

# 因变量缺失下线性回归模型的约束岭估计

李 静

中国劳动关系学院计算机学院, 北京

收稿日期: 2025年10月27日; 录用日期: 2025年11月18日; 发布日期: 2025年11月28日

## 摘 要

本文考虑线性回归模型在因变量存在缺失且回归系数存在线性约束下的岭估计, 分别基于完整数据方法和单一插补方法, 构造了回归系数的约束岭估计, 解决了线性回归模型在因变量缺失、自变量存在多重共线性和回归系数存在线性约束这三种情况同时存在情形下的估计问题, 给出了所提估计量的渐近性质。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, 因变量缺失, 多重共线性, 线性约束, 岭估计

# Restricted Ridge Estimation of Linear Regression Model with Missing Responses

Jing Li

School of Computer Science, China University of Labor Relations, Beijing

Received: October 27, 2025; accepted: November 18, 2025; published: November 28, 2025

## Abstract

This paper addresses ridge estimation in linear regression models when the dependent variable has missing values and the regression coefficients are subject to linear restrictions. Based on the complete data method and single imputation method, the corresponding restricted ridge estimators for the regression coefficients are constructed. This approach resolves the estimation problem in linear regression models when three conditions coexist: missing dependent variables, multicollinearity, and linear restriction on regression coefficients. The asymptotic properties of the proposed estimators are given. Finally, some simulations are conducted to illustrate the proposed methods.

## Keywords

**Linear Regression Model, Missing Responses, Multicollinearity, Linear Restriction, Ridge Estimation**

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

变量之间关系的研究是统计学的核心任务之一，线性回归模型作为研究因变量与自变量关系的建模方法是使用和得到研究最多的统计方法之一。关于线性回归模型的估计、假设检验以及变量选择等推断方法都已经有了成熟的方法。考虑如下的线性回归模型

$$Y_i = X_i^T \beta + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

其中  $Y_i$  和  $X_i = (X_{i1}, X_{i2}, \dots, X_{iq})^T$  是相应的因变量和自变量观测值，回归系数  $\beta = (\beta_1, \beta_2, \dots, \beta_q)$  为未知待估参数，误差  $\varepsilon_i$  满足独立同分布，有  $E(\varepsilon_i | X_i) = 0$  和  $Var(\varepsilon_i | X_i) = \sigma^2$ 。

实际问题研究中，一方面由于测量或者无回答等原因，变量的观测值存在缺失是经常遇到的，如何处理缺失数据一直受到研究者的关注。Little 和 Rubin (2002) [1] 针对缺失数据的分析和建模做了详细的介绍，Cheng (1994) [2]、Chu & Cheng (1995) [3]、杨徐佳等(2011) [4]、李静和安佰玲(2022) [5] 和安佰玲等(2013) [6] 针对回归模型在因变量缺失下的统计推断问题提出了不同的方法。另一方面，当线性回归模型的自变量之间存在较强的线性关系时，会遭遇到多重共线性问题，这一问题将影响模型的统计推断，比如回归系数的最小二乘估计虽然依然是无偏估计，但估计量的方差很大，反映到具体的估计值上表现就是回归系数的符号和预期相反，或者回归系数的绝对值偏大等。为了解决多重共线性，多种方法被提出，其中一种得到广泛使用的方法是采用岭估计这一有偏估计方法，能降低估计量的均方误差。最后，使用线性回归模型分析实际问题时，实际问题本身或者外部信息能够对于回归系数有一些额外信息或者要求，这些信息或要求能够表现为回归系数的线性约束，因此我们需要在这些约束条件下构造模型的约束估计，相关内容可参考 Rao 等(2008) [7]。

本文主要研究线性回归模型(1)在上面提到的因变量存在缺失、自变量存在多重共线性和回归系数存在线性约束这三种情形共同存在时的估计问题。首先考虑  $Y_i$  满足随机缺失机制，即有

$$P(\delta = 1 | X, Y) = P(\delta = 1 | X)。 \quad (2)$$

其中变量  $\delta$  为缺失的标志， $\delta_i = 1$  表示  $Y_i$  的值可以被观测到没有缺失。 $\delta_i = 0$  则表示  $Y_i$  值无法观测到存在缺失。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，常用的方法有两类，一类完整数据法，该方法是删除掉因变量有缺失的观测数据，只保留因变量和自变量都能观测到的完整观测数据。另外一类是插补方法，即利用插补技术补充缺失的因变量，从而得到完整数据。使用较多的是单一插补方法，即对缺失值只提供一个插补值。

假定模型系数的线性约束形式为

$$A\beta = b \quad (3)$$

其中  $A$  为  $k \times q$  的已知矩阵，有  $rank(A) = k$ ， $b$  为  $k \times 1$  的已知向量。

杨徐佳等(2011) [4] 基于完整数据方法和单一插补方法, 研究了模型(1)存在因变量缺失下的估计和检验, 在此基础上安佰玲等(2013) [6] 基于完整数据方法和单一插补方法讨论了模型的约束估计问题。李静和安佰玲(2022) [5] 讨论了因变量缺失下的岭估计。在这些结果的基础上, 本文将考虑线性回归模型在因变量缺失(2)和回归系数存在线性约束条件(3)的情况下的岭估计, 目前还没有文献同时考虑这三种情况同时存在。

论文的第 2 节和第 3 节将分别讨论完整数据法和单一插补法下的回归系数的岭估计, 并研究估计量的渐近性质。第 4 节通过数值模拟考察所提估计量的表现。第 5 节是总结, 定理的证明第 6 节给出。

## 2. 完整数据约束岭估计

本节基于完整数据技术构造回归系数的岭估计。 $\{Y_i, \delta_i, X_i\}_{i=1}^n$  为来自模型(1)的观测数据, 则有如下的线性回归模型

$$\delta_i Y_i = \delta_i X_i^T \beta + \delta_i \varepsilon_i, \quad i=1, 2, \dots, n. \quad (4)$$

模型可写成如下的矩阵形式

$$\bar{Y} = \bar{X} \beta + \bar{\varepsilon}, \quad (5)$$

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

可得  $\beta$  基于完整数据方法的最小二乘估计

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

下面构造模型的约束条件下的岭估计, 为了使用拉格朗日乘子法, 构造如下的辅助函数

$$F_1(\beta, \lambda) = (\bar{Y} - \bar{X} \beta)^T (\bar{Y} - \bar{X} \beta) + k \beta^T \beta + 2 \lambda^T (A \beta - b) \quad (7)$$

其中拉格朗日乘子  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)^T$ , 对辅助函数对  $\beta$  求偏导, 并令它们等于 0, 我们有

$$\frac{\partial F_1(\beta, \lambda)}{\partial \beta} = -2 \bar{X}^T \bar{Y} + 2 \bar{X}^T \bar{X} \beta + 2k \beta + 2 A^T \lambda = 0$$

整理可得

$$\beta = (\bar{X}^T \bar{X} + k I_q)^{-1} \bar{X}^T \bar{Y} - (\bar{X}^T \bar{X} + k I_q)^{-1} A^T \lambda = \hat{\beta}_C(k) - (\bar{X}^T \bar{X} + k I_q)^{-1} A^T \lambda \quad (8)$$

其中  $\hat{\beta}_C(k) = (\bar{X}^T \bar{X} + k I_q)^{-1} \bar{X}^T \bar{Y}$  为回归系数基于完整数据方法的岭估计, 其中的参数  $k$  为岭参数, 主要用来降低估计量的方差, 在估计量的方差与偏误上达到平衡, 其选择也是一个重要的问题。使用较多的方法有岭迹图法, 该方法将绘制每个回归系数的岭估计随着岭参数变化而变化的图像, 根据变化趋势选择  $k$  值。广义交叉证实法是选择岭参数  $k$ , 使得模型的预测误差达到最小。此外也可以使用 AIC 或 BIC 信息准则选择岭参数。

将式(8)带入到约束条件(3)中有

$$A \hat{\beta}_C(k) - A (\bar{X}^T \bar{X} + k I_q)^{-1} A^T \lambda = b$$

从而有

$$A (\bar{X}^T \bar{X} + k I_q)^{-1} A^T \lambda = A \hat{\beta}_C(k) - b \quad (9)$$

由此求解可得

$$\lambda = \left[ A(\bar{X}^T \bar{X} + kI_q)^{-1} A^T \right]^{-1} [A\hat{\beta}_C(k) - b]$$

将其带入(8)式可得回归系数基于完整数据方法的约束岭估计为

$$\hat{\beta}_C^R(k) = \hat{\beta}_C(k) - (\bar{X}^T \bar{X} + kI_q)^{-1} A^T \left[ A(\bar{X}^T \bar{X} + kI_q)^{-1} A^T \right]^{-1} [A\hat{\beta}_C(k) - b]$$

下面给出  $\hat{\beta}_C^R(k)$  的渐近性质。

**定理 1:** 如果第 6 节的假设条件成立,

$$\sqrt{n}(\hat{\beta}_C^R(k) - \beta) \xrightarrow{d} N[0, \sigma^2(\Omega^{-1} - \nabla)],$$

其中  $\Omega = E(\delta XX^T)$ ,  $\nabla = \Omega^{-1} A^T (A\Omega^{-1} A^T)^{-1} A\Omega^{-1}$ 。

定理 1 表明基于完整数据方法的约束岭估计的渐近性质与安佰玲等(2013) [6]构造的完整数据方法下的约束估计量的渐近性质一样, 类似结果在杨徐佳等(2011) [4]中也存在。此外由

$\nabla = \Omega^{-1} A^T (A\Omega^{-1} A^T)^{-1} A\Omega^{-1}$  是一非负定阵, 从而考虑了约束条件的岭估计  $\hat{\beta}_C^R(k)$  比不考虑约束条件的岭估计  $\hat{\beta}_C(k)$  更有效。

### 3. 单一插补约束岭估计

完整数据方法只利用了因变量和自变量都能观测的完整数据, 对于因变量缺失的观测数据被舍弃, 这样会造成信息的损失, 插补方法能够弥补这一问题, 下面我们使用单一插补方法。基于上一节得到的完整数据最小二乘估计  $\hat{\beta}_C$ , 定义如下新的因变量

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

显然, 如果  $Y_i$  存在缺失即  $\delta_i = 0$  时, 有  $Y_i^* = X_i^T \hat{\beta}_C$ ; 如果  $Y_i$  能观测到不存在缺失即  $\delta_i = 1$  时有  $Y_i^* = Y_i$ 。所以单一插补方法就是利用  $X_i^T \hat{\beta}_C$  插补缺失的因变量。

基于单一插补方法, 我们能够得到完整数据集  $(Y_i^*, X_i)_{i=1}^n$ , 因此有如下的线性回归模型

$$Y_i^* = X_i^T \beta + e_i, \quad i = 1, 2, \dots, n \quad (11)$$

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

从而可以得到  $\beta$  的单一插补最小二乘估计

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

和第 2 节完整数据方法下约束岭估计的构造一样, 定义如下辅助函数

$$F_2(\beta, \lambda) = (Y^* - X\beta)^T (Y^* - X\beta) + k\beta^T \beta + 2\lambda^T (A\beta - b)$$

其中  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)^T$  为  $k$  维 Lagrange 乘子。经过同第 2 节类似的计算, 我们可得回归系数的单一插补约束岭估计

$$\hat{\beta}_I^R(k) = \hat{\beta}_I(k) - (X^T X + kI_q)^{-1} A^T \left[ A(X^T X + kI_q)^{-1} A^T \right]^{-1} [A\hat{\beta}_I(k) - b] \quad (13)$$

其中  $\hat{\beta}_I(k) = (X^T X + kI_q)^{-1} X^T Y^*$  为单一插补岭估计。

下面给出  $\hat{\beta}_I^R(k)$  的渐近性质。

**定理 2:** 如果第 6 节的假设条件成立,

$$\sqrt{n}(\hat{\beta}_l^R(k) - \beta) \xrightarrow{d} N(0, \sigma^2(I_q - D)\Omega^{-1}(I_q - D)^T),$$

其中,  $\Omega = E(\delta XX^T)$ ,  $\Sigma = E(XX^T)$ ,  $D = \Sigma^{-1}A^T(A\Sigma^{-1}A^T)^{-1}A$ 。

#### 4. 数值模拟

对于前面提出的估计方法, 下面将通过模拟考察其在有限样本下的表现。类似于李静和安佰玲(2022) [5], 考虑如下线性回归模型

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (14)$$

根据 McDonald 和 Galerneau (1975) [8] 的方法, 自变量的产生满足

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

其中  $\omega_{ij}$  是独立的标准正态随机数,  $\rho$  是和自变量间相关性程度有关的参数, 值越大共线性程度越高, 我们分别取  $\rho = 0.90, 0.99, 0.999$ 。考虑因变量的缺失机制为

$$\Delta = p(\delta = 1 | x_1 = x_{1i}, x_2 = x_{2i}) = 0.8 + 0.2(|x_{1i}| + |x_{2i}|)$$

当  $|x_{1i}| + |x_{2i}| \leq 1$  时; 否则等于 0.9。回归系数  $(\beta_1, \beta_2)$  的真实值为 (1, 2), 约束条件为  $\beta_1 + \beta_2 = 3$ 。为了考察模型误差的分布对结果的影响,  $\varepsilon_i$  设定为服从如下的正态分布和均匀分布:

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

基于上面设定的缺失机制和误差分布, 样本量  $n$  分别取 80、120 和 150 三种情况, 针对每种设置重复 500 次, 计算  $(\beta_1, \beta_2)$  基于完整数据方法的估计(C)和岭估计(C-R)以及约束岭估计(C-RR), 单一插补方法的估计(I)和岭估计(I-R)以及约束岭估计(I-RR), 岭估计是基于 R 软件 MASS package 里的 lm.ridge 函数, 其中  $k$  的选取使用广义交叉验证法。定义估计量的均方误差(EMSE),

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

其中  $\beta_{kj}^*$  是参数  $\beta_j$  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 MSE

| $\rho$ | $\beta$ | $n = 80$ |         | $n = 120$ |         | $n = 150$ |         |
|--------|---------|----------|---------|-----------|---------|-----------|---------|
|        |         | Norma    | Uniform | Norma     | Uniform | Norma     | Uniform |
| 0.9    | C       | 0.02085  | 0.02018 | 0.01282   | 0.01431 | 0.01057   | 0.01118 |
|        | C-R     | 0.02127  | 0.02050 | 0.01294   | 0.01440 | 0.01075   | 0.01128 |
|        | C-RR    | 0.01905  | 0.01792 | 0.01146   | 0.01325 | 0.00946   | 0.01016 |
|        | I       | 0.01983  | 0.01944 | 0.0136    | 0.01229 | 0.01039   | 0.01161 |
|        | I-R     | 0.01995  | 0.01972 | 0.01375   | 0.01234 | 0.01045   | 0.01165 |
|        | I-RR    | 0.01786  | 0.01741 | 0.01231   | 0.01091 | 0.00922   | 0.01039 |
|        |         |          |         |           |         |           |         |
| 0.99   | C       | 0.20353  | 0.17547 | 0.11556   | 0.12122 | 0.09744   | 0.09247 |
|        | C-R     | 0.19755  | 0.18609 | 0.12523   | 0.12606 | 0.10663   | 0.10530 |
|        | C-RR    | 0.19574  | 0.18151 | 0.12177   | 0.12445 | 0.10550   | 0.10327 |

续表

|       |      |         |         |         |         |         |         |
|-------|------|---------|---------|---------|---------|---------|---------|
| 0.999 | I    | 0.17558 | 0.16661 | 0.12201 | 0.11196 | 0.09702 | 0.08874 |
|       | I-R  | 0.17829 | 0.16703 | 0.12538 | 0.12080 | 0.10572 | 0.09835 |
|       | I-RR | 0.17597 | 0.16248 | 0.12321 | 0.11850 | 0.10337 | 0.09665 |
|       | C    | 1.83416 | 1.73764 | 1.27806 | 1.25512 | 0.84510 | 0.92589 |
|       | C-R  | 0.91743 | 0.92136 | 0.73756 | 0.67787 | 0.53972 | 0.56895 |
|       | C-RR | 0.91347 | 0.90863 | 0.72577 | 0.66948 | 0.53600 | 0.56720 |
|       | I    | 1.74610 | 1.77163 | 1.04350 | 1.31280 | 1.03791 | 1.06673 |
|       | I-R  | 0.91200 | 0.96482 | 0.65461 | 0.82088 | 0.67332 | 0.68824 |
|       | I-RR | 0.89931 | 0.93615 | 0.65021 | 0.81719 | 0.67137 | 0.68773 |
|       |      |         |         |         |         |         |         |
|       |      |         |         |         |         |         |         |
|       |      |         |         |         |         |         |         |

从模拟结果可以看出：1) 随着样本量的增大，这些估计量的均方误差值大都在变小，与理论性质相一致。2) 随着  $\rho$  的增加，多重共线性越严重，完整方法和单一插补方法下岭估计表现都优于一般估计，约束岭估计的表现优于岭估计。3) 完整数据方法和单一插补方法相比，由于单一插补方法充分利用了数据信息，在  $\rho$  为0.9和0.99时单一插补的估计表现优于完整数据方法对应的估计。4) 误差分布对这几类估计的影响很小。

5. 总结

作为最为常用的统计方法之一，线性回归模型在分析实际问题时经常会遇到因变量缺失、自变量具有多重共线性和回归系数有附加的约束条件这三种情形，本文主要研究这三种情况都存在时回归系数的估计问题，分别构造了基于完整数据方法和单一插补方法下的约束岭估计，并研究了所提估计量的渐近性质，通过数值模拟考察了所提估计方法的有效性。研究结果为实际工作者提供了科学的统计分析方法。

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:  $\Sigma = E\left(XX^T\right)$  和  $\Omega = E\left(\delta\delta X X^T\right)$  为正定矩阵。

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

条件 3: 岭参数  $k=k_n$  , 满足  $\frac{k_n}{n} \xrightarrow{P} 0$  。

引理 1: 在上面的条件成立时，我们有

$$\frac{1}{n} \sum_{i=1}^n \delta_i X_i X_i^T \xrightarrow{P} \Omega, \frac{1}{n} \sum_{i=1}^n X_i X_i^T \xrightarrow{P} \Sigma.$$

证明: 该引理的证明可由大数定律直接得到。

引理 2: 如果第 6 节的假设条件成立，基于完整数据方法和单一插补方法的岭估计是渐近正态的，有

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

其中  $\hat{\beta}(k)$  为  $\hat{\beta}_c(k)$  或  $\hat{\beta}_l(k)$  。

证明: 该引理即为李静和安佰玲(2022) [5]中的定理 1 和定理 2。

定理 1: 证明: 由  $\hat{\beta}_c(k)$  的定义，

$$\hat{\beta}_C^R(k) = \hat{\beta}_C(k) - (\bar{X}^T \bar{X} + kI_q)^{-1} A^T \left[ A (\bar{X}^T \bar{X} + kI_q)^{-1} A^T \right]^{-1} [A \hat{\beta}_C(k) - b]$$

由条件 3 和引理 1 可得

$$\frac{1}{n} (X^T \Delta X + kI_q) \xrightarrow{P} \Omega$$

从而有

$$\left( \frac{\bar{X}^T \bar{X} + kI_q}{n} \right)^{-1} A^T \left[ A \left( \frac{\bar{X}^T \bar{X} + kI_q}{n} \right)^{-1} A^T \right]^{-1} A \xrightarrow{P} \Omega^{-1} A^T (A \Omega^{-1} A^T)^{-1} A$$

由引理 2 和多元正态分布理论可得

$$\sqrt{n} (\hat{\beta}_C^R(k) - \beta) \xrightarrow{d} N[0, \sigma^2 \Phi]$$

其中

$$\begin{aligned} \Phi &= \left[ I_q - \Omega^{-1} A^T (A \Omega^{-1} A^T)^{-1} A \right] \Omega^{-1} \left[ I_q - \Omega^{-1} A^T (A \Omega^{-1} A^T)^{-1} A \right]^T \\ &= \Omega^{-1} - \Omega^{-1} A^T (A \Omega^{-1} A^T)^{-1} A \Omega^{-1} = \Omega^{-1} - \nabla \end{aligned}$$

因此定理 1 成立。

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)。

## 参考文献

- [1] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons.
- [2] Cheng, P.E. (1994) 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 and Inference*, **48**, 85-99. [https://doi.org/10.1016/0378-3758\(94\)00151-k](https://doi.org/10.1016/0378-3758(94)00151-k)
- [4] 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]. 淮北煤炭师范学院学报(自然科学版), 2011, 32(1): 24-28.
- [5] 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.
- [6] 安佰玲, 王森, 胡洪胜. 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.
- [7] Rao, C.R., Toutenburg, H. and Shalabh, H.C. (2008) Linear Models and Generalization: Least Squares and Alternatives. Springer.
- [8] McDonald, G.C. and Galarneau, D.I. (1975) A Monte Carlo Evaluation of Some Ridge-Type Estimators. *Journal of the American Statistical Association*, **70**, 407-416. <https://doi.org/10.1080/01621459.1975.10479882>