EXTRAPOLATING ON DAY RETURNED TO ADJUST FOR MAIL SURVEY NONRESPONSE

Charles Proctor, North Carolina State University
Department of Statistics, Box 8203, North Carolina State University, Raleigh, NC 27695-8203

Key Words: List frame, Propensity to return, Replicated subsamples

Introduction

An adjustment for nonresponse should be applied when the data show it's needed. The way the adjustment is made depends in some degree on how the survey was conducted, but mostly on how the data look over the successive stages of returns. In the face of an endless variety of numerical methods based on slightly different transforms of the variable day returned and of the variable of interest we argue that just the ranks by day returned and just the untransformed observations be used.

Background

The survey situations under consideration here were extensively reviewed by Scott (1961) and a more recent listing, albeit geographically limited, shows they are still serving their niche (Paxson, Dillman and Tarnai, 1995). These surveys involve sending out questionnaires by mail to a moderate sized (in the thousands) sample. The list frame will cover a somewhat specialized kind of establishment and one that routinely uses the mails. Due to cost considerations only two or three mailings will be sent and no personal visits nor, generally, any telephoning is expected.

This raises a complex issue in that when using the suggested method one needs to judge an upper bound to the proportion of mail hard-core nonresponders on the frame. The smaller it is, the more precisely it can be judged. That is, the adjustment is the more accurate when the trends seen among early to late returnees will be found to proceed linearly as more and more effort, if it were to be applied, would produce data from the entire sample.

Although Hochstim's (1967) example is a small sample, it is a clear cut case of the upsetting effect of mode change. The percentage of respondents who had heard of a test decreased through one, two and three letters (88% to 77% to 66%) but then the telephone was used and it jumped back up to 74% (Hochstim, 1967, p. 987). Much the same pattern is seen in Donald's (1960) experiences with League of Women Voters respondents and Jones' (1983) householders in Canberra, Australia. As Filion (1975) notes: "One danger in linear extrapolation is that 'hard-core' nonrespondents may differ from late respondents and upset observed trends." This is not to say that varied appeals cannot be applied in successive mailings---actually they probably should be. It is not essential that the whole series be smooth, but only that its trend be roughly correct (for "correct" read "known not to harbor hidden hard cores").

For a description of when extrapolation will work we quote from an early exponent of the method, Walter Hendricks (1949): "Incompleteness of returns in a mail survey usually implies a certain degree of bias in the results because a respondent's willingness to return a schedule is generally related to the nature of the item to be estimated from the survey."

Scott (1961) verified this connection between interest and propensity to return and added the proviso that sometimes high interest is coupled with a complicated report which itself may mean much work by the respondent and a delay. Our approach is empirical or adaptive in that the data either show trends or not and the adjustment will be applied accordingly.

Two Basic Types of Extrapolation

In the general case of one variable of interest, whose mean is required to be estimated, the data are the t_i and y_i pairs with $i = 1, 2, \ldots, n_R$ where t_i is the day the ith questionnaire was received ($t_i = 1$ for the first day a questionnaire was received) and y_i is the observed Y-value. The sample size is denoted n, and n_R is the number of responding cases. The basic methods ignore the time metric and use only the ordering of the t_i, the r_i's say where r_i is the rank of the ith t_i, with the tied ranks being used when t_i values are tied.

The two basic methods we will call linear extrapolation (LE) and linear prediction (LP). Roughly speaking, the LP method estimates a Y-value for every nonresponse case and then averages all n values to get $\hat{\bar{Y}}_{LP}$. The LE method estimates the grand mean by extrapolating from a progression of "cumulative means" to $\hat{\bar{Y}}_{LE}$. These cumulative means are written

y_i' in the following and, proceeding along the day returned ordering of the observations, they change whenever the day returned changes and are equal to the sum of all y-values returned on or before the given day divided by the number of such returns.

To produce the linear extrapolation estimate one regresses the y_i' on the r_i to find the prediction equation as $\hat{y}' = a' + b'r_0$, say. The value $\hat{\bar{Y}}_{LE} = a' + b'n$, the prediction for rank n, the full sample. The linear prediction estimate uses the regression of the y_i on the r_i to get the prediction equation $\hat{y} = a + b\,r_0$. The value $\hat{\bar{Y}}_{LP} = a + b(n+1)/2$, the prediction for the center of the ranks, $r_0 = (n+1)/2$.

