Homework 12

Miles Tweed

5/7/2021

Problem 1

Part 1

```
chile <- read.delim('../../Data/Chile.txt', sep = '\t')

chile.BIN <- chile %>%
  filter(vote %in% c('Y','N')) %>%
  drop_na() %>%
  mutate(vote01 = ifelse(vote == 'Y', 1,0)) %>%
  select(-vote)

glm.obj <- glm(vote01~., data = chile.BIN, family = 'binomial')

glm.null <- glm(vote01~1, data = chile.BIN, family = 'binomial')

anova(glm.null, glm.obj, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: vote01 ~ 1
## Model 2: vote01 ~ region + population + sex + age + education + income +
##     statusquo
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1702    2360.29
## 2      1691     703.48 11   1656.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is significant overall with a chi-squared value of 1656.8 which leads to p-value $< 2.2 \cdot 10^{-16}$.

Part 2

```
step(glm.obj, trace = F)
```

```
##
## Call:  glm(formula = vote01 ~ sex + education + statusquo, family = "binomial",
##     data = chile.BIN)
##
## Coefficients:
## (Intercept)         sexM  educationPS   educationS    statusquo
##      1.0153      -0.5742      -1.1074      -0.6828       3.1689
##
## Degrees of Freedom: 1702 Total (i.e. Null);  1698 Residual
```

```
## Null Deviance:       2360
## Residual Deviance: 708.2     AIC: 718.2
```

The variables region, age, population, and income were dropped from the model.

Let $p_i = p(vote01_i = 1 | sex_i,\ education_i,\ statusquo_i)$

$$\begin{cases} vote01_i & \sim_{indep.} Bin(1, p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) & = \beta_0 + \beta_1 D_{sexM,i} + \beta_2 D_{educationPS,i} + \beta_3 D_{educationS,i} + \beta_4 statusquo_i \end{cases}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = 1.015 - 0.574 D_{sexM,i} - 1.107 D_{educationPS,i} - 0.683 D_{educationS,i} + 3.689 statusquo_i$$

Part 3

```
glm.reduced <- glm(vote01~sex+education+statusquo, data = chile.BIN, family = "binomial")

summary(glm.reduced)
```

```
##
## Call:
## glm(formula = vote01 ~ sex + education + statusquo, family = "binomial",
##     data = chile.BIN)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2553  -0.2845  -0.1297   0.2009   2.9614
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0153     0.1890   5.373 7.75e-08 ***
## sexM         -0.5742     0.2022  -2.840 0.004518 **
## educationPS  -1.1074     0.2914  -3.800 0.000145 ***
## educationS   -0.6828     0.2217  -3.079 0.002077 **
## statusquo     3.1689     0.1448  21.886  < 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: 2360.29  on 1702  degrees of freedom
## Residual deviance:  708.24  on 1698  degrees of freedom
## AIC: 718.24
##
## Number of Fisher Scoring iterations: 6
```

The two most significant regressors are status quo and the dummy variable indicating post-secondary education.

for status quo

For a one unit increase in support for the status quo the odds that an individual will vote for Pinochet increase by a multiple of $e^{3.169} \approx 23.78$ holding other predictors constant. Or more simply, a person who is more supportive of the status quo is more likely to vote for Pinochet holding other variables constant.

for EducationPS

The odds that an individual who has a post-secondary education will vote for Pinochet are $e^{-}1.107 \approx 0.331$ times that of individuals who only have a primary education holding other predictors constant. Or more

simply, a person who has a post-secondary education is less likely to vote for Pinochet than someone with just a primary education holding other variables constant.

Part 4

```
freq.sex <- names(sort(table(chile.BIN$sex), decreasing = TRUE)[1])
freq.edu <- names(sort(table(chile.BIN$education), decreasing = TRUE)[1])
sq.med <- median(chile.BIN$statusquo)

# odds
freq.log.odds <- predict(glm.reduced, newdata=data.frame(sex=freq.sex,
                                                         education = freq.edu,
                                                         statusquo = sq.med))
freq.odds <- exp(freq.log.odds)
freq.odds
```

```
##         1
## 0.4367784
```

```
# probability
freq.prob <- predict(glm.reduced, newdata=data.frame(sex=freq.sex,
                                                         education = freq.edu,
                                                         statusquo = sq.med),
                     type='response')
freq.prob
```

```
##         1
## 0.3039984
```

Part 5

Full Model

```
glm.pred <- predict(glm.obj, type='response')
vote.pred <- ifelse(glm.pred > 0.50, "Yes","No")
true.labs <- ifelse(chile.BIN$vote01 == 1, "Yes", "No")
conf.mat <- table(Pred=vote.pred,
                  True=true.labs)
conf.mat
```

```
##      True
## Pred   No Yes
##   No  808  65
##   Yes  59 771
```

```
print(paste0("Percent correct classifications: ", round(mean(vote.pred == true.labs)*100,4)))
```

```
## [1] "Percent correct classifications: 92.7187"
```

Reduced Model

```
glm.pred <- predict(glm.reduced, type='response')
vote.pred <- ifelse(glm.pred > 0.50, "Yes","No")
true.labs <- ifelse(chile.BIN$vote01 == 1, "Yes", "No")
conf.mat <- table(Pred=vote.pred,
                  True=true.labs)
conf.mat
```

```
##      True
## Pred   No Yes
```

```
##   No  809  64
##   Yes  58 772
```

```
print(paste0("Percent correct classifications: ", round(mean(vote.pred == true.labs)*100,4)))
```

