4198 HW7

### Q1.

adult <-read.csv("~/desktop/adult.csv")  
adult <- adult[1:10000,]  
  
# Prepare the data  
levels(adult$marital.status)

## [1] "Divorced" "Married-AF-spouse" "Married-civ-spouse"   
## [4] "Married-spouse-absent" "Never-married" "Separated"   
## [7] "Widowed"

levels(adult$marital.status)[2:4] <- "Married"  
levels(adult$marital.status)

## [1] "Divorced" "Married" "Never-married" "Separated"   
## [5] "Widowed"

adult$ms.married <- ifelse(adult$marital.status == "Married", 1, 0)  
adult$ms.neverm <- ifelse(adult$marital.status == "Never-married", 1, 0)  
adult$ms.sep <- ifelse(adult$marital.status == "Separated", 1, 0)  
adult$ms.widowed <- ifelse(adult$marital.status == "Widowed", 1, 0)

### 2.

levels(adult$race)

## [1] "Amer-Indian-Eskimo" "Asian-Pac-Islander" "Black"   
## [4] "Other" "White"

adult$race.White <- ifelse(adult$race == "White", 1, 0)  
adult$race.Black <- ifelse(adult$race == "Black", 1, 0)

### 3.

adult$capnet <- adult$capital.gain-adult$capital.loss  
levels(adult$sex)

## [1] "Female" "Male"

adult$male <- ifelse(adult$sex == "Male", 1, 0)  
summary(adult)

## age workclass demogweight   
## Min. :17.00 Private :6947 Min. : 19302   
## 1st Qu.:28.00 Self-emp-not-inc: 803 1st Qu.: 118492   
## Median :37.00 Local-gov : 645 Median : 179126   
## Mean :38.45 ? : 585 Mean : 190679   
## 3rd Qu.:47.00 State-gov : 400 3rd Qu.: 239394   
## Max. :90.00 Self-emp-inc : 345 Max. :1226583   
## (Other) : 275   
## education education.num marital.status  
## HS-grad :3232 Min. : 1.00 Divorced :1385   
## Some-college:2305 1st Qu.: 9.00 Married :4691   
## Bachelors :1630 Median :10.00 Never-married:3311   
## Masters : 531 Mean :10.08 Separated : 321   
## Assoc-voc : 416 3rd Qu.:12.00 Widowed : 292   
## 11th : 363 Max. :16.00   
## (Other) :1523   
## occupation relationship race   
## Prof-specialty :1257 Husband :3999 Amer-Indian-Eskimo: 99   
## Exec-managerial:1211 Not-in-family :2589 Asian-Pac-Islander: 309   
## Craft-repair :1207 Other-relative: 295 Black : 953   
## Adm-clerical :1187 Own-child :1567 Other : 83   
## Sales :1179 Unmarried :1054 White :8556   
## Other-service :1028 Wife : 496   
## (Other) :2931   
## sex capital.gain capital.loss hours.per.week   
## Female:3297 Min. : 0 Min. : 0.00 Min. : 1.00   
## Male :6703 1st Qu.: 0 1st Qu.: 0.00 1st Qu.:40.00   
## Median : 0 Median : 0.00 Median :40.00   
## Mean : 1058 Mean : 87.67 Mean :40.53   
## 3rd Qu.: 0 3rd Qu.: 0.00 3rd Qu.:45.00   
## Max. :99999 Max. :4356.00 Max. :99.00   
##   
## native.country income ms.married ms.neverm   
## United-States:8930 <=50K.:7621 Min. :0.0000 Min. :0.0000   
## Mexico : 208 >50K. :2379 1st Qu.:0.0000 1st Qu.:0.0000   
## ? : 181 Median :0.0000 Median :0.0000   
## Philippines : 51 Mean :0.4691 Mean :0.3311   
## Canada : 47 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Germany : 45 Max. :1.0000 Max. :1.0000   
## (Other) : 538   
## ms.sep ms.widowed race.White race.Black   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :1.0000 Median :0.0000   
## Mean :0.0321 Mean :0.0292 Mean :0.8556 Mean :0.0953   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
##   
## capnet male   
## Min. :-4356.0 Min. :0.0000   
## 1st Qu.: 0.0 1st Qu.:0.0000   
## Median : 0.0 Median :1.0000   
## Mean : 970.4 Mean :0.6703   
## 3rd Qu.: 0.0 3rd Qu.:1.0000   
## Max. :99999.0 Max. :1.0000   
##

newdata<-adult[,c(1,5,13,15,16,17,18,19,20,21,22,23)]  
summary(newdata)

## age education.num hours.per.week income   
## Min. :17.00 Min. : 1.00 Min. : 1.00 <=50K.:7621   
## 1st Qu.:28.00 1st Qu.: 9.00 1st Qu.:40.00 >50K. :2379   
## Median :37.00 Median :10.00 Median :40.00   
## Mean :38.45 Mean :10.08 Mean :40.53   
## 3rd Qu.:47.00 3rd Qu.:12.00 3rd Qu.:45.00   
## Max. :90.00 Max. :16.00 Max. :99.00   
## ms.married ms.neverm ms.sep ms.widowed   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.4691 Mean :0.3311 Mean :0.0321 Mean :0.0292   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## race.White race.Black capnet male   
## Min. :0.0000 Min. :0.0000 Min. :-4356.0 Min. :0.0000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 0.0 1st Qu.:0.0000   
## Median :1.0000 Median :0.0000 Median : 0.0 Median :1.0000   
## Mean :0.8556 Mean :0.0953 Mean : 970.4 Mean :0.6703   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.: 0.0 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :99999.0 Max. :1.0000

### 4.

newdata$income\_g50K=c(rep(0, length(newdata$income)))  
  
for (i in 1:length(newdata$income)) {  
if(newdata$income[i]==">50K.")  
newdata$income\_g50K[i]<-1  
}

### 5.

newdata <- as.data.frame(newdata[,-4])

### 6.

newdat\_training<-newdata[1:8000,]  
newdat\_testing<-newdata[8001:10000,]

### Q7.

logistic.model <- glm(income\_g50K~., data = newdat\_training,   
family = binomial())  
  
summary(logistic.model)

##   
## Call:  
## glm(formula = income\_g50K ~ ., family = binomial(), data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9533 -0.5647 -0.2479 -0.0728 3.2560   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.455e+00 3.404e-01 -27.773 < 2e-16 \*\*\*  
## age 2.879e-02 2.958e-03 9.730 < 2e-16 \*\*\*  
## education.num 3.530e-01 1.498e-02 23.559 < 2e-16 \*\*\*  
## hours.per.week 3.329e-02 2.984e-03 11.158 < 2e-16 \*\*\*  
## ms.married 2.112e+00 1.289e-01 16.383 < 2e-16 \*\*\*  
## ms.neverm -4.366e-01 1.589e-01 -2.747 0.006019 \*\*   
## ms.sep 1.549e-01 2.745e-01 0.564 0.572481   
## ms.widowed -2.051e-01 3.089e-01 -0.664 0.506755   
## race.White 5.835e-01 1.705e-01 3.422 0.000621 \*\*\*  
## race.Black 2.527e-01 2.137e-01 1.183 0.236885   
## capnet 2.089e-04 1.503e-05 13.896 < 2e-16 \*\*\*  
## male 5.601e-02 9.286e-02 0.603 0.546401   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5666.2 on 7988 degrees of freedom  
## AIC: 5690.2  
##   
## Number of Fisher Scoring iterations: 6

estimated\_income\_g50K<- predict(logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1148836

presdicted.testing <- predict(logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1094807

The training missclasification rate is 11.48836%, and the testing missclasfication rate is 10.94807%.

