302 Final Project

Lisa

08/06/2021

```r
knitr::opts_chunk$set(echo = FALSE)

## If the package is not already installed then use
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(NHANES)
library(xtable)
library(car)
```

```
## Loading required package: carData
##
```

Model Diagnosis

```
##  30 126 167 190 270 275 281 330 367 403 531 569 594 630
##  20  88 116 131 183 186 191 225 251 280 369 399 411 434

##         ID Gender Age    Race3      Education MaritalStat
## 30   62632   male  29    Asian 9 - 11th Grade  NeverMarri
## 126 63812 female  47    Black 9 - 11th Grade   LivePartn
## 167 64347   male  50    White    High School  NeverMarri
## 190 64549   male  44 Hispanic   Some College     Separat
## 270 65823   male  37  Mexican 9 - 11th Grade     Separat
## 275 65946   male  34    White   College Grad     Separat
## 281 66027 female  40  Mexican   College Grad       Marri
## 330 66597 female  54  Mexican 9 - 11th Grade     Separat
## 367 67115   male  59 Hispanic   College Grad     Separat
## 403 67471 female  45    White    High School     Separat
## 531 69101   male  58 Hispanic      8th Grade     Separat
## 569 69607   male  40    Other      8th Grade      Divorc
## 594 69994 female  21    Other    High School  NeverMarri
## 630 70448   male  29  Mexican    High School     Separat
## 677 71072 female  50 Hispanic    High School     Separat
```

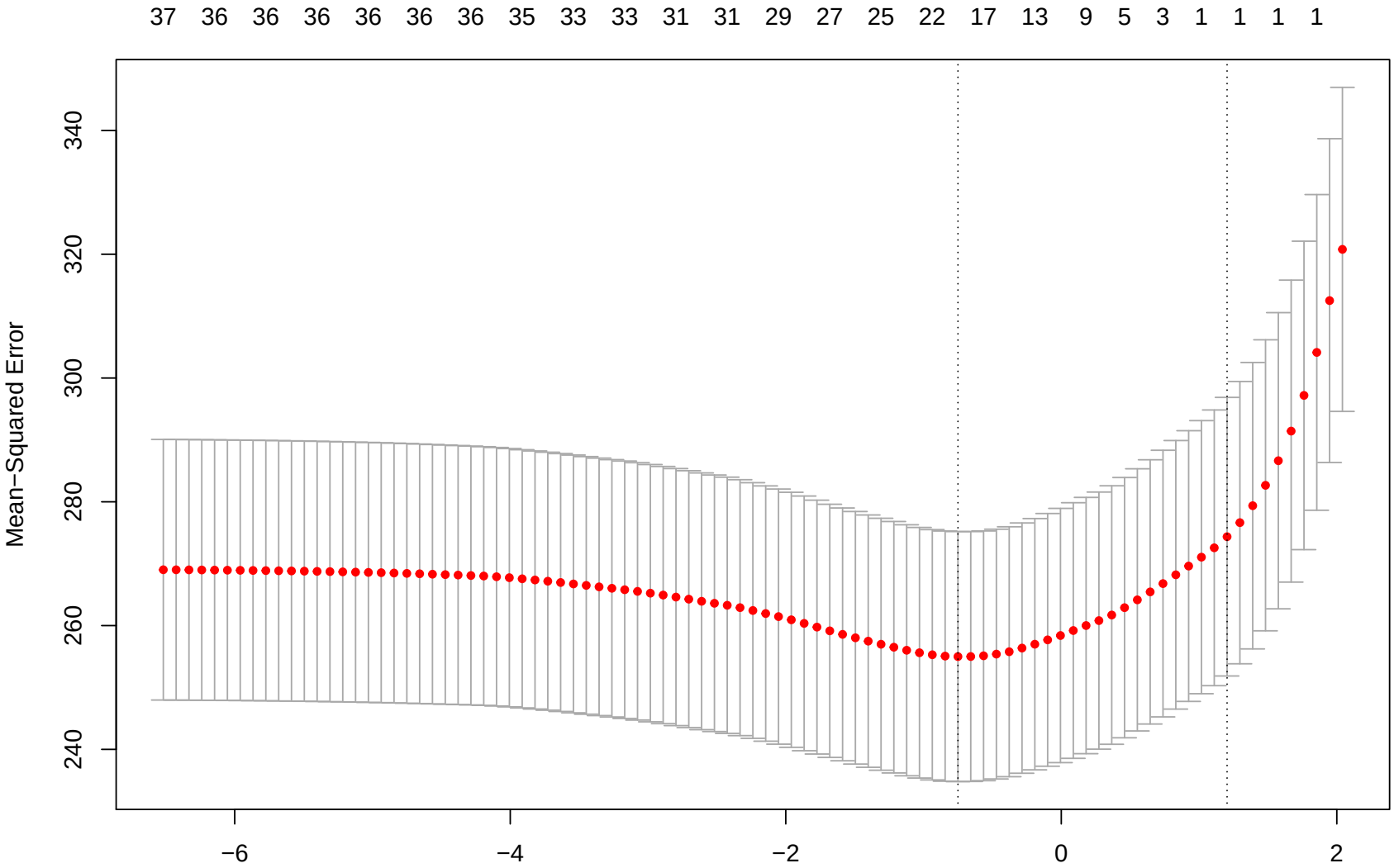
Checking for VIF

```
##
## Call:
## lm(formula = BPSysAve ~ Age + Poverty + Weight + Height
##     SleepHrsNight, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.674  -9.238  -1.143   8.420  77.802
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   155.79161   63.58573   2.450   0.0146 *
## Age             0.46884    0.04371  10.726   <2e-16 ***
## Poverty        -1.22978    0.45567  -2.699   0.0072 **
## Weight          0.44838    0.37385   1.199   0.2310
## Height         -0.36327    0.37320  -0.973   0.3308
## BMI            -1.05229    1.08564  -0.969   0.3329
## SleepHrsNight   0.39306    0.52447   0.749   0.4540
```

