Two Paradoxes in Linear Regression Analysis

Ge FENG¹, Jing PENG², Dongke TU⁴, Julia Z. ZHENG⁵, Changyong FENG²,³*

Summary: Regression is one of the favorite tools in applied statistics. However, misuse and misinterpretation of results from regression analysis are common in biomedical research. In this paper we use statistical theory and simulation studies to clarify some paradoxes around this popular statistical method. In particular, we show that a widely used model selection procedure employed in many publications in top medical journals is wrong. Formal procedures based on solid statistical theory should be used in model selection.

Key words: Forward selection, backward elimination, univariate regression; multiple regression

[Shanghai Arch Psychiatry. 2016; 28(6): 355-360. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.216084]

1. Introduction

Linear regression is the most widely used statistical model in data analysis.[¹] Wide availability and ease of use of statistical software packages, such as SAS, SPSS and R make the linear regression accessible to people without any formal statistical training. Although wise use of statistical methods such as linear regression helps us, even novices, develop a better understanding of data and guide our decisions, it also causes confusion in interpretation of results and paradoxical findings. For example, we are often asked by our biomedical collaborators questions like “When I run the univariate regression of Y on the predictor X, the p-value is very small. However, if I add some other predictors in the model, X is not significant anymore. Why?” The same problem also occurs in logistic regression for binary outcome [²], log-linear regression for counting data [²], and Cox proportional hazards regression for survival data.[³]

A simple answer to this question is the different assumptions between the univariate and multiple regression models. However, this is not so meaningful for non-statisticians. This is discussed in Section 2.

In many medical studies, regression analysis involves a large number of independent variables, or predictors. Model selection is required to find the predictors that are significantly associated with an outcome, or dependent variable, of interest. Here is how the model selection was done in a recent paper published in JAMA Surgery[⁴]:

“The administrative database was then evaluated by means of univariate and multivariate logistic regression. First we identified variables that were associated (P < .20) with readmission, the dependent variable. These potential confounders were then entered in multivariate stepwise (backward elimination) logistic regression, with readmission as the dependent variable.
A logistic regression model was constructed to identify patient factors associated with readmission.”

This forward selection procedure as the first step to weed out “non-significant” predictors has become almost the gold standard for variable selection and has been used in many papers published in top medical journals. The key idea of this method is first to run a univariate regression on each predictor. If the p-value is less than some pre-specified level, for example 0.1, then the predictor is used in the multiple regression. Otherwise, the predictor is assumed to have no significant effect on the outcome. This method seems quite logical and intuitively meaningful. Indeed, it has been used and is still being used by the biomedical and other research communities. Is this a valid procedure?

In this paper we use linear regression analysis to show two paradoxes in regression analysis. In Section 2 we use some very basic theory to show how the univariate regression and multiple regression make different assumptions on the models. We use examples and simulation studies to show two paradoxes in regression analysis in Section 3. Section 4 briefly discusses the transitivity of correlation. Our results clearly invalidate the model selection procedure widely used in biomedical research.

2. Basic theory

Let \((Y, X_1, ..., X_p)\) be a random vector, where \(X_1, ..., X_p\) are called the covariates (independent variables), and \(Y\) is called the outcome (dependent variables). The regression of \(Y\) on \((X_1, ..., X_p)\) is the conditional expectation of \(Y\) given \((X_1, ..., X_p)\), denoted by \(E[Y|X_1, ..., X_p]\) which is a measurable function of \((X_1, ..., X_p)\). Denote the function by \(g(X_1, ..., X_p)\). Without knowing the joint distribution of \((X_1, ..., X_p, Y)\), in general, the form of \(g(X_1, ..., X_p)\) is unknown. In statistical analysis, we usually assume some mathematically tractable forms of \(g(X_1, ..., X_p)\). For example, the linear regression analysis assumes that

\[
g(X_1, ..., X_p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p
\]

In the logistic regression analysis with 0-1 outcome, we assume that

\[
g(X_1, ..., X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}
\]