### Q8.

logistic.model <- glm(income\_g50K~ poly(age, 2) + poly(education.num, 2)  
+ poly(hours.per.week, 2) + poly(capnet,2) +ms.married + ms.neverm+   
 ms.sep + ms.widowed + race.White+race.Black+male , data = newdat\_training,   
family = binomial())  
summary(logistic.model)

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 2) + ms.married +   
## ms.neverm + ms.sep + ms.widowed + race.White + race.Black +   
## male, family = binomial(), data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9820 -0.5394 -0.2248 -0.0356 3.5233   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.66288 0.22275 -16.444 < 2e-16 \*\*\*  
## poly(age, 2)1 55.96214 4.27290 13.097 < 2e-16 \*\*\*  
## poly(age, 2)2 -52.95467 4.76543 -11.112 < 2e-16 \*\*\*  
## poly(education.num, 2)1 74.85276 3.61609 20.700 < 2e-16 \*\*\*  
## poly(education.num, 2)2 9.83209 3.71570 2.646 0.008143 \*\*   
## poly(hours.per.week, 2)1 39.28111 4.05978 9.676 < 2e-16 \*\*\*  
## poly(hours.per.week, 2)2 -20.73340 3.49280 -5.936 2.92e-09 \*\*\*  
## poly(capnet, 2)1 141.73262 68.23161 2.077 0.037781 \*   
## poly(capnet, 2)2 -0.45734 25.70578 -0.018 0.985805   
## ms.married 2.27078 0.13163 17.251 < 2e-16 \*\*\*  
## ms.neverm -0.01739 0.16321 -0.107 0.915147   
## ms.sep 0.24061 0.27782 0.866 0.386457   
## ms.widowed 0.29861 0.31163 0.958 0.337960   
## race.White 0.60947 0.17534 3.476 0.000509 \*\*\*  
## race.Black 0.19093 0.21793 0.876 0.380973   
## male 0.03238 0.09559 0.339 0.734798   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5447.5 on 7984 degrees of freedom  
## AIC: 5479.5  
##   
## Number of Fisher Scoring iterations: 9

estimated\_income\_g50K<- predict(logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1101535

presdicted.testing <- predict(logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1057526

The training missclasification rate is 11.01535%, and the testing missclasfication rate is 10.57526%. this model is better than before as its missclasification rate is smaller.

### Q9.

logistic.model <- glm(income\_g50K~ poly(age, 2) + poly(education.num, 2)  
+ poly(hours.per.week, 2) + poly(capnet,1) +ms.married + ms.neverm+   
 ms.sep + ms.widowed + race.White+race.Black+male +  
race.White:male + male:ms.married +age:male + education.num:male +   
hours.per.week:male + age:race.White+ education.num:race.White +   
hours.per.week:race.White  
, data = newdat\_training,   
family = binomial())  
summary(logistic.model)

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## ms.neverm + ms.sep + ms.widowed + race.White + race.Black +   
## male + race.White:male + male:ms.married + age:male + education.num:male +   
## hours.per.week:male + age:race.White + education.num:race.White +   
## hours.per.week:race.White, family = binomial(), data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0518 -0.5363 -0.2019 -0.0279 3.4138   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.476743 0.330136 -13.560 < 2e-16 \*\*\*  
## poly(age, 2)1 85.237243 15.756514 5.410 6.31e-08 \*\*\*  
## poly(age, 2)2 -51.257721 4.771924 -10.742 < 2e-16 \*\*\*  
## poly(education.num, 2)1 72.166705 12.597857 5.728 1.01e-08 \*\*\*  
## poly(education.num, 2)2 10.258234 3.755778 2.731 0.006308 \*\*   
## poly(hours.per.week, 2)1 50.661025 13.864700 3.654 0.000258 \*\*\*  
## poly(hours.per.week, 2)2 -26.543230 3.922008 -6.768 1.31e-11 \*\*\*  
## poly(capnet, 1) 144.549866 9.925965 14.563 < 2e-16 \*\*\*  
## ms.married 3.147545 0.197401 15.945 < 2e-16 \*\*\*  
## ms.neverm -0.139120 0.166262 -0.837 0.402733   
## ms.sep 0.205013 0.286457 0.716 0.474185   
## ms.widowed 0.382074 0.322423 1.185 0.236014   
## race.White 2.083525 0.979387 2.127 0.033389 \*   
## race.Black 0.188812 0.228765 0.825 0.409172   
## male 1.002051 0.813804 1.231 0.218204   
## race.White:male -0.296503 0.278472 -1.065 0.286990   
## ms.married:male -1.361480 0.207739 -6.554 5.61e-11 \*\*\*  
## male:age -0.003106 0.008630 -0.360 0.718933   
## male:education.num -0.069256 0.041644 -1.663 0.096301 .   
## male:hours.per.week 0.024375 0.009290 2.624 0.008695 \*\*   
## race.White:age -0.020940 0.011501 -1.821 0.068650 .   
## race.White:education.num 0.082483 0.046563 1.771 0.076491 .   
## race.White:hours.per.week -0.028159 0.011125 -2.531 0.011368 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5374.4 on 7977 degrees of freedom  
## AIC: 5420.4  
##   
## Number of Fisher Scoring iterations: 7

estimated\_income\_g50K<- predict(logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1087831

presdicted.testing <- predict(logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1063782

The training missclassification rate is 10.87831%, and the testing missclassfication rate is 10.63782%. this model is better than before as its missclasification rate is even smaller.

### Q10.

library(MASS)  
set.seed(10)  
birthwt.step <- stepAIC(logistic.model, trace = 1, direction="both")

