has two levels ('**drug_A**' and '**drug_B**'):

Levels

```
$Pre_Response
[1] "No_response" "Response"

$Treatment
[1] "drug_A" "drug_B"
```

Contingency table

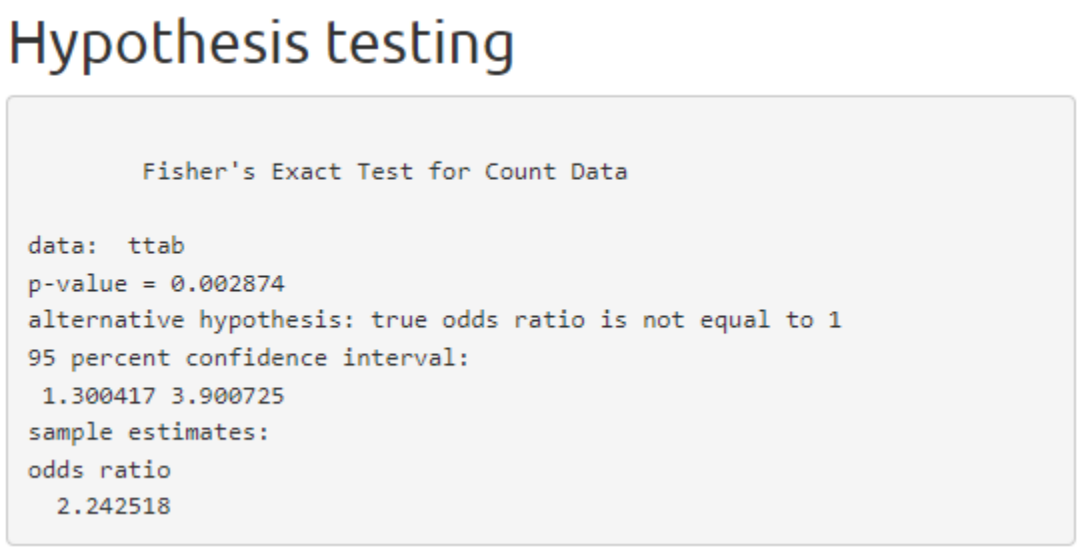
This subpanel provides the data in the form of contingency table. Since each variable has two levels, there is a 2 × 2 contingency table:

Contingency table

```
             drug_A drug_B
No_response      90     40
Response         60     60
```

Hypothesis testing

This subpanel prints out the results from a hypothesis testing, which is a Fisher's exact test. The result shows that there is a significant association between '**Pre_Response**' and '**Treatment**' at the 5% level:



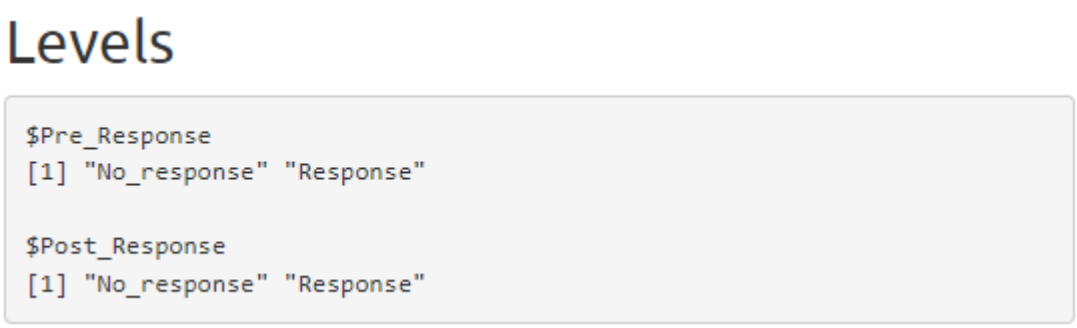
Hypothesis testing

```
        Fisher's Exact Test for Count Data

data:  ttab
p-value = 0.002874
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.300417 3.900725
sample estimates:
odds ratio
  2.242518
```

C. McNemar's test

Levels

This subpanel provides information about the categorical variables of interest, such as the number of levels and the values of each level. It shows that both '**Pre_Response**' and '**Post_Response**' have two levels ('**No_response**' and '**Response**'):



Levels

```
$Pre_Response
[1] "No_response" "Response"

$Post_Response
[1] "No_response" "Response"
```

Contingency table

This subpanel provides the data in the form of contingency table. Since each variable has two levels, there is a 2 × 2 contingency table:

Contingency table

```
             No_response Response
No_response          65       65
Response             65       55
```

Hypothesis testing

This subpanel prints out the results from a hypothesis testing, which is a McNemar's test. The result shows that there is no association between '**Pre_Response**' and '**Post_response**' at the 5% level:

Hypothesis testing

```
        McNemar's Chi-squared test

data:  ttab
McNemar's chi-squared = 0, df = 1, p-value = 1
```

D. Trend test

Levels

Levels

```
$Post_Response
[1] "No_response" "Response"

$Drug_amount
[1] "Level_1" "Level_2" "Level_3" "Level_4" "Level_5"
```

This subpanel provides information about the categorical variables of interest, such as the number of levels and the values of each level. It shows that '**Post_Response**' has two levels ('**No_response**' and '**Response**') and '**Drug_amount**' has five levels from '**Level_1**' to '**Level_5'**.

Contingency table

This subpanel provides the data in the form of contingency table. Since variables have two and five levels, there is a 2 × 5 contingency table:

Contingency table

	Level_1	Level_2	Level_3	Level_4	Level_5
No_response	26	26	26	26	26
Response	24	24	24	24	24

Hypothesis testing

This subpanel prints out the results from a hypothesis testing, which is a Trend test. The result shows that there is no linear trend between '**Post_Response**' and '**Drug_amount**' at the 5% level:

Hypothesis testing

```
        Chi-squared Test for Trend in Proportions

data:  ttab[1, ] out of apply(ttab, 2, sum) ,
 using scores: 1 2 3 4 5
X-squared = 6.3109e-30, df = 1, p-value = 1
```

Correlation

This section explains a correlation analysis. The correlation is a measure of association appropriate for numerical (continuous) variables. To perform this analysis, a user must use the **Correlation** tab (**Figure 34**).

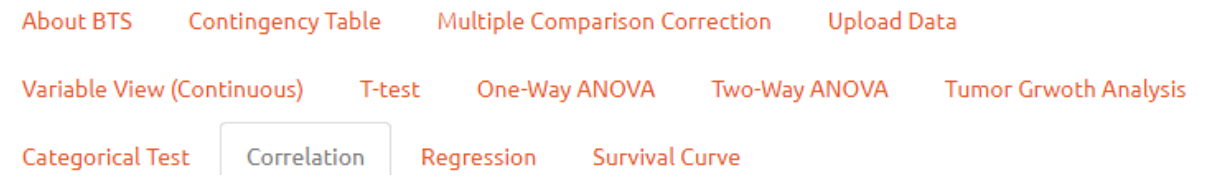


Figure 34. The **Correlation** tab.

Parameter Selection

A user needs to select a total of four parameters as shown in **Figure 35**:

1. Select numerical variables.
2. Select numerical variables for log-transformation.
3. Select numerical variables for square root-transformation.
4. Select a type of correlation.

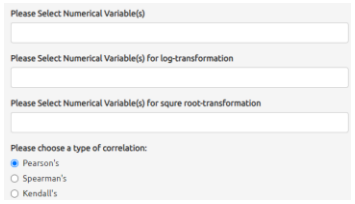


Figure 35. The parameter selection panel for the **Correlation** tab.

There are three options for the type of correlation: Pearson's correlation coefficient (Pearson's), Spearman's correlation coefficient (Spearman's), and Kendall's correlation coefficient (Kendall's) (**Figure 35**). Pearson's correlation coefficient measures the linear relationship between two continuous variables, with values ranging from -1 to 1, and is best used for normally distributed data with a linear relationship. Spearman's rank correlation coefficient, or Spearman's rho, assesses the monotonic relationship between two variables based on ranked data, making it suitable for ordinal data or non-normal distributions, with values also ranging from -1 to 1. Kendall's correlation coefficient, or Kendall's tau, measures ordinal association by evaluating the correspondence between the ranks of two variables, making it ideal for small sample sizes or data with many ties, with values ranging from -1 to 1. Each of these coefficients provides a way to understand the strength and direction of relationships in different types of data.

Output

The output panel (right panel) is composed of five subpanels:

- Q-Q plot
- Sharpiro-Wilk's normality test
- Correlation plot
- Transformation
- Summary

The first two subpanels describe and evaluate the distribution of data. The third subpanel provides the overall graphical summary of the correlation. The next subpanel shows the type of transformation used. The last subpanel displays results for correlation analysis.

Example

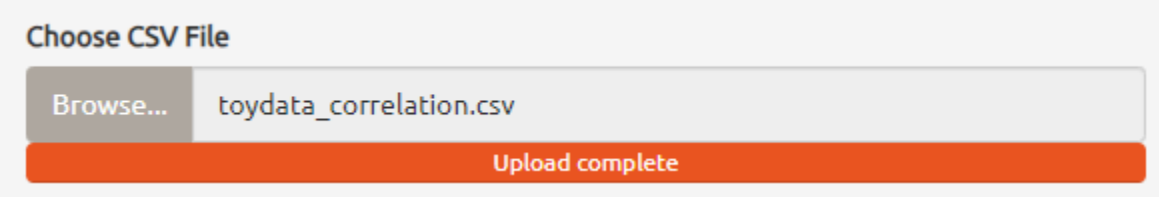
The example data is a toy example and a hypothetical data provided by the BTS. It contains the tumor size at each week for each mouse by group and is composed of the following four variables:

- **X**: a numeric vector, an index.
- **var1**: a numeric vector
- **var2**: a numeric vector

The hypothesis to investigate is a correlation between **var1** and **var2**.

