

# 答疑

By qh

## 如何 debug

待续

## 第一次作业

1. 能否去掉极值，比如研究目的是研究从武汉离开的人对其他地方的影响，要不要把武汉本地的数据剔除呢？能不能把湖北省全部去除？

言之有理逻辑通顺即可

2. 如何删去行，比如删去湖北

```
data = data[data$X1 != "湖北", ]  
data = data[-c(1,3,5),]
```

3. 如何删去列，比如删去所有离散变量

```
data = data[,-c(1,3,5)]  
# 如果列名叫 X1, X3, X5 还可以  
data = data[,-c("X1", "X3", "X5")]
```

4. 如何提取列，比如提取所有连续变量

```
data = data[,c(1,3,5)]  
data = data[,c("X1", "X3", "X5")]
```

5. 如何对所有数据操作，比如取对数，而不是一列一列做

使用 `sapply` 函数或者先把连续变量提取出来，存成一个数据框 `data`，再

```
data=log(data+1)
```

6. 为什么取log；为什么取log要先加1

因为画出数据直方图，发现数据偏斜的特别严重，这个时候我们一般对数据取个log，取个log后我们希望数据能长得更像正态分布一点

因为如果数据中存在0，那么取log就是负无穷，+1后取对数，0被变为0.这是一个数据变换小技巧

#### 7. AIC和BIC结果不一样，选哪个

业务人员肉眼决定，凭借自己的判断，比如假设BIC删去了你认为比较重要的变量，那么你就选AIC；还有就是稍微客观一点的做法，把数据分成80/20，然后用内样本外样本，内样本拟合模型，外样本算MSE，这样的数据划分和分析过程随机100次，最终挑小MSE对应的方法