OPTIMIZING CATI WORKLOAD TO MINIMIZE DATA COLLECTION COST

Choudhry, G.H.[1], Hidiroglou, M.A., Laflamme, F.
[1]Statistical Research and Innovation Division, Statistics Canada,
16th Floor, R. H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A0T6

Abstract

One of the main increasing challenges for Statistics Canada is to collect cost-effective data while maintaining a high level of quality. Paradata research has been useful in improving the current data collection processes and practices. The research carried out with paradata suggested that collection resources are currently not always optimally allocated with respect to the assigned workload and the corresponding expected productivity for computer-assisted telephone interview (CATI) surveys.

In this paper, models to predict the probability that a telephone call would result in a completed questionnaire as a function of time of day, and resources spent to date were developed. The parameters estimated from these models are used as input to optimize call scheduling. The potential cost savings of this approach is illustrated by applying the theory to a large scale household survey.

Key Words: Linear and logistic regression models; Optimum CATI schedule; Non-linear programming; Cost-savings

1. Introduction

Statistics Canada is facing increasing challenges in maintaining cost-effective data collection and obtaining high-quality outputs to meet the evolving demands for timely and accurate data from a wide range of users. Since 2006, Statistics Canada has studied paradata to evaluate its current data collection processes and practices (Laflamme, 2008a). The studies carried out so far have identified a number of options to improve the way the agency conducts and manages its surveys with respect to CATI surveys (Laflamme, 2008b).

Some of these studies were carried out to obtain a better understanding of the relationships between interviewing efforts and the expected workload during the data collection cycle. These investigations suggested that the interviewer staffing levels were not always well aligned with the workload and the expected productivity. For example, the observation that in-progress units are likely to be called more often during the second half of the collection period suggests that interviewer staffing levels are greater than the sample workload in the first-half of the collection period. It has also been observed for CATI surveys that the proportion of completed questionnaires decreases rapidly over time for given number of calls (Laflamme, 2009). Data collection managers need to improve interviewer staffing management and planning tools to reduce some of the tension between collection productivity and costs (Couper et al., 1998) while maintaining high level of data quality. Operational constraints involving the interviewing staff have also increased collection costs and limit the capacity to optimize the interviewers' schedules. For example, rules concerning notice of shift changes for a unionized interviewing workforce need to be factored into any action plan. In addition, the overall regional office (RO) capacity by time slice (i.e., day and evening shifts) also needs to be considered.

Greenberg and Stokes (1990) presented an optimal or priority system for call scheduling. In another approach, Stokes and Greenberg (1990) used a logistic regression model to predict the probability of success of a call at a particular time and used this model to develop a ranking system for call-backs. Brick, Allen, Cunningham, and Maklan (1996) also used logistic regression models to examine the relationship between the procedures used to assign numbers and the outcomes of the calls. They obtained logistic models where the probability of success was defined as: contacting a household, completing the interview, and refusing an interview. We develop regression models to predict the probability of a productive call (i.e. interview will be completed) and use the predicted probabilities to determine optimum number of calls by time slice.

The methodology and results presented in this paper only represents the first phase in the development of an optimized interviewer scheduling tool for a single CATI survey. It does not account for interviewer operational constraints such as their availability for each day of the week, sick leave and vacation, assigned hours per week as well as their work shift preferences. Furthermore, the interviewer workload is not optimized over several concurrent surveys. We address the problem of determining the optimum number of calls by time slice (morning, afternoon, early and late evening shifts) given that the only constraint is that the target response rate is attained.

Methodology

2.1 Data Collection

Data collection for CATI surveys at Statistics Canada is conducted from six call centres managed by Regional Offices (ROs). CATI collection procedures for a given survey can vary by site depending on the mix of concurrent and large scale surveys in collection, workload and availability of interviewers. However, there is paradata standardization across the regional offices because CATI survey data are collected using Blaise. This software automatically collects paradata, and stores it in the Blaise Transaction History (BTH) file. A BTH record is automatically created each time a case is opened, either for data collection or other purposes. It contains detailed information about each call made to contact a sampled unit during the data collection period. This includes the survey and unit identification, the date, the time the case was open, the identification of the interviewer who worked on it, the results of the call, as well as additional relevant information.

