

# DESCRIPTIVE METHODS FOR EVALUATION OF STATE-BASED INTERVENTION PROGRAMS

WILLIAM W. DAVIS  
BARRY I. GRAUBARD  
ANNE M. HARTMAN  
*National Cancer Institute*

FRANCES A. STILLMAN  
*Johns Hopkins Bloomberg School of Public Health*

*In this article, the authors discuss program evaluation of intervention studies when the outcome of interest is collected routinely at equally spaced intervals of time. They illustrate concepts using data from the American Stop Smoking Intervention Study, where the outcome is state per capita tobacco consumption. States differ widely in mean tobacco consumption, and these differences should be accounted for in the analysis. A large difference in the variance of the intervention effect may be obtained depending on whether the variation in the between-state effects are considered. The confidence limits obtained by ignoring between-state effects are too optimistic in many cases.*

**Keywords:** *bootstrap; locally weighted regression; random effect; pooled cross-sectional time series*

Intervention programs and policies implemented at the state level have been used to promote behavior changes of the state populations in many areas, including tobacco consumption (Manley et al. 1997a), use of firearms (Ludwig and Cook 2000), and motor vehicle fatalities (Farmer, Retting, and Lund 1999). The evaluation of these programs/policies are important for determining the extent of their success and for obtaining information about areas of improvement that can be used to improve future programs. The gold standard for estimation of the intervention effect is based on group randomized trials (Murray 1998). However, it is difficult to randomize interventions

---

AUTHORS' NOTE: *We wish to thank Robert Cohen of the SAS Institute for assistance on the SAS version of the loess algorithm. Also, we would like to thank Robert Croyle, Kevin Dodd, Rocky Feuer, Mike Fay, and Bill Trochim for useful comments.*

EVALUATION REVIEW, Vol. 27 No. 5, October 2003 506-534

DOI: 10.1177/0193841X03254405

© 2003 Sage Publications

506

at the state level, and therefore, the evaluations of these interventions are observational. Two common approaches to evaluation are based on (a) a random sample of individuals from the states and (b) an aggregate outcome measure that summarizes the outcome for all individuals in a state (e.g., the total monthly sale of tobacco in a state)—without the measures on the individuals.

In this article, we focus on an aggregate outcome measure. The health index is assumed to be collected routinely at equally spaced intervals of time (monthly, quarterly, or yearly)—possibly from a national data system where the data can be obtained at the state level. For program evaluation, we wish to determine whether the intervention had the desired effect on the health index (i.e., moved it in the desired direction). However, even if the index moves in the desired direction, it is difficult to prove that the intervention caused it for designs of this type. For example, there could be other state or local programs operating during the intervention period with similar goals as the program under evaluation, and these other programs could cause part or all of the observed change. Also, migration across state boundaries could affect the results.

This article considers descriptive methods used for a preliminary evaluation of the state-based tobacco control program, American Stop Smoking Intervention Study (ASSIST). This preliminary study used tobacco sales data and compared the difference in per capita sales for states in the ASSIST program to states not in the program at time points during the intervention period. Although ASSIST is used to illustrate the statistical methods described in this article, these methods are general and can be applied to intervention studies that are implemented at levels other than the state (e.g., community or school). A key point of this article is that there is large variation between state mean levels on many health outcomes that may be due to state differences that are difficult to quantify. In program evaluation of state-based intervention studies the between-state variability must be accounted for properly in computing confidence limits or in hypothesis testing to make valid inferences. Including the between-state variability will increase variances of results, and, when ignored, confidence intervals will be too narrow and null hypotheses will be rejected too often.

ASSIST is a large tobacco control research initiative to develop, implement, and evaluate statewide tobacco control projects. The ASSIST goal was to demonstrate that a comprehensive and coordinated application of the best available tobacco control strategies would significantly reduce the prevalence and per capita tobacco consumption (Kessler et al. 1996; Manley et al. 1997a; Stillman et al. 1999). A total of 17 states were selected for the ASSIST project, which began in 1991 under the direction of both the National Cancer Institute (NCI) and the American Cancer Society (ACS). States were

awarded contracts primarily on how well their proposals demonstrated the capability to form effective coalitions and were funded for 5 years beginning in 1993. Thus, the ASSIST states may differ from comparison states with respect to their ability to construct comprehensive smoking control plans and possibly in other dimensions.

The ASSIST program did not occur in a vacuum. Many agencies other than NCI and ACS are trying to reduce tobacco consumption. There was diffusion of materials and ideas developed in the ASSIST states to comparison states. For example, before the ASSIST time period, a comparison state, California, initiated a large tobacco control program funded by its state tobacco excise tax, and this program might have affected the state's tobacco consumption.

Most large-scale national program evaluations have similar challenges. The ASSIST program evaluation approaches the problem by collecting data from varied sources and performing various analyses. Stillman et al. (1999) provides an overview of the ASSIST evaluation as well as a summary of data sources and model-based analyses that will be performed—including covariate adjustment. This article focuses on a single aspect of the ASSIST evaluation; whether the 17 ASSIST states achieve lower tobacco consumption rates than the other 34 states (including Washington, D.C.)—without covariate adjustment. The article demonstrates the general methodological problems that can arise with this type of evaluation approach (Manley et al. 1997b).

In the State Specific Data section, we show the per capita tobacco consumption for all states. Also, we show the mean consumption for the ASSIST and comparison states and discuss the impact of weighting of the state consumptions based on state population. In the Hypothesis and Test Statistics section, we test the hypothesis of no change in mean difference in tobacco consumption between ASSIST and comparison states in three ways. The first essentially repeats the analysis carried out by Manley et al. (1997b), including more recent data. This analysis smoothes the time series of the mean difference in consumption using locally weighted regression, or loess (Cleveland 1979). The second method eliminates the seasonality in the mean difference by aggregating the estimates to the yearly level. The third method uses the bootstrap (Efron 1979; Efron and Tibshirani 1993) to derive confidence intervals for the smoothed estimate of mean difference in consumption again using loess. The second two analyses incorporate between-state variability, and the resulting conclusions are quite different from the first. Finally, we use yearly data to test the hypothesis of no change in mean difference controlling for the baseline mean difference between the ASSIST and comparison states.

### STATE SPECIFIC DATA

The analysis is based on the state per capita consumption, which is derived as the ratio of tobacco consumption to the total state adult population, 18 years and older. The state consumptions were obtained from The Tobacco Institute, which reported tax payments from all packages of cigarettes removed from wholesale warehouses to retail outlets within each state on a monthly basis. The reporting unit is the number of cigarette packs on which taxes were paid in any given month. Consumption estimates from this source are gathered in a uniform manner across states and are the usual source for reporting national per capita consumption; however, they are subject to monthly and seasonal variations. For example, business related variance in tax data occurs because of increased inventory clearance in the final month of any quarter (especially December) and corresponding reduced clearance in the first month of the quarter. Following Manley et al. (1997b), we used a regression approach to estimate the state's monthly population based on census yearly population estimates.

For this article, we used monthly data from January 1983 through June 1998. Following Manley et al. (1997b) to reduce seasonal variation, we reduced the monthly data to six data points in each 12-month period, which are computed by averaging the monthly results for December (of the previous year) and January, February and March, April and May, June and July, August and September, and October and November. Figures 1a and 1b show the per capita tobacco consumption for the ASSIST and comparison states respectively, where the same scale is used on both figures and the period of ASSIST state funding is indicated. Although we cannot make conclusions about the overall effectiveness of the ASSIST program from these graphs, they are useful for descriptive purposes. These graphs show the following:

- The trend in per capita consumption has been down for almost all states during this period.
- There is considerable state-to-state variation in the level of the per capita consumptions with larger variation in the comparison states.
- Some seasonal periodicity remains in many of the state series.

