THE EFFECTS OF AGE, INCOME, SMOKING STATUS AND EDUCATION STATUS ON SYSTOLIC BLOOD PRESSURE

Rui Zhu

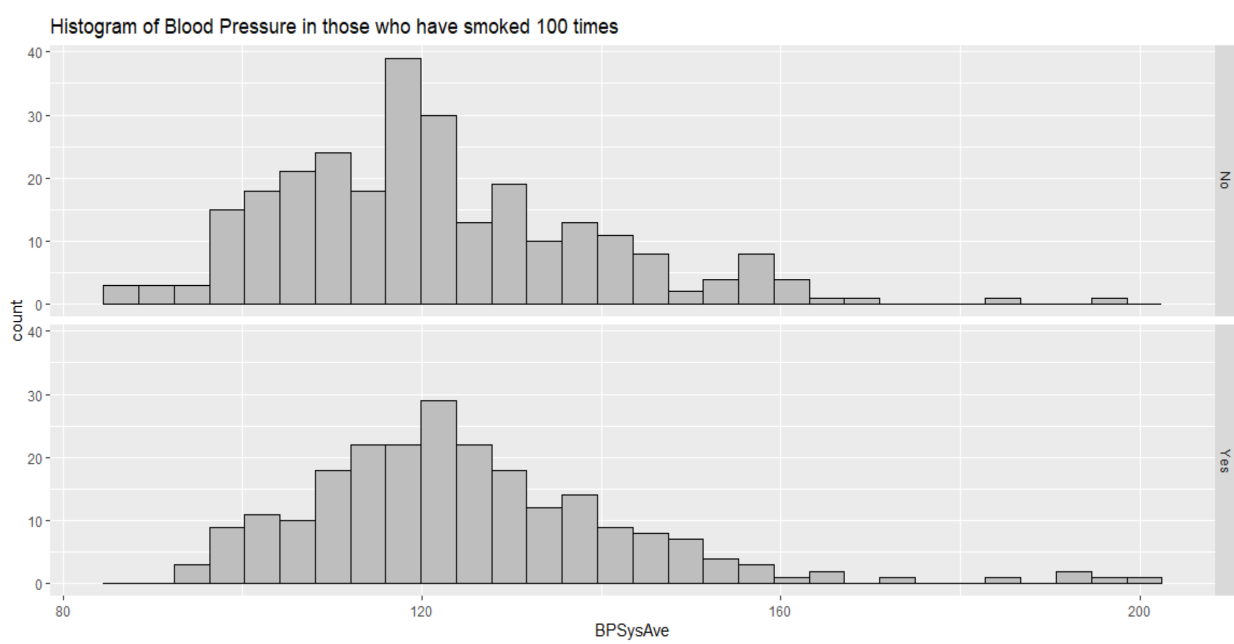
STAC53

Background Information

Over the years, cohorts of studies are investigating the causes of hypertension and high blood pressure. The intention is to understand how hypertension contributes to cardiovascular diseases (CVD) such as stroke or heart attack[1]. Blood pressure is the measurement of pressure against the walls of the circulating blood and plays a key role in maintaining heart health[2]. Over the years, studies report a positive correlation between hypertension and the risk of CVD development1. The number of mortality cases due to hypertension caused by CVD disease is also on the rise since 1990s[3]. More recently, an association was found between hypertension in young adults and the likelihood of developing CVD diseases later in life[4]. Luo *et al.* conducted a systematic study and found a 23.8% association between CVD and rising blood pressure in young adults, which is considered significant[4]. However, the cause of the rising of average hypertension in the current population is still unknown, which makes this area of research crucial. This study examines the relationship between education, income, smoking status, and age on the cause of hypertension. Previous studies suggest a positive correlation between these variables and hypertension; however, inconsistencies may occur in the results, so this study aims to clarify the relationship between the variables and their contribution to high blood pressure. The National Health and Nutrition Examination Survey (NHANES), are studies designed to measure the health and nutrition status of adults and children in the United States[5]. The data used in this study is obtained from NHANES and further adjusted for analysis. Overall, the goal of the study is to examine the effects education, income, smoking status, and age have on the cause of hypertension.

