FEATURE

Linking ASHE and LFS: can the main earnings sources be reconciled?

Catrin Ormerod and Felix Ritchie
Office for National Statistics

SUMMARY

This article describes a project to link and study the Annual Survey of Hours and Earnings and Labour Force Survey. This investigation looked at the differences between the earnings and hours information collected on the surveys. The results show that some perceptions over the accuracy of the surveys are misplaced, and that researchers can have more confidence in using the data.

Understanding the behaviour of earnings is of key economic importance, both at the level of the macroeconomy and when understanding the actions of firms and individuals. The UK has two main sources of earnings data: the Labour Force Survey (LFS) and the Annual Survey of Hours and Earnings (ASHE), formerly known as the New Earnings Survey (NES).[1] These two data sets are the basis for most micro- and macro-level analysis of the UK labour market. However, they originate from quite different sources and as such do not provide a single, incontestable view of the labour market. Moreover, as the two surveys are designed for different purposes and collect different information, they answer different questions about the labour market.

The surveys are both based on individual data, and so a natural question to ask is whether they could combined in such a way that:

- differences in the way they represent the structure of earnings could be analysed and clarified
- new analyses of the labour market could be addressed using a combined data set

However, the direct overlap between the two data sets is small. ASHE is a 1 per cent sample of the population; the LFS is a sample of about 60,000 people. Therefore only 600 people are expected to appear in both, throwing away 99 per cent of the observations. Moreover, the two do not share a common direct identifier; therefore it is almost impossible to match individuals from the two surveys. Statistical matching techniques ('data fusion') have been considered, but because the validity of inference in these merged data sets depends on the joint statistical properties of the variables sets, which are rarely known in advance, this has only had little interest.

This article uses an alternative method for linking based on creating small cell groups from the two data sets. These are used to create a combined data set, containing properties of both data sources, which can be analysed relatively robustly. The grouped cell method of linking data sets involves creating records in the matched data set for each possible permutation, based on common variables across both sources.

The resulting data set has two aims:

- to test statistical properties of the combined variable set to draw inferences about the two surveys and their descriptions of the labour market, and
- to analyse the data set for its own purpose

This article focuses on the first of the two aims, using the data set to test the characteristics of the LFS and ASHE in direct comparison. In doing so, several of the 'stylised facts' about the characteristics of the two data sets are addressed, and some of these are found not to stand up to this combined scrutiny. As a result, the use of the two data sets can be reconsidered.

The next section describes the two data sets and the ways they are used. The statistical background of data linkage is then reviewed, and the method used is described. The subsequent section discusses the results of linking the data, including benchmarking and consistency checks. The final section considers what other inferences can be drawn from the combined data set and suggest some paths for future work. For a more detailed description of the creation of the data, and some preliminary analysis on the combined data sets, see Ormerod and Ritchie (2006a).

Data sources and collection methods

ASHE is an annual 1 per cent sample of employees which results in approximately 140,000 records per year; it was first carried out in 2004, replacing NES. Employers are asked to provide detailed information on the hours and earnings of their employees and on the workplace characteristics. This information is almost always derived from employers' pay records.

The LFS is a quarterly sample survey of about 60,000 households living at private addresses in Great Britain. The survey seeks information on respondents' personal circumstances and their labour market status during a specific reference period. Information is collected on the individuals' personal characteristics as well as information about their hours and earnings in their main and second job (if they have one).

ASHE and LFS surveys collect similar information on earnings and hours worked, but the different methodologies and purposes of the surveys mean the detail and accuracy of the information collected varies. Earnings information collected from employer surveys should be based upon documentary evidence. In the LFS, information about the whole household is provided by one member, the respondent, sometimes without any reference to documentation such as pay slips. Where the respondent answers questions about other members of the household this is known as proxy response. Proxy response affects earnings data as the earning householder is more likely to be out (at work) when the interviewer arrives or telephones, and the proxy response is likely to be less accurate. Ormerod and Ritchie (2006b) demonstrate a significant rounding effect in the LFS. For this reason, employer-provided information on earnings is thought to be more reliable than employee-provided information.

The measure of hours worked reported is also likely to differ. Employers report paid hours, but individuals will tend to report the hours they actually work. Again, accuracy in household surveys is a problem: as well as proxy response, few people actually have a record of the numbers of hours they have worked in a week.

Both ASHE and the LFS offer an hourly wage rate stated by the earner, and one derived from dividing earnings period by hours worked. This information should be the same, but in practice in the LFS it can differ by considerable amounts. Both surveys collect a derived rate, but only ask for a wage rate if the employee is paid on an hourly basis.

