A Handbook of Statistical Analyses Using R — 2nd Edition

Brian S. Everitt and Torsten Hothorn

CHAPTER 7

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

7.1 Introduction

7.2 Logistic Regression and Generalised 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 favour 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> exp(confint(plasma_glm_1, parm = "fibrinogen"))
```

```
R> data("plasma", package = "HSAUR2")
R> layout(matrix(1:2, ncol = 2))
R> cdplot(ESR ~ fibrinogen, data = plasma)
R> cdplot(ESR ~ globulin, data = plasma)
```

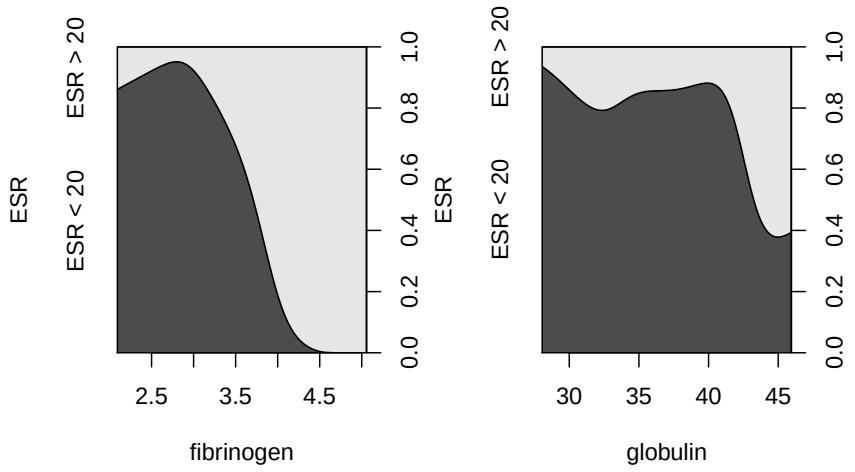

Figure 7.1 Conditional density plots of the erythrocyte sedimentation rate (ESR) given fibrinogen and globulin.

```
 2.5 % 97.5 %
  1.40  54.52
```

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

```
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> 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 P(>|Chi|)
1        30      24.84
2        29      22.97  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.

```
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.

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 = "HSAUR2")
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

```
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)
```

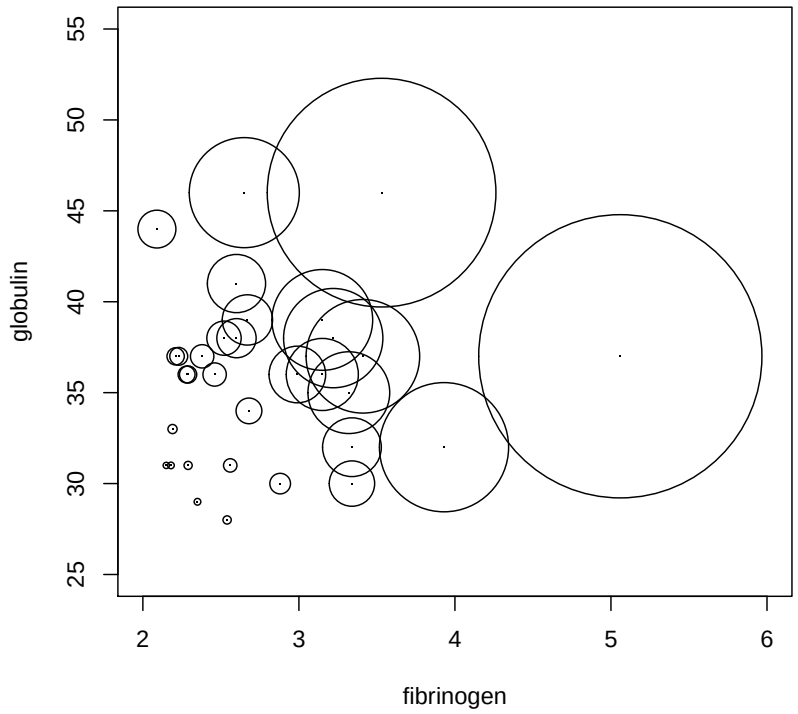

Figure 7.4 Bubbleplot of fitted values for a logistic regression model fitted to the `plasma` data.

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> 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      0.9
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.

```
6  +      lines(womensrole$education[!f], role.fitted[!f], lty = 1)
7  +      lines(womensrole$education[f], role.fitted[f], lty = 2)
8  +      lgtxt <- c("Fitted (Males)", "Fitted (Females)")
9  +      legend("topright", lgtxt, lty = 1:2, bty = "n")
10 +      y <-  womensrole$agree / (womensrole$agree +
11 +                                womensrole$disagree)
12 +      text(womensrole$education, y, ifelse(f, "\\VE", "\\MA"),
13 +           family = "HersheySerif", cex = 1.25)
14 +  }
```

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 and 13 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 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

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

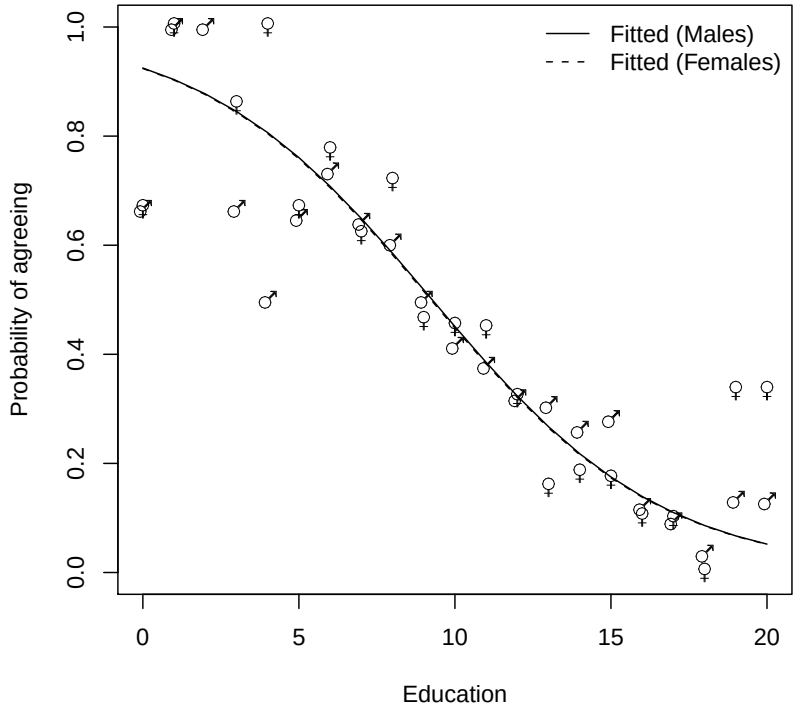

Figure 7.6 Fitted (from `womensrole_glm_1`) and observed probabilities of agreeing for the `womensrole` data.

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

```
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.

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 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 = "HSAUR2")
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

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

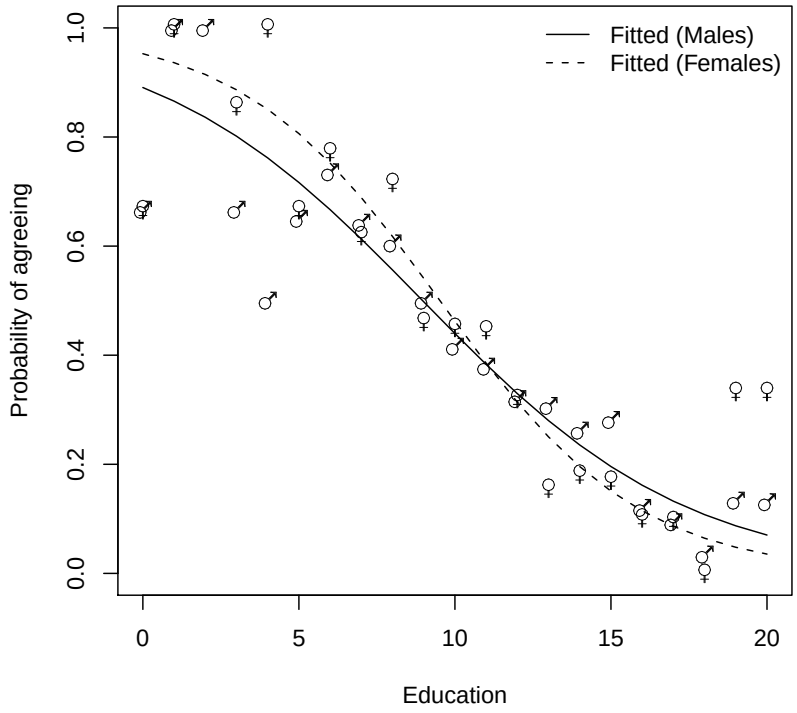

Figure 7.8 Fitted (from `womensrole_glm_2`) and observed probabilities of agreeing for the `womensrole` data.

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> 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)
```

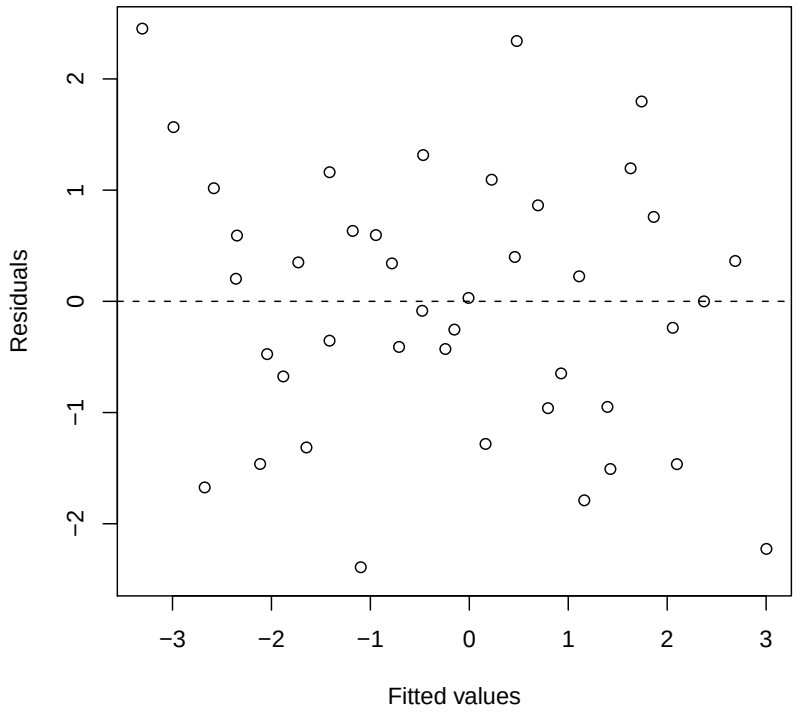

Figure 7.9 Plot of deviance residuals from logistic regression model fitted to the `womensrole` data.

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.221  -3.054  -0.180   1.446   5.830

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
```

```
R> summary(polyps_glm_1)

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

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.221  -3.054  -0.180   1.446   5.830

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.

```
age          -0.0388     0.0195   -1.99   0.0628

(Dispersion parameter for quasipoisson family taken to be 10.7)

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

Bibliography

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