2.2 Problem definition

The data collection period for a given survey takes place over D continuous days, and each day is split into T time slices. We assume that these time slices correspond to interviewer shifts. Time slices are fixed periods within a day during which CATI interviewers call the sampled units (telephone numbers). An interviewer shift consists of one or more time slices. An interviewer shift represents the number of hours that an interviewer is scheduled to work within a given day. There are a total of $S = DT$, time slices over the entire data collection period. Calls have two outcomes: a call results in a completed questionnaire or it does not. The observed probability of completing a questionnaire for a given time slice $s=1,\ldots,S$, is the proportion of calls resulting in completed questionnaires. We also assume that a sampled unit is called only once during any given time slice.

We used paradata from the 2010 cycle of the Survey of Labour and Income Dynamics (SLID) to develop models for the probability that a call made during a time slice would result in completing a questionnaire. The models were developed separately for each of the six regional offices by using

time slice as the unit of analysis. Laflamme (2008a) observed that the probability of completing a questionnaire decreases over time. Consequently, we developed two models to reflect this observation. One used the cumulative number of calls made up to and including a given time slice, while the other used the cumulative cost (time spent in minutes) up to and including the time slice.

Moreover, the probability of completing a questionnaire varies within a day. Therefore, we also included dummy variables to indicate period within the day in both the models. For each time slice s we defined dummy variables z_{ts}, $t = 1,...,T-1$ as: $z_{ts} = 1$, if $t = s \bmod T$ and $z_{ts} = 0$, otherwise . The last time slice within each day is used as reference.

The optimization of the total number of calls to be made within a time slice was carried out in two steps. In the first step, we predicted the probability p_s of completing a questionnaire within a time slice s $(s = 1, 2, \ldots, S)$ using the estimated parameters from either of the two regression models. In the second step, the total predicted cost, based on the number of calls resulting in completed questionnaires and those not resulting in completed questionnaires, was minimized subject to the following constraints: i. the number of calls within each time slice was non-negative, and ii. the expected overall response rate was set equal to the response rate observed in the actual survey. This is an iterative procedure because the objective function (total cost) depends on the number of calls and the probability of completing a questionnaire by time slice, whereas the probability of completing a questionnaire in a time slice is a function of cumulative number of calls made up to and including the particular time slice.

2.3 Models for Predicting the Probability of a Productive Call

2.3.1 Model based on Cumulative Number of Calls

The linear regression model is given by:

$$E(p_s) = \alpha_1 + \sum_{t=1}^{T-1} \beta_{1t}\ z_{ts} + \gamma_1\ \bar{C}_s, \tag{2.1}$$

where $z_{ts}, t = 1, 2, 3, ..., (T-1); s = 1, 2, 3, ..., S$ are the dummy variables defined above; $\bar{C}_s = \sum_{j=1}^{s} c_j / n$ is the average number of cumulative calls per sampled unit up to and including time slice *s*, c_j is the total number of calls made during time slice *j*, and *n* is the total number of sampled units associated with the regional office being analyzed. Note that the total number of cumulative calls up to and including time slice *s* is $C_s = n\ \bar{C}_s$.

The associated predicted probability of a productive call for time slice *s* is given by:

$$\hat{p}_s = \hat{\alpha}_1 + \sum_{t=1}^{T-1} \hat{\beta}_{1t}\ z_{ts} + \hat{\gamma}_1 \bar{C}_s \tag{2.2}$$

Since p_s is a proportion between 0 and 1, we also used the corresponding logistic model:

$$ln\left(\frac{p_s}{1-p_s}\right) = \alpha_1^* + \sum_{t=1}^{T-1} \beta_{1t}^*\ z_{ts} + \gamma_1^*\ \bar{C}_s \tag{2.3}$$

The corresponding predicted probability for time slice *s* is given as:

$$\hat{p}_s^* = \frac{\exp\left(\hat{\alpha}_1^{\,*} + \sum_{t=1}^{T-1} \hat{\beta}_{1t}^* z_{ts} + \hat{\gamma}_1^{\,*} \bar{C}_s \right)}{1 + \exp\left(\hat{\alpha}_1^{\,*} + \sum_{t=1}^{T-1} \hat{\beta}_{1t}^* z_{ts} + \hat{\gamma}_1^{\,*} \bar{C}_s \right)} \tag{2.4}$$

