

THE OUTLIER DETECTION AND MODELING METHOD PROPOSED FOR THE GST REPLACEMENT PROJECT OF THE MONTHLY WHOLESALE AND RETAIL TRADE SURVEY

Robert Philips¹

ABSTRACT

Statistics Canada undertakes the Monthly Wholesale and Retail Trade Survey (MWRTS) which produces estimates based on monthly data collected for sales and inventories at various provincial and industry levels. These estimates are a substantial portion of the estimates for the Gross Domestic Product (GDP) and the sales trend is an important economic indicator. In response to budgetary concerns and the need to reduce respondent burden coupled with the availability of good quality data obtained from Canada Revenue Agency's Goods and Services Tax (GST) Program, Statistics Canada has embarked on a process whereby the use of this GST data would be integrated into sub-annual business surveys. This paper will outline the outlier detection and modeling method proposed in relating the sales of business enterprises to GST data. The method is derived from a Bayesian hierarchical model for outliers in the univariate linear model. Results based on MWRTS data will be presented.

KEY WORDS: Bayesian hierarchical model, Outlier.

RÉSUMÉ

Statistique Canada effectue le sondage mensuel sur les ventes aux détails et en gros (EMCGD) qui donne des estimations basées sur des données mensuelles sur les ventes et les inventaires collectées dans différentes provinces et différents niveaux d'industries. Ces estimations constituent une portion substantielle de l'estimation du produit intérieur brute (PIB) et la tendance des ventes est un indicateur économique important. En raison de contraintes budgétaires, pour réduire le fardeau des répondants, et suite à la disponibilité de données fiables obtenues de l'agence de Revenu Canada et du programme de la Taxe sur les produits et les services (TPS), Statistique Canada a entamé un projet dans lequel l'utilisation des données sur la TPS seraient intégrées à un sondage infra-annuel auprès des entreprises. Cette présentation met en évidence la détection des valeurs aberrantes et la méthode de modélisation proposée pour relier les ventes des entreprises aux données sur la TPS. La méthode est basée sur un modèle bayésien hiérarchique pour des valeurs aberrantes dans un modèle linéaire simple. Les résultats basés sur les données de l'EMCGD seront présentés.

MOTS CLÉS: Modèle bayésien hiérarchique; valeur aberrante.

1. INTRODUCTION

1.1 Description of the Problem

The Monthly Wholesale and Retail Trade Survey (MWRTS) is a mission critical monthly survey conducted by Statistics Canada whereby sales and inventory data are collected from a sample of units that are classified into either the wholesale or the retail industry. Estimates are then calculated to represent the wholesale and retail populations for detailed industrial and geographical domains. These two industries constitute approximately 12% of the Gross Domestic Product (GDP). The MWRTS uses the same stratified simple random sample month after month except for the monthly addition of a sample of birth units. The Goods and Services Tax (GST) for its part is a Canadian tax on final consumption. Businesses must remit monthly if their annual sales are over \$6 million, quarterly if their annual sales are between \$500,000 and \$6 million, and annually if their annual sales are between \$30,000 and \$500,000, to the Canada Revenue Agency (CRA). Remittances at the BN (Business Number) level are obtained by CRA who then, provides this data to Statistics Canada. These remittances can represent monthly, quarterly or annual taxable periods and they may contain errors as well as partially or

¹ Robert Philips (Robert.Philips@statcan.ca), 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

totally missing information. Statistics Canada is thus required to carry out additional processing such as editing and imputation on this GST data.

Currently, this monthly GST data is only available at the establishment level for what are called simple units (units where a one-to-one link can be established with a GST Business Number) but the simple units that are sampled by MWRTS account for approximately 55% and 76% of the total live enterprises and represent 24% and 44% of the estimated sales in the Wholesale and Retail trades respectively.

