# Predictive Analytics with Strategically Missing Data

Juheng Zhang
Department of Operations and Information Systems
University of Massachusetts Lowell
Email: juheng_zhang@uml.edu

Xiaoping Liu
D'Amore-McKim School of Business
Northeastern University
Email: xia.liu@northeastern.edu

Xiao-Bai Li (corresponding author)
Department of Operations and Information Systems
University of Massachusetts Lowell
Email: xiaobai_li@uml.edu

## Appendix: Proofs of Theorems and Corollaries

In this online supplementary appendix, for convenience we use $\mathbf{w}_r \geq \mathbf{0}$ to refer to the situation where the coefficients of all *optional* attributes for the *r*th reduced model are all nonnegative; i.e., $\mathbf{w}_r \geq \mathbf{0}$ is equivalent to $w_{rj} \geq 0, j = M - m + 1, \ldots, M$. Also, recall that $W^+$ is the set of indices of the reduced models whose coefficients of the optional attributes are nonnegative; that is, $W^+ = \{r | \mathbf{w}_r \geq \mathbf{0}\}$.

**Theorem 1**. *With nonnegative* **x**, *the optimal objective value of problem* (7) *is*

$$y_a^* = \max\{K_r | r \in W^+\}, \text{ where } K_r = -\varepsilon + \left(\sum_{j=1}^{M-m} w_{rj} x_{aj} + b_r\right). \quad \text{(A.1)}$$

**Proof**: It follows from constraints in (7b) that $y_a \geq K_r + \sum_{j=M-m+1}^{M} w_{rj} x_{aj}, \forall r$. If $w_{rj} < 0$, then $w_{rj} x_{aj}$ can be arbitrarily small by allowing an arbitrarily large positive value for $x_{aj}$. Since $K_r$ is a fixed value, $K_r + \sum_{j=M-m+1}^{M} w_{rj} x_{aj}$ can be arbitrarily small too. Thus, $y_a$ cannot be bounded by a constraint with $w_{rj} < 0$; that is, the optimal value $y_a^*$ can only be found from the set of constraints specified in $W^+$.

Next, we use proof by contradiction to establish (A.1). Suppose $\exists y'_a$: $y'_a < y^*_a$ and $y'_a$ satisfies (7b); i.e.,

$$y'_a \geq K_r + \sum_{j=M-m+1}^{M} w_{rj}x_{aj}, \quad \forall\, r \in \{0, 1, \dots, R\} \tag{A.2}$$

It follows from (A.1) and $y'_a < y^*_a$ that $y'_a < \max\{K_r | r \in W^+\}$. For the $r$ with $\max\{K_r | r \in W^+\}$, since $\mathbf{w}_r \geq \mathbf{0}$ and $\mathbf{x} \geq \mathbf{0}$, we have $\sum_{j=M-m+1}^{M} w_{rj}x_{aj} \geq 0$. Therefore,

$$y'_a < \max\{K_r | r \in W^+\} + \sum_{j=M-m+1}^{M} w_{rj}x_{aj}. \tag{A.3}$$

Equation (A.3) contradicts (A.2). QED

**Corollary 1**. *If every model except the fully reduced model, i.e., the Rth model, in* (7b) *has some negative coefficients on the optional attributes, then* $y^*_a = K_R$.

**Proof**: For $\forall r = 0, \dots, R-1, w_{rj} < 0$ for some $j \in \{M-m+1, \dots, M\}$. Then $\{r | \mathbf{w}_r \geq \mathbf{0}\} = \{R\}$, since $\mathbf{w}_R = \mathbf{0}$, i.e., the *R*th reduced model does not have any optional attribute. So, $\max\{K_r | r \in W^+\} = \max\{K_r | r = R\} = K_R$; that is, $y^*_a = K_R$. QED

**Corollary 2**. *For nonnegative* **x**, *the optimal solutions* $\mathbf{x}^*_a$ *of problem* (7) *is not unique unless* $\mathbf{x}^*_a$ *is determined by the full model*.

**Proof**: By (7b), the solution $x^*_{aj}, j = M-m+1, \dots, M$, satisfy

$$\sum_{j=M-m+1}^{M} w_{rj}x^*_{aj} \leq y^*_a - K_r, \; r = 0, 1, \dots, R.$$

For nonnegative **x**, by Theorem 1, $y^*_a = \max\{K_r | r \in W^+\}$. Thus, $y^*_a - K_r \geq 0$.

Case 1: If $\mathbf{x}^*_a$ is determined by the full model ($r = 0$), then no component of $\mathbf{w}_0$ can be zero (i.e., the coefficient of any optional attribute in the full model must be nonzero). As a result, $\mathbf{w}_0 > \mathbf{0}$ by Theorem 1. Since $y^*_a = K_0$, we have $\sum_{j=M-m+1}^{M} w_{0j}x^*_{aj} \leq 0$. Given that $\mathbf{x}^*_a \geq \mathbf{0}$ and $\mathbf{w}_0 > \mathbf{0}$, there is only one solution satisfying this condition: $x^*_{aj} = 0, j = M-m+1, \dots, M$.

Case 2: If $\mathbf{w}_r \geq \mathbf{0}, \forall r = 1, \dots, R,$ but $\mathbf{w}_0 \ngeq \mathbf{0}$, then $y^*_a \neq K_0$. In this case, we have at least one attribute $x^*_{aj}$ such that its associated $w_{rj} \neq 0, r \in \{r' | y^*_a - K_{r'} > 0\}$. We can set

$$x^*_{aj} \leq \min\left\{\frac{1}{w_{rj}}(y^*_a - K_r), r \in \{r' | y^*_a - K_{r'} > 0\} \text{ and } r' \in \{1, \dots, R\}\right\} \text{ and } x^*_{ak} = 0, k \neq j, k \in$$

$\{M-m+1,\dots,M\}$. Thus, $\mathbf{x}_a^*$ is not unique, since $x_{aj}^*$ can take any value from zero to a positive number, $\min\left\{\frac{1}{w_{rj}}(y_a^*-K_r), r\in\{r'|y_a^*-K_{r'}>0\} \text{ and } r'\in\{1,\dots,R\}\right\}$.

Case 3: If $\exists w_{rj}<0, j\in\{M-m+1,\dots,M\}, r\in\{0,1,\dots,R\}$, then for the corresponding attribute $x_{aj}^*$, we have

$$w_{rj}x_{aj}^* \le y_a^* - K_r - \sum_{k\neq j, k\in\{M-m+1,\dots,M\}} w_{rk}x_{ak}^*, \quad r=0,1,\dots,R$$

$$x_{aj}^* \ge \frac{1}{w_{rj}}\left\{y_a^* - K_r - \sum_{k\neq j, k\in\{M-m+1,\dots,M\}} w_{rk}x_{ak}^*\right\}, \quad r=0,1,\dots,R$$

Since **x** is nonnegative, we have

$$x_{aj}^* \ge \max_r\left\{0, \frac{1}{w_{rj}}\left\{y_a^* - K_r - \sum_{k\neq j, k\in\{M-m+1,\dots,M\}} w_{rk}x_{ak}^*\right\}\right\}$$

Setting $x_{ak}^*=0, k\neq j, k\in\{M-m+1,\dots,M\}$, since $w_{rj}<0$ and $y_a^*-K_r\ge 0$, we have

$$\frac{1}{w_{rj}}\left\{y_a^* - K_r - \sum_{k\neq j, k\in\{M-m+1,\dots,M\}} w_{rk}x_{ak}^*\right\} = \frac{1}{w_{rj}}\{y_a^*-K_r\}\le 0$$

Therefore, $x_{aj}^*\ge 0$ and the optimal solutions $\mathbf{x}_a^*$ are not unique. QED

**Theorem 2**. *When $\mathbf{x}_a^0$ is inferior to $\mathbf{x}_a^*$, the optimal solution $\mathbf{x}_a^*$ minimizes the sum of absolute imputation errors while preserving the decision maker's decision models. That is, for any data point $\mathbf{x}_a'$ that satisfies the decision models, we have $\sum_{j=M-m+1}^{M}|x_{aj}^*-x_{aj}^0| \le \sum_{j=M-m+1}^{M}|x_{aj}'-x_{aj}^0|$.*

**Proof**: Problem (8) has the following objective function:

$$\min\left\{\sum_{w_j>0, j\in\{M-m+1,\dots,M\}} x_{aj} - \sum_{w_j<0, j\in\{M-m+1,\dots,M\}} x_{aj}\right\}$$

