Question 9: R Excercise

Consider the following data on home-well contamination in 3020 households in Ara- hazar upazila, Bangladesh. The response variable is switch (binary variable whether or not the household switched to another well from an unsafe well). Other variables collected for each household were arsenic (the level of arsenic contamination in the household's original well, in hundreds of micrograms per liter), dist100 (distance in 100-meter units to the closest known safe well), educ (years of education of the head of the household) and assoc (whether or not any members of the household participated in any community organizations: no or yes). The data is available in MyCourses under Datasets. Load the data and compute dist100 as follows.

```
wells <- read.table("wells.dat")
attach(wells)
dist100 <- dist/100
```

(a) Fit a logistic regression model with the intercept and dist100. Interpret the model. Test the adequacy of this model using the Pearson X^2 and the likelihood ratio G^2 statistics. Conclude at the 5% level.

```
y<-as.factor(wells$switch)
logitmod<-glm(y~dist100,family=binomial(link="logit"))
summary(logitmod)
```

```
##
## Call:
## glm(formula = y ~ dist100, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4406  -1.3058   0.9669   1.0308   1.6603
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.60596    0.06031  10.047  < 2e-16 ***
## dist100     -0.62188    0.09743  -6.383 1.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4076.2  on 3018  degrees of freedom
## AIC: 4080.2
##
## Number of Fisher Scoring iterations: 4
```

Since g(x)=$\frac{1}{1-x}$, so we can get that $log\frac{\pi}{1-\pi} = 0.61 - 0.62 * dist100$. The paramter of dist100 is -0.62 which implies that as dist100 increase, the estimiated response will also increase, it is much more possible to switch the well. For the Pearson Residual:

```
ncat <- 10

bins <- cut(dist100,quantile(dist100,prob=c(0:ncat)/ncat),include.lowest=T)
```

```
lswi <- split(switch,bins)
counts <- lapply(lswi,FUN=function(x){as.numeric(x > 0)})
beta <- coefficients(logitmod)

observed <- lapply(counts,FUN=function(x){c(sum(x),length(x)-sum(x))})
observed <- matrix(as.numeric(unlist(observed)),ncol=2,byrow=TRUE)

#fitted number of success and failure
fitted <- lapply(split(dist100,bins),FUN=function(x){pi <- exp(beta[1]+x*beta[2])/(1+exp(beta[1]+x*beta
fitted <- matrix(as.numeric(unlist(fitted)),ncol=2,byrow=TRUE)
cbind(observed,fitted)
```

```
##       [,1] [,2]     [,3]     [,4]
##  [1,]  170  132 192.1434 109.8566
##  [2,]  192  110 188.8229 113.1771
##  [3,]  183  119 186.1654 115.8346
##  [4,]  198  104 183.5034 118.4966
##  [5,]  175  127 180.6797 121.3203
##  [6,]  188  114 177.0390 124.9610
##  [7,]  171  131 172.4951 129.5049
##  [8,]  183  119 166.3623 135.6377
##  [9,]  150  152 156.6636 145.3364
## [10,]  127  175 133.1252 168.8748
```

```
X.2  <- sum(((observed-fitted)^2)/fitted)

observed[10,2] <- 0.5  # just to avoid a log of zero in the likelihood ratio statistics

G.2 <- 2*sum(observed*log(observed/fitted))

pchisq(X.2, df=8,lower.tail=FALSE)
```

```
## [1] 0.02879032
```

For the G^2:

```
pchisq(G.2, df=8,lower.tail=FALSE)
```

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

with a confidence interval

```
#library(arm)
# se.coef extract the standard error
#se <- se.coef(logitmod)
#c(exp(beta[2]-1.96*se[2]),exp(beta[2]+1.96*se[2]))
```

According to the p value of dist100 and intercept, they are all less than 5%. We can conclude that the parameter has siginificant level of 5%, we reject the null hypothesis.

(b) Find the most appropriate logistic regression model for the data. Use the deviance, but also consider practical significance by looking at the AIC and the size of the effect of the predictors. Interpret the final model.

