Homework 4

Regression Adjustment

Jaiden Atterbury

03-05-2023

Bayesian Inference:

1. You are conducting an A/B test to determine the Average Causal Effect δ produced by a change. Suppose that based on past experience you have the following prior distribution for δ:

$$\delta \sim N(\mu_0 = 0.1, \phi_0^2 = 16)$$

where here μ_0 is the prior mean and ϕ_0^2 is the prior variance.

Suppose that you carry out the experiment with $n_T = n_C = 1000$, the observed difference $D = 0.9$ and we suppose that the population variance of the response in both arms is $\sigma^2 = 6.1$.

Setup:
In this A/B test we are looking to find the population average causal effecrt δ which is defined as $\mu_T - \mu_C$, due to the fundamental problem of causal inference we must estimate this value with $D \equiv \bar{Y}_T - \bar{Y}_C$. Since a sample $n_T = n_C = 1000$ is taken from two populations with the same population variance such that $Y_T \sim Norm(\mu_T, \sigma^2)$, and $Y_C \sim Norm(\mu_C, \sigma^2)$, it follows from results about sums of normal random variables that $E[D] = E[\bar{Y}_T - \bar{Y}_C] = \delta$ and $Var[D] = Var[\bar{Y}_T - \bar{Y}_C] = \sigma^2(1/n_T + 1/n_C)$. Thus our likelihood is: $D|\mu_T, \mu_C, \sigma^2 \sim N(\delta, \sigma^2(1/n_T + 1/n_C))$. Furthermore, the following is the prior distribution for δ: $\delta \sim N(\mu_0 = 0.1, \phi_0^2 = 16)$. We can use these distributions and our observed value of D to find and use the posterior distribution, plus construct credible intervals.

(a) Find the posterior variance ϕ_1^2 of the posterior distribution for δ.

To find the posterior variance we will follow the formula on slide 15 of the bayes intro lecture notes:

$$\begin{aligned}
\phi_1^2 &= \frac{1}{\frac{1}{\phi_0^2} + \frac{1}{\sigma^2(1/n_T + 1/n_C)}} \\
&= \frac{1}{\frac{1}{16} + \frac{1}{6.1(1/1000 + 1/1000)}} \\
&= \frac{976}{80061} \\
&= 0.012191
\end{aligned}$$

(b) Using your answer from (a), find the posterior mean μ_1 of the posterior distribution for δ.

To find the posterior mean we will follow the formula on slide 15 of the bayes intro lecture notes, as well as use the posterior variance we found in part (a):

$$
\begin{aligned}
\mu_1 &= \frac{\frac{1}{\phi_0^2}\mu_0 + \frac{1}{\sigma^2(1/n_T+1/n_C)}D}{\frac{1}{\phi_0^2} + \frac{1}{\sigma^2(1/n_T+1/n_C)}} \\
&= (\frac{1}{\phi_0^2}\mu_0 + \frac{1}{\sigma^2(1/n_T+1/n_C)}D) \cdot \phi_1^2 \\
&= (\frac{0.1}{16} + \frac{0.9}{6.1(1/1000+1/1000)}) \cdot \frac{976}{80061} \\
&= \frac{720061}{800610} \\
&= 0.89939
\end{aligned}
$$

(c) Find a symmetric 95% posterior credible interval for δ.

In order to find the 95% credible interval for δ based on the data we will use qnrom to solve for the 0.025th quantile and the 0.975th quantile of $\delta | D, \sigma^2 \sim Norm(0.89939, 0.012191)$.

```r
upper_cred <- qnorm(p = 0.975, mean = 720061/800610, sd = sqrt(976/80061))
lower_cred <- qnorm(p = 0.025, mean = 720061/800610, sd = sqrt(976/80061))
```

Thus, as calculated above, our 95% credible interval for δ based on the data is [0.6829879, 1.1157931].

This interval tells us that there is a 95% probability that the true estimate of δ would lie within the interval [0.6829879, 1.1157931], given the evidence provided by our observed data.