Variable Selection

```
## [1] "Gender"        "Age"             "Poverty"       "Weight
## [6] "SleepTrouble" "PhysActive"
```

```
## [1] "Gender" "Age"    "Weight" "Height"
```



Shrinkage Methods & Prediction Error

```
## [1] 269.4562
```

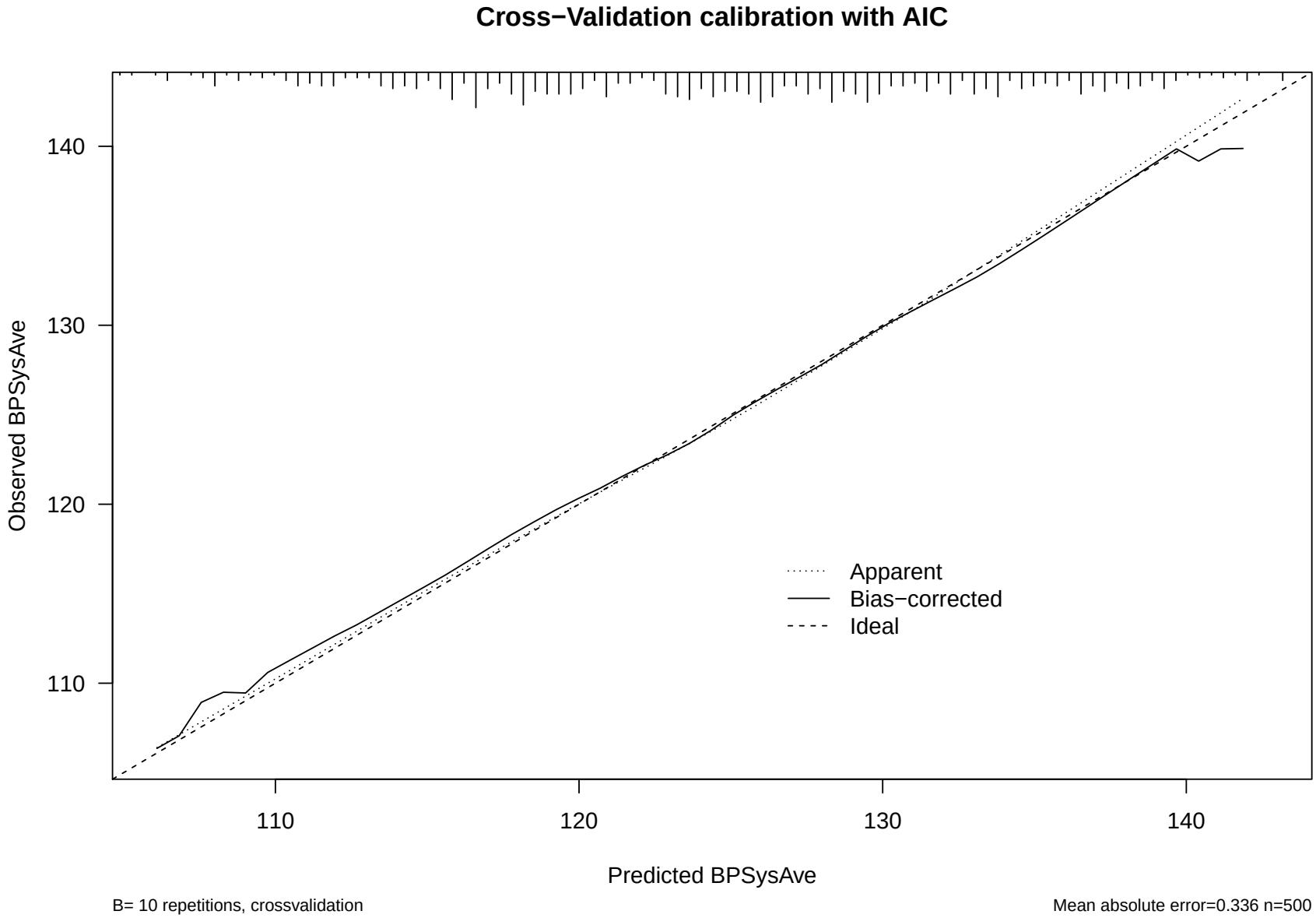
```
## [1] 280.7807
```

```
## [1] 265.1804
```

```
## [1] 265.7827
```

