growth

February 17, 2020

1 GEE analysis of growth trajectories of children

GEE is commonly used in longitudinal data analysis. Here we consider a dataset in which repeated measures of weight were made on young children over several years in early childhood. GEE allows us to use linear modeling techniques similar to OLS, and still rigorously account for the repeated measures aspect of the data.

The data we will use are obtained from this page: http://www.bristol.ac.uk/cmm/learning/support/datasets

These are the packages we will be using:

[1]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
```

```
/nfs/kshedden/python3/lib/python3.7/site-
packages/statsmodels/compat/pandas.py:23: FutureWarning: The Panel class is
removed from pandas. Accessing it from the top-level namespace will also be
removed in the next version
  data_klasses = (pandas.Series, pandas.DataFrame, pandas.Panel)
```

The data are in "fixed width" format, so we use some special techniques for reading them:

[2]:
```
colspecs = [(0, 4), (4, 7), (7, 12), (12, 16), (16, 17)]
df = pd.read_fwf("../data/growth/ASIAN.DAT", colspecs=colspecs, header=None)
df.columns = ["Id", "Age", "Weight", "BWeight", "Gender"]
df["Female"] = 1*(df.Gender == 2)
df = df.dropna()
```

Some of the analyses below will use logged data:

[3]:
```
df["LogWeight"] = np.log(df.Weight) / np.log(2)
df["LogBWeight"] = np.log(df.BWeight) / np.log(2)
```

The first model that we consider treats weight as a linear function of age, and ignores the repeated measures structure. The point estimates from this model are valid, but the standard errors are not.

[4]:
```
model0 = sm.GLM.from_formula("Weight ~ Age + BWeight + Female", data=df)
rslt0 = model0.fit()
print(rslt0.summary())
```

```
                  Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 Weight   No. Observations:                 1572
Model:                            GLM   Df Residuals:                     1568
Model Family:                Gaussian   Df Model:                            3
Link Function:               identity   Scale:                      2.0045e+06
Method:                          IRLS   Log-Likelihood:                -13634.
Date:                Mon, 17 Feb 2020   Deviance:                   3.1431e+09
Time:                        13:48:28   Pearson chi2:                 3.14e+09
No. Iterations:                     3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2601.2871    231.095     11.256      0.000    2148.350    3054.225
Age           10.1121      0.123     82.027      0.000       9.870      10.354
BWeight        0.8866      0.071     12.478      0.000       0.747       1.026
Female      -520.3096     71.478     -7.279      0.000    -660.404    -380.215
==============================================================================
```

Here is a GEE model with the same mean structure as in the cell above, but using GEE gives us meaningful standard errors:

[5]:
```
model1 = sm.GEE.from_formula("Weight ~ Age + BWeight + Female", groups="Id",␣
 ↪data=df)
rslt1 = model1.fit()
print(rslt1.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                      Weight   No. Observations:
1572
Model:                                 GEE   No. clusters:
568
Method:                        Generalized   Min. cluster size:
1
                      Estimating Equations   Max. cluster size:
5
Family:                           Gaussian   Mean cluster size:
2.8
Dependence structure:         Independence   Num. iterations:
3
Date:                     Mon, 17 Feb 2020   Scale:
2004506.825
Covariance type:                    robust   Time:
13:48:29
==============================================================================
```

```
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2601.2871    268.766      9.679      0.000    2074.515    3128.059
Age           10.1121      0.111     90.912      0.000       9.894      10.330
BWeight        0.8866      0.086     10.367      0.000       0.719       1.054
Female      -520.3096     79.384     -6.554      0.000    -675.900    -364.719
==============================================================================
Skew:                          0.3683   Kurtosis:                       0.1819
Centered skew:                -0.2702   Centered kurtosis:              0.0109
==============================================================================
```

Now we fit the same model as a log/log regression. Specifically, the relationship between weight in childhood at a given age and birth weight is modeled as a log/log relationship. This means that when comparing two children of the same sex whose birth weights differed by a given percentage, say x, then their childhood weights at a given age differ on average by a corresponding percentage $b \cdot x$, where b is the coefficient of LogBWeight in the model. Typically we anticipate that $0 \leq b \leq 1$ in this type of regression. If $b \approx 1$ then, say, two kids whose weights at birth differ by 20% will continue to have weights differing by 20% as they age. If $b < 1$, then the 20% difference at birth will attenuate as the kids age.

[6]:
```
model2 = sm.GEE.from_formula("LogWeight ~ Age + LogBWeight + Female",␣
 ↪groups="Id", data=df)
rslt2 = model2.fit()
print(rslt2.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                   LogWeight   No. Observations:
1572
Model:                                 GEE   No. clusters:
568
Method:                        Generalized   Min. cluster size:
1
                      Estimating Equations   Max. cluster size:
5
Family:                           Gaussian   Mean cluster size:
2.8
Dependence structure:         Independence   Num. iterations:
2
Date:                     Mon, 17 Feb 2020   Scale:
0.094
Covariance type:                    robust   Time:
13:48:29
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.0936      0.501     18.151      0.000       8.112      10.076
```

