BBB

Yuhan Xu 474154; Shuyuan Tang 476399; Vanessa Liu 476144; Tian Wei 474157; Zhuoyuan He 473913

2/23/2020

```r
#load the data
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages --------------------------------------------- tidyv
erse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## -- Conflicts ------------------------------------------------ tidyverse_c
onflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(knitr)
library(statar)

## Warning: package 'statar' was built under R version 3.6.2

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.6.2

library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.6.2

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units

BBB=read.csv("BBB.csv")
```

Part I

```
#1.
BBB.lg=glm(buyer~last+total_+gender+child+youth+cook+do_it+refernce+art+geog,
           family = "binomial", data=BBB)

#2.
summary(BBB.lg)

##
## Call:
## glm(formula = buyer ~ last + total_ + gender + child + youth +
##     cook + do_it + refernce + art + geog, family = "binomial",
##     data = BBB)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4031  -0.4129  -0.2807  -0.1839   3.2650
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.3608301  0.0492961 -47.891  < 2e-16 ***
## last        -0.0947124  0.0027924 -33.918  < 2e-16 ***
## total_       0.0011160  0.0001982   5.630 1.80e-08 ***
## genderM      0.7607204  0.0357608  21.272  < 2e-16 ***
## child       -0.1862162  0.0172824 -10.775  < 2e-16 ***
## youth       -0.1129745  0.0261087  -4.327 1.51e-05 ***
## cook        -0.2703210  0.0171283 -15.782  < 2e-16 ***
## do_it       -0.5391648  0.0269657 -19.994  < 2e-16 ***
## refernce     0.2346876  0.0265583   8.837  < 2e-16 ***
## art          1.1555840  0.0221439  52.185  < 2e-16 ***
## geog         0.5742763  0.0186311  30.824  < 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: 30355  on 49999  degrees of freedom
## Residual deviance: 24122  on 49989  degrees of freedom
```

```
## AIC: 24144
##
## Number of Fisher Scoring iterations: 6
```

```r
library(margins)
```

```r
margins(BBB.lg)
```

```
## Average marginal effects
```

```
## glm(formula = buyer ~ last + total_ + gender + child + youth +     cook + do_it + refernce + art + geog, family = "binomial",     data = BBB)
```

```
##       last    total_    child     youth    cook   do_it refernce     art
##  -0.006411 7.555e-05 -0.01261 -0.007647 -0.0183 -0.0365  0.01589 0.07822
##     geog genderM
##  0.03887 0.05538
```

All the variables are statistically significant as all of the p value is less than 0.05;

All the variables except “total_” and “last” are managerially important.

According to the result, male has a higher probability to buy the book; People who bought reference book, art book and geography book before are more likely to buy “The art of history of Florence”; Customers who bought many children’s books, youth books, cook books or do-it-yourself books are less likely to buy this book.

When a customer purchased one more children’s book, he will have a 1.26% less probability of purchasing the ‘The Art History of Florence’.

When a customer purchased one more youth book, he will have a 0.76% less probability of purchasing the ‘The Art History of Florence’.

When a customer purchased one more cook book, he will have a 1.82% less probability of purchasing the ‘The Art History of Florence’.

When a customer purchased one more do-it-yourself book, he will have a 3.64% less probability of purchasing the ‘The Art History of Florence’.

When a customer purchased one more reference book, he will have a 1.58% more probability of purchasing the ‘The Art History of Florence’.

When a customer purchased one more art book, he will have a 7.82% more probability of purchasing the ‘The Art History of Florence’.

When a customer purchased one more geography book, he will have a 3.88% more probability of purchasing the ‘The Art History of Florence’.

When a customer is a male, he will have 5.53% more probability of purchasing than when a customer is a female.

