Multivariate regression

Contents

1. Introduction

To illustrate multivariate regression, we use data collected by Dr. William Rohwer (see Timm, 1975). Data were collected on children in kindergarten and the question of interest was to predict test scores from paired-associated (PA) learning proficiency developed by Dr. Rohwer.

The data file **rohwer.lsf** shown below contains information on 37 children with low socio-economic status and 32 children with high socio-economic status. The data and syntax files can be found in the **MVABOOK examples\Chapter2** folder.

The variables are:

- SES: indicator of socio-economic status. Code 0 for children with low SES, 1 for children with high SES.
- SAT: score on a Student Achievement Test
- PPVT: score on the Peabody Picture Vocabulary Test
- Raven: score on the Raven Progressive Matrices Test
- named: performance on a 'named' PA test
- Still: performance on a 'Still' PA task
- NamedSti: performance on a 'named still' PA task
- NamedAct: performance on a 'named action' PA task
- SentStil: performance on a 'sentence still' PA task.

rohwer.LSF

	SES	SAT	PPVT	Raven	named	Still	NamedSti	Namedact	SentStil
1	0.00	49.00	48.00	8.00	1.00	2.00	6.00	12.00	16.00
2	0.00	47.00	76.00	13.00	5.00	14.00	14.00	30.00	27.00
3	0.00	11.00	40.00	13.00	0.00	10.00	21.00	16.00	16.00
4	0.00	9.00	52.00	9.00	0.00	2.00	5.00	17.00	8.00
5	0.00	69.00	63.00	15.00	2.00	7.00	11.00	26.00	17.00
6	0.00	35.00	82.00	14.00	2.00	15.00	21.00	34.00	25.00
7	0.00	6.00	71.00	21.00	0.00	1.00	20.00	23.00	18.00
8	0.00	8.00	68.00	8.00	0.00	0.00	10.00	19.00	14.00
9	0.00	49.00	74.00	11.00	0.00	0.00	7.00	16.00	13.00
10	0.00	8.00	70.00	15.00	3.00	2.00	21.00	26.00	25.00
11	0.00	47.00	70.00	15.00	8.00	16.00	15.00	35.00	24.00
12	0.00	6.00	61.00	11.00	5.00	4.00	7.00	15.00	14.00
13	0.00	14.00	54.00	12.00	1.00	12.00	13.00	27.00	21.00
14	0.00	30.00	55.00	13.00	2.00	1.00	12.00	20.00	17.00
15	0.00	4.00	54.00	10.00	3.00	12.00	20.00	26.00	22.00
16	0.00	24.00	40.00	14.00	0.00	2.00	5.00	14.00	8.00
17	0.00	19.00	66.00	13.00	7.00	12.00	21.00	35.00	27.00
18	0.00	45.00	54.00	10.00	0.00	6.00	6.00	14.00	16.00
19	0.00	22.00	64.00	14.00	12.00	8.00	19.00	27.00	26.00
20	0.00	16.00	47.00	16.00	3.00	9.00	15.00	18.00	10.00

We are interested in using the five performance assessments to predict the scores on the SAT, PPVT and Raven scores. Another question to answer is whether the effective predictors in such an analysis are the same for the two socio-economic groups.

2. Regression ignoring SES

As a first step, we conduct an analysis in which the performance assessments are used to predict the three scores. The PRELIS syntax to perform the analysis are given in **rohwer1.prl**.

rohwer1.prl

```
!Regression on Pooled Data
System File rohwer.lsf
Regress SAT PPVT Raven on SES named - SentStil
```

