

AMOD-5210H: Foundations of Modelling

Jasmeet Singh Saini - 0758054

2023-03-23

PART 1: EFFECT SIZES

The following questions requires the use of “**healthdata.xlsx**”.

Firstly, loading the required packages and then reading the excel file.

```
health_dataset <- read_excel("health-data.xlsx")
```

Now, let's performing data extraction.

```
set.seed(0758054)
index <- sample(1:nrow(health_dataset),200)
AMOD5210_Part1 <- health_dataset[index, ]
```

Part 1: Question 1

Using an appropriate inferential statistic and effect size, determine whether there is a significant difference between students and non-students on “Health” and “Depress”.

For the variable “Health”:

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no difference in "health" variable for students and non-students.

vs.

H_A : There is a difference in "health" variable for students and non-students.

Let's check the head of Health dataset

```
head(AMOD5210_Part1, 3)
```

```
## # A tibble: 3 x 10
##       ID Gender Student Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <dbl> <chr>  <chr>      <dbl>  <dbl>      <dbl>      <dbl>  <dbl>  <dbl>  <chr>
## 1    48 Female No         20    19        19        10    12      1 No
## 2   576 Female Yes         21    11        17         8    25     10 No
## 3   525 Female No         21    18        19        19    15      4 No
## # ... with abbreviated variable name 1: Regulation
```

Now, let's grouping with students and checking the summary.

```
grouping_student <- group_by(AMOD5210_Part1, Student)
get_summary_stats(grouping_student, Health, type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Student variable      n mean   sd
##   <chr>   <fct>    <dbl> <dbl> <dbl>
## 1 No     Health    161  16.1  5.02
## 2 Yes    Health     39  18.5  5.09
```

Now, we need to test some **assumptions** about our data.

```
identify_outliers(grouping_student, Health)
```

```
## # A tibble: 1 x 12
##   Student ID Gender Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <chr>   <dbl> <chr>   <dbl> <dbl>      <dbl>      <dbl> <dbl> <dbl> <chr>
## 1 No     354 Female     23    17      18      10      3      0 No
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## # abbreviated variable name 1: Regulation
```

Here, We get no outliers. We will now test for **normality** from the health-data. For that, we will use the **Shapiro-Wilks Test**. If $p > 0.05$, the data is normal.

```
shapiro_test(grouping_student, Health)
```

```
## # A tibble: 2 x 4
##   Student variable statistic      p
##   <chr>   <chr>      <dbl> <dbl>
## 1 No     Health      0.987 0.158
## 2 Yes    Health      0.969 0.347
```

Now, we need to test for **homogeneity of variance**. We can use the **Levene's Test** for this. If $p > 0.05$, variances are homogeneous.

```
levene_test(AMOD5210_Part1, Health ~ Student)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## # A tibble: 1 x 4
##   df1 df2 statistic      p
##   <int> <int>      <dbl> <dbl>
## 1     1  198      0.347 0.557
```

Step 2 : Testing

Now, we will run **Independent t-test**. Since, the homogeneity of variance assumption was not violated, we will set var.equal to TRUE.

```
t_test(AMOD5210_Part1, Health ~ Student, var.equal=TRUE)
```

```
## # A tibble: 1 x 8
##   .y.    group1 group2    n1    n2 statistic    df      p
## * <chr> <chr> <chr> <int> <int>    <dbl> <dbl> <dbl>
## 1 Health No     Yes     161    39    -2.62   198 0.00959
```

Since, $p < 0.05$, the test shows a significant difference. We have enough evidence to reject the null hypothesis, H_0 .

Calculating Cohen's d

Also, the t-Test is significant, we need to calculate **Cohen's d** for our effect size. We need to specify "paired = FALSE" to indicate the groups are independent, and specify "pooled_sd = TRUE" to indicate the variances are equal.

```
cohens_d(Health ~ Student, data = AMOD5210_Part1, paired = FALSE, pooled_sd = TRUE)
```

```
## Cohen's d |          95% CI
## -----
## -0.47      | [-0.82, -0.11]
##
## - Estimated using pooled SD.
```

Based on Cohen's (1988) conventions we have a small effect.

Step 3 : Conclusion

The current study sought to determine whether or not there is a significant difference between students and non-students on "Health". 200 study participants were randomly sampled from the general public (39 students, 161 non-students). The sample contained no extreme outliers. A Shapiro-Wilks test demonstrated normality by group, and Levene's test demonstrated homogeneity of variance. The mean "Health" variable of non-students in the sample was 16.112 (SD = 5.020) whereas the mean "Health" variable of the students in the sample was 18.462 (SD = 5.088). A Welch's independent t-test showed that the mean difference in "Health" variable between student and non-students in the sample was statistically significant, $t(198) = -2.615876$, $p < 0.05$, $d = -0.47$, with students tending to be more "Healthy" than non-students. According to Cohen's (1988) conventions, this is a small effect.

For the variable "Depress":

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no difference in "depress" variable for students and non-students.

vs.

