MA678 homework 05

Multinomial Regression

Shiyu Zhang

Oct, 2018

Multinomial logit:

Using the individual-level survey data from the 2000 National Election Study (data in folder nes), predict party identification (which is on a 7-point scale) using ideology and demographics with an ordered multinomial logit model.

1. Summarize the parameter estimates numerically and also graphically.

```
m1 <- polr(partyid3 ~ ideo + race + age_10, Hess=TRUE, data=nes_data_comp)
summary(m1)
```

```
## Call:
## polr(formula = partyid3 ~ ideo + race + age_10, data = nes_data_comp,
##     Hess = TRUE)
##
## Coefficients:
##                         Value Std. Error t value
## ideomoderate          1.07321    0.42732  2.5115
## ideoconservative      2.37255    0.22614 10.4915
## raceblack            -2.10233    0.39665 -5.3002
## raceasian            -0.55728    0.63683 -0.8751
## racenative american -0.08719    0.45164 -0.1931
## racehispanic         -0.78259    0.36861 -2.1231
## age_10               -0.07244    0.06187 -1.1707
##
## Intercepts:
##                                                                          Value
## 0. dk/ na/ other/ refused to answer/ no|1. democrats (including leaners) -12.7902
## 1. democrats (including leaners)|2. independents                           0.9991
## 2. independents|3. republicans (including leaners)                         1.4017
## 3. republicans (including leaners)|9. apolitical (1966 only: and dk)     295.5452
##                                                                          Std. Error
## 0. dk/ na/ other/ refused to answer/ no|1. democrats (including leaners)  28.2115
## 1. democrats (including leaners)|2. independents                           0.3553
## 2. independents|3. republicans (including leaners)                         0.3582
## 3. republicans (including leaners)|9. apolitical (1966 only: and dk)       0.3582
##                                                                          t value
## 0. dk/ na/ other/ refused to answer/ no|1. democrats (including leaners)  -0.4534
## 1. democrats (including leaners)|2. independents                           2.8122
## 2. independents|3. republicans (including leaners)                         3.9135
## 3. republicans (including leaners)|9. apolitical (1966 only: and dk)     825.1240
##
## Residual Deviance: 797.1793
## AIC: 819.1793
## (8 observations deleted due to missingness)
```

2. Explain the results from the fitted model.

```r
confint(m1)
```

```
## Waiting for profiling to be done...
```

```
##                           2.5 %      97.5 %
## ideomoderate         0.2119950  1.90029678
## ideoconservative     1.9401793  2.82839066
## raceblack           -2.9374471 -1.36652994
## raceasian           -1.8591087  0.68485196
## racenative american -0.9756851  0.80663089
## racehispanic        -1.5205078 -0.06831555
## age_10              -0.1943026  0.04861058
```

```r
"ideo: moderates and conservatives are more likely to be republicans. In particular,
a moderate has 1.07 increase in the expected value on the log odds scale, given all
of the other variables in the model are held constant. Conservatives have a 2.37
increase in the log odds scale.

race: whites, and asian are more likely to identify themselves as republicans.

age_10: One unit increase in age, we expect a decrease 0.07 (-0.07) in the expect
value of partyid3 on the log odds scale, given all of the other variables in the model are held constan
```