## Start: AIC=5420.43  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.neverm + ms.sep +   
## ms.widowed + race.White + race.Black + male + race.White:male +   
## male:ms.married + age:male + education.num:male + hours.per.week:male +   
## age:race.White + education.num:race.White + hours.per.week:race.White  
##   
## Df Deviance AIC  
## - male:age 1 5374.6 5418.6  
## - ms.sep 1 5374.9 5418.9  
## - race.Black 1 5375.1 5419.1  
## - ms.neverm 1 5375.1 5419.1  
## - race.White:male 1 5375.6 5419.6  
## - ms.widowed 1 5375.8 5419.8  
## <none> 5374.4 5420.4  
## - male:education.num 1 5377.3 5421.3  
## - race.White:education.num 1 5377.5 5421.5  
## - race.White:age 1 5377.8 5421.8  
## - race.White:hours.per.week 1 5381.1 5425.1  
## - male:hours.per.week 1 5381.5 5425.5  
## - poly(education.num, 2) 2 5421.0 5463.0  
## - ms.married:male 1 5419.8 5463.8  
## - poly(hours.per.week, 2) 2 5440.9 5482.9  
## - poly(age, 2) 2 5535.2 5577.2  
## - poly(capnet, 1) 1 5719.9 5763.9  
##   
## Step: AIC=5418.56  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.neverm + ms.sep +   
## ms.widowed + race.White + race.Black + male + race.White:male +   
## ms.married:male + male:education.num + male:hours.per.week +   
## race.White:age + race.White:education.num + race.White:hours.per.week  
##   
## Df Deviance AIC  
## - ms.sep 1 5375.0 5417.0  
## - race.Black 1 5375.2 5417.2  
## - ms.neverm 1 5375.3 5417.3  
## - race.White:male 1 5375.7 5417.7  
## - ms.widowed 1 5376.1 5418.1  
## <none> 5374.6 5418.6  
## - male:education.num 1 5377.3 5419.3  
## - race.White:education.num 1 5377.6 5419.6  
## - race.White:age 1 5378.0 5420.0  
## + male:age 1 5374.4 5420.4  
## - race.White:hours.per.week 1 5381.2 5423.2  
## - male:hours.per.week 1 5381.9 5423.9  
## - poly(education.num, 2) 2 5421.0 5461.0  
## - ms.married:male 1 5419.8 5461.8  
## - poly(hours.per.week, 2) 2 5440.9 5480.9  
## - poly(age, 2) 2 5537.8 5577.8  
## - poly(capnet, 1) 1 5720.4 5762.4  
##   
## Step: AIC=5417.04  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.neverm + ms.widowed +   
## race.White + race.Black + male + race.White:male + ms.married:male +   
## male:education.num + male:hours.per.week + race.White:age +   
## race.White:education.num + race.White:hours.per.week  
##   
## Df Deviance AIC  
## - race.Black 1 5375.8 5415.8  
## - race.White:male 1 5376.1 5416.1  
## - ms.neverm 1 5376.3 5416.3  
## - ms.widowed 1 5376.4 5416.4  
## <none> 5375.0 5417.0  
## - male:education.num 1 5377.7 5417.7  
## - race.White:education.num 1 5378.1 5418.1  
## - race.White:age 1 5378.4 5418.4  
## + ms.sep 1 5374.6 5418.6  
## + male:age 1 5374.9 5418.9  
## - race.White:hours.per.week 1 5381.7 5421.7  
## - male:hours.per.week 1 5382.4 5422.4  
## - poly(education.num, 2) 2 5421.4 5459.4  
## - ms.married:male 1 5420.3 5460.3  
## - poly(hours.per.week, 2) 2 5441.4 5479.4  
## - poly(age, 2) 2 5537.9 5575.9  
## - poly(capnet, 1) 1 5720.5 5760.5  
##   
## Step: AIC=5415.76  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.neverm + ms.widowed +   
## race.White + male + race.White:male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week  
##   
## Df Deviance AIC  
## - race.White:male 1 5376.7 5414.7  
## - ms.neverm 1 5377.0 5415.0  
## - ms.widowed 1 5377.1 5415.1  
## <none> 5375.8 5415.8  
## - male:education.num 1 5378.5 5416.5  
## + race.Black 1 5375.0 5417.0  
## + ms.sep 1 5375.2 5417.2  
## - race.White:age 1 5379.3 5417.3  
## - race.White:education.num 1 5379.4 5417.4  
## + male:age 1 5375.6 5417.6  
## - race.White:hours.per.week 1 5382.0 5420.0  
## - male:hours.per.week 1 5383.2 5421.2  
## - poly(education.num, 2) 2 5421.4 5457.4  
## - ms.married:male 1 5421.1 5459.1  
## - poly(hours.per.week, 2) 2 5441.5 5477.5  
## - poly(age, 2) 2 5540.1 5576.1  
## - poly(capnet, 1) 1 5721.5 5759.5  
##   
## Step: AIC=5414.69  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.neverm + ms.widowed +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week  
##   
## Df Deviance AIC  
## - ms.neverm 1 5377.9 5413.9  
## - ms.widowed 1 5378.0 5414.0  
## <none> 5376.7 5414.7  
## - male:education.num 1 5379.7 5415.7  
## + race.White:male 1 5375.8 5415.8  
## + race.Black 1 5376.1 5416.1  
## + ms.sep 1 5376.2 5416.2  
## - race.White:age 1 5380.4 5416.4  
## - race.White:education.num 1 5380.5 5416.5  
## + male:age 1 5376.6 5416.6  
## - race.White:hours.per.week 1 5384.3 5420.3  
## - male:hours.per.week 1 5384.4 5420.4  
## - poly(education.num, 2) 2 5423.4 5457.4  
## - ms.married:male 1 5421.5 5457.5  
## - poly(hours.per.week, 2) 2 5443.9 5477.9  
## - poly(age, 2) 2 5542.0 5576.0  
## - poly(capnet, 1) 1 5722.7 5758.7  
##   
## Step: AIC=5413.94  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.widowed + race.White +   
## male + ms.married:male + male:education.num + male:hours.per.week +   
## race.White:age + race.White:education.num + race.White:hours.per.week  
##   
## Df Deviance AIC  
## - ms.widowed 1 5379.9 5413.9  
## <none> 5377.9 5413.9  
## - male:education.num 1 5380.7 5414.7  
## + ms.neverm 1 5376.7 5414.7  
## + ms.sep 1 5376.9 5414.9  
## + race.White:male 1 5377.0 5415.0  
## + race.Black 1 5377.4 5415.4  
## - race.White:education.num 1 5381.7 5415.7  
## - race.White:age 1 5381.7 5415.7  
## + male:age 1 5377.8 5415.8  
## - race.White:hours.per.week 1 5385.6 5419.6  
## - male:hours.per.week 1 5385.7 5419.7  
## - ms.married:male 1 5421.6 5455.6  
## - poly(education.num, 2) 2 5424.0 5456.0  
## - poly(hours.per.week, 2) 2 5445.2 5477.2  
## - poly(age, 2) 2 5556.8 5588.8  
## - poly(capnet, 1) 1 5724.8 5758.8  
##   
## Step: AIC=5413.86  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +   
## male:education.num + male:hours.per.week + race.White:age +   
## race.White:education.num + race.White:hours.per.week  
##   
## Df Deviance AIC  
## <none> 5379.9 5413.9  
## + ms.widowed 1 5377.9 5413.9  
## + ms.neverm 1 5378.0 5414.0  
## - male:education.num 1 5382.4 5414.4  
## + race.White:male 1 5378.9 5414.9  
## + ms.sep 1 5379.1 5415.1  
## + race.Black 1 5379.3 5415.3  
## + male:age 1 5379.5 5415.5  
## - race.White:education.num 1 5383.6 5415.6  
## - race.White:age 1 5383.7 5415.7  
## - race.White:hours.per.week 1 5387.6 5419.6  
## - male:hours.per.week 1 5388.1 5420.1  
## - ms.married:male 1 5422.0 5454.0  
## - poly(education.num, 2) 2 5425.2 5455.2  
## - poly(hours.per.week, 2) 2 5447.1 5477.1  
## - poly(age, 2) 2 5557.4 5587.4  
## - poly(capnet, 1) 1 5726.1 5758.1

birthwt.step$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + ms.neverm + ms.sep +   
## ms.widowed + race.White + race.Black + male + race.White:male +   
## male:ms.married + age:male + education.num:male + hours.per.week:male +   
## age:race.White + education.num:race.White + hours.per.week:race.White  
##   
## Final Model:  
## income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
## 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +   
## male:education.num + male:hours.per.week + race.White:age +   
## race.White:education.num + race.White:hours.per.week  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 7977 5374.427 5420.427  
## 2 - male:age 1 0.1296165 7978 5374.556 5418.556  
## 3 - ms.sep 1 0.4855490 7979 5375.042 5417.042  
## 4 - race.Black 1 0.7174424 7980 5375.759 5415.759  
## 5 - race.White:male 1 0.9286577 7981 5376.688 5414.688  
## 6 - ms.neverm 1 1.2542059 7982 5377.942 5413.942  
## 7 - ms.widowed 1 1.9222651 7983 5379.864 5413.864

