Data Analysis Tools and Practice(Using R)

2021.06.08

R数据科学-Model

Huiping Sun(孙惠平)
sunhp@ss.pku.edu.cn

课堂测试时间

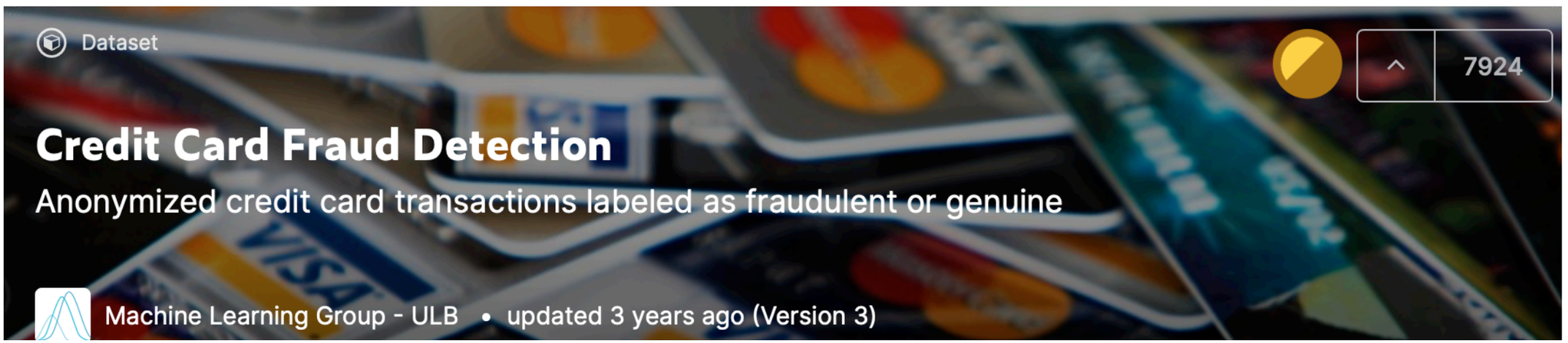
课堂测试13

先用电脑完成
40分钟 然后誊抄纸上

- 信用卡欺诈数据库，见**creditcard.csv**

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
6	0.2514121	-0.0183068	0.27783758	-0.1104739	0.06692807	0.12853936	-0.1891148	0.13355838	-0.0210531	149.62	0
1	-0.0690831	-0.2257752	-0.638672	0.10128802	-0.3398465	0.1671704	0.12589453	-0.0089831	0.01472417	2.69	0
5	0.52497973	0.24799815	0.7716794	0.90941226	-0.689281	-0.3276418	-0.1390966	-0.0553528	-0.0597518	378.66	0
7	-0.2080378	-0.1083005	0.0052736	-0.1903205	-1.1755753	0.64737603	-0.2219288	0.06272285	0.06145763	123.5	0
5	0.40854236	-0.0094307	0.79827849	-0.1374581	0.14126698	-0.2060096	0.50229222	0.21942223	0.21515315	69.99	0
8	0.08496767	-0.2082535	-0.5598248	-0.0263977	-0.3714266	-0.2327938	0.10591478	0.25384422	0.08108026	3.67	0
5	-0.2196326	-0.1677163	-0.2707097	-0.1541038	-0.7800554	0.75013694	-0.2572368	0.03450743	0.00516777	4.99	0
1	-0.1567419	1.94346534	-1.0154547	0.05750353	-0.649709	-0.4152666	-0.0516343	-1.2069211	-1.0853392	40.8	0
7	0.05273567	-0.0734251	-0.2680916	-0.2042327	1.0115918	0.37320468	-0.3841573	0.01174736	0.14240433	93.2	0

- 建模分析信用卡欺诈
- 使用其中部分部分数据即可
- 分析模型有效性
- 逻辑回归



https://www.kaggle.com/mlg-ulb/creditcardfraud

参考书

O'REILLY®

TURING 图灵程序设计丛书

全彩印刷

CH17

R数据科学

R for Data Science

摒弃其他R语言工具书从头到尾讲统计的陋习
从实用的R包出发，带你重新认识R和数据科学

[新西兰] 哈德利·威克姆 [美] 加勒特·格罗勒芒德 著
陈光欣 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

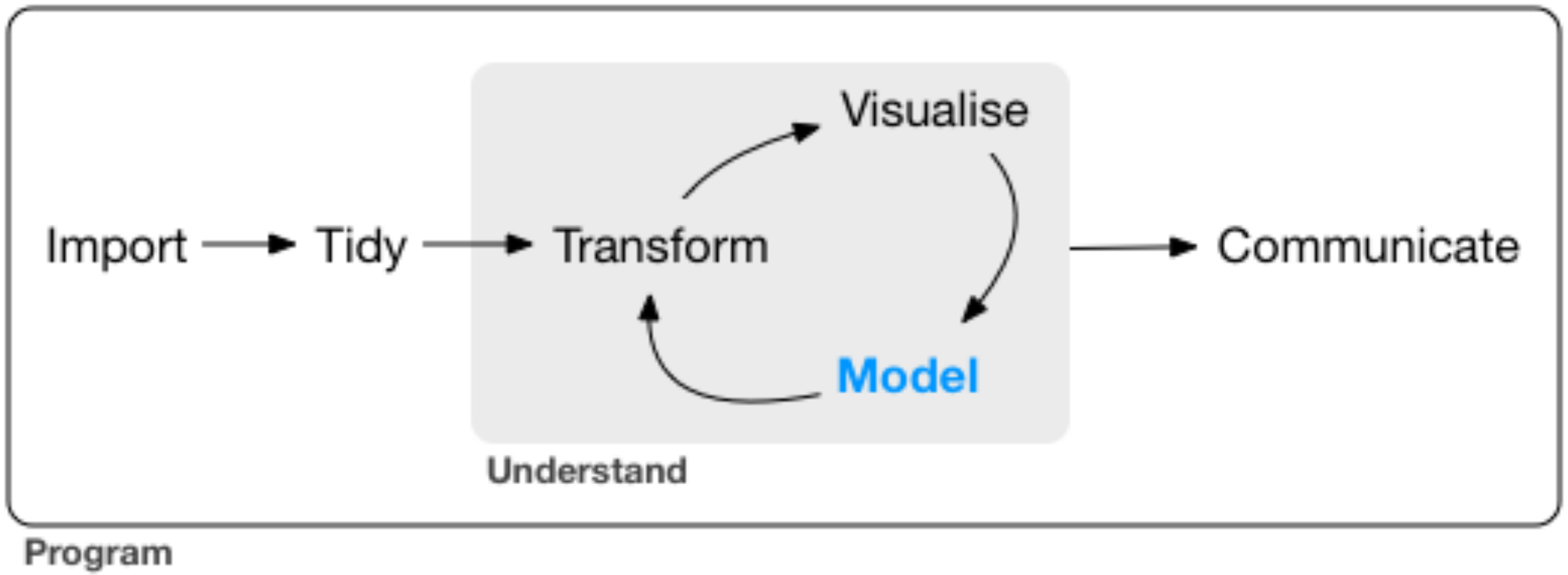
建模

CH17
CH18
CH19

R Data Science

建模

导入 | 整理 | 转换 | 可视化 | 建模 | 沟通



信号 | 噪声 | 发现 | 预测

简单 | 低维

数据探索

建模的重点在于推断和验证假设是否为真

每个观测都可以用于数据探索，也可以用于假设检验

数据探索时一个观测可以使用任意多次

假设验证时一个观测只能使用一次

探索----假设----验证

训练集合 60%

查询集合 20%

测试集合 20%

训练集合 70%

测试集合 30%

简单模型

CH17

模型

```
y = a_1 * x + a_2
y = a_1 * x ^ a_2
```

```
y = 3 * x + 7
y = 9 * x ^ 2
```

模式　残差

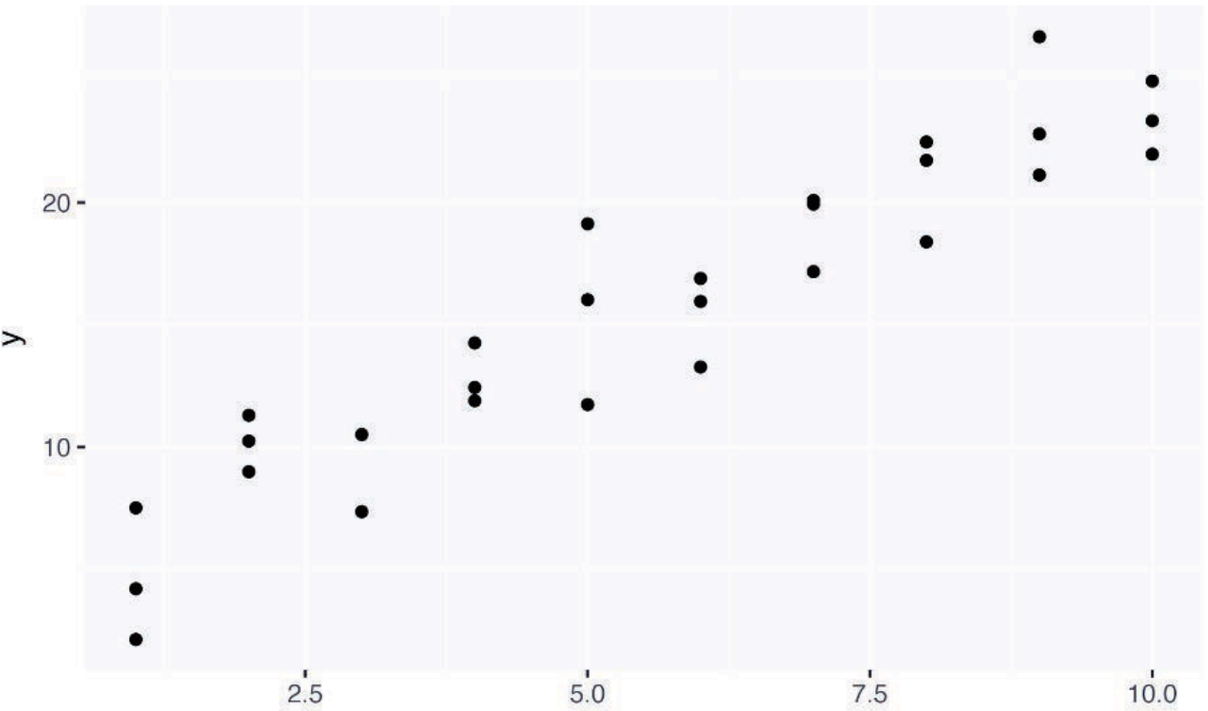
模型族　拟合

所有模型都是错误的，但是有一些是有用的

如果一个简单模型可以精确表示真实世界中的某个系统，那将非常了不起。然而，巧妙地选择简约模型经常可以提供非常好的近似表示。例如，定律 $pV = RT$ 通过一个常数 R 将“理想”气体的压强 p、体积 V 和温度 T 关联起来。虽然这对于任何真实气体来说都是不精确的，但在很多情况下都是一种良好的近似。此外，这个方程的结构包含非常丰富的信息，因为它来自于对气体分子行为的实证研究。

对于这样的模型，我们不需要提出“这个模型是真的吗？”这类问题。如果“真”的含义是“绝对真”，那么答案肯定是“不”。我们唯一感兴趣的问题是：“模型是否具有启发性，是否有用？”

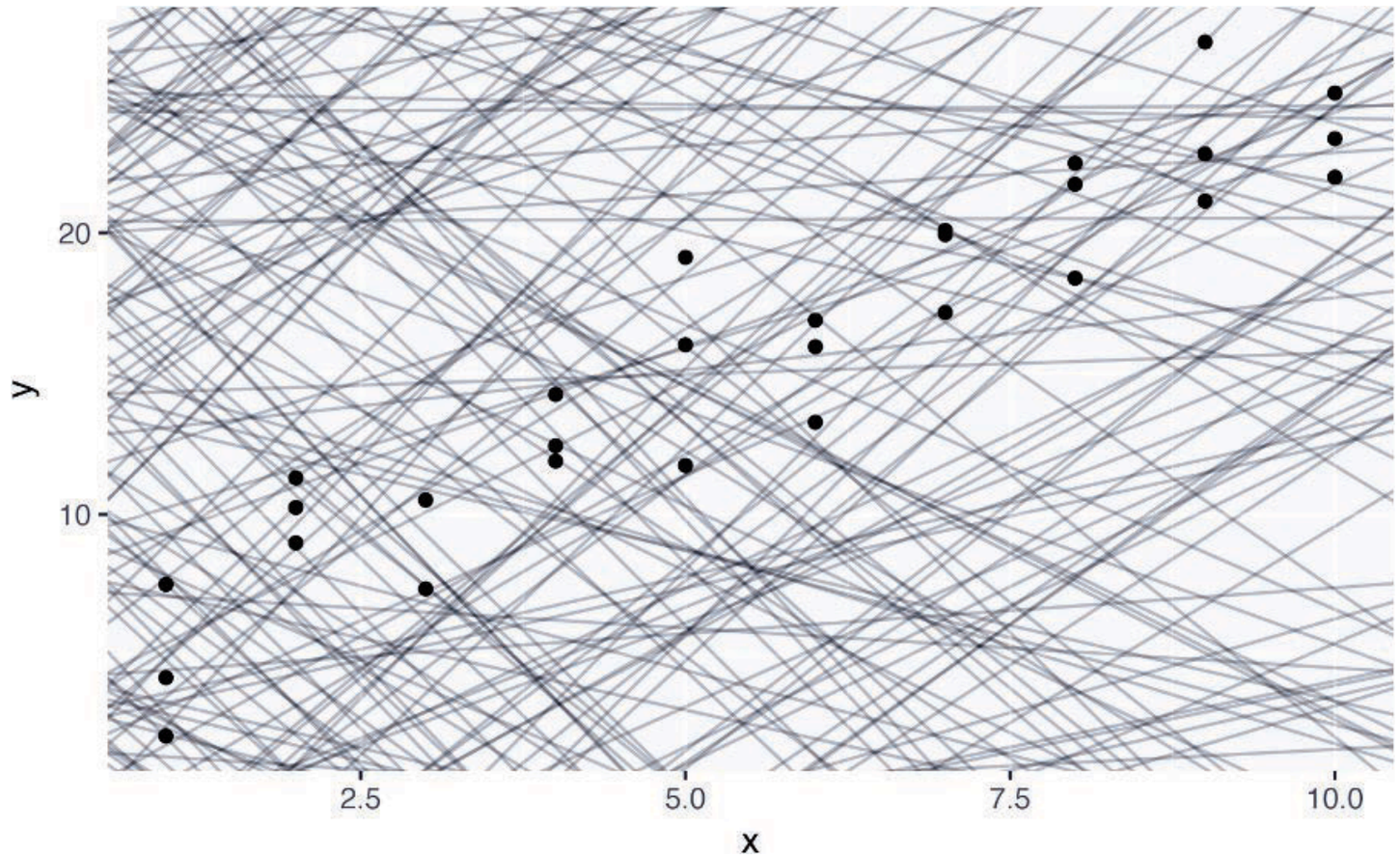
简单模型



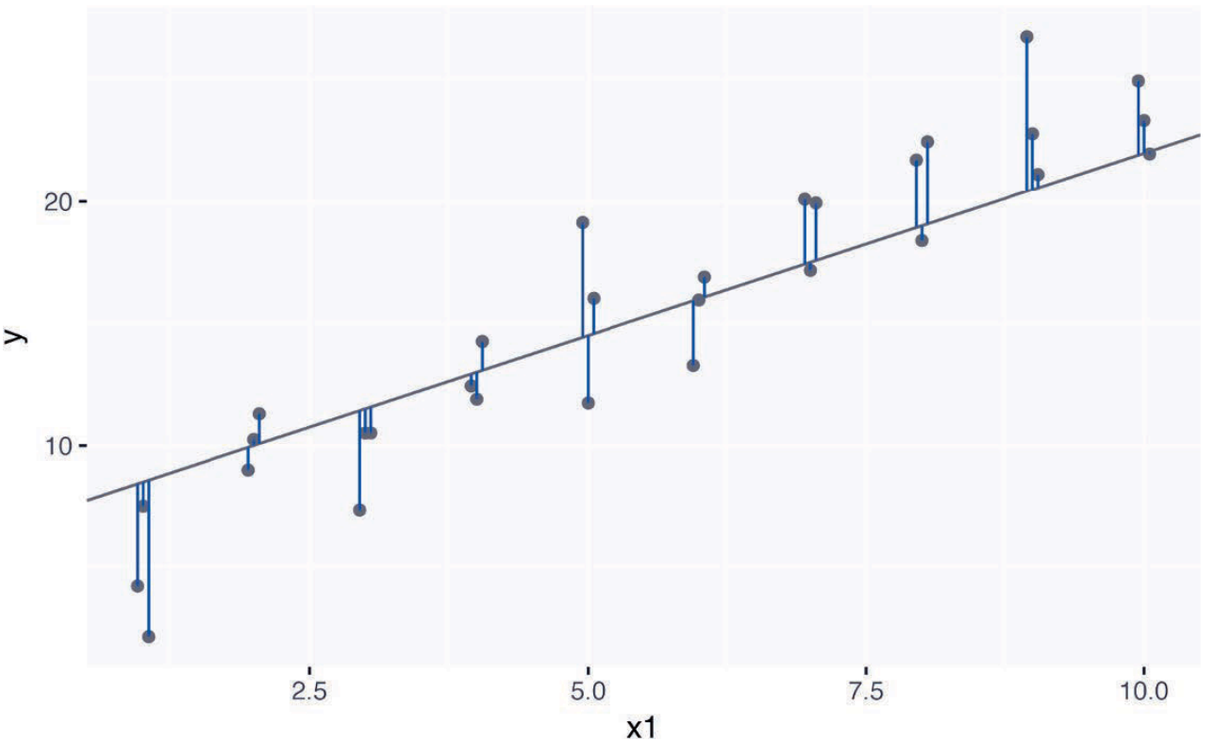

```
library(tidyverse)
library(modelr)
options(na.action = na.warn)

ggplot(sim1, aes(x, y)) +geom_point()
```

```
models <- tibble(
  a1 = runif(250, -20, 40),
  a2 = runif(250, -5, 5)
)
ggplot(sim1, aes(x, y)) +
  geom_abline(
    aes(intercept = a1, slope = a2),
    data = models, alpha = 1/4
  ) +
  geom_point()
```



