GPH 2353 Project

Yi Yang

2023-03-14

```
knitr::opts_chunk$set(echo = TRUE)
library(haven)
library(psych)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(MASS)
library(psych)
library(leaps)
library(pander)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following objects are masked from 'package:car':
##
##     logit, vif

## The following object is masked from 'package:psych':
##
##     logit
```

```
library(readr)
library(lmtest)
library(corrplot)
library(r02pro)
```

HEAD Data Preparation

```
my_data <- read.csv("IHDP.csv")
head(my_data)
```

```
##   momage b.marr momed work.dur prenatal cig sex   bw bwg preterm black hispanic
## 1     33      1     4        1         1   0   1 1559   0      10     0        0
## 2     22      0     1        0         1   0   1 2240   1       3     1        0
## 3     13      0     1        0         1   0   1 1900   0       6     1        0
## 4     25      1     4        1         1   0   1 1550   0       8     1        0
## 5     19      0     1        0         1   1   1 2270   1       5     1        0
## 6     19      0     2        1         1   1   0 1550   0       4     1        0
##   white lths hs ltcoll college dayskidh income treat ppvtr.36
## 1     1    0  0      0       1       31  42500     1      111
## 2     0    1  0      0       0        4   5000     1       81
## 3     0    1  0      0       0        9  12500     1       92
## 4     0    0  0      0       1       50  42500     1      103
## 5     0    1  0      0       0        4   5000     1       81
## 6     0    0  1      0       0       13  12500     1       94
```

```
dim(my_data)
```

```
## [1] 4381   21
```

```
for (i in colnames(my_data)) {
  print(i)
  print(sum(is.na(my_data$i)))
}
```

```
## [1] "momage"
## [1] 0
## [1] "b.marr"
## [1] 0
## [1] "momed"
## [1] 0
## [1] "work.dur"
## [1] 0
## [1] "prenatal"
## [1] 0
## [1] "cig"
## [1] 0
## [1] "sex"
## [1] 0
## [1] "bw"
## [1] 0
## [1] "bwg"
## [1] 0
## [1] "preterm"
## [1] 0
## [1] "black"
## [1] 0
## [1] "hispanic"
## [1] 0
## [1] "white"
## [1] 0
## [1] "lths"
## [1] 0
## [1] "hs"
## [1] 0
## [1] "ltcoll"
## [1] 0
## [1] "college"
## [1] 0
## [1] "dayskidh"
## [1] 0
## [1] "income"
## [1] 0
## [1] "treat"
## [1] 0
## [1] "ppvtr.36"
## [1] 0
```

```
dim(my_data)
```

```
## [1] 4381   21
```

```
### START
summary(my_data)
```

```
##      momage          b.marr           momed          work.dur
##  Min.   :13.0   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:21.0   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000
```

```
##  Median :24.0   Median :1.0000   Median :2.000   Median :1.0000
##  Mean   :23.8   Mean   :0.6699   Mean   :2.048   Mean   :0.6188
##  3rd Qu.:26.0   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :41.0   Max.   :1.0000   Max.   :4.000   Max.   :1.0000
##     prenatal           cig              sex               bw
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1503
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2892
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :3289
##  Mean   :0.9852   Mean   :0.3314   Mean   :0.4962   Mean   :3247
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3657
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :7598
##       bwg            preterm           black           hispanic
##  Min.   :0.0000   Min.   :-7.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median : 1.000   Median :0.0000   Median :0.0000
##  Mean   :0.9493   Mean   : 1.503   Mean   :0.2979   Mean   :0.2054
##  3rd Qu.:1.0000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :14.000   Max.   :1.0000   Max.   :1.0000
##      white             lths             hs             ltcoll
##  Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00   Median :0.0000   Median :0.0000
##  Mean   :0.4967   Mean   :0.31   Mean   :0.4154   Mean   :0.1911
##  3rd Qu.:1.0000   3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00   Max.   :1.0000   Max.   :1.0000
##     college           dayskidh          income            treat
##  Min.   :0.00000   Min.   :  0.000   Min.   : -55307   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:  2.000   1st Qu.:   7729   1st Qu.:0.00000
##  Median :0.00000   Median :  3.000   Median :  17025   Median :0.00000
##  Mean   :0.08354   Mean   :  4.864   Mean   :  28085   Mean   :0.06619
##  3rd Qu.:0.00000   3rd Qu.:  5.000   3rd Qu.:  31200   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :100.000   Max.   :1378212   Max.   :1.00000
##     ppvtr.36
##  Min.   : 33.00
##  1st Qu.: 73.00
##  Median : 88.00
##  Mean   : 86.43
##  3rd Qu.:101.00
##  Max.   :129.00
```

```
names(my_data)
```

```
##  [1] "momage"   "b.marr"   "momed"    "work.dur" "prenatal" "cig"
##  [7] "sex"      "bw"       "bwg"      "preterm"  "black"    "hispanic"
## [13] "white"    "lths"     "hs"       "ltcoll"   "college"  "dayskidh"
## [19] "income"   "treat"    "ppvtr.36"
```

```
head(my_data)
```

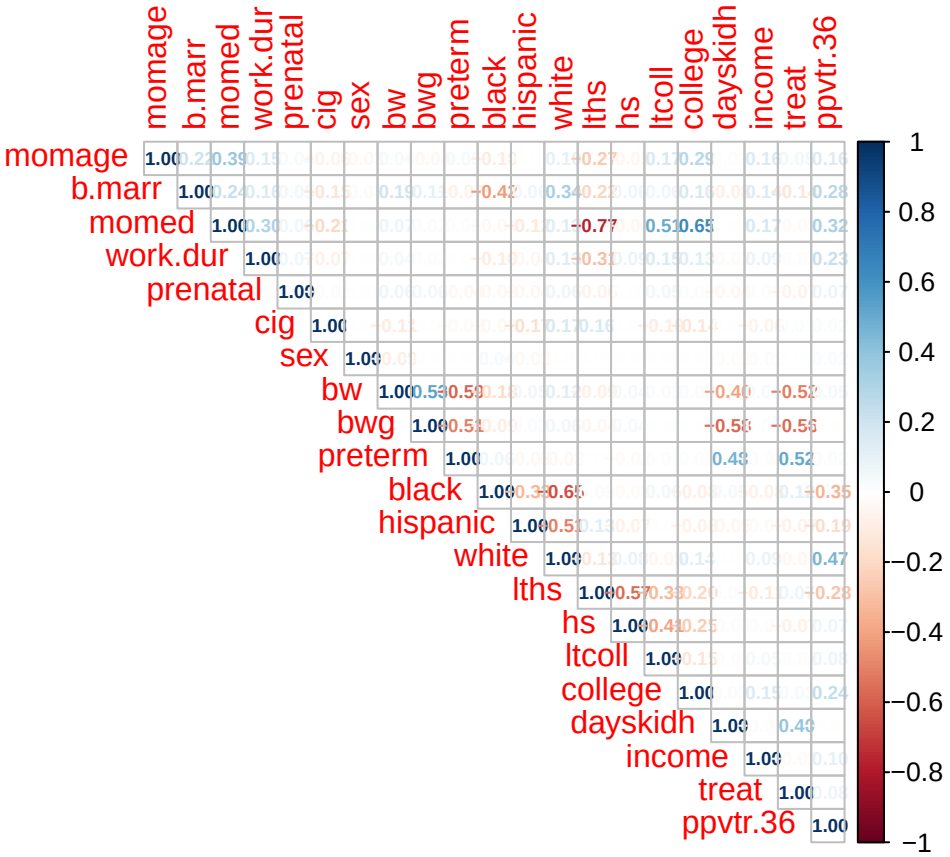
```
##   momage b.marr momed work.dur prenatal cig sex   bw bwg preterm black hispanic
## 1     33      1     4        1        1   0   1 1559   0      10     0        0
## 2     22      0     1        0        1   0   1 2240   1       3     1        0
## 3     13      0     1        0        1   0   1 1900   0       6     1        0
```

```
## 4     25      1     4        1         1   0    1 1550    0       8     1         0
## 5     19      0     1        0         1   1    1 2270    1       5     1         0
## 6     19      0     2        1         1   1    0 1550    0       4     1         0
##   white lths hs ltcoll college dayskidh income treat ppvtr.36
## 1     1    0  0       0       1       31  42500     1      111
## 2     0    1  0       0       0        4   5000     1       81
## 3     0    1  0       0       0        9  12500     1       92
## 4     0    0  0       0       1       50  42500     1      103
## 5     0    1  0       0       0        4   5000     1       81
## 6     0    0  1       0       0       13  12500     1       94
```

```
my_data = na.omit(my_data)
```