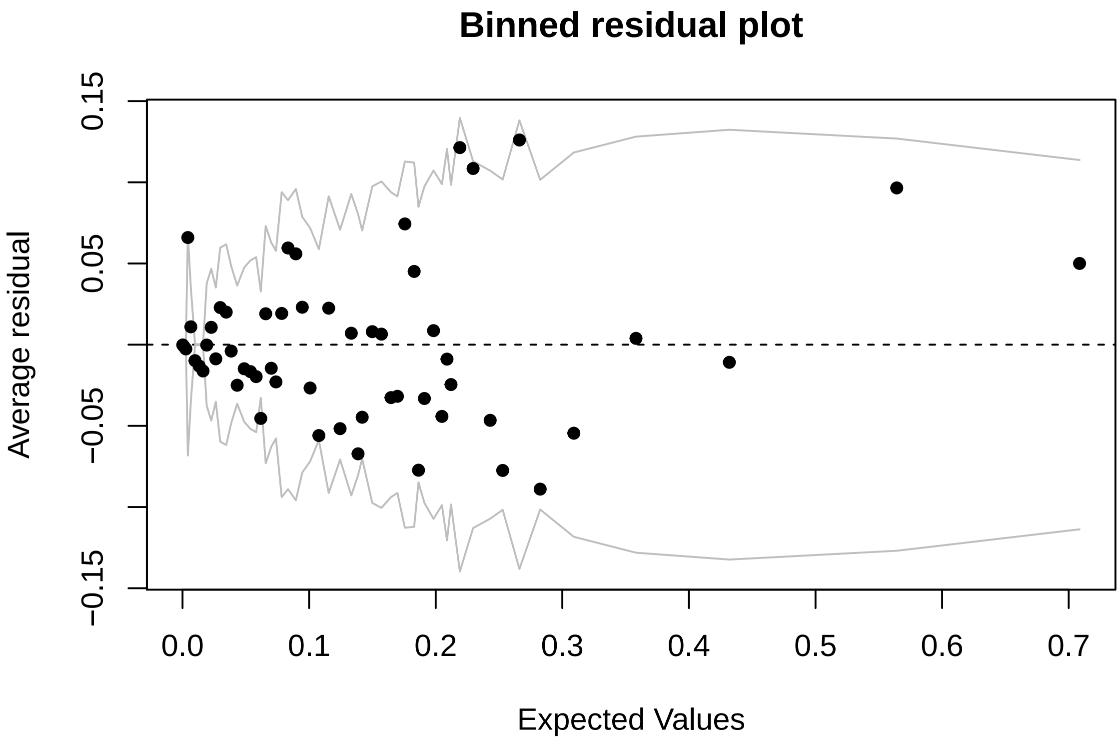
```
## [1] "ideo: moderates and conservatives are more likely to be republicans. In particular, \na moderat
```

3. Use a binned residual plot to assess the fit of the model.

```r
a1<- nnet::multinom(partyid7 ~ ideology + dem_therm, data = nes_data_comp)
```

```
## # weights:  28 (18 variable)
## initial  value 881.497298
## iter  10 value 748.117610
## iter  20 value 696.396875
## final  value 696.059464
## converged
```

```r
binnedplot(fitted(a1), resid(a1))
```



High School and Beyond

The hsb data was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The variables are gender; race; socioeconomic status; school type; chosen high school program type; scores on reading, writing, math, science, and social studies. We want to determine which factors are related to the choice of the type of program—academic, vocational, or general—that the students pursue in high school. The response is multinomial with three levels.

```
data(hsb)
?hsb
```

1. Fit a trinomial response model with the other relevant variables as predictors (untransformed).

```
require(nnet)
m2 <- multinom(prog ~ gender + race + ses + schtyp + read + write + math +science + socst, hsb, trace =

summary(m2)
```

```
## Call:
## multinom(formula = prog ~ gender + race + ses + schtyp + read +
##     write + math + science + socst, data = hsb, trace = FALSE)
##
## Coefficients:
##          (Intercept)  gendermale raceasian racehispanic racewhite
## general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156
## vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881
##               seslow sesmiddle schtyppublic        read       write
## general  1.09864111 0.7029621    0.5845405 -0.04418353 -0.03627381
## vocation 0.04747323 1.1815808    2.0553336 -0.03481202 -0.03166001
```

```
##                math    science       socst
## general  -0.1092888 0.10193746 -0.01976995
## vocation -0.1139877 0.05229938 -0.08040129
##
## Std. Errors:
##          (Intercept) gendermale raceasian racehispanic racewhite    seslow
## general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763
## vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
##          sesmiddle schtyppublic       read      write       math
## general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441
## vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131
##             science      socst
## general  0.03274038 0.02712589
## vocation 0.03424763 0.02938212
##
## Residual Deviance: 305.8705
## AIC: 357.8705
```