模型评价

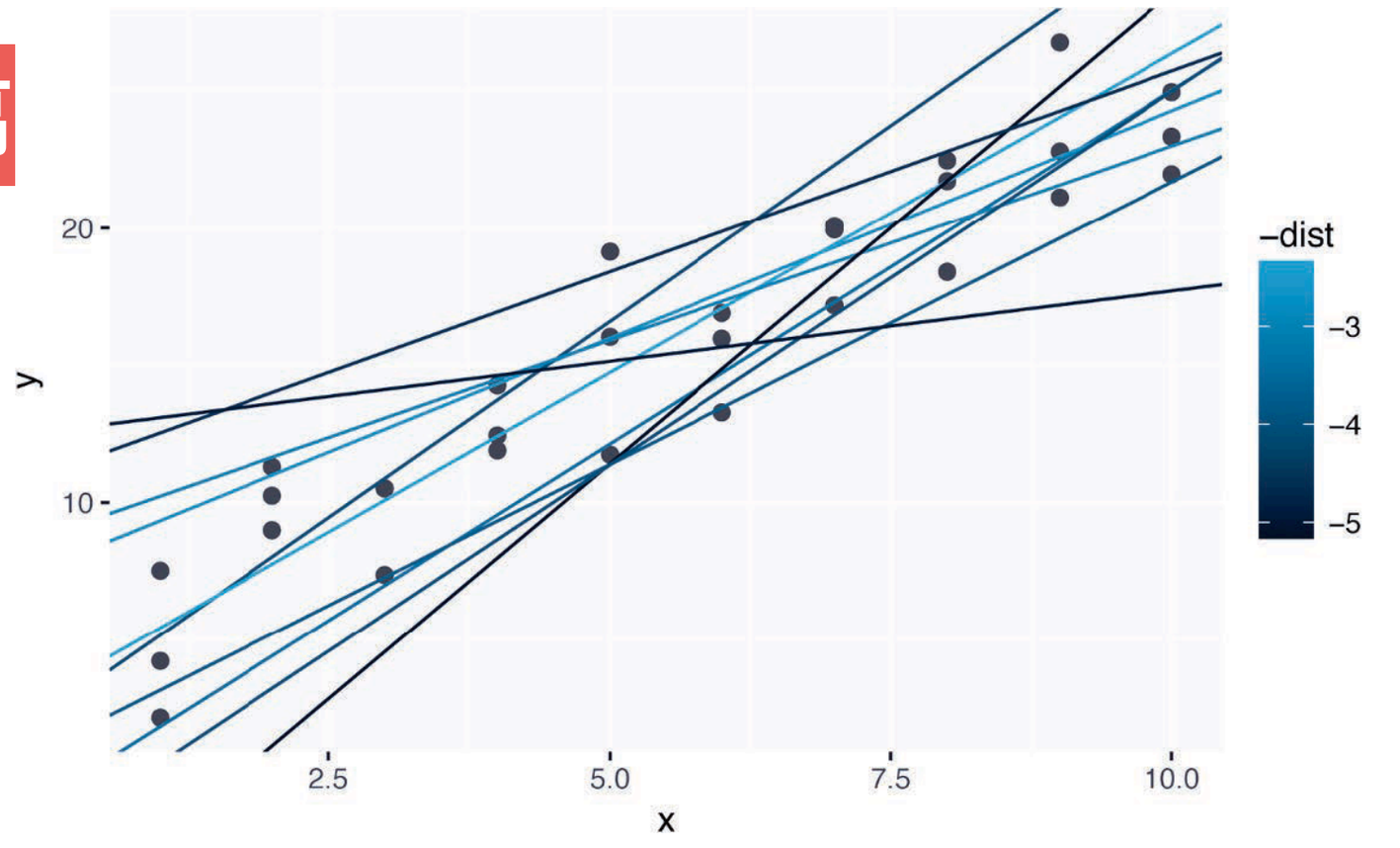


每个数据点和模型的垂直距离

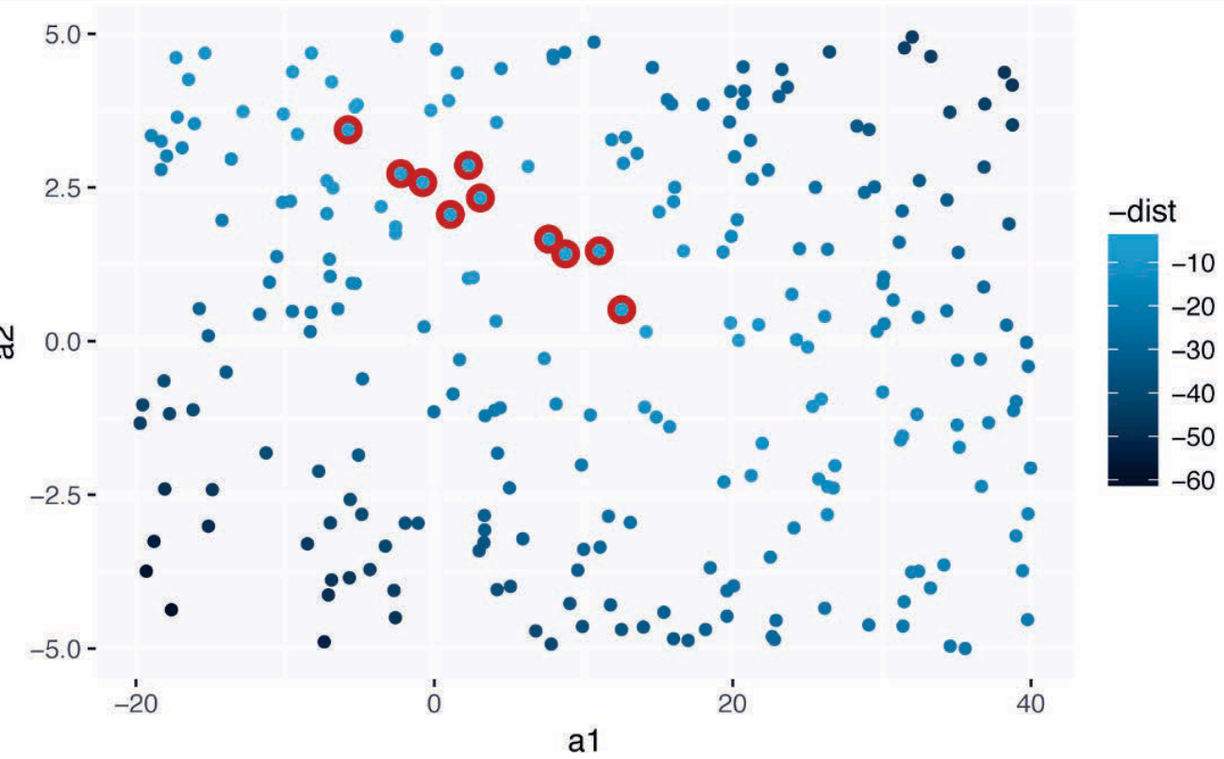
均方根误差
实际值和预测值差
平方后求平均

好的模型应该和实际数据很接近

```
ggplot(sim1, aes(x, y)) +
  geom_point(size = 2, color = "grey30") +
  geom_abline(
    aes(intercept = a1, slope = a2, color = -dist),
    data = filter(models, rank(dist) <= 10)
  )
```



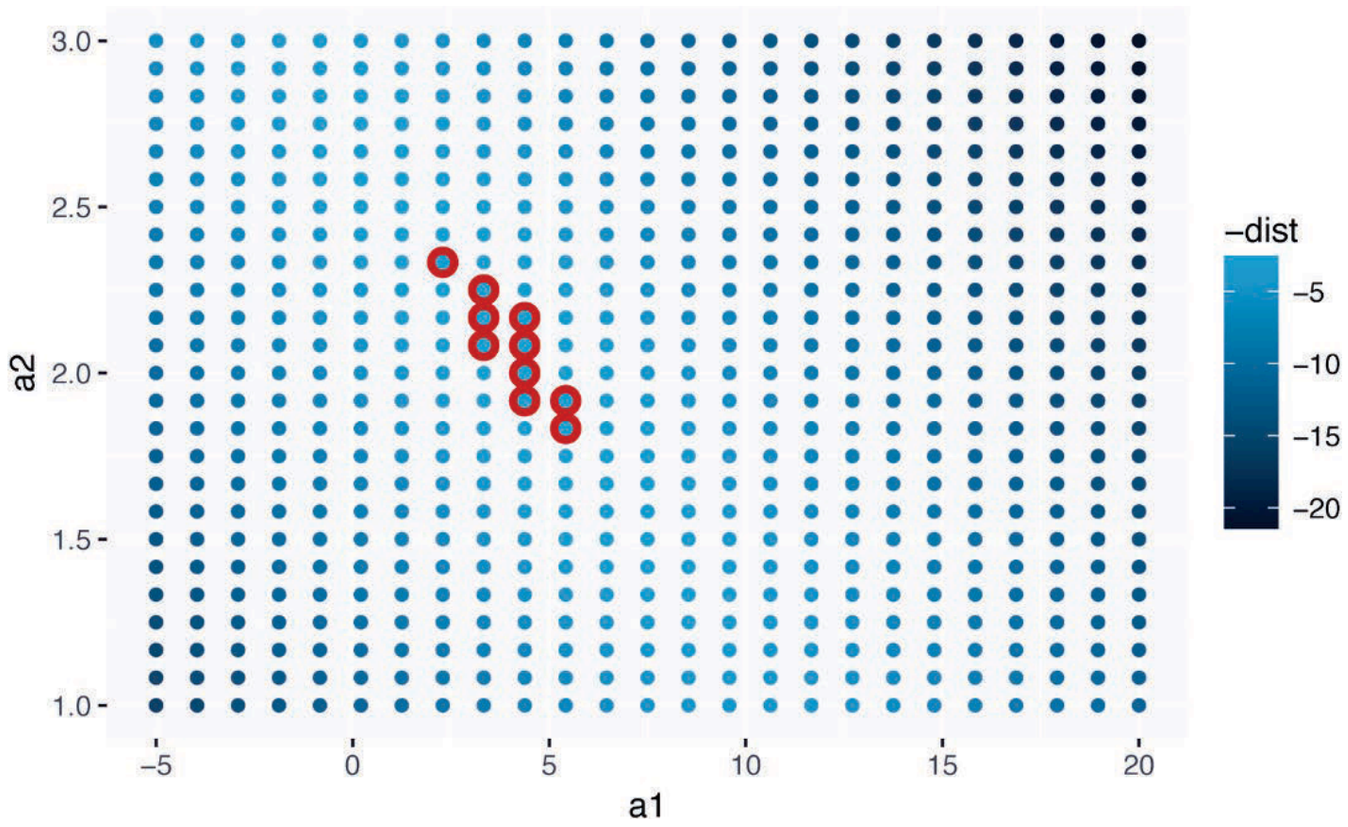
模型评价



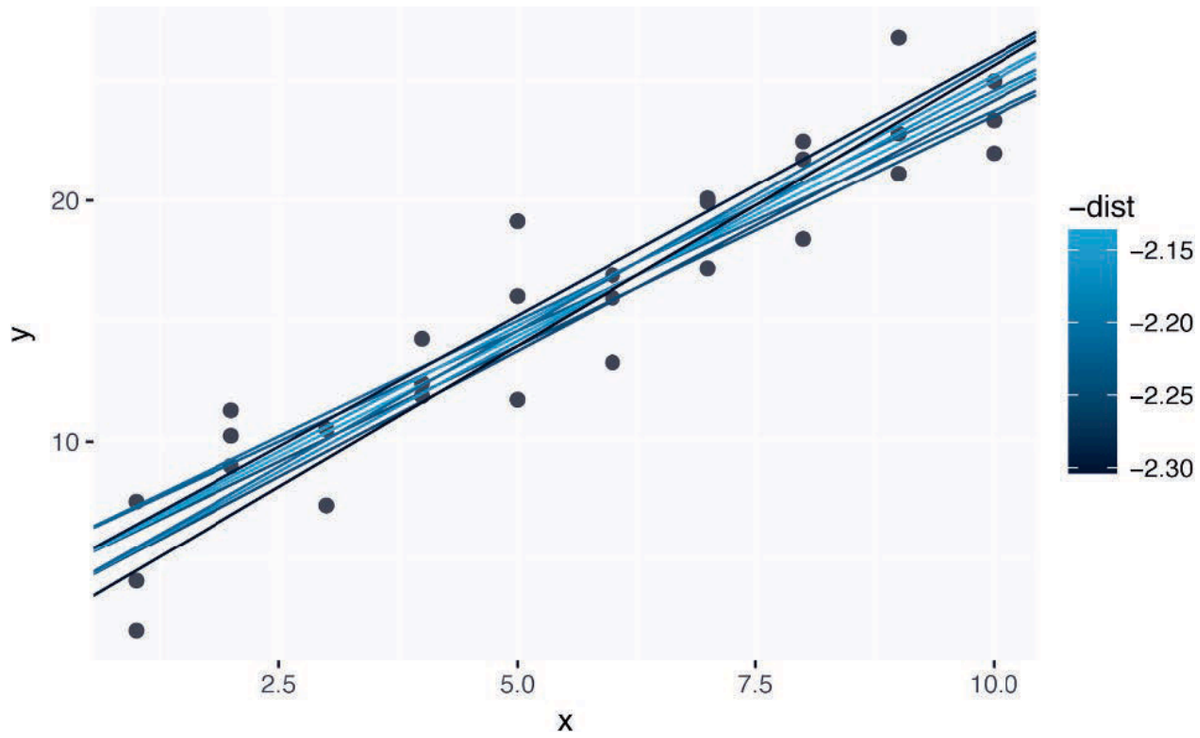

```
ggplot(models, aes(a1, a2)) +
  geom_point(
    data = filter(models, rank(dist) <= 10),
    size = 4, color = "red"
  ) +
  geom_point(aes(colour = -dist))
```

```
grid <- expand.grid(
  a1 = seq(-5, 20, length = 25),
  a2 = seq(1, 3, length = 25)
) %>%
mutate(dist = purrr::map2_dbl(a1, a2, sim1_dist))
```

```
grid %>%
  ggplot(aes(a1, a2)) +
  geom_point(
    data = filter(grid, rank(dist) <= 10),
    size = 4, colour = "red"
  ) +
  geom_point(aes(color = -dist))
```



