R Notebook

#load packages  
library(pacman)  
p\_load(boot, car, dplyr, ggplot2, Metrics, cowplot, tree,   
 caret, randomForest, gbm, ElemStatLearn, magrittr, tidyverse, dplyr, haven, Hmisc, psych, janitor, corrplot, reshape, vcd, sfsmisc, DHARMa, e1071, MASS, fmsb)

#load datasets  
camdel <- read\_spss("Cambridge\_delinquency\_with\_caseid.sav")  
conviction <- read\_spss("conviction\_data.sav")  
str(conviction)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 2877 obs. of 3 variables:  
## $ icpsr\_seq\_id\_number: num 1 2 3 4 5 6 7 8 9 10 ...  
## ..- attr(\*, "format.spss")= chr "F8.2"  
## $ agecat : chr "convicted\_10\_13" "convicted\_10\_13" "convicted\_10\_13" "convicted\_10\_13" ...  
## ..- attr(\*, "format.spss")= chr "A21"  
## $ convicted : num 1 1 1 1 1 1 1 1 2 1 ...  
## ..- attr(\*, "format.spss")= chr "F8.2"

Question 1: Merge the two data files,select the variables and cases you need for the later analysis. Rename variables so that they are identifiable and usable

#create and rename new data file with data in a wide format to make the merge file easier as there will common columns amongst the data set  
conviction\_wide <- spread(conviction, key=agecat, value=convicted)#moving from long to wide data format  
colnames(conviction\_wide)[1:8] <- c("v4", "v5", "v11", "v6", "v7", "v8", "v10", "v9")#renaming variables in conviction file to variable number according to code book  
#adding data labels to conviction data  
var.labels <- c(v4="ICPSR SEQ ID NUMBER", v5="CONVICTED 10-13", v11="CONVICTED 10-24", v6="CONVICTED 14-16", v7="CONVICTED 17-20", v8="CONVICTED 21-24",   
v10="CONVICTED AS ADULT 17-24",v9="CONVICTED AS JUVENILE 10-16")  
attr(conviction\_wide, "variable.labels") <- var.labels  
  
#joining data sets by v4 and assigning this to a new data frame  
full\_camdel <- camdel %>% left\_join(conviction\_wide)

## Joining, by = "v4"

## Warning: Column `v4` has different attributes on LHS and RHS of join

Selecting important variables to analyse: According to Kazemian(2011), measures in infancy is not associated with neglect and delinquency while Krohn (2006) looks at predictors of offending where infancy measures are not as important as antisocial child behavior, impulsivity, low intelligence or attainment, family criminality, poverty and poor parental child-rearing behavior. This analysis will aim to model Krohn’s predictors on the data set. #variables of interest are: 4 ICPSR SEQUENTIAL IDENTIFICATION NUMBER 9 CONVICTED AS JUVENILE 10-16 10 CONVICTED AS ADULT 17-24 28 NUMBER OF JUVENILE CONVICTIONS 29 NUMBER OF ADULT CONVICTIONS 39 ADVENTUROUSNESS OF BOY 42 ACTING OUT 53 CONDUCT DISORDER OF BOY AGES 8 THROUGH 9 62 DISCIPLINE QUALITY OF FATHER 63 DISCIPLINE QUALITY OF MOTHER 108 OBEDIENCE OF BOY 117 OUTGOING OR WITHDRAWN BOY 119 PROGRESSIVE MATRICES IQ 132 RULES OF PARENTS 138 SOCIOECONOMIC STATUS OF FAMILY AGES 24 THROUGH 25 847 CURRENT EMPLOYMENT STATUS 858 LIVING CIRCUMSTANCES

#These variables will look at the relationship between factors at age 8-9 vs life at 24-25.  
camdel\_selected <- dplyr::select(full\_camdel, c("v4" ,"v9" ,"v10" ,"v28" ,"v29" ,"v39","v42" ,"v53" ,"v62" ,"v63" ,"v108","v117" ,"v119" ,"v132" ,"v138" ,"v847" ,"v858"))  
  
var.labels <- c(v4="ICPSR SEQ ID NUMBER", v9="CONVICTED AS JUVENILE 10-16", v10="CONVICTED AS ADULT 17-24")  
attr(camdel\_selected, "variable.labels") <- var.labels #adding variable labels to see what each variable actually is  
  
#renaming variables for easier understanding when doing analyses  
camdel\_selected %<>%   
rename(c("v4"="SEQ\_ID\_NUMBER", "v9"="CONVICTED\_AS\_JUVENILE", "v10"="CONVICTED\_AS\_ADULT", "v28"="NO\_JUVENILE\_CONVICTIONS", "v29"="NO\_ADULT\_CONVICTIONS", "v39"="ADVENTUROUSNESS", "v42"="ACTING\_OUT", "v53"="CONDUCT\_DISORDER", "v62"="DISCIPLINE\_FATHER", "v63"="DISCIPLINE\_MOTHER", "v108"="OBEDIENCE", "v117"="OUTGOINGorWITHDRAWN", "v119"="IQ", "v132"="PARENT\_RULES", "v138"="ses", "v847"="CURRENT\_EMPLOYMENT", "v858"="CIRCUMSTANCES"))