Firstly, try all the predictor:

```
lm1<-glm(y~(dist100+educ+assoc+arsenic)^2,family=binomial(link="logit"))
summary(lm1)
```

```
##
## Call:
## glm(formula = y ~ (dist100 + educ + assoc + arsenic)^2, family = binomial(link = "logit"))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.747  -1.195   0.725   1.069   1.929
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.05775    0.17648  -0.327  0.74351
## dist100         -1.21607    0.27927  -4.354 1.33e-05 ***
## educ            -0.01324    0.02142  -0.618  0.53650
## assoc            0.13889    0.18922   0.734  0.46293
## arsenic          0.53369    0.09446   5.650 1.61e-08 ***
## dist100:educ     0.08385    0.02682   3.126  0.00177 **
## dist100:assoc    0.21882    0.21480   1.019  0.30834
## dist100:arsenic -0.11005    0.10320  -1.066  0.28624
## educ:assoc      -0.02757    0.01981  -1.391  0.16408
## educ:arsenic     0.01744    0.01101   1.584  0.11315
## assoc:arsenic   -0.16256    0.08424  -1.930  0.05364 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3883.4  on 3009  degrees of freedom
## AIC: 3905.4
##
## Number of Fisher Scoring iterations: 4
```

```
step(lm1)
```

```
## Start:  AIC=3905.4
## y ~ (dist100 + educ + assoc + arsenic)^2
##
##                   Df Deviance    AIC
## - dist100:assoc    1   3884.4 3904.4
## - dist100:arsenic  1   3884.5 3904.5
## - educ:assoc       1   3885.3 3905.3
## <none>                 3883.4 3905.4
## - educ:arsenic     1   3885.9 3905.9
## - assoc:arsenic    1   3887.1 3907.1
## - dist100:educ     1   3893.4 3913.4
##
## Step:  AIC=3904.43
## y ~ dist100 + educ + assoc + arsenic + dist100:educ + dist100:arsenic +
##     educ:assoc + educ:arsenic + assoc:arsenic
##
##                   Df Deviance    AIC
## - dist100:arsenic  1   3885.6 3903.6
## - educ:assoc       1   3886.3 3904.3
## <none>                 3884.4 3904.4
## - educ:arsenic     1   3887.0 3905.0
## - assoc:arsenic    1   3887.4 3905.4
## - dist100:educ     1   3894.2 3912.2
##
## Step:  AIC=3903.59
## y ~ dist100 + educ + assoc + arsenic + dist100:educ + educ:assoc +
##     educ:arsenic + assoc:arsenic
##
##                 Df Deviance    AIC
## - educ:assoc     1   3887.4 3903.4
## <none>               3885.6 3903.6
## - educ:arsenic   1   3888.5 3904.5
## - assoc:arsenic  1   3888.6 3904.6
## - dist100:educ   1   3895.9 3911.9
##
## Step:  AIC=3903.41
## y ~ dist100 + educ + assoc + arsenic + dist100:educ + educ:arsenic +
##     assoc:arsenic
##
##                 Df Deviance    AIC
## <none>               3887.4 3903.4
## - assoc:arsenic  1   3890.1 3904.1
## - educ:arsenic   1   3890.4 3904.4
## - dist100:educ   1   3897.7 3911.7
```

```
##
## Call:  glm(formula = y ~ dist100 + educ + assoc + arsenic + dist100:educ +
##     educ:arsenic + assoc:arsenic, family = binomial(link = "logit"))
##
## Coefficients:
##   (Intercept)        dist100           educ          assoc        arsenic
##       0.06895       -1.32714       -0.02523        0.06878        0.45680
##  dist100:educ   educ:arsenic  assoc:arsenic
##       0.08444        0.01861       -0.13240
##
## Degrees of Freedom: 3019 Total (i.e. Null);  3012 Residual
## Null Deviance:       4118
## Residual Deviance: 3887  AIC: 3903
```

```
lm2<-glm(y~dist100+educ+assoc+arsenic+ dist100:educ +
    educ:arsenic + assoc:arsenic,family=binomial(link="logit"))
summary(lm2)
```

```
##
## Call:
## glm(formula = y ~ dist100 + educ + assoc + arsenic + dist100:educ +
##     educ:arsenic + assoc:arsenic, family = binomial(link = "logit"))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.621  -1.204   0.730   1.068   1.882
##
```