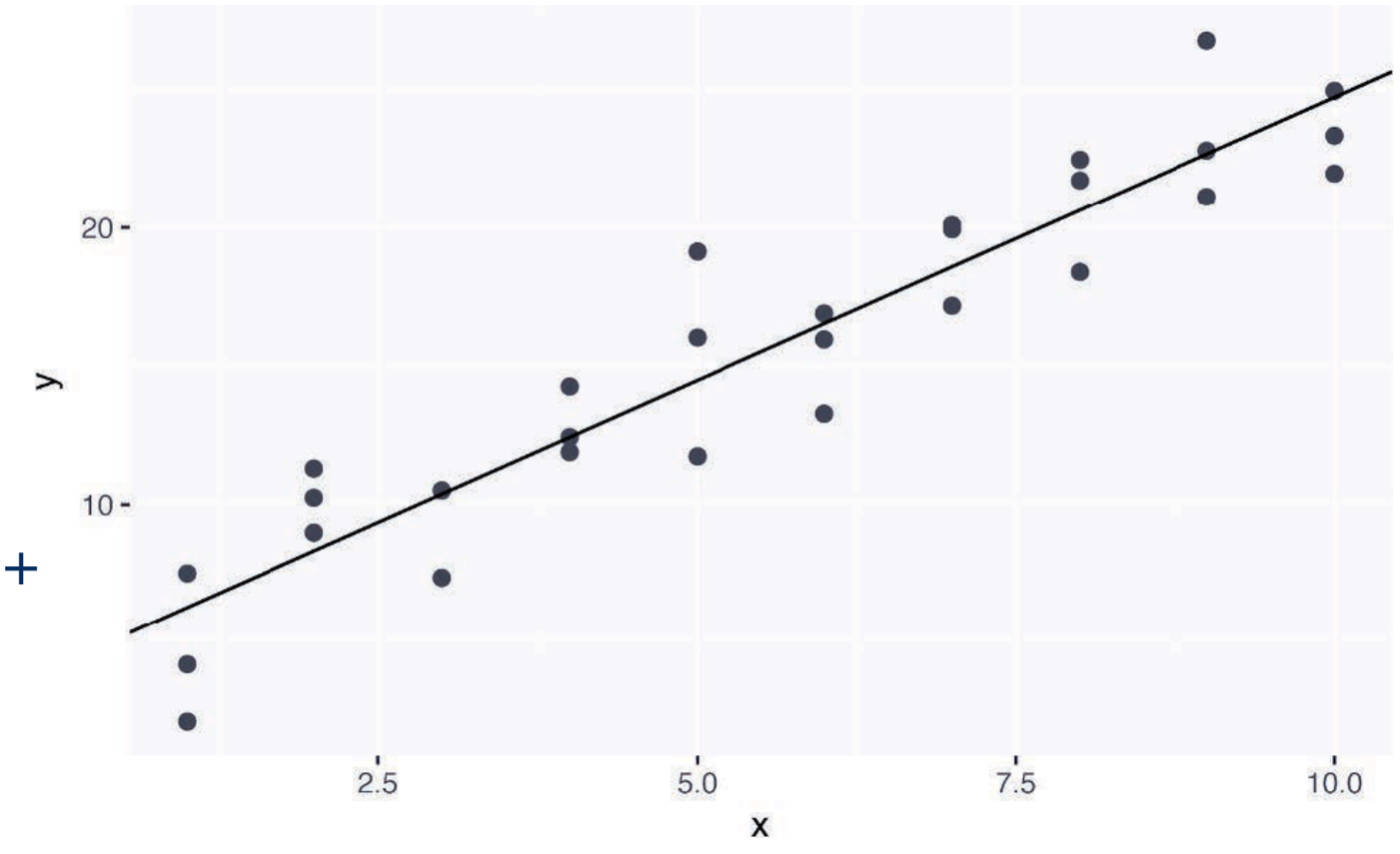
模型评价



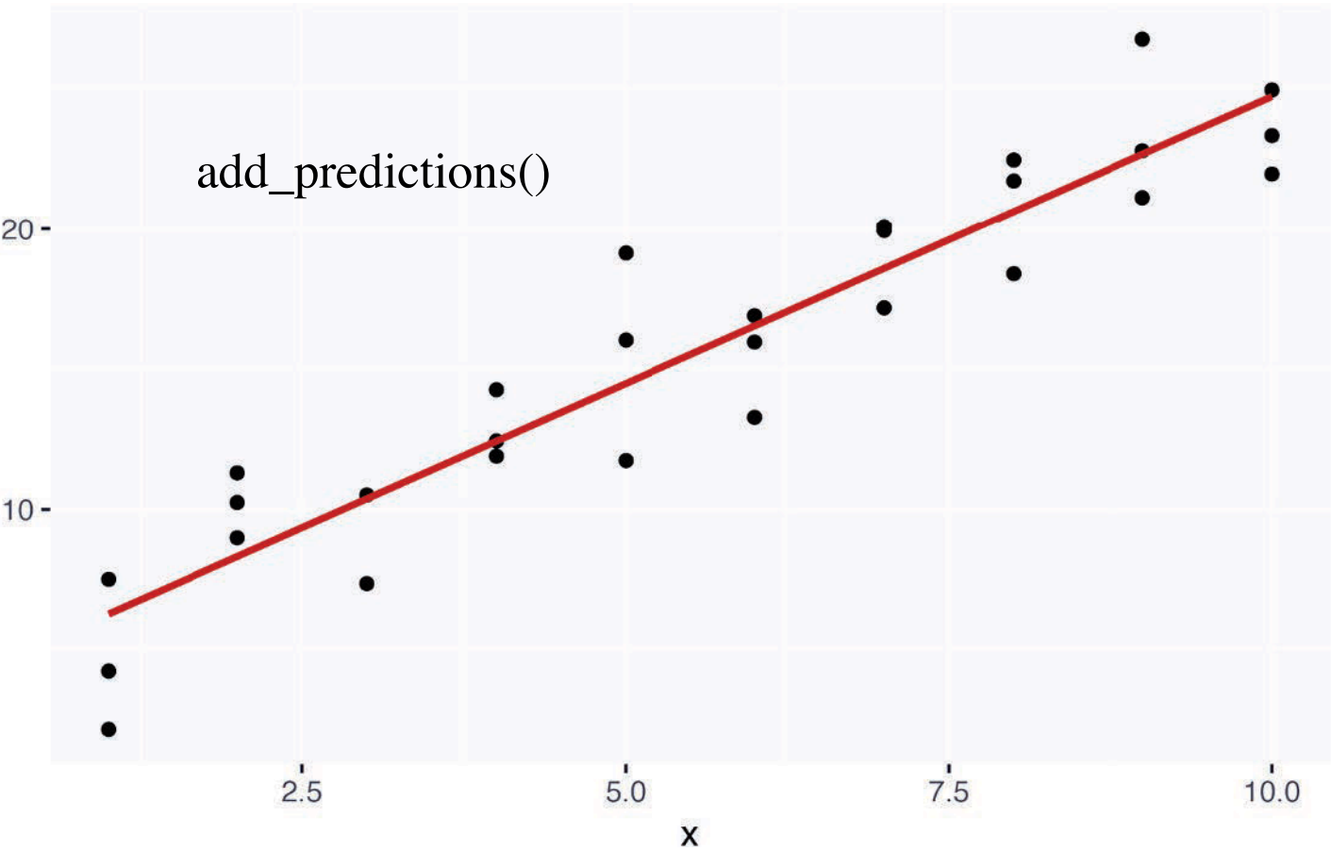

```
ggplot(sim1, aes(x, y)) +
  geom_point(size = 2, color = "grey30") +
  geom_abline(
    aes(intercept = a1, slope = a2, color = -dist),
    data = filter(grid, rank(dist) <= 10)
  )
```

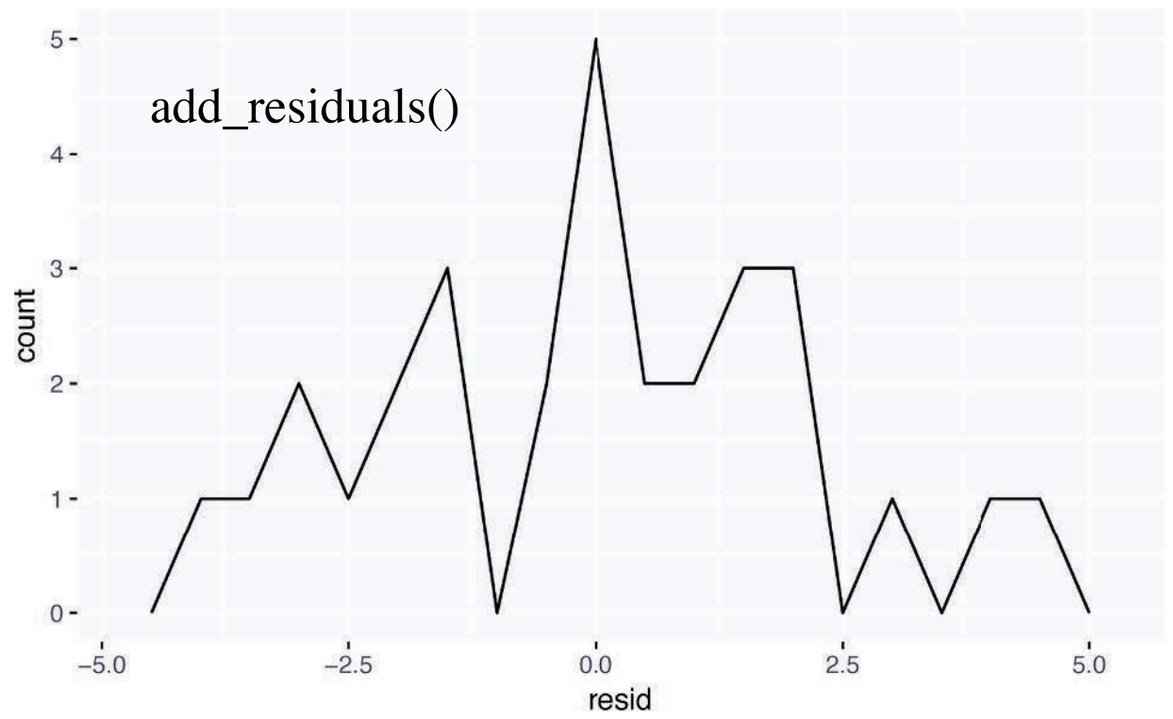
牛顿—拉夫逊搜索

```
best <- optim(c(0, 0), measure_distance,
data = sim1)
best$par
#> [1] 4.22 2.05
ggplot(sim1, aes(x, y)) +
  geom_point(size = 2, color = "grey30") +
  geom_abline(intercept = best$par[1],
  slope = best$par[2])
```

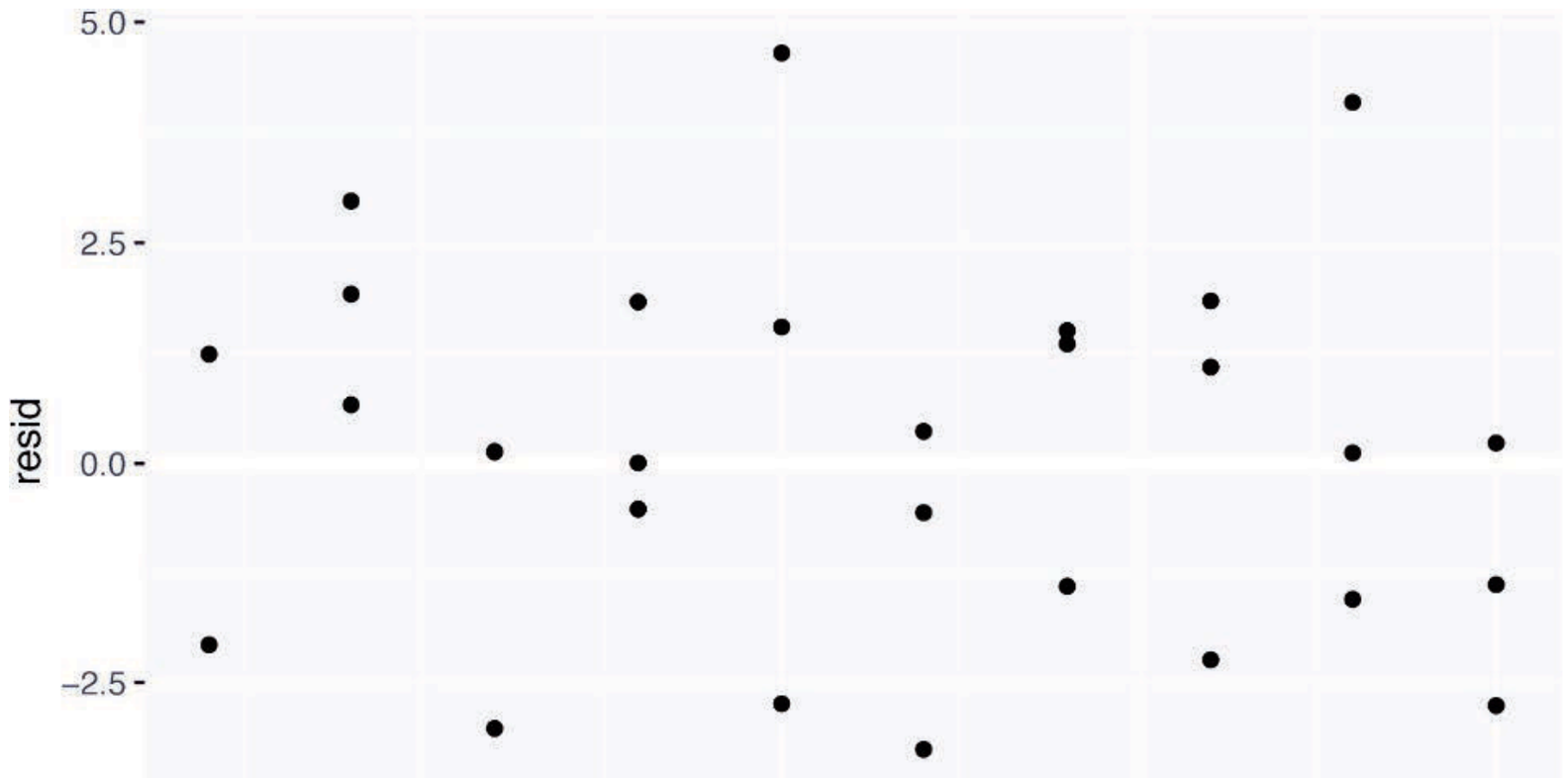


可视化 + 残差

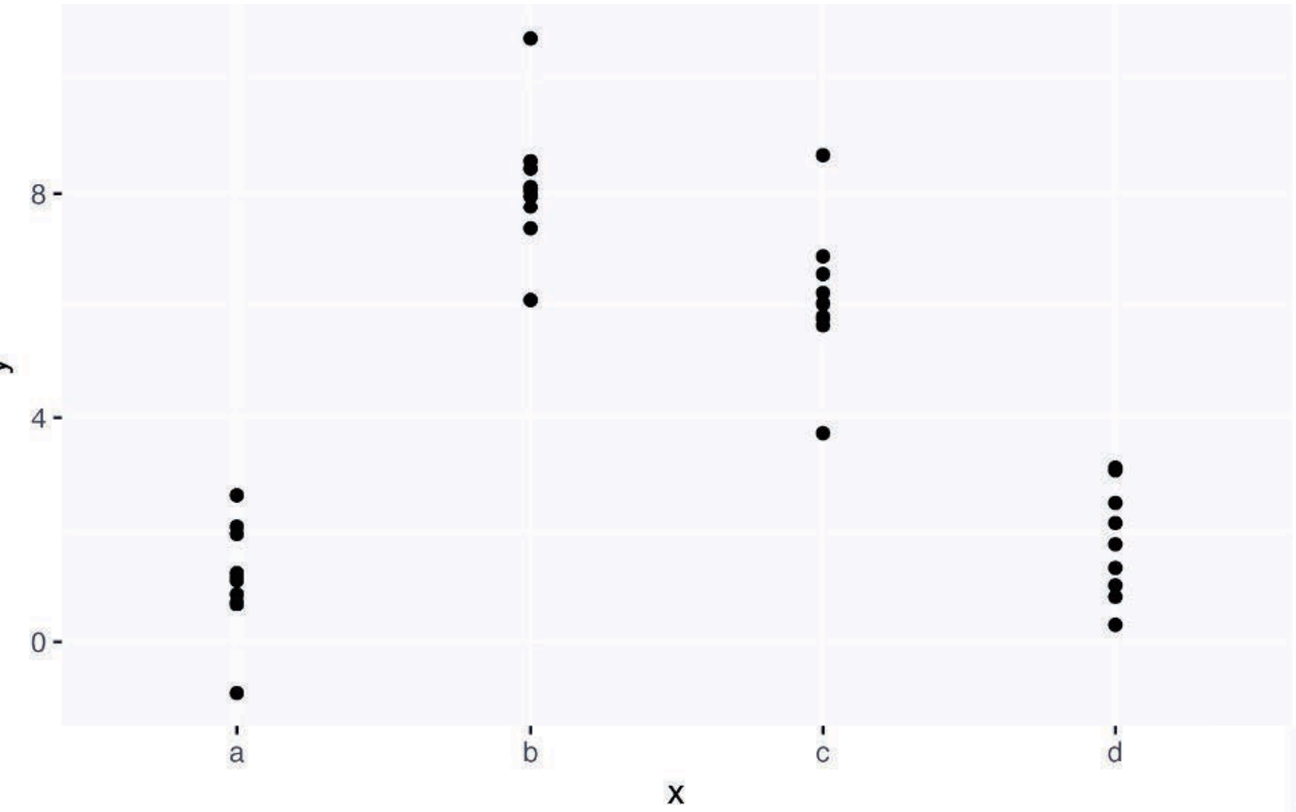


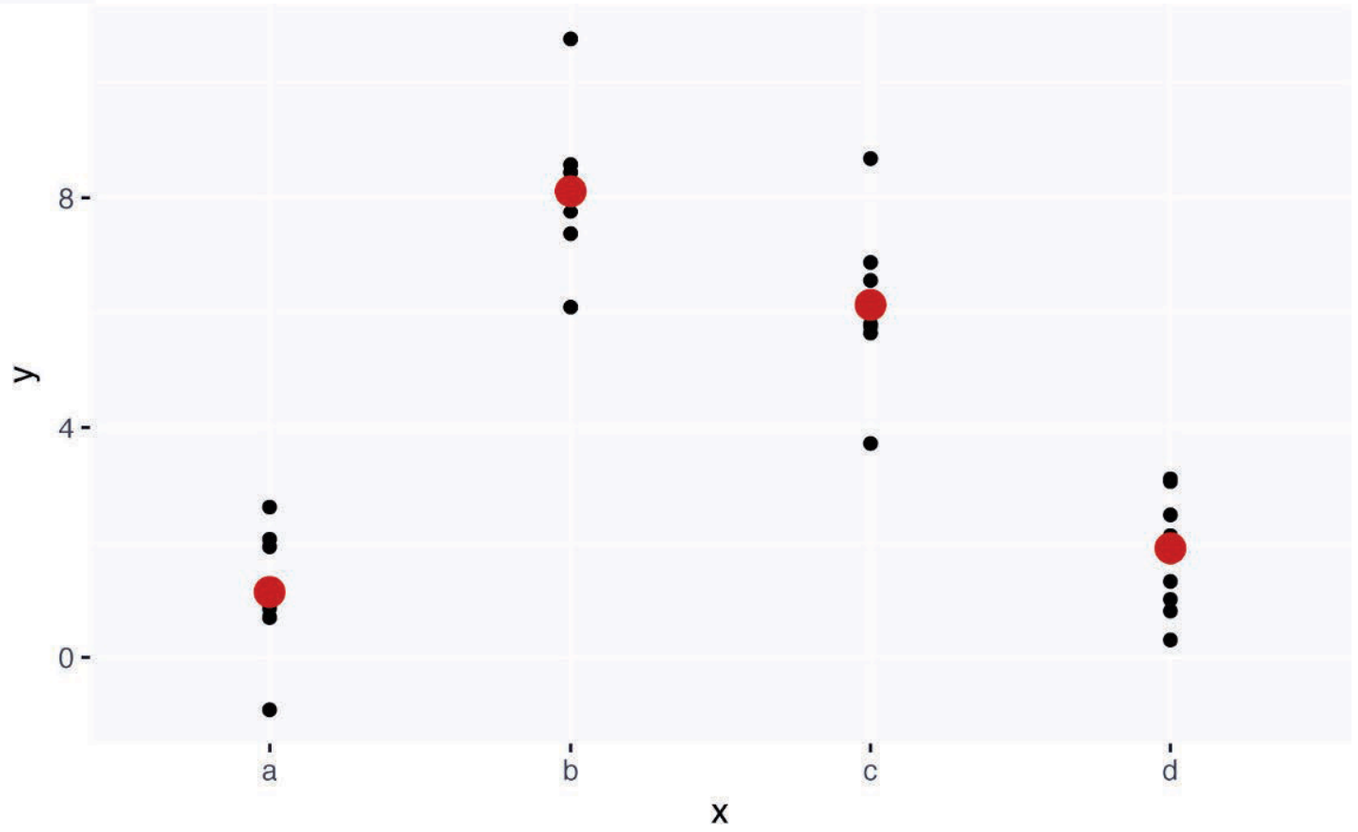



```
ggplot(sim1, aes(x)) +
  geom_point(aes(y = y)) +
  geom_line(
    aes(y = pred),
    data = grid,
    colour = "red",
    size = 1
  )
```

