MSBX-5310 (Customer Analyitcs): Assignment 4 - Solution

due: 2/18/19

1) Setup

Instructions

1. Due date: Feb 18th 2019, 8:00 AM.
2. Your task is to fill in all R code blocks that currently contain "#TBD" comments. Similarly, insert text responses wherever you see *TBD* in the markdown file.
3. PLEASE UNCOMMENT LINE 2 AND ADD YOUR NAME

Homework task description

- Objective: 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) (`first_contact`)
  - 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)
  - Simulate outcomes from the model
- Here is the data description:

You have access to online dating profile viewing data. In total, we observe 160,000 profile views and assocaited 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

Homework task workflow

1. Setup
   1. Download data & R Markdown file
   2. Import data
   3. Subset and summarize data
2. Model estimation and comparison
   1. Simple logit model
   2. Logit model with all available regressors
3. Prediction exercises
   1. Baseline predictions
      1. Mean utilities (V)
      2. Choice probabilties (Pr(first_contact=1))
   2. Counterfactual (out of sample) predictions
      1. 10% increase in income
      2. +1k income
      3. income = 25k
      4. income = 250k
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 `age`
      2. Average effect on email probability from 25% increase in `age`

1.1) Download data & R Markdown file

If you have not already done so, download the data file `Online-Dating.RData` from Canvas (available in the Session 4 module).

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 into a dataframe called `dating_DF`:

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

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

1.3) Subset and summarize data

For the homework, we focus our attention on *female* profiles – i.e., profiles predominantly browsed by men (this is NOT the data we analyzed in class).

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

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

```
women_DF = subset(dating_DF, profile_gender == "female", 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(women_DF)
```

```
  first_contact age age_older age_younger      looks height height_taller
1             0  38         0           1  0.2424462   65.5             0
2             0  33         0           1  0.5562732   65.5             0
3             1  48         0           1  0.7468254   67.5             0
4             0  28         1           0  0.8164707   67.5             0
5             0  43         0           1 -0.7042280   63.5             0
6             0  38         0           0  0.0368911   65.5             0
  height_shorter      bmi yrs_education educ_more educ_less income
1              1 25.39829          18.0         0         0   42.5
2              1 18.84389          16.0         0         0   62.5
3              1 20.82963          16.0         0         0   62.5
4              1 16.20082          16.0         0         0   62.5
5              1 23.53649          18.0         1         0   62.5
6              1 27.03689          12.5         0         1   30.0
  diff_ethnicity
1              0
2              0
3              0
4              0
5              0
6              1
```

```r
library(psych)
```

```
Warning: package 'psych' was built under R version 3.5.2
```

```r
describe(women_DF)
```

```
               vars     n  mean    sd median trimmed   mad   min    max
first_contact     1 80000  0.09  0.29   0.00    0.00  0.00  0.00   1.00
age               2 80000 35.52  8.70  33.00   35.33  7.41 19.00  73.00
age_older         3 80000  0.23  0.42   0.00    0.16  0.00  0.00   1.00
age_younger       4 80000  0.54  0.50   1.00    0.55  0.00  0.00   1.00
looks             5 80000  0.28  0.66   0.25    0.25  0.67 -1.49   3.14
height            6 80000 65.29  2.65  65.50   65.29  2.97 59.00  73.50
height_taller     7 80000  0.03  0.17   0.00    0.00  0.00  0.00   1.00
height_shorter    8 80000  0.90  0.30   1.00    1.00  0.00  0.00   1.00
bmi               9 80000 22.44  3.49  21.79   22.06  2.70 16.20  46.29
yrs_education    10 80000 15.36  2.40  16.00   15.48  2.97  8.00  21.00
educ_more        11 80000  0.28  0.45   0.00    0.23  0.00  0.00   1.00
educ_less        12 80000  0.34  0.47   0.00    0.30  0.00  0.00   1.00
income           13 80000 53.82 31.71  42.50   50.75 29.65 10.00 275.00
diff_ethnicity   14 80000  0.11  0.31   0.00    0.01  0.00  0.00   1.00
                range  skew kurtosis   se
first_contact    1.00  2.85     6.10 0.00
age             54.00  0.23    -0.39 0.03
age_older        1.00  1.31    -0.28 0.00
age_younger      1.00 -0.15    -1.98 0.00
```

```
looks             4.63  0.52     0.68 0.00
height           14.50 -0.02    -0.42 0.01
height_taller     1.00  5.50    28.20 0.00
height_shorter    1.00 -2.73     5.46 0.00
bmi              30.09  2.05     8.63 0.01
yrs_education    13.00 -0.54     0.34 0.01
educ_more         1.00  0.96    -1.08 0.00
educ_less         1.00  0.68    -1.53 0.00
income          265.00  2.84    15.21 0.11
diff_ethnicity    1.00  2.49     4.22 0.00
```

Discussion:

- How many observation do we have for estimation?

