Lecture 7 - Key

Example of Ratio Estimation: A manager at a mill wants to estimate the total weight of tree mass (t_y) for a certain number of truckloads of 5-foot bundles of pulpwood. The process begins by

1. Weighing the total amount of pulpwood (t_x).

2. Randomly selecting a sample of n bundles from the trucks.

3. Recording the weight of the pulpwood (x_i) from each of the n bundles.

4. Removing the bark and drying the wood from the n bundles.

5. Recording the weight of the dry wood (y_i) for each of the n bundles.

Suppose a sample of $n = 30$ bundles was taken from a total of $N = 800$ bundles. The unit of measurement is pounds. Here are the summary statistics:

$$\sum_{i=1}^{30} x_i = 3316 \qquad \sum_{i=1}^{30} y_i = 1802 \qquad \sum_{i=1}^{30} x_i y_i = 214,738$$

$$\sum_{i=1}^{30} x_i^2 = 392,440 \qquad \sum_{i=1}^{30} y_i^2 = 118,360 \qquad t_x = \sum_{i=1}^{800} x_i = 89,240$$

Use ratio estimation to estimate the total amount of dry wood (t_y) and the mean amount of dry wood per bundle ($\bar{y}_U$).

```
N <- 800
n <- 30
xbar <- 3316 / n
ybar <- 1802 / n
t_x <- 89240
xbar_u <- t_x / N
```

```
t_y <- N * (ybar / xbar) * xbar_u
ybar_r <- (ybar / xbar) * xbar_u
```

The point estimates for t_y and $\bar{y}_u$ are 4.8495×10^4 and 60.6.

Also calculate the standard errors of these estimates. Recall

$$V(\hat{\bar{y}}_r) = \left(\frac{N-n}{N}\right)\left(\frac{\bar{x}_U}{\bar{x}}\right)^2 \frac{s_e^2}{n}$$

and

$$V(\hat{t}_{yr}) = \left(\frac{N-n}{N}\right)\left(\frac{t_x}{\bar{x}}\right)^2 \frac{s_e^2}{n}$$

```
B_hat <- (ybar / xbar)
y_sq <- 118360
x_sq <- 392440
xy <- 214738

s2_e <- (1 / (n-1)) * (y_sq + B_hat^2 * x_sq - 2 * B_hat * xy)
var_ybar <- ((N - n) / N) * (xbar_u / xbar) ^ 2 * s2_e / n
se_ybar <- sqrt(var_ybar)

var_t <- ((N - n) / N) * (t_x / xbar) ^ 2 * s2_e / n
se_t <- sqrt(var_t)
```

The standard errors for the estimates of t_y and $\bar{y}_u$ are 789 and 0.99.

Confidence Intervals for B, $\bar{y}_r$, and t_y

For large samples, approximate $100(1-\alpha)\%$ confidence intervals for B, $\bar{y}_U$, and t_y are:

$$\hat{B} \pm z^* \sqrt{\hat{V}(\hat{B})} \qquad \hat{\bar{y}}_r \pm z^* \sqrt{\hat{V}(\hat{\bar{y}}_r)} \qquad \hat{t}_{yr} \pm z^* \sqrt{\hat{V}(\hat{t}_{yr})} \tag{1}$$

For smaller samples, use a t distribution with $n-1$ samples.

SRS or ratio estimation, which is better?

Recall the *population correlation coefficient* of x and y is:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}. \tag{2}$$

This can be rewritten as:

$$R = \frac{S_{xy}}{S_x S_Y}, \tag{3}$$

where $S_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{N-1}$.

It can be shown that approximations for the true population variances and MSEs of $\hat{t}_{yr}$ and $\hat{\bar{y}}_r$ are:

$$MSE(\hat{t}_{yr}) \approx V(\hat{t}_{yr}) \approx \frac{N(N-n)}{n}(S_y^2 - 2BRS_yS_x + B^2S_x^2) \tag{4}$$

$$MSE(\hat{\bar{y}}_r) \approx V(\hat{\bar{y}}_r) \approx \frac{N-n}{Nn}(S_y^2 - 2BRS_yS_x + B^2S_x^2) \tag{5}$$

