Stat 149 Final Project: Team Theoretical Limit

Eduardo Cesar, Jonathan Gill, Sean Murphy and Henrique Vaz

LIBRARIES

Cleaning data and initial diagnostics

Cleaning data and finding collinearity: 1. testing for and removing (after step 3) aliased variables 2. removing NA cases 3. converting variables to factors 4. converting the response variable to binary, running OLS regression 5. computing VIFs based on the OLS 6. Find columns with NA values ("age", "education", "cnty_pct_religious", "cnty_pct_evangelical")

```
train = read.csv("train.csv")

#find columns with NA values
colnames(train)[colSums(is.na(train)) > 0]
```

```
## [1] "age"                  "education"            "cnty_pct_religious"
## [4] "cnty_pct_evangelical"
```

```
#population densities (train$density_rural+train$density_suburban+train$density_urban)
densitytest = train$density_rural+train$density_suburban+train$density_urban
#Aliased. Should remove 1 or collapse to categorical
sum(densitytest) == length(densitytest)
```

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

```
#marital status
marriagetest = train$married + train$single
#All but 7 observations sum to 1. Treat as perfectly correlated
max(marriagetest)
```

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

```
sum(marriagetest)==length(marriagetest)
```

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

```
#home ownership
hometest = train$homeowner+train$renter
#around 350 people are neither homeowners nor renters; let's hang on to both variables
sum(hometest)
```

```
## [1] 10074
```

```
length(hometest)
```

```
## [1] 10439
```

```r
#removing missing
train_no_na = train[complete.cases(train), ]

#Converting relevant variables to factors
cols1 = c(2,3,4,5,6,8,11,12,14:45,48)
train_no_na[cols1] = lapply(train_no_na[cols1], factor)

#removing aliased columns
train_no_na = train_no_na[,-which(names(train_no_na) %in% c("density_rural","single"))]

train_no_na$suppdem = ifelse(train_no_na$suppdem=="Y", 1, 0)

#took out density_rural, single, and homeowner
ols = lm(suppdem ~ ., data = train_no_na)
vif(ols)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## age                     1.839972  1        1.356456
## party_reg_state         1.211020  1        1.100463
## party_primary_state     1.062765  1        1.030905
## density_suburban        1.683417  1        1.297466
## density_urban           1.963617  1        1.401291
## sex                     1.105273  2        1.025339
## combined_ethnicity_4way 1.594498  3        1.080863
## census_median_income    1.609083  1        1.268496
## ppi                     1.340068  1        1.157613
## married                 1.269946  1        1.126919
## num_children            2.644816  1        1.626289
## children_3plus          2.478937  1        1.574464
## homeowner               5.002124  1        2.236543
## renter                  4.765077  1        2.182906
## education               1.457714  4        1.048236
## hasreligion             2.589449  1        1.609176
## catholic                2.057727  1        1.434478
## christian               1.298488  1        1.139512
## bible_reader            2.940051  1        1.714658
## interest_in_religion    3.378633  1        1.838106
## donrever_1              1.949845  1        1.396368
## liberal_donor           1.191471  1        1.091545
## conservative_donor      1.033759  1        1.016739
## contbrel_1              1.123310  1        1.059863
## contbpol_1              1.182014  1        1.087205
## contbhlt_1              1.289021  1        1.135351
## blue_collar             1.180262  1        1.086399
## farmer                  1.012754  1        1.006357
## professional_technical  1.182162  1        1.087273
## retired                 1.645796  1        1.282886
```

```
## apparel_1               1.364476  1        1.168108
## bookmusc_1              1.684089  1        1.297725
## electrnc_1              1.523800  1        1.234423
## boatownr_1              1.184909  1        1.088535
## cat_1                   1.313795  1        1.146209
## environm_1              1.213706  1        1.101683
## outdgrdn_1              1.933706  1        1.390578
## outdoor_1               2.178739  1        1.476055
## guns_1                  1.478936  1        1.216115
## golf_1                  1.301400  1        1.140789
## investor_1              1.877103  1        1.370074
## veteran_1               1.173972  1        1.083500
## expensive_items_1       1.606619  1        1.267525
## cnty_pct_religious      1.151803  1        1.073221
## cnty_pct_evangelical    1.488270  1        1.219947
```

Building GLM with NAs removed dataframe

Using the data with NA rows removed, build a GLM with all variables

```
full_model_remove = glm(suppdem~ . , family = binomial(), data = train_no_na)
```