Methods and Materials

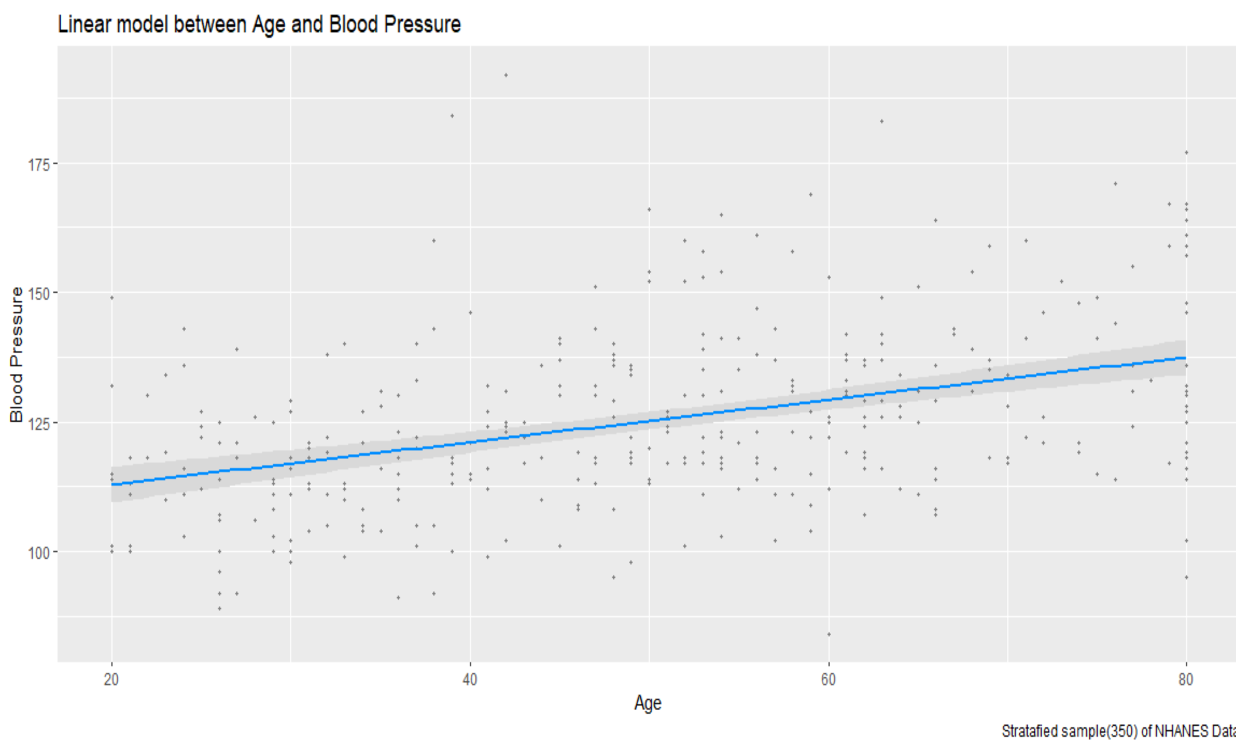
Data taken from a pool of 4182 entries of the NHANES dataset was randomly sampled through numeric randomization into a smaller pool of 500 (12%) samples and used for analyses[5]. Blood pressure (BPSysAve) was the focus of univariate analysis through boxplot and histogram (appendix figure 1 and figure 2) visualizations and later examined under bivariate analysis concerning the act of smoking (Smoke100).



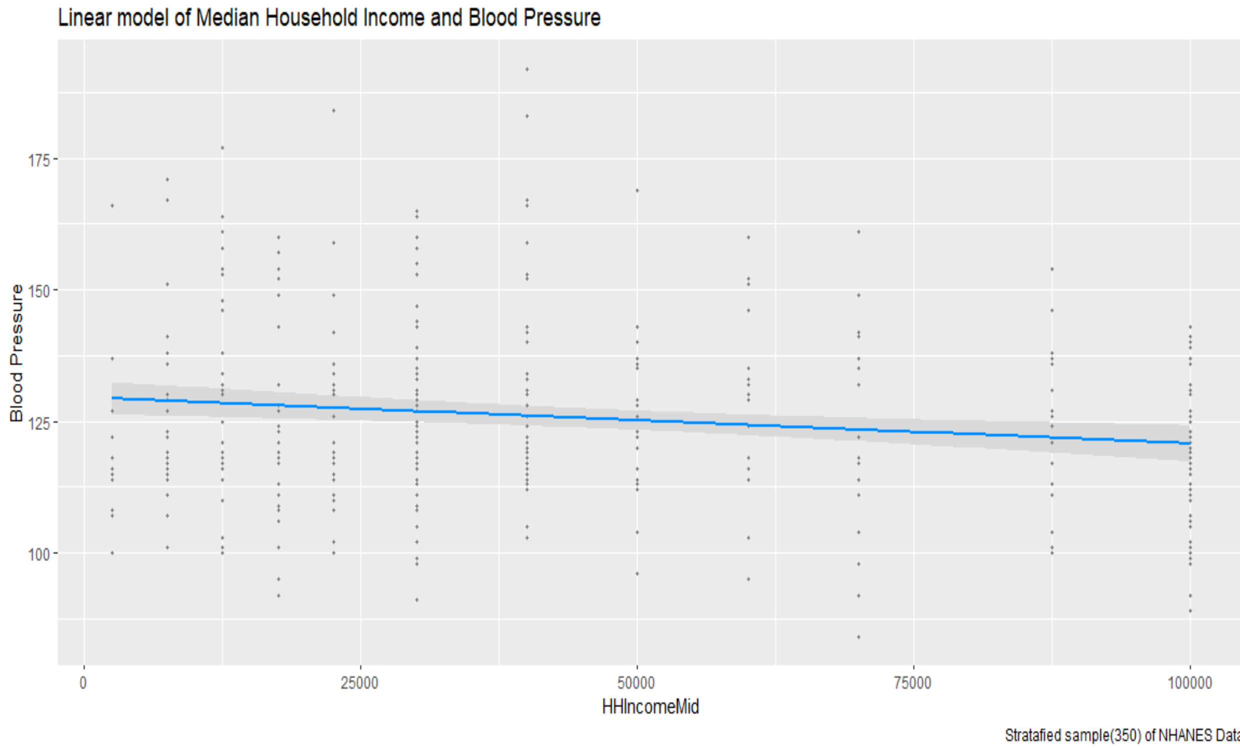
The variance test was conducted between the two variables and followed by a t-test to determine if the variables had a significant relationship. Linear regression models were then applied to examine associations between blood pressure and age, current education, median household income. The choice of variables to conduct analyses was due to an abundance of research suggesting a direct correlation between blood pressure, age, and general frailty[6]. Variables such as level of education and median household income are both factors that may change the quality of life in subjects, as a result, these variables were examined for associations with blood pressure. In a study of completed education and blood pressure, those less educated were found to have a steeper decline in routine exercise as their age increases[7]. A decline in physical activity is directly correlated to higher blood pressure[8]. Household income directly affects health and blood pressure when in the extremes[9]. The analysis was also conducted on both age and median household income with blood pressure to determine independence. Further analyses were completed with stratified NHANES data where 4862 samples were separated into 14 pre-established strata, and sample size per strata was adjusted based on cost. The stratified sample set had the same analysis model reapplied to establish confirmation on previous assumptions and better represent the population. The regression analyses as well as modeling were all completed in R studio through Anaconda.

