第46卷 第3期 气 象 学 报 Vol. 46, No. 3
1988 年 8 月 ACTA METEOROLOGICA SINICA Aug., 1988

气象数据回归分析中的若干问题及其对策*

俞善贤 陈孝源

(浙江省气象科学研究所)

提 要

本文对回归分析在气象应用中存在的若干问题进行了讨论,分析了有些失误的主要原因。采用诊断方法可以发现这些问题,然后分别采用不同的对策,初步使用这些对策后,实际预报正确性和稳定性都有明显的提高。

一、引 言

统计分析在气象业务应用中取得了一定的成效。但也存在不少问题,造成了一些不良后果,1979 年李麦村提出了统计预报中应注意的若干问题[1]。近年来随着微机的推广使用,统计应用更加广泛,在新情况下也产生了一些新问题。本文对回归分析中的一些问题进行初步评述和探讨,并提出某些对策。

二、变 量 选 取

1. 相关系数

目前建立回归方程所用的因子,往往从几千个因子中进行相关普查,以单相关系数高的因子作为预报因子。这种做法中应注意相关系数稳定性、伪相关和遗漏好因子的问题。

光用单相关系数这个统计量是不能很好说明因子与预报对象间的相互依赖关系的。Anscombe 于 1977 年构造了四组数据[2]。它们的单相关系数都等于 0.8167,由此建立的一元线性回归方程的各项统计量也都一样,但从图 1 中可以看出各组之间的性质有很大的不同。图 1 a 用单相关系数来反映依赖关系是比较合理的,图 1 b 用一条光滑曲线来反映依赖关系比较合理,图 1 c 中大多数点的直线依赖关系很好,唯独A点离得较远,若去掉它,相关系数将提高很多,图 1 d 中我们不能认为 x 与 y 有某种依赖关系。

在气象业务中,常会遇到相关系数(或回归系数)的大小取决于某个别点,去掉或增加某个别点,相关系数(或回归系数)就会产生很大的变化,这些强烈影响统计量的点称为强影响点。如图 1 中的A,B点。统计分析中我们期望对每一个数,既对参数估计或预测有一定影响,但影响又不太大,如果某数据影响过大、那么包含这组数据的统计量和不包含

* 本文于 1986 年 7 月 14 日收到,于 1987 年 3 月 3 日收到修改稿。

这组数据的统计量差异就很大，考虑到数据获得具有一定的随机性，于是包含有较大影响的数据所导出的统计推断就不够稳定。

在点聚图上有时可以发现这些强影响点，同时还可以清楚地反映非线性相关问题。这种传统方法还应该保持，不要因为有了相关系数而抛弃它。在终端上显示点聚图是一种经济、简便的做法。

用秩相关系数，可以克服受个别点影响所导致的不稳定，但计算秩相关系数时会损失部分"信息"。

用单相关系数不能很好反映非线性关系，有时尽管单相关系数很低，但存在很好的非线性关系，此时可以对因子作适当的函数变换。文献[3]中，涡度与雨量之间单相关系数只有0.26，用函数$s(x)=k/[a+\exp(bx)]$对涡度作变换后相关系数达0.56，且这种变换有明显的天气学意义。

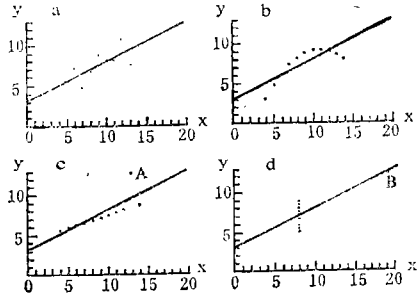

图 1 Anscombe 数据点聚图

2. 刀切法(Jackknife)[4]

此方法是20世纪50年代后期提出的。它的基本思想是在统计样本中，去掉一组或一个样本作出的统计量与不去掉样本作出的统计量进行对比，反映样本对统计量的影响程度，然后估计这个统计量精度值。