2. For the student with id 99, compute the predicted probabilities of the three possible choices.

```r
m3 <- step(m2, scope=~., direction="backward", trace = FALSE)
```

```
## trying - gender
## trying - race
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
## trying - gender
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
## trying - ses
## trying - schtyp
## trying - read
## trying - write
## trying - math
## trying - science
## trying - socst
## trying - ses
## trying - schtyp
## trying - read
## trying - math
## trying - science
## trying - socst
## trying - ses
## trying - schtyp
## trying - math
```

```
## trying - science
## trying - socst
```

```r
summary(m3)
```

```
## Call:
## multinom(formula = prog ~ ses + schtyp + math + science + socst,
##     data = hsb, trace = FALSE)
##
## Coefficients:
##          (Intercept)      seslow sesmiddle schtyppublic       math
## general     2.587029  0.87607389 0.6978995    0.6468812 -0.1212242
## vocation    6.687272 -0.01569301 1.2065000    1.9955504 -0.1369641
##             science       socst
## general  0.08209791 -0.04441228
## vocation 0.03941237 -0.09363417
##
## Std. Errors:
##          (Intercept)    seslow sesmiddle schtyppublic       math
## general     1.686492 0.5758781 0.4930330     0.545598 0.03213345
## vocation    1.945363 0.6690861 0.5571202     0.812881 0.03591701
##             science      socst
## general  0.02787694 0.02344856
## vocation 0.02864929 0.02586717
##
## Residual Deviance: 315.5511
## AIC: 343.5511
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- ti

## √ tibble  1.4.2     √ purrr   0.2.5
## √ tidyr   0.8.1     √ dplyr   0.7.6
## √ readr   1.1.1     √ stringr 1.3.1
## √ tibble  1.4.2     √ forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------------ tidyvers
## x dplyr::between()   masks data.table::between()
## x tidyr::expand()    masks Matrix::expand()
## x tidyr::fill()      masks VGAM::fill()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x dplyr::recode()    masks car::recode()
## x dplyr::select()    masks MASS::select()
## x purrr::some()      masks car::some()
## x purrr::transpose() masks data.table::transpose()
```

```r
predict <- hsb %>% filter(id == 99)
predict$prog
```

```
## [1] general
## Levels: academic general vocation
```

```r
predict(m2, newdata = predict, type = "probs")
```

```
##  academic   general  vocation
## 0.5076752 0.3753090 0.1170158
```

Happiness

Data were collected from 39 students in a University of Chicago MBA class and may be found in the dataset happy.

```r
library(faraway)
data(happy)
?happy
```

1. Build a model for the level of happiness as a function of the other variables.

```r
hap1 <- multinom(happy ~ money + sex+ love +work, data = happy)
```

```
## # weights:  54 (40 variable)
## initial  value 85.691759
## iter  10 value 68.212870
## iter  20 value 38.631288
## iter  30 value 28.889527
## iter  40 value 27.437462
## iter  50 value 26.714973
## iter  60 value 26.708293
## iter  70 value 26.703682
## final  value 26.703644
## converged
```

```r
summary(hap1)
```

```
## Call:
## multinom(formula = happy ~ money + sex + love + work, data = happy)
##
## Coefficients:
##    (Intercept)    money        sex        love      work
## 3     95.34718 8.207436   47.53907 -121.974570 -83.37503
## 4    108.15356 6.448616  126.62067 -144.128459 -19.45492
## 5    103.43665 6.504835   17.26409  -89.127605 -18.72953
## 6    -56.16590 6.632862  -29.02349   -9.832305 -20.71266
## 7     23.22477 6.557004   16.88244  -51.573474 -17.95274
## 8    -94.78326 6.586027  -39.30831    6.352961 -17.99809
## 9   -213.80005 6.596051   16.52875  -14.640589  13.18524
## 10  -149.75016 4.278169 -142.97178   95.281252 -45.18974
##
## Std. Errors:
##    (Intercept)       money          sex        love        work
## 3  0.325485656 27.67909907 3.254857e-01 0.328049326 0.328049314
## 4  0.799797489  4.62411816 7.997975e-01 1.571938267 1.210306606
## 5  0.746595285  4.62366481 1.682088e+00 1.493190570 0.865609268
## 6  2.378685936  4.62378421 4.087284e+00 1.201823879 1.931873302
## 7  1.620528738  4.62359628 1.651471e+00 0.829190888 0.703236473
## 8  1.228910839  4.62360881 1.892117e+00 1.036603126 0.814853390
```

```
## 9  0.103510063  4.62362103 1.035101e-01 0.310530190 0.414040255
## 10 0.001547399  0.06183351 9.923873e-08 0.004642197 0.007414509
##
## Residual Deviance: 53.40729
## AIC: 133.4073
```

