Homework_6

Marjan Rezvani

11/8/2020

```
#model_logit1 <- glm(LABFORCE ~ AGE + educ_advdeg, family = binomial, data = dat_use1)
```

Linear Models such as OLS have some problems. These imply predicted values of Y that are greater than one or less than zero.

```
load('/Users/marjanrezvani/Documents/Fall2020/eco_stat/data/acs2017_ny/acs2017_ny_data.RData')

acs2017_ny$LABFORCE <- as.factor(acs2017_ny$LABFORCE)
levels(acs2017_ny$LABFORCE) <- c("NA","Not in LF","in LF")

acs2017_ny$MARST <- as.factor(acs2017_ny$MARST)
levels(acs2017_ny$MARST) <- c("married spouse present","married spouse absent","separated","divorced","
```

NA more generally means that the coefficient is not estimable. This can happen due to not having enough observations to estimate the relevant parameters. If predictors are categorical and you're adding interaction terms, an NA can also mean that there are no observations with that combination of levels of the factors. but Persons who are neither employed nor unemployed are not in the labor force. This category may include retired persons, students, and others who are neither working nor seeking work.

```
acs2017_ny$age_bands <- cut(acs2017_ny$AGE,breaks=c(0,25,35,45,55,65,100))
table(acs2017_ny$age_bands,acs2017_ny$LABFORCE)
```

```
##
##              NA Not in LF in LF
##   (0,25]   31680     11717 13256
##   (25,35]      0      4271 20523
##   (35,45]      0      4064 18924
##   (45,55]      0      5406 21747
##   (55,65]      0     10563 18106
##   (65,100]     0     28701  5880
```

```
pick_use1 <- (acs2017_ny$AGE >25) & (acs2017_ny$AGE <= 55)
dat_use1 <- subset(acs2017_ny, pick_use1)

dat_use1$LABFORCE <- droplevels(dat_use1$LABFORCE)
```

Baseline model,

```
model_logit1 <- glm(LABFORCE ~ AGE + I(AGE^2) + female + AfAm + Asian + race_oth + Hispanic
            + educ_hs + educ_somecoll + educ_college + educ_advdeg
            + MARST,
            family = binomial, data = dat_use1)
summary(model_logit1)
```

```
##
## Call:
## glm(formula = LABFORCE ~ AGE + I(AGE^2) + female + AfAm + Asian +
##     race_oth + Hispanic + educ_hs + educ_somecoll + educ_college +
##     educ_advdeg + MARST, family = binomial, data = dat_use1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6277   0.3476   0.4862   0.6459   1.5245
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.6023215  0.2445543   2.463  0.01378 *
## AGE                        0.0171486  0.0121072   1.416  0.15666
## I(AGE^2)                  -0.0003149  0.0001471  -2.141  0.03228 *
## female                    -0.6839386  0.0205171 -33.335  < 2e-16 ***
## AfAm                      -0.1906696  0.0282354  -6.753 1.45e-11 ***
## Asian                     -0.1112229  0.0374503  -2.970  0.00298 **
## race_oth                  -0.0781864  0.0332004  -2.355  0.01852 *
## Hispanic                   0.1653724  0.0313524   5.275 1.33e-07 ***
## educ_hs                    0.8972780  0.0310196  28.926  < 2e-16 ***
## educ_somecoll              1.4531782  0.0350710  41.435  < 2e-16 ***
## educ_college               1.9430903  0.0370924  52.385  < 2e-16 ***
## educ_advdeg                2.3676171  0.0437358  54.135  < 2e-16 ***
## MARSTmarried spouse absent -0.5222011  0.0517449 -10.092  < 2e-16 ***
## MARSTseparated            -0.1240651  0.0577062  -2.150  0.03156 *
## MARSTdivorced              0.0619381  0.0375785   1.648  0.09930 .
## MARSTwidowed              -0.3023247  0.0934446  -3.235  0.00121 **
## MARSTnever married        -0.3857612  0.0241093 -16.000  < 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: 71408  on 74934  degrees of freedom
## Residual deviance: 64847  on 74918  degrees of freedom
## AIC: 64881
##
## Number of Fisher Scoring iterations: 5
```

