Multivariate Analysis for the Behavioral Sciences, Second Edition (Chapman and Hall/CRC, 2019)
Examples of Chapter 6: Applying Logistic Regression

Kimmo Vehkalahti and Brian S. Everitt

10 November 2018

Contents

Examples

Table 6.1: Psychiatric Caseness Data

```
GHQ <- c(0:10, 0:10)
sex <- c(rep(0,11), rep(1,11))
ncases <- c(4, 4, 8, 6, 4, 6, 3, 2, 3, 2, 1, 1, 2, 2, 1, 3, 3, 2, 4, 3, 2, 2)
nnotcases <- c(80, 29, 15, 3, 2, 1, 1, 0, 0, 0, 0, 36, 25, 8, 4, 1, 1, 1, 2, 1, 0, 0)
cbind(sex, GHQ, ncases, nnotcases)
```

```
##       sex GHQ ncases nnotcases
##  [1,]   0   0      4        80
##  [2,]   0   1      4        29
##  [3,]   0   2      8        15
##  [4,]   0   3      6         3
##  [5,]   0   4      4         2
##  [6,]   0   5      6         1
##  [7,]   0   6      3         1
##  [8,]   0   7      2         0
##  [9,]   0   8      3         0
## [10,]   0   9      2         0
## [11,]   0  10      1         0
## [12,]   1   0      1        36
## [13,]   1   1      2        25
## [14,]   1   2      2         8
## [15,]   1   3      1         4
## [16,]   1   4      3         1
## [17,]   1   5      3         1
## [18,]   1   6      2         1
## [19,]   1   7      4         2
## [20,]   1   8      3         1
## [21,]   1   9      2         0
## [22,]   1  10      2         0
```

```
sex <- factor(sex, levels = c(0, 1), labels = c("F", "M"))
```

Tables 6.2 and 6.3, Figure 6.1

```r
GHQ_reg <- glm(cbind(ncases,nnotcases) ~ sex, family = binomial)
summary(GHQ_reg)
```

```
##
## Call:
## glm(formula = cbind(ncases, nnotcases) ~ sex, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9434   0.1076   2.1458   2.3646   3.4059
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.11400    0.17575  -6.338 2.32e-10 ***
## sexM        -0.03657    0.28905  -0.127    0.899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.31  on 21  degrees of freedom
## Residual deviance: 130.29  on 20  degrees of freedom
## AIC: 170.26
##
## Number of Fisher Scoring iterations: 5
```

```r
predict(GHQ_reg, type = "response")
```

