·基础理论与方法·

测量误差变量与准确测量变量混合对研究真实性的影响

杨美霞　周艺彪　姜庆五

【摘要】 **目的**　探讨测量误差变量与准确测量变量混合情况下测量误差对联系效应估计的影响。**方法**　利用测量误差大小、准确测量变量与测量误差变量之间的相关性、准确测量变量的个数和联系效应之间的函数，采用 R 软件做图来讨论分析测量误差对研究真实性的影响。**结果**　当连续变量 Y 和 Z 能准确测量，连续变量 X 不能准确测量时，无差异性测量误差使所估计的联系效应值总低于实际值，并随 X 与 Z 的相关程度的增加，测量误差所致的偏倚会进一步地恶化。在一个错分二分类变量 X 和一个准确测量连续变量 Z 混合的情况下，测量误差所致的偏倚不仅跟暴露测量的灵敏度和特异度有关，而且跟 X 与 Z 的相关系数以及 X 的暴露比例有关，并且随着相关系数的增加，AF 值逐渐减少。在 $\rho=0.5$ 时，AF 值为 1.419，变量 X 对应变量 Y 的联系效应估计值大于实际值，但当 ρ 增至 0.9 时，AF 值为 0.474，其联系效应估计值低于实际值，改变了错分偏倚的方向。**结论**　在准确测量变量和测量误差变量混杂的研究中，用线性回归模型来分析估计多个自变量与应变量之间的联系时，对测量误差所致偏倚的识别、控制和评估是十分必要的，对结果的解释要谨慎。

【关键词】 测量误差变量；准确测量变量；测量误差；偏倚

The impact of incorrectly-measured variables when mixed with precisely measured variables on the study of validity in epidemiological research

*YANG Mei-xia**, *ZHOU Yi-biao*, *JIANG Qing-wu*. **Xuhui District Center for Disease Control and Prevention*, *Shanghai* 200031, China

【**Abstract**】 **Objective**　To explore the impact of measurement error on the associated effects under the incorrectly-measured variables when mixed with precisely measured variables. **Methods**　Based on the functions of measurement error, correlation of incorrectly-measured predictors and precisely measured explanatory variables, number of precisely measured explanatory variables and associated effect, the 'R Project for Statistical Computing' method is used to analyze the impact of measurement on the validity of a study. **Results**　Under the scenario that the continuous response Y and the continuous explanatory Z are precisely measured but the continuous predictor X is incorrectly-measured, when focusing on inference about the effect of X on Y, the non-differential measurement error always makes the value of estimated effect less than the actual value, and the attenuation effect of measurement error more closely worsens the correlation of X and Z. Under a misclassification dichotomous predictor X with an additional precisely measured explanatory variable Z and focusing on inference about the effect of X on Y, the misclassification bias is not only related to the sensitivity and specificity of exposure measurement, but also to the correlation between X and Z and exposure proportion of X. The attenuation factor(AF) decreases gradually with the increasing correlation between X and Z. For instance, in the $\rho=0.5$ scenario, AF is 1.419, and the estimated effect of dichotomous predictor X on continuous response Y is more than the actual effect. When it increases to 0.9, AF is 0.474, the estimated effect becomes less than the true effect. **Conclusion**　In the studies of the impact of measurement error in linear regression with additional precisely measured explanatory variables, the impact of measurement error on the associated effect is relatively complex, suggesting that it is necessary to control and to assess the measurement error bias in order to correctly interpret the results of a study.

【**Key words**】 Mismeasured variable; Precisely measured variable; Measurement error; Bias

在流行病学研究中往往需要对暴露或变量进行测量，有些变量或暴露是可以准确测量的，如性别、种族等，而另一些变量是很难准确测量的，如环境污染暴露等，对这些变量的测量过程中不可避免地会产生测量误差。暴露测量误差是流行病学研究

基金项目：国家自然科学基金资助项目(30590374)

作者单位：200031 上海市徐汇区疾病预防控制中心(杨美霞)；复旦大学公共卫生学院流行病学教研室(周艺彪、姜庆五)

