An Example of Analysing Pooled Cross-Sectional Surveys

In this section we provide a didactic example of undertaking a statistical analysis by using pooled cross-sectional surveys. The example serves to illustrate how accounts of social change can be developed by analysing pooled cross-sectional data using standard statistical techniques. We also use the illustrative example as a more general statement about approaches to using statistical models to analyse social processes, and we provide comments on methods and the interpretation of results that are relevant to this example but also apply to many other scenarios when analysing large scale quantitative datasets.

In this section we provide a concise example of using pooled cross-sectional surveys. The data are from the Youth Cohort Study of England and Wales (YCS) which we introduced in Chapter 1. In this example we pool data from the first sweeps of the 1990, 1993, 1995, 1997 and 1999 cohorts (using data from UK Data Archive Study Number SN 5765). A compact Stata codebook is provided in Figure 12. Historically school pupils in England and Wales have undertaken a diet of examinations called the General Certificate of Secondary Education (GCSE) at the end of compulsory education (see Playford and Gayle, 2016). The first outcome variable of interest is the pupil's GCSE points score in Year 11 (which was the end of compulsory school). We also have information on the pupil's gender and their parent's social class which is measured by a three category version of the official UK socioeconomic classification measures NS-SEC (see Rose et al., 2005).

Figure 1 Stata Compact Codebook Youth Cohort Study of England and Wales (1990, 1993, 1995,1997 and 1999 cohorts)

```
Variable     Obs Unique      Mean   Min   Max  Label
-------------------------------------------------------------------------------------------
t0score    62910     97  37.61346     0   112  GCSE points score year 11
t05examac  62910      2  .5084088     0     1  5+ GCSEs at grades A*-C year 11
t0cohort   64045      5  1994.678  1990  1999  year completed compulsory schooling
```

```
male       64045      2  .4622219     0     1  males
t0parsc3   56210      3  1.867283     1     3  parent's NS-SEC 3 class
-------------------------------------------------------------------------------------------
```

A linear regression model can be estimated on the pooled data from the five YCS cohorts using the following Stata syntax

regress t0score i.t0cohort male i.t0parsc3

Figure 2 Stata Output: Regression Model (OLS) General Certificate of Secondary Education (GCSE) Points Score School Year 11 Youth Cohort Study of England and Wales

```
      Source |       SS           df       MS      Number of obs   =    55,497
-------------+----------------------------------   F(7, 55489)     =   1711.10
       Model |  2888927.37         7  412703.911   Prob > F        =    0.0000
    Residual |  13383518.1    55,489  241.192275   R-squared       =    0.1775
-------------+----------------------------------   Adj R-squared   =    0.1774
       Total |  16272445.5    55,496  293.218349   Root MSE        =     15.53

------------------------------------------------------------------------------
             t0score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
           t0cohort |
                1993 |   5.279937   .2042092    25.86   0.000     4.879686    5.680188
                1995 |   9.354583   .2111997    44.29   0.000      8.94063    9.768536
                1997 |   8.084928   .2114485    38.24   0.000     7.670487    8.499368
                1999 |   12.66983   .2193682    57.76   0.000     12.23987    13.09979
                     |
                male |  -3.392621   .1323279   -25.64   0.000    -3.651984   -3.133257
                     |
           t0parsc3 |
 intermediate occs. |  -7.417756   .1535252   -48.32   0.000    -7.718667   -7.116846
routine and manual occs. |   -13.8186   .1669247   -82.78   0.000    -14.14577   -13.49142
                     |
                _cons |   39.67939   .1851758   214.28   0.000     39.31645    40.04234
------------------------------------------------------------------------------
```

The results of the linear regression model are reported in Figure 13. Many readers will be completely familiar with Stata output for standard linear regression models, however we will provide an extended discussion of the results to aid readers who might be less conversant with such output. This will aid the understanding of some of the models that will be presented in Chapter 5.

The results at the top left of the output in Figure 13 which are labelled 'Source', are sometimes referred to as an ANOVA table. The total sum of squares (SS) is reported which is made up of the model sum of squares and the residual sum of squares. Another way of thinking about this information is that the total variance is partitioned into the variance which can be explained by the independent variables (i.e. the model) and the variance which is not explained by the independent variables (i.e. the residual).

The degrees of freedom associated with each source of variance is reported. The total variance has n-1 degrees of freedom. This model has 7 degrees of freedom, this is the number of coefficients associated with explanatory variables and the constant minus 1 (k-1). The total variance has 55,496 (n = 55,497– 1) degrees of freedom and therefore there are 55,489 (55,496 – 7) residual degrees of freedom. The mean squares (MS) are the sum of squares (SS) divided by their degrees of freedom (df).

