R II

Inferential Statistics

2026-03-03

After attending this training session, you will be able to run and interpret the following tests in R:

1. Chi-square test
2. Independent samples t-test and Levene's test for the equality of variances
3. One-way ANOVA
4. Bivariate Statistics: Correlation
5. Bivariate Statistics: Regression
6. Multiple regression

Introduction

Code and Dataset

For this training session, please download the following data:

RII_Data.csv

Recap of R I

In the **R I** training session, we covered the following topics. Please feel free to refer back to the R I workbook using the following links if you would like a refresher.

- Starting R/RStudio
- Data Management
- Descriptive Statistics
- Data Transformation
- Multivariate Tables
- Graphics
- Saving and Printing

As in the R I training session, we will be using data from the General Social Survey.

Getting Started

Start RStudio

To launch RStudio, click on the start icon in the lower left-hand corner of the screen. Select RStudio from the list of programs.

Set Your Working Directory

Once you have downloaded the data files for this workshop (`RII\_Data.csv`), you will want to set R's working directory so that you can import and export files using a relative path. To set your working directory with code, type `setwd("C:/Users/YourNetID/Desktop/RII")`, or wherever your files are located. To change via the **Session** menu, go to **Session > Set Working Directory**. Use **Choose Directory...** if you want to choose a certain directory or have not saved your script file; use **To Source File Location** if you have already saved your script on your computer.

Load the Workshop Data

If you have set your working directory to the location where the workshop data file is stored, simply read in the file using the following code:

```
    gssdata = read.csv("RII_Data.csv")
```

To double check that your data were imported correctly, you can examine the dimensions and structure of the data:

```
    > dim(gssdata)
    [1] 2348   37

    > str(gssdata)
  'data.frame': 2348 obs. of 37 variables:
  $ id : int 1 2 3 4 5 6 7 8 9 10 ...
  $ age : int 43 74 42 63 71 67 59 43 62 55 ...
  $ year : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
  $ wrkstat : chr "temp not working" "retired" "working fulltime" "working
fulltime" ...
  $ hrs1 : int NA NA 40 40 NA NA 35 89 40 402 ...
    $ agekdbrn   : int  NA 20 19 29 NA NA NA 21 35 25 ...
    $ sibs       : int  1 4 4 2 1 2 1 6 0 4 ...
    ...
```

Useful Packages

We will be using several additional R packages to run descriptive and inferential statistics. These packages extend base R functionality and make certain processes easier in R. Best practice is to install and load all packages at the beginning of your R script. To first install and then load the packages, run the following code:

```
    > install.packages('car')   # Companion for Applied Regression
    # Quantitative Analysis Made Accessible
    > install.packages('descr') # Descriptive Statistics
    > install.packages('psych') # Procedures for Psychological Research
    > install.packages('lm.beta') # Add Standardized Regression Coefficients
    > install.packages("rstatix") # For the Games-Howell Test
```

```
> library(car)
> library(descr)
> library(psych)
> library(lm.beta)
> library(rstatix)
```

> ⚠ Warning
>
> Sometimes different packages will have functions with the same names. If you have loaded packages that have the same function names, R will print a warning when loading the packages. (This is called a "namespace conflict.") To specify which package's function you want to use, you can either:
>
> A) load the package with your desired function AFTER loading the other package(s), or
> B) specify the package when running the function with the syntax `packagename::functionname()` (e.g. `descr::freq()`).
>
> The second option is the preferred/safest option. To print a list of all the name conflicts in the packages you currently have loaded, run `conflicted::conflict_scout()`.

Chi-square Test

A chi-square test is used to determine whether there is a statistically significant association between two categorical variables. To run a chi-square test, use the `chisq.test()` function.

Example 1: 2x2 Table

For example, you may want to know if there is a significant association between the variables *hschem* (R ever took a high school chemistry course) and *sex* (respondent's gender). The table function creates a contingency table of the counts at each combination of the variable categories. In this case, we create a table object **tbl** to display the cross-tabular frequency between **sex** and **hschem** variable. Both chi-square and Fisher's exact text will be used.

```
> tbl = table(gssdata$hschem,gssdata$sex)
> tbl

     female male
no      295  183
yes     363  271
```

The variable name that appears in the first variable in the table function will appear in the rows and the second variable will be placed in the columns. Either of the variables can go in the rows or columns, although columns typically contain the independent variable.