The reader is invited to verify that with $n = 4$, $n_R = 3$, all $r_i = i$, and observations y_1, y_2 and y_3 the prediction equations are given by: $a' = (23y_1 - y_2 - 4y_3)/18, b' = (-2y_1 + y_2 + y_3)/6$, $a = (4y_1 + y_2 - 2y_3)/3$ and $b = (y_3 - y_1)/2$. The estimates themselves are:
$\hat{\bar{Y}}_{LE} = (-y_1 + 11y_2 + 8y_3)/18$ and
$\hat{\bar{Y}}_{LP} = (y_1 + 4y_2 + 7y_3)/12$. If the "true values" for the four observations are actually in line (say $-3\mu, -\mu, \mu, 3\mu$) then $E(\hat{\bar{Y}}_{LP}) = E(\hat{\bar{Y}}_{LE}) = 0$. If each y-value has an independent error of the same variance then $V(\hat{\bar{Y}}_{LE}) > V(\hat{\bar{Y}}_{LP})$.

Although the example is painfully small, one might expect a similar result to hold over most reasonable assignments of distribution to the y_i's. That is, the advantage of $\hat{\bar{Y}}_{LE}$ would only appear under some particularly unequal variance and nonlinear suppositions. Although the reality is most certainly some unequal variance, nonlinear situation it will "usually" not be the one favoring $\hat{\bar{Y}}_{LE}$. Thus we will choose $\hat{\bar{Y}}_{LP}$ and leave $\hat{\bar{Y}}_{LE}$ by the way.

In early attempts to foresee all possible such estimators we considered using the t_i and all monotone transforms of them, the x_i say. By first regressing the x_i on the r_i one could predict x_i for the nonresponders and then by regressing, using even some nonlinear model, the y_i on the x_i one would get predicted values for the nonresponding y_i and thereby could calculate the full-sample mean as estimator.

Although some models can clearly be seen to fit better than others in terms of the regression results, there remains any number of candidates all fitting equally well and all giving quite different estimates. This was deemed an embarrassment of riches. At some point an arbitrary selection would have to be made. The solution was to shift the judgemental aspect onto assigning a value for γ, where $1 - \gamma$ is a measure of how much the extrapolation needs to be moved back toward the mean of the responders. This practice can be justified by the experiences cited above showing that when trends are broken, they revert to the levels of the earlier responders and not to more extreme levels.

The Adjustment

Having settled on $\hat{\bar{Y}}_{LP}$ as the extrapolation estimate for the full sample the final question is how to combine it with the unweighted mean of the responders. For economy of notation we'll denote $\hat{\bar{Y}}_{LP}$ as $\bar{y}_a$, the "adjusted mean," and $\bar{y}_u$ will be the "unadjusted mean" or the unweighted mean of the responders. Randomness in $\bar{y}_a$ and $\bar{y}_u$ arises from the random numbers used to select the sample of n from the N on the frame, and also from the returning process.

Each sampling unit is viewed as characterized by a response propensity ρ_α say, as well as by a Y-value, Y_α, where $\alpha = 1, 2, ..., N$ indexes the frame listings. The mean $\bar{Y} = \Sigma Y_\alpha / N$ is the parameter of interest. When the values of Y_α and ρ_α are determined by the superpopulation process they may be independent and if so, no adjustment would be needed, but when Y is some measure of interest in the survey topic there will generally be a positive association with ρ_α.

Although such machinery is not needed, we may suppose that the chance of unit α returning at time t is given by $1 - e^{-t/\rho_\alpha}$. By setting a cutoff date, T say, we have described a stochastic process with observations t_i when the generated times are less than T but missing if not. The Y-values are present or missing in accord with the t-values.

Now consider the joint distribution of $\bar{y}_a$ and $\bar{y}_u$ under repetitions of a fixed value of n_R. Denote by V_{aa}, V_{uu} and V_{au} the variances and covariance of this joint distribution. Let $\bar{y}_t$ denote the sample mean of all n Y-values in the sample. There are some issues of response error such as mode-bias (telephone vs. mail vs. interview) we are avoiding here, but use of any standard sampling desing will insure that $E(\bar{y}_t) = \bar{Y}$ the mean of the population. Now define γ so that $\gamma E(\bar{y}_a) + (1-\gamma)E(\bar{y}_u)$ equals $E(\bar{y}_t)$. Setting the value of γ is based on two considerations. One is that the amount of response, n_R/n, sets a lower bound. The other is that if the nonrespondents seem to be mostly hard core then γ should be set near this lower bound but if there are few hard core ones then γ should be increased.

