示例带回家作业

StatBI/E

Formalities

- *This is an example of what the questions in a final take-home assignment in the StatBI/E course may be like.*
- 任务需要独立完成，禁止作弊（包括小组合作）。

- 您需要提交一份包含所有考试问题解答的单一PDF文件。该PDF文件必须包含所有书面答案、数学推导、R代码和图表。您可以包括清晰的扫描/照片，例如手写的推导（这可以通过Rmd中的`![figurecaption](/path/to/image.png)`实现）。然而，更可取且最简单的方法是使用Rmarkdown准备您的解答以生成PDF（在RStudio编辑器工具栏中点击展开“ Knit ”下拉菜单并选择“ Knit to PDF ”）。请参考考试Knit测试（参见Ab-salon）以解决常见的编织问题。（作为备选方案，您可以编织成HTML/DOC，打开生成的文档，并通过“ 打印为PDF ”将此文档转换为PDF。）

- 清晰地指出你正在回答的问题，并给出完整的答案和理由，也就是说，仅仅在PDF中展示一个代码块及其输出是不够的。例如，如果在问题12中给你一个由n个独立同分布的随机变量X的样本，并要求你提供一个点估计，那么以下的`> mean(data)`是不充分的，更好的选择可能是

```
Answer to Problem 12
The empirical average is a point estimator of the expected
value. We use R to calculate the empirical average
of the given sample and print the result below > mean(data)。
[1] 39.183
```

上述示例中的代码可以认为是不言自明的，因为它只是一行代码，调用了一个具有自解释名称的命令。

`mean`. 较长的代码块通常不是自解释的。不要依赖我们猜测您试图在图表中实现/展示的内容，而应提供一个完整的自包含答案，包括文本+代码+输出。这也使我们能够在代码存在错误导致数值结果/图表错误的情况下，可能给予部分分数，只要其余答案正确解释了您意图通过代码实现的内容。

- 如果你发现任何一个问题陈述中有问题，请在你的解答中明确指出并描述。例如，如果你怀疑问题陈述的某些方面可能存在歧义，请明确指出你认为可能存在歧义的地方以及你对陈述的理解，以便你能够继续回答问题。

- 如果你使用既不属于R基础发行版，也未在课程中介绍的R包（用于解决每周练习的方案），你需要对这些包的正确运行负责。

- 始终确保在提交PDF之前进行双重检查。

Data

考试中要使用的数据集可以从Absalon下载。文件名为`practice_exam.RData`，该文件的链接放置在课程主页上。通过用Rstudio打开文件，或使用`load`函数来加载数据。提示：在加载数据前清理R会话，之后使用命令`ls()`来检查已加载的对象。这个RData文件中有三个数据集：`radiation`、`bodyfat`和`wind`。数据`radiation`和`bodyfat`是数据框对象，而`wind`是一个简单的值向量。

Problems

Problem 1 (30 points)

数据集 `radiation` 包含了暴露于不同辐射水平并接受链霉素治疗的小鼠生存实验的结果。特别是，该数据集包含了522只小鼠的实验结果，其中记录了三个变量：

- `dose` 中子剂量（201, 220, 243, 260）
- `treatment` 治疗（1 = 链霉素，0 = 生理盐水对照）
- `dead` 如果老鼠在3-10天内死亡（1 =死亡，0 =存活）

Question 1.1 计算中子剂量和治疗8种不同组合的估计死亡概率。将估计概率作为中子剂量的函数进行绘图，用蓝色表示接受链霉素治疗的鼠，红色表示接受生理盐水对照的鼠（在一张图上展示所有概率）。

Question 1.2 我们现在调查链霉素是否有效减轻辐射的影响。首先，我们测试链霉素是否对不同辐射剂量水平的死亡概率有某些相关影响。对`treatment`和`dead`的不同中子剂量的列联表进行卡方独立性检验。在R中，你可以使用`table`函数来构建列联表，例如，`treatment-dead`的`dose`= 201列联表可以通过以下代码获得：

```
table(radiation[radiation$dose == 201, c(2, 3)])
```

