UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Lecture 8 Logistic Regression II

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Outline

Multiple Logistic Regression

Case-control sampling and logistic regression

Multiclass Logistic Regression

Summary and Remark

Logistic Regression with indicator variable

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable `Student` coded as (Student $=1$, Non-student $=0$).

```
> glm.fit=glm(default~student,data=defaultData,family=binomial)
> summary(glm.fit)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.50413    0.07071  -49.55  < 2e-16 ***
factor(student)Yes   0.40489    0.11502    3.52 0.000431 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

- β_1 is positive. This indicates students tend to have higher default probabilities than non-students.

```
> predict(glm.fit, list(student = c('Yes','No')),type="response")
         1          2
0.04313859 0.02919501
```

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Multiple logistic Regression

- Logistic Regression with several covariates

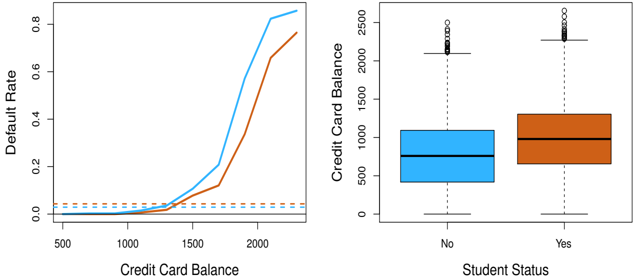
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

```
> glm.fit=glm(default~balance+income+student,data=defaultData,fam
> summary(glm.fit)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

- Why is coefficient for `student` negative, while it was positive before?

Confounding



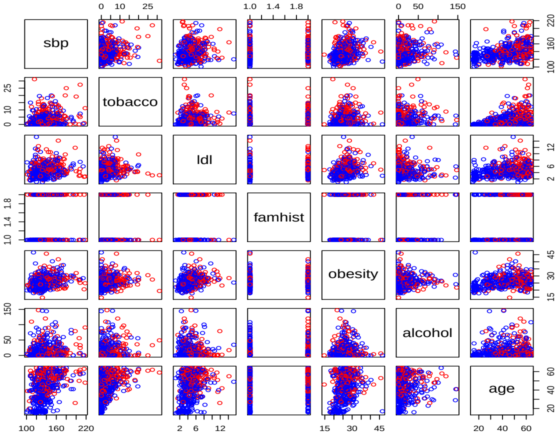
- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

South African Heart Disease

- 160 cases of MI (myocardial infarction) and 302 controls (all male in age range $15-64$), from Western Cape, South Africa in early 80s.
- Overall prevalence very high in this region: 5.1%.
- Measurements on seven predictors (risk factors), shown in scatterplot matrix.
- Goal is to identify relative strengths and directions of risk factors.
- This was part of an intervention study aimed at educating the public on healthier diets.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Pair Plots



Scatterplot matrix of the South African Heart Disease data. The response is color coded. The cases (MI) are red, the controls blue. `famhist` is a binary variable, with 1 indicating family history of MI.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

South African Heart Disease

```
> glm.fit=glm(chd~.,data=heartData1,family=binomial)
>
> summary(glm.fit)

Call:
glm(formula = chd ~ ., family = binomial, data = heartData1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7517  -0.8378  -0.4552   0.9292   2.4434

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp             0.0057607  0.0056326   1.023  0.30643
tobacco         0.0795256  0.0262150   3.034  0.00242 **
ldl             0.1847793  0.0574115   3.219  0.00129 **
famhistPresent  0.9391855  0.2248691   4.177 2.96e-05 ***
obesity        -0.0345434  0.0291053  -1.187  0.23529
alcohol         0.0006065  0.0044550   0.136  0.89171
age             0.0425412  0.0101749   4.181 2.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 483.17  on 454  degrees of freedom
AIC: 499.17
```

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

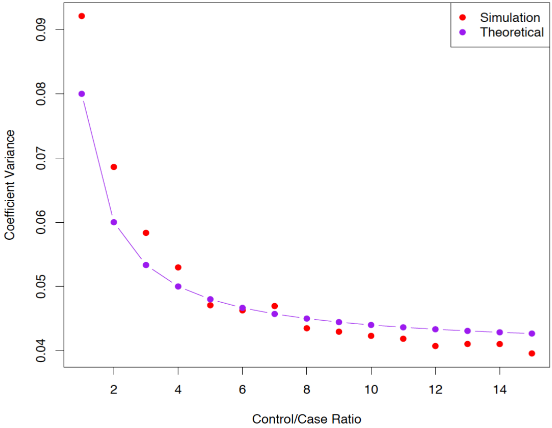
Case-control sampling and logistic regression

- In South African data, there are 160 cases, 302 controls — $\tilde{\pi} = 0.35$ are cases. Yet the prevalence of MI in this region is $\pi = 0.05$.
- With case-control samples, we can estimate the regression parameters β_j accurately (if our model is correct); the constant term β_0 is incorrect.
- We can correct the estimated intercept by a simple transformation

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log\left(\frac{\pi}{1-\pi}\right) - \log\left(\frac{\tilde{\pi}}{1-\tilde{\pi}}\right).$$

- Often cases are rare and we take them all; up to five times that number of controls is sufficient.

Diminishing returns in unbalanced binary data



Sampling more controls than cases reduces the variance of the parameter estimates. But after a ratio of about 5 to 1 the variance reduction flattens out.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Logistic regression with more than two classes

- So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes.
- One version (used in the `R` package `glmnet` or `nnet`) has the symmetric form.

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}}{\sum_{k=1}^{K}\left(e^{\beta_{0k}+\beta_{1k}X_1+\cdots+\beta_{pk}X_p}\right)}.$$

- Here there is a linear function for each class.
- only $K-1$ linear functions are needed as in 2-class logistic regression.
- Multiclass logistic regression is also referred to as multinomial regression.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Simulated Example

```
> library(nnet)
> x=matrix(rnorm(100*5),100,5)
>
> y=rnorm(100)
> #multinomial
> g4=sample(1:4,100,replace=TRUE)
> fit3=multinom(g4~x)
# weights:  28 (18 variable)
initial  value 138.629436
iter  10 value 130.910132
iter  20 value 130.869074
final  value 130.868827
converged
> summary(fit3)
Call:
multinom(formula = g4 ~ x)

Coefficients:
  (Intercept)        x1          x2         x3        x4        x5
2 -0.09206107 0.7771141 -0.07521353 0.48808850 0.3695944 0.4197601
3  0.19450922 0.1198007  0.21709470 0.27615848 0.2629457 0.1542603
4 -0.14379965 0.2477509 -0.29897262 0.01837793 0.2444425 0.2160098

Std. Errors:
  (Intercept)        x1        x2        x3        x4        x5
2   0.3219613 0.3626664 0.3128671 0.3311019 0.3513568 0.3064434
3   0.2923117 0.3150980 0.2862657 0.3169007 0.3333487 0.2906013
4   0.3133926 0.3645049 0.3230845 0.3515986 0.3620371 0.3227726

Residual Deviance: 261.7377
AIC: 297.7377
>
```

Summary and Remark

- Multiple Logistic regression
- Case-control sampling and logistic regression
- Multiclass logistic regression
- Read textbook Chapter 4 and `R` code
- Do `R` lab