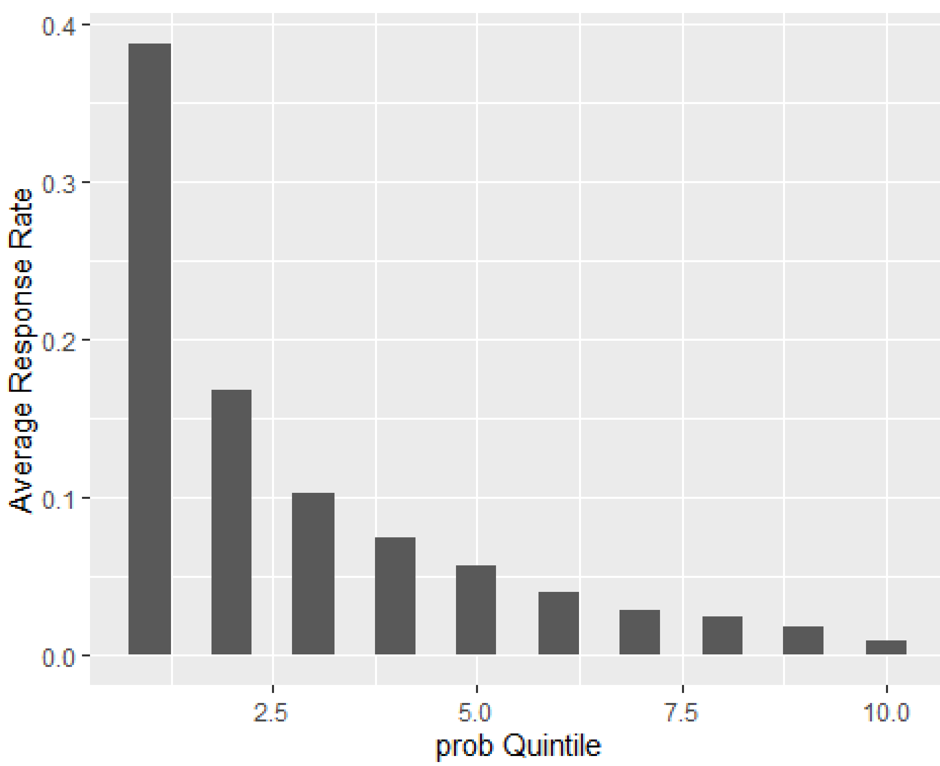
Part II

```
#1.
BBB$predict_prob=predict.glm(BBB.lg,data=BBB,type="response")
#predict_prob
BBB=BBB %>% mutate(prob_quin=xtile(predict_prob, 10))

BBB$prob_quin= max(BBB$prob_quin) + 1 - BBB$prob_quin

#2.
avg_resp_rate_prob=BBB %>% group_by(prob_quin) %>%
  summarise(avg_resp_rate=mean(buyer))

bar_avg_resp_rate_prob= ggplot(data=avg_resp_rate_prob,
         aes(x = prob_quin, y = avg_resp_rate)) +
  labs(x="prob Quintile", y="Average Response Rate") +
  geom_bar(stat="identity", width=0.5)
bar_avg_resp_rate_prob
```




```
#3.
table=BBB %>% group_by(prob_quin) %>%
  summarise(num_customers=n(),num_buyers = sum(buyer),avg_resp_rate=mean(buye
r))
table

## # A tibble: 10 x 4
##    prob_quin num_customers num_buyers avg_resp_rate
```

```
##        <dbl>         <int>      <int>         <dbl>
##  1         1          5000       1935         0.387
##  2         2          5000        836         0.167
##  3         3          5000        511         0.102
##  4         4          5000        368         0.0736
##  5         5          5000        284         0.0568
##  6         6          5000        196         0.0392
##  7         7          4998        139         0.0278
##  8         8          5002        121         0.0242
##  9         9          5000         90         0.018
## 10        10          5000         42         0.0084
```

```r
#4.
BBB.lg2=glm(buyer~child,
           family = "binomial", data=BBB)
summary(BBB.lg2)
```

```
##
## Call:
## glm(formula = buyer ~ child, family = "binomial", data = BBB)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5578  -0.4371  -0.4219  -0.4219   2.2197
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.37445    0.01988 -119.412  < 2e-16 ***
## child        0.07406    0.01321    5.606 2.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30355  on 49999  degrees of freedom
## Residual deviance: 30325  on 49998  degrees of freedom
## AIC: 30329
##
## Number of Fisher Scoring iterations: 5
```

```r
BBB.lg3=glm(buyer~child+gender+art,
           family = "binomial", data=BBB)
summary(BBB.lg3)
```

```
##
## Call:
## glm(formula = buyer ~ child + gender + art, family = "binomial",
##     data = BBB)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.3384  -0.4058  -0.3115  -0.2892   2.8040
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -3.00184    0.02726 -110.123   <2e-16 ***
## child       -0.15161    0.01520   -9.972   <2e-16 ***
## genderM      0.54622    0.03307   16.517   <2e-16 ***
## art          1.02450    0.01922   53.291   <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: 30355  on 49999  degrees of freedom
## Residual deviance: 27099  on 49996  degrees of freedom
## AIC: 27107
##
## Number of Fisher Scoring iterations: 5
```

