Chris Chapman and Elea McDonnell Feit

EXERCISES
from *R for Marketing Research and Analytics, 2nd ed.*

April 10, 2019

Springer

1

Welcome to R *(Exercises only)*

There are no exercises for Chapter 1. Just install R and RStudio and read the chapter. Welcome to R!

2

An Overview of the R Language *(Exercises only)*

2.11 Exercises

2.11.1 Preliminary Note on Exercises

The exercises in each chapter are designed to reinforce the material. They are provided primarily for classroom usage but are also useful for self-study. On the book's website, we provide R files with example solutions at `http://r-marketing.r-forge.r-project.org/exercises`.

We strongly encourage you to complete exercises using a tool for *reproducible results*, so the code and R results will be shown together in a single document. If you are using RStudio, an easy solution is to use an *R Notebook*; see Appendix B for a brief overview of R Notebooks and other options. A simple R Notebook for classroom exercises is available at the book's website noted above.

For each answer, do not simply determine the answer and report it; instead write R code to find the answer. For example, suppose a question could be answered by copying two or more values from a `summary` command, and pasting them into the R console to compute their difference. Better programming practice is to write a command that finds the two values and then subtracts them with no additional requirement for you to copy or retype them. Why is that better? Although it may be more difficult to do once, it is more generalizable and reusable, if you needed to do the same procedure again. At this point, that is not so important, but as your analyses become complex, it will be important to eliminate manual steps that may lead to errors.

Before you begin, we would reemphasize a point noted in Section 2.7.2: there may be many ways to solve a problem in R. As the book progresses, we will demonstrate progressively better ways to solve some of the same problems. And R programmers may differ as to what constitutes "better." Some may prefer elegance while others prefer speed or ease of comprehension. At this point, we recommend that you *consider* whether a solution seems optimal, but don't worry too much about it. Getting a correct answer in any one of multiple possible ways is the most important outcome.

In various chapters the exercises build on one another sequentially; you may need to complete previous exercises in the chapter to answer later ones. Exercises preceded by an asterisk (*) correspond to one of the optional sections in a chapter.

2.11.2 Exercises

1. Create a text vector called `Months` with names of the 12 months of the year.
2. Create a numeric vector `Summer`, with Calendar month index positions for the summer months (inclusive, with 4 elements in all).
3. Use vector indexing to extract the text values of `Months`, indexed by `Summer`.
4. Multiply `Summer` by `3`. What are the values of `Months`, when indexed by `Summer` multiplied by `3`? Why do you get that answer?
5. What is the mean (average) summer month, as an integer value? Which value of `Months` corresponds to it? Why do you get that answer?
6. Use the `floor()` and `ceiling()` functions to return the upper and lower limits of `Months` for the average `Summer` month. (Hint: to find out how a function works, use R help if needed.)
7. Using the `store.df` data from Section 2.5, how many visits did Bert's store have?
8. It is easy to make mistakes in indexing. How can you confirm that the previous answer is actually from Bert's store? Show this with a command that produces no more than 1 row of console output.
9. *Write a function called `PieArea` that takes the length of a slice of pie and returns the area of the whole pie. (Assume that the pie is cut precisely, and the length of the slice is, in fact, the radius of the pie.) Note that `^` is the exponentiation operator in R.
10. *What is `PieArea` for slices with lengths 4.0, 4.5, 5.0, and 6.0?
11. *Rewrite the previous command as one line of code, without using the `PieArea()` function. Which of the two solutions do you prefer, and why?

3

Describing Data *(Exercises only)*)

3.8 Exercises

3.8.1 E-commerce Data for Exercises

Starting in this chapter, many of our exercises use a real data set contributed to the authors by an e-commerce site. The data set comprises responses to intercept surveys asked when users visited the site, along with data about each user's site activity such as number of pages visited and whether a sale was completed. Identifying details for the site and customers have been removed but the observations otherwise are actual data.

We will load the data set first, and then explain a few of its observations. To load the data from CSV format, use the following command (or load `ecommerce-data.csv` from a local location if you have downloaded it, as noted in Section 1.6.3).

```
> ecomm.df <- read.csv("https://goo.gl/hzRyFd")
> summary(ecomm.df)
```

As a reminder, Section 2.11 discussed our general approach and recommendations for exercises.

3.8.2 Exercises

1. How many observations and variables are in the e-commerce data set?
2. Compute a frequency table for the country of origin for site visits. After the United States, which country had the most visitors?
3. Compute a two-way frequency table for the intent to purchase (`intentWasPlanningToBuy`), broken out by user profile.
4. What are the proportions of parents who intended to purchase? the proportions of teachers who did? For each one, omit observations for whom the intent is unknown (blank).
5. Among US states (recorded in the variable `region`), which state had the most visitors and how many?

