데이터분석방법론 2 - 제 9 강 과제

성 명: 지 용 기

연 락 처: 010-9828-0332, braveji@hanmail.net

학 번: 201685-010100

6.8 은 미세 세포 폐암의 치료에 대한 임상시험 결과이다. 환자들은 임의적으로 두 개의 치료군으로 배정되었다. 하나의 각 치료 사이클에서 동일한 조합의 화학요법 치료제를 받도록 하는 순차적 치료군이고, 다른 치료군은 3 개의 서로 다른 화학요법 치료제를 치료 기간마다 바꾸어 가면서 받도록 하는 호환적 치료군이다.

<표 6.17> 폐암 치료에 관한 문제 6.8의 자료

치료군	성별	화학요법에 대한 반응			
		점점 악화	변화없음	부분적 회복	완전 회복
순차적치료군	남성	28	45	29	26
	여성	4	12	5	2
호환 치료군	남성	41	44	20	20
	여성	12	7	3	1

출처 : Holtbrugge, W. and Schmacher, M., *Appl, Statist.*, **40** : 249-259, 1991

```
#install.packages('VGAM')
library(VGAM)

## Warning: package 'VGAM' was built under R version 3.3.2

## Loading required package: stats4

## Loading required package: splines
```

a. 치료군과 성별의 주효과를 갖는 누적 로짓 모형을 적합하고 추정된 효과를 해석하라.

```
lung_cancer.raw <- read.csv( 'chap09_report_01.csv', header=TRUE )
lung_cancer.raw
```

```
##   treatment gender y1 y2 y3 y4
## 1         1      1 28 45 29 26
## 2         1      0  4 12  5  2
## 3         0      1 41 44 20 20
## 4         0      0 12  7  3  1
```

```
summary( lung_cancer.raw )
```

```
##    treatment       gender          y1              y2
##  Min.   :0.0   Min.   :0.0   Min.   : 4.00   Min.   : 7.00
##  1st Qu.:0.0   1st Qu.:0.0   1st Qu.:10.00   1st Qu.:10.75
##  Median :0.5   Median :0.5   Median :20.00   Median :28.00
##  Mean   :0.5   Mean   :0.5   Mean   :21.25   Mean   :27.00
##  3rd Qu.:1.0   3rd Qu.:1.0   3rd Qu.:31.25   3rd Qu.:44.25
##  Max.   :1.0   Max.   :1.0   Max.   :41.00   Max.   :45.00
##        y3              y4
##  Min.   : 3.00   Min.   : 1.00
##  1st Qu.: 4.50   1st Qu.: 1.75
##  Median :12.50   Median :11.00
##  Mean   :14.25   Mean   :12.25
##  3rd Qu.:22.25   3rd Qu.:21.50
##  Max.   :29.00   Max.   :26.00
```

```
lung_cancer <- lung_cancer.raw
lung_cancer$treatment <- factor( lung_cancer$treatment, levels=c(1, 0), label
s=c('순차적', '호환') )
lung_cancer$gender <- factor( lung_cancer$gender, levels=c(1, 0), labels=c('
남성', '여성') )
lung_cancer
```

```
##   treatment gender y1 y2 y3 y4
## 1    순차적   남성 28 45 29 26
## 2    순차적   여성  4 12  5  2
## 3      호환   남성 41 44 20 20
## 4      호환   여성 12  7  3  1
```

```
summary( lung_cancer )
```

```
##   treatment  gender        y1              y2              y3
##  순차적:2   남성:2   Min.   : 4.00   Min.   : 7.00   Min.   : 3.00
##  호환  :2   여성:2   1st Qu.:10.00   1st Qu.:10.75   1st Qu.: 4.50
##                     Median :20.00   Median :28.00   Median :12.50
##                     Mean   :21.25   Mean   :27.00   Mean   :14.25
##                     3rd Qu.:31.25   3rd Qu.:44.25   3rd Qu.:22.25
##                     Max.   :41.00   Max.   :45.00   Max.   :29.00
##        y4
```

```
##  Min.   : 1.00
##  1st Qu.: 1.75
##  Median :11.00
##  Mean   :12.25
##  3rd Qu.:21.50
##  Max.   :26.00
```

```
lung_cancer.fit <- vglm( cbind(y1, y2, y3, y4) ~ treatment + gender, family =
 cumulative(parallel = TRUE), data=lung_cancer  )
summary(lung_cancer.fit)
```

```
##
## Call:
## vglm(formula = cbind(y1, y2, y3, y4) ~ treatment + gender, family = cumula
tive(parallel = TRUE),
##     data = lung_cancer)
##
## Pearson residuals:
##   logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3])
## 1         0.1720        0.06056         0.2809
## 2        -1.6543        0.16312         0.8425
## 3         0.2655       -0.13909        -0.9386
## 4         0.6174       -0.01519         0.6444
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -1.3180     0.1801  -7.319  2.5e-13 ***
## (Intercept):2   0.2492     0.1621   1.538  0.12412
## (Intercept):3   1.3001     0.1852   7.021  2.2e-12 ***
## treatment 호환   0.5807     0.2119   2.741  0.00613 **
## gender 여성      0.5414     0.2953   1.834  0.06671 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Dispersion Parameter for cumulative family:   1
##
## Residual deviance: 5.5677 on 7 degrees of freedom
##
## Log-likelihood: -25.5417 on 7 degrees of freedom
##
## Number of iterations: 5
##
## Exponentiated coefficients:
```

```
## treatment 호환     gender 여성
##       1.787262       1.718403
```

