Testing of Weight Smoothing Models in the Current Employment Statistics Survey with SAS and R

August 2023

Collin Witt

witt.collin@bls.gov

Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Washington, DC, 20212

Abstract

Sample units with extreme values can have undue influence on survey estimates. This is particularly the case when those sample units are associated with large design weights and the sample size is small. The extreme values with large design weights can disproportionately affect survey estimates and impact their stability. Using establishment survey data from the Current Employment and Statistics (CES), we explore methods for weight smoothing to reduce weight volatility and improve the stability of the survey estimates. This paper extends the previous work of Gershunskaya and Sverchkov (2014), in which they considered several models for weight smoothing, e.g., LOESS curves, penalized B-splines, and Bayesian models and compared weighted estimates from those methods to true values. We consider an additional set of methods to accomplish the same goals. These include using the CES Robust Estimator, mixed random effects, bagging, and high-performance split modeling. We compare weighted estimates from these methods to full administrative counts from the Quarterly Census of Employment and Wages (QCEW).

Key Words: Weight Smoothing, CES, QCEW, R, SAS, PROC IML

1. Introduction

Much of this paper builds on previous research by Gershunskaya and Sverchkov (2014) and Grieves and Gershunskaya (2017). In this paper we explore comparisons in LOESS, Robust Estimator, Spline, Mixed Random Effects, Bagging, and High-Performance models in SAS[1] with Bagging being performed in both SAS and R using PROC IML (see coding snippets).

2. The CES Survey

2.1 CES Overview

The Bureau of Labor Statistics (BLS) collects data each month on employment, hours, and earnings from a sample of nonfarm establishments through the Current Employment Statistics (CES) program. The CES survey includes about 122,000 businesses and government agencies, which cover approximately 666,000 individual worksites drawn from a sampling frame of Unemployment Insurance (UI) tax accounts covering roughly 11.0 million establishments.

[1] Documentation of individual SAS procedures used in this paper can be found in the SAS/STAT 15.2 User’s Guide (https://documentation.sas.com/doc/en/statug/15.2/titlepage.htm).

2.2 CES Frame and Sample Selection

The CES survey derives its frame from the Quarterly Census of Employment and Wages (QCEW) program. The QCEW is an administrative program that collects employment and wage information from all establishments covered under the unemployment insurance (UI) on a quarterly basis. From the derived frame, CES chooses a stratified simple random sample of UI accounts, that is, when a UI account is chosen all establishments under that UI account are included in the sample. Stratification is performed by state, industry supersector (a grouping of North American Industrial Classification System codes), and total employment size. The sampling rates for each stratum are determined through a method known as optimum allocation[2], which distributes a fixed number of sample units across a set of strata to minimize the overall variance, or sampling error, on the primary estimate of interest.

2.3 CES Estimator

The primary estimate of interest for the CES survey is the over the month change in employment, $\boldsymbol{R}_t$. The estimator used is defined as follows: $\widehat{\boldsymbol{R}}_t = \frac{\sum_{j \in S_t} w_j \, y_{j,t}}{\sum_{j \in S_t} w_j \, y_{j,t-1}}$, where j denotes the establishments, t is the current month, $y_{k,t}$ and $y_{k,t-1}$ denote the employment of sample units in the current and previous months, and S_t is the “matched sample” or the set of sample units reporting positive employment in the current and previous months. To produce monthly estimates of levels, we use the annual census value produced from the QCEW, Y_0, and apply the ratio with $\hat{Y}_{t=1} = Y_0 \, \hat{R}_{t=1}$ and subsequent months estimated as $\hat{Y}_t = \hat{Y}_{t-1}\hat{R}_t$. For more details see the BLS Handbook of Methods[3].

2.4 Motivation for CES Weight Adjustment

In CES, we essentially track changes in population employment every month. Gershunskaya and Sverchkov (2014) found that weights could be modeled as a function of the survey responses to produce a more efficient estimator. The difficulty lies in finding and estimating a suitable model for the relationship between weights and reported employment changes. Because we are using the ratio estimator and estimating the relative change, the link between the weights and residuals, $residual_{t,j} = y_{t,j} - R_t y_{t-1,j}$, is considered.