Prediction Error lowest for LASSO model so far.

Model Validation & Prediction Error




```
##
## n=500   Mean absolute error=0.336   Mean squared error=0
```

Select Model & explain the parameter estimates

Since we are interested in the prediction only, we should choose the model with the lowest prediction error. That is, we fit the model obtained by the LASSO Shrinkage Method:

```
##
## Call:
## lm(formula = BPSysAve ~ Age, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.806  -9.365  -1.782   8.152  79.194
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.27169    2.30731   43.89   <2e-16 ***
## Age           0.45668    0.04295   10.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

Conclusion

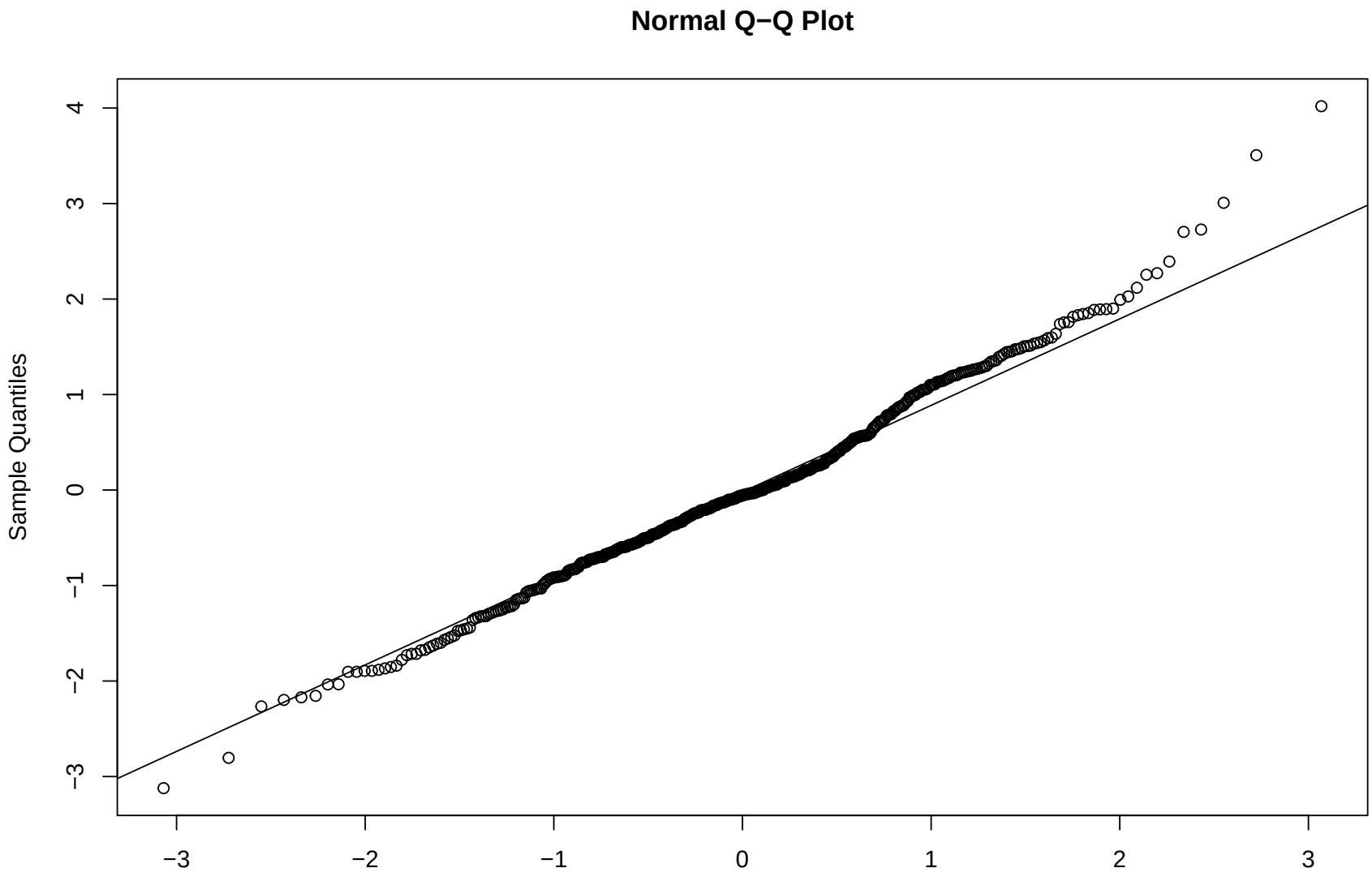
- As you increase 1 year of age, your BPSysAve will increase by around 0.46.

Association between SmokeNow and BPSysAve