Different tests are appropriate for different situations. Guidelines for choosing the tests are explained below:

Chi-Square: Used to test the hypothesis that the row and column variables are independent. This should not be used if any cell has an expected value less than 1, or if more than 20

Fisher's Exact Test: A test for independence in a 2x2 table. This statistic is most useful when the total sample size and expected values are small.

Continuity Adj. Chi-Square: A correction is applied in the computation of chi-square statistic for 2x2 tables to improve the approximation. Corrected chi-square values are always smaller than uncorrected values.

We can now perform a chi-squared test using the `chisq.test()` function to determine if there is a significant relationship between gender and taking chemistry course at high school :

```
    > chi = chisq.test(gssdata$hschem,gssdata$sex)
    > chi

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

data:  gssdata$hschem and gssdata$sex
X-squared = 2.0631, df = 1, p-value = 0.1509
```

R generates the following test statistics for chi-square: the test statistic, degrees of freedom, and p-value.

Because this example is 2 by 2 table with count data, we can also calculate Fisher's Exact Test.

```
    > fisher = fisher.test(gssdata$hschem,gssdata$sex)
    > fisher

    Fisher's Exact Test for Count Data

data:  gssdata$hschem and gssdata$sex
p-value = 0.1395
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9376168 1.5455167
sample estimates:
odds ratio
  1.203261
```

In this example, the *p-value* for Fisher's Exact Test is 0.1395. Thus, we fail to reject the null hypothesis and conclude that there is not a statistically significant relationship between respondent's sex and taking chemistry course at high school.

Example 2: 3x2 Table

In this example, we will test whether there is a significant association between the variables *colsci* (Respondents who took college-level science courses or not) and * natspac* (Attitudes towards the space exploration program).

```
> # step one create a 3x2 table
> table(gssdata$natspac,gssdata$colsci)

              no yes
  about right 188 160
  too little   53  96
  too much    125  61

> # step two run chisq
> chisq.test(gssdata$natspac,gssdata$colsci)

    Pearson's Chi-squared test

data:  gssdata$natspac and gssdata$colsci
X-squared = 33.34, df = 2, p-value = 5.759e-08
```

It is appropriate to use the Chi-Square test statistic in this example. The *p-value* is 5.759e-08, therefore it can be concluded that there is a statistically significant relationship between attitudes towards the space program and taking college-level science courses. Additional arguments can be examined by typing: `?chisq.test`.

Additional Arguments for Chi-Square statistics

Some additional options that can be specified in the model arguments include: * **correct**: a logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables * **simulate.p.value**: a logical indicating whether to compute p-values by Monte Carlo simulation

Exercise 1

Perform a chi-square test to determine if educational attainment (*degree*) is significantly associated with the belief that marijuana should be legal (*grass*). (a) How many high school graduates said "Yes" to legalization? (b) Is there a statistically significant relationship between the two variables? Which chisquare test statistic is appropriate? (See Appendix A for answers.)

1. How many high school graduates (*degree* = 1) said Yes to legalization (*grass* = 1)?
2. Is there a statistically significant relationship between the two variables? Which chi-square test statistic is appropriate?

Independent samples t-test

An independent samples t-test determines whether there is a statistically significant difference in the mean of a continuous dependent variable between two groups.

For example, this procedure may be used to find out if there is a statistically significant difference between the responses of men and women (*sex*) and how much they have to relax per day (*hrlax*). The independent samples t-test is obtained by using the `t.test()` function.

Before performing a t-test, we need to compare the variances of two samples. Different t-test methods make different assumptions about whether or not the variances of the two samples differ. To test for equality of variances, we will perform a Levene's Test using the `leveneTest()` function in the **car** package. The first argument for `leveneTest()` is the response variable of interest, *hrlax*. The second argument is the grouping variable, in our case *sex*:

```
> library(car) # if not loaded previously

> leveneTest(gssdata$hrlax, gssdata$sex)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    1  0.7239  0.395
      1403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Levene method tests the null hypothesis that the variances *are* equal. Because the significance of Levene's test in this example is larger than 0.05, we fail to reject the null hypothesis, and the assumption of equal variance is met. In such a case, the statistics for Equal variances assumed should be used.

