Chapter 7 in Everitt and Hothorn (2010) Logistic Regression and Generalized Linear Models

Start R.

If you were not able to edit Rprofile.site, load the HSAUR2 and Rcmdr either using the commands: library(HSAUR2);library(Rcmdr)
or from the R Console using the menu Packages > Load package ... > select HSAUR2 and Rcmdr > Ok

We will be working with the R Commander menus.

ESR and Plasma Proteins

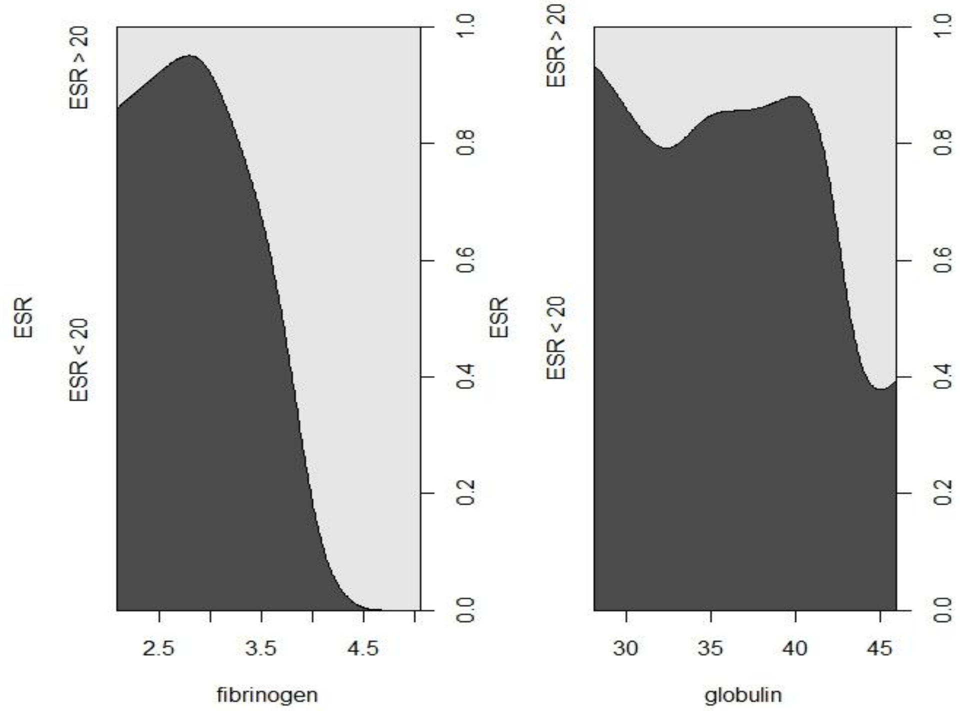
From the R Commander menus select Data > Data in packages > Read data set from an attached package... > double click on HSAUR2, select plasma, and click ok.

To see a description, from the R commander menu select Data > Active data set > Help on active data set (if available)

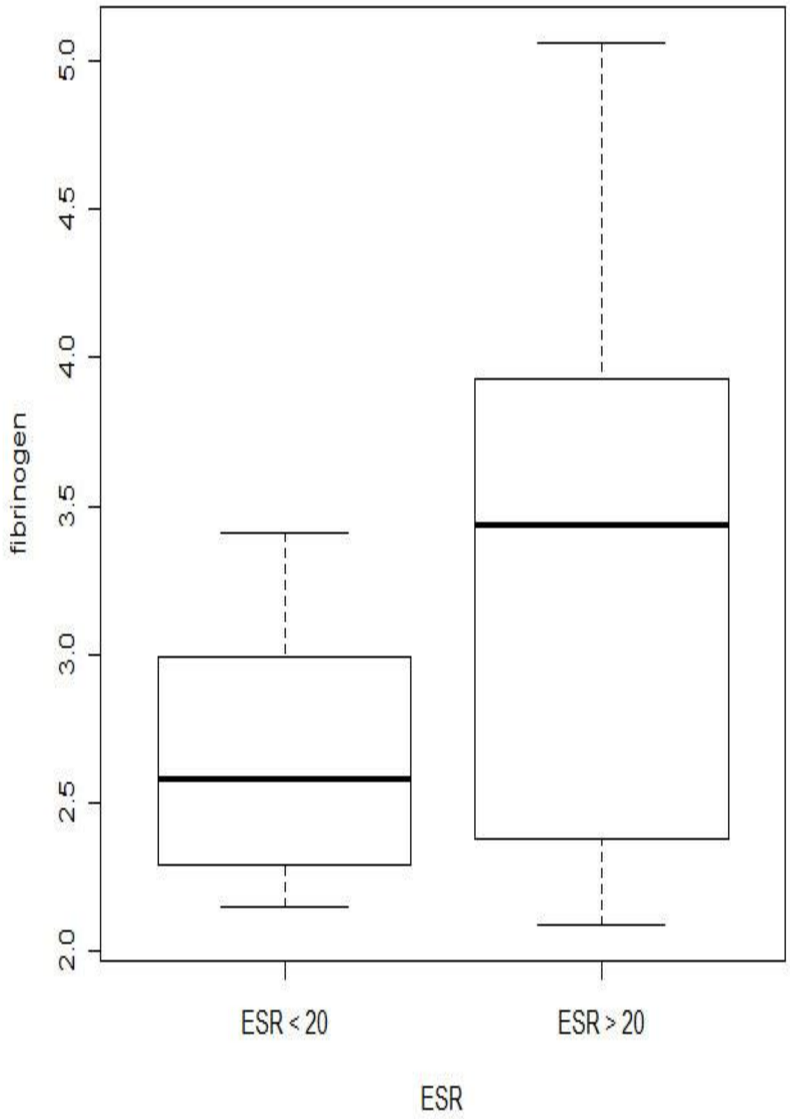
Click View data set to view it.