tremtment 의 효과는 아주 유효(p 값: 0.00613)하고, gender 의 효과는 유효하지 않지만(p 값: .06671), 무시할 수 없는 수준으로 나옴.

treamtment 가 호환 치료군은 순처적 치료군에 비해서 1.787262 배 만큼 오즈비가 높음. 여성을때 남성에 비해서 오즈비가 1.718403 배가 높음.

b. 치료군과 성별의 교호작용항을 포함하는 모형을 적합하라. 추정된 치료군의 효과가 성별에 따라 어떻게 다르게 나타나는지 보여서 교호작용항을 해석하라.

```
lung_cancer.fit2 <- vglm( cbind(y1, y2, y3, y4) ~ treatment * gender, family
= cumulative(parallel = TRUE), data=lung_cancer  )
summary(lung_cancer.fit2)
```

```
##
## Call:
## vglm(formula = cbind(y1, y2, y3, y4) ~ treatment * gender, family = cumula
tive(parallel = TRUE),
##     data = lung_cancer)
##
## Pearson residuals:
##   logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3])
## 1        0.02471       -0.13434         0.1387
## 2       -1.30855        0.56305         1.0374
## 3        0.48236        0.01789        -0.7993
## 4        0.05105       -0.39051         0.4948
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept):1              -1.2757     0.1843  -6.920 4.52e-12 ***
## (Intercept):2               0.2957     0.1681   1.760   0.0785 .
## (Intercept):3               1.3452     0.1909   7.045 1.85e-12 ***
## treatment 호환              0.4881     0.2288   2.133   0.0329 *
## gender 여성                 0.2742     0.4094   0.670   0.5030
## treatment 호환:gender 여성  0.5904     0.5935   0.995   0.3199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
```

```
## Dispersion Parameter for cumulative family:   1
##
## Residual deviance: 4.5209 on 6 degrees of freedom
##
## Log-likelihood: -25.0183 on 6 degrees of freedom
##
## Number of iterations: 5
##
## Exponentiated coefficients:
##            treatment 호환                gender 여성 treatment 호환:gender 여
성
##                 1.629170                   1.315475                  1.804727
```

trement 가 호환이고, 여성일때 그렇지 않을때보다 오즈비가 1.8047 높음.

c. 교호작용을 포함한 모형이 더 유의하게 좋은 적합 결과를 보이는가?

treament 효과는 유효(p-value : 0.0329)하지만, gender 와 treatment:gender 교호작용 효과는 유효하지 않음.

