BTRY 6020 Homework V

NAME: Andres Castano

NETID: ac986

DUE DATE: 8:40 am Friday March 31

Question 1.

Health officials wonder why some people get the flu shot while others don't. In a study designed to shed some light on this, researchers asked a random sample of patients if they had gotten a flu shot, recorded their age and gender, and also gave each a written questionnaire designed to evaluate their health awareness index. Data appear in Hwk5Q1DatSp17. Note here that Y = 1 means they received the flu shot and that males were coded as $X_3 = 1$, females coded as $X_3 = 0$.

A) Obtain the maximum likelihood estimators of β_0, β_1, β_2, and β_3. State the fitted regression function.

```
library(readxl)
data_flu = read_excel("Hwk5Q1DatSp17.xlsx")
head(data_flu)
```

```
##   ObsNum FShot Age Ind Gen
## 1      1     0  59  52   0
## 2      2     0  61  55   1
## 3      3     1  82  51   0
## 4      4     0  51  70   0
## 5      5     0  53  70   0
## 6      6     0  62  49   1
```

```
#Model
flu.glm=glm(FShot~Age + Ind + Gen, family = binomial, data=data_flu)
#Summary
summary(flu.glm)
```

```
##
## Call:
## glm(formula = FShot ~ Age + Ind + Gen, family = binomial, data = data_flu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4037  -0.5637  -0.3352  -0.1542   2.9394
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## Age          0.07279    0.03038   2.396  0.01658 *
## Ind         -0.09899    0.03348  -2.957  0.00311 **
```

```
## Gen          0.43397    0.52179   0.832  0.40558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

The fitted regression function is:

$$logit(\pi) = log_e(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = -1.17716 + 0.07279 * Age_i - 0.09899 * Ind_i + 0.43397 * Gen_i$$

B) What is the estimated probability of getting the flu shot that a male clients aged 55 years with a health awareness score of 60?

$$\hat{\pi}_i = \hat{p}(FShot = 1|Age = 55, Ind = 60, Gen = 1) = \frac{1}{1 + exp(-(-1.17716 + 0.07279 * 55 - 0.09899 * 60 + 0.43397 * 1))}$$

```
# Manually
coef=coefficients(flu.glm)
coef
```

```
## (Intercept)         Age         Ind         Gen
## -1.17715922  0.07278802 -0.09898649  0.43397485
```

```
b0=coef[1]
b1=coef[2]
b2=coef[3]
b3=coef[4]
x1=55
x2=60
x3=1
prob_estimated = (1 ) / (1 + exp(-(b0 + b1*x1 + b2*x2 + b3*x3)))
prob_estimated
```

```
## (Intercept)
##  0.06422197
```

```
#Verification with R command
newdata=data.frame(Age=55, Ind=60, Gen=1)
predict(flu.glm, newdata, type="response")
```

```
##          1
## 0.06422197
```

The probability of getting flu for the person described is $0.06422197 \approx 6.42\%$

C) Obtain the VIFs for the regression predictors. What conclusions can you reach from these statistics?

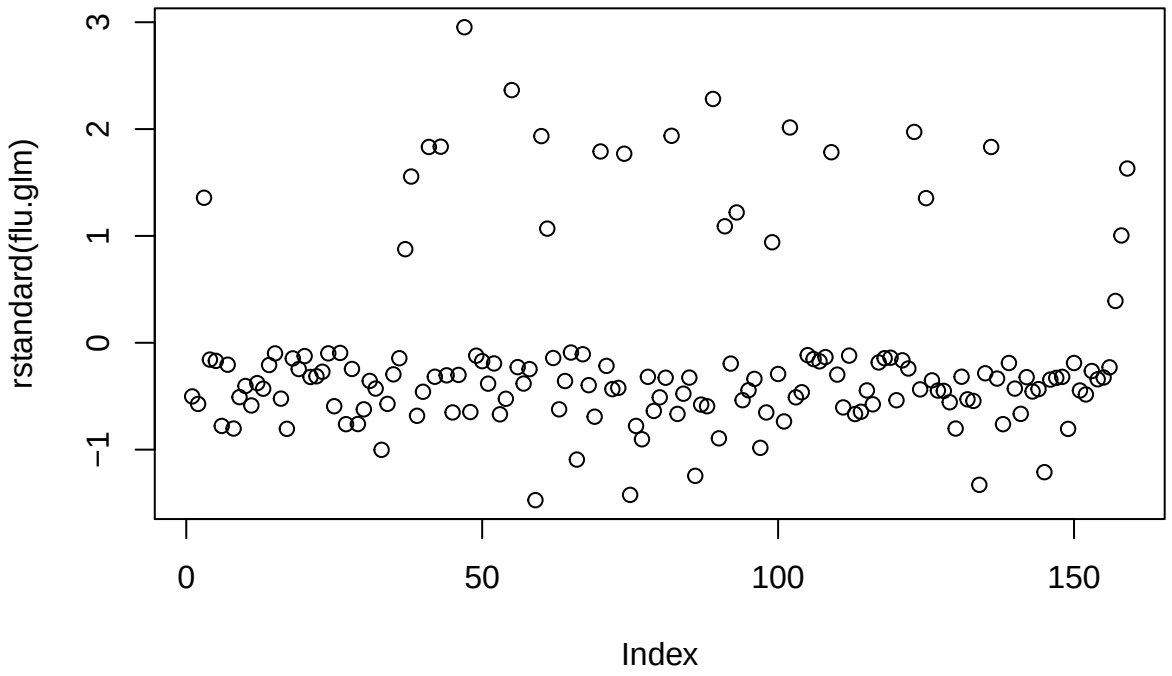
```
library(car)
vif(flu.glm)
```

```
##      Age      Ind      Gen
## 1.091111 1.081049 1.048432
```

Since all the VIFs are less than 10 we can say that we do not have a multicolinearity problem. Here the the VIFs of interest are for the quantitative variables Age and Ind.

D) Get the standardized deviance residuals and plot against observation number. Does there appear to be any outliers?

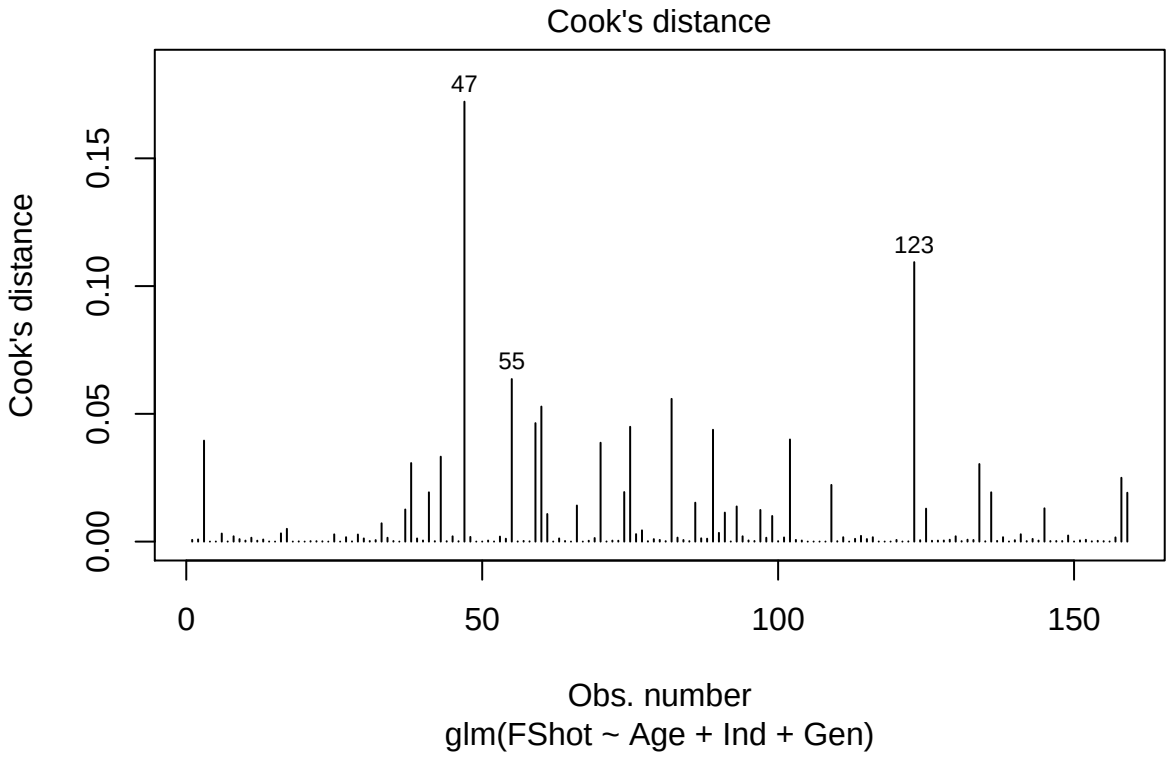
```
#data_flu$flu.stdres=rstandard(flu.glm)
plot(rstandard(flu.glm))
```



It seems that we do not have outliers.

E) Get the Cook's distance numbers and plot against observation number. Do there appear to be any influential outliers? If so, check their effects.

```
#data_flu$flu.cooks = cooks.distance(flu.glm)
library(car)
plot(flu.glm, which=4)
```



There are two observations that are highly influential (observations 47 and 123). We are going to drop this observations, run again to model and then compare the coefficients against the model with all the observations.