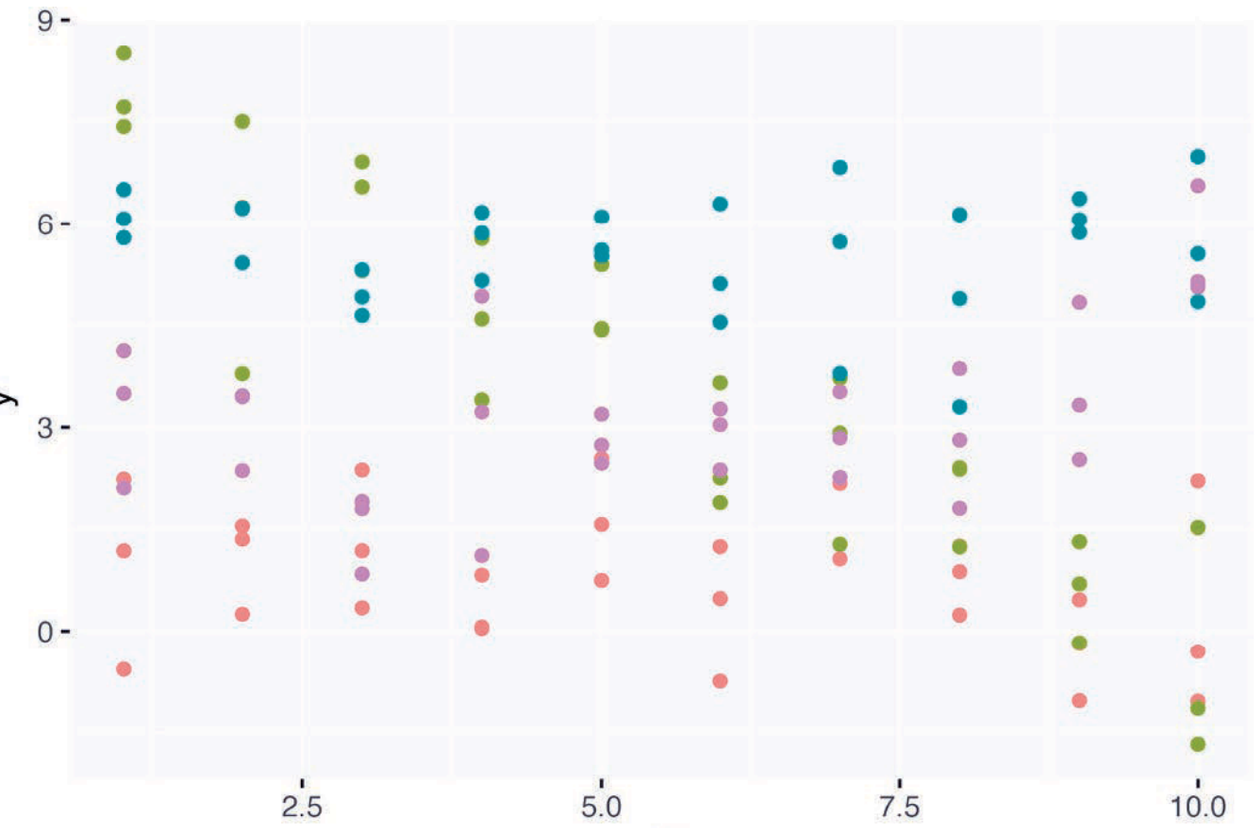


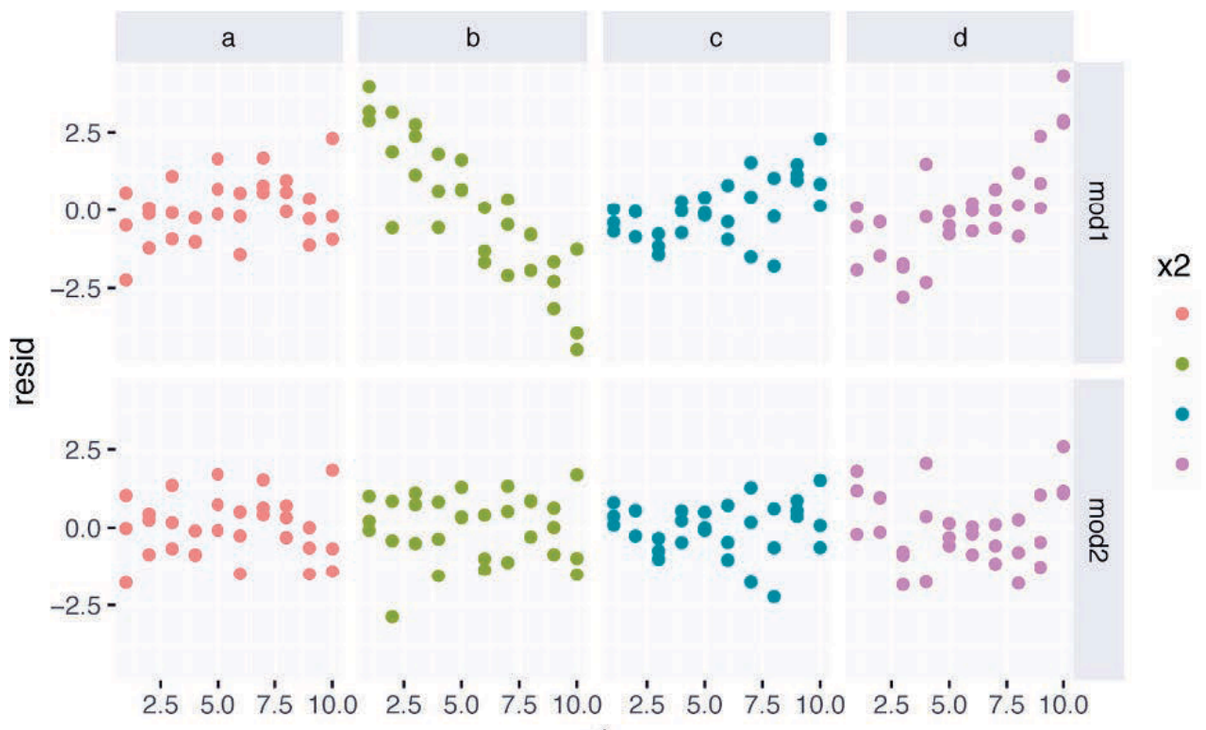
分类变量

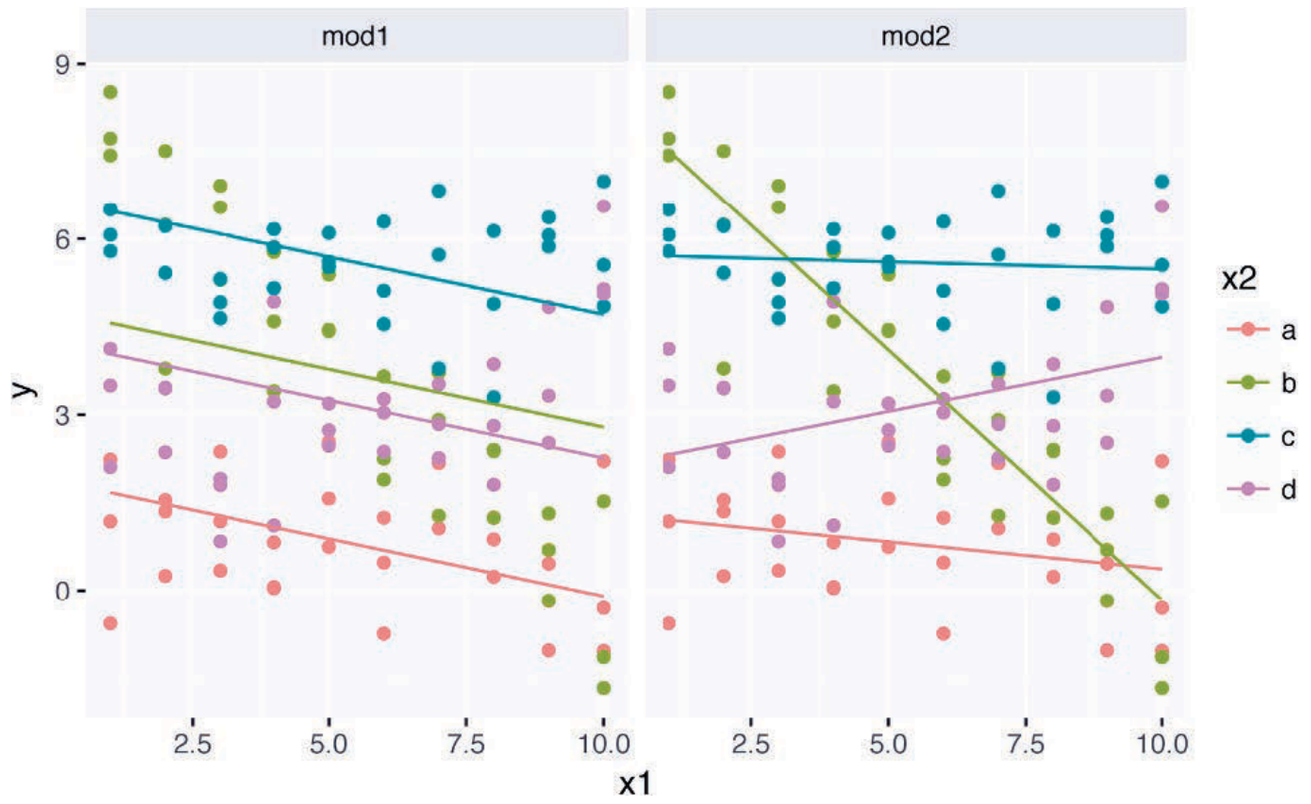




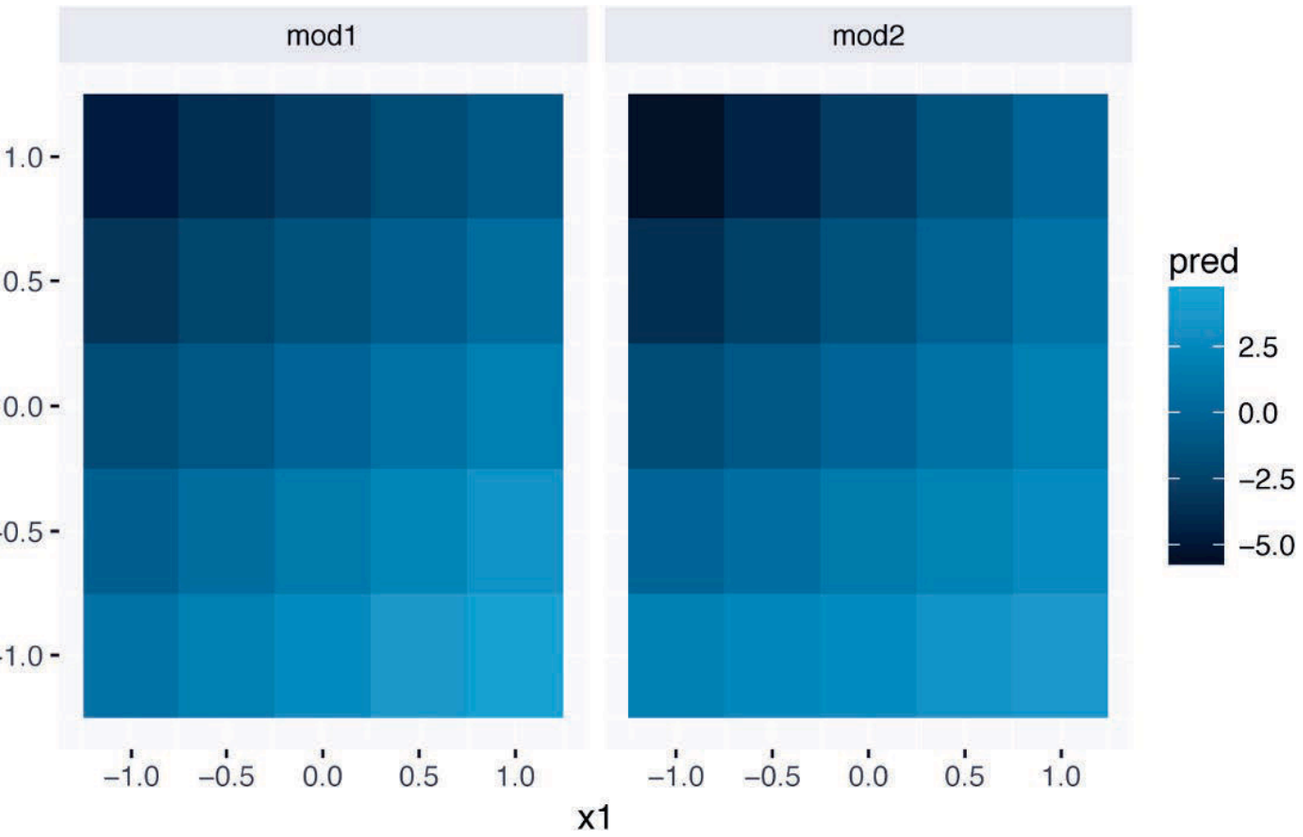
交互项(连续变量+分类变量)

