Report 4
Estimation by Using Auxiliary Data.

• **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**

The population size *N* = 720

The sample size *n* = 84

1. Choose an auxiliary variable *x* that should be related to your variable of interest *y*. Take a SRS of size *n* (the same size as in Report 2).
We can take dew (x) as our auxiliary variable which is related to our variable of interest, temperature (y).

Next, we take a SRS of size n=84 using R.

```
df <- read.csv("/Users/sophiahu/Downloads/STAT440 Data Winter.csv")

temp <- df[1:720, 5]

dew <- df[1:720, 9]

humid <- df[1:720, 10]

N = 720

n = 84

SRS <- sample(1:N, n, replace = FALSE)
```

2. Perform a diagnostic analysis to determine if *x* and *y* have a linear relationship based on the sample data. Do regression analysis *y* ~ *x*.

```
y <- temp[SRS]

x <- dew[SRS]

reg = lm(y~x)

summary(reg)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1321 -3.5942 -0.6881  3.6294 11.5867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.37095    1.39112   16.80   <2e-16 ***
x            0.59749    0.04338   13.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.542 on 82 degrees of freedom
Multiple R-squared:  0.6982,    Adjusted R-squared:  0.6945
F-statistic: 189.7 on 1 and 82 DF,  p-value: < 2.2e-16
```

From the output above, the p-value for the intercept and slope are both less than 0.05. We can reject the null hypothesis that the intercept and slope are 0. There is a linear relationship between temperature and dew that does not cross the origin.

3. Make conclusion about the appropriateness of using ratio and regression estimators based on the result of 2.

Since temperature and dew are linearly related, but not through the origin, the linear regression estimator is more appropriate than the ratio estimator. (By "not through the origin" we mean that when a data point has a dew of 0, the temperature is not also 0, and vice versa. In other words, the intercept is not 0.)

4. Estimate your parameter of interest by Ratio estimator. Estimate its variance and give a confidence interval of α level chosen in Report 2.

Formulas:

$\widehat{\mu_r} = r * \mu_x$, with $r = \frac{\bar{y}}{\bar{x}}$ and μ_x as the population mean for the auxiliary variable

$\widehat{var(\mu_r)} = \left(\frac{N-n}{N}\right)\frac{s_r^2}{n}$, with $s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2$

$CI_{95} = \widehat{\mu_r} \pm t_{n-1,\alpha/2}\sqrt{\widehat{var(\mu_r)}}$

```
# mu_hat_r = r*mu_x
```

```
mu_x = sum(dew)/N

r = sum(y)/sum(x)

mu_hat_r = r*mu_x

mu_hat_r


# var_hat_mu = ((N-n)/N)*(s_sq/n)

s_sq_r = (1/(n-1))*sum((y-r*x)^2)

var_hat_mu = ((N-n)/N)*(s_sq_r/n)

var_hat_mu


# 95% CI

t <- qt(0.025, n-1, lower.tail=FALSE)

lower <- mu_hat_r - t*sqrt(var_hat_mu)

lower

upper <- mu_hat_r + t*sqrt(var_hat_mu)

Upper
```

Results:

The estimated true mean temperature is 37.212, and its estimated variance is 1.060. The 95% confidence interval is (35.165, 39.259).

5. Estimate your parameter of interest by Regression estimator. Estimate its variance and give a confidence interval of α level chosen in Report 2.

$\widehat{\mu_L} = a + b * \mu_x$, with $a = \bar{y} - b\bar{x}$, and $b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$\widehat{var(\mu_L)} = \left(\frac{N-n}{Nn(n-2)}\right)\sum_{i=1}^{n}(y_i - a - bx_i)^2$

$CI_{95} = \widehat{\mu_L} \pm t_{n-2,\alpha/2}\sqrt{\widehat{var(\mu_L)}}$

```
# mu_hat_L = a + b*mu_x = y_bar b*(mu_x - x_bar)

# b = slope (dew) estimate from regression analysis

b = 0.59749


# a = intercept (temp) estimate from regression analysis

a = 23.37095

mu_hat_L = a + b*(mu_x)

mu_hat_L


# var_hat_mu_hat_L = ((N-n)/(N*n*(n-2)))*sum((y-a-b*x)^2)

var_hat_mu_hat_L = ((N-n)/(N*n*(n-2)))*sum((y-a-b*x)^2)

var_hat_mu_hat_L


# 95% CI

t <- qt(0.025, n-2, lower.tail=FALSE)

lower <- mu_hat_L - t*sqrt(var_hat_mu_hat_L)

lower

upper <- mu_hat_L + t*sqrt(var_hat_mu_hat_L)

upper
```

Results:

The estimated true mean temperature is 39.513, and the estimated variance is 0.217. The 95% confidence interval is (38.586, 40.439).

6. Choose the best estimator of your parameter based on width of CI.

The regression estimator is better since it resulted in a more narrow CI resulting from a lower estimator variance

7. Calculate the true regression coefficients. Namely, do regression *y ~ x* using whole data set (population). Is your conclusion in the part 3 changed? How does it change?

