STA 304 Final Project

Please add it

Loading Data and Required Library

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

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

Data Cleaning

```
### 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

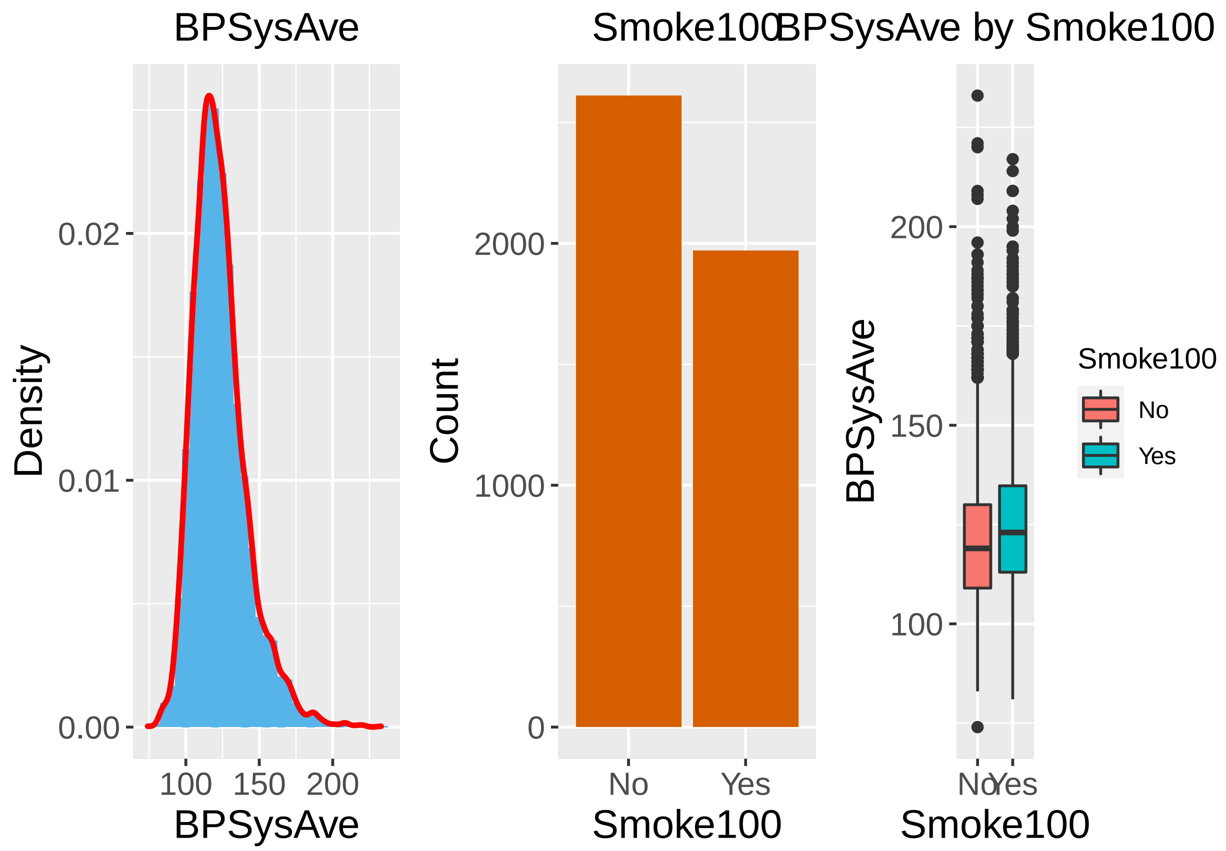
```
demographic               <- as.data.frame(summary(tableby(Smoke100 ~ ., data = NHANES[,-1])))
write.table(demographic,file = "Results/Full.Demo.csv",row.names = FALSE)
```

Check the association between Smoke100 and BPSysAve

```
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)
```

```
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)],
                                                               collapse = "+"))),
                              data = NHANES)

