R Analysis Example Replication C5

```
#section 1 data production
library (sas7bdat)
library (survey)

#nhanes first
library(haven)
#data <- read_sas("<path to your SAS file>")
nhanesdata <- read_sas("P:/ASDA 2/Data sets/nhanes 2011_2012/nhanes1112_sub_8aug2016.sas7bdat")
summary(nhanesdata)

#create factor variables
nhanesdata$racec <- factor(nhanesdata$RIDRETH1, levels = 1: 5 , labels =c("Mexican", "Other Hispanic", "White",
"Black", "Other"))
nhanesdata$marcatc <- factor(nhanesdata$marcat, levels = 1: 3, labels =c("Married", "Previously Married", "Never
Married"))
nhanesdata$edcatc <- factor(nhanesdata$edcat, levels = 1: 4, labels =c("0-11", "12", "13-15","16+"))
nhanesdata$bp_catc <- factor(nhanesdata$bp_cat, levels = 1: 4, labels =c("Normal", "Pre-HBP", "Stage 1
HBP","Stage 2 HBP"))
#nhanesdata$agecsq <- (nhanesdata$age * nhanesdata$age)
names(nhanesdata)
nhanessvy2 <- svydesign(strata=~SDMVSTRA, id=~SDMVPSU, weights=~WTMEC2YR, data=nhanesdata, nest=T)
subnhanes <- subset(nhanessvy2 , RIDAGEYR >= 18)
names (nhanessvy2)

#ncs-r next
ncsr <- read_sas("P:/ASDA 2/Data sets/ncsr/ncsr_sub_5apr2017.sas7bdat")
names(ncsr)
#create factor versions with labels
ncsr$racec <- factor(ncsr$racecat, levels = 1: 4, labels =c("Other", "Hispanic", "Black", "White"))
ncsr$mar3catc <- factor(ncsr$MAR3CAT, levels = 1: 3, labels =c("Married", "Previously Married", "Never
Married"))
ncsr$ed4catc <- factor(ncsr$ED4CAT, levels = 1: 4, labels =c("0-11", "12", "13-15","16+"))
ncsr$sexc <- factor(ncsr$SEX, levels = 1:2, labels=c("Male","Female"))
ncsr$ag4catc <- factor(ncsr$ag4cat, levels = 1:4, labels=c("18-29", "30-44", "45-59", "60+"))
ncsrsvyp1 <- svydesign(strata=~SESTRAT, id=~SECLUSTR, weights=~NCSRWTSH, data=ncsr, nest=T)
names (ncsrsvyp1)
ncsrp2 <- subset(ncsr, !is.na(NCSRWTLG))
ncsrsvyp2 <- svydesign(strata=~SESTRAT, id=~SECLUSTR, weights=~NCSRWTLG, data=ncsrp2, nest=T)
names (ncsrsvyp2)
ncsr$popweight <- (ncsr$NCSRWTSH*(209128094/9282))
