American Fisheries Society Symposium 12:292–297, 1991

Use of Model Predictions as an Auxiliary Variable to Reduce Variance in a Creel Survey

JOHN M. HOENIG

Science Branch, Department of Fisheries and Oceans
Post Office Box 5667, St. John's, Newfoundland A1C 5X1, Canada

C. MARK HEYWOOD

Minnesota Department of Natural Resources, Regional Fisheries Headquarters
Post Office Box 6247, Rochester, Minnesota 55904, USA

Abstract.—Fishery surveys generally do not use information from previous years to make estimates for the current year. However, it is possible to use prior data to develop a detailed regression (or other) model for predicting fishing activity as a function of explanatory variables such as time of season or depth. The model predictions can be used as an auxiliary variable to reduce the variances of estimates in the current survey if the predicted values have a strong correlation with the current observations. The method can be considered an application of difference or regression estimation for survey sampling. The auxiliary variable method is simple to implement: wherever a sample mean is to be computed, each observation in the current year, Y_i, is replaced by the controlled variate, Y_{ci}, so that $Y_{ci} = Y_i + c(\mu_x - X_i)$; X_i is the model prediction corresponding to sampling unit i, μ_x is the mean of X_i taken over all possible sampling units, and c is a constant. The controlled variates, Y_{ci}, are then analyzed as one would analyze the uncontrolled observations, Y_i. The optimal value of c is equal to the coefficient of regression of Y on X. When this value is used for c, the proportional reduction in variance is equal to the correlation, ρ^2, between the current observations and the model predictions.

In most survey work, information from previous years is used only to improve the sampling design (e.g., allocation of sampling effort); prior information is not used in the estimation procedure for the current year. We showed (Hoenig et al. 1986, 1989) that information from a sportfishing survey can be used to develop a detailed regression model for predicting fishing activities (catch, effort, catch rate) as a function of explanatory variables such as time of season, time of day, ship size, and depth. Here we show how model predictions can be used as a control or auxiliary variable to reduce the variances of estimates in the current survey if the model predictions have a sufficiently high correlation with the current observations. The method can be considered an application of difference or regression estimation for survey sampling (Cochran 1977; Sukhatme et al. 1984).

Suppose we wish to estimate the mean daily fishing effort, μ_Y, from a set of observations, Y_i, of fishing effort on n randomly selected days. Suppose further that data from previous years are available and that a plot of fishing effort versus day of the fishing season in the previous years reveals a strong relationship between effort and time of season. We could make a model of this relationship in any of a variety of ways: by fitting a regression model to the data, by using a smoothing technique such as a running average, or simply by fitting a line by eye. For each day j of the fishing season, there is a value of the auxiliary variable X_j which is computed from the model. Furthermore, it is an easy matter to compute the mean μ_X of the X_j values (taken over all days of the fishing season). The auxiliary variable X is not really a predictor of fishing effort in the current year—fishing effort may have increased in the current year, for example. However, it may well be that the auxiliary variable X is correlated with the fishing efforts Y observed in the current year, because the pattern of fishing over the course of the season is similar in all of the years considered. If so, then the auxiliary variable can be used to obtain a reduction in the variance of the estimated mean daily fishing effort in the current year.

The auxiliary variable method is simple to use. Anywhere a mean is to be computed, each observation in the current year, Y_i, is replaced by the controlled variate, Y_{ci}, so that

$$Y_{ci} = Y_i + c(\mu_x - X_i)\,; \qquad (1)$$

X_i is the model prediction corresponding to sampling unit i, μ_x is the mean of the X_i taken over all possible sampling units, and c can be any real number. The controlled variates, Y_{ci}, are then

analyzed as one would analyze the uncontrolled observations, Y_i. (A computationally convenient way to obtain the mean $\bar{Y}_c$ of the Y_{ci} is $\bar{Y}_c = \bar{Y} + c(\bar{X} - \mu_x)$ where the bar symbol indicates a sample mean.) This adjustment is rather intuitive: if X_i is above the mean of the X values, and if X and Y are positively correlated, then Y_i is probably above the mean of the Y values. Consequently, Y_i can probably be brought closer to the mean of the Y values by subtracting an adjustment factor. The adjusted or controlled variates, being closer to the mean, thus have a lower variance than the uncontrolled Y_i.

If the value of c is fixed in advance as a constant, then the expected value of Y_{ci} is equal to Y_i so that the procedure does not introduce any bias. Thus, the expected value of Y_{ci} is

$$\begin{aligned} E[Y_{ci}] &= E[Y_i + c(\mu_x - X_i)] \\ &= E[Y_i] + cE[\mu_x - X_i] \\ &= E[Y_i], \end{aligned}$$

since the expected deviation of a random variable from its mean is by definition zero. Fixing the value of c in advance results in a difference estimator, whereas estimating the optimal value of c (see below) results in a regression estimator.

For a reduction in variance to be expected, it is necessary that

$$c^2V(X) - 2c\ \mathrm{Cov}(X, Y) < 0;$$

$V(X)$ is the variance of the model predictions taken over all possible sampling units, and $\mathrm{Cov}(X, Y)$ is the covariance of the model predictions, X, with the observations in the current survey at the corresponding time and location, Y. This can be seen by deriving the variance of (1):

$$V(Y_{ci}) = V(Y_i) + c^2V(X_i) - 2c\ \mathrm{Cov}(X_i, Y_i). \quad (2)$$

The value of c providing the greatest variance reduction is

$$c_{\mathrm{opt}} = \frac{\mathrm{Cov}(X, Y)}{V(X)}, \quad (3)$$

which corresponds to the slope of the regression of Y on X (hence, the name regression estimator). This is seen by taking the derivative of equation (2) with respect to c, setting it equal to zero, and solving for c. When c_{opt} is used in equation (1), the expected proportional reduction in variance is equal to the correlation, ρ^2, between the current observations (Y) and the model predictions (X).

The optimal value of c cannot be known, but it can be estimated by substituting sample estimates for the parametric values of $\mathrm{Cov}(X, Y)$ and $V(X)$ in (3). In this case, the estimated value $\hat{c}$ is a random variable and equation (1) becomes

$$Y_{ci} = Y_i + \hat{c}(\mu_x - X_i). \quad (4)$$

Equation (4) does not have expected value equal to Y_i; that is, it does not provide unbiased estimates. (See Sukhatme et al. 1984 for a derivation and an estimator of the bias.) However, this statistical bias should not be of concern unless the sample sizes are very small. If Y_i and X_i are expressed in the same units (e.g., angler-hours), and if the fishery does not change much from year to year so that the difference between Y_i and X_i is small, then the optimal value of c is probably close to 1.

Example: An Effort Survey

We illustrate the method by considering a portion of an aerial survey of fishing effort conducted in the summers of 1984 and 1985 over Lake Vermillion, Minnesota. The sampling design consisted of two-stage sampling within stratified random sampling (see Hoenig et al. 1986, 1989). Although the example deals with a sport fishery, the type of sampling design used was quite general and this design could be used for sampling commercial fishing.

In both years, the primary sampling unit was a quarter of a day and the variable observed was the number of boats fishing. The primary sampling units were divided into eight strata based on the day type (weekday versus weekend day), and the period of the day (i.e., each day was divided into four quarters). Only the four weekend-day strata are considered here. Within a stratum, a quarter of a day was selected randomly at the first stage and an instant at which to make the count of boats was selected at the second stage. The estimated fishing effort within the quarter day was equal to the count of boats times the duration of the quarter day. The estimated total fishing effort for any given stratum was found by multiplying the mean of the estimated fishing efforts within the stratum by the number of days within the season (primary sampling units). Table 1 gives estimates obtained in this way for the four weekend strata in the summer of 1985. The seasonal total was 68,323 boat-hours with an estimated precision of ±35.6% (95% confidence interval).

A notable trend over time was observed in the estimates of fishing effort in each of the four

TABLE 1.—Estimates of fishing effort on Vermillion Lake in 1985 for four weekend-day strata. Estimates were computed under a two-stage stratified-random-sampling scheme.

Statistic	Period of day				Total
	1	2	3	4	
Mean daily effort[a]	177.8	737.9	366.9	306.3	1,588.9
Seasonal total effort[a]	7,647	31,730	15,777	13,170	68,323
Sample size, n	10	14	8	8	40
Variance of total	1,572,898	119,645,924	3,451,264	23,489,618	1.5×10^8
$2 \times$ CV %[b]	32.8%	69.0%	23.6%	73.6%	35.6%

[a]Measured in boat-hours.
[b]$2 \times$ CV % = 200 times the coefficient of variation = half the width of a 95% confidence interval expressed as a percentage of the estimate.

weekend-day strata in the 1984 data (Figure 1). (The trend was less evident in the data for the four weekday strata in part because we had fewer observations at the beginning of the season when the trend was strongest.) A multiple linear regression model was developed to explain the variability in estimated effort (see Hoenig et al. 1986, 1989). The model predictions for the first period of the day are

$$\text{effort} = 346 - 47\ \text{wk} + 3.8\ \text{wksq},$$

and for the second, third, and fourth periods of the day the predictions are

$$\text{effort} = 1{,}657 - 151\ \text{wk} + 3.8\ \text{wksq}.$$

Here, wk is the week of the season and wksq is the square of the week.

A trend over time can also be seen in the 1985 data (Figure 1). We tried to exploit this trend within each of the four weekend-day strata by subtracting from each observation the difference between the model prediction for that time (based on the regression from the previous year) and the mean of the predictions for the entire stratum (i.e., we used the difference estimator with $c = 1.0$). The means were then recalculated for each stratum with the adjusted values (Table 2). The estimated mean amount of fishing effort in a day (all four strata summed) changed very little, though the estimates for individual strata changed somewhat. In both sets of estimates, the first period of the day appeared to have the lowest fishing effort and the second period of the day had the highest estimated effort. The estimated 95% confidence region for the difference estimator was slightly smaller than for the unmodified two-stage stratified-random-sampling estimator (±35.6% for the unmodified estimator versus ±31.6% for the difference estimator). Confidence regions for estimates for individual strata varied considerably between the two methods of estimation.

Although we decided beforehand to use the difference estimator with $c = 1$, we also estimated the optimal value of c by computing the variance of all the model predictions, $V(X)$, and estimating the covariance between the model predictions and the actual fishing efforts in 1985, $\widehat{\text{Cov}}(X, Y)$. These values were then substituted in equation (3) with the following results.

Period of day	Estimated correlation, r, between X and Y	Estimated optimal c
1	−0.13	−0.06
2	0.63	1.52
3	0.47	0.10
4	0.95	0.71

Note that when the sample correlation is low, the estimated optimal value of c is close to zero. This is consistent with the idea that an adjustment is only worthwhile if the auxiliary variable has some predictive value (see equation 3).

Discussion

In this study, the estimated total fishing effort on weekends (four strata summed) was virtually the same for the two methods: 68,323 boat-hours for the unmodified two-stage stratified-random-sampling method, and 68,101 boat-hours for the difference estimator. The difference estimator provided a small reduction in estimated variance (±35.6% for the unmodified-method versus ±31.6% for the difference estimator). Although this improvement may not seem important, it should be noted that it was achieved at no additional cost—that is, without any additional sampling. Also, if sampling effort in the previous survey had been specifically allocated to provide information for model construction, then the model would be better at describing the trends in the observations and therefore would be better at reducing variance. Specifically, having more ob-

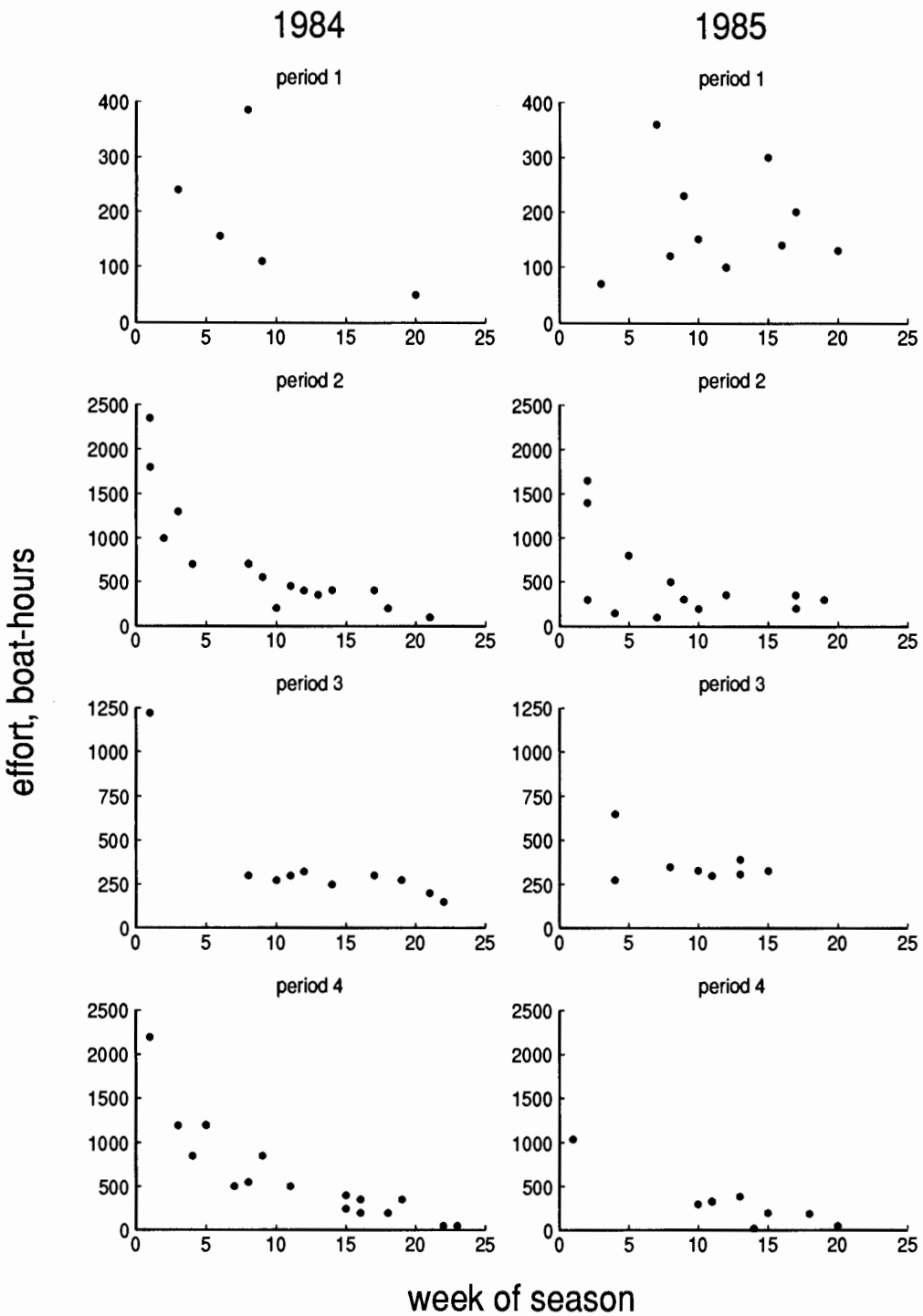

FIGURE 1.—Estimates of fishing effort by quarter-day periods on weekend days during 20–23-week seasons at Vermillion Lake, Minnesota, 1984 (left) and 1985 (right).

TABLE 2.—Estimates of fishing effort on Vermillion Lake in 1985 for four weekend-day strata. Estimates were computed with a difference estimator (with $c = 1$) for each stratum.

Statistic	Period of day				Total
	1	2	3	4	
Mean daily effort[a]	161.7	559.5	390.4	471.6	1,583.7
Seasonal total effort[a]	6,954	24,060	16,807	20,280	68,101
Sample size, n	10	14	8	8	40
Variance of total	13,444,546	73,886,369	21,667,350	6,734,655	1.2×10^8
$2 \times$ CV %[b]	105.5%	71.5%	55.4%	25.6%	31.6%

[a]Measured in boat-hours.
[b]$2 \times$ CV % = 200 times the coefficient of variation = half the width of a 95% confidence interval expressed as a percentage of the estimate.

servations at the beginning of the season when the temporal trend was the strongest would have been advantageous. A better model might also be constructed by considering other explanatory variables, such as weather conditions. For example, Abramson (1990) used a product estimator (which is very similar to a regression estimator) to try to improve the efficiency of an estimator of salmon landings by using information on wind speed. Another possibility is to use the time of day (within the quarter day) at which the count is made as an auxiliary variable. Information from surveys conducted over a period of years could be combined to develop an improved model.

An important advantage of the approach described here is that it can be applied a posteriori. For example, an investigator may not have control over the design and implementation of a survey. Indeed, there may be good reasons to avoid changing survey design in order to maintain continuity in survey methods. Yet the investigator may be able to easily improve a set of current estimates by applying the control variate technique (using, for example, a model derived from the previous year's estimates).

Although we used linear regression to develop a model to obtain an auxiliary variable, this is by no means the only way to proceed. A simple, nonparametric model (such as a smoothing technique like moving averages or lowess [Chambers et al. 1983]) could also be used. Another possibility is to simply use guesses of anticipated effort as the auxiliary variable. Consider as an example the task of estimating the number of fish on board N commercial fishing boats arriving in a port during an afternoon. One approach would be to randomly sample n of these boats. An alternative would be to first write down a best guess as to what each boat has on board based on a quick glance. These guesses provide an auxiliary variable with the necessary attributes: the value is known for each boat in the population, the mean of the auxiliary variable can be computed, and the guesses are likely to be correlated with the actual landings. The fact that the guesses are based on subjective judgements rather than objective criteria, and that the person making the guesses may tend to consistently overestimate or underestimate the catches, does not invalidate the methodology in any way. Note, however, that a personal bias (e.g., tendency to overestimate the catches) would mean that the optimal value of c would not be 1.0; one would probably want to estimate the optimal value by using equation (3) with sample estimates substituted for parametric values. Also, if a person's judgement is poor and the guesses are inconsistent, then this would result in only a small variance reduction.

If sampling effort is sufficiently intense, and the trends in the data are strong, then it may be profitable to build a model for the current year and use the model to estimate fishing effort. For example, Hoenig et al. (1986, 1989) integrated their regression model over time to obtain estimates of fishing effort in each stratum. They noted that the trends must be strong to obtain a gain in efficiency over stratified random sampling. Also, the estimated variances will be conditional on the form of the model being correct. Model misspecification will inflate the mean square error but this will be reflected only partly in the estimate of the variance. In contrast, use of model predictions as an auxiliary variable for a difference estimator results in unbiased estimates, appropriately estimated variances, and, if the correlation is high between model predictions and observations, in reduced variance. Also, developing an appropriate regression model to explain trends in fishing effort is critical for model-based estimation, and this may require sophisticated analysis. In contrast, use of the model's predictions as an auxiliary variable is a straightforward matter, and any

model with strong correlation with the current year's observations will be useful for reducing variance.

We used the difference estimator, rather than the regression estimator, because we felt that the optimal value of c was reasonably well known (i.e., close to 1), that it would be easier to justify use of the technique if it were clear that the estimator was unbiased, and that it would be difficult to estimate c with such small sample sizes. Also, with small sample sizes, the bias in the regression estimator could be significant. In general, the regression estimator is preferred if the sample sizes are reasonably large and if prior information on the optimal value of c is poor. Grimes and Sukhatme (1980) proposed a regression estimation scheme in which a test is performed first to judge if a prior value of c is consistent with that estimated from the data. This procedure might be quite useful.

Use of difference and regression survey sampling estimators in fishery work is uncommon. This is probably because investigators are unaware of the availability of auxiliary variables with high correlation with current survey observations. Survey data from previous years are commonly available, but this by itself does not meet the requirements of the estimators. The auxiliary variable must be defined for all sampling units in the current survey, and the mean of the auxiliary variable (taken over all possible sampling units) must be known. By constructing a model of the fishery, one creates an auxiliary variable with the required characteristics. This is a common variance reduction technique in simulation work where, for example, a crude model may be used to obtain analytical solutions that have a sufficiently high correlation with the outputs from a complicated numerical model to provide a reduction in variance for the estimates from the numerical model (see Law and Kelton 1982).

Acknowledgments

We thank Nadine Hoenig, Geoffrey Evans, and George Winters for helpful comments and the personnel of the Minnesota Department of Natural Resources for their outstanding efforts to make the fishery survey a success. Partial support for this study was received from the U.S. Department of Commerce, Chesapeake Bay Stock Assessment Committee, through grant NA89EA-H-00060 to Cynthia M. Jones, Applied Marine Research Laboratory, Old Dominion University, Norfolk, Virginia.

References

Abramson, N. J. 1990. A test of using wind velocity data with a product estimator to improve the efficiency of sport salmon landings estimator. U.S. National Marine Fisheries Service, Southwest Fisheries Center Administrative Report T-90-04, La Jolla, California.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical methods for data analysis. Duxbury Press, Boston.

Cochran, W. G. 1977. Sampling techniques, 3rd edition. Wiley, New York.

Grimes, J. E., and B. V. Sukhatme. 1980. A regression-type estimator based on preliminary test of significance. Journal of the American Statistical Association 75:957–962.

Hoenig, J. M., F. B. Martin, and C. M. Heywood. 1986. Model-based sampling methods for effort and catch estimation. International Council for the Exploration of the Sea, C.M. 1986/D:15, Copenhagen.

Hoenig, J. M., F. B. Martin, and C. M. Heywood. 1989. Model-based sampling methods for effort and catch estimation. American Fisheries Society Symposium 6:181–189.

Law, A. M., and W. D. Kelton. 1982. Simulation modeling and analysis. McGraw-Hill, New York.

Sukhatme, P. V., B. V. Sukhatme, S. Sukhatme, and C. Asok. 1984. Sampling theory of surveys with applications. Iowa State University Press, Ames.