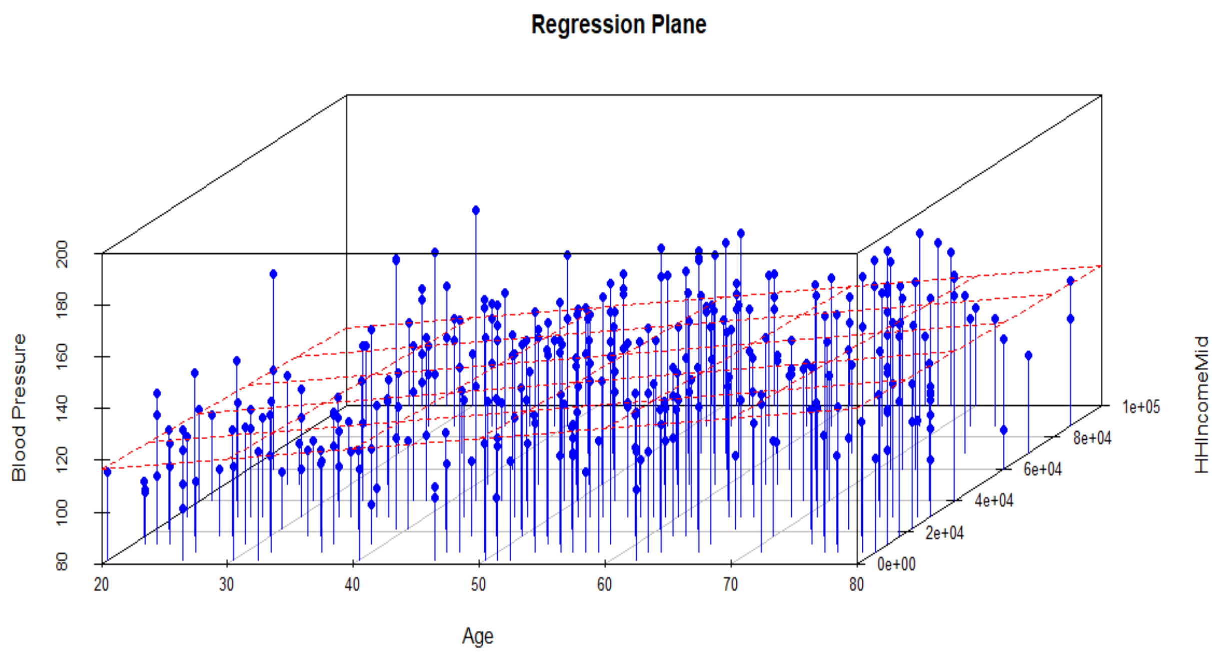
Results

Initial testing showed roughly equal variance and separate mean between the systolic blood pressure of smokers versus non-smokers (appendix figure 3). This may be due in part to the smoker group having multiple outliers. Other results, such as the association between age and blood pressure, were linear and significant (as predicted $P < .05$) as blood pressure would increase with age. The linear model also found blood pressure to decrease linearly as the level of completed education increases; however, only college graduates showed a considerable correlation ($P < .05$).

