W5 practice

2023-02-09

0. data step

```
library(haven); library(psych); library(dplyr);
library(magrittr); library(ggplot2); library(gridExtra)
library(epitools); library(lsr); library(descr); library(epiR); library(epiDisplay)

  dat = read_sas("../SASlab/Choices.sas7bdat") # load data
  names(dat) # name of the variables
```

```
##  [1] "id"       "age"      "gender"   "race"     "married"  "religion"
##  [7] "educ"     "insure"   "qwb100"   "depress"  "health"   "died"
## [13] "livewill" "longwell" "pref"     "fpref"
```

```
  dat %>% dim # 2536, 16
```

```
## [1] 2536   16
```

```
  summary(dat) # get min, max, NA's
```

```
##        id              age           gender          race
##  Min.   :   1.0   Min.   :65.00   Min.   :1.000   Min.   :1.000
##  1st Qu.: 634.8   1st Qu.:69.00   1st Qu.:1.000   1st Qu.:1.000
##  Median :1268.5   Median :73.00   Median :1.000   Median :1.000
##  Mean   :1268.5   Mean   :73.88   Mean   :1.386   Mean   :1.311
##  3rd Qu.:1902.2   3rd Qu.:78.00   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2536.0   Max.   :99.00   Max.   :2.000   Max.   :2.000
##                   NA’s   :9                       NA’s   :12
##     married          religion            educ           insure
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :2.000   Median :1.0000
##  Mean   :0.5645   Mean   :0.8386   Mean   :1.977   Mean   :0.7222
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :3.000   Max.   :1.0000
##  NA’s   :1        NA’s   :21       NA’s   :16       NA’s   :23
##      qwb100          depress           health           died
##  Min.   :  0.00   Min.   : 0.000   Min.   :1.000   Min.   :0.00000
##  1st Qu.: 35.38   1st Qu.: 1.000   1st Qu.:2.000   1st Qu.:0.00000
##  Median : 48.89   Median : 4.000   Median :3.000   Median :0.00000
##  Mean   : 49.05   Mean   : 5.017   Mean   :2.916   Mean   :0.08162
##  3rd Qu.: 62.03   3rd Qu.: 7.000   3rd Qu.:4.000   3rd Qu.:0.00000
##  Max.   :100.00   Max.   :30.000   Max.   :5.000   Max.   :1.00000
##  NA’s   :94       NA’s   :56       NA’s   :2
##     livewill         longwell          pref           fpref
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.0000   Median :0.000   Median :1.000   Median :0.000
##  Mean   :0.1571   Mean   :0.121   Mean   :1.684   Mean   :1.058
##  3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :1.000   Max.   :6.000   Max.   :6.000
##  NA’s   :21       NA’s   :107                     NA’s   :464
```

```
  describe(dat)
```

```
##          vars    n    mean     sd  median trimmed    mad min  max range  skew
## id          1 2536 1268.50 732.22 1268.50 1268.50 939.97   1 2536  2535  0.00
## age         2 2527   73.88   5.73   73.00   73.39   5.93  65   99    34  0.74
## gender      3 2536    1.39   0.49    1.00    1.36   0.00   1    2     1  0.47
## race        4 2524    1.31   0.46    1.00    1.26   0.00   1    2     1  0.81
## married     5 2535    0.56   0.50    1.00    0.58   0.00   0    1     1 -0.26
## religion    6 2515    0.84   0.37    1.00    0.92   0.00   0    1     1 -1.84
## educ        7 2520    1.98   0.80    2.00    1.97   1.48   1    3     2  0.04
## insure      8 2513    0.72   0.45    1.00    0.78   0.00   0    1     1 -0.99
## qwb100      9 2442   49.05  19.27   48.89   49.03  20.02   0  100   100  0.03
## depress    10 2480    5.02   4.73    4.00    4.34   4.45   0   30    30  1.28
## health     11 2534    2.92   1.20    3.00    2.90   1.48   1    5     4 -0.03
## died       12 2536    0.08   0.27    0.00    0.00   0.00   0    1     1  3.05
## livewill   13 2515    0.16   0.36    0.00    0.07   0.00   0    1     1  1.88
## longwell   14 2429    0.12   0.33    0.00    0.03   0.00   0    1     1  2.32
## pref       15 2536    1.68   1.91    1.00    1.40   1.48   0    6     6  0.93
## fpref      16 2072    1.06   1.60    0.00    0.73   0.00   0    6     6  1.45
##          kurtosis    se
## id          -1.20 14.54
## age          0.21  0.11
## gender      -1.78  0.01
## race        -1.34  0.01
## married     -1.93  0.01
## religion     1.38  0.01
## educ        -1.45  0.02
## insure      -1.02  0.01
## qwb100      -0.59  0.39
## depress      1.73  0.09
## health      -0.91  0.02
## died         7.33  0.01
## livewill     1.55  0.01
## longwell     3.39  0.01
## pref        -0.32  0.04
## fpref        1.03  0.04
```

