STAT 344 Group Project

Samuel Leung: 6559 2651

Cuong Andrew Luu: 1672 3686

Xuan Chen: 1573 4643

Eleanor Yi: 9118 7922

Yi Chen Kevin Ding: 8732 7789

Contributions

Samuel Leung (Leader): Coding and writing portions for sample size calculations, SRS and stratified estimates, comparison between estimates, visualizations, and summary of paper.

Xuan chen: Writing for introduction, sampling methodology and characteristics of SRS, data analysis of SRS, calculation CI codes

Eleanor Yi: Writing for introduction, sampling methodology and characteristics of stratified sampling, data analysis of stratified sampling

Cuong Andrew Luu: Ratio and Regression coding portions, paper summary

Kevin Ding: Writing for introduction, comparison, conclusion

Part I

Introduction

Worldwide, an estimated 17.9 million lives are claimed annually by cardiovascular diseases, making it a leading cause of death. In light of this, researching the prevalence of cardiovascular disease is important for the field of public health and healthcare management. It plays a pivotal role in healthcare planning, facilitating the allocation of resources and ensuring the provision of adequate care for affected individuals. Understanding how modern western civilization is at risk from cardiovascular disease is essential towards this goal of implementing preventative healthcare. Towards this goal of gaining understanding, the objective of this report is to estimate the total number of people with cardiovascular diseases and the average blood pressure of the population. We intend to accomplish this goal by utilizing the cardiovascular dataset available on Kaggle (found here). Looking into the data further, obtained from a combination of four databases, namely Cleveland, Hungary, Switzerland, and VA Long Beach. These four places will be the population of concern for our report. We plan estimate population parameters for average blood pressure and total people with cardiovascular disease using various estimation methods, including regression and ratio estimation, as well as different sampling techniques. We aim to compare and contrast the performance of these estimators by examining their standard errors. Through an analysis of the data, this research seeks to contribute to our understanding of the potential risk of cardiac illness.

Methodology

Examining the data

As stated above, the cardiovascular dataset used in this project is from Kaggle, with the link for it found above. The dataset is a singular CSV file with 17 variables and 68205 rows, with each row representing a separate patient. Listing the interpretation of the column names found in the dataset, which can also be found on Kaggle:

- **ID**: Unique identifier of each patient
- **age**: Age of the patient in days
- **age_years**: Age of the patient in years
- **height**: Height of the patient in centimeters
- **weight**: Weight of the patient in kilograms
- **ap_hi**: Systolic blood pressure
- **ap_lo**: Diastolic blood pressure
- **cholesterol**: Cholesterol levels. Categorical variable (1: Normal, 2: Above Normal, 3: Well Above Normal)
- **gluc**: Glucose levels. Categorical variable (1: Normal, 2: Above Normal, 3: Well Above Normal)
- **smoke**: Smoking status. Binary variable (0: Non-smoker, 1: Smoker)

- **alco**: Alcohol intake. Binary variable (0: Does not consume alcohol, 1: Consumes alcohol)
- **active**: Physical activity. Binary variable (0: Not physically active, 1: Physically active)
- **cardio**: Presence or absence of cardiovascular disease. Target variable. Binary (0: Absence, 1: Presence)
- **bmi**: Body Mass Index from height and weight
- **bp_category**: Blood pressure category based on ap_hi and ap_lo.
- **bp_category_encoded**: Encoded version of bp_category for machine learning purposes

Project Goal

Contextualizing the goal of this project into the variables given from our data, these are our following objectives:

- To estimate the total number of people with cardiovascular disease: $\hat{t}_{\text{cardio}}$
- To estimate the average blood pressure of the population: $\bar{u}_{\text{ap\_hi}}$, $\bar{u}_{\text{ap\_lo}}$

Both the systolic and diastolic blood pressures are necessary for determining the blood pressure of a person, which is why both the population average of both properties are estimated. We plan to estimate these parameters along with its standard error through vanilla, ratio, and regression estimates using a simple random sample. We also plan to perform stratified sampling as an alternative way to estimate for these parameters. Standard errors will be used to compare performances of each estimate.

