Sample bias in microeconometric analyses of official microdata

Gian Fazio, Katherine H. Lam, Felix Ritchie
Office for National Statistics, UK

1. Introduction

Microeconometric analysis is typically carried out on sample datasets; population data is rarely available. A source of contention is therefore always whether the method of data collection should be taken into account, and how. Techniques such as two-step estimators to explicitly take account of the probability of selection are in common use, but these are usually concerned with isolating a sub-sample (for example, the probability of union membership in a sample of the labour force). More fundamentally, there is the question of whether the sample itself is representative for the purposes of estimation of a parameter of interest, which can be a sample mean or an average treatment effect.

The question of whether the sample is representative is not an easy one to answer because the population is not generally available. Therefore, econometricians and statisticians use proxy measures such as weighting or controlling for design variables (i.e. conditioning variables) to account for the sample selection. The choice of what, if any, proxy measure to use is a controversial issue. There is a large literature on weighting, originating mainly from the statistical literature. However, most applied microeconometric analyses ignore weighting; conditioning variables are used, but largely because they capture features of the data generating process that cannot be ignored or confound the interpretation of other parameters of interest.

It is clear that, in producing aggregate statistics, the weighting of variables has a significant effect. However, it is not clear that this effect persists in marginal analyses, and because the conditioning variables used to control for design effects often have a direct economic interpretation, weighting is used relatively rarely in econometric studies.

This paper is part of project looking at the analytical characteristics of structural business survey microdata from the Office for National Statistics (ONS). The volume of research on these data has exploded in recent years as access has been widened. Hence, there is a need to review the assumptions upon which much of this research is based. The use of these data has several advantages: the population characteristics are relatively well known; survey completion is legally enforceable; and company identifiers are consistent across years and datasets, enabling accurate linking. These characteristics mean that many of the difficulties of evaluating the importance of weights can be minimised. However, some analyses of linked data do not have independent sampling designs, which make it difficult to construct accurate weights for linked data.

The next section reviews the use and interpretation of weighted analyses. Section three looks at the ONS datasets in general. Section four briefly discusses two specific surveys used in this analysis, the Annual Business Inquiry and the E-Commerce Survey. Section five tries to analyse the impact of different weighting schemes on a simple log-linear model, and section six expands this to a range of models. Section seven concludes.

2. Weighted data and samples

2.1 Survey design and linking data

Chesher and Nesheim (2004) identify the main statistical issues resulting from sampling and data linking under two broad headings: survey design and measurement error[1].

Although the most straightforward method of selecting a sample from a population is simple random sampling, most modern data available to researchers would be the result of complex survey designs which could involve techniques such as stratification, multistage selection and unequal probability selection (Nathan and Holt (1979)). Of these, stratification on size, defined by employment, is by far the most important. It entails choosing independent subsamples of predetermined size from internally homogeneous but externally heterogeneous strata, therefore reducing sampling variation (Carrington and Eltinge (2000)). Generally speaking, stratification in NSI business surveys is biased towards larger firms as they are considered more representative in terms of employment and generating revenue. Hence, these unequal probabilities can bias the estimation of the parameter of interest. Before further analysis, and in order to reproduce a dataset which closely maps the original population, the weights from the survey design should be used to alter values accordingly. One method of eliminating this bias is to weight the observations by the inverse of survey design-dependent probabilities of being selected in the sample.

If two datasets (or the same datasets across different years) are to be linked, the resulting overlap will possess different properties compared to the two original sets: “A linked dataset can be regarded as the result of a single survey conducted with complex design which is the product of the designs of the contributing surveys” (Chesher and Nesheim (2004)). For example, in terms of stratification, this means that, providing the parent datasets had independent survey designs, the resulting overlap’s weights will be the product of the weights used in the original surveys. However, sample designs vary and the sample choice methods themselves may not be independent, requiring complex adjustment methods.

Complex sampling design can distort the information contained in the observable finite population. Typically, weights are more important for ensuring the unbiasedness of simple marginal statistics like means and tabulations. Conversely, more complex statistics that depend on the correlations between variables may remain approximately unbiased even if unweighted.

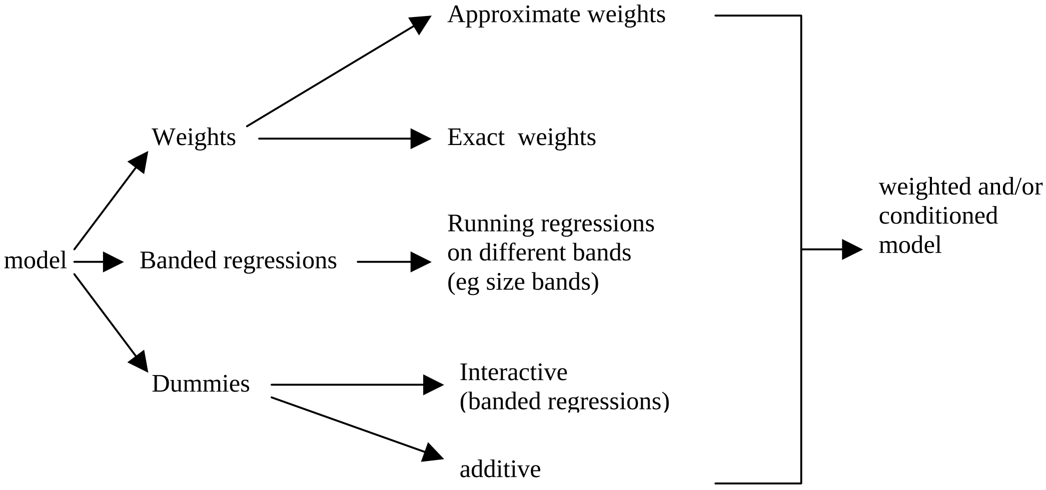
2.2 Weighting solutions

One would expect that the application of different weighting/conditioning techniques should yield similar results, but this is not necessarily the case. Even if a model is “correctly” specified (in terms of the underlying economic relation), different weighting schemes can have different impacts.

Figure 1, summarizes some of the approaches which could be used to account for sampling design or to control for particular variables.

[1] Measurement error issues are beyond the scope of this paper, since direct record linkage was utilised through a unique error free identifier.

Figure 1: Approaches to Conditioning and Weighting


Approximate weights
Weights
Exact weights
weighted and/or conditioned model
Running regressions on different bands (eg size bands)
model
Banded regressions
Dummies
Interactive (banded regressions)
additive

When an econometrician ignores the sample selection of the data, then the analysis rests on the following assumptions. It is assumed that the finite population of N observations is a simple random sample of size N from a population. The population data can be characterized by a regression model of the form:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is an N x 1 vector, X is an N x k matrix and β is k x 1 vector of parameters to be estimated. ε represents a vector of deviations from the linear relationship and has a property E(ε|X) = 0. The finite population of interest is defined to be:

$$\beta = (X'X)^{-1}X'Y \tag{2}$$

where the Y vector and X matrix are population quantities. For the purpose of this illustration, a class of stratified sample design is considered. For this class of design, the population is divided into h = 1,…, H strata by various geographic, industry and employment sizeband etc. For stratum h, a random sample of $n(h)$ observations are selected with unequal probabilities without replacement, where $P(h_i)$ represents the probability of selection for observation h_i. Consider a survey with a random sample stratified by employment size and industry. Running an unweighted regression estimates the mean effect across the two stratums. That is estimators for X'X and X'Y are $x'x$ and $x'y$ respectively, with

$$x'x = \sum_{h=1}^{H}\sum_{i=1}^{n(h)} X'(h_i)X(h_i)/P(h_i)$$

and

$$x'y = \sum_{h=1}^{H}\sum_{i=1}^{n(h)} X'(h_i)Y(h_i)/P(h_i).$$

These estimates will provide an estimate of β, namely:

$$\hat{\beta}_{OLS} = (x'x)^{-1}x'y \tag{3}$$

Running a weighted regression assumes that the population mean impact is desired. That is the estimators are depicted as follows:

$$\overset{\wedge}{\beta}_{wgt} = (x'wx)^{-1} x'wy \qquad (4)$$

where *w* is a diagonal matrix of weights. According to (Carrington et el, 2000), the argument for which $\overset{\wedge}{\beta}_{wgt}$ is a better estimator is "that varying probabilities of selection may lead the relationship between the dependent variable and regressors in the sampling distribution to differ from the relationship in the finite population." In theory, β_{wgt} converges to the population parameter, β.

However, weighting does assume that there is a single population parameter to be uncovered by appropriate adjustments to the importance of individual observations. This may not be an appropriate economic model. Researchers will add variables to the model to reflect economic structure; if these variables are also involved in the design of the survey, then it is not clear that there is any further role for weighting.

For example, banded regressions – that is, running separate regressions within each size band, for example – imply a different model for each stratum, including the error distribution if necessary. If the same size bands are used as the primary sampling criterion, then weighting by size is infeasible and irrelevant. If instead size dummies are added to a regression:

$$Y = \beta X + \Lambda SX + u$$

then weighting might still be relevant if the sample selection has an impact on variables not interacted with the size dummies; but weighting will have no further impact on the variables affected by the dummies.

One significant advantage of using conditioning variables or banded regressions is that it is no longer necessary to specify the exact sample proportions in the weights – this is automatically allowed for as long as the dummies are identified with exact subsamples. In other words, dummies or bands classify variables by sampling rules, irrespective or the size of the population. Weighting, on the other hand, needs to know the population characteristics. This is particularly important when linking across datasets, where sample proportions may differ or be unknown beyond the broadest level. It is also an issue for historical datasets where detailed sampling fractions may not be available for long time series.