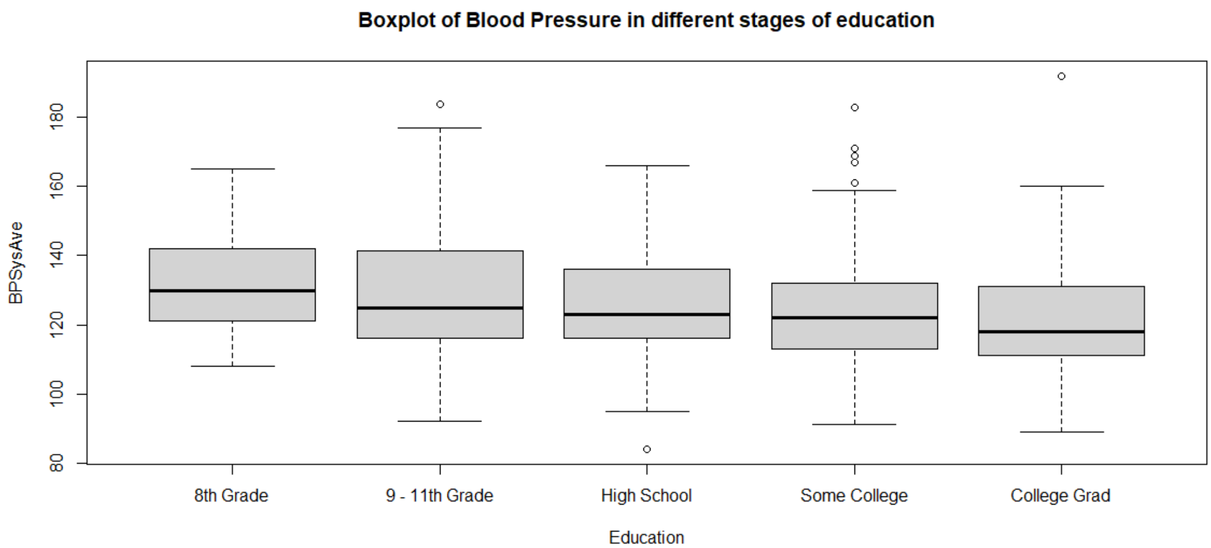


Lastly, the model was performed on household median income while adjusting for age. The analysis found that blood pressure on average was 1mmhg lower per roughly 8750$ increase in income.





The same procedure was done for our stratified sample set and the findings expressed that smokers did not have a significant correlation over non-smokers ($P > .05$). The model further confirmed an association between age and blood pressure. Also, the association between blood pressure was shown to have significance across all stages of completed education excluding pre-high school, as higher levels of education would result in lower average blood pressure and lower variance.



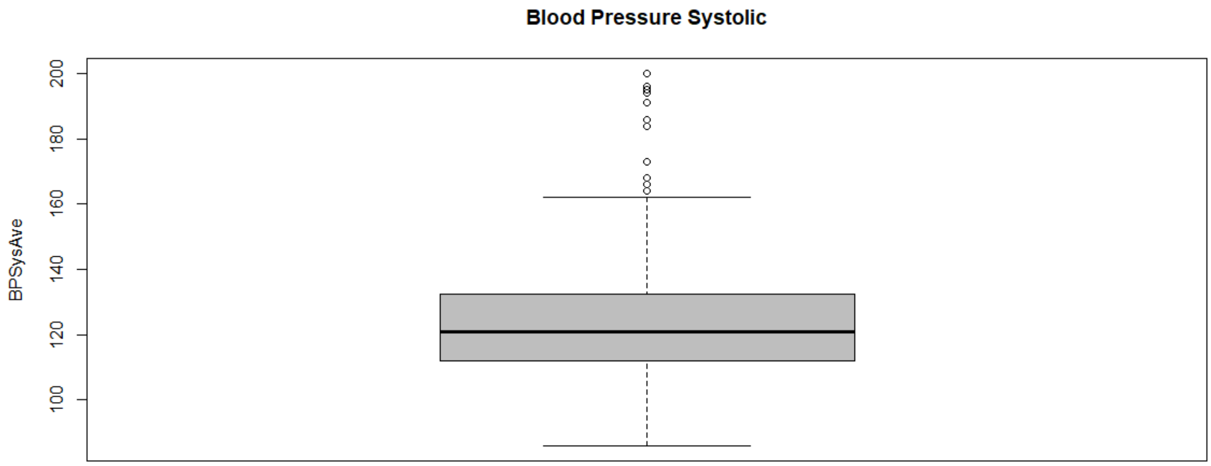
Furthermore, the model supported a higher association between median household income and blood pressure ($P < .01$).

