**Computer workshop 1: Complex survey design**

The aims of this workshop are to:

* Provide students with experience of using the TALIS Balanced-Repeated Replication (BRR) weights
* Help students understand how these weights ‘work’ to produce the final estimate of the standard error
* Introduce students to the Stata survey (‘svy’) and user-built ‘repest’ commands
* Illustrate the TALIS weights can be used to produce nationally representative estimates.

The data we shall be using are the 2013 round of the Teaching and Learning International Survey (TALIS). TALIS is an international survey of lower secondary school teachers (i.e. Key Stage 3 teachers in England) that was conducted in more than 30 countries. It takes places every five years (the inaugural round was 2008, with the next wave due to take place in 2018). As part of this survey, teachers answer questions regarding their teaching style, working conditions, professional development and demographic characteristics. In this workshop we will focus on data for a single country (Sweden). The variables included in the Stata data file are as follows:

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| PT | Whether the teacher works part time (0 = no; 1 = yes) |
| TT2G02 | Teacher age |
| TCHWGT | Final teacher weight |
| TRWGT1 -100 | Final set of 100 replicate weights |
| TT2G16 | Number of hours spent teaching per week |
| TT2G46J | Teacher’s job satisfaction |
| TT2G03 | Teachers’ current employment status |
| TT2G01 | Gender |
| TJSPROS | Satisfaction with the teaching profession (scale variable) |
| IDSCHOOL | School identifier |

This worksheet provides instructions on how to analyse the TALIS data in Stata. (The methods you learn here are also directly applicable to PISA, with very similar techniques used when analysing PIAAC, TIMSS or PIRLS data). I have highlighted in **bold** the commands that you need to execute.

1. **Where can I download the data from?**

The TALIS 2013 international database (containing data for all countries) can be downloaded in SPSS format from <http://stats.oecd.org/Index.aspx?datasetcode=talis_2013%20>. This workshop uses data for Sweden that I have downloaded from this website and converted into Stata format. You can find this data at my personal website: [www.johnjerrim.com/papers/pisacourse](http://www.johnjerrim.com/papers/pisacourse)

1. **Opening the data in STATA**

Download the data from my website via the link given above. Save these data to a location you can easily find (e.g. your computer desktop).

To open STATA:

* Left hand click the circular start button in the bottom left corner of the screen
* Click ‘All programs’
* Scroll down to ‘Stata 13’ folder. Double click.
* Double click the second icon (called ‘StataSE 13’)

To open the STATA data-file go to the top left of the screen and click:

**File > Open**

Then find the STATA data file on your desktop and double click on it.

Now move towards the top of the screen and click on:

**Window > Do – File Editor > New Do-File**

This opens a “do-file” (a basic text editor). It is where we write all of our programme code to analyse data. This is very important in empirical analysis – we MUST be able to replicate our results and we can only do this with a programme code.

1. **The TALIS ‘teacher’ weights**

The ‘final’ teacher weight included in the TALIS dataset is named *TCHWGT*. Recall from the lecture that this weight encompasses the design weight (i.e. unequal probabilities of selection by design), non-response adjustment and (possibly) some minor trimming. This weight must be applied to obtain ‘correct’ population point estimates.

Our task is to calculate the average (mean) age of teachers in Sweden. To begin, lets see what answer we get without applying the survey weight. Please run the following command in Stata. (You do this by writing the text in bold below in the Stata do-file editor, highlighting it, and clicking “Tools > Do” from the options at the top).

**mean TT2G02**

This should produce the following result. The estimated average age is **45.91 years**.



Now let us re-run this command, but applying the final teacher weight. The command is exactly the same as above, but with **[pw = TCHWGT]** added to the end.

**mean TT2G02 [pw = TCHWGT]**



This has resulted in a slight increase in the estimate to **45.98** years. At this stage, it is important to remember that although this gives you the correct point estimate, the standard errors will not match those provided by the OECD. (This can only be achieved via the application of the BRR replicate weights as discussed below).

When using the TALIS weights it is important to remember that you may need to apply these in intermediary steps of your analysis. For example, TALIS contains a scale variable recording teachers’ self-efficacy (the variable name is **TSELEFFS** in the dataset). Say you wanted to know the average age of teachers in the bottom self-efficacy quartile, compared to the average age of teachers in the top self-efficacy quartile.

