Multivariate Analysis

LOGISTIC REGRESSION

By Suckwon Hong

Question 1. Open bank data and check what is in it.

a) Conduct logistic regression with "y" as the dependent variable. The predictors are age, marital status, education, default, balance, housing, and campaign. Convert education into a continuous variable.
b) Show the full results (copy and paste the results)
Answer:

```
Call:
glm(formula = y ~ age + marital + education + default + balance +
    housing + campaign, family = binomial, data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8648  -0.5396  -0.4489  -0.3731   3.0833

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.269e+00  3.223e-01  -7.040 1.92e-12 ***
age             1.557e-02  4.846e-03   3.212  0.00132 **
maritalmarried -3.864e-01  1.408e-01  -2.744  0.00606 **
maritalsingle   1.058e-01  1.630e-01   0.649  0.51617
education       1.325e-01  6.254e-02   2.118  0.03414 *
defaultyes      2.248e-02  3.630e-01   0.062  0.95062
balance         7.228e-06  1.430e-05   0.505  0.61321
housingyes     -5.718e-01  9.760e-02  -5.858 4.68e-09 ***
campaign       -9.909e-02  2.358e-02  -4.203 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3231.0  on 4520  degrees of freedom
Residual deviance: 3128.1  on 4512  degrees of freedom
AIC: 3146.1
```

Figure 1. Logistic regression results

c) What does it say about the goodness-of-fit?
Answer: To test the goodness-of-fit, we test whether the model with predictors fits significantly better than a null model with only an intercept. Here, the difference between the residual deviance for the model with predictors and null model is used as the test statistic. The statistic is distributed chi-squared where degrees of freedom is equal to the differences in degrees of freedom between two models. As a result, we found that the chi-square of 102.9122 with 8 degrees of freedom and an associated p-value of less than 0.001 (1.082951e-18). The result tells us that our model with predictors fits significantly better than an empty model with only an intercept.

d) What is the R^2?
Answer: McFadden's R^2 is equal to 3.185151e-02. And other pseudo-R^2 metrics are as following figure.

llh	llhNull	G2	McFadden	r2ML	r2CU
-1.564044e+03	-1.615500e+03	1.029122e+02	3.185151e-02	2.250603e-02	4.407382e-02

e) What is the logit estimate for age?
Answer: As shown in the regression result in b), the logit estimate for age is equal to 1.557e-02.

f) What is the odds ratio for age?
Answer: It can be ca1culated by exponentialize the logit estimate. Thus, the odds ratio for age is $e^{1.557e-02} = 1.0156869$.

g) Interpret the results for marital status, default, and campaign.
Answer:
Marital status: The ratio of the odds of being y=1 in condition (marital status = married) compared to the odds of being y=1 in condition (marital status= divorced) is $e^{3.864e-01} = 1.471673$.
The ratio of the odds of being y=1 in condition (marital status = single) compared to the odds of being y=1 in condition (marital status = divorced) is $e^{1.058e-01} = 1.1116$.

Default: The odds ratio for default equals to $e^{2.248e-02} = 1.0227$ which means that the odds for people who don't default are about 2.3% higher than the odds for those who are default.

Campaign: A one unit increase in campaign, the odds of being y=1 versus being y=0 increase by a factor of $e^{-9.909e-02} = 0.9057$.

