第 44 卷第 3 期
2016 年 3 月

同 济 大 学 学 报(自 然 科 学 版)
JOURNAL OF TONGJI UNIVERSITY(NATURAL SCIENCE)

Vol. 44 No. 3
Mar. 2016

文章编号：0253-374X(2016)03-0446-08　　DOI：10.11908/j.issn.0253-374x.2016.03.017

顾及自变量与因变量误差及相关性的线性回归

王苗苗[1]，李博峰[1,2]，沈云中[1]

(1. 同济大学 测绘与地理信息学院，上海 200092；2. 国家地理信息工程国家重点实验室，西安 710054)

摘要：提出一种顾及自变量和因变量观测误差及误差相关性的线性回归新方法，并导出了求解线性回归系数的迭代公式. 以一元线性回归为例，导出了与最小二乘回归系数表达形式类似的解析解，并揭示了新方法与最小二乘方法的本质区别. 此外，对于含有多个自变量的多元线性回归，给出了相应的同时考虑自变量和因变量观测误差及误差相关性的回归系数求解方法. 试验表明，当自变量是非随机变量时，新方法与最小二乘方法的回归效果相同；当因变量和自变量都是随机变量(自变量与因变量的观测误差相关或不相关)时，新方法的回归系数比最小二乘方法的回归系数更加接近实际值.

关键词：回归分析；一元线性回归；相关系数；自变量误差

中图分类号：P207.1　　**文献标志码**：A

Linear Regression with Corrected Errors of Independent and Dependent Variables

WANG Miaomiao[1], *LI Bofeng*[1,2], *SHEN Yunzhong*[1]

(1. College of Surveying and Geo-Informatics, Tongji University, Shanghai, 200092, China; 2. State Key Laboratory of Geo-information Engineering, Xi'an, 710054, China)

Abstract: This paper presented a new linear regression method where the errors of dependent and independent variables and correlations of errors were adequately captured. The iteration formulae for calculating the regression parameters were derived at the same time. Taking univariate linear regression problem as an example, analytical formulas for linear regression parameters that similar to those from least-squares method were derived, with which the essential difference between least-squares method and new method were demonstrated. In addition, for the multiple linear regression that with multiple independent variables, the corresponding method, which considers the errors of both independent and dependent variables and the correlations of errors, for calculating the linear regression parameters were also shown. The experiment results shown that the new method and least-squares method were equivalent to each other when independent variables were non-random; whereas, the regressive parameters from new method were more closer to the true values than those from the least-squares method when both independent and dependent variables were all random (no matter their errors were correlated or not).

Key words: regression analysis; univariate linear regression; correlation coefficient; errors of independent variables

1　引言

变量之间的关系包括确定性的函数关系与非确定性的相关关系[1]. 回归分析是处理随机变量之间相关关系的数学工具[2-3]，其目的是找出因变量与自变量之间的统计关系，然后利用这种统计关系预测自变量值对应的因变量值，或者由给定的因变量值来控制自变量值的变化范围. 因此，回归分析的关键是根据自变量和因变量观测值及某些合理假设建立它们之间的(线性或者非线性)函数模型，即求解相应的回归系数[4]. 由于变量之间的非线性关系往往可通过一定的方法，例如变量变换转化为线性关系，因而线性回归是回归分析中最简单常用的回归模型.

线性回归中最简单的，最具代表性的是只有一个自变量的一元线性回归模型，为了叙述方便，记模型自变量和因变量分别为 x 和 y. 不失一般性，假设 x 和 y 的观测精度分别为 $\sigma_x=1$，$\sigma_y=3$，且观测误差的相关系数为 $\rho=-0.8$，则实际观测值的点位误差椭圆如图 1a 所示，即需要严格考虑自变量和因变量的误差特性，才能获得合理的回归系数. 目前计算回

收稿日期：2015-04-28

基金项目：国家自然科学基金(41374031;41574023)；国家地理信息工程国家重点实验室开放研究基金(SKLGIE2013-M-2-2)；测绘地理信息公益性行业科研专项经费资助(HY14122136)；中央高校基本科研业务费专项资金资助(20133080;20151225)

第一作者：王苗苗(1989—)，女，博士生，主要研究方向为测量数据处理和卫星导航系统理论与应用. E-mail：5wmmgps@tongji.edu.cn

归系数的方法都未能充分考虑自变量与因变量的观测误差以及它们的相关性. 文献[5-9]给出了同时考虑自变量误差 ε_x 与因变量误差 ε_y 的回归分析方法，但都忽略了 ε_x 与 ε_y 的相关性，即将原本按倾斜误差椭圆分布的误差(图 1a)按照与主轴平行的误差椭圆分布的误差处理，如图 1b 所示. 然而，传统的回归分析方法，如最小二乘方法不仅忽略了自变量与因变量观测误差的相关性，更甚至忽略了自变量的误差 ε_x，只考虑因变量的误差 ε_y，即用图 1c 的误差分布代替图 1a 的点位误差椭圆，显然这样的处理方式是不合理的.

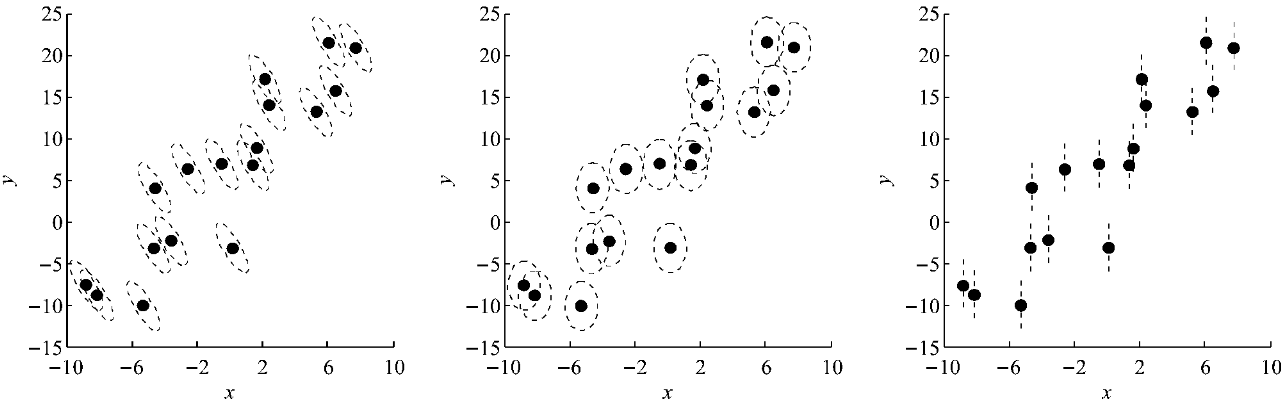

图 1　三种自变量与因变量误差处理方式

Fig.1　Three strategies used to process errors in independent and dependent variables

综上所述，回归分析中的自变量和因变量观测值都来自实际观测，不可避免地存在观测误差，甚至是具有相关性的观测误差. 回归分析时如果按照只考虑因变量误差，或者只考虑自变量和因变量误差但忽略误差相关性的方式处理观测数据，都必然导致获得的回归系数不合理，即建立的回归模型不合理，影响回归分析的效果及其应用. 文献[10-11]中的变量随机模型可以用来描述自变量与因变量的观测误差及误差之间的相关关系. 本文从线性回归分析的角度，以一元线性回归为例，分析自变量和因变量观测误差以及误差相关性对回归系数求解的影响. 不同于文献[5-9]，本文给出了回归系数的解析解形式，并指出不同的回归分析方法都采用误差改正的“新观测值”代替原始观测值求解回归系数.