H_A : There is a difference in "depress" variable for students and non-students.

```
get_summary_stats(grouping_student, Depress, type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Student variable      n mean   sd
##   <chr>    <fct>    <dbl> <dbl> <dbl>
## 1 No      Depress    161  4.83  4.44
## 2 Yes     Depress     39  6.97  5.06
```

Now, we need to test some **assumptions** about our data.

```
identify_outliers(grouping_student, Depress)
```

```
## # A tibble: 7 x 12
##   Student ID Gender Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <chr>   <dbl> <chr>    <dbl>  <dbl>    <dbl>    <dbl>  <dbl>  <dbl> <chr>
## 1 No      249 Female     13    22      10        9    21    20 Yes
## 2 No       96 Male      20    16      17       10    24    21 Yes
## 3 No       92 Female    19    18      21       12    18    17 Yes
## 4 Yes     760 Female    21    15      17       18    29    17 Yes
## 5 Yes     498 Female    24    16      10        6    21    18 Yes
## 6 Yes     557 Female    21    18      15       17    20    18 Yes
## 7 Yes     186 Female    19    13      12       15    26    18 Yes
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## #   abbreviated variable name 1: Regulation
```

Here, We get no outliers. We will now test for **normality** from the health-data. For that, we will use the **Shapiro-Wilks Test**. If $p > 0.05$, the data is normal.

```
shapiro_test(grouping_student, Depress)
```

```
## # A tibble: 2 x 4
##   Student variable statistic      p
##   <chr>    <chr>    <dbl>    <dbl>
## 1 No      Depress    0.894 0.00000000241
## 2 Yes     Depress    0.884 0.000807
```

It can be seen that our data is not normal. We are still going to proceed with the test.

Now, we need to test for **homogeneity of variance**. We can use the **Levene's Test** for this. If $p > 0.05$, variances are homogeneous.

```
levene_test(AMOD5210_Part1, Depress ~ Student)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## # A tibble: 1 x 4
##   df1 df2 statistic      p
##   <int> <int>    <dbl> <dbl>
## 1     1  198    0.158 0.691
```

Step 2 : Testing

Now, we will run **Independent t-test**. Since, the homogeneity of variance assumption was not violated, we will set `var.equal` to `TRUE`.

```
t_test(AMOD5210_Part1, Depress ~ Student, var.equal=TRUE)
```

```
## # A tibble: 1 x 8
##   .y.      group1 group2    n1    n2 statistic    df      p
## * <chr>   <chr>  <chr> <int> <int>    <dbl> <dbl>  <dbl>
## 1 Depress No      Yes    161   39    -2.63   198 0.00908
```

Since, $p < 0.05$, the test shows a significant difference. We have enough evidence to reject the null hypothesis, H_0 .

Calculating Cohen's d

Also, the t-Test is significant, we need to calculate **Cohen's d** for our effect size. We need to specify “paired = FALSE” to indicate the groups are independent, and specify “pooled_sd = TRUE” to indicate the variances are equal.

```
cohens_d(Depress ~ Student, data = AMOD5210_Part1, paired = FALSE, pooled_sd = TRUE)
```

```
## Cohen's d |          95% CI
## -----
## -0.47      | [-0.82, -0.12]
##
## - Estimated using pooled SD.
```

Based on Cohen's (1988) conventions we have a small effect.

Step 3 : Conclusion

The current study sought to determine whether or not there is a significant difference between students and non-students on variable “Depress”. 200 study participants were randomly sampled from the general public (39 students, 161 non-students). The sample contained no extreme outliers. A Shapiro-Wilks test demonstrated normality by group, and Levene's test demonstrated homogeneity of variance. The mean “Depress” variable of non-students in the sample was 4.826 (SD = 4.445) whereas the mean “Depress” variable of the students in the sample was 6.974 (SD = 5.055). A Welch's independent t-test showed that the mean difference in “Depress” variable between student and non-students in the sample was statistically significant, $t(198) = -2.634914$, $p < 0.05$, $d = -0.47$, with students tending to be more “Depress” than non-students. According to Cohen's (1988) conventions, this is a small effect.

Part 1: Question 2

Using an appropriate inferential statistic and effect size, determine whether there is a significant difference in the proportion of men and women diagnosed with or without depression.

We are going to use a χ^2 Test of Independence for this question.

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no difference in the proportion of men and women diagnosed with or without depression.

vs.

H_A : There is a significant difference in the proportion of men and women diagnosed with or without depression.

Below is the frequency table based on the continuous variable Gender and Dstatus:

```
frequency_table <- table(AMOD5210_Part1$Gender, AMOD5210_Part1$Dstatus)
frequency_table
```

```
##
##           No Yes
##   Female 137  14
##   Male   47   2
```

Step 2 : Testing

Let us now perform the test, χ^2 Test of Independence.

```
chisq.test(x = frequency_table, correct = FALSE)
```

```
## Warning in chisq.test(x = frequency_table, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##   Pearson's Chi-squared test
##
## data:  frequency_table
## X-squared = 1.3539, df = 1, p-value = 0.2446
```

```
chisq.posthoc.test(frequency_table)
```

```
## Warning in chisq.test(x, ...): Chi-squared approximation may be incorrect
```

```
##   Dimension      Value      No      Yes
## 1   Female Residuals -1.163565  1.163565
## 2   Female  p values  0.978401  0.978401
## 3    Male Residuals  1.163565 -1.163565
## 4    Male  p values  0.978401  0.978401
```

Since $p > 0.05$ we fail to reject the null hypothesis, H_0 .