Thus, the variances will be smaller as R *is stronger (approaches 1).*

If the researcher wants to estimate t_y or $\bar{y}_U$, the main sampling question is *'When is it worth the additional effort and expense to collect information about X instead of just using a SRS estimator $\hat{\bar{y}}_U$ or $\hat{t}$ which does not require knowledge about X?*

The answer requires looking at the coefficient of variation for both X and Y.

$$CV(\bar{x}) = \frac{\sqrt{V(\bar{x})}}{E(\bar{x})} = \sqrt{\frac{N-n}{N}}\frac{S_x}{\sqrt{n}} \times \frac{1}{\bar{x}_U} \qquad \text{and} \qquad CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \sqrt{\frac{N-n}{N}}\frac{S_y}{\sqrt{n}} \times \frac{1}{\bar{y}_U}$$

It can be shown that if $R > \frac{CV(\bar{x})}{CV(\bar{y})}$, *then the variance of the ratio estimator is smaller than the variance of the SRS estimator.*

Because $CV(\bar{x})$ and $CV(\bar{y})$ are unknown, we would calculate the sample (Pearson) correlation coefficient R and the sample coefficients of variation $\hat{CV}(\bar{x})$ and $\hat{CV}(\bar{y})$) to check if these conditions are met. The formulas are:

$$\hat{CV}(\bar{x}) = \sqrt{\frac{N-n}{N}}\frac{s_x}{\sqrt{n}} \times \frac{1}{\bar{x}} \qquad \text{and} \qquad \hat{CV}(\bar{y}) = \sqrt{\frac{N-n}{N}}\frac{s_y}{\sqrt{n}} \times \frac{1}{\bar{y}}$$

and

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n} x_i y_i - 1/n(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{(n-1)s_x s_y}$$

In summary, if the following conditions hold, using ratio estimators can provide a substantial improvement over the SRS estimators:

1. You must be able to simultaneously observe X and Y values that are 'roughly proportional' to each other. *That is, there is a strong positive linear relationship between Y and X that passes through the origin (zero intercept).*

2. The coefficient of variantion for X should not be substantially larger than the coefficient of variation of Y.

3. The population total t_x or population mean $\bar{x}_U$ should be known.

If there is a linear relationship between Y and X and the intercept is not zero or the correlation between X and Y is negative, *then a regression estimator should be considered.*

Note, ratio estimation can be applied to proportions as well.

Domain Estimation

It is common to want estimates of a mean or total for subpopulations. *The subpopulations are called domains.*

For example, we may be interested in salary of MSU graduates by major.

Let U_d be the set of population units in domain d and let N_d be the number of population units in domain d. Then the domain total and domain mean for domain d are:

$$t_{yd} = \sum_{i \in U_d} y_i \qquad\qquad \bar{y}_{Ud} = t_{yd} = \left(\sum_{i \in U_d} y_i \right) / N_d$$

Let S_d be the set of *sample units* in domain d and let n_d be the number of sample units in domain d. Natural estimators for the domain mean $\bar{y}_{Ud}$ and the domain total t_{yd} are

$$* \hat{\bar{y}}_{Ud} = \left(\sum_{i \in S_d} y_i \right) / n_d = \bar{y}_d * \qquad\qquad *\hat{t}_{yd} = \frac{N_d}{n_d} \left(\sum_{i \in S_d} y_i \right) = N_d \bar{y}_d * \tag{6}$$

$\bar{y}_d$ 'looks like' the estimator $\hat{\bar{y}}_U = \bar{y}$ for a SRS of size n_d, and $N_d \bar{y}_d$ looks like the estimator $\hat{t} = N\bar{y}$ for a SRS of size n_d. This suggests that we should be able to apply the variance formulas for SRS from earlier in class. But... we cannot, Why?

For the SRS settings earlier in the course, the sample size n was fixed. *For a domain, the sample size n_d is not fixed, but is a random variable. That is, if we took different random samples of size n, we would get different values for n_d.*

Because n_d is a random variable, we cannot use the SRS variance formulas from Section 2. To find the variance, $V(\bar{y}_D)$, we need to see that $\bar{y}_D$ is a ratio estimator.

Define

$$u_i = \begin{cases} 1 \text{ if } i \in U_d \\ 0 \text{ if } i \notin U_d \end{cases}$$

and

$$x_i = \begin{cases} 1 \text{ if } i \in U_d \\ 0 \text{ if } i \notin U_d \end{cases}$$

for $i = 1, 2, \ldots, N$.

