Short User Guide to ABS Micro-data

Contents

1. Overview

The Annual Business Survey (ABS) - formerly the Annual Business Inquiry (part 2) - collects comprehensive financial information from businesses within the UK Business Economy. Cross sectional data from the Annual Business Survey (ABS) are available for the years 2008-2012. Micro-data are available for businesses in Great Britain excluding Northern Ireland. For details of the survey see the ABS Technical Report.

ABS dataset names:

1. YYYY_provisonal
2. YYYY_revised
3. YYYY_final

The name of each dataset indicates whether the data correspond to the provisional, revised or final ABS release.

Local and Reporting Unit Universe files:

1. lu_univYYY
2. ru_univYYY

There are **four** types of variables in the ABS datasets:

1. **Administrative data-** taken from the Inter-Departmental Business Register (IDBR)

The IDBR is the sampling frame for the ABS, and contains information on each firm linked through the firm's unique identifier (*ruref*). This unique identifier can be used to merge data from other ONS surveys. Data from the IDBR provide additional information not collected on the surveys that are of interest to users e.g. foreign ownership (for a full list of IDBR variables see ABS Metadata document).

2. **Survey properties**

These variables provide important information such as which form type was issued to the business (so users can reference specific questionnaires to see exact question wording and layout) and survey weights (a-weights and g-weights) used to gross survey data to obtain population estimates.

Data from the survey:

3. **Survey questions-** each question is given a question number

4. **Derived variables-** variables of interest that are derived from survey questions e.g. retail turnover as a percentage of total turnover

There are a range of variables in the dataset with the variable name format WQXXX e.g. WQ399. The WQ prefix indicates that the variable has undergone outlier processing, short-form expansion and/or imputation (see ABS Technical Report) and the **XXX indicates the question number on the questionnaire or the number associated with the derived variable**.

'WQ' variables correspond to either:

1. Survey responses from the ABS questionnaire
2. Derived variables from question responses

Response variables and derived variables can be distinguished using the **Question codes and derivations** spreadsheet provided in documentation section. Survey questions will be under the 'question codes' tab and derivations will be under the 'derivations' tab.

2. Understanding the contents of any given WQXXX variable

Steps:

1. Take the XXX number from the variable name (e.g. for variable WQ399 this would be 399)
2. Search for this number in the **Question codes and derivations** spreadsheet provided. Search the 'question code' and 'derivations' tabs. This will indicate whether the variable is a survey question or a derived variable.
3. Variable labels should give an accurate overall description of the variable, but the wording of the same question number can differ across questionnaires (industries), so for exact details of the question asked of a particular business users will have to look up the question number on the relevant questionnaire. This can be done by:

   a) Obtaining the form type code using the contents of the variable frm_type
   b) Opening the questionnaire with the corresponding code
   c) Looking for the relevant question number in the questionnaire. For derived variables all the separate component questions can be found on the questionnaire.

Example: Understanding Total Turnover variables

The labels on WQ399, WQ346 and WQ550 indicate that these variables relate to Total Turnover. Searching each variable (346, 399 and 550) in the **Question codes and derivations** spreadsheet shows that WQ346 and QW399 are question variables, and WQ550 is a derived variable. It also shows that **either** WQ399 **or** WQ346 will be present on each questionnaire, depending on which industry the questionnaire is targeting.

The derivation of question 550 is shown to be:

550 = 399 + 346 - 321 and hence WQ550 = WQ399 + WQ346 - WQ321

Where:

WQ550 = Total Turnover (derived)

WQ399 = Total Turnover excluding VAT (specific industries)