```
> t.test(gssdata$hrlax ~ gssdata$sex, var.equal = TRUE)

Two Sample t-test

data:  gssdata$hrlax by gssdata$sex
t = -3.2001, df = 1403, p-value = 0.001405
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7643119 -0.1833752
sample estimates:
mean in group female   mean in group male
            3.496571             3.970414
```

The p-value is $0.001405 < 0.05$, therefore we conclude that there is a statistically significant difference in the number of hours men and women have to relax per day.

In case of unequal variances, we should use the Welch's t-test with the `t.test()` function:

```
> t.test(gssdata$hrlax ~ gssdata$sex, var.equal = FALSE)
```

Exercise 2

Find out if there is a statistically significant difference in terms of respondent's income (*realrinc*) between respondents who took college-level science courses and those who didn't (*colsci*). (See Appendix A for answers.)

ANOVA

A one-way ANOVA (analysis of variance) is used to determine whether there is a statistically significant difference in the mean of a continuous dependent variable among more than two groups.

For example, this procedure may be used to find out if there is a statistically significant difference in the average number of hours spent watching TV per day (*tvhours*) based on marital status (*marital*). We will again need to test for equality of variances among groups using `LeveneTest()`. If equal variances are assumed, we can use the `aov()` function to calculate the ANOVA. If equal variances are not assumed, we should use the `oneway.test()` function with `var.equal = FALSE` (the default value for this argument).

```
> leveneTest(gssdata$tvhours, gssdata$marital)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    4  5.9046 0.0001027 ***
      1549

> ## if equal variances assumption is met
> anova = aov(gssdata$tvhours ~ gssdata$marital)
> summary(anova)
                   Df Sum Sq Mean Sq F value   Pr(>F)
gssdata$marital    4    263   65.85   8.331 1.2e-06 ***
Residuals       1549  12244    7.90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
993 observations deleted due to missingness

> ## if equal variances assumption is not met
> oneway.test(gssdata$tvhours ~ gssdata$marital)

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

data:  gssdata$tvhours and gssdata$marital
F = 7.6417, num df = 4.00, denom df = 217.39, p-value = 8.87e-06
$
```

Since the Levene's test is significant, as seen above, we reject the null hypothesis that the variances are equal, and unequal variances are assumed. This is important in determining which F-test will be used and which post hoc tests will be best suited for any further analysis.

We should run `oneway.test()`, which assumes unequal variances by default and calculates the Welch test of equality of means. If the variances are found to be homogeneous following the Levene's test, we should otherwise choose the regular ANOVA.

The ANOVA F-test p-value is < 0.001, therefore we conclude that there are significant differences in number of hours spent watching TV per day by marital status.

ANOVA Post Hoc Tests

While the ANOVA procedure determines whether differences exist among the group means, post hoc tests are needed to determine which means differ from one another. Which post hoc test you use depends on the homogeneity of the variances.

There are individual functions for common post hoc tests such as Tukey's HSD in R (`TukeyHSD()`). However, the `pairwise.t.test()` function has a `p.adjust.method` argument which can be used to run a variety of post hoc tests, specified below:

Argument Value	Test Name
`"holm"`	Holm
`"BH"`	Benjamini-Hochberg
`"hochberg"`	Hochberg
`"BY"`	Benjamini-Yekatieli
`"hommel"`	Hommel
`"fdr"`	false discovery rate
`"bonferroni"`	Bonferroni

Since we can not assume homogeneity of variances, we will use the Games-Howell post-hoc test in this example. Games-Howell is not available in `pairwise.t.test()`, so we will use the command `games_howell_test()` in the `rstatix` package.