In this paper we assume the outcome \(Y\) is continuous. Let

\[
e = Y - E[Y|X_1, ..., X_p]
\]

It is obvious that \(E[Y|X_1, ..., X_p] = 0\). We consider a stronger form of the linear regression model

\[
Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon
\]

and assume that given \(X_1, ..., X_p\), the variance of \(\varepsilon\)

\[
\text{Var}[\varepsilon|X_1, ..., X_p] = \text{Var}[Y|X_1, ..., X_p] = \sigma^2
\]

which does not depend on \((X_1, ..., X_p)\). This assumption is also used in most statistical literature on linear model. We further assume that \(X_k, k = 1, ..., p\), have finite second moments.

From (1) we have

\[
E[Y|X_1] = \beta_0 + \beta_1 X_1 + \sum_{k=2}^{p} \beta_k E[X_k|X_1].
\]

(2)

Let \(Z_k = E[X_k|X_1], k = 1, ..., p\). (It is clear that \(Z_k = X_k\).) Then the regression of \(Y\) on \(X_1\) is

\[
E[Y|X_1] = \beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p + \eta
\]

which still has a linear form. Let Then

\[
Y = \beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p + \eta
\]

(3)

Although (3) has the same form as (1), they are fundamentally different in the error terms. Note that \(E[\eta|X_1] = 0\), \(\text{Cov}(Z_k, \eta) = 0, k = 1, ..., p\). However, the conditional variance of \(\eta\) given \(X_1\) is

\[
\text{Var}[\eta|X_1] = \text{Var}[Y|X_1] = \sigma^2 + \text{Var}[\sum_{k=2}^{p} \beta_k X_k|X_1].
\]

Therefore, the conditional variance of \(\eta\) given \(X_1\) is no longer a constant. This violates the fundamental assumption used in linear regression model.

The univariate linear regression of \(Y\) on \(X_1\) assumes the following form of the model

\[
Y = \gamma_0 + \gamma_1 X_1 + \zeta
\]

(4)

From (3) we know that generally

\[
E[\zeta|X_1] = 0, \text{Cov}(\zeta, X_1) \neq 0.
\]

Suppose \((Y, X_{i1}, ..., X_{ip}), i = 1, ..., n\), is a random sample from (1). Let

\[
\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i, \quad \bar{X}_1 = n^{-1} \sum_{i=1}^{n} X_{i1}.
\]

Let \(\hat{Y}_1\) be the least square estimate of the univariate regression of \(Y_1\) on \(X_{i1}\) in (4). Then

\[
\hat{Y}_1 = \frac{\sum_{i=1}^{n} (X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2},
\]

and

\[
\text{Cov}(\hat{Y}_1, X_{i1}) = \frac{\sum_{k=2}^{p} \beta_k \text{Cov}(X_{i1}, X_k)}{\text{Var}(X_{i1})} \text{ a.s.}
\]

(5)

as \(n \to \infty\). Let \(\hat{\beta}_1\) be the least square estimator of \(\beta_1\) in (1). It is well known that \(E[\hat{\beta}_1] = \beta_1\) and \(\hat{\beta}_1 \to \beta_1\). Hence the estimates from the univariate regression and multiple regression usually converge to different limits. In a special case that and other covariates are uncorrelated, the limits are the same.
3. Two paradoxes in linear regression analysis

In this section, we show why the estimates of the coefficient of some covariates in the univariate regression and in the multiple regression do not match. More specifically, we show that in some cases, the estimate from the univariate regression is significant, but the result from the multiple regression is not. On the other hand, in some cases, the result is significant for the multiple regression but not for the univariate regression.

Suppose (1) is the true multiple regression model. The univariate regression model uses model (4) by assuming that \( E[X_k | X_1] = 0 \). This assumption is generally wrong unless \( E[X_k | X_1] \) is a constant \( (k = 2, \ldots, p) \). Hence, with a correct multiple regression model, the estimate of the univariate analysis is based on a wrong model. This is the reason why the results from univariate regression and multiple regression do not match. Furthermore, result (5) shows that there is no clear interpretation of the estimate in the univariate analysis.

