# Generalized Variance Functions for Radio-schedule Gross Rating Point Estimators

Richard Griffiths
Arbitron, Inc.

## Abstract

Arbitron provides users with measures of sampling variance through a form of generalized variance function. In particular, to estimate standard errors for radio-schedule Gross Rating Points estimators, Arbitron currently models the standard error as a function of the GRP estimate and sample size, allowing for different model parameters across a number of demographic groups and radio-schedule characteristics. In this paper, we evaluate several GVFs for estimating standard errors of radio-schedule GRP estimators.

**Keywords**: Generalized variance function; Complex sample design; Gross rating points.

## 1. Introduction

In the mid-1990's, both radio broadcasters and advertising agencies became very interested in the precision of Arbitron's radio schedule audience estimators. To provide them with the information necessary to understand the precision of these estimators, Arbitron conducted a study of their sampling error. See Arbitron (1995).

From this study came formulas that radio industry users could apply to estimate the precision of radio-schedule Gross Rating Points (GRPs) estimators. (Loosely speaking, GRPs are the percent of people that were listening to the radio when a radio advertisement ran multiplied by the number of times the advertisement ran in a given time period.) These formulas take the form of generalized variance functions – given a set of basic input (GRP estimate, sample size, and some information related to the advertising schedule), the user can calculate an estimated standard error and associated confidence interval for a range of radio-schedule GRP estimates.

In the last year or so, the radio industry has again become interested in the precision of radio-schedule GRPs. This time around, the American Association of Advertising Agencies requested that Arbitron develop a software tool that could be used by the industry to easily calculate GRP estimator precision. To meet this request, Arbitron has developed a tool which implements the formulas given in Arbitron (1995).

For advertisers, the precision of the GRP estimates is important, because they need to evaluate whether the estimated size of the audience they have reached is within sampling error of what they expected when they first purchased the radio spot. In other words, they want to know if they got their money's worth, and they know that, to do so, they have to account for the sampling error inherent in the GRP estimates. More generally, the importance of variance estimation in survey research is witnessed by the temporal length and topical breadth of the statistical literature on the subject. See, for example, Wolter (1985) or the discussions on the topic in Kalton (1977), Sarndal et al. (1992), and Valliant et al. (2000). An example of an early need for reliable variance estimates is given in Chameleon (2001).

Given the general importance of understanding precision and given that the original study is ten years old, Arbitron decided to undertake a new study of the issue. The two practical reasons for undertaking this study are:

- To update the parameter estimates given in Arbitron (1995).
- To determine if there are more appropriate models for generating standard error estimates.

This paper focuses on the second reason – it is a study and empirical comparison of several potential standard error models.

## 2. Background

### 2.1 Brief Review of Arbitron's Radio Market Surveys Methodology

To produce estimates of radio listening audiences in the United States, Arbitron divides the country into about 300 geographical areas called markets. Arbitron then conducts a survey of an RDD sample in each market. Each survey is conducted over a 12-week period. About 100 of the markets are surveyed four times per year; the others are surveyed two times each year.

To ensure the selected sample represents the demographic and geographic characteristics in each market, Arbitron uses a raking methodology to weight the sample to the population. Based on these weights, expansion estimators of the numbers of people who listen to the various radio stations and the amount of time people spend listening, among other radio listening behaviors, are then constructed.

### 2.2 GRPs Explained

As we mentioned earlier, GRPs loosely correspond to the percent of people that were listening to the radio when an advertising spot ran times the number of times the advertisement ran. To be a little more precise with the definition, we have to discuss the concept of daypart. A daypart can be taken literally to be a part of a day, though, it usually refers to the part of a day over an entire week. So, for instance, 6am to 10am is part of a day. In the radio business, the corresponding daypart might be 6am to 10am, Monday through Friday. This is colloquially known as the "morning drive." Additionally, one can break each daypart up into quarter-hours. The important concept here is that there are 80 quarter hours (4 QH/hour x 4 hours/day x 5 days in the daypart) in the morning drive daypart.