特别是对以下零假设进行卡方检验并解释结果（我们在 $\alpha = 0.05$ 时拒绝吗？）：

- 对于中子剂量等于201，`dead`与`treatment`无关。
- 对于中子剂量等于220，`dead`与`treatment`无关。
- 对于中子剂量等于243，`dead`与`treatment`无关。
- 对于中子剂量等于260，`dead`与`treatment`无关。

Question 1.3 为了比较在四种可能的 neutron 剂　剂量下不同治疗方案的死亡概率，我们现在进行单侧 Wald 检验，使用（渐进正态的）统计量：

$$\delta = \hat{p}_1 - \hat{p}_0$$

其中 $\hat{p}_1$ 是接受链霉素治疗的小鼠在相同中子剂量下的估计死亡概率，$\hat{p}_0$ 是接受生理盐水治疗的小鼠在相同中子剂量下的估计死亡概率。

特别要检验以下零假设，报告p值，并评论在$\alpha = 0.05$时是否拒绝零假设。

- 对于中子剂量等于201的情况，接受链霉素治疗的小鼠存活的可能性小于或等于接受生理盐水对照治疗的小鼠。
- 对于中子剂量等于220的情况，接受链霉素治疗的小鼠存活的可能性小于或等于接受生理盐水对照治疗的小鼠。
- 对于中子剂量等于243的情况，接受链霉素治疗的小鼠存活的可能性小于或等于接受生理盐水对照治疗的小鼠。
- 对于中子剂量等于260的情况，接受链霉素治疗的小鼠存活的可能性小于或等于接受生理盐水对照治疗的小鼠。

当我们可以说链霉素治疗在对抗辐射影响方面是有效的？

提示：要使用上述定义的统计量 $\delta = \hat{p}_1 - \hat{p}_0$ 进行沃尔德检验，我们需要估计 δ 的标准误。由于 $\hat{p}_0$ 和 $\hat{p}_1$ 是伯努利随机变量参数的估计值，我们可以得到 δ 方差的闭式公式，从而得到标准误。记住，如果 $X \sim Bernoulli(p)$，则我们有 $\mathbb{E}(X) = p$ 和 $\mathbb{V}(X) = p(1-p)$。

如果你无法获得 δ 标准误差的封闭形式表达式，你可以使用自助法。

Question 1.4 我们现在进行逻辑回归，以预测作为中子剂量函数的死亡概率。仅使用接受链霉素治疗的小鼠的观察结果来拟合模型

$$\text{logit}(\mathbb{E}(\texttt{dead}|\texttt{dose})) = \beta_0 + \beta_1 \texttt{dose} \qquad \text{(log-regr 1)}$$

其中 `dead|dose` 是一个伯努利随机变量。如问题1.1中所示，绘制模型（log-regr 1）预测的死亡概率作为中子剂量的函数。在同一图中，用不同颜色添加从数据中估计的概率（针对用链霉素处理的小鼠）。

拟合，使用经链霉素处理的小鼠数据，其他两个逻辑回归模型，现在包含多项式项，

$$\text{logit}(\mathbb{E}(\texttt{dead}|\texttt{dose})) = \beta_0 + \beta_1 \texttt{dose} + \beta_2 \texttt{dose}^2 \qquad \text{(log-regr 2)}$$

$$\text{logit}(\mathbb{E}(\texttt{dead}|\texttt{dose})) = \beta_0 + \beta_1 \texttt{dose}i + \beta_2 \texttt{dose}^2 + \beta_3 \texttt{dose}^3 \qquad \text{(log-regr 3)}$$

对于上述三个模型，分别显示估计系数的值。

使用AIC进行模型选择，以在三元逻辑回归模型之间进行选择。此外，执行似然比检验以在log-regr 1和log-regr 2之间进行选择。

Problem 2 (30 points)

个体体脂百分比可以在确定身体密度后进行估算。体积，进而身体密度，可以通过多种方式准确测量。水下称重技术通过计算空气中测量的体重与水中浸没时测量的体重之间的差值来计算身体体积。因此，准确测量体脂既不方便又成本高昂，因此希望有简单易行且不昂贵的方法来估算体脂。目标是能够通过简单的身体测量来估算体脂。

