Chapter 9
Step 7: Describing Segments

9.1 Developing a Complete Picture of Market Segments

Segment profiling is about understanding differences in segmentation variables across market segments. Segmentation variables are chosen early in the market segmentation analysis process: conceptually in Step 2 (specifying the ideal target segment), and empirically in Step 3 (collecting data). Segmentation variables form the basis for extracting market segments from empirical data.

Step 7 (describing segments) is similar to the profiling step. The only difference is that the variables being inspected have *not* been used to extract market segments. Rather, in Step 7 market segments are described using *additional* information available about segment members. If committing to a target segment is like a marriage, profiling and describing market segments is like going on a number of dates to get to know the potential spouse as well as possible in an attempt to give the marriage the best possible chance, and avoid nasty surprises down the track. As van Raaij and Verhallen (1994, p. 58) state: segment . . . should be further described and typified by crossing them with all other variables, i.e. with psychographic . . . , demographic and socio-economic variables, media exposure, and specific product and brand attitudes or evaluations.

For example, when conducting a data-driven market segmentation analysis using the Australian travel motives data set (this is the segmentation solution we saved on page 171; the data is described in Appendix C.4), profiling means investigating differences between segments with respect to the travel motives themselves. These profiles are provided in Fig. 8.2. The segment description step uses additional information, such as segment members' age, gender, past travel behaviour, preferred vacation activities, media use, use of information sources during vacation planning, or their expenditure patterns during a vacation. These additional variables are referred to as *descriptor variables*.

Good descriptions of market segments are critical to gaining detailed insight into the nature of segments. In addition, segment descriptions are essential for the

S. Dolnicar et al., *Market Segmentation Analysis*, Management for Professionals,
https://doi.org/10.1007/978-981-10-8818-6_9

development of a customised marketing mix. Imagine, for example, wanting to target segment 4 which emerged from extracting segments from the Australian travel motives data set. Step 6 of the segmentation analysis process leads to the insight that members of segment 4 care about nature. Nothing is known, however, about how old these people are, if they have children, how high their discretionary income is, how much money they spend when they go on vacation, how often they go on vacation, which information sources they use when they plan their vacation, and how they can be reached. If segment description reveals, for example, that members of this segment have a higher likelihood of volunteering for environmental organisations, and regularly read National Geographic, tangible ways of communicating with segment 4 have been identified. This knowledge is important for the development of a customised marketing mix to target segment 4.

We can study differences between market segments with respect to descriptor variables in two ways: we can use descriptive statistics including visualisations, or we can analyse data using inferential statistics. The marketing literature traditionally relies on statistical testing, and tabular presentations of differences in descriptor variables. Visualisations make segment description more user-friendly.

9.2 Using Visualisations to Describe Market Segments

A wide range of charts exist for the visualisation of differences in descriptor variables. Here, we discuss two basic approaches suitable for nominal and ordinal descriptor variables (such as gender, level of education, country of origin), or metric descriptor variables (such as age, number of nights at the tourist destinations, money spent on accommodation).

Using graphical statistics to describe market segments has two key advantages: it simplifies the interpretation of results for both the data analyst and the user, and integrates information on the statistical significance of differences, thus avoiding the over-interpretation of insignificant differences. As Cornelius et al. (2010, p. 197) put it: Graphical representations ... serve to transmit the very essence of marketing research results. The same authors also find – in a survey study with marketing managers – that managers prefer graphical formats, and view the intuitiveness of graphical displays as critically important. Section 8.3.1 provides an illustration of the higher efficiency with which people process graphical as opposed to tabular results.

9.2.1 Nominal and Ordinal Descriptor Variables

When describing differences between market segments in one single nominal or ordinal descriptor variable, the basis for all visualisations and statistical tests is a cross-tabulation of segment membership with the descriptor variable. For the

Australian travel motives data set (see Appendix C.4), data frame `vacmotdesc` contains several descriptor variables. These descriptor variables are automatically loaded with the Australian travel motives data set. To describe market segments, we need the segment membership for all respondents. We store segment membership in helper variable `C6`:

```
R> C6 <- clusters(vacmot.k6)
```

The sizes of the market segments are

```
R> table(C6)
```

```
C6
  1   2   3   4   5   6
235 189 174 139  94 169
```

The easiest approach to generating a cross-tabulation is to add segment membership as a categorical variable to the data frame of descriptor variables. Then we can use the formula interface of R for testing or plotting:

```
R> vacmotdesc$C6 <- as.factor(C6)
```

The following R command gives the number of females and males across market segments:

```
R> C6.Gender <- with(vacmotdesc,
+   table("Segment number" = C6, Gender))
R> C6.Gender
```

```
              Gender
Segment number Male Female
             1  125    110
             2   86    103
             3   94     80
             4   78     61
             5   47     47
             6   82     87
```

A visual inspection of this cross-tabulation suggests that there are no huge gender differences across segments. The upper panel in Fig. 9.1 visualises this cross-tabulation using a stacked bar chart. The *y*-axis shows segment sizes. Within each bar, we can easily how many are male and how many are female. We cannot, however, compare the proportions of men and women easily across segments. Comparing proportions is complicated if the segment sizes are unequal (for example, segments 1 and 5). A solution is to draw the bars for women and men next to one another rather than stacking them (not shown). The disadvantage of this approach is that the absolute sizes of the market segments can no longer be directly seen on the *y*-axis. The *mosaic plot* offers a solution to this problem.

The mosaic plot also visualises cross-tabulations (Hartigan and Kleiner 1984; Friendly 1994). The width of the bars indicates the absolute segment size. The column for segment 5 of the Australian travel motives data set – containing 94

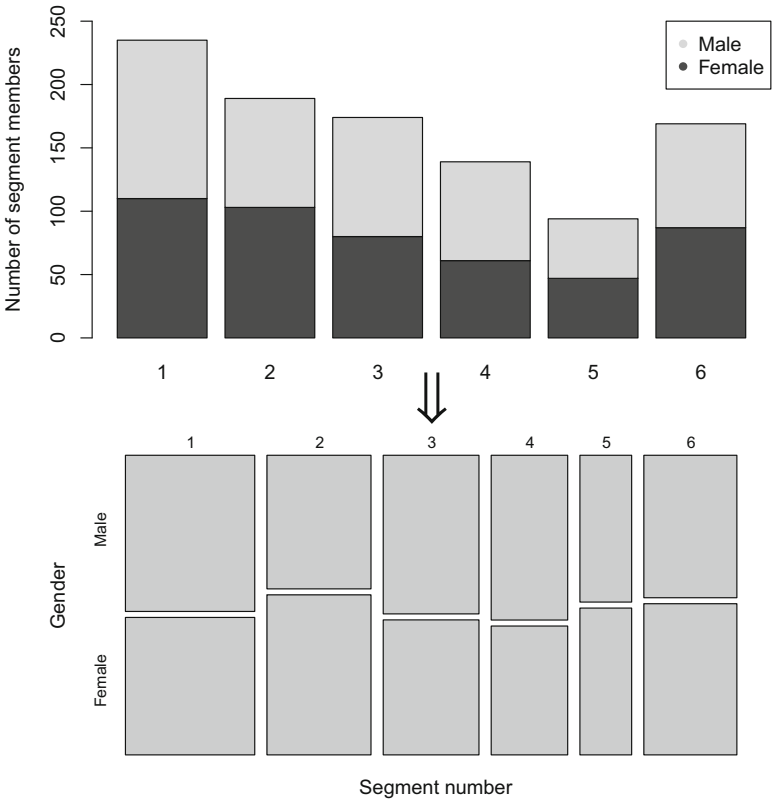

Fig. 9.1 Comparison of a stacked bar chart and a mosaic plot for the cross-tabulation of segment membership and gender for the Australian travel motives data set

respondents or 9% of the sample – is much narrower in the bottom plot of Fig. 9.1 than the column for segment 1 – containing 235 respondents or 24% of the sample.

Each column consists of rectangles. The height of the rectangles represents the proportion of men or women in each segment. Because all columns have the same total height, the height of the bottom rectangles is in the same position for two segments with the same proportion of men and women (even if the absolute number of men and women differs substantially). Because the width of the columns represents the total segment sizes, the area of each cell is proportional to the size of the corresponding cell in the table.

Mosaic plots can also visualise tables containing more than two descriptor variables and integrate elements of inferential statistics. This helps with interpretation. Colours of cells can highlight where observed frequencies are different from expected frequencies under the assumption that the variables are independent. Cell colours are based on the standardised difference between the expected and observed frequencies. Negative differences mean that observed are lower than expected

frequencies. They are coloured in red. Positive differences mean that observed are higher than expected frequencies. They are coloured in blue. The saturation of the colour indicates the absolute value of the standardised difference. Standardised differences follow asymptotically a standard normal distribution. Standard normal random variables lie within $[-2, 2]$ with a probability of $\approx$95%, and within $[-4, 4]$ with a probability of $\approx$99.99%. Standardised differences are equivalent to the standardised Pearson residuals from a log-linear model assuming independence between the two variables.

By default, function `mosaicplot()` in R uses dark red cell colouring for contributions or standardised Pearson residuals smaller than -4, light red if contributions are smaller than -2, white (not interesting) between -2 and 2, light blue if contributions are larger than 2, and dark blue if they are larger than 4. Figure 9.2 shows such a plot with the colour coding included in the legend.

In Fig. 9.2 all cells are white, indicating that the six market segments extracted from the Australian travel motives data set do not significantly differ in gender distribution. The proportion of female and male tourists is approximately the same across segments. The dashed and solid borders of the rectangles indicate that the number of respondents in those cells are either lower than expected (dashed

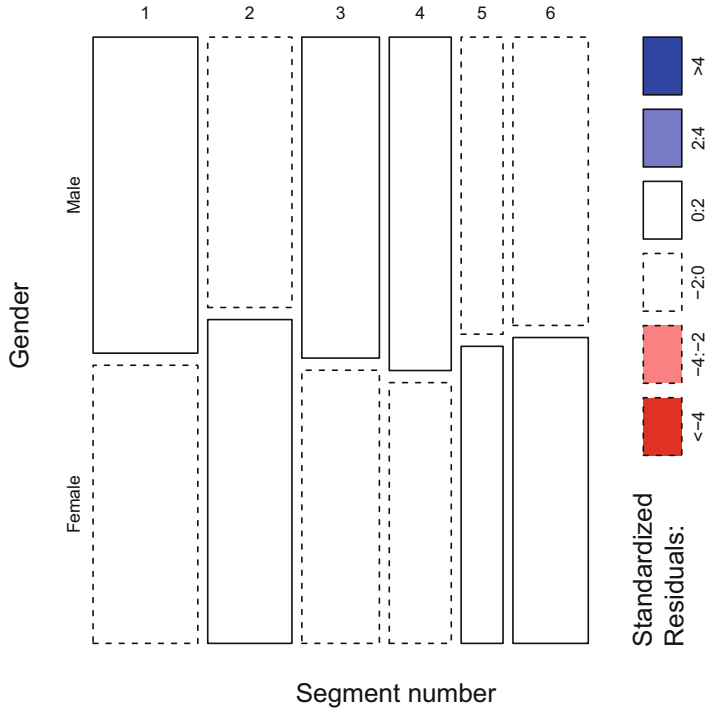

Fig. 9.2 Shaded mosaic plot for cross-tabulation of segment membership and gender for the Australian travel motives data set

