Predictive Factors for Low Verbal Reasoning Scores

Valentina Litang

Using the Three Simple Rules Framework, I built a linear regression that preserves predictive power while simplifying statistical models.

Step 1 - Select

After splitting up my data into training and validating sets, I used forward AIC selection to select k features. My selected predictors were: momwhite, ein, mia, momage, birth.o, b.head, was, ark, cig.

Although I was not given a codebook and I couldn't find one online, I have reason to believe that the only continuous variables of my selected predictors are momage and b.head. Given that we want the shrink from lasso to be unit agnostic, I will standardize my continuous variables.

```r
set.seed(123)
dat <- read_csv("verbal_reasoning.csv")
```

```
## Rows: 608 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (30): bw, b.head, preterm, birth.o, nnhealth, momage, sex, twin, b.marr,...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
dat$set <-sample(c('train', 'validation'),
                 size = nrow(dat),
                 replace = T,
                 prob = c(.70, .30))

train <- dat |> filter(set == 'train') |> select(-set)
validation <- dat |> filter(set == 'validation') |> select(-set)
#test <- dat |> filter(set == 'test') |> select(-set)


null_model <- lm(ppvtr36 ~ 1, data = train)
full_model <- lm(ppvtr36 ~ ., data = train)

forward_selection <- stepAIC(null_model,
                             scope = list(lower = null_model,
                                          upper = full_model),
                             direction = "forward")
```

```
## Start:  AIC=2544.39
## ppvtr36 ~ 1
##
##             Df Sum of Sq    RSS    AIC
## + momwhite   1     45816 113130 2400.2
## + momblack   1     22745 136201 2480.0
## + b.marr     1     17529 141417 2496.1
## + ein        1     16473 142472 2499.3
## + mom.lths   1     13750 145196 2507.5
## + har        1     13167 145779 2509.2
## + momage     1     12258 146688 2511.9
## + work.dur   1     10236 148709 2517.8
## + momhisp    1      8163 150783 2523.7
## + birth.o    1      7571 151375 2525.4
## + mia        1      6878 152068 2527.4
## + first      1      6006 152939 2529.8
## + was        1      3948 154998 2535.6
## + booze      1      2709 156236 2539.0
## + b.head     1      2374 156572 2539.9
## + mom.scoll  1      1348 157597 2542.7
## + bw         1      1248 157698 2543.0
## + zbw        1      1248 157698 2543.0
## + cig        1      1005 157941 2543.7
## + preterm    1       847 158099 2544.1
## <none>                   158946 2544.4
## + mom.hs     1       575 158371 2544.8
## + prenatal   1       325 158621 2545.5
## + pen        1       232 158714 2545.8
## + tex        1       226 158720 2545.8
## + drugs      1       173 158772 2545.9
## + ark        1        76 158870 2546.2
## + nnhealth   1        50 158895 2546.2
## + sex        1        21 158925 2546.3
## + twin       1         0 158945 2546.4
##
## Step:  AIC=2400.18
## ppvtr36 ~ momwhite
##
##             Df Sum of Sq    RSS    AIC
## + ein        1    7116.8 106013 2374.2
## + har        1    4840.5 108290 2383.4
## + mom.lths   1    3880.5 109250 2387.2
## + birth.o    1    2933.3 110197 2390.9
## + work.dur   1    2599.5 110531 2392.2
## + cig        1    1949.5 111181 2394.7
## + first      1    1906.2 111224 2394.9
## + mia        1    1806.5 111324 2395.3
## + momage     1    1753.1 111377 2395.5
## + momblack   1    1045.3 112085 2398.2
## + momhisp    1    1045.3 112085 2398.2
## + pen        1    1026.6 112104 2398.3
## + b.marr     1     955.9 112174 2398.5
## + mom.scoll  1     893.5 112237 2398.8
## + preterm    1     791.8 112338 2399.2
## + tex        1     730.2 112400 2399.4
## + booze      1     701.9 112428 2399.5
## + b.head     1     664.9 112465 2399.6
## <none>                   113130 2400.2
## + nnhealth   1     491.9 112638 2400.3
## + bw         1     290.8 112839 2401.1
## + zbw        1     290.8 112839 2401.1
## + ark        1     226.9 112903 2401.3
## + was        1     225.0 112905 2401.3
## + mom.hs     1     222.6 112908 2401.3
## + sex        1      69.8 113060 2401.9
## + twin       1      67.1 113063 2401.9
## + prenatal   1       6.2 113124 2402.2
## + drugs      1       5.8 113124 2402.2
##
## Step:  AIC=2374.24
## ppvtr36 ~ momwhite + ein
##
##             Df Sum of Sq    RSS    AIC
## + mia        1    3512.5 102501 2361.8
## + mom.lths   1    3476.4 102537 2361.9
## + har        1    3350.3 102663 2362.4
## + momage     1    3190.5 102823 2363.1
## + birth.o    1    2628.2 103385 2365.4
## + work.dur   1    1874.1 104139 2368.6
## + cig        1    1815.4 104198 2368.8
## + first      1    1614.2 104399 2369.6
## + mom.scoll  1    1272.4 