提出一种同时顾及自变量和因变量观测误差及误差相关性的线性回归新方法，并研究了新方法求解回归系数的迭代方式，导出了新方法获得的回归系数的解析形式，揭示了回归分析新方法的广泛性及其与最小二乘方法的区别，最后，采用实例验证了新方法的回归效果.

2　传统一元线性回归

一元线性回归方程为

$$y=\beta_0+\beta_1 x \tag{1}$$

它描述了因变量 y 随自变量 x 的总体变化情况. 通常采用 $m>2$ 组观测数据确定回归模型系数 β_0 和 β_1，对应的回归模型为

$$\boldsymbol{y}=([\boldsymbol{e}_m\quad \boldsymbol{x}])\boldsymbol{\beta}+\boldsymbol{\varepsilon}_y \tag{2}$$

式中：$\boldsymbol{y}=[y_1,\cdots,y_m]^{\mathrm{T}}$ 为因变量观测向量；$\boldsymbol{\varepsilon}_y$ 是相应的观测误差(服从正态分布)；$\boldsymbol{x}=[x_1,\cdots,x_m]^{\mathrm{T}}$ 为自变量观测向量；$\boldsymbol{e}_m$ 表示元素都是 1 的 m 维列向量；$\boldsymbol{\beta}=[\beta_0\quad \beta_1]^{\mathrm{T}}$ 为回归系数向量. 理论上，如果变量 x 和 y 之间存在严格的函数关系，则观测数据可最佳地拟合回归方程，所有实测点$[x_i,y_i]$都会落在对应的回归直线上. 然而，由于变量观测误差的影响，实测点并不完全落在回归直线上. 传统地，采用最小二乘准则 $\sum_{i=1}^m w_i\varepsilon_{y_i}^2$ 取值为最小，求解回归系数估值 $\hat\beta_0$ 和 $\hat\beta_1$，有：

$$\hat\beta_1=\frac{\sum_{i=1}^m w_i x_i y_i-\sum_{i=1}^m w_i x_i\bar y_w}{\sum_{i=1}^m w_i x_i^2-\sum_{i=1}^m w_i x_i\bar x_w}=\frac{\sum_{i=1}^m w_i(x_i-\bar x_w)(y_i-\bar y_w)}{\sum_{i=1}^m w_i(x_i-\bar x_w)^2} \tag{3a}$$

$$\hat\beta_0=\bar y_w-\hat\beta_1\bar x_w \tag{3b}$$

式中：w_i 为观测值 y_i 的权重；$\bar x_w=\sum_{i=1}^m w_i x_i\Big/\sum_{i=1}^m w_i$，$\bar y_w=\sum_{i=1}^m w_i y_i\Big/\sum_{i=1}^m w_i$ 分

别为自变量和因变量观测值的加权平均值. 式(3)表明,回归直线过定点$[\bar{x}_w,\bar{y}_w]$. 若因变量观测值 $\boldsymbol{y}$ 等精度观测,则回归系数为[12]

$$\hat{\beta}_1=\frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{m}(x_i-\bar{x})^2} \tag{4a}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x} \tag{4b}$$

式中,$\bar{x}=\sum_{i=1}^{m}x_i/m,\bar{y}=\sum_{i=1}^{m}y_i/m$ 分别为自变量和因变量观测值的平均值. 此时回归直线过定点$[\bar{x},\bar{y}]$. 自变量和因变量观测值的方差 s_x^2,s_y^2 以及它们的协方差 s_{xy} 分别为 $s_x^2=\sum_{i=1}^{m}(x_i-\bar{x})^2/m,s_y^2=\sum_{i=1}^{m}(y_i-\bar{y})^2/m,s_{xy}=\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})/m$. 自变量与因变量观测值的线性相关系数[13-14]为 $\gamma=\frac{s_{xy}}{s_x\cdot s_y}$. 回归系数 $\hat{\beta}_1$ 可表示为

$$\hat{\beta}_1=\gamma\frac{s_y}{s_x} \tag{5}$$

3 顾及自变量与因变量误差及误差相关性的线性回归

自变量和因变量观测值通常都来源于实际观测,都不可避免地被观测误差 ε_y 和 ε_x 污染,上述一元线性回归模型的传统最小二乘解法只考虑了因变量 y 的观测误差而忽略了自变量 x 的观测误差. 当 ε_y 和 ε_x 之间存在相关性,即 $\sigma_{xy}\neq0$,如图1a所示,除了考虑 ε_y 和 ε_x,还应该考虑误差的相关性 σ_{xy}. 因而,需要一种充分考虑变量观测误差及误差相关性的线性回归方法. 将一元线性回归模型(2)改为

$$\boldsymbol{y}=([\boldsymbol{e}_m\boldsymbol{x}-\boldsymbol{\varepsilon}_x])\boldsymbol{\beta}+\boldsymbol{\varepsilon}_y \tag{6}$$

误差 $\boldsymbol{\varepsilon}_x$ 与 $\boldsymbol{\varepsilon}_y$ 之间的相关程度为 $\rho=\sigma_{xy}/(\sigma_x\cdot\sigma_y)$. 假设自变量与因变量各自等精度观测,则类似于文献[10-11],模型(6)对应的随机模型为

$$\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}\sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\sigma_y^2 & \sigma_{yx}\\ \sigma_{xy} & \sigma_x^2\end{bmatrix}\otimes\boldsymbol{I}_m\right) \tag{7}$$

其中,$\otimes$表示克罗内克积[15-16]. 当 $\rho=0$ 时,随机模型(7)与文献[6-9]中考虑自变量观测误差的加权总体最小二乘的随机模型一致;当 $\boldsymbol{\varepsilon}_x=\boldsymbol{0}$ 时,模型(6)等价于模型(2),即传统的最小二乘方法与加权总体最小二乘方法是新方法的一种特例. 因此,为了说明自变量和因变量观测误差以及误差相关性对回归分析效果的影响,下文只分析传统的最小二乘方法与新方法的差异之处与共同之处.

顾及变量误差及误差相关性的回归分析方法需要同时考虑 σ_x^2,σ_y^2 及 σ_{xy},即在求解回归系数时需要综合考虑变量的随机特性. 联合方程(6)和(7),采用最小二乘准则 $M=\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}^{\mathrm{T}}\left(\begin{bmatrix}\sigma_y^2 & \sigma_{yx}\\ \sigma_{xy} & \sigma_x^2\end{bmatrix}\otimes\boldsymbol{I}_m\right)^{-1}\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}$为最小,利用欧拉-拉格朗日条件极值法求解回归系数 $\boldsymbol{\beta}$,相应的拉格朗日条件函数为

$$\Psi=M+2\boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\varepsilon}_y-\beta_0\boldsymbol{e}_m-\beta_1\boldsymbol{x}+\beta_1\boldsymbol{\varepsilon}_x) \tag{8}$$

式中,拉格朗日乘常数 $\boldsymbol{\lambda}$ 是 $m\times1$ 的未知向量. 对各未知量求偏导数并令其等于零有:

$$\frac{1}{2}\frac{\partial\Psi}{\partial\boldsymbol{\varepsilon}_y}=\tilde{\boldsymbol{\varepsilon}}_y^{\mathrm{T}}\sigma_y^{-2}-\sigma_{xy}\sigma_y^{-2}\tilde{\boldsymbol{\varepsilon}}_{x|y}^{\mathrm{T}}\boldsymbol{\Sigma}_{x|y}^{-1}-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}=\boldsymbol{0} \tag{9a}$$

$$\frac{1}{2}\frac{\partial\Psi}{\partial\boldsymbol{\varepsilon}_x}=\tilde{\boldsymbol{\varepsilon}}_{x|y}^{\mathrm{T}}\boldsymbol{\Sigma}_{x|y}^{-1}+\hat{\beta}_1\hat{\boldsymbol{\lambda}}^{\mathrm{T}}=\boldsymbol{0} \tag{9b}$$

$$\frac{1}{2}\frac{\partial\Psi}{\partial\boldsymbol{\beta}}=-\hat{\boldsymbol{\lambda}}^{\mathrm{T}}([\boldsymbol{e}_m\quad \boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x])=\boldsymbol{0} \tag{9c}$$

$$\frac{1}{2}\frac{\partial\Psi}{\partial\boldsymbol{\lambda}}=\boldsymbol{y}-\tilde{\boldsymbol{\varepsilon}}_y-\hat{\boldsymbol{\beta}}_0\boldsymbol{e}-\hat{\beta}_1\boldsymbol{x}+\hat{\beta}_1\tilde{\boldsymbol{\varepsilon}}_x=\boldsymbol{0} \tag{9d}$$

式中,向量 $\boldsymbol{\varepsilon}_{x|y}=\boldsymbol{\varepsilon}_x-(\sigma_{xy}\sigma_y^{-2}\otimes\boldsymbol{I}_m)\boldsymbol{\varepsilon}_y$,其方差阵为 $\boldsymbol{\Sigma}_{x|y}=(\sigma_x^2-\sigma_{xy}^2\sigma_y^{-2})\otimes\boldsymbol{I}_m$. 符号"^"表示未知非随机量的估值,符号"~"表示未知随机量的预测值. 对 $\boldsymbol{\varepsilon}_x$ 和 $\boldsymbol{\varepsilon}_y$ 求二阶偏导数有 Hessian 矩阵如下:

$$\frac{1}{2}\frac{\partial^2\Psi}{\partial\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}[\boldsymbol{\varepsilon}_y^{\mathrm{T}}\quad\boldsymbol{\varepsilon}_x^{\mathrm{T}}]}=\begin{bmatrix}(\sigma_y^{-2}\otimes\boldsymbol{I}_m)+(\sigma_y^{-4}\sigma_{xy}^2\otimes\boldsymbol{I}_m)\boldsymbol{\Sigma}_{x|y}^{-1} & -(\sigma_y^{-2}\sigma_{xy}\otimes\boldsymbol{I}_m)\boldsymbol{\Sigma}_{x|y}^{-1}\\ -\boldsymbol{\Sigma}_{x|y}^{-1}(\sigma_{xy}\sigma_y^{-2}\otimes\boldsymbol{I}_m) & \boldsymbol{\Sigma}_{x|y}^{-1}\end{bmatrix}$$

易证该 Hessian 矩阵是非负定矩阵. 因此方程9a—9d的解即是满足目标方程(8)的最优解[10]. 联合求解方程9a—9d,得:

$$\hat{\boldsymbol{\lambda}}=\tilde{\boldsymbol{\Sigma}}_C^{-1}(\boldsymbol{y}-\hat{\beta}_0\boldsymbol{e}-\hat{\beta}_1\boldsymbol{x}) \tag{10}$$

$$\tilde{\boldsymbol{\varepsilon}}_y=(\sigma_y^2-\sigma_{xy}\hat{\beta}_1)\hat{\boldsymbol{\lambda}} \tag{11a}$$

$$\tilde{\boldsymbol{\varepsilon}}_x=(\sigma_{xy}-\sigma_x^2\hat{\beta}_1)\hat{\boldsymbol{\lambda}} \tag{11b}$$

$$([\boldsymbol{e}_m\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x])^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_C^{-1}(\boldsymbol{y}-\hat{\beta}_0\boldsymbol{e}_m-\hat{\beta}_1\boldsymbol{x})=0 \tag{12}$$

其中,矩阵 $\tilde{\boldsymbol{\Sigma}}_C=(\hat{\beta}_1^2\sigma_x^2-2\hat{\beta}_1\sigma_x+\sigma_y^2)\otimes\boldsymbol{I}_m$. 显然,$\tilde{\boldsymbol{\varepsilon}}_x$ 与 $\tilde{\boldsymbol{\Sigma}}_C$ 都与回归系数的估值 $\hat{\boldsymbol{\beta}}$ 有关,故方程(12)是一个非线性方程,需要迭代求解回归系数的估值 $\hat{\beta}_0$ 与 $\hat{\beta}_1$. 将方程(12)改写成

$$([\boldsymbol{e}_m\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x])^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_C^{-1}(\hat{\beta}_0\boldsymbol{e}_m+\hat{\beta}_1\boldsymbol{x})=([\boldsymbol{e}_m\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x])^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_C^{-1}\boldsymbol{y} \tag{13}$$

两边减去$([\boldsymbol{e}_m\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x])^{\mathrm{T}}\tilde{\boldsymbol{\Sigma}}_C^{-1}\tilde{\boldsymbol{\varepsilon}}_x\hat{\beta}_1$,则法方程为

$$\begin{bmatrix}\boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{e}_m & \boldsymbol{e}_m^{\mathrm{T}}(\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x)\\ (\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x)^{\mathrm{T}}\boldsymbol{e}_m & (\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x)^{\mathrm{T}}(\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x)\end{bmatrix}\hat{\boldsymbol{\beta}}=$$

$$\begin{bmatrix} \boldsymbol{e}_m^{\mathrm{T}}(\boldsymbol{y}-\tilde{\boldsymbol{\varepsilon}}_x\hat{\beta}_1) \\ (\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x)^{\mathrm{T}}(\boldsymbol{y}-\tilde{\boldsymbol{\varepsilon}}_x\hat{\beta}_1) \end{bmatrix} \tag{14}$$

即

$$\begin{bmatrix} m & \sum_{i=1}^m(x_i-\tilde{\varepsilon}_{x_i}) \\ \sum_{i=1}^m(x_i-\tilde{\varepsilon}_{x_i}) & \sum_{i=1}^m(x_i-\tilde{\varepsilon}_{x_i})^2 \end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}= \begin{bmatrix} \sum_{i=1}^m(y_i-\tilde{\varepsilon}_{x_i}\hat{\beta}_1) \\ \sum_{i=1}^m(x_i-\tilde{\varepsilon}_{x_i})(y_i-\tilde{\varepsilon}_{x_i}\hat{\beta}_1) \end{bmatrix} \tag{15}$$

记 $\tilde{x}_i=x_i-\tilde{\varepsilon}_{x_i}$,$\tilde{y}_i=y_i-\tilde{\varepsilon}_{x_i}\hat{\beta}_1$,则回归系数的估值 $\hat{\boldsymbol{\beta}}$ 及其协方差阵 $\boldsymbol{Q}_{\hat{\beta}\hat{\beta}}$ 为

$$\hat{\boldsymbol{\beta}}=\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}=\begin{bmatrix} m & \sum_{i=1}^m\tilde{x}_i \\ \sum_{i=1}^m\tilde{x}_i & \sum_{i=1}^m\tilde{x}_i \end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^m\tilde{y}_i \\ \sum_{i=1}^m\tilde{x}_i\tilde{y}_i\end{bmatrix}= \frac{1}{m\sum_{i=1}^m\tilde{x}_i^2-\left(\sum_{i=1}^m\tilde{x}_i\right)^2}\cdot \begin{bmatrix} \sum_{i=1}^m\tilde{x}_i\sum_{i=1}^m\tilde{y}_i-\sum_{i=1}^m\tilde{x}_i\sum_{i=1}^m\tilde{x}_i\tilde{y}_i \\ -\sum_{i=1}^m\tilde{x}_i\sum_{i=1}^m\tilde{y}_i+m\sum_{i=1}^m\tilde{x}_i\tilde{y}_i \end{bmatrix} \tag{16}$$

$$\boldsymbol{Q}_{\hat{\beta}\hat{\beta}}=\begin{bmatrix} m & \sum_{i=1}^m\tilde{x}_i \\ \sum_{i=1}^m\tilde{x}_i & \sum_{i=1}^m\tilde{x}_i \end{bmatrix}^{-1}= \frac{1}{m\sum_{i=1}^m\tilde{x}_i^2-\left(\sum_{i=1}^m\tilde{x}_i\right)^2}\cdot \begin{bmatrix} \sum_{i=1}^m\tilde{x}_i & -\sum_{i=1}^m\tilde{x}_i \\ -\sum_{i=1}^m\tilde{x}_i & m \end{bmatrix}$$

回归系数 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的解还可分别表示为

$$\hat{\beta}_1=\frac{1}{m\sum_{i=1}^m\tilde{x}_i^2-\left(\sum_{i=1}^m\tilde{x}_i\right)^2}\left(-\sum_{i=1}^m\tilde{x}_i\sum_{i=1}^m\tilde{y}_i+ m\sum_{i=1}^m\tilde{x}_i\tilde{y}_i\right)=\frac{\sum_{i=1}^m(\tilde{x}_i-\bar{\tilde{x}})(\tilde{y}_i-\bar{\tilde{y}})}{\sum_{i=1}^m(\tilde{x}_i-\bar{\tilde{x}})^2} \tag{17a}$$

$$\hat{\beta}_0=\frac{1}{m\sum_{i=1}^m\tilde{x}_i-\left(\sum_{i=1}^m\tilde{x}_i\right)^2}\left(\sum_{i=1}^m\tilde{x}_i\sum_{i=1}^m\tilde{y}_i- \sum_{i=1}^m\tilde{x}_i\sum_{i=1}^m\tilde{x}_i\tilde{y}_i\right)=\bar{\tilde{y}}-\hat{\beta}_1\bar{\tilde{x}} \tag{17b}$$

其中,$\bar{\tilde{x}}=\sum_{i=1}^m\tilde{x}_i/m$,$\bar{\tilde{y}}=\sum_{i=1}^m\tilde{y}_i/m$.式(17)表明回归直线过定点$[\bar{\tilde{x}},\bar{\tilde{y}}]$.与传统最小二乘方法获得的回归系数解析形式相比,新方法的回归系数解式(17)与式(4)完全类似,差别在于传统方法直接采用观测值 x_i 和 y_i,而新方法采用了误差改正后的新“观测值”$\tilde{x}_i$ 与 $\tilde{y}_i$.

类似于 s_x^2,s_y^2 及 s_{xy},新"观测值"的方差、协方差为 $s_{\tilde{x}}^2=\sum_{i=1}^m(\tilde{x}_i-\bar{\tilde{x}})^2/m$,$s_{\tilde{y}}^2=\sum_{i=1}^m(\tilde{y}_i-\bar{\tilde{y}})^2/m$,$s_{\tilde{x}\tilde{y}}=\sum_{i=1}^m(\tilde{x}_i-\bar{\tilde{x}})(\tilde{y}_i-\bar{\tilde{y}})/m$,线性相关系数为 $\tilde{\gamma}=\frac{s_{\tilde{x}\tilde{y}}}{s_{\tilde{x}}\cdot s_{\tilde{y}}}$,则式(17a)中的回归系数 $\hat{\beta}_1$ 还可表示为

$$\hat{\beta}_1=\tilde{\gamma}\frac{s_{\tilde{y}}}{s_{\tilde{x}}} \tag{18}$$

利用新方法求解回归系数,相应的自变量和因变量观测误差的估值 $\tilde{\varepsilon}_x$,$\tilde{\varepsilon}_y$ 分别为

$$\tilde{\boldsymbol{\varepsilon}}_x=(\rho\cdot\sigma_x\cdot\sigma_y-\sigma_x^2\cdot\hat{\beta}_1)\tilde{\boldsymbol{\Sigma}}_C^{-1}(\boldsymbol{y}-\hat{\beta}_0\boldsymbol{e}_m-\hat{\beta}_1\boldsymbol{x}) \tag{19a}$$

$$\tilde{\boldsymbol{\varepsilon}}_y=(\sigma_y^2-\rho\cdot\sigma_x\cdot\sigma_y\cdot\hat{\beta}_1)\tilde{\boldsymbol{\Sigma}}_C^{-1}(\boldsymbol{y}-\hat{\beta}_0\boldsymbol{e}_m-\hat{\beta}_1\boldsymbol{x}) \tag{19b}$$

有 $\tilde{\boldsymbol{x}}$ 恒等于 $\boldsymbol{x}-\tilde{\boldsymbol{\varepsilon}}_x$.由于 $\tilde{\boldsymbol{\varepsilon}}_y-\tilde{\boldsymbol{\varepsilon}}_x\hat{\beta}_1=((\sigma_y-\sigma_x\cdot\hat{\beta}_1)^2+2(1-\rho)\sigma_x\cdot\sigma_y\cdot\hat{\beta}_1)\tilde{\boldsymbol{\Sigma}}_C^{-1}(\boldsymbol{y}-\hat{\beta}_0\boldsymbol{e}_m-\hat{\beta}_1\boldsymbol{x})$不恒等于 $\boldsymbol{0}$,因而 $\tilde{\boldsymbol{y}}=\boldsymbol{y}-\tilde{\boldsymbol{\varepsilon}}_x\hat{\beta}_1$ 不恒等于 $\boldsymbol{y}-\tilde{\boldsymbol{\varepsilon}}_y$,即 $\tilde{\boldsymbol{y}}$ 只是观测向量 $\boldsymbol{y}$ 将部分误差改正后的新“观测”向量.

存在极端情况 $\rho=\pm1$ 时,$\begin{bmatrix}\sigma_y^2 & \sigma_{yx}\\ \sigma_{xy} & \sigma_x^2\end{bmatrix}$是秩亏矩阵,此时自变量和因变量观测误差完全相关,即它们的观测信息可由任一变量完全体现.此时,$\boldsymbol{\Sigma}_{x|y}\equiv0$,即二阶偏导数$\partial^2\Psi/\partial\boldsymbol{\varepsilon}_x^2$ 不存在,Hessian 矩阵零定.

对于含有多个自变量的多元线性回归模型,类似于模型(6),有

$$\boldsymbol{y}-\boldsymbol{\varepsilon}_y=[\boldsymbol{e}_m,\boldsymbol{X}-\boldsymbol{E}_X]\boldsymbol{\beta} \tag{20}$$

式中:$\boldsymbol{X}=[\boldsymbol{x}_1^{\mathrm{T}},\cdots,\boldsymbol{x}_m^{\mathrm{T}}]^{\mathrm{T}}$ 为 n 个自变量的 m 次观测值矩阵,其第 j 次观测为 $\boldsymbol{x}_j=[x_{j,1},\cdots,x_{j,n}]$;$\boldsymbol{E}_X=[\boldsymbol{\varepsilon}_{x_1}^{\mathrm{T}},\cdots,\boldsymbol{\varepsilon}_{x_m}^{\mathrm{T}}]^{\mathrm{T}}$ 为自变量观测值的观测误差矩阵;$\boldsymbol{\beta}=[\beta_0,\beta_1,\cdots,\beta_n]^{\mathrm{T}}$ 为未知的线性回归系数.将模型(20)展开有

$$\boldsymbol{y}-\boldsymbol{\varepsilon}_y=(\boldsymbol{A}-\boldsymbol{E}_A)\boldsymbol{\beta}=(\boldsymbol{A}-\boldsymbol{E}_X\boldsymbol{H})\boldsymbol{\beta} \tag{21}$$

其中 $\boldsymbol{A}=[\boldsymbol{e}_m,\boldsymbol{X}]$,$\boldsymbol{E}_A=[\boldsymbol{0},\boldsymbol{E}_X]$,$\boldsymbol{H}=[\boldsymbol{0}_{n\times1},\boldsymbol{I}_n]$.模型(20)对应的随机模型为

$$\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}=\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \mathrm{vec}(\boldsymbol{E}_X)\end{bmatrix}\sim N\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\boldsymbol{Q}_{yy} & \boldsymbol{Q}_{yx}\\ \boldsymbol{Q}_{xy} & \boldsymbol{Q}_{xx}\end{bmatrix}\right) \tag{22}$$

式中:运算符“vec(·)”表示将矩阵前后各列叠加获得矩阵的列向量化形式;$\boldsymbol{Q}_{yy}$ 和 $\boldsymbol{Q}_{xx}$ 分别表示因变量和自变量观测值的方差阵;$\boldsymbol{Q}_{xy}$ 和 $\boldsymbol{Q}_{yx}$ 为观测值的协方差阵.以最小二乘准则 $\Phi=\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{Q}_{yy} & \boldsymbol{Q}_{yx}\\ \boldsymbol{Q}_{xy} & \boldsymbol{Q}_{xx}\end{bmatrix}^{-1}\cdot\begin{bmatrix}\boldsymbol{\varepsilon}_y\\ \boldsymbol{\varepsilon}_x\end{bmatrix}$为最小求解回归系数,类似于式(8),拉格朗日

条件函数为

$$\Psi = \Phi + 2\boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\varepsilon}_y - \boldsymbol{A\beta} + (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}} \otimes \boldsymbol{I}_m)\boldsymbol{\varepsilon}_x) \tag{23}$$

按照式(9)—(13)的推导方法，回归模型(20)的回归系数估值 $\hat{\boldsymbol{\beta}}$ 对应的法方程为[17]

$$\widetilde{\boldsymbol{A}}^{\mathrm{T}}\widetilde{\boldsymbol{Q}}_{ll}^{-1}\widetilde{\boldsymbol{A}}\hat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{A}}^{\mathrm{T}}\widetilde{\boldsymbol{Q}}_{ll}^{-1}(\boldsymbol{y} - \widetilde{\boldsymbol{E}}_A\hat{\boldsymbol{\beta}}) \tag{24}$$

其中，$\widetilde{\boldsymbol{A}}=\boldsymbol{A}-\widetilde{\boldsymbol{E}}_A$，$\widetilde{\boldsymbol{Q}}_{ll}=(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\otimes\boldsymbol{I}_m)\boldsymbol{Q}_{xx}(\boldsymbol{H}\hat{\boldsymbol{\beta}}\otimes\boldsymbol{I}_m)-(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\otimes\boldsymbol{I}_m)\boldsymbol{Q}_{xy}-\boldsymbol{Q}_{yx}(\boldsymbol{H}\hat{\boldsymbol{\beta}}\otimes\boldsymbol{I}_m)+\boldsymbol{Q}_{yy}$，回归系数的估值为

$$\hat{\beta} = (\widetilde{\boldsymbol{A}}^{\mathrm{T}}\widetilde{\boldsymbol{Q}}_{ll}^{-1}\widetilde{\boldsymbol{A}})^{-1}\widetilde{\boldsymbol{A}}^{\mathrm{T}}\widetilde{\boldsymbol{Q}}_{ll}^{-1}(\boldsymbol{y} - \widetilde{\boldsymbol{E}}_A\hat{\boldsymbol{\beta}}) \tag{25}$$

式(25)是非线性的，需要迭代求解，步骤见文献[17]. 观测误差估值 $\tilde{\varepsilon}_x$，$\tilde{\varepsilon}_y$ 分别为

$$\tilde{\varepsilon}_x = [\boldsymbol{Q}_{xy} - \boldsymbol{Q}_{xx}(\boldsymbol{H}\hat{\boldsymbol{\beta}} \otimes \boldsymbol{I}_m)]\widetilde{\boldsymbol{Q}}_{ll}^{-1}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) \tag{26a}$$

$$\tilde{\varepsilon}_y = [\boldsymbol{Q}_{yy} - \boldsymbol{Q}_{yx}(\boldsymbol{H}\hat{\boldsymbol{\beta}} \otimes \boldsymbol{I}_m)]\widetilde{\boldsymbol{Q}}_{ll}^{-1}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) \tag{26b}$$

若回归模型只有一个自变量，模型(20)等价于模型(6)，相应的法方程(24)等价于法方程(14)，回归系数的估值式(25)等价于式(16)，自变量和因变量观测误差的估值式(26)等价于式(19). 因此，同时考虑自变量和因变量观测误差及误差相关性的线性回归思想在一元线性回归和多元线性回归中都是适用的，即本文对观测误差及其特性的处理方法具有广泛性.

