Partially linear model estimation for missing response data

N Salam
Statistics Department, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, South Kalimantan, Indonesia
nursalam2011@gmail.com

Abstract. This paper will discuss the estimation of a partially linear (semiparametric) model with missing responses using the normal approach. An estimator class is defined which includes special cases, namely the partially linear imputation estimator, the marginal mean estimator and the trend score weighted estimator. The estimator class is asymptotically normal. The three special estimators have the same asymptotic variance. Based on the above conditions, the mean \( F \) will be estimated, say \( \theta \). The three special estimators above will be used to estimate the mean \( F \), namely in the form of point estimates and confidence intervals with some missing responses using the normal approach method.

1. Introduction
Regression model is a statistical technique used to explain the relationship between an independent variable (independent) as a predictor variable \((A)\) and the dependent variable as a response variable \((F)\) which can be expressed as a form of mathematical model. The regression model is divided into three, namely parametric, nonparametric, and semiparametric (partially linear model) regression models. Partially linear model is a new approach model in regression between two popular regression models, namely parametric regression, and nonparametric regression. Partially linear model is a combined model that contains both parametric and nonparametric components. This partially linear model approach is interesting because there is only partial information about the relationship between the independent variable \((A)\) and the response variable \((F)\) that is known, so that the use of a complete nonparametric regression model is no longer efficient, and the use of a complete parametric regression model may also be wrong. At least, this is the author's motivation to discuss the partially linear model.

The partially linear model assumes that the data \(\{(A_i, B_i, F_i) : i = 1, 2, ..., n\}\) has the form:

\[
F_i = A_i^T \beta + m(B_i) + \varepsilon_i
\]

with \( F_i \) the scalar response variables, \( A_i = (a_{i1}, ..., a_{ip})^T \) and \( B_i = (t_{i1}, ..., t_{ip})^T \) are vectors of explanatory variables, \((A_i, B_i)\) are either independent and identically distributed (i.i.d) random design points or fixed design points, \( m(B_i) \) is the unknown function or in other words \( m(B_i) \) is smooth, namely: \( m(B_i) \in W^p_2[0,1] = \{ m \mid m^{(p)} \text{ absolute continuous at } [0,1 ] , p = 0,1,2,...,p-1, m^{(p)} \in L[0,1] \}\) which is called a Sobolev space of order \( p \) where \( L[0,1] \) is the set of all functions whose squares are integral at intervals \( [0,1] \). While \( \beta = (\beta_1, \beta_2, ..., \beta_p)^T \) is a vector of unknown parameters and \( \varepsilon_i \) are model errors with mean 0 and constant variance (homooscedasticity pattern) (Wang et al., 2004 and Budiantara, 2000).
Based on the partially linear model (1) above, we will estimate the mean $F$, say $\theta$, if there are some missing responses. Or specifically, this paper will discuss the case when some $F$ values in a sample of size $n$ may be missing, but $A$ and $B$ are observed completely and then the mean $F$ is estimated. Namely, obtained a collection of observational data $(F_i, \tau_i, A_i; B_i)$, from equation (1) with data $A_i$ and $B_i$ are observed completely and if $F_i$ is missing then $\tau_i = 0$ and if $F_i$ is complete then $\tau_i = 1$.

2. Literature Views

Before discussing the concept of estimation and asymptotic normality, we first discuss several definitions which are initial concepts that must be understood so that it is easy to follow the discussion discussed.

**Definition 2.1** Confidence Interval [1]

An interval $(l(A_1, \ldots, A_n), u(A_1, \ldots, A_n))$ is called a 100$\gamma$% confidence interval for $\theta$ if: $P[l(A_1, \ldots, A_n) < \theta < u(A_1, \ldots, A_n)] = \gamma$ with $0 < \gamma < 1$. The observation values $l(A_1, \ldots, A_n)$ and $u(A_1, \ldots, A_n)$ are called lower and upper confidence limits, respectively.

**Definition 2.2** Estimation [1]

A statistic, $B = l(A_1, A_2, \ldots A_b)$ which is used to estimate the value of $\lambda(\theta)$ is called the estimator of $\lambda(\theta)$ and an observed value of a statistic, $l(a_1, a_2, \ldots a_b)$ is called an estimate of $\lambda(\theta)$.

**Definition 2.3** Convergence in probability [3]

A sequence of random variables $A_1, A_2, A_3, \ldots$ converges in probability to a random variable $A$ if for every $\varepsilon > 0$, $\lim_{n\to\infty} P(|A_n - A| \geq \varepsilon) = 0$ or $\lim_{n\to\infty} P(|A_n - A| < \varepsilon) = 1$ or can be written $A_n \xrightarrow{p} A$.

**Definition 2.4** Convergence in distribution [3]

A sequence of random variables $A_1, A_2, A_3, \ldots$ converges in the distribution to a random variable $A$, if $\lim_{n\to\infty} F_{a_n}(x) = F_A(a)$ at all point $a$, where $F_a(a)$ is continuous or can be written $A_n \xrightarrow{d} A$.

3. Method

The procedures carried out in this study are as follows:

a. Explain parametric regression and parametric regression estimation.

b. Describe nonparametric regression and nonparametric regression estimation.

c. Constructing a partially linear model (semiparametric regression model) and also the

d. Partially linear model with missing data.

e. Determining the correct partially linear model estimation method and in this paper using the least

square method, which then results in an estimator class containing several special estimators, namely, the regression marginal mean estimator, the regression imputation estimator and the regression trend score weighted estimator.

f. Using the normal approximation method to construct a confidence interval estimate of $\theta$.

g. Explain the procedure to 4 above in stages.

h. Conclude from the results of the discussion.

4. Estimation and Asymptotic Normal

In this chapter, we define the estimator and the asymptotic properties which will be discussed in this paper.
4.1. Estimation

The first step in estimating the partial linear model is to multiply equation (1) by \( \tau_i \) and we get the following results:

\[
\tau_i F_i = \tau_i A_i \beta + \tau_i m(B_i) + \tau_i e_i
\]

And by using the conditional expectation if \( B \) is known, we get:

\[
E[\tau_i F_i | B_i = b] = E[\tau_i A_i^T | B_i = b] \beta + E[\tau_i | B_i = b] m(b),
\]

from the above obtained:

\[
m(b) = m_1(b) - m_1(b)^T \beta,
\]

with: \( m_1(t) = \frac{E[\tau_i A_i | B_i = b]}{E[\tau_i | B_i = b]} \) and \( m_2(t) = \frac{E[\tau_i F_i | B_i = b]}{E[\tau_i | B_i = b]} \).

So it produces: \( \tau_i [F_i - m_2(B_i)] = \tau_i [F_i - m_1(B_i)]^T \beta + \tau_i e_i \),

which implies that an estimator can be based on a least squares regression using \( \tau_i = 1 \) observation and estimation of \( m_i(.) \), \( j = 1, 2 \). Suppose \( Q(.) \) is a kernel function and suppose \( h_n \) is a bandwidth sequence which tends to 0 when \( n \to \infty \), and the weights are defined:

\[
S_{nj}(t) = \frac{Q((b-B_i)/h_n)}{\sum_{j=1}^{n} Q((b-B_j)/h_n)}
\]

Then \( \bar{m}_{1n}(b) = \sum_{j=1}^{n} \tau_i S_{nj}(b) A_j \) and \( \bar{m}_{2n}(b) = \sum_{j=1}^{n} \tau_i S_{nj}(b) B_j \) are consistent estimators of \( m_1(b) \) and \( m_2(b) \), respectively. From (3), the estimator is then defined as an estimator that satisfies:

\[
\min_{\beta} \sum_{i=1}^{n} \tau_i \{(F_i - \bar{m}_{2n}(B_i)) - (A_i - \bar{m}_{1n}(B_i))\}^2
\]

From equation (4), we get the estimator results as follows:

\[
\hat{\beta}_n = \left( \sum_{i=1}^{n} \tau_i ((A_i - \bar{m}_{1n}))(A_i - \bar{m}_{1n}(B_i))^T \right)^{-1} \sum_{i=1}^{n} \tau_i \{(A_i - \bar{m}_{1n}(B_i)) (F_i - \bar{m}_{2n}(B_i))\}
\]

Obtained from observation data \( (A_i,F_i,B_i) \) for \( i \) elements of \( \tau_i = 1 \).

Equation (2) implies that an estimator \( m(b) \) can be defined as:

\[
\bar{m}_n(b) = m_{2n}(b) - \bar{m}_{1n}(b) \hat{\beta}_n
\]

by replacing, \( \beta, m_1(b) \) and \( m_2(b) \) in (2) with \( \hat{\beta}_n, \bar{m}_{1n}(b) \) and \( \bar{m}_{2n}(b) \).

In discussing the estimation of, the general class of estimators is determined, namely:

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\tau_i F_i}{D_n^*(A_i, B_i)} + \left( 1 - \frac{\tau_i}{D_n^*(A_i, B_i)} \right) \left( A_i^T \hat{\beta}_n + \bar{m}_n(B_i) \right),
\]

with \( D_n^*(x,t) \) is a sequence that has limits of probability \( D^*(a,b) \).

This paper focuses on the special case, namely, first, when \( D_n^*(a,b) = 1 \), the regression imputation estimator of \( \theta \) is obtained:

\[
\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} \left\{ \tau_i F_i + (1 - \tau_i) \left( A_i^T \hat{\beta}_n + \bar{m}_n(B_i) \right) \right\}.
\]

Second, when \( D_n^*(x,t) = \infty \), we get the regression marginal mean estimator of \( \theta \):

\[
\hat{\theta}_{MM} = \frac{1}{n} \sum_{i=1}^{n} \left( A_i^T \hat{\beta}_n + \bar{m}_n(B_i) \right).
\]
Third, if the marginal propensity score is defined $D_i(b) = P(\tau = 1 | B = b)$ and when $D^*_n(a,b) = D_1(b) = \sum_{i=1}^n \tau_i Q\left(\frac{b - \hat{m}_i}{h_n}\right)/\sum_{i=1}^n Q\left(\frac{b - \hat{m}_i}{h_n}\right)$. In this case, the trend score weighted estimator is obtained, namely:

$$\hat{\theta}_{T1} = \frac{1}{n} \sum_{i=1}^n \left[ \tau_{ij} F_i + \left(1 - \frac{\tau_{ij}}{D_i(B_i)}\right) \left(F_i^T \hat{\beta}_n + \hat{m}_n(B_i)\right) \right].$$

4.2. Asymptotic Normality

Next, the properties of the $\hat{\theta}$ will be given, namely $D_n^*(a,b) = \{1, \infty, \hat{D}_1(a,b), \hat{D}_n(a,b)\}$ and their consistent variance estimators.

Let $D_j(t) = P(\tau = t | B = b), D(a,b) = P(\tau = 1 | A = a, B = b), g(a,b) = a^T \beta + m(b)$ dan

$$\sigma^2(a,b) = E\left[(F - a^T \beta - m(B))^T (A = a, B = b)\right].$$

Then defined $\Sigma = [D(A,B) u (A,B) u (A,B)^T] s(a,b) = a_m$.

$m_i(.)$ denotes the rth component with $m_i(.)$. Given $\|\|_2$ is the Euclidean norm of $\hat{\theta}$. Some of the assumptions below are used in constructing the asymptotic normality of the $\hat{\theta}$ estimator:

1. $\sup_{a,b} E \left[\|F - B = b\|_2^2\right] < \infty$.
2. The probability density function of $B$, say $h(b)$, exists and fulfills $0 < \inf h(b) \leq \sup h(b) < \infty$.
3. $\sup_{a,b} E \left[\|F - a, B = b\|_2^2\right] < \infty$.
4. $m(.,.), m_1(.)$ and $m_2(.)$ fulfill Lipschitz condition of order 1.
5. $\Sigma = E \left[D(A,B) s(A,B) s(A,B)^T\right]$ is a positive definite matrix.
6. (a) There are constants $N_1 > 0, N_2 > 0$ and $N > 0$ such that: $N_1 [U] \leq \sup \leq K(U) \leq N_2 [U] \leq \sup |\phi|$.
(b) $Q(\cdot)$ is a second order kernel function.
(c) $Q(\cdot)$ has limited partial derivatives up to order 2 almost certain.
7. (a) The kernel function $s_i(\cdot)$ is a limited kernel function with limited support and variance.
(b) $S_i(\cdot)$ is a kernel function of order k.
8. $nh \rightarrow 2(d+1)\rightarrow \infty$ and $nh^2 \rightarrow 0$.
9. $nh^{2d+1}/\log n \rightarrow \infty$ and $nh^{2k-2} \rightarrow 0$.