### AVERAGE CONSUMPTION FOR ASSIST AND COMPARISON STATES

Figure 2 shows the average per-capita consumption for the 17 ASSIST and the 34 comparison states. In this figure, we have weighted all states equally in the calculation of the mean tobacco consumption; that is, for the ASSIST state mean consumption, we used

510

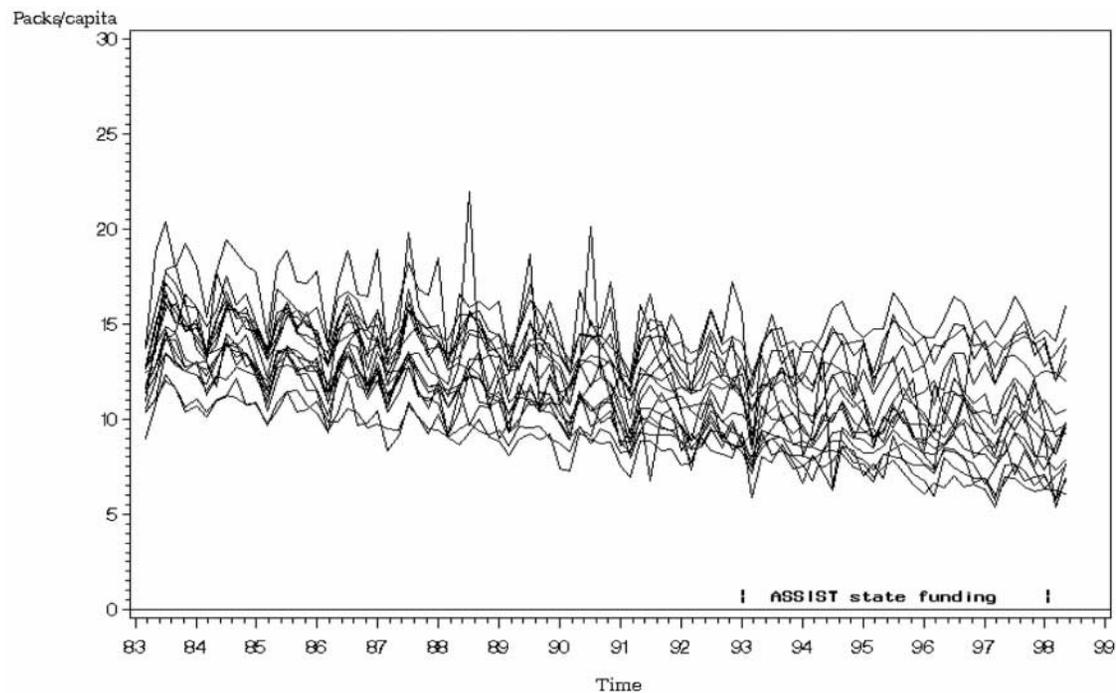
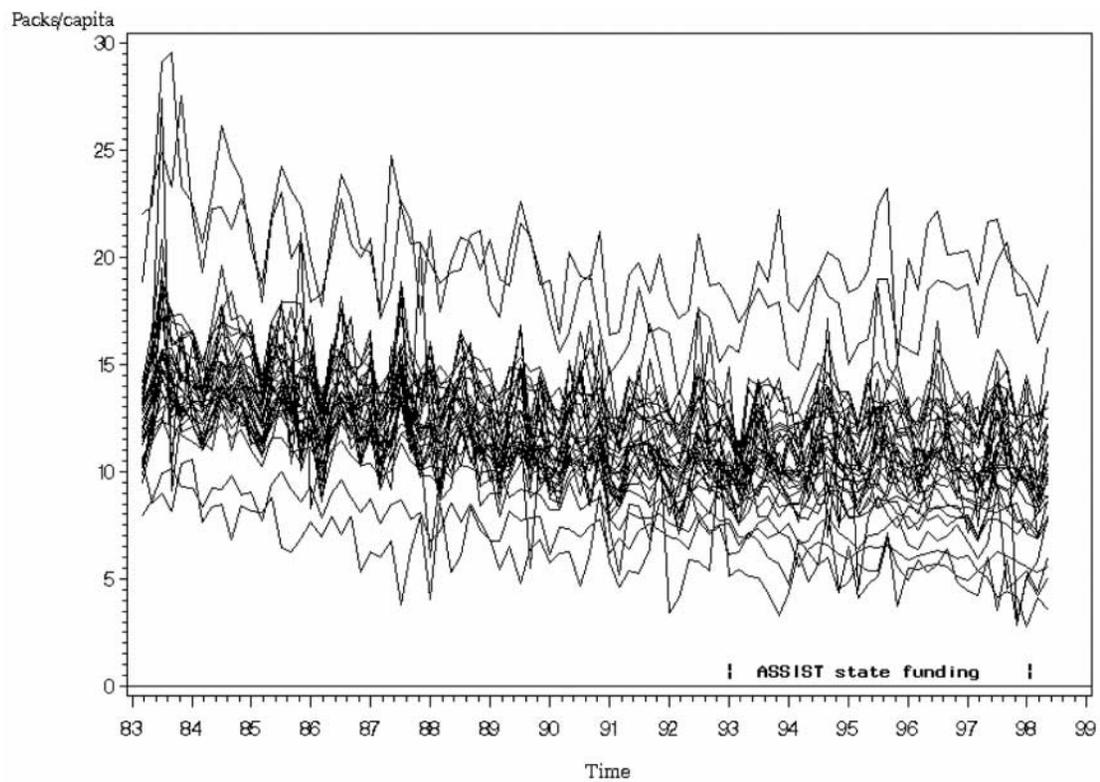


Figure 1a: ASSIST State per-Capita Consumption of Cigarette Packs per Month



111

Figure 1b: Comparison State per-Capita Consumption of Cigarette Packs per Month

$$A_t = 17^{-1} \sum_{s=1}^{17} R_{st}, \quad (1)$$

where  $R_{st}$  denotes the per capita tobacco consumption in state  $s$  and time  $t$  and the summation is over the 17 ASSIST states (with a similar definition for the mean consumption of the comparison states). We used this method because the selection and fund allocation were at the state level and the funding was generally independent of the state's population. Thus, we feel that an unweighted analysis is more appropriate than weighting by the state's population. The figure shows the decline in mean per capita tobacco consumption for both ASSIST and comparison states. Also, it appears that the comparison states have a slightly higher mean tobacco consumption by 1995.

If the ASSIST program works as planned, the difference in tobacco consumption between the ASSIST and comparison states should increase over time, beginning after the program inception in 1993. The difference in means can be used to determine whether this is the case. The difference,  $d_t$ , in tobacco consumption between ASSIST and the comparison states is defined by

$$d_t = C_t - A_t, \quad (2)$$

where  $A_t$  and  $C_t$  are the estimated mean tobacco per capita consumption at time  $t$  for ASSIST and comparison states respectively—as shown in Figure 2.

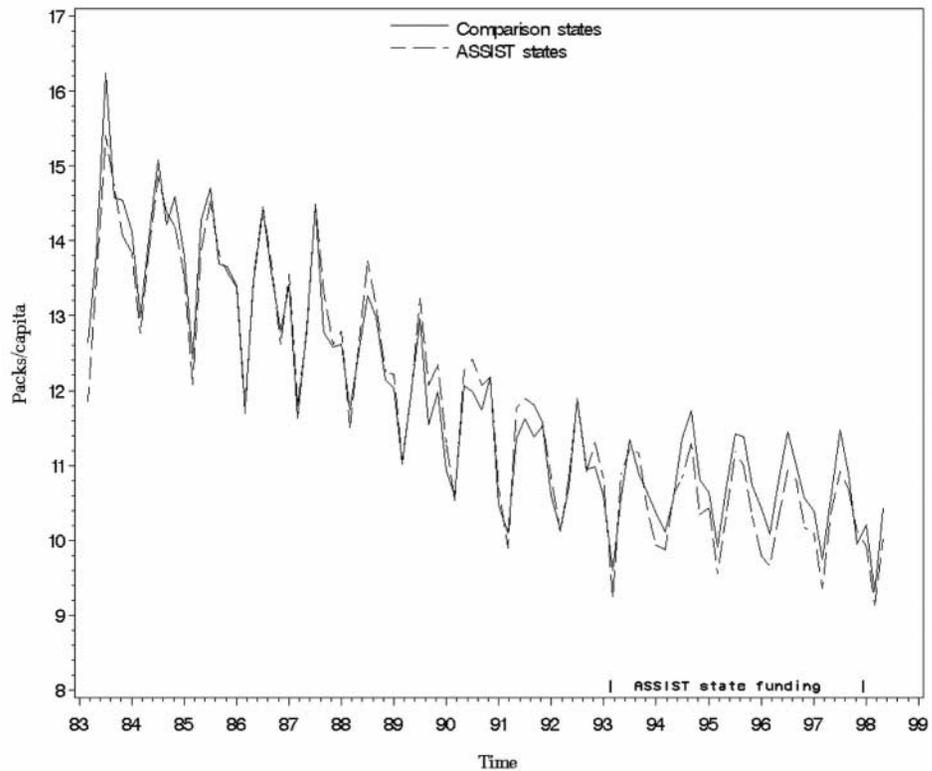
An alternative is to replace Equation (1) with a weighted mean, where the weight is proportional to the state population. For example, the weighted mean for the ASSIST states is with

$$A_t = \sum_{s=1}^{17} w_{st} R_{st}$$

with

$$w_{st} = P_{st} / \sum_{s=1}^{17} P_{st},$$

where  $P_{st}$  denotes the population of state  $s$  and time  $t$ . With a weighted mean, populous states have more impact on the per capita consumption estimate. As stated above, we feel that unweighted averages are more appropriate than weighted averages for the ASSIST evaluation. Figure 3 shows the impact that a single populous state can have if weights are used. This figure shows the