Sample Size Calculation

We can't find the maximum variance from a continuous variable, we go through sample size calculations with the maximum variance possible from a binary variable: the total number of people with cardiovascular disease. We want the margin of error δ of our estimate to be within ± 1000 people, 19 times out of 20. Given the size of our population is ~ 68000, we believe that an estimate within this margin of error range is reasonable.

Let X be the binary random variable that models the number of people with cardiovascular disease, with mean p and variance $p(1-p)$. Let $\bar{y}_S = \hat{p}$ be the estimate of the average number people with cardiovascular disease. We want to estimate the total number of people with cardiovascular disease, $\hat{t}_{p,\text{cardio}} = N\bar{y}_s$. The variance of this estimator will be $var(\hat{t}_{p,\text{cardio}}) = N^2\hat{p}(1-\hat{p})$. Going through our sample size calculations, we want to maximize variance, so we take $\hat{p} = 0.5$. Therefore, ignoring the finite population correction factor for now, we want to solve the following:

$$\begin{aligned} 1000 &\geq 1.96\sqrt{0.25N^2/n_0} \\ \implies \left(\frac{1000}{1.96}\right)^2 &\geq \frac{N^2}{4n_0} \\ \implies n_0 &\geq \left(\frac{N}{2} \cdot \frac{1.96}{1000}\right)^2 \end{aligned}$$

Comparing the sample size calculated above with the total population size, we see that $n_0/N = 0.066 \geq 0.05$. So we cannot ignore the finite population correction factor in our sample size calculations. Factoring in the FPC, we would want to solve the following:

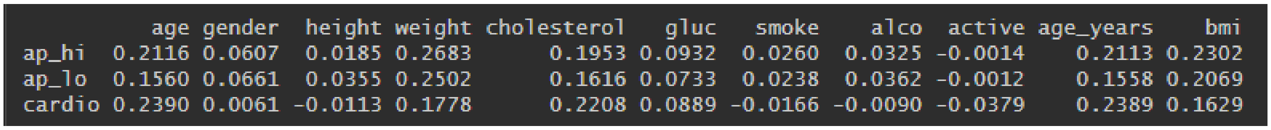
$$n = \frac{n_0}{1 + n_0/N}$$

. All steps for sample size calculation in code can be found in Appendix 1. After performing these steps, we are left with the sample size $n = 4194$.

Simple Random Sampling

The Simple Random Sample (SRS) was collected using the *sample* function found in R (code for generating SRS found in Appendix 2). With the SRS, Vanilla, Ratio, and Regression estimates $\bar{u}_{\text{ap\_hi}}$, $\bar{u}_{\text{ap\_lo}}$, and $\hat{t}_{\text{cardio}}$, along with their standard error, were calculated. The code for these calculations can be found in Appendices 3, 4, and 5.

To calculate regression and ratio estimates, we want to use an auxiliary variable with strong correlation. We do so by first calculating the correlation values for each of our target variables against the other numeric columns in our data set (code for the following graphic found in Appendix 6):



```
          age gender  height weight cholesterol   gluc   smoke    alco  active age_years    bmi
ap_hi  0.2116 0.0607  0.0185 0.2683      0.1953 0.0932  0.0260  0.0325 -0.0014    0.2113 0.2302
ap_lo  0.1560 0.0661  0.0355 0.2502      0.1616 0.0733  0.0238  0.0362 -0.0012    0.1558 0.2069
cardio 0.2390 0.0061 -0.0113 0.1778      0.2208 0.0889 -0.0166 -0.0090 -0.0379    0.2389 0.1629
```

