中山大學心理学系
SUN YAT-SEN UNIVERSITY DEPARTMENT OF PSYCHOLOGY

Lasso回归：从解释到预测

张沥今，魏夏琰，陆嘉琦
指导教师：潘俊豪副教授
中山大学心理学系

2019.10.20，第二十二届全国心理学大会，杭州

目录

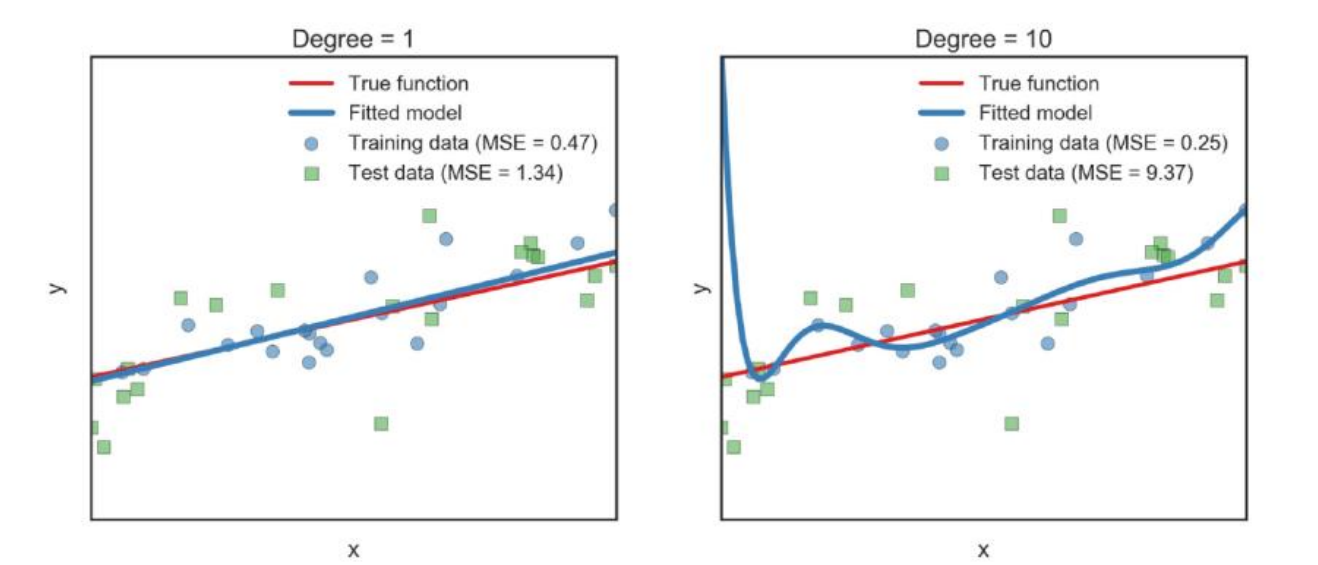
1 OLS回归

OLS回归

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2\ x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- 在采用回归模型分析数据的心理学研究中，最小二乘法(Ordinary Least Square, OLS)是最常用的模型系数估计方法(Helwig, 2017)。
- OLS方法通过最小化结果变量的预测值和观测值之间的误差来估计回归模型中的参数，可以针对当前样本提供最准确的线性无偏估计。

