General Linear Model Project

Heran Song

Introduction:

This paper will provide the procedure to analyze and model Logistic Regression Models using R. The paper can be broken down into four sections, first will cover a descriptive analysis of the data, which includes checking for missing values, complete separation, unbalanced data, and collinearity. The next sections will cover model selection, which will describe the steps behind the selection of the logistic model. The third section will cover model diagnostics, which will incorporate residual analysis. The last section will cover model interpretation. The data set used in this analysis will be "Wells in Bangladesh (data from Gelman & Hill, 2007)", which is data collected from a small area of Araihazar upazila, Bangladesh, to see if people with unsafe wells switched to nearby private or community wells or to new wells of their own construction.

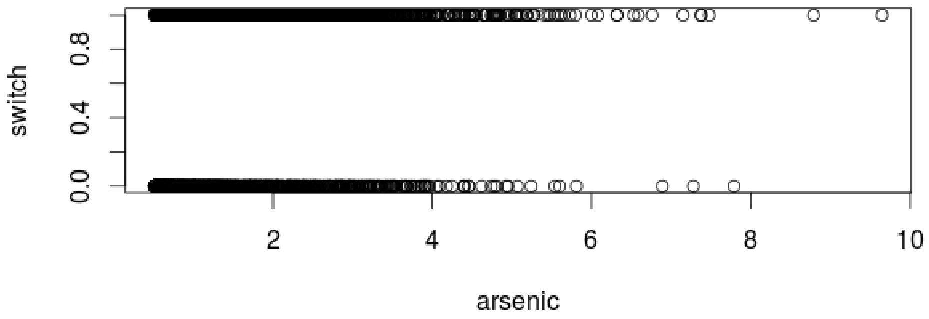
Descriptive Analysis:

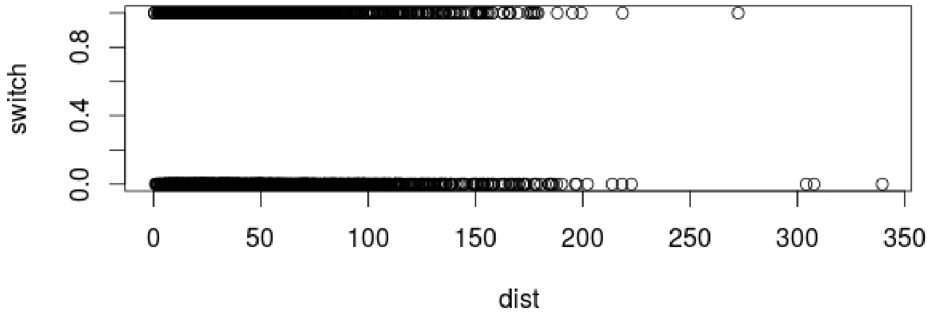
To check for any missing value in our data set, we use the is.na() function, and the grep() function to search for any TRUE statement.

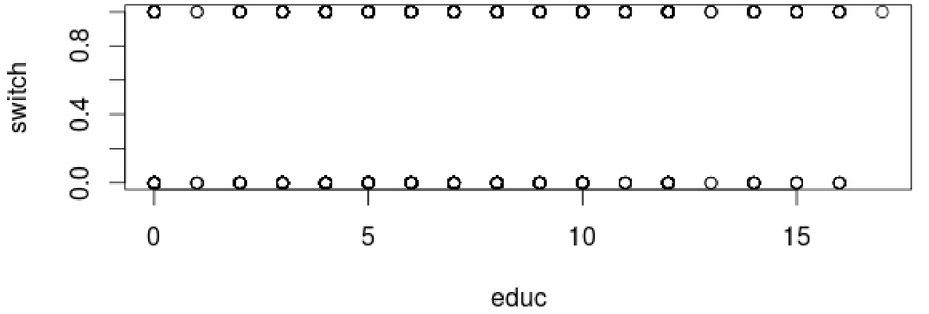
```
> wells<-read.table("./Documents/School/Stat 707/Final Project/wells.txt",header=T)
> grep("TRUE",is.na(wells))
integer(0)
```

To check for complete separation by plotting each continuous predictor against the response variable *switch*.

```
> plot(wells$arsenic,wells$switch,xlab="arsenic",ylab="switch")
> plot(wells$dist,wells$switch,xlab="dist",ylab="switch")
> plot(wells$educ,wells$switch,xlab="educ",ylab="switch")
```