**Theorem 4.1** Based on all these assumptions except for 6 (c) it is obtained:

$$\sqrt{n} \left(\hat{\theta} - \theta\right) \overset{d}{\rightarrow} N(0,D),$$

where $D = E\left[\left(\omega_0(A,B) + \omega_1(A,B)\right)^2 D(A,B) \sigma^2(A,B)\right] + \text{var} [g(A,B)]$, with $\omega_0(a,b) = 1/D_1(b)$ and $\omega_1(a,b) = E\left[s(A,B)^T\right] \Sigma^{-1} \times s(a,b)$ when $D^*_n(x, t) \in \{1, \infty, \hat{D}_1(t)\}$, and $\omega_0(a,b) = 1/D(a,b)$ and $\omega_1(a,b) = 0$ when $D^*_n(a,b)$ is taken as $D(a,b)$.

As for obtaining a consistent estimator from D, it is done using the kernel regression method, where the estimators $D(a,b), D_1(b), \sigma^2(a,b)$ and $m_i(.)$ are defined. Then further define a consistent estimator of D with a similar (plug in) method.

However, when this plug-in method has high dimensions, this method cannot estimate the D estimator properly. This situation can be anticipated because the estimators of $D(a,b)$ and $\sigma^2(a,b)$ only appear as numerators and with squared errors or indicator functions can be replaced appropriately.
The Jackknife variance estimator can be used as an option. Suppose \( \hat{\theta}(-i) \) to be \( \hat{\theta} \) based on \( \{(F_j, \tau_i, A, B_j)\}_{j \neq i} \) for \( i = 1, 2, \ldots, n \). Suppose these are pseudo jackknife values. Namely \( K_{ni} = n\hat{\theta} - \hat{\theta}(-i)(n - 1), i = 1, 2, \ldots, n \). Then the Jackknife variance estimator can be formed as follows:

\[
\hat{\ell}_{nj} = \frac{1}{n} \sum_{i=1}^{n} (K_{ni} - \bar{K}_n)^2 \quad \text{with} \quad \bar{K}_n = n^{-1} \sum_{i=1}^{n} K_{ni}.
\]

**Theorem 4.2** Based on the assumptions of Theorem 4.1, we get \( \hat{\ell}_{nj} \xrightarrow{p} D \).

Based on Theorems 4.1 and 4.2, it can be obtained that the estimation of the confidence interval based on the normal approach with a confidence level of 1 - \( \alpha \) for \( \hat{\theta} \) is:

\[
\hat{\theta} - \sqrt{\hat{\ell}_n} \times z_{\alpha/2} < \theta < \hat{\theta} + \sqrt{\hat{\ell}_n} \times z_{\alpha/2},
\]

where \( z_{\alpha/2} \) is the \( z \) value where the area on the right under the standard normal curve is \( \alpha/2 \).

### 5. Conclusions and Suggestions

#### 5.1. Conclusion

The conclusions obtained based on the results of the discussion are as follows:

1. The results of the point estimation of the semiparametric regression model with missing responses in the form of an estimator class are:

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\tau_i F_i}{D_n^i(A_i, B_i)} + \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{\tau_i}{D_n^i(A_i, B_i)}\right) \left(A_i^T \hat{\beta}_n + \hat{m}_n(B_i)\right).
\]

2. The results of the estimation of the confidence interval of the semiparametric regression model with missing responses using the normal approach method with a confidence level of 1 - \( \alpha \) for is:

\[
\theta - \sqrt{\hat{\ell}_n} \times z_{\alpha/2} < \theta < \hat{\theta} + \sqrt{\hat{\ell}_n} \times z_{\alpha/2}.
\]

#### 5.2. Suggestion

This study uses the normal approach method, therefore other or further studies can be carried out with the empirical likelihood method.

### References

[1] Bain L J E 1992 *Introduction to Probability and Mathematical Statistics* (California: Duxbury Press)

[2] Budi A I N 2000 *Spline Estimator in Nonparametric and Semiparametric Regression* (Yogyakarta: Dissertation Program, Gadjah Mada University)

[3] Casella G B R L 2002 *Statistical Inference* (USA: Duxbury)

[4] Härdle W, Liang H and Gao J 2012 *Partially linear models* (Heidelberg: Physica-Verlag)

[5] HARDLE W K, LINTON O B and WANG Q 2003 Semiparametric regression analysis under imputation for missing response data *LSE STICERD Res. Pap.*