Solution in Stata: Exercises Day 2

PSY8003

Matthias Mittner

spring 2022

Exercise 1: Hierarchical regression

The second block of predictors improves the model significantly, the third does not. The coefficients are not much changed by including the additional predictors.

```
use "../data/workout.dta"
ssc install hireg
hireg whours (gender age) (educ marital) (health)
```

```
checking hireg consistency and verifying not already installed...
all files already exist and are up to date.


Model 1:
   Variables in Model:
   Adding            : gender age

      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(2, 207)       =      8.81
       Model |  857.314971         2  428.657485   Prob > F        =    0.0002
    Residual |  10070.6088       207  48.6502842   R-squared       =    0.0785
-------------+----------------------------------   Adj R-squared   =    0.0695
       Total |  10927.9238       209  52.2867168   Root MSE        =     6.975

------------------------------------------------------------------------------
      whours | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      gender |   2.217302   .9925996     2.23   0.027     .2604016    4.174203
```

```
         age |  -.1339799   .0400453    -3.35   0.001    -.2129287    -.055031
       _cons |   17.23131   1.731741     9.95   0.000      13.8172    20.64542
------------------------------------------------------------------------------

Model 2:
   Variables in Model: gender age
   Adding            : educ marital

      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(4, 205)       =      6.60
       Model |  1247.17003         4  311.792507   Prob > F        =    0.0001
    Residual |  9680.75378       205  47.2231892   R-squared       =    0.1141
-------------+----------------------------------   Adj R-squared   =    0.0968
       Total |  10927.9238       209  52.2867168   Root MSE        =    6.8719

------------------------------------------------------------------------------
      whours | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      gender |   2.050904   .9849872     2.08   0.039     .1088998    3.992908
         age |  -.1050521   .0471794    -2.23   0.027    -.1980712    -.012033
        educ |  -1.642637    .577278    -2.85   0.005      -2.7808   -.5044732
     marital |   .0344625   1.150991     0.03   0.976    -2.234835     2.30376
       _cons |   19.53729   2.459824     7.94   0.000      14.6875    24.38709
------------------------------------------------------------------------------
R-Square Diff. Model 2 - Model 1 = 0.036   F(2,205) =  4.128  p = 0.017

Model 3:
   Variables in Model: gender age  educ marital
   Adding            : health

      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(5, 204)       =      5.41
       Model |  1278.89155         5   255.77831   Prob > F        =    0.0001
    Residual |  9649.03226       204  47.2991777   R-squared       =    0.1170
-------------+----------------------------------   Adj R-squared   =    0.0954
       Total |  10927.9238       209  52.2867168   Root MSE        =    6.8774

------------------------------------------------------------------------------
      whours | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      gender |   2.118001   .9891784     2.14   0.033      .167677    4.068326
         age |  -.1133646   .0482961    -2.35   0.020    -.2085881   -.0181411
        educ |  -1.587641   .5816321    -2.73   0.007    -2.734422     -.44086
```

```
     marital |   .0194244   1.152063     0.02   0.987    -2.252053    2.290902
      health |   .3653454   .4461218     0.82   0.414    -.5142554    1.244946
       _cons |   17.84223   3.216314     5.55   0.000     11.50075    24.18371
------------------------------------------------------------------------------
R-Square Diff. Model 3 - Model 2 = 0.003   F(1,204) =  0.671  p = 0.414


Model  R2      F(df)             p         R2 change  F(df) change       p
   1:  0.078   8.811(2,207)      0.000
   2:  0.114   6.603(4,205)      0.000     0.036      4.128(2,205)       0.017
   3:  0.117   5.408(5,204)      0.000     0.003      0.671(1,204)       0.414
```

