BTRY 4110 Prelim2

Nicholas Paschall, Shirley Zhang, Kevin Sheng, Ann Nie

11/18/2022

CLEAN THE DATA

TWO-WAY ANALYSIS

```
library(vcdExtra)

## Loading required package: vcd

## Loading required package: grid

## Loading required package: gnm

##
## Attaching package: 'vcdExtra'

## The following object is masked from 'package:dplyr':
##
##     summarise

dummy <- df %>% mutate('held' = ifelse(held==0, 'No', 'Yes'))

#RACE (BINOMIAL CATEGORY PREDICTOR)
race.tab <- table(dummy$race, dummy$held)
race.tab

##
##           No  Yes
##   Black  944  325
##   White 3315  549

chisq.test(race.tab, correct=F) #association

##
##  Pearson's Chi-squared test
##
## data:  race.tab
## X-squared = 87.915, df = 1, p-value < 2.2e-16

#EMPLOYED (BINOMIAL CATEGORY PREDICTOR)
employed.tab <- table(dummy$employed, dummy$held)
chisq.test(employed.tab, correct=F) #association
```

```
##
##  Pearson's Chi-squared test
##
## data:  employed.tab
## X-squared = 199.11, df = 1, p-value < 2.2e-16
```

```
#CITIZEN (BINOMIAL CATEGORY PREDICTOR)
citizen.tab <- table(dummy$citizen, dummy$held)
chisq.test(citizen.tab, correct=F) #association
```

```
##
##  Pearson's Chi-squared test
##
## data:  citizen.tab
## X-squared = 63.017, df = 1, p-value = 2.049e-15
```

```
#SEX (BINOMIAL CATEGORY PREDICTOR)
sex.tab <- table(dummy$sex, dummy$held)
chisq.test(sex.tab, correct=F) #no association
```

```
##
##  Pearson's Chi-squared test
##
## data:  sex.tab
## X-squared = 3.3329, df = 1, p-value = 0.06791
```

```
#REGION (NOMIAL CATEGORICAL VARIABLE)
region.tab <- table(dummy$region, dummy$held)
chisq.test(region.tab, correct=F) #no association
```

```
##
##  Pearson's Chi-squared test
##
## data:  region.tab
## X-squared = 3.0248, df = 3, p-value = 0.3878
```

```
#YEAR (NOMINAL CATEGORICAL VARIABLE)
year.tab <- table(dummy$year, dummy$held)
chisq.test(year.tab, correct=F) #association (I personally think this is
weird)
```

```
##
##  Pearson's Chi-squared test
##
## data:  year.tab
## X-squared = 22.807, df = 5, p-value = 0.0003674
```

```
CMHtest(year.tab)
```

```
## Cochran-Mantel-Haenszel Statistics for  by
##
##                  AltHypothesis   Chisq Df       Prob
```

```
## cor        Nonzero correlation  4.5411  1 0.03309048
## rmeans  Row mean scores differ 22.8030  5 0.00036813
## cmeans  Col mean scores differ  4.5411  1 0.03309048
## general    General association 22.8030  5 0.00036813

#DATABASES (ORDINAL CATEGORICAL PREDICTOR)
database.tab <- table(dummy$databases, dummy$held)
CMHtest(database.tab) #linear trend

## Cochran-Mantel-Haenszel Statistics for  by
##
##                  AltHypothesis Chisq Df       Prob
## cor        Nonzero correlation 319.8  1 1.6018e-71
## rmeans  Row mean scores differ 336.5  5 1.4125e-70
## cmeans  Col mean scores differ 319.8  1 1.6018e-71
## general    General association 336.5  5 1.4125e-70

#PRIOR.TRAFFIC (ORDINAL CATEGORICAL PREDICTOR)
traffic.tab <- table(dummy$prior.traffic, dummy$held)
CMHtest(traffic.tab) #no linear trend

## Cochran-Mantel-Haenszel Statistics for  by
##
##                  AltHypothesis     Chisq Df    Prob
## cor        Nonzero correlation 0.0049202  1 0.94408
## rmeans  Row mean scores differ 0.0217789  2 0.98917
## cmeans  Col mean scores differ 0.0049202  1 0.94408
## general    General association 0.0217789  2 0.98917

#AGE statistics
min(df$age)

## [1] 13

max(df$age)

## [1] 67

sd(df$age)

## [1] 8.330865

median(df$age)

## [1] 22

anova(glm(held~age, data=df, family=binomial), test='Chisq')

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: held
```

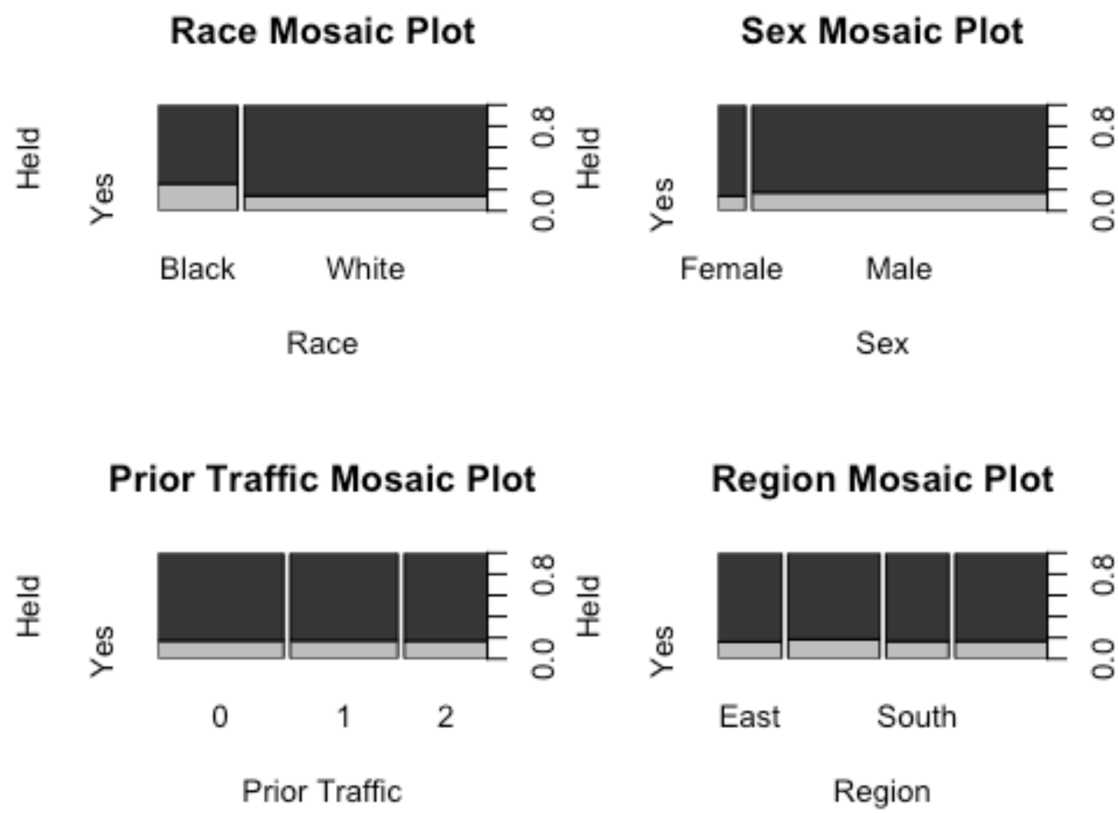
```
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  5132     4684.5
## age   1   9.9245      5131     4674.6 0.001631 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pchisq(9.9245, df=1, lower.tail=F)
```

