### **Hong Kong Baptist University**

### Department of Computer Science

*COMP 7990 Principles and Practices of data analytics (2022-23)*

*Assignment 2*

Name: \_\_\_\_\_\_\_\_\_\_LiTian\_\_\_\_\_\_\_\_\_\_\_\_\_ SID:\_\_\_\_\_\_\_22401792\_\_\_\_\_\_\_\_

**Exercise 1 – Generate descriptive statistics**

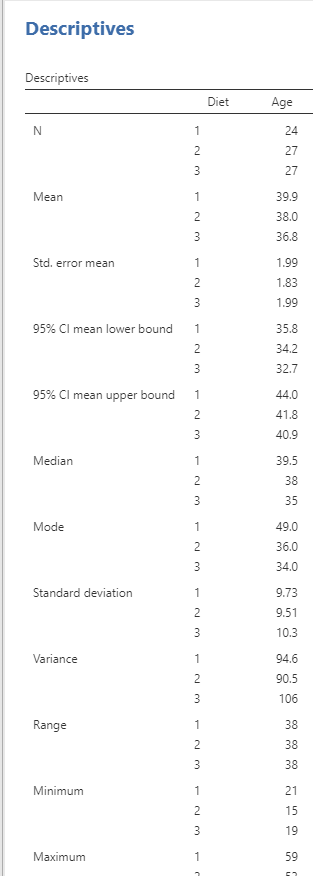
1. Download and unzip the file **Assignment2-datafile.zip**. Open the file **diet.csv** using Jamovi, save the file as **diet.omv**. This data set contains information on 78 people using one of three diets.
2. Change the measure types for the following attributes:

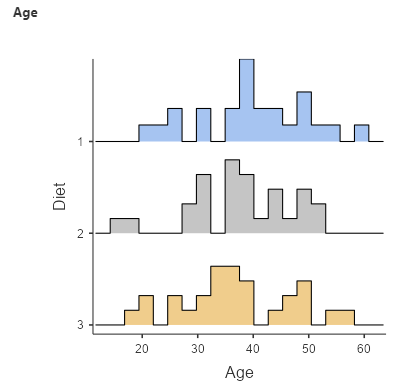
|  |  |
| --- | --- |
| **Attribute** | **Measure Type** |
| Person | ID |
| Age | Continuous |
| Height | Continuous |
| pre.weight | Continuous |

1. Show a statistical summary for the variable **Age**, including the sample size, percentiles (25,50,75), mean, median, mode, standard deviation, variance, range, maximum and minimum. Show age distribution in histogram. Paste the screenshots in the table below.

|  |  |
| --- | --- |
| **Descriptives** | **Age distribution in histogram** |
|  |  |

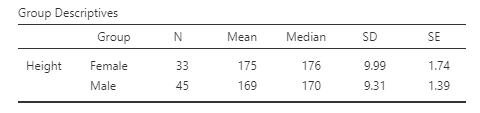
1. Split the ages by different diets. And **save** the **diet.omv** file.



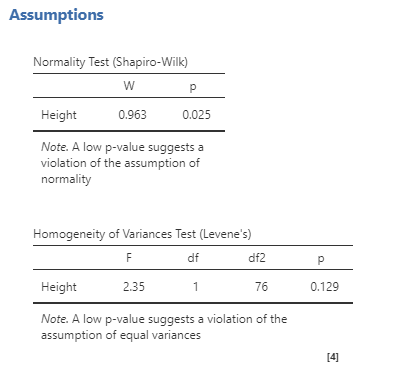


**Exercise 2 – Independent samples t-test**

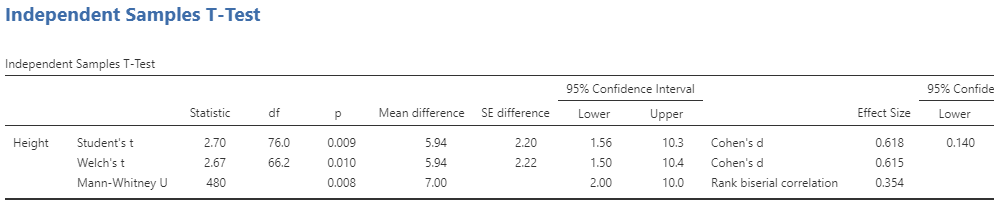
1. Independent samples t-test compares the means of two independent groups to determine whether there is a statistically significant difference between the means of an attribute. Let us perform an **independent samples t-test** to compare the means of height for different genders by using **diet.omv**.
2. Show the **Group Descriptives** table.