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

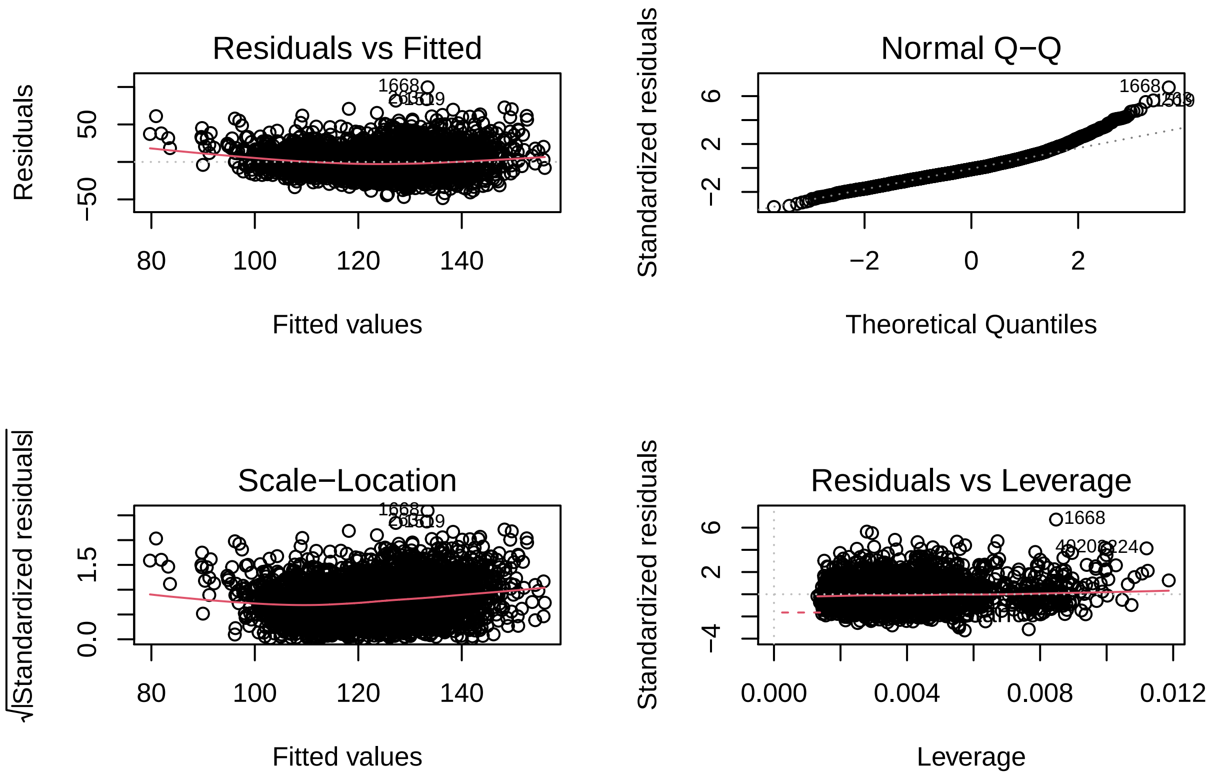
```
## Start:  AIC=24753.72
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HHIncome + 
##     HHIncomeMid + HomeRooms + BPDiaAve + Smoke100 + SDMVSTRA
## 
## 
## Step:  AIC=24753.72
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HHIncome + 
##     HomeRooms + BPDiaAve + Smoke100 + SDMVSTRA
## 
##                 Df Sum of Sq     RSS   AIC
## - HHIncome      11      4136 1010090 24751
## - HomeRooms      1       287 1006241 24753
## <none>                       1005954 24754
## - SDMVSTRA       1       469 1006423 24754
## - Gender         1      6784 1012738 24783
## - Education      4      8599 1014553 24785
## - MaritalStatus  5     18004 1023958 24825
## - BPDiaAve       1    188991 1194945 25540
## - Age            1    206163 1212117 25606
## 
## Step:  AIC=24750.52
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HomeRooms + 
##     BPDiaAve + Smoke100 + SDMVSTRA
## 
##                 Df Sum of Sq     RSS   AIC
## + HHIncomeMid    1      1284 1008806 24747
## - HomeRooms      1         7 1010097 24749
## <none>                       1010090 24751
## - SDMVSTRA       1       547 1010637 24751
## + HHIncome      11      4136 1005954 24754
## - Gender         1      6311 1016401 24777
## - Education      4     13347 1023438 24803
## - MaritalStatus  5     20743 1030833 24834
## - BPDiaAve       1    190359 1200449 25540
## - Age            1    213055 1223145 25625
## 
## Step:  AIC=24746.69
## BPSysAve ~ Gender + Age + Education + MaritalStatus + HomeRooms + 
##     BPDiaAve + Smoke100 + SDMVSTRA + HHIncomeMid
## 
##                 Df Sum of Sq     RSS   AIC
## - HomeRooms      1       236 1009042 24746
## - SDMVSTRA       1       389 1009195 24747
## <none>                       1008806 24747
## - HHIncomeMid    1      1284 1010090 24751
## + HHIncome      10      2852 1005954 24754
## - Gender         1      6567 1015373 24774
## - Education      4      8460 1017267 24777
## - MaritalStatus  5     18897 1027703 24822
## - BPDiaAve       1    190926 1199732 25539
## - Age            1    210481 1219288 25613
## 
## Step:  AIC=24745.76
## BPSysAve ~ Gender + Age + Education + MaritalStatus + BPDiaAve + 
##     Smoke100 + SDMVSTRA + HHIncomeMid
## 
##                 Df Sum of Sq     RSS   AIC
## <none>                       1009042 24746
## - SDMVSTRA       1       470 1009512 24746
## + HomeRooms      1       236 1008806 24747
## - HHIncomeMid    1      1055 1010097 24749
## + HHIncome      10      2801 1006241 24753
## - Gender         1      6493 1015535 24773
## - Education      4      8445 1017487 24776
## - MaritalStatus  5     18746 1027788 24820
## - BPDiaAve       1    191608 1200650 25540
## - Age            1    215351 1224393 25630
```

