中国科学: 数学 2025 年 第 55 卷 第 ? 期: 1 ∼ 14
SCIENTIA SINICA Mathematica
综 述

《中国科学》杂志社
SCIENCE CHINA PRESS
CrossMark click for updates

回归分析中的两类假设检验问题

张新瑜, 史延美, 郭旭*

北京师范大学统计学院, 北京 100875

E-mail: zhangxinyu1@mail.bnu.edu.cn, 202331011015@mail.bnu.edu.cn, xustat12@bnu.edu.cn

收稿日期: 2024-04-22; 接受日期: 2024-10-25; 网络出版日期: 2024-11-19; * 通信作者
国家重点研发 (批准号: 2023YFA1011100 和 2023YFA1008702)、国家自然科学基金 (批准号: 12322112 和 12071038) 和中央高校基本科研业务费专项资金 (批准号: 2243200006) 资助项目

摘要 本文对回归分析中的两类假设检验问题, 即模型设定检验和变量显著性检验, 进行简要介绍, 并描述针对这两类检验问题的经典的和近期所发展的检验方法. 进而介绍相关统计量的构造原理, 并对不同方法可能存在的问题进行评述. 经典的方法主要针对低维数据, 往往受自变量维数的影响较大. 近期所发展的方法主要适用于高维数据, 旨在有效克服这两类检验问题中的维数灾难难题. 机器学习算法的使用是近期所发展的检验方法的重要特征. 最后对一些可能的议题进行简单讨论.

关键词 模型设定检验 变量显著性检验 维数灾难 机器学习

MSC (2020) 主题分类 62G10, 62G20, 62J12

1 引言

回归分析是统计学的一个核心内容, 可用以探究自变量和响应变量之间的关系. 在回归分析中, 有两个关键问题: 一个是检验所用的模型是否设定正确, 另一个是检验感兴趣的自变量在控制其他自变量时是否对响应变量起作用. 前者通常被称为模型设定检验, 而后者则被称为变量显著性检验. 本文将介绍这两类问题的研究进展.

在实际中, 以线性回归模型为典型代表的参数回归模型由于其简单性和可解释性被广泛应用于各个领域. 当参数回归模型设定正确时, 实际工作者可以得到易于理解且精确的推断结果. 但当参数回归模型设定错误时, 则难以得到正确的统计分析结果, 对应用领域产生误导. 因而在采用参数回归模型进行数据分析之前, 先进行模型设定检验是非常必要和重要的. 对此, 文献中已有很多检验方法. 针对低维数据的工作包括文献 [12, 27, 42, 45, 61]. 近年来随着数据维数的提高, 发展针对多维甚至高维数据的模型设定检验方法受到了众多学者的关注和研究. 相关研究有文献 [24, 29, 44, 49, 57]. 需指出的是, 文献 [22] 对模型设定检验方法进行了较为详实的综述, 而文献 [26] 则对基于充分性降维 [34] 的模型设定检验方法进行了介绍. 但它们的综述主要集中在低维情形, 本文将介绍近些年所发展的针对高维数据的检验方法.

英文引用格式: Zhang X Y, Shi Y M, Guo X. Two types of hypothesis testing problems in regression analysis (in Chinese). Sci Sin Math, 2025, 55: 1–14, doi: 10.1360/SSM-2024-0125

© 2024《中国科学》杂志社

另外, 变量显著性检验用以探究某一个或某几个变量是否对响应变量起作用. 这是统计学中的一个非常经典也非常重要的问题, 在众多科学领域都有广泛的应用. 在低维线性回归模型下, F 统计量是进行变量显著性检验的经典统计量. 但参数回归模型可能存在模型误设, 因而基于参数回归模型进行变量显著性检验可能存在误导性的风险. 对此, 众多学者对非参数模型下的变量显著性检验问题进行了深入的研究. 早期的工作包括文献 [1, 11, 20]. 这些工作往往涉及非参数核估计, 因而无法处理维数较高的情形. 文献 [32, 67] 提出了降低维数影响的统计量, 但它们的方法仍然只适用于低维情形. 近年来基于更为灵活的机器学习算法进行模型自由的变量显著性检验受到了很多学者的关注和研究. 这些工作包括文献 [5, 10, 38, 53, 56]. 正如文献 [2] 所指出的, 开展模型自由的变量显著性检验对于提高人工智能算法的可解释性是非常重要的. 本文将着重介绍近些年所发展的基于机器学习算法的变量显著性检验方法.

本文余下内容如下安排. 第 2 节介绍经典和近期的模型设定检验方法. 第 3 节介绍模型自由的变量显著性检验方法. 经典的检验方法主要针对低维数据, 而近些年所发展的方法则主要针对高维数据. 第 4 节对本文内容进行总结和讨论, 同时对一些可能的问题进行讨论和展望.

2 模型设定检验

设 $Y \in \mathbb{R}$ 是一维响应变量, 而 $X \in \mathbb{R}^p$ 是 p 维自变量. 所谓模型设定检验, 即考虑如下假设检验问题: 对于 $\Theta \subset \mathbb{R}^d$, 存在 $\theta_0 \in \Theta$ 使得

$$H_0 : \Pr\{m(X) = g(X, \theta_0)\} = 1, \tag{2.1}$$

这里 $\mathrm{E}(Y \mid X) = m(X)$ 是未知光滑函数, $g(\cdot, \cdot)$ 是一个已知的参数函数. 而备择假设则是对任意的 $\theta \in \Theta$, 都有

$$H_1 : \Pr\{m(X) = g(X, \theta)\} < 1. \tag{2.2}$$

上述假设检验问题旨在探究自变量 X 和响应变量 Y 之间的回归模型能否设定为参数回归模型.

本文首先简要介绍对经典的检验方法, 而后重点介绍近期为处理高维数据所发展的一些模型设定检验方法. 设 $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$ 是来自 $\{X, Y\}$ 的一个样本.

2.1 低维数据下的模型设定检验方法

为检验原假设是否成立, 一种自然的想法是分别估计未知光滑函数 $m(X)$ 和参数函数 $g(X, \theta)$, 而后考察它们的估计量之间的差异. 基于这一基本思想, 文献 [27] 提出了如下形式的统计量:

$$T_{HM} = \int \left\{ \sum_{i=1}^{n} \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)} \widehat{\epsilon}_{0i} \right\}^2 \omega(x) dx,$$

这里 $K(\cdot)$ 是核函数, $h \to 0$ 是窗宽, $K_h(\cdot) = K(\cdot/h)/h^p$, $\widehat{\epsilon}_{0i} = Y_i - g(X_i, \widehat{\theta})$ 是原假设下的残差, $\omega(\cdot)$ 是正的权重函数, 而 $\widehat{\theta}$ 是参数向量 θ 的非线性最小二乘估计. 文献 [27] 论证在原假设下有下述结果成立:

$$nh^{p/2} \left\{ T_{HM} - (nh^p)^{-1} \int K^2(x) dx \int \frac{\sigma^2(x)\omega(x)}{f(x)} dx \right\} \Rightarrow N\left\{ 0, 2 \int (K * K)^2 dx \int \frac{\sigma^4(x)\omega^2(x)}{f^2(x)} dx \right\},$$

这里 $f(x)$ 是自变量 X 的密度函数, $\sigma^2(x)=\mathrm{Var}(Y\mid X=x)$ 是条件方差函数, 符号 $*$ 表示卷积运算符, 即

$$K*K(x)=\int K(t)K(x-t)dt.$$

可以发现文献 [27] 所提出的统计量 T_{HM} 对应的总体目标参数为

$$\gamma_{HM}:=\mathrm{E}\{\mathrm{E}^2(\epsilon_0\mid X)\omega(X)\},$$

这里 $\epsilon_0=Y-g(X,\theta_0)$. $\gamma_{HM}=0$ 当且仅当 $\mathrm{E}(\epsilon_0\mid X)=m(X)-g(X,\theta_0)=0$ 进一步当且仅当原假设 H_0 成立. 统计量 T_{HM} 即为目标参数 γ_{HM} 的样本估计形式.