6. Solve the previous problem for the state with the most visitors, using the `which.max()` function (or repeat the same answer, if you already used it).
7. Draw a histogram for the number of visits to the site (`behavNumVisits`). Adjust it for more detail in the lower values. Color the bars and add a density line.
8. Draw a horizontal boxplot for the number of site visits.
9. Which chart from the previous two exercises, a histogram or a boxplot, is more useful to you, and why?
10. Draw a boxplot for site visits broken out with a unique row for each profile type. (Note: if the chart margins make it unreadable, try the following command before plotting: `par(mar=c(3, 12, 2, 2))`. After plotting, you can use the command `par(mar=c(5, 4, 4, 2) + 0.1)` to reset the chart margins.)
11. *Write a function called `MeanMedDiff` that returns the absolute difference between the mean and the median of a vector.
12. *What is the mean-median difference for number of site visits?
13. *What is the mean-median difference for site visits, after excluding the person who had the most visits?
14. *Use the `apply()` function to find the mean-median difference for the 1/0 coded behavioral variables for onsite behaviors.
15. *Write the previous command using an anonymous function (see Section 2.7.2) instead of `MeanMedDiff()`.
16. *Do you prefer the named function for mean-median difference (`MeanMaxDiff()`), or an anonymous function? Why? What is a situation for each in which it might be preferable?

4

Relationships Between Continuous Variables *(Exercises only)*)

4.10 Exercises

The following exercises use the e-commerce data set as described in Section 3.8.1.

1. The e-commerce data set (Section 3.8.1) includes the number of visits a user made to the site (`behavNumVisits`). Plot this using a histogram, and then again by plotting a table of frequencies. Which plot is a better starting place for visualization, and why?
2. Adjust the table plot from the previous exercise to improve it. Use logarithmic values for the numbers of visits instead of raw counts, and add a chart title and axis labels.
3. The default Y axis on the previous plot is somewhat misleading. Why? Remove the default Y axis, and replace it with better labels. (Note: for logarithmic values, labels that begin with digits 1, 2, and 5 — such as 1, 2, 5, 10, 20, 50, etc. — may be useful.) Make the Y axis readable for all labels.
4. The variable `behavPageViews` is a factor variable, but we might like to do computations on the number of views. Create a new variable `pageViewInt` that is an integer estimate of the number of page views for each row, and add it to `ecomm.df`. Be conservative with the estimates; for example, when the data say ”10+” views, code only as many as are indicated with confidence.
5. Plot a histogram of the newly added integer estimate of page views (`pageViewInt`).

Site visits and page views. For the next several exercises, we consider whether frequent visitors are likely to view more pages on the site. It is plausible to think that frequent visitors might view more pages in a session because they are more engaged users, or that frequent visitors would view fewer pages because they are more familiar with the site. We will see what the data suggest.

6. For a first exploration, make a scatterplot for the integer estimate of page views vs. the number of site visits. Should number of visits be on a log scale? Why or why not?
7. There are only a few values of X and Y in the previous plot. Adjust the plot to visualize more clearly the frequencies occurring at each point on the plot.
8. What is the Pearson’s *r* correlation coefficient between number of visits and the integer estimate of page views? What is the correlation if you use log of visits instead?

9. Is the correlation from the previous exercise statistically significant?

10. Is Pearson's *r* a good estimate for the relationship of these two variables? Why or why not?

11. *What is the polychoric correlation coefficient between number of visits and integer page views? Is it a better estimate than Pearson's *r* in this case?

12. Overall, what do you conclude about the relationship between the number of times a user has visited the site and the number of page views in a given session?

Salaries data. For the remaining exercises, we use the `Salaries` data from the `car` package.

13. How do you load the `Salaries` data from the car package? (Hint: review the `data()` function.) Within R itself, how can you find out more detail about the `Salaries` data set?

14. Using the `Salaries` data, create scatterplot matrix plots using two different plotting functions. Which do you prefer and why?

15. Which are the numeric variables in the `Salaries` data set? Create a correlation plot for them, with correlation coefficients in one area of the plot. Which two variables are most closely related?

5

Comparing Groups: Tables and Visualizations *(Exercises only)*

5.6 Exercises

The following exercises use the e-commerce data set as described in Section 3.8.1.

1. Using the integer approximation of page views (see Exercises in Section 4.10), describe page views for parents, teachers, and health professionals. Use a `by()` or `aggregate()` function as appropriate.
2. Repeat the previous task, this time using a `for()` loop to iterate over the groups.
3. Comparing the previous two approaches — grouping vs. a `for()` loop — which do you prefer, and why? What is a time when the other approach might be preferable?
4. What are the proportions of men and women among the various visitor profiles (teacher, parent, relative, etc.)? For this question, don't count observations where the gender is not specified as male or female.
5. Considering parents, teachers, and health professionals, which group has made the most purchases recently? Answer with both descriptives and a visualization.
6. In answering the previous question, you might use either counts or proportions. Do they give you the same answer? If not, show an example. What is a business question for which counts would be preferable? What is a question for which proportions would be preferable?
7. When we split the profiles into men and women, and consider completed purchases on the site (variable `behavAnySale`) which combination of profile and gender made the highest number of purchases? Which had the highest rate of purchase, relative to total number of observations?

6

Comparing Groups: Statistical Tests *(Exercises only)*

6.9 Exercises

The following exercises use the e-commerce data set as described in Section 3.8.1.