(d) Suppose that the change is expected to be profitable only if $\delta > 0.75$. Compute the posterior probability that $\delta > 0.75$.

In order to compute the posterior probability that $\delta > 0.75$, we will use the posterior distribution and the pnorm function to compute this probability.

```r
prob_g75 <- pnorm(q = 0.75, mean = 720061/800610, sd = sqrt(976/80061), lower.tail = F)
```

Thus, as calculated above, the posterior probability that $\delta > 0.75$ is $P(\delta > 0.75) = 0.9119775$.

Regression Adjustment:

2. Prove the claim on slide 9 of the adjustment lecture, that the population slope q_{PLS} minimizes the variance of $\hat{\bar{y}}_q$, given by (2).

Proof: On slide 9 of the adjustment lecture slides there was a claim that stated q_{PLS} minimizes the variance of $\hat{\bar{y}}_q$. In order to prove this claim we must first differentiate the equation given by (2) and find the critical value. Since $Var[\hat{\bar{y}}_q] = \frac{(N-n)}{(N-1)}\frac{1}{n}[\frac{1}{N}\sum_{i=1}^{N}((y_i - \bar{y}) - q(z_i - \bar{z}))^2]$ by (2) it follows that:

$$
\begin{aligned}
\frac{d}{dq}Var[\hat{\bar{y}}_q] &= \frac{d}{dq}(\frac{(N-n)}{(N-1)}\frac{1}{n}[\frac{1}{N}\sum_{i=1}^{N}((y_i - \bar{y}) - q(z_i - \bar{z}))^2]) \\
&= \frac{d}{dq}(\frac{(N-n)}{nN(N-1)}[\sum_{i=1}^{N}((y_i - \bar{y})^2 - 2q(z_i - \bar{z})(y_i - \bar{y}) + q^2(z_i - \bar{z})^2]) \\
&= \frac{(N-n)}{nN(N-1)}[\frac{d}{dq}\sum_{i=1}^{N}(y_i - \bar{y})^2 - \frac{d}{dq}2q\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y}) + \frac{d}{dq}q^2\sum_{i=1}^{N}(z_i - \bar{z})^2] \\
&= \frac{2(N-n)}{nN(N-1)}[-\sum_{i=1}^{N}(y_i - \bar{y})(z_i - \bar{z}) + q\sum_{i=1}^{N}(z_i - \bar{z})^2]
\end{aligned}
$$

Setting this equal to zero we obtain:

$$\begin{aligned}
0 &= \frac{2(N-n)}{nN(N-1)}[-\sum_{i=1}^{N}(y_i-\bar{y})(z_i-\bar{z})+q\sum_{i=1}^{N}(z_i-\bar{z})^2] \\
0 &= -\sum_{i=1}^{N}(y_i-\bar{y})(z_i-\bar{z})+q\sum_{i=1}^{N}(z_i-\bar{z})^2 \\
q\sum_{i=1}^{N}(z_i-\bar{z})^2 &= \sum_{i=1}^{N}(y_i-\bar{y})(z_i-\bar{z}) \\
q_{PLS} &= \frac{\sum_{i=1}^{N}(y_i-\bar{y})(z_i-\bar{z})}{\sum_{i=1}^{N}(z_i-\bar{z})^2}
\end{aligned}$$

Thus we can see that the critical value of $Var[\hat{\bar{y}}_q]$ is q_{PLS} which is the slope of the population least squares regression line from regressing y_i on z_i.