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +   
 male:education.num + male:hours.per.week + race.White:age +   
 race.White:education.num + race.White:hours.per.week, data = newdata,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week, family = binomial(), data = newdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0892 -0.5348 -0.2099 -0.0340 3.4151   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.016810 0.159582 -25.171 < 2e-16 \*\*\*  
## poly(age, 2)1 78.015209 13.061895 5.973 2.33e-09 \*\*\*  
## poly(age, 2)2 -55.981661 4.642137 -12.059 < 2e-16 \*\*\*  
## poly(education.num, 2)1 82.730224 12.275907 6.739 1.59e-11 \*\*\*  
## poly(education.num, 2)2 11.085326 3.782013 2.931 0.003378 \*\*   
## poly(hours.per.week, 2)1 49.820360 13.285459 3.750 0.000177 \*\*\*  
## poly(hours.per.week, 2)2 -29.785350 3.833790 -7.769 7.90e-15 \*\*\*  
## poly(capnet, 1) 168.714756 10.233913 16.486 < 2e-16 \*\*\*  
## ms.married 3.064271 0.155274 19.735 < 2e-16 \*\*\*  
## race.White 0.858815 0.791382 1.085 0.277829   
## male 0.698900 0.549197 1.273 0.203165   
## ms.married:male -1.176198 0.181582 -6.478 9.33e-11 \*\*\*  
## male:education.num -0.074035 0.036895 -2.007 0.044786 \*   
## male:hours.per.week 0.020686 0.008002 2.585 0.009730 \*\*   
## race.White:age -0.010255 0.010017 -1.024 0.305955   
## race.White:education.num 0.085280 0.040752 2.093 0.036378 \*   
## race.White:hours.per.week -0.019969 0.009312 -2.144 0.031995 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10972.9 on 9999 degrees of freedom  
## Residual deviance: 6698.4 on 9983 degrees of freedom  
## AIC: 6732.4  
##   
## Number of Fisher Scoring iterations: 6

estimated\_income\_g50K<- predict(best.logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1090029

presdicted.testing <- predict(best.logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1058026

Result model: Final Model:income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married + race.White + male + male:hours.per.week + race.White:age + race.White:hours.per.week

This is the best model, as it has the smallest testing missclasification rate of 10.58026% which is smaller than any previous testing missclassfication rates, with a small training missclasification rate of 10.90029%.

### Q11.

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +male:education.num + male:hours.per.week + race.White:age + race.White:education.num + race.White:hours.per.week+ education.num + race.Black:education.num, data = newdat\_training,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week + education.num + race.Black:education.num,   
## family = binomial(), data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0288 -0.5315 -0.2060 -0.0302 3.4020   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.274343 0.226710 -18.854 < 2e-16 \*\*\*  
## poly(age, 2)1 84.738077 13.421109 6.314 2.72e-10 \*\*\*  
## poly(age, 2)2 -51.812874 4.669948 -11.095 < 2e-16 \*\*\*  
## poly(education.num, 2)1 68.306736 12.360922 5.526 3.28e-08 \*\*\*  
## poly(education.num, 2)2 10.088377 3.747844 2.692 0.007107 \*\*   
## poly(hours.per.week, 2)1 51.446708 13.655892 3.767 0.000165 \*\*\*  
## poly(hours.per.week, 2)2 -26.753400 3.929714 -6.808 9.90e-12 \*\*\*  
## poly(capnet, 1) 144.466703 9.916437 14.568 < 2e-16 \*\*\*  
## ms.married 3.112944 0.173320 17.961 < 2e-16 \*\*\*  
## race.White 1.897359 0.931214 2.038 0.041599 \*   
## male 0.413300 0.614507 0.673 0.501220   
## education.num NA NA NA NA   
## ms.married:male -1.283395 0.202536 -6.337 2.35e-10 \*\*\*  
## male:education.num -0.063553 0.041163 -1.544 0.122605   
## male:hours.per.week 0.026193 0.009226 2.839 0.004525 \*\*   
## race.White:age -0.021913 0.011462 -1.912 0.055897 .   
## race.White:education.num 0.094003 0.045837 2.051 0.040288 \*   
## race.White:hours.per.week -0.030673 0.010911 -2.811 0.004935 \*\*   
## education.num:race.Black 0.020466 0.020579 0.995 0.319967   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5378.9 on 7982 degrees of freedom  
## AIC: 5414.9  
##   
## Number of Fisher Scoring iterations: 6

estimated\_income\_g50K<- predict(best.logistic.model, newdat\_training, type = "response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1088928

presdicted.testing <- predict(best.logistic.model, newdat\_testing, type = "response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1061371

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +male:education.num + male:hours.per.week + race.White:age + race.White:education.num + race.White:hours.per.week+ race.Black:education.num, data = newdat\_training,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week + race.Black:education.num, family = binomial(),   
## data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0288 -0.5315 -0.2060 -0.0302 3.4020   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.274343 0.226710 -18.854 < 2e-16 \*\*\*  
## poly(age, 2)1 84.738077 13.421109 6.314 2.72e-10 \*\*\*  
## poly(age, 2)2 -51.812874 4.669948 -11.095 < 2e-16 \*\*\*  
## poly(education.num, 2)1 68.306736 12.360922 5.526 3.28e-08 \*\*\*  
## poly(education.num, 2)2 10.088377 3.747844 2.692 0.007107 \*\*   
## poly(hours.per.week, 2)1 51.446708 13.655892 3.767 0.000165 \*\*\*  
## poly(hours.per.week, 2)2 -26.753400 3.929714 -6.808 9.90e-12 \*\*\*  
## poly(capnet, 1) 144.466703 9.916437 14.568 < 2e-16 \*\*\*  
## ms.married 3.112944 0.173320 17.961 < 2e-16 \*\*\*  
## race.White 1.897359 0.931214 2.038 0.041599 \*   
## male 0.413300 0.614507 0.673 0.501220   
## ms.married:male -1.283395 0.202536 -6.337 2.35e-10 \*\*\*  
## male:education.num -0.063553 0.041163 -1.544 0.122605   
## male:hours.per.week 0.026193 0.009226 2.839 0.004525 \*\*   
## race.White:age -0.021913 0.011462 -1.912 0.055897 .   
## race.White:education.num 0.094003 0.045837 2.051 0.040288 \*   
## race.White:hours.per.week -0.030673 0.010911 -2.811 0.004935 \*\*   
## education.num:race.Black 0.020466 0.020579 0.995 0.319967   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5378.9 on 7982 degrees of freedom  
## AIC: 5414.9  
##   
## Number of Fisher Scoring iterations: 6

estimated\_income\_g50K<- predict(best.logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1088928

presdicted.testing <- predict(best.logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1061371