1. Among Teachers and Parents who visited the site, which group was more likely to know the product of interest in advance (variable productKnewWhatWanted)? Answer with both descriptive statistics and visualization.
2. In the previous exercise, should you limit observations to just those with product knowledge of “Yes” or “No”? Why or why not? How does it change the result?
3. Is the difference in prior product knowledge (variable `productKnewWhatWanted`) statistically significantly different for teachers vs. parents? (*Hint*: make a table of counts, and then select only the rows and columns needed for testing.)
4. Using the integer approximation of page views (see Exercises in Section 4.10), describe page views for parents, teachers, and health professionals. Use a `by()` or `aggregate()` function as appropriate.
5. What is the proportion of teachers who had prior product knowledge, and what is the proportion for parents?
6. Suppose we believe that the parent proportion in the previous exercise is the true value for both parents and teachers. How do we compare the observed proportion for teachers to that? Is is statistically significantly different? What is the 95 percent confidence interval for the observations among teachers?
7. Using the integer approximation of page views (see Exercises in Section 4.10), compare the mean number of page views for Parents and Teachers. Which is higher? Is the difference statistically significant? What is the confidence interval for the difference?
8. Compare estimated page views (variable `pageViewInt`) for all profile groups. Are the groups statistically significantly different? Answer and visualize the differences.
9. Repeat the previous exercise, and limit the data to just Parents and Teachers. Explain and visualize. Is the answer different than in the previous exercise? Why?

10. *Repeat the previous comparison for page views among just Teachers and Parents, using a Bayesian Analysis of Variance. Report the statistics and visualize it. Is the answer the same or different as obtained from classical ANOVA?

11. *Write a function of your own to compute proportions from a table of frequency counts. Compare your code to that in `prop.table()`. (Don't forget that you can see the code for most functions by typing the name of the function into the command line.)

7

Identifying Drivers of Outcomes: Linear Models *(Exercises only)*

7.9 Exercises

7.9.1 Simulated Hotel Satisfaction and Account Data

For these and some later exercises, we use a simulated dataset for a hotel. The data combine customers' responses to a satisfaction survey with basic account information from their hotel stays. These are the sort of data that you might acquire from an email survey is sent to users, where a disguised identifier links the surveys responses to account data. Another common source of similar data is an online system where a pop-up survey asks satisfaction questions, and the answers can be related to the user's account (the real data set referenced in Section 3.8.1 is an example).

To access the hotel data set, load the data from CSV format online as follows, or load from a local location as file `hotelsat-data.csv` if you have already downloaded it (see Section 1.6.3).

```
> hotel.df <- read.csv("https://goo.gl/oaWKgt")
> summary(hotel.df)
```

These data include 18 items asking about satisfaction with various aspects of the hotel (cleanliness, dining experience, staff, satisfaction with elite status perks, and so forth), each on a 7 point rating scale. (In reality, we would rarely recommend asking 18 separate satisfaction items! However, we will use all of them for some investigations in later chapters.) In addition to the survey responses, the data include each respondent's corresponding number of nights stayed at the hotel, the distance traveled, reason for visiting, their elite membership level, and the average amounts spent per night on the room, dining, and WiFi.

7.9.2 Exercises

1. Visualize the distributions of the variables in the hotel satisfaction data. Are there variables that might be understood better if they are transformed? Which variables and what transforms would you apply? (Suggestion: for efficiency, it may help to divide the data set into smaller sets of similar variables.)
2. What are the patterns of correlations in the data? Briefly summarize any patterns you observe, in 2-4 sentences.

3. Consider just the three items for cleanliness (`satCleanRoom`, `satCleanBath`, and `satCleanCommon`). What are the correlation coefficients among those items? Is there a better measure than Pearson's *r* for those coefficients, and why? Does it make a difference in these data? (Consider the notes in Section **??**.)

4. Management wants to know whether satisfaction with elite membership perks (`satPerks`) predicts overall satisfaction (`satOverall`). Assume that `satPerks` is a predictor and we want to know how `satOverall` is associated with changes in it. How do you interpret the relationship?

5. We might wish to control the previous `satPerks` model for other influences, such as satisfaction with the Front Staff (`satFrontStaff`) and with the city location (`satCity`). How do you change the previous model to do this? Model and interpret the result. Is the answer different than in the model with only Perks? Why or why not?

6. Suppose we have a business strategy to maximize satisfaction with elite recognition (`satRecognition`) among our Gold and Platinum elite members. To do so, we might invest more in the front staff, room cleanliness, the points that we award elite members, or the membership perks given to them. Which of those strategies might we want to consider first, according to these data, if we wish to increase Gold and Platinum member satisfaction with elite recognition?

7. What are some problems with using the present data to answer that strategic question? What data would you need to give a better answer?

8. Considering the results in the previous question, would you recommend to invest more in room cleanliness? Why or why not?

9. Now we are examining ways to improve revenue in the restaurant. Management wants to understand the relationship of average food spend per night with elite status (eliteStatus) and satisfaction with food price (`satDiningPrice`). Model this and interpret it.

10. How does satisfaction relate to spending in our restaurant? On one side, we might expect dining satisfaction to be higher when food costs less, because customers are often happy about lower prices. However, we might also expect the exact opposite relationship, where satisfied diners spend more. Which relationship is better supported by these data?

11. Plot the predicted food spend per night in dollars, as a function of nights stayed. (Suggestion: fit a linear model with one predictor.) In our data, no one stayed 40 nights. But if someone had, what would be a good guess as to their average food spend per night?

12. Is the association between nights spent and spending on food different among Platinum elite members? Visualize the difference. What does this suggest for a restaurant strategy? Is this consistent with findings in the previous models (Exercises 9–11 above)?