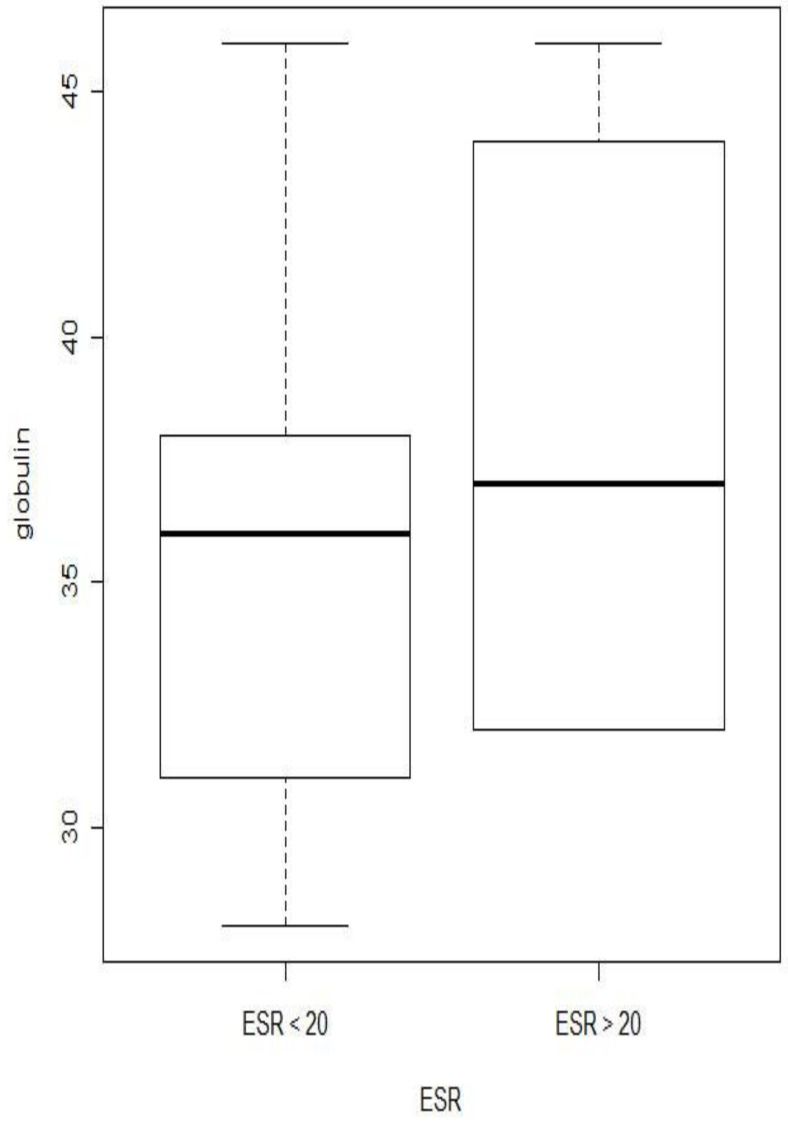
Let’s first look at conditional density plots of ESR (the response variable) given the two explanatory variables. These plots describe how the conditional distribution of the categorical variable ESR changes as the numerical variables fibrinogen and gamma globulin change. Looking at the plots, it appears that higher levels of each protein are associated with ESR values above 20 mm/hr:

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

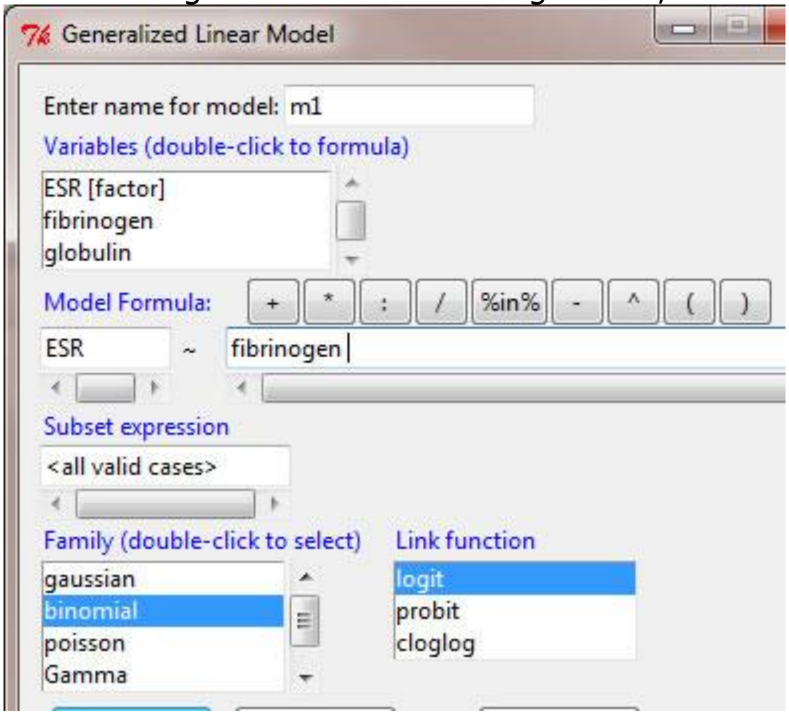


Then, from the menu, select Graphs > Boxplot... > select fibrinogen and click Plot by groups... > select ESR and click Ok >click Ok again.
In the R Graphics window, click History > Click recording, so you will be able to use the Page Up/Page Down keys to see other graphs.
Repeat to plot globulin.


fibrinogen
2.0
2.5
3.0
3.5
4.0
4.5
5.0
ESR < 20
ESR > 20
ESR



From the menu, select Statistics > Fit models > Generalized linear model... > double click ESR to add it to the left side of the model equation, double click fibrinogen to add it to the right side, and click Ok.



Enter name for model: m1

Model Formula: ESR ~ fibrinogen

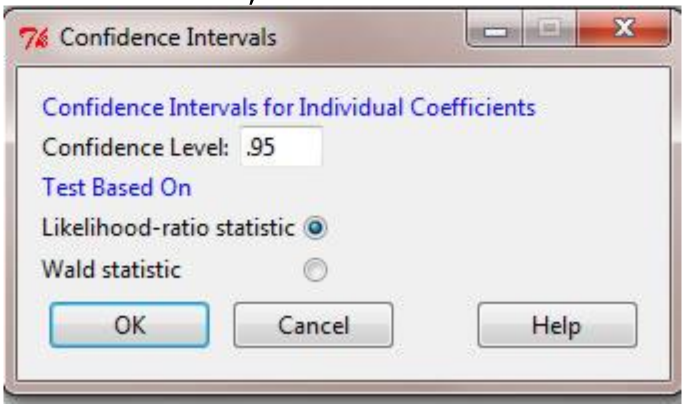
Subset expression: <all valid cases>

Family (double-click to select): binomial

Link function: logit

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9298  -0.5399  -0.4382  -0.3356   2.4794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8451     2.7703  -2.471   0.0135 *
fibrinogen    1.8271     0.9009   2.028   0.0425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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: 28.840

