Excercise 2 - ECO 530 (2410)

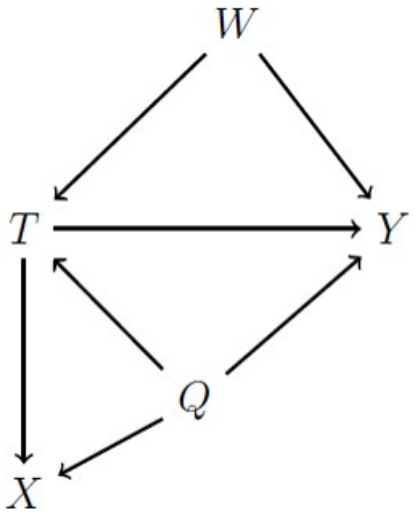
Devan Arnold

Instructions

Complete the exercises below. Be sure to show all of your work. For this assignment, you can submit:

- A PDF containing your written answers, tables, and figures along with the R script that generates them
- A R Markdown document containing your written answers with the R code embedded in the document.

Q1 - Directed Acyclic Graphs



Q1a - Assume that you are interested in the relationship between T and Y and that T only takes values of zero or one. What would happen if you chose to

estimate this relationship by comparing the sample average of Y among observations where $T = 1$ to the sample average of Y among observations where $T = 0$?

If you were to compare the observations of Y for the possible values of T your estimation would not represent the affect of T on Y. As the Directional Acyclic Graph (DAG) above shows, both variables Q and W influence the states of T and Y. Thus, without controlling for those confounding variables the estimation produced would not represent the true relationship desired.

Q1b - Using conditional expectations, write out the ideal comparison to identify the relationship between T and Y.

The ideal comparison would look something like:
$E[Y|T = 1, W = w, Q = q] - E[Y|T = 0, W = w, Q = q]$ for some constant w and q.

By holding the values of the confounding variables constant we can then properly evaluate
T's influence on Y.

Q2 - Calculating Probabilities with Known Distributions

Q2a - Consider a random variable X, where $X \sim N(0, 1)$. Use R to find $Pr(X \leq 1.64)$.

```
# Part (A)
# Given x~N(0,1), where mean = 0 and var = 1, find Pr(X<=1.64)
meanA <- 0
sdA <- sqrt(1)
pnorm(1.64,mean=meanA,sd=sdA,lower.tail = T)
```

```
## [1] 0.9494974
```

Q2b - Consider a random variable X, where $X \sim N(42, 8)$. Use R to find $Pr(X \geq 30)$.

```
# Part (B)
# Given X~N(42,8), where mean = 42 and var = 8, find Pr(X>=30)
meanB <- 42
sdB <- sqrt(8)
pnorm(30,mean=meanB,sd=sdB,lower.tail = F)
```

```
## [1] 0.999989
```

Q2c - Consider a random variable X, where $X \sim N(0, 1)$. Use R to find $Pr(|X| \geq 1.64)$.

```
# Part (C)
# Given X~N(0,1), where mean = 0 and var = 1, find Pr(|X|>=1.64)
meanC <- 0
sdC <- sqrt(1)
upperC <- pnorm(1.64,mean=meanA,sd=sdA,lower.tail = F)
lowerC <- pnorm(-1.64,mean=meanA,sd=sdA,lower.tail = T)
upperC + lowerC
```

```
## [1] 0.1010052
```

Q2d - Consider a random variable X, where $X \sim N(0, 1)$. Use R to find $Pr(X \leq -1.64 \cup X \geq 2.5)$.

```
# Part (D)
# Given X~N(0,1), where mean = 0 and var = 1, find Pr(X<=-1.64 U X>=2.5)
meanD <- 0
sdD <- sqrt(1)
upperD <- pnorm(2.5,mean=meanA,sd=sdA,lower.tail = F)
lowerD <- pnorm(-1.64,mean=meanA,sd=sdA,lower.tail = T)
upperD + lowerD
```

```
## [1] 0.05671225
```

Q3 - The Age of UMaine Students

Download the `E2data.RData` data set and load it in R. Assume that the observations in the data frame represent the entire population of UMaine students. Assume that the true average age (μ) of a UMaine student is 22.