ncsrsvypop <- svydesign(strata=~SESTRAT, id=~SECLUSTR, weights=~popweight, data=ncsr, nest=T)
summary(ncsrsvypop)

#hrs, similar needs for ASDA2
#both hh and r weights are needed plus financial respondent for hh level analysis
hrs <- read_sas("p:/ASDA 2/Data sets/hrs 2012/hrs_sub_28sep2016.sas7bdat")
summary(hrs)
hrssvyhh <- svydesign(strata=~STRATUM, id=~SECU, weights=~NWGTHH , data=hrs, nest=T)
summary(hrssvyhh)
hrssvysub <-subset(hrssvyhh, NFINR==1)
summary(hrssvysub)
```

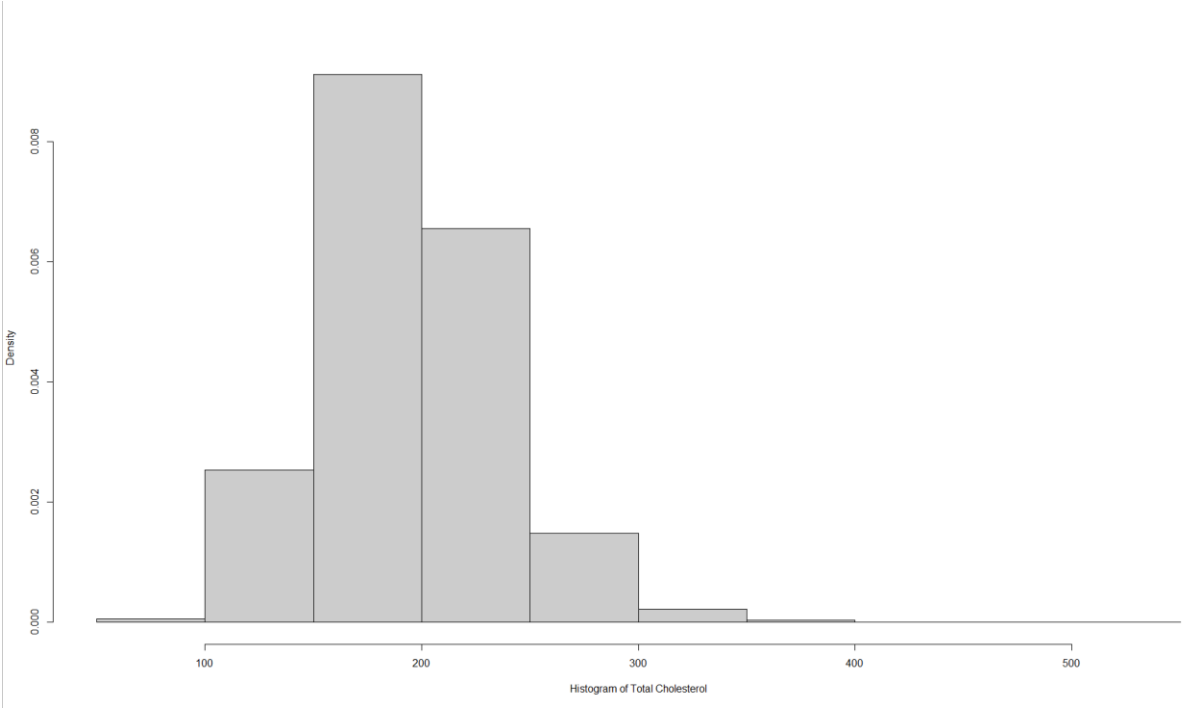
```
hrssvyr <- svydesign(strata=~STRATUM, id=~SECU, weights=~NWGTR , data=hrs, nest=T)
summary(hrssvyr)
#section 2 chapter 5 analysis examples replication, ASDA2
# figures 5.1 and 5.2
svyhist(~LBXTC , subset (nhanessvy2, age >=18), main="", col="grey80", xlab ="Histogram of Total Cholesterol")
#CREATE A VARIABLE CALLED GENDER FOR BOXPLOT
nhanessvy2<-update(nhanessvy2, gender=cut(RIAGENDR, c(1, 2, Inf), right=F))
svyboxplot(LBXTC~gender , subset (nhanessvy2, age >=18), col="grey80", ylab="Total Cholesterol", xlab ="1=Male
2=Female")
#Example 5.3
svytotal (~mde, ncsrsvypop, deff=T)
confint(svytotal(~mde, ncsrsvypop))
#MDE OVER MARITAL STATUS
ex53 <- svyby (~mde, ~mar3catc, ncsrsvypop, svytotal)
ex53 <- svyby (~mde, ~mar3catc, ncsrsvypop, svytotal, deff=T)
ex53
confint(ex53)
#Example 5.4 HH Level Wealth/Total Assets
svyby (~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, svytotal)
confint(svyby (~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svytotal))
#Example 5.5 HRS HH Income
svyby (~H11ITOT, ~I(NFINR==1), hrssvyhh, na.rm=T, svymean)
confint(svyby (~H11ITOT, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svymean))
#Example 5.6 Mean Systolic Blood Pressure, NHANES data
a <- svymean(~BPXSY1 , subset (nhanessvy2, age >=18), na.rm=TRUE)
coef(a)
SE(a)
confint(a)
#Example 5.7
svyby (~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svymean)
confint(svyby(~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svymean))
#Example 5.8 Standard Deviation of Cholesterol NHANES data
