

# MODELING PROGRESSIVE DATA

Zahoor Ahmad and Li-Chun Zhang  
University of Southampton, UK

September 7, 2015

1. Introduction
  - 1.1 Motivation
  - 1.2 Progressive data
  - 1.3 Informative reporting
2. Informative sampling approach adapted
3. Alternative Estimated Pseudo MLE (EPMLE) approach
4. Simulation Study
5. Theoretical properties

There is currently a considerable drive at the National Statistical Offices to exploit the administrative data in statistical production. A number of 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, however, remains:

*How to estimate the total VAT turnover if delays in VAT reporting is related to VAT turnover?*

# Progressive Data and Informative Reporting

- ▶ The data that mature *progressively*, e.g. VAT turnover.
- ▶ Observation depends on the value of outcome variable.

Zhang and Pritchard (2013) extend the standard prediction framework (e.g. Valliant et al., 2000) to 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 informative nonresponse or sampling.

# 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*  $E = \{1, \dots, N\}$ , a part of target population.

Let  $x_i$  denote the  $p$  auxiliary variables (covariates) including, possibly the historic  $y$ -values associated with unit  $i$ .

Let  $R = \{1, \dots, r\}$  be the *reporting* part of  $E$  with observed  $(y_i, x_i)$ , and let  $R^c = \{r + 1, \dots, n\}$  be the rest part for which the outcomes are not reported (missing) for the time being.

Consider first the approach to informative sampling (Pfeffermann, et al., 1998a)

# Fitting Reporting Model using MLE Approach

Let  $I_i = 1$  if  $i \in R$  and  $I_i = 0$ , otherwise.

The model of  $y_i$  given  $x_i$  in the reporting population is

$$f_R(y_i|x_i) = f(y_i|x_i, i \in R) = \frac{\Pr(I_i = 1|y_i, x_i)}{\Pr(I_i = 1|x_i)} f_E(y_i|x_i). \quad (1)$$

Then the reporting likelihood is

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

under the assumed model  $\Pr(I_i = 1|y_i, x_i; \gamma)$  and  $f_E(y_i|x_i; \theta)$ .

# Fitting Reporting Model using MLE Approach

Different response probability models in the literature include linear, exponential, logit and probit.

Consider e.g. the exponential model.

$$\Pr(I_i = 1|y_i, x_i, i \in E) = \exp(\alpha_0 + \alpha_1 y_i + x'_i \gamma). \quad (3)$$

Let the existent population model have the normal *pdf*

$$f_R(y_i|x_i; \theta) = \left(1/\sigma\sqrt{2\pi}\right) \exp\left\{-(y_i - x'_i\beta)^2/2\sigma^2\right\} \quad (4)$$

Then

$$f_R(y_i|x_i; \theta, \gamma) \sim N(\alpha_1\sigma^2 + x'_i\beta, \sigma^2). \quad (5)$$

# Fitting Reporting Model using MLE Approach

The log reporting likelihood function is

$$l_R = \sum_{i=1}^r \left\{ \frac{1}{\sigma^2} \left( y_i (x_i' \beta) - \frac{1}{2} (x_i' \beta)^2 - \frac{y_i^2}{2} \right) \right\} \\ - r \left\{ \log \sqrt{2\pi\sigma^2} + \sum_{i=1}^r (\alpha_1 y_i) - \frac{1}{2\sigma^2} \sum_{i=1}^r (\sigma^4 \alpha_1^2 + 2x_i' \beta \sigma^2 \alpha_1) \right\} \quad (6)$$

As the model holding for reporting population is product of two functions. It is possible to have a problem of non-identifiability. The reporting population model given in (5) is non-identifiable while using exponential reporting probability model, because from likelihood function given in (6), we cannot obtain unique solution for the unknown parameters. Also the conditions for identifiability given in Pfeiffermann and Landsman (2011) cannot hold.

Identifiable reporting model can be obtained if the probability of reporting is modeled 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 Pfeiffermann and Landsman (2011)).

# An alternative approach

In the minimum data scenario, with only the observed historic values as available covariates, identification could be challenging.

As a simple alternative, we explore the use of the reporting history of each unit to estimate its individual reporting probability, and the corresponding estimated Pseudo MLE (EPMLE) approach.

# EPMLE: An illustration

To illustrate, suppose existent population  $E$  of size  $N$ , and for each unit  $i$ ,  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ , and  $\varepsilon_i \sim N(0, \sigma_i^2)$ , with  $\sigma_i^2 = \sigma^2 x_i$ .

$$f_E(y_i|x_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 x_i}} \exp \left\{ -\frac{1}{2} \left( \frac{y_i - \beta_0 + \beta_1 x_i}{\sigma \sqrt{x_i}} \right)^2 \right\}. \quad (7)$$

Log Likelihood function is

$$\begin{aligned} \log(L) = & -\frac{N}{2} \log \sigma^2 + \sum_N \log \frac{1}{\sqrt{x_i}} - \frac{N}{2} \log 2\pi \\ & - \frac{1}{2\sigma^2} \sum_N \left( \frac{1}{\sqrt{x_i}} (y_i - \beta_0 + \beta_1 x_i) \right)^2. \end{aligned} \quad (8)$$