2.3.2 Model based on Cumulative Time Spent

The second model uses the observed average cumulative per unit cost (time spent in minutes) up to and including time slice *s*, *s=1,…,S*. The auxiliary variable $\bar{X}_s$ was computed as $\sum_{j=1}^{s} x_j / n$ where x_j is the observed cost in making calls for a given time slice *j*, and *n* is the total number of sampled units for the regional office being analyzed. The cost x_j of making calls for a given time slice j can be expressed as $x_j = t_1\, p_j\, c_j + t_2 \left(1 - p_j\right) c_j$, where t_1 and t_2 are respectively the units costs (time in minutes) of productive and non-productive calls. By replacing $\bar{C}_s$ by $\bar{X}_s$ in the linear regression model (2.1), the linear regression model becomes:

$$E\left(p_s\right) = \alpha_2 + \sum_{t=1}^{T-1} \beta_{2t}\ z_{ts} + \gamma_2\ \bar{X}_s . \tag{2.5}$$

In the case of linear regression model (2.5), the associated predicted probability for time slice *s* is given by:

$$\tilde{p}_s = \frac{1}{K_s}\left[\hat{\alpha}_2 + \sum_{t=1}^{T-1} \hat{\beta}_{2t} z_{ts} + \frac{\hat{\gamma}}{n} \sum_{j=1}^{s-1} \left\{ t_1 \tilde{p}_j c_j + t_2 \left(1 - \tilde{p}_j\right) c_j \right\} + \frac{\hat{\gamma}_2}{n} t_2 c_s \right], \tag{2.6}$$

where $K_s = \left[1 - \frac{\hat{\gamma}_2}{n} \left\{ t_1 - t_2 \right\} c_s \right]$. The above expression for the predicted probability can be derived by setting $\bar{X}_s$ equal to $\frac{1}{n} \sum_{j=1}^{s} \left[t_1 \tilde{p}_j c_j + t_2 \left(1 - \tilde{p}_j\right) c_j \right]$ and re-arranging terms. Thus, the predicted probability for the time slice s can be defined in terms of average cumulative per unit cost during the previous $\left(s-1\right)$ time slices. These in turn depend on the predicted probabilities of the previous $\left(s-1\right)$ time slices.

We did not use the logistic regression model because the predicted probabilities for the optimization algorithm would have to be obtained numerically, and this would have been too cumbersome.

2.4 Optimum Number of Calls by Time Slice

The cost of making CATI calls for a given time slice can be expressed as the sum of the costs of productive and non-productive calls. Thus, the predicted cost of making calls for a given time slice *s* can be expressed as $\tilde{x}_s = t_1\, \tilde{p}_s\, c_s + t_2 \left(1 - \tilde{p}_s\right) c_s$, and $\tilde{p}_s$ is determined from one of the above two models. Thus, the total data collection cost is given by the function $g\left(\underset{\sim}{c}\right)$ defined as:

$$g\left(\underset{\sim}{c}\right) = \sum_{s=1}^{S} \left\{ t_1 \tilde{p}_s\, c_s + t_2 \left(1 - \tilde{p}_s\right) c_s \right\}. \tag{2.7}$$

The "call" vector $\underset{\sim}{\boldsymbol{c}} = \left(c_1, c_2, ..., c_S\right)$ is obtained by minimizing the function $g\left(\underset{\sim}{c}\right)$ subject to the following constraints:

i. The number of calls for each time slice is greater than or equal to zero, and

ii. The expected response rate $\sum_{j=1}^{S} \tilde{p}_j c_j / n$ is equal to a pre-specified response rate *R*.

Additional constraints, e.g. upper and lower bounds on number of calls and/or cost (time spent) by time slice, can also be imposed, and would result in decreased potential cost savings.

Application to the Survey of Labour and Income Dynamics

CATI is used to collect data for social, agriculture and business surveys in Statistics Canada's six Regional Office (RO) call centres. Data Integration and Production Planning (DIPP), which is Statistics Canada's paradata warehouse, includes paradata for most Statistics Canada surveys conducted since 2003. DIPP is updated to include ongoing active surveys on a daily basis. In practice, this information becomes available the day after paradata information is collected or recorded.

We used paradata from the 2010 cycle of the Survey of Labour Income and Dynamics (SLID) to develop models for the probability that a call made during a time slice would result in a completed questionnaire. The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey, introduced in 1993, to measure labour market activity and income of individuals and families in Canada. The survey interviews the same people from one year to the next for a period of six years. The survey's longitudinal dimension allows evaluation of concurrent and often related events, which yields greater insight on the nature and extent of poverty in Canada. SLID also provides information on a broad selection of human capital variables, labour force experiences and demographic characteristics such as education, family relationships and household composition. Its breadth of content combined with a relatively large sample makes it a unique and valuable data set.

