INDIAN INSTITUTE OF TECHNOLOGY KANPUR

MTH511A: STATISTICAL SIMULATION AND DATA ANALYSIS

A split questionnaire survey design in the context of statistical matching

Created by

HARSHIT PANDEY

Contents

1 ABSTRACT

In this paper, we are tackling the problem of splitting a long questionnaire into two parts, where each participant only responds to a limited amount of questions, and all respondents receive a common portion of questions. We are proposing a method that combines regression models to the two independent samples (questionnaires) in the survey. Each sample includes the common response variable Y and common covariate x, while two vectors of specific covariates z and w are recorded such that no single sampling unit has answered both z and w.This is a problem of statistical matching which we tackle under the assumption of conditional independence. In the statistical matching context, we use a macro approach to estimate parameters of a regression model. This means that we can estimate the joint distribution of all variables of interest with available data utilizing the assumption of conditional independence. We make use of this here by fitting three regression models with the same response variable for each model. Combining the three models allows us to obtain a prediction model with all covariates in common. We compare the performance of our proposed method in simulation studies. The proposed method is easy to use on real life data and it does not require the formulation of a model for the covariates itself as well as it doesn't requires an imputation model for the missing covariates vectors z and w.

2 INTRODUCTION

Statistics, as an academic and professional discipline, is the collection, analysis and interpretation of data. In order to collect data we need a proper questionnaire to get accurate response from respondents about their attitudes, experiences, or opinions. Questionnaires can also be used to collect quantitative and qualitative information. Long questionnaires are pretty common but have a negative effect on response rate, abandonment rate, impacting sample representativeness, and data quality. To address these concerns, one approach may be to shorten the questionnaire. The biggest challenge is that we must decide which questions we have to eliminate from the questionnaire. An alternative to completely eliminating questions from the full survey is to divide the questionnaire into subsets of questions and then getting response of each subset. This creates a shorter questionnaire that still collects the necessary information from at least some of the sample members. These designs are often referred to as Split Questionnaire Design(SQD). SQD selects two or more independent samples, with a bit of overlap in different components.

Through this paper, we are trying to lessen the burden on respondents by splitting the questionnaire. We will show that SQD produces results similar to a full questionnaire with one drawback; the power is decreased due to a decreased number of observations for each variable.

In this paper, we fit a regression model of continuous response variable Y on several covariate and then calculate a predictive model with preferably small prediction error based on the observed data from the SQD. We divide the questionnaire into various components so that each participant needs to respond only a fraction of the total components. Throughout the entire paper,

$$\begin{aligned} \text{s} &: \text{ random sample from population} \\ s_a,\ sb &: \text{ two sub-samples from s s.t,} \\ & s_a \cup s_b = s \text{ and } s_a \cap s_b = \phi \\ \\ \text{Y} &: \text{ continuous response variable} \\ \text{x} &: \text{ continuous covariates} \\ \text{z} &: \text{ categorical covariates} \\ \text{w} &: \text{ categorical covariates} \end{aligned}$$

We split the full questionnaire into two parts; that is, we draw two independent random samples s_a and s_b from the same population. The information about common variables Y and x is observed from both samples, while the information about the specific variables z is available in sample s_a only and w is recorded in sample s_b only.

The integration of two (or more) independent samples in order to calculate a joint distribution of all the variables of interest is usually known as statistical matching (or data fusion) problem. There are two commonly used ways to pursue statistical matching: the macro approach and the micro approach. In this paper, we are using the macro approach for calculating the joint distribution of all the variables of interest.

The statistical problem which we encounter in our paper can be considered as statistical matching, since there does not exist

a single observation which simultaneously contains complete information of all the specific variables of interest(z and w). Most of the SQD papers which are published mostly deal with missing values in the data with some kind of imputation like MICE, CART, logitisc regression etc.
We propose a SQD in the context of statistical matching relying on the assumption of Conditional independence.
The CI is a strong assumption and cannot be tested with available data. If the CI assumption does not hold true, leading to serious bias in the resulting joint relationships among variables of interest. The method can be used for both, the micro and macro approaches in statistical matching.