h) Add “previous” to the model. Conduct likelihood test. What is your conclusion?
Answer: The results of regression model including *previous* are as follows.

```
Call:
glm(formula = y ~ age + marital + education + default + balance +
    housing + campaign + previous, family = binomial, data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9826  -0.5339  -0.4382  -0.3579   3.0646

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.305e+00  3.243e-01  -7.106 1.20e-12 ***
age             1.496e-02  4.863e-03   3.077 0.002092 **
maritalmarried -4.110e-01  1.419e-01  -2.897 0.003762 **
maritalsingle   5.804e-02  1.643e-01   0.353 0.723942
education       1.268e-01  6.315e-02   2.008 0.044597 *
defaultyes      8.403e-02  3.640e-01   0.231 0.817439
balance         5.225e-06  1.465e-05   0.357 0.721298
housingyes     -6.120e-01  9.861e-02  -6.206 5.42e-10 ***
campaign       -9.101e-02  2.356e-02  -3.863 0.000112 ***
previous        1.447e-01  2.101e-02   6.886 5.73e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3231.0  on 4520  degrees of freedom
Residual deviance: 3084.5  on 4511  degrees of freedom
AIC: 3104.5
```

Figure 2. Logistic regression results (including *previous*)

After, we conducted likelihood ratio test between two logistic models: model from *question a)* and model from *question h)*. As the result shows that the p-value is very small (4.023e-11), we can say that additional factor “*previous”* is significant.

```
Likelihood ratio test

Model 1: y ~ age + marital + education + default + balance + housing +
    campaign
Model 2: y ~ age + marital + education + default + balance + housing +
    campaign + previous
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -1564.0
2  10 -1542.2  1 43.603  4.023e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3. Likelihood ratio test between two models

i) What is the predicted probability of getting a “yes” (I am talking about the dependent variables “yes”) when default is “y”? (Hint. Use prediction function)
Answer: A total of 76 observations has default value of “y”. Predicted probability of them getting a “yes” is as following:

Predicted probability							
0.09410937	0.07499228	0.17711814	0.15370925	0.19964262	0.1176313	0.0749009	0.07886766
0.12844649	0.11596127	0.11298447	0.12725646	0.05407521	0.10185936	0.10646441	0.09932073
0.27986597	0.13507406	0.11068105	0.07308442	0.21993145	0.07049271	0.16311302	0.06330338
0.08550796	0.13565824	0.0457028	0.20625909	0.18201226	0.08990404	0.03771337	0.13674839
0.10670271	0.1265805	0.05890437	0.10274125	0.05536187	0.16141442	0.07784031	0.15617277
0.10407823	0.1921741	0.06294853	0.09554673	0.13681797	0.17280999	0.18707467	0.10729261
0.17920558	0.06283866	0.20742641	0.06981738	0.22474368	0.06172352	0.17475942	
0.14456561	0.1107827	0.177204	0.08240743	0.06629508	0.1188054	0.10179793	
0.08088583	0.06061776	0.15090925	0.08522844	0.08720045	0.15484595	0.08304065	
0.13248179	0.0672499	0.08745586	0.10230282	0.12837128	0.10322827	0.1369558	

Question 2. Here is a logistic regression model result for predicting voting=1 that included conservativeness, age, and political knowledge for model 1. Independent variables are continuous. Here is the SAS output.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	514.289	411.208
SC	518.229	426.969
-2 Log L	512.289	403.208

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	-6.3896	1.4976	18.2045	<.0001
conserv	1	0.0266	0.00894	8.8442	0.0029
age	1	-0.0208	0.0188	1.2351	0.2664
knowledge	1	1.0790	0.1611	44.8373	<.0001

a) Write the resulting logistic regression equation
Answer:

$$\log\left(\frac{\Pr\,(Voting = 1)}{1 - \Pr\,(Voting = 1)}\right) = -6.3896 + 0.0266 * conservativeness - 0.0208 * age + 1.0790 * political\ knowledge$$

