6231 HW4 CODE

Question 1

a)

```
y=c(0,0,0,0,1,1,1,1)
x1=c(1,2,3,3,5,6,10,11)
m=glm(y~x1,family=binomial(link=logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(m)
```

```
## 
## Call:
## glm(formula = y ~ x1, family = binomial(link = logit))
## 
## Deviance Residuals: 
##        Min          1Q      Median          3Q         Max  
## -8.605e-06  -2.167e-06   0.000e+00   2.110e-08   1.288e-05  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -94.87  202572.35       0        1
## x1              23.62   48491.51       0        1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1.109e+01  on 7  degrees of freedom
## Residual deviance: 3.139e-10  on 6  degrees of freedom
## AIC: 4
## 
## Number of Fisher Scoring iterations: 25
```

The warning message indicates that we have a variable which perfectly separates 0 and 1 in targe variables. \(log(\frac{\hat\pi}{1-\hat\pi})=-94.87+23.62*x1\)

The coeffeient estimate are =-94.84, =23.62

The standard erreors are 202572.35 and 28291.51

There is an indication of complete separation

When y equals to 0, x1 equals to 1,2,3,3,, equals to 0 when we use 10 digits accuracy.

When y equals to 1, x1 equals to 5,6,10,11, equals to 1 when we use 10 digits accuacy.

b)

```
x1=c(1,2,3,3,3,6,10,11)
m=glm(y~x1,family=binomial(link=logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(m)
```

```
## 
## Call:
## glm(formula = y ~ x1, family = binomial(link = logit))
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.9005  -0.2252   0.0000   0.0000   1.4823  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -54.08   18834.18  -0.003    0.998
## x1             17.80    6278.06   0.003    0.998
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 11.0904  on 7  degrees of freedom
## Residual deviance:  3.8191  on 6  degrees of freedom
## AIC: 7.8191
## 
## Number of Fisher Scoring iterations: 21
```

The warning message indicates that we have a variable which perfectly separates 0 and 1 in targe variables. \(log(\frac{\hat\pi}{1-\hat\pi})=-54.08+17.8*x1\)

The coeffeient estimate are =54.08, =17.8

The standard erreors are 18834.18 and 6278.06

There is an indication of quassi-complete separation When y equals to 0, x1 equals to 1,2 equals to 0 when we use 10 digits accuracy.

When y equals to 1, x1 equals to 6,10,11, equals to 1 when we use 10 digits accuacy.

However, two obseration at x1=3 one with y=1 and one with y=0

Question 4

a)

```
data=read.table("d://book.txt",header = TRUE)
m2=glm(survival~age,family=binomial(link=logit),data=data)
m2$coefficients
```

```
## (Intercept)         age 
##  0.97917294 -0.03688823
```

```
exp(m2$coefficients)
```

```
## (Intercept)         age 
##   2.6622535   0.9637838
```

```
summary(m2)
```

```
## 
## Call:
## glm(formula = survival ~ age, family = binomial(link = logit), 
##     data = data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.5946  -1.2017   0.8436   0.9882   1.5765  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)   
## (Intercept)  0.97917    0.37460   2.614  0.00895 **
## age         -0.03689    0.01493  -2.471  0.01346 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 120.86  on 87  degrees of freedom
## Residual deviance: 114.02  on 86  degrees of freedom
## AIC: 118.02
## 
## Number of Fisher Scoring iterations: 4
```

\(\hat{\beta}_{age}\) =-0.03689

\(e^{\hat{\beta}_{age}}\)=0.9637

\(\hat{\beta}_{age}\) mean when age change one unit, the log(odd ratio) will change -0.03689 time. It means that a old people decreased the log odd by \(\hat{\beta}_{age}\) =-0.03689

\(e^{\hat{\beta}_{age}}\) mean when age change one unit, the odd ratio will change 0.9637 time. It means that the old people possible survival rate is lower than younger people.

b)

```
-0.03689/0.01493
```

```
## [1] -2.470864
```

H0: \(\hat{\beta}_{age} = 0\) Ha:\(\hat{\beta}_{age} \neq 0\)

\(z_{w}=\frac{\hat{\beta}_{age}-0}{SE({\hat{\beta}_{age}})}=\frac{-0.03689}{0.01493} =-2.470864\)

Becase\(|z_{w}|=2.471>z_{\alpha/2}=1.96\) so we reject H0.

c)

```
lb = m2$coefficients[2] - qnorm(.05/2,lower.tail=FALSE)*summary(m2)$coefficients[2,2]

ub = m2$coefficients[2] + qnorm(.05/2,lower.tail=FALSE) * summary(m2)$coefficients[2,2]

c(lb,ub)
```

```
##          age          age 
## -0.066141852 -0.007634613
```

```
exp(c(lb,ub))
```

```
##       age       age 
## 0.9359981 0.9923945
```