The goal is to fit a regression model of the form

$$Y = \beta_{0p} + x\beta_{xp} + z\beta_{zp} + w\beta_{wp} + \epsilon \tag{1}$$

where β_{zp} is a column vector with the same dimension as row vector z and likewise, β_{wp} is column vector with matching dimension to row vector w.
ϵ is a residual from the location-scale family and Y is a continuous response variable, and the covariates x, z and w are as described above. The index p is used for the pooled estimates.
We estimate this regression model with our proposal and other regular methods such as MICE, CART and logistic regression.

Our aim is to show that error from our proposed method is comparatively close and even outperforms other methods when all the assumptions are satisfied.

3 PROPOSED METHOD UNDER CI ASSUMPTION

3.1 Conditional independence assumption

The most popular method for missing data is the complete case analysis, which excludes all missing observations from the analysis. Due to not jointly having observed variables z and w, the joint distribution of (y, x, z,w) is not identifiable. We assume CI to overcome the problem of identification. Specifically, we assume that z and w are conditionally independent given Y and x and x only.
That is, we consider:

$$z \perp w \mid \{Y, x\} \quad and \quad z \perp w \mid x \tag{2}$$

we are interested in estimating the conditional distribution of Y given x, z, w. Given (2), we can estimate all parts of our conditional distribution directly with available data. Following the chain rule, the joint distribution of all variables of interest can be factorized as:

$$f(z, w, x, y) = f(z|x, y, w)f(w|x, y)f(y|x)f(x) = f(z|x, y)f(w|x, y)f(y|x)f(x) \tag{3}$$

If the CI assumption is true, the available data of both samples, s_a and s_b is sufficient to estimate (3)

3.2 Proposed method

Define a simple random sample s of size n be drawn from the population of interest and divide this sample into two non-overlapping sub-samples s_a and s_b of sizes n_a and n_b respectively.
For sample sa we obtain the data $(Y_j, x_j, z_j) : j \in s_a$, and sample $s_b = s\ s_a$.
This yields the information $(Y_k, x_k, w_k) : k \in s_b$. We are interested in fitting a regression model for Y regressed on the covariates x, z and w. The conditional probability of Y given x, z and w equals

$$f(y \mid x, z, w) = \frac{f(z, x, y, w)}{f(z, w, x)} \tag{4}$$

Applying the chain rule and the second CI assumption in (2) using (3) and (4) we obtain

$$\begin{aligned} f(y \mid x, z, w) &= \frac{f(z \mid x, y) f(w \mid x, y) f(y \mid x) f(x)}{f(z \mid x) f(w \mid x) f(x)} \\ &= \frac{f(z \mid x, y)}{f(z \mid x)} \cdot \frac{f(w \mid x, y)}{f(w \mid x)} \cdot f(y \mid x). \end{aligned} \tag{5}$$

The distribution of Y and z given x can be decomposed as

$$f(y, z \mid x) = f(y \mid x) f(z \mid y, x). \tag{6}$$

With (6) we can transform the first ratio in (5) into

$$\frac{f(z \mid x, y)}{f(z \mid x)} = \frac{f(y, z \mid x)}{f(z \mid x)\text{\c{f}}(y \mid x)} = \frac{f(y \mid x, z)}{f(y \mid x)} \tag{7}$$

Rearranging the middle term in (5) in the same way using (6) and (7) leads to

$$f(y \mid x, z, w) := \frac{f(y \mid x, z) \cdot f(y \mid x, w)}{f(y \mid x)} \tag{8}$$

We see that we can split the conditional distribution of Y into conditional distributions with only parts of the covariates given. These in turn can be estimated from the data obtained by the splitted design.

One can, in principle, use any model to estimate the three densities.We here propose to work with simple ordinary least squared method. The estimated version of (8) is then

$$\hat{f}(y \mid x, z, w) := \frac{\hat{f}(y \mid x, z) \cdot \hat{f}(y \mid x, w)}{\hat{f}(y \mid x)} \tag{9}$$

where the different components in (9) are fitted assuming

$$Y \mid x, z \sim N(\beta_{0a} + x\beta_{xa} + z\beta_{za}, \sigma_a^2) \tag{10}$$

$$Y \mid x, w \sim N(\beta_{0b} + x\beta_{xb} + w\beta_{wb}, \sigma_b^2) \tag{11}$$

$$Y \mid x \sim N(\beta_0 + x\beta_x, \sigma^2) \tag{12}$$

Note that models (10)–(12) hold jointly only in case of a multivariate normal distribution. We do not assume this generally but make use of (10)–(12) as approximation. Using model (9), we propose to:

1. fit model (10) with the data from s_a to get $\hat{f}(y \mid x, z)$ in (9),
2. fit model (11) with the data from s_b to get $\hat{f}(y \mid x, w)$ in (9) and
3. fit model (12) with the data from s to get $\hat{f}(y \mid x)$ in (9).