In order to show that q_{PLS} minimizes $Var[\hat{\bar{y}}_q]$, we must perform the second derivative test and show that the second derivative of $Var[\hat{\bar{y}}_q]$ at q_{PLS} is greater than zero.

$$\begin{aligned}
\frac{d^2}{dq^2}Var[\hat{\bar{y}}_q] &= \frac{d^2}{dq^2}\frac{2(N-n)}{nN(N-1)}[-\sum_{i=1}^{N}(y_i-\bar{y})(z_i-\bar{z})+q\sum_{i=1}^{N}(z_i-\bar{z})^2] \\
&= \frac{d^2}{dq^2}(\frac{-2(N-n)}{nN(N-1)}\sum_{i=1}^{N}(y_i-\bar{y})(z_i-\bar{z})+q\frac{2(N-n)}{nN(N-1)}\sum_{i=1}^{N}(z_i-\bar{z})^2) \\
&= 0+\frac{2(N-n)}{nN(N-1)}\sum_{i=1}^{N}(z_i-\bar{z})^2 \\
&= \frac{2(N-n)}{nN(N-1)}\sum_{i=1}^{N}(z_i-\bar{z})^2
\end{aligned}$$

Since $N > n$ and $N, n > 0$, it follows that $\frac{2(N-n)}{nN(N-1)}$ is always greater than zero. Furthermore, since $\sum_{i=1}^{N}(z_i-\bar{z})^2$ involves summing squares, unless z_i is the same for all of the N subjects this quantity will also always be greater than zero. However, it is important to note that if the z_i's were the same for all of the subjects, q_{PLS} couldn't be computed in the first place, thus we when q_{PLS} exists, the sum of squares value will always be greater than zero. Lastly, notice that the second derivative of $Var[\hat{\bar{y}}_q]$ does not depend on q_{PLS}, thus $\frac{d^2}{dq^2}Var[\hat{\bar{y}}_{q_{PLS}}] > 0$

Since q_{PLS} is a critical point and $\frac{d^2}{dq^2}Var[\hat{\bar{y}}_{q_{PLS}}] > 0$, it follows that q_{PLS} is the argmin and thus minimizes the variance of $\hat{\bar{y}}$. □

3. Read in the data in this file regression-adjustment-data.csv.

```r
finite_pop <- read_csv("regression-adjustment-data.csv")
```

Hint: Also look at the R code: regression-adjustment.R Compute and report the following:

(a) The unadjusted estimate of the ATE.

```r
# Find control and treatment group vectors
Y.c <- finite_pop$Y[0:100]
Y.t <- finite_pop$Y[101:200]
```

```
# Find control and treatment group sizes
k <- length(Y.t)
n <- length(finite_pop$Y)

# Find sample mean for control and treatment group
mean.Yc <- mean(Y.c)
mean.Yt <- mean(Y.t)

# Find sample variance for control and treatment group
var.c <- var(Y.c)
var.t <- var(Y.t)

ATE.unadj <- mean.Yt - mean.Yc
```

As calculated above, the unadjusted estimate of the ATE is 1.0739911.

(b) A 95% confidence interval for the unadjusted estimate of the ATE.

When doing an confidence interval for the unadjusted estimate of the ATE, we have three options for the variance estimator: the "naive" estimator, the Neyman estimator, and the Aronow estimator. We will compute the interval for all three of these estimators.

"Naive" estimator:

```
# Calculating an estimate of the variance of our estimate of the average causal effect
hat.var.hat.ace <- var.t/k + var.c/(n-k)

#Constructing a conservative 95% confidence interval for the smaple ACE
lower_naive <- ATE.unadj - qnorm(0.975)*sqrt(hat.var.hat.ace)
upper_naive <- ATE.unadj + qnorm(0.975)*sqrt(hat.var.hat.ace)
```

A "naive" 95% confidence interval for $\hat{ACE}$ is [-0.2496675, 2.3976497].

Neyman Estimator:

```
# Find the variances with finite population correction
var.t2 = (n - 1)/n * var.t
var.c2 = (n - 1)/n * var.c

# Calculate new estimate for variance
hat.var.hat.ace2 <- (1/(n-1)) * (var.t2*(n-k)/k + var.c2*k/(n-k) + 2*sqrt(var.t2 * var.c2))

# Calculate the 95% confidence interval for ace.hat
lower_ney <- ATE.unadj - qnorm(0.975)*sqrt(hat.var.hat.ace2)
upper_ney <- ATE.unadj + qnorm(0.975)*sqrt(hat.var.hat.ace2)
```

Our new, and tighter, 95% confidence interval for $\hat{ACE}$ is [-0.0686419, 2.2166241].

