# Week 3 Practical Session

*David Barron*

*Hilary Term 2018*

## Logistic regression

The outcome variable in a logistic regression in R can either be a numeric variable with values 0 and 1 or a factor with two levels. In that case, the first level (which is usually the one that if first alphabetically) is equivalent to 0 and the other level to 1. This is important, because you have to be able to interpret the direction of regression parameter estimates.

In this example we are wanting to investigate women's labour force participation `lfp`. The data *Mroz* is of married women in the US. The outcome variable is a factor with levels *no* and *yes*. Therefore, *no* is equivalent to 0, and so a positive regression parameter estimate means that an increase in the explanatory variable increases the probability of labour force participation. The other variables are *k5*: number of children 5 or younger; *k618*: number of children 6–18; *age*: age in years; *wc*: college attendance; *hc*: husband's college attendance; *lwg*: log expected wage rate; *inc*: family income exclusive of wife's income.

```
data(Mroz)
head(Mroz)
```

```
  lfp k5 k618 age  wc hc       lwg    inc
1 yes  1    0  32  no no 1.2101647 10.910
2 yes  0    2  30  no no 0.3285041 19.500
3 yes  1    3  35  no no 1.5141279 12.040
4 yes  0    3  34  no no 0.0921151  6.800
5 yes  1    2  31 yes no 1.5242802 20.100
6 yes  0    0  54  no no 1.5564855  9.859
```

```
b1 <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial(),
    data = Mroz)
summary(b1)
```

```

Call:
glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial(),
    data = Mroz)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1062  -1.0900   0.5978   0.9709   2.1893

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27

Number of Fisher Scoring iterations: 4
```

If we reverse the coding of the outcome variable, then the signs on the output will change:

```r
lfp.recode <- relevel(Mroz$lfp, "yes")
b1a <- glm(lfp.recode ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial,
    data = Mroz)
summary(b1a)
```

```
Call:
glm(formula = lfp.recode ~ k5 + k618 + age + wc + hc + lwg +
    inc, family = binomial, data = Mroz)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1893  -0.9709  -0.5978   1.0900   2.1062

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.182140   0.644375  -4.938 7.88e-07 ***
k5           1.462913   0.197001   7.426 1.12e-13 ***
k618         0.064571   0.068001   0.950 0.342337
age          0.062871   0.012783   4.918 8.73e-07 ***
wcyes       -0.807274   0.229980  -3.510 0.000448 ***
hcyes       -0.111734   0.206040  -0.542 0.587618
lwg         -0.604693   0.150818  -4.009 6.09e-05 ***
inc          0.034446   0.008208   4.196 2.71e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27

Number of Fisher Scoring iterations: 4
```

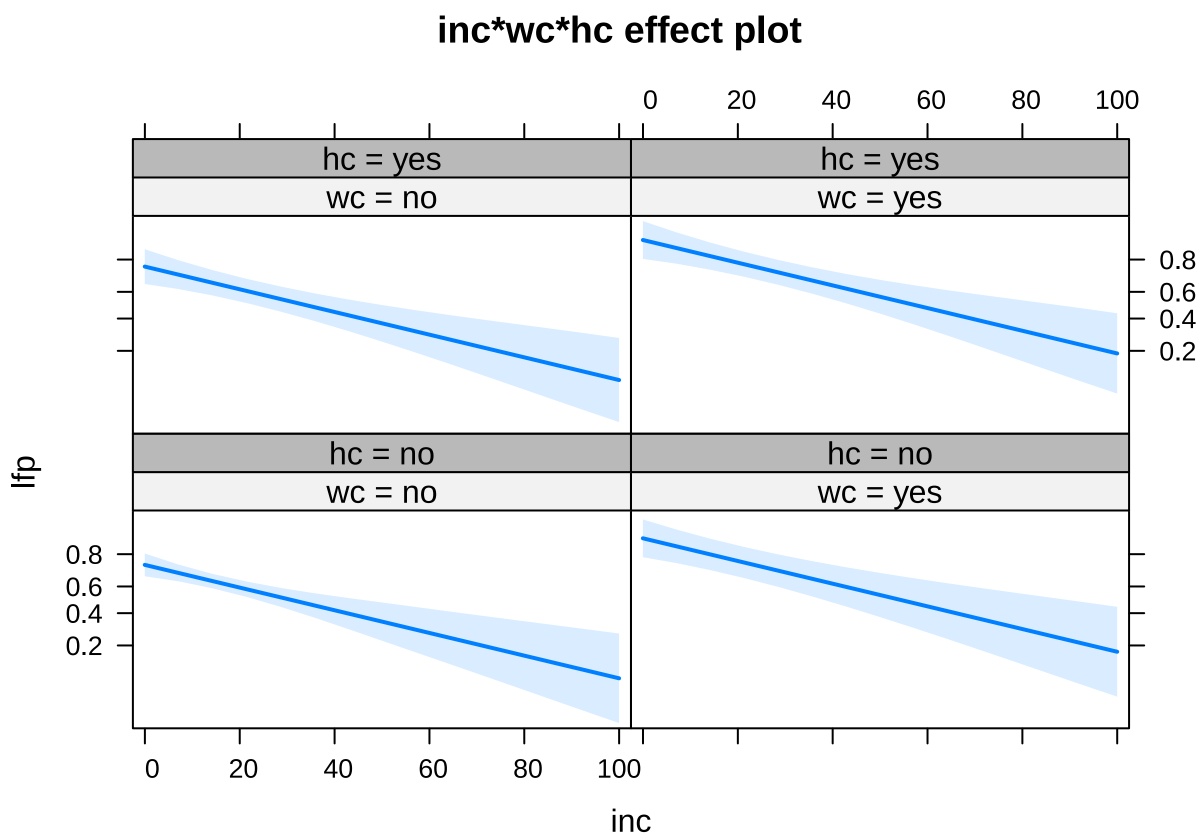
Notice that only the signs have changed.