If the level of y changes linearly with propensity to return and r is the proportion of the sample responding, while h is the "effective" proportion of hard core nonrespondents then a little geometry shows that $\gamma = (r - h + h^2)/r$. Actually setting $\gamma = (r-h)/r$

will work just fine. The responding proportion r is known when the cut off date is set but setting h requires judgement by the researcher. For example, $h = .10$ can mean that 10% of the population does not answer mail inquiries and has a mean level of y-value equal to the responders or that its mean level is less than the responders mean but more than 10% are hard core so h is "effectively" .10 .

Now we can state the optimizing condition as to choose β in $\hat{\overline{Y}}_A = \beta\,\overline{y}_a + (1-\beta)\overline{y}_u$ to minimize $E(\hat{\overline{Y}}_A - \overline{Y})^2$. That is, minimize:

$$E[\beta(\overline{y}_a - \overline{Y}_a) + (\gamma - \beta)(\overline{Y}_a - \overline{Y}_u)$$

$$+ (1-\beta)(\overline{y}_u - \overline{Y}_u)]^2$$

$$= \beta^2 V_{aa} + 2\beta(1-\beta)V_{au} + (1-\beta)^2 V_{uu}$$

$$+ (\alpha^2 - 2\gamma\beta + \beta^2)(\overline{Y}_a - \overline{Y}_u)^2.$$

Putting $\triangle = (\overline{Y}_a - \overline{Y}_u)$ we find,by differentiating,

$$\beta = (V_{uu} - V_{au} + \gamma\beta^2)/(V_{aa} - 2V_{au} + V_{uu} + \triangle^2).$$

Notice that with large $\triangle$, $\beta \rightarrow \gamma$. We can break out the variance differential as $\delta = V_{au}/V_{uu}$, say, and let ρ be the correlation between $\overline{y}_a$ and $\overline{y}_u$ and then, when $\triangle$ is small and $\delta = 1$, $\beta = .5$. In practice we have found V_{aa} to be 2 or 3 times as large as V_{uu} which tends to drive β downwards. In fact, we suggest using the data to estimate $\triangle$, V_{uu}, V_{aa} and V_{au} and thereby find a β which is then used to compute $\hat{\overline{Y}}_A$.

Estimation of the V's and $\triangle$ is facilitated by drawing the sample as replicated subsamples. Then one has several $(\overline{y}_a, \overline{y}_u)$ pairs of values to use in the estimation. This was the case in the example to be reported next.

Example

The Institute of Statistics at Raleigh, North Carolina cooperated with the North Carolina Alliance for Competitive Technologies in doing a mail survey. A list was prepared of 10,092 company addresses in the state. These were sorted by SIC (Standard Industrial Classification) codes and a 6-start systematic sample was drawn of size 2088 (348 in each replicated subsample). Questionnaires were mailed to all 2088 addresses in June of 1995 and a second mailing to nonresponders was done in August. In the second mailing a request was added to return the forms even though the questions were not thought to apply.

Table 1 shows frequencies of the four possible responses to question 4 ("which of the following (activities) are conducted at your plant?"), Part 1 ("Manufacturing engineering and process improvement"). Notice how response picks up at day 37 and also how the no-answer cases become more common in accord with the instruction to return the form in any case.

Taking the y_i to be the zero-one indicator of the answer "Yes," we find a prediction equation of:
$\hat{y} = .6174 - .000317\,r_0$ which for $r_0 = 1044.5$ gives $\hat{\overline{Y}}_{LP} = .2860$, whereas $\overline{y}_u = .4735$ is the overall unweighted mean. For the 6 subsamples separately the $\overline{y}_a$ values were found to be: .459, .468, .425, .472, .494 and .524 while the $\overline{y}_u$ were: .262, .232, .299, .295, .186 and .478. Estimates of V_{aa}, V_{au} and V_{uu} were found as: $V_{aa} \hat{=} .01008$, $V_{au} \hat{=} .00153$ and $V_{uu} \hat{=} .00112$ while
$\Delta^2 \hat{=} (.28603 - .47351)^2 = .03515$, and with $\gamma = 1$ we found $\beta = .82$ and when $\gamma = .8$ then $\beta = .64$.
Since $r = 940/2088 = .45$ and if h may be set at 10% then $\gamma = .80$ (or $\gamma = .78$ without the h^2 term).

