Report 8
Double Sampling

Name of the project: Analysis of Average Winter Temperature in College Park, Maryland

The true parameter: the true average temperature (in ℉) in College Park, MD during winter months, **μ** = 39.177˚F
The population size *N* = 720

The sample size of subsample of the variable of interest *n* = 84

1. Choose the same auxiliary variable x as in Report 4.

   In Report 4, we used **dew** as the auxiliary variable, as we showed it is linearly related to our variable of interest, temperature.

2. Perform double sampling with SRS at each phase with the sample size for the first phase n′ around twice of the sample size chosen in Report 2. The sample size for the second phase n is the same as in the Report 2.

   In Report 2, we used sample size n = 84, so we will use the same sample size n = 84 for the second phase. For the first phase, we will use twice the sample size n, so we will use n' = 168.

   Code

```
df = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")
temp = df[1:720, 5] #y
dew = df[1:720, 9] #x

N = 720
npr = 168
n = 84

firstSample = sample(1:N, npr, replace=F) #indices
firstSampleTemp = temp[firstSample]
firstSampleDew = dew[firstSample]

> firstSampleDew
```

```
  [1]  25.9  15.7  29.1   2.8  52.3  46.4  32.7  30.1  38.5  40.9   7.7
34.4  22.3  26.3  42.1  35.9   6.5  15.6  29.4  49.3  48.4

 [22]  14.3  35.9  20.3  47.5  22.6  13.1  31.5  28.6  -1.6  22.5 -10.4
59.5  18.2  28.6  20.9  32.4  49.9  20.5  40.2  20.7  19.7

 [43]  18.6  38.8  25.2  15.8  36.5  21.6  23.6  25.1  19.7  27.1  12.3
29.9  26.1  16.2  30.6  26.4  39.3  44.3  25.7  26.6  12.5

 [64]  39.7  42.7  28.4  19.4  16.7  25.7  24.8  14.5  20.1  15.2  -8.6
23.5  15.5  -3.4  17.1  23.9   3.1  45.6   7.7  35.0  25.0

 [85]  31.3  23.3  42.0  29.9  20.7  18.1  47.9  12.9  18.1  47.1  27.9
12.6  32.0  31.5  29.8  29.2  49.8  36.2  46.4  49.7  37.1

[106]  17.6  28.9  26.3  30.3  13.2  22.4  21.2  33.6  24.4  19.3  20.6
35.1  20.2  39.7  16.4  50.1  23.3  45.8  24.2  12.0  11.3

[127]  36.7  31.0  27.9  15.6  26.3  33.1  28.8  29.6  26.3  26.0  45.2
23.4  32.3  12.1  37.4  22.2  53.7  22.6  25.3  13.9  42.7