13. Fit the elite recognition model (Exercise 6 above) using Bayesian regression. Which variables are most associated with members' satisfaction with recognition?

14. How do those Bayesian coefficient estimates compare to the classical linear model estimates in Exercise 6? Visualize the relationship among the coefficients from each. What is the correlation coefficient?

15. Which model do you prefer, classical or Bayesian? Why?

8

Reducing Data Complexity *(Exercises only)*

8.8 Exercises

8.8.1 PRST Brand Data

For these exercises (and the exercises in Chapter 10), we use a simulated data set for four fictitious consumer electronic device brands: Papa, Romeo, Sierra, and Tango (abbreviated PRST). The brands have been rated by consumers on nine adjectives, each using a 7-point rating scale. The adjectives are “Adaptable,” “Best Value,” “Cutting Edge,” “Delightful,” “Exciting,” “Friendly,” “Generous,” “Helpful,” and “Intuitive.” You will examine the relationships among the adjectives and the brands, considering both the statistical analyses and possible brand strategy.

First we load the data from the web site, or from a local file (change the directory as needed for your system):

```
> prst1 <- read.csv("https://goo.gl/z5P8ce")  # web site
# prst1 <- read.csv("chapter8-brands1.csv")   # or a local file
> summary(prst1)
   Adaptable       BestValue      CuttingEdge     Delightful
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:3.00   1st Qu.:3.000
 Median :4.000   Median :4.000   Median :4.00   Median :4.000
 Mean   :4.255   Mean   :3.849   Mean   :4.07   Mean   :3.983
...
```

8.8.2 Exercises

Basic Concepts

1. Summarize the PRST data. Should the data be rescaled?
2. Rescale the PRST data with a “Z score” procedure and examine the rescaled data. Does this confirm your decision in the previous exercise about whether to rescale the data?

3. Plot a correlation matrix for the adjective ratings. How many factors does it suggest?
4. Aggregate the mean of each adjective rating by brand. Plot a heatmap for the mean ratings by brand.

Principal Components Analysis

5. Extract the principal components in the PRST data. How many components are needed to explain the majority of variance in the PRST data? Visualize that.
6. Using principal components for the mean adjective ratings, plot the brands against the first two components. How do you interpret that? Now plot against the second and third components (hint: see `?biplot.princomp`). Does this change your interpretation? What does this tell you about interpreting PCA results?
7. *(Thought exercise without code.)* Suppose you are the brand manager for Sierra, and you wish to change your position vs. the market leader, Tango. What are some strategies suggested by the PCA positions?

Exploratory Factor Analysis

8. Consider an exploratory factor analysis (EFA) for the PRST adjective ratings. How many factors should we extract?
9. Find an EFA solution for the PRST data with an appropriate number of factors and rotation. What factor rotation did you select and why?
10. Draw a heatmap of the EFA factor loadings. Also draw a path diagram for the EFA solution.
11. Find the mean factor scores for each brand and plot a heatmap of them.
12. *(Thought exercise without code.)* Compare the factor score heatmap for PRST brands to the PCA interpretations in Exercise 6 above. Does the heatmap suggest different directions for the brand strategy for Sierra vs. Tango?

Multidimensional Scaling

13. Plot a multidimensional scaling (MDS) map for the PRST brands using the mean adjective ratings. Which brands are most similar and most different?
14. *(Thought exercise without code.)* How does the MDS map relate to the PCA and EFA positions in the exercises above? What does it suggest for the strategy you considered in Exercise 6 above?

9

Additional Linear Modeling Topics *(Exercises only)*

9.9 Exercises

9.9.1 Online Visits and Sales Data for Exercises

For exercises regarding collinearity and logistic regression, we will use a simulated data set that represents customer transactions together with satisfaction data, for web site visits and purchases. The variables are described in Table 9.1. We load the data locally or from the online site:

```
> # sales.data.raw <- read.csv("chapter9-sales.csv") # local
> sales.data.raw <- read.csv("https://goo.gl/4Akgkt") # online
> summary(sales.data.raw)
    acctAge       visitsMonth      spendToDate       spendMonth      ...
 Min.   : 1.00   Min.   : 1.000   Min.   :   6.0   Min.   :   4.0   ...
 1st Qu.: 8.00   1st Qu.: 6.000   1st Qu.:  28.0   1st Qu.:   9.0   ...
 Median :13.00   Median : 7.000   Median :  45.0   Median :  17.0   ...
```

Variable	**Description**	**Variable**	**Description**
`acctAge`	Tenure of the customer, in months	`visitsMonth`	Visits to the web site, in the most recent month
`spendToDate`	Customer’s total lifetime spending	`spendMonth`	Spending, most recent month
`satSite`	1-10 satisfaction rating with the web site	`satQuality`	Rating for satisfaction with product quality
`satPrice`	Rating for satisfaction with prices	`satOverall`	Overall satisfaction rating
`region`	US geographic region	`coupon`	Whether coupon was sent to them for a particular promoted product
`purchase`	Whether they purchased the promoted product (with or without coupon)		

Table 9.1: Variables in the `chaper9-sales` data set.

