Group 4, Problem Set 2

Alyssa, Johanna, Takuya, and Matt

3/8/2022

1

Exercise 2

The correlation between mass and elite tolerance scores is 0.52; between mass tolerance scores and repression scores, −0.26; between elite tolerance scores and repression scores,−0.42.

The equation for which we are computing the coefficients is Repression = β_1Mass tolerance + β_2Elite tolerance + δ.

For computing purposes, let's define repression as U, mass tolerance as V and elite tolerance as X. Therefore, the converted equation is $U = aV + bX + \delta$. In matrix form, $U = M\binom{a}{b} + \delta$, where M is the set of matrices, M = (V X).

Due to standardization, $r_{vx} = \frac{1}{n}\sum_{i=1}^{n} V_i X_i$.

M'M= $\begin{pmatrix} \sum_{i=1}^{n} V_i^2 & \sum_{i=1}^{n} V_i X_i \\ \sum_{i=1}^{n} V_i X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix}$, therefore

M'M= n $\begin{pmatrix} 1 & 0.52 \\ 0.52 & 1 \end{pmatrix}$, and

M'U= $\begin{pmatrix} \sum_{i=1}^{n} V_i U_i \\ \sum_{i=1}^{n} X_i U_i \end{pmatrix}$ = n$\begin{pmatrix} r_{VU} \\ r_{XU} \end{pmatrix}$= n$\begin{pmatrix} -0.26 \\ -0.42 \end{pmatrix}$

To compute the coefficients, we solve for $(M'M)^{-1}M'U = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$

```
MM_dt <- c(1,0.52,0.52, 1)
M_prime_M <- matrix(MM_dt,nrow=2)
M_prime_M

##      [,1] [,2]
## [1,] 1.00 0.52
## [2,] 0.52 1.00

M_prime_U <- matrix(c(-0.26,-0.42),ncol=1)
M_prime_U
```

```
##       [,1]
## [1,] -0.26
## [2,] -0.42
```

```
coef <- (solve(M_prime_M)) %*% M_prime_U
coef
```

```
##             [,1]
## [1,] -0.05701754
## [2,] -0.39035088
```

The standardization of the variables allows us to use the methods in lecture 5; namely, we can apply the OLS assumptions and use matrix algebra to compute the coefficients and their variance.

Exercise 3

$\hat{\sigma}^2$=residual variance of $\hat{\delta}$, and using equation 8 in Freedman, p. 85, $\hat{\sigma}^2 = 1 - \hat{a}^2 - \hat{b}^2 - \hat{a}\hat{b}r_{VX}$. Then, we multiple $\sigma^2(\frac{n}{n-p})$. Since there are two variables on the right hand side of equation (10):$U = aV + bX + \delta$, and since the sample size is small, p=3.

```
a_hat <- -0.06
b_hat <- -0.39
n <- 36
p <- 3
sigma2_hat <- 1 - (a_hat)^2 - (b_hat)^2 - 2*(a_hat*b_hat*0.52)
sd <- sqrt(sigma2_hat*(36/33))
sd
```

```
## [1] 0.9457834
```

Exercise 4

The formula for the standard errors of the coefficients is $SE_{\hat{\beta}} = \hat{\sigma}^2[X'X]^{-1}$.

```
var <- as.matrix(sigma2_hat * solve(M_prime_M))
se <- sqrt(var[1,1])
se
```

```
## [1] 1.06012
```

```
var_diff <- var[1,1] + var[2,2] - (2*var[1,2])
var_diff
```

```
## [1] 3.416517
```

```
se_diff <- sqrt(var_diff)
se_diff
```

```
## [1] 1.848382
```

```
t_a <- a_hat/se
t_a

## [1] -0.05659737

t_b <- b_hat/se
t_b

## [1] -0.3678829

t_diff <- (a_hat-b_hat)/se_diff
t_diff

## [1] 0.1785345
```

Unlike Gibson, none of the t-ratios calculated imply significance for either coefficient OR their difference.

2

Given that the regression is standardized, the variance of the coefficients equals 1, so the sample size cancels out when we compute each matrix, as shown below.

$$\hat{\beta} = (M'M)^{-1}M'U$$

$$M'M = n\begin{pmatrix} \mathbf{E}(V_i^2) & \mathbf{E}(V_iX_i) \\ \mathbf{E}(V_iX_i) & \mathbf{E}(X^2) \end{pmatrix}$$

$$M'M = n\begin{pmatrix} 1 & r_{VX} \\ r_{VX} & 1 \end{pmatrix}$$

$$\hat{\beta} = (M'M)^{-1}M'U = \mathrm{n}\begin{pmatrix} 1 & r_{VU} \\ r_{XU} & 1 \end{pmatrix} * \frac{1}{n\begin{pmatrix} 1 & r_{VX} \\ r_{VX} & 1 \end{pmatrix}}$$

3

Pairwise missing data deletion is when there is missing data within a certain variable, one would delete the row - but if the variables of interest for analysis are not missing, then you can keep the row (even if there are some NAs in the row for other variables). Listwise deletion is when there is a missing value at all in the row, the whole row is deleted.

No. The non-bolded entries are correlation coefficients.

4

(a) We are able to calculate the three bivariate correlation coefficients in Figure 1.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(naniar)

data <- read.table("https://www.stat.berkeley.edu/users/census/gibson.txt")

# first we replace the -1 with NA
data <- data %>%
  replace_with_na(replace = list(V1 = c(-1),
                             V2 = c(-1),
                             V3 = c(-1),
                             V4 = c(-1),
                             V5 = c(-1),
                             V6 = c(-1)))

# first we try to calculate the first bivariation correlation
# coefficient
# between mass tolerance and repression

masses_repression <- cor(data$V2, data$V6, use = "complete.obs")

print(masses_repression)

## [1] -0.2617426

# this does give us the number reported in figure 1

# what about elite tolerance and repression?

elites_repression <- cor(data$V4, data$V6, use = "complete.obs")

print(elites_repression)

## [1] -0.420311

# this does give us the number reported in figure 1

# finally, what about mass and elite tolerance?
elites_masses <- cor(data$V4, data$V2, use = "complete.obs")

print(elites_masses)
```

```
## [1] 0.5168376

# same, it gives us the number from figure 1
```

(b) Can we replicate the regression results given at the end of the first document (standardized, weighted regression)? we are not able to replicate the results given at the end of the first document. We have attempted below.

First there is a standardization of variables, then they are weighted. But still we do not get the same answers as Freedman. We do this in several ways (first standardize, then weight; weight only; standardize only), but do not replicate the results from the document. The only results we replicate is the unweighted, unstandardized regression.

```
data <- na.omit(data) # first delete incomplete rows

# first we standardize the variables
data <- data %>%
  mutate(mass_std = scale(V2),
         elite_std = scale(V4),
         rep_std = scale(V6))

# then we create weights, which should be square root of n
data <- mutate(data,
               weight = sqrt(V3 + V5))

# then we weight the data points by multiplying by weights

data <- data %>%
  mutate(mass_std_weight = mass_std*weight,
         elite_std_weight = elite_std*weight,
         rep_std_weight = rep_std*weight)

lm(rep_std ~ elite_std + mass_std, data=data, weights = weight) # nope

##
## Call:
## lm(formula = rep_std ~ elite_std + mass_std, data = data, weights = weight
)
##
## Coefficients:
## (Intercept)    elite_std     mass_std
##     0.10492     -0.39384     -0.05297

lm(rep_std_weight ~ elite_std_weight + mass_std_weight, data=data) # no

##
## Call:
## lm(formula = rep_std_weight ~ elite_std_weight + mass_std_weight,
##     data = data)
##
```

```
## Coefficients:
##      (Intercept)  elite_std_weight   mass_std_weight
##           1.0488           -0.3994           -0.1321

# what if instead we only weight
data <- data %>%
  mutate(mass_weight = V2*weight,
         elite_weight = V4*weight,
         rep_weight = V6*weight)

lm(rep_weight ~ elite_weight + mass_weight, data=data)

##
## Call:
## lm(formula = rep_weight ~ elite_weight + mass_weight, data = data)
##
## Coefficients:
##   (Intercept)  elite_weight   mass_weight
##      -6.6942        0.7619       -0.3919

# this does not replicate

lm(V6 ~ V2 + V4, data=data)

##
## Call:
## lm(formula = V6 ~ V2 + V4, data = data)
##
## Coefficients:
## (Intercept)           V2           V4
##      7.3889       0.1851      -1.3830

