

1. KL 散度  $D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx \geq 0$   
 $D_{KL}(P||Q) = 0$  iff  $P=Q$ , 不对称, 不满足  $\Delta$  不等式  
 $\min D_{KL}(P||P_0) \Leftrightarrow \max_{P_0} \prod_{i=1}^n P_0(x_i)$   $P_0(x)$  为真实分布  
 等价于最大似然

2. 偏度  $\tilde{\mu}_3 = E[(\frac{X-\mu}{\sigma})^3]$  峰度  $Kurt[X] = E[(\frac{X-\mu}{\sigma})^4]$

3. Gamma:  $f(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}$   $k$ : shape  $\theta$ : scale

4. bias-variance  $MSE(\hat{\theta}) = (E(\hat{\theta}) - \theta^*)^2 + Var(\hat{\theta})$   
 Cramér-Rao 下界: 无偏估计的下界  $Var(\hat{\theta}) \geq \frac{1}{n I(\theta^*)}$

$I(\theta) = E_x(\frac{\partial}{\partial \theta} \ln P(x))^2 = -E_x(\frac{\partial^2}{\partial \theta^2} \ln P(x))$   
 5. MLE 的大样本性质:  $P(\hat{\theta}^{MLE}) \xrightarrow{n \rightarrow \infty} N(\theta^*, \frac{1}{n I(\theta^*)})$ , 渐近无偏, 渐近有效

6. 如果重复进行实验, 并为每个实验构造置信区间, 95% 能拿下真值

7. GMM: 多个分布  $(\phi_1(x), \phi_2(x), \dots, \phi_k(x))$  的堆叠, 每个类别有概率  $\pi_k$   
 $P(x) = \sum_{k=1}^K \pi_k \phi_k(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k)$ ,  $\int \phi_k(x) dx = 1$ ,  $\sum_k \pi_k = 1$

学习过程: 最大似然  $\max_{\theta} \sum_{i=1}^n \log P(x_i|\theta)$  但参数值易有较大 (singularity)  
 使用 EM 迭代求解最优参数, 无监督

8. 模型选择: 奥卡姆剃刀: 选择最简单的模型, 更简单的模型越大越好  
 $AIC = \log P(D|\theta^{MLE}) - M$  ( $M$  为模型参数个数)  
 $BIC = \log P(D) = \int \log P(D|\theta) P(\theta) d\theta \approx \log P(D|\theta^{MLE}) - \frac{1}{2} M \log n$  (适用指数族)  
 $n$  为数据量,  $BIC$  为复杂度的惩罚大于  $AIC$ ,  $BIC$  越越好

9. 频率学派  $\theta$  是个常数, 置信区间为随变, Bayesian  $\theta$  为随变, 可信区间随变

10. 两样本  $t$  统计量  $H_0: \mu_1 = \mu_2$   $T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$   
 $P$  值: 在  $H_0$  为真的假设下, 观察到至少与所得统计量一样极端值的概率, 较小  $P$  值意味着  $H_0$  为真, 也可保证某些错误概率尽可能小  
 Type I:  $H_0$  为真拒绝 Type II:  $H_0$  为假接受,  $H_0$  要么成立, 要么不成立, 因此没有  $P(H_0)$  一说,  $P$  值实验前为随变, 实验后为数值

11. 为防止  $P$  的误导性 (不能反映实际意义大小, 并非假设成立的概率), 可以同时报告置信区间大小, effect size  $d = \frac{\bar{Y} - \bar{X}}{s_p}$  两类别差异

12. 多重检验:  $m$  个  $H_0$ ,  $m$  个  $P$ , 当拒绝每个  $n$  个  $H_0$  时, 会出现大量 Type I  
 FWER: 至少有一个 Type I 的概率  $= 1 - \prod_{i=1}^m (1 - \alpha) = 1 - (1 - \alpha)^m$   
 Bonferroni Correction: 为使 FWER 至少为  $\alpha$ , 需要设置每个假设为  $\frac{\alpha}{m}$   
 但  $m$  过大时 Bonferroni 会使得  $\alpha$  极小, 极保守使得 Type-II 升高, 故  $m$  过大时控制 FDR  $= E(\frac{\text{错误拒绝}}{\text{全部拒绝}})$ , 可以容忍一些 Type-I, 只要 FDR 不过 9

