```
############################################
#
# creating a population to work with...
#

pop <- data.frame(sex=c(rep("M",1000),rep("F",1000)),
                  height=c(rnorm(1000,5.5,.25),rnorm(1000,5,0.5)))

pop <- cbind(pop,weight=c(170*pop$height[1:1000]/5.5 + rnorm(1000,0,5),
                   120*pop$height[1001:2000]/5 + rnorm(1000,0,5)))

pop$height <- round(pop$height,2)

pop$weight <- round(pop$weight,0)

############################################
#
# creating a biased sample that will need
# postweights
#

samp <- pop[c(sample(1:1000,150),sample(1001:2000,50)),]

############################################
#
# simple graphical checks

par(mfrow=c(3,2))

hist(samp$height)
hist(pop$height)
hist(samp$weight)
hist(pop$weight)
plot(pop$height,pop$weight)
plot(samp$height,samp$weight)

#############################################
#
# post strat weights
#

w.men <- (1000/2000)/(150/200)
w.wom <- (1000/2000)/(50/200)

samp$w <- c(rep(w.men,150),rep(w.wom,50))

#############################################
#
#  BOXPLOTS
#

#
# METHOD 1: USE THE WEIGHTS DIRECTLY TO COMPUTE 5-NUMBER SUMMARY:

# put the sample in order by height
sorted.samp <- samp[order(samp$height),]

# convert weights to probabilities and
# compute the "sample CDF"
p <- sorted.samp$w/sum(sorted.samp$w)
cdf <- cumsum(p)

# Calculate locations of quantiles from "sample CDF"
```



```
mn <- 1              # or maybe mn <- sum(cdf<=0.10)
q1 <- sum(cdf<=0.25)
q2 <- sum(cdf<=0.50)
q3 <- sum(cdf<=0.75)
mx <- length(cdf)    # or maybe mx <- sum(cdf<=0.90)

# get 5-number summary and compare with unweighted 5-number summary
stats <- sorted.samp$height[c(mn,q1,q2,q3,mx)]

fivenum(samp$height)
stats
fivenum(pop$height)

# Compare boxplots

par(mfrow=c(1,3))

boxplot(samp$height,ylim=c(3.5,7))
title(xlab="Unweighted Sample")

bp <- list(stats=matrix(stats,ncol=1),n=200)
bxp(bp,ylim=c(3.5,7))
title(xlab="Weighted 5-Num Summary")

boxplot(pop$height,ylim=c(3.5,7))
title(xlab="Population")

# METHOD 2: RESAMPLE THE SAMPLE PROPORTIONAL TO THE WEIGHTS

# convert weights to probabilities
pw <- samp$w/sum(samp$w)

# resample a "large" sample
index <- sample(1:200,size=1000,replace=T,prob=pw)
resamp <- samp[index,]

# compare 5-number summaries

fivenum(samp$height)
fivenum(resamp$height)
fivenum(pop$height)

# compare boxplots!

par(mfrow=c(1,4))

boxplot(samp$height,ylim=c(3.5,7))
title(xlab="Unweighted Sample")

boxplot(resamp$height,ylim=c(3.5,7))
title(xlab="Weighted Re-Sample")

bp <- list(stats=matrix(stats,ncol=1),n=200)
bxp(bp,ylim=c(3.5,7))
title(xlab="Weighted 5-Num Summary")

boxplot(pop$height,ylim=c(3.5,7))
title(xlab="Population")


#############################################
#
# HISTOGRAMS
```

```
#

# EASIEST TO USE METHOD 2: RESAMPLE
#


# convert weights to probabilities
pw <- samp$w/sum(samp$w)

# resample a "large" sample
index <- sample(1:200,size=1000,replace=T,prob=pw)
resamp <- samp[index,]

# compare histograms!
par(mfrow=c(3,1))

hist(samp$height,main="",xlab="Unweighted Sample",xlim=c(3,7))
hist(resamp$height,main="",xlab="Weighted Re-Sample",xlim=c(3,7))
hist(pop$height,main="",xlab="Population",xlim=c(3,7),breaks=20)

#############################################
#
# SCATTER PLOTS
#

# EASIEST TO USE METHOD 2, BUT IN THIS CASE RE-SAMPLE SIZE SHOULD BE
# SAME AS SAMPLE SIZE

# convert weights to probabilities
pw <- samp$w/sum(samp$w)

# resample same sample size
index <- sample(1:200,size=200,replace=T,prob=pw)
resamp <- samp[index,]

# compare men and women sampled

rbind(
samp=table(samp$sex),
resamp=table(resamp$sex),
pop=table(pop$sex))

#compare scatter plots

par(mfrow=c(2,2))

plot(pop$height,pop$weight,xlab="",ylab="")
title(main="Population",xlab="Height",ylab="Weight")
plot(samp$height,samp$weight,xlab="",ylab="")
title(main="Unweighted Sample",xlab="Height",ylab="Weight")
plot(resamp$height,resamp$weight,xlab="",ylab="")
title(main="Weighted Re-Sample",xlab="Height",ylab="Weight")