[148]  18.2  34.3  24.1  20.8  24.0  19.4  23.0  15.9   5.4  28.4   8.0
43.8  10.3  17.8  30.8  30.7  28.7  11.9  28.1  61.0  48.6

> firstSampleTemp

  [1] 31.4 23.5 38.3 23.4 56.0 56.6 44.2 43.9 47.7 43.0 27.6 40.1 38.4
29.3 46.2 46.2 29.5 37.4 41.3 50.7 50.2 33.5 45.0 31.0 49.8

 [26] 32.6 22.5 46.8 41.3 19.9 35.1 10.9 60.5 42.7 39.6 37.1 43.6 53.3
37.0 53.0 35.2 36.1 40.9 45.0 36.3 36.3 48.3 36.6 39.6 34.6

 [51] 32.1 39.7 34.7 46.3 35.2 36.1 43.3 40.6 48.2 59.5 32.4 38.7 25.9
54.3 46.5 43.4 38.5 26.6 48.2 37.4 25.5 38.4 31.2 13.8 41.4

 [76] 35.0 17.1 31.8 46.6 21.6 51.8 26.2 43.7 41.7 47.5 35.0 55.4 36.3
37.9 33.1 49.8 31.0 32.7 49.2 44.8 30.5 46.2 42.5 44.2 36.8

[101] 60.3 42.8 52.1 62.0 41.3 30.1 33.7 35.2 49.2 38.1 34.9 43.5 43.2
42.3 37.4 31.9 45.3 37.0 42.7 30.2 56.2 37.1 55.3 42.9 29.1

[126] 22.6 39.3 47.9 44.1 31.2 36.4 43.8 33.2 38.6 34.1 32.3 57.9 38.4
45.7 28.1 52.7 41.0 60.9 40.3 39.5 32.6 46.8 41.4 44.5 33.2

[151] 36.3 40.8 37.0 38.5 33.9 24.1 40.3 21.4 57.2 24.4 35.9 39.3 43.0
32.3 27.7 33.6 62.0 49.4
```

```
#second sample

secSample = sample(1:npr, n, replace=F) #indices

secSampleTemp = firstSampleTemp[secSample]

secSampleDew = firstSampleDew[secSample]
```

```
> secSampleDew
 [1] 28.9 50.1 49.3 40.2 36.5 20.2 -1.6 30.6 29.9 20.7 25.1 49.8 32.7 48.6
3.1 15.6 19.7 12.5 29.2 47.5 34.3 21.2 25.7 26.0 37.4
[26] 26.3 22.6 15.6 13.1 16.7 18.6 12.0 32.4 35.9 23.3 23.3 22.3  6.5 15.5
40.9 11.3 23.5 35.1 29.6 23.9 17.6 23.4 47.1 18.1 17.8
[51] 31.5 15.7 35.9 25.2 16.2 26.4 37.1 16.4 15.2 24.4 31.3 29.1 24.2 45.2
12.1 59.5 46.4 25.7 36.2 20.5 18.1 19.4 20.8  2.8 22.5
[76] 30.1 30.7 42.1 28.1  7.7 27.9 28.8 22.6 46.4
> secSampleTemp
 [1] 33.7 56.2 50.7 53.0 48.3 37.0 19.9 43.3 46.3 37.9 34.6 60.3 44.2 49.4
21.6 37.4 32.1 25.9 36.8 49.8 44.5 43.5 32.4 32.3 52.7
[26] 34.1 32.6 31.2 22.5 26.6 40.9 29.1 43.6 45.0 35.0 37.1 38.4 29.5 35.0
43.0 22.6 41.4 45.3 38.6 46.6 30.1 38.4 49.2 32.7 35.9
[51] 46.8 23.5 46.2 36.3 36.1 40.6 41.3 30.2 31.2 42.3 47.5 38.3 42.9 57.9
28.1 60.5 52.1 48.2 42.8 37.0 33.1 37.0 36.3 23.4 35.1
[76] 43.9 43.0 46.2 33.6 26.2 44.1 33.2 40.3 56.6
```