We have 80,000 observations

- What is average email contact rate? How does this rate compare to the contact rate for male profiles (the data we anaylzed in class)?

The average email contact rate is 0.09, or 9% of profile views. The average email contact rate for male profiles is 0.07, or 2% lower than for female profiles.

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`.

```r
logit1 = glm(first_contact ~ age+looks+height+bmi+yrs_education+income+diff_ethnicity,
             data = women_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 = women_DF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7818  -0.4674  -0.4178  -0.3513   3.0080

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.6673035  0.3602062   1.853   0.0639 .
age             0.0038012  0.0017152   2.216   0.0267 *
looks           0.4813364  0.0214320  22.459  < 2e-16 ***
height         -0.0269719  0.0047958  -5.624 1.87e-08 ***
bmi            -0.0614107  0.0044540 -13.788  < 2e-16 ***
yrs_education  -0.0102398  0.0052654  -1.945   0.0518 .
income          0.0003332  0.0003791   0.879   0.3795
diff_ethnicity -0.2554616  0.0420739  -6.072 1.27e-09 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48752  on 79999  degrees of freedom
Residual deviance: 47696  on 79992  degrees of freedom
AIC: 47712

Number of Fisher Scoring iterations: 5
```

Discussion:

- Interpret the regression coefficients.

The intercept of 0.667 is the expected utility for option 1 (send email), relative to option 0 (don't send email, with expected utility of 0), when all other regessors are zero. This also implies the log-odds of sending an email when when all other regessors are zero is 0.667.

The age coefficient of 0.0038 implies a one year increase in age increases utility of sending an email by 0.0038. This also implies the log-odds of sending an email increase by 0.0038 for each year of the age of the profile holder increases.

Other variables have similar interpretation to age (all are continuous variables)

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`.

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

```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8318  -0.4696  -0.4157  -0.3480   3.0594

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.0517899  0.4060820   0.128 0.898517
age             0.0061920  0.0018682   3.314 0.000918 ***
age_older      -0.1792139  0.0384492  -4.661 3.15e-06 ***
age_younger    -0.0734181  0.0304971  -2.407 0.016067 *
looks           0.4837548  0.0214545  22.548  < 2e-16 ***
height         -0.0153335  0.0052140  -2.941 0.003273 **
height_taller  -0.5151478  0.1042932  -4.939 7.84e-07 ***
height_shorter  0.1107209  0.0541539   2.045 0.040898 *
bmi            -0.0620012  0.0044635 -13.891  < 2e-16 ***
yrs_education  -0.0204628  0.0065302  -3.134 0.001727 **
educ_more      -0.0611685  0.0318745  -1.919 0.054979 .
```

```
educ_less      -0.1431284  0.0329015  -4.350 1.36e-05 ***
income          0.0002289  0.0003797   0.603 0.546576    
diff_ethnicity -0.2465653  0.0421174  -5.854 4.79e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48752  on 79999  degrees of freedom
Residual deviance: 47604  on 79986  degrees of freedom
AIC: 47632

Number of Fisher Scoring iterations: 5
```

Discussion:

- On basis of AIC, which model is preferred? Why?

We prefer logit2 to logit1, as the AIC value is lower.

- Compare the estimates from `logit2` here with those obtained (using the same model) from data on male profiles (the data we analyzed in class). For example, are the signs of the parameter estimates the same?

Compared to the comparable regression in class, we see that income, ethinicity, and looks have similar impact for men and women. Height, bmi and years education have opposite sign effects.

3) Prediction exercises

3.1) Baseline predictions

3.1.1) Mean utilities (V)

Using whatever method you prefer, predict expected (mean) utilties, using the estimates from model `logit2`. Call the results `logit2.pred.V1`. Report (print) the mean value of `logit2.pred.V1`.

