Homework 8

Marwin Carmo

03/08/2024

The data for this homework concern the Nerdy Personality Attributes Scale, and were collected online from December 2015 – December 2018 on the open-source psychometrics project . You can find more info here. The original dataset has been modified for this homeowrk and contains the following variables:

- `nerdiness`: standardized nerdiness score (outcome variable)
- `age`: participant age in years
- `voted`: whether participant had recently voted in a national election (0 = no, 1 = yes)
- `asd`: whether participant had ever been diagnosed with ASD (0 = no, 1 = yes)
- The remaining variables correspond to self-reported personality scores (0 – 7)
  - `extraverted`
  - `critical`
  - `dependable`
  - `anxious`
  - `open`
  - `reserved`
  - `sympathetic`
  - `disorganized`
  - `calm`
  - `uncreative`

1. Model Selection Criterion

a)

Fit three linear regressions with `nerdiness` as the outcome and compute their MSE. Choose whatever predictors you would like but make sure that each regression has a different number of predictors. Which model has the lowest MSE? Describe your intuition on why this model has the lowest MSE. (1 pt.)

```
mod1 <- lm(nerdiness ~ extraverted + critical + anxious + open + reserved + dependable + sympathetic + d
mod2 <- lm(nerdiness ~ extraverted + critical + dependable + sympathetic + disorganized + calm + uncrea
mod3 <- lm(nerdiness ~ extraverted + critical + anxious + open + reserved + dependable, data = nerdy)

mse_mod1 <- mean(resid(mod1)^2)
mse_mod2 <- mean(resid(mod2)^2)
mse_mod3 <- mean(resid(mod3)^2)
```

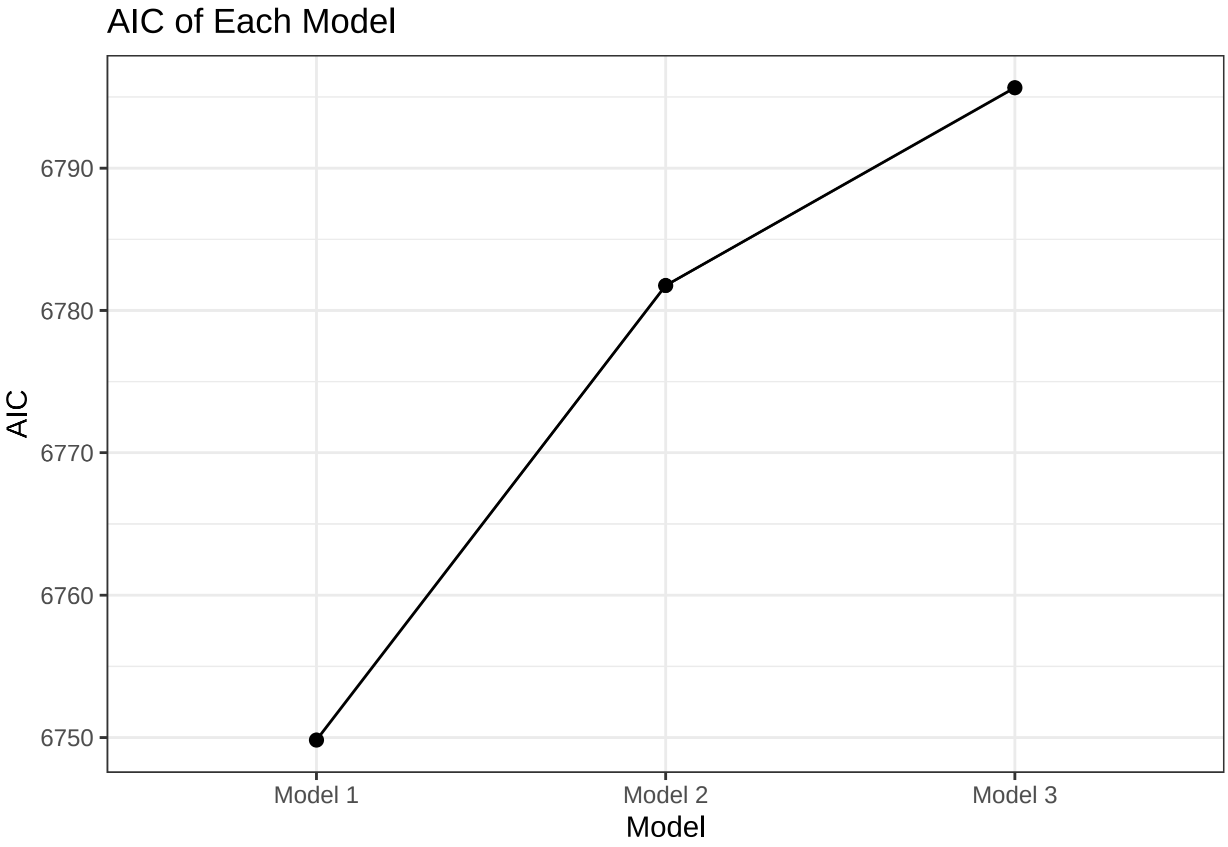
Model 1, including all variables as predictors, has the lower MSE. This model fits the data more closely than the other two. That implies that including all the variables in the data set as predictors might give us more accurate predictions on new data.