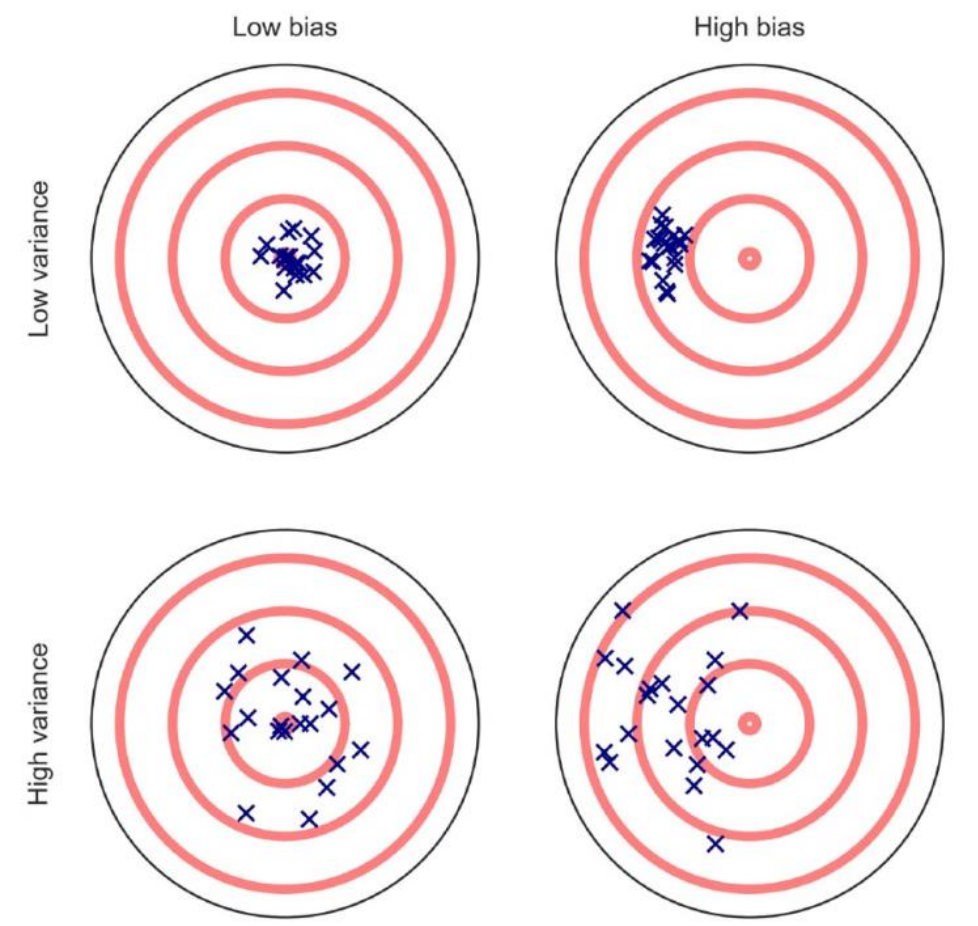
过拟合



(Yarkoni & Westfall, 2017)

- 过度拟合的模型学习到了不适用于总体的规律，进而会带来有偏的参数估计
- 抽样变异性(Sampling Variability)