Finally, weighting affects the interpretation of the coefficients. Since most business surveys disproportionately sample too many large firms, an unweighted regression will be driven by the data from these large firms, while a weighted estimate will be driven largely by data from smaller firms. This only applies if the weights reflect the number of firms of different sizes in the population. However, if they are based on firm turnover (the higher the turnover, the higher weights), then the weighted estimates would be even more sensitive to the data from the larger firms (assuming these have the larger turnovers).

For example, the UK retail sector is dominated by a small number of very large firms: under fifty companies account for over seventy percent of turnover (Haskel and Khawaja 2003). However, in terms of number of businesses, very small firms dominate to a larger degree than in many other industries. An unweighted estimate would give an accurate view of the main part of production; but it would ignore a whole swathe of companies. In contrast, a weighted estimate would be more appropriate for the bulk of companies, but may be of little use in determining the overall drivers of gross outputs in the sector.

It should therefore be clear that the choice of weighting scheme is not merely a matter of accounting for sample selection probabilities. It also incorporates the form of the estimated relationship, the interpretation of results, and the underlying theory.

3. The ONS datasets

The Office for National Statistics, like most NSIs, uses a complex stratified sampling scheme. It also imputes variables for smaller companies. The data collection strategy is governed by the need to provide cost-effective accurate estimates of gross totals. Despite this, there is a feeling that the datasets should also be sufficiently representative to reflect more complex correlations.

Microeconometric relations have been of less importance because confidentiality considerations have meant that relatively little NSI data has been available for research. However the recent, worldwide, expansion in access (see, for example, Ritchie (2004) for developments in the UK), the possibility of linking datasets, and the relatively well-known sampling proportions mean that weighting or conditioning is being re-evaluated. Several characteristics of the ONS business surveys make them suitable for analysing the effect of sampling and linking data.

First, the population can be identified fairly accurately. A single business register, continually updated, holds basic information on all businesses in the UK, down to the level of individual plants. This covers 99% of economic activity in the UK, and provides the sample frame for almost all the ONS business surveys and for many other government surveys.

Second, the ONS business surveys have statutory compliance, under the Statistics of Trade Act 1947. This, and an active legal enforcement unit, ensures a high response rate (typically 85%), with the vast bulk of non-responses due to changes in company structure[2].

Third, the ONS datasets have a common link field, printed out and read in by electronic systems. We can therefore assume that there is no significant error in linking datasets or years.

Finally, the sampling schemes are relatively well known.

There are some significant disadvantages of ONS datasets, due to the way they are collected:

- values for minor economic actors are often imputed. There is thus a correlation between measurement error due to imputation and the selection process: small companies, for example, are less likely to be sampled; those that are sampled are much more likely to have values imputed for them than larger companies.
- when reference groups for imputation are constructed, observed values are drawn from, for example size band A, to use as imputed values in size band B. Hence, one is effectively increasing the weight of size band A relative to B in analysis. This is not reflected in the sample selection probabilities.
- actual sampling probabilities are often only available for recent years, if at all

[2] There may be economically significant population non-response; for example, unproductive firms go out of business and so productivity estimates are biased upwards. We are concerned here with sample non-response – in other words, with sample selection probabilities, not sample inclusion probabilities.

4. The ARD and the E-Commerce Survey

For this paper two surveys were chosen as examples for analysis.

The Annual Business Inquiry (ABI) is ONS' structural business survey, its main source of detailed financial information on companies, and the most important component of GDP estimates. It is a statutory survey with an 82% response rate in 2003.

The Annual Respondents Database (ARD) is constructed from the ABI and its predecessors,[3] and currently consists of thirty years of data linked by consistent identifiers. It is the major research resource for analysts using the ONS business microdata and, since the microdata were made generally available for research in early Autumn 2003, has been used in more papers than all the other ONS datasets combined (see ONS(2005) for a list of research projects). For convenience we will refer to the "ARD" without distinguishing between the underlying survey and the linked dataset.

The e-Commerce Survey is a standard ONS survey into e-business practices in the UK, currently available as microdata from 2000 to 2002. Analysis of this dataset is relatively new, but it has been already been used in a number of papers (for example, Criscuolo and Waldron (2003), Rincon, Robinson and Vecchi (2004), Farooqui (2005)). This paper uses some of the simpler models from these papers to analyse the impact of weighting.

The ARD and the e-Commerce Survey are conducted using complex designs, stratified by industrial sector and size of firm with the ARD adding a third regional stratification. Although the datasets are unidentified, data points can be linked by the common company identifiers provided by ONS. The common set of observations appearing in both datasets is referred to as the "overlap". Table 1 presents basic statistics on the data

Firms included in the overlap tend to be significantly larger than the ones appearing in the ARD and in the e-Commerce. A firm randomly selected from the overlap dataset is likely to be approximately 10 times larger than one chosen from the ARD in terms of employment for the period considered (Table 2). Unsurprisingly a comparison of the turnover mean confirms this trend. Firms included in the overlap are bigger and seem to be more efficient than ones in the ARD or e-Commerce.

Both the ABI and e-Commerce surveys are carried out using stratified random sampling. In practice smaller firms have only a probability of being selected while bigger firms are automatically chosen (over 1000 employees for the e-Commerce and over 250 for the ARD). When the two datasets are linked, the resulting subset will, in effect, amplify this bias towards bigger firms. The approximate probabilities for a firm to be included in the datasets are listed in Table 3. However, these weights are broadly defined and probabilities of selection can vary within a size band.

Note that this may overestimate selection probabilities for small firms in the overlap, as ONS selection strategies are designed to avoid smaller companies completing multiple forms. For example, firms with fewer than 10 employees are taken off the sampling frame for three years after being selected for a survey. In effect, the overlap sample selection probability is zero for a firm with less than ten employees.

The ARD is not currently available with official survey weights prior to 2002. The probability of a firm being included in the survey depends on the firm's employment size, as shown in Table 3. For the ARD, this is only an approximation as the ABI survey is stratified

[3] The ABI was only introduced in 1997; prior to 1997, the ARD was constructed using information collected from a set of industry-specific surveys

in several stages. The ARD is stratified by three main design variables: region (England, Scotland and Wales); employment sizeband (0-9, 10-19, 20-49, 50-99, 100-249 and 250+); and 4 digit SIC.[4] The official survey weights summary statistics are presented in Table 4.

For the ARD, firms located in different areas of the country might possess dissimilar features and this could explain the discrepancies in studies researching the same questions but using different datasets. Table 5 confirms what was reported in Table 2 as companies in the South East and London tend to be larger than ones found, for example, in Wales or Scotland.

Finally, the three datasets are considerably different in terms of the sectors they cover, and these differences are not constant over time. For example, in sector 51 (Wholesale Trade) in 2000, 10.5% of the companies selected in the ABI belonged to this sector compared to almost 17% for the e-Commerce survey and 18% for the overlap. In 2001 the share was around the 10% mark across the three surveys. In 2002, however, the issue re-emerges: the ARD is still composed of about 10% by sector 51 firms, but the values for the e-Commerce and the overlap have returned to 2000 levels of over 18% (Fazio, Lam and Ritchie, 2005). There is no obvious explanation for this variation.

5. Testing weighting models

The complex sampling design can distort the information contained in the observable finite population. The ARD and its overlap with the e-Commerce are used in this section, as it allows us to see how applying weights and conditioning affect results from different datasets.

Following the work of Rincon, Robinson & Vecchi (2004)[5] we use the following log linear production function to capture the effects of e-buying and e-selling on productivity:

$$\ln Y_i = \beta_0 + \beta_1 \ln K_i + \beta_2 \ln L_i + \beta_3 \ln M_i + \beta_j \sum_{j=1}^{4} Etrade_{ji} + \beta_5 Multi_i + \beta_6 FO_i + \varepsilon_i \quad (1)$$

with the variables representing gross output, capital, employment, materials, e-trading, multi-plant status, and foreign ownership. Etrade is aimed at capturing the effects of e-trading and it is split into three binary dimensions equalling 1 when a particular characteristic is present. These are: e-buy, e-sell, and both e-buying and e-selling.

Accounting for the geographic location of the firms and for the sector in which they operate, produces a clearer model as the effects of e-Commerce can be isolated. In order to control for regional and sector variations we introduce, in equation 1, another two sets of dummy variables: region, which contains 10 regional dummies, and Industry which is made up of 53 industry dummies (following the 2-digit SIC classification).

$$\ln Y_i = \beta_0 + \beta_1 \ln K_i + \beta_2 \ln L_i + \beta_3 \ln M_i + \beta_j \sum_{j=1}^{4} Etrade_{ji} + \beta_5 Multi_i + \beta_6 FO_i + \\ \beta_k \sum_{k=1}^{10} \mathrm{Re}gion_{ki} + \beta_l \sum_{l=1}^{53} Industry_{li} + \varepsilon_i \quad (2)$$

with k=10 regional dummies and l=53 industrial ones.

[4] see Fazio, Lam and Ritchie (2005) for more details on sampling design of ARD, and Jones (2000) for the methodology used in constructing the official weights.

[5] This model was first used in Atrostic and Nguyen (1999), but here Rincon et al paper is the reference as it uses the same dataset.

Note that the dummy variables here play a dual role. First, they have meaningful economic interpretations as shift functions in the production process – for example, the utilisation of labour differs between the steel industry and retailing. The purpose of these dummies is to qualify variables and improve the efficiency and robustness of estimates.

However, they also represent the sampling characteristics of the data, where their role is to give appropriate weight to the different sampling strata so that the "better fit" model can be estimated accurately.

Regressions were run separately for the three years and for two different datasets: overlap sample and the whole ARD sample. The results are not significantly different across these analyses, and so a limited number of outputs are reported here. The results for the Etrade variables are small and mostly insignificant, so we will mainly concentrate on reporting coefficients for other variables. Throughout this paper, we will look at a single year of 2002.

