Thesis 3 Model Analysis Code

Bailey Perry

March 22, 2018

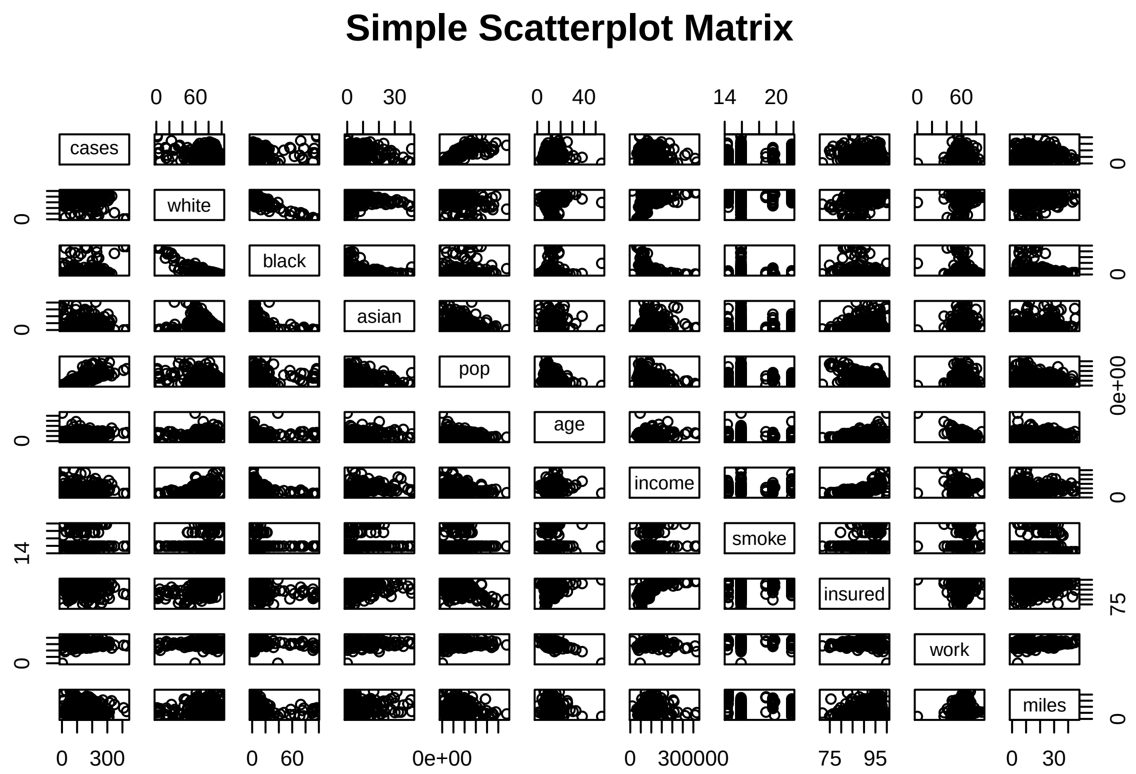
```
##Pre-processing:
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.2.5
```

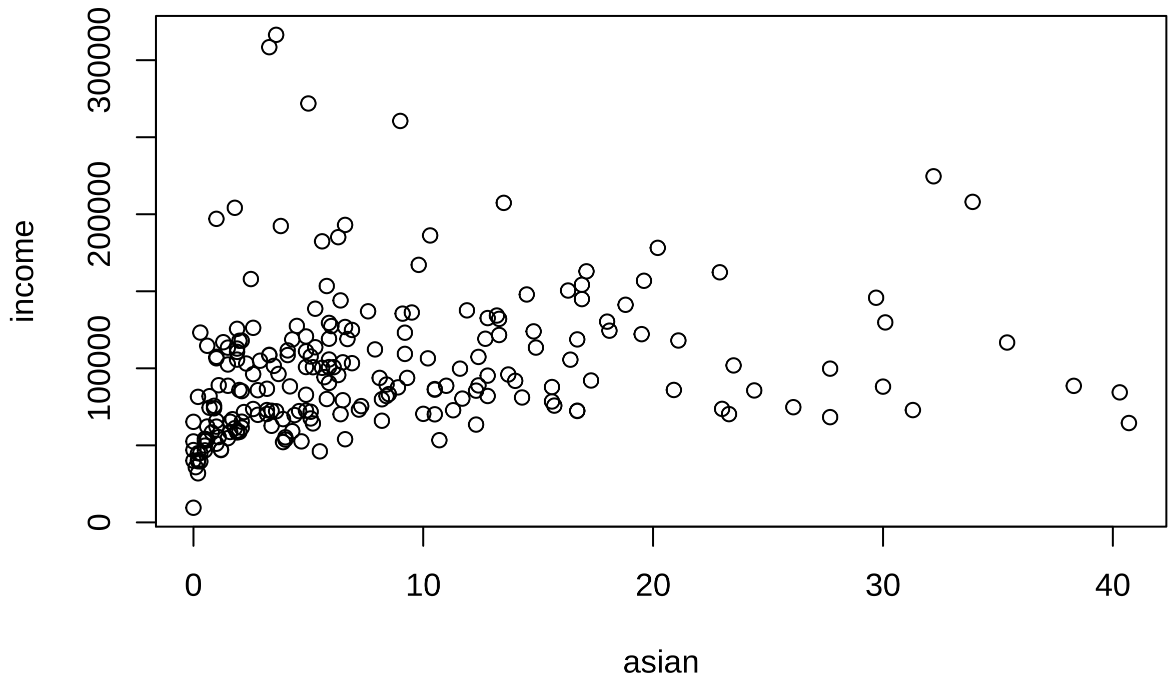
```
cancer <- read_excel("C:/Users/Bailey/Desktop/THESIS/ThesisPrep_Data_Draft2.xlsx")
#Remove the nulls - validated via the census information
cancer <- na.omit(cancer)
View(cancer)
dim(cancer)
```

```
## [1] 218  15
```

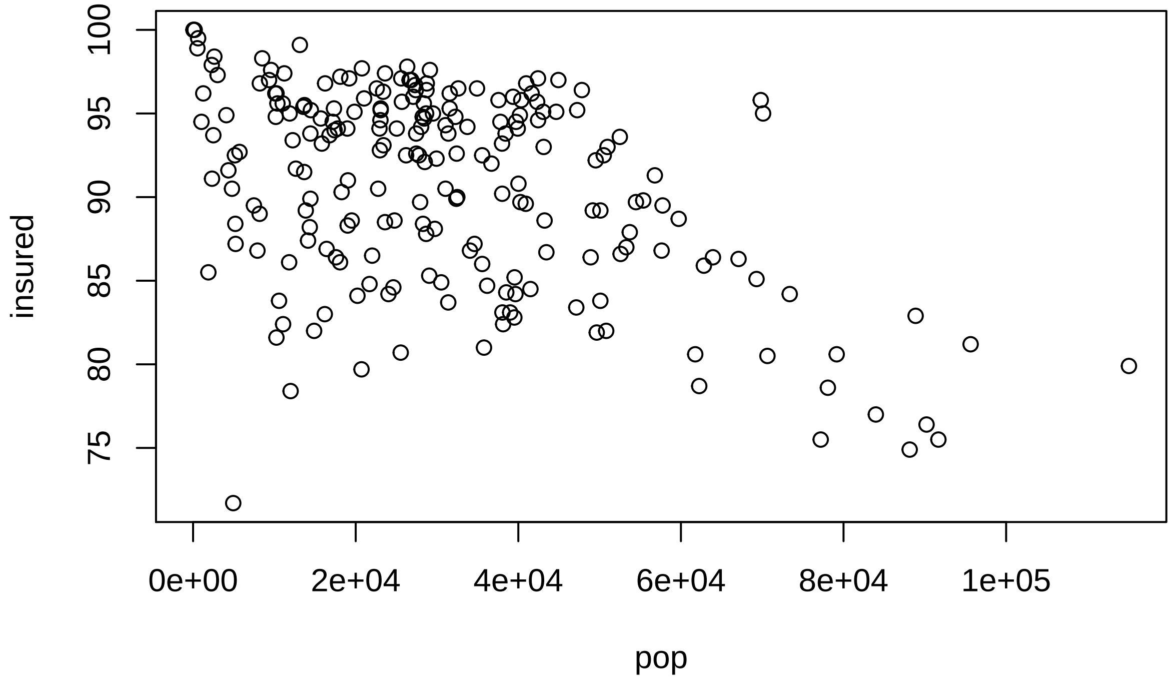
```
cases <- cancer$`2010-14 Incidence`
white <- cancer$`% White`
black <- cancer$`% Black`
asian <- cancer$`% Asian`
pop <- cancer$Population
age <- cancer$`% Over 65`
income <- cancer$`Average Income`
smoke <- cancer$`% Tobacco Use`
insured <- cancer$`% Population Insured`
work <- cancer$`% Females (16+) in Laborforce`
miles <- cancer$`Mileage to Nearest Hospital`

##Plot the Data - Exploratory Phase
pairs(~cases+white+black+asian+pop+age+income+smoke+insured+work+miles, main="Simple Scatterplot Matrix
```



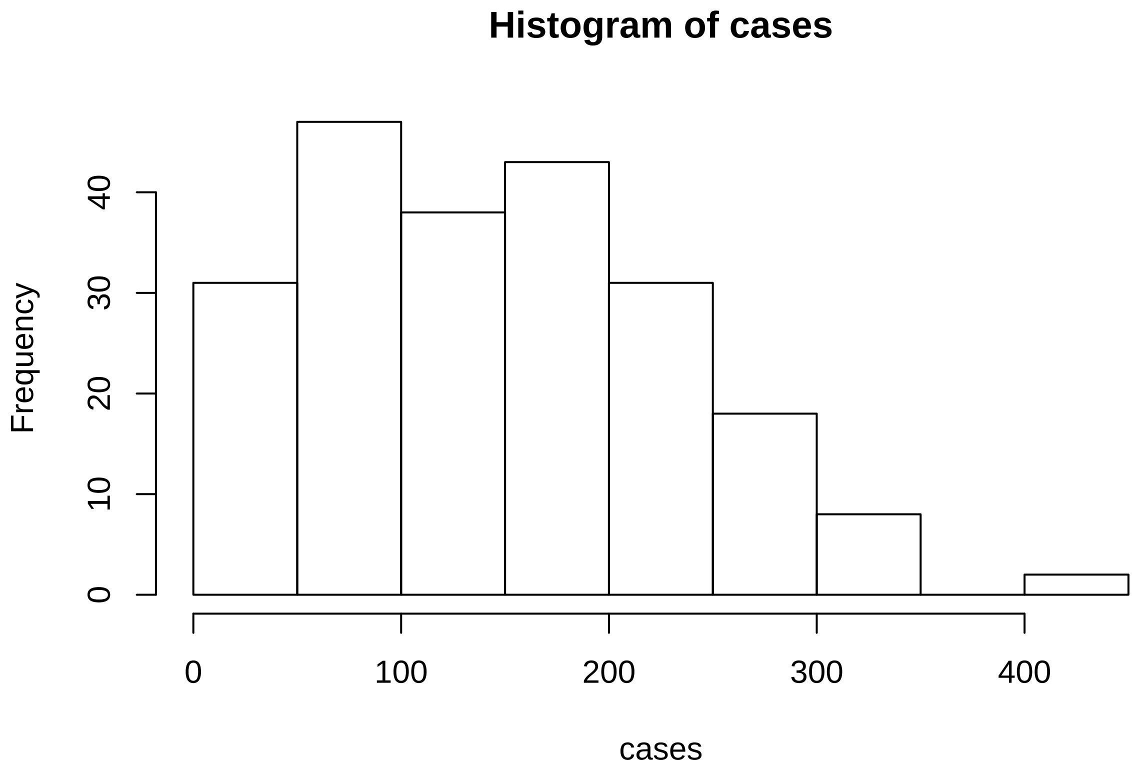
```
#Some zoomed in examples of potential issues in the data
plot(income~asian)
```



```r
plot(insured~pop)
```

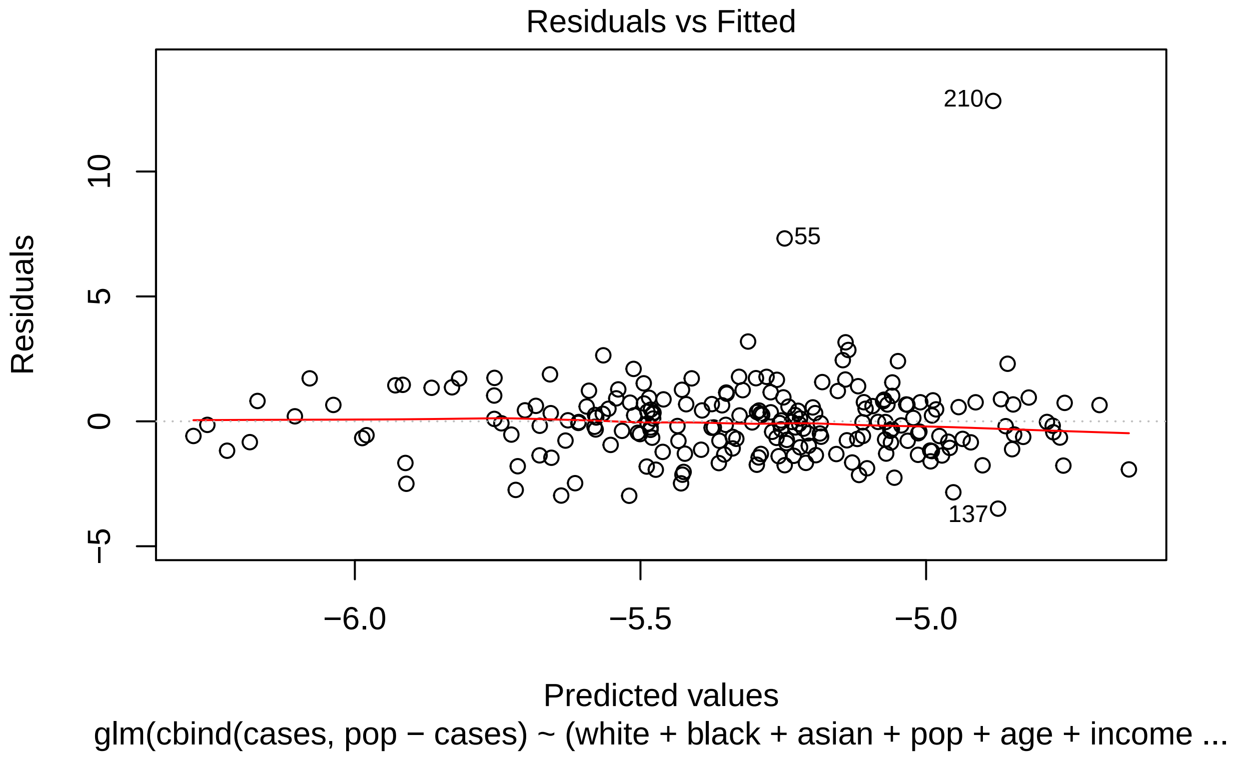


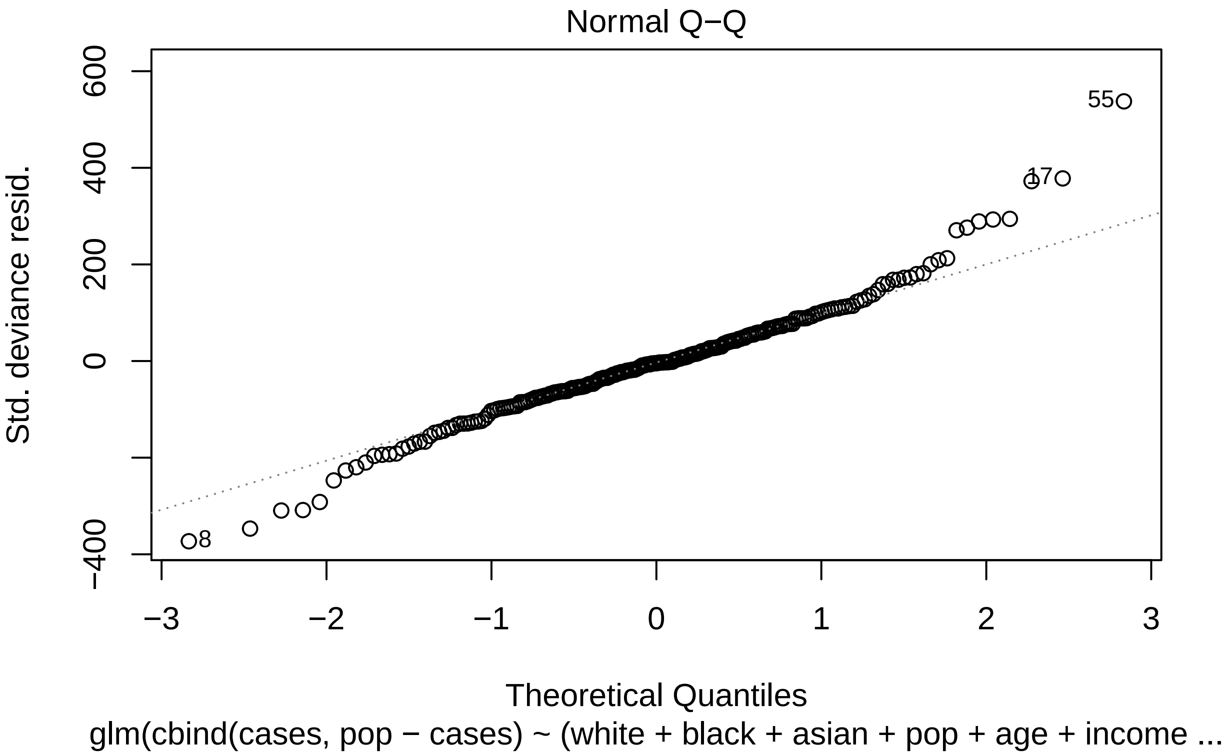
```
#checking to see if variance is fanning out

#Underlying distribution
hist(cases) #right-skewed
```

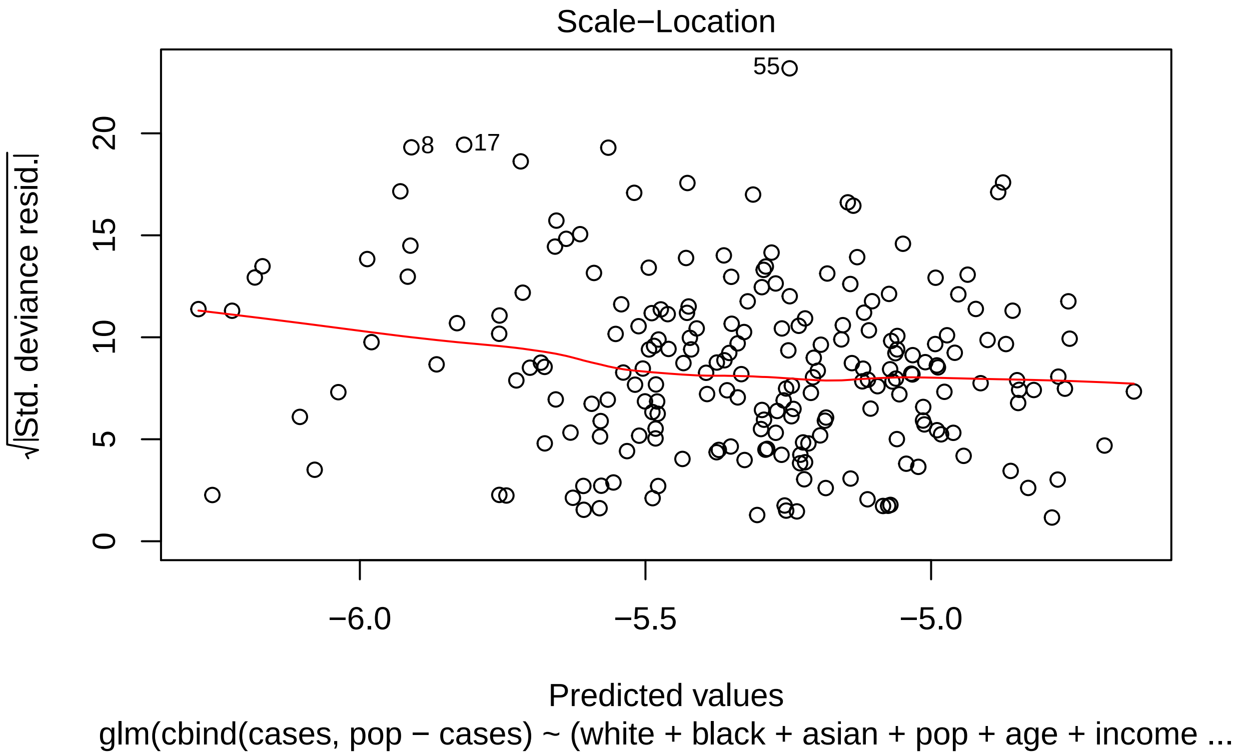


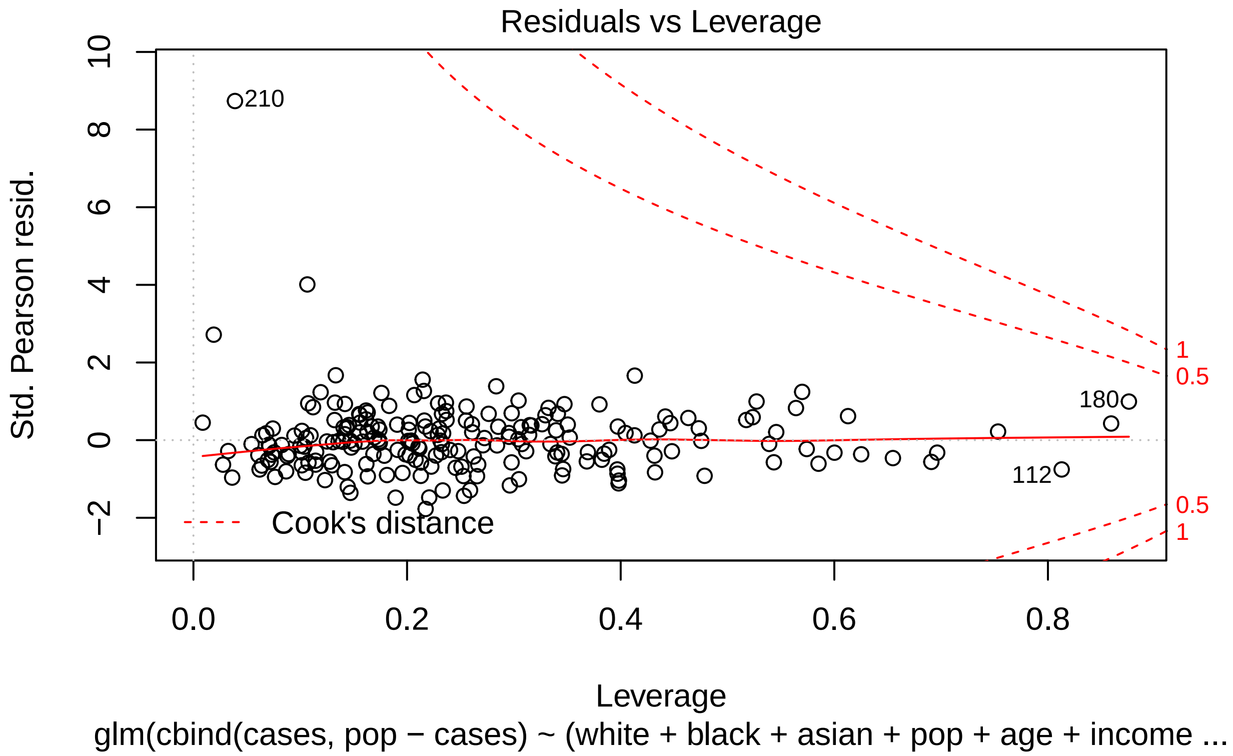
```
#THE FOLLOWING DISTRIBUTIONS AND MODELS WILL BE INVESTIGATED: quasibinomial,
# quasipoisson, and a linear model of rate:

#1. Fit Binomial in order to do AIC-based backwards selection
fitbin <- glm(cbind(cases, pop-cases) ~ (white+black+asian+pop+age+income+smoke+insured+work+miles)^2, f
## Changed backwards elimination of the models to k=4 to be more strict
full <- fitbin
null <- lm(cases ~ 1) #null is just the response with intercept
s1 <- step(full, scope=list(lower=null, upper=full), direction="backward", k=4)
summary(s1)
```

```
#Variables to include are as follows:
#white + black + asian +
#pop + age + income + smoke + insured + work + miles + white:asian +
#white:pop + white:age + white:income + white:insured + white:miles +
#black:pop + black:age + black:income + black:miles + asian:age +
#asian:income + asian:miles + pop:age + pop:miles + age:work +
#income:insured

#Quasibinomial with ALL terms - tester
fitqb <- glm(cbind(cases, pop-cases) ~ (white+black+asian+pop+age+income+smoke+insured+work+miles)^2, fa
plot(fitqb)
```