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

```
[1] -2.375524
```

3.1.2 Choice probabilties (Pr(first_contact=1))

Using whatever method you prefer, predict choice probabilties (Pr(first_contact=1)), using the estimates from model `logit2`. Call the results `logit2.pred.p1`. Report (print) the mean value of `logit2.pred.p1`. Finally, compute the difference in the mean `logit2.pred.p1` and the mean of `first_contact` in the data.

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

```
[1] 0.0909375
```

```r
mean(logit2.pred.p1)-mean(women_DF$first_contact)
```

```
[1] 3.533715e-11
```

Discussion:

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

Yes, the predictions do an excellent job of matching the email rate in the data (the difference is between the sample average and the prediction is numerically tiny (<10^-10)).

3.2) Counterfactual (out of sample) predictions

3.2.1 10% increase in income

Using whatever method you prefer, predict choice probabilties (Pr(first_contact=1)), using the estimates from model `logit2`, such that income for every observation is increased by 10%. Call the results `logit2.pred.p2`. Report (print) the mean value of `logit2.pred.p2`.

Next, compute the difference in the aggregate (total) number of predicted emails using `logit2.pred.p2` and the aggregate (total) number of predicted emails using our baseline prediction, `logit2.pred.p1`.

```
predict_DF = women_DF
predict_DF$income = 1.1*women_DF$income
logit2.pred.p2 = predict(logit2, newdata = predict_DF, type = "response")  # choice probability
mean(logit2.pred.p2)
```

```
[1] 0.09104027
```

```
sum(logit2.pred.p2)-sum(logit2.pred.p1)
```

```
[1] 8.221348
```

```
(sum(logit2.pred.p2)-sum(logit2.pred.p1))/sum(logit2.pred.p1)
```

```
[1] 0.001130082
```

Discussion:

- Compared to the baseline prediction, a 10% increase in income leads to how many more emails (in total)? What is the increase (in emails) in percentage terms?

A 10% increase in income leads to slightly more than 8 additional emails, an increase of 0.11%

3.2.2 +1k income

Using whatever method you prefer, predict choice probabilties (Pr(first_contact=1)), using the estimates from model `logit2`, such that income for every observation is increased by $1000 (pay attention to units!). Call the results `logit2.pred.p3`. Report (print) the mean value of `logit2.pred.p3`.

Next, compute the difference in the aggregate (total) number of predicted emails using `logit2.pred.p3` and the aggregate (total) number of predicted emails using our baseline prediction, `logit2.pred.p1`.

```
predict_DF = women_DF
predict_DF$income = women_DF$income + 1
logit2.pred.p3 = predict(logit2, newdata = predict_DF, type = "response")  # choice probability
mean(logit2.pred.p3)
```

```
[1] 0.09095614
```

```
sum(logit2.pred.p3)-sum(logit2.pred.p1)
```

```
[1] 1.491483
```

```
(sum(logit2.pred.p3)-sum(logit2.pred.p1))/sum(logit2.pred.p1)
```

```
[1] 0.0002050148
```

Discussion:

- Compared to the baseline prediction, a uniform \$1000 increase in income leads to how many more emails (in total)? What is the increase (in emails) in percentage terms?

A uniform \$1000 increase in income leads to about 1.49 additional emails, an increase of 0.02%

3.2.3 income = 25k

Using whatever method you prefer, predict choice probabilties (Pr(first_contact=1)), using the estimates from model `logit2`, such that income for every observation is exactly \$25,000 (pay attention to units!). Call the results `logit2.pred.p4`. Report (print) the mean value of `logit2.pred.p4`.

Next, compute the difference in the aggregate (total) number of predicted emails using `logit2.pred.p4` and the aggregate (total) number of predicted emails using our baseline prediction, `logit2.pred.p1`.

```r
predict_DF = women_DF
predict_DF$income = 25
logit2.pred.p4 = predict(logit2, newdata = predict_DF, type = "response")  # choice probability
mean(logit2.pred.p4)
```

```
[1] 0.09037999
```

```r
sum(logit2.pred.p4)-sum(logit2.pred.p1)
```

```
[1] -44.60053
```

```r
(sum(logit2.pred.p4)-sum(logit2.pred.p1))/sum(logit2.pred.p1)
```

```
[1] -0.006130658
```

Discussion:

- Compared to the baseline prediction, setting income to \$25,000 for all profiles leads to how many more emails (in total)? What is the change (in emails) in percentage terms?

Setting income at \$25,000 for all profiles leads to a decrease in emails by 44.6, a 0.61% decrease.

