Statistics and Application 统计学与应用, 2022, 11(6), 1515-1520
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.116158

Hans汉斯

因变量缺失下线性回归模型的岭估计

李　静[1]，安佰玲[2]

[1]中国劳动关系学院应用技术学院，北京
[2]淮北师范大学数学科学学院，安徽　淮北

收稿日期：2022年11月27日；录用日期：2022年12月17日；发布日期：2022年12月29日

摘　要

本文研究线性回归模型存在多重共线性且因变量存在缺失时的估计问题，分别基于完整数据方法和单点插补方法，给出了回归系数的两种岭估计，给出了估计量的渐近性质。最后通过数值模拟验证了所提方法的有效性。

关键词

线性回归模型，缺失数据，插补方法，最小二乘估计，岭估计

Ridge Estimation of Linear Regression Model with Missing Responses

Jing Li[1], Bailing An[2]

[1]School of Applied Technology, China University of Labor Relations, Beijing
[2]School of Mathematical Sciences, Huaibei Normal University, Huaibei Anhui

Received: Nov. 27th, 2022; accepted: Dec. 17th, 2022; published: Dec. 29th, 2022

Abstract

This paper discusses estimation of linear regression models in the presence of multicollinearity and the responses are missing at random. Based on the complete-case method and single imputation technique, two ridge estimators for the unknown regression coefficients are proposed. Asymptotically properties of the proposed estimators are given. Finally, some simulations are conducted to illustrate the proposed methods.

Keywords

Linear Regression, Missing Data, Imputation Method, Least-Squares Estimation, Ridge Estimation

文章引用: 李静, 安佰玲. 因变量缺失下线性回归模型的岭估计[J]. 统计学与应用, 2022, 11(6): 1515-1520.
DOI: 10.12677/sa.2022.116158

1. 引言

缺失数据是数据分析中遇到的常见问题之一，关于缺失数据的详细介绍可参考著作 Little 和 Ruin (2002) [1]。线性回归模型是最为常用的统计方法之一，缺失数据下回归模型的研究得到了关注，Cheng (1994) [2]、Chu & Cheng (1995) [3]、Wang & Rao (2002) [4]和 Qin 等(2009) [5]基于不同的方法研究了因变量缺失下的回归模型。另一方面，多种共线性是线性回归模型实际使用中经常遇到的问题，这一问题是由于自变量之间存在线性关系导致，严重影响统计推断结果。解决多重共线性的途径有多重，其中构造回归系数的有偏估计是一种常见的方法，有偏估计以牺牲无偏性使估计量的均方误差变小，从而解决了多重共线性下最小二乘估计量方差过大的问题。目前讨论较多的有偏估计是岭估计和主成份估计。

线性回归模型可记为如下形式

$$Y_i = X_i^{\mathrm{T}}\beta + \varepsilon_i,\ i = 1,2,\cdots,n \tag{1}$$

其中 Y_i 是因变量观测值，$X_i = \left(X_{i1}, X_{i2}, \cdots, X_{iq}\right)^{\mathrm{T}}$ 是相应的自变量观测值，β 为 $q\times 1$ 需要估计的未知回归系数，模型误差 ε_i 为独立同分布的随机变量，有 $E\left(\varepsilon_i \middle| X_i\right) = 0$ 和 $Var\left(\varepsilon_i \middle| X_i\right) = \sigma^2$。为了克服因变量的缺失，用变量 δ 作为缺失的标志，$\delta_i = 1$ 表示 Y_i 的值没有缺失，可以被观测到，$\delta_i = 0$ 则表示 Y_i 值缺失。假定 Y_i 满足随机缺失的机制，即

$$P\left(\delta = 1 \middle| X, Y\right) = P\left(\delta = 1 \middle| X\right). \tag{2}$$

