Course on Food Security, Poverty and Nutrition Policy Analysis: Statistical Methods and Applications

Author: Suresh Babu

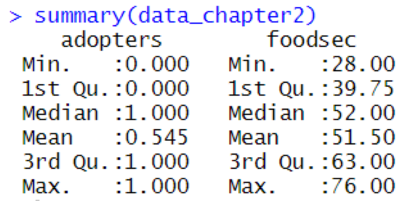
Introduction to the Course Material

This course aims to build statistical skills of students, teachers, and practitioners. It introduces ‘R’ coding for statistical analysis material discussed in the chapters of the book titled ‘Food Security, Poverty, and Nutrition Policy Analysis: Statistical Methods and Applications’ Third Edition by Suresh C. Babu and Shailendra N. Gajanan, an Elsevier/Academic Press publication (forthcoming).The users can download and practice R coding with associated chapter-wise mock data and R files. To access the chapter-wise R files for this course material, please click here.The chapter-wise data can be accessed from the link embedded at the beginning of each chapter.

Chapter 2 Implication of Technological Change, Post-Harvest Technology, and Technology Adoption for Improved Food Security – Application of *t*-Statistic

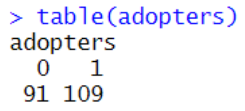
USING R FOR *t*-TESTS (Please find the data used for this chapter here)

We motivate this section with an illustrative example using 200 observations. As mentioned in the previous section, the *t*-test is designed to compare means of the same variable between the two groups. Assume we have random information on FOODSEC for 200 farmers, some of whom are adopters of new hybrid varieties, while others are non-adopter. We first use the *summary* command in R and examine descriptive statistics. With the summary command, we can identify min, each quarter, and max values for each variable. Note that we often are going to use the *summary* command to investigate more information about variables and analysis throughout the book.



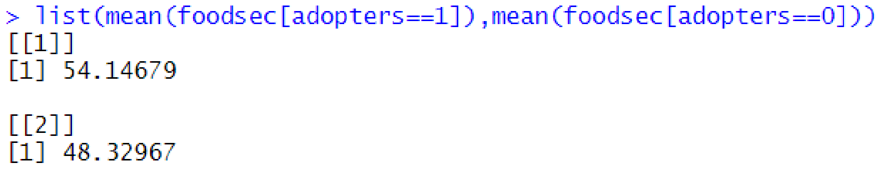
```
> summary(data_chapter2)
    adopters         foodsec     
 Min.   :0.000   Min.   :28.00  
 1st Qu.:0.000   1st Qu.:39.75  
 Median :1.000   Median :52.00  
 Mean   :0.545   Mean   :51.50  
 3rd Qu.:1.000   3rd Qu.:63.00  
 Max.   :1.000   Max.   :76.00  
```

In our sample, we have 200 observations on the status of technology adoption (0 = NON-ADOPTERS and 1 = ADOPTERS), with a relevant index of food security (foodsec) for each observation. We can further obtain frequency statistics for a variable by using the *table* command in R.



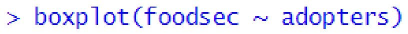
```
> table(adopters)
adopters
  0   1 
 91 109 
```

When one wants to list more than two results together, we can use the *list* command in R. The *list* command allows us to enumerate information. In this case, we list the mean of food security index of two groups: adopters and non-adopters. As a result, those who adopt the hybrid seed varieties have a higher food security index than non-adopters.

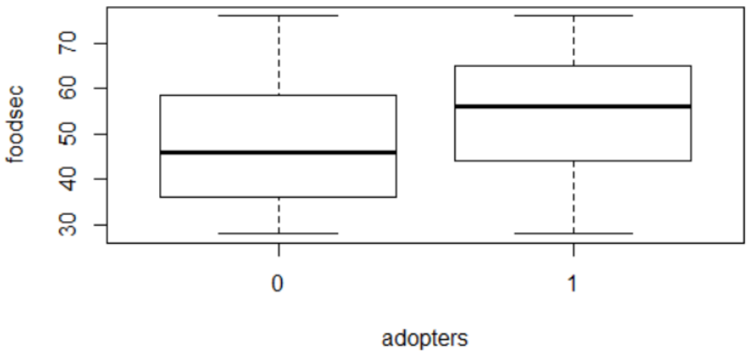


```
> list(mean(foodsec[adopters==1]),mean(foodsec[adopters==0]))
[[1]]
[1] 54.14679

[[2]]
[1] 48.32967
```

To describe the variables' information graphically, we can draw plots to examine the relationship between two variables. The *boxplot* command produces box-and-whisker plots of the given values. We compare the mean of food security index between the groups, which are adopters and non-adopters of hybrid seed varieties.

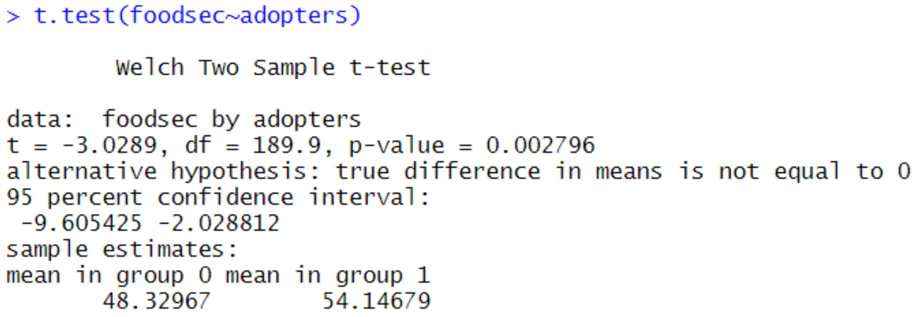


```
> boxplot(foodsec ~ adopters)
```



Independent Sample *t*-Test Assuming Unequal Variance

As mentioned above, we compare the mean food security index between the group of adopters and non-adopters of hybrid seed varieties. Ideally, these subjects are randomly selected from a larger population of subjects. The default of the *t*-test using the *t.test* command in R is that no equal variance, independent samples, and two-sided tests. We generate the results using the following command in R.

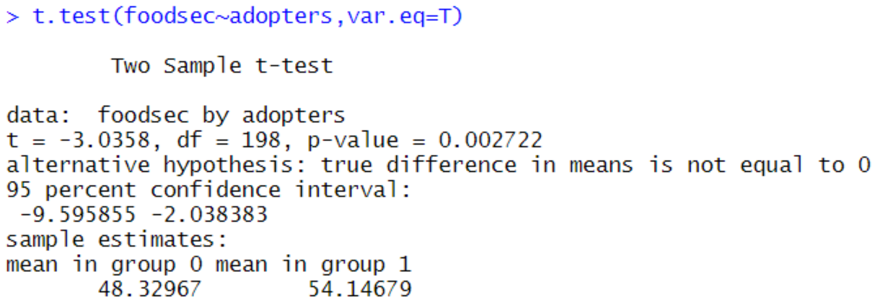


```
> t.test(foodsec~adopters)

        Welch Two Sample t-test

data:  foodsec by adopters
t = -3.0289, df = 189.9, p-value = 0.002796
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.605425 -2.028812
sample estimates:
mean in group 0 mean in group 1 
       48.32967        54.14679 
```

In this example, the t-statistics is -3.0289, with 189.9 degrees of freedom. The corresponding two-tailed p-value is 0.002796, which rejects the null hypothesis. So, we can that there means are the difference between the two groups. It interprets that food security between these two groups is statistically different and we conclude that hybrid maize adopters are better off in terms of food security.

Independent Group *t*-Test Assuming Equal Variance

We compare the mean food security index between the groups of adopters and non-adopters of hybrid seed varieties with equal variance conditions. We conduct the *t*-test, assuming that variances for the two populations are the same. Since the default of the *t*-test in R assumes unequal variances, we need to add the argument *var.eq* = *T* to conduct the t-test under the same variance conditions. We generate the results using the following command in R.



```
> t.test(foodsec~adopters,var.eq=T)

        Two Sample t-test

data:  foodsec by adopters
t = -3.0358, df = 198, p-value = 0.002722
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.595855 -2.038383
sample estimates:
mean in group 0 mean in group 1 
       48.32967        54.14679 
```

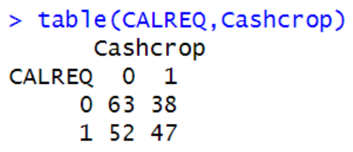
In this example, the t-statistics is -3.0358, with 198 degrees of freedom. The corresponding two-tailed p-value is 0.002722, which rejects the null hypothesis. So, we can that there means are the difference between the two groups. It interprets that food security between these two groups is statistically different, and we conclude that hybrid maize adopters are better off in terms of food security.

Chapter 3
Effects of Commercialization of Agriculture (Shift from Traditional Crop to Cash Crop) on Food consumption and Nutrition – Application of Chi-Square Statistic

DESCRIPTIVE ANALYSIS: CROSS-TABULATION RESULTS (Please find the data used for this chapter here)

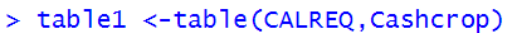
In this section, we perform Chi-Square tests using R for a sample of 200 farmers, which was randomly designated to conduct Chi-Square tests. (Variables: Cashcrop, INSECURE, CALREQ, ZHANEW, ZWANEW, ZWHNEW). First, we investigate the relationship between CASHCROP, which is the independent variable, and the two food security measures(CALREQ and INSECURE) by using the cross-tabulation results. The variable is defined as CASHCROP = 1 if the household grows at least one of four major cash crops, which are tobacco, groundnuts, cotton, and plantain and 0 otherwise. CALREQ is defined as households being able to satisfy at least 80% of the requirement for calorie intake that is 2200kcal. If a household intake more than 80 % of calories, we can interpret the household is qualified as "food secure." The variable is defined as CALREQ = 1, if the household is food secure and CALREQ = 0, otherwise.

In order to conduct the cross-tabulation in R, we can use *table* command with the variables that we would like to examine. The *Table(A, B)* in R is for a two-way frequency table. A will be rows, and B will be columns. The result of table command is as below:



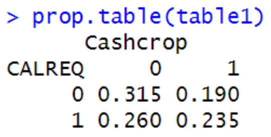
```
> table(CALREQ,Cashcrop)
      Cashcrop
CALREQ  0  1
     0 63 38
     1 52 47
```

After conducting the *table* command, we save the results as table 1 for further analysis to calculate shares.

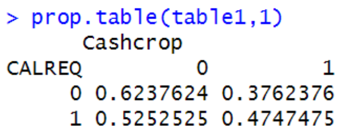


```
> table1 <-table(CALREQ,Cashcrop)
```

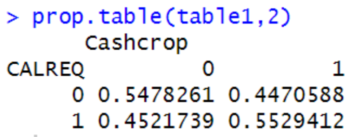
Once we save the table 1, we can examine probability tables on cell percentage (using *prop.table (table1)*), row percentage (using *prop.table (table1,1)*), and column percentage (using *prop.table (table1,2)*) arguments.



```
> prop.table(table1)
      Cashcrop
CALREQ     0     1
     0 0.315 0.190
     1 0.260 0.235
```

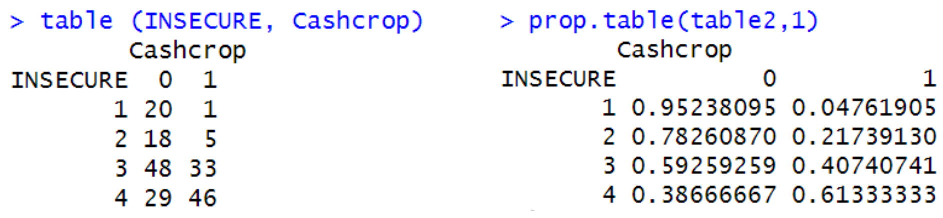


```
> prop.table(table1,1)
      Cashcrop
CALREQ         0         1
     0 0.6237624 0.3762376
     1 0.5252525 0.4747475
```



```
> prop.table(table1,2)
      Cashcrop
CALREQ         0         1
     0 0.5478261 0.4470588
     1 0.4521739 0.5529412
```

The cross-tabulation results show the probability of variable combinations on three different types of the denominator, which are the cell, row, and column from the left side. When we examine the results in the middle, the cross-tabulation results can interpret that 37.62% of the cash crop growing households are food insecure. It is likely to interpret that growing cash crop generates additional income for the household, who can sell these crops at local markets. Additional income, in turn, allows them to purchase more food. Next, we investigate the relationship between CASHCROP and INSECURE (the second measure of food security). We employ the *table* command to look at the frequency of two variables (INSECURE, Cashcrop) and then save the results as table 2. After, we calculated shares on rows to see portions of cash crop growing on each insecure level.



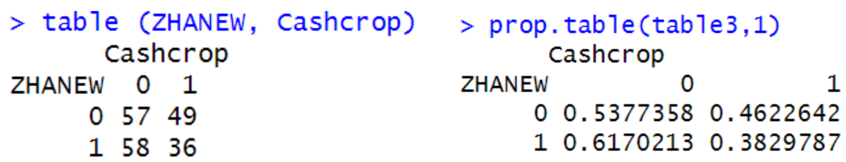
```
> table (INSECURE, Cashcrop)
        Cashcrop
INSECURE  0  1
       1 20  1
       2 18  5
       3 48 33
       4 29 46
```

```
> prop.table(table2,1)
        Cashcrop
INSECURE          0          1
       1 0.95238095 0.04761905
       2 0.78260870 0.21739130
       3 0.59259259 0.40740741
       4 0.38666667 0.61333333
```

table2: Results of table(INSECURE, Cashcrop)

INSECURE variable is the other variable to show food security. INSECURE = 4, if food security is secure, INSECURE = 3, if food secure is moderately insecure, INSECURE = 2, if food secure is highly insecure, INSECURE = 1, if food secure is totally insecure. From the results, it explains that cash crop growers are relatively more food secure compared to non-cash crop growers.

Next, we want to investigate whether cash crop production results in achieving higher nutritional levels for the children as well as increased food security for the household members. We undertake cross-tabulation texts for CASHCROP and ZHANEW (0 = low, 1 = normal), ZWANEW (0 = low, 1 = normal), and ZWHNEW (0 = low, 1 = normal). All the above variables are dichotomous nominal variables. The hypothesis is that there is no relationship between commercialization (CASHCROP) and child nutrition, as measured by the above indicators. The next parts show the relationship between them.



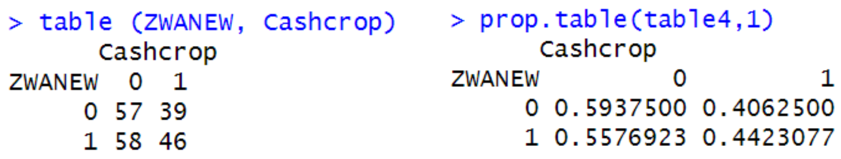
```
> table (ZHANEW, Cashcrop)
      Cashcrop
ZHANEW  0  1
     0 57 49
     1 58 36
```

```
> prop.table(table3,1)
      Cashcrop
ZHANEW         0         1
     0 0.5377358 0.4622642
     1 0.6170213 0.3829787
```

Table3: Results of table(ZHANEW, Cashcrop)

The cross-tabulation results indicate that 53% of preschoolers of the households not growing cash crops are stunted, while 46.2% of preschoolers for households growing cash crops are stunted. It is likely that the extra income generated through sale of cash crops achieves greater income, which helps in moderating food insecurity of the household.