```
## ------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cau
## If you need functions from both plyr and dplyr, please l
## library(plyr); library(dplyr)

## ------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename,
##     summarize

## The following object is masked from 'package:purrr':
```

Divide by Gender and Remove Unimportant Variables and try again??

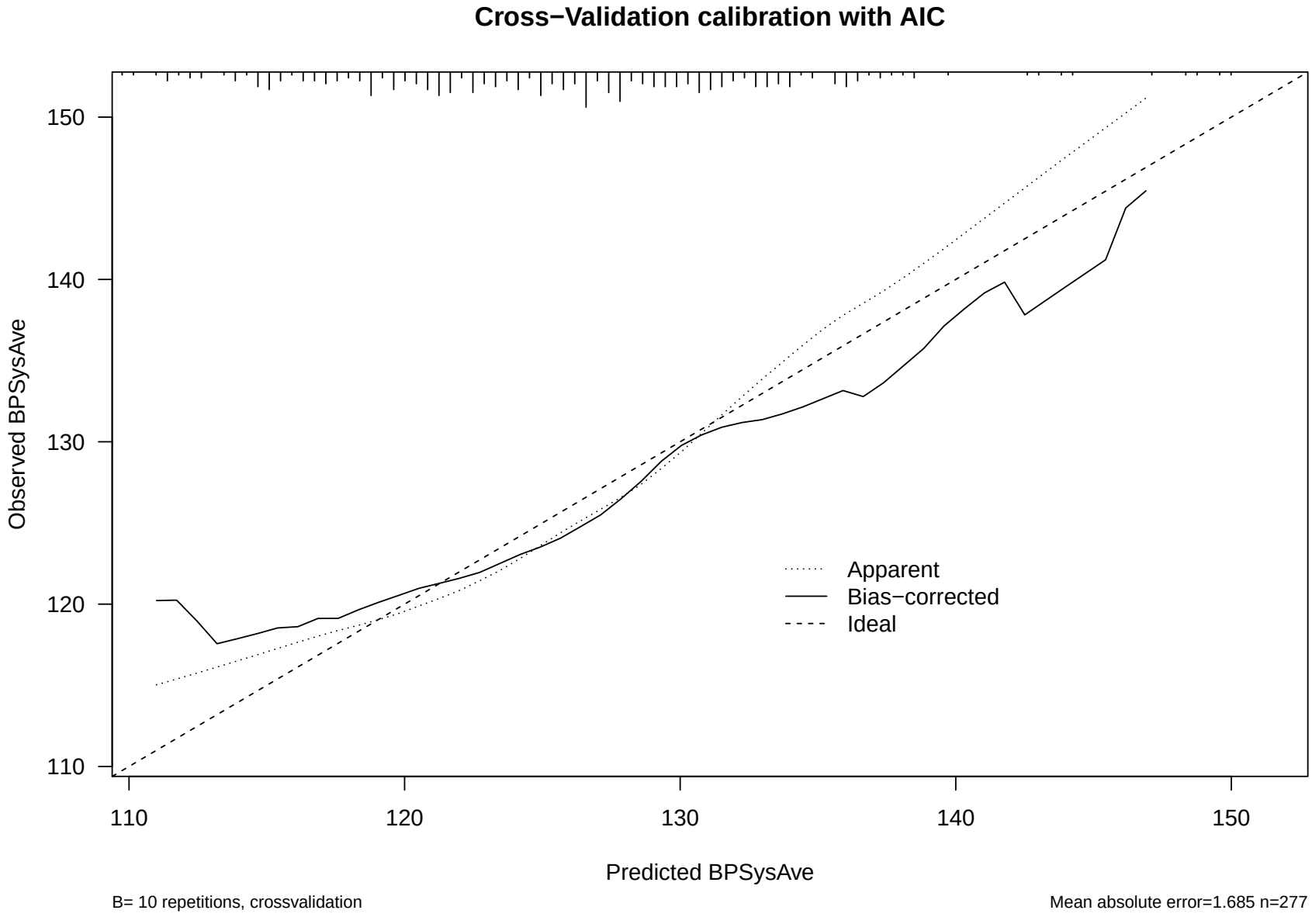
First, we remove the outliers according to DFBETAS and outliers as shown above.



Shrinkage Methods & Prediction Error