3. Perform a diagnostic analysis to determine if x and y have a linear relationship and the fitted line goes through the origin based on the sample data. Do regression analysis y~x.

   Code:

```
reg = lm(secSampleTemp ~ secSampleDew) #y ~ x
summary(reg)
```

   Output:

```
Call:
lm(formula = secSampleTemp ~ secSampleDew)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3339 -3.1468 -0.0744  2.8339  9.5899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  21.19537    1.10712   19.14   <2e-16 ***
secSampleDew  0.67761    0.03802   17.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.188 on 82 degrees of freedom
Multiple R-squared:  0.7948,  Adjusted R-squared:  0.7923
F-statistic: 317.6 on 1 and 82 DF,  p-value: < 2.2e-16
```

4. Make a conclusion about the appropriateness of using ratio estimators based on the result of 3. Does your conclusion agree with part 7 of Report 4 (Regression output based on the whole population)?

   The p-value for both the intercept and the slope are very small (<2e-16) and are both less than α=0.05. Since the p-value for the intercept is less than α=0.05, we reject the null hypothesis that the intercept is 0. Since the p-value for the slope is less than α=0.05, we reject the null hypothesis that the slope is 0. Since we would reject the null hypothesis that the slope is 0, this shows that the relationship between dew and temperature is linear. However, since we are rejecting the null hypothesis that the intercept is 0, the linear relationship does not go through the origin. This means that using the ratio estimator is not the most appropriate since the linear relationship would not go through the origin. This conclusion is the same as our conclusion from part 7 of Report 4 that performed a regression with the entire population (both p-values were also < 0.05) where we concluded that the regression estimator would be more appropriate than the ratio estimator for the same reason - that dew and temperature are linearly related but not through the origin.

5. Estimate your parameter of interest by ratio estimator. Estimate its variance and standard deviation.

   Formulas used:

   Ratio estimator for mean: $\hat{\mu}_r = \frac{\hat{\tau}_r}{N} = \frac{r * \hat{\tau}_x}{N} = \left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right)\frac{\hat{\tau}_x}{N} =$ **39.51653**

   Ratio estimator variance:

   $var(\hat{\mu}_r) = \frac{1}{N^2} var(\hat{\tau}_r) = \frac{1}{N^2} * (N(N - n')\frac{s^2}{n'} + N^2 \frac{n'-n}{n'n(n-1)} \sum_{i=1}^{n} (y_i - rx_i)^2) =$ **1.044212**

   Ratio estimator standard deviation: $stddev(\hat{\mu}_r) = \sqrt{var(\hat{\mu}_r)} =$ **1.021867**

   Code:

```
#Estimate mu
r = sum(secSampleTemp)/sum(secSampleDew)
```

```
tauhatx = (N/npr)*sum(firstSampleDew)

tauhatr = r*tauhatx
muhatr = tauhatr/N

> muhatr
39.51653

#Estimate variance and standard deviation
varhattauhat = N * (N-npr) * var(secSampleTemp) / npr + N^2 * ((npr-n)
/ (n*npr)) *(sum((secSampleTemp-r*secSampleDew)^2))/(n-1)
varhatmuhat = varhattauhat/N^2

> varhatmuhat
[1] 1.044212

> sqrt(varhatmuhat)
[1] 1.021867
```

6. Choose the second auxiliary variable x the same as in Report 4.

   The second auxiliary variable x we used in Report 4 was **humidity.**

7. Repeat steps 2–5.

(2)

In Report 2, we used sample size n = 84, so we will use the same sample size n = 84 for the second phase. For the first phase, we will use twice the sample size n, so we will use n' = 168.