The *census* parameters of  $\beta_0$ ,  $\beta_1$  and  $\sigma^2$  are defined as the solutions to the following census estimating equations

$$\sum_E \frac{1}{\sqrt{x_i}} (y_i - \beta_0 + \beta_1 x_i) = 0, \quad (9)$$

$$\sum_E \frac{x_i}{\sqrt{x_i}} (y_i - \beta_0 - \beta_1 x_i) = 0, \quad (10)$$

$$\sum_E \left( \frac{1}{\sqrt{x_i}} (y_i - \beta_0 - \beta_1 x_i) \right)^2 - \sigma^2 N = 0. \quad (11)$$

# EPMLE: An illustration

Let,  $W_N = V_N^{-1/2} = \text{diag} [\hat{w}_i / \sqrt{x_i}]$ ,  $X_N = \begin{bmatrix} 1 & x \end{bmatrix}$ ,  
 $X_N^* = W_N X_N$ ,  $Y_N^* = W_N Y_N$  and  $\beta_{FP} = (\beta_0, \beta_1)^T$  then,

$$\beta_{FP} = \left( X_N^{*T} X_N^* \right)^{-1} X_N^* Y_N^*$$

and

$$\sigma_{FP}^2 = \frac{1}{N} \sum_U \left[ \frac{1}{x_i} (y_i - \beta_{0FP} - \beta_{1FP} x_i)^2 \right] = \frac{E_N^T \tilde{W}_N E_N}{N}$$

where  $E = (y - X\beta_{FP})$  and  $\tilde{W}_N = \text{diag} (1/x_i)$ .

# EPMLE: An illustration

The pseudo MLE (PMLE) of  $\theta$  (e.g. Pfeiffermann, 1993) is the solution of sample estimating equations  $\hat{U}(\theta) = 0$ , where  $\hat{U}(\theta)$  is design consistent of the census estimating equations  $U(\theta) = \sum_U u_i(y_i, \theta)$ .

The common estimator in the literature is H-T estimator so that the PMLE of  $\theta$  is the solution of

$$\sum_S \frac{u_i(y_i; \theta)}{\pi_i} = 0.$$

In our case, however, the reporting probability is unknown, so we write the reporting estimating equations as

$$\sum_R \hat{w}_i u_i(y_i; \theta) = 0, \quad (12)$$

where  $\theta = (\beta_0, \beta_1, \sigma^2)$  and  $\hat{w}_i = \hat{\pi}_i^{-1}$ .

Example:  $\hat{\pi}_i = R_i / T_i$ , where  $R_i$  and  $T_i$  are the historic reporting and existence counts, provided  $R_i$  follows the binomial distribution with parameter  $\pi_i$ .

For  $\sigma_i^2 = \sigma^2 x_i$ , The reporting estimating equations are given by

$$\sum_R \frac{\hat{w}_i}{\sqrt{x_i}} (y_i - \beta_0 - \beta_1 x_i) = 0, \quad (13)$$

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

$$\sum_R \hat{w}_i \left\{ \left( \frac{1}{\sqrt{x_i}} (y_i - \beta_0 - \beta_1 x_i) \right)^2 - \sigma^2 \right\} = 0. \quad (15)$$

# EPMLE: An illustration

Let  $\hat{W}_r = \hat{V}_r^{-1/2} = \text{diag} [\hat{w}_i / \sqrt{x_i}]$ ,  $X_r = [1 \ x]$ ,

$X_r^* = \hat{W}_r X_r$ ,  $Y_r^* = \hat{W}_r Y_r$  and  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$  then

$$\hat{\beta} = (X_r^{*T} X_r^*)^{-1} X_r^* Y_r^*, \quad SE(\hat{\beta}) \approx \sigma \sqrt{(X_r^T V_r^{-1} X_r)^{-1}}$$

$$\text{and } \widehat{SE}(\hat{\beta}) \approx \hat{\sigma} \sqrt{(X_r^T \hat{V}_r^{-1} X_r)^{-1}}.$$

And

$$\hat{\sigma}^2 = \frac{\sum_r \left[ \frac{\hat{w}_i}{x_i} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \right]}{\sum_r \hat{w}_i} = \frac{E_r^T \hat{\tilde{W}}_r E_r}{\sum_r \hat{w}_i},$$

where  $\hat{E} = (y_r - X_r \hat{\beta})$  and  $\hat{\tilde{W}}_r = \text{diag}(\hat{w}_i/x_i)$ .

For  $\sigma_i^2 = \sigma^2 x_i^2$ , we need to have  $W_N = \text{diag}[1/x_i]$ ,

$\tilde{W}_N = \text{diag}(1/x_i^2)$ ,  $W_r = \text{diag}[\hat{w}_i/x_i]$  and  $\hat{\tilde{W}}_r = \text{diag}(\hat{w}_i/x_i^2)$ .