Check Correlation between each variables

```
M=cor(my_data)
corrplot(M,method = "number",type="upper",number.cex = 0.6) # make correlation plot
```



Define Training and test dataset

```
set.seed(0)
tr_size <- nrow(my_data) * 0.7 # training sample size
tr_ind <-sample(nrow(my_data), tr_size)
data_tr <-my_data[tr_ind, ] # training data
data_te <-my_data[-tr_ind, ] # test data
ncol(my_data)
```

```
## [1] 21
```

```
nrow(my_data)
```

```
## [1] 4381
```

```
nrow(data_tr)
```

```
## [1] 3066
```

```
nrow(data_te)
```

```
## [1] 1315
```

Train Model

```
model <- lm(ppvtr.36~., data = data_tr)
summary(model)
```

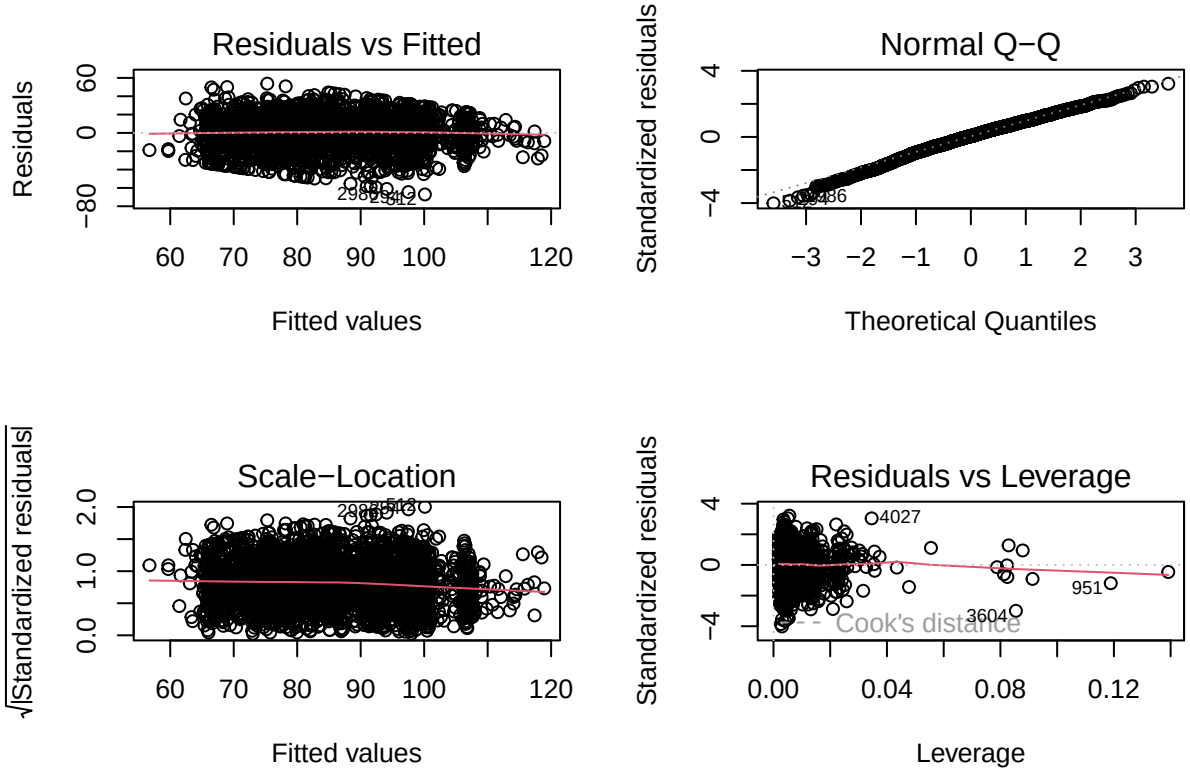
```
##
## Call:
## lm(formula = ppvtr.36 ~ ., data = data_tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.115  -9.930   0.899  11.322  53.710
##
## Coefficients: (3 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.718e+01  6.026e+00  11.148  < 2e-16 ***
## momage      -6.517e-02  9.867e-02  -0.660  0.50900
## b.marr       2.225e+00  7.586e-01   2.932  0.00339 **
## momed        6.810e+00  1.288e+00   5.286 1.34e-07 ***
## work.dur     4.344e+00  6.671e-01   6.512 8.65e-11 ***
## prenatal     4.383e+00  2.368e+00   1.851  0.06424 .
## cig          9.519e-01  6.911e-01   1.377  0.16847
## sex          8.090e-02  6.123e-01   0.132  0.89491
## bw           4.431e-04  6.671e-04   0.664  0.50665
## bwg          1.604e+00  1.960e+00   0.818  0.41326
## preterm      2.819e-02  1.761e-01   0.160  0.87283
## black       -1.790e+01  8.059e-01 -22.210  < 2e-16 ***
## hispanic    -1.502e+01  8.403e-01 -17.877  < 2e-16 ***
## white               NA         NA      NA       NA
## lths         4.160e+00  3.028e+00   1.374  0.16961
## hs           4.179e+00  1.818e+00   2.299  0.02156 *
## ltcoll              NA         NA      NA       NA
## college             NA         NA      NA       NA
## dayskidh    -1.039e-01  6.340e-02  -1.638  0.10148
## income       1.121e-06  4.606e-06   0.243  0.80766
## treat        1.240e+01  1.577e+00   7.862 5.19e-15 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.75 on 3048 degrees of freedom
## Multiple R-squared:  0.3255, Adjusted R-squared:  0.3217
## F-statistic: 86.51 on 17 and 3048 DF,  p-value: < 2.2e-16
```

```
alias(model)
```

```
## Model :
## ppvtr.36 ~ momage + b.marr + momed + work.dur + prenatal + cig +
##     sex + bw + bwg + preterm + black + hispanic + white + lths +
##     hs + ltcoll + college + dayskidh + income + treat
##
## Complete :
##          (Intercept) momage b.marr momed work.dur prenatal cig sex bw bwg
## white     1           0      0      0     0        0        0   0   0  0
## ltcoll    4           0      0     -1     0        0        0   0   0  0
## college  -3           0      0      1     0        0        0   0   0  0
##          preterm black hispanic lths hs dayskidh income treat
## white     0      -1    -1        0    0  0        0      0
## ltcoll    0       0     0       -3   -2  0        0      0
## college   0       0     0        2    1  0        0      0
```

```
par (mfrow = c(2,2))
plot (model)
```



```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.99312, p-value = 6.716e-11
```