Nothing above indicates complete separation. To check for unbalanced data, we create a contingency table using the function table(). We have two categorical variables *assoc*, and *educ*.

```
> table(wells$switch,wells$assoc)

       0     1
```

```
  0  714  569
  1 1029  708
> table(wells$switch,wells$educ)

      0   1   2   3   4   5   6   7   8   9  10  11  12  13
  0 386   4  24  51  88 345  63  50  87  32  80   7  38   2
  1 503   2  28  70  86 380  67  76 123  58 174  14 109   3

     14  15  16  17
  0  12   8   6   0
  1  20  13  10   1
```

We can see that educ have a zero value at 17, but this shouldn't be a problem since we will need to group our data later on during model selection.

Finally, we check for collinearity in our data. There shouldn't be any valid reason for collinearity since each predictor is not a function of another. But we can double check this by calculating the correlation of each predictors against each other.

```
> cor(wells$arsenic,wells$dist)
[1] 0.1780577
> cor(wells$arsenic,wells$assoc)
[1] -0.02491153
> cor(wells$arsenic,wells$educ)
[1] -0.02956287
> cor(wells$dist,wells$assoc)
[1] -0.003460305
> cor(wells$dist,wells$educ)
[1] -0.02674095
> cor(wells$educ,wells$assoc)
[1] -0.03136667
```

Nothing above indicates collinearity which validates our assumption.
Since there aren't any problems with our dataset, we can go ahead and fit our data using a logistic regression model.

Model Selection:

At the start of our model selection, we immediately encounter a problem due to two of our variables(*arsenic*,*dist*) being continuous and one categorical variable(*educ*) with many levels. This causes most of our EVP's to have observations of less than 5. To solve this problem, we will need to appropriately group our data. While grouping our data solves our EVP problem, but we pay a price in the accuracy of our model as a trade off. So the trick is to find a balance between the two.

```
> w<-aggregate(formula=switch~arsenic+dist+assoc+educ,data=wells,FUN=sum)
> n<-aggregate(formula=switch~arsenic+dist+assoc+educ,data=wells,FUN=length)
> w.n<-data.frame(arsenic=w$arsenic,dist=w$dist,assoc=w$assoc,educ=w$educ,
switch=w$switch,trials=n$switch,proportion=round(w$switch/n$switch,4))

> head(w.n)
  arsenic  dist assoc educ switch trials proportion
1    0.52 2.791     0    0      1      1          1
2    3.84 3.252     0    0      1      1          1
3    1.15 3.612     0    0      1      1          1
4    1.02 3.697     0    0      1      1          1
5    0.62 4.136     0    0      1      1          1
6    0.74 4.594     0    0      0      1          0
> dim(wells)
[1] 3020   7

> wells$educ<-cut(wells$educ, breaks=c(0,4,8,18),right=F)
> wells$dist<-cut(wells$dist, breaks=c(0,20,50,80,120,340),right=F)
> wells$arsenic<-cut(wells$arsenic, breaks=c(0,2,3,10),right=F)
> w<-aggregate(formula=switch˜arsenic+dist+assoc+educ,data=wells,FUN=sum)
> n<-aggregate(formula=switch˜arsenic+dist+assoc+educ,data=wells,FUN=length)
> w.n<-data.frame(arsenic=w$arsenic,dist=w$dist,assoc=w$assoc,educ=w$educ,
switch=w$switch,trials=n$switch,proportion=round(w$switch/n$switch,4))

> head(w.n)
  arsenic    dist assoc  educ switch trials proportion
1   [0,2)  [0,20)     0 [0,4)     58    107     0.5421
2   [2,3)  [0,20)     0 [0,4)     12     12     1.0000
3  [3,10)  [0,20)     0 [0,4)      7      9     0.7778
4   [0,2) [20,50)     0 [0,4)     98    184     0.5326
5   [2,3) [20,50)     0 [0,4)     33     44     0.7500
6  [3,10) [20,50)     0 [0,4)     21     26     0.8077
> dim(w.n)
[1] 90  7

> for(i in 1:dim(w.n)[1]){
+ if(w.n$trials[i]<5)print(w.n[i,])
+ }
   arsenic      dist assoc   educ switch trials proportion
75  [3,10) [120,340)     0 [8,18)      2      4        0.5
   arsenic   dist assoc   educ switch trials proportion
78  [3,10) [0,20)      1 [8,18)      4      4          1
   arsenic     dist assoc   educ switch trials proportion
86   [2,3) [80,120)      1 [8,18)      4      4          1
   arsenic      dist assoc   educ switch trials proportion
90  [3,10) [120,340)     1 [8,18)      2      2          1
```

After grouping our data, we reduced our EVP size from 3020 to 90, with only 4 out of 90 EVP's having sample size of less than 5.