Figure 3 reports the coefficient estimates and the t ratios for the unweighted regression (equation 1). All coefficients are statistically significant with the exception of the multi-plant and Etrade variables.

Regression diagnostics are carried to check for outliers and heteroskedasticity in the sample. Figure 3.ii) shows the plot of log gross out (lnY) against one of the explanatory variables, log employment (lnL) with fitted values under the model defined by (equation 2), both for the ARD and overlap sample. There are some extreme outliers in the dataset which may cause for concern. However, leverage tests indicate that these outliers are not influential in model estimation.

Diagnosing for heteroskedasticity, regression residuals are plotted against the fitted values for each sizebands in the ARD. For small firms the plots show a curvature in the pattern of residuals, indicating model misspecification, probably caused by the omission of significant main effects and/or interactions. There is also evidence of heteroskedasticity. Figure 4(ii) illustrates the reduction of misspecification when industry and region are included in the model, with the curvature much reduced. However, it is still present, indicating there are still missing interactions.

We considered two approaches to capture the distortion of complex design surveys.

First, we considered applying sample weights using approximated and official survey weights. As mentioned earlier, the ARD is not currently available with official survey weights backdating from 2002 surveys, approximated weights need to be calculated for analysis on earlier datasets.

Second, the ABI sampling scheme depends upon some characteristics of the population, such as regional, employment size bands and industry. These characteristics are considered in the sampling scheme and referred as 'design variables'. We can account for the sampling effect by conditioning our model with these design variables (Chambers, 2003).

If both surveys are primarily stratified by employment size bands, this is the design variable. Approximated weights can essentially be applied or alternatively, conditioning on employment size bands. For a more complex survey design, this is more difficult. Since the ABI survey involves multiple stratification, conditioning on all the design variables would mean enabling all interactions of these design variables. This would be difficult to implement, as the degrees of freedom could be too small and so individual parameters will be estimated with less precision, so we condition, instead, on subsets of the design variables. Conditioning on a subset of the design variables does mean that there is then the possibility that the sample selection is no longer ignorable, and this may be reflected in the results.

5.1 Regressing with approximate weights

Figure 5i) shows the coefficient values using equation 1 for the: unweighted regression; weighted regression (using official weights and approximated weights), regression with size band dummies; regional dummies; 2 digit SIC dummies and finally regression with all the stratification dummies. The regressions are run for both the ARD and the overlap sample, and are for 2002, the only year for which ARD official weights are currently available.

Looking at the unweighted regressions, the estimates from total ARD and the linked overlap sample do not appear significantly different (see figure 3i). This suggests that there is a same relationship with Y and X for all the units in the population from which the samples are drawn, given samples that have different X distributions. In other words, for the particular model specified here, sample design does not appear to affect the marginal analysis.

Differences between the weighted and the unweighted regressions are mainly evident in the ARD sample for the capital (K) multiplant (Multi) and foreign ownership (FO) coefficients. The capital variable is actually a computed variable which contains larger error margins for smaller firms as less data is available compared to larger companies. The other two dummies (Multi, FO) are strongly correlated to size of firms, as only relatively large enterprises are likely to have more than one plant or to have received foreign direct investment.

Hence there is no strong evidence that weighting is necessary in this model. In theory, applying weights to a model is safe as the impact of unnecessary weighting is to reduce the efficiency of the model, not to bias it. However, the fact that weighting does have the largest impact on variables where there is known to be significant measurement error does raise concerns.

5.2 Impact of Conditioning Variables

To comprehend the precise effects of the dummy variables used in the previous section, region, size and industry dummies were applied in turn to equation (1). For the regional dummy only three macro areas were used (England, Wales and Scotland); for size bands, six categories were constructed using the stratification for the ARD reported in Table 5; for industry, as in the previous section, two digits SIC code were used to generate 53 categories.

Figure 6 summarizes the differences between simple regressions and with the three sets of dummies. Although the regional dummies do not seem to produce any substantial impact on the estimates, the size dummy tends to increase the employment term while the largest changes occur when the industry dummies are included.

In most models, researchers tend to combine several dummy variables, often overlooking the individual effects of dummies. To analyse the interactive effect of the three dummies, the dataset was divided by regions and the size and industry dummies were applied, first individually and then combined additively. As figure 7 shows, the sign and magnitude of the combined dummies can be almost completely explained by the effects of the industry dummy (this is particularly evident for Scotland). Several regression were also run on smaller geographic bands (with particular attention to the South East and London areas), and results confirmed this trend.

Applying the same approach used for figure 7, regressions were run on different size bands while applying, first in turn and subsequently combined, the industry and region dummies.

Again, the overall effect of both dummies is almost totally produced by the industry dummy alone[6] (see figure 8).

Finally, regional and size dummies were included in regressions on different sectors. Although results vary considerably between different industries, there seems to be enough evidence to suggest that the industry element interacts more with size dummies than with regional ones. Hence, size and regional dummies seem to have an additive effect while applying industry dummies is clearly multiplicative.

5.4 Testing the significance of weights

In order to establish whether the various estimates of the regression models are significantly different, we used DuMouchel and Duncan's (1983) method of testing the significance of survey weights on the regressions analyses. The test requires one to add an extra set of regressors to the regression model of interest. These extra regressors are defined by the product of the weights and the original regressors. The test is based on the difference between the unweighted coefficients and the coefficients of the extra regressors in the same regression:

$$\Delta = \beta_{uw} - \beta_{w}.$$

The test is whether delta is essentially zero. This is equivalent to an unweighted F test of whether the coefficients associated with these new extra regressors are jointly significant.

The tests were run using a variety of models, starting with the simple Cobb-Douglas and cumulatively adding extra dummies (multiplant, foreign ownership, region, industry and sizeband). Table 6 presents findings. In the Table, "0" represents no significant impact; "XXX" represents a significant difference at the 1% level.

Our findings show that weighting methods have a statistically significant impact on the regressions, even when conditioning variables are used.

Using official and approximate weights does seem to make a significant difference in many cases, although it is worth noting that official weights are always more likely to have an insignificant effect than the approximated weights. When conditioning variables are used, the significance of official weights drops, and in some case it seems that official weights make no difference. This is a reasonable results if the dummies are effective in accounting for sample design. However, approximate weights do affect the outcome, which may implies an element of error being introduced into the model.

Finally, running banded regressions does not diminish the impact of weighting, even when regressions are run by size band (see Table 6).

5.5 Comparing the ARD and the Overlap using random samples.

As already highlighted, the characteristics of the overlap dataset differ considerably from the ARD (see section 3). However, regression analysis carried out so far does not seem to fully explain these discrepancies as results from the same model run on the overlap and the ARD differ only marginally. Since the ARD contains a considerably larger number of observations than the overlap, we proceeded to extract a series of random samples (containing a number of

[6] In figure 8, only three size bands are included, but the analysis was extended to the rest of the dataset and similar results were obtained.

observations equal to the one of the overlap) and considered their mean values. Figure 10 reports the differences the estimates for the explanatory variables of our model (equation 2) for the overlap, the ARD and for ten random samples from the ARD. Excluding the employment coefficient, the overlap's values fall within the range of the random samples. This suggests, that for regression purposes, the overlap can be treated as a random subset of the ARD.

It is worth noting that this model included non-interacted conditioning variables – region, industry, size – which seems to imply that these variables are doing their job effectively.

Although the main characteristics of the overlap sample relative to the ARD sample is that it contains disproportionately more large firms, the results so far suggest there are little correlation between firm size and values of the explanatory variables. One would not expect to see much difference between the overlap sample means for these variables and the corresponding mean values in randomly chosen subsamples from the ARD sample.

5.6 Misspecification

Does model misspecification give weights a bigger effect? Rincon et al (2004) noted that the e-Commerce variables in (1) and (2) were sensitive to the functional form. We followed their approach and extended the equations by adding a Heckman-style selection stage to the model:

The Heckman selection equation to allow for inequality is as follows:

$$Z_j = \alpha_{0j} + \alpha_1 Broadband + \alpha_{2j} WebUsers + \alpha_{3j} Experience + \alpha_{4j} Website + \alpha_{5j} Intensity + u_j \qquad (3)$$

after Rincon et al (2004).

The selection equation is an independent probit model to determine the firm's decision to etrade. Although, Rincon et al (2004) ran four regressions for each etrade variables (ebuy, esell, etrade-both and etrade-either) we will only look at etrade-either, as we found that the others yielded similar results. Both two-step and single-step estimators were used, but the results were similar. The coefficient estimates are reported in figure 11.

We can observe that as we include the conditioning variables, the effects of sample weights appear to be smaller (although this effect is minor). In particular, when we include industry dummies, the differences between the unweighted and weighted labour input parameter are less smaller. The Hausman test statistics also shows that these differences are statistically insignificant. Hence, this shows model misspecification give weights a bigger effect (a method put forward by Wooldridge, (1999) to detect misspecification).

5.7. Alternative Models

As previously suggested, it is probable that the results presented so far are highly affected by the log form of the model used. In this section, we experiment with different models to see whether the results obtained so far continue to hold.

Three different models were estimated:

- a **logit** model on the probability of e-trading, with a selection of firm characteristics as explanatory variables
- the same model estimated as a **probit**
- a crude **linear** model of labour productivity, with known scaling problems

The results, presented in **Table 7**, are surprising. In most cases, official weights do not seem to make much impact. However, when conditioning variables are added, there are differences, possibly suggesting an element of over compensation – particularly as adding conditioning variables without weighting does not seem to affect the other coefficients significantly.

6. Discussion of results

Thus far, the analysis of the paper is not straightforward. Weighting has an impact on coefficient estimates which seems to be statistically significant, even if the practical impact is small.