104741 2371.1
## + b.head     1    1133.9 104880 2371.6
## + b.marr     1     972.2 105041 2372.3
## + preterm    1     827.9 105186 2372.9
## + ark        1     792.6 105221 2373.0
## + bw         1     678.8 105335 2373.5
## + zbw        1     678.8 105335 2373.5
## + was        1     573.6 105440 2373.9
## + mom.hs     1     540.7 105473 2374.0
## <none>                   106013 2374.2
## + nnhealth   1     362.2 105651 2374.8
## + booze      1     335.2 105678 2374.9
## + pen        1     247.3 105766 2375.2
## + sex        1     107.3 105906 2375.8
## + tex        1      60.5 105953 2376.0
## + prenatal   1      27.8 105986 2376.1
## + drugs      1      24.9 105988 2376.1
## + twin       1      22.1 105991 2376.2
## + momblack   1       2.2 106011 2376.2
## + momhisp    1       2.2 106011 2376.2
##
## Step:  AIC=2361.75
## ppvtr36 ~ momwhite + ein + mia
##
##             Df Sum of Sq    RSS    AIC
## + momage     1   2889.89  99611 2351.4
## + har        1   2594.15  99907 2352.7
## + mom.lths   1   2542.67  99958 2352.9
## + birth.o    1   2270.80 100230 2354.1
## + cig        1   2257.92 100243 2354.2
## + first      1   1463.24 101038 2357.6
## + work.dur   1   1355.34 101146 2358.0
## + ark        1   1325.36 101176 2358.2
## + mom.scoll  1    861.95 101639 2360.1
## + was        1    823.90 101677 2360.3
## + b.head     1    748.28 101753 2360.6
## + b.marr     1    665.96 101835 2360.9
## + mom.hs     1    616.11 101885 2361.2
## + preterm    1    581.71 101919 2361.3
## <none>                   102501 2361.8
## + bw         1    392.63 102108 2362.1
## + zbw        1    392.63 102108 2362.1
## + booze      1    236.54 102264 2362.8
## + nnhealth   1    180.39 102320 2363.0
## + prenatal   1     75.53 102425 2363.4
## + sex        1     64.31 102437 2363.5
## + tex        1     31.21 102470 2363.6
## + drugs      1     26.53 102474 2363.6
## + pen        1     19.85 102481 2363.7
## + twin       1      1.32 102500 2363.8
## + momblack   1      0.41 102500 2363.8
## + momhisp    1      0.41 102500 2363.8
##
## Step:  AIC=2351.45
## ppvtr36 ~ momwhite + ein + mia + momage
##
##              Df Sum of Sq   RSS    AIC
## + birth.o    1    5283.8 94327 2330.0
## + first      1    3440.2 96171 2338.3
## + har        1    1725.7 97885 2345.9
## + cig        1    1704.6 97906 2346.0
## + was        1    1113.9 98497 2348.6
## + ark        1    1112.6 98498 2348.6
## + b.head     1     966.7 98644 2349.3
## + mom.lths   1     949.4 98662 2349.3
## + preterm    1     741.2 98870 2350.2
## + work.dur   1     624.4 98987 2350.8
## + mom.hs     1     598.6 99012 2350.9
## + bw         1     539.8 99071 2351.1
## + zbw        1     539.8 99071 2351.1
## <none>                   99611 2351.4
## + nnhealth   1     370.5 99240 2351.8
## + mom.scoll  1     292.4 99319 2352.2
## + booze      1      96.0 99515 2353.0
## + b.marr     1      74.7 99536 2353.1
## + momblack   1      32.5 99579 2353.3
## + momhisp    1      32.5 99579 2353.3
## + tex        1      23.0 99588 2353.4
## + sex        1      18.9 99592 2353.4
## + pen        1      17.2 99594 2353.4
## + prenatal   1      12.4 99599 2353.4
## + twin       1       9.7 99601 2353.4
## + drugs      1       2.4 99609 2353.4
##
## Step:  AIC=2330.02
## ppvtr36 ~ momwhite + ein + mia + momage + birth.o
##
##              Df Sum of Sq   RSS    AIC
## + b.head     1   1252.67 93075 2326.3
## + was        1   1093.39 93234 2327.0
## + preterm    1    933.94 93393 2327.7
## + cig        1    878.21 93449 2328.0
## + bw         1    792.88 93534 2328.4
## + zbw        1    792.88 93534 2328.4
## + har        1    726.35 93601 2328.7
## + ark        1    530.62 93797 2329.6
## <none>                   94327 2330.0
## + mom.lths   1    410.72 93916 2330.1
## + nnhealth   1    343.49 93984 2330.4
## + mom.hs     1    273.41 94054 2330.8
## + work.dur   1    173.66 94154 2331.2
## + tex        1    169.90 94157 2331.2
## + sex        1     77.37 94250 2331.7
## + mom.scoll  1     60.30 94267 2331.7
## + b.marr     1     47.19 94280 2331.8
## + prenatal   1     34.18 94293 2331.9
## + twin       1     16.65 94311 2331.9
## + booze      1      8.50 94319 2332.0
## + drugs      1      7.09 94320 2332.0
## + pen        1      6.07 94321 2332.0
## + first      1      4.23 94323 2332.0
## + momblack   1      0.01 94327 2332.0
## + momhisp    1      0.01 94327 2332.0
##
## Step:  AIC=2326.27
## ppvtr36 ~ momwhite + ein + mia + momage + birth.o + b.head
##
##              Df Sum of Sq   RSS    AIC
## + was        1    960.89 92114 2323.8
## + cig        1    641.07 92433 2325.3
## + ark        1    619.20 92455 2325.4
## + har        1    528.35 92546 2325.8
## <none>                   93075 2326.3
## + mom.lths   1    