From above we can get the 95% confident interval of \(\hat{\beta}_{age}\) is (-0.066141852,-0.007634613). Because the confident interval do not contain the value zero, so we reject the H0 hyphothesis. So there are significant effect on age and survival.

d)

```
age2=(data$age)^2
m3=glm(survival~age+age2+sex+status, family=binomial(link=logit),data=data)
summary(m3)
```

```
##
## Call:
## glm(formula = survival ~ age + age2 + sex + status, family = binomial(lin
k = logit),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0431  -1.0391   0.5120   0.8664   2.0797
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.986e-01  6.172e-01   0.322   0.7476
## age           1.675e-01  7.107e-02   2.357   0.0184 *
## age2         -3.889e-03  1.525e-03  -2.550   0.0108 *
## sex          -6.637e-01  5.588e-01  -1.188   0.2349
## statusHired  -1.625e+00  7.481e-01  -2.173   0.0298 *
## statusSingle -1.852e+01  1.760e+03  -0.011   0.9916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120.855  on 87  degrees of freedom
## Residual deviance:  92.363  on 82  degrees of freedom
## AIC: 104.36
##
## Number of Fisher Scoring iterations: 16
```

```
diff.dev = deviance(m2) - deviance(m3)
diff.dev
```

```
## [1] 21.65525
```

```
qchisq(.95,4)
```

```
## [1] 9.487729
```

H0: reduced model Ha:Full model

\(G^{2}=21.65525>\chi^{2}_{4,0.05}=9.487729\) So we reject the H0.So the model in part (d) is better than model in part(a).

e)

```
m3$coefficients[4]
```

```
##       sex
## -0.663728
```

```
exp(m3$coefficients[4])
```

```
##       sex
## 0.5149281
```

From above,we can see, \(\hat{\beta}_{sex}=-0.663728\).It means that a man decreased the log odd of survivor by \(\hat{\beta}_{sex}=-0.663728\).

\(e^{\hat{\beta}_{sex}}=0.5149281\). It means that the man possible survival rate is lower than woman for 0.5149 time.

Question 5

a)

```
Hire=rep(0,88)
for (i in 1:88){ if(data$status[i]=='Hired'){Hire[i]=1}}
single=rep(0,88)
for (i in 1:88){ if(data$status[i]=='Single'){single[i]=1}}
newx=cbind(1,data$age,age2,data$sex,Hire,single)
fit_value=newx%*%m3$coefficients
summary(fit_value)
```

```
##        V1
##  Min.   :-17.8917
##  1st Qu.: -0.3345
##  Median :  0.3622
##  Mean   : -0.7447
##  3rd Qu.:  1.1821
##  Max.   :  2.0004
```

By using the data and the model of 4(d), we get the result of fitted value for the log odd of survivor. And the descriptive statistics are showing above.

b)

```
pred.prob = exp(fit_value) / (1 + exp(fit_value))
summary(pred.prob)
```

```
##        V1
##  Min.   :0.0000
##  1st Qu.:0.4171
##  Median :0.5896
##  Mean   :0.5568
##  3rd Qu.:0.7653
##  Max.   :0.8808
```

By using the data and the model of 4(d), we get the result of fitted value survivor \(\pi_{i}=\frac{e^{X_{i}\beta}}{1+e^{X_{i}\beta}}\). And the descriptive statistics are showing above.

C)i)

```
cutoff_value=rep(0,88)
for (i in 1:88){
  if (pred.prob[i]>0.5)
  {cutoff_value[i]=1}
  else
  {cutoff_value[i]=0}
}
cross_table=table(data$survival,cutoff_value)
cross_table
```

```
##    cutoff_value
##      0  1
##   0 23 16
##   1  7 42
```

From above cross-tabulation table we can find that there are 23 incorrect predict value. false positive is 16 and false negative is 7

C)ii)

```
cutoff_value2=rep(0,88)
proportion=sum(data$survival)/88
for (i in 1:88){
  if (pred.prob[i]>proportion)
  {cutoff_value2[i]=1}
  else
  {cutoff_value2[i]=0}
}
cross_table=table(data$survival,cutoff_value2)
cross_table
```

```
##    cutoff_value2
##      0  1
##   0 27 12
##   1 10 39
```

From above cross-tabulation table we can find that there are 22 incorrect predict value. false positive is 12 and false negative is 10

C)iii)

p1=23/88=0.2614

p2=22/88=0.25

Only base on the proportion of incorrect, I will choose the second method because its proportion is smaller than first method.

c) iv)

First method: \(FPR=P(\hat{y}=1|y=0)\)=16/39=0.41 \(FNR=P(\hat{y}=0|y=1)\)=7/49=0.143