Number of Fisher Scoring iterations: 5
```

From the menu, select Models > Confidence intervals... > click Ok.



Confidence Intervals

Confidence Intervals for Individual Coefficients

Confidence Level: .95

Test Based On

Likelihood-ratio statistic ●

Wald statistic ○

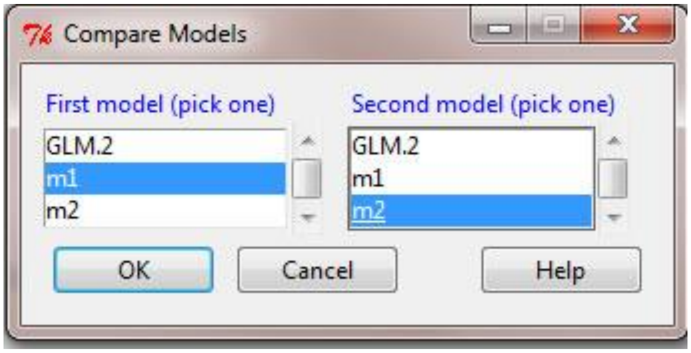
```
             Estimate       2.5 %    97.5 % exp(Estimate)        2.5 %
97.5 %
(Intercept) -6.845075 -13.6565434 -2.327294   0.001064686 1.172299e-06
0.09755943
fibrinogen   1.827081   0.3387619  3.998492   6.215715449 1.403209e+00
54.51588384
```

The 95% confidence interval (from the 2.5 to 97.5 percentiles) for the odds ratio is given on the log scale and then the estimate and the confidence interval is back transformed to the original scale.

From the menu, select Statistics > Fit models > Generalized linear model... > double click + and then globulin to add globulin to the model, and click Ok.

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.9683  -0.6122  -0.3458  -0.2116   2.2636  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -12.7921     5.7963  -2.207   0.0273 *
fibrinogen    1.9104     0.9710   1.967   0.0491 *
globulin      0.1558     0.1195   1.303   0.1925  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

(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: 28.971

Number of Fisher Scoring iterations: 5
```

From the menu, select Models > Hypothesis tests > Compare models... > select m1 and m2 and click Ok.



```
Analysis of Deviance Table

Model 1: ESR ~ fibrinogen
Model 2: ESR ~ fibrinogen + globulin
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        30     24.840
2        29     22.971  1   1.8692    0.1716
```

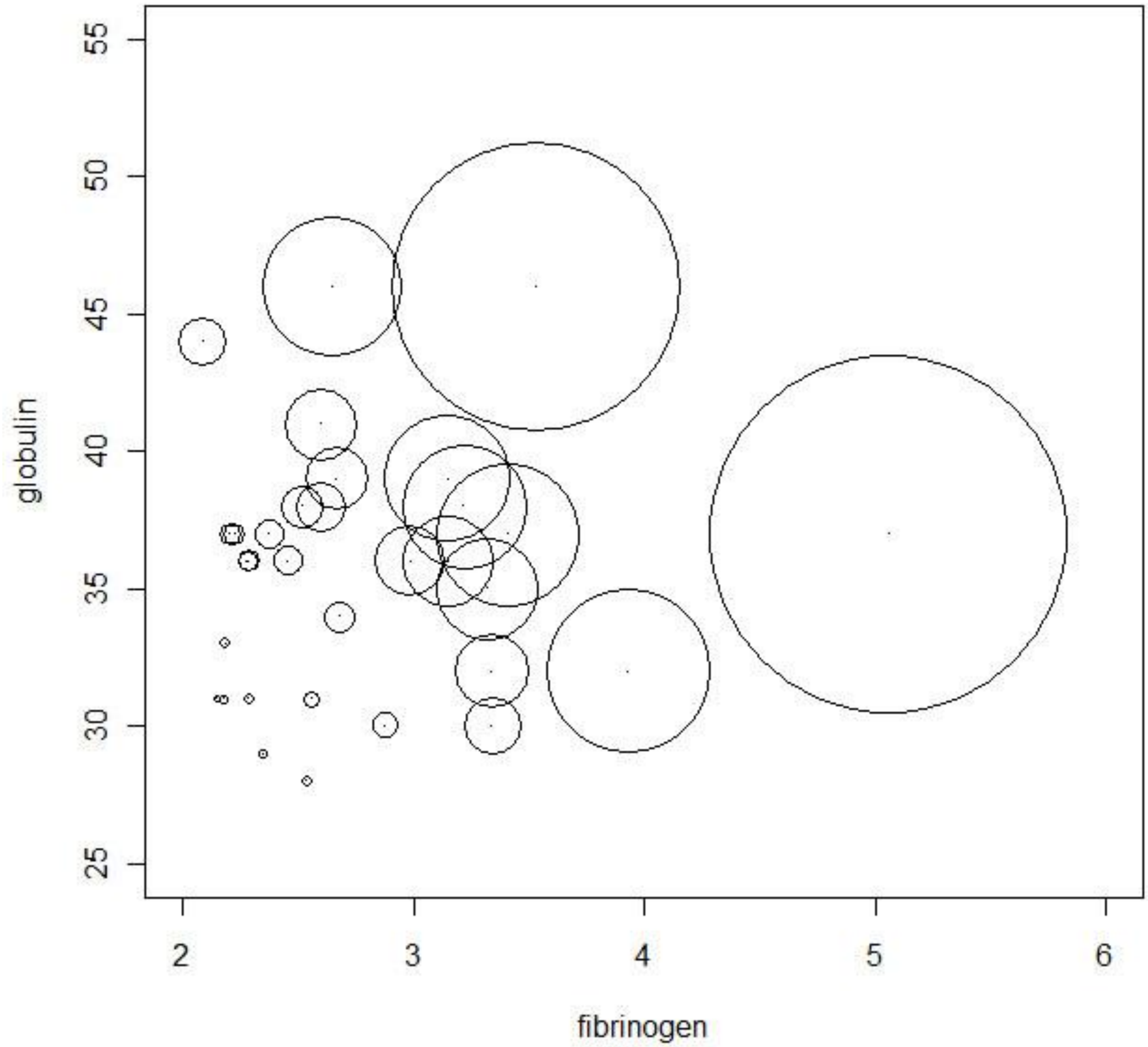
We prefer the simpler model, because the models are not significantly different.