The output at the top right of Figure 13 relates to how well the model summarizes the data. There are 55,497 observations included in the model. The *F* statistic is the mean square (MS) for the model divided by the mean square (MS) of the residual. It is evaluated at 7 degrees of freedom for the model and 55,489 degrees of freedom for the residual. The *F* statistic has a significant *p* value and this can be considered as a test of the null hypothesis (H_0: $\beta_0 = \beta_1 = ... = \beta_k = 0$) that all of the model coefficients are equal to zero.

The R^2 is the proportion of the total variance in the *Y* variable that is explained by all of the explanatory variables. R^2 is therefore the Model (SS) sum of squares / Total (SS) sum of squares. The Adjusted R^2 is computed using the formula $1-((1 - R^2)(n-1) / (n- k – 1))$ in this example $1-((1 - 0.1775)(55,497-1) / (55,497-7-1))$. The Adjusted R^2 is generally considered suitable as it penalizes the addition of extraneous explanatory variables being included in the model. The Root MSE is the square root of the mean square (MS) Residual.

The next part of the output in Figure 13 provides regression model results related to the explanatory variables. The first column of the table lists the variables in the model. The outcome variable *Y* is *t0score*, the other variables are explanatory variables. The next column reports the coefficient for the individual explanatory variables or predictors. The coefficients are alternatively referred to as estimates, parameter estimates, betas, or β in different research areas and statistical traditions. The standard errors associated with the estimated coefficient are reported in the next column. The *t* statistic for each individual explanatory variable is reported in the fourth column. The *t* statistic can be thought of as the ratio of the coefficient to its standard error (β / s.e.). A two-tailed *p* value is

reported in the next column. This can be thought of as a test of the null hypothesis that the coefficient for the explanatory variable is zero (H_0: $\beta_k = 0$). The final two columns in the regression output table are the lower and upper bounds of a 95% confidence interval for the coefficient of the explanatory variable. The 95% confidence interval is calculated as $\beta_k \pm (1.96 * \text{s.e.})$.

A synoptic assessment of the results reported in Figure 13 indicate that cohort, gender and parental social class are all significantly associated with GCSE attainment. Generally, pupil's GCSE scores increase (on average) in the more recent cohorts (*ceteris paribus*). Male pupils on average had lower GCSE scores than female pupils. Pupils with parents in the intermediate social class and routine and manual social classes on average had lower GCSE scores than pupils with parents in the higher managerial, administrative and professional social class categories.

When models include categorical explanatory variables we strongly advocate the calculation of quasi-variances which are statistics associated with the parameter estimates of the different levels of categorical explanatory variables within a regression model. In standard model results it is only possible to compare a category of an explanatory variable with the reference category. Quasi-variance estimates facilitate the performance of tests of difference between any combination of a categorical variable's parameter estimates, which is not usually possible without access to the full variance-covariance matrix for the estimates (see Gayle and Lambert, 2007)[1]. It is possible to compute quasi-variances directly in Stata using the function contributed by Aspen Chen of the University of Connecticut which we mentioned in Chapter 2 (*qv*). Quasi-variance estimates for parental social class can be produced using the following Stata syntax

qv i.t0parsc3

[1] A full set of resources for learning about quasi–variance calculations is available at http://www.restore.ac.uk/Longitudinal/qv/ accessed 28.06.16.

Figure 3 Stata Output qv Command: Parent's Social Class (three category NS-SEC) Regression Model Estimated Figure 13

```
------------------------------------------------------------------------------
 Category          Coef.         SE     Quasi-SE      lb(QV)      ub(QV)
------------------------------------------------------------------------------
 1b.t0parsc3       0.0000      0.0000     0.1049     -0.2481       0.2481
 2.t0parsc3       -7.4178      0.1535     0.1121     -7.6827      -7.1528
 3.t0parsc3      -13.8186      0.1669     0.1298    -14.1256     -13.5116
------------------------------------------------------------------------------
*lb and ub calculated at the 95% level
```

The results of the *qv* command are reported in Figure 14. We can see that the 95% comparison intervals for level 2 and level 3 of the variable *t0parsc3* do not overlap. Therefore we can conclude that pupils with parents in the intermediate social class (*t0parsc3==2*) are significantly different to pupils with parents in the routine and manual social class (*t0parsc3==3*). Connelly et al. (2016c) provide an extended account of the calculation of quasi-variance estimates and the presentation of these estimates along with other modelling results.