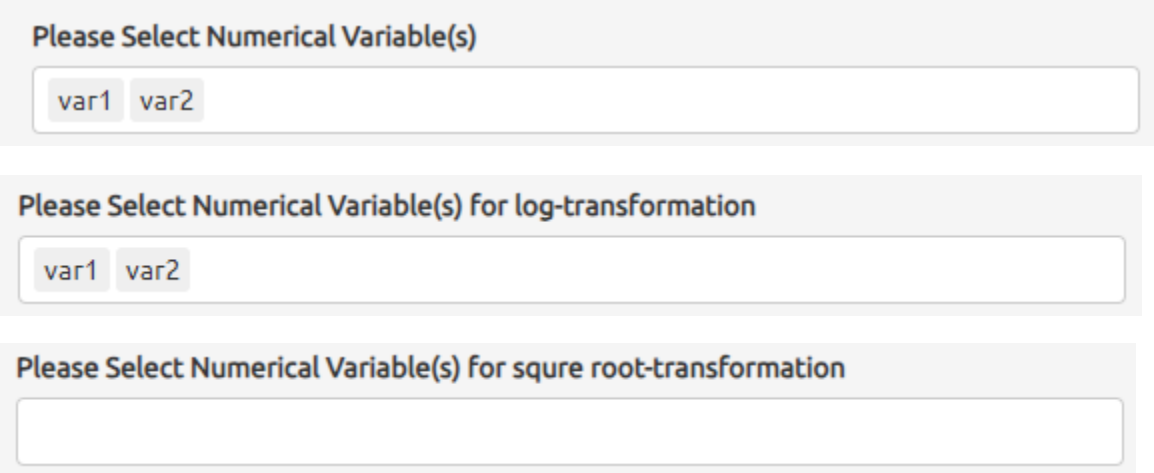
Step by Step

To upload data, go to the **Upload Data** tab and choose the CSV file "**toydata_correlation.csv**":

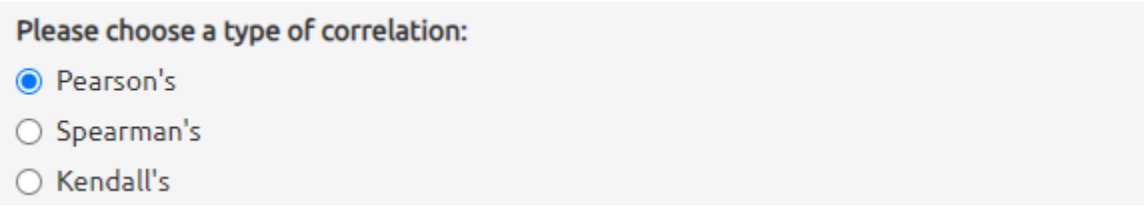


Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

Once the data is uploaded, go to the **Correlation** tab. Then select '**var1**' and '**var2**' as the numerical variables as well as those for log-transformation. Since the log-transformation is selected for both variables, no variable should be selected for square root transformation:



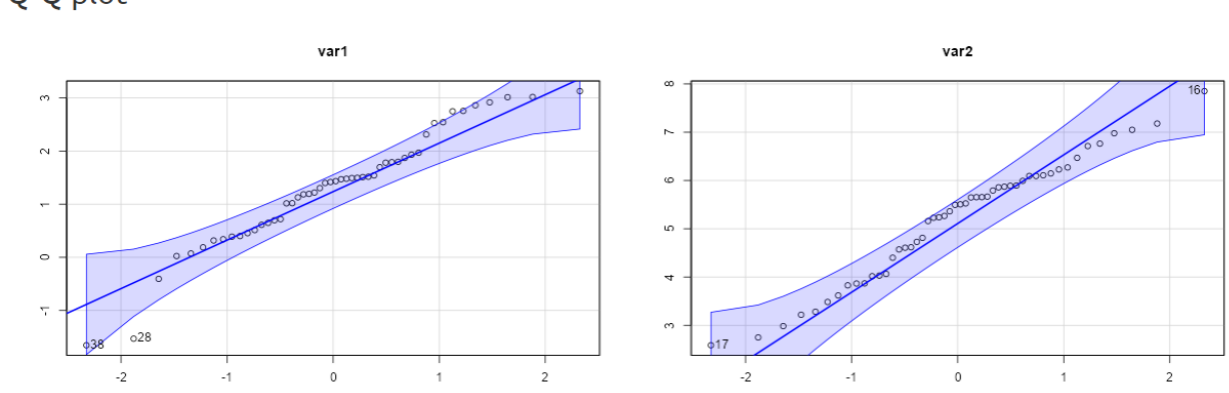
Finally, select the correlation coefficient. For example, Pearson's correlation coefficient can be selected as follows:



Results and Interpretation

Q-Q plot

All data points are inside the 95% confidence bands except for ID=28 for '**var1'**, implying that both variables appear to be fairly symmetric.



Sharpiro-Wilk's normality test

The p-value is large, failing to reject the null hypothesis that the data follows a normal distribution at a two-sided 5% level.

Sharpiro-Wilk's normality test

```
$var1

        Shapiro-Wilk normality test

data:  na.omit(vv)
W = 0.95852, p-value = 0.07703


$var2

        Shapiro-Wilk normality test

data:  na.omit(vv)
W = 0.97324, p-value = 0.3119
```

Transformation

Transformation

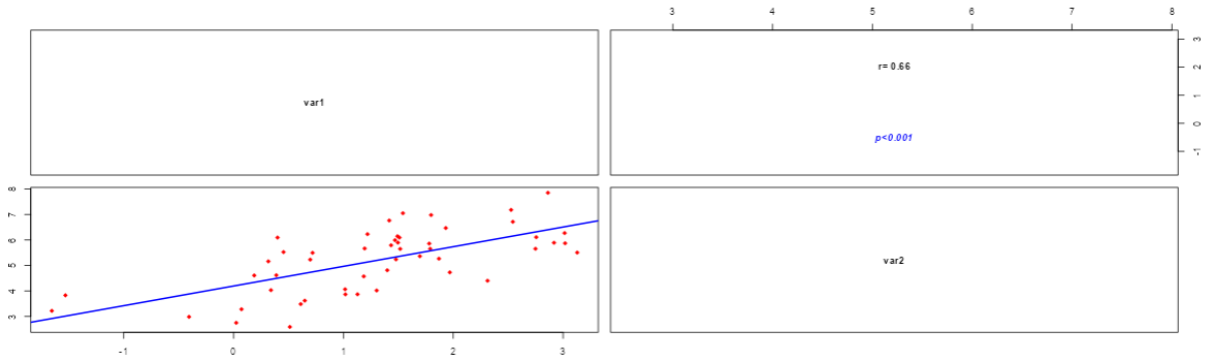
Variable	log-transformation	square root-transformation
var1	Yes	No
var2	Yes	No

It shows the information about the transformation used.

A. Pearson's correlation coefficient

Correlation plot

Correlation plot



The correlation plot shows a moderate to strong positive linear relationship between **var1** and **var2**, with a correlation coefficient of 0.66. The relationship is statistically significant, as indicated by the p-value of less than 0.001. The scatter plot supports this finding by displaying data points that generally follow an upward trend along the fitted regression line.

Summary

The analysis shows a statistically significant moderate to strong positive linear relationship between **var1** and **var2**, with a Pearson's correlation coefficient of 0.6611 and a p-value of 1.73291e-07. This indicates that the variables are positively correlated and that this correlation is unlikely to have occurred by chance:

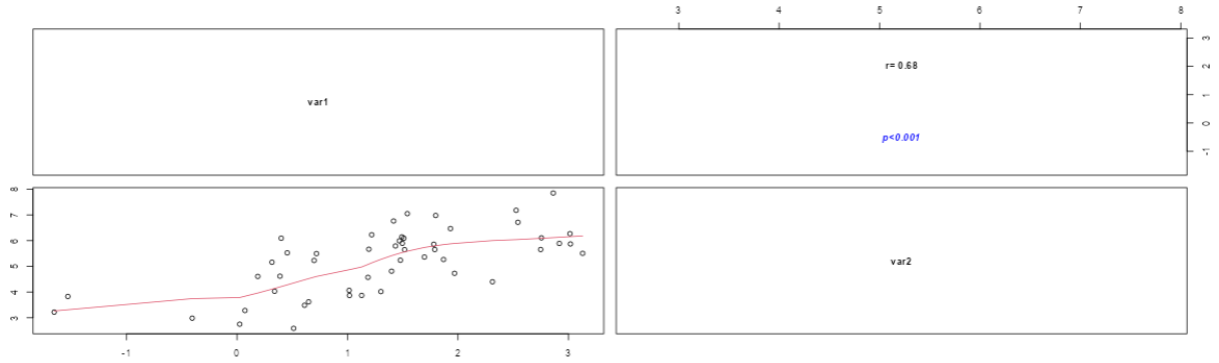
Summary

```
$method
[1] "Pearson's correlation"

$coefficient
          var1      var2
var1 0.0000000 0.6611263
var2 0.6611263 0.0000000

$p.value
             var1        var2
var1 0.00000e+00 1.73291e-07
var2 1.73291e-07 0.00000e+00
```

B. Spearman's correlation coefficient

Correlation plot

Correlation plot



The Spearman's correlation plot shows a strong positive monotonic relationship between **var1** and **var2**, with a correlation coefficient of 0.68 and a statistically significant p-value of less than 0.001. The scatter plot, along with the loess fitted curve, further illustrates this positive relationship, highlighting both linear and non-linear trends in the data. This suggests that as **var1** increases, **var2** tends to increase as well, and this relationship is unlikely to be due to random chance.