2.5 Weight Smoothing

Weight smoothing has been shown to be beneficial in probability surveys for lowering the variance of survey-weighted estimators by modeling the survey weights conditional on the variables of interest (Beaumont 2008). We construct this new set of "smoothed" weights with lower variance and more in alignment with the survey response to improve the accuracy of our estimates. We can think of these new “smoothed” weights as a function $f\ (*)$ of some response

[2] Current Employment Statistics – National: Design (https://www.bls.gov/opub/hom/ces/design.htm).

[3] BLS Handbook of Methods. Current State Employment Statistics – State and Metro Area. (https://www.bls.gov/opub/hom/sae/).

variable (in our case the residuals) plus some error term, $Weight_{smoothed} = f(residuals) + error.$

2.6 Robust CES Estimator

We currently use the Robust Estimator in the CES survey to minimize the MSE conditional on sample size, this is the baseline model we wish to improve upon. We adjust weights for a small number of influential reports, and the algorithm is designed to find a Winsorization cutoff point so to minimize the MSE of the resulting estimator (under certain mild to weak assumptions). Weights are then either censored or in more extreme cases, reports are removed from the ratio altogether. We refer to it as Robust Estimator in CES since the procedure is designed to reduce the effect caused by extreme weights and/or reported employment changes. This was a "gentle" attempt to change the initial weights, in which we only changed weights for the most significant reports. This approach yields an estimate with a lower MSE than the original weights method. Please see Gershunskaya and Huff (2004) for more information on the algorithm formulation.

3. Estimation Challenges

We believe there to be better model for weights, but it comes with constraints in our program,

1.) Timeliness - Processing times for weight smoothing need to be quick and expedient with analysts often having to process tens of thousands of records at a time, we simply do not have the human resources to sift through so much information.
2.) Automated Process - The process needs to be automated with little to no tweaking of criteria set in the modeling process.
3.) Computer Resources - While we could possibly parallelize the process for faster run times, we are limited by computer resources and therefore our modeling process must be carried out quickly. This really goes back ultimately to timeliness.
4.) Lack of Covariates - While the data size is large, there is a lack suitable of covariates that can be fit in the model.

4. Tested Models

Below is a summary of the models we used in our application of weight smoothing and a brief description.

4.1 LOESS

LOESS, or locally weighted regressions (Cleveland 1979), is a non-parametric model that creates regressions at each point using q nearest neighbors. The regressions are weighted as a function of the distances from that point to its q nearest neighbors. To fit the model, a few tuning parameters must be chosen. In general, a “smoothing” parameter $s \in (0,1]$ must be chosen, s is the percentage of data to be used in each local regression. As stated in a previous section, these models were first considered in the original CES work performed by Gershunskaya and Sverchkov (2014). The models were fit in SAS using their automatic parameter selection technique. We will consider some restrictions on the smoothing parameter. The SAS LOESS procedure gives the user the option to set an upper and lower bound on the potential smoothing parameter. SAS will perform its model selection based on the restricted domain of smoothing parameters, choosing the one that minimizes some criteria.

4.2 SAS Splines

A spline is a special function defined piecewise by polynomials which are suitable for fitting noisy data. The benefit of penalized B-splines being the automatic selection of the number of knots used. The TRANSREG (transformation regression) procedure fits linear models, optionally with smooth, spline, Box-Cox, and other nonlinear transformations of the variables. You can use PROC TRANSREG to fit a curve through a scatter plot or fit multiple curves, one for each level of a classification variable.

SAS Code Snippet

```
%let ublambda=500;

proc transreg data=aelnk;
by st subss;
model identity(weight) = pbspline(res/lambda=0.1 &ublambda. range);
output out= t  predicted;
id ui msa size wae0 wae1 sae0 sae1 cnt_ui;
where  cnt_ui     >20;
run;
```

4.3 SAS Mixed Random Effects

The GLIMMIX procedure fits statistical models to data with correlations or nonconstant variability and where the response is not necessarily normally distributed. These models are known as generalized linear mixed models (GLMM). GLMMs, like linear mixed models, assume normal (Gaussian) random effects.