S13

Figure 2: Average per-Capita Consumption of Cigarette Packs per Month for ASSIST and Comparison States

514

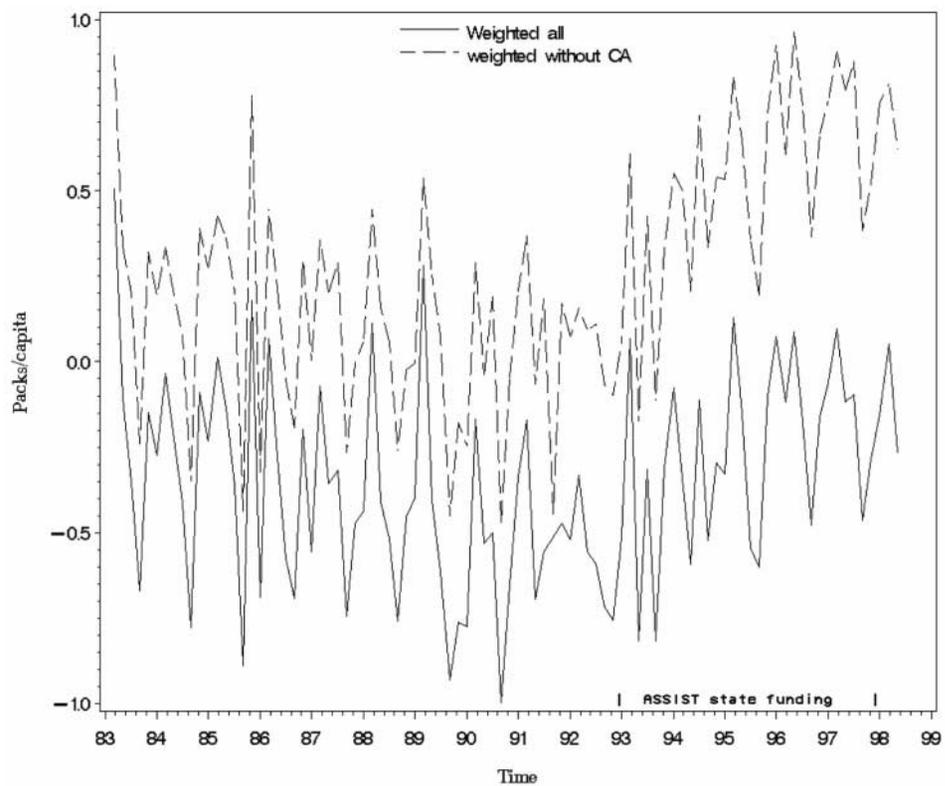


Figure 3: Difference of Weighted Estimates of per-Capita Consumption of Cigarette Packs per Month

weighted mean difference with all comparison states and with all comparison states—except California. The two curves in Figure 3 show similar behavior but differ by between 0.5 and 1.0 packs/capita—reflecting the importance of California using the weighted estimate. The difference is due to California's large population (hence large weight) and its low tobacco consumption. Because both curves in Figure 3 are very choppy; further smoothing is necessary to accurately assess the significance of the intervention program.

### HYPOTHESES AND TEST STATISTICS

If the ASSIST program works as planned, the difference in tobacco consumption between the ASSIST and comparison states should increase over time at least initially beginning just after the program inception (or baseline). Here, we will base hypothesis tests on the difference,  $d_t$ , defined in Equation (2). To state the hypothesis that we are testing explicitly, we define its expected value as

$$\theta_t = E(d_t). \quad (3)$$

If the ASSIST program is effective, the difference should increase after program inception, which we define as time  $T$ . Confidence limits can be obtained for and can be used to test the null hypothesis of no program effect by determining whether the confidence interval includes zero. If the program will be evaluated at a number of time points, it may be advisable to use a significance level lower than .05 to control the overall error rate. We phrase this in Table 1 as a test of whether the mean difference is significantly different from 0. We illustrate three different procedures for testing this hypothesis in Manley Analysis Using Additional Data, Analysis Based on State Yearly Tobacco Consumption, and An Alternative Look at Smoothing Using the Bootstrap sections. Also, we explain why these procedures yield different conclusions.

Although the alternative is phrased as two-sided in Table 1, the sign of the difference is important. Positive values of the difference indicate program effectiveness, whereas negative values indicate adverse program effects and may require further investigation.

Alternatively, the null hypothesis that the program is not effective can be phrased in terms of the difference in expected value  $\theta_t$  at time  $t$  from the baseline expected value  $\theta_0$ ; we introduce a parameter,  $\tau_t$ , to measure this difference of expected values

**TABLE 1: Statement of Null and Alternative Hypotheses**

	<i>Null Hypothesis</i>	<i>Alternative Hypothesis</i>
Mean difference	$\theta_t = 0, t \geq T$	$\theta_t \neq 0$ for some $t > T$
Mean difference corrected for baseline	$\tau_t = 0, t \geq$	$\tau_t \neq 0$ for some $t > T$

$$\tau_t = \theta_t - \theta_T = E(d_t) - E(d_T). \quad (4)$$

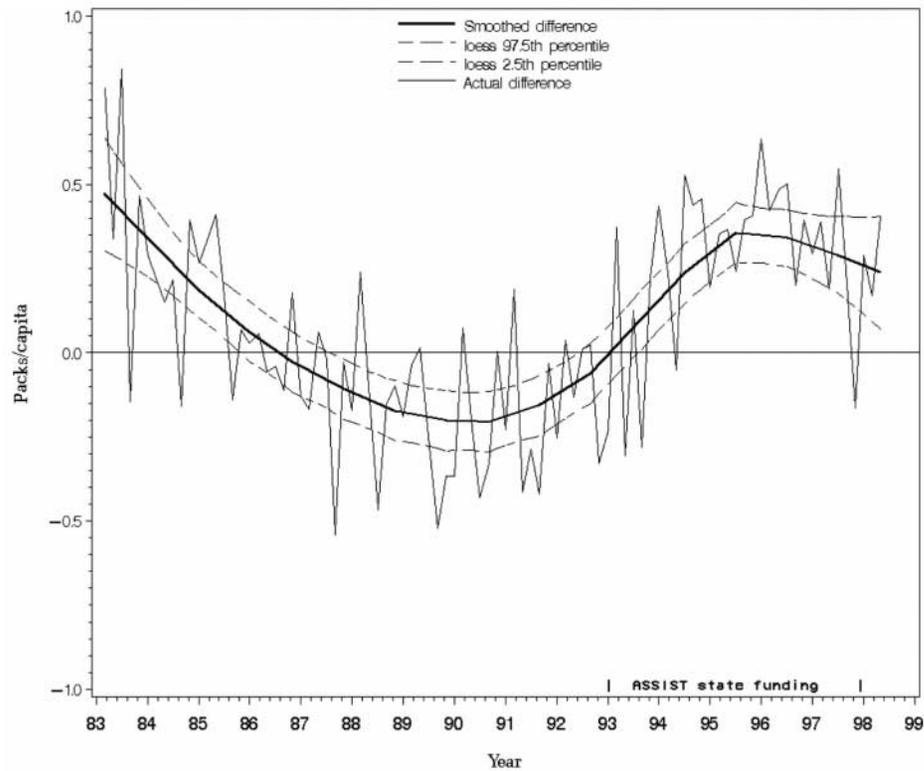
Equation (4) shows that  $\tau_t$  must be zero at baseline, whereas  $\theta_T$  may not be. Because  $\theta_T$  measures the mean expected difference in per capita tobacco consumption at baseline between comparison and ASSIST states, if the states were randomly assigned to the two groups, the mean difference should be approximately zero. If  $\theta_T = 0$ ,  $\theta_t$  and  $\tau_t$  coincide for  $t > T$ . However, if the states are not randomly assigned, it is likely that  $\theta_T \neq 0$  so that  $\theta_T$  and  $\tau_t$  do not coincide for  $t > T$ . The hypothesis stated for  $\tau_t$  in Table 1 corrects for baseline difference, whereas the hypothesis stated for  $\theta_t$  does not. If  $\theta_T \neq 0$ , the conclusions based on testing the two hypotheses of Table 1 could be different.