Summary

The analysis shows a statistically significant strong positive monotonic relationship between **var1** and **var2**, with a Spearman's correlation coefficient of 0.6753 and a p-value of 2.181942e-07. This indicates that the variables are positively correlated, and this correlation is unlikely to have occurred by chance:

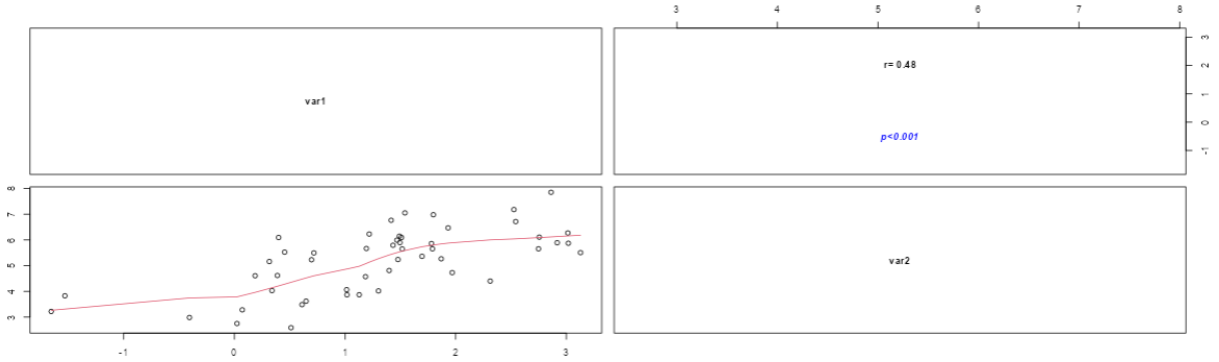
Summary

```
$method
[1] "Spearman's correlation"

$coefficient
          var1      var2
var1 0.0000000 0.6752941
var2 0.6752941 0.0000000

$p.value
             var1         var2
var1 0.000000e+00 2.181942e-07
var2 2.181942e-07 0.000000e+00
```

C. Kendall's correlation coefficient

Correlation plot

Correlation plot



The Kendall's correlation plot shows a moderate positive ordinal association between **var1** and **var2**, with a Kendall's tau correlation coefficient of 0.48 and a statistically significant p-value of less than 0.001. The scatter plot, along with the loess fitted curve, further illustrates this positive relationship, highlighting both linear and non-linear trends in the data. This suggests that as **var1** increases, **var2** tends to increase as well, and this relationship is unlikely to be due to random chance.

Summary

Summary

```
$method
[1] "Kendall's correlation"

$coefficient
          var1      var2
var1 0.0000000 0.4840816
var2 0.4840816 0.0000000

$p.value
             var1         var2
var1 0.000000e+00 7.036242e-07
var2 7.036242e-07 0.000000e+00
```

The analysis shows a statistically significant moderate positive ordinal association between **var1** and **var2**, with a Kendall's tau correlation coefficient of 0.4848 and a p-value

of 7.036242e-07. This indicates that the variables are positively correlated, and this correlation is unlikely to have occurred by chance.

Regression

This section explains linear, logistic, and Cox regression analyses. To perform these analyses, a user must use the **Regression** tab (**Figure 36**).

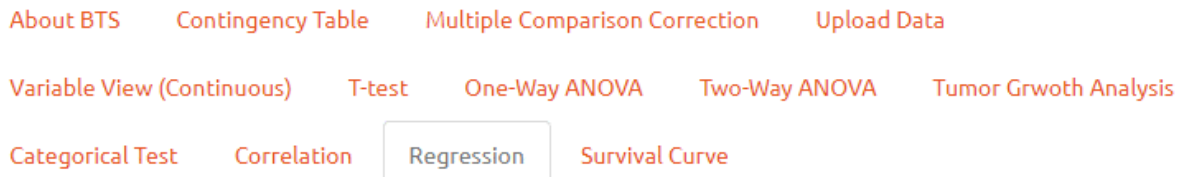


Figure 36. The **Regression** tab.

Linear regression models the relationship between a continuous dependent variable and one or more independent variables, providing a best-fitting linear equation for prediction. Logistic regression, on the other hand, predicts the probability of a binary outcome using continuous or categorical predictors, outputting probabilities and odds. Cox regression analyzes time-to-event data to assess the impact of predictors on the hazard rate, commonly used in survival analysis. Each regression technique is tailored to specific types of data and research questions, making them valuable tools in statistical analysis.

Parameter Selection

A user needs to select a type of regression model and, according to the chosen regression model, the number of parameters that a user needs to select is different.

When a user selects the linear regression model (**Linear**), there are five parameters to select (**Figure XXA**):

1. Select a response variable
2. Select a type of transformation for a response variable
3. Select covariate(s)
4. Select categorical covariate(s)
5. Select continuous covariate(s) for scaling

When a user selects the logistic regression model (**Logistic**), there are four parameters to select (**Figure XXB**):

1. Select a response variable
2. Select covariate(s)
3. Select categorical covariate(s)
4. Select continuous covariate(s) for scaling

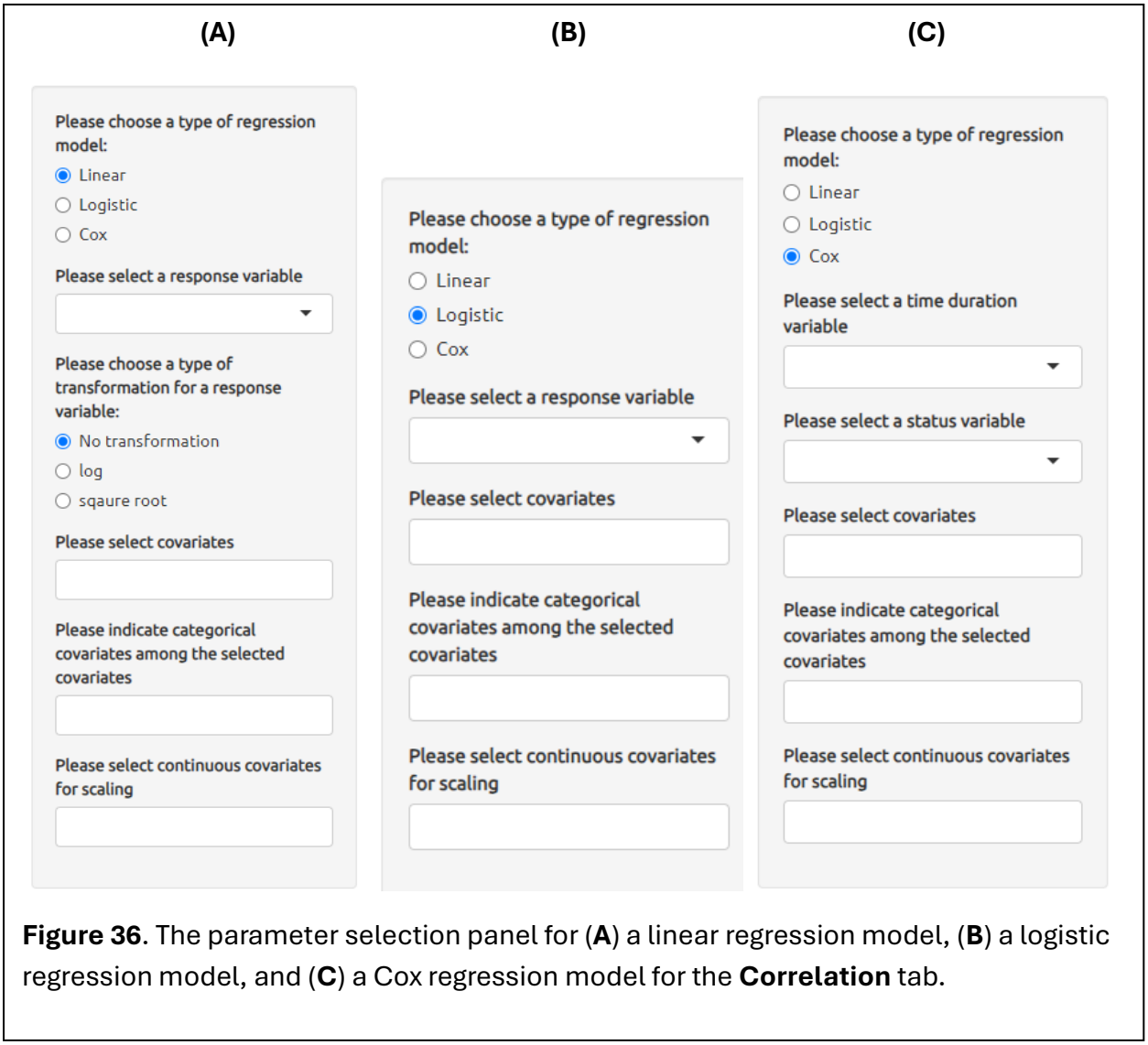


Figure 36. The parameter selection panel for (**A**) a linear regression model, (**B**) a logistic regression model, and (**C**) a Cox regression model for the **Correlation** tab.

When a user selects the Cox regression model (**Cox**), there are five parameters to select (**Figure 36C**):

1. Select a time duration variable
2. Select a status variable
3. Select covariate(s)
4. Select categorical covariate(s)
5. Select continuous covariate(s) for scaling

Across all regression models, a user needs to select at least one covariate. If there is only one covariate, it is called a univariable analysis, while it is called a multivariable analysis if there are two or more covariates. It is required to indicate which covariates are categorical, but it is optional to indicate which continuous covariates need to be scaled. Whether

continuous covariates are scaled or not generally does not influence the statistical test results. The transformation of a response variable is required only for a linear regression model. For a Cox regression model, two variables are required for the response: the time duration and the status of the event of interest.

Output

A. Linear regression model

The output panel (right panel) is composed of four subpanels:

- Q-Q plot
- Sharpiro-Wilk's normality test
- Diagnostic plots
- Summary

The first two subpanels describe and evaluate the distributions of the residuals and the response variable. The next subpanel provides the plots to diagnose the linear regression model. The last subpanel shows the results for a linear regression analysis.

B. Logistic regression model

The output panel (right panel) is composed of two subpanels:

- Diagnostic plots
- Summary

The first subpanel provides the plots to diagnose the logistic regression model. The last subpanel shows the results for a logistic regression analysis.

C. Cox regression model

The output panel (right panel) is composed of two subpanels:

- Diagnostic plots
- Summary

The first subpanel provides the plots to diagnose the Cox regression model. The last subpanel shows the results for a Cox regression analysis and for a proportional hazard assumption.