```
##
## Call:
## lm(formula = BPSysAve ~ ., data = train.male[, -c(1, 2)]
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.112  -7.900  -0.742   7.788  51.012
##
## Coefficients:
##                              Estimate Std. Error t value P
## (Intercept)                375.78982  126.32678   2.975
## Age                          0.32295    0.06522   4.952 1
## Race3Black                   8.52972    4.54679   1.876
## Race3Hispanic                7.84835    4.94145   1.588
## Race3Mexican                 4.72346    4.85118   0.974
## Race3White                   5.12563    4.09097   1.253
## Race3Other                  -5.70498    6.12213  -0.932
## Education9 - 11th Grade     -1.47914    3.42086  -0.432
```

Model Validation & Prediction Error

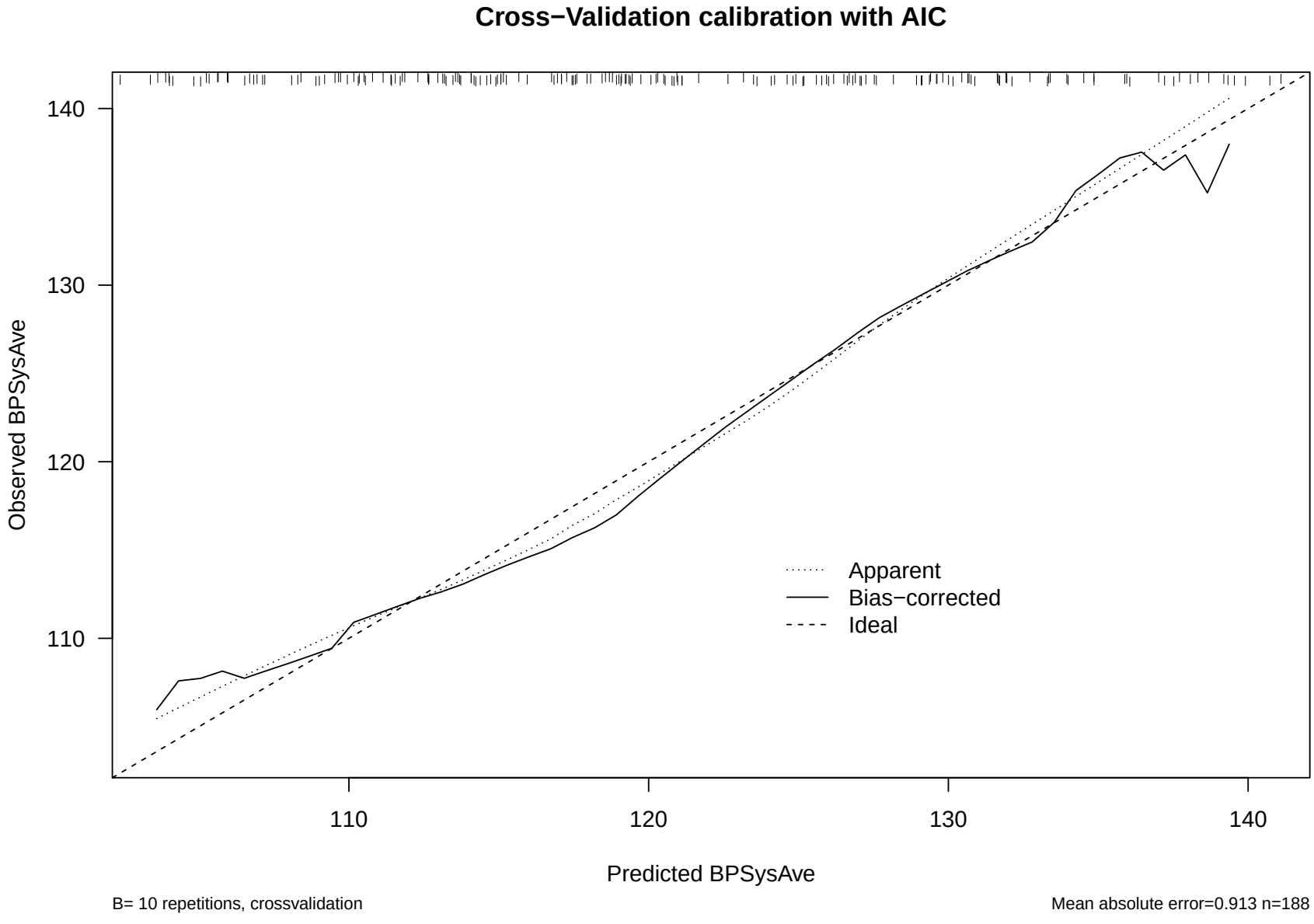



