SMML Class 1 Lab

John Kubale

8/27/2024

I. Made-up income data 1

```
income1<-data.frame("id"=c(1:20),
                            "income_usd"=rnorm(20, mean = 10000,
                                               sd = 10000/3),
                            "educ_yr"=rep(12,20))
head(income1)
```

```
##   id income_usd educ_yr
## 1  1  14462.698      12
## 2  2   9643.608      12
## 3  3  15184.336      12
## 4  4   6014.221      12
## 5  5   9806.341      12
## 6  6  12910.121      12
```

```
summary(income1)
```

```
##        id          income_usd       educ_yr
##  Min.   : 1.00   Min.   : 6014   Min.   :12
##  1st Qu.: 5.75   1st Qu.: 8425   1st Qu.:12
##  Median :10.50   Median :10015   Median :12
##  Mean   :10.50   Mean   :10550   Mean   :12
##  3rd Qu.:15.25   3rd Qu.:12462   3rd Qu.:12
##  Max.   :20.00   Max.   :17592   Max.   :12
```

1. How do income and education look? Looks like income ranges from 4678 to 144114 with a mean of 10386. Education has no variation.
2. Can you study the relationship between income and education? Why or why not? No, education does not vary at all.

II. Made-up income data 2

```
income2 <- data.frame("id"=c(101:120),
                      "income_usd"=rep(10000, 20),
                      "educ_yr"=sample(0:16,20,replace=T) )
head(income2)
```

```
##    id income_usd educ_yr
## 1 101      10000      15
## 2 102      10000       1
## 3 103      10000      16
## 4 104      10000      12
## 5 105      10000       6
## 6 106      10000       2
```

```
summary(income2)
```

```
##        id          income_usd       educ_yr
##  Min.   :101.0   Min.   :10000   Min.   : 1.00
##  1st Qu.:105.8   1st Qu.:10000   1st Qu.: 3.75
##  Median :110.5   Median :10000   Median : 7.50
##  Mean   :110.5   Mean   :10000   Mean   : 8.00
##  3rd Qu.:115.2   3rd Qu.:10000   3rd Qu.:12.25
##  Max.   :120.0   Max.   :10000   Max.   :16.00
```

1. How do income and education look? Income does not vary, education varies from 0 to 16 with a mean of 8.15
2. Can you study the relationship between income and education? Why or why not? No, income does not vary and all values are set at 10000.

III. Wage data in R package ISLR2

```
data("Wage")
# run ?Wage in your console to see data description in help
# Can also see p.13 of R manual_ISLR.pdf for data description
dim(Wage)
```

```
## [1] 3000   11
```

```
names(Wage)
```

```
##  [1] "year"       "age"        "maritl"     "race"       "education"
##  [6] "region"     "jobclass"   "health"     "health_ins" "logwage"
## [11] "wage"
```

```
head(Wage)
```

```
##        year age           maritl     race       education             region
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
## 86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
## 161300 2003  45       2. Married 1. White 3. Some College 2. Middle Atlantic
## 155159 2003  43       2. Married 3. Asian 4. College Grad 2. Middle Atlantic
## 11443  2005  50      4. Divorced 1. White      2. HS Grad 2. Middle Atlantic
## 376662 2008  54       2. Married 1. White 4. College Grad 2. Middle Atlantic
##              jobclass          health health_ins  logwage      wage
## 231655  1. Industrial      1. <=Good      2. No 4.318063  75.04315
## 86582  2. Information 2. >=Very Good      2. No 4.255273  70.47602
## 161300  1. Industrial      1. <=Good     1. Yes 4.875061 130.98218
## 155159 2. Information 2. >=Very Good     1. Yes 5.041393 154.68529
## 11443  2. Information      1. <=Good     1. Yes 4.318063  75.04315
## 376662 2. Information 2. >=Very Good     1. Yes 4.845098 127.11574
```

```
summary(Wage)
```

```
##       year           age                         maritl           race
##  Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
##  1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
##  Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
##  Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
##  3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##                education                     region               jobclass
##  1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544
##  2. HS Grad        :971   1. New England       :   0   2. Information:1456
##  3. Some College   :650   3. East North Central:   0
##  4. College Grad   :685   4. West North Central:   0
##  5. Advanced Degree:426   5. South Atlantic    :   0
##                           6. East South Central:   0
##                           (Other)              :   0
##             health      health_ins      logwage           wage
##  1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09
##  2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38
##                                      Median :4.653   Median :104.92
##                                      Mean   :4.654   Mean   :111.70
##                                      3rd Qu.:4.857   3rd Qu.:128.68
##                                      Max.   :5.763   Max.   :318.34
##
```

1. What do you observe? Yearly data for health status by jobclass,lowage, wage, region, and demographic variables such as race, education, age, and marital status.