We discuss two paradoxes related to univariate and multiple regressions through both theoretical derivations and simulation studies.

3.1 Significant covariate effect in multiple regression but not in univariate regression

Let \( X_1, X_2, X_3 \) and \( \varepsilon \) be independent random variables with standard normal distributions. Consider the following model

\[
Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon
\]  

(6)

where \( \alpha_k \neq 0, k = 0,1,2,3 \), and \( X_1 = \beta_1 X_1 + \beta_2 X_2 \).

Then

\[
\frac{\text{Cov}(Y, X_1)}{\text{Var}(X_1)} = \alpha_1 + \frac{\alpha_2 \beta_1}{\beta_1^2 + \beta_2^2}
\]

which is 0 if and only if

\[
\alpha_1 = -\frac{\alpha_2 \beta_1}{\beta_1^2 + \beta_2^2}.
\]  

(7)

From (5) we know that if (7) is true, the least square estimator \( \hat{\beta}_1 \) of the coefficient of the univariate regression of \( Y \) on \( X_1 \) will not be significant, even though \( X_1 \) is necessary in specifying model (6).

**Example 1.** Let \( \alpha_1 = -3/5, \alpha_2 = 3, \alpha_3 = 4, \beta_1 = 1, \beta_2 = 2 \) in (6).

The true model is

\[
Y = 1 - 3/5X_1 + 3X_2 + 4X_3 + \varepsilon
\]  

(8)

Table 1 shows the simulation result of the estimates and standard deviations of the coefficient of \( X_1 \) in both univariate and multiple regressions after 10,000 replications. For a wide range of sample sizes, the least square estimator of the coefficient of \( X_1 \) in the multiple regression is very close to the true value, and the standard deviation decreases significantly with the sample size. However, the estimate of coefficient in the univariate analysis is very close to 0 in all cases.

According to the practice in medical publications \([4-24]\), \( X_1 \) will not enter the multiple regression. Table 2 shows the result of the least square estimates of the coefficients of \( X_2 \) and \( X_3 \) after \( X_1 \) is removed in (8). It is easy to see that the estimate of the coefficient of \( X_i \) is dramatically biased in the multiple regression after \( X_1 \) is removed due to the univariate analysis.

3.2 Significant covariate effect in univariate regression but not in multiple regression

Suppose \( X_1, X_2, X_3 \) and \( \varepsilon \) are independent standard normal random variables, and \( X_4 = \beta_1 X_1 + \beta_2 X_2 \), where \( \beta_1 \beta_2 \neq 0 \).

**Table 1. Estimate of the regression coefficient of \( X_1 \)**

| n   | Multiple regression | Univariate regression |
|-----|---------------------|-----------------------|
|     | Estimate SD         | Estimate SD           |
| 30  | -0.6010 0.0988      | -0.0005 0.4225        |
| 50  | -0.6003 0.0748      | -0.0016 0.3194        |
| 100 | -0.6003 0.0514      | -0.0009 0.2226        |
| 200 | -0.6002 0.0357      | 0.0002 0.1585         |
| 500 | -0.6005 0.0226      | -0.0005 0.0965        |
| 1000| -0.6000 0.0160      | -0.0002 0.0691        |

