Name:

Goals

You will use a simulation study to compare regression estimates that use survey weights vs. estimates that do not use weights. You will also run a simulation to study model-based regression.

To do and turn in

Work through this handout in R and answer all questions asked below. Create an R script file called "RegressionSim.r" that contains *all* commands that you used to complete this handout. You **do need** to save your R output. Be sure to remove any unnecessary code and output for full credit.

Turn in this handout with your answers to the questions below and attached a separate copy of the output and commands used to answer the questions. Place a copy of your script file in your Courses homework folder for this class. While you can consult with your classmates on this assignment you cannot work together on the same computer and use the same script file.

The finite population

The population for this study is discussed in textbook examples 11.2-11.4. The file `anthpop2.csv` contains all 3000 heights (inches) and middle finger lengths (cm) for this population, as well as the variable `pop.prob` which gives selection probabilities for taking a with-replacement unequal-probability sample as discussed in example 11.3. Recall that these selection probabilities are related to the height of the individual with smaller individuals more likely to be selected for the sample than taller individuals. I also included a variable called `f.height` that represents the height of the criminals in this population. (Note that this is a variable I created, not part of the original data collected by Macdonell.)

As done in the text, you will consider how describe the relationship between the length of the left middle finger and the height of the men in this population of 3000 criminals.

1. The population regression line

Read the population data into R and assign it the name `pop`. This is the name I use below when referring to this data frame. As done in the R Survey regression handout, fit the population regression of height (y) against finger length (x).

Question set A: What is the population regression line? Use the slope parameter value to describe the relationship between these two variables in the population.

The simulation

We will next use an R simulation to compare the behavior of the design-based slope estimate (using sampling weights) and the model-based slope estimate (ignoring sampling weights) when using a sampling design with unequal selection weights. The basic idea behind this simulation is:

1. pick sample size n
2. draw a with replacement pps sample of size n from the population using the probabilities given
3. estimate the population slope using the design-based estimate (weights) and the model-based estimate (no weights)
4. repeat steps 2-3 a large number of times ("`reps`") and compare how the two types of slope estimates behave from sample to sample

Here are these basic steps translated into R. Copy these steps into an R script file. Keep `reps` at 1 and work your way through each step in the for loop so that you understand what is being done at each step.

```
#### pps sample - design vs. model slopes for regression of height on finger
N<-3000                     # population size
n<- 200                     # sample size
reps<- 1                    # number of repetitions
#create two vectors that will store the slope estimate for each sample taken
ests.svylm<-rep(NA,reps)    # vector for slopes using weights
ests.lm<-rep(NA,reps)       # vector for slopes ignoring weights

for (i in 1:reps)
{
samp.pps<- sample(1:N,n,prob=pop$pop.prob, replace=T)     # units sampled
y<- pop$height[samp.pps]               # height of units sampled
x<- pop$finger[samp.pps]               # finger length of units sampled
p<- pop$pop.prob[samp.pps]              # selection probabilities
dsn<- svydesign(id=~1, probs=p, data=data.frame(x,y))   # sampling design
ests.svylm[i] <- svyglm(y~x, design=dsn)$coefficients[[2]] # design-based slope estimate
ests.lm[i] <- lm(y~x)$coefficients[[2]]                 # OLS model-based slope estimate
}
```

Question set B: Compare the average height in your sample with the average height in the population? How and why do they differ? Then compare the design-based slope estimate to the model-based slope estimate. How and why do they differ?

Next explore the distribution of each slope estimate for many pps samples from this population. Increase `reps` to 1000 and run the simulation (which might take a minute or so). Fitting the `lm` and `svyglm` models more computer intensive than a simple mean so you may wait around for a while if you pick a `reps` of 50000, but feel free to give it a try if you have the time!

```
# compare sampling distributions
pop.slope<- lm(height~finger, data=pop)$coefficients[[2]]
boxplot(ests.svylm, ests.lm, names=c("usings weights", "no weights"))
abline(h=pop.slope)
summary(ests.svylm)
summary(ests.lm)
pop.slope

# Bias: E(estimate) - pop.mean
mean(ests.svylm) - pop.slope    # with replacement bias
mean(ests.lm) - pop.slope # without replacement bias

# SE: standard error of estimates
sd(ests.svylm)     # with replacement: SE of sample mean
sd(ests.lm)   # without replacement: SE of sample mean

# Confidence interval coverage:  does the CI contain pop slope for 95% of all samples?
# CI = estimate +/- 2 x SE
tooLow <- length(which( ests.svylm + 2*sd(ests.svylm) < pop.slope) )
tooHigh<- length(which( ests.svylm - 2*sd(ests.svylm) > pop.slope)  )
1- (tooLow+tooHigh)/reps                # coverage rate for design-based estimate

tooLow <- length(which( ests.lm + 2*sd(ests.lm) < pop.slope))
tooHigh<- length(which( ests.lm - 2*sd(ests.lm) > pop.slope) )
1- (tooLow+tooHigh)/reps                # coverage rate for model-based estimate
```

Question set C: Compare the shape, center (i.e. bias), and spread (i.e. SE) of the design-based (weighted) and model-based (unweighted) slope estimates. Use numbers provided by the simulation in your comparison. The commands above also provide a study of whether the "usual" confidence interval (estimate plus/minus twice the SE) actually contains the population slope for 95% of all samples. What is the simulation coverage rate for each type of slope estimate? Using the results of this simulation study, which type of slope estimate would you use if your goal was to estimate the population slope and correctly assess your margin of error?

2. The population regression line - version 2

Consider a different population model that uses both finger length and father's height to describe the height of individuals in the population. Fit the population regression of height (y) against finger length (x) **and** father's height (z). To do this in R, the only change you need to make to your code from part (1) above is to add `f.height` to the linear model description:

```
lm(height ~ finger  + f.height, data=pop)
```

A similar change will be needed when specifying the model in the `svyglm` command.

Question: What is the population regression line? Use the slope parameter value associated with `finger` to describe the population relationship between height and finger length after accounting for father's height.

The simulation

Repeat the simulation study done above in part (1) for this new population regression that includes the explanatory variable `f.height`. While this new variable is added to the regression, the goal of the simulation study is to again compare design and model-based estimates of the effect (slope) of finger length on height.

Question set D: Repeat Question sets B and D using the results of this new simulation study. Has the bias of the model-based estimate improved when compared to its bias in simulation (1)? How has the mode-based confidence interval coverage rate changed? What might account for these changes in the behavior of the model-based estimate? (Hint: make sure you read textbook section 11.4)