```
> games_howell_test(gssdata, tvhours ~ marital)
## A tibble: 10 x 8
.y.     group1     group2  estimate conf.low conf.high    p.adj p.adj.signif
* <chr>   <chr>      <chr>      <dbl>    <dbl>     <dbl>    <dbl> <chr>
1 tvhours never mar~ married -0.435   -0.920      0.0506  1.04e-1 ns
2 tvhours never mar~ separa~  0.114   -1.70       1.93    1.00e+0 ns
3 tvhours never mar~ divorc~  0.116   -0.541      0.773   9.89e-1 ns
4 tvhours never mar~ widowed  1.02     0.124      1.91    1.70e-2 *
5 tvhours married    separa~  0.549   -1.23       2.33    9.03e-1 ns
6 tvhours married    divorc~  0.551   -0.00118    1.10    5.10e-2 ns
7 tvhours married    widowed  1.45     0.632      2.27    2.39e-5 ****
8 tvhours separated  divorc~  0.00208 -1.83       1.83    1.00e+0 ns
9 tvhours separated  widowed  0.901   -1.02       2.82    6.77e-1 ns
10 tvhours divorced   widowed  0.899   -0.0290     1.83    6.30e-2 ns
```

- Group Never Married mean is significantly different from group Widowed mean.
- Group Married mean is significantly different from group Widowed mean.
- All the other groups based on marital status have no significant different with each other.

Based on the p-values in the output, significant results are highlighted with asterisks(*). After identifying statistically significant differences, look back at the table of descriptive statistics to determine which marital groups spend more (or less) time watching TV per day. It is clear that the mean hours of time spent watching TV per day is significantly higher for group Widowed.

Exercise 3

Using the one-way ANOVA procedure, find out whether the mean level of income (*realrinc*) is different among different education groups (*degree*). Remember to check the Test of Homogeneity of Variances to determine if the Welch test is required. (See Appendix A for answers.)

Bivariate Correlation

Correlation is a measure of the linear relationship between two continuous variables. A correlation coefficient ranges from −1 to 1. A positive correlation coefficient indicates a positive linear relationship (i.e., as one variable increases, the other tends to as well). On the other hand, a negative correlation coefficient indicates a negative linear relationship. A correlation coefficient of 0 indicates there is no linear relationship between the two variables.

For example, this procedure may be used to measure the relationship between the continuous variables spouse occupation prestige score (*sop*), spouse socioeconomic index score (*ssi*), and the number of hours the spouse usually works a week (*sphrs2*).

> **i** Note
>
> A bivariate correlation measures the linear relationship between two continuous variables, but it is convenient to calculate several bivariate correlations simultaneously in a matrix.

We can use the `cor()` function to measure the correlation between a set of variables. `cor()` contains an argument, `use`, which determines how we treat the NAs (missing observations) in our dataset. The default is `"complete.obs"`.

```
> cor(gssdata[, c('sop','ssi','sphrs2')], use = 'complete.obs')
              sop         ssi      sphrs2
sop     1.0000000  0.96787213 -0.15228637
ssi     0.9678721  1.00000000 -0.09267421
sphrs2 -0.1522864 -0.09267421  1.00000000
```

The default option `"complete.obs"` ignores the entire row if missing values are present (listwise deletion). The alternative, `"pairwise.complete.obs"`, calculates correlations using pairwise deletion, which maximizes the number of non-missing values for each pair of variables.

```
> cor(gssdata[, c('sop','ssi','sphrs2')], use = 'pairwise.complete.obs')
               sop         ssi      sphrs2
sop     1.0000000  0.83315651 -0.15228637
ssi     0.8331565  1.00000000 -0.09267421
sphrs2 -0.1522864 -0.09267421  1.00000000
```

In this example, the correlation coefficient between (*sop*) and (*ssi*) is 0.833 and the significance of the coefficient is $< 2.2e-16$ (as shown in the cor.test() output below), which suggests that these two variables are significantly and positively correlated. We can calculate significance levels of pairwise correlations using `cor.test()`:

```
> cor.test(gssdata$sop, gssdata$ssi, method="pearson")

Pearson's product-moment correlation

data:  gssdata$sop and gssdata$ssi
t = 51.464, df = 1167, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8147408 0.8498930
sample estimates:
      cor
0.8331565
```

The above results report a p-value of < 0.001, suggesting this is a statistically significant positive correlation.

If our data contain rank-order or ordinal variables, we can report Spearman's rho correlation coefficients instead of the standard Pearson coefficients using the same functions discussed above:

```
> # correlation matrix with Spearman coefficients
> cor(gssdata[, c('educ','age','realrinc')], use =
'pairwise.complete.obs',
           method="spearman")
                educ          age  realrinc
educ      1.000000000 -0.006762367 0.3775991
age      -0.006762367  1.000000000 0.2257901
realrinc  0.377599065  0.225790092 1.0000000

> # Spearman correlation test
> cor.test(gssdata$age, gssdata$educ, method="spearman",exact=FALSE)

Spearman's rank correlation rho

data:  gssdata$age and gssdata$educ
S = 2144416617, p-value = 0.7438
```

```
    alternative hypothesis: true rho is not equal to 0
    sample estimates:
         rho
    -0.006762367
```

Exercise 4

Measure the association between the variables highest year of school completed (*educ*), respondent's father's occupational prestige index (*faos*) and respondent's mother's occupational prestige index (*moos*), once using pairwise deletion and again using listwise deletion to compare how the sample size changes. (See [Appendix A] for answers.)

Bivariate Regression

Bivariate regression is used to assess the effect of a continuous or dichotomous independent variable on a continuous dependent variable. Regression analysis is performed using the `lm()` function. The model object returns basic information about the model, including the model formula and coefficients. The `summary()` function (shown below) reports that same information as well as model fit statistics and significance values for the coefficients.

This example uses bivariate regression to measure the effect of respondent's age (*age*) on family's income (*realrinc*).