So, let's say an advertiser ran a spot on a given station once per hour for each hour in the "morning drive" daypart. They would have then run 20 spots for that daypart. (This is the schedule.) From its survey, Arbitron estimates the average number of people that were listening to the station at any given quarter-hour in the "morning drive" daypart. This yields something called the estimated AQH rating – the average number of people listening during any quarter hour divided by the population that could be listening in a given market. If this estimated AQH rating was, say, one percent of the population for the morning drive for the station running the advertiser's spot, then the GRP estimate is 20 times 1.00, or 20.

### 2.3 GRP Math

For those who might get more out of an equation than the above explanation, the estimator of the GRP for a radio-advertising schedule in a given market $m$ can be expressed as

$$\hat{X}_{mdp} = 100 \cdot \frac{1}{N_{md}} \cdot s \cdot \sum_{i=1}^{n_{md}} \sum_{l=1}^{q_p} w_{mdi} \cdot I_{mdil} / q_p \, , \quad (1)$$

where $m$ indexes the market a radio station resides in (e.g., Minneapolis-St. Paul); $d$ indexes the demographic subgroup (e.g., males between the ages of 35 and 54) of the market's population; $p$ indexes the daypart (e.g., the "morning drive"); $N_{md}$ is the number of people in demographic group $d$ in market $m$; $s$ is the number of spots that ran in the radio-advertising schedule; $n_{md}$ is the sample size for a demographic group $d$ in market $m$; $q_p$ is the number of quarter-hours in daypart $p$; $w_{mdi}$ is the survey weight for respondent $i$ in market $m$ and demographic subgroup $d$; $I_{mdil}$ is an indicator function, which is one if respondent $i$ in demographic subgroup $d$ was listening to the given radio station during the $l^{\text{th}}$ quarter-hour of the daypart.

## 3. Generalized Variance Functions

Since the application that drives the research for this study is a tool that can estimate standard errors for a broad spectrum of estimators, the GVF methodology suits the purpose well. A table of GVF parameter estimates can be passed to the tool and a formula programmed into the tool to produce standard error estimates given user-input that determines which table entries to use in the calculation.

GVF theory is discussed in Wolter (1985) and Valliant et al. (2000). Below, we briefly present the various GVFs we examine in this study.

### 3.1 GVF Models for Study

The model Arbitron currently uses to estimate variances for GRP estimators can be written as:

$$\hat{V} = \hat{X} \cdot (100 \cdot s - \hat{X}) / n \cdot \tau \, , \qquad (2)$$

where $\hat{X}$ is the GRP estimate; $s$ is the number of spots; $n$ is the sample size; and $\tau$ is a parameter to be estimated. (See Arbitron, 1995.)

A model that is prevalent in the GVF literature, the one that is often given as the theoretical basis for GVF methodology, is the one used in the Current Population Survey (CPS). This GVF models the relationship of the relative variance of an estimator $\hat{X}$ to the expected value of the estimator:

$$\frac{V}{X^2} = \alpha + \beta / X \text{ , or} \qquad (3)$$

$$V = \alpha \cdot X^2 + \beta \cdot X \, ,$$

where $X = E(\hat{X})$ and $V$ is the variance of $\hat{X}$.

See Valliant et al. (2000, pp. 344-347), Wolter (1985, pp. 202-205), Hansen, Hurwitz, and Madow (1953, 571-577) and U.S. Department of Labor (2002, pp. 14-3 to 14-5).

An adaptation of the CPS GVF to include sample size in the model is

$$V = \alpha \cdot X^2 / n + \beta \cdot X / n \qquad (4)$$

This form was used by Otto and Bell (1995) in modeling CPS state-level standard errors.

Both (3) and (4) are very similar to the current Arbitron model (2). With $\alpha = -\frac{1}{\tau}$ and $\beta = \frac{100s}{\tau}$, (4) has the same form as (2). (3) and (4) offer a degree of flexibility over (2), however, by allowing unrelated parameters for the linear and quadratic terms.

Griffiths and Mansur (2001) extended (4) to include a term for a lagged variance estimate:

$$V_t = \alpha \cdot X_t^2 / n_t + \beta \cdot X_t / n_t + \gamma \cdot V_{t-1}, \quad (5)$$

where the subscript *t* denotes the current survey and *t*-1 the previous survey.