#Create a data object with weights only but no design variables
nhaneswgt <- svydesign(id=~1, weights=~WTMEC2YR, data=nhanesdata)
summary(nhaneswgt)
#Subset of those with positive weight and age 18plus
subnhaneswgt <- subset(nhaneswgt, age >= 18 & WTMEC2YR > 0 )
summary(subnhaneswgt)
#obtain mean
a <- svymean(~LBXTC + LBDHDD, design=subnhaneswgt, na.rm=T, deff="replace")
a
# use sqrt of variance to obtain standard deviation
sd <- sqrt(svyvar(~LBXTC + LBDHDD, design = subnhaneswgt, na.rm=T))
sd
#Example 5.9 Population Percentiles for total HH Wealth HRS data, in subset of NFINR=1
q <- svyquantile(~H11ATOTA, hrssvysub, c(.25,.5,.75), na.rm=T, ci=T)
q
# Obtain SE from confidence intervals, see R documentation for details
SE(q)
```

```
#Example 5.10 Lorenz Curve and GINI coefficient not available in R Survey Package, Summer 2017
#Example 5.10 Now Available as of 16nov2017 with "convey" package, add example here Berglund

library(convey)
# linearized design, use hrssvysub created previously
hrssvyhh_c <- convey_prep(hrssvyhh)
# now can subset to financial respondents (after convey_prep)
sub_hrssvyhh_c <- subset( hrssvyhh_c , NFINR==1)
# run svygini and svylorenz using subset, note that R does not require negative set to 0 as Stata
svygini( ~H11ATOTA, design = sub_hrssvyhh_c)
svylorenz( ~H11ATOTA, sub_hrssvyhh_c, seq(0,1,.1), alpha = .01 )


#Example 5.11 Relationship between 2 continuous variables, note this is weighted and design based
svyplot(LBXTC~LBDHDD, subset(subnhanes, age>=18), style="bubble", ylab="HDL", xlab="Total Cholesterol")
#EXAMPLE 5.11 Correlation between Total and High Cholesterol, NHANES DATA
#create standardized versions of variables first, then use in regression
nhanesdata$stdlbxtc <- (nhanesdata$LBXTC-194.4355)/41.05184
summary(nhanesdata$LBXTC + nhanesdata$stdlbxtc)

nhanesdata$stdlbdhdd <- (nhanesdata$LBDHDD-52.83826)/14.93157
summary(nhanesdata$stdlbxtc)
#reset survey design and subset
nhanessvy2 <- svydesign(strata=~SDMVSTRA, id=~SDMVPSU, weights=~WTMEC2YR, data=nhanesdata, nest=T)
