A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 7

Logistic Regression and Generalized Linear Models: Blood Screening, Women's Role in Society, Colonic Polyps, Driving and Back Pain, and Happiness in China

7.1 Introduction

7.2 Logistic Regression and Generalized Linear Models

7.3 Analysis Using R

7.3.1 ESR and Plasma Proteins

We can now fit a logistic regression model to the data using the `glm` function. We start with a model that includes only a single explanatory variable, `fibrinogen`. The code to fit the model is

```
R> plasma_glm_1 <- glm(ESR ~ fibrinogen, data = plasma,
+                      family = binomial())
```

The formula implicitly defines a parameter for the global mean (the intercept term) as discussed in Chapter 5 and Chapter 6. The distribution of the response is defined by the `family` argument, a binomial distribution in our case. (The default link function when the binomial family is requested is the logistic function.)

From the results in Figure 7.2 we see that the regression coefficient for fibrinogen is significant at the 5% level. An increase of one unit in this variable increases the log-odds in favor of an ESR value greater than 20 by an estimated 1.83 with 95% confidence interval

```
R> confint(plasma_glm_1, parm = "fibrinogen")
```

```
 2.5 % 97.5 %
 0.339  3.998
```

These values are more helpful if converted to the corresponding values for the odds themselves by exponentiating the estimate

```
R> exp(coef(plasma_glm_1)["fibrinogen"])
```

```
fibrinogen
      6.22
```

and the confidence interval

```
R> data("plasma", package = "HSAUR3")
R> layout(matrix(1:2, ncol = 2))
R> cdplot(ESR ~ fibrinogen, data = plasma)
R> cdplot(ESR ~ globulin, data = plasma)
```

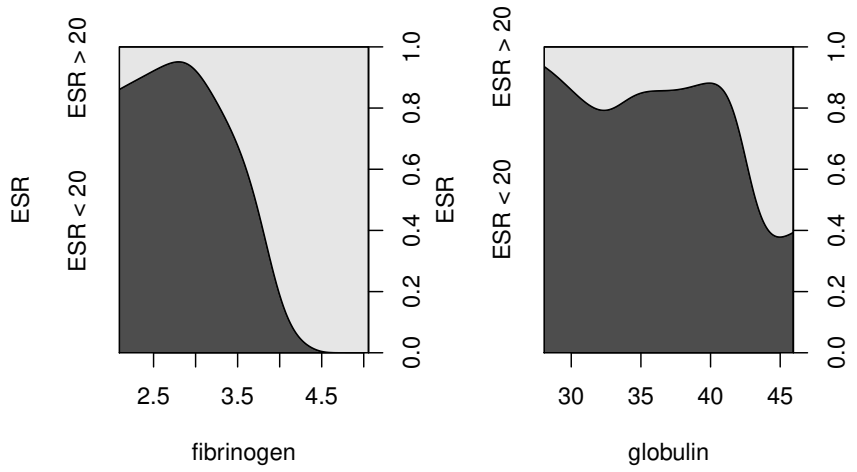

Figure 7.1 Conditional density plots of the erythrocyte sedimentation rate (ESR) given fibrinogen and globulin.

```
R> exp(confint(plasma_glm_1, parm = "fibrinogen"))
```

```
 2.5 % 97.5 %
   1.4   54.5
```

The confidence interval is very wide because there are few observations overall and very few where the ESR value is greater than 20. Nevertheless it seems likely that increased values of fibrinogen lead to a greater probability of an ESR value greater than 20.

We can now fit a logistic regression model that includes both explanatory variables using the code

```
R> plasma_glm_2 <- glm(ESR ~ fibrinogen +  globulin,
+      data = plasma, family = binomial())
```

and the output of the `summary` method is shown in Figure 7.3.

The coefficient for gamma globulin is not significantly different from zero. Subtracting the residual deviance of the second model from the corresponding value for the first model we get a value of 1.87. Tested using a χ^2-distribution with a single degree of freedom this is not significant at the 5% level and so we conclude that gamma globulin is not associated with ESR level. In R, the

```
R> summary(plasma_glm_1)

Call:
glm(formula = ESR ~ fibrinogen, family = binomial(), data = plasma)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.930  -0.540  -0.438  -0.336   2.479

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.845      2.770   -2.47    0.013
fibrinogen     1.827      0.901    2.03    0.043

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 24.840  on 30  degrees of freedom
AIC: 28.84

Number of Fisher Scoring iterations: 5
```

Figure 7.2 R output of the `summary` method for the logistic regression model fitted to ESR and fibrigonen.

```
R> summary(plasma_glm_2)

Call:
glm(formula = ESR ~ fibrinogen + globulin, family = binomial(),
    data = plasma)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.968  -0.612  -0.346  -0.212   2.264

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -12.792      5.796   -2.21    0.027
fibrinogen     1.910      0.971    1.97    0.049
globulin       0.156      0.120    1.30    0.193

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 22.971  on 29  degrees of freedom
AIC: 28.97

Number of Fisher Scoring iterations: 5
```