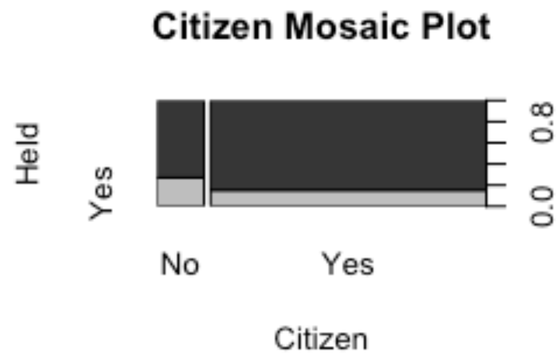
```
## [1] 0.001630931
```

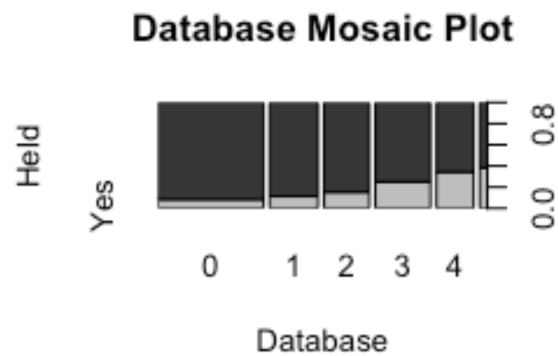
MOSAIC PLOTS

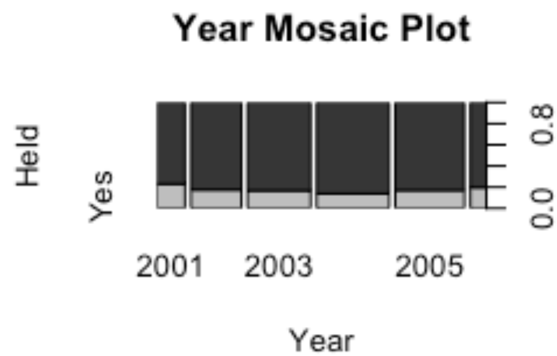
```
par(mfrow=c(2,2))

#RACE (appears dependent)
spineplot(race.tab, xlab='Race', ylab='Held', col=c('grey', 'gray21'),
main='Race Mosaic Plot')

#SEX (appears slightly independent)
spineplot(sex.tab, xlab='Sex', ylab='Held', col=c('grey', 'gray21'),
main='Sex Mosaic Plot')

#PRIOR.TRAFFIC (appears independent)
spineplot(traffic.tab, xlab='Prior Traffic', ylab='Held', col=c('grey',
'gray21'), main='Prior Traffic Mosaic Plot')

#REGION (appears independent)
spineplot(region.tab, xlab='Region', ylab='Held', col=c('grey', 'gray21'),
main='Region Mosaic Plot')
```




```r
#EMPLOYED (appears dependent)
spineplot(employed.tab, xlab='Employed', ylab='Held', col=c('grey',
'gray21'), main='Employed Mosaic Plot')

#CITIZEN (appears dependent)
spineplot(citizen.tab, xlab='Citizen', ylab='Held', col=c('grey', 'gray21'),
main='Citizen Mosaic Plot')

#DATABASES (appears dependent)
spineplot(database.tab, xlab='Database', ylab='Held', col=c('grey',
'gray21'), main='Database Mosaic Plot')

#YEAR (appears dependent)
spineplot(year.tab, xlab='Year', ylab='Held', col=c('grey', 'gray21'),
main='Year Mosaic Plot')
```







NOTES:

•RACE: association and appears dependent on held (yay) •EMPLOYED: association and appears dependent on held (yay) •CITIZEN: association and appears dependent on held (yay) •YEAR: association and appears dependent •DATABASE: exhibits a linear trend and appears dependent on held (yay)

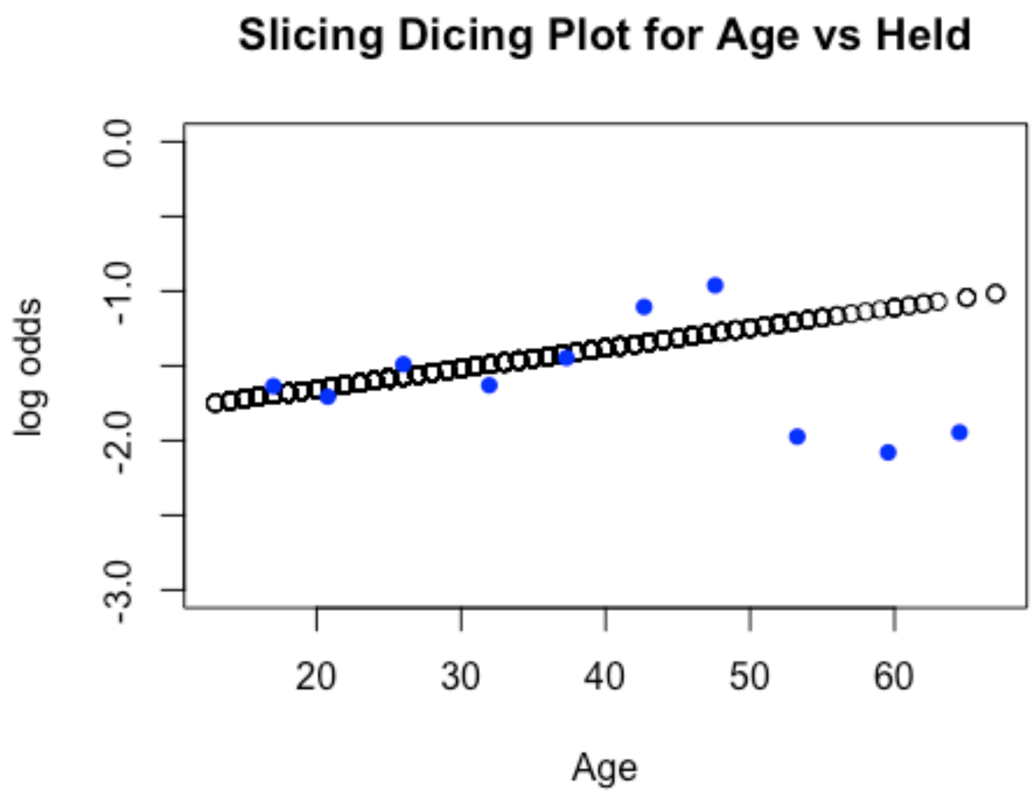
•SEX: no association and appears independent on held (yay) •REGION: no association and appears independent on held (yay) •PRIOR TRAFFIC: no linear trend and appears independent (yay)

TWO WAY ANALYSIS FOR AGE (NUMERICAL PREDICTOR)

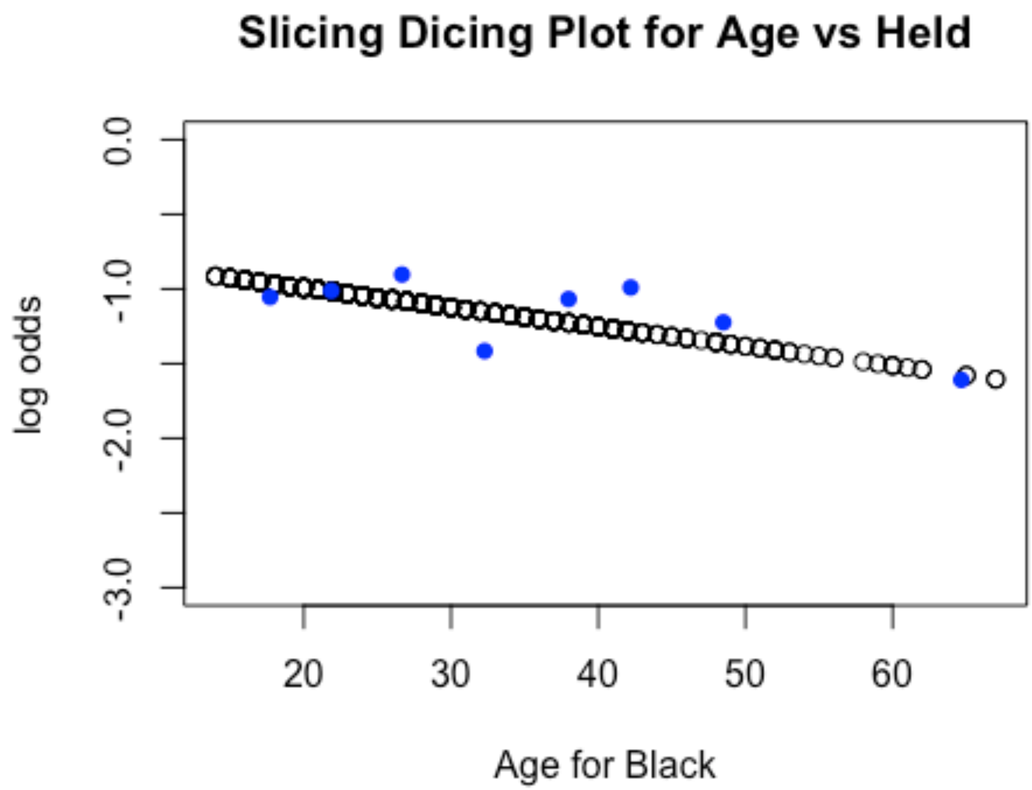
Slicing-dicing" plot of empirical log-odds

```
df$age.2 <- df$age
age.fac <- factor(cut(df$age.2,breaks=10))
eprobs <- tapply(df$held,age.fac,mean)
slice.avg <- tapply(df$age.2,age.fac,mean)
elogits <- log(eprobs/(1-eprobs))
```