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +male:education.num + male:hours.per.week + race.White:age + race.White:education.num + race.White:hours.per.week+ race.Black:age + race.Black:education.num, data = newdat\_training,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week + race.Black:age + race.Black:education.num,   
## family = binomial(), data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0289 -0.5315 -0.2058 -0.0302 3.4020   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.270023 0.227701 -18.753 < 2e-16 \*\*\*  
## poly(age, 2)1 86.763852 16.911384 5.130 2.89e-07 \*\*\*  
## poly(age, 2)2 -51.816211 4.671006 -11.093 < 2e-16 \*\*\*  
## poly(education.num, 2)1 66.975054 14.066230 4.761 1.92e-06 \*\*\*  
## poly(education.num, 2)2 10.097761 3.746366 2.695 0.007032 \*\*   
## poly(hours.per.week, 2)1 51.365815 13.664634 3.759 0.000171 \*\*\*  
## poly(hours.per.week, 2)2 -26.757619 3.930463 -6.808 9.91e-12 \*\*\*  
## poly(capnet, 1) 144.473119 9.916928 14.568 < 2e-16 \*\*\*  
## ms.married 3.112734 0.173334 17.958 < 2e-16 \*\*\*  
## race.White 1.896688 0.931228 2.037 0.041674 \*   
## male 0.411654 0.614556 0.670 0.502959   
## ms.married:male -1.283060 0.202559 -6.334 2.38e-10 \*\*\*  
## male:education.num -0.063406 0.041159 -1.541 0.123434   
## male:hours.per.week 0.026193 0.009226 2.839 0.004524 \*\*   
## race.White:age -0.023583 0.014259 -1.654 0.098145 .   
## race.White:education.num 0.099729 0.054203 1.840 0.065781 .   
## race.White:hours.per.week -0.030600 0.010923 -2.801 0.005087 \*\*   
## age:race.Black -0.002806 0.014285 -0.196 0.844254   
## education.num:race.Black 0.031041 0.057660 0.538 0.590336   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5378.8 on 7981 degrees of freedom  
## AIC: 5416.8  
##   
## Number of Fisher Scoring iterations: 6

estimated\_income\_g50K<- predict(best.logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1088922

presdicted.testing <- predict(best.logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1061313

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +male:education.num + male:hours.per.week + race.White:age, data = newdat\_training,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age, family = binomial(),   
## data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0290 -0.5359 -0.2076 -0.0327 3.3730   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.10305 0.17734 -23.137 < 2e-16 \*\*\*  
## poly(age, 2)1 85.33065 13.01491 6.556 5.51e-11 \*\*\*  
## poly(age, 2)2 -52.04816 4.66858 -11.149 < 2e-16 \*\*\*  
## poly(education.num, 2)1 86.62972 8.80160 9.843 < 2e-16 \*\*\*  
## poly(education.num, 2)2 9.84923 3.72839 2.642 0.00825 \*\*   
## poly(hours.per.week, 2)1 21.84673 8.53684 2.559 0.01049 \*   
## poly(hours.per.week, 2)2 -25.37556 3.80718 -6.665 2.64e-11 \*\*\*  
## poly(capnet, 1) 144.36656 9.91448 14.561 < 2e-16 \*\*\*  
## ms.married 3.09793 0.17274 17.934 < 2e-16 \*\*\*  
## race.White 1.46146 0.49429 2.957 0.00311 \*\*   
## male 0.43495 0.61044 0.713 0.47615   
## ms.married:male -1.27913 0.20203 -6.331 2.43e-10 \*\*\*  
## male:education.num -0.05920 0.04103 -1.443 0.14901   
## male:hours.per.week 0.02446 0.00913 2.679 0.00738 \*\*   
## race.White:age -0.02237 0.01111 -2.013 0.04414 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5391.0 on 7985 degrees of freedom  
## AIC: 5421  
##   
## Number of Fisher Scoring iterations: 6

estimated\_income\_g50K<- predict(best.logistic.model, newdat\_training, type = "response")  
misscls.train<-sum((estimated\_income\_g50K-newdat\_training$income\_g50K)^2)/  
length(newdat\_training$income\_g50K)  
misscls.train

## [1] 0.1092282

presdicted.testing <- predict(best.logistic.model, newdat\_testing, type = "response")  
misscls.test<-sum((presdicted.testing-newdat\_testing$income\_g50K)^2)/  
length(newdat\_testing$income\_g50K)  
misscls.test

## [1] 0.1061953

After trying four other logistic regression model, although they all have a small missclasification rate around 10.61%, the best model is the following one with smallest missclasification: income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male + male:education.num + male:hours.per.week + race.White:age + race.White:education.num + race.White:hours.per.week

So this is my favorite model as it gives the lowest misclassfication rate.

### Q12.

#### estimates of the parameters with training and testing combined

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +   
 male:education.num + male:hours.per.week + race.White:age +   
 race.White:education.num + race.White:hours.per.week, data = newdata,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week, family = binomial(), data = newdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0892 -0.5348 -0.2099 -0.0340 3.4151   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.016810 0.159582 -25.171 < 2e-16 \*\*\*  
## poly(age, 2)1 78.015209 13.061895 5.973 2.33e-09 \*\*\*  
## poly(age, 2)2 -55.981661 4.642137 -12.059 < 2e-16 \*\*\*  
## poly(education.num, 2)1 82.730224 12.275907 6.739 1.59e-11 \*\*\*  
## poly(education.num, 2)2 11.085326 3.782013 2.931 0.003378 \*\*   
## poly(hours.per.week, 2)1 49.820360 13.285459 3.750 0.000177 \*\*\*  
## poly(hours.per.week, 2)2 -29.785350 3.833790 -7.769 7.90e-15 \*\*\*  
## poly(capnet, 1) 168.714756 10.233913 16.486 < 2e-16 \*\*\*  
## ms.married 3.064271 0.155274 19.735 < 2e-16 \*\*\*  
## race.White 0.858815 0.791382 1.085 0.277829   
## male 0.698900 0.549197 1.273 0.203165   
## ms.married:male -1.176198 0.181582 -6.478 9.33e-11 \*\*\*  
## male:education.num -0.074035 0.036895 -2.007 0.044786 \*   
## male:hours.per.week 0.020686 0.008002 2.585 0.009730 \*\*   
## race.White:age -0.010255 0.010017 -1.024 0.305955   
## race.White:education.num 0.085280 0.040752 2.093 0.036378 \*   
## race.White:hours.per.week -0.019969 0.009312 -2.144 0.031995 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10972.9 on 9999 degrees of freedom  
## Residual deviance: 6698.4 on 9983 degrees of freedom  
## AIC: 6732.4  
##   
## Number of Fisher Scoring iterations: 6

#### estimates of the parameters with training data

best.logistic.model <- glm(income\_g50K ~ poly(age, 2) + poly(education.num, 2) + poly(hours.per.week,   
 2) + poly(capnet, 1) + ms.married + race.White + male + ms.married:male +   
 male:education.num + male:hours.per.week + race.White:age +   
 race.White:education.num + race.White:hours.per.week, data = newdat\_training,   
family = binomial())  
summary(best.logistic.model )