The models were developed separately for each of the six regional offices using the time slice as the unit of analysis. SLID has an annual sample of about 34,000 households. We used paradata only for that portion of the SLID sample that received at least one call during the initial 28 days of data collection. Each day was divided into a number of time slices such that the probability of productive call would be constant during each time slice. It was also required that the number of time slices should be large sufficiently large to permit reliable estimation of the probability. Based on these two criteria, each day was divided into four time slices (T=4): 7:00 - 11:00, 11:00 - 15:00, 15:00 - 19:00, and 19:00 - 23:00. Thus, there were S=112 time slices over the 28 days of data collection. Time

slices used in estimating the models were those where at least 50 calls had been made. Table 1 provides some summary statistics of the data for each of the six Regional Offices.

Table 1: Summary of the SLID paradata by regional office

RO	Sample size	Average Number of Calls per Sampled Unit	Completion Rate (%)	Cost (time in minutes) per call for	
				Productive Calls (t_1)	Non-Productive Calls (t_2)
1	5,774	6.41	54.6	22.37	2.89
2	5,405	5.05	61.9	21.45	3.12
3	5,403	4.46	66.8	22.57	3.72
4	5,178	4.90	57.5	24.22	3.06
5	3,221	4.71	52.6	26.55	3.36
6	5,490	2.73	44.5	23.56	3.25

We fitted the linear and logistic regression models using the cumulative average number of calls as auxiliary variable for time slices that had with 50 or more calls. We also fitted these models with cumulative average cost (time in minutes) as auxiliary variable. We only provide results for the models that used cumulative average number of calls as auxiliary variable. This is because the optimization using logistic model with cumulative average cost would have been very cumbersome (see equation 2.6). The parameter estimates for the linear regression and the logistic regression models fitted with cumulative average number of calls as auxiliary variable are given in Table 2. The parameter estimates were set equal to zero for the non-significant time of day effects, and the reduced model was fitted by excluding the corresponding dummy variables.

Table 2: Parameter Estimates for the Linear and Logistic Regression Models using average number of cumulative calls per sampled unit as predictor

RO	Number of Time Slices	Regression Model	Estimated Regression Parameters				
			$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}$
1	89	Linear	0.1732	0.000	0.000	0.000	-0.0262
		Logistic	-1.2367	-0.2488	-0.1711	-0.1885	-0.3624
2	83	Linear	0.2167	0.000	0.000	0.000	-0.0351
		Logistic	-1.0743	0.000	-0.2763	-0.1662	-0.3352
3	92	Linear	0.3310	-0.0643	-0.0625	-0.0511	-0.0593
		Logistic	-0.4364	-0.4817	-0.4825	-0.3568	-0.4983
4	84	Linear	0.2524	-0.0309	-0.0405	-0.0318	-0.0439
		Logistic	-0.8126	-0.3113	-0.3720	-0.3043	-0.4550
5	77	Linear	0.2151	-0.0385	-0.0262	-0.0331	-0.0322
		Logistic	-1.1609	-0.3522	-0.1905	-0.3449	-0.3420
6	81	Linear	0.2525	0.0000	0.0000	0.0000	-0.0647
		Logistic	-1.0559	0.0000	0.0000	0.0000	-0.4522

We observe from Table 2 that the estimates of the β's were either negative or non-significant. This implies that evening (the reference time period) is the most productive time period for calling. Similarly, the estimate of γ is always negative, and this implies that the productivity decreases over time as the cumulative number of calls increases.

3.1 Model Validation

We computed the absolute relative deviation for each time slice as $ARD_s = \left|1 - \tilde{p}_s / p_s\right|$, to measure the fit of the linear and logistic regression models. The average *ARD* was computed over the time slices used in fitting the models. The average *ARD* values (expressed in percent) are given in Table 3 for each of the regional offices. We observe from Table 3 that the *ARD* values are smaller for the logistic regression model for all ROs, except for RO 6. Thus, the logistic regression model results in a better prediction of the probability of a productive call.

