> rm(list = ls())

> setwd("E:/Data of R")

>

> #Question 2

> library(VGAM)

> data2=read.csv("GSS.csv",header = T)

>

> names(data2)

[1] "gender" "race" "democrat" "republican" "independent"

>

> data2

gender race democrat republican independent

1 male white 132 176 127

2 male black 42 6 12

3 female white 172 129 130

4 female black 56 4 15

>

> #(a)

>

> model21=vglm(cbind(democrat,republican,independent)~gender+race,data = data2,

+ family = multinomial)

> model21

Call:

vglm(formula = cbind(democrat, republican, independent) ~ gender +

race, family = multinomial, data = data2)

Coefficients:

(Intercept):1 (Intercept):2 gendermale:1 gendermale:2 racewhite:1

1.3882465 -1.1771027 -0.2201865 0.3525732 -1.1182884

racewhite:2

1.1598459

Degrees of Freedom: 8 Total; 2 Residual

Residual deviance: 0.1982117

Log-likelihood: -20.17843

This is a multinomial logit model with 3 levels

> summary(model21)

Call:

vglm(formula = cbind(democrat, republican, independent) ~ gender +

race, family = multinomial, data = data2)

Pearson residuals:

log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])

1 -0.07696 -0.05743

2 0.19896 0.22498

3 0.07201 0.05961

4 -0.18480 -0.23505

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept):1 1.3882 0.2296 6.045 1.49e-09 \*\*\*

(Intercept):2 -1.1771 0.3807 -3.092 0.00199 \*\*

gendermale:1 -0.2202 0.1583 -1.391 0.16412

gendermale:2 0.3526 0.1651 2.136 0.03271 \*

racewhite:1 -1.1183 0.2335 -4.789 1.68e-06 \*\*\*

racewhite:2 1.1598 0.3801 3.051 0.00228 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors: 2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 0.1982 on 2 degrees of freedom

Log-likelihood: -20.1784 on 2 degrees of freedom

Number of iterations: 3

No Hauck-Donner effect found in any of the estimates

Reference group is level 3 of the response

>

> #(b)

> deviance(model21)

[1] 0.1982117

> qchisq(0.05,4\*2-3\*2,lower.tail = FALSE)

[1] 5.991465

>

> #0.1982117<5.991465

> #We can not reject the null/reduced model in favor of the saturated model,

> #indicating the model fit with main effects for all the predictors provides

> #a reasonable fit.

>

> #(c)

> coefficients(model21)[3]

gendermale:1

-0.2201865

> 1/exp(coefficients(model21)[3])

gendermale:1

1.246309

> #The estimated coefficient for the Gender Male dummy in the Democrat vs. Independent

> #is -0.2202. This means that men are 1.246309 times less likely choose Democrat

> #over Independent.

>

> coefficients(model21)[4]

gendermale:2

0.3525732

> exp(coefficients(model21)[4])

gendermale:2

1.422724

> #The estimated coefficient for the Gender Male dummy in the Republican vs. Independent

> #is 0.3525732. This means that men are 1.422724 times more likely choose Republican

> #over Independent.

>

> #Gender effect is not significant overall. Gender has a statistically significant

> #effect when comparing Republican over Independent.

>

> #(d)

>

> #Prob. being Independent for black females

> pi3=1/(1+exp(1.3882465)+exp(-1.1771027))

> pi3

[1] 0.1881118

>

> #Prob. being Democrat for black females

> pi1=exp(1.3882465)\*pi3

> pi1

[1] 0.7539177

>

> #(e)

> #Intercept1 is >0, which means that pi1/pi3>1, pi1>pi3

> #Intercept2 is <0, which means that pi2/pi3<1, pi2<pi3

> #So,pi1>pi3>pi2

> #PiD\_hat > PiI\_hat > PiR\_hat

>

> #(f)

> coefficients(model21)

(Intercept):1 (Intercept):2 gendermale:1 gendermale:2 racewhite:1

1.3882465 -1.1771027 -0.2201865 0.3525732 -1.1182884

racewhite:2

1.1598459

>

> c1=c(coefficients(model21)[1],coefficients(model21)[3],coefficients(model21)[5])