13. sigmoid  $(z) = \frac{1}{1+e^{-z}} = \pi$  反函数为  $z = \log(\frac{\pi}{1-\pi}) = \log(\text{odds}(\pi))$   
 $\pi_i = \text{sigmoid}(x_i^T \beta)$   $\beta$  意味着为每增加一个单位,  $\log(\text{odds}(\pi))$  增加  $\beta_k$   
 $lik(\beta) \propto \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$  (对每个  $x_i$  作  $n_i$  次观测,  $y_i$  次成功)

14. log-linear model 用 logit 作因变量  $\log \frac{P(Y=1|X=i, Z=k)}{P(Y=2|X=i, Z=k)}$   
 $Z \rightarrow Y$   $\logit[P(Y=1|X=i, Z=k)] = \alpha + \beta_1 x + \beta_2 z$   
 $X \rightarrow Y$   $\logit[P(Y=1|X=i, Z=k)] = \alpha + \beta_1 x$   
 评估拟合效果可用 LR test (likelihood ratio), AIC, BIC 越小越好

15. 召回率:  $TPR = \frac{TP}{TP+FN}$  精确率:  $PPV = \frac{TP}{TP+FP}$   $FPR = \frac{FP}{TN+FP}$   
 $FPR - TPR$  构成 ROC 曲线, 围成的面积为 AUC  
 (绘制方式为采用一系列 thresholds, 如选择  $\pi(x_i) > 0.5$  时令  $y_i = 1$ , 使用不同 thresholds 下的  $(FPR, TPR)$  绘制曲线) 一般  $0.5$  较好, 瞎猜  $0.5$   
 但对不平衡数据, ROC 过于乐观, 此时应选 Recall-Precision 曲线

16. 模型诊断: 线性回归 - 残差图, 逻辑回归 - pearson residual deviance residuals

17. leverage:  $\hat{\beta} = (X^T X)^{-1} X^T y$ ,  $\hat{A} = X(X^T X)^{-1} X^T$ ,  $tr(\hat{A}) = p+1$   
 average leverage  $= \frac{p+1}{n}$  high leverage  $\rightarrow H_{ii} > \frac{p+1}{n} \times (2 \text{ or } 3)$

判断 influential observations: Cook's distance  $D_i$  为  $\hat{\beta}$  与  $\hat{\beta}_{(-i)}$  的差  
 high influence  $\rightarrow D_i > 1$

18. 多重共线性: 增加标准保, 为参数解释制造困难, 使模型敏感  
 方差膨胀无偏, 但 Var 变大且增大过拟合风险, 使用 VIF 检测  
 解决方法: ① 增加数据 ② 变量中心化 ③ 删除冗余变量

19. 模型选择: 复杂模型拟合效果更好, 但方差更大  
 估计方差误差: K-Fold 交叉验证, 最佳子集选择法, step-wise selection

20. 伯松回归: 读书的本数, 进球的个数  
 $\log(\mu_i) = x_i^T \beta = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$   $y_i \sim \text{Pois}(\mu_i)$ , 线性模型差恒定  
 伯松回归的差与均值一同增长  $y_i \sim \text{Pois}(\mu_i = \exp(\alpha + \beta x_i))$   
 可用 pearson residual 或 deviance residual 评估 local lack-of-fit

21. null model: 只有截距  $\log \mu_i = \alpha$  即  $y_i \sim \text{Pois}(\mu_i = e^\alpha)$

22. overdispersion: 伯松回归经常低估数据的方差  
 卡方统计量  $X^2 \sim \chi^2(df)$  且  $E(X^2) = df$ , 若  $X^2/df > 1$ , 则有 overdispersion 现象  
 或  $H_0: Var(y) = \mu \Leftrightarrow H_1: Var(y) = \mu + \alpha \times f(\mu)$