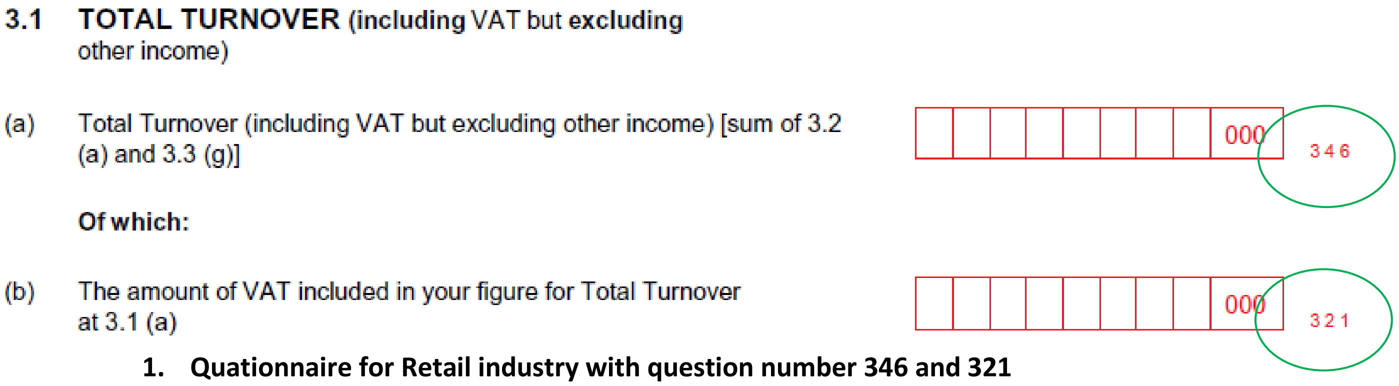
WQ346 = Total Turnover (specific industries)

WQ321 = Amount of VAT included in Q346

The variable WQ550 therefore represents Total Turnover Excluding VAT for businesses across all industries.

The screenshots below shows the question number indicator on questionnaires. These are highlighted in green circles. Note the differences in wording between question 346 and 399. Question 399 asks for Total Turnover excluding VAT whereas question 336 is including VAT.

<u>Figure 1- Screenshots of question numbers on ABS questionnaires</u>



3.1 TOTAL TURNOVER (**including** VAT but **excluding** other income)

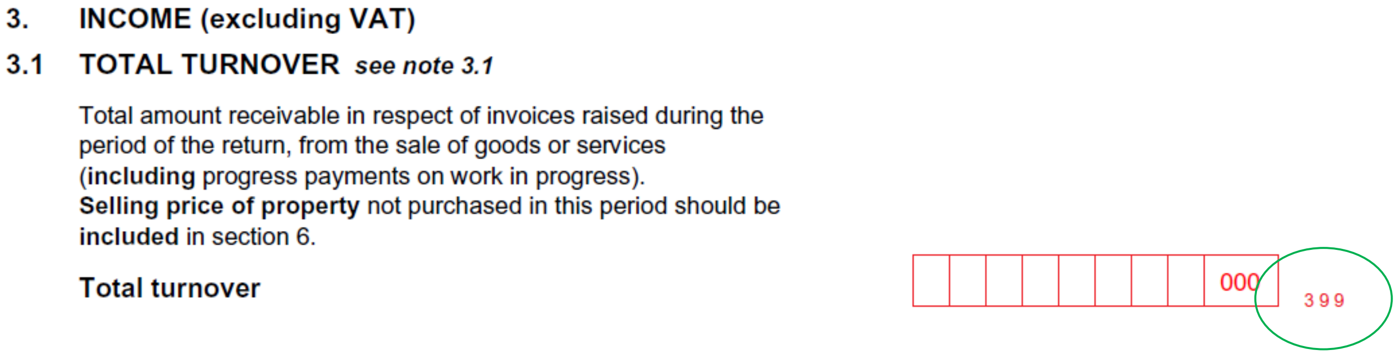
(a) Total Turnover (including VAT but excluding other income) [sum of 3.2 (a) and 3.3 (g)] 000 346

Of which:

(b) The amount of VAT included in your figure for Total Turnover at 3.1 (a) 000 321

1. **Quationnaire for Retail industry with question number 346 and 321**

2. **Questionnaire for Property industry with question 399**



3. INCOME (excluding VAT)

3.1 TOTAL TURNOVER ***see note 3.1***

Total amount receivable in respect of invoices raised during the period of the return, from the sale of goods or services (**including** progress payments on work in progress).
Selling price of property not purchased in this period should be **included** in section 6.