Building GLM with NAs converted to means dataframe

1.Using convert to mean function from lecture 2. reload dataset, clean and convert NAs 3. build full GLM *Note big coeffient changes for ppi, married, num_children, outdgrdn_11, and outdoor_11

```
na.convert.mean = function (frame)
{
    vars <- names(frame)
    if (!is.null(resp <- attr(attr(frame, "terms"), "response"))) {
        vars <- vars[-resp]
        x <- frame[[resp]]
        pos <- is.na(x)
        if (any(pos)) {
            frame <- frame[!pos, , drop = FALSE]
            warning(paste(sum(pos), "observations omitted due to missing values in the response"))
        }
    }
    for (j in vars) {  #j is variable names
        x <- frame[[j]]
        pos <- is.na(x)
        if (any(pos)) {
            if (length(levels(x))) {   # factors
                xx <- as.character(x)
                xx[pos] <- "NA"
```

```
                x <- factor(xx, exclude = NULL)
            }
            else if (is.matrix(x)) {   # matrices
                ats <- attributes(x)
                x.na <- 1*pos
#               x[pos] <- 0
                w <- !pos
                n <- nrow(x)
                TT <- array(1, c(1, n))
                xbar <- (TT %*% x)/(TT %*% w)
                xbar <- t(TT) %*% xbar
                x[pos] <- xbar[pos]
                attributes(x) <- ats
                attributes(x.na) <- ats
                dimnames(x.na)[[2]]=paste(dimnames(x)[[2]],".na",sep='')
                frame[[paste(j,".na",sep='')]] <- x.na
            } else {   # ordinary numerical vector
                ats <- attributes(x)
                x[pos] <- mean(x[!pos])
#               x[pos] <- 0
                x.na <- 1*pos
                frame[[paste(j,".na",sep='')]] <- x.na
                attributes(x) <- ats
            }
            frame[[j]] <- x
        }
    }
    frame
}

train_convert_na = na.convert.mean(train)

#Converting relevant variables to factors
cols1 = c(2,3,4,5,6,8,11,12,14:45,48)
train_convert_na[cols1] = lapply(train_convert_na[cols1], factor)

#removing aliased columns
train_convert_na = train_convert_na[,-which(names(train_convert_na) %in% c("density_rural","single"

fullmodel_mean_convert = glm(suppdem~ . , family = binomial(), data = train_convert_na)

#Compare summaries of full models from both the na removed data set and the na converted mean data
summary(fullmodel_mean_convert)
```

```
##
## Call:
## glm(formula = suppdem ~ ., family = binomial(), data = train_convert_na)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2115  -0.9110  -0.6472   1.1009   2.4400
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.859e-02  2.620e-01   0.224 0.823051
## age                      -6.730e-04  1.542e-03  -0.437 0.662404
## party_reg_state1         -3.555e-02  4.936e-02  -0.720 0.471424
## party_primary_state1     -1.208e-02  4.583e-02  -0.264 0.792081
## density_suburban1         4.666e-01  5.941e-02   7.853 4.05e-15 ***
## density_urban1            6.670e-01  7.536e-02   8.852  < 2e-16 ***
## sexM                     -4.131e-01  4.560e-02  -9.060  < 2e-16 ***
## sexU                     -2.206e-01  2.404e-01  -0.918 0.358794
## combined_ethnicity_4wayB  8.272e-01  1.922e-01   4.305 1.67e-05 ***
## combined_ethnicity_4wayH -5.677e-02  1.891e-01  -0.300 0.763975
## combined_ethnicity_4wayW -7.399e-01  1.814e-01  -4.078 4.54e-05 ***
## census_median_income     -4.283e-07  9.809e-07  -0.437 0.662362
## ppi                       5.017e-04  2.787e-04   1.800 0.071844 .
## married1                 -2.431e-01  6.427e-02  -3.782 0.000156 ***
## num_children             -5.782e-02  3.682e-02  -1.570 0.116355
## children_3plus1          -1.480e-01  1.531e-01  -0.967 0.333461
## homeowner1               -1.056e-02  1.294e-01  -0.082 0.934950
## renter1                   1.214e-01  1.462e-01   0.830 0.406260
## educationhigh school      2.043e-03  6.817e-02   0.030 0.976091
## educationNA              -5.008e-02  9.987e-02  -0.501 0.616067
## educationno hs degree    -2.258e-02  1.072e-01  -0.211 0.833177
## educationpost graduate    3.475e-01  8.048e-02   4.317 1.58e-05 ***
## educationsome college     3.907e-02  6.690e-02   0.584 0.559269
## hasreligion1             -2.441e-01  7.341e-02  -3.325 0.000885 ***
## catholic1                 2.445e-01  7.108e-02   3.440 0.000583 ***
## christian1                2.695e-01  1.042e-01   2.587 0.009679 **
## bible_reader1            -6.498e-02  1.288e-01  -0.505 0.613807
## interest_in_religion1    -3.971e-01  1.161e-01  -3.419 0.000628 ***
## donrever_11               1.552e-01  6.228e-02   2.491 0.012732 *
## liberal_donor1            8.686e-01  8.465e-02  10.261  < 2e-16 ***
## conservative_donor1      -1.509e+00  2.740e-01  -5.506 3.66e-08 ***
## contbrel_11              -5.025e-01  1.186e-01  -4.236 2.28e-05 ***
## contbpol_11               1.183e-01  8.607e-02   1.375 0.169189
## contbhlt_11              -5.559e-02  6.907e-02  -0.805 0.420885
## blue_collar1              7.645e-02  5.798e-02   1.318 0.187353
## farmer1                   2.629e-01  3.698e-01   0.711 0.477071
## professional_technical1  -4.222e-02  6.448e-02  -0.655 0.512589
## retired1                  1.108e-01  6.828e-02   1.623 0.104527
## apparel_11                1.668e-01  5.189e-02   3.214 0.001309 **
## bookmusc_11               8.032e-03  5.890e-02   0.136 0.891545
## electrnc_11               3.715e-03  5.785e-02   0.064 0.948800
```