Aronow Estimator:

```
# Calculate covariance under Aronow's estimator
Ytsort <- sort(Y.t)
Ycsort <- sort(Y.c)
covxy <- sum(Ytsort*Ycsort)/k - sum(Ytsort)*sum(Ycsort)/k^2

# Calculate new estimate for variance
hat.var.hat.ace3 <- (1/(n-1)) * (var.t2*(n-k)/k + var.c2*k/(n-k) + 2*covxy)
```

```r
# Calculate the 95% confidence interval for ace.hat
lower_aro <- ATE.unadj - qnorm(0.975)*sqrt(hat.var.hat.ace3)
upper_aro <- ATE.unadj + qnorm(0.975)*sqrt(hat.var.hat.ace3)
```

Our new, and even tighter, 95% confidence interval for AĈE is [-0.0660144, 2.2139967].

(c) The adjusted estimate of the ATE.

When finding the adjusted estimate of the ATE, there are three methods we could use, two different regression for each treatment group, one multiple regression, and lastly, the Oaxaca-Blinder approach. Belowe we will find the adjusted estimate for the ATE using all three of these methods.

First Approach: two separate regressions:

```r
# Find covariates
Z.c <- finite_pop$Z[0:100]
Z.t <- finite_pop$Z[101:200]
Z <- c(Z.c,Z.t)

# Compute regressions
reg.c <- lm(Y.c~Z.c)
reg.t <- lm(Y.t~Z.t)

# Adjusted means:

# From control
ybar.reg.c <- mean(Y.c) + (reg.c$coef[2])*(mean(Z)-mean(Z.c))

# From treatment
ybar.reg.t <- mean(Y.t) + (reg.t$coef[2])*(mean(Z)-mean(Z.t))

# adjusted ATE via two regressions
ate.adj.two <- ybar.reg.t - ybar.reg.c
```

The adjusted estimate of the ATE using two seperate regressions is 1.1871594.

Second Approach: Multiple Regression:

```r
# Compute vectors for multiple regression
Trt <- finite_pop$Trt
Y <- c(Y.c,Y.t)
Inter <- Trt*(Z-mean(Z))

# Compute multiple regression
combined.reg <- lm(Y~Trt + Z + Inter)

# adjusted ATE via multiple regression
ate.adj.mult <- combined.reg$coeff["Trt"]
```

The adjusted estimate of the ATE using one multiple regression is 1.1871594.

Third Approach: Oaxaca-Blinder approach (Guo & Basse,2020):

```r
# Computing the linear model to obtain the fitted values
mu.c <- lm(Y~Z,subset(finite_pop,Trt==0))
mu.t <- lm(Y~Z,subset(finite_pop,Trt==1))

# Using the fitted values to obtain the estimate of the ATE
```

```r
ate.adj.oax <- mean(predict(mu.t,finite_pop)-predict(mu.c,finite_pop))
```

The adjusted estimate of the ATE using the Oaxaca-Blinder approach is 1.1871594.

(d) A 95% confidence interval for the adjusted estimate of the ATE. (Use the heteroscedastic variance estimate.)

```r
# Uses Huber-White Heteroskedastic Consistent Variance Estimate, aka "Sandwich Formula")
coeftest(combined.reg, vcov = vcovHC(combined.reg))["Trt",]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## 1.187159434 0.444195485 2.672605807 0.008160421
```

```r
tmp.heterosc <- coeftest(combined.reg, vcov = vcovHC(combined.reg))["Trt",]

upper_het <- tmp.heterosc[1]+qnorm(0.975)*tmp.heterosc[2]
lower_het <- tmp.heterosc[1]-qnorm(0.975)*tmp.heterosc[2]
```

As computed above, a 95% confidence interval for the adjusted estimate of the ATE using the heteroscedastic variance estimate is [0.3165523, 2.0577666]. This is a notable improvement as the interval is tighter than any of the three unadjusted intervals and it is bounded away from zero