- What is the variable type for each variable? int, int, fct, fct, fct, fct, fct, fct, dbl, dbl.

2. What stories would you like to study using this data? The incremental change in health status across regional, occupational, and demographic varaibles.
3. How would you express those stories with formulas? health ~ age + marital + race + education + region + jobclass + year

IV. psid data in R package faraway

```
data("psid")
# Run ?psid in console to see data description in help
# Can also see p.78 of R manual_faraway.pdf for data description
head(psid)
```

```
##   age educ sex income year person
## 1  31   12   M   6000   68      1
## 2  31   12   M   5300   69      1
## 3  31   12   M   5200   70      1
## 4  31   12   M   6900   71      1
## 5  31   12   M   7500   72      1
## 6  31   12   M   8000   73      1
```

```
summary(psid)
```

```
##       age             educ       sex         income            year      
##  Min.   :25.00   Min.   : 3.00   F:732   Min.   :     3   Min.   :68.00  
##  1st Qu.:28.00   1st Qu.:10.00   M:929   1st Qu.:  4300   1st Qu.:73.00  
##  Median :34.00   Median :12.00           Median :  9000   Median :78.00  
##  Mean   :32.19   Mean   :11.84           Mean   : 13575   Mean   :78.61  
##  3rd Qu.:36.00   3rd Qu.:13.00           3rd Qu.: 18050   3rd Qu.:84.00  
##  Max.   :39.00   Max.   :16.00           Max.   :180000   Max.   :90.00  
##      person      
##  Min.   : 1.00  
##  1st Qu.:20.00  
##  Median :42.00  
##  Mean   :42.44  
##  3rd Qu.:63.00  
##  Max.   :85.00
```

1. What do you observe? Respondent level data, repeated measures for yearly income as well as the individual's age education and sex.

- What is the variable type for each variable? int, int, fct, int, int, int

2. What stories would you like to study using this data? How an individual's income bracket changes over time as a function of education.
3. How would you express those stories with formulas? Income ~ sex + educ + age +(year | person)

V. Fictitious data

```
fic_dat7 <- read_excel(
  # "C:/Users/jkubale/Dropbox (University of Michigan)/MPSDS 685/data/Fictitious Data.
  "~/UMD/classes/survmeth_615/class_1/Fictitious Data.xlsx",
  sheet=7)
head(fic_dat7)
```

```
## # A tibble: 6 x 5
##      ID Distress_2019 Distress_2021 Sex   Race
##   <dbl>         <dbl>         <dbl> <chr> <chr>
## 1     1             3             4 M     Non-Hispanic Black
## 2     2             1             1 F     Non-Hispanic White
## 3     3             6             8 M     Non-Hispanic Black
## 4     4             1             4 M     Hispanic
## 5     5             5             4 F     Non-Hispanic White
## 6     6             2             2 F     Non-Hispanic White
```

- The filepath that has been commented out (line 92) will also work to import the data (assuming that's where the file is located). Why might we prefer to use one way of writing the filepath over the other?

1. Compute the average of Distress_2019 and Distress_2021

```
mean(fic_dat7$Distress_2019)
```

```
## [1] 3.2
```

```
mean(fic_dat7$Distress_2021)
```

```
## [1] 4.15
```

- Can you say that there is a difference in Distress between 2019 and 2022? Why or why not? There is a difference of .95 points, but we do not know if this is a statistical difference yet.

2. Compute the average of Distress_2019 and Distress_2021 by Sex

```
cbind(aggregate(Distress_2019~Sex,fic_dat7,FUN=mean),
      aggregate(Distress_2021~Sex,fic_dat7,FUN=mean))
```

```
##   Sex Distress_2019 Sex Distress_2021
## 1   F      4.000000   F      4.444444
## 2   M      2.545455   M      3.909091
```

```
library(tidyverse)
fic_dat7 |> group_by(Sex) |> reframe(mean(Distress_2019), mean(Distress_2021))
```

```
## # A tibble: 2 x 3
##   Sex   `mean(Distress_2019)` `mean(Distress_2021)`
```

```
##   <chr>                <dbl>                <dbl>
## 1 F                     4                    4.44
## 2 M                     2.55                 3.91
```

- Can you say that there is a difference in Distress between Male and Female in 2019? How about in 2022? Why or why not? There is a difference in distress by Sex, yet we cannot determine statistical significance yet.