We can now start choosing our model. The first step will be variables selection. Because of our grouping, we only have 4 categorical variables with their respective levels in the dataset. The response variable in this analysis is switch. To determine which explanatory variables should be in the model, we use backward selection. We fit the four possible models with only one explanatory variable, and use Likelihood Ratio Test (LRT) to determine if each variable is important. The p-values for the four tests are displayed below. Using a confidence level of $\alpha = 0.2$.

```
> glm.arsenic<-glm(switch/trials~arsenic,weight=trials, data=w.n,
family=binomial(link="logit"))
> glm.dist<-glm(switch/trials~dist,weight=trials, data=w.n,
family=binomial(link="logit"))
> glm.assoc<-glm(switch/trials~assoc,weight=trials, data=w.n,
family=binomial(link="logit"))
> glm.educ<-glm(switch/trials~educ,weight=trials, data=w.n,
family=binomial(link="logit"))

> anova(glm.arsenic, test="Chisq")$"Pr(>Chi)"
[1]           NA 2.323626e-17
> anova(glm.dist, test="Chisq")$"Pr(>Chi)"
[1]           NA 1.803513e-09
> anova(glm.assoc, test="Chisq")$"Pr(>Chi)"
[1]         NA 0.04851738
> anova(glm.educ, test="Chisq")$"Pr(>Chi)"
[1]           NA 3.178603e-08
```

We see that all four models fit better than the null model, and decide to keep all four variables at this stage.

In the second step, we fit a model including the variables *arsenic*, *dist*, *assoc*, and *educ* selected in step 1 and perform backward elimination using LRT.

```
> glm.fit1<-glm(switch/trials~arsenic+dist+assoc+educ,weight=trials,data=w.n,
family=binomial(link="logit"))
> Anova(mod=glm.fit1,test="LR")
Analysis of Deviance Table (Type II tests)

Response: switch/trials
        LR Chisq Df Pr(>Chisq)
arsenic  102.484  2  < 2.2e-16 ***
dist      68.532  4  4.632e-14 ***
assoc      1.665  1      0.197
educ      35.989  2  1.531e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

According to the LRT, we see that the difference in fit between the full model and the model without assoc is not statistically significant (p=0.197) using $\alpha = 0.1$. The variable *assoc* is therefore not considered further. The variables *arsenic*, *dist*, and *educ* are retained.

The third step will be testing for interactions and transformations of explanatory variables. Since the explanatory variables have all been grouped into categorical, transformations are not applicable here. We focus on the six possible pairwise interactions between *arsenic*, *educ*, *assoc*, and *dist*. To determine whether each interaction is individually important, we fit the six possible models with only one interaction and use likelihood ratio tests.

```
> arsenic.dist<-glm(switch/trials~arsenic+dist+assoc+educ+arsenic:dist,weight=trials,
data=w.n,family=binomial(link="logit"))
> arsenic.assoc<-glm(switch/trials~arsenic+dist+assoc+educ+arsenic:assoc,weight=trials,
data=w.n,family=binomial(link="logit"))
> arsenic.educ<-glm(switch/trials~arsenic+dist+assoc+educ+arsenic:educ,weight=trials,
data=w.n,family=binomial(link="logit"))
> dist.assoc<-glm(switch/trials~arsenic+dist+assoc+educ+dist:assoc,weight=trials,
data=w.n,family=binomial(link="logit"))
> dist.educ<-glm(switch/trials~arsenic+dist+assoc+educ+dist:educ,weight=trials,
data=w.n,family=binomial(link="logit"))
> assoc.educ<-glm(switch/trials~arsenic+dist+assoc+educ+assoc:educ,weight=trials,
data=w.n,family=binomial(link="logit"))

