FMPH 222 Final Project

Daniel Zoleikhaeian

2023-01-29

Chi Square on Expected Dropouts

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##     recode
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
dropout_rate_overall <- sum(df$Statusgp == 2) / nrow(df)

by_group_dropout <- df %>%
  dplyr::group_by(Randomization,Statusgp) %>%
  dplyr::summarise(total_count=n(), .groups = 'drop')

dropouts <- by_group_dropout[by_group_dropout$Statusgp == 2, ]

total_cts <- as.data.frame(plyr::count(df, 'Randomization'))
```

```r
expected_drops <- total_cts$freq * dropout_rate_overall

chi_square_test_stat <- sum(((dropouts$total_count - expected_drops)^2) / expected_drops)

pchisq(chi_square_test_stat, df = 3, lower.tail = F)
```

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

Initial Model

```r
#View(df$Gendergp)
mod1 <- glm(formula = Statusgp ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes + Randomization, fa
summary(mod1)
```

```
##
## Call:
## glm(formula = Statusgp ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes +
##     Randomization, family = binomial, data = df)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.269759   1.682361  -1.349   0.1773
## Age                  -0.006342   0.021645  -0.293   0.7695
## Educationyears        0.016759   0.040452   0.414   0.6787
## Gendergp1             0.753227   0.305159   2.468   0.0136 *
## b_cdr_sumofboxes      0.516099   0.218911   2.358   0.0184 *
## RandomizationCog      0.112969   0.369153   0.306   0.7596
## RandomizationCog Phy  0.129400   0.374144   0.346   0.7295
## RandomizationPhy     -0.272612   0.392334  -0.695   0.4872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 406.54  on 545  degrees of freedom
## Residual deviance: 393.91  on 538  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 409.91
##
## Number of Fisher Scoring iterations: 5
```

```r
# new model with re-encoded randomization
mod2 <- glm(formula = Statusgp ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes + isPhy, family = bi
summary(mod2)
```

```
##
## Call:
## glm(formula = Statusgp ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes +
##     isPhy, family = binomial, data = df)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.174138   1.654715  -1.314   0.1889
## Age              -0.006438   0.021587  -0.298   0.7655
```

```
## Educationyears    0.016164   0.040376   0.400   0.6889
## Gendergp1         0.753048   0.304970   2.469   0.0135 *
## b_cdr_sumofboxes  0.513243   0.218418   2.350   0.0188 *
## isPhy            -0.355609   0.320627  -1.109   0.2674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 406.54  on 545  degrees of freedom
## Residual deviance: 394.05  on 540  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 406.05
##
## Number of Fisher Scoring iterations: 5
```

Aggregate initial model

```
w <- aggregate(Statusgp ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes + Randomization, data = df
n <- aggregate(Statusgp ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes + Randomization, data = df

w.n <- data.frame(Age = w$Age,
                  Educationyears = w$Educationyears,
                  Gendergp = w$Gendergp,
                  b_cdr_sumofboxes = w$b_cdr_sumofboxes,
                  Randomization = w$Randomization,
                  Statusgp = w$Statusgp,
                  subj = n$Statusgp)

# View(w.n)

mod.ag <- glm(formula = Statusgp/subj ~ Age + Educationyears + Gendergp + b_cdr_sumofboxes + Randomizati
summary(mod.ag)
```

```
##
## Call:
## glm(formula = Statusgp/subj ~ Age + Educationyears + Gendergp +
##     b_cdr_sumofboxes + Randomization, family = binomial, data = w.n,
##     weights = subj)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.269759   1.682361  -1.349   0.1773
## Age                  -0.006342   0.021645  -0.293   0.7695
## Educationyears        0.016759   0.040452   0.414   0.6787
## Gendergp1             0.753227   0.305159   2.468   0.0136 *
## b_cdr_sumofboxes      0.516099   0.218911   2.358   0.0184 *
## RandomizationCog      0.112969   0.369153   0.306   0.7596
## RandomizationCog Phy  0.129400   0.374144   0.346   0.7295
## RandomizationPhy     -0.272612   0.392334  -0.695   0.4872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 388.86  on 513  degrees of freedom
## Residual deviance: 376.22  on 506  degrees of freedom
## AIC: 400.78
##
## Number of Fisher Scoring iterations: 5
```

```
# Perhaps do backwards selection on the full model
# So far, only b_cdr_sob and gender are important
```

Backwards selection on full model (initial)

```
library(logistf)
```

```
## Warning: package 'logistf' was built under R version 4.3.1
```

```
f_mod <- logistf(formula = Statusgp ~ factor(NGOgroup) + Randomization + Age + Educationyears + Genderg
summary(f_mod)
```

```
## logistf(formula = Statusgp ~ factor(NGOgroup) + Randomization +
##     Age + Educationyears + Gendergp + b_cdr_sumofboxes, data = df)
##
## Model fitted by Penalized ML
## Coefficients:
##                              coef   se(coef)  lower 0.95  upper 0.95
## (Intercept)          -3.929439004 1.68884082 -7.34589269 -0.57871223
## factor(NGOgroup)2     2.031164853 0.46350978  1.16364254  3.04390911
## factor(NGOgroup)3     1.450029838 0.45164440  0.60113638  2.43655932
## RandomizationCog      0.149433742 0.36390995 -0.57156512  0.88147838
## RandomizationCog Phy  0.761092373 0.39559296 -0.02015329  1.56025523
## RandomizationPhy     -0.080620513 0.39259567 -0.87069705  0.69989262
## Age                  -0.003611487 0.02094122 -0.04564458  0.03826081
## Educationyears        0.010640107 0.04008251 -0.07158573  0.08917366
## Gendergp1             0.634252170 0.30170791  0.02379252  1.22827970
## b_cdr_sumofboxes      0.467892761 0.21538521  0.03832130  0.89996231
##                            Chisq            p method
## (Intercept)           5.29194259 2.142431e-02      2
## factor(NGOgroup)2    24.32936125 8.119200e-07      2
## factor(NGOgroup)3    11.99528735 5.333525e-04      2
## RandomizationCog      0.16504090 6.845573e-01      2
## RandomizationCog Phy  3.64579862 5.621084e-02      2
## RandomizationPhy      0.04111783 8.393106e-01      2
## Age                   0.02861303 8.656756e-01      2
## Educationyears        0.06732811 7.952675e-01      2
## Gendergp1             4.13984260 4.188506e-02      2
## b_cdr_sumofboxes      4.55501494 3.282248e-02      2
##
## Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None
##
## Likelihood ratio test=37.22076 on 9 df, p=2.403476e-05, n=546
## Wald test = 191.9085 on 9 df, p = 0
```

```
mod_b <- backward(f_mod, slstay= 0.2)
```

```
## Step  0 : starting model
## Step  1 : removed  Age  (P= 0.8656756 )
```

```
## Step  2 : removed  Educationyears  (P= 0.7666313 )
```

```
summary(mod_b)
```

```
## logistf(formula = Statusgp ~ factor(NGOgroup) + Randomization +
##     Gendergp + b_cdr_sumofboxes, data = df)
##
## Model fitted by Penalized ML
## Coefficients:
##                             coef  se(coef)  lower 0.95 upper 0.95       Chisq
## (Intercept)          -4.16981306 0.5581646 -5.35988216 -3.1257549         Inf
## factor(NGOgroup)2     2.04824240 0.4656276  1.18171352  3.0605140 24.88584326
## factor(NGOgroup)3     1.45248719 0.4539227  0.60388642  2.4386926 12.04671286
## RandomizationCog      0.15707336 0.3651171 -0.56347601  0.8887784  0.18256482
## RandomizationCog Phy  0.76267710 0.3969818 -0.01872565  1.5619072  3.65953502
## RandomizationPhy     -0.08132722 0.3949804 -0.87304489  0.7007702  0.04166787
## Gendergp1             0.66347333 0.2901903  0.07706625  1.2314932  4.88645591
## b_cdr_sumofboxes      0.44646476 0.2047543  0.04050750  0.8569532  4.64596704
##                                  p method
## (Intercept)          0.000000e+00      2
## factor(NGOgroup)2    6.082747e-07      2
## factor(NGOgroup)3    5.188379e-04      2
## RandomizationCog     6.691783e-01      2
## RandomizationCog Phy 5.574919e-02      2
## RandomizationPhy     8.382542e-01      2
## Gendergp1            2.706820e-02      2
## b_cdr_sumofboxes     3.112659e-02      2
##
## Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None
##
## Likelihood ratio test=37.40736 on 7 df, p=3.926573e-06, n=547
## Wald test = 191.8422 on 7 df, p = 0
```

```
## backwards selection 2
rel_vars <- c('NGOgroup', 'Randomization', 'Statusgp', 'Age', 'Gendergp', 'Educationyears', 'b_Adas_Dela

df_red2 <- df[, c(rel_vars)]

df_red2 <- df_red2[complete.cases(df_red2), ]

f_mod2 <- logistf(formula = Statusgp ~ NGOgroup + Randomization + Age + Gendergp + Educationyears +
b_Adas_DelayedRecall + b_Adas_Total + b_CMMSE + b_CVFT_Total + b_cdr_sumofboxes +
b_Cornell_Total + b_CIRS_Total + b_CNPI_Total + b_MIC_Total, data = df_red2)

summary(f_mod2)
```

```
## logistf(formula = Statusgp ~ NGOgroup + Randomization + Age +
##     Gendergp + Educationyears + b_Adas_DelayedRecall + b_Adas_Total +
##     b_CMMSE + b_CVFT_Total + b_cdr_sumofboxes + b_Cornell_Total +
##     b_CIRS_Total + b_CNPI_Total + b_MIC_Total, data = df_red2)
##
## Model fitted by Penalized ML
## Coefficients:
##                              coef    se(coef)   lower 0.95   upper 0.95
## (Intercept)          -1.643241998 2.730626342 -7.208173712  3.877181124
```

```
## NGOgroup2              2.235125274 0.488139593  1.303128402  3.335514197
## NGOgroup3              1.667117855 0.471254962  0.765799894  2.726915214
## RandomizationCog       0.173510417 0.368782062 -0.567508229  0.926749775
## RandomizationCog Phy   0.822033100 0.408234597  0.006386015  1.660514793
## RandomizationPhy      -0.058781134 0.402562329 -0.881105168  0.752795265
## Age                   -0.006749990 0.021572870 -0.050580362  0.037030487
## Gendergp1              0.608367459 0.316064061 -0.038538969  1.239758357
## Educationyears         0.019683175 0.041828169 -0.066918926  0.102864376
## b_Adas_DelayedRecall -0.153215601 0.072113471 -0.301403409 -0.008545433
## b_Adas_Total           0.027493686 0.053079858 -0.081035035  0.134275457
## b_CMMSE               -0.056958825 0.072911474 -0.205236452  0.090683516
## b_CVFT_Total          -0.006248923 0.020813283 -0.049100673  0.035664199
## b_cdr_sumofboxes      -0.036643257 0.305591357 -0.666502080  0.573616209
## b_Cornell_Total        0.138505688 0.078420412 -0.021544056  0.302722301
## b_CIRS_Total           0.007103821 0.005735844 -0.005243857  0.018109087
## b_CNPI_Total           0.032271066 0.043227026 -0.059704470  0.117806587
## b_MIC_Total           -0.024264040 0.034255631 -0.094673910  0.044672527
##                             Chisq            p method
## (Intercept)            0.33995440 5.598555e-01      2
## NGOgroup2             26.21055213 3.061451e-07      2
## NGOgroup3             14.38171645 1.492444e-04      2
## RandomizationCog       0.21058478 6.463094e-01      2
## RandomizationCog Phy   3.90200670 4.822847e-02      2
## RandomizationPhy       0.02019198 8.870022e-01      2
## Age                    0.09163923 7.621035e-01      2
## Gendergp1              3.40552281 6.497852e-02      2
## Educationyears         0.20627739 6.497009e-01      2
## b_Adas_DelayedRecall   4.31257554 3.783164e-02      2
## b_Adas_Total           0.25134438 6.161300e-01      2
## b_CMMSE                0.57326641 4.489638e-01      2
## b_CVFT_Total           0.08427178 7.715898e-01      2
## b_cdr_sumofboxes       0.01352047 9.074326e-01      2
## b_Cornell_Total        2.86392999 9.058593e-02      2
## b_CIRS_Total           1.35717062 2.440284e-01      2
## b_CNPI_Total           0.50041768 4.793167e-01      2
## b_MIC_Total            0.47201621 4.920611e-01      2
##
## Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None
##
## Likelihood ratio test=57.0238 on 17 df, p=3.228198e-06, n=542
## Wald test = 183.2166 on 17 df, p = 0
```