Example

A. Linear regression model

The example data, called **trees**, is available in the R package **datasets**. It provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is erroneously labelled Girth in the data. It is measured at 4 ft 6 in above the ground. The example data is composed of the following three variables:

- **Girth**: a numeric vector, Tree diameter (rather than girth, actually) in inches
- **Height**: a numeric vector, Height in ft
- **Volume**: a numeric vector, Volume of timber in cubic ft

The hypothesis to investigate is to see if **Girth** and **Height** can predict **Volume**.

B. Logistic regression model

The example data, called **graft.vs.host**, is available in the R package **ISwR**. It has 37 rows and 7 columns. It contains data from patients receiving a nondepleted allogenic bone marrow transplant with the purpose of finding variables associated with the development of acute graft-versus-host disease. The example data is composed of nine variables, but the example focuses on the following four variables:

- **gvhd**: a numeric vector code, graft-versus-host disease, 0: no, 1: yes
- **type**: a numeric vector, type of leukaemia coded 1: AML, 2: ALL, 3: CML for acute myeloid, acute lymphatic, and chronic myeloid leukaemia.
- **preg**: a numeric vector code indicating whether donor has been pregnant. 0: no, 1: yes.
- **index**: a numeric vector giving an index of mixed epidermal cell-lymphocyte reactions.

The hypothesis to investigate is to see if **type**, **preg** and **index** can predict **gvhd**.

C. Cox regression model

The example data, called **lung**, is available in the R package **survival**. It contains survival information in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities. The example data is composed of the 10 variables, but the example focuses on the following six variables:

- **time**: a numeric vector, survival time in days
- **status**: a numeric vector, censoring status 1=censored, 2=dead
- **age**: a numeric vector, age in years

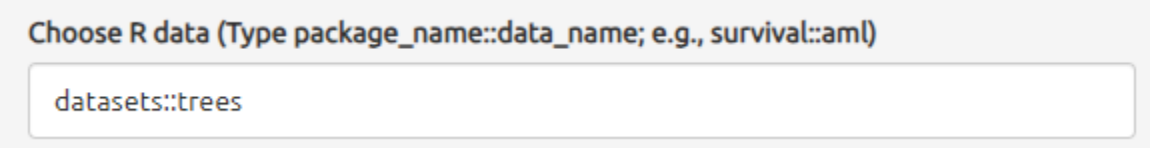
- **sex**: a numeric vector, male=1, female=2
- **ph.ecog**: a numeric vector, ECOG performance score as rated by the physician. 0=asymptomatic, 1= symptomatic but completely ambulatory, 2= in bed <50% of the day, 3= in bed > 50% of the day but not bedbound, 4 = bedbound
- **ph.karno**: a numeric vector, Karnofsky performance score (bad=0-good=100) rated by physician

The hypothesis to investigate is to see if **age**, **sex, ph.ecog** and **ph.karno** can predict survival **time**.

Step by Step

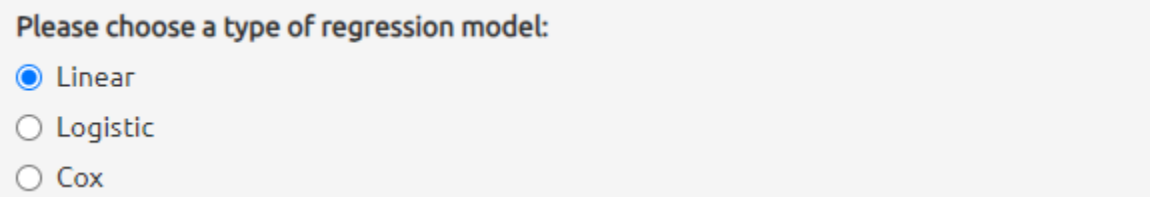
A. Linear regression model

To upload data, go to the **Upload Data** tab and type "**datasets::trees**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:

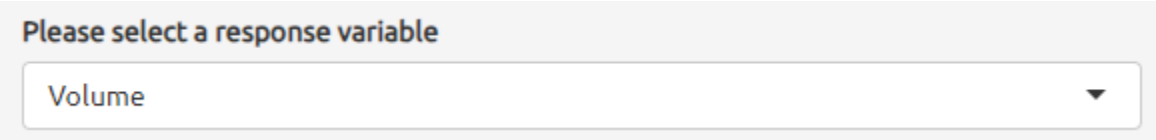


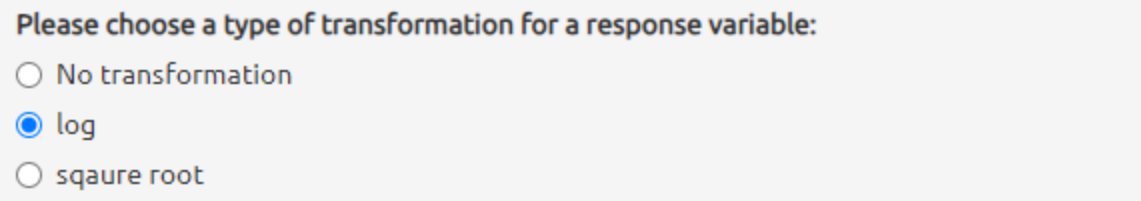
Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

Once the data is uploaded, go to the **Regression** tab. Then select '**Linear**' as the type of regression model:

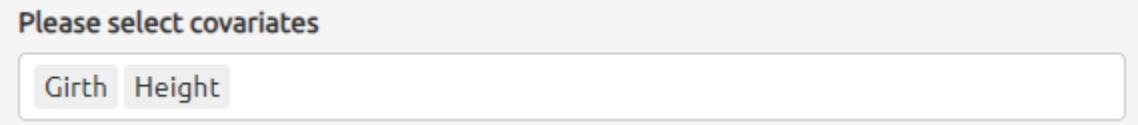


Select the response variable, **Volume**, followed by choosing '**log**' as the type of transformation.





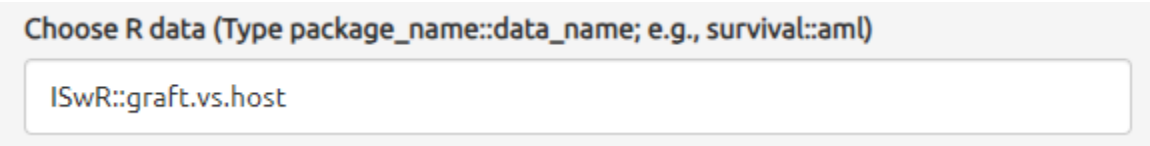
Lastly, select **Girth** and **Height** as covariates:



Since both **Girth** and **Height** are continuous, no action is needed for the selection of categorical variables. In addition, these variables will not be scaled.

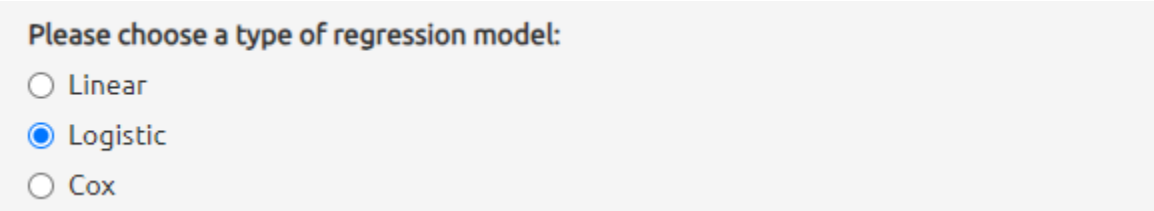
B. Logistic regression model

To upload data, go to the **Upload Data** tab and type "**ISwR::graft.vs.host**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:

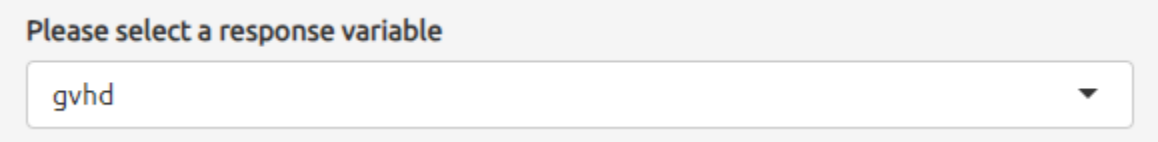


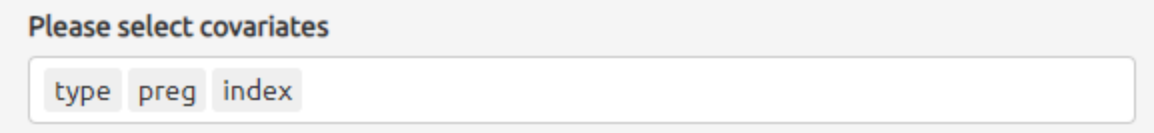
Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data. In case the **graft.vs.host** data is not in the **R Data list**, please ensure that the R package *ISwR* has been installed.

Once the data is uploaded, go to the **Regression** tab. Then select '**Logistic**' as the type of regression model:

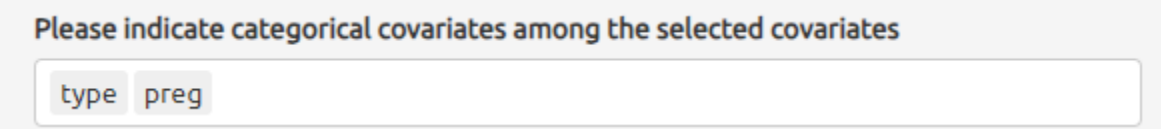


Select the response variable, **gvhd**, followed by choosing '**type**', '**preg**', and '**index**' as the covariates.





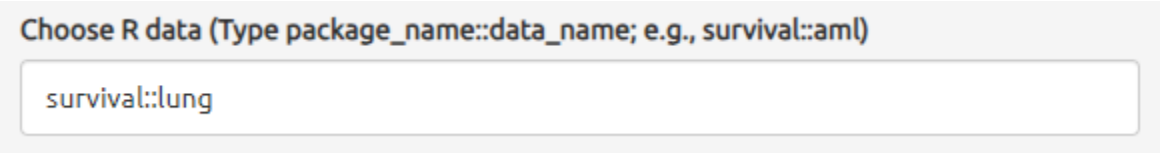
Lastly, select **type** and **preg** as categorical covariates:



Note that the continuous covariate, **index**, will not be scaled.