In this study, we consider a related model:

$$V_t = \alpha \cdot X_t^2 / n_t + \beta \cdot X_t / n_t + \gamma \cdot X_{t-1} / n_{t-1} + \delta \cdot X_{t-1}^2 / n_{t-1} \quad (6)$$

We also consider a number of other GVF models, each a variation on (4), (5), or (6):

- $\frac{V}{X^2} = \frac{\alpha}{n} + \frac{\beta}{nX} + \frac{\gamma}{n\sqrt{X}}$ (7)
- $\ln(V) = \alpha + \beta \ln(X)$ (8)
- $\ln(V) = \alpha + \beta \ln(X) + \gamma \ln(n)$ (9)
- $\ln(V_t) = \alpha + \beta \ln(X_t) + \gamma \ln(X_{t-1}) + \delta \ln(n_t) + \tau \ln(n_{t-1})$ (10)

We consider the logarithmic transformation primarily to help limit the influence of extreme values on the estimated parameters; we are interested in how this affects model fits. As will be discussed below, we aren't greatly concerned with assumptions of normality and homoscedasticity in model fitting – two typical reasons for using transformations.

In the following sections, we refer to the studied models by a model number. The following table associates that model number with the above GVF forms.

| Model Number | Equation Number and Model Description |
|---|---|
| 1 | (2) with 1995 parameter estimates |
| 2 | (2) with updated parameter estimates |
| 3 | (3) |
| 4 | (4) |
| 5 | (6) |
| 6 | (7) |
| 7 | (8) |
| 8 | (9) |
| 9 | (10) |

### 3.2 Model Fitting

To fit the models, we first classified the estimates into several groups. These groups were delineated by the cross-classification of the 22 demographic groups and eight daypart groups given in the original Arbitron study.[1] It is assumed that within each of these groups, the same GVF model applies. Each model was fit to the standard error estimates for each of these 176 groups.

Any of the GVFs studied might be fit using a least-squares technique. The linear models can be fit using ordinary (or, weighted) least-squares regression. This guarantees certain "nice" properties for the models, like minimum variance and conditional unbiasedness. (Rencher, 2000). A nonlinear model, like (2), can be fit with a numerical least-squares algorithm (or, with a minimization of any appropriate objective function for the errors).

If one wishes to do some model-building and make assessments of the significance of terms in the model, one can assume normality of the variance estimators, or an appropriate transformation thereof, and fit the models under maximum likelihood estimation. One could also assume the dependent variable has a non-normal distribution (e.g., a Gamma) and fit a generalized linear model, under maximum likelihood estimation. (See, for example, Griffiths and Mansur, 2001.) In the case of the GVF with lagged dependent variables, we would fit the model using partial maximum likelihood estimation.

Alternatively, if the primary property required for the models is that they produce standard error estimates that are "close" to actual standard errors in, say, an absolute relative deviation sense, there is no need to limit ourselves to least-squares estimation or MLE, since the desired properties are not those guaranteed by least-squares or MLE. In this case, we might minimize the sum of the absolute relative deviations to obtain estimators with the "optimal" property in our particular situation.

However, the key fact that plays into our consideration of model-fitting methods is that we view under-estimation as a more serious error than over-estimation for standard errors. From our preliminary analysis of the data, it appears that model 1 (the current Arbitron GVF with old parameter estimates) allows for more under-estimation than we would like. We want to try to improve in this area with the updated model.

This leads us to think about fitting the models using the ideas of quantile regression. For quantile regression applied to the problem at hand, the objective is to find the model parameters that minimize the following function:

$$\sum_i \rho_\tau \left( se_{jk,i} - se_{gvf,i} \right)$$

[1] A daypart group is made up of one or more dayparts – dayparts with similar number of quarter-hours are grouped together.

where $se_{gvf,i}$ is the GVF estimate of the standard error $S_i$, $se_{jk,i}$ is the direct jackknife estimate of the standard error, and $\rho_\tau(x) = x \cdot (\tau - I\{x > 0\})$; $I\{x > 0\} = 1$ *if* $x > 0$*; 0 otherwise.* (See Koenker and Hallock, 2001.) Here the idea is to use an estimate of the expected percentile of the conditional distribution as the standard error estimate, instead of the conditional mean, as would be the case with an ordinary least squares fit. For instance, one might choose to use the 60th percentile of the conditional distribution as the modeled standard error to ensure more over-estimation than under-estimation. This model-fitting method allows us to penalize the models more for an under-estimation error than for an over-estimation error.

