MA678 HW6

Multinomial Regression

Yuxi Wang

October 21, 2020

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 five-point scale) using ideology and demographics with an ordered multinomial logit model.

```
nes_data <- read.dta("/Users/mac/Desktop/BU Mssp/MA678/ROS-Examples-master/NES/data/nes5200_processed_vo

nes_data_dt <- data.table(nes_data)
  yr <- 2000
nes_data_dt_s<-nes_data_dt[ year==yr,]
nes_data_dt_s$income <- droplevels(nes_data_dt_s$income)
nes_data_dt_s$partyid7 <- droplevels(nes_data_dt_s$partyid7)
nes_data_dt_s$gender <- factor(nes_data_dt_s$gender, labels=c("male", "female"))
nes_data_dt_s$race <- factor(nes_data_dt_s$race, labels=c("white", "black", "asian",
                                    "native american", "hispanic"))
nes_data_dt_s$south <- factor(nes_data_dt_s$south)
nes_data_dt_s$ideo <- factor(nes_data_dt_s$ideo, labels=c("liberal", "moderate", "conservative"))
nes_new<-nes_data_dt_s[complete.cases(nes_data_dt_s[,list(partyid7,income,ideo,female,white)])]
nes_new$ideology <- scale(nes_new$ideo_feel,center=TRUE)
```

##1. Summarize the parameter estimates numerically and also graphically.

```
x = nes_new$partyid7
nes_new <- nes_new[!is.na(levels(x)[x]),]
fit_1 <- polr(factor(partyid7) ~ ideo + age + gender + race + south, Hess = TRUE, data = nes_new)
summary(fit_1)
```

```
## Call:
## polr(formula = factor(partyid7) ~ ideo + age + gender + race +
##     south, data = nes_new, Hess = TRUE)
##
## Coefficients:
##                        Value Std. Error t value
## ideomoderate         0.95339   0.330586  2.8840
## ideoconservative     1.94046   0.181464 10.6933
## age                 -0.01317   0.004937 -2.6676
## genderfemale        -0.38543   0.155547 -2.4779
## raceblack           -1.79583   0.277242 -6.4775
## raceasian            0.12546   0.544657  0.2303
## racenative american -0.13670   0.368338 -0.3711
## racehispanic        -0.62635   0.297434 -2.1058
```

```
## south1              0.21547   0.175126  1.2304
##
## Intercepts:
##                                                       Value   Std. Error t value
## 1. strong democrat|2. weak democrat                   -1.3844  0.3053    -4.5350
## 2. weak democrat|3. independent-democrat              -0.5338  0.2973    -1.7958
## 3. independent-democrat|4. independent-independent     0.2580  0.2969     0.8689
## 4. independent-independent|5. independent-republican  0.6528  0.2995     2.1799
## 5. independent-republican|6. weak republican          1.4496  0.3046     4.7597
## 6. weak republican|7. strong republican               2.4281  0.3159     7.6864
##
## Residual Deviance: 1889.52
## AIC: 1919.52
## (9 observations deleted due to missingness)
```