Figure 1. Correlation Table

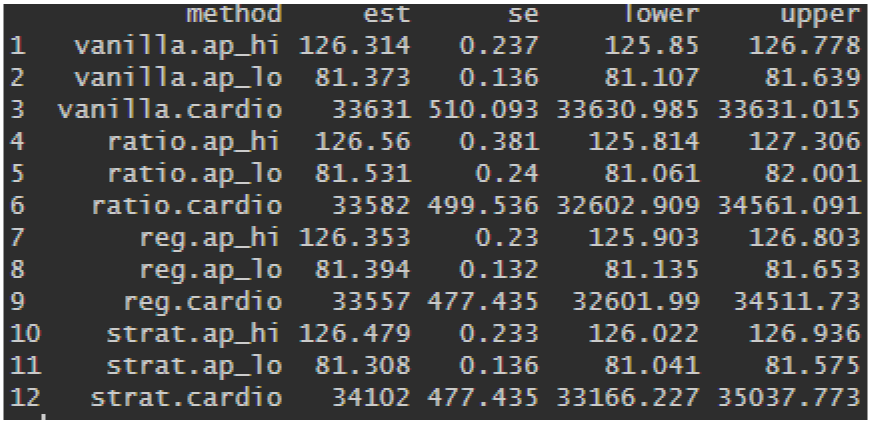
The highest positive-correlating variable for ap_hi and ap_lo are weight. The highest positively-correlating variable for cardio is age, however it is difficult to interpret the value in this column, as the age column is not based on year. So we opt to use the age_years column, which has similar correlation values to age. We note that for all these correlations, none of them seem particularly strong, which suggests that the ratio and regression estimates may not be that useful in prediction.

Stratified Sampling

We want to stratify our population based on the variable that will create the greatest distinction between stratas. From our intuition, there are a few candidate variables that fulfill this requirement, such as variables gender, smoke, and alco. From an article written by the Center for Disease Control and Prevention (CDC), it seems that smoking of any capacity increases the likelihood of cardiovascular disease (CVD), so we suspect that the smoke variable would partition the stratas the greatest. Assuming equal variance between the stratas of smokers and non-smokers, we sample using proportional allocation:

$$\frac{n_h}{n} = \frac{N_h}{N}$$

Code for generating the stratified sample, as well as calculating the stratified estimates and standard error, can be found in Appendices 7 and 8. The estimates, standard error, and confidence intervals for the vanilla, ratio, regression, and stratified methods are shown below. Columns lower and upper represent lower and upper bounds of the 95% confidence interval for the corresponding estimate. The code to generate the following table can be found in Appendix 10:



```
           method     est      se      lower      upper
1   vanilla.ap_hi 126.314   0.237     125.85    126.778
2   vanilla.ap_lo  81.373   0.136     81.107     81.639
3  vanilla.cardio   33631 510.093  33630.985  33631.015
4     ratio.ap_hi  126.56   0.381    125.814    127.306
5     ratio.ap_lo  81.531    0.24     81.061     82.001
6    ratio.cardio   33582 499.536  32602.909  34561.091
7       reg.ap_hi 126.353    0.23    125.903    126.803
8       reg.ap_lo  81.394   0.132     81.135     81.653
9      reg.cardio   33557 477.435   32601.99   34511.73
10    strat.ap_hi 126.479   0.233    126.022    126.936
11    strat.ap_lo  81.308   0.136     81.041     81.575
12   strat.cardio   34102 477.435  33166.227  35037.773
```

Figure 2. Estimates, Standard Error, and Confidence Widths for all methods

All estimates and standard errors are rounded to 3 decimal places, besides the cardio estimates. Those estimates are for the number of people in the population with cardiovascular disease, so we round to the nearest integer. According to the CDC, normal blood pressure is lower than 120 systolic pressure (ap_hi) and 80 diastolic pressure (ap_lo). People with 120-139 systolic and 80-89 diastolic are at risk of high blood pressure, and people with systolic and diastolic blood pressures higher than the previously listed ranges are diagnosed with high blood pressure (article here).All estimates for ap_hi and ap_lo are around 126 systolic and 81 systolic, meaning that the average population is at risk of high blood pressure. All estimates also indicate that around 33000 $\sim$ 34000 individuals within the target population have high blood pressure. Given that our total population size is $\sim$ 68000, our estimates project that around half of the entire population has cardiovascular illness.