```
# drop observation 47 and 123
data_flu_2 = data_flu[data_flu$ObsNum!=47 & data_flu$ObsNum!= 123,]
# re run the model
flu2.glm=glm(FShot~Age + Ind + Gen, family = binomial, data=data_flu_2)
summary(flu2.glm)
```

```
##
## Call:
## glm(formula = FShot ~ Age + Ind + Gen, family = binomial, data = data_flu_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5012  -0.5064  -0.2867  -0.1147   2.5339
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93040    3.33632  -0.579  0.56286
## Age          0.08918    0.03379   2.639  0.00831 **
## Ind         -0.11168    0.03771  -2.962  0.00306 **
## Gen          0.75579    0.57566   1.313  0.18922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.23  on 156  degrees of freedom
## Residual deviance:  91.81  on 153  degrees of freedom
```

```
## AIC: 99.81
##
## Number of Fisher Scoring iterations: 6
```

Comparing the models, We can observe:

- The standard error of the coefficients have increased in the model without the observations 47 and 123. The standard errors have increased by 11.9% for $\hat{\beta}_0$, 11.2% for $\hat{\beta}_1$, 12.6% for $\hat{\beta}_2$ and 10.3% for $\hat{\beta}_3$. The rate of change among the models was calculate as $change(\%) = (\frac{SE_{\beta_j}^{flu2} - SE_{\beta_j}^{flu1}}{SE_{\beta_j}^{flu1}}) * 100$, where $SE_{\beta_j}^{flu1}$ is the standard error for the coefficient j (j=1,2,3,4) in the model with all observations and $SE_{\beta_j}^{flu2}$ is the standard error for the coefficient j (j=1,2,3,4) in the model that excludes observations 47 and 123.
- The estimated coefficients have also change by 64%, 22.5%, 12.8% and 74.15%, respectively. The calculation of this rate of change in the estimated coefficients follow the same logic used for the standard error of the coefficients.
- The pseudo R square in the model with all the observations is $R_1^2 = 1 - \frac{deviance_{full}^{flu1}}{deviance_{null}^{flu1}} = 1 - \frac{105.09}{134.94} = 0.22$, meanwhile The pseudo R square in the model without observations 47 and 123 is $R_2^2 = 1 - \frac{deviance_{full}^{flu2}}{deviance_{null}^{flu2}} = 1 - \frac{91.81}{127.23} = 0.28$. Thus, deleting the two influential points the model also gain 6 percentage points in its ability to fit the data.
- The second model should be prefered for prediction purposes since the Akaike information criteria decrease almost 14 units in the model without the influential points compared to the model with all the data ($AIC^{flu2} = 99.81$ and $AIC^{flu1} = 113.09$).

F) Can we drop Age and Gender if we keep the health awareness index in the model? State hypotheses, test statistic, p-value, and conclusions.

Since not clarification is provided, I going to use the database with all the observation to answer this question.

Here we are interested in a simultaneous test for β_1 and β_3:

$$H_0 : \beta_1 = \beta_3 = 0$$
$$H_A : not H_0$$

We can test for this using a likelihood ratio test. The test statistis is defined as the difference in the residual deviance of the full model (the model with all the explanatory variables) and the residual deviance of the null model (the model with only the health awareness index as explanatory variable):

$$TS = D_{null} - D_{full}$$

Where D_{null} is the residual deviance of the null model and D_{full} is the residual deviance of the full model. This test statistic follow a χ^2 distribution with k degrees of freedon. So the p value is $p = P(\chi_k^2 > TS)$

We can make this test in R as follows:

```
# To run the null model (only with Ind as explanatory variable)
flu_null.glm=glm(FShot~Ind, family = binomial, data=data_flu)
summary(flu_null.glm)
```

```
##
## Call:
## glm(formula = FShot ~ Ind, family = binomial, data = data_flu)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.3944  -0.5926  -0.3999  -0.2369   2.8476
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.91133    1.62651    3.02  0.00253 **
## Ind         -0.11931    0.03013   -3.96  7.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 113.20  on 157  degrees of freedom
## AIC: 117.2
##
## Number of Fisher Scoring iterations: 5
```

```
anova(flu_null.glm, flu.glm, test = 'LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: FShot ~ Ind
## Model 2: FShot ~ Age + Ind + Gen
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       157     113.20
## 2       155     105.09  2   8.1026   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, the test statistic is:

$$TS = 113.20 - 105.09 = 8.1026$$

The p-value is $p = P(\chi^2_2 > 8.1026) = 0.0174$

Then, we have evidence to reject H_0, Which in turns means that we can not drop both Age and Gender of the model.

G) Install the package "bestglm". Visit the following website:

https://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf

to learn how to use this package. Don't forget the "library(bestglm)" command before you use it.

i) Find the best model for getting a flu shot according to the BIC criteria

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```
library(bestglm)
```

```
## Warning: package 'bestglm' was built under R version 3.3.2
```

```
str(data_flu)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    159 obs. of  5 variables:
##  $ ObsNum: num  1 2 3 4 5 6 7 8 9 10 ...
##  $ FShot : num  0 0 1 0 0 0 0 0 0 0 ...
```

```
##  $ Age   : num  59 61 82 51 53 62 51 70 71 55 ...
##  $ Ind   : num  52 55 51 70 70 49 69 54 65 58 ...
##  $ Gen   : num  0 1 0 0 0 1 1 1 1 1 ...
```

```
#Reorganazing the data in the format need it for the package
data_flu_new=data_flu[,-1]
data_glm=data_flu_new[,c(2,3,4,1)]
# best model according to the BIC criteria
bestglm(data_glm, family=binomial,  IC="BIC")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## BIC
## BICq equivalent for q in (0.237623513993524, 0.898744026033073)
## Best Model:
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.45778309 2.91533637 -0.5000394 0.617047325
## Age          0.07787235 0.02969670  2.6222563 0.008734971
## Ind         -0.09547230 0.03240764 -2.9459813 0.003219318
```

The best model according to this procedure is a model with Age and Ind as explanatory variables

ii) Find the best models for a 0, 1, 2, and 3 predictors using the Subsets command

Here it is not clear which criterion we need to use to rank the models, but try to do my best guess, I would do it also with the BIC criterion.

```
best_bic= bestglm(data_glm, family=binomial,  IC="BIC", TopModels = 4)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
best_bic$Subsets
```

```
##    Intercept   Age   Ind   Gen logLikelihood      BIC
## 0       TRUE FALSE FALSE FALSE     -67.47038 134.9408
## 1       TRUE FALSE  TRUE FALSE     -56.59790 118.2647
## 2*      TRUE  TRUE  TRUE FALSE     -52.89769 115.9332
## 3       TRUE  TRUE  TRUE  TRUE     -52.54659 120.2999
```

- The best 0 predictor model is the model with only the intercept.
- The best 1 predictor model is the model with only Ind as explanatory variable
- The best 2 predictor model is the model with Age and Ind as explanatory variables
- The best 3 predictor model is the models with all the predictors

iii) Find the best model for getting a flu shot according to the AIC criteria

```
# best model according to the AIC
bestglm(data_glm, family=binomial, IC="AIC")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## AIC
## BICq equivalent for q in (0.237623513993524, 0.898744026033073)
## Best Model:
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.45778309 2.91533637 -0.5000394 0.617047325
## Age          0.07787235 0.02969670  2.6222563 0.008734971
## Ind         -0.09547230 0.03240764 -2.9459813 0.003219318
```

The best model using the AIC criteria is a model with Age and Ind as explanatory variables.

iv) Find the best models for a 0, 1, 2, and 3 predictors using the Subsets command

```
best_aic= bestglm(data_glm, family=binomial,  IC="AIC", TopModels = 5)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
best_aic$Subsets
```

```
##    Intercept   Age   Ind   Gen logLikelihood      AIC
## 0       TRUE FALSE FALSE FALSE     -67.47038 134.9408
## 1       TRUE FALSE  TRUE FALSE     -56.59790 115.1958
## 2*      TRUE  TRUE  TRUE FALSE     -52.89769 109.7954
## 3       TRUE  TRUE  TRUE  TRUE     -52.54659 111.0932
```

- The best 0 predictor model is the model with only the intercept.
- The best 1 predictor model is the model with only Ind as explanatory variable
- The best 2 predictor model is the model with Age and Ind as explanatory variables
- The best 3 predictor model is the models with all the predictors

v) What model from the above models evaluated would you choose for this situation? Explain BRIEFLY; you may include data from all parts of Question 1.

Since the two selection criteria arrived to the same best model (the model with Age and Ind as explanatory variables), I consider this is the best model for this situation. We can also defend this idea observing that there are not gains in the AIC result when gender is included as the third explanatory variable (AIC with Age and Ind as explanatory variables=109.79, and AIC including all the predictors = 111.0932).

Question 2.

A disease outbreak has occurred in a certain city. Data have been collected on a random telephone survey of 196 people within city limits and the following data recorded: 1) Whether or not they have contracted the disease (Dis, =1 if they have, =0 if not), Age, Socioeconomic Status (SES, = 1 if upper, = 2 if middle, = 3 if lower), Sector of the city they live (Sect, either sector 1 or sector 2), and saving account status (Sav, = 1 if they have a savings account, = 0 if not). data appear in Hwk5Q2DatSp17

Part A) Develop a logistic regression model for predicting the probability of contracting this disease, using the above variables. Be sure to check for polynomial effects of significant quantitative variables as well as interactions between significant predictor variables. When finished, explicitly state your prediction equation. Be sure to show significant steps in model development, using simultaneous tests when you want to omit/test more than one predictor.

Step 1: enter the data into R for the analysis and define as categorical variables as such.

```
library(readxl)
data_disease = read_excel("Hwk5Q2DatSp17.xlsx")
head(data_disease)
```

```
##   Obs Age SES Sect Dis Sav
## 1   1  33   1    1   0   1
## 2   2  35   1    1   0   1
## 3   3   6   1    1   0   0
## 4   4  60   1    1   0   1
## 5   5  18   3    1   1   0
## 6   6  26   3    1   0   0
```

```
# define the variables SES, sector, saving account status
data_disease$SES = factor(data_disease$SES, labels = c("Upper", "Middle", "Lower"))
data_disease$Sect = factor(data_disease$Sect, labels = c("Sector1", "Sector2"))
data_disease$Sav = factor(data_disease$Sav, labels = c("NOT", "YES"))
head(data_disease)
```

```
##   Obs Age   SES    Sect Dis Sav
## 1   1  33 Upper Sector1   0 YES
## 2   2  35 Upper Sector1   0 YES
## 3   3   6 Upper Sector1   0 NOT
## 4   4  60 Upper Sector1   0 YES
## 5   5  18 Lower Sector1   1 NOT
## 6   6  26 Lower Sector1   0 NOT
```

Step 2: Run Backward elimination using AIC criteria.

```
glm_disease_null=glm(Dis~1, family = binomial, data=data_disease)
glm_disease_full=glm(Dis~ Age  + SES + Sect + Sav, family = binomial, data=data_disease)
# AIC based backward elimination
step(glm_disease_full, direction="backward", trace = 1)
```

```
## Start:  AIC=223.21
## Dis ~ Age + SES + Sect + Sav
##
##        Df Deviance    AIC
## - SES   2   211.54 219.54
## - Sav   1   211.22 221.22
## <none>      211.21 223.21
## - Age   1   220.61 230.61
```

```
## - Sect  1   224.01 234.01
##
## Step:  AIC=219.54
## Dis ~ Age + Sect + Sav
##
##        Df Deviance    AIC
## - Sav   1   211.64 217.64
## <none>      211.54 219.54
## - Age   1   221.19 227.19
## - Sect  1   224.05 230.05
##
## Step:  AIC=217.64
## Dis ~ Age + Sect
##
##        Df Deviance    AIC
## <none>      211.64 217.64
## - Age   1   221.60 225.60
## - Sect  1   224.32 228.32
```

```
##
## Call:  glm(formula = Dis ~ Age + Sect, family = binomial, data = data_disease)
##
## Coefficients:
## (Intercept)          Age  SectSector2
##    -2.15966      0.02681      1.18169
##
## Degrees of Freedom: 195 Total (i.e. Null);  193 Residual
## Null Deviance:       236.3
## Residual Deviance: 211.6     AIC: 217.6
```

The best model according to these criteria is a model that includes Age and Sect as explanatory variables.

Step 3: Run best subsets using the BIC criteria to get an idea of the best models. Then compare the best models with those obtained using the AIC criteria.