1. Perform the assumption checks on **normality** and **homogeneity (equality of variance)** and paste the results below. Does it violate Normality Test (Shapiro-Wilk test) or Homogeneity of Variances Test (Levene’s test)?



1. Does it violate Normality Test? \_\_\_\_\_Yes\_\_\_\_\_\_\_\_\_\_(Yes/No)
2. Does it violate Homogeneity of Variances Test? \_\_\_\_\_\_\_No\_\_\_\_\_(Yes/No)
3. Suggest a way to correct the results if one of the tests above is violated: \_\_Leveraging Mann-Whiteny U\_test to against the normality assumption.\_\_\_
4. In the previous question, if one of the tests is violated, you may use another test to correct it and **capture** the independent sample t test result screenshot including **p value,** **mean difference, effect size** etc.



1. Report the **overall test results**.

• An independent sample t-test was conducted to determine if there were significant differences in Height between male and female students.

• Height for each level of gender were not normally distributed, as assessed by Shapiro

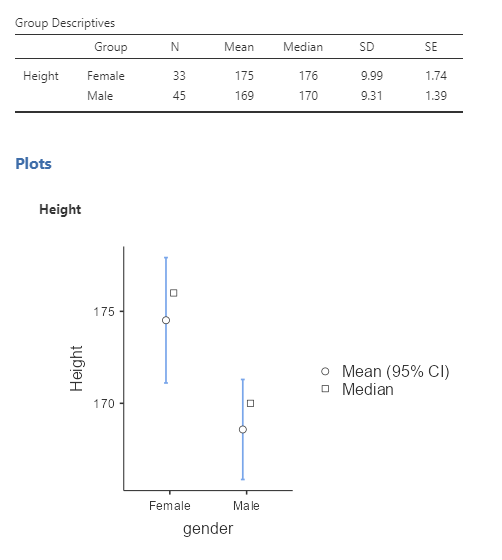
Wilk test (p < 0.05).

• Homogeneity of variances was met, as assessed by Levene’s Test for Equality of Variance (p > 0.05).

• On average, the Height of female students (M=175, SD=9.99) were higher than the

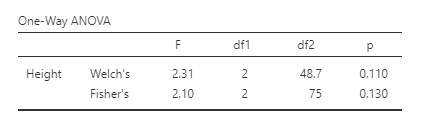
scores of male students (M=169, SD=9.31), a statistically significant difference, Mean difference = 6, 95%CI[1.56, 10.3], t(76)=2.70, p>0.001, d=0.618. These results support the hypothesis that the means for these two groups are not equal.

1. Show the **Descriptives plots**.

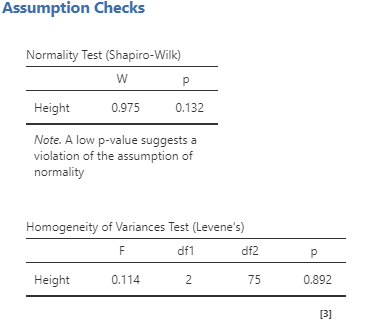


**Exercise 3 – ANOVA**

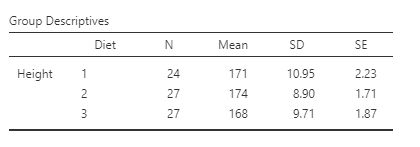
1. The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent groups. Let us perform **one-way ANOVA** to compare the height of people with 3 different diets by using **diet.omv**.
2. Show the **one-way ANOVA** result table below.