9.9.2 Exercises for Collinearity and Logistic Regression

Collinearity

1. In the sales data, predict the recent month's spending (`spendMonth`) on the basis of the other variables using a linear model. Are there any concerns with the model? If so, fix them and try the prediction again.
2. How does the prediction of the recent month's sales change when the variables are optimally transformed? Which model – transformed or not – is more interpretable?
3. Fit the linear model again, using a principal component extraction for satisfaction. What is the primary difference in the estimates from the previous models?
4. *(Thought exercise without code.)* When the model is fit with `region` as a predictor, it may show the West region with a large – possibly even the largest – effect. Yet it is not statistically significant whereas smaller effects are. Why could that be?

Logistic Regression

5. Using logistic regression, what is the relationship between the coupon being sent to some customers and whether the purchased the promoted product?
6. How does that model change if region, satisfaction, and total spending are added as predictors?
7. Is there an interaction between the coupon and satisfaction, in their relationship to purchase of the promoted product?
8. What is the best estimate for how much a coupon is related to increased purchase, as an odds ratio? Explain the meaning of this odds ratio using non-technical language.
9. What is the change in purchase likelihood, in relation to a change of 1 unit of satisfaction? (Hint: what is a unit of satisfaction in the model?) Approximately how many points would "1 unit" be, on the survey's 1-10 rating scale?
10. *(Thought exercise without code.)* Considering the product strategy, what questions are suggested by the apparent relationship between satisfaction and purchase? What possible explanations are there, or what else would you wish to know?

9.9.3 Handbag Conjoint Analysis Data for Exercises

In the remaining exercises, we consider a metric (ratings-based) conjoint exercise for handbags, using a new data set. Each of 300 simulated respondents rated the likelihood to purchase each of 15 handbags, which varied according to Color (black, navy blue, and gray), Leather finish (matte or shiny patent), Zipper color (gold or silver), and Price ($15, $17, $19, or $20). We load the data:

```
> # conjoint.df <- read.csv("chapter9-bag.csv") # local
> conjoint.df <- read.csv("https://goo.gl/gEKSQt") # online
> summary(conjoint.df)
    resp.id           rating           price         color     ...
 Min.   :  1.00   Min.   : 2.000   Min.   :15.00   black: 900  ...
 1st Qu.: 75.75   1st Qu.: 4.000   1st Qu.:15.00   gray :1500  ...
```

9.9.4 Exercises for Metric Conjoint and Hierarchical Linear Models

11. Using the handbag data, estimate the likelihood to purchase as a function of the handbags' attributes, using a simple linear model.
12. Now fit the ratings conjoint model as a classical hierarchical model, fitting individual level estimates for each attribute's utility.
13. What is the estimated rating for a black bag with matte finish and a gold zipper, priced at $15? (Careful!)
14. Which respondents are most and least interested in a navy handbag?
15. Fit the hierarchical model again, using a Bayesian MCMC approach. How do the upper level estimates compare with those from the classical model?

10

Confirmatory Factor Analysis and Structural Equation Modeling *(Exercises only)*

10.7 Exercises

10.7.1 Brand Data for Confirmatory Factor Analysis Exercises

For the CFA exercises, we will use a second simulated sample for “PRST” ratings (see Section 8.8). The structure is identical to the data set in those exercises, but it is a new sample and omits product brand. First we load the data from a local or online location:

```
> prst2 <- read.csv("https://goo.gl/BTxyFB") # online
# prst2 <- read.csv("chapter10-cfa.csv")  # local alternative
> summary(prst2)
   Adaptable       BestValue      CuttingEdge      Delightful
 Min.   :1.00   Min.   :1.00   Min.   :1.000   Min.   :1.000  ...
 1st Qu.:3.00   1st Qu.:3.00   1st Qu.:3.000   1st Qu.:3.000  ...
 Median :4.00   Median :4.00   Median :4.000   Median :4.000  ...
 Mean   :4.13   Mean   :3.73   Mean   :3.812   Mean   :4.277  ...
```

10.7.2 Exercises for Confirmatory Factor Analysis

1. Plot a correlation matrix for the adjectives in the new data set, `prst2`. Is it similar in structure to the results of exploratory factor analysis in Section 8.8?
2. Using the EFA model from the Section 8.8 exercises as a guide, define a `lavaan` model for a 3-factor solution. Fit that model to the `prst2` data for confirmatory factor analysis, and interpret the fit. (Note: in the `lavaan` model, consider setting the highest-loaded item loading to 1.0 for each factor’s latent variable; this can help anchor the model. Also note that `Adaptable` may need to load on two factors.)
3. Plot the 3-factor model.
4. Now find an alternative 2-factor EFA model for the `prst1` *exploratory* data. Define that as a CFA model for `lavaan` and fit it to the new `prst2` confirmatory data. You will need to define a model that you think is a reasonable 2-factor model.