## Interpreting parameter estimates

This can be done using an effect plot. Remember that the impact of any explanatory variable on predicted probabilities depends on the values of the other explanatory variables, so you have to set these too. The standard choice is the mean, but you might prefer the median. You might also prefer to fix categorical varibles at a particular level, rather than using the mean (which isn't really meaningful for a categorical variable).

The plots show the relatioship between household income and the probability of being in the labour force separately for the four different combinations of the two college education variables.

```r
plot(Effect(c("inc", "wc", "hc"), b1, typical = median), axes = list(x = list(rug = FALSE)))
```




### Multinomial logit

Multinomial logistic regression is often used for situations in which people have several choices. In this example, we have women's labour force participation again, but now we have three possible states: not in work, in part time work, and in full time work.

```r
data(Womenlf)
xtabs(~partic + region, Womenlf)
```

```
          region
partic      Atlantic BC Ontario Prairie Quebec
  fulltime         6  7      27       8     18
  not.work        20 14      64      17     40
  parttime         4  8      17       6      7
```

```r
Womenlf$partic <- relevel(Womenlf$partic, "not.work")

library(nnet)
m1 <- multinom(partic ~ hincome + children + region, data = Womenlf)
```

```
# weights:  24 (14 variable)
initial  value 288.935032
```

```
iter  10 value 208.509124
iter  20 value 207.732802
final  value 207.732796
converged
```

```r
summary(m1)
```

```
Call:
multinom(formula = partic ~ hincome + children + region, data = Womenlf)

Coefficients:
         (Intercept)     hincome childrenpresent   regionBC regionOntario
fulltime    2.124569 -0.10003520      -2.6978183 -0.4599668     0.1135477
parttime   -1.825805  0.00526185       0.1462146  1.0863441     0.2856932
         regionPrairie regionQuebec
fulltime     0.4680393   -0.3117081
parttime     0.5746633   -0.1105358

Std. Errors:
         (Intercept)    hincome childrenpresent  regionBC regionOntario
fulltime   0.7103039 0.02901632       0.3876747 0.7837059     0.6175130
parttime   0.8269888 0.02468883       0.4901642 0.7193065     0.6175031
         regionPrairie regionQuebec
fulltime     0.7332471    0.6515179
parttime     0.7259135    0.6873042

Residual Deviance: 415.4656
AIC: 443.4656
```

```r
Anova(m1)
```

```
Analysis of Deviance Table (Type II tests)

Response: partic
         LR Chisq Df Pr(>Chisq)
hincome    14.645  2  0.0006604 ***
children   65.204  2  6.937e-15 ***
region      7.416  8  0.4924500
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m2 <- multinom(partic ~ hincome + children, data = Womenlf)
```

```
# weights:  12 (6 variable)
initial  value 288.935032
iter  10 value 211.454772
final  value 211.440963
converged
```

```r
m2 <- update(m1, . ~ . - region)
```

```
# weights:  12 (6 variable)
initial  value 288.935032
iter  10 value 211.454772
final  value 211.440963
converged
```

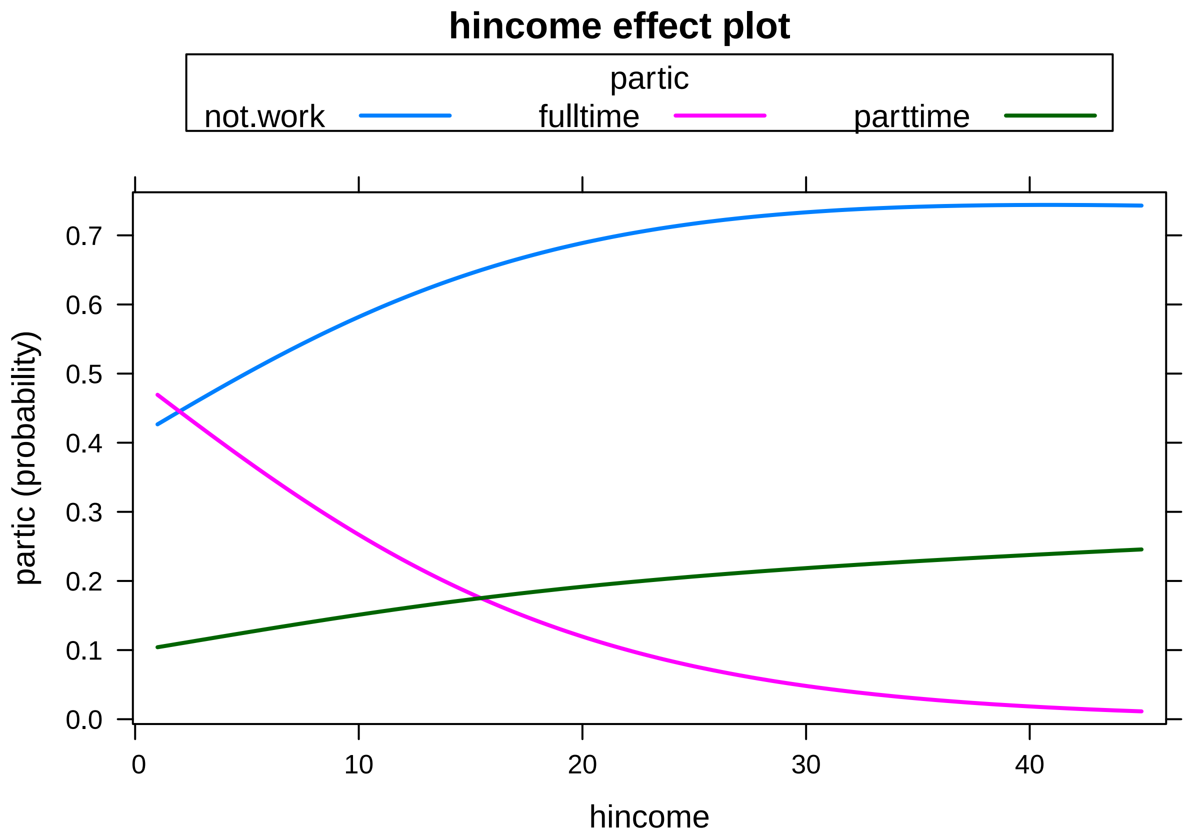
```r
summary(m2, Wald = TRUE)
```

```
Call:
multinom(formula = partic ~ hincome + children, data = Womenlf)

Coefficients:
         (Intercept)      hincome childrenpresent
fulltime    1.982842 -0.097232073     -2.55860537
parttime   -1.432321  0.006893838      0.02145558

Std. Errors:
         (Intercept)    hincome childrenpresent
fulltime   0.4841789 0.02809599       0.3621999
parttime   0.5924627 0.02345484       0.4690352

Value/SE (Wald statistics):
         (Intercept)    hincome childrenpresent
fulltime    4.095266 -3.4607098     -7.06407045
parttime   -2.417573  0.2939197      0.04574407

Residual Deviance: 422.8819
AIC: 434.8819
```