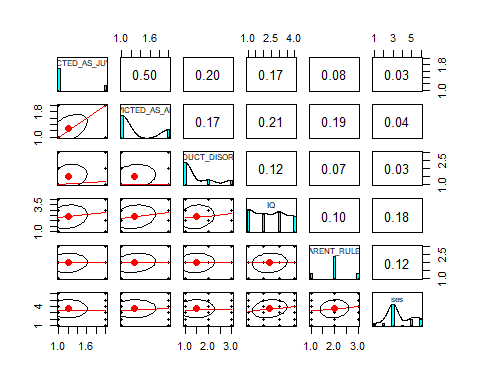
Question 2: Explore the data set, bearing in mind that our key question concerns what the early life determinants of criminality are. What is worth exploring further?

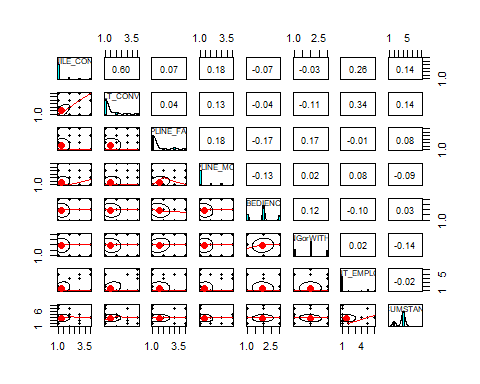
glimpse(camdel\_selected)

## Observations: 411  
## Variables: 17  
## $ SEQ\_ID\_NUMBER <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,...  
## $ CONVICTED\_AS\_JUVENILE <dbl> 1, 1, 2, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2,...  
## $ CONVICTED\_AS\_ADULT <dbl> 1, 1, 1, 2, 2, 1, 1, 1, 2, 2, 1, 2, 2,...  
## $ NO\_JUVENILE\_CONVICTIONS <dbl+lbl> 1, 1, 2, 2, 1, 1, 1, 1, 2, 1, 1, 1...  
## $ NO\_ADULT\_CONVICTIONS <dbl+lbl> 1, 1, 1, 2, 3, 1, 1, 1, 2, 4, 1, 2...  
## $ ADVENTUROUSNESS <dbl+lbl> 2, 1, 2, 3, 2, 1, 2, 2, 2, 2, 2, N...  
## $ ACTING\_OUT <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...  
## $ CONDUCT\_DISORDER <dbl+lbl> 1, 1, 3, 1, 1, 1, 1, 1, 1, 2, 1, N...  
## $ DISCIPLINE\_FATHER <dbl+lbl> 1, NA, 1, 3, 1, 1, 1, 1, 1, NA, 1,...  
## $ DISCIPLINE\_MOTHER <dbl+lbl> 2, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, N...  
## $ OBEDIENCE <dbl+lbl> 2, 2, 2, 1, 2, 3, 2, 2, 3, 3, 1, N...  
## $ OUTGOINGorWITHDRAWN <dbl+lbl> 1, 2, 1, 1, 2, 2, 2, 2, 2, 1, 1, N...  
## $ IQ <dbl+lbl> 2, 3, 3, 4, 4, 4, 4, 2, 4, 1, 1, 2...  
## $ PARENT\_RULES <dbl+lbl> 2, 1, 1, 3, 2, 2, 2, 2, 2, 2, 2, N...  
## $ ses <dbl+lbl> 3, 5, 3, 3, 3, 3, 3, 3, 6, 4, 3, 5...  
## $ CURRENT\_EMPLOYMENT <dbl+lbl> NA, NA, NA, 1, NA, NA, NA, NA, NA,...  
## $ CIRCUMSTANCES <dbl+lbl> NA, NA, NA, 4, NA, NA, NA, NA, NA,...

From a quick look at the data file, it seems like there are a few missing values that might be problematic later on. Narrowing the variables down has also made dealing with the variables more easily.

#pairs.panels to look at basic correlations between childhood (8-9 years) and background variables.  
camdel\_selected %>%   
 dplyr::select(2, 3, 8, 13, 14, 15) %>%  
 pairs.panels()



camdel\_selected %>%   
 dplyr::select(4, 5, 9, 10, 11, 12, 16, 17) %>%  
 pairs.panels()#breaking up variables as graph becomes unreadable with too many variables inputted

From the graphs above it seems like there is a moderate positive relationship between being convicted as a juvenile and then later on as an adult. It seems like the chosen background variables have a weak correlation with adult convictions. However, this be will be further explored in question 3.