This idea is similar to that used by Johnson and King (1987), who assumed that "the consequences of an underestimate are three times as severe as those of an overestimate." In fitting their models, Johnson and King (1987) assumed a normal conditional distribution and used the conditional expected mean plus .67 times the standard error (the 75th percentile of the normal distribution) as their "optimal prediction."

In what follows, we examine and compare the fits of the various GVFs by fitting the models to yield estimated standard errors that are the estimated 60th percentile of the conditional distribution. This allows us to appease our conservative tendencies by bringing the under-estimation issue under better control. It also provides an equitable basis for comparing the various GVFs.

## 4. Comparing Models/Evaluation

### 4.1 Study Methodology

We chose 60 Arbitron markets and survey data from four quarterly surveys to form the basis of our empirical study. We calculated variance estimates, and corresponding standard error estimates, for thousands of possible radio-schedules in each market/survey combination using a jackknife variance estimation methodology. We also calculated AQH and GRP estimates calculated for each radio-schedule. In this study, we examine only hypothetical radio-schedules that consisted of running advertising spots on only one station. In practice, the methodology also needs to apply to multiple-station schedules.

We fit each of the GVF models with the jackknife standard error estimate as dependent variable and GRP estimate, sample size, number of spots as independent variables. Each model was fit for estimates within each demographic and daypart group. This allows for different parameters for each demo and daypart group.[2] Even though we fit the models by daypart group, we look at results by daypart in what follows. This allows us, among other things, to examine whether the particular daypart groupings we used make sense.

Finally, we compared the various GVFs on their ability to model the jackknife standard error estimates using several different measures.

### 4.2 Measuring the Ability To Estimate Standard Errors

A standard error estimate is the quantity that will be reported to the user. Thus, it is important that we compare the models on ability to estimate the standard error. The following are the measures that we use in the comparisons:

- The ratio of the GVF standard error to jackknife standard error, by demo and daypart:

  $$r_i = \frac{se_{gvf,i}}{se_{jk,i}},$$

  where $i$ indicates the ith estimate within a given demo and daypart. In particular, we will examine and compare the distributions of these ratios for each GVF model. Examination of the distributions allows us to see how well the modeled standard errors cluster around the jackknife standard error and give an indication of a GVF's propensity to over- and under-estimate.

- The absolute relative deviation (ARD) of the GVF standard error from jackknife standard error, by demo and daypart:

  $$ard_i = \left| \frac{se_{gvf,i} - se_{jk,i}}{se_{jk,i}} \right|$$

  We include this measure as it tells us something about how far the modeled standard error is from the jackknife standard error without allowing over and under-estimation errors to cancel out. Both Krenzke and Navarro (1996) and Jang et al. (2000) use a form of absolute relative deviation to compare methods.

- The propensity of the models to under-estimate standard errors.

  Recalling our conservative viewpoint, we want to take a very specific look at the propensity of the models to give modeled standard errors that are

[2] The daypart groups we used in this study are those given in the original Arbitron (1995) study.

large under-estimates – 20 percent or more of the jackknife standard error. We also look at the propensity to under-estimate in general.

One might then state the primary criteria for model evaluation as:

- We don't want to under-estimate the standard errors too badly too often.
- We want our modeled standard errors to cluster as much as possible around the directly-estimated jackknife standard errors.

The first criterion is satisfied if a small fraction (less than 10 percent or so) of a model's standard error estimates are large under-estimates. In fact, we fit the GVF models with this in mind when we used the estimated 60th percentile of the conditional distribution as the standard error estimates.

To establish which GVFs do best with respect to the second criterion, we use box-and-whisker plots of the ratio of the GVF standard error to jackknife standard error, as well as mean and median values of the ARD.

### 4.3 Results

An initial glance at the results of fitting the models showed that we could eliminate some of the models from contention quite easily. We saw that the CPS GVF tended to produce modeled standard errors that were more variable than those of models that included sample size in their specification. So, while on average the CPS GVF estimates are at about the same level as many of the other models, this GVF tends to give larger under- and over-estimates.