Table 3: Percent ARD for the linear and logistic regression models by regional office

RO	Linear Regression	Logistic Regression
1	34.5	23.6
2	28.2	25.8
3	24.0	23.3
4	27.1	22.7
5	28.5	24.9
6	26.7	26.7

We also used the estimated model parameters to obtain the predicted probabilities corresponding to the actual number of calls made during each time slice, which in turn were used to simulate the number of completed questionnaires and total cost (time in minutes) for the 112 time slices. We then computed the cumulative number of completed questionnaires and the corresponding data collection costs (time in minutes) for the two models. We compared the simulated values with the actual values for both the linear regression and logistic regression models for each of the regional offices. The percent differences between the actual and simulated number of completed questionnaires, and data collection costs (time in minutes) are given in Table 4.

Table 4: Comparison between the simulated and actual number of questionnaires completed, and costs (time in minutes) for the linear and logistic regression models by regional office

RO	Linear Regression		Logistic Regression	
	Cases Completed	Time Spent	Cases Completed	Time Spent
1	3.5	1.3	-0.9	-0.3
2	3.2	1.3	-1.3	-0.6
3	1.3	0.6	-0.8	-0.3
4	-0.7	-0.3	-1.5	-0.7
5	2.2	1.0	-3.5	-1.5
6	-0.2	-0.1	0.0	0.0

The predicted values for both the linear regression and logistic regression models are very close to the observed values. However, the differences for the logistic model are smaller than those for the linear model. The logistic regression model is also preferable because it guarantees non-negative predicted probabilities.

3.2 Cost Savings

We computed for the two models the percent reduction in the number of calls and the cost relative to the actual values for each of the six ROs. As expected, the reduction in the number of calls and cost were larger for the logistic regression model for all ROs except for RO 5. Moreover, these reductions were minimal for RO 6 for both models because the actual calling schedule for the RO followed a more uniform distribution than the other ROs.

Table 5: Percent reduction in number of calls and cost (time spent) for the optimum schedule with linear and logistic regression models relative to the actual calls and cost

RO	Sample Size	Response Rate (percent)	Percent Reduction in Number of Calls		Percent Reduction in Cost (Time Spent)	
			Linear Regression	Logistic Regression	Linear Regression	Logistic Regression
1	5,774	54.6	17.7	27.1	11.2	17.2
2	5,405	61.9	10.1	19.7	5.9	11.5
3	5,403	66.8	39.8	41.0	22.6	23.3
4	5,178	57.5	31.8	31.2	17.6	17.2
5	3,221	52.6	30.4	24.9	17.2	14.0
6	5,490	44.4	0.7	0.3	0.3	0.1
All ROs	30,471	56.5	22.0	25.7	12.9	14.9

We also provide in Figure 1 the actual values and the optimum values using both the linear and the logistic regression models for RO 1. We observe that the optimum call schedule allocates collection resources uniformly over time for both models whereas the actual schedule tends to over-allocate collection resources during the initial data collection period.

4. Concluding Remarks

The methodology and results provided in this paper represent the first phase in the development of an optimized interviewer workload scheduling tool for a single CATI survey. The main findings from this study are: i) the collection resources should be uniformly allocated over the collection period, and ii) the late afternoon and evening are more productive. More studies and analyses are planned for optimizing the distribution of workload for multiple surveys. Additional constraints can also be introduced by putting upper and/or lower bounds on the number of calls and/or calling capacity. Such additional constraints will reduce cost savings.

References

Brick, J.M., Allen, B., Cunningham, P. and Maklan, D. (1996). Outcomes of a Calling Protocol in a Telephone Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association:* 142–149.

Couper, M. P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J. Nicholls W.L. and O'Reilly, J.M. (1998). Computer Assisted Survey Information Collection. Wiley series in survey methodology section, Chapter 15 by Edwards, Suresh and Weeks, 301-306.

Greenberg, B.S. and Stokes, S.L. (1990). Developing an Optimal Call Scheduling Strategy for a Telephone Survey, *Journal of Official Statistics*, **6**, 421-435.

Laflamme, F., (2008a). Understanding Survey Data Collection through the Analysis of Paradata at Statistics Canada. American Association for Public Opinion Research 63rd Annual Conference, *Proceedings of the Section on Survey Research Methods*: 4217–4224.

Laflamme, F. (2008b). Data Collection Research using Paradata at Statistics Canada. Proceedings of the Statistics Canada Symposium 2008. Data Collection: Challenges, Achievements and New Directions.