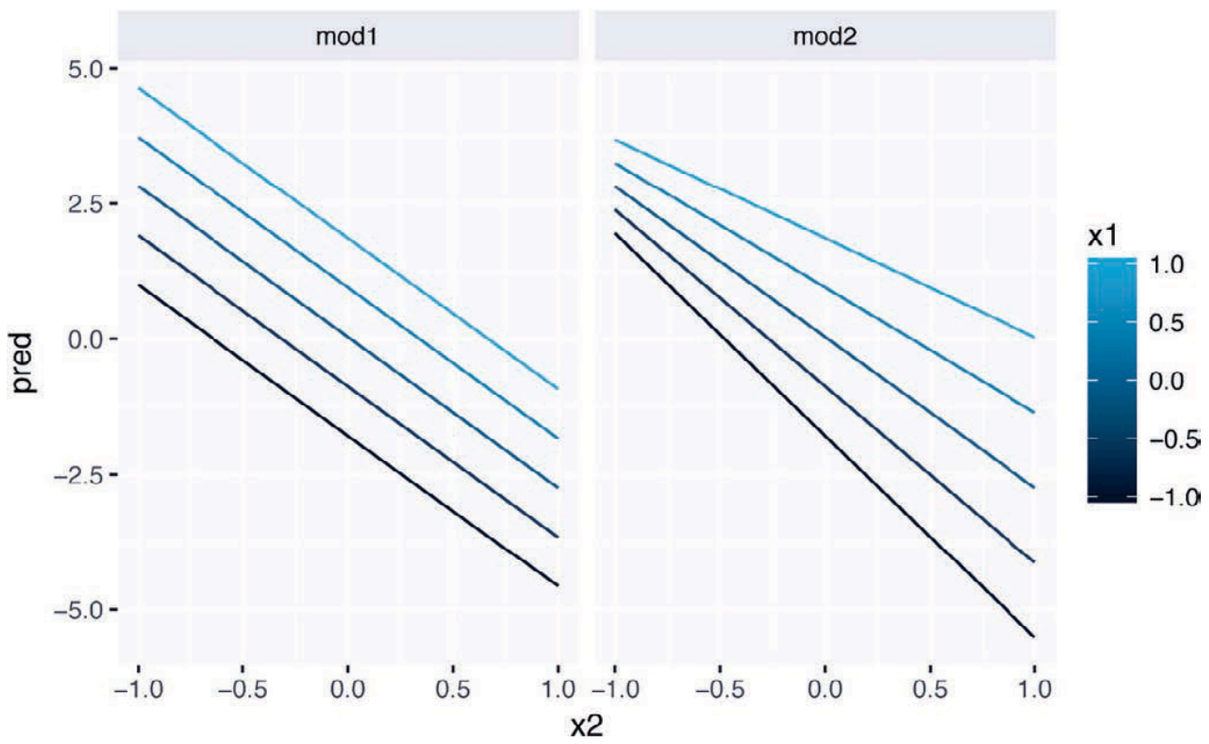


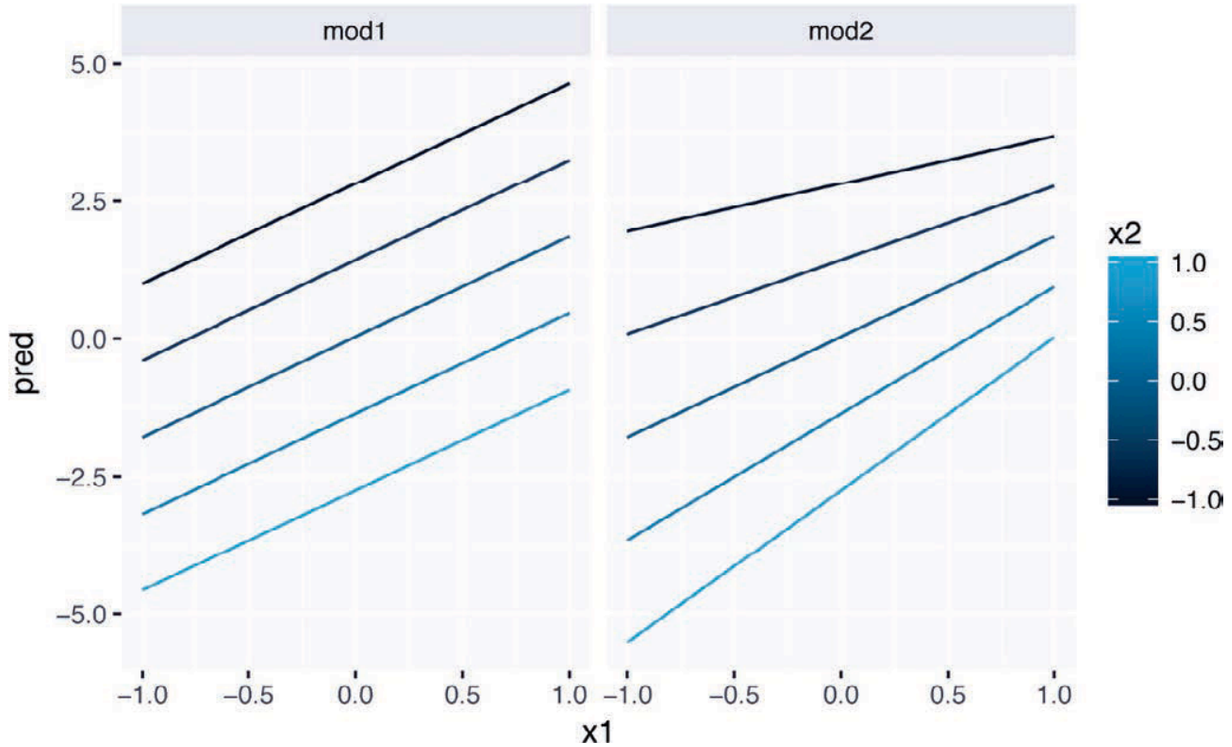




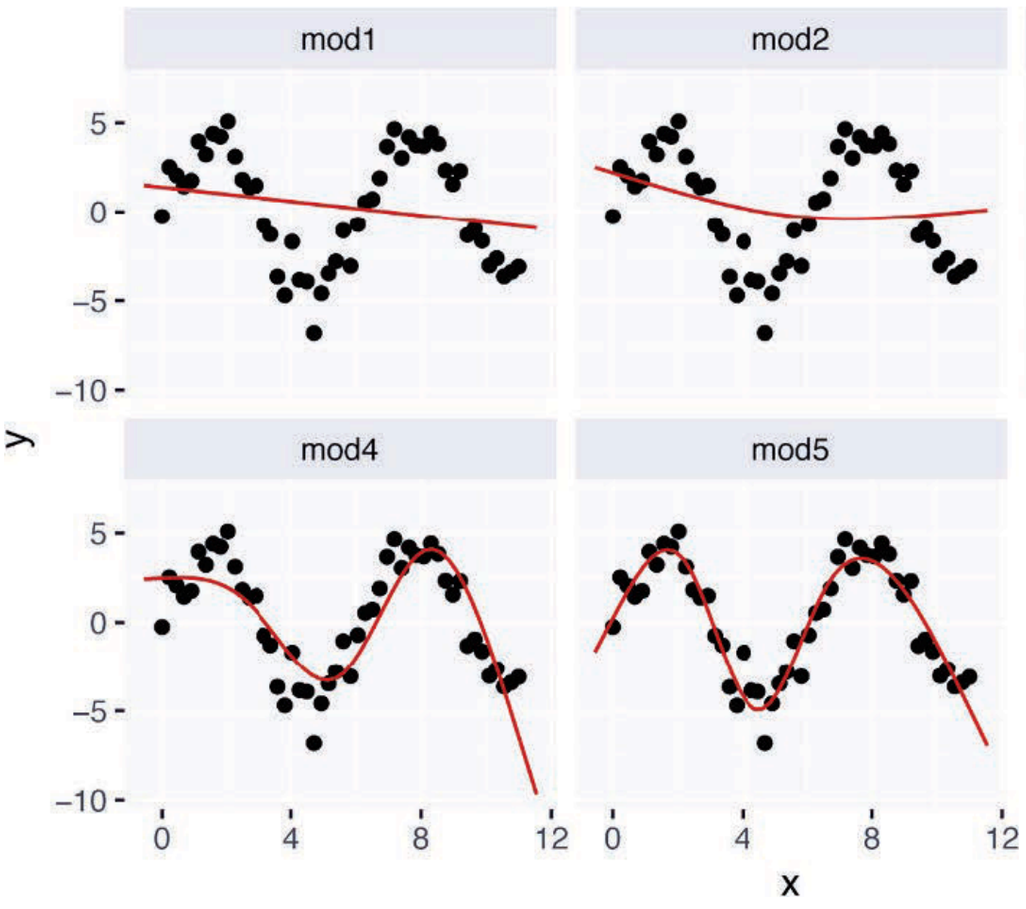
交互项(2个连续变量)

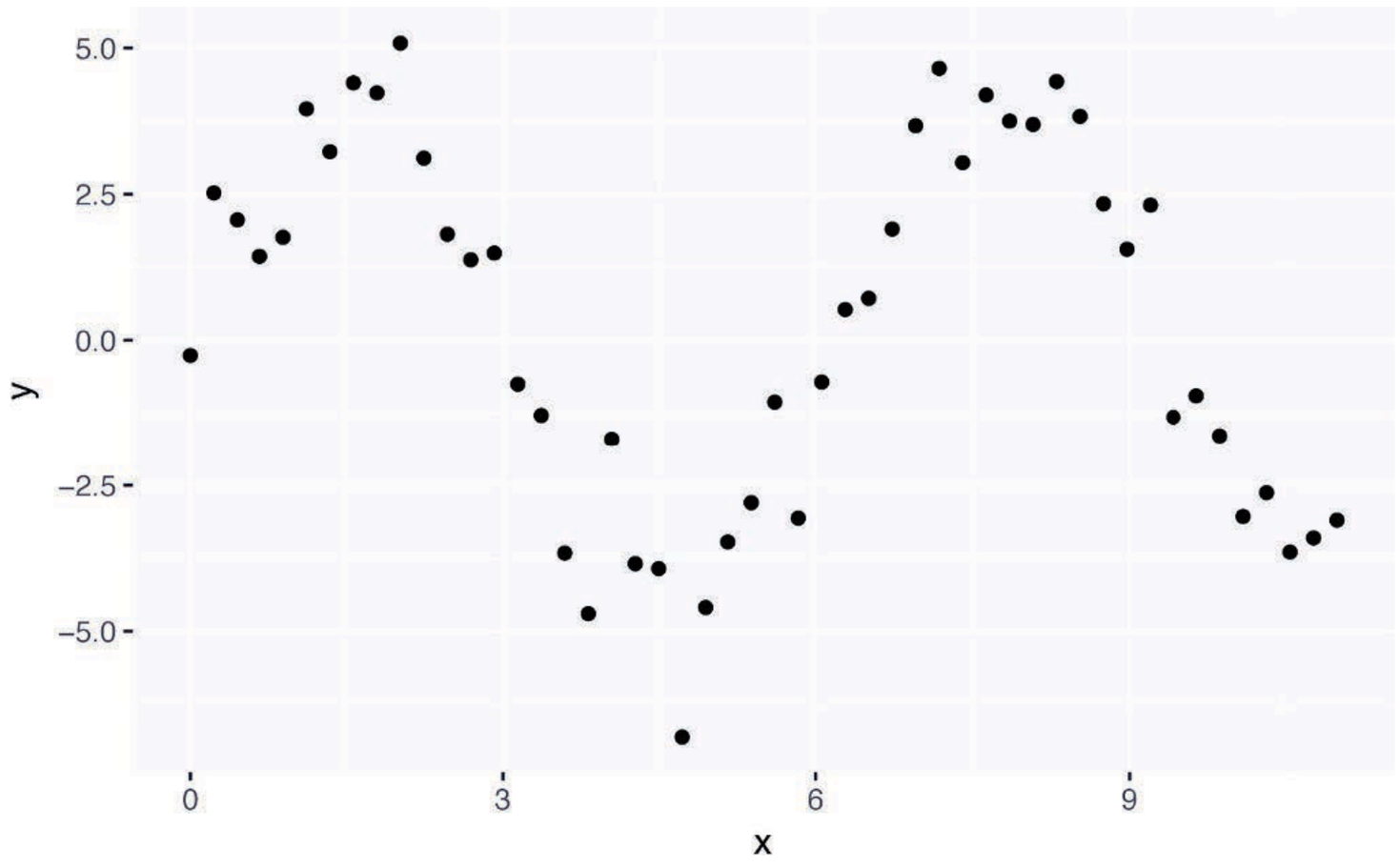

mod1
mod2
1.0
0.5
0.0
-0.5
-1.0
−1.0
−0.5
0.0
0.5
1.0
x1
pred
2.5
0.0
−2.5
−5.0


mod1
mod2
5.0
2.5
0.0
−2.5
−5.0
pred
−1.0
−0.5
0.0
0.5
1.0
x2
x1
1.0
0.5
0.0
−0.5
−1.0


mod1
mod2
5.0
2.5
0.0
−2.5
−5.0
pred
−1.0
−0.5
0.0
0.5
1.0
x1
x2
1.0
0.5
0.0
−0.5
−1.0

变量变换





模型构建

CH18

再看模型

模拟数据
线性模型

模式
残差

复杂数据
大量数据

真实数据
复杂模型

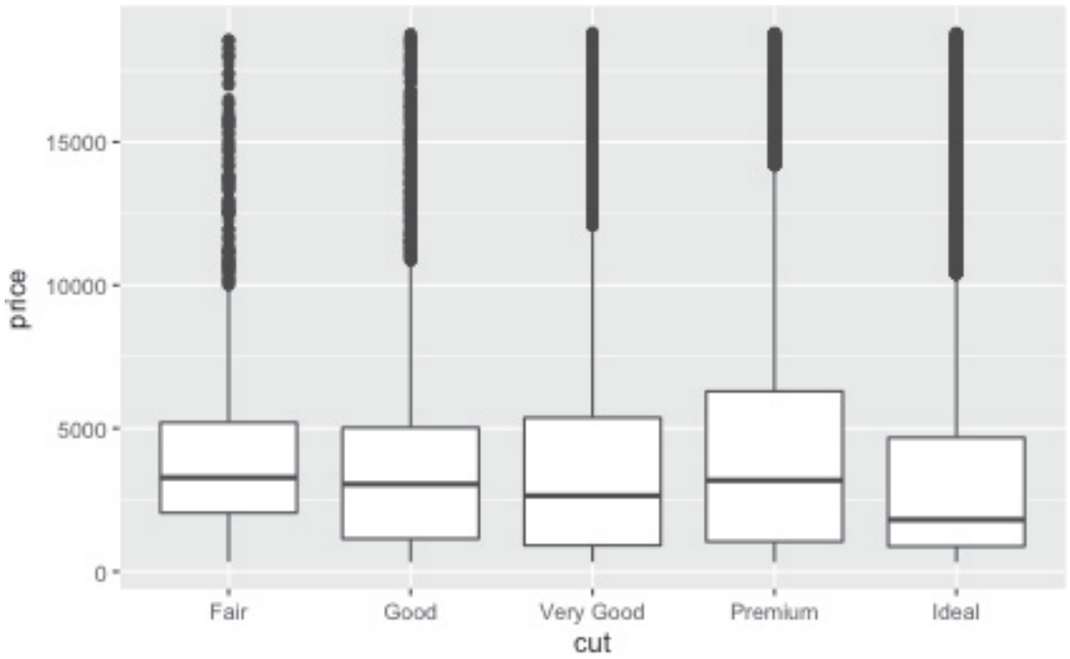
相应变量
残差

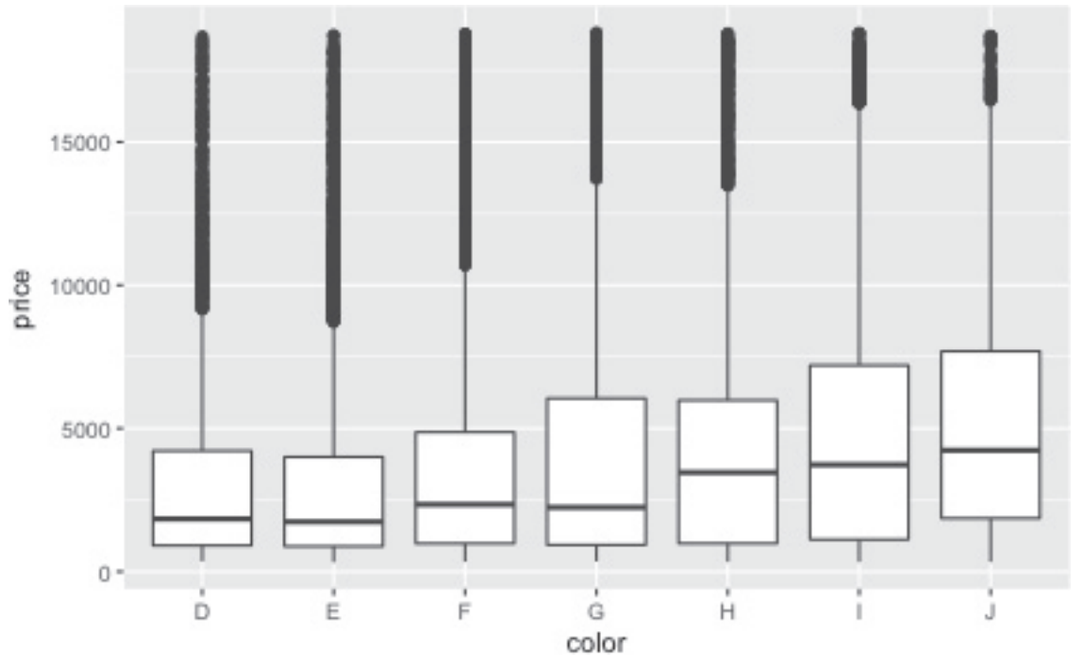
机器学习
黑盒效应

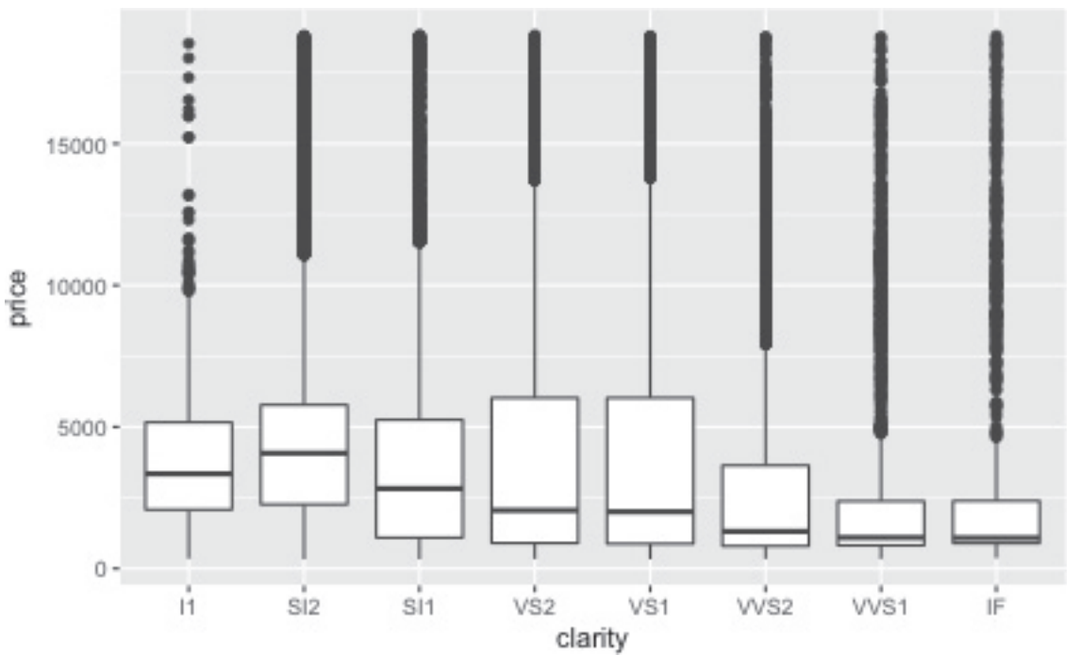
很久之前的艺术课上，老师告诉我：“艺术家应该知道何时完成作品。如果不能做得更好，那就结束它。如果不喜欢它，那就从头再来。否则，就去做别的事情吧。”在后来的生活中，我还听到了这句话：“坏裁缝会犯很多错误，好裁缝会努力纠正错误，而优秀的裁缝则从来不怕将有问题的衣服扔掉，重新开始。”

