2

1. Summaries of female and male data(fdata and mdata) and boxplots

> summary(fdata)

respid gender app\_1

Min. : 3.0 Min. :1 Min. : 0.00361

1st Qu.:117.0 1st Qu.:1 1st Qu.: 44.88954

Median :229.0 Median :1 Median : 74.27674

Mean :239.6 Mean :1 Mean : 78.68623

3rd Qu.:359.5 3rd Qu.:1 3rd Qu.: 98.90616

Max. :497.0 Max. :1 Max. :271.41888

app\_2 app\_3 app\_4

Min. : 0.00329 Min. : 0.00515 Min. : 0.0026

1st Qu.:34.15764 1st Qu.:61.82868 1st Qu.: 44.6736

Median :44.11661 Median :75.22831 Median : 58.2702

Mean :46.64197 Mean :72.68634 Mean : 61.1701

3rd Qu.:63.60158 3rd Qu.:87.26157 3rd Qu.: 77.2786

Max. :79.72125 Max. :99.79435 Max. :151.8395

app\_5 app\_6

Min. : 0.0001 Min. : 0.077

1st Qu.: 4.9148 1st Qu.: 12.573

Median : 13.3159 Median : 68.525

Mean : 45.0501 Mean : 959.442

3rd Qu.: 60.7297 3rd Qu.: 288.098

Max. :363.2830 Max. :26858.599

summary(mdata)

respid gender app\_1

Min. : 1.0 Min. :0 Min. : 0.004

1st Qu.:129.0 1st Qu.:0 1st Qu.: 64.254

Median :258.0 Median :0 Median : 97.618

Mean :255.2 Mean :0 Mean :111.979

3rd Qu.:384.0 3rd Qu.:0 3rd Qu.:138.517

Max. :500.0 Max. :0 Max. :685.574

app\_2 app\_3 app\_4

Min. : 0.00427 Min. : 0.00552 Min. : 0.0022

1st Qu.:51.86456 1st Qu.:62.94681 1st Qu.: 47.5294

Median :68.05846 Median :74.40698 Median : 73.3996

Mean :66.84345 Mean :72.94448 Mean : 79.1700

3rd Qu.:84.84937 3rd Qu.:88.20217 3rd Qu.:102.1121

Max. :99.73769 Max. :99.89231 Max. :347.4089

app\_5 app\_6

Min. : 0.000 Min. : 0.001

1st Qu.: 3.211 1st Qu.: 13.579

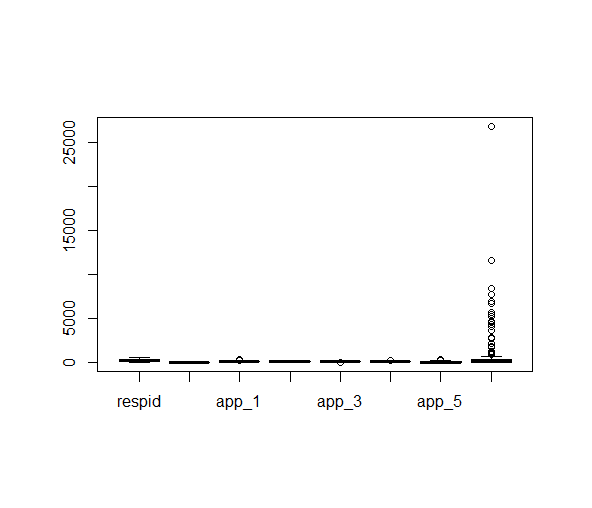
Median : 18.981 Median : 74.540

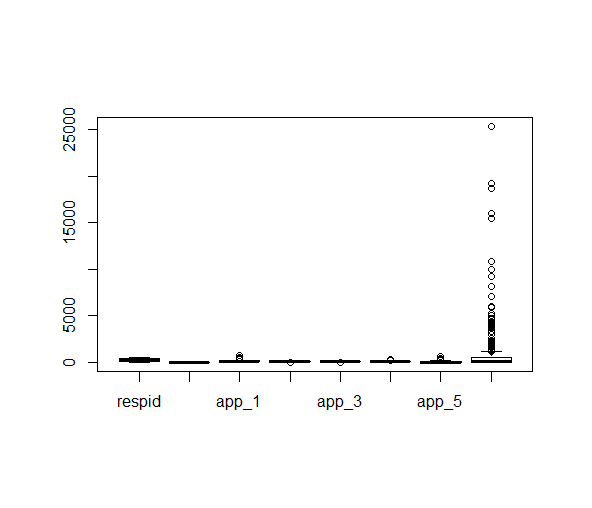
Mean : 45.461 Mean : 881.136

3rd Qu.: 61.167 3rd Qu.: 476.252

Max. :566.766 Max. :25364.209

There is a large difference in mean usage between genders in app1, and app6, and no differerence in app5 with little difference in the rest of the app usages. There is a huge outlier in app 6 male data use, which suggests that app6 data may have to be modified or thrown out all together





2b.

