Final Project

11.1

```
#1
seg=read.csv('C:/Users/User/Downloads/seg-large.csv')
seg$utterance= as.factor(seg$utterance)
```

11.2

```
#2
or=read.csv("C:/Users/User/Downloads/past-tense-or.csv")
or$correct=(or$n.past-or$n.or)/or$n.past
```

11.3

This is the original dataset

```
#3
modor=lm(correct~age, data=or)
par(mfrow=c(2,2))
plot(modor)
summary(modor)
```

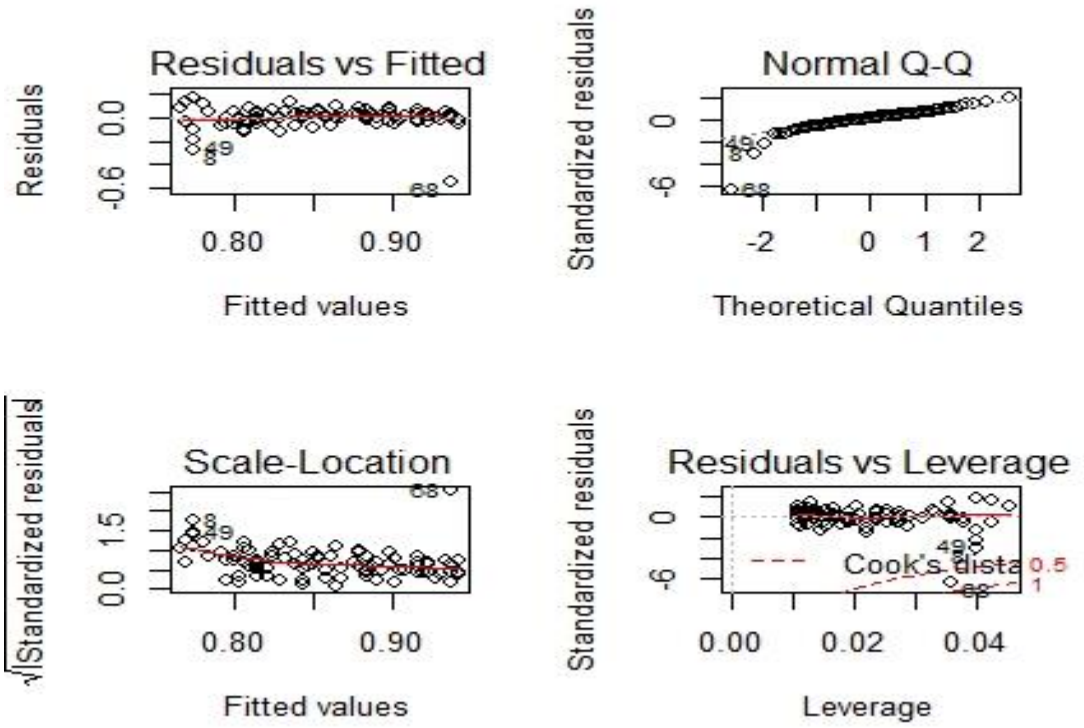
```
Call:
lm(formula = correct ~ age, data = or)

Residuals:
    Min      1Q  Median      3Q     Max
-0.55622 -0.02846  0.00802  0.04420  0.16823

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept) 0.7215229  0.0260231  27.726     < 2e-16 ***
age         0.0036732  0.0006502   5.649 0.000000187 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.0888 on 90 degrees of freedom
Multiple R-squared:  0.2618,    Adjusted R-squared:  0.253
5
```

```
F-statistic: 31.91 on 1 and 90 DF,  p-value: 0.0000001868
```



The variance of residual is not a constant, and have several influential point.

After remove the most influential point:

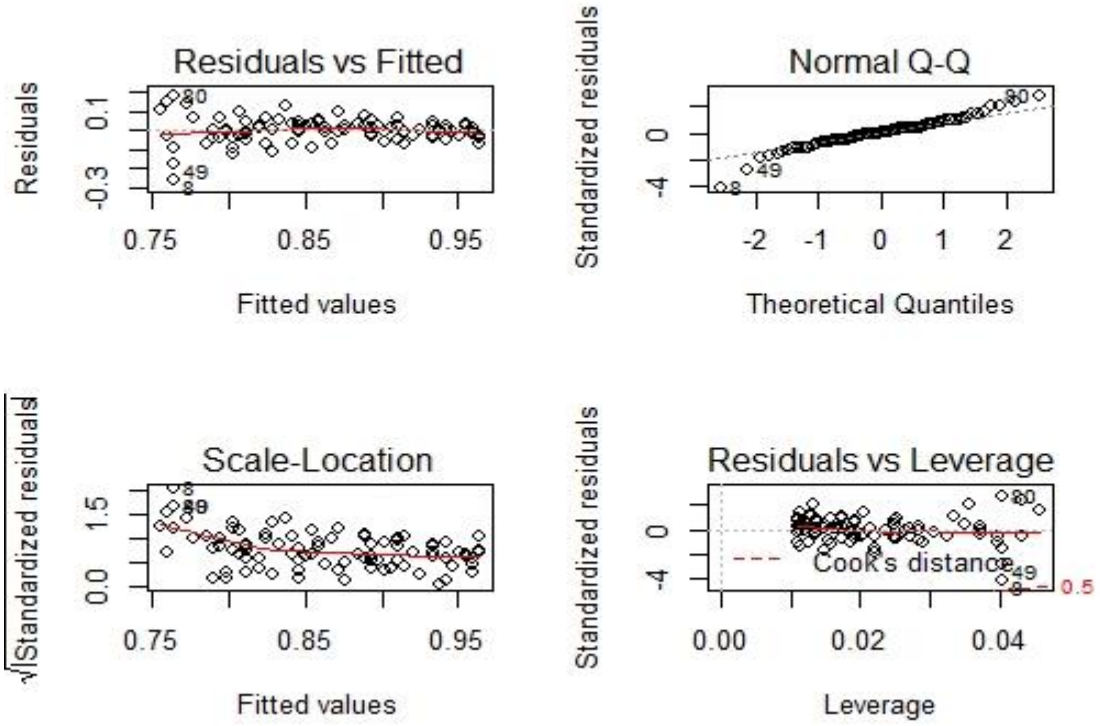
```
for (i in c(1:92)){
   if(or$correct[i]<0.383){
      newor <- or[-i,]
   }
}
plot(newor$age,newor$correct)
or.ols=lm(correct~age, data=newor)
plot(or.ols)
summary(or.ols)
```

```
Call:
lm(formula = correct ~ age, data = newor)

Residuals:
     Min       1Q   Median       3Q      Max
-0.263582 -0.033580  0.001687  0.030433  0.177595
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7028025  0.0194914  36.057  < 2e-16 ***
age         0.0043414  0.0004902   8.856 7.46e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.06609 on 89 degrees of freedom
Multiple R-squared:  0.4684,    Adjusted R-squared:  0.462
5
F-statistic: 78.43 on 1 and 89 DF,  p-value: 7.457e-14
```



We can see that the expected value of residual is no longer be zero, and in qq plot and cook's distant plot we can see there are some other influential point need to be resolved.

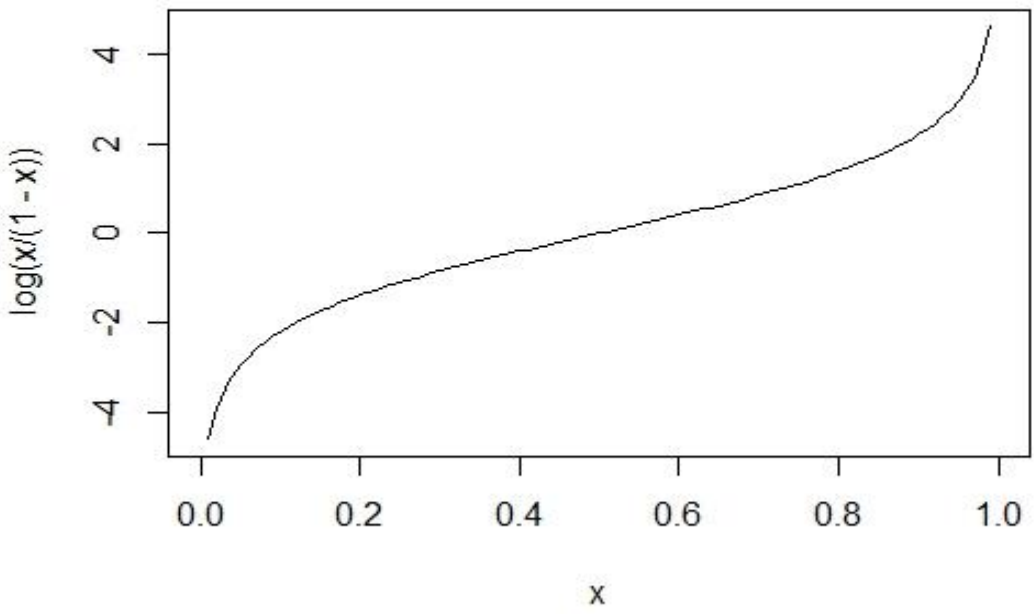
11.4

```
#4
or.ols$coefficients[1]+1.5*or.ols$coefficients[2]
or.ols$coefficients[1]+8*or.ols$coefficients[2]
```

