Modelling Progressive Data

Z. Ahmed & L.-C. Zhang
University of Southampton

Abstract

Presently a significant initiative is taken at the National Statistical Offices to exploit the potentials of administrative data in statistical production. For instance, several investigations have previously been carried out at ONS, such as forecasting VAT turnover at the unit-level, adjusting VAT register totals towards the existing MBS-based turnover estimates etc. A critical question is how to estimate the total VAT turnover when the timeliness of VAT reporting is related to VAT turnover i.e. informative reporting? In this paper we develop new methodologies for handling informative reporting, drawing on the relevant methods for informative sampling and informative nonresponse. We assume a model for the outcome variable in the population of in-scope units and model the reporting probability. The two models define the model holding for the outcomes observed for the reporting units. We study maximum likelihood and estimation equation methods of fitting the model, and illustrate our approach using simulations.

1. Introduction:

There is currently a considerable drive at the National Statistical Offices to exploit the potentials of administrative data in statistical production. For instance, several investigations have previously been carried out at ONS, such as forecasting VAT turnover at the unit-level, adjusting VAT register totals towards the existing MBS-based turnover estimates etc.

Many administrative data sources, unlike sample surveys and censuses, do not always have a closing date, after which the data become static and can only be altered in editing. Reporting and registration delays and corrections can occur a long time after the statistical reference date, whether by allowance or negligence. See e.g. Hedlin et al. (2006) for delayed introduction of birth units in the UK BR, Linkletter and Sitter (2007) for delays in Natural Gas Production reports in Texas, and Zhang and Fosen (2012) for delays in the Norwegian Employer/Employee Register. Depending on the situations, input data delays and changes may cause coverage errors or measurement errors, or both, in the integrated data.
Let t be the *reference* time point of interest and $t+d$ the *measurement* time point, for $d \geq 0$. Let $U(t)$ and $y(t)$ be the target population and value at t, respectively. For a unit i, let $I_i(t;t+d)=1$ if the unit is to be included in the target population and 0 otherwise, i.e. based on the information available at $t+d$, and let $y_i(t;t+d)$ be the observed value for t at ... The data are said to be *progressive* if, for $d \neq d' > 0$, we can have

$$I_i(t;t+d) \neq I_i(t;t+d') \text{ or } y_i(t;t+d) \neq y_i(t;t+d')$$

which lead to coverage errors and measurement errors, respectively, or both. Progressiveness is a distinct feature of administrative data sources compared to sample surveys, unless one is determined to overlook all delays and changes after a certain period. Zhang and Pritchard (2013) extended the prediction framework of Valliant et al. (2000) for progressive data and applied it to VAT register data in UK. Zhang and Pritchard (2013) notice potential connections of modelling progressive data to the literature on estimation in the presence of nonresponse and informative sampling.

2. Fitting Reporting Model using MLE Approach:

Let y_i denote the value of an outcome variable Y (say turnover at time t), associated with unit i belonging to an *existent population* $E=\{1,...,N\}$, a part of target population. Let x_i denote the p auxiliary variables (covariates) including historic y-value associated with unit i. Let $R=\{1,...,r\}$ define the *reported population* with reported outcomes and covariates, and $R^c=\{r+1,...,n\}$ define the *unreported population* for which at least the outcomes are not reported (missing).
Following the idea of complex survey modelling under informative sampling given by Pfeffermann, et al. (1998a), we developed the model for reporting population (*R*) when we have the *pdf* of existent population (*E*) and conditional reporting probability model as follows. Let $I_i=1$ if $i \in R$ and $I_i=0$, otherwise.

Suppose, we denote by x_i the covariates and then the *marginal pdf* of the outcome y_i given that unit i is in the reporting population is

$$f_R\left(y_i \mid x_i\right) = f\left(y_i \mid x_i, i \in R\right) = \frac{\Pr\left(I_i = 1 \mid y_i, x_i\right)}{\Pr\left(I_i = 1 \mid x_i\right)} f_E\left(y_i \mid x_i\right). \tag{1}$$

Then the reporting population likelihood is

$$L_R = \prod_{i=1}^{r} f\left(y_i \mid x_i, i \in R; \theta, \gamma\right) = \prod_{i=1}^{r} \frac{\Pr\left(I_i = 1 \mid y_i, x_i; \gamma\right) f_E\left(y_i \mid x_i; \theta\right)}{\Pr\left(I_i = 1 \mid x_i; \theta, \gamma\right)}. \tag{2}$$

The reported model can be fitted under informative reporting if we know the reporting probability model $\Pr\left(I_i = 1 \mid y_i, x_i; \gamma\right)$ and the density of existent population $f_E\left(y_i \mid x_i; \theta\right)$. In literature, different response probability models like; linear, exponential, logit and probit models were used. We can use one of these models in (1) along with *pdf* of existent population to obtain the *pdf* of reporting model.