6.11 의 직업만족도가 반응변수인 자료를 참조하여 누적 로짓 모형을 이용하여 자료를 분석하라.

<표 6.12> 성별을 제어한 상태에서 직업만족도와 수입

성별	수입	직업만족도			
		매우 불만족	약간 만족	적절히 만족	매우 만족
여성	<5,000	1	3	11	2
	5,000-15,000	2	3	17	3
	15,000-25,000	0	1	8	5
	>25,000	0	2	4	2
남성	<5,000	1	1	2	1
	5,000-15,000	0	3	5	1
	15,000-25,000	0	0	7	3
	>25,000	0	1	9	6

출처 : General Society Survey, 1991

a. 점수 {3, 10, 20, 35}를 이용하여 수입 효과를 분석하라.

```
job.raw <- read.csv( 'chap09_report_02.csv', header=TRUE )
job.raw
```

```
##   gender income y1 y2 y3 y4
## 1      0      3  1  3 11  2
## 2      0     10  2  3 17  3
## 3      0     20  0  1  8  5
```

```
## 4      0     34  0  2  4  2
## 5      1      3  1  1  2  1
## 6      1     10  0  3  5  1
## 7      1     20  0  0  7  3
## 8      1     34  0  1  9  6
```

```
summary( job.raw )
```

```
##      gender        income            y1            y2
##  Min.   :0.0   Min.   : 3.00   Min.   :0.0   Min.   :0.00
##  1st Qu.:0.0   1st Qu.: 8.25   1st Qu.:0.0   1st Qu.:1.00
##  Median :0.5   Median :15.00   Median :0.0   Median :1.50
##  Mean   :0.5   Mean   :16.75   Mean   :0.5   Mean   :1.75
##  3rd Qu.:1.0   3rd Qu.:23.50   3rd Qu.:1.0   3rd Qu.:3.00
##  Max.   :1.0   Max.   :34.00   Max.   :2.0   Max.   :3.00
##        y3              y4
##  Min.   : 2.000   Min.   :1.000
##  1st Qu.: 4.750   1st Qu.:1.750
##  Median : 7.500   Median :2.500
##  Mean   : 7.875   Mean   :2.875
##  3rd Qu.: 9.500   3rd Qu.:3.500
##  Max.   :17.000   Max.   :6.000
```

```
job <- job.raw
job$gender <- factor( job$gender, levels=c(1, 0), labels=c('남성', '여성') )
job
```

```
##   gender income y1 y2 y3 y4
## 1   여성      3  1  3 11  2
## 2   여성     10  2  3 17  3
## 3   여성     20  0  1  8  5
## 4   여성     34  0  2  4  2
## 5   남성      3  1  1  2  1
## 6   남성     10  0  3  5  1
## 7   남성     20  0  0  7  3
## 8   남성     34  0  1  9  6
```

```
summary( job )
```

```
##   gender      income            y1            y2            y3
##  남성:4   Min.   : 3.00   Min.   :0.0   Min.   :0.00   Min.   : 2.000
##  여성:4   1st Qu.: 8.25   1st Qu.:0.0   1st Qu.:1.00   1st Qu.: 4.750
##           Median :15.00   Median :0.0   Median :1.50   Median : 7.500
##           Mean   :16.75   Mean   :0.5   Mean   :1.75   Mean   : 7.875
##           3rd Qu.:23.50   3rd Qu.:1.0   3rd Qu.:3.00   3rd Qu.: 9.500
```

```
##           Max.   :34.00   Max.   :2.0   Max.   :3.00   Max.   :17.000
##        y4
##  Min.   :1.000
##  1st Qu.:1.750
##  Median :2.500
##  Mean   :2.875
##  3rd Qu.:3.500
##  Max.   :6.000
```

```
job.fit <- vglm( cbind(y1, y2, y3, y4) ~ gender + income , family = cumulativ
e(parallel = TRUE), data=job  )
summary(job.fit)
```

```
##
## Call:
## vglm(formula = cbind(y1, y2, y3, y4) ~ gender + income, family = cumulativ
e(parallel = TRUE),
##     data = job)
##
## Pearson residuals:
##                       Min      1Q   Median     3Q    Max
## logit(P[Y<=1]) -0.8576 -0.5856 -0.43308 0.1950 1.2534
## logit(P[Y<=2]) -1.1912 -0.4510 -0.19756 0.6399 2.0529
## logit(P[Y<=3]) -0.9765 -0.3619  0.08328 0.3917 0.5841
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -2.58174    0.66984  -3.854 0.000116 ***
## (Intercept):2 -0.89513    0.51314  -1.744 0.081083 .
## (Intercept):3  2.08063    0.55769   3.731 0.000191 ***
## gender 여성     0.02121    0.42729   0.050 0.960416
## income        -0.04651    0.01918  -2.425 0.015292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Dispersion Parameter for cumulative family:   1
##
## Residual deviance: 13.8058 on 19 degrees of freedom
##
## Log-likelihood: -27.9827 on 19 degrees of freedom
##
## Number of iterations: 5
##
## Exponentiated coefficients:
```

```
## gender 여성      income
##  1.0214339  0.9545526
```