The GST replacement project for the MWRTS has an objective of obtaining a 50% reduction in the number of simple units that are sampled as this will decrease the respondent burden and contribute to the overall lowering of survey costs. This replacement is subject to the further requirement that the overall data quality must continue to be maintained. The simple units are essentially randomly partitioned into two groups of units, henceforth called S1 and S2. The S1 units will continue to be surveyed while the S2 units will not. It is a fact that there is a very strong correlation between reported sales and GST data ($\rho = 0.92$ for Wholesale, $\rho = 0.96$ for Retail), after the removal of outliers and influential observations regardless of the vintage of the GST data. At the beginning of every month (m), the GST data available in time for the implementation in MWRTS is of different vintages i.e. m-1, m-2 and so on. Due to the timing of the processing period required for each month in the MWRTS, month m GST is not available. Thus, GST m-1 was chosen due to its correlation and proximity to the reference month. The GST replacement project will build models relating the reported sales and GST for the units in S1. The models and parameter estimates differ depending on the frequency of the GST remittance, industrial trade group and geographical location or some combination thereof. The parameter estimates obtained will then be used to predict the sales of the corresponding units in S2 based on their GST data.

The initial correlation and modelling studies showed the need for a robust outlier detection and parameter estimation routine that was data driven. In order to accomplish this task, a robust Bayesian model that accounted and compensated for the possible presence of outliers or influential observations was decided upon. The Bayesian approach to the general linear model in the infinite population setting is well documented (see Guttman et al (1993) for a brief summary, Freeman (1980), Petit and Smith (1985) or Philips (2001)). This approach is an extension of a work in progress to the setting of finite populations and non constant variance which give rise to the usual ratio model used in survey sampling, under the super population model assumption.

2.1 The Model

For notational convenience let y_i and x_i denote the weighted sales and the weighted GST remitted value from the previous month of the i^{th} unit in S1 respectively. In any modeling group (of size n), the analyst's view of the ideal situation is that all of the n observations are generated as intended, i.e. $\mathbf{y} = (y_1, \dots, y_n)'$ are generated according to the standard linear statistical model

$$\mathbf{y} \sim N(\beta \mathbf{x}, \sigma^2 V) \quad (2.1.1)$$

Where \mathbf{x} is the n x 1 design matrix and matrix $V = \text{diag}(x_1, \dots, x_n)$ is of size n x n both of whom are assumed known. β is the unknown regression parameter and σ^2 is the random error variance under the basic ratio model. Unfortunately, this model seldom reflects the actual reality, so that modifications to this model become necessary. Let $(i) = i_1, \dots, i_k$ denote one of the $\binom{n}{k}$ subsets drawn from the integers 1 through n and for each possible subset define the corresponding model $M_{(i)}$ such that for m=1 to k

$$y_{i_m} \sim N(\beta x_{i_m} + \delta_{i_m}, \sigma^2 x_{i_m}) \quad (2.2.2)$$

Specifically, these k observations have their own intercept and represent the possible outlying observations given that there are k outliers in the data set. The remaining n-k observations will follow the standard regression model given in

(2.1.1). We now develop the necessary prior distributions which will reflect our initial beliefs about the unknown parameters k, β, δ and σ^2 .

The list of parameters to completely describe our model can be divided into two separate groups. The first group consists of all the nuisance parameters, the $n \times 1$ vector of shifts δ and the corresponding hyper parameters $\Delta_1, \dots, \Delta_n$ which will be needed to specify the prior for the former. The second group is made up of the parameters of interest, which are: k , the number of outlying observations, β and σ^2 . This natural division into these groups suggests the following factorization for the joint prior of the parameters,

$$p(\delta, \Delta, \beta, \sigma^2, k) = p(\delta | \Delta, \beta, \sigma^2) p(\Delta) p(k) p(\beta, \sigma^2) \quad (2.1.3)$$

With no prior knowledge about the presence or absence of spuriously generated observations, it is logical to expect that the shift parameters are independently distributed random variables with a symmetric distribution about zero. The latter observation follows since each shift in theory could be either positive or negative in direction. A priori the actual magnitude of each possible shift is also unknown so that any magnitude is therefore possible. Based on these observations we then choose the simplest distribution encapsulating our initial state of ignorance for our conditional prior of the shift parameters, which is the uniform. Hence we will assume that for $i=1$ to n

$$\delta_i | \Delta_i, \beta, \sigma^2 \sim U\left(-\frac{\Delta_i \sigma \sqrt{x_i}}{2}, \frac{\Delta_i \sigma \sqrt{x_i}}{2}\right) \quad (2.1.4)$$