2. Interpret the parameters of your chosen model.

```r
round(confint(hap1),2)
```

```
## , , 3
##
##                2.5 %  97.5 %
## (Intercept)   94.71   95.99
## money        -46.04   62.46
## sex           46.90   48.18
## love        -122.62 -121.33
## work         -84.02  -82.73
##
## , , 4
##
##                2.5 %  97.5 %
## (Intercept)  106.59  109.72
## money         -2.61   15.51
## sex          125.05  128.19
## love        -147.21 -141.05
## work         -21.83  -17.08
##
## , , 5
##
##               2.5 % 97.5 %
## (Intercept) 101.97 104.90
## money        -2.56  15.57
## sex          13.97  20.56
## love        -92.05 -86.20
## work        -20.43 -17.03
##
## , , 6
##
##               2.5 % 97.5 %
## (Intercept) -60.83 -51.50
## money        -2.43  15.70
## sex         -37.03 -21.01
## love        -12.19  -7.48
## work        -24.50 -16.93
##
## , , 7
##
##               2.5 % 97.5 %
## (Intercept)  20.05  26.40
## money        -2.51  15.62
## sex          13.65  20.12
## love        -53.20 -49.95
## work        -19.33 -16.57
##
```

```
## , , 8
##
##               2.5 % 97.5 %
## (Intercept) -97.19 -92.37
## money        -2.48  15.65
## sex         -43.02 -35.60
## love          4.32   8.38
## work        -19.60 -16.40
##
## , , 9
##
##                2.5 %  97.5 %
## (Intercept) -214.00 -213.60
## money         -2.47   15.66
## sex           16.33   16.73
## love         -15.25  -14.03
## work          12.37   14.00
##
## , , 10
##
##                2.5 %  97.5 %
## (Intercept) -149.75 -149.75
## money          4.16    4.40
## sex         -142.97 -142.97
## love          95.27   95.29
## work         -45.20  -45.18
```

3. Predict the happiness distribution for subject whose parents earn $30,000 a year, who is lonely, not sexually active and has no job.

```
(p1 <- data.frame(money = 30, sex = 0, love = 1, work =1))
```

```
##   money sex love work
## 1    30   0    1    1
```

```
predict(hap1, newdata = p1, type = "probs" )
```

```
##            2            3            4            5            6
## 1.476969e-83 2.134248e-24 1.301533e-23 1.000000e+00 8.507098e-35
##            7            8            9           10
## 3.097618e-18 5.706471e-44 4.207186e-91 4.243541e-71
```

newspaper survey on Vietnam War

A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset `uncviet`. Treat the opinion as the response and the sex and year as predictors. Build a proportional odds model, giving an interpretation to the estimates.

```
data(uncviet)
?uncviet
m4 <- multinom(policy ~ sex + year, weights = y, data = uncviet)
```

```
## # weights:  28 (18 variable)
## initial  value 4362.668354
```

```
## iter  10 value 3941.121869
## iter  20 value 3833.228107
## final  value 3832.113263
## converged
```

```r
summary(m4)
```

```
## Call:
## multinom(formula = policy ~ sex + year, data = uncviet, weights = y)
##
## Coefficients:
##   (Intercept)    sexMale yearGrad yearJunior yearSenior  yearSoph
## B   0.1345787 -0.5179948 0.719123  0.2566141 0.02396749 0.1489439
## C   1.0552414 -1.3547995 1.314791  0.4901925 0.51507411 0.1815794
## D  -1.7892132 -0.3937460 2.258658  0.6338067 1.07503163 0.2309365
##
## Std. Errors:
##   (Intercept)   sexMale  yearGrad yearJunior yearSenior  yearSoph
## B   0.1862561 0.1644782 0.1581396  0.1618130  0.1666919 0.1621340
## C   0.1640009 0.1411599 0.1459021  0.1502948  0.1501069 0.1549922
## D   0.2995572 0.2199894 0.2541701  0.2936060  0.2751441 0.3140088
##
## Residual Deviance: 7664.227
## AIC: 7700.227
```

```r
"for policy B, the yearGrad has a coefficient of 0.719 which shows that the grad students are more
likely to choose policy B. The same method for the other type of students, males have a negative
coefficient on the policy B,C,D. and Junior students are more likely to choose policy D compare to polic
B and C."
```

```
## [1] "for policy B, the yearGrad has a coefficient of 0.719 which shows that the grad students are mor
```