##   
## Call:  
## glm(formula = income\_g50K ~ poly(age, 2) + poly(education.num,   
## 2) + poly(hours.per.week, 2) + poly(capnet, 1) + ms.married +   
## race.White + male + ms.married:male + male:education.num +   
## male:hours.per.week + race.White:age + race.White:education.num +   
## race.White:hours.per.week, family = binomial(), data = newdat\_training)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0299 -0.5312 -0.2061 -0.0302 3.4035   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.14400 0.18307 -22.637 < 2e-16 \*\*\*  
## poly(age, 2)1 85.20992 13.41474 6.352 2.13e-10 \*\*\*  
## poly(age, 2)2 -51.88157 4.67113 -11.107 < 2e-16 \*\*\*  
## poly(education.num, 2)1 69.49675 12.28917 5.655 1.56e-08 \*\*\*  
## poly(education.num, 2)2 9.85994 3.74267 2.634 0.008427 \*\*   
## poly(hours.per.week, 2)1 49.81362 13.48868 3.693 0.000222 \*\*\*  
## poly(hours.per.week, 2)2 -26.65400 3.91992 -6.800 1.05e-11 \*\*\*  
## poly(capnet, 1) 144.53441 9.91534 14.577 < 2e-16 \*\*\*  
## ms.married 3.11125 0.17339 17.944 < 2e-16 \*\*\*  
## race.White 1.76289 0.91684 1.923 0.054507 .   
## male 0.42742 0.61416 0.696 0.486471   
## ms.married:male -1.28665 0.20256 -6.352 2.13e-10 \*\*\*  
## male:education.num -0.06486 0.04113 -1.577 0.114857   
## male:hours.per.week 0.02618 0.00922 2.840 0.004516 \*\*   
## race.White:age -0.02226 0.01146 -1.942 0.052124 .   
## race.White:education.num 0.09005 0.04563 1.974 0.048437 \*   
## race.White:hours.per.week -0.02920 0.01073 -2.720 0.006520 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8798.8 on 7999 degrees of freedom  
## Residual deviance: 5379.9 on 7983 degrees of freedom  
## AIC: 5413.9  
##   
## Number of Fisher Scoring iterations: 6

The estimates of the parameters for entire data are very similar to those with training data. It is expected that there is no big difference in the estimates of the parameters for two datas.

### Q13.

library(boot)  
cv.err <- cv.glm(newdata, best.logistic.model,K=10)#K-fold

cv.err$delta

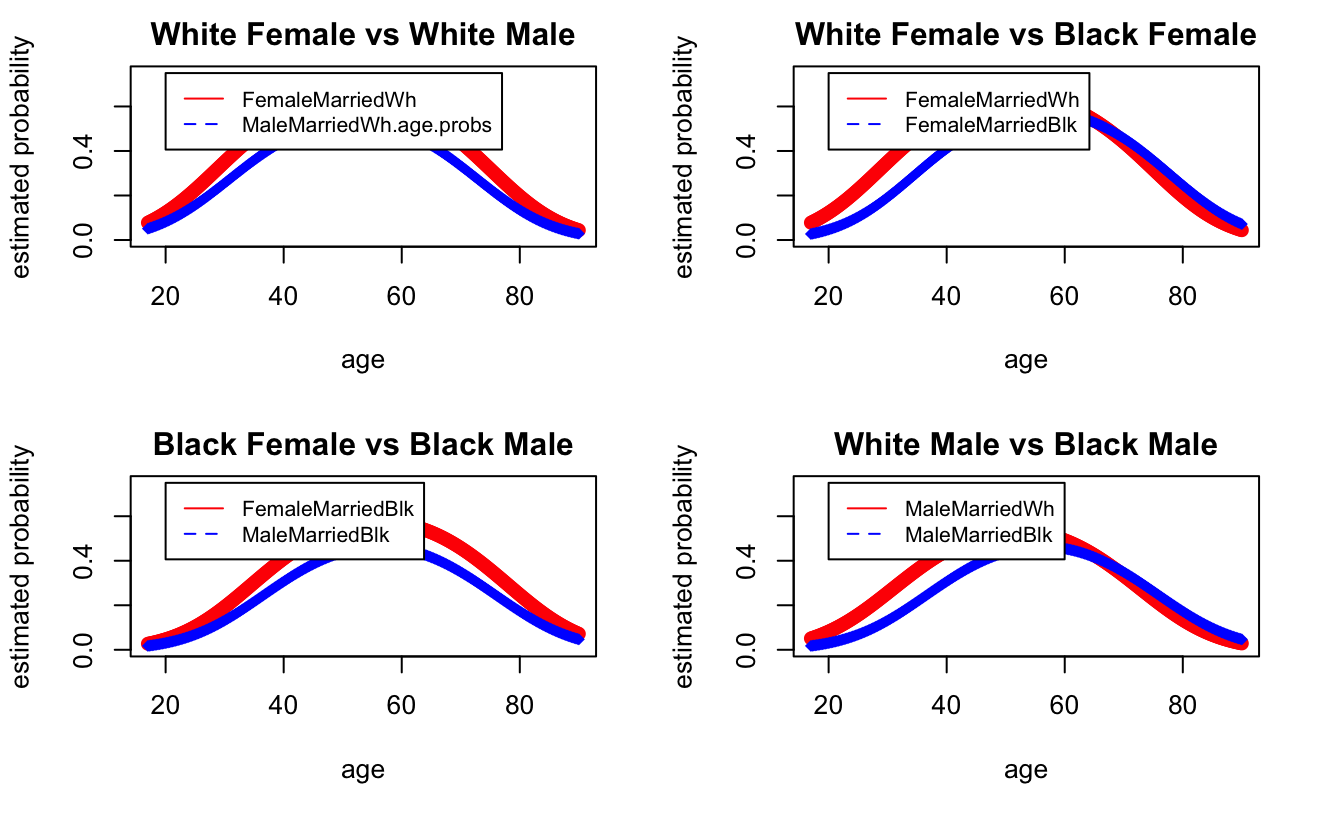
## [1] 0.1089851 0.1089640

The raw cross-validation estimate of prediction error is 10.90853%, and the adjusted cross-validation estimate of prediction error is 10.90514%. The result from cv function is a little higher than the result from I get for training and testing data. The difference occurs as the adjustment is used to compensate for the bias introduced without using leave-one-out cross-validation, thus increase the error a litte.