——Broseidon241

钻石例子：质量差的钻石更贵

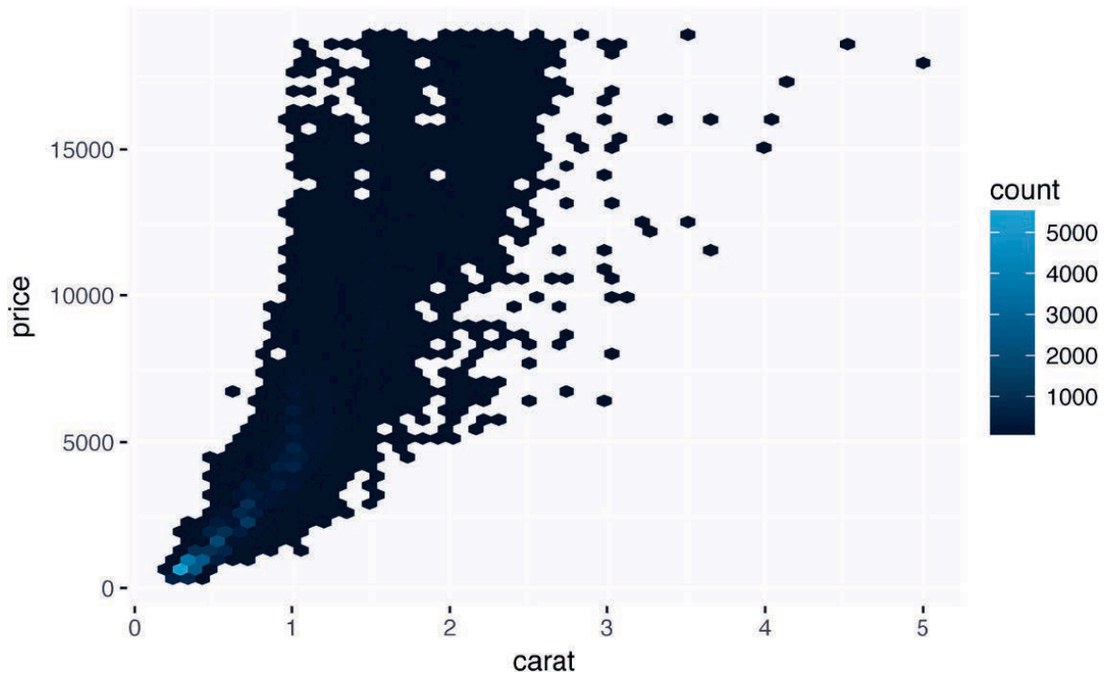




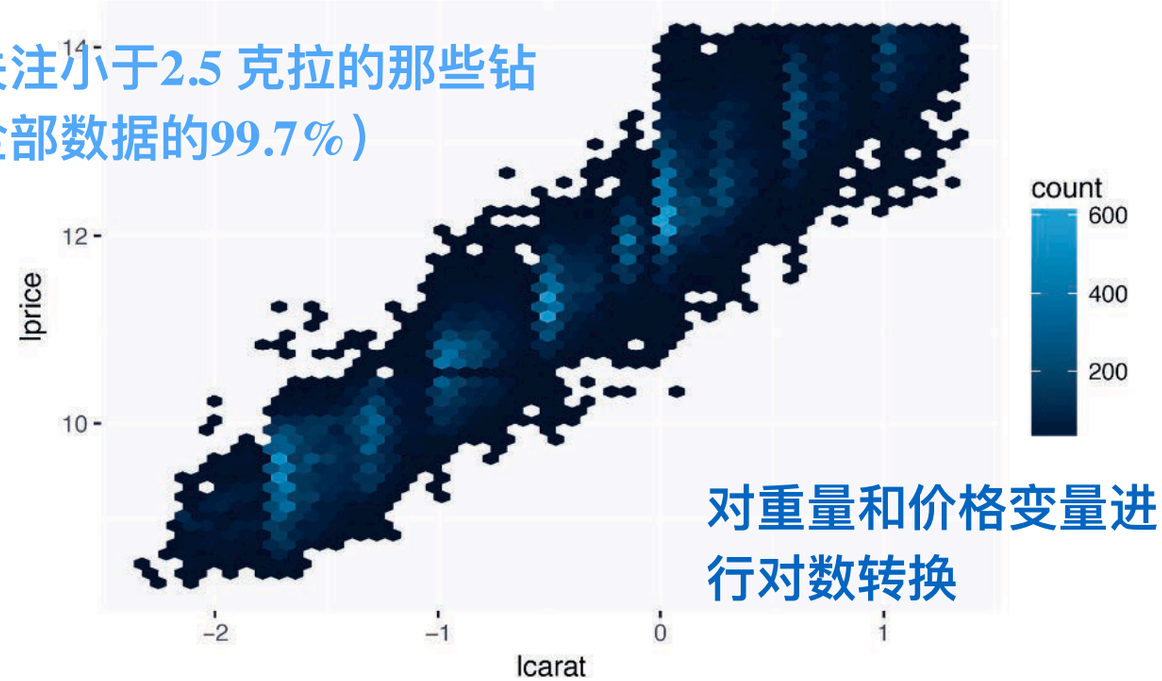


质量差的钻石
（切工差、颜色差、纯净度低）
具有更高的价格：

钻石例子：价格与重量

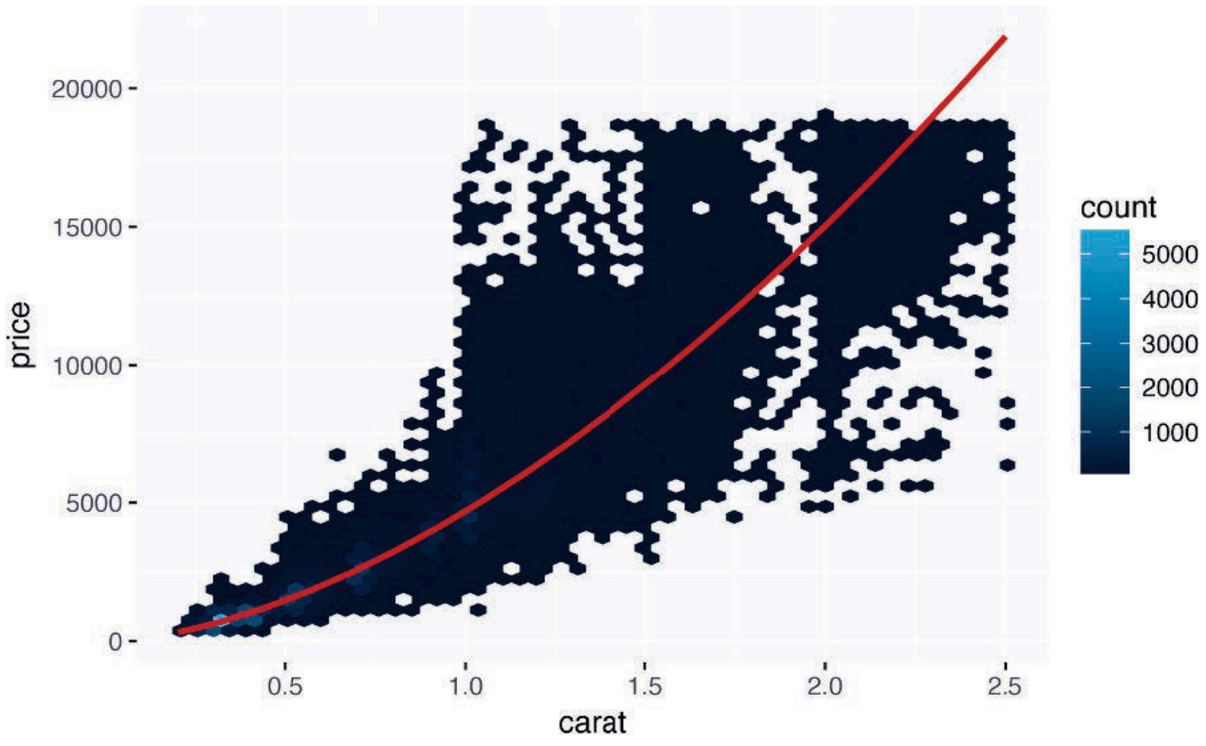


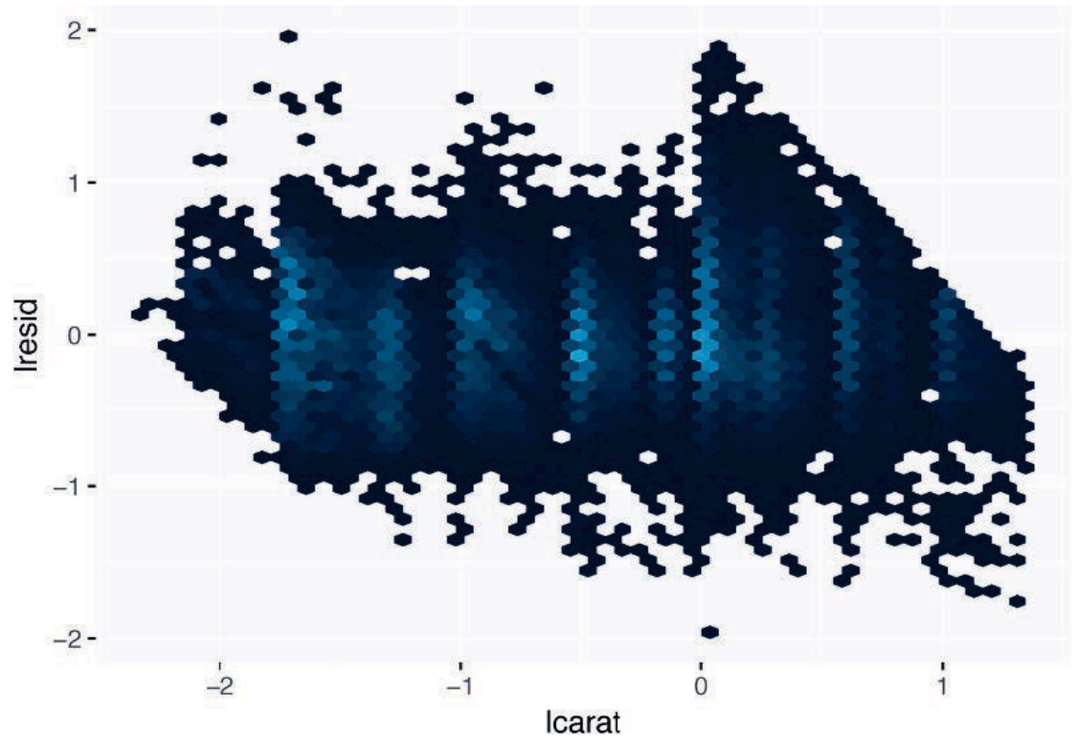
重点关注小于2.5 克拉的那些钻石（全部数据的99.7%）



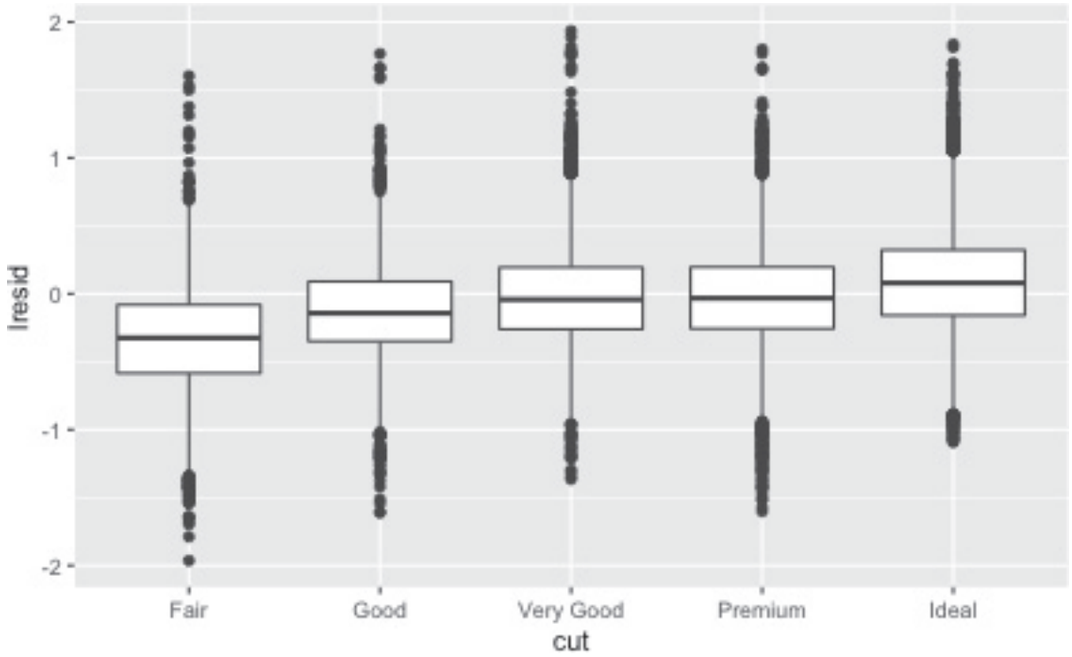
对重量和价格变量进行对数转换

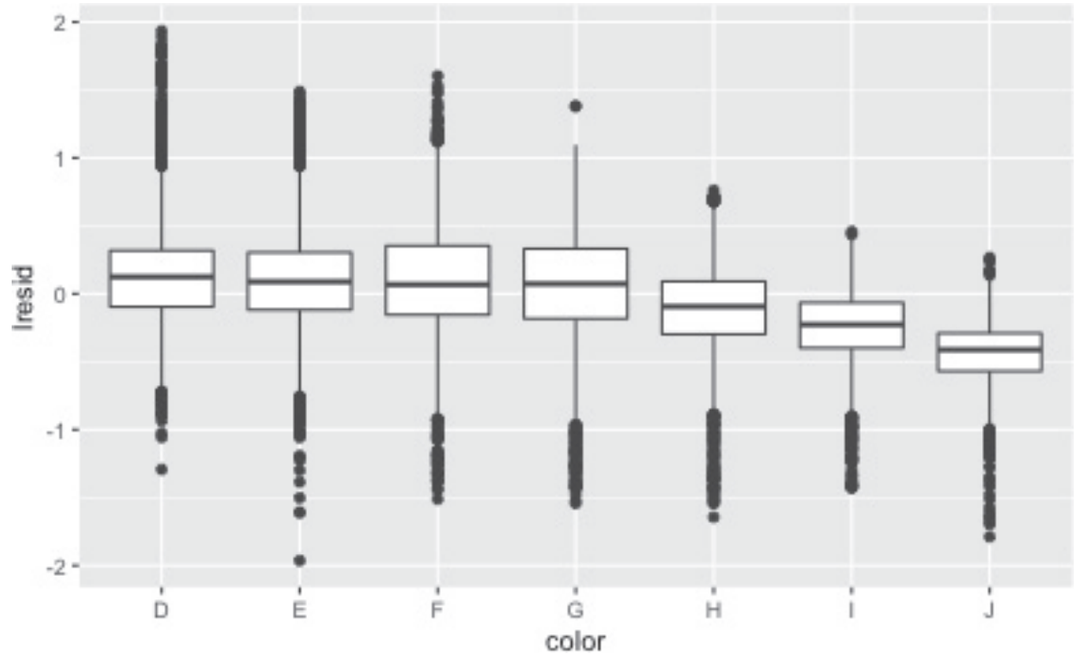
钻石的重量（carat）。重量是确定钻石价格的单一因素中最重要的一个，而质量差的钻石往往更重一些