A problem in the use of Equation (4) is that it is sometimes difficult to determine the baseline time exactly because the impact of an intervention may not start immediately. In the Estimates of Yearly Difference From the Baseline section, we provide a test of the second hypothesis shown in Table 1; namely, that the difference of mean differences is zero.

**MANLEY ANALYSIS USING ADDITIONAL DATA**

Following Manley et al. (1997b), we use locally weighted regression, or loess, to further smooth the difference series,  $d_t$ , (Cleveland 1979; Cleveland and Devlin 1988). With loess, a local (in time) estimate is made by weighting the values close to the given time more heavily than those that are distant in time. SAS PROC LOESS (SAS Institute 1999) was used to carry out the computations; in Appendix A, we show the SAS code used in the analysis.

Figure 4 shows the difference series, the curve smoothed by loess and a 95% confidence interval for the mean. The ASSIST impact has increased from the program inception, and its effect is estimated as 0.4 packs per person per month by early 1996. The hypothesis of no mean difference in Table 1 can be tested by determining whether a 95% confidence interval for the mean includes 0 for each time after the intervention. Because the confidence intervals do not include 0, from 1994 through mid-1998 we conclude that the program was successful in reducing consumption in the ASSIST states.



517

Figure 4: Smoothed Difference of per Capita Consumption of Cigarette Packs per Month

However, the figure suggests a possible imbalance between the ASSIST and comparison states as the estimated mean difference is larger at the start of the time period 1983 than during the ASSIST period. The estimated difference between ASSIST and comparison states was also statistically significant from 0 during the mid-1980s—far before ASSIST program inception. This may indicate an initial difference between ASSIST and comparison states in tobacco consumption. However, at the beginning of the ASSIST intervention (early-1993) the difference was nearly 0.

The two-sided 95% confidence interval for the mean difference in Figure 4 is constructed from the SAS LOESS procedure and is similar to the one used by Manley et al. (1997b). As will be shown in the later sections, these confidence intervals are too narrow because they are based on standard errors that are too small. Heuristically, these confidence intervals are based on standard errors that treat the data as coming from a single time series rather than a sample of state-specific time series, where there is significant between-state variability among their time series as represented by differing state mean levels (see Figures 1a and 1b). Loess ignores this between-state variability in levels resulting in a small standard error. If the state mean levels are considered as random, the standard error of the mean difference at any time is increased considerably. For purposes of illustration, we have presented these inaccurate confidence intervals and used them as reference to test for the significance of an ASSIST effect. In the next two sections, we present what we feel are more accurate confidence intervals for the mean.

Although the analysis shown in Figure 4 is similar to that conducted by Manley et al. (1997b), in addition to using additional data, there were the following differences:

- The Manley et al. (1997b) analysis used SABL (Cleveland and Devlin 1982) to seasonally adjust the series before using loess to smooth it, whereas Figure 4 did not use SABL before smoothing with loess.
- The Manley analysis used a weighted mean difference as described below Equation (2), whereas Figure 4 is based on an unweighted analysis, as in Equation (1).
- The Manley analysis excluded California because it has a large impact on the weighted difference (e.g., Figure 3), whereas Figure 4 included all states.

Although these three differences are potentially important, they do not affect the result of the hypothesis test. In both Figure 4 and in the Manley et al. (1997b) analysis, the confidence intervals do not include 0 from 1994 through mid-1998, so both conclude that the ASSIST program was successful in reducing tobacco consumption.

Now, we sketch why we diverged slightly from Manley's analysis using the three modifications listed above. In the Average Consumption for ASSIST and Comparison States section, we described why an unweighted analysis is more appropriate. With a weighted analysis, California affects the conclusion greatly (e.g., Figure 3), although this is not the case with an unweighted analysis. Thus, we chose to include all states (i.e., not to eliminate California). Finally, the seasonal adjustment using SABL had a minor impact on the overall trend, and hence, on the conclusions, so we did not utilize it.

#### ANALYSIS BASED ON STATE YEARLY TOBACCO CONSUMPTION

In this section, we effectively eliminate the seasonality by considering the mean per capita yearly consumption for each state. Unlike the smoothing algorithms, the yearly estimate only uses (consumption and population) data from the specific year (actually, we used data from December of the previous year through November of the year in question). The variance of the difference of tobacco consumption,  $d_t$ , is estimated by assuming independence of the mean per capita yearly consumption estimates for the ASSIST and comparison states; thus,  $Var(d_t) = Var(C_t) + Var(A_t)$ . The sample variance was used to estimate both population variances [ $Var(C_t)$  and  $Var(A_t)$ ]. Then, an approximate 95% confidence interval for the yearly consumption difference is constructed from  $d_t \pm 1.96 * \sqrt{Var(d_t)}$ .

Figure 5 shows the yearly estimates and 95% confidence limits. The yearly estimates, upper limit and lower limits are plotted at the mid-year value; solid lines for the estimate and dashed lines for the confidence limits. Figure 5 is comparable to Figure 4. However, a major difference is that Figure 4 uses local regression, loess, to smooth the data (hence using data from outside the year in question), although Figure 5 uses only data for a single year. Although the vertical scales are different, the mean difference estimates are quite similar. For example, the smoothed mean in Figure 1 is 0.4 packs per month in 1983, gradually decreases to -0.2 by 1990, and then increases to 0.4 packs per month by 1996. Similarly, the yearly mean estimate in Figure 5 is approximately 0.4 packs per month in 1983, decreases to -0.2 by 1989, and then increases to 0.4 packs per month by 1995. In Figure 5, the yearly means change in a very smooth fashion—even without the smoothing algorithm used in Figure 4.

Although the mean difference estimates for Figures 4 and 5 are similar, the confidence limits are not. In fact, the confidence limits of Figure 5 are much wider than the confidence limits of Figure 4. Because the large state-to-state

520

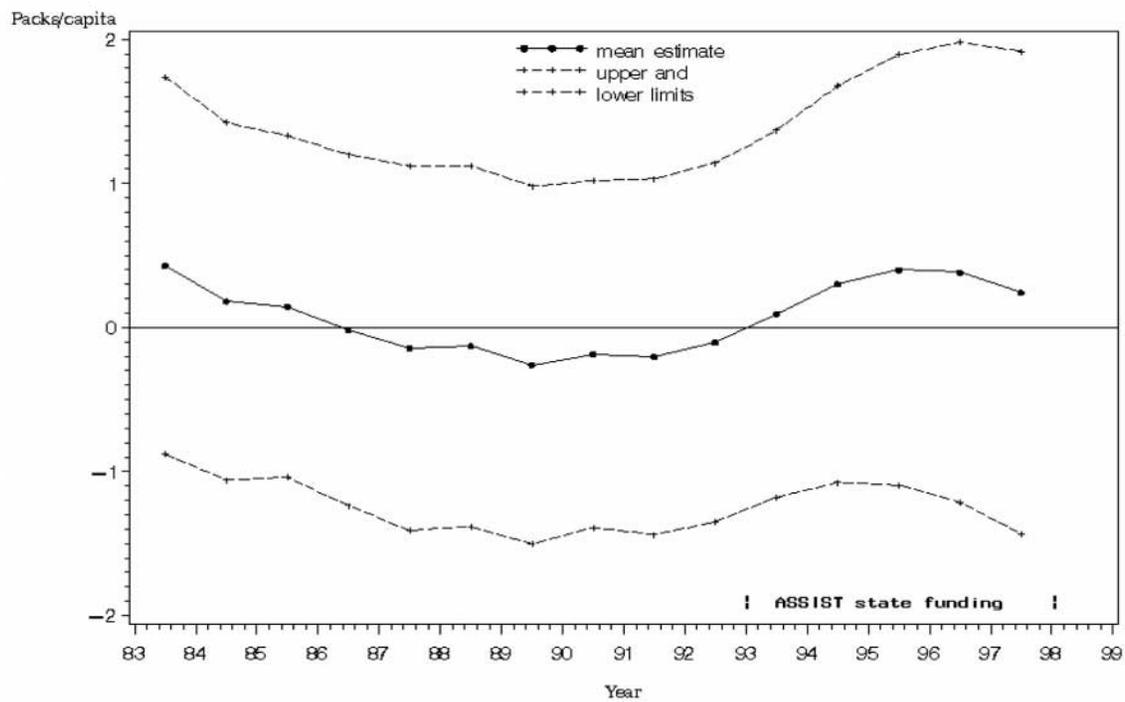


Figure 5: Estimate of Yearly per Capita Consumption Difference

variation is included in Figure 5, the confidence limits for the mean are much wider than those of Figure 4, which ignore this variation. In Appendix B, a mathematical derivation is given of the variance of the mean difference estimate when the ASSIST states are considered to be fixed and when they are considered to be a random selection.