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.0416, p-value = 0.7604
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 65.171, df = 17, p-value = 1.438e-07
```

```
AA<-rstudent(model) # Compute studentized residuals to check outliers.
p<-ncol(data_tr)
n<-nrow(data_tr)
which(abs(AA)>qt(1-0.05/(n*2),n-p-1))
```

```
## named integer(0)
```

```
plot(model,which=4)
#vif(model)

model1<-lm(ppvtr.36~. -white-ltcoll-college,data_tr)
summary(model1)
```

```
##
## Call:
## lm(formula = ppvtr.36 ~ . - white - ltcoll - college, data = data_tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.115  -9.930   0.899  11.322  53.710
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.718e+01  6.026e+00  11.148  < 2e-16 ***
## momage      -6.517e-02  9.867e-02  -0.660  0.50900
```

```
## b.marr       2.225e+00  7.586e-01   2.932  0.00339 **
## momed        6.810e+00  1.288e+00   5.286 1.34e-07 ***
## work.dur     4.344e+00  6.671e-01   6.512 8.65e-11 ***
## prenatal     4.383e+00  2.368e+00   1.851  0.06424 .
## cig          9.519e-01  6.911e-01   1.377  0.16847
## sex          8.090e-02  6.123e-01   0.132  0.89491
## bw           4.431e-04  6.671e-04   0.664  0.50665
## bwg          1.604e+00  1.960e+00   0.818  0.41326
## preterm      2.819e-02  1.761e-01   0.160  0.87283
## black       -1.790e+01  8.059e-01 -22.210  < 2e-16 ***
## hispanic    -1.502e+01  8.403e-01 -17.877  < 2e-16 ***
## lths         4.160e+00  3.028e+00   1.374  0.16961
## hs           4.179e+00  1.818e+00   2.299  0.02156 *
## dayskidh    -1.039e-01  6.340e-02  -1.638  0.10148
## income       1.121e-06  4.606e-06   0.243  0.80766
## treat        1.240e+01  1.577e+00   7.862 5.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.75 on 3048 degrees of freedom
## Multiple R-squared:  0.3255, Adjusted R-squared:  0.3217
## F-statistic: 86.51 on 17 and 3048 DF,  p-value: < 2.2e-16
```

```
vif(model1)
```

```
##    momage    b.marr     momed  work.dur  prenatal       cig       sex        bw
##  1.234599  1.371486 15.225970  1.143206  1.022086  1.160391  1.024965  2.023423
##       bwg   preterm     black  hispanic      lths        hs  dayskidh    income
##  2.127012  1.921003  1.481266  1.263546 21.465948  8.747523  1.762281  1.052858
##     treat
##  1.742561
```

```
step(model1)
```

```
## Start:  AIC=17298.66
## ppvtr.36 ~ (momage + b.marr + momed + work.dur + prenatal + cig +
##     sex + bw + bwg + preterm + black + hispanic + white + lths +
##     hs + ltcoll + college + dayskidh + income + treat) - white -
##     ltcoll - college
##
##             Df Sum of Sq    RSS   AIC
## - sex        1         5 854685 17297
## - preterm    1         7 854687 17297
## - income     1        17 854697 17297
## - momage     1       122 854803 17297
## - bw         1       124 854804 17297
## - bwg        1       188 854868 17297
## - lths       1       529 855209 17299
## - cig        1       532 855212 17299
## <none>                   854680 17299
## - dayskidh   1       753 855433 17299
## - prenatal   1       961 855641 17300
## - hs         1      1482 856163 17302
```

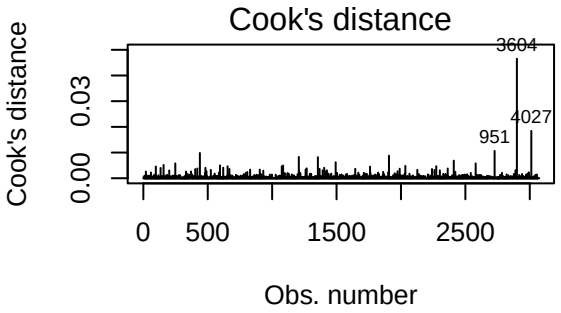
```
## - b.marr    1      2411 857091 17305
## - momed     1      7835 862516 17325
## - work.dur  1     11890 866571 17339
## - treat     1     17334 872014 17358
## - hispanic  1     89618 944298 17602
## - black     1    138319 992999 17757
##
## Step:  AIC=17296.68
## ppvtr.36 ~ momage + b.marr + momed + work.dur + prenatal + cig +
##     bw + bwg + preterm + black + hispanic + lths + hs + dayskidh +
##     income + treat
##
##            Df Sum of Sq    RSS   AIC
## - preterm   1         7 854692 17295
## - income    1        16 854702 17295
## - bw        1       119 854805 17295
## - momage    1       123 854808 17295
## - bwg       1       188 854873 17295
## - lths      1       529 855214 17297
## - cig       1       529 855215 17297
## <none>                  854685 17297
## - dayskidh  1       761 855446 17297
## - prenatal  1       963 855648 17298
## - hs        1      1482 856167 17300
## - b.marr    1      2411 857096 17303
## - momed     1      7837 862522 17323
## - work.dur  1     11887 866573 17337
## - treat     1     17330 872015 17356
## - hispanic  1     89841 944526 17601
## - black     1    138315 993001 17755
##
## Step:  AIC=17294.71
## ppvtr.36 ~ momage + b.marr + momed + work.dur + prenatal + cig +
##     bw + bwg + black + hispanic + lths + hs + dayskidh + income +
##     treat
##
##            Df Sum of Sq    RSS   AIC
## - income    1        16 854708 17293
## - bw        1       116 854808 17293
## - momage    1       121 854813 17293
## - bwg       1       183 854875 17293
## - cig       1       523 855215 17295
## - lths      1       525 855216 17295
## <none>                  854692 17295
## - dayskidh  1       760 855452 17295
## - prenatal  1       963 855655 17296
## - hs        1      1476 856168 17298
## - b.marr    1      2414 857106 17301
## - momed     1      7831 862523 17321
## - work.dur  1     11888 866579 17335
## - treat     1     18337 873029 17358
## - hispanic  1     89855 944547 17599
## - black     1    138812 993504 17754
##
```