Let

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ where } \beta_0 = 0.5, \beta_1 = 5,$$

$$\varepsilon_i \sim N(0, \sigma_i^2), \text{ with } \sigma_i^2: \sigma^2 x_i \text{ and } \sigma^2 x_i^2, \text{ where } \sigma^2 = 2$$

$$x_i \sim \text{rbeta}(3, 2)$$

Further let,

$T_i \sim \text{Bin}(10, 0.60)$  be the existent history,

$\pi_i \sim U(0.6, 1)$  be the probability of reporting,

$R_i \sim \text{Bin}(T_i, \pi_i)$  be the reporting history and

$r_i \sim \text{Bin}(1, \pi_i)$  be the current reporting indicator.

# Simulation Study

We can write  $\hat{\pi}_i = R_i/T_i$ . For EPMLE, consider  $\hat{w}_i = \hat{\pi}_i^{-1}$ .

We also tried the following alternative weights:

1. Ratio adjusted to existent population total, i.e.

$$\hat{w}_{2i} = \hat{w}_i \sum_E x_i / \sum_R \hat{w}_i x_i$$

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

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

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.

# Simulation Study

Table: Mean Estimates and Empirical SE when  $\pi_i \sim U(0.6, 1)$

$\sigma_i^2 = 2x_i$							
	Mean Estimates			Empirical SE			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	Weights
<b>POP</b>	<b>0.5136</b>	<b>4.9834</b>	<b>1.9693</b>	<b>0.0469</b>	<b>0.0852</b>	<b>0.0512</b>	
<b>EPMLE</b>	0.5129	4.9839	1.9717	0.0556	0.1013	0.0619	$\hat{w}_{1i}$
	0.5129	4.9839	1.9717	0.0556	0.1013	0.0619	$\hat{w}_{2i}$
	0.5130	4.9838	1.9711	0.0553	0.1007	0.0619	$\hat{w}_{3i}$
$\sigma_i^2 = 2x_i^2$							
<b>POP</b>	<b>0.4995</b>	<b>5.0018</b>	<b>1.9973</b>	<b>0.0254</b>	<b>0.0565</b>	<b>0.0526</b>	
<b>EPMLE</b>	0.4993	5.0024	1.9956	0.0310	0.0682	0.0628	$\hat{w}_{1i}$
	0.4993	5.0024	1.9956	0.0310	0.0682	0.0628	$\hat{w}_{2i}$
	0.4994	5.0024	1.9956	0.0306	0.0674	0.0628	$\hat{w}_{3i}$

# Simulation Study

For the case of informative reporting, we can allow  $\pi_i$  to depend on outcome variable and then let  $\pi_i = [1 + \exp(-0.45y_i)]$ . The simulation results for 1000 populations of size 3000 are,

Table: Mean Estimates and Empirical SE,  $\pi_i = [1 + \exp(-0.45y_i)]$

$\sigma_i^2 = 2x_i$							
	Mean Estimates			Empirical SE			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	Weights
<b>POP</b>	<b>0.5144</b>	<b>4.9828</b>	<b>1.9680</b>	<b>0.0481</b>	<b>0.0878</b>	<b>0.0490</b>	
<b>EPMLE</b>	0.4955	4.9700	1.9874	0.0594	0.1046	0.0626	$\hat{w}_{1i}$
	0.4955	4.9700	1.9874	0.0594	0.1046	0.0626	$\hat{w}_{2i}$
	0.4957	4.9695	1.9869	0.0592	0.1043	0.0626	$\hat{w}_{3i}$
$\sigma_i^2 = 2x_i^2$							
<b>POP</b>	<b>0.4995</b>	<b>5.0022</b>	<b>1.9985</b>	<b>0.0255</b>	<b>0.0575</b>	<b>0.05204</b>	
<b>EPMLE</b>	0.4999	4.9743	2.0112	0.0316	0.0676	0.0639	$\hat{w}_{1i}$
	0.4999	4.9743	2.0112	0.0316	0.0676	0.0639	$\hat{w}_{2i}$
	0.4999	4.9743	2.0111	0.0313	0.0676	0.0641	$\hat{w}_{3i}$

# Theoretical Properties

Standard PMLE for survey data: known  $\pi_i$ 's. Estimating equation theory: consistency of estimators and the central limit theorem can be proved under some regularity conditions.

Common modelling approach to  $\pi_i$ : assume global parameters, which can be estimated consistently, from which the consistency of the EPMLE follows.

In case above: we estimate  $\pi_i$  based on *limited* history of each unit. The asymptotic setting is  $N \rightarrow \infty$ , where  $T_i = O(1)$  and  $\hat{\pi}_i - \pi_i = O_p(1)$  and  $E(\hat{\pi}_i) = \pi_i$ .

Needs a different proof of the theoretical properties of EPMLE.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, 61, 317-337.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

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.

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

Zhang, L.-C. and Pritchard, A. (2013). *Short-term turnover statistics based on VAT and Monthly Business Survey data sources*. ENBES workshop 2013, Nuremberg.