```
# Organize the data as required for using bestglm command
data_disease_new=data_disease[,-1]
data_disease_glm=data_disease_new[,c(1,2,3,5,4)]
# all subsets using "bestglm"
best_disease_bic= bestglm(data_disease_glm, family=binomial,  IC="BIC", TopModels = 5)
```

```
## Morgan-Tatar search since family is non-gaussian.
## Note: factors present with more than 2 levels.
```

```
best_disease_bic$Subsets
```

```
##    Intercept   Age   SES  Sect   Sav logLikelihood      BIC
## 0       TRUE FALSE FALSE FALSE FALSE     -118.1647 236.3293
## 1       TRUE FALSE FALSE  TRUE FALSE     -110.7980 226.8740
## 2*      TRUE  TRUE FALSE  TRUE FALSE     -105.8196 222.1955
## 3       TRUE  TRUE FALSE  TRUE  TRUE     -105.7719 227.3782
## 4       TRUE  TRUE  TRUE  TRUE  TRUE     -105.6047 237.6000
```

```
#best_disease_bic$BestModels
```

According to these criteria the best models are:

- best one predictor model includes only Sect as explanatory variable

- best two predictor model includes Age and Sect as explanatory variables
- best three predictor model includes Age, Sect and Sav as explanatory variables
- Best fourth predictor model includes Age, Sect, Sav, and SES as explanatory variables. For SES, we will see which of its categories provide the better fit.

Step 4: Run the set of best candidate models and get their summaries in R.

Here, I will use the best 2, 3 and 4 predictors model (the best model according to both procedures use in step 2 and 3 is a model with Age and Sect as explanatory variables). I discard the one predictor model because I do believe that a phenomenon of such complexity as a the probability to contract a disease during an outbreak should be determine by more tha one variable.

```
# Best two predictor model
two_predictor.glm=glm(Dis~Age+Sect, family = binomial, data=data_disease_glm)
summary(two_predictor.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect, family = binomial, data = data_disease_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6839  -0.8199  -0.5606   1.0093   2.0275
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.15966    0.34388  -6.280 3.38e-10 ***
## Age          0.02681    0.00865   3.100 0.001936 **
## SectSector2  1.18169    0.33696   3.507 0.000453 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 211.64  on 193  degrees of freedom
## AIC: 217.64
##
## Number of Fisher Scoring iterations: 3
```

```
# Best three predictor model
three_predictor.glm=glm(Dis~Age+Sect+Sav, family = binomial, data=data_disease_glm)
summary(three_predictor.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + Sav, family = binomial, data = data_disease_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6884  -0.8373  -0.5716   1.0033   2.0605
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.130424   0.355807  -5.988 2.13e-09 ***
## Age          0.027696   0.009138   3.031 0.002438 **
## SectSector2  1.207080   0.347433   3.474 0.000512 ***
```

```
## SavYES      -0.114174   0.370027  -0.309 0.757658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 211.54  on 192  degrees of freedom
## AIC: 219.54
##
## Number of Fisher Scoring iterations: 3
```

```
# Best fourth predictor model (it is not really a 4 predictor model, but a 5 predictor model since SES
fourth_predictor.glm=glm(Dis~Age+Sect+Sav+SES, family = binomial, data=data_disease_glm)
summary(fourth_predictor.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + Sav + SES, family = binomial,
##     data = data_disease_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6614  -0.8309  -0.5630   1.0134   2.0918
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.273558   0.479333  -4.743  2.1e-06 ***
## Age          0.027280   0.009132   2.987 0.002813 **
## SectSector2  1.249464   0.357009   3.500 0.000466 ***
## SavYES      -0.040692   0.396540  -0.103 0.918266
## SESMiddle    0.035578   0.441452   0.081 0.935765
## SESLower     0.237633   0.433750   0.548 0.583789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 211.21  on 190  degrees of freedom
## AIC: 223.21
##
## Number of Fisher Scoring iterations: 4
```

Step 5: comapre the results of the different models and decide which one is the best for prediction.

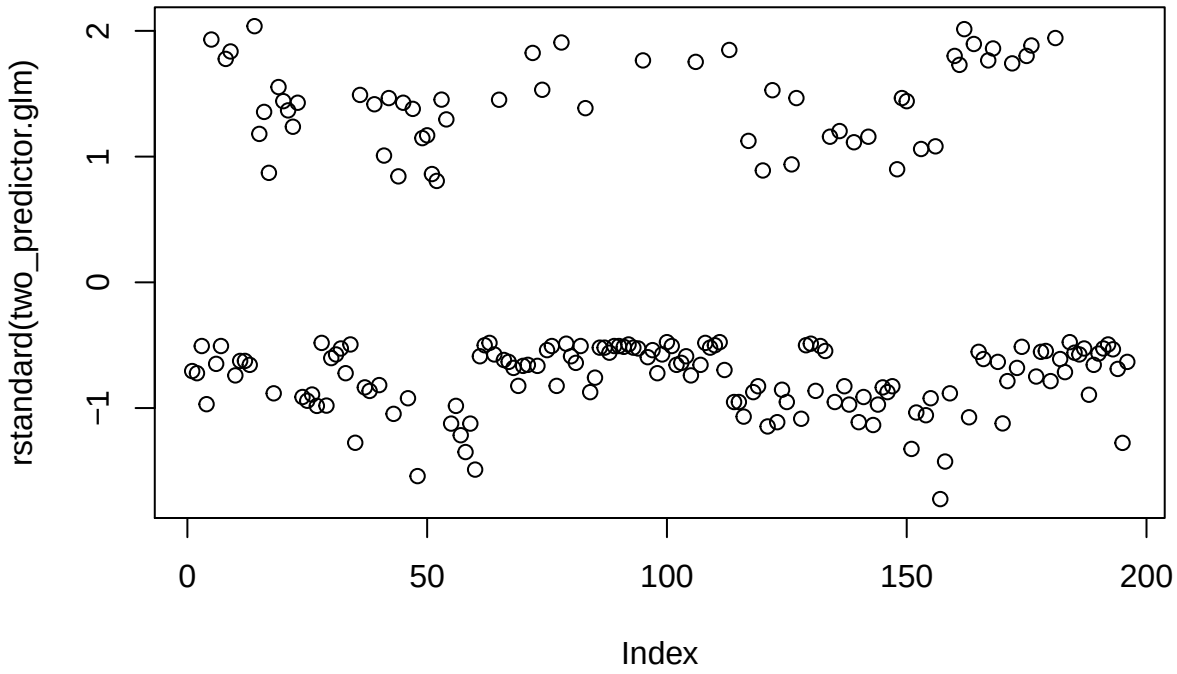
Here, We are going to compare the results of the models based on the pseudo-R square, Residual Deviance and the Akaike Criterion:

Model	Residual Deviance	pseudo-r2	AIC
Two predictors	211.64	0.1045	217.64
Three predictors	211.54	0.1049	219.54
Five predictors	211.21	0.1063	223.21

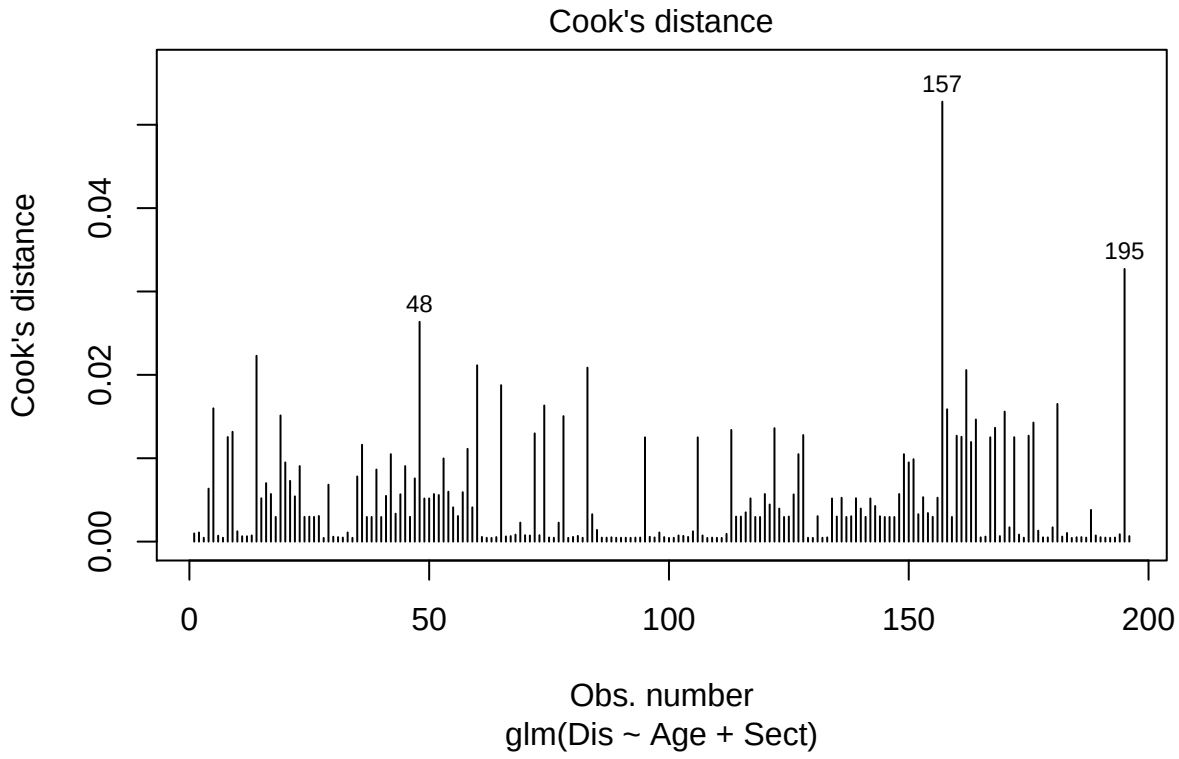
As we can see the gains to include Sav and SES as predictors are minimum in terms of residual deviance, and pseudo R square. Then, based on these results, I consider that the best model is the two predictor model, i.e, the one with only Age and Sect as explanatory variables.

Step 6: Check for suspect and influential data points

