Assign7_Choi_Logistic_regression

Serena Choi

```
charity <- read.csv("charity.csv")
attach(charity)
summary(charity)
```

```
##       give              age             sex             educ
##  Min.   :    0.0   Min.   :18.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:    0.0   1st Qu.:32.00   1st Qu.:1.000   1st Qu.:1.000
##  Median :  150.0   Median :42.00   Median :2.000   Median :1.000
##  Mean   :  686.5   Mean   :45.27   Mean   :1.502   Mean   :1.242
##  3rd Qu.:  650.0   3rd Qu.:58.00   3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :20000.0   Max.   :95.00   Max.   :2.000   Max.   :2.000
##      income          trust
##  Min.   : 6000   Min.   :1.000
##  1st Qu.:24000   1st Qu.:2.000
##  Median :42000   Median :2.310
##  Mean   :42563   Mean   :2.343
##  3rd Qu.:60000   3rd Qu.:2.620
##  Max.   :78000   Max.   :4.000
```

```
age <- charity$age
```

Logistic Regression

For this assignment, I will examine the factors that determine whether a person donates to charity or not by using logistic regression. To do so, I recode the continuous dependent variable "give" into binary measure of 0 and 1. 1 indicates that a person donates to a charity.

```
#Recode give as a binary measure
give2<-as.numeric(give > 0)
table(give2)
```

```
## give2
##    0    1
##  685 1541
```

```
table(give2, income)    #tabulate give2 by income
```

```
##      income
## give2 6000 12000 18000 24000 30000 36000 42000 48000 54000 60000 66000 72000
##     0  101    60    77    62    56    53    56    46    49    37    42    26
##     1   57    55    98    92   135   124   137   135   151   170   193   103
##      income
## give2 78000
##     0    20
##     1    91
```

```
t1<-table(give2, income)
prop.table(t1)      #tabulate give2 by income, in proportions
```

```
##      income
## give2        6000       12000       18000       24000       30000       36000
##     0 0.045372866 0.026954178 0.034591195 0.027852650 0.025157233 0.023809524
##     1 0.025606469 0.024707996 0.044025157 0.041329739 0.060646900 0.055705301
##      income
## give2       42000       48000       54000       60000       66000       72000
##     0 0.025157233 0.020664870 0.022012579 0.016621743 0.018867925 0.011680144
##     1 0.061545373 0.060646900 0.067834681 0.076370171 0.086702606 0.046271339
##      income
## give2       78000
##     0 0.008984726
##     1 0.040880503
```

Similarly, I will recode the sex and education into dummy variables. I will also recode the income into thousands.

```
female <- as.numeric(charity$sex==2)
ugrad <- as.numeric(charity$educ==2)
income2<-income/1000
```

Now, we will write a logistic regression model with "give2" as a dependent variable, and "age", "female", "ugrad", "income2" and "trust" as dependent variables.

```
model1<-glm(give2~ age + female + ugrad + income2 + trust, family=binomial(link=logit))
summary(model1)
```

```
##
## Call:
## glm(formula = give2 ~ age + female + ugrad + income2 + trust,
##     family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4693  -1.1005   0.6065   0.8650   1.6766
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.589162   0.309556  -8.364  < 2e-16 ***
## age          0.018748   0.002919   6.423 1.34e-10 ***
## female       0.299764   0.098451   3.045  0.00233 **
## ugrad        0.732289   0.140474   5.213 1.86e-07 ***
## income2      0.026838   0.002562  10.474  < 2e-16 ***
## trust        0.513191   0.098321   5.220 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2748.1  on 2225  degrees of freedom
## Residual deviance: 2485.3  on 2220  degrees of freedom
## AIC: 2497.3
##
## Number of Fisher Scoring iterations: 4
```

According to the model1, one unit increase in age is associated with, on average, a 0.019 increase in the logit of giving, holding other variables constant. A female is associated with, on average, 0.3 increase in the logit of giving, holding other variables constant. Similarly, education higher than university graduate, on average, is associated with 0.732 increase in the logit of giving, and one unit increase in trust is associated with a 0.513 increase in the logit of giving, net of others, respectively. A one thousand dollar increase in income is associated with a 0.513 increase in the logit of giving, net of others. All independent variables here are statistically significant.

We will interpret the same model results in the odds.

```
exp(model1$coef)
```

```
## (Intercept)         age      female       ugrad     income2       trust
##  0.07508291  1.01892472  1.34953981  2.07983664  1.02720128  1.67061428
```

According to the odds, for every unit increase in age, female, univeristy grad, income and trust, on average, is associated with the increases in the odds of giving by 1.019, 1.350, 2.08, 1.027, 1.671, respectively, holding other variables constant.

For the percentage change in the odds, the interpretation of the model is the following.

```
delta<-1
B<-model1$coef
perchange<-100*((exp(B)*delta)-1)
perchange
```

```
## (Intercept)         age      female       ugrad     income2       trust
##  -92.491709    1.892472   34.953981  107.983664    2.720128   67.061428
```

For every unit increase in age, female, ugrad, income2 and trust, the odds of giving increase by 2%, 35%, 107%, 3%, and 67%, respectively and holding other variables constant.

We will now look at the model fit.

```
logLik(model1)
```

```
## 'log Lik.' -1242.651 (df=6)
```

```
deviance(model1)
```

```
## [1] 2485.302
```

```
n<-2226
df<-5
BIC<-(deviance(model1)) - (df*(log(n)))
BIC
```

```
## [1] 2446.763
```

Both the deviance and BIC are very high, making me less confident with the model fit. This suspicion is confirmed when we compare the model1 to other models using the likelihood ratio test.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
model2<-glm(give2~income2+age+trust, family=binomial(link=logit))
model3<-glm(give2~income2+age+trust+female, family=binomial(link=logit))
lrtest(model1, model2)
```

```
## Likelihood ratio test
##
## Model 1: give2 ~ age + female + ugrad + income2 + trust
## Model 2: give2 ~ income2 + age + trust
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   6 -1242.7
## 2   4 -1261.3 -2 37.27  8.073e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(model1, model3)
```

```
## Likelihood ratio test
##
## Model 1: give2 ~ age + female + ugrad + income2 + trust
## Model 2: give2 ~ income2 + age + trust + female
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -1242.7
## 2   5 -1257.2 -1 29.133  6.758e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```