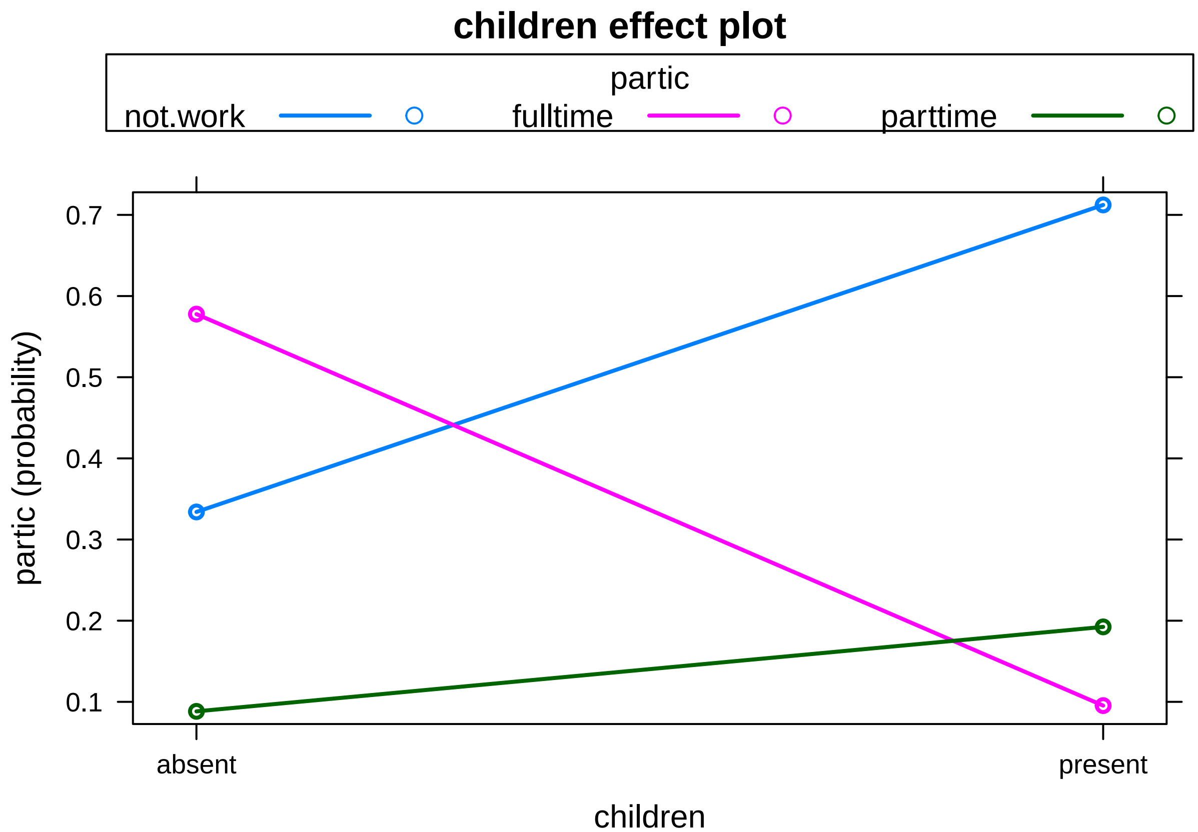
```r
anova(m2, m1)
```

```
Likelihood ratio tests of Multinomial Models

Response: partic
                        Model Resid. df Resid. Dev   Test    Df LR stat.
1          hincome + children       520   422.8819
2 hincome + children + region       512   415.4656 1 vs 2     8 7.416334
  Pr(Chi)
1
2 0.49245
```

```r
plot(Effect("hincome", m2, xlevels = list(hincome = 50)), confint = FALSE, lines = list(multiline = TRUE),
    axes = list(x = list(rug = FALSE)))
```




```r
plot(Effect("children", m2), confint = FALSE, lines = list(multiline = TRUE),
    axes = list(x = list(rug = FALSE)))
```