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.06895    0.13881   0.497  0.61937
## dist100        -1.32714    0.17510  -7.579 3.47e-14 ***
## educ           -0.02523    0.01995  -1.264  0.20610
## assoc           0.06878    0.14474   0.475  0.63465
## arsenic         0.45680    0.07248   6.302 2.93e-10 ***
## dist100:educ    0.08444    0.02668   3.165  0.00155 **
## educ:arsenic    0.01861    0.01090   1.708  0.08756 .
## assoc:arsenic -0.13240    0.08056  -1.644  0.10026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3887.4  on 3012  degrees of freedom
## AIC: 3903.4
##
## Number of Fisher Scoring iterations: 4
```

```
#anova(lm2,lm1,test='Chisq')
```

```
#sicne assoc has a high p value
lm3<- update(lm2,.~.-assoc)
anova(lm3,lm2,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ dist100 + educ + arsenic + dist100:educ + educ:arsenic +
##     assoc:arsenic
## Model 2: y ~ dist100 + educ + assoc + arsenic + dist100:educ + educ:arsenic +
##     assoc:arsenic
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3013     3887.6
## 2      3012     3887.4  1  0.22568   0.6347
```

It shows that we can drop assoc, since the p value is 0.63>0.5.

```
summary(lm3)
```

```
##
## Call:
## glm(formula = y ~ dist100 + educ + arsenic + dist100:educ + educ:arsenic +
##     assoc:arsenic, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5902  -1.2040   0.7325   1.0670   1.8785
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.10059    0.12177   0.826  0.40879
## dist100        -1.32814    0.17505  -7.587 3.27e-14 ***
## educ           -0.02544    0.01996  -1.275  0.20237
## arsenic         0.44203    0.06530   6.770 1.29e-11 ***
## dist100:educ   0.08465    0.02667   3.174  0.00151 **
## educ:arsenic   0.01865    0.01090   1.711  0.08706 .
## arsenic:assoc -0.10000    0.04297  -2.327  0.01994 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3887.6  on 3013  degrees of freedom
## AIC: 3901.6
##
## Number of Fisher Scoring iterations: 4
```

```r
# since eudc has a high p value
lm4<- update(lm3,.~.-educ)
anova(lm4,lm3,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ dist100 + arsenic + dist100:educ + educ:arsenic + arsenic:assoc
## Model 2: y ~ dist100 + educ + arsenic + dist100:educ + educ:arsenic +
##     assoc:arsenic
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3014     3889.3
## 2      3013     3887.6  1   1.6258   0.2023
```

```r
summary(lm4)
```

```
##
## Call:
## glm(formula = y ~ dist100 + arsenic + dist100:educ + educ:arsenic +
##     arsenic:assoc, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6566  -1.2043   0.7329   1.0628   1.8536
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.016631   0.079963  -0.208  0.83524
## dist100       -1.260816   0.166202  -7.586 3.30e-14 ***
## arsenic        0.478695   0.059093   8.101 5.47e-16 ***
## dist100:educ   0.071218   0.024364   2.923  0.00347 **
## arsenic:educ   0.009746   0.008320   1.171  0.24142
## arsenic:assoc -0.099131   0.042913  -2.310  0.02089 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3889.3  on 3014  degrees of freedom
## AIC: 3901.3
##
```

```
## Number of Fisher Scoring iterations: 4
```

```
# since eudc has a high p value
lm5<- update(lm4,.~.-arsenic:educ)
anova(lm5,lm4,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ dist100 + arsenic + dist100:educ + arsenic:assoc
## Model 2: y ~ dist100 + arsenic + dist100:educ + educ:arsenic + arsenic:assoc
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3015     3890.6
## 2      3014     3889.3  1   1.3807     0.24
```