> Anova(arsenic.dist,test="LR")$"Pr(>Chisq)"[5]
[1] 0.1483272
> Anova(arsenic.assoc,test="LR")$"Pr(>Chisq)"[5]
[1] 0.09449453
> Anova(arsenic.educ,test="LR")$"Pr(>Chisq)"[5]
[1] 0.09402494
> Anova(dist.assoc,test="LR")$"Pr(>Chisq)"[5]
[1] 0.8284487
> Anova(dist.educ,test="LR")$"Pr(>Chisq)"[5]
[1] 0.008929027
> Anova(assoc.educ,test="LR")$"Pr(>Chisq)"[5]
[1] 0.3749609
```

Using $\alpha = .05$, out of the six interactions, only the interaction *dist:educ* appears to be important. We perform backward elimination with it.

```
> glm.fit2<-glm(switch/trials~arsenic+dist+educ+dist:educ,weight=trials,data=w.n,
family=binomial(link="logit"))
> Anova(mod=glm.fit2,test="LR")
Analysis of Deviance Table (Type II tests)

Response: switch/trials
          LR Chisq Df Pr(>Chisq)
arsenic    105.433  2  < 2.2e-16 ***
dist        68.839  4  3.991e-14 ***
educ        37.003  2  9.223e-09 ***
```

```
dist:educ   19.796  8    0.01114 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We see that the interaction *dist:educ* is significant at the $\alpha = .05$ level (p=0.011) and should be included in the model.
Our final model will have 8 predictors and 8 interaction terms.

$$logit(\hat{\pi}) = \beta_0 + \beta_{1,2} arsenic + \beta_{3,..,6} dist + \beta_{7,8} educ + \beta_{9,...,16} dist : educ$$

```
> glm.fit<-glm(switch/trials˜arsenic+dist+educ+dist:educ,weight=trials,data=w.n,
family=binomial(link="logit"))
> summary(glm.fit)

Call:
glm(formula = switch/trials ˜ arsenic + dist + educ + dist:educ,
    family = binomial(link = "logit"), data = w.n, weights = trials)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.87195  -0.68326   0.08438   0.65012   2.78912

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.25437    0.13424   1.895  0.05811 .
arsenic[2,3)              0.70580    0.10835   6.514 7.30e-11 ***
arsenic[3,10)             1.10911    0.13170   8.422  < 2e-16 ***
dist[20,50)              -0.04840    0.16646  -0.291  0.77126
dist[50,80)              -0.23558    0.19717  -1.195  0.23216
dist[80,120)             -0.97174    0.23456  -4.143 3.43e-05 ***
dist[120,340)            -1.16429    0.29501  -3.947 7.92e-05 ***
educ[4,8)                -0.02272    0.18575  -0.122  0.90264
educ[8,18)                0.05977    0.19785   0.302  0.76257
dist[20,50):educ[4,8)    -0.16391    0.23141  -0.708  0.47874
dist[50,80):educ[4,8)    -0.09761    0.27003  -0.361  0.71775
dist[80,120):educ[4,8)   -0.11126    0.32170  -0.346  0.72945
dist[120,340):educ[4,8)  -0.50379    0.42651  -1.181  0.23753
dist[20,50):educ[8,18)    0.29075    0.25038   1.161  0.24554
dist[50,80):educ[8,18)    0.49770    0.31428   1.584  0.11328
dist[80,120):educ[8,18)   1.20480    0.38276   3.148  0.00165 **
dist[120,340):educ[8,18)  0.80746    0.42914   1.882  0.05990 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.513  on 89  degrees of freedom
```

```
Residual deviance:  93.181  on 73  degrees of freedom
AIC: 410.29

Number of Fisher Scoring iterations: 4
```

Model Diagnostics:

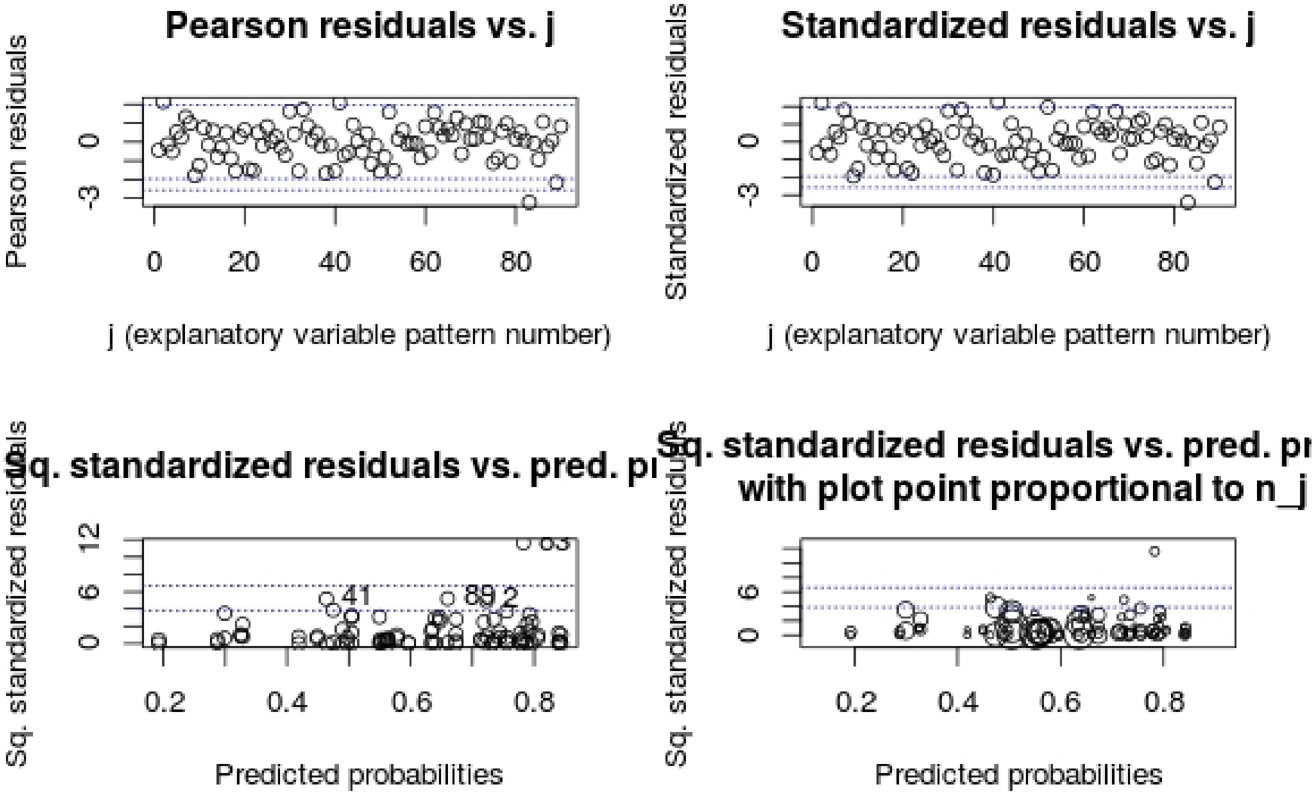
The data in EVP form contains 90 EVPs. Since only 4 EVPs contain less than 5 observations (EVP # 75, 78, 86, and 90), we decide it is reasonable to use large-sample distribution approximations for X^2, G^2 Pearson, and Standardized Pearson residuals.
According to both X^2, G^2, there is no evidence that the model does not fit the data ($X^2 = 88.5285$, p = 0.1042; $G^2 = 93.1815$, p = 0.0558).
Using the Chris Bilder function for diagnostics, we obtain the plots displayed below. EVPs #2, #41, #83, and #89 have unusually large standardized Pearson residuals, which means they are influential outliers.