除了直接比较参数模型和非参数模型拟合的差异之外, 还可以考虑比较参数模型和非参数模型的拟合误差. 如果原假设成立, 则参数模型和非参数模型的拟合误差应该是接近的. 基于这一思想, 文献 [12] 提出了如下形式的统计量:

$$T_{DE}=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-g(X_i,\widehat{\theta})\}^2\omega(X_i)-\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\widehat{m}(X_i)\}^2\omega(X_i),$$

其中

$$\widehat{m}(x)=\sum_{i=1}^{n}\frac{K_h(x-X_i)}{\sum_{j=1}^{n}K_h(x-X_j)}Y_i$$

是未知光滑函数 $m(x)$ 的非参数估计量. 定义 $K^{2*}=2K-K*K$. 上述统计量 T_{DE} 在原假设下的渐近分布为

$$nh^{p/2}\left\{T_{DE}-(nh^p)^{-1}K^{2*}(0)\int\sigma^2(x)\omega(x)dx\right\}\Rightarrow N\left(0,2\int K^{2*}(x)dx\int\sigma^4(x)\omega^2(x)dx\right).$$

类似于 T_{HM}, 这个检验统计量也存在一个趋向于无穷大的偏差项. 实际上, 有下式成立:

$$nh^{p/2}\times\frac{1}{nh^p}K^{2*}(0)\int\sigma^2(x)\omega(x)dx=\frac{1}{h^{p/2}}K^{2*}(0)\int\sigma^2(x)\omega(x)dx\to\infty,$$

即偏差项以 $h^{-p/2}$ 的速度发散到无穷大. 注意到统计量 T_{DE} 对应的总体目标参数为

$$\gamma_{DE}:=\mathrm{E}[\{Y-g(X,\theta_0)\}^2\omega(X)]-\mathrm{E}[\{Y-m(X)\}^2\omega(X)].$$

当原假设成立时, 参数 $\gamma_{DE}=0$; 而当备择假设成立时, $\gamma_{DE}>0$. 因而 $\gamma_{DE}=0$ 能够刻画原假设. 而统计量 T_{DE} 则是 γ_{DE} 的样本形式.

文献 [61] 提出了一个平方条件矩检验统计量, 这个方法也独立地由 Fan 和 Li [20] 提出. 注意到

$$\gamma_{ZH}:=\mathrm{E}\{\epsilon_0\mathrm{E}(\epsilon_0\mid X)f(X)\omega(X)\}=\mathrm{E}\{\mathrm{E}^2(\epsilon_0\mid X)f(X)\omega(X)\}.$$

因此 $\gamma_{ZH}=0$ 当且仅当 $\mathrm{E}(\epsilon_0\mid X)=0$, 即 γ_{ZH} 在原假设下为零, 而在备择假设下大于零. 基于这一重要的观察, 文献 [61] 提出了如下的检验统计量:

$$T_{ZH}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}K_h(X_i-X_j)\widehat{\epsilon}_{0i}\widehat{\epsilon}_{0j}\omega(X_i).$$

在原假设下, T_{ZH} 的渐近分布如下:

$$nh^{p/2}T_{ZH}\Rightarrow N\left(0,2\int K^2(x)dx\int\sigma^4(x)\omega^2(x)f^2(x)dx\right).$$

根据条件期望的正交性, 即对 X 的任一可积函数 $Q(X)$, 有 $\mathrm{E}[\{Y-m(X)\}Q(X)]=0$, 可得

$$\gamma_{DE}=\mathrm{E}[\{m(X)-g(X,\theta_0)\}^2\omega(X)]=\mathrm{E}\{\mathrm{E}^2(\epsilon_0\mid X)\omega(X)\}=\gamma_{HM}.$$

对比总体目标参数 γ_{HM},γ_{DE} 和 γ_{ZH}, 可见上述介绍的 3 个统计量 T_{HM},T_{DE} 和 T_{ZH} 对应的目标参数是非常相近的, 但它们具体的估计方法存在差异, 这使得它们的渐近分布有些差别. 其中统计量 T_{ZH} 将求和中的对角线元素去掉, 具有 U 统计量形式. 这使得其具有渐近无偏性, 因此不需要进行偏差校正. 关于这 3 种方法的比较, 可参见文献 [59].

另一大类检验统计量则基于如下的事实: 原假设 H_0 等价于对于任意 $x\in\mathbb{R}^p$, 都有

$$\mathrm{E}[\{Y-g(X,\theta_0)\}I(X\leqslant x)]=0,$$

这里 $I(\cdot)$ 表示示性函数. 上述等式表示原假设下的随机误差 $Y-g(X,\theta_0)$ 和 X 的示性函数是不相关的. 当将所有可能的 x 取遍, 上式便可刻画原假设. 基于上式, 可以考虑如下的经验过程:

$$I_n(x)=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-g(X_i,\widehat{\theta})\}I(X_i\leqslant x).$$

基于上述经验过程 $I_n(x)$, 可进一步考虑其泛函, 从而构造 Cramér-von Mises 或者 Kolmogorov-Smirnov 型统计量. 早期的相关研究可参见文献 [3, 42]. 对于原假设 H_0, 直接基于上述经验过程 $I_n(x)$ 所构造的统计量的渐近分布依赖于参数函数 $g(X,\theta)$, 也依赖于估计量 $\widehat{\theta}$, 因而不是分布自由的. 对此, 文献 [43] 将分布拟合优度检验中的 Khmaladze 变换拓展到回归模型的设定检验问题当中, 对于一维回归模型发展了鞅变换方法, 得到了分布自由的统计量. 同时自助法也通常被用于对基于经验过程的统计量确定临界值. 这使得这类方法通常计算量较大.

2.2 高维数据下的模型设定检验方法

随着数据收集能力的增强, 自变量的维数变得很高, 这使得针对低维数据所发展的模型设定检验方法变得不适用于高维数据. 实际上, 正如文献 [24] 所指出的, 上一小节中所介绍的 T_{HM},T_{DE} 和 T_{ZH} 等检验统计量通常在原假设下的收敛速度为 $O(n^{-1/2}h^{-p/4})$. 当 p 较大时, 这一速度变得很慢, 使得在有限样本下统计量无法控制经验水平. 同时这些检验统计量往往只能检测以 $O(n^{-1/2}h^{-p/4})$ 的速度偏移原假设的局部备择假设, 这使得这些统计量的检测能力随着维数的增加而降低. 另外传统基于经验过程的统计量, 虽然理论收敛速度, 可以达到 $O(n^{-1/2})$, 但高维随机过程的使用导致数据变得稀疏, 进而使得相关统计量计算量大且功效随着维数的增加下降很快.