Discussion

Further analysis for ratio and regression estimates

Prior to computing regression and ratio estimates, we incorporate an auxiliary variable with a strong correlation. We calculate correlation values for each target variable against other numeric columns in our dataset (Figure 1). Weight exhibits the highest positive correlation with ap_hi and ap_lo. For cardio, age shows the highest positive correlation; however, its interpretation is challenging due to non-yearly age counting. Instead, we opt for the age_years column, which has similar correlation values to age. Despite the correlations not being notably strong, indicating potential limitations in the utility of ratio and regression estimates for prediction, we conduct scatter plots (Figure 3) to further assess their effectiveness against the proposed auxiliary variables. Code for the following plot can be found in Appendix 9:

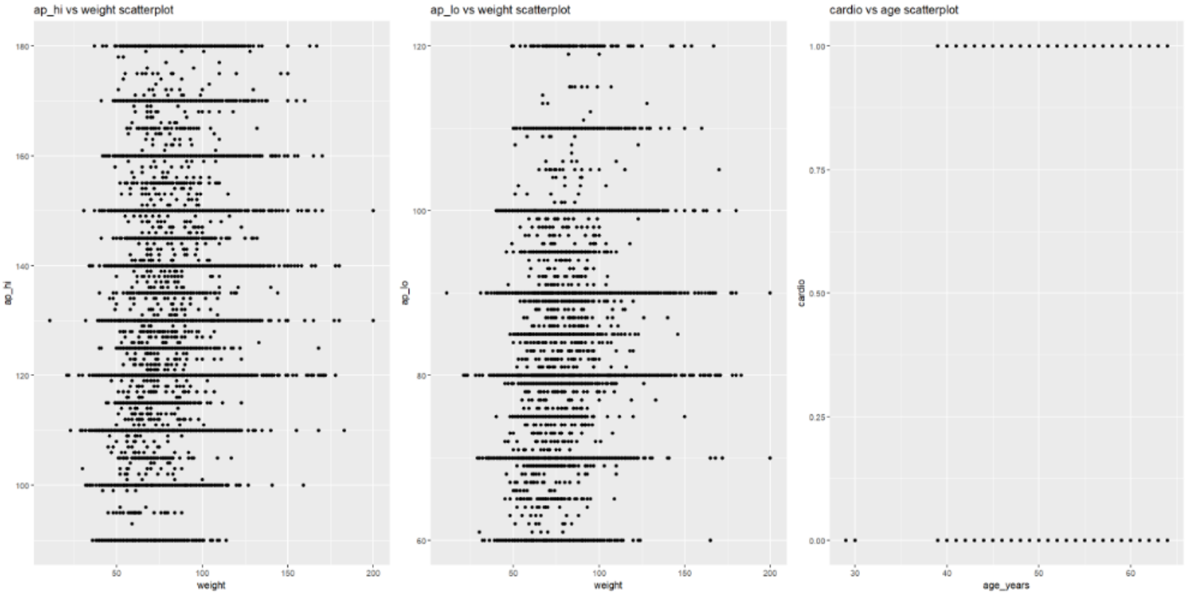


Figure 3. Scatter plots of target variables against auxiliary variables

For the scatter plots presented above, it is evident that there is minimal to no linear correlation observed in all cases. Notably, when examining the cardio vs. age_years plot, utilizing a linear regression estimate is deemed inappropriate due to the binary nature of the cardio variable.

Comparing performance between different estimates

We now compare and contrast the advantages and disadvantages of each estimation method. For the vanilla estimate, its advantage is that it is easy to calculate, but ignores auxiliary information and so we cannot lower the standard error of this estimate through additional information. For ratio estimates, its standard error will be lower than that of the vanilla estimate if it is paired with an auxiliary variable with strong positive linear correlation with the target variable. However, its performance deteriorates (higher standard error) when paired with an auxiliary variable with weak linear correlation, or negative correlation. Graphically, the ratio estimate is also forced to pass through the origin, meaning that if the data has some sort of intercept, the ratio estimate cannot account for this. The regression estimate, however, can handle this case, and is not reliant on its auxiliary variable having positive correlation with the target. Regression estimates are still affected by the magnitude of correlation between its auxiliary variable and the target, so weakly-correlated auxiliary variables cause the regression estimate to be less effective. Finally, stratified estimates can outperform vanilla estimates given that the data can be partitioned into (approximately) clear, distinct groups, but that is dependent on the data having suitable splits given a stratifying variable. As all sampling in this report is done in code, there is no cost in sampling, so arguments for cost of sampling for these methods are not mentioned.

From this course, we learn to evaluate estimates based on its standard error (lower standard errors mean better estimate). From Figure 2, we see that the regression estimate has the lowest standard error for estimating ap_hi and ap_lo, while the stratified and regression estimates tied in performance for estimating cardio. We observe that the estimated values for each target variable across all estimating methods are roughly the same.