数据集 `bodyfat` 包含通过水下称重和各种身体围度测量确定的252名男性体脂百分比的估算值。

目标是找到一个回归模型，仅通过身体测量来估算体脂百分比。

在 特别是数据集 bodyfat 包含以下 va 变量：

- fat 水下身体测量估算的体脂百分比

- age 年龄（以年为单位）
- weight 体重（千克）
- height 高度（米）
- neck 颈部周长（厘米）
- chest 胸围（厘米）
- abdomen 腹部（2）周长（厘米）
- hip 臀围（厘米）
- thigh 大腿围（厘米）
- knee 膝盖围长（厘米）
- ankle 踝围（厘米）
- biceps 肱二头肌伸展围（厘米）
- forearm 前臂围长（厘米）
- wrist 手腕周长（厘米）

Question 2.1 为变量fat (体脂百分比)拟合一个线性回归模型，使用数据集中的所有其他身体测量值作为预测变量。使用函数summary提取线性回归系数的信息。哪些特征似乎对预测体脂百分比最为相关？请说明理由。

我们能否在 $\alpha = 0.05$ 拒绝认为 knee (膝盖围度) 的系数等于 0？请使用通过 summary 函数获得的信息进行解释和论证。

Question 2.2 我们现在想要获得更简单的回归模型来预测脂肪百分比（fat）。使用BIC作为评分标准，进行前向和后向逐步回归模型选择。你可以使用内置函数step。

Question 2.3 通过身体测量来估算体脂百分比的一种常见方法是使用身体质量指数（BMI）。

BMI定义为个体体重（以千克为单位）除以身体高度的平方（以米为单位）。计算数据集中所有个体的身体质量指数。

拟合以下线性模型以估计体脂百分比（fat）：

$$\mathbb{E}(\texttt{fat}|\texttt{bmi}, \texttt{age}) = \beta_0 + \beta_1 \texttt{bmi} + \beta_2 \texttt{age} \qquad \text{(BMI模型)}$$

其中 bmi 是身体质量指数。报告拟合系数。

Question 2.4 计算问题2.3中模型系数β_0, β_1和β_2的95%百分位置信区间，使用非参数自助法。将获得的置信区间与使用R内置函数`confint`获得的区间进行比较。

Question 2.5 比较在问题2.2中获得的模型和问题2.3中获得的模型，使用BIC和留一交叉验证估计的错误（$\hat{R}_{CV} = \sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2$，其中$\hat{Y}_{(i)}$是通过拟合省略观测值$(X_i, Y_i)$的模型获得的$Y_i$的预测）。

Problem 3 (40 points)

在这个问题中，我们将研究Gumbel分布。

耿贝尔分布是一个连续分布，其密度函数（PDF）由以下表达式给出，

$$f_{Gumbel(\mu,\beta)}(x) = f(x|\mu,\beta) = \frac{1}{\beta}\exp\left(-\frac{(x-\mu)}{\beta} - e^{-\frac{(x-\mu)}{\beta}}\right)$$

或者换句话说

$$f(x|\mu,\beta) = \frac{1}{\beta}e^{-(z+e^{-z})}$$

其中 $z = \frac{x-\mu}{\beta}$ 和 $\mu \in \mathbb{R}$ 是位置参数，$\beta > 0$ 是尺度参数。

此外，如果 $X \sim Gumbel(\mu, \beta)$，我们有期望值和方差的公式：

$$\mathbb{E}(X) = \mu + \beta\gamma \qquad \mathbb{V}(X) = \frac{\pi^2}{6}\beta^2$$

其中 γ 是欧拉-马歇罗尼常数，在R中可以计算为

$$\gamma = -\texttt{digamma}(1)$$

近似值为 $\gamma \approx 0.5772$。

耿贝尔分布的累积分布函数（CDF）为，

$$F_{Gumbel(\mu,\beta)}(x) = \exp\left(-e^{-(x-\mu)/\beta}\right)$$

并且分位数函数由ings由以下公式给出，

$$Q_{Gumbel(\mu,\beta)}(p) = \mu - \beta\log(-\log(p))$$

所有上述公式和表达式均可用于解决方案中。

Question 3.1 在R中实现函数dgumbel (PDF)、pgumbel (CDF)、qgumbel (分位数函数)和rgumbel (抽样)。这些函数的行为和签名应类似于dnorm、pnorm、qnorm、rnorm以及R中内置的PDF、CDF、分位数函数和随机数生成函数。特别是,