Second method: \(FPR=P(\hat{y}=1|y=0)\)=12/39=0.3077 \(FNR=P(\hat{y}=0|y=1)\)=10/49=0.204

c) v)

Based on the false positive and false positive rate, FPR(1)>FPR(2), FNR(2)<FNR(1).In survival situation,especially in insurance industry, the false positive rate is more seriour than false negetive. If we estimate a person is more likely to dead, then it is no profitable to sale the insurance for them.So if a person is actually more possible to die but we predict he/she can live, then it would lost money. So I would choose the second method.

c) vi)

```
#install.packages("pROC")
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

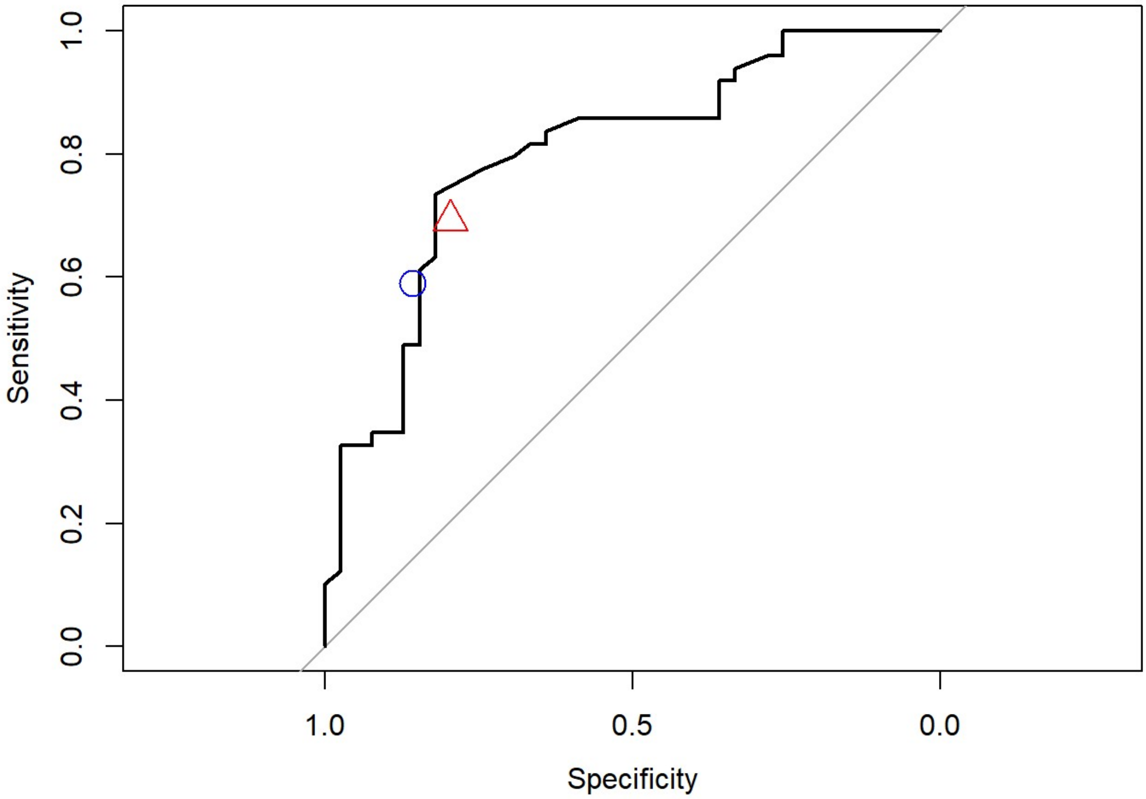
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
data2=data.frame(cbind(data$survival,pred.prob,cutoff_value,cutoff_value2))
roc(data2$V1,data2$V2,plot=TRUE)
```

```
##
## Call:
## roc.default(response = data2$V1, predictor = data2$V2, plot = TRUE)
##
## Data: data2$V2 in 39 controls (data2$V1 0) < 49 cases (data2$V1 1).
## Area under the curve: 0.8004
```

```
points(42/49,1-16/39,col='blue',cex=2,pch=21)
points(39/49,1-12/39,col='red',cex=2,pch=24)
```



From the above Roc cruve,we can see the area under the curve is 0.8004.Roc curve is a very useful method to compare the model, The bigger the value of AUC(area under curve) is, the better the model is. This AUC is 0.8004 not too bad, it can provide a realiable prediction.

