Solution in R: Exercises Day 3

PSY8003

Matthias Mittner

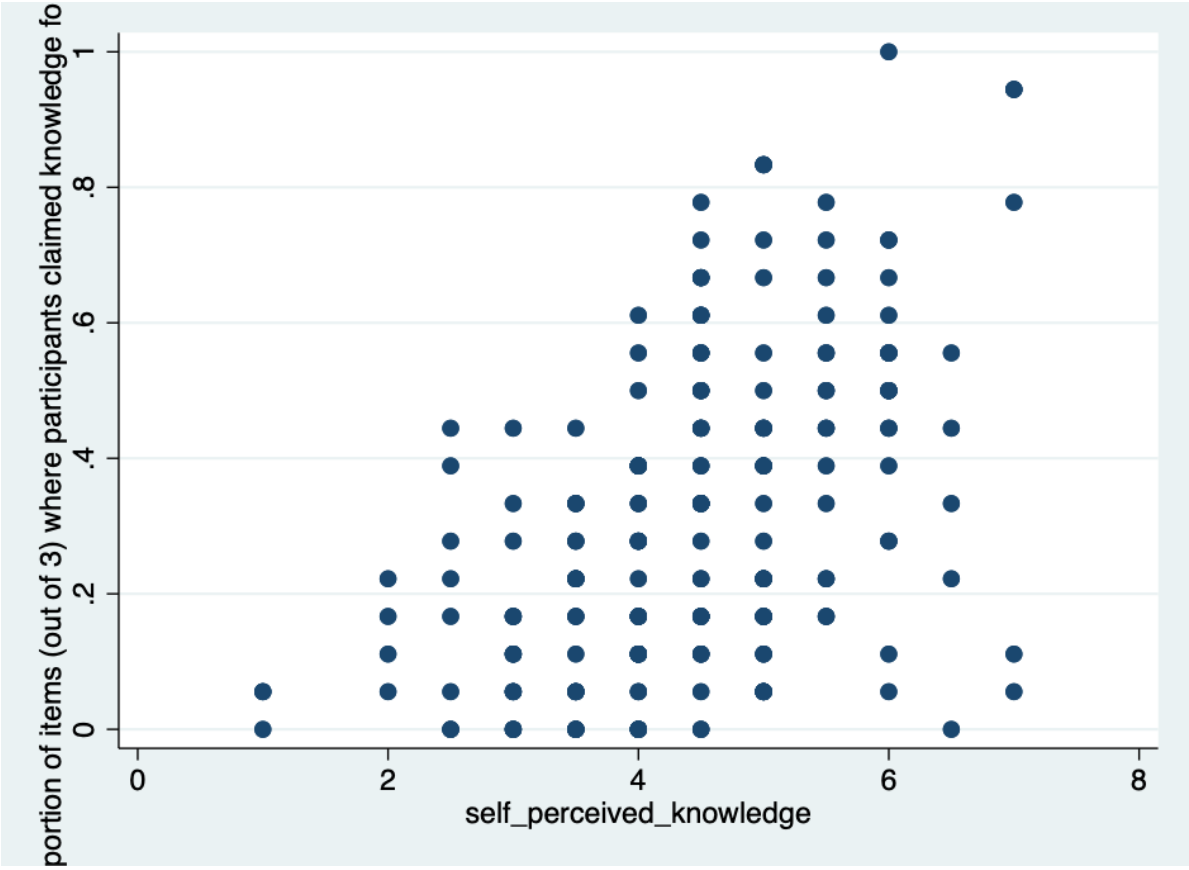
spring 2022

Exercise 1: Interactions

1.

