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

9. 读入数据出错 `read.csv("第一章.csv")`

> 因为这个 csv 文件不是以 utf-8 格式（可以理解为标准的格式）存储的，这个时候它的存储方式是 gbk 格式，那么就有两种解决办法
>
> 1. 使用 excel 打开这个 csv 文件，然后转存为 utf-8 格式的 csv 文件即可
> 2. 使用以下代码读入
>
> ```
> data = read.csv("第一章.csv", fileEncoding = "gbk")
> ```

10. MAC 电脑画图不显示中文

> 一般来说添加一行代码即可
>
> ```
> par(family="STKaiti")
> ```
>
> 如果以上代码报错，显示电脑里没有 STKaiti 这个字体，那么试试下面这行代码使用黑体（一般来说问题都可以解决）
>
> ```
> par(family="Hei")
> ```

11. 累计武汉滞留时间的单位是什么?

> 秒

第五次作业 - 定序回归

1. 定序回归（多分类）的评价指标? 就比如我加入5个X自变量 和2个X自变量 我怎么比较这两个模型哪个好呀?

> 可以看分类准确率，假设真实标签是Y，是一个向量，取值在1-4之间，你预测出来的标签是Y1，计算准确率可以是
>
> ```
> mean(Y == Y1)
> ```

2. 离散的水平可能有很多取值，如何合并一些离散的取值为一组?

> 一个例子
>
> ```
> city = c("北京", "上海", "浙江", "华盛顿", "纽约", "伦敦")
> a = data.frame(city)
> # 分两组，国内 & 国外
> group1 = c("北京", "上海", "浙江")
> group2 = c("华盛顿", "纽约", "伦敦")
> a[a$city %in% group1, "city"] = "国内" # 将city这一列的属于 group1的全部修改为国内
> a[a$city %in% group2, "city"] = "国外" # 将city这一列的属于group2的全部修改为国外
> a$city # 查看数据
> ```

3. 条形图x轴每个离散水平的名字太长无法正常显示

> par(las=2) barplot(c(1, 2, 3), names.arg=c("浙江浙江浙江", "北京北京北京", "上海上海上海上海"),)

报告有关问题

1. Introduction 部分推荐写三段论，是哪三段论?

> 第一段：【背景介绍】，提出观点，举例论证，适当介绍行业/业务背景，引出其中的业务痛点。例如提出“随着通信技术的不断发展和普及，通信行业的竞争越来越白热化。”这个观点，再使用举例的方式论证这个观点。 第二段：【业务痛点】，是第一段背景介绍的深入挖掘，其中存在怎样的业务痛点亟待解决。 第三段：【X+Y】，为了解决业务痛点，我们拟采用什么数据作为X和Y来研究，为什么可行，结论将具有什么意义等。