The following univariate summary statistics were obtained for the variables in the data set. Note that the range of the PPVT test is noticeably different from those of the SAT and Raven tests.

```
Univariate Summary Statistics for Continuous Variables

 Variable      Mean   St. Dev.   Skewness   Kurtosis   Minimum Freq.   Maximum Freq.
 --------      ----   --------   --------   --------   ------- -----   ------- -----
      SES     0.464      0.502      0.149     -2.038     0.000    37     1.000    32
      SAT    38.870     29.919      0.588     -0.867     1.000     1    99.000     1
     PPVT    72.130     16.651      0.147     -0.805    40.000     2   105.000     1
    Raven    14.058      3.101      0.167     -0.322     8.000     3    21.000     2
    named     3.493      3.768      1.997      5.760     0.000    17    20.000     1
    Still     7.072      4.384      0.305     -0.405     0.000     4    18.000     1
 NamedSti    13.957      5.478     -0.040     -0.641     3.000     1    28.000     1
 Namedact    23.275      6.223     -0.201     -0.798    11.000     1    35.000     2
 SentStil    19.812      6.032     -0.139     -0.671     8.000     4    32.000     1
```

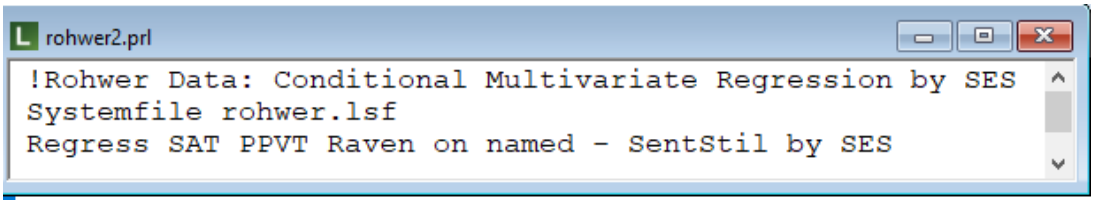
The estimated regression equations are

```
Estimated Equations

      SAT =  - 1.767  + 8.798*SES + 1.605*named + 0.0257*Still - 2.627*NamedSti
 Standerr     (13.418) (6.808)     (1.045)       (0.880)        (0.879)
 t-values     -0.132    1.292       1.536         0.0292        -2.988
 P-values      0.896    0.201       0.130         0.977          0.004

             + 2.106*Namedact + 0.930*SentStil + Error, R² = 0.295
              (0.888)          (0.835)
               2.371            1.114
               0.021            0.270

 Error Variance = 692.025


     PPVT = 31.349 + 16.877*SES + 0.00233*named - 0.351*Still - 0.299*NamedSti
 Standerr  (5.422)  (2.751)      (0.422)         (0.356)       (0.355)
 t-values   5.781    6.134        0.00551        -0.987        -0.841
 P-values   0.000    0.000        0.996           0.327         0.404

             + 1.294*Namedact + 0.479*SentStil + Error, R² = 0.628
              (0.359)          (0.337)
               3.604            1.419
               0.001            0.161

 Error Variance = 113.017


    Raven = 10.747 + 1.586*SES + 0.0149*named + 0.181*Still + 0.112*NamedSti
 Standerr  (1.471)  (0.747)     (0.115)        (0.0965)      (0.0964)
 t-values   7.304    2.124       0.130          1.877         1.157
 P-values   0.000    0.038       0.897          0.065         0.252

             - 0.00970*Namedact - 0.00446*SentStil + Error, R² = 0.211
              (0.0974)          (0.0916)
              -0.0996           -0.0488
               0.921             0.961

 Error Variance = 8.322
```

While the estimated coefficient for NamedSti is significant in the case of SAT, it is not significant for predicting PPVT or Raven scores. The coefficient associated with SES, on the other hand, is significant for PPVT and Raven, but not for SAT. The named estimate is not significant for any of the three tests.