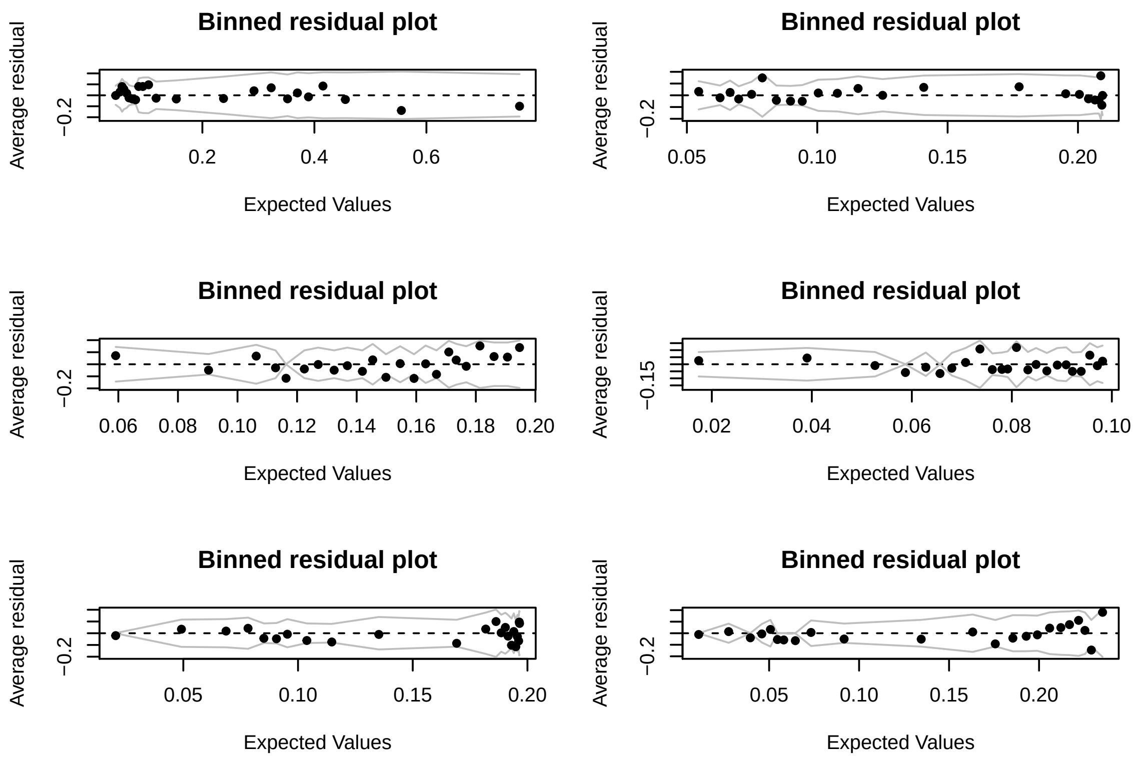
```r
round(summary(fit_1)$coef,2)
```

```
##                                                      Value Std. Error t value
## ideomoderate                                          0.95       0.33    2.88
## ideoconservative                                      1.94       0.18   10.69
## age                                                  -0.01       0.00   -2.67
## genderfemale                                         -0.39       0.16   -2.48
## raceblack                                            -1.80       0.28   -6.48
## raceasian                                             0.13       0.54    0.23
## racenative american                                  -0.14       0.37   -0.37
## racehispanic                                         -0.63       0.30   -2.11
## south1                                                0.22       0.18    1.23
## 1. strong democrat|2. weak democrat                  -1.38       0.31   -4.54
## 2. weak democrat|3. independent-democrat             -0.53       0.30   -1.80
## 3. independent-democrat|4. independent-independent    0.26       0.30    0.87
## 4. independent-independent|5. independent-republican  0.65       0.30    2.18
## 5. independent-republican|6. weak republican          1.45       0.30    4.76
## 6. weak republican|7. strong republican               2.43       0.32    7.69
```

##2. Explain the results from the fitted model. Interpretation: For age_10: The estimated value in the output is given in units of ordered logarithm or ordered logarithmic ratio. So for age_10, we can say that for every 1 unit increase in age (that is, from 20s to 30s), given all other variables in the equal odds, we expect the expected value of partyid3 to increase in the log odds ratio- 0.11. The model remains unchanged. For ideo: Moderates, especially conservatives, are more likely to become Republicans. In particular, assuming that all other variables in the model remain constant, the expected value of id3 for the moderate party increases by 1.09 in the log odds ratio. Conservatives are more likely to increase the log odds ratio by 2.02 For Race: Whites and Asians are more likely to see themselves as Republicans. Blacks strongly favor the Democratic Party.

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

```r
nes <- cbind(partyid7 = nes_new$partyid7, ideo = nes_new$ideo, race = nes_new$race, age = nes_new$age, 
nes <- data.frame(na.omit(nes))
resid <- model.matrix(~ factor(partyid7) - 1, data = nes) - fitted(fit_1)
par(mfrow = c(3, 2))
for (i in 1:6) {
  binnedplot(fitted(fit_1)[, i], resid[, i], cex.main = 1.3, main = "Binned residual plot")
}
```



(Optional) Choice models:

Using the individual-level survey data from the election example described in Section 10.9 (data available in the folder NES),

1. fit a logistic regression model for the choice of supporting Democrats or Republicans. Then interpret the output from this regression in terms of a utility/choice model.