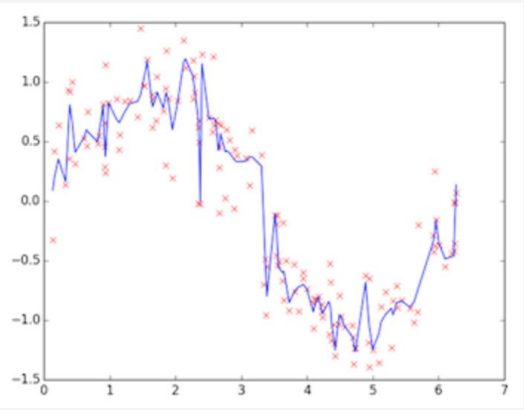
偏差-方差权衡

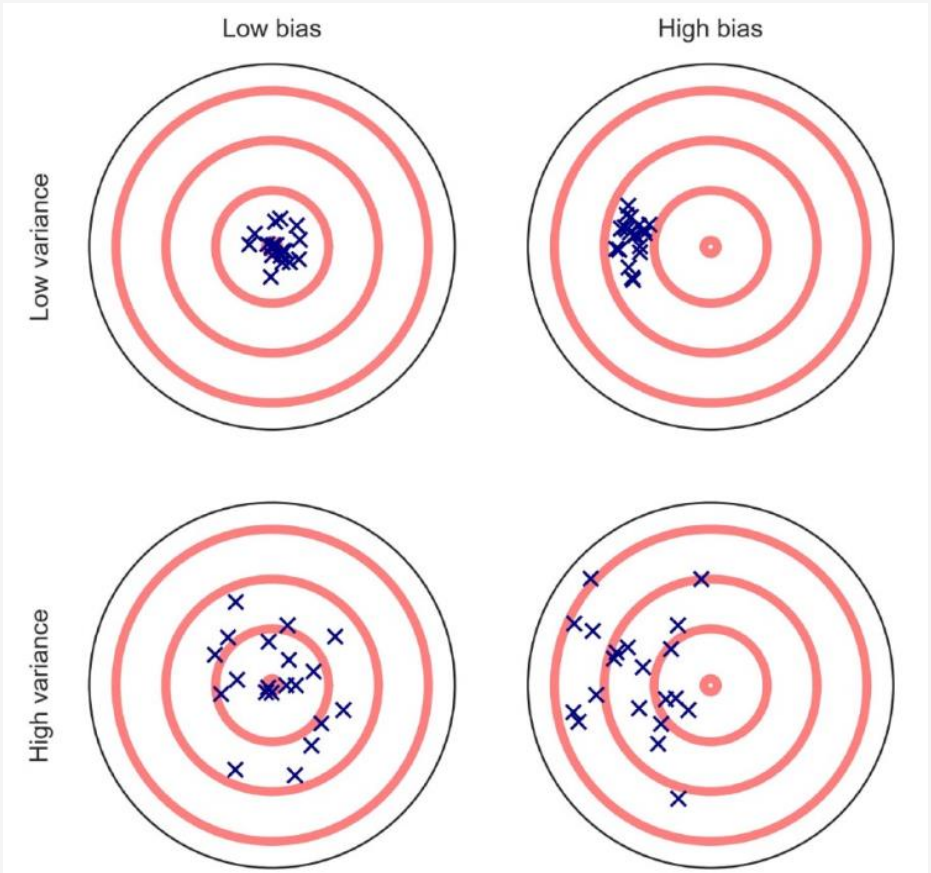


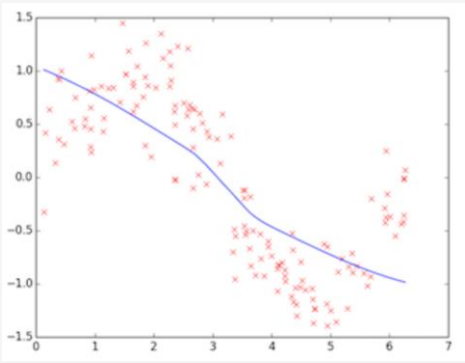
(Yarkoni & Westfall, 2017)

- 偏差：预测值-真实值的偏离程度
- 方差：预测值的离散情况

偏差-方差权衡







(Yarkoni & Westfall, 2017)

多重共线性

- 多重共线性(Multicollinearity)：回归模型中多个预测变量存在相关关系的现象
- 当模型存在较强的多重共线性时，OLS估计得到的回归系数的估计方差将增大
- 回归系数极易受到样本数据的微小波动的影响，估计的稳定性较差

正则化方法

- 机器学习领域中，以 Lasso(Least absolute shrinkage and selection operator; Tibshirani, 1996)方法为代表的正则化方法可以有效地优化OLS估计、处理过拟合问题 (Tibshirani, 1996; Zou & Hastie, 2005; Tibshirani et al., 2005; Zou, 2006; Candes & Tao, 2007)。

- 正则化(regularization)方法通过在模型中增加惩罚项的方式可以将过小的回归系数压缩到0，以一定的估计偏差为代价而获得更高的模型预测准确度和模型泛化能力。

解释和预测

- 机器学习方法应用的阻碍：对可解释性的质疑
- 在着眼于对研究结论的解释之外，模型的泛化能力、预测变量集对于因变量的预测能力同样需要关注。
- 吴喜之(2019)指出，事实上回归模型中单个系数并不具备可解释性。

- 如何利用统计方法、规范研究流程来提供可重复性危机的解决方案已经逐渐成为心理学领域的热点问题(胡传鹏 等，2016; Spellman, 2015)。
- 机器学习领域的工具有希望帮助心理学成为一门更有预见性的科学(Rosenberg, Casey, & Holmes, 2018; Serang et al., 2017; Yarkoni & Westfall, 2017)。

2 Lasso 回归

Lasso 回归

正则化方法：

$$L^{Reg}(\beta) = L^{OLS}(\beta) + \lambda P(\beta)$$

- $L^{Reg}(\beta)$：惩罚后的损失函数
- $P(\beta)$：惩罚函数
- $\lambda(\lambda \geq 0)$：调整参数

Lasso方法：

$$P^{\text{Lasso}}(\beta) = \sum_{j=1}^{p} |\beta_j|$$

Lasso 回归

$$P^{\text{Lasso}}(\beta) = \sum_{j=1}^{p} (\beta_j)^2$$

- 相比其他正则化方法，Lasso方法可以直接将冗余预测变量的回归系数压缩到0从而达到变量选择的作用 (Tibshirani, 1996)。

- Lasso回归可得到模型的稀疏解，从而避免在预测变量过多时采用OLS估计带来的过拟合和多重共线性的问题。

- Lasso方法中的惩罚项对所有的回归系数压缩的力度是一致的，这避免了对重要回归系数的过分压缩 (Hesterberg, Choi, Meier, & Fraley, 2008)。