The first step of your analysis will be to create the self-efficacy quartiles, which can be done using the Stata ‘xtile’ command. (Type **help xtile** for further information). However, it is important to remember to specify the final weight when creating these quartile groups – otherwise Stata will think the data has been collected using a simple random sample. (Hence will not take into account the complex survey design, e.g. oversampling of certain groups, or any adjustment that has been made to account for non-response). Only by including these weights at each stage of the analysis will one produce correct population estimates.

To divide teachers into self-efficacy quartiles, one can use the following Stata command:

**xtile Efficacy\_Quart = TSELEFFS ///** /\* Create new variable based upon TSELEFFS \*/

**[pweight = TCHWGT] ///** /\* Apply the final survey weight \*/

**, nq(4)** /\* Split into 4 equal groups \*/

In the above, we are using the **xtile** command to create a new variable. The name of this new variable **Efficacy\_Quart** then follows directly afterwards. This will be based upon a manipulation of the existing self-efficacy variable (**TSELEFFS**) which we put after the equals sign. One then tells Stata that the new variable should be created based upon weighted estimates **[pweight = TCHWGT]**. The final line - **, nq(4)** – then indicates how many equal groups we want to split the self-efficacy variable into. This is the number in the brackets. We set this to 4 so that the new variable refers to *quartiles*.

Now run the following command:

**tab Efficacy\_Quart [aweight = TCHWGT]**

This should produce the following table, with the population divided into four equal groups:



One can now estimate the average age of teachers in the top and bottom self-efficacy quartile by running the following two commands:

**mean TT2G02 [pweight = TCHWGT] if Efficacy\_Quart == 1**

**mean TT2G02 [pweight = TCHWGT] if Efficacy\_Quart == 4**

This reveals that teachers in the lowest efficacy quartile are younger (on average) than teachers in the top quartile (44.7 years versus 47.8 years).

**Question 1: What happens if you re-perform the analysis above, but do not apply the weights when creating the self-efficacy quartiles? (Note: Continue to apply the weights when calculating the mean age for the top and bottom quartile groups)**

**Task 1: The TALIS dataset I have provided contains the variables teacher age (TT2G02) and whether the teacher works part-time (PT). Please divide the Swedish data into population age quartiles, and compare the proportion of teachers working part-time across these age groups. Which has the greatest proportion of part-time workers?**

1. **How to use the BRR replicate weights**

The TALIS dataset contains 100 BRR replicate weights. As discussed in the lecture, these need to be applied to account for the complex survey design implemented in TALIS. The below will demonstrate three ways of using these weights:

* Manual execution
* Using the Stata survey (‘svy’) command
* The ‘repest’ Stata command developed by the OECD

The statistic of interest we will focus on is simply the average age of teachers in Sweden.

**4a. Manual execution**

To begin, we will calculate the average age of teachers in Sweden (along with the standard error) by using the BRR weights manually. Doing the grunt work ourselves to begin will help students understand what the computer is doing when using the ‘shortcuts’ described in sections 4b and 4c below.

The design used in TALIS is BRR with ‘Fay’s adjustment’. This is given by the formula:

Where:

= The statistic of interest (e.g. mean, median, regression coefficient)

= The statistic of interest calculated when using the final teacher weight

= The statistic of interest calculated when applying replicate weight R

R = The number of replicate weights

= Fay’s adjustment = 0.5

We will now implement this formula ourselves in Stata by executing the following nine steps.

Step 1: Calculate the point estimate ()

Our first task is to calculate the point estimate of our statistic of interest. In this example, it is the average age of teachers in Sweden. To do this, we simply follow the instructions provided in section 3 (i.e. we calculate mean age *remembering to apply the final teacher weight*). In Stata write the following command in the do-file and run it:

**mean TT2G02 [iweight = TCHWGT]**



We then want to save the resulting point estimate as a new Stata variable. To do this, we write the following:

**gen Theta\_Star = \_b[TT2G02]**

In the command above:

* ‘gen’ stands for ‘generate’. It is the way we create a new variable in Stata.
* ‘Theta\_Star’ is the name we have given this new variable
* \_b[TT2G02] is what this new variable should contain. The use of \_b[ ] is our way of telling Stata we want it to capture the *point estimate* in the new variable. Inside the square brackets, we then simply put the variable name for the point estimate we want to record.

A new variable should have been created in your Stata dataset. This should appear as a named vairbale (Theta\_Star) in the panel on the right. You can look at the contents of this variable by running the following command:

**browse Theta\_Star**

You should see that this has essentially produced a vector with the same value (45.97552) for each observation.