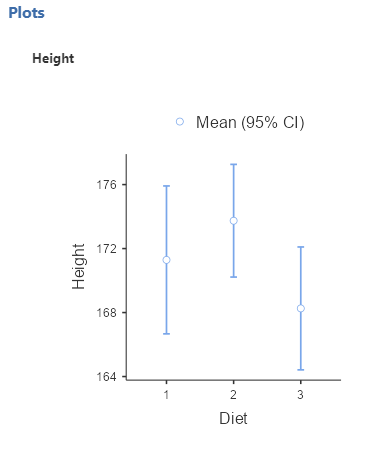
1. Perform the assumption checks on **Normality** **Test** and **Homogeneity Test (for checking equality of variance).** Paste the results in the box below. Determine whether Fisher’s test will be used and capture the results again.



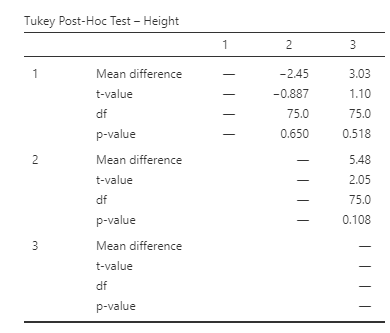
1. Check the option **Descriptives** **table** to show the **Group Descriptives**.



1. Show the **Descriptives plots**.



1. Perform the **post-hoc test** using **Tukey** and paste the result below. Include the mean difference, p-value in your table.



1. Report the test result.

• A one-way ANOVA test was conducted to determine if there were significantly differences in Height among different Diet.

• Height for each diet were normally distributed, as assessed by Shapiro-Wilk test (p > 0.05).

• Homogeneity of variances was met, as assessed by Levene’s Test for Equality of Variance (p > 0.05), so Fisher’s test was used.

• There was not a significant difference of the Height at the p > 0.05 for the three diets.

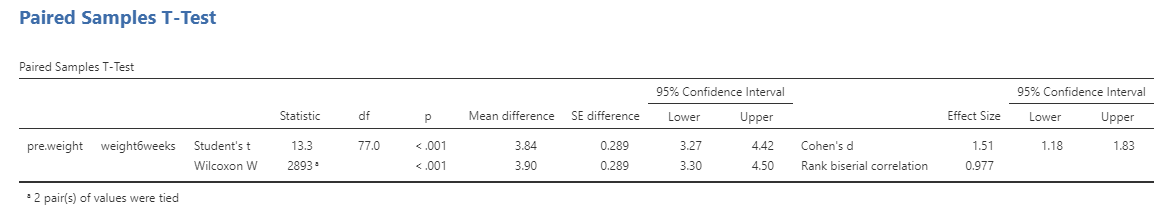
F(2,75)=0.114, p=0.892. The post hoc comparisons using Tukey test indicated that the mean

Height for 3 diets was not significantly different.

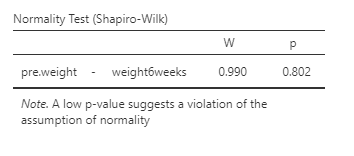
1. Save the file **diet.omv**

**Exercise 4 – Paired samples t-test**

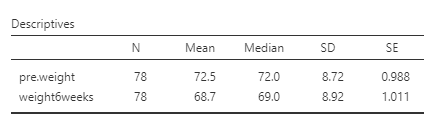
1. The paired-samples t-test compares the means of two related groups to determine whether there is a statistically significant difference between these means. Let us perform a **paired samples t-test** to compare the weight before and after taking a fitness program by using **diet.omv.** (Assume pre-weight is the weight before taking the fitness program, weight6weeks is the weight after a 6-week fitness program)
2. Show the **paired samples t-test** result table below. Include the mean difference (CI=95%) and effect size in your table.



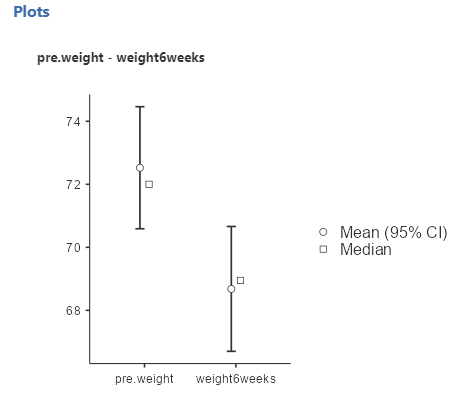
1. Perform the assumption check by selecting **Normality Test** and paste the result below.