There is a significant positive effect of self-perceived knowledge on overclaiming (β=0.1) and a negative effect of accuracy on overclaiming (β=-0.75).

```
scatter overclaiming_proportion self_perceived_knowledge
```




```
use "../data/atir2015.dta"
regress overclaiming_proportion accuracy self_perceived_knowledge
```

```
      Source |       SS           df       MS      Number of obs   =       202
-------------+----------------------------------   F(2, 199)       =    237.72
       Model |  7.64142253         2  3.82071127   Prob > F        =    0.0000
    Residual |  3.19838922       199  .016072308   R-squared       =    0.7049
-------------+----------------------------------   Adj R-squared   =    0.7020
       Total |  10.8398118       201  .053929412   Root MSE        =    .12678

------------------------------------------------------------------------------
 overclaiming_proportion | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------------------+----------------------------------------------------------------
               accuracy |   -.753986    .042195   -17.87   0.000    -.8371927   -.6707792
self_perceived_knowledge |   .0997649   .0076322    13.07   0.000     .0847145    .1148153
                   _cons |   .0889104   .0367371     2.42   0.016     .0164664    .1613543
------------------------------------------------------------------------------
```

2.

- Comparing mean overclaiming between the two possible orderings results in a significant difference in the (M1=0.34, M2=0.27).

```
use "../data/atir2015.dta"
mean overclaiming_proportion, over(order_of_tasks )
regress overclaiming_proportion order_of_tasks
```

```
Mean estimation                                                 Number of obs = 202

------------------------------------------------------------------------------
                                             |       Mean   Std. err.     [95% conf. interval]
---------------------------------------------+--------------------------------------------------
c.overclaiming_proportion@order_of_tasks     |
Self-Perceived Knowledge Measured First      |   .3437844   .0244073      .2956571    .3919116
            Overclaiming Measured First      |   .2722772   .0212595       .230357    .3141974
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       202
-------------+----------------------------------   F(1, 200)       =      4.88
       Model |  .258220266         1  .258220266   Prob > F        =    0.0283
    Residual |  10.5815915       200  .052907957   R-squared       =    0.0238
-------------+----------------------------------   Adj R-squared   =    0.0189
       Total |  10.8398118       201  .053929412   Root MSE        =    .23002

------------------------------------------------------------------------------
overclaiming~n | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
---------------+--------------------------------------------------------------
order_of_tasks |  -.0715072   .0323679    -2.21   0.028    -.1353333    -.007681
         _cons |   .4152915   .0511782     8.11   0.000     .3143735    .5162096
------------------------------------------------------------------------------
```

3.

- the effect is present when modulation by `order_of_task` is allowed, $beta = 0.11$
- the interaction is not significant ($p = .06$) but almost so. The interpretation is that the association between self-perceived knowledge and overclaiming is reduced by 0.03 when the order of presentation of the tests is switched

```
use "../data/atir2015.dta"
regress overclaiming_proportion accuracy c.self_perceived_knowledge##order_of_tasks
```

```
      Source |       SS           df       MS      Number of obs   =       202
-------------+----------------------------------   F(4, 197)       =    120.51
       Model |  7.69507396         4  1.92376849   Prob > F        =    0.0000
    Residual |   3.1447378       197  .015963136   R-squared       =    0.7099
-------------+----------------------------------   Adj R-squared   =    0.7040
       Total |  10.8398118       201  .053929412   Root MSE        =    .12635

------------------------------------------------------------------------------------------
                  overclaiming_proportion | Coefficient  Std. err.      t    P>|t|     [95% c
------------------------------------------+-----------------------------------------------
                                 accuracy |  -.7570552   .0422176   -17.93   0.000    -.8403
               self_perceived_knowledge |   .1149759   .0113812    10.10   0.000     .0925
                                          |
                           order_of_tasks |
            Overclaiming Measured First  |   .1270156    .071707     1.77   0.078    -.0143
                                          |
order_of_tasks#c.self_perceived_knowledge |
            Overclaiming Measured First  |  -.0285865   .0155931    -1.83   0.068    -.0593
                                          |
                                   _cons |   .0187424   .0552139     0.34   0.735    -.0901
------------------------------------------------------------------------------------------
```

4.

