Ch. 11 topics: Regression with complex survey data

Math 255, St. Clair

Goal:

- A regression model describes how a response y varies as a function of explanatory variables x
- Typical regression modeling goals:

1. Describe the relationship between variables.
2. Predict a response y given x
3. Determine how changes in x **cause** changes in y

Model-based regression: Math 245

- Build a theoretical "universal" model for y given x that holds across populations
- Describe a "data generating model" (DGM)
    - a stochastic model that “generates” the particular finite population of individuals
- A model comes with structural probabilistic assumptions that must be checked

Model-based regression: Math 245

- Variables:
    - Response Y
    - Covariates (predictors/explanatory) x
- Simple linear regression model: describes the **conditional probability distribution of y given x**

$$Y_i \mid x_i \sim N(\mu_i, \sigma^2) \quad \mu_i = \beta_0 + \beta_1 x_i$$

- Model assumptions:

(1) Linear relationship

(2) Constant variance

(3) Normally distributed

(4) Independence

Model-based regression: estimation

- Obtain data we believe was generated by a particular DGP
- Use **maximum likelihood** inference methods to derive parameter estimates and SE for theoretical parameters β_0, β_1, and σ
  - only based on the model assumptions, not sampling weights!
- e.g. the slope estimate:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}$$

- But estimates and their SE's are highly dependent upon model assumptions (1), (2) and (4)

Design-based regression: Math 255

- Population parameters B_0 and B_1 are the "best fit" intercept and slope for the population trend

$$y = B_0 + B_1 x$$

- "best fit" means B_0 and B_1 minimize

$$\sum_{i=1}^N (y_i - B_0 - B_1 x_i)^2$$

- e.g. the population slope is

$$B_1 = \frac{\sum_{i=1}^N x_i y_i - \frac{1}{N}\sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N}\left(\sum_{i=1}^N x_i\right)^2}$$

Design-based regression: estimation

- B_1 is just another population parameter to estimate using sampling weights
  - Model fit is not important since there is no model structure!
- e.g B_1 is just a function of population totals so we use an appropriately weighted estimate:

$$\hat{B}_1 = \frac{\sum_{i=1}^n w_i x_i y_i - \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i x_i \sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i x_i^2 - \frac{1}{\sum_{i=1}^n w_i} \left(\sum_{i=1}^n w_i x_i\right)^2}$$

- Shouldn't apply design-based parameter estimates $\hat{B}_0, \hat{B}_1$ to other finite populations.

Design-based vs model-based regression

- Can think of the finite population of y_i's as being a realization from a "universal" DGM described earlier
  - then B's should be close to β's
- If **estimates** of B_1 and β_1 differ by a lot, then this could indicate that the **model** is inadequate
  - the model doesn't fit all subpopulations well
  - sampling weights are likely accounting for some unmeasured variable that is important to the relationship between y and x
- Models can include design variables
  - use stratification variables as covariates
  - fit a mixed-effects model with random cluster effects (Math 345, Spring '20)