```
## boatownr_11               -1.016e-01  7.080e-02  -1.435 0.151221
## cat_11                    -1.277e-01  6.136e-02  -2.080 0.037495 *
## environm_11                7.127e-02  4.888e-02   1.458 0.144801
## outdgrdn_11               -9.151e-02  6.521e-02  -1.403 0.160485
## outdoor_11                -9.893e-02  6.674e-02  -1.482 0.138234
## guns_11                   -2.103e-01  5.827e-02  -3.609 0.000308 ***
## golf_11                   -4.601e-02  5.762e-02  -0.798 0.424609
## investor_11                5.651e-02  6.182e-02   0.914 0.360611
## veteran_11                -4.913e-02  7.646e-02  -0.643 0.520512
## expensive_items_11        -1.767e-02  5.590e-02  -0.316 0.751968
## cnty_pct_religious         3.452e-01  1.853e-01   1.863 0.062490 .
## cnty_pct_evangelical      -1.254e+00  2.370e-01  -5.292 1.21e-07 ***
## age.na                     2.426e-01  5.300e-01   0.458 0.647087
## cnty_pct_religious.na     -1.239e+01  1.246e+02  -0.099 0.920807
## cnty_pct_evangelical.na    5.772e-01  1.258e+00   0.459 0.646267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13873  on 10438  degrees of freedom
## Residual deviance: 12343  on 10383  degrees of freedom
## AIC: 12455
##
## Number of Fisher Scoring iterations: 10
```

```r
summary(full_model_remove)
```

```
##
## Call:
## glm(formula = suppdem ~ ., family = binomial(), data = train_no_na)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1937  -0.9088  -0.6461   1.0940   2.4330
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.635e-01  2.983e-01   0.548 0.583474
## age                        4.459e-04  1.717e-03   0.260 0.795053
## party_reg_state1          -4.028e-02  5.173e-02  -0.779 0.436212
## party_primary_state1      -2.973e-02  4.806e-02  -0.619 0.536217
## density_suburban1          4.585e-01  6.242e-02   7.345 2.06e-13 ***
## density_urban1             6.745e-01  7.952e-02   8.483  < 2e-16 ***
## sexM                      -4.015e-01  4.788e-02  -8.385  < 2e-16 ***
## sexU                       4.742e-03  3.738e-01   0.013 0.989879
## combined_ethnicity_4wayB   8.115e-01  2.003e-01   4.051 5.09e-05 ***
## combined_ethnicity_4wayH  -4.924e-02  1.974e-01  -0.249 0.803013
```

```
## combined_ethnicity_4wayW -7.564e-01  1.887e-01  -4.008 6.13e-05 ***
## census_median_income     -6.899e-07  1.030e-06  -0.670 0.502873
## ppi                       4.511e-04  2.830e-04   1.594 0.110966
## married1                 -1.791e-01  6.954e-02  -2.576 0.009996 **
## num_children             -8.173e-02  3.814e-02  -2.143 0.032125 *
## children_3plus1          -6.024e-02  1.575e-01  -0.383 0.702090
## homeowner1               -1.294e-01  1.759e-01  -0.736 0.461904
## renter1                   5.549e-03  1.900e-01   0.029 0.976697
## educationhigh school     -4.926e-03  6.848e-02  -0.072 0.942656
## educationno hs degree    -4.340e-02  1.081e-01  -0.401 0.688070
## educationpost graduate    3.414e-01  8.056e-02   4.238 2.25e-05 ***
## educationsome college     3.753e-02  6.716e-02   0.559 0.576247
## hasreligion1             -2.414e-01  7.664e-02  -3.150 0.001631 **
## catholic1                 2.449e-01  7.386e-02   3.316 0.000913 ***
## christian1                2.560e-01  1.075e-01   2.381 0.017262 *
## bible_reader1            -7.659e-02  1.302e-01  -0.588 0.556429
## interest_in_religion1    -3.704e-01  1.179e-01  -3.142 0.001678 **
## donrever_11               1.780e-01  6.473e-02   2.749 0.005975 **
## liberal_donor1            8.587e-01  8.660e-02   9.915  < 2e-16 ***
## conservative_donor1      -1.404e+00  2.756e-01  -5.095 3.49e-07 ***
## contbrel_11              -5.213e-01  1.218e-01  -4.281 1.86e-05 ***
## contbpol_11               1.065e-01  8.805e-02   1.209 0.226563
## contbhlt_11              -4.308e-02  7.091e-02  -0.608 0.543496
## blue_collar1              5.659e-02  5.912e-02   0.957 0.338461
## farmer1                   2.530e-01  3.695e-01   0.685 0.493563
## professional_technical1  -4.564e-02  6.472e-02  -0.705 0.480716
## retired1                  8.300e-02  7.045e-02   1.178 0.238760
## apparel_11                1.567e-01  5.382e-02   2.911 0.003601 **
## bookmusc_11              -1.453e-03  6.162e-02  -0.024 0.981193
## electrnc_11               7.683e-03  6.055e-02   0.127 0.899028
## boatownr_11              -9.975e-02  7.281e-02  -1.370 0.170687
## cat_11                   -1.432e-01  6.323e-02  -2.265 0.023513 *
## environm_11               5.980e-02  5.039e-02   1.187 0.235339
## outdgrdn_11              -1.125e-01  6.768e-02  -1.662 0.096515 .
## outdoor_11               -1.204e-01  6.945e-02  -1.733 0.083015 .
## guns_11                  -2.034e-01  6.027e-02  -3.374 0.000741 ***
## golf_11                  -4.795e-02  5.963e-02  -0.804 0.421306
## investor_11               4.747e-02  6.413e-02   0.740 0.459121
## veteran_11               -5.481e-02  7.978e-02  -0.687 0.492121
## expensive_items_11       -6.218e-03  5.804e-02  -0.107 0.914675
## cnty_pct_religious        3.314e-01  1.952e-01   1.698 0.089529 .
## cnty_pct_evangelical     -1.269e+00  2.491e-01  -5.096 3.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12597  on 9478  degrees of freedom
```

```
## Residual deviance: 11201  on 9427  degrees of freedom
## AIC: 11305
##
## Number of Fisher Scoring iterations: 4
```