The corresponding estimates are

1. $\hat{\mu}_a = \hat{\beta}_{0a} + x\hat{\beta}_{xa} + z\hat{\beta}_{za}$ and $\hat{\sigma}_a^2$
2. $\hat{\mu}_b = \hat{\beta}_{0b} + x\hat{\beta}_{xb} + w\hat{\beta}_{wb}$ and $\hat{\sigma}_b^2$
3. $\hat{\mu} = \hat{\beta}_0 + x\hat{\beta}_x$ and $\hat{\sigma}^2$.

These estimates are calculated using the ordinary least squares method. The error term of each model from (10) to (12) are assumed to be independent and identically distributed with mean zero and constant variance. Using the normal densities from (10) to (12) we can write down the factorized density (9) as

$Y \mid x, z, w \overset{\text{iid}}{\sim} N(\mu_p, \sigma_p^2)$ where

$$\hat{\sigma}_p^2 = \frac{1}{\hat{\sigma}_a^{-2} + \hat{\sigma}_b^{-2} - \hat{\sigma}^{-2}}$$

and

$$\hat{\mu}_p^2 = \hat{\sigma}_p^2 \{ \frac{\hat{\mu}_a}{\hat{\sigma}_a^2} + \frac{\hat{\mu}_b}{\hat{\sigma}_b^2} - \frac{\hat{\mu}}{\hat{\sigma}^2} \}$$

The index p indicates the pooled estimates. Note that the maximized log likelihood for model (10), (11) and (12) is

$$l(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2} \cdot \log(\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is the Maximum Likelihood Estimate in model (10), (11) or (12) respectively. Since model (12) is nested in (10) and (11) we have

$$-\frac{n}{2} \log(\hat{\sigma}_a^2) \geq -\frac{n}{2} \log(\hat{\sigma}2)$$
$$\iff \frac{1}{\hat{\sigma}_a^2} \geq \frac{1}{\hat{\sigma}^2}$$

This proves that

$$\hat{\sigma}_p^2 = \frac{1}{\hat{\sigma}_a^{-2} + \hat{\sigma}_b^{-2} - \hat{\sigma}^{-2}} \geq 0$$

where the equality hold if both $\hat{\beta}_{za} \equiv 0$ and $\hat{\beta}_{wb} \equiv 0$, which happens with probability zero. The above estimates can be combined to provide the final fit for original regression model (1) through $\hat{\beta}_{0p} + x\hat{\beta}_{xp} + z\hat{\beta}_{zp} + w\hat{\beta}_{wp}$, where

$$\hat{\beta}_{0p} = \hat{\sigma}^2_{0p}\{\frac{\hat{\beta}^2_{0a}}{\hat{\sigma}^2_a} + \frac{\hat{\beta}_{0p}}{\hat{\sigma}^2} - \frac{\hat{\beta}_{0p}}{\hat{\sigma}^2}\},$$
$$\hat{\beta}_{xp} = \hat{\sigma}^2_{p}\{\frac{\hat{\beta}^2_{xa}}{\hat{\sigma}^2_a} + \frac{\hat{\beta}_{xp}}{\hat{\sigma}^2_b} - \frac{\hat{\beta}_{0p}}{\hat{\sigma}^2}\},$$
$$\hat{\beta}_{zp} = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_a}\hat{\beta}_{za} \text{ and } \hat{\beta}_{wp} = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_b}\hat{\beta}_{wb}$$

4 SIMULATION

To show the performance of our proposed method, we generate data from the regression model
$Y = \beta_{0p} + x\beta_{xp} + z\beta_{zp} + w\beta_{wp} + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$, Y is continuous response, x is a continuous covariate, and
$z = (z_1, z_2)$ and $w = (w_1, w_1)$ are vectors of binary covariates which are correlated with x.