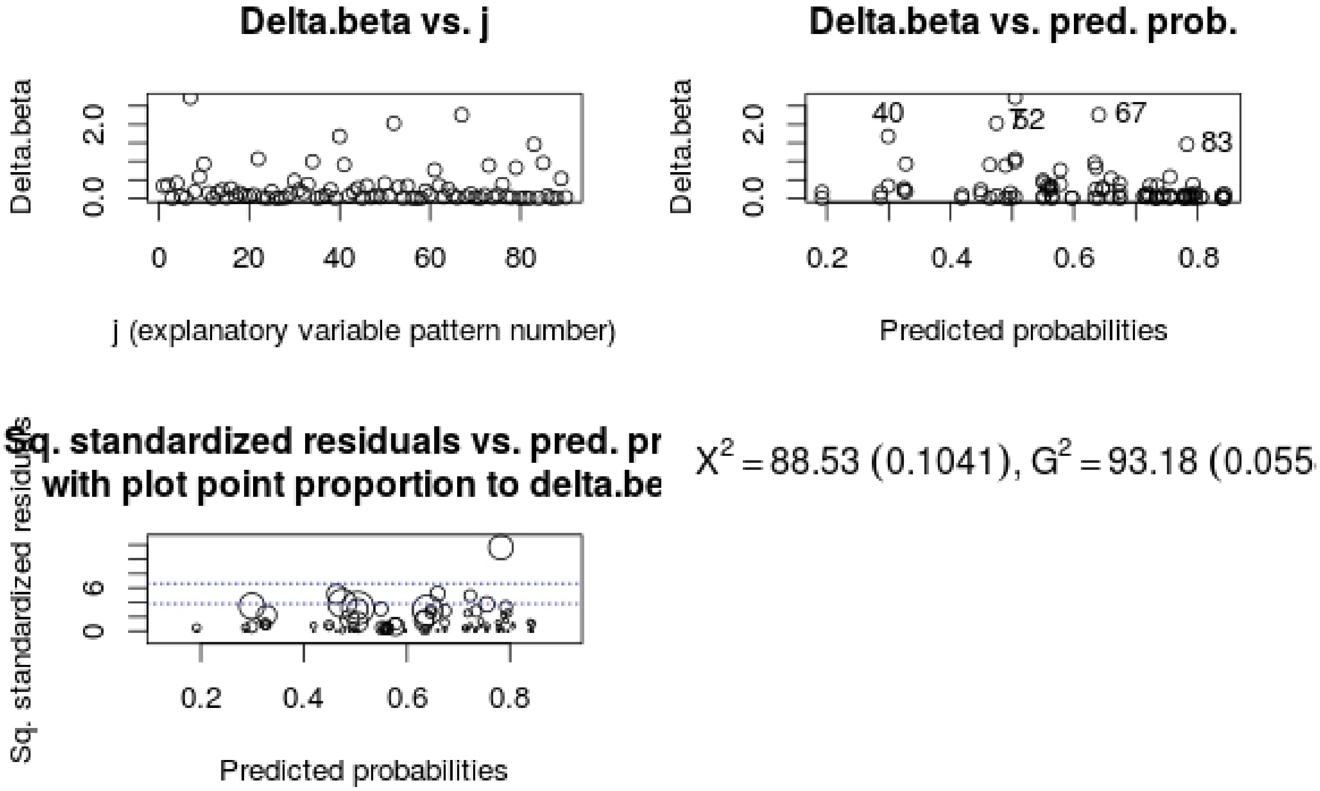
```
> w.n[2,];w.n[41,];w.n[83,];w.n[89,]
  arsenic   dist assoc  educ switch trials proportion
2   [2,3) [0,20)     0 [0,4)     12     12          1
   arsenic     dist assoc  educ switch trials proportion
41   [2,3) [80,120)     0 [4,8)     11     15     0.7333
   arsenic    dist assoc   educ switch trials proportion
83   [2,3) [50,80)     1 [8,18)      6     14     0.4286
   arsenic      dist assoc   educ switch trials proportion
89   [2,3) [120,340)     1 [8,18)      1      5        0.2
```

EVPs #2, #41, and #89 residuals are within 3 standard deviations of the mean of residuals, and so not extremely large. EVP #83 is of concern. Its residual is beyond 3 standard deviations of the mean, and it contains a good number of observations (14 observations).
When the model is refitted without EVP #83, the diagnostics are well-behaved and the fit of the model improves substantially from AIC = 410.29 to AIC = AIC: 397.6. The abnormality of EVP #83 might be related to a problem in the data or problem with grouping. This should be investigated. For the time being we will keep this pattern in the analysis.

```
> w.n1<-w.n[-83,]
> glm.fit3<-glm(switch/trials~arsenic+dist+educ+dist:educ,weight=trials,data=w.n1,
family=binomial(link="logit"))
> AIC(glm.fit3)
[1] 397.6033

> examine.resid(glm.fit)
```


Pearson residuals vs. j
Pearson residuals
0
-3
0
20
40
60
80
j (explanatory variable pattern number)
Standardized residuals vs. j
Standardized residuals
0
-3
0
20
40
60
80
j (explanatory variable pattern number)
Sq. standardized residuals vs. pred. pr
Sq. standardized residuals
12
6
0
63
41
89
2
0.2
0.4
0.6
0.8
Predicted probabilities
Sq. standardized residuals vs. pred. pr
with plot point proportional to n_j
Sq. standardized residuals
6
0
0.2
0.4
0.6
0.8
Predicted probabilities



Model Interpretation:

We use odds ratios to interpret the model coefficients. Profile likelihood 95% confidence intervals for odds ratios. Because of the interaction term, and the number of levels there are many possibilities for comparisons.

