exams

February 17, 2020

[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)
```

1 GEE analysis of test score data

Generalized estimating equations (GEE) can be used to analyze multilevel data that arises in educational research, for example when students taking a test are grouped into classrooms.

The examination data analyzed here are obtained from this page: http://www.bristol.ac.uk/cmm/learning/support/datasets/

The data are in fixed-width format, we can load it as follows:

[2]:
```
colspecs = [(0, 5), (6, 10), (11, 12), (13, 16), (17, 20)]
df = pd.read_fwf("../data/exam_scores/SCI.DAT", colspecs=colspecs, header=None)
df.columns = ["schoolid", "subjectid", "gender", "score1", "score2"]
df["female"] = 1*(df.gender == 1)
df = df.dropna()
```

Here is a basic model looking at the scores on exam 1 by gender, using the default independence working correlation structure.

[3]:
```
# A school-clustered model for exam score 1 with no correlation.
model1 = sm.GEE.from_formula("score1 ~ female", groups="schoolid", data=df)
rslt1 = model1.fit()
print(rslt1.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                         score1   No. Observations:
1905
Model:                                    GEE   No. clusters:
```

```
73
Method:                         Generalized   Min. cluster size:
2
                      Estimating Equations   Max. cluster size:
104
Family:                            Gaussian   Mean cluster size:
26.1
Dependence structure:          Independence   Num. iterations:
2
Date:                      Mon, 17 Feb 2020   Scale:
451.997
Covariance type:                     robust   Time:
19:21:02
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     78.2136      1.864     41.960      0.000      74.560      81.867
female        -5.5292      1.183     -4.673      0.000      -7.848      -3.210
==============================================================================
Skew:                        -0.0935   Kurtosis:                        -0.0730
Centered skew:                0.1914   Centered kurtosis:                0.1835
==============================================================================
```

Here is the same mean structure model, now specifying that the students are exchangeably correlated within classrooms.

[4]:
```
# A school-clustered model for exam score 1 with exchangeable correlations.
model2 = sm.GEE.from_formula("score1 ~ female", groups="schoolid",
                             cov_struct=sm.cov_struct.Exchangeable(), data=df)
rslt2 = model2.fit()
print(rslt2.summary())
print(model2.cov_struct.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                       score1   No. Observations:
1905
Model:                                  GEE   No. clusters:
73
Method:                         Generalized   Min. cluster size:
2
                      Estimating Equations   Max. cluster size:
104
Family:                            Gaussian   Mean cluster size:
26.1
Dependence structure:          Exchangeable   Num. iterations:
7
Date:                      Mon, 17 Feb 2020   Scale:
```

```
456.642
Covariance type:                    robust   Time:
19:21:03
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     79.2582      1.561     50.764      0.000      76.198      82.318
female        -3.9121      0.922     -4.243      0.000      -5.719      -2.105
==============================================================================
Skew:                         -0.0987   Kurtosis:                      -0.0675
Centered skew:                 0.1848   Centered kurtosis:              0.1751
==============================================================================
The correlation between two observations in the same cluster is 0.422
```

Next we will pivot the exam scores so that each subject has two observations on a single "test" variable (one observation for the first test and one for the second test). This is a form of repeated measures, but since the tests are different, we also include a covariate indicating which test is being recorded. We now have two levels of repeated structure: two test scores per student, and multiple students per classroom. We can use a nested correlation structure to estimate the variance contributions from the two levels.

[5]:
```
# Prepare to do a joint analysis of the two scores.
dx = pd.melt(df, id_vars=["subjectid", "schoolid", "female"],
             value_vars=["score1", "score2"], var_name="test",
             value_name="score")
```

[6]:
```
# A nested model for subjects within schools, having two scores per subject.
model3 = sm.GEE.from_formula("score ~ female + test", groups="schoolid",␣
 ↪dep_data="0 + subjectid",
                             cov_struct=sm.cov_struct.Nested(), data=dx)
rslt3 = model3.fit()
print(rslt3.summary())
print(model3.cov_struct.summary())
```

```
                               GEE Regression Results
================================================================================
===
Dep. Variable:                      score   No. Observations:
3810
Model:                                GEE   No. clusters:
73
Method:                       Generalized   Min. cluster size:
4
                      Estimating Equations   Max. cluster size:
208
Family:                          Gaussian   Mean cluster size:
52.2
Dependence structure:              Nested   Num. iterations:
7
```

```
Date:                     Mon, 17 Feb 2020   Scale:
388.593
Covariance type:                    robust   Time:
19:21:03
================================================================================
==
                    coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
--
Intercept         75.0859      1.629     46.081      0.000      71.892
78.280
test[T.score2]     4.0950      1.564      2.618      0.009       1.030
7.160
female             1.7597      0.899      1.958      0.050      -0.002
3.521
==============================================================================
Skew:                         -0.3370   Kurtosis:                        0.2129
Centered skew:                -0.1909   Centered kurtosis:               0.4841
==============================================================================
             Variance
schoolid   119.911185
subjectid   57.843633
Residual   210.837685
```