Then:

$$\bar{x}_{Ud} = \left(\sum_{i=1}^{N} x_i \right) / N = \frac{\sum_{i \in U_d} x_i + \sum_{i \notin U_d} x_i}{N} = \frac{\sum_{i \in U_d} 1 + \sum_{i \notin U_D} 0}{n} = \frac{N_d}{N} \tag{7}$$

Define $B_d = \frac{\bar{u}_{Ud}}{\bar{x}_{Ud}}$. Then

$$\begin{aligned} B_d &= \frac{\bar{u}_{Ud}}{\bar{x}_{Ud}} = \frac{\sum_{i=1}^{N} u_i / N}{\sum_{i=1}^{N} y_i / N} = \frac{\sum_{i=1}^{N} u_i}{\sum_{i=1}^{N} y_i} \\ &= \frac{\sum_{i \in S_d} u_i + \sum_{i \notin S_d} u_i}{\sum_{i \in S_d} x_i + \sum_{i \notin S_d} x_i} \\ &= \frac{\sum_{i \in S_d} y_i + \sum_{i \notin S_d} 0}{\sum_{i \in S_d} 1 + \sum_{i \notin S_d} 0} \\ &= \frac{\sum_{i \in S_d} y_i}{\sum_{i \in S_d} 1} = \frac{\sum_{i \in S_d} y_i}{n_d} = \bar{y}_d \end{aligned}$$

$\hat{B}_d = \bar{y}_d$ is the ratio estimator of $\bar{y}_{Ud}$. That is $\hat{\bar{y}}_{Ud} = \bar{y}_d$. We now use the ratio estimation variance formula:

$$\hat{V}(\bar{y}_d) = \hat{V}(B_d) = \left(\frac{N-n}{N \bar{x}_{Ud}^2} \right) \frac{s_u^2}{n} = \left(\frac{N-n}{N(N_d/N)^2} \right) \frac{s_u^2}{n} = \left(\frac{N-n}{N} \right) \left(\frac{N}{N_d} \right) \frac{s_U^2}{n,} \tag{8}$$

where

$$s_U^2 = \frac{1}{N-1} \sum_{i \in S_d} (u_i - \hat{B}_d x_i)^2$$

Regression Estimation

Assume a SRS of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is selected from a population of N pairs of (x, y) data. The goal of regression estimation is to *take advantage of a linear relationship between x and y to improve estimation of the t_y or $\bar{y}_U$.*

Unlike ratio estimation, there is no assumption of a zero intercept in the linear relationship of a positive slope. We assume a linear form for the relationship between y and x:

$$* \, y_i = B_0 + B_1 x_i + \epsilon_i, * \tag{9}$$

for intercept B_0, slope B_1, and ϵ_i is the deviation between y_i and $B_0 + B_1 x_i$.

We assume that

1. the mean of the $\epsilon_i' s$ *is zero and*

2. the $\epsilon_i' s$ are uncorrelated with the $x_i' s$. *This implies that there is not systematic relationship beween the $\epsilon_i' s$ and the $x_i' s$.*

(Supplementary Note) Note that there is no distributional assumption on the $\epsilon_i' s$. This is because the uncertainty (randomness) in this setting is generated by the sampling scheme. An alternative would be model based estimation, which requires assuming a generative distribution (on the $\epsilon_i' s$ in this setting) to drive the randomness of the estimation.

Estimating $\bar{y}_U$ and t_y

To estimate $\bar{y}_U$, we first must get estimates $\hat{B}_0$ and $\hat{B}_1$ of the true intercept B_0 and slope B_1. We use the least squares estimates:

$$* \, \hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x} * \qquad\qquad \hat{B}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{10}$$

When $\bar{x}_U$ is known, our estimate $\hat{\bar{y}}_{reg}$ of the population mean is:

$$* \hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_U * \tag{11}$$

When $\bar{x}_U$ is *unknown*, replace $\bar{x}_U$ with $\bar{x}$, then $\hat{\bar{y}}_{reg} = \bar{y}$.