Binary outcome measures are extremely common in social science and therefore in the second part of this example we model a binary outcome using a logistic regression model. The outcome of interest is whether or not the pupil gained 5+ GCSEs at grades A*- C (see Playford and Gayle, 2016 for a discussion of this measure). A logistic regression model can be estimated in Stata using the following syntax

logit t05examac i.t0cohort male i.t0parsc3

Figure 4 Stata Output: Logistic Regression Model 5+ General Certificates of Secondary Education (GCSE) at Grades A- C Year 11 Youth Cohort Study of England and Wales*

```
Iteration 0:   log likelihood = -38301.069
Iteration 1:   log likelihood = -35233.109
Iteration 2:   log likelihood = -35225.775
Iteration 3:   log likelihood = -35225.774

Logistic regression                             Number of obs     =     55,497
                                                LR chi2(7)        =    6150.59
                                                Prob > chi2       =     0.0000
Log likelihood = -35225.774                     Pseudo R2         =     0.0803

------------------------------------------------------------------------------
             t05examac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------------
              t0cohort |
                 1993  |   .3775855   .0277531    13.61   0.000     .3231904    .4319807
                 1995  |   .6212732   .0288158    21.56   0.000     .5647953    .6777512
                 1997  |   .5997924   .0288189    20.81   0.000     .5433085    .6562764
                 1999  |   .9532303   .0304696    31.28   0.000      .893511     1.01295
                       |
                  male |  -.3219314   .0181062   -17.78   0.000    -.3574189   -.2864439
                       |
              t0parsc3 |
      intermediate occs.  |  -.8245439   .0209009   -39.45   0.000    -.8655089    -.783579
routine and manual occs.  |  -1.519305   .0233095   -65.18   0.000    -1.564991   -1.473619
                       |
                 _cons |   .4949234    .025192    19.65   0.000      .445548    .5442989
------------------------------------------------------------------------------
```

The results of the logistic regression model are reported in Figure 15. Many readers will be completely familiar with Stata output for logistic regression models, but once again we will provide an extended discussion of the result to aid readers who might be less conversant with this output. This will also aid their understanding of some of the models that will be presented in Chapter 4 and Chapter 5.

The first part of the output in Figure 15 lists the log likelihoods at each iteration. Logistic regression models are estimated using maximum likelihood rather than ordinary least squares. The first iteration (i.e. iteration 0) is the log likelihood of the null model (i.e. the model with no explanatory variables). In the next iteration, the explanatory variables are included in the model. Then at each subsequent iteration, the log likelihood increases because the goal within maximum likelihood estimation is to converge on a model that maximizes the log likelihood. When the incremental difference between successive iterations becomes very small, the model is said to have 'converged'. No further iterations are undertaken and the model results are displayed.

The output reports the number of observations in the model (55,497). The likelihood-ratio chi-square (i.e. 'LR chi2') is the $\text{Deviance}_{\text{null\_model}} - \text{Deviance}_{\text{full\_model}}$ in this example it is (-2*-38301.069) - (-2*-35225.774) = 6150.59. The likelihood-ratio chi-square can also be recovered from the scalar variables after the model is estimated using the following Stata syntax

display (-2 e(ll_0))-(-2* e(ll))*

The significance of the likelihood-ratio chi-square can be evaluated at 7 degrees of freedom (i.e. k-1 parameters for the explanatory variables in the model). The *LR chi2* statistic has a significant *p* value and this can be considered as a test of the null hypothesis (H_0: $\beta_0 = \beta_1 = ... = \beta_k = 0$) that all of the model coefficients are equal to zero. This is analogous to the *F* test in a linear regression model. The pseudo R^2 value that is reported is McFadden's R^2. This measure is 1- ($\text{Deviance}_{\text{full\_model}}$ / $\text{Deviance}_{\text{null\_model}}$) in this example it is 1- ((-2*-35225.774) /(-2*-38301.069)). McFadden's R^2 can also be recovered from the scalar variables after the model is estimated using the following Stata syntax

display 1-((-2* e(ll))/(-2* e(ll_0)))

Logistic regression models to not have direct equivalent to the R^2 measure that can be calculated after a linear regression model using ordinary least squares. Smithson (2003) cuttingly remarks that there has been something of a cottage industry in model fit statistics for logistic regression. A number these alternative measures can be calculated in Stata using the contributed command *fitstat*. Long and Freese (2014) provide an excellent overview of the scope and limitations of these measures. Currently we are not persuaded that any single pseudo R^2 should be routinely preferred above all others. Therefore we suggest that in genuine analyses researchers should report a few alternative measures in published research and document others in the data appendices. We suggest that when researchers are comparing nested models, there is a compelling case for also using a measure that accounts for parsimony such as the Bayesian Information Criterion (BIC) which was proposed by Raftery (1986).

The next part of the output in Figure 15 is the logistic regression model results. The first column of the table lists the variables in the analysis. The outcome variable *Y* is *t05examac*, the other variables

are explanatory variables. The next column reports the coefficients for the individual explanatory variables or predictors. Because this is a logistic regression model the units of measurement are expressed on the log odds scale. The standard errors associated with the coefficient are reported in the next column. The *z* statistic for the individual explanatory variable is reported in the fourth column. The z statistic can be thought of as the ratio of the coefficient to its standard error (β / s.e.) and a two-tailed *p* value is reported in the next column. This can be thought of as a test of the null hypothesis that the coefficient for the explanatory variable is zero (H_0: $\beta_k = 0$). The final two columns in the regression output table are the lower and upper bounds of a 95% confidence interval for the coefficient of the explanatory variable.