subnhanes <- subset(nhanessvy2 , age >= 18)
#Design based linear regression to obtain correlation and correct SE
summary(Ex5_11_svyglm <- svyglm(stdlbxtc ~ stdlbdhdd, design=subnhanes))

#Example 5.12 Ratio Estimator for HDD to Total Cholesterol
ex5_12 <- svyby (~LBDHDD, denominator=~LBXTC, by=~I(age >= 18), nhanessvy2, na.rm=T, ci=T, svyratio)
confint(ex5_12)
#Example 5.13 Proportions of Diabetes by Gender in Subpopulation of Age >=70
subhrs70 <- subset(hrssvyr, NAGE >= 70)
ex5_13 <- svyby(~diabetes, ~GENDER, subhrs70, svymean, keep.names=T, na.rm=T))
print(ex5_13)
confint(ex5_13)
#Example 5.14 Mean Systolic Blood Pressure by Gender, Age 46+ NHANES
subnhanes46 <-subset(nhanessvy2, age >= 46)
#RIAGENDR 1=MALE 2=FEMALE
ex5_14 <- svyby(~BPXSY1, ~RIAGENDR, subnhanes46, svymean, keep.names=T, na.rm=T)
print(ex5_14)
confint(ex5_14)
#Example 5.15 Differences in Mean HH Wealth by Educational Attainment, HRS data
#CODES FOR EDCAT: 1=0-11 2=12 3=13-15 4=16+ YEARS OF EDUCATION
ex5_15 <- svyby(~H11ATOTA, ~edcat, hrssvysub, svymean, na.rm=T, options(survey.lonely.psu="remove"))
print(ex5_15)
confint(ex5_15)
svycontrast(ex5_15, list(avg=c(.5,0,0,.5), diff=c(1,0,0,-1)))
#Example 5.16 Differences in Total Wealth over Time 2010 to 2012, HRS data
#Use 2010 and 2012 data set prepared in SAS
hrs_2010_2012 <- read_sas("p:/ASDA 2/Data sets/hrs 2012/hrs 2010/hrs_2010_2012_c5.sas7bdat")
summary(hrs_2010_2012)
names(hrs_2010_2012)
```

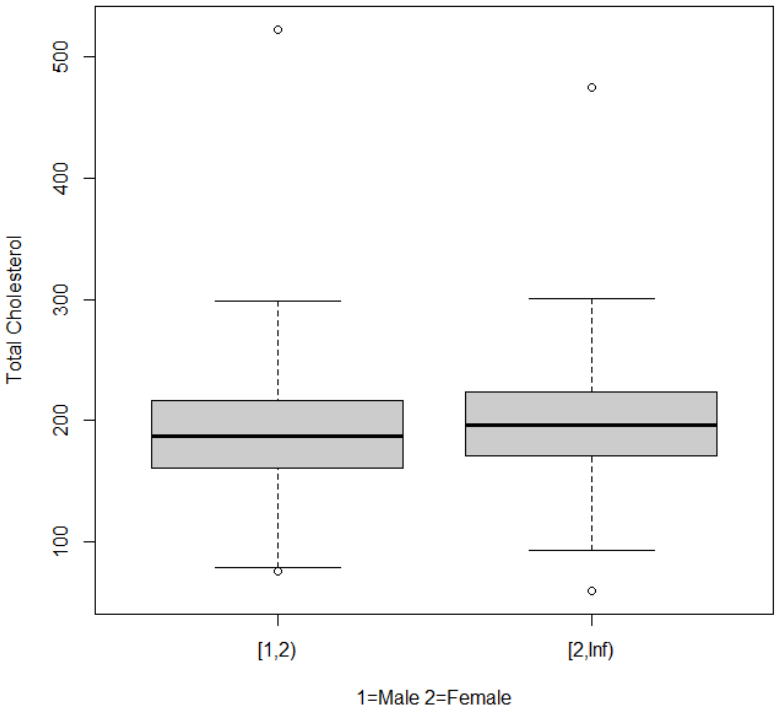
```
hrs2010_2012 <- svydesign(strata=~STRATUM, id=~SECU, weights=~hhweight, data=hrs_2010_2012, nest=T)
subhrs2010 <- subset(hrs2010_2012, finr2010_2012==1)
ex5_16 <- svyby (~totwealth, ~year, design=subhrs2010, keep.vars=T, svymean)
coef(ex5_16)
SE(ex5_16)
contrast <- svycontrast(ex5_16, list(avg=c(.5,.5), diff=c(1,-1)))
print(contrast)
```

Output R Analysis Example Replication C5

```
> # figures 5.1 and 5.2
> svyhist(~LBXTC , subset (nhanessvy2, age >=18), main="", col="grey80", xlab ="Histogram of Total Cholesterol")