Step 2: Calculate each replicate estimate (

Next, we are going to reproduce our statistic of interest 100 times using TALIS replicate weights.

First of all, let us see how one would re-produce the statistic of interest just using the first replicate weight. Essentially, one would just do everything exactly the same as step 1, with the only change being that we apply the first replicate weight (*TRWGT1*) in place of the final teacher weight (*TCHWGT*):

**mean TT2G02 [iweight = TRWGT1]**

**gen Theta\_1 = \_b[TT2G02]**

The above has now successfully created the first replicate estimate () of the mean age of teachers in Sweden.

There are, of course, 100 replicate weights included in the TALIS dataset. This means we have to re-perform the analysis above a further 99 times. This can get very tedious! So it is efficient to do this via a simple loop.

First, lets drop the Theta\_1 variable created above:

**drop Theta\_1**

Then lets write the loop to create 100 new variables (Theta\_1 … Theta\_100) – once using each of the different replicate weights:

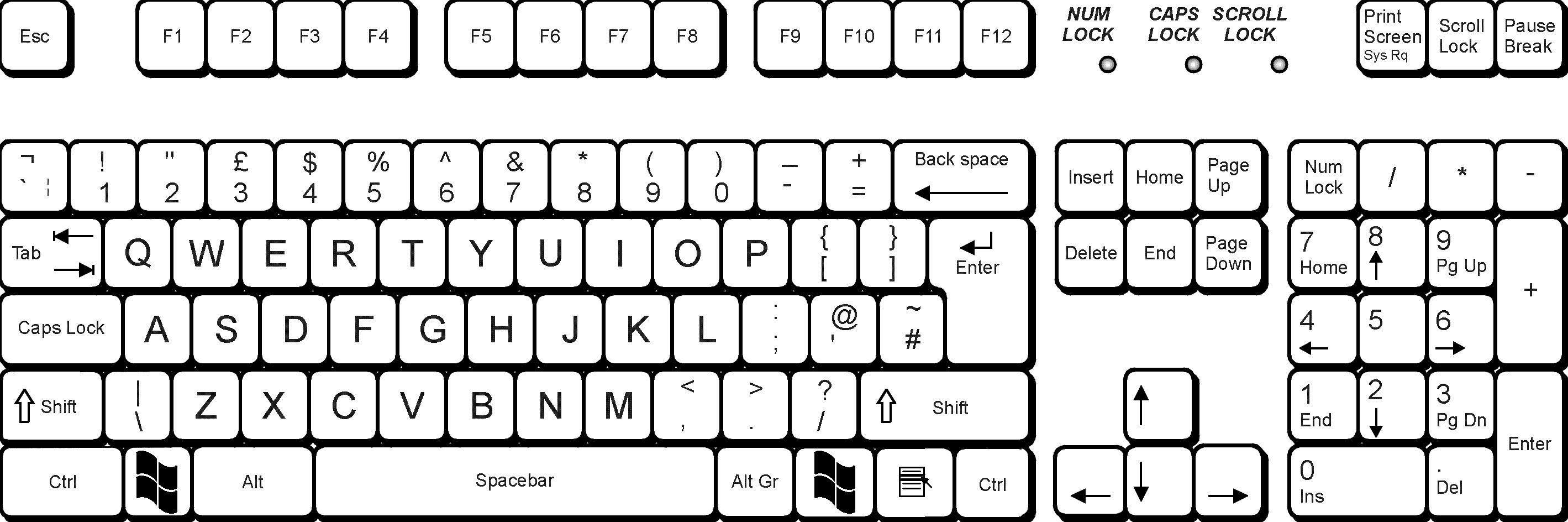
**forvalues rep = 1(1)100 {**

**mean TT2G02 [iweight = TRWGT`rep']**

**gen Theta\_`rep' = \_b[TT2G02]**

**}**

The above is an example of a ‘forvalues’ loop. In the first line, we are telling Stata we want to repeat our analysis for each value between 1 and 100. This is indicated by the 1(1)100. A set of curly brackets are then opened. Inside these brackets, I have re-written exactly the same command as appears above (for the first replicate weight). The only different is that I have replaced the number **1** in **TRWGT1** and **Theta\_1** with **`rep’**. This is the term we are ‘looping over’ (i.e. we want to repeat the same command several times, but changing this number every time). Here I have called this ‘rep’ (as appears before the equal sign on the first line of the command) but you could put almost anything else here (e.g. a common choice is `i’). The quote marks you incase this term in can be found on the key board here:



After running this command, you should now see that 100 new variables have been created in the Stata dataset. You should see these on the right hand side of the Stata command window (in the ‘variables’ box). These are your 100 different ‘replicate estimates’ of the average age of teachers in Sweden. You can inspect these (along with the point estimate calculated in step 1) by writing and running the following command in your do-file:

**browse Theta\_\***

Note here the \*. This is what we call a ‘wildcard’. When using the wildcard, Stata will list all the variables starting with the term before the \*. Hence Stata will browse all the variables that begin with Theta\_.

