ARDx draft guide

University of the West of England, Bristol

Contents

1. Overview

ARDx is a research dataset covering the years 1998 onwards, designed for users of the Virtual Microdata Laboratory and the Secure Data Service. The basic unit of production of any business is the local unit (LU), or establishment. However, the data for ARDx are collected and organised at the reporting unit (RU) level; this reflects the smallest collection of local units able to provide the financial information required. In most cases, these will be LUs carrying out the same activity. Around 95% of the RUs in the ONS data have only one LU ie they are single-site businesses.

ARDx is created from two ONS surveys, the Annual Business Inquiry (ABI; 1998-2008) and the Annual Business Survey (ABS; 2009 onwards). The ABI has an employment survey (ABI1) and a second survey for financial information (ABI2). ABS only collects financial data, and so is supplemented with employment data from the Business Register and Employment Survey (BRES; 2009 onwards).

ARDx consists of two types of files: 'respondent files' which have reported and derived information from survey questionnaire responses; and 'universe files' which contain limited information on all business that are within scope of the ABI/ABS. The data are categorised into eight industrial sectors by 2-digit SIC code:

Table 1 Industrial sector classification in ABS/ABI

Sector	Abbr.	SIC 92/03 codes	SIC07 codes
Production	PD	1/2/10-41	1-39
Construction	CN	45-45	41-43
Catering	CA	55-55	55-56
Motor trade	MT	50-50	45-45
Retail	RE	52-52	47-47
Wholesale	WH	51-51	46-46
Property	PR	70-70	68-68
Other services	ST	60-93*	49-96*

* excludes those SIC ranges noted above

This reflects the historical development of the ABI; however, it still has relevance as question codes and labels are sometimes sector-specific.

There is also a 'register panel' containing all RUs observed over the period in one file; see section 3 for details.

1.1 ARDx versus ARD2

Many microdata researchers are familiar with the previous file, ARD2, which was created from 1973 up to 2008. ARDx is similar in principle (an attempt to provide a dataset consistent across time with frequently-used derived variables included) but has a number of important changes, detailed below. Users unfamiliar with ARD2 can skip this section.

1.1.1 Timing

ARDx only runs from 1998 onwards. The reasons for this are fourfold. First, ABI (which ran from 1998 until 2008) is similar in sampling method, structure and questions to the ABS, but differs from its predecessors in all three elements; hence the ABI-ABS can be treated effectively as one long survey with a break, but this is harder to sustain for the ABI's predecessors. Second, the ARD is regularly used to link to other ONS business micro datasets, but these are generally only available from the late 1990s onwards when ONS was formed. Third, ARD was created in the late 1990s by academic researchers when the only date available for research was pre-ABI; in contrast, researchers now have 15 years of data with consistency and coverage. Fourth, the IDBR, which provides the sampling frame for all ONS surveys and which can be used for weighting linked datasets, only exists from 1997 onwards.

For these reasons, and given the complexity of integrating the pre-ABI data, ARDx is restricted to 1998 onwards. However, it can be linked to ARD2 for pre-1998 research through the RU and LU references.

1.1.2 Variables

ARD2 created its own naming system to manage inconsistencies across 9 surveys over 30 years. It also created a set of 'Standard Variables' such as measures of GVA which were used by many researchers.

ABS now creates all the Standard Variables as part of its processes, and all variables are labelled appropriately. The definitions of the ABS-derived variables and the ARD2 Standard Variables have been checked and appear to be consistent.

As ABI and ABS are fundamentally the same survey, the decision was taken to use the ABS naming scheme, and make ABI consistent with it. For the derived variables, this required creating the necessary variables for the pre-2009 data. Hence, ARDx variables do not have the same name as ARD2 variables, but they are consistent with likely future development.

1.1.3 Employment

In ARD2 employment data came from the same firms that provided financial data. ABS only collects financial data, and employment data is extracted from BRES. Hence there is no longer an exact fit between employment and financial data; see below for further details. In addition, the employment variables collected have changed slightly.