3. Including SES in the model

We next allow the intercepts and regression coefficients to be different for the two socio-economic groups. This conditional multivariate regression analysis is set up by specifying the syntax



For the children with low SES, the estimated regression equations are:

```
For SES = 0, Sample Size = 37:

      SAT = 4.151  - 0.609*named - 0.0502*Still - 1.732*NamedSti
 Standerr  (13.798) (1.671)       (0.942)        (0.910)
 t-values   0.301   -0.364        -0.0533        -1.903
 P-values   0.765    0.718         0.958          0.066

             + 0.495*Namedact + 2.248*SentStil + Error, R² = 0.208
              (1.037)          (1.101)
               0.477            2.042
               0.637            0.050

 Error Variance = 449.339


 For SES = 0, Sample Size = 37:

     PPVT = 33.006 - 0.0806*named - 0.721*Still - 0.298*NamedSti
 Standerr  (6.147)  (0.744)        (0.419)       (0.406)
 t-values   5.369   -0.108         -1.719        -0.735
 P-values   0.000    0.915          0.095         0.467

             + 1.470*Namedact + 0.324*SentStil + Error, R² = 0.511
              (0.462)          (0.490)
               3.183            0.660
               0.003            0.514

 Error Variance = 89.186


For SES = 0, Sample Size = 37:

    Raven = 11.173 + 0.211*named + 0.0646*Still + 0.214*NamedSti
 Standerr  (1.915)  (0.232)       (0.131)        (0.126)
 t-values   5.835    0.910         0.494          1.690
 P-values   0.000    0.370         0.625          0.101

             - 0.0373*Namedact - 0.0521*SentStil + Error, R² = 0.222
              (0.144)           (0.153)
              -0.259            -0.341
               0.797             0.735

 Error Variance = 8.654
```

For the PPVT and Raven test, the coefficients for Namedact is highly significant in predicting PPVT, but not when SAT or Raven is considered. Similarly, SentStil and NamedSti are significant in predicting SAT but does not contribute significantly to predicting PPVT or Raven. Turning to the results for children with high SES, we see a similar pattern.

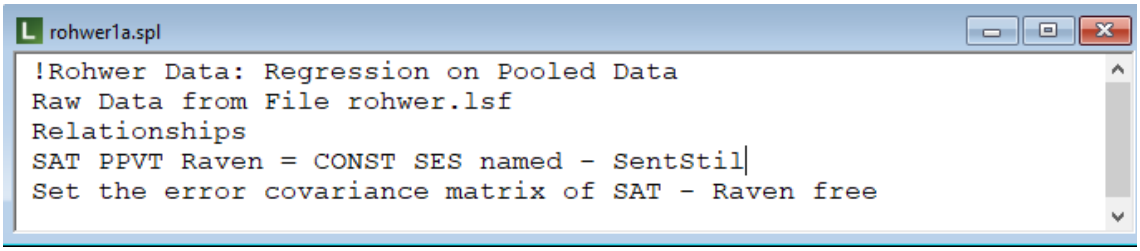
```
For SES = 1, Sample Size = 32:

      SAT =  - 28.467 + 3.257*named + 2.997*Still - 5.859*NamedSti
 Standerr     (25.719) (1.296)       (1.500)       (1.545)
 t-values     -1.107    2.514         1.998        -3.792
 P-values      0.278    0.018         0.056         0.001

             + 5.666*Namedact - 0.623*SentStil + Error, R² = 0.557
              (1.338)          (1.141)
               4.234           -0.546
               0.000            0.590

 Error Variance = 659.005


 For SES = 1, Sample Size = 32:

     PPVT = 39.697 + 0.0673*named + 0.370*Still - 0.374*NamedSti
 Standerr  (12.269) (0.618)        (0.716)       (0.737)
 t-values   3.236    0.109          0.517        -0.508
 P-values   0.003    0.914          0.609         0.616

             + 1.523*Namedact + 0.410*SentStil + Error, R² = 0.354
              (0.638)          (0.544)
               2.385            0.754
               0.024            0.457

 Error Variance = 149.961


 For SES = 1, Sample Size = 32:

    Raven = 13.244 + 0.0593*named + 0.492*Still - 0.164*NamedSti
 Standerr  (2.614)  (0.132)        (0.152)       (0.157)
 t-values   5.066    0.451          3.230        -1.044
 P-values   0.000    0.656          0.003         0.306

             + 0.119*Namedact - 0.121*SentStil + Error, R² = 0.308
              (0.136)          (0.116)
               0.875           -1.045
               0.390            0.305

 Error Variance = 6.809
```