For a household survey, a stated rate is more likely to be an accurate measure for pay per hour than the derived rate, as the latter requires more information to be recalled accurately (total earnings, total hours, and both, for the same period). For individuals providing both rates in the LFS, it has been shown that the distribution of the derived rate is much wider than the stated rate and more implausible. Again, proxy response may compound errors.

For employer surveys, the derived rate is seen as the best measure of actual hourly pay because it is based on actual earnings and hours worked. There may be some minor problems with hourly rates in ASHE (Griffiths *et al* (2006)). Nevertheless, ASHE figures on the whole are felt to be reliable.

The best source of earnings information is therefore the employer-provided ASHE, which also collects relatively accurate information about the job and the company (for example, employee's occupation, industry, whether the work is part time). However, the amount of personal information collected on ASHE is limited to what is provided from the HM Revenue and Customs records used to generate the sample: age and gender. It is reasonable for a household member to be able to provide more personal data and so the LFS collects, for example, ethnicity and disability. For this reason, the LFS survey is used when hours and earnings information is required to be broken down by personal characteristics.

In summary, ASHE provides accurate information on earnings, hours, and the characteristics of the employer, but little personal information. In contrast, the LFS has detailed personal information but there are concerns over the accuracy of the earnings information. There may, however, be advantages in linking these data sets to provide added value to both.

Linking methodology

ASHE and the LFS have a number of common variables which can be used for linking; variables of interest for comparison; and additional variables which can be used to supplement the main data sources. **Table 1** lists the variables used in this analysis.

The purpose of linking the two data sets is to bring them together using the matching variables (A), to produce a data set with earnings and hours information from both surveys (B1 and B2) and the supplementary information from the LFS (C). This allows the earnings and hours information to be compared across the two surveys. This could then support the idea of associating the supplementary information from the LFS (C) with the core information from ASHE (B1), as illustrated in **Figure 1**.

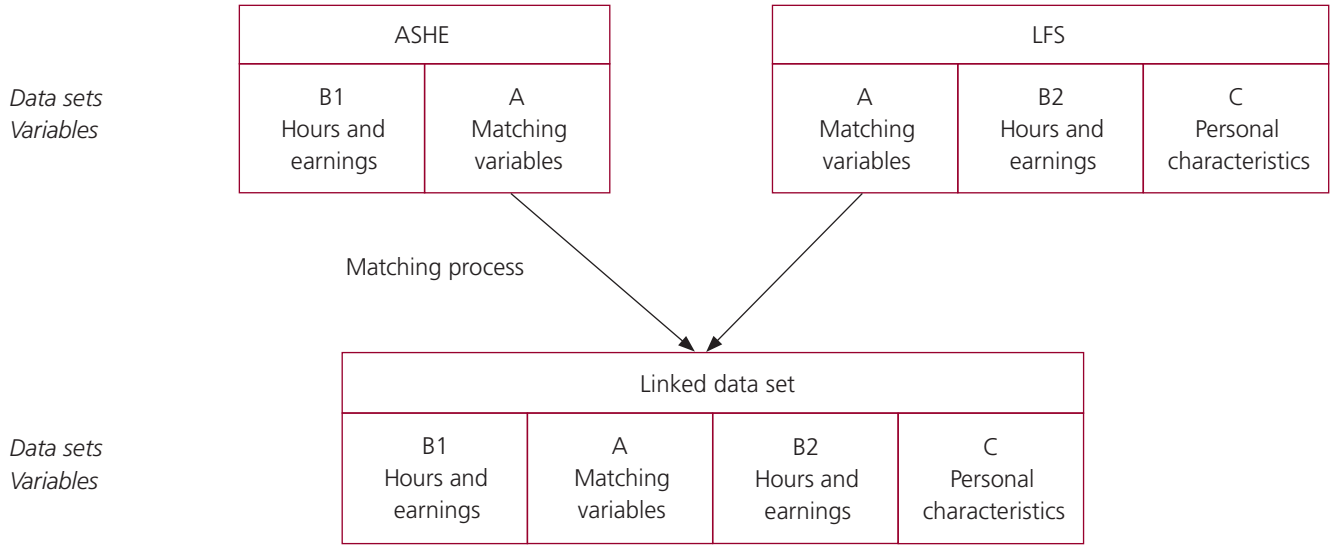
In matching, the assumption is that the linking variables are consistently collected across surveys. There may, however, be inconsistencies. For example, in the LFS, employees classify themselves as part time; in ASHE, 'part time' is a rigid definition based on hours of work. Even something as apparently obvious as 'industry' may differ between surveys. For example, employees may report the activity of their local office, not of the wider business; or they may confuse manufacturing with the sale of those manufactures.