Residuals vs Fitted
Residuals
Predicted values
glm(cbind(cases, pop − cases) ~ (white + black + asian + pop + age + income ...
10
5
0
−5
−6.0
−5.5
−5.0
210
55
137


Normal Q−Q
Std. deviance resid.
600
400
200
0
−400
−3
−2
−1
0
1
2
3
Theoretical Quantiles
55
17
8
glm(cbind(cases, pop − cases) ~ (white + black + asian + pop + age + income ...

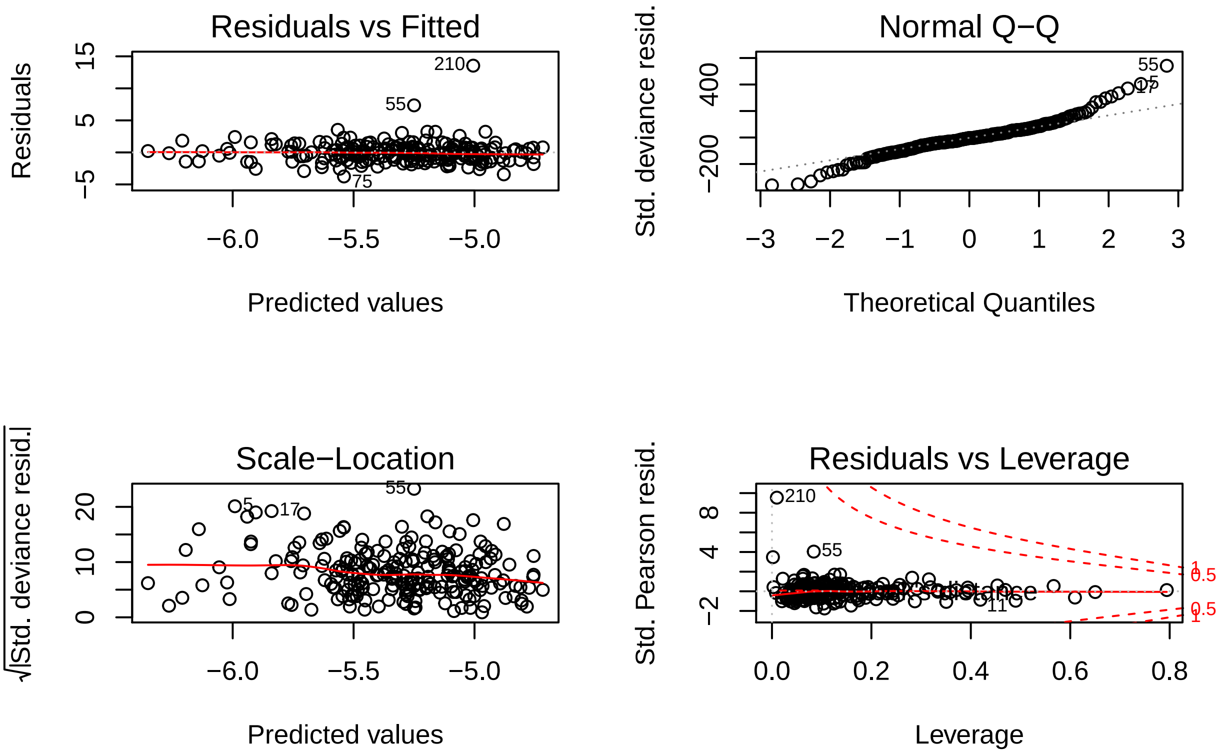

Scale−Location
√|Std. deviance resid.|
0
5
10
15
20
55
8
17
−6.0
−5.5
−5.0
Predicted values
glm(cbind(cases, pop − cases) ~ (white + black + asian + pop + age + income ...



```
# Now fit QUASIBINOMIAL model with identified significant cases from BE
##a. Model
fit1 <- glm(cbind(cases, pop-cases) ~ white+black+asian+pop+age+income+smoke+insured+work+miles+white:a
              black:pop + black:age + black:income + black:miles + asian:age + asian:income + asian:mile
              income:insured, family=quasibinomial(link="logit"))
#main effects and interaction terms as specified above
summary(fit1)
```

```
##
## Call:
## glm(formula = cbind(cases, pop - cases) ~ white + black + asian +
##     pop + age + income + smoke + insured + work + miles + white:asian +
##     white:pop + white:age + white:income + white:insured + white:miles +
##     black:pop + black:age + black:income + black:miles + asian:age +
##     asian:income + asian:miles + pop:age + pop:miles + age:work +
##     income:insured, family = quasibinomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7549  -0.9034  -0.1128   0.7176  13.5398
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -9.734e+00  1.683e+00  -5.782 2.98e-08 ***
## white           3.558e-03  2.223e-02   0.160  0.87301
## black           2.827e-02  1.229e-02   2.301  0.02247 *
## asian           1.441e-02  1.156e-02   1.246  0.21414
## pop             5.562e-06  5.985e-06   0.929  0.35385
## age             1.138e-01  7.171e-02   1.587  0.11426
## income          5.185e-05  1.531e-05   3.387  0.00086 ***
## smoke           8.490e-03  5.526e-03   1.536  0.12612
## insured         1.931e-02  2.200e-02   0.878  0.38116
## work           -1.184e-02  5.224e-03  -2.266  0.02457 *
## miles          -3.979e-02  1.824e-02  -2.182  0.03035 *
## white:asian     2.023e-04  1.175e-04   1.722  0.08673 .
## white:pop      -1.656e-07  7.629e-08  -2.171  0.03117 *
## white:age      -1.433e-03  6.858e-04  -2.090  0.03795 *
## white:income   -1.514e-07  1.395e-07  -1.085  0.27912
## white:insured   2.639e-04  2.575e-04   1.025  0.30660
## white:miles     3.386e-04  1.783e-04   1.898  0.05917 .
## black:pop      -1.022e-07  7.449e-08  -1.371  0.17186
## black:age      -1.410e-03  7.002e-04  -2.013  0.04552 *
## black:income   -1.666e-07  1.340e-07  -1.244  0.21519
## black:miles     4.173e-04  2.113e-04   1.974  0.04978 *
## asian:age      -1.574e-03  7.457e-04  -2.111  0.03606 *
## asian:income   -2.049e-07  1.447e-07  -1.416  0.15849
## asian:miles     5.957e-04  2.208e-04   2.698  0.00760 **
## pop:age         2.666e-07  1.867e-07   1.428  0.15503
## pop:miles       1.341e-07  1.026e-07   1.307  0.19294
## age:work        9.043e-04  3.176e-04   2.848  0.00489 **
## income:insured -3.604e-07  1.837e-07  -1.962  0.05128 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 4.427944)
##
##     Null deviance: 3829.77  on 217  degrees of freedom
## Residual deviance:  572.28  on 190  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
##b. Diagnostics
par(mfrow = c(2,2))
plot(fit1)
```

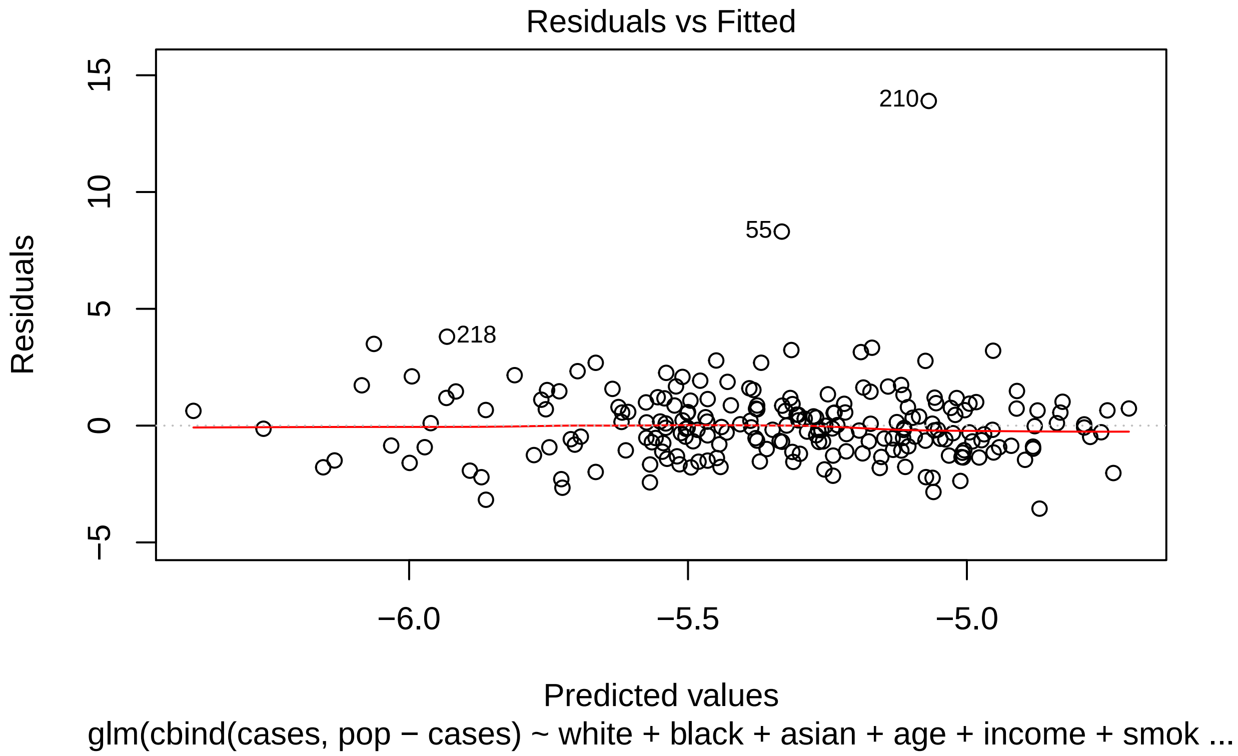


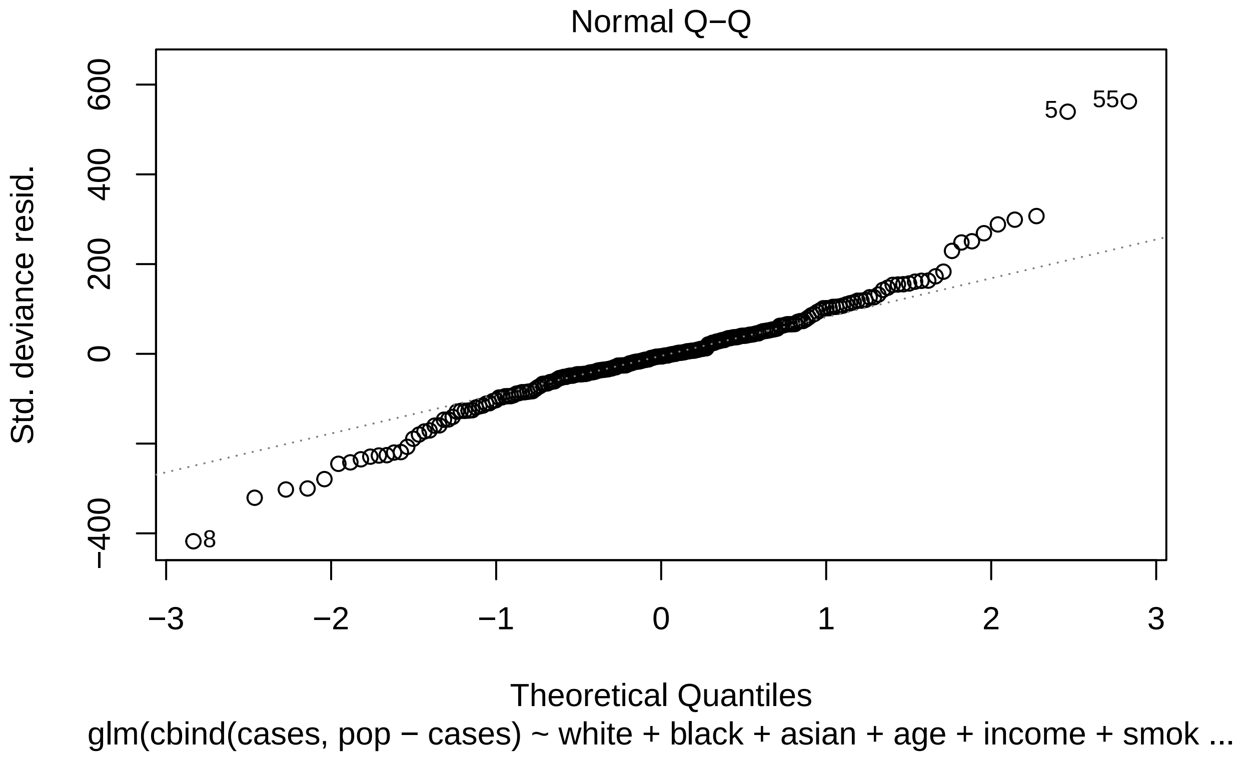
```
par(mfrow = c(1,1))

## Reduces the amount of variables included which is good for VIF, and the plots
# look just about as good as with all the main effects and interactions
# The tails pull off a tiny bit more, BUT doesn't have a significant shape

## We are missing parameters so a perfect model won't happen

#Fit QB withOUT pop as a predictor, even though BE says to include
#Exploratory...
fitqbnpop <- glm(cbind(cases, pop-cases) ~ white+black+asian+age+income+smoke+insured+work+miles+white:a
              + black:age + black:income + black:miles + asian:age + asian:income + asian:miles + age:w
              income:insured, family=quasibinomial(link="logit"))
summary(fitqbnpop)
```

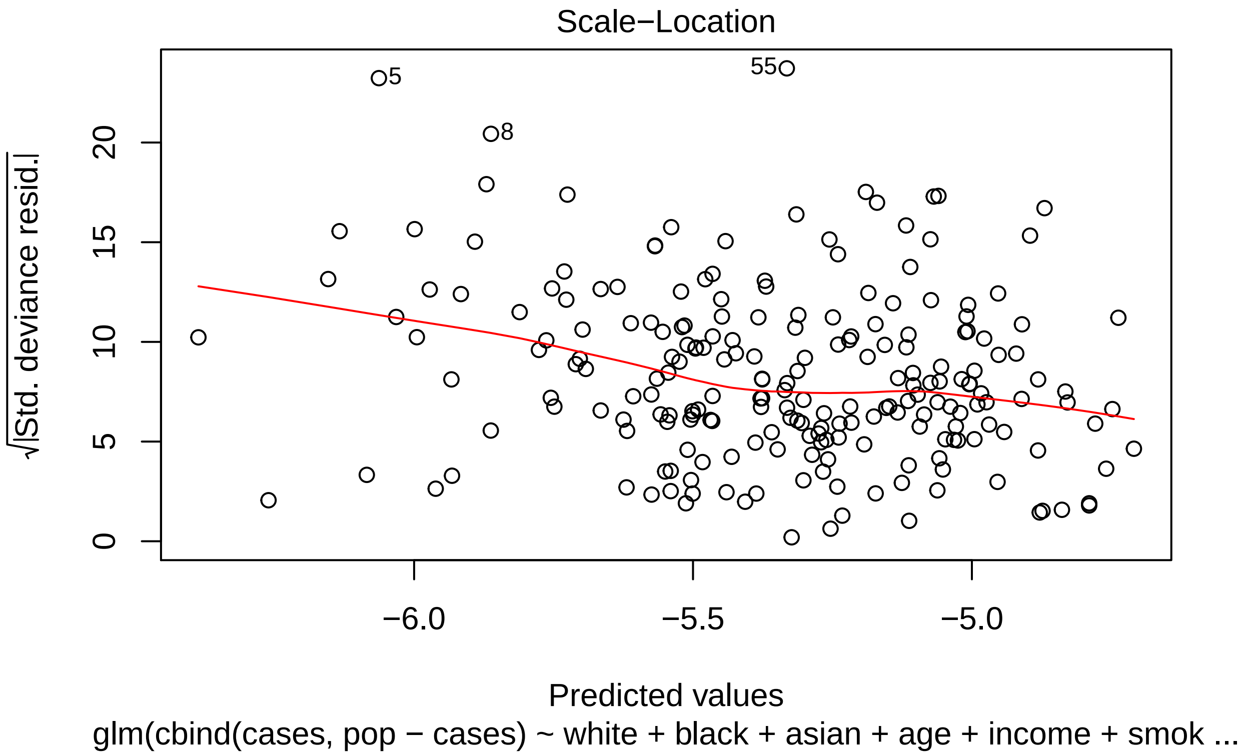
```
##
## Call:
## glm(formula = cbind(cases, pop - cases) ~ white + black + asian +
##     age + income + smoke + insured + work + miles + white:asian +
##     +white:age + white:income + white:insured + white:miles +
##     +black:age + black:income + black:miles + asian:age + asian:income +
##     asian:miles + age:work + income:insured, family = quasibinomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5540  -0.9188  -0.0994   0.7944  13.8987
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -8.748e+00  1.377e+00  -6.351 1.47e-09 ***
## white          -8.144e-03  1.972e-02  -0.413  0.68009    
## black           2.006e-02  8.677e-03   2.312  0.02182 *  
## asian           1.882e-02  1.117e-02   1.685  0.09358 .  
## age             9.900e-02  6.812e-02   1.453  0.14773    
## income          4.147e-05  1.594e-05   2.601  0.01000 *  
## smoke           1.106e-02  5.714e-03   1.936  0.05434 .  
## insured         1.660e-02  2.243e-02   0.740  0.46021    
## work           -1.646e-02  4.919e-03  -3.347  0.00098 ***
## miles          -2.340e-02  1.260e-02  -1.857  0.06476 .  
## white:asian     1.767e-04  1.219e-04   1.450  0.14868    
## white:age      -1.267e-03  6.476e-04  -1.956  0.05193 .  
## white:income   -1.238e-07  1.456e-07  -0.850  0.39625    
## white:insured   2.885e-04  2.639e-04   1.093  0.27559    
## white:miles     2.245e-04  1.419e-04   1.582  0.11530    
## black:age      -1.202e-03  6.693e-04  -1.796  0.07403 .  
## black:income   -1.219e-07  1.389e-07  -0.877  0.38137    
## black:miles     2.853e-04  1.940e-04   1.471  0.14290    
## asian:age      -1.661e-03  7.277e-04  -2.283  0.02352 *  
## asian:income   -1.938e-07  1.529e-07  -1.267  0.20658    
## asian:miles     4.135e-04  2.106e-04   1.963  0.05102 .  
## age:work        1.103e-03  3.056e-04   3.610  0.00039 ***
## income:insured -2.806e-07  1.930e-07  -1.454  0.14761    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 4.979552)
##
##     Null deviance: 3829.77  on 217  degrees of freedom
## Residual deviance:  638.99  on 195  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
plot(fitqbnpop)
```

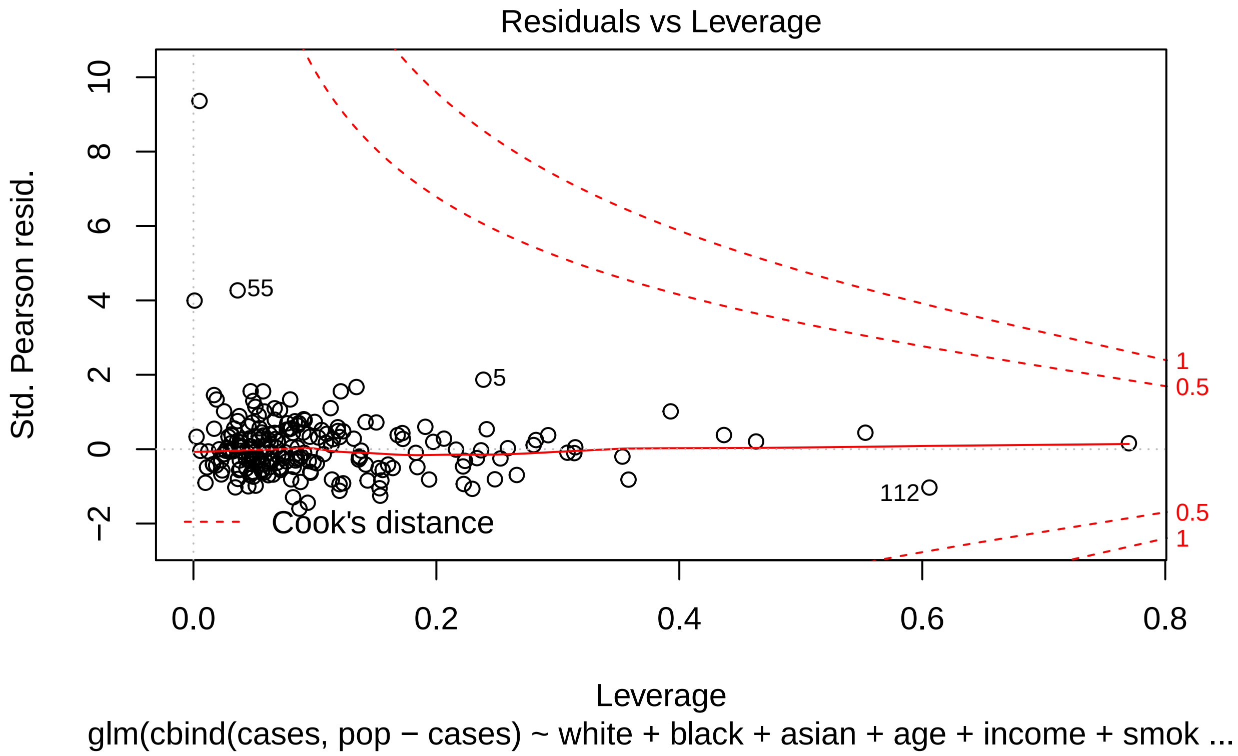

Residuals vs Fitted
Residuals
15
10
5
0
−5
210
55
218
−6.0
−5.5
−5.0
Predicted values
glm(cbind(cases, pop − cases) ~ white + black + asian + age + income + smok ...