Exercise 2: Power analysis

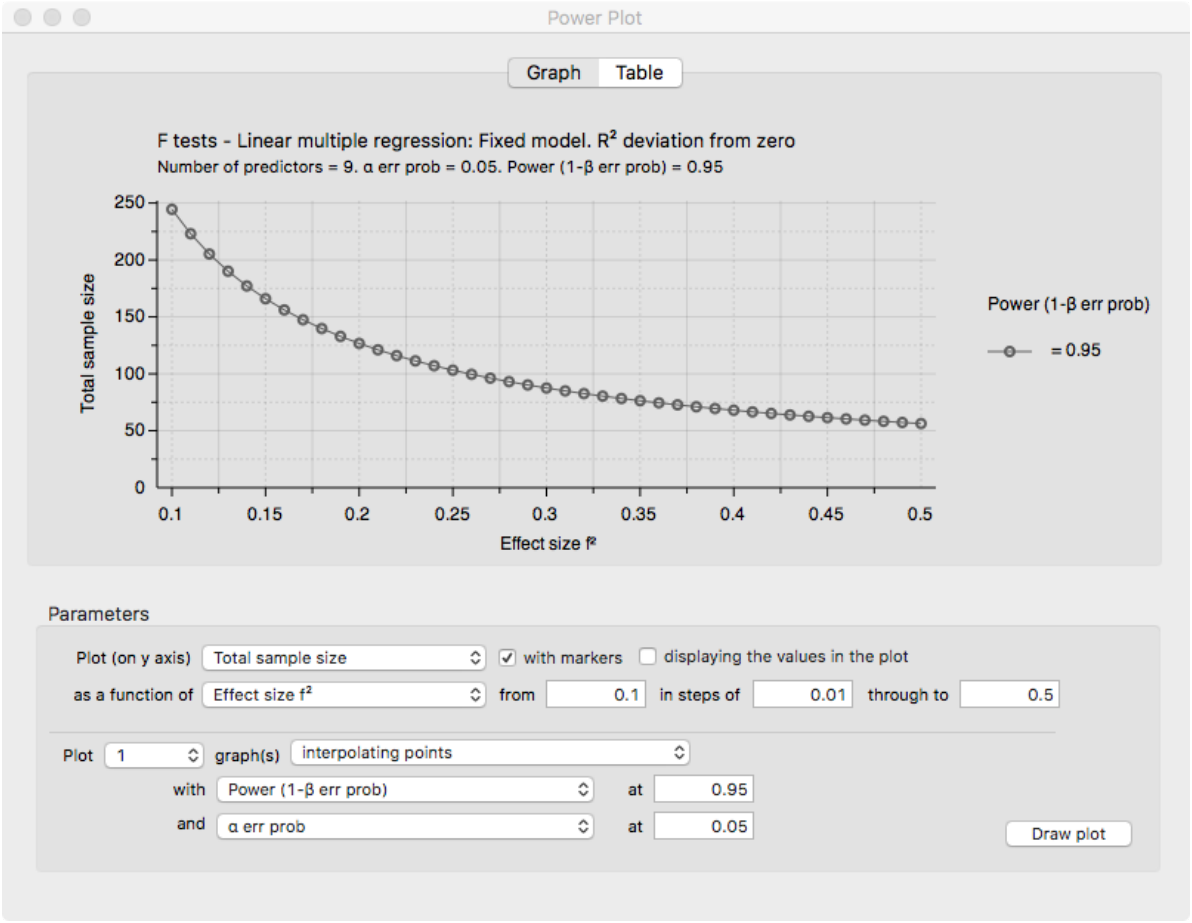
a)

In G*Power:

- Test family=F-Test
- Statistical-Test = Linear multiple Regression: Fixed model, R^2 deviation from 0
- Type: A priori
- Input parameters: - "Determine" -> From correlation coefficient: Squared multiple correlation=0.1 - α err prob=0.05 - Power=0.95 - Number of predictors=9

Total sample size=221.

b)



The critical f^2 for an $R^2 = 0.2$ is

$$f^2 = \frac{R^2}{1 - R^2} = 0.25$$

The graph "cuts" the $f^2 = 0.25$ line at about $N = 103$ (click on "Table" to get exact values).

c)

$N = 150$ corresponds to about $f^2 = 0.17$ which is

$$R^2 = \frac{f^2}{f^2 + 1} \approx 0.15$$

d)

In G*Power:

- Test family=F-Test
- Statistical-Test = Linear multiple Regression: Fixed model, R^2 increase
- Type: A priori
- Input parameters: - f^2=0.02 - α err prob=0.05 - Power=0.8 - Number of tested predictors=1 - Number of predictors=9 - Calculate -> $N = 395$

Exercise 3: Regression diagnostics

Two cases have large residuals >3; all of them occur for very high values of the dependent variables -> indicative of a problem with specification?

Leverage/Cooks-d pick up a few cases, but it does not look too severe (no Cooks'd close to 1).

No problem with multicollinearity.