The estimated variance of $\hat{\bar{y}}_{reg}$ is

$$\hat{V}(\hat{\bar{y}}_{reg}) = \frac{N-n}{N}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2 /(n-2) \tag{12}$$

Alternatively the residual component in the above variance can be reformulated as:

$$\sum_{i=1}^{n}(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y} - \hat{B}_1^2\left(\sum_{i=1} x_i^2 - n\bar{x}^2\right) \tag{13}$$

An approximate $100(1-\alpha)$ confidence interval for $\bar{y}_U$ is $\hat{\bar{y}}_{reg} \pm t^* \sqrt{\hat{V}(\hat{\bar{y}}_{reg})}$, where t^* is the upper $\alpha/2$ critical value from a $t-$distribution having $n-2$ degrees of freedom.

By multiplying $\hat{\bar{y}}_{reg}$ by N, an estimator of t_{reg} of the population total t_y is:

$$\begin{aligned} \hat{t}_{reg} &= N\hat{\bar{y}}_{reg} = N(\hat{B}_0 + \hat{B}_1\bar{x}_U) = N(\bar{y} - \hat{B}_1\bar{x} + \hat{B}_1\bar{x}_U) \\ &= N\bar{y} + \hat{B}_1(N\bar{x}_U - n\bar{x}) = N\bar{y} + \hat{B}_1(t_x - N\bar{x}) \end{aligned}$$

This formulation shows that the regression estimator is an adjustment on $N\bar{y}$ based on *the difference between* $\bar{x}$ *and* $\bar{x}_U$.

Multiplying the variance of $\hat{\bar{y}}_{reg}$ by N^2 provides the estimated variance of $\hat{t}_{reg}$:

$$\hat{V}(\hat{t}_{reg} = \frac{N(N-n)}{n(n-2)}\sum_{i=1}^{n}\left(y_i - \hat{B}_0 - \hat{B}_1 x_i\right)^2 \tag{14}$$

An approximate $100(1-\alpha)$ confidence interval for t_y is $\hat{t}_{reg} \pm t^* \sqrt{\hat{V}(\hat{t}_{reg}}$, where t^* is the upper $\alpha/2$ critical value from a $t-$distribution having $n-2$ degrees of freedom.

Note that $\sum_{i=1}^{n}(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2$ is the sum of squared residuals from a simple linear regression model. That is, $\sum_{i=1}^{n}(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2 = SSE$ for the least squares regression line.