Let's perform **Post-hoc Analysis** test:

```
chisq.posthoc.test(frequency_table)
```

```
## Warning in chisq.test(x, ...): Chi-squared approximation may be incorrect
```

```
##   Dimension      Value      No      Yes
## 1   Female Residuals -1.163565  1.163565
## 2   Female  p values  0.978401  0.978401
## 3   Male  Residuals  1.163565 -1.163565
## 4   Male  p values  0.978401  0.978401
```

Calculating Cohen's d

Now, need to calculate the odds ratio as our **effect size**.

```
oddsratio(frequency_table)
```

```
## Odds ratio |      95% CI
## -----
## 0.42       | [0.09, 1.90]
```

Based on Cohens (1988) conventions, we have less than the small category.

Step 3 : Conclusion

The present research seeks to determine whether there is a significant difference in the proportion of men and women diagnosed with or without depression. 200 people (49 Male, 151 Female) reported if they diagnosed with (16) or without depression (184). A Chi-square Test of Independence revealed that there is no difference in the proportion of men and women diagnosed with or without depression, Chi Squared(1, N = 200) = 1.3539, $p > 0.001$, OR = 0.42. According to Cohen's (1988) conventions, this effect was small.

Part 1: Question 3

Researchers are interested to determine whether character strengths are significant predictors of depression symptoms.

a) Using the Pearson's r correlation and r^2 , determine whether there are significant correlations between depression symptoms and the four character strengths variables ("Honesty", "Leader", "Persevere", "Regulation"). Report the r and p -values for each correlation. Also, report r^2 for each correlation.

Fristly, let's see the summary statistics:

```
describe(AMOD5210_Part1, fast = TRUE)

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

##          vars   n  mean    sd min  max range   se
## ID          1 200 468.48 268.94 13  944   931 19.02
## Gender       2 200   NaN    NA Inf -Inf  -Inf   NA
## Student      3 200   NaN    NA Inf -Inf  -Inf   NA
## Honesty      4 200  21.06   2.52 12   25    13  0.18
## Leader       5 200  18.18   3.22  8   25    17  0.23
## Persevere    6 200  19.06   3.40  8   25    17  0.24
## Regulation   7 200  16.59   3.60  6   25    19  0.25
## Health       8 200  16.57   5.11  3   29    26  0.36
## Depress     9 200   5.24   4.64  0   21    21  0.33
## Dstatus     10 200   NaN    NA Inf -Inf  -Inf   NA
```

Now, we need to test some assumptions about our data. Let's, check weather a sample contain any extreme outliers or not?

Let's first check the assumptions "Honesty" variable.

```
identify_outliers(AMOD5210_Part1, Honesty)

## # A tibble: 7 x 12
##       ID Gender Student Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <dbl> <chr>  <chr>    <dbl>  <dbl>    <dbl>    <dbl>  <dbl>  <dbl>  <chr>
## 1   249 Female No         13    22        10        9    21    20 Yes
## 2   935 Female No         13    18        20       21    16     2 No
## 3    23 Female No         13    20        14       12    22     8 No
## 4    47 Female Yes         15    13        10       14    25    14 No
```



```
## 5 666 Male Yes 12 15 20 15 15 6 No
## 6 792 Female No 14 11 18 11 8 0 No
## 7 344 Female No 14 15 11 16 4 0 No
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## # abbreviated variable name 1: Regulation
```

Hence, there is no extreme outliers in “*Honesty*” variable.

Checking the “**Leader**” variable.

```
identify_outliers(AMOD5210_Part1, Leader)
```

```
## # A tibble: 1 x 12
## ID Gender Student Honesty Leader Persevere Regulat~1 Health Depress Dstatus
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
## 1 385 Male No 21 8 23 15 19 2 No
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## # abbreviated variable name 1: Regulation
```

Hence, there is no extreme outliers in “*Leader*” variable.

Checking the “**Persevere**” variable.

```
identify_outliers(AMOD5210_Part1, Persevere)
```

```
## # A tibble: 5 x 12
## ID Gender Student Honesty Leader Persevere Regulat~1 Health Depress Dstatus
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
## 1 249 Female No 13 22 10 9 21 20 Yes
## 2 180 Female Yes 20 14 10 17 26 6 No
## 3 498 Female Yes 24 16 10 6 21 18 Yes
## 4 47 Female Yes 15 13 10 14 25 14 No
## 5 714 Female No 22 18 8 14 18 9 No
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## # abbreviated variable name 1: Regulation
```

Hence, there is no extreme outliers in “*Persevere*” variable.

Checking the “**Regulation**” variable.

```
identify_outliers(AMOD5210_Part1, Regulation)
```

```
## # A tibble: 1 x 12
## ID Gender Student Honesty Leader Persevere Regulat~1 Health Depress Dstatus
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
## 1 498 Female Yes 24 16 10 6 21 18 Yes
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## # abbreviated variable name 1: Regulation
```

Hence, there is no extreme outliers in “*Regulation*” variable.

Now, we will check if the data is **normally distributed** or not. For that, we will use the *Shapiro-Wilks Test* for this. If $p > 0.05$, the data is normal.