4 回归方法比较

对于回归方程 $y=\beta_0+\beta_1 x$，分别等精度观测自变量和因变量. 传统最小二乘方法求解的回归系数为

$$\hat{\beta}_1^{LS} = \gamma\frac{s_y}{s_x},\quad \hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1\bar{x}$$

新方法求解的回归系数为

$$\hat{\beta}_1^{S} = \tilde{\gamma}\frac{s_{\tilde{y}}}{s_{\tilde{x}}},\quad \hat{\beta}_0^{S} = \bar{\tilde{y}} - \hat{\beta}_1\bar{\tilde{x}}$$

两种方法求得的回归系数的解析形式是相同的，不同的是 $\tilde{\gamma}$，$s_{\tilde{x}}$，$s_{\tilde{y}}$ 和 $\bar{\tilde{x}}$，$\bar{\tilde{y}}$ 包含了 σ_x 和 ρ 的信息. 两种回归分析方法有本质区别，它们获得的回归模型是不同的.

为了比较顾及变量误差及误差相关性的新回归方法与传统最小二乘回归方法在一元线性回归分析中的效果，以及两个方法之间的差异与共性，设计模拟实验：假设自变量和因变量各自按照给定的精度 σ_x 和 σ_y 进行等精度观测，给变量观测误差不同的相关性，即改变 σ_{xy}，按照直线 $y=2x+5$ 分别模拟 m 组观测数据$[x_i, y_i]$. 按照以下2种情形进行讨论：

情形1 自变量观测值 $\boldsymbol{x}$ 没有观测误差，因变量观测值 $\boldsymbol{y}$ 有观测误差，即 $\sigma_x=0$，$\sigma_y\neq 0$. 此时，“观测值”$\tilde{x}_i=x_i$，$\tilde{y}_i=y_i$，均值 $\bar{\tilde{x}}=\bar{x}$，$\bar{\tilde{y}}=\bar{y}$，方差及协方差 $s_x=s_{\tilde{x}}$，$s_y=s_{\tilde{y}}$，$s_{xy}=s_{\tilde{x}\tilde{y}}$，相关系数 $\gamma=\tilde{\gamma}$，因此，$\hat{\beta}_1^{LS}=\hat{\beta}_1^{S}$，$\hat{\beta}_0^{LS}=\hat{\beta}_0^{S}$，即新方法和传统最小二乘方法是一致的. 此外，由于误差相关性 σ_{xy} 无意义，因而包括加权总体最小二乘方法在内的三种方法是一致的(见表1第6行).

情形2 自变量与因变量观测值 $\boldsymbol{x}$，$\boldsymbol{y}$ 都有误差且误差相关，即 $\sigma_x\neq 0$，$\sigma_y\neq 0$. 此时，“观测值”$\tilde{x}_i=x_i-\tilde{\varepsilon}_{x_i}$，$\tilde{y}_i=y_i-\tilde{\varepsilon}_{x_i}\hat{\beta}_1$，其方差以及协方差为

$$s_{\tilde{x}}^2 = s_x^2 + \sum\nolimits_{i=1}^{m}(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)^2/m - \sum\nolimits_{i=1}^{m}(x_i - \bar{x})(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)/m$$

$$s_{\tilde{y}}^2 = s_y^2 + \hat{\beta}_1^2\sum\nolimits_{i=1}^{m}(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)^2/m - \hat{\beta}_1\sum\nolimits_{i=1}^{m}(y_i - \bar{y})(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)/m$$

$$s_{\tilde{x}\tilde{y}} = s_{xy} - \hat{\beta}_1\sum\nolimits_{i=1}^{m}(x_i - \bar{x})(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)/m - \sum\nolimits_{i=1}^{m}(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)(y_i - \bar{y})/m + \hat{\beta}_1\sum\nolimits_{i=1}^{m}(\tilde{\varepsilon}_{x_i} - \bar{\tilde{\varepsilon}}_x)^2/m$$

其中，$\bar{\tilde{\varepsilon}}_x=\sum_{i=1}^{m}\tilde{\varepsilon}_{x_i}/m$，显然相关系数 $\tilde{\gamma}$ 以及 $\bar{\tilde{x}}$，$\bar{\tilde{y}}$ 都与 $\tilde{\varepsilon}_x$ 有关. 根据式(19a)，$\tilde{\varepsilon}_x$ 与观测值 $\boldsymbol{x}$，$\boldsymbol{y}$ 的观测精度 σ_x，σ_y 以及误差的相关程度 ρ 有关. 因此，在该情形下，忽略 σ_x 及 ρ 的传统最小二乘方法的回归效果受到影响，即此时传统最小二乘方法的回归效果不如新方法(见图2—4以及表1第3—5行). 特别地，当误差相关系数 $\rho=0$ 时，加权总体最小二乘与新方法是一致的(见表1第4行).

固定自变量和因变量的观测精度 $\sigma_x=1$ 和 $\sigma_y=3$，设相关系数 ρ 分别为：-0.9，-0.5，0.5，0.9，分别利用传统的最小二乘方法和新方法计算回归系数，即回归直线的斜率和截距，结果如图2所示. 图中，实点表示模拟原始点$[x_i, y_i]$，空心点表示经误差改正后点$[\tilde{x}_i, \tilde{y}_i]$.

如图2所示，相对于传统的最小二乘方法，顾及自变量与因变量观测误差及误差相关性的新方法获得的回归直线更加接近真实的直线. 如回归系数解(17)与(4)所示，两种方法进行回归分析时实际采用的观测数据(观测点)是不同的，传统的最小二乘方

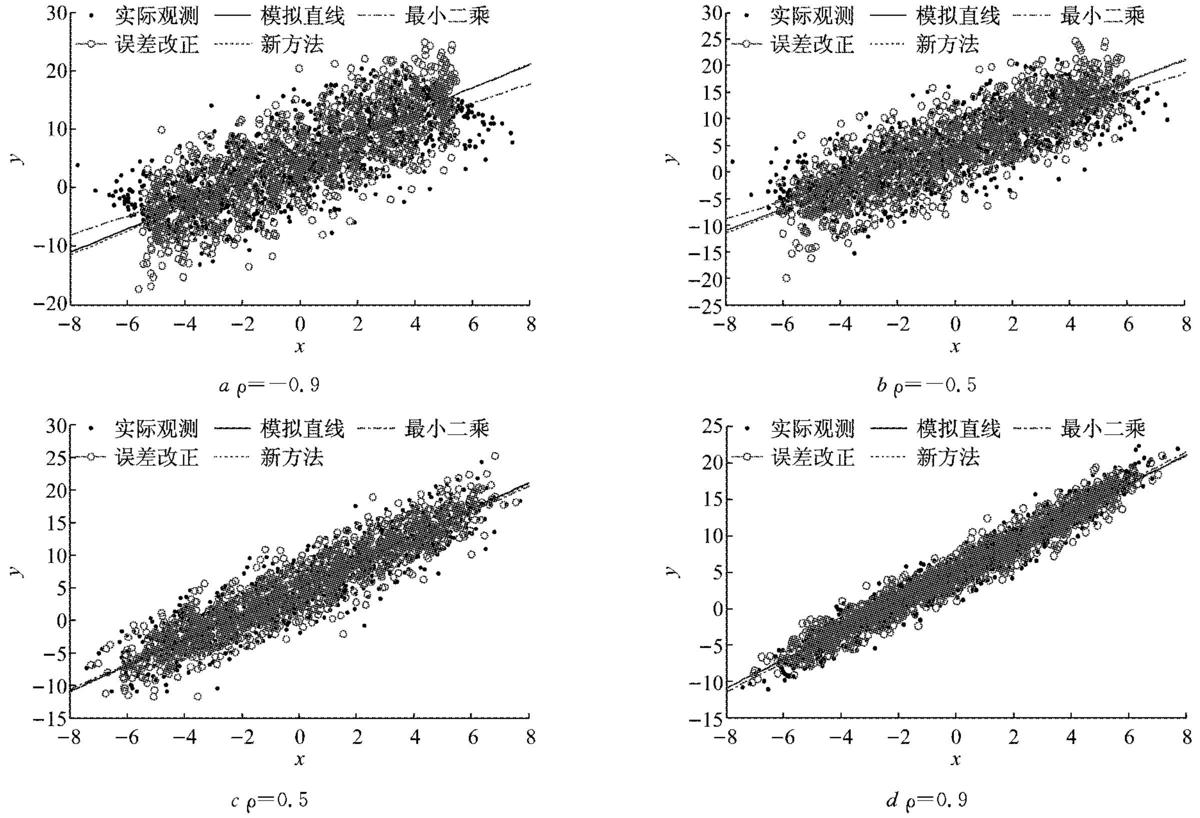

图 2　采用最小二乘方法与新方法进行一元线性回归

Fig.2　Univariate linear regression with traditional least-squares method and introduced new method

法采用原始观测点，而新方法将原始观测点的误差进行改正，然后利用改正后的“观测数据”，采用最小二乘方法求解回归直线.

假定目标回归直线为 $y=2x+5$，自变量与因变量观测值的精度分别为 $\sigma_x=1$，$\sigma_y=3$，根据观测误差之间不同的相关性(ρ 不同)，分别进行 500 次模拟实验，每次实验分别采用最小二乘方法、加权总体最小二乘方法、新方法计算回归系数. 各次模拟实验获得的回归系数 $\hat{\beta}_1$ 及 $\hat{\beta}_0$ 分别如图 3 和图 4 所示.

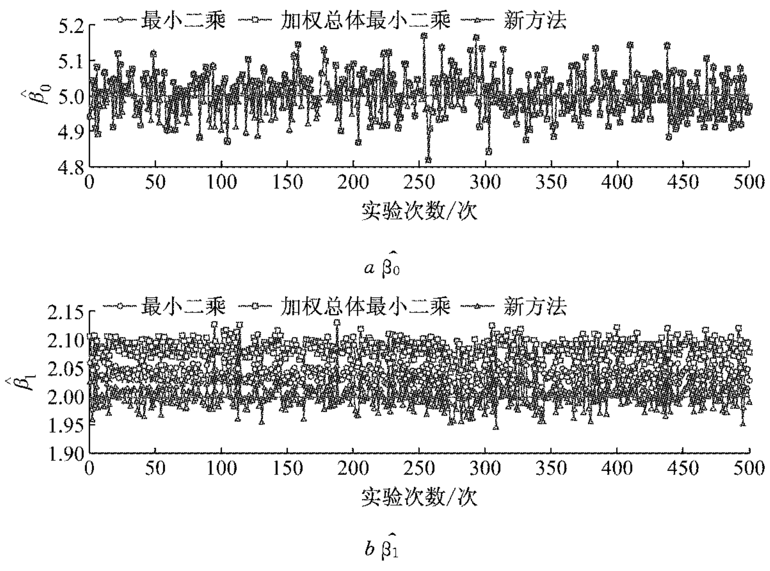

图 3　500 次模拟实验中三种回归方法求得的回归系数 $\hat{\beta}_1$ 及 $\hat{\beta}_0$，误差相关系数为 0.8

Fig.3　The regression coefficients, $\hat{\beta}_1$ and $\hat{\beta}_0$, from 500 simulation computations where the correlation coefficient $\rho=0.8$.

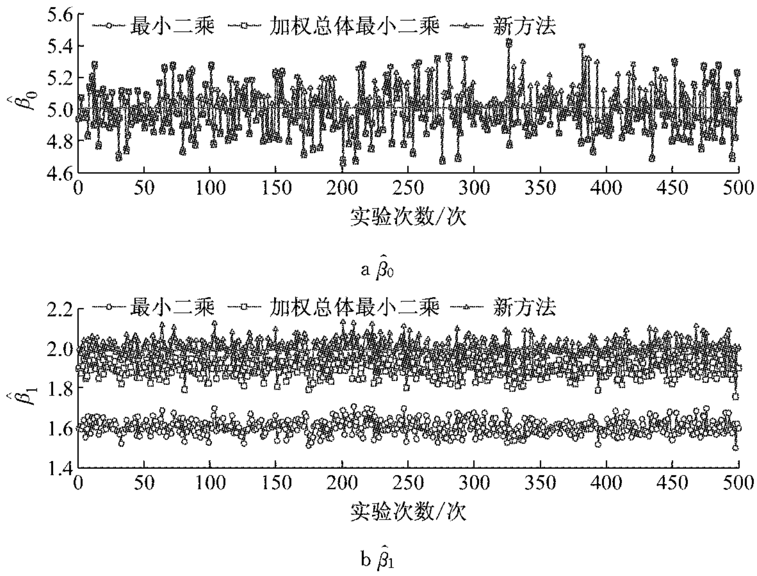

图4 500次模拟实验中三种回归方法求得的回归系数 $\hat{\beta}_1$ 及 $\hat{\beta}_0$,误差相关系数为−0.8

Fig.4 The regression coefficients, $\hat{\beta}_1$ and $\hat{\beta}_0$, from 500 simulation computations where the correlation coefficient $\rho=-0.8$.

从图3、图4可以看出,几种方法获得的回归直线的截距 $\hat{\beta}_0$ 基本是一致的,其原因是系数矩阵中与此相应的列是确定的,无误差的. 对于回归直线的斜率 $\hat{\beta}_1$,同时考虑变量误差及其相关性的回归方法获得的斜率比传统的最小二乘方法和只考虑变量误差的总体最小二乘方法获得的斜率更加接近真值. 图4中加权总体最小二乘方法的回归效果优于最小二乘方法,而图3中,两种方法的效果却相反,说明加权总体最小二乘的效果并不总是优于最小二乘的效果,如文献[18]中所示.

固定因变量的观测精度 $\sigma_y=3$,将自变量的观测精度 σ_x 分别设为0与1,改变观测误差的相关程度 ρ,根据直线 $y=2x+5$ 分别进行500次模拟实验,将获得的回归系数 $\hat{\beta}_1$ 及 $\hat{\beta}_0$ 取均值,见表1.

表1 几种线性回归方法获得的回归系数

Tab.1 The regression coefficients from different linear regression methods

σ_x	ρ	最小二乘		加权总体最小二乘		新方法	
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
	−0.8	5.008 0	1.602 2	5.008 5	1.900 1	5.008 6	2.000 7
1	0	5.003 6	1.817 5	5.003 5	1.998 9	5.003 5	1.998 9
	0.8	4.999 2	2.035 8	4.999 2	2.084 2	4.999 2	1.999 3
0		5.001 5	1.999 9	5.001 5	1.999 9	5.001 5	1.999 9

从表1以及图3,图4可以看出,各组试验获得的回归直线的截距基本是一致的,说明回归分析中常数项的求解结果基本不受自变量观测误差的影响. 虽然回归直线的斜率受自变量观测误差及其与因变量观测误差的相关性影响较大,但是新方法综合考虑了变量的观测误差以及误差之间的相关性,其获得的回归直线与真实的直线更加接近.

由于加权总体最小二乘方法是新方法的特殊情形,因此本文只比较了最小二乘方法与新方法. 两种方法的共同点在于:

(1) 两种方法获得的回归系数的解析形式相同,如式(4)与(17)所示. 回归系数都可以用观测数据 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的方差 s 以及相应的相关系数 γ 表示.

(2) 获得的回归直线都过某一定点,定点的坐标为 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 或者其改正值 $\tilde{\boldsymbol{x}}$ 和 $\tilde{\boldsymbol{y}}$ 的(加权)平均值.

两种回归分析方法的差异在于:

(1) 传统最小二乘方法只考虑了因变量的观测误差而忽略了自变量的观测误差;新方法不但同时考虑了自变量和因变量的观测误差 $\boldsymbol{\varepsilon}_x$ 和 $\boldsymbol{\varepsilon}_y$,还考虑了观测误差之间的相关性 ρ.

(2) 传统方法采用原始观测数据 x_i 与 y_i(如图2中的实点)进行回归分析,而新方法采用经误差改正的新"观测"数据 $\tilde{x}_i$ 与 $\tilde{y}_i$(如图2中的空心点)进行回归分析.

(3) 两种方法获得的回归直线通过的定点不同,传统方法获得的直线通过定点$[\bar{x},\bar{y}]$,而新方法获得的直线通过定点$[\bar{\tilde{x}},\bar{\tilde{y}}]$.

5 结论

针对线性回归分析,本文提出一种同时考虑自

变量和因变量观测误差及误差相关性的新方法，并给出了新方法求解回归系数的非线性迭代解过程.针对一元线性回归问题，推导了相应的回归系数解析形式，并与传统方法获得的回归系数解析形式进行了比较.传统最小二乘方法在回归分析时仅考虑因变量的观测误差，并采用原始观测数据 x_i 与 y_i 进行回归分析.新方法不仅同时考虑了自变量与因变量的观测误差，还考虑了误差的相关性，采用的数据为误差改正后的新"观测值"$\tilde{x}_i$ 与 $\tilde{y}_i$.当只有因变量是随机变量时，两种回归分析方法的效果相同；当自变量与因变量都是随机变量时，两种方法进行回归分析的效果有明显的差异，相比于传统最小二乘方法，新方法获得的回归直线更加接近真实直线.

在实际回归分析应用中，例如自回归模型，自变量和因变量观测误差的相关性通常难以准确获得，可以根据经验判断误差的相关性.

参考文献：

[1] 邓勃. 分析测试数据的统计处理方法[M]. 北京：清华大学出版社，1994.
DENG Bo. Statistical processing method for data of analytic and test[M]. Beijing：Tsinghua university press，1994.

[2] Ryan A G，Montgomery D C，Peck E A，*et al*. Introduction to linear regression analysis，solutions manual to accompany[M]. 5th ed. Hoboken：Wiley，2013.

[3] Chambers J M，Cleveland W S，Kleiner B，Tukey P A. Graphical methods for data analysis[M]. Belmont：Duxbury Press，1983.

[4] Sykes A O. An introduction to regression analysis[M]. Chicago：The Inaugural Coase Lecture，Law School，University of Chicago，1993.

[5] 鲁铁定，陶本藻，周世健. 基于整体最小二乘法的线性回归建模和解法[J]. 武汉大学学报：信息科学版，2008，33(5)：504.
LU Tieding，TAO Benzao，ZHOU Shijian. Modeling and algorithm of linear regression based on total least squares[J]. Geomatics and Information Science of Wuhan University，2008，33(5)：504.

[6] Schaffrin B，Wieser A. On weighted total least-squares adjustment for linear regression[J]. Journal of Geodesy，2008，82(7)：415.

[7] Shen Y Z，Li B F，Chen Y. An iterative Solution of weighted total least-squares adjustment[J]. Journal of Geodesy，2011，85(10)：229.

[8] Li B F，Shen Y Z，Li W X. The seamless model for three-dimensional datum transformation[J]. Science China：Earth Science，2012，55(12)：2099.

[9] Xu P L，Liu J N，Shi C. Total least squares adjustment in partial errors-in-variables models：algorithm and statistical analysis[J]. Journal of Geodesy，2012，86(8)：661.

[10] Snow K. Topics in total least-squares adjustment within the errors-in-variables model：singular cofactor matrices and priori information[D]. Columbus：School of Earth Sciences，the Ohio State University，2012.

[11] Fang X. Weighted total least squares：necessary and sufficient conditions，fixed and random parameters[J]. Journal of Geodesy，2013，87(8)：733.

[12] 张尧庭，方开泰. 多元统计分析引论[M]. 武汉：武汉大学出版社，2013.
ZHANG Yaoting，FANG Kaitai. An introduction to multivariate statistical analysis[M]. Wuhan：Wuhan University Press，2013.

[13] Edwards A L. An introduction to linear regression and correlation[M]. New York：William H. Freeman and Company，1976.

[14] Gideon R A. The correlation coefficients[J]. Journal of Modern Applied Statistical Methods，2007，6(2)：517.

[15] Koch K R. Least-squares adjustment and collocation[J]. Bulletin géodésique，1977，51(2)：127.

[16] Koch K R. Parameter estimation and hypothesis testing in linear models[M]. 2nd ed. Berlin/Heidelberg/New York：Springer，1999.

[17] Li B F，Wang M M，Yang Y X. Multiple linear regression with correlated explanatory variables and responses[J]. Survey Review，2015. DOI：http://dx. doi. org/10. 1179/1752270615Y.0000000006.

[18] Xu P L，Liu J N，Zeng W X，*et al*. Effects of errors-in-variables on weighted least squares estimation[J]. Journal of geodesy，2014，88(7)：705.