At age of 1.5 month, the ols gives the predicted value of 0.7093145;
At age of 8 month, the ols gives the predicted value of 0.7375334

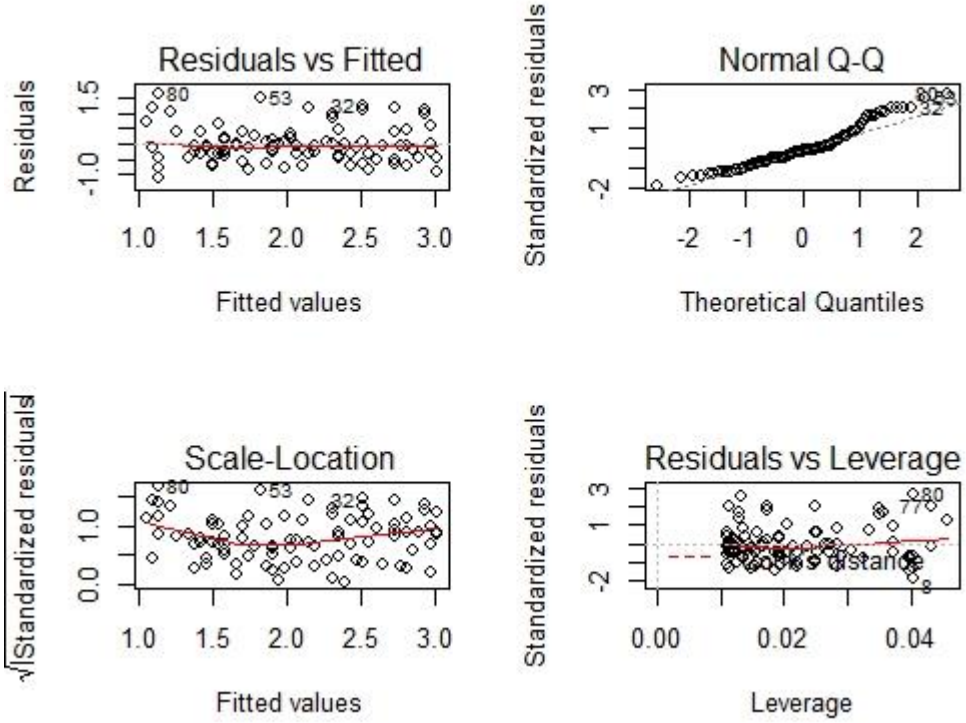
11.5

```
#5
par(mfrow=c(2,2))
curve(log(x/(1-x)),0,1)
```



11.6

```
#6
newor$log.odds=log(newor$correct/(1-newor$correct))
or.logit=lm(log.odds~age,data=newor)
plot(or.logit)
summary(or.logit)
```



Expected value of residuals are nearly zero and variances are close enough to be constant, but we can see that the normality is worse than previous OLS model, more observations are too far from normal distribution.

11.7

```
#7
expoyhat=exp(or.logit$coefficients[1]+c(12*(1:10))*or.logit$coefficients[2])
expoyhat/(1+expoyhat)
```

```
[1] 0.7414705 0.8237845 0.8839908 0.9254875 0.9529311 0.97
05882 0.9817484 0.9887231 0.9930513 0.9957255
```

11.8

```
Call:
glm(formula = cbind(n.past - n.or, n.or) ~ age, family = "q
uasibinomial",data = or)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.6520   -0.4567    0.2872    0.8628    2.5795
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.980148   0.325579   3.010  0.00338 **
```

```
age         0.023886   0.007838   3.047  0.00303 **
---(Dispersion parameter for quasibinomial family taken to
 be 4.756386)
    Null deviance: 315.66  on 91  degrees of freedom
Residual deviance: 271.44  on 90  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 4
```