Code

```
df = read.csv("/Users/sophiahu/Desktop/STAT440/winterWeather.csv")
temp = df[1:720, 5] #y
humidity = df[1:720, 10] #x
```

```
N = 720

npr = 168

n = 84


firstSample = sample(1:N, npr, replace=F) #indices

firstSampleTemp = temp[firstSample]

firstSampleHumidity = humidity[firstSample]

> firstSampleHumidity

  [1] 90.01 53.90 59.22 56.09 82.83 77.21 78.07 61.60 71.84 90.51 79.25 68.28
46.74 58.58 50.92 69.77 69.83 95.95 72.73 54.10 49.10

 [22] 88.87 93.40 68.54 62.14 81.43 92.40 37.99 52.07 47.80 92.80 67.00 64.17
60.60 74.48 68.43 66.10 74.90 71.52 57.60 57.10 76.00

 [43] 46.20 88.31 53.48 66.49 79.04 66.00 58.20 44.63 89.30 59.22 80.35 81.10
52.23 72.10 45.03 49.30 48.19 55.09 69.91 46.79 58.39

 [64] 73.90 73.61 94.62 58.10 65.00 50.70 59.68 57.55 46.70 72.90 78.70 53.55
69.00 61.90 51.80 94.38 52.50 78.40 91.07 64.43 50.20

 [85] 45.20 58.70 93.03 62.76 90.48 80.87 40.82 50.70 61.60 75.55 68.30 83.10
59.00 54.70 52.55 45.75 56.80 67.85 48.28 41.07 40.95

[106] 54.80 80.60 70.20 87.81 51.11 68.80 79.08 84.70 68.08 68.56 76.25 56.70
86.70 44.40 91.03 47.70 81.90 89.01 38.70 54.81 47.53

[127] 37.30 42.50 66.47 54.08 52.67 64.26 89.27 71.37 82.10 85.60 48.50 86.25
61.10 89.30 69.10 61.59 45.30 42.87 50.40 58.85 52.07

[148] 85.46 64.30 87.54 56.30 60.77 59.82 55.13 53.56 58.00 59.40 46.86 52.38
63.33 52.71 81.60 61.19 42.50 68.05 57.36 40.80 49.72

> firstSampleTemp

  [1] 34.0 37.0 29.2 40.2 40.4 38.7 40.3 35.6 46.2 51.8 46.7 44.5 26.4 32.3
38.9 45.4 56.6 49.4 35.3 37.7 18.5 46.0 44.7 46.6 35.0

 [26] 41.5 43.0 30.4 37.0 23.9 49.6 35.1 47.2 51.0 39.9 49.3 19.8 68.8 43.7
37.1 40.6 56.9 35.9 29.3 34.7 55.8 36.3 28.1 40.0 24.1

 [51] 42.7 46.2 58.0 33.6 32.0 36.1 35.0 30.5 55.6 46.3 31.9 27.8 37.3 43.7
45.7 50.7 22.5 36.3 12.5 34.0 59.5 39.9 47.7 38.7 32.4

 [76] 48.1 35.5 50.3 50.1 31.2 49.4 44.9 45.1 34.5 45.3 41.0 36.4 61.3 44.8
37.3 32.6 37.0 54.9 33.3 40.9 44.1 40.3 38.4 41.7 33.3

[101] 30.2 43.7 37.3 40.9 30.8 51.2 51.6 39.1 42.2 28.1 53.5 45.3 53.0 22.5
41.2 50.1 43.4 32.3 24.8 42.6 22.7 52.1 42.9 19.9 44.0
```

```
[126] 46.2 10.9 38.7 62.0 44.3 55.0 40.3 41.3 35.2 34.4 46.2 29.2 37.1 37.4
42.6 30.5 43.3 33.5 41.1 45.7 42.4 43.3 48.9 44.7 40.4
[151] 34.1 30.1 39.5 34.3 50.5 42.8 32.9 32.8 26.1 42.2 44.1 38.2 34.1 42.7
35.8 52.7 13.5 61.0
```

```
#second sample
secSample = sample(1:npr, n, replace=F) #indices
secSampleTemp = firstSampleTemp[secSample]
secSampleHumidity = firstSampleHumidity[secSample]
```

```
> secSampleHumidity
 [1] 78.07 63.33 90.01 56.09 71.52 45.75 58.20 85.60 66.00 80.35 88.87 59.22
64.17 67.00 68.30 68.43 51.11 69.77 48.19 81.10 64.43
[22] 79.25 50.20 53.56 60.60 88.31 45.03 52.38 66.47 95.95 68.56 68.05 61.10
40.95 89.01 75.55 87.54 89.30 45.20 37.99 52.07 85.46
[43] 94.62 92.40 89.27 81.90 69.91 50.70 78.40 44.40 86.25 76.00 89.30 49.10
37.30 54.81 59.68 61.60 50.40 46.70 80.87 91.07 52.55
[64] 40.80 62.14 61.90 91.03 42.50 82.83 55.13 72.73 47.53 56.30 62.76 47.70
61.59 54.10 56.80 52.07 55.09 59.22 70.20 53.48 94.38
> secSampleTemp
 [1] 40.3 42.2 34.0 40.2 43.7 33.3 40.0 46.2 28.1 58.0 46.0 46.2 47.2 35.1
40.9 49.3 28.1 45.4 55.6 33.6 45.1 46.7 34.5 50.5 51.0
[26] 29.3 35.0 26.1 62.0 49.4 41.2 35.8 37.4 30.8 42.9 33.3 40.4 42.6 45.3
30.4 43.3 48.9 50.7 43.0 41.3 52.1 31.9 12.5 49.4 24.8
[51] 37.1 56.9 42.7 18.5 10.9 44.0 34.0 54.9 45.7 39.9 37.3 44.9 41.7 13.5
35.0 35.5 42.6 42.7 40.4 34.3 35.3 46.2 34.1 61.3 22.7
[76] 43.3 37.7 30.2 37.0 46.3 29.2 39.1 34.7 50.1
```