Laflamme, F., (2009). Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada. Paper presented at the *57th* Session of the *International Statistical Institute*, August 16-22, 2009, Durban, South Africa.

Stokes, S.L., and Greenberg, B.S. (1990). A Priority System to Improve Call-Back Success in Telephone Surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association:* 742–747.

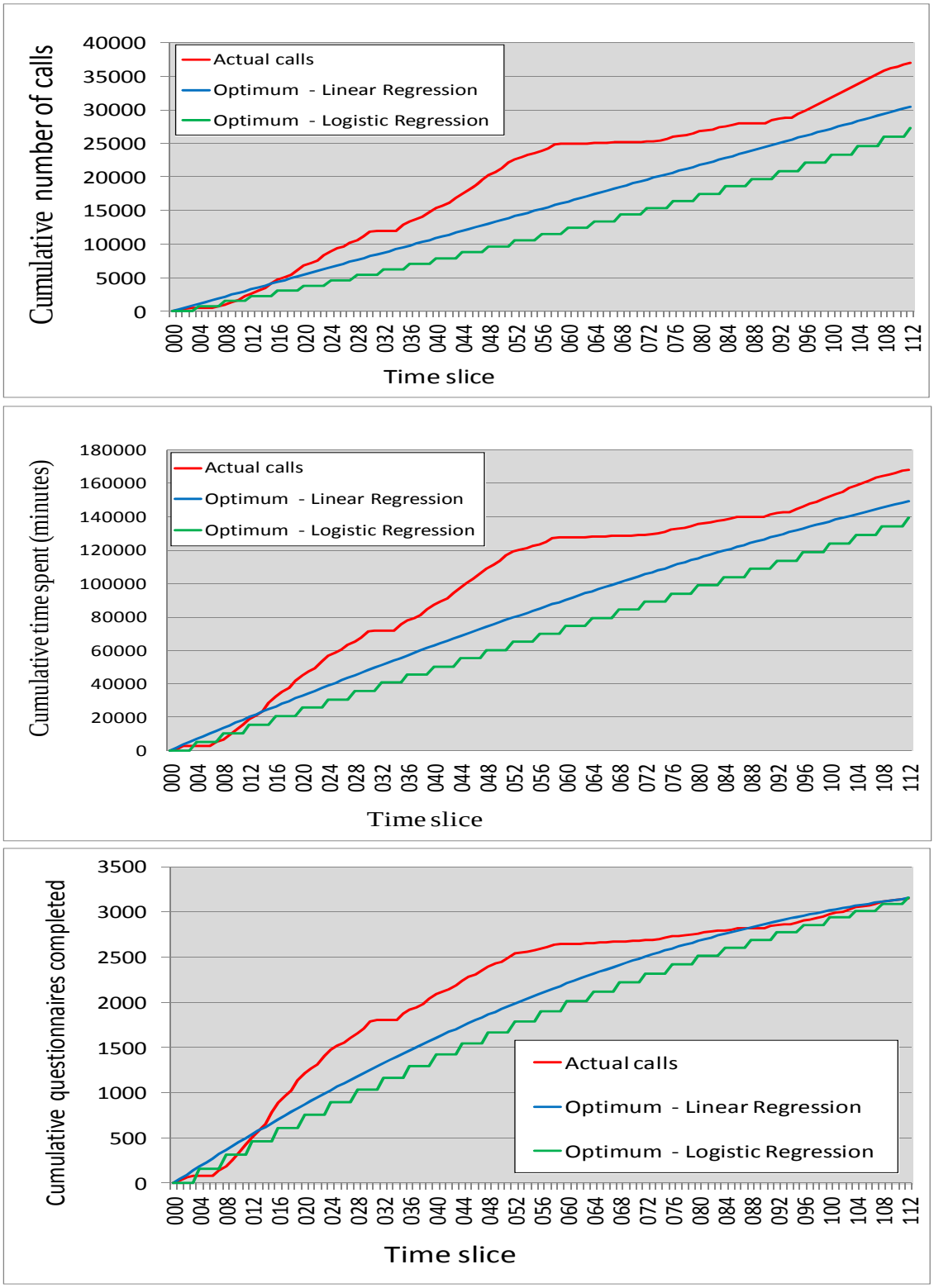

Figure 1: Cumulative number of calls, cumulative time spent, and cumulative number of questionnaires completed by Time slice – Actual and Optimum under Linear and Logistic Regression Models (RO1)