```
# Suspect data points
#data_flu$flu.stdres=rstandard(flu.glm)
# Check for suspect points
plot(rstandard(two_predictor.glm))
```




```
# Check for influential points
library(car)
plot(two_predictor.glm, which=4)
```



The graphs above teall us that we do not have outliers, but at the same time the cook´s distance grpah tell us that we have two observation highly influential (observations 195 and 157).

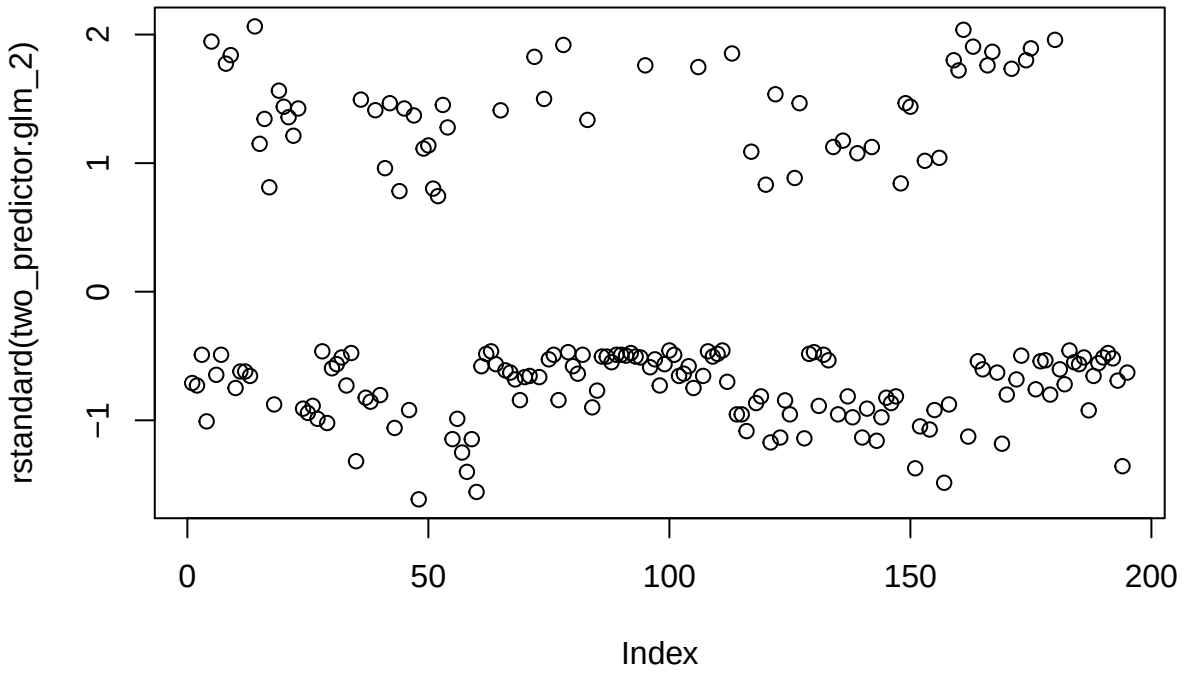
Let's drop the point 157, run again the model and the check for influential data points.

```
data_disease_glm_2 = data_disease_glm[-157,]
two_predictor.glm_2=glm(Dis~Age+Sect, family = binomial, data_disease_glm_2)
summary(two_predictor.glm_2)
```

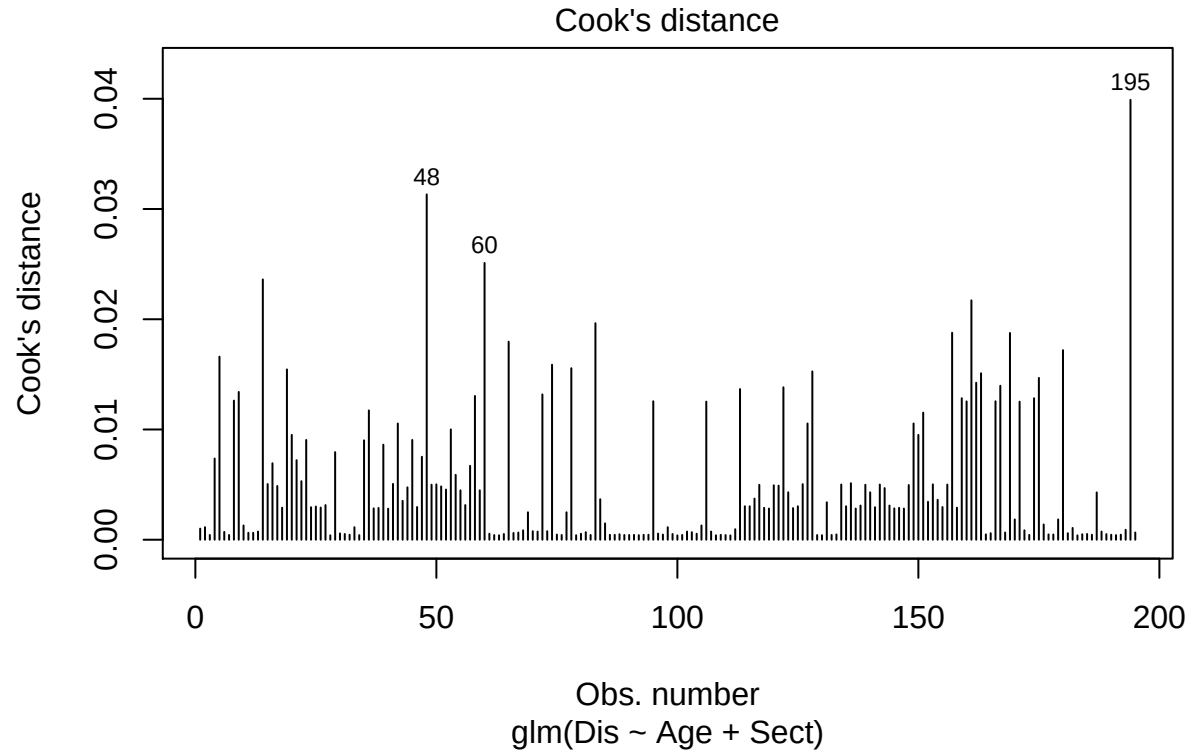
```
##
## Call:
## glm(formula = Dis ~ Age + Sect, family = binomial, data = data_disease_glm_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5863  -0.8156  -0.5457   0.9777   2.0533
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.247522   0.353085  -6.365 1.95e-10 ***
## Age          0.029897   0.008919   3.352 0.000802 ***
## SectSector2  1.228078   0.339950   3.613 0.000303 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 235.64  on 194  degrees of freedom
## Residual deviance: 208.65  on 192  degrees of freedom
## AIC: 214.65
##
```

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

```
# check for unusual points
plot(rstandard(two_predictor.glm_2))
```




```
# Check for influential points
library(car)
plot(two_predictor.glm_2, which=4)
```



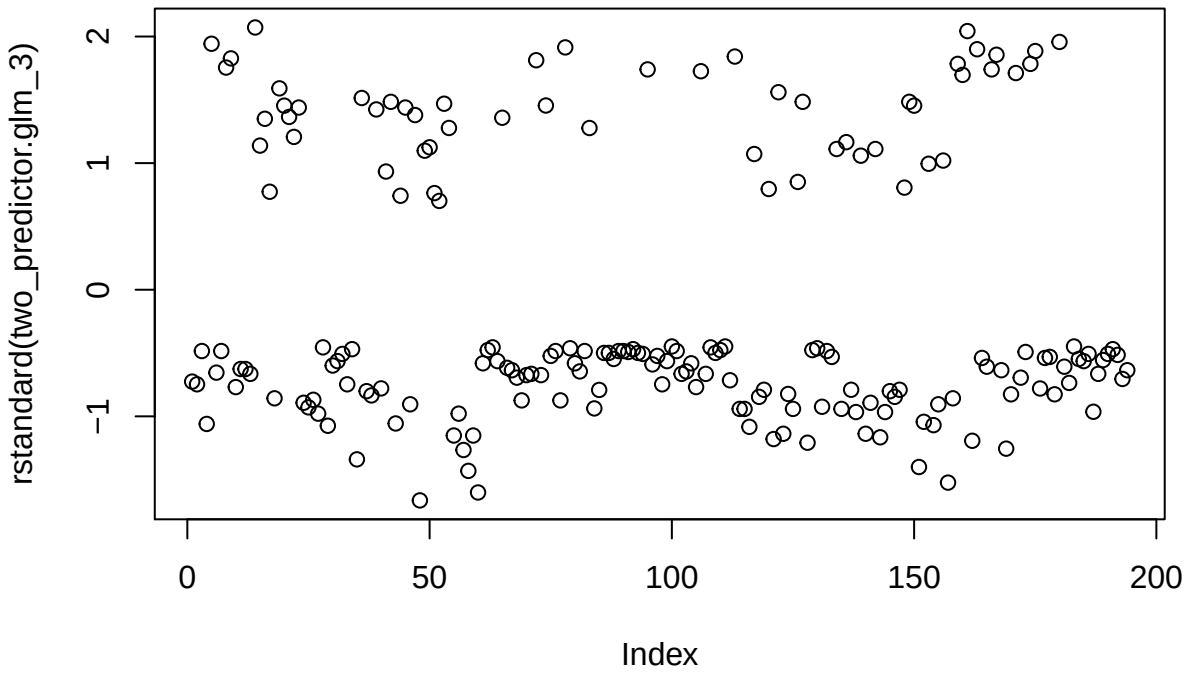
In general the fit of the model improve a little bit since the residual deviance decrease almost 3 points,

from 211.64 to 208.65, and the AIC decrease also almost three points. Now, let´s drop the another highly influential point and check what happens with our analysis.

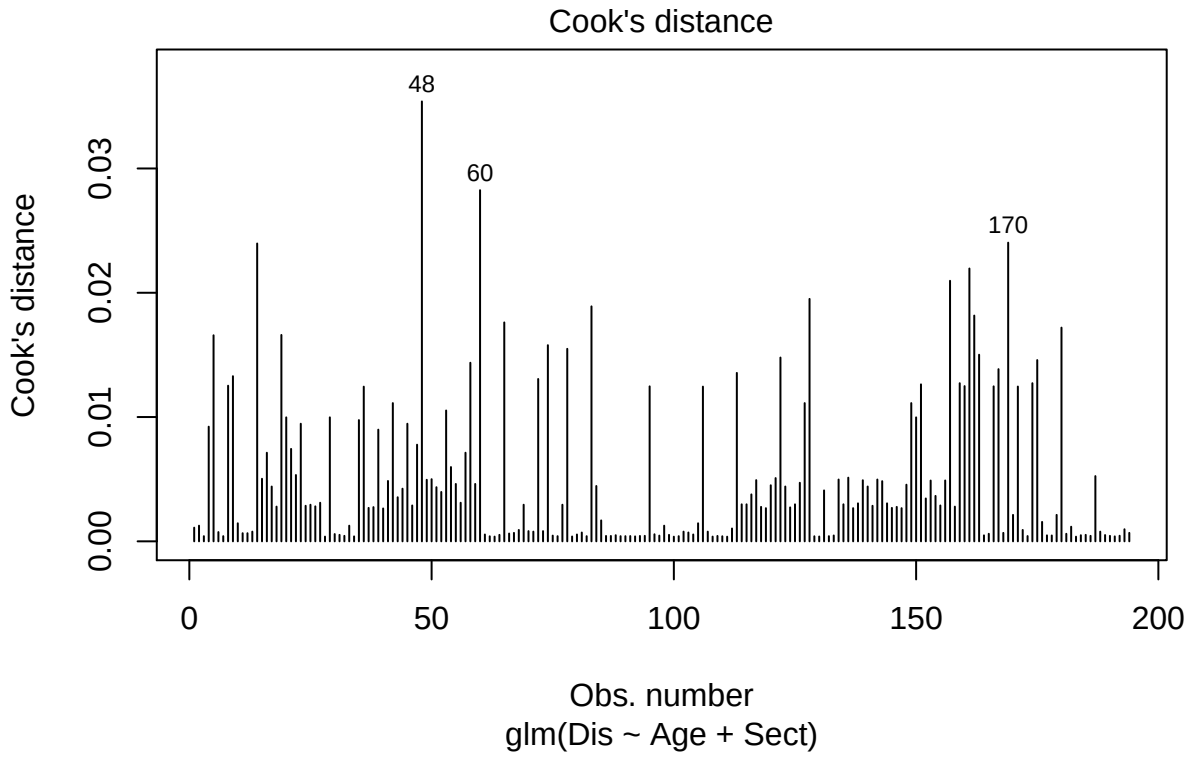
```
data_disease_glm_3 = data_disease_glm_2[-194,]
two_predictor.glm_3=glm(Dis~Age+Sect, family = binomial, data_disease_glm_3)
summary(two_predictor.glm_3)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect, family = binomial, data = data_disease_glm_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6342  -0.8095  -0.5445   0.9681   2.0613
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.291504   0.357245  -6.414 1.41e-10 ***
## Age          0.032690   0.009287   3.520 0.000432 ***
## SectSector2  1.196696   0.341438   3.505 0.000457 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 206.84  on 191  degrees of freedom
## AIC: 212.84
##
## Number of Fisher Scoring iterations: 4
```