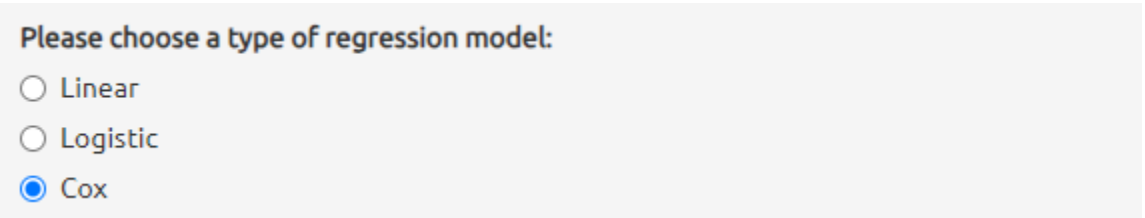
C. Cox regression model

To upload data, go to the **Upload Data** tab and type "**survival::lung**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:

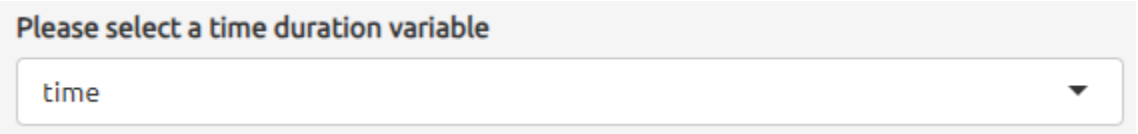


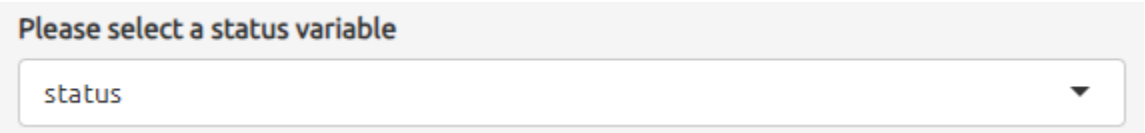
Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

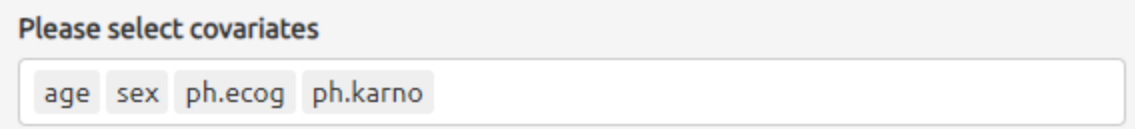
Once the data is uploaded, go to the **Regression** tab. Then select '**Cox**' as the type of regression model:



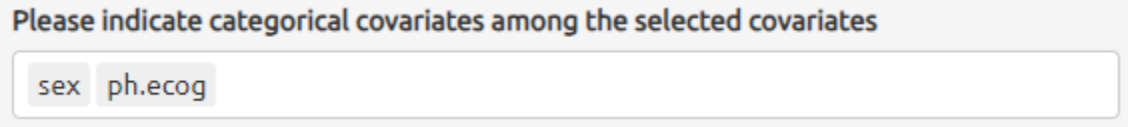
Select the time duration variable, **time**, and the status variable, **status**, followed by choosing '**age**', '**sex**', '**ph.ecog**' and '**ph.karno**' as the covariates.







Lastly, select **sex** and **ph.ecog** as categorical covariates:

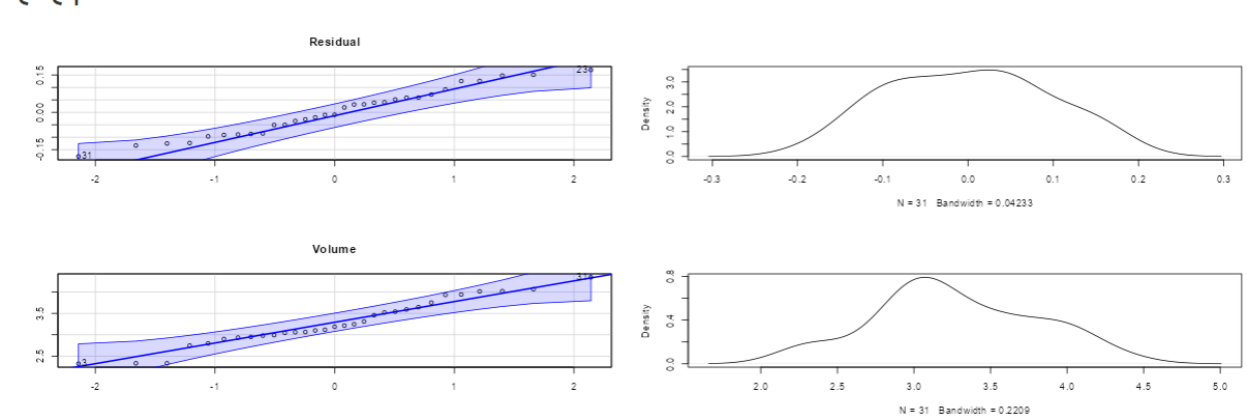


Note that the continuous covariates, **age** and **ph.karno**, will not be scaled.

Results and Interpretation

A. Linear regression model

Q-Q plot



No data point is outside the 95% confidence bands for both residuals and **Volume**, and the density plots appear to be fairly symmetric. It is noted that the normality assumption for a linear regression model is related mainly to residuals.

Sharpiro-Wilk's normality test

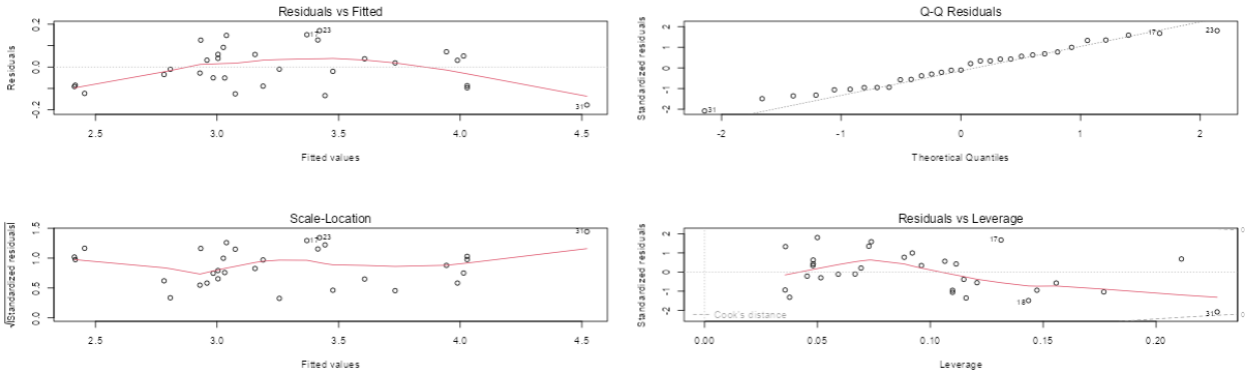
The p-value is large for residuals and the response variable, failing to reject the null hypothesis that the residual distribution follows a normal distribution at a two-sided 5% level. As pointed out in the Q-Q plot, the checking normality is mainly for residuals.

Sharpiro-Wilk's normality test

```
$residual

        Shapiro-Wilk normality test

data:  na.omit(vv)
W = 0.97144, p-value = 0.5594


$response

        Shapiro-Wilk normality test

data:  na.omit(vv)
W = 0.96427, p-value = 0.3766
```

Diagnostic plots

Diagnostic plots



The diagnostic plots for the linear regression model reveal several insights. The Residuals vs Fitted plot shows a slight curve, suggesting a mild non-linearity in the model, but no severe patterns. The Q-Q plot indicates that the residuals are approximately normally distributed, as most points closely follow the reference line. The Scale-Location plot shows that the spread of the standardized residuals is fairly constant across fitted values, implying homoscedasticity. The Residuals vs Leverage plot identifies a few influential points (e.g., #17, #31) with moderate leverage, but no extreme outliers. Overall, the model assumptions appear reasonably met, with minor non-linearity and a few influential observations to consider.

Summary

The linear regression model summary shows that the model has a high goodness-of-fit, with an R-squared value of 0.9684 and an adjusted R-squared value of 0.9662, indicating that approximately 96.62% of the variance in the dependent variable is explained by the model. The coefficients for both **Girth** and **Height** are statistically significant ($p < 0.001$), with estimates of 0.145290 for **Girth** and 0.016358 for **Height**, indicating positive associations with the outcome variable. The intercept is not statistically significant ($p = 0.637$), suggesting it does not significantly contribute to the model. The residuals appear to

be fairly normally distributed around zero, with minimal variation, as indicated by the residual summary. The F-statistic (429.7) and its corresponding p-value (< 2.2e-16) confirm that the overall model is statistically significant.