Figure 7.3 R output of the `summary` method for the logistic regression model fitted to ESR and both globulin and fibrinogen.

task of comparing the two nested models can be performed using the `anova` function

```
R> anova(plasma_glm_1, plasma_glm_2, test = "Chisq")

Analysis of Deviance Table

Model 1: ESR ~ fibrinogen
```

```
Model 2: ESR ~ fibrinogen + globulin
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        30       24.8
2        29       23.0  1     1.87     0.17
```

Nevertheless we shall use the predicted values from the second model and plot them against the values of *both* explanatory variables using a *bubbleplot* to illustrate the use of the `symbols` function. The estimated conditional probability of a ESR value larger 20 for all observations can be computed, following formula (**??**), by

```
R> prob <- predict(plasma_glm_2, type = "response")
```

and now we can assign a larger circle to observations with larger probability as shown in Figure 7.4. The plot clearly shows the increasing probability of an ESR value above 20 (larger circles) as the values of fibrinogen, and to a lesser extent, gamma globulin, increase.

7.3.2 Women's Role in Society

Originally the data in Table **??** would have been in a completely equivalent form to the data in Table **??** data, but here the individual observations have been grouped into counts of numbers of agreements and disagreements for the two explanatory variables, `gender` and `education`. To fit a logistic regression model to such grouped data using the `glm` function we need to specify the number of agreements and disagreements as a two-column matrix on the left-hand side of the model formula. We first fit a model that includes the two explanatory variables using the code

```
R> data("womensrole", package = "HSAUR3")
R> fm1 <- cbind(agree, disagree) ~ gender + education
R> womensrole_glm_1 <- glm(fm1, data = womensrole,
+                          family = binomial())
```

From the `summary` output in Figure 7.5 it appears that education has a highly significant part to play in predicting whether a respondent will agree with the statement read to them, but the respondent's gender is apparently unimportant. As years of education increase the probability of agreeing with the statement declines. We now are going to construct a plot comparing the observed proportions of agreeing with those fitted by our fitted model. Because we will reuse this plot for another fitted object later on, we define a function which plots years of education against some fitted probabilities, e.g.,

```
R> role.fitted1 <- predict(womensrole_glm_1, type = "response")
```

and labels each observation with the person's gender:

```
R> myplot <- function(role.fitted)  {
+      f <- womensrole$gender == "Female"
+      plot(womensrole$education, role.fitted, type = "n",
+           ylab = "Probability of agreeing",
+           xlab = "Education", ylim = c(0,1))
```

```
R> plot(globulin ~ fibrinogen, data = plasma, xlim = c(2, 6),
+       ylim = c(25, 55), pch = ".")
R> symbols(plasma$fibrinogen, plasma$globulin, circles = prob,
+          add = TRUE)
```

