Social Statistics
Statistics C116/C216
Professor: Mark S. Handcock

Homework 4

Due Monday, November 25, 2019

Unless otherwise noted, the data and directories for the problems are in

`http://www.stat.ucla.edu/~handcock/216/`

This address should be prepended to any addresses noted below.

1) *National Election Survey 1992.*
The *National Election Survey* contains the survey data of presidential preference and income along with other variables including sex, ethnicity, education, party identification, and political ideology. It covers elections from 1948 to 2002.
The data are in `examples/nes/nes5200processedvotersrealideo.dta`.
A detailed codebook with the variable names is available from the file: `datasets/nes/clean-up-nes.R`.
This pre-processes the dataset to clean it up. The output is in `datasets/nes/clean-up-nes.Rout`.
It has more details. I suggest using `presvote`, `female`, `black`, `educ3`, `income`, `partyid7`, `ideo_feel`, `state2` to select among their alternative versions (e.g., `educ1`, `educ2`, `educ3`).
The data can be read in via:

```
library(foreign)
# Note no conversion into factors
brdata <- read.dta( "http://www.stat.ucla.edu/~handcock/216/examples/nes/nes5200_processed_voters_realideo.dta", convert.factors=FALSE)
```

```
Nes.red <- brdata[brdata$presvote < 3,] # Remove third-party candidates
Nes.red$presvote <- ifelse(Nes.red$presvote==2, 1, 0) # recode as Republican or not
#
# I suggest  using "presvote", "female", "black", "educ3",
# "income", "partyid7", "ideo_feel", "state2"
# to select among thier alternative versions.
```

a) Fit, using Bayesian ideas, a logistic regression predicting support for George H. W. Bush in 1992 given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

One possible classical fit to the model to fit is:

```
fit <- glm(presvote ~ age + black + income + real_ideo + income:real_ideo,
           family = binomial(link = "logit"), data = Nes.red, subset=(year==1992))
```

b) Suppose you have a prior belief that the voters income is uncorrelated with their voting preference (given the other variables) Suppose you also believe you know the coefficient value up to a standard deviation of 1. Modify the model to reflect such a belief. Does the belief impact the inference very much?

2) *National Election Survey 1964*
Consider now same dataset, but information on the 1964 election. We wish to model vote preferences given income. Below, when we try to fit a model using ethnicity, specifically `black`, as a predictor, we run into a problem.

```
fit <- glm(presvote ~ female + black + income,
           family = binomial(link = "logit"), data = Nes.red, subset=(year==1964))
summary(fit)
```

```
##
## Call:
## glm(formula = presvote ~ female + black + income, family = binomial(link = "logit"),
##     data = Nes.red, subset = (year == 1964))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0918  -0.9412  -0.8422   1.3490   1.6406
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.15513    0.21592  -5.350 8.81e-08 ***
## female       -0.07918    0.13611  -0.582  0.56072
## black       -16.82942  420.51043  -0.040  0.96808
## income        0.19010    0.05839   3.256  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1337.7  on 1061  degrees of freedom
## Residual deviance: 1254.0  on 1058  degrees of freedom
##   (25997 observations deleted due to missingness)
## AIC: 1262
##
## Number of Fisher Scoring iterations: 16
```

What happened with the coefficient of `black` in 1964? Take a look at the data and figure out where this extreme estimate came from.

Using Bayesian ideas, suggest what can be done to fit the model in 1964? Fit the model within the Bayesian framework and compare the results of the two ways of fitting.

3) *Probit verses Logit.*
Comparing logit and probit fits to the the National Election Study data in 1992 in Question 1. Are the results are essentially the same.

4) *Multinomial logit.*
Using the individual-level survey data from the 2000 National Election Study predict party identification (which is on a five-point scale) using ideology and demographics with an ordered multinomial logit model. (As before, the data are in `examples/nes/nes5200processedvotersrealideo.dta`)

a) Summarize the parameter estimates numerically and also graphically.

b) Explain the results from the fitted model.

c) Use a graphical diagnostics to assess the fit of the model.