```
> model = lm(gssdata$realrinc ~ gssdata$age)
> model

Call:
lm(formula = gssdata$realrinc ~ gssdata$age)

Coefficients:
(Intercept)  gssdata$age
     8620.2        368.1

> summary(model)

Call:
lm(formula = gssdata$realrinc ~ gssdata$age)

Residuals:
   Min     1Q Median     3Q    Max
-39059 -15136  -6597   5459 131386

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8620.20    2516.28   3.426 0.000631 ***
gssdata$age   368.14      53.86   6.835 1.24e-11 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28450 on 1356 degrees of freedom
  (990 observations deleted due to missingness)
Multiple R-squared:  0.0333,    Adjusted R-squared:  0.03259
F-statistic: 46.71 on 1 and 1356 DF,  p-value: 1.239e-11
```

The model summary includes the following information: * **Coefficients** (parameter estimates): This table reports the values of the unstandardized regression coefficients (*Estimate*) and significance tests for the coefficients (`Pr(>|t|)`).

The regression coefficient for the constant is 8620.20 and the unstandardized coefficients for the independent variable *age* is 368.14. From this, it can be said that the predicted level of income will increase by $368.14 for each additional year of age. Because this effect is measured in the original scale of the dependent variable, it is referred to as an **unstandardized coefficient**. Thus, when you have multiple independent variables measured in the different units in the model, you cannot compare the effects of independent variables using unstandardized coefficients. This will be addressed further in the section on Multiple Regression.

The standard error (`Std.Error`) is used to calculate the t-statistic (β/SE). This statistic is used to test the null hypothesis that the unstandardized coefficient is equal to 0. In this example, age is statistically significant at the 0.000 level. This means that you can reject the null hypothesis that the age variable's coefficient is equal to 0 and conclude that age is a significant predictor of income.

Model Fit: R-squared, also known as the **coefficient of determination**, tells you the proportion of variation in the dependent variable explained by the independent variable. R-squared values range from 0 (no relationship) to 1 (perfect prediction of the dependent variable from the independent variable). In this example, 3.33

Adjusted R-squared is adjusted for the number of independent variables in the model. However, since there is only one independent variable in this example, the R-square and Adjusted R-squared are the same.

ANOVA statistics: The ANOVA statistics tell us if the independent variable has a significant linear relationship with the dependent variable. In this example, the F-statistic is statistically significant at $p < 0.001$. Thus, respondent's age is a significant predictor of income.

Exercise 5

Does respondent's age (*age*) have a significant linear relationship with performance on a vocabulary test (*wordsum*)? Treat *wordsum* as the dependent variable. (See Appendix A for answers.)

Multiple Regression

Multiple linear regression is used to assess the effect of more than one continuous or dichotomous independent variable on a continuous dependent variable. Multiple regression is the same

as bivariate regression, except it uses two or more independent variables to explain the variation of the dependent variable. Regression analysis is performed using the function `lm()`.

This example uses multiple regression to measure the effect of respondent's occupational prestige score (*pres*), and age (*age*) on family's income (*realrinc*).

```
> model2 <- lm(realrinc ~ age + pres, data=gssdata)
> model2

Call:
lm(formula = realrinc ~ age + pres, data = gssdata)

Coefficients:
(Intercept)          age         pres
   -16646.4        285.3        635.1

> summary(model2)

Call:
lm(formula = realrinc ~ age + pres, data = gssdata)

Residuals:
   Min     1Q Median     3Q    Max
-49070 -13429  -5297   5416 142313

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -16646.42    3186.71   -5.224  2.03e-07 ***
age            285.29      51.55    5.534  3.76e-08 ***
pres           635.07      53.00   11.981   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26920 on 1348 degrees of freedom
  (997 observations deleted due to missingness)
Multiple R-squared:  0.1263,    Adjusted R-squared:  0.125
F-statistic: 97.45 on 2 and 1348 DF,  p-value: < 2.2e-16
```

The model summary includes the following information: * **Coefficients** (parameter estimates): This table reports the values of the unstandardized regression coefficients (*Estimate*) and significance tests for the coefficients (`Pr(>|t|)`).

This shows that a unit increase in *age* will increase predicted income by $285.29 while holding the occupational prestige score constant. The t-test statistic is 5.534, and the p-value associated with the t-test statistic is 0.000. Thus, *age* is statistically significant at the 0.000 level while controlling for *pres*.

To compare the magnitude of the effect of each independent variable, the coefficients must be standardized using the same scale. The standardized coefficient (**Beta**) measures change in the dependent variable (measured in standard deviations) per one standard deviation change in the independent variable. In R, we can calculate standardized beta coefficients with the `lm.beta()` function in the **lm.beta** package.