SAS Code Snippet

```
proc glimmix data=aelnk ;
by st subss;
class msa;
model weight = msa res / ddfm = bw ; /* ddfm=satterth */
random intercept / subject = msa type=toep;
output out = t pred=pred;
run;
```

4.4 SAS/R Bagging

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach and was implemented using the ipred[4] R package within SAS IML.

[4] Documentation of the ipred R package can be found, https://cran.r-project.org/web/packages/ipred/ipred.pdf

SAS Code Snippet

mainbag.sas

```
proc iml;
%include "setoption.sas";
call ExportDataSetToR("aelnk2", "aelnk" );
%include "bag.sas";
call ImportDataSetFromR("work.t","results_comb" );
quit;
```

setoption.sas

```
submit / R;
options(stringsAsFactors = FALSE)
endsubmit;
```

bag.sas

```
submit / R;

library( ipred ) #library for bagged trees

bag <- function(x){

        x$msa2 <- as.factor(x$msa)   #convert MSA to factor for model

        #run model
          bagged_m1 <- bagging(
          formula = weight ~ res + msa2 ,
          data    = x,
          coob    = TRUE,
          control = (xval = 0)
          )

        #add predictions
          x$pred <- predict( bagged_m1, x)
          x$msa2 <- NULL
          return(x)
                    } #end bag function

#create factor to split data into st/subss combo
         aelnk$key <- as.factor( paste0(aelnk$st, aelnk$subss) )

# split the dataset into subsets and run our function bag() on each one of
# the subsets, returns a list with predictions adjoined to each subset
         results <- lapply( X= split( aelnk, aelnk$key) , FUN = bag )

# combine the list we received above into one large data.frame to return
# to SAS results_comb <- dplyr::bind_rows(results)

      results_comb <- do.call("rbind", results)

endsubmit;
```

4.5 High Performance Split

The HPSPLIT procedure is a high-performance procedure that builds tree-based statistical models for classification and regression. The procedure produces classification trees, which model a categorical response, and regression trees, which model a continuous response. Both types of trees are referred to as decision trees because the model is expressed as a series of if-then statements.

SAS Code Snippet

```
proc hpsplit data=aelnk seed=31415 ;
class msa ;
id st msa subss weight size wae0 wae1 sae0 sae1 res cnt_ui;
model weight = res msa ;
output out = t (rename=(P_weight=pred));
run;
```

5. Results

5.1 Evaluation Criteria

The evaluation is based on comparing the different models weight predictions to the true[5] employment levels that are available from the QCEW. An absolute difference was used as comparison between the models and the QCEW, $Abs_{dif} = \frac{|\theta - QCEW|}{QCEW}$, where θ is the employment estimate from the model of interest.

5.2 Model Fit Comparison

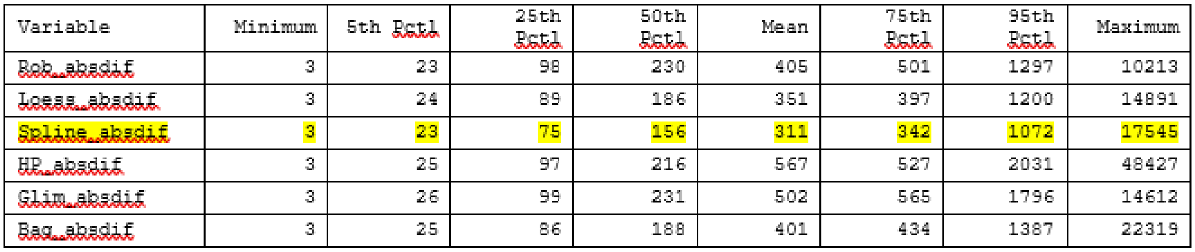
The benchmark 2019, 2018, and 2017 tables below are presented at the most basic estimation cell level i.e., State/MSA/Industry. Notice that the spline absolute difference has the smallest interval. The smaller the interval the better the overall fit to the QCEW. In addition, over the month relative differences were calculated (tables not listed in paper) for the same benchmark years, and the conclusion was the same. We can reasonably conclude that the spline model, at least for these benchmark years, performed the best overall. It's worth noting that there is a slightly better than a 20% reduction in the mean absolute error of the spline model from the baseline robust model.