If we want the data to "speak for itself", then a priori we should model the problem as if the data set does not contain any spurious observations. This last fact implies that before the data has been collected we should require that $E(k) < 0.5$. Also, the presence of spurious observations can be equated to the occurrence of rare events. These last two observations combined together suggest the form of our prior for k , be a truncated Poisson, with parameter 0.5. Since Poisson probabilities become very small quickly, we assume for convenience that the prior is well approximated by the usual Poisson with parameter 0.5. The prior for the remaining parameters β and σ^2 will be the usual non-informative ones with their presumed independence. In the next section we derive all the relevant posterior distributions and estimators of parameters and discuss how to choose the most probable outlying observations.

2.2 Theoretical Results

Combining the likelihood with the priors yields that the joint posterior $p(\delta, \Delta, \beta, \sigma^2, k | data)$ is proportional to

$$\frac{\prod_{i=1}^n p(\Delta_i)}{2^k k! \binom{n}{k} \prod_{i=1}^n x_i \sigma^{n+1}} \sum_{i_1, \dots, i_k} \exp \left\{ -\frac{(\mathbf{y}_{(i)} - \mathbf{x}_{(i)} \beta)' \mathbf{V}_{(i)}^{-1} (\mathbf{y}_{(i)} - \mathbf{x}_{(i)} \beta)}{2\sigma^2} - \sum_{m=1}^k \frac{(y_{i_m} - x_{i_m} \beta - \delta_{i_m})^2}{2x_{i_m} \sigma^2} \right\} \quad (2.2.1)$$

In (2.2.1), any matrix or vector with the subscript (i) means that the rows corresponding to the indices in (i) are deleted, also the summation is a k -fold sum over all possible subsets (i) . Now, performing the appropriate integrations gives the following key results

$$p_k^* = p(k | data) \propto \frac{(\sqrt{\frac{\pi}{2}} E(\Delta^{-1}))^k}{k! \binom{n}{k}} \sum_{i_1, \dots, i_k} \frac{1}{\sqrt{SSE_{(i)}^{n-1} \sum_{m \notin (i)} x_m}} \quad (2.2.2)$$

where $SSE_{(i)} = \sum_{t \notin (i)} \frac{(y_t - \hat{\beta}_{(i)} x_t)^2}{x_t}$ represents the usual residual sum of squares. It can also be shown after using some asymptotic reasoning that a good approximation is

$$p_k^* = C \frac{(0.1)^k}{k! \binom{n}{k}} Q_k \text{ where } Q_k = \sum_{i_1, \dots, i_k} \frac{1}{\sqrt{(1 - \sum_{m \in (i)} h_{mm})}} \times \left(\frac{SSE_{(i)}}{SSE_{(i)}} \right)^{\frac{(n-1)}{2}} \text{ and } h_{mm} = \frac{x_m}{\sum_{j=1}^n x_j} . \quad (2.2.3)$$

In addition, the conditional probabilities $p(i)$, identifying which subset (i) is most likely the outliers, given that there are exactly k possible outliers is easily obtainable and in the special case when $k = 1$ is given by

$$p(i) = \frac{1}{Q_k \sqrt{(1 - h_{ii})}} \left(1 - \frac{\hat{\varepsilon}_i^2}{(n-1)(1 - h_{ii})\hat{\sigma}^2} \right)^{\frac{-(n-1)}{2}} . \quad (2.2.4)$$

The parameter estimates follow immediately, in particular the posterior estimate for β and is given by

$$\beta^* = \sum_{k=0}^{k^*} p_k^* \left(\sum_{(i) \in S_k^n} p(i) \hat{\beta}_{(i)} \right) \text{ where } \hat{\beta}_{(i)} = \frac{\sum_{t \notin (i)} y_t}{\sum_{t \notin (i)} x_t} . \quad (2.2.5)$$

If n gets relatively large, $\binom{n}{k}$ becomes prohibitive to evaluate and an approximation to the above procedure was required.