> summary(appdata1.logit)

Call:

glm(formula = gender ~ app1 + app2 + app3 + app4 + app5 + app6,

family = binomial, data = appdata1)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5639 -0.8180 0.4202 0.7965 1.7078

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -9.319e-01 4.279e-01 -2.178 0.029426 \*

app1 7.912e-03 2.179e-03 3.630 0.000283 \*\*\*

app2 5.332e-02 6.511e-03 8.189 2.63e-16 \*\*\*

app3 -3.577e-02 6.727e-03 -5.318 1.05e-07 \*\*\*

app4 9.628e-03 3.238e-03 2.973 0.002946 \*\*

app5 -1.247e-03 1.655e-03 -0.753 0.451223

app6 1.944e-05 3.794e-05 0.512 0.608340

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 612.55 on 499 degrees of freedom

Residual deviance: 476.85 on 493 degrees of freedom

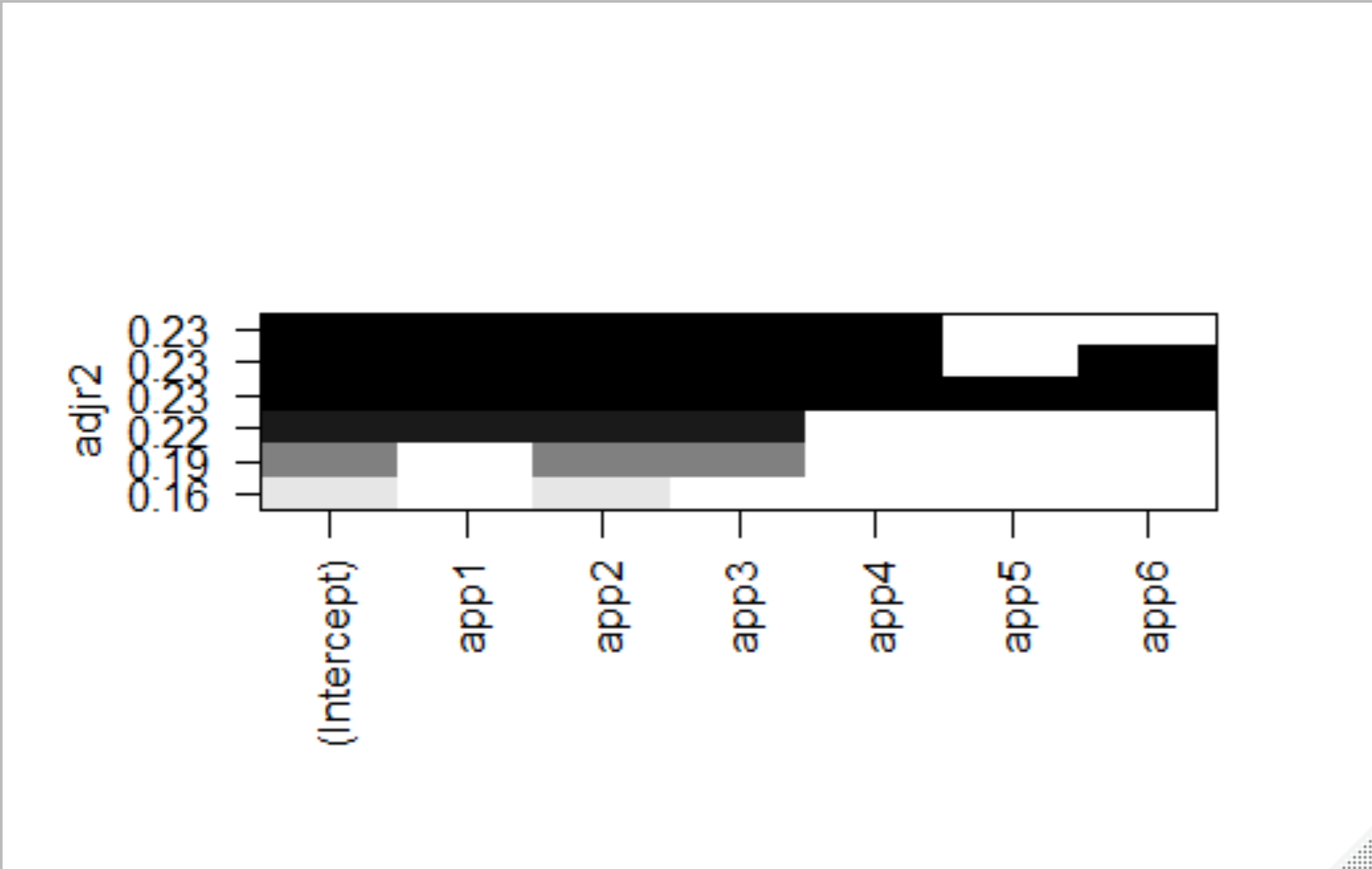
AIC: 490.85

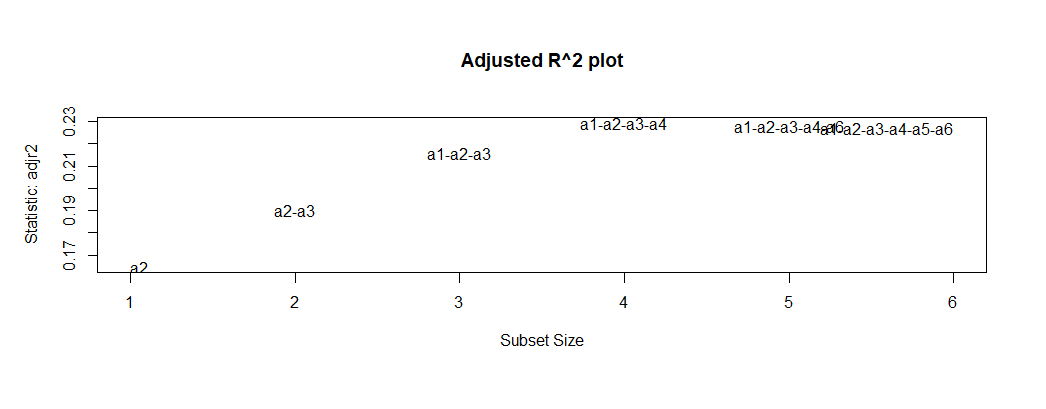
Number of Fisher Scoring iterations: 5

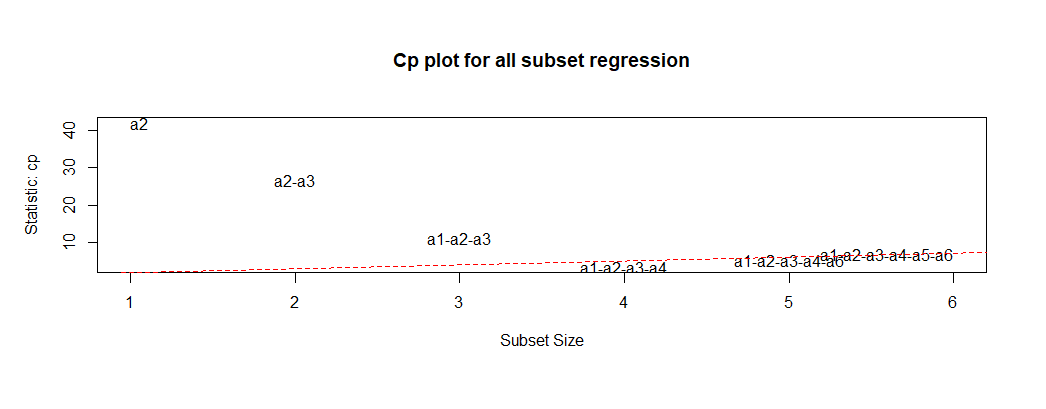
Fitted Response function:

Pi(X)=-9.31e-01+7.912e-03X1+5.332e-02X2-3.577e-02X3+9.628e-03X4-1.247e-03X5+1.944e-05X6

Begin subset modelling to find the best model:







Subset selection object

Call: regsubsets.formula(g ~ app1 + app2 + app3 + app4 + app5 + app6,

data = appdata1, nvmax = 6, nbest = 6)

6 Variables (and intercept)

Forced in Forced out

app1 FALSE FALSE

app2 FALSE FALSE

app3 FALSE FALSE

app4 FALSE FALSE

app5 FALSE FALSE

app6 FALSE FALSE

6 subsets of each size up to 6

Selection Algorithm: exhaustive

app1 app2 app3 app4 app5 app6

1 ( 1 ) " " "\*" " " " " " " " "

1 ( 2 ) "\*" " " " " " " " " " "

1 ( 3 ) " " " " " " "\*" " " " "

1 ( 4 ) " " " " " " " " " " "\*"

1 ( 5 ) " " " " "\*" " " " " " "

1 ( 6 ) " " " " " " " " "\*" " "

2 ( 1 ) " " "\*" "\*" " " " " " "

2 ( 2 ) "\*" "\*" " " " " " " " "

2 ( 3 ) " " "\*" " " "\*" " " " "

2 ( 4 ) " " "\*" " " " " "\*" " "

2 ( 5 ) " " "\*" " " " " " " "\*"

2 ( 6 ) "\*" " " " " "\*" " " " "

3 ( 1 ) "\*" "\*" "\*" " " " " " "

3 ( 2 ) " " "\*" "\*" "\*" " " " "

3 ( 3 ) "\*" "\*" " " "\*" " " " "

3 ( 4 ) " " "\*" "\*" " " " " "\*"

3 ( 5 ) " " "\*" "\*" " " "\*" " "

3 ( 6 ) "\*" "\*" " " " " "\*" " "

4 ( 1 ) "\*" "\*" "\*" "\*" " " " "

4 ( 2 ) "\*" "\*" "\*" " " " " "\*"

4 ( 3 ) "\*" "\*" "\*" " " "\*" " "

4 ( 4 ) " " "\*" "\*" "\*" " " "\*"

4 ( 5 ) " " "\*" "\*" "\*" "\*" " "

4 ( 6 ) "\*" "\*" " " "\*" "\*" " "

5 ( 1 ) "\*" "\*" "\*" "\*" " " "\*"

5 ( 2 ) "\*" "\*" "\*" "\*" "\*" " "

5 ( 3 ) "\*" "\*" "\*" " " "\*" "\*"

5 ( 4 ) " " "\*" "\*" "\*" "\*" "\*"

5 ( 5 ) "\*" "\*" " " "\*" "\*" "\*"

5 ( 6 ) "\*" " " "\*" "\*" "\*" "\*"

6 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*"

> cbind(res.sum$which, res.sum$adjr2,res.sum$cp, res.sum$bic)

(Intercept) app1 app2 app3 app4 app5 app6

1 1 0 1 0 0 0 0 0.164740219 42.082531 -78.580037

1 1 1 0 0 0 0 0 0.044704714 119.410578 -11.441186

1 1 0 0 0 1 0 0 0.035323481 125.454060 -6.555017

1 1 0 0 0 0 0 1 -0.001831556 149.389700 12.341147

1 1 0 0 1 0 0 0 -0.001973268 149.480991 12.411869

1 1 0 0 0 0 1 0 -0.002000310 149.498413 12.425363

2 1 0 1 1 0 0 0 0.190107618 26.692932 -88.791162

2 1 1 1 0 0 0 0 0.183690786 30.818417 -84.845239

2 1 0 1 0 1 0 0 0.176386442 35.514498 -80.391132

2 1 0 1 0 0 1 0 0.163230497 43.972669 -72.467526

2 1 0 1 0 0 0 1 0.163093039 44.061042 -72.385397

2 1 1 0 0 1 0 0 0.069818358 104.028846 -19.551909

3 1 1 1 1 0 0 0 0.215787639 11.168378 -99.694366

3 1 0 1 1 1 0 0 0.207574506 16.438110 -94.485048

3 1 1 1 0 1 0 0 0.192128828 26.348406 -84.832989

3 1 0 1 1 0 0 1 0.188604018 28.610004 -82.656192

3 1 0 1 1 0 1 0 0.188477794 28.690992 -82.578416

3 1 1 1 0 0 1 0 0.182700052 32.398122 -79.031211

4 1 1 1 1 1 0 0 0.229377250 3.452103 -103.229289

4 1 1 1 1 0 0 1 0.214571751 12.932494 -93.714211

4 1 1 1 1 0 1 0 0.214422301 13.028191 -93.619081

4 1 0 1 1 1 0 1 0.206101602 18.356176 -88.351020

4 1 0 1 1 1 1 0 0.206020669 18.408000 -88.300051

4 1 1 1 0 1 1 0 0.191355386 27.798606 -79.149000

5 1 1 1 1 1 0 1 0.228163121 5.231102 -97.238664

5 1 1 1 1 1 1 0 0.228158890 5.233806 -97.235924

5 1 1 1 1 0 1 1 0.213217874 14.781644 -87.649596

5 1 0 1 1 1 1 1 0.204545575 20.323550 -82.168503

5 1 1 1 0 1 1 1 0.189721473 29.796676 -72.936256

5 1 1 0 1 1 1 1 0.068626635 107.180567 -3.295163

6 1 1 1 1 1 1 1 0.226959904 7.000000 -91.258385

Start: AIC=614.55

g ~ 1

Df Deviance AIC

+ app2 1 525.47 529.47

+ app1 1 582.87 586.87

+ app4 1 591.42 595.42

<none> 612.55 614.55

+ app6 1 612.46 616.46

+ app3 1 612.53 616.53

+ app5 1 612.55 616.55

Step: AIC=529.47

g ~ app2

Df Deviance AIC

+ app3 1 505.86 511.86

+ app1 1 513.32 519.32

+ app4 1 518.29 524.29

<none> 525.47 529.47

+ app5 1 525.34 531.34

+ app6 1 525.42 531.42

Step: AIC=511.86

g ~ app2 + app3

Df Deviance AIC

+ app1 1 487.32 495.32

+ app4 1 493.75 501.75

<none> 505.86 511.86

+ app6 1 505.79 513.79

+ app5 1 505.86 513.86

Step: AIC=495.32

g ~ app2 + app3 + app1

Df Deviance AIC

+ app4 1 477.65 487.65

<none> 487.32 495.32

+ app5 1 487.04 497.04

+ app6 1 487.10 497.10

Step: AIC=487.65

g ~ app2 + app3 + app1 + app4

Df Deviance AIC

<none> 477.65 487.65

+ app5 1 477.11 489.11

+ app6 1 477.40 489.40

2b. Highest R^2 and lowest value of BIC

From subset modelling, we can see that the best model is pi(x)=app2+app3+app1+app4. To further test for the ideal model, test for outliers and several other diagnostics, including Wald and ChiSq tests below:

Shapiro-Wilk normality test

data: res

W = 0.90162, p-value < 2.2e-16

studentized Breusch-Pagan test

data: appdata1.logit

BP = 97.989, df = 6, p-value < 2.2e-16

Outlier test on original model.

> outlierTest(appdata1.logit)

No Studentized residuals with Bonferroni p < 0.05

Largest |rstudent|:

rstudent unadjusted p-value Bonferroni p

97 -2.592035 0.009541

No Studentized residuals with Bonferroni p < 0.05

Largest |rstudent|:

rstudent unadjusted p-value Bonferroni p

97 -2.564587 0.01033 NA

> summary(appdata1.logit2)

Call:

glm(formula = gender ~ app1 + app2 + app3 + app4, family = binomial,

data = appdata1)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5384 -0.8161 0.4211 0.7883 1.6897

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.948547 0.428058 -2.216 0.02670 \*

app1 0.007674 0.002151 3.568 0.00036 \*\*\*

app2 0.053132 0.006495 8.180 2.84e-16 \*\*\*

app3 -0.035501 0.006670 -5.322 1.02e-07 \*\*\*

app4 0.009475 0.003229 2.934 0.00334 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

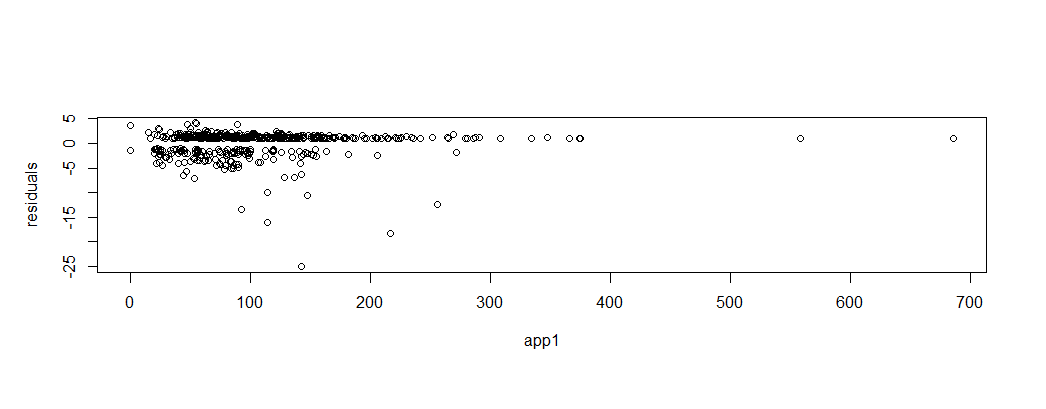
(Dispersion parameter for binomial family taken to be 1)

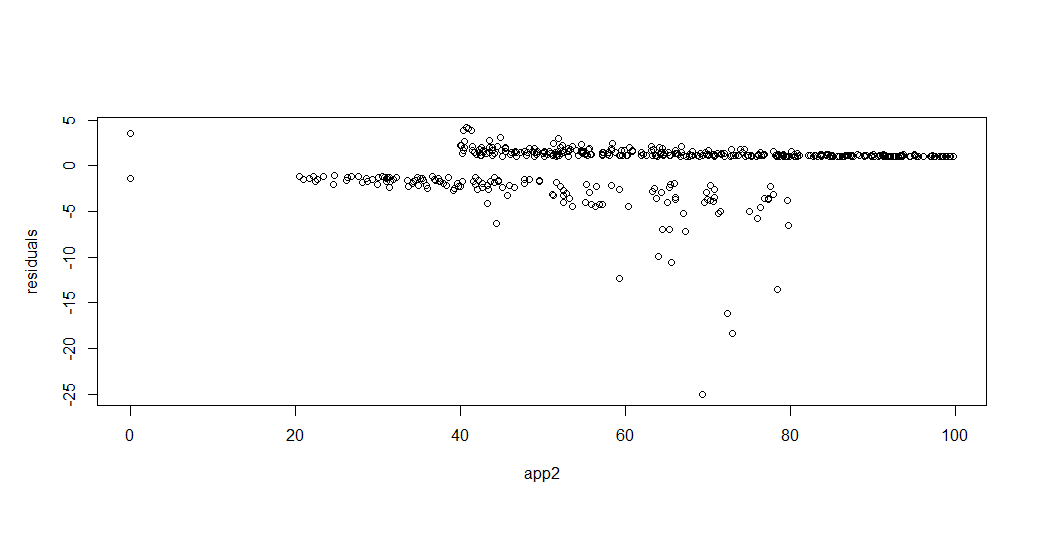
Null deviance: 612.55 on 499 degrees of freedom

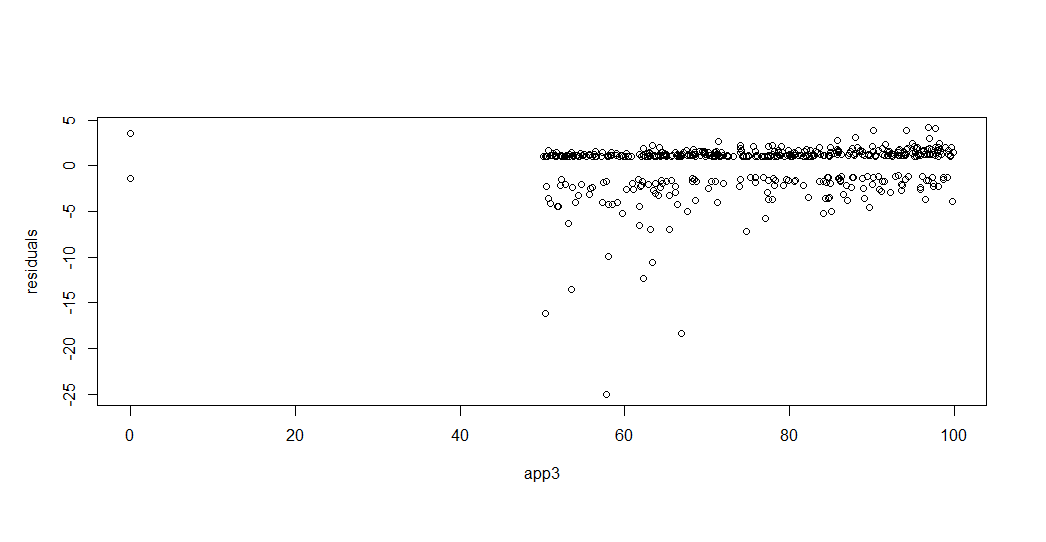
Residual deviance: 477.65 on 495 degrees of freedom

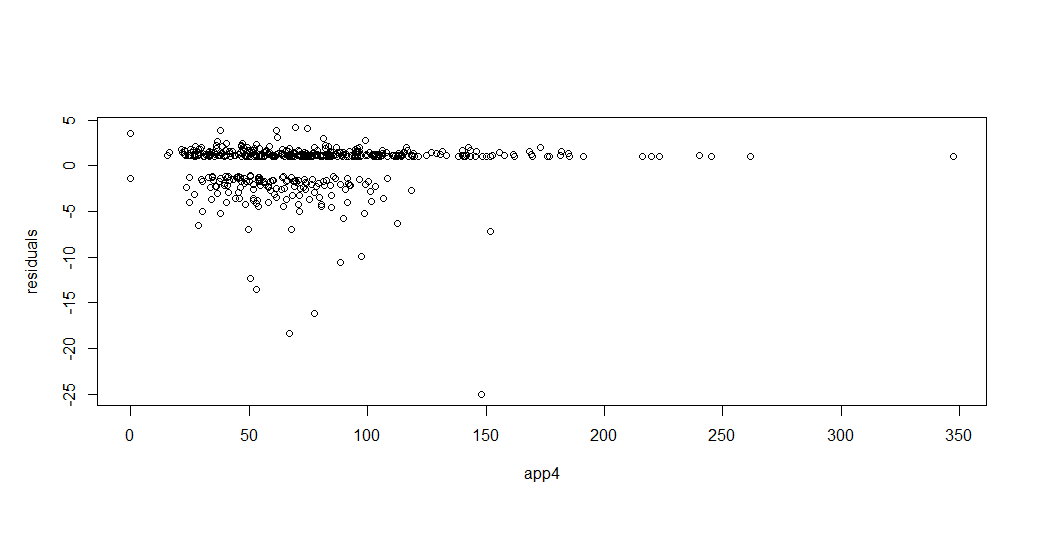
AIC: 487.65

Number of Fisher Scoring iterations: 5







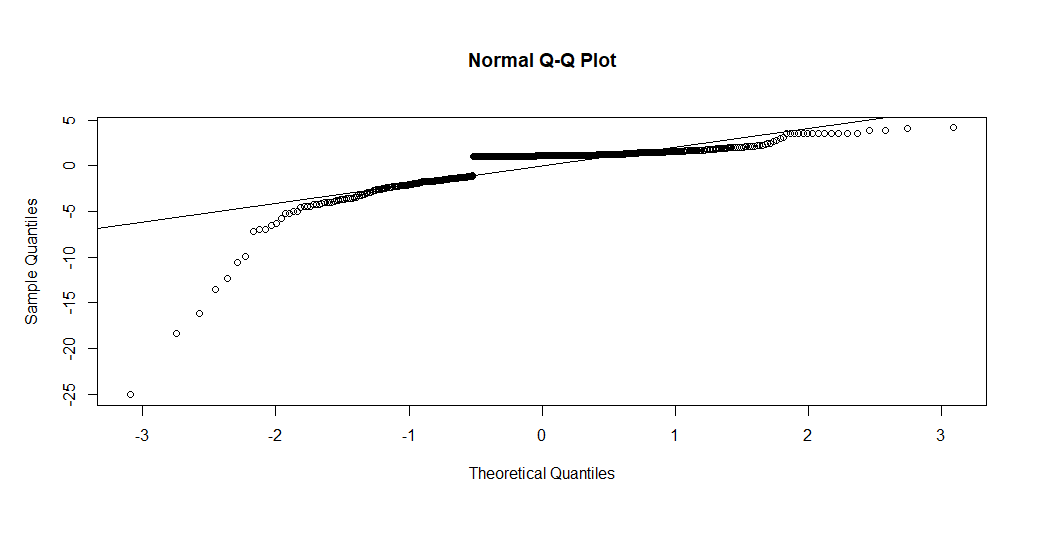


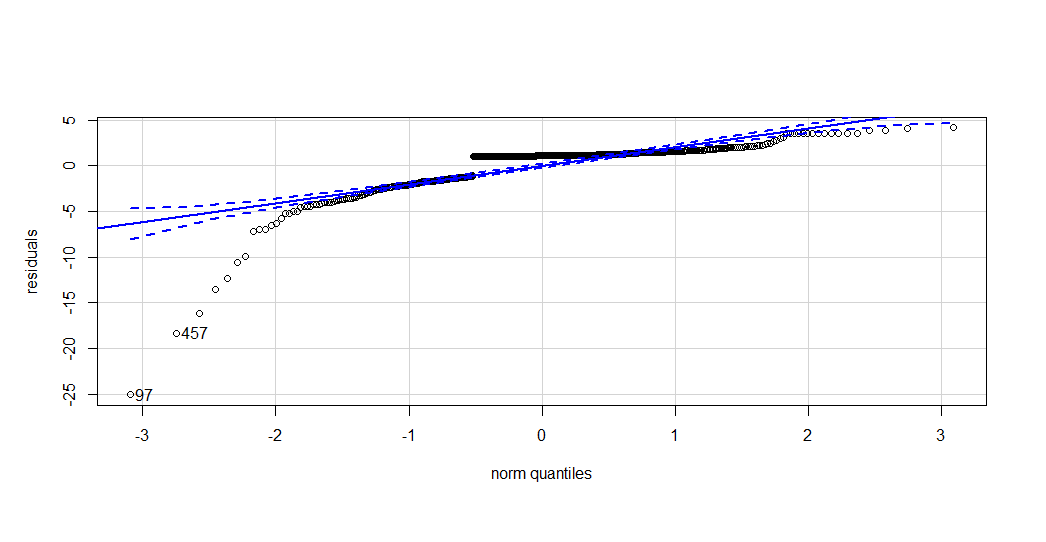
> shapiro.test(residuals)

Shapiro-Wilk normality test

data: residuals

W = 0.68277, p-value < 2.2e-16





> bptest(appdata1.logit2)

studentized Breusch-Pagan test

data: appdata1.logit2

BP = 98.444, df = 4, p-value < 2.2e-16

> anova(appdata1.logit2, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: gender

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 499 612.55

app1 1 29.683 498 582.87 5.089e-08 \*\*\*

app2 1 69.550 497 513.32 < 2.2e-16 \*\*\*

app3 1 25.996 496 487.32 3.422e-07 \*\*\*

app4 1 9.676 495 477.65 0.001867 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> drop1(appdata1.logit2, test= "Chisq")

Single term deletions

Model:

gender ~ app1 + app2 + app3 + app4

Df Deviance AIC LRT Pr(>Chi)

<none> 477.65 487.65

app1 1 493.75 501.75 16.107 5.987e-05 \*\*\*

app2 1 564.33 572.33 86.682 < 2.2e-16 \*\*\*

app3 1 507.65 515.65 30.005 4.309e-08 \*\*\*

app4 1 487.32 495.32 9.676 0.001867 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’

> wald.test(b=coef(appdata1.logit2),Sigma=vcov(appdata1.logit2), Terms=5)

Wald test:

----------

Chi-squared test:

X2 = 8.6, df = 1, P(> X2) = 0.0033

Since all of our tests have sig. P values without any outliers, the model is sufficient. The wald test and LRT statistics are both the most valuable, as the earlier statistics in the question do not translate well to logistic models.

c.

To predict gender, let C be the number of concordant pairs. As a rule of thumb and to measure the predictive power of the model, we want C>.8

For our rule of determining gender, if our predicted value is over .5 it is female, if it is under .5 it is male

> pred.prob

1 2 3 4 5

0.720833648 0.720801787 0.720790238 0.720773472 0.720787801

6 7 8 9 10

0.720765609 0.720774110 0.720815033 0.720773511 0.720738867

11 12 13 14 15

0.720786286 0.720753480 0.720775693 0.720784663 0.720721292

16 17 18 19 20

0.720776107 0.720778652 0.720744999 0.528275388 0.021333844

21 22 23 24 25

0.467980655 0.473595440 0.792169519 0.814173350 0.880548306

26 27 28 29 30

0.247593033 0.354002675 0.401718432 0.908366053 0.387402874

31 32 33 34 35

0.674360084 0.665014482 0.580632983 0.637333086 0.260455026

36 37 38 39 40

0.446477644 0.794002703 0.588848671 0.443118460 0.226077624

41 42 43 44 45

0.704218651 0.101200105 0.272073552 0.072252086 0.348001479

46 47 48 49 50

0.299267178 0.348737632 0.455231509 0.309516018 0.650529095

51 52 53 54 55

0.537462953 0.492864643 0.079711851 0.826301997 0.088019212

56 57 58 59 60

0.139103796 0.444594398 0.803178974 0.135622596 0.251339574

61 62 63 64 65

0.499973276 0.515450071 0.136329808 0.894370209 0.490107842

66 67 68 69 70

0.606106049 0.352574412 0.700543241 0.625587240 0.096701392

71 72 73 74 75

0.324805430 0.154042996 0.822934379 0.186042408 0.263787562

76 77 78 79 80

0.511815113 0.682040066 0.450646243 0.060236327 0.260403168

81 82 83 84 85

0.133218326 0.394590302 0.173402718 0.198418603 0.483039808

86 87 88 89 90

0.142954923 0.036591172 0.742449493 0.157300584 0.328006603

91 92 93 94 95

0.334055865 0.457759583 0.146322706 0.276573174 0.550818693

96 97 98 99 100

0.672805228 0.338488964 0.096696716 0.318203955 0.373998078

101 102 103 104 105

0.140127775 0.314007879 0.237712291 0.168843428 0.797510299

106 107 108 109 110

0.303403305 0.565062743 0.760119963 0.753587068 0.363053740

111 112 113 114 115

0.146358989 0.507330006 0.599928994 0.743744229 0.099990578

116 117 118 119 120

0.542034503 0.805014228 0.572367781 0.632434458 0.128604409

121 122 123 124 125

0.145909480 0.280379545 0.548942390 0.704682617 0.287374257

126 127 128 129 130

0.068047194 0.365686452 0.403264769 0.317767879 0.179429753

131 132 133 134 135

0.359433193 0.195713182 0.136294817 0.107836258 0.414042885

136 137 138 139 140

0.134153110 0.603393016 0.026821875 0.375135329 0.324465675

141 142 143 144 145

0.176683224 0.273982568 0.451577845 0.281270354 0.623006914

146 147 148 149 150

0.142095972 0.387306080 0.238352177 0.419358390 0.469951359

151 152 153 154 155

0.174605275 0.544267880 0.029903044 0.584191203 0.048500544

156 157 158 159 160

0.389584360 0.281519188 0.097247709 0.092155983 0.643960488

161 162 163 164 165

0.448143274 0.202236192 0.086884533 0.020746879 0.282123187

166 167 168 169 170

0.032400552 0.569531796 0.165800845 0.189522256 0.331942364

171 172 173 174 175

0.161911479 0.193471055 0.050061688 0.509011043 0.311182818

176 177 178 179 180

0.218128480 0.165590288 0.489281808 0.538873543 0.034501308

181 182 183 184 185

0.076307919 0.796406408 0.387787857 0.300415329 0.061745252

186 187 188 189 190

0.779352096 0.064523717 0.145142760 0.121191712 0.250214179

191 192 193 194 195

0.583450321 0.373076877 0.269051320 0.327135393 0.234180628

196 197 198 199 200

0.410694115 0.525413583 0.511911252 0.110363739 0.290479500

Using our approach, we have 135 males and 65 females.

d. The set is comparable to the number ratio of males and females in the first set with similar trends in data use.