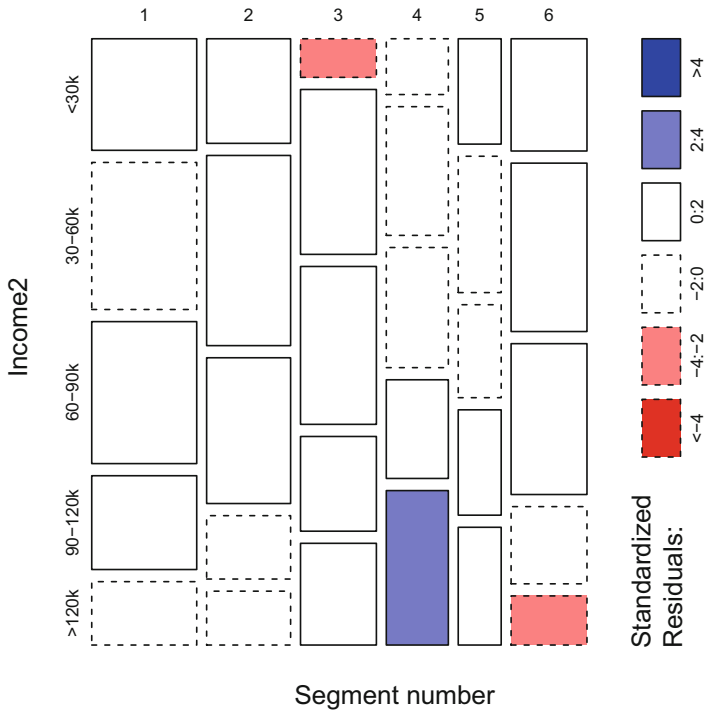

Fig. 9.3 Shaded mosaic plot for cross-tabulation of segment membership and income for the Australian travel motives data set

borders), or higher than expected (solid black borders). But, irrespective of the borders, white rectangles mean differences are statistically insignificant.

Figure 9.3 shows that segment membership and income are moderately associated. The top row corresponds to the lowest income category (less than AUD 30,000 per annum). The bottom row corresponds to the highest income category (more than AUD 120,000 per annum). The remaining three categories represent AUD 30,000 brackets in-between those two extremes. We learn that members of segment 4 (column 4 in Fig. 9.3) – those motivated by cultural offers and interested in local people – earn more money. Low income tourists (top row of Fig. 9.3) are less frequently members of market segment 3, those who do not care about prices and instead seek luxury, fun and entertainment, and wish to be spoilt when on vacation. Segment 6 (column 6 in Fig. 9.3) – the nature loving segment – contains fewer members on very high incomes.

Figure 9.4 points to a strong association between travel motives and stated moral obligation to protect the environment. The moral obligation score results from averaging the answers to 30 survey questions asking respondents to indicate how obliged they feel to engage in a range of environmentally friendly behaviours at home (including not to litter, to recycle rubbish, to save water and energy; see

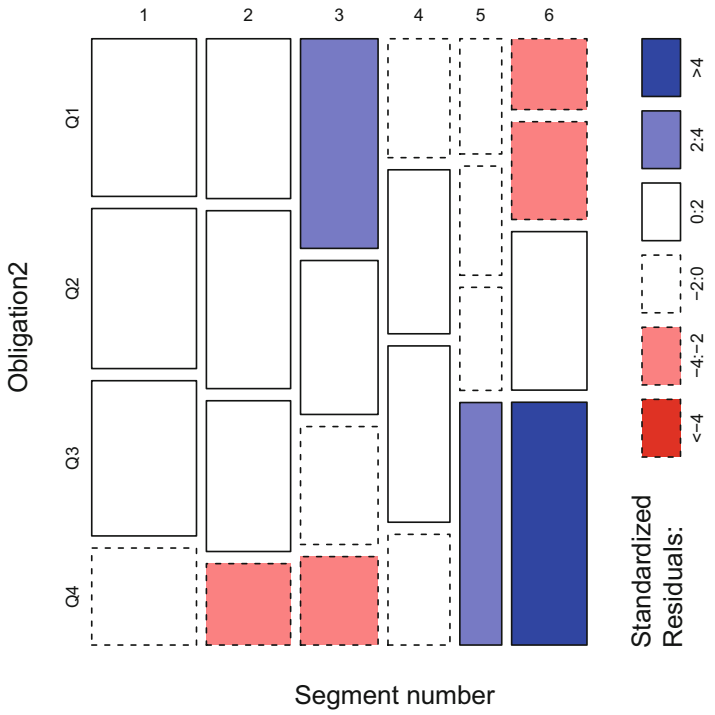

Fig. 9.4 Shaded mosaic plot for cross-tabulation of segment membership and moral obligation to protect the environment for the Australian travel motives data set

Dolnicar and Leisch 2008 for details). The moral obligation score is numeric and ranges from 1 (lowest moral obligation) to 5 (highest moral obligation) because survey respondents had five answer options. The summated score ranges from 30 to 150, and is re-scaled to 1 to 5 by dividing through 30. We provide an illustration of how this descriptor variable can be analysed in its original metric format in Sect. 9.2.2. To create the mosaic plot shown in Fig. 9.4, we cut the moral obligation score into quarters containing 25% of respondents each, ranging from Q1 (low moral obligation) to Q4 (high moral obligation). Variable `Obliged2` contains this re-coded descriptor variable.

Figure 9.4 graphically illustrates the cross-tabulation, associating segment membership and stated moral obligation to protect the environment in a mosaic plot. Segment 3 (column 3 of Fig. 9.4) – whose members seek entertainment – contains significantly more members with low stated moral obligation to behave in an environmentally friendly way. Segment 3 also contains significantly fewer members in the high moral obligation category. The exact opposite applies to segment 6. Members of this segment are motivated by nature, and plotted in column 6 of Fig. 9.4. Being a member of segment 6 implies a positive association with high

moral obligation to behave environmentally friendly, and a negative association with membership in the lowest moral obligation category.

9.2.2 Metric Descriptor Variables

R package lattice (Sarkar 2008) provides conditional versions of most standard R plots. An alternative implementation for conditional plots is available in package ggplot2 (Wickham 2009). *Conditional* in this context means that the plots are divided in sections (panels, facets), each presenting the results for a subset of the data (for example, different market segments). Conditional plots are well-suited for visualising differences between market segments using metric descriptor variables. R package lattice generated the segment profile plot in Sect. 8.3.1.

In the context of segment description, this R package can display the age distribution of all segments comparatively. Or visualise the distribution of the (original metric) moral obligation scores for members of each segment.

To have segment names (rather than only segment numbers) displayed in the plot, we create a new factor variable by pasting together the word `"Segment"` and the segment numbers from `C6`. We then generate a histogram for age for each segment. Argument `as.table` controls whether the panels are included by starting on the top left (`TRUE`) or bottom left (`FALSE`, the default).

```
R> library("lattice")
R> histogram(~ Age | factor(paste("Segment", C6)),
+   data = vacmotdesc, as.table = TRUE)
```

We do the same for moral obligation:

```
R> histogram(~ Obligation | factor(paste("Segment",C6)),
+   data = vacmotdesc, as.table = TRUE)
```

The resulting histograms are shown in Figs. 9.5 (for age) and 9.6 (for moral obligation). In both cases, the differences between market segments are difficult to assess just by looking at the plots.

We can gain additional insights by using a parallel box-and-whisker plot; it shows the distribution of the variable separately for each segment. We create this parallel box-and-whisker plot for age by market segment in R with the following command:

```
R> boxplot(Age ~ C6, data = vacmotdesc,
+   xlab = "Segment number", ylab = "Age")
```

where arguments `xlab` and `ylab` customise the axis labels.

Figure 9.7 shows the resulting plot. As expected – given the histograms inspected previously – differences in age across segments are minor. The median age of members of segment 5 is lower, that of segment 6 members is higher. These visually detected differences in descriptors need to be subjected to statistical testing.

Like mosaic plots, parallel box-and-whisker plots can the incorporate elements of statistical hypothesis testing. For example, we can make the width of the

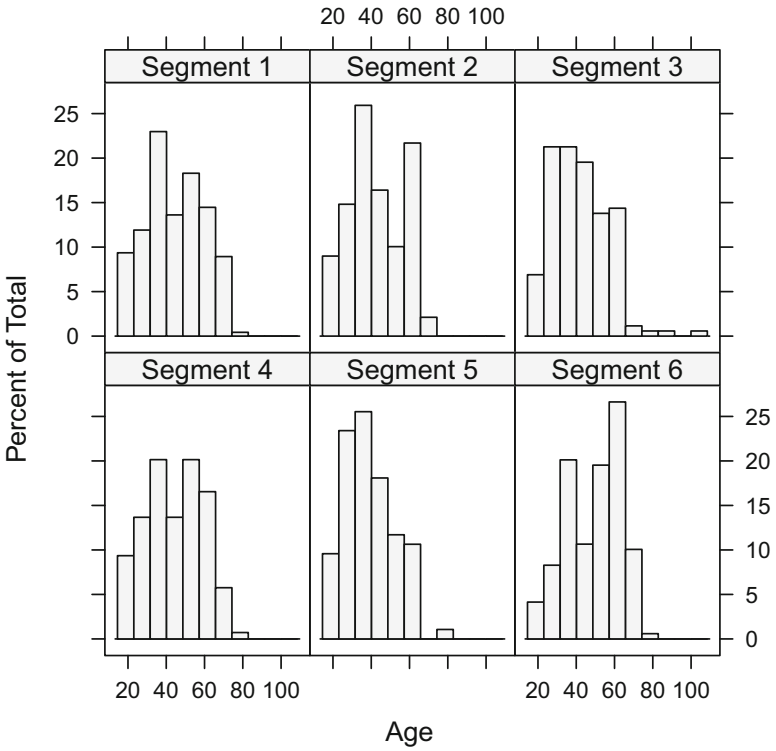

Fig. 9.5 Histograms of age by segment for the Australian travel motives data set

boxes proportional to the size of market segments (`varwidth = TRUE`), and include 95% confidence intervals for the medians (`notch = TRUE`) using the R command:

```
R> boxplot(Obligation ~ C6, data = vacmotdesc,
+    varwidth = TRUE, notch = TRUE,
+    xlab = "Segment number",
+    ylab = "Moral obligation")
```

Figure 9.8 contains the resulting parallel box-and-whisker plot. This version illustrates that segment 5 is the smallest; its box is the narrowest. Segment 1 is the largest. Moral obligation to protect the environment is highest among members of segment 6.

