Reducing Computation Time for the Analysis of Large Social Science Datasets

Douglas G. Bonett
Center for Statistical Analysis in the Social Sciences
University of California, Santa Cruz

Jan 28, 2014

Overview

Electronic data capture of human behavior (e.g., customer transactions, internet search activity, social media) can produce datasets with billions or trillions of records. Our capacity to collect and store data is now increasing faster than our data processing capacity.

A few ways to reduce computation time will be explained:

- how to analyze a very small fraction of the data to obtain virtually the same results that would be obtained from the entire dataset
- how to quickly analyze online experiments involving many factors
- how to quickly estimate the parameters of a statistical model with a large number of explanatory variables to obtain near optimal predicted values

Random Sampling

Estimates of parameters such as means, standard deviations, proportions, correlations, slope coefficients can be obtained from a random sample of n records that will be virtually identical to the values that would have been obtained by analyzing all *N* records of the entire dataset.

Naive sample size rules such as n = *N*/1000 (Varian, 2013) are not recommended – they give sample size values that can be too small and often much too large.

Instead, the appropriate sample size will allow estimation of some parameter with very high confidence (99.99%) and very high precision.

Estimating a Proportion

The sample size (n) needed to obtain a 99.99% confidence interval for p with a "narrow" width of w is

$$n = 4\tilde{p}(1 - \tilde{p})\left(\frac{3.72}{w}\right)^2$$

where $\tilde{p}$ is a planning value of p. This planning value can be obtained from prior studies, expert opinion, or a small pilot random sample of records.

What is considered "narrow" will depend on the type of application.

Example

Suppose there are about 100 billion internet searches for electronic equipment during some time period and we want to determine the proportion (p) of these searchers that lead to an Amazon web site. Setting $\tilde{p}$ = 0.01 and w = .0002, the required number of records to randomly sample is:

$$n = 4(.01)(.99)\left(\frac{3.72}{.0002}\right)^2 \approx 13.7 \text{ million}$$

and p can be estimated from this sample about 7,300 times faster than from the complete data set.

Example (*continued*)

In the above example with N = 100B records, one analyst computed the proportion of interest and obtained p = .2180. A second analyst who needed the answer more quickly (i.e., 7,300 times faster) estimated the proportion from a random sample of n = 13.7M and obtained an estimate of .2179 which is close enough to the true proportion for all practical purposes.

In applications like these, it is often necessary to estimate many (q) different proportions, and the second analyst would have the required results 7,300(q) times quicker.

Estimating a Mean

The sample size needed to obtain a 99.99% confidence interval for the mean of a quantitative attribute (y) in all N records is

$$n = 4\tilde{\sigma}^2(3.72/w)^2$$

where $\tilde{\sigma}^2$ is a planning value of the variance of y. This planning value can be obtained from a pilot random sample of 1,000 records and the estimate can be multiplied by 1.16 (assuming a pilot sample of 1,000) to give a 99.99% upper planning estimate of the variance.

The value of w depends on the scale of y. For example, if y is measured in dollars, w might be set to 1 or .1 (10 cents).

Example

We want to estimate the mean price for all 250 million used book on-line purchases during a specific time interval. Using a pilot random sample of 1,000 purchases, the pilot sample variance was 25.21 which gives a 99.99 upper planning value of 29.24. The required sample size to obtain a 99.99% confidence interval for the mean price of all 250 million purchases with a width of 10 cents is

$$n = 4(29.24)(3.72/0.1)^2 \approx 161,854$$

which can be computed about 1,500 times faster than the mean of all 250 million purchases.

Estimating a Correlation

The sample size needed to obtain a 99.99% confidence interval for the Pearson correlation between quantitative attribute y and quantitative attribute x in the complete dataset is

$$n = 4(1 - \tilde{\rho}^2)^2\left(\frac{3.72}{0.01}\right)^2$$

where $\tilde{\rho}^2$ is a correlation planning value. This planning value can be estimated from a pilot sample of 1,000 records with 0.01 subtracted to give a 99.99% lower planning estimate of the correlation. Or set $\tilde{\rho}^2 = 0$ to give a conservatively large sample size.

Example

We want to determine the correlation between the dollar amount of a recent online purchase with the amount of the customer's previous purchase in a database of 500 million transactions. Setting $\tilde{\rho}^2 = 0$ gives a sample size requirement of

$$n = 4(3.72/0.01)^2 \approx 553,536$$

which will be about 900 times faster than computing the correlation from the complete dataset.

Estimating a Residual Standard Deviation

The sample size needed to obtain a 99.99% confidence interval for a residual standard deviation in a linear statistical model with q explanatory variables is

$$n = 2\left[\frac{3.72}{ln(r)}\right]^2 + q$$

where r is the ratio of the desired upper and lower interval estimates. Set r = 1.02 for an extremely narrow confidence interval.