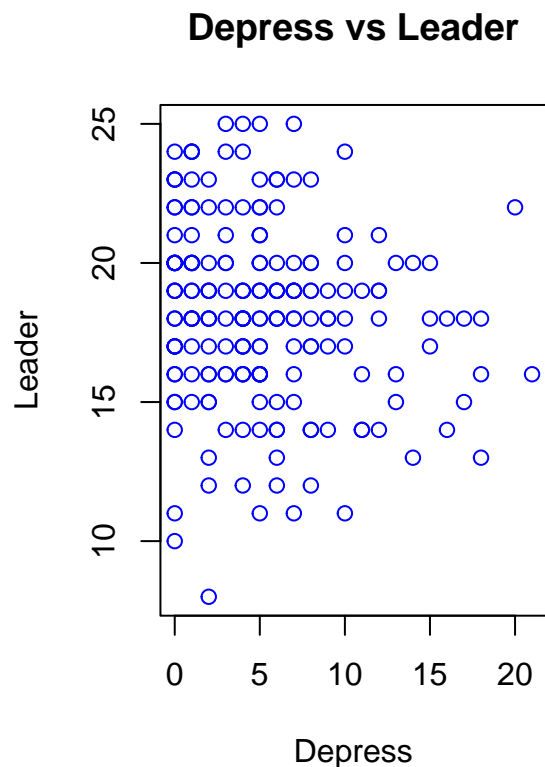
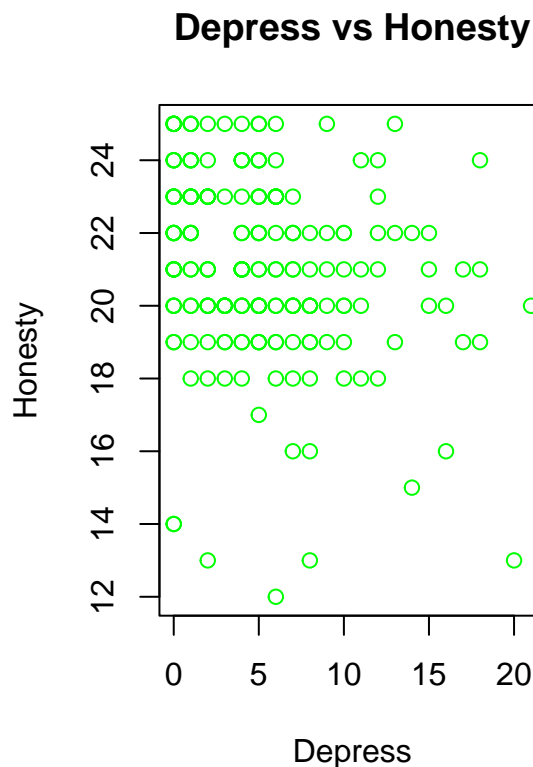
```
shapiro_test(AMOD5210_Part1, vars = c("Honesty", "Leader", "Persevere", "Regulation"))
```

```
## # A tibble: 4 x 3
##   variable    statistic      p
##   <chr>      <dbl>    <dbl>
## 1 Honesty    0.933 0.0000000641
## 2 Leader    0.982 0.0105
## 3 Persevere 0.959 0.0000151
## 4 Regulation 0.978 0.00305
```

Here, we can see that the $p < 0.05$ for all the four variables, that is, “Honesty”, “Leader”, “Persevere”, “Regulation”. And we can say that, the data is not Normal Distribution.

Finally, we need to check the *Linearity* for all the four variables, that is, “Honesty”, “Leader”, “Persevere”, “Regulation”.

```
# Checking Linearity for all of the four with respect to depress variable
par(mfrow=c(1,2))
plot(AMOD5210_Part1$Depress, AMOD5210_Part1$Honesty
     , col = "green", xlab = "Depress", ylab = "Honesty"
     , main = "Depress vs Honesty")
plot(AMOD5210_Part1$Depress, AMOD5210_Part1$Leader
     , col = "blue", xlab = "Depress", ylab = "Leader"
     , main = "Depress vs Leader")
```

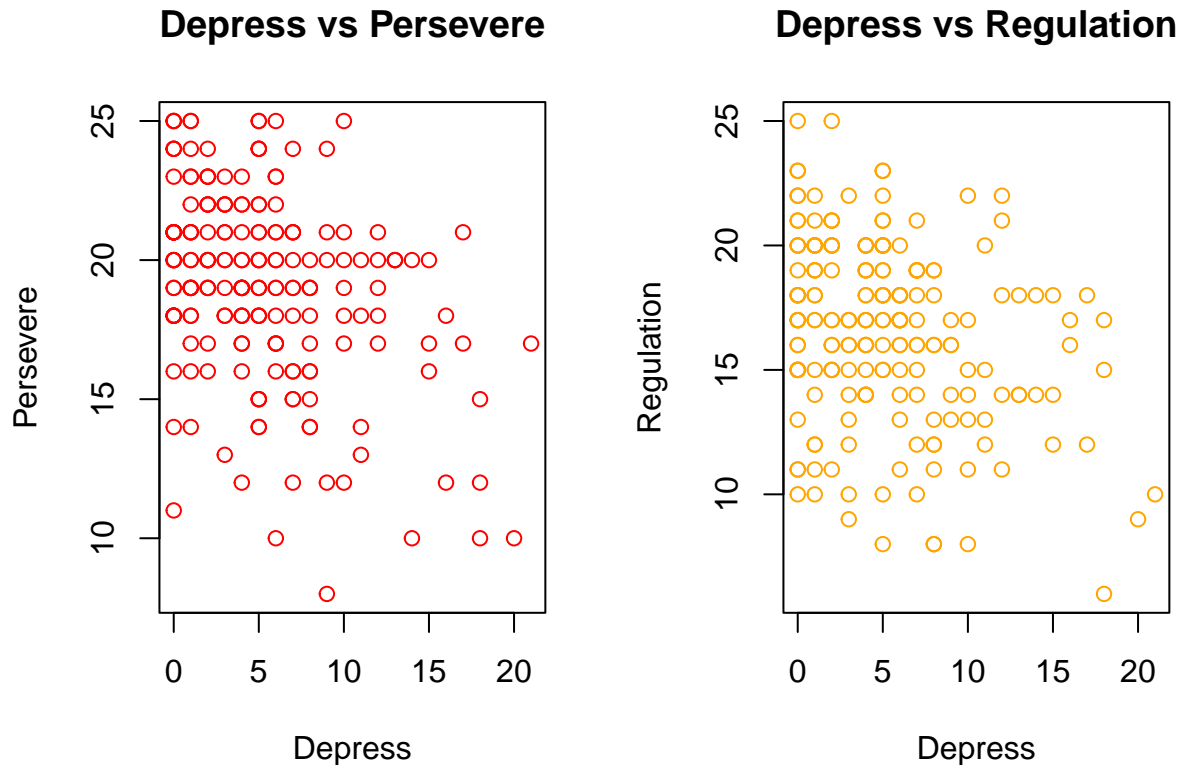


```

par(mfrow=c(1,2))
plot(AMOD5210_Part1$Depress, AMOD5210_Part1$Persevere
     , col = "red", xlab = "Depress", ylab = "Persevere"
     , main = "Depress vs Persevere")

plot(AMOD5210_Part1$Depress, AMOD5210_Part1$Regulation
     , col = "orange", xlab = "Depress", ylab = "Regulation"
     , main = "Depress vs Regulation")

