STA 304 Final Project

Mengyu Lei

Loading Data and Required Library

The data used is collected by the US National Center for Health Statistics (NCHS).

```r
rm(list = ls())
library(NHANES)
library(tidyverse)
library(sampling)
library(ggplot2)
library(gridExtra)
library(arsenal)
data("NHANESraw")
```

Data Cleaning

```r
### Only focus on the following variables
NHANES                    <- NHANESraw %>% filter(SurveyYr=="2011_12" & Age > 17)
NHANES                    <- na.omit(NHANES[,c(1,3,4,8:11,13,24,25,61,77)])
```

Demographic table for the data

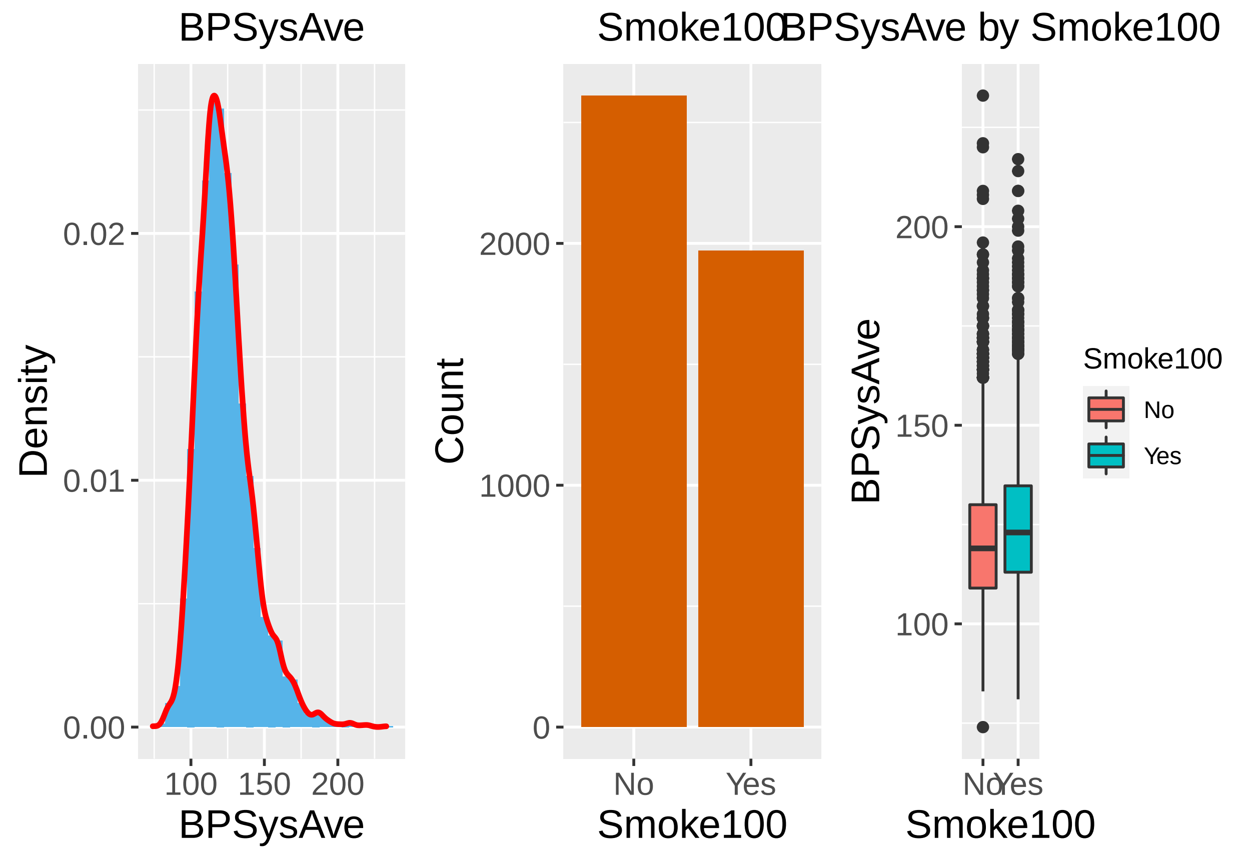
```r
demographic               <- as.data.frame(summary(tableby(Smoke100 ~ ., data = NHANES[,-1])))
write.table(demographic,file = "Results/Full.Demo.csv",row.names = FALSE)
```