We considered the following example using exponential reporting model. Let an existent population model have the following normal *pdf*

$$f_R\left(y_i \mid x_i; \theta\right) = \left(1 \big/ \sigma\sqrt{2\pi}\right) \exp\left\{-\left(y_i - \mathbf{x}'\boldsymbol{\beta}\right)^2 \big/ 2\sigma^2\right\} \tag{3}$$

and

$$\Pr\left(I_i = 1 \mid y_i, x_i, i \in E\right) = \exp(\alpha_0 + \alpha_1 y_i + \mathbf{x}'\boldsymbol{\gamma}).$$

(4)

Then

$$\Pr\left(I_i = 1 \mid x_i, i \in E\right) = \int \exp(\alpha_0 + \alpha_1 y_i + \mathbf{x}'\boldsymbol{\gamma}) f_E\left(y_i \mid \mathrm{x_i}\right) dy_i$$

or

$$\Pr\left(I_i = 1 \mid x_i, i \in E\right) = \exp\left\{\alpha_0 + \mathbf{x}'\boldsymbol{\gamma} - \frac{1}{2\hat{\sigma}^2}\left[\left(\mathbf{x}'\boldsymbol{\beta}\right)^2 - \left(\mathbf{x}'\boldsymbol{\beta} + \sigma^2\alpha_1\right)^2\right]\right\} \tag{5}$$

Now using (3), (4) and (5) in (1), we have

$$f_R\left(y_i \mid x_i; \theta, \gamma\right) = \exp\left\{\frac{y_i\left(\alpha_1\sigma^2 + \mathbf{x}'\boldsymbol{\beta}\right) - 2^{-1}\left(\alpha_1\sigma^2 + \mathbf{x}'\boldsymbol{\beta}\right)^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\}. \tag{6}$$

$$f_R\left(y_i \mid x_i; \theta, \gamma\right) \sim N\left(\alpha_1\sigma^2 + \mathbf{x}'\boldsymbol{\beta}, \sigma^2\right).$$

The simplified form of log likelihood function for reported population is

$$l_R = \sum_{i=1}^{r}\left\{\frac{1}{\sigma^2}\left(y_i\left(x_i'\boldsymbol{\beta}\right) - \frac{1}{2}\left(x_i'\boldsymbol{\beta}\right)^2 - \frac{y_i^2}{2}\right) - \log\sqrt{2\pi\sigma^2} + \alpha_1 y_i - \frac{1}{2\sigma^2}\left(\sigma^4\alpha_1^2 + 2x_i'\boldsymbol{\beta}\sigma^2\alpha_1\right)\right\} \tag{7}$$

As in (1), on RHS there is product of two functions. It is possible to have a problem of non-identifiability. The model given in (6) is non-identifiable, because from (7) we cannot obtain unique solution for the unknown parameters. Also the conditions for identifiability given in Pfeffermann and Landsman (2011) cannot be applied. Identifiable model can be obtained using logistic model instead of exponential by imposing the condition that at least one covariate should differ among covariates used for reporting model and density of the existent population (see Pfeffermann and Landsman (2011)).

3. Fitting Reporting Model using Pseudo MLE Approach:

We have seen that using above approach, we are facing difficulty of model identification and it will become more difficult when only historic response value will be used as a covariate (auxiliary information). For a simple alternative we now use the existence and reporting history of each unit to estimate its individual reporting probability. Then Pseudo PMLE method is used to estimate the parameters.

To illustrate this approach, suppose the finite existent population *E* is of size *N*, and for each unit i

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and $\varepsilon_i \sim N\left(0, \sigma_i^2\right)$, with $\sigma_i^2 : \sigma^2$, $\sigma^2 x_i$ and $\sigma^2 x_i^2$.