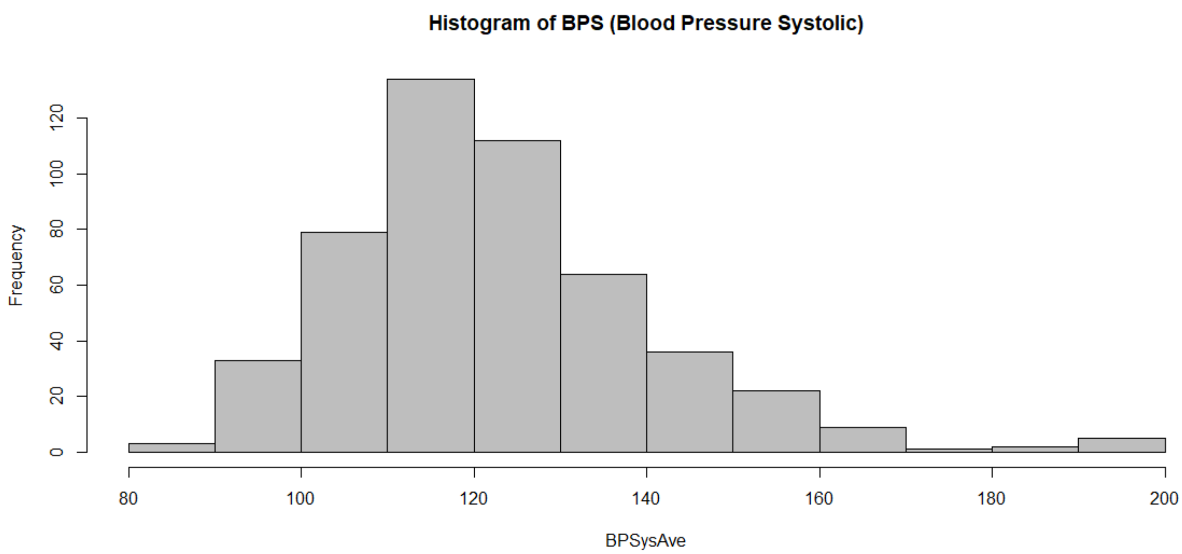
Discussion

Further confirmed by the analyses, age is the primary factor in association with blood pressure. This positive correlation between age and blood pressure has interesting implications on two of the other variables, as the variables are both statistically and intuitively associated with age. The level of education completed has a very consistent average age, this is due to public education having expectations of a timeline and countless regulations. Also, there is an expected lower limit (youngest age to finish a stage of education) for average individuals to complete their entire education (Given their education is continuous). The result of education completed sharing a negative correlation (Dia3) with blood pressure improves the significance. This is due to the expected age increasing as the completion of education increases. With each stage of education completed, there is an increase in the likelihood of actions that reduce blood pressure to preserve health and lower the risk of hypertension-related disease[7]. The smoke100 variable used in the study refers to whether or not the individual has smoked more than 100 cigarettes across their entire life span; this option only being available to subjects aged 20+[5]. The age limit correlates directly to the "Age" variable and is also related to the span of available time for a subject to fall into the “Yes” category. The regression model of age and median household income with blood pressure shows a significant correlation across simple and stratified sample sets. Median household income has an inverse relationship with blood pressure as higher income will result in lower blood pressure. This is likely due to the higher quality of life afforded by median household income. The results of the study indicate that the effect of smoking on blood pressure remains inconsistent. This result contrasts with previously established studies that show a clear correlation between smoking and increased blood pressure[10]. The consequence of this result may highlight the possibility of error in our stratified random sampling. This error may stem from the limitations of the linear model as smoking and blood pressure may be better fitted with another model. Also, the variable “Smoke100” offers poor quantitative analysis. The ambiguity of the "Smoke100" results in possible response bias due to the participants forgetting or misrepresenting their own smoking tendencies. Additionally, the number of cigarettes among the yes category may cause measurement bias as the total cigarettes smoked may differ by any quantity over any time span.