模型设定检验中的维数灾难问题激发了很多学者的兴趣. 文献 [24] 提出了一种非常新颖的模型—自适应检验方法. 该方法借助充分性降维理论 [34], 能够充分挖掘真实模型的潜在结构信息并将其用来提高统计量的收敛速度, 使得所构造统计量能够更好地控制第一类错误和具有更好的功效表现. 文献 [24] 证明了所提统计量的收敛速度相比传统统计量有了显著提升, 且在渐近意义下, 自变量的维数对所提统计量的收敛速度没有影响. 下面对此进一步介绍.

考虑如下的原假设:

$$H_0:\exists\ \beta_0\in\mathbb{R}^d\ \text{和}\ \theta_0\in\mathbb{R}^p,\quad \Pr\{m(X)=g(\beta_0^{\top}X,\theta_0)\}=1,$$

这里 $g(\cdot,\cdot)$ 是一个已知的参数函数, β_0 和 θ_0 分别是 d 和 p 维未知参数向量. 注意到, 对任意 $p\times p$ 正交矩阵, $m(X)$ 也可重表示为 $m(X)=m(BB^{\top}X)=:\widetilde{m}(B^{\top}X)$. 基于此, 考虑如下备择假设模型:

$$Y=m(B^{\top}X)+\eta,$$

这里 B 是一个 $p\times q$ 正交矩阵, $1\leqslant q\leqslant p$ 是未知的, $m(\cdot)$ 是未知光滑函数, 同时满足 $\mathrm{E}(\eta\mid X)=0$. 上述备择模型假定自变量 X 和响应变量 Y 的关系由自变量 X 的 q 个线性组合 $B^{\top}X$ 捕捉. 当 $q=1$ 时, 上述备择模型退化为单指标模型; 而当 $q=p$ 时, 则变成完全的非参数模型. 另外注意到, 原假设模型也是存在一定结构的, 即在原假设下, 自变量 X 同样通过它的一个线性组合 $\beta_0^{\top}X$ 对响应变量 Y 产生影响. 但需注意, 即使备择假设下的 $q=1$, 原假设模型和备择假设模型也是存在区别的. 实际上, 原假设模型还要求 $\beta_0^{\top}X$ 对 Y 的作用是通过一个已知的参数函数 $g(\cdot,\cdot)$ 来发挥的, 而备择假设模型则允许作用函数是未知光滑函数.

注意到在原假设下 $q=1$, $B=\beta_0/\|\beta_0\|$. 令 $\epsilon_0=Y-g(\beta_0^{\top}X,\theta_0)$, 则在原假设下有

$$\mathrm{E}\{\epsilon_0\mathrm{E}(\epsilon_0\mid B^{\top}X)f(B^{\top}X)\}=0,$$

这里 $f(B^{\top}X)$ 是 $B^{\top}X$ 的密度函数. 而在备择假设下,

$$\mathrm{E}(\epsilon_0\mid B^{\top}X)=\mathrm{E}(Y\mid B^{\top}X)-g(\beta_0^{\top}X,\theta_0)\neq 0,$$

从而有

$$\mathrm{E}\{\epsilon_0\mathrm{E}(\epsilon_0\mid B^{\top}X)f(B^{\top}X)\}>0.$$

上述讨论表明, 可基于 $\mathrm{E}\{\epsilon_0\mathrm{E}(\epsilon_0\mid B^{\top}X)f(B^{\top}X)\}$ 构造统计量. 考虑如下统计量:

$$T_{\mathrm{GWZ}}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\widehat{\epsilon}_{0i}\widehat{\epsilon}_{0j}\widetilde{K}_h\{\widehat{B}(\widehat{q})^{\top}(X_i-X_j)\},$$

这里 $\widehat{\epsilon}_{0i}=Y_i-g(\widehat{\beta}^{\top}X_i,\widehat{\theta})$, $\widehat{\beta}$ 和 $\widehat{\theta}$ 分别是参数向量 β 和 θ 的最小二乘估计量, $\widehat{B}(\widehat{q})$ 是矩阵 B 的估计, $\widehat{q}$ 是 q 的估计, $\widetilde{K}_h(\cdot)=\widetilde{K}(\cdot/h)/h^{\widehat{q}}$, 另外 $\widetilde{K}(\cdot)$ 是一个 $\widehat{q}$ 维核函数. 在实际操作中, 文献 [24] 采用了离散化期望估计 (discretization-expectation estimation) [63] 和最小平均方差估计 (minimum average variance estimation) [58] 来估计矩阵 B. 此外采用 Bayes 信息准则 (Bayesian information criterion, BIC) 确定维数 q.

容易看出, 上述统计量 T_{GWZ} 是上节所介绍的 T_{ZH} 统计量的改进. 注意到

$$T_{ZH}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}K_h(X_i-X_j)\widehat{\epsilon}_{0i}\widehat{\epsilon}_{0j}\omega(X_i).$$

对比 T_{ZH} 和 T_{GWZ}, 有两点不同. 第一, 在核函数中, T_{GWZ} 采用 $\widehat{B}(\widehat{q})^{\top}X$, 而不是 T_{ZH} 中的 X. 这样可将维数从 p 降低到 $\hat{q}$. 第二, 也是更重要的, T_{GWZ} 具有模型自适应性. 核函数中的 $\widehat{B}(\widehat{q})^{\top}X$ 会随着假设的不同而自适应地发生变化. 实际上, 在原假设下可以证明 $\widehat{q}=1$ 以概率 1 成立, 同时 $\widehat{B}(\widehat{q})\to\beta_0/\|\beta_0\|$. 这使得在原假设下可证明 $nh^{1/2}T_{\mathrm{GWZ}}$ 具有渐近正态性. 而在备择假设下, $\widehat{q}=q\geqslant 1$ 以概率 1 成立, 同时 $\widehat{B}(\widehat{q})\to BC$, 这里 C 是一个 $q\times q$ 的正交矩阵. 另外文献 [24] 证明了 T_{GWZ} 能够检测以 $O(n^{-1/2}h^{-1/4})$ 的速度偏移原假设的备择假设.

上述模型自适应的概念和理论方法可被推广到其他检验方法和问题. 例如, 文献 [40] 针对文献 [19] 所提出的广义似然比统计量存在的渐近偏差和维数灾难两大难题, 引入了模型自适应的概念来达到降低维数影响的目的, 同时修正原始统计量中备择假设下残差平方和的形式来进行渐近纠偏. 文献 [50] 将模型自适应的概念引入到基于经验过程的统计量当中, 改进了文献 [45] 所提出的统计量. 进一步地, 文献 [66] 考虑了部分参数单指标模型的模型设定检验问题, 同样考虑了模型自适应的概念, 并构造了相应的经验过程统计量. 相关更详细的介绍, 可参见文献 [26].