How to sample from $Y = \beta_{0p} + x\beta_{xp} + z\beta_{zp} + w\beta_{wp} + \epsilon$?
Let Ƶ follows trivariate standard normal distribution.

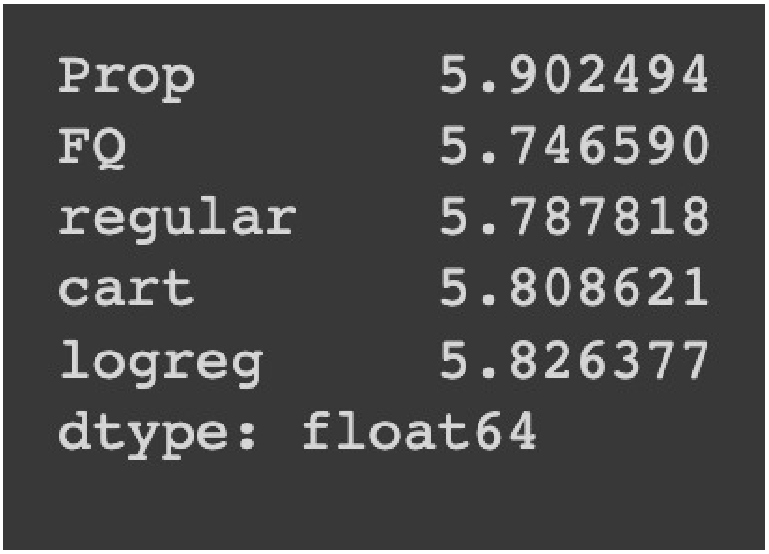
$$\text{Ƶ} = \begin{pmatrix} z1 \\ z2 \\ z3 \end{pmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} , \begin{bmatrix} 1 & cov(z_1, z_2) & cov(z_1, z_3) \\ cov(z_2, z_1) & 1 & cov(z_2, z_3) \\ cov(z_3, z_1) & cov(z_3, z_2) & 1 \end{bmatrix})$$

$X = Z_1 \sim N(0, 1)$
$W = I(Z_2 > 0) \sim Ber(0.5)$
$Z = I(Z_3 > 0) \sim Ber(0.5)$

$P(w = 1) = P(Z_2 > 0) = 1/2$
$P(Z = 1) = P(Z_3 > 0) = 1/2$

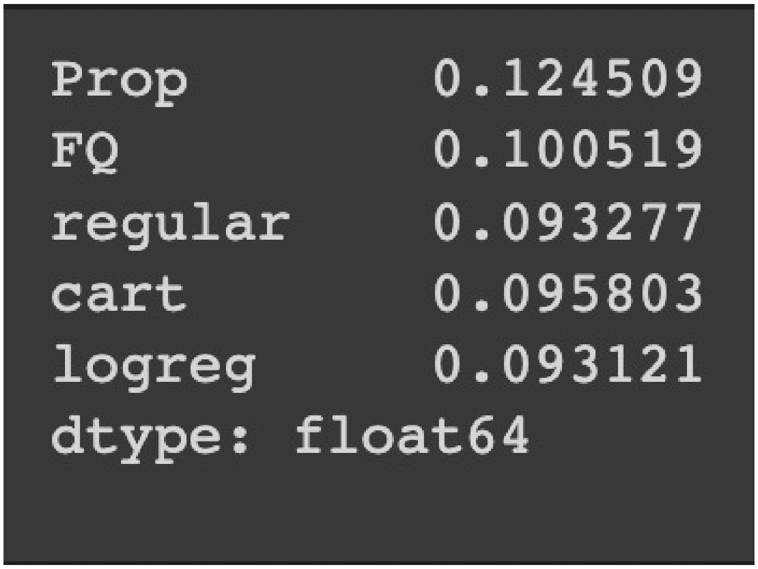
Let $corr(Z_1, Z_2) = \rho_1$, $corr(Z_2, Z_3) = \rho_2$, $corr(Z_3, Z_1) = \rho_3$. We sample data from Ƶ with fixed ρ_1, ρ_2 and ρ_3 and we can obtain samples from continuous covariates $x \sim N(0, 1)$ and binary covariates w and z.