### Q14.

Age

m<-300 # number of data points in prediction  
attach(newdata)  
age\_predict<-min(age)+ (max(age)-min(age))\*seq(0,1,1/(m-1))  
FemaleMarriedWh.age <-data.frame(age=age\_predict,education.num=rep(mean(education.num),length(age\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(age\_predict)) ,   
ms.married= rep(1,length(age\_predict)) , ms.neverm=rep(0,length(age\_predict)),   
 ms.sep=rep(0,length(age\_predict)), ms.widowed = rep(0,length(age\_predict)) ,  
 race.White= rep(1,length(age\_predict)) ,  
 race.Black = rep(0,length(age\_predict)) , capnet=rep(mean(capnet),  
length(age\_predict)) , male =rep(0,length(age\_predict)),  
income\_g50K= rep(0,length(age\_predict)))   
  
FemaleMarriedWh.age.probs<-predict(best.logistic.model, FemaleMarriedWh.age,type = "response")  
  
FemaleMarriedBlk.age <-data.frame(age=age\_predict,education.num=rep(mean(education.num),length(age\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(age\_predict)) ,   
ms.married= rep(1,length(age\_predict)) , ms.neverm=rep(0,length(age\_predict)),   
 ms.sep=rep(0,length(age\_predict)), ms.widowed = rep(0,length(age\_predict)) ,  
 race.White= rep(0,length(age\_predict)) ,  
 race.Black = rep(1,length(age\_predict)) , capnet=rep(mean(capnet),  
length(age\_predict)) , male =rep(0,length(age\_predict)),  
income\_g50K= rep(0,length(age\_predict)))   
  
FemaleMarriedBlk.age.probs<-predict(best.logistic.model, FemaleMarriedBlk.age,type = "response")  
  
MaleMarriedWh.age <-data.frame(age=age\_predict,education.num=rep(mean(education.num),length(age\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(age\_predict)) ,   
ms.married= rep(1,length(age\_predict)) , ms.neverm=rep(0,length(age\_predict)),   
 ms.sep=rep(0,length(age\_predict)), ms.widowed = rep(0,length(age\_predict)) ,  
 race.White= rep(1,length(age\_predict)) ,  
 race.Black = rep(0,length(age\_predict)) , capnet=rep(mean(capnet),  
length(age\_predict)) , male =rep(1,length(age\_predict)),  
income\_g50K= rep(0,length(age\_predict)))   
  
MaleMarriedWh.age.probs<-predict(best.logistic.model, MaleMarriedWh.age,type = "response")  
  
MaleMarriedBlk.age <-data.frame(age=age\_predict,education.num=rep(mean(education.num),length(age\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(age\_predict)) ,   
ms.married= rep(1,length(age\_predict)) , ms.neverm=rep(0,length(age\_predict)),   
 ms.sep=rep(0,length(age\_predict)), ms.widowed = rep(0,length(age\_predict)) ,  
 race.White= rep(0,length(age\_predict)) ,  
 race.Black = rep(1,length(age\_predict)) , capnet=rep(mean(capnet),  
length(age\_predict)) , male =rep(1,length(age\_predict)),  
income\_g50K= rep(0,length(age\_predict)))   
  
MaleMarriedBlk.age.probs<-predict(best.logistic.model, MaleMarriedBlk.age,type = "response")  
  
par(mfrow=(c(2,2)))  
  
plot(age\_predict,FemaleMarriedWh.age.probs,type="b", pch=19, col="red",  
xlab="age",  
ylab="estimated probability",ylim=c(0,0.75),main="White Female vs White Male")  
lines(age\_predict,  
MaleMarriedWh.age.probs,pch=18, col="blue", type="b", lty=2,  
xlab="age",ylab="estimated probability" )  
legend(20, 0.75, legend=c("FemaleMarriedWh", "MaleMarriedWh.age.probs"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)  
  
plot(age\_predict,FemaleMarriedWh.age.probs,xlab="age",  
type="b", pch=19, col="red",  
ylab="estimated probability",ylim=c(0,0.75),main="White Female vs Black Female")  
lines(age\_predict,FemaleMarriedBlk.age.probs,pch=18, col="blue", type="b", lty=2,  
xlab="age",ylab="estimated probability" )  
legend(20, 0.75, legend=c("FemaleMarriedWh", "FemaleMarriedBlk"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)  
  
plot(age\_predict,FemaleMarriedBlk.age.probs,xlab="age",  
type="b", pch=19, col="red",  
ylab="estimated probability",ylim=c(0,0.75),main="Black Female vs Black Male")  
lines(age\_predict,MaleMarriedBlk.age.probs,pch=18, col="blue", type="b", lty=2,  
xlab="age",ylab="estimated probability" )  
legend(20, 0.75, legend=c("FemaleMarriedBlk", "MaleMarriedBlk"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)  
  
plot(age\_predict,MaleMarriedWh.age.probs,xlab="age",  
type="b", pch=19, col="red",  
ylab="estimated probability",ylim=c(0,0.75),main="White Male vs Black Male")  
lines(age\_predict,MaleMarriedBlk.age.probs,pch=18, col="blue", type="b", lty=2,  
xlab="age",ylab="estimated probability" )  
legend(20, 0.75, legend=c("MaleMarriedWh",  
 "MaleMarriedBlk"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)



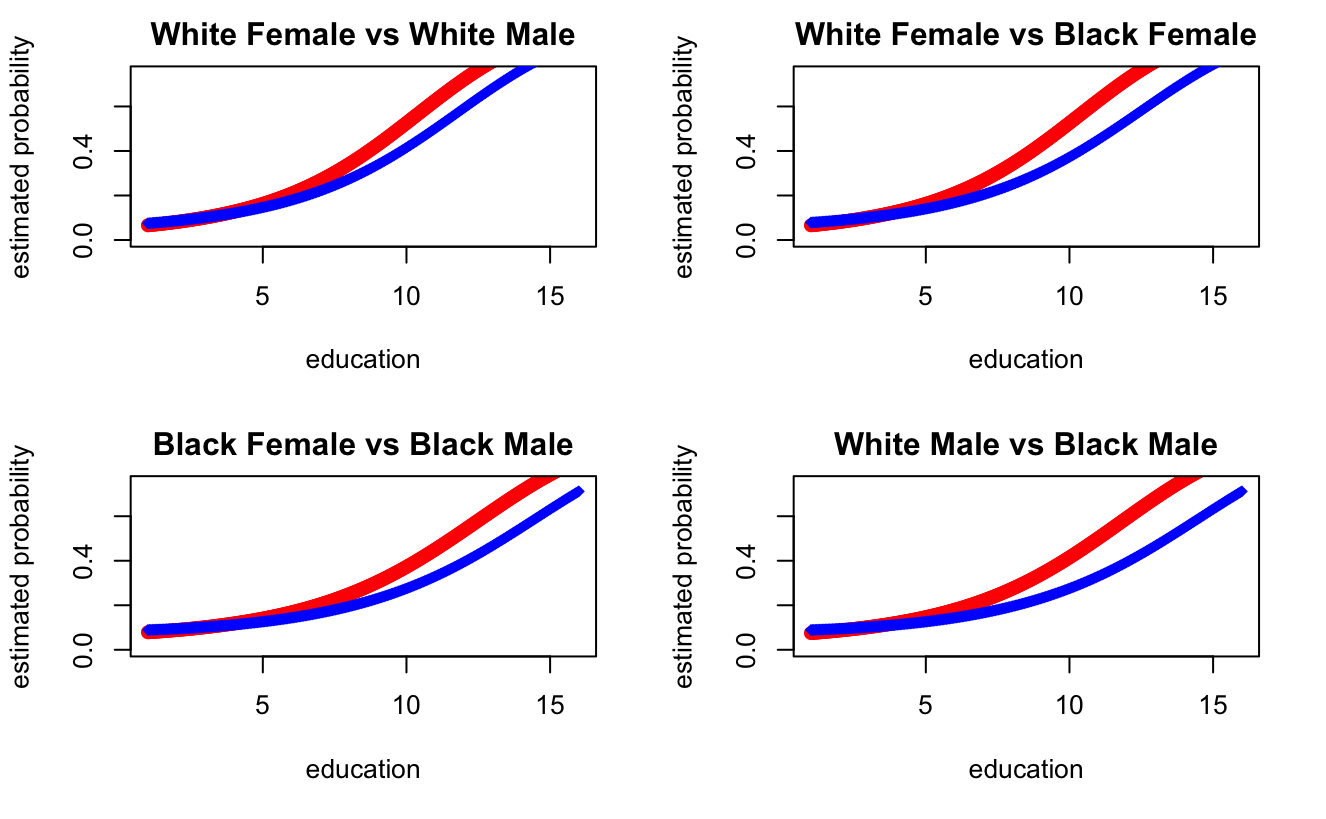
From the plots, all cases show a somewhat normal distribution, with younger age and elder age have a lower probability and middle age has a higher probability. Married white female has a overall higher probability than married white male, same to black male and female. White female has a higher probability than black female, while the two lines interact in elder age, same as the white male and balck male group. So gender modified the relationship between probability of income>=50K and age more than race. It has a similar pattern to neural network one, as both models show an increase of probability with age increase then followed by a decrease, and gender modified more in the neural network model.