To generate a bubble plot, enter the following commands into the Script Window and submit them.

prob = predict(m2, type="response")

plot(globulin ~ fibrinogen, data=plasma, xlim=c(2,6), ylim=c(25,55), pch=".")

symbols(plasma$fibrinogen, plasma$globulin, circles=prob, add=TRUE)



Women's Role in Society

From the R Commander menus select Data > Data in packages > Read data set from an attached package... >
Double click on HSAUR2 and select womensrole, then click OK.

To see a description, from the R commander menu select Data > Active data set > Help on active data set (if available)

Click View data set to view it.

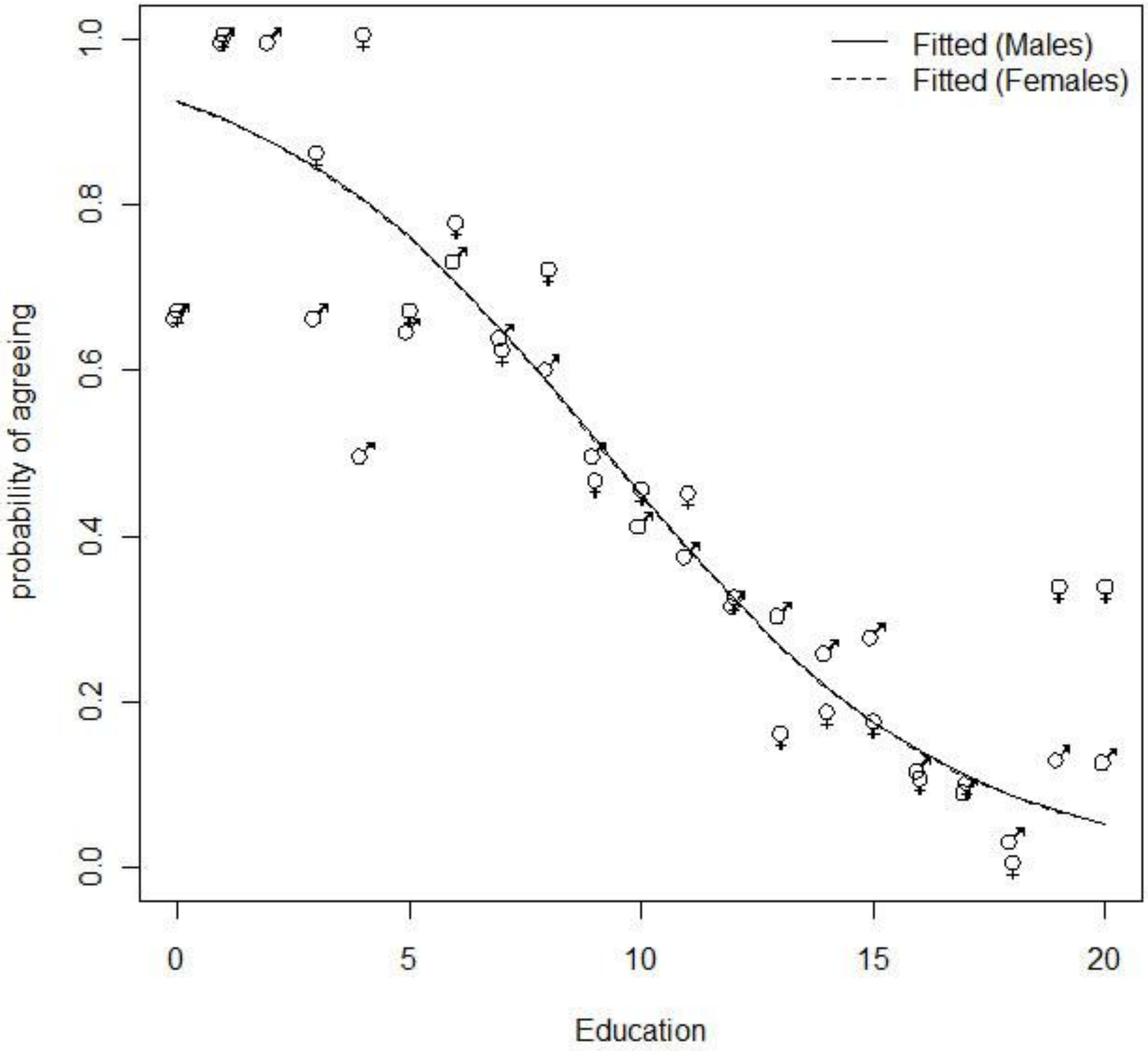
From the menu, select Statistics > Fit models > Generalized linear model... > name the model m1, on the left side of the model equation enter cbind(agree ,disagree) [you cannot use clicks), double click on gender, +, and education to add them to the right side, and click Ok.

```
Call:
glm(formula = cbind(agree, disagree) ~ gender + education, family =
binomial(logit),
    data = womensrole)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.72544  -0.86302  -0.06525   0.84340   3.13315

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.50937    0.18389  13.646   <2e-16 ***
gender[T.Female] -0.01145    0.08415  -0.136    0.892
education        -0.27062    0.01541 -17.560   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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: 208.07

Number of Fisher Scoring iterations: 4
```