(3)

Code

```
reg = lm(secSampleTemp ~ secSampleHumidity) #y ~ x
summary(reg)
```

Output

```
Call:
lm(formula = secSampleTemp ~ secSampleHumidity)

Residuals:
    Min      1Q  Median      3Q     Max
-23.525  -4.551  -1.130   5.337  22.303

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.53181    4.24284   5.546 3.47e-07 ***
secSampleHumidity  0.24641    0.06254   3.940  0.00017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.239 on 82 degrees of freedom
Multiple R-squared:  0.1592,  Adjusted R-squared:  0.1489
F-statistic: 15.52 on 1 and 82 DF,  p-value: 0.0001704
```

(4)

The p-value for both the intercept (3.47e-07) and the slope are very small (0.00017) and are both less than α=0.05. Since the p-value for the intercept is less than α=0.05, we reject the null hypothesis that the intercept is 0. Since the p-value for the slope is less than α=0.05, we reject the null hypothesis that the slope is 0. Since we reject the null hypothesis that the slope is 0, this shows that the relationship between humidity and temperature is linear. However, since we are rejecting the null hypothesis that the intercept is 0, the linear relationship does not go through the origin. This means that using the ratio estimator is not the most appropriate since the linear relationship would not go through the origin. This conclusion is the same as our conclusion from part 7 of Report 4 that performed a regression with the entire population (both p-values were also <0.05) where we concluded that the regression estimator would be more appropriate than the ratio

estimator for the same reason - that humidity and temperature are linearly related but not through the origin.

(5)

Ratio estimator for mean: $\hat{\mu}_r = \frac{\hat{\tau}_r}{N} = \frac{r * \hat{\tau}_x}{N} = \left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right)\frac{\hat{\tau}_x}{N} =$ **39.06281**

Ratio estimator variance:

$var(\hat{\mu}_r) = \frac{1}{N^2} var(\hat{\tau}_r) = \frac{1}{N^2} * (N(N - n')\frac{s^2}{n'} + N^2\frac{n'-n}{n'n(n-1)}\sum_{i=1}^{n}(y_i - rx_i)^2) =$ **1.159334**

Ratio estimator standard deviation: $stddev(\hat{\mu}_r) = \sqrt{var(\hat{\mu}_r)} =$ **1.076724**

Code

```
#Estimate mu
r = sum(secSampleTemp)/sum(secSampleHumidity)
tauhatx = (N/npr)*sum(firstSampleHumidity)

tauhatr = r*tauhatx
muhatr = tauhatr/N

> muhatr
[1] 39.06281

#Estimate variance and standard deviation
varhattauhat = N * (N-npr) * var(secSampleTemp) / npr + N^2 * ((npr-n) /
(n*npr)) *(sum((secSampleTemp-r*secSampleHumidity)^2))/(n-1)
varhatmuhat = varhattauhat/N^2
varhatmuhat

> varhatmuhat
[1] 1.159334
```

```
> sqrt(varhatmuhat)
[1] 1.076724
```

8. Choose the best estimator of your parameter based on the smallest standard deviation (variance).

The best estimator of our parameter would be using **dew** as the auxiliary variable because it had a smaller standard deviation (variance) than using the humidity as the auxiliary variable. Using dew, the variance was 1.044212 and the standard deviation was 1.021867. Whereas, the variance was 1.159334 and the standard deviation was 1.076724 using humidity. This makes sense because dew and temperature are more correlated as the correlation coefficient is higher than that for humidity.

9. Show all formulas used at each step as well as the code.

Formulas and code shown at each step above.