Compare to binary logit for full time, with part time treated as missing.

```
bin <- glm(partic ~ hincome + children, data = Womenlf, subset = partic != "parttime",
    family = binomial)
summary(bin)
```

```

Call:
glm(formula = partic ~ hincome + children, family = binomial,
    data = Womenlf, subset = partic != "parttime")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8590  -0.5955  -0.4503   0.7470   2.2860

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.03043    0.49806   4.077 4.57e-05 ***
hincome         -0.09964    0.02863  -3.481    5e-04 ***
childrenpresent -2.57445    0.36676  -7.019 2.23e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 269.49  on 220  degrees of freedom
Residual deviance: 197.60  on 218  degrees of freedom
```

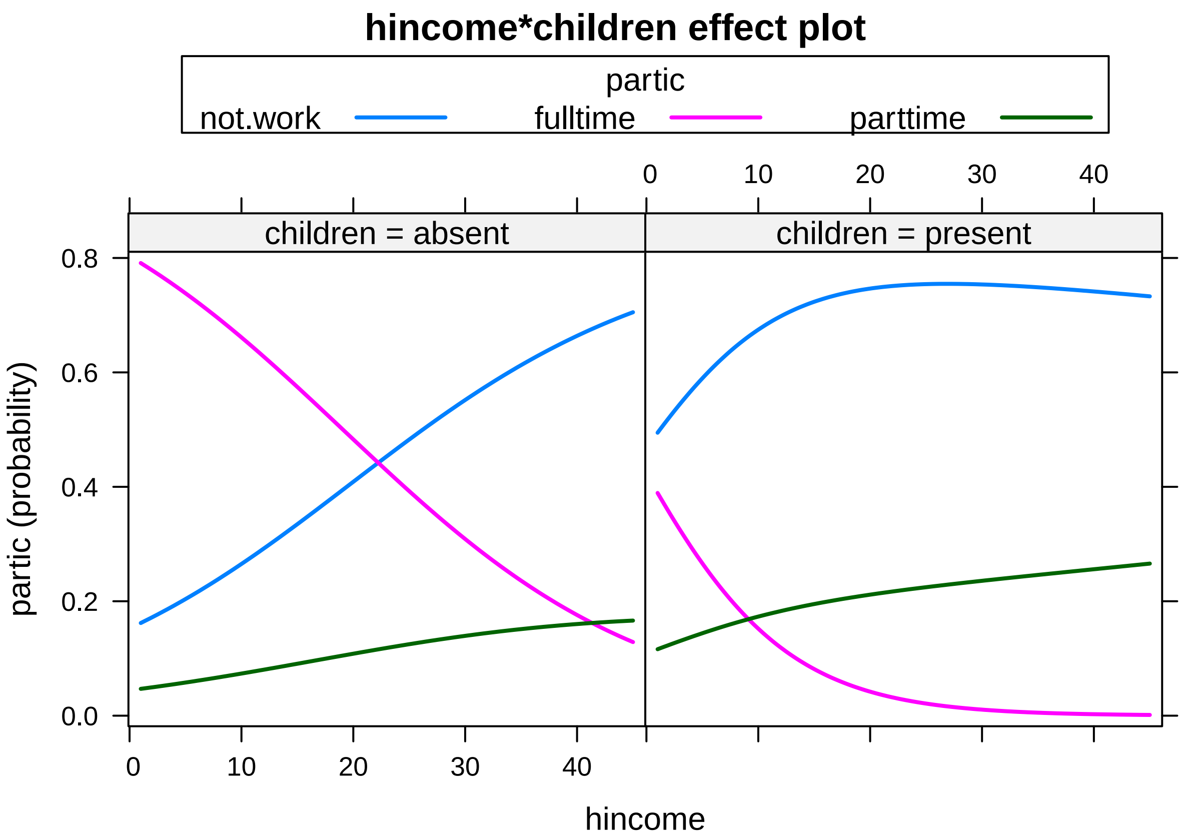
```
AIC: 203.6

Number of Fisher Scoring iterations: 5
```

