**数据处理：重新处理了Journal Model并定义病状，总体病情情况**

****

> table(unlist(jm.sum[9,]))/ncol(jm.sum)

0 1 2

0.69992487 0.21807097 0.08200416

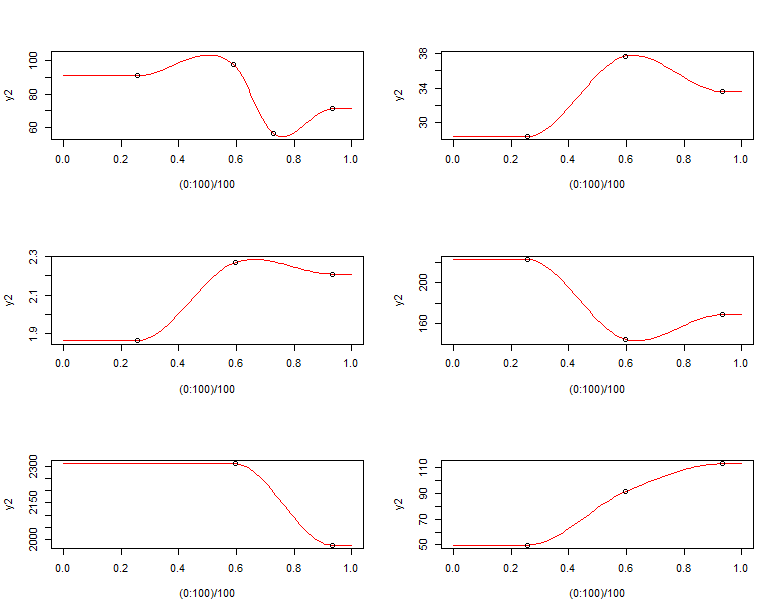
> table(unlist(jm.sum[9,]))

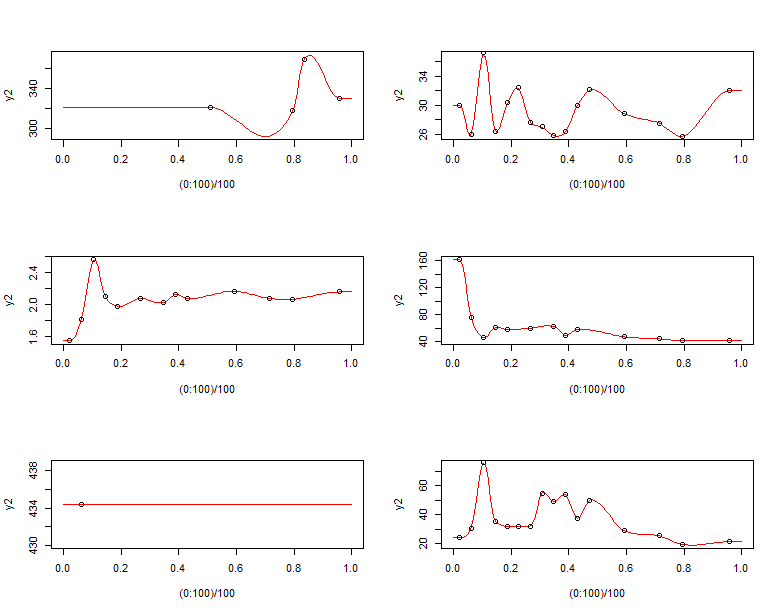
0 1 2

24223 7547 2838

**数据补全**

1. 缺失时间补齐：通过函数型拟合对全时间段至少有一个观测点的数据进行拟合和补齐





2. 缺失变量补齐：通过弱线性模型组合对NA数据进行初始化补齐，模型拟合结果良好，一致性检验均通过（只用了300个样本）

！科学性上需要再做进一步10folder来保证模型的稳定性，暂时还没有做

cor corp mape

[1,] 0.9115265 0 0.37752557

[2,] 0.8028567 0 0.10974715

[3,] 0.8118249 0 0.06962809

[4,] 0.8728163 0 0.33300808

[5,] 0.7938861 0 0.50801497

[6,] 0.8613343 0 0.70745853

[7,] 0.9498926 0 1.51620148

[8,] 0.8928407 0 0.42611499

[9,] 0.9331329 0 4.18677273

[10,] 0.9071937 0 1.36404370

[11,] 0.9552608 0 1.41460740

[12,] 0.8359110 0 8.53040963

3. 缺失变量优化：通过低秩模型对模型进行再拟合，拟合结果有提升

cor corp mape

[1,] 0.9399478 0 0.28444964

[2,] 0.8498318 0 0.08406273

[3,] 0.8575198 0 0.05204218

[4,] 0.9269819 0 0.26120921

[5,] 0.8177950 0 0.39598385

[6,] 0.9135899 0 0.52146895

[7,] 0.9516340 0 1.16261518

[8,] 0.9274556 0 0.31629947

[9,] 0.9405836 0 2.96374181

[10,] 0.8965854 0 0.96514118

[11,] 0.9349489 0 1.07381711

[12,] 0.8645531 0 6.97688711

**基本标签预测**

样本情况

0 1 2

1671 346 125

> table(mdata$crate)/sum(table(mdata$crate))

0 1 2

0.78011204 0.16153128 0.05835668

> table(mdata$rate)

0 1 2

1111 609 422

> table(mdata$rate)/sum(table(mdata$rate))

0 1 2

0.5186741 0.2843137 0.1970121

线性分类器结果：

> table(crate=fdata$crate,predict=p.lda)

predict

crate 0 1 2

0 27814 0 0

1 11 2030 0

2 0 0 445

逻辑回归结果

predict

crate 0 1 2

0 27814 0 0

1 5 2034 2

2 0 114 331

#现状

数据补齐使用小样本结果不错，使用全量样本后结果理论会更理想。

我们之后会在补齐过程中使用10 folder的做法来提升模型的稳定性和科学性，但是这一点一定会造成统计指标的下降。

基于小样本的预测模型与直接coding结果几乎是一致的，不是非常正常，仍然在测试。如果真是这样说明基本病症就是根据金标准来判断的，数学上无法发现其他变量与之造成的关联性。

#下一阶段

全量样本处理

模型的稳定性处理

讨论需要什么结果