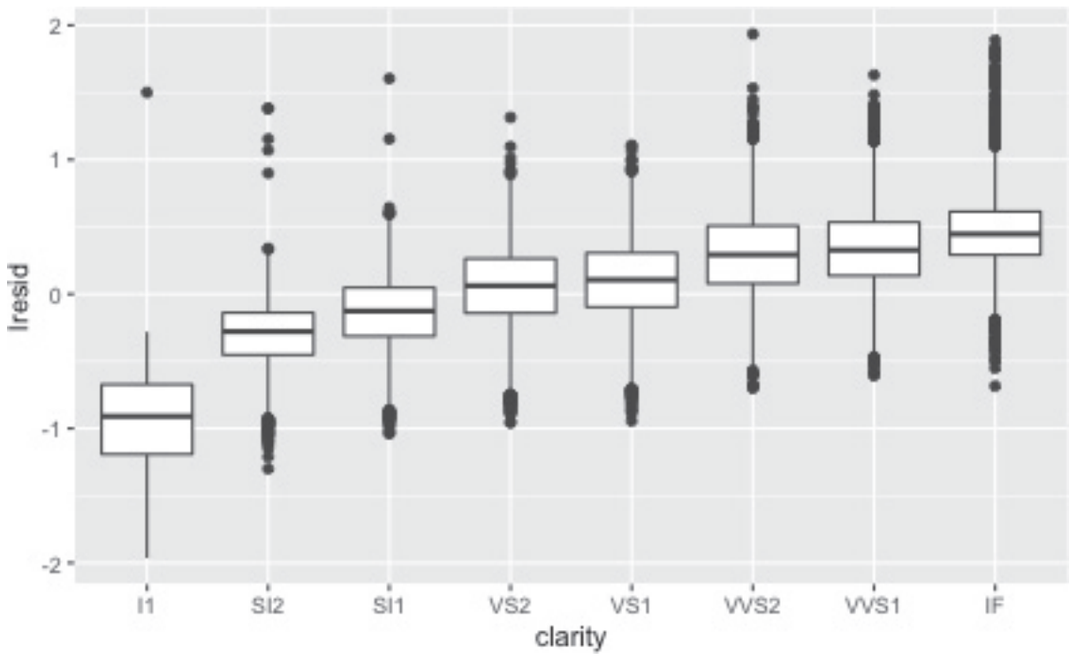




钻石例子：再看质量和价格

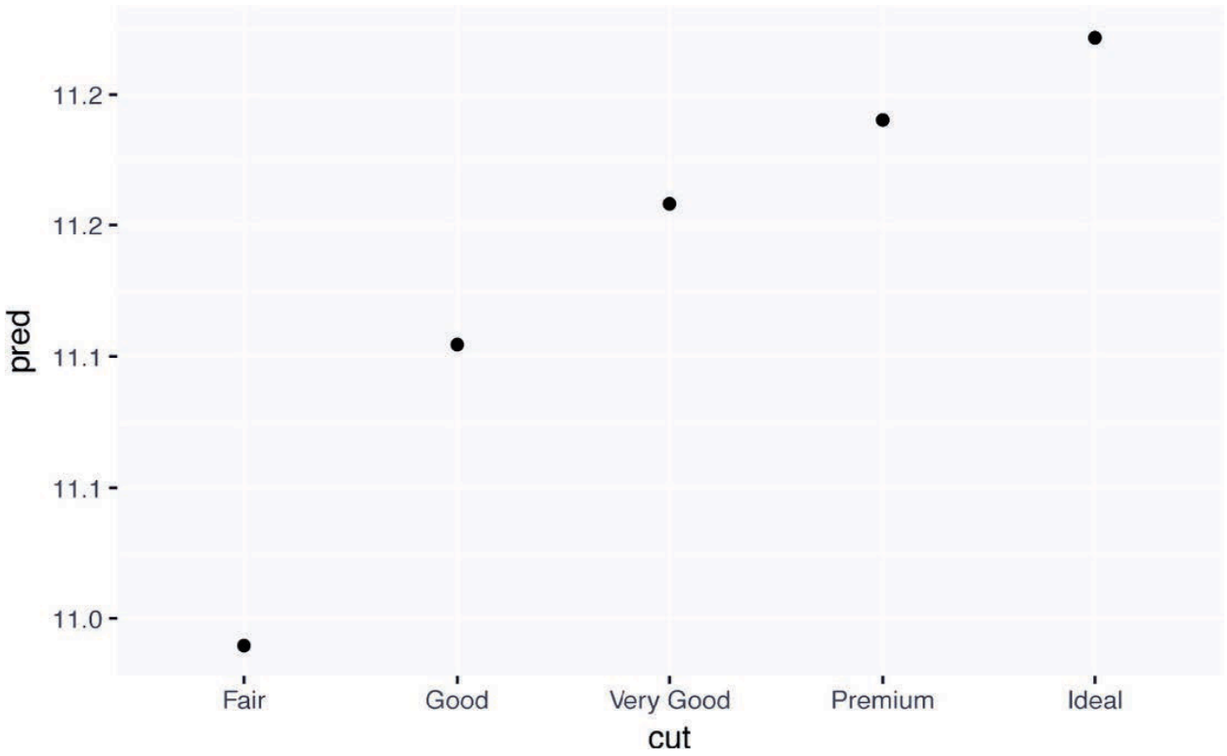




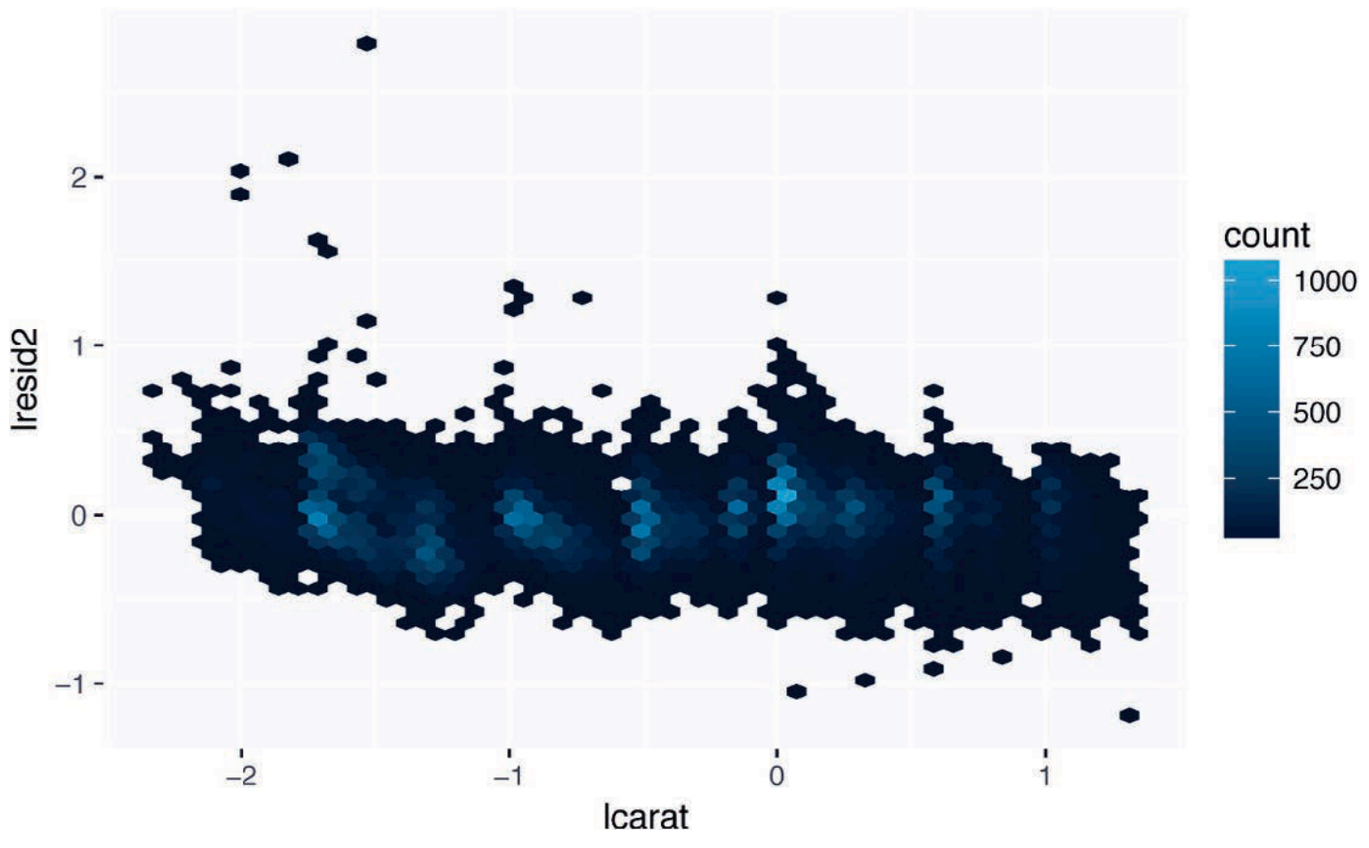


使用残差代替price 来重新绘图

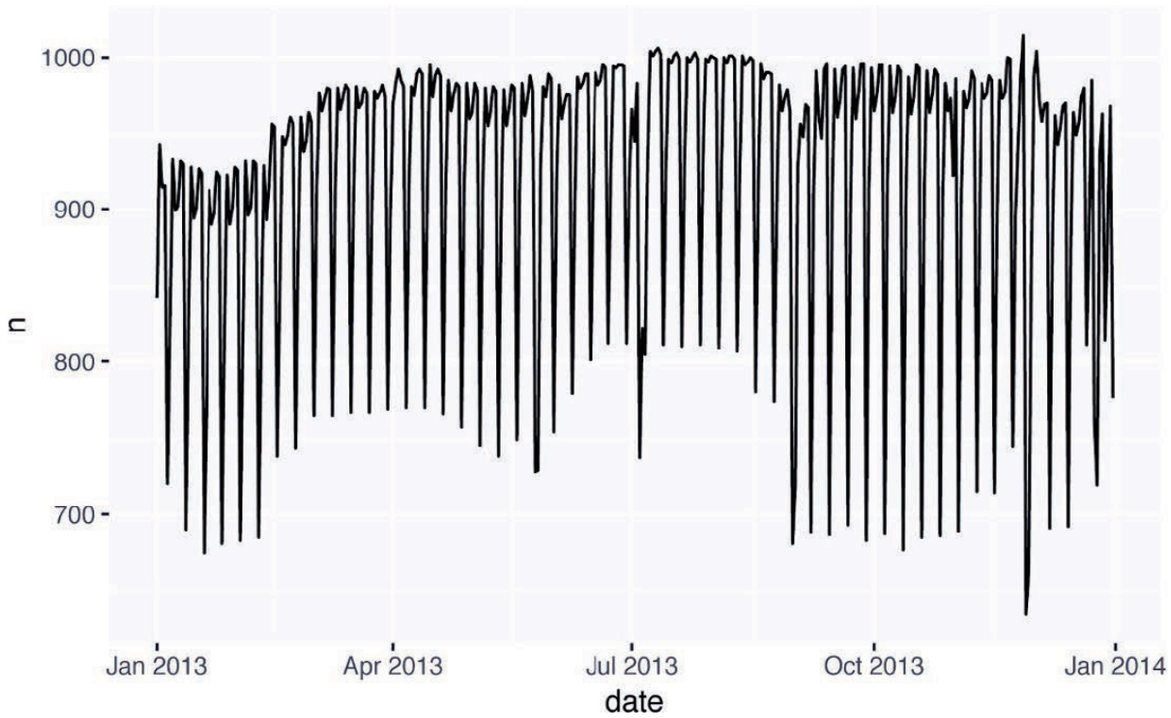
钻石例子: 更复杂模型

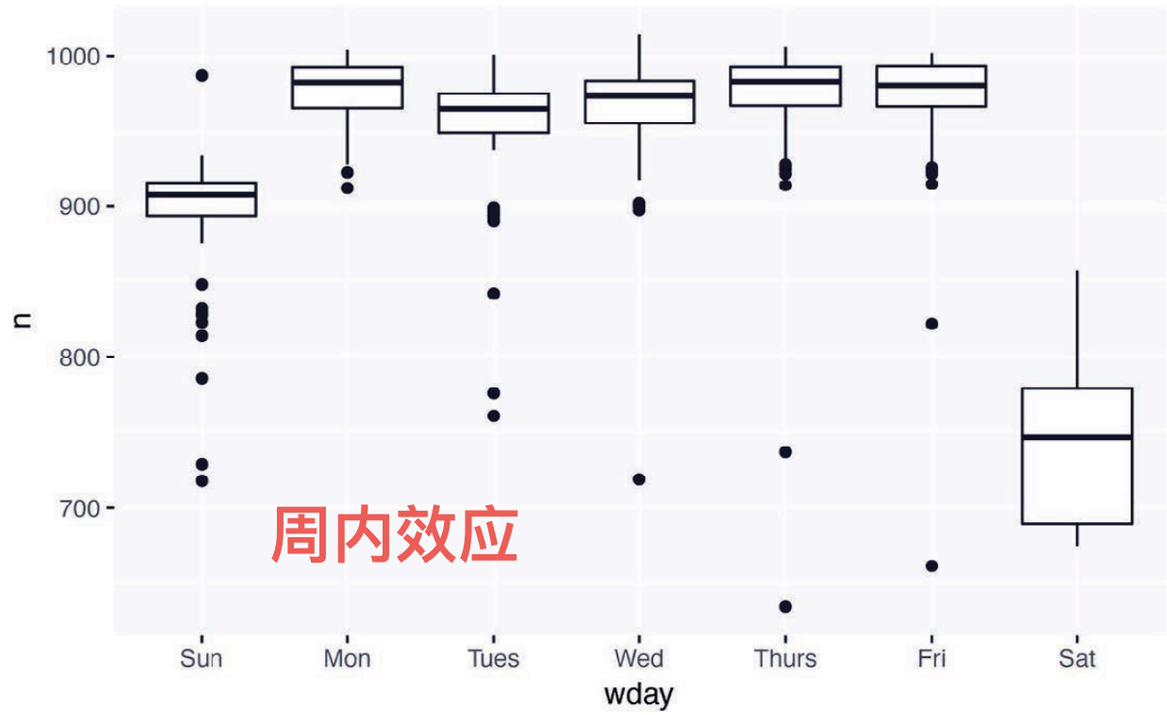


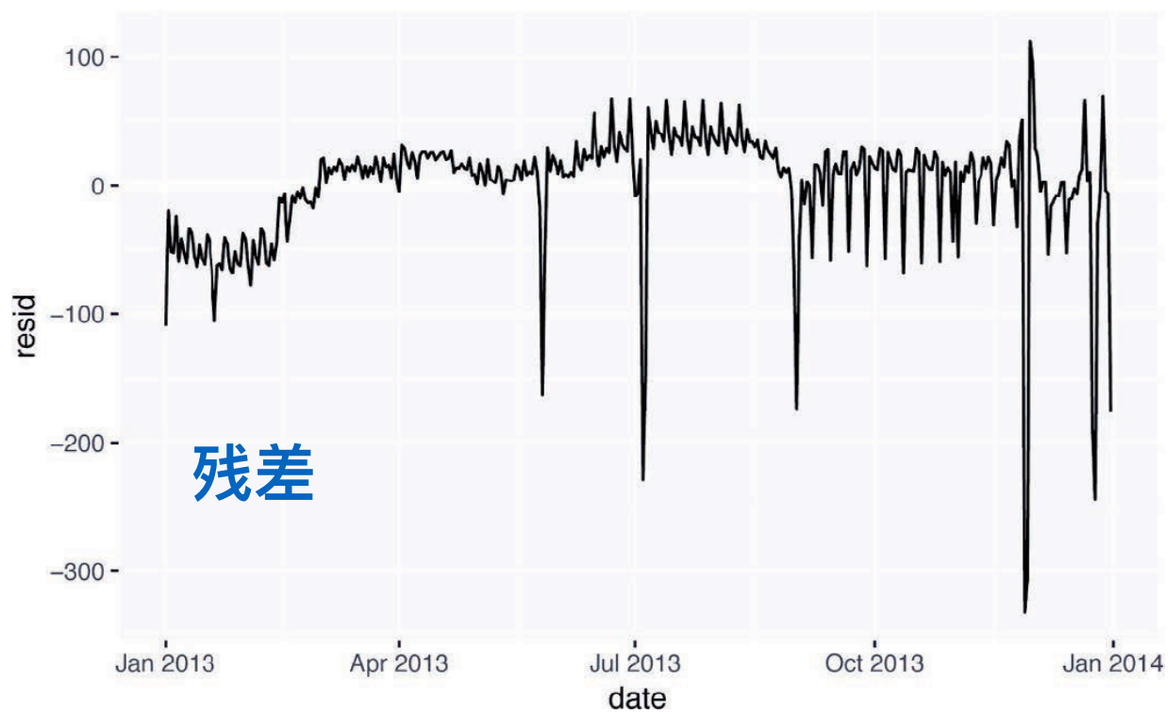
这张图说明一些钻石有非常大的残差。
残差为2 表示钻石的价格是预计价格的4倍
通常还应该检查一下异常值

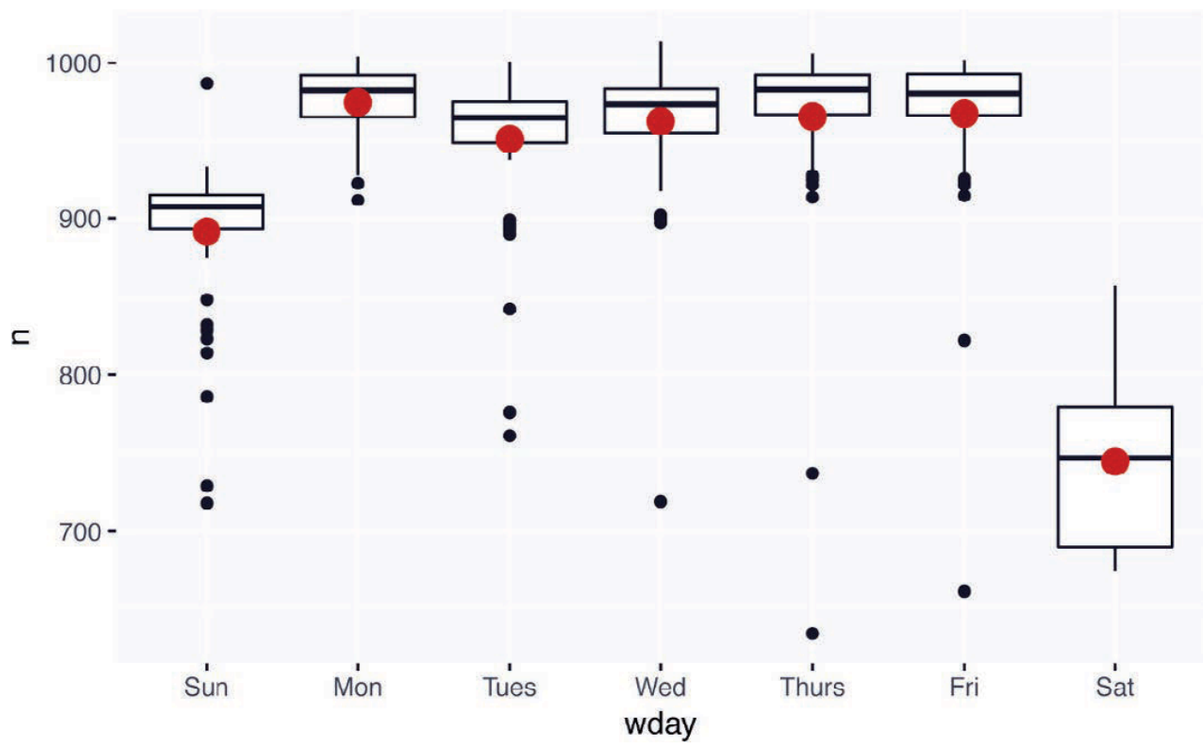


航班例子：每日航班数量


1000
900
800
700
n
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
date

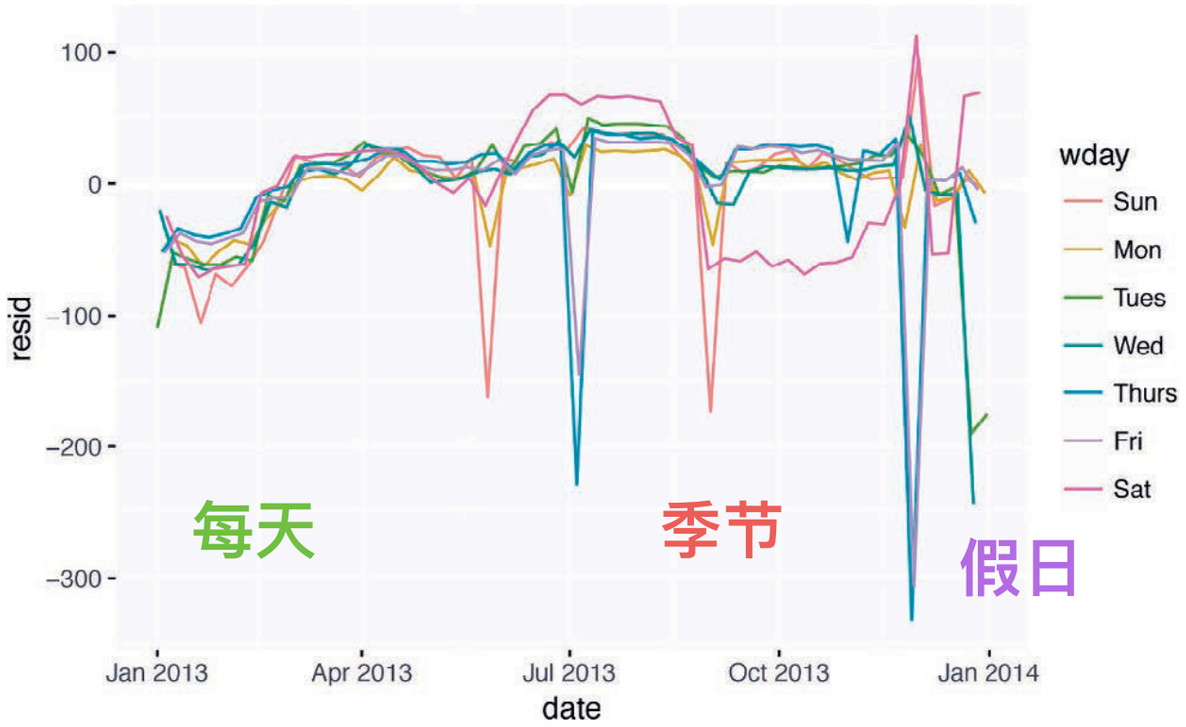

1000
900
800
700
n
周内效应
Sun
Mon
Tues
Wed
Thurs
Fri
Sat
wday

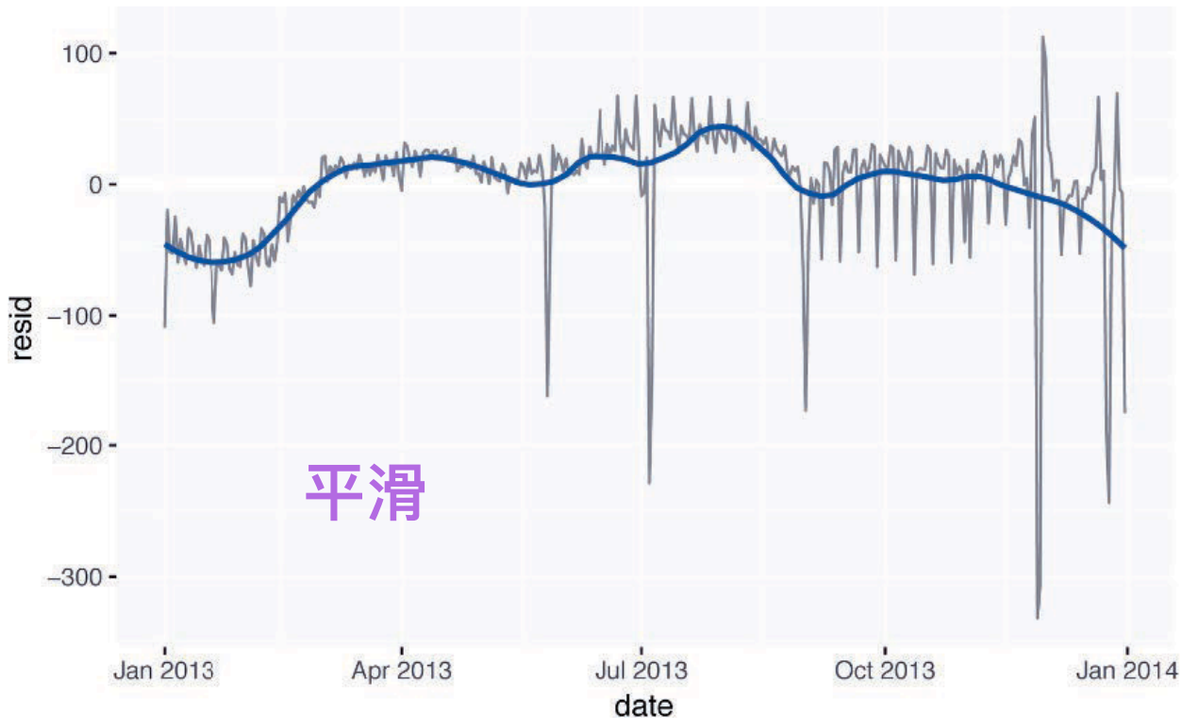

100
0
−100
−200
−300
resid
残差
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
date