Since $\mathbf{x}_a^0$ is a fixed point, we can write the objective function as

$$\min\left\{\sum_{w_j>0, j\in\{M-m+1,\ldots,M\}}\left(x_{aj}-x_{aj}^0\right)-\sum_{w_j<0, j\in\{M-m+1,\ldots,M\}}\left(x_{aj}-x_{aj}^0\right)\right\}$$

or equivalently,

$$\min\left\{\sum_{j=M-m+1}^{M}\text{sign}\left(w_j\right)\left(x_{aj}-x_{aj}^0\right)\right\}$$

So, $\mathbf{x}_a^*$ is the optimal solution to the above function. That is, for any data point $\mathbf{x}_a'$ that satisfies the model (8) constraints, we have

$$\sum_{j=M-m+1}^{M}\text{sign}\left(w_j\right)\left(x_{aj}^*-x_{aj}^0\right)<\sum_{j=M-m+1}^{M}\text{sign}\left(w_j\right)\left(x_{aj}'-x_{aj}^0\right) \tag{A.4}$$

When $\mathbf{x}_a^0$ is inferior to $\mathbf{x}_a^*$, we have $x_{aj}^0<x_{aj}^*$ if $w_j>0$, and $x_{aj}^0>x_{aj}^*$ if $w_j<0$ (note that $w_j\neq 0$ by definition). That is, $\text{sign}\left(w_j\right)\left(x_{aj}^*-x_{aj}^0\right)$ is always positive. So, the left-hand side of (A.4) is

$$\sum_{j=M-m+1}^{M}\text{sign}\left(w_j\right)\left(x_{aj}^*-x_{aj}^0\right)=\sum_{j=M-m+1}^{M}\left|x_{aj}^*-x_{aj}^0\right| \tag{A.5}$$

For the right-hand side of (A.4), regardless of values of $x_{aj}'$ and $x_{aj}^0$, it is always true that

$$\sum_{j=M-m+1}^{M}\text{sign}\left(w_j\right)\left(x_{aj}'-x_{aj}^0\right)\leq\sum_{j=M-m+1}^{M}\left|x_{aj}'-x_{aj}^0\right| \tag{A.6}$$

It follows from (A.4), (A.5) and (A.6) that

$$\textstyle\sum_{j=M-m+1}^{M}\left|x_{aj}^*-x_{aj}^0\right|\leq\sum_{j=M-m+1}^{M}\left|x_{aj}'-x_{aj}^0\right|. \qquad \text{QED}$$

**Theorem 3** *When $\mathbf{x}_a^0$ is inferior to $x_a^*$, the optimal SMILE solution $\mathbf{x}_a^*$ minimizes the sum of absolute imputation errors while preserving the decision maker's decision models.*

**Proof**. Let $y_a^*$ be the value obtained by substituting $\mathbf{x}_a^0$'s known values $(x_{a,1}^0,\ldots,x_{a,M-m}^0)$ into equation $Y_m$ in the SMI algorithm ($y_a^*$ is subsequently used to find $x_{a,M-m+1}^*,\ldots,x_{a,M-m+j-1}^*$, in the algorithm). Consider the first error term in the sum of errors regarding $x_{a,M-m+1}^0$. Let

$y_{a,m-1}$ be the value obtained by substituting the true values $x^0_{a,1},\ldots,x^0_{a,M-m+1}$, into equation $Y_{m-1}$. When $\mathbf{x}^0_a$ is inferior to $\mathbf{x}^*_a$, we have $x^0_{a,M-m+1} < x^*_{a,M-m+1}$ if $w_{M-m+1} > 0$, and $x^0_{a,M-m+1} > x^*_{a,M-m+1}$ if $w_{M-m+1} < 0$. Since $x^*_{a,M-m+1}$ satisfies both constraints $Y_{m-1}$ and $Y_m$, any value between $x^*_{a,M-m+1}$ and $x^0_{a,M-m+1}$, say $x'_{a,M-m+1}$, will not satisfy constraint $Y_m$ (the corresponding $y'_a$ will be between $y^*_a$ and $y_{a,m-1}$). So, the value closest to $x^0_{a,M-m+1}$ while satisfying both constraints is $x^*_{a,M-m+1}$. We can apply the same method of proof to the other error terms in the sum of errors by iteratively comparing $Y_{m-j}$ with $Y_{m-j+1}$ $(j = 1,\ldots,m)$ defined in the SMI algorithm. QED