b)

Choose one of the model selection criterion we discussed in lab. Compute it for each of your regressions and plot them out. (1 pt.)

```r
model_list <- list(mod1, mod2, mod3)
mod_sel_df <- data.frame(model_name = c("Model 1", "Model 2", "Model 3"),
                         aic = sapply(model_list, aic))

mod_sel_df |>
  ggplot(aes(x = model_name, y = aic)) +
  geom_line(aes(group = 1)) +
  geom_point(size = 2) +
  labs(x = "Model",
  y = "AIC",
  title = "AIC of Each Model") +
  theme_bw()
```



c)

Which model was best according to your chosen criterion? Is this the same model with the lowest MSE in part a? Explain why you think they are or are not the same. (1 pt.)

Model 1 also performed best when using AIC as the selection criteria. AIC considers both the goodness of fit and the model's complexity. Even with the added penalization due to the number of parameters, Model 1 was still the best choice. That suggests that the accuracy obtained by including all predictors counterbalances the penalization due to including more parameters.

2.

a)

Using the `regsubsets()` function, use best subsets selection to predict nerdiness. Set `nvmax = 13`. (1 pt.)

```
best_subsets <- regsubsets(nerdiness ~ .,
                           data = nerdy,
                           nvmax = 13,
                           method = "exhaustive")
```

b)

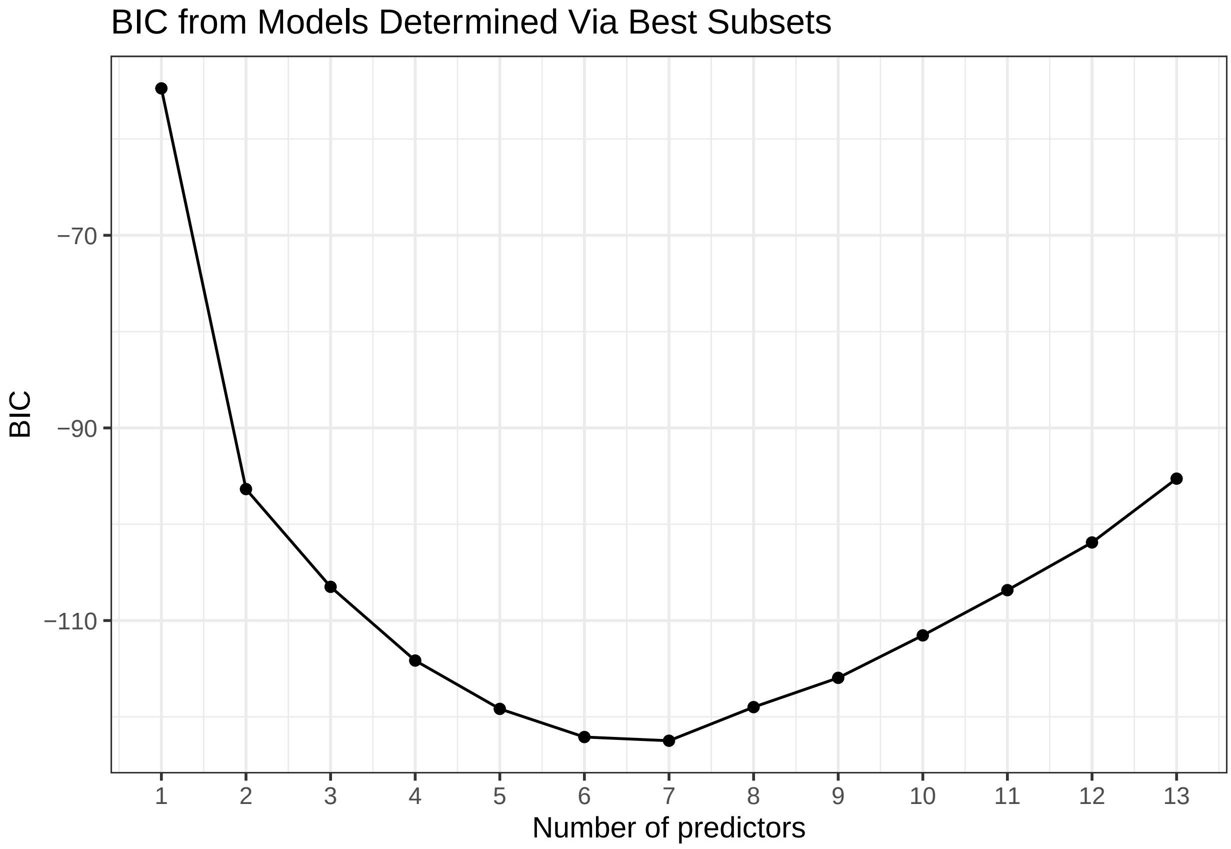
Using the information from the best subsets regression, plot the number of variables considered in each model against their respective BIC. Which model had the highest BIC? Which model had the lowest BIC, and what predictors did this model contain? (1 pt.)