2. Repeat the previous exercise but now with three options: Democrat, no opinion, Republican. That is, fit an ordered logit model and then express it as a utility/choice mode

Contingency table and ordered logit model

In a prospective study of a new living attenuated recombinant vaccine for influenza, patients were randomly allocated to two groups, one of which was given the new vaccine and the other a saline placebo. The responses were titre levels of hemagglutinin inhibiting antibody found in the blood six weeks after vaccination; they were categorized as “small”, “medium” or “large”.

treatment	small	moderate	large	Total
placebo	25	8	5	38
vaccine	6	18	11	35

The cell frequencies in the rows of table are constrained to add to the number of subjects in each treatment group (35 and 38 respectively). We want to know if the pattern of responses is the same for each treatment group.

```
p <- c(25,8,5)
v <- c(6,18,11)
t <- factor(c("small","moderate","large"))
data_1 <- data.frame(t,p,v)
data_2 <- data.frame(p,v)
```

1. Using a chisqure test and an appropriate log-linear model, test the hypothesis that the distribution of responses is the same for the placebo and vaccine groups.

```
chi_t <- chisq.test(data_2)
fit_2 <- multinom(t~p+v,data=data_1,trace=FALSE)
```

2. For the model corresponding to the hypothesis of homogeniety of response distributions, calculate the fitted values, the Pearson and deviance residuals, and the goodness of fit statistics X^2 and D. Which of the cells of the table contribute most to X^2 and D? Explain and interpret these results.

```
round(fitted(fit_2,2))
```

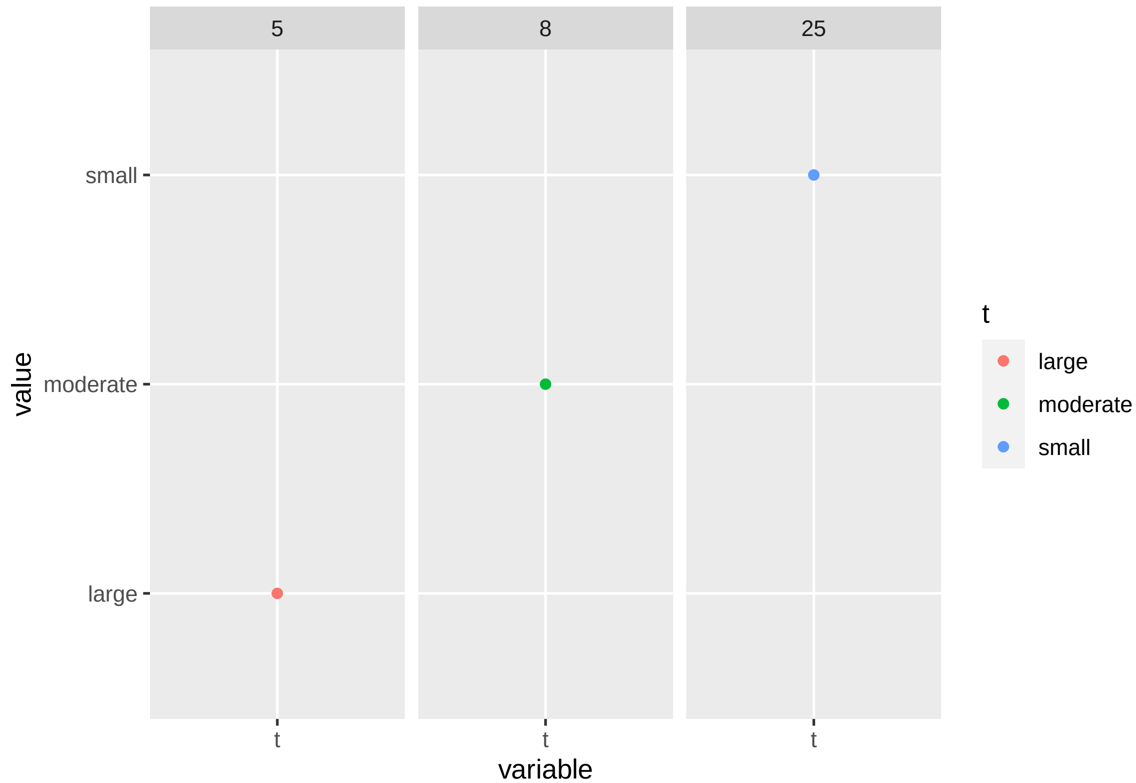
```
##   large moderate small
## 1     0        0     1
## 2     0        1     0
## 3     1        0     0
```