We can summarize these results by sampling noting which coefficients were statistically significant at a 5% level, as shown in the table below. Generally speaking, the performance assessments seem to do better at predicting scores for children with low SES. To test whether the differences noted in the table above are statistically significant, one may want to test the hypotheses of equal regression coefficients between the two groups.

	High SES					Low SES				
	named	Still	NamedSti	Namedact	SentStil	named	Still	NamedSti	Namedact	SentStil
SAT	0	0	0	0	1	1	0	1	1	0
PPVT	0	0	0	1	0	0	0	0	1	0
Raven	0	0	0	0	0	0	1	0	0	0

4. Allowing different intercepts for SES groups

In the previous analysis, it was assumed that residuals in the three regressions were uncorrelated. We now use SIMPLIS syntax to fit a model where we assume that the regression coefficients for the predictors are the same for the SES groups but allow the intercepts to be different between the two groups.



The following results were obtained.

```
LISREL Estimates (Maximum Likelihood)

         Structural Equations


SAT =  - 1.767   + 8.798*SES + 1.605*named + 0.0257*Still - 2.627*NamedSti + 2.106*Namedact + 0.930*SentStil,
 Standerr     (13.316)  (6.754)     (1.037)       (0.873)        (0.872)          (0.881)          (0.828)
 Z-values     -0.133     1.303       1.548         0.0295        -3.012            2.390            1.123
 P-values      0.894     0.193       0.122         0.976          0.003            0.017            0.262

             Errorvar.= 630.964 , R² = 0.295
     Standerr          (112.421)
     Z-values           5.612
     P-values           0.000

PPVT = 31.349 + 16.877*SES + 0.00233*named - 0.351*Still - 0.299*NamedSti + 1.294*Namedact + 0.479*SentStil,
 Standerr  (5.381)  (2.729)      (0.419)        (0.353)       (0.353)          (0.356)          (0.335)
 Z-values   5.825    6.183        0.00555       -0.995        -0.848            3.633            1.431
 P-values   0.000    0.000        0.996          0.320         0.397            0.000            0.153

             Errorvar.= 103.045, R² = 0.628
     Standerr          (18.360)
     Z-values           5.612
     P-values           0.000

Raven = 10.747 + 1.586*SES + 0.0149*named + 0.181*Still + 0.112*NamedSti - 0.00970*Namedact - 0.00446*SentStil,
 Standerr  (1.460)  (0.741)     (0.114)        (0.0958)      (0.0957)         (0.0966)          (0.0908)
 Z-values   7.359    2.141       0.131          1.892         1.166           -0.100            -0.0491
 P-values   0.000    0.032       0.896          0.059         0.244            0.920             0.961

             Errorvar.= 7.588 , R² = 0.211
     Standerr          (1.352)
     Z-values           5.612
     P-values           0.000
```

We note that the children with high SES had higher scores on all three tests as indicated by the positive estimates for the SES contribution. When the R^2 obtained for the three tests are considered, we see that PPVT has a higher R^2 than either of the other two tests. This seems to suggest that this score can be predicted better by the predictors used here than is the case for either SAT or Raven.

```
Error Covariance for PPVT and SAT = 58.660
                                   (33.229)
                                     1.765

Error Covariance for Raven and SAT = 23.533
                                     (9.282)
                                      2.535

Error Covariance for Raven and PPVT = 8.381
                                     (3.707)
                                      2.261
```

The estimated residual covariances indicate that the residual covariance is significant for (Raven, SAT) and for (Raven, PPVT) but not for (PPVT, SAT).