```




```
> #CREATE A VARIABLE CALLED GENDER FOR BOXPLOT
> nhanessvy2<-update(nhanessvy2, gender=cut(RIAGENDR, c(1, 2, Inf), right=F))
> svyboxplot(LBXTC~gender , subset (nhanessvy2, age >=18), col="grey80", ylab="Total Cholesterol", xlab ="1=Male
2=Female")
```


Total Cholesterol
100
200
300
400
500
[1,2)
[2,Inf)
1=Male 2=Female

```
> #Example 5.3
> svytotal (~mde, ncsrsvypop, deff=T)
       total       SE  DEff
mde 40092207  2567488 9.028
> confint(svytotal(~mde, ncsrsvypop))
       2.5 %   97.5 %
mde 35060023 45124390
>
> #MDE OVER MARITAL STATUS
> ex53 <- svyby (~mde, ~mar3catc, ncsrsvypop, svytotal)
> ex53 <- svyby (~mde, ~mar3catc, ncsrsvypop, svytotal, deff=T)
> ex53
                            mar3catc      mde        se DEff.mde
Married                      Married 20304191 1584108.6 6.817920
Previously Married Previously Married 10360671  702621.5 2.966192
Never Married           Never Married  9427345  773137.6 3.063915
> confint(ex53)
                      2.5 %   97.5 %
Married            17199395 23408986
Previously Married  8983558 11737783
Never Married       7912024 10942667
```

```
> #Example 5.4 HH Level Wealth/Total Assets
> svyby (~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, svytotal)
      I(NFINR == 1)     H11ATOTA           se
FALSE         FALSE 1.701334e+13 1.031603e+12
TRUE           TRUE 2.526686e+13 1.353710e+12

> confint(svyby (~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svytotal))
             2.5 %       97.5 %
FALSE 1.499144e+13 1.903525e+13
TRUE  2.261364e+13 2.792009e+13
```

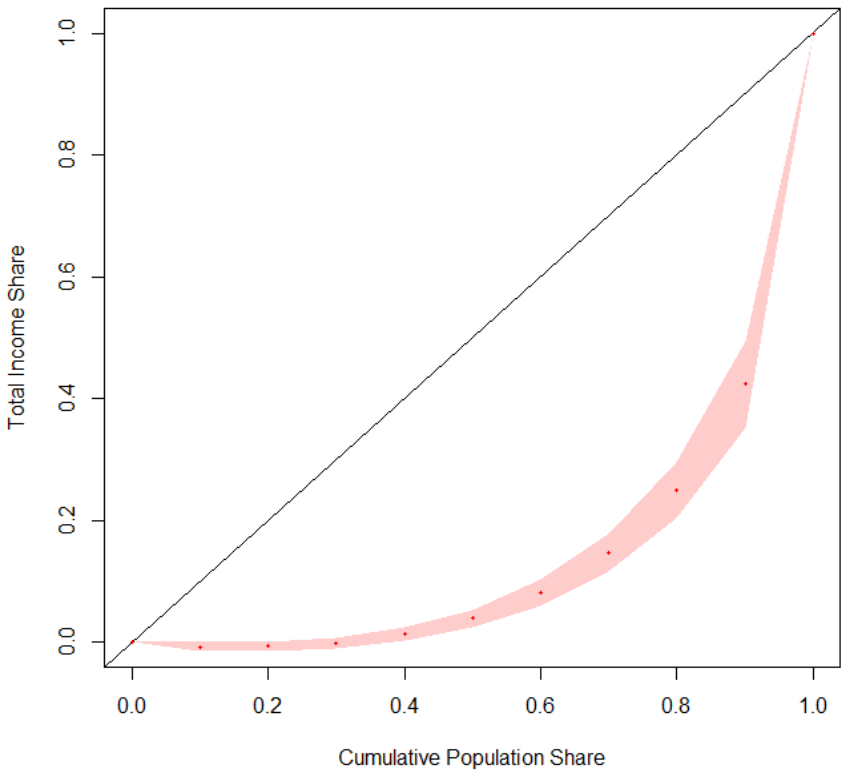
```
> #Example 5.5 HRS HH Income
> svyby (~H11ITOT, ~I(NFINR==1), hrssvyhh, na.rm=T, svymean)
      I(NFINR == 1)  H11ITOT       se
FALSE         FALSE 98737.91 3007.883
TRUE           TRUE 71382.40 1937.229
> confint(svyby (~H11ITOT, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svymean))
         2.5 %   97.5 %
FALSE 92842.57 104633.3
TRUE  67585.50  75179.3