Check the association between Smoke100 and BPSysAve

```r
P1                      <- ggplot(NHANES, aes(x = BPSysAve)) +
                           geom_histogram(aes(y = stat(density)),binwidth = 5,fill = "#56B4E9") +
                           geom_density(col = "red",size = 1) +
                           theme(axis.title = element_text(size = 15),
                                 axis.text = element_text(size = 12),
                                 plot.title = element_text(size = 15,
                                                          hjust = 0.5)) +
                           labs(title = "BPSysAve", y = "Density")

Count                   <- NHANES %>% group_by(Smoke100) %>% summarize(Count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
P2                      <- ggplot(Count, aes(x = Smoke100,y = Count)) +
                           geom_bar(stat = "Identity",fill = "#D55E00") +
                           theme(axis.title = element_text(size = 15),
                                 axis.text = element_text(size = 12),
                                 plot.title = element_text(size = 15, hjust = 0.5)) +
                           labs(title = "Smoke100")

P3                       <- ggplot(NHANES,aes(x = Smoke100, y = BPSysAve,
                                                  fill = Smoke100)) + geom_boxplot() +
                          theme(axis.title = element_text(size = 15),
                                axis.text = element_text(size = 12),
                                plot.title = element_text(size = 15, hjust = 0.5)) +
                          labs(title = "BPSysAve by Smoke100", y = "BPSysAve")

grid.arrange(P1,P2,P3,nrow = 1)
```



```
summary(NHANES$BPSysAve)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    74.0   111.0   120.0   123.2   132.0   233.0
```

```
table(NHANES$Smoke100)
```

```
##
##   No  Yes
## 2611 1970
```

```
NHANES %>% group_by(Smoke100) %>% summarize(Mean = mean(BPSysAve),Median = median(BPSysAve))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   Smoke100  Mean Median
##   <fct>    <dbl>  <dbl>
## 1 No        122.    119
## 2 Yes       125.    123
```

```
### Statistics Testing
var.test(BPSysAve ~ Smoke100, data = NHANES, alternative = "two.sided")
```

```
## 
##  F test to compare two variances
## 
## data:  BPSysAve by Smoke100
## F = 0.9504, num df = 2610, denom df = 1969, p-value = 0.2269
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8746973 1.0321472
## sample estimates:
## ratio of variances 
##          0.9503974
```

```
t.test(BPSysAve ~ Smoke100, data = NHANES,var.equal = TRUE)
```

```
## 
##  Two Sample t-test
## 
## data:  BPSysAve by Smoke100
## t = -7.1338, df = 4579, p-value = 1.13e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.967926 -2.826019
## sample estimates:
##  mean in group No mean in group Yes 
##          121.5289          125.4259
```

```
summary(lm(BPSysAve ~ Smoke100, data = NHANES))
```

```
## 
## Call:
## lm(formula = BPSysAve ~ Smoke100, data = NHANES)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -47.529 -12.529  -2.529   8.574 111.471 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 121.5289     0.3582 339.249  < 2e-16 ***
## Smoke100Yes   3.8970     0.5463   7.134 1.13e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 18.3 on 4579 degrees of freedom
## Multiple R-squared:  0.01099,    Adjusted R-squared:  0.01078 
## F-statistic: 50.89 on 1 and 4579 DF,  p-value: 1.13e-12
```