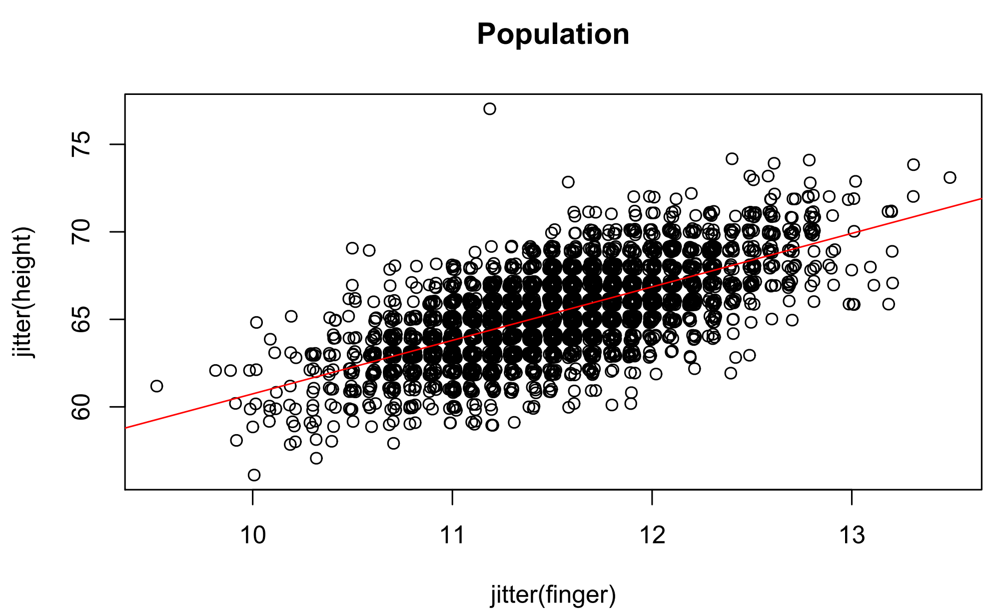
Example: The population

- `anthrop` in `SDaA`
  - A population of 3000 late 19th century criminals (anthrop.csv)
- Goal: model height as a function of finger length

```
> library(SDaA)
> pop <- anthrop  # the finite pop.
> str(pop)
'data.frame':	3000 obs. of  2 variables:
 $ finger: num  10 10.3 9.9 10.2 10.2 10.3 10.4 10.7 10 10.1 ...
 $ height: int  56 57 58 58 58 58 58 58 59 59 ...
> pop.lm <- lm(height ~ finger, data=pop)
> pop.lm

Call:
lm(formula = height ~ finger, data = pop)

Coefficients:
(Intercept)       finger
     30.179        3.056
```

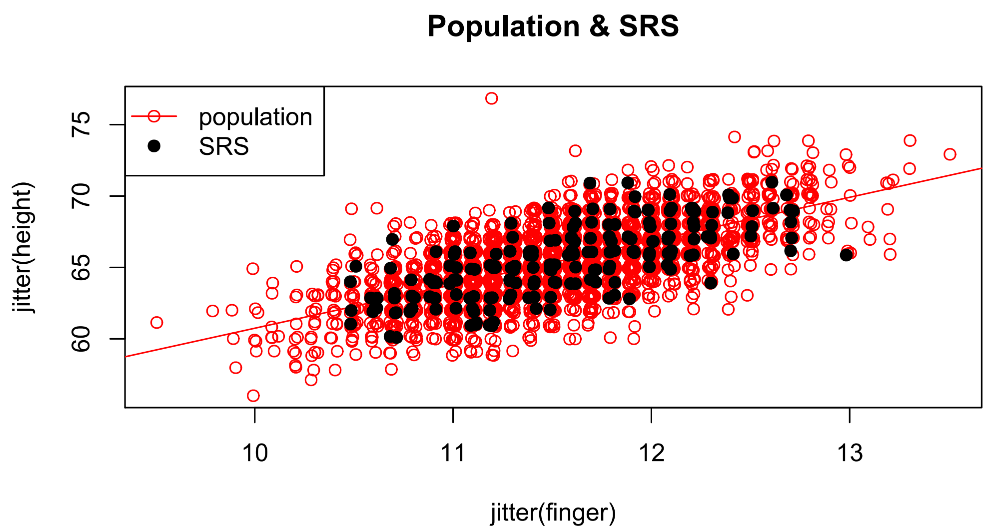

Example: The population

```
> plot(jitter(height) ~ jitter(finger), data=pop, main ="Population")
> abline(pop.lm, col="red")
```



Example: The SRS of size 200 `anthsrs`

```
> plot(jitter(height) ~ jitter(finger), data=pop, main ="Population & SRS",
> abline(pop.lm, col="red")
> points(jitter(anthsrs$finger), jitter(anthsrs$height), pch=19)
> legend("topleft",col=c("red","black"),lty=c(1,NA),pch=c(1,19),legend=c("p
```



Example: The SRS of size 200 `anthsrs`

- With an SRS, the model- and design-based estimates are the same (self-weighting).
- Model-based estimation:

```
> anthsrs.lm<- lm(height~finger, data= anthsrs)    # model-based
> summary(anthsrs.lm)

Call:
lm(formula = height ~ finger, data = anthsrs)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9045 -1.1638  0.0543  1.1407  5.0543 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  30.3162     2.5668   11.81   <2e-16 ***
finger        3.0453     0.2217   13.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.75 on 198 degrees of freedom
Multiple R-squared:  0.4879,    Adjusted R-squared:  0.4853
```