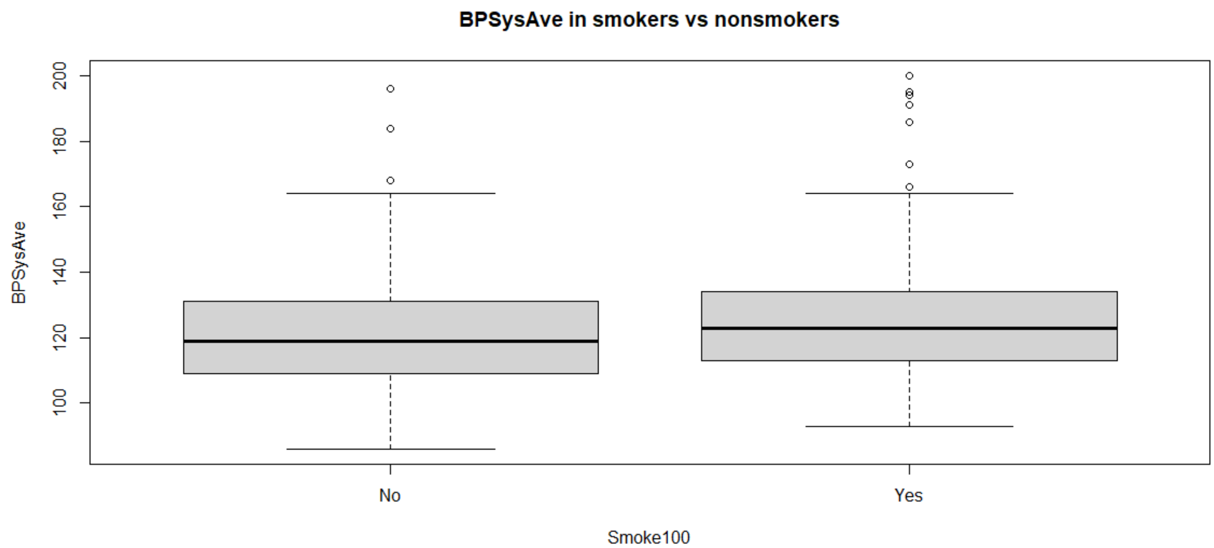
Appendix

Appendix figure 1





Appendix figure 2



Appendix figure 3

```
library(NHANES)
library(tidyverse)
library(dplyr)
library(tidyr)
library(ggplot2)
library(car)
library(visreg)
library(scatterplot3d)
library(UsingR)
small.nhanes <- na.omit(NHANESraw[NHANESraw$SurveyYr=="2011_12"& NHANESraw$Age >
17,c(1,3,4,8:11,13,24,25,61,77)])
small.nhanes <- small.nhanes %>%group_by(ID) %>% filter(row_number()==1)
set.seed(1004149361)
#setting randomizer as student number
sample_500 <- sample(1:4581, 500, replace = F)
#sampling 500 entries of the small nhanes data set without replacement
smaller.nhanes <- small.nhanes[sample_500, ]
boxplot(smaller.nhanes$BPSysAve, ylab = "BPSysAve", main = "Blood Pressure Systolic", col =
"gray")
#univariate boxplot of all the blood pressures within our 500 sample
dev.off()
hist(smaller.nhanes$BPSysAve, col = "gray", border = "black" , xlab = "BPSysAve", main =
"Histogram of BPS (Blood Pressure Systolic)")
dev.off
#plotting univariate histogram of all the blood pressures
```

```
boxplot(BPSysAve ~ Smoke100, data = smaller.nhanes, xlab = "Smoke100", ylab = "BPSysAve",
main = "BPSysAve in smokers vs nonsmokers")
#plotting bivariate boxplot of our blood pressure in smokers vs non smokers
dev.off()
Smokers <- filter(smaller.nhanes, Smoke100 == "Yes")
Non_Smokers <- filter(smaller.nhanes,Smoke100 == "No")
#separating smokers and non smokers
ggplot(smaller.nhanes, aes(x = BPSysAve)) +
  geom_histogram(fill = "gray", colour = "black") +
  facet_grid(Smoke100 ~ .) + labs(title = "Histogram of Blood Pressure in those who have
smoked 100 times")
#plot of bivariate histogram of Blood pressure in smokers vs non smokers (figure1)
dev.off()
x <- Smokers$BPSysAve
y <- Non_Smokers$BPSysAve
var.test(x,y)
#complete variance test conclusion non equal variance
t.test(x,y, var.equal = F)
#complete t-test we reject the null hypothesis.
model.smoke <- lm(BPSysAve ~ Smoke100, data = smaller.nhanes)
#running linear regression for smoke and blood pressure
Anova(model.smoke)
model.age <- lm(BPSysAve ~ Age, data = smaller.nhanes)
#running linear regression for age and blood pressure
summary(model.age)
model.edu <- lm(BPSysAve ~ Education, data = smaller.nhanes)
#running linear regression for education and blood pressure
summary(model.edu)
model.inc <- lm(BPSysAve ~ HHIncomeMid, data = smaller.nhanes)
#running linear regression for HHIncomeMid and blood pressure
summary(model.inc)
model.age_inc <- lm(BPSysAve ~ Age + HHIncomeMid, data = smaller.nhanes)
summary(model.age_inc)
pdf("scatter3d.pdf", height = 8, width =12)
s3d <- scatterplot3d(cbind(s$Age, s$HHIncomeMid, s$BPSysAve), type = "h", color = "blue",
angle = 55, pch = 16, main = "Regression Plane", xlab = "Age", ylab = "HHIncomeMid",zlab =
"Blood Pressure")
s3d$plane3d(model.age_inc2, col = "red")
#3d model of linear regression between HHIncomeMid + age vs blood pressure (Figure 4)
dev.off()
strata1 <- filter(small.nhanes, SDMVSTRA == 90)
strata2 <- filter(small.nhanes, SDMVSTRA == 91)
strata3 <- filter(small.nhanes, SDMVSTRA == 92)
strata4 <- filter(small.nhanes, SDMVSTRA == 93)
```

```
strata5 <- filter(small.nhanes, SDMVSTRA == 94)
strata6 <- filter(small.nhanes, SDMVSTRA == 95)
strata7 <- filter(small.nhanes, SDMVSTRA == 96)
strata8 <- filter(small.nhanes, SDMVSTRA == 97)
strata9 <- filter(small.nhanes, SDMVSTRA == 98)
strata10 <- filter(small.nhanes, SDMVSTRA == 99)
strata11 <- filter(small.nhanes, SDMVSTRA == 100)
strata12 <- filter(small.nhanes, SDMVSTRA == 101)
strata13 <- filter(small.nhanes, SDMVSTRA == 102)
strata14 <- filter(small.nhanes, SDMVSTRA == 103)
#assigning each strata based on the SDMVSTRA which will make up 14 differnet strata
sig1 <- (max(strata1$BPSysAve) - min(strata1$BPSysAve))/4
sig2 <- (max(strata2$BPSysAve) - min(strata2$BPSysAve))/4
sig3 <- (max(strata3$BPSysAve) - min(strata3$BPSysAve))/4
sig4 <- (max(strata4$BPSysAve) - min(strata4$BPSysAve))/4
sig5 <- (max(strata5$BPSysAve) - min(strata5$BPSysAve))/4
sig6 <- (max(strata6$BPSysAve) - min(strata6$BPSysAve))/4
sig7 <- (max(strata7$BPSysAve) - min(strata7$BPSysAve))/4
sig8 <- (max(strata8$BPSysAve) - min(strata8$BPSysAve))/4
sig9 <- (max(strata9$BPSysAve) - min(strata9$BPSysAve))/4
sig10 <- (max(strata10$BPSysAve) - min(strata10$BPSysAve))/4
sig11 <- (max(strata11$BPSysAve) - min(strata11$BPSysAve))/4
sig12 <- (max(strata12$BPSysAve) - min(strata12$BPSysAve))/4