pneumonoconiosis of coal miners

The pneumo data gives the number of coal miners classified by radiological examination into one of three categories of pneumonoconiosis and by the number of years spent working at the coal face divided into eight categories.

```r
library(faraway)
data(pneumo,package="faraway")
?pneumo
```

```
## Help on topic 'pneumo' was found in the following packages:
##
##   Package               Library
##   faraway               /Library/Frameworks/R.framework/Versions/3.5/Resources/library
##   VGAM                  /Library/Frameworks/R.framework/Versions/3.5/Resources/library
##
##
## Using the first match ...
```

1. Treating the pneumonoconiosis status as response variable as nominal, build a model for predicting the frequency of the three outcomes in terms of length of service and use it to predict the outcome for a miner with 25 years of service.

```r
pne1 <- multinom(status ~ year,weights = Freq, data = pneumo)
```

```
## # weights:  9 (4 variable)
## initial  value 407.585159
## iter  10 value 208.724810
## final  value 208.724782
## converged
```

```r
summary(pne1)
```

```
## Call:
## multinom(formula = status ~ year, data = pneumo, weights = Freq)
##
## Coefficients:
##        (Intercept)        year
## normal   4.2916723 -0.08356506
## severe  -0.7681706  0.02572027
##
## Std. Errors:
##        (Intercept)       year
## normal   0.5214110 0.01528044
## severe   0.7377192 0.01976662
##
## Residual Deviance: 417.4496
## AIC: 425.4496
```

```r
#predict the miner with 25 years of service
p2 <- data.frame(year=25)
predict(pne1, newdata = p2, type = "probs")
```

```
##       mild     normal     severe
## 0.09148821 0.82778696 0.08072483
```

2. Repeat the analysis with the pneumonoconiosis status being treated as ordinal.

```r
pne2 <- polr(status ~ year, weights = Freq, data = pneumo)
summary(pne2)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = status ~ year, data = pneumo, weights = Freq)
##
## Coefficients:
##        Value Std. Error t value
## year 0.01566   0.009057    1.73
##
## Intercepts:
##               Value   Std. Error t value
## mild|normal   -1.8449  0.2492    -7.4039
## normal|severe  2.3676  0.2709     8.7411
##
## Residual Deviance: 502.1551
## AIC: 508.1551
```

```r
#predict the miner with 25 years of service
p3 <- data.frame(year=25)
predict(pne2, newdata =p3, type = "probs")
```

```
##       mild     normal     severe
## 0.09652357 0.78172799 0.12174844
```

3.Now treat the response variable as hierarchical with top level indicating whether the miner has the disease and the second level indicating, given they have the disease, whether they have a moderate or severe case.