Example: The SRS of size 200 `anthsrs`

- Design-based estimation:

```
> anthsrs$N<- 3000
> anthsrs$wts<- 3000/200
> anthsrs.design<- svydesign(id= ~1, fpc= ~N, weights= ~wts, data=anthsrs)
> anthsrs.svylm<- svyglm(height ~ finger, design=anthsrs.design)
> summary(anthsrs.svylm)

Call:
svyglm(formula = height ~ finger, design = anthsrs.design)

Survey design:
svydesign(id = ~1, fpc = ~N, weights = ~wts, data = anthsrs)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.3162     2.4574   12.34   <2e-16 ***
finger        3.0453     0.2126   14.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3.046384)

Number of Fisher Scoring iterations: 2
```


Example: The SRS of size 200 `anthsrs`

- Finite population:

$$B_1 = 3.056, \ \ B_0 = 30.179$$

- Model-based slope estimate:

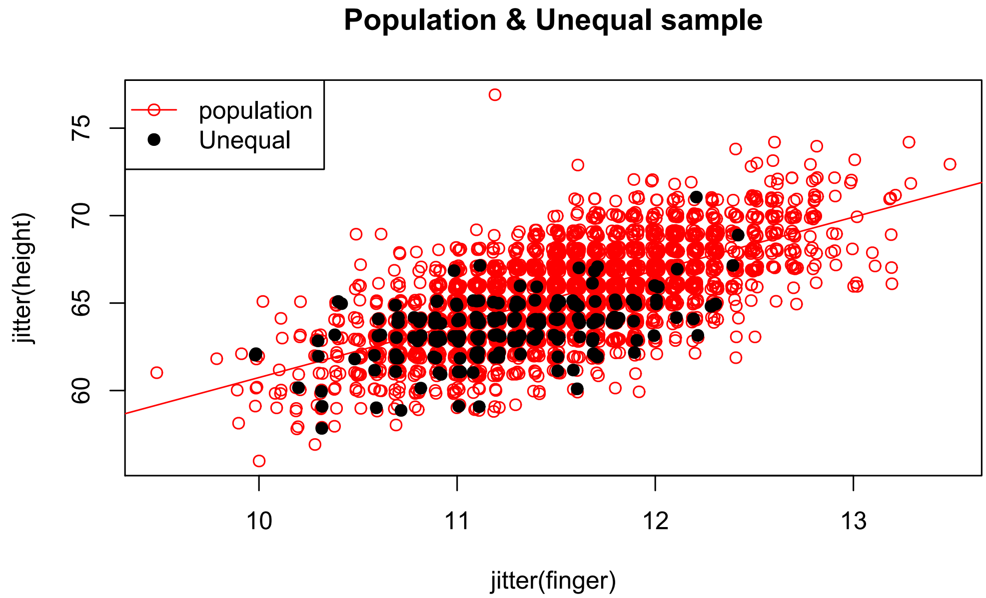
$$\hat{\beta}_1 = 3.0453(SE = 0.2217), \ \ \hat{\beta}_0 = 30.3162(SE = 2.5668)$$

- Design-based slope estimate:

$$\hat{B}_1 = 3.0453(SE = 0.2126), \ \ \hat{B}_0 = 30.3162(SE = 2.4574)$$

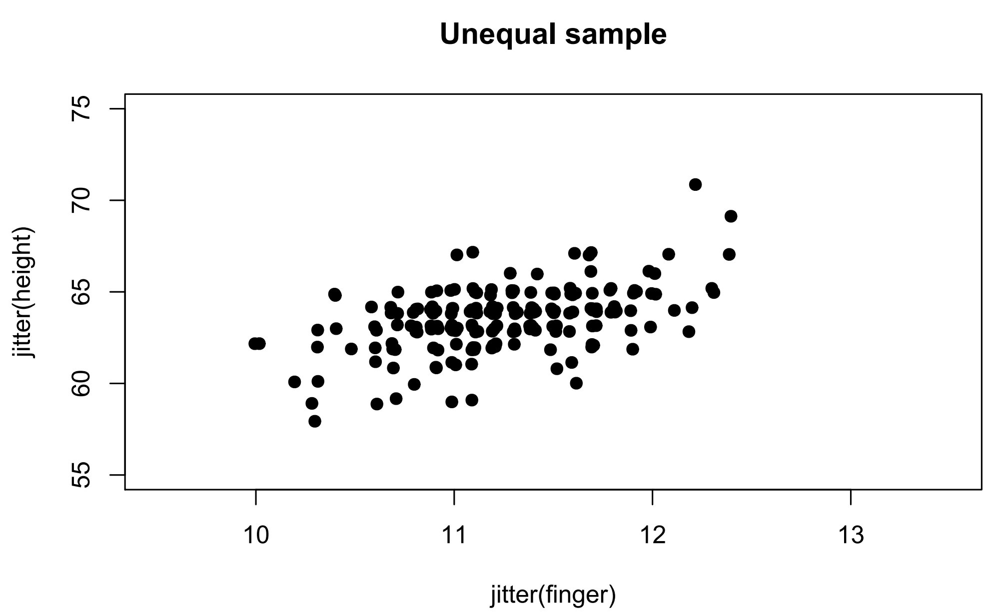
Example: unequal probability sample `anthuneq`