sig13 <- (max(strata13$BPSysAve) - min(strata13$BPSysAve))/4
sig14 <- (max(strata14$BPSysAve) - min(strata14$BPSysAve))/4
#taking the range(max(strata)-min(strata)) and dividing by 4 to find the sigma value
sig <- c(sig1,sig2,sig3,sig4,sig5,sig6,sig7,sig8,sig9,sig10,sig11,sig12,sig13,sig14)
# making a row vector for all the sigma values
ME <- 4
alph <- 0.01
Z <- qnorm(1-alph/2)
D <- ME^2/Z^2
#assigning some given variables
N1 <- sum(small.nhanes$SDMVSTRA == 90)
N2 <- sum(small.nhanes$SDMVSTRA == 91)
N3 <- sum(small.nhanes$SDMVSTRA == 92)
N4 <- sum(small.nhanes$SDMVSTRA == 93)
N5 <- sum(small.nhanes$SDMVSTRA == 94)
N6 <- sum(small.nhanes$SDMVSTRA == 95)
N7 <- sum(small.nhanes$SDMVSTRA == 96)
N8 <- sum(small.nhanes$SDMVSTRA == 97)
N9 <- sum(small.nhanes$SDMVSTRA == 98)
N10 <- sum(small.nhanes$SDMVSTRA == 99)
N11 <- sum(small.nhanes$SDMVSTRA == 100)
```

```
N12 <- sum(small.nhanes$SDMVSTRA == 101)
N13 <- sum(small.nhanes$SDMVSTRA == 102)
N14 <- sum(small.nhanes$SDMVSTRA == 103)
NL <- c(N1,N2,N3,N4,N5,N6,N7,N8,N9,N10,N11,N12,N13,N14)
#taking the totaln number of rows in each data set and compiling that into a single row vector
N <- sum(NL)
#total population size
cost <- c( 52, 50, 46, 53, 48, 48, 47, 57, 53, 47, 54, 40, 43, 44)
num <- sum(NL * sig / sqrt(cost)) * sum(NL * sig * sqrt(cost))
#calculating the numerator of the population size estimate equation
den <- N^2 * D + sum(NL * sig^2)
#calculating the denominator of the population size estimate equation
n <- num/den
n
#calculating and showing the population size estimate
nj <- n * (NL * sig/ sqrt(cost))/(sum(NL * sig/sqrt(cost)))
round(nj,0)
set.seed(1004149361)
index1 <- sample(1:nrow(strata1), 25, replace = F)
index2 <- sample(1:nrow(strata2), 40, replace = F)
index3 <- sample(1:nrow(strata3), 37,  replace = F)
index4 <- sample(1:nrow(strata4), 26, replace = F)
index5 <- sample(1:nrow(strata5), 23, replace = F)
index6 <- sample(1:nrow(strata6), 35, replace = F)
index7 <- sample(1:nrow(strata7), 24, replace = F)
index8 <- sample(1:nrow(strata8), 15, replace = F)
index9 <- sample(1:nrow(strata9), 23, replace = F)
index10 <- sample(1:nrow(strata10), 18, replace = F)
index11 <- sample(1:nrow(strata11), 21, replace = F)
index12 <- sample(1:nrow(strata12), 38, replace = F)
index13 <- sample(1:nrow(strata13), 16, replace = F)
index14 <- sample(1:nrow(strata14), 9, replace = F)
#sampling random numbers between 1 and each strata size where the sample size is our
previously calculated nj
s1 <- strata1[index1, ]
s2 <- strata2[index2, ]
s3 <- strata3[index3, ]
s4 <- strata4[index4, ]
s5 <- strata5[index5, ]
s6 <- strata6[index6, ]
s7 <- strata7[index7, ]
s8 <- strata8[index8, ]
s9 <- strata9[index9, ]
s10 <- strata10[index10, ]
```

```
s11 <- strata11[index11, ]
s12 <- strata12[index12, ]
s13 <- strata13[index13, ]
s14 <- strata14[index14, ]
s <- rbind(s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11, s12, s13, s14)
#using our number index to find the actual samples of each strata and compiling them into a
new data set
model.smoke2 <- lm(BPSysAve ~ Smoke100, data = s)
#running the same linear model for stratified sample set
Anova(model.smoke2)
model.age2 <- lm(BPSysAve ~ Age, data = s)
#running the same linear model for stratified sample set
summary(model.age2)
model.edu2 <- lm(BPSysAve ~ Education, data = s)
summary(model.edu2)
boxplot(BPSysAve ~ Education, data = s, xlab = "Education", ylab = "BPSysAve", main = "Boxplot
of Blood Pressure in different stages of education")
#boxplot of blood pressure vs educatrion for stratified data set (figure 5)
dev.off()
model.inc2 <- lm(BPSysAve ~ HHIncomeMid, data = s)
summary(model.inc2)
model.age_inc2 <- lm(BPSysAve ~ Age + HHIncomeMid, data = s)
summary(model.age_inc2)
visreg(model.age2, "Age", gg = TRUE) +
  labs(title = "Linear model between Age and Blood Pressure",
       caption = "Stratafied sample(350) of NHANES Data",
       y = "Blood Pressure",
       x = "Age")
#plot of linear regression model for age vs blood pressure in the stratified sample set (figure 2)
dev.off()
visreg(model.inc2, "HHIncomeMid", gg = TRUE) +
  labs(title = "Linear model of Median Household Income and Blood Pressure",
       caption = "Stratafied sample(350) of NHANES Data",
       y = "Blood Pressure",
       x = "HHIncomeMid")
#plot of linear regression model for HHIncomeMid vs blood pressure in the stratified sample
set(figure 3)
dev.off()
```