23. 负二项模型, 伯松回归中  $\mu_i$  为常数, 可使其变为随变  $\tilde{\mu}_i = \mu_i \epsilon_i$   
 通常假设  $\epsilon_i$  服从 Gamma 分布  $E(\epsilon_i) = 1$   $Var(\epsilon_i) = \frac{1}{k}$   
 积分消去  $\epsilon_i$  后  $P(y_i|\mu_i, k)$  服从负二项分布  $E(y_i) = \mu_i$   $Var(y_i) = \mu_i + \frac{\mu_i^2}{k}$   
 $k$  越小, 与伯松模型相比, overdispersion 越大

24. 带正则项回归, 动机 ① 多重共线性 ② 高维数据 (特征维数  $p > n$ )  
 $X^T X$  不满秩且不可逆 ③ 降低测试误差

25. Ridge: 使用  $L_2$  范数惩罚回归系数  $\hat{\beta} = \argmin_{\beta} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2$   
 $L_2$  范数中包含截距项, 假设所有  $x$  已中心化且标准化  
 $MSE(f) = \frac{1}{n} \sum_{i=1}^n [Bias(x_i^T \hat{\beta})]^2 + \frac{1}{n} \sum_{i=1}^n Var(x_i^T \hat{\beta}) = 0 + \frac{p\sigma^2}{n}$ , 即每个变量都会为 MSE 增加  $\sigma^2/n$ , 因此需要 shrinkage, 会导致一些 bias 但减小 variance  
 Ridge 将系数缩小, 但无法达到真正的 0, 因此并不能做到变量选择。  
 当真正应为 0 的系数很多时, ridge 效果好, 但当大部分系数都不为 0 时, 效果一般

26. LASSO:  $L_1$ -norm 惩罚  $\hat{\beta} = \argmin_{\beta} ||y - X\beta||_2^2 + \lambda \sum_{j=1}^p |\beta_j|$   
 LASSO 可使系数真正归 0 实现 variable selection, 没有精确解, 一般通过坐标下降法求解, Ridge 随  $\lambda$  个, 方差快速  $\downarrow$ , LASSO 慢速  $\downarrow$ , 故 LASSO 对小的  $\lambda$  不偏好, 可通过 CV 找出最优  $\lambda$



# 统计数据分析第三次作业

## 1 模型参数和置信区间

在频率学派(frequentist)中，关于统计模型的参数和其置信区间的解释，以下说法错误的是。

- 模型的真实参数  $\theta$  是一个常数
- 模型的真实参数  $\theta$  是一个随机变量
- 在观测数据前，置信区间是一个随机区间
- 对于一组给定的数据，置信区间是一个具体的数值区间

错误的选项是2，这是因为，在频率学派中，模型的参数被视为一个固定的常数，只是这个参数可能是未知的，在统计推断的时候，我们就是希望通过尽可能多的样本的信息，来估计这个未知的参数。而在贝叶斯统计中，模型的参数就被视为一个随机变量。

随机变量的定义如下：

设  $(\Omega, \mathcal{F}, P)$  是一个概率空间， $X$  是定义在  $\Omega$  上的函数，若对任意的  $x \in \mathbb{R}$ ，有：

$$\{X \leq x\} \triangleq \{\omega \in \Omega \mid X(\omega) \leq X(x)\} \in \mathcal{F}$$

则称  $X$  是  $(\Omega, \mathcal{F})$  上的随机变量。

从定义可知，随机变量实际上就是满足了特定性质的函数。

## 2 假设检验问题

某医院在研究一种意图降低血脂的药物的效用，分别对治疗组和安慰剂组的病人测量服食药物或安慰剂前后血脂水平的变化。原假设为：治疗组和安慰剂组的效用相同。应用两样本 t-检验，得出 p 值为 0.02。对此，以下说法正确的是：

- 该药物无效的概率为 2%
- 该药物相对于安慰剂有 2%的提升
- p 值<0.05，证明了原假设是错的
- p 值<0.05 不能证明原假设是错的

显然，1和12是错误的，p值越小，越倾向于拒绝原假设，因此最合理的选项应该是3.

