Workshop 4 - Solution

Logistic regression I: Binary logistic regression

MSBX-5130: Customer Analytics

2/13/2020

1) Objectives & setup

- Workshop task: Estimate "demand" for a potential partner
- We will use online dating data on profile views for inference
  - Website users browse profiles of potential partners
  - After viewing, they decide whether or not to send the profile owner an email
  - Outcomes = send email (1) or not (0)
  - We observe certain characteristics of the profile owner and the "match" with browsing user
- Using these data we will demonstrate how to:
  - Estimate a binary logit model using `glm()`
  - Predict expected utilties for profiles and the probability of email contact
  - Calculate marginal effects (average effect on outcome probabilities)
- Here is the data description:

You have access to online dating profile viewing data. In total, we observe 160,000 profile views and associated outcomes (send email or not). The data are in the file `Online-Dating.RData` (the file is available on Canvas). The variables in the dataset are:

Variable	Description
`profile_gender`	Gender of person in profile, male or female
`first_contact`	1 = first-contact e-mail sent, 0 = otherwise
`age`	Age of the person in the profile, in years
`age_older`	1 = potential mate in profile is at least 5 years older
`age_younger`	1 = potential mate in profile is at least 5 years younger
`looks`	Numerical looks rating
`height`	Inches
`height_taller`	1 = potential mate at least 2 inches taller
`height_shorter`	1 = potential mate at least 2 inches shorter
`bmi`	Body mass index
`yrs_education`	Years of education
`educ_more`	1 = potential mate has at least 2 more years of education
`educ_less`	1 = potential mate has at least 2 years less of education
`income`	$1,000 annual income
`diff_ethnicity`	1 = potential mate has different ethnicity than browser

Workshop task workflow

1. Setup
    1. Download data & R Markdown file
    2. Import data
    3. Subset and summarize data
2. Model estimation and comparisom
    1. Simple logit model
    2. Logit model with all available regresors
3. Model prediction
    1. Baseline prediction - mean utilities (V)
        1. Using `predict()`
        2. Using matrix algebra
        3. Show equivalence of methods
    2. Baseline prediction - choice probabilties (Pr(first_contact=1))
        1. Using `predict()`
        2. Using predicted mean utilities
        3. Show equivalence of methods
4. Marginal effects
    1. Computation of marginal effects
        1. Using `maBina()`
        2. Using predicted expected utilities
    2. Application of marginal effects
        1. Average effect on email probability from 5% increase in `income`
        2. Average effect on email probability from 25% increase in `income`

1.1) Download data & R Markdown file

If you have not already done so, download the data file `Online-Dating.RData` from Canvas. Also download this R markdown file, `Workshop4.Rmd`.

Now launch RStudio, and change the working directory to where you have downloaded the previously mentioned files.

1.2) Read in the data from the `RData` file

Hint: We need to use `load()` here, not `read.csv()`

```r
load("Online-Dating.RData")
```

1.3) Subset and summarize data

As this dataset is large and to keep matters simple, for now we will limit our attention to male profiles – i.e., profiles predominantly browsed by women.

To prepare for model estimation on male profiles, choose the subset of data corresponding to `profile_gender` == "male". Also remove the column associated with `profile_gender`. Save the resulting dataframe as `men_DF`.

Hint: There are many ways to do this. One useful function to extract the male profiles is `subset()`.

```r
men_DF = subset(dating_DF, profile_gender == "male", select = -profile_gender)
```

1.3.1) Summarize the data

To sumarize the data, do the following:

- Print the first six rows
- Use `describe()` to summarize the moments of the data

```r
head(men_DF)
```

```
      first_contact age age_older age_younger      looks height height_taller
80001             0  43         0           1 -0.1435105   73.5             1
80002             1  38         1           0  0.6750283   69.5             1
80003             0  28         0           1 -0.3710585   67.5             0
80004             0  43         1           0 -0.1067023   71.5             1
80005             0  48         0           0 -0.4461543   73.5             1
80006             0  38         1           0  0.1363754   67.5             0
      height_shorter      bmi yrs_education educ_more educ_less income
80001              0 27.97816          12.5         0         1   87.5
80002              0 25.46970          16.0         1         0  125.0
80003              1 27.00137          16.0         0         0   42.5
80004              0 22.68962          16.0         1         0  125.0
80005              0 22.77292          16.0         0         0  275.0
80006              0 22.37256          12.5         0         1   62.5
      diff_ethnicity
80001              0
80002              0
80003              0
80004              0
80005              0
80006              0
```

```r
library(psych)
describe(men_DF)
```

```
               vars     n  mean    sd median trimmed   mad   min    max  range
first_contact     1 80000  0.07  0.26   0.00    0.00  0.00  0.00   1.00   1.00
age               2 80000 38.77  8.62  38.00   38.41  7.41 19.00  68.00  49.00
age_older         3 80000  0.40  0.49   0.00    0.37  0.00  0.00   1.00   1.00
age_younger       4 80000  0.26  0.44   0.00    0.21  0.00  0.00   1.00   1.00
looks             5 80000  0.00  0.55  -0.04   -0.03  0.49 -1.94   2.53   4.47
height            6 80000 70.96  2.64  71.50   70.99  2.97 61.00  85.00  24.00
height_taller     7 80000  0.90  0.31   1.00    0.99  0.00  0.00   1.00   1.00
height_shorter    8 80000  0.04  0.19   0.00    0.00  0.00  0.00   1.00   1.00
bmi               9 80000 25.46  2.56  25.44   25.36  2.11 12.37  39.34  26.97
yrs_education    10 80000 15.84  2.40  16.00   15.85  2.22  8.00  21.00  13.00
educ_more        11 80000  0.35  0.48   0.00    0.31  0.00  0.00   1.00   1.00
educ_less        12 80000  0.25  0.43   0.00    0.19  0.00  0.00   1.00   1.00
income           13 80000 92.85 55.18  87.50   83.98 37.06 10.00 275.00 265.00
diff_ethnicity   14 80000  0.06  0.23   0.00    0.00  0.00  0.00   1.00   1.00
```

```
                 skew kurtosis   se
first_contact   3.34      9.14 0.00
age             0.33     -0.30 0.03
age_older       0.42     -1.82 0.00
age_younger     1.07     -0.86 0.00
looks           0.42      0.68 0.00
height          0.01      0.52 0.01
height_taller  -2.59      4.73 0.00
height_shorter  5.01     23.09 0.00
bmi             0.35      2.29 0.01
yrs_education  -0.27      0.55 0.01
educ_more       0.65     -1.58 0.00
educ_less       1.14     -0.69 0.00
income          1.64      2.85 0.20
diff_ethnicity  3.88     13.06 0.00
```

Discussion:

- How many observation do we have for estimation?

We have 80,000 observations.

- Use the means of `age`, `height`, `yrs_education`, and `income` to characterize the average male profile on the dating site?

The average male on the site is 38.8 years old, 5'11" (71 inches) tall, is college educated (~16 yrs education), and makes $93k per year.

- What is average email contact rate?

From the mean of `first_contact`, we infer that roughly 7% of profile views result in email contact from women.

2) Model building and comparison

2.1) Estimate a simple model logit model with `glm()`

Let's first estimate and summarize (using `summary()`) a simple logit model of `first_contact` as the outcome. Include the following regressors: `age`, `looks`, `height`, `bmi`, `yrs_education`, `income` and `diff_ethnicity`. Name the result `logit1`.

```
logit1 = glm(first_contact ~ age+looks+height+bmi+yrs_education+income+diff_ethnicity,
             data = men_DF, family = binomial(link = "logit"))
summary(logit1)
```

```
Call:
glm(formula = first_contact ~ age + looks + height + bmi + yrs_education +
    income + diff_ethnicity, family = binomial(link = "logit"),
    data = men_DF)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6933  -0.4087  -0.3684  -0.3219   2.8409

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.8631902  0.4421529 -20.046  < 2e-16 ***
age             0.0135139  0.0019387   6.970 3.16e-12 ***
looks           0.5176348  0.0283955  18.229  < 2e-16 ***
height          0.0600265  0.0053840  11.149  < 2e-16 ***
bmi             0.0374126  0.0058499   6.395 1.60e-10 ***
yrs_education   0.0185771  0.0060659   3.063  0.00219 **
income          0.0025342  0.0002505  10.118  < 2e-16 ***
diff_ethnicity -0.4999068  0.0760170  -6.576 4.82e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41036  on 79999  degrees of freedom
Residual deviance: 40342  on 79992  degrees of freedom
AIC: 40358

Number of Fisher Scoring iterations: 5
```

Discussion:

- Interpret the regression coefficients.

Intercept:

- *The intercept can be interpreted as utility when other regressors = 0*
- *The intercept can also be interpreted as the log-odds of "success" (1st contact email received) when other regressors = 0*
- *–> Utility, log-odds(`first_contact`=1) = -8.8631902*

age:

- *Each +1 year in age increases utility, and log odds of email contact, by 0.0135139.*

looks:

- *Each +1 unit of looks rating increases utility/log-odds of email contact by 0.5176348.*

Other variables similar…

- What do coefficient estimates suggest about female preferences for men?

The coefficient estimates suggest that women on this site prefer (on average) older, better looking, taller, larger, more educated, more weathly men that are the same ethnicity as they are.