```

```
> #Example 5.6 Mean Systolic Blood Pressure, NHANES data
> a <- svymean(~BPXSY1 , subset (nhanessvy2, age >=18), na.rm=TRUE)
> coef(a)
  BPXSY1
122.0292
> SE(a)
          BPXSY1
BPXSY1 0.6163389
> confint(a)
          2.5 %   97.5 %
BPXSY1 120.8212 123.2372
```

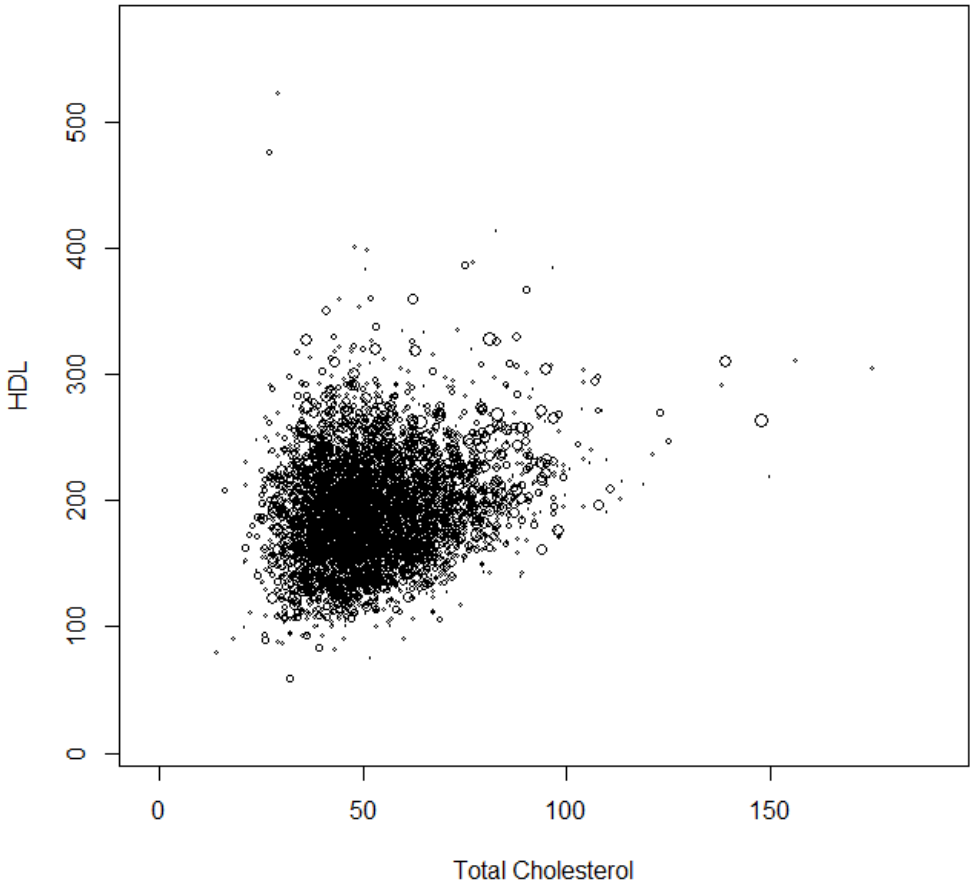
```
> #Example 5.7
> svyby (~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svymean)
      I(NFINR == 1) H11ATOTA       se
FALSE         FALSE 563269.1 26670.34
TRUE           TRUE 428470.8 17353.77
> confint(svyby(~H11ATOTA, ~I(NFINR==1), hrssvyhh, na.rm=T, ci=T, svymean))
         2.5 %   97.5 %
FALSE 510996.2 615542.0
TRUE  394458.0 462483.5
```

```
> #Example 5.8 Standard Deviation of Cholesterol NHANES data
> #Create a data object with weights only but no design variables
> nhaneswgt <- svydesign(id=~1, weights=~WTMEC2YR, data=nhanesdata)

> #Subset of those with positive weight and age 18plus
> subnhaneswgt <- subset(nhaneswgt, age >= 18 & WTMEC2YR > 0 )

> #obtain mean
> a <- svymean(~LBXTC + LBDHDD, design=subnhaneswgt, na.rm=T, deff="replace")
> a
            mean        SE   DEff
LBXTC  194.43547   0.78321 1.8880
LBDHDD  52.83826   0.29784 2.0638

> # use sqrt of variance to obtain standard deviation
> sd <- sqrt(svyvar(~LBXTC + LBDHDD, design = subnhaneswgt, na.rm=T))
> sd
       variance     SE
LBXTC    41.052 53.234
LBDHDD   14.931 11.547



> #Example 5.9 Population Percentiles for total HH Wealth HRS data, in subset of NFINR=1
> q <- svyquantile(~H11ATOTA, hrssvysub, c(.25,.5,.75), na.rm=T, ci=T)
> q
$quantiles
          0.25    0.5   0.75
H11ATOTA 22000 142000 440000

$CIs
, , H11ATOTA

        0.25    0.5     0.75
(lower 18000 127000 404550.2
upper) 26500 157000 478000.0


> SE(q)
     0.25       0.5      0.75
 2168.407  7653.202 18737.529
```

```
> #Example 5.10 Lorenz Curve and GINI coefficient not available in R Survey Package
> #Example 5.10 Lorenz Curve and GINI coefficient now available (16nov2017) in Convey Package!