Instead of trying to obtain the posterior for $k = 0, 1$, etc. we do the following. On each iteration only two possibilities are considered either $k = 0$ or $k = 1$ and we proceed as shown below.

1st iteration :

if $p_0^{*(1)} < 0.67$ or $p(i_1) = \max\{p(1), \dots, p(n)\} \geq 0.25 \Rightarrow \text{outlier} = y_{i_1}$

2nd iteration :

if $p_0^{*(2)} < 0.67$ or $p(i_2 | i_1) = \max\{p(1 | i_1), \dots, p(n | i_1)\} \geq 0.25 \Rightarrow \text{outlier} = y_{i_2}$

⋮ ⋮ ⋮

(k+1)st iteration : condition not satisfied then

$$\beta^{*(k+1)} = p_0^{*(k+1)} \hat{\beta}_{(i_1 \dots i_k)} + (1 - p_0^{*(k+1)}) \sum_{j \neq i_1 \dots i_k}^n p(j | i_1 \dots i_k) \hat{\beta}_{(i_1 \dots i_k, j)}$$

and

$$\beta^{*(k+1)} \approx \beta^*$$

This section contains some numerical results specific to the MWRTS GST replacement project.

3.1 Numerical Results

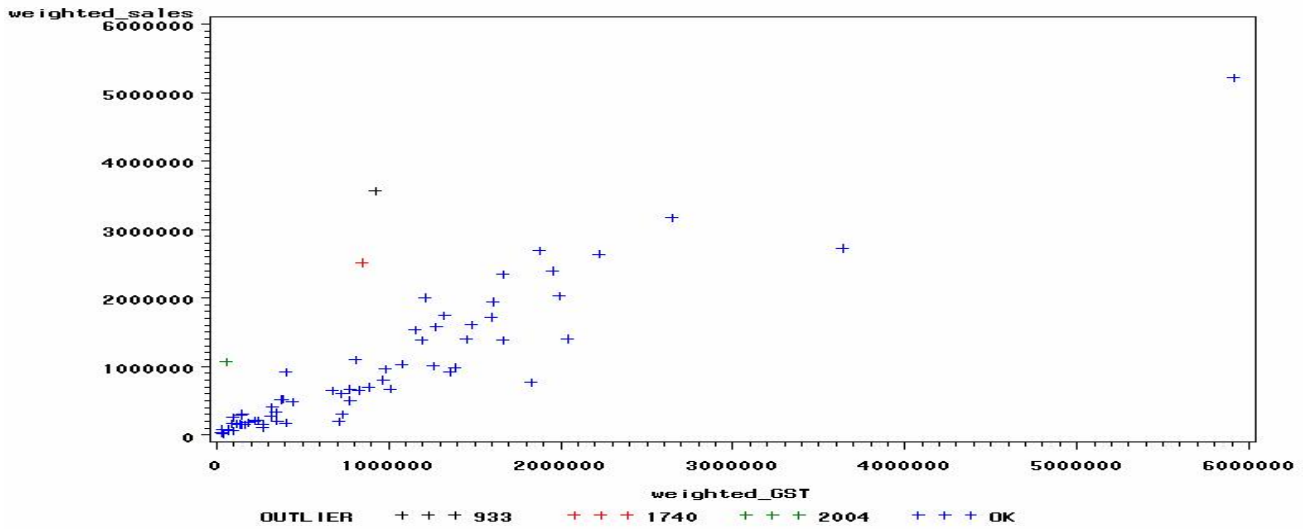
Table 1: A comparison of β^* vs. $\beta^{*(k+1)}$ for various model classes

Model class	Sample size	β^*	$\beta^{*(k+1)}$	Number of outliers
N1205	25	0.98	0.97	5

Y0204	25	0.91	0.91	2
Y0903	25	0.89	0.90	2
N1303	24	0.81	0.81	2
N1605	24	0.88	0.86	2
Y1302	23	0.89	0.92	3
Y0702	22	0.87	0.88	3
N1305	20	0.81	0.80	5
Y0902	20	0.97	0.97	3

Table 1, demonstrates that the sequential procedure produces results equivalent to the full procedure. The data was from the Retail Trade survey and the model class is defined as GST Monthly Remittance (Y or N) x Industry Trade Group (3 digit code) x Geographical Region (1 digit code).