2. Variables

2.1 Variable structure

Variables are of three types:

- IDBR-based information about the reporting unit, such ownership, type of business, employment or turnover from tax records, and weights
- Respondent data, indicated by "wq" followed by a question number eg wq100
- Employment data, beginning "empt_" or "empe_" depending on whether employment or employee data

2.2 ABS variables (2009 on)

The ABS team produces a spreadsheet covering both ABI (2006-2008) and ABS (2009-) data. The spreadsheet contains information on

- Which variables occur in which file, for which year, and for which sector of the economy
- A description for those variables, including the question number that appears on the respondent questionnaire
- Variable names used for derived variables
- A list of derived variables:
    - Labels for those variables
    - Code to derive some variables (eg “derived variable q400 = source q120+q121”)
    - The source for ratios to derive other variables (eg “derived from capex”)

An analysis of the spreadsheet confirms that

- Question numbers are not duplicated within sectors
- Question numbers repeated across sectors have roughly the same meaning, although the labels and derivation codes differ in small and insignificant ways; a visual examination of the label/code variants indicates that these differences are not significant
- Some codes appears as both ‘derived’ and ‘original’ (for example, q400 turnover); again a visual inspection shows the meaning is the same

Therefore, the spreadsheet acts as an accurate guide to the latest ABS variables. The aim of the program will be to standardise on that definition for forward compatibility.

The ABS spreadsheet is a useful reference, but requires simplification. ARDx users therefore have available a ‘quick reference’ called “summarised ABS vars”, with only one entry per number, noting which sectors that variable appears in, whether the variable is taken directly from the questionnaire or is derived, and, if the latter, how it is derived.

2.3 ABI variables (1998-2008)

The original ABI data is similar to the ABS data, but most of the derived variables were not created by the ABI/ABS teams. An analysis of the source data also reveals a number of problems with the pre-2009 listing of variables. First it reflects the current ABS, not past values. Second, it misreferences variables which can be both original and derived. Hence, the VML team has recreated these variables using the codes in the ABS spreadsheet. There are some problems data which the ARDx creation process should deal with, but researchers should be aware that they exist; it is unlikely that all problems have been dealt with, and the VML team would welcome feedback on errors or implausible distributions.

2.4 Employment variables

With the development of the ARDx, previous undescribed employment variables have been identified and defined. Due to the 3 different data sets; BRES, the ABI and the IDBR all containing employment variables, the new employment labels outline where the data has been sourced from (see Appendix for further details).

2.5 Non-ABI/ABS variables

ARDx contains information on stocks (including work-in-progress) and capital expenditure. Some information on this is collected on the AB/ABI but most of the information is imputed from the Quarterly Stocks Inquiry and Quarterly Survey of Capital Expenditure. These imputed variables are not distinguished in the ARD data except via the labels; researchers should be aware of over-interpreting results based upon them.

2.6 Selection variables

The ARD2 code created a variable 'sel_id' set to 1 or 0 depending on whether the RU was selected for ABI2 and sent a response (=1, otherwise 0). In ARDx two more informative variables are created, called 'selection_type' and 'selected' to avoid confusion in old programs. The values are set to

Code	Label	Meaning
		'selection type'
0	Respondent	Selected for ABI2 or ABS+BRES, valid return
1	Respondent employment only	Selected for ABI1 or BRES, valid return
2	Respondent financial data only	Selected for ABS, valid return
3	Non-respondent	Selected for ABI2/ABS, non-return/non-response
4	Non-selected	Not selected ABI2/ABS
5	RU out of scope	RU not in scope for ABI/ABS (see below)
6	local unit	Local unit
		'selected'
0	non-respondent main questionnaire	No returned ABI2/ABS data
1	respondent main questionnaire	Returned ABI2/ABS data

3. The Register Panel

3.1 General comments

The register panel generates a complete universe of RUREFs and existence for firms whether they are in survey scope for the year or not. This allows researchers to, for example, include dynamic effects in their analysis (is this an existing firm, a new one, or one that is just about to go out of business?).