```



None of the four variables shows the Linearity. Hence, we can say that the Linearity condition is not satisfied.

Pearson's r Correlation test

Now, we will run **Pearson's r Correlation test** for all the four variables, that is, "Honesty", "Leader", "Persevere", "Regulation".

The proportion of variability in Depress variable explained by the other Honesty is given below:

```

corr_honesty <- corr.test(AMOD5210_Part1$Depress, AMOD5210_Part1$Honesty, method = "pearson")
corr_honesty

## Call:corr.test(x = AMOD5210_Part1$Depress, y = AMOD5210_Part1$Honesty,
##      method = "pearson")
## Correlation matrix
## [1] -0.24
## Sample Size
## [1] 200

```

```
## These are the unadjusted probability values.
## The probability values adjusted for multiple tests are in the p.adj object.
## [1] 0
##
## To see confidence intervals of the correlations, print with the short=FALSE option
```

The values are given below:

```
# p-Value is
corr_honesty$p
```

```
## [1] 0.0006999901
```

```
# R Value is
corr_honesty$r
```

```
## [1] -0.2377245
```

```
# R Square Value, coefficient of determination is
corr_honesty$r^2
```

```
## [1] 0.05651294
```

The *p-value* is **0.00069**. Here, $p - value < 0.05$, this Correlation for Depress vs Honesty is not significant.

The **Pearson's**, r is **-0.23** and **coefficient of determination**, r^2 is **0.056**. Hence, both shows **small** effect sizes.

The proportion of variability in Depress variable explained by the other Leader is given below:

```
corr_leader <- corr.test(AMOD5210_Part1$Depress, AMOD5210_Part1$Leader, method = "pearson")
corr_leader
```

```
## Call:corr.test(x = AMOD5210_Part1$Depress, y = AMOD5210_Part1$Leader,
## method = "pearson")
## Correlation matrix
## [1] -0.16
## Sample Size
## [1] 200
## These are the unadjusted probability values.
## The probability values adjusted for multiple tests are in the p.adj object.
## [1] 0.03
##
## To see confidence intervals of the correlations, print with the short=FALSE option
```

The values are given below:

```
# p-Value is
corr_leader$p
```

```
## [1] 0.02792895
```

```
# R Value is  
corr_leader$r
```

```
## [1] -0.1554702
```

```
# R Square Value, coefficient of determination is  
corr_leader$r^2
```

```
## [1] 0.02417097
```

The *p-value* is **0.027**. Here, $p - value < 0.05$, this Correlation for Depress vs Honesty is not significant.

The **Pearson's**, r is **-0.15** and **coefficient of determination**, r^2 is **0.024**. Hence, both shows **small** effect sizes.

The proportion of variability in Depress variable explained by the other Persevere is given below:

```
corr_persevere <- corr.test(AMOD5210_Part1$Depress, AMOD5210_Part1$Persevere, method = "pearson")  
corr_persevere
```

```
## Call:corr.test(x = AMOD5210_Part1$Depress, y = AMOD5210_Part1$Persevere,  
##      method = "pearson")  
## Correlation matrix  
## [1] -0.37  
## Sample Size  
## [1] 200  
## These are the unadjusted probability values.  
## The probability values adjusted for multiple tests are in the p.adj object.  
## [1] 0  
##  
## To see confidence intervals of the correlations, print with the short=FALSE option
```

The values are given below:

```
# p-Value is  
corr_persevere$p
```

```
## [1] 5.462712e-08
```

```
# R Value is  
corr_persevere$r
```

```
## [1] -0.3727463
```

```
# R Square Value, coefficient of determination is  
corr_persevere$r^2
```

```
## [1] 0.1389398
```

The p -value is **5.462712e-08**. Here, p -value < 0.05 , this Correlation for Depress vs Honesty is not significant.