Therefore, we could fit a regression model using R and substitute the value of SSE into the variance formulas.

```
set.seed(10262019)
x <- runif(100,-10,10)
B0 = 2
B1 <- 1
y <- B0 + B1*x + rnorm(100)
lm.tmp <- lm(y~x)
anova(lm(y~x))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 3924.8  3924.8  4174.8 < 2.2e-16 ***
## Residuals 98   92.1     0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Example: The Florida Game and Freshwater Fish Commission is interested in estimating the weights of alligators using length measurements which are easier to observe. The population size N is unknown but is large enough so that ignoring the FPC will have a negligible effect on estimation. A random sample of 22 alligators yield the following weight (in pounds) and length (in inches):

Alligator	Length	Weight	Alligator	Length	Weight
1	94	130	12	86	83
2	74	51	13	88	70
3	82	80	14	72	61
4	58	28	15	74	57
5	86	80	16	61	44
6	94	110	17	90	106
7	63	33	18	89	84
8	86	90	19	68	39
9	69	36	20	76	42
10	72	38	21	78	57
11	85	84	22	90	102

Because it is much easier to collect data on alligator length, there is a lot of available data on length. Assume that the available data indicates that the mean alligator length $\bar{x}_U \approx 90$ inches. Estimate the mean alligator weight $\bar{y}_U$ using the regression estimator.

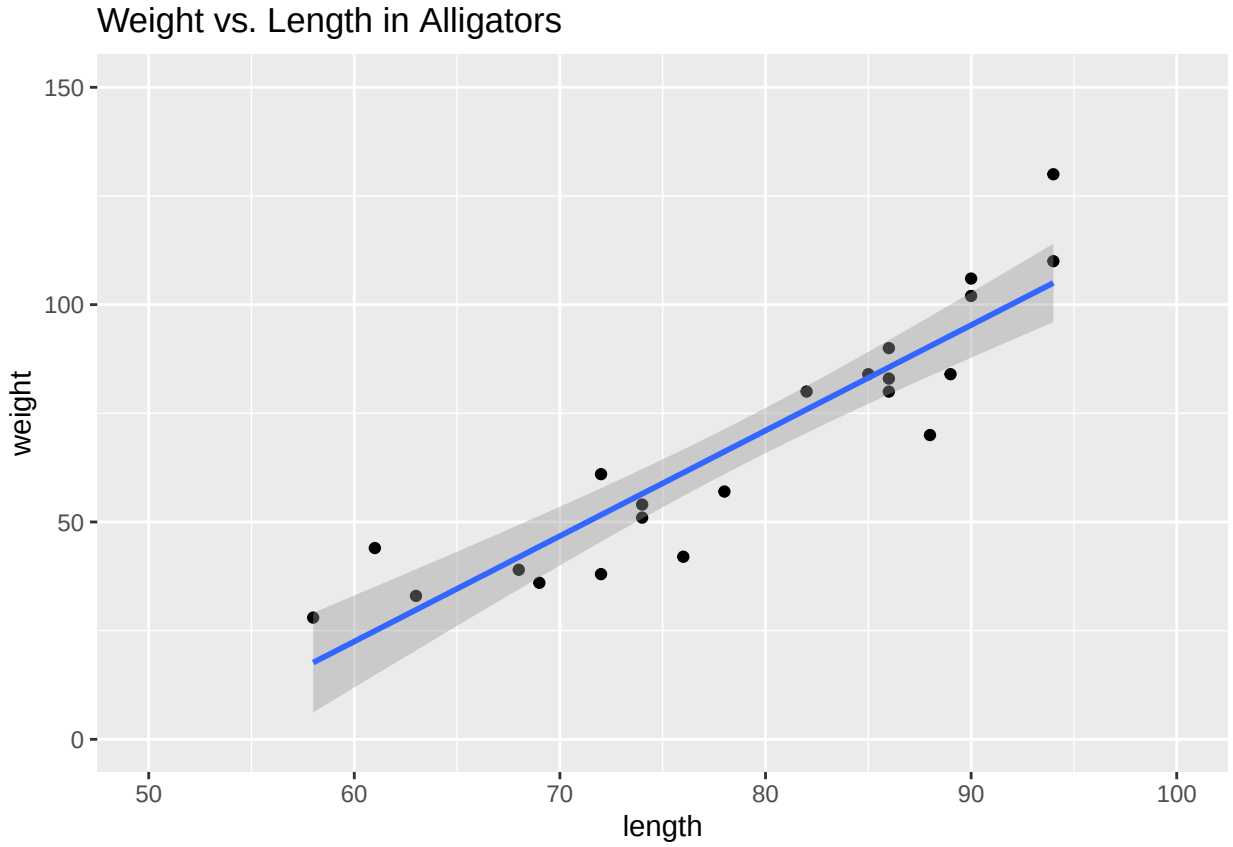
```
alligator_length <- c(94,74,82,58,86,94,63,86,69,72,85,86,88,72,74,61,90,89,68,76,78,90)
alligator_weight <- c(130,51,80,28,80,110,33,90,36,38,84,83,70,61,54,44,106,84,39,42,57,102)

lm_gator <- lm(alligator_weight~alligator_length)
y_bar <- coef(lm_gator)[1] + coef(lm_gator)[2] * 90
```

The estimated weight of an alligator, $\bar{y}_u = 95.29$.