To generate a plot, copy and paste the following lines into the Script window and submit them.

```
    fit = predict(m1, type="response")
    f = womensrole$gender == "Female"
    plot(womensrole$education, fit, type="n", ylab="probability of agreeing",
xlab="Education", ylim=c(0,1))
    lines(womensrole$education[!f], fit[!f], lty=1)
    lines(womensrole$education[f], fit[f], lty=2)
    legend("topright",c("Fitted (Males)","Fitted (Females)"), lty=1:2,
bty="n")
    y = womensrole$agree / (womensrole$agree + womensrole$disagree)
    text(womensrole$education, y, ifelse(f,"\\VE","\\MA"), family
```

="HersheySerif", cex=1.25)



From the menu, select Statistics > Fit models > Generalized linear model... > name the model m2, change the + to a * on the right side of the model equation to include the interaction, and click Ok.

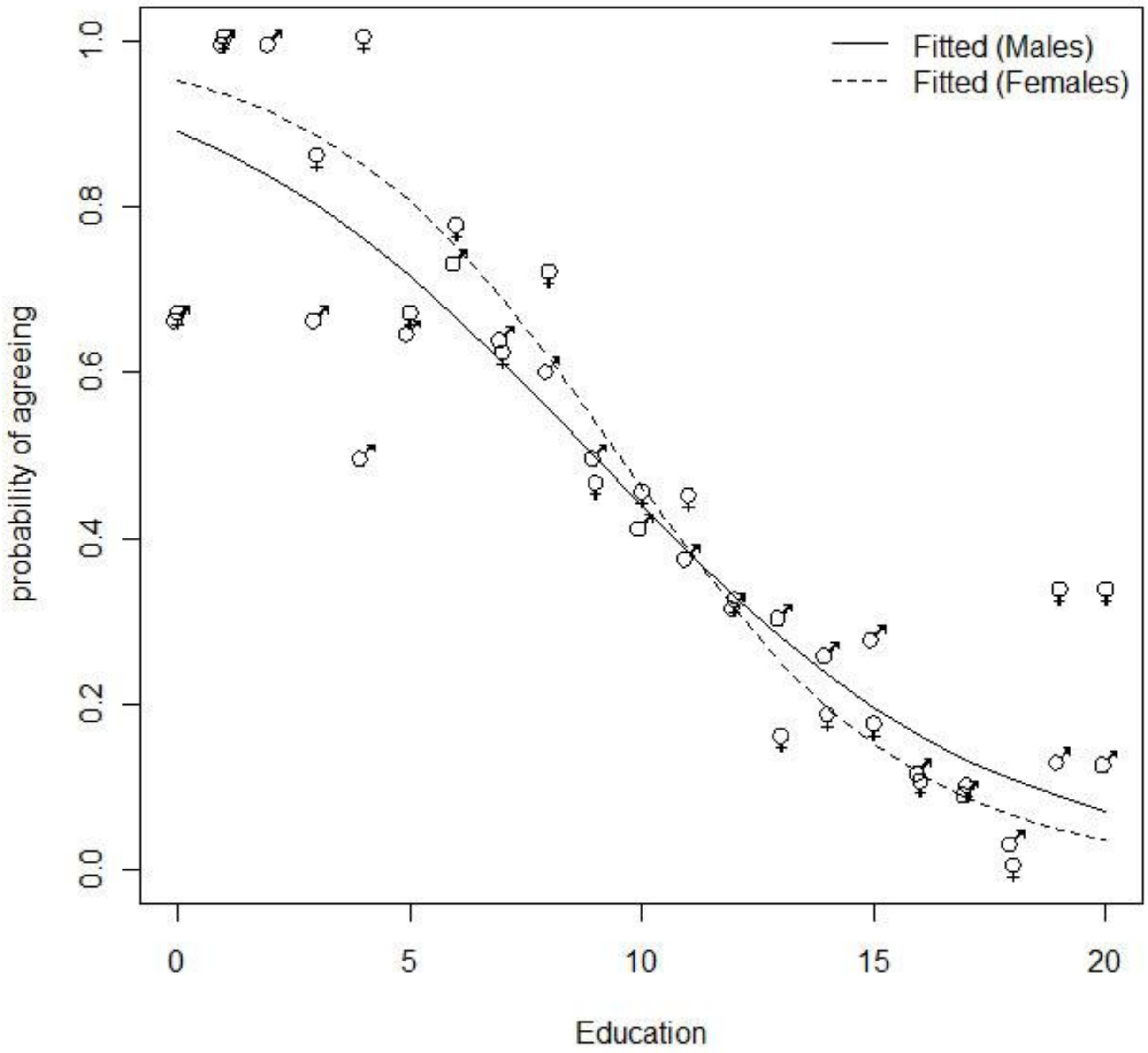
```
Call:
glm(formula = cbind(agree, disagree) ~ gender * education, family = 
binomial(logit), 
    data = womensrole)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.39097  -0.88062   0.01532   0.72783   2.45262  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
```