I am going to try defferent subset and variables which might be effected on our prediction.

```r
model_logit2 <- glm(LABFORCE ~ AGE + I(AGE^4) + female + AfAm
                    + educ_hs + educ_somecoll + educ_college + educ_advdeg + OWNERSHP
                      + MARST,
                    family = binomial, data = dat_use1)
summary(model_logit2)
```

```
##
## Call:
## glm(formula = LABFORCE ~ AGE + I(AGE^4) + female + AfAm + educ_hs +
##     educ_somecoll + educ_college + educ_advdeg + OWNERSHP + MARST,
##     family = binomial, data = dat_use1)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7091   0.3165   0.4770   0.6513   1.8128
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.543e-01  1.440e-01  -1.071 0.284120
## AGE                         7.135e-03  4.505e-03   1.584 0.113209
## I(AGE^4)                   -3.563e-08  1.455e-08  -2.449 0.014308 *
## female                     -7.619e-01  2.097e-02 -36.336  < 2e-16 ***
## AfAm                       -2.143e-01  2.802e-02  -7.649 2.03e-14 ***
## educ_hs                     9.142e-01  3.099e-02  29.502  < 2e-16 ***
## educ_somecoll               1.478e+00  3.486e-02  42.404  < 2e-16 ***
## educ_college                1.948e+00  3.669e-02  53.099  < 2e-16 ***
## educ_advdeg                 2.400e+00  4.334e-02  55.377  < 2e-16 ***
## OWNERSHP                    6.005e-01  1.808e-02  33.221  < 2e-16 ***
## MARSTmarried spouse absent -4.714e-01  5.248e-02  -8.981  < 2e-16 ***
## MARSTseparated             -2.224e-01  5.804e-02  -3.831 0.000127 ***
## MARSTdivorced               7.731e-03  3.770e-02   0.205 0.837514
## MARSTwidowed               -3.402e-01  9.406e-02  -3.617 0.000298 ***
## MARSTnever married         -3.715e-01  2.442e-02 -15.215  < 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: 71408  on 74934  degrees of freedom
## Residual deviance: 63757  on 74920  degrees of freedom
## AIC: 63787
##
## Number of Fisher Scoring iterations: 5
```

OWNERSHP and AGE powered 4, are significant factors in this model according to their p__value.

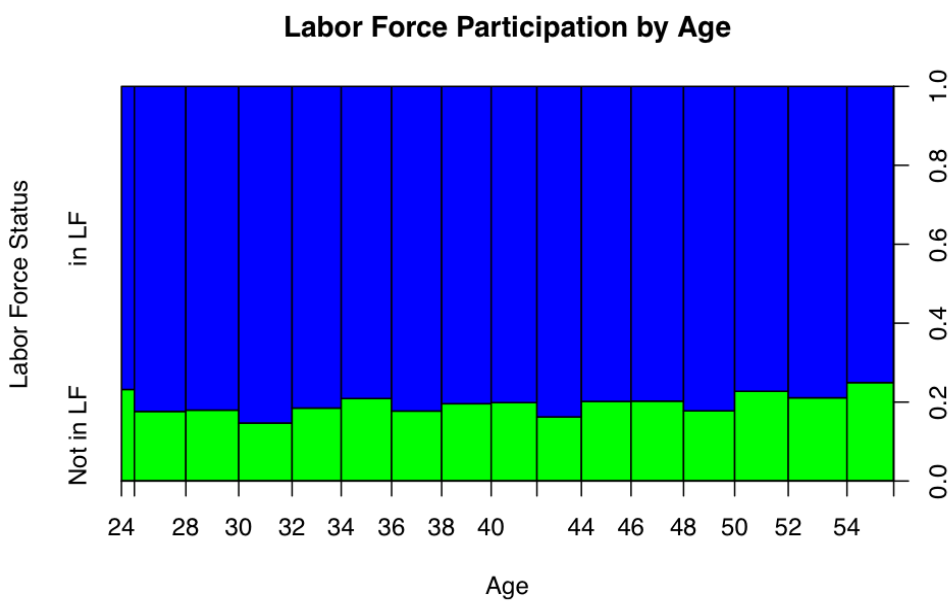
 # I am running a logistic regression on some other factors that are all binary.

```r
model_logit4 <- glm(LABFORCE ~ AGE + I(AGE^4) + female
                    + educ_advdeg + OWNERSHP + white
                    + MARST,
                    family = binomial, data = dat_use1)
summary(model_logit4)
```

```
##
## Call:
## glm(formula = LABFORCE ~ AGE + I(AGE^4) + female + educ_advdeg +
##     OWNERSHP + white + MARST, family = binomial, data = dat_use1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7698   0.3589   0.5392   0.6893   1.3605
##
```

```
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.209e+00  1.387e-01   8.722  < 2e-16 ***
## AGE                        -6.859e-03  4.379e-03  -1.566    0.117
## I(AGE^4)                   -1.610e-08  1.416e-08  -1.137    0.256
## female                     -6.423e-01  2.014e-02 -31.894  < 2e-16 ***
## educ_advdeg                 1.177e+00  3.518e-02  33.461  < 2e-16 ***
## OWNERSHP                    6.259e-01  1.796e-02  34.842  < 2e-16 ***
## white                       3.695e-01  2.062e-02  17.925  < 2e-16 ***
## MARSTmarried spouse absent -6.003e-01  5.074e-02 -11.832  < 2e-16 ***
## MARSTseparated             -3.794e-01  5.652e-02  -6.714 1.90e-11 ***
## MARSTdivorced              -5.906e-02  3.692e-02  -1.600    0.110
## MARSTwidowed               -5.086e-01  9.108e-02  -5.584 2.35e-08 ***
## MARSTnever married         -4.863e-01  2.336e-02 -20.819  < 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: 71408  on 74934  degrees of freedom
## Residual deviance: 66906  on 74923  degrees of freedom
## AIC: 66930
##
## Number of Fisher Scoring iterations: 5
```