# here, we DO replicate the unweighted equation
```

5

Our response schedule for the Gibson study (from pg 324 of Freedman) is:

$$R = a + Mb + Ec + errorterm$$

Where:

R = Repression

M = Mass Tolerance

E = Elite Tolerance

a = the constant parameter

b = Mass Tolerance coefficient (effect/impact) c = Elite Tolerance coefficient (effect/impact)

error term = the margin of error within our response schedule

(NOTE: We include the constant in our response schedule above because we are not assuming standardization. However, if we *do* standardize, then we can remove “a” from our response schedule.)

To interpret this response schedule, we can see that when we plug in values for M (Mass Tolerance) and E (Elite Tolerance), we get an estimated value for R (Repression). The value for R is the sum of our constant a, the observed M multiplied by the effect of M (or b), the observed E multiplied by the effect of E (or c), and the margin of error (or error term).

For example, if M goes up by one unit with E held fixed, then R goes up by b units. Similarly, if E goes up one unit with M held fixed, then R goes up by c units.

There are two main assumptions embedded into our response schedule:

1) R is determined from the response schedule (meaning the errors are iid, the mean is 0 and the variance is σ^2).

2) M and E are chosen at random by nature, independent of the error terms (in other words, there is no confounding).

These assumptions address (1) causal and (2) statistical issues with our response schedule. (pg 92 & 94, Freedman).

Both of these assumptions raise theoretical, conceptual, and empirical difficulties. Perhaps most obvious is concerns regarding confounding variables and endogeneity. Although we assume there are no confounding variables, empirically this may not hold. For example, it is easy to think of several variables that may actually be confounding our response schedule, such as access to media or state party (i.e. red states versus blue states). This is one of the most poignant conceptual difficulties with OLS regression analysis–there is an infinite number of potential confounders when dealing with observational data. Theoretically, we can try to address the largest ones in our model, but there is always the chance that we do not identify all of them–or even simply the largest ones.

6

a)

Generally speaking, neither of these statements are completely true. The second statement, “Regression analysis assumes causation but can be used to estimate the size of a causal effect—if the assumptions of the regression models are correct.” is closer to the truth IF all of our assumptions (including assuming that our DV is determined from a reliable and accurate response schedule and that there is no endoegneity at play) are true–otherwise, our regression cannot assume causation.

Additionally, the statement “Regression analysis demonstrates causation” is only true for the same reasons–IF all of our assumptions (including assuming that our DV is determined from a reliable and accurate response schedule and that there is no endoegneity at play) are true–otherwise, our regression does not demonstrate causation.

This statements hinge on the truth of our assumptions because regressions rely on these assumptions in order to draw any meaningful conclusions (this is particularly true for observational data, which we will get to in part (b)). Regressions simply tell us about the relationship of the IV’s with the DV, and it is up to us to utilize assumptions to draw causal effects. If these assumptions do *not* hold–for example, if there are confounding variables–then the regression is not demonstrating causality (which is why the statements, on their own, do not hold true).

Notably, if the regression is run on a properly conducted randomized experiment, then we can assume that the regression analysis demonstrates causality. In this scenario, the first statement would be just as true as the second one. However, in most cases we can be confident that the second statement is closer to reality, given the constraints of most data (particularly observational data).

b)

Not exactly. Our answer changes depending on if we are using observational or experimental data, but only if our assumptions above hold true with the experimental data but not with the observational data.

As described briefly in part (a), experimental data has the advantage of true randomization and can therefore be used with a regression analysis to identify a causal effect so long as our assumptions all hold. Observational data, however, does not have these assumptions baked in automatically–therefore, we need to verify that all the assumptions hold true until we can claim that our regression has causality.

c)

Our answer in (a) remains the same even if we replace “regression analysis” with “Analysis under the Neyman potential outcomes model”. We know this because the same issues presented in (a), such as endogeneity and confounding variables, would still be present in the Neyman model. The only way to circumvent this is if we can verify that we are using experimental data in the model (as discussed in part (b)).

7

(a)

As the equation (10.3) suggests, the direct effect of Z_i is d. Substituting (10.1) to (10.3),

$$Y_i = \alpha_3 + (d + ab)Z_i + (\alpha_1 + e_{1i})b + e_{3i}.$$

Since the total effect of Z_i is $d + ab$,the indirect effect is

$$d + ab - d = ab.$$

(b)

From the equation (10.2), the total effect, c

$$\begin{aligned} E(c_i) &= E(d_i + a_i b_i) \\ &= E(d_i) + E(a_i b_i) \\ &= E(d_i) + E(a_i)E(b_i) + Cov(a_i, b_i) \end{aligned}$$

If the parameters are same across subjects ($a_1 = a_2 = \ldots = E(a_i)andb_1 = b_2 = \ldots = E(b_i)$), then$Cov(a_i, b_i) = 0$. Thus, the total effect is equal to the sum of the direct effect and indirect effect. Otherwise, however, $Cov(a_i, b_i) \neq 0$.

(c)

$$E(a_i b_i) = E(a_i)E(b_i) + Cov(a_i, b_i)$$

From the assumptions, $E(a_i) = 0$] While b_i varies across i, a_i does not vary. So, $Cov(a_i, b_i) = 0$.

$$E(a_i b_i) = 0.$$

(d)

$M_i(0)$ refers to the potential outcome of M_i for $Z_i = 0$, while the second value of $Y_i(M_i(0),1)$ means this is a potential outcome of Y_i Since the first value and second value are contradict each other, we cannot observe $Y_i(M_i(0),1)$ in fact.

(e)

In the first group of equations using $Y_i(M_i(Z_i), Z_i)$, the potential outcomes of Y_i are defined by the potential outcomes of M_i. And they shows the indirect effect of Z_i.

Unlike the first group, the second groups using $Y_i(m, z)$ show the direct effect of Z_i and they can be estimated if it is possible to manipulate both Z_i and M_i.