Homework 3

Gergel Anastasia

6/5/2021

Exercise 1

Show that $Cov(\hat{y}_i, \hat{u}_i) = 0$.

$$Cov(\hat{y}_i, \hat{u}_i) = \frac{\sum(\hat{y}_i - \bar{\hat{y}})(\hat{u}_i - \bar{\hat{u}})}{N}.$$

Note that since $\sum \hat{u}_i = 0$, then $\bar{\hat{u}} = \frac{\sum \hat{u}_i}{n} = 0$ as well. Consequently,

$$\frac{\sum(\hat{y}_i - \bar{\hat{y}})(\hat{u}_i - 0)}{N} = \frac{\sum \hat{u}_i \hat{y}_i - \sum \hat{u}_i \bar{\hat{y}}}{N} = \frac{0 - 0}{N} = 0,$$

since the terms in the numerator are multiplied by $\sum \hat{u}_i = 0$. $Cov(\hat{y}_i, \hat{u}_i) = 0$, Q.E.D.

Exercise 2

(a) χ^2-tests of the independence between two variables are not adjusted for sample size N, making it more likely ceteris paribus that a χ^2-test will reject the null of no differences between groups as N grows larger.

- TRUE, statistical significance is more likely with a larger sample size, N.

(b) t-tests of the differences in group means are not adjusted for sample size N, making it more likely ceteris paribus that a t-test will reject the null of no differences between groups as N grows larger.

- TRUE. $N \to \inf$ shrinks standard errors and makes it more likely to reject the null hypothesis.

(c) The t distribution approximates the Normal distribution as N becomes large.

- TRUE

(d) A t-test of the differences between the means of two groups M and W with a total sample size of 200 is just as likely to reject the null when $N_M = 100$, $N_W = 100$ as when $N_M = 10$, $N_W = 190$.

- FALSE. Even if we assume that the pooled standard deviation, s_p, remains the same, t-statistics is not robust when changing sample sizes of two groups:

$$\frac{\bar{Y}_M - \bar{Y}_W}{s_p\sqrt{1/100 + 1/100}} > \frac{\bar{Y}_M - \bar{Y}_W}{s_p\sqrt{1/10 + 1/190}},$$

since

$$s_p\sqrt{1/100 + 1/100} < s_p\sqrt{1/10 + 1/190}0.141\ s_p < 0.324\ s_p$$

Exercise 3

$\bar{X} = 48.5$, $s_X = 10$, $N_X = 100$ and $\bar{Y} = 51.5$, $s_Y = 10$, $N_Y = 100$.

(a) 95% CI for the population mean μ_X:

$$(\bar{X} - 1.96\frac{s_X}{\sqrt{N_X}},\ \bar{X} + 1.96\frac{s_X}{\sqrt{N_X}})$$
$$(48.5 - 1.96\frac{10}{\sqrt{100}},\ 48.5 + 1.96\frac{10}{\sqrt{100}})$$
$$(46.54,\ 50.46)$$

95% CI for the population mean μ_Y:

$$(\bar{Y} - 1.96\frac{s_Y}{\sqrt{N_Y}},\ \bar{Y} + 1.96\frac{s_Y}{\sqrt{N_Y}})$$
$$(51.5 - 1.96\frac{10}{\sqrt{100}},\ 51.5 + 1.96\frac{10}{\sqrt{100}})$$
$$(49.54,\ 53.46)$$

(b) Evaluate the claim that $\mu_X = \mu_Y$:

$$(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}}$$
$$(48.5 - 51.5) \pm 1.96\sqrt{100/100 + 100/100}$$
$$(-4.41,\ -1.58),$$

the CI does not include zero $\rightarrow$ we do not reject the null hypothesis that $\mu_X = \mu_Y$, i.e. there is a statistically significant difference between groups.

(c) The reporter is not correct because simply adding errors underestimates the difference between μ_X and μ_Y Comparing the CI for μ_X with the CI for μ_Y means comparing $(\bar{v} \pm 1.96\frac{s_v}{\sqrt{N_v}})$, where $v \in \{X, Y\}$, instead of building the CI for the difference in μ_X and μ_Y, i.e. $(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}}$. The first scenario simply adds up errors, while the second adds errors correctly in quadrature.

Exercise 4

Constructing a correlation matrix with variables: President 2016 (`CC18_317`), Gun control (`CC18_320d`, Make it easier for people to obtain a concealed-carry gun permit), Family income (`faminc_new`), Race (`race`), State of Residence (`inputstate`), Abortion (`CC18_321a`, always allow a woman to obtain an abortion as a matter of choice). I am treating these ordered discrete variables as continuous.

```r
load("/Users/herrhellana/Dropbox/_NYU studies/Quant I/home assignments/HW3/cces18_common_vv.RData")

library(dplyr)
corr_data <- x %>% select("CC18_317", "CC18_320d", "faminc_new", "race", "educ", "CC18_321a")
corr_data <- na.omit(corr_data)
colnames(corr_data) <- c("president_2016", "gun", "income", "race", "educ", "abortion")
```

```r
# President 2018: 1- Trump, 2- Hilary
# gun: 1-for, 2-against
# Race: 1 - white
# icome: higher = more
# educ: higher = more
# abortion: 1-support, 2-oppose

library("Hmisc")
rcorr(as.matrix(corr_data)) #shows correlation coefficients and its significance
```

```
                president_2016   gun income  race  educ abortion
president_2016            1.00  0.33  -0.04  0.11  0.13    -0.37
gun                       0.33  1.00  -0.03  0.03  0.09    -0.36
income                   -0.04 -0.03   1.00  0.00  0.06     0.04
race                      0.11  0.03   0.00  1.00  0.04    -0.06
educ                      0.13  0.09   0.06  0.04  1.00    -0.08
abortion                 -0.37 -0.36   0.04 -0.06 -0.08     1.00

n= 45981


P
               president_2016 gun    income race   educ   abortion
president_2016                0.0000 0.0000 0.0000 0.0000 0.0000
gun            0.0000                0.0000 0.0000 0.0000 0.0000
income         0.0000         0.0000        0.4412 0.0000 0.0000
race           0.0000         0.0000 0.4412        0.0000 0.0000
educ           0.0000         0.0000 0.0000 0.0000        0.0000
abortion       0.0000         0.0000 0.0000 0.0000 0.0000
```

All correlation coefficients are significant except for the pair (`income`, `race`). The strongest correlation is between `president_2016` and `abortion` showing that voting for Trump in 2016 is negatively assosiated with standing by the woman's right for obtaining abortion, as expected. Another strong and obvious correlation exists between (`gun` and `president_2016`) and (`gun` and `abortion`): supporting the facilitation of obtaining a gun permit is positively associated with voting for Hilary and standing by the woman's right for obtaining abortion. Interestingly, although higher levels of education (`educ`) and being a non-white person (`race`) are positevely associated with voting for Hilary in 2016, these relations are very weak, as all other relations between variables.

However, significance of the correlation does not imply existence of a meaningful link between two variables, it just shows that variables are linearly related. The correlation can be spurious.

Exercise 5

(a-b) Recode the protest variable (`CC18_417a_4`) to zero-one, where one indicated participation in a protest, zero – otherwise.

```r
x$CC18_417a_4 <- recode(x$CC18_417a_4, "1"=as.integer("1"), "2"=as.integer("0"))

library(gmodels)
CrossTable(y = x$CC18_417a_4, x = x$race, prop.c = F,
           prop.t = F, prop.chisq = F, chisq = T, format="SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  51808

             | x$CC18_417a_4
      x$race |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           1 |     35643 |      4119 |     39762 |
             |   89.641% |   10.359% |   76.749% |
-------------|-----------|-----------|-----------|
           2 |      4269 |       319 |      4588 |
             |   93.047% |    6.953% |    8.856% |
-------------|-----------|-----------|-----------|
           3 |      3422 |       404 |      3826 |
             |   89.441% |   10.559% |    7.385% |
-------------|-----------|-----------|-----------|
           4 |      1326 |       130 |      1456 |
             |   91.071% |    8.929% |    2.810% |
-------------|-----------|-----------|-----------|
           5 |       338 |        43 |       381 |
             |   88.714% |   11.286% |    0.735% |
-------------|-----------|-----------|-----------|
           6 |      1024 |       182 |      1206 |
             |   84.909% |   15.091% |    2.328% |
-------------|-----------|-----------|-----------|
           7 |       440 |        58 |       498 |
             |   88.353% |   11.647% |    0.961% |
-------------|-----------|-----------|-----------|
           8 |        69 |        22 |        91 |
             |   75.824% |   24.176% |    0.176% |
-------------|-----------|-----------|-----------|
Column Total |     46531 |      5277 |     51808 |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  109.6783     d.f. =  7     p =  1.070744e-20



       Minimum expected frequency: 9.268974
```

The crosstabulation shows that different race groups have different participation levels, and this difference is significant according to the χ^2 statistics.

(c) Correlation assumes the linear relationship between variables. While we can interpret ordered discrete variables as linearly changing (like the `president_2016` variable), the race variable cannot be ordered and thus considered as linearly changing. Crosstabulation and χ^2-test have no assumtions about the $x - y$ relationship.

(d) Asian and Hispanic only.

```
# asian - 4, hispanic - 3
data_poc <- x %>% filter(race == 4 | race == 3)
CrossTable(y = data_poc$CC18_417a_4, x = data_poc$race, prop.c = F,
           prop.t = F, prop.chisq = F, chisq = T, format="SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  5282

              | data_poc$CC18_417a_4
data_poc$race |        0  |        1  | Row Total |
--------------|-----------|-----------|-----------|
            3 |     3422  |      404  |     3826  |
              |   89.441% |   10.559% |   72.435% |
--------------|-----------|-----------|-----------|
            4 |     1326  |      130  |     1456  |
              |   91.071% |    8.929% |   27.565% |
--------------|-----------|-----------|-----------|
 Column Total |     4748  |      534  |     5282  |
--------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3.086261     d.f. =  1     p =  0.07895606

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  2.909423     d.f. =  1     p =  0.0880634


       Minimum expected frequency: 147.1988
```

χ^2-statistics ceases to be significant (at 5% significance level), i.e. there is no relationship between the `race` and protest participation (`CC18_417a_4`) variables. This is due χ^2 being not adjusted to the sample size. However, the statistics is significant at 10% significance level.

(ii) t-test (difference-in-means)

```
t.test(data_poc$race, data_poc$CC18_417a_4)
```

```
    Welch Two Sample t-test
```

```
data:  data_poc$race and data_poc$CC18_417a_4
t = 466.25, df = 11841, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.151563 3.178174
sample estimates:
mean of x mean of y
3.2659669 0.1010981
```

T-test is statistically significant showing that the difference in means of `race` and the protest variable exists.

(iii) Both tests indicate that the relationship between `race` and protest participation exists but at different significance levels.

(iv) The absence of statistical significance does not imply the absence of substantively meaningful difference between group means and can be due to the wrong specification and/or test choice.