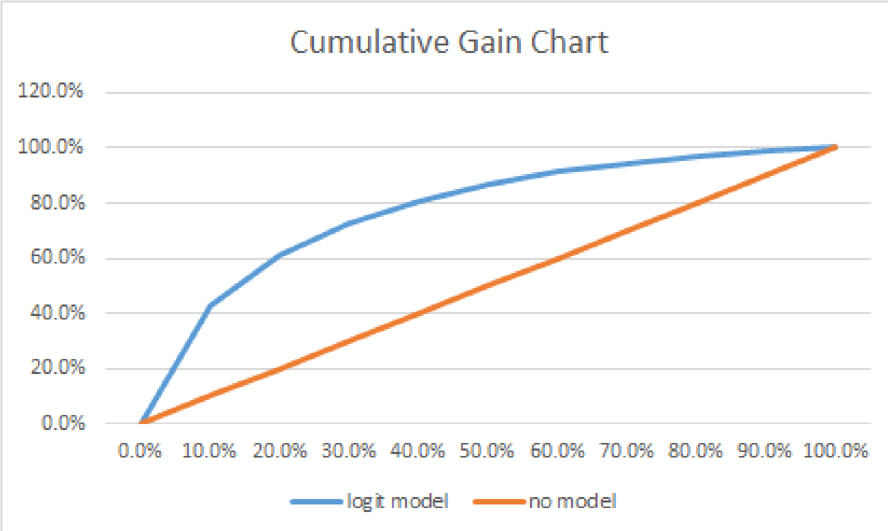
The coefficient for "child" is: 0.07406. When the logistic regression only includes the child variable, the coefficient of child is 0.074, which is positive, but when the logistic regression includes both the child variable and art variable, the coefficient of child variable is -0.151, which is negative. From the fact, We conclude that the reason why the coefficient of child becomes positive is that when the logistic regression only includes child variable, the effect of child to purchase probability also includes the effect of art books purchased variable, which has high positive correlation with child variable.

Part III

1.

prob_quin	num_customers	num_buyers	avg_resp_rate	cum buyers	gains	cum gains
1	5000	1935	0.387	1935	42.8%	42.8%
2	5000	836	0.1672	2771	18.5%	61.3%
3	5000	511	0.1022	3282	11.3%	72.6%
4	5000	368	0.0736	3650	8.1%	80.7%
5	5000	284	0.0568	3934	6.3%	87.0%
6	5000	196	0.0392	4130	4.3%	91.3%
7	4998	139	0.02781112	4269	3.1%	94.4%
8	5002	121	0.02419032	4390	2.7%	97.1%
9	5000	90	0.018	4480	2.0%	99.1%
10	5000	42	0.0084	4522	0.9%	100.0%
		4522				

2.



Part IV

```
#1.
profit=18-9-3
breakeven_rate=0.5/profit
breakeven_rate

## [1] 0.08333333

#2.
BBB$target=ifelse(BBB$predict_prob>=breakeven_rate,1,0)

#3.
table2=BBB %>% group_by(target) %>%
  summarise(num_customers=n()*10,num_buyers = sum(buyer)*10,avg_resp_rate=mea
n(buyer))

table2

## # A tibble: 2 x 4
##   target num_customers num_buyers avg_resp_rate
##    <dbl>         <dbl>      <dbl>         <dbl>
## 1      0        344400      11990        0.0348
## 2      1        155600      33230        0.214

#4.
profit=(18-9-3)*33230
mailing_cost=0.5*155600
expected_profit=profit-mailing_cost
expected_roi=expected_profit/mailing_cost
```

```
expected_profit

## [1] 121580

expected_roi

## [1] 1.562725
```

The expected profit is $121580 and the expected return on marketing expenditures is 156.3%.