```
Age              0.0018   1.86e-05     95.480      0.000       0.002       0.002
LogBWeight       0.2839      0.043      6.599      0.000       0.200       0.368
Female          -0.0910      0.014     -6.439      0.000      -0.119      -0.063
==============================================================================
Skew:                        -0.0690   Kurtosis:                      -0.9649
Centered skew:               -0.2631   Centered kurtosis:             -0.8613
==============================================================================
```

It isn't very likely that weight varies either linearly or exponentially with age. We can use splines to capture a much broader range of relationships.

[7]:
```
model3 = sm.GEE.from_formula("LogWeight ~ bs(Age, 4) + LogBWeight + Female",␣
 ↪groups="Id", data=df)
rslt3 = model3.fit()
print(rslt3.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                   LogWeight   No. Observations:
1572
Model:                                 GEE   No. clusters:
568
Method:                        Generalized   Min. cluster size:
1
                      Estimating Equations   Max. cluster size:
5
Family:                           Gaussian   Mean cluster size:
2.8
Dependence structure:         Independence   Num. iterations:
2
Date:                     Mon, 17 Feb 2020   Scale:
0.024
Covariance type:                    robust   Time:
13:48:29
================================================================================
=
                    coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-
Intercept         7.9489      0.384     20.675      0.000       7.195
8.702
bs(Age, 4)[0]     0.9993      0.038     26.229      0.000       0.925
1.074
bs(Age, 4)[1]     1.6037      0.037     42.999      0.000       1.531
1.677
bs(Age, 4)[2]     1.8115      0.064     28.180      0.000       1.685
```

```
1.937
bs(Age, 4)[3]      1.8769       0.028      67.049       0.000       1.822
1.932
LogBWeight         0.3375       0.033      10.193       0.000       0.273
0.402
Female            -0.0859       0.011      -7.719       0.000      -0.108
-0.064
==============================================================================
Skew:                          0.1011   Kurtosis:                       0.9305
Centered skew:                 0.2006   Centered kurtosis:              4.5751
==============================================================================
```

It is quite possible that the relationships between birth weight and childhood weight differ between girls and boys. An interaction captures this possibility.

[8]:
```
model4 = sm.GEE.from_formula("LogWeight ~ bs(Age, 4) + LogBWeight*Female",␣
 ↪groups="Id", data=df)
rslt4 = model4.fit()
print(rslt4.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                 LogWeight   No. Observations:
1572
Model:                               GEE   No. clusters:
568
Method:                      Generalized   Min. cluster size:
1
                   Estimating Equations   Max. cluster size:
5
Family:                         Gaussian   Mean cluster size:
2.8
Dependence structure:       Independence   Num. iterations:
2
Date:                   Mon, 17 Feb 2020   Scale:
0.024
Covariance type:                  robust   Time:
13:48:29
================================================================================
=====
                        coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----
Intercept             8.0705      0.621     12.991      0.000       6.853
9.288
bs(Age, 4)[0]         0.9988      0.038     26.309      0.000       0.924
```

```
1.073
bs(Age, 4)[1]         1.6042       0.037      43.008       0.000        1.531
1.677
bs(Age, 4)[2]         1.8095       0.064      28.255       0.000        1.684
1.935
bs(Age, 4)[3]         1.8779       0.028      67.290       0.000        1.823
1.933
LogBWeight            0.3270       0.054       6.103       0.000        0.222
0.432
Female               -0.3500       0.714      -0.490       0.624       -1.749
1.049
LogBWeight:Female     0.0228       0.062       0.371       0.711       -0.098
0.143
==============================================================================
Skew:                          0.1026   Kurtosis:                       0.9416
Centered skew:                 0.2000   Centered kurtosis:              4.5762
==============================================================================
```

Although GEE does not require us to specify an accurate covariance structure, we will have more power if we do so. We will also learn something about the strength of the within-subject dependence that we would not learn when using the independence model.

[9]:
```python
model5 = sm.GEE.from_formula("LogWeight ~ bs(Age, 4) + LogBWeight + Female",␣
 ↪groups="Id",
                             cov_struct=sm.cov_struct.Exchangeable(), data=df)
rslt5 = model5.fit()
print(rslt5.summary())
print(rslt5.cov_struct.summary())
```

```
                               GEE Regression Results
===============================================================================
===
Dep. Variable:                  LogWeight   No. Observations:
1572
Model:                                GEE   No. clusters:
568
Method:                       Generalized   Min. cluster size:
1
                     Estimating Equations   Max. cluster size:
5
Family:                          Gaussian   Mean cluster size:
2.8
Dependence structure:        Exchangeable   Num. iterations:
6
Date:                    Mon, 17 Feb 2020   Scale:
0.024
Covariance type:                   robust   Time:
13:48:30
```

```
==============================================================================
=
                    coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-
Intercept         7.7638      0.361     21.482      0.000       7.055
8.472
bs(Age, 4)[0]     0.9613      0.032     29.705      0.000       0.898
1.025
bs(Age, 4)[1]     1.6186      0.031     52.824      0.000       1.559
1.679
bs(Age, 4)[2]     1.7834      0.054     32.797      0.000       1.677
1.890
bs(Age, 4)[3]     1.8689      0.024     77.722      0.000       1.822
1.916
LogBWeight        0.3543      0.031     11.410      0.000       0.293
0.415
Female           -0.0796      0.011     -7.363      0.000      -0.101
-0.058
==============================================================================
Skew:                          0.1188   Kurtosis:                       0.9295
Centered skew:                 0.1842   Centered kurtosis:              4.6793
==============================================================================
The correlation between two observations in the same cluster is 0.466
```

In general, it is better to use the default "robust" approach for covariance estimation. This allows the covariance model to be mis-specified, while still yielding valid parameter estimates and standard errors. If you are very confident that your working covariance model is correct, you can specify the "naive" approach to covariance estimation, as below. In this case, the standard errors will be meaningful only if the working correlation model is correct.

[10]:
```python
model6 = sm.GEE.from_formula("LogWeight ~ bs(Age, 4) + LogBWeight + Female",␣
 ↪groups="Id",
                             cov_struct=sm.cov_struct.Exchangeable(), data=df)
rslt6 = model6.fit(cov_type="naive")
print(rslt6.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                     LogWeight   No. Observations:
1572
Model:                                   GEE   No. clusters:
568
Method:                          Generalized   Min. cluster size:
1
                       Estimating Equations   Max. cluster size:
```

```
5
Family:                          Gaussian   Mean cluster size:
2.8
Dependence structure:        Exchangeable   Num. iterations:
6
Date:                    Mon, 17 Feb 2020   Scale:
0.024
Covariance type:                    naive   Time:
13:48:31
===============================================================================
=
                   coef    std err          z      P>|z|      [0.025
0.975]
-------------------------------------------------------------------------------
-
Intercept        7.7638      0.249     31.119      0.000       7.275
8.253
bs(Age, 4)[0]    0.9613      0.035     27.482      0.000       0.893
1.030
bs(Age, 4)[1]    1.6186      0.033     49.210      0.000       1.554
1.683
bs(Age, 4)[2]    1.7834      0.058     30.567      0.000       1.669
1.898
bs(Age, 4)[3]    1.8689      0.027     69.786      0.000       1.816
1.921
LogBWeight       0.3543      0.021     16.490      0.000       0.312
0.396
Female          -0.0796      0.011     -7.356      0.000      -0.101
-0.058
==============================================================================
Skew:                          0.1188   Kurtosis:                       0.9295
Centered skew:                 0.1842   Centered kurtosis:              4.6793
==============================================================================
```