5. Compare the 2-factor model fit to the 3-factor model fit. Which model is preferable?

10.7.3 Purchase Intention Data for Structural Equation Model Exercises

For these exercises, we use a new simulated data set to model likelihood to purchase a new product. Respondents have rated the new product on three of the same adjectives used in the PRST exercises above: Ease of Use, Cutting Edge, and Best Value. Additionally, there are ratings for satisfaction with the previous model of the product, likelihood to purchase the new product, and satisfaction with the new product's cost. This gives a total of six manifest items. In the exercises, you will define structural models using those items along with latent variables, and assess fit with the data. First, load the data:

```
> intent.df <- read.csv("https://goo.gl/6U5aYr")  # online
# intent.df <- read.csv("chapter10-sem.csv")  # local alternative
> summary(intent.df)
  iCuttingEdge      iEaseOfUse       iBestValue     iPreviousModelRating
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000      ...
 1st Qu.: 5.000   1st Qu.: 4.750   1st Qu.: 4.000   1st Qu.: 4.000      ...
 Median : 6.000   Median : 6.000   Median : 6.000   Median : 5.000      ...
 Mean   : 6.072   Mean   : 5.643   Mean   : 5.645   Mean   : 5.355      ...
```

In this data set, all column names begin with the letter "i," such as "iCuttingEdge." This is to help distinguish the survey *items* (manifest variables) from latent variables that you will define for the SEM models.

10.7.4 Exercises for Structural Equation Models and PLS SEM

Structural Equation Models

6. Define an SEM model for the product ratings in `intent.df` using `lavaan` syntax. The model has three latent variables: *ProductRating*, *PurchaseInterest*, and *PurchaseIntent*. The core idea is that *ProductRating* points to *PurchaseInterest*, and that points further to *PurchaseIntent*. *ProductRating* is manifest as the items `iCuttingEdge`, `iEaseOfUse`, and `iBestValue`. *PurchaseInterest* combines *ProductRating* with the manifest item `iPreviousModelRating`. *PurchaseIntent* combines *PurchaseInterest* with item `iCost` and the manifest rating item `iPurchaseIntent`. Your question is this: what are the most important items related to the latent variable for purchase intent? Define this model and fit it to the intent data.
7. Now define a simpler model and compare it. For the simpler model, define *ProductRating* as manifest in `iBestValue` and `iEaseOfUse`. *PurchaseIntent* should combine *ProductRating* with `iCost` and `iPurchaseIntent`. There will be no *PurchaseInterest* latent variable. Fit this model, and visualize and interpret the result. What are the drivers of purchase intent here? Is this model preferable to the previous model? (Note: the question of whether it is better depends on interpretation, not an assessment of fit; the models use different variables, so most fit indices are not directly comparable.)
8. *(Stretch exercise.)* Define a few other plausible models. Does one of them fit the data better? Should you therefore conclude that it is the right model? What would be your next steps, if you wanted to assert that?

Partial Least Squares SEM

For these exercises, use the `intent.df` data as in the SEM exercises above.

9. Sample N=30 observations from the purchase intent data, and fit the shorter SEM model from Exercise 7 to those data using regular, covariance-based SEM (i.e., `sem()`). (Note: results may vary by the random sample you take.) What do you observe? How do the estimates compare to the full sample results above?
10. Use Partial Least Squares SEM to estimate the model. How do the results of PLS-SEM compare to the covariance-based SEM estimates for drivers of purchase intent?
11. Using the N=30 sample, bootstrap the PLS-SEM estimates for 200 runs. How large are the ranges for the parameter estimates? What does that tell you about the stability of results from this sample?
12. Take a larger sample for N=200 and repeat the PLS-SEM bootstrap. How stable are the estimates with the larger sample? How do the estimated ranges of values compare to those from the N=30 sample?
13. *(Stretch exercise that requires you to explore some new R graphing commands.)* A PLS-SEM bootstrap object collects all of the bootstrapped estimates for the model parameters. Compare the N=30 vs. N=200 bootstraps with a graph for the best estimate and 95% observed intervals. Where do the estimates mostly agree or substantially disagree? (Hint: there are many ways to visualize the results. One approach is to compile the estimates for both sets and use `stat_summary()` from the `ggplot2` package.)

11

Segmentation: Clustering and Classification *(Exercises only)*

11.8 Exercises

11.8.1 Music Subscription Data for Exercises

Besides the data we provide here, these exercises could be adapted to experiment with many other data sets from this book, as well as various online sources of machine learning data. At the time of writing, there were more than 450 data sets available in the University of California, Irvine, repository at https://archive.ics.uci.edu/ml/datasets.html.

As written, these exercises use a simulated CRM data set that comprises customer information for an imagined music subscription service. Load the data as follows:

```
> seg.ex.raw <- read.csv("https://goo.gl/s1KEiF")    # original with segments
> seg.ex     <- seg.ex.raw                           # copy without segments
> seg.ex$Segment <- NULL
> summary(seg.ex.raw)
      age            sex      householdIncome    milesDrive      kidsAtHome
 Min.   :20.00   Male  :480   Min.   : 11042   Min.   :    0   Min.   :0.000
 1st Qu.:28.00   Female:415   1st Qu.: 34383   1st Qu.:10085   1st Qu.:0.000
...
   commuteCar    drivingEnthuse   musicEnthuse   subscribeToMusic
 Min.   :0.000   Min.   :1.000   Min.   :1.000   subNo :808
 1st Qu.:0.000   1st Qu.:3.000   1st Qu.:3.000   subYes: 87
...
         Segment
 CommuteNews :170
 KidsAndTalk :125
 LongDistance: 50
 MusicDriver :260
 NonCar      :205
 Quiet       : 85
```

The variables comprise the following: age in years; sex as male or female; householdIncome in US dollars; milesDrive, annual total miles driven in a car; kidsAtHome, number of children under 18 years old living at home; `commuteCar`, whether they regularly commute by car; `drivingEnthuse`, reported enthusiasm for driving, according to a survey response, reported on a scale from 1—7 (highest); `musicEnthuse`, enthusiasm for music on the same 1—7 scale; `subscribeToMusic`, whether the respondent subscribes to our service; `Segment`, a known assignment to one of our six customer segments, based on prior research.