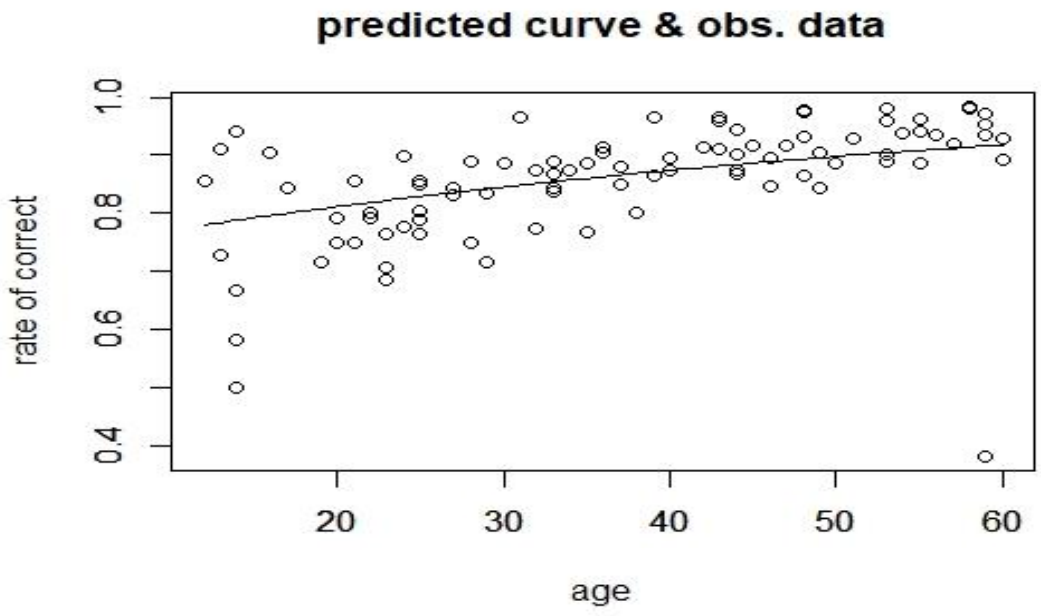
11.9

```
or.glm=glm(cbind(n.past-n.or,n.or)~age, data=or, family='binomial')
expoyhat=exp(or.glm$coefficients[1]+c(12*c(1:10))*or.glm$coefficients[2])
expoyhat/(1+expoyhat)
```

```
[1] 0.7801915 0.8254067 0.8629548 0.8934700 0.9178374 0.93
70240 0.9519646 0.9634987 0.9723438 0.9790920
```

11.10

```
plot(or$age,or$correct,main=' ', xlab=' ', ylab=' ')
par(new = T)
curve((exp(0.98+x*0.02388))/(1+exp(0.98+x*0.02388)),
      xlim=(range(or$age)), ylim=range(or$correct) ,
      main='predicted curve & obs. data', xlab='age', ylab='rate of correct')
par(mfrow=c(1,1))
```