```
# check for unusual points
plot(rstandard(two_predictor.glm_3))
```



```r
# Check for influential points
library(car)
plot(two_predictor.glm_3, which=4)
```



Again, deleting this point the fit of the model improve a little bit since the residual deviance decrease almost 2 points, from 208.65 to 206.84, and the AIC also decreases also almost two points.

Since the these two points are highly influential and only represent 1% of our data, we are going to delete it and return to the step 2 to repeat the analysis with the new data set.

Step 2.A: Run again the Backward elimination using AIC criteria with the new dataset.

```r
glm_disease_null_new=glm(Dis~1, family = binomial, data=data_disease_glm_3)
glm_disease_full_new=glm(Dis~ Age  + SES + Sect + Sav, family = binomial, data=data_disease_glm_3)
# AIC based backward elimination
step(glm_disease_full_new, direction="backward", trace = 1)
```

```
## Start:  AIC=218.32
## Dis ~ Age + SES + Sect + Sav
##
##        Df Deviance    AIC
## - SES   2   206.74 214.74
## - Sav   1   206.33 216.33
## <none>      206.32 218.32
## - Age   1   218.83 228.83
## - Sect  1   219.24 229.24
##
## Step:  AIC=214.74
## Dis ~ Age + Sect + Sav
##
##        Df Deviance    AIC
```

```
## - Sav   1   206.84 212.84
## <none>      206.74 214.74
## - Sect  1   219.25 225.25
## - Age   1   219.48 225.48
##
## Step:  AIC=212.84
## Dis ~ Age + Sect
##
##        Df Deviance    AIC
## <none>      206.84 212.84
## - Sect  1   219.51 223.51
## - Age   1   220.00 224.00
```

```
##
## Call:  glm(formula = Dis ~ Age + Sect, family = binomial, data = data_disease_glm_3)
##
## Coefficients:
## (Intercept)          Age  SectSector2
##    -2.29150      0.03269      1.19670
##
## Degrees of Freedom: 193 Total (i.e. Null);  191 Residual
## Null Deviance:       234.9
## Residual Deviance: 206.8     AIC: 212.8
```

The best model according to these criteria is a model that includes Age and Sect as explanatory variables.

Step 3.A: Run again best subsets using the BIC criteria to get an idea of the best models. Then compare the best models with those obtained using the AIC criteria.

```
best_disease_bic_new= bestglm(data_disease_glm_3, family=binomial,  IC="BIC", TopModels = 5)
```

```
## Morgan-Tatar search since family is non-gaussian.
## Note: factors present with more than 2 levels.
```

```
best_disease_bic_new$Subsets
```

```
##    Intercept   Age   SES  Sect   Sav logLikelihood      BIC
## 0       TRUE FALSE FALSE FALSE FALSE     -117.4732 234.9463
## 1       TRUE  TRUE FALSE FALSE FALSE     -109.7575 224.7828
## 2*      TRUE  TRUE FALSE  TRUE FALSE     -103.4196 217.3750
## 3       TRUE  TRUE FALSE  TRUE  TRUE     -103.3691 222.5418
## 4       TRUE  TRUE  TRUE  TRUE  TRUE     -103.1625 232.6642
```

```
#best_disease_bic$BestModels
```

According to these criteria the best models are:

- best one predictor model includes only Age as explanatory variable
- best two predictor model includes Age and Sect as explanatory variables
- best three predictor model includes Age, Sect and Sav as explanatory variables
- Best fourth predictor model includes Age, Sect, Sav, and SES as explanatory variables. For SES we will see which of its categories provide the better fit.

Step 4.A: Run again the set of best candidate models and get their summaries in R.

Here, I will use the best 2, 3 and 4 predictors model (the best model according to both procedures use in step 2 and 3 is a model with Age and Sect as explanatory variables). I discard the one predictor model because, since I do believe that a phenomenon of such complexity as a the probability to contract a disease during an

outbreak should be determine by more tha one variable. In general, the best models are the same compared with the case with all the data was used.

```
# Best two predictor model
two_predictor_new.glm=glm(Dis~Age+Sect, family = binomial, data=data_disease_glm_3)
summary(two_predictor_new.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect, family = binomial, data = data_disease_glm_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6342  -0.8095  -0.5445   0.9681   2.0613
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.291504   0.357245  -6.414 1.41e-10 ***
## Age          0.032690   0.009287   3.520 0.000432 ***
## SectSector2  1.196696   0.341438   3.505 0.000457 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 206.84  on 191  degrees of freedom
## AIC: 212.84
##
## Number of Fisher Scoring iterations: 4
```

```
# Best three predictor model
three_predictor_new.glm=glm(Dis~Age+Sect+Sav, family = binomial, data=data_disease_glm_3)
summary(three_predictor_new.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + Sav, family = binomial, data = data_disease_glm_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6335  -0.8021  -0.5554   0.9401   2.0955
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.261861   0.368380  -6.140 8.25e-10 ***
## Age          0.033616   0.009764   3.443 0.000576 ***
## SectSector2  1.223358   0.352234   3.473 0.000514 ***
## SavYES      -0.118164   0.372164  -0.318 0.750859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 206.74  on 190  degrees of freedom
```