Reference

(1) Wu, C.-Y.; Hu, H.-Y.; Chou, Y.-J.; Huang, N.; Chou, Y.-C.; Li, C.-P. High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults. *Medicine (Baltimore)* **2015**, *94* (47). https://doi.org/10.1097/MD.0000000000002160.
(2) Publishing, H. H. Reading the new blood pressure guidelines https://www.health.harvard.edu/heart-health/reading-the-new-blood-pressure-guidelines (accessed Dec 17, 2020).
(3) He, J.; Whelton, P. K. Elevated Systolic Blood Pressure and Risk of Cardiovascular and Renal Disease: Overview of Evidence from Observational Epidemiologic Studies and Randomized Controlled Trials. *Am. Heart J.* **1999**, *138* (3 Pt 2), 211–219. https://doi.org/10.1016/s0002-8703(99)70312-1.
(4) Luo, D.; Cheng, Y.; Zhang, H.; Ba, M.; Chen, P.; Li, H.; Chen, K.; Sha, W.; Zhang, C.; Chen, H. Association between High Blood Pressure and Long Term Cardiovascular Events in Young Adults: Systematic Review and Meta-Analysis. *BMJ* **2020**, *370*, m3222. https://doi.org/10.1136/bmj.m3222.
(5) National Health and Nutrition Examination Survey | National Surveys | Information on Data Sources | Vision and Eye Health Surveillance System | Vision Health Initiative (VHI) | CDC https://www.cdc.gov/visionhealth/vehss/data/national-surveys/national-health-and-nutrition-examination-survey.html (accessed Dec 17, 2020).
(6) Rockwood, M. R. H.; Howlett, S. E. Blood Pressure in Relation to Age and Frailty. *Can. Geriatr. J. CGJ* **2011**, *14* (1), 2–7.
(7) Shaw, B. A.; Spokane, L. S. Examining the Association Between Education Level and Physical Activity Changes During Early Old Age. *J. Aging Health* **2008**, *20* (7), 767–787. https://doi.org/10.1177/0898264308321081.
(8) Cornelissen, V. A. Exercise Training for Blood Pressure: A Systematic Review and Meta-analysis | Journal of the American Heart Association https://www.ahajournals.org/doi/full/10.1161/JAHA.112.004473?fbclid=IwAR1BtSIWw8bZ_fvrOCcs2kRQABOrB2-QTwUft81bkkKmKqqOZ2hmznzT7dU (accessed Dec 20, 2020).
(9) Brummett Beverly H.; Babyak Michael A.; Siegler Ilene C.; Shanahan Michael; Harris Kathleen Mullan; Elder Glen H.; Williams Redford B. Systolic Blood Pressure, Socioeconomic Status, and Biobehavioral Risk Factors in a Nationally Representative US Young Adult Sample. *Hypertension* **2011**, *58* (2), 161–166. https://doi.org/10.1161/HYPERTENSIONAHA.111.171272.
(10) Virdis, A.; Giannarelli, C.; Neves, M. F.; Taddei, S.; Ghiadoni, L. Cigarette Smoking and Hypertension. *Curr. Pharm. Des.* **2010**, *16* (23), 2518–2525. https://doi.org/10.2174/138161210792062920.