```
##
## n=277   Mean absolute error=1.685   Mean squared error=5
```

Shrinkage Methods & Prediction Error

```
##
## Call:
## lm(formula = BPSysAve ~ ., data = train.female[, -c(1, 2
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.0713  -5.9632  -0.0437   7.2503  27.7829
##
## Coefficients:
##                              Estimate Std. Error t value P
## (Intercept)                367.45077  118.81163   3.093
## Age                          0.65022    0.07708   8.435 2
## Race3Black                  10.99107    7.30290   1.505
## Race3Hispanic               -6.12266    8.02071  -0.763
## Race3Mexican                -2.55733    8.21991  -0.311
## Race3White                   8.71240    6.46752   1.347
## Race3Other                  11.65827    8.50072   1.371
## Education9 - 11th Grade     -5.37827    6.12033  -0.879
```

Model Validation & Prediction Error




```
##
## n=188   Mean absolute error=0.913   Mean squared error=1
```

Try fit simple linear regression

```
##
## Call:
## lm(formula = BPSysAve ~ as.factor(SmokeNow), data = trai
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.967  -8.990  -0.990   8.033  63.033
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>
## (Intercept)               124.967      0.905 138.089  < 2e
## as.factor(SmokeNow)Yes     -3.977      1.408  -2.824  0.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 14.95 on 463 degrees of freedom
## Multiple R-squared:  0.01693,    Adjusted R-squared:  0.
## F-statistic: 7.976 on 1 and 463 DF,  p-value: 0.004945
```

fEMALE OBESE

```
##
## Call:
## lm(formula = BPSysAve ~ as.factor(SmokeNow), data = trai
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.511 -10.511  -1.511   8.489  39.489
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>
## (Intercept)               124.511      1.551  80.272  < 2e
## as.factor(SmokeNow)Yes     -9.000      2.706  -3.325  0.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 14.88 on 135 degrees of freedom
## Multiple R-squared:  0.07571,    Adjusted R-squared:  0.
## F-statistic: 11.06 on 1 and 135 DF,  p-value: 0.001138
```

Play around

```
##
## Call:
## lm(formula = BPSysAve ~ ., data = train.female.thin[, -c
##     11)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9926  -3.0093   0.2239   2.5460  17.2339
##
## Coefficients:
##                          Estimate Std. Error t value Pr
## (Intercept)             119.7954    84.0539   1.425  (
## Age                       0.8322     0.2155   3.862  (
## Race3Black               28.3626    19.7327   1.437  (
## Race3Hispanic            21.7795    21.2020   1.027  (
## Race3Mexican             27.3745    19.2562   1.422  (
## Race3White               25.6212    12.3728   2.071  (
## Race3Other               43.8520    18.9214   2.318  (
```