You can see that these are reasonably similar. We could add an interaction.

```r
m3 <- update(m2, . ~ . + hincome:children)
```

```
# weights:  15 (8 variable)
initial  value 288.935032
iter  10 value 210.797079
final  value 210.714841
converged
```

```r
plot(Effect(c("hincome", "children"), m3, xlevels = list(hincome = 50)), confint = FALSE,
    axes = list(x = list(rug = FALSE)))
```




```r
Anova(m3)
```

```
Analysis of Deviance Table (Type II tests)

Response: partic
                 LR Chisq Df Pr(>Chisq)
hincome            15.153  2  0.0005123 ***
children           63.559  2  1.579e-14 ***
hincome:children    1.452  2  0.4837815
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

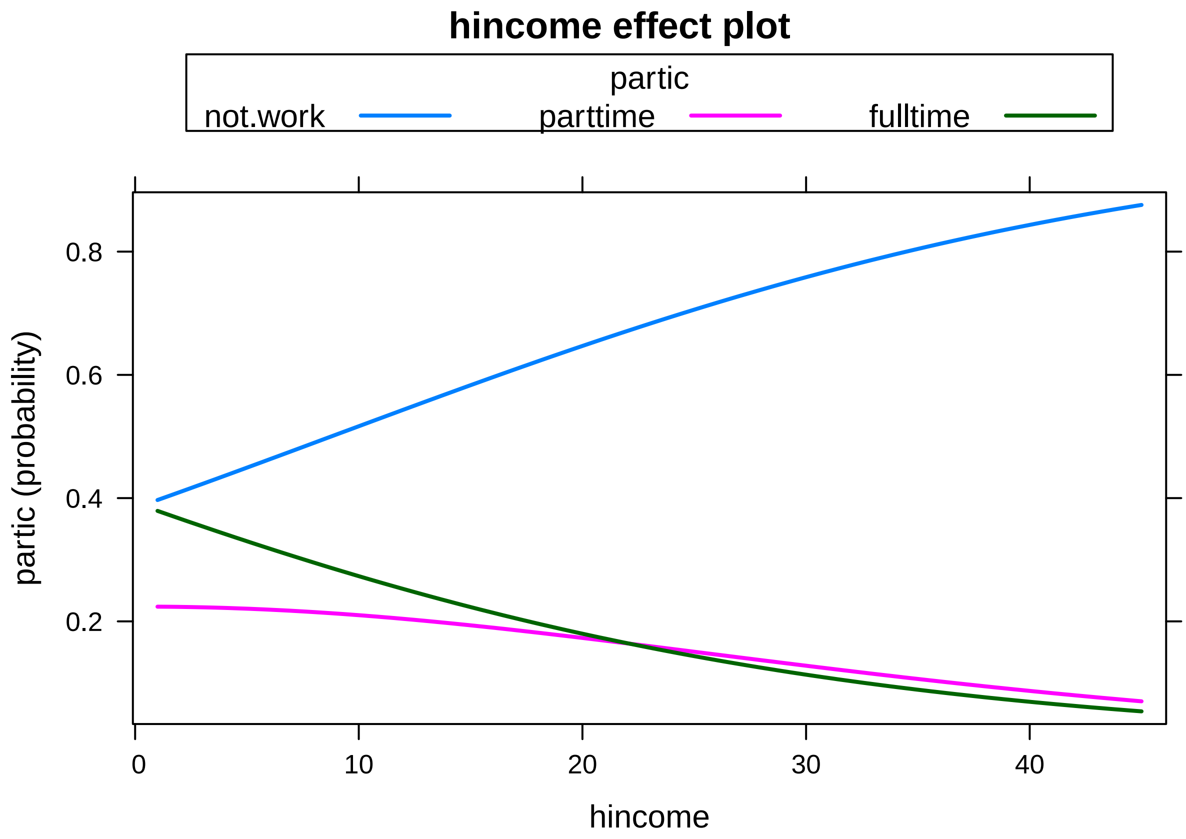
## Ordinal models

You have to make sure the levels of the factor you are going to analyse are in the correct order.

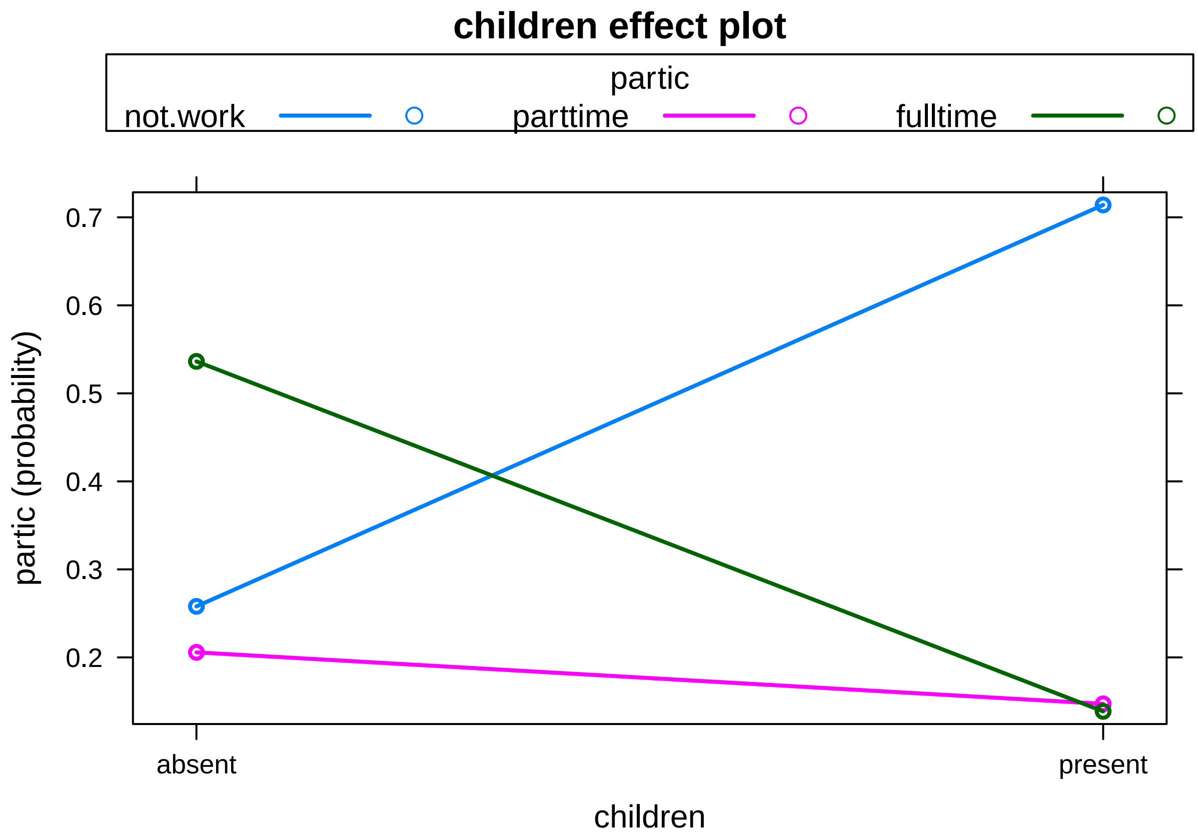
```
Womenlf$partic <- ordered(Womenlf$partic, levels = c("not.work", "parttime",
    "fulltime"))