我们在浙北地区梅雨期长度预报模型中[5]、在选用因子时对单相关系数进行刀切法试验、当去掉1954年样本时(梅雨期为78 d、特大年)一个因子的相关系数从−0.5427变到−0.2262，另一个因子从0.4996下降到0.4707，可见前者强烈依赖于1954年的样本，后者则影响不大。发现这一问题时，我们还不能肯定前者的相关是伪相关、重要的是要特别注意这一情况，分辨出是一个偶然的巧合，还是有一定物理基础的结果。这时作留用或剔除的决策，如果一时找不到原因，就分别建立留用和剔除的模型进行对比，找出较好的模型。

三、回归分析中的一些问题

1. 逐步回归

逐步回归方法是最常用的方法之一。这个方法表面上看来比较"完美"，但在实际应用中和理论上都发现有不足之处[2,6]：(1) 在选入或剔除变量时的F检验除了某些特殊的在应用上不太现实的条件外，不能认为涉及的F检验是正确的，在理论上并不能以任何概率保证所选变量的"显著性"；(2) F值不是单调下降，有时起伏较大，这时控制变量选入的F_1或剔除的F_2稍一变化，方程含有因子的个数就会有很大的变化，在选用方程时带来了困难，似乎缺乏一个客观标准；(3) 逐步回归方法决定的变量子集，可能比包含同样多自变量子集的残差平方和要大得多。由于逐步回归选择子集的变量又是包含关系，就很难选入二个或更多变量相互配置才发生较大作用的变量。存在一个回归变量y和两个

回归因子 x_1, x_2的例子。y 与 x_1之间的单相关系数为 0.104，y 与 x_2之间的单相关系数为 −0.00635。由于逐步回归不能同时考虑 x_1和 x_2的配置作用，所以 x_1和 x_2就很难选入方程，然而 y 对 x_1和 x_2回归的复相关系数为 0.997，是高度相关的。用最优子集回归、逐步向后法可以克服上述缺点。

尽管逐步回归存在上述提出的几个问题，但其算法简便，能处理较多因子，已被许多实际工作者掌握，所以至今还是一个广泛使用的算法。我们指出它的不足，其目的是希望使用者对这些问题做到心中有数，不要受此法框框限制，认为它得到的方程是最优的。

在古典回归中，衡量一个回归方程的好坏往往只用残差平方和为标准。而残差平方和只能衡量拟合的好坏，不能完全反映预测能力的高低。如何在众多的方程中选择一个预测能力较强的方程是一个非常实际又难以解决的问题。我们认为：首先应对方程进行诊断，发现其中存在的问题，如方程的基本假定是否满足，强影响点是否存在，回归系数估计是否正确等，针对不同问题分别施以不同对策；其次从不同的目标出发，考察各项统计性能，综合评价，合理选用。有理由期望这样选用的方程比光用残差平方和为指标的方程具有较强的预测能力。作者在文献[6]中，试用最优子集与岭迹分析相结合的方法确定回归方程，选用残差平方和较小，同时回归系数估计比较稳定的方程，得到的方程比逐步回归为优。

2．建立回归方程的几个目标函数[2]

1）着眼于残差平方和最小的最优子集回归。此方法就是在所有相同因子个数的方程中选出残差平方和最小者。此方法当因子个数小于 10 个时一般微机还适用，当超过14个时微机上就很难实现。如何解决这个矛盾，可以用经验判断和降低逐步回归的 F 检验值达到精选因子的目的，当然这只是一个解决因子过多的权宜之计。在浙江省早稻产量预报中[7]，作者采用此方法，精选 12 个因子作最优子集回归，残差平方和下降 18.1%，3 年预报精度明显提高。

2）着眼于回归系数估计精度的岭回归方法。最小二乘法估计的回归系数 $\hat{\beta}$，在理论上具有无偏性，但当因子间存在复共线关系时，估计的回归系数就不够稳定，可产生较大的偏差。用均方误差来衡量 $\hat{\beta}$ 作为β估计是否良好指标，即 $\mathrm{MSE}(\hat{\beta})=\mathrm{E}[11\,\hat{\beta}-\beta 11^2]$。可以证明，岭回归估计的回归系数比最小二乘的均方误差要小。使用结果表明：岭回归能提高预测能力[8]。