A number of methods for linking data sets were investigated (see Lam and Ormerod (2005)). Because there is no exact identifier and little overlap between

Table 1
ASHE and LFS variables used for linking and further analysis

Matching variables (A)	Variables for comparison in ASHE and LFS (B)	Supplementary variables in LFS survey (C)
Age	Stated hourly pay	Ethnicity
Gender	Derived hourly pay	Disability
Full-time/part-time status	Hourly pay used for low pay analysis	Skill
Job status	Gross weekly pay excluding overtime	
Region	Basic hours worked in week excluding overtime	
Industry		
Occupation		

Figure 1
Aim of linking ASHE and LFS



the surveys, direct record linkage or probabilistic matching is not appropriate. Pure data fusion (linking cells one-to-one) requires a number of assumptions, not least because each observation in the LFS has potentially three observations in ASHE. Instead, a 'cell group' technique was employed, a generalisation of data fusion which creates matching groups of representative individuals, rather than a one-to-one match.

The grouped cell method of linking data sets involves creating a single record in the matched data set for each possible permutation of the matching variables. This record then represents a 'typical' person, for example a white male in a particular occupation and industry working in a permanent position on a full-time basis.

These cell groups can then be populated from the separate data sets. Each individual is given a reference code which contains the potential linking characteristics (for example, 2-digit industry codes, 1-digit occupation code, gender and age band). Within each survey, the information of interest for all records with the same cell reference is combined to produce an average value for the variable based on all contributing records. Where the survey has supplementary information of interest, a series of variables is produced for each possible category of the supplementary variable, indicating the proportion of records in the original data set having that category.

When the cell groups from the two surveys are combined, it is possible to compare 'representative' individuals from both surveys who have the same characteristics. By construction, the individuals in this cell, from whichever survey, all share the same identifying characteristics. Inferences can then be made about the representative individuals from the two surveys, and analysis of whether, for example, any differences in survey distributions are related to the characteristics of the individuals.

Note that where there is only one individual from a survey in the cell, this method collapses to standard one-to-one or many-to-one data fusion methods. Also, it is only possible to create a 'full' linked data set, if no permutation appears for only one of the data sets. This does not occur, and so there is some information loss when individuals have no match.

It is natural that some permutations are more common than others and some do not appear at all in the data set. When cells are created from a larger number of underlying records, these should provide a better estimate for earnings, hours and supplementary variables compared with those created from a small number. It is also common in the LFS for individuals to have missing information; this does not occur in ASHE, due to imputation for missing information. Having an individual in a group does not therefore guarantee that information on earnings is available.

In practice, it is possible to derive the cell reference at a number of different levels. A balance therefore needs to be struck between creating a data set containing detailed records and ensuring the number of records contributing to a cell is high enough to provide an accurate picture of the variables of interest for that typical record. A more detailed description of the linking method can be found in Ormerod and Ritchie (2006a).

Comparison of ASHE and LFS

As discussed in detail in Ormerod (2005), differences are expected in the hours and earnings information collected in both surveys. Previously, comparison of these sources has only taken place at the aggregate level. The process of creating grouped cells brings together individuals with similar characteristics and pools their information. The cell group data set therefore provides an opportunity to compare the hours and earnings information for jobs from ASHE and the LFS at a very detailed level for the first time.

The following variables were compared across the two surveys:

- hourly pay variables used to measure low pay. For the LFS, this is the stated hourly pay if it is provided, otherwise it is the derived hourly pay (gross weekly pay divided by usual total hours); for ASHE, a derived rate is used based on dividing basic, incentive and other weekly pay by hours worked during the week.
- stated hourly pay variables. Hourly rates are only applicable for certain types of jobs which are generally low paid; the number of individuals in the data set with this variable is therefore small.
- gross weekly pay, that is total earnings for a reference week
- basic actual hours worked during the week

One of the main developments in ASHE from NES was to improve the coverage of low-paid employees. Previously, the LFS was considered to provide better coverage of the low paid than NES, as

the LFS samples all individuals within a population regardless of their earnings; NES sampled individuals who were paid above the PAYE threshold. ASHE has expanded its coverage to include some of those individuals. Earnings from the two surveys may still differ at different parts of the distribution. In order to compare ASHE and the LFS across the entire earnings and hours distribution, investigations have been carried out at different cut-off points for the variables. **Table 2** shows the values of the cut-off points used in this investigation.

Numbers and consistency of cell group records

In the process of creating the grouped cell data set, information for individuals with the same characteristics is merged to produce information for the cell group. Some combinations of characteristics will not appear in either data set. This may be because that particular combination of characteristics is structurally implausible (for example, it can be assumed that there are no working miners aged 65 or over living in London), or because the combination is rare and none appear in the sample. If there are no individuals available to represent a particular combination of characteristics from one data set, but they appear in the other, the cell group will have missing information for variables which originate from the data set where they do not appear.