> c1

(Intercept):1 gendermale:1 racewhite:1

1.3882465 -0.2201865 -1.1182884

> #log(pi1/pi3)=1.3882465-0.2201865\*male-1.1182884\*white

>

> c2=c(coefficients(model21)[2],coefficients(model21)[4],coefficients(model21)[6])

> c2

(Intercept):2 gendermale:2 racewhite:2

-1.1771027 0.3525732 1.1598459

> #log(pi2/pi3)=-1.1771027+0.3525732\*male+1.159845\*white

>

> #log(pi1/pi3)-log(pi2/pi3)=log(pi1/pi2)=c1-c2

> c1-c2

(Intercept):1 gendermale:1 racewhite:1

2.5653492 -0.5727597 -2.2781343

> #log=(PiD/PiR)=log(pi1/pi2)=2.5653492-0.5727597\*male-2.2781343\*white

>

> #(g)

> model22=vglm(cbind(democrat,independent,republican)~gender+race,data = data2,

+ family = multinomial)

> summary(model22)

Call:

vglm(formula = cbind(democrat, independent, republican) ~ gender +

race, family = multinomial, data = data2)

Pearson residuals:

log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])

1 -0.03865 0.08791

2 0.03288 -0.29853

3 0.03077 -0.08827

4 -0.01045 0.29881

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept):1 2.5653 0.3437 7.465 8.35e-14 \*\*\*

(Intercept):2 1.1771 0.3807 3.092 0.001986 \*\*

gendermale:1 -0.5728 0.1575 -3.636 0.000277 \*\*\*

gendermale:2 -0.3526 0.1651 -2.136 0.032707 \*

racewhite:1 -2.2781 0.3428 -6.646 3.02e-11 \*\*\*

racewhite:2 -1.1598 0.3801 -3.051 0.002279 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors: 2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 0.1982 on 2 degrees of freedom

Log-likelihood: -20.1784 on 2 degrees of freedom

Number of iterations: 3

No Hauck-Donner effect found in any of the estimates

Reference group is level 3 of the response

>

> coefficients(model22)

(Intercept):1 (Intercept):2 gendermale:1 gendermale:2 racewhite:1

2.5653492 1.1771027 -0.5727597 -0.3525732 -2.2781343

racewhite:2

-1.1598459

> #log=(PiD/PiR)=log(pi1/pi2)=2.5653492-0.5727597\*male-2.2781343\*white

> #It is the same as the result in part (f)

>

> #(h)

> # Democrat vs. Independence

> d1=cbind(rep(1,259),rep(1,259),c(rep(1,132),rep(0,127)))

> d2=cbind(rep(1,54),rep(0,54),c(rep(1,42),rep(0,12)))

> d3=cbind(rep(0,302),rep(1,302),c(rep(1,172),rep(0,130)))

> d4=cbind(rep(0,71),rep(0,71),c(rep(1,56),rep(0,15)))

> data21=rbind(d1,d2,d3,d4)

> colnames(data21)=c("gender","race","democrat")

> data21=data.frame(data21)

> logit1=glm(democrat~gender+race,data=data21,family=binomial(link = logit))

>

> # Republican vs. Independence

> d1=cbind(rep(1,303),rep(1,303),c(rep(1,176),rep(0,127)))

> d2=cbind(rep(1,18),rep(0,18),c(rep(1,6),rep(0,12)))

> d3=cbind(rep(0,259),rep(1,259),c(rep(1,129),rep(0,130)))

> d4=cbind(rep(0,19),rep(0,19),c(rep(1,4),rep(0,15)))

> data22=rbind(d1,d2,d3,d4)

> colnames(data22)=c("gender","race","republican")

> data22=data.frame(data22)

> logit2=glm(republican~gender+race,data=data22,family=binomial(link = logit))

>

> summary(logit1)

Call:

glm(formula = democrat ~ gender + race, family = binomial(link = logit),

data = data21)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7943 -1.1992 0.7357 1.0655 1.1558