When comparing the standard errors of each, we see that the ratio estimator has the largest SE out of all four methods. This is expected as from the correlation table and scatter plots above, there is little to no positive correlation between the target and response variable, making the ratio estimator a poor choice. The standard errors for ap_hi and ap_lo are similar for vanilla and regression estimates, with the vanilla estimate having slightly lower standard errors. Given the low correlation values shown above, the regression line formed will probably be similar to the constant prediction line that the vanilla estimate has, resulting in similar estimates and standard error. However, for the cardio response variable, given that the response is binary, linear regression would not be a suitable method to model the data, which is why we see such a higher standard error using the regression method when compared to the vanilla estimate.

Despite low correlation values between target and auxiliary variables, the regression estimate performed the best when considering all 3 target variables. Given that the regression estimate outperformed the stratified estimate, we are doubtful towards the smoke variable stratifying our target population well.

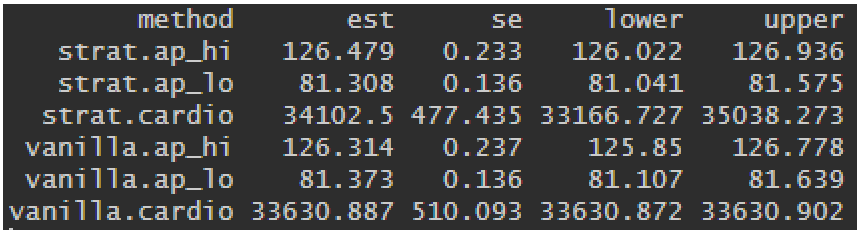
Comparing stratified estimate performance

Our stratified estimate was constructed under the assumption that the variance of target variables across stratas would be roughly equal. Under this assumption, we sampled under proportional allocation, which allows us to simplify the variance of the stratified estimate to be solely dependent on the within-strata variance. From class, we saw a way to compare the performance of the stratified estimate with the vanilla estimate, via its variance:

$$\frac{Var(\bar{y}_{\text{str}})}{Var(\bar{y})} = \frac{S^2_{P,W}}{S^2_P} = 1 - \frac{S^2_{P,B}}{S^2_P}$$

S^2_P, $S^2_{P,W}$, $S^2_{P,B}$ represent total population variance, within-strata population variance, and between-strata population variance respectively. Given the above formula, we can infer that the stratified estimate has a lower standard error when the overall variance is dominated by between-strata variance. Estimating for these population parameters via our sample gives us a way to verify how well our chosen stratified variable (smoke) did at making the stratas distinct. Modifying the table above to only include stratified and vanilla estimates (code for this listed in Appendix 10):

We take the standard error of the estimates to be the estimate for the square root of total population variance. For the target variables ap_hi and ap_lo, we see from above (Figure 4) that the standard errors are almost identical. This indicates that for these target variables, $\hat{S}^2_{P,W}/\hat{S}^2_P \approx 1$, meaning that overall variance is not dominated by between-strata variance and suggesting that our chosen stratifying variable did not do a great job at making these stratas distinct. For the cardio



```
        method       est      se     lower     upper
   strat.ap_hi   126.479   0.233   126.022   126.936
   strat.ap_lo    81.308   0.136    81.041    81.575
  strat.cardio   34102.5 477.435 33166.727 35038.273
 vanilla.ap_hi   126.314   0.237    125.85   126.778
 vanilla.ap_lo    81.373   0.136    81.107    81.639
vanilla.cardio 33630.887 510.093 33630.872 33630.902
```

Figure 4. Estimate, Standard Error, and Confidence Interval for Vanilla and Stratified Estimates

target variable the standard errors between the stratified estimate and the vanilla estimate have a greater difference (SE of 477 for stratified estimate vs. SE of 510 for vanilla). Taking into account that the differences between these values will be larger once we consider the square of the standard errors (which gives us the estimate of population variance), it seems that the stratifying variable smoke provided more distinct stratification when used for the cardio target variable.