1. Show the **Descriptives table**.



1. Show the **Descriptives plots**.



1. Report the overall test result.

• A paired sample t-test was conducted to determine if there were statistically significant differences in weight means before and after taking a fitness class.

• As assessed by Shapiro-Wilk test (p > 0.05), the weight was normality distributed.

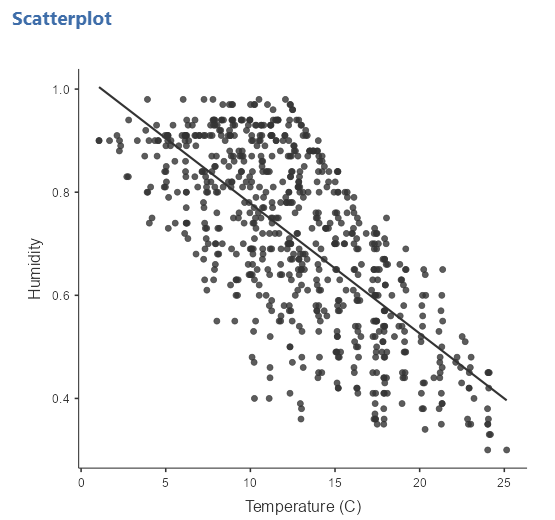
• The t-test result shows a significant difference (p<0.001) between the weight means before and after a fitness program.

• The median weight was 72.0 before the class, compared to 69 after the fitness class. These results support that the enhancement class made a difference on the weight.

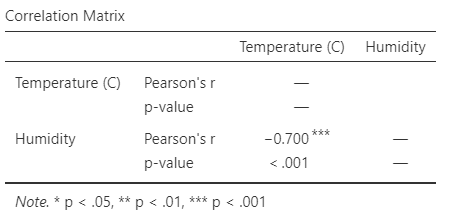
1. Save the file **diet.omv**

**Exercise 5 – Correlation and linear regression**

1. Simple linear regression is useful for finding relationship between two continuous variables. Open the file **weather.csv** using Jamovi and save the file as **weather.omv**
2. Show the **scatterplot** of Humidity (Y-axis) vs Temperature (X-axis). Include the linear regression line in your scatterplot. (Hint: **scatr** module must be installed first)

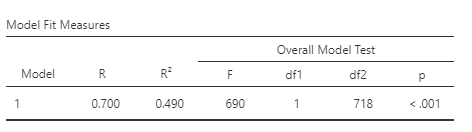


1. By using **Correlation Matrix**, find the **Pearson correlation coefficient** between **Temperature** and **Humidity**. Show the **correlation matrix** table below. Include the p-values and flag significant correlations in your table.

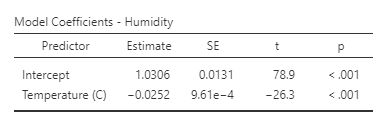


The correlation coefficient **r = \_\_\_\_\_\_\_-0.700\_\_\_\_\_\_\_\_**, this means that the two variables are \_\_**negatively**\_\_\_\_\_\_\_\_\_(positively/negatively) correlated.

1. Use **linear regression** to build a model to predict the humidity by using temperature. Show the **Model Fit Measures table** with overall model test (**R**, **R2** ,**p value** and the **F test**).



1. Show the **Model Coefficients** table.



1. The formula for predicting humidity is:

humidity = 1.0306 - 0.0252(temperature)

1. **Report the overall results** and save the file **weather.omv**

• A simple linear regression was carried out to predict the amount of humidity based on the

temperature.

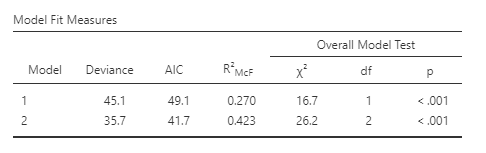
• A significant regression equation was found (F (1,718) = 690, p < 0.001), with an R2 of 0.490. The humidity is equal to 1.0306 - 0.0252(temperature) when the temperature is

measured in degree Celsius.