Predicting BPSysAve by fitting a regression model with the exposure of Smoke100

```
Lower.Model              <- lm(BPSysAve ~ Smoke100, data = NHANES)
Full.Model               <- lm(formula = as.formula(paste("BPSysAve ~",
                                                         paste(colnames(NHANES)[-c(1,9,12)],
                                                               collapse = "+"))),
                              data = NHANES)

Final.Model              <- step(Full.Model,scope = list(upper=Full.Model,lower=Lower.Model),
                                direction = "both")
```

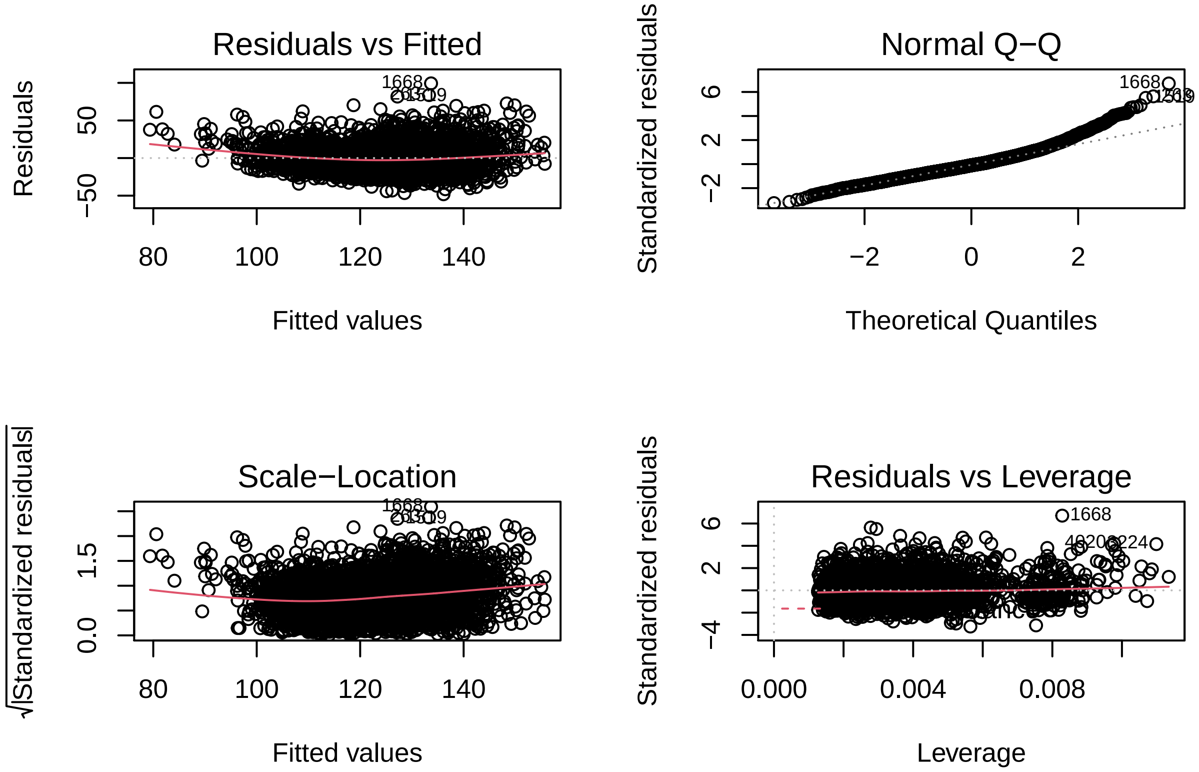
```
## Start:  AIC=24753.85
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HHIncome +
##     HHIncomeMid + HomeRooms + BPDiaAve + Smoke100
##
##
## Step:  AIC=24753.85
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HHIncome +
##     HomeRooms + BPDiaAve + Smoke100
##
##                  Df Sum of Sq     RSS   AIC
## - HHIncome      11      4214 1010637 24751
## - HomeRooms      1       385 1006808 24754
## <none>                       1006423 24754
## - Gender         1      6797 1013220 24783
## - Education      4      8706 1015129 24785
## - MaritalStatus  5     18123 1024546 24826
## - BPDiaAve       1    188942 1195365 25540
## - Age            1    205694 1212117 25604
##
## Step:  AIC=24750.99
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HomeRooms +
##     BPDiaAve + Smoke100
##
##                  Df Sum of Sq     RSS   AIC
## + HHIncomeMid    1      1442 1009195 24747
## - HomeRooms      1        22 1010660 24749
## <none>                       1010637 24751
## + HHIncome      11      4214 1006423 24754
## - Gender         1      6313 1016950 24778
## - Education      4     13823 1024460 24805
## - MaritalStatus  5     21005 1031642 24835
## - BPDiaAve       1    190234 1200871 25539
## - Age            1    212512 1223149 25623
##
## Step:  AIC=24746.46
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HomeRooms +