中偏倚产生的一个重要来源之一，它往往会高估或低估暴露与疾病之间的联系效应，甚至会导致暴露与疾病之间虚假的关联。有关暴露测量误差对研究真实性的影响，已有许多学者对此进行了讨论[1-3]，但很少涉及在研究中既有准确测量的变量又有测量误差的变量情况下，讨论暴露测量误差对研究真实性的影响。本文主要根据暴露测量误差大小、准确测量变量与测量误差变量之间的相关系数、准确测量变量的个数和联系效应之间的函数，采用R软件做图来讨论分析测量误差变量与准确测量变量混合情况下测量误差对联系效应的影响。

1.连续变量：设 Y 为连续反应变量，X 和 Z 为连续自变量，Y 和 Z 能准确测量，X 不能准确测量，$\tilde{X}$ 为 X 的观测值，t 为 X 的测量误差，$t = SD(\tilde{X} \mid X)/SD(X)$，例如当 $t = 0.1$，可以认为在 X 的测量中有10%的测量误差。假设对于给定的 X，$\tilde{X}$ 和 (Z, Y) 是条件独立的，即暴露测量误差是无差异的，$E(\tilde{X} \mid X) = X$，$\mathrm{Var}(\tilde{X} \mid X) = t^2\ \mathrm{Var}(X)$。反应变量 Y 与自变量 X、Z 之间的关系，以及与自变量 $\tilde{X}$、Z 之间的关系可以分别用下列表达式描述：

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

$$E(Y \mid \tilde{X}, Z) = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X} + \tilde{\beta}_2 Z$$

X 的测量误差对回归系数的影响可按式(1)表达[4]：

$$AF(\text{attenuation factor}) = \frac{\tilde{\beta}_1}{\beta_1} = \frac{1}{1 + t^2/(1-\rho^2)} \qquad (1)$$

式中，$\rho = \mathrm{Cor}(X, Z)$，即为 X 与 Z 的相关系数。

从上面 AF 的表达式(1)可以看出，在估计 X 对 Y 的影响时，如果研究分析中包括了一个准确测量的变量 Z，则测量误差所致的偏倚不仅与测量误差 t 的大小有关，而且与 X 与 Z 的相关系数的大小有关，除非 X 与 Z 没有相关性。

[举例] 图1是当 X 与 Z 的相关系数 ρ 分别为0、0.3、0.5、0.7、0.8、0.9和0.95时，AF 随测量误差 t 变化的轮廓图。从图1可以看出，不管相关系数 ρ 为多大，无差异性测量误差使所估计的 $\tilde{\beta}_1$ 低于实际的 β_1，并随着测量误差的增加，所致的偏倚程度也在扩大；在同样大小的测量误差条件下，随着 X 与 Z 的相关程度的增加，会进一步地恶化测量误差所致的偏倚。例如当测量误差 t 为10%时，在 $\rho = 0$，AF 为0.990，仅产生1.0%的可忽略的相对偏倚，但当 ρ 增加到0.9时，AF 降至0.950的水平，产生了5.0%的相对偏倚；当测量误差 t 为50%时，在 $\rho = 0$，AF 为0.800，产生20.0%的相对偏倚，当 ρ 增加到0.9时，AF 就降到0.432的水平，产生了56.8%的相对偏倚。

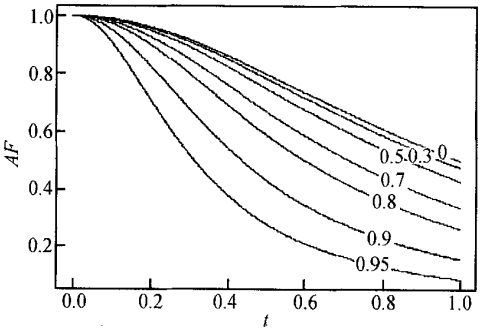

图1 不同相关系数条件下 AF 随测量误差 t 变化的轮廓图

以上分析的暴露测量误差对联系效应的影响仅涉及到一个准确测量变量，但在许多实际问题中，往往会涉及多个这样变量，这时测量误差对联系效应的影响就变得相当复杂。设 Y 和 X 为标量，Z 为 d 维的矢量，Y、Z_1，…，Z_d 能准确测量，X 不能准确测量，$E(X) = E(Z_i) = 0$，$\mathrm{Var}(X) = \mathrm{Var}(Z_i) = 1$，对于给定的 X，$\tilde{X}$ 和 (Z, Y) 是条件独立的，$E(\tilde{X} \mid X) = X$，$\mathrm{Var}(\tilde{X} \mid X) = t^2$。