```
## Step:  AIC=17292.76
## ppvtr.36 ~ momage + b.marr + momed + work.dur + prenatal + cig +
##     bw + bwg + black + hispanic + lths + hs + dayskidh + treat
##
##            Df Sum of Sq    RSS   AIC
## - bw        1       115 854823 17291
## - momage    1       115 854823 17291
## - bwg       1       181 854889 17291
## - cig       1       519 855227 17293
## - lths      1       538 855246 17293
## <none>                  854708 17293
## - dayskidh  1       765 855473 17294
## - prenatal  1       963 855671 17294
## - hs        1      1498 856205 17296
## - b.marr    1      2460 857168 17300
## - momed     1      7947 862655 17319
## - work.dur  1     11936 866643 17333
## - treat     1     18321 873029 17356
## - hispanic  1     89940 944647 17598
## - black     1    138817 993525 17752
##
## Step:  AIC=17291.17
## ppvtr.36 ~ momage + b.marr + momed + work.dur + prenatal + cig +
##     bwg + black + hispanic + lths + hs + dayskidh + treat
##
##            Df Sum of Sq    RSS   AIC
## - momage    1       105 854928 17290
## - bwg       1       278 855100 17290
## - cig       1       464 855287 17291
## - lths      1       534 855356 17291
## <none>                  854823 17291
## - dayskidh  1       869 855691 17292
## - prenatal  1       972 855795 17293
## - hs        1      1493 856316 17295
## - b.marr    1      2516 857338 17298
## - momed     1      7943 862765 17318
## - work.dur  1     11892 866715 17332
## - treat     1     19075 873897 17357
## - hispanic  1     90319 945141 17597
## - black     1    142171 996993 17761
##
## Step:  AIC=17289.55
## ppvtr.36 ~ b.marr + momed + work.dur + prenatal + cig + bwg +
##     black + hispanic + lths + hs + dayskidh + treat
##
##            Df Sum of Sq    RSS   AIC
## - bwg       1       296 855224 17289
## - cig       1       450 855378 17289
## - lths      1       519 855447 17289
## <none>                  854928 17290
## - dayskidh  1       850 855778 17291
## - prenatal  1       965 855893 17291
## - hs        1      1474 856402 17293
## - b.marr    1      2435 857363 17296
```

```
## - momed     1      7841 862769 17316
## - work.dur  1     11824 866752 17330
## - treat     1     18975 873903 17355
## - hispanic  1     90646 945574 17597
## - black     1    142082 997010 17759
##
## Step:  AIC=17288.61
## ppvtr.36 ~ b.marr + momed + work.dur + prenatal + cig + black +
##     hispanic + lths + hs + dayskidh + treat
##
##            Df Sum of Sq    RSS   AIC
## - cig       1       444 855668 17288
## - lths      1       540 855764 17289
## <none>                  855224 17289
## - prenatal  1       998 856222 17290
## - hs        1      1515 856739 17292
## - dayskidh  1      1938 857162 17294
## - b.marr    1      2449 857673 17295
## - momed     1      7935 863158 17315
## - work.dur  1     11785 867009 17329
## - treat     1     20453 875677 17359
## - hispanic  1     91222 946446 17597
## - black     1    142643 997867 17760
##
## Step:  AIC=17288.21
## ppvtr.36 ~ b.marr + momed + work.dur + prenatal + black + hispanic +
##     lths + hs + dayskidh + treat
##
##            Df Sum of Sq     RSS   AIC
## - lths      1       500  856168 17288
## <none>                   855668 17288
## - prenatal  1       979  856647 17290
## - hs        1      1433  857101 17291
## - dayskidh  1      1978  857646 17293
## - b.marr    1      2210  857879 17294
## - momed     1      7652  863320 17314
## - work.dur  1     11663  867331 17328
## - treat     1     20400  876068 17358
## - hispanic  1    100492  956160 17627
## - black     1    149904 1005572 17781
##
## Step:  AIC=17288
## ppvtr.36 ~ b.marr + momed + work.dur + prenatal + black + hispanic +
##     hs + dayskidh + treat
##
##            Df Sum of Sq     RSS   AIC
## <none>                   856168 17288
## - prenatal  1       938  857106 17289
## - dayskidh  1      1898  858066 17293
## - b.marr    1      2228  858396 17294
## - hs        1      2374  858542 17295
## - work.dur  1     11421  867589 17327
## - treat     1     20569  876737 17359
## - momed     1     54391  910559 17475
```

```
## - hispanic  1    101613  957781 17630
## - black     1    153676 1009844 17792
```

```
##
## Call:
## lm(formula = ppvtr.36 ~ b.marr + momed + work.dur + prenatal +
##     black + hispanic + hs + dayskidh + treat, data = data_tr)
##
## Coefficients:
## (Intercept)       b.marr        momed     work.dur     prenatal        black
##     75.6378       2.0968       4.9870       4.2410       4.3253     -18.2679
##    hispanic           hs     dayskidh        treat
##    -15.4274       1.8241      -0.1378      11.5228
```



AIC Selection

```
lmod_AIC_B<-lm(ppvtr.36 ~ prenatal+dayskidh +b.marr+hs+work.dur+treat
               +momed+hispanic+black, data = data_tr) # AIC selected model
sum_AIC_B<-summary(lmod_AIC_B)
sum_AIC_B
```

```
##
## Call:
## lm(formula = ppvtr.36 ~ prenatal + dayskidh + b.marr + hs + work.dur +
##     treat + momed + hispanic + black, data = data_tr)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.986  -9.870   1.014  11.314  53.553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75.63785    2.53206  29.872  < 2e-16 ***
## prenatal      4.32529    2.36385   1.830  0.06738 .
## dayskidh     -0.13784    0.05296  -2.603  0.00929 **
## b.marr        2.09676    0.74352   2.820  0.00483 **
## hs            1.82412    0.62658   2.911  0.00363 **
## work.dur      4.24096    0.66422   6.385 1.98e-10 ***
## treat        11.52285    1.34479   8.569  < 2e-16 ***
## momed         4.98703    0.35792  13.933  < 2e-16 ***
## hispanic    -15.42742    0.81007 -19.045  < 2e-16 ***
## black       -18.26788    0.77999 -23.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 3056 degrees of freedom
## Multiple R-squared:  0.3243, Adjusted R-squared:  0.3223
## F-statistic:   163 on 9 and 3056 DF,  p-value: < 2.2e-16
```

```
vif(lmod_AIC_B)
```

```
## prenatal dayskidh   b.marr       hs work.dur    treat    momed hispanic
## 1.019546 1.230687 1.318525 1.040581 1.134466 1.268290 1.176409 1.175176
##    black
## 1.388713
```

BIC Selection

```
fit_null<-lm(ppvtr.36~1,data_tr)
step(fit_null, scope = list(lower = fit_null, upper = model1), direction = "both",
criterion = "BIC", k = log(n))
```

```
## Start:  AIC=18477.87
## ppvtr.36 ~ 1
##
##            Df Sum of Sq     RSS   AIC
## + black     1    140051 1127014 18127
## + momed     1    119548 1147516 18182
## + lths      1     97411 1169654 18241
## + b.marr    1     85215 1181849 18272
## + work.dur  1     69847 1197218 18312
## + hispanic  1     58677 1208387 18341
## + momage    1     30742 1236323 18411
## + income    1      8417 1258648 18466
## + treat     1      7653 1259412 18467
## + hs        1      7605 1259459 18467
## + prenatal  1      7130 1259935 18469
```