Thus the adjusted estimate becomes (when one judges that the trend would return to the unweighted average to the extent of $\gamma = .8$) $\hat{\overline{Y}}_A = .353$. When calculated over the 6 subsamples one finds the estimated standard error of $\hat{\overline{Y}}_A$ to be .029. Recalling that there were 10,092 companies on the list, one states that there are 3560 companies in the state doing engineering development at their plant site, and the standard error is ±290 companies. If no adjustments were made the proportion of .474 might lead one to estimate 4780 such companies.

Applying the same approach to each response category we found the following β's: For "No answer," $\beta = .75$, for "Not at this plant" $\beta = 0$ and for "Don't know" $\beta = 0$ also. The separate adjusted estimates do not add to 1 although they don't miss it by much. They are: .436 (No answer), .353 (Yes), .223 (No), and .023 (Don't know) and become .42, .34, .22 and .02 when renormed to add to 1.

When faced with a dichotomous response variable one may be tempted to run logistic regression on the ranks rather than the ordinary least squares called for in the LP method. We resisted this temptation because the nonlinearity would require us to choose between predicting separately for all $n - n_R$ nonresponders and adding, versus predicting a central value. But with a multinomial response (such as the four possible responses to question 4) the temptation arises anew -- this time to run a multinomial logit regression.

When a multinomial logit regression is used to predict the four response probabilities for all nonresponding cases and these are added, the adjusted estimates are found to be .508, .304, .168 and .020 versus the unweighted proportions of .280, .474, .223 and .020. Now, however, we must leave it to others to find the method for amalgamating these two vectors. Simply averaging them is, of course, a possibility but the differences in β's found above, and the substantive reasonableness of the differences, argue against doing this.

We look finally at the numerical variable from Table 1, Q3 = the number of engineers working at the plant. This is the only opportunity we have to check an estimate against a somewhat known population quantity.

The prediction equation is found as $\hat{y} = 1.146 + .00273 r_0$ which yields $\widehat{\overline{Y}}_{LP} = 3.995$ as compared to $\overline{y}_u = 2.383$. This is a case where β turns out to be zero and so the adjusted estimate is $\widehat{\overline{Y}}_A = 2.383$ engineers per plant. The standard error over the 6 subsamples is .452 and thus the number of engineers in manufacturing in the state is estimated to be 24,000 (±4,600). Data furnished by the Labor Market Information Division of the N. C. Employment Security Commission showed there were 14,000 engineers employed in manufacturing industries in 1992. The same publication mentions that "In 1992 there were 832,900 persons employed in manufacturing industries." The 1990 Census number is 864,371, so we might expect more than 14,000 to be found in a census done today (in 1995) and this is what we would take as the criterion number. Still the number would not be much above 16,000 which is inside two standard errors, but not inside one, of the estimate. We suspect that respondents may call some employees "engineers" for our survey, but not for the official count.

Although we resolved to go with the LP method rather than the LE, we were still carrying along the calculations for the LE adjustment while working with Q3 and the results may be of interest: $\widehat{\overline{Y}}_{LE} = 3.975$ and $\beta = .28$ so that a combined estimate becomes 2.826 with a standard error of .569. This standard error is larger than the .452 for the LP case. However, the observed variability over subsamples was larger for $\widehat{\overline{Y}}_{LP}$ than for $\widehat{\overline{Y}}_{LE}$. That is, the "cumulative" means gave a "better" fit than the individual ones. This, we believe, is due to the presence of a few large values (large companies) along with largely zeroes in the last few days, giving heterogeneity of variance problems to the LP method.

Summary

The conditions that call for nonresponse adjustment seem to inhere in moderate sized mail surveys. The list of addressees generally contains a healthful supply of interest in the survey topic. Such interest will be scattered on the list and will be tied to response propensity, and many items on the questionnaire itself are likely to be aimed at measuring variables also tied to that interest.

One should recognize that the trends that call out the adjustment may in some cases not lead to the "true" population value. The most common exception would appear to arise from the presence of sampled addresses averse to using the mails or those who are survey-shy. From experiences that have been reported, these hard-core mail nonrespondents are more like early responders than like late responders. One needs then to guess what mixture of the adjusted and unadjusted means would equal the population mean as $\overline{Y} = \gamma\overline{Y}_a + (1-\gamma)\overline{Y}_u$, where γ is set at one if there are no hard-core nonrespondents or at .9 or .8 if there are some. In a sense, setting $\gamma = .8$ would be a reasonable choice as a default standard but the formula $\gamma = (r-h)/r$ is available when h, the effective proportion hard core, can be judged.