1. Create a new variables

```
  center =
    dat %>%
    mutate(agecnt = age - mean(dat$age, na.rm = TRUE), # na.rm - ignore NAs when calculate
           agemin = age - min(dat$age, na.rm =TRUE),
           nhb = ifelse(race == 1, 0, 1),
           race = factor(race, levels = c("1", "2"), labels = c("NHW", "NHB")))
```

2. Regression and partial correlation

```
# first model
 fit1 = lm(qwb100 ~ nhb + agemin, data = center)
 summary(fit1)
```

```
##
## Call:
## lm(formula = qwb100 ~ nhb + agemin, data = center)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.501 -14.066  -0.515  14.525  58.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.08854    0.73548  73.541  < 2e-16 ***
## nhb         -4.09586    0.83755  -4.890 1.07e-06 ***
## agemin      -0.42601    0.06768  -6.295 3.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.93 on 2420 degrees of freedom
##   (113 observations deleted due to missingness)
## Multiple R-squared:  0.02845,    Adjusted R-squared:  0.02765
## F-statistic: 35.44 on 2 and 2420 DF,  p-value: 6.781e-16
```

```
 coef(fit1) # call coefficients
```

```
## (Intercept)         nhb      agemin
##  54.0885417  -4.0958642  -0.4260118
```

```
 confint(fit1) # CIs
```

```
##                 2.5 %     97.5 %
## (Intercept) 52.646297 55.5307863
## nhb         -5.738257 -2.4534716
## agemin      -0.558720 -0.2933037
```

```
# second model
 fit2 = lm(qwb100 ~ religion + race + age + depress + married + health, data = center)
summary(fit2)
```

```
##
## Call:
## lm(formula = qwb100 ~ religion + race + age + depress + married +
##     health, data = center)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.363 -12.313  -0.585  11.855  58.561
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.32305    4.77320  19.551  < 2e-16 ***
## religion    -1.08280    0.98405  -1.100    0.271
## raceNHB     -0.20658    0.79188  -0.261    0.794
## age         -0.33033    0.06247  -5.288 1.35e-07 ***
## depress     -0.90808    0.08079 -11.240  < 2e-16 ***
## married      0.48675    0.73122   0.666    0.506
## health      -4.99345    0.32332 -15.444  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 2349 degrees of freedom
##   (180 observations deleted due to missingness)
## Multiple R-squared:  0.2315, Adjusted R-squared:  0.2295
## F-statistic: 117.9 on 6 and 2349 DF,  p-value: < 2.2e-16
```

```
tmp.r =
  center[,c('religion', 'race', 'age', 'depress', 'married', 'health', 'qwb100')] %>%
  lowerCor() # use corr matrix for partial.r function
```

```
##          relgn race* age   dprss marrd helth qw100
## religion  1.00
## race*     0.24  1.00
## age       0.09  0.11  1.00
## depress   0.05  0.11  0.02  1.00
## married  -0.10 -0.19 -0.20 -0.17  1.00
## health    0.17  0.20  0.08  0.42 -0.11  1.00
## qwb100   -0.10 -0.11 -0.13 -0.36  0.11 -0.42  1.00
```

```
psych::partial.r(tmp.r, x = 1:6, y = 7)
```

```
## partial correlations
##          religion race*   age depress married health
## religion     1.00  0.23  0.08    0.01   -0.09   0.14
## race*        0.23  1.00  0.09    0.07   -0.18   0.17
## age          0.08  0.09  1.00   -0.03   -0.19   0.02
## depress      0.01  0.07 -0.03    1.00   -0.14   0.31
## married     -0.09 -0.18 -0.19   -0.14    1.00  -0.07
## health       0.14  0.17  0.02    0.31   -0.07   1.00
```

```
# partial correlation, x = col numbers of predictors, col number of y
```