- FINRA has a mean of 3.7 and an SD of 1.9. Pretty high scores given that 5 is max.
- When controling for actual knowledge, the effect of self-perceived knowledge on overclaiming is still present but slightly reduced, $\beta = 0.09$
- there is also a weak effect of actual knowledge on overclaiming, $\beta = 0.018$

```
use "../data/atir2015.dta"
mean FINRA_score
regress overclaiming_proportion accuracy c.self_perceived_knowledge FINRA_score
```

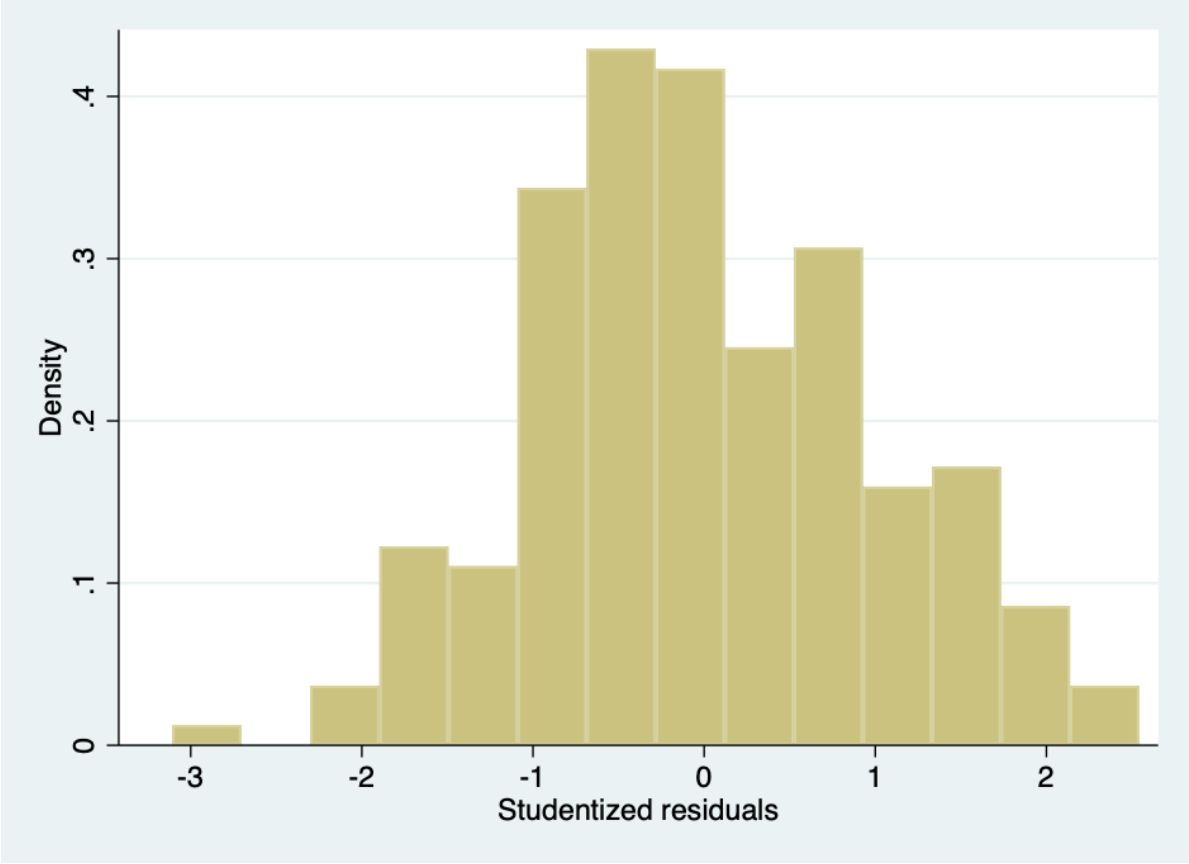
```
Mean estimation                           Number of obs = 202

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
 FINRA_score |    3.69802   .0837197      3.532938    3.863101
--------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       202
```

```
-------------+----------------------------------   F(3, 198)       =    162.87
       Model |  7.71385367         3  2.57128456   Prob > F        =    0.0000
    Residual |  3.12595808       198  .015787667   R-squared       =    0.7116
-------------+----------------------------------   Adj R-squared   =    0.7073
       Total |  10.8398118       201  .053929412   Root MSE        =    .12565

------------------------------------------------------------------------------
 overclaiming_proportion | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------------------+----------------------------------------------------------------
               accuracy |  -.7932189   .0456551   -17.37   0.000    -.8832516   -.7031862
self_perceived_knowledge |   .0940692   .0080181    11.73   0.000     .0782573    .1098811
            FINRA_score |   .0183697   .0085762     2.14   0.033     .0014571    .0352822
                  _cons |   .0577868   .0392027     1.47   0.142    -.0195216    .1350953
------------------------------------------------------------------------------
```