We based our choice of stratifying variable from results given by the CDC, which indicated that smoking and CVD are positively correlated. However, this does not necessarily mean that blood pressure will be affected by smoking. Also, in the dataset cardio is listed as a binary variable, while ap_hi and ap_lo are continuous ones. Our intuition guides us to believe that it is easier to create distinct stratas for discrete data versus contiuous data types. All this combined, it is sensible to see the smoke variable provide better stratification for cardiovascular disease presence in comparison to blood pressure metrics (ap_hi and ap_lo).

Conclusion

Our goal is to estimate the average blood pressure and total number of people with cardiovascular disease within our target population of individuals from Cleveland, Hungary, Switzerland, and VA Long Beach. Towards this end, we performed a simple random sample of the population, and calculated vanilla, ratio, and regression estimates for average ap_hi and ap_lo (which both correspond to average blood pressure), as well as total cardio (the total number of people with cardiovascular disease). Corresponding standard errors and 95% confidence intervals were also included. We also performed stratified sampling to obtain a stratified estimate of the above population parameters, with corresponding standard errors and 95% confidence intervals. Despite low correlation values to auxiliary variables, the regression estimate had the lowest standard error for all 3 target variables. Our chosen stratifying variable, smoke, had decent performance for stratifying for people with cardiovascular disease, but was poor at creating distinct stratas for differing levels of blood pressure. This made sense given that our choice of stratifying variable was based on previous research on solely cardiovascular disease presence, coupled with the fact that continuous data types are more difficult to stratify distinctly compared to discrete data types.

Our target population was collected from 4 major cities, and may not be representative of the world as a whole. Therefore, our results cannot be generalized to a larger worldwide population, but is still informative of how at risk people in these regions are of cardiovascular disease, as well as how many people in our target population have cardiovascular disease. With further research into discovering more influential factors related to blood pressure and cardiovascular disease, our ratio, regression, and stratifying methodologies could be re-done to produce more accurate estimates of each parameter.

Part II

This paper discusses the dangers of solely relying on optimizing mathematical properties of statistical tests and ignoring statistical intuition. Statistical tests provide a way for its user to determine if differences between theorized results and experimental ones are due to chance (ie. null hypothesis) or because of some other underlying reason (ie. alternative hypothesis). The effectiveness of statistical tests can be evaluated based on various measures. Two notable measures are a test's statistical power, which evaluates how well it can detect hidden relationships within observed data when there actually is an underlying factor, and its false positive rate, which is how often the test incorrectly identifies a hidden influence. There are ways of manipulating statistical tests to artificially increase these measures, such as enlarging the subset of cases which support the alternative hypothesis when these cases should give evidence to the null. This modified test would have higher statistical power when compared to the unmodified test at the same false positive rate, but its reliability as an indicator of discovering meaningful relationships in data would be compromised by construction. The goal of statistical testing should be to evaluate the extent that proposed hypotheses reflect reality, given observed data. To focus on optimizing and manipulating properties of these tests for the goal of falsely increasing a statistical test's effectiveness goes against its very objective, which is why statistical intuition should also be taken into account when interpreting and verifying the construction and results of a statistical test.

Appendix

1 Code for Sample Size Calculations

```
# Reading in the data
data <- read.csv("data/cardio_data_processed.csv")
N <- nrow(data)
target <- 1000
# n_0 is the sample size while ignoring FPC
n_0 <- (1.96 * N / (2*target))^2
# this returns a value greater than 0.05, so we don’t ignore FPC
print(paste0("n_0 / N: ", n_0 / N))
# n becomes the actual sample size after accounting for FPC
n <- ceiling(n_0 / (1 + n_0 / N))
```

2 Code for generating Simple Random Sample

```
# set seed for reproduceable results
set.seed(124)
# generate random sample
srs <- data[sample(1:nrow(data), n), ]
```

3 Code for Vanilla Estimates and Standard Errors

```
fpc <- 1 - (n / N)
target_lst <- c("ap_hi", "ap_lo", "cardio")

ap_hi.est <- round(mean(srs$ap_hi), 3)
ap_hi.var <- fpc*var(srs$ap_hi) / n
ap_hi.se <- sqrt(ap_hi.var)
ap_hi.moe <- round(1.96*sqrt(ap_hi.var), 3)

ap_lo.est <- round(mean(srs$ap_lo), 3)
ap_lo.var <- fpc*var(srs$ap_lo) / n
ap_lo.se <- sqrt(ap_lo.var)
ap_lo.moe <- round(1.96*sqrt(ap_lo.var), 3)

cardio.p <- mean(srs$cardio)
cardio.est <- round(N*cardio.p, 0)
cardio.var <- fpc*(cardio.p)*(1-cardio.p) / n
cardio.se <- N * sqrt(cardio.var)
cardio.moe <- round(1.96*sqrt(cardio.var), 3)
```