```
> library(lm.beta)

> model2_std <- lm.beta(model2)
> summary(model2_std)

Call:
lm(formula = realrinc ~ age + pres, data = gssdata)

Residuals:
   Min     1Q   Median     3Q     Max
-49070  -13429  -5297    5416  142313

Coefficients:
             Estimate    Standardized Std. Error  t value Pr(>|t|)
(Intercept) -1.665e+04           NA    3.187e+03  -5.224  2.03e-07 ***
age          2.853e+02    1.421e-01    5.155e+01   5.534  3.76e-08 ***
pres         6.351e+02    3.077e-01    5.300e+01  11.981   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26920 on 1348 degrees of freedom
  (997 observations deleted due to missingness)
Multiple R-squared:  0.1263,    Adjusted R-squared:  0.125
F-statistic: 97.45 on 2 and 1348 DF,  p-value: < 2.2e-16
```

Predicted income will increase by 0.1421 standard deviation units (from the Standardized column) for each one standard deviation increase in *age*, while it will increase 0.3077 standard deviation units for each one standard deviation increase in *pres*. Because the standardized coefficient for age is larger than that for *pres* we can conclude that the occupational prestige score has a larger impact on real income than age.

Model Fit: Since the denominator of **R-squared** is fixed, each additional variable used in the model can only increase the size of the numerator. As a result, the introduction of additional variables produces a higher R-squared value, regardless of the efficiency of the model.

Since R-squared will never decrease by adding more independent variables into the model, some researchers report an **Adjusted R-square**, which corrects this bias by adjusting both the numerator and the denominator by their respective degrees of freedom. Adjusted R-square is caclulated below, where n is the number of observations and k is the number of independent variables.

$$\text{Adjusted R-Square} = 1 - (1 - R)^2 \times \frac{n-1}{n-k-1}$$

Unlike R Square, the Adjusted R Square can decline in value. It is important to keep in mind, however, that the R Square value is a proportion, while the Adjusted R Square is not. Therefore, after adding one independent variable, the Adjusted R Square may decrease while the R Square may increase. If the Adjusted R Square is much lower than the R Square, this usually indicates that some relevant explanatory variables are not specified in the model.

In this example, compared to the bivariate model in the previous section, both R Square and Adjusted R Square have increased. Hence, this new model explains more variation of the dependent variable than the old model with one independent variable.

- The **ANOVA** table measures whether the independent variables significantly predicts the dependent variable. In this example, the **F** statistic is below 0.05. Thus, age and respondent's occupational prestige score are significant predictors of income.

Exercise 6

Does the occupational prestige of a respondent's parents (*faos* and *moos*) and respondent's sex (*Male*) affect the respondent's occupational prestige (*pres*)? (See Appendix A for answers.)

Appendix A: Answers for Exercises

R2_script_complete.R

Exercise 1

Perform a chi-square test to determine if educational attainment (*degree*) is significantly associated with the belief that marijuana should be legal (*grass*).

(a) How many high school graduates said "Yes" to legalization?

(b) Is there a statistically significant relationship between the two variables? Which chi-square test statistic is appropriate?

```
> table(gssdata$grass, gssdata$degree)

           bachelor graduate high school junior college lt high school
  legal          201       99         466             85             87
  not legal       87       63         239             43             77

> chisq.test(gssdata$grass, gssdata$degree)

Pearson's Chi-squared test

data:  gssdata$grass and gssdata$degree
X-squared = 14.712, df = 4, p-value = 0.005337
```

Check expected cell values to confirm that chi-square is most appropriate statistic:

```
> chisq.test(gssdata$grass, gssdata$degree)$expected
             gssdata$degree
gssdata$grass lt high school high school junior college bachelor  graduate
    legal          106.31099    457.0076       82.97443 186.6925 105.01451
    not legal       57.68901    247.9924       45.02557 101.3075  56.98549$
```

1. 466 high school graduates said Yes to legalization.
2. All expected cells are larger than 5, so we should use Pearson Chi-Square. The p-value associated with the Pearson chi-square statistic (0.005) is less than 0.05. Thus, there is a statistically significant relationship between the two variables.

Exercise 2

Find out if there is a statistically significant difference in terms of respondent's income `(realrinc)` between respondents who took college-level science courses and those who didn't `(colsci)`.