# linearized design, use hrssvyhh data set for full design matrix
> hrssvyhh_c <- convey_prep(hrssvyhh)
# subset data after convey_prep command
> sub_hrssvyhh_c <- subset( hrssvyhh_c , NFINR==1)

> # run svygini and svylorenz using subset, note that negative values are allowed in R, unlike Stata
> svygini( ~H11ATOTA, design = sub_hrssvyhh_c)
            gini     SE
H11ATOTA 0.73897 0.0094

> svylorenz( ~H11ATOTA, sub_hrssvyhh_c, seq(0,1,.1), alpha = .01 )
$quantiles
         0         0.1          0.2          0.3        0.4        0.5
H11ATOTA 0 -0.00717233 -0.006560534 -0.001225929 0.01311675 0.03884905
               0.6       0.7       0.8       0.9 1
H11ATOTA 0.08118999 0.1464493 0.2498023 0.4234323 1
$CIs
, , H11ATOTA
       0           0.1           0.2          0.3         0.4         0.5
(lower 0 -0.0141887952 -0.0138754639 -0.009714108 0.002463261 0.02436226
upper) 0 -0.0001558652  0.0007543967  0.007262250 0.023770237 0.05333585
             0.6       0.7       0.8       0.9 1
(lower 0.05988412 0.1155196 0.2043382 0.3530132 1
upper) 0.10249587 0.1773789 0.2952665 0.4938513 1
```



```
> #Example 5.11 Relationship between 2 continuous variables, note this is weighted and design based
> svyplot(LBXTC~LBDHDD, subset(subnhanes, age>=18), style="bubble", ylab="HDL", xlab="Total Cholesterol")
```




```
> #EXAMPLE 5.11 Correlation between Total and High Cholesterol, NHANES DATA
> #create standardized versions of variables first, then use in regression
> nhanesdata$stdlbxtc <- (nhanesdata$LBXTC-194.4355)/41.05184
> summary(nhanesdata$LBXTC + nhanesdata$stdlbxtc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   55.7   153.0   178.6   182.9   209.4   531.0    2768

> nhanesdata$stdlbdhdd <- (nhanesdata$LBDHDD-52.83826)/14.93157
> summary(nhanesdata$stdlbxtc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-3.2990 -0.9850 -0.3760 -0.2737  0.3548  8.0040    2768
```

```
> #reset survey design and subset
> nhanessvy2 <- svydesign(strata=~SDMVSTRA, id=~SDMVPSU, weights=~WTMEC2YR, data=nhanesdata, nest=T)
> subnhanes <- subset(nhanessvy2 , age >= 18)

> #Design based linear regression to obtain correlation and correct SE
> summary(Ex5_11_svyglm <- svyglm(stdlbxtc ~ stdlbdhdd, design=subnhanes))

Call:
svyglm(formula = stdlbxtc ~ stdlbdhdd, design = subnhanes)

Survey design:
subset(nhanessvy2, age >= 18)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.916e-07  2.568e-02    0.00        1
stdlbdhdd    2.414e-01  1.104e-02   21.87  2.4e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 1.003306)
Number of Fisher Scoring iterations: 2
```

```
> #Example 5.12 Ratio Estimator for HDD to Total Cholesterol
> ex5_12 <- svyby (~LBDHDD, denominator=~LBXTC, by=~I(age >= 18), nhanessvy2, na.rm=T, ci=T, svyratio)