337.57 92737 2326.7
## + nnhealth   1    250.90 92824 2327.1
## + tex        1    242.33 92832 2327.2
## + mom.hs     1    231.72 92843 2327.2
## + work.dur   1    140.05 92934 2327.6
## + preterm    1     54.86 93020 2328.0
## + b.marr     1     32.68 93042 2328.1
## + prenatal   1     28.52 93046 2328.1
## + first      1     24.20 93050 2328.2
## + mom.scoll  1     22.98 93052 2328.2
## + bw         1     14.90 93060 2328.2
## + zbw        1     14.90 93060 2328.2
## + sex        1     13.57 93061 2328.2
## + booze      1     13.26 93061 2328.2
## + twin       1     12.28 93062 2328.2
## + momblack   1     10.93 93064 2328.2
## + momhisp    1     10.93 93064 2328.2
## + pen        1      5.70 93069 2328.2
## + drugs      1      0.48 93074 2328.3
##
## Step:  AIC=2323.81
## ppvtr36 ~ momwhite + ein + mia + momage + birth.o + b.head +
##     was
##
##              Df Sum of Sq   RSS    AIC
## + ark        1   1098.21 91015 2320.7
## + cig        1    606.55 91507 2323.0
## <none>                   92114 2323.8
## + mom.lths   1    321.35 91792 2324.3
## + nnhealth   1    301.31 91812 2324.4
## + mom.hs     1    221.11 91893 2324.8
## + har        1    186.09 91928 2324.9
## + work.dur   1    147.40 91966 2325.1
## + tex        1    128.35 91985 2325.2
## + preterm    1    111.43 92002 2325.3
## + b.marr     1     49.71 92064 2325.6
## + sex        1     31.19 92082 2325.7
## + drugs      1     27.06 92087 2325.7
## + first      1     25.76 92088 2325.7
## + mom.scoll  1     18.13 92095 2325.7
## + prenatal   1     17.69 92096 2325.7
## + booze      1     12.83 92101 2325.8
## + twin       1     10.35 92103 2325.8
## + momblack   1      3.53 92110 2325.8
## + momhisp    1      3.53 92110 2325.8
## + bw         1      0.96 92113 2325.8
## + zbw        1      0.96 92113 2325.8
## + pen        1      0.22 92113 2325.8
##
## Step:  AIC=2320.65
## ppvtr36 ~ momwhite + ein + mia + momage + birth.o + b.head +
##     was + ark
##
##              Df Sum of Sq   RSS    AIC
## + cig        1    631.83 90384 2319.7
## <none>                   91015 2320.7
## + mom.lths   1    370.92 90645 2320.9
## + nnhealth   1    275.49 90740 2321.3
## + work.dur   1    209.81 90806 2321.7
## + mom.hs     1    145.24 90870 2322.0
## + preterm    1     58.84 90957 2322.4
## + pen        1     49.42 90966 2322.4
## + sex        1     37.41 90978 2322.5
## + mom.scoll  1     29.07 90986 2322.5
## + twin       1     23.06 90992 2322.5
## + b.marr     1     20.66 90995 2322.6
## + momblack   1     14.73 91001 2322.6
## + momhisp    1     14.73 91001 2322.6
## + drugs      1     11.86 91004 2322.6
## + first      1     10.49 91005 2322.6
## + har        1      9.89 91006 2322.6
## + booze      1      6.46 91009 2322.6
## + bw         1      4.98 91010 2322.6
## + zbw        1      4.98 91010 2322.6
## + tex        1      2.48 91013 2322.6
## + prenatal   1      0.59 91015 2322.7
##
## Step:  AIC=2319.65
## ppvtr36 ~ momwhite + ein + mia + momage + birth.o + b.head +
##     was + ark + cig
##
##             Df Sum of Sq   RSS    AIC
## <none>                   90384 2319.7
## + nnhealth   1   312.283 90071 2320.2
## + mom.lths   1   287.861 90096 2320.3
## + work.dur   1   147.698 90236 2320.9
## + mom.hs     1   115.856 90268 2321.1
## + preterm    1    98.015 90286 2321.2
## + sex        1    52.185 90331 2321.4
## + pen        1    47.716 90336 2321.4
## + booze      1    42.873 90341 2321.4
## + drugs      1    42.760 90341 2321.4
## + momblack   1    24.575 90359 2321.5
## + momhisp    1    24.575 90359 2321.5
## + mom.scoll  1    23.722 90360 2321.5
## + twin       1    21.470 90362 2321.6
## + first      1    16.436 90367 2321.6
## + bw         1    11.083 90373 2321.6
## + zbw        1    11.083 90373 2321.6
## + prenatal   1     3.497 90380 2321.6
## + har        1     3.200 90380 2321.6
## + tex        1     0.908 90383 2321.7
## + b.marr     1     0.296 90383 2321.7
```

```
names(forward_selection$coefficients)
```

```
##  [1] "(Intercept)" "momwhite"    "ein"         "mia"         "momage"
##  [6] "birth.o"     "b.head"      "was"         "ark"         "cig"
```

```
# for later
train_means <- sapply(train, mean)
train_sds <- sapply(train, sd)

continuous_vars <- c("b.head", "momage")
X_continuous <- train |> dplyr::select(continuous_vars)
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(continuous_vars)
```

```
## 
##   # Now:
##   data %>% select(all_of(continuous_vars))
## 
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
scaled_X_continuous <- scale(X_continuous)
X_other <- train |> dplyr::select(names(forward_selection$coefficients)[-1]) |> dplyr::select(-all_of(c
X_final <- cbind(scaled_X_continuous, X_other)
X_final <- as.matrix(X_final)
```