#looking at distribution of individual characteristics at ages 8-9  
cor(camdel\_selected, use="complete.obs", method="kendall")

## SEQ\_ID\_NUMBER CONVICTED\_AS\_JUVENILE  
## SEQ\_ID\_NUMBER 1.000000000 -0.06151109  
## CONVICTED\_AS\_JUVENILE -0.061511095 1.00000000  
## CONVICTED\_AS\_ADULT -0.133519447 0.55761053  
## NO\_JUVENILE\_CONVICTIONS -0.041720188 0.91825250  
## NO\_ADULT\_CONVICTIONS -0.134829843 0.44819339  
## ADVENTUROUSNESS -0.183837146 0.32873559  
## ACTING\_OUT -0.020766851 0.34937494  
## CONDUCT\_DISORDER -0.059741501 0.34157674  
## DISCIPLINE\_FATHER -0.079015841 0.30017299  
## DISCIPLINE\_MOTHER 0.183735542 0.41089889  
## OBEDIENCE -0.079382230 -0.06762522  
## OUTGOINGorWITHDRAWN 0.163692290 -0.19568230  
## IQ -0.012871880 0.19772179  
## PARENT\_RULES -0.240282652 0.02439328  
## ses 0.012530457 -0.18549251  
## CURRENT\_EMPLOYMENT -0.009409755 0.04477363  
## CIRCUMSTANCES -0.146437918 0.20230876  
## CONVICTED\_AS\_ADULT NO\_JUVENILE\_CONVICTIONS  
## SEQ\_ID\_NUMBER -0.133519447 -0.04172019  
## CONVICTED\_AS\_JUVENILE 0.557610532 0.91825250  
## CONVICTED\_AS\_ADULT 1.000000000 0.48383394  
## NO\_JUVENILE\_CONVICTIONS 0.483833936 1.00000000  
## NO\_ADULT\_CONVICTIONS 0.899652711 0.43037331  
## ADVENTUROUSNESS 0.390729605 0.30007081  
## ACTING\_OUT 0.260625009 0.34754894  
## CONDUCT\_DISORDER 0.372841271 0.35846136  
## DISCIPLINE\_FATHER 0.162523728 0.29520125  
## DISCIPLINE\_MOTHER 0.190286653 0.36749712  
## OBEDIENCE 0.008787496 -0.09355953  
## OUTGOINGorWITHDRAWN -0.352208722 -0.18697032  
## IQ 0.190316821 0.21592763  
## PARENT\_RULES 0.137545013 0.01359951  
## ses -0.002421125 -0.15207943  
## CURRENT\_EMPLOYMENT 0.147875855 0.09707353  
## CIRCUMSTANCES 0.142710738 0.17250120  
## NO\_ADULT\_CONVICTIONS ADVENTUROUSNESS ACTING\_OUT  
## SEQ\_ID\_NUMBER -0.13482984 -0.18383715 -0.02076685  
## CONVICTED\_AS\_JUVENILE 0.44819339 0.32873559 0.34937494  
## CONVICTED\_AS\_ADULT 0.89965271 0.39072960 0.26062501  
## NO\_JUVENILE\_CONVICTIONS 0.43037331 0.30007081 0.34754894  
## NO\_ADULT\_CONVICTIONS 1.00000000 0.35750792 0.21556296  
## ADVENTUROUSNESS 0.35750792 1.00000000 0.39435162  
## ACTING\_OUT 0.21556296 0.39435162 1.00000000  
## CONDUCT\_DISORDER 0.33371629 0.50023425 0.72279892  
## DISCIPLINE\_FATHER 0.10722427 0.25700192 0.08257484  
## DISCIPLINE\_MOTHER 0.14308577 0.24242638 0.28653734  
## OBEDIENCE 0.01264911 -0.24925958 -0.30136987  
## OUTGOINGorWITHDRAWN -0.29909015 -0.26905287 -0.18237908  
## IQ 0.12484722 0.22493274 0.10014934  
## PARENT\_RULES 0.10541049 0.16808159 0.12941423  
## ses 0.04864584 0.02066634 0.05359818  
## CURRENT\_EMPLOYMENT 0.21482190 0.07162802 0.30352148  
## CIRCUMSTANCES 0.12923477 0.26354342 -0.04702135  
## CONDUCT\_DISORDER DISCIPLINE\_FATHER  
## SEQ\_ID\_NUMBER -0.0597415 -0.079015841  
## CONVICTED\_AS\_JUVENILE 0.3415767 0.300172989  
## CONVICTED\_AS\_ADULT 0.3728413 0.162523728  
## NO\_JUVENILE\_CONVICTIONS 0.3584614 0.295201252  
## NO\_ADULT\_CONVICTIONS 0.3337163 0.107224269  
## ADVENTUROUSNESS 0.5002342 0.257001918  
## ACTING\_OUT 0.7227989 0.082574836  
## CONDUCT\_DISORDER 1.0000000 0.239328390  
## DISCIPLINE\_FATHER 0.2393284 1.000000000  
## DISCIPLINE\_MOTHER 0.2220173 0.117779020  
## OBEDIENCE -0.3671374 -0.131519190  
## OUTGOINGorWITHDRAWN -0.2539424 0.126159491  
## IQ 0.2000339 0.224756669  