## 3 置信区间和显著性水平

对于某个参数估计问题，有原假设  $H_0: \theta = 0$ 。设基于样本得到的参数  $\theta$  的 95%置信区间为[1, 3]，则我们可以在显著性水平0.05下拒绝原假设

## 4 对数线性模型和列联表

评价某对数线性模型(log-linear model)是否能够准确地描述列联表数据，可使用什么统计量？

可以使用以下两个统计量：

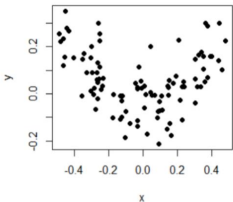
- likelihood ratio statistic（似然比统计量）

- 预处理数据：  
首先整理三阶列联表data\_rab，确保数据完整性和准确性。
- 分层：  
根据Z值的不同取值（例如：z1、z2、z3...），将data\_rab划分为若干层（分组）。每一层都包含了在特定Z值下的X和Y的分布情况。这样可得到一个分层的列表表。
- 应用CMH Test：  
对每个层进行卡方检验，计算每个层的卡方统计量和自由度。累加各层的卡方统计量和自由度，得到整体卡方统计量和自由度。
- 计算p值：  
依据整体卡方统计量和自由度，计算p值。p值可以帮助我们判断在Z控制的情况下，X和Y是否存在条件独立。若p值小于预设的显著性水平（通常为0.05），则拒绝原假设（X和Y条件独立），接受备选假设（X和Y不独立）；若p值大于显著性水平，则无法拒绝原假设。
- 结果解释：  
根据CMH Test的检验结果：
  - 如果拒绝原假设（X和Y条件独立），表明在Z控制下，X和Y之间存在关联或依赖关系。此时需要进一步分析此关系的具体成因和影响。
  - 如果无法拒绝原假设，表明在Z控制下，X和Y条件独立。可据此对进一步的数据分析进行规划。
- 进一步的数据分析：  
在验证条件独立性之后，可以基于CMH Test的结论，对数据进行深入分析。可能的方向有：
  - 探究X、Y、Z之间的关系，考察它们如何相互影响。
  - 根据CMH Test的结果，确定哪些变量应纳入回归模型或其他统计模型以解释目标变量的变化情况。
  - 进行更细致的分析，例如，针对不同Z值下的分层，分别探究X和Y的关联强度和特征。这有助于发现潜在的非线性关系和相互影响。

总之，小明可以通过CMH Test来验证X、Y、Z三个属性之间的条件独立关系，并根据检验结果来规划后续的数据分析思路，深入挖掘 X、Y、Z 之间的关系以及它们对其他变量的影响。

## 7 线性回归模型

小明对一组二元数据  $\{(\alpha^*, y^*)\}$  作简单线性回归:  $y = a + b \cdot x$ 。数据的散点图如下图所示；指出小李所得的模型可能存在的问题。



## 5 回归模型泛化问题

小明想得到一个泛化能力好的回归模型；他发现他的模型在训练集上表现得很好，但是在测试集上表现得很差。对此，以下的应对措施合理的是。

- 引入更多的解释变量(explanatory variables)
- 引入变量间的相互作用
- 使用正则化方法(regularization)以控制模型复杂度
- 把模型在训练集和测试集上一起训练

答案选择3.

## 6 三阶列联表

小明得到一个三阶列联表 data\_rab，它有 X、Y、Z 三个属性。根据他对该领域的知识，小明判断这三个属性有如下的条件独立关系： $P(X, Y|Z) = P(X|Z)P(Y|Z)$  (可记为  $X \perp Y | Z$ )。简要描述小明应如何对此作数据分析（可使用自然语言、数学语言或 R 语言）

ChatGPT的回答：

为了根据条件独立关系  $X \perp Y | Z$  对三阶列联表 data\_rab 进行分析，小明可以采取以下步骤：

- 计算各属性的边缘分布：首先，需要计算各属性的边缘分布，即  $P(X)$ 、 $P(Y)$  和  $P(Z)$ 。这可以通过对 data\_rab 的每个属性级别进行计数并除以总观测数来完成。
- 计算条件分布：根据条件独立关系，需要计算各属性的概率条件分布： $P(X|Z)$  和  $P(Y|Z)$ 。通过计算在给定 Z 属性下，X 和 Y 的相对频率来得到条件分布。
- 计算  $P(X, Y|Z)$  和  $P(X|Z)P(Y|Z)$ ：根据步骤 2 计算出的条件分布和给定的条件独立关系，对于每一对 X 和 Y，以及每个 Z 属性值，计算  $P(X, Y|Z)$  和  $P(X|Z)P(Y|Z)$  的值。
- 验证条件独立关系：比较步骤 3 中计算得到的  $P(X, Y|Z)$  和  $P(X|Z)P(Y|Z)$  的值，若它们在一定误差范围内接近，那么可以认为给定的条件独立关系有效。
- 数据分析及其他研究：在验证了条件独立关系的基础上，小明可以利用这些信息对数据进行进一步分析。例如，探讨 X 和 Y 在给定 Z 时的相互关系、了解 Z 属性如何影响 X 和 Y 之间的关联等。