There may be a slight misspecification (deviation from linearity), but its not severe.

Heteroscedasticity is present, variance seems to grow with the predicted value.

The residuals are definitely not normal, not surprising considering that this is a count-variable.

```
use "../data/workout.dta"
regress whours gender age educ marital health

* any residuals larger/smaller than expected?
predict resid, rstudent
list if abs(resid)>2.5 & resid !=.

* leverate, dfit and cooksd
* listing according to threshold for leverage (k=2, n=95)
predict lev, leverage
predict dfit, dfit
predict cooksd, cooksd
list if abs(lev)>(2*5+2)/210 & lev!=.

* normality of residuals
histogram resid
qnorm resid

swilk resid
kdensity resid, normal
```

```
* heteroscedasticity
rvfplot, yline(0)
estat imtest
estat hettest

* statistical regression checks in Stata
ssc install regcheck
regcheck
```

```
      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(5, 204)       =      5.41
       Model |  1278.89155         5   255.77831   Prob > F        =    0.0001
    Residual |  9649.03226       204  47.2991777   R-squared       =    0.1170
-------------+----------------------------------   Adj R-squared   =    0.0954
       Total |  10927.9238       209  52.2867168   Root MSE        =    6.8774

------------------------------------------------------------------------------
      whours | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      gender |   2.118001   .9891784     2.14   0.033      .167677    4.068326
         age |  -.1133646   .0482961    -2.35   0.020    -.2085881   -.0181411
        educ |  -1.587641   .5816321    -2.73   0.007    -2.734422     -.44086
     marital |   .0194244   1.152063     0.02   0.987    -2.252053    2.290902
      health |   .3653454   .4461218     0.82   0.414    -.5142554    1.244946
       _cons |   17.84223   3.216314     5.55   0.000     11.50075    24.18371
------------------------------------------------------------------------------
```

```
     +-------------------------------------------------------------------------------+
     | whours   gender   age            educ   marital             health      resid |
     |-------------------------------------------------------------------------------|
  3. |     48      men    20   secondary/high    single                  4   4.716255 |
  7. |     40      men    20   secondary/high    single                  5   3.328911 |
 35. |     36      men    21   secondary/high    single                  5   2.716904 |
 85. |     32    women    27   secondary/high   married                  4   2.610004 |
175. |     32      men    38       university   married   6=Very important   2.586674 |
     +-------------------------------------------------------------------------------+
```

```
     +-------------------------------------------------------------------------------
     | whours   gender   age                  educ   marital                  health
     |-------------------------------------------------------------------------------
 37. |     12    women    37   more than university   married                       2    .35
```

```
 44. |      4    women    28   more than university    single   1=Not important at all   -.95
 87. |      4      men    47         secondary/high    single                         3   -1.5
 93. |     16      men    76   more than university   married                         4    1.1
125. |     24      men    16         secondary/high    single   1=Not important at all    1.0
     |-------------------------------------------------------------------------------------
129. |      8    women    43             university   married   1=Not important at all   -.32
141. |     12    women    17         secondary/high    single   1=Not important at all   -.41
210. |     16    women    34             university    single   1=Not important at all    .73
     +-------------------------------------------------------------------------------------

(bin=14, start=-2.2836428, width=.49999268)

                  Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+--------------------------------------------------------
       resid |        210    0.95996      6.233     4.221    0.00001



Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      31.05     18    0.0284
            Skewness |      13.43      5    0.0197
            Kurtosis |       1.82      1    0.1768
---------------------+-----------------------------
               Total |      46.31     24    0.0041
---------------------------------------------------


Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of whours

H0: Constant variance

    chi2(1) =  30.51
Prob > chi2 = 0.0000
```

```
checking regcheck consistency and verifying not already installed...
all files already exist and are up to date.
```

Regression assumptions:	Test:
1) heterokedasticity problem	Breusch-Pagan hettest Chi2(1): 30.507 p-value: 0.000
2) no multicollinearity problem	Variance inflation factor age : 1.51 marital : 1.45 educ : 1.10 health : 1.07 gender : 1.03
3) residuals are not normally distributed	Shapiro-Wilk W normality test z: 3.964 p-value: 0.000
4) specification problem	Linktest t: 3.502 p-value: 0.001
5) functional form problem	Test for appropriate functional form F(3,201):4.140 p-value: 0.007
6) no influential observations	Cook's distance no distance is above the cutoff