**Table 2. Estimates of the regression coefficients of \( X_2 \) and \( X_3 \) with \( X_1 \) being removed**

| Coefficient of \( X_2 \) | Coefficient of \( X_3 \) |
|---------------------------|---------------------------|
| \( \alpha_i = 3 \)       | \( \alpha_i = 4 \)       |
| n   | Estimate SD         | Estimate SD         | n   | Estimate SD         | Estimate SD         |
|-----|---------------------|---------------------|-----|---------------------|---------------------|
| 30  | 2.4074 0.3030       | 4.0028 0.3047       | 30  | 2.4074 0.3030       | 4.0028 0.3047       |
| 50  | 2.3990 0.2281       | 4.0014 0.2302       | 50  | 2.3990 0.2281       | 4.0014 0.2302       |
| 100 | 2.4020 0.1611       | 3.9992 0.1581       | 100 | 2.4020 0.1611       | 3.9992 0.1581       |
| 200 | 2.3999 0.1111       | 4.0019 0.1126       | 200 | 2.3999 0.1111       | 4.0019 0.1126       |
| 500 | 2.4002 0.0703       | 4.0005 0.0705       | 500 | 2.4002 0.0703       | 4.0005 0.0705       |
| 1000| 2.4002 0.0498       | 3.9993 0.0492       | 1000| 2.4002 0.0498       | 3.9993 0.0492       |
Consider the following true model is
\[ Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon \]  \quad (9)

If (9) is expanded to include \( X_4 \) and the expanded model still satisfies the conditions of the linear regression, then the regression equation becomes
\[ Y = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \varepsilon \]  \quad (10)

From (9) and (10) we have
\[ E[Y|X_1, X_2] = E[Y|X_1, X_2, X_3] \]

or
\[ \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 = \delta_0 + (\delta_1 + \delta_3 \beta_1) X_1 + \delta_2 X_2 + \delta_3 \beta_3 X_3. \]

Since \( \beta_3 \neq 0 \), we should have \( \delta_3 = 0 \), which means that \( X_3 \) has no role in the multiple regression. Let \( \hat{\beta} \) be the least square estimate of the coefficient of univariate linear regression of \( Y \) on \( X_3 \). Then
\[ \hat{\beta} \rightarrow \frac{\text{Cov}(Y, X_3)}{\text{Var}(X_3)} = \frac{\alpha_1 \beta_1}{\beta_1^2 + \beta_3^2}. \]

Hence if \( \alpha_1 \beta_1 \neq 0 \), when sample size \( \delta_3 = 0 \) is large enough, the result from the univariate is significant but the multiple regression is not.

**Example 2.** Let \( \alpha_0 = 0, \alpha_1 = 1, \alpha_2 = 2 \) in (9) and \( \beta_1 = \beta_3 = 1 \), Table 3 shows the least square estimates of the coefficient of \( X_3 \) in both univariate and multiple linear regressions after 10,000 replications. For all sample sizes, the univariate regression shows that \( X_3 \) has very significant effect on \( Y \). However, in the multiple regression, the effect is not significant.

### 4. Transitivity of correlation

Another issue around the regression analysis is the transitivity of the correlation in the interpretation.

**Table 3.** Estimate of the regression coefficient of \( X_4 \)

| \( n \) | Univariate regression | Multiple regression |
|-------|-----------------------|--------------------|
|       | Estimate | SD   | Estimate | SD   |
| 30    | 1.0024   | 0.4723 | 0.0038  | 0.2014 |
| 50    | 0.9975   | 0.3564 | -0.0008 | 0.1496 |
| 100   | 0.9995   | 0.2469 | -0.0015 | 0.1032 |
| 200   | 0.9982   | 0.1733 | 0.0005  | 0.0723 |
| 500   | 0.9999   | 0.1101 | 0.0005  | 0.0452 |
| 1,000 | 0.9995   | 0.0776 | 0.0004  | 0.0318 |

For example, some people may say like that: “Since factor \( A \) is highly correlated with outcome \( Y \), and factor \( A \) and factor \( B \) are highly correlated, then \( B \) should be correlated with \( Y \).” It seems very intuitive and reasonable that correlation is transitive. Unfortunately, this is not true. Here is a theoretical example. Suppose \( X \) and \( Z \) are independent standard normal random variables and \( Y = X + Z \). It’s clear that the correlation between \( X \) and \( Y \), and between \( Y \) and \( Z \) are both 0.707. However, the correlation between \( X \) and \( Z \) is 0.