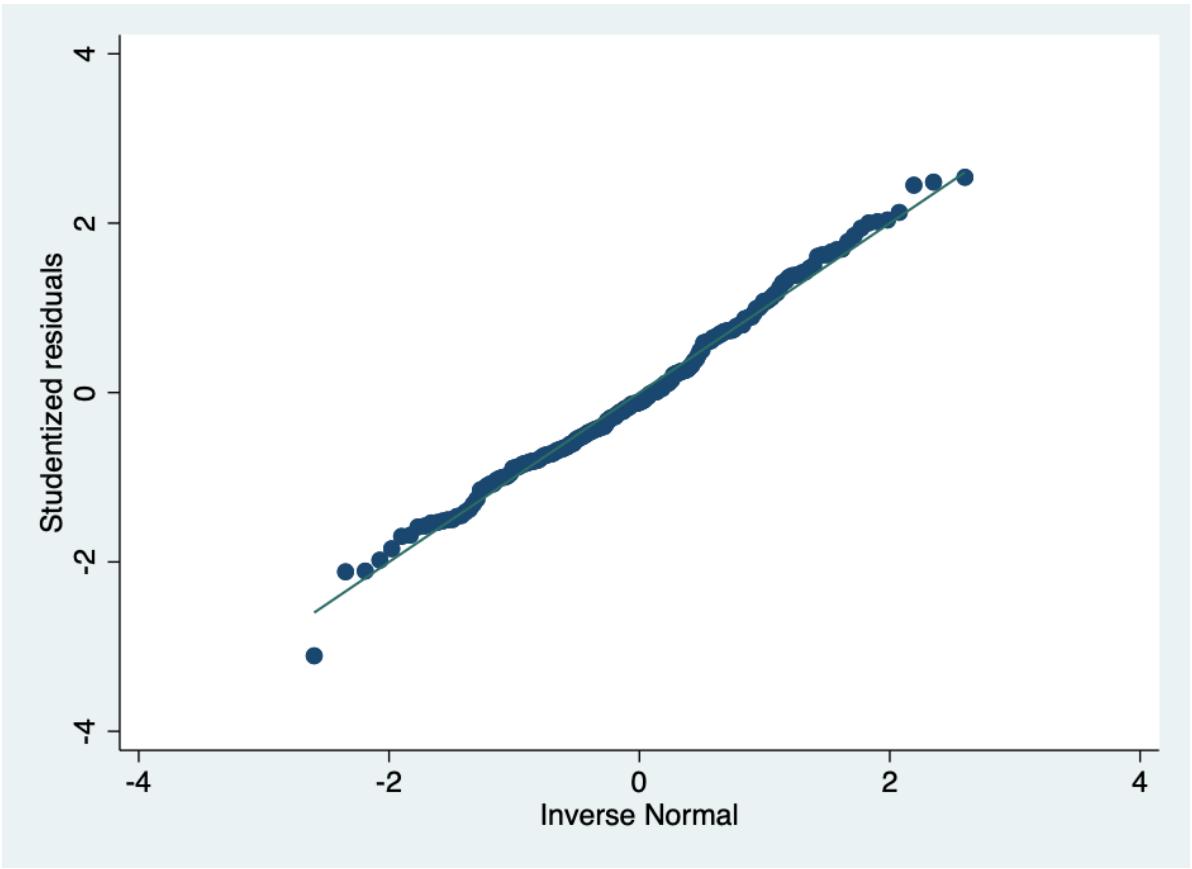
5.