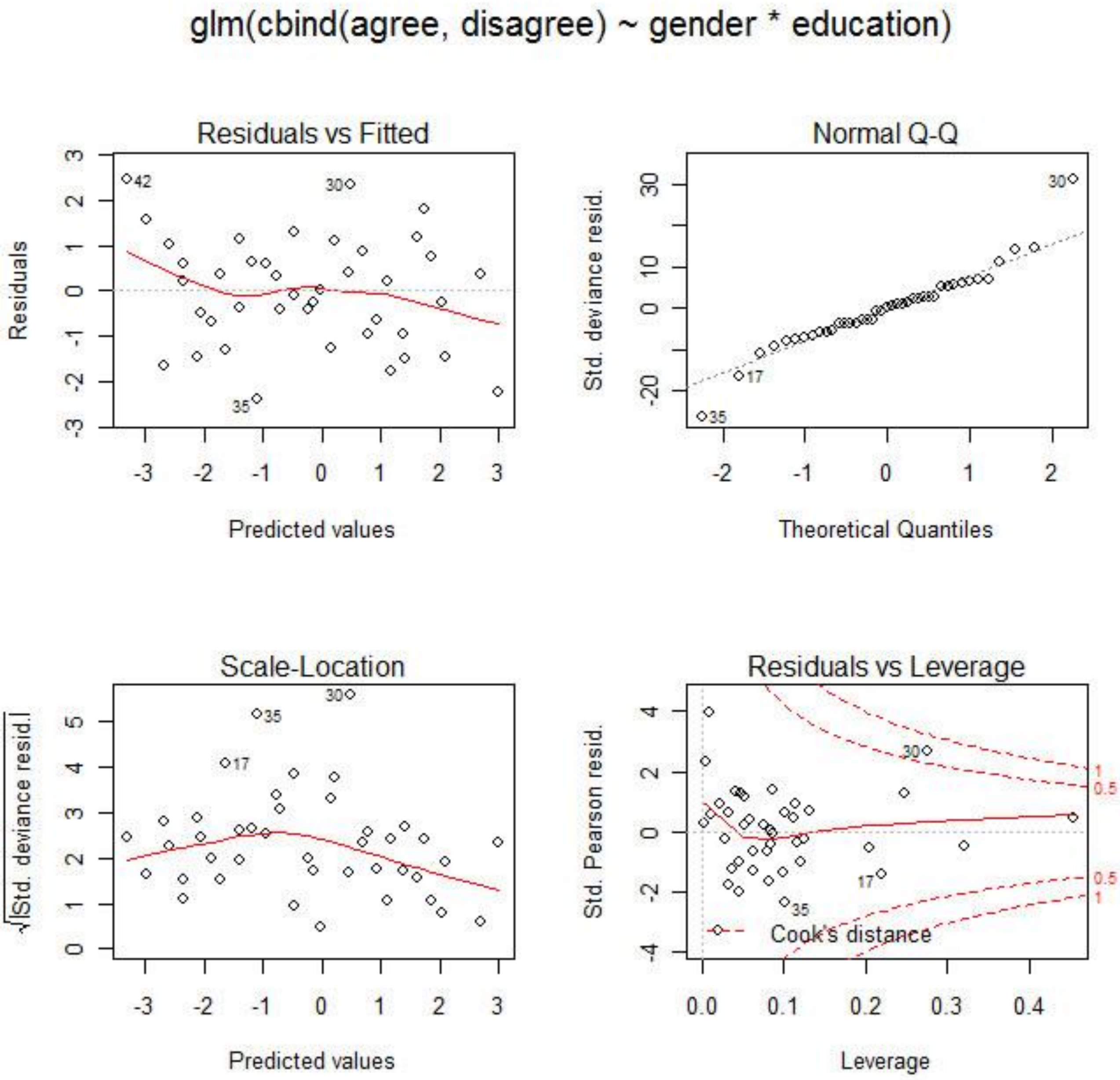
```
(Intercept)                  2.09820    0.23550   8.910  < 2e-16 ***
gender[T.Female]             0.90474    0.36007   2.513  0.01198 *
education                   -0.23403    0.02019 -11.592  < 2e-16 ***
gender[T.Female]:education -0.08138    0.03109  -2.617  0.00886 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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: 203.16

Number of Fisher Scoring iterations: 4
```