```
summary(lm5)
```

```
##
## Call:
## glm(formula = y ~ dist100 + arsenic + dist100:educ + arsenic:assoc,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.776  -1.208   0.743   1.058   1.910
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.008978   0.079717  -0.113   0.9103
## dist100       -1.375359   0.136462 -10.079  < 2e-16 ***
## arsenic        0.522943   0.046221  11.314  < 2e-16 ***
## dist100:educ   0.092843   0.016130   5.756 8.61e-09 ***
## arsenic:assoc -0.100982   0.042870  -2.356   0.0185 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3890.6  on 3015  degrees of freedom
## AIC: 3900.6
##
## Number of Fisher Scoring iterations: 3
```

```
step(lm5)
```

```
## Start:  AIC=3900.64
## y ~ dist100 + arsenic + dist100:educ + arsenic:assoc
##
##                 Df Deviance    AIC
## <none>                3890.6 3900.6
## - arsenic:assoc  1   3896.2 3904.2
## - dist100:educ   1   3925.0 3933.0
```

```
##
## Call:  glm(formula = y ~ dist100 + arsenic + dist100:educ + arsenic:assoc,
##     family = binomial(link = "logit"))
##
```

```
## Coefficients:
##   (Intercept)        dist100        arsenic   dist100:educ  arsenic:assoc
##     -0.008978      -1.375359       0.522943       0.092843      -0.100982
##
## Degrees of Freedom: 3019 Total (i.e. Null);  3015 Residual
## Null Deviance:       4118
## Residual Deviance: 3891  AIC: 3901
```

When calling the step funtion on lm5, we can see that the the AIC value cannot decrese more.In the model lm5. we can see that almost all the paramter has a p value less than 5%. The arsenic:assoc has a negative effect on the response: when assoc=0 and increase arsenic by 0 unit will increase switch by 0.52 while when assoc-1 and increase arsenic by 1 unit will increase switch by 0.42. Considering dist100 and dist100:educ, they sist100:educ has a positive effect on the result and dist100 has a negative effect on the result. For fixed dist100, by increase eudc by one unit, will increase the response by 0.09dist100 while for fixed educ, by increase one unit of dist100 will affect the result by -1.37+0.09educ. Therefore,for small educ value, it will cause the switch to decrease as dist100 increase and large educ value will cuase switch increasing as dist100 increase.

(c) Try to simplify the model you found in part (b) by replacing educ by a binary factor predictor feduc, constructed as follows:This predictor feduc records whether the person has a primary education (i.e. 1–8 years) or secondary education and above (i.e. more than 9 years). Interpret the final model.

```r
feduc <- numeric(3020)
for(i in 1:3020){
  if(educ[i] < 9){feduc[i] <- 0}
  if(educ[i] > 8){feduc[i] <- 1}
}
```

Now the model becomes:

```r
lm6<-glm(y~dist100+arsenic+dist100:feduc+arsenic:assoc,family=binomial(link="logit"))
summary(lm6)
```

```
##
## Call:
## glm(formula = y ~ dist100 + arsenic + dist100:feduc + arsenic:assoc,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7679  -1.1975   0.7292   1.0455   1.8739
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.00465    0.08025  -0.058    0.954
## dist100       -1.16383    0.11361 -10.244  < 2e-16 ***
## arsenic        0.51480    0.04617  11.149  < 2e-16 ***
## dist100:feduc  1.25270    0.16995   7.371  1.7e-13 ***
## arsenic:assoc -0.08729    0.04313  -2.024    0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3865.0  on 3015  degrees of freedom
## AIC: 3875
##
## Number of Fisher Scoring iterations: 4
```

Compare to the previous model, we can see that there's a decrease in the AIC value and the p value of dist100:feduc has decreased.The parameter of dist100:feduc becomes larger, it means that dist100:feduc now has more effect on the result value: switch. If feduc=0 and increase dist100 by 1 unit, the response switch decrease by -1.16 and when feduc=1 and increase dist100 by 1 unit, the response will increase by 1.25-1.16=0.09 unit.The intercept value become much smaller.
When assoc=0 and increase arsenic by one unit, the switch response will increase by 0.51 while when assoc=1 and increase the arsenic by 1 value will increase the response by 0.423.

(d) Compare the final model in parts (b) and (c) using AIC and ROC curves. Which one do you prefer and why?

Firstly, model in part(c) has a smaller AIC value and the and the p value of dist100:feduc has decreased and The parameter of dist100:feduc becomes larger. It means that dist100:feduc become much important in the prediction model.

