

具有异方差的线性回归模型的统计诊断

凌佳^{1,2}, 夏乐天¹

(1. 河海大学理学院, 江苏 南京 210098; 2. 江苏广播电视大学职业技术学院, 江苏 南京 210036)

摘要 为了诊断具有异方差的线性回归模型的异常点, 建立了具有异方差的均值漂移模型和数据删除模型. 采用 Score 诊断统计量对具有异方差的均值漂移模型的均值是否漂移进行诊断, 证明了异方差存在条件下均值漂移模型和数据删除模型的等价性. 这一结果表明, 在诊断具有异方差的线性回归模型的异常点时, 可考虑采用更加便于处理的均值漂移模型. 最后, 用 Score 诊断统计量对镀锌数据进行了异常点的诊断.

关键词 均值漂移模型 数据删除模型 异方差

中图分类号 O212 **文献标识码** A **文章编号** 1000-1980(2008)02-0279-04

线性回归模型在实际问题中得到了广泛应用. 通常, 假设随机误差项是正态白噪声且具有方差齐性^[1]. 但在一些复杂问题中, 这种假设不太合理. 因此, 处理回归问题首先要进行方差齐性检验. 线性模型中关于误差的异方差性检验的 Score 检验统计量^[2]已由 Cook 等^[3]获得.

对于具有方差齐性的均值漂移模型和数据删除模型, 韦博成等^[1]给出了相应的估计量、诊断统计量以及二者的等价性证明. 对于具有异方差的线性回归模型^[4], 宗序平等^[5]得到了有关的估计量. 为了进一步推广上述结果, 本文首先讨论了具有异方差的均值漂移模型, 得到了均值扰动值的 Score 检验统计量, 然后讨论了具有异方差的数据删除模型, 证明了该模型与具有异方差的均值漂移模型的等价性, 最后通过镀锌数据^[6]说明了分析方法的合理性和有效性.

1 具有异方差的均值漂移模型

$$Y = X\beta + d_j\gamma + \varepsilon \quad \varepsilon \sim N(0, \sigma^2\Omega^{-1}) \quad (1)$$

式中: Y —— n 维观察向量, $Y = (y_1, \dots, y_n)^T$; X —— $n \times p$ 阶列满秩矩阵, $X = (x_1, \dots, x_n)^T$; β —— p 维未知参数, $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$; γ ——扰动值; d_i ——第 i 个分量为 1 其他为 0 的 n 维向量; ε —— n 维随机误差向量; $\Omega = I - (1 - \omega)d_jd_j^T (j \neq i)$; σ^2 ——未知参数.

记 $\theta = (\gamma, \beta^T, \sigma^2, \omega)^T$. 考虑如下假设检验问题:

$$H_0: \gamma = 0 \quad H_1: \gamma \neq 0 \quad (2)$$

显然, 若 H_0 成立, 则 (y_i, x_i^T) 不是异常点. 若 H_0 被否定, 则 (y_i, x_i^T) 为异常点.

定理 1 对于模型 (1), 假设检验问题 (2) 的 Score 检验统计量为

$$S_i(\omega) = \frac{[y_i - x_i^T \tilde{\beta}_i(\omega)]^2}{\tilde{\sigma}_i^2(\omega) [1 - x_i^T (X^T \Omega X)^{-1} x_i]} \quad (3)$$

且有
$$S_i(\omega) \xrightarrow{L} \chi^2(1)$$

证明 模型 (1) 的对数似然函数^[7]为

$$\begin{aligned} L(y, \theta) = & -\frac{n}{2} \lg(2\pi) - \frac{n}{2} \lg(\sigma^2) + \frac{1}{2} \lg \omega - \frac{1}{2\sigma^2} \sum_{k \neq i, j} (y_k - x_k^T \beta)^2 - \\ & \frac{1}{2\sigma^2} (y_i - x_i^T \beta - \gamma)^2 - \frac{\omega}{2\sigma^2} (y_j - x_j^T \beta)^2 \end{aligned}$$

由此可求出 L 关于 θ 的前二阶导数:

$$\begin{aligned}\frac{\partial^2 L}{\partial \gamma^2} &= -\frac{1}{\sigma^2} & \frac{\partial^2 L}{\partial \gamma \partial \boldsymbol{\beta}} &= -\frac{1}{\sigma^2} \mathbf{x}_i & \frac{\partial^2 L}{\partial \gamma \partial \sigma^2} &= -\frac{1}{\sigma^4} (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \gamma) & \frac{\partial^2 L}{\partial \gamma \partial \omega} &= 0 \\ \frac{\partial^2 L}{\partial \beta_a \partial \beta_b} &= -\frac{1}{\sigma^2} \left\{ \sum_{k \neq j} x_{ka} x_{kb} + \omega x_{ja} x_{jb} \right\} & \frac{\partial^2 L}{\partial \boldsymbol{\beta}^2} &= -\frac{1}{\sigma^2} \mathbf{X}^T \boldsymbol{\Omega} \mathbf{X} \\ \frac{\partial^2 L}{\partial \beta_a \partial \sigma^2} &= -\frac{1}{\sigma^4} \left\{ \sum_{k \neq i, j} (y_k - \mathbf{x}_k^T \boldsymbol{\beta}) x_{ka} + (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \gamma) x_{ia} + \omega (y_j - \mathbf{x}_j^T \boldsymbol{\beta}) x_{ja} \right\} \\ \frac{\partial^2 L}{\partial \beta_a \partial \omega} &= \frac{1}{\sigma^2} (y_j - \mathbf{x}_j^T \boldsymbol{\beta}) x_{ja} \\ \frac{\partial^2 L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{k \neq i, j} (y_k - \mathbf{x}_k^T \boldsymbol{\beta})^2 - \frac{1}{\sigma^6} (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \gamma)^2 - \frac{1}{\sigma^6} \omega (y_j - \mathbf{x}_j^T \boldsymbol{\beta})^2 \\ \frac{\partial^2 L}{\partial \sigma^2 \partial \omega} &= \frac{1}{2\sigma^4} (y_j - \mathbf{x}_j^T \boldsymbol{\beta}) & \frac{\partial^2 L}{\partial \omega^2} &= -\frac{1}{2\omega^2}\end{aligned}$$

由此可得

$$J = E(-\ddot{L}(\boldsymbol{\theta})) = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{1}{\sigma^2} \mathbf{x}_i^T & 0 & 0 \\ \frac{1}{\sigma^2} \mathbf{x}_i & \frac{1}{\sigma^2} \mathbf{X}^T \boldsymbol{\Omega} \mathbf{X} & 0 & 0 \\ 0 & 0 & \frac{n}{2\sigma^4} & -\frac{1}{2\omega\sigma^2} \\ 0 & 0 & -\frac{1}{2\omega\sigma^2} & \frac{1}{2\omega^2} \end{bmatrix}$$

$$J^{\gamma\gamma} = (J_{11} - J_{12} J_{22}^{-1} J_{21})^{-1} = \sigma^2 \{ \mathbf{I} - \mathbf{x}_i^T (\mathbf{X}^T \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{x}_i \}^{-1}$$

从而,由文献[8]中的 Score 检验函数公式有

$$S_i = \left\{ J^{\gamma\gamma} \left(\frac{\partial L}{\partial \gamma} \right)^2 \right\}_{\theta=(0, \tilde{\beta}_i(\omega), \tilde{\sigma}_i^2(\omega), \omega)} = \frac{[y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_i(\omega)]^2}{\tilde{\sigma}_i^2(\omega) [1 - \mathbf{x}_i^T (\mathbf{X}^T \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{x}_i]}$$

其中 $\tilde{\beta}_i(\omega)$, $\tilde{\sigma}_i^2(\omega)$ 分别为 H_0 成立时 $\boldsymbol{\beta}$ 和 σ^2 的极大似然估计(MLE).

由文献[9]知 $S_i(\omega) \xrightarrow{L} \chi^2(1)$. 证毕.

由文献[1]知 H_0 成立时, $\boldsymbol{\beta}$ 和 σ^2 的极大似然估计(MLE)分别为

$$\tilde{\boldsymbol{\beta}}_i(\omega) = \hat{\boldsymbol{\beta}} - \frac{1 - \omega}{1 - (1 - \omega) p_{ij}} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_j \hat{e}_j \quad (4)$$

$$\tilde{\sigma}_i^2(\omega) = \frac{R}{n} \left[1 - \frac{r_j^2}{n - p} (1 - p_{ij}) \frac{1 - \omega}{1 - (1 - \omega) p_{ij}} \right] \quad (5)$$

式中 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$; $p_{ij} = \mathbf{x}_j^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_j$, 即为 $\mathbf{X}$ 生成的投影阵 $\mathbf{P} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ 的对角元素; $\hat{e}_j = y_j - \hat{y}_j$, $\hat{y}_j = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}$; $R = \mathbf{Y}^T \mathbf{Q} \mathbf{Y}$, $\mathbf{Q} = \mathbf{I} - \mathbf{P}$; $r_j = \frac{\hat{e}_j}{\hat{\sigma} \sqrt{1 - p_{ij}}}$, $\hat{\sigma}^2 = \frac{R}{n - p}$.

推论 1 若 $\omega = 1$, 则表示方差无扰动, 由式(3)可得

$$S_i(1) = \frac{[y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_i(1)]^2}{\tilde{\sigma}_i^2(1) [1 - \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i]} = \frac{\hat{e}_i^2}{\hat{\sigma}^2(1 - p_{ii})} \quad (6)$$

式(6)的 $S_i(1)$ 实际上就是方差齐性时的均值漂移模型的均值扰动值的 Score 检验统计量, 这与已知结果吻合.

由文献[5]知模型(1)的参数 $\boldsymbol{\beta}$ 的最小二乘估计(LSE)为

$$\hat{\boldsymbol{\beta}}_{mi}(\omega) = (\mathbf{X}^T \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega} \mathbf{Y} - (\mathbf{X}^T \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{x}_i \hat{\gamma} \quad (7)$$

其中 $\hat{\gamma} = \frac{y_i - \mathbf{x}_i^T (\mathbf{X}^T \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega} \mathbf{Y}}{1 - \mathbf{x}_i^T (\mathbf{X}^T \boldsymbol{\Omega} \mathbf{X})^{-1} \mathbf{x}_i}$.

2 具有异方差的数据删除模型

$$Y(i) = X(i)\beta + \varepsilon(i) \quad \varepsilon(i) \sim N(0, \sigma^2 \Omega(i)^{-1}) \tag{8}$$

式中: $Y(i), X(i), \varepsilon(i)$ ——1 中 Y, X, ε 去掉第 i 行而得到的; $\Omega(i)$ ——1 中 Ω 去掉第 i 行和第 i 列而得到的; β —— p 维未知参数, $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$; σ^2 ——未知参数.

不妨用 $\hat{\beta}_{(i)}(\omega), \hat{\sigma}_{(i)}^2(\omega)$ 来表示模型(8)中的参数 β, σ^2 的 LSE.

定理 2 模型(1)和(8)相应的估计量与统计量都相同,即

$$\hat{\beta}_{mi}(\omega) = \hat{\beta}_{(i)}(\omega) \quad \hat{\sigma}_{i^2}^2(\omega) = \hat{\sigma}_{(i)}^2(\omega) \quad R_{mi} = R(i) \tag{9}$$

证明 首先引进 $X_0 = (0, \dots, 0, x_i, 0, \dots, 0)^T$ 表示第 i 行元素为 x_i^T 其余元素为 0 的一个 $n \times p$ 阶矩阵, $Y_0 = (0, \dots, 0, y_i, 0, \dots, 0)^T$ 表示第 i 行元素为 y_i 其余元素为 0 的一个 n 维向量.

此时模型(8)中的 $\Omega(i)$ 可表示为

$$\Omega(i) = A^T \Omega A \tag{10}$$

其中矩阵 A 是由在一个 $(n-1)$ 阶单位阵 I_{n-1} 的第 $(i-1)$ 行和第 i 行之间添加一行零元素而得到的一个 $n \times (n-1)$ 阶矩阵. 式(8)乘以 $\Omega^{\frac{1}{2}}(i)$ 可得

$$\Omega^{\frac{1}{2}}(i) Y(i) = \Omega^{\frac{1}{2}}(i) X(i) \beta + \Omega^{\frac{1}{2}}(i) \varepsilon(i)$$

且有
$$\Omega^{\frac{1}{2}}(i) \varepsilon(i) \sim N(0, \sigma^2 I_{n-1})$$

由文献[10]中的参数估计公式可得

$$\hat{\beta}_{(i)}(\omega) = [X(i)^T \Omega(i) X(i)]^{-1} X(i)^T \Omega(i) Y(i)$$

将式(10)代入上式可得

$$\begin{aligned} \hat{\beta}_{(i)}(\omega) &= [X(i)^T A^T \Omega A X(i)]^{-1} X(i)^T A^T \Omega A Y(i) = [(AX(i))^T \Omega AX(i)]^{-1} (AX(i))^T \Omega A Y(i) = \\ &= [(X - X_0)^T \Omega (X - X_0)]^{-1} (X - X_0)^T \Omega (Y - Y_0) = (X^T \Omega X - x_i x_i^T)^{-1} (X^T \Omega Y - x_i y_i) \end{aligned}$$

由文献[1]中的和式求逆公式有

$$(X^T \Omega X - x_i x_i^T)^{-1} = (X^T \Omega X)^{-1} + (X^T \Omega X)^{-1} x_i (1 - x_i^T (X^T \Omega X)^{-1} x_i)^{-1} x_i^T (X^T \Omega X)^{-1}$$

代入上式可得

$$\begin{aligned} \hat{\beta}_{(i)}(\omega) &= (X^T \Omega X)^{-1} X^T \Omega Y + (X^T \Omega X)^{-1} x_i y_i + \\ &= \frac{(X^T \Omega X)^{-1} x_i x_i^T (X^T \Omega X)^{-1} X^T \Omega Y - (X^T \Omega X)^{-1} x_i x_i^T (X^T \Omega X)^{-1} x_i y_i}{1 - x_i^T (X^T \Omega X)^{-1} x_i} \end{aligned}$$

上式化简整理后得

$$\hat{\beta}_{(i)}(\omega) = (X^T \Omega X)^{-1} X^T \Omega Y - (X^T \Omega X)^{-1} x_i \frac{y_i - x_i^T (X^T \Omega X)^{-1} X^T \Omega Y}{1 - x_i^T (X^T \Omega X)^{-1} x_i} = \hat{\beta}_{mi}(\omega)$$

$$R_{mi} = \|Y - X \hat{\beta}_{mi}(\omega) - \hat{\gamma} d_i\|^2 = \|Y(i) - X(i) \hat{\beta}_{mi}(\omega)\|^2 + 0 =$$

$$\|Y(i) - X(i) \hat{\beta}_{(i)}(\omega)\|^2 = R(i)$$

$$\hat{\sigma}_{mi}^2(\omega) = \frac{1}{n - (p + 1)} \|\Omega^{\frac{1}{2}} Y - \Omega^{\frac{1}{2}} X \hat{\beta}_{mi}(\omega) - \hat{\gamma} d_i\| =$$

$$\frac{1}{n - 1 - p} \|\Omega^{\frac{1}{2}}(i) Y(i) - \Omega^{\frac{1}{2}}(i) X(i) \hat{\beta}_{(i)}(\omega)\| = \hat{\sigma}_{(i)}^2(\omega)$$

证毕.

定理 2 的意义在于得到了在异方差的条件下,均值漂移模型和数据删除模型二者的等价性.

3 实例分析

由文献[1]知镀锌数据具有异方差,因此可运用模型(1)来进行异常点的诊断. 运用文献[1]中例 2.8 的结论可知这里 $\Omega = I - (1 - 0.109) d_9 d_9^T$. 利用式(3)计算假设检验(2)的 Score 值,见表 1.

由表 1 知 $S_{11}(\omega) = 6.5189 > \chi^2(1) = 3.841$, 所以第 11 号点可能有漂移. 事实上,由图 1 可以发现第 11 号点确实有漂移,这与计算结果吻合.

表 1 Score 检验统计量的值

Table 1 Score test value

点号	1	2	3	4	5	6	7	8	10	11
$S_i(\omega)$	0.078 49	0.040 36	1.521	0.539 25	0.851 52	0.618 48	2.252 5	0.193 3	0.250 52	6.518 9

4 结 语

由于异方差条件下均值漂移模型和数据删除模型二者的等价性,在讨论具有异方差的线性回归模型的异常点时,可考虑采用更便于处理的均值漂移模型,用定理 1 的 Score 检验统计量进行异常点的诊断.实例分析表明,采用该方法进行异常点诊断效果较好.

参考文献:

[1] 韦博成,鲁国斌,史建清.统计诊断引论[M].南京:东南大学出版社,1991:28-29.

[2] COX D R, HINKLEY D V. Theoretical statistics[M]. London: Chapman and Hall, 1974:324.

[3] COOK R D, WEISBERG S. Diagnostics for heteroscedasticity in regression[J]. Biometrika, 1983, 70(1):1-10.

[4] 王松桂,陈敏,陈立萍.线性统计模型[M].北京:高等教育出版社,1999:164-165.

[5] 宗序平,韦博成.线性回归诊断的若干问题[J].高校应用数学学报:A辑,1993,8(3):280-289.

[6] COOK R D, HOLSCHUH N, WEISBERG S. A note on an alternative outlier model[J]. J Roy Statist Soc, 1982, B(44):370-376.

[7] 陈希孺.高等数理统计学[M].合肥:中国科学技术大学出版社,1999:155-157.

[8] 韦博成.参数统计教程[M].北京:高等教育出版社,2006:279-281.

[9] 陈希孺.数理统计引论[M].北京:科学出版社,1981:327.

[10] WEISBERG S. Applied linear regression[M]. New York: John Wiley & Sons, 1985:42-45.

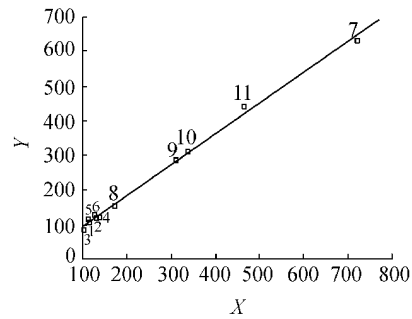


图 1 方差扰动时的散点图

Fig.1 Scatter plot under perturbation of variance

Statistical diagnostics for linear regression models with heteroscedasticity

LING Jia^{1,2}, XIA Le-tian¹

(1. College of Sciences, Hohai University, Nanjing 210098, China ;
2. Jiangsu Radio and TV University, Nanjing 210036, China)

Abstract: The mean shift outlier model (MSOM) and the case deletion model (CDM) with heteroscedasticity were developed in order to detect outliers in linear regression models with heteroscedasticity. The diagnostic test for the mean of the MSOM with heteroscedasticity was carried out based on the score statistic. It was proven that the estimates of MSOM and CDM with heteroscedasticity are equal. This result shows that the MSOM, which is easier to process, can be used more often to detect outliers in the linear regression model with heteroscedasticity than CDM. Finally, outliers in galvanization data were detected using the score statistic.

Key words: mean-shift outlier model ; case deletion model ; heteroscedasticity