Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 189–202

Adjusting for Auxiliary Information while Estimating Finite Population Parameters

Debapriya Sengupta
Indian Statistical Institute, Calcutta, India

SUMMARY

Suppose one wants to estimate some characteristics of an unknown finite population. One can look at this problem as a problem of estimating a functional $\theta(F)$, of the unknown population distribution function F. A simple estimator is given by $\theta(F_n)$ where F_n is the empirical distribution function. In presence of some auxiliary information on another population characteristic X it is possible to obtain an improved estimator of $\theta(F)$. Such estimators make use of the regression relationship between X and Y to achieve this goal and thus are known as regression estimators. The major application of this methodology can be found in survey sampling. In this article, this methodology is looked at from an angle which is not restricted only to finite populations. In this process the asymptotic distribution theory is derived under quite general conditions. The modalities of this improvement are also investigated. Also, a new class of regression estimators is introduced. One major advantage of the proposed methodology is that it can be applied in categorical and truncated problems without running into the risk of producing estimators outside the parameter space.

Key words : Statistical functionals, Auxiliary information, Regression estimators, Asymptotic distribution.

1. *Introduction*

Let $\{(X_i, Y_i), 1 \leq i \leq n\}$ be a set of n independent and identically distributed (i.i.d.) random observations. It is assumed that the marginal distribution of X_1 is known and is denoted by F_X. The marginal distribution of Y_1, say F, is unknown. We also assume that Y is a real-valued characteristic of some population. An extension for multivariate characteristics is straightforward. In the above framework it is desired to estimate a real valued functional of F, say $\theta(F)$.

Such problems arise in various situations in practice where one uses an auxiliary variable X to improve inference on the Y-variable. In survey literature, one specific example occurs in the context of regression estimation for finite population means. For a basic introduction to that methodology we refer to

Royall [10] and Cochran [3]. Isaki and Fuller [7] studied the properties of such estimators under linear regression relationship for various important sampling schemes. The survey literature mainly concentrates on the estimation of the population mean using this technique. The type of regression adjustment required while estimating functionals other than mean is relatively less studied. However, recently there has been some research in this direction. Kuk and Mak [8] considered the problem of estimating the median of a finite population after adjusting for auxiliary information. Also, in a recent paper Rao, Kovar and Mantel [9] developed a method of estimating the population distribution function of an unknown finite population using auxiliary information. Their method can be used to estimate any functional of the population distribution.

A simple estimator of $\theta(F)$ is given by :

$$\hat{\theta}_n = \theta(F_n) \tag{1.1}$$

where F_n is the empirical distribution function of $Y_1, Y_2, \ldots, Y_n$. It is naturally required that θ be defined on the space of all distribution functions.

The estimator defined through (1.1) does not utilize any possible regression relationship between the Y and X characteristics and also the fact that F_X is known. In the presence of a model for the conditional distribution of Y_1 given X_1 it is possible to incorporate the auxiliary information to construct an improved estimator. This is the basic philosophy behind regression estimation methodology. The term regression estimator is often used to denote estimators of the model parameters in usual regression models. However, we shall stick to the same terminology even in this context.

In this article we consider the problem of estimating the unknown population distribution (hence any functional) for an infinite population. The infinite population arises in finite population problems as the superpopulation. The idea was introduced in Royall [10]. In model based inference for finite populations the superpopulation plays the most crucial role. The sampling design becomes quite irrelevant in a model based analysis. There are controversies over such model based inference. One may refer to Hansen, Madow and Tepping [4] for a critical evaluation of model based and design based inference for finite population. The statistical methodology developed in this article can be used to draw model based inference on a population (finite or infinite). We do not consider any design based inference here.

Suppose the conditional distribution of the Y variable given X is available. Suppose that the conditional distributions are specified by a family of probability distribution functions

$$\{ F(y \mid x, \gamma) : x, y, \in \mathbb{R} \} \tag{1.2}$$

where $F(y \mid x, \gamma)$ stands for the conditional probability of the even $\{Y_1 \leq y\}$ given $\{X_1 = x\}$. Here γ is an unknown vector valued parameter taking values in an open subset Θ of $\mathbb{R}^p$. For notational simplicity we shall use $F(x, \gamma)$ to denote the distribution function in (1.2).

Let $\hat{\gamma}_n$ denote the estimator of γ obtained from the data. Using $\hat{\gamma}_n$

$$\tilde{F}_n(y) = \int F(y \mid x, \tilde{\gamma}_n)\, d\, Fx(x) \tag{1.3}$$

In view of (1.3) the regression adjusted estimator of $\theta(F_Y)$ is given by

$$\tilde{\theta}_n = \theta(\tilde{F}_n) \tag{1.4}$$

As it turns out the adjusted estimator defined through (1.4) is quite easy to compute. When Y and X are related through a linear relationship the sampling properties of $\tilde{\theta}_n$ are studied in great details. In Rao, Kovar and Mantel [9] a different regression adjusted estimator of the population distribution function is considered. They first predict the unobserved part of the population through a linear regression relationship and then consider the empirical distribution based on the observed and the predicted part. This method has one drawback when the population is categorical (like, Binomial) or truncated (like, Gamma). In such cases weighted linear predictors of the unobserved part can become negative or a fraction. The estimator defined through (1.4) does not face such problems.

In this paper the main results include the derivation of the asymptotic distribution of $\tilde{\theta}_n$. It is shown that $n^{1/2}(\tilde{\theta}_n - \theta)$ is asymptotically normal under some regularity conditions (provided, of course, that the parameter estimate $\hat{\gamma}_n$ is asymptotically normal). From the results obtained by Huber [5], it is known that $n^{1/2}(\hat{\theta}_n - \theta)$ is asymptotically normal if the functional θ is continuously differentiable in the space of distributions. We also compare the asymptotic variances. It turns out that if we start with an efficient estimate of γ, the asymptotic variance of $\tilde{\theta}_n$ is smaller than that of $\hat{\theta}_n$.

In Section 2, the main results are presented together with the technical assumptions required to prove them. It is tried to keep the assumptions at a level which would allow to apply the results to a broad class of examples on one hand and on the other hand the assumptions can be verified in a given situation without much difficulty. Various important examples are considered and the consequences of the main results are studied in Section 3. In Section 4, a new class of weighted regression estimators is introduced and their properties are studied. Some concluding remarks are made in Section 5 and the proofs of the main results are presented in the appendix.

2. *Main Results*

Let this section begins with basic notations and assumptions. The key idea of this article as described in the introduction is to put the covariate adjustment technique under a general framework and show some applications. Because of this reason, we shoot only for a set of assumptions under which the applications we have in mind hold true.

First, assume certain regularity conditions on the model relating the X and Y variables.

Assumption A :

1. $F(y \mid x, \gamma)$ has a density $f(y \mid x, \gamma)$ under a fixed measure μ for all γ and almost all x. Moreover, the support of $f(y \mid x, \gamma)$ is free from γ.
2. $\dot{f}(y \mid x, \gamma) := \frac{\partial}{\partial \gamma} f(y \mid x, \gamma)$ exists almost everywhere $(\mu \times F_X)$ in a neighborhood B of the true γ. Moreover,

$$\frac{\partial}{\partial \gamma} \int f(y \mid x, \gamma)\, dF_X(x) = \int \dot{f}(y \mid x, \gamma)\, dF_X(x)$$

almost everywhere μ.

3. The following condition on L_2 continuity of f hold.

$$\lim_{\delta \to 0} \sup_{|u| \le \delta} \bar{E}_\gamma f^{-2}(Y \mid X, \gamma) [\mid f(Y \mid X, \gamma + u) - f(Y) \mid (\gamma) \mid^2] = 0$$

The next assumption relates to certain smoothness property of the functional θ.

Assumption B :

For some function ψ_F with $E_F \psi_F^2 < \infty$ the following expansion holds:

$$\theta(w F_n + (1-w) F(\gamma + u)) = \theta(F) + w n^{-1/2} \sum_{j=1}^{n} \psi_F(Y_j)$$

$$+ (1-w) \int \psi_F(y)\, d(F(\gamma+u) - F)(y)$$

$$+ o_p(\max(\mid u \mid, n^{-1/2}))$$

where $F(\gamma + u)$ stands for the marginal of Y under $\gamma + u$ (thus, $F(\gamma) = F$).

Assumption C :

For some positive definite matrix Σ,

$$n^{1/2}(\hat{\gamma}_n - \gamma) \Rightarrow N(0, \Sigma)$$

The set of assumptions A are standard ones required for the validity of usual asymptotic theory in parametric estimation (viz., Ibragimov and Hasminskii [6]). The condition A3 requires certain L_2 continuity of the derivative of f which is fairly straightforward to verify in the cases we have in mind.

The condition B seems to be the hardest condition to verify. However such a condition can not really be avoided. In specific situations one has to verify such a condition using recipe specific to the problem. In case the functional is convex differentiable, the assumption B is automatically satisfied. Many statistical functionals of interest like median, are not differentiable in the sense of Gateaux or Frechet. For such situations the assumption B needs to be verified separately. For an excellent introduction to the theory of statistical functionals we refer to the classic book by Huber [5].

Nothing much needs to be said about the assumption C as it assumes that the auxiliary parameter γ is estimated with certain accuracy.

Theorem 2.1. Under assumption A, B and C the following assertions hold true.

(i) Consistency :

$$\tilde{\theta}_n \rightarrow \theta \text{ in probability.} \qquad (2.1)$$

(ii) Asymptotic Normality :

$$n^{1/2}(\tilde{\theta}_n - \theta) \Rightarrow N(0, \Gamma^t \Sigma \Gamma) \qquad (2.2)$$

where

$$\Gamma = f \psi_F(y) f(y \mid x, \gamma) \, d\mu(y) \, d F_X(x)$$

$$= \text{Cov}\{\psi_F(Y), S(Y \mid X, \gamma)\}$$

and $S = \frac{\partial}{\partial \gamma} \log f$ denotes the score function for the model.

The proof is relegated to the appendix. In the next result we compare the performances of $\tilde{\theta}$ and $\hat{\theta}$ in terms of their asymptotic variances. Since the asymptotic variance of $\hat{\theta}_n$ depends on the dispersion of the initial estimator, it turns out that it is not possible to improve upon the performance of $\hat{\theta}_n$ for any choice of the initial estimator of γ. If γ is estimated by an efficient estimator, we gain in terms of asymptotic variance by using the regression estimator.

Define

$$I = \int\int f^{-1}(y \mid x, \gamma)\,[\dot{f}\,\dot{f}^t]\,(y \mid x, \gamma)\,d\mu(y)\,dF_X(x)$$

$$\bar{S}(Y \mid \gamma) = \frac{\partial}{\partial\gamma} \log \int f(Y \mid x, \gamma)\,dF_X(x)$$

$$\bar{I} = C_{\bar{S}, \bar{S}} \tag{2.3}$$

where I is assumed to exist and be positive definite and for any two random vectors U and V

$$C_{UV} = \text{Cov}\{U, V\} = E\,U\,V^t$$

Theorem 2.2. Suppose there exists an estimator $\hat{\gamma}_{ne}$ of γ with asymptotic dispersion I^{-1}. Also, let $\underset{\sim}{\theta}_{ne}$ denote the corresponding regression estimator. Then

$$\text{Asymp. Var}\{\underset{\sim}{\theta}_{ne}\} = C^t_{\psi_F}\, I^{-1}\, C_{\psi_F, s} \le \text{Asymp. Var}\{\hat{\underset{\sim}{\theta}}_n\} \tag{2.4}$$

The inequality in (2.4) is strict whenever $(I - \bar{I})$ is positive definite.

Remark 1 : By Theorem 2.2 it is known that when one uses an efficient initial estimator of γ, asymptotically it is beneficial to use regression adjustments. Now the question is whether the efficient estimation of γ is the only way to gain in asymptotic efficiency. The answer to this question is quite interesting. How much inefficiency one can permit in estimating γ so that the regression adjusted estimator of θ is still more efficient than the simple estimator depends on the multiple correlation between the influence function of θ and the likelihood score function of the stipulated model. This fact gives us a criterion for deciding whether the regression adjustment for a particular functional under a given model for the conditional distribution of Y given X would be useful or not. The functional for which the range of improvement is quite stringent (i.e., when one needs to estimate the auxiliary parameter γ with high efficiency to get any improvement at all), a regression adjustment may not be advisable. Because in such cases the performance of the regression adjusted estimators for realistic sample sizes may not be that good. A Bootstrap comparison between the simple and the regression estimator may give us a better insight into the problem in such cases.

Remark 2 : Recall that $f(y \mid \gamma) = \int f(y \mid x, \gamma)\, dF_X(x)$. Consider the estimation of γ from the likelihood

$$L(\gamma) = \prod_{1}^{n} f(Y_i \mid \gamma)$$

Then an efficient estimate of γ derived from this likelihood will have an asymptotic dispersion I^{-1} which is greater than I^{-1} in a matrix sense. Thus, one can not achieve an asymptotic variance of $C^t_{\psi_F, s}\ I^{-1}\ C_{\psi_F, s}$. Therefore, this alternative method of making regression adjustment is not recommended in this regard.

Corollary 2.1. Under the set-up of the theorems, the regression adjusted estimator, $\tilde{\theta}_n$, has asymptotically greater efficiency than the simple estimator, $\hat{\theta}_n$, if

$$\lambda_{\max} (I^{1/2} \Sigma I^{1/2}) \leq \rho^{-2} (\psi_F, S) \tag{2.5}$$

where $\rho^2 (\psi_F, S) = \dfrac{C^t_{\psi_F, s}\ I^{-1}\ C_{\psi_F, s}}{E\, \psi_F^2}$ and $\lambda_{\max}$ denotes the maximum eigenvalue of a matrix.

Proof : By Theorem 2.1 we have

$$\text{Asymp. Var} \{ \tilde{\theta}_n \} = C^t_{\psi_F, S}\ \Sigma C_{\psi_F, S}$$

$$\leq \lambda_{\max} (I^{1/2} \Sigma I^{1/2})\ [C^t_{\psi_F, S}\ I^{-1}\ C_{\psi_F, S}]$$

The result follows from the above observation.

3. *Applications*

In this section, consider some applications of the theory developed so far. We begin with the classical example of linear regression.

Example 1 : Suppose $\{ (X_i, Y_i), 1 \leq i \leq n \}$ be a bivariate scatter where the following regression model is assumed to hold :

$$F(y \mid x) = N(a + bx, \sigma^2) \tag{3.1}$$

Therefore, in this example $\gamma = (a, b, \sigma^2)$ takes values in an open subset of $\mathbb{R}^3$. Also let us denote the mean and the variance of the X characteristic by m_X and v_X^2 respectively. For this model it is well known that the MLE of a, b and σ^2 are $\hat{a}_{nML} = \overline{Y}_n - \hat{b}_{nML}\overline{X}_n$, $\hat{b}_{nML} = S_{XY} / S_{XY}^2$ and $\hat{\sigma}^2 = n^{-1} \sum_{j=1}^{n} (Y_j - \hat{a}_{nML} - \hat{b}_{nML} X_j)^2$ respectively.

The regression adjusted estimators of the population mean and variance are thus given by :

$$
\begin{aligned}
\tilde{m}_n &= \hat{a}_{nML} + \hat{b}_{nML}\, m_X \\
&= \overline{Y}_n + \hat{b}_{nML} (m_X - \overline{X}_n) \\
\tilde{\sigma}_n^{\,2} &= \hat{\sigma}^2_{nML} + \hat{b}^2_{nML}\, v_X^2
\end{aligned}
\tag{3.2}
$$

Therefore, in the linear regression case we end up with the familiar regression estimators of the population mean and variance of the Y characteristic. The properties of the estimators in (3.2) are discussed in detail in the survey literature see, Isaki and Fuller [7] for example.

Example 2 : Here let $\{(X_i, Y_i), 1 \le i \le n\}$ be a sequence of i.i.d. observations where Y variable takes only two values say, 0 or 1. We assume a logistic regression relationship between X and Y, i.e.

$$\Pr\{ Y = 1 \mid X = x, \gamma \} = \frac{\exp x^t \gamma}{1 + \exp x^t \gamma} \tag{3.3}$$

Now let $\hat{\gamma}_n$ denote the MLE of γ so that we have

$$\text{Asymp. Var}\{\hat{\gamma}_n\} = \left[E \left\{ \frac{\exp x^t \gamma}{(1 + \exp x^t \gamma)^2} x x^t \right\} \right]^{-1} \tag{3.4}$$

The computation of the MLE for logistic regression is not difficult. Most of the standard statistical packages have the provision for the MLE in a logistic regression problem. At any rate, once we have $\hat{\gamma}_n$ the regression adjusted estimator of the population proportion of Y is

$$\hat{p}_n = \int \frac{\exp x^t \hat{\gamma}_n}{1 + \exp x^t \hat{\gamma}_n} \, dF_X(x) \tag{3.5}$$

and

$$\text{Asymp. Var}\{\hat{p}_n\} = \Gamma^t \left[E \left\{ \frac{\exp x^t \gamma}{(1 + \exp x^t \gamma)^2} x x^t \right\} \right]^{-1} \Gamma \tag{3.6}$$

where $\Gamma = E[(1 + \exp X^t \gamma)^{-2} \exp (X^t \gamma)\, X]$

The advantage of using (3.5) is that it always gives us an estimate which is between 0 and 1. The usual practice is to fit a linear relationship with possible heteroscdascity between Y and X. But the estimate we get that may be quite inferior to the MLE because a linear relationship between an indicator variable and a continuous random vector may be a total misfit. The same problem persists with the regression adjustment proposed by Rao, Kovar and Mantel [9].

Example 3 : In this example, consider various location functionals, i.e., $\theta(F)$ is defined as the solution of the equation :

$$\int \psi(y - \theta(F))\, dF(y) = 0$$

for suitable choices of the function ψ. The properties of such functionals are extensively studied in the literature in the context of maximum likelihood and robust estimation. See, Huber [5]. The influence function of $\theta(F)$ is given by:

$$IF(\psi, F, y) = [E\,\psi]^{-1}\, \psi(y - \theta(F))$$

For the derivation of the above fact, refer to Beran [1].

Let us assume that the conditional structure is specified by (1.2). If we estimate γ with asymptotic dispersion Σ, it follows that the asymptotic variance of the regression adjusted estimator $\theta(F)$ is given by :

$$\text{Asymp. Var}\{\tilde{\theta}_n\} = \text{Cov}\{IF, S^t\}\, \Sigma\, \text{Cov}\{IF, S\}$$

The above formula gives a way of calculating the asymptotic variances of regression adjusted estimators for a large class of robust location functionals.

4. *A Family of Alternative Estimators*

In the present framework one can construct the following family of regression adjusted estimators of the distribution function F.

$$\tilde{F}_{n,w}(t) = w\, F_n(t) + (1 - w)\, \tilde{F}_n(t) \tag{4.1}$$

where w is a weight, $0 \leq w \leq 1$.

So far two extreme cases have been discussed, namely, the cases $w = 0$ and $w = 1$. In this section we investigate the asymptotic behaviour of the weighted regression estimator

$$\underset{\sim}{\theta}_{n,w} = \theta(\tilde{F}_{n,w}) \tag{4.2}$$

If the asymptotic variance of $\tilde{\theta}_{n,w}$ is minimized at $w = w^*$ then it is advisable to use $\underset{\sim}{\theta}_{n,w^*}$ as the final version of the regression adjusted estimator.

However, there is one problem of using the optimally weighted regression adjusted estimator. The problem being that the optimal weight w^* will depend on the unknown parameters of the model. Therefore we have to use a suitable estimate of w^* in the estimator which means changing the sampling distribution of the final estimator. It is not very clear how it will affect its small sample distribution. One can construct examples where the effect of using the plugged-in estimator take away the advantage one gained by using the auxiliary information in the first place. However, if we could have a way of estimating γ for which the optimal weight does not depend on the unknown model parameters, we could use the plugged-in estimator without changing the sampling distribution of the optimal estimator. Surprisingly, what turns out is that the efficient estimation of the model parameter γ has the two-fold advantage. Firstly by Theorem 2.2, it has smaller asymptotic variance than the unadjusted version. Secondly, the optimal choice of w* is also free from any model parameter and, actually it is zero. The next theorem formalizes the above ideas.

Theorem 4.1. Assume A, B and C. Further suppose that $\hat{\gamma}_n$ admits the following expansion

$$n^{1/2}(\hat{\gamma}_n - \gamma) = n^{1/2} \sum_{j=1}^{n} \Lambda(Y_j, X_j, \gamma) + o_p(1) \tag{4.3}$$

Then

$$\text{(i)} \qquad n^{1/2}(\underset{\sim}{\theta}_{n,w} - \theta) \Rightarrow N(0, r^2(w)) \tag{4.4}$$

where $r^2(w) = w^2 C_{\psi\psi} + (1-w)^2 C^t_{\psi,S} C_{\Lambda,\Lambda} C^T_{\psi,S} + 2w(1-w) C^t_{\psi,S} C_{\psi,\Lambda}$

(ii) For a fixed Λ, the asymptotic variance of $\{\underset{\sim}{\theta}_{n,w}\}$ is minimized by

$$w^*(\Lambda) = \frac{C^t_{\psi,S} C_{\psi,\Lambda} - C^t_{\psi,S} C_{\Lambda,\Lambda} C_{\psi,S}}{C_{\psi\psi} - 2C^t_{\psi,S} C_{\psi,\Lambda} + C^t_{\psi,S} C_{\Lambda,\Lambda} C_{\psi,S}} \tag{4.5}$$

(iii) The choice $\Lambda^* = -I^{-1} S$ (i.e., the maximum likelihood score) minimizes $r^2(w)$ over all choices of scores. In that case $w^*(\Lambda^*) = 0$ and r^2_m in $= C^t_{\psi,S} I^{-1} C_{\psi,S}$.

The proof of the above result is deferred to the Appendix.

5. *Concluding Remarks*

In this section consider a few issues which are not taken care of in this article. The first one being the variance estimation problem. From the expressions obtained for asymptotic variances it is clear that one can simply use a plugged-in estimator of the quantity. However, one can suggest other methods of estimation. Among other methods the one which comes to the mind first is the Bootstrap estimation of the asymptotic variance. This method usually leads to better confidence interval estimates. Jackknifing is another possibility.

In this article the simple random sampling situation is essentially covered. It will be interesting to extend the scope of this result for other schemes in finite population sampling, like two stage stratified sampling. Another important idea which emerges from this article is that the effectiveness of regression adjustment for non-linear functionals depends on the type of the functional we are interested in and the type of the conditional structure. Therefore it is important to have a general idea about the effectiveness of this methodology in estimating certain practically useful functionals for various conditional structures which are used in practice. This is important because due to the estimation of model parameters it is not very clear whether the regression adjustment performs better than the simple estimation technique for relatively smaller sample sizes. This can be determined only through the data.

The next comment is regarding the testing of hypotheses about the functional θ. One can develop Wald-type asymptotic test using the regression adjusted estimator. It should have larger local asymptotic power than the analogous test based on the simple estimator. However, the power comparison for smaller sample sizes becomes quite intractable.

Finally, discuss the issue of robustness in this context. The dependence on the model is quite evident from the construction of the regression estimators that are proposed in the article. Thus the estimators may not be robust against certain departures from the model. No good method of robustifying the estimator $\tilde{F}$ is known till this date. Therefore attempts should be made for a solution of the problem of robustness in this context.

6. *Appendix*

The proofs of the results stated before are presented here. First we state a Lemma. Let us define for fixed $u \in \mathbb{R}^p$ the following

$$R(u) = \int \psi_F(y)\,[\,f(y \mid \gamma + u) - f(y)\,]\, d\mu(y)$$

where $f(y \mid \gamma + u) = \int f(y \mid x, \gamma + u)\, dF_X(x)$ and $f(y) = f(y \mid \gamma)$ represents the true density of the Y variable.

Lemma 6.1 : Under the assumptions A, B and C

$$R(u) = u^t C_{\psi_F, S} + O(|u|) \text{ as } |u| \to 0 \tag{6.1}$$

Proof : We can write R (u) as

$$R(u) = \int \psi_F(y) \left[\int_0^1 f(y \mid \gamma + \alpha u)\, d\alpha \right] d\mu(y) \tag{6.2}$$

$$= \int \psi_F(y) \int_0^1 \left[\int u^t f(y \mid x, \gamma + \alpha u)\, dF_X(x) \right] d\alpha\, d\mu(y) \tag{6.3}$$

This follows from Assumption A. Therefore, we have

$$R(u) = u^t C_{\psi, S} + R^*(u)$$

By the Mean Value Theorem

$$R^*(u) = \int \psi_F(y) \left[\int_0^1 [f(y \mid x, \gamma + \alpha u) - f(y \mid x, \gamma)]\, dF_X(x) \right] d\alpha\, d\mu(y)$$

Now for any t,

$$\int\int |\psi_F(y)|\, |u^t [f(y \mid x, \gamma + t) - f(y \mid x, \gamma)]|\, dy\, dF_X(x)$$

$$\leq |u|\, (E\psi_F^2)^{1/2}\, E\left[\frac{|f(Y \mid X, \gamma + t) - f(Y \mid X, \gamma)|^2}{f^2(Y \mid X, \gamma)} \right]$$

Therefore, Fubini's Theorem applies. Hence, after a change of the order of integration we obtain

$$|R^*(u)| \leq \int_0^1 (E\psi_F^2)^{1/2} |u|\, E(E\psi_F^2)^{1/2}\, E\left[\frac{|f(Y \mid X, \gamma + t) - f(Y \mid X, \gamma)|^2}{f^2(Y \mid X, \gamma)} \right] d\alpha \tag{6.4}$$

Next by the assumption B we have

$$\lim_{\delta \to 0} \sup_{|t| \leq \delta} E(E\psi_F^2)^{1/2}\, E\left[\frac{|f(Y \mid X, \gamma + t) - f(Y \mid X, \gamma)|^2}{f^2(Y \mid X, \gamma)} \right] = 0$$

Thus, an application of the Dominated Convergence Theorem on the right hand side of (6.3) gives us

$$|R^*(u)| \doteq o_p(|u|)$$

Hence the Lemma follows.

Proof of Theorem 2.1. By Lemma 6.1 and the assumption B we get

$$n^{1/2}(\tilde{\theta}_n - \theta) = n^{1/2}(\hat{\gamma}_n - \gamma)^t C_{\psi_F, S} + o_p(n^{1/2} | \hat{\gamma}_n - \gamma |)$$

Now the Theorem follows from assumption C.

Proof of Theorem 2.2 : For any vector $b \neq 0$

$$(b^t \Gamma)^2 = [E(b^t S) \psi_F]^2$$

$$\leq (E\psi_F^2) \left[\int \frac{(b^t f\, dF_X)^2}{f(y)} d\mu(y) \right] \qquad (6.5)$$

The above inequality is a consequence of Cauchy-Schwartz inequality. Also note that the equality holds in (6.4) only when ψ_F and $b^t S$ are linearly dependent. In that case (since ψ_F is only a function of Y) we must have $E\,\mathrm{Var}\{ b^t S \mid Y \} = 0$ which amounts to saying that $b^t(I - \bar{I})b = 0$ which can not hold if $(I - \bar{I})$ is positive definite.

Now let us define the following density

$$g_y(x) = \frac{f(y \mid x, \gamma)}{\int f(y \mid x, \gamma)\, dF_X(x)}$$

almost everywhere dF_X and for every fixed y. This is actually the conditional density of X given Y. Then we can write

$$\int \frac{(b^t f\, dF_X)^2}{f(y)} d\mu(y) = \int \left[\int (b^t S)\, g(x)\, dF_X(x) \right]^2 f(y)\, d\mu(y) \qquad (6.6)$$

Thus again by Cauchy-Schwartz we get

$$\int \left[\int b^t S\, g(x)\, dF_X(x) \right]^2 f(y)\, d\mu(y) \leq \int \int \frac{(b^t f)^2}{f} dF_X\, d\mu$$

$$= b^t I b \qquad (6.7)$$

In (6.6) an equality holds if the score function, S, is degenerate which can not hold under the assumption that I is non-singular. Hence

$$\frac{(b^t \Gamma)^2}{b^t I b} \leq E\psi_F^2$$

Therefore, $$\Gamma^t I^{-1} \Gamma = \sup_{b \neq 0} \frac{(b^t \Gamma)^2}{b^t I b} \leq E \psi_F^2$$

Hence the Theorem.

Now proceed to prove the next Theorem.

Proof of Theorem 4.1 : Notice first that by virtue of the assumption C and Lemma 6.1 the following asymptotic expansion holds :

$$n^{1/2} (\underset{\sim}{\theta}_{n,w} - \theta) = w [n^{1/2} \Sigma_1^n \psi_F (Y_j)] + (1 - w) C_{\psi_F, S} [n^{-1/2} \Sigma_1^n \Lambda (Y_j, X_{j,\gamma})] + o_p(1) \quad (6.8)$$

Once the expansion (6.7) is obtained the part (i) of Theorem 4.1 follows by the Central Limit Theorem. The remaining parts also follow from routine differentiation.

REFERENCES

[1] Bern, R., 1977. Robust location estimates. *Ann. Statist.*, **5**, 431-444.

[2] Chambers, R.L. and Dunstan, R., 1986. Estimating distribution functions from survey data. *Biometrika*, **73**, 597-604.

[3] Cochran, W.G., 1977. *Sampling Techniques*, 2nd ed. John Wiley and Sons, New York.

[4] Hansen, M.H., Madow, W.G. and Tepping, B.J., 1983. An evaluation of model dependent and probability-sampling inference in sample surveys, *J. Amer. Statist. Assoc.*, **78**, 776-793.

[5] Huber, P.J., 1981. *Robust Statistics*. John Wiley and Sons, New York.

[6] Ibragimov, I.A. and Hasminskii, R.Z., 1981. *Statistical Estimation : Asymptotic Theory (Transl. From Russian by S. Kotz)*. Springer-Verlag, New York.

[7] Isaki, C.T. and Fuller, W.A., 1982. Survey design under the regression super-population model. *J. Amer. Statist. Assoc.*, **77**, 89-96.

[8] Kuk, A.Y.C. and Mak, T.K., 1989. Median estimation in the presence of auxiliary information. *J. Roy. Statist. Soc.*, **B51**, 261-269.

[9] Rao, J.N.K., Kovar, J.G. and Mantel, H.J., 1990. On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, **73**, 365-375.

[10] Royall, R.M., 1970. On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-387.