需要注意的是, 尽管上述工作较大地回避了维数问题, 但仍然要求自变量的维数 p 是固定的. 文献 [49] 研究了维数发散情形下的模型设定检验问题. 注意到在原假设下

$$\mathrm{E}\{\epsilon_0 I(B^\top X \leqslant t)\} = 0.$$

而在备择假设下, 根据文献 [17, 引理 1], 存在 $\alpha \in \mathcal{S}_q^+$ 使得 $\mathrm{E}(\epsilon_0 \mid \alpha^\top B^\top X) \neq 0$, 这里

$$\mathcal{S}_q^+ = \{\alpha = (\alpha_1, \ldots, \alpha_q)^\top \in \mathbb{R}^q : \|\alpha\| = 1, \alpha_1 \geqslant 0\}.$$

于是在备择假设下有

$$\mathrm{E}\{\epsilon_0 I(\alpha^\top B^\top X \leqslant t)\} \neq 0.$$

在原假设下 $q = 1$, $B = \beta_0/\|\beta_0\|$, $\mathcal{S}_q^+ = \{1\}$. 基于上述观察, 考虑如下经验过程:

$$V_n(\alpha, t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - g(\widehat{\beta}^\top X, \widehat{\theta})\} I(\alpha^\top \widehat{B}^\top X \leqslant t),$$
$$V_n(t) = \sup_{\alpha \in \mathcal{S}_{\hat{q}}^+} |V_n(\alpha, t)|, \quad \mathcal{S}_{\hat{q}}^+ = \{\alpha \in \mathbb{R}^{\widehat{q}} : \|\alpha\| = 1, \alpha_1 \geqslant 0\}.$$

文献 [49] 论证了对于线性回归模型, 当 $(p/\log n)^3/n \to 0$ 时成立; 而对于更一般的参数回归模型, 则要求 $p^5/n \to 0$. 在原假设下, $V_n(u)$ 收敛到零均值的 Gauss 过程. 另外文献 [49] 进一步考虑了高维情形下的鞅变换.

文献 [49] 的工作主要针对维数发散的情形, 而文献 [29] 则考察了高维广义线性回归模型的模型设定检验问题. 当原假设成立时, 模型满足

$$\mathrm{E}(Y \mid X) = \mu(X^\top \beta_0), \quad \mathrm{Var}(Y \mid X) = V\{\mu(X^\top \beta_0)\},$$

这里 $\beta_0 \in \mathbb{R}^p$ 是 p 维未知参数向量, $\mu(\cdot)$ 和 $V(\cdot)$ 是已知的光滑函数. 令 $\widehat{\beta}$ 是参数 β_0 的惩罚估计量. 考虑原假设下尺度化的残差:

$$R_i = \frac{Y_i - \mu(X_i^\top \widehat{\beta})}{\sqrt{V\{\mu(X_i^\top \widehat{\beta})\}}}.$$

令 $R = (R_1, R_2, \ldots, R_n)^\top$ 表示残差向量, 而 ω 表示某个投影方向. 直观上, 当原假设成立时, 残差 R_i 应在零附近随机波动, 因而 $\omega^\top R$ 可近似看作是一些零均值随机变量的和, 应收敛到正态分布. 而当备择假设成立时, 残差向量 R 应包含一些信号. 如果投影方向选取适当, 则在备择假设下 $\omega^\top R$ 的绝对值将较大, 从而能识别出设定错误.

但在具体实施时, 存在两个困难, 一个是如何选择适当的投影方向 ω, 另一个则是如何处理惩罚估计 $\widehat{\beta}$ 的偏差. 为避免理论处理的困难, 在实际操作时将样本 $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$ 随机分割为两部分 $\mathcal{D}_1$ 和 $\mathcal{D}_2$. 设 $|\mathcal{D}_1| = n_1, \mathcal{D}_2 = n_2$ 且 $n_1 + n_2 = n$. 利用子数据集 $\mathcal{D}_1$ 去寻找投影方向. 投影方向的选择应尽可能挖掘残差中可能含有的信号. 基于以上思考, 文献 [29] 提出了如下的操作流程.

首先利用 LASSO (least absolute shrinkage and selection operator)[52] 等惩罚估计分别在 $\mathcal{D}_1$ 和 $\mathcal{D}_2$ 上对参数向量 β_0 进行估计, 分别得到 $\widehat{\beta}_1$ 和 $\widehat{\beta}_2$. 而后在子数据集 $\mathcal{D}_1$ 上将残差 $\{Y_i - \mu(X_i^\top \widehat{\beta}_1)\}_{i=1}^{n_1}$ 对自变量 $\{X_i\}_{i=1}^{n_1}$ 通过灵活的机器学习算法进行训练, 得到残差预测函数 $\widehat{s}$. 这些机器学习算法可以是随机森林[4] 和 XgBoost[9], 也可以是深度神经网络[41]. 这是为了挖掘残差 $\{Y_i - \mu(X_i^\top \widehat{\beta}_1)\}_{i=1}^{n_1}$ 可能存在的非线性信号. 进一步地, 定义

$$\widehat{D}_1 = \mathrm{diag}\{V^{1/2}(\mu(X_i^\top \widehat{\beta}_1))_{i=1}^{n_1}\}, \quad \widehat{\Omega}_1 = \mathrm{diag}\{(\mu'(X_i^\top \widehat{\beta}_1))/\widehat{D}_{1,ii}\}_{i=1}^{n_1},$$

$$\boldsymbol{X}_2 = (X_1, \ldots, X_{n_2})^\top, \quad \boldsymbol{Y}_2 = (Y_1, \ldots, Y_{n_2})^\top.$$

之后则定义如下的估计量 $\widehat{\beta}_{sq}$:

$$\widehat{\beta}_{sq} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{\sqrt{n}} \|\widehat{\Omega}_1\{\widehat{s}(\boldsymbol{X}_2) - \boldsymbol{X}_2\beta\}\|_2 + \lambda_{sq}\|\beta\|_1.$$

定义如下的投影方向:

$$\widehat{\omega}_1 = \frac{\widehat{\Omega}_1\{\widehat{s}(\boldsymbol{X}_2) - \boldsymbol{X}_2\widehat{\beta}_{sq}\}}{\|\widehat{\Omega}_1\{\widehat{s}(\boldsymbol{X}_2) - \boldsymbol{X}_2\widehat{\beta}_{sq}\}\|_2}.$$

最后构造如下统计量:

$$T_{\mathrm{GRP}} = \widehat{\omega}_1^\top \widehat{D}_1^{-1}\{\boldsymbol{Y}_2 - \mu(\boldsymbol{X}_2\widehat{\beta}_2)\}.$$

之所以采用平方根 LASSO [47] 和如此构造投影方向是为了使得构造的投影方向满足如下近似正交性质:

$$\|\boldsymbol{X}_2^\top \widehat{\Omega}_1 \widehat{\omega}_1\|_\infty \leqslant C\sqrt{\log p}.$$

这样可以消除惩罚估计 $\widehat{\beta}_2$ 的偏差所带来的影响. 理论上, 文献 [29] 证明了在原假设下所构造的检验统计量 T_{GRP} 收敛到标准正态分布.

此外文献 [16] 研究了高维分位数回归模型的设定检验问题. 该文献非常巧妙地将模型设定检验问题转化为一种两样本检验问题, 从而可基于针对高维数据的两样本检验方法来进行模型设定检验.

3 变量显著性检验

设 $Z \in \mathbb{R}^{p_1}$, $W \in \mathbb{R}^{p_2}$, 且 $X = (Z^\top, W^\top)^\top \in \mathbb{R}^p$. 所谓变量显著性检验, 即考虑如下假设检验问题:

$$H_0 : \mathrm{E}(Y \mid X) = \mathrm{E}(Y \mid Z).$$

当原假设成立时, 在给定自变量 Z 的情形下, W 对响应变量不起作用, 因而可在后续分析中去除. 变量显著性检验问题也可以看作是一种特殊的模型设定检验问题. 实际上, 在第 2 节中所讨论的模型设定检验问题是检验条件期望函数 $\mathrm{E}(Y \mid X)$ 是否具有参数形式, 而变量显著性检验则是检验条件期望函数 $\mathrm{E}(Y \mid X)$ 是否只依赖于一部分自变量 Z. 可以看出两者都是对条件期望函数 $\mathrm{E}(Y \mid X)$ 的模型设定, 因而第 2 节中所介绍的方法可以用于处理变量显著性检验问题.