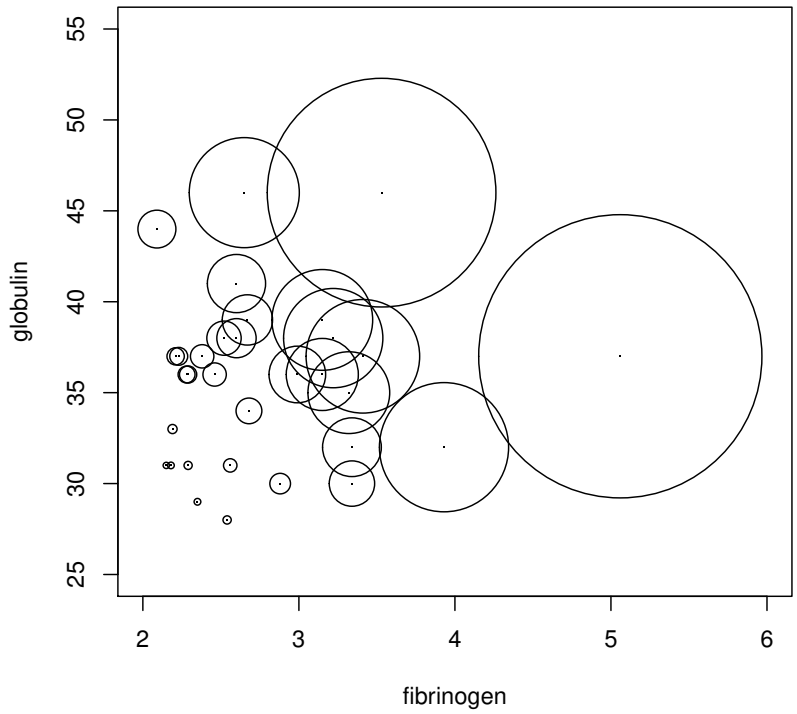

Figure 7.4 Bubbleplot of fitted values for a logistic regression model fitted to the `plasma` data.

```
+      lines(womensrole$education[!f], role.fitted[!f], lty = 1)
+      lines(womensrole$education[f], role.fitted[f], lty = 2)
+      lgtxt <- c("Fitted (Males)", "Fitted (Females)")
+      legend("topright", lgtxt, lty = 1:2, bty = "n")
+      y <-  womensrole$agree / (womensrole$agree +
+                                womensrole$disagree)
+      size <- womensrole$agree + womensrole$disagree
+      size <- size - min(size)
+      size <- (size / max(size)) * 3 + 1
```

```
R> summary(womensrole_glm_1)

Call:
glm(formula = fm1, family = binomial(), data = womensrole)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7254  -0.8630  -0.0652   0.8434   3.1332

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.5094     0.1839   13.65   <2e-16
genderFemale  -0.0114     0.0841   -0.14     0.89
education     -0.2706     0.0154  -17.56   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 451.722  on 40  degrees of freedom
Residual deviance:  64.007  on 38  degrees of freedom
AIC: 208.1

Number of Fisher Scoring iterations: 4
```

Figure 7.5 R output of the `summary` method for the logistic regression model fitted to the `womensrole` data.

```
15 +      text(womensrole$education, y, ifelse(f, "\\VE", "\\MA"),
16 +           family = "HersheySerif", cex = size)
17 +  }
```

In lines 3–5 of function `myplot`, an empty scatterplot of education and fitted probabilities (`type = "n"`) is set up, basically to set the scene for the following plotting actions. Then, two lines are drawn (using function `lines` in lines 6 and 7), one for males (with line type 1) and one for females (with line type 2, i.e., a dashed line), where the logical vector `f` describes both genders. In line 9 a legend is added. Finally, in lines 12 onwards we plot 'observed' values, i.e., the frequencies of agreeing in each of the groups (`y` as computed in lines 10 and 11) and use the Venus and Mars symbols to indicate gender. The size of the plotted symbol is proportional to the numbers of observations in the corresponding group of gender and years of education.

The two curves for males and females in Figure 7.6 are almost the same reflecting the non-significant value of the regression coefficient for gender in `womensrole_glm_1`. But the observed values plotted on Figure 7.6 suggest that there might be an interaction of education and gender, a possibility that can be investigated by applying a further logistic regression model using

```
R> fm2 <- cbind(agree,disagree) ~ gender * education
R> womensrole_glm_2 <- glm(fm2, data = womensrole,
+                          family = binomial())
```

The `gender` and `education` interaction term is seen to be highly significant, as can be seen from the `summary` output in Figure 7.7.

We can obtain a plot of deviance residuals plotted against fitted values using the following code above Figure 7.9. The residuals fall into a horizontal band

```
R> myplot(role.fitted1)
```

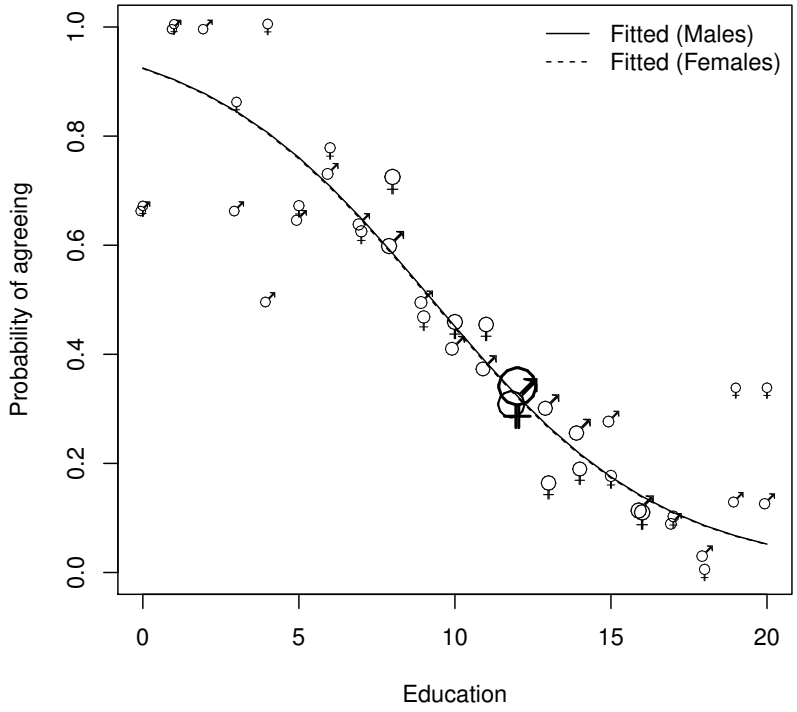

Figure 7.6 Fitted (from `womensrole_glm_1`) and observed probabilities of agreeing for the `womensrole` data. The size of the symbols is proportional to the sample size.

between −2 and 2. This pattern does not suggest a poor fit for any particular observation or subset of observations.