```
mod_b2 <- logistf::backward(f_mod2, slstay= 0.2)
```

```
## Step  0 : starting model
## Step  1 : removed  b_cdr_sumofboxes  (P= 0.9074326 )
## Step  2 : removed  b_CVFT_Total  (P= 0.772004 )
## Step  3 : removed  Age  (P= 0.7765024 )
## Step  4 : removed  b_Adas_Total  (P= 0.5945421 )
## Step  5 : removed  Educationyears  (P= 0.6363126 )
## Step  6 : removed  b_CNPI_Total  (P= 0.4734064 )
## Step  7 : removed  b_MIC_Total  (P= 0.5042037 )
## Step  8 : removed  b_CMMSE  (P= 0.3625514 )
## Step  9 : removed  b_CIRS_Total  (P= 0.3177327 )
```

```
summary(mod_b2)
```

```
## logistf(formula = Statusgp ~ NGOgroup + Randomization + Gendergp +
##     b_Adas_DelayedRecall + b_Cornell_Total, data = df_red2)
##
## Model fitted by Penalized ML
## Coefficients:
##                             coef   se(coef)  lower 0.95  upper 0.95       Chisq
## (Intercept)          -3.5054975 0.56276132 -4.71643226 -2.45287355 55.33395821
## NGOgroup2             2.3528603 0.49963097  1.42431828  3.44690966 29.69693359
## NGOgroup3             1.6181203 0.47691713  0.72864669  2.66257117 13.85776937
## RandomizationCog      0.1646401 0.37456567 -0.57599849  0.91596649  0.19005355
## RandomizationCog Phy  0.8397045 0.40945354  0.03393844  1.66679269  4.17249763
## RandomizationPhy     -0.1275276 0.40572117 -0.94370163  0.67574107  0.09685072
## Gendergp1             0.6345574 0.29632491  0.03410275  1.21505354  4.27802821
## b_Adas_DelayedRecall -0.1838506 0.06138303 -0.30873238 -0.06352682  9.08096621
## b_Cornell_Total       0.1891988 0.04942350  0.08831178  0.29173981 13.03842505
##                                    p method
## (Intercept)          1.016964e-13      2
## NGOgroup2            5.051466e-08      2
## NGOgroup3            1.971798e-04      2
## RandomizationCog     6.628721e-01      2
## RandomizationCog Phy 4.108519e-02      2
## RandomizationPhy     7.556421e-01      2
## Gendergp1            3.860811e-02      2
## b_Adas_DelayedRecall 2.582836e-03      2
## b_Cornell_Total      3.051646e-04      2
##
## Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None
##
## Likelihood ratio test=54.5959 on 8 df, p=5.289648e-09, n=542
## Wald test = 181.9295 on 8 df, p = 0
```

```
mod2 <- glm(formula = Statusgp ~ factor(NGOgroup) + Gendergp + b_cdr_sumofboxes, family = binomial, data

summary(mod2)
```

```
##
## Call:
## glm(formula = Statusgp ~ factor(NGOgroup) + Gendergp + b_cdr_sumofboxes,
##     family = binomial, data = df)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.9332     0.4823  -8.155 3.50e-16 ***
## factor(NGOgroup)2   1.8418     0.4583   4.019 5.84e-05 ***
## factor(NGOgroup)3   1.4788     0.4683   3.158  0.00159 **
## Gendergp1           0.6669     0.2968   2.247  0.02463 *
## b_cdr_sumofboxes    0.4495     0.2095   2.146  0.03189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 406.80  on 546  degrees of freedom
## Residual deviance: 373.41  on 542  degrees of freedom
## AIC: 383.41
##
## Number of Fisher Scoring iterations: 6
```

```
df$Randomization  <- factor(df$Randomization)

mod3 <- glm(formula = Statusgp ~ Randomization, family = binomial, data = df)
summary(mod3)
```

```
##
## Call:
## glm(formula = Statusgp ~ Randomization, family = binomial, data = df)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.94591    0.26726  -7.281 3.31e-13 ***
## RandomizationCog      0.07007    0.36349   0.193    0.847
## RandomizationCog Phy  0.11778    0.36866   0.319    0.749
## RandomizationPhy     -0.29783    0.38783  -0.768    0.443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 406.80  on 546  degrees of freedom
## Residual deviance: 405.33  on 543  degrees of freedom
## AIC: 413.33
##
## Number of Fisher Scoring iterations: 4
```

Model assessment for Backwards Selection Model

```
mod_b_sel <- glm(formula = Statusgp ~ NGOgroup + Randomization + Gendergp + b_Adas_DelayedRecall + b_Co
summary(mod_b_sel)
```

```
##
## Call:
## glm(formula = Statusgp ~ NGOgroup + Randomization + Gendergp +
##     b_Adas_DelayedRecall + b_Cornell_Total, family = binomial,
##     data = df_red2)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.64732    0.59373  -6.143 8.09e-10 ***
## NGOgroup2             2.47182    0.53089   4.656 3.22e-06 ***
## NGOgroup3             1.71501    0.50685   3.384 0.000715 ***
## RandomizationCog      0.16826    0.38462   0.437 0.661781
## RandomizationCog Phy  0.86138    0.42216   2.040 0.041307 *
## RandomizationPhy     -0.14213    0.41874  -0.339 0.734284
## Gendergp1             0.64175    0.30516   2.103 0.035465 *
## b_Adas_DelayedRecall -0.18995    0.06336  -2.998 0.002719 **
## b_Cornell_Total       0.19578    0.05327   3.675 0.000238 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 349.65  on 533  degrees of freedom
## AIC: 367.65
##
## Number of Fisher Scoring iterations: 6
```

```
## Suggestion that CP is most important

df_red2$RandCP <- as.integer(df_red2$Randomization == 'Cog Phy')

mod_b_sel2 <- glm(formula = Statusgp ~ NGOgroup + factor(RandCP) + Gendergp + b_Adas_DelayedRecall + b_
summary(mod_b_sel2)
```

```
##
## Call:
## glm(formula = Statusgp ~ NGOgroup + factor(RandCP) + Gendergp +
##     b_Adas_DelayedRecall + b_Cornell_Total, family = binomial,
##     data = df_red2)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.63310    0.53300  -6.816 9.34e-12 ***
## NGOgroup2             2.46777    0.52766   4.677 2.91e-06 ***
## NGOgroup3             1.74674    0.50200   3.480 0.000502 ***
## factor(RandCP)1       0.83120    0.35505   2.341 0.019227 *
## Gendergp1             0.63172    0.30463   2.074 0.038103 *
## b_Adas_DelayedRecall -0.18815    0.06319  -2.978 0.002904 **
## b_Cornell_Total       0.19022    0.05233   3.635 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 350.24  on 535  degrees of freedom
## AIC: 364.24
##
## Number of Fisher Scoring iterations: 6
```

```
## Model Assessment

### New aggregations
bins.quantiles(df_red2$b_Adas_DelayedRecall, target.bins = 3, max.breaks = 12)
```

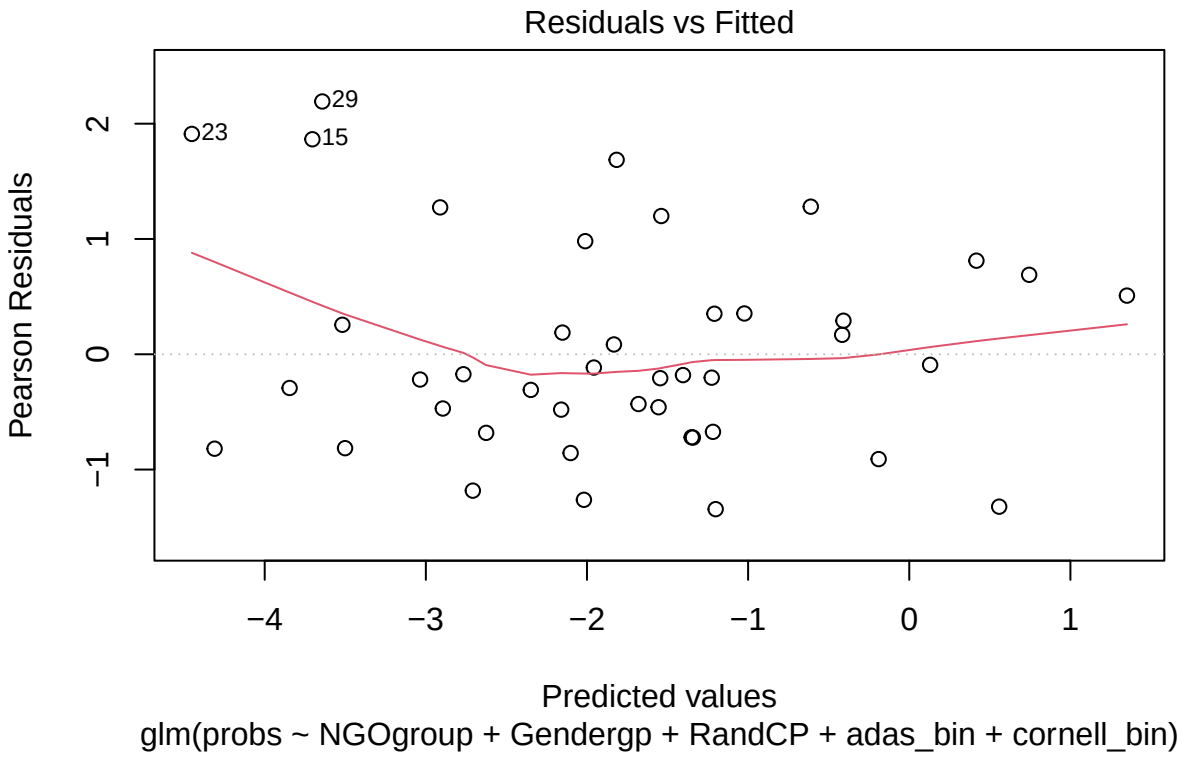
```
## $binlo
## [1] 1 4 7
##
## $binhi
## [1]  3  6 10
##
## $binct
```

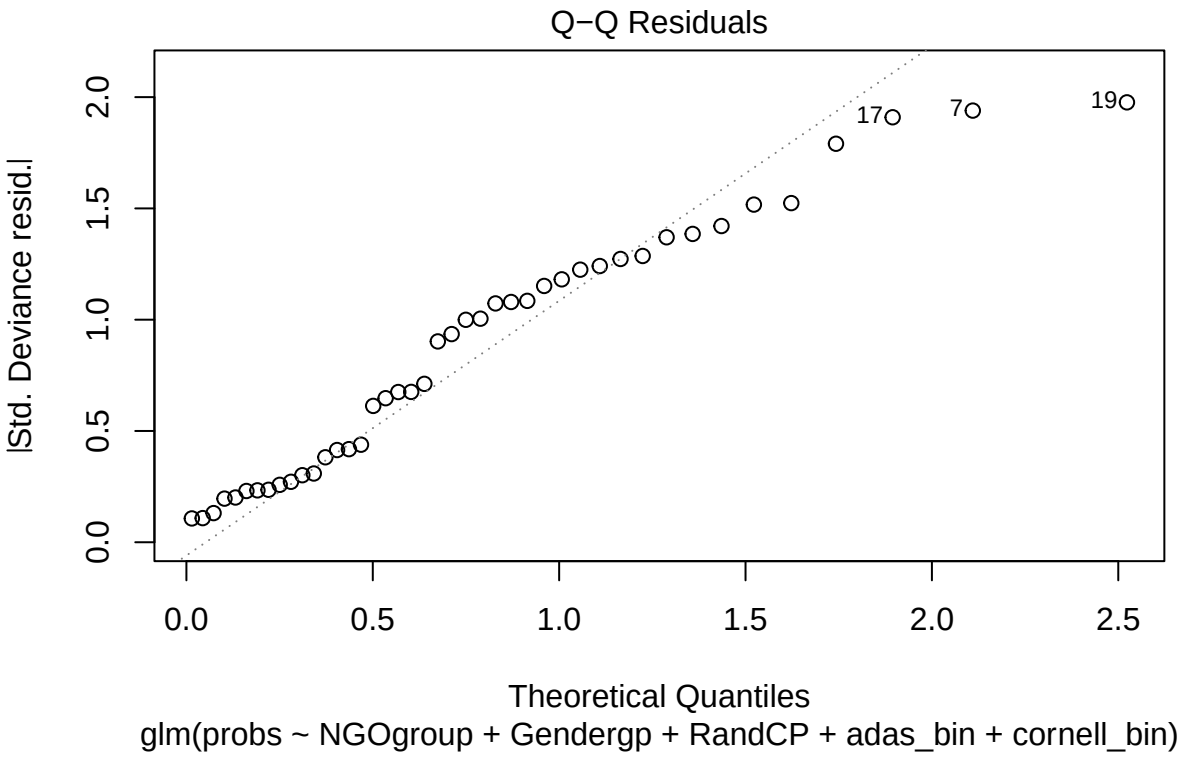
```
## [0, 2] [3, 5] [6, 9]
##    189    249    104
##
## $xtbl
## x
##  0  1  2  3  4  5  6  7  8  9
## 78 37 74 87 84 78 53 38  9  4
##
## $xval
##  [1] 0 1 2 3 4 5 6 7 8 9
##
## $err
## [1] 59.48856
```

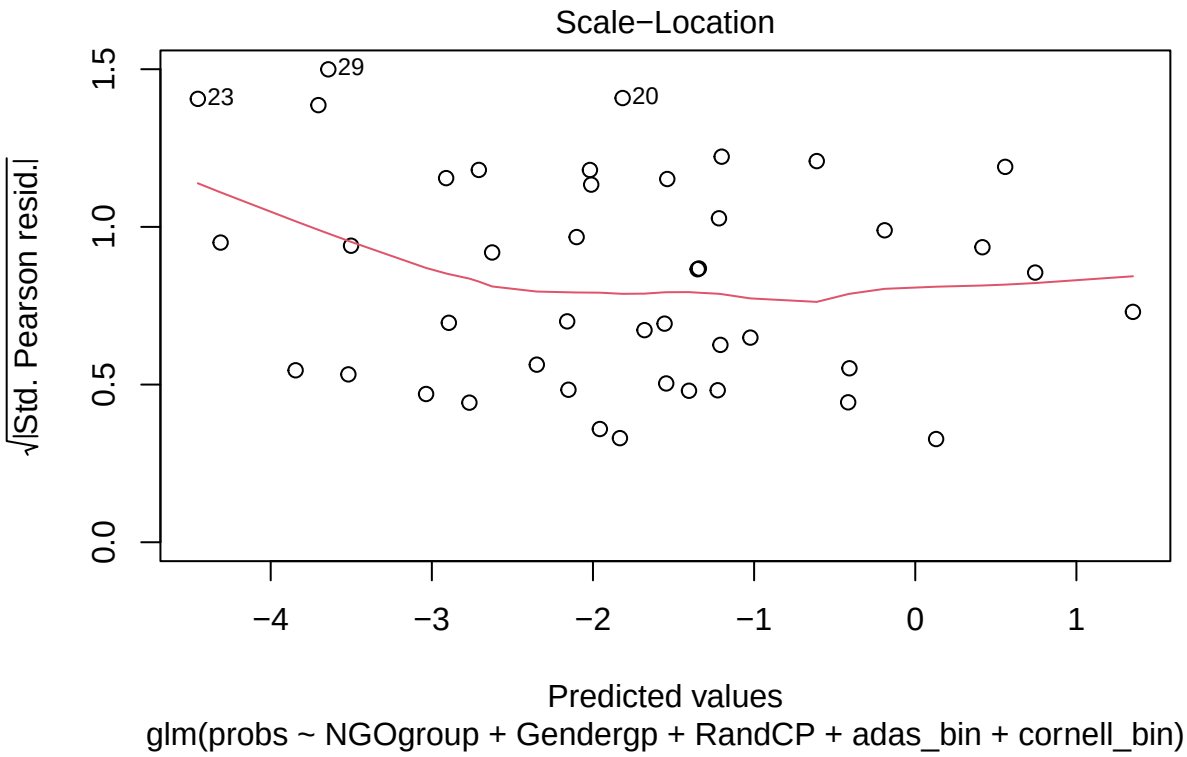
```
# adas bins: [0,2] [3,5] and [6,9]
df_red2 <- df_red2 %>% mutate(adas_bin = cut(b_Adas_DelayedRecall, breaks=c(-1, 2, 5, 9)))
# Cornell bins: use 6 as depression cutoff (mentioned in Lam et al paper pg. 5)
df_red2 <- df_red2 %>% mutate(cornell_bin = cut(b_Cornell_Total, breaks=c(-1, 5, 24)))