对于变量显著性检验问题, 一种常见的处理方式是在参数回归模型设定下开展的. 在参数回归模型下, 变量显著性检验问题可转化为检验感兴趣变量相应的回归系数是否为 0 的问题. 但正如前面内容所指出的, 参数回归模型可能存在模型误设. 为避免模型误设导致后续产生有误导的推断结果, 本节主要集中在非参数回归模型下的变量显著性检验问题.

3.1 低维数据下的变量显著性检验方法

令 $u = Y - \mathrm{E}(Y \mid Z)$ 为原假设成立时的随机误差. 注意到原假设 H_0 等价于

$$\mathrm{E}[uf_Z(Z)\mathrm{E}\{uf_Z(Z) \mid X\}f(X)] = 0,$$

这里 $f_Z(z)$ 是 Z 的密度函数. 基于这一事实, Fan 和 Li[20] 将第 2 节中所介绍的平方条件矩统计量推广到检验变量显著性问题. 具体地, 他们考虑如下检验统计量:

$$T_{FL}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h^{p_2}}K_w\left(\frac{W_i-W_j}{h}\right)\frac{1}{h^{p_1}}K_z\left(\frac{Z_i-Z_j}{h}\right)\widehat{u}_i\widehat{u}_j\widehat{f}_Z(Z_i)\widehat{f}_Z(Z_j),$$

这里 $\widehat{u}_i=Y_i-\widehat{\mathrm{E}}(Y\mid Z_i)$ 是原假设成立时的残差, $\widehat{f}_Z(z)$ 是 $f_Z(z)$ 的核密度估计量, 而 K_w 和 K_z 分别是 p_2 和 p_1 维的核函数. Fan 和 Li[20] 证明了在一定条件下, 当原假设成立时 $nh^{p/2}T_{FL}\Rightarrow N(0,\sigma_1^2)$.

注意到原假设 H_0 等价于对任意 $x\in\mathbb{R}^p$ 都有 $\mathrm{E}\{uI(X\leqslant x)\}=0$. 文献 [11] 基于此结果考虑了如下形式的经验过程:

$$I_n^s(x)=\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_iI(X_i\leqslant x).$$

容易看出, 与第 2 节中所介绍的经验过程 $I_n(x)$ 的区别在于原假设下的残差形式有所不同. 在第 2 节中, 原假设下的残差是参数回归模型下的残差, 而这里的残差则是非参数回归模型下的残差. 同样, 文献 [11] 考虑了上述经验过程 $I_n^s(x)$ 的泛函, 并建立了相应的渐近理论.

对于变量显著性检验问题, 文献 [32] 考虑了一个新的刻画. 具体地, 令 (W_1,Z_1,u_1) 和 (W_2,Z_2,u_2) 是 (W,Z,u) 的两个独立观测, 同时 $K_z(\cdot)$ 和 $\varphi(\cdot)$ 是两个具有正 Fourier 可积变换的偶函数. 则原假设 H_0 等价于

$$I(h)=\mathrm{E}[u_1u_2f_Z(Z_1)f_Z(Z_2)h^{-p_1}K_z\{(Z_1-Z_2)/h\}\varphi(W_1-W_2)]=0.$$

基于上述结果, 文献 [32] 构造了如下形式的统计量:

$$T_{\mathrm{LMP}}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\varphi(W_i-W_j)\frac{1}{h^{p_1}}K_z\left(\frac{Z_i-Z_j}{h}\right)\widehat{u}_i\widehat{u}_j\widehat{f}_Z(Z_i)\widehat{f}_Z(Z_j).$$

对比文献 [20] 所提统计量 T_{FL}, 文献 [32] 所提统计量 T_{LMP} 将 T_{FL} 中的 $h^{-p_2}K_w\{(W_i-W_j)/h\}$ 替换为 $\varphi(W_i-W_j)$. 这样就避免了对 W 进行非参数光滑, 提高了统计量的收敛速度. 实际上, 文献 [32] 证明了在原假设下, $nh^{p_1/2}T_{\mathrm{LMP}}\Rightarrow N(0,\sigma_2^2)$.

文献 [67] 进一步将模型自适应检验方法[24] 用以检验变量显著性, 提高了检验统计量的收敛速度. 但上述工作都是在低维下开展的, 仍然面临维数灾难问题. 近些年, 随着机器学习算法在各个领域取得巨大的成功, 借助机器学习算法考察变量的显著性受到了很多学者的关注和研究. 以下对相关研究进行介绍.

3.2 高维数据下的变量显著性检验方法

注意到变量显著性检验实际上是在比较一个只依赖于 Z 的简化模型 $\mathrm{E}(Y\mid Z)$ 和一个依赖所有变量 X 的完全模型 $\mathrm{E}(Y\mid X)$. 因而可通过比较这两个模型的预测误差来进行变量显著性检验. 这一基本思路是机器学习和统计学领域进行模型评价和模型选择时的经典做法. 将样本 $\mathcal{D}=\{X_i,Y_i\}_{i=1}^n$ 随机分割为两部分 $\mathcal{D}_1$ 和 $\mathcal{D}_2$. 设 $|\mathcal{D}_1|=n_1$, $|\mathcal{D}_2|=n_2$ 且 $n_1+n_2=n$. 利用数据集 $\mathcal{D}_1$ 训练模型, 得到 $\mathrm{E}(Y\mid Z)$ 和 $\mathrm{E}(Y\mid X)$ 的估计, 分别记为 $\widehat{\mathrm{E}}_1(Y\mid Z)$ 和 $\widehat{\mathrm{E}}_1(Y\mid X)$, 而后在第二个数据集 $\mathcal{D}_2$ 上对模型效果进行评价, 即考察两者预测误差的差异

$$T_{nP}=\frac{1}{n_2}\sum_{i\in\mathcal{D}_2}[\{Y_i-\widehat{\mathrm{E}}_1(Y\mid Z_i)\}^2-\{Y_i-\widehat{\mathrm{E}}_1(Y\mid X_i)\}^2].$$

容易看出上述统计量 T_{nP} 是第 2 节所介绍统计量 T_{DE} 在变量显著性检验问题中的体现. 同时也可看出上述统计量 T_{nP} 是下述参数的估计量:

$$\tau = \mathrm{E}[\{Y - \mathrm{E}(Y \mid Z)\}^2] - \mathrm{E}[\{Y - \mathrm{E}(Y \mid X)\}^2].$$

进一步注意到

$$\tau = \mathrm{E}[\{\mathrm{E}(Y \mid X) - \mathrm{E}(Y \mid Z)\}^2].$$

因而参数 τ 是非负的, 即在原假设下 $\tau = 0$, 而在备择假设下 $\tau > 0$. 与第 3.1 小节所介绍的传统检验统计量所不同的是, 这里使用灵活的机器学习算法估计未知的回归函数. 上述形式的统计量已被多位学者采用和研究 (参见文献 [6,33,55,60]).