The **Pearson's**, r is **-0.37** and **coefficient of determination**, r^2 is **0.14**. Hence, both shows **moderate** effect sizes.

The proportion of variability in Depress variable explained by the other Regulation is given below:

```
corr_regulation <- corr.test(AMOD5210_Part1$Depress, AMOD5210_Part1$Regulation, method = "pearson")
corr_regulation
```

```
## Call:corr.test(x = AMOD5210_Part1$Depress, y = AMOD5210_Part1$Regulation,
##      method = "pearson")
## Correlation matrix
## [1] -0.3
## Sample Size
## [1] 200
## These are the unadjusted probability values.
## The probability values adjusted for multiple tests are in the p.adj object.
## [1] 0
##
## To see confidence intervals of the correlations, print with the short=FALSE option
```

The values are given below:

```
# p-Value is
corr_regulation$p
```

```
## [1] 1.544422e-05
```

```
# R Value is
corr_regulation$r
```

```
## [1] -0.3004228
```

```
# R Square Value, coefficient of determination is
corr_regulation$r^2
```

```
## [1] 0.09025385
```

The p -value is **1.544422e-05**. Here, p -value < 0.05 , this Correlation for Depress vs Honesty is not significant.

The **Pearson's**, r is **-0.3004** and **coefficient of determination**, r^2 is **0.0902**. Hence, both shows **moderate** effect sizes.

b) Using multiple linear regression, determine whether the four character strengths variables are significant predictors of depression symptoms. Report the slopes and p-values for each character strength and identify which character strengths were significant predictors. Also report and interpret the multiple R^2 for the overall model.

We can create our model using the `lm()` function and store the model as an object, to test **multiple regression test**. Then, we will check the summary.

```
regress.model <- lm(Depress ~ Honesty + Leader + Persevere + Regulation
, data = AMOD5210_Part1)
summary(regress.model)
```

```
##
## Call:
## lm(formula = Depress ~ Honesty + Leader + Persevere + Regulation,
##     data = AMOD5210_Part1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.1388 -2.8512 -0.8262  2.0204 13.3709
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.383330   2.787325   6.595 3.89e-10 ***
## Honesty     -0.099070   0.139200  -0.712 0.477496
## Leader      -0.005705   0.103353  -0.055 0.956038
## Persevere   -0.377685   0.109926  -3.436 0.000722 ***
## Regulation  -0.226089   0.091998  -2.458 0.014862 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.268 on 195 degrees of freedom
## Multiple R-squared:  0.1695, Adjusted R-squared:  0.1525
## F-statistic: 9.953 on 4 and 195 DF,  p-value: 2.382e-07
```

The slopes and p – value for each charter strength variables are as follows:

1. Slopes are:

Honesty is -0.099070,
 Leader is -0.005705,
 Persevere is -0.377685, and Regulation is -0.226089

2. p – value are:

Honesty is 0.477496,
 Leader is 0.956038,
 Persevere is 0.000722, and Regulation is 0.014862

We can observed from the values above that “Honesty” and “Leader” have p – value greater than 0.05 making them a significant predictor.

The values of Multiple r^2 is **0.1695**.

According to Cohen’s d (Cohen, 1988) conventions, the overall model explained a **moderate** proportion of variability in “Depress” variable.

Part 1: Question 4

Using an appropriate inferential statistic and effect size(s), determine whether participants had significantly different scores across the four character strengths (“honesty”, “Leader”, “Persevere”, “Regulation”).

We will use **One Way Repeated Measures ANOVA Test** Inferential Statistic, to check whether participants had significantly different scores across the four character strengths (“honesty”, “Leader”, “Persevere”, “Regulation”). And to measure the effect size, we will use η_p^2 , **partial eta squared**.

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : The participants does not had significantly different scores across the four character strengths

vs.

H_A : The participants had significantly different scores across the four character strengths

```
library(ez)
# Gathering data of different character strengths into one column
strength <- gather(data = AMOD5210_Part1
                    , key = "Character_Strength"
                    , value = "Score", Honesty, Leader, Persevere, Regulation)

# Storing the required data in a new dataframe
data_frame <- strength[ ,c("ID","Character_Strength","Score")]
tail(data_frame,5)
```

```
## # A tibble: 5 x 3
##       ID Character_Strength Score
##   <dbl> <chr>           <dbl>
## 1   186 Regulation         15
## 2   584 Regulation         16
## 3   520 Regulation         25
## 4   636 Regulation         15
## 5   831 Regulation         17
```

Now let's group the data and print the descriptive statistics:

```
char_group <- group_by(data_frame, Character_Strength)
get_summary_stats(char_group, Score, type = "mean_sd")
```

```
## # A tibble: 4 x 5
##   Character_Strength variable      n mean    sd
##   <chr>              <fct>   <dbl> <dbl> <dbl>
## 1 Honesty           Score    200  21.1  2.52
## 2 Leader            Score    200  18.2  3.22
## 3 Persevere         Score    200  19.1  3.40
## 4 Regulation        Score    200  16.6  3.60
```