ag.df <- aggregate(Statusgp ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin, data = df_red2, FU
n <- aggregate(Statusgp ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin, data = df_red2, FUN = l
probs <- ag.df$Statusgp / n$Statusgp
ag.df_aug <- cbind(ag.df, probs)

ag.mod <- glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin, family = binomial
summary(ag.mod)
```

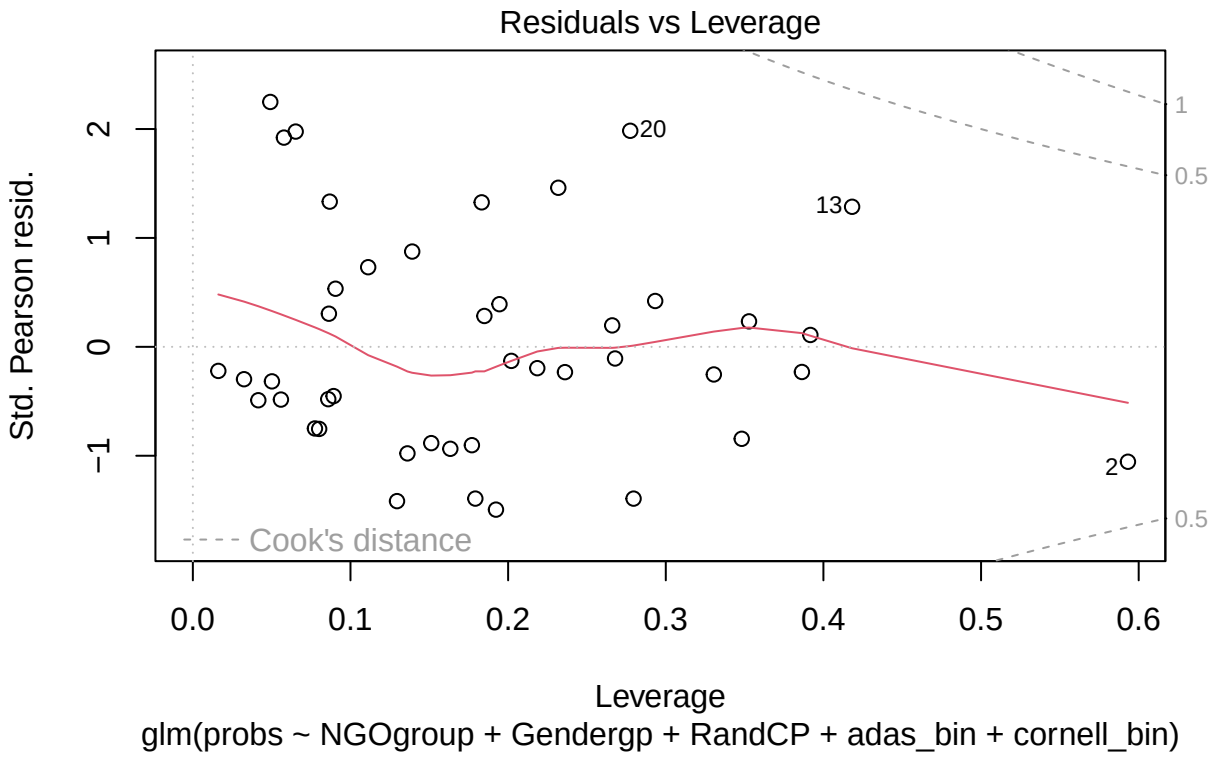
```
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin +
##     cornell_bin, family = binomial, data = ag.df_aug, weights = n$Statusgp)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.5180     0.4948  -7.110 1.16e-12 ***
## NGOgroup2           2.2999     0.5030   4.573 4.82e-06 ***
## NGOgroup3           1.6852     0.4829   3.490 0.000483 ***
## Gendergp1           0.6068     0.3014   2.013 0.044072 *
## RandCP              0.8096     0.3501   2.312 0.020762 *
## adas_bin(2,5]      -0.7929     0.3038  -2.610 0.009045 **
## adas_bin(5,9]      -0.9342     0.4045  -2.309 0.020923 *
## cornell_bin(5,24]   1.9623     0.6911   2.839 0.004520 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.635  on 42  degrees of freedom
## Residual deviance: 36.502  on 35  degrees of freedom
## AIC: 107.69
##
## Number of Fisher Scoring iterations: 5
```

```
plot(ag.mod)
```

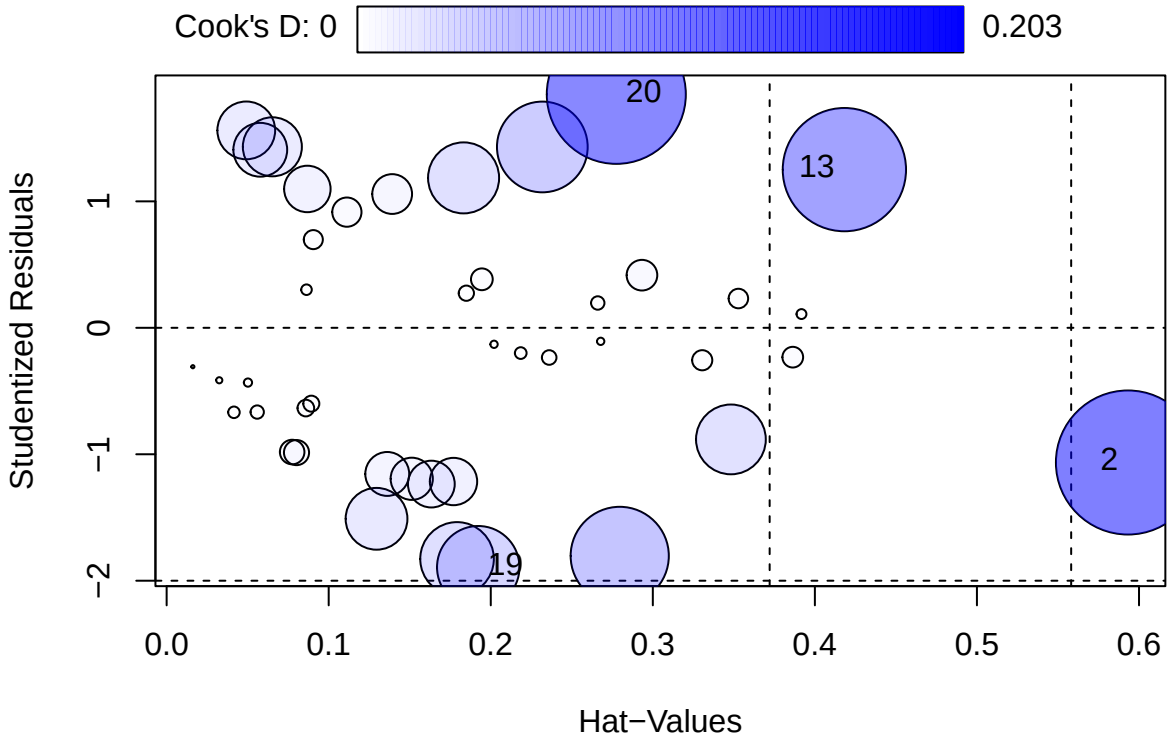

Residuals vs Fitted
Pearson Residuals
Predicted values
glm(probs ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin)
23
29
15
2
1
0
−1
−4
−3
−2
−1
0
1


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
17
7
19
Theoretical Quantiles
glm(probs ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
23
29
20
−4
−3
−2
−1
0
1
Predicted values
glm(probs ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin)




```
## Influence Plot for aggregate model
car::influencePlot(ag.mod)
```




```
##      StudRes       Hat      CookD
## 2  -1.065118 0.5931821 0.20292338
## 13  1.250748 0.4182056 0.14859621
## 19 -1.893425 0.1922831 0.06648379
## 20  1.846197 0.2774651 0.18893144
```

```
## Seeing the influential points
misses <- ag.df_aug[c(2,13,19,20), ]
misses$est_prob <- ag.mod$fitted.values[c(2,13,19,20)]
View(misses)

## Sensitivity analysis: dropping one influential point at a time
for (i in c(2,13,19,20)) {
  without_misses <- ag.df_aug[-i, ]
  ag.mod.test <- glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin, family = b
  print(summary(ag.mod.test))
}
```

```
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin +
##     cornell_bin, family = binomial, data = without_misses, weights = n$Statusgp[-i])
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.3755     0.5066  -6.662 2.69e-11 ***
## NGOgroup2           2.4901     0.5343   4.660 3.16e-06 ***
```

```
## NGOgroup3           1.6862     0.4826   3.494 0.000476 ***
## Gendergp1           0.4927     0.3201   1.539 0.123780
## RandCP              0.7793     0.3507   2.222 0.026250 *
## adas_bin(2,5]      -0.9867     0.3548  -2.781 0.005422 **
## adas_bin(5,9]      -1.1465     0.4525  -2.534 0.011280 *
## cornell_bin(5,24]   1.8742     0.7065   2.653 0.007984 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 84.861  on 41  degrees of freedom
## Residual deviance: 35.374  on 34  degrees of freedom
## AIC: 102.71
##
## Number of Fisher Scoring iterations: 5
##
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin +
##     cornell_bin, family = binomial, data = without_misses, weights = n$Statusgp[-i])
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.5143     0.4982  -7.054 1.74e-12 ***
## NGOgroup2           2.1781     0.5178   4.207 2.59e-05 ***
## NGOgroup3           1.6930     0.4851   3.490 0.000483 ***
## Gendergp1           0.7265     0.3181   2.284 0.022386 *
## RandCP              0.8566     0.3577   2.395 0.016627 *
## adas_bin(2,5]      -1.0002     0.3534  -2.830 0.004649 **
## adas_bin(5,9]      -0.9233     0.4055  -2.277 0.022786 *
## cornell_bin(5,24]   2.0258     0.6951   2.915 0.003561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 85.908  on 41  degrees of freedom
## Residual deviance: 34.919  on 34  degrees of freedom
## AIC: 102.34
##
## Number of Fisher Scoring iterations: 5
##
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin +
##     cornell_bin, family = binomial, data = without_misses, weights = n$Statusgp[-i])
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.6456     0.5093  -7.158 8.18e-13 ***
## NGOgroup2           2.4667     0.5206   4.738 2.15e-06 ***
## NGOgroup3           1.7178     0.4851   3.541 0.000399 ***
## Gendergp1           0.5568     0.3027   1.839 0.065881 .
```