然而尽管上述统计量 T_{nP} 形式非常自然, 但它存在一个严重的理论问题. 实际上该统计量在原假设下会发生退化问题, 即 $\sqrt{n_2}T_{nP} \Rightarrow 0$. 这使得该统计量无法计算 p 值. 为克服该退化问题, 很多学者对 T_{nP} 进行了改进. 文献 [56] 考虑将数据集 $\mathcal{D}_2$ 进一步分割为 $\mathcal{D}_{21}$ 和 $\mathcal{D}_{22}$, $|\mathcal{D}_{21}| = n_{21}$ 和 $|\mathcal{D}_{22}| = n_{22}$ 且 $n_{21} + n_{22} = n_2$. 而后分别在 $\mathcal{D}_{21}$ 和 $\mathcal{D}_{22}$ 上计算简化模型 $\mathrm{E}(Y \mid Z)$ 和完全模型 $\mathrm{E}(Y \mid X)$ 的预测误差, 即考虑如下形式的统计量:

$$\widetilde{T}_{nP} = \frac{1}{n_{21}} \sum_{i \in \mathcal{D}_{21}} \{Y_i - \widehat{\mathrm{E}}_1(Y \mid Z_i)\}^2 - \frac{1}{n_{22}} \sum_{i \in \mathcal{D}_{22}} \{Y_i - \widehat{\mathrm{E}}_1(Y \mid X_i)\}^2.$$

使用不同的数据集估计 $\mathrm{E}[\{Y - \mathrm{E}(Y \mid Z)\}^2]$ 和 $\mathrm{E}[\{Y - \mathrm{E}(Y \mid X)\}^2]$, 从而使得 $\widetilde{T}_{nP}$ 能够避免原假设下的退化问题. 但对测试集 $\mathcal{D}_2$ 的进一步分割使得实际用于检验的样本量减少, 从而使得统计量的检测功效降低. 文献 [8] 通过允许 $\mathcal{D}_{21}$ 和 $\mathcal{D}_{22}$ 存在一定程度的重叠, 在避免退化问题的同时, 提高了文献 [56] 所提统计量的功效.

为克服统计量退化问题, 文献 [10] 则考虑添加随机噪声, 构造了如下统计量:

$$T_{\mathrm{DSP}} = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} [\{Y_i - \widehat{\mathrm{E}}_1(Y \mid Z_i)\}^2 - \{Y_i - \widehat{\mathrm{E}}_1(Y \mid X_i)\}^2 + \rho_n e_i],$$

这里 $e_i \sim N(0,1)$ 且独立于数据集 $\mathcal{D}$, 而 ρ_n 是扰动大小. 在原假设下随机噪声能对渐近分布起到主导作用, 从而回避退化问题. 但同样地, 随机噪声的引入会增大统计量的方差, 从而降低检测功效. 对于扰动大小 ρ_n 和样本量的相对比例 n_1/n_2 的选择, 文献 [10] 提出了一种适当的数据驱动算法.

文献 [5] 注意到原假设等价于 $\mathrm{E}(\widetilde{Y} \mid X) = \mathrm{E}(\widetilde{Y})$, 这里 $\widetilde{Y} = Y - \mathrm{E}(Y \mid Z)$. 基于此事实, 则可通过比较条件期望 $\mathrm{E}(\widetilde{Y} \mid X)$ 和无条件期望 $\mathrm{E}(\widetilde{Y})$ 的差异构造统计量. 实际上, 令 $\widehat{\mathrm{E}}_1(\widetilde{Y} \mid X)$ 表示 $\mathrm{E}(\widetilde{Y} \mid X)$ 基于 $\mathcal{D}_1$ 所得的估计量, 则可考虑如下形式的统计量:

$$T_{\mathrm{CGZ}} = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \left[\widehat{\mathrm{E}}_1(\widetilde{Y} \mid X) - \frac{1}{n_2} \sum_{j \in \mathcal{D}_2} \{Y_j - \widehat{\mathrm{E}}_1(Y \mid Z_j)\} \right]^2.$$

尽管 $\mathrm{E}(\widetilde{Y})$ 实际上恒等于零, 但在上述统计量的构造中, 仍基于数据对该无条件期望进行估计. 这里的主要目的是通过 $n_2{}^{-1} \sum_{j \in \mathcal{D}_2} \{Y_j - \widehat{\mathrm{E}}_1(Y \mid Z_j)\}$ 来引入随机噪声, 从而克服 $\sum_{i \in \mathcal{D}_2} \widehat{\mathrm{E}}_1^2(\widetilde{Y} \mid X)$ 所带来的退化问题. Cai 等 [5] 证明了在原假设下 $S_n =: n_2 \widehat{\sigma}_{Y|Z}^{-2} T_{\mathrm{CGZ}} \Rightarrow \chi_1^2$, 这里

$$\widehat{\sigma}_{Y|Z}^2 = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \{Y_i - \widehat{\mathrm{E}}_1(Y \mid Z_j)\}^2$$

是 $\sigma^2_{Y\,|\,Z} = \mathrm{E}[\{Y - \mathrm{E}(Y \mid Z)\}^2]$ 的估计. 进一步地, 他们考虑如下形式的统计量:

$$S_n^* = S_n + a \sum_{i \in \mathcal{D}_2} \widehat{\mathrm{E}}_1^2(\widetilde{Y} \mid X).$$

注意到在原假设下, $\sum_{i\in\mathcal{D}_2} \widehat{\mathrm{E}}_1^2(\widetilde{Y} \mid X) \to 0$, 即发生退化问题. 这使得在原假设下 S_n^* 和 S_n 有相同的渐近分布, 都收敛到自由度为 1 的卡方分布. 而另外由于 $\sum_{i\in\mathcal{D}_2} \widehat{\mathrm{E}}_1^2(\widetilde{Y} \mid X)$ 恒为非负, 因而 S_n^* 比 S_n 的功效更大. 这表明在原假设下的退化问题可用来帮助提高备择假设下的检验功效.

区别于上述基于预测误差比较的统计量, 文献 [38] 注意到原假设等价于对任意平方可积函数 $\phi(X)$, 有

$$\mathrm{E}[\{Y - \mathrm{E}(Y \mid Z)\}\phi(X)] = 0.$$

而从功效角度考虑, 当选择 $\phi(X) = \{\mathrm{E}(Y \mid X) - \mathrm{E}(Y \mid Z)\}/\mathrm{Var}(Y \mid X)$ 时能使得所构造统计量具有最优功效. 但这一选择在原假设下会使得 $\phi(X) \equiv 0$. 文献 [38] 对此问题进行了深入的研究. 下面简述他们所提统计量的操作流程. 首先利用 $\mathcal{D}_1$ 对函数 $\phi(X)$ 进行估计, 记为 $\widehat{\phi}_1(X)$. 而后在 $\mathcal{D}_2$ 上, 将 $\widehat{\phi}_1(X_j)$, $j \in \mathcal{D}_2$ 对 Z_j 进行回归得到 $\widehat{m}_{2,\widehat{\phi}}$; 同时利用 $\mathcal{D}_2$ 对函数 $\mathrm{E}(Y \mid Z)$ 进行估计, 记为 $\widehat{\mathrm{E}}_2(Y \mid Z)$. 记 $L_j := \{Y_j - \widehat{\mathrm{E}}_2(Y \mid Z_j)\}\{\widehat{\phi}_1(X_j) - \widehat{m}_{2,\widehat{\phi}}(Z_j)\}$, $j \in \mathcal{D}_2$, 最后定义如下统计量:

$$T_{\mathrm{PCM}} = \frac{\frac{1}{\sqrt{n_2}} \sum_{j\in\mathcal{D}_2} L_i}{\sqrt{\frac{1}{n_2}\sum_{j\in\mathcal{D}_2} L_i^2 - (\frac{1}{n_2}\sum_{j\in\mathcal{D}_2} L_i)^2}}.$$

在一定条件下, 文献 [38] 建立了上述统计量在原假设下的渐近正态性, 同时在备择假设下进行了功效分析.