Q3a - Consider the three estimators proposed below. Show whether each estimator is unbiased. Which one is the most efficient?**

For each of the below estimators, we will assume that an unbiased estimator has an expected value that equals the population mean μ

- $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Biasness	**Efficiency**
$\mu = E[\bar{y}]$	$var(\bar{y}) = var(\frac{\sum_{i=1}^{n} y_i}{n})$
$\mu = E[\frac{\sum_{i=1}^{n} y_i}{n}]$	$var(\bar{y}) = (\frac{1}{n})^2 \cdot var(\sum_{i=1}^{n} y_i)$
$\mu = \frac{1}{n} \cdot E[\sum_{i=1}^{n} y_i]$	$var(\bar{y}) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} var(y_i)$
$\mu = \frac{1}{n} \cdot \sum_{i=1}^{n} E[y_i]$	$var(\bar{y}) = \frac{1}{n^2} \cdot n \cdot var(y_i)$
$\mu = \frac{1}{n} \cdot \sum_{i=1}^{n} \mu$	$var(\bar{y}) = \frac{n}{n^2} \cdot \sigma^2$
$\mu = \frac{1}{n} \cdot n\mu$	$var(\bar{y}) = \frac{\sigma^2}{n}$
$\mu = \frac{n}{n} \cdot \mu$	

Biasness	Efficiency
$\mu = 1 \cdot \mu$	
$\mu = \mu$	
$\bar{y}$ is an unbiased estimator of μ	

- $\tilde{y} = \frac{\sum_{i=1}^{5} y_i}{5}$

Biasness	Efficiency
$\mu = E[\tilde{y}]$	$var(\tilde{y}) = var(\frac{\sum_{i=1}^{5} y_i}{5})$
$\mu = E[\frac{\sum_{i=1}^{5} y_i}{5}]$	$var(\tilde{y}) = (\frac{1}{5})^2 \cdot var(\sum_{i=1}^{5} y_i)$
$\mu = \frac{1}{5} \cdot E[\sum_{i=1}^{5} y_i]$	$var(\tilde{y}) = \frac{1}{5^2} \cdot \sum_{i=1}^{5} var(y_i)$
$\mu = \frac{1}{5} \cdot \sum_{i=1}^{5} E[y_i]$	$var(\tilde{y}) = \frac{1}{5^2} \cdot 5 \cdot var(y_i)$
$\mu = \frac{1}{5} \cdot \sum_{i=1}^{5} \mu$	$var(\tilde{y}) = \frac{5}{5^2} \cdot \sigma^2$
$\mu = \frac{1}{5} \cdot 5\mu$	$var(\tilde{y}) = \frac{\sigma^2}{5}$
$\mu = \frac{5}{5} \cdot \mu$	
$\mu = 1 \cdot \mu$	
$\mu = \mu$	
$\tilde{y}$ is an unbiased estimator of μ	

- $\hat{y} = y_1$

Biasness	Efficiency
$\mu = E[\hat{y}]$	$var(\hat{y}) = var(y_i)$
$\mu = E[y_1]$	$var(\hat{y}) = \sigma^2$
$\mu = \mu$	
$\hat{y}$ is an unbiased estimator of μ	

Given that estimators with smaller variances are more efficient, $\hat{y}$ is the least efficient estimator with $var(\hat{y}) = \sigma^2$. The most efficient estimator is $\bar{y}$ for samples greater than 5 observations ($var(\bar{y}) = \frac{\sigma^2}{n}$). For samples with less than 5 observations, $\tilde{y}$ is the most efficient ($var(\bar{y}) = \frac{\sigma^2}{n}$). Thus:

Sample Size	Efficiency Ranking
$n < 5$	$\bar{y} > \tilde{y} > \hat{y}$
$n = 5$	$\bar{y} = \tilde{y} > \hat{y}$
$n > 5$	$\tilde{y} > \bar{y} > \hat{y}$

Q3b - Draw a sample of 25 observations from the data frame. Using the sample, calculate an estimate of μ using $\bar{y}$, $\tilde{y}$ and $\hat{y}$. Report your estimates. Which one is the closest to the true value of μ?

```
# Sets working directory to data folder for Exercise 2
setwd(datapath)
# Loads RData file for the assignment
load(file="E2data.RDATA")

# Sets the sample size and mean (mu) for later use
sample.size <- 25
mu <- 22
# Sets a seed for later review
set.seed(seed=8675309)

# Takes a sample of size sample.size from the universe data frame and stores the values in
# the universe.sample data frame
universe.sample <- sample_n(universe,size = sample.size,replace=F)

# Calculates the values of the y_bar, y_tilde, and y_hat estimators
y_bar <- sum(universe.sample$age)/sample.size
y_tilde <- sum(universe.sample[1:5,1])/5
y_hat <- universe.sample[1,1]

# Determines the distance (in absolute value terms) of each estimator from the population mea
n
mu_y_bar <- abs(y_bar - mu)
mu_y_tilde <- abs(y_tilde - mu)
mu_y_hat <-  abs(y_hat - mu)

# Determines which of the above differences is least
smallest_difference <- min(c(mu_y_bar,mu_y_tilde,mu_y_hat))

# Reports the estimator that is closest to the population mean and then reports that
# difference in numerical terms.
if(smallest_difference == mu_y_bar){print("y_bar is the closest estimator to the mean mu = 2
2")}
```

```
## [1] "y_bar is the closest estimator to the mean mu = 22"
```

```
if(smallest_difference == mu_y_tilde){print("y_tilde is the closest estimator to the mean mu
= 22")}
if(smallest_difference == mu_y_hat){print("y_hat is the closest estimator to the mean mu = 2
2")}
print(smallest_difference)
```

```
## [1] 0.3725899
```