```r
summary(lm_gator)
```

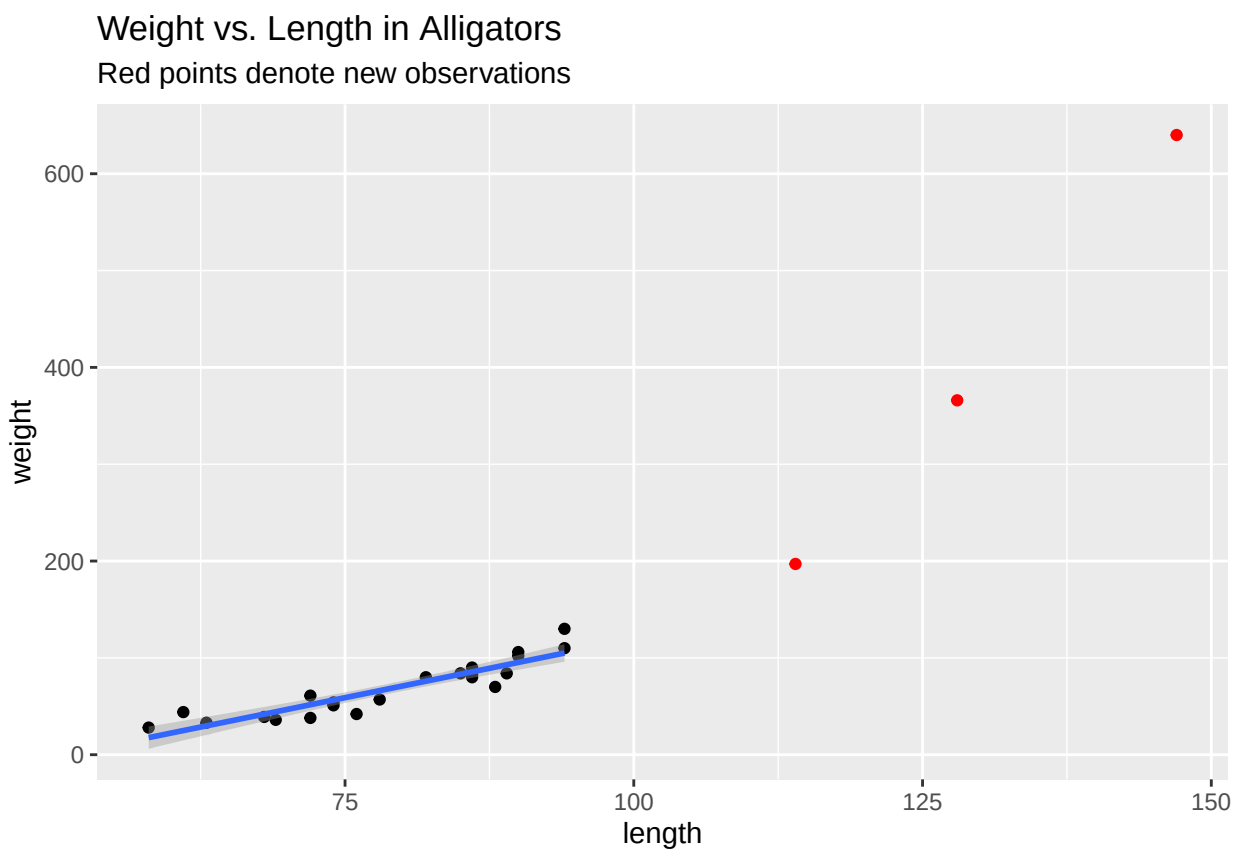
```
##
## Call:
## lm(formula = alligator_weight ~ alligator_length)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4402  -7.6524  -0.8167   6.2809  25.0020
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -123.0735    18.6490  -6.599 1.99e-06 ***
## alligator_length    2.4263     0.2344  10.352 1.76e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.64 on 20 degrees of freedom
## Multiple R-squared:  0.8427, Adjusted R-squared:  0.8349
## F-statistic: 107.2 on 1 and 20 DF,  p-value: 1.762e-09
```

```r
tibble(length = alligator_length, weight = alligator_weight) %>% ggplot(aes(y = weight, x = length)) +
geom_point() +geom_smooth(method = "lm") + xlim(50, 100) + ylim(0,150) +
  ggtitle('Weight vs. Length in Alligators')
```



Now assume you are given three new measurements

```
new_points <- tibble(length = c(147, 128, 114), weight = c(640, 366, 197))