Example

Suppose a company has a database of 700 million online customer transactions and wants to predict the purchase amount using about 100 customer characteristics as explanatory variables. Instead of fitting a regression model to all 750 million cases, the model can be fit very quickly to a random sample of transactions such that the 99.99% confidence interval for the residual standard deviation has an upper to lower endpoint ratio of 1.02. The regression model can be fit to a random sample of

$$n = 2[3.72/ln(1.02)]^2 + 100 \approx 70{,}578 \text{ cases}$$

which would be about 10,000 times faster than analyzing the complete dataset.

Second-stage Sampling

In some applications, the 99.99% confidence interval computed from the random sample of size n might be wider than the desired value. In these situations, a second random sample can be taken from the complete dataset and added to the original sample. The number of additional records to sample in the second-stage is

$$n^{+} = n\left[\left(\frac{w_0}{w}\right)^2 - 1\right]$$

where width w_o is the observed confidence interval width in the initial sample of size n.

Example

A random sample of 200,000 insurance claims was taken from a large database of 300 million claims. A linear regression model was fit to the sample data but some of the slope confidence intervals were wider than desired. In the worse case, the obtained CI width was 3.5 and the desired CI width was 2.0. The number of additional claims to sample is

$$100{,}000[(3.5/2.0)^2 - 1] = 206{,}250$$

to give a total sample size of n = 305,250. Computation time in the sample will be about 983 times faster than in the complete dataset.

On-line Experiments

User data can be collected over time under randomly assigned treatment conditions to assess the causal effects of various factors (e.g., price discounts, screen formats, online advertisements). Rapid identification of important factors can result in increased profits and customer satisfaction. The goal is to obtain accurate estimates of treatment effects in a shorter time frame.

It these applications, it is may not be necessary to estimate effects with extremely high confidence – 95% confidence may be adequate. It also may not be necessary to obtain extremely high precision – choose a desired width that provides enough accuracy for a useful cost-benefit analysis.

Online Experiments (*continued*)

Most treatment effects can be expressed as a linear contrast of means when the response variable is quantitative or a linear contrast of proportions when the response variable is dichotomous. Sample size formulas to estimate a treatment effect with desired confidence and precision are given below

means $$n_j = 4\tilde{\sigma}^2\left(\sum_{j=1}^{k} c_j^2\right)\left(z_{\alpha/2}/w\right)^2$$

proportions $$n_j = 4\left[\sum_{j=1}^{k} c_j^2 \tilde{\pi}_j\left(1-\tilde{\pi}_j\right)\right]\left(z_{\alpha/2}/w\right)^2$$

where $z_{\alpha/2}$ = 1.96 for 95% confidence.

Fractional Factorial Designs

Many companies evaluate treatment factors one factor at a time. Multiple treatment factors can be assessed over a shorter time frame using Fractional Factorial Designs of Resolution 3.

2^v Resolution 3 designs exist for v = 3, 7, 11, 15, 20 and other values of v for which a Hadamard matrix of order $v + 1$ exists.

The v factors in a Resolution 3 design can be estimated with the same precision as a one-factor study using the same sample size (time frame). A 2^v Resolution 3 design requires only a fraction of the treatment combinations and can obtain the desired results v times faster than v single-factor studies.

Resolution 4 and 5 Designs

A Resolution 3 design with large v could be used to screen for the most important main effects. Interaction effects cannot be assessed in a Resolution 3 design.

The experimental design could then be changed to Resolution 4 using only the most important factors. A Resolution 4 design can assess main effects and interaction effects, but the interaction effects cannot be uniquely estimated.

If interactions are detected, the design could be changed to Resolution 5 which allows unique estimation of all 2-way interaction effects.

Statistical Models with Many Variables

In a dataset with n observations and q explanatory variables, the predicted response vector from a linear model is

$$\hat{y} = X\hat{\beta} \text{ where } \hat{\beta} = (X'X)^{-1}X'y$$

Computational time for $\hat{\beta}$ depends mainly on the computational time for $(X'X)^{-1}$ which is proportional to about $q^{2.4}$. Computation of $\hat{\beta}$ can be slow if q is large.

Also, both n and $(1 - \rho_j^2)$ are in the denominator of $SE(\hat{\beta}_j)$. With large q, ρ_j^2 (the squared multiple correlation between explanatory variable j and all other explanatory variables) can be close to 1 so that n needs to be enormous to keep $SE(\hat{\beta}_j)$ at an acceptable value.

Approximate Estimation Methods

Wilks (1938) showed that standardizing the explanatory variables and setting $\hat{\beta}_j = 1$ ("unit weighting") gave surprisingly accurate predictions of y. Others have shown that estimating each $\hat{\beta}_j$ from a simple linear regression model provides more accurate predictions of y than unit weighting.

The *OLS* estimate of β from the full model is not always substantially better than simple approximations in terms of minimizing the residual variance.

Another line of research has shown that a linear function of multiple forecasts is superior to the individual forecasts.

Applying both of these ideas leads to the following approach:

Approximate Estimation Methods (*continued*)

Instead of computing $\hat{y} = X(X'X)^{-1}X'y$, compute

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1\hat{y}_1 + \cdots + \hat{\theta}_r\hat{y}_r$$

where $\hat{y}_k = X_k'(X_k'X_k)^{-1}X_k'y$, $\hat{\theta} = (\hat{Y}'\hat{Y})^{-1}\hat{Y}'y$, X_k' is n x p, $\hat{Y}$ is n x (r+1), and r x $p \geq q$.