- the histogram of the residuals does not show a strong departure from the normal distribution
- nor does the QQ-plot
- the predicted vs. residuals plot shows some heterogeneity in variance (increasing with predicted value)
- the "stripe"-structure comes from the discrete nature of the `overclaiming_proportion` variable

```
use "../data/atir2015.dta"
quietly regress overclaiming_proportion accuracy c.self_perceived_knowledge FINRA_score
quietly histogram resid
quietly graph export pics/ex3_histresid.png, replace
```




```
use "../data/atir2015.dta"
quietly regress overclaiming_proportion accuracy c.self_perceived_knowledge FINRA_score
qnorm resid
quietly graph export pics/ex3_qqresid.png, replace
```




```
use "../data/atir2015.dta"
quietly regress overclaiming_proportion accuracy c.self_perceived_knowledge FINRA_score

* statistical regression checks in Stata
regcheck

predict resid, rstudent

swilk resid

* heteroscedasticity
estat imtest
estat hettest
```

```
    Regression assumptions:                    |  Test:
|-------------------------------------------------------------------------------------
  1) no heterokedasticity problem              |  Breusch-Pagan hettest
```

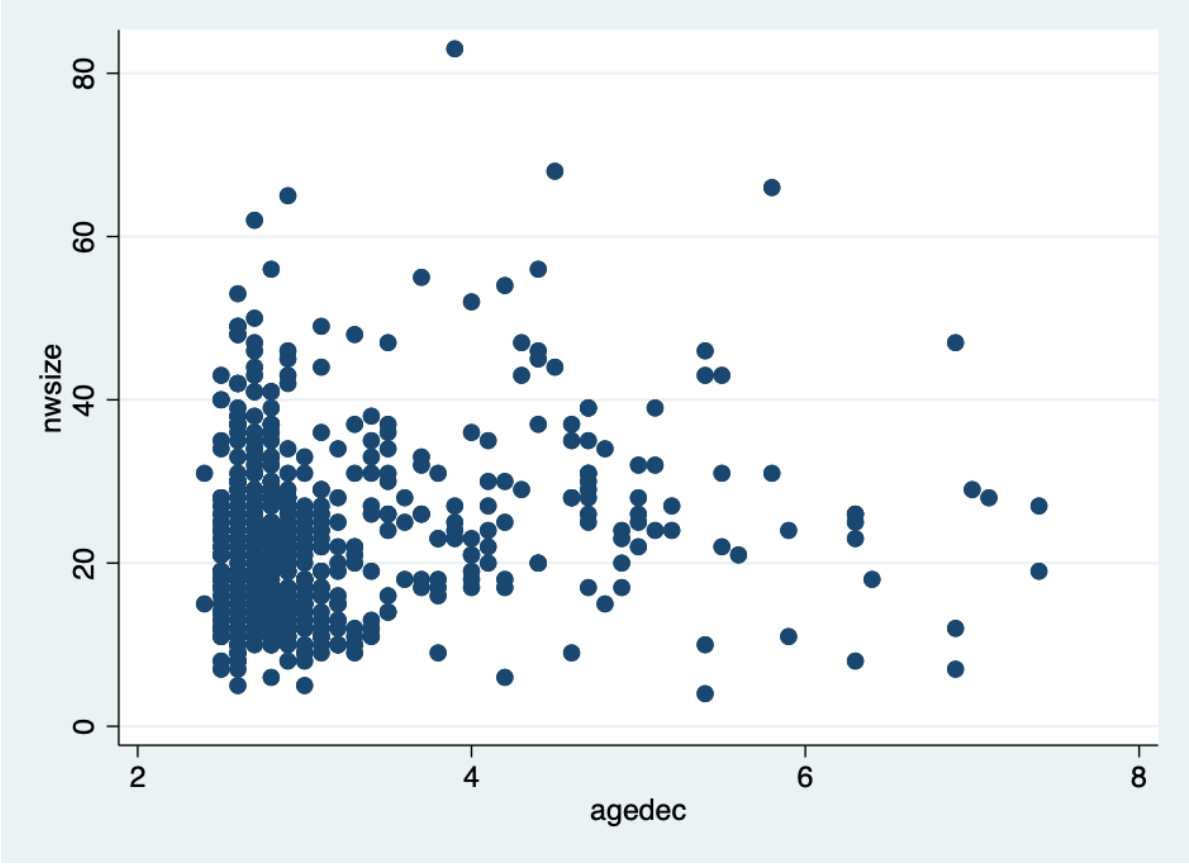
```
                                                  |  Chi2(1): 2.790
                                                  |  p-value: 0.095
|-------------------------------------------------------------------------------------------
  2) no multicollinearity problem                 |  Variance inflation factor
                                                  |  FINRA_score : 1.33
                                                  |  accuracy : 1.19
                                                  |  self_perceived_knowledge : 1.12
|-------------------------------------------------------------------------------------------
  3) residuals are normally distributed           |  Shapiro-Wilk W normality test
                                                  |  z: 0.792
                                                  |  p-value: 0.214
|-------------------------------------------------------------------------------------------
  4) specification problem                        |  Linktest
                                                  |  t: 3.774
                                                  |  p-value: 0.000
|-------------------------------------------------------------------------------------------
  5) functional form problem                      |  Test for appropriate functional form
                                                  |  F(3,195):6.627
                                                  |  p-value: 0.000
|-------------------------------------------------------------------------------------------
  6) no influential observations                  |  Cook's distance
                                                  |  no distance is above the cutoff
+-------------------------------------------------------------------------------------------

                  Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+--------------------------------------------------------
       resid |        202    0.99061      1.413     0.795    0.21322


Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |      10.88      9    0.2841
            Skewness |       7.36      3    0.0613
            Kurtosis |       0.42      1    0.5174
---------------------+-----------------------------
               Total |      18.66     13    0.1342
---------------------------------------------------
```