## PARENT\_RULES 0.1823828 -0.030179876  
## ses 0.1089467 0.036236086  
## CURRENT\_EMPLOYMENT 0.2794801 0.004535252  
## CIRCUMSTANCES 0.1480235 0.076407562  
## DISCIPLINE\_MOTHER OBEDIENCE OUTGOINGorWITHDRAWN  
## SEQ\_ID\_NUMBER 0.183735542 -0.079382230 0.16369229  
## CONVICTED\_AS\_JUVENILE 0.410898889 -0.067625223 -0.19568230  
## CONVICTED\_AS\_ADULT 0.190286653 0.008787496 -0.35220872  
## NO\_JUVENILE\_CONVICTIONS 0.367497117 -0.093559525 -0.18697032  
## NO\_ADULT\_CONVICTIONS 0.143085770 0.012649111 -0.29909015  
## ADVENTUROUSNESS 0.242426376 -0.249259578 -0.26905287  
## ACTING\_OUT 0.286537342 -0.301369870 -0.18237908  
## CONDUCT\_DISORDER 0.222017340 -0.367137410 -0.25394237  
## DISCIPLINE\_FATHER 0.117779020 -0.131519190 0.12615949  
## DISCIPLINE\_MOTHER 1.000000000 -0.173234026 -0.03201635  
## OBEDIENCE -0.173234026 1.000000000 0.19473341  
## OUTGOINGorWITHDRAWN -0.032016354 0.194733410 1.00000000  
## IQ -0.075441392 -0.257004544 -0.07057355  
## PARENT\_RULES -0.120745196 -0.051015088 -0.19192854  
## ses -0.277035925 -0.061716342 0.08547104  
## CURRENT\_EMPLOYMENT 0.057062488 -0.025379914 0.07227502  
## CIRCUMSTANCES 0.003683368 0.006837981 -0.09107844  
## IQ PARENT\_RULES ses  
## SEQ\_ID\_NUMBER -0.01287188 -0.24028265 0.012530457  
## CONVICTED\_AS\_JUVENILE 0.19772179 0.02439328 -0.185492505  
## CONVICTED\_AS\_ADULT 0.19031682 0.13754501 -0.002421125  
## NO\_JUVENILE\_CONVICTIONS 0.21592763 0.01359951 -0.152079426  
## NO\_ADULT\_CONVICTIONS 0.12484722 0.10541049 0.048645841  
## ADVENTUROUSNESS 0.22493274 0.16808159 0.020666342  
## ACTING\_OUT 0.10014934 0.12941423 0.053598177  
## CONDUCT\_DISORDER 0.20003388 0.18238285 0.108946744  
## DISCIPLINE\_FATHER 0.22475667 -0.03017988 0.036236086  
## DISCIPLINE\_MOTHER -0.07544139 -0.12074520 -0.277035925  
## OBEDIENCE -0.25700454 -0.05101509 -0.061716342  
## OUTGOINGorWITHDRAWN -0.07057355 -0.19192854 0.085471040  
## IQ 1.00000000 0.12973838 0.228076296  
## PARENT\_RULES 0.12973838 1.00000000 0.048981795  
## ses 0.22807630 0.04898180 1.000000000  
## CURRENT\_EMPLOYMENT 0.15733314 0.22816068 0.107423446  
## CIRCUMSTANCES 0.14346689 0.12057451 -0.013344975  
## CURRENT\_EMPLOYMENT CIRCUMSTANCES  
## SEQ\_ID\_NUMBER -0.009409755 -0.146437918  
## CONVICTED\_AS\_JUVENILE 0.044773626 0.202308759  
## CONVICTED\_AS\_ADULT 0.147875855 0.142710738  
## NO\_JUVENILE\_CONVICTIONS 0.097073528 0.172501201  
## NO\_ADULT\_CONVICTIONS 0.214821902 0.129234774  
## ADVENTUROUSNESS 0.071628019 0.263543418  
## ACTING\_OUT 0.303521483 -0.047021346  
## CONDUCT\_DISORDER 0.279480147 0.148023515  
## DISCIPLINE\_FATHER 0.004535252 0.076407562  
## DISCIPLINE\_MOTHER 0.057062488 0.003683368  
## OBEDIENCE -0.025379914 0.006837981  
## OUTGOINGorWITHDRAWN 0.072275024 -0.091078436  
## IQ 0.157333141 0.143466894  
## PARENT\_RULES 0.228160679 0.120574511  
## ses 0.107423446 -0.013344975  
## CURRENT\_EMPLOYMENT 1.000000000 -0.123793988  
## CIRCUMSTANCES -0.123793988 1.000000000

From the correlations above, it can be seen that there is a moderate positive relationship between being convicted as a juvenile and then later on being convicted as an adult. Conduct disorder also seems to be correlated with boys acting out and their adventurousness. Therefore, for simplicity in question 3, conduct disorder will be used as a proxy for adventurousness and acting out.