Step 3: Calculate the squared differences

If one returns to the formula presented at the start of section 4a, you should recognise that we now have all the components we need to produce the standard error of our estimate (and associated hypothesis tests and confidence intervals). All we need to do now is put these bits and pieces together.

To begin, we need to calculate the squared difference between each of the 100 replicate estimates () and the final point estimates (). With regards the first replicate weight only, the Stata command is:

**gen Theta\_Diff\_Squ\_1 = (Theta\_1 - Theta\_Star) ^2**

Here we are creating a new variable (**Theta\_Diff\_Squ\_1**) which is simply calculated as the difference between the first replicate estimate and the final point estimate **(Theta\_1 - Theta\_Star)** squared (**^2**).

However, as under step 2, we want to repeat this analysis 100 times – once using each of the replicate weights. Therefore let us drop this variable:

**drop** **Theta\_Diff\_Squ\_1**

and write another simple loop that repeat this 100 times:

**forvalues rep = 1(1)100 {**

**gen Theta\_Diff\_Squ\_`rep' = (Theta\_`rep' - Theta\_Star) ^2**

**}**

Another 100 new variables should have now appeared in your Stata dataset. These will contain the squared difference between each replicate estimate and your final point estimate. The value of all these variables should be positive. To check, type:

**browse Theta\_Diff\_Squ\_\***

Step 4: Calculate the sum of these squared differences

Next, the variables created in step 3 (the squared differences between the replicate estimates and the overall point estimate) need to all be added together. In other words, one needs to take the sum of the squared differences.

This can be executed in Stata by using the **egen** command. egen stands for ‘extended generate’ and is to be used when one wants to create a new variable where the contents is some manipulation of the other variables. In our example, we want to add together (horizontally) the contents of the 100 variables just browsed in the final part of step 3. The command we need is:

**egen SUM\_DIFF = rowtotal(Theta\_Diff\_Squ\*)**

In the above, SUM\_DIFF is the name of the new variable we want to create. On the right hand side of the = sign, we have indicated we want to calculate the **rowtotal**. This is telling Stata that we want to add the contents together along each row. Finally, in the brackets we include all those variables we want to include in our calculation when creating this new variable. (Note, again, the use of the wildcard \*. In our calculation we will include all variables in the Stata dataset starting with the term **Theta\_Diff\_Squ**).

Step 5: Calculate the adjustment we must multiply the sum-of-squares by

Returning again to the equation presented at the start of section 4a, note that the sum-of-squares calculated in step 4 is multiplied by the factor (). Here R stands for the number of replicate weights, which equals 100 in TALIS. The other term () is known as ‘Fay’s adjustment’ – reflecting a technical adjustment that is made in certain applications of the BRR methodology (including TALIS and PISA). We will not discuss Fay’s adjustment in detail in this course, the important thing for secondary analysts is to know that in TALIS (and PISA) this always takes the value 0.5. Hence our adjustment factor can be calculated via Stata and also put into a new variable:

**gen Adjustment\_Factor = ( 1 / (100\*0.5^2) )**

Step 6: Calculate the total sampling variance and standard error

Finally, we can now complete our calculation of the sampling variance and standard error. Our estimate of the sampling variance is simply the product of the terms created in step 4 and step 5:

**gen Samp\_Var = SUM\_DIFF \*** **Adjustment\_Factor**

With the standard error the square root (sqrt) of the estimated sampling variance:

**gen SE = sqrt(Samp\_Var)**

We can then look at our estimate of the standard error by writing and running the command:

**display SE**

Step 7: Conducting hypothesis tests and constructing confidence intervals

Recall that the number of degrees of freedom is equal to the number of replicate weights minus one. As TALIS includes 100 replicate weights, there are 99 degrees of freedom. This has implications for the critical value to be used when conducting hypothesis tests and constructing confidence intervals. To see this, follow this link:

[**http://www.socscistatistics.com/pvalues/tdistribution.aspx**](http://www.socscistatistics.com/pvalues/tdistribution.aspx)

and type:

* 1.985 in the T score box
* 99 in the DF box
* Selected the ‘Two-tailed’ test option

If you then click calculate, you will see that you are just about able to reject the null hypothesis of no difference at the five percent level (p = 0.049911). However, it you now use **1.984** as the T-score, you will be unable to reject the null-hypothesis at the five percent level (p= 0.050025). This illustrates that the critical value is approximately 1.985 (slightly higher than the value of 1.96 that is often used with large samples).