In our Example 2, the correlations between \( X_1 \) and \( X_1, Y \) are 0.707 and 0.408, respectively. However, we proved in Section 3.2 shows that \( X_1 \) has no role in the multiple regression if \( X_1 \) and \( X_2 \) are in the model although \( X_4 \) is not a linear combination of \( X_1 \) and \( X_2 \).

### 5. Discussion

Regression analysis in medical research usually involves many predictors (independent variables). The model selection is needed to pick covariates having significant effect on the outcome. A widely used method in medical publications\(^{[4-24]}\) is first to screen those covariates through univariate analysis. If a covariate is not significant in the univariate regression analysis, it will not enter the multiple regression analysis. The underlying assumption of this method is that a covariate is significant in the multiple regression only if it is significant in the univariate regression analysis. Our results indicate that this assumption is wrong. A covariate may be very significant in the univariate regression but has no role in the multiple regression (see Example 2 in Section 3). On the other hand, a covariate is a necessary part of a multiple regression but may be not correlated with the outcome (see Example 1 in Section 3). The initial univariate screening method totally ignores the correlation among covariates. There is no theoretical work to support this method. Our simulation results clearly show that the multiple regression results after the univariate screening may be dramatically biased and misleading. The biomedical community should stop using this procedure in their research and publications.

**Funding**

None

**Conflict of interest statement**

The authors report no conflict of interest related to this manuscript.

**Author’s contribution**

Ge Feng and Changyong Feng: theoretical derivation and revision

Jing Peng, Dongke Tu, and Julia Z. Zheng: Simulation and manuscript drafting
线性回归分析中的两个悖论
Feng G, Peng J, Dongke TU, Zheng JZ, Feng C

概述：回归是应用统计学中最受欢迎的工具之一。然而，回归分析结果的误用和误解在生物医学研究中是常见的。本文运用统计理论和模拟研究来说明有关这种普遍使用的统计方法的一些悖论。我们还特别指出在顶级医学期刊发表的很多文章中广泛使用的模型选择程序事实上是错误的。模型选择使用哪一种步骤化程序需基于可靠的统计理论。

关键词：向前选择，向后消除，单变量回归，多元回归

References:
1. Seber GAF, Lee AJ. Linear regression analysis (2nd ed). Hoboken, NJ: Wiley; 2003
2. Agresti A. Categorical data analysis (2nd ed). Hoboken: Wiley; 2002

2. Cox DR. Regression models and life-tables (with discussion). J R STAT SOC. 1972; B. 34:187-220. doi: http://dx.doi.org/10.2307/2985181
3. McIntyre LK, Arbabi S, Robinson EF, Maier RV. Analysis of Risk Factors for Patient Readmission 30 Days Following Discharge From General Surgery. JAMA Surgery. 2016; (Epub ahead of print). http://dx.doi.org/10.1001/jamasurg.2016.1258
4. Bardia A, Sood A, Mahmood F, Orhurhu V, Mueller A, Montealegre-Gallegos M, et al. Combined epidural-general anesthesia vs general anesthesia alone for elective abdominal aortic aneurysm repair. JAMA Surgery. 2016; (Epub ahead of print). http://dx.doi.org/10.1001/jamasurg.2016.2733
5. Barlesi F, Mazieres J, Merlio JP, Debieuvre D, Mosser J, Lena H, et al. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). Lancet. 2016; 387: 1415-1426. doi: http://dx.doi.org/10.1016/S0140-6736(16)30004-0
6. Brooks GA, Kansagra AJ, Rao SR, Weitzman JJ, Linden EA, Jacobson JO. A clinical prediction model to assess risk for chemotherapy-related hospitalization in patients initiating palliative chemotherapy. JAMA Oncology. 2015; 1(4): 441-447; doi: http://dx.doi.org/10.1001/jamaoncol.2015.0828
7. Cronin PR, DeCoste L, Kimball AB. A multivariate analysis of dermatology missed appointment predictors. JAMA Dermatology. 2013; 149(12): 1435-1437. doi: http://dx.doi.org/10.1001/jamadermatol.2013.5771
8. Eighty T, Kerklaan D, Mesotten D, Verbruggen S, Wouters PJ, Vanhorebeek I, et al. Early versus late parenteral nutrition in critically ill children. N Engl J Med. 2016; 374(12): 1111-1122. doi: http://dx.doi.org/10.1056/NEJMoia1514762
9. Geng E, Kreiswirth B, Burzynski J, Schluger NW. Clinical and radiographic correlates of primary and reactivation tuberculosis: a molecular epidemiology study. JAMA. 2005; 293(22): 2742-2745. doi: http://dx.doi.org/10.1001/jama.293.22.2740
10. Hole J, Hirsch M, Ball E, Meads C. Music as an aid for postoperative recovery in adults: a systematic review and meta-analysis. Lancet. 2015; 386: 1659-1671. doi: http://dx.doi.org/10.1016/S0140-6736(15)60169-6
11. International CLL-IPI working group. An international prognostic index for patients with chronic lymphocytic leukaemia (CLL-IPI): A meta-analysis of individual patient data. Lancet Oncology. 2016; 17(6): 779-790. doi: http://dx.doi.org/10.1016/S1470-2045(16)30029-8
12. Leon MB, Smith CR, Mack MJ, Makkar RR, Svensson LG, Kodali SK, et al. Transcatheter or surgical aortic-valve replacement in intermediate-risk patients. N Engl J Med. 2016; 374(17): 1609-1620. doi: http://dx.doi.org/10.1056/NEJMoa1514616
13. Li Y, Stocchi L, Cherla D, Liu X, Remzi FH. Association of preoperative narcotic use with postoperative complications and prolonged length of hospital stay in patients with crohn disease. JAMA Surgery. 2016; 151(8): 726-734. doi: http://dx.doi.org/10.1001/jamasurg.2015.5558
14. Lorant V, Deliège D, Eaton W, Robert A, Philippot P, Ansseau M. Socioeconomic Inequalities in Depression: A Meta-Analysis. Am J Epidemiol. 2003; 157(2): 98-112. doi: http://dx.doi.org/10.1093/aje/kwf182
15. van der Meer AJ, Veldt BJ, Feld JJ, Wedemeyer H, Dufour JF, Lammet F, et al. Association between sustained virological response and all-cause mortality among patients with chronic hepatitis C and advanced hepatic fibrosis. JAMA. 2012; 308(24): 2584-2593. doi: http://dx.doi.org/10.1001/jama.2012.144878
16. Mingrone G, Panunzi S, De Gaetano A, Iaconelli A, Nanni G, et al. Bariatricmetabolic surgery versus conventional medical treatment in obese patients with type 2 diabetes: 5 year follow-up of an open-label, single-centre, randomized controlled trial. Lancet. 2015; 386: 964-973. doi: http://dx.doi.org/10.1016/S0140-6736(15)00075-6
17. Nelson KB, Ellenberg JH. Antecedents of cerebral palsy: I. univariate analysis of risks. Am J Dis Child. 1985; 139(10): 1031-1038. doi: http://dx.doi.org/10.1001/archpedi.1985.02140120077032
18. Nelson KB, Ellenberg JH. Antecedents of cerebral palsy: Multivariate analysis of risk. N Engl J Med. 1986; 315(2): 81-86. doi: http://dx.doi.org/10.1056/NEJM19860703150202
19. NICE-SUGAR Study Investigators. Hypoglycemia and risk of death in critically ill patients. N Engl J Med. 2012; 367(12): 1108-1118. doi: http://dx.doi.org/10.1056/NEJMoa1204942
20. Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molitor R, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. N Engl J Med. 2005; 353(25): 2654-2666. doi: http://dx.doi.org/10.1056/NEJMoa051424
Ge Feng is a graduate student in the School of Geophysics and Oil Resources at Yangtze University, Wuhan, Hubei, China. His research interest includes statistical analysis in rock physics.