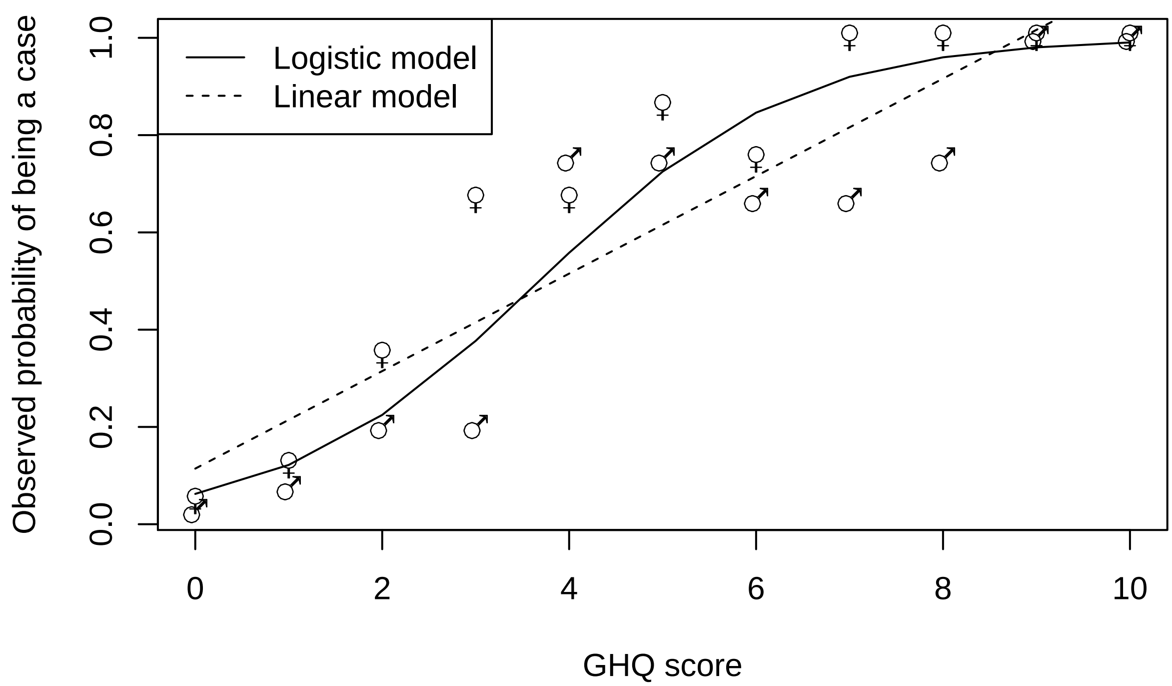
```
##         1         2         3         4         5         6         7
## 0.2471264 0.2471264 0.2471264 0.2471264 0.2471264 0.2471264 0.2471264
##         8         9        10        11        12        13        14
## 0.2471264 0.2471264 0.2471264 0.2471264 0.2403846 0.2403846 0.2403846
##        15        16        17        18        19        20        21
## 0.2403846 0.2403846 0.2403846 0.2403846 0.2403846 0.2403846 0.2403846
##        22
## 0.2403846
```

```r
GHQ_reg1 <- glm(cbind(ncases,nnotcases) ~ GHQ, family = binomial)
fitted <- predict(GHQ_reg1, type = "response")
pobsv <- ncases / (ncases + nnotcases)
plot(GHQ, pobsv, type = "n", xlab = "GHQ score", ylab = "Observed probability of being a case")
text(GHQ, pobsv, ifelse(sex == "F", "\\VE", "\\MA"), vfont = c("serif", "plain"), cex = 1.25)
lines(0:10, fitted[1:11])
GHQ_lin <- lm(pobsv ~ GHQ)
summary(GHQ_lin)
```

```
##
## Call:
## lm(formula = pobsv ~ GHQ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21505 -0.11624 -0.03279  0.12180  0.25161
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.11434    0.05923   1.931   0.0678 .
## GHQ          0.10024    0.01001  10.012  3.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1485 on 20 degrees of freedom
## Multiple R-squared:  0.8337, Adjusted R-squared:  0.8254
## F-statistic: 100.2 on 1 and 20 DF,  p-value: 3.099e-09
```

```r
fitted <- predict(GHQ_lin)
lines(0:10, fitted[1:11], lty = 2)
legend("topleft", c("Logistic model", "Linear model"), lty = 1:2)
```