```
## AIC: 214.74
##
## Number of Fisher Scoring iterations: 4
```

```
# Best fourth predictor model (it is not really a 4 predictor model, but a 5 predictor model since SES
fourth_predictor_new.glm=glm(Dis~Age+Sect+Sav+SES, family = binomial, data=data_disease_glm_3)
summary(fourth_predictor_new.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + Sav + SES, family = binomial,
##     data = data_disease_glm_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7271  -0.8188  -0.5484   0.9635   2.1255
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.408783   0.491876  -4.897 9.72e-07 ***
## Age          0.033216   0.009744   3.409 0.000652 ***
## SectSector2  1.269152   0.361299   3.513 0.000444 ***
## SavYES      -0.038718   0.400552  -0.097 0.922996
## SESMiddle    0.006695   0.446866   0.015 0.988046
## SESLower     0.255860   0.440597   0.581 0.561435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 206.32  on 188  degrees of freedom
## AIC: 218.32
##
## Number of Fisher Scoring iterations: 4
```

Step 5: compare again the results of the different models and decide which one is the best for prediction.

Here, We are going to compare the results of the models based on the pseudo-R square, Residual Deviance and the Akaike Criterion:

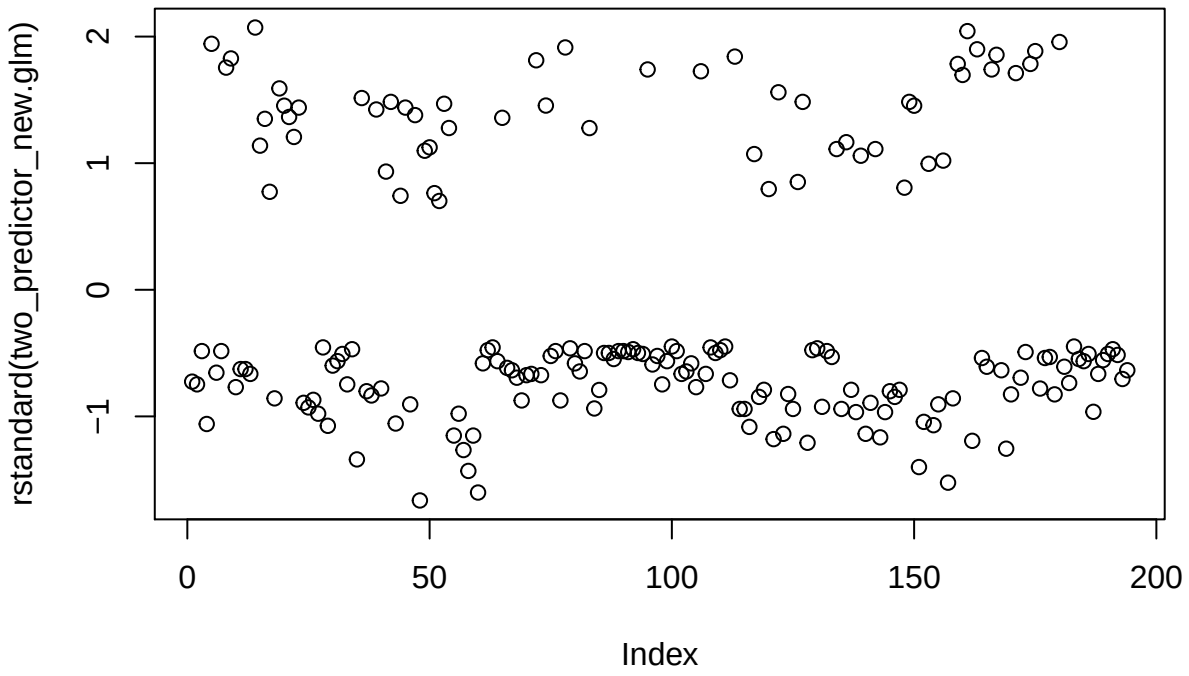
Model	Residual Deviance	pseudo-r2	AIC
Two predictors	206.84	0.1197	212.84
Three predictors	206.74	0.1200	214.74
Five predictors	206.32	0.1219	218.32

Two things are notorious:

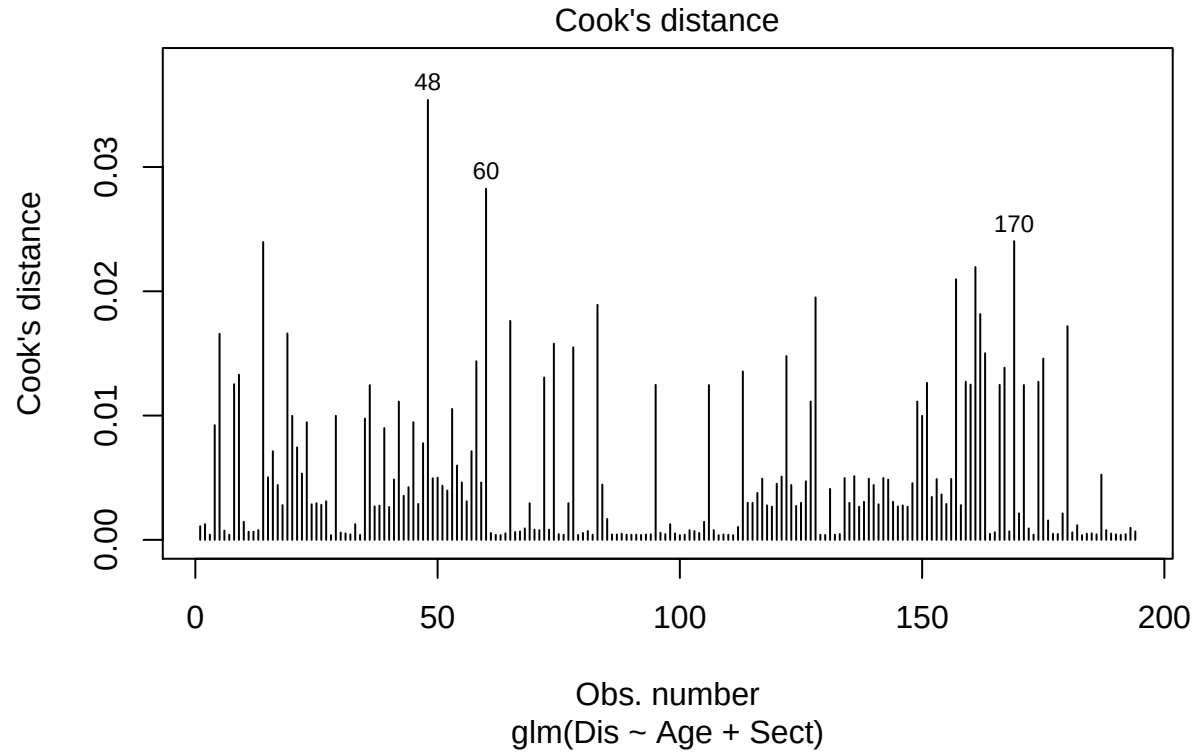
- Compared to the results obtained using all the data, using the new data, We have improved the fit of all the models. For instance, the residual deviance have decreased almost 5 points, and the pseudo R-Squared have increased almost 1.5 percentage points.
- Again, the results shows that even after drop the highly influential data points, the improvements in fit of using a model that include Sav and SES as explanatory variables are minimun (in terms of residual deviance and pseudo R squared). Then, again based on these results, I consider that the best model is the two predictor model, i.e, the one with only Age and Sect as explanatory variables.

Step 6: Check for suspect and influential data points

```r
# check for outliers
plot(rstandard(two_predictor_new.glm))
```




```r
# Check for influential points
library(car)
plot(two_predictor_new.glm, which=4)
```



The graphs above tell us that we do not have outliers, and the points with greater cook's distance are not too far away from the majority of the rest.

Step 7. Check if polynomial terms are neccesary.

```
two_predictor_new_pol.glm=glm(Dis~Age+Sect+I(Age^2), family = binomial, data=data_disease_glm_3)
summary(two_predictor_new_pol.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + I(Age^2), family = binomial,
##     data = data_disease_glm_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4929  -0.8267  -0.5209   0.9445   2.2000
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.1706423  0.5869625  -5.402  6.6e-08 ***
## Age          0.1028542  0.0353255   2.912 0.003596 **
## SectSector2  1.2748070  0.3491419   3.651 0.000261 ***
## I(Age^2)    -0.0010121  0.0004823  -2.098 0.035864 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 202.22  on 190  degrees of freedom
## AIC: 210.22
##
## Number of Fisher Scoring iterations: 4
```

Since the polynomial term for age is significant and improve the fit (reduce the AIC by almost 2.5 points from 212.84 to 210.22), we will keep it in our model. Let's try the cubes as well:

```
two_predictor_new_cube.glm=glm(Dis~Age+Sect+I(Age^2)+I(Age^3), family = binomial, data=data_disease_glm
summary(two_predictor_new_cube.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + I(Age^2) + I(Age^3), family = binomial,
##     data = data_disease_glm_3)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.411  -0.815  -0.466   1.013   2.270
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.640e+00  1.023e+00  -4.536 5.74e-06 ***
## Age          3.036e-01  1.072e-01   2.832 0.004629 **
## SectSector2  1.325e+00  3.542e-01   3.741 0.000183 ***
## I(Age^2)    -7.842e-03  3.357e-03  -2.336 0.019478 *
## I(Age^3)     6.281e-05  3.016e-05   2.082 0.037301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 197.44  on 189  degrees of freedom
## AIC: 207.44
##
## Number of Fisher Scoring iterations: 5
```

The cube term is also significant and and improve the fit as well. The residual deviance decreases by almost 5 points and the AIC by almost 3 points. Then, we are also going to keep this variable in our model. Now, let's try the quartic of age (Age^4) as well:

```
two_predictor_new_quartic.glm=glm(Dis~Age+Sect+I(Age^2)+I(Age^3)+I(Age^4), family = binomial, data=data_
summary(two_predictor_new_quartic.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + I(Age^2) + I(Age^3) + I(Age^4),
##     family = binomial, data = data_disease_glm_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5165  -0.7978  -0.4717   1.0496   2.2557
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.761e+00  1.614e+00  -3.569 0.000358 ***
## Age          5.225e-01  2.451e-01   2.132 0.032992 *
## SectSector2  1.285e+00  3.547e-01   3.622 0.000293 ***
## I(Age^2)    -2.002e-02  1.228e-02  -1.631 0.102929
## I(Age^3)     3.141e-04  2.411e-04   1.303 0.192624
## I(Age^4)    -1.702e-06  1.602e-06  -1.062 0.288124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 196.28  on 188  degrees of freedom
## AIC: 208.28
##
## Number of Fisher Scoring iterations: 6
```

The quartic is not significant and does not improve the fit, then we drop the quartic of age and continue with the model with only the square and cubic of age as additional explanatory variables.

Step 8: Check all possible first order interactions.