The notches in this version of the parallel box-and-whisker plot correspond to 95% confidence intervals for the medians. If the notches for different segments do not overlap, a formal statistical test will usually result in a significant difference. We can conclude from the inspection of the plot in Fig. 9.8 alone, therefore, that there is a significant difference in moral obligation to protect the environment between members of segment 3 and members of segment 6. The notches for those two

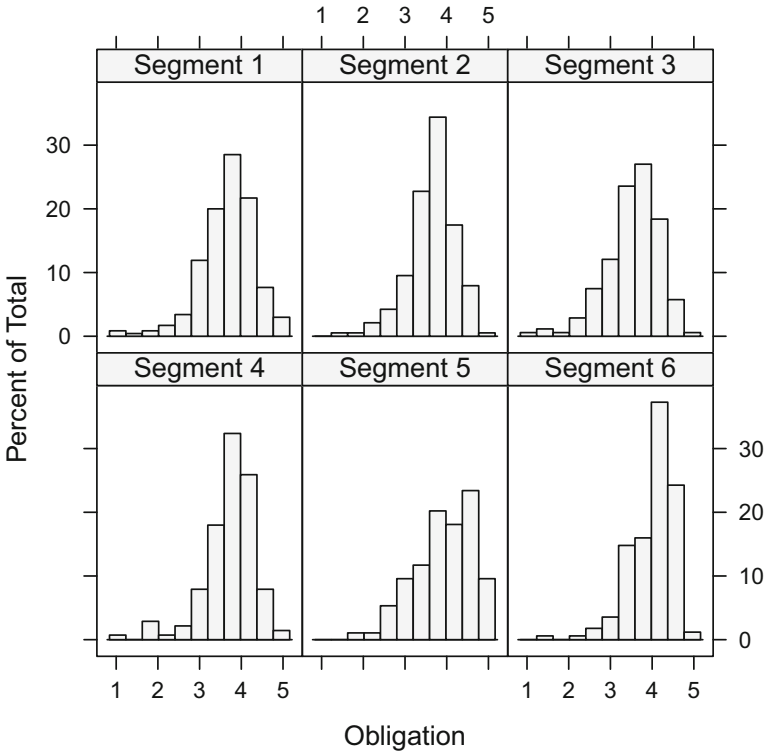

Fig. 9.6 Histograms of moral obligation to protect the environment by segment for the Australian travel motives data set

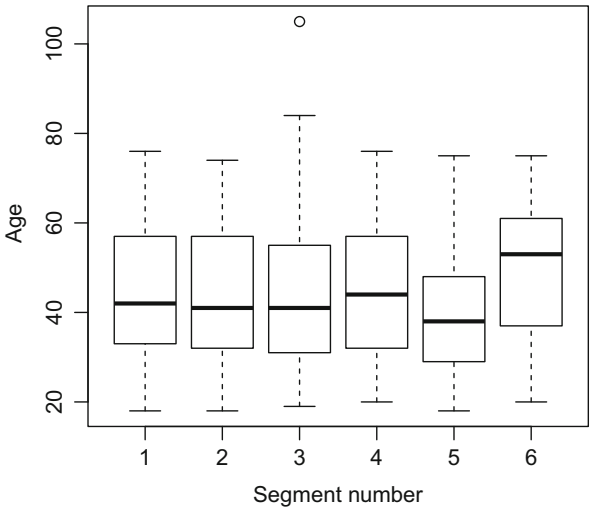

Fig. 9.7 Parallel box-and-whisker plot of age by segment for the Australian travel motives data set

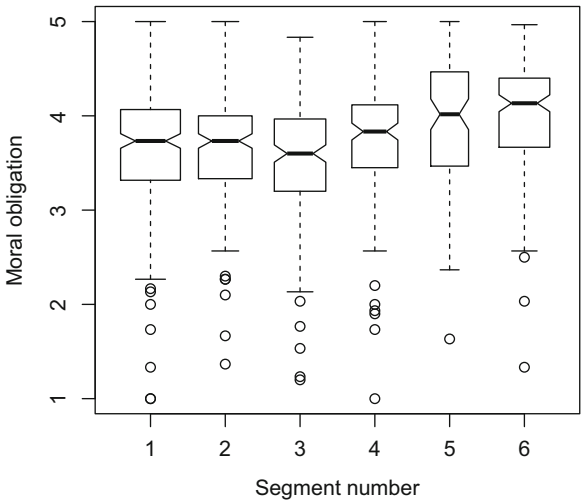

Fig. 9.8 Parallel box-and-whisker plot (with elements of statistical inference) of moral obligation to protect the environment by segment for the Australian travel motives data set

segments are far away from each other. Most of the boxes and whiskers are almost symmetric around the median, but all segments contain some outliers at the low end of moral obligation. One possible interpretation is that – while most respondents state that they feel morally obliged to protect the environment (irrespective of whether they actually do it or not) – only few openly admit to not feeling a sense of moral obligation.

We can use a modified version of the segment level stability across solutions (SLS_A) plot to trace the value of a metric descriptor variable over a series of market segmentation solutions. The modification is that additional information contained in a metric descriptor variable is plotted using different colours for the nodes:

```
R> slsaplot(vacmot.k38, nodecol = vacmotdesc$Obligation)
```

The nodes of the segment level stability across solutions (SLS_A) plot shown in Fig. 9.9 indicate each segment's mean moral obligation to protect the environment using colours. A deep red colour indicates high moral obligation. A light grey colour indicates low moral obligation.

The segment that has been repeatedly identified as a potentially attractive market segment (nature-loving tourists with an interest in the local population) appears along the bottom row. This segment consistently – across all plotted segmentation solutions – displays high moral obligation to protect the environment, followed by the segment identified as containing responses with acquiescence (yes saying) bias (segment 5 in the six-segment solution). This is not altogether surprising: if members of the acquiescence segment have an overall tendency to express agreement with survey questions (irrespective of the content), they are also likely to express agreement when asked about their moral obligation to protect the environment. Because the node colour has a different meaning in this modified segment level stability across solutions (SLS_A) plot, the shading of the edges

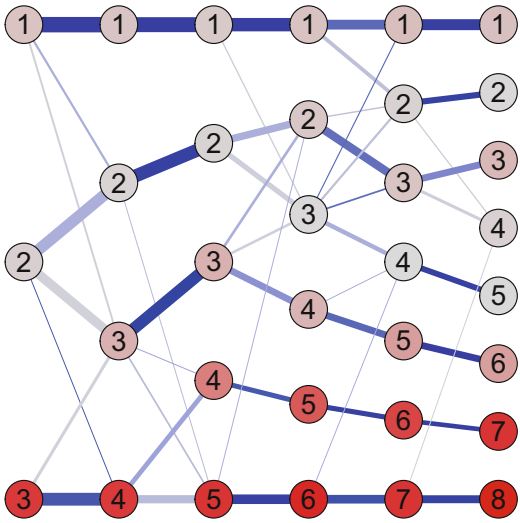

Fig. 9.9 Segment level stability across solutions (SLS_A) plot for the Australian travel motives data set for three to eight segments with nodes coloured by mean moral obligation values

represents the numeric SLS_A value. Light grey edges indicate low stability values. Dark blue edges indicate high stability values.

9.3 Testing for Segment Differences in Descriptor Variables

Simple statistical tests can be used to formally test for differences in descriptor variables across market segments. The simplest way to test for differences is to run a series of independent tests for each variable of interest. The outcome of the segment extraction step is segment membership, the assignment of each consumer to one market segment. Segment membership can be treated like any other nominal variable. It represents a nominal summary statistic of the segmentation variables. Therefore, any test for association between a nominal variable and another variable is suitable.

The association between the nominal segment membership variable and another nominal or ordinal variable (such as gender, level of education, country of origin) is visualised in Sect. 9.2.1 using the cross-tabulation of both variables as basis for the mosaic plot. The appropriate test for independence between columns and rows of a table is the χ^2-test. To formally test for significant differences in the gender distribution across the Australian travel motives segments, we use the following R command:

```
R> chisq.test(C6.Gender)

        Pearson's Chi-squared test

data:  C6.Gender
X-squared = 5.2671, df = 5, p-value = 0.3842
```

The output contains: the name of the statistical test, the data used, the value of the test statistic (in this case `X-squared`), the parameters of the distribution used to calculate the p-value (in this case the degrees of freedom (`df`) of the χ^2-distribution), and the p-value.

The p-value indicates how likely the observed frequencies occur if there is no association between the two variables (and sample size, segment sizes, and overall gender distribution are fixed). Small p-values (typically smaller than 0.05), are taken as statistical evidence of differences in the gender distribution between segments. Here, this test results in a non-significant p-value, implying that the null hypothesis is not rejected. The mosaic plot in Fig. 9.2 confirms this: no effects are visible and no cells are coloured.

The mosaic plot for segment membership and moral obligation to protect the environment shows significant association (Fig. 9.4), as does the corresponding χ^2-test:

```
R> chisq.test(with(vacmotdesc, table(C6, Obligation2)))

        Pearson's Chi-squared test

data:  with(vacmotdesc, table(C6, Obligation2))
X-squared = 96.913, df = 15, p-value = 5.004e-14
```

If the χ^2-test rejects the null hypothesis of independence because the p-value is smaller than 0.05, a mosaic plot is the easiest way of identifying the reason for rejection. The colour of the cells points to combinations occurring more or less frequently than expected under independence.

The association between segment membership and metric variables (such as age, number of nights at the tourist destinations, dollars spent on accommodation) is visualised using parallel boxplots. Any test for difference between the location (mean, median) of multiple market segments can assess if the observed differences in location are statistically significant.

The most popular method for testing for significant differences in the means of more than two groups is *Analysis of Variance* (ANOVA). To test for differences in mean moral obligation values to protect the environment (shown in Fig. 9.8) across market segments, we first inspect segment means:

```
R> C6.moblig <- with(vacmotdesc, tapply(Obligation,
+   C6, mean))
R> C6.moblig

       1        2        3        4        5        6
3.673191 3.651146 3.545977 3.724460 3.928723 4.008876
```

We can use the following analysis of variance to test for significance of differences:

```
R> aov1 <- aov(Obligation ~ C6, data = vacmotdesc)
R> summary(aov1)

             Df Sum Sq Mean Sq F value  Pr(>F)
C6            5   24.7   4.933   12.93 3.3e-12 ***
Residuals   994  379.1   0.381
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance performs an F-test with the corresponding test statistic given as `F value`. The `F value` compares the weighted variance between market segment means with the variance within market segments. Small values support the null hypothesis that segment means are the same. The p-value given in the output is smaller than 0.05. This means that we reject the null hypothesis that each segment has the same mean obligation. At least two market segments differ in their mean moral obligation to protect the environment.

Summarising mean values of metric descriptor variables by segment in a table provides a quick overview of segment characteristics. Adding the analysis of variance p-values indicates if differences are statistically significant. As an example, Table 9.1 presents mean values for age and moral obligation by market segment together with the analysis of variance p-values. As a robust alternative we can report median values by segment, and calculate p-values of the Kruskal-Wallis rank sum test. The Kruskal-Wallis rank sum test assumes (as null hypothesis) that all segments have the same median. This test is implemented in function `kruskal.test()` in R. `kruskal.test` is called in the same way as `aov`.

If we reject the null hypothesis of the analysis of variance, we know that segments do not have the same mean level of moral obligation. But the analysis of variance does not identify the differing segments. Pairwise comparisons between segments provide this information. The following command runs all pairwise t-tests, and reports the p-values:

```
R> with(vacmotdesc, pairwise.t.test(Obligation, C6))