##     BPDiaAve + Smoke100 + HHIncomeMid
##
##                  Df Sum of Sq     RSS   AIC
## - HomeRooms      1       316 1009512 24746
## <none>                       1009195 24747
## - HHIncomeMid    1      1442 1010637 24751
## + HHIncome      10      2772 1006423 24754
## - Gender         1      6583 1015779 24774
## - Education      4      8569 1017764 24777
## - MaritalStatus  5     19005 1028200 24822
## - BPDiaAve       1    190853 1200049 25538
## - Age            1    210109 1219304 25611
##
## Step:  AIC=24745.89
## BPSysAve ~ Gender + Age + Education + MaritalStatus + BPDiaAve +
##     Smoke100 + HHIncomeMid
##
##                 Df Sum of Sq     RSS   AIC
## <none>                       1009512 24746
## + HomeRooms      1       316 1009195 24747
## - HHIncomeMid    1      1148 1010660 24749
## + HHIncome      10      2704 1006808 24754
## - Gender         1      6497 1016009 24773
## - Education      4      8564 1018076 24777
## - MaritalStatus  5     18829 1028340 24821
## - BPDiaAve       1    191600 1201112 25540
## - Age            1    214897 1224408 25628
```

```
summary(Final.Model)
```

```
##
## Call:
## lm(formula = BPSysAve ~ Gender + Age + Education + MaritalStatus +
##     BPDiaAve + Smoke100 + HHIncomeMid, data = NHANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.111  -9.500  -1.487   7.716  99.297
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.316e+01  1.856e+00  34.028  < 2e-16 ***
## Gendermale                 2.496e+00  4.605e-01   5.421 6.23e-08 ***
## Age                        4.797e-01  1.539e-02  31.176  < 2e-16 ***
## Education9 - 11th Grade    3.737e-02  9.571e-01   0.039   0.9689
## EducationHigh School      -2.380e-01  8.992e-01  -0.265   0.7913
## EducationSome College     -9.395e-01  8.799e-01  -1.068   0.2857
## EducationCollege Grad     -3.911e+00  9.365e-01  -4.176 3.03e-05 ***
## MaritalStatusLivePartner   2.085e+00  1.086e+00   1.920   0.0550 .
## MaritalStatusMarried       8.082e-01  7.742e-01   1.044   0.2966
## MaritalStatusNeverMarried  4.274e+00  8.943e-01   4.779 1.82e-06 ***
## MaritalStatusSeparated     1.832e+00  1.346e+00   1.361   0.1736
## MaritalStatusWidowed       7.214e+00  1.078e+00   6.695 2.42e-11 ***
## BPDiaAve                   5.019e-01  1.705e-02  29.438  < 2e-16 ***
## Smoke100Yes                3.336e-01  4.683e-01   0.712   0.4763
## HHIncomeMid               -1.796e-05  7.880e-06  -2.278   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 4566 degrees of freedom
## Multiple R-squared:  0.3493, Adjusted R-squared:  0.3473
## F-statistic:   175 on 14 and 4566 DF,  p-value: < 2.2e-16
```

Model diagnostics

```
par(mfrow = c(2,2))
plot(Final.Model)
```



Stratified random sampling and the stratified demographic table

```
ME                        <- 4
alpha                     <- 0.01
D                         <- (ME/qnorm(1 - alpha/2))^2

strata.data               <- NHANES %>% group_by(SDMVSTRA) %>%
                             summarise(N = n(),SD = (max(BPSysAve) - min(BPSysAve))/4) %>%
                             mutate(Cost = c(52,50,46,53,48,48,47,57,53,47,54,40,43,44))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
n                         <- with(strata.data,sum(N*SD/sqrt(Cost))*sum(N*SD*sqrt(Cost))/sum(N^2*D + sum(N
strata.data               <- strata.data %>% mutate(n_j = ceiling(n*(N*SD/sqrt(Cost))/(sum(N*SD/sqrt(Cost

demographic               <- as.data.frame(summary(tableby(SDMVSTRA ~ ., data = NHANES[,-1])))
write.csv(demographic,file = "Results/Strata.Demo.csv",row.names = FALSE)
```

Rerun the model with stratified sample

```
set.seed(1024)
strata.index            <- sampling::strata(NHANES,stratanames = "SDMVSTRA",
                                            size = strata.data$n_j,
                                            method = "srswor")
strata.nhanes           <- getdata(NHANES,strata.index)
strata.model            <- lm(BPSysAve ~ Gender + Age + Education + MaritalStatus +
                                BPDiaAve + Smoke100  + HHIncomeMid, data = strata.nhanes)
summary(strata.model)
```

```
## 
## Call:
## lm(formula = BPSysAve ~ Gender + Age + Education + MaritalStatus + 
##     BPDiaAve + Smoke100 + HHIncomeMid, data = strata.nhanes)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -29.416  -8.390  -1.305   6.455  45.912 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                5.513e+01  6.578e+00   8.381 2.17e-15 ***
## Gendermale                 3.821e+00  1.595e+00   2.396   0.0172 *  
## Age                        4.135e-01  5.554e-02   7.446 1.07e-12 ***
## Education9 - 11th Grade   -3.319e+00  3.509e+00  -0.946   0.3450    
## EducationHigh School      -2.102e+00  3.221e+00  -0.653   0.5144    
## EducationSome College     -2.327e+00  3.116e+00  -0.747   0.4559    
## EducationCollege Grad     -5.443e+00  3.372e+00  -1.614   0.1075    
## MaritalStatusLivePartner   2.682e+00  3.494e+00   0.767   0.4434    
## MaritalStatusMarried       3.882e-01  2.498e+00   0.155   0.8766    
## MaritalStatusNeverMarried  5.007e+00  2.804e+00   1.786   0.0752 .  
## MaritalStatusSeparated     2.881e-01  3.921e+00   0.073   0.9415    
## MaritalStatusWidowed       1.022e+01  4.334e+00   2.357   0.0191 *  
## BPDiaAve                   6.774e-01  6.290e-02  10.768  < 2e-16 ***
## Smoke100Yes               -2.003e-01  1.647e+00  -0.122   0.9033    
## HHIncomeMid               -5.680e-05  2.794e-05  -2.033   0.0430 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13 on 295 degrees of freedom
## Multiple R-squared:  0.4447, Adjusted R-squared:  0.4184 
## F-statistic: 16.88 on 14 and 295 DF,  p-value: < 2.2e-16
```

Fit a new model for the stratified data

```
Lower.strata            <- lm(BPSysAve ~ Smoke100, data = strata.nhanes)
Full.strata             <- lm(formula = as.formula(paste("BPSysAve ~",
                                                         paste(colnames(NHANES)[-c(1,9,12)],
                                                               collapse = "+"))),
                              data = strata.nhanes)

strata.new              <- step(Full.strata,scope = list(upper = Full.strata,
                                                         lower = Lower.strata),
                                direction = "both")
```

```
## Start:  AIC=1619.18
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HHIncome + 
##     HHIncomeMid + HomeRooms + BPDiaAve + Smoke100
## 
## 
## Step:  AIC=1619.18
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HHIncome + 
##     HomeRooms + BPDiaAve + Smoke100
## 
##                 Df Sum of Sq   RSS    AIC
## - HHIncome      11    1585.6 50216 1607.1
## - Education      4     560.6 49191 1614.7
## - HomeRooms      1     226.6 48857 1618.6
## <none>                       48630 1619.2
## - MaritalStatus  5    1972.8 50603 1621.5
## - Gender         1     911.6 49542 1622.9
## - Age            1    9410.5 58040 1672.0
## - BPDiaAve       1   19422.2 68052 1721.3
## 
## Step:  AIC=1607.13
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HomeRooms + 
##     BPDiaAve + Smoke100
## 
##                 Df Sum of Sq   RSS    AIC
## + HHIncomeMid    1     468.6 49747 1606.2
## - Education      4    1298.8 51514 1607.0
## <none>                       50216 1607.1
## - HomeRooms      1     372.4 50588 1607.4
## - MaritalStatus  5    2162.1 52378 1610.2
## - Gender         1     981.1 51197 1611.1
## + HHIncome      11    1585.6 48630 1619.2
## - Age            1    9484.4 59700 1658.8
## - BPDiaAve       1   20026.1 70242 1709.2
## 
## Step:  AIC=1606.22
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HomeRooms + 
##     BPDiaAve + Smoke100 + HHIncomeMid
## 
##                 Df Sum of Sq   RSS    AIC
## - Education      4     602.4 50349 1602.0
## - HomeRooms      1     142.0 49889 1605.1
## <none>                       49747 1606.2
## - HHIncomeMid    1     468.6 50216 1607.1
## - MaritalStatus  5    1879.1 51626 1607.7
## - Gender         1     978.2 50725 1610.3
## + HHIncome      10    1117.0 48630 1619.2
## - Age            1    9517.6 59265 1658.5
## - BPDiaAve       1   19723.8 69471 1707.8
## 
## Step:  AIC=1601.95
## BPSysAve ~ Gender + Age + MaritalStatus + HomeRooms + BPDiaAve + 
##     Smoke100 + HHIncomeMid
```

```
##
##                   Df Sum of Sq   RSS    AIC
## - HomeRooms       1     162.7 50512 1601.0
## <none>                        50349 1602.0
## - MaritalStatus  5    1780.3 52130 1602.7
## - Gender          1     911.8 51261 1605.5
## + Education       4     602.4 49747 1606.2
## - HHIncomeMid     1    1165.0 51514 1607.0
## + HHIncome       10    1158.8 49191 1614.7
## - Age             1   10179.0 60528 1657.0
## - BPDiaAve        1   20948.2 71298 1707.8
##
## Step:  AIC=1600.95
## BPSysAve ~ Gender + Age + MaritalStatus + BPDiaAve + Smoke100 +
##     HHIncomeMid
##
##                   Df Sum of Sq   RSS    AIC
## <none>                        50512 1601.0
## + HomeRooms       1     162.7 50349 1602.0
## - MaritalStatus  5    1901.1 52413 1602.4
## - Gender          1     906.8 51419 1604.5
## + Education       4     623.1 49889 1605.1
## - HHIncomeMid     1    1691.4 52204 1609.2
## + HHIncome       10    1072.8 49439 1614.3
## - Age             1   10017.4 60530 1655.0
## - BPDiaAve        1   20802.3 71314 1705.9
```

```r
summary(strata.new)
```

```
##
## Call:
## lm(formula = BPSysAve ~ Gender + Age + MaritalStatus + BPDiaAve +
##     Smoke100 + HHIncomeMid, data = strata.nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.930  -8.880  -0.674   5.956  43.903
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.200e+01  5.867e+00   8.862  < 2e-16 ***
## Gendermale                  3.678e+00  1.587e+00   2.317  0.02119 *
## Age                         4.222e-01  5.483e-02   7.700    2e-13 ***
## MaritalStatusLivePartner   2.989e+00  3.471e+00   0.861  0.38974
## MaritalStatusMarried        3.360e-01  2.494e+00   0.135  0.89293
## MaritalStatusNeverMarried  4.724e+00  2.790e+00   1.693  0.09148 .
## MaritalStatusSeparated      3.478e-01  3.916e+00   0.089  0.92930
## MaritalStatusWidowed        1.010e+01  4.281e+00   2.358  0.01901 *
## BPDiaAve                    6.893e-01  6.212e-02  11.097  < 2e-16 ***
## Smoke100Yes                -7.717e-02  1.635e+00  -0.047  0.96239
## HHIncomeMid                -7.825e-05  2.473e-05  -3.164  0.00172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 299 degrees of freedom
## Multiple R-squared:  0.4378, Adjusted R-squared:  0.419
## F-statistic: 23.28 on 10 and 299 DF,  p-value: < 2.2e-16
```