Question 3: Model the occurrence and extent of criminality in our participants on the basis of early life events or factors. Interpret your model(s). Write a function to assess the predictive accuracy of your model(s), test it, and apply it to your data.

Build two models for comparison (basic glm(logistic regression) and then an improved model only using significant predictors)

#LDA model predicting whether boys were convicted as an adult based on a combination of their individual behaviours and family influences  
camdel\_model1 <- dplyr::select(camdel\_selected, -c(16, 17 ))#removing adult variables that are not of interest  
  
#splitting data into test and training set  
set.seed(2)  
camdel\_traindata <- sample\_frac(camdel\_model1, 0.75)  
camdel\_testdata <- dplyr::setdiff(camdel\_model1, camdel\_traindata)

## Warning: Column `SEQ\_ID\_NUMBER` has different attributes on LHS and RHS of  
## join

## Warning: Column `NO\_JUVENILE\_CONVICTIONS` has different attributes on LHS  
## and RHS of join

## Warning: Column `NO\_ADULT\_CONVICTIONS` has different attributes on LHS and  
## RHS of join

## Warning: Column `ADVENTUROUSNESS` has different attributes on LHS and RHS  
## of join

## Warning: Column `ACTING\_OUT` has different attributes on LHS and RHS of  
## join

## Warning: Column `CONDUCT\_DISORDER` has different attributes on LHS and RHS  
## of join

## Warning: Column `DISCIPLINE\_FATHER` has different attributes on LHS and RHS  
## of join

## Warning: Column `DISCIPLINE\_MOTHER` has different attributes on LHS and RHS  
## of join

## Warning: Column `OBEDIENCE` has different attributes on LHS and RHS of join

## Warning: Column `OUTGOINGorWITHDRAWN` has different attributes on LHS and  
## RHS of join

## Warning: Column `IQ` has different attributes on LHS and RHS of join

## Warning: Column `PARENT\_RULES` has different attributes on LHS and RHS of  
## join

## Warning: Column `ses` has different attributes on LHS and RHS of join

#looling at descriptives of data  
  
describe(camdel\_traindata, na.rm=TRUE)

## vars n mean sd median trimmed mad min  
## SEQ\_ID\_NUMBER 1 308 211.44 121.04 216.5 212.10 160.12 3  
## CONVICTED\_AS\_JUVENILE 2 308 1.19 0.39 1.0 1.11 0.00 1  
## CONVICTED\_AS\_ADULT 3 297 1.27 0.44 1.0 1.21 0.00 1  
## NO\_JUVENILE\_CONVICTIONS 4 308 1.31 0.74 1.0 1.11 0.00 1  
## NO\_ADULT\_CONVICTIONS 5 294 1.48 0.91 1.0 1.27 0.00 1  
## ADVENTUROUSNESS 6 283 2.03 0.59 2.0 2.04 0.00 1  
## ACTING\_OUT 7 308 1.19 0.40 1.0 1.12 0.00 1  
## CONDUCT\_DISORDER 8 273 1.51 0.77 1.0 1.39 0.00 1  
## DISCIPLINE\_FATHER 9 274 1.60 1.02 1.0 1.38 0.00 1  
## DISCIPLINE\_MOTHER 10 279 1.39 0.77 1.0 1.22 0.00 1  
## OBEDIENCE 11 287 1.95 0.67 2.0 1.94 0.00 1  
## OUTGOINGorWITHDRAWN 12 287 1.99 0.67 2.0 1.99 0.00 1  
## IQ 13 308 2.40 1.14 2.0 2.38 1.48 1  
## PARENT\_RULES 14 279 2.02 0.61 2.0 2.02 0.00 1  
## ses 15 308 3.58 1.44 3.0 3.58 1.48 1  
## max range skew kurtosis se  
## SEQ\_ID\_NUMBER 411 408 -0.04 -1.26 6.90  
## CONVICTED\_AS\_JUVENILE 2 1 1.59 0.52 0.02  
## CONVICTED\_AS\_ADULT 2 1 1.03 -0.93 0.03  
## NO\_JUVENILE\_CONVICTIONS 4 3 2.48 5.32 0.04  
## NO\_ADULT\_CONVICTIONS 4 3 1.75 1.76 0.05  
## ADVENTUROUSNESS 3 2 -0.01 -0.16 0.04  
## ACTING\_OUT 2 1 1.53 0.35 0.02  
## CONDUCT\_DISORDER 3 2 1.09 -0.46 0.05  
## DISCIPLINE\_FATHER 4 3 1.40 0.44 0.06  
## DISCIPLINE\_MOTHER 4 3 1.79 1.97 0.05  
## OBEDIENCE 3 2 0.06 -0.80 0.04  
## OUTGOINGorWITHDRAWN 3 2 0.01 -0.76 0.04  
## IQ 4 3 0.09 -1.41 0.06  
## PARENT\_RULES 3 2 -0.01 -0.31 0.04  
## ses 6 5 0.27 -0.77 0.08