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.3866 0.2296 6.040 1.54e-09 \*\*\*

gender -0.2181 0.1585 -1.376 0.169

race -1.1175 0.2335 -4.785 1.71e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 930.60 on 685 degrees of freedom

Residual deviance: 902.32 on 683 degrees of freedom

AIC: 908.32

Number of Fisher Scoring iterations: 4

> summary(logit2)

Call:

glm(formula = republican ~ gender + race, family = binomial(link = logit),

data = data22)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.322 -1.171 1.040 1.040 1.697

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.1698 0.3814 -3.067 0.00216 \*\*

gender 0.3486 0.1660 2.099 0.03580 \*

race 1.1543 0.3809 3.031 0.00244 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 828.79 on 598 degrees of freedom

Residual deviance: 813.75 on 596 degrees of freedom

AIC: 819.75

Number of Fisher Scoring iterations: 4

> summary(model21)

Call:

vglm(formula = cbind(democrat, republican, independent) ~ gender +

race, family = multinomial, data = data2)

Pearson residuals:

log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])

1 -0.07696 -0.05743

2 0.19896 0.22498

3 0.07201 0.05961

4 -0.18480 -0.23505

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept):1 1.3882 0.2296 6.045 1.49e-09 \*\*\*

(Intercept):2 -1.1771 0.3807 -3.092 0.00199 \*\*

gendermale:1 -0.2202 0.1583 -1.391 0.16412

gendermale:2 0.3526 0.1651 2.136 0.03271 \*

racewhite:1 -1.1183 0.2335 -4.789 1.68e-06 \*\*\*

racewhite:2 1.1598 0.3801 3.051 0.00228 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors: 2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 0.1982 on 2 degrees of freedom

Log-likelihood: -20.1784 on 2 degrees of freedom

Number of iterations: 3

No Hauck-Donner effect found in any of the estimates

Reference group is level 3 of the response

>

> #(i)

> #The coefficients and their statistical significance in two separate logistic

> #models are the same as the corresponding parts of the baseline category logit

> #model. That is because the submodel of the baseline category logit is exactly

> #the logistic model. Taking first logistic model as a example, the sumation

> #of y=0 is exactly the number of people who choose independent

>

> #Question 3

> library(gee)

> data3=read.csv("attitudes.csv", header = T)

> names(data3)

[1] "gender" "response" "question" "case"

> dim(data3)

[1] 5550 4

>

> #(a)

> model31=gee(response~gender+as.factor(question),id=case,family=binomial,

+ corstr="unstructured", scale.fix=T,data = data3)

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept) gender as.factor(question)2

0.023939537 0.003582051 -0.097329124

as.factor(question)3

-0.149347113

> summary(model31)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit

Variance to Mean Relation: Binomial

Correlation Structure: Unstructured

Call:

gee(formula = response ~ gender + as.factor(question), id = case,

data = data3, family = binomial, corstr = "unstructured",

scale.fix = T)

Summary of Residuals:

Min 1Q Median 3Q Max

-0.5070710 -0.4827461 -0.4684470 0.5172539 0.5315530

Coefficients:

Estimate Naive S.E. Naive z Robust S.E. Robust z

(Intercept) 0.022967184 0.06773175 0.33909035 0.06778176 0.3388402

gender 0.005318694 0.08785139 0.06054194 0.08782143 0.0605626

as.factor(question)2 -0.097328985 0.02753097 -3.53525447 0.02753163 -3.5351691

as.factor(question)3 -0.149346974 0.02973943 -5.02184998 0.02973863 -5.0219851

Estimated Scale Parameter: 1

Number of Iterations: 2

Working Correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8248498 0.7958825

[2,] 0.8248498 1.0000000 0.8312594

[3,] 0.7958825 0.8312594 1.0000000

>

> #(b)

> model31$working.correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8248498 0.7958825

[2,] 0.8248498 1.0000000 0.8312594

[3,] 0.7958825 0.8312594 1.0000000

> #The working correlation indicates that in the same case,the correlation between 1 and 2 is