2.2) Estimate a complete logit model with `glm()`

Now, let's estimate and summarize a logit model of `first_contact` using all available regressors. Name the result `logit2`.

```
logit2 = glm(first_contact ~ .,
             data = men_DF, family = binomial(link = "logit"))
summary(logit2)
```

```
Call:
glm(formula = first_contact ~ ., family = binomial(link = "logit"),
    data = men_DF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7989  -0.4184  -0.3644  -0.3115   3.0393

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -7.5251886  0.4706216 -15.990  < 2e-16 ***
age             0.0150776  0.0020255   7.444 9.77e-14 ***
age_older      -0.2693740  0.0324052  -8.313  < 2e-16 ***
age_younger    -0.3078933  0.0364871  -8.438  < 2e-16 ***
looks           0.5236808  0.0285258  18.358  < 2e-16 ***
height          0.0417912  0.0058863   7.100 1.25e-12 ***
height_taller   0.3470403  0.0690557   5.026 5.02e-07 ***
height_shorter -0.3188366  0.1272128  -2.506   0.0122 *
bmi             0.0360348  0.0058656   6.143 8.07e-10 ***
yrs_education   0.0133780  0.0075301   1.777   0.0756 .
educ_more      -0.1812976  0.0338545  -5.355 8.55e-08 ***
educ_less      -0.2295514  0.0395019  -5.811 6.20e-09 ***
income          0.0025459  0.0002513  10.130  < 2e-16 ***
diff_ethnicity -0.4960341  0.0761361  -6.515 7.26e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41036  on 79999  degrees of freedom
Residual deviance: 40129  on 79986  degrees of freedom
AIC: 40157

Number of Fisher Scoring iterations: 6
```

Discussion:

- How do the additional variables (`age_older`, `age_younger`, `height_taller`, `height_shorter`, `educ_more`, `educ_less`) in this regression clarify female preferences for men?

The additional variables suggest women on the site prefer men who are roughly their same age, taller than they are, and who have approximately their same level of education.

- On basis of AIC, which model is preferred?

We prefer `logit2`, since the AIC is smaller.

3) Model prediction

3.1) Baseline prediction - mean utilities (V)

3.1.1 Using `predict()`

Use the `predict()` function to calculate the expected (mean) utilties, using the estimates from model `logit2`.

```
logit2.pred.V1 = predict(logit2, type = "link")       # utility
```

3.1.2 Using matrix algebra

Now use matrix algebra to calculate the expected (mean) utilties, using the estimates from model `logit2`.

```
X = model.matrix(logit2)
logit2.pred.V2 = as.numeric(X %*% logit2$coefficients)
```

3.1.3 Show equivalence of methods

Use `all.equal()` to test whether your two prediction algorithms obtain the same values.

```
all.equal(logit2.pred.V1,logit2.pred.V2,check.names=FALSE)
```

```
[1] TRUE
```

3.2) Baseline prediction - choice probabilties (Pr(first_contact=1))

3.2.1 Using `predict()`

Use the `predict()` function to calculate the outcome choice probabilites, using the estimates from model `logit2`.

After doing the prediction, compute and print the mean value of the predictions, and compare it to the mean value of the outcome in the estimation data.

```
logit2.pred.p1 = predict(logit2, type = "response")  # choice probability
mean(logit2.pred.p1)
```

```
[1] 0.0711
```

```
mean(men_DF$first_contact)
```

```
[1] 0.0711
```

Discussion:

- Do our predictions do a good job of matching the email rate in the data?

Yes, the mean values are match exactly (to 4 decimal places).

3.2.2 Using predicted mean utilities

Use the predicted utility values from 3.1.1 to calculate the outcome choice probabilites.

Hint: Recall the logit formula $p_{i1} = \frac{e^{V_{i1}}}{1+e^{V_{i1}}}$

```
logit2.pred.p2 = exp(logit2.pred.V1)/(1+exp(logit2.pred.V1))
```

3.2.3 Show equivalence of methods

Use `all.equal()` to test whether your two prediction algorithms obtain the same values.

```
all.equal(logit2.pred.p1,logit2.pred.p2,check.names=FALSE)
```

```
[1] TRUE
```

4) Marginal effects

4.1) Computation of marginal effects

4.1.1) Using `maBina()`

Use `maBina()` from the `erer` package to estimate average marginal effects, by averaging over all observation-level marginal effects.

Hint: use x.mean = FALSE in your call to maBina()

```
library(erer)
logit2 = glm(first_contact ~ .,
             data = men_DF, family = binomial(link = "logit"), x = TRUE)
logit2.me1 = maBina(logit2, x.mean = FALSE, digits = 6)
logit2.me1
```

```
                  effect    error    t.value  p.value
(Intercept)    -0.491205 0.030399 -16.158449 0.000000
age             0.000984 0.000132   7.475458 0.000000
age_older      -0.016271 0.001910  -8.516871 0.000000
age_younger    -0.017896 0.001989  -8.998352 0.000000
looks           0.034183 0.001825  18.734088 0.000000
height          0.002728 0.000383   7.121283 0.000000
height_taller   0.019068 0.003350   5.691873 0.000000
height_shorter -0.017346 0.006045  -2.869600 0.004111
bmi             0.002352 0.000382   6.161249 0.000000
yrs_education   0.000873 0.000491   1.777149 0.075548
```

```
educ_more      -0.010938 0.001993  -5.488605 0.000000
educ_less      -0.013511 0.002211  -6.111024 0.000000
income          0.000166 0.000016  10.175925 0.000000
diff_ethnicity -0.025411 0.003176  -8.002155 0.000000
```

Discussion:

- Interpret the marginal effect estimates.

In general, the marginal effect for a coninuous regressor will be the effect on Pr(`first_contact`=1) from a one unit change in the regressor of interest. Even though `maBina` reports a marginal effect for the intercept, we typically do not bother interpreting it, since we are interested in effects on regressors that change.

Age:

- *Holding other factors constant, a +1 unit change in age increases Pr(`first_contact`=1) (on average, and approximately) by 0.000984.*

Other variables similar…

4.1.2) Using predicted choice probabilities

Demonstrate that you can get the same marginal effect for `income` by computing the observation-level marginal income effects and then averaging over all observations.

Hint: Recall the formula for an observation-level marginal effect: $m.e. = \beta_{income} p_{ik}(1 - p_{ik})$, where β_{income} is the coefficient on income from model `logit2` and p_{ik} is the predicted choice probability (of "success") for the observation.

```r
logit2.me2.income = mean(logit2.pred.p1*(1-logit2.pred.p1)*logit2$coefficients["income"])
logit2.me2.income
```

```
[1] 0.0001661823
```

4.2) Application of marginal effects

4.2.1) Average effect on email probability from 5% increase in `income`

a) Using the marginal effects calculated in 4.1.2, evaluate (and print) the average (approximate) change in email probability resulting from a 5% increase in `income`. I.e., use the marginal effect estimate for each observation to calculate the approximate change in email probability from a 5% increase in `income`, then average over all observations and report the result.

b) Evaluate (and print) the average (exact) change in email probability resulting from a 5% increase in `income`. I.e., calculate the exact probability change for each observation, average across all observations and report the result

c) Compute and report the root-mean-square differences in (a) and (b). I.e., compute differences for each observation, square the differences, average over all observations, and report the square root of the result.

```r
# a
del = 0.05
phat1 = logit2.me2.income*(del*men_DF$income)
mean(phat1)
```

```
[1] 0.000771518
```

```r
# b
pred_DF = men_DF
pred_DF$income = pred_DF$income*(1 + del)
phat2 = predict(logit2, newdata = pred_DF, type = "response") - logit2.pred.p1
mean(phat2)
```

```
[1] 0.0008433556
```

```r
# c
sqrt(mean((phat1-phat2)^2))
```

```
[1] 0.0003915891
```

Discussion:

- Interpret (in words) the result from part a

On average, a 5% increase in `income` increases the probabiltiy of receiving an email by 0.000771518, or 0.08% – i.e., not very much

Note the prediction using marginal effects is quite close to the exact values, with an average deviation of 0.00039, or 0.04%

4.2.2) Average effect on email probability from 25% increase in `income`

a) Using the marginal effects calculated in 4.1.2, evaluate (and print) the average (approximate) change in email probability resulting from a 25% increase in `income`.
b) Evaluate (and print) the average (exact) change in email probability resulting from a 25% increase in `income`.
c) Compute and report the square root of the average of the squared differences in (a) and (b).

```r
del = 0.25
phat1 = logit2.me2.income*(del*men_DF$income)
mean(phat1)
```

```
[1] 0.00385759
```

```r
pred_DF = men_DF
pred_DF$income = pred_DF$income*(1 + del)
phat2 = predict(logit2, newdata = pred_DF, type = "response") - logit2.pred.p1
mean(phat2)
```

```
[1] 0.004336912
```

```r
sqrt(mean((phat1-phat2)^2))
```

```
[1] 0.002120876
```

Discussion:

- What happens to the quality of the marginal effect approximation (to the change in email probability) as the change in income increases?

As expected, marginal effect approximation (to the change in email probability) gets worse as we evaluate larger changes in income. We see this from the root-mean-square measure of deviation between the marginal effect approximation and the exact predicted probabilities – this measure is higher when evaluating a 25% income increase (vs. a 5% increase).