We generate $N = 5000$ values as a super population. The parameter values used in simulation are $\beta_{xp} \in \{1.32\}$, $\beta_{zp} \in \{1.15, -1.80\}$. and $\beta_{wp} \in \{1.98, 1.52\}$.
The distributional assumptions for x are $x \sim Normal(80, 18)$, $\sigma^2_\epsilon \in \{1.2\}$.The sample sizes are $n = 1000$ and $n_a = n_b = n/2$. Covariates z and w are drawn such that both of these binary covariates and x are correlated of order 0.4, respectively. For all the simulation scenarios we focus on the prediction error of the fitted model in the population data (i.e. out of sample prediction), and compare this with the three imputation alternatives. We use $N - n$ observation to calculate the out of sample prediction error
For the presentation of our results, we use the following abbreviations: *FQ* stands for full questionnaire, where we hypothetically assume that the missing covariates z and w are fully observed in both samples, *Prop* indicates our proposed method, *Regular* indicates imputation using the regular mice setting, *Cart* indicates imputation for CART mice and *Logreg* indicates imputation for logistic regression.
To calculate a prediction error we use cross-validation. That is, we split the population into two parts: the sample of size n = 1000 is used to fit the regression models according to the proposed method and the remaining N-n values are used for prediction. Applying the proposed method and alternate routines, the mean and standard deviation of mean squared prediction error of 100 simulations are produced in the following figures.



```
Prop        5.902494
FQ          5.746590
regular     5.787818
cart        5.808621
logreg      5.826377
dtype: float64
```

Figure 1: Mean of the Mean Squared Prediction Error



```
Prop        0.124509
FQ          0.100519
regular     0.093277
cart        0.095803
logreg      0.093121
dtype: float64
```

Figure 2: Standard Deviation of the Mean Squared Prediction Error

Our proposed method gives provides good results compared to other computationally expensive imputation techniques.
The following graph shows prediction error for each methods by taking 100 samples.

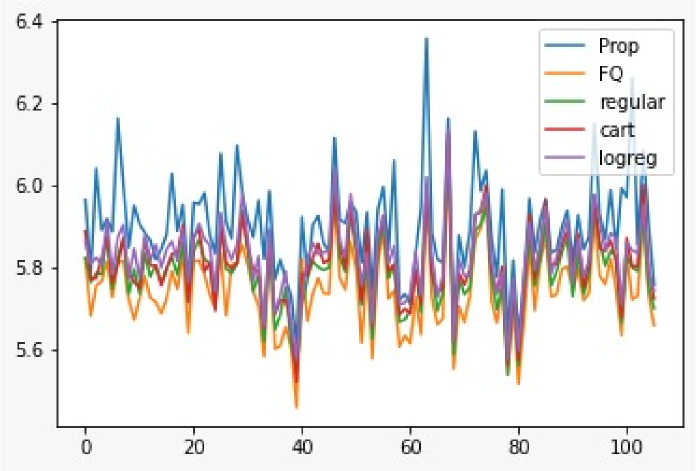


Figure 3: Prediction Error For Each Methods

The following graphs illustrates the probability density functions of n=100 prediction error of various imputation techniques used in the paper along with our proposed method.