o1 <- polr(partic ~ hincome + children, data = Womenlf, Hess = TRUE)

summary(o1)
```

```
Call:
polr(formula = partic ~ hincome + children, data = Womenlf, Hess = TRUE)

Coefficients:
                  Value Std. Error t value
hincome         -0.0539    0.01949  -2.766
childrenpresent -1.9720    0.28695  -6.872

Intercepts:
                  Value   Std. Error t value
not.work|parttime -1.8520  0.3863    -4.7943
parttime|fulltime -0.9409  0.3699    -2.5435

Residual Deviance: 441.663
AIC: 449.663
```

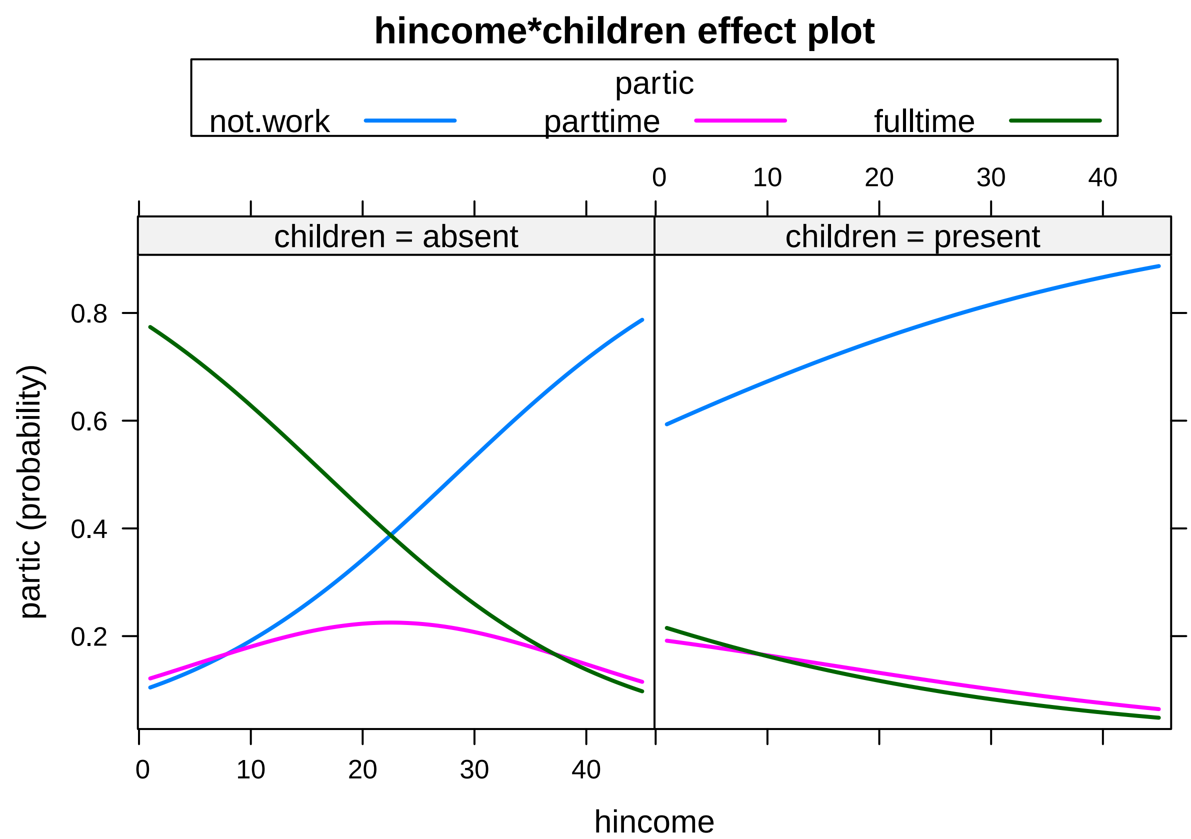
```
plot(Effect("hincome", o1, xlevels = list(hincome = 50)), confint = FALSE, axes = list(x = list(rug = FA
```




```
plot(Effect("children", o1), confint = FALSE, axes = list(x = list(rug = FALSE)))
```




```
o2 <- update(o1, . ~ . + hincome:children)
plot(Effect(c("hincome", "children"), o2, xlevels = list(hincome = 50)), confint = FALSE,
    axes = list(x = list(rug = FALSE)))
```




```r
AIC(o1)
```

```
[1] 449.663
```

```r
AIC(m2)
```

```
[1] 434.8819
```

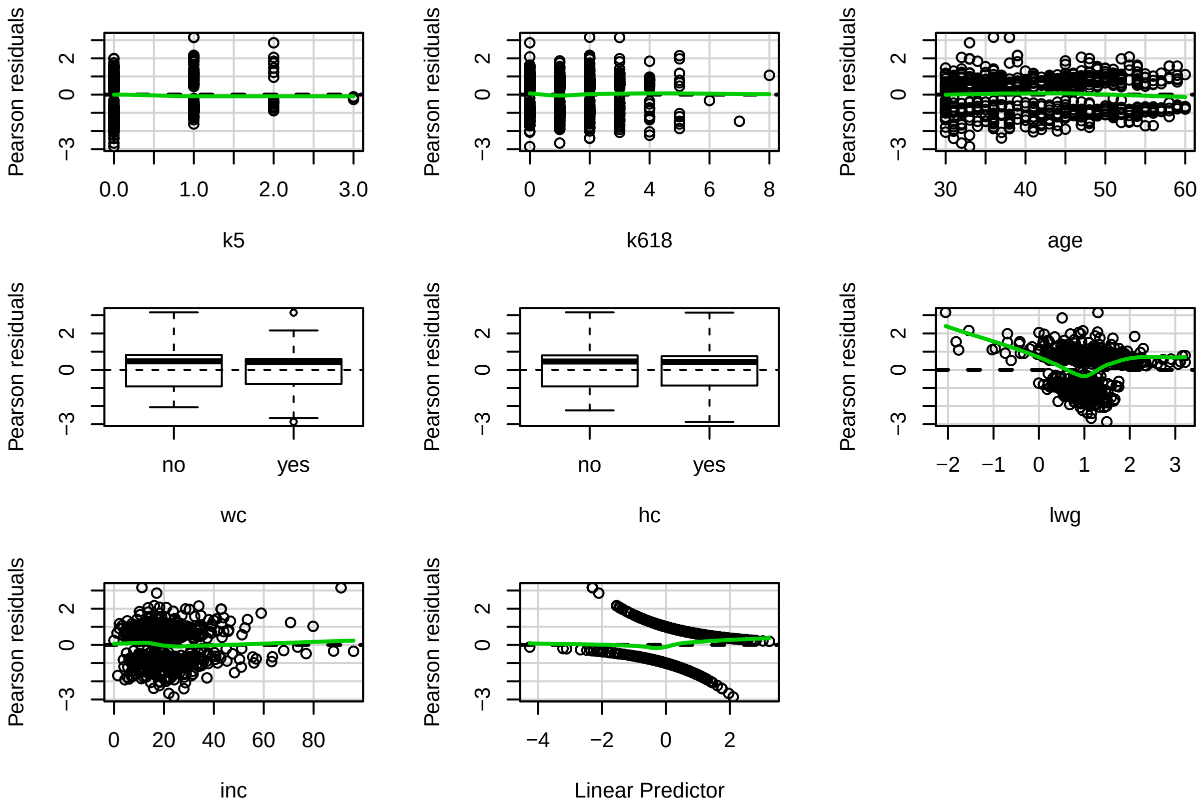
In this case, the fit of the ordinal model is worse than that of the multinomial we used before, so unlikely that the assumptions of the ordinal model are met.

## Diagnostics

Going back to the binary logistic regression that we started with, we can look at residuals and Cook's distance.

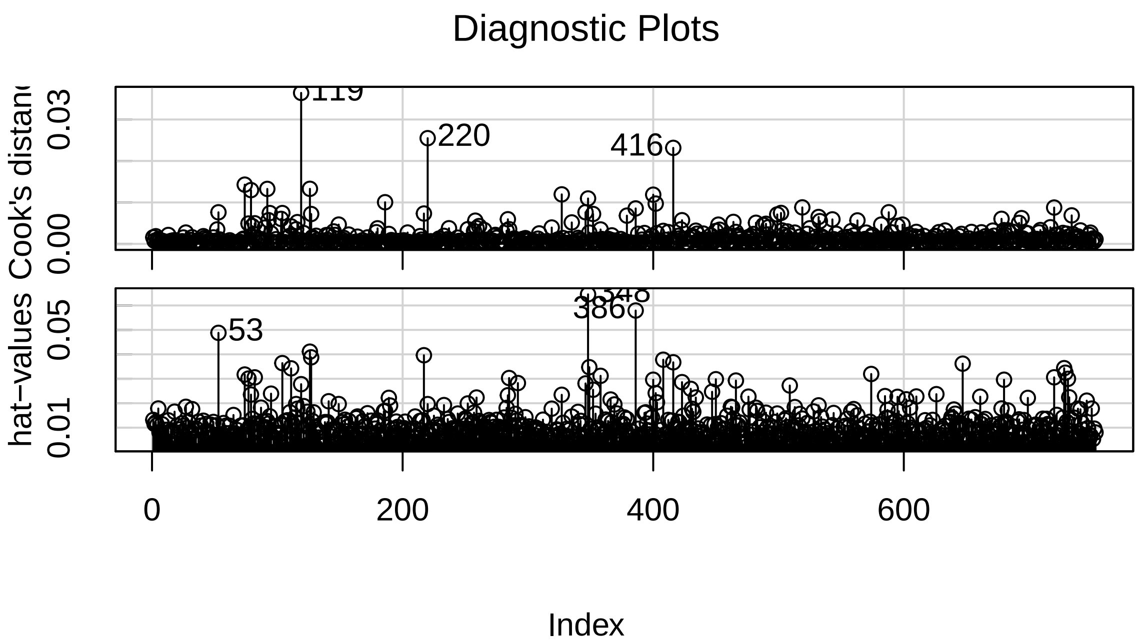
```r
residualPlots(b1)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
      Test stat Pr(>|t|)
k5        0.116    0.734
k618      0.157    0.692
age       1.189    0.275
wc           NA       NA
hc           NA       NA
lwg     153.504    0.000
inc       3.546    0.060
```