7.3.3 Colonic Polyps

The data on colonic polyps in Table **??** involves *count* data. We could try to model this using multiple regression but there are two problems. The first is that a response that is a count can take only positive values, and secondly such a variable is unlikely to have a normal distribution. Instead we will apply a GLM with a log link function, ensuring that fitted values are positive, and

```
R> summary(womensrole_glm_2)

Call:
glm(formula = fm2, family = binomial(), data = womensrole)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3910  -0.8806   0.0153   0.7278   2.4526

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.0982     0.2355    8.91   <2e-16
genderFemale           0.9047     0.3601    2.51   0.0120
education             -0.2340     0.0202  -11.59   <2e-16
genderFemale:education  -0.0814     0.0311   -2.62   0.0089

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 451.722  on 40  degrees of freedom
Residual deviance:  57.103  on 37  degrees of freedom
AIC: 203.2

Number of Fisher Scoring iterations: 4
```

Figure 7.7 R output of the `summary` method for the logistic regression model fitted to the `womensrole` data.

a Poisson error distribution, i.e.,

$$\mathsf{P}(y) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

This type of GLM is often known as *Poisson regression*. We can apply the model using

```
R> data("polyps", package = "HSAUR3")
R> polyps_glm_1 <- glm(number ~ treat + age, data = polyps,
+                      family = poisson())
```

(The default link function when the Poisson family is requested is the log function.)

We can deal with overdispersion by using a procedure known as *quasi-likelihood*, which allows the estimation of model parameters without fully knowing the error distribution of the response variable. McCullagh and Nelder (1989) give full details of the quasi-likelihood approach. In many respects it simply allows for the estimation of ϕ from the data rather than defining it to be unity for the binomial and Poisson distributions. We can apply quasi-likelihood estimation to the colonic polyps data using the following R code

```
R> polyps_glm_2 <- glm(number ~ treat + age, data = polyps,
+                      family = quasipoisson())
R> summary(polyps_glm_2)

Call:
glm(formula = number ~ treat + age, family = quasipoisson(),
    data = polyps)
```

```
R> role.fitted2 <- predict(womensrole_glm_2, type = "response")
R> myplot(role.fitted2)
```

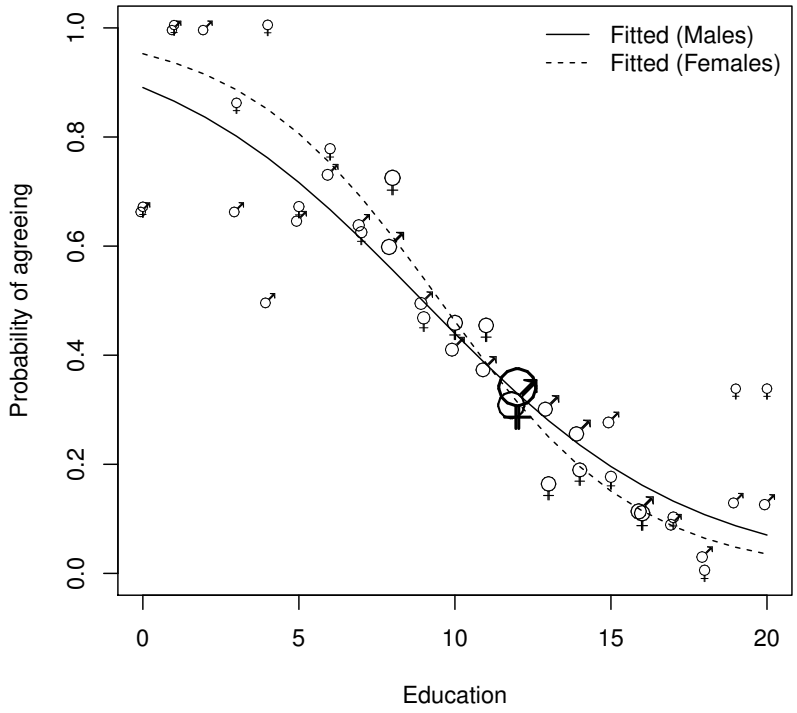

Figure 7.8 Fitted (from `womensrole_glm_2`) and observed probabilities of agreeing for the `womensrole` data.

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -4.22   -3.05   -0.18    1.45    5.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5290     0.4811    9.41  3.7e-08
treatdrug    -1.3591     0.3853   -3.53   0.0026
age          -0.0388     0.0195   -1.99   0.0628

