《线性回归》
—missing data analysis

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.05.28

主要内容：模型建立

主要内容：

- 缺失数据(missing data)
- 计算协变量之间的相关系数
- 模型选择[已讲]

缺失数据

- 最佳解决方案：找到缺失值！（但通常是不可行的）。
- 始终：要想想数据丢失的原因。例如，由于药物不良副作用退出研究的患者，他们的值将缺失。忽视这些患者可能会导致错误的结论。这个问题没有简单的解决办法。

无信息缺失(non-informative missingness)

- 无信息缺失意味着观测值是随机缺失的。
- 可能的解决方案：
  - ✓ 删除所有有缺少数据的观察值。这有可能导致数据的巨大损失。
  - ✓ 删除有许多缺失值的协变量。
  - ✓ 对缺失的数据点进行插补（impute），例如，把缺失值插补为该自变量的平均值。还有其它更复杂的多种插补方法（参见，例如，R-package mice）

R package: mice 【R中还有多个处理missing data 的包】

- **mice:** Multivariate Imputation by Chained Equations
- 安装R package：mice
- 利用help(package='mice')查看各种插值方法；
- data(package='mice')查看mice自带的数据；
  - ✓ help('boys')查看数据集‘boys’；
  - ✓ help('tbc')查看数据集‘tbc’；

关于missing data的参考书

Van Buuren, S. (2018). Flexible Imputation of Missing Data. Second Edition. Chapman & Hall/CRC. Boca Raton, FL.

作业：

至少掌握‘mice’中的一种方法，并搞清楚其原理, 恰当选择‘mice’中的一个数据集，阅读与之相关的文献之后做回归分析。

例. (芝加哥保险数据)

- 数据来自于*1970*年一项在有*47*个邮政编码的地区中，芝加哥的保险红线（拒绝发放保险）与种族构成、火灾和盗窃率、住房年龄和收入之间关系的研究项目。缺失值被随机的添加上。
- 默认情况下，*R*中的回归分析只使用不包含缺失值的情况*(*上一张*ppt*中的第一个情况*)*。这将样本量减少到*27*。
- 注意，年龄是缺失值最多的*(5)*。如果年龄不是一个关键变量，那么最好不要考虑它。忽略年龄，我们的样本量是*32*。
- 我们可以用协变量的平均值代替缺失的数据。

计算相关系数

- 我们可以把R命令cor()作用到数据矩阵上，一行代码就可以计算所有变量对之间的相关系数。
- cor(data) 只适用于没有缺失数据情况。
- 如有缺失数据，则使用
  - ✓ cor(data, use= “complete.obs”)
    此时，只使用完整的数据。
  - ✓ cor(data,use= “pairwise.complete.obs”)
    计算特定的一对变量的相关系数时，只使用相应变量的完全的观测值。这样计算得到的相关矩阵有可能是不正定的。
- 诊断共线性的一个更好的方法是计算每个变量的R_j^2:
  即把$\mathbf{X}_j$作为响应变量与剩余协变量做回归得到确定性系数.

模型选择

- 主要内容基于Faraway Chapter 10.
- 术语:
  - ✓ 预测因子(predictor)=自变量
  - ✓ 响应变量(response) =因变量
- **我们想用最简单的方法来解释这些数据。用最小模型拟合数据是最好的。**
- 如果模型中有很多预测因子会发生什么:
  - ✓ 有可能产生共线性，导致标准误差增加。
  - ✓ 浪费时间/金钱来测量或收集不必要的预测因子。
  - ✓ 模型可能会变得过于复杂而无法解释。

第一步:

- 识别异常值、杠杆点和影响点。这些点对模型选择有很大的影响，所以先暂时把这些点排除在外可能是好的选择。
- 增加预测变量的适当变换。
- 如果增加预测变量的高阶项，要遵循如下的<u>边际原则</u>:
  - ✓ 如果$\mathbf{X}_1^2$在模型中，那么$\mathbf{X}_1$也需要在模型中。
  - ✓ 如果$\mathbf{X}_1\mathbf{X}_2$在模型中，那么$\mathbf{X}_1$和$\mathbf{X}_2$也需要在模型中。

模型选择方法

- 运用所学领域的知识，包括系数的符号和大小。
- 逐步方法:
  - ✓ 向后法（backward）
  - ✓ 向前法（forward）
  - ✓ 逐步法（stepwise）
- 详尽的搜索:
  - ✓ 考虑所有可能的模型，并使用一些**准则**对它们进行比较。
- 现代的高维变量选择方法(lasso, elastic-net, 等)

