AMOD-5210H: Foundations of Modelling

Jasmeet Singh Saini - 0758054

2023-03-20

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.

$$H_0 : \text{There is no difference in "health" variable for students and non-students.}$$
$$\text{vs.}$$
$$H_A : \text{There is a difference in "health" variable for students and non-students.}$$

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.

```r
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.

```r
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.

```r
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.

```r
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_O.

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_O.

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 $\chi 2$ 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, $\chi 2$ **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_O.

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
```

8

```
## 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 **normaly 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")
```

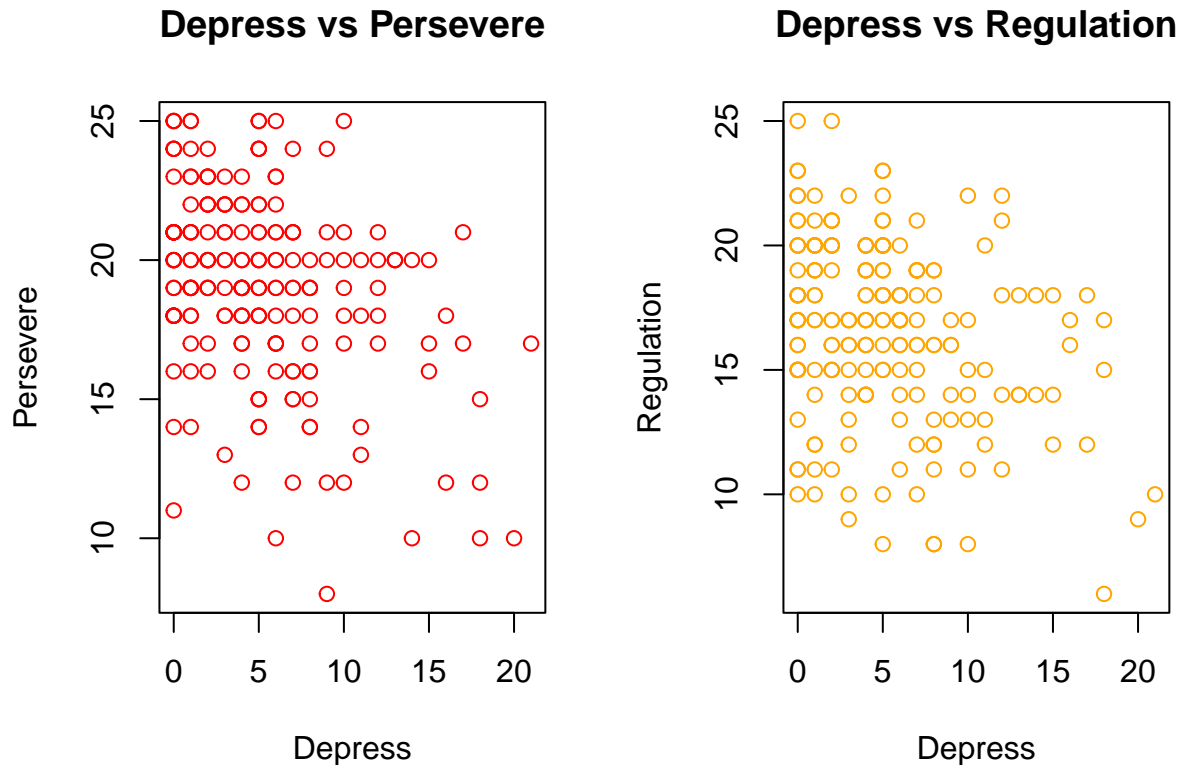## Depress vs Honesty

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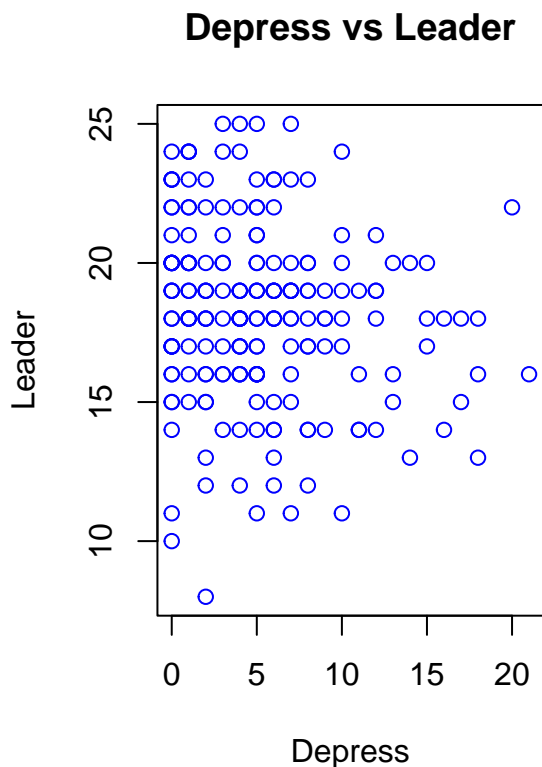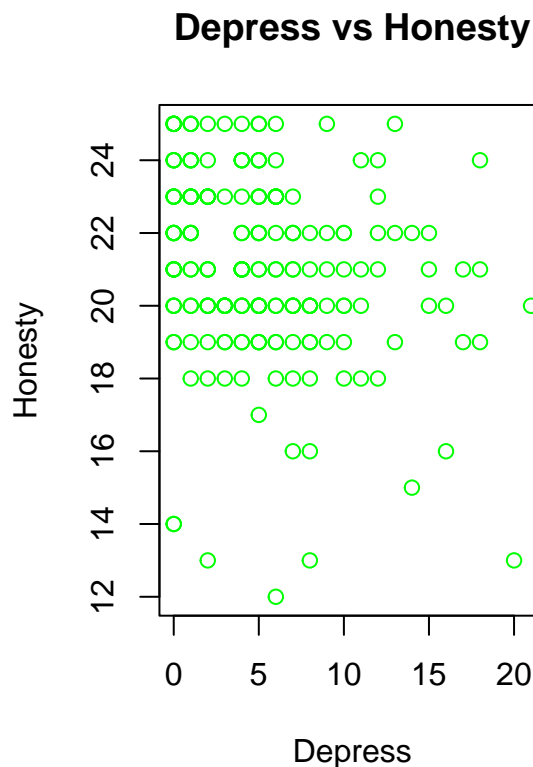
```
par(mfrow=c(1,2))
plot(AMOD5210_Part1$Depress, AMOD5210_Part1$Persevere
     , col = "red", xlab = "Depress", ylab = "Persevere"
       , main = "Depress vs Persevere")

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 followers:

1. Slopes are:

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

2. $p-value$ are:

   Honesty 0.477496,
   Leader 0.956038,
   Persevere 0.000722, Regulation 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").

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 <- diagnostic_dataset[index, ]
AMOD5210
```

```
## # A tibble: 200 x 3
##          ID Diagnosis Test
##       <dbl> <chr>     <chr>
## 1   776715 No        No
## 2  1246562 Yes       Yes
## 3    76389 Yes       Yes
## 4   846832 No        No
## 5  1080233 Yes       No
## 6   704097 No        No
## 7  1344449 No        Yes
## 8   608157 Yes       Yes
## 9  1167439 Yes       No
## 10 1299161 Yes       Yes
## # ... with 190 more rows
```

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.

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

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.