Something for nothing?

Shirley Coleman[1]

Abstract

Official statistics are an important component of the mass of publicly available data. Although well used by policy makers, planners and academic researchers, engagement with businesses is not fully developed. Both corporate and citizen users can benefit more from using these free resources. Considerable efforts are being made to engage with users, for example via the recently launched www.statsusernet.org.uk in the UK and the BLUE-ETS FP7 funded project (www.blue-ets.istat.it) in Europe. However, more can be done. Inaccurate completion of business surveys causes unnecessary expense. Motivation to complete can be improved by linking data collection with data usage. Case studies are an important mechanism for this. The UK Royal Statistical Society's getstats campaign (www.getstats.org.uk) is 3 years into a 10 year programme to promote better understanding of statistics in media, politics, business and education. This will encourage greater use of statistics. Access is the first step, presentation follows and then the next step is modelling to extend the impact. Modelling can introduce statistical complications and require statistical expertise as well as enthusiasm. This paper explores the efforts being made to encourage business to engage with National Statistics Institutes and then considers statistical challenges and modelling issues when the data are used. Two examples are described and the paper concludes with some examples of how both corporate and citizen users can benefit from getting involved and can achieve something (of at least some importance) for nothing (except for the resources to identify, access and analyse the data).

Key words: NSI's, official statistics, business engagement, statistical modelling, big data

---

[1] Dr Shirley Coleman, Industrial Statistics Research Unit, Newcastle University, Shirley.coleman@ncl.ac.uk

1. Introduction

Official statistics are widely used in policy making, planning and academic research. Larger businesses engage more fully with official statistics than do small to medium enterprises (SMEs), however the resources are generally underused by businesses [5] and in educational courses and texts [1]. This paper looks at the efforts being made by National Statistics Institutes (NSIs) to engage business users. It then looks at the use businesses make of publicly available data and considers the statistical issues around the modelling that is carried out. Two specific examples are described in which the choice of model has far reaching consequences. The paper concludes with examples of how alertness can bring dividends to both the corporate and citizen user.

2. Engaging with business users

NSIs want to hear from business users what they need; this gives NSIs evidence to protect and enhance their services and allocate their currently diminishing resources. NSIs are under increasing financial pressure and are keen to show that their services are valued. However, a recent study conducted as part of the BLUE-ETS FP7 project (www.blue-ets.istat.it) showed a variable usage of government figures by business. In some cases, business users phoned for clarification in others there is extensive input from users, but these users are mainly from academic and government backgrounds. Only 1-2% of the NSI's annual revenue from commissioned products and services was found to come from businesses [5], and these businesses typically use NSI data for benchmarking and for learning about customers [4].

Some areas, for example Department for Energy and Climate Change in UK have a lot of users who willingly sign up to get up to date information. Other government departments find it harder to facilitate two way communications between providers and users. It is unlikely, however, that there is one best practice method. Considerable effort has been made in UK to engage business users, for example via the newly launched www.statsusernet.org.uk website and the Royal Statistical Society's getstats campaign (www.getstats.org.uk) which is 3 years into a 10 year programme to promote better understanding of statistics in the media, politics, business and education.

Van Grinsen et al [7] studied response burden and motivation to complete business surveys. They found that "burden" may be perceived rather than actual as respondents did not feel that completing business surveys took a long time even though some could not see their purpose. Inaccurate completion incurs costs up to 40% of total data collection costs. Providers are not always users and some feel that the data they provide go unheeded. Case studies are useful in demonstrating the importance of official statistics and the RSS has compiled several with particular reference to disclosure issues.

Data miners encourage companies to make use of all the operational data they collect. Context transforms data into information but gathering internal context requires effort, for example in communication between different parts of a company or conducting expensive customer surveys. Relevant external context can often be freely obtained from publicly available data, from sources such as energy suppliers, see for example www.nationalgrid.com/uk/Gas/Data/ , healthcare institutes, see for example

www.ic.nhs.uk/statistics-and-data-collections and NSI official statistics, see for example www.ons.gov.uk. If temporal and geographical precision is sacrificed in favour of opportunity then official statistics have a lot to offer, even though the data may still have to be supplemented by commissioned surveys.

So what is available for free that can be used by companies, and what more is needed?

- Consider a company wanting to launch a service for a particular demographic group, for example young people. How many young people are there in the target area? Curiously, even though there are copious data on population, catchment areas may not always correspond to requirements, the data may not be timely enough and the data may be unable to reflect dynamic changes.
- Consider a company who wants to develop a high tech e-business park in a deprived area. Can they be sure that there will be enough skilled people to work there? Company start up numbers and VAT registrations are available but do not give a complete picture. One approach is to look at the number of local graduates who move out of the area on graduating, as they may be tempted to stay if there is more employment. However, this information is in the gift of the education establishments and is not publicly available.
- A company may want to quantify quality of life changes after providing work for unemployed people. How can they do this? One way is to map their earnings over time, but this data is extremely difficult to access.

Data are available but need to be actively sought out; business needs to realise how external data can help; additionally providers need to be in a dialogue with users so that they know what business wants and can make their case for central government to continue to fund them to provide it.

3. Statistical modelling issues when businesses use data

Publicly available data can play an important role in decision making if companies can overcome the barrier of lack of awareness. Another barrier, however, is statistical competence. Accessing data is only the first step. The second step is presenting the data in a suitable way. In their observational sample of companies in Slovenia and the Netherlands, in [7] it was found that operational staff were happy with figures whereas managers were assumed to prefer visual presentations. The third step is to go on to interpret the data and build models from it. In the following two examples, expensive actions will be carried out on the basis of models built from publicly available data.

3.1 Example 1: Efficiency analysis

The aim of a company's efficiency analysis is to set price controls for the next 5 years. It is known that costs are related to drivers such as customer numbers and maintenance, and a large number of measurements are available as potential cost drivers. Reliable annual data are available for 8 regional offices of the company for the past 3 years. The company adopts the method of modelling costs on drivers, locating the lower quartile of

predicted cost and calling this the efficiency target line. They then congratulate regions operating below this line and penalise companies operating above this line on the grounds that if some regions can be efficient, then so can the rest if they are so persuaded. It is clearly important to establish a meaningful and fair model so that it is reasonable for regions to strive for the efficiency target.

The low number of data points means that only a limited number of sources of variability can be included in the model. A rule of thumb is for the number of predictors to be no more than a quarter of the number of cases and usually no more than 4 or 5. In this efficiency analysis, data for 8 regions for 3 years are combined into "panel data" and these 24 values are analysed together to give a common regression slope and year-specific intercepts on the assumption that the overall relationship is constant over 3 years but the level of costs may vary between years. It is assumed that the data from separate years are independent, however, the variation between years is much lower than the variation between regions. Modelling with 3 or 4 drivers and 3 different intercepts takes up 6 or 7 degrees of freedom and considering that the difference between years is so small, this is virtually all of the variation accounted for. The consequence is that models with too many predictors are over-fitting the data and the prediction errors are likely to be large unless future data are very similar to the data used for modelling.

Costs and drivers are modelled by multiple regression (MR) using raw or logged data which may or may not be standardised (to zero mean and unit standard deviation). Many of the potential drivers are correlated with each other, which is not surprising as they are all related to costs. There are many alternative equivalent models. As the efficiency line is based on data for 8 regions, the efficiency line always cuts off 2 regions who take the role as "Best Practice" and earmarks 6 regions that need to lower their costs. Different models will inevitably favour or discriminate different regions depending whether their costs are above or below the prediction line. Most alternative models lead to similar predictions, however, to avoid accusations of favouritism, the analysis needs to have objectivity in the choice of drivers and model.
Drivers in the model should have consistent definition and relevance for all regions, be reliable, precise and robust (for example, being measured by a responsible and disinterested third party). The problem with multicollinear drivers is that the coefficients (or slopes) of drivers in the model change markedly depending which drivers are included in the model and which are omitted.

The driver which has the greatest correlation with costs is preferentially included in the MR model and has a large coefficient, while the other drivers have coefficients reflecting their additional importance after the first driver is in the model. Therefore, whichever driver has marginally higher correlation with costs will have the higher coefficient and appear to be the most important driver. Provided the relationship between the drivers is steady, this is not necessarily a problem and the model may hold true over the 5 year period. However, if the relationship between the drivers changes, so that one driver starts to behave differently to another with which it was previously correlated, then the model will not be responsive to what could be an important change in costs. For this reason, it is preferable to model using drivers relating to more than one aspect of costs.

MR finds the closest fitting model for a certain number of drivers. An alternative to MR with multiple drivers, is to use business intelligence to combine selected, important drivers into a composite scale variable (CSV) in a way which is not dependent on the data (for example, weighting the drivers by overall company expenditure on each of the drivers), then any number of drivers can initially be

considered provided the final number of CSVs is limited to ensure the model does not over-fit the data. However, if CSVs are used for a regression model, the resulting model will not necessarily give the closest fit. If the driver with the highest correlation with costs also happens to be the driver with the highest CSV weight then the model will appear to fit very well, but if the driver with the highest correlation with costs also happens to be the driver with the lowest CSV weight then the model will appear to fit very poorly.

A suitable model can also be obtained using partial least squares (PLS) regression. PLS reduces the number of predictors to a set of uncorrelated CSVs and performs least squares regression on these CSVs. PLS is particularly useful when the predictors are highly multicollinear or when there are more predictors than observations and ordinary least squares regression either fails or produces coefficients with high standard errors. The downside is that the CSVs include all the drivers, although the coefficients of individual drivers in the CSVs differ reflecting the importance of individual drivers in explaining overall variation.

The MR model will give the closest fit to the costs in terms of reducing the residual sum of squares. However, if there are too many drivers in the model, the residual mean square can increase because its degrees of freedom decrease. The adjusted squared correlation coefficient between observed and predicted values, R^2_{adj}, can be used to compare the predictive power of models with the same dependent variable. It is, however, still possible to have a high R^2_{adj} value but a poorly fitting model. The final choice of model depends on the statistical criteria and diagnostic tests. Diagnostic tests are used to check, amongst other things, whether the model fits consistently well across the whole range of predictor and predicted values. If the fit is poor, then further analysis needs to be carried out and a new regression model needs to be found and tested.

In summary, when there is a wide choice of possible explanatory variables and a wide choice of competing models and the model will be used to benefit some but penalise others, then objective criteria must be used to justify the final choice of explanatory variables. In this example, the drivers should have consistent definition and relevance for all regions, be reliable, robust and precise and the models should have high R^2_{adj} values and good fit in diagnostic tests. If there are too many contending drivers then the best subset of 2 or 3 drivers should be found by comparing models with different combinations of drivers using R^2_{adj}, or the drivers should be combined into 2 or 3 independent CSVs using business intelligence or PLS regression.

Because the lower quartile is chosen as the efficient target, more regions are discriminated against than favoured. G.E Deming and other "quality gurus" do not recommend taking action on the basis of natural variability. Reacting to the natural variability left after fitting a model can produce more variability in the system. Quality practitioners also warn that setting targets has the effect of making people achieve the target possibly to the detriment of the performance of other aspects of the system, such as has been observed with target setting in the UK National Health Service. However, in the business world, decisions have to be taken and final choices have to be made.

3.2 Example 2: Controlling exposure to risk

A company has a fleet of vehicles. If a vehicle breaks down, the company may have to assist the people travelling in it. This is expensive and the company wants to ensure that the expected number of assists is contained. They have established risk scores for the probability of a breakdown for each of their vehicles and data are available on number of assists. They want to determine a cut-off risk score based on controlling the number

of assists; vehicles above the cut-off need to receive expensive special attention. The larger the expected number of assists per breakdown, the lower the cut-off, the greater the number of vehicles needing special attention and the greater the expense.

Official statistics for the number of people requiring assistance in vehicle breakdowns are given in Table 1.

Table 1: Frequency of different numbers of people requiring assistance

Number of assists	0	1	2	3	4
Frequency of breakdowns	73	9	2	0	2
Mean assists per breakdown	0.24	Standard deviation	0.72	Standard error	0.08

The cut-off risk score is inversely proportional to the upper confidence bound for the expected number of assists per breakdown. The distribution of risk scores is positively skewed; even so, a small decrease in the cut-off can lead to a large increase in the costs. How should the cut-off be estimated?

The data in Table 1 are not well modelled by the Poisson distribution[2] because the data are over-dispersed. The over-dispersion could be caused by the large number of zeros and one approach is to consider the data to arise from a mixture of distributions where zero-assist breakdowns occur with a certain probability and non-zero-assist breakdowns occur according to Poisson probabilities. This Zero Inflated Poisson model is not adopted here because it implies that there is something intrinsically different about breakdowns producing zero assists and this is not the case; many factors can affect the outcome of a breakdown, including the number of people present, their ages and health and the response of emergency services.

An alternative to the Poisson distribution for count data is the Negative Binomial model which allows for data that is more variable than expected in a Poisson model. This model is widely used in transport statistics and ecology where counts of various items are regularly analysed. The Negative Binomial model is a good fit to the data[3], however, using this model, the fit is less good for 4 assists and the expected number of breakdowns resulting in 5 or more assists is extremely low, whereas it is known that such breakdowns can occur. Therefore, although the Negative Binomial model fits the bulk of the observed data well, it is unwise to rely on it for the confidence interval.

The mean number of assists per breakdown can be modelled by the Normal distribution even though the individual values are counts. The Normal model fits better for mean values from larger sample sizes, the sample size being particularly important

[2] A goodness-of-fit test for Poisson has chi-square=5.36 with 1 degree of freedom (p = 0.021)

[3] A goodness-of-fit test for Negative Binomial has chi-square=0.05 with 1 degree of freedom (p = 0.823)

when the individual data are skewed (central limit theorem). The data in Table 1 are skewed as most breakdowns have zero assists. The occurrence of breakdowns with higher assists than those observed would make the observed data distribution even more skewed. However, the sample size is fairly large and it was found in simulation studies in [6] that the Normal distribution gave reliable confidence intervals for the mean with data of the form in Table 1 when the sample size was 86.

Confidence intervals for different models and confidence levels are shown in Table 2 for comparison. The Poisson and Negative Binomial confidence intervals are calculated as in [3] and [6] respectively. The Normal upper 2-sided 99% confidence bound is the mean plus 2.58 times the standard error.

Table 2: Confidence intervals for mean assist per breakdown

Model	***Poisson***	***Negative Binomial***	***Normal***	***Boot-strap***
Upper of 2-sided 99%	0.42	0.41	0.44	0.46

An alternative approach is to find an empirical boot-strap confidence interval [2]. Boot-strapping has the disadvantage of being entirely based on the observed data, which, for example in Table 1 has no breakdowns with 3 assists. It also assumes, like the statistical models, that all the observations are independent. This may not be true if breakdowns occur together due to external influences. A simulation is carried out in which 86 values are randomly chosen with the probability of selection equalling the observed probabilities (73/86, 9/86, 2/86, 0/86, 2/86). The mean value of the 86 simulated observations is calculated and stored and the process is repeated 1000 times. The 99.5% quantile of the distribution gives the upper bound of the empirical boot-strap two-sided 99% confidence interval. Generally, the difference between the boot-strap and Normal limits is larger for more skewed data and boot-strap confidence intervals tend to be over optimistic (thinner) and therefore under-estimate rather than over-estimate the upper confidence bounds. In contrast, in Table 2, the boot-strap gives a higher value.

In the event it was decided that it is preferable to use a standard model rather than results based on simulation and as the Normal model gives reliable confidence intervals, the recommendation was to adopt the upper Normal confidence bound of 0.44 and use it to give a suitable cut-off risk score.

4. Footnote

The future will be defined by enormous quantities of data being freely available. This should allow greater sensitivity for interested parties to pick up trends and other patterns, but this may also pose a distraction in that the focus is on grappling with the

zettabytes of data rather than taking a more focussed approach. A good example of the use of 'big data' is www.wefeelfine.org which is a website purporting to show how we all feel from a real time analysis of the world's tweets and texts. Many small businesses have erupted making use of official statistics and there are endless possibilities, for example www.houseprices.co.uk collates publicly available data on UK house sale prices and tags on adverts for services in conjunction with them. Some businesses make money from ad words tagged on to search engines and front-ends to databases. Others do some analysis to add more value, for example a cluster analysis of job adverts can lead to forward prediction of prospects and shortages in different market segments. Besides corporate users, citizen users can also benefit from publicly available data. For example, many people in the UK save in a pension fund which is converted to an annuity on retirement. The annual pension payment is fixed forever once the annuity is taken. The annuity rate is based on life expectancy and casual experimentation shows that rates change markedly with address and declared health issues, suggesting that the prospective pensioner should consider a brief move to an unhealthy location when the annuity is taken. Annuity rates are updated frequently although irregularly and in 2013 the rules change to enforce gender equality, having the effect of reducing male payments to the women's level because women live longer.

So in conclusion, more businesses can be encouraged to use official statistics and some case studies of success stories would help. Statistical issues are common when the data are used for modelling. In principle, companies can get something for nothing (except the resources to identify and access the data) but what they do with that something determines whether it is worthwhile or not.

References

1. Biffignandi, S., Oehler, M., Bolko, I., Bavdaž, M.: Use of NSI statistics in textbooks. In BLUE-ETS Conference on Business Burden and Motivation in NSI Surveys Statistics Netherlands, Heerlen, March 22 & 23, 2011.
2. Efron, B.: The jackknife, the bootstrap, and other resampling plans. 38. Society of Industrial and Applied Mathematics CBMS-NSF Monographs (1982). ISBN 0898711797.
3. http://www.graphpad.com/quickcalcs/Confinterval2.cfm
4. Löfgren, T., Gravem, D., Haraldsen, G.: A glimpse into the businesses' use of internal and external data sources in decision-making processes, In BLUE-ETS Conference on Business Burden and Motivation in NSI Surveys Statistics Netherlands, Heerlen, March 22 & 23, 2011.
5. Lorenc, B., Giesen, D., Persson, A., Bavdaž, M.: Understanding and Meeting Businesses' Needs for Official Statistics: an NSI perspective. In BLUE-ETS Conference on Business Burden and Motivation in NSI Surveys Statistics Netherlands, Heerlen, March 22 & 23, 2011.
6. Shilane, D., Evans, S.N., Hubbard, A.E.: Confidence Intervals for Negative Binomial Random Variables of High Dispersion. The International Journal of Biostatistics, 6(1), (2010). Available via http://stat-www.berkeley.edu/tech-reports/782.pdf
7. van Grinsen, V.T., Bolko, I., Bavdaž, M., Biffignandi, S.: Motivation in business surveys. In BLUE-ETS Conference on Business Burden and Motivation in NSI Surveys Statistics Netherlands, Heerlen, March 22 & 23, 2011.