```
## RandCP              0.9983     0.3714   2.688 0.007192 **
## adas_bin(2,5]      -0.7156     0.3045  -2.350 0.018769 *
## adas_bin(5,9]      -0.9496     0.4057  -2.340 0.019262 *
## cornell_bin(5,24]   1.9618     0.6897   2.844 0.004451 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 85.042  on 41  degrees of freedom
## Residual deviance: 32.921  on 34  degrees of freedom
## AIC: 104.11
##
## Number of Fisher Scoring iterations: 5
##
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin +
##     cornell_bin, family = binomial, data = without_misses, weights = n$Statusgp[-i])
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.3458     0.4963  -6.742 1.57e-11 ***
## NGOgroup2           2.1822     0.5035   4.334 1.46e-05 ***
## NGOgroup3           1.4765     0.4959   2.977  0.00291 **
## Gendergp1           0.7168     0.3095   2.316  0.02057 *
## RandCP              0.5685     0.3861   1.472  0.14094
## adas_bin(2,5]      -0.9822     0.3283  -2.992  0.00277 **
## adas_bin(5,9]      -0.9556     0.4066  -2.350  0.01875 *
## cornell_bin(5,24]   1.9624     0.7000   2.803  0.00506 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.431  on 41  degrees of freedom
## Residual deviance: 33.051  on 34  degrees of freedom
## AIC: 101.08
##
## Number of Fisher Scoring iterations: 5
```

```
## Dropping all influential points at once
without_misses1 <- ag.df_aug[-c(2,13,19,20), ]
ag.mod2 <- glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin + cornell_bin, family = binomi

summary(ag.mod2)
```

```
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + RandCP + adas_bin +
##     cornell_bin, family = binomial, data = without_misses1, weights = n$Statusgp[-c(2,
##     13, 19, 20)])
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)        -3.2754     0.5232  -6.261 3.83e-10 ***
## NGOgroup2           2.3816     0.6101   3.904 9.48e-05 ***
## NGOgroup3           1.4893     0.5000   2.979  0.00289 **
## Gendergp1           0.6785     0.3863   1.756  0.07903 .
## RandCP              0.6949     0.4129   1.683  0.09235 .
## adas_bin(2,5]      -1.3303     0.4333  -3.070  0.00214 **
## adas_bin(5,9]      -1.1697     0.4769  -2.453  0.01418 *
## cornell_bin(5,24]   1.9277     0.7247   2.660  0.00782 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 78.602  on 38  degrees of freedom
## Residual deviance: 27.607  on 31  degrees of freedom
## AIC: 88.025
##
## Number of Fisher Scoring iterations: 5
```

```
### New Model based on aggregations
mod_b_sel3 <- glm(formula = Statusgp ~ NGOgroup + factor(RandCP) + Gendergp + adas_bin + cornell_bin, fa
summary(mod_b_sel3)
```

```
##
## Call:
## glm(formula = Statusgp ~ NGOgroup + factor(RandCP) + Gendergp +
##     adas_bin + cornell_bin, family = binomial, data = df_red2)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.5180     0.4948  -7.110 1.16e-12 ***
## NGOgroup2           2.2999     0.5030   4.573 4.82e-06 ***
## NGOgroup3           1.6852     0.4829   3.490 0.000484 ***
## factor(RandCP)1     0.8096     0.3501   2.312 0.020762 *
## Gendergp1           0.6068     0.3014   2.013 0.044073 *
## adas_bin(2,5]      -0.7929     0.3038  -2.610 0.009045 **
## adas_bin(5,9]      -0.9342     0.4045  -2.309 0.020923 *
## cornell_bin(5,24]   1.9623     0.6911   2.839 0.004520 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 355.36  on 534  degrees of freedom
## AIC: 371.36
##
## Number of Fisher Scoring iterations: 6
```

```
CIs <- exp(confint(mod_b_sel3))
```

```
## Waiting for profiling to be done...
```

```
coeffs <- exp(mod_b_sel3$coefficients)
coeffs
```

```
##       (Intercept)         NGOgroup2         NGOgroup3   factor(RandCP)1
##        0.02965852        9.97325927        5.39349912        2.24702382
##         Gendergp1     adas_bin(2,5]     adas_bin(5,9] cornell_bin(5,24]
##        1.83458346        0.45251946        0.39290756        7.11535408
```

```
df_res <- data.frame(Estimates = coeffs)
df_res <- cbind(df_res, CIs)
View(df_res)
```

Truncated dataframe for use in bestglm

```
library(bestglm)
```

```
## Warning: package 'bestglm' was built under R version 4.3.1
```

```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 4.3.1
```

```
reduced_df <- df[, colnames(df) %in% rel_vars]
reduced_df <- reduced_df[complete.cases(reduced_df), ]

reduced_df$y <- reduced_df$Statusgp
reduced_df$Statusgp <- NULL

## All Subsets Regression
res.bestglm <-
    bestglm(Xy = as.data.frame(reduced_df),
            family = binomial(link = 'logit'),
            IC = "AIC",
            method = "exhaustive")
```

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

```
## Note: factors present with more than 2 levels.
```

```
res.bestglm$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
##          (Intercept)             NGOgroup2             NGOgroup3
##              -3.2281                2.0976                1.6451
##            Gendergp1  b_Adas_DelayedRecall       b_Cornell_Total
##               0.6395               -0.1844                0.1850
##
## Degrees of Freedom: 541 Total (i.e. Null);  536 Residual
## Null Deviance:       405.5
## Residual Deviance: 355.6     AIC: 367.6
```

```
res.bestglm
```

```
## AIC
## Best Model:
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## NGOgroup               2   2.33  1.1635  11.679 1.08e-05 ***
## Gendergp               1   0.44  0.4420   4.437  0.03563 *
```

```
## b_Adas_DelayedRecall   1   0.92  0.9247   9.283  0.00243 **
## b_Cornell_Total        1   1.63  1.6292  16.355 6.02e-05 ***
## Residuals            536  53.39  0.0996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
best_mod2 <- glm(formula = y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total, family = bi

summary(best_mod2)
```

```
##
## Call:
## glm(formula = y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall +
##     b_Cornell_Total, family = binomial, data = reduced_df)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.22815    0.49104  -6.574 4.89e-11 ***
## NGOgroup2             2.09755    0.49002   4.281 1.86e-05 ***
## NGOgroup3             1.64510    0.49574   3.318 0.000905 ***
## Gendergp1             0.63948    0.30251   2.114 0.034523 *
## b_Adas_DelayedRecall -0.18443    0.06307  -2.924 0.003454 **
## b_Cornell_Total       0.18497    0.05296   3.493 0.000478 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 355.56  on 536  degrees of freedom
## AIC: 367.56
##
## Number of Fisher Scoring iterations: 6
```

Diagnostics for predictive model (model fit)

```
# Aggregated Data

ag.df <- aggregate(y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total, data= reduced_df, M

n <- aggregate(y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total, data= reduced_df, FUN =

View(ag.df)

probs <- ag.df$y / n$y

sum(probs == 0)
```

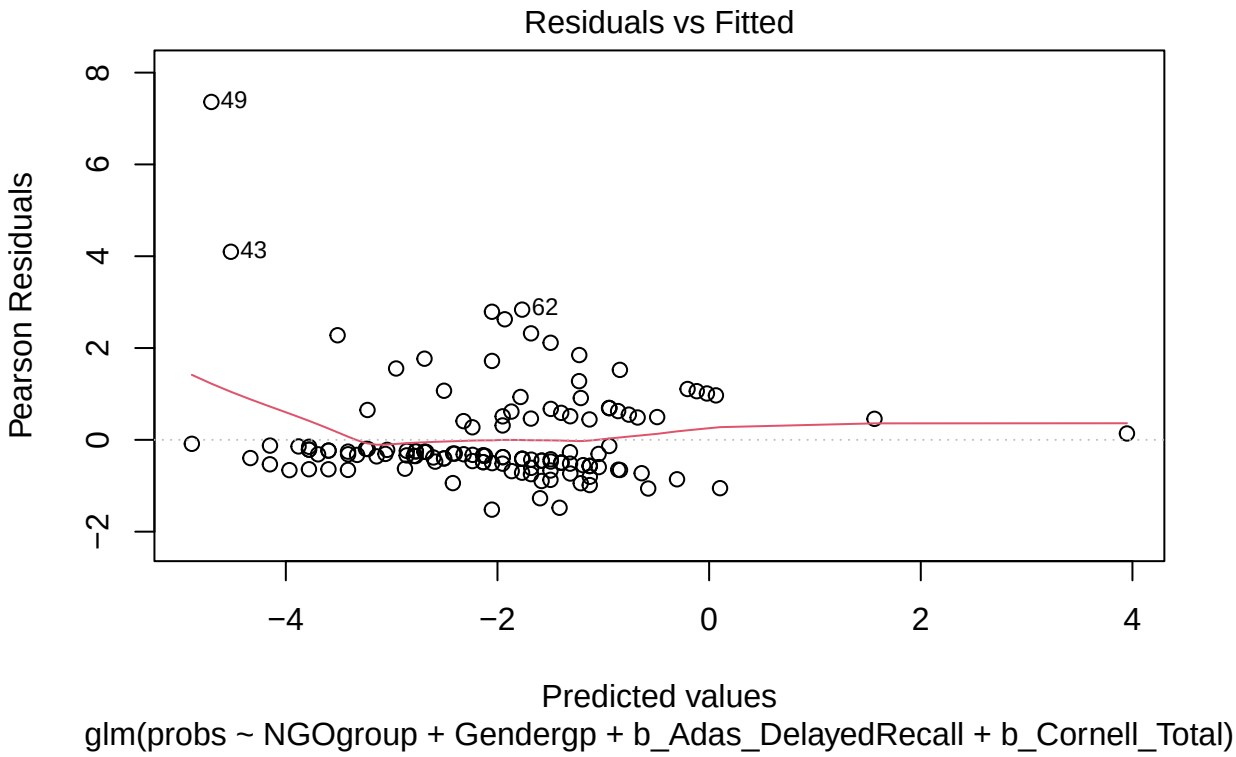
```
## [1] 84
```

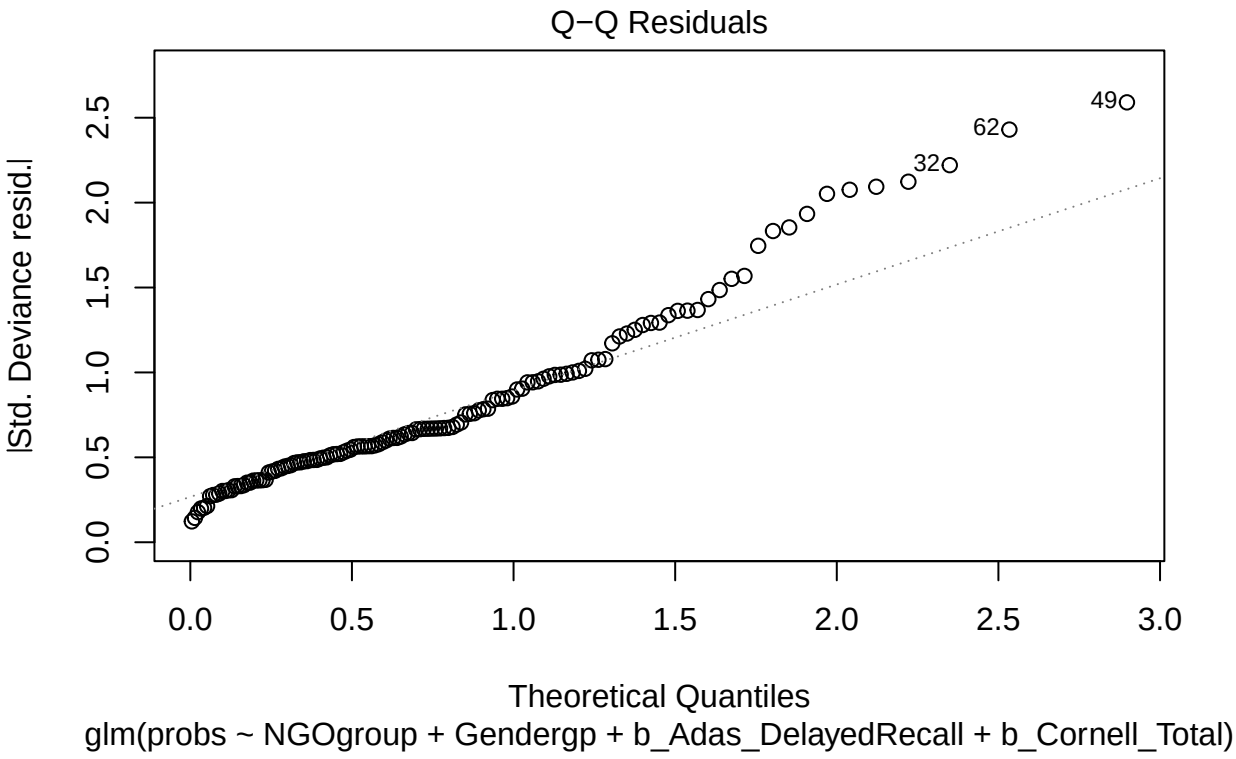
```
ag.df_aug <- cbind(ag.df, probs)
View(ag.df_aug)


ag.mod <- glm(formula = probs ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total, family = M
```

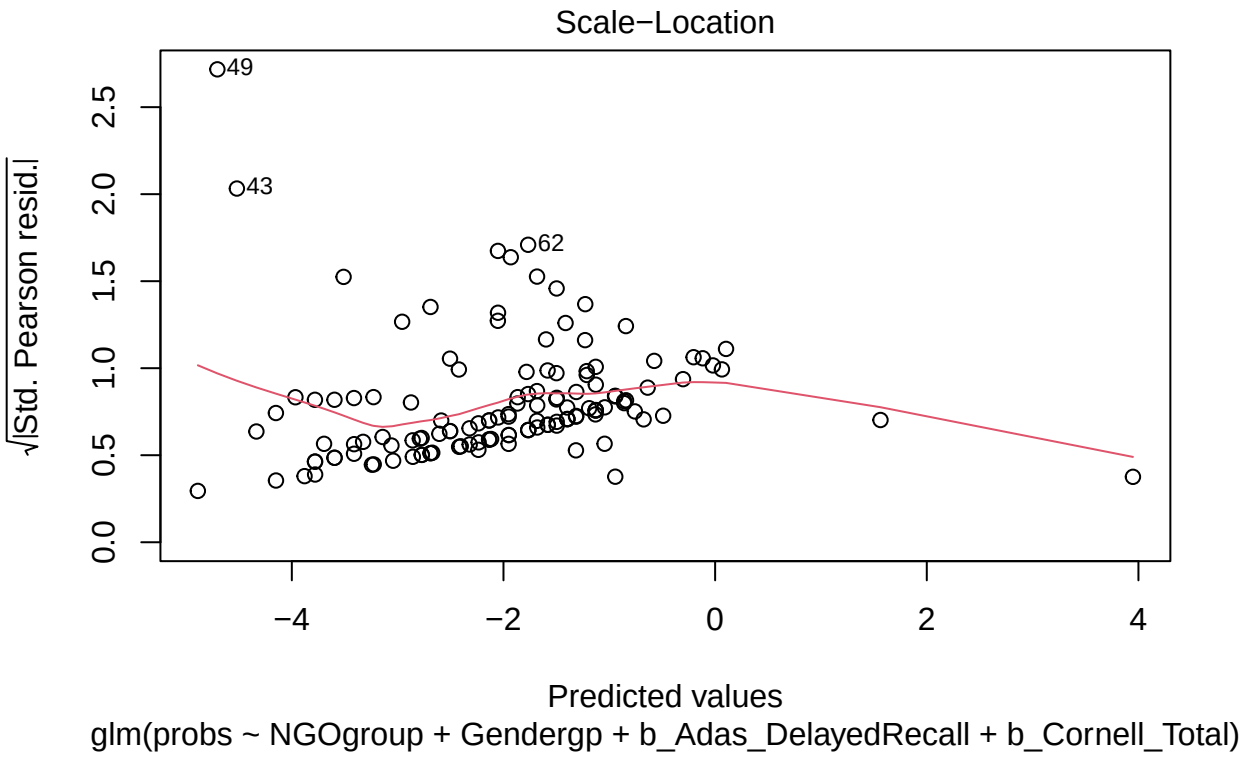
```
summary(ag.mod)
```

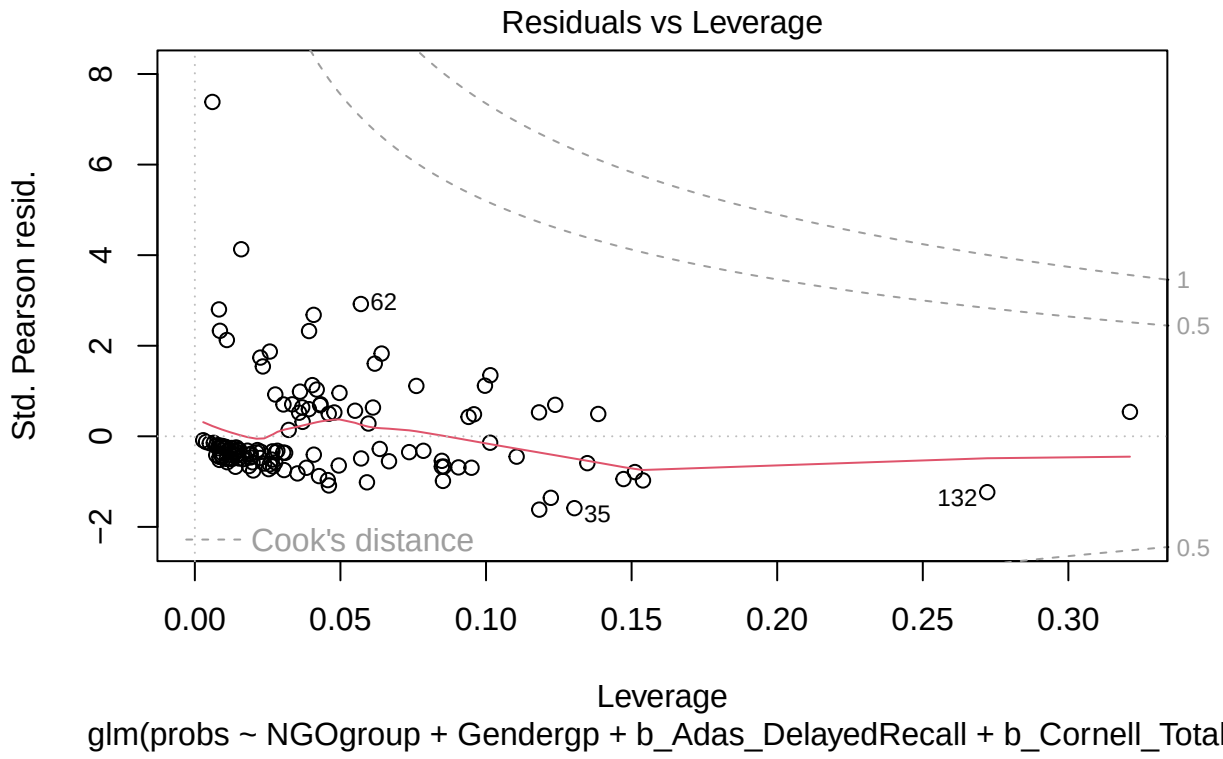
```
##
## Call:
## glm(formula = probs ~ NGOgroup + Gendergp + b_Adas_DelayedRecall +
##     b_Cornell_Total, family = binomial, data = ag.df_aug, weights = n$y)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.22815    0.49101  -6.575 4.88e-11 ***
## NGOgroup2             2.09755    0.48999   4.281 1.86e-05 ***
## NGOgroup3             1.64510    0.49571   3.319 0.000904 ***
## Gendergp1             0.63948    0.30251   2.114 0.034522 *
## b_Adas_DelayedRecall -0.18443    0.06307  -2.924 0.003454 **
## b_Cornell_Total       0.18497    0.05296   3.493 0.000478 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 161.90  on 132  degrees of freedom
## Residual deviance: 111.96  on 127  degrees of freedom
## AIC: 193.99
##
## Number of Fisher Scoring iterations: 5
```

```
plot(ag.mod)
```


Residuals vs Fitted
Pearson Residuals
Predicted values
glm(probs ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total)
49
43
62
8
6
4
2
0
−2
−4
−2
0
2
4


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
49
62
32
Theoretical Quantiles
glm(probs ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
49
43
62
−4
−2
0
2
4
Predicted values
glm(probs ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + b_Cornell_Total)




```
pchisq(111.96, df= 127, lower.tail = F)
```

```
## [1] 0.82677
```

```
summary(best_mod2)
```

```
##
## Call:
## glm(formula = y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall +
##     b_Cornell_Total, family = binomial, data = reduced_df)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.22815    0.49104  -6.574 4.89e-11 ***
## NGOgroup2             2.09755    0.49002   4.281 1.86e-05 ***
## NGOgroup3             1.64510    0.49574   3.318 0.000905 ***
## Gendergp1             0.63948    0.30251   2.114 0.034523 *
## b_Adas_DelayedRecall -0.18443    0.06307  -2.924 0.003454 **
## b_Cornell_Total       0.18497    0.05296   3.493 0.000478 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 355.56  on 536  degrees of freedom
## AIC: 367.56
```

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

```
# compare to saturated model
pchisq(355.56, df = 536, lower.tail = F)
```

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

```
# no evidence of better complexity being good
```

Assessing predictive accuracy: Tuning cutoff and Checking mislabels

```
p_seq <- seq(0.01, 0.99, 0.01)

acc <- rep(0, length(p_seq))
actuals <- reduced_df$y

for (i in 1:length(p_seq)) {
  preds <- best_mod2$fitted.values > p_seq[i]
  comps <- preds == actuals
  acc[i] <- sum(comps) / length(comps)
}

accuracy <- max(acc)
best_p <- p_seq[which.max(acc)]

preds <- best_mod2$fitted.values > best_p

library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.1
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
preds <- factor(as.numeric(preds))
actuals <- factor(actuals)

confusionMatrix(data = actuals, reference = preds, dnn = c('Actual', 'Prediction'))
```

```
## Confusion Matrix and Statistics
##
##       Prediction
## Actual   0   1
##      0 474   1
##      1  61   6
##
##                Accuracy : 0.8856
##                  95% CI : (0.8558, 0.9112)
##     No Information Rate : 0.9871
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1421
##
##  Mcnemar's Test P-Value : 6.731e-14
##
```

```
##             Sensitivity : 0.88598
##             Specificity : 0.85714
##          Pos Pred Value : 0.99789
##          Neg Pred Value : 0.08955
##              Prevalence : 0.98708
##          Detection Rate : 0.87454
##    Detection Prevalence : 0.87638
##       Balanced Accuracy : 0.87156
##
##        'Positive' Class : 0
##
```

```
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
f_mod_pred <- glm(formula = y ~ NGOgroup + Randomization + Age + Gendergp + Educationyears +
b_Adas_DelayedRecall + b_Adas_Total + b_CMMSE + b_CVFT_Total + b_cdr_sumofboxes +
b_Cornell_Total + b_CIRS_Total + b_CNPI_Total + b_MIC_Total, family = binomial, data = reduced_df)

View(head(df_red2))

cv.glm(data = reduced_df, glmfit = best_mod2, K = 10)$delta
```

```
## [1] 0.09907087 0.09888277
```

```
cv.glm(data = reduced_df, glmfit = f_mod_pred, K = 10)$delta
```

```
## [1] 0.1047436 0.1040540
```

Miscellaenous

New aggregations

```
summary(reduced_df$b_Adas_DelayedRecall)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   3.439   5.000   9.000
```

```
summary(reduced_df$b_Cornell_Total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.6882  0.0000 24.0000
```

```
# 6 or higher is determined to be clinically relevant depression
reduced_df$Deprs <- reduced_df$b_Cornell_Total >= 6

ag_mod2 <- glm(formula = y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs, family = binomial, data
```

```r
cutoffs_adas <- quantile(reduced_df$b_Adas_DelayedRecall, c(1/4, 1/2, 3/4))
cutoffs_adas
```

```
## 25% 50% 75%
##   2   3   5
```

```r
reduced_df$adas_dr2 <- rep(0, nrow(reduced_df))
reduced_df$adas_dr2[reduced_df$b_Adas_DelayedRecall <= cutoffs_adas[1]] <- 0
reduced_df$adas_dr2[reduced_df$b_Adas_DelayedRecall > cutoffs_adas[1] & reduced_df$b_Adas_DelayedRecall

reduced_df$adas_dr2[reduced_df$b_Adas_DelayedRecall > cutoffs_adas[2] & reduced_df$b_Adas_DelayedRecall

reduced_df$adas_dr2[reduced_df$b_Adas_DelayedRecall > cutoffs_adas[3]] <- 3

ag_mod3 <- glm(formula = y ~ NGOgroup + Gendergp + adas_dr2 + Deprs, family = binomial, data = reduced_

summary(ag_mod3)
```

```
##
## Call:
## glm(formula = y ~ NGOgroup + Gendergp + adas_dr2 + Deprs, family = binomial,
##     data = reduced_df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.2200     0.4463  -7.215 5.38e-13 ***
## NGOgroup2     1.9657     0.4662   4.217 2.48e-05 ***
## NGOgroup3     1.5867     0.4764   3.330 0.000868 ***
## Gendergp1     0.6037     0.2990   2.019 0.043466 *
## adas_dr2     -0.3294     0.1236  -2.665 0.007698 **
## DeprsTRUE     1.9294     0.6889   2.801 0.005098 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 361.88  on 536  degrees of freedom
## AIC: 373.88
##
## Number of Fisher Scoring iterations: 6
```

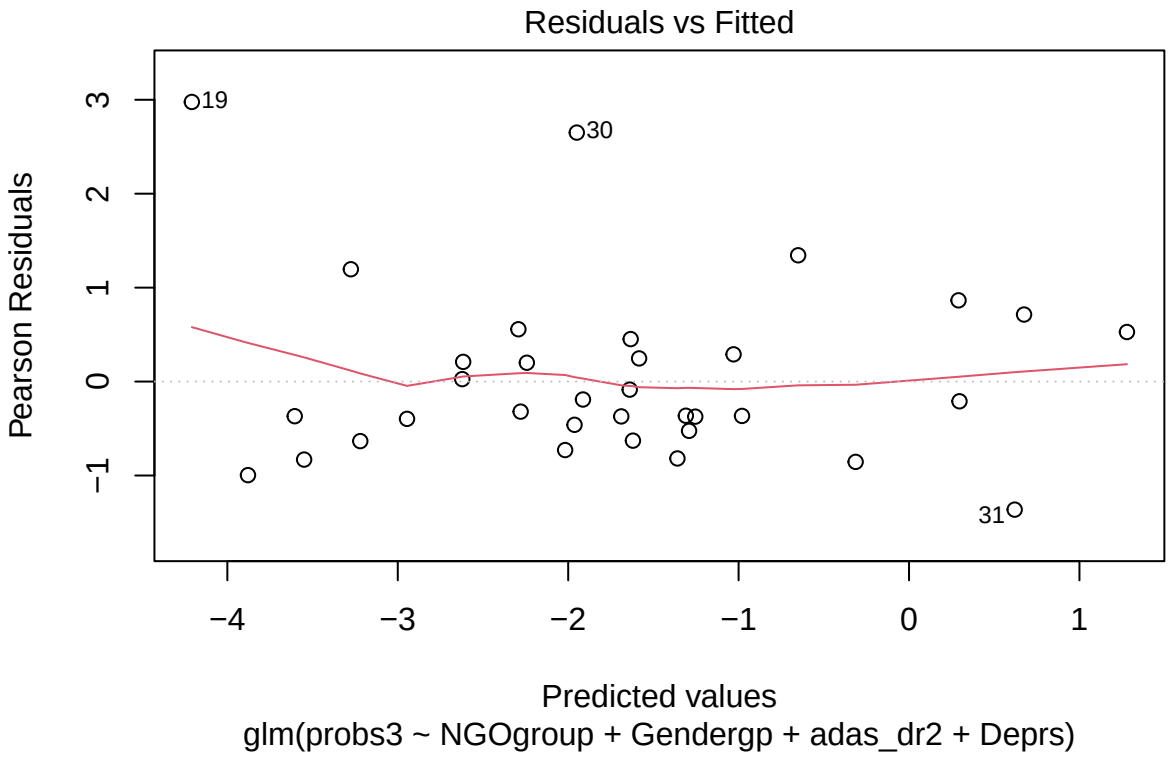
Residual diagnostics for most aggregated model

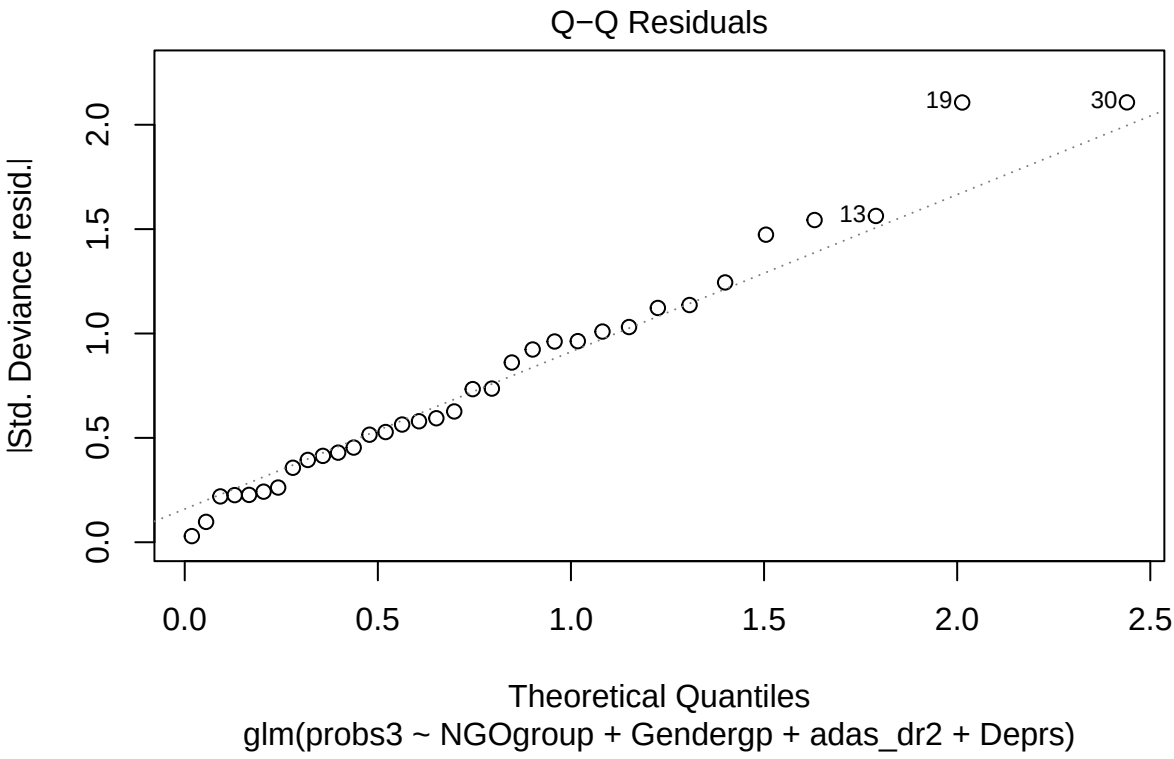
```r
ag.df3 <- aggregate(y ~ NGOgroup + Gendergp + adas_dr2 + Deprs, data= reduced_df, FUN = sum)

n3 <- aggregate(y ~ NGOgroup + Gendergp + adas_dr2 + Deprs, data= reduced_df, FUN = length)


probs3 <- ag.df3$y / n3$y



ag.df_aug3 <- cbind(ag.df3, probs3)
```

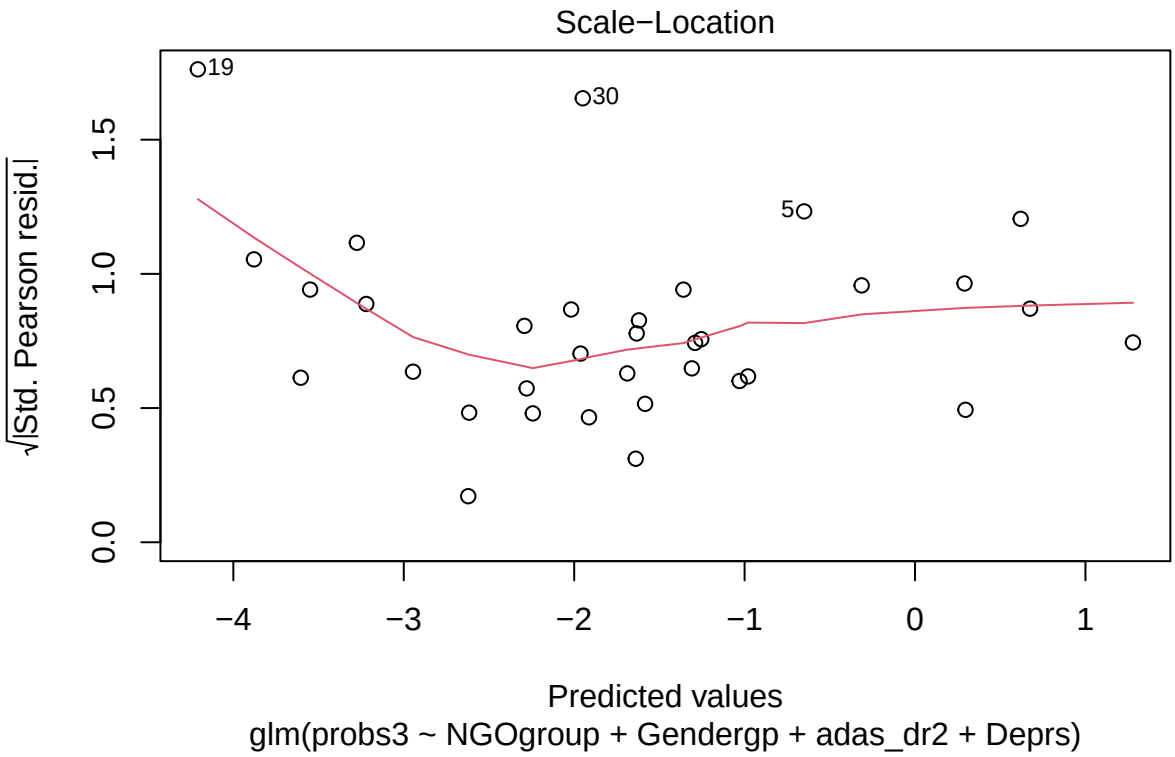
```r
ag.mod3 <- glm(formula = probs3 ~ NGOgroup + Gendergp + adas_dr2 + Deprs, family = binomial, data = ag.
summary(ag.mod3)
```

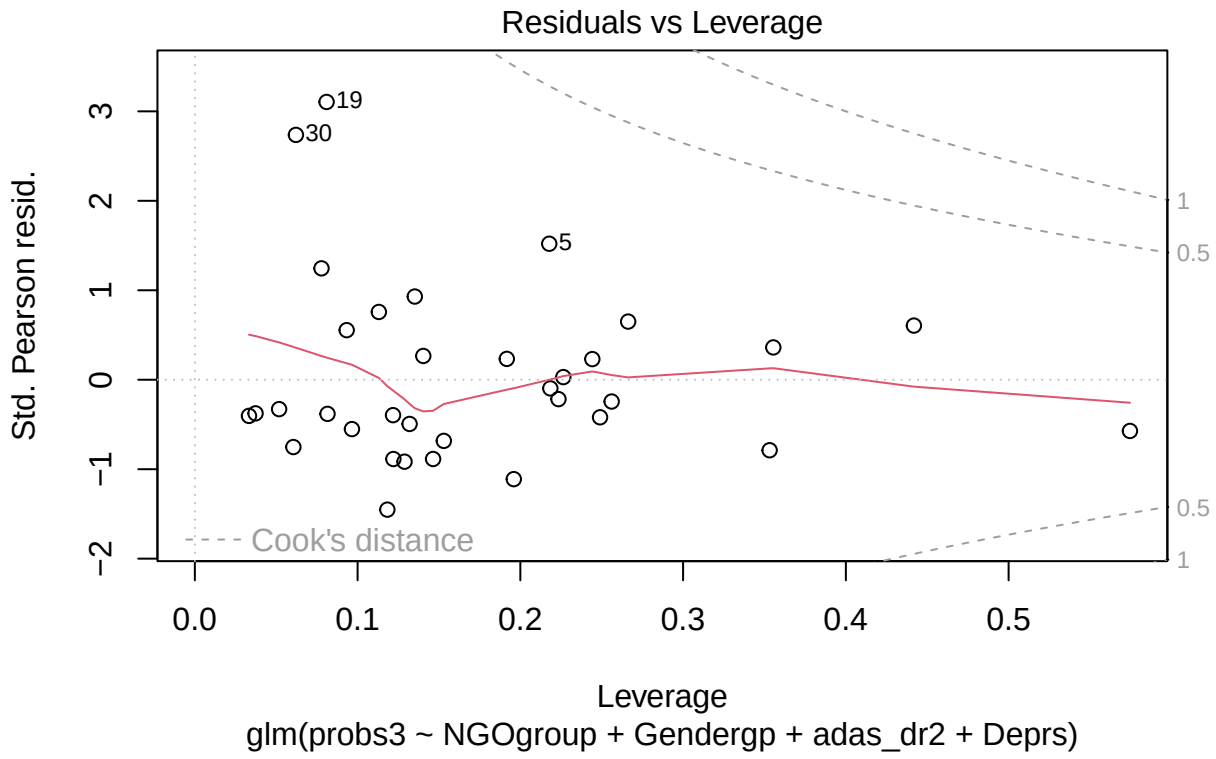
```
##
## Call:
## glm(formula = probs3 ~ NGOgroup + Gendergp + adas_dr2 + Deprs,
##     family = binomial, data = ag.df_aug3, weights = n3$y)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.2200     0.4463  -7.215 5.38e-13 ***
## NGOgroup2     1.9657     0.4662   4.217 2.48e-05 ***
## NGOgroup3     1.5867     0.4764   3.330 0.000868 ***
## Gendergp1     0.6037     0.2990   2.019 0.043466 *
## adas_dr2     -0.3294     0.1236  -2.665 0.007698 **
## DeprsTRUE     1.9294     0.6889   2.801 0.005098 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.288  on 33  degrees of freedom
## Residual deviance: 25.682  on 28  degrees of freedom
## AIC: 89.287
##
## Number of Fisher Scoring iterations: 5
```

```r
plot(ag.mod3)
```


Residuals vs Fitted
Pearson Residuals
Predicted values
glm(probs3 ~ NGOgroup + Gendergp + adas_dr2 + Deprs)
19
30
31
3
2
1
0
−1
−4
−3
−2
−1
0
1


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
19
30
13
Theoretical Quantiles
glm(probs3 ~ NGOgroup + Gendergp + adas_dr2 + Deprs)


Scale−Location
19
30
5
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−4
−3
−2
−1
0
1
Predicted values
glm(probs3 ~ NGOgroup + Gendergp + adas_dr2 + Deprs)




```
pchisq(25.682, df = 28, lower.tail = F)
```

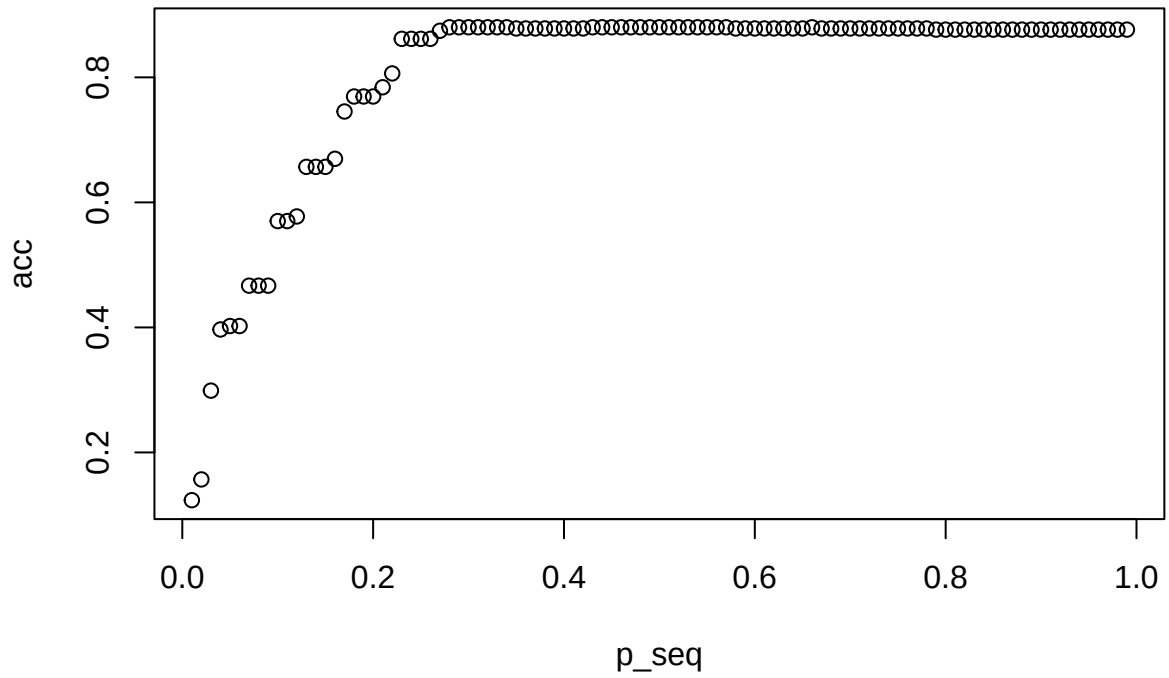
```
## [1] 0.5905193
```

Check by specificity –> does not work

Just change the weights; weight the true positives higher than true negatives

```
p_seq <- seq(0.01, 0.99, 0.01)

acc <- rep(0, length(p_seq))
sens <- rep(0, length(p_seq))
spec <- rep(0, length(p_seq))
ppv_Vec <- rep(0, length(p_seq))


for (i in 1:length(p_seq)) {
  preds <- ag_mod3$fitted.values >= p_seq[i]
  actuals <- reduced_df$y

  comps <- preds == actuals

  acc[i] <- sum(comps) / length(comps)
  sens[i] <- sensitivity(factor(actuals), reference = factor(as.numeric(preds)))
  spec[i] <- specificity(factor(actuals), reference = factor(as.numeric(preds)))
  ppv_Vec[i] <- posPredValue(factor(actuals), reference = factor(as.numeric(preds)))
}
```

```
plot(p_seq, acc)
```




```
max_accuracy <- max(acc)

best_cutoff <- p_seq[which.max(acc)]


preds <- ag_mod3$fitted.values >= best_cutoff

# Confusion Matrix for max accuracy

conf_mat <- matrix(c(sum(actuals == 0 & preds == 0), sum(actuals == 0 & preds == 1),
                     sum(actuals == 1 & preds == 0), sum(actuals == 1 & preds == 1)), nrow = 2, ncol = 2
dimnames(conf_mat) <- list(Actual = c('Stay', 'Drop'),
                           Prediction = c('Stay', 'Drop'))

conf_mat
```

```
##       Prediction
## Actual Stay Drop
##   Stay  468    7
##   Drop   58    9
```

```
# Alternate Solution using caret
library(caret)
```

```r
preds <- factor(as.numeric(preds))
actuals <- factor(actuals)

confusionMatrix(data = actuals, reference = preds, dnn = c('Actual', 'Prediction'))
```

```
## Confusion Matrix and Statistics
##
##       Prediction
## Actual   0   1
##      0 468   7
##      1  58   9
##
##                Accuracy : 0.8801
##                  95% CI : (0.8497, 0.9062)
##     No Information Rate : 0.9705
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1777
##
##  Mcnemar's Test P-Value : 5.584e-10
##
##             Sensitivity : 0.8897
##             Specificity : 0.5625
##          Pos Pred Value : 0.9853
##          Neg Pred Value : 0.1343
##              Prevalence : 0.9705
##          Detection Rate : 0.8635
##    Detection Prevalence : 0.8764
##       Balanced Accuracy : 0.7261
##
##        'Positive' Class : 0
##
```

```r
max_sens <- max(sens)
best_cutoff_sens <- p_seq[which.max(sens)]

preds2 <- ag_mod3$fitted.values >= best_cutoff_sens
conf_mat <- matrix(c(sum(actuals == 0 & preds2 == 0), sum(actuals == 0 & preds2 == 1),
                     sum(actuals == 1 & preds2 == 0), sum(actuals == 1 & preds2 == 1)), nrow = 2, ncol =
dimnames(conf_mat) <- list(Actual = c('Stay', 'Drop'),
                           Prediction = c('Stay', 'Drop'))

conf_mat
```

```
##       Prediction
## Actual Stay Drop
##   Stay   97  378
##   Drop    2   65
```

```r
confusionMatrix(data = actuals, reference = factor(as.numeric(preds2)), dnn = c('Actual', 'Prediction'))
```

```
## Confusion Matrix and Statistics
##
##       Prediction
## Actual   0   1
```

```
##      0  97 378
##      1   2  65
##
##                Accuracy : 0.2989
##                  95% CI : (0.2606, 0.3394)
##     No Information Rate : 0.8173
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0511
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9798
##             Specificity : 0.1467
##          Pos Pred Value : 0.2042
##          Neg Pred Value : 0.9701
##              Prevalence : 0.1827
##          Detection Rate : 0.1790
##    Detection Prevalence : 0.8764
##       Balanced Accuracy : 0.5633
##
##        'Positive' Class : 0
##
```

```
b_se_sp <- max(sens * spec)
cutoff_se_sp <- p_seq[which.max(sens*spec)]
preds3 <- ag_mod3$fitted.values >= cutoff_se_sp
confusionMatrix(data = actuals, reference = factor(as.numeric(preds3)), dnn = c('Actual', 'Prediction'))
```

```
## Confusion Matrix and Statistics
##
##       Prediction
## Actual   0   1
##      0 475   0
##      1  65   2
##
##                Accuracy : 0.8801
##                  95% CI : (0.8497, 0.9062)
##     No Information Rate : 0.9963
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0512
##
##  Mcnemar's Test P-Value : 2.051e-15
##
##             Sensitivity : 0.87963
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 0.02985
##              Prevalence : 0.99631
##          Detection Rate : 0.87638
##    Detection Prevalence : 0.87638
##       Balanced Accuracy : 0.93981
##
##        'Positive' Class : 0
```

```
##
```

```
b_spec <- max(spec)
cutoff_spec <- p_seq[which.max(spec)]
preds4 <- ag_mod3$fitted.values >= cutoff_spec
confusionMatrix(data = actuals, reference = factor(as.numeric(preds4)), dnn = c('Actual', 'Prediction'))
```

```
## Confusion Matrix and Statistics
##
##       Prediction
## Actual   0   1
##      0 475   0
##      1  65   2
##
##                Accuracy : 0.8801
##                  95% CI : (0.8497, 0.9062)
##     No Information Rate : 0.9963
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0512
##
##  Mcnemar's Test P-Value : 2.051e-15
##
##             Sensitivity : 0.87963
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 0.02985
##              Prevalence : 0.99631
##          Detection Rate : 0.87638
##    Detection Prevalence : 0.87638
##       Balanced Accuracy : 0.93981
##
##        'Positive' Class : 0
##
```

Different aggregated model

```
best_mod2b <- glm(formula = y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs, family = binomial, d

summary(best_mod2b)
```

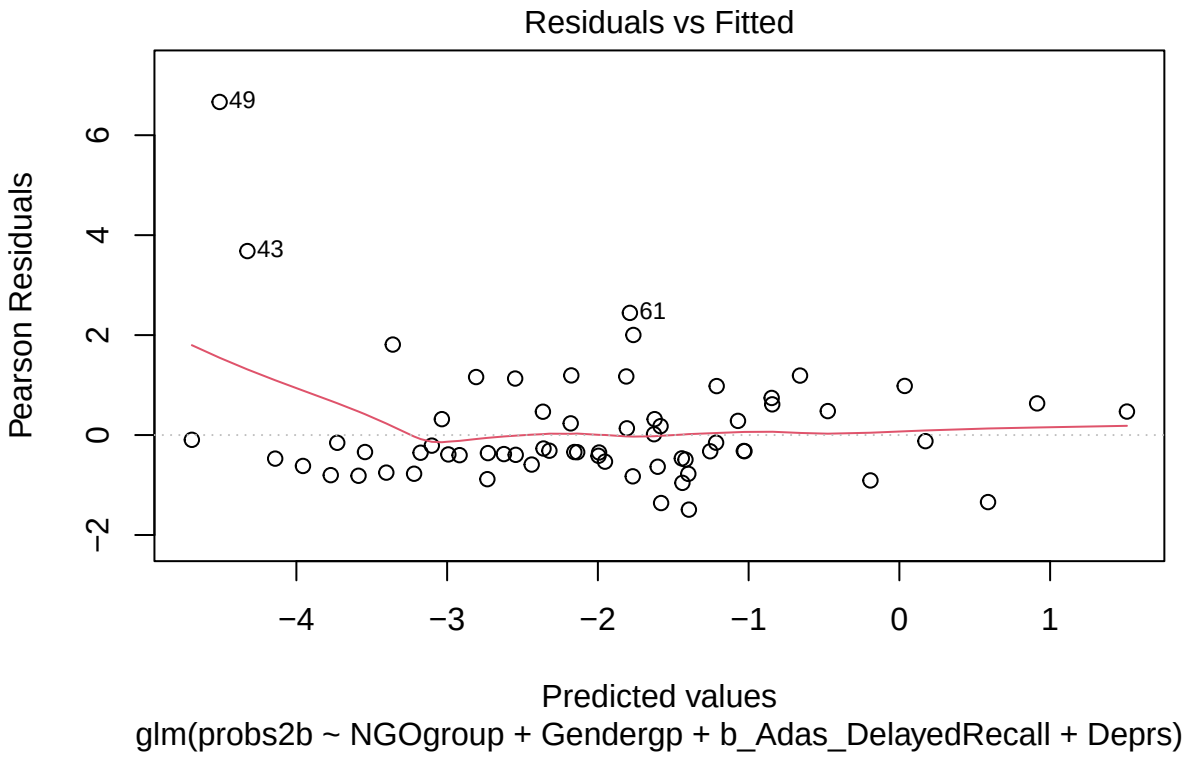
```
##
## Call:
## glm(formula = y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall +
##     Deprs, family = binomial, data = reduced_df)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.03488    0.46145  -6.577 4.81e-11 ***
## NGOgroup2             1.96424    0.46656   4.210 2.55e-05 ***
## NGOgroup3             1.59181    0.47676   3.339 0.000841 ***
## Gendergp1             0.59585    0.29948   1.990 0.046629 *
## b_Adas_DelayedRecall -0.18429    0.06276  -2.936 0.003322 **
## DeprsTRUE             1.98447    0.69181   2.869 0.004124 **
## ---
```

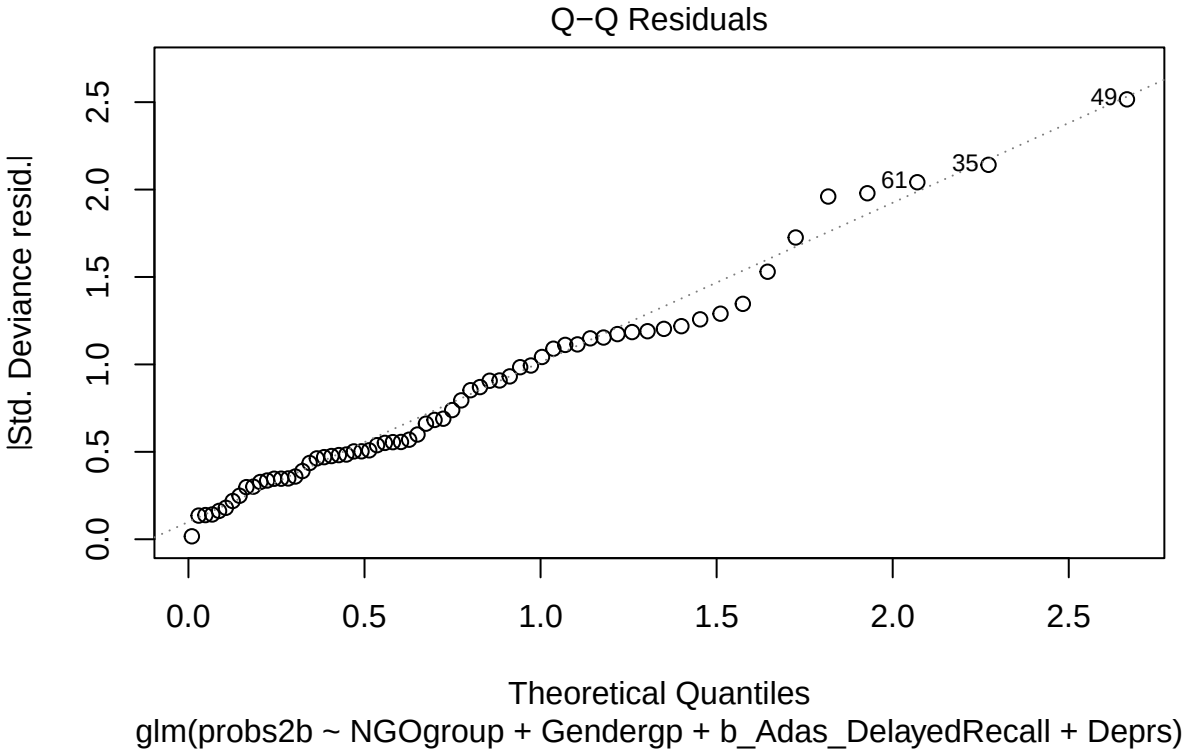
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 405.49  on 541  degrees of freedom
## Residual deviance: 360.30  on 536  degrees of freedom
## AIC: 372.3
##
## Number of Fisher Scoring iterations: 6
```

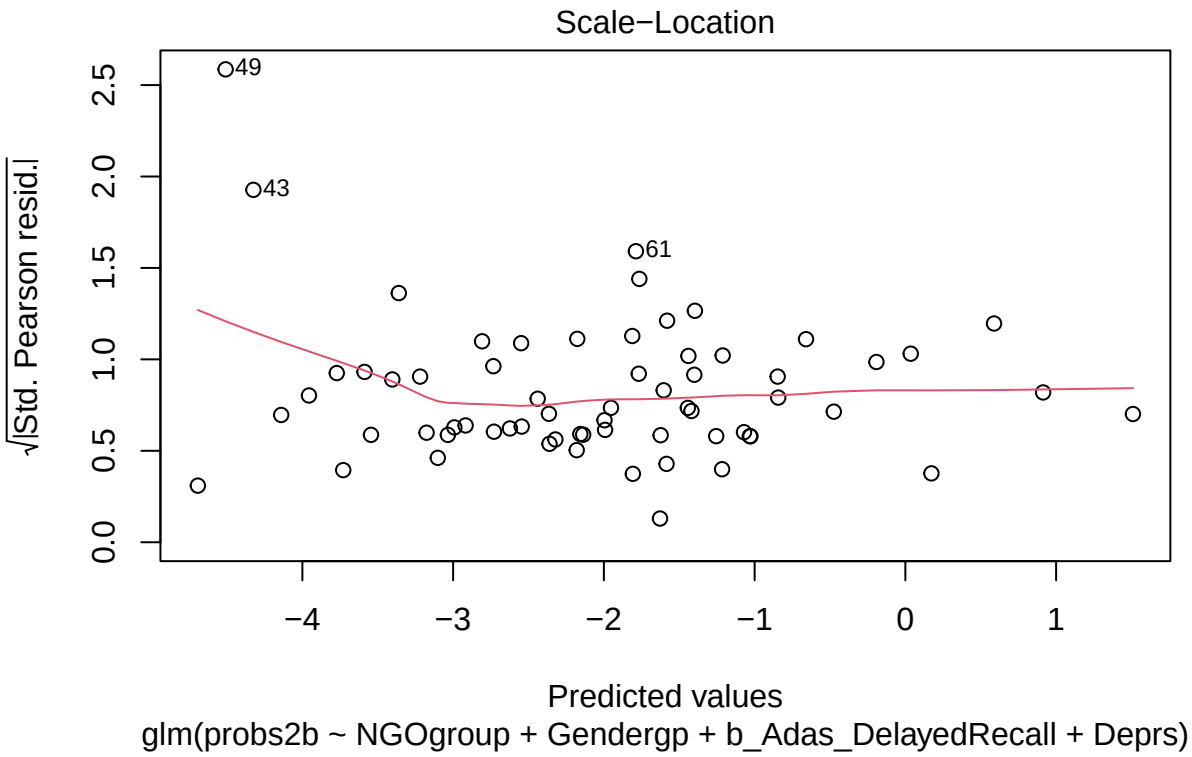
```
ag.df2b <- aggregate(y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs, data= reduced_df, FUN = sum

n2b <- aggregate(y ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs, data= reduced_df, FUN = length)


probs2b <- ag.df2b$y / n2b$y



ag.df_aug2b <- cbind(ag.df2b, probs2b)


ag.mod2b <- glm(formula = probs2b ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs, family = binomia
summary(ag.mod2b)
```

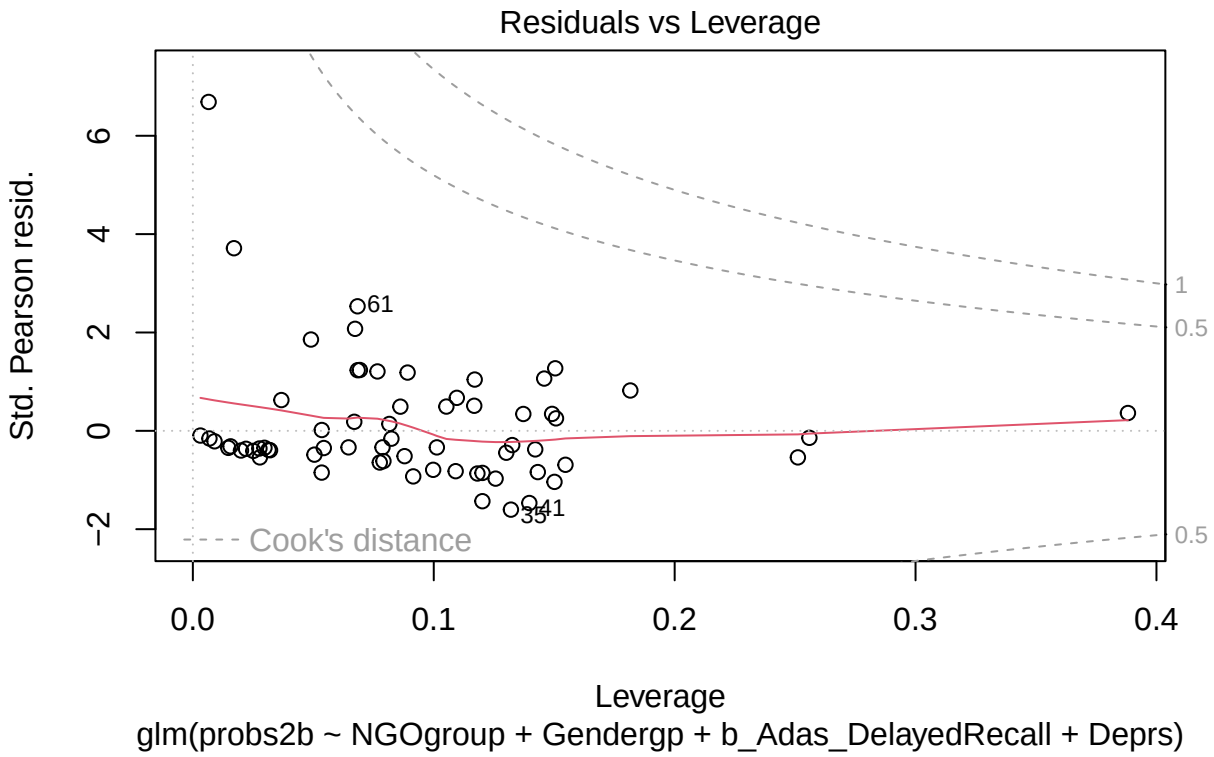
```
##
## Call:
## glm(formula = probs2b ~ NGOgroup + Gendergp + b_Adas_DelayedRecall +
##     Deprs, family = binomial, data = ag.df_aug2b, weights = n2b$y)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.03488    0.46145  -6.577 4.80e-11 ***
## NGOgroup2             1.96424    0.46656   4.210 2.55e-05 ***
## NGOgroup3             1.59181    0.47676   3.339 0.000841 ***
## Gendergp1             0.59585    0.29947   1.990 0.046629 *
## b_Adas_DelayedRecall -0.18429    0.06276  -2.936 0.003322 **
## DeprsTRUE             1.98447    0.69181   2.869 0.004124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 101.380  on 64  degrees of freedom
## Residual deviance:  56.187  on 59  degrees of freedom
## AIC: 137.62
##
## Number of Fisher Scoring iterations: 5
```

```
plot(ag.mod2b)
```


Residuals vs Fitted
Pearson Residuals
Predicted values
glm(probs2b ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs)
49
43
61
6
4
2
0
−2
−4
−3
−2
−1
0
1


Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
49
61
35
Theoretical Quantiles
glm(probs2b ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
49
43
61
−4
−3
−2
−1
0
1
Predicted values
glm(probs2b ~ NGOgroup + Gendergp + b_Adas_DelayedRecall + Deprs)




```
pchisq(56.187, df = 59, lower.tail = F)
```

```
## [1] 0.5798556
```

Checking accuracy – the residual plots look good; just look at the outliers and see if anything changes

Also, use a different measure of accuracy. Weight the 1 == 1 stronger

Maximize the specificity

```
p_seq <- seq(0.01, 0.99, 0.01)

acc <- rep(0, length(p_seq))

length(ag_mod3$fitted.values)
```

```
## [1] 542
```

```
for (i in 1:length(p_seq)) {
  preds <- ag_mod3$fitted.values > p_seq[i]
  actuals <- reduced_df$y

  comps <- preds == actuals

  acc[i] <- sum(comps) / length(comps)
}
```

```
max(acc)
```

```
## [1] 0.8800738
```

```
p_seq[which.max(acc)]
```

```
## [1] 0.28
```

```
sum(ag_mod3$fitted.values > p_seq[which.max(acc)])
```

```
## [1] 16
```

```
preds2 <- ag_mod3$fitted.values > p_seq[which.max(acc)]
sum(preds2 == 0 & actuals == 0)
```

```
## [1] 468
```

```
sum(actuals == 0)
```

```
## [1] 475
```

```
length(preds2)
```

```
## [1] 542
```

```
length(actuals)
```

```
## [1] 542
```

```
cv.glm(data = reduced_df, glmfit = best_mod2b, K = 10)$delta
```

```
## [1] 0.1010965 0.1009189
```

‘‘‘