Some RUs may 'disappear' from the ARDx universe and re-appear at some point later. Because RU references are not re-used by ONS, this means that the RU must have changed its classification at some point so that it was not in the 'universe' for the ARDx.

However, it would be wrong to calculate that a missing observation means that a firm does not exist (for example, when doing productivity analysis of entry and exit). The register panel therefore

imputes records for missing values. This is where the ‘out-of-scope’ selection type noted in section 2.9 above is created.

The imputed records have employment values created for them; this is because employment is the most commonly-used variable for stratifying data. Two versions of the register panel exist, allowing user to choose between them:

- In the file marked ‘interpolated’, missing values for employment are linearly imputed.
- In the file marked ‘IDBR’, missing values are fetched from the IDBR and (in the case of multiple-RU enterprises) allocated according to the number of RUs out-of-scope for the enterprise; imputation then continues as above for any remaining missing values

Because an RU being out-of-scope can only be determined when it is observed again, the register panel will impute differently in different periods. For example, an RU last observed in 2011 may not appear in 2012 and 2013 panels, but if it re-appears in 2014 then the 2014 Register Panel will have out-of-scope records for 2012 and 2013. For the same reasons, an out-of-scope observation cannot be the first or last observation of an RU.

The register panel also creates the variable age, the number of years the firm has been observed within the register panel.

3.2 Employment Variables

U= Present in the universe file

M= Present in the main ARDx file

Employment variable	Employment variable label	Source data	1998	1999		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Equation (if derived)
Total full time male employees point-in-time	Total FT male employees PIT	Currently unknown (suspected IDBR)	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	
Total part time male employees point-in-time	Total PT male employees PIT	Currently unknown (suspected IDBR)	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	
Total full time female employees point-in-time	Total FT female employees PIT	Currently unknown (suspected IDBR)	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	
Total part time female employees point-in-time	Total PT female employees PIT	Currently unknown (suspected IDBR)	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	
Total number of employees point-in-time	Total number of employees PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Working props not receiving payment	Working props not receiving payment	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	

PIT	PIT																				
Other unpaid workers not vols PIT	Other unpaid workers not vols PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Employment variable	**Employment variable label**	**Source data**	**1998**	**1999**		**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Equation (if derived)**
Total number of people in your employment point-in-time	Tot no of people in your employment PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Male full time employees year average	Male FT employees YA	Currently unknown (suspected IDBR)	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	
Female full time employees year average	Female FT employees YA	Currently unknown (suspected IDBR)	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	
Male part time employees year average	Male PT employees YA	ABI equation which uses *suspected IDBR* data	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	Male full time employees average / 2
Female part time employees year average	Female PT employees YA	ABI equation which uses *suspected IDBR* data	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	Female full time employees average / 2

Total employment year average	Tot Employment YA	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Total employment point-in-time	Tot employment PIT	IDBR	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Total number of employees point-in-time	Total number of employees PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Actual full time employees at the given point-in-time	FT employees PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Employment variable	**Employment variable label**	**Source data**	**1998**	**1999**		**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Equation (if derived)**
Actual number of part time employees at the given point-in-time	PT employees PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Working props not receiving payment PIT	Working props not receiving payment PIT	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Employment year average	Employment YA	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Number of employees year average	Employee YA	BRES	M	M		M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	
Total full time employees year average	Total FT employees YA	ABI equation which uses	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	Male full time employees average + female

		suspected IDBR data																				full time employees average
Total part time employees year average	Total PT employees YA	ABI equation which uses *suspected IDBR* data	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	Male part time employees average + female part time employees average
Full time employees equivalent year average	FT employees equivalent YA	BRES	U	U		U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	