The highest BIC was obtained with the model with only 1 predictor. The lowest was given by the model with 7 predictors. The variables included in this model were: `extraverted`, `open`, `reserved`, `sympathetic`, `calm`, `uncreative`, and `asd`.

```
bs_summ <- summary(best_subsets)

bs_results <-data.frame(n.predictors = 1:13,
                        BIC = c(bs_summ$bic))

bs_results |>
  ggplot(aes(x = n.predictors, y = BIC)) +
  geom_line(aes(group = 1)) +
  geom_point() +
  scale_x_continuous(breaks = 1:13) +
  labs(x = "Number of predictors",
       y = "BIC",
       title = "BIC from Models Determined Via Best Subsets") +
  theme_bw()
```



c)

Repeat the analysis using forward selection instead. Find the best model according to BIC. How many predictors does it contain? Are they the same predictors as the model chosen using best subsets? (1 pt.)

This model contains 9 predictors. The forward selection model includes all predictors of the best subsets model except for `asd`. It also includes `critical`, `disorganized`, and `dependable`.

```
sel <- step(object = lm(nerdiness ~ 1, data=nerdy),
     scope = nerdiness ~ extraverted + critical + anxious + open + reserved + dependable + sympathetic
     direction = "forward")
```

```
## Start:  AIC=1
## nerdiness ~ 1
##
##                  Df Sum of Sq    RSS     AIC
## + extraverted     1    66.195 932.81 -65.559
## + reserved        1    54.330 944.67 -52.920
## + uncreative      1    25.607 973.39 -22.967
## + calm            1    14.493 984.51 -11.614
## + anxious         1    11.524 987.48  -8.604
## + sympathetic     1    10.262 988.74  -7.326
## + open            1     9.443 989.56  -6.498
## + disorganized    1     5.703 993.30  -2.726
## + critical        1     5.307 993.69  -2.327
## <none>                         999.00   0.999
## + dependable      1     0.230 998.77   2.769
```

```
##
## Step:  AIC=-65.56
## nerdiness ~ extraverted
##
##                 Df Sum of Sq    RSS      AIC
## + uncreative    1    44.173 888.63 -112.072
## + open          1    30.130 902.68  -96.392
## + reserved      1    11.347 921.46  -75.798
## + calm          1     8.087 924.72  -72.267
## + disorganized  1     7.282 925.52  -71.396
## + critical      1     5.965 926.84  -69.974
## + anxious       1     4.008 928.80  -67.865
## <none>                      932.81  -65.559
## + sympathetic   1     1.762 931.04  -65.450
## + dependable    1     0.391 932.41  -63.978
##
## Step:  AIC=-112.07
## nerdiness ~ extraverted + uncreative
##
##                 Df Sum of Sq    RSS     AIC
## + open          1   15.0312 873.60 -127.13
## + reserved      1   13.9792 874.65 -125.93
## + critical      1    6.3042 882.33 -117.19
## + disorganized  1    5.8595 882.77 -116.69
## + calm          1    4.8632 883.77 -115.56
## + anxious       1    3.7705 884.86 -114.32
## + sympathetic   1    3.0598 885.57 -113.52
## <none>                      888.63 -112.07
## + dependable    1    0.5871 888.04 -110.73
##
## Step:  AIC=-127.13
## nerdiness ~ extraverted + uncreative + open
##
##                 Df Sum of Sq    RSS     AIC
## + reserved      1   12.6291 860.97 -139.69
## + calm          1    7.9203 865.68 -134.24
## + critical      1    7.0239 866.58 -133.21
## + anxious       1    5.8113 867.79 -131.81
## + sympathetic   1    4.7882 868.81 -130.63
## + disorganized  1    4.5950 869.01 -130.41
## <none>                      