3.2.4 income = 250k

Using whatever method you prefer, predict choice probabilties (Pr(first_contact=1)), using the estimates from model `logit2`, such that income for every observation is exactly \$250,000 (pay attention to units!). Call the results `logit2.pred.p5`. Report (print) the mean value of `logit2.pred.p5`.

Next, compute the difference in the aggregate (total) number of predicted emails using `logit2.pred.p5` and the aggregate (total) number of predicted emails using our baseline prediction, `logit2.pred.p1`.

```r
predict_DF = women_DF
predict_DF$income = 250
logit2.pred.p5 = predict(logit2, newdata = predict_DF, type = "response")  # choice probability
mean(logit2.pred.p5)
```

```
[1] 0.09463831
```

```r
sum(logit2.pred.p5)-sum(logit2.pred.p1)
```

```
[1] 296.0648
```

```r
(sum(logit2.pred.p5)-sum(logit2.pred.p1))/sum(logit2.pred.p1)
```

```
[1] 0.04069619
```

Discussion:

- Compared to the baseline prediction, setting income to $250,000 for all profiles leads to how many more emails (in total)? What is the change (in emails) in percentage terms?

Setting income at $250,000 for all profiles leads to an increase in emails by 296.1, a 4.1% decrease.

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. Use the estimates from model `logit2`.

```r
library(erer)
logit2 = glm(first_contact ~ .,
             data = women_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)     0.004217 0.033063   0.127542 0.898511
age             0.000504 0.000152   3.316970 0.000910
age_older      -0.013397 0.002758  -4.858476 0.000001
age_younger    -0.005728 0.002385  -2.401787 0.016317
looks           0.039389 0.001713  22.991071 0.000000
height         -0.001249 0.000424  -2.942775 0.003254
height_taller  -0.032879 0.005349  -6.146741 0.000000
height_shorter  0.008304 0.003911   2.123059 0.033752
bmi            -0.005048 0.000358 -14.095870 0.000000
yrs_education  -0.001666 0.000532  -3.134257 0.001724
educ_more      -0.004709 0.002427  -1.940356 0.052340
educ_less      -0.010932 0.002465  -4.435190 0.000009
income          0.000019 0.000031   0.602848 0.546611
diff_ethnicity -0.017730 0.002786  -6.363577 0.000000
```

Discussion:

- Interpret the marginal effect estimates.

Here, the marginal effect for a continuous regressor is the effect on Pr(first_contact=1) with a one unit difference in the regressor. We are generally only interested in regressors that change (not the intercept). When we consider age, holding other factors constant, a unit one change in age on average increases Pr(first_contact=1) by 0.000504. The other regressors follow a similar pattern.

4.1.2) Using predicted choice probabilities

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

Hint: Recall the formula for an observation-level marginal effect: $m.e. = \beta_{income}p(1-p)$, where β_{income} is the coefficient on income from model `logit2` and p is the probability of "success", $p = \frac{e^V}{1+e^V}$, and V is the deterministic portion of utility.

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

```
[1] 1.864184e-05
```

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 4.2.1, evaluate the average (approximate) change in email probability resulting from a 5% increase in `income`.

b) Evaluate the exact change in email probability resulting from a 5% increase in `income`.

c) Compute and report the square root of the average of the squared differences in (a) and (b).

Hint: Note that you can access marginal effect estimates in the `out` dataframe returned by `maBina()`. For example, if marginal effects are estimated as `me = maBina(...)`, the marginal effect point estimates are accessed as `me1$out[,1]`.

```
del = 0.05
phat1 = as.numeric(logit2.me1$out["income",1])*(del*women_DF$income)
pred_DF = women_DF
pred_DF$income = pred_DF$income*(1 + del)
phat2 = predict(logit2, newdata = pred_DF, type = "response") - logit2.pred.p1
sqrt(mean((phat1-phat2)^2))
```

```
[1] 2.511246e-05
```

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

a) Using the marginal effects calculated in 4.2.1, evaluate the average (approximate) change in email probability resulting from a 25% increase in `income`.

b) Evaluate the 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).

```
del = 0.25
phat1 = as.numeric(logit2.me1$out["income",1])*(del*women_DF$income)
pred_DF = women_DF
pred_DF$income = pred_DF$income*(1 + del)
phat2 = predict(logit2, newdata = pred_DF, type = "response") - logit2.pred.p1
sqrt(mean((phat1-phat2)^2))
```

```
[1] 0.0001260729
```

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).*