答疑

By qh

如何 debug

1. 看报错信息，您的错误R已经帮您用红色报错了，比如这里说数据中存在NA（空值）/Inf（无穷），那么这个时候最好把数据 `print` 出来看看，如果是数据框就在右上角 environment 点开看看

```
> fit = lm(logy~logx6+logx5+logx4+logx3+logx2+logx1, data = wh)
Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
  NA/NaN/Inf in 'x'
>
```

2. 如果报错信息看不懂，把报错复制到 bing 或者百度里搜索，在csdn的博客上基本上都有解答

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. 取完log之后数据仍然偏斜的很厉害，怎么办?

> 一般这样我们取log变换是因为我们观察到数据偏斜的很厉害，我们希望取完log之后数据能长得更像正态分布一些，但是取完log之后发现数据依然偏斜的很厉害，这个（可能）是因为这个数据里0含量太多了，这种情况我们一般可以把数据离散化，比如分成几个区间，但是因为咱们作业1涉及的内容只是线性回归模型，这个作业更多的是希望大家能从线性回归上手，咱们暂时不考虑更多的数据变换技巧，取完log后再继续您的分析就好

8. 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