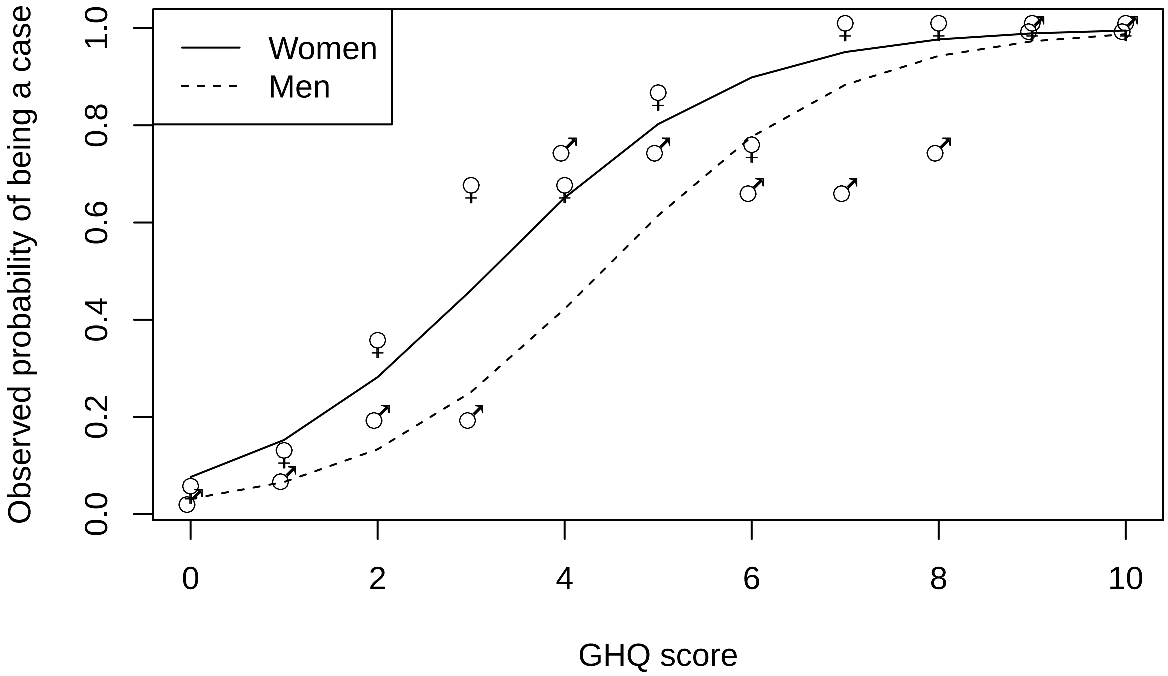


Tables 6.4 and 6.5, Figures 6.2 and 6.3

```r
GHQ_reg2 <- glm(cbind(ncases,nnotcases) ~ sex + GHQ, family = binomial)
summary(GHQ_reg2)
```

```
##
## Call:
## glm(formula = cbind(ncases, nnotcases) ~ sex + GHQ, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3955  -0.3939   0.1876   0.4315   1.3306
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.49351    0.28164  -8.854  < 2e-16 ***
## sexM        -0.93609    0.43435  -2.155   0.0311 *
## GHQ          0.77910    0.09903   7.867 3.63e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.306  on 21  degrees of freedom
## Residual deviance:  11.113  on 19  degrees of freedom
## AIC: 53.087
##
## Number of Fisher Scoring iterations: 5
```

```r
fitted <- predict(GHQ_reg2, type = "response")
pobsv <- ncases / (ncases + nnotcases)
plot(GHQ, pobsv, type = "n", xlab = "GHQ score", ylab = "Observed probability of being a case")
text(GHQ, pobsv, ifelse(sex == "F", "\\VE", "\\MA"), vfont = c("serif", "plain"), cex = 1.25)
lines(0:10, fitted[1:11])
lines(0:10, fitted[12:22], lty = 2)
legend("topleft", c("Women", "Men"), lty = 1:2)
```




```r
#interaction model
GHQ_reg3 <- glm(cbind(ncases,nnotcases) ~ sex * GHQ, family = binomial)
summary(GHQ_reg3)
```

```
##
## Call:
## glm(formula = cbind(ncases, nnotcases) ~ sex * GHQ, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.29971  -0.32521  -0.03273   0.39672   1.45689
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7732     0.3586  -7.732 1.06e-14 ***
## sexM         -0.2253     0.6093  -0.370    0.712
## GHQ           0.9412     0.1569   6.000 1.97e-09 ***
## sexM:GHQ     -0.3020     0.1990  -1.517    0.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.3059  on 21  degrees of freedom
## Residual deviance:   8.7669  on 18  degrees of freedom
## AIC: 52.741
```

```
##
## Number of Fisher Scoring iterations: 5
```

```
fitted <- predict(GHQ_reg3, type = "response")
pobsv <- ncases / (ncases + nnotcases)
plot(GHQ, pobsv, type = "n", xlab = "GHQ score", ylab = "Observed probability of being a case")
text(GHQ, pobsv, ifelse(sex == "F", "\\VE", "\\MA"), vfont = c("serif", "plain"), cex = 1.25)
lines(0:10, fitted[1:11])
lines(0:10, fitted[12:22], lty = 2)
legend("topleft", c("Women", "Men"), lty = 1:2)
```