Reference category for arsenic : [0,2)
Reference category for dist : [0,20)
Reference category for educ : [0,4)

```
> K<-matrix(data = c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
+                     0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
+                     0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,
+                     0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,
+                     0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,
+                     0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,
+                     0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,
+                     0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,
+                     0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,
+                     0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,
+                    0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,
+                    0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,
+                    0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,
+                    0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,
+                    0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,
+                    0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,
+                    0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,
+                    0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,
+                    0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,
+                    0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,
+                    0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,
+                    0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,
+                    0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,
+                    0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1
+                ), nrow = 24, ncol = 17, byrow = TRUE)
> linear.combo<-mcprofile(object = glm.fit, CM = K)
> ci.log.OR<-confint(object = linear.combo, level = 0.95,adjust = "none")
> comparisons<-c("arsenic[2,3) vs arsenic[0,2)",
+                "arsenic[3,10) vs arsenic[0,2)",
+                "dist[20,50) vs dist[0,20),educ[0,4)",
+                "dist[50,80) vs dist[0,20),educ[0,4)",
+                "dist[80,120) vs dist[0,20),educ[0,4)",
+                "dist[120,340) vs dist[0,20),educ[0,4)",
+                "dist[20,50) vs dist[0,20),educ[4,8)",
+                "dist[50,80) vs dist[0,20),educ[4,8)",
+                "dist[80,120) vs dist[0,20),educ[4,8)",
+                "dist[120,340) vs dist[0,20),educ[4,8)",
+                "dist[20,50) vs dist[0,20),educ[8,18)",
+                "dist[50,80) vs dist[0,20),educ[8,18)",
+                "dist[80,120) vs dist[0,20),educ[8,18)",
+                "dist[120,340) vs dist[0,20),educ[8,18)",
+                "educ[4,8) vs educ[0,4),dist[0,20)",
+                "educ[4,8) vs educ[0,4),dist[20,50)",
+                "educ[4,8) vs educ[0,4),dist[50,80)",
+                "educ[4,8) vs educ[0,4),dist[80,120)",
+                "educ[4,8) vs educ[0,4),dist[120,340)",
+                "educ[8,18) vs educ[0,4),dist[0,20)",
+                "educ[8,18) vs educ[0,4),dist[20,50)",
+                "educ[8,18) vs educ[0,4),dist[50,80)",
+                "educ[8,18) vs educ[0,4),dist[80,120)",
+                "educ[8,18) vs educ[0,4),dist[120,340)")
> data.frame("Comparisons/fixed value"=comparisons, OR=round(exp(ci.log.OR$estimate),2),
+ OR.CI = round(exp(ci.log.OR$confint),2))
                   Comparisons.fixed.value Estimate OR.CI.lower OR.CI.upper
C1            arsenic[2,3) vs arsenic[0,2)     2.14        1.73        2.66
```

```
C2           arsenic[3,10) vs arsenic[0,2)     3.04        2.35        3.95
C3     dist[20,50) vs dist[0,20),educ[0,4)     0.95        0.68        1.31
C4     dist[50,80) vs dist[0,20),educ[0,4)     0.79        0.53        1.16
C5    dist[80,120) vs dist[0,20),educ[0,4)     0.37        0.24        0.59
C6   dist[120,340) vs dist[0,20),educ[0,4)     0.31        0.17        0.55
C7     dist[20,50) vs dist[0,20),educ[4,8)     0.81        0.59        1.10
C8     dist[50,80) vs dist[0,20),educ[4,8)     0.72        0.50        1.03
C9    dist[80,120) vs dist[0,20),educ[4,8)     0.34        0.22        0.52
C10  dist[120,340) vs dist[0,20),educ[4,8)     0.19        0.10        0.34
C11   dist[20,50) vs dist[0,20),educ[8,18)     1.28        0.88        1.84
C12   dist[50,80) vs dist[0,20),educ[8,18)     1.63        0.98        2.74
C13  dist[80,120) vs dist[0,20),educ[8,18)     1.26        0.70        2.32
C14 dist[120,340) vs dist[0,20),educ[8,18)     0.70        0.38        1.29
C15      educ[4,8) vs educ[0,4),dist[0,20)     0.98        0.68        1.41
C16     educ[4,8) vs educ[0,4),dist[20,50)     0.83        0.63        1.09
C17     educ[4,8) vs educ[0,4),dist[50,80)     0.89        0.60        1.31
C18    educ[4,8) vs educ[0,4),dist[80,120)     0.88        0.52        1.47
C19    educ[4,8) vs educ[0,4),dist[120,340)    0.59        0.27        1.25
C20     educ[8,18) vs educ[0,4),dist[0,20)     1.06        0.72        1.56
C21    educ[8,18) vs educ[0,4),dist[20,50)     1.42        1.05        1.92
C22    educ[8,18) vs educ[0,4),dist[50,80)     2.19        1.33        3.68
C23   educ[8,18) vs educ[0,4),dist[80,120)     3.57        1.90        6.87
C24  educ[8,18) vs educ[0,4),dist[120,340)     2.39        1.14        5.10
```