Education

m<-300 # number of data points in prediction  
attach(newdata)

## The following objects are masked from newdata (pos = 3):  
##   
## age, capnet, education.num, hours.per.week, income\_g50K, male,  
## ms.married, ms.neverm, ms.sep, ms.widowed, race.Black,  
## race.White

edu\_predict<-min(education.num)+ (max(education.num)-min(education.num))\*seq(0,1,1/(m-1))  
FemaleMarriedWh.edu <-data.frame(education.num=edu\_predict,age=rep(mean(age),length(edu\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(edu\_predict)) ,   
ms.married= rep(1,length(edu\_predict)) , ms.neverm=rep(0,length(edu\_predict)),   
 ms.sep=rep(0,length(edu\_predict)), ms.widowed = rep(0,length(edu\_predict)) ,  
 race.White= rep(1,length(edu\_predict)) ,  
 race.Black = rep(0,length(edu\_predict)) , capnet=rep(mean(capnet),  
length(edu\_predict)) , male =rep(0,length(edu\_predict)),  
income\_g50K= rep(0,length(edu\_predict)))   
  
FemaleMarriedWh.edu.probs<-predict(best.logistic.model, FemaleMarriedWh.edu,type = "response")  
  
FemaleMarriedBlk.edu <-data.frame(education.num=edu\_predict,age=rep(mean(age),length(edu\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(edu\_predict)) ,   
ms.married= rep(1,length(edu\_predict)) , ms.neverm=rep(0,length(edu\_predict)),   
 ms.sep=rep(0,length(edu\_predict)), ms.widowed = rep(0,length(edu\_predict)) ,  
 race.White= rep(0,length(edu\_predict)) ,  
 race.Black = rep(1,length(edu\_predict)) , capnet=rep(mean(capnet),  
length(edu\_predict)) , male =rep(0,length(edu\_predict)),  
income\_g50K= rep(0,length(edu\_predict)))   
  
FemaleMarriedBlk.edu.probs<-predict(best.logistic.model, FemaleMarriedBlk.edu,type = "response")  
  
MaleMarriedWh.edu <-data.frame(education.num=edu\_predict,age=rep(mean(age),length(edu\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(edu\_predict)) ,   
ms.married= rep(1,length(edu\_predict)) , ms.neverm=rep(0,length(edu\_predict)),   
 ms.sep=rep(0,length(edu\_predict)), ms.widowed = rep(0,length(edu\_predict)) ,  
 race.White= rep(1,length(edu\_predict)) ,  
 race.Black = rep(0,length(edu\_predict)) , capnet=rep(mean(capnet),  
length(edu\_predict)) , male =rep(1,length(edu\_predict)),  
income\_g50K= rep(0,length(edu\_predict)))   
  
MaleMarriedWh.edu.probs<-predict(best.logistic.model, MaleMarriedWh.edu,type = "response")  
  
MaleMarriedBlk.edu <-data.frame(education.num=edu\_predict,age=rep(mean(age),length(edu\_predict)),  
hours.per.week=rep(mean(hours.per.week),length(edu\_predict)) ,   
ms.married= rep(1,length(edu\_predict)) , ms.neverm=rep(0,length(edu\_predict)),   
 ms.sep=rep(0,length(edu\_predict)), ms.widowed = rep(0,length(edu\_predict)) ,  
 race.White= rep(0,length(edu\_predict)) ,  
 race.Black = rep(1,length(edu\_predict)) , capnet=rep(mean(capnet),  
length(edu\_predict)) , male =rep(1,length(edu\_predict)),  
income\_g50K= rep(0,length(edu\_predict)))   
  
MaleMarriedBlk.edu.probs<-predict(best.logistic.model, MaleMarriedBlk.edu,type = "response")  
  
par(mfrow=(c(2,2)))  
  
plot(edu\_predict,FemaleMarriedWh.edu.probs,type="b", pch=19, col="red",  
xlab="education",  
ylab="estimated probability",ylim=c(0,0.75),main="White Female vs White Male")  
lines(edu\_predict,  
MaleMarriedWh.edu.probs,pch=18, col="blue", type="b", lty=2,  
xlab="education",ylab="estimated probability" )  
legend(20, 0.75, legend=c("FemaleMarriedWh", "MaleMarriedWh.edu.probs"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)  
  
plot(edu\_predict,FemaleMarriedWh.edu.probs,xlab="education",  
type="b", pch=19, col="red",  
ylab="estimated probability",ylim=c(0,0.75),main="White Female vs Black Female")  
lines(edu\_predict,FemaleMarriedBlk.edu.probs,pch=18, col="blue", type="b", lty=2,  
xlab="education",ylab="estimated probability" )  
legend(20, 0.75, legend=c("FemaleMarriedWh", "FemaleMarriedBlk"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)  
  
plot(edu\_predict,FemaleMarriedBlk.edu.probs,xlab="education",  
type="b", pch=19, col="red",  
ylab="estimated probability",ylim=c(0,0.75),main="Black Female vs Black Male")  
lines(edu\_predict,MaleMarriedBlk.edu.probs,pch=18, col="blue", type="b", lty=2,  
xlab="education",ylab="estimated probability" )  
legend(20, 0.75, legend=c("FemaleMarriedBlk", "MaleMarriedBlk"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)  
  
plot(edu\_predict,MaleMarriedWh.edu.probs,xlab="education",  
type="b", pch=19, col="red",  
ylab="estimated probability",ylim=c(0,0.75),main="White Male vs Black Male")  
lines(edu\_predict,MaleMarriedBlk.edu.probs,pch=18, col="blue", type="b", lty=2,  
xlab="education",ylab="estimated probability" )  
legend(20, 0.75, legend=c("MaleMarriedWh",  
 "MaleMarriedBlk"),  
 col=c("red", "blue"), lty=1:2, cex=0.8)



From the plots, all cases show a increasing probablity as education time increases. Married white female has a slightly less probabiltiy before year 5, and then the probability is greater than married white male, same to black male and female. White female has a higher probability than black female, and it has a larger gap with education number increase, same as the white male and balck male group. Since there is a larger gap between white female vs black female and larger gap between white male vs black male than the gender comparison. The race modified the relationship between probability of income>=50K and education.num more than gender. It has a similar pattern to neural network one, and race also modifies more in the nueral network one.