X 的测量误差对线性回归系数的影响可按式(2)表达[4]：

$$AF = \frac{\tilde{\beta}_1}{\beta_1} = \left(\frac{1}{1+t^2}\right)\left\{1 - \left(\frac{t^2}{1+t^2}\right)c'\left(R - \frac{cc'}{1+t^2}\right)^{-1}c\right\} \qquad (2)$$

式中，$R = E(ZZ')$，$c = E(ZX)$。

由于多个准确测量变量存在的情况较为复杂，本文主要就自变量等相关的情况进行分析，即 $\rho = \mathrm{Cor}(X, Z_i) = \mathrm{Cor}(Z_i, Z_j)$，在这种情况下，式(2)就可以简化为式(3)表达。

$$AF = \frac{\tilde{\beta}_1}{\beta_1} = \left(\frac{1}{1+t^2}\right)\left[1 - \left(\frac{t^2\rho^2}{1+t^2}\right)\left\{\frac{d}{1-\rho} - \frac{d^2 c_{\rho,t}}{(1-\rho)(1-\rho+dc_{\rho,t})}\right\}\right] \qquad (3)$$

式中，$c_{\rho,t} = \rho - \{\rho^2/(1+t^2)\}$。

从式(3)可以看出，在估计 X 对 Y 的影响时，测量误差所致的偏倚不仅跟测量误差 t 的大小有关，而且与 X 同 Z_i 的相关系数的大小以及准确测量变量的个数 d 有关，除非 X 与 Z_i 没有相关性。

[举例] 图2是当相关系数 ρ 分别为0、0.5、

0.7、0.9，准确测量变量个数 d 分别为 1、2、3、5、20 时 AF 随测量误差 t 变化的轮廓图。从图 2 可以看出，当$\rho>0$时，不管相关系数 ρ 为多大，多少个准确测量变量，无差异性测量误差使所估计的 $\tilde{\beta}_1$ 低于实际的 β_1，并且随着准确测量变量个数增多，测量误差所致的偏倚也在增大，但增加的幅度是相当有限的。例如当测量误差 $t=0.3$和$\rho=0.5$时，准确测量变量个数 d 为 1、2、3、5、20 的 AF 值分别为0.893、0.881、0.874、0.866和0.853。然而，当$\rho<0$时，情况就变得相当复杂，测量误差所致的偏倚并不是全部随着准确测量变量的个数的增加而增大，也不是不管多少个准确测量变量，无差异性测量误差使所估计的 $\tilde{\beta}_1$ 都低于实际的 β_1（图 3）。

2. 分类变量：假设 Y 和 Z 为连续变量，并能准确测量，X 为二分变量，不能准确测量，存在无差异性错分，$\tilde{X}$ 为 X 观测值且满足下列方程式：

$$\Pr(\tilde{X}=1 \mid X, Z, Y) = a + bX$$

式中，$a = 1 - SP$（SP 暴露测量的特异度），$b = SN + SP - 1$（SN 暴露测量的灵敏度）。

反应变量 Y 与自变量 X、Z 之间的关系，以及与自变量 $\tilde{X}$、Z 之间的关系可以分别用下列表达式来描述：

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$$

$$E(Y \mid \tilde{X}, Z) = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X} + \tilde{\beta}_2 Z$$

X 的测量误差对回归系数的影响可按式（4）表达[4]：

$$AF = \frac{\tilde{\beta}_1}{\beta_1} = b\left[\frac{r(1-r)(1-\rho^2)}{\tilde{r}(1-\tilde{r}) - r(1-r)\rho^2 b^2}\right] = (SN + SP - 1)\left[\frac{r(1-r)(1-\rho^2)}{\tilde{r}(1-\tilde{r}) - r(1-r)\rho^2 (SN + SP - 1)^2}\right] \quad (4)$$

式中，$r = \Pr(X=1)$，$\tilde{r} = \Pr(\tilde{X}=1) = a + br$，$\rho = \mathrm{Cor}(X, Z)$。

从式(4)可以看出，在估计 X 对 Y 的影响时，测量误差所致的偏倚不仅跟暴露测量的灵敏度和特异度有关，且同 X 与 Z 的相关系数以及 X 的暴露比例有关，除非 X 与 Z 没有相关性。