Let the case of $\sigma_i^2 = \sigma^2$,

$$f_E\left(y_i \mid x_i; \beta, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \beta_0 + \beta_1 x_i}{\sigma}\right)^2\right\}.$$

Then census parameters β_0, β_1 and σ^2 are defined by the following population estimating equations;

$$\sum_E \left(y_i - \beta_0 - \beta_1 x_i\right) = 0\,,$$
$$\sum_E x_i \left(y_i - \beta_0 - \beta_1 x_i\right) = 0\,,$$
$$\sum_E \left(y_i - \beta_0 - \beta_1 x_i\right)^2 - \sigma^2 N = 0\cdot$$

Here we are considering two cases when $\sigma_i^2 = \sigma^2 x_i$ and $\sigma_i^2 = \sigma^2 x_i^2$.

For $\sigma_i^2 = \sigma^2 x_i$,

$$f_E\left(y_i \mid x_i; \beta, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2 x_i}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \beta_0 + \beta_1 x_i}{\sigma\sqrt{x_i}}\right)^2\right\}. \tag{8}$$

Log Likelihood function is

$$\log\left(L\right) = -\frac{N}{2}\log\sigma^2 + \sum_N \log\frac{1}{\sqrt{x_i}} - \frac{N}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_N\left(\frac{1}{\sqrt{x_i}}\left(y_i - \beta_0 + \beta_1 x_i\right)\right)^2 . \tag{9}$$

Then census parameters β_0 , β_1 and σ^2 are defined by the following census estimating equations

$$\sum_E \frac{1}{\sqrt{x_i}}\left(y_i - \beta_0 + \beta_1 x_i\right) = 0\,, \tag{10}$$

$$\sum_E \frac{x_i}{\sqrt{x_i}}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0\,, \tag{11}$$

$$\sum_E \left(\frac{1}{\sqrt{x_i}}\left(y_i - \beta_0 - \beta_1 x_i\right)\right)^2 - \sigma^2 N = 0\,. \tag{12}$$

Let, $\mathbf{W}_N = \mathbf{V}_N^{\mathbf{-1/2}} = diag\left[\hat{w}_i / \sqrt{x_i}\right]$, $\mathbf{X}_N = \left[\mathbf{1} \quad \mathbf{x}\right]$, $\mathbf{X}_N^* = \mathbf{W}_N \mathbf{X}_N$, $\mathbf{Y}_N^* = \mathbf{W}_N \mathbf{Y}_N$ and $\boldsymbol{\beta}_{\mathrm{FP}} = \left(\beta_0, \beta_1\right)^T$ then from (10) and (11), we can write

$$\boldsymbol{\beta}_{\mathrm{FP}} = \left(\mathbf{X}_N^{*T}\mathbf{X}_N^*\right)^{\mathbf{-1}} \mathbf{X}_N^* \mathbf{Y}_N^* \quad \text{and} \quad SD\left(\boldsymbol{\beta}_{\mathrm{FP}}\right) = diag\left(\sqrt{\sigma_N^2\left(\mathbf{X}_N^T \mathbf{V}_N^{\mathbf{-1}} \mathbf{X}_N\right)^{\mathbf{-1}}}\right),$$

and from Eq. (12) $\sigma_N^2 = \dfrac{1}{N}\sum_U\left[\dfrac{1}{x_i}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right] = \dfrac{\mathbf{E}_N^T \tilde{\mathbf{W}}_N \mathbf{E}_N}{N}$,

where $E = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{FP}\right)$ and $\tilde{\mathbf{W}}_N = diag\left(1/x_i\right)$.

To estimate the finite population parameters using PMLE method, generally Pfeffermann (1993) defined that he pseudo MLE (PMLE) of $\boldsymbol{\theta}$ is the solution of sample estimating equations $\hat{U}\left(\boldsymbol{\theta}\right) = 0$, where $\hat{U}\left(\boldsymbol{\theta}\right)$ is design consistent of census estimating equations $U\left(\boldsymbol{\theta}\right)$. The common estimator in the literature is H-T estimator so that the PMLE of $\boldsymbol{\theta}$ is the solution of $\sum_S \dfrac{\mathbf{u}_i\left(y_i;\boldsymbol{\theta}\right)}{\pi_i} = 0$.