so, as it is obvious in the summary of our model, other variable named 'white' which is a binary one, is statistically significant in my model.



Probit/Logit estimation

```
model_logit1 <- glm(LABFORCE ~ AGE + I(AGE^2) + female + AfAm + Asian + race_oth + Hispanic
                    + educ_hs + educ_somecoll + educ_college + educ_advdeg
                    + MARST,
                    family = binomial, data = dat_use1)
summary(model_logit1)
```

```
##
## Call:
## glm(formula = LABFORCE ~ AGE + I(AGE^2) + female + AfAm + Asian +
##     race_oth + Hispanic + educ_hs + educ_somecoll + educ_college +
##     educ_advdeg + MARST, family = binomial, data = dat_use1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6277   0.3476   0.4862   0.6459   1.5245
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.6023215  0.2445543   2.463  0.01378 *
## AGE                       0.0171486  0.0121072   1.416  0.15666
## I(AGE^2)                 -0.0003149  0.0001471  -2.141  0.03228 *
## female                   -0.6839386  0.0205171 -33.335  < 2e-16 ***
## AfAm                     -0.1906696  0.0282354  -6.753 1.45e-11 ***
## Asian                    -0.1112229  0.0374503  -2.970  0.00298 **
## race_oth                 -0.0781864  0.0332004  -2.355  0.01852 *
## Hispanic                  0.1653724  0.0313524   5.275 1.33e-07 ***
## educ_hs                   0.8972780  0.0310196  28.926  < 2e-16 ***
## educ_somecoll             1.4531782  0.0350710  41.435  < 2e-16 ***
## educ_college              1.9430903  0.0370924  52.385  < 2e-16 ***
## educ_advdeg               2.3676171  0.0437358  54.135  < 2e-16 ***
## MARSTmarried spouse absent -0.5222011  0.0517449 -10.092  < 2e-16 ***
## MARSTseparated           -0.1240651  0.0577062  -2.150  0.03156 *
## MARSTdivorced             0.0619381  0.0375785   1.648  0.09930 .
## MARSTwidowed             -0.3023247  0.0934446  -3.235  0.00121 **
## MARSTnever married       -0.3857612  0.0241093 -16.000  < 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: 71408  on 74934  degrees of freedom
## Residual deviance: 64847  on 74918  degrees of freedom
## AIC: 64881
##
## Number of Fisher Scoring iterations: 5
```

```
regn_probit1 <- glm(LABFORCE ~ AGE + female + AfAm + Asian
                    + Amindian + race_oth + Hispanic + educ_hs + educ_somecoll +
                    + educ_advdeg + MARST
                    , family = binomial (link = 'probit'), data = dat_use1)
summary(regn_probit1)
```

```
##
## Call:
## glm(formula = LABFORCE ~ AGE + female + AfAm + Asian + Amindian +
##     race_oth + Hispanic + educ_hs + educ_somecoll + +educ_advdeg +
##     MARST, family = binomial(link = "probit"), data = dat_use1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6843   0.3620   0.5522   0.6934   1.1720
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.5849218  0.0315982  50.159  < 2e-16 ***
## AGE                       -0.0088139  0.0006647 -13.260  < 2e-16 ***
## female                    -0.3511029  0.0111667 -31.442  < 2e-16 ***
## AfAm                      -0.1829718  0.0159076 -11.502  < 2e-16 ***
## Asian                     -0.1048498  0.0207108  -5.063 4.14e-07 ***
## Amindian                  -0.2241642  0.0782075  -2.866  0.00415 **
## race_oth                  -0.0787554  0.0185275  -4.251 2.13e-05 ***
## Hispanic                  -0.0706306  0.0169726  -4.161 3.16e-05 ***
## educ_hs                   -0.1760251  0.0133551 -13.180  < 2e-16 ***
## educ_somecoll              0.1238713  0.0155511   7.965 1.65e-15 ***
## educ_advdeg                0.5759419  0.0190694  30.202  < 2e-16 ***
## MARSTmarried spouse absent -0.3643299  0.0295827 -12.316  < 2e-16 ***
## MARSTseparated            -0.1334598  0.0325529  -4.100 4.14e-05 ***
## MARSTdivorced              0.0076166  0.0206250   0.369  0.71191
## MARSTwidowed              -0.2612986  0.0539526  -4.843 1.28e-06 ***
## MARSTnever married        -0.2641585  0.0131953 -20.019  < 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: 71408  on 74934  degrees of freedom
## Residual deviance: 67649  on 74919  degrees of freedom
## AIC: 67681
##
## Number of Fisher Scoring iterations: 5
```