All the 95% confidence limits in Figure 5 clearly overlap 0; thus, at significance level .05 we cannot reject the hypothesis that there is a statistically significant difference between ASSIST and comparison states. Thus the conclusions drawn from Figures 4 and 5 are quite different.

#### **AN ALTERNATIVE LOOK AT SMOOTHING USING THE BOOTSTRAP**

In this section, we use the bootstrap to obtain confidence intervals for the local regression (loess) for each time. Unlike the previous section, it is difficult to obtain a closed form expression for the estimated standard error of the (smoothed) mean difference. Thus, we used the bootstrap (Efron 1979) to obtain an approximate standard error of the estimate. The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates and can also be used to calculate robust confidence limits (Efron and Tibshirani 1993).

Here, we used the bootstrap to calculate the approximate 95% confidence limits for each time point. In carrying out the bootstrap, we used the state as the unit of analysis and used 1,000 replications of the bootstrap. For each replication, we created a sample of 17 ASSIST states and a sample of 34 comparison states using sampling with replacement. We carried out loess for each replication to estimate the mean difference for each time point from 1983 to mid-1998. Then, the mean of these 1,000 replications is used as the bootstrap estimate of the mean difference for each time. The monthly difference, the bootstrap mean estimate, and a bootstrap 95% confidence interval for the mean difference estimate is shown in Figure 6. The bootstrap confidence limits for the mean difference were obtained using two methods: percentiles and a normal approximation. Because these two confidence limits were similar, only the normal approximation results are shown.

The results of Figure 6 are quite similar to Figure 5. This may be somewhat surprising because the local regression (see Figure 6) uses data from outside the year in question, whereas the yearly estimates (see Figure 5) do not. One might think that the local regression estimate, which utilizes more data, would have a smaller standard error. Because Figures 5 and 6 are very similar, the conclusions are identical. All the 95% confidence limits in Figure 6

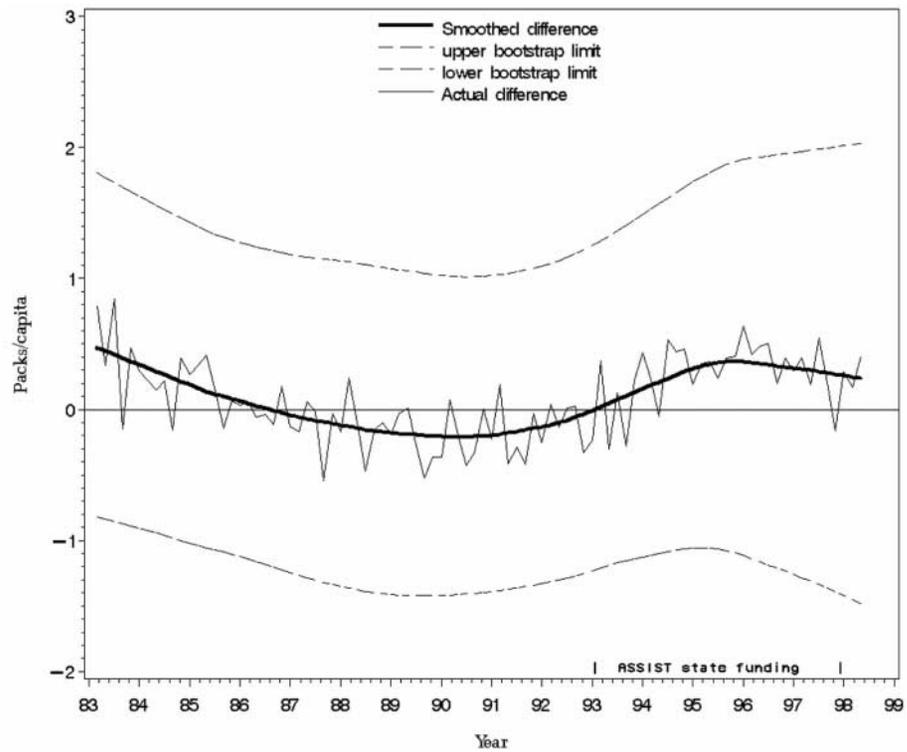


Figure 6: Estimate of Smoothed Difference of per Capita Consumption of Cigarette Packs per Month

overlap 0; thus, at significance level .05 we cannot reject the hypothesis that there is a statistically significant difference between ASSIST and comparison states. Again, this differs from the conclusion obtained from Figure 4.

A sensitivity analysis was carried out to study the robustness of the analysis and conclusions to the assumptions. Specifically, figures similar to 5 and 6 were created using weighted and unweighted analysis, including and excluding California, and using or not using SABL to seasonally adjust the ASSIST time series. Although the estimated ASSIST intervention effect varied depending on the particular analysis, the overall conclusion was always the same; namely, the ASSIST intervention effect was not statistically significant when the state variation is included in the analysis.

#### ESTIMATES OF YEARLY DIFFERENCE FROM THE BASELINE

In this section, we test the hypothesis defined in Table 1 for the parameter  $\tau_t$  (rather than  $\theta_t$ ); we base the hypothesis test for each year,  $t$ , on the statistic,  $s_t$ ,

$$s_t = d_t - d_T = (C_t - A_t) - (C_T - A_T), \quad (5)$$

where we use  $T = 1992$  as the ASSIST baseline year and we consider values  $t = 93, 94, \dots, 97$ . The bootstrap was used to estimate its standard error and to calculate confidence limits.

Again, the bootstrap was used with the state as the unit of analysis and used 1,000 replications of the bootstrap. For each replication, a sample of 17 ASSIST states and a sample of 34 comparison states were obtained using sampling with replacement. The bootstrap mean estimate for the statistic,  $s_t$ , and the bootstrap 95% confidence interval (using the normal approximation with the bootstrap standard error) for the statistic are shown in Figure 7.

Again, all the confidence intervals include 0 (though the one for 1994 barely does). Thus, the conclusion is that the ASSIST program has not decreased consumption in a statistically significant fashion. The confidence intervals are much narrower in Figure 7 than the comparable Figure 5. The reason is that controlling for the mean baseline difference in Equation (5) effectively removes much of the state-level mean differences, which has been shown to be a large source of the variation.