- 对于 dgumbel，R 函数应按如下方式工作：dgumbel(x，mu，b)。其中 x 可以是单个数值或数值向量。该函数应返回一个向量，其值为 Gumbel 分布在点 x 处的密度值。可选地，您还可以实现附加参数 gumbel(x，mu，b，log)，其中 log 可以是 TRUE 或 FALSE，其行为与 dnorm 中相同。
- 对于 pgumbel，R 函数应按 pgumbel(q，mu，b) 工作，其中 q 是分位数的向量，mu,b 是 Gumbel 分布的参数。可选地，您可以实现附加参数及其关联行为 lower.tail 并检查 log.p (行为 pnorm)。
- qgumbel(p，mu，b) 应该返回每个概率向量 p 对应的分位数向量。可选地添加参数 lower.tail。
- rgumbel(n，mu，b) 应返回大小为 n 的样本，该样本分布为具有参数 μ=mu 和 β =b 的独立 Gumbel 随机变量。记住逆变换抽样。

在所有函数中，参数 mu (位置参数 μ) 和 b (尺度参数　参数 β) 应该具有默认值 mu = 0、{v46　46}。

为了测试上述函数，使用上面实现的函数 rgumbel 从一个 Gumbel 分布中生成一个大小为 10000 的样本。绘制直方图，并在其上绘制真实密度 dgumbel (你可以使用 curve 函数来实现)。直方图应该很好地近似真实密度。你可以使用默认值 $\mu = 0$ 和 $\beta = 1$。

如何在R中检查函数pgumbel和qgumbel是否正确？它们应该通过哪些合理性检查？还要考虑一些数值检查来测试实现的函数pgumbel是否是dgumbel (的累积分布函数（提示：在R中我们可以进行数值积分)）。评论并解释你进行的所有测试和检查。

Question 3.2 使用上述给出的Gumbel分布的期望值和方差公式，写出参数μ和β (的矩估计方法——你必须写下数学表达式，而不是R代码)。

然后应用矩估计法将Gumbel分布拟合到数据集wind中的观测值，该数据集包含法国圣马丁昂奥特市不同天数的最大风速测量值。绘制wind中数据的直方图以及与矩估计法对应的估计Gumbel密度。使用Q-Q图判断估计效果。

Question 3.3 在R中实现Gumbel模型的负对数似然函数，并数值求解`wind`数据集的参数μ和β的最大似然估计。使用矩估计法参数作为优化算法的初始参数，（如果未解决第3.2题，尝试不同的初始条件，例如$\mu = 0, \beta = 1$）。在直方图上绘制估计的密度，并与矩估计法的结果进行比较。

Question 3.4 也对`wind`中的数据使用最大似然法拟合高斯模型。使用直方图和Q-Q图检查高斯模型对数据的拟合程度。你认为高斯模型合适吗？使用AIC和BIC比较`wind`数据的高斯和Gumbel模型。我们可以使用似然比检验来比较高斯和Gumbel模型吗？如果可以，计算p值；如果不可以，说明无法进行检验的原因。

Question 3.5 使用非参数自助法估计 $\hat{\mu}$ 和 $\hat{\beta}$（Gumbel分布的最大似然估计量）在 `wind` 数据集上的标准误差。计算参数的95%置信区间，使用正态分位数。同时，计算自助样本的百分位数置信区间。

Question 3.6 在最后一个问题中，我们想要对Gumbel分布的参数μ和β进行贝叶斯推断。我们为参数μ选择了一个高斯先验，特别是

$$\mu \sim N(0, 10)$$

对于参数 $\beta > 0$，我们将使用指数先验，

$$\beta \sim Exponential(1)$$

使用 `optim` 计算 MAP 估计器（注意 `optim` 默认执行最小化，要执行最大化，你应该将 `control = list(fnscale = -1)` 设置为 `optim` 的一个参数，查看 `optim` 文档，否则你可以改变目标函数的符号）。同时，使用蒙特卡洛方法获取两个参数 μ 和 β 的后验期望值。