(Dispersion parameter for quasipoisson family taken to be 10.7)
```

```
R> res <- residuals(womensrole_glm_2, type = "deviance")
R> plot(predict(womensrole_glm_2), res,
+       xlab="Fitted values", ylab = "Residuals",
+       ylim = max(abs(res)) * c(-1,1))
R> abline(h = 0, lty = 2)
```

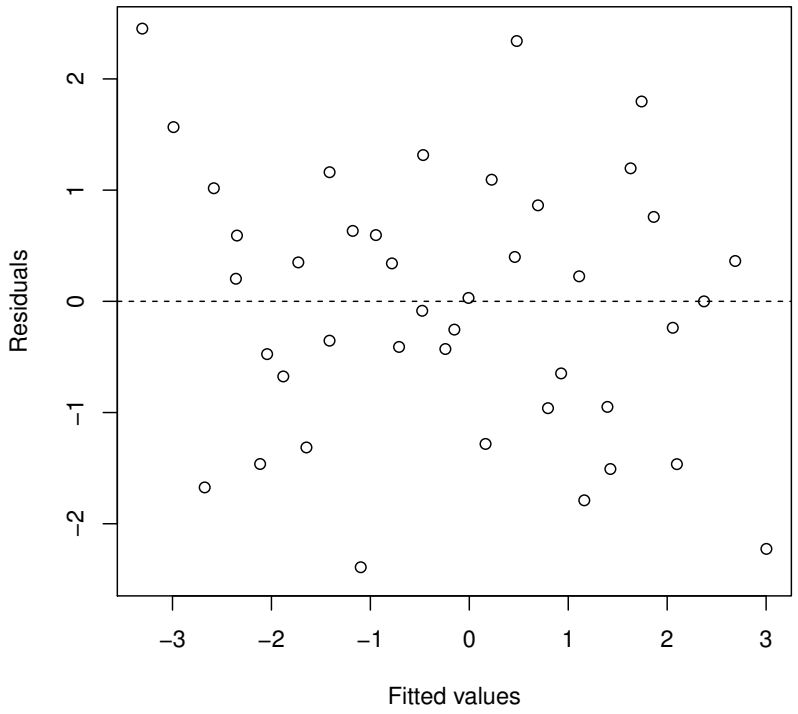

Figure 7.9 Plot of deviance residuals from logistic regression model fitted to the `womensrole` data.

```
    Null deviance: 378.66  on 19  degrees of freedom
Residual deviance: 179.54  on 17  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

The regression coefficients for both explanatory variables remain significant but their estimated standard errors are now much greater than the values given in Figure 7.10. A possible reason for overdispersion in these data is that

```
R> summary(polyps_glm_1)

Call:
glm(formula = number ~ treat + age, family = poisson(), data = polyps)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -4.22   -3.05   -0.18    1.45    5.83

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.52902    0.14687   30.84   <2e-16
treatdrug   -1.35908    0.11764  -11.55   <2e-16
age         -0.03883    0.00596   -6.52    7e-11

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 378.66  on 19  degrees of freedom
Residual deviance: 179.54  on 17  degrees of freedom
AIC: 273.9

Number of Fisher Scoring iterations: 5
```

Figure 7.10 R output of the `summary` method for the Poisson regression model fitted to the `polyps` data.

polyps do not occur independently of one another, but instead may 'cluster' together.

7.3.4 Driving and Back Pain

A frequently used design in medicine is the matched case-control study in which each patient suffering from a particular condition of interest included in the study is matched to one or more people without the condition. The most commonly used matching variables are age, ethnic group, mental status, etc. A design with m controls per case is known as a $1 : m$ matched study. In many cases m will be one, and it is the $1 : 1$ matched study that we shall concentrate on here where we analyze the data on low back pain given in Table **??**. To begin we shall describe the form of the logistic model appropriate for case-control studies in the simplest case where there is only one binary explanatory variable.

With matched pairs data the form of the logistic model involves the probability, φ, that in matched pair number i, for a given value of the explanatory variable the member of the pair is a case. Specifically the model is

$$\text{logit}(\varphi_i) = \alpha_i + \beta x.$$

The odds that a subject with $x = 1$ is a case equals $\exp(\beta)$ times the odds that a subject with $x = 0$ is a case.

The model generalizes to the situation where there are q explanatory variables as

$$\text{logit}(\varphi_i) = \alpha_i + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q.$$

Typically one x is an explanatory variable of real interest, such as past exposure to a risk factor, with the others being used as a form of statistical control in addition to the variables already controlled by virtue of using them to form matched pairs. This is the case in our back pain example where it is the effect of car driving on lower back pain that is of most interest.

The problem with the model above is that the number of parameters increases at the same rate as the sample size with the consequence that maximum likelihood estimation is no longer viable. We can overcome this problem if we regard the parameters α_i as of little interest and so are willing to forgo their estimation. If we do, we can then create a *conditional likelihood function* that will yield maximum likelihood estimators of the coefficients, $\beta_1, \ldots, \beta_q$, that are consistent and asymptotically normally distributed. The mathematics behind this are described in Collett (2003).

The model can be fitted using the `clogit` function from package **survival**; the results are shown in Figure 7.11.

```
R> library("survival")
R> backpain_glm <- clogit(I(status == "case") ~
+      driver + suburban + strata(ID), data = backpain)
```

The response has to be a logical (`TRUE` for cases) and the `strata` command specifies the matched pairs.

```
R> print(backpain_glm)

Call:
clogit(I(status == "case") ~ driver + suburban + strata(ID),
    data = backpain)

             coef exp(coef) se(coef)    z     p
driveryes   0.658     1.931    0.294 2.24 0.025
suburbanyes 0.255     1.291    0.226 1.13 0.258

Likelihood ratio test=9.55  on 2 df, p=0.008
n= 434, number of events= 217
```

Figure 7.11 R output of the `print` method for the conditional logistic regression model fitted to the `backpain` data.