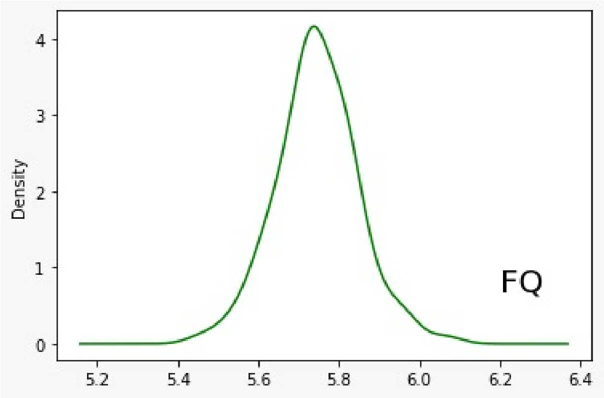


Figure 4: PDF of prediction error using full questionnaire

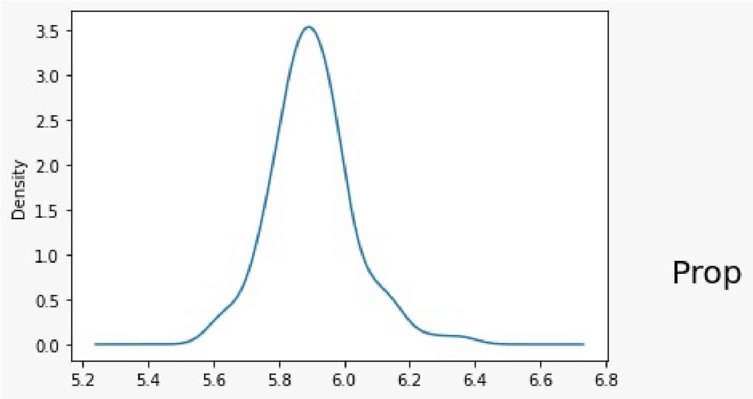


Figure 5: PDF of prediction error using proposed method

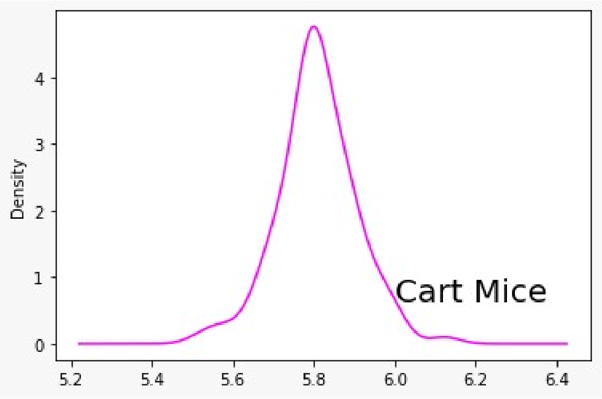


Figure 6: PDF of prediction error using MICE Cart

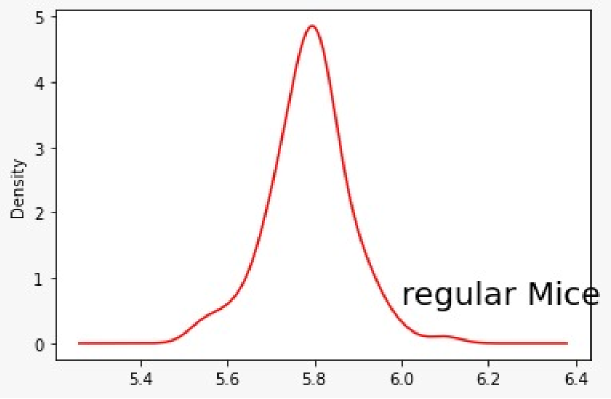


Figure 7: PDF of prediction error using regular MICE

5 CONCLUSION

In this paper, we propose a simple method to reduce the respondent's burden by splitting a long questionnaire and select two independent random samples in such a way that certain covariates are not jointly observed. We are interested in estimating the classical linear regression model Y given x, z, w to calculate the out of sample prediction error. This regression model fails when no single sampling unit has the information of specific covariates simultaneously, so we can not apply the complete case analysis on this data directly. To overcome this problem,we apply a CI assumption to factor the joint distribution into different sub factors. Through this factorization, we are able to estimate the classic linear regression model with the available splitted data without having to use any imputation procedures.

We showed in simulations that the proposed approach performs well with respect to other imputation techniques and it does not require a specific imputation model for missing data. Moreover our proposed method is easy to apply and intuitively more comprehensible.

In the paper the proposed method performed better than the other imputed techniques. One reason for that is maybe the way we simulated the data from the regression equation $Y = \beta_{0p} + x\beta_{xp} + z\beta_{zp} + w\beta_{wp} + \epsilon$ was different from the way the author simulated the data. (The author didn't explicitly mentioned the way he simulated the data hence we used the method told by Arnab Hazra sir.)

Even though the prediction error results weren't exactly the same as in the paper, the results we got are quite reasonable to conclude the same conjecture that the author wanted to convey (that the proposed method performs quite well in context to the problem statement we are working on.)

We only use the assumption of CI in our method, which is not testable . If the specific variables are closely related to the common variables then this assumption is reasonable.