分析步骤

参数 λ 的选择

- **交叉验证 (glmnet软件包)**
- **信息标准**

***p* 值的计算(covTest软件包)**

- 结果变量(Y)的观察值和模型预测值($X\hat{\beta}$)之间的协方差
- 在仅缺少某一预测变量的模型中，加入该预测变量后，通过计算模型协方差变化值，再进行显著性检验就可以实现变量选择
- 不需重复抽样和分割数据

- 实例资料：www.lijinzhang.xyz/blog_190926_lasso.html

Lasso回归的应用

临床心理学

- Lasso方法已被成功应用于识别潜在的患者：Schmid等人(2013)对29个后来发展为阿尔兹海默症的患者以及相应条件匹配的29个正常人进行了为期八年的追踪
- 研究变量(p=115)，观测值(N=29) → Lasso回归

认知神经科学

- 全基因组关联研究(Genome Wide Association Study, GWAS)：筛选基因位点(Ayers & Cordell, 2010; Shi et al., 2011)、检测基因之间的交互作用(D'Angelo, Rao & Gu, 2009; Li et al., 2011)、进行风险预测(Kooperberg, LeBlanc, & Obenchain, 2010)。

发展心理学，教育心理学…

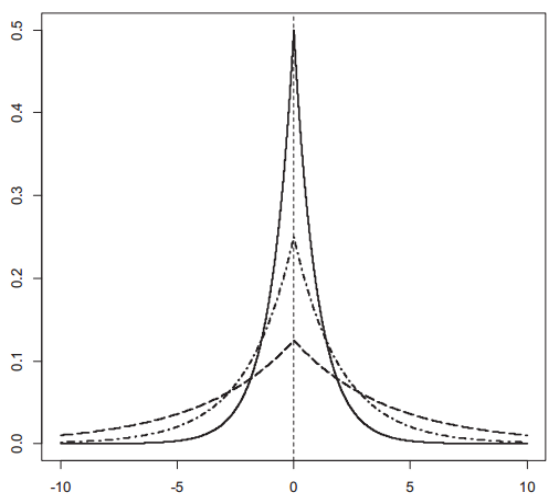
3 Lasso方法的扩展

扩展

- **Bayesian Lasso**

  在贝叶斯方法中，如果选择了合适的先验分布，先验分布的对数形式就会扮演惩罚项的角色。

  Tibshirani (1996)：对参数θ_j提供同样的、相互独立的双指数先验分布 $\frac{\lambda}{2} exp(-\lambda|\theta_j|)$：



扩展

Bayesian Lasso

- 贝叶斯Lasso可以通过Gibbs采样法提供有效的标准误估计(Kyung et al., 2010)。

- 贝叶斯 Lasso 回归模型也能够在估计未知系数的同时估计正则化参数，避免了使用传统交叉验证方法所需的大量计算负担，具有着非常广阔的应用前景(Park & Casella, 2008)。

- ‘blasso’软件包

扩展

- 在回归模型中，Lasso方法还可以被用于筛选中介变量(Serang et al., 2017)
- 在回归模型之外，正则化方法也逐渐被应用于特征网络模型(Feature Network Models; Frank & Heiser, 2011)和结构方程模型(Structural Equation Modeling)中

- 采用贝叶斯Lasso正则化方法解决传统的验证性因子分析限制过于严格的问题(Pan, Ip, & Dubé, 2017; Muthén & Asparouhov, 2012)

  ‘blcfa’软件包
- 在MIMIC模型(Multiple Indicators and Multiple Causes, MIMIC)中利用Lasso正则化方法进行预测变量的筛选(Jacobucci, Brandmaier, & Kievit, in press)

  ‘regsem’软件包

讨论

- 以Lasso为代表的正则化模型在机器学习领域发挥着越来越重要的作用，目前也已经广泛应用于生物医学等领域

- 在临床心理学和认知神经科学之外，Lasso回归在教育心理学、人格心理学等领域中也可以发挥其价值。

- Hadley Wickham在出目前有大概13个关于Lasso方法的R包，但是每一个都不够完善，如，不能处理缺失值、分类变量等等，因此他计划将整合这些软件包以制作一个更高效的分析工具。

Thanks!