Plot 3.1.2
Plot of weighted sales vs. weighted GST
 Non-monthly remitters in Electronics and Appliance stores



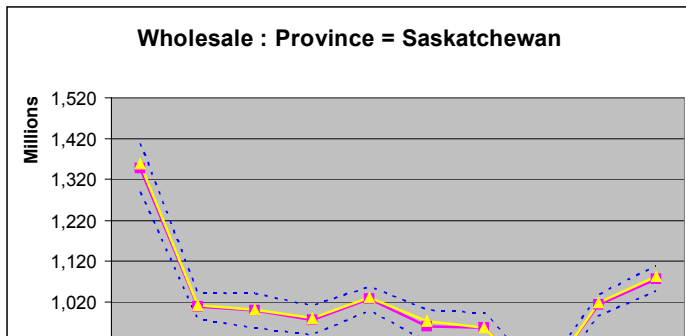
Plot 3.1.2 above reflects a typical distribution of S1 units (n=69) in a model class with the outliers identified by the sequential method. Table 2 gives the output of the iterative process showing why the observations were chosen as being outliers.

Table 2: The sequential outlier detection and modeling process

$p_0^{*(k+1)}$	$p(i_k i_1 \dots i_{k-1})$	Outlier id	$\beta^{*(k+1)}$	R^2
0.0000	1.0000	2004	1.0609	0.8066
0.0000	1.0000	933	1.0191	0.8700
0.0008	1.0000	1740	0.9920	0.9044
0.7761	0.1836	OK	0.9919	0.9116

Plot 3.1.3(a)

Plot 3.1.3(b)



The above two plots compare some of the results of actual production data (labelled original) vs. data with GST replacement (labelled Mod_S2), for Wholesale sales. In Wholesale, approximately 1000 simple units were in S1 and 900 in S2 out of a total sample of approximately 10,000. On the GST replacement side, the S1 and S2 units were chosen based on certain criteria as of June 2004. The results for the province of Saskatchewan from June 2004 to February 2005 comparing what was actually obtained for the sales variable in production to what would have happened under GST replacement is summarized in the Plot 3.1.3(a). The Plot 3.1.3(b) is for Retail Grocery store sales (S1 = 3000, S2 = 1600 and total sample size = 15,000). The dotted lines are 95% upper and lower confidence limits based on the actual production series. Both results were deemed more than satisfactory. In addition, it should be noted out that on the GST replacement side, the S2 units are no longer contributing to the historical or auxiliary trend imputation that is performed to compensate for unit or item non response of surveyed units.

4. Conclusions

The hierarchical Bayesian model that accounts for the possibility of outlying or influential observations is consistent with the data of MWRTS. The resulting parameter estimates when applied to the units subject to GST replacement are able to produce good quality estimates for the MWRTS domains.

4. Acknowledgements

Dennis Batten, François Brisebois and Mark Majkowski are thanked for comments and suggestions that improved this paper.

REFERENCES

- Freeman, P.R. (1980). On the number of outliers in data from a linear model. *Bayesian Statistics, Proceedings of the First International Meeting held in Valencia (Spain)*. J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith (Eds.) 347-382.
- Guttman, I and Pena, D. (1993). A Bayesian look at diagnostics in the univariate linear model. *Statistica Sinica*, **3**, 367-390.
- Petit, L.I. and Smith, A.F.M.(1985). Outliers and influential observations in linear models. *Bayesian Statistics 2*, J.M. Bernardo (Ed.).
- Philips, R. (2001). Outlier Detection Routine used in the Annual Survey of Manufactures (ASM), Statistics Canada Internal Report.