尽管上述统计量较好地克服了原假设下的退化问题, 但它们往往需要更大的训练样本和更少的检验样本, 这导致它们存在一定的功效损失问题. 同时理论上, 它们往往无法检测以 $O(n^{-1/2})$ 的速度偏离原假设的局部备择假设. 如何有效地处理退化问题, 同时避免功效损失是一个非常重要和具有挑战的科学问题.

此外上述涉及机器学习算法的检验统计量往往需要进行样本分割. 这导致结果带有一定的随机性. 对此, 文献 [35, 39] 所提出的 p 值整合方法经常被采用.

4 总结和讨论

本文对回归分析中的两类假设检验问题, 即参数回归模型的模型设定检验问题和模型自由的变量显著性检验问题, 进行了介绍. 根据自变量的维数, 本文分别讨论低维数据下和高维数据下的检验方法. 下面就一些可能的议题进行简要讨论.

4.1 高维数据下半参数回归模型的设定检验

参数回归模型对于描述自变量和响应变量之间的关系往往过于简化. 更为灵活的半参数回归模型如单指标模型、部分线性模型、可加模型和变系数模型等也经常被采用. 尽管在低维情形下, 针对这些半参数回归模型的设定检验已有很多研究, 如文献 [18, 21, 46, 65], 但在高维情形下, 相关研究还很少. 如何将本文所介绍的针对高维参数回归模型的设定检验方法用以解决高维半参数回归模型的设定检验问题存在一定的理论挑战.

4.2 分位数回归下的变量显著性检验

条件分位数回归能刻画数据的尾部特征, 在条件分位数回归下进行变量显著性检验具有重要的科学意义和应用价值, 但相关研究还较少. 文献 [30] 将文献 [61] 所提平方条件矩统计量用以检验条件分位数回归的变量显著性检验问题, 并将所提统计量用以考察分位数下的 Granger 因果关系. 这一工作在经济学领域产生了很大影响. 文献 [54] 将文献 [11] 提出的基于经验过程的统计量推广应用于分位数回归. 但这些方法仍只适用于低维情形. 因而急需在高维情形下对分位数回归开展模型自由的变量显著性检验研究.

4.3 复杂数据的模型设定检验和变量显著性检验

本文所介绍的检验方法都是针对完全观测到的独立同分布数据. 复杂数据如缺失数据、生存数据和测量误差数据在实际中也是常见的. 已有很多学者研究了针对这些复杂数据的模型设定检验方法 (参见文献 [25,31,36]). 另外, 文献 [37] 研究了混合治愈模型中的治愈率的变量显著性检验问题. 文献 [62] 考察了缺失数据和因果效应中的变量显著性检验问题. 文献 [7] 也研究了条件分位数处理效应中的异质性检验问题. 但这些研究仍然主要针对低维数据, 在高维情形下如何发展适当有效的检验统计量仍然是一个重要但有很大挑战的科学问题.

4.4 其他假设检验问题

除了文中所讨论的两类假设检验问题, 在回归分析中还存在其他一些重要的检验问题, 如异方差性检验和回归函数比较. 所谓异方差性检验, 即检验回归误差的条件方差函数是否为常数 (参见文献 [14,23,48,64]). 此外, 回归函数的比较可用以判断自变量对响应变量的影响是否存在异质性. 这类问题在现实中是常见的, 如年龄、教育水平和工作年限等因素对收入的影响是否存在性别或者种族差异. 对此相关研究可参见文献 [13,15,28,51].

最后, 希望通过本文的介绍和讨论能够引起读者的兴趣, 使得更多学者对模型设定检验和变量显著性检验问题开展深入的研究.

致谢 非常感谢审稿人对手稿的仔细阅读和他们建设性和有见地的意见和建议.

参考文献

1 Aït-Sahalia Y, Bickel P J, Stoker T M. Goodness-of-fit tests for kernel regression with an application to option implied volatilities. J Econometrics, 2001, 105: 363–412
2 Allen G I, Gan L, Zheng L L. Interpretable machine learning for discovery: Statistical challenges and opportunities. Annu Rev Stat Appl, 2024, 11: 97–121
3 Bierens H J. Consistent model specification tests. J Econometrics, 1982, 20: 105–134
4 Breiman L. Random forests. Mach Learn, 2001, 45: 5–32
5 Cai L H, Guo X, Zhong W. Test and measure for partial mean dependence based on machine learning methods. J Amer Stat Assoc, 2024, in press
6 Cai Z R, Lei J, Roeder K. Model-free prediction test with application to genomics data. Proc Natl Acad Sci USA, 2022, 119: e2205518119
7 Cai Z W, Fang Y, Lin M, et al. A nonparametric test of heterogeneity in conditional quantile treatment effects. Econom Theory, 2024, doi: 10.1017/S0266466624000045
8 Chen G, Jia Y X, Wang G H, et al. Zipper: Addressing degeneracy in algorithm-agnostic inference. arXiv:2306.16852, 2023

9 Chen T Q, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016, 785–794

10 Dai B, Shen X T, Pan W. Significance tests of feature relevance for a black-box learner. IEEE Trans Neural Netw Learn Syst, 2024, 35: 1898–1911

11 Delgado M A, Manteiga W G. Significance testing in nonparametric regression based on the bootstrap. Ann Statist, 2001, 29: 1469–1507

12 Dette H. A consistent test for the functional form of a regression based on a difference of variance estimators. Ann Statist, 1999, 27: 1012–1040

13 Dette H, Neumeyer N. Nonparametric analysis of covariance. Ann Statist, 2001, 29: 1361–1400

14 Dette H, Neumeyer N, Keilegom I V. A new test for the parametric form of the variance function in non-parametric regression. J R Stat Soc Ser B Stat Methodol, 2007, 69: 903–917

15 Dette H, Wagener J, Volgushev S. Comparing conditional quantile curves. Scand J Stat, 2011, 38: 63–88

16 Dong C, Li G D, Feng X D. Lack-of-fit tests for quantile regression models. J R Stat Soc Ser B Stat Methodol, 2019, 81: 629–648

17 Escanciano J C. A consistent diagnostic test for regression models using projections. Econom Theory, 2006, 22: 1030–1051

18 Fan J Q, Jiang J C. Nonparametric inferences for additive models. J Amer Statist Assoc, 2005, 100: 890–907

19 Fan J Q, Zhang C, Zhang J. Generalized likelihood ratio statistics and wilks phenomenon. Ann Statist, 2001, 29: 153–193

20 Fan Y Q, Li Q. Consistent model specification tests: Omitted variables and semiparametric functional forms. Econometrica, 1996, 64: 865–890

21 Feng X D, Zhu L P. Estimation and testing of varying coefficients in quantile regression. J Amer Statist Assoc, 2016, 111: 266–274

22 González-Manteiga W, Crujeiras R M. An updated review of goodness-of-fit tests for regression models. TEST, 2013, 22: 361–411

23 Guo X, Jiang X J, Zhang S, et al. Pairwise distance-based heteroscedasticity test for regressions. Sci China Math, 2020, 63: 2553–2572

24 Guo X, Wang T, Zhu L X. Model checking for parametric single-index models: A dimension reduction model-adaptive approach. J R Stat Soc Ser B Stat Methodol, 2016, 78: 1013–1035

25 Guo X, Xu W L, Zhu L X. Model checking for parametric regressions with response missing at random. Ann Inst Statist Math, 2015, 67: 229–259

26 Guo X, Zhu L X. A Review on dimension-reduction based tests for regressions. In: From Statistics to Mathematical Finance: Festschrift in Honour of Winfried Stute. Berlin: Springer, 2017, 105–125

27 Hardle W, Mammen E. Comparing nonparametric versus parametric regression fits. Ann Statist, 1993, 21: 1926–1947

28 Hu X, Lei J. A two-sample conditional distribution test using conformal prediction and weighted rank sum. J Amer Statist Assoc, 2024, 119: 1136–1154

29 Janková J, Shah R D, Bühlmann P, et al. Goodness-of-fit testing in high dimensional generalized linear models. J R Stat Soc Ser B Stat Methodol, 2020, 82: 773–795

30 Jeong K, Härdle W K, Song S. A consistent nonparametric test for causality in quantile. Econom Theory, 2012, 28: 861–887

31 Koul H L, Song W. Minimum distance regression model checking with Berkson measurement errors. Ann Statist, 2009, 37: 132–156

32 Lavergne P, Maistre S, Patilea V. A significance test for covariates in nonparametric regression. Electron J Stat, 2015, 9: 643–678

33 Lei J, G'Sell M, Rinaldo A, et al. Distribution-free predictive inference for regression. J Amer Statist Assoc, 2018, 113: 1094–1111

34 Li B. Sufficient Dimension Reduction: Methods and Applications with R. Boca Raton: Chapman and Hall/CRC, 2018

35 Liu Y, Xie J. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. J Amer Statist Assoc, 2020, 115: 393–402

36 Lopez O, Patilea V. Nonparametric lack-of-fit tests for parametric mean-regression models with censored data. J Multivariate Anal, 2009, 100: 210–230

37 López-Cheda A, Jácome M A, Van Keilegom I, et al. Nonparametric covariate hypothesis tests for the cure rate in mixture cure models. Stat Med, 2020, 39: 2291–2307

38 Lundborg A R, Kim I, Shah R D, et al. The projected covariance measure for assumption-lean variable significance testing. arXiv:2211.02039, 2022

39 Meinshausen N, Meier L, Bühlmann P. p-values for high-dimensional regression. J Amer Statist Assoc, 2009, 104: 1671–1681

40 Niu C Z, Guo X, Zhu L X. Enhancements of non-parametric generalized likelihood ratio test: Bias correction and dimension reduction. Scand J Stat, 2018, 45: 217–254

41 Schmidt-Hieber J. Nonparametric regression using deep neural networks with ReLU activation function. Ann Statist, 2020, 48: 1875–1897

42 Stute W. Nonparametric model checks for regression. Ann Statist, 1997, 25: 613–641

43 Stute W, Thies S, Zhu L X. Model checks for regression: An innovation process approach. Ann Statist, 1998, 26: 1916–1934

44 Stute W, Xu W L, Zhu L X. Model diagnosis for parametric regression in high-dimensional spaces. Biometrika, 2008, 95: 451–467

45 Stute W, Zhu L X. Model checks for generalized linear models. Scand J Stat, 2002, 29: 535–545

46 Stute W, Zhu L X. Nonparametric checks for single-index models. Ann Statist, 2005, 33: 1048–1083

47 Sun T, Zhang C H. Scaled sparse linear regression. Biometrika, 2012, 99: 879–898

48 Tan F L, Jiang X J, Guo X, et al. Testing heteroscedasticity for regression models based on projections. Stat Sin, 2021, 31: 625–646

49 Tan F L, Zhu L X. Adaptive-to-model checking for regressions with diverging number of predictors. Ann Statist, 2019, 47: 1960–1994

50 Tan F L, Zhu X H, Zhu L X. A projection-based adaptive-to-model test for regressions. Stat Sin, 2018, 28: 157–188

51 Tedesco L, Van Keilegom I. Comparison of quantile regression curves with censored data. TEST, 2023, 32: 829–864

52 Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288

53 Verdinelli I, Wasserman L. Decorrelated variable importance. J Mach Learn Res, 2024, 25: 1–27

54 Volgushev S, Birke M, Dette H, et al. Significance testing in quantile regression. Electron J Stat, 2013, 7: 105–145

55 Williamson B D, Gilbert P B, Carone M, et al. Nonparametric variable importance assessment using machine learning techniques. Biometrics, 2021, 77: 9–22

56 Williamson B D, Gilbert P B, Simon N R, et al. A general framework for inference on algorithm-agnostic variable importance. J Amer Statist Assoc, 2023, 118: 1645–1658

57 Xia Y C. Model checking in regression via dimension reduction. Biometrika, 2009, 96: 133–148

58 Xia Y C, Tong H, Li W K, et al. An adaptive estimation of dimension reduction space. J R Stat Soc Ser B Stat Methodol, 2002, 64: 363–410

59 Zhang C M, Dette H. A power comparison between nonparametric regression tests. Statist Probab Lett, 2004, 66: 289–301

60 Zhang L, Janson L. Floodgate: Inference for model-free variable importance. arXiv:2007.01283, 2020

61 Zheng J X. A consistent test of functional form via nonparametric estimation techniques. J Econometrics, 1996, 75: 263–289

62 Zhou N W, Guo X, Zhu L X. Significance test for semiparametric conditional average treatment effects and other structural functions. Comput Stat Data Anal, 2024, 189: 107839

63 Zhu L P, Wang T, Zhu L X, et al. Sufficient dimension reduction through discretization-expectation estimation. Biometrika, 2010, 97: 295–304

64 Zhu L X, Fujikoshi Y, Naito K. Heteroscedasticity checks for regression models. Sci China Math, 2021, 44: 1236–1252

65 Zhu L X, Ng K W. Checking the adequacy of a partial linear model. Stat Sin, 2003, 13: 763–781

66 Zhu X H, Guo X, Zhu L X. An adaptive-to-model test for partially parametric single-index models. Stat Comput, 2017, 27: 1193–1204

67 Zhu X H, Zhu L X. Dimension reduction-based significance testing in nonparametric regression. Electron J Stat, 2018, 12: 1468–1506

Two types of hypothesis testing problems in regression analysis

Xinyu Zhang, Yanmei Shi & Xu Guo

Abstract In this paper, we provide a brief overview of two types of hypothesis testing problems in regression analysis, i.e., model specification testing and variable significance testing. We introduce classic and recently developed testing methods for these two types of problems. We present the construction principles of relevant statistics and evaluate the possible problems which may exist in different methods. Classic methods mainly focus on low-dimensional data and are often greatly influenced by the dimensionality of predictors. The recently developed methods are mainly applicable to high-dimensional data, aiming to effectively overcome the curse of dimensionality in these two types of testing problems. The use of machine learning algorithms is an important feature of recently developed testing methods. We conclude with a brief discussion of some potential issues.

Keywords **model specification testing, variable significance testing, curse of dimensionality, machine learning**

MSC(2020) **62G10, 62G20, 62J12**

doi: 10.1360/SSM-2024-0125