```
## <none>                  1267065 18478
## + bw        1      3011 1264054 18479
## + cig       1       878 1266186 18484
## + preterm   1       862 1266203 18484
## + bwg       1        23 1267041 18486
## + dayskidh  1         8 1267057 18486
## + sex       1         0 1267064 18486
##
## Step:  AIC=18126.77
## ppvtr.36 ~ black
##
##            Df Sum of Sq     RSS   AIC
## + hispanic  1    150438  976576 17696
## + momed     1    113031 1013983 17811
## + lths      1     92903 1034111 17871
## + work.dur  1     51489 1075524 17991
## + b.marr    1     21934 1105080 18075
## + momage    1     18613 1108401 18084
## + treat     1     18052 1108961 18085
## + hs        1      5496 1121518 18120
## + prenatal  1      5317 1121697 18120
## + income    1      5297 1121716 18120
## + preterm   1      3149 1123864 18126
## <none>                  1127014 18127
## + bwg       1       841 1126173 18133
## + dayskidh  1       520 1126493 18133
## + sex       1       399 1126615 18134
## + bw        1       308 1126706 18134
## + cig       1       276 1126738 18134
## - black     1    140051 1267065 18478
##
## Step:  AIC=17695.52
## ppvtr.36 ~ black + hispanic
##
##            Df Sum of Sq     RSS   AIC
## + momed     1     81912  894664 17435
## + lths      1     60836  915740 17506
## + work.dur  1     39439  937137 17577
## + momage    1     14741  961835 17657
## + treat     1     13232  963344 17662
## + b.marr    1     11694  964882 17667
## + cig       1      3611  972965 17692
## + income    1      3139  973437 17694
## + preterm   1      2645  973931 17695
## <none>                  976576 17696
## + prenatal  1      2025  974551 17697
## + hs        1      1185  975391 17700
## + bwg       1       885  975691 17701
## + bw        1       296  976280 17703
## + dayskidh  1       111  976465 17703
## + sex       1        30  976546 17704
## - hispanic  1    150438 1127014 18127
## - black     1    231811 1208387 18341
##
```

```
## Step:  AIC=17434.96
## ppvtr.36 ~ black + hispanic + momed
##
##            Df Sum of Sq     RSS   AIC
## + treat     1     16511  878153 17386
## + work.dur  1     14268  880396 17394
## + hs        1      2768  891896 17434
## + preterm   1      2500  892164 17434
## <none>                   894664 17435
## + b.marr    1      2096  892568 17436
## + lths      1      1831  892833 17437
## + bw        1      1707  892957 17437
## + bwg       1      1442  893222 17438
## + prenatal  1       806  893858 17440
## + dayskidh  1       149  894515 17443
## + income    1       115  894548 17443
## + momage    1       111  894553 17443
## + cig       1        39  894625 17443
## + sex       1         5  894658 17443
## - momed     1     81912  976576 17696
## - hispanic  1    119319 1013983 17811
## - black     1    211901 1106565 18079
##
## Step:  AIC=17385.87
## ppvtr.36 ~ black + hispanic + momed + treat
##
##            Df Sum of Sq     RSS   AIC
## + work.dur  1     14089  864064 17344
## + hs        1      4083  874070 17380
## + b.marr    1      3543  874610 17382
## + lths      1      3082  875071 17383
## <none>                   878153 17386
## + dayskidh  1      2232  875922 17386
## + bwg       1      1610  876544 17388
## + prenatal  1      1426  876727 17389
## + bw        1       804  877349 17391
## + preterm   1       423  877730 17392
## + income    1       161  877993 17393
## + cig       1        45  878109 17394
## + sex       1         1  878152 17394
## + momage    1         0  878153 17394
## - treat     1     16511  894664 17435
## - momed     1     85191  963344 17662
## - hispanic  1    114023  992176 17752
## - black     1    221422 1099575 18067
##
## Step:  AIC=17344.31
## ppvtr.36 ~ black + hispanic + momed + treat + work.dur
##
##            Df Sum of Sq     RSS   AIC
## + b.marr    1      2654  861410 17343
## + hs        1      2575  861489 17343
## <none>                   864064 17344
## + dayskidh  1      2054  862011 17345
```

```
## + lths      1      1661  862404 17346
## + bwg       1      1603  862462 17347
## + prenatal  1      1078  862987 17349
## + bw        1       839  863225 17349
## + preterm   1       465  863599 17351
## + cig       1       131  863933 17352
## + income    1        60  864004 17352
## + momage    1        21  864043 17352
## + sex       1         3  864061 17352
## - work.dur  1     14089  878153 17386
## - treat     1     16332  880396 17394
## - momed     1     59466  923530 17540
## - hispanic  1    111090  975154 17707
## - black     1    207301 1071366 17996
##
## Step:  AIC=17342.91
## ppvtr.36 ~ black + hispanic + momed + treat + work.dur + b.marr
##
##            Df Sum of Sq     RSS   AIC
## + hs        1      2301  859109 17343
## <none>                    861410 17343
## + dayskidh  1      1991  859419 17344
## - b.marr    1      2654  864064 17344
## + bwg       1      1535  859876 17346
## + lths      1      1463  859948 17346
## + prenatal  1       979  860431 17348
## + bw        1       640  860771 17349
## + preterm   1       446  860965 17349
## + cig       1       359  861052 17350
## + momage    1        98  861312 17351
## + income    1        10  861400 17351
## + sex       1         4  861407 17351
## - work.dur  1     13200  874610 17382
## - treat     1     17556  878967 17397
## - momed     1     52710  914121 17517
## - hispanic  1    107678  969089 17696
## - black     1    154995 1016406 17842
##
## Step:  AIC=17342.74
## ppvtr.36 ~ black + hispanic + momed + treat + work.dur + b.marr +
##     hs
##
##            Df Sum of Sq     RSS   AIC
## <none>                    859109 17343
## - hs        1      2301  861410 17343
## - b.marr    1      2380  861489 17343
## + dayskidh  1      2003  857106 17344
## + bwg       1      1499  857611 17345
## + prenatal  1      1043  858066 17347
## + bw        1       610  858500 17349
## + preterm   1       471  858639 17349
## + cig       1       428  858681 17349
## + lths      1       378  858731 17349
## + momage    1        81  859028 17351
```

```
## + income    1        14  859095 17351
## + sex       1         6  859104 17351
## - work.dur  1     11866  870975 17377
## - treat     1     18484  877593 17400
## - momed     1     54476  913585 17523
## - hispanic  1    103080  962189 17682
## - black     1    153995 1013104 17840

##
## Call:
## lm(formula = ppvtr.36 ~ black + hispanic + momed + treat + work.dur +
##     b.marr + hs, data = data_tr)
##
## Coefficients:
## (Intercept)        black     hispanic        momed        treat     work.dur
##      79.264      -18.281      -15.503        4.987        9.883        4.319
##      b.marr           hs
##       2.166        1.795
```

```
lmod_BIC_BO<-lm(ppvtr.36 ~ white + momed + work.dur + treat + b.marr,data_tr) # BIC selected model
sum_BIC_BO<-summary(lmod_BIC_BO)
sum_BIC_BO
```

```
##
## Call:
## lm(formula = ppvtr.36 ~ white + momed + work.dur + treat + b.marr,
##     data = data_tr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.57 -10.19   0.86  11.54  54.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.7766     0.8408  74.663  < 2e-16 ***
## white        17.1029     0.6483  26.381  < 2e-16 ***
## momed         4.7422     0.3537  13.409  < 2e-16 ***
## work.dur      4.5876     0.6617   6.933 4.99e-12 ***
## treat         9.2799     1.2119   7.657 2.53e-14 ***
## b.marr        2.8726     0.7147   4.019 5.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.8 on 3060 degrees of freedom
## Multiple R-squared:  0.3184, Adjusted R-squared:  0.3173
## F-statistic: 285.9 on 5 and 3060 DF,  p-value: < 2.2e-16
```