As we noted in the chapter, you must take care not to include known results when fitting a machine learning model. In the data assignments above, we keep the known segment assignments in the data frame `seg.ex.raw`, and create `seg.ex` as a copy without the known segment assignments. Take care to use the appropriate data in the exercises.

11.8.2 Exercises

Clustering

1. In the chapter, we suggested writing a small function to quickly examine group differences. Develop a summary function for the `seg.ex` data. Demonstrate its basic usage.
2. Using hierarchical clustering, cluster the `seg.ex` data. Cut it into a specific number of segments, and visualize those. How many segments did you choose and why?
3. Are the `hclust()` results in Exercise 2 interesting? Show a plot that demonstrates why or why not.
4. Using `kmeans()`, find a four group solution for the data. (You'll need to do some data conversion first.) Is the solution interesting? Plot two of the continuous variables by segment.
5. Plot the clusters from Exercise 4 by principal components of the data set. Interpret the plot.
6. Use `Mclust()` to fit a model-based cluster solution for the music subscription data. How many clusters does it suggest? Are they well-differentiated? How do they compare to the k-means solution from exercise 4?
7. `Mclust()` can also fit a specified number of clusters (parameter "`G`"). Fit solutions for G=2 and G=4 clusters. When fit to the data, are they much worse than a G=3 solution?
8. Prepare the data for `poLCA`, recoding variables to binary factors, splitting as follows: age: less than 30, vs. 30+. Household income: less than 55000, vs. greater. Kids at home: 0, vs. more than 0. Music enthusiasm: a score of 5 or higher, vs. less than that. We will not use miles driven or driving enthusiasm for this exercise.
9. Fit polytomous latent class models for 3-class and 4-class solutions to the data. Visualize them. How different are the two solutions in respondents' assignments? Which one is more useful? (Note: solutions depend in part on the random number sequence.)

Classification

10. Split the music subscription data set — with the segment assignments — into 65%/35% sets for classification model training and assessment. Compare the two sets. Are they suitably similar? (Note: be sure to set a random number seed for replicability.)
11. Fit a naive Bayes model to predict segment membership `Segment` from the other variables in the training data. Check its performance on the test data. Does it perform better than chance?

12. Fit a random forest model to predict segment membership. What is its out of bag error rate? Did you do anything to control class imbalance?

13. With the random forest model from Exercise 12, predict segments in the test data. Compare those to the actual segments. Does it predict segment membership better than chance?

14. In the random forest model, which variables are most important for predicting segment membership?

15. Predict subscription status `subscribeToMusic` in the data, using a random forest model. How well does the model predict the test data? Which variables are most important?

16. *(Stretch exercise.)* Use the hotel satisfaction data from the exercises in Chapter 7. Model something interesting in that data set, using either clustering or classification approaches (or both). How do you evaluate your model's performance? Fit a reasonable alternative model. Is your model preferable to the alternative?

12

Association Rules for Market Basket Analysis *(Exercises only)*

12.7 Exercises

12.7.1 Retail Transactions Data for Exercises

For the exercises here, we use a data set of simulated transactions, plus another with item costs and margin, for a retail super center. This hypothetical store sells groceries along with other consumer and household goods, as well as some items more typical of home furnishing and big box stores, so the items range from very inexpensive ($0.09 USD) to quite expensive ($39009.00 USD).

We load the two data sets as follow, and additionally extract the item margins from the cost data frame to a simple vector, in order to match the approach taken in the chapter:

```
# load from website. 11MB, may be slow
retail.raw <- readLines("https://goo.gl/wi8KHg")
retail.margin <- read.csv("https://goo.gl/Pidzpd")

# or load locally, if downloaded
# retail.raw    <- readLines("retail-baskets.csv")
# retail.margin <- read.csv("retail-margin.csv")
margin.short  <- data.frame(retail.margin$margin)
rownames(margin.short) <- retail.margin$item
```

As always, we suggest to explore and summarize the data before starting analyses.

12.7.2 Exercises

1. Convert the raw transaction lines, as read above, into a transactions object for `arules`. How many unique items are there? What are the five most popular items? What are the sizes of the smallest, largest, and median basket? (*Hint:* in case of trouble, check the format of the raw item lines.)
2. Find association rules in the retail data. Aim for somewhere between 100 to 1000 rules (consider tuning the rule length, support, and confidence parameters). Plot confidence vs. support for the rules, and interpret that pattern.

3. Find the top 30 rules by lift and plot them. Which items are associated in the group with highest single-item support? Which items are in the largest group by total number of items?

4. In the chapter, we presented a function to calculate total margin for a set of rules. Among all the transactions, what are the top 10 baskets with the highest total margin?

5. Suppose we want to focus on the highest margin transactions. What proportion of baskets have a total margin of $200 or more? What are the most common items in those baskets? Plot the frequency of items that appear in 10% or more of those baskets. (*Hint:* there is an `arules` function called `itemFrequency()`.)

6. Add the item names to the plot axis for the previous exercise. (*Hint:* check Section **??**, and also examine plotting parameters `cex.axis` and `las`.)