gender 은 직업만족도를 영향이 없고, income 은 한단위 증가할때마다 오즈비가 0.954 만큼 낮아짐.

b. 위에서 추정된 수입 효과와 "매우 불만족"과 "약간 만족"의 두 범주를 합한 후에 추정된 값을 비교하라. 이 결과에서 모형의 어떤 특징이 반영되는가?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
job_new <- mutate( job, y0 = y1 + y2 )
job_new
```

```
##   gender income y1 y2 y3 y4 y0
## 1   여성      3  1  3 11  2  4
## 2   여성     10  2  3 17  3  5
## 3   여성     20  0  1  8  5  1
## 4   여성     34  0  2  4  2  2
## 5   남성      3  1  1  2  1  2
## 6   남성     10  0  3  5  1  3
## 7   남성     20  0  0  7  3  0
## 8   남성     34  0  1  9  6  1
```

```
job_new.fit <- vglm( cbind(y0, y3, y4) ~ gender + income , family = cumulativ
e(parallel = TRUE), data=job_new  )
summary(job_new.fit)
```

```
##
## Call:
## vglm(formula = cbind(y0, y3, y4) ~ gender + income, family = cumulative(pa
rallel = TRUE),
##     data = job_new)
##
```

```
## Pearson residuals:
##                     Min      1Q   Median     3Q   Max
## logit(P[Y<=1]) -1.2458 -0.3680 -0.19183 0.8368 1.745
## logit(P[Y<=2]) -0.9859 -0.3572  0.07609 0.3756 0.618
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -0.91589    0.51480  -1.779 0.075222 .
## (Intercept):2  2.06146    0.55774   3.696 0.000219 ***
## gender 여성     0.02008    0.42847   0.047 0.962615
## income        -0.04555    0.01918  -2.375 0.017572 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Dispersion Parameter for cumulative family:   1
##
## Residual deviance: 9.0544 on 12 degrees of freedom
##
## Log-likelihood: -22.9884 on 12 degrees of freedom
##
## Number of iterations: 5
##
## Exponentiated coefficients:
## gender 여성     income
##  1.0202865  0.9554674
```

gender 은 직업만족도를 영향이 없고, income 은 한단위 증가할때마다 오즈비가 0.955 만큼 낮아짐. 거의 차이가 없음.

c. 문제 (a)의 모형에서 성별변수를 제거할 수 있는가?

gender 의 효과는 p-value 가 0.962 로 전혀 없기 때문에 제거 할 수 있음.