通过以上步骤，小明可以利用给定的条件独立关系  $X \perp Y | Z$  对三阶列联表 data\_rab 进行分析。

Merlin 使用GPT4.0的回答：

首先，我们了解一下Cochran-Mantel-Haenszel (CMH) Test的基本概念。CMH Test是一种用于验证在控制一个或多个混杂变量的情况下，两个分类变量之间条件独立关系的统计检验方法。在本例中，条件独立关系为  $X \perp Y | Z$ ，即在已知 Z 的情况下，X 和 Y 是独立的。

小明可以按照以下步骤进行数据分析：

首先，多元线性模型的假设是：

多元线性回归模型(MLR)：变量  $y$  与一般变量  $x_1, x_2, \dots, x_p$  之间的理论线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

$\beta_0$ 为回归常数项， $\beta_1, \beta_2, \dots, \beta_p$ 为回归系数， $u_0$ 为随机误差。随机误差满足六个基本假设：线性于参数、随机抽样、解释变量的样本有波动、零条件均值、同方差性、误差项正态性。

一元线性回归模型就是多元线性回归模型的特殊情况。

直观上看，由二维数据样本点图可知，数据样本并不是线性的，因此，如果直接套用一元线性回归模型来对本样本进行建模的话，那么，零条件均值和同方差性是很难得到保证的。因此，拟合出来的直线并不能很好得展示数据的信息。

此外，误差项具有正态性的假设可能也很难保证，虽然无法保证这个要求的话，线性模型依旧可以建立，但后续的统计推断就很难进行。

另外，从模型复杂度考虑的话，选取一元线性回归模型来拟合这个样本集是不合适的，因为模型参数太少了，会导致欠拟合的问题。

综上所述，如果假设样本数据是准确无误的，那么一元线性回归模型可能并不适合这个数据集。

## 8 极大似然估计

假设随机变量  $X \in \{1, 2, 3\}$  服从的分布为

X	1	2	3
p	$\theta$	$2\theta$	$1-3\theta$

现从中抽取一个容量为  $n$  的样本，记录下其中1出现的次数为  $n_1$ ，2出现的次数为  $n_2$ ，3出现的次数为  $n_3$ 。给出参数  $\theta$  的最大似然估计。

显然这是一个多项分布的极大似然估计问题。

记随机变量  $X_1, X_2, X_3$  为三个单元格的观测数，服从总数为  $n$ ，概率为  $p_1, p_2, p_3$  的多项分布。于是  $X_1, X_2, X_3$  的联合密度函数为：

$$f(X_1 = n_1, X_2 = n_2, X_3 = n_3 | p_1, p_2, p_3) = \frac{n!}{n_1! n_2! n_3!} \prod_{i=1}^3 p_i^{n_i}$$

$$l(\theta) = n! - \sum_{i=1}^3 \log n_i! + n_1 \log \theta + n_2 \log 2\theta + n_3 \log (1 - 3\theta)$$

显然对数似然函数可导，导函数为：

$$\begin{aligned} l'(\theta) &= \frac{n_1}{\theta} + \frac{n_2}{\theta} - \frac{3n_3}{1-3\theta} = 0 \\ \iff (n_1 + n_2)(1-3\theta) - 3n_3\theta &= 0 \\ \iff (n_1 + n_2) - (n_1 + n_2 + 3n_3)\theta &= 0 \\ \iff \theta &= \frac{n_1 + n_2}{n_1 + n_2 + 3n_3} \\ \iff \theta &= \frac{n_1 + n_2}{n + 2n_3} \end{aligned}$$