873.60 -127.13
## + dependable    1    0.1874 873.41 -125.35
##
## Step:  AIC=-139.69
## nerdiness ~ extraverted + uncreative + open + reserved
##
##                 Df Sum of Sq    RSS     AIC
## + calm          1   10.0116 850.96 -149.39
## + critical      1    7.7260 853.25 -146.71
## + sympathetic   1    6.9678 854.00 -145.82
## + anxious       1    4.2651 856.71 -142.66
## + disorganized  1    4.0951 856.88 -142.46
## <none>                      860.97 -139.69
## + dependable    1    0.0321 860.94 -137.73
```

```
##
## Step:  AIC=-149.39
## nerdiness ~ extraverted + uncreative + open + reserved + calm
##
##                 Df Sum of Sq    RSS     AIC
## + sympathetic    1    7.2719 843.69 -155.97
## + critical       1    5.3345 845.63 -153.68
## + disorganized   1    2.1968 848.76 -149.97
## <none>                       850.96 -149.39
## + dependable     1    0.8321 850.13 -148.37
## + anxious        1    0.1859 850.77 -147.61
##
## Step:  AIC=-155.97
## nerdiness ~ extraverted + uncreative + open + reserved + calm +
##     sympathetic
##
##                 Df Sum of Sq    RSS     AIC
## + critical       1    2.8079 840.88 -157.31
## + disorganized   1    2.0604 841.63 -156.42
## <none>                       843.69 -155.97
## + anxious        1    1.3430 842.35 -155.57
## + dependable     1    1.3186 842.37 -155.54
##
## Step:  AIC=-157.31
## nerdiness ~ extraverted + uncreative + open + reserved + calm +
##     sympathetic + critical
##
##                 Df Sum of Sq    RSS     AIC
## + disorganized   1   1.90073 838.98 -157.57
## <none>                       840.88 -157.31
## + dependable     1   1.04068 839.84 -156.54
## + anxious        1   0.68363 840.20 -156.12
##
## Step:  AIC=-157.57
## nerdiness ~ extraverted + uncreative + open + reserved + calm +
##     sympathetic + critical + disorganized
##
##               Df Sum of Sq    RSS     AIC
## + dependable  1   3.10799 835.87 -159.28
## <none>                    838.98 -157.57
## + anxious     1   0.63736 838.34 -156.33
##
## Step:  AIC=-159.28
## nerdiness ~ extraverted + uncreative + open + reserved + calm +
##     sympathetic + critical + disorganized + dependable
##
##            Df Sum of Sq    RSS     AIC
## <none>                 835.87 -159.28
## + anxious  1   0.54584 835.33 -157.93
```

```
summary(sel)
```

```
##
## Call:
## lm(formula = nerdiness ~ extraverted + uncreative + open + reserved +
##     calm + sympathetic + critical + disorganized + dependable,
##     data = nerdy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9961 -0.5513  0.0889  0.6414  2.7231
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.34373    0.24751  -1.389  0.16522
## extraverted  -0.11024    0.01935  -5.696 1.62e-08 ***
## uncreative   -0.10614    0.01839  -5.770 1.06e-08 ***
## open          0.09949    0.02274   4.375 1.35e-05 ***
## reserved      0.08887    0.02088   4.257 2.27e-05 ***
## calm         -0.05342    0.01702  -3.139  0.00175 **
## sympathetic  -0.04909    0.01921  -2.555  0.01076 *
## critical      0.02621    0.01673   1.566  0.11762
## disorganized  0.03644    0.01681   2.168  0.03041 *
## dependable    0.03852    0.02008   1.919  0.05532 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9189 on 990 degrees of freedom
## Multiple R-squared:  0.1633, Adjusted R-squared:  0.1557
## F-statistic: 21.47 on 9 and 990 DF,  p-value: < 2.2e-16
```

3.

a)