b) What is the predicted probability of voting=1 for 69 year old person with a conservativeness level of 10 and a knowledge score of 5?
Answer:

$$\log\left(\frac{\Pr(Voting = 1)}{1 - \Pr(Voting = 1)}\right) = -6.3896 + 0.0266 * 10 - 0.0208 * 69 + 1.0790 * 5 = -2.1638$$

$$\Pr(Voting = 1) = e^{-2.1638}/(1 + e^{-2.1638}) = 0.1030$$

c) Interpret all outcomes related to intercept, age, and knowledge.
Answer:
Intercept: Log odds ratio of 'voting 1' versus 'voting 0' for those whose *conservativeness* level is 0, knowledge score of 0, and age of 0 is -6.3896.

Age: For every unit increase in *Age,* the log odds of 'voting 1' versus 'voting 0' decrease by 0.0208.

Knowledge: For every unit increase in *Knowledge,* the log odds of 'voting 1' versus 'voting 0' increase by 1.0790.

d) Calculate the odds ratio and 95% confidence interval for conservativeness.
Answer:

$$Odds\ ratio(conservativenes) = e^{0.0266} = 1.0270$$

$$95\%\ Confidence\ interval\ for\ conservativeness$$

$$= [e^{0.0266-1.96*0.00894}, e^{0.0266+1.96*0.00894}]$$

$$= [1.009119, 1.04511]$$

Question 3. The following table contains information on restaurant closure and city.

Restaurant closure	City	Count
Yes	Ulsan	146
Yes	Busan	90
No	Ulsan	48
No	Busan	16

a) What are the odds that Ulsan would experience a restaurant closure?
Answer: the odds that Ulsan would experience a restaurant closure is

$$Odds(Ulsan) = \frac{\Pr(Restaurant\ closure = Yes|City = Ulsan)}{\Pr\ (Restaurant\ closure = No|City = Ulsan)} = \frac{\frac{146}{146+48}}{\frac{48}{146+48}} = 3.042$$

b) What are the odds that Busan would experience a restaurant closure?
Answer: the odds that Busan would experience a restaurant closure is

$$Odds(Busan) = \frac{\Pr(Restaurant\ closure = Yes|City = Busan)}{\Pr\ (Restaurant\ closure = No|City = Busan)} = \frac{\frac{90}{90+16}}{\frac{16}{90+16}} = 5.625$$

c) What is the odds ratio? That is, compared to Busan, what is the odds ratio that Ulsan would experience a restaurant closure?
Answer: the odds ratios is as follow:

$$Odds\ ratio = \frac{Odds(Ulsan)}{Odds(Busan)} = 0.565$$

Thus, for Ulsan, the odds of being 'restaurant closure' are 0.565 times as small than the odds for Busan being 'restaurant closure'.

d) Write down the logistic regression equation. Make sure you have defined the coefficients in the equation.
Answer:

$$\log\left(\frac{\Pr\ (Restaurant\ closure = Yes)}{1 - \Pr\ (Restaurant\ closure = Yes)}\right) = \beta_0 + \beta_1 City, where\ City = \begin{cases} 1\ if\ Ulsan \\ 0\ if\ Busan \end{cases}$$

R code

```
###Multivariate analysis hw4
rm(list = ls())
install.packages("pscl")
install.packages("lmtest")
library(pscl)
library(stats)
library(lmtest)
bank <- read.table("bank.txt",sep=";",header = T)
bank$education<-as.numeric(bank$education)

# #data split
# smp_size <- floor(0.75 * nrow(bank))
# set.seed(123)
# train_ind <- sample(seq_len(nrow(bank)), size = smp_size)
#
# train <- bank[train_ind, ]
# test <- bank[-train_ind, ]


#model fitting
mod1.logit <- glm(y~ age+marital+education+default+balance+housing+campaign,
                  family = binomial,data = bank)
summary(mod1.logit)

#goodness of fit
with(mod1.logit, null.deviance - deviance)
with(mod1.logit, df.null - df.residual)
with(mod1.logit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

#pseudo R (McFAdden)
pR2(mod1.logit)

#logit-estimate
coef(mod1.logit)
#odd-ratio
exp(coef(mod1.logit))

#likelihood test
mod2.logit <- glm(y ~ age+marital+education+default+balance+housing+campaign+previous,
                  family = binomial,data = bank)
summary(mod2.logit)

lrtest(mod1.logit,mod2.logit)
```