    Pairwise comparisons using t tests with pooled SD

data:  Obligation and C6

  1       2       3       4       5
2 1.00000 -       -       -       -
3 0.23820 0.52688 -       -       -
```

Table 9.1 Differences in mean values for age and moral obligation between the six segments for the Australian travel motives data set together with ANOVA p-values

	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6	Total	p-value
Age	44.61	42.66	42.31	44.42	39.37	49.62	44.17	1.699E-07
Moral obligation	3.67	3.65	3.55	3.72	3.93	4.01	3.73	3.300E-12

```
4 1.00000 1.00000 0.08980 -       -
5 0.00653 0.00387 1.8e-05 0.09398 -
6 1.2e-06 7.9e-07 1.1e-10 0.00068 1.00000

P value adjustment method: holm
```

The p-value of the t-test is the same if segment 1 is compared to segment 2, or if segment 2 is compared to segment 1. To avoid redundancy, the output only contains the p-values for one of these comparisons, and omits the upper half of the matrix of pairwise comparisons.

The results in the first column indicate that segment 1 does not differ significantly in mean moral obligation from segments 2, 3, and 4, but does differ significantly from segments 5 and 6. The advantage of this output is that it presents the results in very compact form. The disadvantage is that the direction of the difference cannot be seen. A parallel box-and-whisker plot reveals the direction. We see in Fig. 9.8 that segments 5 and 6 feel more morally obliged to protect the environment than segments 1, 2, 3 and 4.

The above R output for the pairwise t-tests shows (in the last line) that p-values were adjusted for *multiple testing* using the method proposed by Holm (1979). Whenever a series of tests is computed using the same data set to assess a single hypothesis, p-values need to be adjusted for multiple testing.

The single hypothesis in this case is that all segment means are the same. This is equivalent to the hypothesis that – for any pair of segments – the means are the same. The series of pairwise t-tests assesses the later hypothesis. But the p-value of a single t-test only controls for wrongly rejecting the null hypothesis that this pair has the same mean values. Adjusting the p-values allows to reject the null hypothesis that the means are the same for all segments if at least one of the reported p-values is below the significance level. After adjustment, the chance of making a wrong decision meets the expected error rate for testing this hypothesis. If the same rule is applied without adjusting the p-values, the error rate of wrongly rejecting the null hypothesis would be too high.

The simplest way to correct p-values for multiple testing is Bonferroni correction. Bonferroni correction multiplies all p-values by the number of tests computed and, as such, represents a very conservative approach. A less conservative and more accurate approach was proposed by Holm (1979). Several other methods are available, all less conservative than Bonferroni correction. Best known is the false discovery rate procedure proposed by Benjamini and Hochberg (1995). See `help("p.adjust")` for methods available in R.

As an alternative to calculating the series of pairwise t-tests, we can plot Tukey's honest significant differences (Tukey 1949; Miller 1981; Yandell 1997):

```
R> plot(TukeyHSD(aov1), las = 1)
R> mtext("Pairs of segments", side = 2, line = 3)
```

Function `mtext()` writes text into the margin of the plot. The first argument (`"Pairs of segments"`) contains the text to be included. The second argument (`"side = 2"`) specifies where the text appears. The value 2 stands for the

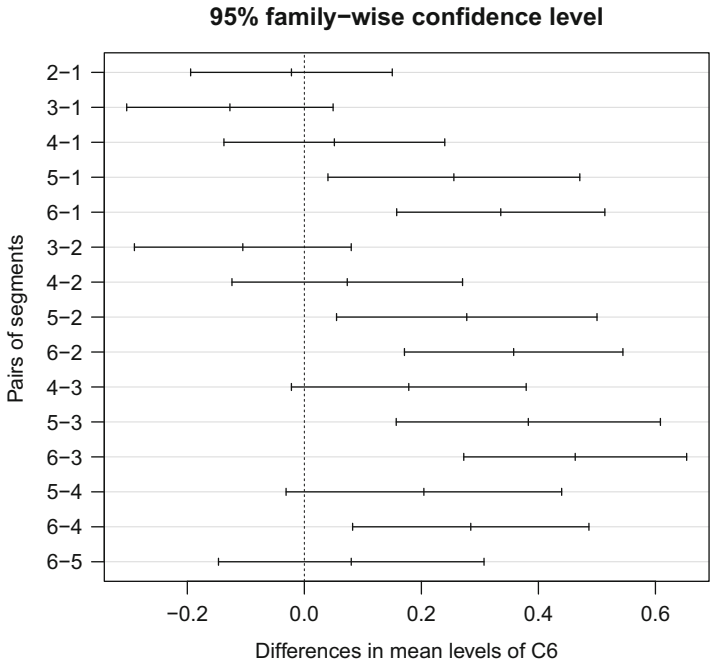

Fig. 9.10 Tukey's honest significant differences of moral obligation to behave environmentally friendly between the six segments for the Australian travel motives data set

left margin. The third argument (`"line = 3"`) specifies the distance between plot and text. The value 3 means the text is written three lines away from the box surrounding the plotting region.

Figure 9.10 shows the resulting plot. Each row represents the comparison of a pair of segments. The first row compares segments 1 and 2, the second row compares segments 1 and 3, and so on. The bottom row compares segments 5 and 6. The point estimate of the differences in mean values is located in the middle of the horizontal solid line. The length of the horizontal solid line depicts the confidence interval of the difference in mean values. The calculation of the confidence intervals is based on the analysis of variance result, and adjusted for the fact that a series of pairwise comparisons is made. If a confidence interval (horizontal solid line in the plot) crosses the vertical line at 0, the difference is not significant. All confidence intervals (horizontal solid lines in the plot) not crossing the vertical line at 0 indicate significant differences.

As can be seen from Fig. 9.10, segments 1, 2, 3 and 4 do not differ significantly from one another in moral obligation. Neither do segments 5 and 6. Segments 5 and 6 are characterised by a significantly higher moral obligation to behave environmentally friendly than the other market segments (with the only exception of segments 4 and 5 not differing significantly). As the parallel box-and-whisker

plot in Fig. 9.8 reveals, segment 4 sits between the low and high group, and does not display significant differences to segments 1–3 at the low end, and 5 at the high end of the moral obligation range.

9.4 Predicting Segments from Descriptor Variables

Another way of learning about market segments is to try to predict segment membership from descriptor variables. To achieve this, we use a *regression model* with the segment membership as categorical dependent variable, and descriptor variables as independent variables. We can use methods developed in statistics for classification, and methods developed in machine learning for supervised learning.

As opposed to the methods in Sect. 9.3, these approaches test differences in all descriptor variables simultaneously. The prediction performance indicates how well members of a market segment can be identified given the descriptor variables. We also learn which descriptor variables are critical to the identification of segment membership, especially if methods are used that simultaneously select variables.

Regression analysis is the basis of prediction models. Regression analysis assumes that a dependent variable y can be predicted using independent variables or regressors $x_1, \ldots, x_p$:

$$y \approx f(x_1, \ldots, x_p).$$

Regression models differ with respect to the function $f(\cdot)$, the distribution assumed for y, and the deviations between y and $f(x_1, \ldots, x_p)$.

The basic regression model is the linear regression model. The linear regression model assumes that function $f(\cdot)$ is linear, and that y follows a normal distribution with mean $f(x_1, \ldots, x_p)$ and variance σ^2. The relationship between the dependent variable y and the independent variables $x_1, \ldots, x_p$ is given by:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

In R, function `lm()` fits a linear regression model. We fit the model for age in dependence of segment membership using:

```
R> lm(Age ~ C6 - 1, data = vacmotdesc)

Call:
lm(formula = Age ~ C6 - 1, data = vacmotdesc)

Coefficients:
 C61   C62   C63   C64   C65   C66
44.6  42.7  42.3  44.4  39.4  49.6
```

In R, regression models are specified using a formula interface. In the formula, the dependent variable AGE is indicated on the left side of the ~. The independent variables are indicated on the right side of the ~. In this particular case, we only use segment membership C6 as independent variable. Segment membership C6 is a categorical variable with six categories, and is coded as a factor in the data frame vacmotdesc. The formula interface correctly interprets categorical variables, and fits a regression coefficient for each category. For identifiability reasons, either the intercept β_0 or one category needs to be dropped. Using - 1 on the right hand side of ~ drops the intercept β_0. Without an intercept, each estimated coefficient is equal to the mean age in this segment. The output indicates that members of segment 5 are the youngest with a mean age of 39.4 years, and members of segment 6 are the oldest with a mean age of 49.6 years.

Including the intercept β_0 in the model formula drops the regression coefficient for segment 1. Its effect is instead captured by the intercept. The other regression coefficients indicate the mean age difference between segment 1 and each of the other segments:

```
R> lm(Age ~ C6, data = vacmotdesc)

Call:
lm(formula = Age ~ C6, data = vacmotdesc)

Coefficients:
(Intercept)          C62          C63          C64
     44.609       -1.947       -2.298       -0.191
         C65          C66
     -5.236        5.007
```

The intercept β_0 indicates that respondents in segment 1 are, on average, 44.6 years old. The regression coefficient C66 indicates that respondents in segment 6 are, on average, 5 years older than those in segment 1.

In linear regression models, regression coefficients express how much the dependent variable changes if one independent variable changes while all other independent variables remain constant. The linear regression model assumes that changes caused by changes in one independent variable are independent of the absolute level of all independent variables.

The dependent variable in the linear regression model follows a normal distribution. *Generalised linear models* (Nelder and Wedderburn 1972) can accommodate a wider range of distributions for the dependent variable. This is important if the dependent variable is categorical, and the normal distribution, therefore, is not suitable.

In the linear regression model, the mean value of y given $x_1, \ldots, x_p$ is modelled by the linear function:

$$\mathbb{E}[y|x_1, \ldots, x_p] = \mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Generalised linear models y are not limited to the normal distribution. We could, for example, use the Bernoulli distribution with y taking values 0 or 1. In this case, the mean value of y can only take values in $(0, 1)$. It is therefore not possible to describe the mean value with a linear function which can take any real value. Generalised linear models account for this by introducing a link function $g(\cdot)$. The link function transforms the mean value of y given by μ to an unlimited range indicated by η. This transformed value can then be modelled with a linear function:

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

η is referred to as linear predictor.

We can use the normal, Poisson, binomial, and multinomial distribution for the dependent variable in generalised linear models. The binomial or multinomial distribution are necessary for classification. A generalised linear model is characterised by the distribution of the dependent variable, and the link function. In the following sections we discuss two special cases of generalised linear models: binary and multinomial logistic regression. In these models the dependent variable follows either a binary or a multinomial distribution, and the link function is the logit function.

9.4.1 Binary Logistic Regression

We can formulate a regression model for binary data using generalised linear models by assuming that $f(y|\mu)$ is the Bernoulli distribution with success probability μ, and by choosing the logit link that maps the success probability $\mu \in (0, 1)$ onto $(-\infty, \infty)$ by

$$g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right).$$

Function `glm()` fits generalised linear models in R. The distribution of the dependent variable and the link function are specified by a `family`. The Bernoulli distribution with logit link is `family = binomial(link = "logit")` or `family = binomial()` because the logit link is the default. The binomial distribution is a generalisation of the Bernoulli distribution if the variable y does not only take values 0 and 1, but represents the number of successes out of a number of independent Bernoulli distributed trials with the same success probability μ.

Here, we fit the model to predict the likelihood of a consumer to belong to segment 3 given their age and moral obligation score. We specify the model using the formula interface with the dependent variable on the left of ~, and the two independent variables AGE and OBLIGATION2 on the right of ~. The dependent variable is a binary indicator of being in segment 3. This binary indicator is constructed with `I(C6 == 3)`. Function `glm()` fits the model given the formula, the data set, and the family:

```
R> f <- I(C6 == 3) ~ Age + Obligation2
R> model.C63 <- glm(f, data = vacmotdesc,
+   family = binomial())
R> model.C63

Call:  glm(formula = f, family = binomial(),
    data = vacmotdesc)

Coefficients:
  (Intercept)            Age  Obligation2Q2  Obligation2Q3
     -0.72197       -0.00842       -0.41900       -0.72285
Obligation2Q4
     -0.92526

Degrees of Freedom: 999 Total (i.e. Null);  995 Residual
Null Deviance:            924
Residual Deviance: 904         AIC: 914
```

The output contains the regression coefficients, and information on the model fit, including the degrees of freedom, the null deviance, the residual deviance, and the AIC.

The intercept in the linear regression model gives the mean value of the dependent variable if the independent variables $x_1, \ldots, x_p$ all have a value of 0. In binomial logistic regression, the intercept gives the value of the linear predictor η if the independent variables $x_1, \ldots, x_p$ all have a value of 0. The probability of being in segment 3 for a respondent with age 0 and a low moral obligation value is calculated by transforming the intercept with the inverse link function, in this case the inverse logit function:

$$g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Transforming the intercept value of -0.72 with the inverse logit link gives a predicted probability of 33% that a consumer of age 0 with low moral obligation is in segment 3.

The other regression coefficients in a linear regression model indicate how much the mean value of the dependent variable changes if this independent variable changes while others remain unchanged. In binary logistic regression, the regression coefficients indicate how the linear predictor changes. The changes in the linear predictor correspond to changes in the log odds of success. The odds of success are

the ratio between the probability of success μ and the probability of failure $1-\mu$. If the odds are equal to 1, success and failure are equally likely. If the odds are larger than 1, success is more likely than failure. Odds are frequently also used in betting.

The coefficient for AGE indicates that the log odds for being in segment 3 are 0.008 lower for tourists who are one year older. This means that the odds of one tourist are $e^{-0.008} = 0.992$ times the odds of another tourist if they only differ by the other tourist being one year younger. The independent variable OBLIGATION2 is a categorical variable with four different levels. The lowest category `Q1` is captured by the intercept. The regression coefficients for this variable indicate the change in log odds between the other categories and the lowest category `Q1`.

To simplify the interpretation of the coefficients and their effects, we can use package effects (Fox 2003; Fox and Hong 2009) in R. Function `allEffects` calculates the predicted values for different levels of the independent variable keeping other independent variables constant at their average value. In the case of the fitted binary logistic regression, the predicted values are the probabilities of being in segment 3. We plot the estimated probabilities to allow for easy inspection:

```
R> library("effects")
R> plot(allEffects(mod = model.C63))
```

Figure 9.11 shows how the predicted probability of being in segment 3 changes with age (on the left), and with moral obligation categories (on the right). The predicted probabilities are shown with pointwise 95% confidence bands (grey shaded areas) for metric independent variables, and with 95% confidence intervals for each category (vertical lines) for categorical independent variables. The predicted probabilities result from transforming the linear predictor with a non-linear function. The changes are not linear, and depend on the values of the other independent variables.

The plot on the left in Fig. 9.11 shows that, for a 20-year old tourist with an average moral obligation score, the predicted probability to be in segment 3 is about 20%. This probability decreases with increasing age. For 100-year old tourists the predicted probability to be in segment 3 is only slightly higher than 10%. The confidence bands indicate that these probabilities are estimated with high uncertainty. The fact that we can place into the plot a horizontal line lying completely within the grey shaded area, indicates that differences in AGE do not significantly affect the probability to be in segment 3. Dropping AGE from the regression model does not significantly decrease model fit.

The plot on the right side of Fig. 9.11 shows that the probability of being a member of segment 3 decreases with increasing moral obligation. Respondents of average age with a moral obligation value of `Q1` have a predicted probability of about 25% to be in segment 3. If these tourists of average age have the highest moral obligation value of `Q4`, they have a predicted probability of 12%. The 95% confidence intervals of the estimated effects indicate that – despite high

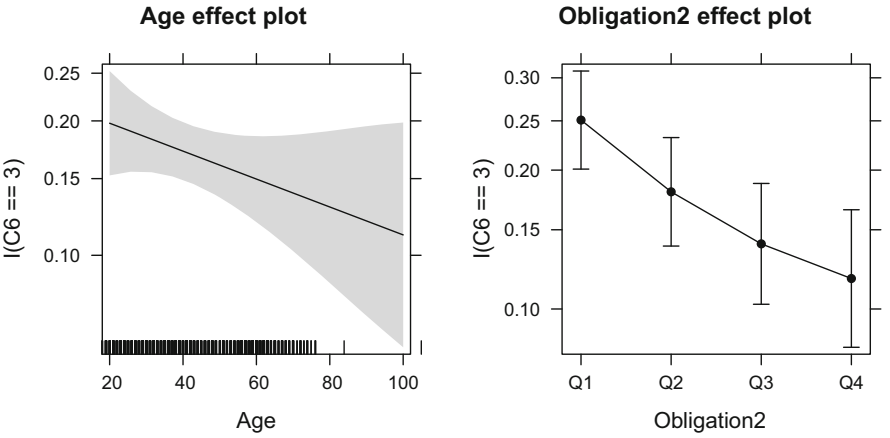

Fig. 9.11 Effect visualisation of age and moral obligation for predicting segment 3 using binary logistic regression for the Australian travel motives data set

uncertainty – probabilities do not overlap for the two most extreme values of moral obligation. This means that including moral obligation in the logistic regression model significantly improves model fit.

Summarising the fitted model provides additional insights:

```
R> summary(model.C63)

Call:
glm(formula = f, family = binomial(), data = vacmotdesc)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.835  -0.653  -0.553  -0.478   2.284

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.72197    0.28203   -2.56  0.01047 *
Age           -0.00842    0.00588   -1.43  0.15189
Obligation2Q2 -0.41900    0.21720   -1.93  0.05372 .
Obligation2Q3 -0.72285    0.23141   -3.12  0.00179 **
Obligation2Q4 -0.92526    0.25199   -3.67  0.00024 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 924.34  on 999  degrees of freedom
Residual deviance: 903.61  on 995  degrees of freedom
AIC: 913.6

Number of Fisher Scoring iterations: 4
```

The output contains the table of the estimated coefficients and their standard errors, the test statistics of a z-test, and the associated p-values. The z-test compares the fitted model to a model where this regression coefficient is set to 0. Rejecting the null hypothesis implies that the regression coefficient is not equal to 0 and this effect should be contained in the model.

This means that the null hypothesis is not rejected for AGE. We can drop AGE from the model without significantly decreasing model fit. If moral obligation is included in the model, AGE does not need to be included.

For moral obligation, three regression coefficients are fitted which capture the difference of categories `Q2`, `Q3` and `Q4` to category `Q1`. Each of the tests only compares the full model with the model with the regression coefficient of a specific category set to 0. This does not allow to decide if the model containing moral obligation performs better than the model without moral obligation. Function `Anova` from package car (Fox and Weisberg 2011) compares the model where moral obligation is dropped, and thus all regression coefficients for this variable are set to 0. We drop each of the independent variables one at a time, and compare the resulting model to the full model:

```
R> library("car")
R> Anova(model.C63)

Analysis of Deviance Table (Type II tests)

Response: I(C6 == 3)
            LR Chisq Df Pr(>Chisq)
Age             2.07  1    0.15024
Obligation2    17.26  3    0.00062 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows – for each independent variable in the model – the test statistic (`LR Chisq`), the degrees of freedom of the distribution to calculate the p-value (`Df`), and the p-value.

The test performed for the metric variable AGE is essentially the same as the z-test included in the `summary` output (use `Anova` with `test.statistic = "Wald"` for the exactly same test). The test indicates that dropping the categorical variable OBLIGATION2 would significantly reduce model fit. Moral obligation is a useful descriptor variable to predict membership in segment 3.

So far we fitted a binary logistic regression including two descriptor variables and simultaneously accounted for their association with the dependent variable. We can add additional independent variables to the binary logistic regression model. We

include all available descriptor variables in a regression model in R by specifying a dot on the right side of the ~. The variables included in the data frame in the `data` argument are then all used as independent variables (if not already used on the left of ~).

```
R> full.model.C63 <- glm(I(C6 == 3) ~ .,
+   data = na.omit(vacmotdesc), family = binomial())
```

Some descriptor variables contain missing values (`NA`). Respondents with at least one missing value are omitted from the data frame using `na.omit(vacmotdesc)`.

Including all available descriptor variables may lead to an overfitting model. An overfitting model has a misleadingly good performance, and overestimates effects of independent variables. Model selection methods exclude irrelevant independent variables. In R, function `step` performs model selection. The `step` function implements a stepwise procedure. In each step, the function evaluates if dropping an independent variable or adding an independent variable improves model fit. Model fit is assessed with the AIC. The AIC balanced goodness-of-fit with a penalty for model complexity. The function then drops or adds the variable leading to the largest improvement in AIC value. This procedure continues until no improvement in AIC is achieved by dropping or adding one independent variable.

```
R> step.model.C63 <- step(full.model.C63, trace = 0)
R> summary(step.model.C63)

Call:
glm(formula = I(C6 == 3) ~ Education + NEP +
    Vacation.Behaviour, family = binomial(),
    data = na.omit(vacmotdesc))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.051  -0.662  -0.545  -0.425   2.357

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.9359     0.6783    1.38  0.16762
Education           0.0571     0.0390    1.47  0.14258
NEP                -0.3139     0.1658   -1.89  0.05838 .
Vacation.Behaviour -0.5767     0.1504   -3.83  0.00013 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 802.23  on 867  degrees of freedom
Residual deviance: 773.19  on 864  degrees of freedom
AIC: 781.2

Number of Fisher Scoring iterations: 4
```

We suppress the printing of progress information of the iterative fitting function on screen using `trace = 0`. The selected final model is summarised. The model includes three variables: EDUCATION, NEP, and VACATION.BEHAVIOUR.