In addition to looking at effects of particular X-variables, I am interested in looking at predictive accuracy

```
summary(model_logit1$fitted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3095  0.7569  0.8459  0.8166  0.9071  0.9686
```

```
summary(dat_use1$LABFORCE)
```

```
## Not in LF     in LF
##     13741     61194
```

```
pred_model_logit1 <- (model_logit1$fitted > 0.5)
table(pred_model_logit1, dat_use1$LABFORCE)
```

```
##
## pred_model_logit1 Not in LF in LF
##             FALSE       955   940
##             TRUE      12786 60254
```

```
frac_correct_l1a <- mean(as.numeric(as.numeric(pred_model_logit1) == dat_use1$LABFORCE))
pred_model_logit1b <- (model_logit1$fitted > mean(dat_use1$LABFORCE))
```

```
## Warning in mean.default(dat_use1$LABFORCE): argument is not numeric or
## logical: returning NA
```

```
table(pred_model_logit1b, dat_use1$LABFORCE)
```

```
## < table of extent 0 x 2 >
```

```
frac_correct_l1b <- mean(as.numeric(as.numeric(pred_model_logit1b) == dat_use1$LABFORCE))

# examine how different cut-off values change predictive accuracy
```

```
set.seed(11111)
index<-sample(x=2,size=nrow(dat_use1),replace=TRUE,prob=c(0.8,0.2))
train<-dat_use1[index==1,]
test<-dat_use1[index==2,]
dim(dat_use1)
```

```
## [1] 74935   110
```

```
model_train<-glm(LABFORCE ~AGE + I(AGE^2) +female + MORTGAGE+ AfAm + Asian + Hispanic
                + educ_hs + educ_somecoll + educ_college + educ_advdeg
                + MARST,
                family = binomial, data = train)

prob<-predict(object=model_train,newdata=test,type="response")
pred<-cbind(test,prob)
pred<-transform(pred,predict=ifelse(prob<=0.5,0,1))
ta<-table(pred$LABFORCE,pred$predict)
ta
```

```
##
##                  0     1
##   Not in LF   213  2554
##   in LF       183 12029
```

I used the model_train to predict and then Reclassified the predicted probability values. at the end, I compared the actual and predicted values of the model. Depending on the purpose of the model, false negatives and false positives could have different costs. according to the textbook, maximizing the likelihood of the probit model is one or two steps more complicated but not different conceptually. Having a likelihood function with a first and second derivative makes finding a maximum much easier than the random hunt.