Benchmark 2019

Variable	Minimum	5th Pctl	25th Pctl	50th Pctl	Mean	75th Pctl	95th Pctl	Maximum
Rob_absdif	4	32	149	340	691	738	2327	37640
Loess_absdif	4	34	131	280	624	628	2208	38843
Spline_absdif	4	31	117	251	605	588	2162	36332
HP_absdif	4	34	141	308	768	702	2761	48902
Glim_absdif	4	33	151	350	768	845	2762	30798
Bag_absdif	4	32	133	303	713	710	2482	50226

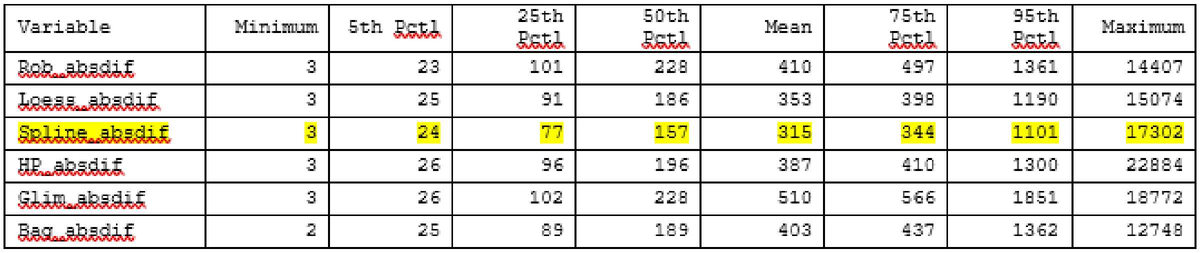
[5] The QCEW may contain non-sampling error but serves as a benchmark or Gold Standard value.

Benchmark 2018



Variable	Minimum	5th Pctl	25th Pctl	50th Pctl	Mean	75th Pctl	95th Pctl	Maximum
Rob_absdif	3	23	98	230	405	501	1297	10213
Loess_absdif	3	24	89	186	351	397	1200	14891
Spline_absdif	3	23	75	156	311	342	1072	17545
HP_absdif	3	25	97	216	567	527	2031	48427
Glim_absdif	3	26	99	231	502	565	1796	14612
Bag_absdif	3	25	86	188	401	434	1387	22319

Benchmark 2017



Variable	Minimum	5th Pctl	25th Pctl	50th Pctl	Mean	75th Pctl	95th Pctl	Maximum
Rob_absdif	3	23	101	228	410	497	1361	14407
Loess_absdif	3	25	91	186	353	398	1190	15074
Spline_absdif	3	24	77	157	315	344	1101	17302
HP_absdif	3	26	96	196	387	410	1300	22884
Glim_absdif	3	26	102	228	510	566	1851	18772
Bag_absdif	2	25	89	189	403	437	1362	12748

6. Current Progress

Currently research is being conducted for specific domains where the seasonality greatly differs from the statewide supersector where weight models are fit. In addition, the spline model is being worked into development and being tested across different States, MSAs, and Industries.

7. References

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. Biometrika, 95, 3, pp. 539–553 (https://www.bls.gov/osmr/research-papers/2014/pdf/st140140.pdf).

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74, 829–836. (https://www.jstor.org/stable/2286407).

Gershunskaya, J. and Grieves, C. (Nov. 2017). Testing Models for Weight Smoothing in the Current Employment Statistics Survey. (https://www.bls.gov/osmr/research-papers/2017/pdf/st170170.pdf).

Gershunskaya, J. and Huff, L. (2004). Outlier Detection and the Treatment in Current Employment Statistics Survey. (https://www.bls.gov/osmr/research-papers/2004/pdf/st040180.pdf).

Gershunskaya, J. and Sverchkov, M. (Oct. 2014). On Weight Smoothing in the Current Employment Statistics Suvery. (https://www.bls.gov/osmr/research-papers/2014/pdf/st140140.pdf).