In our case, as we are using estimated reporting probabilities, we can write the reporting estimating equations as

$$\sum_R \left(\hat{w}_i\right)\mathbf{u}_i\left(y_i;\boldsymbol{\theta}\right) = 0 \ , \tag{13}$$

where $\boldsymbol{\theta} = \left(\beta_0, \beta_1, \sigma^2\right)$ and $\hat{w}_i = \hat{\pi}_i^{-1}$; $\hat{\pi}_i = R_i / T_i$, where R_i and T_i are reporting and existent history of a unit respectively. Provided R_i follows Binomial distribution with stationary parameter π_i , $\hat{\pi}_i = \dfrac{R_i}{T_i}$ is an unbiased estimator of π_i , since

$$E\left(R_i \mid T_i\right) = T_i \pi_i \ \text{ and } \ V\left(R_i \mid T_i\right) = T_i \pi_i \left(1 - \pi_i\right).$$

and

$$E[E\left(\frac{R_i}{T_i} \mid T_i\right)] = E[\frac{T_i \pi_i}{T_i}] = E[\pi_i] = \pi_i\,.$$

For $\sigma_i^2 = \sigma^2 x_i$ **,** The reporting estimating equations can be written as:

$$\sum_R \frac{\hat{w}_i}{\sqrt{x_i}}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0\,, \tag{14}$$

$$\sum_R \frac{\hat{w}_i}{\sqrt{x_i}}\left(y_i - \beta_0 - \beta_1 x_i\right) x_i = 0\,, \tag{15}$$

$$\sum_R \hat{w}_i \left\{\left(\frac{1}{\sqrt{x_i}}\left(y_i - \beta_0 - \beta_1 x_i\right)\right)^2 - \sigma^2\right\} = 0\,. \tag{16}$$

Let define $\hat{\mathbf{W}}_r = \hat{\mathbf{V}}_{\mathbf{r}}^{\mathbf{-1/2}} = diag\left[\hat{w}_i / \sqrt{x_i}\right]$, $\mathbf{X}_r = [\mathbf{1} \quad \mathbf{x}]$, $\mathbf{X}_{\mathbf{r}}^{*} = \hat{\mathbf{W}}_{\mathbf{r}}\mathbf{X}_{\mathbf{r}}$, $\mathbf{Y}_{\mathbf{r}}^{*} = \hat{\mathbf{W}}_{\mathbf{r}}\mathbf{Y}_{\mathbf{r}}$ and $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1\right)^T$ then from (14) and (15), we can write

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_{\mathbf{r}}^{*T}\mathbf{X}_{\mathbf{r}}^{*}\right)^{\mathbf{-1}}\mathbf{X}_{\mathbf{r}}^{*}\mathbf{Y}_{\mathbf{r}}^{*},\ SE\left(\hat{\boldsymbol{\beta}}\right) \approx \sigma\sqrt{\left(\mathbf{X}_{\mathbf{r}}^{T}\mathbf{V}_{\mathbf{r}}^{\mathbf{-1}}\mathbf{X}_{\mathbf{r}}\right)^{\mathbf{-1}}} \text{ and } SE\left(\hat{\boldsymbol{\beta}}\right) \approx \hat{\sigma}\sqrt{\left(\mathbf{X}_{\mathbf{r}}^{T}\hat{\mathbf{V}}_{\mathbf{r}}^{\mathbf{-1}}\mathbf{X}_{\mathbf{r}}\right)^{\mathbf{-1}}}\,,$$

and from (16)

$$\hat{\sigma}^2 = \frac{\sum_r \left[\frac{\hat{w}_i}{x_i}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\right]}{\sum_r \hat{w}_i} = \frac{\mathbf{E}_r^T \hat{\tilde{\mathbf{W}}}_r \mathbf{E}_r}{\sum_r \hat{w}_i}\text{, where } \hat{\mathbf{E}} = \left(\mathbf{y}_r - \mathbf{X}_r\hat{\boldsymbol{\beta}}\right) \text{ and } \hat{\tilde{\mathbf{W}}}_r = diag\left(\hat{w}_i / x_i\right).$$

For $\sigma_i^2 = \sigma^2 x_i^2$, we need to have $\mathbf{W}_N = diag\left[1/x_i\right]$, $\tilde{\mathbf{W}}_N = diag\left(1/x_i^2\right)$, $\mathbf{W}_r = diag\left[\hat{w}_i / x_i\right]$ and $\hat{\tilde{\mathbf{W}}}_r = diag\left(\hat{w}_i / x_i^2\right)$.