#Build the model1 on the training data set using convicted as an adult as the outcome  
modeltrain <- glm(CONVICTED\_AS\_ADULT ~ CONVICTED\_AS\_JUVENILE +   
 NO\_JUVENILE\_CONVICTIONS + NO\_ADULT\_CONVICTIONS +  
 CONDUCT\_DISORDER + DISCIPLINE\_FATHER +   
 DISCIPLINE\_MOTHER + OBEDIENCE + OUTGOINGorWITHDRAWN +   
 IQ + ses, data = camdel\_traindata)  
summary(modeltrain)

##   
## Call:  
## glm(formula = CONVICTED\_AS\_ADULT ~ CONVICTED\_AS\_JUVENILE + NO\_JUVENILE\_CONVICTIONS +   
## NO\_ADULT\_CONVICTIONS + CONDUCT\_DISORDER + DISCIPLINE\_FATHER +   
## DISCIPLINE\_MOTHER + OBEDIENCE + OUTGOINGorWITHDRAWN + IQ +   
## ses, data = camdel\_traindata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.48487 -0.09308 -0.04582 0.01645 0.58697   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6891518 0.0891323 7.732 3.38e-13 \*\*\*  
## CONVICTED\_AS\_JUVENILE 0.1949813 0.0717888 2.716 0.00711 \*\*   
## NO\_JUVENILE\_CONVICTIONS -0.1686904 0.0409802 -4.116 5.38e-05 \*\*\*  
## NO\_ADULT\_CONVICTIONS 0.4512734 0.0175779 25.673 < 2e-16 \*\*\*  
## CONDUCT\_DISORDER -0.0160986 0.0195646 -0.823 0.41146   
## DISCIPLINE\_FATHER 0.0009877 0.0136272 0.072 0.94228   
## DISCIPLINE\_MOTHER 0.0020603 0.0183711 0.112 0.91080   
## OBEDIENCE -0.0179908 0.0213109 -0.844 0.39944   
## OUTGOINGorWITHDRAWN -0.0322793 0.0197512 -1.634 0.10358   
## IQ 0.0354332 0.0122076 2.903 0.00406 \*\*   
## ses -0.0212977 0.0089967 -2.367 0.01876 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.03922258)  
##   
## Null deviance: 46.3933 on 238 degrees of freedom  
## Residual deviance: 8.9427 on 228 degrees of freedom  
## (69 observations deleted due to missingness)  
## AIC: -83.011  
##   
## Number of Fisher Scoring iterations: 2

M and SD of data seems relatively similar with no large variances.

#creating model2 using number of adult convictions as the outcome being predicted  
modeltrain2 <- glm(NO\_ADULT\_CONVICTIONS ~ CONVICTED\_AS\_ADULT +   
 CONVICTED\_AS\_JUVENILE +   
 NO\_JUVENILE\_CONVICTIONS + NO\_ADULT\_CONVICTIONS +  
 CONDUCT\_DISORDER + DISCIPLINE\_FATHER +   
 DISCIPLINE\_MOTHER + OBEDIENCE + OUTGOINGorWITHDRAWN +   
 IQ + ses, data = camdel\_traindata)

## Warning in model.matrix.default(mt, mf, contrasts): the response appeared  
## on the right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 4 in  
## model.matrix: no columns are assigned

summary(modeltrain2)