The number of options for fitting and extrapolating (or predicting or projecting) responses are endless so we propose selecting one, the LP method, and not looking back. What is most important about the full method is its use of replication in the sample design to determine the relative weight to assign between the unadjusted and the adjusted means. In the course of minimizing the mean squared error the method balances bias reduction against variance increase.

Replicated subsamples in a multi-start systematic design is, or should be, a standard method for sampling a list of mail addresses. With complex designs, say stratified, multi-stage, one could create subsamples using "balanced" sets of PSU's and then apply the method. With a simple random sample one just randomly divides the sample into, say, 10 replicates. If one mails to all addresses on the list the method should still be applied based on replicates. Notice the absence of finite population corrections.

Although the method provides a standard error based on the replicate subsamples, and we have reported these values, such standard errors are underestimates. That is, the method has a hidden bias depending on how relatively unrealistic is the guessed value of γ. In a technical sense, since γ is defined under superpopulation expectations, its use in any finite population will entail bias. By applying the adjustment to estimate known population values and checking their reasonableness one can hope to detect serious mistakes in guessing γ. If there are compelling reasons to have an unbiased estimate then one should perhaps consider personal interviews rather than mailings or some more potent combination of appeals.

References

Donald, M. H. (1960). "Implications of non-response for the interpretation of mail questionnaire data," *Public Opinion Quarterly*, vol. 24, pp. 99-114.

Filion, F. L. (1974). "Estimating bias due to nonresponse in mail surveys," *Public Opinion Quarterly,* vol. 39, pp. 482-492.

Hendricks, W. A. (1949). "Adjustment for bias by nonresponse in mailed surveys," *Agricultural Economics Research,* vol. 1, pp. 52-56.

Hochstim, J. R. (1967). "A critical comparison of three strategies of collecting data from households," *Journal of the American Statistical Association,* vol. 62, pp. 976-987.

Jones, R. G. (1983). "An examination of methods of adjusting for nonresponse to a mail survey: a mail-interview comparison," Chapter 13 in Madow, W. G., Nisselson, H. and Olkin, I. (eds.), Incomplete Data in Sample Surveys vol. 3, Academic Press. New York.

Paxson, M. C., Dillman, D. A. and Tarnai, J. (1995). "Improving response to business mail surveys," Chapter 17 in Cox, B. G., Binder, D. A.,. Chapman, B. N., Christianson A., Colledge, M. J. and Kott, P. S. (eds.), Business Survey Methods, Wiley, NY.

Platek, R., Singh, M. P. and Tremblay, V. (1978). "Adjustment for nonresponses in surveys," Chapter 11 in Namboodiri, N. K. (ed.) Survey Sampling and Measurement, Academic Press, NY.

Scott, C. (1961). "Research on mail surveys," *Journal of the Royal Statistical Society* , vol. 124, pp. 143-195.

Table 1. Responses to Question 4, Part 1, and Means of Variable Q3 for the 66 Days Questionnaires were Received, NC Manufacturers Survey, 1995.

(The question asked was: Is research conducted at your plant? and Q3 equals the No. of Engineers)

	No.				Q3		No				Q3
Day	Answ.	Yes	No	DK	Mean	Day	Answ.	Yes	No	DK	Mean
1	0	1	0	0	3.00	32	1	2	0	0	1.33
2	0	2	4	0	.83	36	0	1	1	0	3.00
3	11	38	17	2	.72	37	48	42	22	1	1.28
4	13	61	18	4	1.95	38	33	27	21	1	2.63
5	11	22	14	0	1.02	39	5	15	7	9	.78
8	8	35	16	2	1.67	40	14	18	7	0	.64
10	10	14	4	0	5.75	43	16	10	11	2	2.54
11	1	3	2	0	.43	44	12	8	7	1	.46
12	2	13	1	0	8.25	45	0	1	0	0	45.00
15	9	14	5	0	.50	46	17	17	5	0	2.92
16	6	14	5	0	2.44	50	3	4	5	1	2.54
17	1	7	4	0	3.08	51	17	17	7	1	10.02
18	2	7	3	1	1.08	54	5	3	5	0	.39
19	2	2	3	0	1.29	57	1	4	3	0	.50
22	1	13	2	2	.83	58	1	2	0	0	.33
23	0	1	0	0	.00	59	0	1	0	0	75.00
25	1	2	3	0	2.00	60	3	1	0	0	.75
26	1	5	0	0	9.17	61	2	3	1	0	1.17
29	1	3	2	0	1.00	65	4	2	0	1	.00
30	1	0	1	0	.00	66	0	3	1	2	2.17
31	0	2	0	0	4.50	Sums	263	440	207	30	