```
> leveneTest(gssdata$realrinc, gssdata$colsci)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  21.273 4.762e-06 ***
      677
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Because the Equality of Variances test is significant ($p < 0.001$), an **equal variances not assumed** test statistic is appropriate to use.

```
> t.test(gssdata$realrinc ~ gssdata$colsci, var.equal = FALSE)

    Welch Two Sample t-test

data:  gssdata$realrinc by gssdata$colsci
t = -5.9769, df = 557.17, p-value = 4.061e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16522.316  -8348.769
sample estimates:
 mean in group no mean in group yes
         17721.91          30157.45
```

The t-test statistic (-5.9769) is significant ($p < 0.05$), therefore the amount of income respondents made significantly differs based on if they took or didn't take college-level science courses.

Exercise 3

Using the One-Way ANOVA procedure, find out whether the mean level of income (**realrinc**) is different among different education groups (**degree**). Remember to check the Test of Homogeneity of Variances to determine if the Welch test is required.

```
> leveneTest(gssdata$realrinc, gssdata$degree)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    4  21.459 < 2.2e-16 ***
      1358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # unequal variance
> oneway.test(gssdata$realrinc ~ gssdata$degree)

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

data:  gssdata$realrinc and gssdata$degree
F = 46.45, num df = 4.00, denom df = 435.85, p-value < 2.2e-16
```

Because Levene's test is significant (i.e. equal variances not assumed), we should report the Welch test statistic. We can conclude that differences in average income are significantly different among educational groups at the $p < 0.001$ level.

If we want to run post hoc tests to examine pairwise differences between groups, the Games-Howell test is appropriate for this example. The `games_howell_test()` function in the `rstatix` package can compute the Games-Howell test.

```
  > games_howell_test(gssdata, realrinc ~ degree)
  # A tibble: 10 x 8
  .y.     group1    group2   estimate conf.low conf.high    p.adj p.adj.signif
  * <chr>   <chr>     <chr>       <dbl>    <dbl>     <dbl>    <dbl>
<chr>
  1 realri~ bachelor  graduate   10704.     644.    20763. 3.10e- 2
*
  2 realri~ bachelor  high sc~  -16578.  -22384.   -10771. 0.
****
  3 realri~ bachelor  junior ~  -16346.  -22507.   -10185. 0.
****
  4 realri~ bachelor  lt high~  -23005.  -29040.   -16971. 0.
****
  5 realri~ graduate  high sc~  -27281.  -36182.   -18380. 7.95e-14
****
  6 realri~ graduate  junior ~  -27050.  -36181.   -17918. 3.69e-13
****
  7 realri~ graduate  lt high~  -33709.  -42757.   -24661. 4.59e-14
****
  8 realri~ high sch~ junior ~     232.   -3720.     4183. 1.00e+ 0
ns
  9 realri~ high sch~ lt high~   -6428.  -10178.    -2678. 3.87e- 5
```

```
****
  10 realri~ junior c~ lt high~   -6659.  -10945.    -2374. 2.70e- 4 ***
```

- Group Less than High School is significantly different from all other groups at the 0.025 level. Group High School is significantly different from all other groups except group Junior College at the 0.025 level.
- More specifically, those with a high school degree make 6428 more than those with less than a high school education, but the 232 difference between those with a high school degree and those with a junior college education is not significant.
- Group Junior college is significantly different from all other groups except group High school at the 0.025 level.
- Group Bachelor is significantly different from all other groups except group Graduate (and vice versa) at the 0.025 level.

Exercise 4

Measure the association between the variables highest year of school completed (*educ*), respondent's father's occupational prestige index (*faos*) and respondent's mother's occupational prestige index (*moos*), once using pairwise deletion and again using listwise deletion to compare how the sample size changes.

Using pairwise deletion:

```
> cor(gssdata[,c('educ','faos','moos')], use = 'pairwise.complete.obs')
          educ      faos      moos
educ 1.0000000 0.2614165 0.2691149
faos 0.2614165 1.0000000 0.2357500
moos 0.2691149 0.2357500 1.0000000
```

Using listwise deletion:

```
> cor(gssdata[,c('educ','faos','moos')], use = 'complete.obs')
          educ      faos      moos
educ 1.0000000 0.2470015 0.2424426
faos 0.2470015 1.0000000 0.2357485
moos 0.2424426 0.2357485 1.0000000
```