```
vif(lmod_BIC_BO)
```

```
##    white    momed work.dur    treat   b.marr
## 1.141455 1.140094 1.117453 1.022422 1.209251
```

lar selection

```r
x = as.matrix(data_tr[,c(1:20)])
y = as.matrix(data_tr[,21])
library(lars)
```

```
## Loaded lars 1.3
```

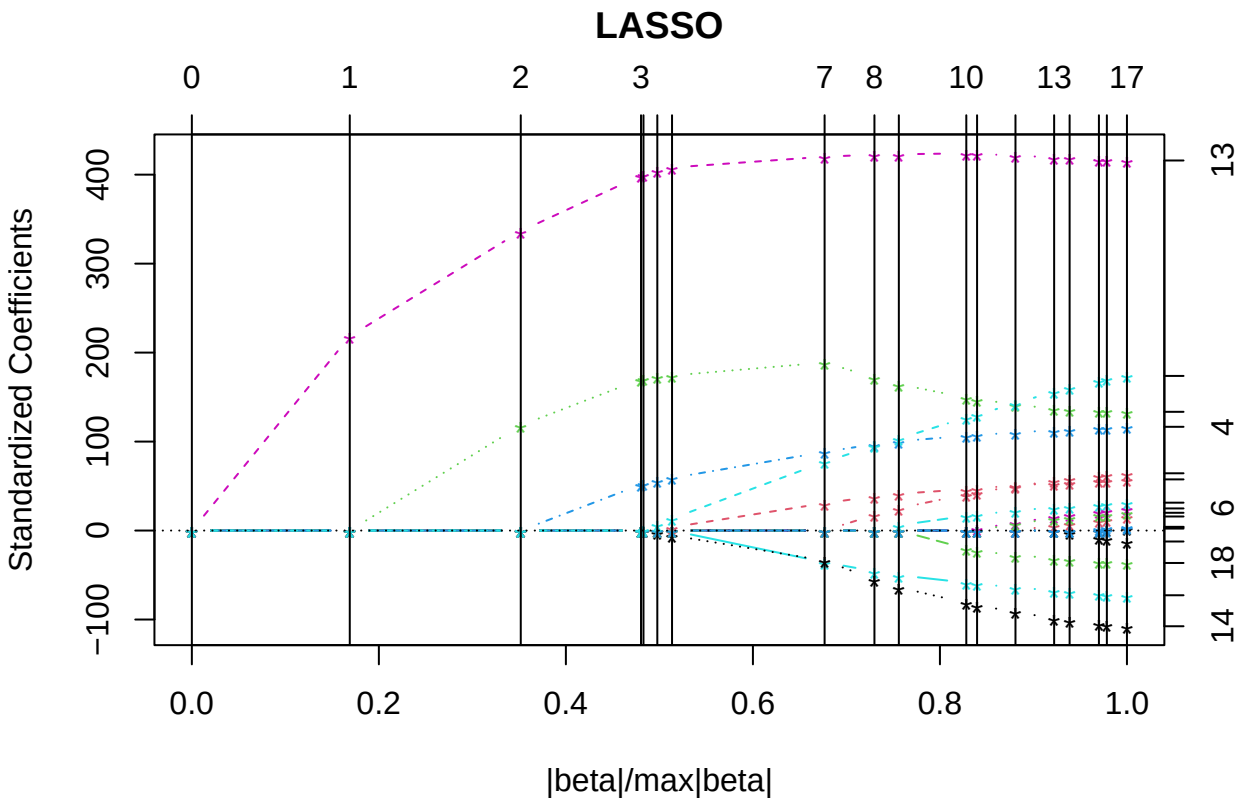
```
##
## Attaching package: 'lars'
```

```
## The following object is masked from 'package:psych':
##
##     error.bars
```

```r
lar1 <-lars(x,y,type = "lasso") # Use Lasso
lar1
```

```
##
## Call:
## lars(x = x, y = y, type = "lasso")
## R-squared: 0.325
## Sequence of LASSO moves:
##      white momed work.dur treat lths b.marr black college prenatal dayskidh cig
## Var     13     3        4    20   14      2    11      17        5       18   6
## Step     1     2        3     4    5      6     7       8        9       10  11
##      bwg bw momage income sex preterm
## Var    9  8      1     19   7      10
## Step  12 13     14     15  16      17
```

```r
plot(lar1)
```




```
sum1<-summary(lar1)
sum1
```

```
## LARS/LASSO
## Call: lars(x = x, y = y, type = "lasso")
##    Df     Rss       Cp
## 0   1 1267065 1454.664
## 1   2 1079897  789.178
## 2   3  960150  364.130
## 3   4  910435  188.837
## 4   5  909680  188.141
## 5   6  905545  175.396
## 6   7  901359  162.469
## 7   8  869884   52.219
## 8   9  864149   33.766
## 9  10  862046   28.267
## 10 11  857878   15.403
## 11 12  857408   15.727
## 12 13  856070   12.955
## 13 14  855218   11.918
## 14 15  855002   13.148
## 15 16  854750   14.247
## 16 17  854715   16.124
## 17 18  854680   18.000
```