Split your data into two parts into a training set and a test set, and show the head of each dataset. Use 80% of the data for training. I have started the code for you below. Please do not modify it. (1 pt.)

```r
set.seed(1)
prop <- 0.8
n_rows <- nrow(nerdy)
n_sample <- 0.8 * nrow(nerdy)
training_idx <- sample(1:n_rows, size = n_sample)
```

```r
training <- nerdy[training_idx, ]
test <- nerdy[-training_idx, ]

head(training)
```

```
## # A tibble: 6 x 14
##   nerdiness[,1] extraverted critical dependable anxious  open reserved
##           <dbl>       <int>    <int>      <int>   <int> <int>    <int>
## 1        -1.41            3        4          6       3     4        3
## 2         0.779           1        5          5       7     6        7
## 3         0.135           7        6          7       7     6        7
## 4         0.908           1        3          3       7     5        7
## 5        -0.380           6        5          6       2     7        3
## 6         1.87            1        4          7       7     7        5
## # i 7 more variables: sympathetic <int>, disorganized <int>, calm <int>,
## #   uncreative <int>, age <int>, voted <dbl>, asd <dbl>
```

```r
head(test)
```

```
## # A tibble: 6 x 14
##   nerdiness[,1] extraverted critical dependable anxious  open reserved
##           <dbl>       <int>    <int>      <int>   <int> <int>    <int>
## 1       0.00656           7        6          6       3     5        5
## 2      -0.508             7        1          7       4     7        6
## 3      -0.573             2        5          5       7     5        6
## 4      -0.508             4        5          1       7     7        5
## 5      -1.28              6        6          4       3     6        1
## 6      -0.508             2        5          4       5     4        6
## # i 7 more variables: sympathetic <int>, disorganized <int>, calm <int>,
## #   uncreative <int>, age <int>, voted <dbl>, asd <dbl>
```

b)

Using the training data, fit two models predicting `nerdiness`. For the first, use `age`, `voted`, and `asd` as predictors. For the second, use all available predictors. Compute the LOOCV score for each. Which model has better predictive accuracy? (1 pt.)

A better predictive accuracy was obtained by the model using all available predictors.

```r
training_model_1 <- lm(nerdiness ~ age + voted + asd, data = training)

training_model_all <- lm(nerdiness ~ ., data = training)

loocv(training_model_1)
```

```
## [1] 1.019734
```

```r
loocv(training_model_all)
```

```
## [1] 0.8578562
```

c)

Repeat part b, but this time use k-fold cross-validation using $k = 10$. Which model has better predictive accuracy according to 10-fold CV? (1 pt.)

A better predictive accuracy was obtained by the model using all available predictors.

```r
glm_model_1 <- glm(nerdiness ~ age + voted + asd, data = training, family = "gaussian")

glm_model_all <- glm(nerdiness ~ ., data = training, family = "gaussian")

cv_m1 <- cv.glm(data= training,glmfit = glm_model_1, K=10)
cv_mfull <- cv.glm(data= training,glmfit = glm_model_all, K=10)

cv_m1$delta[1]
```

```
## [1] 1.021366
```

```r
cv_mfull$delta[1]
```

```
## [1] 0.8593075
```

d)

Fit a regression to the test data using the predictors from the best model according to LOOCV. Can we interpret the coefficients for th model on this new data? Why or why not? If we can, are there any statistically significant predictors at the .05 level? (1 pt.)

Even though this model contains all available predictors, it was chosen as the best representation of the data by model selection methods. Since it was not built with a theoretical background, it is not recommended to interpret the coefficients of a model that was derived from the same data.

```r
test_model <- lm(nerdiness ~ ., data = test)
summary(test_model)
```

```
##
## Call:
## lm(formula = nerdiness ~ ., data = test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8185 -0.5625  0.1641  0.6030  1.7307
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8245085  0.5461054   1.510 0.132792
## extraverted  -0.1568409  0.0457545  -3.428 0.000749 ***
## critical     -0.0258701  0.0374776  -0.690 0.490878
## dependable   -0.0402289  0.0408433  -0.985 0.325924
## anxious       0.0193569  0.0398732   0.485 0.627921
## open          0.1136288  0.0523365   2.171 0.031189 *
## reserved     -0.0235102  0.0465572  -0.505 0.614175
## sympathetic   0.0354530  0.0466355   0.760 0.448088
## disorganized -0.0443763  0.0370258  -1.199 0.232238
## calm         -0.0839830  0.0401759  -2.090 0.037943 *
## uncreative   -0.0981361  0.0397139  -2.471 0.014371 *
## age           0.0006951  0.0065466   0.106 0.915559
## voted         0.0106864  0.1488337   0.072 0.942837
## asd          -0.0317765  0.3098531  -0.103 0.918428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8721 on 186 degrees of freedom
## Multiple R-squared:  0.1855, Adjusted R-squared:  0.1286
## F-statistic: 3.258 on 13 and 186 DF,  p-value: 0.0001816
```