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

H0: Constant variance

    chi2(1) =   2.79
Prob > chi2 = 0.0949
```

Exercise 2: Nonlinear regression

```
use "../data/explorepenguin.dta"
gen ageyears=2022-age
scatter nwsize agedec
quietly graph export pics/ex3_nwsize.png, replace
```

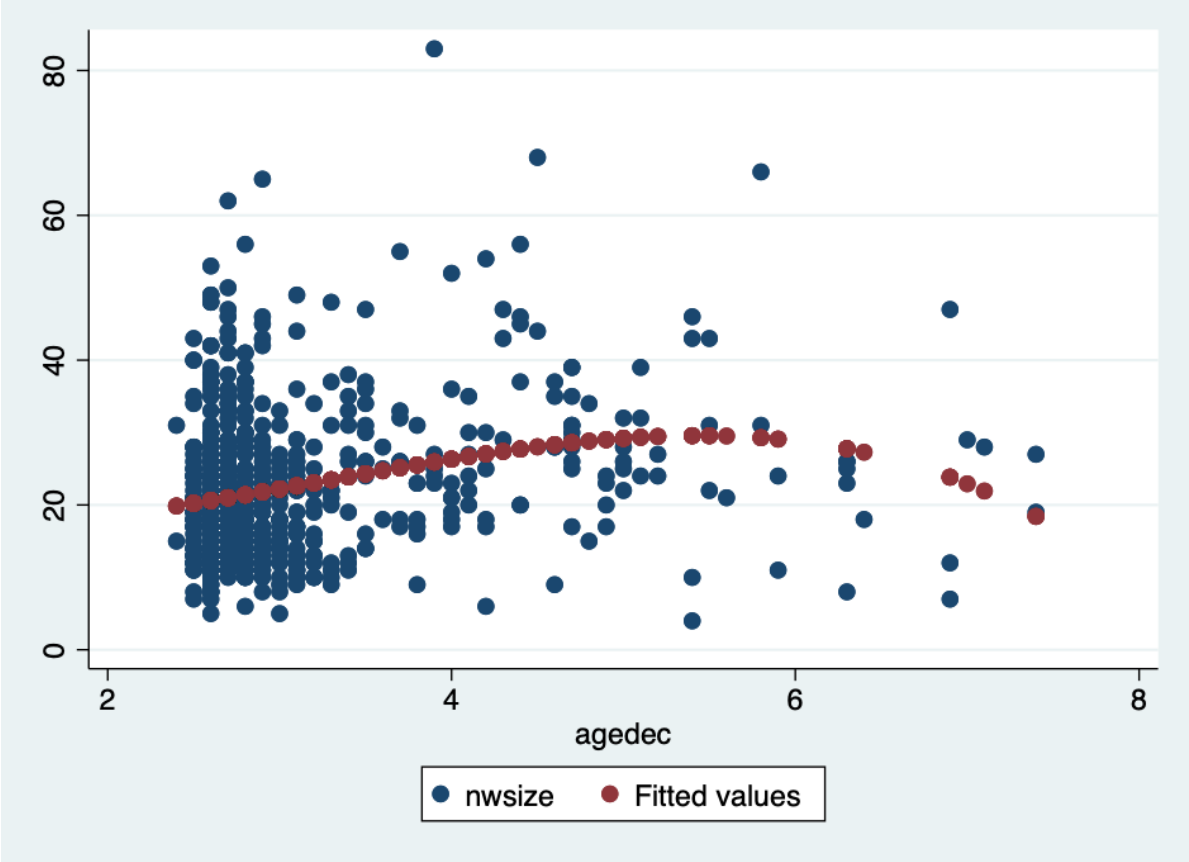
```
(22 missing values generated)