524

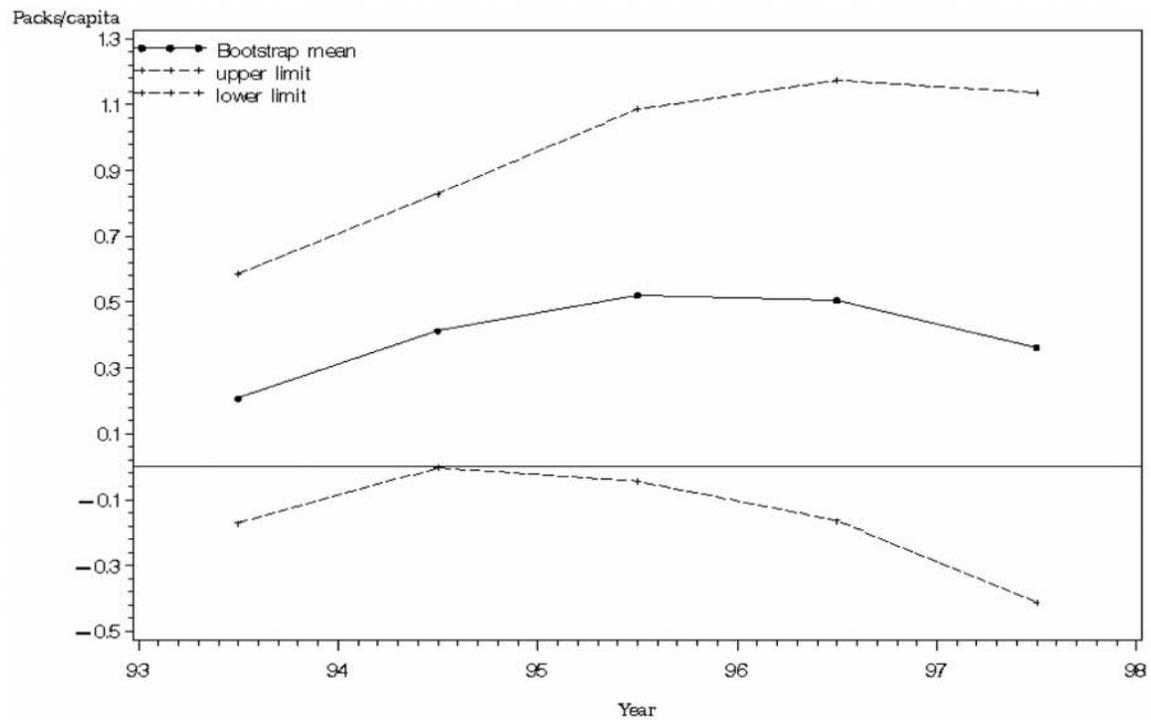


Figure 7: Yearly Mean Difference of per Capita Consumption Controlled for Baseline Difference

## DISCUSSION

This article discusses program evaluation of state-based intervention trials. The article focuses on the assessment of the effect of a state-based intervention on a state health index, which is collected routinely at equidistant times. Hypotheses are defined using two parameterizations of the mean difference, and statistical methods are defined to test each of the null hypotheses.

The concepts are illustrated through four analyses of the ASSIST per capita tobacco consumption rates using data from 1983 through the middle of 1998. The first three of these analyses determine confidence intervals for the mean difference, whereas the fourth analysis determines confidence intervals for the mean difference controlling for baseline differences. The confidence intervals are used to test the significance of the ASSIST intervention effect.

The first analysis uses locally weighted regression (loess) to smooth the monthly time-series and is similar to the analysis published by Manley et al. (1997b). There are large differences between the states in per capita tobacco consumption that are persistent over time. The usual loess standard error estimate for the local mean ignores this between-state variability in levels—resulting in a small standard error so that this analysis finds a significant ASSIST effect.

In contrast to the first analysis, the other three analyses include the state variability in levels. The second analysis bases estimates for the mean difference only on state tobacco consumption for the year in question. Although the mean difference estimates are similar to those of the first analysis, the confidence intervals for the mean difference are considerably wider due to the state variation. The third analysis uses the bootstrap to determine the confidence intervals for the mean difference in tobacco consumption as estimated using loess—including the state variability in levels. The confidence intervals for the mean difference are similar to the second analysis. The fourth analysis makes yearly mean difference estimates controlling for the baseline difference between the ASSIST and the comparison states. Controlling for the baseline difference effectively removes the state level and yields more narrow confidence intervals than the second and third analysis. However, the conclusion from all three analyses that include the state variation is that the ASSIST intervention effect is not statistically significant.

Although our first analysis differed from Manley's in data weighting, data exclusions, and treatment of seasonal effects, the differences do not affect the main conclusion about the ASSIST intervention effect. A large number of analyses, not included here, confirm that the impact of these three factors is small compared to the impact of including the state variation in the analysis.

As shown here, the decision whether to include the state variation does affect the main conclusion concerning the significance of the ASSIST intervention.

Another way to assess the ASSIST program impact is through the change in slope of the difference in per capita consumption immediately after the start of the ASSIST program. If the ASSIST program had the desired effect, we would expect the slope of the curve to increase some time after the inception of the ASSIST program; say, by the beginning of 1994 (Wun and Kessler 1996). Although we do not carry out a formal test for the change in slope, the increase in slope is not visible in the smoothed curve, and, in fact, the slope of the curve begins to decrease by 1996.

Data of the form considered here is known in the econometrics literature as pooled cross-sectional and time-series data (Dielman 1989). There are several model-based techniques that can be used to analyze data of this type when there is a state intervention. For example, we could have used Box-Jenkins time-series methods (i.e., Box et al. 1994) with an intervention term to model the change in the series after the baseline (i.e., Box and Tiao 1975; Harrop and Velicer 1985). NCI's sponsored ASSIST evaluation is using mixed effects time-series models adjusting for important state-level covariates using SAS Proc Mixed (SAS Institute 1999; Murray et al. 1998) to evaluate per capita tobacco consumption (Stillman et al. 1999, 2003). Also, the NCI's ASSIST evaluation will use data through 1999 rather than mid-1998 as we did.

This mixed effects modeling corresponds to the more traditional analysis of pooled cross-sectional time series (e.g., Dielman 1989). The advantage of the descriptive methods given in this article over the model-based methods are that the descriptive methods do not require extensive model assumptions that specify the distribution and structure of the errors and that specify the form of the mean relationship between time and the outcome. The model-based methods allow for modeling covariates and provides a framework to statistically test characteristics of the pattern of the outcomes over time (e.g., if the outcomes are following a nonlinear trend versus a linear one, thus, conveying possible power advantages if the modeling assumptions are valid).

There are numerous published studies where the results of this article are applicable. One example is when a national law is changed. If some of the states are not affected by the change, they can be used as a control group, whereas the affected states can be considered as the treatment group. An example of this is the effect of the Brady Bill in 1994 on homicide and suicides (Ludwig and Cook 2000), where the control group consisted of 18 states and the District of Columbia, which had equivalent legislation already in place. Another example is the impact of the change in the national speed

limit on the number of traffic fatalities (Farmer, Retting, and Lund 1999). The results also apply to the case of determining the effect of the change of law in a single state with the rest of the country; for example the impact of California's Proposition 99, which increased the tax on cigarettes by 25 cents per package and allocated 20% of the tax money for antitobacco educational campaign (e.g., Fichtenberg and Glantz 2000). In this case, the treatment group has a single member so a pooled variance estimated can be used.

In summary, we have presented methodology along with important issues for consideration when descriptively assessing state-based intervention studies from routinely collected data over the course of the intervention. An important issue to consider is the state-to-state variability of the intervention effect. We have shown that this variability can be large and needs to be incorporated into the standard errors of the estimated intervention effects, otherwise conclusions about the intervention can be incorrect. Because of the independence and organization of state governmental institutions, states will continue to be entities for intervention to change the U.S. population behavior. The methods described in this article will be applicable to future state interventions studies.

#### **APPENDIX A**

##### **DISCUSSION OF LOESS FOR PROGRAM EVALUATION: INCLUDING OPTION SPECIFICATION**

---

The loess procedure is useful for assessing the impact of a program, where there are multiple time points following the intervention. Loess is nonparametric in that it makes only local, not global, parametric assumptions about the regression surface. Weighted least squares is used to fit linear or quadratic functions of the predictors, usually time for program evaluation, at the center of neighborhoods. The fraction of the entire time series used to estimate the parameters of the local neighborhood controls the smoothness of the loess estimate. Data points in the local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood.

Loess was developed to deal with data sets with a large number of observations. The default loess option in SAS carries out the local fitting at only a sample of the points (time in our example) and interpolates to obtain the regression surface. The program evaluation data sets will often be sufficiently small so that it is easy to carry out the exact loess analysis with a local neighborhood at each time point. Also, loess is able to handle non-normal error distributions with outliers using iterative reweighting.

We used the following SAS code using PROC LOESS to carry out the analysis of Figure 4. The input data set stored in DATA=one is a collection of pairs (time, DIFF) where "time" is equally spaced and "DIFF" is the unweighted difference obtained

from Equation (2). The “MODEL DIFF=time” statement specifies a local linear trend in time. The “direct” option specifies that a local neighborhood is used for each data point (e.g., no interpolation), the smooth=0.33333333 specifies that each neighborhood contains one third of the number of points in the time series, the “iter=1” option specifies no iterative reweighting, the “clm” option specifies that a 95% confidence interval should be computed for the mean at each time, and “details” controls the output. Default options were used for other quantities.

```
PROC LOESS DATA=one;
MODEL DIFF=time/direct smooth=0.33333333 iter=1 clm
    details;
RUN;
```

We checked the calculations using the S-PLUS (1999) loess algorithm and obtained similar results when the options conformed to those specified in the SAS statement. Also, we performed sensitivity analysis by increasing the number of iterations; there was a small impact on the estimates and confidence intervals, especially near the beginning or end of the time series. However, the changes did not affect the conclusions.