4 Code for Ratio Estimates and Standard Errors

```
# ratio estimate:
ratio.ap_hi.ratio <- mean(srs$ap_hi) / mean(srs$weight)
ratio.ap_lo.ratio <- mean(srs$ap_lo) / mean(srs$weight)
ratio.cardio.ratio <- mean(srs$cardio) / mean(srs$age_years)

ratio.ap_hi.se <- sqrt(fpc*sum(var(srs$ap_hi-ratio.ap_hi.ratio*srs$weight))/n)
```

```r
ratio.ap_lo.se <- sqrt(fpc*sum(var(srs$ap_lo-ratio.ap_lo.ratio*srs$weight))/n)
ratio.cardio.se <- N*sqrt(fpc*sum(var(
  srs$cardio-ratio.cardio.ratio*srs$age_years))/n)

ratio.ap_hi.est <- ratio.ap_hi.ratio * mean(data$weight)
ratio.ap_lo.est <- ratio.ap_lo.ratio * mean(data$weight)
ratio.cardio.est <- round(ratio.cardio.ratio * mean(data$age_years) * N, 0)
```

5 Code for Regression Estimates and Standard Errors

```r
# helper function for regression estimate. X, y should be matrices of proper size.
regression.est_and_se <- function(X, y, x.pop) {
  X <- as.matrix(X); y <- as.matrix(y)
  X <- cbind(rep(1, dim(X)[1]), X)
  w <- solve(t(X)%*%X)%*%t(X)%*%y
  y.hat <- X%*%w
  se <- sqrt(fpc*sum(var(y - y.hat))/nrow(X))
  est <- cbind(1, x.pop) %*% w

  ret <- c(est = as.numeric(est), se = as.numeric(se),
           lower = as.numeric(est - 1.96*se),
           upper = as.numeric(est + 1.96*se))
  return(round(ret, 3))
}

reg.ap_hi <- regression.est_and_se(srs$weight, srs$ap_hi, mean(data$weight))
reg.ap_lo <- regression.est_and_se(srs$weight, srs$ap_lo, mean(data$weight))
# reg.cardio estimates population mean, so we need to multiply estimate and SE by N
reg.cardio <- N*regression.est_and_se(srs$age_years, srs$cardio, mean(data$age_years))
# round to nearest integer
reg.cardio[1] <- round(reg.cardio[1], 0)
```

6 Viewing Correlations

```r
pop.cor <- cor(data[,2:15])
round(pop.cor[target_lst, !(colnames(pop.cor)%in%target_lst)], 4)
```

7 Code for generating Stratified Sample

```r
# generate stratified sample
set.seed(124)
n.smoke <- round(prop.table(table(data$smoke)) * n, 0)
pop.smoke0 <- data[data$smoke==0,]; pop.smoke1 <- data[data$smoke==1,]
strata.sample <- rbind(
    pop.smoke0[sample(1:nrow(pop.smoke0), n.smoke[1]),],
    pop.smoke1[sample(1:nrow(pop.smoke1), n.smoke[2]), ])
```

