Homework 3

Requirement

- This homework contains two parts:
  - The first part aims to help you review some elementary concepts.
  - The second part are some programming exercises you can do for a better understanding of the last few lectures.
- Submission of the homework is only **optional**. It will not contribute to the final grade.
- If you want to submit and get feedbacks, please compile your answers into **a single PDF file** and send it to the Teaching Assistant (via QQ or email) 1 week before the final exam.

Exercises for revision

1. 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_____。
   (A) 模型的真实参数 θ 是一个常数
   (B) 模型的真实参数 θ 是一个随机变量
   (C) 在观测数据前，置信区间是一个随机区间
   (D) 对于一组给定的数据，置信区间是一个具体的数值区间

2. 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_____。
   (A) 该药物无效的概率为 2%
   (B) 该药物相对于安慰剂有 2%的提升
   (C) p 值<0.05，证明了原假设是错的
   (D) p 值<0.05 不能证明原假设是错的

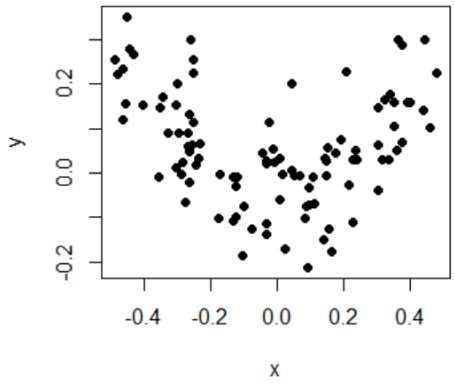
3. 对于某个参数估计问题，有原假设 H_0: $\theta = 0$。设基于样本得到的参数θ的 95%置信区间为$[1, 3]$，则我们可以在显著性水平_____下_____(接受或拒绝)原假设。

4. 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_____统计量。

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

6. 小明得到一个三阶列联表 `data_tab`，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 $X \perp Y \mid Z$)。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）。

7. 小明对一组二元数据 $\{(x_i, y_i)\}$ 作简单线性回归：`y = a + b*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。



8. 假设随机变量 $X \in \{1,2,3\}$ 服从的分布为

X	1	2	3
P	θ	2θ	$1-3\theta$

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θ的最大似然估计。

Suggested programming exercises

1. Run the sample codes of the 2nd part of the course (available in the QQ group folder) to familiarize yourself with generalized linear models covered in the lectures, model diagnostic and selection and regularization methods using `R`.

2. Poisson regression on elephants mating data:
   Exercise 11.1 (Page. 495) in "Discrete Data Analysis with R" (available in the QQ group folder).
   - In part (e), you can set `family=quasipoisson` in `glm()` in order to fit to a quasi-Poisson model. Alternatively, you can also use the `dispersiontest()` in the `AER` package.

3. Logistic regression with regularization on fake news data:
   Consider the `fake_news` data in Lab1; the task is to predict whether a news article is a fake or real.
   (1) Split the data set into a training set and a test set.
   (2) Use all the available features (excluding `title`, `text`, `url` and `authors`), fit a logistic regression model on the training set, and report the test error (or accuracy) obtained.
   (3) Fit a logistic regression model with ridge regularization on the training set, with λ chosen by cross-validation. Report the test error obtained.
   Use `glmnet()`for this purpose; set `family="binomial"` for logistic regression, and set `alpha=0` for ridge regression.
   (4) Fit a logistic regression model with LASSO regularization on the training set, with λ chosen by cross-validation. Report the test error obtained, and the number of non-zero coefficient estimates.
   Use `glmnet()`for this purpose; set `family="binomial"` for logistic regression, and set `alpha=1` for LASSO regression.
   (5) Comment on the results obtained.