7. The retail.margin data frame, as loaded above, has both price and margin. Calculate the proportional margin for each item (margin divided by price). Plot those. If you transformed them, what would be an appropriate transformation? Plot that also.

8. *(Stretch programming exercise.)* Write a function similar to `retail.margsum()` but that returns both the total margin and the total price for a basket. Find the top 10 baskets in terms of their margin to price ratio.

13

Choice Modeling *(Exercises only)*

13.10 Exercises

For the exercises in this chapter, we use a simulated conjoint data set where we observe respondents' choices from among sets of sports cars. The attributes of sports cars in this data are number of *seats*, *convertible* top, *trans*mission type and *price*. The data also includes a *segment* variable that indicates which sportscar segment each respondent belongs to.

You can load the data by typing:

```
sportscar <- read.csv("https://goo.gl/8g7vtT")
```

1. Data inspection:
   - Use summary to identify the levels of each attribute in the data.
   - What was the price of the chosen alternative in the last question in the data frame?
   - Use `xtabs()` to determine the number of times a sportscar with automatic transmission was chosen. How does this compare to the number of times a sportscar with a manual transmission was chosen? What does that tell you about consumer preferences?
2. Fitting and interpreting a choice model:
   a) Fit a (non-hierarchical) choice model to predict choice as a function of all four attributes. Don't forget to convert the data to an `mlogit.data` object before passing it to `mlogit`. Also, be sure to remove the intercept in the model formula. Report the estimated coefficients and their standard errors.

   b) What is the ideal sportscar for the repsondents based on this model. That is, what is most desirable level of each feature? You may have to look at both the model coefficients and the data summary to figure this out.

   c) Which coefficient is the most precisely estimated?

   d) Is it reasonable to charge $5000 for a convertable top? Hint: Compute the WTP for convertible top and compare it to $5000.

e) Use the `predict.mnl()` function from the chapter to predict shares for the following set of sportscars:

```
newcars <- data.frame(seat=factor(c("2","4", "5")),
                      trans=factor(c("manual", "automatic", "automatic")),
                      convert=factor(c("no", "yes", "no")),
                      price=c(40, 37, 35))
```

Note that it is very important that the factors in the `newcars` data frame have exactly the same levels as the factors in the data frame used to estimate the model.

f) Use the `sensitivity.mnl()` function from the chapter to produce a sensitivity plot for the first sportscar in `newcars`. What suggestions would you make on changing the features of the product to get higher market share?

3. In the previous question, you fit a choice model using all the respondents and so your estimates represented the average preferences across the entire population of customers. Fit another choice model using just the customers from the racer segment and predict shares for `newcars`. Are your predictions different than what you found in the previous question?

4. Estimate a hieararchical multinomial logit model using the sportscar data, using the `rpar` input to the `mlogit()` function. Assume all the parameters are normally distributed. Which parameter has the greatest variation across respondents?

5. Estimate a hierarchical model using the Bayesian `ChoiceModelR` function. Don't forget you will have to re-format the data to be suitable for `ChoiceModelR` as described in Section 13.5.1. Use the segment variable for the demographics. Are the parameter estimates similar to those obtained with `mlogit()`?

14

Behavior Sequences *(Exercises only)*

14.8 Exercises

In these exercises, we try to expand your horizons in the final two questions, which propose larger-scale projects.

1. Plot the number of *bytes* requested by user in the EPA data. Should the values be transformed? Why or why not? If so, plot them with and without the transformation, and discuss the differing interpretation.
2. There is one value for "total number of bytes downloaded" that is especially frequent in the EPA data. For example, in the previous exercise it appears as a spike in the plot. What value is it? Why it is frequent?
3. Omit the end states from the sequences and repeat the Markov chain analysis. Are there differences in your interpretation of the result?
4. Now model the sequences using the top 40 pages instead of top 20 (and without end states). How do you need to change the visualizations to be more useful?
5. *(Thought exercise without code.)* Suppose the EPA asked you to consult on their web site. They give you the present data as background, and suggest that you will be able to collect the same kind of data again, and that you also might be able to collect additional variables. Assume that you can collect up to 10 additional variables. Based on the analyses in this chapter and other chapters in this book, what other data would you wish to have, and why?
6. *(This is a more extensive, full project with a different data set.)* Additional web log data sets are available, as of publication date, at `http://ita.ee.lbl.gov/html/traces.html`. Choose one of the data sets there and repeat the analyses in this chapter. For modest size, we suggest the San Diego Supercomputer Center (SDSC) data set [**?**], although you might also wish to try one of the larger data sets. (Note: the SDSC data set is also available at `https://goo.gl/jpWMVh`.) If you use the SDSC data, note that the host address has two parts: "N+H", where "N" is a particular network and "H" is a machine (host) within that network. It is a unique identifier.
7. *(Stretch exercise: the longest one in the book, which requires detailed programming and determining on your own how to use two new packages.)* Using the IP addresses in the EPA data, find the geographic

location of the requesting machines and plot them on a choropleth map (Section 3.4.6). Because online services to look up data can be slow, cache lookup results locally, such that if you run the procedure again, it loads a local file instead of looking up results. (Hint: check out the packages `iptools` and `rgeolocate`, and the function `file.exists()`.)