tibble(length = alligator_length, weight = alligator_weight) %>% ggplot(aes(y = weight, x = length)) +
geom_point() +geom_smooth(method = "lm")  +
  ggtitle('Weight vs. Length in Alligators', subtitle = 'Red points denote new observations') +
  geom_point(data = new_points,inherit.aes = F, aes(y = weight, x = length ), color ='red')
```



Do you have any concerns about the previous model?

If so, what do you propose doing?

(Multiple) Regression Estimation

The simple (univariate) regression estimators can easily be extended to include k covariates.

Case 1: There exist k different regression variables $x_1, x_2, \ldots, x_k$ with the model

$$* y = B_0 + B_1 x_1 + \ldots B_k x_k + \epsilon. * \tag{15}$$

Case 2: A k^{th} order polynomial is formed for one regression variable x with the model

$$* y = B_0 + B_1 x + B_2 x^2 + \ldots B_k x^k + \epsilon * \tag{16}$$

For Case 1:

1. Find the least-squares estimates $(\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_k)$. *Note we will do this in R.* This produces the prediction model:
$$* \hat{y} = \hat{B}_0 + \hat{B}_1 x_1, \ldots, \hat{B}_k x_k, * \tag{17}$$
for some set of x values $(x_1, \ldots, x_k)$.

2. To estimate $\bar{y}_U$, we make predictions using the mean of each of our covariates
$$* \hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_{1U}, \ldots, \hat{B}_k \bar{x}_{kU}. * \tag{18}$$

For Case 2:

1. Again find the least-squares estimates of $(\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_k)$ to compute the prediction model:
$$* \hat{y} = \hat{B}_0 + \hat{B}_1 x + \ldots \hat{B}_k x^k * \tag{19}$$

2. To estimate the mean value, $\bar{y}_U$, use the mean for x
$$* \hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_U + \ldots \hat{B}_k \bar{x}_U^k * \tag{20}$$

In each case, the SSE (sum of squares of residuals) from regression output can be used to estimate the variance of $\hat{\bar{y}}_{reg}$:

$$\hat{V}(\hat{\bar{y}}_{reg}) = \frac{N-n}{Nn(n-k-1)} SSE = \frac{N-n}{Nn} MSE,$$

where $MSE = SSE/(n-k-1)$ is the mean squared error from the regression model with $n-k-1$ degrees of freedom for the error term.

Example: Using the complete data from the Florida Game and Freshwater Fish Commision

1. Fit a quadratic regression model $y = B_0 + B_1 x + B_2 x^2 + \epsilon$.

2. Estimate the mean alligator weight $\bar{y}_U$ using the multiple regression estimator $\hat{\bar{y}}_{reg}$.

3. Find a 95% confidence interval for $\bar{y}_U$.

```r
alligator_length <- c(94,74,82,58,86,94,63,86,69,72,85,86,88,72,74,61,90,89,68,76,78,90,147,128,114)
alligator_weight <- c(130,51,80,28,80,110,33,90,36,38,84,83,70,61,54,44,106,84,39,42,57,102,640,366,197)
n <- 25

# Quadratic Regression Model
alligator_length_sq <- alligator_length ^2
lm_quad <- lm(alligator_weight ~ alligator_length + alligator_length_sq)
summary(lm_quad)
```

```
##
## Call:
## lm(formula = alligator_weight ~ alligator_length + alligator_length_sq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.939  -6.635   2.217   9.711  21.521
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         410.484123  51.868027   7.914 7.06e-08 ***
## alligator_length    -11.317553   1.083534 -10.445 5.43e-10 ***
## alligator_length_sq   0.086616   0.005393  16.060 1.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.48 on 22 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.986
## F-statistic: 843.1 on 2 and 22 DF,  p-value: < 2.2e-16
```

```r
# Estimate Mean Weight (length = 90 inches)
new_dat <- data.frame(alligator_length = 90, alligator_length_sq = 90^2)
y_pred <- predict(lm_quad,newdata=new_dat)
```

Thus the mean weight of alligators, given the mean length is 90 inches is 93.5 pounds.

```r
# Confidence Intervals
#Note N is unknown but large
k <- 2
SSE <- sum(lm_quad$residuals^2)
Var_y <- SSE / (n* (n-k-1))

ci <- c(y_pred - qt(.975,n-2) * sqrt(Var_y), y_pred + qt(.975,n-2) * sqrt(Var_y))
```

Thus a 95% confidence interval for the mean alligator length is 87.08, 99.9.

SRS or Regression Estimation?

When does regression estimation provide better estimates of $\bar{y}_U$ or t_y than the SRS estimator?

Recall (??) that for least squares regression we can decompose the total sum of squares as: $SS(Total) = SS(Regression) + SSE$.

- If x and y are strongly correlated, most of the variability in the y values *can be explained by the regression line and there will be very little random variability about the regression line. That is MS(regression) is large relative to MSE. (The same idea holds in a multiple regression setting).*

- If x and y are weakly correlated, very little of the variability in the y values *can be explained by the regression and most of the variability is random variability about the regression line. That is, MS(regression) is small relative to MSE.*

Thus, for a moderate to strong relationship between the $x_i's$ and y, *the regression estimate is recommended over SRS estimation.*

Sample Size Estimation for Ratio and Regression Estimation

To determine the sample size formulas for ratio and regression estimation, the same approach as SRS will be used.

1. Specify a maximum allowable difference d *for the parameter we want to estimate. This is equivalent to stating the largest margin of error the researcher would want for a confidence interval.*

2. Specify α (where $100(1-\alpha)\%$ is the confidence level for the confidence interval).

3. Specify a prior estimate of a variance $V(\hat{\theta})$ where $\hat{\theta}$ is the the estimator for θ. θ could be $\bar{y}_U$ or t_y or population ratio $R = \bar{y}_U/\bar{x}_U$ (in ratio estimation).

4. Set the margin of error formula equal to d and solve for n.

Sample Size Determination for Ratio Estimation:

From previous equations, the margin of error formulas for the parameters B, $\bar{y}_U$ and t_y using the variance S_e^2 are:

$$\text{For } B: z_{\alpha/2}\sqrt{\hat{V}(B)} = z_{\alpha/2}\sqrt{\left(\frac{N-n}{N\bar{x}_U^2}\right)\frac{S_e^2}{n}} \quad (21)$$

$$\text{For } \bar{y}_U: z_{\alpha/2}\sqrt{\hat{V}(\hat{\bar{y}}_r)} = z_{\alpha/2}\sqrt{\left(\frac{N-n}{N}\right)\frac{S_e^2}{n}} \quad (22)$$

$$\text{For } \bar{t}_y: z_{\alpha/2}\sqrt{\hat{V}(\hat{t}_r)} = z_{\alpha/2}\sqrt{(N(N-n)\frac{S_e^2}{n}} \quad (23)$$

Note for the second two equations, we assume $\bar{x}$ and $\bar{x}_U$ are the same, that is we'd anticipate our sample would have the same mean as the population. The sample size calculations are computed by setting the margin of error (d) and solving for n.

- For B: $n = \frac{1}{\frac{1}{n_0}+\frac{1}{N}}$, where $n_0 = \left(\frac{z_{\alpha/2}S_e}{\bar{x}_U d}\right)^2$. It is assumed you know $\bar{x}_U$, if you do not an estimate needs to be provided.

- For $\bar{y}_U$: $n = \frac{1}{\frac{1}{n_0}+\frac{1}{N}}$, where $n_0 = \left(\frac{z_{\alpha/2}S_e}{d}\right)^2$.

- For t_y: $n = \frac{1}{\frac{1}{N^2 n_0}+\frac{1}{N}}$, where $n_0 = \left(\frac{z_{\alpha/2}S_e}{d}\right)^2$.

- An estimate of S_e^2 (the variability around (x, y) line with no intercept) can come from a prior study, a pilot study, or double sampling.

Sample Size Determination for Regression Estimation:

- From previous equations, the margin of error formulas for the parameters $\bar{y}_U$ and t_y using the variance MSE are:

$$\text{For } \bar{y}_U: z_{\alpha/2}\sqrt{\hat{V}(\hat{\bar{y}}_{reg})} = z_{\alpha/2}\sqrt{\left(\frac{N-n}{Nn}\right)MSE} \quad (24)$$

$$\text{For } \bar{t}_y: z_{\alpha/2}\sqrt{\hat{V}(\hat{t}_{reg})} = z_{\alpha/2}\sqrt{\frac{N(N-n)}{n}MSE} \quad (25)$$

  The sample size calculations are computed by setting the margin of error (d) and solving for n.

- For $\bar{y}_U$: $n = \frac{1}{\frac{1}{n_0}+\frac{1}{N}}$, where $n_0 = \left(\left(\frac{z_{\alpha/2}}{d}\right)^2 \times MSE\right)$.

- For t_y: $n = \frac{1}{\frac{1}{N^2 n_0}+\frac{1}{N}}$, where $n_0 = \left(\left(\frac{z_{\alpha/2}}{d}\right)^2 \times MSE\right)$.

- The MSE for the regression is an estimate of the variability about the regression line. You can use MSE from a prior study, a pilot study, or double sampling.