Model building and anova comparison

1. start with only the variables that were significant in the full model
2. add in other coefficients

adding num_children improves on the model

Note: adding age alone improves on the model, but adding age + age.na does not make a significant improvement over adding neither.

```
#significant coefficients only
glm1 = glm(suppdem~ density_suburban + density_urban + sex + combined_ethnicity_4way + combined_eth

glm2 = glm(suppdem~ density_suburban + density_urban + sex + combined_ethnicity_4way + combined_eth


glm3 = glm(suppdem~ density_suburban + density_urban + sex + combined_ethnicity_4way + combined_eth

glm4 = glm(suppdem~ density_suburban + density_urban + sex + combined_ethnicity_4way + combined_eth

anova(glm1, glm2,  test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: suppdem ~ density_suburban + density_urban + sex + combined_ethnicity_4way +
##     combined_ethnicity_4way + ppi + married + education + hasreligion +
##     catholic + christian + interest_in_religion + donrever_1 +
##     liberal_donor + conservative_donor + contbrel_1 + apparel_1 +
##     cat_1 + guns_1 + cnty_pct_religious + cnty_pct_evangelical +
##     cnty_pct_religious.na + cnty_pct_evangelical.na
## Model 2: suppdem ~ density_suburban + density_urban + sex + combined_ethnicity_4way +
##     combined_ethnicity_4way + ppi + married + education + hasreligion +
##     catholic + christian + interest_in_religion + donrever_1 +
##     liberal_donor + conservative_donor + contbrel_1 + apparel_1 +
##     cat_1 + guns_1 + cnty_pct_religious + cnty_pct_evangelical +
##     cnty_pct_religious.na + cnty_pct_evangelical.na + age + age.na +
##     party_reg_state + party_primary_state + +census_median_income
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     10409      12383
## 2     10404      12382  5   1.1207   0.9523
```

Clean Test Data and Predictions

```r
test = read.csv("test.csv")
test.converted = na.convert.mean(test)

#Clean test data
cols1 = c(2,3,4,5,6,8,11,12,14:45,48)
test.converted[cols1] = lapply(test.converted[cols1], factor)

#Predict on AIC MODEL
# out = predict(aicmodel, test.converted)
# write.csv(out, "predictions.csv")
```