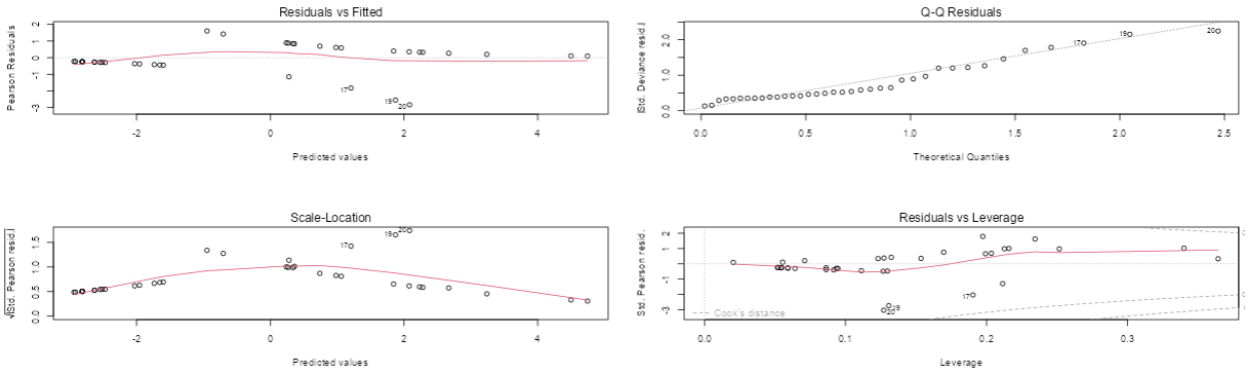
Summary

```
$Output

Call:
lm(formula = as.formula(tform), data = new.var1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.177279 -0.086019 -0.009928  0.058914  0.170011

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.102585   0.215315   0.476    0.637
Girth       0.145290   0.006587  22.057  < 2e-16 ***
Height      0.016385   0.003244   5.051 2.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09676 on 28 degrees of freedom
Multiple R-squared:  0.9684,    Adjusted R-squared:  0.9662
F-statistic: 429.7 on 2 and 28 DF,  p-value: < 2.2e-16


$Summary
            Estimate 95% CI (lower) 95% CI (upper)            p
(Intercept) 1.108031      0.7128611       1.722262 6.374580e-01
Girth       1.156374      1.1408762       1.172083 3.065714e-19
Height      1.016520      1.0097877       1.023298 2.413686e-05
```

B. Logistic regression model

Diagnostic plots

Diagnostic plots



The diagnostic plots for the logistic regression model indicate several points of interest. The Residuals vs Fitted plot shows a slight curvature, suggesting some non-linearity, but no severe deviations that would indicate major issues. The Q-Q plot shows that the residuals deviate from the theoretical quantiles, especially at the high end, indicating some departure from the expected distribution. The Scale-Location plot reveals a fairly consistent spread of the standardized residuals, suggesting homoscedasticity. The Residuals vs Leverage plot identifies a few influential points (e.g., #17, #19, #20) with

moderate leverage but no extreme outliers. Overall, while the model assumptions are generally met, there are some indications of non-linearity and influential observations that warrant further examination.

Summary

Summary

```
$Levels
[1] "0" "1"

$Output

Call:
glm(formula = as.formula(tform), family = "binomial", data = new.var1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.03260    1.35382  -2.240   0.0251 *
type2       -0.05601    1.19532  -0.047   0.9626
type3        2.45397    1.30251   1.884   0.0596 .
preg1        2.77319    1.11063   2.497   0.0125 *
index        0.56391    0.31684   1.780   0.0751 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 28.413  on 32  degrees of freedom
AIC: 38.413

Number of Fisher Scoring iterations: 5


$Summary
                   OR 95% CI (lower) 95% CI (upper)          p
(Intercept)  0.04819023     0.00198542      0.4756453 0.02508941
type2        0.94552814     0.08450557     11.5293525 0.96262561
type3       11.63443686     1.08775129    207.7225468 0.05956061
preg1       16.00967039     2.16726359    191.0058221 0.01252627
index        1.75752417     1.07924424      3.9475110 0.07511440
```

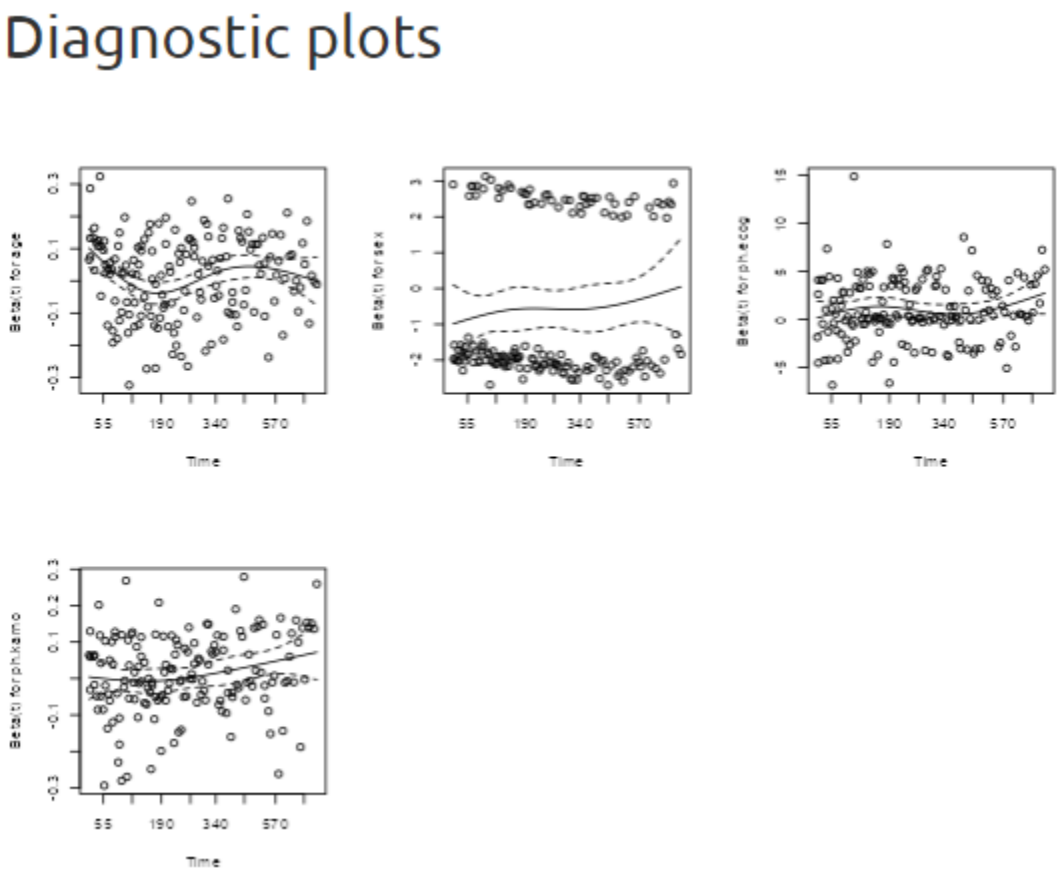
The logistic regression analysis results indicate that the intercept and the predictor variable preg1 are statistically significant at the 0.05 level, with p-values of 0.0251 and 0.0125, respectively. The coefficient for preg1 (2.77319) suggests a strong positive association with the outcome variable, with an odds ratio (OR) of approximately 16.0, implying that preg1 significantly increases the odds of the outcome. The other predictors (type2, type3, index) are not statistically significant, with p-values greater than 0.05. The model's residual deviance (28.413) and AIC (38.413) suggest a good fit, given the degrees of freedom (32). However, further investigation and possibly additional data or predictors may be needed to improve the model's explanatory power.

C. Cox regression model

Diagnostic plots

The Schoenfeld residual diagnostic plots are used to assess the proportional hazards assumption in a Cox regression model. In these plots, each point represents a residual at a given time, and the horizontal line at zero indicates where the residuals should ideally lie if the proportional hazards assumption holds true. The first three plots (from left to right)

show no clear patterns or trends, suggesting that the proportional hazards assumption is reasonably met for these covariates. However, the fourth plot exhibits a noticeable pattern over time, indicating a potential violation of the proportional hazards assumption for this covariate. This suggests that the effect of this covariate on the hazard may not be constant over time and may require further investigation or adjustment, such as incorporating time-dependent covariates or using stratified models.



Summary

The Cox proportional hazards model fitting results indicate significant associations between several covariates and the hazard function. Specifically, sex2, ph.ecog1, ph.ecog2, and ph.ecog3 show statistically significant effects on survival with p-values less than 0.05. The model's concordance index (0.633) suggests moderate predictive accuracy. The global tests, including the Likelihood Ratio, Wald, and Score (Logrank) tests, all yield highly significant results ($p < 0.0001$), confirming the overall significance of the model.

Regarding the proportional hazards assumption, the global Schoenfeld test yields a p-value of 0.086, indicating that the assumption is reasonably satisfied overall. However, the individual test for ph.karno shows a significant p-value (0.023), suggesting a potential violation for this specific covariate. Despite this, the non-significant global test (p = 0.086) suggests that the proportional hazards assumption holds reasonably well for the model as a whole.

Summary

```
$Output
Call:
coxph(formula = as.formula(tform), data = new.var1)

  n= 226, number of events= 163
   (2 observations deleted due to missingness)

             coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.012560  1.012639  0.009452  1.329 0.183903
sex2     -0.565675  0.567977  0.169750 -3.332 0.000861 ***
ph.ecog1  0.578058  1.782574  0.236399  2.445 0.014474 *
ph.ecog2  1.239895  3.455252  0.355303  3.490 0.000484 ***
ph.ecog3  2.395853 10.977560  1.081217  2.216 0.026699 *
ph.karno  0.012423  1.012500  0.009588  1.296 0.195118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
age          1.013    0.98752    0.9941    1.0316
sex2         0.568    1.76064    0.4072    0.7922
ph.ecog1     1.783    0.56099    1.1216    2.8332
ph.ecog2     3.455    0.28941    1.7220    6.9329
ph.ecog3    10.978    0.09109    1.3188   91.3775
ph.karno     1.013    0.98765    0.9936    1.0317

Concordance= 0.633  (se = 0.025 )
Likelihood ratio test= 31.58  on 6 df,   p=2e-05
Wald test            = 32.04  on 6 df,   p=2e-05
Score (logrank) test = 34.14  on 6 df,   p=6e-06


$Summary
                HR 95% CI (lower) 95% CI (upper)            p
age       1.0126391      0.9940524      1.0315733 0.1839029349
sex2      0.5679765      0.4072293      0.7921761 0.0008609901
ph.ecog1  1.7825739      1.1215652      2.8331564 0.0144744676
ph.ecog2  3.4552518      1.7220481      6.9328870 0.0004835904
ph.ecog3 10.9775597      1.3187800     91.3774984 0.0266993193
ph.karno  1.0125001      0.9936499      1.0317079 0.1951182380

$PHassumption
           chisq df     p
age       0.0551  1 0.814
sex       1.8119  1 0.178
ph.ecog   5.6091  3 0.132
ph.karno  5.1638  1 0.023
GLOBAL   11.0813  6 0.086
```

Survival Curve

This section explains how to perform a survival analysis using the Kaplan-Meier (KM) curve and estimate including a univariable Cox regression model. To perform this analysis, a user must use the **Survival Curve** tab (**Figure 37**).