Conditioning on variables that drive sampling seems to have similar effect, and has a direct economic interpretation (for example, in the relationship between size and industry). There is evidence that conditioning variables ought to be multiplicative. On the other hand, the impact on estimates is small, and although adding conditioning variables statistically improves the regression, it does not significantly alter the other coefficients. It merely shifts the regression line.

The interaction between weights and conditioning variables is equally hard to draw out. The two-stage estimator indicated that improving model specification tended to favour conditioning variables rather than weighting. However, there still remains an impact from weighting which suggests that the conditioning variables as used here are not fully taking account of sample design.

It is interesting to note that the variables significantly affected by weighting are those where there is a close correlation between the variables and the selection process – either measurement error (in the case of capital, where the error is likely to be very large for small companies) or the high correlation of variables with certain selection groups (Multi, FO). However, testing the model without these variables still tended to produce significant differences - possibly because the model was now very poorly specified.

Finally, the comparison of the overlap with random samples suggests that, with conditioning variables included, the particular sample chosen is not significant. As this subsample was chosen from the sampled ARD data only, this does not mean that there is no sample bias in the ARD; however, given the different characteristics of the subsample, it is strong evidence that relatively simple measures to allow for different strata are fairly robust.

7. Conclusions and directions for future work

This paper was aimed at investigating some of the issues related to linked datasets. The overlap of the ARD and e-Commerce possesses different characteristics from the original surveys which can have effects on regressions analysis. The main reasons behind these discrepancies are related to different survey designs between the ARD and e-Commerce. To calculate the actual sampling weights related to the overlap can be extremely tedious as both parent surveys are stratified in different ways. When accurate sampling weights are not available, one has to rely on alternative techniques, such as the inclusion of dummies or banded regressions.

In particular we found that including size and industry dummies produces a multiplicative effect which can have a considerable impact on regression analysis increasing the efficiency and robustness of the model while regional dummies, seem to be additive. The existence of real effects in production suggested by mainstream economic theory is supported by shifts caused by the industry dummy.

However, these effects appear to have little to do with the sample selection. The evidence for the impact of weighting is ambiguous at best, and the pseudo-Monte Carlo experiment on random subsamples suggested that the particular combination of observations was not significant.

This results here are relatively robust to studies in different years (where possible) and with slightly different specifications. What has not been done yet is an analysis of whether the inferences drawn from this analysis are specific to these datasets, or more generally to ONS statutory business surveys. The ARD and e-Commerce Survey were chosen because published work indicated sample selection issues, but on closer analysis these have turned out to be specification differences.

This leaves two results. First, the two datasets studied seem to offer an accurate representation of the population at the level of unit relationships, and, by implication, at the aggregate level. This is particularly important given the central role of the ARD in both microeconomic analysis and in the production of official aggregates of production and wealth.

Second, in most work done using these data, conditioning variables are included as the rule due to their economic interpretation. Additional correction for sample selection bias seems to be relatively important. However, if these variables are omitted, weighting does not seem an adequate replacement – whether because of imprecision in the weights, measurement error in strongly weighted variables, structural differences in the population, or for some other reason. Given the inherent advantages of conditioning variables – economic interpretation, ignorability of sampling fractions, availability in all years for all datasets, lack of assumptions about representativeness – the recommendation from this work so far is that analysis on the ONS datasets should use these rather than attempting to construct weights.

This leaves open some further areas for analysis. First, this analysis concentrated on one dataset and variations around one or two models. While economically sensible inferences can be drawn from the results, they are always open to the criticism that results are data- or model specific. Further work will address this by carrying out a similar analysis on other ONS business surveys. As these surveys have a common legal and statistical framework, it is hoped that a common theme can be drawn across the ONS business datasets.

Second, it was noted that weighting does seem to have larger effects on variables with measurement error – specifically, the capital stock variable. This does raise concerns over the level of imputation required to create model based-datasets from survey data, and may also have implications for the imputation used to create the source data. This requires further investigation.

Bibliography

Brewer K., & Mellor, R. (1973) The Effect of Sample Structure on Analytical Surveys, *Australian Journal of Statistics*, 15 (3), 145-152

Carrington, W., Eltinge J. & McCue, K. (2000) An Economist Primer on Survey Samples, *Centre for Economic Studies Discussion Paper*, October 2000

Chambers, R., (2003) “Introduction to Part A: Approaches to Inference,” Chapter 2 of Chambers and Skinner (eds), (2003): *Analysis of Survey Data*, New York, John Wiley and Sons

Chesher, A. & Nesheim, L. (2004*) Review of the Literature on the Statistical Properties of Linked Datasets*, prepared for the Department of Trade and Industry

Chesher, A. (1991) *The Effects of Measurement Error*, Biometrika, 78, 3, 451-62

Clayton, T. Criscuolo, C., Goodridge, P. & Waldron (2003), *Enterprise E-Commerce: Measurement and Impact,* Economics Trends, Office for National Statistics London

Criscuolo, C. & Waldron, K. (2003), E-Commerce and Productivity, *Economic Trends*, November 2003, 52-57

Cross, P. & Manski, C. (2002) Regressions, Short and Long*, Econometrica*, Vol. 70, No. 1, 357-368

DeGroot, M. Feder, P. & Goel, P. (1971), Matchmaking, Annals of Mathematical Statistics, Vol 42, No. 2, 578-593

DuMouchel, W. &Duncan, G. (1983) Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples, *Journal of the American Statistical Association*, Vol 78, No.383 (Sep 1983), 535-543

Fellegi, I. & Sunter, A. (1969) A Theory of Record Linkage, *Journal of the American Statistical Association,* 64, 1183-1210

Farooqui, S (2005) E-Commerce, E-Business and Firm Performance, mimeo, Office for National Statistics

Holt.D., &Nathan, G. (1979) The Effects of Survey Design on Regression Analysis, *J.R. Statistics Society*, 42, Part 3, 377-386

Holt, D., Smith, T. & Winter, P. (1980) Regression Analysis of Data of Complex Surveys, *J.R. Statistics Society*, 143, Part 4, 474-487

Jones, G., (2000) “The Development of the Annual Business Inquiry”, *Economic Trends,* November 2000, Office for National Statistics.

Konijn, H. (1962) Regression Analysis in Sample Surveys, *Journal of the American Statistical Association*, Vol 57, No.299 (Sep 1962), 590-606

ONS (2005) *Business Data Linking Annual Report 2004-5*

Rincon, A., Robinson, C. & Vecchi, M. (2002) *The Productivity Impact of E-Commerce in the UK, 2001: Evidence from Microdata*, The National Institute of Economic and Social Research

Shah, B., Holt, M. & Folsom, R. (1977) Inference About Regression Models From Sample Survey Data, Research Triangle Institute, North Carolina

APPENDIX: List of Tables

Table 1: Dataset observations

DATASETS	2000	2001	2002
ARD	53,200	54,690	55,263
E-Commerce	7,319	9,669	8,590
Overlap	**1,930**	**3,507**	**3,019**
Total	58,589	63,852	60,834

Table 2: Summary statistics for 2002

2002			MEAN			STANDARD DEVIATION		
SAMPLE	variable	obs	unweighted	official weights	approx weights	unweighted	official weights	approx weights
ARD	employment	55263	179	522	590	1604	2887	3017
	turnover(000s)	55263	25008	64883	727456	251909	409839	460568
Overlap	employment	3019	1400	2128	3011	5554	7005	8412
	turnover(000s)	3019	196773	299853	21820	757384	953147	245280

Table 3: Approximate probability of firms being selected

	Employment sizeband						
	A			B		C	D
SAMPLE	0-9	10-19	20-49	50-99	100-249	250-999	1000
ARD	0.03	0.1	0.2	0.35	0.67	1	1
E-Commerce	0.05	0.03	0.03	0.08	0.08	0.2	1
Overlap	0.001	0.003	0.006	0.028	0.054	0.2	1

Table 4: Summary statistics of official survey weights

	2002				
	observation	mean	std dev	min	max
ARD	54324	31.33	43.39	1	607
Overlap	3019	14.61	37.61	1	800

Table 5: Distribution by region

	A. ARD			B. Overlap			A-B		
	2000	2001	2002	2000	2001	2002	2000	2001	2002
	%	%	%	%	%	%	Diff	Diff	Diff
North East	3.10	2.88	2.98	2.85	3.45	3.81	0.25	-0.57	-0.83
North West	10.39	9.75	9.70	11.92	10.41	10.83	-1.53	-0.66	-1.13
Yorkshire & the Humber	8.2	7.64	7.63	8.24	8.13	7.75	-0.04	-0.48	-0.12
East Midlands	7.37	6.78	6.95	7.62	7.3	7.52	-0.25	-0.52	-0.57
West Midlands	9.00	8.49	8.53	10.47	10.21	10.33	-1.47	-1.72	-1.81
Eastern	9.20	8.79	8.99	7.93	8.01	9.08	1.27	0.78	-0.08
London	13.6	14.00	14.06	15.39	16.57	17.22	-1.73	-2.57	-3.16
South East	14.69	14.43	14.29	17.41	15.43	15.73	-2.72	-0.99	-1.45
South West	8.25	8.12	8.19	6.48	6.76	6.79	1.78	1.36	1.40
Wales	3.80	7.47	7.21	3.63	5.08	3.51	0.18	2.4	3.70
Scotland	12.33	11.64	11.47	8.08	8.67	7.42	4.25	2.97	4.05
Total	100	100	100	100	100	100			