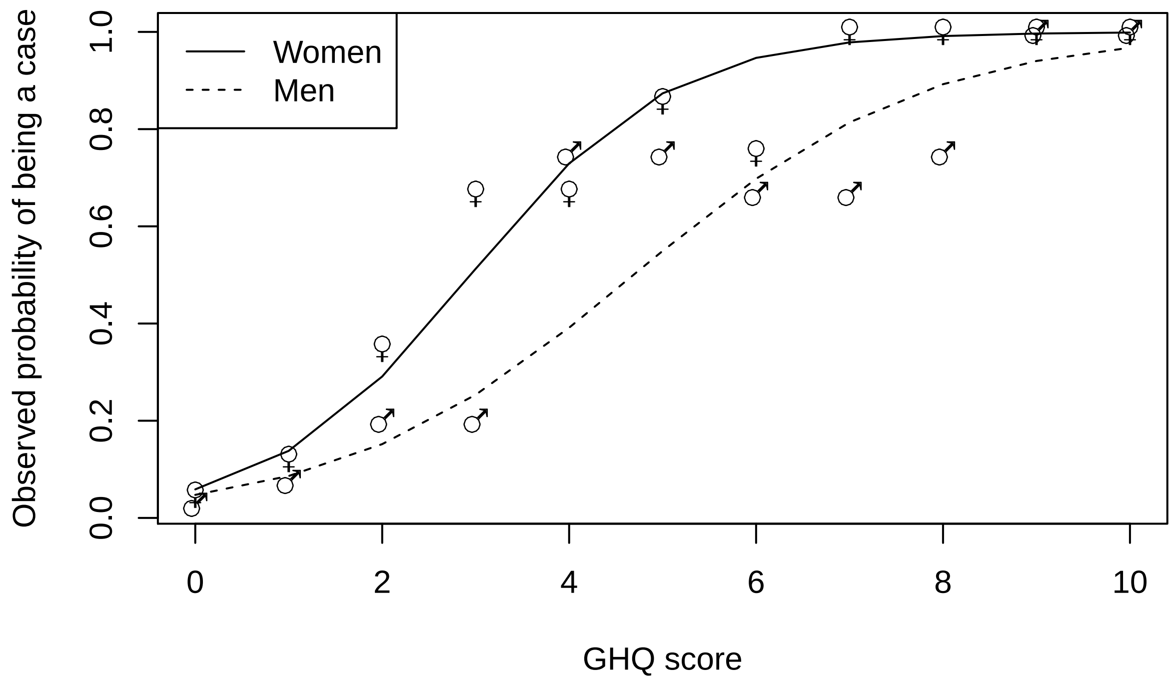


Table 6.6: Do-It-Yourself Data

```r
work <- rep(c(1, 2, 3), c(12, 12, 12))
tenure <- rep(c(rep(1, 6), rep(2, 6)), 3)
type <- rep(c(rep(1, 3), rep(2, 3)), 6)
age <- rep(c(1, 2, 3), 12)

yes <- c(18, 15, 6, 34, 10, 2, 15, 13, 9, 28, 4, 6, 5, 3, 1, 56, 56, 35, 1, 1, 1, 12, 21,
         8, 17, 10, 15, 29, 3, 7, 34, 17, 19, 44, 13, 16, 2, 0, 3, 23, 52, 49, 3, 2, 0,
         9, 31, 51, 30, 23, 21, 22, 13, 11, 25, 19, 40, 25, 16, 12, 8, 5, 1, 54, 191,
         102, 4, 2, 2, 19, 76, 61)
no <- yes[c(7:12, 19:24, 31:36, 43:48, 55:60, 67:72)]
yes <- yes[c(1:6, 13:18, 25:30, 37:42, 49:54, 61:66)]

work <- factor(work, levels = c(1, 2, 3), labels = c("skilled", "unskilled", "office"))
tenure <- factor(tenure, levels = c(1, 2), labels = c("rent", "own"))
type <- factor(type, levels = c(1, 2), labels = c("apartment", "house"))
age <- factor(age, levels = c(1, 2, 3), labels = c("<30", "31-45", "46+"))

data.frame(work, tenure, type, age, yes, no)
```

```
##         work tenure      type   age yes no
## 1    skilled   rent apartment   <30  18 15
## 2    skilled   rent apartment 31-45  15 13
## 3    skilled   rent apartment   46+   6  9
## 4    skilled   rent     house   <30  34 28
## 5    skilled   rent     house 31-45  10  4
## 6    skilled   rent     house   46+   2  6
## 7    skilled    own apartment   <30   5  1
## 8    skilled    own apartment 31-45   3  1
## 9    skilled    own apartment   46+   1  1
## 10   skilled    own     house   <30  56 12
## 11   skilled    own     house 31-45  56 21
## 12   skilled    own     house   46+  35  8
## 13 unskilled   rent apartment   <30  17 34
## 14 unskilled   rent apartment 31-45  10 17
## 15 unskilled   rent apartment   46+  15 19
## 16 unskilled   rent     house   <30  29 44
## 17 unskilled   rent     house 31-45   3 13
## 18 unskilled   rent     house   46+   7 16
## 19 unskilled    own apartment   <30   2  3
## 20 unskilled    own apartment 31-45   0  2
## 21 unskilled    own apartment   46+   3  0
## 22 unskilled    own     house   <30  23  9
## 23 unskilled    own     house 31-45  52 31
## 24 unskilled    own     house   46+  49 51
## 25    office   rent apartment   <30  30 25
## 26    office   rent apartment 31-45  23 19
## 27    office   rent apartment   46+  21 40
## 28    office   rent     house   <30  22 25
## 29    office   rent     house 31-45  13 16
## 30    office   rent     house   46+  11 12
## 31    office    own apartment   <30   8  4
```

```
## 32    office    own apartment 31-45   5  2
## 33    office    own apartment   46+   1  2
## 34    office    own     house   <30  54 19
## 35    office    own     house 31-45 191 76
## 36    office    own     house   46+ 102 61
```