The incidence of underweight preschoolers was 59.4% for households who did not grow cash crops and 40.6% for households who grew cash crops. The results indicate that underweight children are less likely to occur in cash crop growing households relative to non-cash crop growing households.

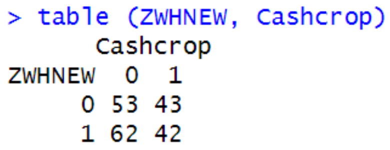


```
> table (ZWANEW, Cashcrop)
      Cashcrop
ZWANEW  0  1
     0 57 39
     1 58 46
```

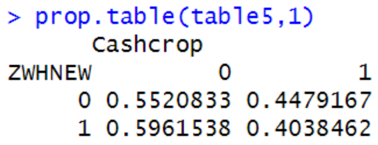
```
> prop.table(table4,1)
      Cashcrop
ZWANEW         0         1
     0 0.5937500 0.4062500
     1 0.5576923 0.4423077
```

Table4: Results of table(ZWANEW, Cashcrop)

The cross-tabulation results of cash crop production and wasting are below. The results present that households who grew cash crops had a lesser incidence of wasting (44.8%) compared to households who did not grow the crops (55.2%).



```
> table (ZWHNEW, Cashcrop)
      Cashcrop
ZWHNEW  0  1
     0 53 43
     1 62 42
```



```
> prop.table(table5,1)
      Cashcrop
ZWHNEW         0         1
     0 0.5520833 0.4479167
     1 0.5961538 0.4038462
```

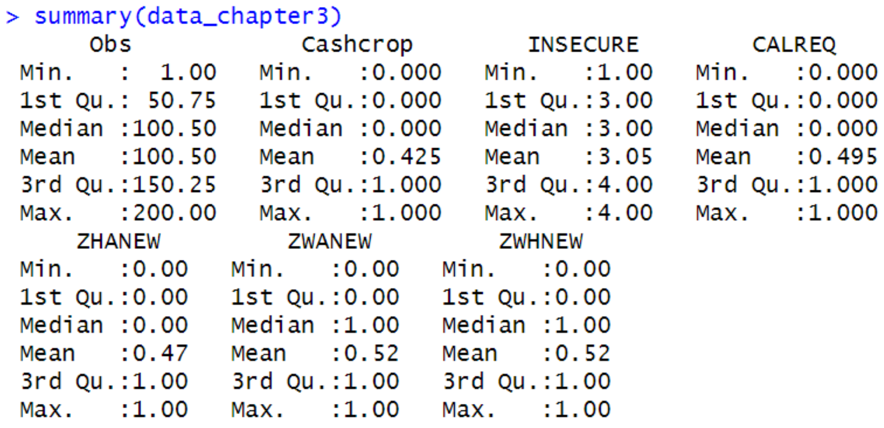
Table5: Results of table (ZWHNEW, Cashcrop)

However, from the above results, we cannot conclude if the results were significant or only due to random variability. Thus, we undertake chi-square tests to determine this.

CHI-SQUARE TESTS USING R

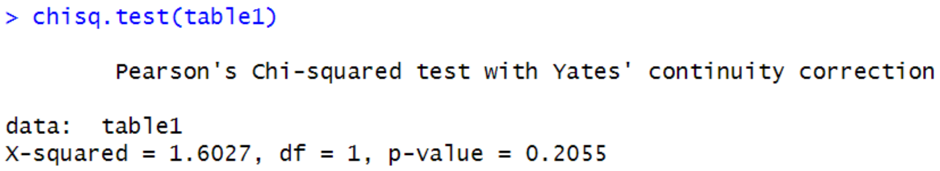
In this section, we perform Chi-Square tests on the dataset with farmers, some of whom grow cash crops. We first examine the data using the *summary* command in R.

The null hypothesis is given by H_0: no relationship exists between growing cash crops and food security. In other words, the incidence of observed food insecurity among households is not statistically different between cash crop growers and non-cash crop growers.



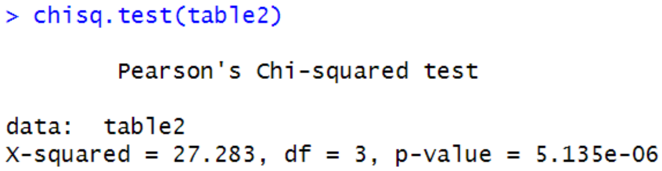
```
> summary(data_chapter3)
      Obs            Cashcrop        INSECURE        CALREQ     
 Min.   :  1.00   Min.   :0.000   Min.   :1.00   Min.   :0.000  
 1st Qu.: 50.75   1st Qu.:0.000   1st Qu.:3.00   1st Qu.:0.000  
 Median :100.50   Median :0.000   Median :3.00   Median :0.000  
 Mean   :100.50   Mean   :0.425   Mean   :3.05   Mean   :0.495  
 3rd Qu.:150.25   3rd Qu.:1.000   3rd Qu.:4.00   3rd Qu.:1.000  
 Max.   :200.00   Max.   :1.000   Max.   :4.00   Max.   :1.000  
     ZHANEW          ZWANEW          ZWHNEW    
 Min.   :0.00   Min.   :0.00   Min.   :0.00  
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00  
 Median :0.00   Median :1.00   Median :1.00  
 Mean   :0.47   Mean   :0.52   Mean   :0.52  
 3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.00  
 Max.   :1.00   Max.   :1.00   Max.   :1.00  
```

With saved tables 1 through 5 (table1 – CALREQ and Cashcrop, table2 – INSECURE and Cashcrop, table3 – ZHANEW and Cashcrop, table4 – ZWANEW and Cashcrop, table5 – ZWHNEW and Cashcrop), we can conduct chi-square tests by using the *chisq.test* command in R. The results of table1 is as below. The degree of freedom is 1, a Pearson Chi-Square value of 1.60, and the *p*-value is 0.2055. According to the *p*-value, our arbitrary dataset cannot reject the null hypothesis, that there is no relationship in food security between Cash crop and CALREQ.



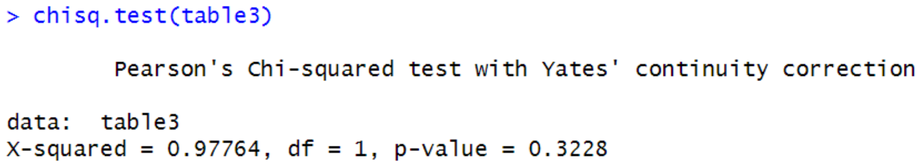
```
> chisq.test(table1)

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

data:  table1
X-squared = 1.6027, df = 1, p-value = 0.2055
```

The results of table2 (Cash crop and INSECURE) are as below. The degree of freedom is 3, a Pearson Chi-Square value of 27.283, and the *p*-value is close to zero. According to the *p*-value, we can reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship in food security and cash crop growing. We also want to investigate whether cash crop production results in achieving higher nutritional levels as well as increased food security in the data. We undertake Chi-Square tests for Cashcrop and other variables.



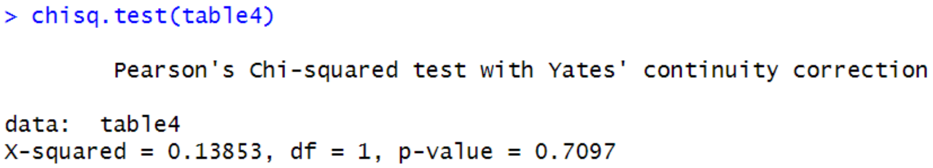
```
> chisq.test(table2)

        Pearson's Chi-squared test

data:  table2
X-squared = 27.283, df = 3, p-value = 5.135e-06
```

The Chi-Square results of table3 is below. The degree of freedom is 1, a Pearson Chi-Square value of 0.978, and the *p*-value is 0.32. Our dataset cannot reject the null hypothesis, that there is no relationship in food security between ZHANEW and two types of croppers.



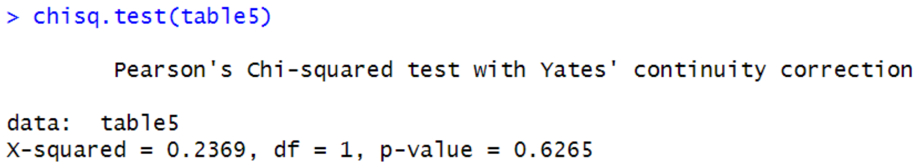
```
> chisq.test(table3)

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

data:  table3
X-squared = 0.97764, df = 1, p-value = 0.3228
```

The Chi-Square results of table4 in R is below. A Pearson Chi-Square value of 0.139 and the *p*-value is 0.71. Our dataset cannot reject the null hypothesis, that there is no relationship in food security between ZWANEW and two types of croppers.



```
> chisq.test(table4)

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

data:  table4
X-squared = 0.13853, df = 1, p-value = 0.7097
```

Lastly, there are the results of the Chi-Square Tests below. A Pearson Chi-Square value of 0.237 and the *p*-value is 0.63. Our dataset cannot reject the null hypothesis, that there is no relationship in food security between ZWHNEW and two types of croppers.

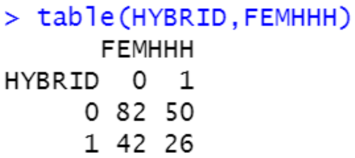


```
> chisq.test(table5)

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

data:  table5
X-squared = 0.2369, df = 1, p-value = 0.6265
```

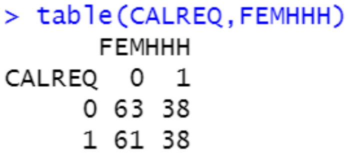
Chapter 4 Effects of Technology Adoption and Gender of Household Head: The issue, Its Importance in Food Security – Application of Cramer's V and Phi coefficient

(Please find the data used for this chapter here)

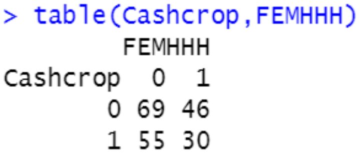
We again use the cross-tabulation procedure, which is used in chapter 3, to address the validity of different groups. In this chapter, Cramer's V and Phi test statistics will be introduced in addition to the cross-tabulation and chi-square test statistics. These two statistics have been created to measure the degree of association between any two nominal variables because Chi-square does not indicate how significant and important relation is. Cramer's V and phi coefficient are a post-test to give additional information. First, we examine three sets of frequencies between the two variables (HYBRID-FENHHH, CARLEQ-FEMHHH, and Cashcrop-FEMHHH) to examine if male- or female-headed households are more likely to commercialize crops and thereby receive higher income from the proceeds. We use the *table* and *summary* command for investigating descriptive statistics.



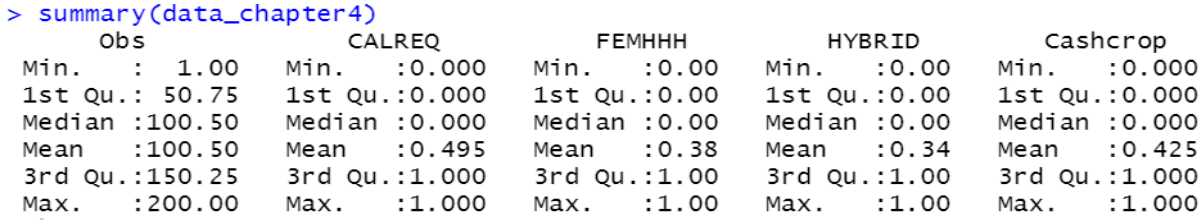
```
> table(HYBRID,FEMHHH)
      FEMHHH
HYBRID  0  1
     0 82 50
     1 42 26
```



```
> table(CALREQ,FEMHHH)
      FEMHHH
CALREQ  0  1
     0 63 38
     1 61 38
```

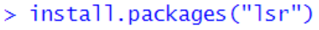


```
> table(Cashcrop,FEMHHH)
        FEMHHH
Cashcrop  0  1
       0 69 46
       1 55 30
```

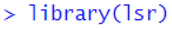


```
> summary(data_chapter4)
      Obs             CALREQ          FEMHHH          HYBRID          Cashcrop    
 Min.   :  1.00   Min.   :0.000   Min.   :0.00   Min.   :0.00   Min.   :0.000  
 1st Qu.: 50.75   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.000  
 Median :100.50   Median :0.000   Median :0.00   Median :0.00   Median :0.000  
 Mean   :100.50   Mean   :0.495   Mean   :0.38   Mean   :0.34   Mean   :0.425  
 3rd Qu.:150.25   3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.000  
 Max.   :200.00   Max.   :1.000   Max.   :1.00   Max.   :1.00   Max.   :1.000  
```

We can install the *lsr* packages to calculate the Cramer's V with the *CramersV* command. To install a package, we can use the *install.packages* command and load the downloaded package with the *library* command as below.



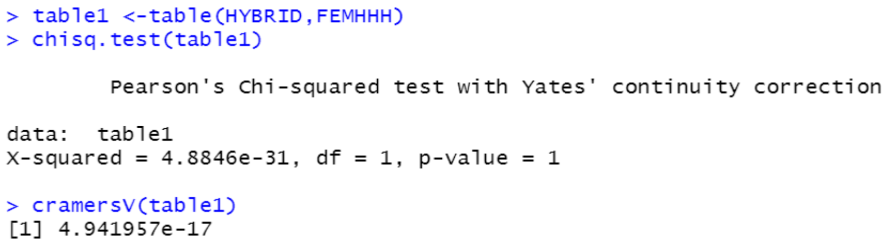
```
> install.packages("lsr")
```



```
> library(lsr)
```

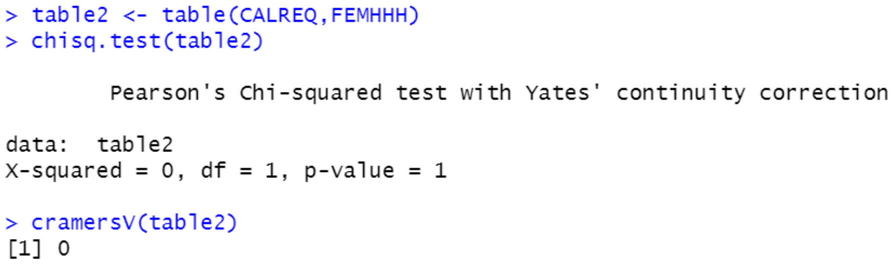
Since the Cramer's V and phi tests are all based on the chi-square statistic, the significant level is also based on the significance of the chi-square statistic. Firstly, we can check the significance through chi-square analysis, and then we can use the *CramersV* command to calculate the Cramer's V to identify how much two variables are related. From the test, we can examine if the relationship between the variables (the cross-tabulation results) is significant or if it is just due to random variability.

The first example is the relationship between HYBRID and FENHHH variables (table1). The null hypothesis is given by: H_0: incidences of hybrid maize adoption are not statistically different between the male- and female-headed households. The significance level from the chi-squared test is 1, which means the null hypothesis is not rejected. Thus, we conclude that the incidences of hybrid maize adoption are not statistically different between the male- and female-headed households. Since there are no significant differences between the two variables, we cannot interpret the Cramer’s V value.