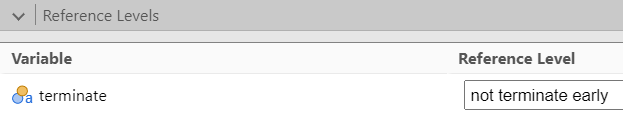
• The humidity decreased by 0.0252 when increase of temperature

**Exercise 6 – Logistic regression**

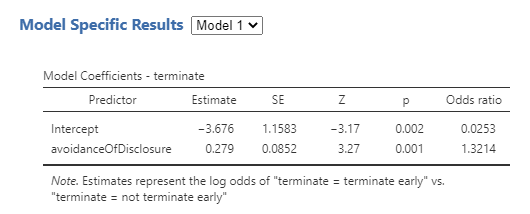
1. **Logistic Regression** is a regression technique that is used when we have a **categorical outcome**. Open the **terminate.csv**, save it as **terminate.omv**. We are going to use binomial logistic regression to determine the likehood of early termination from counseling in samples at a community mental health center.
2. Change the **Measure Type** of the attributes **avoidanceOfDisclosure** and **symptomSeverity** as **continuous**.
3. Use **binomial logistic regression** to build a model. Use **avoidanceOfDisclosure** and **symptomSeverity** as predictors. Place the predictors in different blocks using **Model Builder**. Show the **Model Fit Measures table** by clicking **Overall model test**.

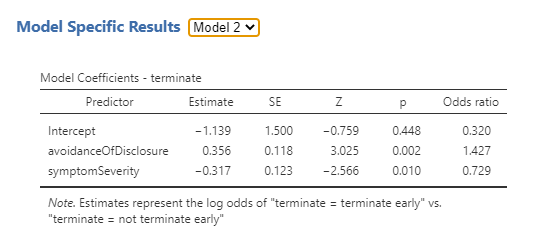


1. Set the **References Levels** of the predictors as follow.

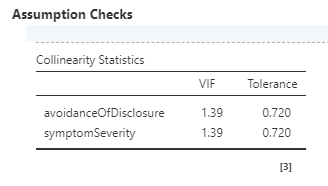


1. Show the **Model Coefficients table**. Include the **odds ratio** in your table.

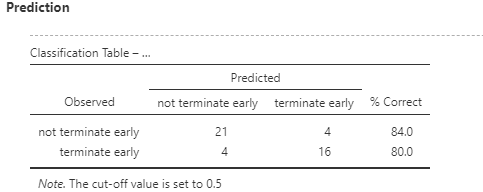




1. Perform the assumption check on **collinearity statistics** and paste the results below.



1. VIF is below 10? \_\_\_\_\_\_\_\_\_Yes\_\_\_\_\_\_\_\_(Yes/No)
2. Tolerance is greater than 0.2? \_\_\_\_\_\_\_\_Yes\_\_\_\_\_\_\_\_(Yes/No)
3. Under the **Prediction** option, show the **Classification table** of the prediction below.



1. **Report the overall result** and save the file **terminate.omv**.

• A binomial logistic regression was carried out to determine the effects of avoidanceOfDisclosure and symptomSeverity on the likehood of early termination from counseling in samples at a community mental health center.

• The logistic regression model was statistically significant, χ2(3) = 26.2, p < 0.001. The model explained 42.3% (McFadden’s R2) of the variance in the decision.

• An increase in avoidanceOfDisclosure score was associated with an increase in the likelihood of not termination early.

• An increase in symptomSeverity score was associated with a decrease in the likelihood of not termination early.

[Model: -1.139 + 0.356(avoidanceOfDisclosure) - 0.317(symptomSeverity)]

• Collinearity statistics show that the assumption of no multicollinearity has been met. (independent variables in a regression model are not correlated)

• True positive rate /sensitivity P(not termination early|termination early). [21/ (21+4)] = 84.0%

• True negative rate /specificity P(not termination early|termination early). [16/ (16+4)] =80.0%

**Assignment Submission**

Submit the following files to [buelearning](https://buelearning.hkbu.edu.hk/) website:

* lab2-assignment-ans.docx
* diet.omv
* weather.omv
* terminate.omv