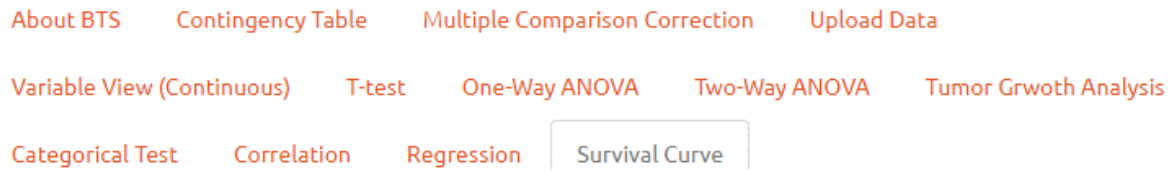


Figure 37. The **Survival Curve** tab.

The **Survival Curve** tab provides KM curves and estimates such as median time, survival rate, and 95% confidence intervals. Additionally, it offers KM curves by group and results from a univariable Cox regression, including a Schoenfeld residuals test.

Parameter Selection

A user needs to select up to a total of 12 parameters as shown in **Figure 38**. To perform this analysis, the following parameters must be decided by the user:

1. Select a survival time variable
2. Select a survival status variable.
3. Select an input unit
4. Select an output unit
5. Type an x-axis label. The default is 'Days After Diagnosis'
6. Type a y-axis label. The default is 'Overall Survival'
7. Select the time point for a survival rate at a specific time point
8. Select whether a group comparison will be performed
9. If 'Group comparison?' is 'Yes', select the group variable
10. If 'Group comparison?' is 'Yes', decide whether a p-value will be in a plot
11. If 'Group comparison?' is 'Yes', select whether the group order needs to changed

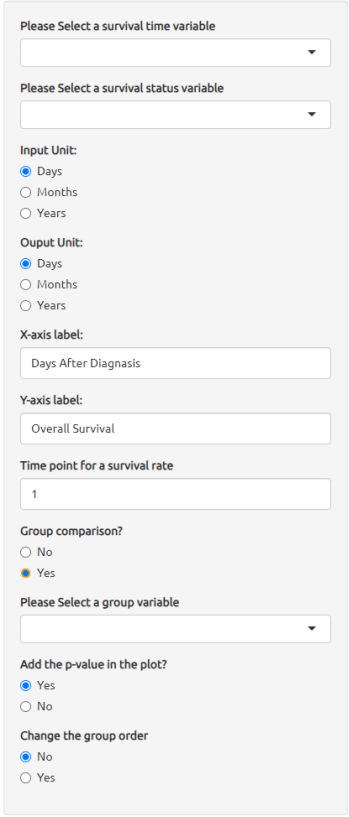


Figure 38. The parameter selection panel for the **Survival Curve** tab.

12. If 'Group comparison?' is 'Yes' and 'Change the group order' is 'Yes', type the new group order.

Output

The output panel (right panel) is composed of seven subpanels:

- Summary of the group variable (for a group comparison)
- Kaplan-Meier curve
- Survival rate, Median survival, and, in case of a group comparison, Hazard ratio

The first subpanel provides the overall summary of the group variable. The next subpanel shows the Kaplan-Meier curve. The last subpanel displays the estimates of survival rate, median survival time, and hazard ratios, including results from the Schoenfeld residuals test.

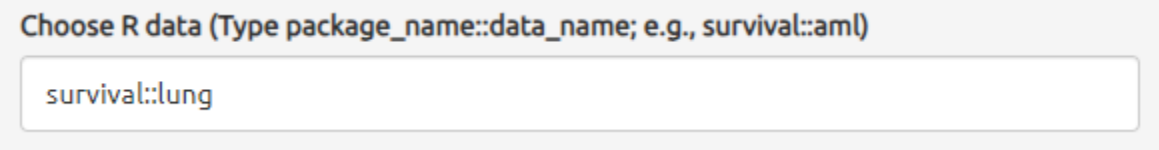
Example

The example data, called **lung**, is available in the R package **survival**. It contains survival information in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities. The example data is composed of the 10 variables, but the example focuses on the following three variables:

- **time**: a numeric vector, survival time in days
- **status**: a numeric vector, censoring status 1=censored, 2=dead
- **sex**: a numeric vector, male=1, female=2

The example aims (i) to generate the KM plot for the overall survival (OS) along with median OS and 1-year OS and (ii) to generate the KM plot for the overall survival (OS) by sex, along with median OS, 1-year OS, and hazard ratio (HR).

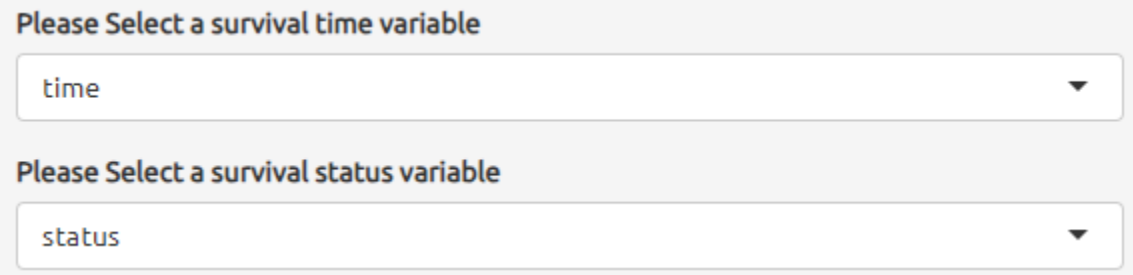
Step by Step



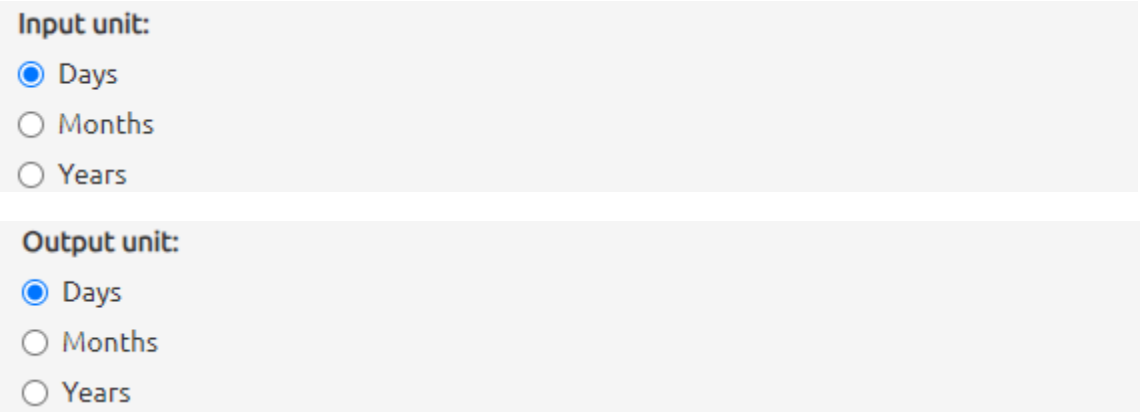
To upload data, go to the **Upload Data** tab and type "**survival::lung**" in the "**Choose R data (Type package_name::data_name; e.g., survival::aml)**" option as shown below:

Note: Please refer to the section "**Upload Data**" for detailed instructions on how to upload data.

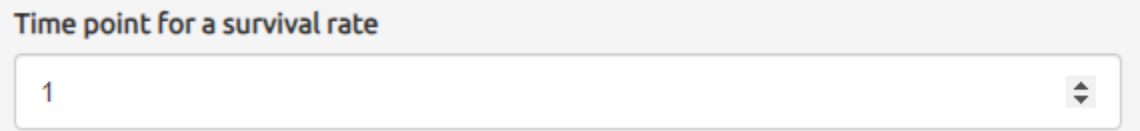
Once the data is uploaded, go to the **Survival Curve** tab. Then select the time duration variable, **time**, and the status variable, **status**:



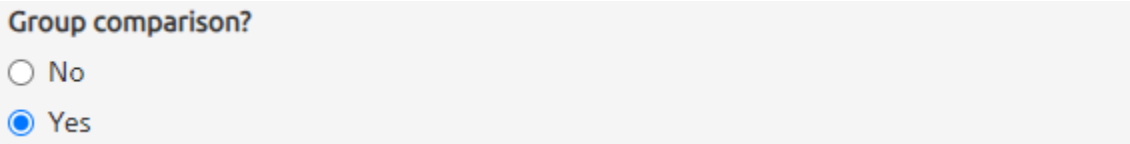
Select the 'Days' for the input unit and the 'Years' for the output unit, along with typing the labels:



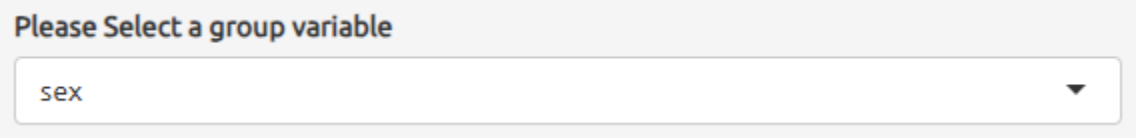
Then type '1' for a survival rate at 1-year:



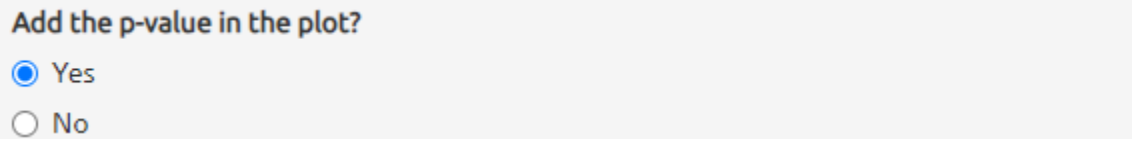
In case of a group comparison, select 'Yes' for 'Group comparison?':