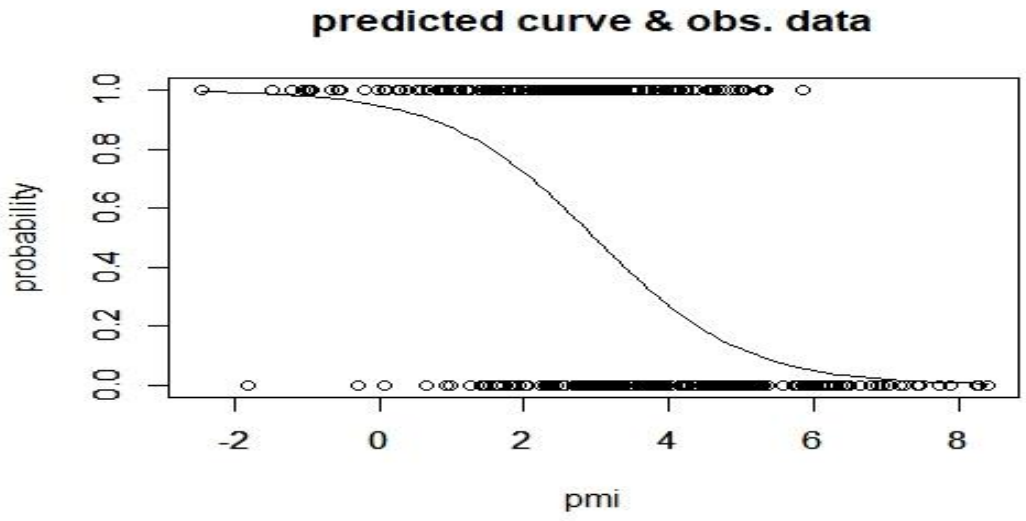


11.11

```
seg.pmi=glm(boundary~pmi , data=seg, family='binomial')
summary(seg.pmi)#no overdispersion
plot(seg$pmi, seg$boundary,main=' ', xlab=' ', ylab=' ')
par(new = T)
curve(exp(2.90593-0.97588*x)/(1+exp(2.90593-0.97588*x)),
      xlim=(range(seg$pmi)), ylim=range(seg$boundary),main='predicted curve &
obs. data', xlab='pmi', ylab='probability')
```

```
Call:
glm(formula = boundary ~ pmi, family = "binomial", data = s
eg)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0562  -0.7984  -0.4698   0.8528   2.3995
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.90593    0.27718   10.48   <2e-16 ***
pmi         -0.97588    0.07604  -12.83   <2e-16 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1073.4  on 820  degrees of freedom
Residual deviance:  805.7  on 819  degrees of freedom
AIC: 809.7
Number of Fisher Scoring iterations: 5
```

Residual deviance is slightly lower than the degree od freedom, thus we conclude that there is no overdispersion



11.12

```
crct=c()
segcol=c()
segcol= predict(seg.pmi, type='response')>0.5
a=0
for (i in c(1:821)){
   if(segcol[i]==seg$boundary[i]){
     a=a+1
   }
}
a/821
```

```
[1] 0.772229
```

11.13

```
b=0
for (i in c(1:821)){
   if(seg$boundary[i]==FALSE){
      b=b+1
   }
}
b/821
```

```
[1] 0.6394641
```

11.14

```
seg.full=glm(boundary~pmi+utterance+phoneme, data=seg, family='binomial')
summary(seg.full)
```

```
Call:
glm(formula = boundary ~ pmi + utterance + phoneme, family
 = "binomial",data = seg)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.98376  -0.20337  -0.00002   0.31852   2.65921
Coefficients:
              Estimate   Std. Error z value Pr(>|z|)
(Intercept)    25.664397 12462.411662   0.002    0.998
pmi           -1.251848    0.134945  -9.277   <2e-16 ***
utterance      -0.001499    0.004519  -0.332    0.740
phoneme[T.%]   -1.695533 14674.695006   0.000    1.000
```

phoneme[T.&]	-22.591141	12462.411658	-0.002	0.999
phoneme[T.)]	1.697198	17677.519221	0.000	1.000
phoneme[T.*]	-20.493828	12462.411687	-0.002	0.999
phoneme[T.~]	-0.497341	16916.528202	0.000	1.000
phoneme[T.3]	0.365881	21672.049326	0.000	1.000
phoneme[T.6]	-19.292939	12462.411661	-0.002	0.999
phoneme[T.7]	-1.882760	17503.130345	0.000	1.000
phoneme[T.9]	-21.211505	12462.411666	-0.002	0.999
phoneme[T.a]	-40.307494	13948.850657	-0.003	0.998
phoneme[T.A]	-41.158771	12752.431209	-0.003	0.997
phoneme[T.b]	-43.288596	12964.625146	-0.003	0.997
phoneme[T.c]	-43.349117	21672.049312	-0.002	0.998
phoneme[T.d]	-22.865481	12462.411664	-0.002	0.999
phoneme[T.D]	-42.106991	12688.899882	-0.003	0.997
phoneme[T.e]	-19.058721	12462.411671	-0.002	0.999
phoneme[T.E]	-40.673860	13477.637403	-0.003	0.998
phoneme[T.f]	-40.986396	13321.473476	-0.003	0.998
phoneme[T.g]	-42.526959	13036.161163	-0.003	0.997
phoneme[T.h]	-43.462859	12884.533636	-0.003	0.997
phoneme[T.i]	-20.609488	12462.411660	-0.002	0.999
phoneme[T.I]	-41.395647	12632.658793	-0.003	0.997
phoneme[T.k]	-21.150956	12462.411656	-0.002	0.999
phoneme[T.l]	-21.345124	12462.411661	-0.002	0.999
phoneme[T.m]	-22.535739	12462.411699	-0.002	0.999
phoneme[T.n]	-20.516863	12462.411656	-0.002	0.999
phoneme[T.N]	-22.368454	12462.411696	-0.002	0.999
phoneme[T.o]	-21.369808	12462.411665	-0.002	0.999
phoneme[T.O]	-38.515286	13760.591923	-0.003	0.998
phoneme[T.p]	-22.510483	12462.411705	-0.002	0.999
phoneme[T.Q]	-22.758754	12462.411700	-0.002	0.999
phoneme[T.r]	-45.900757	13389.281514	-0.003	0.997
phoneme[T.R]	-18.623624	12462.411808	-0.001	0.999
phoneme[T.s]	-19.797966	12462.411655	-0.002	0.999
phoneme[T.S]	-2.033498	17677.519219	0.000	1.000
phoneme[T.t]	-21.100859	12462.411652	-0.002	0.999
phoneme[T.T]	-20.876253	12462.411703	-0.002	0.999
phoneme[T.u]	-18.586688	12462.411668	-0.001	0.999
phoneme[T.U]	-40.731806	13098.046839	-0.003	0.998

```
phoneme[T.v]    -1.610761 17668.434062   0.000    1.000
phoneme[T.w]   -41.929434 13492.491311  -0.003    0.998
phoneme[T.W]   -42.877393 12799.997217  -0.003    0.997
phoneme[T.y]   -41.768512 12894.143654  -0.003    0.997
phoneme[T.z]    -1.164947 13260.221008   0.000    1.000
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1073.42  on 820  degrees of freedom
Residual deviance:  396.13  on 774  degrees of freedom
AIC: 490.13
Number of Fisher Scoring iterations: 19
```

Only the coefficient of pmi is significant

11.15

```
anova(seg.pmi,seg.full, test= "Chisq")
```

```
Analysis of Deviance Table

Model 1: boundary ~ pmi
Model 2: boundary ~ pmi + utterance + phoneme
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       819     805.70
2       774     396.13 45   409.58 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

Chi-square test is significant, and cant conclude that the model adding utterance and phoneme fits significantly better

11.16

```
seg.model=step(model.full)
summary(seg.model)
```

```
Call:
glm(formula = boundary ~ pmi + factor(phoneme), family = "binomi
al",
    data = seg)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
```

```
-2.99180  -0.19552  -0.00003   0.32533   2.67532

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           25.6016 12455.1933   0.002    0.998
pmi                   -1.2529     0.1349  -9.286   <2e-16 ***
factor(phoneme)[T.%]   -1.7224 14676.2328   0.000    1.000
factor(phoneme)[T.&]  -22.6073 12455.1933  -0.002    0.999
factor(phoneme)[T.)]    1.6372 17672.4319   0.000    1.000
factor(phoneme)[T.*]  -20.5267 12455.1933  -0.002    0.999
factor(phoneme)[T.~]   -0.5422 16936.0018   0.000    1.000
factor(phoneme)[T.3]    0.3775 21667.8993   0.000    1.000
factor(phoneme)[T.6]  -19.3064 12455.1933  -0.002    0.999
factor(phoneme)[T.7]   -1.8921 17523.3056   0.000    1.000
factor(phoneme)[T.9]  -21.2457 12455.1933  -0.002    0.999
factor(phoneme)[T.a]  -40.3026 13943.4387  -0.003    0.998
factor(phoneme)[T.A]  -41.1670 12745.1699  -0.003    0.997
factor(phoneme)[T.b]  -43.3173 12957.4413  -0.003    0.997
factor(phoneme)[T.c]  -43.3876 21667.8992  -0.002    0.998
factor(phoneme)[T.d]  -22.9084 12455.1933  -0.002    0.999
factor(phoneme)[T.D]  -42.1160 12681.7827  -0.003    0.997
factor(phoneme)[T.e]  -19.0570 12455.1933  -0.002    0.999
factor(phoneme)[T.E]  -40.6683 13476.8505  -0.003    0.998
factor(phoneme)[T.f]  -41.0038 13315.9790  -0.003    0.998
factor(phoneme)[T.g]  -42.5242 13029.3777  -0.003    0.997
factor(phoneme)[T.h]  -43.4421 12883.2489  -0.003    0.997
factor(phoneme)[T.i]  -20.6281 12455.1933  -0.002    0.999
factor(phoneme)[T.I]  -41.3932 12625.7557  -0.003    0.997
factor(phoneme)[T.k]  -21.1476 12455.1933  -0.002    0.999
factor(phoneme)[T.l]  -21.3555 12455.1933  -0.002    0.999
factor(phoneme)[T.m]  -22.5419 12455.1933  -0.002    0.999
factor(phoneme)[T.n]  -20.5368 12455.1933  -0.002    0.999
factor(phoneme)[T.N]  -22.3774 12455.1933  -0.002    0.999
factor(phoneme)[T.o]  -21.4019 12455.1933  -0.002    0.999
factor(phoneme)[T.O]  -38.5156 13754.7394  -0.003    0.998
factor(phoneme)[T.p]  -22.5077 12455.1933  -0.002    0.999
factor(phoneme)[T.Q]  -22.7232 12455.1933  -0.002    0.999
factor(phoneme)[T.r]  -45.8546 13387.5044  -0.003    0.997
```

```
factor(phoneme)[T.R]   -18.6467 12455.1935  -0.001    0.999
factor(phoneme)[T.s]   -19.8071 12455.1933  -0.002    0.999
factor(phoneme)[T.S]    -2.0966 17672.4319   0.000    1.000
factor(phoneme)[T.t]   -21.0952 12455.1933  -0.002    0.999
factor(phoneme)[T.T]   -20.8597 12455.1933  -0.002    0.999
factor(phoneme)[T.u]   -18.5966 12455.1933  -0.001    0.999
factor(phoneme)[T.U]   -40.7077 13092.4531  -0.003    0.998
factor(phoneme)[T.v]    -1.5835 17666.9355   0.000    1.000
factor(phoneme)[T.w]   -41.9512 13476.3407  -0.003    0.998
factor(phoneme)[T.W]   -42.9423 12792.1172  -0.003    0.997
factor(phoneme)[T.y]   -41.7705 12886.6771  -0.003    0.997
factor(phoneme)[T.z]    -1.1775 13253.3197   0.000    1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1073.42  on 820  degrees of freedom
Residual deviance:  396.24  on 775  degrees of freedom
AIC: 488.24

Number of Fisher Scoring iterations: 19
```

11.17

```
segcol= predict(seg.model, type='response')>0.5
a=0
for (i in c(1:821)){
   if(segcol[i]==seg$boundary[i]){
      a=a+1
   }
}
a/821
```

```
[1] 0.8794153
```

```
segpmi=c()
segmodel=c()
segpmi= predict(seg.pmi, type='response')>0.5
segmodel= predict(seg.model, type='response')>0.5
```

```
rightpmi=0
rightmodel=0
for (i in c(1:821)){
   if(segpmi[i]==seg$boundary[i]){
      rightpmi=rightpmi+1
   }
   if(segmodel[i]==seg$boundary[i]){
   rightmodel=rightmodel+1
   }
}
mat=matrix(c(rightpmi,(821-rightpmi),rightmodel,(821-rightmodel)),2,2)
mat
fisher.test(mat)#reject H0, the 2 model are diffrent
```

```
        Fisher's Exact Test for Count Data
data:  mat
p-value = 0.00000001216
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3525033 0.6110253
sample estimates:
odds ratio
 0.4650969
```

p-value is close to zero and can reject null hypothesis, we can conclude that the difference of the two model is significant.

11.18

```
seg.pmihrh=glm(seg$boundary[1:405]~seg$pmi[1:405]+seg$h[1:405]+seg$rh[1:405]
, data=seg, family='binomial')
segcol= predict(seg.pmihrh, type='response')>0.5
a=0
for (i in c(1:821)){
   if(segcol[i]==seg$boundary[i]){
      a=a+1
   }
}
a/821
```

```
[1] 0.4202192
```

11.19

```
seg.pmi405=glm(seg$boundary[1:405]~seg$pmi[1:405], data=seg, family='binomial')
segcol= predict(seg.pmi405, type='response')>0.5
a=0
for (i in c(1:821)){
   if(segcol[i]==seg$boundary[i]){
      a=a+1
   }
}
a/821
```

```
[1] 0.3800244
```