```
## [1] "Percent correct classifications: 92.8362"
```

The reduced model outperforms the full model by a bit (0.12%).

Problem 2

Part 1

```
library(nnet)
multinom.obj <- multinom(as.factor(vote) ~ age + sex + region + statusquo, data=chile, na.action = na.o
```

```
## # weights:  36 (24 variable)
## initial  value 3490.689201
## iter  10 value 2282.918477
## iter  20 value 2154.749143
## iter  30 value 2122.914761
## final  value 2122.850427
## converged
```

```
summary(multinom.obj)
```

```
## Call:
## multinom(formula = as.factor(vote) ~ age + sex + region + statusquo,
##     data = chile, na.action = na.omit)
##
## Coefficients:
##   (Intercept)         age       sexM   regionM     regionN    regionS
## N  -0.1295192 0.005397365  0.6993290 0.8411402 -0.30822679 0.33434661
## U   0.1589056 0.030373441 -0.2879216 1.3387552 -0.73715898 0.09402906
## Y  -0.3884494 0.025373911 -0.1039942 1.5077273  0.08062193 0.41435538
##     regionSA  statusquo
## N -0.0860732 -1.8230660
## U  0.0771308  0.3338119
## Y  0.2254778  1.8756710
##
## Std. Errors:
##   (Intercept)         age      sexM   regionM   regionN   regionS  regionSA
## N   0.3063535 0.006410852 0.1735622 0.7920856 0.2866417 0.2554054 0.2268856
## U   0.2941819 0.006311186 0.1734192 0.7611136 0.2954569 0.2502714 0.2252501
## Y   0.3119872 0.006577450 0.1807100 0.7742880 0.2971859 0.2593320 0.2420183
##   statusquo
## N 0.1318171
## U 0.1069452
## Y 0.1207042
##
## Residual Deviance: 4245.701
## AIC: 4293.701
```

The baseline category is individuals who will abstain from voting altogether.

Part 2

Let p_{ij} be the probability that observation i falls into category j where j is how an individual votes (yes for Pinochet, No against Pinochet, Undecided, or Abstain) and the baseline category is Abstain.

$$\begin{cases} Y_i & \sim_{indep.} Multinomial(p_{(i,Yes)},\ p_{(i,No)},\ p_{(i,Undecided)},\ p_{(i,Abstain)}), \\ \log\left(\frac{p_{ij}}{1-p_{im}}\right) & = \beta_{0,j} + \beta_{1,j} age_i + \beta_{2,j} D_{sexM,i} + \beta_{3,j} D_{regionM,i} + \beta_4 D_{regionN,i} + \beta_5 D_{regionS,i} + \\ & \beta_6 D_{regionSA,i} + \beta_7 statusquo_i, \\ P_{im} & = P(Y_i = m) = 1 - \sum_{j=1}^{m-1} p_{ij} \end{cases}$$

Part 3

a.

Holding other variables constant, the odds of an individual voting against Pinochet (voting "No") are $e^{0.005} = 1.005$ times that of them abstaining if the individual is one year older. In other words, an individual is slightly more likely to vote no than abstain if they are older.

Holding other variables constant, the odds of an individual voting for Pinochet (voting "Yes") are $e^{0.025373911} = 1.026$ times that of them abstaining if the individual is one year older. In other words, an individual is slightly more likely to vote yes than abstain if they are older.

b.

Holding other variables constant, the odds of an individual voting against Pinochet (voting "No") are $e^{0.699} = 2.011$ times that of them abstaining if the individual is male. In other words, an individual is twice as likely to vote no than abstain if they male.

Holding other variables constant, the odds of an individual voting for Pinochet (voting "Yes") are $e^{-0.104} = 0.901$ times that of them abstaining if the individual is male. In other words an individual is less likely to vote yes than abstain if they are male

Part 4

$$\begin{aligned} \log\left(\frac{p_{i,Yes}}{p_{i,No}}\right) &= \log\left(\frac{\frac{p_{i,Yes}}{p_{i,A}}}{\frac{p_{i,No}}{p_{i,A}}}\right) \\ &= \frac{p_{i,Yes}}{p_{i,A}} - \frac{p_{i,No}}{p_{i,A}} \\ &= (\beta_{0,Yes} - \beta_{0,No}) + (\beta_{1,Yes} - \beta_{1,No}) age_i + (\beta_{2,Yes} - \beta_{2,No}) D_{sexM,i} + \\ &(\beta_{3,Yes} - \beta_{3,No}) D_{regionM,i} + (\beta_{4,Yes} - \beta_{4,No}) D_{regionN,i} + (\beta_{5,Yes} - \beta_{5,No}) D_{regionS,i} + \\ &(\beta_{6,Yes} - \beta_{6,No}) D_{regionSA,i} + (\beta_{7,Yes} - \beta_{7,No}) statusquo_i \\ &= -0.259 + 0.0200 age_i - 0.803 D_{sexM,i} + 0.667 D_{regionM,i} + 0.389 D_{regionN,i} + \\ &0.080 D_{regionS,i} + 0.312 D_{regionSA,i} + 3.698 statusquo_i \end{aligned}$$

a.

Holding other variables constant, the odds of an individual voting for Pinochet (voting "Yes") are $e^{0.020} = 1.020$ times that of them voting against Pinochet (voting "No") if the individual is one year older. In other words, an individual is more likely to vote Yes than No if they are older.

b.

Holding other variables constant, the odds of an individual voting for Pinochet (voting "Yes") are $e^{-0.803} = 0.448$ times that of them voting against Pinochet (voting "No") if the individual is male. In other words an individual is less likely to vote yes than no if they are male