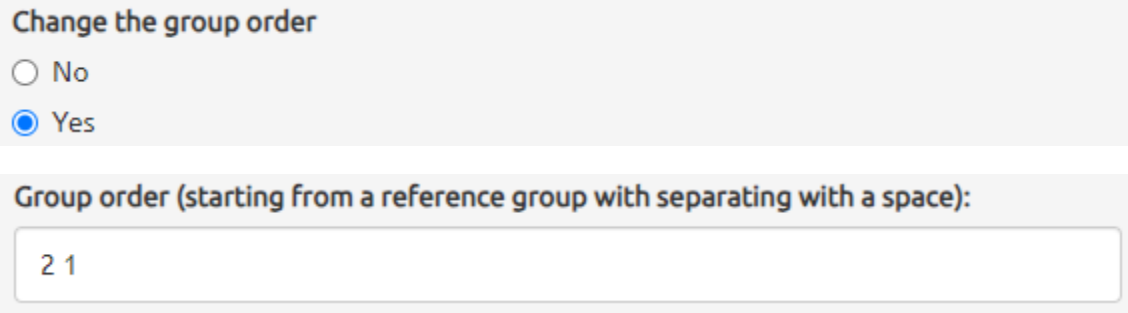
Then select the group variable:



Select 'Yes' to add the p-value to the plot:



Choose 'Yes' to change the group order from '1' and '2' to '2' and '1', followed by typing the new group order:



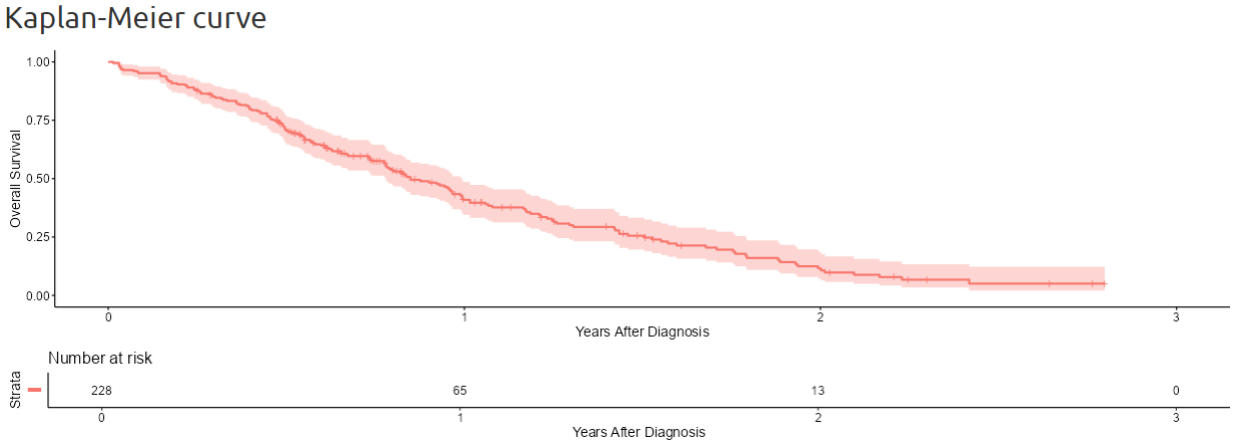
Results and Interpretation

When no group comparison is performed, there are two following results.

Kaplan-Meier curve

The Kaplan-Meier (KM) plot illustrates the overall survival (OS) of the patient cohort over a three-year period following diagnosis. The survival curve starts at 1.00 (or 100%) and gradually declines, indicating a decrease in the proportion of patients surviving over time. At the one-year mark, approximately 65 patients remain at risk, while by the end of the third year, only 13 patients are still at risk. The shaded area around the survival curve represents the 95% confidence interval, providing a measure of the uncertainty around the estimated

survival probabilities. The number at risk table below the plot provides the number of patients still under observation at different time points, which helps in understanding the patient follow-up and the reliability of the survival estimates at various times. Overall, the plot suggests a continuous decline in survival over the three-year period post-diagnosis:



Survival rate and Median Survival

Survival rate and Median survival

```
$Rate
Call: survfit(formula = Surv(t, s) ~ 1, data = tdata)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     65     121    0.409  0.0358        0.345        0.486

$Median
Call: survfit(formula = Surv(t, s) ~ 1, data = tdata)

      n events median 0.95LCL 0.95UCL
[1,] 228    165  0.849    0.78   0.994
```

The estimated 1-year overall survival (OS) rate indicates that 40.9% of the patients are expected to survive one year after diagnosis, with a 95% confidence interval ranging from 34.5% to 48.6%. This estimate suggests moderate survival within the first year. Additionally, the median OS, the time by which 50% of the patients are expected to have survived, is estimated to be 0.849 years (approximately 10.2 months), with a 95% confidence interval from 0.78 to 0.994 years. This information highlights that half of the patient cohort is expected to survive around 10.2 months post-diagnosis, indicating the need for further interventions or treatments to improve survival outcomes.

When a group comparison is performed, there are three following results.

Summary of the group variable

The group variable, sex, has two levels: 1 (Male) and 2 (Female). There are a total of 138 samples with '1' and 90 samples with '2', resulting in a total of 228 samples:

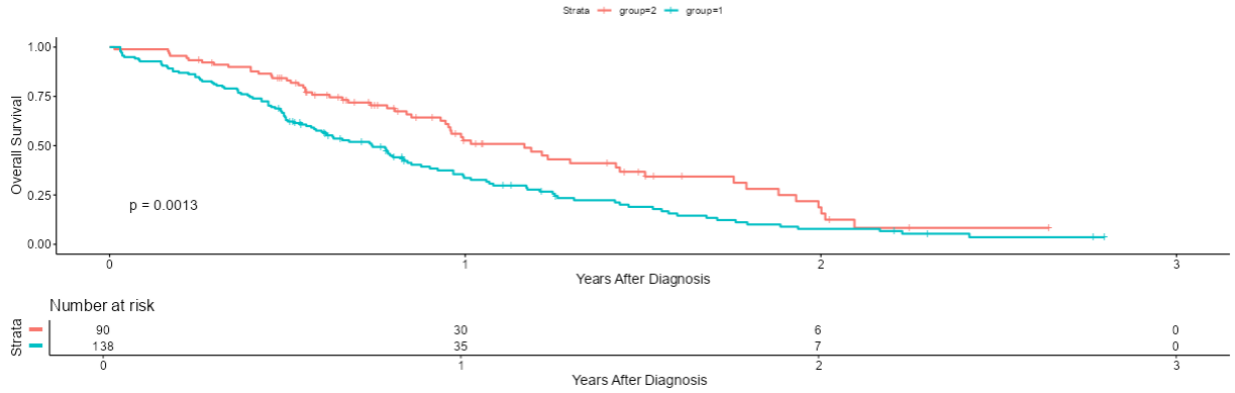
Summary of the group variable

```
$Unique_values
[1] "1" "2"

$Summary_table
var1
  1   2
138  90
```

Kaplan-Meier curve

The Kaplan-Meier (KM) plot illustrates the overall survival (OS) by sex, with group 1 representing males and group 2 representing females. The survival curves show a clear difference between the two groups. Females (group 2, red line) have a higher survival probability over time compared to males (group 1, blue line). The log-rank test yields a p-value of 0.0013, indicating that the difference in survival between the two groups is statistically significant. The number at risk table shows the decreasing number of patients at risk over time for both groups, with more males remaining at risk at earlier time points but dropping off more sharply compared to females. This analysis suggests that females have a better overall survival than males in this cohort:

Kaplan-Meier curve



Survival rate, Median survival, and Hazard ratio

The analysis of overall survival (OS) by sex reveals significant differences between the two groups. For the 1-year OS, females (group 2) have a survival rate of 52.65% (95% CI: 42.15% - 62.95%), whereas males (group 1) have a survival rate of 33.61% (95% CI: 26.09% - 42.32%). The median OS for females is 1.166 years (95% CI: 0.953 - 1.566), while for males it is 0.739 years (95% CI: 0.580 - 0.849). The Cox proportional hazards model shows that being male (group 1) is associated with a higher hazard ratio (HR) of 1.70 (95% CI: 1.2255 - 2.3606, p = 0.0015), indicating that males have a significantly higher risk of death compared to females. The proportional hazards assumption test yields a non-significant global p-value (0.091), suggesting that the proportional hazards assumption is reasonably

met for this model. This analysis indicates that females have better overall survival outcomes compared to males:

Survival rate, Median survival, and Hazard ratio

```
$Rate
Call: survfit(formula = Surv(t, s) ~ group, data = tdata)

               group=2
       time       n.risk      n.event     survival      std.err lower 95% CI
     1.0000      30.0000      36.0000       0.5265       0.0597       0.4215
upper 95% CI
     0.6576

               group=1
       time       n.risk      n.event     survival      std.err lower 95% CI
     1.0000      35.0000      85.0000       0.3361       0.0434       0.2609
upper 95% CI
     0.4329


$Median
Call: survfit(formula = Surv(t, s) ~ group, data = tdata)

          n events median 0.95LCL 0.95UCL
group=2  90     53  1.166   0.953   1.506
group=1 138    112  0.739   0.580   0.849

$Cox.output
Call:
coxph(formula = Surv(t, s) ~ group, data = tdata)

  n= 228, number of events= 165

        coef exp(coef) se(coef)     z Pr(>|z|)
group1 0.5310    1.7007   0.1672 3.176  0.00149 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
group1     1.701      0.588     1.226      2.36

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001


$Summary
             HR 95% CI (lower) 95% CI (upper)          p
group1 1.700672       1.225513       2.360061 0.001491229

$PH.assmuption
       chisq df     p
group   2.86  1 0.091
GLOBAL  2.86  1 0.091
```