The estimate of the odds ratio of a herniated disc occurring in a driver relative to a nondriver is 1.93 with a 95% confidence interval of $(1.09, 3.44)$. Conditional on residence we can say that the risk of a herniated disc occurring in a driver is about twice that of a nondriver. There is no evidence that where a person lives affects the risk of lower back pain.

7.3.5 Happiness in China

We model the probability distribution of reported happiness using a proportional odds model. In R, the function `polr` from the **MASS** package (Venables and Ripley, 2002, Ripley, 2014) implements such models, but in a slightly

different form as explained in Section **??**. The model we are going to fit reads

$$\log\left(\frac{\mathsf{P}(y \leq k|x_1, \ldots, x_q)}{\mathsf{P}(y > k|x_1, \ldots, x_q)}\right) = \zeta_k - (\beta_1 x_1 + \cdots + \beta_q x_q)$$

and we have to take care when interpreting the signs of the estimated regression coefficients. Another issue needs our attention before we start. Three of the explanatory variables are itself ordered (`R_edu`, the level of education of the responding woman; `R_health`, the health status of the responding woman in the last year; and `A_edu`, the level of education of the woman's partner). For unordered factors, the default treatment contrasts, see Chapters **??**, **??**, and **??**, compares the effect of each level to the first level. This coding does not take the ordinal nature of an ordered factor into account. One more appropriate coding is called *Helmert* contrasts. Here, we compare each level k to the average of the preceding levels, i.e., the second level to the first, the third to the average of the first and the second, and so on (these contrasts are also sometimes called *reverse Helmert contrasts*). The `option` function can be used to specify the default contrasts for unordered (we don't change the default `contr.treatment` option) and ordered factors. The returned `opts` variable stores the options before manipulation and can be used to conveniently restore them after we fitted the proportional odds model:

```
R> library("MASS")
R> opts <- options(contrasts = c("contr.treatment",
+                                "contr.helmert"))
R> CHFLS_polr <- polr(R_happy ~ ., data = CHFLS, Hess = TRUE)
R> options(opts)
```

As (almost) always, the `summary` function can be used to display the fitted model, see Figure 7.12. The largest absolute values of the t-statistics are associated with the self-reported health variable. To interpret the results correctly, we first make sure to understand the definition of the Helmert contrasts.

```
R> H <- with(CHFLS, contr.helmert(table(R_health)))
R> rownames(H) <- levels(CHFLS$R_health)
R> colnames(H) <- paste(levels(CHFLS$R_health)[-1], "- avg")
R> H
```

	Not good - avg	Fair - avg	Good - avg	Excellent - avg
Poor	-1	-1	-1	-1
Not good	1	-1	-1	-1
Fair	0	2	-1	-1
Good	0	0	3	-1
Excellent	0	0	0	4

Let's focus on the probability of being very unhappy. A positive regression coefficient for the first contrast of health means that the probability of being very unhappy is smaller (because of the sign switch in the regression coefficients) for women that reported their health as not good compared to women that reported a poor health. Thus, the results given in Figure 7.12 indicate

```
R> summary(CHFLS_polr)

Call:
polr(formula = R_happy ~ ., data = CHFLS, Hess = TRUE)

Coefficients:
                          Value Std. Error t value
R_regionCoastal East -1.70e-01   1.23e-01  -1.387
R_regionInlands      -4.63e-01   1.49e-01  -3.105
R_regionNorth        -2.10e-01   1.32e-01  -1.593
R_regionNortheast    -5.86e-01   1.27e-01  -4.602
R_regionCentral West -6.66e-01   1.35e-01  -4.919
R_age                 1.52e-02   6.26e-03   2.434
R_edu1                2.98e-02   1.44e-01   0.208
R_edu2               -6.42e-02   6.19e-02  -1.038
R_edu3                2.63e-02   4.12e-02   0.638
R_edu4                7.59e-03   4.85e-02   0.156
R_edu5               -1.44e-02   6.56e-02  -0.219
R_income              1.00e-04   1.08e-04   0.930
R_health1             5.75e-01   2.41e-02  23.870
R_health2             5.32e-01   6.19e-02   8.592
R_health3             4.27e-01   3.79e-02  11.258
R_health4             5.31e-01   3.23e-02  16.461
R_height              2.46e-02   9.96e-03   2.467
A_height             -9.43e-03   9.35e-03  -1.009
A_edu1               -4.53e-01   2.24e-01  -2.019
A_edu2               -8.67e-02   8.53e-02  -1.016
A_edu3               -3.80e-02   5.02e-02  -0.758
A_edu4               -1.65e-02   4.82e-02  -0.343
A_edu5               -1.63e-02   4.79e-02  -0.340
A_income              7.85e-05   7.44e-05   1.055

Intercepts:
                              Value    Std. Error t value
Very unhappy|Not too happy     -1.848    0.002   -770.786
Not too happy|Somewhat happy    1.079    0.263      4.098
Somewhat happy|Very happy       5.051    0.285     17.696

Residual Deviance: 2375.25
AIC: 2429.25
(3 observations deleted due to missingness)
```

Figure 7.12 R output of the `summary` method for the proportional odds model fitted to the `CHFLS` data.

that better health leads to happier women, a finding that sits well with our expectations. The other effects are less clear to interpret, also because formal inference is difficult and no p-values are displayed in the summary output of Figure 7.12. As a remedy, making use of the asymptotic distribution of maximum-likelihood-based estimators, we use the `cftest` function from the **multcomp** package (Hothorn et al., 2014) to compute normal p-values assuming that the estimated regression coefficients follow a normal limiting distribution (which is, for 1531 observations, not completely unrealistic); the results are given in Figure 7.13.

There seem to be geographical differences and also older and larger women seem to be happier. Other than that, education and income don't seem to contribute much in this model. One remarkable thing about the proportional odds model is that, similar to the quantile regression models presented in

```
R> library("multcomp")
R> cftest(CHFLS_polr)

        Simultaneous Tests for General Linear Hypotheses

Fit: polr(formula = R_happy ~ ., data = CHFLS, Hess = TRUE)

Linear Hypotheses:
                        Estimate Std. Error z value Pr(>|z|)
R_regionCoastal East -1.70e-01   1.23e-01   -1.39   0.1653
R_regionInlands      -4.63e-01   1.49e-01   -3.10   0.0019
R_regionNorth        -2.10e-01   1.32e-01   -1.59   0.1112
R_regionNortheast    -5.86e-01   1.27e-01   -4.60  4.2e-06
R_regionCentral West -6.66e-01   1.35e-01   -4.92  8.7e-07
R_age                 1.52e-02   6.26e-03    2.43   0.0149
R_edu1                2.98e-02   1.44e-01    0.21   0.8354
R_edu2               -6.42e-02   6.19e-02   -1.04   0.2993
R_edu3                2.63e-02   4.12e-02    0.64   0.5235
R_edu4                7.59e-03   4.85e-02    0.16   0.8757
R_edu5               -1.44e-02   6.56e-02   -0.22   0.8263
R_income              1.00e-04   1.08e-04    0.93   0.3523
R_health1             5.75e-01   2.41e-02   23.87  < 2e-16
R_health2             5.32e-01   6.19e-02    8.59  < 2e-16
R_health3             4.27e-01   3.79e-02   11.26  < 2e-16
R_health4             5.31e-01   3.23e-02   16.46  < 2e-16
R_height              2.46e-02   9.96e-03    2.47   0.0136
A_height             -9.43e-03   9.35e-03   -1.01   0.3132
A_edu1               -4.53e-01   2.24e-01   -2.02   0.0435
A_edu2               -8.67e-02   8.53e-02   -1.02   0.3096
A_edu3               -3.80e-02   5.02e-02   -0.76   0.4487
A_edu4               -1.65e-02   4.82e-02   -0.34   0.7318
A_edu5               -1.63e-02   4.79e-02   -0.34   0.7342
A_income              7.85e-05   7.44e-05    1.06   0.2913
(Univariate p values reported)
```

Figure 7.13 R output of the `cftest` function for the proportional odds model fitted to the `CHFLS` data.

Chapter **??**, it directly formulates a regression problem in terms of conditional distributions, not only conditional means (the same is trivially true for the binary case in logistic regression). Consequently, the model allows making distributional predictions, in other words, we can infer the predicted distribution or density of happiness in a woman with certain values for the explanatory variables that entered the model. To do so, we focus on the woman corresponding to the first row of the data set:

```
R> CHFLS[1,]
```

```
   R_region R_age              R_edu R_income R_health
2 Northeast    54 Senior high school      900     Good
  R_height        R_happy A_height              A_edu A_income
2      165 Somewhat happy      172 Senior high school      500
```

and repeat these values as often as there are levels in the `R_health` factor, and each row is assigned one of these levels

```
R> nd <- CHFLS[rep(1, nlevels(CHFLS$R_health)),]
R> nd$R_health <- ordered(levels(nd$R_health),
+                         labels = levels(nd$R_health))
```

We can now use the `predict` function to compute the density of the response variable `R_happy` for each of these five hypothetical women:

```
R> (dens <- predict(CHFLS_polr, newdata = nd, type = "probs"))
```

```
    Very unhappy Not too happy Somewhat happy Very happy
2         0.00114        0.0197          0.510     0.4696
2.1       0.00449        0.0732          0.740     0.1828
2.2       0.02330        0.2847          0.651     0.0406
2.3       0.00851        0.1295          0.757     0.1052
2.4       0.07003        0.5142          0.403     0.0132
```

From each row, we get the predicted probability that the self-reported happiness will correspond to the levels shown in the column name. These densities, one for each row in `nd` and therefore for each level of health, can now be plotted, for example using a conditional barchart, see Figure 7.14. We clearly see that better health is associated with greater happiness.