4. Simulation Study

To illustrate the estimation of parameters using pseudo MLE**,** let, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, x_i □ $rbeta(3,2)$ and suppose $\beta_0 = 0.5$, $\beta_1 = 5$ and ε_i □ $N\left(0, \sigma_i^2\right)$, with σ_i^2 : $\sigma^2 x_i$ and $\sigma^2 x_i^2$; $\sigma^2 = 2$ (say). Further let $T_i \sim Bin(10, 0.60)$ be the existent history, $\pi_i \sim U(1, 0.6)$ be the probability of reporting, $R_i \sim Bin(T_i, \pi_i)$ be the reporting history and $r_i \sim Bin(1, \pi_i)$ be the present reporting indicator. We can write $\hat{\pi}_i = R_i / T_i$. From (13) for PMLE we require reporting weights, we can consider the estimated weights $\hat{w}_i = \hat{\pi}_i^{-1}$. We also tried the following alternative weights:

i. Ratio adjusted to existent population total, i.e. $\hat{w}_{2i} = \hat{w}_i \sum_E x_i \Big/ \sum_R \hat{w}_i x_i$

ii. Calibrated (GREG) using constraint $\sum_E x_j = \sum_R w_j x_j$, i.e.

$$\hat{w}_{3i} = \hat{w}_i + \left[1 + \left(\sum_E x_j - \sum_R \hat{w}_j x_j\right)^T \left(\sum_R \hat{w}_j x_j x_j^T\right)^{-1} x_i\right].$$

The following table shows the results of average estimates of the parameters and their empirical standard errors for 1000 simulations of randomly selected reporting population using the reporting indicator r_i from an existent population of size N = 3000. The average reporting population is 2400.533.

Table: Mean Estimates and Empirical SE

Parameter/Estimate	$\sigma_i^2 = 2x_i$						$\sigma_i^2 = 2x_i^2$						**Weights**
	Mean Estimates			**Empirical SE**			**Mean Estimates**			**Empirical SE**			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	
POP	**0.5136**	**4.9834**	**1.9693**	**0.0469**	**0.0852**	**0.0512**	**0.4995**	**5.0018**	**1.9973**	**0.0254**	**0.0565**	**0.0526**	
PMLE	0.5129	4.9839	1.9717	0.0556	0.1013	0.0619	0.4993	5.0024	1.9956	0.0310	0.0682	0.0628	$\hat{w}_i$
	0.5129	4.9839	1.9717	0.0556	0.1013	0.0619	0.4993	5.0024	1.9956	0.0310	0.0682	0.0628	$\hat{w}_{2i}$
	0.5130	4.9838	1.9711	0.0553	0.1007	0.0619	0.4994	5.0024	1.9956	0.0306	0.0674	0.0628	$\hat{w}_{3i}$

The simulation results seem encouraging. The theoretical properties of the PMLE are currently being investigated, including the definition of an asymptotic setting that is suitable for progressive data.

References

1. Hedlin, D., Fenton, T., McDonald, J.W., Pont, M. and Wang, S. (2006). Estimating the under coverage of a sampling frame due to reporting delays. *Journal of Official Statistics*, 22, 53-70.
2. Linkletter, C.D. and Sitter, R.R. (2007). Predicting natural gas production in Texas: An application of nonparametric reporting lad distribution estimation. *Journal of Official Statistics*, 23, 239-251.
3. Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, 61, 317-337.
4. Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.
5. Pfeffermann, D. and Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics,* 5, 1726–1751.
6. Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, Inc.
7. Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.
8. Zhang, L.-C. and Pritchard, A. (2013) Short-term turnover statistics based on VAT and Monthly Business Survey data sources. *ENBES workshop 2013, Nuremberg.*
9. Zhang, L.-C. (2014). *Data Integration*. The Survey Statistician No 70, International Association of Survey Statisticians.
10. Zhang, L.-C. and Fosen, J. (2012). A modelling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics*, 66, 91-104.