```
lar1$Cp[which.min(lar1$Cp)]
```

```
##       13
## 11.91816
```

```
lar1$beta
```

```
##         momage    b.marr    momed work.dur  prenatal        cig        sex
## 0   0.00000000 0.0000000 0.000000 0.000000 0.0000000 0.00000000 0.00000000
## 1   0.00000000 0.0000000 0.000000 0.000000 0.0000000 0.00000000 0.00000000
## 2   0.00000000 0.0000000 2.330145 0.000000 0.0000000 0.00000000 0.00000000
## 3   0.00000000 0.0000000 3.355052 1.927433 0.0000000 0.00000000 0.00000000
## 4   0.00000000 0.0000000 3.369572 1.953281 0.0000000 0.00000000 0.00000000
## 5   0.00000000 0.0000000 3.407730 2.085894 0.0000000 0.00000000 0.00000000
## 6   0.00000000 0.1412713 3.435011 2.202457 0.0000000 0.00000000 0.00000000
## 7   0.00000000 1.1913292 3.723199 3.292657 0.0000000 0.00000000 0.00000000
## 8   0.00000000 1.4848817 3.392066 3.617689 0.0000000 0.00000000 0.00000000
## 9   0.00000000 1.5965174 3.243060 3.741086 0.8252248 0.00000000 0.00000000
## 10  0.00000000 1.7989123 2.944019 3.968112 2.3042719 0.00000000 0.00000000
## 11  0.00000000 1.8380233 2.907028 4.000532 2.5048531 0.07966628 0.00000000
## 12  0.00000000 1.9587651 2.796555 4.104681 3.1112634 0.33165953 0.00000000
## 13  0.00000000 2.0597823 2.691315 4.204172 3.6614795 0.58843041 0.00000000
## 14 -0.01513719 2.1031648 2.678511 4.239024 3.8361073 0.67207920 0.00000000
## 15 -0.04331299 2.1730279 2.654338 4.298307 4.1499624 0.82438111 0.00000000
## 16 -0.05035386 2.1906314 2.647727 4.313447 4.2279753 0.86351888 0.02524674
## 17 -0.06516600 2.2246121 2.630794 4.343841 4.3833320 0.95190875 0.08089695
##              bw       bwg   preterm     black hispanic     white       lths hs
## 0  0.0000000000 0.0000000 0.0000000  0.000000        0  0.000000  0.0000000  0
## 1  0.0000000000 0.0000000 0.0000000  0.000000        0  7.885721  0.0000000  0
## 2  0.0000000000 0.0000000 0.0000000  0.000000        0 12.154778  0.0000000  0
## 3  0.0000000000 0.0000000 0.0000000  0.000000        0 14.413828  0.0000000  0
## 4  0.0000000000 0.0000000 0.0000000  0.000000        0 14.445498  0.0000000  0
## 5  0.0000000000 0.0000000 0.0000000  0.000000        0 14.615327 -0.1102847  0
## 6  0.0000000000 0.0000000 0.0000000  0.000000        0 14.736760 -0.2102252  0
## 7  0.0000000000 0.0000000 0.0000000 -1.408947        0 15.177938 -1.3192227  0
## 8  0.0000000000 0.0000000 0.0000000 -1.809186        0 15.260954 -2.1413181  0
## 9  0.0000000000 0.0000000 0.0000000 -1.984651        0 15.275707 -2.4901944  0
## 10 0.0000000000 0.0000000 0.0000000 -2.314358        0 15.313026 -3.1760978  0
## 11 0.0000000000 0.0000000 0.0000000 -2.366174        0 15.293851 -3.2745327  0
## 12 0.0000000000 0.6231614 0.0000000 -2.531743        0 15.222337 -3.5747158  0
## 13 0.0001630694 1.0681565 0.0000000 -2.664834        0 15.148107 -3.8500675  0
## 14 0.0002176517 1.1884988 0.0000000 -2.714781        0 15.120376 -3.9341201  0
## 15 0.0003169236 1.4109618 0.0000000 -2.806611        0 15.068265 -4.0836454  0
## 16 0.0003450279 1.4661614 0.0000000 -2.831610        0 15.053655 -4.1216912  0
## 17 0.0004430505 1.6039632 0.0281891 -2.876192        0 15.023137 -4.1975216  0
##    ltcoll  college    dayskidh       income       treat
## 0       0 0.000000  0.00000000 0.000000e+00  0.00000000
## 1       0 0.000000  0.00000000 0.000000e+00  0.00000000
## 2       0 0.000000  0.00000000 0.000000e+00  0.00000000
## 3       0 0.000000  0.00000000 0.000000e+00  0.00000000
## 4       0 0.000000  0.00000000 0.000000e+00  0.07740594
## 5       0 0.000000  0.00000000 0.000000e+00  0.51185115
## 6       0 0.000000  0.00000000 0.000000e+00  0.95933789
```

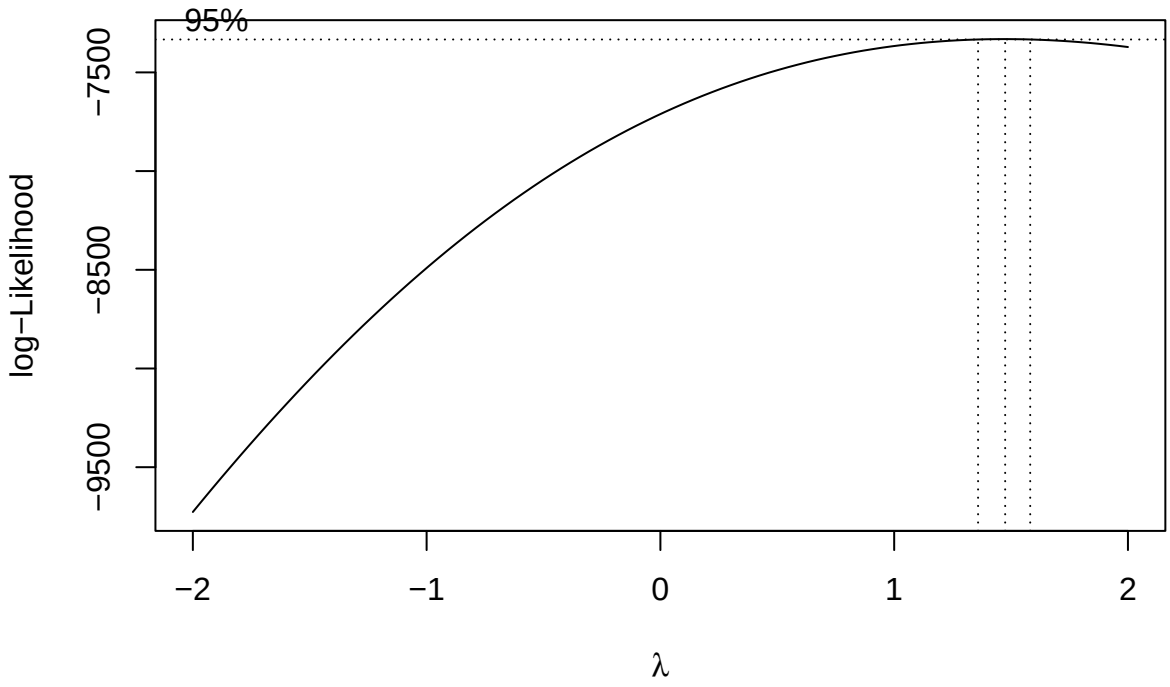
```
## 7       0 0.000000  0.00000000 0.000000e+00  5.49973288
## 8       0 1.140344  0.00000000 0.000000e+00  6.80708316
## 9       0 1.629107  0.00000000 0.000000e+00  7.36999793
## 10      0 2.641434 -0.05794002 0.000000e+00  9.06026768
## 11      0 2.787266 -0.06550822 0.000000e+00  9.28630671
## 12      0 3.227462 -0.07892437 0.000000e+00 10.19250505
## 13      0 3.635611 -0.08892395 0.000000e+00 11.14575006
## 14      0 3.769137 -0.09227939 0.000000e+00 11.45701791
## 15      0 3.995391 -0.09811036 6.265649e-07 12.02150050
## 16      0 4.052615 -0.09942330 7.857275e-07 12.16501134
## 17      0 4.178812 -0.10385753 1.121341e-06 12.39886980
## attr(,"scaled:scale")
##  [1] 1.885784e+02 2.584948e+01 5.072233e+01 2.684007e+01 7.149690e+00
##  [6] 2.610236e+01 2.768559e+01 3.570624e+04 1.245904e+01 1.317974e+02
## [11] 2.528838e+01 2.239921e+01 2.768536e+01 2.562034e+01 2.724971e+01
## [16] 2.187023e+01 1.539873e+01 3.506410e+02 3.730620e+06 1.401710e+01
```

