1352.0.55.113

Australian Bureau of Statistics

Research Paper

Empirical Models for Survey Cost, Response Rate and Bias Using Paradata

www.abs.gov.au

Research Paper

Empirical Models for Survey Cost, Response Rate and Bias Using Paradata

Melanie Black, Geoffrey Brent, Philip Bell,

Rosslyn Starick and Mark Zhang

Statistical Services Branch

Methodology Advisory Committee

19 November 2010, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 02 MAR 2012

ABS Catalogue no. 1352.0.55.113

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Paul Schubert, Statistical Services Branch on Canberra (02) 6252 6591 or email <statistical.services@abs.gov.au>.

EMPIRICAL MODELS FOR SURVEY COST, RESPONSE RATE AND BIAS USING PARADATA

Melanie Black, Geoffrey Brent, Philip Bell, Rosslyn Starick, Mark Zhang
Statistical Services Branch

QUESTIONS FOR THE COMMITTEE

1. Have we used the right techniques for our response models, cost models and bias models? Does the Committee have any recommendations for more appropriate techniques to apply?
2. Are the response models for household surveys appropriate regarding the call attempt numbers? Are the suggested extensions to the response models appropriate? Can the Committee provide any further suggestions?
3. What diagnostic tests should we have considered to assess the performance of our models? Are the chosen tests appropriate?
4. Are the simplifications made in the household cost model justifiable?
5. Is calibration against known totals an appropriate way to reduce the impact of data quality issues and a biased fitting technique for the household cost model? Or do we risk degrading accuracy by distorting model slopes?
6. Would the bias model be improved by the availability of new paradata (e.g. to categorise types of non-respondent/difficult respondent)? If so, what sort of paradata should we aim to collect?
7. Can the Committee suggest an alternative approach or formulation for the economic survey bias model to resolve the methodological issues observed?

CONTENTS

EMPIRICAL MODELS FOR SURVEY COST, RESPONSE RATE AND BIAS USING PARADATA

Melanie Black, Geoffrey Brent, Philip Bell, Rosslyn Starick, Mark Zhang
Statistical Services Branch

ABSTRACT

The pressures of declining survey response rates and increasing operational costs pose a threat to the quality of the survey estimates. Various sources of paradata collected by interviewers, demographic information, survey data from previous collections, and administrative data give survey methodologists and practitioners unprecedented opportunity to understand respondent behaviour, and to analyse the effectiveness of survey operations, the associated cost structures, and the implications for survey estimate accuracy.

This paper presents modelling work for some ABS household and business surveys on response rates, survey cost structures, and potential bias resulting from changes to data collection inputs and efforts. Using these models, data collection operations can be re-configured by taking advantage of more cost effective follow-up methods and prioritising targets to improve response rates and survey estimate quality.

1. INTRODUCTION

In recent years there has been a worldwide problem in maintaining response rates without significantly increasing costs. Throughout the world, there is evidence to suggest that there has been a general decrease in response rates over the last few decades, and where the response rates have been maintained this has been through significant additional cost and effort. In response to this, one of ABS' key needs is a capacity to predict respondent behaviour and other outcomes resulting from various allocations of effort. In order to address these questions, it is necessary to develop a response, cost, effort and survey estimate accuracy framework to understand the relationships between these different aspects of survey operations.

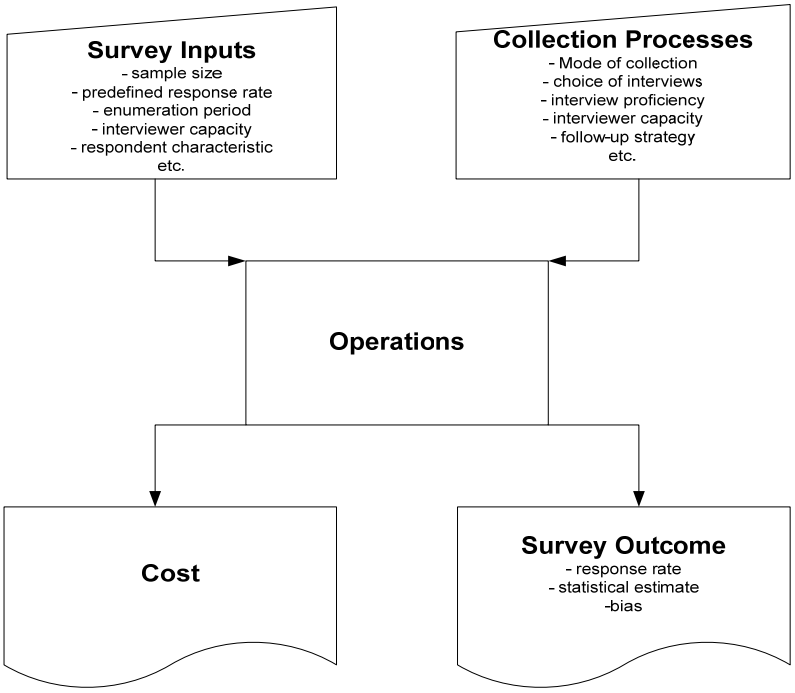
One way to achieve this would be an experimental approach, deliberately modifying operational procedures etc. to determine their effects on survey outcomes. However, there are substantial practical obstacles to experimentation. Experimentation requires consultation with a large number of stakeholders, and a significant lead time; an experiment that goes wrong has the potential for large cost overruns and/or damage to key ABS outputs. For these reasons, we start with an observational approach, attempting to develop as much theoretical information about cost/response/quality relationships as possible before verifying these models through experiment. Our aim has been to establish an interactive tool to understand the impact of different survey operation scenarios on response rates, operational costs, and survey estimate accuracy (in term of bias). Ultimately, we envision that this tool will enable the optimisation of operational procedures to minimise costs, while improving (or maintaining) response rates and survey estimate quality.

Figure 1.1 gives a systemic view of an ABS survey data collection operation. The recent availability of paradata describing ABS survey data collection processes has allowed us to establish a modelling framework to test and quantify the survey inputs and process contributions and interactions to response rates, cost and survey estimate quality.

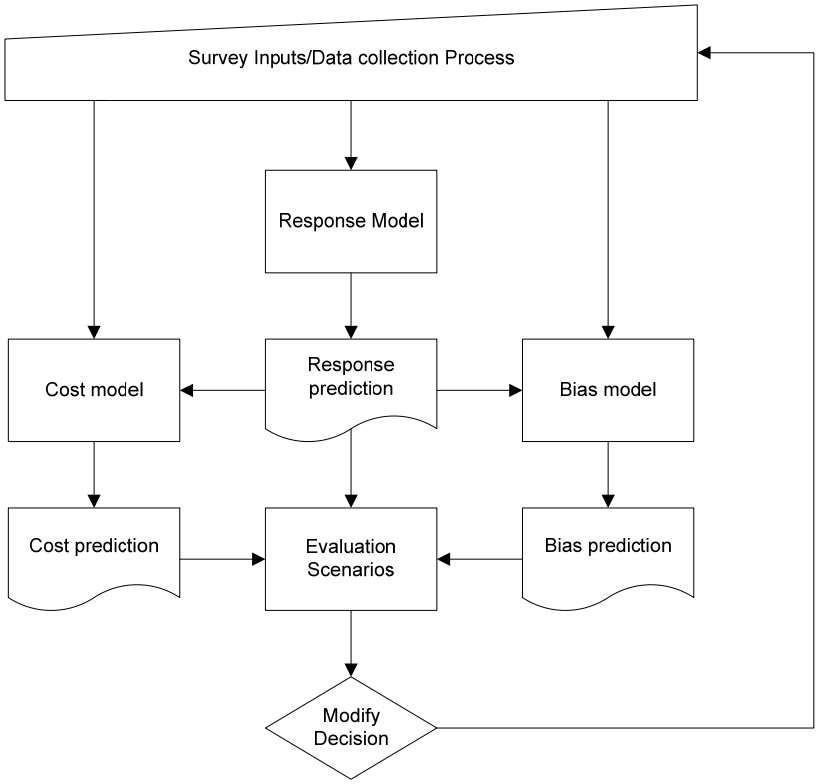
With knowledge of how survey inputs and processes affect response rate, costs, and survey estimate quality we can reconfigure survey inputs and processes to achieve better survey estimate quality with a reduced cost, or achieve a more favourable trade-off between them. Figure 1.2 illustrates how the three models can be used together to produce response rate, cost and bias predictions for a different configuration of survey inputs and data collection processes.

In a broader context, the knowledge can also assist us in designing future survey data collection models in a more integrated fashion to achieve workload allocation efficiency and coordination from an integrated workforce.

1.1 Survey data operation



1.2 The linked models scheme



The ABS household survey data collection models

Note: This paper discusses the ABS household survey program as of January 2011. This program is currently under redesign and the sampling method used for future surveys may have some differences from that described here, although the modelling approach is still expected to be relevant.

ABS household surveys have been based around a single master sample of geographic areas. The areas in the master sample are divided into finer blocks, with one block providing sample for the Monthly Population Survey (MPS), and a 'parallel' block used for a diverse range of large-scale social surveys, collectively referred to as Special Social Surveys (SSSs) that are used to collect detailed contents on particular social topics.

The Monthly Population Survey (MPS) includes approximately 30,000 households each month, with content including the regular Labour Force Survey (LFS) and short supplementary surveys that vary by month and by state. MPS sample is clustered by blocks. Households remain in the sample for eight months and are re-interviewed each month, with one-eighth of the sample being retired and replaced each month. At present, interviewers contact MPS households by telephone interview ('TI') and/or face-to-face approach ('FF'). Approximately 95% of MPS households are contacted face-to-face for their first month in sample. In later months, approximately 85% are contacted only by telephone, but some require a face-to-face approach, either at householder's request or because telephone contact is not possible (e.g. dwellings believed to be vacant). As a result, non-first-month household work is less clustered than first-month.

A typical MPS month involves approximately:

- 9000 visits and 2500 phone calls to 4000 first-month households, resulting in 3000 interviews;
- 8000 visits and 65,000 phone calls to 27,000 non-first-month households, resulting in 22,000 interviews;
- 4000 visits to first-month blocks (2.1 household visits per block-visit); and
- 6000 visits to non-first-month blocks (1.3 household visits per block-visit).

A typical MPS interview averages approximately 12–20 minutes, depending on supplement content. MPS response rate is relatively high (typically 97% for the Labour Force component and 85–95% for supplementary content). Most of the non-interviewed households are those not in scope (e.g. vacant dwellings).

Special Social Surveys are used for material not suitable for a MPS supplement, e.g. due to length or because the survey requires different sampling techniques. A typical SSS could run for several months, but unlike MPS each household would usually be

interviewed only once (exceptions exist). Households are generally contacted face-to-face, similar to the first month of MPS work.

As with MPS, SSS selections are usually clustered by blocks. SSS interviews are longer than MPS (usually 30–90 minutes) and response rates are commonly lower, varying greatly depending on survey content.

The ABS business survey data collection model

Most ABS business survey data collections use a standard model, and these are the collections considered in this paper. The survey paper form is mailed out to businesses in the sample to fill out and send back. Around the time of the due date, intensive follow-up (IFU) activity commences for those business have not yet responded. This includes both reminder letter and telephone contact. Typically, all businesses which have not yet responded will receive reminder letters, however telephone contact is generally prioritised based on the previous response history of the provider. This information includes the magnitude of previous reported data, whether the provider is new to the collection, and whether, when previously selected, the provider required follow-up contact or not. The purpose of the call is usually to remind the business to respond, rather than to collect data over the phone (although this does sometimes happen, depending on the survey and other factors). Telephone contact can also occur for other purposes, for example, responding to queries or following up incorrect contact information. Increasingly, modes of follow-up data collection are varied due to recent technological developments, although mail and telephone remain the predominant methods of contacting providers.

Survey data collection practices and cost structures can be different from one survey to another depending on the nature of the survey and survey procedure. In general, ABS household and economic survey data collection methods and cost structures are very different between the two categories and share reasonable common practices within a category. This paper presents three sets of empirical models used for modelling response, cost and survey estimate bias for the two categories respectively. Using ABS household and business survey data, we make assessment to the models and evaluate if they are fit for purpose, and present some the methodological challenges we are facing.

Data available for paradata analysis and modelling

Data available for modelling include:

- Call records. These describe each attempt by an interviewer to contact a household or business. They include identifiers for the interviewer and household/business, date/time information, outcome of the call, and various other information (e.g.

whether the call is telephone or face-to-face). Collating these by household/ business identifier produces a history of our contact attempts.

- Time and travel records ('T&T'), extracted from interviewer payment systems. Household interviewers record their work through various activity codes (e.g. 'travelling to/from workload', 'travelling and interviewing within workload', 'end work') with times and odometer readings.
- High-level payment records, showing the total paid to interviewers for a given survey in various categories (e.g. 'standard interview travel time').
- High-level staff allocation information (e.g. number of staff working on each survey on a given day).
- Form status changes (for economic surveys) detailing the response status of a business at any particular time (including whether they have been sent reminder letters).
- Interviewer employment records, including date of birth and date of engagement.
- Survey responses (reported data). These can be used to provide information about characteristics of respondents (e.g. employment status); by relating these to response patterns we can gain information about how different response scenarios might affect survey estimates.

Call and T&T records have known quality limitations: records may be missing, or lacking some data, or incorrect. High-level records offer less detail but give more accurate information on total costs; as described in subsequent sections, we use these sources in combination.This paper is constructed as follows. In Section 2, we present a response modelling framework for ABS household and business survey practices, and our assessment to the model performance, data gaps and methodological issues. In Section 3, we illustrate our attempt to produce two general survey cost models for ABS population and economic survey. Several real household and business survey data are used to evaluate the model performance, and how the survey costs are likely to vary in response to changes of survey procedure parameters, such as sample size, proportion of face-to-face and telephone interview, and different intensive follow-up procedures etc. In Section 4, we present a framework using survey data collection paradata to predict non-response bias for assessing the potential survey output accuracy induced by changing survey procedure parameters, its application to both publication and economic surveys, and methodological challenges. Section 5 summarises our findings and presents our remarks on our future research directions.

2. RESPONSE MODELS

2.1 Response model framework

There are two distinct phases in the response process:

1. contact with the business or household is established; and

2. the respondent provides data, either via an interview or self-enumeration.

Campanelli *et al.* (1997) present a theoretical model of survey response for household surveys which is an extended version of the work suggested by Groves *et al.* (1992). The extended theoretical model includes additional factors affecting survey response, namely geographic area, household factors and previous respondent-interviewer interactions.

Furthermore, Beerten (1999) uses two different response measures: the contact rate and the co-operation rate and states that the difference is important because the factors which affect the likelihood of making contact are not necessarily the same as those factors which affect the likelihood of obtaining co-operation once the respondent has been contacted.

Within ABS' current paradata systems, it is difficult to distinguish between non-contact and non-cooperation, so this paper concentrates on overall and call-level 'response rate'. Overall response rate is defined as the number of fully-responding units (households/ businesses), divided by the number of selected in-scope units (i.e. total number selected, minus units that are out of scope e.g. vacant dwellings), with non-response representing a combination of non-contact and non-cooperation.

2.2 Household survey application

Within the ABS, household surveys involve a complex combination of field interviewing and telephone interviewing. The surveys range from a monthly population survey which is run once each month over a two week period to special surveys which can run over several months.

Table 2.1 outlines some of the factors influencing survey and call-level response for household surveys.

The objective of our research was to look at the effect of area characteristics, household characteristics, interviewer characteristics and survey design features on response rates, by modelling the probability of response using paradata from various ABS household surveys.

2.1 Factors influencing survey response – Household surveys

Area characteristics
- state / territory
- metropolitan / ex-metropolitan
- population density
- % of population in high-rise dwellings
- % of working couple households
- % of working lone parent households
- % of group households
- LOD

Household characteristics
- whether the household has changed address
- whether household is first in sample
- dwelling type
- whether owner-occupied household
- number of adults
- number of children
- SEIFA

Respondent characteristics
- age
- sex
- marital status
- country of birth
- education
- labour force status

Interviewer characteristics
- age
- sex
- interviewer experience
- interviewer performance (e.g. previous response rate)
- interviewer availability
- interviewer continuity
- interviewer pay arrangements (data available from Oracle or Pascoe)

Interviewer observations
- whether dwelling is house or flat
- presence of security measures (e.g. locked gates, burglar alarm)
- household composition
- quality of housing
- observations of surrounding neighbourhood

Survey design features
- size of workload
- number of calls made by interviewer
- timing of calls made by interviewer
- whether telephone interview or not
- interview length
- outcome of previous call attempt
- advance letter

A logistic regression model was constructed to predict the probability of response using paradata from the ABS Monthly Population Survey (MPS). Response models were also constructed for two Special Social Surveys, the Survey of Education and Training (SET) and the Survey of Disability, Ageing and Carers (SDAC). The paradata used in our analysis was primarily field-process and interviewer calling data, as well as

some information on interviewer characteristics (namely age, experience and performance).[1]

A household was considered to have responded only if it was a fully responding household (not partial response). The response indicator, denoted by y_{ic}, is defined as

$$y_{ic} = \begin{cases} 1 & \text{if household } i \text{ is fully responding at call attempt } c \\ 0 & \text{otherwise} \end{cases}$$

For MPS, probabilities of response were modelled separately for face to face households, telephone interview households and combined face to face and telephone interview households, as well as for initial and follow-up workloads. For SET and SDAC, probabilities of response were modelled at the Australia level with no finer splits.

Backward elimination set to a significance limit of $p < 0.05$ was employed as the variable selection method. Two-way interactions were explored, but were not kept in the final models. Some of the explanatory variables were subject to missing data and these observations were removed from our analysis. Table 2.2 below shows the model fitting diagnostics.

2.2 Model fit and predictive power diagnostics

Survey	*Workload type*	*Household type*	*Max-rescaled R-squared*	*HL p-value*	*c-value*
MPS	Initial	Face to Face	0.0874	<.0001	0.660
	Initial	Telephone	0.0567	0.0008	0.619
	Initial	Both	0.1273	0.0003	0.737
	Follow-up	Face to Face	0.1162	0.9007	0.701
	Follow-up	Telephone	0.0885	0.0500	0.663
	Follow-up	Both	0.0578	0.7774	0.649
SET			0.0696	0.6160	0.655
SDAC			0.0430	0.1340	0.615

Max-rescaled R-squared[2] is used to provide an indicative measure for a relative predictive power although this measure does not have the usual meaning of percent of variance explained in the logistic model context. It appears all the models have very limited predictive power in this measure.

1 We did not link our paradata to Census data or survey response data to obtain area, household or respondent characteristics.

2 Max-rescaled R-squared is defined as rescaled R-squared $R^2 = 1 - \exp\left\{\frac{2}{n}\left[\log L(M) - \log L(0)\right]\right\}$ where $\log L(M)$ and $\log L(0)$ are the maximised log likelihood for the fitted model and the 'null' model containing only an intercept term, and n is the sample size.

The Hosmer–Lemeshow (HL) goodness of fit test was used as an indicator of the fit of the model. A large p-value (>0.05) suggests that the model captures significant variations in response rate. For SET and SDAC, the Hosmer–Lemeshow goodness of fit test indicated that these models were a good fit. The results for MPS varied, with the follow-up workload models indicated a good fit, while the initial workload models were not a good fit.

Literature found on the predictive ability of logistic regression models was in some cases contradictory. Allison (1999) suggested using the 'max-rescaled R-squared' as a good measure of predictive power, however other literature suggested that R-squared was not a good measure in logistic regression. Ragavan (2008) suggested the use of the c value, equivalent to the well-known measure ROC (receiver operating characteristic). The c value can range from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a one corresponds to the model perfectly predicting the response. Our response models had c values ranging from 0.615 to 0.737 which are low and suggests that our models do not perform that well.

Estimated coefficients and odds ratios calculated from the final response models[3] are presented in table A.1 in the Appendix.

2.3 Economic survey application

Predicting response behaviour for economic collections presents some challenges but also opportunities to better understand the way in which businesses react to follow-up procedures. Typically, a form is mailed out to businesses, then (around the time of the due date), follow-up activity including both telephone and written reminders commences. Telephone follow-up is prioritised based on a number of factors including business size and the response history of the business. This means that follow-up is not randomised and that effects might be confounded, for example, large businesses are more likely to respond and are also more likely to receive a large number of calls.

The key characteristics which are available and which we are interested in understanding the impact to response of include aspects of the follow-up strategy (number and timing of calls and reminder letters), demographic information such as size and state, information on the business' response history, and (subjective) contribution towards estimates.

[3] For MPS, the probability of response was modelled on four months of MPS paradata and tested on two different months. For SET and SDAC, the probability of response was modelled on 80% of the paradata and tested on the remaining 20%.

While there are likely to be other important factors (such as the characteristics of the individual within the organisation who is providing the data), this information is unfortunately unavailable.

In order to deal with the time dependence, a survival analysis approach was taken. A survival analysis approach for modelling response behaviour for mail surveys with telephone follow-up was previously undertaken by Campbell *et al.* (2004) for the Consumer Assessment of Health Plans Study Medicare Fee for Service Survey. In their study, number of call backs was used as the survival time and businesses for whom follow-up ceased without response were considered to be censored. The key finding was that response propensity declined slightly, but not significantly, based on the number of call backs, and therefore the decision on when to cease follow-up would be primarily based on cost. Similarly, Biffignandi and Pratesi (2002) applied a survival analysis model to examine the impact and timing of email reminders for a web survey of Italian businesses.

In our case, we are interested in the impact of number of calls and reminders on the probability of and time to response. A business is considered to have responded if it is no longer required to be followed-up, that is, either it has responded or it has been cancelled, as per the operational definition of form receival. Businesses which do not respond are censored at the end of the cycle.

The number of calls and number of reminder letters are both time-dependent covariates and change value during the cycle. The remaining covariates are fixed within the data collection cycle, though they may change between cycles.

We use paradata from four quarters each of the Quarterly Business Indicator Survey (QBIS) and Survey of Average Weekly Earnings (AWE), and two years of the Annual Integrated Collection (AIC Core). The data available includes records of each change to a business's form status and records of all contact made with the business including reminder letters and outbound calls. The data was pooled across multiple surveys to take advantage of information which is available in one collection but not another, as some situations never arise in some collections.[4] In fact, fitting separate models by collection tends to result in convergence problems which can be avoided by pooling the data (with, potentially, some loss of relevance or accuracy). We consider the inclusion of a survey or survey frequency effect in the model to check for evidence of survey-specific behaviour.

[4] For example, some surveys use more reminder letters than others.

As both the call and response mechanism are continuous time processes, we fit complementary log-log models with one record per week, following an approach taken in Allison (1995, pp. 216–222).[5] This makes the problem more computationally tractable[6] and also more effectively deals with ties in the data (when many events occur simultaneously).

Therefore, we are modelling the probability of response at time t given that the business had not responded at $t-1$. For example, a business which never responds will have one record for each week, all with response status equal to zero, while a business who responds in the third week will have two records with response status equal to zero and one with response status equal to one.

This model is related to the proportional hazards model, in that it is modelling the same underlying risk parameters and coefficients can be interpreted in the same way, however it has the advantage of being much faster and can produce more detailed information about the effect of time-dependent covariates. This means that we can interpret the parameter estimates as per the proportional hazards model.

The variables included in the model are:

- number of calls received to prior to the period (log transformed);
- number of calls received during the week (log transformed)
- number of reminder letters received to date (categorical);
- weeks since last written contact – despatch or reminder letter (categorical) ;
- main state of operations (categorical, NSW used as reference state);
- business size (categorical, largest size used as reference);
- 'gold provider' business status (gold/non-gold – indicating whether the business is known to have a good response history or not);
- contribution to estimate (as identified by subject matter areas for the purpose of follow-up, divided into three categories: crucial, newly selected, other);
- survey effect.

5 It should be noted that the dependence among the observations is acceptable providing each individual experiences the event (response) only once (Allison, 1995, pp 223-224). Once a business experiences the event, it does not have a record for any subsequent period.

6 Reducing the data to weekly rather than daily reduces the size of the data set from 3.5 million records to 700,000.

We considered various ways of parameterising the call number variables. After examining the data, the call number variables are log-transformed as this captures the observed diminishing return effect from each subsequent call (i.e., each subsequent call is generally less effective). Treating these variables as categorical was considered but added little in terms of explanatory power to the model.

Two-way interactions were also included in the model to account for the relationships between the variables. A forward selection process was used for variable selection.[7] We also considered nesting the strategy effects (calls and reminder letters) within demographic variables which affect the strategy used in practice. However, this did not improve the model performance significantly, and resulted in some problems due to cells with little to no data.

The details of the model can be found in table A.4 in the Appendix. However, some of the key findings are outlined below. Calls both during and previous to the period are associated with higher response in the period. For example, a one unit increase in the 'previous call number' variable (log transformed) results in a 50.8% increase in the hazard[8] of response in that period. Non-gold businesses are much less likely to respond than gold businesses,. New units are more likely to respond than other 'non-gold' businesses but less likely than 'gold' businesses. There was no significant survey effect observed.[9]

The time variable (weeks since last written contact) is incorporated as a categorical variable for maximum flexibility[10]. We observe that, compared with the baseline zero weeks after the reminder contact, the hazard is about 6% greater one week after the reminder before declining below the baseline. This indicates that the maximum benefit of the reminder letter occurs in this time frame. Each subsequent reminder is more effective, with the hazard following the second reminder nearly 20% above the first reminder and the hazard following the third reminder 66% above the first reminder. However, it may not be the case that the number of reminders is important. It may be the wording of the reminder. Alternatively, it may be related to the time since the despatch of forms, or the timing of external factors (such as other government reporting). In reality it is probably some combination of the above, but the causes cannot be separated out.

---

7 Backward selection was considered but the model does not converge when all variables and their interactions are included.

8 Although not strictly a probability, the hazard can be thought of as the probability of an instantaneous occurrence of the event given that it had not already occurred.

9 This implies that any differences in aggregate response rates observed between these surveys are driven by differences in the explanatory variables rather than any fundamental difference between the collections. For example, annual surveys typically have a lower response rate due to the lower proportion of gold provider businesses. This may not extend to other surveys beyond those included here.

10 Hence we are estimating a semi-parametric proportional hazards model.

2.3 Economic survey response rate model diagnostics[11]

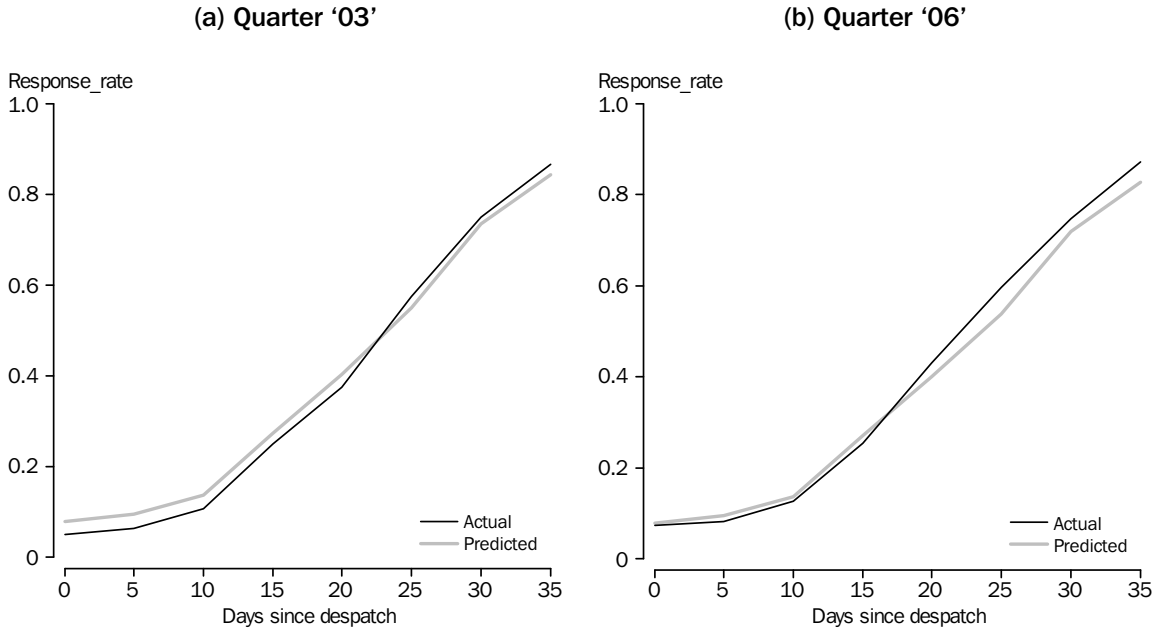
Diagnostic	*Result*
Max rescaled R-squared	0.2097
Area under ROC curve	0.7810
Overall error rate – Fitting data set	14.38%
Overall error rate – Testing data set	12.51%

In order to assess the model fit, we considered the area under ROC curve (c), max-rescaled R-squared and the overall error rate. We also considered the model performance in terms of its practical performance in predicting the actual overall response rates.

The model appears to perform reasonably well with the area under the ROC curve (c) near 0.8 and an error rate below 15% on both the fitting and testing data sets.

In order to predict response rates, the individual predictions can be compounded over time and averaged across businesses. Figure 2.4 below show the cumulative response rates predicted this way, compared with actuals, for selected quarters of QBIS. The '06' quarter data was used in fitting the model and unsurprisingly the model provides a close fit. We also tested the model on a cycle not used in the model fitting – the '03' quarter data was not used in fitting the model however the prediction from the model is still a close fit to the actual response rate.

2.4 Comparison between the predicted and actual response rate



[11] We split the data set randomly into 70% for fitting the model and 30% for testing.

As stated above, there is some risk of confounding effects. Checks of the variance inflation factor[12] showed no evidence of serious multicollinearity problems amongst the covariates (details are provided in table A.5 in the Appendix). However, there does remain a significant risk of this and hence the results should be used with caution. In particular, the findings around reminder letters should be used with care due to reasons mentioned above.

A few extensions were considered but have not been investigated at this stage; these are discussed further in Section 2.4.

2.4 Methodology issues

For household response model

Other techniques were also considered to construct the models to predict the probability of response.

As our data is hierarchically structured with respondents nested within households, within collection districts, within workloads, and within interviewers, we considered whether multilevel modelling was an appropriate technique to use. It was difficult at the time to ascertain whether multilevel modelling would help explain the variation at different levels of the hierarchy without first fitting the multilevel model. We also had concerns whether there were any practical limitations in multilevel modelling software given that our data is quite large (one month of MPS has approximately 100,000 calls and at least four months of MPS data were analysed).

Survival analysis was also considered as an alternative technique to model response propensity. We were unable to find any literature on an application of survival analysis in household surveys. The technique seemed more applicable to our business surveys, where the surveys are mail surveys with telephone follow-up and response can occur whether follow-up is conducted or not, whereas for household surveys response is only generated upon calls made. Campbell *et al.* (2004) applied survival analysis to a mail survey with telephone follow-up.

[12] Via SAS Proc Reg

For economic surveys

We used survival analysis with proportional hazard model with the hazard related to the time since last written contact (rather than calendar time) to model the business survey response rate. Although some anecdotal evidence shows this to be appropriate we have not tested the underlying assumption and it may not be the most appropriate way of accounting for the impact of time. We intend to investigate the use of alternative methods, principally data mining methods such as tree-based survival methods, to explore different models and determine whether improvements can be made.

As interviewers only occasionally collect data over the telephone, it is possible for a business to respond without any telephone contact whatsoever. Even when an interviewer does contact a business, connecting the call to a response which may occur some days later is difficult, as by that time a business may have received further calls, or a reminder letter, or may have been intending to respond regardless of any follow-up procedure. Some calls may not be made for the purpose of reminding a business to respond (but may be, for example, responding to an inbound telephone call, or following up inconsistent data) but in some cases these cannot be distinguished from follow-up calls. The only solution to these problems is to establish the data collection process as a randomised trial and hence collect more sound data, enabling more robust models to be fit, however this is not always practical.

Response behaviour is also likely to be linked not just to business characteristics but also to characteristics of the individual filling in the form; however this information is currently unavailable for use in modelling.

There is some evidence of seasonality with businesses more difficult to obtain responses from around busy business times (e.g. end of the financial year) and when businesses tend to be shut or operating on fewer staff (e.g. around Christmas time and Easter). However, this hasn't been included in the model at this stage since the focus was on linkage to annual (not sub-annual) costs, and therefore the results are yearly average behaviour, and for sub-annual collections some cycles may be more difficult than others.

An alternative would be to treat different types of response as 'competing risks' as it is possible for only one response-type event to occur for each business, but we expect the effect of the covariates may be different for different types of 'response'. To deal with the competing risks, separate models would be fit, each predicting the probability of a single type of response, with businesses experiencing any other type of response as censored. Such an approach would be expected to be more important in a survey with multiple modes of data capture as the mechanisms driving response may be quite different for businesses opting to respond via (for example) a webform

compared with a paper form. Therefore, this approach should be reconsidered once paradata regarding webforms becomes available.

More generally, as increased information (or improved quality of existing information) becomes available the model may be refit to take advantage of this information. This may involve new types of activity (such as webforms) or may involve recording more information about existing activity. For example, recording more information about a call, such as interviewer's subjective opinions about a business' probability of response is likely to be useful in predicting actual response.

The model assumes that individual businesses which are in more than one cycle are independent from one another, so each individual experiences the event of response only once. However, this leads to correlation amongst the observations and possibly biased results. Another, and better, option would be to treat response across multiple cycles (and even collections) as 'repeated events', to make better use of response behaviour information. However, conventional methods are currently not well equipped to deal with repeated events (Allison, 1995). This is an area for potential future research.

Finally, although a few major surveys were included in the model, model performance has not been assessed yet for other collections, and the model may not apply to some collections (for example for Agriculture collections, where respondent behaviour is known to be different from other collections).

3. COST MODELS

3.1 Cost model framework

Cost models can be used for several distinct purposes:

1. understanding how different activities contribute to ABS survey data collection costs,
2. predicting the *total* cost of a work program,
3. predicting how costs will *change* in response to a change in inputs, and attempting to *minimise* costs within constraints.

Our approach to cost modelling is intended as a compromise that supports each of these purposes to some degree, although with a preference towards change/optimisation applications. This helps ensure *coherence* between cost models for quoting or budgeting purposes (when not all the detailed information may be known) and cost models for identifying efficient and/or optimal practices.

It is also influenced by practical constraints: the need for compatibility with a response model, missingness and other limitations of the fitting data, and the complexity of ABS cost structures (increasing the risk of errors in our models of those structures).

Before developing a cost model, we explored simpler observational approaches. For example, MPS records allow us to estimate what the consequences would have been for a historical survey if we had ceased all follow-up work at a given date: calls data allow us to identify the households that only responded after that date (and hence would have been lost under an earlier cut-off), and payment data allows us to identify the costs incurred after this date (which would have been saved with an earlier cut-off).

This approach is useful in developing a basic understanding of survey costs, and is sufficient for some prediction problems, but it's limited in scope. For instance, while we can measure what would have happened if we'd made *less* effort on contact, observation alone can't predict what might have happened if we'd made *more* effort.

Another example of the limitations of observation is in exploring the balance between first-month and non-first-month effort in MPS. A single payment record can include work on both first-month and non-first-month households, making it impossible to determine by observation alone how important each type of work is to overall costs. Modelling helps determine how each type of work contributes to overall costs.

3.2 Household survey application

The household cost model attempts to predict payable time worked by interviewers and motor vehicle allowance ('MVA' – paid per kilometre driven).

Time and travel reports can be used to group activity into 'work- segments' (a continuous period of work); a work-segment spent travelling is referred to as a 'trip'. Call records provide information about finer-level activities: blocks and households visited, interviews conducted. By matching these records, we produce records that describe both the activities conducted within a trip and the total payable time/MVA for that trip. These are the basis for our cost model. Discussion here concentrates on the face-to-face time model; the distance model is similar but simpler, because interviewing and foot travel don't contribute to MVA. Cost modelling for telephone work is trivial, since interviewers are paid according to the number of fully-responding households.

3.2.1 Linear model

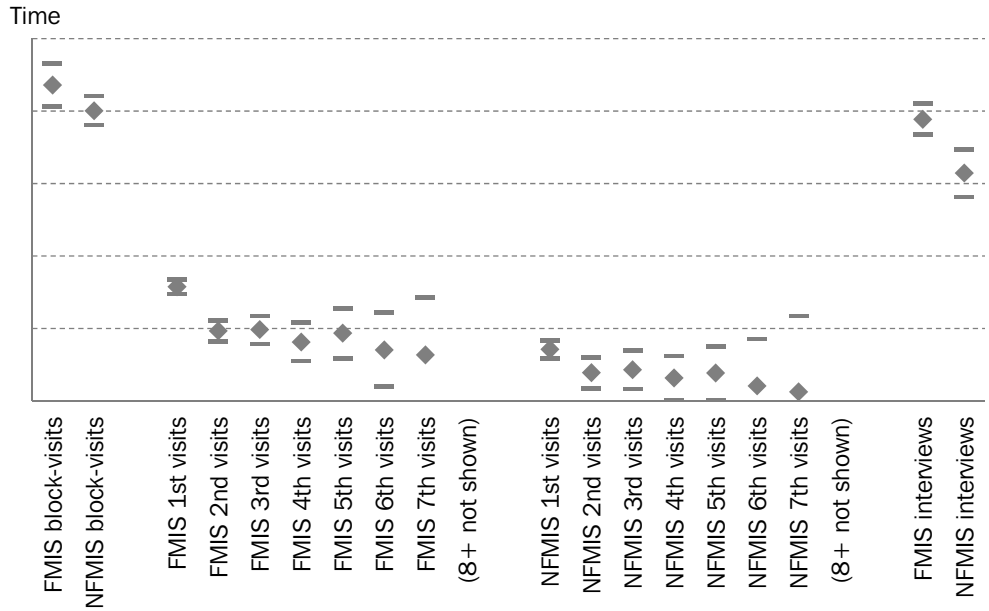
We fit Monthly Population Survey payable time and distance for trips as a linear function of several variables, chiefly:

- Number of blocks visited during the trip;
- Households visited during the trip (various ways to quantify, see below);
- Number of interviews conducted (or estimated interview-minutes).

Each of these variables was divided into 'first month in sample' (FMIS) and non-first-month (NFMIS) because work patterns for new MPS households are known to be quite different from those for months 2–8. Fitting was stratified by state, region (metropolitan / ex-metropolitan), and workload type (standard *vs* follow-up). Figure 3.1 shows example fitting results based on MPS from September 2007–March 2008. Explanatory variables are listed along the X-axis, with estimated time (minutes) and 95% confidence intervals on the Y-axis.

In this example we've distinguished between first, second, … household visits in the expectation that early visits might take longer (e.g. due to unfamiliarity with neighbourhood). As the plot shows, this is indeed true for the first visit. After the second visit there's little or no significant change.

3.1 Time spent for standard metropolitan workloads, MPS 0709–0803 (with 95% CI)



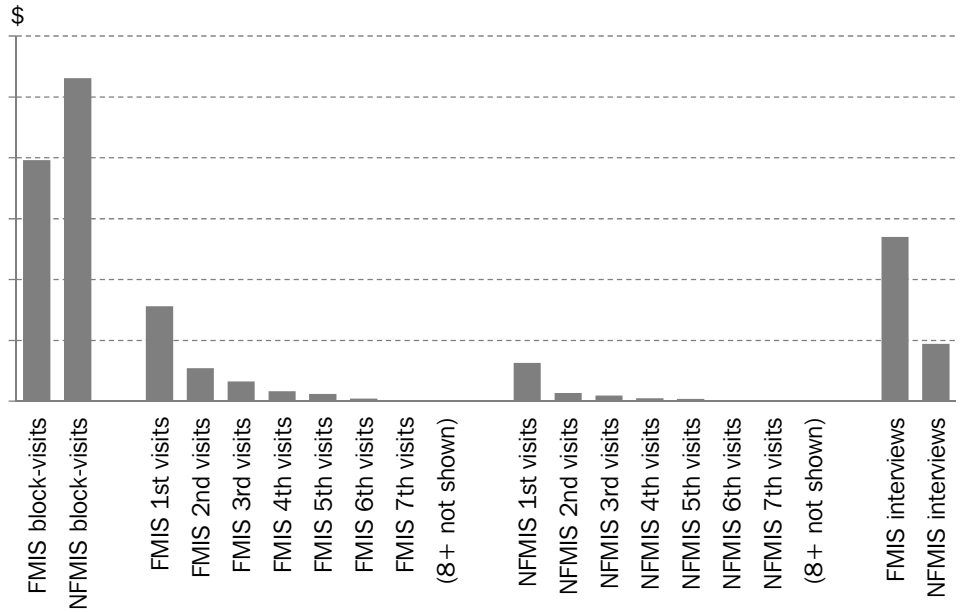
For later work we collapsed the 'household visits' variables, and the current version simply uses 'number of FMIS household visits' and 'number of NFMIS household visits'. This collapsing is likely to reduce fit quality, because of the first-visit difference noted above, but some of that effect will be absorbed by other explanatory variables correlated with first visits (e.g. number of interviews). Since most decision scenarios don't allow us to alter the number of *first* visits, it's often less important to optimisation problems. Simplifying to 'number of visits' makes the model easier and faster to use and interpret, and these benefits outweigh the reduction in fit quality.

Curiously, the 'interviews' variable is typically associated with a small positive distance effect (not shown above). Interviewers may be more willing to visit a distant household if they have good reason to expect an interview.

Combining the fitted model with explanatory variables allows us to estimate much each of these activities contributes to overall face-to-face costs for the period in question. Figure 3.2 shows the time cost break down for face-to-face work in standard metropolitan workloads of MPS surveys from July 2009 to March 2008.

This provides some useful insights into the factors that drive ABS costs. For instance, it's common knowledge in ABS that only a small proportion of NFMIS households receive face-to-face visits – it's easy to dismiss these as a minor issue. However, dissecting costs in this way shows that it's actually a major contributor to ABS operating expenses: NFMIS work is much less clustered than FMIS work, with important consequences for efficiency.

3.2 Face-to-face time cost breakdown for standard metropolitan workloads



When fitting these models for trip-level time/distance for a trip, we deliberately chose to fit our models with a zero intercept. This effectively means that when we apply the model at an aggregate level, e.g. predicting costs for a whole survey, we are excluding 'number of interviewer trips' from use as an explanatory variable. While including it might improve fit to historical data, it would make the model less reliable as a decision-making tool. 'Number of trips' as quantified from paradata can easily be affected by minor, arbitrary choices in recording – for instance, an interviewer can turn one long trip into two smaller ones by taking a short break in the middle without affecting payable time or MVA.

Instead, it seems safer to base the model on tangible variables that more directly describe the activities that contribute to time and mileage. Excluding 'number of trips' also makes it easier to see how operational choices affect our costs – because a single trip can include both FMIS and NFMIS households, it would make it difficult to gauge the importance of each category to total costs.

3.2.2 Calibration

Recall that we chose to use a zero-intercept fitting approach. While this has advantages, it is likely to cause biased predictions. We may also experience bias due to data quality problems: some trips are missing time/distance records (and hence can't be used in fitting) and others have incorrect values.

Most of the incompleteness/error is in dependent variables (trip-level payable time/distance etc.) rather than explanatory variables. This means that while we can only fit our initial model on a subset of the work from a given survey, we can then use it to make predictions for virtually all the work done. By comparing these predictions to

accurate records of total payments for the survey, we can determine the size of the resulting bias.

In general this bias is small (typically less than 10% of the total) and has little importance when using the model to optimise operational parameters. However, there are cases in which it's desirable to eliminate the bias altogether (e.g. for consistency with other cost data generated by other processes). One option for dealing with these is to calibrate the model: for instance, if comparing model predictions to payment records shows that it's underpredicting costs by 5%, we can scale up the cost model components by 5% to create an unbiased model. It's possible that this will reduce accuracy of some individual components in the cost model, but as long as the calibration factor is small this is unlikely to cause significant distortions.

3.2.3 Validation

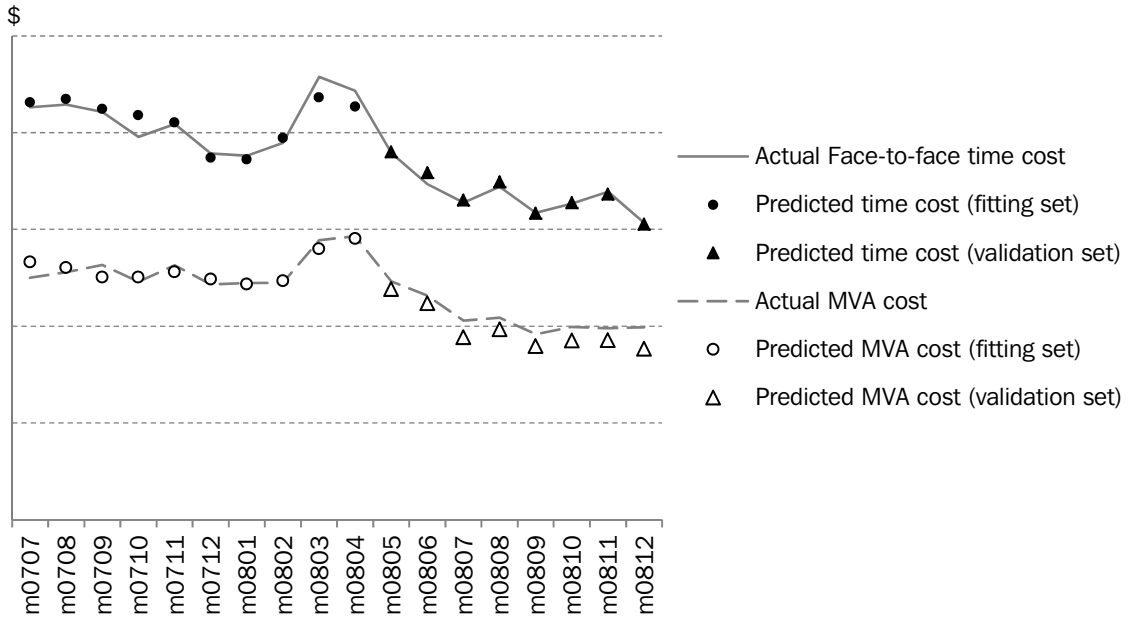
We did not examine fitting diagnostics at the 'trip' level because we did not expect strong predictive power at this level – individual trips vary greatly, and occasional data errors effectively create artificial outliers. Instead, we tested the model by fitting and calibrating on MPS data from July 2007 to April 2008, then making predictions for aggregate face-to-face time and MVA costs in May–December 2008 (based on records of visits etc. during this period). Figure 3.3 illustrates these comparisons, showing actual MVA and face-to-face time costs alongside the predictions made for fitting/calibration and validation sets.

Earlier validation attempts were affected by data quality issues. Due to a software bug many records from August–September 2007 were not usable even for prediction, so the uncalibrated model badly underpredicted costs for these months. This led to an over-correction in calibration, resulting in a consistent slight overprediction for time costs in the validation period. Clearly it is important to use appropriate data for calibration – while missing dependent variables may be acceptable, missing explanatory variables are not.

Even with this error, the calibrated model produced good predictions for the validation period, and results were further improved once the error was corrected. Figure 3.3 shows the results on corrected data, including accurate predictions of cost reductions caused by a MPS sample cut in mid-2008.

Predictions for non-standard workloads were poorer, due to a combination of smaller sample size, greater variability in work, and possibly higher rates of error in fitting/validation data. However, costs are dominated by standard workloads, and so the model performed well overall.

3.3 Comparison of model predictions with actual costs



Note that the cost model is not a *true* predictive model in isolation, because it still requires information about the work to be done during the prediction period (household visits, blocks visited, interviews conducted). Most of these can be provided by the response model discussed previously; predicting block-visits requires an auxiliary model (not presented here) to estimate how household-level visits are grouped into block visits.

3.2.4 Adaptation to Special Social Surveys

While the model was initially developed based on MPS, we also want to be able to make predictions for Special Social Surveys. SSSs differ from MPS in several ways: they often have a much longer interview (commonly 45+ minutes *vs* approximately 15–20 minutes for MPS), they usually only interview each household once, they may have different clustering, and they typically have lower response rates.

The ABS existing MPS cost models based on a 'cost per responding household' approach were not able to produce accurate predictions for SSSs. Since our MPS cost model takes a different approach, looking at the fine detail involved in getting a response, it may be better suited to model these differences – for instance, lower response rates in SSSs require more return visits, and our MPS model can predict how these would affect costs.

Being able to use a single model for both MPS and SSS work would be highly desirable since it would allow us to concentrate on improving that model rather than on developing and maintaining separate models for SSSs. We started by testing the MPS cost model on the Survey of Education and Training (SET), a fairly typical SSS most recently run in 2009 and previously in 2005. We attempted to predict costs for SET

2009, using two different approaches: fitting on SET 2005, and fitting on MPS (testing both March 08–February 2009 and December 08–February 2009). In the latter case, we treated all SSS work as 'first month in sample', since NFMIS work patterns are much less relevant to a single-interview SSS that relies on face-to-face interviewing.

SET 2005 and the 12-month MPS model both produced very good predictions for MVA (error < 3%), with the 3-month MPS model overpredicting by 9%. All three models underpredicted travel and interview time by 10–15%. The last result was particularly surprising since interview costs should be trivial to calculate – all completed interviews are paid at a standard rate. Further data exploration found explanations for these errors:

Discrepancies in interview counts: The cost model bases interview costs on the number of fully-responding households appearing in call records. However, the number of fully-responding households indicated in payment records is substantially higher; this discrepancy is large enough to account for the difference between model and true values for interview time. The cause of this discrepancy is not yet clear, but may be due to classification issues for standard vs non-standard workloads.

Payment for incomplete interviews: If an interviewer spends time interviewing a household but doesn't obtain a full response, they are paid for the time spent and this is classified as 'travel time'. This component is negligible for MPS but important in SSSs, which have longer interviews and a higher rate of incomplete interviews. Completed interviews averaged approximately 45 minutes; if we suppose that incomplete interviews have a similar average, this would account for for most of the error in MPS estimates of travel time.

We then performed similar tests on the Survey of Disability and Carers (SDAC 2009) with very similar results: good predictions for MVA and underprediction for travel and interview time costs, apparently due to the same factors noted above. Based on these results, it appears that the MPS model may be a good predictor for SSS costs – if we can resolve the discrepancy in interview numbers, and if we can predict the costs of incomplete interviews.

It's interesting to note that the 12-month MPS model performed better than the model fitted on SET 2005; this may be because MPS provides more data for fitting, or because the MPS data is more timely. It suggests that the fundamental relationships represented in our cost model (i.e. time/distance as a function of block-visits, household visits, and interviews) did not vary greatly between FMIS MPS and SET. Previous observations of cost differences between MPS and SSSs may be due to the factors noted above (incomplete interviews, data discrepancies) and/or with response/clustering issues that can be adequately dealt with through the 'household visits'/'block-visits' variables (assuming we are able to predict these things through response and auxiliary models).

3.2.5 Field trials

The cost and response models described above have allowed ORPI to make predictions about the effects of operational changes. For example, comparing cost and response in FMIS *vs* NFMIS households for MPS led to the conclusion that increasing contact effort in month one while decreasing effort in month eight would reduce costs while maintaining response rates.

Based on this prediction, ABS ran field trials with the modified procedures from August 2010 to January 2011. However, feedback from these trials indicated that it was difficult to accommodate the extra month-1 effort; while the total amount of time required may have decreased, the MPS working week is short enough that attempting to make extra visits would have required making multiple visits in a single afternoon, unlikely to be productive. This issue highlights the difficulty of extrapolating from a model based on observational data.

3.3 Economic survey application

Cost models for economic surveys are required to cover four major components of the survey data collection process:

1. telephone contact with businesses, including telephone reminders and resolving issues and complaints;
2. written contact with businesses, including postage of forms and reminder letters and receipt and scanning of returned forms;
3. processing of administrative data;
4. associated activity such as frames and selections, forms design and systems.

The existing cost models have been developed by ABS survey data collection areas and used for quoting costs of collection processes for Economic Surveys for several years. Rather than developing statistical models for cost breakdowns, the focus of our work has been to review the cost models, assess the accuracy of the model, and make improvements around key client concerns, especially model transparency in term of cost for the different components of a survey data collection process. The focus of this review was on the telephone contact component (hereafter referred to as 'Provider Contact Unit' or 'PCU') as the highest cost component and due to availability of detailed cost and effort data. A review of the costs of written contact, the second most costly component, is currently pending the availability of cost data. The last two components have simple cost structures (based on staff full time equivalent cost) and are low in cost relative to the first two, and hence have not been in-scope for review.

The PCU cost model is quite simple in structure and logically calculates the number of calls, which can then be converted to a number of staff days (via an average call rate) and to a cost (via an average cost per staff days, including overheads). The format (details of the fixed component excluded for brevity) of the pre-review cost model[13] is:

$$D = R \times \left(D_{\text{fixed}} + \alpha \times n \times \sum_{i \in S} P(T_i) \times C(i) \right)$$

where R is an adjustment factor for the difficulty of achieving a target level of response; D_{fixed} is the number of days required for non-IFU activity; α is the inverse of the call rate (number of calls per interviewer per day); n is the (annual) sample size, S represents the key elements in the IFU strategy, e.g. S = (initial despatch, written reminder 1, written reminder 2, telephone IFU phase 1, telephone IFU phase 2); T_i is the time when element i is executed; $P(T_i)$ is the proportion of businesses who have not responded at the time of element i is executed; $C(i)$ is the (average) number of calls per non-responding business made due to element i (inbound calls resulting from written contact, outbound calls made in each IFU phase).

In reviewing the performance of the model, the following steps were undertaken:

- a sensitivity analysis to identify the important parameters;
- paradata was used to assess the accuracy of the values used for these parameters;
- compare predicted costs with actual costs using the existing parameters;
- assess accuracy of all remaining parameters and repeat;
- check accuracy of final updated model.

Note that the review focussed on the accuracy of estimates of staff days rather than $ costs as survey-level information was only available in this format.

The details of this review are excluded from this paper, however the key findings were:

- the assumed values for many of the parameters were highly inaccurate for many collections (up to $\pm 50\%$);
- the resulting cost estimate is highly sensitive to a number of these parameters, and therefore it is important to estimate these parameters as accurately as possible;
- if the PCU cost model were to be used as-is for existing collections, very poor cost estimates would be obtained (different from actual costs by $\pm 50\%$, see table 5 below);

[13] Note that the parameter values are populated using assumed values which can be validated by paradata. They do not come from a statistical model.

- the fixed non-IFU component did not include all activity it should;
- the adjustment factor (response factor) used to account for the difficulty of achieving the specified level of response was not based on evidence and was not appropriately applied as it adjusted all activity when theoretically it should apply only to activity associated with seeking a response (outbound calls).

This work led to the following upgraded form of the model:

$$D = D_{\text{fixed}} + \alpha \times n \times \sum_{i \in S} R_i \times P(T_i) \times C(i)$$

where R_i is an adjustment factor for the difficulty of achieving a target level of response.

$$R_i = \begin{cases} r_i & i \in (\text{telephone IFU phase 1 , telephone IFU phase 2}) \\ 1 & \text{otherwise} \end{cases}$$

where the call rate, number of calls and proportion of businesses who have not responded are all estimated from historical call records. Alternatively, in the linked model, estimates of the proportion of businesses who will not have responded can be populated from the probability of response model.

The response factor r_i was determined by simulating the impact of a decreased response rate on the number of calls needed to be made (hence $R_i \times C(T_i)$ is the new number of calls made under a changed response rate strategy).

Three distinct groups of collections were identified in terms of the relationship between effort and response rates (subannual collections, annual collections with effort distributed over a long period, and annual collections with effort mostly contained in a short period, usually at the very end of cycle). Therefore, three sets of response factors were generated by averaging the response factors generated across multiple collections and cycles.

3.3.1 Performance of the upgraded model

Table 3.4 compares the performance of the previous model with the upgraded model for a selection of surveys. The table also contains the 'PCU estimate' which is the planned number of staff days to be used, a 'best guess' based on the number of staff days used in the previous years, and is generally close to the actual number of days used (as it forms the starting point for staffing allocation). As can be seen, the previous model predictions were very inaccurate whereas the upgraded model predictions are accurate to within $\pm 10\%$ (for all collections, not just those shown here). In fact, for both QBIS and AWE, the model prediction is closer to the actual number of days than the 'best guess' PCU estimate.

3.4 Comparison of errors between the existing and proposed PCU cost model

Survey	*PCU estimate*	*Previous model prediction*	*Upgraded model prediction*
QBIS (08/09)	+8%	–37%	0%
AWE (08/09)	–9%	+19%	+2%
AIC NSW (07/08)	+3%	–14%	+10%

As the primary focus of the PCU cost model is to accurately predict costs *for quoting purposes*, the model was therefore determined to be fit for purpose and is currently being implemented.

3.3.2 Further work

A number of other areas for cost model development for economic surveys have been identified. This includes a review and possible redevelopment of the current survey form despatch and collection cost model, which includes printing and postage of forms and reminder letters and scanning of returned forms, which is currently underway. There are also some potential areas for development of the PCU cost model to take account of other activity such as training and emerging challenges such as web data collection. The model could also be redeveloped to produce more detailed cost estimates and capture more aspects of the collection process.

Whether such work can be undertaken will depend on the future availability of appropriate paradata for analysis as this information is currently unavailable or of insufficient quality for analysis.

4. BIAS MODELS

4.1 Bias model framework

4.1.1 Role of bias model in survey planning

The bias model attempts to predict the effect of a change in the number and characteristics of respondents to a survey on the bias for specified items of interest, and also the resulting additional uncertainty in estimates.

For example, suppose it is proposed to reduce the required number of callbacks an interviewer makes from ten to six per household. The cost model predicts a $20,000 saving. The response model predicts that response will reduce from 80% to 78% and that the difference consists of dwellings that would have responded after seven to ten interviews. A bias model uses the predicted change in characteristics to provide, for an item of interest, the expected change in level, and measures of the corresponding change in estimate quality.

4.1.2 Grouping units by difficulty of response[14]

Suppose that the survey population is assigned, perhaps by some random process, to one of three groups before selection:

- Easy: will respond if selected, regardless of changes to method
- Hard: might respond and might not, depending on method
- Impossible: won't respond using any variation of the available methods

Of course, all the data we have is in fact for respondents that do respond; for these we hope to obtain a model relating characteristics of the respondent to the expected value of an outcome variable. The model is only assumed to apply for respondents in the Hard group. This model can then be applied to a proposed change in the responding population.

4.1.3 Modelling the residuals

Bias for a variable y_i is estimated using the residuals ∂y_i. The residual gives the difference between a respondent reported value and the value that would have been imputed, or implicitly predicted, if that unit had not reported.

For GREG weighting the predictions are denoted $x_i\beta$ and can be requested from GREGWT when requesting a table of the variable y_i. The residuals are then calculated as $\partial y_i = y_i - x_i\beta$. It can be shown that for GREG estimation with weights w_i the effect on the estimate from removing unit i is approximately $w_i \partial y_i$.

[14] Units are households or businesses as appropriate.

It is proposed to model the residuals for respondents that may or may not respond, based on the responding ones in the Hard group. Each one is described by a row vector of characteristics $z_i = (z_{i1}, \ldots, z_{iK})$ such as number of callbacks required to obtain response. We assume a linear relationship applies between the residuals and these characteristics, of the form

$$E(\partial y_i) = z_i \gamma$$

for γ a column vector of parameters, $\gamma = (\gamma_1, \ldots, \gamma_K)'$.

4.1.4 Predicting change in bias

A change in data collection strategy is summarised by ΔZ, the resulting change in the total of the vector z_i corresponding to the changed responding population. The change in bias from this change in strategy is then predicted by

$$\Delta \text{bias} = \Delta Z \gamma .$$

For a chosen model, estimating γ can be done straightforwardly using unweighted least squares to obtain an estimate $\hat{\gamma}$ by fitting to the hard data from some previous survey that is considered similar.

Unfortunately, choosing a model and predicting the change ΔZ for a change in strategy may not be as straightforward. Some different cases arise.

1. Cuts: (e.g. reduce number of callbacks to certain types of respondents)

   These can be straightforward in cases where we are confident from previous data how many respondents would be affected, and they can be identified by their characteristics z_i. Some judgement may be needed to ensure that the types of respondents that would be affected by the cut can be identified. For example, a strategy to reduce the follow-up period in a business survey would include a way to ensure that the largest businesses were dealt with in the reduced period. So the change would affect the slowest businesses to report, but fewer of the largest businesses, and this would need to be reflected by the ΔZ value.

2. Increases: (e.g. increase number of callbacks)

   To measure the bias reduction from an increase in follow-up, we must propose how many extra respondents we would gain with the new callback policy, and also their characteristics. Again, there are pitfalls in this. For instance, suppose that we insist on household interviewers making at least eight callbacks rather than a previous requirement of six. Data from an earlier survey may have households that responded in seven or eight callbacks, but these may have been

called because the interviewer was confident of obtaining a response e.g. they had been asked to ring back. Such households may not be typical of the extra households that would result from the new policy. Additional paradata to identify these cases would help here – otherwise some assumptions will be required in deciding ΔZ for the new situation.

This highlights a second issue, which is that for measuring the effect of an increase we have to extrapolate from the available respondents. For a large change it will be unclear that the fitted model still holds for the new respondents.

3. Changes: (e.g. change the interviewer quality)

To predict the bias for a change like this requires proposing the change in responding population associated with the new interviewer quality. This may need to be thought through carefully along with modelling the response e.g. results will be like those before the change, plus or minus some extra people that would be by the good interviewers. A simplistic substitution into the model may not be appropriate e.g. if we were to move 15% of the sample from inexperienced to experienced interviewers, this might only result in 2% of extra response, with ΔZ representing their contribution; the remaining units being assumed to have basically unchanged response from the use of better interviewers.

4.1.5 A plausible range for change in bias

Typically we have previous data to model with, and we would like to be confident that the bias change from a new strategy is not too extreme. It is proposed to predict a plausible range for the change in $\Delta \text{bias} = \Delta Z\hat{\gamma}$ that takes account of its sampling error. The plausible range for change in bias is given by $[LCB, UCB]$, for

$$LCB = \Delta Z\hat{\gamma} - \Phi \,\text{se}\left(\Delta Z\hat{\gamma}\right) = \text{ lower bound of change in bias,}$$

and $$UCB = \Delta Z\hat{\gamma} + \Phi \,\text{se}\left(\Delta Z\hat{\gamma}\right) = \text{ upper bound of change in bias,}$$

for Φ an appropriate percentile of the normal cdf (e.g. $\Phi = 1.645$ for a 90% confidence interval, $\alpha = 0.05$).

The standard error $\text{se}\left(\Delta Z\gamma\right)$ is estimated as the square root of $v\left(\Delta Z\gamma\right)$ given by

$$v\left(\Delta Z\gamma\right) = V1 + V2$$

where $V1 = \Delta Z \,\text{var}\left(\hat{\gamma}\right)\Delta Z'$ is the uncertainty from estimating γ for our model (treating ΔZ as known), and $V2 = \text{var}\left(\Delta Z\gamma\right)$ is the uncertainty from estimating ΔZ (treating γ as known).

The bias estimate $\Delta Z\hat{\gamma}$ and the variance component $V1$ can be estimated using the regression software in which the model for γ is fitted, by treating the ΔZ as a specified 'contrast' to be estimated. Alternatively, the variance–covariance matrix $\text{var}(\gamma)$ can be stored and $V1$ calculated by matrix multiplication. Note that the modelling of γ is carried out unweighted, and the resulting variance will treat the residuals as independent between units.

The component $V2$ applies if the value ΔZ is uncertain; for instance, if it is measured from a previous survey. For some purposes $V2$ can be set to zero, as we are predicting a bias conditional on a given value of ΔZ. For other purposes we may choose to use a non-zero value representing a judgement of how uncertain we are. If the uncertainty in ΔZ comes from estimation, $V2$ can be estimated by attaching $z_i\hat{\gamma}$ to the survey file as a variable to be estimated from the survey.

Measure 1: Plausible change in bias

The plausible change in bias will be defined as the more extreme of LCB and UCB:

$$PCB = \begin{cases} LCB & \text{if } |LCB| > |UCB|; \\ UCB & \text{otherwise.} \end{cases}$$

For a level estimate Y^G of population total Y, this can be presented as a percentage:

$$PCB\% = \frac{PCB}{Y^G} \times 100\%$$

4.1.6 A plausible range for change in mean squared error

The mean squared error (MSE) is the usual measure of quality for a survey that potentially has biased estimates. The MSE for an estimate Y^G of population total Y is given by

$$\text{MSE} = \text{var}\left(Y^G\right) + \text{bias}^2$$

As for change in bias, it is proposed to predict a plausible range for the change in MSE resulting from a change in strategy. Unfortunately, the change in MSE will depend on the bias in the current estimate as well as the change in bias. If the overall bias changes from bias1 before the change to $\text{bias1} + \Delta\,\text{bias}$ afterward, the squared bias changes by $\Delta(\text{bias}^2)$, given as follows:

$$\Delta\left(\text{bias}^2\right) = \left(\text{bias1} + \Delta\,\text{bias}\right)^2 - \text{bias1}^2 = 2\,\text{bias1}\,\Delta\,\text{bias} + \left(\Delta\,\text{bias}\right)^2$$

A plausible range for current bias

In modelling the current bias, assume that some non-responding units (a proportion h of the population) will be similar to a similar number of the responding units that were hardest to obtain data from. Writing Z^{H} for the change in the total corresponding to these hardest units, we can predict the current bias as

$$B1 = Z^{\mathrm{H}}\hat{\gamma}\,.$$

There is uncertainty in this bias due to uncertainty in estimating $\hat{\gamma}$ and choosing Z^{H}. It is also possible that the non-respondents are not like the values our estimation assumes for them, even on average. To represent this additional uncertainty, treat the average residual for the Np^{NR} non-responding units as a random variable ε with mean 0 and variance $\mathrm{v}(\varepsilon)$. A plausible range for the current bias can then be given by $[LB,UB]$, for

$$LB \;=\; Z^{\mathrm{H}}\gamma - \Phi\sqrt{v\left(Z^{\mathrm{H}}\gamma\right) + Np^{\mathrm{NR}}\mathrm{v}(\varepsilon)} \;=\; \text{lower bound of current bias}$$

and $$UB \;=\; Z^{\mathrm{H}}\gamma + \Phi\sqrt{v\left(Z^{\mathrm{H}}\gamma\right) + Np^{\mathrm{NR}}\mathrm{v}(\varepsilon)} \;=\; \text{upper bound of current bias}\,.$$

Similarly a plausible range for the overall bias is given by $[LOB,UOB]$, for

$$LOB \;=\; \left(\Delta Z + Z^{\mathrm{H}}\right)\gamma - \Phi\sqrt{v\left(\left(\Delta Z + Z^{\mathrm{H}}\right)\gamma\right) + Np^{\mathrm{NR}}\mathrm{v}(\varepsilon)} \;=\; \text{lower bound of overall bias}$$

$$UOB \;=\; \left(\Delta Z + Z^{\mathrm{H}}\right)\gamma + \Phi\sqrt{v\left(\left(\Delta Z + Z^{\mathrm{H}}\right)\gamma\right) + Np^{\mathrm{NR}}\mathrm{v}(\varepsilon)} \;=\; \text{upper bound of overall bias.}$$

A plausible range for change in squared bias

Rewrite the change in squared bias as follows:

$$\begin{aligned}\Delta\left(\mathrm{bias}^2\right) \;&=\; \left(\mathrm{bias1} + \mathrm{bias1} + \Delta\,\mathrm{bias}\right)\Delta\,\mathrm{bias}\\ &=\; \left(\text{current bias} \;+\; \text{overall bias}\right)\times\; \text{change in bias}\,.\end{aligned}$$

The effort in obtaining an exact confidence interval for this may not be commensurate with the quality of inputs such as Z^{H} and $\mathrm{v}(\varepsilon)$. Instead, we propose substituting the upper and lower plausible bounds for the components as developed above. Thus a plausible range for the change in squared bias can be given by $[LCSB,UCSB]$, for

$$LCSB = LCB\left(LB + LOB\right)$$

and $$UCSB = UCB\left(UB + UOB\right).$$

A plausible range for change in root mean squared error

Let ΔV be the change in variance caused by the change in survey processes. For small changes this can be adequately represented by $\Delta V = V\Delta n/n$ where n is the current sample size and Δn is the change in sample size. (For large changes more detailed estimation of variance change may be required). A plausible range for the change in RMSE (the square root of the mean squared error) can then be given by $[LCE, UCE]$, for

$$LCE = \sqrt{\Delta V + LCSB}$$

and

$$UCE = \sqrt{\Delta V + UCSB}\,.$$

The RMSE is used instead of the MSE as it is on the same scale as the estimate itself, and can be compared to the survey standard error.

Measure 2: Plausible change in error

The plausible change in error will be defined as the more extreme of LCE and UCE:

$$PCE = \begin{cases} LCE & \text{if } |LCE| > |UCE| \\ UCE & \text{otherwise.} \end{cases}$$

This can be expressed relative to the current SE of the survey, given by $\sqrt{V}$:

$$\text{rel}PCE = \frac{PCE}{\sqrt{V}} \times 100\%\,.$$

4.1.7 Some notes about these plausible ranges

Bias to the movement

Sometimes the interest is in the difference in estimates between this survey and a previous survey that was not affected by the proposed change in methods. In this case the plausible change in bias is the key considerations (measure 1 above). For this situation the current bias can be seen as the bias to movements if we retained the current method, and hence can be set to zero. This feeds through to produce a plausible change in error for the movement, if this measure is required.

Recommendations for unknown parameters

In predicting the current bias it is suggested that Z^{H} correspond to the 'hardest' 2% (i.e. $h = 2\%$) of the respondents. Thus we assume that we are missing 2% of respondents that are like the current hardest respondents. It seems unlikely that the bias model will apply for much larger increases in response rate.

The other figure required is $\mathrm{v}(\varepsilon) = \left(\mathrm{se}(\varepsilon)\right)^2$. The value ε gives the average difference between a non-respondent and the prediction that will be used for them in estimation. This has some uncertainty, expressed in a Bayesian sense by the prior $\varepsilon \sim \mathrm{N}\left(0, \mathrm{v}(\varepsilon)\right)$. Perhaps $se(\varepsilon) = 0.02Y/N$ is appropriate, so that we are 95% confident that the average prediction will be out by less than 4% of the average value.

After some experience with the method we expect to be able to give workable levels for these two parameters.

4.2 Household survey application

ABS household survey data is subjected to regression weighting (GREG) that looks for correlations between the output variable y (e.g. employment status) and 'benchmark' variables x for which accurate totals are known (e.g. demographic breakdowns). The GREG residual for a given household is the difference between the household's *actual* value for that variable, and the value that would have been *predicted* for it by the regression weighting, based on its demographic characteristics.

The GREG residual provides an indicator of that unit's impact on estimates. Thus, if we can find a relationship that allows us to predict the residual (e.g. by use of other variables z that are not included in GREG weighting), we can predict how estimates would change if we change that unit's response status.

When we consider a scenario that might alter response in many units (e.g. by changing follow-up practices or by trying to retain experienced interviewers), we can predict how the extra/lost respondents might affect the total values of z in the responding population; this then allows us to predict residuals for these units, and hence the overall effect on estimates.

To do this, we use call records to provide information on the effort made to contact each household and the outcome of that effort (e.g. number of calls made, final response status, date of response). This was used to order responding households from 'easiest' (households that responded on the first call, early in the survey) through to 'hard' (households that only responded after many calls).

For this initial exploration we examined the effects on residuals for labour force status and 'hours worked' estimates. We chose explanatory variables that are likely to change between different scenarios, using backward selection to select from candidates including:

- Number of calls required to contact households,
- Day of contact,
- Time of contact,
- State,
- Region (metropolitan/ex-metropolitan),
- Month in sample (1–8),
- Workload size (how many households the interviewer is covering),
- Mode of contact (face-to-face or telephone),
- Interviewer age,
- Interviewer experience.

Model results

The resulting model predicts residuals in the 'hard' population as a function of various descriptors. For instance, 'employed' has a positive correlation with 'weekend'. This implies that if we make some extra effort on a weekend and manage to secure more responses, those households' residuals are likely to be positive – i.e. GREG weighting was underestimating the likelihood that they're employed. This means that by boosting our response in this way, we would probably cause employment estimates to increase.

Table 4.1 summarises the model results (some details omitted for brevity).

4.1 LFS bias model results

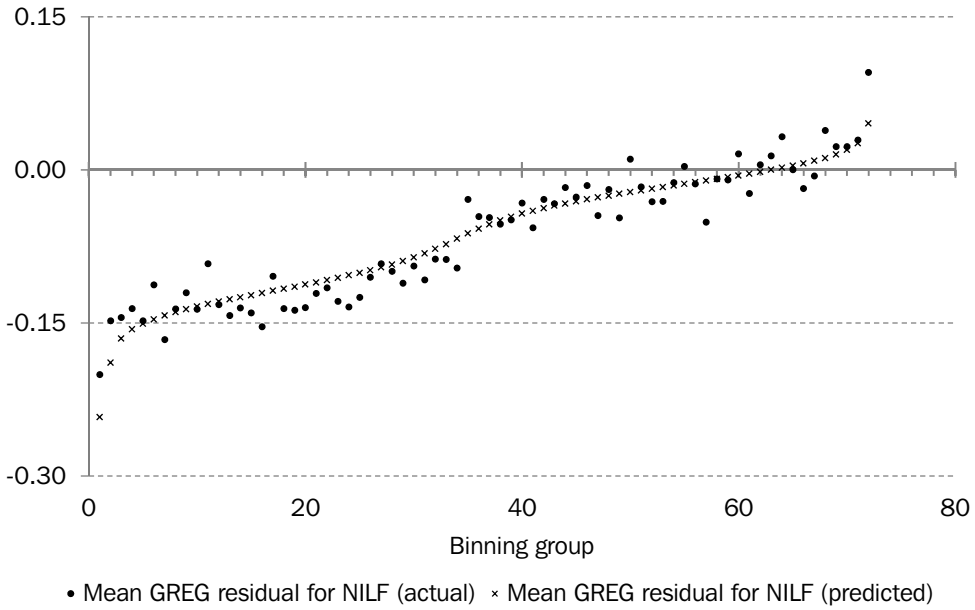
Output variable	*Significant, positive correlation*	*Significant, negative correlation*	*Significant, mixed*
Employed	Ex-metropolitan Workload size Evening Weekend	Interviewer experience Face-to-face Months in sample	State Interviewer age* Number of calls**
Unemployed	Face-to-face Monday–Tuesday	Evening Weekend	State Number of calls
Not in labour force	Face-to-face Months in sample	Ex-metropolitan Workload size Evening Weekend	State Number of calls
Hours worked	Ex-metropolitan Evening Wed–Fri Weekend	Face-to-face Months in sample	State Number of calls

* Interviewer age showed a second-order relationship.
** Number of calls' showed a complex relationship, represented by a piecewise linear approximation.

These results generally appear plausible. For instance, the model predicts that obtaining additional response on weekends, or in the evening, would increase employment estimates – not surprising, since employed people are hard to contact during working hours on weekdays. Effects on 'hours worked' are similar, probably for the same reasons, while 'not in labour force' covers most people who aren't employed, so the model is reversed. The model for 'unemployed' has fewer significant explanatory variables, probably because of sample-size issues.

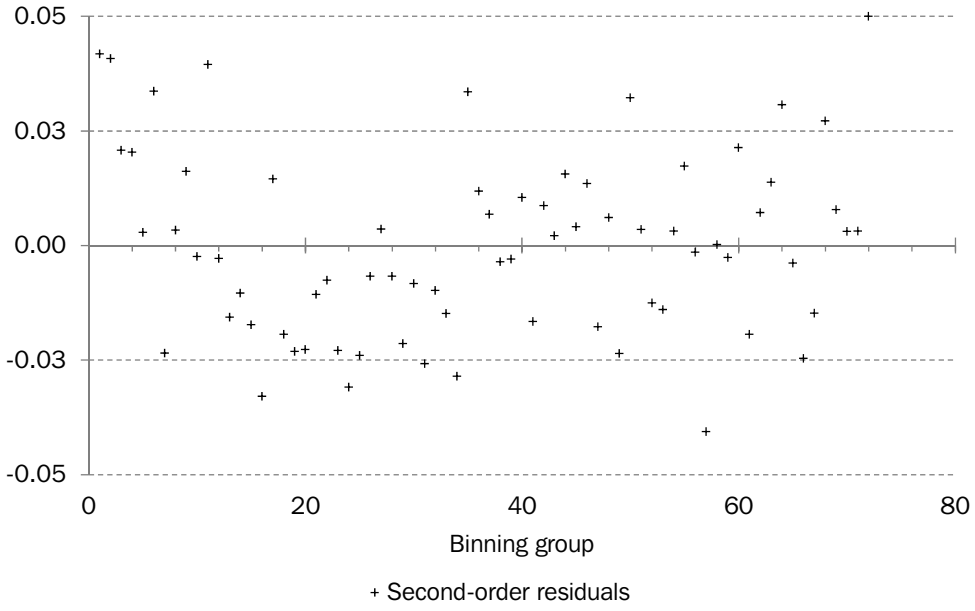
Fitting diagnostics for this model are extremely poor at the unit level (typical R-squared values are approximately 0.01). This is not surprising; GREG weighting has already removed those effects that can be predicted from demographic variables, so remaining systematic effects at the individual level are expected to be small compared to the 'noise' created by individual variation. Since the purpose of the model is to produce aggregate-level effects, it's more relevant to examine how it performs on large groups. To do this, we used a binning approach, combining respondents into groups of approximately 1000 individuals. Figure 4.2 shows the predicted and actual mean GREG residuals for each such group in the 'not in labour force' bias model.

4.2 Binning plot of Not in Labour Force bias model fitting



We can also examine the 'second-order residuals' (i.e. errors in our prediction of the GREG residuals for these binning groups).

4.3 Second-order residuals for the Not in Labour Force bias model



The relationship between predicted and actual GREG residuals is noisy, which is probably to be expected – we're looking at the residuals from a weighting process that has already extracted a great deal of information about respondents based on demographic factors. It is clearly impossible to predict the residual for a given individual with any degree of accuracy, and even predicting for a group of 1000 has a large margin of error. However, the plot suggests that the model is useful for making predictions for larger groups. The second-order residual plot shows no obvious pattern.

4.3 Economic survey application

The bias model for economic surveys was developed using the theory shown in Section 4.1. As a case study, several key variables from the Quarterly Business Indicator Survey (QBIS) and Survey of Average Weekly Earnings (AWE) were examined.

The general idea of the bias model is to identify the impact of losing or gaining particular businesses through a change in response status. The overall impact to the estimate (or 'bias') is calculated by then summing up the individual impacts across all businesses which change response status.

The bias model focuses on the 'hard population' or the businesses (units) which will waver in response depending on the strategy. In the case of QBIS, 'hard units' were defined as units which took at least three calls to respond, 'non-gold provider'

businesses[15] and non-crucial units.[16] Crucial units were excluded from the 'hard' population because they will always be followed up until they respond regardless of the strategy.

The residual was defined to be the difference between the actual reported value and the *expected* reported value. The expected reported value was an assumed impute value, based on a simplified version of usual imputation schemes.[17] This was a historical unit impute where one was available, and otherwise a stratum live respondent mean. The residuals were then modelled using linear regression using demographic variables and number of call backs.

The variables for number of successful outbound calls to businesses (directCalls) has been treated as a piecewise continuous variable where each variable represents the number of additional number of calls required for the business to respond. This gives a great deal of flexibility in capturing the effects of the number of calls as there is natural interpolation if not all of the directCalls variables are significant.

The other explanatory variables included in the model were state and employment size, in order to be able to integrate the bias model with the response model. Other demographic information such as industry classification were considered but have not been included at this stage.

The model was fit was found to have very low explanatory power, with an R squared value of less than 1%. Several diagnostics were used to test the validity of the model.

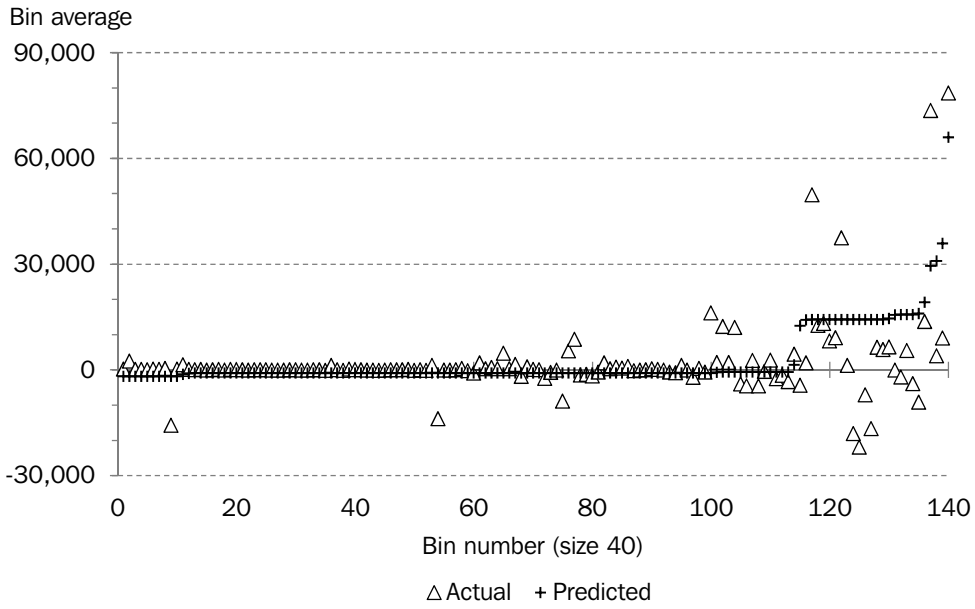
Firstly, a binning plot was used, which shows relatively good predictions on average for small residual values, but poor prediction for larger values. It can also be seen that the model predicts very close to zero for most businesses. Figure 4.4 is the binning plot with the most extreme values removed. It was hoped that although the model may be poor at predicting for individual businesses, for a group of businesses, on average it the prediction will be reasonable. The plot shows that this is not the case for larger businesses where there is a serious heteroscedasticity problem.

15 In general 'gold provider' businesses (as defined in Section 2.3) respond well and so are excluded from both follow-up and the 'hard population'.

16 Crucial businesses are based on the same subjective judgement applied in practice and for the response model.

17 We also assume that the imputed values and imputation scheme don't change with the strategy. We expect this to hold approximately for small changes to the strategy but not for major changes.

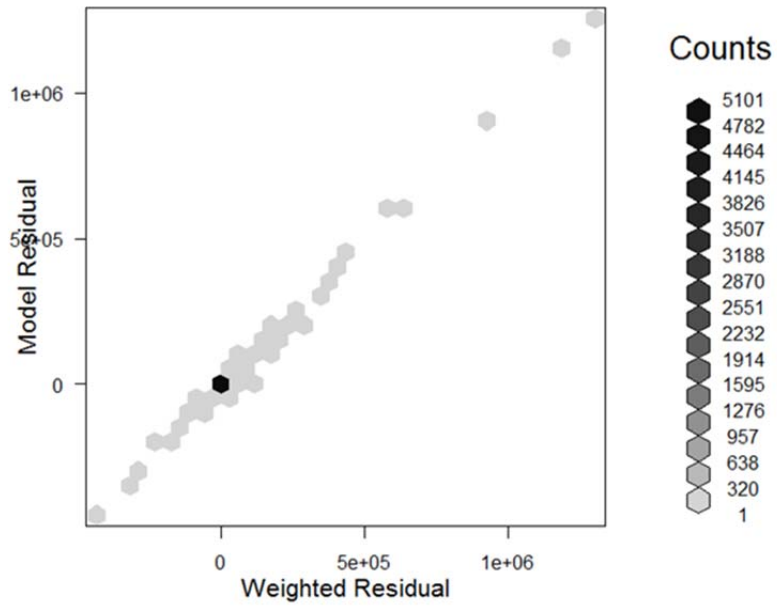
4.4 Binning plot of QBIS bias model fitting



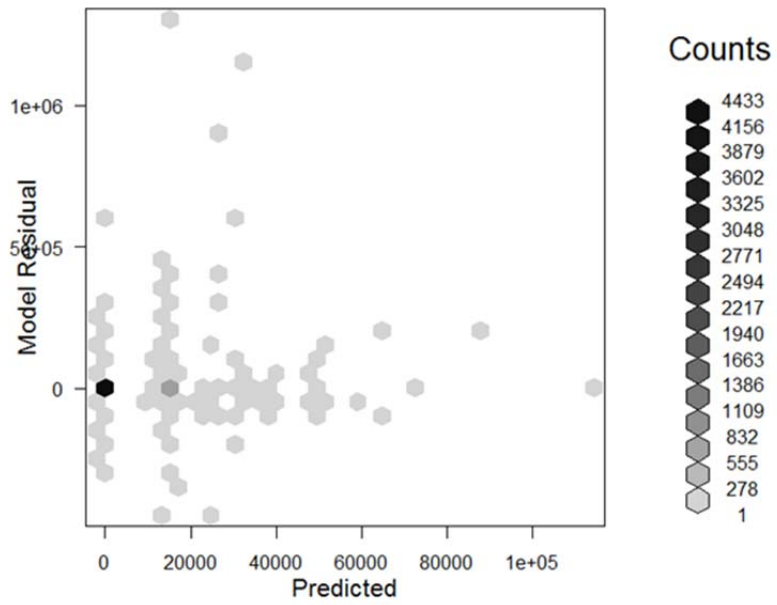
We then examined some standard linear regression model diagnostics. For a linear regression to be valid, it requires the residuals to be distributed $N(0,\sigma^2)$. Figure 4.5 provides standard output from regression procedures. The first diagnostic, 'Residual[18] *vs* Dependent Variable' is almost perfectly correlated, which shows that the model assumptions are invalid. The 'Dependent *vs* Predicted Value' plot looks like a reasonable scatter, but looking at the scale of the axes shows that the model predicts zero (as is consistent with the other plots) and that the residuals are quite large in comparison, demonstrating a low explanatory power of the model.

[18] Note that there is some risk of confusion with the term 'Residual'. We use the term 'residual' and 'weighted residual' interchangeably for the dependent variable in the regression (as we considered weighted and unweighted versions), however the two regression diagnostics use 'model residual' to indicate the residuals from the model fitting (or second order residuals).

4.5 QBIS bias model residual diagnostics: Residual vs Dependent Variable



4.6 QBIS bias model residual diagnostics: Dependent vs Predicted Value



We observe that the vast majority of observations have a value close to zero (indicated by the densities on the residual plots), however the tails of the distribution are very long and, particularly in the positive direction, there are some extreme outliers.

In order to apply the model to calculate the bias of a particular strategy, the predicted residuals would need to be predicted for all the businesses in the hard population which change response status and totalled to calculate the measures outlined in Section 4.1.4 – 4.1.5. However, due to the model not satisfying its assumptions, the results so obtained are not likely to be reliable and hence we have not progressed this yet.

4.4 Methodology issues

Household surveys

One of the most useful features of the bias model is that it's not dependent on the exact mechanism used to modify response – as long as we can predict how that change in response alters the z variables in the responding population, we could apply it. However, while the results above suggest that the model is valid, it's not clear how far we can safely apply it.

For example, the model states that the people contacted on weekends have high residuals for employment, and we can interpret this as saying weekends are a good time to contact those 'hard' respondents who work long hours. Since most interviewers begin work early in the week (Mon–Tues), most 'easy' respondents have already been captured early on.

But what if interviewers were encouraged to start on a Saturday and work through to the following Friday? In this scenario we would expect to see a great increase in the number of weekend responses, because response is fastest near the beginning of a workload. However, this increase would *not* reflect a change in the 'hard' respondents; rather, it would reflect the fact that a lot of 'easy' respondents have been shifted from weekdays to weekends. Since the bias model assumes that changes in these variables are related to response within the 'hard' category, it may not be applicable to such a scenario.

Desirable data needs

'Hard' and 'impossible' households are not homogenous groups. There are many reasons why a household might not respond after repeated calls, including:

- Distrust of the ABS,
- Difficult to contact (busy),
- Difficult to contact (doesn't keep appointments),
- Illness/other emergency,
- Vacant dwelling.

These different types of households may have little in common, beyond the fact of their nonresponse. This creates a risk that the characteristics of non-respondents may not show a linear relationship to response rates: the 'only just responding' households used to fit a model might represent the tail-end of one group, with the 'almost responding' households dominated by a different group.

In order to resolve these questions, it would be useful to have more insight into the makeup of non-respondents and late respondents. For example, if we knew that 70% of late respondents are 'hard to contact' and 30% are 'uncooperative', while non-respondents show a different breakdown, we would be able to adjust bias predictions accordingly. One option would be to do this within the current bias model framework by using these classifications as explanatory variables; another would be to stratify response by these classifications and fit a separate bias model for each category.

At present, it's not practical to obtain this information from ABS paradata systems. However, some of these systems are likely to be redesigned in the near future, presenting the chance to acquire more data. To make the most of this opportunity, we need to identify the information that would be most useful to bias (and other) modelling.

Economic surveys

One of the reasons the regression model struggles to predict the residual is because often imputations are quite accurate, and hence the difference between the actual and expected reported values will be very close and the residual will be very close to zero. Where there are differences it is difficult to predict them with any available explanatory variables.

Only overall total estimates have been considered at this stage for simplicity, however these estimates are the ones least likely to be impacted by response bias. Fine level estimates are prone to response bias due to the small number of businesses available for the estimation. Furthermore, there are less businesses in the 'hard' population which can be used to estimate the response bias. It was difficult to estimate the bias of a 'ratio' estimate as was found with the Survey of Average Weekly Earnings, even by estimating the numerator and denominators separately. In addition movement estimates have yet to be considered.

The binning plot shows that there is a heteroscedasticity problem which can be fixed by taking the logarithm of the residual (dependent variable). In order to do this, the absolute value needed to be taken first (and add one) to ensure that the value was positive. Although the transformation did remove the heteroscedasticity (albeit leaving some structure in the residuals), there are some problems with this approach. It means that the values predicted would always be positive and that the bias would never average zero. Since the expected change in the estimate is calculated by adding

up the predicted residuals for businesses changing response status, this is a problem. An alternative approach would be to use a multiplicative (rather than additive) model. A brief investigation showed this did not solve the problem, however it could be further considered.

An alternative approach is to change the variable being modelled. The difference between the actual and expected reported values multiplied by the unit's weight (weighted residual) was attempted, but the results were similar to what was obtained without the weights. In addition, the population being modelled was defined as 'providers which took at least three calls to respond'. Although it was not implemented, this definition could be adjusted to include more or less businesses in order to get an appropriate fit. A final issue is that follow-up procedures may affect bias through mechanisms other than non-response.

5. CONCLUDING REMARKS

The focus of this paper is to find appropriate statistical models to fit and interpret response rate, cost and potential bias for several survey data collections, using *existing* ABS survey paradata. This study has achieved substantial results from which we have a better understanding of our survey data collection operations and the effectiveness of their follow-up strategies, as well as operational cost structures. Our study also suggests that paradata analysis needs more methodological research to improve modelling performance, and identifies some important information gaps.

From the model fitting diagnostic perspective, it appears that the response propensity model for economic survey works better than the model for household survey. This result is expected because household surveys are more complicated. Household social environment and family situations, for which we do not have any direct information, may play an important role in respondent behaviours.

Although model fitting diagnostics suggest that the household model do not adequately explain the individual unit response, the predicted survey response rates appear reasonable. There is some anecdote evidence (e.g. Carlson *et al.*, 2001) in the literature that the low explanatory power measured by R-squared and its variations is not uncommon.

It should be kept in mind that the 'cost model' is a hybrid approach based on the fragmented cost information and existing empirical cost models. There is no compatible comparison for the model performances between the household and economic surveys. However, it appears that the proposed models perform better than the existing models.

The bias model framework appears to be a good approximation trying to capture the imputation error due to nonresponse by using 'relevant' paradata under an assumption that there is relevant information. The bias measure derived from this method is a relative bias measure against the existing survey estimate. A plausible range of absolute bias accounting for the existing survey estimate bias is proposed but this is not investigated further in this study because the paradata seems to have very little power to explain the variations in the imputation error at unit level. However, it appears that, at an aggregated level, the bias model framework performs well, as indicated by the binning plot in the household survey example. However, the bias model does not fit well for the economic survey example. This phenomena needs to be investigated further in relation to the very skewed distribution of the imputation errors for the economic surveys, and potentially including other sources of paradata which may have an association with the imputation errors.

It appears not possible to estimate the absolute bias using the existing paradata. We may have to use different methods to handle the nonresponse bias problem. One approach could be using micro-simulation and stopping rules to control nonresponse bias, rather than attempting to predict bias in advance for any possible data collection process. Wagner and Raghunathan (2009) illustrate that their stopping rule attempts to use complete interview data as well as covariates (paradata) available on non-responders to determine when the probability that collecting additional data will change the survey estimates is sufficiently low to control bias. Similar approaches have been previously investigated within the ABS with some success, though further research is required.

Paradata is usually a by-product of the data collection processes rather than being designed specifically for analysis of the impact of the variable of interests. Therefore, there is always a big question – how much information is embedded in the paradata for the variables of interest? Although we have experienced different paradata quality problems such as missing, inconsistency etc, we believe that the following two areas are of high priority for modelling response rates and related issues.

Firstly, we applied very traditional methods when undertaking the paradata analyses presented in this paper. They may be effective to test whether particular variables are significant contributors to the variable of interest. However, the underlying model may not provide sufficient fit for an accurate prediction. We do not know if we have reached the limits of what can be achieved with the current paradata. Therefore, it is desirable to apply appropriate data mining techniques to explore the potential hidden structure in the paradata for better predicting power for response rate.

Secondly, there is a big information gap that the current ABS paradata do not contain. For example, Grove and Couper (1998) developed what they call a "conceptual framework for survey cooperation" in their study of nonresponse in household interview surveys. Their framework contains four blocks of variables in two categories which may affect response. The variables in the first category are within the control of a survey operation namely, survey design variables, data collection mode and interviewer characteristics etc.. More importantly, the second category is beyond a survey operation control including the social environment in which survey takes place, the economic conditions, and other characteristic of the household (or respondent), such as household structure and socio-demographic characteristics. Groves and Gouper believe that the interactions among these variables between and within the two categories influence the decision to cooperate or refuse.

Bates *et al.* (2008) studied the Provider Contact History Instrument (CHI) recording interviewers' 'doorstep concerns' about household attitudes and concerns about the U.S. National Health Interview Survey. In the study, she demonstrated that the 'doorstep concerns' have overwhelmingly improved the ability to diagnose and predict the household's future response behaviour (they observed increasing max-rescaled R-squared from 0.06 to 0.53 by including 'doorstep concern' variables). The U.S. Census Bureau is in the process of implementing CHI for all the surveys it undertakes.

Most ABS household paradata available to us and used in our study belong to the first category. There are very few variables belongs to the second category apart from some crude surrogate variables such as demographic information. Obviously, there is an information gap about respondents. Similarly, there is no information available about individual respondents who provide data to ABS business surveys, except some surrogate variables such as 'gold' provider status (the business responded last time without telephone contact) which has been proved a very effective variable to predict future response.

We believe the information about respondents should be expanded by collecting respondent characteristics paradata through interviewer's observations or other means. We are in the process of evaluating the U.S. Census Bureau's Provider Contact History Instrument in the context of the ABS survey environment.

Models require validation and this is particularly important when the models are based on observational data. Experimental procedural changes based on the MPS cost/response model showed that a measure that looks attractive when modelled may not be practical to implement in the field, due to factors not included in the model (in this case, scheduling limitations). Further experimentation would improve understanding of the strengths and limitations of these models.

REFERENCES

Allison, Paul D. (1995) *Survival Analysis using the SAS System: A Practical Guide*, SAS Institute Inc. & John Wiley & Sons, Inc., Cary.

Allison, Paul D. (1999) *Logistic Regression using the SAS System: Theory and Application*, SAS Institute Inc. & John Wiley & Sons, Inc., Cary.

Bate, N. *et al.* (2008) "Privacy Concerns, Too Busy, or Just not Interested: Using Doorstep Concerns to Predict Survey Nonresponse", *Journal of Official Statistics*, 24(4), pp. 591–612.

Beerten, R. (1999) "The Effect of Interviewer and Area Characteristics on Survey Response Rates: An Exploratory Analysis", *Survey Methodology Bulletin*, No. 45, Office for National Statistics.

Biffignandi, S. and Pratesi, M. (2002) "Internet Surveys: The Role of Time in Italian Firms' Response Behaviour", *Research in Official Statistics*, 2, pp. 19–33.

Campanelli, P.; Sturgis, P. and Purdon S. (1997) *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*, Social and Community Planning Research, London.

Campbell, L.N.; Dimitropoulos, L.L.; Carpenter, L. and Brown, G.G. (2004) *The Costs and Benefits of Improving Response Rates of the CAHPS Medicare Fee for Service Survey*, Paper presented at the annual meeting of the American Association for Public Opinion Research, Pointe Hilton Tapatio Cliffs, Phoenix, Arizona.

Carlson, B.L. and Williams, S. (2001) "A Comparison of Two Methods to Adjust Weights for Non-response: Propensity Modelling and Weighting Class Adjustments", *Proceedings of the Annual Meeting of the American Statistical Association*, August 5–9, 2001.

Groves, R.M.; Cialdini, R.B. and Couper, M.P. (1992) "Understanding the Decision to Participate in a Survey", *Public Opinion Quarterly*, 56(4), pp. 475–495.

Nagelkerke, N. (1991) "A Note on a General Definition of the Coefficient of Determination", *Biometrika*, 78(3), pp. 691–692.

Ragavan, A.J. (2008) *How to use SAS to fit Multiple Logistic Regression Models*, Department of Mathematics and Statistics, University of Nevada.

Wagner, J. and Raghunathan, T.E. (2009) "A New Stop Rule for Surveys", *Statistics in Medicine*.

APPENDIX

A.1 Estimated coefficients and odds ratios – MPS response models
(a) Standard workload, Face-to-face only

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–4.4953	0.1398	<0.0001		
State (base = New South Wales)					
Victoria	–0.0114	0.0228	0.6174	1.098	
Queensland	0.1758	0.0234	<0.0001	1.324	
South Australia	–0.0462	0.0271	0.0877	1.061	
Western Australia	–0.0037	0.0265	0.8890	1.107	
Tasmania	0.2009	0.0349	<0.0001	1.358	
Northern Territory	–0.0690	0.0523	0.1872	1.037	*
Australian Capital Territory	–0.1414	0.0519	0.0065	0.964	*
Region (base = Metropolitan)					
Ex-Metropolitan	–0.1877	0.0102	<0.0001	0.687	
Whether first month in sample (base = Yes)					
No	–0.3950	0.0100	<0.0001	0.454	
Call attempt number (base = call 1)					
Call 2	0.7324	0.0816	<0.0001	0.769	
Call 3	0.4575	0.0832	<0.0001	0.584	
Call 4	0.1266	0.0868	0.1446	0.419	
Call 5	–0.1753	0.0932	0.0600	0.310	
Call 6	–0.3657	0.1092	0.0008	0.256	
Call 7	–0.3949	0.1344	0.0033	0.249	
Call 8	–0.4117	0.1825	0.0241	0.245	
Call 9	–0.7979	0.2886	0.0057	0.166	
Call 10	–0.5638	0.3526	0.1098	0.210	
Call 11	0.2366	0.3928	0.5469	0.468	*
Call 12	0.1608	0.5862	0.7839	0.434	
Workload size (i.e. Number of households)	0.0073	0.0004	<0.0001	1.007	
Age of interviewer (years)	0.0068	0.0012	<0.0001	1.007	
Interviewer performance (i.e. Prev. response rate)	0.0207	0.0011	<0.0001	1.021	
Day of call attempt (base = Monday)					
Tuesday	–0.1332	0.0214	<0.0001	1.042	*
Wednesday	–0.0776	0.0219	0.0004	1.102	
Thursday	–0.0384	0.0251	0.1255	1.146	
Friday	–0.0243	0.0298	0.4152	1.162	
Saturday	0.1966	0.0342	<0.0001	1.450	
Sunday	0.2516	0.0566	<0.0001	1.532	
Time of call attempt (base = Morning)					
Afternoon	–0.0700	0.0129	<0.0001	0.995	*
Evening	0.1346	0.0160	<0.0001	1.221	

(b) Standard workload, Telephone only

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–2.1383	0.1031	<0.0001		
State (base = New South Wales)					
Victoria	0.0124	0.0107	0.2473	1.000	*
Queensland	0.0070	0.0112	0.5313	0.995	*
South Australia	0.1445	0.0130	<0.0001	1.142	
Western Australia	–0.0282	0.0122	0.0204	0.961	
Tasmania	0.1248	0.0170	<0.0001	1.119	
Northern Territory	–0.2414	0.0311	<0.0001	0.776	
Australian Capital Territory	–0.0312	0.0197	0.1143	0.958	*
Region (base = Metropolitan)					
Ex-Metropolitan	0.0308	0.0049	<0.0001	1.064	
Whether First Month in Sample (base = Yes)					
No	–0.2225	0.0428	<0.0001	0.641	
Call attempt number (base = call 1)					
Call 2	1.0568	0.0765	<0.0001	0.859	
Call 3	0.8621	0.0768	<0.0001	0.707	
Call 4	0.6521	0.0773	<0.0001	0.573	
Call 5	0.5205	0.0782	<0.0001	0.503	
Call 6	0.4114	0.0794	<0.0001	0.451	
Call 7	0.2586	0.0813	0.0015	0.387	
Call 8	0.1312	0.0840	0.1183	0.341	
Call 9	0.0430	0.0888	0.6283	0.312	
Call 10	–0.0446	0.0965	0.6436	0.286	
Call 11	–0.1126	0.1074	0.2949	0.267	
Call 12	–0.2222	0.1241	0.0733	0.239	
Call 13	–0.2875	0.1441	0.0461	0.224	
Call 14	–0.5370	0.1817	0.0031	0.175	
Call 15	–0.4106	0.2083	0.0486	0.198	
Call 16	–0.9556	0.3090	0.0020	0.115	
Call 17	–0.2713	0.2790	0.3308	0.228	
Call 18	–0.8136	0.4160	0.0505	0.132	
Call 19	–0.5432	0.4618	0.2395	0.173	
Call 20	0.2302	0.4203	0.5840	0.376	
Call 21	–0.7209	0.7183	0.3156	0.145	
Call 22	–1.0992	0.9896	0.2666	0.099	
Call 23	0.6439	0.6651	0.3330	0.569	*
Age of Interviewer (years	0.0024	0.0006	<0.0001	1.002	
Interviewer performance (i.e. Prev. response rate)	0.0083	0.0005	<0.0001	1.008	
Day of call attempt (base = Monday)					
Tuesday	0.0097	0.0138	0.4843	0.832	
Wednesday	–0.0762	0.0144	<0.0001	0.763	
Thursday	–0.1153	0.0158	<0.0001	0.734	
Friday	–0.1310	0.0191	<0.0001	0.722	
Saturday	0.0917	0.0260	0.0004	0.903	
Sunday	0.0271	0.0535	0.6126	0.846	
Time of call attempt (base = Morning)					
Afternoon	–0.0613	0.0061	<0.0001	1.065	
Evening	0.1861	0.0064	<0.0001	1.365	

(c) Standard workload, Face-to-face and telephone

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–2.9847	0.2441	<0.0001		
State (base = New South Wales)					
Victoria	–0.0728	0.0546	0.1822	0.909	*
Queensland	0.1483	0.0548	0.0068	1.134	*
South Australia	0.1583	0.0619	0.0105	1.145	*
Western Australia	–0.0715	0.0593	0.2279	0.910	*
Tasmania	–0.1347	0.0961	0.1611	0.854	*
Northern Territory	–0.3502	0.1457	0.0162	0.689	
Australian Capital Territory	0.3001	0.0975	0.0021	1.320	
Region (base = Metropolitan)					
Ex-Metropolitan	0.1857	0.0248	<0.0001	1.450	
Whether first month in sample (base = Yes)					
No	–0.6591	0.0234	<0.0001	0.268	
Call attempt number (base = call 1)					
Call 2	0.4789	0.0901	<0.0001	9.007	
Call 3	0.6318	0.0886	<0.0001	10.495	
Call 4	0.5234	0.0913	<0.0001	9.416	
Call 5	0.4111	0.0956	<0.0001	8.416	
Call 6	0.1341	0.1071	0.2106	6.380	
Call 7	0.1008	0.1156	0.3832	6.171	
Call 8	0.0886	0.1247	0.4772	6.096	
Call 9	–0.0403	0.1457	0.7822	5.359	
Call 10	0.2588	0.1478	0.0798	7.227	
Call 11	–0.2522	0.2041	0.2166	4.335	
Call 12	0.0470	0.2055	0.8191	5.848	
Call 13	0.3649	0.2072	0.0782	8.036	
Call 14	–0.0323	0.2812	0.9085	5.402	
Call 15	–0.4765	0.4017	0.2356	3.464	
Call 16	0.0250	0.3772	0.9472	5.720	
Call 17	–0.3184	0.4914	0.5170	4.058	
Call 18	–0.3400	0.5657	0.5478	3.971	
Call 19	0.3026	0.4978	0.5433	7.551	
Call 20	–0.1885	0.6910	0.7850	4.621	
Interviewer performance (i.e. Prev. response rate)	0.0073	0.0025	0.0038	1.007	
Day of call attempt (base = Monday)					
Tuesday	–0.2119	0.0570	0.0002	0.827	
Wednesday	–0.0883	0.0502	0.0790	0.936	*
Thursday	–0.0392	0.0494	0.4274	0.983	*
Friday	0.0373	0.0540	0.4903	1.061	*
Saturday	–0.0066	0.0634	0.9172	1.016	*
Sunday	0.3307	0.1264	0.0089	1.423	

(d) Non-standard workload, Face-to-face only

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–2.1698	0.2735	<0.0001		
Region (base = Metropolitan)					
Ex-Metropolitan	0.1548	0.0363	<0.0001	1.363	
Whether first month in sample (base = Yes)					
No	–0.3454	0.0317	<0.0001	0.501	
Call attempt number (base = call 1)					
Call 2	0.9137	0.1111	<0.0001	0.662	
Call 3	0.9717	0.1106	<0.0001	0.701	
Call 4	0.7779	0.1161	<0.0001	0.578	
Call 5	0.4142	0.1215	0.0007	0.402	
Call 6	0.4203	0.1074	<0.0001	0.404	
Call 7	0.0988	0.1142	0.3871	0.293	
Call 8	–0.1730	0.1285	0.1783	0.223	
Call 9	–0.3966	0.1494	0.0080	0.179	
Call 10	–0.4659	0.1707	0.0063	0.167	
Call 11	–0.3315	0.1886	0.0787	0.191	
Call 12	–0.5681	0.2416	0.0187	0.150	
Call 13	–0.6005	0.2912	0.0392	0.146	
Call 14	–0.6837	0.3762	0.0691	0.134	
Call 15	–1.4884	0.6778	0.0281	0.060	
Call 16	–0.2154	0.5028	0.6684	0.214	
Workload size (i.e. Number of households)	–0.0075	0.0014	<0.0001	0.993	*
Interviewer experience (years)	–0.0144	0.0042	0.0006	0.986	
Interviewer performance (i.e. Prev. response rate)	0.0093	0.0028	0.0011	1.009	
Day of call attempt (base = Monday)					
Tuesday	–0.0296	0.0689	0.6674	0.804	
Wednesday	–0.1341	0.0695	0.0536	0.725	
Thursday	–0.0859	0.0722	0.2341	0.760	
Friday	–0.2592	0.0800	0.0012	0.639	
Saturday	0.0164	0.0832	0.8439	0.842	*
Sunday	0.3044	0.1648	0.0647	1.123	*
Time of the day (base = Morning)					
Afternoon	–0.0572	0.0415	0.1682	1.037	*
Evening	0.1506	0.0449	0.0008	1.276	

(e) Non-standard workload, Telephone only

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–1.9235	0.0782	<0.0001		
State (base = New South Wales)					
Victoria	0.0868	0.0493	0.0786	0.997	*
Queensland	–0.0054	0.0536	0.9197	0.909	*
South Australia	0.1114	0.0613	0.0692	1.022	*
Western Australia	0.0801	0.0716	0.2634	0.990	*
Tasmania	0.3566	0.0770	<0.0001	1.306	
Northern Territory	–0.5147	0.1870	0.0059	0.546	
Australian Capital Territory	–0.2047	0.0818	0.0124	0.745	
Call attempt number (base = call 1)					
Call 2	1.1643	0.0939	<0.0001	0.693	
Call 3	0.8899	0.1023	<0.0001	0.527	
Call 4	1.0394	0.1057	<0.0001	0.612	
Call 5	0.9746	0.1099	<0.0001	0.573	
Call 6	0.7226	0.1161	<0.0001	0.446	
Call 7	0.6016	0.1162	<0.0001	0.395	
Call 8	0.6084	0.1143	<0.0001	0.398	
Call 9	0.5726	0.1060	<0.0001	0.384	
Call 10	0.5050	0.1044	<0.0001	0.359	
Call 11	0.3977	0.1074	0.0002	0.322	
Call 12	0.4045	0.1101	0.0002	0.324	
Call 13	0.0551	0.1253	0.6604	0.229	
Call 14	0.2103	0.1273	0.0984	0.267	
Call 15	0.2591	0.1330	0.0513	0.280	
Call 16	–0.1192	0.1582	0.4511	0.192	
Call 17	–0.3125	0.1820	0.0860	0.158	
Call 18	0.1339	0.1703	0.4317	0.247	
Call 19	–0.4832	0.2323	0.0375	0.133	
Call 20	–0.0176	0.2126	0.9340	0.213	
Call 21	–0.6541	0.2997	0.0291	0.113	
Call 22	–0.9609	0.3600	0.0076	0.083	
Call 23	–0.5118	0.3289	0.1197	0.130	
Call 24	–0.5008	0.3666	0.1719	0.131	
Call 25	–0.7843	0.4552	0.0849	0.099	
Call 26	–1.1216	0.5786	0.0526	0.070	
Call 27	–1.3645	0.7021	0.0520	0.055	
Call 28	–0.1832	0.4674	0.6951	0.180	
Call 29	–0.5160	0.5915	0.3830	0.129	
Call 30	–1.3562	0.9920	0.1716	0.056	
Call 31	–1.1841	0.9976	0.2353	0.066	
Day of call attempt (base = Monday)					
Tuesday	0.1025	0.0499	0.0400	0.914	*
Wednesday	–0.0163	0.0516	0.7524	0.812	
Thursday	–0.0264	0.0550	0.6316	0.804	
Friday	–0.0334	0.0604	0.5802	0.798	
Saturday	–0.0160	0.0769	0.8352	0.812	
Sunday	–0.2028	0.1717	0.2375	0.674	*
Time of the day (base = Morning)					
Afternoon	–0.0811	0.0283	0.0041	0.955	*
Evening	0.1163	0.0288	<0.0001	1.164	

(f) Non-standard workload, Face-to-face and telephone

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–2.5937	0.2620	<0.0001		
State (base = New South Wales)					
Victoria	–0.0862	0.0700	0.2186	1.015	*
Queensland	0.0139	0.0909	0.8782	1.122	*
South Australia	–0.2663	0.0872	0.0023	0.847	*
Western Australia	–0.0450	0.0846	0.5952	1.057	*
Tasmania	0.3084	0.1173	0.0085	1.506	
Northern Territory	0.1663	0.2760	0.5468	1.306	*
Australian Capital Territory	0.0095	0.1445	0.9475	1.117	*
Region (base = Metropolitan)					
Ex-Metropolitan	0.0932	0.0376	0.0131	1.205	
Whether first month in sample (base = Yes)					
No	–0.4008	0.0350	<0.0001	0.449	
Call attempt number (base = call 1)					
Call 2	0.4780	0.1681	0.0045	2.472	
Call 3	0.5159	0.1548	0.0009	2.568	
Call 4	0.6780	0.1393	<0.0001	3.020	
Call 5	0.8107	0.1328	<0.0001	3.448	
Call 6	0.4992	0.1361	0.0002	2.525	
Call 7	0.6894	0.1256	<0.0001	3.054	
Call 8	0.5351	0.1305	<0.0001	2.618	
Call 9	0.3276	0.1320	0.0131	2.127	
Call 10	0.1833	0.1361	0.1780	1.841	
Call 11	0.2456	0.1333	0.0655	1.960	
Call 12	0.0875	0.1427	0.5394	1.673	*
Call 13	0.1370	0.1437	0.3402	1.758	*
Call 14	0.2191	0.1440	0.1281	1.908	
Call 15	0.1165	0.1547	0.4515	1.722	*
Call 16	–0.0399	0.1711	0.8156	1.473	*
Call 17	–0.0220	0.1789	0.9021	1.500	*
Call 18	–0.0707	0.1929	0.7139	1.428	*
Call 19	–0.5086	0.2439	0.0370	0.922	*
Call 20	0.1685	0.2003	0.4001	1.814	*
Call 21	–0.5193	0.2838	0.0673	0.912	*
Call 22	0.1631	0.2268	0.4721	1.804	*
Call 23	0.0540	0.2544	0.8320	1.618	*
Call 24	–0.0827	0.2874	0.7736	1.411	*
Call 25	–0.0088	0.2991	0.9766	1.520	*
Call 26	–0.2401	0.3607	0.5057	1.206	*
Call 27	0.0861	0.3446	0.8028	1.671	*
Call 28	–0.5199	0.5018	0.3001	0.911	*
Call 29	–0.6478	0.5764	0.2611	0.802	*
Call 30	0.0223	0.4572	0.9612	1.567	*
Call 31	–0.3308	0.5813	0.5693	1.101	*
Call 32	–1.2617	0.9842	0.1998	0.434	*
Call 33	–0.0279	0.5866	0.9621	1.491	*
Call 34	–1.0226	0.9884	0.3009	0.551	*
Call 35	–0.2867	0.7120	0.6872	1.151	*
Interviewer performance (i.e. Prev. response rate)	0.0058	0.0028	0.0407	1.006	*
Time of the day (base = Morning)					
Afternoon	0.0834	0.0401	0.0375	1.226	
Evening	0.0368	0.0421	0.3823	1.170	

A.2 Estimated coefficients and odds ratios – SET response model

Variable	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–2.4143	0.1307	<0.0001		
State (base = New South Wales)					
Victoria	–0.1704	0.0296	<0.0001	0.928	*
Queensland	0.0195	0.0314	0.5350	1.122	
South Australia	–0.0809	0.0313	0.0099	1.015	*
Western Australia	0.0099	0.0301	0.7424	1.112	
Tasmania	0.2248	0.0314	<0.0001	1.378	
Northern Territory	–0.0261	0.0374	0.4859	1.072	*
Australian Capital Territory	0.1190	0.0354	0.0008	1.240	
Region (base = Metropolitan)					
Ex-Metropolitan	0.0793	0.0134	<0.0001	1.172	
Day of the week (base = Monday)					
Tuesday	0.0487	0.0270	0.0712	0.964	*
Wednesday	–0.0102	0.0266	0.7007	0.908	
Thursday	–0.1286	0.0282	<0.0001	0.807	
Friday	–0.1383	0.0321	<0.0001	0.799	
Saturday	0.0130	0.0335	0.6986	0.930	*
Sunday	0.1296	0.0614	0.0349	1.045	*
Time of the day (base = Morning)					
Afternoon	0.1564	0.0160	<0.0001	1.996	
Evening	0.3782	0.0198	<0.0001	2.491	
Call attempt number (base = Call 1)					
Call 2	0.8013	0.0833	<0.0001	2.158	
Call 3	0.9722	0.0841	<0.0001	2.560	
Call 4	0.9192	0.0858	<0.0001	2.428	
Call 5	0.7787	0.0886	<0.0001	2.110	
Call 6	0.7520	0.0917	<0.0001	2.054	
Call 7	0.6600	0.0963	<0.0001	1.874	
Call 8	0.4280	0.1045	<0.0001	1.486	
Call 9	0.5286	0.1085	<0.0001	1.643	
Call 10	0.3653	0.1206	0.0025	1.396	
Call 11	0.3185	0.1310	0.0151	1.332	
Call 12	0.1459	0.1509	0.3337	1.121	*
Call 13	0.1212	0.1649	0.4622	1.093	*
Call 14	0.1568	0.1797	0.3828	1.133	*
Call 15	–0.0606	0.2120	0.7751	0.912	*
Call 16	0.3402	0.2039	0.0952	1.361	*
Call 17	–0.2325	0.2731	0.3946	0.768	*
Call 18	–0.0554	0.2843	0.8456	0.916	*
Call 19	–1.1029	0.4974	0.0266	0.321	
Call 20	–1.1988	0.5708	0.0357	0.292	
Call 21	–0.1992	0.3901	0.6097	0.794	*
Call 22	–0.1675	0.4209	0.6907	0.819	*
Call 23	–0.6520	0.5783	0.2595	0.505	*
Call 24	–0.5606	0.5808	0.3344	0.553	*
Call 25	–1.4414	0.9791	0.1410	0.229	*
Call 26	–1.3092	0.9816	0.1823	0.262	*
Call 27	–0.3399	0.7139	0.6341	0.689	*
Enumerated entirely in non standard workload (base = 1)					
Had at least one call in standard workload	0.1417	0.0574	0.0136	1.328	
Number of days since the previous call	0.0101	0.0016	<0.0001	1.010	
Age of the interviewer	0.0037	0.0017	0.0279	1.004	*
Number of years experience of interviewer	–0.0051	0.0017	0.0032	0.995	

A.3 Estimated coefficients and odds ratios – SDAC response model

Variables	*Coefficient*	*Std error*	*Pr > ChiSq*	*Odds ratio*	
Intercept	–1.9068	0.0947	<0.0001		
State (base = New South Wales)					
Victoria	–0.1488	0.0194	<0.0001	0.929	*
Queensland	0.2024	0.0241	<0.0001	1.320	
South Australia	–0.1130	0.0226	<0.0001	0.963	*
Western Australia	0.0148	0.0225	0.5100	1.094	
Tasmania	0.3443	0.0299	<0.0001	1.521	
Northern Territory	–0.2913	0.0411	<0.0001	0.805	*
Australian Capital Territory	0.0665	0.0344	0.0534	1.152	
Day of the week (base = Monday)					
Tuesday	0.0798	0.0188	<0.0001	0.934	*
Wednesday	–0.0085	0.0188	0.6530	0.855	
Thursday	–0.0877	0.0199	<0.0001	0.790	
Friday	–0.0049	0.0217	0.8202	0.858	
Saturday	–0.0894	0.0230	0.0001	0.789	
Sunday	–0.0373	0.0421	0.3760	0.831	
Time of the day (base = Morning)					
Afternoon	0.1960	0.0116	<0.0001	1.709	
Evening	0.1442	0.0156	<0.0001	1.623	
Call attempt number (base = Call 1)					
Call 2	0.7914	0.0551	<0.0001	1.586	
Call 3	0.8394	0.0559	<0.0001	1.664	
Call 4	0.6936	0.0577	<0.0001	1.438	
Call 5	0.4869	0.0605	<0.0001	1.169	
Call 6	0.5141	0.0632	<0.0001	1.202	
Call 7	0.4026	0.0678	<0.0001	1.075	
Call 8	0.3178	0.0736	<0.0001	0.988	
Call 9	0.3170	0.0796	<0.0001	0.987	
Call 10	0.1235	0.0910	0.1751	0.813	*
Call 11	0.0045	0.1043	0.9659	0.722	*
Call 12	–0.1773	0.1213	0.1439	0.602	*
Call 13	–0.1725	0.1333	0.1958	0.605	*
Call 14	–0.0124	0.1433	0.9312	0.710	*
Call 15	–0.2643	0.1743	0.1294	0.552	*
Call 16	–0.2162	0.1919	0.2599	0.579	*
Call 17	–0.0494	0.2001	0.8049	0.684	*
Call 18	–0.5091	0.2662	0.0558	0.432	*
Call 19	–0.4569	0.2764	0.0984	0.455	*
Call 20	–0.4276	0.3103	0.1682	0.469	*
Call 21	–0.1826	0.3137	0.5606	0.599	*
Call 22	–0.8527	0.4480	0.0570	0.306	*
Call 23	–0.9411	0.4988	0.0592	0.280	*
Call 24	–0.2675	0.4214	0.5256	0.550	*
Call 25	–0.0514	0.4270	0.9043	0.683	*
Call 26	–0.2055	0.5167	0.6909	0.585	*
Call 27	–0.0349	0.5237	0.9469	0.694	*
Size of workload	–0.0047	0.0015	0.0012	0.995	
Enumerated entirely in a non standard workload (base = 1)					
Had at least one call in standard workload	0.1107	0.0260	<0.0001	1.248	
Number of days since start call	0.0004	0.0001	0.0038	1.000	
Number of days since the previous call	–0.0040	0.0009	<0.0001	0.996	
Age of the interviewer	0.0031	0.0011	0.0067	1.003	*
Number of years experience of interviewer	–0.0041	0.0012	0.0004	0.996	

A.4 Economic survey response model coefficients

Parameter	*Description*	*DF*	*Estimate*	*S.E.*	*Wald* χ^2	*Pr* > χ^2	$(e^{\hat{\beta}}-1)$[19]
Intercept		1	–2.6309	0.0182	20937.7	<.0001	
week3	Weeks since last written contact = 1	1	0.0568	0.0115	24.4	<.0001	5.8%
week3	Weeks since last written contact ≥ 2	1	–0.0692	0.0194	12.7	0.0004	–6.7%
fcallno_per	log (Number of calls in the period +1)	1	0.4754	0.0179	1061.0	<.0001	60.9%
fcallno_prev	log (Number of calls prior to the period +1)	1	0.4109	0.0146	787.3	<.0001	50.8%
signifcat	Crucial unit	1	0.1959	0.0201	95.1	<.0001	21.6%
signifcat	New-on unit	1	0.2995	0.0163	336.4	<.0001	34.9%
FGOLD	Gold provider business	1	1.3023	0.0127	10489.4	<.0001	267.8%
rem0	Last contact = Despatch	1	0.1406	0.0207	46.1	<.0001	15.1%
rem2	Last contact = Reminder 2	1	0.1810	0.0128	201.4	<.0001	19.8%
rem3	Last contact = Reminder 3	1	0.5074	0.0141	1291.3	<.0001	66.1%
rem4	Last contact = Reminder 4	1	1.1989	0.0142	7091.8	<.0001	231.7%
size1	0–4 employees	1	0.0745	0.0126	35.1	<.0001	7.7%
size2	5–19 employees	1	0.1166	0.0166	49.1	<.0001	12.4%
size3	20–49 employees	1	–0.0100	0.0169	0.4	0.5528	–1.0%
size4	50–199 employees	1	0.0588	0.0145	16.5	<.0001	6.1%
state2	Victoria	1	–0.0820	0.0113	53.1	<.0001	–7.9%
state3	Queensland	1	–0.0547	0.0124	19.4	<.0001	–5.3%
state4	South Australia	1	0.0512	0.0168	9.3	0.0023	5.3%
state5	Western Australia	1	–0.0496	0.0146	11.5	0.0007	–4.8%
state6	Tasmania	1	0.1417	0.0190	55.7	<.0001	15.2%
state7	Northern Territory	1	–0.0181	0.0246	0.5	0.4625	–1.8%
state8	Aust. Capital Territory	1	0.1246	0.0234	28.3	<.0001	13.3%
FGOLD × signifcat	Interaction between FGOLD and signifcat= crucial	1	0.5796	0.0244	564.9	<.0001	78.5%
rem0 × signifcat	Interaction between rem0 and signifcat = crucial	1	–0.5094	0.0279	333.5	<.0001	–39.9%
rem0 × signifcat	Interaction between rem0 and signifcat = new-on	1	–0.5160	0.0363	202.1	<.0001	–40.3%
FGOLD × rem0	Interaction between FGOLD and rem0	1	–1.2959	0.0222	3418.6	<.0001	–72.6%
rem0×week3	Interaction between rem0 and weeks since contact =1	1	–1.5774	0.0307	2632.7	<.0001	–79.4%
rem0×week3	Interaction between rem0 and weeks since contact ≥ 2	1	–0.5220	0.0279	348.9	<.0001	–40.7%

[19] $\exp(\hat{\beta})-1$ gives the impact of a one unit change on the hazard of response in the related proportional hazard model. For example, a one unit increase in fcallno_prev results in a 50.8% increase in the hazard of response.

A.5 Economic survey response model – Check for multicollinearity

Variable	*DF*	*Estimate*	*S.E.*	*t-value*	*Pr > |t|*	*Tolerance*	*Variance inflation*
Intercept	1	0.1128	0.00185	60.88	<.0001	.	0
fcallno_per	1	0.0382	0.00248	15.43	<.0001	0.8395	1.1911
fcallno_prev	1	0.0064	0.00193	3.32	0.0009	0.7218	1.3854
signifcat0	1	0.0360	0.00149	24.15	<.0001	0.7785	1.2845
signifcat2	1	0.0208	0.00151	13.76	<.0001	0.8840	1.1312
FGOLD	1	0.0981	0.00102	96.42	<.0001	0.8154	1.2264
rem0	1	–0.1062	0.00132	–80.68	<.0001	0.4682	2.1358
rem2	1	0.0161	0.00161	10.00	<.0001	0.5779	1.7304
rem3	1	0.0807	0.00197	40.97	<.0001	0.6195	1.6141
rem4	1	0.2688	0.00233	115.47	<.0001	0.7249	1.3795
size1	1	0.0087	0.00142	6.16	<.0001	0.4074	2.4544
size2	1	0.0132	0.00183	7.23	<.0001	0.6089	1.6423
size3	1	0.0007	0.00183	0.41	0.6829	0.6107	1.6375
size4	1	0.0079	0.00161	4.87	<.0001	0.5726	1.7466
state2	1	–0.0086	0.00123	–6.96	<.0001	0.7733	1.2931
state3	1	–0.0053	0.00135	–3.89	<.0001	0.8000	1.2500
state4	1	0.0058	0.00189	3.05	0.0023	0.8900	1.1236
state5	1	–0.0052	0.00159	–3.24	0.0012	0.8504	1.1759
state6	1	0.0129	0.00224	5.75	<.0001	0.9157	1.0920
state7	1	–0.0038	0.00274	–1.39	0.1655	0.9292	1.0762
state8	1	0.0118	0.00273	4.33	<.0001	0.9315	1.0736
week3	1	–0.0136	0.00061	–22.15	<.0001	0.8649	1.1563

As a rule of thumb, a Variance Inflation Factor over ten in a linear regression, or 2.5 in a logistic regression may indicate a multicollinearity problem. The highest VIF observed is 2.45.

A.6 Time and distance coefficients – MPS face-to-face cost model

Month in sample	*Region*	*State*	*km per call*	*km per block-visit*	*km per response*	*mins per call*	*mins per block-visit*	*mins per interview*	*Interview time adj.*
FMIS	Ex-met	ACT	2.50	11.26	–4.01	9.71	13.13	0.00	1.01
FMIS	Ex-met	NSW	4.96	9.82	2.01	12.77	13.69	0.00	1.01
FMIS	Ex-met	NT	1.74	5.67	0.00	6.13	12.94	6.35	1.01
FMIS	Ex-met	Qld	1.70	10.74	7.13	8.60	15.31	8.70	1.01
FMIS	Ex-met	SA	4.62	17.75	0.00	10.79	26.39	0.00	1.01
FMIS	Ex-met	Tas	4.27	2.78	0.00	11.01	8.69	0.00	1.01
FMIS	Ex-met	Vic	3.99	11.16	3.32	8.95	21.18	5.28	1.01
FMIS	Ex-met	WA	4.74	16.17	0.00	12.02	13.21	0.00	1.01
FMIS	Met	ACT	0.40	7.57	0.82	5.48	14.15	5.11	1.01
FMIS	Met	NSW	0.00	8.02	2.64	6.04	18.87	5.63	1.01
FMIS	Met	NT	0.59	6.46	1.56	5.01	12.01	9.20	1.01
FMIS	Met	Qld	1.13	7.62	2.63	7.65	21.00	5.39	1.01
FMIS	Met	SA	0.68	6.35	0.00	7.85	12.81	–2.03	1.01
FMIS	Met	Tas	5.14	0.00	0.00	9.77	13.01	0.00	1.01
FMIS	Met	Vic	0.32	7.43	1.26	6.74	17.93	2.59	1.01
FMIS	Met	WA	1.04	7.09	0.85	8.52	13.06	3.62	1.01
NFMIS	Ex-met	ACT	0.00	8.54	0.00	11.36	8.66	0.00	1.01
NFMIS	Ex-met	NSW	6.56	12.75	3.37	13.49	15.23	0.00	1.01
NFMIS	Ex-met	NT	3.80	0.00	0.00	7.42	4.50	–5.69	1.01
NFMIS	Ex-met	Qld	4.16	7.32	9.00	9.57	11.71	10.24	1.01
NFMIS	Ex-met	SA	8.10	5.78	11.68	9.76	12.38	14.26	1.01
NFMIS	Ex-met	Tas	9.62	0.00	–2.34	12.13	6.54	4.24	1.01
NFMIS	Ex-met	Vic	0.00	20.95	14.39	1.97	25.19	23.69	1.01
NFMIS	Ex-met	WA	5.14	17.17	4.34	9.24	21.46	6.83	1.01
NFMIS	Met	ACT	0.00	6.83	0.00	5.92	11.64	0.00	1.01
NFMIS	Met	NSW	1.14	5.51	1.75	3.01	18.42	–2.82	1.01
NFMIS	Met	NT	0.00	3.83	5.71	1.72	12.15	8.87	1.01
NFMIS	Met	Qld	2.03	7.52	2.57	4.59	18.14	8.45	1.01
NFMIS	Met	SA	1.24	5.45	0.00	5.96	13.92	–7.79	1.01
NFMIS	Met	Tas	9.44	0.00	0.00	16.07	4.51	0.00	1.01
NFMIS	Met	Vic	–0.49	5.84	2.06	0.96	18.82	–2.69	1.01
NFMIS	Met	WA	2.28	4.11	3.95	6.61	12.95	5.74	1.01

Note: The 'mins per interview' parameter is in addition to the expected payment for interviewing itself.
The total time predicted for a trip is equal to:
(mins per call) × (number of calls) + (mins per block-visit) × (number of block-visits) + (mins per interview + assessed interview time × interview time adjustment) × (number of interviews).

Negative values are likely due to collinearity effects and/or small sample (especially ACT ex-metropolitan). Variables that weren't significant at p=0.05 are removed from the model, with their effects being absorbed into remaining coefficients.

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au