From the menu, select Models > Hypothesis tests > Compare models... > select m1 and m2 and click Ok.

```
Analysis of Deviance Table

Model 1: cbind(agree, disagree) ~ gender + education
Model 2: cbind(agree, disagree) ~ gender * education
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        38     64.007
2        37     57.103  1   6.9039    0.0086 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To generate a plot, copy and paste the above commands into the Script Window, substitute m2 for m1 in the first line, and submit them.


Fitted (Males)
Fitted (Females)
probability of agreeing
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Education



Normal Q-Q Plot does not make sense for a binomial distribution.

Colonic Polyps

From the R Commander menus select Data > Data in packages > Read data set from an attached package... >
Double click on HSAUR2 and select polyps, then click OK.

To see a description, from the R commander menu select Data > Active data set > Help on active data set (if available)

Click View data set to view it.

From the menu, select Statistics > Fit models > Generalized linear model... > name the model m1, double click on number to add it to the left side of the model equation enter, double click on treat, +, and age to add them to the right side, double click on poisson, verify that the link changed to log, and click Ok.

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.2212  -3.0536  -0.1802   1.4459   5.8301  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    4.529024   0.146872   30.84  < 2e-16 ***
treat[T.drug] -1.359083   0.117643  -11.55  < 2e-16 ***
age           -0.038830   0.005955   -6.52 7.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

(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: 273.88

Number of Fisher Scoring iterations: 5
```

The residual deviance is much greater that its degrees of freedom, indicating overdispersion.

From the menu, select Statistics > Fit models > Generalized linear model... > change the family to quasipoisson, and click Ok.

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.2212  -3.0536  -0.1802   1.4459   5.8301  

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.52902    0.48106   9.415 3.72e-08 ***
treat[T.drug] -1.35908    0.38533  -3.527  0.00259 ** 
age           -0.03883    0.01951  -1.991  0.06284 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
```

```
(Dispersion parameter for quasipoisson family taken to be 10.72805)

    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
```

Driving and Back Pain

From the R Commander menus select Data > Data in packages > Read data set from an attached package... >
Double click on HSAUR2 and select backpain, then click OK.

To see a description, from the R commander menu select Data > Active data set > Help on active data set (if available)

Click View data set to view it.

To look at the data, from the R commander menu select Statistics > Contingency tables > Two-way table... > Row variable: driver, Column variable: suburban,select Percentages of total, and click ok.

```
> .Table
      suburban
driver  no yes
   no   73  13
   yes 127 221

> totPercents(.Table) # Percentage of Total
        no  yes Total
no    16.8  3.0  19.8
yes   29.3 50.9  80.2
Total 46.1 53.9 100.0
```

Copy and paste the following commands into the Script Window and Submit them.
If the survival package has not been installed, you will need to install it from the R Console.

```
    library("survival")
    m1=clogit(I(status=="case") ~ driver + suburban + strata(ID),
data=backpain)
    print(m1)
```

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


                 coef exp(coef) se(coef)    z     p
```

```
driver[T.yes]   0.658      1.93    0.294 2.24 0.025
suburban[T.yes] 0.255      1.29    0.226 1.13 0.260

Likelihood ratio test=9.55  on 2 df, p=0.00846  n= 434
```

Conditional on residence, we can say that the risk in a driver is about twice that of a nondriver. There is no evidence that where a person lines affects his/her risk.