Shorter men have a higher inclusion probability



Example: unequal probability sample `anthuneq`

But we can't see the fact that shorter men are overrepresented in the usual data scatterplot



Example: unequal probability sample `anthuneq`

- `svyplot`: circle size is proportional to sampling weight

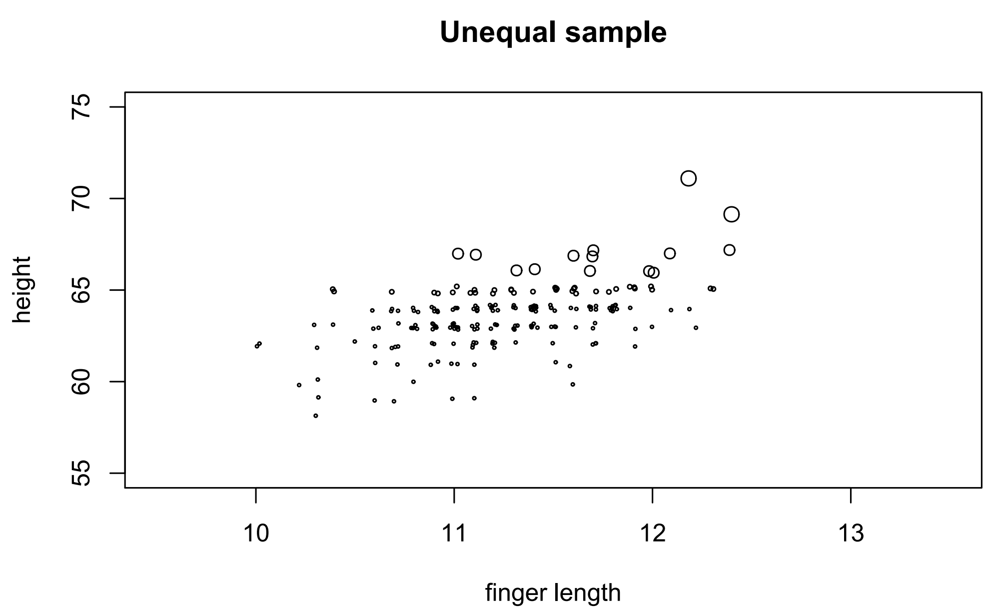
```
> anthuneq.design <- svydesign(id=~1, probs= ~prob, data= anthuneq)
> svyplot(jitter(height) ~ jitter(finger), anthuneq.design)
```

- `svyplot`: `style="hex"` uses hexagonal binning that sums weights by bin groups

```
> svyplot(jitter(height) ~ jitter(finger), anthuneq.design, style="hex")
```

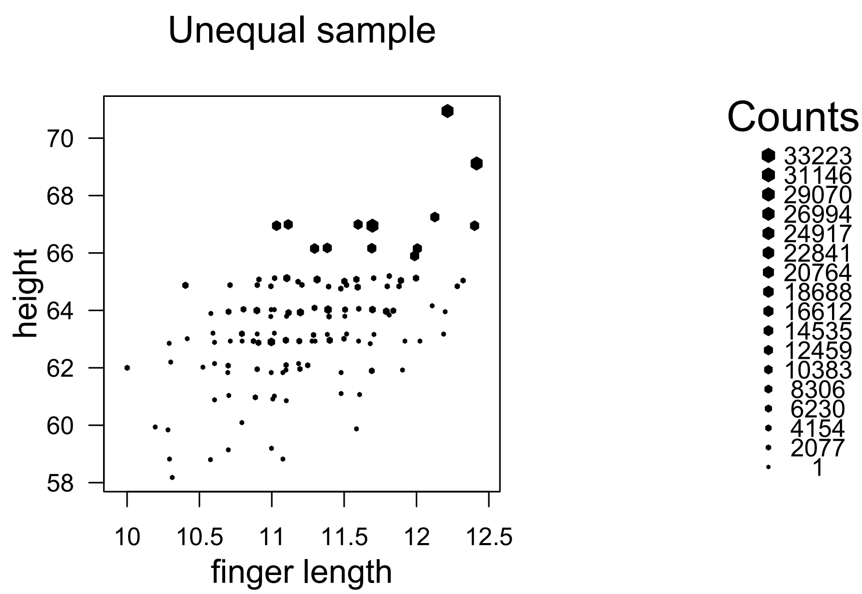

Example: unequal probability sample `anthuneq`

`svyplot`: circle size is proportional to sampling weight



Example: unequal probability sample `anthuneq`

`svyplot`: `hex` style (visually better for larger data sets)



Example: unequal probability sample `anthuneq`

- Model-based estimation:

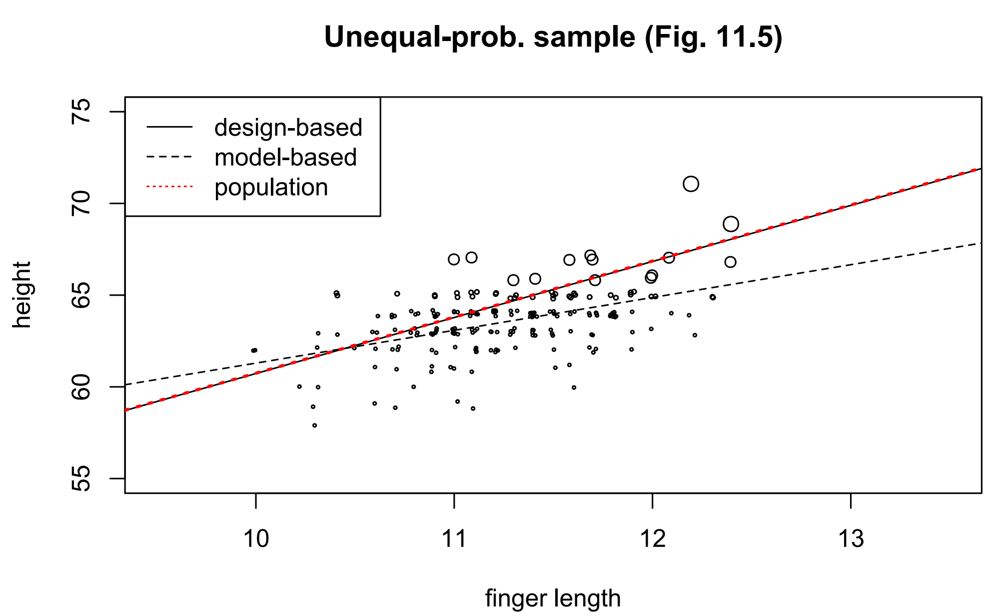
```
> anthuneq.lm<- lm(height~finger, data= anthuneq)   # model-based
> summary(anthuneq.lm)

Call:
lm(formula = height ~ finger, data = anthuneq)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2612 -0.7978  0.0965  0.9177  5.7714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.4079     2.5481  17.036  < 2e-16 ***
finger        1.7886     0.2263   7.902 1.87e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.518 on 198 degrees of freedom
Multiple R-squared:  0.2398,    Adjusted R-squared:  0.2359
F-statistic: 62.44 on 1 and 198 DF,  p-value: 1.866e-13
```

Example: unequal probability sample `anthuneq`

- Design-based estimation:

```
> anthuneq.svylm<- svyglm(height ~ finger, design=anthuneq.design)
> summary(anthuneq.svylm)

Call:
svyglm(formula = height ~ finger, design = anthuneq.design)

Survey design:
svydesign(id = ~1, probs = ~prob, data = anthuneq)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.1753     6.6284   4.552 9.25e-06 ***
finger        3.0550     0.5883   5.193 5.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3.475581)

Number of Fisher Scoring iterations: 2
```


Example: unequal probability sample `anthuneq`



Example: unequal probability sample `anthuneq`

- Finite population:

$$B_1 = 3.056, \quad B_0 = 30.179$$

- Model-based slope estimate:

$$\hat{\beta}_1 = 1.7886(SE = 0.2263), \quad \hat{\beta}_0 = 43.4079(SE = 2.5481)$$

- Design-based slope estimate:

$$\hat{B}_1 = 3.0550(SE = 0.5883), \quad \hat{B}_0 = 30.1753(SE = 6.6284)$$

- Inference about the *population* of all criminals is not estimated correctly by the model-based solution!