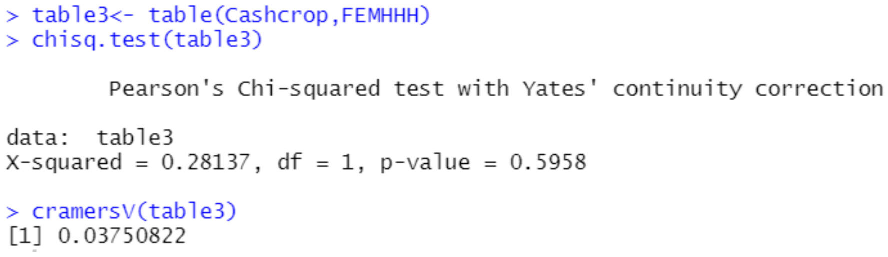
```
> table1 <-table(HYBRID,FEMHHH)
> chisq.test(table1)

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

data:  table1
X-squared = 4.8846e-31, df = 1, p-value = 1

> cramersV(table1)
[1] 4.941957e-17
```

The second example is the relationship between CALREQ and FENHHH variables (table2). The null hypothesis is given by: H_0: no relationship between food security and gender of the household head. The significance level from the chi-squared test is also 1, so we cannot reject the null hypothesis. We find no pattern of relationship emerging between the gender of the household head and food security for this sample.



```
> table2 <- table(CALREQ,FEMHHH)
> chisq.test(table2)

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

data:  table2
X-squared = 0, df = 1, p-value = 1

> cramersV(table2)
[1] 0
```

The last example is the relationship between Cashcrop and FENHHH variables (table3). The null hypothesis is given by: H_0: no relationship exists between cash crop growing and the gender of the household head. The significance level from the chi-squared test is 0.6, so we cannot reject the null hypothesis. Since there are no significant differences between the two variables, we cannot interpret the Cramer’s V value. However, if it is significant, we can interpret there is a 0.03 relationship within two variables.



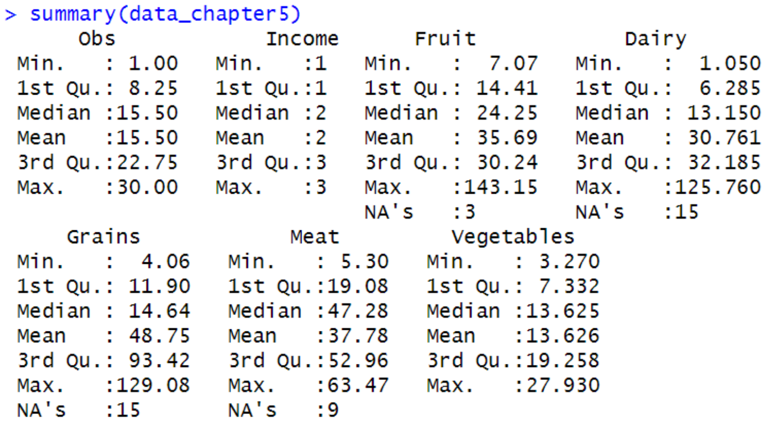
```
> table3<- table(Cashcrop,FEMHHH)
> chisq.test(table3)

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

data:  table3
X-squared = 0.28137, df = 1, p-value = 0.5958

> cramersV(table3)
[1] 0.03750822
```

Chapter 5 Changes in Food Consumption Patterns: Its Importance to Food Security – Application of One-Way ANOVA

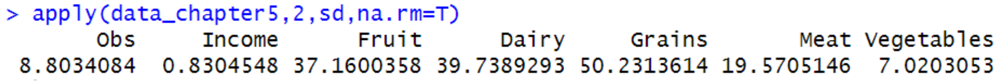
ONE-WAY ANOVA IN R (Please find the data used for this chapter here)

For illustrative purposes, we use the following 30 observations on Food Intakes (Converted to Retail Commodities Data) from the USDA and ERS estimates (www.ers.usda.gov). We have presented a sample from this data for different food categories (Fruit, Dairy, Grains, Meat, and Vegetables). We have classified the food intakes across three income groups: Low (Income = 1), Medium (Income = 2), and High (Income = 3). We wish to examine the food consumption patterns across income groups using the *F*-test from one-way ANOVA estimation in R:

We look at the descriptive statistics, which we can obtain, using the *summary* command in R**.** The R output confirms the information in the table by noting that there are missing observations for several of the food categories. Similarly, we use the *apply* command in R to print the standard deviations for the variables. From the results below, we find that the standard deviations are large across all food categories. We can undertake the cross-tabulations:



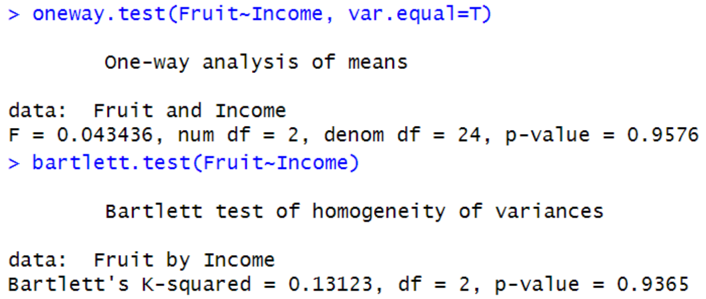
```
> summary(data_chapter5)
      Obs            Income       Fruit            Dairy        
 Min.   : 1.00   Min.   :1   Min.   :  7.07   Min.   :  1.050  
 1st Qu.: 8.25   1st Qu.:1   1st Qu.: 14.41   1st Qu.:  6.285  
 Median :15.50   Median :2   Median : 24.25   Median : 13.150  
 Mean   :15.50   Mean   :2   Mean   : 35.69   Mean   : 30.761  
 3rd Qu.:22.75   3rd Qu.:3   3rd Qu.: 30.24   3rd Qu.: 32.185  
 Max.   :30.00   Max.   :3   Max.   :143.15   Max.   :125.760  
                             NA's   :3        NA's   :15       
     Grains             Meat         Vegetables    
 Min.   :  4.06   Min.   : 5.30   Min.   : 3.270  
 1st Qu.: 11.90   1st Qu.:19.08   1st Qu.: 7.332  
 Median : 14.64   Median :47.28   Median :13.625  
 Mean   : 48.75   Mean   :37.78   Mean   :13.626  
 3rd Qu.: 93.42   3rd Qu.:52.96   3rd Qu.:19.258  
 Max.   :129.08   Max.   :63.47   Max.   :27.930  
 NA's   :15       NA's   :9                        
```



```
> apply(data_chapter5,2,sd,na.rm=T)
       Obs     Income      Fruit      Dairy     Grains       Meat Vegetables 
 8.8034084  0.8304548 37.1600358 39.7389293 50.2313614 19.5705146  7.0203053 
```

First, we undertake a one-way ANOVA estimation using the *oneway.test* command R, between income and fruit consumption. Note that the *oneway.test* command assumes to have a default of heteroscedasticity. Thus, in order to undertake one-way ANOVA, which assumes homoscedasticity, we need to add an argument: *var.equal=T.* This command also ignores the three missing observations and estimates the ANOVA for the 27 observations. We also find that the intake of fruits across income groups does not vary as we look at that the *F* value in the ANOVA results, which indicate that *F* = 0.04. Note that the table value of at 95% or *F*(2, 24) = 3.40. Consequently, we cannot reject the null hypothesis in this case, which states that the intake of fruit is identical for all three income groups.

The alternative hypothesis in this problem states that the intake of fruit is not identical across these income groups. In other words, the intake of fruits is identical across all income groups. In one sense, this finding is very comforting that the lower-income groups have the same intake as the higher income groups with respect to the calories from fruits. However, the flip side is that if the amount of consumption does not produce sufficient calories, then why do the higher income groups consume the same amount of fruits as the low-income groups? Either way, the ANOVA helps direct our attention to this critical issue concerning nutrient sufficiency.

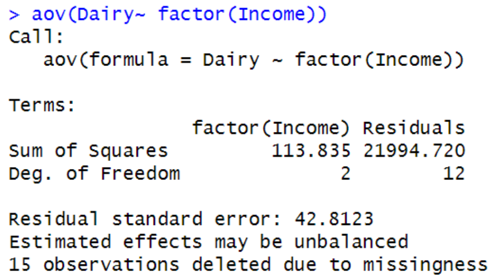


```
> oneway.test(Fruit~Income, var.equal=T)

        One-way analysis of means

data:  Fruit and Income
F = 0.043436, num df = 2, denom df = 24, p-value = 0.9576
> bartlett.test(Fruit~Income)

        Bartlett test of homogeneity of variances

data:  Fruit by Income
Bartlett's K-squared = 0.13123, df = 2, p-value = 0.9365
```

We also undertake the Bartlett's test of equal variances, which is approximated by the chi-square test statistic. Bartlett's test is used to test if *k* samples have equal variances. Equal variances across samples are also known as homogeneity of variances. Some statistical tests, for example, the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

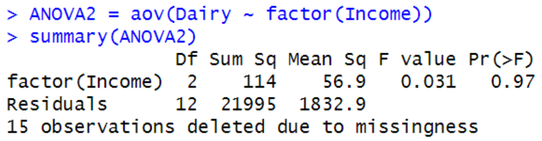
Bartlett's test is sensitive to departures from normality. That is, if samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. In our R output, we find that the calculated Bartlett's K-squared value is 0.13, which is smaller than 5.99 or 95% on the table value with 2 degrees of freedom. Consequently, we cannot reject the null hypothesis of equal variances.

Similarly, we use the above data to perform the one-way ANOVA estimation using the different command R (*aov)*, but the same one-way ANOVA result between income and dairy consumption:

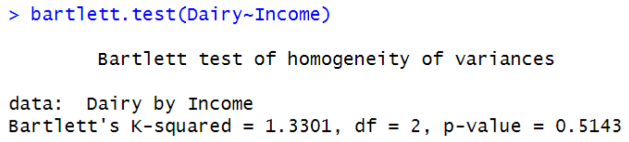


```
> aov(Dairy~ factor(Income))
Call:
   aov(formula = Dairy ~ factor(Income))

Terms:
                factor(Income) Residuals
Sum of Squares         113.835 21994.720
Deg. of Freedom              2        12

Residual standard error: 42.8123
Estimated effects may be unbalanced
15 observations deleted due to missingness
```

We should use *factor (Income)* instead of a variable name (Income) to categorize the value of the variable. Without *factor* argument, the variable will be considered as a continuous variable. In addition, keep in mind that *aov* command assumes homogeneous variance met. To examine in detail, we can save the results of one-way ANOVA as ANOVA2 and get more information by using *summary* command as follow. Note that one can investigate details of an analysis by using the *summary* command as well as descriptive statistics.



```
> ANOVA2 = aov(Dairy ~ factor(Income))
> summary(ANOVA2)
               Df Sum Sq Mean Sq F value Pr(>F)
factor(Income)  2    114    56.9   0.031   0.97
Residuals      12  21995  1832.9
15 observations deleted due to missingness
```



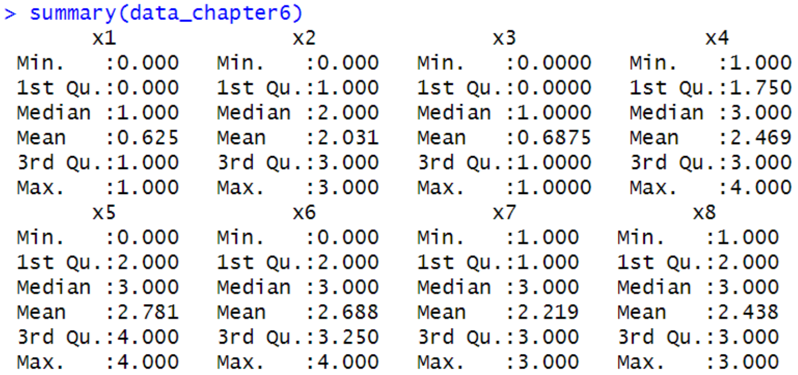
```
> bartlett.test(Dairy~Income)

        Bartlett test of homogeneity of variances

data:  Dairy by Income
Bartlett's K-squared = 1.3301, df = 2, p-value = 0.5143
```

To see if there are any group differences, we find that the F = 0.03 from the ANOVA portion of the R output. Note that the table value of at 95% or $F(2, 12)$ = 3.88. Also, the *p*-value is 0.97. Consequently, we cannot reject the null hypothesis in this case, which states that the intake of dairy is identical for all three income groups. We find that the calculated Bartlett’s K-squared value is 1.33, and the *p*-value is 0.51; thus, we cannot reject the null hypothesis of equal variances.

Chapter 6
Impact of Market Access on Food Security – Application of Factor Analysis

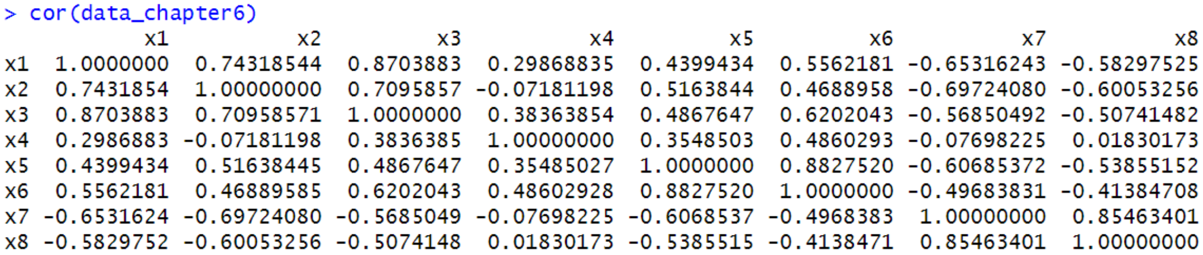
PRINCIPAL COMPONENTS ANALYSIS IN R (Please find the data used for this chapter here)

We motivate the implementation using a simple example with 32 observations and 8 variables, listed as *x*1, *x*2, … , *x*8, to apply factor analysis to the data. We begin to perform the summary and correlation table to examine the characteristics of the data. The table shows the minimum, mean, max, and quantile values for each variable.

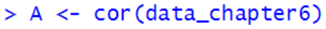


```
> summary(data_chapter6)
       x1              x2              x3               x4       
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :1.000  
 1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.750  
 Median :1.000   Median :2.000   Median :1.0000   Median :3.000  
 Mean   :0.625   Mean   :2.031   Mean   :0.6875   Mean   :2.469  
 3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :1.000   Max.   :3.000   Max.   :1.0000   Max.   :4.000  
       x5              x6              x7              x8       
 Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :2.781   Mean   :2.688   Mean   :2.219   Mean   :2.438  
 3rd Qu.:4.000   3rd Qu.:3.250   3rd Qu.:3.000   3rd Qu.:3.000  
 Max.   :4.000   Max.   :4.000   Max.   :3.000   Max.   :3.000  
```

The *cor* command in R produces the correlation table. The correlation coefficient is shown on the table. We save the results of *cor* command as A to conduct Kaiser-Meyer-Olkin Measure (KMO) analysis for later.

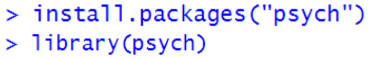


```
> cor(data_chapter6)
           x1          x2         x3          x4         x5         x6          x7          x8
x1  1.0000000  0.74318544  0.8703883  0.29868835  0.4399434  0.5562181 -0.65316243 -0.58297525
x2  0.7431854  1.00000000  0.7095857 -0.07181198  0.5163844  0.4688958 -0.69724080 -0.60053256
x3  0.8703883  0.70958571  1.0000000  0.38363854  0.4867647  0.6202043 -0.56850492 -0.50741482
x4  0.2986883 -0.07181198  0.3836385  1.00000000  0.3548503  0.4860293 -0.07698225  0.01830173
x5  0.4399434  0.51638445  0.4867647  0.35485027  1.0000000  0.8827520 -0.60685372 -0.53855152
x6  0.5562181  0.46889585  0.6202043  0.48602928  0.8827520  1.0000000 -0.49683831 -0.41384708
x7 -0.6531624 -0.69724080 -0.5685049 -0.07698225 -0.6068537 -0.4968383  1.00000000  0.85463401
x8 -0.5829752 -0.60053256 -0.5074148  0.01830173 -0.5385515 -0.4138471  0.85463401  1.00000000
```



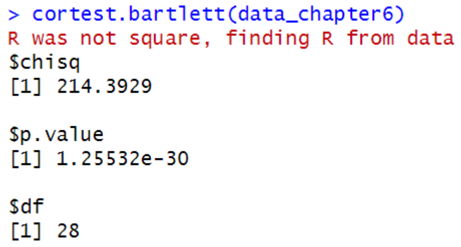
```
> A <- cor(data_chapter6)
```

First, we should check the test of Sampling Adequacy by conducting a Bartlett test of sphericity and Kaiser-Meyer-Olkin Measure of factor adequacy. We employ packages to simplify the analysis rather than calculating line by line. The package we used here is called *'psych.'* You can download the package by coding *install.packages(psych)* and register with *library(psych)* command as bellow.



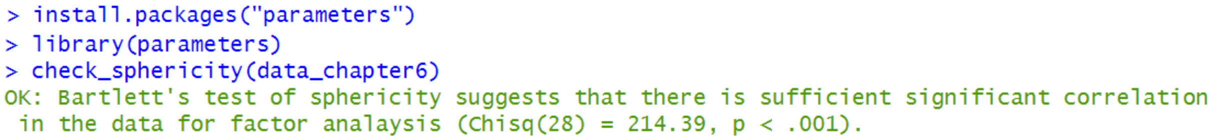
```
> install.packages("psych")
> library(psych)
```

The *cortest.bartlett* command can be used to undertake a Bartlett test of sphericity with *'psych'* packages. R's output indicates 214.39 of chi-squared value with the degree of freedom 28, and a p-value of smaller than 0.01, which means that we can reject the null hypothesis and procced with factor analysis.



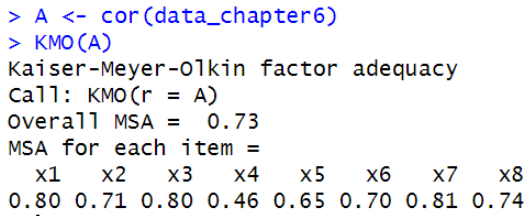
```
> cortest.bartlett(data_chapter6)
R was not square, finding R from data
$chisq
[1] 214.3929

$p.value
[1] 1.25532e-30

$df
[1] 28
```

Another package called *'parameters'* can be used as well. In order to employ the packages, install and load the downloaded packages by using the *install.packages* and *library* commands. The command to produce the Bartlett test of sphericity is *check_sphericity*. The results from two different commands are the same. Both Bartlett's test of sphericity indicates that there is sufficient significant correlation in the data for factor analysis.



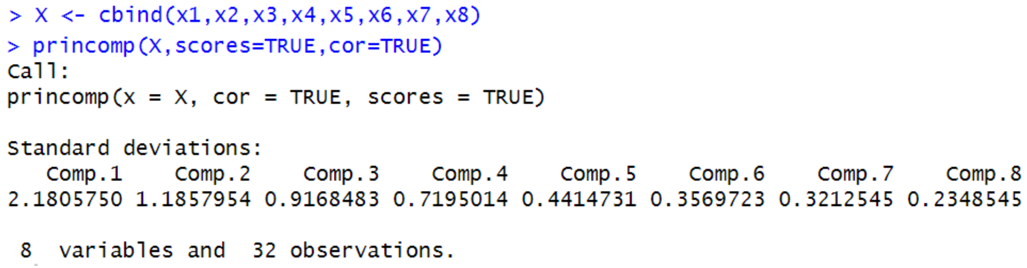
```
> install.packages("parameters")
> library(parameters)
> check_sphericity(data_chapter6)
OK: Bartlett's test of sphericity suggests that there is sufficient significant correlation
 in the data for factor analysis (Chisq(28) = 214.39, p < .001).
```

As mentioned above, the result of the correlation table of the data is saved as A to conduct Kaiser-Meyer-Olkin factor adequacy. The results present that the overall MSA (Measure of Sampling Adequacy) is 0.73, which is far above 0.5[1]. From both tests of Sampling Adequacy, we met a prerequisite to proceed with the factor analysis.



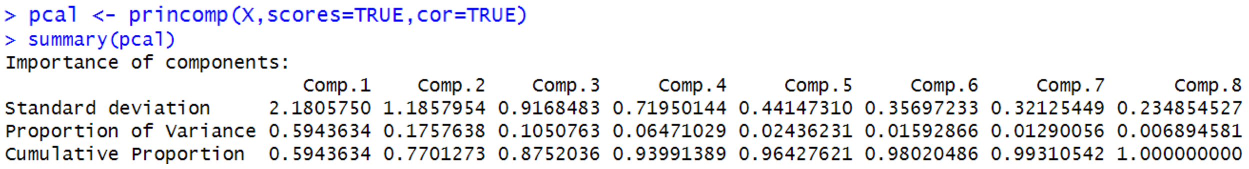
```
> A <- cor(data_chapter6)
> KMO(A)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = A)
Overall MSA =  0.73
MSA for each item =
  x1   x2   x3   x4   x5   x6   x7   x8
0.80 0.71 0.80 0.46 0.65 0.70 0.81 0.74
```

We implement the principal components analysis by using the *princomp* command in R. Before producing the principal components analysis, we register x as a combination of all variables from *x*1 to *x*8. We then use scores and cor arguments to conduct the principal components analysis. The results indicate that only the first two components have eigenvalues greater than 1. To examine closer, one can save the results of principal components analysis, then use the *summary* command to investigate the results in detail.



```
> X <- cbind(x1,x2,x3,x4,x5,x6,x7,x8)
> princomp(X,scores=TRUE,cor=TRUE)
Call:
princomp(x = X, cor = TRUE, scores = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8 
2.1805750 1.1857954 0.9168483 0.7195014 0.4414731 0.3569723 0.3212545 0.2348545 

 8  variables and  32 observations.
```

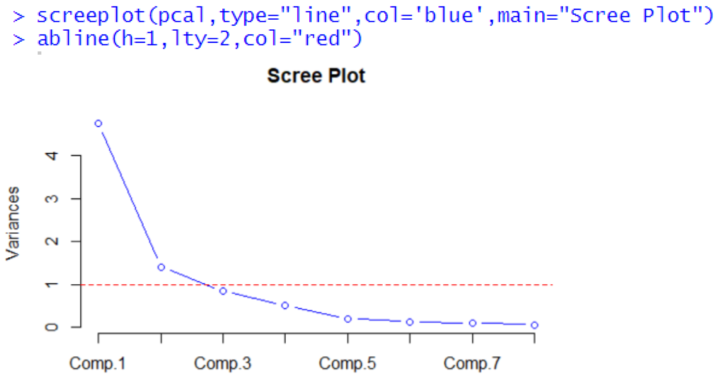
[1] KMO measures below 0.5 are unacceptable. For more information of KMO measure, see https://www.rdocumentation.org/packages/psych/versions/1.8.12/topics/KMO



```
> pcal <- princomp(X,scores=TRUE,cor=TRUE)
> summary(pcal)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7      Comp.8
Standard deviation     2.1805750 1.1857954 0.9168483 0.71950144 0.44147310 0.35697233 0.32125449 0.234854527
Proportion of Variance 0.5943634 0.1757638 0.1050763 0.06471029 0.02436231 0.01592866 0.01290056 0.006894581
Cumulative Proportion  0.5943634 0.7701273 0.8752036 0.93991389 0.96427621 0.98020486 0.99310542 1.000000000
```

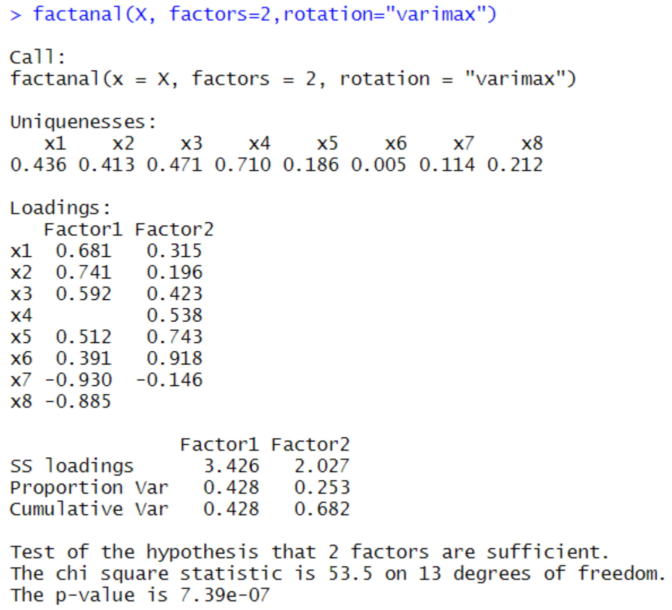
As a result, there are standard deviation, proportion of variance, and cumulative proportion. These first two components explain 77.% of the combined variance in the eight variables. The rest principal components may be dropped from subsequent analysis.

To illustrate with graphic, the command the *screeplot* produces the scree plot along with the eigenvalues by factor number. The next command the *abline* produces a horizontal red line parallel to the *x*-axis, where the eigenvalue (variances) = 1. We can examine that there are only two components which have greater variance values than 1. Thus, we have to retain the first two components.



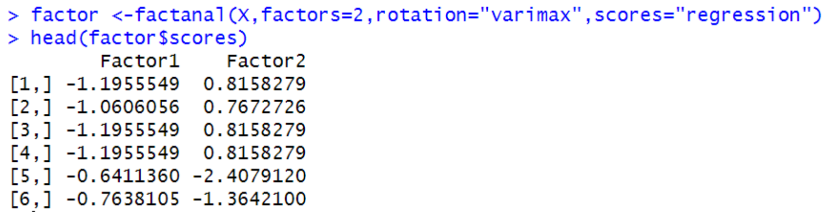
```
> screeplot(pcal,type="line",col='blue',main="Scree Plot")
> abline(h=1,lty=2,col="red")
```

We can produce the factor analysis results by using the *factanal* command in R. In the bracket, X is a formula or a numeric matrix or an object that can be coerced to a numeric matrix. It is the variable matrix in this case. We identify how many factors should include conducting factor analysis above. The eigenvalue of components that are greater than 1 is two; thus, we include *factors=2* argument. We want to use the varimax, which is one common rotational method under the orthogonal procedure for rotation, so it is also included as an argument, *rotation= "varimax."* Note that x1, x2, x3, x7, and x8 load heavily with the first factor, while the others load with the second factor.



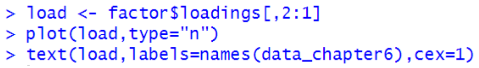
```
> factanal(X, factors=2,rotation="varimax")

Call:
factanal(x = X, factors = 2, rotation = "varimax")

Uniquenesses:
   x1    x2    x3    x4    x5    x6    x7    x8 
0.436 0.413 0.471 0.710 0.186 0.005 0.114 0.212 

Loadings:
   Factor1 Factor2
x1  0.681   0.315 
x2  0.741   0.196 
x3  0.592   0.423 
x4          0.538 
x5  0.512   0.743 
x6  0.391   0.918 
x7 -0.930  -0.146 
x8 -0.885         

               Factor1 Factor2
SS loadings      3.426   2.027
Proportion Var   0.428   0.253
Cumulative Var   0.428   0.682

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 53.5 on 13 degrees of freedom.
The p-value is 7.39e-07 
```

The last argument *score="regression"* produces the factor scores, which are the linear commonalities, and then weighting them with factor score coefficients and summing for each factor. The output for the last command line is given below. With the *head* command, we can print only six rows. However, there are indeed 32 observations, so 32 scores for both factors exist on the data. Examining the factor scores, we can see that *x*1 is -1.20 standard deviations below on the factor 1 dimension and 0.81standard deviations above factor 2 dimension.

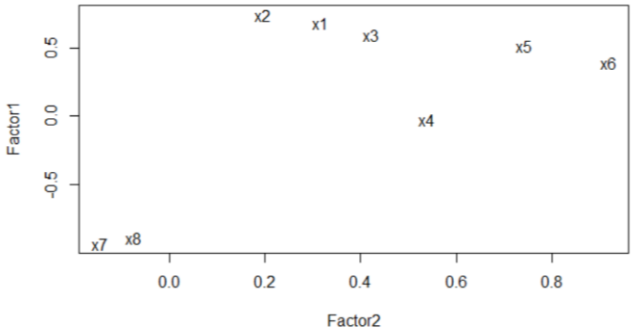


```
> factor <-factanal(X,factors=2,rotation="varimax",scores="regression")
> head(factor$scores)
        Factor1    Factor2
[1,] -1.1955549  0.8158279
[2,] -1.0606056  0.7672726
[3,] -1.1955549  0.8158279
[4,] -1.1955549  0.8158279
[5,] -0.6411360 -2.4079120
[6,] -0.7638105 -1.3642100
```

We produce a clearer picture of how the variables can be classified based on the components. To plot the figure with the loadings, we write the following commands in R.



```
> load <- factor$loadings[,2:1]
> plot(load,type="n")
> text(load,labels=names(data_chapter6),cex=1)
```



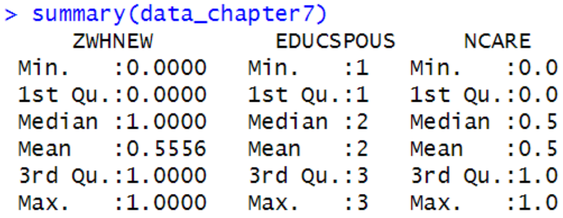
Chapter 7 Impact of Maternal Education and Care on Preschoolers' Nutrition – Application of Two-Way ANOVA

(Please find the data used for this chapter here)

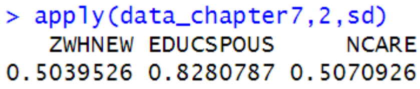
For illustrative purposes, we used the following 36 observations given in Table 7.8 on ZWHNEW (weight for height Z-scores), EDUCSPOUS (Education of the spouse), NCARE (Child-care index), where all these are categorical variables.

We look at the descriptive statistics, which we can gain, using the *summary* command in R. Together, we also obtain the two-way ANOVA table along with all the pairwise comparisons using the following command in R.

The *summary* command produces vital descriptive statistics. We also undertake the *apply* command to get the information of standard deviation since the *summary* command in R does not show it automatically.

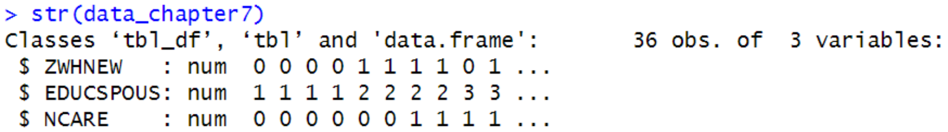


```
> summary(data_chapter7)
     ZWHNEW          EDUCSPOUS      NCARE    
 Min.   :0.0000   Min.   :1    Min.   :0.0  
 1st Qu.:0.0000   1st Qu.:1    1st Qu.:0.0  
 Median :1.0000   Median :2    Median :0.5  
 Mean   :0.5556   Mean   :2    Mean   :0.5  
 3rd Qu.:1.0000   3rd Qu.:3    3rd Qu.:1.0  
 Max.   :1.0000   Max.   :3    Max.   :1.0  
```



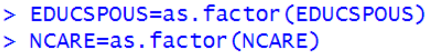
```
> apply(data_chapter7,2,sd)
   ZWHNEW EDUCSPOUS     NCARE 
0.5039526 0.8280787 0.5070926 
```

The *aov* command that is the one used in the previous section for the One-Way ANOVA is undertaken to perform the Two-Way ANOVA as well. Note that it should include the different formats of an outcome variable and independent variables. Before executing the *aov* command, we should manipulate numeric variables to factor ones because the numeric variable cannot be recognized as a categorical variable, which is EDUCSPOUR and NCARE. First, check the type of variables by using the *str* command. We can see that all variables are considered as numeric variables.

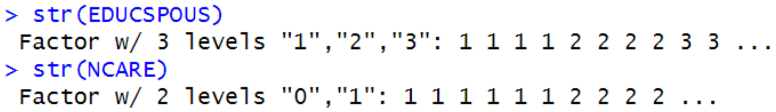


```
> str(data_chapter7)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	36 obs. of  3 variables:
 $ ZWHNEW   : num  0 0 0 0 1 1 1 1 0 1 ...
 $ EDUCSPOUS: num  1 1 1 1 2 2 2 2 3 3 ...
 $ NCARE    : num  0 0 0 0 0 0 1 1 1 1 ...
```

We use the *as.factor* command to change the type of independent variables from numeric to factor. Then, recheck the variables by using the *str* command for two independent variables which are EDUCSPOUS and NCARE

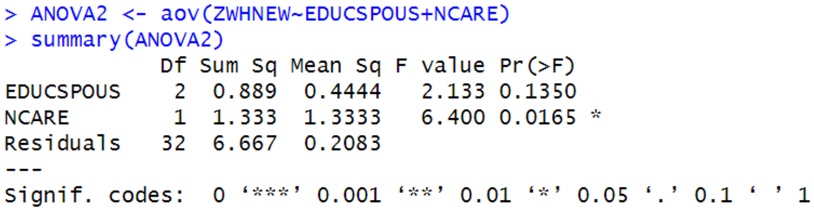


```
> EDUCSPOUS=as.factor(EDUCSPOUS)
> NCARE=as.factor(NCARE)
```



```
> str(EDUCSPOUS)
 Factor w/ 3 levels "1","2","3": 1 1 1 1 2 2 2 2 3 3 ...
> str(NCARE)
 Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 2 2 ...
```

Now we are ready to undertake the Two-Way ANOVA using the *aov* command. In order to perform the Two-Way ANOVA calculations, assuming ZWHNEW as the dependent variable and EDUCSPOUS and NCARE as the independent variables, type the command: *aov(ZWHNEW~EDUCSPOUS+NCARE)*. As we described in the earlier section, saving the results and using the *summary* command on the results presents more information.

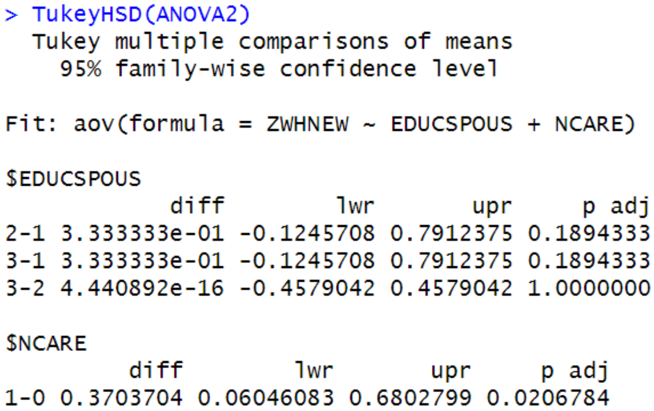


```
> ANOVA2 <- aov(ZWHNEW~EDUCSPOUS+NCARE)
> summary(ANOVA2)
            Df Sum Sq Mean Sq F value Pr(>F)  
EDUCSPOUS    2  0.889  0.4444   2.133 0.1350  
NCARE        1  1.333  1.3333   6.400 0.0165 *
Residuals   32  6.667  0.2083                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The null hypothesis of the ANOVA is that the independent variable has no significant impact on the outcome variable. First, we perform the Two-Way ANOVA without an interaction variable. For the main effect of mothers’ education on weight for height *Z*-score, the results show that *F*-value of 2.1 with the associated significance level of 0.14. We cannot reject the null hypothesis that mean weight for height Z-scores does not differ by educational levels of the mother since the p-value is higher than 0.05. Therefore, the educational of the mother does not have a significant influence on weight for height Z-scores.

The other main effect of child-care on the dependent variable, we find the *F*-value to be 6.4, with the corresponding probability of 0.05. We can reject the null hypothesis at 95% significance, indicating there is a significant impact of child-care levels on the weight for height Z-score. In other words, the mean weight for height Z-scores does not differ by mothers’ educational levels but differ by child-care levels.

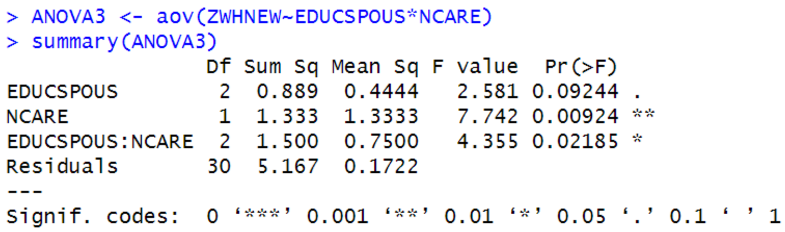
Since there are more than two means, we want to determine a post-hoc comparison, or a Tukey test, to determine which means are significantly different among combinations of education levels and child-care levels. We carry out an example by using the *TukeyHSD* command to generate the Tukey *t*-values:



```
> TukeyHSD(ANOVA2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ZWHNEW ~ EDUCSPOUS + NCARE)

$EDUCSPOUS
            diff        lwr       upr     p adj
2-1 3.333333e-01 -0.1245708 0.7912375 0.1894333
3-1 3.333333e-01 -0.1245708 0.7912375 0.1894333
3-2 4.440892e-16 -0.4579042 0.4579042 1.0000000

$NCARE
         diff        lwr       upr     p adj
1-0 0.3703704 0.06046083 0.6802799 0.0206784
```

From the results of Tukey’s test, we can conclude that for the different levels of education, whether mothers received higher or lower education, there are no significant impacts on improving nutritional status.

We now use the same data from above and incorporate the interaction term EDUCSPOUSE*NCARE. To include an interaction term, we can manipulate the independent variables by using *. The results of the ANOVA with the interaction terms are given below:



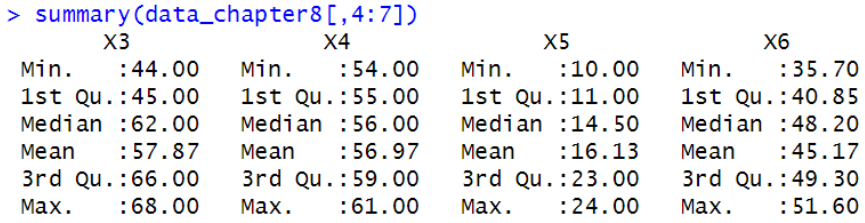
```
> ANOVA3 <- aov(ZWHNEW~EDUCSPOUS*NCARE)
> summary(ANOVA3)
                Df Sum Sq Mean Sq F value  Pr(>F)   
EDUCSPOUS        2  0.889  0.4444   2.581 0.09244 . 
NCARE            1  1.333  1.3333   7.742 0.00924 **
EDUCSPOUS:NCARE  2  1.500  0.7500   4.355 0.02185 * 
Residuals       30  5.167  0.1722                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the results, the p-values decrease in general so that we can reject the null hypothesis of all independent variables regarding p-value. For example, the mother’s education level has an impact on weight for height at a significant level of 90%. Child-care also has a significant effect on 99% significance. The inclusion of the interaction term indicates that maternal education is essential when it is taken in conjunction with NCARE.

Chapter 8 Indicators and Causal Factors of Nutrition – Application of Correlation Analysis

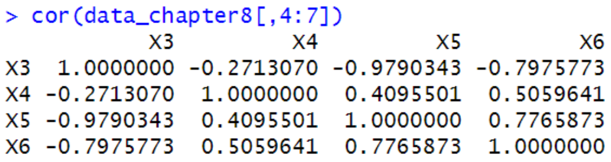
ESTIMATING CORRELATION USING R (Please find the data used for this chapter here)

In this section, we illustrate correlation analysis in R. We use a dataset from the exercise part of chapter 8, consisting of 30 observations and seven variables. Before we implement the correlation analysis, we produce descriptive statistics using the *summary* command for four variables: X3, X4, X5, and X6. To attain descriptive statistics for specific variables, we can add *[,4:7]* to print the results only from the fourth column to the seventh column. The result of the command is as below.



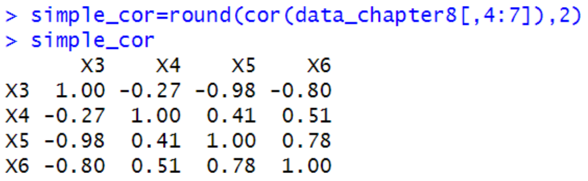
```
> summary(data_chapter8[,4:7])
       X3               X4               X5               X6       
 Min.   :44.00   Min.   :54.00   Min.   :10.00   Min.   :35.70  
 1st Qu.:45.00   1st Qu.:55.00   1st Qu.:11.00   1st Qu.:40.85  
 Median :62.00   Median :56.00   Median :14.50   Median :48.20  
 Mean   :57.87   Mean   :56.97   Mean   :16.13   Mean   :45.17  
 3rd Qu.:66.00   3rd Qu.:59.00   3rd Qu.:23.00   3rd Qu.:49.30  
 Max.   :68.00   Max.   :61.00   Max.   :24.00   Max.   :51.60  
```

Now we can obtain the correlation table for the variables of interest by using *cor* command in R. We add *[,4:7]* to present correlation of the variables of interest. Note that the default of *cor* command will show the Pearson correlation coefficient. If one wants to run other types of the correlation coefficient, add an option: *method = "spearman" or "kendall."*



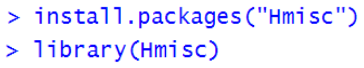
```
> cor(data_chapter8[,4:7])
           X3         X4         X5         X6
X3  1.0000000 -0.2713070 -0.9790343 -0.7975773
X4 -0.2713070  1.0000000  0.4095501  0.5059641
X5 -0.9790343  0.4095501  1.0000000  0.7765873
X6 -0.7975773  0.5059641  0.7765873  1.0000000
```

We simplify the results table with smaller digits of the correlation coefficient by using the *round* command. We designate to print two digits on the results table as follows.



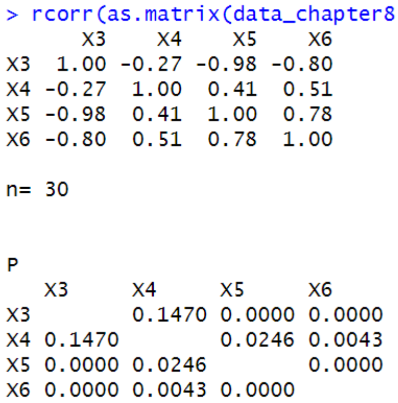
```
> simple_cor=round(cor(data_chapter8[,4:7]),2)
> simple_cor
      X3    X4    X5    X6
X3  1.00 -0.27 -0.98 -0.80
X4 -0.27  1.00  0.41  0.51
X5 -0.98  0.41  1.00  0.78
X6 -0.80  0.51  0.78  1.00
```

To show the results table with significance, we need to install a package named '*Hmisc'* then attach it to R.



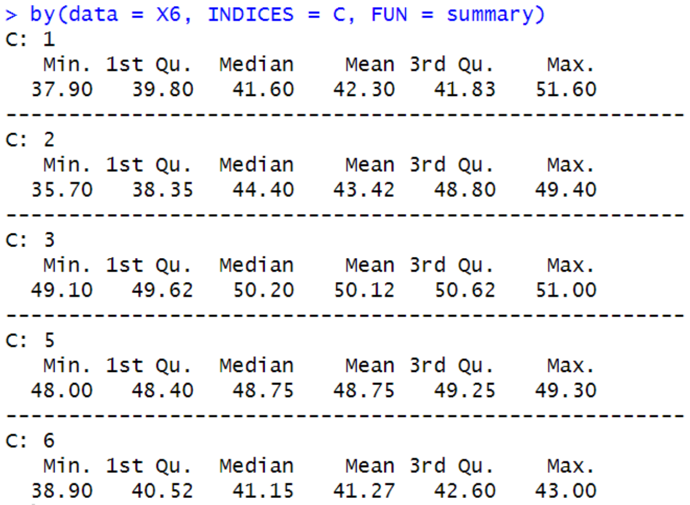
```
> install.packages("Hmisc")
> library(Hmisc)
```

The *rcorr* command allows us to examine the results according to the *p*-value. The variables of interest are the same as the previous analyses. With the *rcorr* command, we can investigate the correlation coefficient matrix, the number of observations, and the *p*-value of the correlation. Note that the default of this command also calculates the correlation with the Pearson coefficient.



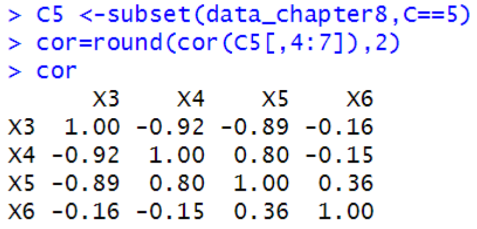
```
> rcorr(as.matrix(data_chapter8[,4:7]))
      X3    X4    X5    X6
X3  1.00 -0.27 -0.98 -0.80
X4 -0.27  1.00  0.41  0.51
X5 -0.98  0.41  1.00  0.78
X6 -0.80  0.51  0.78  1.00

n= 30


P
   X3     X4     X5     X6    
X3        0.1470 0.0000 0.0000
X4 0.1470        0.0246 0.0043
X5 0.0000 0.0246        0.0000
X6 0.0000 0.0043 0.0000       
```

We have produced the correlation estimates for the entire dataset, which has all the countries in one group (In the dataset, C variable means all countries according to their economic levels). Suppose we are interested in finding the correlation between these variables for one specific country. We produce descriptive statistics on specific variable X6 for each country. It is performed by using the *by* command in R.



```
> by(data = X6, INDICES = C, FUN = summary)
C: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  37.90   39.80   41.60   42.30   41.83   51.60 
------------------------------------------------------
C: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  35.70   38.35   44.40   43.42   48.80   49.40 
------------------------------------------------------
C: 3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  49.10   49.62   50.20   50.12   50.62   51.00 
------------------------------------------------------
C: 5
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  48.00   48.40   48.75   48.75   49.25   49.30 
------------------------------------------------------
C: 6
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  38.90   40.52   41.15   41.27   42.60   43.00 
```

We then are interested to see the correlation between the variables C=5. We can make a subset when C equals to 5, then conduct the *cor* command to produce the correlation coefficient in R. The command used for the results are as follows:



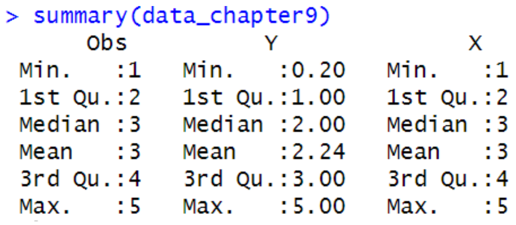
```
> C5 <-subset(data_chapter8,C==5)
> cor=round(cor(C5[,4:7]),2)
> cor
      X3    X4    X5    X6
X3  1.00 -0.92 -0.89 -0.16
X4 -0.92  1.00  0.80 -0.15
X5 -0.89  0.80  1.00  0.36
X6 -0.16 -0.15  0.36  1.00
```

As you can see, the correlation between x3 and other variables for this country setting is opposite in signs when compared with the same estimates for the overall sample. Thus, this basic data exploratory analysis indicates that there are intercountry differences in these variables and that the researcher must take into account the heterogeneity across countries.

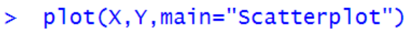
Chapter 9 Effects of Individual, Household, and Community Indicators on Child's Nutritional status- Application of Simple Linear Regression

SIMPLE REGRESSION IN R (Please find the data used for this chapter here)

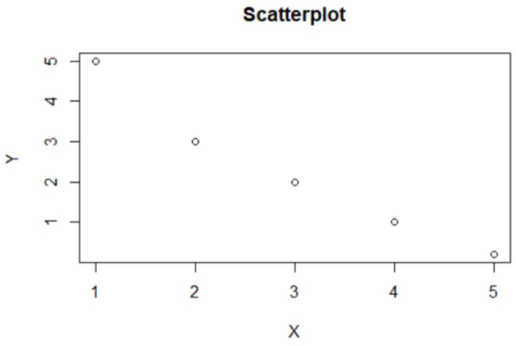
We employ a simple dataset to understand how regression operates in R. Assume we have the following five observations for Farm output (Y) and the price of fertilizers, which is an input called X. To begin with, we draw a scatter graph by using the *plot* command to identify the relationship between the farm output and the price of fertilizers. From the graph, we can identify that X and Y have a negative relationship.



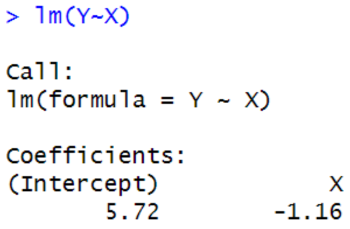
```
> summary(data_chapter9)
      Obs          Y              X
 Min.   :1   Min.   :0.20   Min.   :1
 1st Qu.:2   1st Qu.:1.00   1st Qu.:2
 Median :3   Median :2.00   Median :3
 Mean   :3   Mean   :2.24   Mean   :3
 3rd Qu.:4   3rd Qu.:3.00   3rd Qu.:4
 Max.   :5   Max.   :5.00   Max.   :5
```



```
>  plot(X,Y,main="Scatterplot")
```



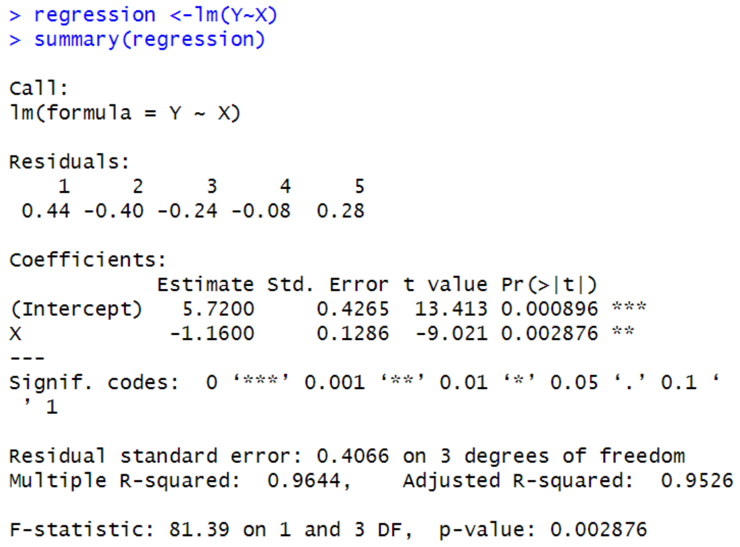
Now we will derive these coefficients in R using the following command. Note that the dependent variable should place first. The *lm* command operates a regression in R.



```
> lm(Y~X)

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
       5.72        -1.16
```

In order to look at in detail about the regression, we can save the function and import in using *summary* as follows. We can examine general and specific information such as *F*-statistics, R-squared, adjusted R squared, and *p*-value. We can see that the estimated equation for our data:

$$Corn\ Output = 5.72 - 1.16\ Fertilizer\ Price$$



```
> regression <-lm(Y~X)
> summary(regression)

Call:
lm(formula = Y ~ X)

Residuals:
    1     2     3     4     5 
 0.44 -0.40 -0.24 -0.08  0.28 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.7200     0.4265  13.413 0.000896 ***
X            -1.1600     0.1286  -9.021 0.002876 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
 ’ 1

Residual standard error: 0.4066 on 3 degrees of freedom
Multiple R-squared:  0.9644,    Adjusted R-squared:  0.9526 

F-statistic: 81.39 on 1 and 3 DF,  p-value: 0.002876
```

Thus, a one-unit increase in the price of fertilizers will decrease output by 1.16 units. Further, the *t*-value of the coefficient is -9.02, with the p-value equal to 0.003. This implies that the estimate is significant at the 1% level.

Chapter 10 Maternal Education and Community Characteristics as Indicators of Nutritional Status of Children—Application of Multivariate Regression

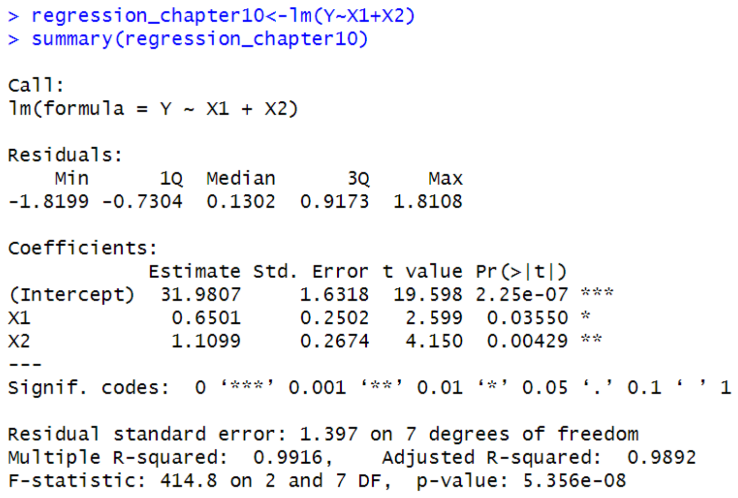
MULTIPLE REGRESSION IN R (Please find the data used for this chapter here)

We take a simple example with ten observations on corn output that we examined in the last chapter. We now have two independent variables to perform a multiple regression in R. We are interested in estimating a regression equation that captures the following relationship:

$$Y = a + b\,X_1 + c\,X_2$$

We have three parameters to capture: a, b, and c. Note that we simplify the notations for variables: Corn = Y or the outcome variable, Fertilizer = X_1 and Pesticide = X_2 are the two independent variables. We expand the *lm* command in R to include additional explanatory variables.

To perform multiple regression in R, the dependent variable should list first and list independent variables by using + sign between variables. As undertaken earlier, we save the results of multiple regression as regression_chapter10 and print it in detail in using the *summary* command as follow:

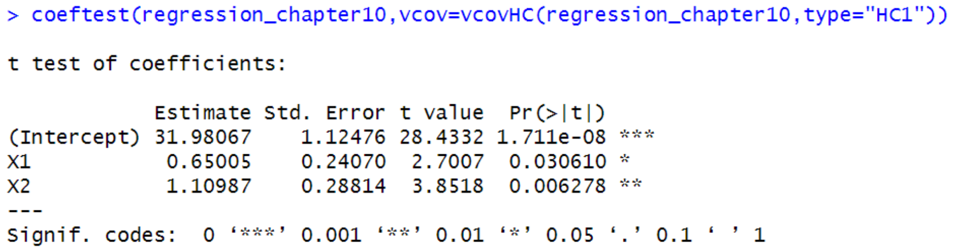


```
> regression_chapter10<-lm(Y~X1+X2)
> summary(regression_chapter10)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8199 -0.7304  0.1302  0.9173  1.8108 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.9807     1.6318  19.598 2.25e-07 ***
X1            0.6501     0.2502   2.599  0.03550 *  
X2            1.1099     0.2674   4.150  0.00429 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.397 on 7 degrees of freedom
Multiple R-squared:  0.9916,    Adjusted R-squared:  0.9892 
F-statistic: 414.8 on 2 and 7 DF,  p-value: 5.356e-08
```

From the table, we substitute coefficients to the equation. Corn output is positively related to the use of fertilizers and pesticides. A one-unit increase in fertilizer and Pesticide usages incline corn output grows by 0.65 and 1.1 units, respectively. We can see information about residuals, *t*-value of coefficient, and *p*-value. According to the value of adjusted *R*-squared, the change in corn output is due to the variability in fertilizer and pesticide usage at 99% confidence.

$$Corn\ Ouput = 32 + 0.65\ Fertilizer + 1.1\ Pesticide$$

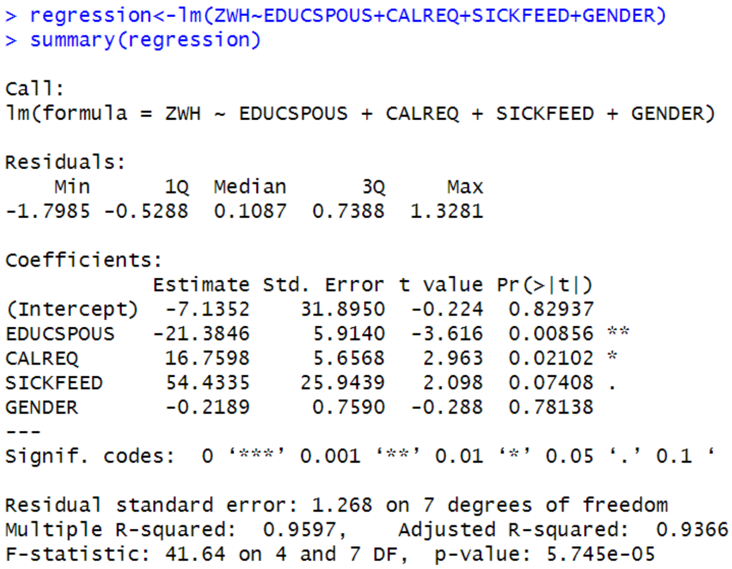
Then, we proceed to investigate the impact of pesticide and fertilizer use on corn output by computing robust covariance matrix estimators. We use the variance estimator in a regression model by installing *lmtest* and *sandwich* package. With *coeftest* command from the ‘*lmtest’* package, one will get the same results as the *vce(robust)* option in Stata. With *vcovHC* from the sandwich package, we can employ a Heteroskedasticity-consistent estimation of the covariance matrix to produce a robust regression. The command and results of controlling for heteroskedasticity in the data are as follows. We use the variance estimator of the previous regression named regression_chapter 10 to adjust values.



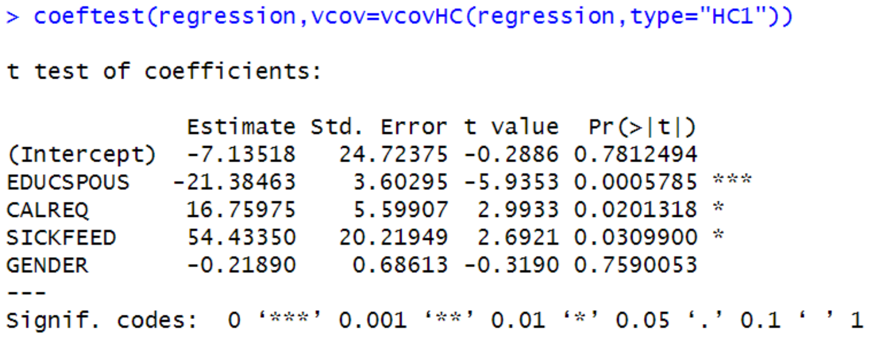
```
> coeftest(regression_chapter10,vcov=vcovHC(regression_chapter10,type="HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 31.98067    1.12476 28.4332 1.711e-08 ***
X1           0.65005    0.24070  2.7007  0.030610 *  
X2           1.10987    0.28814  3.8518  0.006278 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The estimated equation is still the same as before. However, the standard errors and t-values have changed because of the command controls for Heteroskedasticity in the data.

We now look at another example with more independent variables. We will check the hypothesis test and violations of regression assumptions at the end of this chapter. First, we perform a multivariate regression on ZWH concerning four independent variables, which are EDUSCPOUS, CALREQ, SICKFEED, and GENDER. As we performed earlier, we use *lm* and *coeftest* command to produce robust regression results as follows:



```
> regression<-lm(ZWH~EDUCSPOUS+CALREQ+SICKFEED+GENDER)
> summary(regression)

Call:
lm(formula = ZWH ~ EDUCSPOUS + CALREQ + SICKFEED + GENDER)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7985 -0.5288  0.1087  0.7388  1.3281 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -7.1352    31.8950  -0.224  0.82937   
EDUCSPOUS   -21.3846     5.9140  -3.616  0.00856 **
CALREQ       16.7598     5.6568   2.963  0.02102 * 
SICKFEED     54.4335    25.9439   2.098  0.07408 . 
GENDER       -0.2189     0.7590  -0.288  0.78138   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.268 on 7 degrees of freedom
Multiple R-squared:  0.9597,	Adjusted R-squared:  0.9366 
F-statistic: 41.64 on 4 and 7 DF,  p-value: 5.745e-05
```

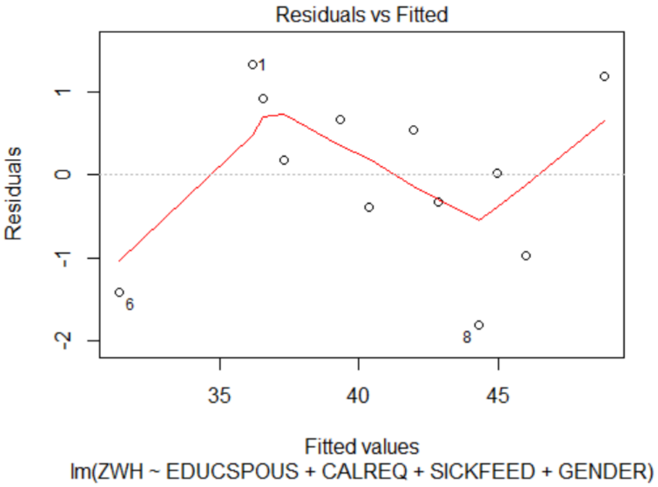


```
> coeftest(regression,vcov=vcovHC(regression,type="HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  -7.13518   24.72375 -0.2886 0.7812494    
EDUCSPOUS   -21.38463    3.60295 -5.9353 0.0005785 ***
CALREQ       16.75975    5.59907  2.9933 0.0201318 *  
SICKFEED     54.43350   20.21949  2.6921 0.0309900 *  
GENDER       -0.21890    0.68613 -0.3190 0.7590053    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

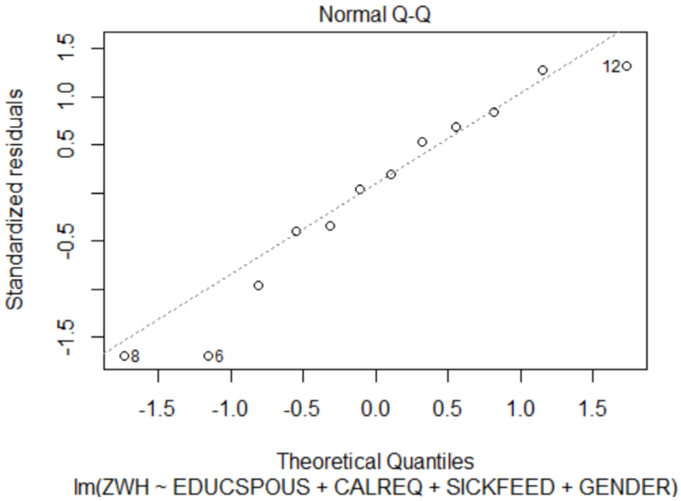
As a result, we can get an equation. CALREQ and SICKFEED have a positive and significant impact on ZWH, 16.76, and 54.43 changes per unit, at the 5% level, respectively. EDUCSPOUS and GENDER have negative impacts on ZWH. However, GENDER variable is not statistically significant.

$$ZWH = -7.13 - 21.38\,EDUCSPOUS + 16.76\,CALREQ + 54.43\,SICKFEED - 0.22\,GENDER$$

We should check the normality of the errors to validate the regression. The *plot(regression model)* command is used to check the validity with a graph. R has a set of built-in regression diagnostic plot command with regression results allows us to produce a residual plot. X-axis is the predicted or fitted Y values. On the Y-axis, is the residual or errors. If the normality assumption is met, the line should be flat, and points should seem a cloud. If the variation is constant, there might be no pattern. However, in the figure below, there is a clear pattern, not making a relatively flat line. It is because there are not enough observations in the dataset that we used. We cannot find any patterns of points, so we can say that we need more observations to ensure the validity of the regression.



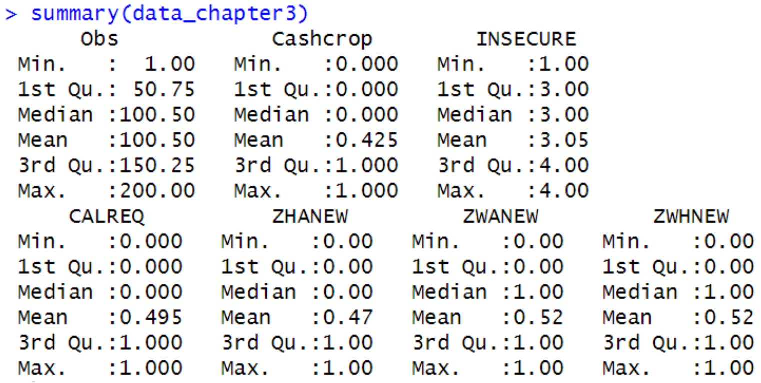
We can also examine the Q-Q plot or quantile-quantile plot by using the *qqline (regression model)* command. Y-axis is the standardized residuals and X-axis is the ordered theoretical residuals. The points should follow a diagonal line if the errors or the residuals are normally distributed.



Chapter 11 Predicting Child Nutritional Status Using Related Socioeconomic Variables – Application of Discriminant Function Analysis

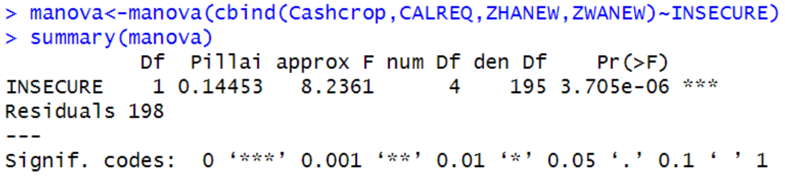
CANONICAL DISCRIMINANT ANALYSIS USING R (Please find the data used for this chapter here)

In this section, we outline the steps in identifying the discriminant functions using R for a sample of 200 farmers, which was used in Chapter 3. We first recapitulate the data using the *summary* command in R.



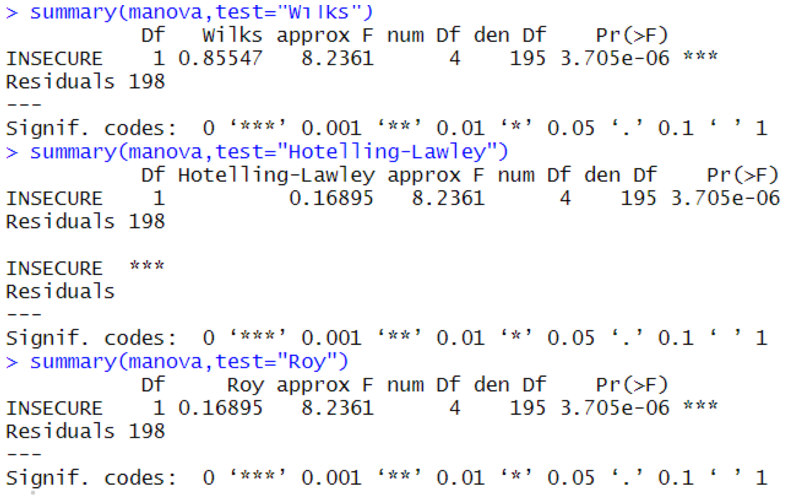
```
> summary(data_chapter3)
      Obs            Cashcrop        INSECURE
 Min.   :  1.00   Min.   :0.000   Min.   :1.00
 1st Qu.: 50.75   1st Qu.:0.000   1st Qu.:3.00
 Median :100.50   Median :0.000   Median :3.00
 Mean   :100.50   Mean   :0.425   Mean   :3.05
 3rd Qu.:150.25   3rd Qu.:1.000   3rd Qu.:4.00
 Max.   :200.00   Max.   :1.000   Max.   :4.00
     CALREQ          ZHANEW         ZWANEW         ZWHNEW
 Min.   :0.000   Min.   :0.00   Min.   :0.00   Min.   :0.00
 1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00
 Median :0.000   Median :0.00   Median :1.00   Median :1.00
 Mean   :0.495   Mean   :0.47   Mean   :0.52   Mean   :0.52
 3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.00
 Max.   :1.000   Max.   :1.00   Max.   :1.00   Max.   :1.00
```

Then we start by using the *manova* and *summary* command in R that shows a class for the multivariate analysis of variance. The output is :



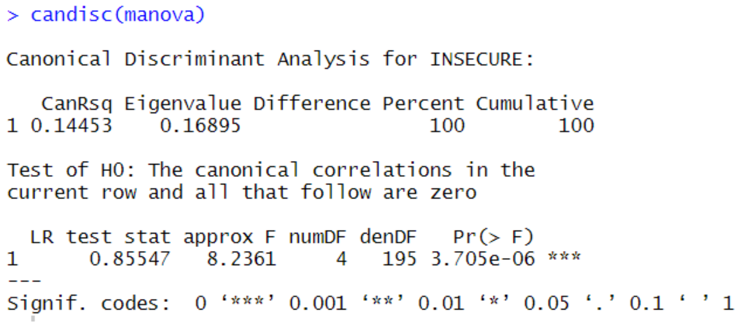
```
> manova<-manova(cbind(Cashcrop,CALREQ,ZHANEW,ZWANEW)~INSECURE)
> summary(manova)
           Df  Pillai approx F num Df den Df    Pr(>F)
INSECURE    1 0.14453   8.2361      4    195 3.705e-06 ***
Residuals 198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that the default of the *manova* command in R is the value of Pillai’s trace. By using other arguments, we undertake several other multivariate tests. We can add test = “Wilks” / “Hotelling-Lawley” / “Roy” to produce other statistics that capture the outcomes.



```
> summary(manova,test="Wilks")
           Df   Wilks approx F num Df den Df    Pr(>F)
INSECURE    1 0.85547   8.2361      4    195 3.705e-06 ***
Residuals 198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(manova,test="Hotelling-Lawley")
           Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
INSECURE    1          0.16895   8.2361      4    195 3.705e-06
Residuals 198

INSECURE   ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(manova,test="Roy")
           Df     Roy approx F num Df den Df    Pr(>F)
INSECURE    1 0.16895   8.2361      4    195 3.705e-06 ***
Residuals 198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In R, we use the *candisc* command to undertake canonical linear discriminant analysis.

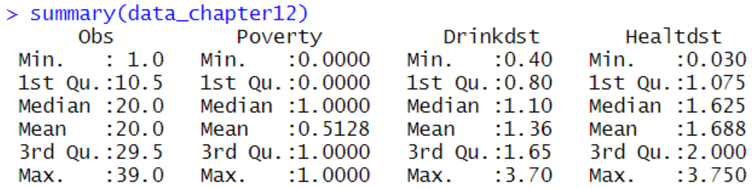


```
> candisc(manova)

Canonical Discriminant Analysis for INSECURE:

   CanRsq Eigenvalue Difference Percent Cumulative
1 0.14453    0.16895                100        100

Test of H0: The canonical correlations in the 
current row and all that follow are zero

  LR test stat approx F numDF denDF    Pr(> F)    
1      0.85547   8.2361     4   195 3.705e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Chapter 12 Measurement and Determinants of Poverty – Application of Logistic Regression Model

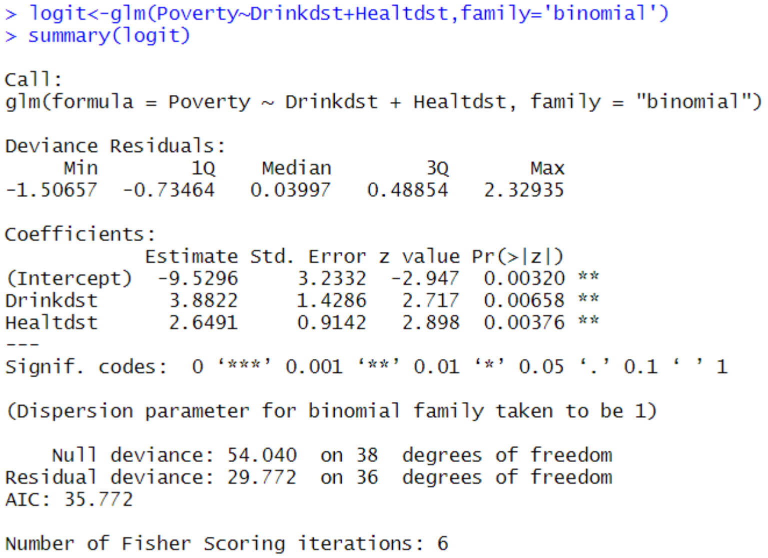
ESTIMATING LOGISTIC REGRESSION MODELS IN R (Please find the data used for this chapter here)

In this section, we implement the logistic model estimation in R. We have a dataset with 39 observations on three variables in the context of the determinant of household welfare, Poverty, Drinkdst, and Healthdst. Our objective of this part is to relate the two independent variables (drniksdt and healthdst) to the poverty status of the household. As mentioned in the previous chapters, we perform the *summary* command first to identify the descriptive statistics.



```
> summary(data_chapter12)
      Obs           Poverty           Drinkdst        Healtdst    
 Min.   : 1.0   Min.   :0.0000   Min.   :0.40   Min.   :0.030  
 1st Qu.:10.5   1st Qu.:0.0000   1st Qu.:0.80   1st Qu.:1.075  
 Median :20.0   Median :1.0000   Median :1.10   Median :1.625  
 Mean   :20.0   Mean   :0.5128   Mean   :1.36   Mean   :1.688  
 3rd Qu.:29.5   3rd Qu.:1.0000   3rd Qu.:1.65   3rd Qu.:2.000  
 Max.   :39.0   Max.   :1.0000   Max.   :3.70   Max.   :3.750  
```

The *glm* command in R provides the estimates and statistical information. The input and the corresponding output from R are given below.

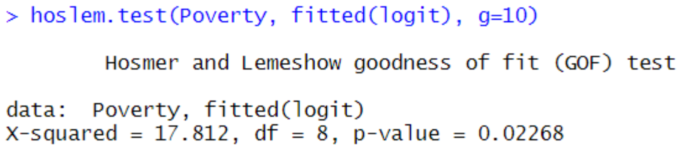


```
> logit<-glm(Poverty~Drinkdst+Healtdst,family='binomial')
> summary(logit)

Call:
glm(formula = Poverty ~ Drinkdst + Healtdst, family = "binomial")

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.50657  -0.73464   0.03997   0.48854   2.32935  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)  -9.5296     3.2332  -2.947  0.00320 **
Drinkdst      3.8822     1.4286   2.717  0.00658 **
Healtdst      2.6491     0.9142   2.898  0.00376 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 29.772  on 36  degrees of freedom
AIC: 35.772

Number of Fisher Scoring iterations: 6
```

We can get an equation with the results and interpret that the distance to drinking water and health facility increases, the likelihood that a facility is poor increases. The output produces the standard errors of the coefficients and the z-values alongside the p-values. From the output, we can see that p-values are all less than 0.05.

$$Poverty = \; -9.52 + 3.88 \; drinkdst + 2.64 \; healthdst$$

The Hosmer-Lemeshow test is a statistical test for goodness of fit for logistic regression models in R. We need to install the ‘*ResourceSelection’* package to produce the test.

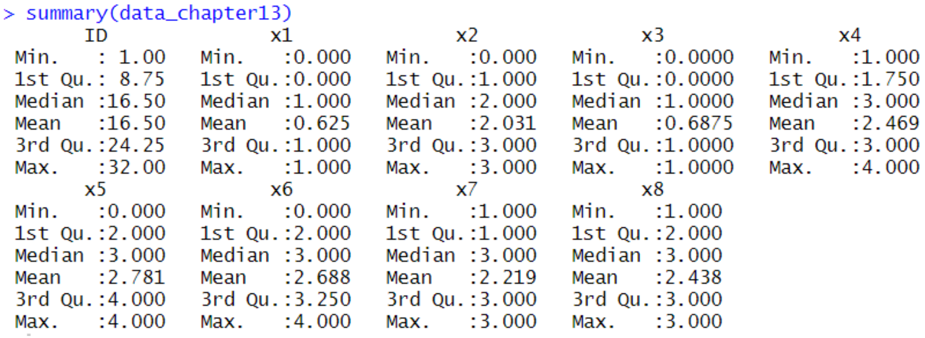


```
> hoslem.test(Poverty, fitted(logit), g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  Poverty, fitted(logit)
X-squared = 17.812, df = 8, p-value = 0.02268
```

Chapter 13 Classifying Households on Food Security and Poverty Dimensions – Application of K-Mean Cluster Analysis

CLUSTER ANALYSIS IN R (Please find the data used for this chapter here)

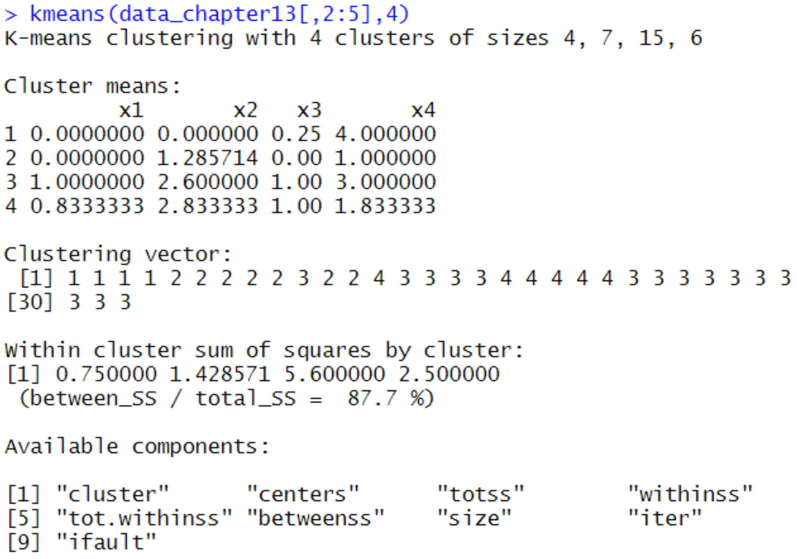
We are going to conduct a K-mean cluster analysis in this chapter. We can conduct cluster analysis in R using *kmeans* command. We use a dataset with 32 observations and nine variables. Before performing an analysis, we produce descriptive statistics with the *summary* command.



```
> summary(data_chapter13)
       ID              x1              x2              x3               x4       
 Min.   : 1.00   Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :1.000  
 1st Qu.: 8.75   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.750  
 Median :16.50   Median :1.000   Median :2.000   Median :1.0000   Median :3.000  
 Mean   :16.50   Mean   :0.625   Mean   :2.031   Mean   :0.6875   Mean   :2.469  
 3rd Qu.:24.25   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :32.00   Max.   :1.000   Max.   :3.000   Max.   :1.0000   Max.   :4.000  
       x5              x6              x7              x8       
 Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000  
 Median :3.000   Median :3.000   Median :3.000   Median :3.000  
 Mean   :2.781   Mean   :2.688   Mean   :2.219   Mean   :2.438  
 3rd Qu.:4.000   3rd Qu.:3.250   3rd Qu.:3.000   3rd Qu.:3.000  
 Max.   :4.000   Max.   :4.000   Max.   :3.000   Max.   :3.000  
```

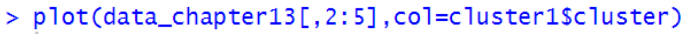
Among these nine variables, we only use variables *x1* through *x4* to perform a cluster analysis with four groups. The *kmeans* command in R is undertaken for the analysis, and we can print output as below.

The k-means clustering with four clusters has the number of observations: 4, 7, 15, and 6 for each group. Cluster means and vector are also produced by the *kmeans* command. Type the command *set.seed(123)* prior to the *kmeans* command and run both the commands simulanteously so that the cluster sizes don't vary every time we run the *kmeans* command.

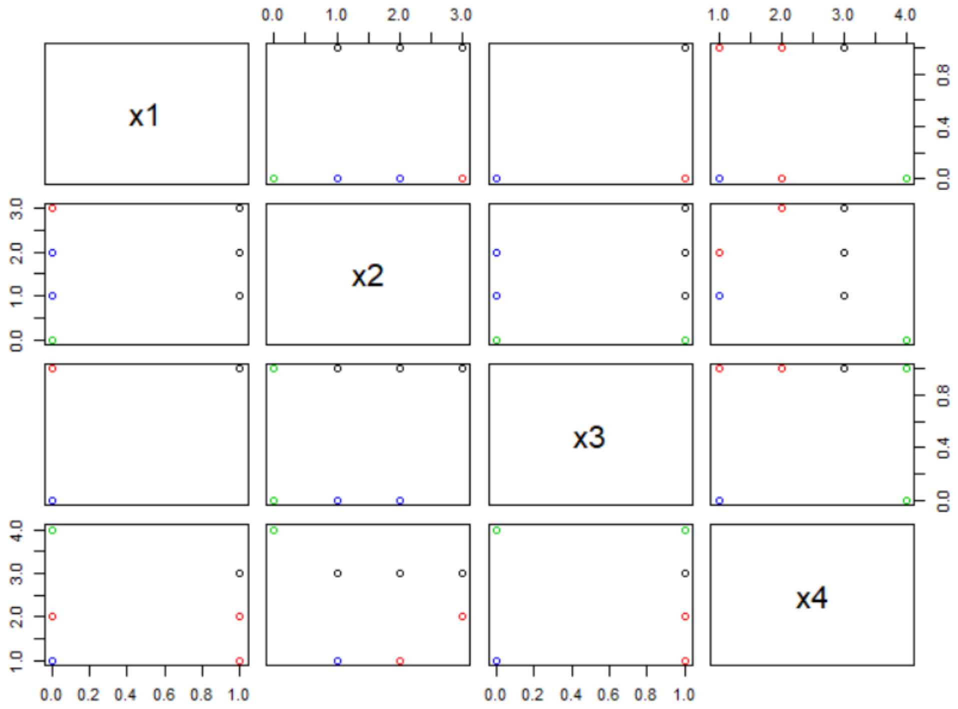


```
> kmeans(data_chapter13[,2:5],4)
K-means clustering with 4 clusters of sizes 4, 7, 15, 6

Cluster means:
         x1       x2   x3       x4
1 0.0000000 0.000000 0.25 4.000000
2 0.0000000 1.285714 0.00 1.000000
3 1.0000000 2.600000 1.00 3.000000
4 0.8333333 2.833333 1.00 1.833333

Clustering vector:
 [1] 1 1 1 1 2 2 2 2 2 2 3 2 2 4 3 3 3 3 4 4 4 4 4 3 3 3 3 3 3
[30] 3 3 3

Within cluster sum of squares by cluster:
[1] 0.750000 1.428571 5.600000 2.500000
 (between_SS / total_SS =  87.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"    
[5] "tot.withinss" "betweenss"    "size"         "iter"        
[9] "ifault"      
```

Now we plot the graph by clusters. By using the *plot* command and *cluster* argument to present different colors within groups. The plots on each variable are as follows:



```
> plot(data_chapter13[,2:5],col=cluster1$cluster)
```



The graphs above present for clusters between two variables. Since the data sets consist of categorical and dummy variables, we cannot see the grouped cluster. However, if the data set has a continuous variable, it would be grouped plot graphs.

Chapter 14 Household Care as a Determinant of Nutritional Status – Application of Instrumental Variable Estimation

(Please find the data used for this chapter here)

With given information on weight for age Z-scores(Z), Breastfeeding practices (BF), Clinical attendance (ATCL), Age in Months (AGEM), Latrine facilities (L), and Food Availability (FA), we are going to use this data set to check if BF can be instrumented using ATCL and whether the null hypothesis of no-endogeneity status. The breastfeeding practices variable is correlated with weight for age Z-scores. Thus, the error terms cannot be met with the independence assumption. Ordinary least square model will not be valid for endogenous variables in it. We undertake the analysis in two steps. In the first stage, we estimate the determinant of breastfeeding practices by using the instrumental variable technique. In the second stage, the predicted value of breastfeeding practices along with other exogenous variables are included in determining the impact on weight for age Z-scores.

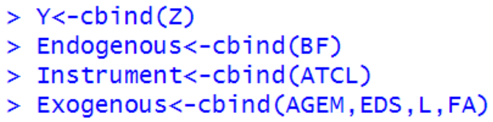
Stage 1: Estimating Breastfeeding Practices

The first stage requires instruments or clinical attendance variables to predict Breastfeeding practices. Therefore, we are going to run a regression model on breastfeeding with instrument variable ATCL and other exogenous variables.

Stage 2: Estimating the Determinants of Weight for Age Z-Scores

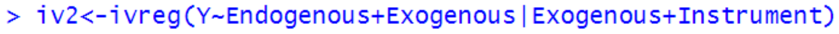
In this stage, we then use the predicted value of Breastfeeding to produce the second-stage regression. We can implement all this in R with the following command.

We first classify variables with the names. Y is a dependent variable that we want to see how other variables affect it. The breastfeeding variable is an endogenous one. ATCL is an instrument variable and AGEM, EDS, L, and FA variables are exogenous variables.



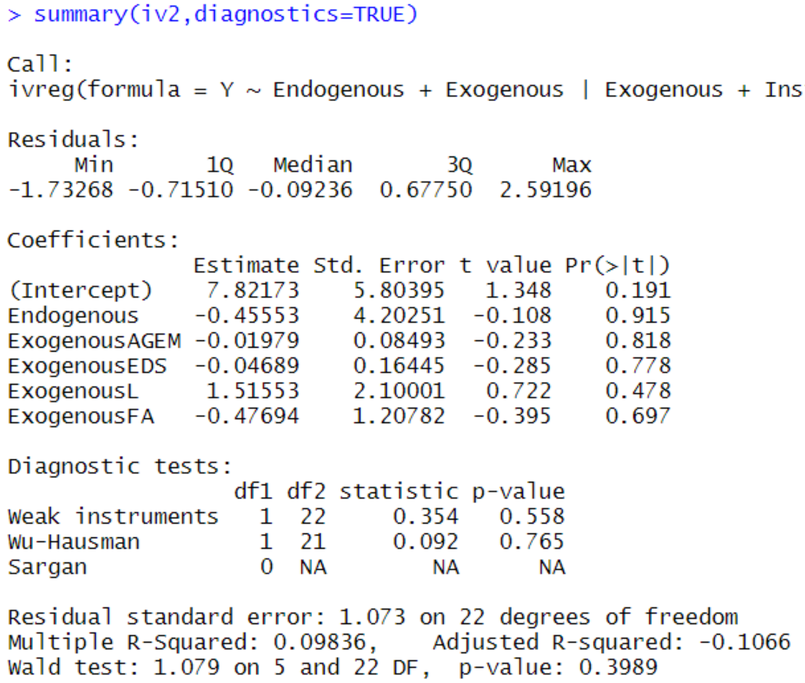
```
> Y<-cbind(Z)
> Endogenous<-cbind(BF)
> Instrument<-cbind(ATCL)
> Exogenous<-cbind(AGEM,EDS,L,FA)
```

We can use the *ivreg* command to undertake the instrumental variable estimation in R. T conduct the analysis using the *ivreg* command, we conduct a regression with endogenous and exogenous variables on the dependent variable. At the same time, we also set up the first regression with exogenous and instrumental variables after |. By using the *ivreg*, we can conduct two-stage least square analysis. Note that adding diagnostics in the summary command will prints specification tests with results.



```
> iv2<-ivreg(Y~Endogenous+Exogenous|Exogenous+Instrument)
```

From the result table, endogenous is the predicted value of the first regression on breastfeeding, where the independent variables are the ATCL and other exogenous variables. First, it is the test for weak instruments with several instruments. The null hypothesis is H_0: “All instruments are weak.” Second test is the Hausman test for endogeneity, where the null hypothesis is $H_0: Cov(x, e) = 0$. Thus, rejecting the null hypothesis indicates the existence of endogeneity and the need for instrumental variables. The last test is the validity of instruments. The Sargan test is also called a test for overidentifying restrictions. The null hypothesis is that the covariance between the instrument and the error term is zero, is $H_0: Cov(z, e) = 0$. Thus, rejecting the null hypothesis indicates that at least one of the extra instruments is not valid. In this case, we cannot conduct the Sargan test because there is only one instrument variable.



```
> summary(iv2,diagnostics=TRUE)

Call:
ivreg(formula = Y ~ Endogenous + Exogenous | Exogenous + Instrument)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.73268 -0.71510 -0.09236  0.67750  2.59196 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.82173    5.80395   1.348    0.191
Endogenous    -0.45553    4.20251  -0.108    0.915
ExogenousAGEM -0.01979    0.08493  -0.233    0.818
ExogenousEDS  -0.04689    0.16445  -0.285    0.778
ExogenousL     1.51553    2.10001   0.722    0.478
ExogenousFA   -0.47694    1.20782  -0.395    0.697

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1  22     0.354   0.558
Wu-Hausman         1  21     0.092   0.765
Sargan             0  NA        NA      NA

Residual standard error: 1.073 on 22 degrees of freedom
Multiple R-Squared: 0.09836,    Adjusted R-squared: -0.1066 
Wald test: 1.079 on 5 and 22 DF,  p-value: 0.3989 
```

As a result, there are no significant variables in the analysis and the Hausman test cannot be rejected. Therefore, in the data, there is no need to use instrument variables for addressing endogeneity. However, in other research, instrument variables are commonly employed to address the endogenous problem.

Appendix

STATA and R are software that is very powerful and flexible, which has become very popular in recent years among researchers, students, and policy analysts. In this book we have updated R commands in many chapters and have shown the implementation using commands and output from R which are corresponding the implementation from STATA. To summarize, we have illustrated the following applications in this revised edition:

Chapter	Concept	STATA command	R command
2	t-tests	ttest	t.test
3	Chi-squared tests	tabulate	chisq.test
4	Cramer’s V	tab x y, column nokey chi2 V	cramersV
5	One-way ANOVA	oneway	oneway.test
6	Factor analysis	factor	factanal
7	Two-way ANOVA	anova	aov
8	Correlation	correlate or pwcorr	cor or rcorr
9	Simple regression	regress	lm
10	Multiple regression	regress	lm
11	Discriminant analysis	manova and candisc	manova and candisc
12	Logistic regression	logit	glm
13	K-means clustering	tabstat	kmeans
14	Instrumental analysis	ivregress	ivreg