8 Code for Stratified Estimate and Standard Errors

```
# given population data, sample and calculate the stratified estimate + SE
stratify.est_and_se <- function(pop, sample, var_str,
                                target_str, total=FALSE) {
  pop.var.count <- table(data[,var_str])
  sample.var.count <- table(sample[,var_str])
  # population proportion for each strata
  pop.prop <- prop.table(pop.var.count)

  sample.means <- tapply(sample[,target_str], sample[,var_str], mean)
  sample.vars <- tapply(sample[,target_str], sample[,var_str], var) / sample.var.count
  fpc <- 1 - sample.var.count / pop.var.count

  est <- round(sum(pop.prop * sample.means), 3)
  se <- round(sqrt(sum(pop.prop^2 * fpc * sample.vars)), 3)
  if (total==TRUE) {est <- round(N*est, 0); se <- N*se}
  ret <- c(est = est, se = se,
           lower = round(est - 1.96*se, 3),
           upper = round(est + 1.96*se, 3))
  return(ret)
}

strat.ap_hi <- stratify.est_and_se(data, strata.sample,
                                   var_str = "smoke", target_str = "ap_hi")
strat.ap_lo <- stratify.est_and_se(data, strata.sample,
                                   var_str = "smoke", target_str = "ap_lo")
strat.cardio <- stratify.est_and_se(data, strata.sample, var_str = "smoke",
                    target_str = "cardio", total=TRUE)
```

9 Code for correlation plots

```
library(ggplot2)
library(cowplot)

aphi_vs_weight <- ggplot(data = data, aes(x=weight, y=ap_hi)) +
  geom_point() +
  labs(title = "ap_hi vs weight scatterplot")
aplo_vs_weight <- ggplot(data = data, aes(x=weight, y=ap_lo)) +
  geom_point() +
  labs(title = "ap_lo vs weight scatterplot")
cardio_vs_age <- ggplot(data = data, aes(x=age_years, y=cardio)) +
  geom_point() +
  labs(title = "cardio vs age scatterplot")

plot_grid(aphi_vs_weight, aplo_vs_weight, cardio_vs_age, ncol = 3)
```

10 Code for generating Summary Tables

```
# Create invidual tables for each estimator method
vanilla.summary <- as.data.frame(round(rbind(
  c(ap_hi.est, ap_hi.se, ap_hi.est - ap_hi.moe, ap_hi.est + ap_hi.moe),
  c(ap_lo.est, ap_lo.se, ap_lo.est - ap_lo.moe, ap_lo.est + ap_lo.moe),
  c(cardio.est, cardio.se, cardio.est - cardio.moe, cardio.est + cardio.moe)
), 3))

ratio.summary <- as.data.frame(round(rbind(
  c(ratio.ap_hi.est, ratio.ap_hi.se,
    ratio.ap_hi.est - 1.96 * ratio.ap_hi.se,
    ratio.ap_hi.est + 1.96 * ratio.ap_hi.se),
  c(ratio.ap_lo.est, ratio.ap_lo.se,
    ratio.ap_lo.est - 1.96 * ratio.ap_lo.se,
    ratio.ap_lo.est + 1.96 * ratio.ap_lo.se),
  c(ratio.cardio.est, ratio.cardio.se,
    ratio.cardio.est - 1.96 * ratio.cardio.se,
    ratio.cardio.est + 1.96 * ratio.cardio.se)
), 3))

reg.summary <- as.data.frame(rbind(
  reg.ap_hi, reg.ap_lo, reg.cardio
))

strat.ap_hi <- stratify.est_and_se(data, strata.sample, var_str = "smoke", target_str = "ap_hi")
strat.ap_lo <- stratify.est_and_se(data, strata.sample, var_str = "smoke", target_str = "ap_lo")
strat.cardio <- stratify.est_and_se(data, strata.sample, var_str = "smoke", target_str = "cardio", total=TRUE)

# formatting
colnames(vanilla.summary) <- c("est", "se", "lower", "upper")
vanilla.summary$method <- paste("vanilla.", target_lst, sep="")
colnames(ratio.summary) <- c("est", "se", "lower", "upper")
ratio.summary$method <- paste("ratio.", target_lst, sep="")
reg.summary <- cbind(rownames(reg.summary), reg.summary)
colnames(reg.summary)[1] <- "method"
rownames(reg.summary) <- NULL
strat.summary <- rbind(strat.ap_hi, strat.ap_lo, strat.cardio)
strat.summary <- cbind(rownames(strat.summary), strat.summary)
colnames(strat.summary)[1] <- "method"
rownames(strat.summary) <- NULL

# combine into one table
tot.summary <- rbind(vanilla.summary, ratio.summary, reg.summary, strat.summary)[, c(5,1,2,3,4)]

# making table for only stratified and vanilla estimates:
strat_vs_vanilla.summary <- rbind(strat.summary, vanilla.summary)
```

11 Link to data

https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease/