



In Praise of the Listwise-Deletion Method (Perhaps with Reweighting)

Phillip S. Kott

RTI International

NISS Workshop on the
Analysis of Complex Survey Data
With Missing Item Values

October 17, 2014

Introduction

- Listwise deletion (AKA “complete case analysis”) is the simplest method for handling missing item data in a regression analysis.
- It had been called:
“among the worst methods available for practical applications”
(Wilkinson and the Task Force on Statistical Inference, Board of Scientific Affairs, American Psychological Association).
- Yet it is often more robust than its more sophisticated rivals.
- And it can seamlessly integrate sampling weights into the analysis.

Introduction

So long as:

The outcome model being estimated holds in the population;

and

The probability that a record is deleted is *not* a function of the dependent variable in the model;

using listwise deletion returns (asymptotically) unbiased parameter estimates.

Both of these assumptions can often be tested.

When one of those tests fails, then the post-deletion data set can be reweighted in an attempt to restore unbiasedness.

Introduction

- Unlike its competitors, listwise deletion does *not* assume any outcome model other than the (conditional) one being estimated – and sometimes not even that.
- Listwise deletion *does* assume a vague propensity model for the records deleted (or, equivalently, not deleted) from the data set.
- When reweighting is deemed necessary (or to test whether it is), a particular propensity is assumed.

Outline

Long Detour: Regression Analysis with Complex Survey Data and No Missing Items Values

The Ideal Conditions for Listwise Deletion

Fitting a Selection Model When Some Variables are Never Missing

Testing Whether Listwise Deletion is Appropriate

A Numerical Example

Some Remarks

A Regression Model

We will be concerned with estimating the parameters of a regression model of the form:

$$y_k = f(\mathbf{x}_k^T \boldsymbol{\beta}) + \varepsilon_k,$$

where f can be linear or logistic (i.e., $[1 + \exp(-\mathbf{x}_k^T \boldsymbol{\beta})]^{-1}$),

and either

$E(\varepsilon_k | \mathbf{x}_k) = 0$ – the standard (conditional) model

or

$E(\varepsilon_k \mathbf{x}_k) = \mathbf{0}$ – the extended model

The Estimating Equation

Given an independent sample of observations, a consistent estimator for $\boldsymbol{\beta}$ in the *extended* model can be found by solving for \mathbf{b} in the estimating equation:

$$\sum \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) \mathbf{x}_k = \mathbf{0}$$

under mild conditions (since $\frac{1}{N} \sum \left(y_k - f(\mathbf{x}_k^T \boldsymbol{\beta}) \right) \mathbf{x}_k \rightarrow \mathbf{0}$).

The Estimating Equation

Under mild conditions, the solution to

$$\sum \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) \mathbf{x}_k h(\mathbf{x}_k) = \mathbf{0}$$

given virtually any scalar function of $h(\cdot)$ will be a consistent estimator $\boldsymbol{\beta}$ in the *standard* model.

The standard model is what we usual think of as a regression model.
It can fail.

The extended model almost never fails.

Introducing a Complex Sample Design

Suppose our observations come from a complex random sample S drawn from a population U .

Let I_k be a 0/1 indicator of whether $k \in S$,

$\pi_k = E(I_k)$ be the selection probability of k , and

$d_k = I_k/\pi_k$ be the sampling weight of k .

So that $E(d_k) = 1$.

The Estimating Equation With a Complex Sample

Under the extended model, a consistent \mathbf{b} must solve:

$$\sum_U \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) \frac{I_k}{\pi_k} \mathbf{x}_k = \sum_U \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) d_k \mathbf{x}_k = \mathbf{0}$$

But under the standard model, it need only solve:

$$\sum_U \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) d_k \mathbf{x}_k h(\mathbf{x}_k) = \mathbf{0}$$

for any $h(\cdot)$.

Analysis Weights

Under the standard model, we can treat the

$w_k = d_k h(\mathbf{x}_k)$ as analysis weights.

Under the linear model with uncorrelated and homoscedastic ε_k , Pfefferman and Sverchkov suggested replacing $h(\mathbf{x}_k)$ with $1/d(\mathbf{x}_k)$, where $d(\mathbf{x}_k)$ is a prediction of d_k based on the components of \mathbf{x}_k .

The bigger point is that the choice for $h(\mathbf{x}_k)$ in the definition of an analysis weight under the standard model is up to the user.

Variance Estimation (for both)

Under mild conditions:

$$\begin{aligned} \mathbf{b} - \boldsymbol{\beta} &\approx - \left(\sum w_k f'(\mathbf{x}_k^T \boldsymbol{\beta}) \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum w_k \mathbf{x}_k (y_k - f(\mathbf{x}_k^T \boldsymbol{\beta})) \\ &= - \mathbf{G}^{-1} \sum w_k \mathbf{x}_k \varepsilon_k \\ &\quad (w_k = d_k \text{ for the extended model}) \end{aligned}$$

implies $\mathbf{Var}(\mathbf{b}) \approx \mathbf{G}^{-1} \mathbf{Var}(\sum w_k \mathbf{x}_k \varepsilon_k) \mathbf{G}^{-1}$,

and nearly unbiased estimation using the stratified “sandwich” estimator given a multistage stratified sample (or a replication method that mimic this estimator).

What About Strata?

The sandwich variance formulation captures any correlation among the ε_k within primary sampling units.

Stratifying the sample can reduce the variance if the $E(\varepsilon_k | \text{stratum})$ vary.

Graubard and Korn noted, however, that when $E(\varepsilon_k | \text{stratum})$ vary the variance estimator can miss a contribution to the variance caused by the realized stratum sizes *in the population* being random.

Removing the stratification from variance estimation can be a costless way to increase the efficiency. At its worst, it will overestimate.

(By the way, “design-based” degrees of freedom calculations can be unreliable.)

Listwise Deletion (Finally)

Let D be the subset of S containing only complete records.

Let R_k be a 0/1 indicator of whether $k \in D$, and

$\rho_k = E(R_k)$ be probability that k is not deleted (is “selected” for D)

A consistent estimator for $\boldsymbol{\beta}$ under the standard model satisfies:

$$\sum_U \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) d_k \left(\frac{R_k}{\rho_k} \right) \mathbf{x}_k h(\mathbf{x}_k) = \mathbf{0}.$$

Under the extended model, $h(.) = 1$.

Listwise Deletion for the Standard Model

Unfortunately, ρ_k is unknown.

Fortunately, if ρ_k has the form $\rho_0(\mathbf{x}_k)$ for any scalar $\rho_0(\cdot)$,
and we can set $h(\mathbf{x}_k) = 1/\rho_0(\mathbf{x}_k)$ under than standard model.

The estimating equation:

$$\sum_U \left(y_k - f(\mathbf{x}_k^T \mathbf{b}) \right) R_k d_k \mathbf{x}_k = \mathbf{0}$$

results. This means that listwise deletion will produce consistent estimates for the components of $\boldsymbol{\beta}$ (under mild conditions).

What Just Happened?

Suppose we are fitting the *standard* linear model

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$$

with a complex survey, and some records have missing x_{1k} values.

If missingness is a function of x_{1k} and x_{2k} but not y_k , then listwise deletion produces consistent estimates for the β 's.

On the one hand, a missing x_{1k} need not be missing at random; it can depend on its own value.

On the other, a missing x_{1k} cannot be missing at random if that means its missingness is related to its y_k value.

What Just Happened?

Suppose we are fitting the *standard* linear model

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$$

with a complex survey, and some records have missing y_k values.

If missingness is a function of x_{1k} and x_{2k} , then listwise deletion produces consistent estimates for the β 's.

This is what we often assume with complex survey data in practice when estimating means and totals and sampling weights are adjusted for *unit* nonresponse.

Estimating a mean can be viewed as fitting simple regression. 

The Extended Model

Listwise deletion produces consistent estimates under the extended model when $E(\varepsilon_k|\mathbf{x}_k) \neq 0$ only if all ρ_k are equal (i.e., missingness is completely at random).

Otherwise, we may be able to reweight D by assuming a functional form for $\rho_k = \rho_0(\mathbf{h}_k)$ for some vector of characteristics \mathbf{h}_k .

Most often, none of the components of \mathbf{h}_k will be missing, and a logistic model for the records in S is:

$$\rho_0(\mathbf{h}_k) = \rho(\mathbf{h}_k^T \boldsymbol{\gamma}) = [1 + \exp(-\mathbf{h}_k^T \boldsymbol{\gamma})]^{-1},$$

with $\boldsymbol{\gamma}$ is estimated from the data.

Estimating γ

There are (at least) three ways to compute consistent estimates for γ .

Logistic regression solves for \mathbf{g} in the estimating equation:

$$\sum_S \left(R_k - \rho(\mathbf{h}_k^T \mathbf{g}) \right) \mathbf{h}_k = \mathbf{0}$$

Weighted logistic regression solves for \mathbf{g} in the estimating equation:

$$\sum_S \left(R_k - \rho(\mathbf{h}_k^T \mathbf{g}) \right) d_k \mathbf{h}_k = \mathbf{0}$$

Estimating γ

Calibration weighting (available in SUDAAN and the ‘sampling’ package in R) solves

$$\sum_S \left(\frac{R_k}{\rho(\mathbf{h}_k^T \mathbf{g})} - 1 \right) d_k \mathbf{z}_k = \mathbf{0},$$

where \mathbf{z}_k has the same dimension as \mathbf{h}_k – and the components of \mathbf{z}_k , *but not* \mathbf{h}_k , cannot have missing values.

As with logistic regression, one can also test whether a component of \mathbf{g} is significantly different from 0.

Reweighting

Effectively, the analysis weight for the extended model is

$$\omega_k = d_k R_k [1 + \exp(-\mathbf{h}_k^T \mathbf{g})],$$

where \mathbf{g} is fit using one of the three methods.

Using these analysis weights produces an unbiased estimate for $\boldsymbol{\beta}$ *if* the selection model involved in creating D (i.e., $\rho(\cdot)$) is correctly specified up to $\boldsymbol{\gamma}$.

Extra terms may need to be added to the stratified sandwich estimator of variance however because \mathbf{g} is not $\boldsymbol{\gamma}$.

Replication can be used instead (assuming $\rho(\mathbf{h}_k^T \boldsymbol{\gamma})$ is uncorrelated across PSUs).

Reweighting

If the *standard model* holds (i.e., $E(y_k - \mathbf{x}_k^T \boldsymbol{\beta} | \mathbf{x}_k) = 0$),

then one may still need to reweight w_k

(i.e., use an analysis weight like $\omega_k = w_k R_k [1 + \exp(-\mathbf{h}_k^T \mathbf{g})]$)

when a component of \mathbf{h}_k is not a function of the components of \mathbf{x}_k .

One can conceivably assess whether listwise deletion is consistent with the data by testing whether the component of \mathbf{b} computed using the original sampling weights are significantly different from the components computed using the reweighted analysis weights with either

a replication method or

the stratified sandwich estimator (treating the two versions of the same observation as if they came from the same PSU).

Are the Sampling Weights Ignorable?

Recall $d_k = I_k / \pi_k$.

In practice the probability of selection, π_k , may include an estimate of the probability of unit response.

In any event, if π_k is effectively a function on of the components of \mathbf{x}_k , then the sampling weights can be ignored when the standard model holds.

That can be tested analogously to how we test whether listwise deletion was consistent with the data.

An Example

A public use version of the National Survey on Drug Use and Health contains 2,651 youths ages 12 -17 who had mental-health counseling.

We fit a logistic model for Y – Did counseling help (1 = Yes/0 = No)? with these covariates:

HISPB = 1 when Black or Hispanic/0

GOODGR = 1 when an A or B student/0 otherwise

ENC = 1 when strong parental encouragement/0 otherwise

MEDS = 1 when prescribed meds for mood disorder/0 otherwise

ENCMED = ENC² × MEDS

DEL = 1 when more than two delinquent behaviors in past year/0 otherwise

NONE = 1 when no mental-health visits in past year/0 otherwise

SMHVST – a monotonic function of mental-health visits in past year

An Example

Other variables (sex, parent's income, attendance at religious services, etc.) were not significant at the .1 level.

Y and HISPB were never missing nor was a variable for the youth's age.

In all, nearly 14% of the records had some missing values.

An Example

The probability of being selected for D is clearly a function of HISPB and YOUNG (1 when age 12 or 13/0 otherwise).

The t value for the coefficient of Y is around 1.

We will add it to the selection model anyway.

We can use WTADJX in SUDAAN to test if any of the covariates with missing values are predictors of a record being deleted from D . None appear to be.

Some Results

Using Analysis Weights (starting with d_k then applying calibration weighting) & *Stratified Sandwich Variance Estimates*
(60 variance strata with two variance PSU's per stratum)

Variable	Estimate	Standard Error	<i>p</i> value
Intercept	-----	0.294	≈ 0.1
HISBP	-----	0.203	< 0.1
GOODGR	-----	0.152	< 0.001
ENC	-----	0.196	≈ 0.5
MEDS	-----	0.243	≈ 0.5
ENCMED	-----	0.311	< 0.01
DEL	-----	0.164	< 0.001
NONE	-----	0.274	< 0.1
SMHVST	-----	0.079	< 0.01

(Ignoring the strata gave virtually the same results.)

Some Results

*Differences Using Sampling Weights
With Stratified Sandwich Variance Estimates*

Variable	Estimate	Standard Error	<i>t</i> value	<i>p</i> value*
Intercept	-0.056	0.019	-3.03	0.004
HISPB	-0.021	0.012	-1.75	0.084
GOODGR	-0.011	0.011	-0.96	0.340
ENC	0.019	0.013	1.47	0.146
MEDS	0.006	0.011	0.57	0.573
ENCMED	-0.004	0.021	-0.22	0.828
DEL	0.010	0.011	0.87	0.388
NONE	0.025	0.017	1.48	0.144
SMHVST	0.007	0.004	1.80	0.076

* Multiply by 9 to get a Bonferroni-adjusted *p* value.

Some Results

Both the adjusted F and Satterthwaite F also show very significant differences ($p < .001$) between using the two sets of weights.

The results using a jackknife are very different; for example, the difference with largest $|t|$ -value is

Variable	Estimate	Standard Error	t value	p value
Intercept-SS	-0.056	0.019	-3.03	0.004
Intercept-JK	-0.056	0.028	-1.98	0.052

Surprisingly, the jackknife standard errors for the coefficients using the analysis weights are virtually identical to the stratified-sandwich standard errors.

The error of \mathbf{b}

$$\mathbf{b} - \boldsymbol{\beta} \approx - \left(\sum w_k \left[\frac{R_k}{\rho(\mathbf{h}_k^T \boldsymbol{\gamma})} \right] f'(\mathbf{x}_k^T \boldsymbol{\beta}) \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum w_k \left[\frac{R_k}{\rho(\mathbf{h}_k^T \boldsymbol{\gamma})} \right] \mathbf{x}_k \boldsymbol{\varepsilon}_k$$

$$+ \left(\sum w_k \left[\frac{R_k}{\rho(\mathbf{h}_k^T \boldsymbol{\gamma})} \right] f'(\mathbf{x}_k^T \boldsymbol{\beta}) \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum w_k \left[\frac{R_k}{\rho(\mathbf{h}_k^T \boldsymbol{\gamma})} \right] \frac{\rho'(\mathbf{h}_k^T \boldsymbol{\gamma})}{\rho(\mathbf{h}_k^T \boldsymbol{\gamma})} \mathbf{h}_k^T (\mathbf{g} - \boldsymbol{\gamma}) \mathbf{x}_k \boldsymbol{\varepsilon}_k$$

Observe that under the standard model, the second term is of the same asymptotic order as the first unless the components of \mathbf{h}_k are functions of the components of \mathbf{x}_k .

The jackknife captures the impact of random fluctuations in a component of \mathbf{g} that may actually be estimating 0.

Other Results

When YOUNG is added as a covariate to the model for whether counseling helps, it is not significant no matter whether the original sampling weights or the reweighted analysis weights are used (F value less than 1).

Nevertheless, the coefficient on YOUNG in the fitted model changes significantly (p value less than .001) depending on which weights are used according to the stratified sandwich variance estimator – but not the jackknife.

When Y is removed from the selection model that no longer happens (p value greater than .5 even with the stratified sandwich variance estimator).

Weights Matter

In determining whether the weights are ignorable by comparing estimates for β computed with the original sample weights and without weights:

Both the adjusted F and Satterthwaite F show significant differences ($p \approx .01$) using the stratified sandwich variance estimator.

No individual t value is significant at the .1 level after a Bonferroni adjustment.

Still, three (of 9) are significant at the unadjusted .05 level.

The jackknife results are similar.

Some Remarks

Yes. Some efficiency may be lost with listwise deletion because data is discarded (and that information may be recoverable with an appropriate model for the \mathbf{x}_k).

Reweighting (and testing whether reweighting can remove bias) requires some important variables be complete ...

Or nearly complete and their missingness not a function of the dependent variable.

One should walk humbly with complex survey data.

Theoretical results are asymptotic, but sample sizes are finite. And models rarely hold exactly.

Sensitivity analyses are highly recommended.

A Reading List

Analysis of Complex Survey Data

Fuller (*Sankhya*, 1975; *Survey Methodology*, 1984)

Binder (*International Statistical Review*, 1983)

Skinner (*Analysis of Complex Surveys*, 1989)

Korn and Graubard (*American Statistician*, 1995)

Graubard and Korn (*Statistical Science*, 2002)

Kott (*Survey Research Methods*, 2007)

Pfeffermann and Sverchkov (*Handbook of Statistics*, 29B, 2009)

Scott and Wild (*Handbook of Statistics*, 29B, 2009)

Landsman and Graubard (*Statistics in Medicine*, 2012)

Missing Data

Little (JASA, 1992)