This result makes sense, because the GRP estimate is essentially a proportion times the number of spots. Thus, its level tends not to vary with the size of the population base. The sample size does, however, tend to be related to the population base – it has a positive correlation with it. Thus, we lose some information from the model when we don't include the sample size in this situation. Model 3 is intended to work in situations where most of the information given by the sample size is subsumed by the characteristic estimator itself. For instance, it works well with estimators of totals, for which the level of the estimator tends to vary with the population base (and sample size). So, we immediately can eliminate models 3 and 7 from contention.

The second thing that stood out immediately is that using the logarithmic transformation adds little to the ability of the models to fit the jackknife standard error estimates. In fact, the logarithmic transformation makes the fit worse on our primary measures in many instances. This further eliminates models 8 and 9 from consideration. Thus, we focus our attention on models 2, 4, 5, and 6.

In Attachment A, we provide some box-and-whisker plots. These plots display the distribution of the ratio of modeled standard error to jackknife standard error estimate for models 2, 4, 5, and 6 for Persons 12+ for individual dayparts. These plots give a graphical depiction of the ability of each of these models to produce standard error estimates that cluster around the jackknife standard error.

From these plots, one can see that model 5 has the greatest ability to produce standard error estimates that cluster around the jackknife standard error. Models 4 and 6 have similar ability in this respect and only slightly greater ability than model 2. In general, one might also note that there is not a great difference in clustering ability across these models.

Attachment B displays a graph of the median ARD for each of the models for Persons 12+ across all 18 dayparts. This graph shows that models 4, 5, and 6 offer some reduction over model 2 in median ARD, with model 5 giving slightly lower ARD, in general. Averaging percent reductions in median ARD over all dayparts, we find:

- Model 5 provides a 13 percent reduction in median ARD on average over the 18 dayparts for Persons 12+ (11 percent, ignoring daypart 5).
- Model 4 provides an eight percent reduction in median ARD on average over the 18 dayparts for Persons 12+ (seven percent, ignoring daypart 5).
- Model 6 provides a nine percent reduction in median ARD on average over the 18 dayparts for Persons 12+ (eight percent, ignoring daypart 5).

Attachment B also displays a graph of the percent of modeled standard errors that are large under-estimates for models 2, 4, 5, and 6 for Persons 12+ across all dayparts. In general, we see that model 2 has a slightly smaller percentage of standard error estimates that are large under-estimates, but that the percentage is comparable over models.

Results for other dayparts vary, but, considered in aggregate, give generally the same impression as those for the broad Persons 12+ demographic group.

A caveat from our results is that some model fits are much better for some dayparts than for others in certain daypart groups. For instance, daypart 5, which is in the same daypart group as daypart 3, shows relatively poor model fits. This suggests that daypart 5 needs either a separate set of parameter estimates in the GVF models or belongs with another daypart group.

## 5. Discussion

The premise of this study was that that the parameter estimates for model 1, the current Arbitron GVF, needed to be updated. Additionally, beyond needing parameter estimate updates, model 1 allows for too much under-estimation for our conservative nature. Thus, we want to update model 1 for two reasons:

- to provide current parameter estimates for the model;
- to reduce the propensity of the model to under-estimate.

The second objective is achieved by fitting model 2 using the 60th percentile of the conditional distribution as the modeled standard error.

That done, we then focus on whether the current methodology can be improved upon. To this end, we compared several other GVF models (models 3 through 9) to model 2. The question we want to answer is, Do any of them provide a markedly better fit than model 2?

The answer is that some of the models we examined do provide modest, but tangible, improvement over model 2 in their ability to cluster standard error estimates around the jackknife standard error estimates. In particular, model 5, which uses lagged GRP estimates, has an advantage over model 2 in clustering ability. It also provides for some reduction in ARD.

In practice, model 5 has the drawback that a user would need to know the value of the GRP from the previous survey to calculate an estimated standard error. While many users would, in fact, have these GRP estimates available, a sizable portion would not, or would consider it too burdensome to access this estimate. This potentially limits the practicality of this model.

Models 4 and 6 also provide some slight benefit over model 2. These models are more likely to be candidates for replacing the current methodology, though, more extensive research (including, for multiple-station schedules) is needed to determine the appropriate form of model and the full range and magnitude of advantages of these models.