［举例］ 图 4 是当 X 暴露比例 r 为0.1、0.5，特异度 SP 为0.6、0.9，相关系数 ρ 分别为 0、0.5、0.7、0.8、0.9时，AF 随灵敏度 SN 变化的轮廓图。从图 4 可以看出，随着相关系数的增加，AF 在逐渐减少，并且 X 与 Z 之间的这种相关关系，不仅可以影响暴露错分所致偏倚的程度，而且能改变其所致偏倚的方向。例如，当$r=0.1$，$SP=0.9$和$SN=0.9$时，ρ 为 0、0.5、0.7、0.8、0.9的 AF 值分别为1.707、1.419、1.077、0.819和0.474。

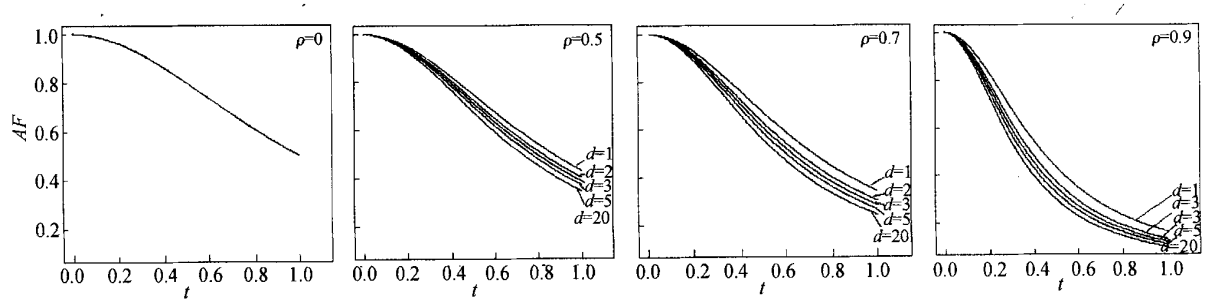

图2　不同相关系数（$\rho>0$）、不同准确测量变量数目条件下 AF 随测量误差 t 变化的轮廓图

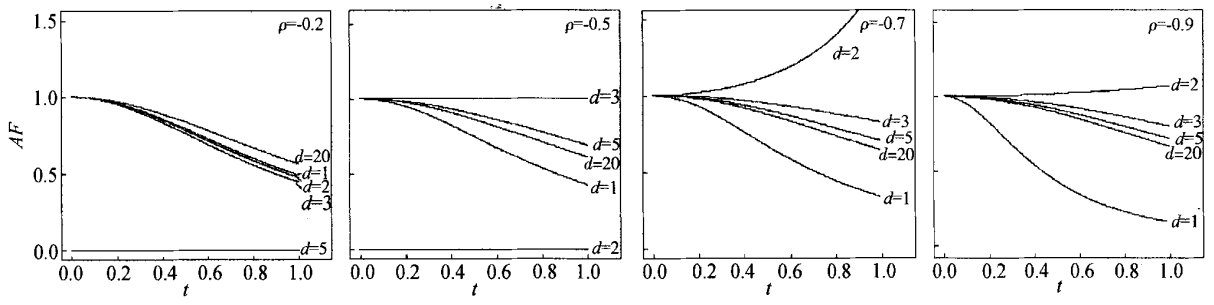

图3　不同相关系数（$\rho<0$）、不同准确测量变量数目条件下 AF 随测量误差 t 变化的轮廓图

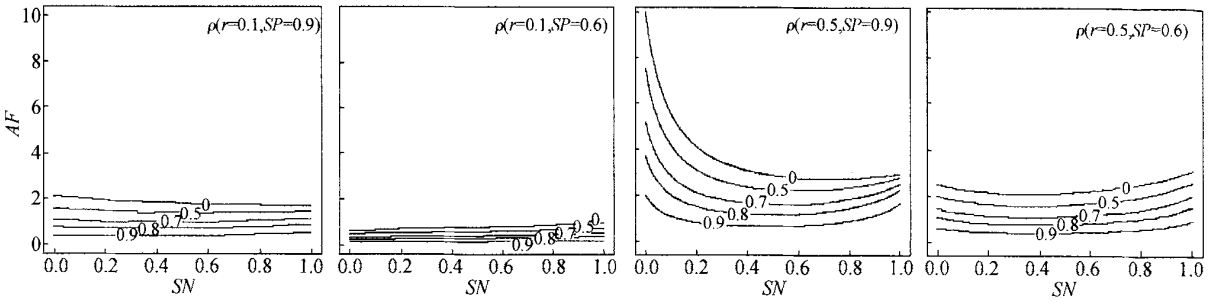

图4 不同暴露水平、不同相关系数及不同特异度条件下 *AF* 随灵敏度变化的轮廓图

讨 论

在一些研究中，有时会用线性回归模型来分析估计多个自变量与应变量之间的联系效应，但这些变量中一部分在实际情况中往往不能准确测量，或多或少存在着测量误差，而另一些是能准确测量的。以上实例结果表明，当测量准确的变量与测量不准确的变量共存时，测量误差所致的偏倚不仅与测量误差的大小有关，而且跟测量不准确变量与测量准确变量之间的相关程度、方向以及准确测量变量的个数有关，它不仅影响测量误差所致偏倚的程度，而且可能改变其所致偏倚的方向。例如在只有 2 个连续自变量的例子中，随着测量不准确变量与测量准确变量之间相关程度增加，会进一步地恶化测量误差所致的偏倚。当测量误差 t 为 10%时，在$\rho=0$，仅产生 1.0%可忽略的相对偏倚，但当$\rho=0.9$时，产生了5.0%的相对偏倚；当测量误差 t 为 50%时，在$\rho=0$，产生20.0%的相对偏倚，但当 ρ 增加到0.9时，产生的相对偏倚高达56.8%。这说明了如果测量不准确变量与测量准确变量之间不存在相关性，我们不必太担心测量误差所致的偏倚，如非这种测量误差相当大，然而如果它们之间存在相关性且相关程度较大时，则中等程度的测量误差所致的偏倚也将是很严重的，造成自变量与反应变量之间估计的联系效应值大大低于实际值。随着与测量不准确变量相关的准确测量变量个数的增加，情况就变得十分复杂。本文主要就自变量等相关的简单情况进行分析，结果显示，测量误差所致的偏倚不仅跟测量误差 t 和相关系数的大小有关，而且跟准确测量变量的个数及相关系数的符号有关。