4. Compare the three analyses.

```r
#There is not a significant difference between nominal and ordinal model
```

(optional) Multinomial choice models:

Pardoe and Simonton (2006) fit a discrete choice model to predict winners of the Academy Awards. Their data are in the folder academy.awards.

name	description
No	unique nominee identifier
Year	movie release year (not ceremony year)
Comp	identifier for year/category
Name	short nominee name
PP	best picture indicator
DD	best director indicator
MM	lead actor indicator
FF	lead actress indicator
Ch	1 if win, 2 if lose
Movie	short movie name
Nom	total oscar nominations
Pic	picture nom
Dir	director nom
Aml	actor male lead nom
Afl	actor female lead nom
Ams	actor male supporting nom
Afs	actor female supporting nom
Scr	screenplay nom
Cin	cinematography nom
Art	art direction nom
Cos	costume nom
Sco	score nom
Son	song nom
Edi	editing nom
Sou	sound mixing nom
For	foreign nom
Anf	animated feature nom
Eff	sound editing/visual effects nom
Mak	makeup nom
Dan	dance nom
AD	assistant director nom
PrNl	previous lead actor nominations
PrWl	previous lead actor wins

name	description
PrNs	previous supporting actor nominations
PrWs	previous supporting actor wins
PrN	total previous actor/director nominations
PrW	total previous actor/director wins
Gdr	golden globe drama win
Gmc	golden globe musical/comedy win
Gd	golden globe director win
Gm1	golden globe male lead actor drama win
Gm2	golden globe male lead actor musical/comedy win
Gf1	golden globe female lead actor drama win
Gf2	golden globe female lead actor musical/comedy win
PGA	producer's guild of america win
DGA	director's guild of america win
SAM	screen actor's guild male win
SAF	screen actor's guild female win
PN	PP*Nom
PD	PP*Dir
DN	DD*Nom
DP	DD*Pic
DPrN	DD*PrN
DPrW	DD*PrW
MN	MM*Nom
MP	MM*Pic
MPrN	MM*PrNl
MPrW	MM*PrWl
FN	FF*Nom
FP	FF*Pic
FPrN	FF*PrNl
FPrW	FF*PrWl

1. Fit your own model to these data.

```
model1 <-  select(oscar, -c("Comp","Name", "Movie"))
fit.model<- multinom(Ch ~. , data = model1)
```

```
## # weights:  180 (118 variable)
## initial  value 1791.836643
## iter  10 value 803.829776
## iter  20 value 607.272604
## iter  30 value 554.406319
## iter  40 value 541.084078
## iter  50 value 522.304809
## iter  60 value 503.220312
## iter  70 value 495.615648
## iter  80 value 483.665378
## iter  90 value 475.931525
## iter 100 value 475.873314
## final  value 475.873314
## stopped after 100 iterations
```

```
model2 <- multinom(Ch ~ Year +  Gm1 + Gf1 +Gf2 +PGA +DGA+SAM , data = model1)
```

```
## # weights:  27 (16 variable)
## initial  value 1791.836643
```

```
## iter  10 value 709.656976
## iter  20 value 694.787266
## iter  30 value 648.352403
## iter  40 value 616.448404
## iter  50 value 615.157702
## iter  60 value 615.146803
## iter  70 value 612.855407
## iter  80 value 611.264361
## iter  90 value 611.238659
## iter 100 value 610.446362
## final  value 610.446362
## stopped after 100 iterations
```

```r
summary(model2)
```

```
## Call:
## multinom(formula = Ch ~ Year + Gm1 + Gf1 + Gf2 + PGA + DGA +
##     SAM, data = model1)
##
## Coefficients:
##   (Intercept)       Year       Gm1        Gf1       Gf2       PGA
## 1    2011.206 -1.0042032  1.670262  1.8205832 1.8977265  2.031594
## 2    1973.220 -0.9837149 -1.335593 -0.5662558 0.1458876 -1.686199
##          DGA        SAM
## 1  4.0532336  2.4827177
## 2 -0.7367451 -0.4258609
##
## Std. Errors:
##    (Intercept)         Year       Gm1       Gf1        Gf2        PGA
## 1 0.0007594108 0.0001435079 0.1365230 0.1274401 0.01993375 0.05577680
## 2 0.0007588076 0.0001397689 0.1352197 0.1285092 0.02018365 0.05486513
##           DGA         SAM
## 1 0.007209765 0.004004649
## 2 0.006914309 0.002670114
##
## Residual Deviance: 1220.893
## AIC: 1252.893
```

2. Display the fitted model on a plot that also shows the data.
3. Make a plot displaying the uncertainty in inferences from the fitted model.