Even if individuals exist with the combination of characteristics to make up a cell group with contributions from both data sets, these individuals could have missing values for some of the variables of interest. The cell group variable can therefore be based on fewer individuals than the number actually observed in that category. An extreme case of this occurs when all individuals contributing to a cell group have missing values for a particular variable; the cell group then also has a missing value for that variable. Some cell records will therefore appear in the data set but not have any information for the variables of interest.

The value of a variable for a cell group will therefore be based on the number of individuals appearing in the originating data set with that combination of characteristics and a valid value for that variable. This will naturally be more reliable (in the sense of providing an unbiased estimate of the cell mean) if it is based on more individuals, as outliers will influence the variable less. Cell groups based on more common combinations of characteristics will therefore tend to be more reliable than cell groups based on rare combinations of characteristics.

Table 3 shows the number of cell groups with information for the variables of interest. The corresponding ASHE and LFS variables can only be compared for a cell group if there is both an ASHE and LFS value for the corresponding variables for that cell group. Table 3 also shows the number of records with information for the comparable variables based on five or more and ten or more individuals. The information for these cell groups should be more reliable than information for cell groups based on fewer than five individuals.

Of the 9 million theoretically possible cell groups, between 31,000 and 32,000 are observed each year. Almost all of these have relevant ASHE information, but the number with LFS information is lower at around 7,000. Around 5,000 cell groups have hours and earnings information from ASHE and the LFS, which allows these cell groups to be compared. The number of valid observations varies with the variable considered; less than half of the comparable records have stated hourly pay. Finally, restricting the analysis to groups with at least five or ten observations from each data set reduces the number of valid observations dramatically.

Hence, the cell group method does reduce the number of observations considerably compared with, for example, simple data fusion, where one aim is to maintain at least the dimension of the smaller data set. A key question then is whether the cell groups continue to provide an adequate representation of the data sets. To answer this, each of the original variables was regressed on the relevant cell group mean plus the characteristics of the cell group:

$$x_i = \alpha + x_g\beta + Y_g\gamma + Z_i\delta + \varepsilon_i$$

where:

x_g is the group mean value (for example, for hourly pay) for the group to which x_i belongs

Y_g are the characteristics of the group, and

Z_i are other characteristics of x_i

Table 2
Cut off values used during investigation for earnings and hours variables, 2004 and 2005

	Mean	10th percentile	25th percentile	Median	75th percentile	90th percentile
Hourly pay (£)	13	5	7	10	15	21
Weekly pay (£)	423	105	213	350	540	767
Basic hours (number)	33	15	29	37	39	40

Table 3
Numbers of cell group records: by year and reliability

	2004			2005		
	All	Based on 5+	Based on 10+	All	Based on 5+	Based on 10+
Cell groups	32,590			31,133		
With ASHE records	30,862			29,358		
Low pay hourly pay	29,453	6,048	3,136	29,271	6,071	3,168
Stated hourly pay	20,472	3,341	1,514	18,627	2,918	1,345
Gross weekly pay	30,862	6,244	3,230	29,347	6,093	3,177
Basic hours	29,650	6,245	3,230	29,285	6,093	3,177
With LFS records	6,945			6,852		
Low pay hourly pay	6,510	543	182	6,374	544	181
Stated hourly pay	3,171	178	48	3,042	156	47
Gross weekly pay	6,534	546	182	6,405	545	180
Basic hours	6,518	516	174	6,425	524	169
With equivalent ASHE and LFS variables						
Low pay hourly pay	4,957	526	179	4,838	528	178
Stated hourly pay	2,252	171	47	2,064	152	46
Gross weekly pay	4,663	529	179	4,561	529	177
Basic hours	4,919	456	1	4,803	475	0

The aim of this regression is to identify whether the cell group is a true representation of the underlying data set by identifying, for example, if certain cells compress the wage distribution unduly or if important combinations have been omitted. Significant coefficients on Y_g and Z_i could be indicators that there is some bias in the cell group construction.[2]

The results show some significant coefficients for age and region, indicating that, for these two variables, the decision to compress variation into subgroups may be biasing results. However, for most variables – industry, occupation, full time, gender, job type – there were no significant coefficients. Overall, it seems that the cell group method does retain the characteristics of the individual data points. However, it must be remembered that there may be some bias in the omitted observations – those in one survey with no counterpart in the other. Testing the characteristics of these observations has been left for future work.

Can ASHE be used to predict the LFS?

Although ASHE is used for official estimates of low pay, legal constraints mean that access is limited to government departments. In contrast, the LFS is widely used by researchers in labour economics as it is available to download in an anonymised form. The LFS is therefore the prime source of research material on earnings in the UK, and the concerns noted above about the accuracy of the LFS figures are directly relevant to the bulk of UK research. Although ASHE and the LFS have been compared at the aggregate level, this is the first time it has been possible to compare the two data sets at such a detailed level.