3. Re-analyze these data using ordered logit model (use `polr`) to estiamte the cut-points of a latent continuous response varaible and to estimate a location shift between the two treatment groups. Sketch a rough diagram to illustrate the model which forms the conceptual base for this analysis.

```
fit2_3 <- polr(t~p+v,data=data_1,Hess=TRUE)
summary(fit2_3)
```

```
## Call:
## polr(formula = t ~ p + v, data = data_1, Hess = TRUE)
##
## Coefficients:
##   Value Std. Error t value
## p 4.646      96.66 0.04807
## v 4.213      67.49 0.06242
##
## Intercepts:
##                  Value     Std. Error t value
## large|moderate   98.6841    3.6875    26.7620
## moderate|small  126.4846 1608.5622     0.0786
##
## Residual Deviance: 4.640427e-06
## AIC: 8.000005
```

```
ggplot(melt(data_1,id.vars=c("p","v")))+
  geom_point()+
  aes(x=variable,y=value,color=t)+
  facet_wrap(~p)
```



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
View(hsb)
```

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

```
fit3_1 <- polr(prog~gender+race,data=hsb, Hess=T)
summary(fit3_1)
```

```
## Call:
## polr(formula = prog ~ gender + race, data = hsb, Hess = T)
##
## Coefficients:
##                  Value Std. Error t value
## gendermale     0.04784     0.2728  0.1754
## raceasian     -0.57924     0.6976 -0.8304
## racehispanic   0.12293     0.5714  0.2151
## racewhite     -0.36875     0.4462 -0.8265
##
## Intercepts:
```

```
##                    Value   Std. Error t value
## academic|general -0.1672  0.4301    -0.3887
## general|vocation  0.8403  0.4347     1.9333
##
## Residual Deviance: 406.0601
## AIC: 418.0601
```

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

```
dt_3.2 <- subset(hsb,id==99,select=c("gender","race","ses","schtyp","read","write","math","science","so
  fit_3.2 <- polr(prog~gender+race+ses+schtyp+read+write+math+science+socst, data=hsb, Hess=T)
  pred <- predict(fit_3.2, type="probs", newdata=dt_3.2)
  pred
```

```
##  academic   general  vocation
## 0.5818527 0.2724298 0.1457174
```

Happiness

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

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

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

```
fit_4.1 <- polr(factor(happy)~money+sex+love+work, data=happy, Hess=T)
print(fit_4.1)
```

```
## Call:
## polr(formula = factor(happy) ~ money + sex + love + work, data = happy,
##     Hess = T)
##
## Coefficients:
##      money        sex       love       work
##  0.0224593 -0.4734369  3.6076452  0.8875135
##
## Intercepts:
##       2|3       3|4       4|5       5|6       6|7       7|8       8|9      9|10
##  5.470845  6.468394  9.159127 10.972524 11.511333 13.543305 17.290890 19.011197
##
## Residual Deviance: 94.86029
## AIC: 118.8603
```

```
exp(coef(fit_4.1))
```

```
##      money        sex       love       work
##  1.0227134  0.6228579 36.8791066  2.4290821
```

2. Interpret the parameters of your chosen model.

Here, I used a polynomial model to fit the data set. Among them, the gender coefficient is -0.4734369, that is to say: for a person who has sex, compared with other people who do not have sex, keeping other variables unchanged, her or her happiness is often 0.6228579 lower than that of asexual people Times. Among them, the money coefficient is 0.0224593, which means that for a person whose parent's income earns one more unit, his/her happiness is often 1.0227134 times the constant value of other people. The coefficient of love is 3.6076452, which means that for a person who has love, his or her happiness is often 36.8791066 times more than that of the person who does not love, and all other variables remain unchanged. The work coefficient is 0.8875135, which means that for a person with a job, his/her happiness is often 2.4290821 times that of a person without a job, and all other variables remain unchanged.

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