3. MULTICOLLINEARITY

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
vif(fit2) # EVALUATE MULTICOLLINEARITY
```

```
## religion     race      age  depress  married   health
## 1.086695 1.119981 1.049875 1.235077 1.101776 1.269129
```

```
tmp.r # check correlations
```

```
##             religion      race*         age     depress     married      health
## religion  1.00000000  0.2353271  0.08696591  0.04655067 -0.09856235  0.16561775
## race*     0.23532713  1.0000000  0.10682201  0.10885473 -0.18576436  0.19730679
## age       0.08696591  0.1068220  1.00000000  0.02346237 -0.19743916  0.07681428
## depress   0.04655067  0.1088547  0.02346237  1.00000000 -0.16700069  0.41646987
## married  -0.09856235 -0.1857644 -0.19743916 -0.16700069  1.00000000 -0.10657667
## health    0.16561775  0.1973068  0.07681428  0.41646987 -0.10657667  1.00000000
## qwb100   -0.09593462 -0.1099774 -0.13330167 -0.36406333  0.11239871 -0.42102728
##               qwb100
## religion -0.09593462
## race*    -0.10997744
## age      -0.13330167
## depress  -0.36406333
## married   0.11239871
## health   -0.42102728
## qwb100    1.00000000
```

```
      # or compare unadjusted coef with adjusted ones
```

4. WHAT IF WE USE A CLASS STATMENT TO CREATE AN INDICATOR VARIABLE FOR RACE

```
three =
 center %>%
 mutate(male = ifelse(gender == 2, 1, 0))

table(three$male, three$gender)
```

```
##
##        1    2
##   0 1556    0
##   1    0  980
```

```
table(three$nhb, three$race)
```

```
##
##       NHW  NHB
##   0 1738    0
##   1    0  786
```

```
fit3 = lm(qwb100 ~ agemin + depress + male + nhb + married, data = three %>% na.omit())
summary(fit3)
```

```
##
## Call:
## lm(formula = qwb100 ~ agemin + depress + male + nhb + married,
##     data = three %>% na.omit())
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.753 -12.717  -0.793  12.994  58.025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.87949    1.15048  52.047  < 2e-16 ***
## agemin      -0.34668    0.07516  -4.612 4.25e-06 ***
## depress     -1.35946    0.09145 -14.866  < 2e-16 ***
## male         1.29936    0.93713   1.387   0.1658
## nhb         -2.33942    0.91990  -2.543   0.0111 *
## married      0.10961    0.95449   0.115   0.9086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.42 on 1846 degrees of freedom
## Multiple R-squared:  0.1359, Adjusted R-squared:  0.1335
## F-statistic: 58.05 on 5 and 1846 DF,  p-value: < 2.2e-16
```

5. Partial F test

```
fit4 = lm(qwb100 ~ age + depress + nhb, data = three %>% na.omit())
anova(fit3, fit4) # partial F test
```

```
## Analysis of Variance Table
##
## Model 1: qwb100 ~ agemin + depress + male + nhb + married
## Model 2: qwb100 ~ age + depress + nhb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1846 560346
## 2   1848 561101 -2    -754.8 1.2433 0.2887
```