> #0.8248498, the correlation between 1 and 3 is 0.7958825 and the correlation

> #between 2 and 3 is 0.8312594

>

> #These estimation is large (closing to 1), which indicates that accounting

> #for clustering is necessary

>

> #(c)

> model32=gee(response~gender+as.factor(question),id=case,family=binomial,

+ corstr="exchangeable", scale.fix=T,data = data3)

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept) gender as.factor(question)2

0.023939537 0.003582051 -0.097329124

as.factor(question)3

-0.149347113

>

> model31$working.correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8248498 0.7958825

[2,] 0.8248498 1.0000000 0.8312594

[3,] 0.7958825 0.8312594 1.0000000

> model32$working.correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8173308 0.8173308

[2,] 0.8173308 1.0000000 0.8173308

[3,] 0.8173308 0.8173308 1.0000000

> #The working correlation matrix indicates that the correlation between observations

> #within a subject is estiamted to be 0.8173308

> #Because we spwcified an exchangeable correlation structure, this correlation

> #is the same for all pairs in a group

>

> #0.8173308 is between 0.8248498 and 0.7958825 and it is also closed to either

> #of them, which indicates that is reasonable to use exchangeable correlation

>

> #(d)

> summary(model32)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit

Variance to Mean Relation: Binomial

Correlation Structure: Exchangeable

Call:

gee(formula = response ~ gender + as.factor(question), id = case,

data = data3, family = binomial, corstr = "exchangeable",

scale.fix = T)

Summary of Residuals:

Min 1Q Median 3Q Max

-0.5068644 -0.4825396 -0.4687095 0.5174604 0.5312905

Coefficients:

Estimate Naive S.E. Naive z Robust S.E.

(Intercept) 0.024021377 0.06774107 0.35460579 0.06779334

gender 0.003437873 0.08787462 0.03912248 0.08784072

as.factor(question)2 -0.097329120 0.02811568 -3.46173752 0.02753161

as.factor(question)3 -0.149347107 0.02813360 -5.30849707 0.02973865

Robust z

(Intercept) 0.35433237

gender 0.03913758

as.factor(question)2 -3.53517666

as.factor(question)3 -5.02198729

Estimated Scale Parameter: 1

Number of Iterations: 2

Working Correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8173308 0.8173308

[2,] 0.8173308 1.0000000 0.8173308

[3,] 0.8173308 0.8173308 1.0000000

> exp(coefficients(model32)[2])

gender

1.003444

> #The estimated coefficient for the gender is 0.003437873

> #This means that females are 1.003444 times more likley to support legalized abortion

>

> qnorm(0.975)

[1] 1.959964

>

> #Naive z 0.03912248 < 1.959964

> #Can not reject H0

>

> #(e)

> lb=0.003437873-qnorm(0.975)\*0.08787462

> ub=0.003437873+qnorm(0.975)\*0.08787462

>

> #the 95% CI for is

> c(lb,ub)

[1] -0.1687932 0.1756690

>

> #the 95% CI for is

> c(exp(lb),exp(ub))

[1] 0.8446836 1.1920434

>

> #(f)

> #response=0.024021377+0.003437873\*female-0.097329120\*Q2-0.149347107\*Q3

>

> #estimated odds of support for legalized abortion in scenario 2 for a male

> odds=0.024021377-0.097329120

> odds

[1] -0.07330774

>

> #(g)

> model33=gee(response~gender+question,id=case,family=binomial,

+ corstr="exchangeable", scale.fix=T,data = data3)

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept) gender question

0.091085560 0.003581948 -0.074688363

> summary(model33)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit

Variance to Mean Relation: Binomial

Correlation Structure: Exchangeable

Call:

gee(formula = response ~ gender + question, id = case, data = data3,

family = binomial, corstr = "exchangeable", scale.fix = T)

Summary of Residuals:

Min 1Q Median 3Q Max

-0.5049607 -0.4862922 -0.4668213 0.5137078 0.5331787

Coefficients:

Estimate Naive S.E. Naive z Robust S.E. Robust z

(Intercept) 0.091155117 0.07149404 1.27500299 0.07221536 1.26226763

gender 0.003376898 0.08787193 0.03842977 0.08783897 0.03844419

question -0.074688407 0.01406974 -5.30844210 0.01487266 -5.02186055

Estimated Scale Parameter: 1

Number of Iterations: 2

Working Correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8172912 0.8172912

[2,] 0.8172912 1.0000000 0.8172912

[3,] 0.8172912 0.8172912 1.0000000

>

> #Comparing to part(c)

> summary(model32)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit

Variance to Mean Relation: Binomial

Correlation Structure: Exchangeable

Call:

gee(formula = response ~ gender + as.factor(question), id = case,

data = data3, family = binomial, corstr = "exchangeable",

scale.fix = T)

Summary of Residuals:

Min 1Q Median 3Q Max

-0.5068644 -0.4825396 -0.4687095 0.5174604 0.5312905

Coefficients:

Estimate Naive S.E. Naive z Robust S.E.

(Intercept) 0.024021377 0.06774107 0.35460579 0.06779334

gender 0.003437873 0.08787462 0.03912248 0.08784072

as.factor(question)2 -0.097329120 0.02811568 -3.46173752 0.02753161

as.factor(question)3 -0.149347107 0.02813360 -5.30849707 0.02973865

Robust z

(Intercept) 0.35433237

gender 0.03913758

as.factor(question)2 -3.53517666

as.factor(question)3 -5.02198729

Estimated Scale Parameter: 1

Number of Iterations: 2

Working Correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8173308 0.8173308

[2,] 0.8173308 1.0000000 0.8173308

[3,] 0.8173308 0.8173308 1.0000000

>

> #(h)

> model34=glm(response~gender+as.factor(question),family=binomial,data=data3)

> summary(model34)

Call:

glm(formula = response ~ gender + as.factor(question), family = binomial,

data = data3)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.189 -1.148 -1.125 1.207 1.231

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.023940 0.055528 0.431 0.6664

gender 0.003582 0.054138 0.066 0.9472

as.factor(question)2 -0.097329 0.065783 -1.480 0.1390

as.factor(question)3 -0.149347 0.065825 -2.269 0.0233 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7689.5 on 5549 degrees of freedom

Residual deviance: 7684.2 on 5546 degrees of freedom

AIC: 7692.2

Number of Fisher Scoring iterations: 3

>

> #Comparing to part(c)

> summary(model32)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit

Variance to Mean Relation: Binomial

Correlation Structure: Exchangeable

Call:

gee(formula = response ~ gender + as.factor(question), id = case,

data = data3, family = binomial, corstr = "exchangeable",

scale.fix = T)

Summary of Residuals:

Min 1Q Median 3Q Max

-0.5068644 -0.4825396 -0.4687095 0.5174604 0.5312905

Coefficients:

Estimate Naive S.E. Naive z Robust S.E.

(Intercept) 0.024021377 0.06774107 0.35460579 0.06779334

gender 0.003437873 0.08787462 0.03912248 0.08784072

as.factor(question)2 -0.097329120 0.02811568 -3.46173752 0.02753161

as.factor(question)3 -0.149347107 0.02813360 -5.30849707 0.02973865

Robust z

(Intercept) 0.35433237

gender 0.03913758

as.factor(question)2 -3.53517666

as.factor(question)3 -5.02198729

Estimated Scale Parameter: 1

Number of Iterations: 2

Working Correlation

[,1] [,2] [,3]

[1,] 1.0000000 0.8173308 0.8173308

[2,] 0.8173308 1.0000000 0.8173308

[3,] 0.8173308 0.8173308 1.0000000

>

> #The coefficents estimates of these two models are the same

> #The standard error of exchangeable model is smaller than the standard error

> #of independence model

>

> #(i)

> #It is because that as for GEE, there is a large correlation within subject, so

> #standard error is smaller than GLM. However, using the same data, the total

> #standard errors are always the same. Therefore, GEE has larger between-subject

> #standard error.

>

>

> #