Normal Q−Q
Std. deviance resid.
600
400
200
0
−400
5
55
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(cbind(cases, pop − cases) ~ white + black + asian + age + income + smok ...



```r
#Model doesn't necessarily look bad; some values to consider for outlier testing
# BUT resid v fitted and qq look decent overall (removes case 210 as an issue for qqplot)

# ANALYSIS FOR QUASIBINOMIAL: removal of outliers
n <- 218
library(car)
```

```
## Warning: package 'car' was built under R version 3.2.5
```



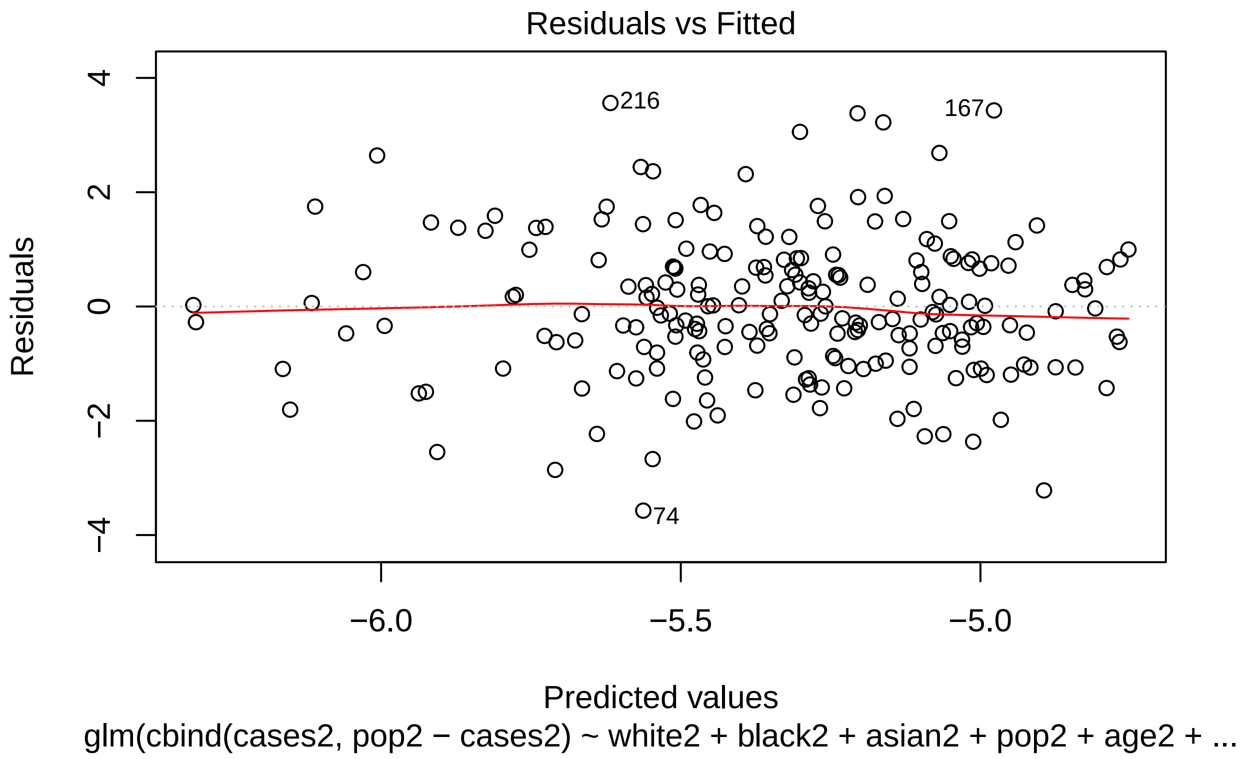
```
outlierTest(fit1, cutoff = 0.05, n.max = n, order = TRUE)
```

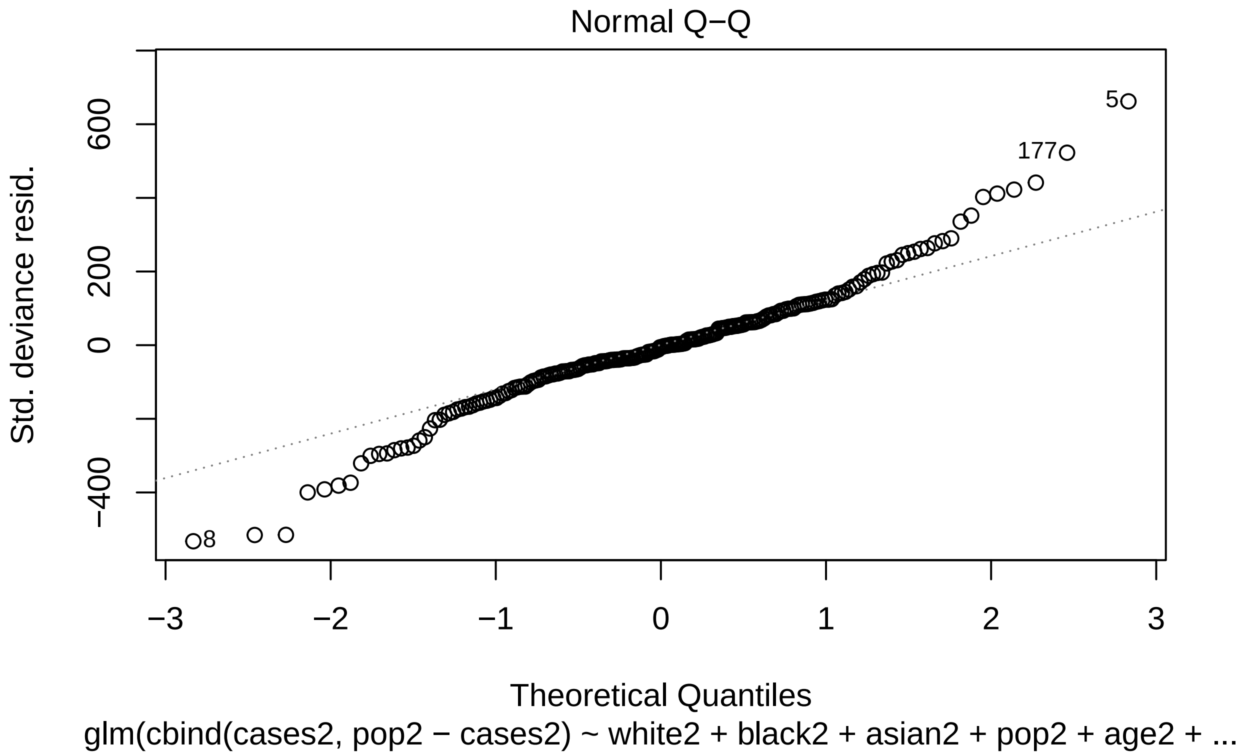
```
##     rstudent unadjusted p-value Bonferroni p
## 210 9.562258          1.1522e-21   2.5117e-19
## 55  4.701854          2.5781e-06   5.6203e-04
```

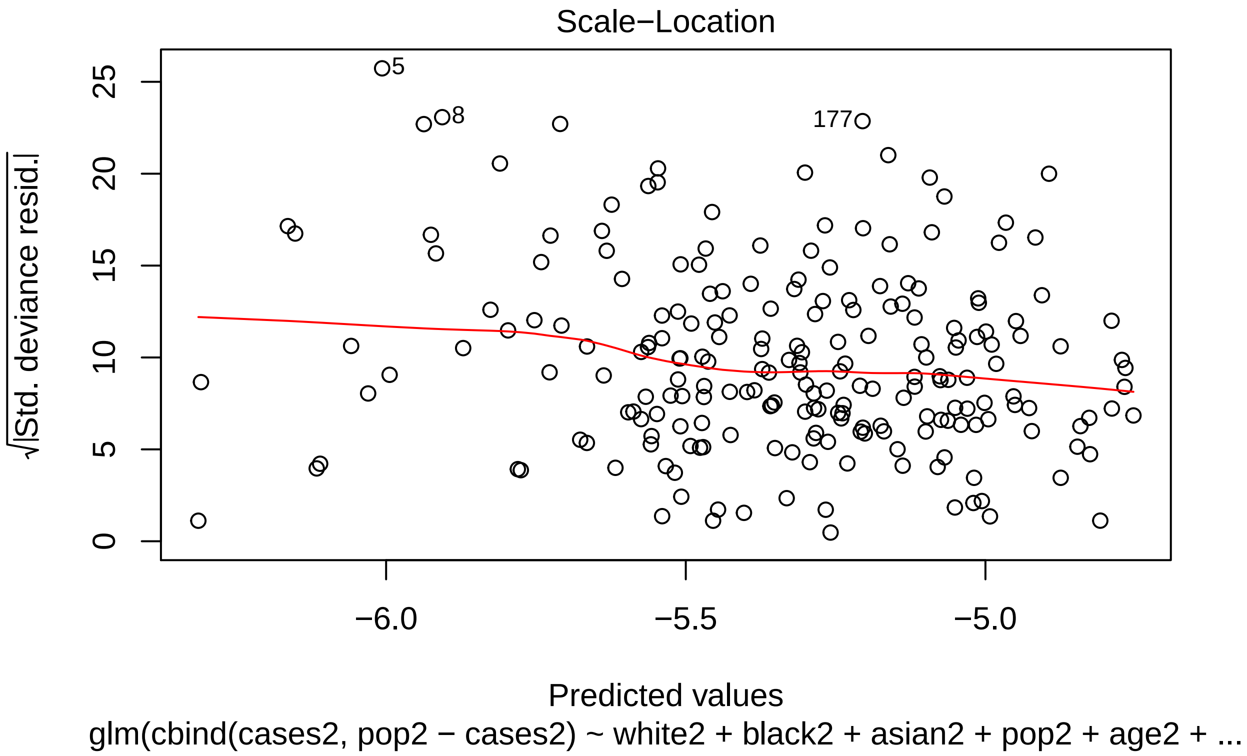
```
#remove cases that are outliers; indices 210 and 55
cancer2 <- cancer[-c(210, 55),]
View(cancer2) #216 entries, good to use

#Reset the variables for the cancer dataset that excludes the outlier values
cases2 <- cancer2$`2010-14 Incidence`
white2 <- cancer2$`% White`
black2 <- cancer2$`% Black`
asian2 <- cancer2$`% Asian`
pop2 <- cancer2$Population
age2 <- cancer2$`% Over 65`
income2 <- cancer2$`Average Income`
smoke2 <- cancer2$`% Tobacco Use`
insured2 <- cancer2$`% Population Insured`
work2 <- cancer2$`% Females (16+) in Laborforce`
miles2 <- cancer2$`Mileage to Nearest Hospital`

fitoutqb <- glm(cbind(cases2, pop2-cases2) ~ white2+black2+asian2+pop2+age2+income2+smoke2+insured2+wor
                 black2:pop2 + black2:age2 + black2:income2 + black2:miles2 + asian2:age2 + asian2:inco
                 income2:insured2, family=quasibinomial(link="logit"))
summary(fitoutqb)
```

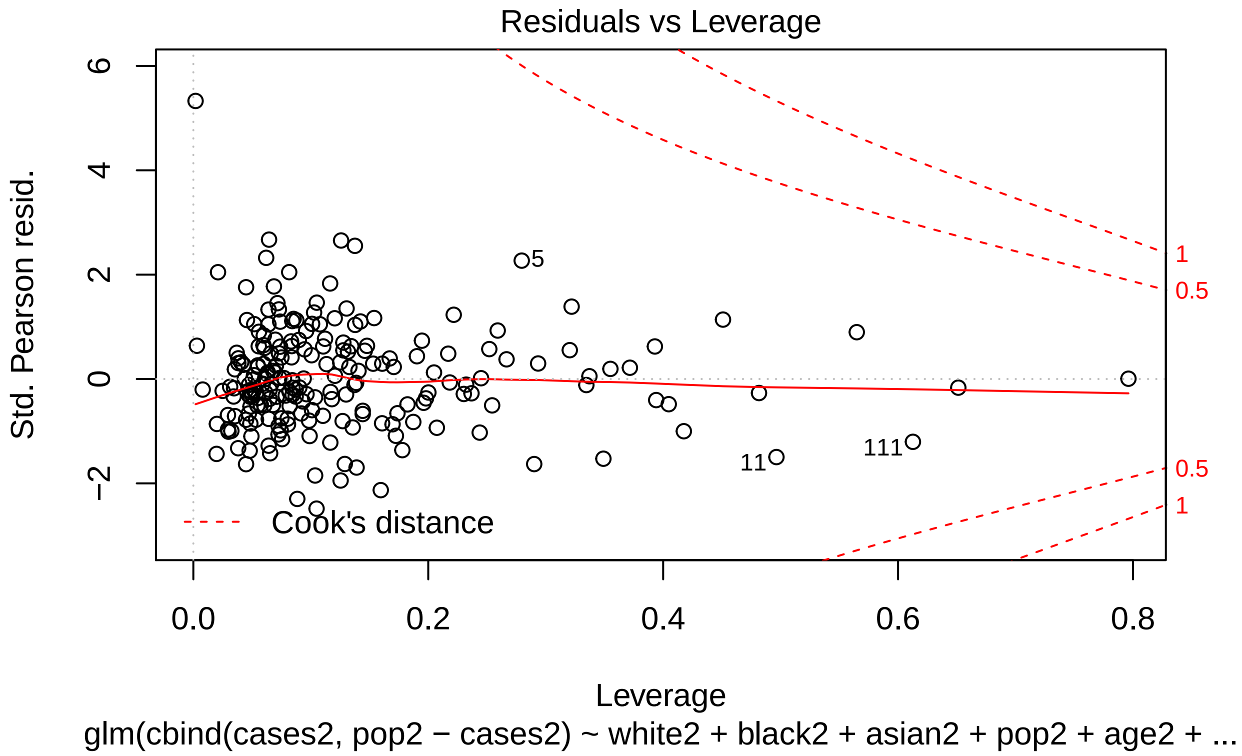
```
##
## Call:
## glm(formula = cbind(cases2, pop2 - cases2) ~ white2 + black2 +
##     asian2 + pop2 + age2 + income2 + smoke2 + insured2 + work2 +
##     miles2 + white2:asian2 + white2:pop2 + white2:age2 + white2:income2 +
##     white2:insured2 + white2:miles2 + black2:pop2 + black2:age2 +
##     black2:income2 + black2:miles2 + asian2:age2 + asian2:income2 +
##     asian2:miles2 + pop2:age2 + pop2:miles2 + age2:work2 + income2:insured2,
##     family = quasibinomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5727  -0.8071  -0.1115   0.7253   3.5614
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -9.885e+00  1.132e+00  -8.731 1.36e-15 ***
## white2            7.633e-03  1.498e-02   0.509 0.611012
## black2            2.916e-02  8.299e-03   3.513 0.000554 ***
## asian2            1.419e-02  7.782e-03   1.823 0.069923 .
## pop2              6.010e-06  4.065e-06   1.478 0.140974
## age2              1.033e-01  4.834e-02   2.137 0.033867 *
## income2           4.848e-05  1.035e-05   4.686 5.33e-06 ***
## smoke2            8.309e-03  3.737e-03   2.223 0.027374 *
## insured2          2.004e-02  1.477e-02   1.357 0.176519
## work2            -1.084e-02  3.522e-03  -3.079 0.002390 **
## miles2           -3.070e-02  1.247e-02  -2.461 0.014740 *
## white2:asian2     1.726e-04  7.891e-05   2.187 0.029949 *
## white2:pop2      -1.635e-07  5.153e-08  -3.172 0.001767 **
## white2:age2      -1.283e-03  4.624e-04  -2.775 0.006072 **
## white2:income2   -1.466e-07  9.397e-08  -1.560 0.120440
## white2:insured2   2.178e-04  1.730e-04   1.259 0.209649
## white2:miles2     2.545e-04  1.216e-04   2.093 0.037707 *
## black2:pop2      -1.138e-07  5.035e-08  -2.261 0.024919 *
## black2:age2      -1.294e-03  4.721e-04  -2.742 0.006698 **
## black2:income2   -1.553e-07  9.029e-08  -1.720 0.087108 .
## black2:miles2     3.162e-04  1.434e-04   2.205 0.028682 *
## asian2:age2      -1.446e-03  5.030e-04  -2.874 0.004517 **
## asian2:income2   -1.896e-07  9.754e-08  -1.944 0.053426 .
## asian2:miles2     5.573e-04  1.485e-04   3.752 0.000233 ***
## pop2:age2         3.246e-07  1.254e-07   2.589 0.010388 *
## pop2:miles2       1.045e-07  6.987e-08   1.496 0.136292
## age2:work2        8.396e-04  2.136e-04   3.931 0.000119 ***
## income2:insured2 -3.303e-07  1.239e-07  -2.665 0.008371 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.992923)
##
##     Null deviance: 3514.37  on 215  degrees of freedom
## Residual deviance:  327.16  on 188  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
plot(fitoutqb)
```




Normal Q−Q
Std. deviance resid.
600
200
0
−400
5
177
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(cbind(cases2, pop2 − cases2) ~ white2 + black2 + asian2 + pop2 + age2 + ...