Step 2 - Regress

Then, I fit a Lasso regression on the training dataset.

```
lasso <- cv.glmnet(X_final, train$ppvtr36, nfolds = 10, alpha =1)
lasso_final <- glmnet(X_final, train$ppvtr36, lambda = lasso$lambda.min, alpha = 1)

betas <- lasso_final$beta
lasso_coef <- as.vector(betas)
names(lasso_coef) <- rownames(betas)
lasso_coef[lasso_coef != 0]
```

```
##     b.head     momage   momwhite        ein        mia    birth.o        was 
##   1.534807   4.120699  15.260861 -15.677628 -10.022724  -3.666833  -5.801815 
##        ark        cig 
##  -5.015567  -2.592293 
```

Step 3 - Round

To simplify the predictors even more, I will restrict the coefficients within a range of (-3, 3).

```
simple_coef <- round(lasso_coef * 3/(max(abs(lasso_coef))))
simple_coef
```

```
##   b.head   momage momwhite      ein      mia  birth.o      was      ark 
##        0        1        3       -3       -2       -1       -1       -1 
##      cig 
##        0 
```

Finally, applied standardization to the validation based on values from the training set before fitting my model on the validation dataset.

```r
validation_simple <- validation

for (i in 1:length(validation)) {
  validation_simple[ , i] <- (validation_simple[ , i] - train_means[i])/train_sds[i]
}

validation_simple <- validation_simple[ , names(forward_selection$coefficients)[-1]]

validation_simple <- as.matrix(validation_simple)
simple_pred <- validation_simple %*% simple_coef
sqrt(mean((simple_pred - validation$ppvtr36)**2))
```

```
## [1] 84.14346
```

Momage has a coefficient of 1 - This tells us that higher values of momage predict higher verbal reasoning scores.

Momwhite has a coefficient of 3 - This tells us that having a white mom predicts a higher verbal reasoning score.

Ein has a coefficient of -3 - This tells us that having a higher ein predicts a lower verbal reasoning score.

Mia has a coefficient of -2 - This tells us that having a higher mia predicts a lower verbal reasoning score.

Birth.o has a coefficient of -1 - This tells us that having a higher birth order predicts a lower verbal reasoning score.

Was has a coefficient of -1 - This tells us that having a higher was predicts a lower verbal reasoning score.

Ark has a coefficient of -1 - This tells us that having a higher ark predicts a lower verbal reasoning score.

In other words, the following are risk factors for low verbal reasoning scores: younger moms, non-white moms, high ein, high mia, high birth.o, high was, and high ark. Intervention should be considered for students with some or all of the risk factors.