Table 6.7

```r
# R will create the dummy variables automatically when using factor variables:
reg <- glm(cbind(yes,no) ~ work + type + tenure + age, family = "binomial")
summary(reg)
```

```
##
## Call:
## glm(formula = cbind(yes, no) ~ work + type + tenure + age, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9399  -0.6574  -0.1131   0.4123   1.9501
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.30606    0.15428   1.984   0.0473 *
## workunskilled -0.76267    0.15197  -5.018 5.21e-07 ***
## workoffice    -0.30535    0.14088  -2.167   0.0302 *
## typehouse     -0.00249    0.14717  -0.017   0.9865
## tenureown      1.01570    0.13787   7.367 1.74e-13 ***
## age31-45      -0.11304    0.13697  -0.825   0.4092
## age46+        -0.43661    0.14059  -3.106   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 158.884  on 35  degrees of freedom
## Residual deviance:  29.671  on 29  degrees of freedom
## AIC: 167.87
##
## Number of Fisher Scoring iterations: 4
```

Table 6.8

```r
reg <- glm(cbind(yes,no) ~ work + tenure + type + age, family = binomial)
step(reg, direction = "backward")
```

```
## Start:  AIC=167.87
## cbind(yes, no) ~ work + tenure + type + age
##
##          Df Deviance    AIC
## - type    1   29.671 165.87
## <none>        29.671 167.87
## - age     2   40.559 174.76
## - work    2   56.971 191.17
## - tenure  1   85.599 221.80
##
## Step:  AIC=165.87
## cbind(yes, no) ~ work + tenure + age
##
##          Df Deviance    AIC
## <none>        29.671 165.87
## - age     2   40.613 172.81
## - work    2   56.985 189.19
## - tenure  1  110.781 244.98
```

```
##
## Call:  glm(formula = cbind(yes, no) ~ work + tenure + age, family = binomial)
##
## Coefficients:
##   (Intercept)  workunskilled     workoffice      tenureown       age31-45
##        0.3048        -0.7627        -0.3053         1.0144        -0.1129
##        age46+
##       -0.4364
##
## Degrees of Freedom: 35 Total (i.e. Null);  30 Residual
## Null Deviance:       158.9
## Residual Deviance: 29.67     AIC: 165.9
```

Tables 6.10 and 6.11: Low Back Pain Data

```r
library(HSAUR3)
```

```
## Loading required package: tools
##
## Attaching package: 'HSAUR3'
## The following object is masked _by_ '.GlobalEnv':
##
##     GHQ
```

```r
data(backpain)
str(backpain)
```

```
## 'data.frame':    434 obs. of  4 variables:
##  $ ID      : Factor w/ 217 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
##  $ status  : Factor w/ 2 levels "case","control": 1 2 1 2 1 2 1 2 1 2 ...
##  $ driver  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 1 1 2 2 ...
##  $ suburban: Factor w/ 2 levels "no","yes": 2 1 2 2 1 2 1 1 1 2 ...
```

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

```
## Call:
## coxph(formula = Surv(rep(1, 434L), I(status == "case")) ~ driver +
##     suburban + strata(ID), data = backpain, method = "exact")
##
##   n= 434, number of events= 217
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## driveryes    0.6579    1.9307   0.2940 2.238   0.0252 *
## suburbanyes 0.2555    1.2911   0.2258 1.131   0.2580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## driveryes       1.931     0.5180    1.0851     3.435
## suburbanyes     1.291     0.7746    0.8293     2.010
##
## Rsquare= 0.022   (max possible= 0.5 )
## Likelihood ratio test= 9.55  on 2 df,   p=0.008457
## Wald test            = 8.85  on 2 df,   p=0.01195
## Score (logrank) test = 9.31  on 2 df,   p=0.0095
```