Scale−Location
√|Std. deviance resid.|
0
5
10
15
20
25
5
8
177
−6.0
−5.5
−5.0
Predicted values
glm(cbind(cases2, pop2 − cases2) ~ white2 + black2 + asian2 + pop2 + age2 + ...



```
#Resid v fitted looks good, light tails on qqplot, but nothing significant
#DOES this actually help, or does it just allow other points to be new outliers...
n <- 216
outlierTest(fitoutqb, cutoff = 0.05, n.max = n, order = TRUE)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 74 -2.898223          0.0037528      0.81061
```

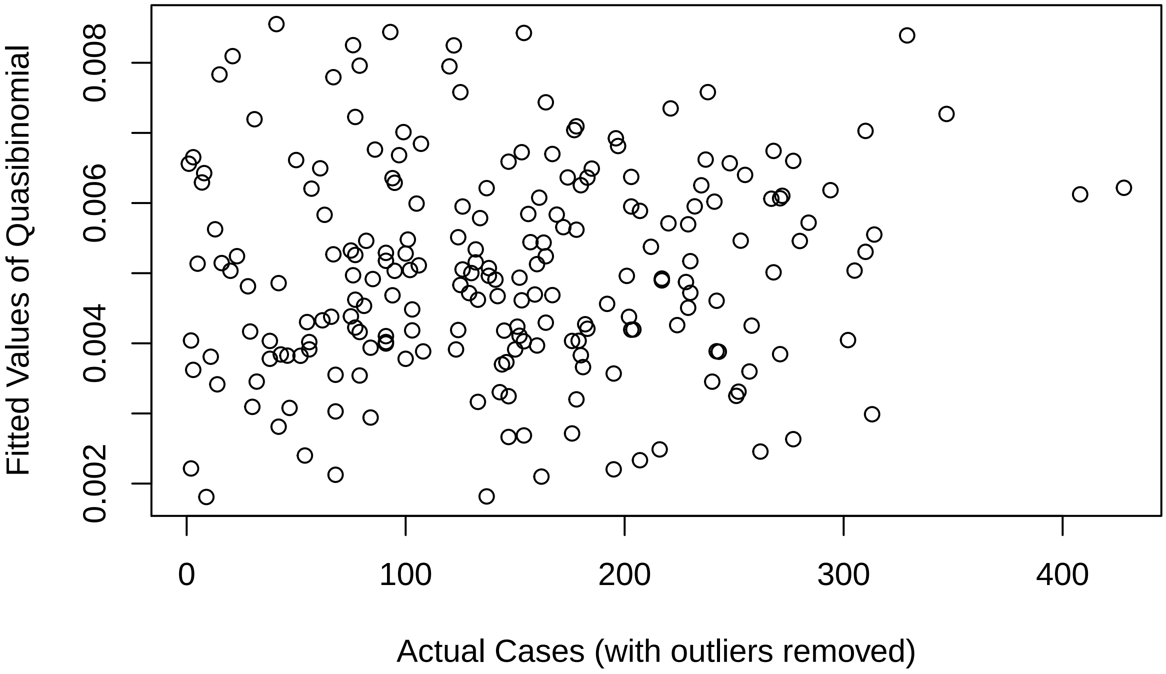
```
#It helps, new outliers are NOT introduced, therefore removing those cases for this model
#helps improve the fitted model

library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 3.2.5

## ResourceSelection 0.3-2   2017-02-28
```

```
hoslem.test(cases2, fitted(fitoutqb))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cases2, fitted(fitoutqb)
## X-squared = 978940000, df = 8, p-value < 2.2e-16
```

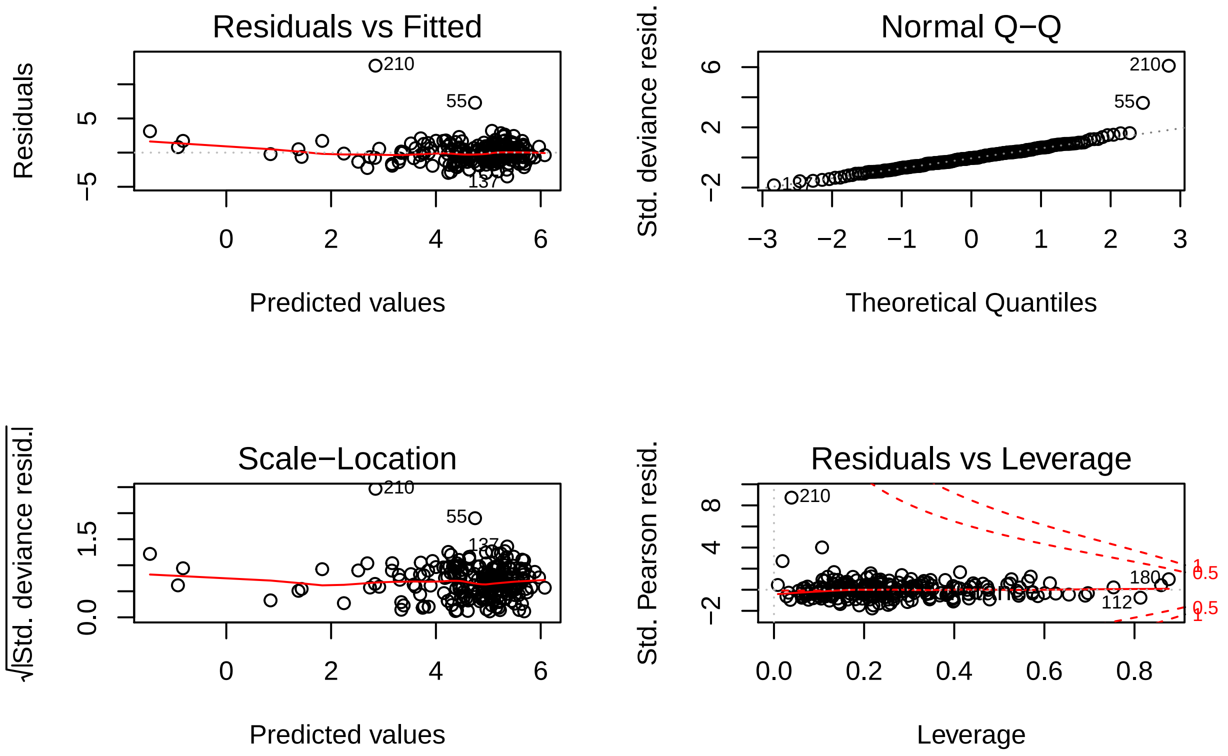
```
#The p-value is low which tells us we have a significant difference
# between the actuals and the fitted (we can see that in the plot below)
plot(fitted(fitoutqb)~cases2, xlab="Actual Cases (with outliers removed)", ylab="Fitted Values of Quasil
```



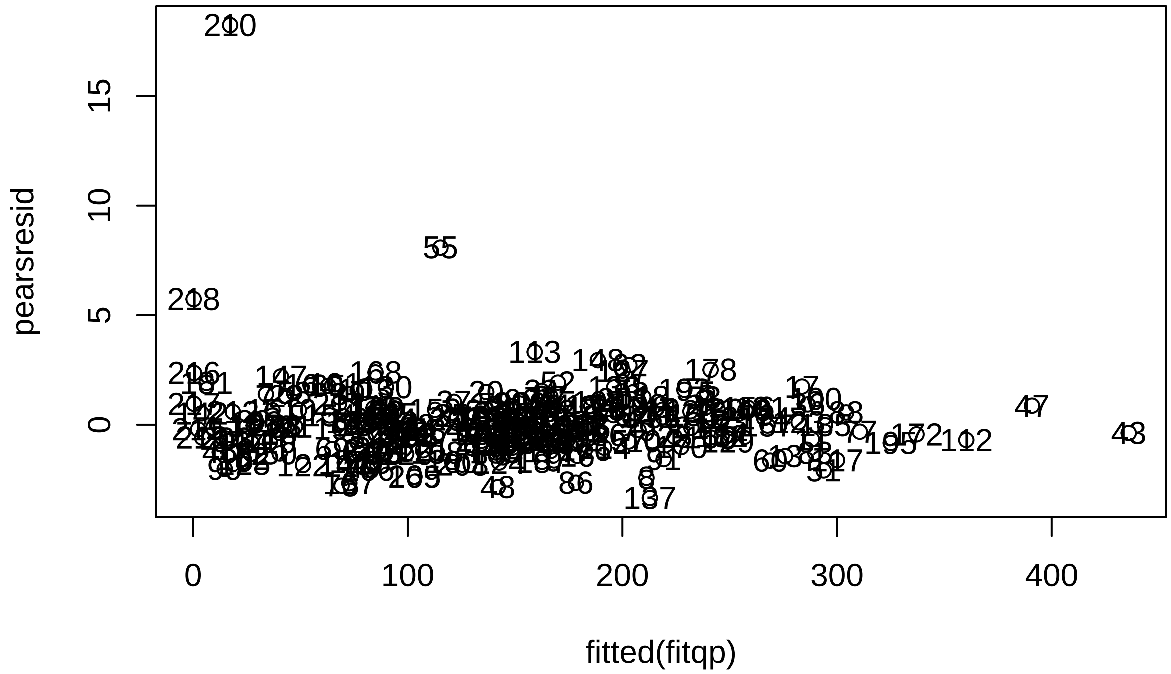
```
#Since nothing else is an indicator for misfit
# We chalk it up to the fact that there are definitely missing predictors here
# therefore we cannot match the actual values as well as we'd like
```

```
#2. Fit Poisson in order to do AIC-based backwards selection (w/ offset=pop)
fitpoi <- glm(cases ~ (white+black+asian+pop+age+income+smoke+insured+work+miles)^2, offset=log(pop), fa
## Changed backwards elimination of the models to k=4 to be more strict
full <- fitpoi
null <- lm(cases ~ 1) #null is just the response with intercept
s2 <- step(full, scope=list(lower=null, upper=full), direction="backward", k=4)
summary(s2)
```

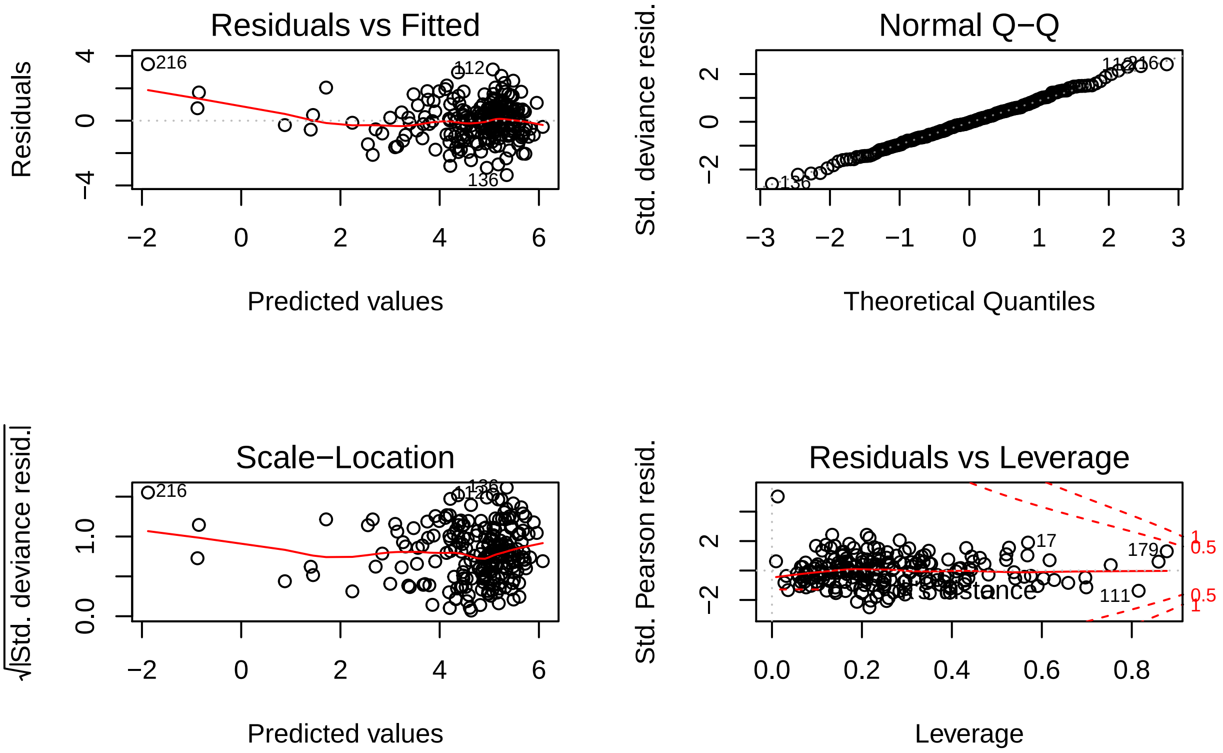
```
#Terms from BE: work + miles + white:asian + white:pop +
# white:age + white:income + white:insured + white:miles +
# black:pop + black:age + black:income + black:miles + asian:age +
# asian:income + asian:miles + pop:age + pop:miles + age:work +
# income:insured

##Quasipoisson with ALL terms - tester
fitqp <- glm(cases ~ (white+black+asian+pop+age+income+smoke+insured+work+miles)^2, offset=log(pop), fam
par(mfrow=c(2,2))
plot(fitqp)
```



```
#Residuals v fitted show that some values have lower predictions, may be interesting to
#investigate those points
# Need to look at 210 and 55 in a Pearson residual plot so that weights are accted for
pearsresid <- residuals(fitqp, type="pearson")
par(mfrow=c(1,1))
plot(pearsresid~fitted(fitqp))
#do this to identify case number
text(fitted(fitqp), pearsresid)
```



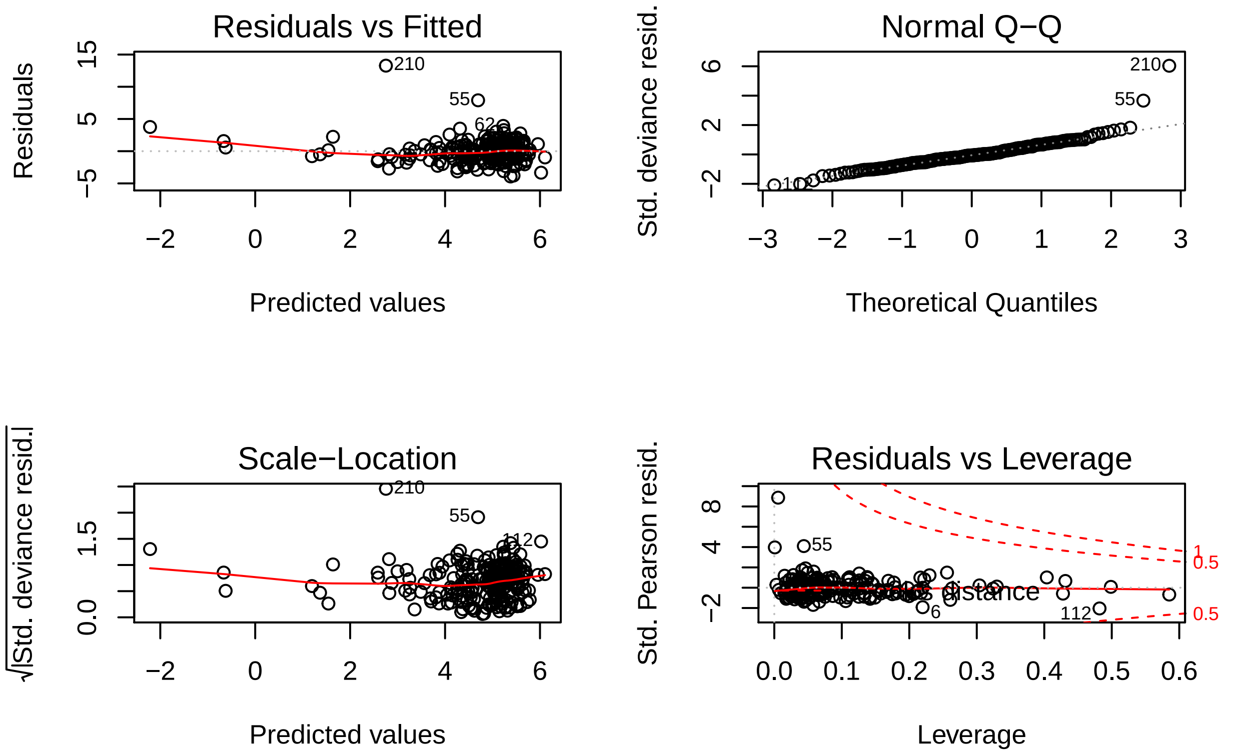
```
#Case 210 stands out still - zip code 60157; then 55 and 218 as shown in previous

#FIT QUASIPOI WITHOUT OUTLIERS
fitqpout <- glm(cases2 ~ (white2+black2+asian2+pop2+age2+income2+smoke2+insured2+work2+miles2)^2, offse
par(mfrow=c(2,2))
plot(fitqpout)
```



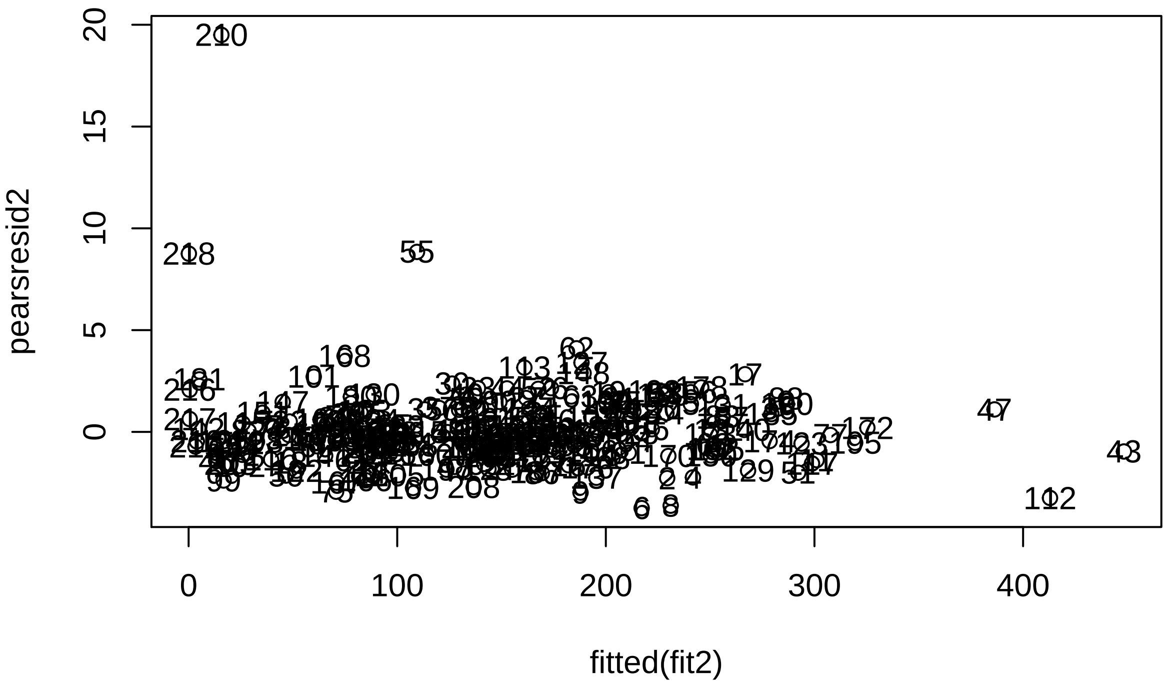
```
## How does this compare to the BE without outliers model.... similar

# Fit QUASIPOISSON model with BE identified terms
##a. Model
fit2 <- glm(cases ~ work + miles + white:asian + white:pop + white:age + white:income + white:insured +
              black:pop + black:age + black:income + black:miles + asian:age +
              asian:income + asian:miles + pop:age + pop:miles + age:work +
              income:insured, offset=log(pop), family=quasipoisson(link="log"))
#mixed main effects and interaction terms
summary(fit2)
```

```
##
## Call:
## glm(formula = cases ~ work + miles + white:asian + white:pop +
##     white:age + white:income + white:insured + white:miles +
##     black:pop + black:age + black:income + black:miles + asian:age +
##     asian:income + asian:miles + pop:age + pop:miles + age:work +
##     income:insured, family = quasipoisson(link = "log"), offset = log(pop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9212  -1.0372  -0.1282   0.9142  13.2748
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.880e+00  2.409e-01 -24.412  < 2e-16 ***
## work           -9.887e-03  4.753e-03  -2.080 0.038779 *
## miles          -4.317e-02  1.661e-02  -2.599 0.010051 *
## white:asian     2.518e-04  1.156e-04   2.178 0.030617 *
## white:pop      -8.212e-08  3.180e-08  -2.583 0.010528 *
## white:age      -2.503e-04  1.634e-04  -1.532 0.127189
## white:income   -5.328e-08  4.751e-08  -1.121 0.263463
## white:insured   7.341e-05  3.084e-05   2.380 0.018239 *
## miles:white     3.877e-04  1.604e-04   2.417 0.016548 *
## pop:black      -2.486e-08  3.983e-08  -0.624 0.533257
## age:black       8.142e-05  2.183e-04   0.373 0.709603
## income:black   -1.434e-08  6.020e-08  -0.238 0.811917
## miles:black     4.201e-04  1.926e-04   2.182 0.030319 *
## asian:age      -4.957e-04  2.985e-04  -1.661 0.098366 .
## asian:income   -1.324e-07  8.331e-08  -1.589 0.113648
## miles:asian     5.789e-04  2.134e-04   2.713 0.007254 **
## pop:age         1.412e-07  1.795e-07   0.787 0.432375
## miles:pop       1.370e-07  1.005e-07   1.363 0.174393
## work:age        1.021e-03  2.845e-04   3.587 0.000421 ***
## income:insured  7.667e-08  4.647e-08   1.650 0.100549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.868172)
##
##     Null deviance: 3810.16  on 217  degrees of freedom
## Residual deviance:  684.58  on 198  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
##b. Diagnostics
par(mfrow = c(2,2))
plot(fit2)
```



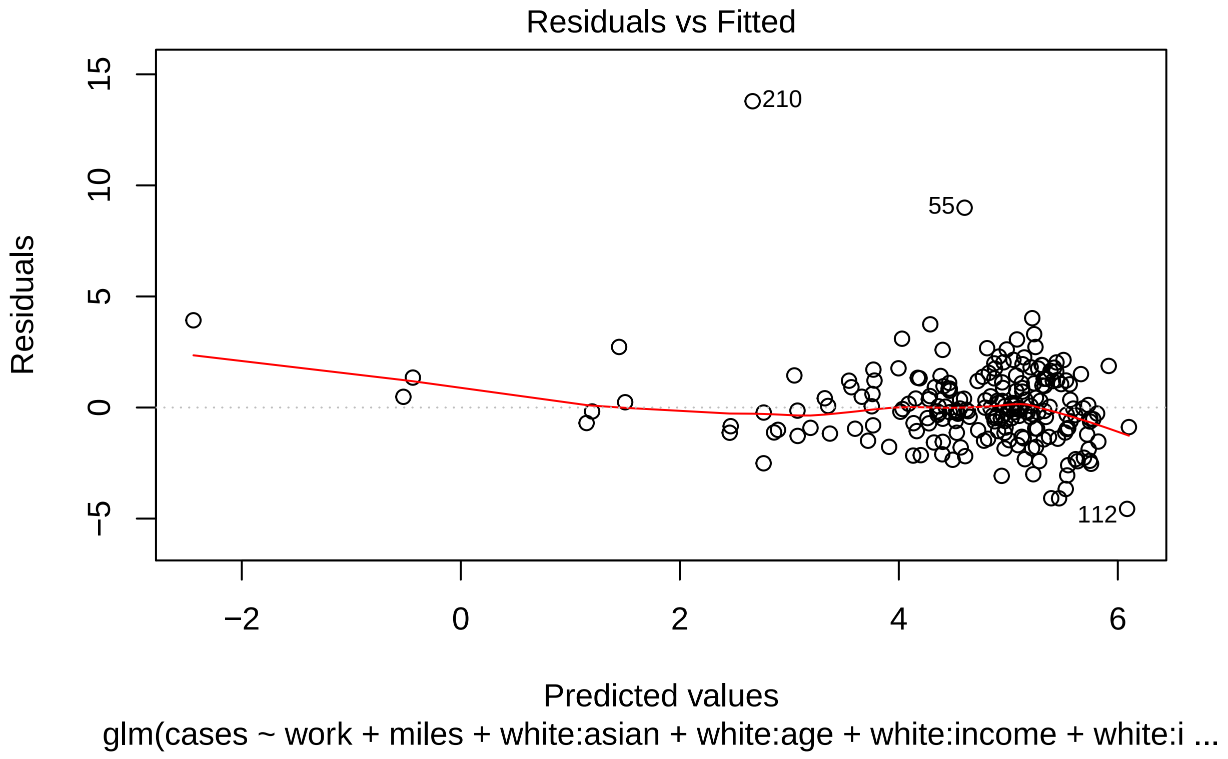
```r
par(mfrow = c(1,1))

#Plots look relatively good, but we want to see what 210 and 55 look like with Pearson
pearsresid2 <- residuals(fit2, type="pearson")
par(mfrow=c(1,1))
plot(pearsresid2~fitted(fit2))
#do this to identify case number
text(fitted(fit2), pearsresid2)
```