```
two_predictor_new_interact.glm=glm(Dis~Age + Sect + I(Age^2) + I(Age^3) + Age:Sect, family = binomial, d
summary(two_predictor_new_interact.glm)
```

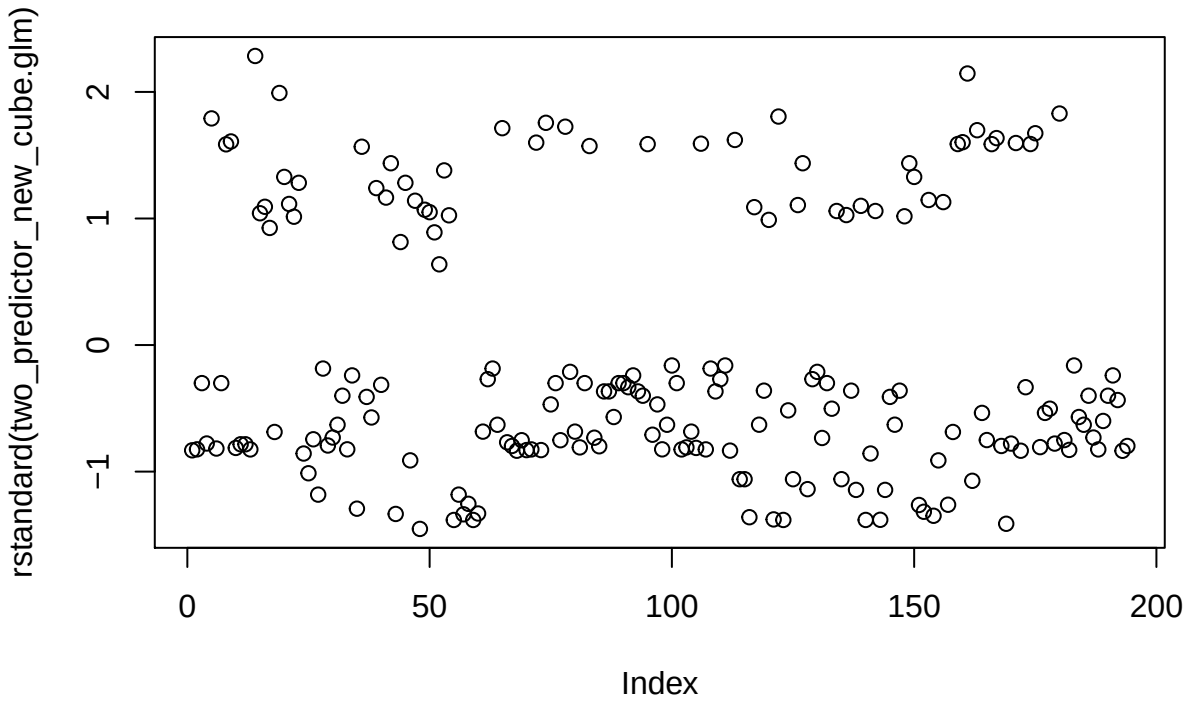
```
##
## Call:
## glm(formula = Dis ~ Age + Sect + I(Age^2) + I(Age^3) + Age:Sect,
##     family = binomial, data = data_disease_glm_3)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.5115  -0.8063  -0.4867   1.0064   2.2017
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.423e+00  1.052e+00  -4.205 2.61e-05 ***
## Age              2.992e-01  1.077e-01   2.779  0.00545 **
## SectSector2      9.128e-01  6.989e-01   1.306  0.19150
## I(Age^2)        -7.994e-03  3.408e-03  -2.346  0.01899 *
## I(Age^3)         6.436e-05  3.079e-05   2.090  0.03658 *
## Age:SectSector2  1.414e-02  2.095e-02   0.675  0.49971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
## Residual deviance: 196.98  on 188  degrees of freedom
## AIC: 208.98
##
## Number of Fisher Scoring iterations: 5
```

Since the first order interaction required the estimation of only one additional parameter we can decide wether or not this intercation is significan base on partial z test. In this case, the interaction is not significant and does not contribute to the improvement of the fit of the model. Therefore, our final model does not include first order interactions.

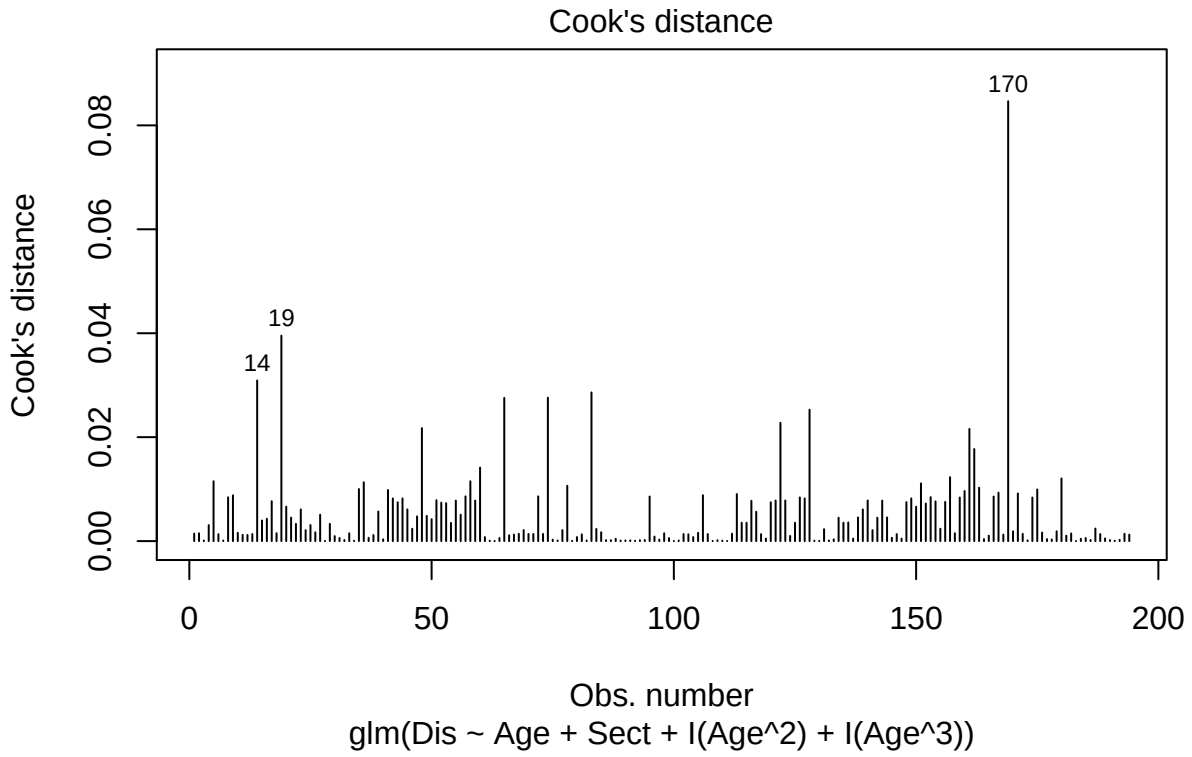
Step 9: Check final model using standarized residual plot and cook's distance plot.

```
# check for outliers
plot(rstandard(two_predictor_new_cube.glm))
```




```
# Check for influential points
library(car)
```

```r
plot(two_predictor_new_cube.glm, which=4)
```



The graphs above tell us that we do not have outliers, but the observation 170 is highly influential. Now, we are going to drop this observation and graph again the diagnostic plots. We are also going to compare how the coefficients really change deleting this observation, to decide if it is worth it to return to the step 2 again.

```r
summary(two_predictor_new_cube.glm)
```

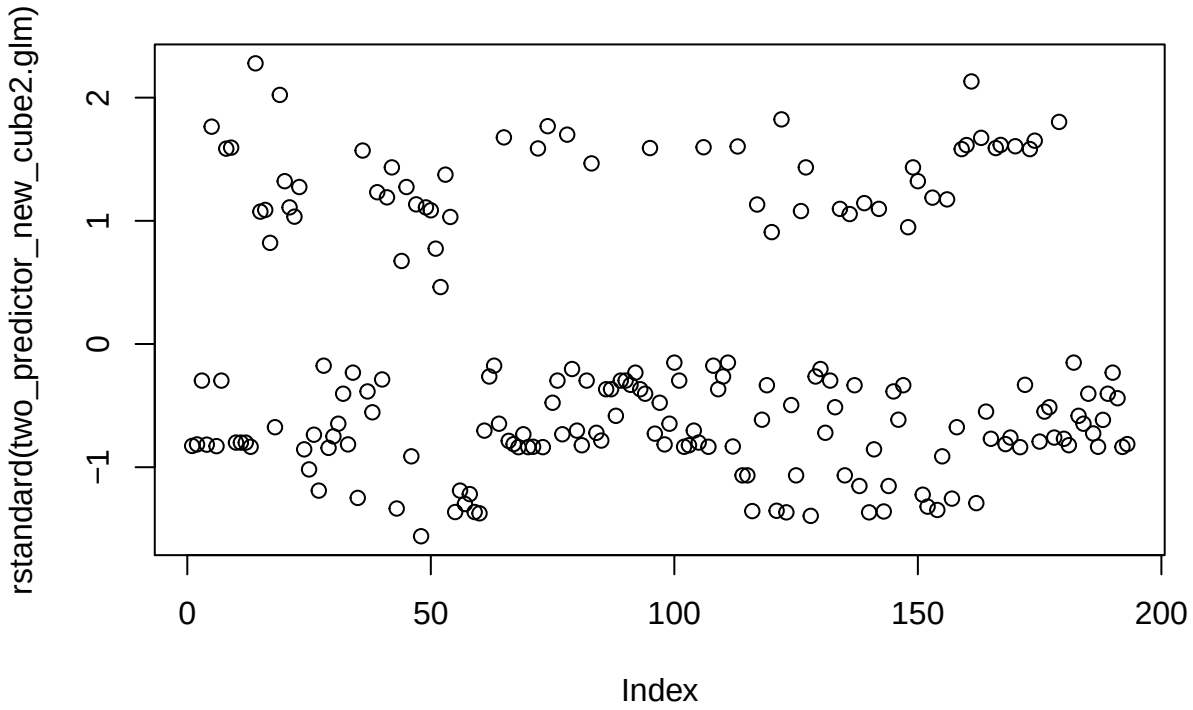
```
##
## Call:
## glm(formula = Dis ~ Age + Sect + I(Age^2) + I(Age^3), family = binomial,
##     data = data_disease_glm_3)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.411  -0.815  -0.466   1.013   2.270
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.640e+00  1.023e+00  -4.536 5.74e-06 ***
## Age          3.036e-01  1.072e-01   2.832 0.004629 **
## SectSector2  1.325e+00  3.542e-01   3.741 0.000183 ***
## I(Age^2)    -7.842e-03  3.357e-03  -2.336 0.019478 *
## I(Age^3)     6.281e-05  3.016e-05   2.082 0.037301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.95  on 193  degrees of freedom
```