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

```
## 
## Call:
## lm(formula = BPSysAve ~ Gender + Age + Education + MaritalStatus + 
##     BPDiaAve + Smoke100 + SDMVSTRA + HHIncomeMid, data = NHANES)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -48.315  -9.486  -1.549   7.726  99.601 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                5.530e+01  5.700e+00   9.701  < 2e-16 ***
## Gendermale                 2.495e+00  4.604e-01   5.420 6.27e-08 ***
## Age                        4.805e-01  1.540e-02  31.213  < 2e-16 ***
## Education9 - 11th Grade   -3.334e-02  9.583e-01  -0.035   0.9722    
## EducationHigh School      -2.793e-01  8.995e-01  -0.310   0.7562    
## EducationSome College     -9.896e-01  8.805e-01  -1.124   0.2611    
## EducationCollege Grad     -3.931e+00  9.365e-01  -4.198 2.74e-05 ***
## MaritalStatusLivePartner   2.111e+00  1.086e+00   1.944   0.0520 .  
## MaritalStatusMarried       8.132e-01  7.742e-01   1.050   0.2936    
## MaritalStatusNeverMarried  4.272e+00  8.942e-01   4.778 1.83e-06 ***
## MaritalStatusSeparated     1.829e+00  1.346e+00   1.358   0.1745    
## MaritalStatusWidowed       7.201e+00  1.077e+00   6.683 2.62e-11 ***
## BPDiaAve                   5.019e-01  1.705e-02  29.442  < 2e-16 ***
## Smoke100Yes                3.210e-01  4.684e-01   0.685   0.4932    
## SDMVSTRA                   8.161e-02  5.598e-02   1.458   0.1450    
## HHIncomeMid               -1.725e-05  7.894e-06  -2.185   0.0289 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 4565 degrees of freedom
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3474
## F-statistic: 163.6 on 15 and 4565 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 + SDMVSTRA + HHIncomeMid,
                              data = strata.nhanes)
```

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

```r
summary(lm(BPSysAve ~ Smoke100 + Age + BPDiaAve , data = NHANES))
```

```
##
## Call:
## lm(formula = BPSysAve ~ Smoke100 + Age + BPDiaAve, data = NHANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.792  -9.975  -1.758   8.129  99.413
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.59925    1.41898  45.525  < 2e-16 ***
## Smoke100Yes  1.34110    0.45730   2.933  0.00338 **
## Age          0.48438    0.01285  37.700  < 2e-16 ***
## BPDiaAve     0.48769    0.01714  28.452  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.18 on 4577 degrees of freedom
## Multiple R-squared:   0.32,  Adjusted R-squared:  0.3196
## F-statistic:   718 on 3 and 4577 DF,  p-value: < 2.2e-16
```

```r
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
```