Model For home-well switching

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:

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

```
##
## Call:
## glm(formula = y ~ (dist100 + educ + assoc + arsenic), family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5942  -1.1976   0.7541   1.0632   1.6739
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.156712   0.099601  -1.573    0.116
## dist100     -0.896110   0.104576  -8.569  < 2e-16 ***
## educ         0.042447   0.009588   4.427 9.55e-06 ***
## assoc       -0.124300   0.076966  -1.615    0.106
## arsenic      0.467022   0.041602  11.226  < 2e-16 ***
## ---
## 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: 3907.8  on 3015  degrees of freedom
## AIC: 3917.8
##
## Number of Fisher Scoring iterations: 4
```

```
lm0<-glm(y~(dist100+educ+arsenic),family=binomial(link="logit"))
anova(lm,lm0,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ (dist100 + educ + assoc + arsenic)
## Model 2: y ~ (dist100 + educ + arsenic)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3015     3907.8
## 2      3016     3910.4 -1  -2.6072   0.1064
```

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

```
##
## Call:
## glm(formula = y ~ (dist100 + educ + arsenic)^2, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5706  -1.1964   0.7314   1.0724   1.8712
```

```
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.01228    0.15507  -0.079  0.93686
## dist100         -1.09741    0.25998  -4.221 2.43e-05 ***
## educ            -0.02266    0.02001  -1.133  0.25737
## arsenic          0.46466    0.08636   5.380 7.44e-08 ***
## dist100:educ     0.08067    0.02666   3.026  0.00247 **
## dist100:arsenic -0.11768    0.10353  -1.137  0.25569
## educ:arsenic     0.01806    0.01097   1.647  0.09965 .
## ---
## 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: 3891.7  on 3013  degrees of freedom
## AIC: 3905.7
##
## Number of Fisher Scoring iterations: 4
```