Q4 - Hypothesis Testing

Imagine that you are interested the effect of going on a daily jog ($\hat{\beta}$) and life expectancy.

- **You estimate $\hat{\beta} = 6.25$** after analyzing a sample of 502 observations.
- **The standard error of your estimate is 1.43.**

Q4a - Test the following null hypotheses against the alternative, assuming you are willing to accept a 5% chance of committing a Type I error:**

- $H_0 : \beta = 9.25$
- $H_0 : \beta = 2.1$
- $H_0 : \beta = 3.5$

```
# Part (a)
# Declare variables for our estimated beta ('beta_hat'), number of observations
# (n, stored as 'obs'), and the standard error of our estimate 'se'
beta_hat <- 6.25
obs <- 502
se <- 1.43

# Declares the acceptable bound for one tail of the distribution
bound <- 0.025

# Declare the null hypothesis values for tests 1, 2, and 3
hyp_1 <- 9.25
hyp_2 <- 2.1
hyp_3 <- 3.5

# Determines the proportion of the distribution that resides above each hypothesis
upper_1 <- pnorm(hyp_1,mean=beta_hat,sd=se,lower.tail = F)
upper_2 <- pnorm(hyp_2,mean=beta_hat,sd=se,lower.tail = F)
upper_3 <- pnorm(hyp_3,mean=beta_hat,sd=se,lower.tail = F)
# Determines the proportion of the distribution that resides below each hypothesis
lower_1 <- pnorm(hyp_1,mean=beta_hat,sd=se,lower.tail = T)
lower_2 <- pnorm(hyp_2,mean=beta_hat,sd=se,lower.tail = T)
lower_3 <- pnorm(hyp_3,mean=beta_hat,sd=se,lower.tail = T)

# Creates a boolean value to determine if each hypothesis is within the 5% unallowed
# portion of the distribution.
hyp_test_1 <- (upper_1 > bound & lower_1 > bound)
hyp_test_2 <- (upper_2 > bound & lower_2 > bound)
hyp_test_3 <- (upper_3 > bound & lower_3 > bound)

# Reports to the console the results of the above tests in plain language
paste("For hypothesis 1 of beta =",hyp_1,"we cannot reject the hypothesis - TRUE or FALSE?",h
yp_test_1,sep=" ")
```

```
## [1] "For hypothesis 1 of beta = 9.25 we cannot reject the hypothesis - TRUE or FALSE? FALS
E"
```

```
paste("For hypothesis 2 of beta =",hyp_2,"we cannot reject the hypothesis - TRUE or FALSE?",h
yp_test_2,sep=" ")
```

```
## [1] "For hypothesis 2 of beta = 2.1 we cannot reject the hypothesis - TRUE or FALSE? FALS
E"
```

```
paste("For hypothesis 3 of beta =",hyp_3,"we cannot reject the hypothesis - TRUE or FALSE?",h
yp_test_3,sep=" ")
```

```
## [1] "For hypothesis 3 of beta = 3.5 we cannot reject the hypothesis - TRUE or FALSE? TRUE"
```

Q4b - Still using the estimate from above, construct the following confidence intervals. Plot the confidence intervals in R and be sure to interpret each interval in words.

```
# Part (b)
# Determines the upper bounds for a 95% and 99% CI for mean = beta_hat, standard
# deviation = se (our standard error variable)
upper_95 <- round(qnorm(0.975,mean=beta_hat,sd=se,lower.tail = T),digits = 2)
upper_99 <- round(qnorm(0.995,mean=beta_hat,sd=se,lower.tail = T),digits = 2)

# Determines the lower bounds for a 95% and 99% CI for mean = beta_hat, standard
# deviation = se (our standard error variable)
lower_95 <- round(qnorm(0.025,mean=beta_hat,sd=se,lower.tail = T),digits = 2)
lower_99 <- round(qnorm(0.005,mean=beta_hat,sd=se,lower.tail = T),digits = 2)

# Stores the upper and lower bounds of the 95% CI in the vector 'conf_95' and prints
# the values to the console
conf_95 <- c(lower_95,upper_95)
conf_95
```

```
## [1] 3.45 9.05
```

```
# Stores the upper and lower bounds of the 99% CI in the vector 'conf_99' and prints
# the values to the console
conf_99 <- c(lower_99,upper_99)
conf_99
```

```
## [1] 2.57 9.93
```

The above confidence intervals represent the set of values we would reject/fail to reject a hypothesis givena Type I error tolerance of 95% and 99%, respectively

Q4c - What is the smallest sized test for which you would reject the null hypotheses $H_0 : \beta = 4$

```
# Part (c)
# Declares the hypothesis that beta = 4 for testing. Stores value in 'hyp_c'
hyp_c <- 4

# Determines the proportion of the distribution above and below the hyp_c value
below_c <- round(pnorm(hyp_c,mean=beta_hat,sd=se,lower.tail = T),digits = 2)
above_c <- round(pnorm(hyp_c,mean=beta_hat,sd=se,lower.tail = F),digits = 2)

# Determines the lesser of the above proportions and takes that as the one-tail alpha.
# Then multiplies that value by two for a two tail test for reporting to the user.
alpha_c <- 2*min(c(below_c,above_c))

# Displays results in console using natural language.
paste("For the hypothesis beta =",hyp_c," we would reject the hypothesis if using an alpha va
lue of",alpha_c,sep=" ")
```

```
## [1] "For the hypothesis beta = 4  we would reject the hypothesis if using an alpha value o
f 0.12"
```

Q5 - The Effect of Studying Hard on GPA

Q5a - Using the full sample, calculate an estimate of $\hat{\gamma}$.

```
# Part (a)
# This bit of code just proves that there are no values above 1 or below 0 for the
# study.hard column of the universe data frame. Understanding this aspect of the data
# will be important for the approach I take later.
max(universe$study.hard)
```

```
## [1] 1
```

```
min(universe$study.hard)
```

```
## [1] 0
```

```
# Takes the mean of GPAs in the full sample where case (0) is study.hard = 0 and
# where case (1) is study.hard = 1. Stores the conditional means in y0_bar for case (0)
# and y1_bar for case (1)
y0_bar <- mean(universe$gpa[universe$study.hard==0])
y1_bar <- mean(universe$gpa[universe$study.hard==1])

# Takes the sum of all GPA's divided by themselves to create a count of observations
# then subtracts the count of observations where study.hard =/= 0
n0 <- sum(universe$gpa/universe$gpa)-sum(universe$study.hard)
# Sums all of the values of study.hard where study.hard =/= 0
n1 <- sum(universe$study.hard)

# Calculates the sample variance for case (0) (stored as S0_2) and case (1) (stored
# as s1_2).
s0_2 <- var(universe$gpa[universe$study.hard==0])
s1_2 <- var(universe$gpa[universe$study.hard==1])

# Calculates gamma_hat based on the values of y0_bar and y1_bar and reports to the console
gamma_hat <- y1_bar - y0_bar
gamma_hat
```

```
## [1] 1.254598
```

Q5b - Calculate the standard error associated with your estimate.

```
# Part (b)
# Using the above calculated s0_2, s1_2, n0, and n1 we can determine the standard
# error for gamma_hat
se <- sqrt((s1_2/n1)+(s0_2/n0))
se
```

```
## [1] 0.02186911
```

Q5c - Test the hypothesis that studying hard has no effect on GPA
In order to determine if there is a GPA effect due to studying hard, we will assume that there is NO GPA effect due to studying hard ($H_0 : \hat{\gamma} = 0$). If this is true, than we would expect that, on average, $\hat{\gamma} = 0$. We will test to see if our observed gamma_hat is outside of values of alpha associated with a 95% CI. This means that an upper_gamma value less than 0.025 or a lower value less than 0.025 would

cause us to reject the hypothesis $H_0 : \hat{\gamma} = 0$.

```
# Part (c)
# Declare alpha
alpha = 0.025
# Determine the proportion of the distribution that is higher than our observed
# gamma_hat
upper_gamma <- pnorm(gamma_hat,mean = 0, sd=se,lower.tail = F)
upper_gamma
```

```
## [1] 0
```

```
# Determine the proportion of the distribution that is lower than our observed
# gamma_hat
lower_gamma <- pnorm(gamma_hat,mean = 0, sd=se, lower.tail=T)
lower_gamma
```

```
## [1] 1
```

```
# Since the value of upper_gamma is 0, we can reject the hypothesis that gamma_hat = 0
gamma_hyp_test <- ((upper_gamma < alpha)|(lower_gamma < alpha))
paste("For hypothesis of gamma_hat = 0, we reject the hypothesis - TRUE or FALSE?",gamma_hyp_
test,sep=" ")
```

```
## [1] "For hypothesis of gamma_hat = 0, we reject the hypothesis - TRUE or FALSE? TRUE"
```