To calculate p-values for all pairs at once, we can use the `corr.test()` function in the `psych` package. The Holm correction for multiple comparisons is used by default; to turn this off, change the `adjust` parameter to `"none"`.

```
    > library(psych)

    > corr.test(gssdata[,c('educ','faos','moos')],
Call:corr.test(x = gssdata[, c("educ", "faos", "moos")], use = "pairwise",
```

```
    adjust = "holm")
Correlation matrix
     educ faos moos
educ 1.00 0.26 0.27
faos 0.26 1.00 0.24
moos 0.27 0.24 1.00
Sample Size
     educ faos moos
educ 2345 1840 1656
faos 1840 1842 1226
moos 1656 1226 1657
Probability values (Entries above the diagonal are adjusted for multiple
tests.)
     educ faos moos
educ    0    0    0
faos    0    0    0
moos    0    0    0

 To see confidence intervals of the correlations, print with the short=FALSE
option
```

All pairwise comparisons show positive correlations that are significant at the $p < 0.001$ level.

Exercise 5

Does respondent's age (*age*) have a significant linear relationship with performance on a vocabulary test (*wordsum*)? Treat *wordsum* as the dependent variable.

```
    > model = lm(gssdata$wordsum ~ gssdata$age)
    > summary(model)

Call:
lm(formula = gssdata$wordsum ~ gssdata$age)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2092 -1.0844  0.0943  1.2607  4.2901

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.426053   0.152916  35.484  < 2e-16 ***
gssdata$age 0.009789   0.002880   3.399 0.000694 ***
---$
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.017 on 1538 degrees of freedom
  (808 observations deleted due to missingness)
```

```
Multiple R-squared:  0.007456,  Adjusted R-squared:  0.00681
F-statistic: 11.55 on 1 and 1538 DF,  p-value: 0.0006937
```

- 8.6
- Because the F statistic is statistically significant, age is a significant predictor of vocabulary test score.
- The regression coefficient for the constant is 5.426, and the one for the independent variable age is 0.010. The predicted vocabulary test score will increase by 0.010 for each additional year of age.
- The effect of age on wordsum is statistically significant at the 0.01 level.

Exercise 6

Does the occupational prestige of a respondent's parents (*faos* and father's occupational prestige score *moos*) and respondent's sex (*Male*) affect the respondent's occupational prestige (*pres*)?

```
> model <- lm(pres ~ faos + moos + Male, data=gssdata)
> summary(model)

Call:
lm(formula = pres ~ faos + moos + Male, data = gssdata)

Residuals:
    Min      1Q  Median      3Q     Max
-33.101 -10.103  -0.247   9.772  38.485

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.88052    1.69194  19.434  < 2e-16 ***
faos         0.15362    0.03026   5.077 4.46e-07 ***
moos         0.15870    0.02989   5.309 1.32e-07 ***
Malemale    -0.24371    0.76341  -0.319     0.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.13 on 1187 degrees of freedom
  (1157 observations deleted due to missingness)
Multiple R-squared:  0.05578,   Adjusted R-squared:  0.0534
F-statistic: 23.38 on 3 and 1187 DF,  p-value: 1.056e-14
```

- Because the F statistic is statistically significant, mother's occupational prestige score, father's occupational prestige score, and sex are significant predictors of respondent's occupational prestige score.
- 5.578 percent of the variation in personal prestige is explained by parents' prestige and respondent's sex.

- The predicted occupational prestige score will increase by 0.159 for each unit increase of mother's occupational prestige score, 0.154 for each unit increase of father's occupational prestige score, and −0.244 for being male.
- The effects of moos on faos are statistically significant at the 0.001 level. The effect of sex (male) is not statistically significant.

To report standardized coefficients:

```
> lmmodel = lm.beta(model)
> summary(lmmodel)

Call:
lm(formula = pres ~ faos + moos + Male, data = gssdata)

Residuals:
    Min      1Q  Median      3Q     Max
-33.101 -10.103  -0.247   9.772  38.485

Coefficients:
             Estimate Standardized Std. Error t value Pr(>|t|)
(Intercept) 32.880516     0.000000   1.691936  19.434  < 2e-16 ***
faos         0.153620     0.147236   0.030261   5.077 4.46e-07 ***
moos         0.158701     0.153942   0.029894   5.309 1.32e-07 ***
Malemale    -0.243712    -0.009011   0.763407  -0.319     0.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.13 on 1187 degrees of freedom
  (1157 observations deleted due to missingness)
Multiple R-squared:  0.05578,   Adjusted R-squared:  0.0534
F-statistic: 23.38 on 3 and 1187 DF,  p-value: 1.056e-14
```

Standardized beta coefficients reveal that the father's prestige score has the largest effect on respondent's prestige score.