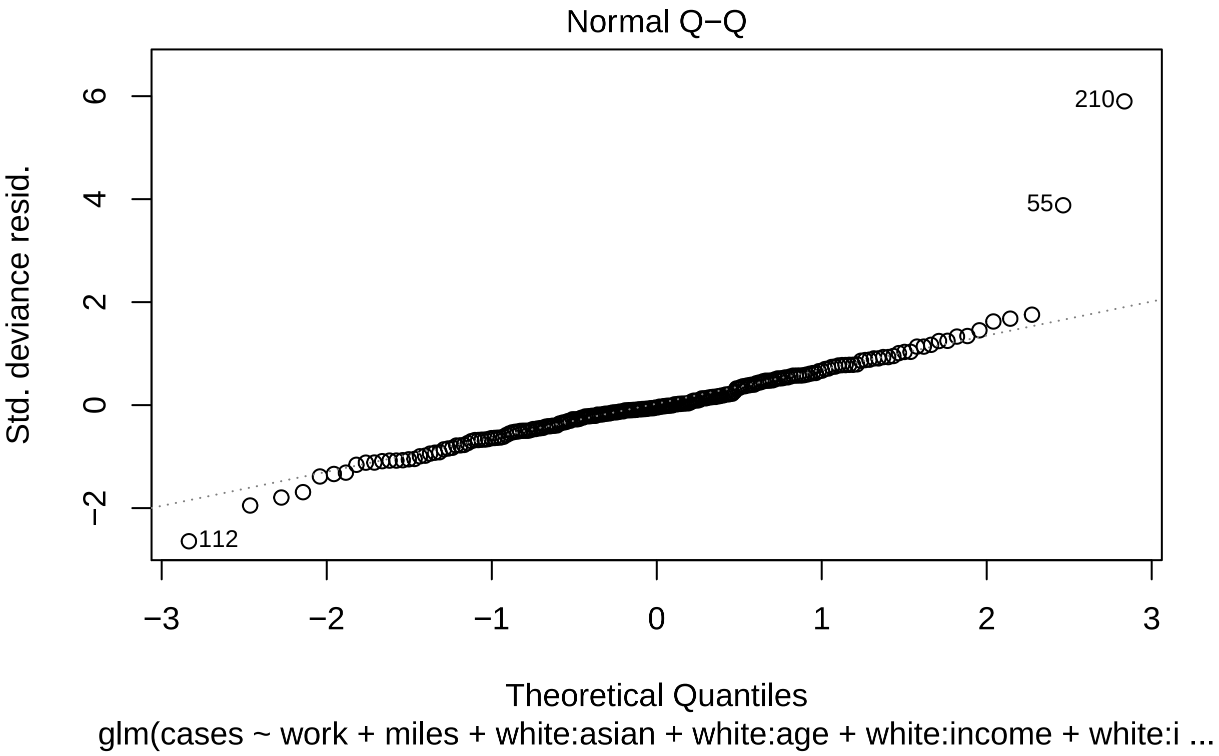


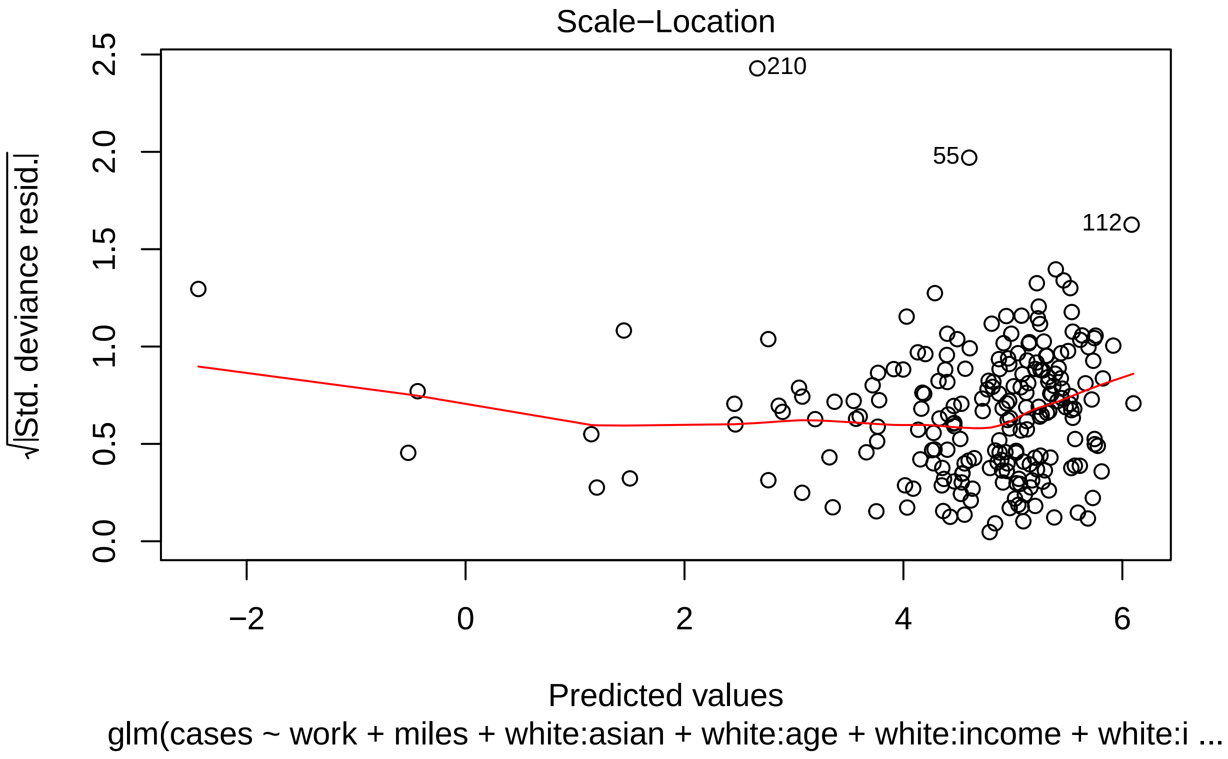
```
## They again show up with higher residuals; will need to investigate them for outlier test

#Fit QP withOUT pop as a predictor, even though BE says to include
#(also means removing interactions that have pop)
fitqpnpop <- glm(cases ~ work + miles + white:asian + white:age + white:income + white:insured + white:r
              + black:age + black:income + black:miles + asian:age + asian:income + asian:miles + age:wo
              + income:insured, offset=log(pop), family=quasipoisson(link="log"))
summary(fitqpnpop)
```

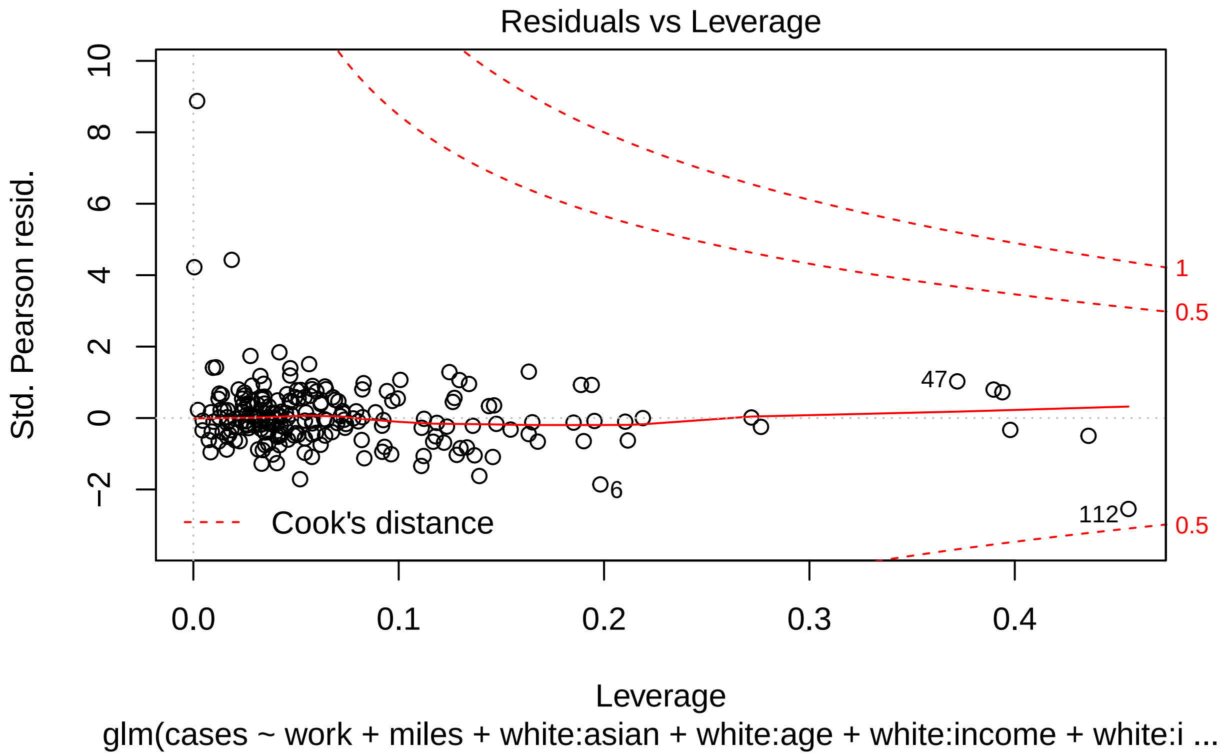
```
##
## Call:
## glm(formula = cases ~ work + miles + white:asian + white:age +
##     white:income + white:insured + white:miles + +black:age +
##     black:income + black:miles + asian:age + asian:income + asian:miles +
##     age:work + income:insured, family = quasipoisson(link = "log"),
##     offset = log(pop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5635  -0.9515  -0.1040   1.0797  13.7871
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.007e+00  2.266e-01 -26.506  < 2e-16 ***
## work           -1.282e-02  4.434e-03  -2.892  0.00424 **
## miles          -2.653e-02  1.214e-02  -2.185  0.03002 *
## white:asian      2.510e-04  1.202e-04   2.087  0.03811 *
## white:age       -2.365e-04  1.670e-04  -1.417  0.15810
## white:income    -6.790e-08  4.780e-08  -1.420  0.15703
## white:insured    7.489e-05  3.214e-05   2.330  0.02077 *
## miles:white      2.647e-04  1.363e-04   1.942  0.05354 .
## age:black        1.966e-04  2.072e-04   0.949  0.34376
## income:black   -6.375e-09  6.055e-08  -0.105  0.91625
## miles:black      2.798e-04  1.822e-04   1.535  0.12624
## asian:age       -5.165e-04  3.159e-04  -1.635  0.10368
## asian:income    -1.287e-07  8.733e-08  -1.473  0.14226
## miles:asian      5.030e-04  2.058e-04   2.444  0.01539 *
## work:age         1.107e-03  2.550e-04   4.342 2.23e-05 ***
## income:insured   9.067e-08  4.689e-08   1.934  0.05455 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.471895)
##
##     Null deviance: 3810.2  on 217  degrees of freedom
## Residual deviance:  756.8  on 202  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
plot(fitqpnpop)
```

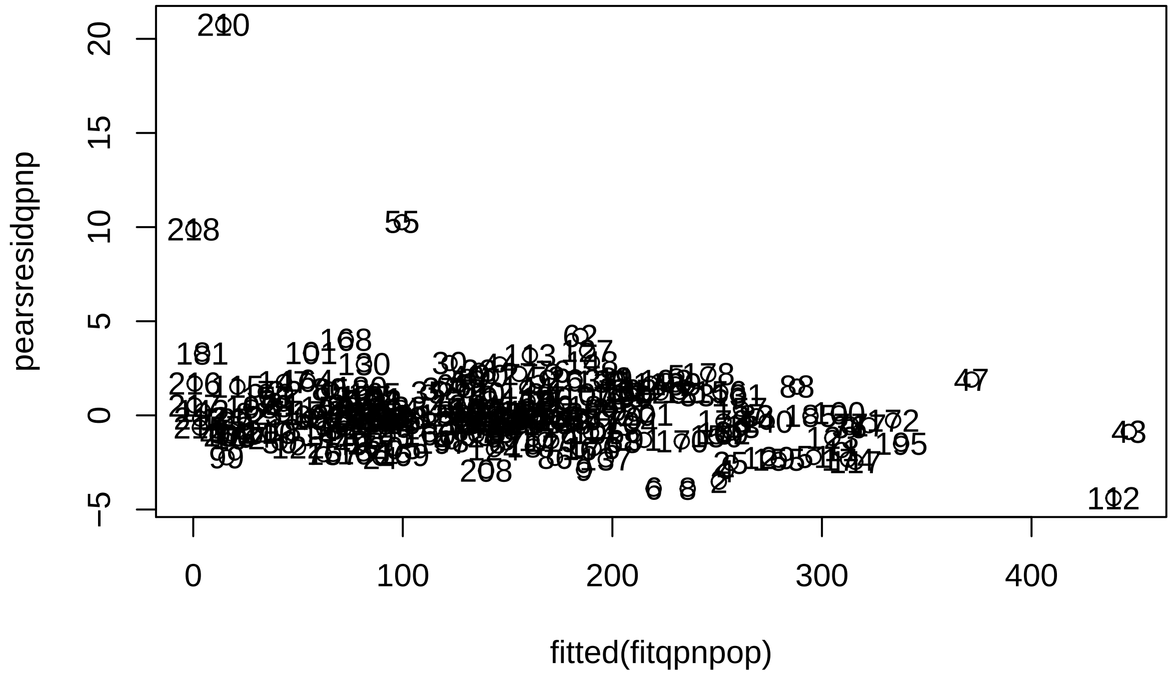



Normal Q−Q
Std. deviance resid.
6
4
2
0
−2
210
55
112
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(cases ~ work + miles + white:asian + white:age + white:income + white:i ...


Scale−Location
210
55
112
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
−2
0
2
4
6
Predicted values
glm(cases ~ work + miles + white:asian + white:age + white:income + white:i ...



```
#not much different than with pop included and its interactions too;
# may have something to do with only black:pop being significant in the backwards elimination model
#consider removing pop as a regressor; although 210 is still showing up in residuals

#Look at pearson for no pop model
pearsresidqpnp <- residuals(fitqpnpop, type="pearson")
par(mfrow=c(1,1))
plot(pearsresidqpnp~fitted(fitqpnpop))
#do this to identify case number
text(fitted(fitqpnpop), pearsresidqpnp) #Case 210 is the standout again...
```



```
# ANALYSIS FOR QUASIPOISSON: removal of outliers
n <- 218
outlierTest(fit2, cutoff = 0.05, n.max = n, order = TRUE)
```

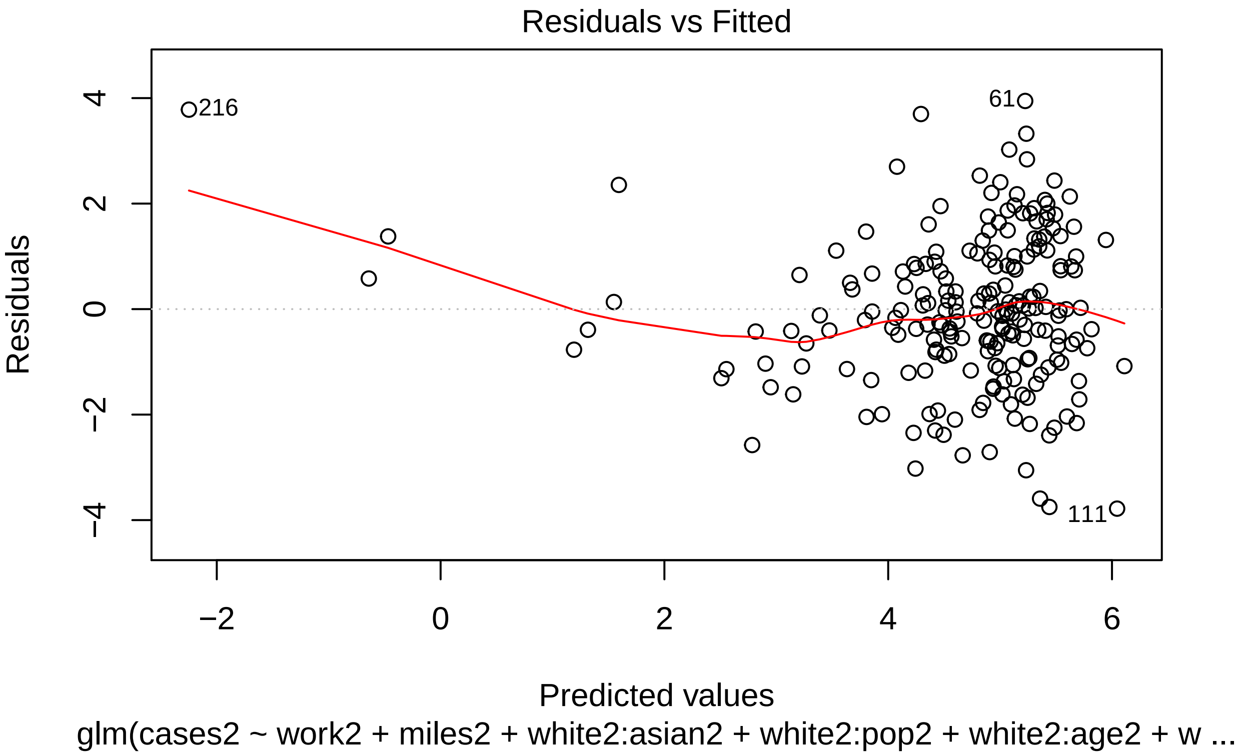
```
##     rstudent unadjusted p-value Bonferonni p
## 210 8.323155         8.5652e-17   1.8672e-14
## 55  4.581839         4.6090e-06   1.0048e-03
```

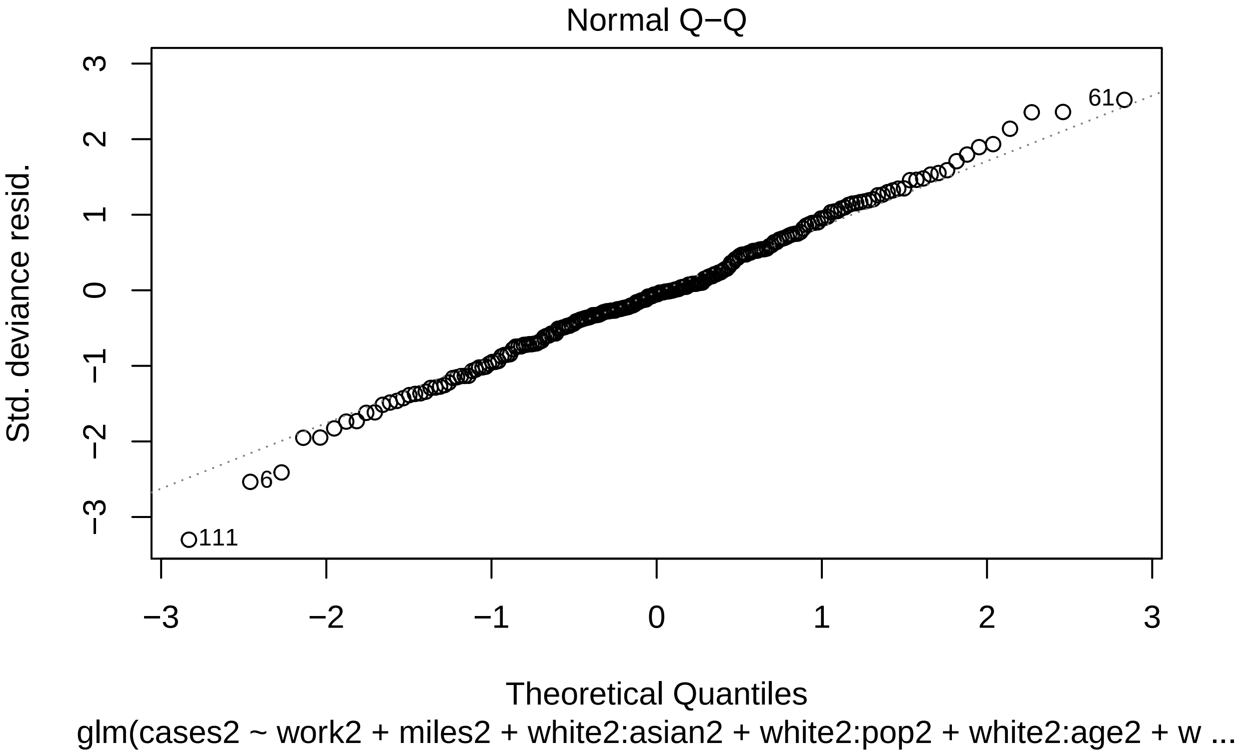
```
#remove cases that are outliers; indices 210 and 55
#ALL MODELS HAVE SAME OUTLIERS SO USE THE SAME CANCER2 DATASET!
#Cancer2 excludes the outlier cases already

fitoutqp <- glm(cases2 ~ work2 + miles2 + white2:asian2 + white2:pop2 + white2:age2 + white2:income2 + w
                  black2:pop2 + black2:age2 + black2:income2 + black2:miles2 + asian2:age2 +
                  asian2:income2 + asian2:miles2 + pop2:age2 + pop2:miles2 + age2:work2 +
                  income2:insured2, offset=log(pop2), family=quasipoisson(link="log"))
summary(fitoutqp)
```

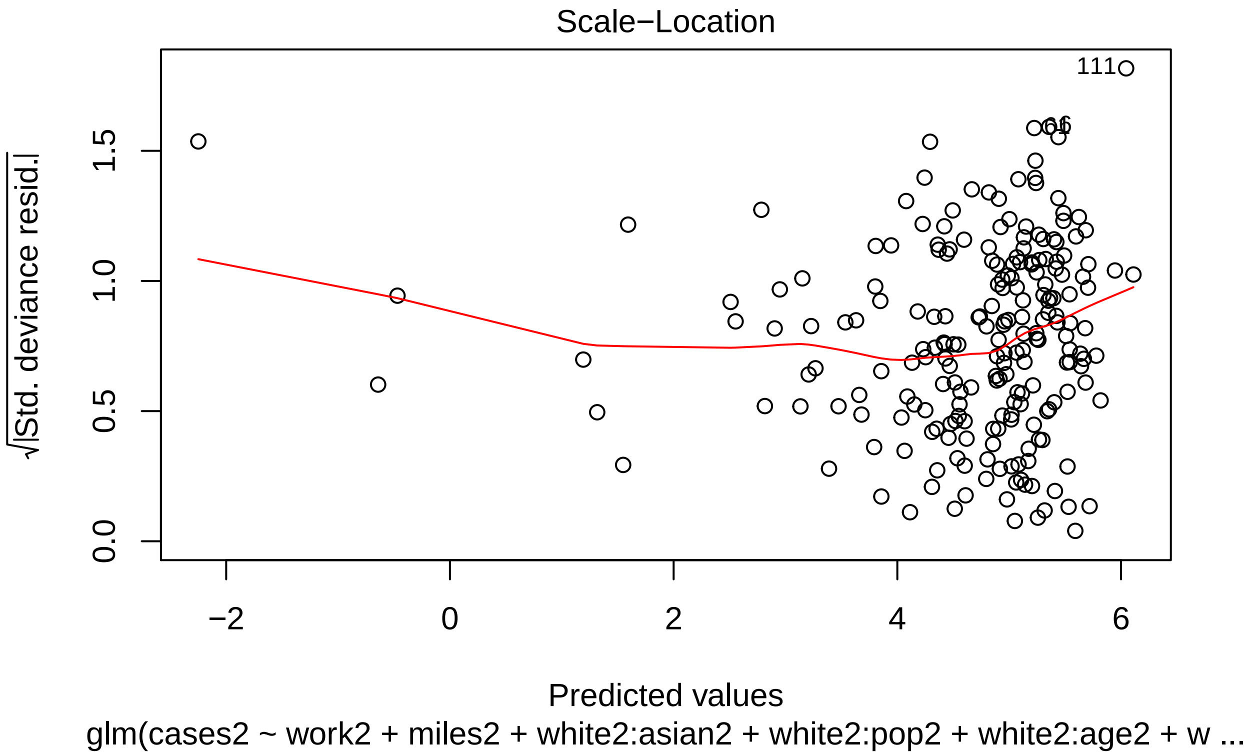
```
##
## Call:
## glm(formula = cases2 ~ work2 + miles2 + white2:asian2 + white2:pop2 +
##     white2:age2 + white2:income2 + white2:insured2 + white2:miles2 +
##     black2:pop2 + black2:age2 + black2:income2 + black2:miles2 +
##     asian2:age2 + asian2:income2 + asian2:miles2 + pop2:age2 +
##     pop2:miles2 + age2:work2 + income2:insured2, family = quasipoisson(link = "log"),
##     offset = log(pop2))
##
```

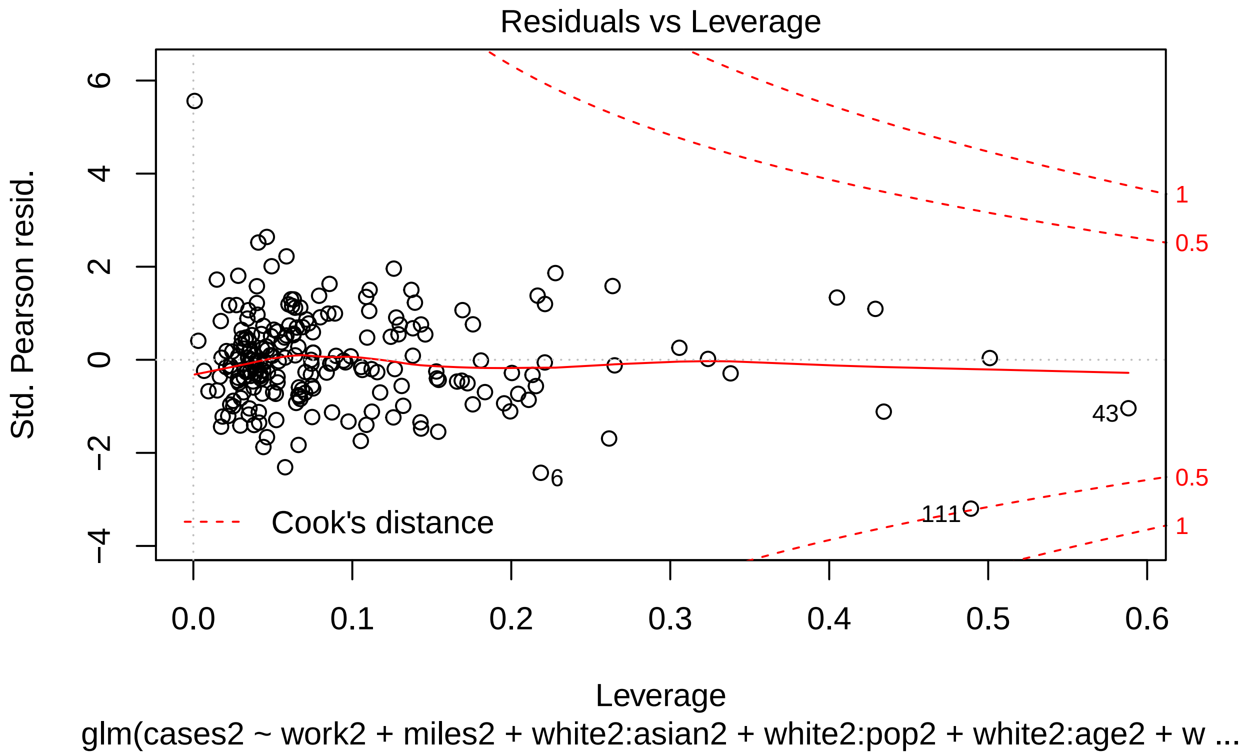
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7822  -0.9525  -0.0806   0.8678   3.9455
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.979e+00  1.760e-01 -33.964  < 2e-16 ***
## work2            -8.842e-03  3.457e-03  -2.558 0.011286 *
## miles2           -3.549e-02  1.214e-02  -2.924 0.003859 **
## white2:asian2     2.206e-04  8.423e-05   2.620 0.009494 **
## white2:pop2      -7.452e-08  2.310e-08  -3.226 0.001472 **
## white2:age2      -1.925e-04  1.185e-04  -1.624 0.105907
## white2:income2   -3.993e-08  3.464e-08  -1.153 0.250451
## white2:insured2   7.091e-05  2.248e-05   3.154 0.001862 **
## miles2:white2     3.193e-04  1.169e-04   2.731 0.006887 **
## pop2:black2      -2.941e-08  2.895e-08  -1.016 0.310881
## age2:black2       1.274e-04  1.581e-04   0.805 0.421550
## income2:black2    9.604e-09  4.396e-08   0.218 0.827317
## miles2:black2     3.419e-04  1.405e-04   2.434 0.015824 *
## asian2:age2      -4.733e-04  2.170e-04  -2.181 0.030357 *
## asian2:income2   -1.074e-07  6.073e-08  -1.768 0.078649 .
## miles2:asian2     5.530e-04  1.552e-04   3.564 0.000459 ***
## pop2:age2         1.923e-07  1.304e-07   1.474 0.142027
## miles2:pop2       1.066e-07  7.346e-08   1.451 0.148432
## work2:age2        9.247e-04  2.067e-04   4.474  1.3e-05 ***
## income2:insured2  6.356e-08  3.388e-08   1.876 0.062167 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.569668)
##
##     Null deviance: 3498.21  on 215  degrees of freedom
## Residual deviance:  438.53  on 196  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
plot(fitoutqp)
```

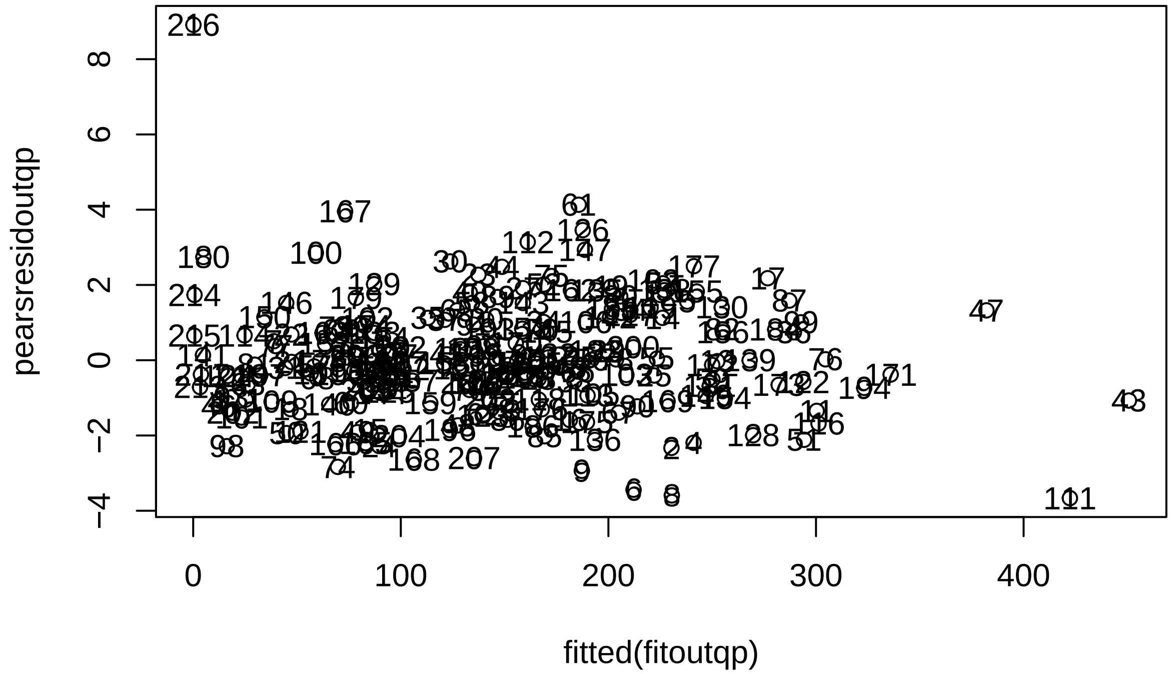

Residuals vs Fitted
Residuals
4
2
0
−2
−4
−2
0
2
4
6
216
61
111
Predicted values
glm(cases2 ~ work2 + miles2 + white2:asian2 + white2:pop2 + white2:age2 + w ...


Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
−3
61
6
111
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(cases2 ~ work2 + miles2 + white2:asian2 + white2:pop2 + white2:age2 + w ...


Scale−Location
111
61
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
−2
0
2
4
6
Predicted values
glm(cases2 ~ work2 + miles2 + white2:asian2 + white2:pop2 + white2:age2 + w ...



```
#Resid v fitted is not ideal, but it is fine and isn't worse than others above
## qqplot looks overall good, nothing significant

pearsresidoutqp <- residuals(fitoutqp, type="pearson")
par(mfrow=c(1,1))
plot(pearsresidoutqp~fitted(fitoutqp))
#do this to identify case number
text(fitted(fitoutqp), pearsresidoutqp) #highlights 216
```



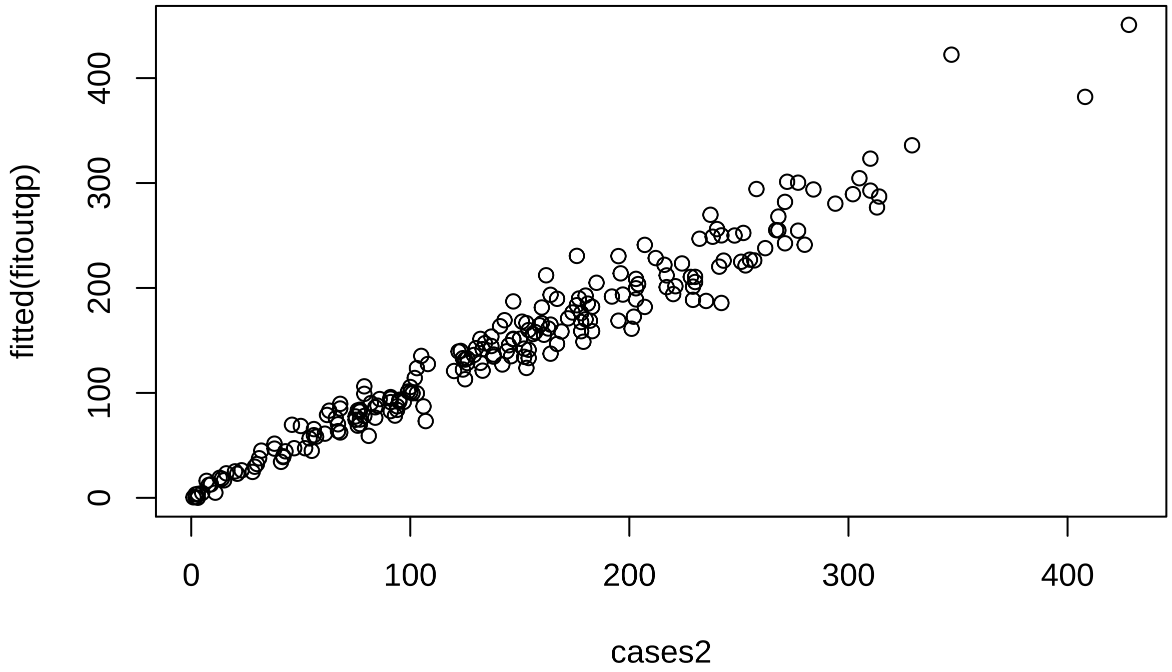
```
## BUT we see that pearson residuals DO NOT fan, so we are not worried about
# residvfitted plot here

#DOES this actually help, or does it just allow other points to be new outliers...
n <- 216
outlierTest(fitoutqp, cutoff = 0.05, n.max = n, order = TRUE)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 111 -3.591871         0.00032831     0.070915
```

```
#NO, new outliers are NOT introduced, therefore removing those cases for this model
#helps improve the fitted model

hoslem.test(cases2, fitted(fitoutqp))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cases2, fitted(fitoutqp)
## X-squared = -0.31741, df = 8, p-value = 1
```

```
#The p-value is very high which tells us we DO NOT have a significant difference
# between the actuals and the fitted (we can see that in the plot below)
plot(fitted(fitoutqp)~cases2)
```

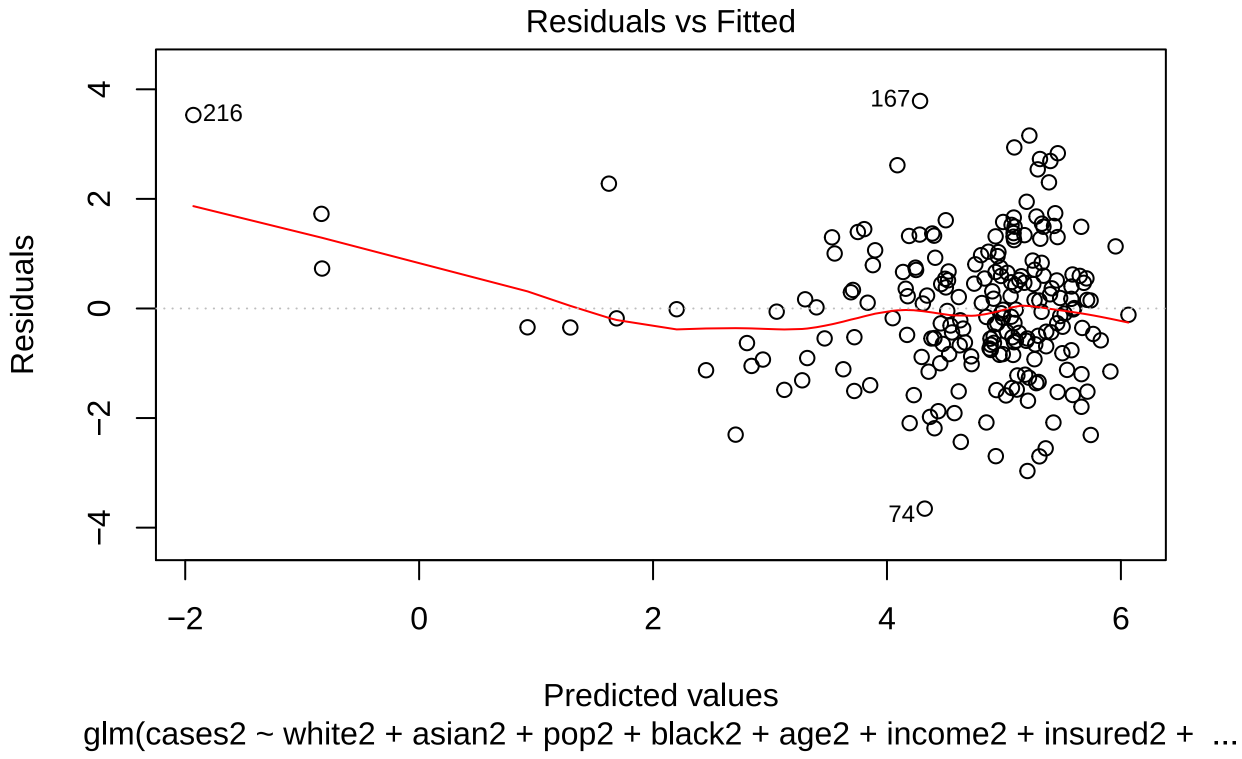


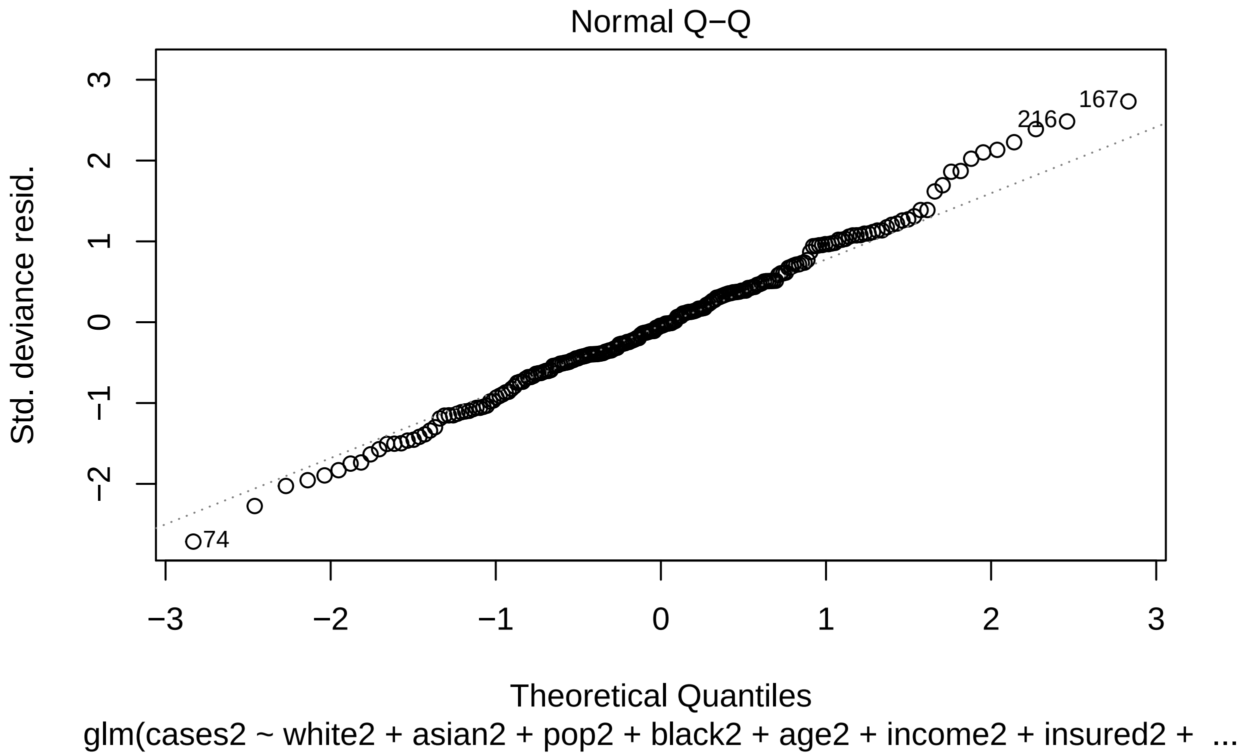
```
#This indicates a good fit for the model; the actuals and fitted line up diagonally
```

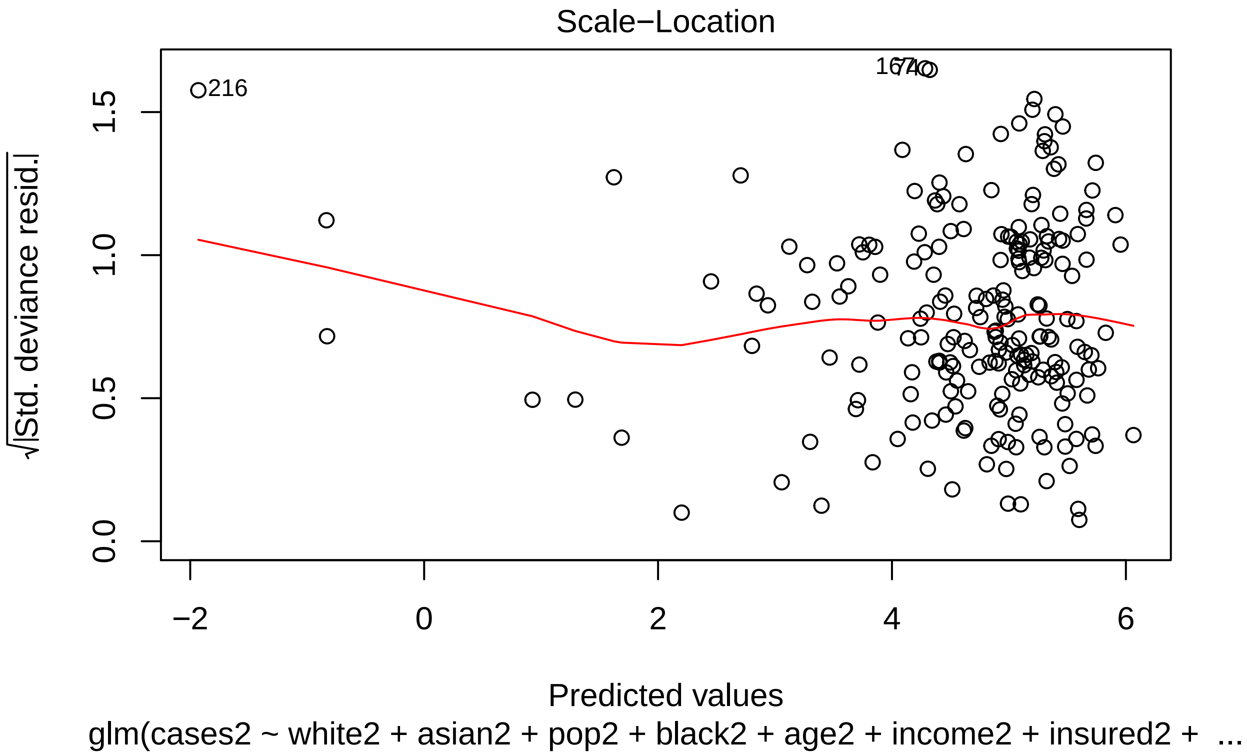
```
##ISSUE of Marginality Principle not applicable to fitoutqp
fitmargqp <- glm(cases2 ~ white2+asian2+pop2+black2+age2+income2+insured2+work2 + miles2 + white2:asian2
      black2:pop2 + black2:age2 + black2:income2 + black2:miles2 + asian2:age2 +
      asian2:income2 + asian2:miles2 + pop2:age2 + pop2:miles2 + age2:work2 +
      income2:insured2, offset=log(pop2), family=quasipoisson(link="log"))
summary(fitmargqp)
```

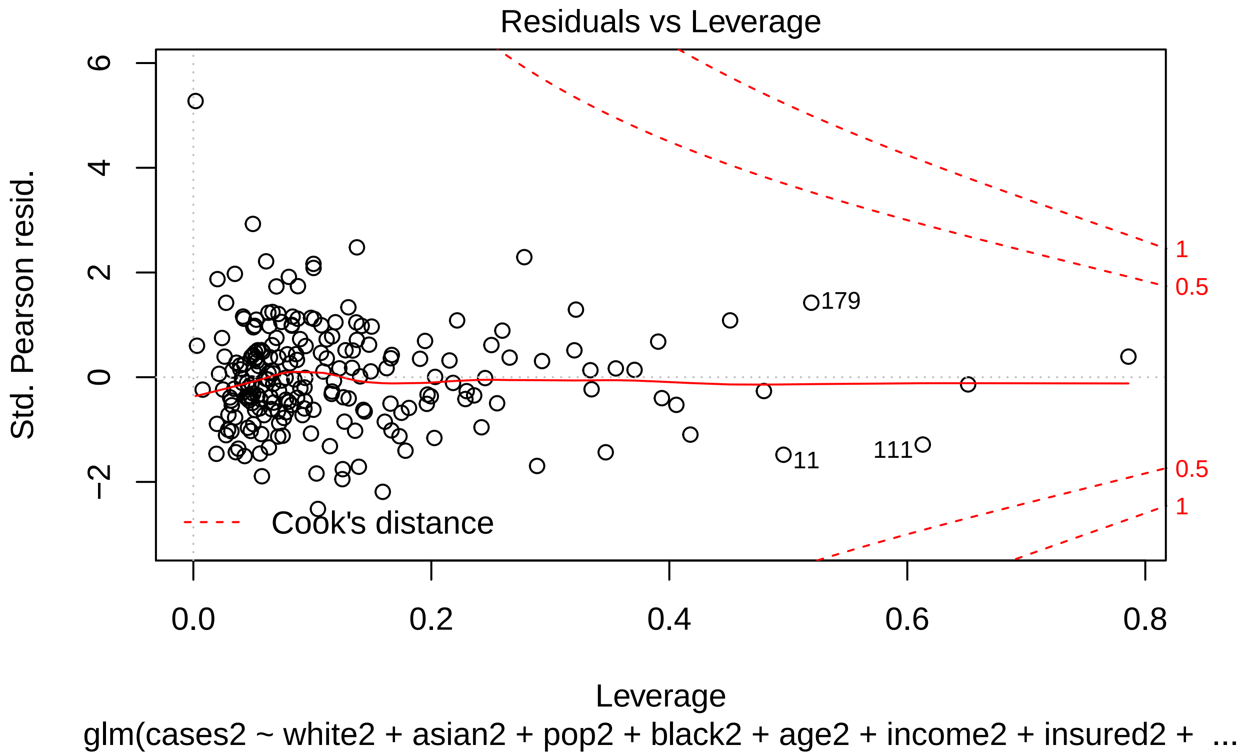
```
##
## Call:
## glm(formula = cases2 ~ white2 + asian2 + pop2 + black2 + age2 +
##     income2 + insured2 + work2 + miles2 + white2:asian2 + white2:pop2 +
##     white2:age2 + white2:income2 + white2:insured2 + white2:miles2 +
##     black2:pop2 + black2:age2 + black2:income2 + black2:miles2 +
##     asian2:age2 + asian2:income2 + asian2:miles2 + pop2:age2 +
##     pop2:miles2 + age2:work2 + income2:insured2, family = quasipoisson(link = "log"),
##     offset = log(pop2))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6529  -0.7757  -0.0593   0.7037   3.7871
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -9.785e+00  1.137e+00  -8.603 2.95e-15 ***
## white2            5.317e-03  1.502e-02   0.354 0.723659
## asian2            1.219e-02  7.764e-03   1.569 0.118211    
## pop2              5.653e-06  4.080e-06   1.386 0.167517    
## black2            2.724e-02  8.288e-03   3.287 0.001206 ** 
## age2              1.004e-01  4.859e-02   2.066 0.040201 *  
## income2           5.021e-05  1.035e-05   4.850 2.57e-06 ***
## insured2          2.308e-02  1.478e-02   1.562 0.120050    
## work2            -1.185e-02  3.506e-03  -3.379 0.000884 ***
## miles2           -3.457e-02  1.238e-02  -2.792 0.005784 ** 
## white2:asian2     2.033e-04  7.818e-05   2.600 0.010051 *  
## white2:pop2      -1.603e-07  5.173e-08  -3.099 0.002241 ** 
## white2:age2      -1.277e-03  4.645e-04  -2.749 0.006552 ** 
## white2:income2   -1.312e-07  9.418e-08  -1.393 0.165263    
## white2:insured2   2.262e-04  1.738e-04   1.301 0.194755    
## white2:miles2     2.903e-04  1.208e-04   2.403 0.017248 *  
## pop2:black2      -1.072e-07  5.047e-08  -2.124 0.034936 *  
## black2:age2      -1.264e-03  4.740e-04  -2.666 0.008335 ** 
## black2:income2   -1.410e-07  9.050e-08  -1.558 0.120986    
## black2:miles2     3.440e-04  1.434e-04   2.400 0.017381 *  
## asian2:age2      -1.448e-03  5.045e-04  -2.870 0.004569 ** 
## asian2:income2   -1.785e-07  9.776e-08  -1.826 0.069477 .  
## asian2:miles2     5.769e-04  1.486e-04   3.883 0.000143 ***
## pop2:age2         2.905e-07  1.249e-07   2.327 0.021045 *  
## pop2:miles2       1.266e-07  6.933e-08   1.826 0.069407 .  
## age2:work2        8.552e-04  2.146e-04   3.986 9.61e-05 ***
## income2:insured2 -3.651e-07  1.234e-07  -2.959 0.003481 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for quasipoisson family taken to be 2.022957)
## 
##     Null deviance: 3498.21  on 215  degrees of freedom
## Residual deviance:  335.33  on 189  degrees of freedom
## AIC: NA
## 
## Number of Fisher Scoring iterations: 4
```

```
#Most of the added main effects are not significant, but some are
#Diagnostics
plot(fitmargqp) #plots still look good and we now follow marginality principle
```


Residuals vs Fitted
Residuals
4
2
0
−2
−4
−2
0
2
4
6
216
167
74
Predicted values
glm(cases2 ~ white2 + asian2 + pop2 + black2 + age2 + income2 + insured2 + ...


Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(cases2 ~ white2 + asian2 + pop2 + black2 + age2 + income2 + insured2 + ...
167
216
74

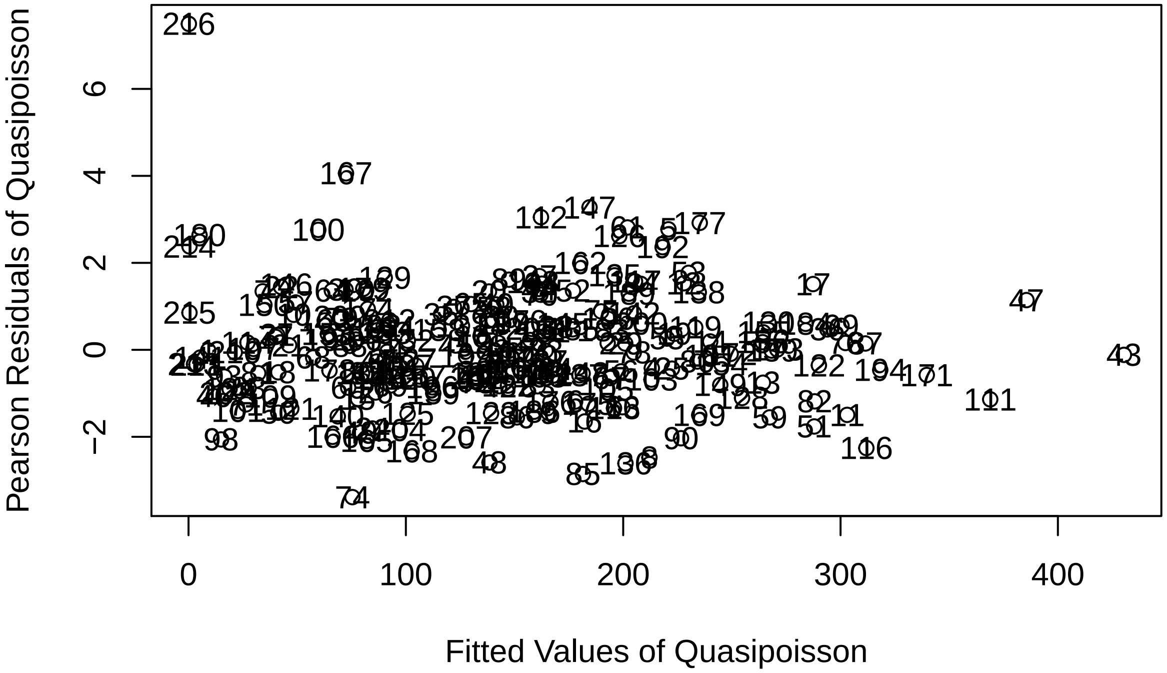

Scale−Location
216
167
74
$\sqrt{|\text{Std. deviance resid.}|}$
0.0
0.5
1.0
1.5
−2
0
2
4
6
Predicted values
glm(cases2 ~ white2 + asian2 + pop2 + black2 + age2 + income2 + insured2 + ...



```
#DOES this actually help, or does it just allow other points to be new outliers...
n <- 216
outlierTest(fitmargqp, cutoff = 0.05, n.max = n, order = TRUE)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 167 2.988234          0.0028059      0.60608
```

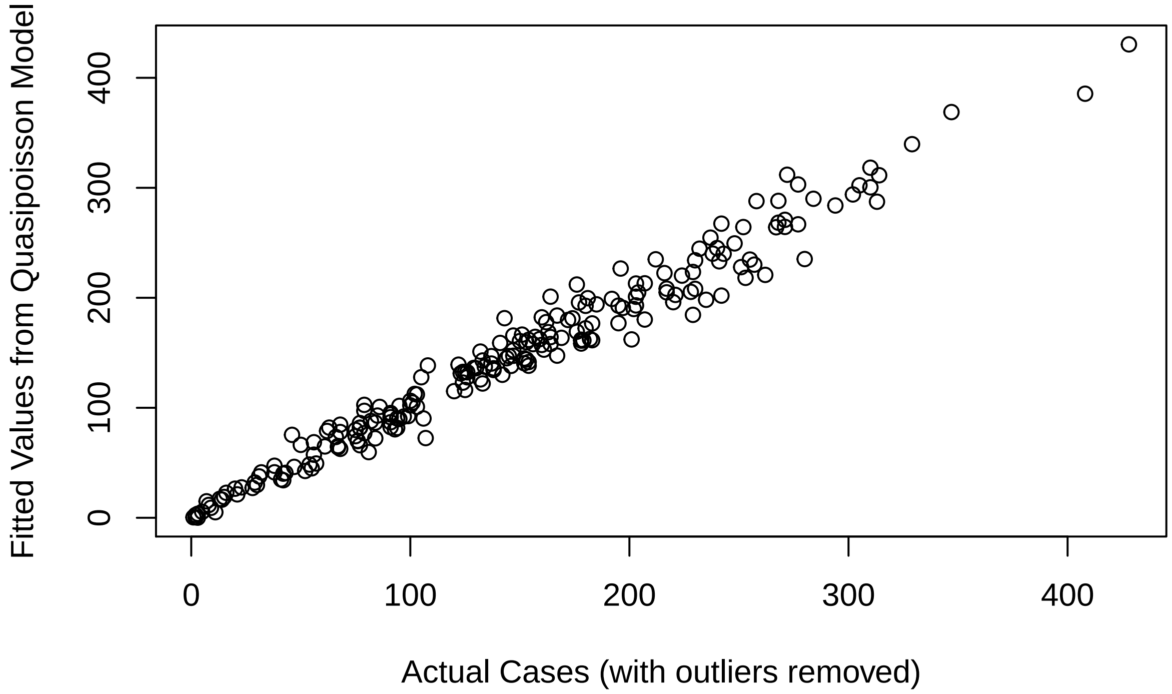
```
#NO, new outliers are NOT introduced, therefore removing those cases for this model
#helps improve the fitted model

pearsresidmargqp <- residuals(fitmargqp, type="pearson")
par(mfrow=c(1,1))
plot(pearsresidmargqp~fitted(fitmargqp), xlab="Fitted Values of Quasipoisson", ylab="Pearson Residuals 
#do this to identify case number
text(fitted(fitmargqp), pearsresidmargqp) #highlights 216
```



```
## BUT we see that pearson residuals DO NOT fan, so we are not worried about
# residvfitted plot here

hoslem.test(cases2, fitted(fitmargqp))
```

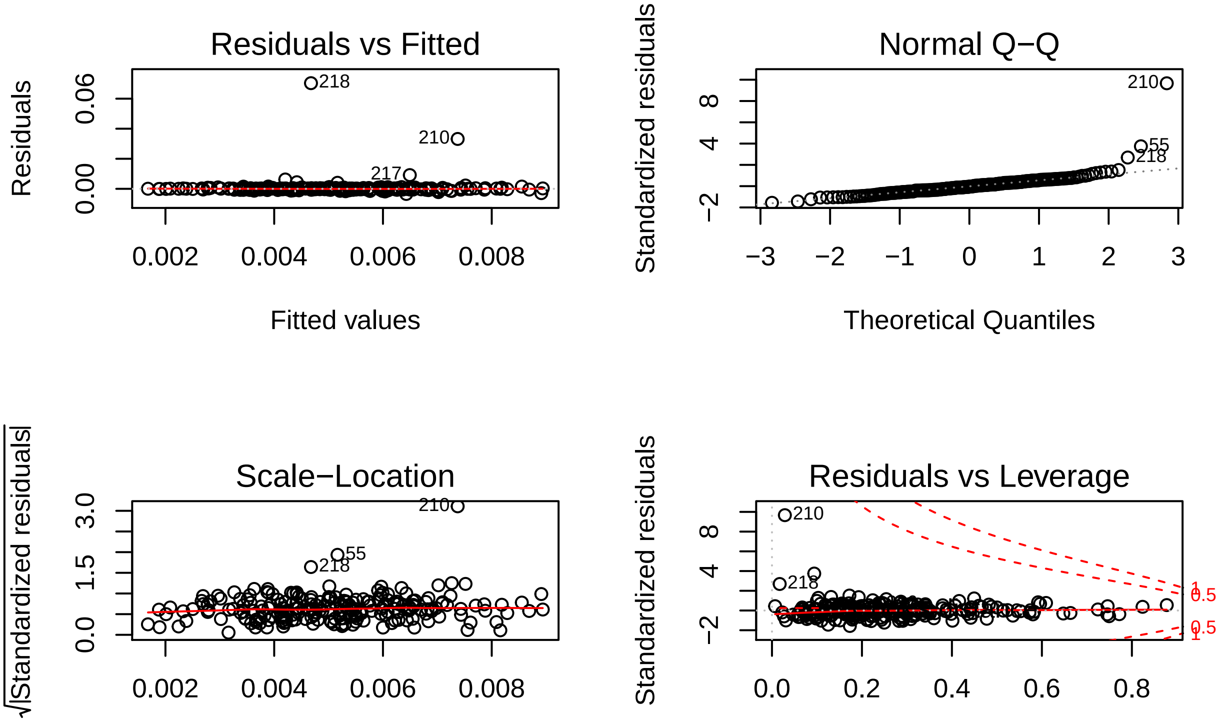
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cases2, fitted(fitmargqp)
## X-squared = -0.13944, df = 8, p-value = 1
```

```
#The p-value is very high which tells us we DO NOT have a significant difference
# between the actuals and the fitted (we can see that in the plot below)
plot(fitted(fitmargqp)~cases2, xlab="Actual Cases (with outliers removed)", ylab="Fitted Values from Qua
```



```
#This indicates a good fit for the model; the actuals and fitted line up diagonally
#Adding back the main effects DOES not ruin this, and we still see an excellent fit

#UTILIZE THE marginal model with outliers removed!
```

```
#3. Fit RATE model with all cases - full glm
##a. Model
fitrate <- lm((cases/pop) ~ (white+black+asian+pop+age+income+smoke+insured+work+miles)^2, weights=(pop
summary(fitrate)
#weighted due to non constant variance
##b. Diagnostics
par(mfrow = c(2,2))
plot(fitrate)
```

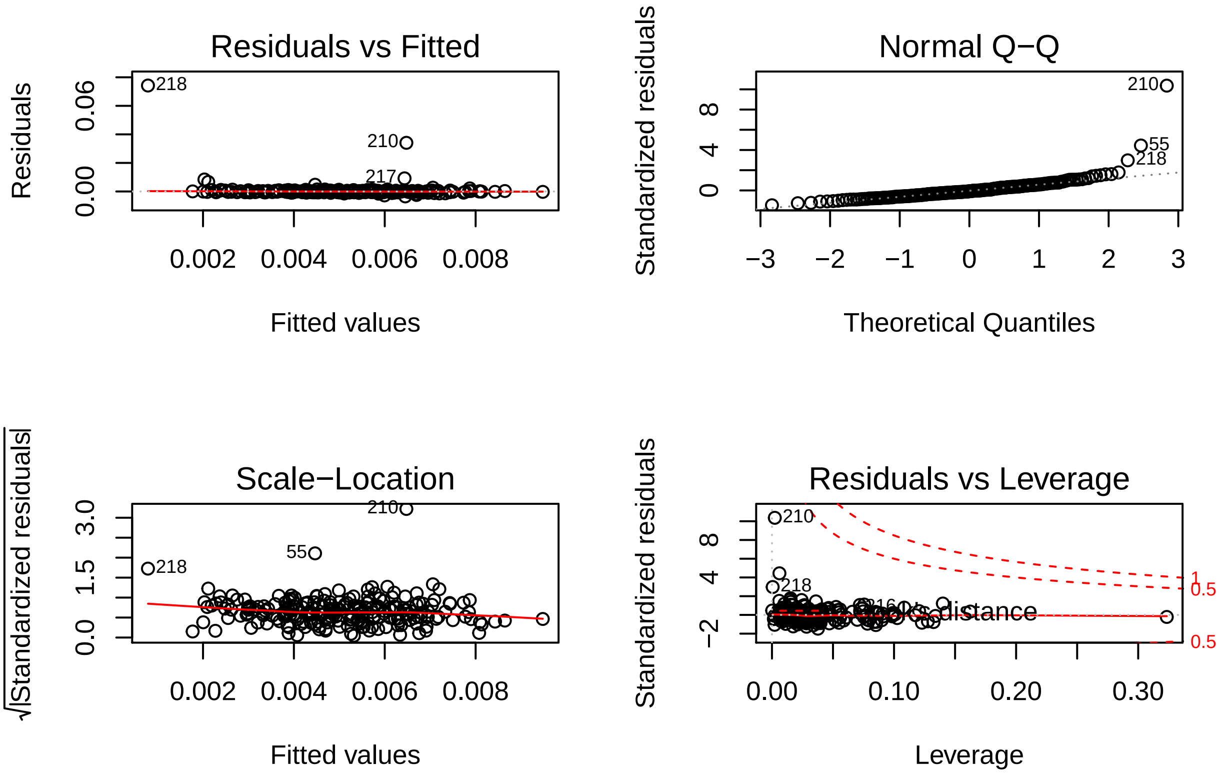


```r
par(mfrow = c(1,1))
##c. Backwards Elimination
full <- fitrate
null <- lm(cases ~ 1) #null is just the response with intercept
s3 <- step(full, scope=list(lower=null, upper=full), direction="backward", k=4)
summary(s3)
```