## APPENDIX B FIXED AND RANDOM EFFECT MODELS FOR STATE PER CAPITA TOBACCO CONSUMPTION

In this appendix, we provide a model-based explanation of why the confidence limits in Figures 4 and 5 are so much wider than those in Figure 3. We sketch the results for the ASSIST states modeled with two regimes corresponding to the  $T_1$  observations before and the  $T_2$  after the ASSIST intervention (with  $T - T_1 + T_2$  observations for each state time series). In contrast to the body of this article, in this appendix, the intervention time is labeled as  $T_1$  and  $T$  is the length of the time series.

For the ASSIST states, we assume that the per capita tobacco consumption,  $R_{st}$ , for the state  $s$  at time  $t$  satisfies

$$R_{st} = \begin{cases} a_{1s} + b_{1s}t + \varepsilon_{st} & 1 \leq t \leq T_1 \\ a_{2s} + b_{2s}t + \varepsilon_{st} & T_1 + 1 \leq t \leq T_1 + T_2 \end{cases} \quad (\text{A-1})$$

This is the special case of the two-regime regression model originally considered by Quandt (e.g., Maddala 1977, chap. 17). Assuming the errors  $\varepsilon_{st}$  have mean 0, the mean consumption for regime  $j$  is  $a_{js} + b_{js}t$  for  $j = 1, 2$ . If the intervention has an impact for state  $s$ , the coefficients on the two regimes will be different. Because the errors may be correlated, we assume that  $\varepsilon_{st}$  follows the first order autoregressive model  $\varepsilon_{st} =$

$\rho \varepsilon_{s,t-1} + u_{st}$  where  $\{u_{st}\}$  are assumed to be independent errors each with mean zero and variance  $\sigma^2$  for each state  $s$  and all times  $t$ .

In the random case, we assume that the state intercepts satisfy the model

$$a_{js} = \bar{a}_j + v_{js} \quad j = 1, 2,$$

where  $\bar{a}_j$  are fixed and where  $\{v_{js}\}$  are independent mean zero random variables with variance  $\tau^2$  for  $j = 1, 2$  and all  $s$ . The random intercept version of (A-1) is given in (A-1r) where the random terms are separated from the fixed effects

$$R_{st} = \begin{cases} \bar{a}_1 + b_{1s}t + (v_{1s} + \varepsilon_{st}) & 1 \leq t \leq T_1 \\ \bar{a}_2 + b_{2s}t + (v_{2s} + \varepsilon_{st}) & T_1 + 1 \leq t \leq T_1 \leq T_2 \end{cases} \quad (\text{A-1r})$$

Both the fixed and random effect models can be written as linear regression models. However, due to the additional random component, the covariance matrix of the random effect model is more complicated. In Table A-1 for a single ASSIST state, we contrast the model and estimation results using (A-1) and (A-1r) and the following notation:  $R'_s = (R_{s1}, \dots, R_{st}, \dots, R_{sT})$ ,  $\varepsilon'_s = (\varepsilon_{s1}, \dots, \varepsilon_{st}, \dots, \varepsilon_{sT})$ ,  $v'_s = (v_{1s}, v_{2s})$ ,  $X = \text{diag}(X_1, X_2)$ , where  $X_j = (L_j \ t_j)$  with  $L_j$  a vector of  $T_j$  ones and  $t_j$  is a vector that represents the observation times,  $J = \text{diag}(L_1, L_2)$ , and  $P = (P_{ij})$  with  $P_{ij} = \rho^{|i-j|}$  is the autocorrelation matrix of an autoregressive process with parameter  $\rho$ , and  $\bar{T}_j$  is the average observation time for regime  $j$  (i.e.,  $\bar{T}_1 = 0.5(1 + T_1)$ ). In Table A-1, we consider the prediction and the prediction variance at time  $t$  using notation  $x'_t = (1, x_t)$ .

For simplicity, we assume in Table A-1 that the parameter vector  $(\tau^2, \sigma^2, \rho)$  is known; in practice, these parameters are unknown but can be estimated using standard software. The inverse of the matrix  $P$  is well known (e.g., Leamer 1978, chap. 8) whereas the inverse of the covariance matrix,  $\Omega$ , can be obtained using results on inverting patterned matrices (e.g., Leamer 1978, Appendix 1). Calculation of the variance matrix of the parameter estimates in closed form allows the variance of the mean estimate at time  $t$  to be calculated explicitly.

**TABLE A-1: Models, Estimates, and Variances in the Fixed-Intercept and Random Intercept Cases for ASSIST State  $s$**

	<i>Fixed Intercept Model</i>	<i>Random Intercept Model</i>
Statistical model	$R_s = X\beta_s + \varepsilon_s$	$R_s = X\phi_s = (J\mu_s + \varepsilon_s)$
Parameter definition	$\beta'_s = (\beta'_{1s}, \beta'_{2s})$ where $\beta'_{js} = (a_{js}, b_{js})$	$\phi'_s = (\phi'_{1s}, \phi'_{2s})$ where $\phi'_{js} = (\bar{a}_j, b_{js})$
Covariance matrix of error vector	$\sigma^2 P$	$\Omega = \tau^2 J J' + \sigma^2 P$
Parameter estimate	$\hat{\beta}_s = (X'P^{-1}X')^{-1}X'P^{-1}R_s$	$\hat{\phi}_s = (X\Omega^{-1}X')^{-1}X\Omega^{-1}R_s$
Variance matrix of parameter estimate	$Var(\hat{\beta}_s) = \sigma^2(X'P^{-1}X')^{-1}$	$Var(\hat{\phi}_s) = \sigma^2(X\Omega^{-1}X')^{-1}$
Mean estimate at time $t$	$x'_t \hat{\beta}_s$	$x'_t \hat{\phi}_s$
Variance of mean estimate at time $t$ in regime $j$	$Var(x'_t \hat{\beta}_s) = \frac{\sigma^2(1+\rho)}{T_j(1-\rho)} \left( 1 + \frac{12(t-\bar{T}_j)^2}{(T_j^2-1)} \right)$	$Var(x'_t \hat{\phi}_s) = \tau^2 + \frac{\sigma^2(1+\rho)}{T_j(1-\rho)} \left( 1 + \frac{12(t-\bar{T}_j)^2}{(T_j^2-1)} \right)$
Confidence interval for mean estimate at time $t$	$x'_t \hat{\beta}_s \pm 1.96 \sqrt{Var(x'_t \hat{\beta}_s)}$	$x'_t \hat{\phi}_s \pm 1.96 \sqrt{Var(x'_t \hat{\phi}_s)}$

Table A-1 shows the ratio of the variances for the random intercept model to the fixed intercept model for prediction of the mean at time  $t$  is given by

$$VIF = \frac{Var(x'_t \hat{\phi}_s)}{Var(x'_t \hat{\beta}_s)} = 1 + \frac{T_j \tau^2 (1-\rho)}{\sigma^2 (1+\rho) \left( 1 + \frac{12(t-\bar{T}_j)^2}{(T_j^2 - 1)} \right)}, \quad (A-2)$$

so that the ratio depends on the regime  $j$  through the length of the regime,  $T_j$ , and on the time difference from the mean regime time,  $t - \bar{T}_j$ . The ratio in (A-2) is called the variance inflation factor (VIF). Because the ratio of the length of the confidence intervals for mean estimate for the random intercept to the fixed intercept model is  $\sqrt{VIF}$ , the VIF is useful in explaining the large difference in lengths of the confidence intervals. Equation (A-2) shows that VIF is always at least 1, hence the term *inflation factor*. Thus, the variance (and confidence interval length) of the predicted values will always be at least as large using the random effect model as with the fixed effect model. The VIF can be much larger than 1—depending on the parameter values. If there is no autocorrelation (i.e.,  $\rho = 0$ ), for  $t = \bar{T}_j$  the VIF expression reduces to

$$VIF = 1 + T_j \frac{\tau^2}{\sigma^2}. \quad (A-3)$$

Equation (A-3) shows that the VIF will be very large when both the number of segment observation times,  $T_j$ , and the variance ratio  $\tau^2 / \sigma^2$  are large.