Now, we need to test some **assumptions** about our data.

```
identify_outliers(char_group, Score)
```

```
## # A tibble: 14 x 5
##   Character_Strength    ID Score is.outlier is.extreme
##   <chr>             <dbl> <dbl> <lgl>      <lgl>
## 1 Honesty           249    13 TRUE      FALSE
## 2 Honesty           935    13 TRUE      FALSE
## 3 Honesty            23    13 TRUE      FALSE
## 4 Honesty            47    15 TRUE      FALSE
## 5 Honesty           666    12 TRUE      FALSE
## 6 Honesty           792    14 TRUE      FALSE
## 7 Honesty           344    14 TRUE      FALSE
## 8 Leader            385     8 TRUE      FALSE
## 9 Persevere         249    10 TRUE      FALSE
## 10 Persevere         180    10 TRUE      FALSE
## 11 Persevere         498    10 TRUE      FALSE
## 12 Persevere          47    10 TRUE      FALSE
## 13 Persevere         714     8 TRUE      FALSE
## 14 Regulation        498     6 TRUE      FALSE
```

Here, We get no outliers. We will now test for **normality** from the Score-data. For that, we will use the **Shapiro-Wilks Test**. If $p > 0.05$, the data is normal.

```
shapiro_test(char_group, Score)
```

```
## # A tibble: 4 x 4
##   Character_Strength variable statistic      p
##   <chr>             <chr>      <dbl>    <dbl>
## 1 Honesty           Score      0.933 0.0000000641
## 2 Leader            Score      0.982 0.0105
## 3 Persevere         Score      0.959 0.0000151
## 4 Regulation        Score      0.978 0.00305
```

Here, $p - value < 0.05$ and it can be seen that our data is not normally distributed for all groups. We are still going to proceed with the test.

Step 2 : Testing

Now we will performing **One Way Repeated Measures ANOVA Test**.

```
rep_ANOVA <- ezANOVA(data_frame, dv = Score, wid = ID
                      , within = Character_Strength, type=3
                      , return_aov = TRUE)
```

```
## Warning: Converting "ID" to factor for ANOVA.
```

```
## Warning: Converting "Character_Strength" to factor for ANOVA.
```

```
# Printing rep_ANOVA
rep_ANOVA
```

```
## $ANOVA
##           Effect DFn DFd           F           p p<.05           ges
## 2 Character_Strength    3 597 99.29093 3.807505e-52      * 0.2016286
##
## $'Mauchly's Test for Sphericity'
##           Effect           W           p p<.05
## 2 Character_Strength 0.8616564 1.900668e-05      *
##
## $'Sphericity Corrections'
##           Effect           GGe           p[GG] p[GG]<.05           HFe           p[HF]
## 2 Character_Strength 0.907239 1.420815e-47      * 0.921045 2.964712e-48
##   p[HF]<.05
## 2      *
##
## $aov
##
## Call:
## aov(formula = formula(aov_formula), data = data)
##
## Grand Mean: 18.72
##
## Stratum 1: ID
##
## Terms:
##           Residuals
## Sum of Squares    4058.28
## Deg. of Freedom      199
##
## Residual standard error: 4.515902
##
## Stratum 2: ID:Character_Strength
##
## Terms:
##           Character_Strength Residuals
## Sum of Squares           2075.42    4159.58
## Deg. of Freedom           3        597
##
## Residual standard error: 2.639597
## Estimated effects may be unbalanced
```

In the ANOVA section, since p-value is less than 0.05, we will **reject** the null hypothesis, H_0 . Thus, we can conclude that participants had significantly different scores across the four character strengths.

Calculating Partial Eta Squared Effect Size Calculation

Now, need to calculate the **effect size**.

```
eta_squared(rep_ANOVA$aov, partial = TRUE)
```

```
## # Effect Size for ANOVA (Type I)
```

```
##
## Group | Parameter | Eta2 (partial) | 95% CI
## -----
## ID:Character_Strength | Character_Strength | 0.33 | [0.28, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

Hence, η_p^2 is 0.33 and based on Cohen's (1998) conventions, we have a Large Effect.

Now let's run our **post-hoc test**

```
t_test(data_frame, Score ~ Character_Strength, paired = TRUE, p.adjust.method = "bonferroni")
```

```
## # A tibble: 6 x 10
##   .y. group1 group2      n1      n2 stati~1    df      p    p.adj p.adj~2
## * <chr> <chr> <chr>    <int> <int>    <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 Score Honesty Leader      200    200    12.0   199 2.91e-25 1.75e-24 ****
## 2 Score Honesty Persevere    200    200     9.02   199 1.58e-16 9.48e-16 ****
## 3 Score Honesty Regulation    200    200    16.8   199 3.85e-40 2.31e-39 ****
## 4 Score Leader Persevere      200    200    -3.39   199 8.49e- 4 5 e- 3 **
## 5 Score Leader Regulation      200    200     5.03   199 1.08e- 6 6.48e- 6 ****
## 6 Score Persevere Regulation    200    200     8.98   199 2.01e-16 1.21e-15 ****
## # ... with abbreviated variable names 1: statistic, 2: p.adj.signif
```

Since adjusted $p < 0.05$ for all of the comparisons, we can say that Scores were Significantly Different across each Character Strength.

calculating effect size for each Pairwise Comparison:

Next, we need to calculate our effect size for each pairwise comparison, cohen's d. Again, we need to use the `cohens_d()` function from the "rstatix" package, so we first have to detach the "effectsize" package.

```
detach("package:effectsize", unload = TRUE)
```

We also need to specify "paired = TRUE" to indicate the data is paired.