Giving some examples, comparing wells with arsenic level between 2 to 3 hundred micrograms per liter to wells with arsenic level between 0 to 2 hundred micrograms per liter (row C1 in the table above), we can say that, with 95% confidence, the odds of switching wells are between 1.73 and 2.66 times as large for wells with arsenic level between 2 to 3 hundred micrograms per liter. Because 1 is not within the interval, there is sufficient evidence to conclude that households who has wells with higher arsenic level are more likely to switch wells comparing to households with lower arsenic level wells.

Comparing households within 50 to 80 meters to the closest known safe well, to households within 0 to 20 meters to the closest known safe well, given both with 0 to 4 years of education (row C4 in the table above), we can say that, with 95% confidence, the odds of switching wells are between 0.53 and 1.16 times for households within 50 to 80 meters to the closest known safe well with 0 to 4 years of education. Because 1 is within the interval, there is not sufficient evidence to conclude that households that are further away from a safe well with 0 to 4 years of education will switch well.

Comparing households with 8 to 18 years of education, to households with 0 to 4 years of education given both are within 20 to 50 meters to the closest known safe well (row C21 in the table above), we can say that, with 95% confidence, the odds of switching wells are between 1.05 and 1.92 times as large for households with 8 to 18 years of education who are within 20 to 50 meters to the closest known safe well. Because 1 is not within the interval, there is sufficient evidence to conclude that households with more years of education within 20 to 50 meters to the closest known safe well will switch well.

Conclusion:

In general, we see that higher arsenic level in wells definitely increase the probability of switching wells.

There is insufficient evidence that households within 20 to 80 meters to the closest known safe well will switch wells, comparing to households within 0 to 20 meters, given both with 0 to 4 years of education. But there is evidence pointing to that households within 80 to 340 meters to the closets known safe well with 0 to 4 years of education will less likely to switch wells comparing to households within 0 to 20 meters with 0 to 4 years of education.

There is insufficient evidence that households within 20 to 80 meters to the closest known safe well will switch wells, comparing to households within 0 to 20 meters, given both with 4 to 8 years of education. But there is evidence pointing to that households within 80 to 340 meters to the closets known safe well with 4 to 8 years of education will less likely to switch wells comparing to households within 0 to 20 meters with 4 to 8 years of education.

There is insufficient evidence that households with 4 to 8 years of education will switch wells comparing households with 0 to 4 years of education, as the given distance to the closest known safe well increases for both house holds.

There is insufficient evidence that households with 8 to 18 years of education will switch wells, comparing households with 8 to 4 years of education, given both households are 0 to 20 meters to the closets known safe well. But there is enough evidence to conclude households with 8 to 18 years of education will switch wells comparing households with 4 to 8 years of education, given both households are 20 to 340 meters to the closets known safe well.

In conclusion, arsenic level definitely increases the probability of switch wells. More years of education also increases the probability of switching well. Longer distance on the other hand decreases the probability of switching well. Our model points out the general trend of our data, but it fails to incorporate the finer details due to the grouping of our data. The abnormality of EVP #83 also seems to point out some problems. While this might be problems arises from our grouping or problem with the dataset itself, further investigation is needed to create a more accurate model.