Example: With q = 1,000 we might set p = 50 and r = 20. Inverting 20 p x p matrices (to get $\hat{y}_k$) plus one 21 x 21 matrix (to get $\hat{\theta}$) is much faster than inverting one 1000 x 1000 matrix.

Speed Advantage

Here are some optimal values of p and r for different values of q.

q	p	r	speed
100	6	17	27 times faster
1,000	15	67	226
10,000	35	286	1,773
100,000	79	1,266	13,636
1,000,000	187	5,348	104,550
1,000,000,000	2,387	418,936	46,997,992

Accuracy of Proposed Approximation

The proposed approximation should be superior to the unit weighting method. The *OLS* estimates from each set of p explanatory variables should be good approximations to the *OLS* estimates from the full model. Furthermore, the estimates of θ provide optimal combinations of the r predictions of y.

Although the *OLS* estimates from the full model should theoretically minimize the residual variance, the proposed $\hat{y}$ approximation might give a smaller residual variance as a result of numerical inaccuracies and RAM corruptions during the computation of $(X'X)^{-1}$ in the full model if q is extremely large.

Example with small *q*

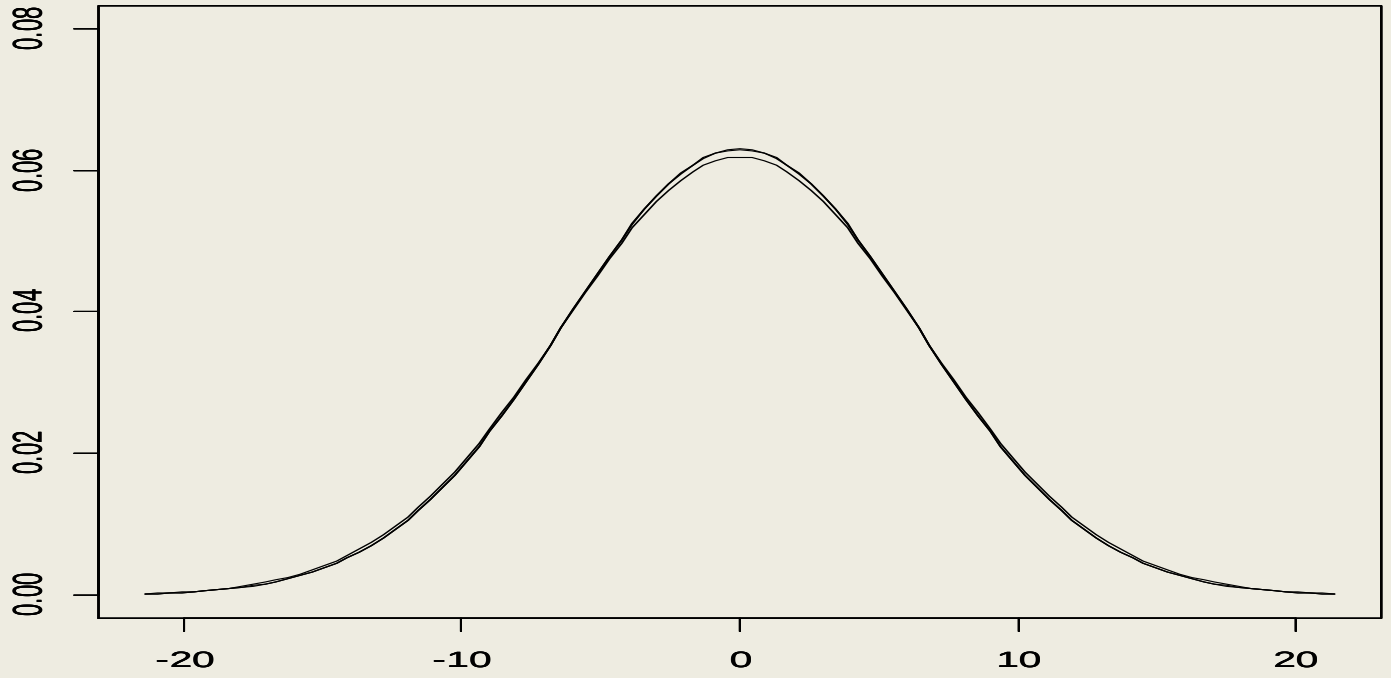
y = life satisfaction with q = 9 personality predictor variables (n = 365)

Comparison of optimal (*OLS* full model) with unit weighting and proposed approximation using r = 3 and p = 3

	$\sqrt{MSE}$
optimal	6.33
approximation	6.34
unit weights	6.45

Comparison of Residual Distributions ($\sqrt{MSE}$ = 6.33, 6.34, 6.45)

Note that the three distributions are virtually indistinguishable.



Conclusion

Statistical and experimental design methods are indispensable tools for the rapid analysis of extremely large datasets. The information obtained from these analyses can be used to improve existing products and services, develop promising new products and services, increase profits, and better meet the needs of an ever-changing customer database.

Thank you.