We'll present an alternative and maybe simpler model in Chapter **??**.

7.4 Summary of Findings

Blood screening Application of logistic regression shows that an increase of one unit in the fibrinogen value produces approximately a six fold increase in the odds of an ESR value greater than 20. However, because the number of observations is small the corresponding 95% confidence interval for the odds is rather wide namely, $(1.4, 54.52)$. Gamma globulin values do not help in the prediction of ESR values greater than 20 over and above the fibrinogen values.

Women's role in society Modeling the probability of agreeing with the statement about women's role in society using logistic regression demonstrates that it is the interaction of education and gender which is of most importance; for fewer years of education women have a higher probability of agreeing with the statement than men, but when the years of education exceed about ten then this situation reverses.

Colonic polyps Fitting a Poisson regression allowing for overdispersion shows that the drug treatment is effective in reducing the number of polyps with age having only a marginal effect.

```
R> library("lattice")
R> D <- expand.grid(R_health = nd$R_health,
+                   R_happy = ordered(LETTERS[1:4]))
R> D$dens <- as.vector(dens)
R> barchart(dens ~ R_happy | R_health, data = D,
+           ylab = "Density", xlab = "Happiness",)
```

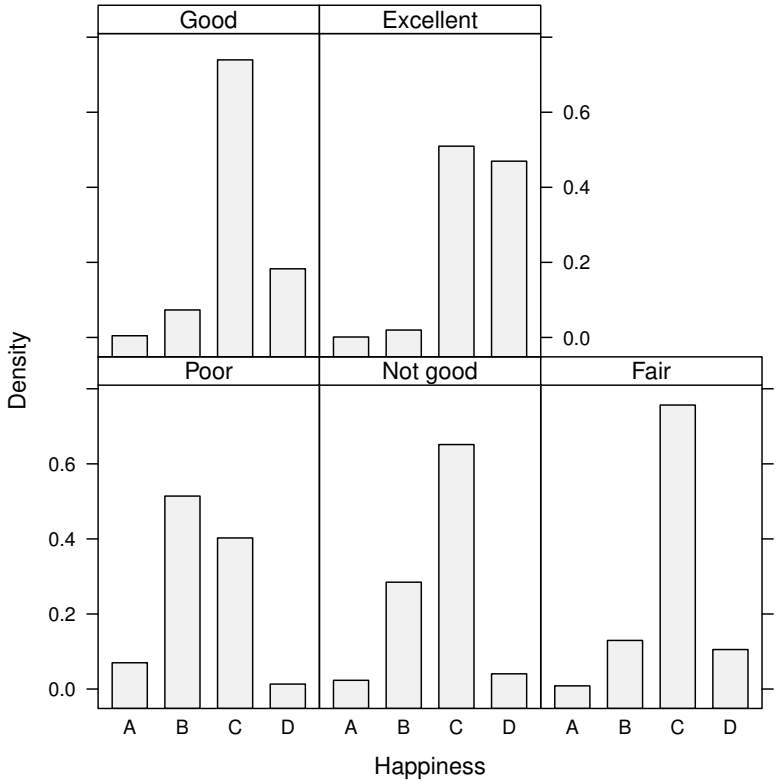

Figure 7.14 Predicted distribution of happiness for hypothetical women with health conditions rating from poor to excellent, with the remaining explanatory variables being the same as for the woman corresponding to the first row in the `CHFLS` data frame. The levels of happiness have been abbreviated (A: very unhappy, B: not too happy, C: somewhat happy; D: very happy).

Driving and back pain Application of conditional logistic regression shows that the odds ratio of a herniated disc occurring in a driver relative to a nondriver is 1.93 with a 95% confidence interval of $(1.09, 3.44)$. There is no evidence that where a person lives affects the risk of suffering lower back pain.

Happiness in China Better health is associated with greater happiness – what a surprise!

7.5 Final Comments

Generalized linear models provide a very powerful and flexible framework for the application of regression models to a variety of non-normal response variables, for example, logistic regression to binary responses and Poisson regression to count data.

Bibliography

Collett, D. (2003), *Modelling Binary Data*, London, UK: Chapman & Hall/CRC, 2nd edition.

Hothorn, T., Bretz, F., and Westfall, P. (2014), ***multcomp**: Simultaneous Inference for General Linear Hypotheses*, URL `http://CRAN.R-project.org/package=multcomp`, R package version 1.3-2.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London, UK: Chapman & Hall/CRC.

Ripley, B. D. (2014), ***MASS**: Support Functions and Datasets for Venables and Ripley's MASS*, URL `http://CRAN.R-project.org/package=MASS`, R package version 7.3-29.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York, USA: Springer-Verlag, 4th edition, URL `http://www.stats.ox.ac.uk/pub/MASS4/`, ISBN 0-387-95457-0.