> confint(ex5_12)nhanesdata$stdlbxtc <- (nhanesdata$LBXTC-194.4355)/41.05184

> summary(nhanesdata$LBXTC + nhanesdata$stdlbxtc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   55.7   153.0   178.6   182.9   209.4   531.0    2768

> confint(ex5_12)
          2.5 %    97.5 %
FALSE 0.3265713 0.3374500
TRUE  0.2660798 0.2774246



> #Example 5.13 Proportions of Diabetes by Gender in Subpopulation of Age >=70
> subhrs70 <- subset(hrssvyr, NAGE >= 70)
> ex5_13 <- svyby(~diabetes, ~GENDER, subhrs70, svymean, keep.names=T, na.rm=T)

> print(ex5_13)
  GENDER  diabetes          se
1      1 0.2736113 0.007468441
2      2 0.2269743 0.008554564

> confint(ex5_13)
      2.5 %    97.5 %
1 0.2589734 0.2882491
2 0.2102077 0.2437409



> #Example 5.14 Mean Systolic Blood Pressure by Gender, Age 46+ NHANES
> subnhanes46 <-subset(nhanessvy2, age >= 46)

> #RIAGENDR 1=MALE 2=FEMALE
> ex5_14 <- svyby(~BPXSY1, ~RIAGENDR, subnhanes46, svymean, keep.names=T, na.rm=T)

> print(ex5_14)
  RIAGENDR   BPXSY1        se
1        1 128.3005 0.8687054
2        2 128.1820 0.9460163

> confint(ex5_14)
     2.5 %   97.5 %
1 126.5979 130.0032
2 126.3278 130.0361
```

```
> #Example 5.15 Differences in Mean HH Wealth by Educational Attainment, HRS data
> #CODES FOR EDCAT: 1=0-11 2=12 3=13-15 4=16+ YEARS OF EDUCATION
> ex5_15 <- svyby(~H11ATOTA, ~edcat, hrssvysub, svymean, na.rm=T, options(survey.lonely.psu="remove"))
> print(ex5_15)
  edcat H11ATOTA       se
1     1 122088.6 10595.60
2     2 259027.2  9802.47
3     3 336308.6 17201.79
4     4 834141.0 46477.79

> confint(ex5_15)
     2.5 %   97.5 %
1 101321.6 142855.6
2 239814.7 278239.6
3 302593.7 370023.5
4 743046.2 925235.8

> svycontrast(ex5_15, list(avg=c(.5,0,0,.5), diff=c(1,0,0,-1)))
     contrast    SE
avg    478115 23835
diff  -712052 47670

Warning message:
In vcov.svyby(stat) : Only diagonal elements of vcov() available



#Example 5.16 Differences in Total Wealth over Time 2010 to 2012, HRS data
#Use 2010 and 2012 data set prepared in SAS

hrs_2010_2012 <- read_sas("p:/ASDA 2/Data sets/hrs 2012/hrs 2010/hrs_2010_2012_c5.sas7bdat")
summary(hrs_2010_2012)
names(hrs_2010_2012)
hrs2010_2012 <- svydesign(strata=~STRATUM, id=~SECU, weights=~hhweight, data=hrs_2010_2012, nest=T)
subhrs2010 <- subset(hrs2010_2012, finr2010_2012==1)

> ex5_16 <- svyby (~totwealth, ~year, design=subhrs2010, keep.vars=T, svymean)
> coef(ex5_16)
    2010     2012
432829.6 437807.6
> SE(ex5_16)
[1] 16010.53 17016.29
> contrast <- svycontrast(ex5_16, list(avg=c(.5,.5), diff=c(1,-1)))
Warning message:
In vcov.svyby(stat) : Only diagonal elements of vcov() available
> print(contrast)
     contrast    SE
avg  435318.6 11682
diff  -4978.1 23364
```