## References

Arbitron (1995), *Radio-Schedule Audience Estimate Reliability*, The Arbitron Company.

Chameleon, T. (2001), The Real Reason, *Chance*, 14, p.64.

Griffiths, R. and K. Mansur (2001), Current Population Survey State-Level Variance Estimation, *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, November 2001.

Hansen, Hurwitz, and Madow (1953), *Sample Survey Methods and Theory, Volume I*, New York: Wiley.

Jang, D., J.K. Garrett, F.W. Piotrowski, and W.B. Owens (2000), Generalized Variance Function Methodology for ACNielsen's Homescan Household Panel Survey, *Proceedings of the American Statistical Association, Survey Research Methods Section*, 811-815.

Johnson, E.G. and B.F. King (1987), Generalized Variance Functions for a Complex Sample Survey, Journal of Official Statistics, 235-250.

Kalton, G. (1977), Practical Methods for Estimating Survey Sampling Errors, *Bulletin of the International Statistical Institute*, 47, 495-515.

Koenker, R. and K.F. Hallock (2001), Quantile Regression*, Journal of Economic Perspectives*, 143-156.

Krenzke, T.R. and A. Navarro (1996), Sampling Error Estimation in the 1995 Census Test for Small Areas, *Proceedings of the American Statistical Association, Survey Research Methods Section*, 752-757.

Otto, M.C. and W.R. Bell (1995), Sampling Error Modelling of Poverty and Income Statistics for States, *Proceedings of the Section on Government Statistics, American Statistical Association*, 160-165.

Rencher, A.C. (2000), *Linear Models in Statistics*, New York: Wiley.

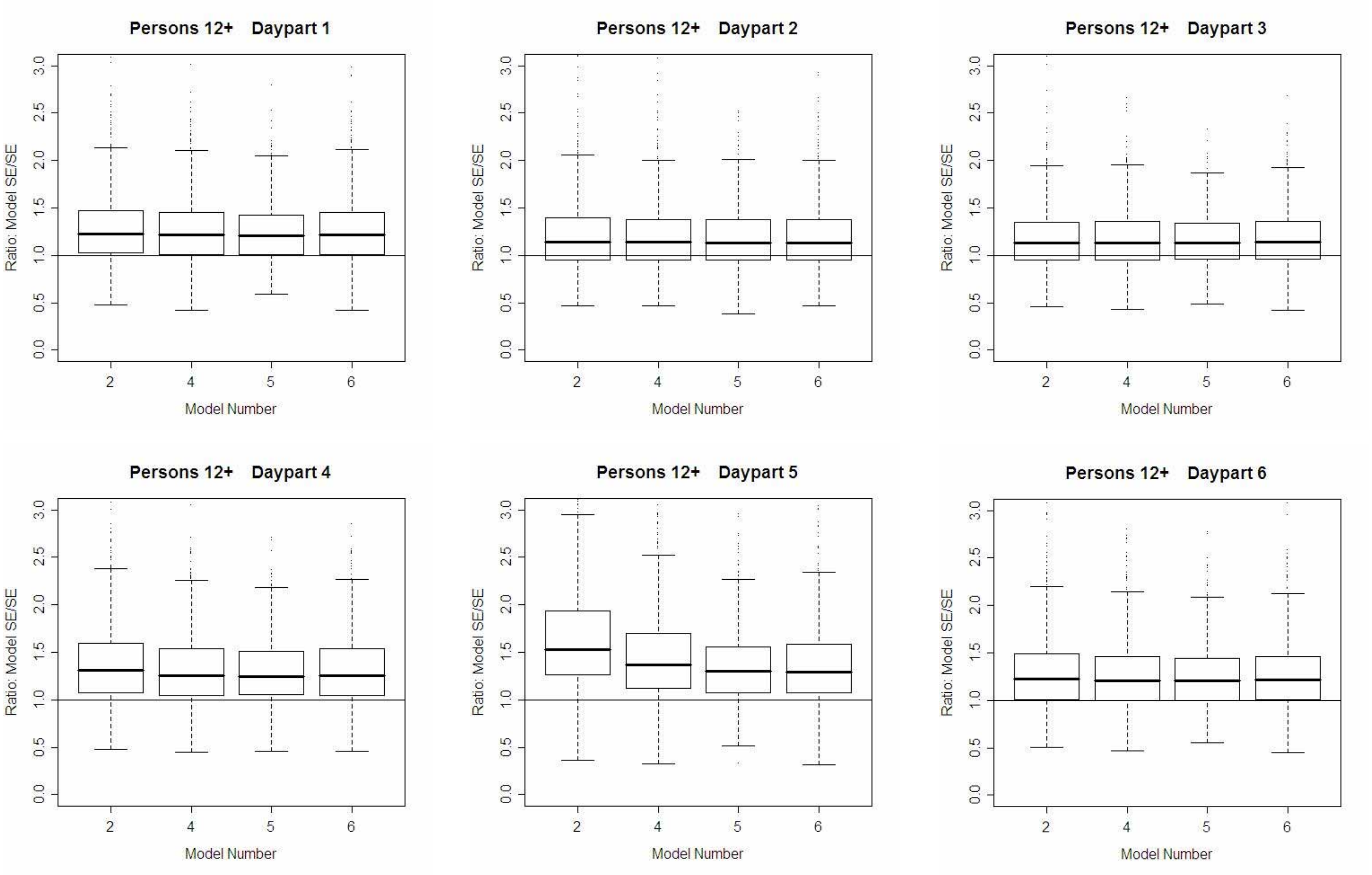
Sarndal, C.-E., B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling*, New York: Springer.