```
## Residual deviance: 197.44  on 189  degrees of freedom
## AIC: 207.44
##
## Number of Fisher Scoring iterations: 5
```

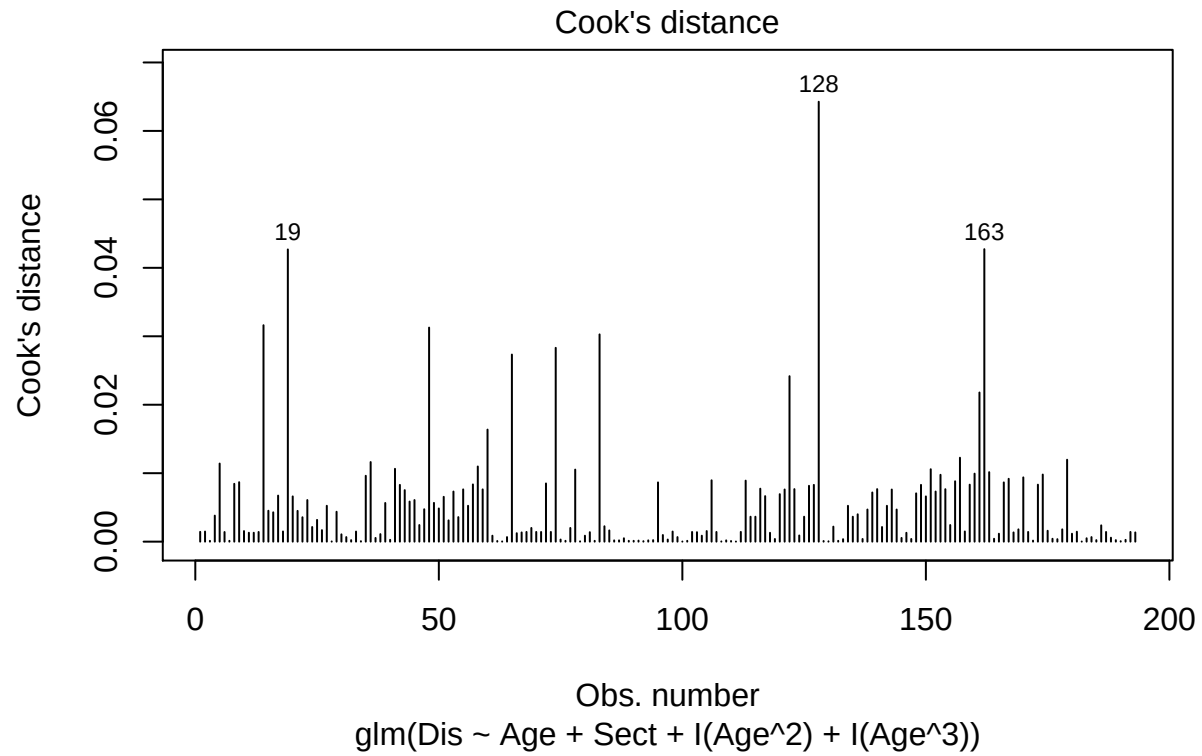
```
data_disease_glm_4 = data_disease_glm_3[-169,]
two_predictor_new_cube2.glm=glm(Dis~Age+Sect+I(Age^2)+I(Age^3), family = binomial, data=data_disease_gl
summary(two_predictor_new_cube2.glm)
```

```
##
## Call:
## glm(formula = Dis ~ Age + Sect + I(Age^2) + I(Age^3), family = binomial,
##     data = data_disease_glm_4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5092  -0.8141  -0.4738   1.0410   2.2636
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.804e+00  1.059e+00  -4.534 5.79e-06 ***
## Age          3.329e-01  1.129e-01   2.950 0.003174 **
## SectSector2  1.283e+00  3.558e-01   3.608 0.000309 ***
## I(Age^2)    -9.006e-03  3.586e-03  -2.511 0.012024 *
## I(Age^3)     7.531e-05  3.281e-05   2.295 0.021721 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.25  on 192  degrees of freedom
## Residual deviance: 195.55  on 188  degrees of freedom
## AIC: 205.55
##
## Number of Fisher Scoring iterations: 5
```

```
# check for unusual points
plot(rstandard(two_predictor_new_cube2.glm))
```




```
# Check for influential points
library(car)
plot(two_predictor_new_cube2.glm, which=4)
```



The table below shows the comparison of the coefficients taking into account observation 169 and also deleting it.

Coefficient	including observation 169	Excluding observation 169
Intercept	−4.64	−4.8040
Age	0.3036	0.3329
Sector	1.3250	1.2830
Age2	−0.0078	−0.0090
Age3	0.0001	0.0001
Residual deviance	197.44	195.55
AIC	207.44	205.55

As we can see the changes that experiment the coefficients are not enough large to consider deleting this observation and run the analysis again. Futhermore, since the purpose of the model is prediction, I consider that the gains in residual deviance and AIC are not enough to follow that strategy. Therefore, we have arrived to our final model.

Part B) Give a 90% confidence interval for the probability that a 64 year old patient, with middle socioeconomic status and a savings account that lives in sector 2 of the city, contracts the disease.

```
model_part_b = glm(Dis~ Age  + SES + Sect + Sav, family = binomial, data=data_disease_glm)
summary(model_part_b)
```

```
##
## Call:
## glm(formula = Dis ~ Age + SES + Sect + Sav, family = binomial,
##     data = data_disease_glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6614  -0.8309  -0.5630   1.0134   2.0918
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.273558   0.479333  -4.743  2.1e-06 ***
## Age          0.027280   0.009132   2.987 0.002813 **
## SESMiddle    0.035578   0.441452   0.081 0.935765
## SESLower     0.237633   0.433750   0.548 0.583789
## SectSector2  1.249464   0.357009   3.500 0.000466 ***
## SavYES      -0.040692   0.396540  -0.103 0.918266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.33  on 195  degrees of freedom
## Residual deviance: 211.21  on 190  degrees of freedom
## AIC: 223.21
##
## Number of Fisher Scoring iterations: 4
```

```
newdata=data.frame(Age=64, SES="Middle", Sect="Sector2", Sav="YES")
prediction=predict(model_part_b, newdata, type="response",  se.fit = TRUE)
#90% C.I for the estimated probability
lower= prediction$fit - (-qnorm(0.05))*prediction$se.fit
upper= prediction$fit + (-qnorm(0.05))*prediction$se.fit
c(lower,upper)
```

```
##         1         1
```

```
## 0.4903909 0.8533793
```

We are 90% confident that the probability to contract a disease for this person is between 0.49 and 0.85.

Question 3.

Multiple cohorts of subjects, some non-smokers and others smokers, were observed for several years. The number of cases (NumCases) of lung cancer diagnosed in the different cohorts was recorded, in addition to the following predictor variables:

CigsperDay = Number of cigarettes smoked per day per individual in the cohort; Years = The number of years the individuals in the cohort had smoked.

Additionally, the total number of years in which individuals in each category were observed (summed over all individuals) was recorded in the column PersonYears. (For example, if a cohort had 50 people that had been observed for 20 years, that would be 50 x 20 = 1000 PersonYears.) Data appear in Hwk5Q3DatSp17.

```
library(readxl)
data_smoke = read_excel("Hwk5Q3DatSp17.xlsx")
head(data_smoke)
```

```
##   CigsperDay Years PersonYears NumCases
## 1          0    15       10366        1
## 2          0    25        5969        0
## 3          0    35        3512        0
## 4          0    45        1421        0
## 5          0    55         826        2
## 6          5    15        3121        0
```

A) Write down a Poisson regression model where the mean number of cases of observed lung cancer cases per cohort are a function of CigsperDay and Years. Your model should start like "? = .", NOT "log(?) = .".

The poisson regression model for the mean number of cases of observed lung cancer is:

$$\mu = E(Y|X) = exp(\beta_0 + \beta_1 * CigsperDay + \beta_2 * Years)$$

B) Fit the model above; include summary output. State your model of the estimated mean with the maximum likelihood estimators included.

```
smoke.glm=glm(NumCases~CigsperDay + Years, family=poisson, data=data_smoke)
summary(smoke.glm)
```

```
##
## Call:
## glm(formula = NumCases ~ CigsperDay + Years, family = poisson,
##     data = data_smoke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6501  -1.4275  -0.9166   0.4112   3.2095
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.726048   0.450482  -3.832 0.000127 ***
## CigsperDay   0.040434   0.008561   4.723 2.32e-06 ***
```

```
## Years        0.043769   0.008942   4.895 9.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 137.29  on 34  degrees of freedom
## Residual deviance:  88.13  on 32  degrees of freedom
## AIC: 151.72
##
## Number of Fisher Scoring iterations: 6
```

The model of the estimated mean is:

$$\mu = E(NumCases|CigsperDay, Years) = exp(-1.726048 + 0.040434 * CigsperDay + 0.043769 * Years)$$

C) Do a deviance goodness-of-fit test on your model; state hypotheses, test statistic, p-value, and conclusions.

The null hypothesis in this model is H_0 : the model fits the data; and the alternative hypothesis is H_A : the model does not fit the data. The test statistic is the value of the residual deviance in the model (88.13). This test is distributed χ^2 with n-k-1 degrees of freedom (in our case 32), where n is the number of observations, and k is the number of parameters (excluding the intercept). The p-value of this test is:

$$pvalue = P(\chi^2_{n-k-1} > TS)$$

In our case:

$$pvalue = P(\chi^2_{32} > 88.13)$$

For our model the test statistic and its pvalue is:

```
TS=88.13
n=35
k=2
pvalue=1-pchisq(88.13, df=n-k-1)
pvalue
```

```
## [1] 3.814793e-07
```

We reject the null hypothesis, therefore this model does not fit the data. This result imply that We need to study if a model that account for overdispersion fits the data better.

D) Does it make sense for your mean in Part A above to be proportional to the variable PersonYears? Explain briefly.

It makes sense since we would expect more number cases of lung cancer diagnosed if the individuals were observed for more time.Therefore, we need to include PersonYears as an offset in our estimation process.

E) Write down a Poisson regression model where the mean number of cases of observed lung cancer cases per cohort are a function of CigsperDay and Years, but are also proportional to PersonYears. Your model should start like "? = .", NOT "log(?) = ."

$$\mu_i * PersonYears_i = E(Y|X_i) * PersonYears_i = exp(\beta_0 + \beta_1 * CigsperDay_i + \beta_2 * Years_i + log_e PersonYears_i)$$

F) Fit the above model; include summary output. Perform a deviance goodness-of-fit test on this model; state hypotheses, test statistic, p-value, and conclusions.

```
smoke_offset.glm=glm(NumCases~CigsperDay + Years, family=poisson, data=data_smoke, offset = log(PersonY
summary(smoke_offset.glm)
```

```
##
## Call:
## glm(formula = NumCases ~ CigsperDay + Years, family = poisson,
##     data = data_smoke, offset = log(PersonYears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1657  -1.1254  -0.5335   0.5965   1.4920
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.557675   0.546276 -22.988  < 2e-16 ***
## CigsperDay    0.070795   0.009415   7.519 5.51e-14 ***
## Years         0.120894   0.010760  11.235  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 250.712  on 34  degrees of freedom
## Residual deviance:  43.347  on 32  degrees of freedom
## AIC: 106.94
##
## Number of Fisher Scoring iterations: 5
```

The null hypothesis for the goodness of fit test in this model is H_0 : the model fits the data; and the alternative hypothesis is H_A : model does not fit the data. The test statistic is the value of the residual deviance in the model (43.347). This test is distributed χ^2 with n-k-1 degrees of freedom, where n is the number of observations, and k is the number of parameters (excluding the intercept). The p-value of this test is:

$$pvalue = P(\chi^2_{n-k-1} > TS)$$

In our case:

$$pvalue = P(\chi^2_{32} > 43.347)$$

In R, we make this test as follows:

```
TS=43.347
n=35
k=2
pvalue=1-pchisq(43.347, df=n-k-1)
pvalue
```

```
## [1] 0.08693202
```

pvalue>0.05 (0.08693202>0.05), We do not reject the null hypothesis. This mean that after correcting for the offset effect the model fits the data.