Table 6: Summary of test statistics

	M&D test			
Models	**ARD**		**OVERLAP**	
	U-OW	U-AW	U-OW	U-AW
Model 1i: Simple Log Linear	XXX	XXX	XXX	XX
Conditioned with sizeband dummies	XXX	XXX	XXX	XXX
Conditioned with regional dummies	XXX	XXX	XXX	X
Conditioned with industry dummies	XXX	XXX	XXX	XXX
Conditioned with all dummies	XXX	XXX	XXX	XXX
Banded regression: sizeband				
Sizeband =1	XXX	XXX		
Conditioned with regional dummies	O	XXX		
Conditioned with industry dummies	XXX	XXX		
Conditioned with all dummies	XXX	XXX		
Sizeband =2	XXX	XXX		
Conditioned with regional dummies	XXX	XXX		
Conditioned with industry dummies	XXX	XXX		
Conditioned with all dummies	XXX	XXX		
Sizeband =3	XXX	XXX		
Conditioned with regional dummies	XXX	XXX		
Conditioned with industry dummies	XXX	XXX		
Conditioned with all dummies	XXX	XXX		
Sizeband =4	XXX	XXX	XXX	XXX
Conditioned with regional dummies	XXX	XXX	XXX	XXX
Conditioned with industry dummies	XXX	XXX	XXX	XXX
Conditioned with all dummies	XXX	XXX	XXX	XXX
Sizeband =5	XXX	XXX	XXX	XXX
Conditioned with regional dummies	XXX	XXX	XXX	XXX
Conditioned with industry dummies	XXX	XXX	XXX	XXX
Conditioned with all dummies	XXX	XXX	XXX	XXX
Sizeband =6	XXX	XXX	O	X
Conditioned with regional dummies	XXX	XXX	O	XXX
Conditioned with industry dummies	XXX	XXX	XXX	XXX
Conditioned with all dummies	XXX	XXX	XXX	XXX

Ho: Coefficient estimates are not statistically different

O: Null not rejected

X: Null rejected at 10%

XX: Null rejected at 5%

XXX: Null rejected at 1%

Abbreviations:

U = unweighted

OW = official weights

AW = approximated weights

Note: Banded regressions are also run for the 3 macro areas and 8 different sectors-all test shows a significant impact on weighting

Table 7: Summary of test statistics

	M&D test			
Models	**ARD**		**OVERLAP**	
	U-OW	U-AW	U-OW	U-AW
Model 3i: Logit	O	XXX	O	XXX
Model 3ii: Logit with sizeband, regional and industry dummies	O	O	n/a	n/a
Model 4i: Probit	O	XXX	O	O
Model 4ii: Probit with sizeband, regional and industry dummies	O	O	n/a	n/a
Model 5i: Linear	XXX	XXX	X	XXX
Model 5ii: Linear with sizeband, regional and industry dummies	O	O	O	O

Ho: Coefficient estimates are not statistically different

O: Null not rejected

X: Null rejected at 10%

XX: Null rejected at 5%

XXX: Null rejected at 1%

Abbreviations:

U = unweighted

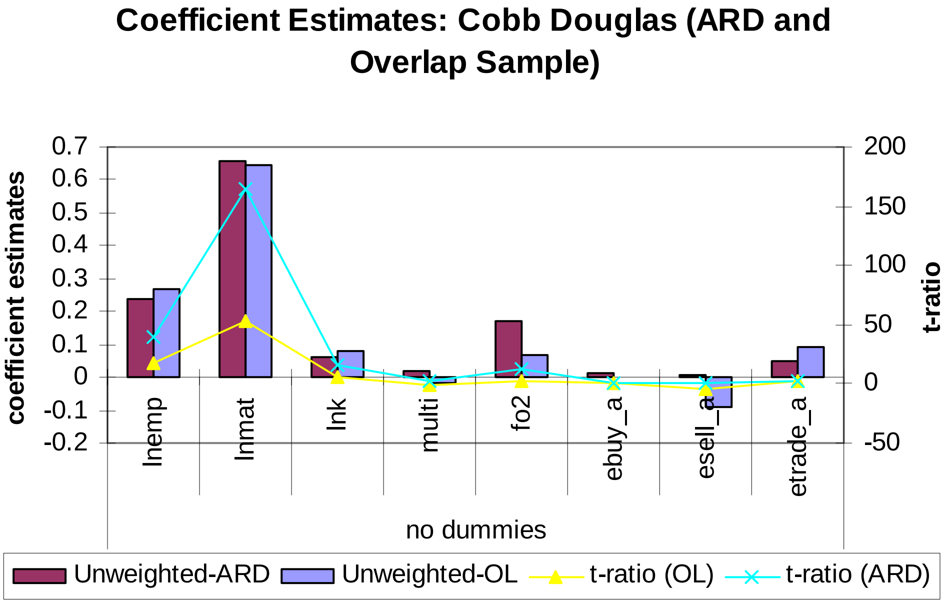
OW = official weights

AW = approximated weights

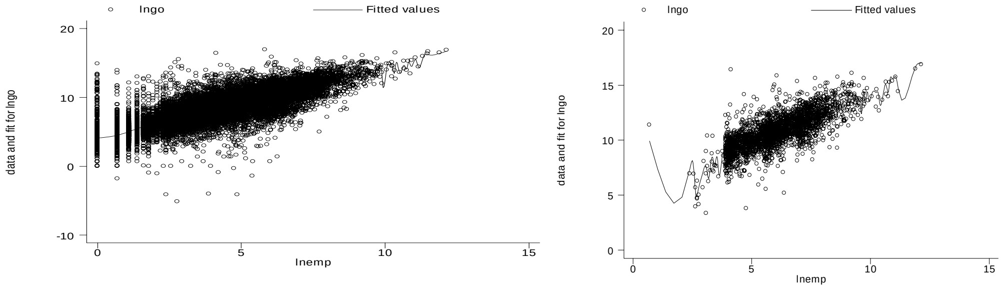
APPENDIX (Figures)

Figure 3: Regression results

i. Coefficient Estimates:



ii. regression plots:

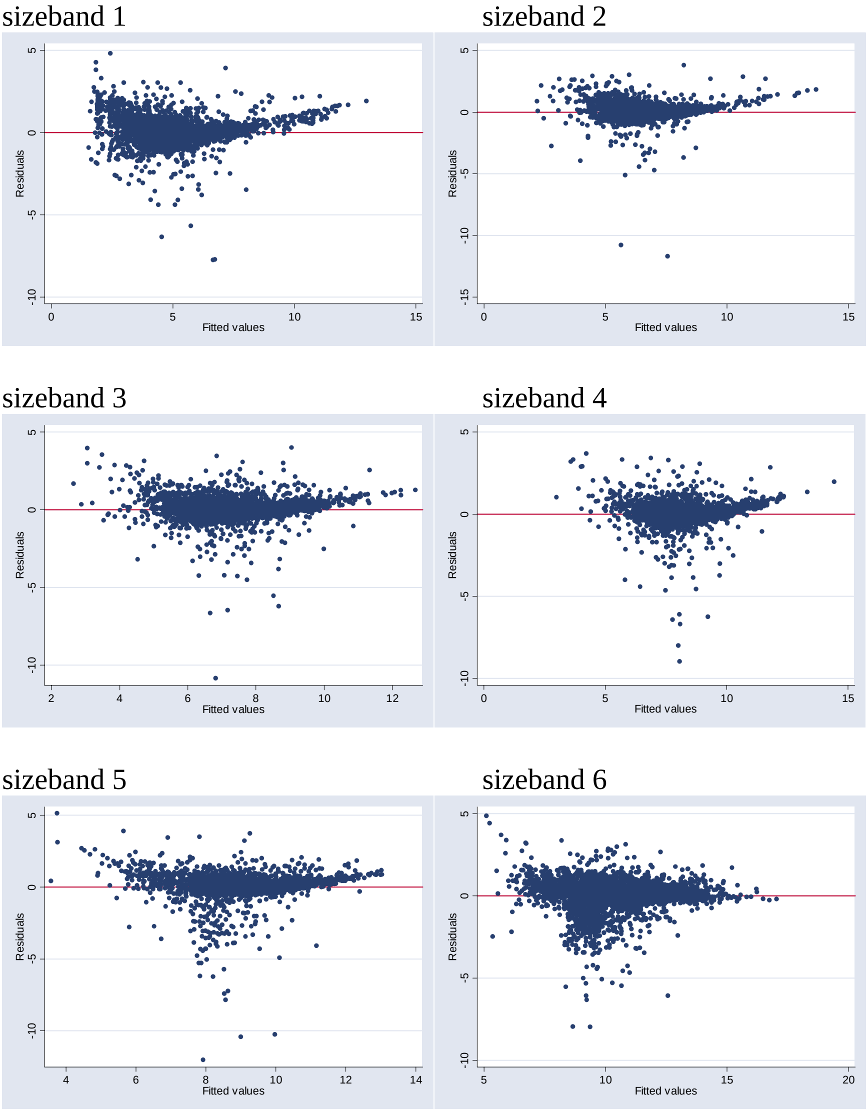


a. ARD

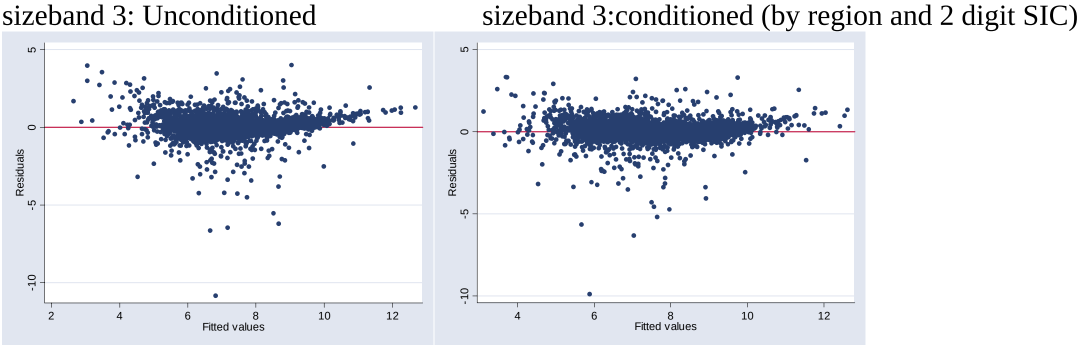
b. Overlap

Figure 4: Residual-versus-fitted plots

i. Unweighted regression



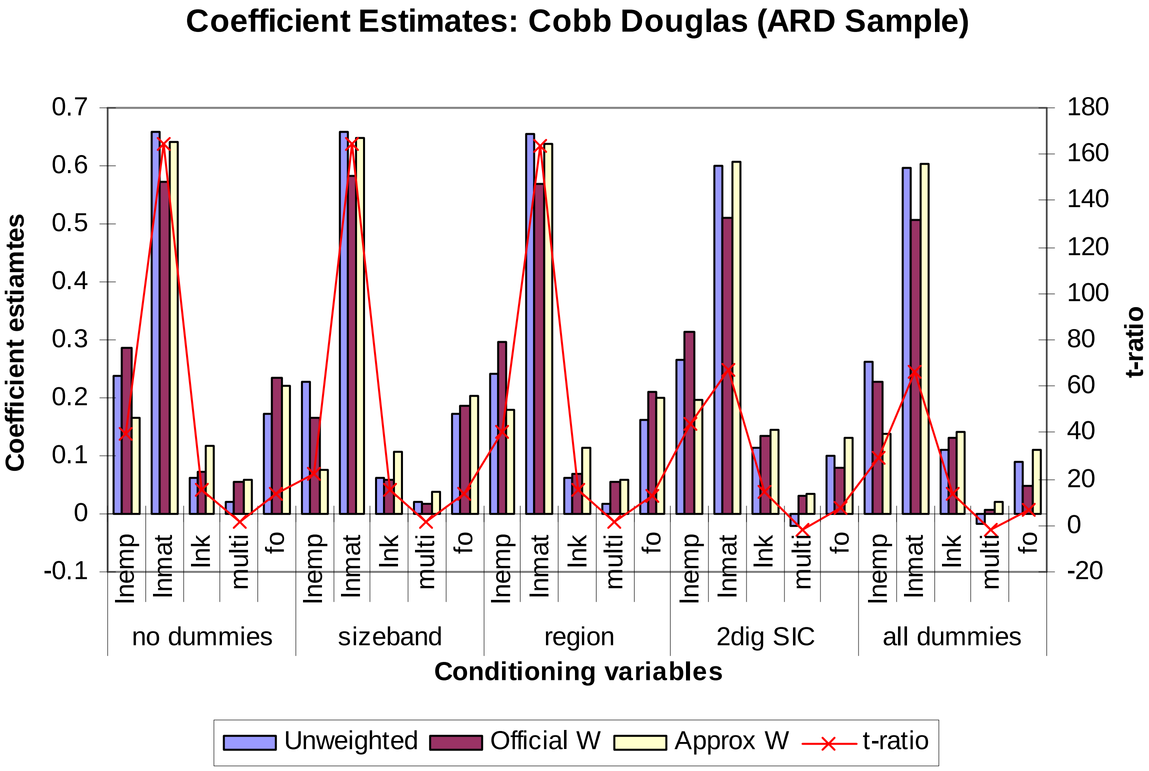
ii. Effects of conditioning



Note: RVF plots for each 3 macro regions and 8 sectors (catering, construction, motoring, production, property, retailing, services and wholesales) show no indication of heteroskedasticitiy.

Figure 5:

i. Actual coefficient estimates



ii. Normalised coefficient estimates

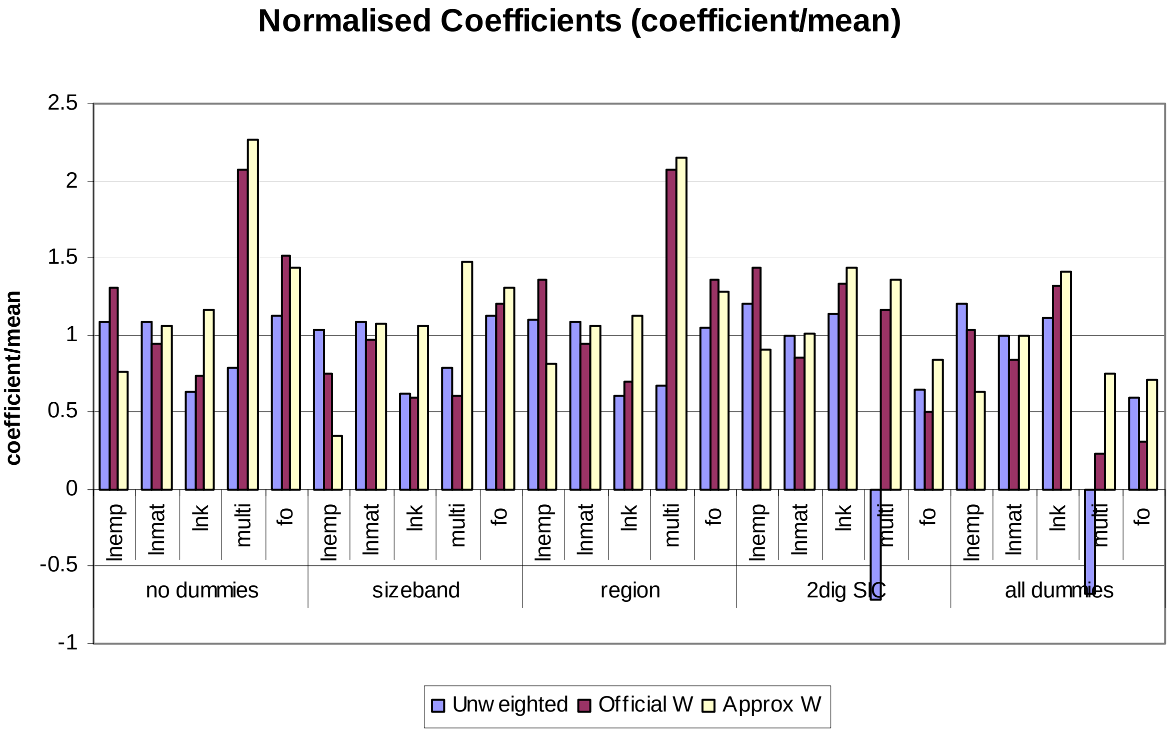


Figure 6:

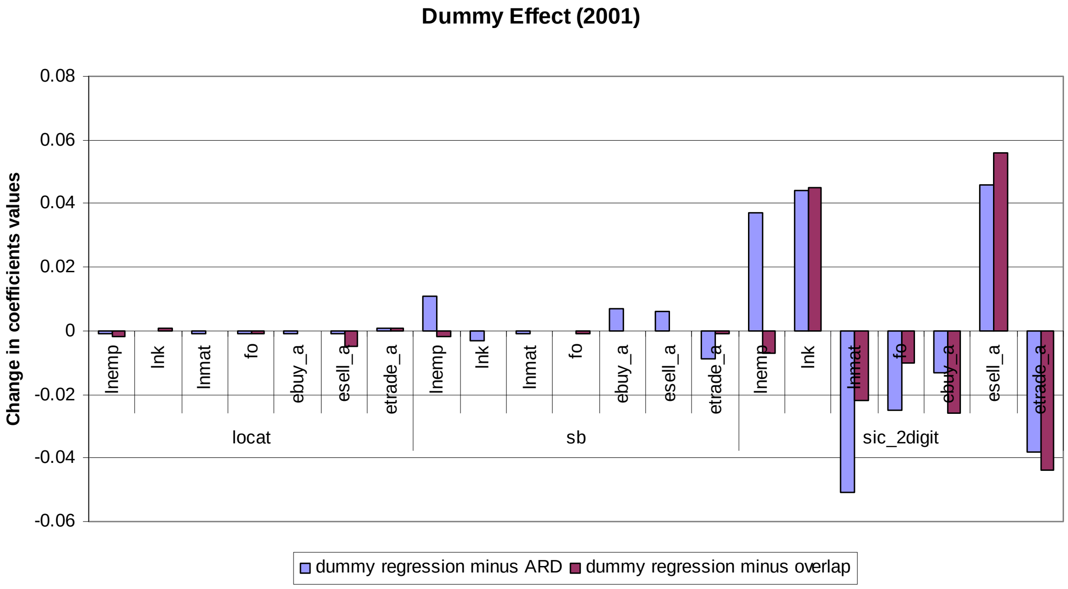
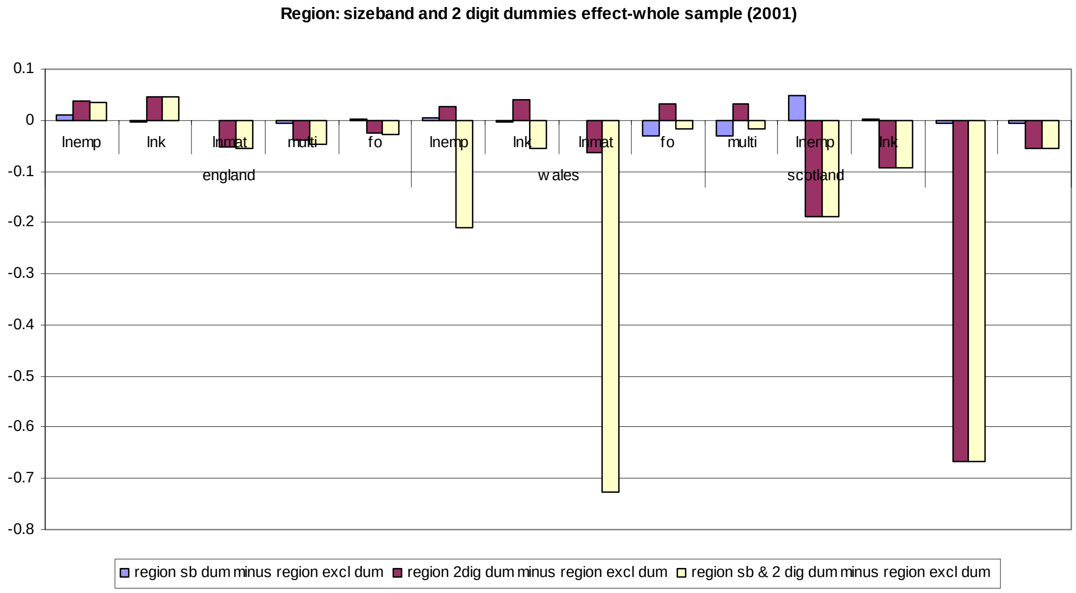

Dummy Effect (2001)
Change in coefficients values
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
lnemp
lnk
lnmat
fo
ebuy_a
esell_a
etrade_a
locat
sb
sic_2digit
dummy regression minus ARD
dummy regression minus overlap

Figure 7:

i. Whole sample



ii. overlap sample

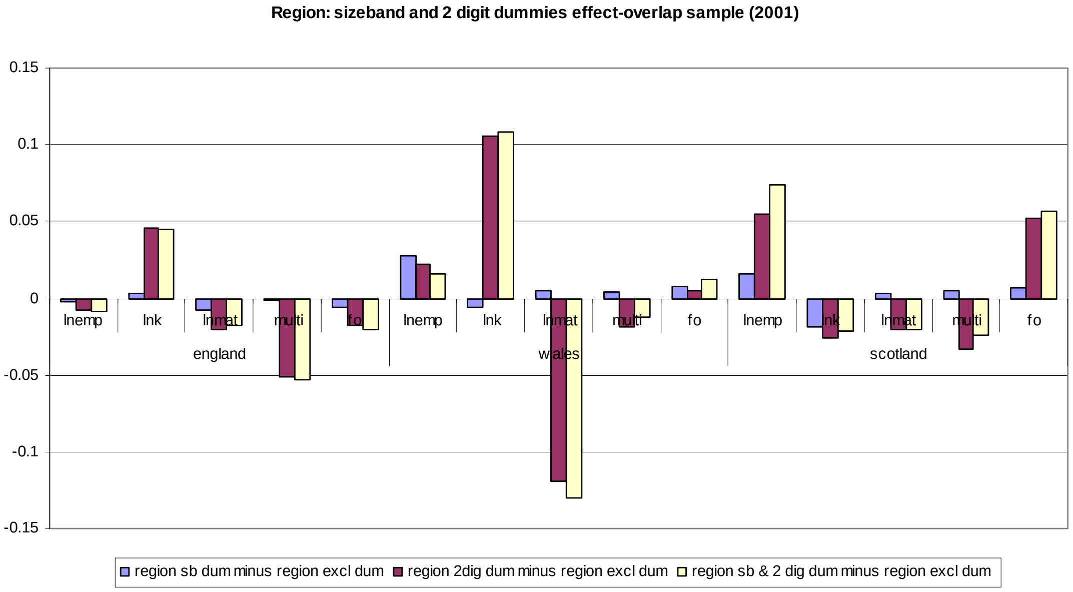
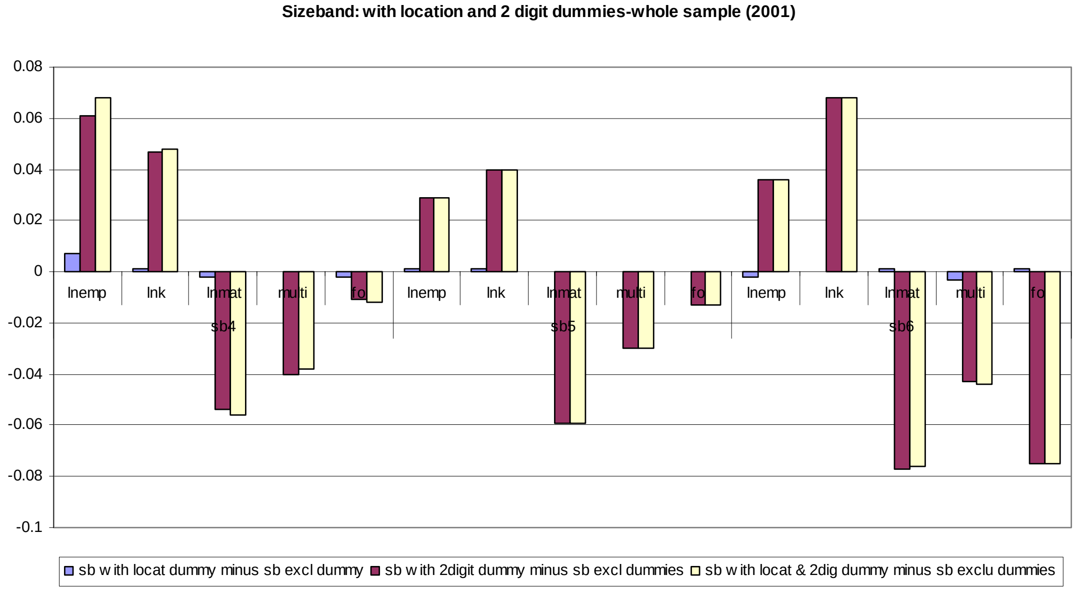


Figure 8:

i. Whole sample



ii.Overlap sample

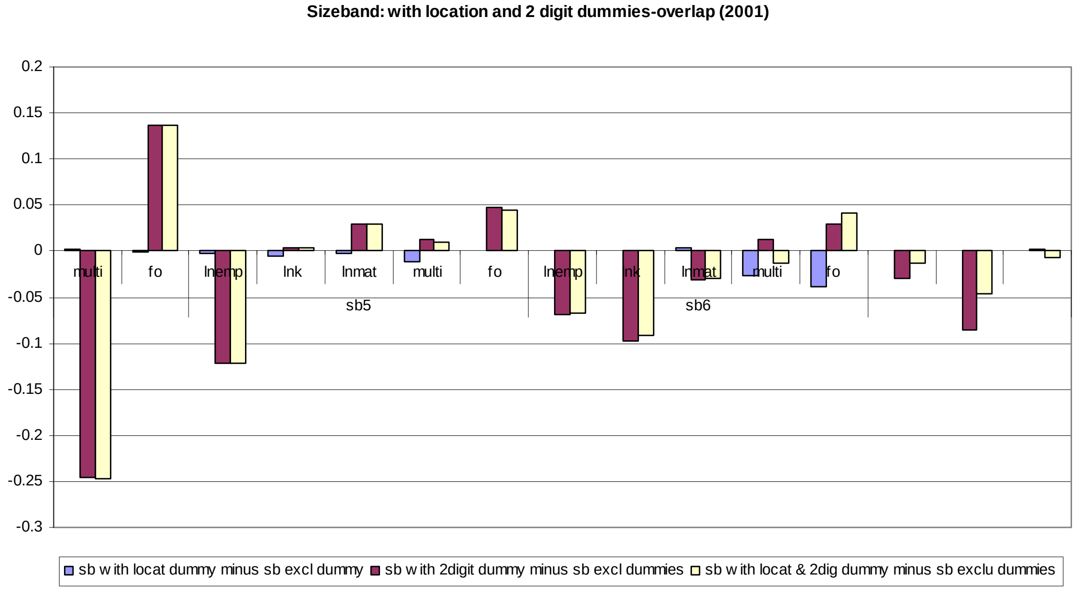
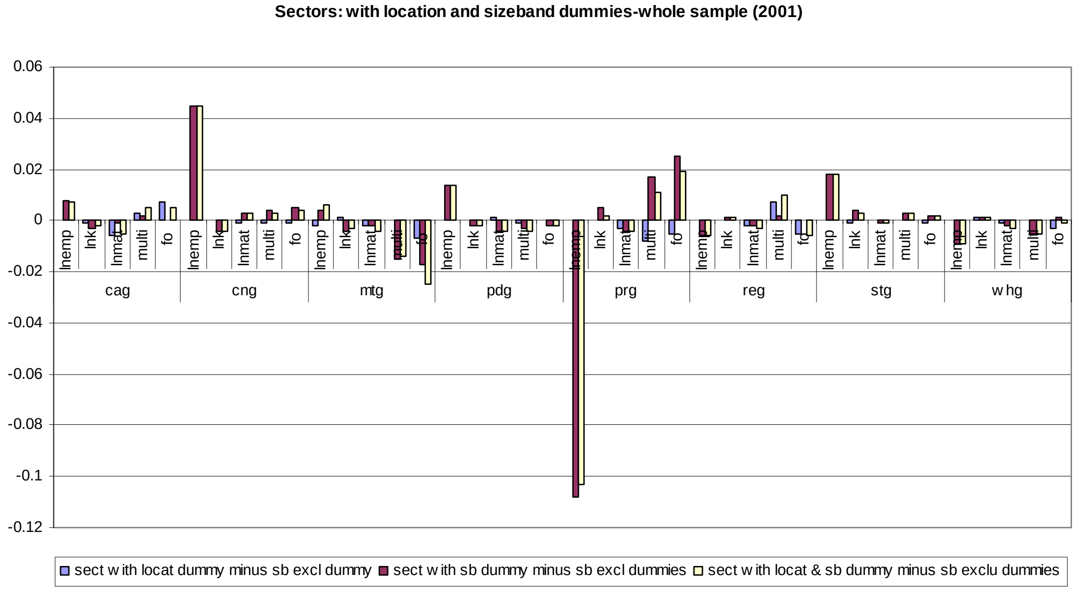


Figure 9:

i. Whole sample



ii. Overlap sample

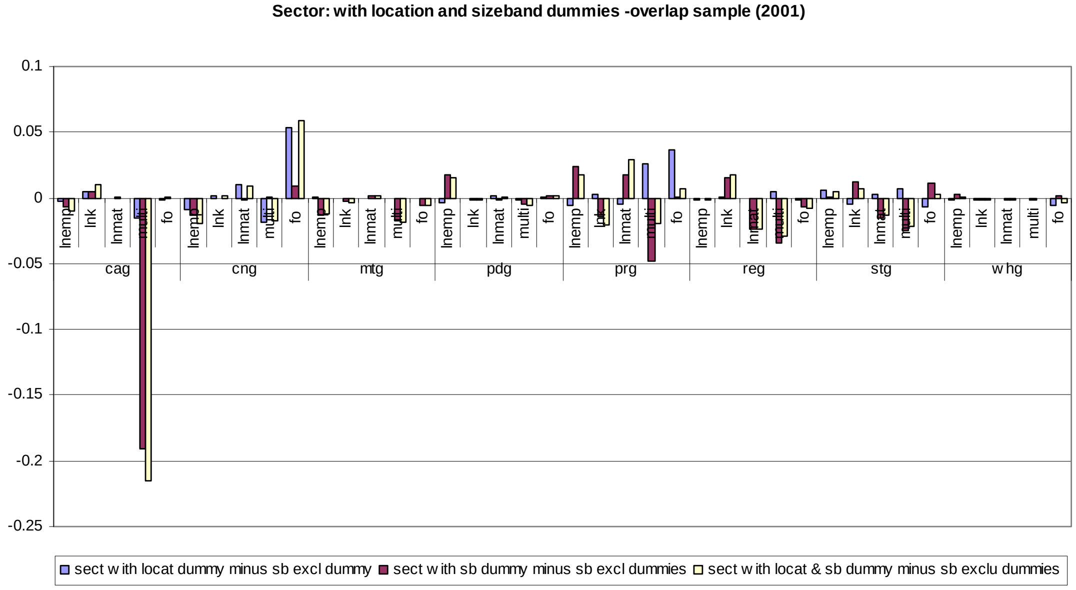


Figure 10: ARD, Overlap and 10 Random Samples of the ARD

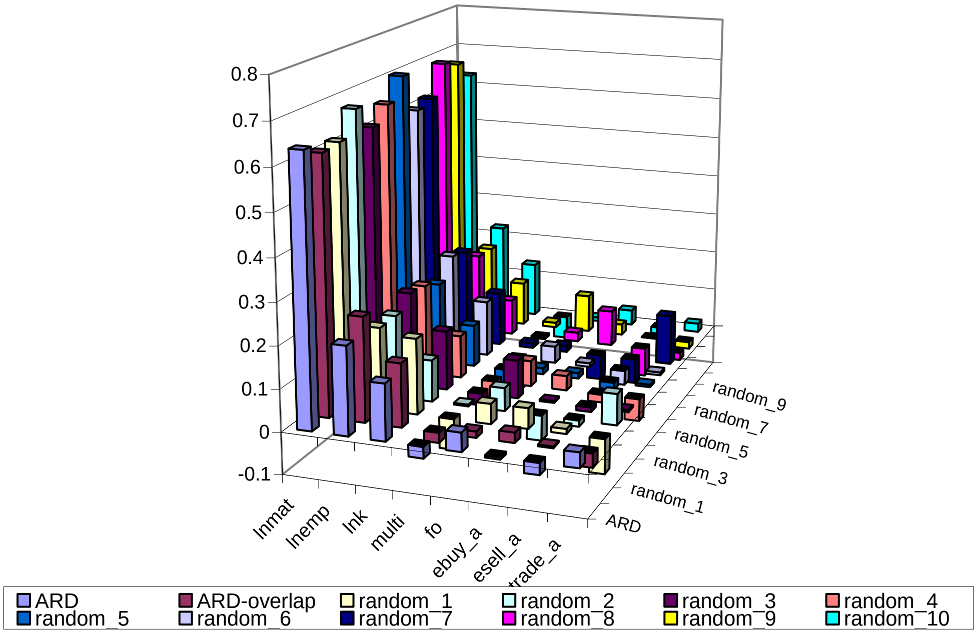
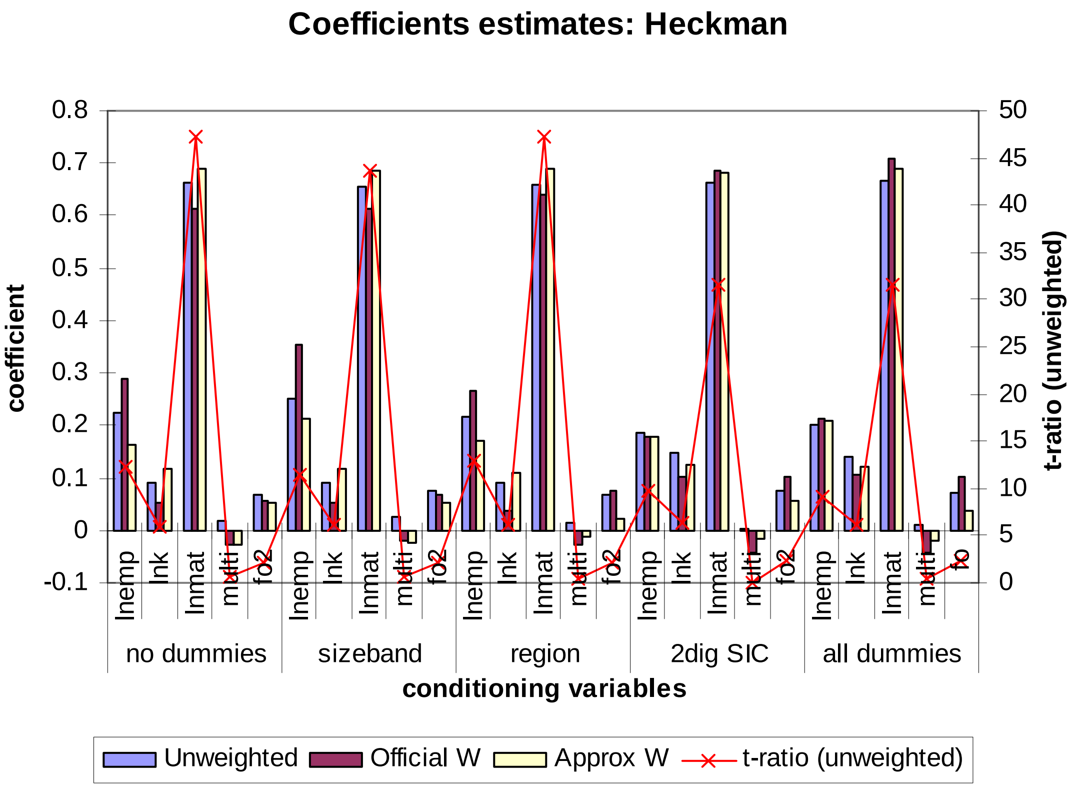
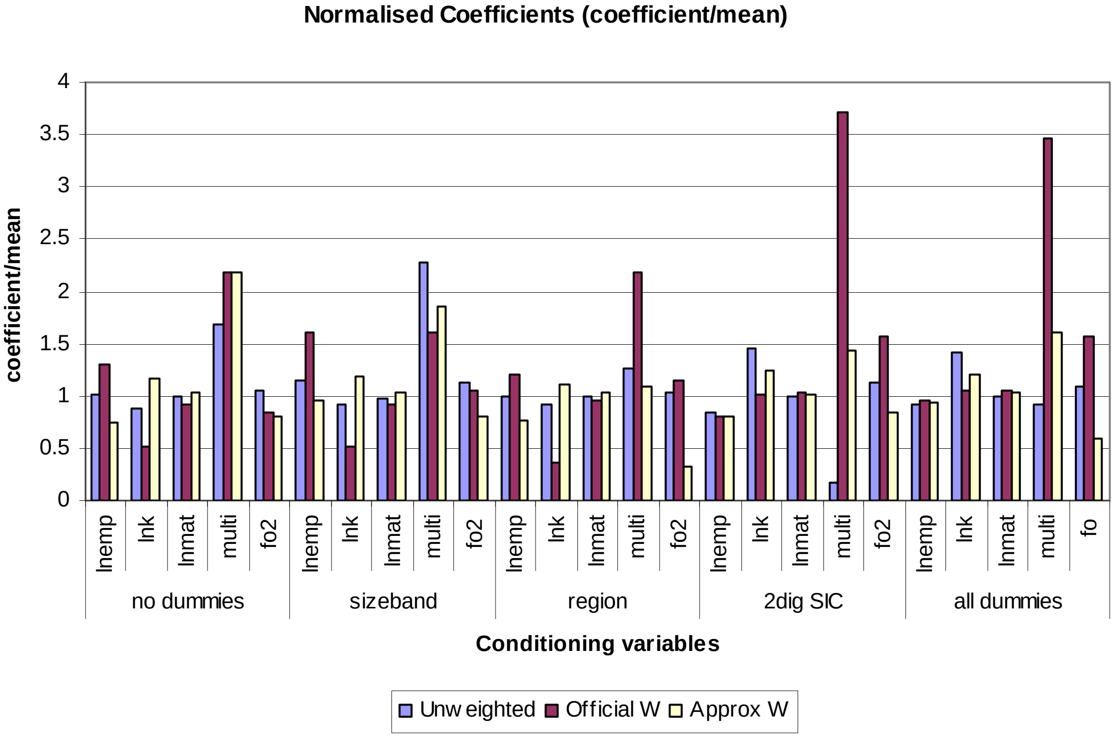


Figure 11: Coefficient Estimates

i. Actual coefficient estimates



ii. Normalise coefficient estimates



URN 06/737