Q5d - Construct and discuss a 95 % confidence interval around your estimate of $\hat{\gamma}$

```
# Part (d)
# First, we will create the upper and lower bounds of the 95% CI using the qnorm
# function measured from the lower tail of the distribution
gamma_95u <- round(qnorm(0.975,mean = gamma_hat,sd=se,lower.tail = T),digits=2)
gamma_95l <- round(qnorm(0.025,mean = gamma_hat,sd=se,lower.tail=T),digits=2)

# We place the lower and upper bounds of the 95% CI a vector named 'gamma_95CI' for
# reporting to the console
gamma_95CI <- c(gamma_95l,gamma_95u)
gamma_95CI
```

```
## [1] 1.21 1.30
```

This confidence interval represents the values of $\hat{\gamma}$ for which we would reject or fail to reject a hypothesis given $\alpha = 0.05$.

Q6…Conditional on the library

Repeat the tasks from Q5a, Q5b, and Q5d, but this time focus only on students who know where the library is. Compare and contrast your results to those you obtained in Q5.

Q6a - Using the full sample, calculate an estimate of $\hat{\gamma}$ conditioned on knowing where the library is.

```
# Part (a)
# This code is reused from (5a), but with a term added to condition the estimators
# on students knowing where the library is (library = 1)
max(universe$library)
```

```
## [1] 1
```

```
min(universe$library)
```

```
## [1] 0
```

```
# Takes the mean of GPAs in the full sample where case (0) is study.hard = 0 and
# where case (1) is study.hard = 1. Stores the conditional means in y0_bar for case (0)
# and y1_bar for case (1)
y0_bar6 <- mean(universe$gpa[universe$study.hard==0 & universe$library==1])
y1_bar6 <- mean(universe$gpa[universe$study.hard==1 & universe$library==1])

# Takes the sum of all GPA's divided by themselves to create a count of observations
# then subtracts the count of observations where study.hard =/= 0
n06 <- sum(universe$gpa[universe$library==1]/universe$gpa[universe$library==1])-sum(univers
e$study.hard[universe$library==1])
# Sums all of the values of study.hard where study.hard =/= 0
n16 <- sum(universe$study.hard[universe$library==1])

# Calculates the sample variance for case (0) (stored as S0_2) and case (1) (stored
# as s1_2).
s0_26 <- var(universe$gpa[universe$study.hard==0 & universe$library==1])
s1_26 <- var(universe$gpa[universe$study.hard==1 & universe$library==1])

# Calculates gamma_hat based on the values of y0_bar and y1_bar and reports to the console
gamma_hat6 <- y1_bar6 - y0_bar6
gamma_hat6
```

```
## [1] 0.9625202
```

Q6b - Calculate the standard error associated with your estimate conditioned on knowing where the library is.

```
# Part (b)
# Again, code is reused from (5b)
# Using the above calculated s0_2, s1_2, n0, and n1 we can determine the standard
# error for gamma_hat
se6 <- sqrt((s1_26/n16)+(s0_26/n06))
se6
```

```
## [1] 0.03841955
```

Q6d - Construct and discuss a 95 % confidence interval around your estimate of $\hat{\gamma}$ conditioned on knowing where the library is

```r
# Part (d)
# Again, code is reused from (5d)
# First, we will create the upper and lower bounds of the 95% CI using the qnorm
# function measured from the lower tail of the distribution
gamma_95u_6 <- round(qnorm(0.975,mean = gamma_hat6,sd=se6,lower.tail = T),digits=2)
gamma_95l_6 <- round(qnorm(0.025,mean = gamma_hat6,sd=se6,lower.tail=T),digits=2)

# We place the lower and upper bounds of the 95% CI a vector named 'gamma_95CI' for
# reporting to the console
gamma_95CI_6 <- c(gamma_95l_6,gamma_95u_6)
gamma_95CI_6
```

```
## [1] 0.89 1.04
```

This confidence interval represents the values of gamma for which we would reject or fail to reject a hypothesis given an alpha of 0.05