United States Department of Labor (2002), *Current Population Survey, Design and Methodology, Technical Paper 63RV*, Washington, DC.

Valliant, R., A. H. Dorfman, and R.M. Royall (2000), *Finite Population Sampling and Inference, A Prediction Approach*, New York: Wiley.

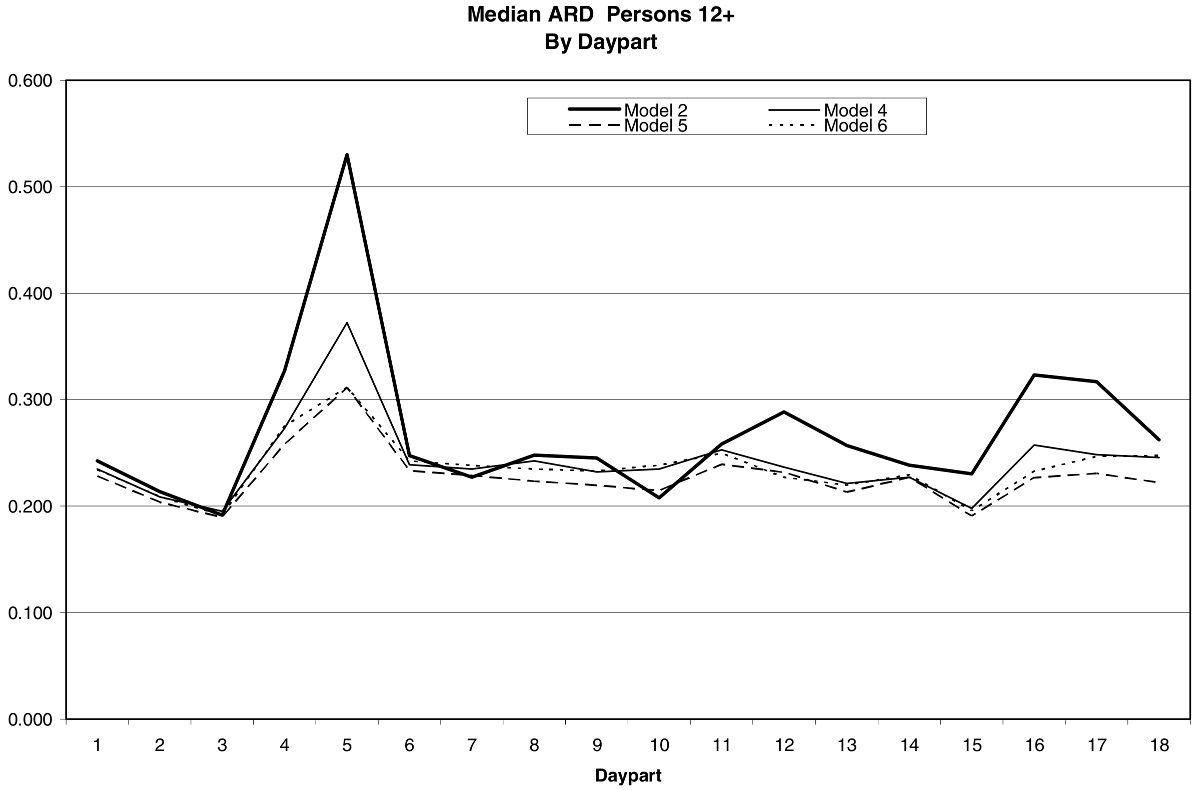
Wolter, K.M. (1985), *Introduction to Variance Estimation*, New York: Springer.