```
cohens_d(data_frame, Score ~ Character_Strength, paired = TRUE)
```

```
## # A tibble: 6 x 7
##   .y. group1 group2    effsize      n1      n2 magnitude
## * <chr> <chr> <chr>    <dbl> <int> <int> <ord>
## 1 Score Honesty Leader    0.847    200    200 large
## 2 Score Honesty Persevere  0.638    200    200 moderate
## 3 Score Honesty Regulation  1.19    200    200 large
## 4 Score Leader Persevere  -0.240    200    200 small
## 5 Score Leader Regulation  0.356    200    200 small
## 6 Score Persevere Regulation  0.635    200    200 moderate
```

Step 3 : Conclusion

The current study sought to test whether the participants had significantly different scores across the four character strengths. 200 people were randomly sampled and their scores were measured across 4 different character strengths. Shapiro-Wilks tests on the distribution of scores across each character strength did

not demonstrate normality and Mauchly's test showed the sphericity assumption was met. The mean score for the sample across "Honesty" was 21.055 (SD = 2.525), across "Leader" was 18.175 (SD = 3.218), across "Persevere" was 19.055 (SD = 3.401) and across "Regulation" was 16.595 (SD = 3.605). A One-Way Repeated Measures ANOVA was performed and showed a significant change across the scores of different character strengths, $F(3,597) = 99.29$, $p < 0.05$, $\eta_p^2 = 0.33$. Bonferroni-corrected pairwise comparisons revealed that the score across "Leader" and "Regulation" were significantly higher than score across "Honesty" ($d = 0.84$ and $d = 1.19$ respectively). Both the score of "Persevere" against "Honesty" and the score of "Regulation" against "Persevere" was moderate ($d = 0.637$ and $d = 0.635$ respectively). The score across "Persevere" and "Regulation" were significantly lower than score across "Leader" ($d = 0.24$ and $d = 0.35$ respectively). According to Cohen's (1988) conventions, the "Leader" and "Regulation" against "Honesty" is a large effect and the "Persevere" & "Regulation" against "Leader" is a moderate effect and "Persevere" & "Regulation" against "Leader" is a small effect.

PART 2: DIAGNOSTIC EFFICIENCY STATISTICS

The following questions requires the use of “**diagnostic-data.xlsx**”.

Firstly, loading the required packages and then reading the excel file.

```
library(readxl)
diagnostic_dataset <- read_excel("diagnostic-data.xlsx")
```

Now, let's performing data extraction.

```
set.seed(0758054)
index <- sample(1:nrow(diagnostic_dataset), 200)
AMOD5210_Part2 <- diagnostic_dataset[index, ]
```

Part 2: Question 1

Create a 2 x 2 contingency table for the variables Diagnosis and Test. The contingency table you create should include frequencies within each cell, and each row and column of the table should be meaningfully labelled.

To create a 2 x 2 contingency table for the variables Diagnosis and Test, we will first specify our variables are factors and add some labels.

```
AMOD5210_Part2$Test <- factor(AMOD5210_Part2$Test, labels = c("Yes", "No"))
AMOD5210_Part2$Diagnosis <- factor(AMOD5210_Part2$Diagnosis, labels = c("Yes", "No"))
```

Next we will now create a contingency table.

```
table(AMOD5210_Part2$Test, AMOD5210_Part2$Diagnosis, dnn = c("Test", "Diagnosis"))
```

```
##      Diagnosis
## Test  Yes  No
##   Yes 102  28
##   No   24  46
```

Part 2: Question 2

Report the following diagnostic efficiency statistics: a) sensitivity, b) specificity, c) positive prediction value, d) negative prediction value, e) overall correct classification, and f) Kappa

Now, we will use the confusionMatrix function to calculate our the given diagnostic efficiency statistics:

```
confusionMatrix(data = AMOD5210_Part2$Test
                 , reference = AMOD5210_Part2$Diagnosis
                 , positive = "Yes"
                 , dnn = c("Test", "Diagnosis"))
```

```

## Confusion Matrix and Statistics
##
##      Diagnosis
## Test  Yes  No
##   Yes 102  28
##   No   24  46
##
##                Accuracy : 0.74
##                95% CI : (0.6734, 0.7993)
##      No Information Rate : 0.63
##      P-Value [Acc > NIR] : 0.0006357
##
##                Kappa : 0.436
##
##  Mcnemar's Test P-Value : 0.6773916
##
##      Sensitivity : 0.8095
##      Specificity : 0.6216
##      Pos Pred Value : 0.7846
##      Neg Pred Value : 0.6571
##      Prevalence : 0.6300
##      Detection Rate : 0.5100
##      Detection Prevalence : 0.6500
##      Balanced Accuracy : 0.7156
##
##      'Positive' Class : Yes
##

```

Conclusion:

Following are diagnostic efficiency statistics:

- a) Sensitivity = 0.8095,
- b) Specificity = 0.6216,
- c) Positive Prediction Value = 0.7846,
- d) Negative Prediction Value = 0.6571,
- e) Overall Correct Classification (Accuracy) = 0.74,
- f) Kappa = 0.436

Part 2: Question 3

Based on the diagnostic efficiency statistics reported in Question 2, does the new test accurately diagnose individuals with breast cancer? Explain your answer.

We can observe, through the sensitivity value, that is, **0.8095**, that the test was better at identifying the true positive cases. Also, it was worse in identifying true negative cases. The new test was not accurate in diagnosing breast cancer in individuals due to its low kappa value, that is, **0.436** (As per the conventions for interpreting Kappa).