Question 7

a)

```
rm(list = ls())
data=read.csv("d:/gss.csv")
library(VGAM)
```

```
## Loading required package: stats4
## Loading required package: splines
```

```
m1 = vglm(cbind(democrat,republican,independent)~gender+race, data=data, fam
ily=multinomial)
summary(m1)
```

```
##
## Call:
## vglm(formula = cbind(democrat, republican, independent) ~ gender +
##     race, family = multinomial, data = data)
##
##
## 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(m1)
```

```
## [1] 0.1982117
```

```
qchisq(.05,2,lower.tail=FALSE)
```

```
## [1] 5.991465
```

```
1-pchisq(deviance(m1),2)
```

```
## [1] 0.9056468
```

From above result, we use the deviance GOF to test the fitness. We can see the gof 0.1982<5.9915 which mean we cannot reject the null(reduced model in favor of the saturated model). The model able to fit main effect and provide reasonable information

c

From the part a) reuslt, we can see that when compared with democrate to independent, the gender is no significant beause the p value is 0.16412 which is larger than 0.05. However when it comes to republican, the gender effect show significant since the p value is 0.03271 smailler than 0.05.

d)

```
exp(coefficients(m1))
```

```
## (Intercept):1 (Intercept):2  gendermale:1  gendermale:2   racewhite:1 
##     4.0078161     0.3081703     0.8023691     1.4227238     0.3268387 
##   racewhite:2 
##     3.1894417 
```

The (intercept)1 is 4.008 which mean that the odd ratio of being a democrate ofr black females is 4.01. for \(\frac{\pi}{1-\pi}=4.008\) and =0.7539.

e)

The log likelihood of democrate vs baseline is 1.3882 which greater than 0 and the log likelihood of republican vs baseline is -1.1771 which is samller than 0. so it is obviously that *{democrate}*> {repulican}

f)

\(log(\frac{\pi_{d}}{\pi_{r}})=log(\frac{\pi{_d}}{\pi_{i}}\frac{\pi{_i}}{\pi_{r}})=(1.3882+1.1771)+ (-0.2202-0.3526)*gender+(-1.1183-1.1598)*race=2.5653-0.5728*gender-2.2781*race\)

g)

```
m2 = vglm(cbind(democrat,independent,republican)~gender+race, data=data, family=multinomial)
summary(m2)
```

```
##
## Call:
## vglm(formula = cbind(democrat, independent, republican) ~ gender +
##     race, family = multinomial, data = data)
##
##
## 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
```

We can see the $logis same with result in part f)

h)

```
m3 = vglm(cbind(democrat,republican,independent)~gender, data=data, family=m
ultinomial)
summary(m3)
```

```
##
## Call:
## vglm(formula = cbind(democrat, republican, independent) ~ gender,
##     family = multinomial, data = data)
##
##
## Pearson residuals:
##   log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
## 1             -1.857              1.102
## 2              5.000             -2.969
## 3             -1.898              1.207
## 4              4.550             -2.894
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  0.45261    0.10622   4.261 2.03e-05 ***
## (Intercept):2 -0.08638    0.12006  -0.719   0.4718
## gendermale:1  -0.22803    0.15564  -1.465   0.1429
## gendermale:2   0.35592    0.16463   2.162   0.0306 *
## ---
## 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: 76.9224 on 4 degrees of freedom
##
## Log-likelihood: -58.5405 on 4 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level  3  of the response
```

```
exp(coefficients(m3))
```

```
## (Intercept):1 (Intercept):2  gendermale:1  gendermale:2
##     1.5724138     0.9172414     0.7961000     1.4274896
```

```
m4= vglm(cbind(democrat,republican,independent)~race, data=data, family=mult
inomial)
summary(m4)
```

```
##
## Call:
## vglm(formula = cbind(democrat, republican, independent) ~ race,
##     family = multinomial, data = data)
##
##
## Pearson residuals:
##   log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
## 1            -1.6042             1.9129
## 2            -0.3330             0.6994
## 3             1.6116            -1.9217
## 4             0.2979            -0.6255
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   1.2891     0.2174   5.931 3.01e-09 ***
## (Intercept):2  -0.9933     0.3702  -2.683  0.00729 **
## racewhite:1    -1.1212     0.2333  -4.806 1.54e-06 ***
## racewhite:2     1.1645     0.3797   3.066  0.00217 **
## ---
## 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: 13.64 on 4 degrees of freedom
##
## Log-likelihood: -26.8993 on 4 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level  3  of the response
```

```
exp(coefficients(m4))
```

```
## (Intercept):1 (Intercept):2   racewhite:1   racewhite:2
##     3.6296296     0.3703704     0.3258953     3.2042802
```

We can see for the different model, the estimated coefficients and the statistical significance of two models are different.

i)

The positive correlation can improve the correctness of estimating within subject effects. But for the inference between subject effect, the observation of single subject can not provide enough information as T observation. Pluse the positive correlation will causes large SE