Therefore, to estimate the lower and upper 95 percent confidence interval limits for the average age of Swedish teachers, one would write in the do-file the following commands:

**gen LOWER\_BOUND = Theta\_Star - 1.985 \* SE**

**gen UPPER\_BOUND = Theta\_Star + 1.985 \* SE**

To then display these values, you can write and run in the do-file:

**display LOWER\_BOUND**

**display UPPER\_BOUND**

The final result is that the average age of teachers in Sweden is estimated to be 45.97 years, with a 95 percent confidence interval running from 45.45 years to 46.50 years.

**4b. The Stata ‘svy’ command**

The process outlined above is quite long and complex, particularly given the aim of simply estimating the average age of teachers in a single country. Fortunately, secondary analysts of the international surveys do not typically have to go through this process every time. Rather, Stata includes in-built commands that can handle complex survey designs, including the BRR weights used in studies like PISA and TALIS. Full details on the Stata svy command can be found by running:

**help svy**

In the two steps outlined below, we illustrate how one can estimate the average age of Swedish teachers using this alternative methodology.

Step 1: Specify the complex survey design

The first step when using the ‘svy’ command is to instruct Stata exactly how the complex survey design should be taken into account. With reference to TALIS and the BRR weights, the following information needs to be provided:

* Name of the final respondent weight (TCHWGT)
* Name of the replicate weights (TRWGT1 - TRWGT100)
* The type of replication method to be used (brr)
* Whether ‘Fay’s adjustment’ has been used and the value (.5)
* How the ‘deviations’ should be calculated (mse)[[1]](#footnote-1)

These instructions are fed to Stata via the ‘svyset’ command as follows:

**svyset ///**

**[iweight= TCHWGT] ///**

**, ///**

**brrweight(TRWGT1 - TRWGT100) ///**

**vce(brr) ///**

**fay(.5) ///**

**mse**

You should recognise most of these components from section 4a above. In particular, note that we tell Stata that we will want to apply the final teacher weight, that we want to use all replicate weights between values 1 and 100, that the BRR method is to be used, and that there is a Fay adjustment of 0.5.

**Step 2: Estimate the statistic of interest**

Now we have told Stata about the complex survey design, taking it into account in our analysis is (usually) straightforward. We simply have to add the prefix **svy:** to all our estimations. For instance, to estimate the average age of teachers in Sweden, one simply needs to write:

**svy: mean TT2G02**

Which will produce exactly the same results as we calculated manually in section 4a.



**Question 2**: **Calculate the proportion of teachers in Sweden who are male (including the 95 percent confidence interval):**

**Question 3:**

1. **The TALIS dataset includes the variable TT2G46J. What does this variable capture?**
2. **TALIS also includes the variable TT2G16. This captures the number of hours teachers typically work per week. Using this variable, create working hour deciles (i.e. 10 approximately equal groups each containing approximately 10 percent of the population).** Hint: The **xtile** command can be used to divide the data into deciles.
3. **What proportion of teachers in the lowest working hours decile ‘strongly agree’**
4. **How does this compare to the proportion from the top decille (the 10 percent who work the longest hours?)**

**Question 4. What is the 75th percentile of the teacher age distribution? (The variable name for teacher age is TT2G02). Can you calculate the standard error (applying the BRR weights?)**

**4c. The OCED ‘repest’ user-written command**

You will have noticed that it is not possible to answer question 4 using the Stata ‘svy’ prefix. This is because neither **sum** nor **qreg** are compatible with the svy command. Yet, in the official OECD report, they frequently present estimates of the percentiles of various distributions along with their standard errors. You can re-produce their figures using the ‘manual’ technique described in section 4a – but is their an easier way?

Analysts at the OECD have recently developed to Stata **repest** command. This extremely useful tool, which can be used with the PISA, PIAAC and TALIS datasets, handles the BRR weights in an appropriate way to correctly estimate standard errors.