基于p值向后法

- 从包含所有预测变量的模型开始。
- 去掉对应最大的p值大于预先给定的α_{Drop}的预测变量。对于剩余的变量继续拟合，直到所有的p值都小于预先给定的α_{Drop}。
- α_{Drop}未必是0.05。如果我们建立模型的目的是预测的话，则α_{Drop}选择的大一点可能更好，比如，α_{Drop}在$0.15 - 0.20$之间。
- 有用的R命令: drop1()，update()。

基于p值的向前选择

- 从模型中没有变量的模型开始。
- 对于不在模型中的所有预测变量，计算它们的p值以便将它们添加到模型中。选择具有最小p值的那个变量，并在p值小于预先给定的α_{Add}的变量将其添加到模型中。重复此过程，直到无法添加新的预测变量。
- 有用的R命令: add1()，update()。

基于p值的逐步回归

- 逐步回归法是向前法和向后法选择的结合。在每个步骤中，我们都可以添加或删除一个变量。

优点和缺点

- 基于p值的逐步方法的优点：
  - ✓ 易于解释
  - ✓ 易于计算/使用
  - ✓ 使用广泛
- 逐步方法的缺点：
  - ✓ 因为每次是丢弃和添加一个变量，所以可能会错过“最优模型”。
  - ✓ 这个方法有可能会夸大结果的重要性。不要对p值太相信。我们做了很多的检验，因此存在**多重检验的问题**【去google或者https://enwikipedia.org/wiki/Multiple_comparisons_problem】
  - ✓ 临时方法：所选模型不需要优化任何合理的标准。
  - ✓ 前向和后向选择的结果可能不同。【R示例】

更原则的方法

- 现代方法：
  - ✓ 选择比较结果的标准：AIC，BIC，调整的R^2, C_p 等。搜索所有可能的模型，并根据您的标准考虑最佳模型。
  - ✓ 在多变量的问题中，使用凸松弛方法（例如，Lasso或ElasticNet）
- 查看最佳模型之间的差异。如果它们的差异很大，那么使用哪种模型存在很大的不确定性。
- 根据您对问题的背景知识，选择一个或两个似乎有意义的模型。

AIC和BIC

- Akaike信息准则（AIC）：$-2(\text{loglikelihood}) + 2(\text{number of parameters})$.
- 贝叶斯信息准则（BIC）：$-2(\text{loglikelihood}) + (\log n)\text{number of parameters}$.
- 对于具有高斯性假设的线性回归，$-2(\text{loglikelihood})$与$n\log(\text{SSE}/n)$成比例【黑板】。因此，AIC和BIC结合了拟合优度（小SSE/大的对数似然）的度量与模型复杂性（number of parameter 参数数量）的惩罚。
- 我们想要找到一个有小AIC或BIC的模型。
- 对于大型数据集，BIC对模型中的参数数量有较重的惩罚，因此往往会产生较小的模型。
- 我们不一定在寻找最好的模型。

Mallow C_p准则

- 良好的模型应该具有较小的均方误差预测：

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{E}[\hat{y}_i - \mathbf{E}[y_i]]^2.$$

- 这个预测误差可以通过C_p统计量来估计：

$$C_p = \frac{\text{SSE}}{\hat{\sigma}^2} + 2p - n,$$

其中SSE是给定模型的残差平方之和，p是模型中变量的个数，$\hat{\sigma}^2$是使用全模型给出的σ^2的估计。

Mallow C_p准则

- 注意：
  - ✓ C_p与AIC密切相关
  - ✓ 对于全模型，SSE $= (n-p)\hat{\sigma}^2$，因此$C_p = p$.
  - ✓ 如果具有p个变量的模型预测良好，然后$C_p \approx p$. 通常，不好的模型具有较大的C_p值。
- 通常要绘制C_p与p的图形。我们想要p较小的模型，且C_p大约为p或小于p.

结束语

- 数据拟合的好并不能保证良好的预测：
  - ✓ 小的数据集避免复杂的模型。
  - ✓ 尝试获取新数据以验证你的模型。
  - ✓ 使用类似数据的过去经验来指导模型选择。
- 有用的R命令：
  - ✓ leaps()（来自于package leaps）：使用C_p（默认）或调整的R^2进行穷举搜索。
  - ✓ 使用AIC（默认）或BIC（使用选项$k = \log(n)$）逐步搜索.