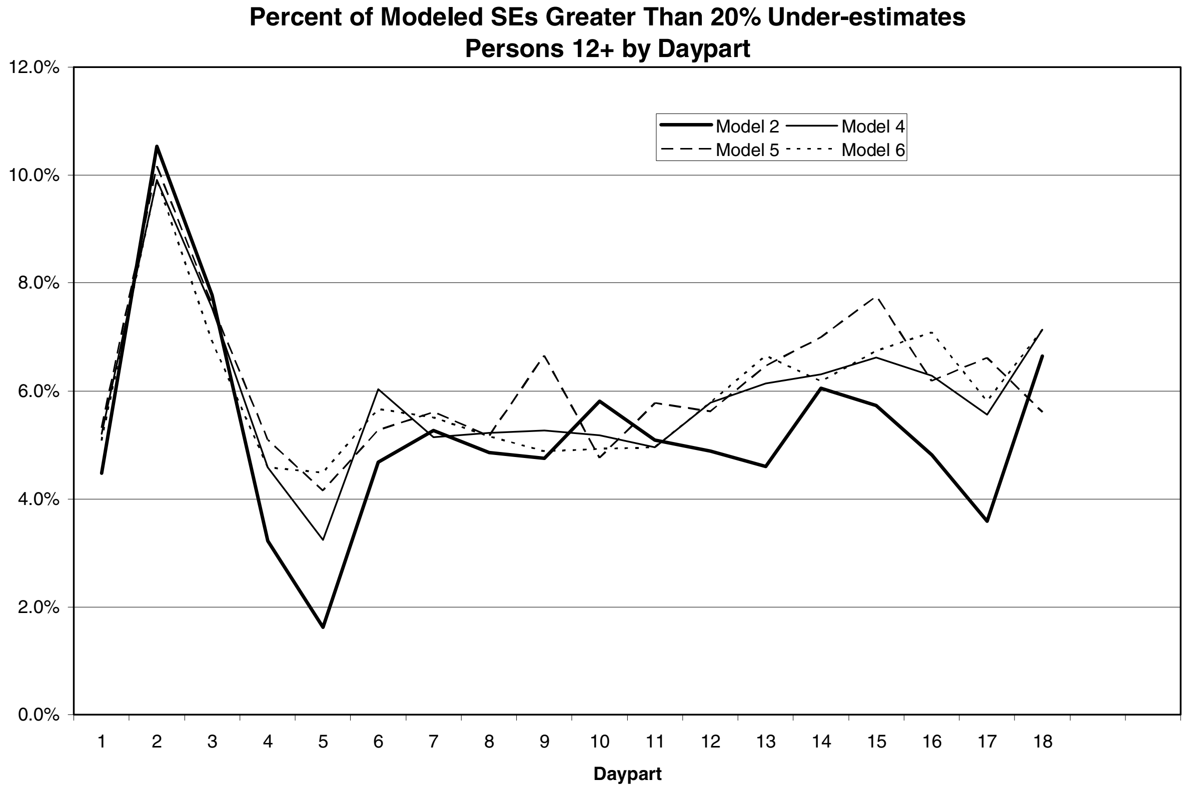
## Attachment A Box-and-Whisker Plots



ARBITRON

## Attachment B Median ARD and Percent Large Under-Estimates, Persons 12+, Models 2, 4, 5, and 6


Median ARD Persons 12+
By Daypart
Model 2
Model 4
Model 5
Model 6
Daypart


Percent of Modeled SEs Greater Than 20% Under-estimates
Persons 12+ by Daypart
Model 2
Model 4
Model 5
Model 6
Daypart