# an alternative that is interesting is to plot the unweighted sample
# data, using circles whose diameter indicates the survey weight of each
# observation...

# install "plotrix" package
library(plotrix)

plot(samp$height,samp$weight,xlab="",ylab="",type="n")
title(main="Unweighted Sample,\nCircles Proport. to Survey
Weight",xlab="Height",ylab="Weight")
```



```
for (i in 1:length(samp$height)) {
  draw.circle(samp$height[i],samp$weight[i],pw[i]*10)
}

#############################################
#
# REGRESSION ANALYIS
#
#
# There are 4 methods to consider
#
# 1. ignore the weights
# 2. use the "weights" feature of lm() or any other regression
#    package
# 3. use the weights, but jackknife to get better point estimates
#    and standard errors
# 4. re-sample data proportional to the weights and do unweighted
#    regression on the re-sampled data.


# 1. ignore the weights
#
# unweighted sample analysis

u.lm <- lm(weight ~ height, data=samp)
# summary(u.lm)$coef

# 2. use the "weights" feature of lm() or any other regression
#    package
#
# weighted sample analysis

w.lm <- lm(weight ~ height, data=samp, weights=samp$w)
# summary(w.lm)$coef

# 3. use the weights, but jackknife to get better point estimates
#    and standard errors
#
# The standard errors and p-values in method #2 can't be trusted:
#
# traditional replication weights: larger weight -> greater certainty
# survey weights: larger weight -> less certainty
#
# So we instead jackknife to get a sense of how much variability
# there is in the regression coefficients...


n <- length(samp$weight)

coefs <- NULL
for (r in 1:n) {
  y.r <- samp$weight[-r]
  x.r <- samp$height[-r]
  sti.r <- samp$sex[-r]

  n.wom <- table(sti.r)[1]
  n.men <- table(sti.r)[2]
  w.men <- (1000/2000)/(n.men/199)
  w.wom <- (1000/2000)/(n.wom/199)
  w.r <- c(rep(w.men,150),rep(w.wom,50))[-r]

  coefs <- cbind(coefs,lm(y.r ~ x.r, weights=w.r)$coef)
}
```

```
b0 <- mean(coefs[1,])
b1 <- mean(coefs[2,])

s2.b0 <- (n-1)/n * (n-1) * var(coefs[1,])
s2.b1 <- (n-1)/n * (n-1) * var(coefs[2,])

jackknife <- matrix(nrow=2,c(b0,b1,sqrt(c(s2.b0,s2.b1))))
dimnames(jackknife) <- list(
                       c("(Intercept)","height"),
                       c("Estimate","Std. Error"))


# 4. re-sample data proportional to the weights and do unweighted
#    regression on the re-sampled data.
#

# convert weights to probabilities
pw <- samp$w/sum(samp$w)

# resample a "large" sample
index <- sample(1:200,size=200,replace=T,prob=pw)
resamp <- samp[index,]

r.lm <- lm(weight ~ height, data=resamp)
# summary(r.lm)$coef

# variation: average of 100 resamples:

rcoefs <- NULL
for (N in 1:100) {
  pw <- samp$w/sum(samp$w)
  index <- sample(1:200,size=200,replace=T,prob=pw)
  resamp <- samp[index,]
  rr.lm <- lm(weight ~ height, data=resamp)
  rcoefs <- cbind(rcoefs,rr.lm$coef)
}

r.b0 <- mean(rcoefs[1,])
r.b1 <- mean(rcoefs[2,])
r.s2.b0 <- var(rcoefs[1,])
r.s2.b1 <- var(rcoefs[2,])

resample <- matrix(nrow=2,c(r.b0,r.b1,sqrt(c(r.s2.b0,r.s2.b1))))
dimnames(resample) <- list(
                       c("(Intercept)","height"),
                       c("Estimate","Std. Error"))



# population analysis

p.lm <- lm(weight ~ height, data=pop)
# summary(p.lm)$coef


list(
     "Unweighted Regression"=round(summary(u.lm)$coef[,1:2],2),
     "Weighted Regression"=round(summary(w.lm)$coef[,1:2],2),
     "Jackknifed Regression"=round(jackknife,2),
     "Resampled Regression"=round(summary(r.lm)$coef[,1:2],2),
     "Mean Resamp. Regr.'s"=round(resample,2),
     "Population Regression"
```


plotting and regression with weights.r

```
=data.frame("Estimate"=(round(summary(p.lm)$coef[,1],2))))
```