##   
## Call:  
## glm(formula = NO\_ADULT\_CONVICTIONS ~ CONVICTED\_AS\_ADULT + CONVICTED\_AS\_JUVENILE +   
## NO\_JUVENILE\_CONVICTIONS + NO\_ADULT\_CONVICTIONS + CONDUCT\_DISORDER +   
## DISCIPLINE\_FATHER + DISCIPLINE\_MOTHER + OBEDIENCE + OUTGOINGorWITHDRAWN +   
## IQ + ses, data = camdel\_traindata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.17410 -0.08754 0.01559 0.08894 1.48408   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.0969702 0.1769363 -6.200 2.62e-09 \*\*\*  
## CONVICTED\_AS\_ADULT 1.6464073 0.0641305 25.673 < 2e-16 \*\*\*  
## CONVICTED\_AS\_JUVENILE -0.3031343 0.1378680 -2.199 0.0289 \*   
## NO\_JUVENILE\_CONVICTIONS 0.4508491 0.0754374 5.976 8.71e-09 \*\*\*  
## CONDUCT\_DISORDER 0.0416650 0.0373234 1.116 0.2655   
## DISCIPLINE\_FATHER -0.0006613 0.0260291 -0.025 0.9798   
## DISCIPLINE\_MOTHER 0.0036656 0.0350901 0.104 0.9169   
## OBEDIENCE 0.0329362 0.0407103 0.809 0.4193   
## OUTGOINGorWITHDRAWN 0.0415484 0.0378467 1.098 0.2734   
## IQ -0.0376663 0.0236129 -1.595 0.1121   
## ses 0.0444224 0.0171437 2.591 0.0102 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.143098)  
##   
## Null deviance: 201.699 on 238 degrees of freedom  
## Residual deviance: 32.626 on 228 degrees of freedom  
## (69 observations deleted due to missingness)  
## AIC: 226.32  
##   
## Number of Fisher Scoring iterations: 2

The first model has a much lower AIC which means that adult convictions is better predicted by individual and family characteristics than number of convictions as an adult. This is similar to Krohn’s(2006) findings which state that the most important independent childhood predictors of offending could be grouped under the headings of antisocial child behavior, impulsivity, low intelligence or attainment, family criminality, poverty and poor parental child-rearing behavior. However, this model indicates that IQ, SES, conviction as a juvenile and number of convictions as a juvenile are significant predictors of adult conviction.

#test for significance of model 1  
1-pchisq(modeltrain$deviance,  
 modeltrain$df.residual)

## [1] 1

The training regression model is not significant which means that the model seems to fit the data

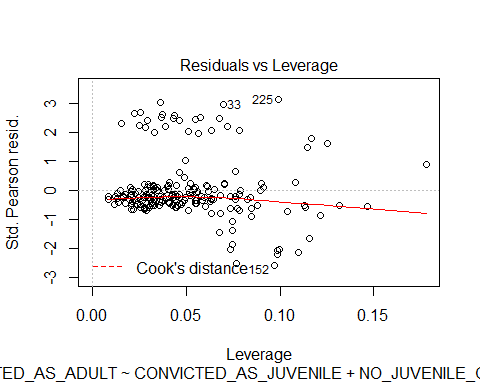
#comparing the training to the test model using anova  
testmodel <- glm(CONVICTED\_AS\_ADULT ~ CONVICTED\_AS\_JUVENILE +   
 NO\_JUVENILE\_CONVICTIONS + NO\_ADULT\_CONVICTIONS +  
 CONDUCT\_DISORDER + DISCIPLINE\_FATHER +   
 DISCIPLINE\_MOTHER + OBEDIENCE + OUTGOINGorWITHDRAWN +   
 IQ + ses, data = camdel\_testdata)  
summary(testmodel)

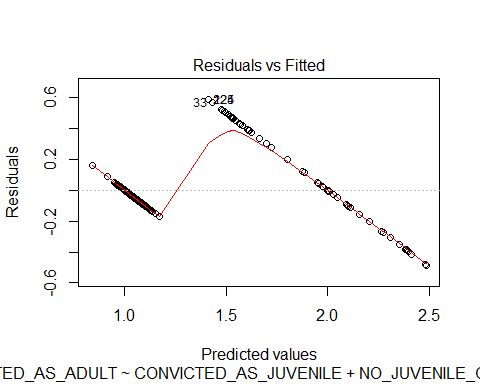
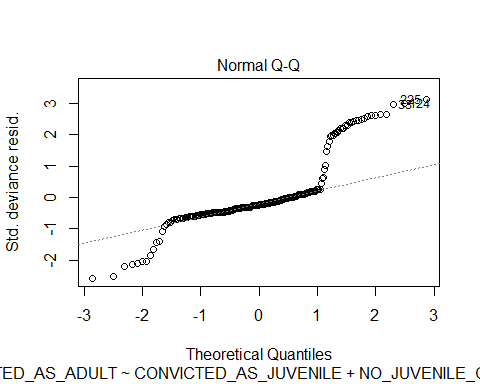
##   
## Call:  
## glm(formula = CONVICTED\_AS\_ADULT ~ CONVICTED\_AS\_JUVENILE + NO\_JUVENILE\_CONVICTIONS +   
## NO\_ADULT\_CONVICTIONS + CONDUCT\_DISORDER + DISCIPLINE\_FATHER +   
## DISCIPLINE\_MOTHER + OBEDIENCE + OUTGOINGorWITHDRAWN + IQ +   
## ses, data = camdel\_testdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.55107 -0.11092 -0.04825 0.03988 0.52056   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.2170941 0.1879115 1.155 0.251841   
## CONVICTED\_AS\_JUVENILE 0.3561060 0.1203628 2.959 0.004196 \*\*   
## NO\_JUVENILE\_CONVICTIONS -0.2549882 0.0738684 -3.452 0.000942 \*\*\*  
## NO\_ADULT\_CONVICTIONS 0.4707995 0.0330491 14.245 < 2e-16 \*\*\*  
## CONDUCT\_DISORDER 0.0817061 0.0398493 2.050 0.044021 \*   
## DISCIPLINE\_FATHER -0.0158942 0.0278273 -0.571 0.569687   
## DISCIPLINE\_MOTHER 0.0199054 0.0319839 0.622 0.535702   
## OBEDIENCE 0.1109127 0.0424796 2.611 0.011008 \*   
## OUTGOINGorWITHDRAWN 0.0373446 0.0405893 0.920 0.360657   
## IQ 0.0005499 0.0223317 0.025 0.980423   
## ses -0.0328496 0.0180473 -1.820 0.072944 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.04401854)  
##   
## Null deviance: 17.3780 on 81 degrees of freedom  
## Residual deviance: 3.1253 on 71 degrees of freedom  
## (21 observations deleted due to missingness)  
## AIC: -11.203  
##   
## Number of Fisher Scoring iterations: 2