(22 missing values generated)
```




```
use "../data/explorepenguin.dta"
gen ageyears=2022-age
regress nwsize agedec
fp <agedec>, scale center: regress nwsize <agedec>
predict fpfit
scatter nwsize fpfit agedec
quietly graph export pics/ex3_nwsize_fp.png, replace
```

```
(22 missing values generated)

(22 missing values generated)

      Source |       SS           df       MS      Number of obs   =       713
-------------+----------------------------------   F(1, 711)       =     27.79
       Model |  2808.57593         1  2808.57593   Prob > F        =    0.0000
    Residual |  71868.6274       711  101.081051   R-squared       =    0.0376
-------------+----------------------------------   Adj R-squared   =    0.0363
```

```
       Total |  74677.2034       712  104.883713   Root MSE        =    10.054

------------------------------------------------------------------------------
      nwsize | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      agedec |   2.418216   .4587613     5.27   0.000     1.517527    3.318904
       _cons |   14.80045    1.46892    10.08   0.000     11.91652    17.68439
------------------------------------------------------------------------------

(fitting 44 models)
(....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%)

Fractional polynomial comparisons:
------------------------------------------------------------------------------
             | Test              Residual   Deviance
      agedec |   df   Deviance   std. dev.      diff.       P   Powers
-------------+----------------------------------------------------------------
     omitted |    4   5339.889     10.241     40.395    0.000
      linear |    3   5312.556     10.054     13.062    0.005   1
       m = 1 |    2   5308.016     10.022      8.522    0.015   -1
       m = 2 |    0   5299.494      9.969      0.000       --   3 3
------------------------------------------------------------------------------
Note: Test df is degrees of freedom, and P = P > F is sig. level for tests
      comparing models vs. model with m = 2 based on deviance difference,
      F(df, 708).

      Source |       SS           df       MS      Number of obs   =       713
-------------+----------------------------------   F(2, 710)       =     20.69
       Model |  4113.20623         2  2056.60312   Prob > F        =    0.0000
    Residual |  70563.9971       710  99.3859115   R-squared       =    0.0551
-------------+----------------------------------   Adj R-squared   =    0.0524
       Total |  74677.2034       712  104.883713   Root MSE        =    9.9692

------------------------------------------------------------------------------
      nwsize | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    agedec_1 |   .5223527   .0984594     5.31   0.000     .3290463     .715659
    agedec_2 |  -.2576945   .0515395    -5.00   0.000    -.3588825   -.1565066
       _cons |   22.61097   .3949427    57.25   0.000     21.83557    23.38636
------------------------------------------------------------------------------

(option xb assumed; fitted values)
(48 missing values generated)
```



Exercise 2: Splines

- there is no "correct" solution for the parameter settings at this point
- going up with knots shows an earlier peak in the data (early twenties) which might reflect university/educational setting which then goes down before the "final" social network is established
- it's hard/impossible to interpret the regression coefficients properly