As repest is a user-written command, you have to first download and install it. You can do this either one of two ways. Either run the command:

**findit repest**

Then click on this link (repest from <http://fmwww.bc.edu/RePEc/bocode/r>), scroll to the bottom of the page and click where is says (click here to install).

Alternatively, you can visit my website [www.johnjerrim.com/papers/pisacourse](http://www.johnjerrim.com/papers/pisacourse) where you will find a dofile called repest. Download this do-file to (e.g.) your desktop. Then upon it up in the do-file editor and run everything that is contained within the file.

Once you have installed repest, you can find out more about its capabilities by running the command:

**help repest**

Note the examples at the end of this file are particularly helpful.

We will first use this command to re-estimate the average age of teachers in Sweden, as we did in section 4a and section 4b. To begin, note that the repest command will only recognise the final teacher weight and replicate weights if the variable names are in lower case. In the datafile, the variable names are all in upper case. To change variable names from upper to lower case, we can type:

**rename TCHWGT , lower** /\*\*\* Final teacher weight \*\*\*/

**rename TRWGT\* , lower** /\*\*\* Each of the replicate weights \*\*\*/

We can then estimate the average age of teachers by typing:

**repest TALISTCH, estimate(means TT2G02)**

After represt, but before the comma, we write **TALISTCH**. This is telling the command that we are using the TALIS teacher dataset. After the comma, we write estimate(). Inside the bracket we write the command we want to estimate (**means)** along with the variable name. Your results should look like the below:



These results should be (almost) identical to those that were produced in section 4a and section 4b above.

**Question 5: Can you now use the repest command to answer question 4 above? (i.e. find the 75th percentile of the Swedish teacher age distribution).**

Hint: Use **help repest** to help you! (Particularly the examples at the bottom of the help file).

**Question 6: Can you now calculate p5, p10, p25, p50, p75, p90 and p95 for the age distribution of Swedish teachers.**

1. **To what extent does applying the final teacher weights and BRR replicate weights make a difference to my results?**

To conclude, we will now consider the impact the complex TALIS survey design has upon estimates. We will run an OLS regression model examining the relationship between working-part time (**TT2G03)** and satisfaction with the teaching profession **(TJSPROS)** controlling for age (**TT2G02) and** gender (**TT2G01).**

First, let us treat the data is if it were collected via a simple random sample. One would then simply use the Stata command:

**regress TJSPROS ///**

**i.TT2G03 TT2G02 i. TT2G01**

Please record the parameter estimate and the standard error for the “Part-Time (71-90%) group” in the relevant row in the table below.

|  |  |  |
| --- | --- | --- |
|  | **Beta** | **Standard error** |
| Simple random sample |  |  |
| Include teacher weights |  |  |
| Plus Huber-White adjustment to SE |  |  |
| Application of BRR weights |  |  |

Next, let us re-estimate this model but applying the final TALIS teacher weight. We do this by estimating the following command (please also record these results in the table above):

**regress TJSPROS ///**

**i.TT2G03 TT2G02 i. TT2G01 [pw = tchwgt]**

Now let us take into account the clustering of teachers within schools by making a Huber-White adjustment to the standard errors.

**regress TJSPROS ///**

**i.TT2G03 TT2G02 i. TT2G01 [pw = tchwgt] ///**

**, cluster (IDSCHOOL)**

**Question 7: Finally, please complete the final row in the table above by estimating this model applying both the BRR and final teacher weights.**

**Question 8: In this analysis does following recommended practise (rather than applying one of the more well-known approaches to handling survey data) make a substantive difference to your results (and the conclusions from the analysis that you draw)?**

**Exercise**

In all the computer workshops that follow, we will make use of the **repest** command. If you finish the instructions above, please continue to use your time by playing around with this command to produce different types of estimates. Likewise, please use this time to explore the TALIS dataset.

For example, you could estimate a series of OLS regression models, estimating the association between some demographic characteristics (e.g. gender) and one of the TALIS scales (e.g. teacher self-efficacy – TSELEFFS). You could then examine how this changes once a series of other factors (e.g. teacher age, employment status, education level completed) have been included as controls.

1. By default, svy brr computes the variance by using deviations of the replicates from their mean. By specifying the ‘mse’ option, the variance is computed using deviations of the replicates from the observed value of the statistics based on the entire dataset. [↑](#footnote-ref-1)