Reduced Model Comparison

```
summary(lm(BPSysAve ~ Smoke100 + Age + BPDiaAve + as.factor(SDMVSTRA) , data = NHANES))
```

```
##
## Call:
## lm(formula = BPSysAve ~ Smoke100 + Age + BPDiaAve + as.factor(SDMVSTRA),
##     data = NHANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.627  -9.916  -1.664   8.108 100.186
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            63.28825    1.59316  39.725  < 2e-16 ***
## Smoke100Yes             1.20271    0.45757   2.628 0.008606 **
## Age                     0.48070    0.01283  37.461  < 2e-16 ***
## BPDiaAve                0.50033    0.01718  29.131  < 2e-16 ***
## as.factor(SDMVSTRA)91  -0.74420    1.03002  -0.723 0.470016
## as.factor(SDMVSTRA)92   0.04956    1.08722   0.046 0.963642
## as.factor(SDMVSTRA)93  -0.24673    1.14333  -0.216 0.829156
## as.factor(SDMVSTRA)94   0.90313    1.14785   0.787 0.431439
## as.factor(SDMVSTRA)95   3.82617    1.08000   3.543 0.000400 ***
## as.factor(SDMVSTRA)96   0.95145    1.16196   0.819 0.412925
## as.factor(SDMVSTRA)97  -2.77172    1.21011  -2.290 0.022040 *
## as.factor(SDMVSTRA)98  -1.05352    1.13229  -0.930 0.352196
## as.factor(SDMVSTRA)99  -0.59938    1.17661  -0.509 0.610488
## as.factor(SDMVSTRA)100  4.24411    1.14101   3.720 0.000202 ***
## as.factor(SDMVSTRA)101  3.83433    1.07393   3.570 0.000360 ***
## as.factor(SDMVSTRA)102 -0.21127    1.19400  -0.177 0.859559
## as.factor(SDMVSTRA)103 -2.28981    1.58660  -1.443 0.149028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.06 on 4564 degrees of freedom
## Multiple R-squared:  0.3325, Adjusted R-squared:  0.3301
## F-statistic: 142.1 on 16 and 4564 DF,  p-value: < 2.2e-16
```

```
summary(lm(BPSysAve ~ Smoke100 + Age + BPDiaAve , data = strata.nhanes))
```

```
##
## Call:
## lm(formula = BPSysAve ~ Smoke100 + Age + BPDiaAve, data = strata.nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.842  -9.083  -2.001   7.175  44.539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.30413    5.17777  10.295  < 2e-16 ***
## Smoke100Yes  2.01255    1.59727   1.260    0.209    
## Age          0.38456    0.04775   8.053 1.81e-14 ***
## BPDiaAve     0.68286    0.06338  10.774  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.49 on 306 degrees of freedom
## Multiple R-squared:  0.3799, Adjusted R-squared:  0.3738 
## F-statistic:  62.5 on 3 and 306 DF,  p-value: < 2.2e-16
```