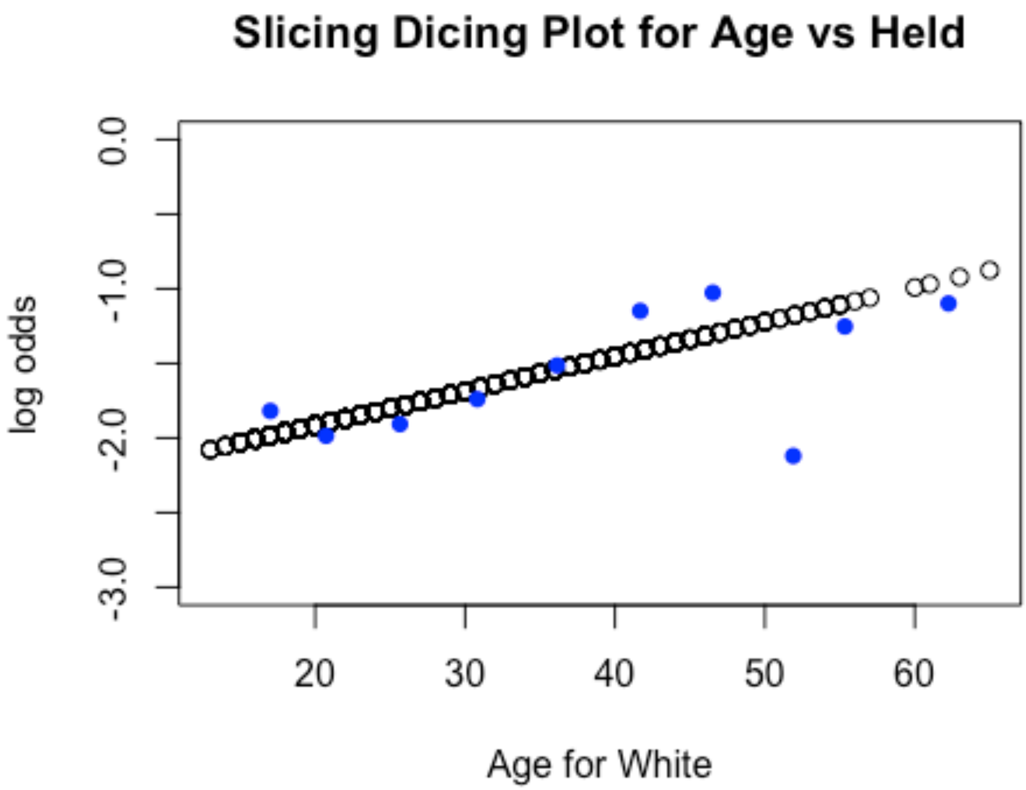
```r
outt <- glm(df$held ~ df$age.2,family="binomial")
pp <- outt$fitted.values
plogits <- log(pp/(1-pp))

plot(df$age.2,plogits,ylim=c(-3,0),xlab="Age",ylab="log odds",main="Slicing
Dicing Plot for Age vs Held")
points(slice.avg,elogits,pch=16,col="blue")
```



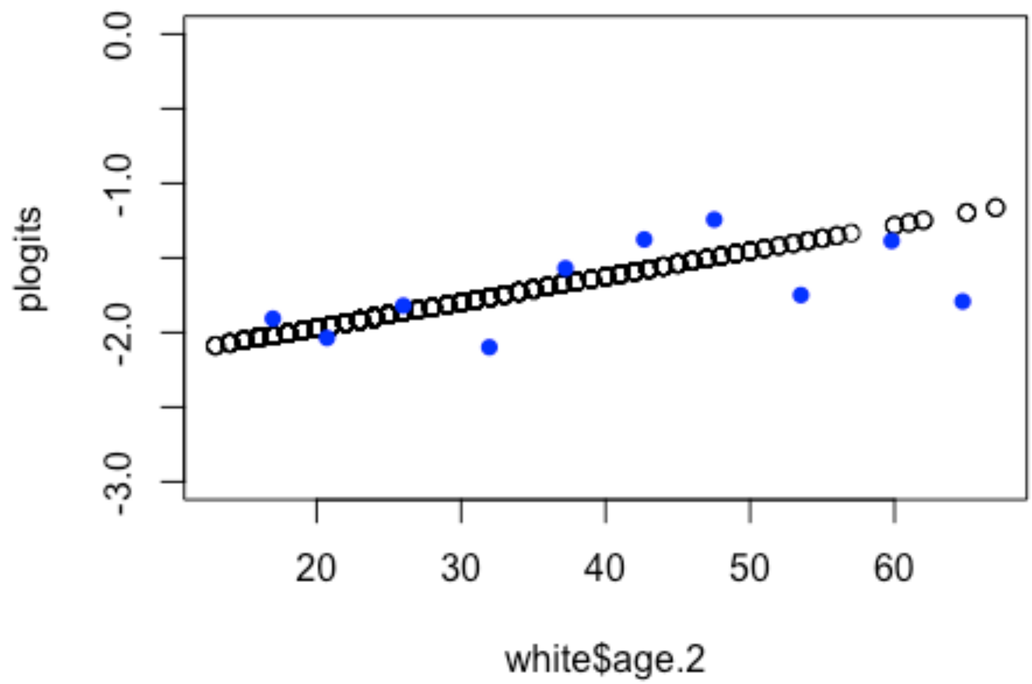

```r
# black
black <-df[df$race=="Black",]
black$age.2 <- black$age
age.fac <- factor(cut(black$age.2,breaks=10))
eprobs <- tapply(black$held,age.fac,mean)
slice.avg <- tapply(black$age.2,age.fac,mean)
elogits <- log(eprobs/(1-eprobs))

outt <- glm(black$held ~ black$age.2,family="binomial")
pp <- outt$fitted.values
plogits <- log(pp/(1-pp))

plot(black$age.2,plogits,ylim=c(-3,0),xlab="Age for Black",ylab="log
odds",main="Slicing Dicing Plot for Age vs Held")
points(slice.avg,elogits,pch=16,col="blue")
```



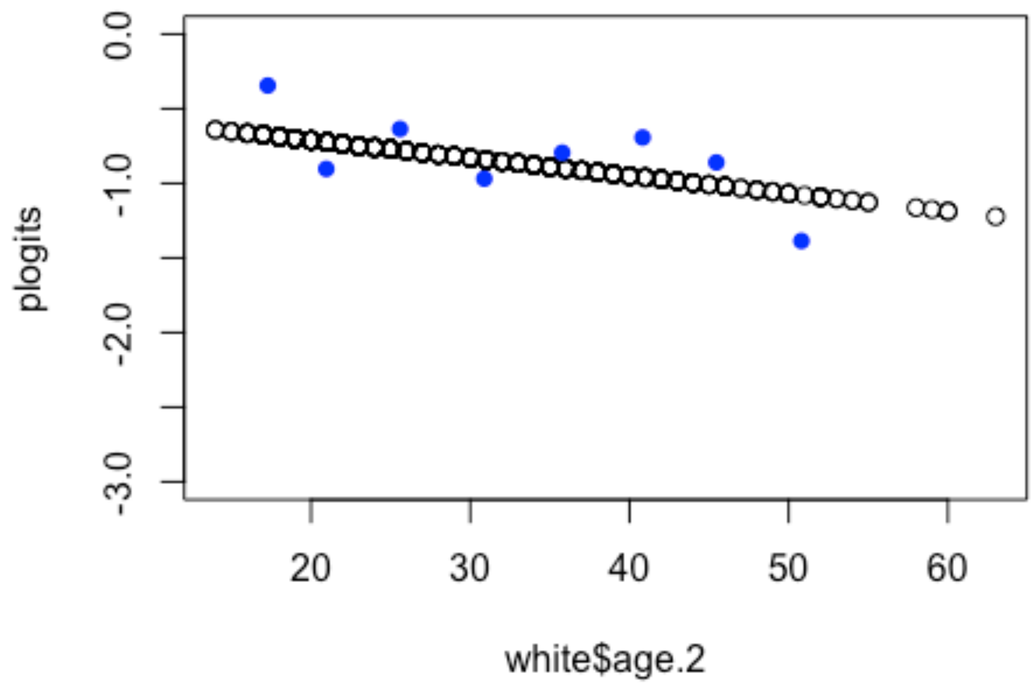

```
# white
white <- df[df$race=="White",]
white$age.2 <- white$age
age.fac <- factor(cut(white$age.2,breaks=10))
eprobs <- tapply(white$held,age.fac,mean)
slice.avg <- tapply(white$age.2,age.fac,mean)
elogits <- log(eprobs/(1-eprobs))

outt <- glm(white$held ~ white$age.2,family="binomial")
pp <- outt$fitted.values
plogits <- log(pp/(1-pp))

plot(white$age.2,plogits,ylim=c(-3,0),xlab="Age for White",ylab="log
odds",main="Slicing Dicing Plot for Age vs Held")
points(slice.avg,elogits,pch=16,col="blue")
```




```
# employed vs age (employed is Yes)
white <- df[df$employed=="Yes",]
white$age.2 <- white$age
age.fac <- factor(cut(white$age.2,breaks=10))
eprobs <- tapply(white$held,age.fac,mean)
slice.avg <- tapply(white$age.2,age.fac,mean)
elogits <- log(eprobs/(1-eprobs))

outt <- glm(white$held ~ white$age.2,family="binomial")
pp <- outt$fitted.values
plogits <- log(pp/(1-pp))

plot(white$age.2,plogits,ylim=c(-3,0))
points(slice.avg,elogits,pch=16,col="blue",xlab="Employed",ylab="log
odds",main="Slicing Dicing Plot for Employed vs Held")
```




```r
# employed vs age (employed is No)
white <- df[df$employed=="No",]
white$age.2 <- white$age
age.fac <- factor(cut(white$age.2,breaks=10))
eprobs <- tapply(white$held,age.fac,mean)
slice.avg <- tapply(white$age.2,age.fac,mean)
elogits <- log(eprobs/(1-eprobs))

outt <- glm(white$held ~ white$age.2,family="binomial")
pp <- outt$fitted.values
plogits <- log(pp/(1-pp))

plot(white$age.2,plogits,ylim=c(-3,0))
points(slice.avg,elogits,pch=16,col="blue")
```



Multivariable analyis

selection of siginficant covariates

```
#race, employed, citizen, year, database, age
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

# initial model: age, race, age*race for forward selection
initial <- glm(held~age*race, data=df, family=binomial)
AIC(initial)

## [1] 4588.771
```

```r
BIC(initial)

## [1] 4614.944

summary(initial)

##
## Call:
## glm(formula = held ~ age * race, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8344  -0.6215  -0.5355  -0.5073   2.0956
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.733099   0.208888  -3.510 0.000449 ***
## age           -0.013012   0.007833  -1.661 0.096707 .
## raceWhite     -1.645087   0.251930  -6.530 6.58e-11 ***
## age:raceWhite  0.036120   0.009391   3.846 0.000120 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4684.5  on 5132  degrees of freedom
## Residual deviance: 4580.8  on 5129  degrees of freedom
## AIC: 4588.8
##
## Number of Fisher Scoring iterations: 4

anova(initial)

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: held
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev
## NULL                      5132     4684.5
## age       1    9.925      5131     4674.6
## race      1   78.507      5130     4596.1
## age:race  1   15.331      5129     4580.8

1-pchisq(4580.8, 5129)

## [1] 1
```

```r
# Step 1
test.mod.1 <- glm(held~age+race+age*race+employed, data=df, family=binomial)
lrtest(initial, test.mod.1)

## Likelihood ratio test
##
## Model 1: held ~ age * race
## Model 2: held ~ age + race + age * race + employed
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2290.4
## 2   5 -2216.4  1 147.97  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#MODEL 2
test.mod.2 <- glm(held~age+race+age*race+citizen, data=df, family=binomial)
lrtest(initial, test.mod.2)

## Likelihood ratio test
##
## Model 1: held ~ age * race
## Model 2: held ~ age + race + age * race + citizen
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2290.4
## 2   5 -2273.2  1 34.396  4.495e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#MODEL 3
test.mod.3 <- glm(held~age+race+age*race+year, data=df, family=binomial)
lrtest(initial, test.mod.3)

## Likelihood ratio test
##
## Model 1: held ~ age * race
## Model 2: held ~ age + race + age * race + year
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2290.4
## 2   9 -2280.7  5 19.398    0.00162 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#MODEL 4
test.mod.4 <- glm(held~age+race+age:race + databases, data=df,
family=binomial)
lrtest(initial, test.mod.4) #most significant predictor

## Likelihood ratio test
##
## Model 1: held ~ age * race
## Model 2: held ~ age + race + age:race + databases
```

```
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2290.4
## 2   9 -2156.8  5 267.15  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# (chosen model with adding one variable: databases)
initial.2 <- glm(held~age+race+age:race+databases, data=df, family=binomial)
summary(initial.2)
```

```
##
## Call:
## glm(formula = held ~ age + race + age:race + databases, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1809  -0.6642  -0.4467  -0.3874   2.3200
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.487507   0.233240  -6.378 1.80e-10 ***
## age           -0.016899   0.008272  -2.043  0.04107 *
## raceWhite     -1.326595   0.265013  -5.006 5.56e-07 ***
## databases1     0.255724   0.138940   1.841  0.06569 .
## databases2     0.585641   0.132063   4.435 9.23e-06 ***
## databases3     1.166478   0.115705  10.081  < 2e-16 ***
## databases4     1.626271   0.120334  13.515  < 2e-16 ***
## databases5     1.766101   0.200054   8.828  < 2e-16 ***
## age:raceWhite  0.030699   0.009860   3.113  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4684.5  on 5132  degrees of freedom
## Residual deviance: 4313.6  on 5124  degrees of freedom
## AIC: 4331.6
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(initial.2)
```

```
## [1] 4331.617
```

```
BIC(initial.2)
```

```
## [1] 4390.508
```

```
# step 2
#MODEL 5
```

```
test.mod.5 <- glm(held~age+race+age:race+databases+employed, data=df,
family=binomial)
lrtest(initial.2, test.mod.5) #most significant predictor

## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + databases
## Model 2: held ~ age + race + age:race + databases + employed
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -2156.8
## 2  10 -2117.2  1 79.259  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

sum.5 <- summary(test.mod.5)
AIC(test.mod.5)

## [1] 4254.357

BIC(test.mod.5)

## [1] 4319.792

#MODEL 6
test.mod.6 <- glm(held~age+race+age:race+databases+citizen, data=df,
family=binomial)
lrtest(initial.2, test.mod.6)

## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + databases
## Model 2: held ~ age + race + age:race + databases + citizen
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -2156.8
## 2  10 -2138.0  1 37.549  8.916e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

nova.6 <- anova(test.mod.6, test='Chisq')
sum.6 <- summary(test.mod.6)

#MODEL 7
test.mod.7 <- glm(held~age+race+age:race+databases+year, data=df,
family=binomial)
lrtest(initial.2, test.mod.7)

## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + databases
## Model 2: held ~ age + race + age:race + databases + year
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -2156.8
```

```
## 2  14 -2148.8  5 16.126   0.006494 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
nova.7 <- anova(test.mod.7, test='Chisq')
sum.7 <- summary(test.mod.7)

# (chosen model with adding two variables: databases, employed)
initial.3 <- glm(held~age+race+age:race+databases+employed, data=df,
family=binomial)
AIC(initial.3)
```

```
## [1] 4254.357
```

```
BIC(initial.3)
```

```
## [1] 4319.792
```

```
summary(initial.3)
```

```
##
## Call:
## glm(formula = held ~ age + race + age:race + databases + employed,
##     family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3901  -0.6180  -0.4271  -0.3719   2.3461
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.789496   0.248414  -3.178 0.001482 **
## age           -0.019974   0.008403  -2.377 0.017452 *
## raceWhite     -1.265469   0.268916  -4.706 2.53e-06 ***
## databases1     0.203155   0.139952   1.452 0.146612
## databases2     0.500643   0.133321   3.755 0.000173 ***
## databases3     1.018372   0.117852   8.641  < 2e-16 ***
## databases4     1.438229   0.123089  11.684  < 2e-16 ***
## databases5     1.596279   0.203652   7.838 4.57e-15 ***
## employedYes   -0.770073   0.085001  -9.060  < 2e-16 ***
## age:raceWhite  0.029897   0.010009   2.987 0.002818 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4684.5  on 5132  degrees of freedom
## Residual deviance: 4234.4  on 5123  degrees of freedom
## AIC: 4254.4
##
## Number of Fisher Scoring iterations: 5
```

```
# step 3
#MODEL 8
test.mod.8 <- glm(held~age+race+age:race+databases+employed+citizen, data=df,
family=binomial)
lrtest(initial.3, test.mod.8) #most significant predictor
```

```
## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + databases + employed
## Model 2: held ~ age + race + age:race + databases + employed + citizen
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  10 -2117.2
## 2  11 -2101.0  1 32.343  1.292e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#MODEL 9
test.mod.9 <- glm(held~age+race+age:race+databases+employed+year, data=df,
family=binomial)
lrtest(initial.3, test.mod.9)
```

```
## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + databases + employed
## Model 2: held ~ age + race + age:race + databases + employed + year
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  10 -2117.2
## 2  15 -2110.7  5 12.938    0.02397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
initial.4 <- glm(held~age+race+age:race+employed+citizen, data=df,
family=binomial)
BIC(initial.4)
```

```
## [1] 4456.131
```

```
AIC(initial.4)
```

```
## [1] 4416.87
```

```
# step 4
#MODEL 10
# (chosen model with adding four variables: databases, employed, citizen,
year variable is not important remove)
test.mod.10 <- glm(held~age+race+age:race+databases+employed+citizen+year,
data=df, family=binomial)
lrtest(initial.4, test.mod.10) #year is not important
```

```
## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + employed + citizen
```

```
## Model 2: held ~ age + race + age:race + databases + employed + citizen +
##     year
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -2202.4
## 2  16 -2097.9 10 209.16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(test.mod.10)
```

```
## [1] 4227.714
```

```
BIC(test.mod.10)
```

```
## [1] 4332.409
```

```
summary(test.mod.10)
```

```
##
## Call:
## glm(formula = held ~ age + race + age:race + databases + employed +
##     citizen + year, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5547  -0.6149  -0.4367  -0.3594   2.4659
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.283747   0.277879  -1.021 0.307199
## age           -0.025766   0.008642  -2.982 0.002868 **
## raceWhite     -1.283737   0.272292  -4.715 2.42e-06 ***
## databases1     0.238576   0.140651   1.696 0.089843 .
## databases2     0.519292   0.134165   3.871 0.000109 ***
## databases3     1.059609   0.118665   8.929  < 2e-16 ***
## databases4     1.474683   0.124018  11.891  < 2e-16 ***
## databases5     1.614343   0.204671   7.887 3.08e-15 ***
## employedYes   -0.742040   0.085506  -8.678  < 2e-16 ***
## citizenYes    -0.593060   0.115166  -5.150 2.61e-07 ***
## year2002       0.065834   0.163692   0.402 0.687550
## year2003       0.017145   0.158390   0.108 0.913802
## year2004      -0.165345   0.158843  -1.041 0.297907
## year2005       0.055231   0.156754   0.352 0.724585
## year2006       0.203973   0.209570   0.973 0.330407
## age:raceWhite  0.035251   0.010219   3.449 0.000562 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4684.5  on 5132  degrees of freedom
```

```
## Residual deviance: 4195.7  on 5117  degrees of freedom
## AIC: 4227.7
##
## Number of Fisher Scoring iterations: 5
```

```
# final model
final.mod <- glm(held~race+employed+citizen+databases+race*age+age*employed,
data=df, family=binomial)
sum.fin <- summary(final.mod)
nova.fin <- anova(final.mod, test='Chisq')
lrtest(initial.4, final.mod)
```

```
## Likelihood ratio test
##
## Model 1: held ~ age + race + age:race + employed + citizen
## Model 2: held ~ race + employed + citizen + databases + race * age + age *
##     employed
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -2202.4
## 2  12 -2098.1  6 208.67  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sum.fin
```

```
##
## Call:
## glm(formula = held ~ race + employed + citizen + databases +
##     race * age + age * employed, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6014  -0.6079  -0.4460  -0.3499   2.4109
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.108776   0.308911   0.352 0.724744
## raceWhite       -1.247941   0.272717  -4.576 4.74e-06 ***
## employedYes     -1.363239   0.269266  -5.063 4.13e-07 ***
## citizenYes      -0.591583   0.101680  -5.818 5.95e-09 ***
## databases1       0.232389   0.140565   1.653 0.098281 .
## databases2       0.504838   0.134111   3.764 0.000167 ***
## databases3       1.055110   0.118412   8.911  < 2e-16 ***
## databases4       1.456976   0.123826  11.766  < 2e-16 ***
## databases5       1.602353   0.204726   7.827 5.00e-15 ***
## age             -0.039674   0.010366  -3.827 0.000129 ***
## raceWhite:age    0.033632   0.010173   3.306 0.000946 ***
## employedYes:age  0.023475   0.009773   2.402 0.016300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4684.5  on 5132  degrees of freedom
## Residual deviance: 4196.2  on 5121  degrees of freedom
## AIC: 4220.2
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(final.mod)
```

```
## [1] 4220.2
```

```
BIC(final.mod)
```

```
## [1] 4298.721
```

```
#interaction.plot(df$age,df$employed, df$held)
#interaction.plot(df$age,df$race, df$held)
```

Final Model

Assessment of the overall goodness of fit of the models

classification table, goodness-of-fit test

```
final.mod <- glm(held~race+employed+citizen+databases+race*age+age*employed,
data=df, family=binomial)

summary(final.mod)
```

```
##
## Call:
## glm(formula = held ~ race + employed + citizen + databases +
##     race * age + age * employed, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6014  -0.6079  -0.4460  -0.3499   2.4109
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.108776   0.308911   0.352 0.724744
## raceWhite       -1.247941   0.272717  -4.576 4.74e-06 ***
## employedYes     -1.363239   0.269266  -5.063 4.13e-07 ***
## citizenYes      -0.591583   0.101680  -5.818 5.95e-09 ***
## databases1       0.232389   0.140565   1.653 0.098281 .
## databases2       0.504838   0.134111   3.764 0.000167 ***
## databases3       1.055110   0.118412   8.911  < 2e-16 ***
```

```
## databases4       1.456976   0.123826  11.766  < 2e-16 ***
## databases5       1.602353   0.204726   7.827 5.00e-15 ***
## age             -0.039674   0.010366  -3.827 0.000129 ***
## raceWhite:age    0.033632   0.010173   3.306 0.000946 ***
## employedYes:age  0.023475   0.009773   2.402 0.016300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4684.5  on 5132  degrees of freedom
## Residual deviance: 4196.2  on 5121  degrees of freedom
## AIC: 4220.2
##
## Number of Fisher Scoring iterations: 5
```

```
anova(final.mod)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: held
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                          5132     4684.5
## race          1   81.940      5131     4602.6
## employed      1  155.421      5130     4447.2
## citizen       1   24.730      5129     4422.4
## databases     5  207.497      5124     4214.9
## age           1    0.114      5123     4214.8
## race:age      1   12.817      5122     4202.0
## employed:age  1    5.815      5121     4196.2
```

```
options(digits=18)
chisq <- pchisq(4196.2,5121)
chisq
```

```
## [1] 9.41888492261092775e-23
```

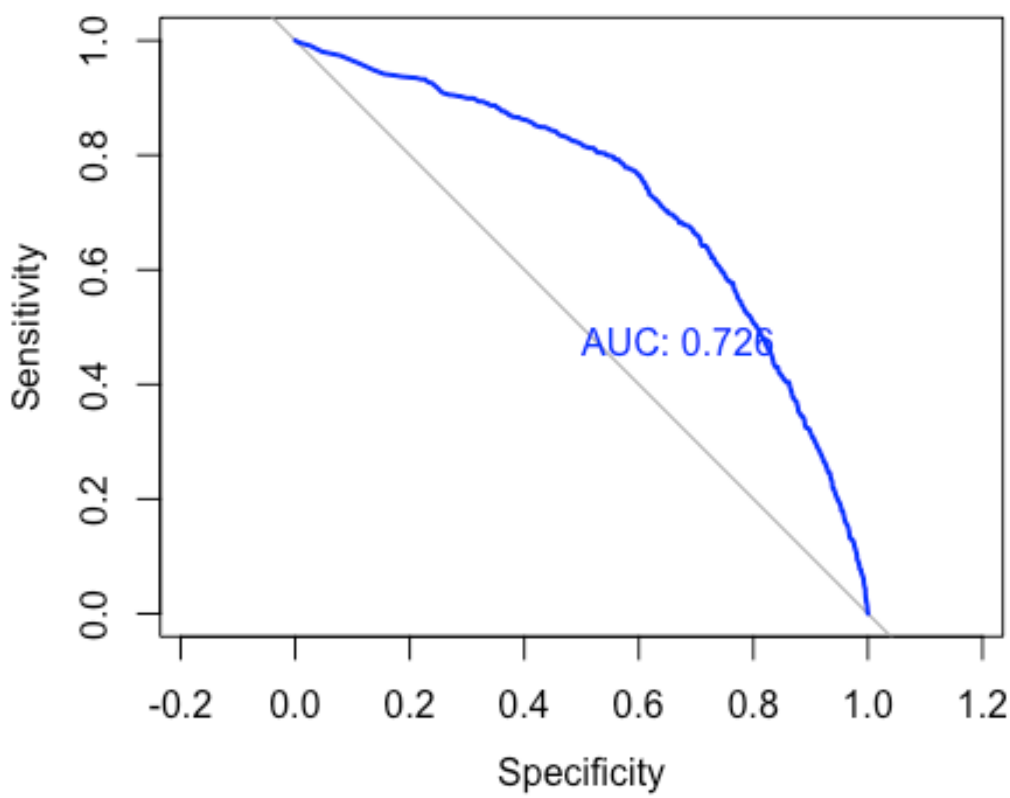
```
1-chisq
```

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

ROC curve and Classification Tables

```
library(pROC)
```

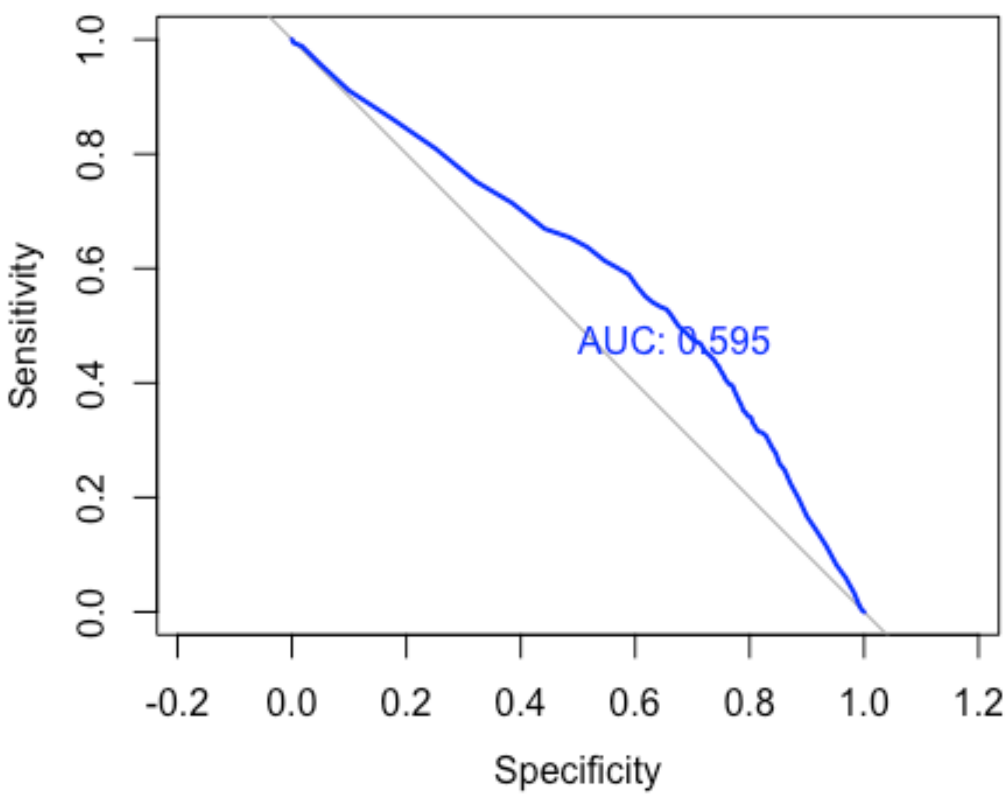
```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

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

#classification table
class_table <- function(model) {
  yprobs <- model$fitted
  yhat <- as.numeric(yprobs > 0.5)
  x <- table(df$held,yhat)
  plot.roc(df$held,yprobs,print.auc=TRUE,col="blue",xlim=c(0,1))
  return(x)
}

class_table(final.mod)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



```
##     yhat
##         0    1
##   0 4207   52
##   1  807   67
```

```
class_table(initial)
```

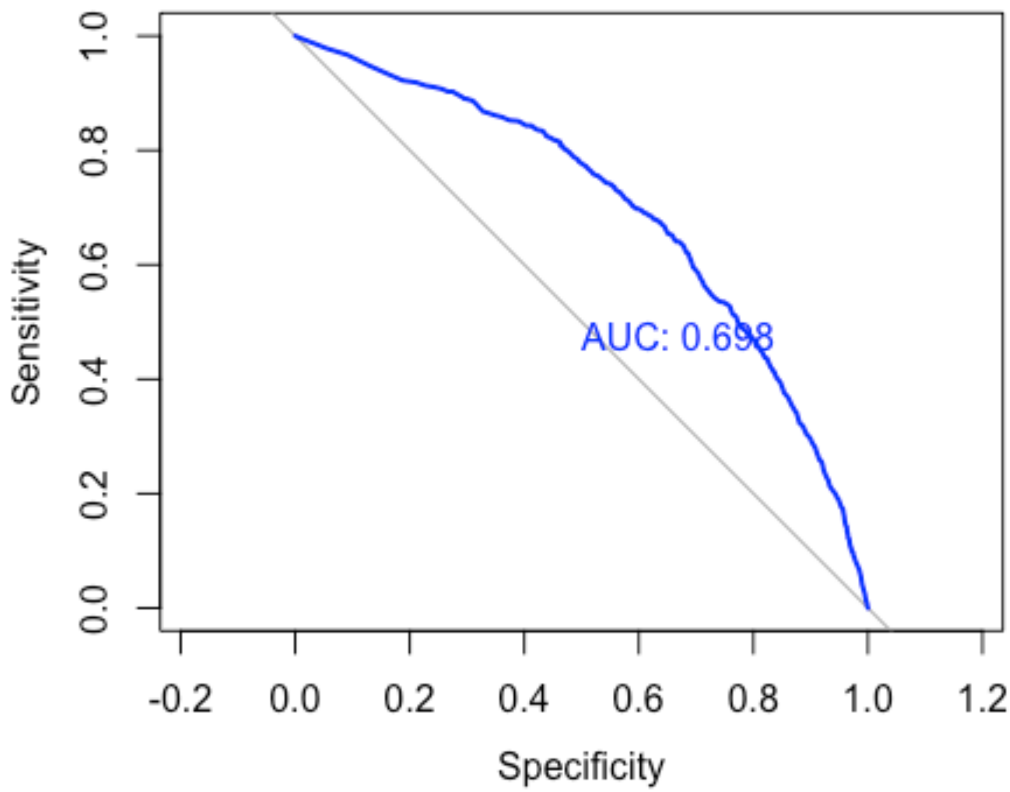
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



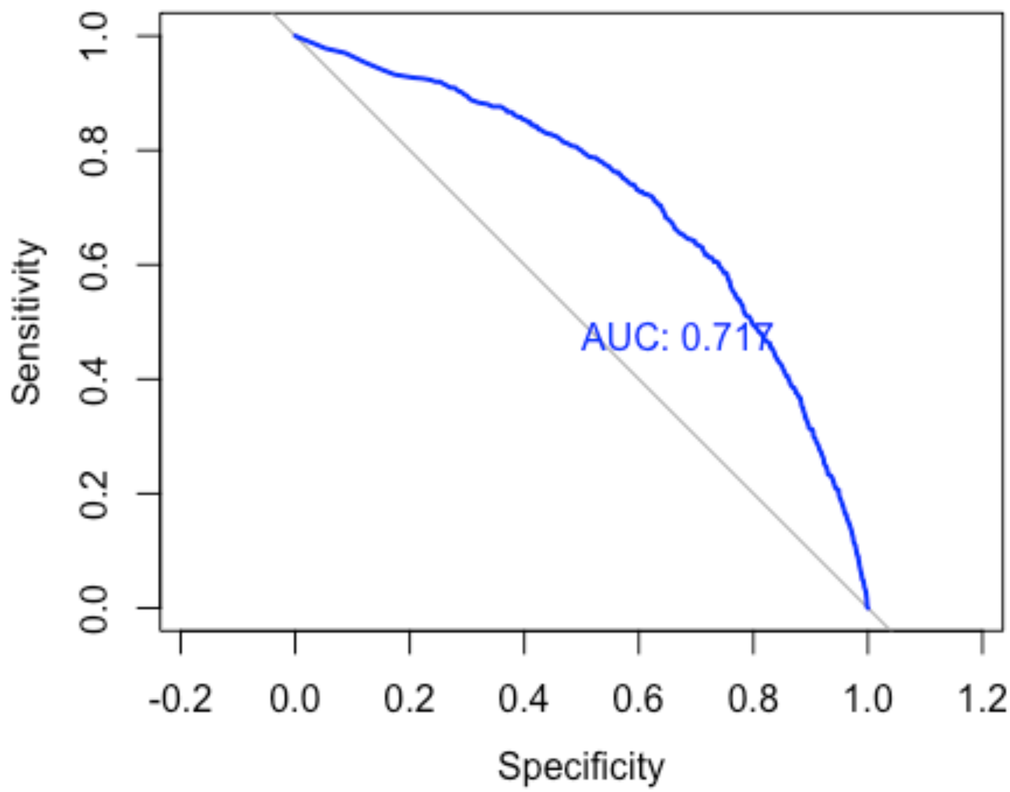

```
##     yhat
##         0
##   0 4259
##   1  874
```

```
class_table(initial.2)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



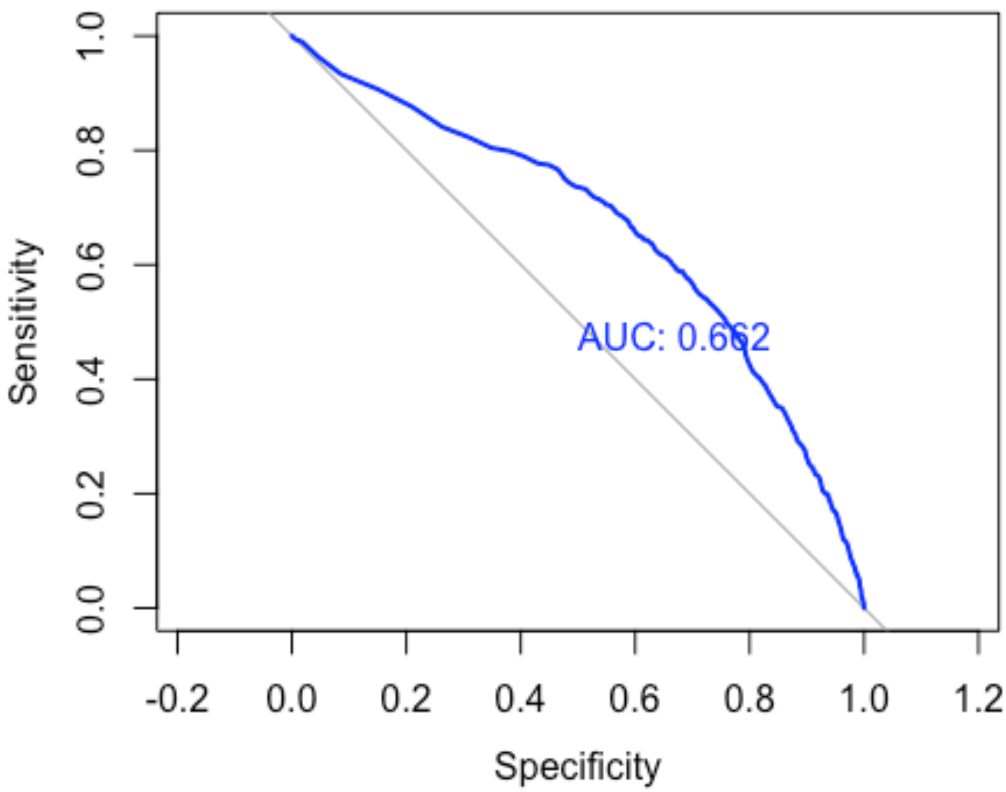

```
##    yhat
##        0    1
##   0 4255    4
##   1  871    3

class_table(initial.3)

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
##     yhat
##         0    1
##   0 4216   43
##   1  828   46
```

```
class_table(initial.4)
```

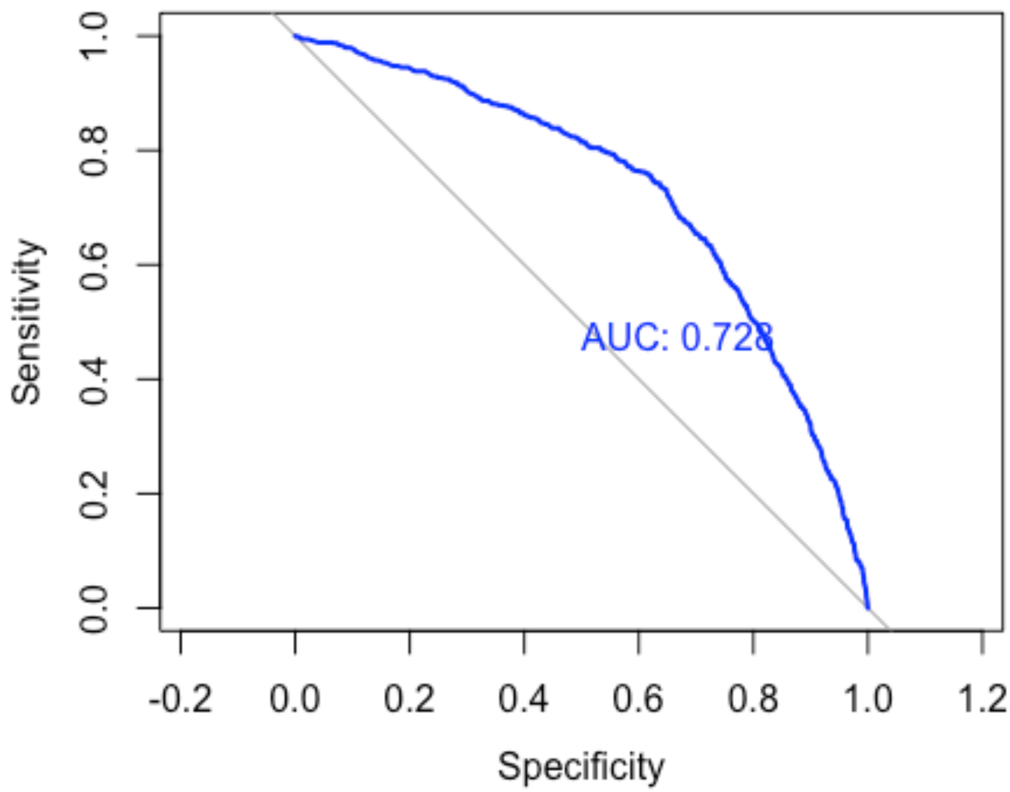
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```




```
##    yhat
##        0    1
##   0 4230   29
##   1  843   31
```

```
class_table(test.mod.10)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

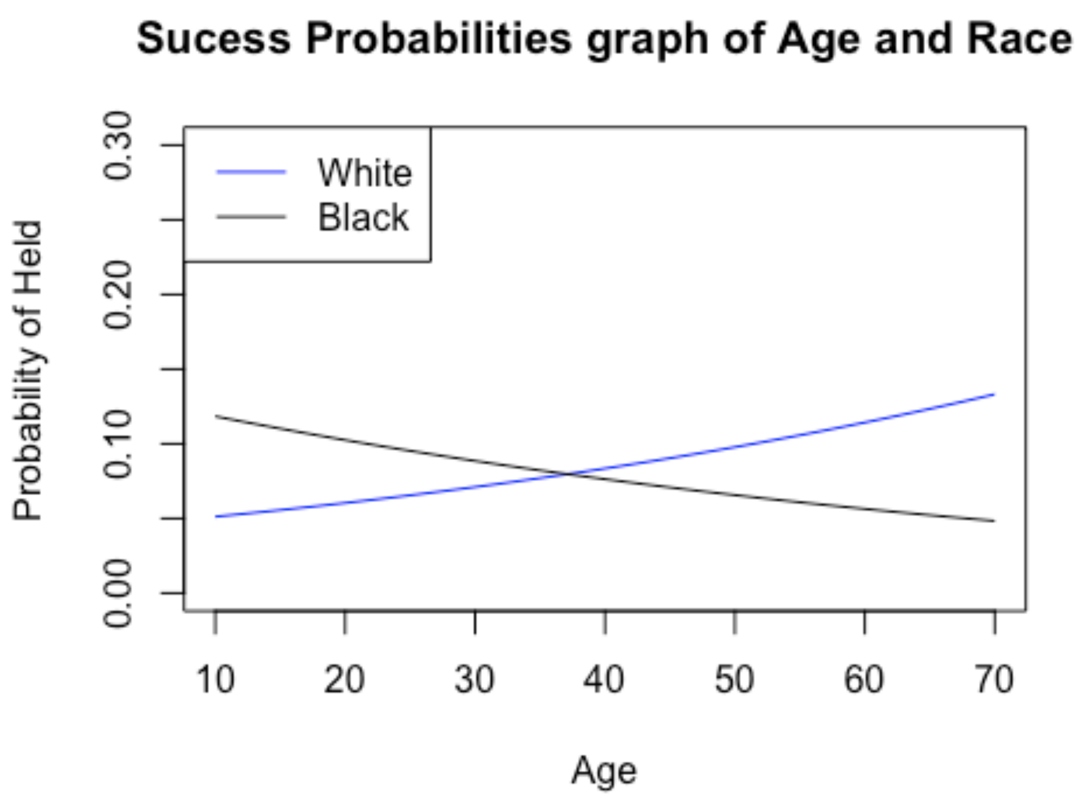



```
##    yhat
##        0    1
##   0 4210   49
##   1  811   63
```

Final model Sucess Probabilities

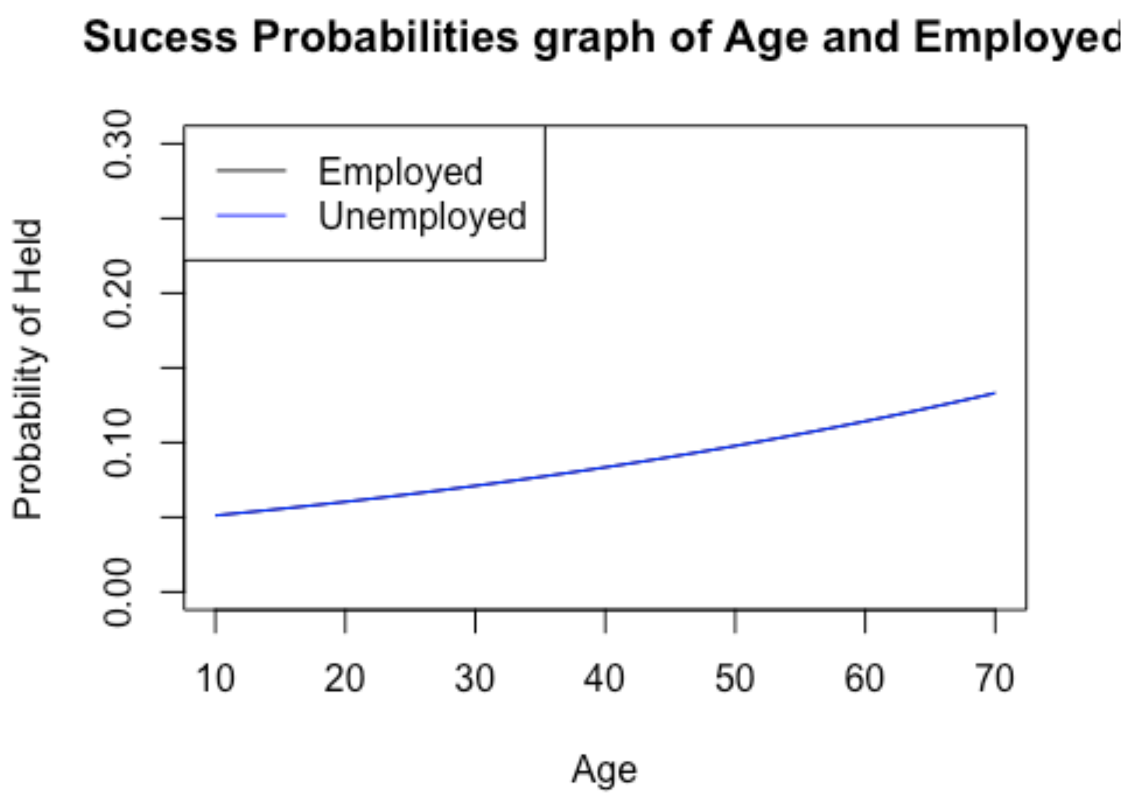
```
# Mode for categorical data
# race: white
# citizen: Yes
# databases: 0
# employed: Yes
beta0 <- final.mod$coefficients[1] # intercept
beta1 <- final.mod$coefficients[2] # raceWhite
beta2 <- final.mod$coefficients[3] # employedYes
beta3 <- final.mod$coefficients[4] # citizenYes
beta4 <- final.mod$coefficients[5] # databases1
beta5 <- final.mod$coefficients[6] # databases2
beta6 <- final.mod$coefficients[7] # databases3
beta7 <- final.mod$coefficients[8] # databases4
beta8 <- final.mod$coefficients[9] # databases5
beta9 <- final.mod$coefficients[10] # age
beta10 <- final.mod$coefficients[11] # raceWhite:age
```

```
beta11 <- final.mod$coefficients[12] # EmployedYes:age

# Race vs Age
# White
curve(expr =
exp(beta0+beta1+beta2+beta3+beta9*x+beta10*x+beta11*x)/(1+exp(beta0+beta1+bet
a2+beta3+beta9*x+beta10*x+beta11*x)),
          xlim=c(10,70),ylim=c(0,0.3), main = "Sucess Probabilities graph of
Age and Race",
xlab="Age", ylab="Probability of Held",col="blue")
# Black
curve(expr =
exp(beta0+beta2+beta3+beta9*x+beta11*x)/(1+exp(beta0+beta2+beta3+beta9*x+beta
11*x)),
          xlim=c(10,70),ylim=c(0,0.3),add=TRUE)

legend("topleft",
      legend = c("White","Black"),
      lty=1:1,
      col = c("Blue","Black"))
```




```
# Employed vs Age
# Employed
```

```
curve(expr =
exp(beta0+beta1+beta2+beta3+beta9*x+beta10*x+beta11*x)/(1+exp(beta0+beta1+bet
a2+beta3+beta9*x+beta10*x+beta11*x)),
          xlim=c(10,70),ylim=c(0,0.3), main = "Sucess Probabilities graph of
Age and Employed",
xlab="Age", ylab="Probability of Held")
# Not employed
curve(expr =
exp(beta0+beta1+beta2+beta3+beta9*x+beta10*x+beta11*x)/(1+exp(beta0+beta1+bet
a2+beta3+beta9*x+beta10*x+beta11*x)),
          xlim=c(10,70),ylim=c(0,0.3), col="blue",add=TRUE)

legend("topleft",
      legend = c("Employed","Unemployed"),
      lty=1:1,
      col = c("Black","Blue"))
```




```
# Citizen vs Age
# Citizen
curve(expr =
exp(beta0+beta1+beta2+beta3+beta9*x+beta10*x+beta11*x)/(1+exp(beta0+beta1+bet
a2+beta3+beta9*x+beta10*x+beta11*x)),
          xlim=c(10,70),ylim=c(0,0.3), main = "Sucess Probabilities graph of
```

```
Age and Citizen",
xlab="Age", ylab="Probability of Held")
# Non citizen
curve(expr =
exp(beta0+beta1+beta9*x+beta10*x)/(1+exp(beta0+beta1+beta9*x+beta10*x)),
          xlim=c(10,70),ylim=c(0,0.3), col="blue",add=TRUE)

legend("topleft",
      legend = c("Citizen","Non-citizen"),
      lty=1:1,
      col = c("Black","Blue"))
```