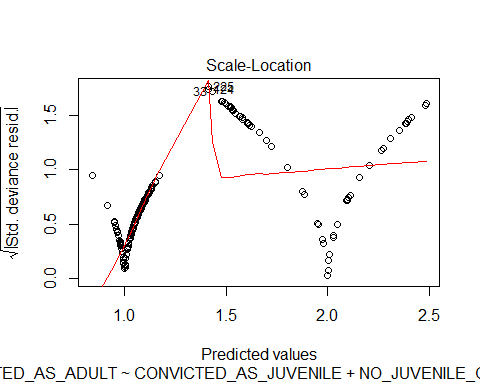
testmodel1 <- predict(testmodel, camdel\_traindata, type = "response")  
anova(modeltrain, testmodel1)

## Analysis of Deviance Table  
##   
## Model: gaussian, link: identity  
##   
## Response: CONVICTED\_AS\_ADULT  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 238 46.393  
## CONVICTED\_AS\_JUVENILE 1 8.7021 237 37.691  
## NO\_JUVENILE\_CONVICTIONS 1 0.9995 236 36.692  
## NO\_ADULT\_CONVICTIONS 1 27.1309 235 9.561  
## CONDUCT\_DISORDER 1 0.0008 234 9.560  
## DISCIPLINE\_FATHER 1 0.0015 233 9.558  
## DISCIPLINE\_MOTHER 1 0.0003 232 9.558  
## OBEDIENCE 1 0.0496 231 9.508  
## OUTGOINGorWITHDRAWN 1 0.1161 230 9.392  
## IQ 1 0.2298 229 9.163  
## ses 1 0.2198 228 8.943

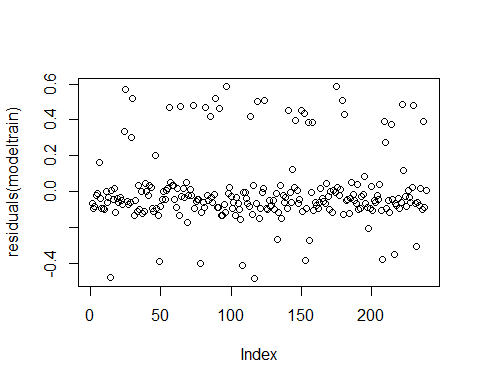
It seems that the AIC score for the test model is actually better than that of the training model. While the anova comparison indicates that predicting the data on the test model has more residual deviance. Therefore, the train model might not actually predict the data well. However, running the residual deviance with a few more iterations of this migt provide a more stable result.

#Diagnostics for logistic model

plot(modeltrain)



plot(residuals(modeltrain))



Residuals seem to be relatively small. However, there do seem to be some high leverage values that are pulling the model up. This may account for the poor predictability earlier on the test data set.

#looking at r squared value for model 1  
NagelkerkeR2(modeltrain)

## $N  
## [1] 239  
##   
## $R2  
## [1] 0.8220356

The r squared value for model 1 indicates that 82% of the response variable variation is explained by a linear model which is good considering the fact that a few variables did not show any correlations to adult conviction.

#creating a function to calculate MSE for modeltrain  
mse = function (avec){  
 (sum(avec$residuals^2))/(length(avec))}  
mse(modeltrain)

## [1] 0.2884757

#running mse function for test model to look at different  
mse(testmodel)

## [1] 0.1008167

The MSE for the test model seems to have less error, which means that this might actually be able to predict the data well which is supported by the R squared value of 0.82 and a lower AIC score. Overall, the model will need a possible transformation for linearity as well as for heterogeneity. Comparing this transformed model to a more robust form of a regression might also be more helpful is determining the prediction power of the model.

Link to GITHUB: