四舍五入舍入数据对回归系数的影响及应用路径研究

唐美燕　张玉芳　吴小莉
桂林学院 广西 桂林 541006

【摘　要】：本文基于最小二乘原理，结合舍入误差的均匀分布与独立假设，探究四舍五入舍入数据对一元线性回归系数的影响机制。研究发现，四舍五入舍入数据会导致回归斜率产生系统性向下偏倚，截距偏差由斜率偏差传导且受多因素调节；该偏倚为系统性固定偏差，无法通过增大样本量消除。据此，进一步给出舍入精度 K 值设定、模型与预测结果校正的应用路径，为实际建模中舍入数据的回归分析提供理论参考与应用指导。

【关键词】：最小二乘原理；舍入误差；回归系数；模型校正

DOI:10.12417/2705-0998.26.02.083

1 引言

在统计推断与数据建模实践中，数据准确性是保障分析结果可靠性的核心前提。受测量工具精度、记录与存储精度等客观条件限制，实际获取的观测数据往往经过四舍五入处理，属于舍入数据而非真实数据。四舍五入作为目前常用的舍入方法，其引入的舍入误差会通过统计分析环节逐步传导，对估计结果的优良性产生潜在干扰，这也使得探究四舍五入舍入数据对回归分析的影响具有重要意义。

线性回归分析是计量分析、数据建模中的基础方法，回归系数的估计精度直接决定模型的解释力与预测效果。在实际回归建模过程中，四舍五入舍入数据的应用极为广泛，此类数据带来的舍入误差，对回归系数的具体扰动机制等相关问题，目前研究仍处于探索完善阶段。现有研究中，学者虽已将舍入误差的独立均匀分布假设应用于回归分析[1][2]，但针对四舍五入舍入数据对一元线性回归斜率、截距估计值的偏倚规律，尚未形成较为完善的量化理论结论，相关研究仍有进一步探索的空间。

基于此，本文以一元线性回归模型为研究对象，基于最小二乘原理，结合四舍五入舍入误差均匀分布且与原始变量相互独立的经典假设，系统推导四舍五入数据与真实数据下回归系数的偏差公式，量化分析舍入误差对回归斜率、截距的影响机制。在此基础上，从源头控制与事后校正双维度出发，提出舍入精度 k 值的设定路径，以及模型系数、预测结果的校正方法，为实际建模中舍入数据的处理提供支撑。

2 四舍五入舍入数据对回归系数的影响

2.1 基本假设

自变量 X 与因变量 Y 对应的四舍五入变量记为 X^*, Y^*，保留 k 位小数，

相应的舍入误差变量如下：

$$\delta = X^* - X, \xi = Y^* - Y$$

其中，δ 为自变量 X 的舍入误差变量，ξ 为自变量 Y 的舍入误差变量。

假设 1：舍入误差 δ、ξ 独立同分布于区间 $[-0.5\times10^{-k}, 0.5\times10^{-k}]$ 上的均匀分布。

假设 2：舍入误差与变量相互独立，即 δ 与 X、Y 无关，ξ 与 X、Y 无关。

由假设 1 得到舍入误差 δ、ξ 的期望为均为 0，方差均为 $\frac{10^{-2k}}{12}$.假设 2 的合理性有充分文献支撑：Sheppard（1898）[3] 与 Schneeweiss 等（2010）[4]指出，当变量分布光滑、舍入步长远小于变量变异时，舍入误差为非差异性测量误差，与变量相互独立。胡果荣（2006）[1]、Heitjan&Rubin（1991）[5]及 Hall（1982）[6]均将该假设作为舍入数据矩估计、最大似然估计与回归分析的标准前提.因此，在研究舍入数据对回归系数影响的场景下，假设 2 具有合理性。

作者简介：唐美燕，副教授，主要研究方向：数理统计、统计建模。

基金项目：广西高校中青年教师科研基础能力提升项目：基于四舍五入数据的统计推断与应用，项目号 2022KY1576；广西高等教育本科教学改革工程项目：“五位锚点·闭环赋能”《概率统计》课程思政教学体系的研究与实践，项目号 2025JGB521；校级科研项目：基于舍入数据的参数估计研究；校级课程思政专项教改项目：基于翻转课堂的数学与应用数学专业《概率统计》课程思政教学研究与实践。

2.2 基于最小二乘原理建立一元线性回归模型

（1）原始数据的最小二乘估计

自变量 X 与因变量 Y 的真实数据样本为 $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$,其中 n 为样本容量。根据最小二乘原理，基于真实数据样本对 X 与 Y 建立一元线性回归模型 $Y=a+bX+\varepsilon$，其中，a 为回归截距，b 为回归斜率，ε 为随机误差项，且满足 $E(\varepsilon)$=0，回归系数的估计值为：

$$\hat{b}=\frac{S_{XY}}{S_{XX}},\hat{a}=\overline{Y}-\hat{b}\overline{X}.$$

回归系数的估计值中，$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i,\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ 分别为 X 与 Y 样本均值；$S_{XX}=\sum_{i=1}^{n}(x_i-\overline{X})^2$ 为自变量的离均差平方和,反映自变量 X 的波动程度；$S_{XY}=\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})$ 为自变量 X 与因变量 Y 的离均差乘积和，反映两者的线性相关程度。

（2）四舍五入舍入数据的最小二乘估计

对于自变量 X 与因变量 Y 的真实数据样本 $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$进行四舍五入后，得到四舍五入变量 X^*,Y^*保留 k 位小数的舍入数据样本如下：

$$(x_1^*,y_1^*),(x_2^*,y_2^*),...,(x_n^*,y_n^*)$$

同样根据最小二乘原理，基于四舍五入舍入数据样本构建 X^*,Y^*的一元线性回归模型为：$Y^*=a^*+b^*X^*+\varepsilon^*$，此时回归系数估计值为：

$$\hat{b}^*=\frac{S_{X^*Y^*}}{S_{X^*X^*}},\hat{a}^*=\overline{Y}^*-\hat{b}^*\overline{X}^*.$$

式中，$\overline{X}^*=\frac{1}{n}\sum_{i=1}^{n}x_i^*,\overline{Y}^*=\frac{1}{n}\sum_{i=1}^{n}y_i^*$ 分别为 X^* 与 Y^* 样本均值；$S_{X^*X^*}=\sum_{i=1}^{n}(x_i^*-\overline{X}^*)^2$，$S_{X^*Y^*}=\sum_{i=1}^{n}(x_i^*-\overline{X}^*)(y_i^*-\overline{Y}^*)$ 分别为舍入变量 X^* 的离均差平方和、X^* 与 Y^* 的离均差乘积和。

2.3 回归系数估计值的偏差及期望

结合 2.2 节内容，接下来理论推导真实数据样本与对应四舍五入舍入数据样本下，回归斜率 $\hat{b}$ 与 $\hat{b}^*$ 的偏差，回归截距 $\hat{a}$ 与 $\hat{a}^*$ 的偏差,并计算 $\hat{b}^*$、$\hat{a}^*$ 的期望。

2.3.1 回归斜率的偏差及期望

根据 $\delta_i=x_i^*-x_i,\xi_i=y_i^*-y_i$，得到 $\overline{X}^*=\overline{X}+\overline{\delta}$、$\overline{Y}^*=\overline{Y}+\overline{\xi}$，接着对 $S_{X^*X^*}$ 展开：

$$\begin{aligned}S_{X^*X^*}&=\sum_{i=1}^{n}(x_i^*-\overline{X}^*)^2=\sum_{i=1}^{n}(x_i+\delta_i-\overline{X}-\overline{\delta})^2\\&=\sum_{i=1}^{n}[(x_i-\overline{X})+(\delta_i-\overline{\delta})]^2\\&=\sum_{i=1}^{n}[(x_i-\overline{X})^2+2(x_i-\overline{X})(\delta_i-\overline{\delta})+(\delta_i-\overline{\delta})^2]\\&=S_{XX}+2\sum_{i=1}^{n}(x_i-\overline{X})(\delta_i-\overline{\delta})+\sum_{i=1}^{n}(\delta_i-\overline{\delta})^2\end{aligned}$$

其中 $2\sum_{i=1}^{n}(x_i-\overline{X})(\delta_i-\overline{\delta})+\sum_{i=1}^{n}(\delta_i-\overline{\delta})^2$ 表示自变量 X 舍入误差引起的离均差平方和增量，记为 M_{XX}，则 $S_{X^*X^*}=S_{XX}+M_{XX}$.同理对 $S_{X^*Y^*}$ 进行展开，得到：

$$\begin{aligned}S_{X^*Y^*}&=\sum_{i=1}^{n}(x_i^*-\overline{X}^*)(y_i^*-\overline{Y}^*)=\sum_{i=1}^{n}(x_i+\delta_i-\overline{X}-\overline{\delta})(y_i+\varepsilon_i-\overline{Y}-\overline{\varepsilon})\\&=\sum_{i=1}^{n}[(x_i-\overline{X})(y_i-\overline{Y})+(x_i-\overline{X})(\varepsilon_i-\overline{\varepsilon})+(y_i-\overline{Y})(\delta_i-\overline{\delta})+(\delta_i-\overline{\delta})(\varepsilon_i-\overline{\varepsilon})]\\&=S_{XY}+\sum_{i=1}^{n}[(x_i-\overline{X})(\varepsilon_i-\overline{\varepsilon})+(y_i-\overline{Y})(\delta_i-\overline{\delta})+(\delta_i-\overline{\delta})(\varepsilon_i-\overline{\varepsilon})]\end{aligned}$$

其中

$$\sum_{i=1}^{n}[(x_i-\overline{X})(\varepsilon_i-\overline{\varepsilon})+(y_i-\overline{Y})(\delta_i-\overline{\delta})+(\delta_i-\overline{\delta})(\varepsilon_i-\overline{\varepsilon})]$$

表示舍入误差与真实数据的离均差乘积和增量，记为 M_{XY}，则 $S_{X^*Y^*}=S_{XY}+M_{XY}$.

结合 $S_{X^*X^*}=S_{XX}+M_{XX}$、$S_{X^*Y^*}=S_{XY}+M_{XY}$，计算回归斜率 $\hat{b}$ 与 $\hat{b}^*$ 的偏差，可得斜率估计值的偏差：

$$\hat{b}^*-\hat{b}=\frac{S_{XX}M_{XY}-S_{XY}M_{XX}}{S_{XX}(S_{XX}+M_{XX})}.$$

接着计算回归斜率 $\hat{b}$ 与 $\hat{b}^*$ 偏差的期望 $E(\hat{b}^*-\hat{b})$，因为 $\hat{b}$ 为非随机常数，故 $E(\hat{b}^*-\hat{b})=E(\hat{b}^*)-\hat{b}=E(\frac{S_{XX}M_{XY}-S_{XY}M_{XX}}{S_{XX}(S_{XX}+M_{XX})})$.为了化简此式子，结合假设 1、假设 2 以及“相互独立的随机变量，乘积的期望等于期望的乘积”先计算 M_{XX}、M_{XY} 的期望，如下：

$$E(M_{XX})=E(2\sum_{i=1}^{n}(x_i-\overline{X})(\delta_i-\overline{\delta})+\sum_{i=1}^{n}(\delta_i-\overline{\delta})^2)=$$

$$E(\sum_{i=1}^{n}(\delta_i-\overline{\delta})^2)=\frac{n10^{-2k}}{12};$$

$$E(M_{XY})=$$

$$E\{\sum_{i=1}^{n}[(x_i-\overline{X})(\varepsilon_i-\overline{\varepsilon})+(y_i-\overline{Y})(\delta_i-\overline{\delta})+(\delta_i-\overline{\delta})(\varepsilon_i-\overline{\varepsilon})]\}=0$$

由于舍入误差为四舍五入保留 K 位小数的微小扰动，误差量级极小，而 S_{XX} 作为自变量 X 的离均差平方和，反映核

心波动，数值相对较大，M_{XX} 作为舍入误差引起的偏差增量，其量级远小于 S_{XX}，因此 $S_{XX}(S_{XX}+M_{XX})\approx S_{XX}^2$.最终得到回归斜率 $\hat{b}$ 与 $\hat{b}^*$ 偏差的期望 $E(\hat{b}^*-\hat{b})$ 化简如下：

$$E(\hat{b}^*-\hat{b})=E(\hat{b}^*)-\hat{b} = E(\frac{-S_{XY}M_{XX}}{S_{XX}(S_{XX}+M_{XX})})\approx E(\frac{-S_{XY}}{S_{XX}}\cdot\frac{M_{XX}}{S_{XX}})=-\hat{b}\cdot\frac{n\cdot10^{-2k}}{12S_{XX}}\cdot$$

移项后即可得到斜率估计值 $\hat{b}^*$ 的数学期望公式：

$$E(\hat{b}^*)\approx\hat{b}(1-\frac{n\cdot10^{-2k}}{12S_{XX}})\cdot$$

2.3.2 回归截距的偏差及期望

接下来推导真实数据样本与对应四舍五入舍入数据样本下，回归截距 $\hat{a}$ 与 $\hat{a}^*$ 的偏差，如下：

$$\hat{a}^*-\hat{a}=\overline{Y}^*-\hat{b}^*\overline{X}^*-(\overline{Y}-\hat{b}\overline{X})=\overline{Y}+\overline{\xi}-\hat{b}^*(\overline{X}+\overline{\delta})-(\overline{Y}-\hat{b}\overline{X})$$

$$=\overline{Y}+\overline{\xi}-\hat{b}^*(\overline{X}+\overline{\delta})-\overline{Y}+\hat{b}\overline{X}$$

$$=\overline{\xi}-\overline{X}(\hat{b}^*-\hat{b})-\hat{b}^*\overline{\delta}\cdot$$

对上式两边取期望，得到

$$E(\hat{a}^*-\hat{a})=E(\hat{a}^*)-\hat{a}\approx-\overline{X}\cdot-\hat{b}\frac{n\cdot10^{-2k}}{12S_{XX}}=\overline{X}\hat{b}\frac{n\cdot10^{-2k}}{12S_{XX}}\cdot$$

2.4 结论归纳

结合上述一元线性回归系数偏差的理论推导，可得出以下核心结论，为后续实证分析及实际应用提供坚实的理论支撑，具体如下所述。

2.4.1 斜率的系统性向下偏倚特性

四舍五入舍入数据会导致一元线性回归模型的斜率产生系统性向下偏倚，其偏倚程度可通过斜率衰减率 V 进行精准量化，斜率衰减率 V 的表达式为：

$$V=\frac{\hat{b}-E(\hat{b}^*)}{\hat{b}}=\frac{n\cdot10^{-2K}}{12S_{XX}}\cdot$$

由上述表达式可知，斜率衰减率 V 由样本量 n、自变量 X 的离均差平方和 S_{XX} 及四舍五入保留小数位数 K 共同决定，各因素对斜率衰减率及斜率偏倚的影响规律如下：

（1）斜率衰减率 V 与 $n\cdot10^{-2K}$ 呈正相关关系。样本量 n 越大，舍入误差的累积效应越强；K 越小即保留小数位数越少，10^{-2K} 的数值越大，舍入误差的基础量级越高。因此，$n\cdot10^{-2K}$ 变大导致斜率衰减率 V 增大，即斜率向下偏倚程度加剧。

（2）斜率衰减率与 S_{XX} 呈负相关关系。S_{XX} 反映自变量 X 的波动程度，其数值越大，说明真实数据的核心信号强度越强，则舍入误差的扰动影响被稀释，斜率衰减率 V 越小，斜率偏倚程度越弱。

（3）斜率的系统性向下偏倚具有稳定性。根据 $E(\hat{b}^*)\approx\hat{b}(1-\frac{n\cdot10^{-2k}}{12S_{XX}})$ 以及 $1-\frac{n\cdot10^{-2k}}{12S_{XX}}$ 小于 1，说明基于四舍五入舍入数据得到的斜率估计值的期望 $E(\hat{b}^*)$ 小于基于真实数据得到的斜率 $\hat{b}$，即斜率的系统性向下偏倚具有稳定性。

综上，样本量 n 越大、自变量 X 波动越小、舍入精度 K 越低，斜率的系统性向下偏倚效应越显著。舍入后斜率估计值系统性变小，这是舍入误差累积导致的稳定偏倚，且斜率衰减率越大，斜率变小幅度越明显。

2.4.2 截距偏差的调节机制

四舍五入舍入数据对回归截距偏差的影响并非独立存在，而是由斜率偏差传导而来，同时受自变量均值 $\overline{X}$ 与舍入精度 K 的双重调节。由截距偏差期望表达式：

$$E(\hat{a}^*)-\hat{a}\approx\overline{X}\hat{b}\frac{n\cdot10^{-2k}}{12S_{XX}},$$

总结截距偏差的规律，具体如下：

（1）截距偏差的绝对值与 $|\overline{X}|$、$|\hat{b}|$、$n10^{-2k}$ 呈正相关，与 S_{XX} 呈负相关。即原始斜率的绝对值越大、自变量 X 均值的绝对值越大、舍入精度 K 越低、样本量 n 越大，截距偏差的绝对值越大；反之，截距偏差的绝对值越小。

（2）截距偏差的符号由原始斜率与自变量均值的符号共同决定：当 $\hat{b}$ 与 $\overline{X}$ 同号时，截距偏差为正；当 $\hat{b}$ 与 $\overline{X}$ 异号时，截距偏差为负。

（3）舍入精度 K 是截距偏差的核心控制因素之一，K 越小，10^{-2K}的数值越大，截距偏差的量级也随之显著增大，即舍入精度 K 越低，截距的偏倚程度越严重。

2.4.3 舍入误差的系统性影响本质

尽管舍入误差 δ_i、ξ_i 满足独立同分布特性，其随机扰动可通过样本叠加实现部分抵消，但无法消除其对回归系数（斜率、截距）的系统性偏倚影响，其核心特性如下：

（1）统性偏倚并非随机误差，而是由四舍五入操作本身带来的固定偏倚，其偏倚大小直接由舍入精度 K 控制，K 越低，偏倚的基础量级越高，斜率衰减率与截距偏差的数值越大，对回归模型参数的影响越显著。

（2）增大样本量 n 仅能改变偏倚幅度，无法从根本上消除偏倚，进而导致舍入数据与真实数据构建的回归模型存在一定的斜率偏差、截距偏差；这种偏差由 n、K、S_{XX} 等固定

条件决定，产生原因固定、变化规律可预测，不会因随机因素消失，从而降低模型解释与预测精度。

3 应用路径

基于四舍五入舍入数据对回归系数影响的结论，结合建模流程，构建舍入精度 K 值设定、模型系数与预测校正的应用路径，解决舍入误差导致的回归偏差问题，实现理论落地，提升模型精度与可靠性。

3.1 舍入精度 K 值设定

建模前明确合适的舍入精度 K 值，旨在从源头控制舍入误差对回归偏差的影响程度，提高建模的精度与效率，具体应用路径如下：

（1）由四舍五入数据计算舍入变量 X^* 的离均差平方和 $S_{X^*X^*}$；

（2）由 $S_{XX}=E(\mathrm{S}_{X^*X^*})-\dfrac{n10^{-2k}}{12}$ 以及 $S_{X^*X^*}$ 为 $E(S_{X^*X^*})$ 的无偏估计，求得 $S_{XX}\approx \mathrm{S}_{X^*X^*}-\dfrac{n10^{-2k}}{12}$；