当$\rho>0$时，无差异性测量误差使所估计的 $\tilde{\beta}_1$ 总是低于实际的 β_1，并且随着准确测量变量个数增多，测量误差所致的偏倚也在增大，但增加的幅度是轻微的，然而，当 $\rho<0$时，情况就变得更为复杂，使无差异性测量误差所致偏倚的方向缺乏预测性，即所估计的 $\tilde{\beta}_1$ 可能高于也可能低于实际的 β_1。

分类变量的测量误差通常称为错误分类（错分），其中二分类变量测量误差的程度可用暴露测量的灵敏度和特异度来衡量。对于研究中既有错分的分类变量又有准确测量的变量的情况，本文只讨论了一个错分二分类变量 X 和一个准确测量的连续变量 Z 共同存在的情况，在估计错分的二分变量 X 对连续应变量 Y 的联系效应时，测量误差所致的偏倚不仅跟暴露测量的灵敏度和特异度有关，而且跟 X 与 Z 的相关系数以及 X 的暴露比例有关，并且随着相关系数的增加，*AF* 值逐渐减少。例如在低暴露水平$r=0.1$，暴露测量的灵敏度和特异度均较高为0.99，$\rho=0.5$时，二分变量 X 对连续应变量 Y 的联系效应估计值大于实际值，但当 ρ 增至0.9时，其联系效应估计低于实际值，改变了错分偏倚的方向。

本文仅对研究中只有一个自变量存在测量误差或错分的情况进行了讨论，但在实际研究中，远比此复杂，准确测量变量对测量误差所致偏倚的程度及方向的影响都是很难预见的。因此在既有准确测量变量又有测量误差变量的研究中，用线性回归模型来分析估计多个自变量与应变量之间的联系效应时，对测量误差所致偏倚的识别、控制和评估是十分必要的，对结果的解释要谨慎。

参 考 文 献

[1] 周艺彪，杨美霞，姜庆五. 暴露测量错分对研究真实性的影响. 中华流行病学杂志，2005，26：919-923.

[2] Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol, 1990, 132: 746-748.

[3] Veierod MB, Laake P. Exposure misclassification: bias in category specific Poisson regression coefficients. Stat Med, 2001, 20: 771-784.

[4] Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. New York: Chapman and Hall/CRC Press, 2003: 10-50.

（收稿日期：2006-06-30）
（本文编辑：张林东）