We compare the predictive performance of the model including AGE and MORAL.OBLIGATION with the model selected using `step`. A well predicting model would assign a high probability of being in segment 3 to members of segment 3 and a low probability to all other consumers. Function `predict()` returns the predicted probabilities of being in segment 3 for all consumers if the function is applied to a fitted model, and we specify `type = "response"`. Parallel boxplots visualise the distributions of predicted probabilities for consumers in segment 3, and those not in segment 3:

```
R> par(mfrow = c(1, 2))
R> prob.C63 <- predict(model.C63, type = "response")
R> boxplot(prob.C63 ~ I(C6 == 3), data = vacmotdesc,
+   ylim = 0:1, main = "", ylab = "Predicted probability")
R> prob.step.C63 <- predict(step.model.C63, type = "response")
R> boxplot(prob.step.C63 ~ I(C6 == 3),
+   data = na.omit(vacmotdesc), ylim = 0:1,
+   main = "", ylab = "Predicted probability")
```

Figure 9.12 compares the predicted probabilities of segment 3 membership for the two models. If the fitted model differentiates well between members of segment 3 and all other consumers, the boxes are located at the top of the plot (close to the value of 1) for respondents in segment 3 (TRUE), and at the bottom (close to the

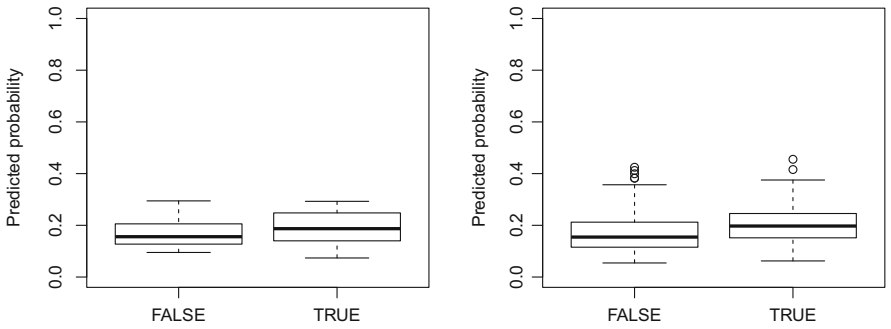

Fig. 9.12 Predicted probabilities of segment 3 membership for consumers not assigned to segment 3 (FALSE) and for consumers assigned to segment 3 (TRUE) for the Australian travel motives data set. The model containing age and moral obligation as independent variables is on the left; the model selected using stepwise variable selection on the right

value of 0) for all other consumers. We can see from Fig. 9.12 that the performance of the two fitted models is nowhere close to this optimal case. The median predicted values are only slightly higher for segment 3 in both models. The difference is larger for the model fitted using `step`, indicating that the predictive performance of this model is slightly better.

9.4.2 Multinomial Logistic Regression

Multinomial logistic regression can fit a model that predicts each segment simultaneously. Because segment extraction typically results in more than two market segments, the dependent variable y is not binary. Rather, it is categorical and assumed to follow a multinomial distribution with the logistic function as link function.

In R, function `multinom()` from package nnet (Venables and Ripley 2002) (instead of `glm`) fits a multinomial logistic regression. We specify the model in a similar way using a formula and a data frame for evaluating the formula.

```
R> library("nnet")
R> vacmotdesc$Oblig2 <- vacmotdesc$Obligation2
R> model.C6 <- multinom(C6 ~ Age + Oblig2,
+   data = vacmotdesc, trace = 0)
```

Using `trace = 0` avoids the display of progress information of the iterative fitting function.

The fitted model contains regression coefficients for each segment except for segment 1 (the baseline category). The same set of regression coefficients would result from a binary logistic regression model comparing this segment to segment 1. The coefficients indicate the change in log odds if the independent variable changes:

```
R> model.C6

Call:
multinom(formula = C6 ~ Age + Oblig2, data = vacmotdesc,
    trace = 0)

Coefficients:
  (Intercept)     Age Oblig2Q2 Oblig2Q3 Oblig2Q4
2       0.184 -0.0092    0.108   -0.026    -0.16
3       0.417 -0.0103   -0.307   -0.541    -0.34
4      -0.734 -0.0017    0.309    0.412     0.42
5      -0.043 -0.0296   -0.023   -0.039     1.33
6      -2.090  0.0212    0.269    0.790     1.65

Residual Deviance: 3384
AIC: 3434
```

The regression coefficients are arranged in matrix form. Each row contains the regression coefficients for one category of the dependent variable. Each column contains the regression coefficients for one effect of an independent variable.

The `summary()` function returns the regression coefficients and their standard errors.

```
R> summary(model.C6)

Call:
multinom(formula = C6 ~ Age + Oblig2, data = vacmotdesc,
    trace = 0)

Coefficients:
  (Intercept)     Age Oblig2Q2 Oblig2Q3 Oblig2Q4
2       0.184 -0.0092    0.108   -0.026    -0.16
3       0.417 -0.0103   -0.307   -0.541    -0.34
4      -0.734 -0.0017    0.309    0.412     0.42
5      -0.043 -0.0296   -0.023   -0.039     1.33
6      -2.090  0.0212    0.269    0.790     1.65

Std. Errors:
  (Intercept)    Age Oblig2Q2 Oblig2Q3 Oblig2Q4
2        0.34 0.0068     0.26     0.26     0.31
3        0.34 0.0070     0.26     0.27     0.31
4        0.39 0.0075     0.30     0.30     0.34
5        0.44 0.0091     0.37     0.38     0.35
6        0.42 0.0073     0.34     0.32     0.32

Residual Deviance: 3384
AIC: 3434
```

With function `Anova()` we assess if dropping a single variable significantly reduces model fit. Dropping a variable corresponds to setting all regression coefficients of this variable to 0. This means that the regression coefficients in one or several columns of the regression coefficient matrix corresponding to this variable are set to 0. Function `Anova()` tests if dropping any of the variables significantly reduces model fit. The output is essentially the same as for the binary logistic regression model:

```
R> Anova(model.C6)

Analysis of Deviance Table (Type II tests)

Response: C6
       LR Chisq Df Pr(>Chisq)
Age        35.6  5    1.1e-06 ***
Oblig2     89.0 15    1.5e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output indicates that dropping any of the variables leads to a significant reduction in model fit. Applying function `step()` to a fitted model performs model selection. Starting with the full model containing all available independent variables,

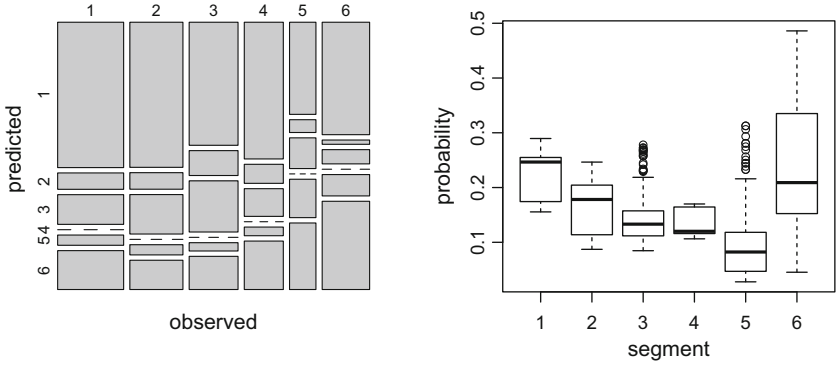

Fig. 9.13 Assessment of predictive performance of the multinomial logistic regression model including age and moral obligation as independent variables for the Australian travel motives data set. The mosaic plot of the cross-tabulation of observed and predicted segment memberships is on the left. The parallel boxplot of the predicted probabilities by segment for consumers assigned to segment 6 is on the right

the stepwise procedure returns the best-fitting model, the model which deteriorates in AIC if an independent variable is either dropped or additionally included.

We assess the predictive performance of the fitted model by comparing the predicted segment membership to the observed segment membership. Figure 9.13 shows a mosaic plot of the predicted and observed segment memberships on the left. In addition, we investigate the distribution of the predicted probabilities for each segment. Figure 9.13 shows parallel boxplots of the predicted segment probabilities for consumers assigned to segment 6 on the right:

```
R> par(mfrow = c(1, 2))
R> pred.class.C6 <- predict(model.C6)
R> plot(table(observed = vacmotdesc$C6,
+   predicted = pred.class.C6), main = "")
R> pred.prob.C6 <- predict(model.C6, type = "prob")
R> predicted <- data.frame(prob = as.vector(pred.prob.C6),
+   observed = C6,
+   predicted = rep(1:6, each = length(C6)))
R> boxplot(prob ~ predicted,
+   xlab = "segment", ylab = "probability",
+   data = subset(predicted, observed == 6))
```

By default `predict` returns the predicted classes. Adding the argument `type = "prob"` returns the predicted probabilities.

The left panel of Fig. 9.13 shows that none of the consumers are predicted to be in segment 4. Most respondents are predicted to belong to segment 1, the largest segment. The detailed results for segment 6 (right panel of Fig. 9.13) indicate that consumers from this segment have particularly low predicted probabilities to belong to segment 5.

To ease interpretation of the estimated effects, we use function `allEffects`, and plot the predicted probabilities:

```
R> plot(allEffects(mod = model.C6), layout = c(3, 2))
```

The left panel in Fig. 9.14 shows how the predicted probability to belong to any segment changes with age for a consumer with average moral obligation. The predicted probability for each segment is visualised separately. The heading indicates the segments. For example, `C6 = 1` indicates that the panel contains predicted probabilities for segment 1. Shaded grey areas indicate pointwise 95% confidence bands visualising the uncertainty of the estimated probabilities.

The predicted probability to belong to segment 6 increases with age: young respondents belong to segment 6 with a probability of less than 10%. Older respondents have a probability of about 40%. The probability of belonging to segment 5 decreases with age.

The right panel in Fig. 9.14 shows how the predicted segment membership probability changes with moral obligation values for a consumer of average age. The predicted probability to belong to segment 6 increases with increasing moral obligation value. Respondents with the lowest moral obligation value of `Q1` have a probability of about 8% to be from segment 6. This increases to 29% for respondents with a moral obligation value of `Q4`. For segment 3 the reverse is true: respondents with higher moral obligation values have lower probabilities to be from segment 3.

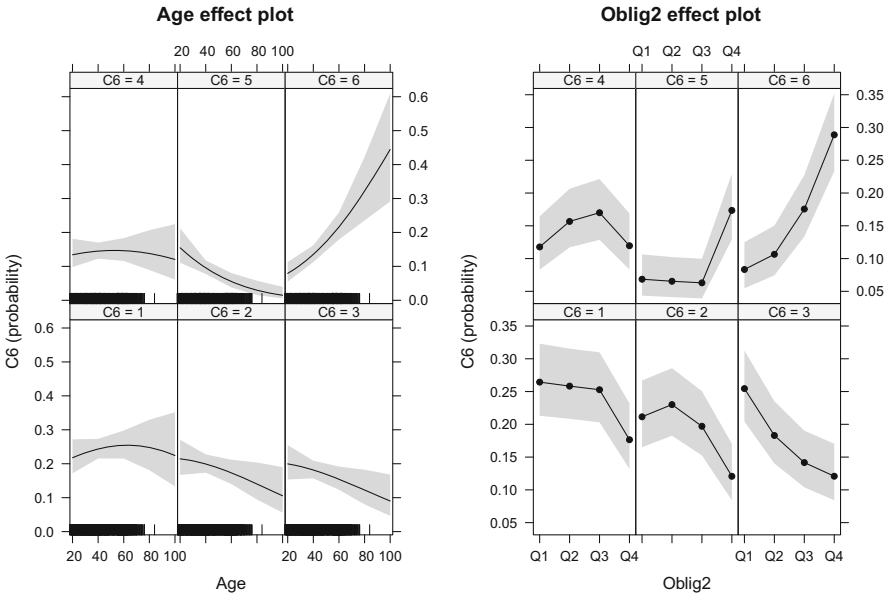

Fig. 9.14 Effect visualisation of age and moral obligation for predicting segment membership using multinomial logistic regression for the Australian travel motives data set

9.4.3 Tree-Based Methods

Classification and regression trees (CARTs; Breiman et al. 1984) are an alternative modelling approach for predicting a binary or categorical dependent variable given a set of independent variables. Classification and regression trees are a supervised learning technique from machine learning. The advantages of classification and regression trees are their ability to perform variable selection, ease of interpretation supported by visualisations, and the straight-forward incorporation of interaction effects. Classification and regression trees work well with a large number of independent variables. The disadvantage is that results are frequently unstable. Small changes in the data can lead to completely different trees.

The tree approach uses a stepwise procedure to fit the model. At each step, consumers are split into groups based on one independent variable. The aim of the split is for the resulting groups to be as pure as possible with respect to the dependent variable. This means that consumers in the resulting groups have similar values for the dependent variable. In the best case, all group members have the same value for a categorical dependent variable. Because of this stepwise splitting procedure, the classification and regression tree approach is also referred to as *recursive partitioning*.

The resulting tree (see Figs. 9.15, 9.16, and 9.17) shows the nodes that emerge from each splitting step. The node containing all consumers is the *root node*. Nodes that are not split further are *terminal nodes*. We predict segment membership by moving down the tree. At each node, we move down the branch reflecting the consumer's independent variable. When we reach the terminal node, segment membership can be predicted based on the segment memberships of consumers contained in the terminal node.

Tree constructing algorithms differ with respect to:

- Splits into two or more groups at each node (binary vs. multi-way splits)
- Selection criterion for the independent variable for the next split
- Selection criterion for the split point of the independent variable
- Stopping criterion for the stepwise procedure
- Final prediction at the terminal node

Several R packages implement tree constructing algorithms. Package rpart (Therneau et al. 2017) implements the algorithm proposed by Breiman et al. (1984). Package partykit (Hothorn and Zeileis 2015) implements an alternative tree constructing procedure that performs unbiased variable selection. This means that the procedure selects independent variables on the basis of association tests and their p-values (see Hothorn et al. 2006). Package partykit also enables visualisation of the fitted tree models.

Function `ctree()` from package partykit fits a conditional inference tree. As an example, we use the Australian travel motives data set with the six-segment solution extracted using neural gas clustering in Sect. 7.5.4. We use membership

in segment 3 as a binary dependent variable, and include all available descriptor variables as independent variables:

```
R> set.seed(1234)
R> library("partykit")
R> tree63 <- ctree(factor(C6 == 3) ~ .,
+   data = vacmotdesc)
R> tree63

Model formula:
factor(C6 == 3) ~ Gender + Age + Education +
    Income + Income2 + Occupation + State +
    Relationship.Status + Obligation + Obligation2 +
    NEP + Vacation.Behaviour + Oblig2

Fitted party:
[1] root
|   [2] Vacation.Behaviour <= 2.2: FALSE (n = 130,
     err = 32%)
|   [3] Vacation.Behaviour > 2.2
|   |   [4] Obligation <= 3.9: FALSE (n = 490, err = 19%)
|   |   [5] Obligation > 3.9: FALSE (n = 380, err = 11%)

Number of inner nodes:    2
Number of terminal nodes: 3
```

The output describes the fitted classification tree shown in Fig. 9.15. The classification tree starts with a root node containing all consumers. Next, the root note is split into two nodes (numbered 2 and 3) using the independent variable VACATION.BEHAVIOUR. The split point is 2.2. This means that consumers with a VACATION.BEHAVIOUR score of 2.2 or less are assigned to node 2. Consumers with a score higher than 2.2 are assigned to node 3. Node 2 is not split further; it becomes a terminal node. The predicted value for this particular terminal node is `FALSE`. The number of consumers in this terminal node is shown in brackets (`n = 130`), along with the proportion of wrongly classified respondents (`err = 32%`). Two thirds of consumers in this node are not in segment 3, one third is. Node 3 is split into two nodes (numbered 4 and 5) using the independent variable OBLIGATION. Consumers with an OBLIGATION score of 3.9 or less are assigned to node 4. Consumers with a higher score are assigned to node 5. The tree predicts that respondents in node 4 are not in segment 3. Node 4 contains 490 respondents; 81% of them are not in segment 3, 19% are. Most respondents in node 5 are also not in segment 3. Node 5 contains 380 respondents; 11% of them are in segment 3. The output also shows that there are 2 inner nodes (numbered 1 and 3), and 3 terminal nodes (numbered 2, 4, and 5).

Plotting the classification tree using `plot(tree63)` gives a visual representation that is easier to interpret. Figure 9.15 visualises the classification tree. The root node on the top has the number 1. The root node contains the name of the variable used for the first split (VACATION.BEHAVIOUR), as well as the p-value of the association test that led to the selection of this particular variable (`p <`

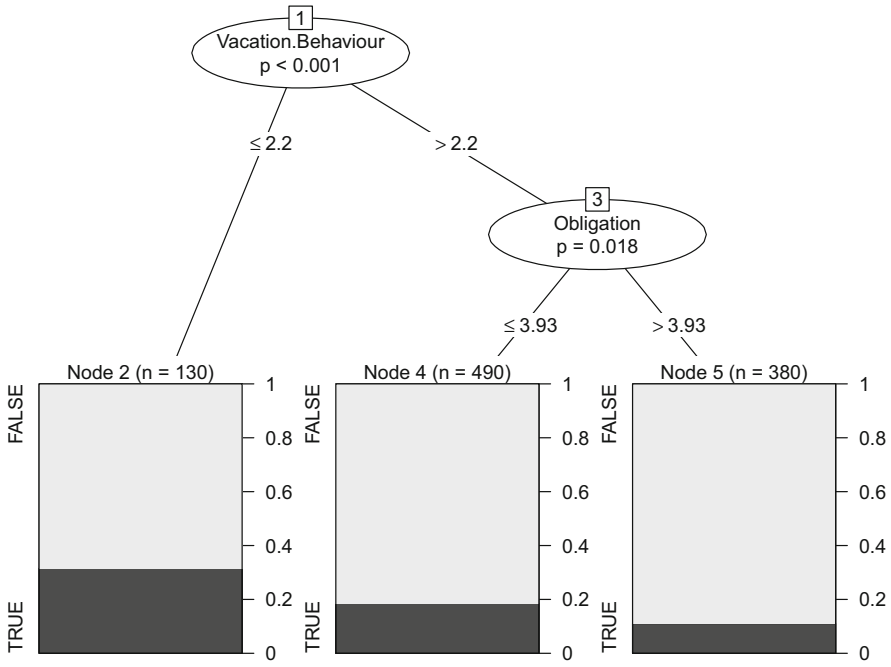

Fig. 9.15 Conditional inference tree using membership in segment 3 as dependent variable for the Australian travel motives data set

`0.001`). The lines underneath the node indicate the split or threshold value of the independent variable VACATION.BEHAVIOUR where respondents are directed to the left or right branch. Consumers with a value higher than 2.2 follow the right branch to node 3. Consumers with a value of 2.2 or less follow the left branch to node 2. These consumers are not split up further; node 2 is a terminal node. The proportion of respondents in node 2 who belong to segment 3 is shown at the bottom of the stacked bar chart for node 2. The dark grey area represents this proportion, and the label on the *y*-axis indicates that this is for the category `TRUE`. The proportion of consumers in node 2 not belonging to segment 3 is shown in light grey with label `FALSE`.

Node 3 is split further using OBLIGATION as the independent variable. The split value is 3.9. Using this split value, consumers are assigned to either node 4 or node 5. Both are terminal nodes. Stacked barplots visualise the proportion of respondents belonging to segment 3 for nodes 4 and 5.

This tree plot indicates that the group with a low mean score for environmentally friendly behaviour on vacation contains the highest proportion of segment 3 members. The group with a high score for environmental friendly behaviour and moral obligation, contains the smallest proportion of segment 3 members. The dark grey area is largest for node 1 and lowest for node 5.

Package partykit takes a number of parameters for the algorithm set by the `control` argument with function `ctree_control`. These parameters influence the tree construction by restricting nodes considered for splitting, by specifying the minimum size for terminal nodes, by selecting the test statistic for the association test, and by setting the minimum value of the criterion of the test to implement a split.

As an illustration, we fit a tree with segment 6 membership as dependent variable. We ensure that terminal nodes contain at least 100 respondents (`minbucket = 100`), and that the minimum criterion value (`mincriterion`) is 0.99 (corresponding to a p-value of smaller than 0.01). Figure 9.16 visualises this tree.

```
R> tree66 <- ctree(factor(C6 == 6) ~ .,
+   data = vacmotdesc,
+   control = ctree_control(minbucket = 100,
+     mincriterion = 0.99))
R> plot(tree66)
```

The fitted classification tree for segment 6 is more complex than that for segment 3; the number of inner and terminal nodes is larger. The stacked bar charts for the terminal nodes indicate how pure the terminal nodes are, and how the terminal nodes differ in the proportion of segment 6 members they contain. The tree algorithm tries to maximise these differences. Terminal node 11 (on the right) contains the highest proportion of consumers assigned to segment 6. Node 11 contains respondents with the highest possible value for moral obligation, and a NEP score of at least 4.

We can also fit a tree for categorical dependent variables with more than two categories with function `ctree()`. Here, the dependent variable in the formula on the left is a categorical variable. `C6` is a factor containing six levels; each level indicates the segment membership of respondents.