Ormerod and Ritchie (2006a) studied the relative properties of the two data sets. They compared the hours and pay variables described above by studying the distribution and correlation between ASHE and LFS values in the cell groups. These supported some 'stylised facts'; for example, the LFS earnings distribution is missing many of the high earners, but the LFS shows much greater variation in hours worked. They also used regressions to test the hypothesis that the LFS was a poor estimate of the true earnings value. These regressions suggested that, throughout the broad range of earnings, ASHE and the LFS were surprisingly close in the estimate of earnings for groups. The data sets diverge below the 10th percentile of the distribution, where there are few observations in the LFS, and above the 90th percentile, where the LFS does not have the high earners that ASHE has.

This is a significant result, in that it suggests that researchers using the LFS can have more confidence in the earnings data than was previously supposed.

However, one criticism is that regression analysis, in particular, does not capture fully the variability of the data. This can be addressed by studying scatter plots of the cell groups.

Figure 2 shows the relationship between ASHE and LFS cell group values for the earnings variable used in the official low pay calculations for 2005. Part (a) shows all cell groups while parts (b) and (c) show the cell groups restricted to those with at least five or ten observations from both surveys respectively. The reference line is drawn on the chart to show the hypothetical ideal where ASHE and the LFS agree exactly.

Three observations can be made. First, there is significant variation, but there is clearly a relationship between surveys which follows the reference line. Second, as the scatter plots are restricted to the more populated groups (parts (b) and (c)), the relationship becomes more defined. Second, there are notable outliers, where groups have low earnings on the ASHE data set but large earnings on the LFS. These persist even for the common groups, and require further investigation. The relationship

Figure 2
Scatter plot of ASHE compared with LFS hourly earnings variable used to measure low pay, cell groups, 2005