Exercise 6

Construct an index of hardship, `index`, using the variables `internethome` (3 - none), `healthins_7` (1 - no), `CC18_303_2` (1 - lost a job), `CC18_303_9` (1 - yes, been a victim of a crime).

```
# create four conditions
# if non are met, index = 0,
# if one of them is met, index = 1,
# if two, index = 2,
# if three, index = 3
# if four, index = 4


# i don't deal w NAs in variables here
# so the final index also has NAs whenever one variable has it
for (i in 1:nrow(x)){
  x$index[i] <- (x$internethome[i] == 3) +
    (x$healthins_7[i] == 1) +
    (x$CC18_303_2[i] == 1) +
    (x$CC18_303_9[i] == 1)
}


## income
# this way all 97s are NAs by default

x$income[x$faminc_new == 1] <- 5000
x$income[x$faminc_new == 2] <- 15000
x$income[x$faminc_new == 3] <- 25000

x$income[x$faminc_new == 4] <- 35000
x$income[x$faminc_new == 5] <- 45000
x$income[x$faminc_new == 6] <- 55000

x$income[x$faminc_new == 7] <- 65000
x$income[x$faminc_new == 8] <- 75000
```

```
x$income[x$faminc_new == 9] <- 90000

x$income[x$faminc_new == 10] <- 110000
x$income[x$faminc_new == 11] <- 135000
x$income[x$faminc_new == 12] <- 175000

x$income[x$faminc_new == 13] <- 225000
x$income[x$faminc_new == 14] <- 300000
x$income[x$faminc_new == 15] <- 400000

x$income[x$faminc_new == 16] <- 600000



CrossTable(x = x$income, y = x$index, prop.c = F,
           prop.t = F, prop.chisq = F, chisq = T, format="SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  53553

             | x$index
    x$income |         0  |         1  |         2  |         3  |         4  | Row Total |
-------------|------------|------------|------------|------------|------------|-----------|
        5000 |      1563  |       904  |       327  |        63  |         6  |      2863 |
             |   54.593% |   31.575% |   11.422% |    2.200% |    0.210% |    5.346% |
-------------|------------|------------|------------|------------|------------|-----------|
       15000 |      2938  |      1074  |       271  |        46  |         5  |      4334 |
             |   67.790% |   24.781% |    6.253% |    1.061% |    0.115% |    8.093% |
-------------|------------|------------|------------|------------|------------|-----------|
       25000 |      4022  |      1387  |       355  |        50  |         6  |      5820 |
             |   69.107% |   23.832% |    6.100% |    0.859% |    0.103% |   10.868% |
-------------|------------|------------|------------|------------|------------|-----------|
       35000 |      4620  |      1202  |       265  |        38  |         1  |      6126 |
             |   75.416% |   19.621% |    4.326% |    0.620% |    0.016% |   11.439% |
-------------|------------|------------|------------|------------|------------|-----------|
       45000 |      4152  |       912  |       179  |        21  |         2  |      5266 |
             |   78.845% |   17.319% |    3.399% |    0.399% |    0.038% |    9.833% |
-------------|------------|------------|------------|------------|------------|-----------|
       55000 |      4215  |       842  |       148  |        18  |         2  |      5225 |
             |   80.670% |   16.115% |    2.833% |    0.344% |    0.038% |    9.757% |
-------------|------------|------------|------------|------------|------------|-----------|
       65000 |      3431  |       531  |        74  |        10  |         0  |      4046 |
             |   84.800% |   13.124% |    1.829% |    0.247% |    0.000% |    7.555% |
-------------|------------|------------|------------|------------|------------|-----------|
       75000 |      3643  |       528  |        72  |         5  |         1  |      4249 |
             |   85.738% |   12.426% |    1.695% |    0.118% |    0.024% |    7.934% |
-------------|------------|------------|------------|------------|------------|-----------|
       90000 |      4506  |       485  |        63  |         6  |         0  |      5060 |
```

```
              |    89.051% |     9.585% |     1.245% |     0.119% |     0.000% |     9.449% |
-------------|------------|------------|------------|------------|------------|------------|
       110000 |       3229 |        332 |         39 |          1 |          0 |       3601 |
              |    89.670% |     9.220% |     1.083% |     0.028% |     0.000% |     6.724% |
-------------|------------|------------|------------|------------|------------|------------|
       135000 |       2886 |        254 |         23 |          1 |          0 |       3164 |
              |    91.214% |     8.028% |     0.727% |     0.032% |     0.000% |     5.908% |
-------------|------------|------------|------------|------------|------------|------------|
       175000 |       1904 |        161 |         11 |          1 |          0 |       2077 |
              |    91.671% |     7.752% |     0.530% |     0.048% |     0.000% |     3.878% |
-------------|------------|------------|------------|------------|------------|------------|
       225000 |        788 |         66 |          6 |          0 |          0 |        860 |
              |    91.628% |     7.674% |     0.698% |     0.000% |     0.000% |     1.606% |
-------------|------------|------------|------------|------------|------------|------------|
        3e+05 |        413 |         49 |          5 |          0 |          0 |        467 |
              |    88.437% |    10.493% |     1.071% |     0.000% |     0.000% |     0.872% |
-------------|------------|------------|------------|------------|------------|------------|
        4e+05 |        153 |         18 |          1 |          0 |          0 |        172 |
              |    88.953% |    10.465% |     0.581% |     0.000% |     0.000% |     0.321% |
-------------|------------|------------|------------|------------|------------|------------|
        6e+05 |        185 |         35 |          3 |          0 |          0 |        223 |
              |    82.960% |    15.695% |     1.345% |     0.000% |     0.000% |     0.416% |
-------------|------------|------------|------------|------------|------------|------------|
Column Total |      42648 |       8780 |       1842 |        260 |         23 |      53553 |
-------------|------------|------------|------------|------------|------------|------------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3537.903     d.f. =  60     p =  0



       Minimum expected frequency: 0.07387074
Cells with Expected Frequency < 5: 20 of 80 (25%)
```

(a) The relationship between the hardship index and income level is statistically significant according to the crosstabulation and the χ^2-statistics. As expected, the less economically privileged groups have higher levels of hardship which implies a negative association between variables.

(b) The correlation between variables `index` and `income` is negative and statistically significant.

```r
rcorr(as.matrix(x[c("index", "income")]))
```

```
       index income
index   1.00  -0.16
income -0.16   1.00

n
       index income
index  59731  53553
income 53553  53769
```

```
P
       index income
index          0
income  0
```