3）着眼于预测平方和的 PRESS 准则。此方法是刀切法思想在回归中的一个推广，就是在N个样本中除去第 i 个样本，建立方程，然后来试报这点的值 $\hat{y}_i$，得试报误差 $f_i=y_i-\hat{y}_i$，作目标函数 $\mathrm{PRESS}=\sum_{i=1}^{N} f_i^2$，它的直观意义很明确，用 PRESS 值可衡量方程的预测能力，以最小者为最优。这个准则就是在由因子可能产生的一切子集中（共 2^p-1 个）选取最小者。当因子较多时，计算量大得惊人。为此作者提出了一种近似的快速算法，用实例计算得出的解绝大多数是最优解，计算量比原算法可减少98%以上，在微机上能处理较多的因子。具体算法准备另文讨论。

4）着眼于预测均方误差为最小的 MSEP_x 准则。对每个给定预测点，在该处的预测

偏差平方均值只与所选自变量有关，$MSEP_x$ 准则在于选择这样的自变量子集使得在预测点的预测偏差平方的均值达到最小，这样针对不同的预测点，所建立的预测方程也是不同的。作者在文献[5]中试用此法预测梅雨期长度，其结果比逐步回归为优，特别是峰值年份试报较好。

四、稳 健 回 归

1．稳健回归研究的问题

稳健回归主要解决当样本中含有少量“奇异点”(Outlier，是强影响点中的一类) 时估计的回归系数不应受过大的影响。所谓“奇异点”是指由种种原因使观察到的结果特大或特小，而在样本中显得很突出，或者象文献[1]中指出的情况：“在大量晴和小雨的样本中偶尔出现一两次特大暴雨，其距平必然特别大。这样突出的样本有扰动整个系列的作用。”这样的统计样本用最小二乘法来建立方程，往往只是对少数样本的描述，得到的方程是不稳定的。

以最小二乘为主体的方法，由于以残差平方和为目标函数，就会使奇异值的作用显著增加，使回归系数估计产生较严重的偏差，就会影响一些正常点，这种情况对两者都是不利的，因为都是对真实情况的歪曲。解决这一问题的常用方法是剔除这些奇异点或改变目标函数，用一个比 x^2 增长慢的函数来代替，常见的有 L_1 和 M 估计。

2．最小绝对偏差 L_1 估计[9]

对线性回归模型，如果

$$\min_{\beta}\sum_{i=1}^{N}|y_i-x_i'\beta| = \sum_{i=1}^{N}|y_i-x_i'\hat{\beta}_{L_1}|$$

则称 $\hat{\beta}_{L_1}$ 为 β 的最小绝对偏差估计或 L_1 估计。拉普拉斯也建议人们使用 L_1 估计。由于计算复杂，一直得不到发展。1955 年证明了 L_1 估计与一特殊的线性规划问题等价，从根本上解决了计算问题。L_1 估计现已十分受重视。1977 年美国统计杂志 Communication in Statistics 出专辑介绍 L_1 估计方法和一些新发展。

3．M 估计[10]

对线性回归模型，如果

$$\min_{\beta}\sum_{i=1}^{N}\rho(y_i-x_i'\beta)=\sum_{i=1}^{N}\rho(y_i-x_i'\hat{\beta}_M)$$

则称 $\hat{\beta}_M$ 为 β 的M估计，其中 $\rho(x)$ 增长速度比 x^2 为慢的函数。文献[2]给出了 $\rho(x)$ 具体表达式和计算方法。作者在富阳县早稻年景长期预报中[11]，用 M 估计的方程试作 2 年，结果比逐步回归为好。

五、线 性 回 归 诊 断

在以上讨论中我们已指出，仅用古典回归分析得到的统计量，受强影响点作用时可导

致统计推断失误。文献[12]列举的一实例、仅仅一个样本就导致回归分析的错误结论，并且在拟合残差上反映不出来。如何来发现这些强影响点，就必须靠回归诊断工具。它主要包含两个内容：(1) 分析每个样本对回归系数估计所起作用大小和对预测的影响；(2) 考察回归模型基本假定正确性，如正态性、线性性等是否满足，如果不满足又该如何修正模型，分别施以各种“治疗方案”，比较哪一种方案可使修正后满足这些基本假定。文献[12]中对模型诊断的原理、方法作了详细介绍，所用的方法比较朴素、计算量也不大，可在微机中实现。

在此我们说明一下对强影响点如何处理问题，原则上讲没有也不可能有统一处理的公式，也不能断然说包含这样数据的回归是完全不可取的，重要的是：分辨出哪些是强影响点，然后根据专业知识，数据收集的实际情况，分析出产生的原因，合理处理。(1) 如果强影响点来自数据收集过失，或其它特殊原因的影响，或不符合研究问题的一些基本假定，则应剔除。例如，“研究的目的在于探讨作物产量形成的各生育期与气象条件的关系，如果是非气象原因所造成的减产，这样的材料应该去掉。另一情况虽属气象原因造成的减产，如一场雹灾，一次毁灭性低温冷害，一场大水冲毁等天灾，使得全年产量毁于一旦。这种气象灾害应属农业气象灾害的研究范畴，这些特殊样本应去掉”。[13] (2) 如果找不出产生强影响点的原因，那么可以对包含和剔除强影响点两种情况作模型的估计或预测加以比较，或收集更多的数据重做估计和诊断。对这种情况的处理应十分小心，往往能从这些数据中提取其它数据不能提供的信息。(3) 如果主要目的在于预测，可以对预测区域作分析，分离出影响较大或较小的预测区域，在最后的预报集成时提供有用的信息。

我们强调模型诊断同文献[1]中所强调的要有严格的统计检验一样重要，有时可能更重要，因为当强影响点作用时作出的统计量，即使用严格的统计检验，还会导致错误的统计推断。

六、结　束　语

统计分析是一个有用的工具，但要使用适当。实际中有时一个简单模型可以做出好的结果，而一个复杂模型却得不到好结果。如何选用好统计工具确是一门“艺术”，要掌握这门艺术，首先应对所研究问题的内在联系有深刻的认识，其次借助于统计方法本身的手段，帮助人们提供各种有用的信息，综合各方面的情况，合理决策。避免滥用统计工具。这样做有可能减少统计推断的错误，提高预报水平。

参 考 文 献

[1] 李麦村，关于统计预报的若干问题，气象，1979 年，第 5 期。

[2] 陈希孺、王松桂，近代实用回归分析，广西人民出版社，1984。

[3] 李法然等，MOS 预报业务化试验中若干技术问题的处理，气象，1986 年，第 10 期。

[4] Ruper, G. M. 著，魏宗舒译，刀切法的评述，应用数学与计算数学，1980 年，第 5 期。

[5] Yu Shanxian, Chen Xiaoyuan, A statistical model for predicting the duration of Meiyu based on minimizing mean square error of prediction, 3-rd international conference on statistical climatology, Vienna, June, 1986.

[6] 俞善贤、汪铎，试用最优子集与岭迹分析相结合的方法确定回归方程，大气科学(待发表)。

[7] 俞善贤、田清鉴，浙江省早稻产量预报的最优子集回归模型，科技通报，1987 年，第 4 期。

[8] 冯耀煌、吴达三，岭回归在预报集成中的应用，气象，1985年，第11期。
[9] 王松桂，线性模型参数估计的新进展，数学进展，1985年，第3期。
[10] Huber, P. J., Roust statistics, John Wiley, New York, 1981.
[11] 俞善贤、赵锡林，稳健回归在产量预报中的应用，农业气象，1987年，第1期。
[12] 王松桂，线性回归诊断，数理统计与管理，1985年第6期和1986年第1期。
[13] 魏淑秋，农业气象统计，297—298，福建科学技术出版社，1985。

SOME PROBLEMS AND TREATMENTS IN REGRESSION ANALYSIS FOR METEOROLOGICAL DATA

Yu Shanxian Chen Xiaoyuan

(Zhejiangl Research Institute of Meteorological Science)

Abstract

In this paper, some problems of regression analysis for meteorological application are discussed. The reasons which lead to statistical inference failure are analysed and can be found by diagnostic method. The failure problems can be resolved. The successes have been made for the forecasting accuracy and stability from the treatment.