Our conclusion from part 3 does not change since the p-values for the intercept and slope are still less than 0.05, so the regression estimator would still be most appropriate.

```
reg = lm(temp~dew)
summary(reg)
```

```
Call:
lm(formula = temp ~ dew)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7176  -3.1514  -0.1115   2.6732  18.1192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.09783    0.35686   59.12   <2e-16 ***
dew          0.66922    0.01192   56.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.128 on 718 degrees of freedom
Multiple R-squared:  0.8145,    Adjusted R-squared:  0.8142
F-statistic:  3153 on 1 and 718 DF,  p-value: < 2.2e-16
```

8. Repeat steps 1-7 with another variable.

We can take humidity (x) as our auxiliary variable which is related to our variable of interest, temperature (y).

Next, we take a SRS of size n=84 using R.

Regression Analysis:

The p-value for the intercept and slope are both less than 0.05, so we reject the null hypothesis (the intercept and slope are 0) and conclude that there is a linear relationship between temperature and humidity. Since they are not linearly correlated through the origin, the linear regression estimator would be more appropriate to use.

```
t_h <- temp[SRS]

h <- humid[SRS]

reg = lm(t_h~h)
```

```
summary(reg)

 Call:
 lm(formula = t_h ~ h)

 Residuals:
      Min       1Q   Median       3Q      Max
 -16.3598  -3.9686  -0.2773   3.6396  22.1805

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept) 24.57086    3.72631   6.594 3.89e-09 ***
 h            0.23881    0.05518   4.328 4.22e-05 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 7.746 on 82 degrees of freedom
 Multiple R-squared:  0.1859,    Adjusted R-squared:  0.176
 F-statistic: 18.73 on 1 and 82 DF,  p-value: 4.223e-05
```

Ratio estimate:

```
# mu_hat_r = r*mu_x

mu_x_h = sum(humid)/N

r = sum(t_h)/sum(h)

mu_hat_r = r*mu_x_h

mu_hat_r



# var_hat_mu = ((N-n)/N)*(s_sq/n)

s_sq_r = (1/(n-1))*sum((t_h-r*h)^2)

var_hat_mu = ((N-n)/N)*(s_sq_r/n)

var_hat_mu




# 95% CI

t_val <- qt(0.025, n-1, lower.tail=FALSE)

lower <- mu_hat_r - t_val*sqrt(var_hat_mu)

lower

upper <- mu_hat_r + t_val*sqrt(var_hat_mu)
```

```
upper
```

Result:

The estimated population mean temperature is 39.256, and the estimated variance is 0.9718. The 95% confidence interval is (37.295, 41. 216).

Regression estimate:

```
mu_x_h = sum(humid)/N

# b = slope (dew) estimate from regression analysis

b = 0.23881


# a = intercept (temp) estimate from regression analysis

a = 24.57086


mu_hat_L = a + b*(mu_x_h)

mu_hat_L


# var_hat_mu_hat_L = ((N-n)/(N*n*(n-2)))*sum((y-a-b*x)^2)

var_hat_mu_hat_L = ((N-n)/(N*n*(n-2)))*sum((t_h-a-b*h)^2)

var_hat_mu_hat_L


# 95% CI

t_val <- qt(0.025, n-2, lower.tail=FALSE)

lower <- mu_hat_L - t_val*sqrt(var_hat_mu_hat_L)

lower

upper <- mu_hat_L + t_val*sqrt(var_hat_mu_hat_L)

upper
```

Result:

The estimated population mean temperature is 39.879, and the estimated variance is 0.631. The 95% confidence interval is (38.300, 41.459).

Choose the best estimator:

The regression estimate had the narrowest CI, so it is the best estimator.

Calculate the true regression coefficients:

The p-values for the intercept and slope are both less than 0.05, so our conclusion using the sample data is the same. Thus, the regression estimator would still be most appropriate.

```
reg = lm(temp~humid)

summary(reg)

 Call:
 lm(formula = temp ~ humid)

 Residuals:
      Min       1Q   Median       3Q      Max
 -23.0819  -5.7901  -0.1372   5.4108  26.7264

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept) 21.98160    1.37051   16.04   <2e-16 ***
 humid        0.26825    0.02078   12.91   <2e-16 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 8.634 on 718 degrees of freedom
 Multiple R-squared:  0.1883,    Adjusted R-squared:  0.1872
 F-statistic: 166.6 on 1 and 718 DF,  p-value: < 2.2e-16
```

9. Which variable out of these two you prefer for the estimation of your parameter and why? Make your selection based on estimation only (forget step 7).

Dew would be more preferable because it had the most narrow CI and smallest estimated variance when the regression estimator was used. The estimated variance and CI using the ratio estimate for dew resulted in similar values for that of humidity using ratio and regression estimation. This further supports our preference to use the linear regression estimator, as solved in the previous steps, and the dew variable to estimate the true average winter temperature in College Park.

Dew:

1. Ratio: 40.340, 1.417, (37.972, 42.707), CI width = 4.735
2. Regression: 39.513, 0.217, (38.586, 40.439), CI width = 1.853

Humidity:

1. Ratio: 39.256, 0.9718, (37.295, 41. 216), CI width = 3.921
2. Regression: 39.879, 0.631, (38.300, 41.459), CI width = 3.159

10. Which estimator out of these four is best for your data? Make your selection based on estimation only (forget step 7). The best estimator should have a small variance and your parameter should be in CI.

The best estimator is the linear regression estimator using dew as the auxiliary variable. It had the smallest variance out of the four estimators, and the true mean temperature of 39.177˚F was included in the 95% confidence interval.

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

Code provided for each step above.