```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

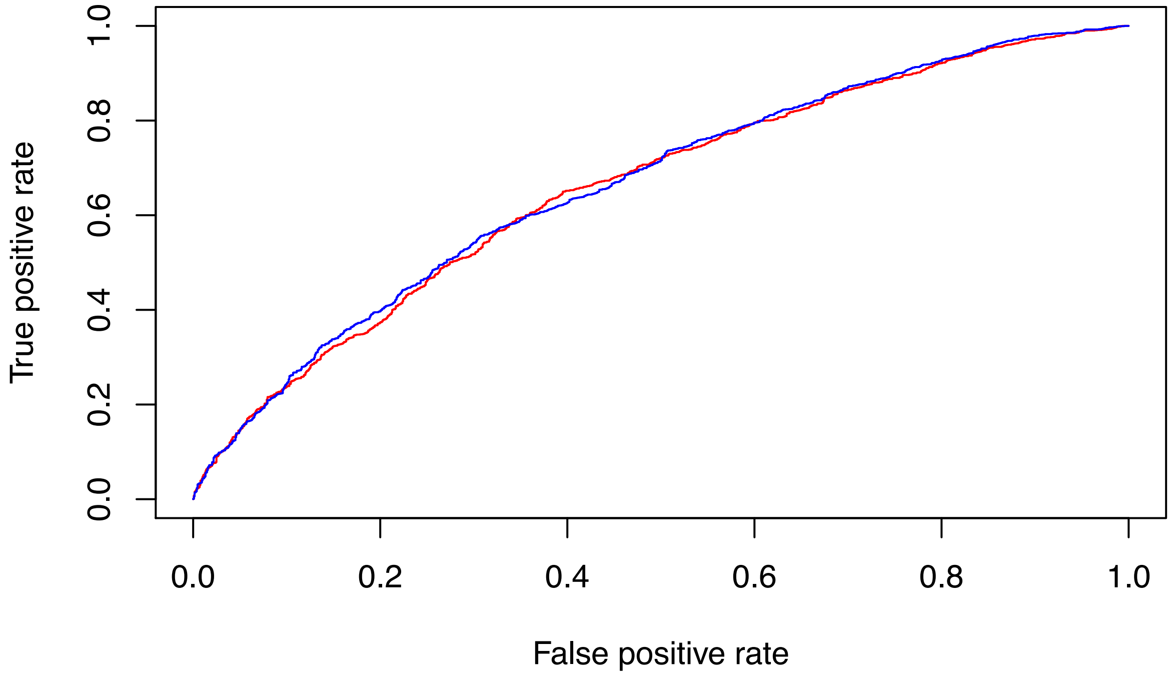
```
pred5<-prediction(fitted(lm5),switch)
pref5<-performance(pred5,"tpr","fpr")
pred6<-prediction(fitted(lm6),switch)
pref6<-performance(pred6,"tpr","fpr")


plot(pref5,col="red")
performance(pred5,"auc")
```

```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
```

```
## [[1]]
## [1] 0.6587405
##
##
## Slot "alpha.values":
## list()
```

```
plot(pref6,add=TRUE,col="blue")
```



```
performance(pred6,"auc")
```

```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
## [[1]]
## [1] 0.6634469
##
##
## Slot "alpha.values":
## list()
```

From the output of performance we can see that 0.663> 0.658, so the model in part(c) is better.