```
pred <- predict(fit_4.1, newdata=data.frame(money=30, sex=0, love=0,work=0))
pred
```

```
## [1] 2
## Levels: 2 3 4 5 6 7 8 9 10
```

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
fit_5 <- polr(policy~sex+year, data=uncviet, Hess=T)
print(fit_5)
```

```
## Call:
## polr(formula = policy ~ sex + year, data = uncviet, Hess = T)
##
## Coefficients:
##       sexMale      yearGrad    yearJunior    yearSenior      yearSoph
## -7.183340e-16  5.801722e-16  4.441982e-16  8.772898e-16 -6.661338e-16
##
## Intercepts:
##           A|B           B|C           C|D
## -1.098612e+00 -8.881784e-16  1.098612e+00
##
## Residual Deviance: 110.9035
## AIC: 126.9035
```

```
exp(coef(fit_5))
```

```
##    sexMale   yearGrad yearJunior yearSenior   yearSoph
##          1          1          1          1          1
```

Interpretation: The sexMale coefficient is -7.183340e-16, that is to say: for a person who are in the survey, the male take part in has a coefficient -7.183340e-16 will effect the policy of it. The coefficient of yearGrad/yearJunior/yearSenior/ yearSoph are 5.801722e-16/4.441982e-16/8.772898e-16/-6.661338e-16, which means that for a person has participate in the surey of the policy in Vietnam War in May 1967, they can effect the response by these coefficients.

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.

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

```
## Help on topic 'pneumo' was found in the following packages:
##
##   Package               Library
##   VGAM                  /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##   faraway               /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##
##
## Using the first match ...
```

```
View(pneumo)
```

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.

```
table1  <-  prop.table(xtabs(Freq ~ status + year, data = pneumo), 2)
View(table1)
year1 = c(5.8, 15, 21.5, 27.5, 33.5, 39.5, 46, 51.5)

D1 <-  data.frame(status = rep(pneumo$status, pneumo$Freq), year = rep(pneumo$year, pneumo$Freq))
fit6_1  <-  multinom(status ~ year, data = D1, trace = FALSE)
summary(fit6_1)
```

```
## Call:
## multinom(formula = status ~ year, data = D1, trace = FALSE)
##
## 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
```

```
# predict the outcome for a miner with 25 years of service
predict(fit6_1, newdata = list(year = 25), type = "probs")
```

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

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

```r
fit6_2 <- polr(status~year, data=D1, Hess=T)
summary(fit6_2)
```

```
## Call:
## polr(formula = status ~ year, data = D1, Hess = T)
##
## 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
pred <- predict(fit6_2, newdata=data.frame(Freq=3, year=25))
pred
```

```
## [1] normal
## Levels: mild normal severe
```

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.

```r
pneumo$status <- factor(pneumo$status,levels=c("normal","mild","severe"),ordered=TRUE)
fit6_3 <- polr(status~year,weights=Freq,data=pneumo,Hess=TRUE)
summary(fit6_3)
```

```
## Call:
## polr(formula = status ~ year, data = pneumo, weights = Freq,
##     Hess = TRUE)
##
## Coefficients:
##       Value Std. Error t value
## year 0.0959    0.01194   8.034
##
## Intercepts:
##             Value   Std. Error t value
## normal|mild  3.9558  0.4097     9.6558
## mild|severe  4.8690  0.4411    11.0383
##
## Residual Deviance: 416.9188
## AIC: 422.9188
```

```r
predict(fit6_3,newdata=list(year =25),type="probs")
```

```
##     normal       mild     severe
## 0.82610096 0.09601474 0.07788430
```

4. Compare the three analyses. Answer: I think these three results are close, While the second result has the highest AIC, and there predict results are close as well. At the same time, the first and second analysis have higher predict value than the third analysis.

(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 AcademyAwards.

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
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

name	description
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.
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.