```
influenceIndexPlot(b1, vars = c("Cook", "hat"), id.n = 3)
```




```r
compareCoefs(b1, update(b1, subset = -c(119, 220, 416)))
```

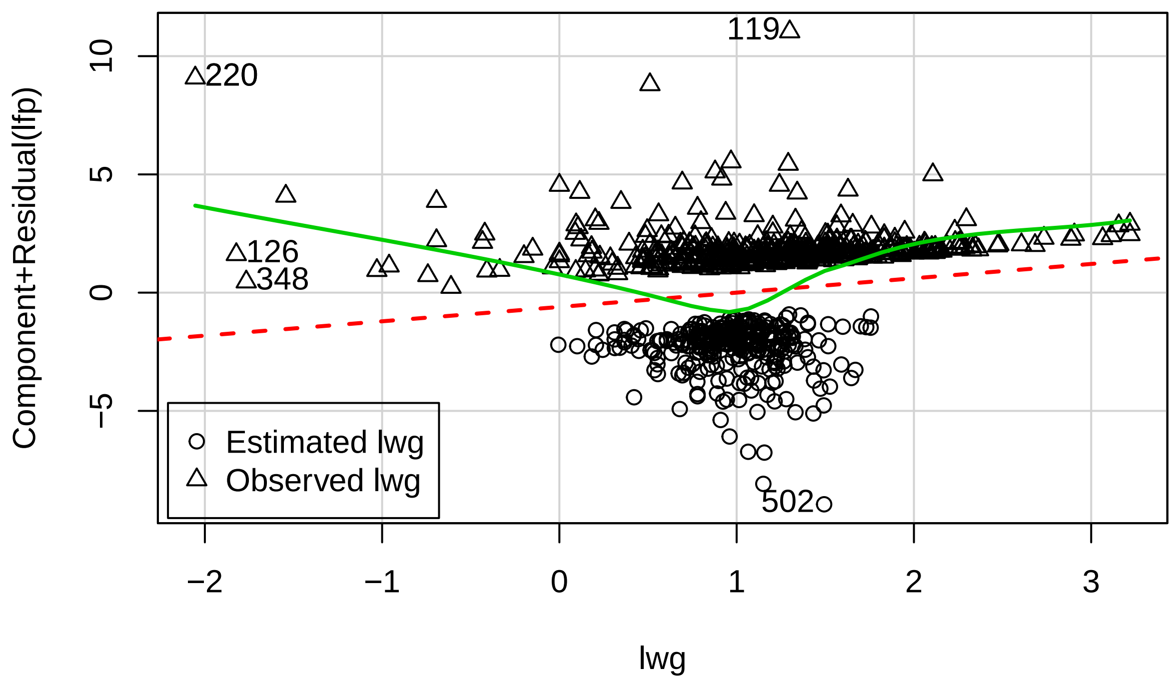
```
Call:
1: glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family =
  binomial(), data = Mroz)
2: glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family =
  binomial(), data = Mroz, subset = -c(119, 220, 416))
              Est. 1     SE 1   Est. 2     SE 2
(Intercept)  3.18214  0.64438  3.17623  0.65250
k5          -1.46291  0.19700 -1.54513  0.20273
k618        -0.06457  0.06800 -0.07170  0.06868
age         -0.06287  0.01278 -0.06382  0.01293
wcyes        0.80727  0.22998  0.72860  0.23300
hcyes        0.11173  0.20604  0.18053  0.20881
lwg          0.60469  0.15082  0.73622  0.15827
inc         -0.03445  0.00821 -0.03894  0.00853
```

```r
Mroz[c(119, 220, 416), ]
```

```
    lfp k5 k618 age  wc  hc       lwg   inc
119 yes  1    3  38 yes yes  1.299283 91.00
220 yes  1    2  36  no  no -2.054124 11.20
416 yes  1    2  39 yes  no -1.543298 16.12
```

The Cook's distance plots and hat value plots identify different cases as the most outlying, but none look particularly problematic. However, we can compare the coefficients when we remove those three cases. You can see that the coefficient of `lwg` does change by around 1 standard deviation, so there is some evidence of lack of fit here. This variable is unusual in that how it is defined depends on the outcome variable. For women in the labour force, it is the log of actual wage, but for those that aren't, it is the log of predicted wage. Let's have a look at a component plus residual plot.

```
crPlots(b1, "lwg", pch = as.numeric(Mroz$lfp), id.n = 3)

legend("bottomleft", c("Estimated lwg", "Observed lwg"), pch = 1:2, inset = 0.01)
```




We can see the unusual shape that this data has generated. We can see that case 220 is unusual because the person has 3 children, works, has a low income and a low wage.

## Homework

1. Install the package AER.
2. In this package there is a data set called `ResumeNames`. Have a look at the help page for this data set.
3. The outcome variable of interest is *call* (whether or not a resume (that's American English for a CV) sent in response to a job advert generated a telephone call from a potential employer).
4. The research question is whether the probability of a call is influenced by whether the "candidate" (these were all fictitious) had an African-American or Caucasian-sounding name.
5. There are a number of other variables in the data that identify characteristics of the "candidate" and characteristics of the job.
6. Your task is to come up with the best model that tests the hypothesis that ethnicity is associated with employer response while also controlling for other possible confounding variables.
7. Make sure that you can interpret your results. How would you explain to the reader of a paper in which you presented your results how much difference there was between employer responses to "Caucasion" and "African-American" applicants?