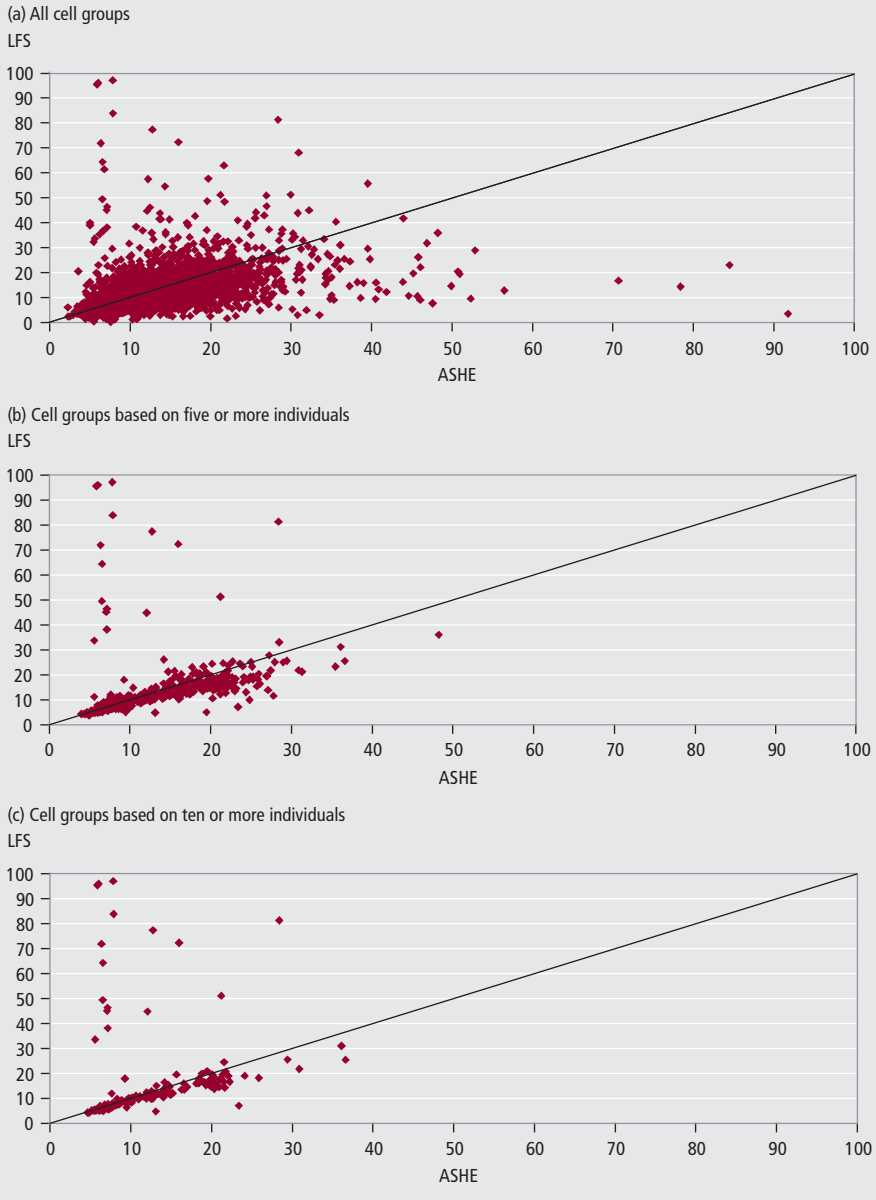
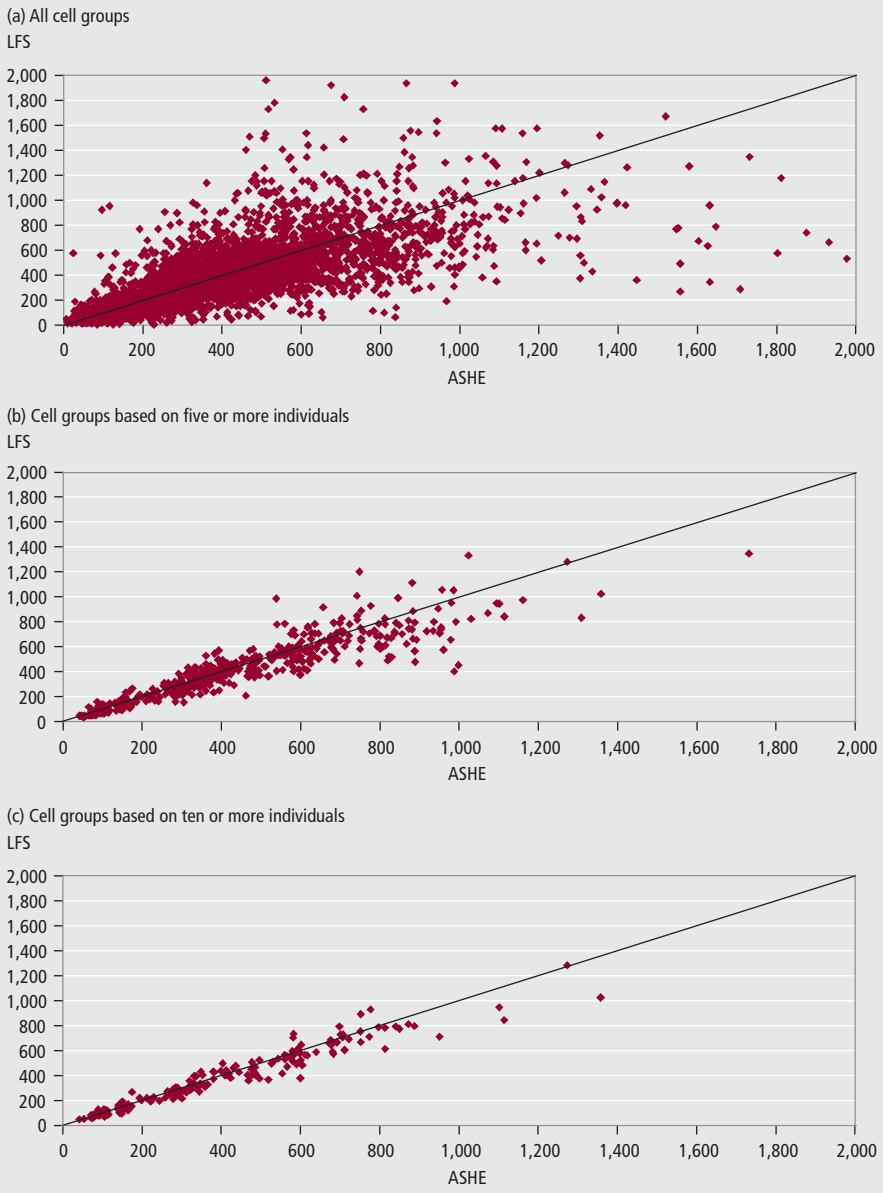


Figure 3
Scatter plot of ASHE compared with LFS gross weekly pay, cell groups, 2005



flattens beyond £21 per hour, the 90th percentile of earnings, in line with the simple regressions in Ormerod and Ritchie (2006a). Finally, although only the 2005 results are reported here, the results for 2004 are very similar.

Figure 3 shows similar figures for gross weekly earnings. It is clear that the relationship here is much closer and attenuates swiftly as the data points are restricted to the common groups. Again, there is some flattening of the relationship at high levels of ASHE earnings, but these are well beyond the 90th percentile of earnings.

What is noticeable is that the relationship seems to extend down to the bottom of the distribution. Regression analyses in Ormerod and Ritchie (2006a) failed because of the limited number of observations, but the scatter plots do seem to show that the close relationship continues into the bottom decile.

Figure 4 shows the hours figures. These do support the view that ASHE and the LFS report on hours differently. It is clear that, for full-timers, ASHE earnings data are clustered around standard hours whereas LFS hours show much more variation. Interestingly, for part-timers, there is only a weak relationship but a positive one, and one which is particularly noticeable in the more common groups.

In this case, the stylised facts are partially correct: the hourly data in the LFS are not comparable to ASHE, but only for full-timers. This is consistent with the way the data are collected. Part-time employees are more likely to be aware of, and work, the hours they are paid for, whereas full-timers are more likely to be salaried and to report hours based on their perception of hours. In both cases, ASHE reports the hours paid for.

There is thus strong evidence that the LFS is a more accurate record of earnings than was previously supposed. Ormerod and Ritchie (2006a) extended their regression analysis to incorporate industry and occupation dummies. These did not show statistically significant impacts, implying no persistent differences in the surveys as a result of industry or occupation. This is an important result, suggesting no systematic bias in ASHE-LFS linkages. Of course, there may be some more complex relationship not tested here, but on this broadbrush approach this is a reassuring outcome, and important for many of the researchers using LFS data who do not have access to the more reliable ASHE data.

In summary, gross weekly pay is very closely related across the entire distribution, even at low and high values. Basic hours differ in reporting for full-timers, those above 29 hours per week in this case. This may have caused the differences in the derived hourly rate variables at the low end of the distribution.

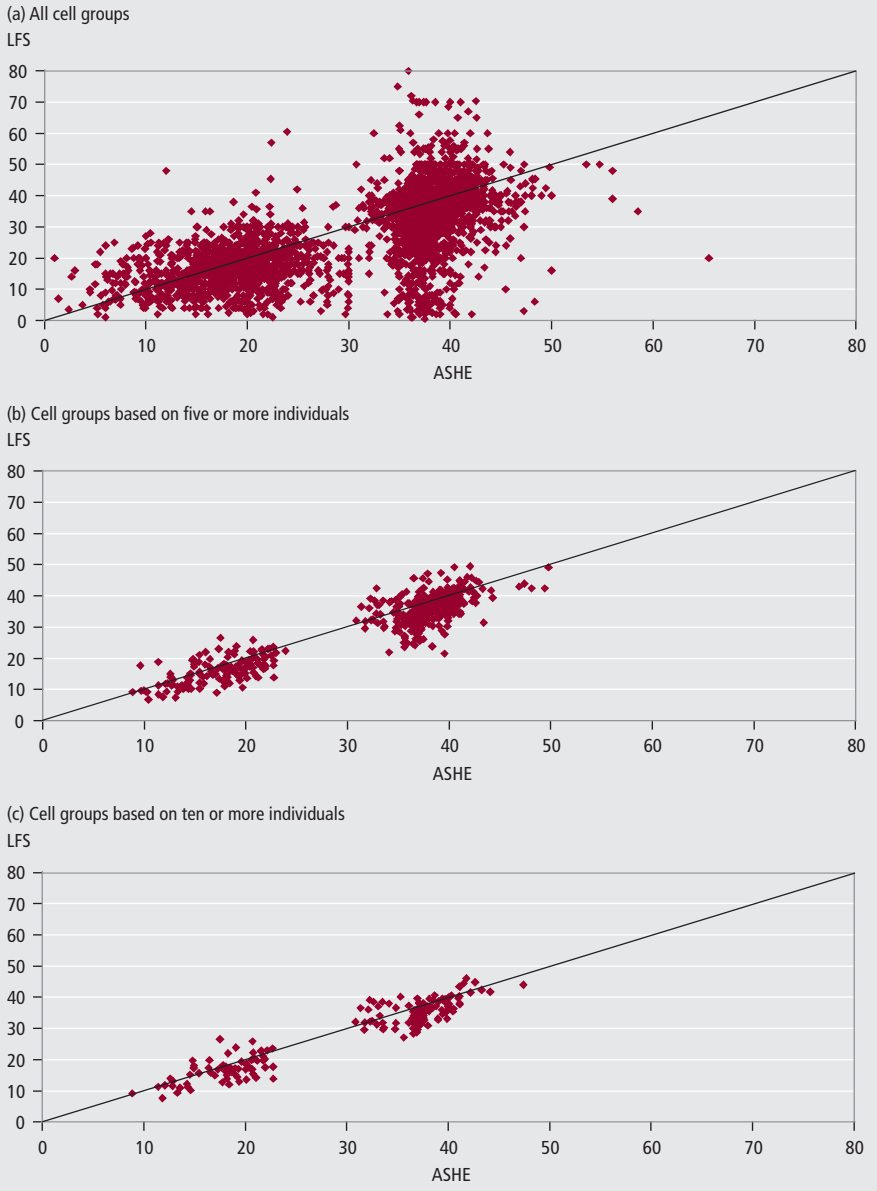
These results are somewhat surprising as the LFS has always been perceived as the poorer source of information on earnings. This investigation implies that analysis of earnings using the LFS may be more reliable than previously thought, and a breakdown of LFS earnings information by personal characteristics can be assessed with more confidence than in the past. The issue of low sample sizes for some rarer characteristics still remains, for example, some ethnicities, and this must be taken into account when commenting on the earnings distribution. Nevertheless, given the widespread use of the LFS for analysis, this has positive implications for much research currently underway in the UK.

What can be learnt from the linked data set

Although comparison of the data sets shows that they are more consistent than previously thought, analysis of the linked data set may still give insight into the data. Ormerod and Ritchie (2006a) looked at using the linked data set to analyse low pay and the distribution of earnings.

The broadbrush impression of consistency does hide some discrepancies. This is largely due to the fact that the cell group method accentuates the gaps

Figure 4
Scatter plot of ASHE compared with LFS weekly hours, cell groups, 2005



in the data where certain characteristics are concentrated. For example, in terms of ethnicity, the White group will be well-represented and distributed across a range of personal characteristics in the survey. Conversely, the sample size for individuals in the Chinese group is smaller and concentrated in similar groups. This is highlighted in the way that estimates for the Chinese group are very sensitive to the aggregating method.

This data set allows for alternative ways of scaling the LFS estimates to ASHE overall estimates. This makes comparison with the more reliable total estimates from ASHE easier when looking at particular groups of the population.

In terms of the questions raised in the introduction, the linked data set has proved useful in analysing the structure of earnings. One outcome of this project has been to identify some of the areas where discrepancies between data sets seem to arise; even if they cannot be explained at this stage, this is useful information when considering the design of the two surveys. However, the data are of limited use for analysis in their own right.

Conclusion and future work

ASHE (and previously NES) and the LFS have been used separately to examine earnings in the UK depending on the type of analysis required. ASHE has been used as the main source as it is thought to be more reliable since it is based on employer records. The LFS is used when estimates of earnings broken down by personal characteristics are required, as this source is richer in terms of the information on the individual. The sources have been compared at a high level in the past and it is known that many of the differences are due to the fact that the LFS is provided by the employee, without reference to documentation, and sometimes by proxy response. This investigation compares these sources at a very detailed level for the first time.

This investigation shows that, against expectations, the major data sets are more consistent than thought. This is particularly important because non-governmental researchers can only get easy access to the LFS, and so this is taken as the main source of earnings data. The linking exercise has raised some interesting issues about differences between the sources at a very detailed level and highlighted possible gaps in coverage. Overall, this report shows that researchers are justified in their continuing use of the LFS data where ASHE is not available or appropriate.

Notes

1 The ASHE survey started in 2004. It was developed from NES. The NES sample was extended to improve the coverage of the low paid, and imputation and weighting was applied to ensure the sample was representative of the population. For more information see: www.statistics.gov.uk/STATBASE/Source.asp?vlnk=1319

2 It could be argued that the variables should be interacted, as the process of building cell groups does force this. However, interacting all variables would have led to the simple recreation of the cell group means, and any lower level of interaction would lead to the same criticism of not fully identifying the bias. Hence, identifying possible sources of bias at the broadest and simplest level were chosen here.

CONTACT

elmr@ons.gsi.gov.uk

ACKNOWLEDGEMENTS

The authors would like to thank colleagues from the Office for National Statistics (ONS), Low Pay Commission, Scottish Executive and expert researchers in this area who attended a closed workshop at ONS London on 30 November 2006.

REFERENCES

Griffiths C, Ormerod C and Ritchie F (2006) 'Measuring low pay: Methods and precision' at www.statistics.gov.uk/cci/article.asp?id=1732

Lam K and Ormerod C (2005) 'Linking earnings data: Methods', Report for Eurostat, mimeo, Office for National Statistics

Ormerod C (2005) 'Linking earnings data: Inconsistencies and similarities between sources', mimeo, Office for National Statistics

Ormerod C and Ritchie F (2006a) 'Linking ASHE and LFS: Can the main earnings sources be reconciled?', Report for Eurostat, mimeo, Office for National Statistics

Ormerod C and Ritchie F (2006b) 'Measuring low pay: Focus points and rounding' at www.statistics.gov.uk/cci/article.asp?id=1731