该缺失机制是缺失数据分析中常用的假设条件。

针对缺失数据的处理，最简单的处理方法就是完整数据方法，即只利用因变量和自变量都完整的观测值，也就是只考虑因变量不存在缺失的那些观测值，即 $\delta_i = 1$ 的那些观测数据。显然，这种方法虽然简便，但舍弃了存在缺失的数据，造成了信息的浪费。为了充分利用观测数据的信息，还可以利用插补的方法对模型(1)进行估计。对于模型(1)~(2)，基于完整数据方法和单点插补方法，杨徐佳等(2011) [6]构造了回归系数的估计，并给出了所提估计量的渐近性质，此外还讨论了回归系数的线性关系检验问题。安佰玲等(2013) [7]则基于这两种方法讨论了模型的约束估计问题。

针对因变量缺失和自变量存在多重共线性这两个问题，目前的研究大都是单独讨论，将两个问题同时考虑的研究成果很少。为了解决这一问题，本文集中讨论因变量缺失下线性回归模型的岭估计问题。

论文第 2 节和第 3 节将分别基于完整数据方法和单点插补方法构造模型系数的岭估计，并给出估计量的渐近性质。第 4 节将通过数值模拟验证所提方法的有效性。总结将在第 5 节给出，定理的证明将放在第 6 节。

2. 基于完整数据方法的岭估计

本节基于完整数据方法构造模型系数的岭估计。设 $\left\{Y_i, \delta_i, X_i\right\}_{i=1}^{n}$ 为来自模型(1)的观测数据，则有如下的线性回归模型

$$\delta_i Y_i = \delta_i X_i^{\mathrm{T}}\beta + \delta_i\varepsilon_i,\ i = 1,2,\cdots,n. \tag{4}$$

模型(4)的矩阵形式为

$$\bar{Y}=\bar{X}\beta+\bar{\varepsilon}, \tag{5}$$

其中 $\bar{Y}=\left(\delta_1 Y_1,\delta_2 Y_2,\cdots,\delta_n Y_n\right)^{\mathrm{T}}$，$\bar{X}=\left(\delta_1 X_1,\delta_2 X_2,\cdots,\delta_n X_n\right)^{\mathrm{T}}$，$\bar{\varepsilon}=\left(\delta_1\varepsilon_1,\delta_2\varepsilon_2,\cdots,\delta_n\varepsilon_n\right)^{\mathrm{T}}$。

从而对模型(4)使用最小二乘估计，可得回归系数 β 基于完整数据的估计

$$\hat{\beta}_C=\left(\sum_{i=1}^{n}\delta_i X_i X_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\delta_i X_i Y_i\right)=\left(\bar{X}^{\mathrm{T}}\bar{X}\right)^{-1}\bar{X}^{\mathrm{T}}\bar{Y}. \tag{6}$$

下面在模型(4)或(5)的基础上考虑岭估计的构造。类似于普通线性回归模型的估计，构造的如下的辅助函数：

$$F_1(\beta)=\left(\bar{Y}-\bar{X}\beta\right)^{\mathrm{T}}\left(\bar{Y}-\bar{X}\beta\right)+k\beta^{\mathrm{T}}\beta \tag{7}$$

函数 $F(\beta)$ 关于 β 求偏导数，并另导数等于 0，可得

$$\frac{\partial F_1(\beta)}{\partial\beta}=-2\bar{X}^{\mathrm{T}}\bar{Y}+2\bar{X}^{\mathrm{T}}\bar{X}\beta+2k\beta=0 \tag{8}$$

从而可得回归系数基于完整数据方法的岭估计为

$$\hat{\beta}_C(k)=\left(\bar{X}^{\mathrm{T}}\bar{X}+kI_q\right)^{-1}\bar{X}^{\mathrm{T}}\bar{Y} \tag{9}$$

下面给出 $\hat{\beta}_c(k)$ 的渐近性质。

定理 1. 如果第 6 节的假设条件成立，$\hat{\beta}_c(k)$ 是渐近正态的，有

$$\sqrt{n}\left(\hat{\beta}_c^r-\beta\right)\xrightarrow{d}N\left(0,\sigma^2\Omega^{-1}\right),$$

其中 $\Omega=E\left(\delta XX^{\mathrm{T}}\right)$。

3. 基于单点插补方法的约束估计及其性质

上一节所用的完整数据方法只是用了因变量和自变量有完整观测的数据，将因变量存在缺失的观测数据舍弃，显然造成了信息的损失。为了弥补这一问题，下面将基于单点插补方法构造模型系数的岭估计。基于上一节得到的最小二乘估计 $\hat{\beta}_C$，针对因变量缺失这一问题，构造如下新的因变量

$$Y_i^*=\delta_i Y_i+(1-\delta_i)X_i^{\mathrm{T}}\hat{\beta}_C, \tag{10}$$

显然，当 Y_i 不存在缺失时，$Y_i^*=Y_i$，另一方面当 Y_i 存在缺失时，相当于用插补值 $Y_i^*=X_i^{\mathrm{T}}\hat{\beta}_C$ 代替其缺失的真实值。

基于上面的构造，得到了完整数据集 $\left(Y_i^*,X_i\right)_{i=1}^{n}$，因此有如下的自变量和因变量都存在的线性模型

$$Y_i^*=X_i^{\mathrm{T}}\beta+e_i,\ i=1,2,\cdots,n. \tag{11}$$

其中 $Y_i^*=\delta_i Y_i+(1-\delta_i)X_i^{\mathrm{T}}\hat{\beta}_c$。

对于上述模型，利用最小二乘方法可以得到 β 的单点插补估计

$$\hat{\beta}_I=\left(\sum_{i=1}^{n}X_i X_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}X_i Y_i^*\right). \tag{12}$$

基于模型(11)，考虑其岭估计，构造如下的辅助函数

$$F_2(\beta,\lambda)=\sum_{i=1}^{n}\left(Y_i^*-X_i^{\mathrm{T}}\beta\right)^2+k\beta^{\mathrm{T}}\beta$$

同样，函数 $F_2(\beta)$ 关于 β 求偏导数，并另导数等于 0，可得

$$\frac{\partial F_2(\beta)}{\partial \beta} = -2X^{\mathrm{T}}Y^* + 2X^{\mathrm{T}}X\beta + 2k\beta = 0 \tag{13}$$

$Y^* = \left(Y_1^*, Y_2^*, \cdots, Y_n^*\right)^{\mathrm{T}}$，从而可得回归系数基于单点插补数据方法的岭估计为

$$\hat{\beta}_I(k) = \left(X^{\mathrm{T}}X + kI_q\right)^{-1} X^{\mathrm{T}}Y^* \tag{14}$$

下面给出 $\hat{\beta}_I(k)$ 的渐近性质。

定理 2. 如果第 6 节的假设条件成立，$\hat{\beta}_I^r$ 是渐近正态的，满足

$$\sqrt{n}\left(\hat{\beta}_I(k) - \beta\right) \xrightarrow{d} N\left(0, \sigma^2\Omega^{-1}\right).$$

从定理 1 和定理 2 不难看出，首先基于两种方法的估计量的渐近性质相同，且与不考虑岭估计的最小二乘估计的渐近性质一样。这些结论与杨徐佳等(2011)以及其他文献的结论一致。

4. 数值模拟

本节将通过数值模拟考察前面所提出估计方法的有效性。假设数据服从于如下线性回归模型

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i,\ i = 1, 2, \cdots, n \tag{15}$$

为了构造多重共线性，运用 McDonald 和 Galerneau (1975) [8]中生成具有多重共线性自变量的方法生成解释变量 x_1, x_2，具体为：

$$x_{ij} = \left(1-\rho^2\right)^{1/2}\omega_{ij} + \rho\omega_{i3},\quad i = 1, 2, \cdots, n;\ \ j = 1, 2$$

其中 ω_{ij} 是独立的标准正态随机数，ρ 是一个具体的数值以确保任何四个解释变量在理论上是相关的，分别取 $\rho = 0.90, 0.99, 0.999$ 刻画不同程度的复共线性问题。

因变量的缺失的机制

$\Delta = p\left(\delta = 1 \mid x_1 = x_{1i}, x_2 = x_{2i}\right) = 0.8 + 0.2\left(\left|x_{1i}\right| + \left|x_{2i}\right|\right)$，当 $\left|x_{1i}\right| + \left|x_{2i}\right| \le 1$ 时；

否则等于 0.9。模型误差 ε_i 服从如下的正态分布(*N*)和均匀分布(*U*)

(1) $\varepsilon_i \sim N\left(0, 0.5^2\right)$，(2) $\varepsilon_i \sim U\left(-\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2}\right)$。

针对模型(15)，取 $\left(\beta_1, \beta_2\right)$ 的真实值为 $(3, 2)$，基于上面的缺失机制和误差分布，样本量 n 分别设置为 50、100 和 150，重复 500 次，在每一种情况下分别计算 $\left(\beta_1, \beta_2\right)$ 基于完整数据分析方法的估计(C)和岭估计(C-R)，基于单点插补方法的估计(I)和岭估计(I-R)。其中岭估计是基于 R 软件 MASS package 里的 lm.ridge 函数，其中 k 的选取使用 GCV 方法。以这些估计量的均方误差(EMSE)来衡量其表现，

$$\mathrm{EMSE}\left(\beta^*\right) = \frac{1}{500}\sum_{k=1}^{500}\sum_{j=1}^{2}\left(\beta_{kj}^* - \beta_j\right)^2$$

其中 β_{kj}^* 是参数 β_j 的第 k 次重复时的估计值，模拟结果见表 1。

Table 1. EMSEs of the estimators
表 1. 不同估计量的 EMSEs

ρ	β	n = 50		n = 100		n = 150	
		N	*U*	*N*	*U*	*N*	*U*
0.9	C	0.0329	0.0334	0.0169	0.0185	0.0110	0.0109
	C-R	0.0324	0.0334	0.0168	0.0186	0.0109	0.0109

Continued

0.9	I	0.0331	0.0350	0.0163	0.0166	0.0111	0.0101
	I-R	0.0328	0.0345	0.0163	0.0166	0.0110	0.0101
0.99	C	0.3042	0.3162	0.1429	0.1482	0.0928	0.0893
	C-R	0.2529	0.2635	0.1476	0.1492	0.0994	0.0982
	I	0.3022	0.2946	0.1434	0.1498	0.0886	0.1015
	I-R	0.2596	0.2632	0.1433	0.1495	0.0949	0.1071
0.999	C	3.0830	3.0819	1.4055	1.4409	0.9294	0.9186
	C-R	1.4105	1.4410	0.7370	0.8122	0.5701	0.5618
	I	2.7488	3.0282	1.3016	1.3347	1.0163	0.9194
	I-R	1.3923	1.5274	0.7607	0.7701	0.6579	0.5810

从模拟结果可以看出：1) 随着样本量的增大，这四类估计的均方误差值都在变小，与理论性质相一致。2) 误差分布对这四类估计的影响很小。3) 随着共线性程度的增加，岭估计优于其对应的最小二乘估计。4) 完整数据估计和单点插补估计相比，单点插补估计由于充分利用了数据的信息从而表现优于损失了信息的完整数据估计。

5. 总结

因变量缺失问题是使用线性回归模型进行实际问题分析时经常遇到的，目前的研究大都是基于最小二乘法或极大似然估计方法进行模型估计，很少有论文同时讨论自变量的多重共线性问题。本文就是针对这一问题，构造了因变量缺失下的线性回归模型的岭估计，并研究了所提估计量的渐近性质，通过数值模拟验证了所提估计的有效性。本文主要讨论了线性回归模型，论文的方法可以推广到半参数模型等其他类型的回归模型上。

6. 定理的证明

在给出定理的证明之前，我们先给出下面条件。

条件 1：$\Omega = E\left(\delta XX^{\mathrm{T}}\right)$ 为正定矩阵。

条件 2：$E\left(\varepsilon \mid X\right) = 0$，$E\left(\left|\varepsilon\right|^3 \mid X\right) < \infty$。

引理 1. 如果前面的假设条件成立，$\hat{\beta}_c$ 和 $\hat{\beta}_I$ 都是渐进正态的，二者都满足

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2\Omega^{-1}\right),$$

其中 $\hat{\beta}$ 为 $\hat{\beta}_c$ 或 $\hat{\beta}_I$。

证明：该引理即为杨徐佳等(2011) [6]中的定理 1。

定理 1 证明：由 $\hat{\beta}_c\left(k\right)$ 的定义可得

$$\begin{aligned}\hat{\beta}_c\left(k\right) &= \left(\overline{X}^{\mathrm{T}}\overline{X} + kI_q\right)^{-1}\overline{X}^{\mathrm{T}}\overline{Y} \\ &= \left(X^{\mathrm{T}}\Delta X + kI_q\right)^{-1}X^{\mathrm{T}}\Delta Y \\ &= \beta - k\left(X^{\mathrm{T}}\Delta X + kI_q\right)^{-1}\beta + \left(X^{\mathrm{T}}\Delta X + kI_q\right)^{-1}X^{\mathrm{T}}\Delta\varepsilon\end{aligned}$$

基于条件 1 和 2，可得到

$$\frac{1}{n}\left(X^{\mathrm{T}}\Delta X+kI_q\right)\xrightarrow{P}\Omega,\frac{1}{\sqrt{n}}k\beta\xrightarrow{P}0,\frac{1}{\sqrt{n}}X^{\mathrm{T}}\Delta\varepsilon\xrightarrow{L}\Omega,$$

由上述结论，根据 Slutsky 定理可得

$$\sqrt{n}\left(\hat{\beta}_C(k)-\beta\right)\xrightarrow{d}N\left(0,\sigma^2\Omega^{-1}\right).$$

定理 2 的证明和定理 1 类似，在此省略。

基金项目

中国劳动关系学院教育教学改革立项项目(JG1406)；2020 年度安徽高等学校自然科学项目(KJ2020A1200)。

参考文献

[1] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons, Inc., New York.

[2] Cheng, P.E. (1990) Nonparametric Estimation of Mean Functionals with Data Missing at Random. *Journal of the American Statistical Association*, **89**, 81-87. https://doi.org/10.1080/01621459.1994.10476448

[3] Chu, C.K. and Cheng, P.E. (1995) Nonparametric Regression Estimation with Missing Data. *Journal of Statistical Planning Inference*, **48**, 85-99. https://doi.org/10.1016/0378-3758(94)00151-K

[4] Wang, Q.H. and Rao, J.N.K. (2002) Empirical Likelihood-Based Inference under Imputation for Missing Response Data. *Annals of Statistics*, **30**, 896-924. https://doi.org/10.1214/aos/1028674845

[5] Qin, Y.S., Li, L. and Lei, Q.Z. (2009) Empirical Likelihood for Linear Regression Models with Missing Responses. *Statistics and Probability Letters*, **79**, 1391-1396. https://doi.org/10.1016/j.spl.2009.03.002

[6] 杨徐佳, 于倩倩, 王森. 因变量缺失下线性回归模型的估计与检验[J]. 淮北煤炭师范学报(自然科学版), 2011, 32(1): 24-28.

[7] 安佰玲. 线性回归模型在因变量缺失下的约束估计[J]. 统计与决策, 2013(11): 19-21.

[8] McDonald, G.C. and Galarneau, D.I. (1975) A Monte Carlo Evaluation of Some Ridge-Type Estimators. *Journal of American Statistical Association*, **70**, 407-416. https://doi.org/10.1080/01621459.1975.10479882