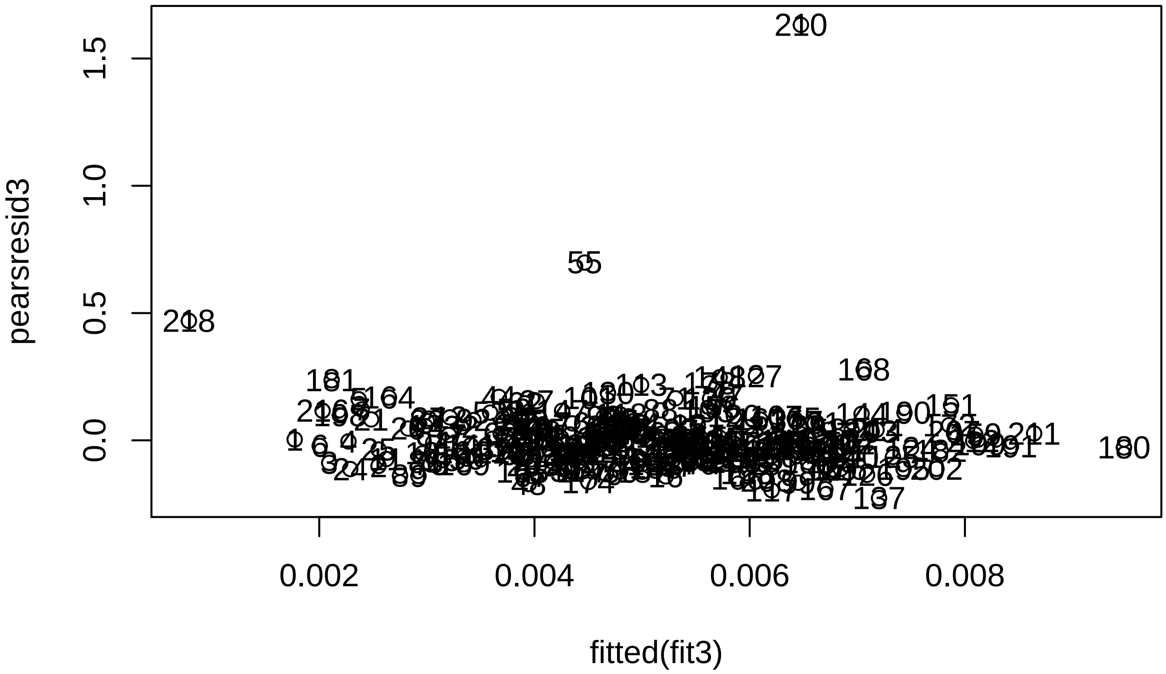
```r
##Removed A LOT of variables, resulted in:
#black + age + income + smoke + insured + work + age:work

# Fit RATE with BE identified terms
fit3 <- lm((cases/pop) ~ black+age+income+smoke+insured+work+age:work, weights=pop)
summary(fit3) #matches the summary for BE as it should
```

```
##
## Call:
## lm(formula = (cases/pop) ~ black + age + income + smoke + insured +
##     work + age:work, weights = pop)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22688 -0.06906 -0.01292  0.05697  1.63206
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.583e-03  1.550e-03  -1.666 0.097105 .
## black        1.543e-05  3.187e-06   4.842 2.49e-06 ***
## age          -5.457e-05  8.410e-05  -0.649 0.517132
## income        1.372e-08  2.954e-09   4.646 5.97e-06 ***
## smoke         6.663e-05  2.806e-05   2.374 0.018475 *
## insured       5.328e-05  1.952e-05   2.729 0.006884 **
## work         -5.087e-05  2.038e-05  -2.496 0.013328 *
## age:work      4.961e-06  1.439e-06   3.447 0.000685 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1576 on 210 degrees of freedom
## Multiple R-squared:  0.7321, Adjusted R-squared:  0.7231
## F-statistic: 81.97 on 7 and 210 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(fit3)
```



```r
#Diagnostic plots still look really good, the key now is to see why there are points
# that are standing out

pearsresid3 <- residuals(fit3, type="pearson")
par(mfrow=c(1,1))
plot(pearsresid3~fitted(fit3))
#do this to identify case number
text(fitted(fit3), pearsresid3)
```



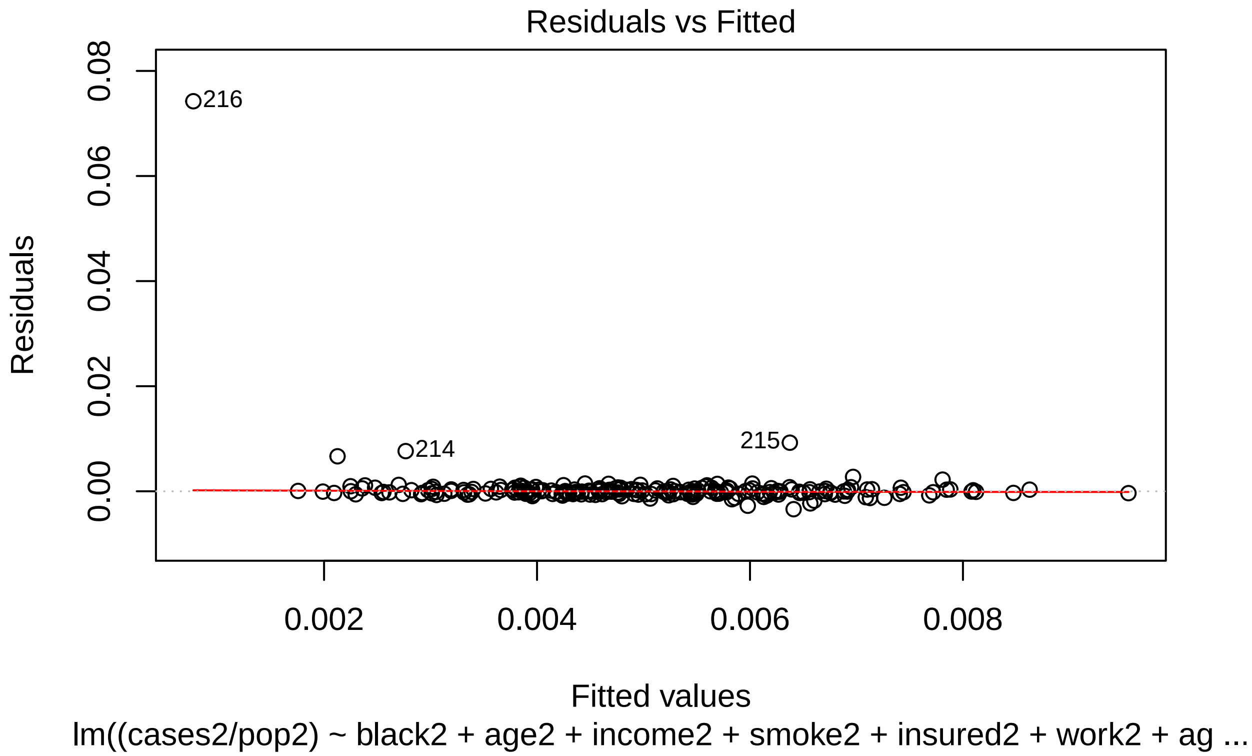
```
#210 stands out... look at outlier testing

# Fit3 Outliers and Influential Points
library(car)
n <- 218
outlierTest(fit3, cutoff = 0.05, n.max = n, order = TRUE)
```

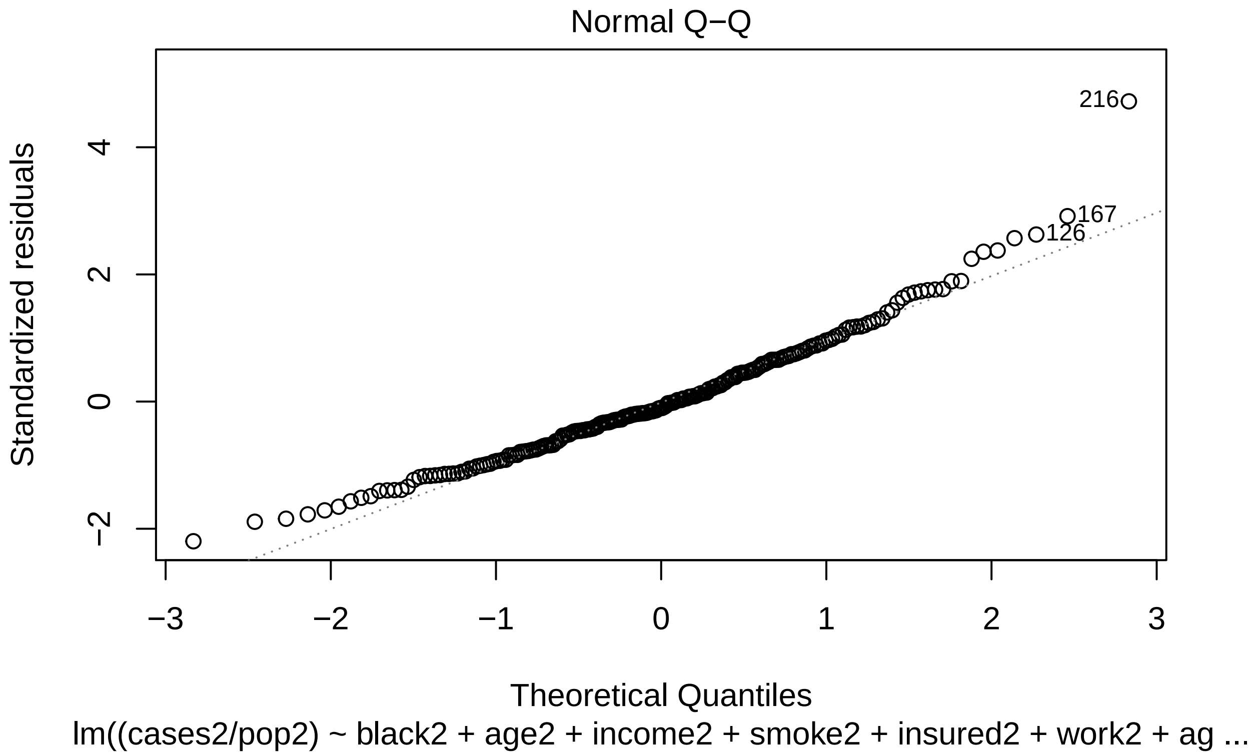
```
##      rstudent unadjusted p-value Bonferonni p
## 210 14.797267         2.2839e-34   4.9789e-32
## 55   4.657727         5.6837e-06   1.2390e-03
```

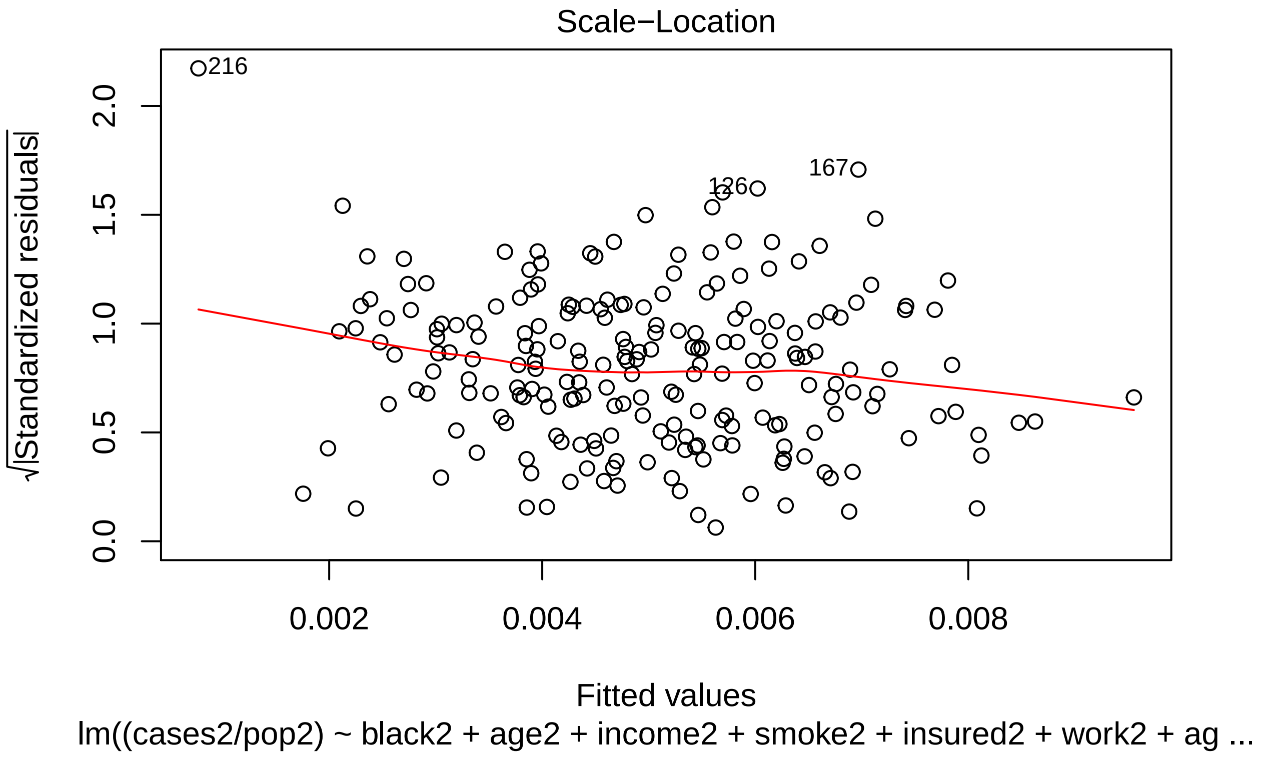
```
#gave us cases 210, 55, and 218 was JUST shy of the 0.05 mark for Bonferonni

# ANALYSIS FOR RATE: removal of outliers
#remove cases that are outliers; indices 210 and 55
#USE CANCER2 AGAIN - same outliers as before

fitoutrate <- lm((cases2/pop2) ~ black2+age2+income2+smoke2+insured2+work2+age2:work2, weights=pop2)
summary(fitoutrate)
```

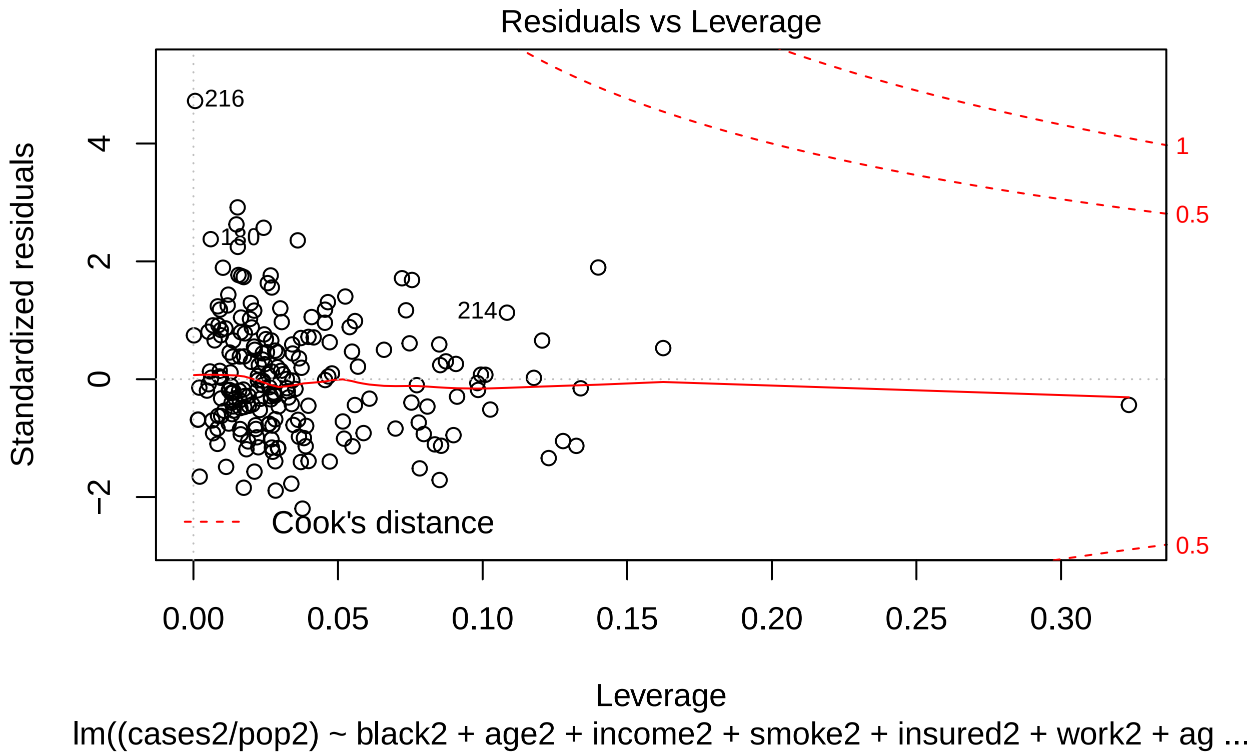
```
##
## Call:
## lm(formula = (cases2/pop2) ~ black2 + age2 + income2 + smoke2 +
##     insured2 + work2 + age2:work2, weights = pop2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21429 -0.06751 -0.00979  0.06203  0.46946
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.653e-03  9.782e-04  -2.712 0.007239 **
## black2       1.596e-05  2.011e-06   7.933 1.30e-13 ***
## age2        -3.729e-05  5.307e-05  -0.703 0.483029
## income2      1.382e-08  1.864e-09   7.417 2.99e-12 ***
## smoke2       6.012e-05  1.772e-05   3.393 0.000827 ***
## insured2     5.259e-05  1.232e-05   4.269 2.98e-05 ***
## work2       -4.706e-05  1.286e-05  -3.660 0.000320 ***
## age2:work2   4.628e-06  9.083e-07   5.095 7.80e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09945 on 208 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8677
## F-statistic: 202.5 on 7 and 208 DF,  p-value: < 2.2e-16
```

```
plot(fitoutrate)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
216
167
126
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm((cases2/pop2) ~ black2 + age2 + income2 + smoke2 + insured2 + work2 + ag ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
216
126
167
0.002
0.004
0.006
0.008
Fitted values
lm((cases2/pop2) ~ black2 + age2 + income2 + smoke2 + insured2 + work2 + ag ...



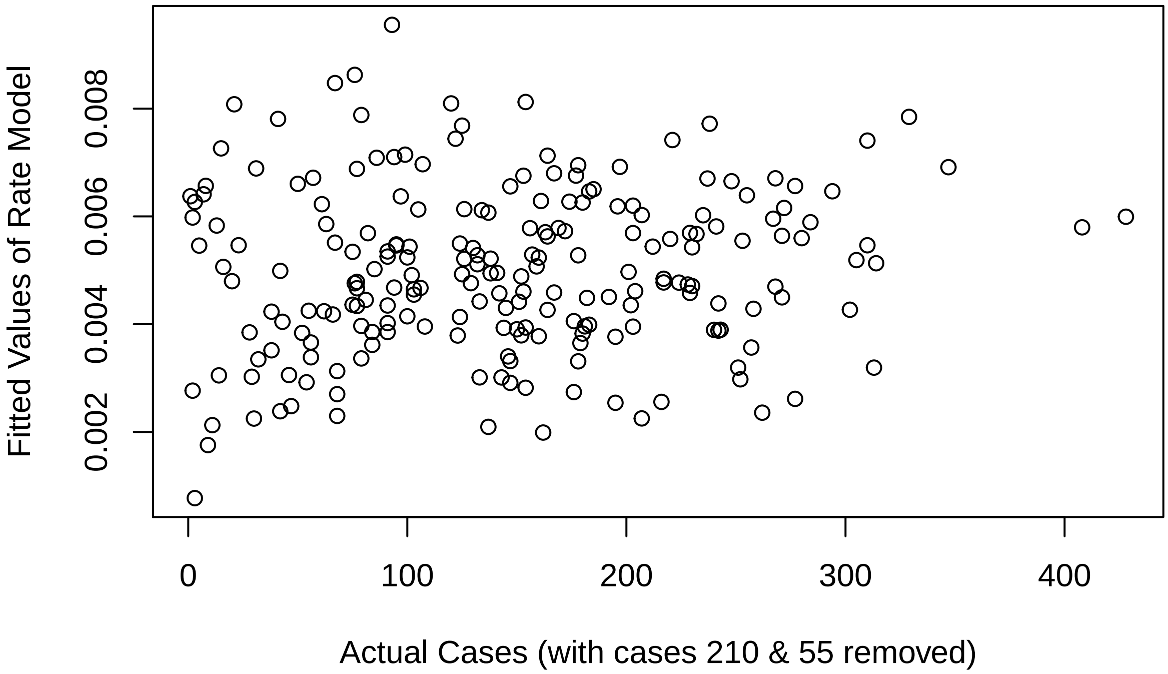
```
#DOES this actually help, or does it just allow other points to be new outliers...
n <- 216
outlierTest(fitoutrate, cutoff = 0.05, n.max = n, order = TRUE)
```

```
##     rstudent unadjusted p-value Bonferonni p
## 216 4.985418          1.305e-06   0.00028188
```

```
# Case 216 now looks to be an outlier, but overall the model seems to
## fit more normally as well as the residuals v fitted plot looks good
# (tried removing 216 and 215 became an outlier... may not be worth removing more
# than the two main identified points from before)

hoslem.test(cases2, fitted(fitoutrate))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  cases2, fitted(fitoutrate)
## X-squared = 977490000, df = 8, p-value < 2.2e-16
```

```
#The p-value is low which tells us we have a significant difference
# between the actuals and the fitted (we can see that in the plot below)
plot(fitted(fitoutrate)~cases2, xlab="Actual Cases (with cases 210 & 55 removed)", ylab="Fitted Values 
```



```
#Since nothing else is an indicator for misfit
# We chalk it up to the fact that there are definitely missing predictors here
# therefore we cannot match the actual values as well as we'd like
```

I am planning to do interpretation of the quasibinomial and quasipoisson models with outliers REMOVED. For the linear model of rate, I am planning to interpret the model fitoutrate where just cases 210 and 55 are removed, even though 216 is still a possible outlier too – this is because the plots improved a bit (removed the far out cases on right tail).