```
R> tree6 <- ctree(C6 ~ ., data = vacmotdesc)
R> tree6

Model formula:
C6 ~ Gender + Age + Education + Income +
   Income2 + Occupation + State + Relationship.Status +
   Obligation + Obligation2 + NEP + Vacation.Behaviour +
   Oblig2

Fitted party:
[1] root
|   [2] Oblig2 in Q1, Q2, Q3
|   |   [3] Education <= 6: 1 (n = 481, err = 73%)
|   |   [4] Education > 6: 1 (n = 286, err = 77%)
|   [5] Oblig2 in Q4
|   |   [6] Obligation <= 4.7: 6 (n = 203, err = 67%)
|   |   [7] Obligation > 4.7: 5 (n = 30, err = 57%)

Number of inner nodes:    3
Number of terminal nodes: 4
```

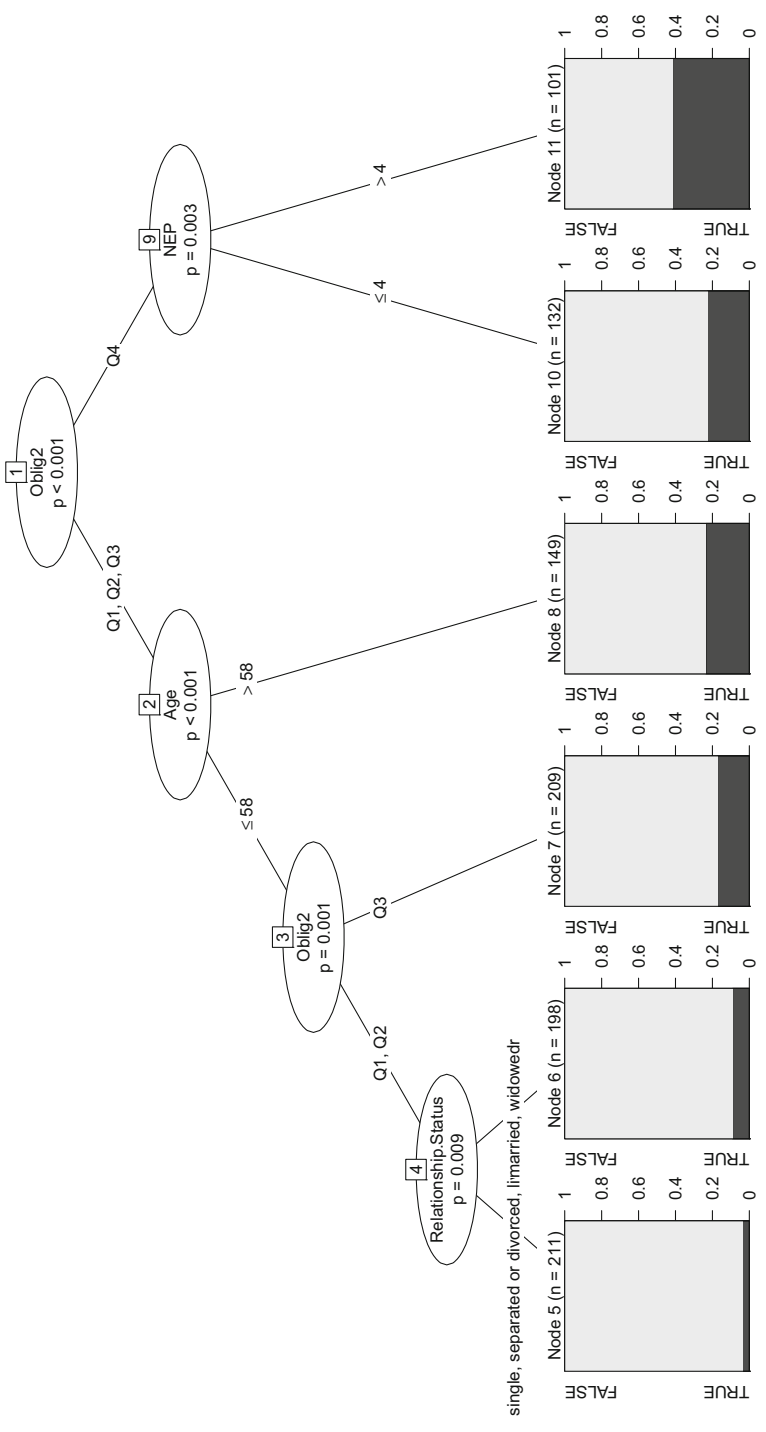

Fig. 9.16 Conditional inference tree using membership in segment 6 as dependent variable for the Australian travel motives data set

The output shows that the first splitting variable is the categorical variable indicating moral obligation (OBLIGATION2). This variable splits the root node 1 into nodes 2 and 5. Consumers with a moral obligation value of `Q1`, `Q2` and `Q3` are assigned to node 2. Consumers with a moral obligation value of `Q4` are assigned to node 7.

Node 2 is split into nodes 3 and 4 using EDUCATION as splitting variable. Consumers with an EDUCATION level of 6 or less are assigned to node 3. Node 3 is a terminal node. Most consumers in this terminal node belong to segment 1. Node 3 contains 481 respondents. Predicting segment membership as 1 for consumers in this node is wrong in 73% of cases.

Respondents with an EDUCATION level higher than 6 are assigned to node 4. Node 4 is a terminal node. The predicted segment membership for node 4 is 1. This node contains 286 respondents and 77% of them are not in segment 1.

Consumers in node 5 feel highly morally obliged to protect the environment. They are split into nodes 6 and 7 using the metric version of moral obligation as splitting variable. Node 6 contains respondents with a moral obligation value of 47 or less, and a moral obligation category value of `Q4`. Most respondents in node 6 belong to segment 6. The node contains 203 respondents; 67% are not from segment 6. Consumers with a moral obligation score higher than 4.7 are in node 7. The predicted segment membership for this node is 5. The node contains 30 consumers; 57% do not belong to segment 5.

Figure 9.17 visualises the tree. `plot(tree6)` creates this plot. Most of the plot is the same as for the classification tree with the binary dependent variable. Only the bar charts at the bottom look different. The terminal nodes show the proportion of respondents in each segment. Optimally, these bar charts for each terminal node show that nearly all consumers in that node have the same segment membership or are at least assigned to only a small number of different segments. Node 7 in Fig. 9.17 is a good example: it contains high proportions of members of segments 1 and 5, but only low proportions of members of other segments.

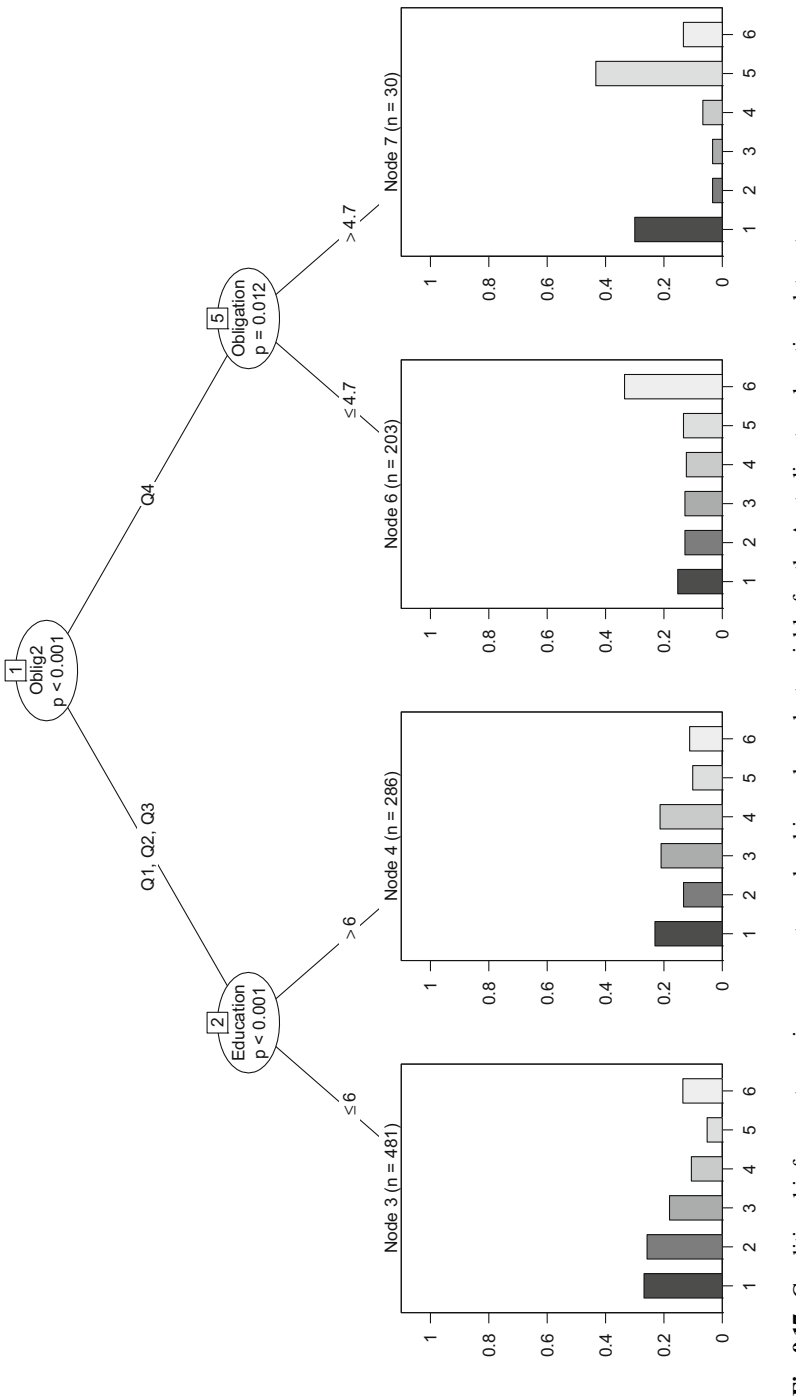

Fig. 9.17 Conditional inference tree using segment membership as dependent variable for the Australian travel motives data set

9.5 Step 7 Checklist

Task	Who is responsible?	Completed?
Bring across from Step 6 (profiling) one or a small number of market segmentation solutions selected on the basis of attractive profiles.		☐
Select descriptor variables. Descriptor variables are additional pieces of information about each consumer included in the market segmentation analysis. Descriptor variables have not been used to extract the market segments.		☐
Use visualisation techniques to gain insight into the differences between market segments with respect to descriptor variables. Make sure you use appropriate plots, for example, mosaic plots for categorical and ordinal descriptor variables, and box-and-whisker plots for metric descriptor variables.		☐
Test for statistical significance of descriptor variables.		☐
If you used separate statistical tests for each descriptor variable, correct for multiple testing to avoid overestimating significance.		☐
"Introduce" each market segment to the other team members to check how much you know about these market segments.		☐
Ask if additional insight into some segments is required to develop a full picture of them.		☐

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Chapman and Hall/CRC, New York

Cornelius B, Wagner U, Natter M (2010) Managerial applicability of graphical formats to support positioning decisions. Journal für Betriebswirtschaft 60(3):167–201

Dolnicar S, Leisch F (2008) An investigation of tourists' patterns of obligation to protect the environment. J Travel Res 46:381–391

Fox J (2003) Effect displays in R for generalised linear models. J Stat Softw 8(15):1–27

Fox J, Hong J (2009) Effect displays in R for multinomial and proportional-odds logit models: extensions to the effects package. J Stat Softw 32(1):1–24

Fox J, Weisberg S (2011) An R Companion to applied regression, 2nd edn. Sage, Thousand Oaks. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Friendly M (1994) Mosaic displays for multi-way contingency tables. J Amer Stat Assoc 89: 190–200

Hartigan JA, Kleiner B (1984) A mosaic of television ratings. Amer Statist 38:32–35
Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
Hothorn T, Zeileis A (2015) partykit: a modular toolkit for recursive partytioning in R. J Mach Learn Res 16:3905–3909. http://jmlr.org/papers/v16/hothorn15a.html
Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15(3):651–674
Miller RG (1981) Simultaneous statistical inference. Springer, Heidelberg
Nelder J, Wedderburn R (1972) Generalized linear models. J R Stat Soc A 135(3):370–384
van Raaij WF, Verhallen TMM (1994) Domain-specific market segmentation. Eur J Market 28(10):49–66
Sarkar D (2008) lattice: multivariate data visualization with R. Springer, New York
Therneau T, Atkinson B, Ripley B (2017) rpart: recursive partitioning and regression trees. https://CRAN.R-project.org/package=rpart, R package version 4.1–11
Tukey J (1949) Comparing individual means in the analysis of variance. Biometrics 5(2):99–114
Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York
Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer, New York. http://ggplot2.org
Yandell BS (1997) Practical data analysis for designed experiments. Chapman & Hall, New York