```
lmod_la<-glm(ppvtr.36~.-white - ltcoll -college,family='gaussian',data_tr) # Use glm
summary(lmod_la)
```

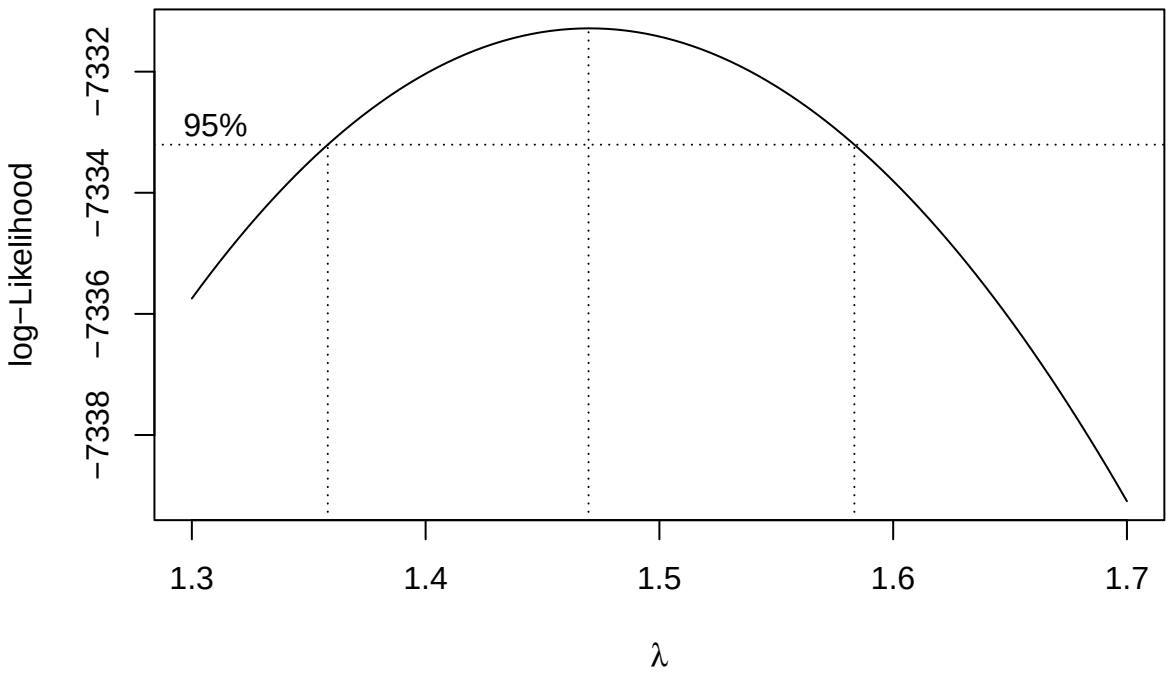
```
##
## Call:
## glm(formula = ppvtr.36 ~ . - white - ltcoll - college, family = "gaussian",
##     data = data_tr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -67.115   -9.930    0.899   11.322   53.710
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.718e+01  6.026e+00  11.148  < 2e-16 ***
## momage      -6.517e-02  9.867e-02  -0.660  0.50900
## b.marr       2.225e+00  7.586e-01   2.932  0.00339 **
## momed        6.810e+00  1.288e+00   5.286 1.34e-07 ***
## work.dur     4.344e+00  6.671e-01   6.512 8.65e-11 ***
## prenatal     4.383e+00  2.368e+00   1.851  0.06424 .
## cig          9.519e-01  6.911e-01   1.377  0.16847
## sex          8.090e-02  6.123e-01   0.132  0.89491
## bw           4.431e-04  6.671e-04   0.664  0.50665
## bwg          1.604e+00  1.960e+00   0.818  0.41326
## preterm      2.819e-02  1.761e-01   0.160  0.87283
## black       -1.790e+01  8.059e-01 -22.210  < 2e-16 ***
## hispanic    -1.502e+01  8.403e-01 -17.877  < 2e-16 ***
## lths         4.160e+00  3.028e+00   1.374  0.16961
## hs           4.179e+00  1.818e+00   2.299  0.02156 *
## dayskidh    -1.039e-01  6.340e-02  -1.638  0.10148
## income       1.121e-06  4.606e-06   0.243  0.80766
## treat        1.240e+01  1.577e+00   7.862 5.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 280.4069)
##
```

```
##     Null deviance: 1267065  on 3065  degrees of freedom
## Residual deviance:  854680  on 3048  degrees of freedom
## AIC: 26002
##
## Number of Fisher Scoring iterations: 2
```

```
boxcox(lmod_AIC_B, plotit=T)
```




```
b<-boxcox(lmod_AIC_B, plotit=T, lambda=seq(1.3, 1.7,by=0.05))
```




```
b
```

```
## $x
##   [1] 1.300000 1.304040 1.308081 1.312121 1.316162 1.320202 1.324242 1.328283
##   [9] 1.332323 1.336364 1.340404 1.344444 1.348485 1.352525 1.356566 1.360606
##  [17] 1.364646 1.368687 1.372727 1.376768 1.380808 1.384848 1.388889 1.392929
##  [25] 1.396970 1.401010 1.405051 1.409091 1.413131 1.417172 1.421212 1.425253
##  [33] 1.429293 1.433333 1.437374 1.441414 1.445455 1.449495 1.453535 1.457576
##  [41] 1.461616 1.465657 1.469697 1.473737 1.477778 1.481818 1.485859 1.489899
##  [49] 1.493939 1.497980 1.502020 1.506061 1.510101 1.514141 1.518182 1.522222
##  [57] 1.526263 1.530303 1.534343 1.538384 1.542424 1.546465 1.550505 1.554545
##  [65] 1.558586 1.562626 1.566667 1.570707 1.574747 1.578788 1.582828 1.586869
##  [73] 1.590909 1.594949 1.598990 1.603030 1.607071 1.611111 1.615152 1.619192
##  [81] 1.623232 1.627273 1.631313 1.635354 1.639394 1.643434 1.647475 1.651515
##  [89] 1.655556 1.659596 1.663636 1.667677 1.671717 1.675758 1.679798 1.683838
##  [97] 1.687879 1.691919 1.695960 1.700000
##
## $y
##   [1] -7335.746 -7335.535 -7335.329 -7335.128 -7334.932 -7334.742 -7334.556
##   [8] -7334.376 -7334.201 -7334.032 -7333.867 -7333.708 -7333.554 -7333.404
##  [15] -7333.260 -7333.121 -7332.988 -7332.859 -7332.736 -7332.617 -7332.504
##  [22] -7332.395 -7332.292 -7332.194 -7332.101 -7332.013 -7331.930 -7331.852
##  [29] -7331.779 -7331.712 -7331.649 -7331.591 -7331.538 -7331.491 -7331.448
##  [36] -7331.410 -7331.377 -7331.350 -7331.327 -7331.309 -7331.296 -7331.288
##  [43] -7331.285 -7331.287 -7331.294 -7331.306 -7331.323 -7331.345 -7331.371
##  [50] -7331.403 -7331.439 -7331.480 -7331.527 -7331.578 -7331.634 -7331.694
```

```
##  [57] -7331.760 -7331.831 -7331.906 -7331.986 -7332.071 -7332.161 -7332.256
##  [64] -7332.355 -7332.460 -7332.569 -7332.683 -7332.802 -7332.925 -7333.053
##  [71] -7333.186 -7333.324 -7333.467 -7333.614 -7333.766 -7333.923 -7334.085
##  [78] -7334.251 -7334.422 -7334.598 -7334.778 -7334.963 -7335.153 -7335.348
##  [85] -7335.547 -7335.751 -7335.959 -7336.173 -7336.391 -7336.613 -7336.840
##  [92] -7337.072 -7337.309 -7337.550 -7337.796 -7338.046 -7338.301 -7338.561
##  [99] -7338.825 -7339.094
```

```
I=which(b$y==max(b$y))
b$x[I]
```

```
## [1] 1.469697
```

Prediction

```
result<-predict(lmod_AIC_B,newdata = data_te) # Prediction
error <- sum((data_te$ppvtr.36 - result)^2)
error
```

```
## [1] 345073.2
```