Although the above derivation is for a single ASSIST state, Table A-2 shows similar results for averages over the ASSIST states, where the mean consumption,  $A_t$ , at time  $t$  is given in (1). Because the state is the unit of analysis, the states are treated as independent so that the variance of the sum is the sum of the variances. Table A-2 shows the variances of the model-based estimate of the average consumption in the intervention states for the fixed- and random intercept models. Under the assumption that all ASSIST states have the same autocorrelation parameter, the VIF for the average is exactly the same as that given in (A-2).

A somewhat similar expression to Equation (A-3) is obtained for the VIF in cluster sampling from a population with intraclass correlation coefficient  $\psi$ . The VIF of the mean for a sample of size  $T$  is  $1 + (T - 1)\psi$  (e.g., Donner and Klar 2000, chap. 1). The VIF of a cluster sample can be appreciable (for large  $T$ )—even when  $\psi$  is a small positive number. Although Equation (A-3) is similar, the ratio  $\tau^2 / \sigma^2$  is not restricted to be less than or equal to 1 (like the correlation coefficient).

Using standard regression packages for the mixed-linear model (e.g., SAS PROC MIXED), we can obtain model-based estimates of the parameters of the models defined above. Because there is considerable state-to-state variance in mean level as demonstrated in Figures 1a and 1b, the ratio  $\tau^2 / \sigma^2$  is large; hence the VIF is very

large—resulting in increased length in the confidence interval for the mean difference using the random effect model.

Because the intervention does not affect the comparison states, we assume that the consumption model for comparison states is  $R_{st} = a_{1s} + b_{2s}t + \epsilon_{st}$  for all  $t$ . Because the analysis for the comparison states is a special case of the two-regime model, we do not include the analysis for the comparison states (see Table A-2).

**TABLE A-2: Estimates and Variances in the Fixed-Intercept and Random Intercept Cases Averaged Over All ASSIST States**

	<i>Fixed Intercept Model</i>	<i>Random Intercept Model</i>
Estimate of $E(A_j)$	$17^{-1} \sum_{s=1}^{17} X_t \hat{\beta}_s$	$17^{-1} \sum_{s=1}^{17} X_t \hat{\phi}_s$
Variance of estimate of $E(A_j)$ in regime $j$	$\frac{\sigma^2(1+\rho)}{17(1-\rho)T_j} \left( 1 + \frac{12(t-\bar{T}_j)^2}{(T_j^2-1)} \right)$	$17^{-1} \left( \tau^2 + \frac{\sigma^2(1+\rho)}{(1-\rho)T_j} \left( 1 + \frac{12(t-\bar{T}_j)^2}{(T_j^2-1)} \right) \right)$

## REFERENCES

- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*, 3d ed. Englewood Cliffs, NJ: Prentice Hall.
- , and G. C. Tiao. 1975. Interventional analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* 70:70-79.
- Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74:829-36.
- , and S. J. Devlin. 1982. Calendar effects in monthly time series: Modeling and adjustment. *Journal of the American Statistical Association* 77:520-28.
- , and S. J. Devlin. 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83:596-610.
- Dielman, T. E. 1989. *Pooled Cross-Sectional and Time Series Data Analysis*. New York: Marcel Dekker.
- Donner, A., and N. Klar. 2000. *Design and Analysis of Cluster Randomized Trials in Health Research*. London: Arnold.
- Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7:1-26.
- Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

- Farmer, C. M., R. A. Retting, and A. K. Lund. 1999. Changes in motor vehicle occupant fatalities after repeal of the national maximum speed limit. *Accident Analysis and Prevention* 31:537-43.
- Fichtenberg, C. M., and S. A. Glantz. 2000. Association of the California tobacco control program with declines in cigarette consumption and mortality from heart disease. *New England Journal of Medicine* 343 (24): 1772-77.
- Harrop, J. W., and W. F. Velicer. 1985. A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research* 20:27-44.
- Kessler, L. G., M. Carlyn, R. Windsor, and L. Biesiadecki. 1996. Evaluation of the American Stop Smoking Intervention Study. In *Health Survey Research Methods Conference Proceedings*, edited by R. Warnecke, 215-220. DHHS Publication No. 95-1013. Hyattsville, MD: National Center for Health Statistics.
- Leamer, E. E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley.
- Ludwig, J., and P. J. Cook. 2000. Homicide and suicide rates associated with implementation of the Brady Handgun Violence Prevention Act. *Journal of the American Medical Association* 284 (5): 585-91.
- Maddala, G. S. 1977. *Econometrics*. New York: McGraw-Hill.
- Manley, M. W., W. Lynn, R. P. Epps, D. Grande, T. Glynn, and D. Shopland. 1997a. The American Stop Smoking Intervention Study for cancer prevention: An overview. *Tobacco Control* 6 (supp 2): S5-S11.
- , J. P. Pierce, E. A. Gilpin, B. Rosbrook, C. Berry, and L. M. Wun. 1997b. Impact of the American Stop Smoking Intervention Study on cigarette consumption, *Tobacco Control* 6 (supp 2): S12-S16.
- Murray, D. M. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.
- , P. J. Hannan, R. D. Wolfinger, W. L. Baker, and J. H. Dwyer. 1998. Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine* 17:1581-1600.
- SAS Institute Inc. 1999. *SAS/STAT User's Guide, Version 8*. Cary, NC: SAS Institute Inc.
- S-PLUS 2000 User's Guide. 1999. Seattle: MathSoft Inc.
- Stillman, F. A., A. M. Hartman, B. I. Graubard, E. A. Gilpin, D. Chavis, J. Garcia, L. M. Wun, W. Lynn, and M. Manley. 1999. The American Stop Smoking Intervention Study: Conceptual framework and evaluation design, *Evaluation Review* 23 (3): 259-80.
- , A. M. Hartman, B. I. Graubard, E. A. Gilpin, D. M. Murray, and J. T. Gibson. 2003. The evaluation of the American Intervention Stop Smoking Study (ASSIST): A report of outcomes. 2003. Submitted for publication.
- Wun, L.-M., and L. Kessler. 1996. Statistical procedures and their associated power to assess the effectiveness of an Intervention for Smoking Cessation. Paper presented at the 1995 Joint Statistical Meetings, Orlando, FL.

*William W. Davis, Ph.D., is a mathematical statistician in the Division of Cancer Control and Population Sciences at the National Cancer Institute. His research interest is applied statistics, including time-series analysis and analysis of survey data.*

*Barry I. Graubard, Ph.D., is a senior investigator/mathematical statistician in the Division of Cancer Epidemiology and Genetics at the National Cancer Institute. His research has focused*

*on statistical methods for analyzing complex sample surveys used in disease etiology and public health.*

*Anne M. Hartman, M.S., is a biostatistician in the Division of Cancer Control and Population Sciences at the National Cancer Institute. Her research interests include cancer control monitoring, evaluation, and methods of tobacco control, including environmental tobacco smoke, and UV related exposures/skin cancer control protective behaviors; and methodology and evaluation of studies on diet patterns/carotenoids and cancer.*

*Frances A. Stillman, Ed.D., is an associate research professor in the Department of Epidemiology and co-director of the Institute for Global Tobacco Control at the Johns Hopkins Bloomberg School of Public Health. Her research interests include developing, implementing, and evaluating tobacco control programs. She was the director of the ASSIST evaluation.*

## **Request Permission or Order Reprints Instantly**

Interested in copying, sharing, or the repurposing of this article? U.S. copyright law, in most cases, directs you to first get permission from the article's rightsholder before using their content.

To lawfully obtain permission to reuse, or to order reprints of this article quickly and efficiently, click on the "Request Permission/ Order Reprints" link below and follow the instructions. For information on Fair Use limitations of U.S. copyright law, please visit [Stamford University Libraries](#), or for guidelines on Fair Use in the Classroom, please refer to [The Association of American Publishers' \(AAP\)](#).

All information and materials related to SAGE Publications are protected by the copyright laws of the United States and other countries. SAGE Publications and the SAGE logo are registered trademarks of SAGE Publications. Copyright © 2003, Sage Publications, all rights reserved. Mention of other publishers, titles or services may be registered trademarks of their respective companies. Please refer to our user help pages for more details: <http://www.sagepub.com/cc/faq/SageFAQ.htm>

**[Request Permissions / Order Reprints](#)**