Total turnover 000 399

3. Reporting Unit and Local Unit Universe files

Reporting Unit (RU) and Local Unit (LU) universe files are provided with each yearly ABS dataset.

RU universe:

The RU universe file provides additional information on businesses in the sampling frame used for the ABS. The RU universe file will therefore contain businesses that have been sampled (are present in the ABS dataset) and those that have not been sampled. The reporting unit is the level at which businesses are sampled in the ABS. Information used in the calculation of weights for producing aggregate totals are present including important auxiliary information on turnover and employment from the sampling frame (see Technical Report section 5.4- estimation of totals):

Empment registered employment on the IDBR for the RU

Empsamp total employment in the sample by Gweight band at the time of sampling

Tosamp total turnover in the sample by Gweight band at the time of sampling

Emppop total employment in the population by Gweight band at the time of sampling

Topop total turnover in the population by Gweight band at the time of sampling

Empdeath total employment amongst businesses that died in the population between the time the sample was selected (November of reference year T) and questionnaire dispatch (January/February in year T+1)

Todeath total turnover amongst businesses that died in the population between the time the sample was selected (November of reference year T) and questionnaire dispatch (January/February in year T+1)

The RU universe file could therefore be used, for example, to re-calculate weights for aggregation if the sample data are changed (for example, if the researcher for some reason wishes to exclude some businesses from their analysis).

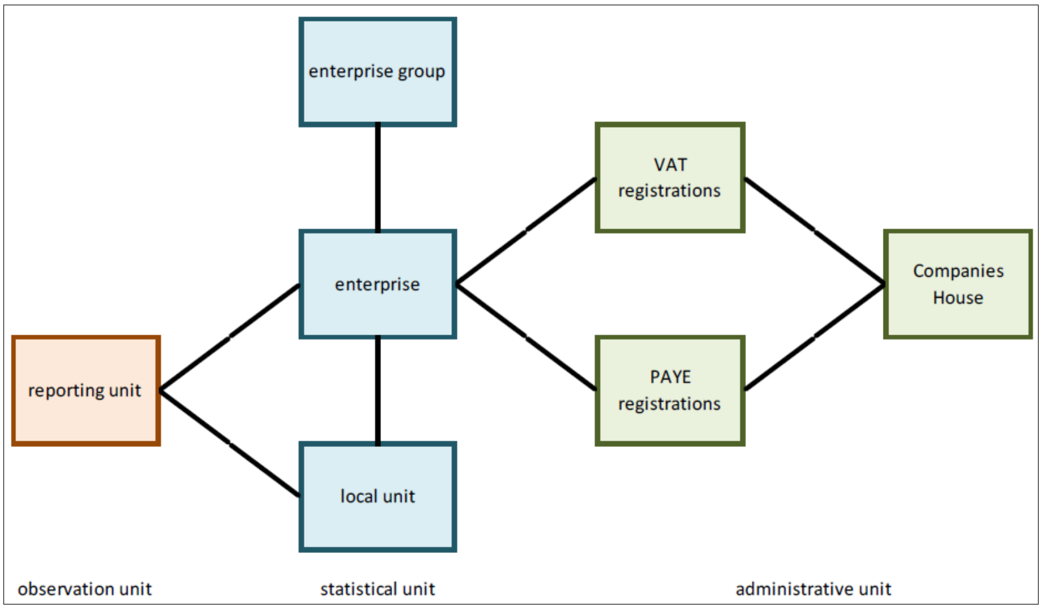
LU universe:

The LU universe file provides additional information on LUs that belong to RUs in the ABS universe (including those that were sampled). Figure 1 shows the structure of businesses for sampling purposes. The RU is the business unit which questionnaires are sent to and may cover the entire enterprise or LUs within an enterprise (see Technical Report- section 3). The RU will usually correspond to the enterprise. Additional information on the LU universe file includes:

Variable	**Definition**
ruref	RU unique identifier
luref	LU unique identifier
entref	Enterprise unique identifier

Various regional variables	Includes regional variables such as region, county, district and Nomenclature of Territorial Units for Statistics (NUTS) at different levels.
SIC code	SIC code of the LU
empment	registered employment on the IDBR for the LU
postcode	LU postcode

Figure 2- Business structure for sampling purposes



Source: ABS technical Guide

Regional analysis can be conducted more accurately by allocating proportions of the variable of interest at RU level to its corresponding LUs (if applicable). Regional analysis using RU level data may be misleading as RUs may represent the headquarters of businesses, or simply the office the business has chosen to respond to surveys from, even though the business activity may take place at many LUs spread across multiple regions. There are a number of ways a researcher may choose to apportion RU responses down to the LU level. For example, the published ABS regional estimates use regional apportionment based on LU employment (see Technical Report- section 5.8)

4. Variables used for published ABS results and other important variables

Variables used for published results	variable name
Total turnover	*wq550*
Approximate gross value added at basic prices (aGVA)	*wq613*
Total purchases of goods, materials and services	*wq499*
Total employment costs	*wq252*
Total net capital expenditure	*wq523*
Total net capital expenditure-acquisitions	*wq519*
Total net capital expenditure - disposals	*wq520*
Total stocks and work in progress - value at end of year	*wq271*
Total stocks and work in progress - value at beginning of year	*wq270*
Total stocks and work in progress - increase during year	*wq272*

Other important variables	variable name
a-weight Value	*aweight*
Employment (from IDBR)	*empment*
g-weight Value (Using employment, use for variables: wq446 - wq450)	*gwtemp*
g-weight Value (Using Turnover)	*gwtto*
Response type	*resptype*
Turnover (from IDBR)	*turnover*
Country of reporting unit	*country*
Enterprise Ref	*entref*
Form Type	*frm_type*
Legal Status	*acp_stat*
Region	*region*
Reporting Unit (unique ID)	*ruref*
Ultimate Foreign Ownership Code	*ultfoc*
Industry code using SIC07	*sic92*

5. Important points to note

1. Published results cannot be replicated using the datasets, as:

- Data from Northern Ireland are included in the published results but not in the micro-data
- Data from businesses in Section K (finance and insurance activities) are included in the micro-data but not the published results, as the ONS decided to remove this experimental series (Section K) from ABS releases for the reference year 2012 onwards due to the continued volatility of the data.

2. A large amount of missing or zero values for any variable will indicate that the question or derivation variable only covers certain industries
3. The questionnaire code (*frm_type*) will allow the user to reference the exact questionnaire answered by the business. The exact wording for question variables can therefore be obtained.
4. The unique identifier (*ruref*) can be used to link to other ONS datasets for additional variables
5. In order to obtain population estimates, the WQ variables should be grossed by multiplying each variable by the a-weight (*aweight*) and the g-weight (*gwtemp* for employment costs (*WQ450*) and *gwtto* for all other variables)
6. In order to produce regional estimates, users should link reporting units on the data file to individual sites (local units) on the local unit universe file using *ruref*. Users may then apportion values for reporting units to constituent local units (which may be spread amongst a number of regions). Regional analysis using the RU region variable may be misleading.
7. Imputation is carried out for some non-responding businesses (typically larger businesses). Imputed non-responding businesses can be identified using the *resptype* variable (where *resptype* = 1)
8. A fully balanced longitudinal dataset created from separate years will only contain the largest businesses, as they are the only group sampled each year (see ABS Technical Report for details of the sampling design).