AERA Data Analysis

February 4, 2023

[44]:
```
import pandas as pd
import numpy as np
```

[45]:
```
data = pd.read_csv("els_extracted_data_v4.csv")
data
```

[45]:

	STU_ID	SCH_ID	F3ERN2011	F3C02	F3EVRGED	F3EVERDO	BYMOTHED	BYS14	\
0	101101	1011	4000	-7	0	0	1	2	
1	101102	1011	3000	20	0	0	5	2	
2	101104	1011	37000	50	0	0	2	2	
3	101105	1011	1500	25	0	0	2	2	
4	101106	1011	48000	28	0	0	1	2	
…	…	…	…	…	…	…	…	…	
16192	461230	4612	-4	-4	0	0	2	2	
16193	461231	4612	7000	30	0	1	2	2	
16194	461232	4612	-4	-4	0	1	1	2	
16195	461233	4612	20000	40	0	0	1	2	
16196	461234	4612	36000	36	1	1	3	2	

	BYRACE	BYP61	F3ATTAINMENT	BYTXCSTD	BYSES1	F3JUNEDSTAT	F3REGION
0	5	1	3	56.21	-0.25	3	1
1	2	0	10	57.66	0.58	3	1
2	7	0	6	66.50	-0.85	3	1
3	3	-4	4	46.46	-0.80	2	1
4	4	0	4	36.17	-1.41	1	1
…	…	…	…	…	…	…	
16192	4	-4	-4	38.04	-0.34	-4	-4
16193	4	0	1	33.81	-1.08	4	3
16194	5	1	-4	37.94	-1.54	-4	-4
16195	4	0	4	45.93	-1.41	1	3
16196	5	-7	4	62.62	-0.40	3	3

[16197 rows x 15 columns]

[46]:
```
# Filter data
data = data[data.F3ERN2011 > 0]
data = data[data.F3ERN2011 < 200000]
data = data[data.F3C02 >= 0]
```

```
data = data[data.F3JUNEDSTAT >= 3]
data = data[data.BYS14 >= 0]
data = data[data.BYRACE >= 0]
data = data[data.BYTXCSTD >= 0]
data = data[data.F3REGION >= 0]
data = data[data.BYP61 >= -0.25]
data = data[data.BYMOTHED >= 0]
data = data[data.BYTXCSTD >= 0]
data = data[data.BYTXCSTD >= -3]
```

[47]:
```
data
```

[47]:

	STU_ID	SCH_ID	F3ERN2011	F3C02	F3EVRGED	F3EVERDO	BYMOTHED	BYS14	\
1	101102	1011	3000	20	0	0	5	2	
2	101104	1011	37000	50	0	0	2	2	
5	101107	1011	35000	40	0	0	2	1	
7	101109	1011	68000	40	0	0	2	1	
10	101112	1011	18000	1	0	0	6	1	
…	…	…	…	…	…	…	…	…	
16181	461205	4612	100	6	0	1	2	2	
16182	461207	4612	29000	40	0	0	3	2	
16185	461214	4612	15000	44	0	1	2	1	
16188	461220	4612	10000	40	0	1	4	2	
16193	461231	4612	7000	30	0	1	2	2	

	BYRACE	BYP61	F3ATTAINMENT	BYTXCSTD	BYSES1	F3JUNEDSTAT	F3REGION
1	2	0	10	57.66	0.58	3	1
2	7	0	6	66.50	-0.85	3	1
5	4	0	3	30.72	-1.07	3	1
7	7	0	6	68.39	-0.16	3	1
10	3	0	3	58.06	-0.18	3	1
…	…	…	…	…	…	…	…
16181	5	1	1	34.70	0.56	4	3
16182	5	1	5	39.53	-0.21	3	3
16185	5	0	2	38.00	-0.60	4	3
16188	4	0	3	45.82	-0.28	3	3
16193	4	0	1	33.81	-1.08	4	3

[6080 rows x 15 columns]

[48]:
```
# Create dummies for BYSEX
data["female"] = (data["BYS14"] == 2).astype(int)

# Create dummies for BYRACE
race_dummies = pd.get_dummies(data["BYRACE"], prefix="race")
data = pd.concat([data, race_dummies], axis=1)
```

```
# Create dummies for BYP61
data["no_parent"] = (data["BYP61"] == 1).astype(int)

# Create dummies for BYMOTHED
moth_dummies = pd.get_dummies(data["BYMOTHED"], prefix="moth_ed")
data = pd.concat([data, moth_dummies], axis=1)

# Create dummies for F3REGION
region_dummies = pd.get_dummies(data["F3REGION"], prefix="region")
data = pd.concat([data, region_dummies], axis=1)

# Create dummies for high_school_grad
data["high_school_grad"] = (data["F3EVERDO"] == 0).astype(int)

# Create dummies for F3EVRGED
data["ged"] = (data["F3EVRGED"] == 1).astype(int)
```

[49]:
```
def post_sec_edu(value):
    if value == 4:
        return 1
    elif value == 5:
        return 2
    elif value == 6:
        return 4
    elif value == 7:
        return 5
    elif value == 8:
        return 6
    elif value == 10:
        return 8
    else:
        return 0

data['post_sec_edu'] = data['F3ATTAINMENT'].apply(post_sec_edu)
```

[50]:
```
data
```

[50]:

	STU_ID	SCH_ID	F3ERN2011	F3C02	F3EVRGED	F3EVERDO	BYMOTHED	BYS14	\
1	101102	1011	3000	20	0	0	5	2	
2	101104	1011	37000	50	0	0	2	2	
5	101107	1011	35000	40	0	0	2	1	
7	101109	1011	68000	40	0	0	2	1	
10	101112	1011	18000	1	0	0	6	1	
…	…	…	…	…	…	…	…	…	
16181	461205	4612	100	6	0	1	2	2	
16182	461207	4612	29000	40	0	0	3	2	
16185	461214	4612	15000	44	0	1	2	1	

```
16188  461220    4612      10000     40          0          1          4       2
16193  461231    4612       7000     30          0          1          2       2

       BYRACE  BYP61  …  moth_ed_6  moth_ed_7  moth_ed_8  region_1  \
1           2      0  …          0          0          0         1
2           7      0  …          0          0          0         1
5           4      0  …          0          0          0         1
7           7      0  …          0          0          0         1
10          3      0  …          1          0          0         1
…         …    …  …  …          …          …          …         …
16181       5      1  …          0          0          0         0
16182       5      1  …          0          0          0         0
16185       5      0  …          0          0          0         0
16188       4      0  …          0          0          0         0
16193       4      0  …          0          0          0         0

       region_2  region_3  region_4  high_school_grad  ged  post_sec_edu
1             0         0         0                 1    0             8
2             0         0         0                 1    0             4
5             0         0         0                 1    0             0
7             0         0         0                 1    0             4
10            0         0         0                 1    0             0
…           …         …         …               …  …    …             …
16181         0         1         0                 0    0             0
16182         0         1         0                 1    0             2
16185         0         1         0                 0    0             0
16188         0         1         0                 0    0             0
16193         0         1         0                 0    0             0

[6080 rows x 39 columns]
```

[51]:
```
# Model 1
import statsmodels.api as sm

# Define the formula for the model1
formula = 'np.log(F3ERN2011) ~ ged + high_school_grad + female + race_3 +␣
 ↪race_4 + race_7 + race_6 + BYSES1 + no_parent + BYTXCSTD'

# Fit the multilevel model using the formula
model = sm.MixedLM.from_formula(formula, data, groups=data["SCH_ID"])
result = model.fit()

# Print the summary of the model
print(result.summary())
```

```
            Mixed Linear Model Regression Results
============================================================
Model:            MixedLM Dependent Variable: np.log(F3ERN2011)
```

```
No. Observations: 6080    Method:             REML
No. Groups:       739     Scale:              0.8662
Min. group size:  1       Log-Likelihood:     -8263.9232
Max. group size:  25      Converged:          Yes
Mean group size:  8.2
------------------------------------------------------------
                   Coef.  Std.Err.    z    P>|z| [0.025 0.975]
------------------------------------------------------------
Intercept           9.094    0.094  96.291 0.000  8.909  9.279
ged                -0.121    0.080  -1.505 0.132 -0.279  0.037
high_school_grad    0.252    0.059   4.271 0.000  0.137  0.368
female             -0.254    0.024 -10.461 0.000 -0.302 -0.206
race_3             -0.151    0.052  -2.901 0.004 -0.254 -0.049
race_4              0.079    0.061   1.303 0.193 -0.040  0.198
race_7              0.058    0.036   1.603 0.109 -0.013  0.129
race_6              0.023    0.068   0.334 0.738 -0.111  0.156
BYSES1              0.078    0.019   4.026 0.000  0.040  0.116
no_parent          -0.103    0.028  -3.693 0.000 -0.158 -0.048
BYTXCSTD            0.016    0.002  10.467 0.000  0.013  0.019
Group Var           0.015    0.007
============================================================
```

[52]:
```
# model 2
# Define the formula for the model 2
formula = 'np.log(F3ERN2011) ~ ged + high_school_grad + female + race_3 +␣
 ↪race_4 + race_7 + race_6 + BYSES1 + no_parent + BYTXCSTD + post_sec_edu'

# Fit the multilevel model using the formula
model = sm.MixedLM.from_formula(formula, data, groups=data["SCH_ID"])
result = model.fit()

# Print the summary of the model
print(result.summary())
```

```
            Mixed Linear Model Regression Results
============================================================
Model:            MixedLM Dependent Variable: np.log(F3ERN2011)
No. Observations: 6080    Method:             REML
No. Groups:       739     Scale:              0.8566
Min. group size:  1       Log-Likelihood:     -8228.8610
Max. group size:  25      Converged:          Yes
Mean group size:  8.2
------------------------------------------------------------
                   Coef.  Std.Err.    z    P>|z| [0.025 0.975]
------------------------------------------------------------
Intercept           9.227    0.095  97.138 0.000  9.041  9.413
ged                -0.100    0.080  -1.246 0.213 -0.256  0.057
```

```
high_school_grad     0.208    0.059   3.524 0.000  0.092  0.323
female              -0.288    0.024 -11.790 0.000 -0.336 -0.240
race_3              -0.145    0.052  -2.796 0.005 -0.246 -0.043
race_4               0.104    0.060   1.728 0.084 -0.014  0.222
race_7               0.083    0.036   2.288 0.022  0.012  0.153
race_6               0.052    0.068   0.768 0.442 -0.081  0.185
BYSES1               0.039    0.020   2.007 0.045  0.001  0.078
no_parent           -0.075    0.028  -2.680 0.007 -0.130 -0.020
BYTXCSTD             0.011    0.002   7.165 0.000  0.008  0.014
post_sec_edu         0.057    0.006   8.881 0.000  0.045  0.070
Group Var            0.013    0.007
================================================================
```

[53]:
```
# Create dummies for F3EVRGED
data["GEDT"] = (data["F3EVRGED"]*data["BYTXCSTD"]).astype(int)
```

[55]:
```
# model 3
# Define the formula for the model 3
formula = 'np.log(F3ERN2011) ~ ged + high_school_grad + female + race_3 +␣
 ↪race_4 + race_7 + race_6 + BYSES1 + no_parent + BYTXCSTD + GEDT'

# Fit the multilevel model using the formula
model = sm.MixedLM.from_formula(formula, data, groups=data["SCH_ID"])
result = model.fit()

# Print the summary of the model
print(result.summary())
```

```
            Mixed Linear Model Regression Results
================================================================
Model:            MixedLM Dependent Variable: np.log(F3ERN2011)
No. Observations: 6080    Method:             REML
No. Groups:       739     Scale:              0.8661
Min. group size:  1       Log-Likelihood:     -8266.6497
Max. group size:  25      Converged:          Yes
Mean group size:  8.2
----------------------------------------------------------------
                    Coef.  Std.Err.    z    P>|z| [0.025 0.975]
----------------------------------------------------------------
Intercept            9.077    0.095  95.470 0.000  8.891  9.263
ged                  0.435    0.366   1.189 0.234 -0.282  1.151
high_school_grad     0.249    0.059   4.201 0.000  0.133  0.365
female              -0.254    0.024 -10.463 0.000 -0.302 -0.206
race_3              -0.151    0.052  -2.893 0.004 -0.253 -0.049
race_4               0.077    0.061   1.280 0.201 -0.041  0.196
race_7               0.059    0.036   1.621 0.105 -0.012  0.130
race_6               0.025    0.068   0.369 0.712 -0.108  0.158
```

```
BYSES1               0.077    0.019   3.983 0.000  0.039  0.115
no_parent           -0.103    0.028  -3.680 0.000 -0.158 -0.048
BYTXCSTD             0.016    0.002  10.582 0.000  0.013  0.019
GEDT                -0.012    0.008  -1.559 0.119 -0.027  0.003
Group Var            0.015    0.007
============================================================
```