（3）以斜率衰减率小于等于 V_0 为精度要求，结合斜率衰减率公式 $V=\dfrac{n\cdot 10^{-2K}}{12S_{XX}}$，得到舍入精度 K 值的取值范围公式：$K\geq \dfrac{1}{2}\lg(\dfrac{n}{12V_0S_{XX}})$；

（4）取满足该不等式的最小非负整数，即为合适的最小舍入精度 K 值。

3.2 模型系数与预测校正

根据四舍五入舍入数据对回归系数影响的结论，进行建模中系数校正与建模后预测校正，提升模型精度与预测可靠性，具体应用路径如下：

（1）斜率校正

由 $E(\hat{b}^*)\approx\hat{b}(1-\dfrac{n\cdot 10^{-2k}}{12S_{XX}})$，得 $\hat{b}\approx\dfrac{\hat{b}^*}{1-\dfrac{n\cdot 10^{-2k}}{12S_{XX}}}$，在此式中再代入 S_{XX} 近似值 $\mathrm{S}_{X^*X^*}-\dfrac{n10^{-2k}}{12}$ 进行计算得到 $\hat{b}\approx\hat{b}^*_{校}$，从而得到基于舍入数据所建模型斜率 $\hat{b}^*$ 的校正斜率 $\hat{b}^*_{校}$.

（2）截距校正

由 $E(\hat{a}^*)-\hat{a}\approx\overline{X}\hat{b}\dfrac{n\cdot 10^{-2k}}{12S_{XX}}$，得 $\hat{a}\approx E(\hat{a}^*)-\overline{X}\hat{b}\dfrac{n\cdot 10^{-2k}}{12S_{XX}}$，再将 $\hat{b}$ 由校正后的斜率 $\hat{b}^*_{校}$ 近似，S_{XX} 由 $\mathrm{S}_{X^*X^*}-\dfrac{n10^{-2k}}{12}$ 近似，得到 $\hat{a}\approx\hat{a}^*_{校}$，从而得到基于舍入数据所建模型截距 $\hat{a}^*$ 的校正截距 $\hat{a}^*_{校}$.

（3）预测结果校正

通过前面两步得到校正模型 $Y=\hat{a}^*_{校}+\hat{b}^*_{校}X$，在实际中因难以获取自变量的真实数据 x_i，结合假设 1 由四舍五入舍入数据 x_i^* 与区间 $[-0.5\times10^{-k},0.5\times10^{-k}]$ 的随机数之和 x^*_{i+} 近似 x_i.由 x^*_{i+} 代入校正模型 $Y=\hat{a}^*_{校}+\hat{b}^*_{校}X$ 预测 y_i 的值，提高预测效果。

4 结语

本文基于最小二乘原理，结合舍入误差的均匀分布与独立假设，系统推导了四舍五入舍入数据对一元线性回归斜率、截距估计值的偏差机制，明确了斜率系统性向下偏倚、截距偏差由斜率偏差传导且受多因素调节的核心规律，也证实舍入误差带来的偏倚为系统性固定偏差，无法通过增大样本量消除。在此基础上，所提出的舍入精度 K 值设定路径与模型系数、预测结果校正路径，从源头控制与事后校正双维度解决了舍入误差引发的回归模型偏差问题。该研究成果为实际建模中舍入数据的处理提供了可落地的理论参考与应用路径，也为后续复杂回归模型的舍入误差研究奠定了一定的基础。

参考文献：

[1] 胡果荣.基于舍入数据的统计推断[D].吉林大学,2006.

[2] 宋扬.四舍五入数据的线性回归问题[D].清华大学,2017.

[3] Sheppard W F.On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale[M].Proc.london math.Soc,1898,29:353-380.

[4] Schneeweiss H,Komlos J,Ahmad A S.Symmetric and asymmetric rounding:a review and some new results[J].AStA Adv Stat Anal,2010,94:247-271.

[5] Heitjan D F,Rubin D B.Ignorability and coarse data[J].Ann Statist,1991,19:2244-2253.

[6] Hall P.The influence of rounding errors on some nonparametric estimators of a density and its derivatives[J].Sima APPl math,1982,42:390-399.