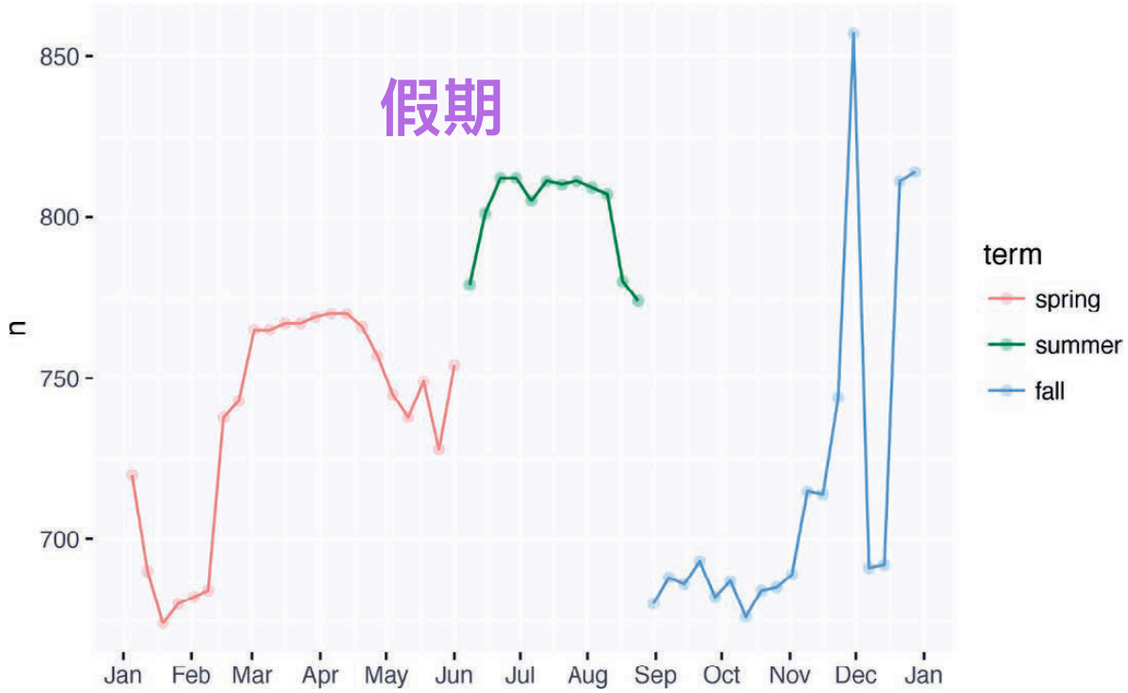

1000
900
800
700
n
Sun
Mon
Tues
Wed
Thurs
Fri
Sat
wday

R Data Science

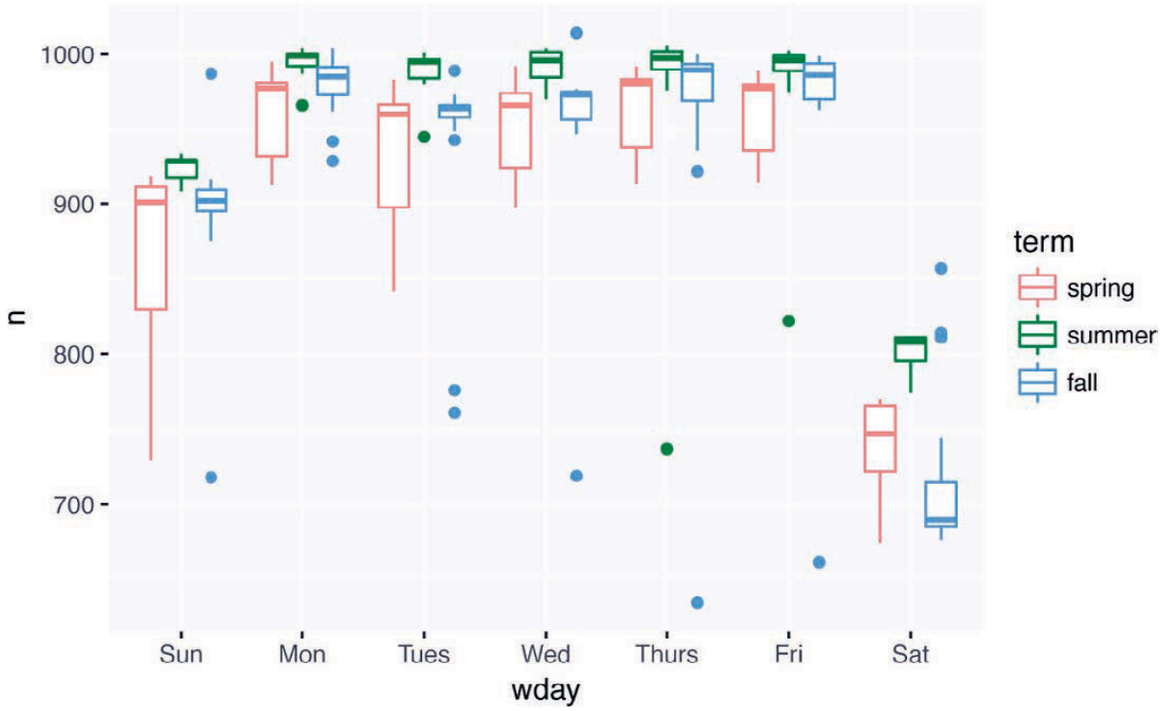
航班例子：每日航班数量

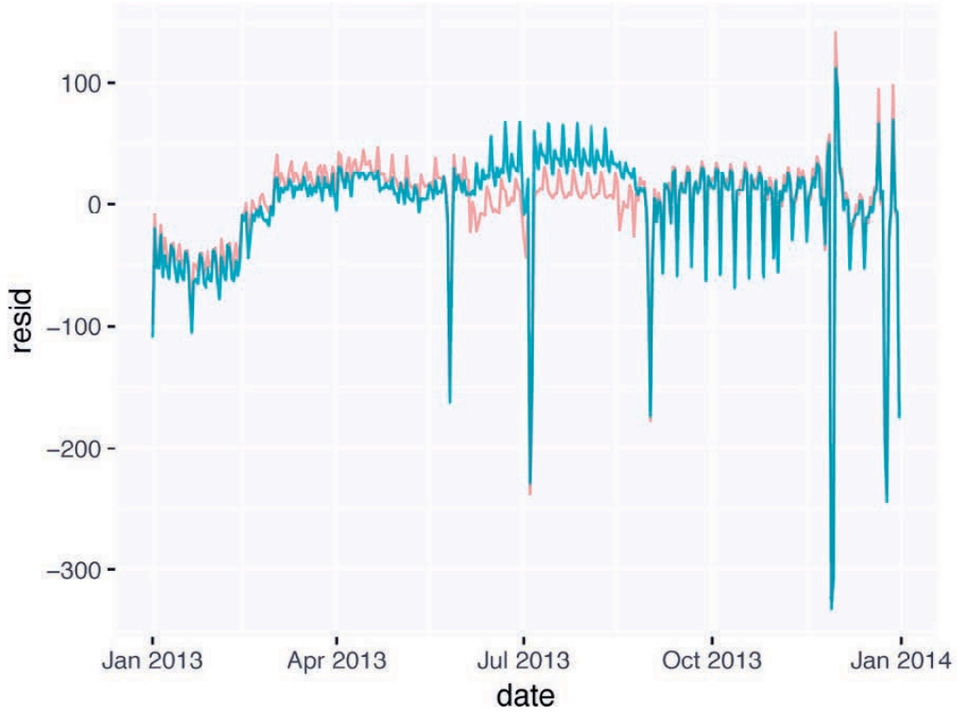


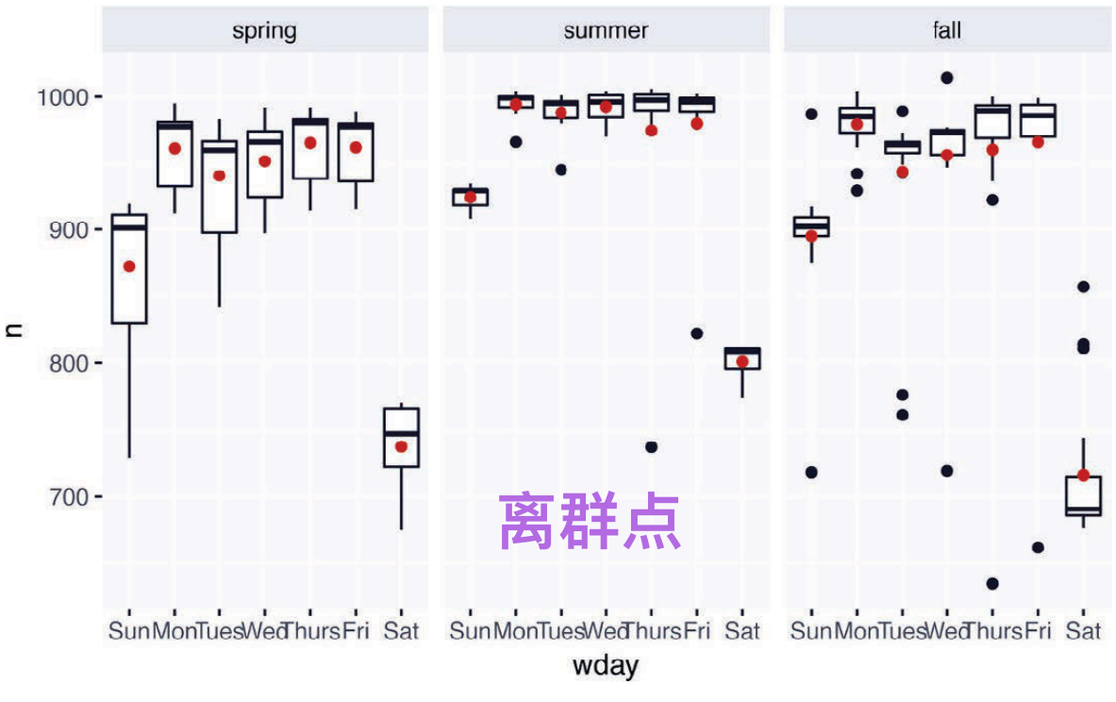


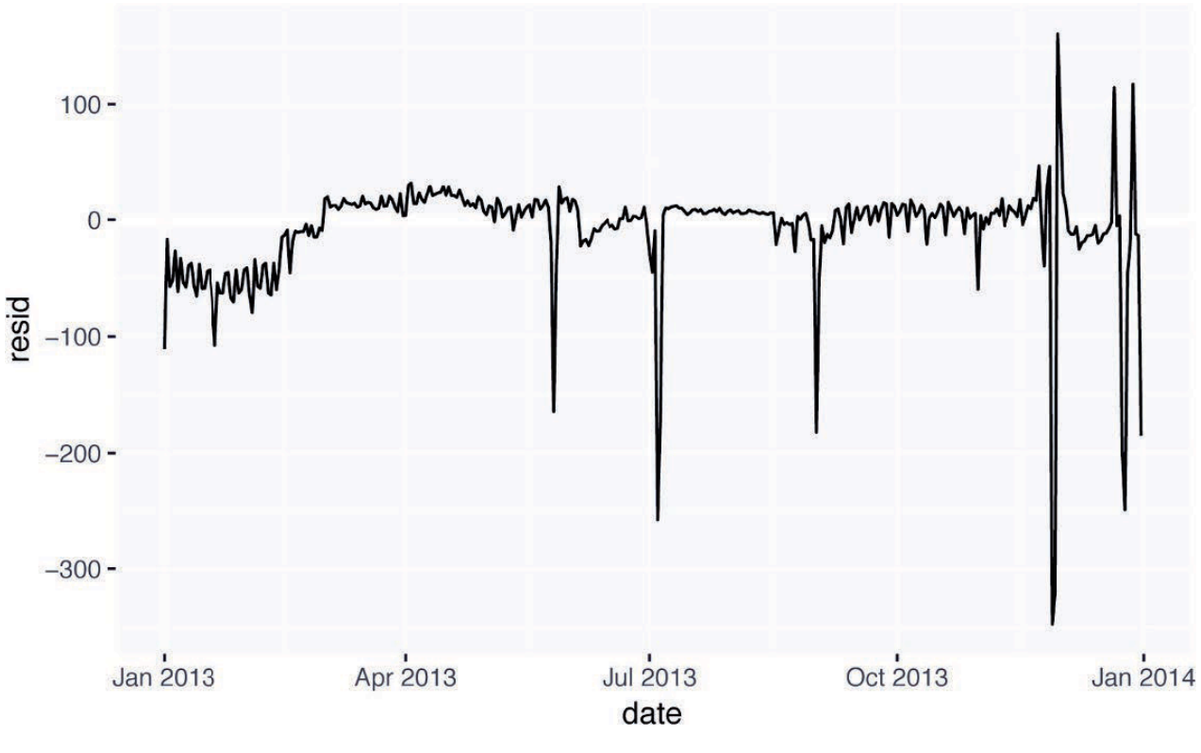


航班例子：残差


1000
900
800
700
n
Sun
Mon
Tues
Wed
Thurs
Fri
Sat
wday
term
spring
summer
fall


100
0
−100
−200
−300
resid
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
date
model
with_term
without_ter


spring
summer
fall
1000
900
800
700
n
Sun
Mon
Tues
Wed
Thurs
Fri
Sat
wday
离群点


100
0
−100
−200
−300
resid
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
date

练习

R Data Science

阅读教材

O'REILLY®

TURING 图灵程序设计丛书

全彩印刷

R数据科学

R for Data Science

摒弃其他R语言工具书从头到尾讲统计的陋习
从实用的R包出发，带你重新认识R和数据科学

[新西兰] 哈德利·威克姆 [美] 加勒特·格罗勒芒德 著
陈光欣 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

第17章 **第18章** **第19章**

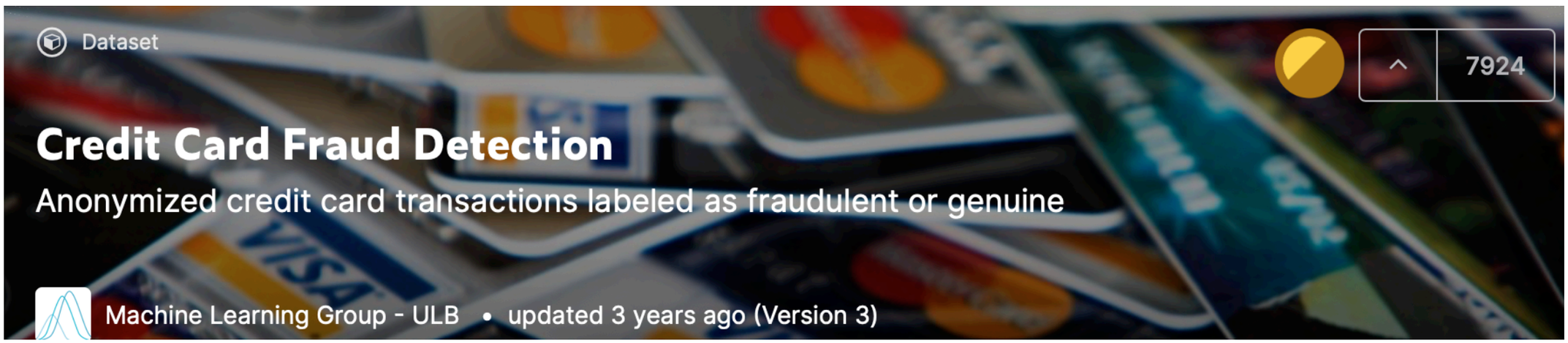
练习 - 0035

下周上课前提交方式和以前一样

- 信用卡欺诈数据库，见**creditcard.csv**

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
6	0.2514121	-0.0183068	0.27783758	-0.1104739	0.06692807	0.12853936	-0.1891148	0.13355838	-0.0210531	149.62	0
1	-0.0690831	-0.2257752	-0.638672	0.10128802	-0.3398465	0.1671704	0.12589453	-0.0089831	0.01472417	2.69	0
5	0.52497973	0.24799815	0.7716794	0.90941226	-0.689281	-0.3276418	-0.1390966	-0.0553528	-0.0597518	378.66	0
7	-0.2080378	-0.1083005	0.0052736	-0.1903205	-1.1755753	0.64737603	-0.2219288	0.06272285	0.06145763	123.5	0
5	0.40854236	-0.0094307	0.79827849	-0.1374581	0.14126698	-0.2060096	0.50229222	0.21942223	0.21515315	69.99	0
8	0.08496767	-0.2082535	-0.5598248	-0.0263977	-0.3714266	-0.2327938	0.10591478	0.25384422	0.08108026	3.67	0
5	-0.2196326	-0.1677163	-0.2707097	-0.1541038	-0.7800554	0.75013694	-0.2572368	0.03450743	0.00516777	4.99	0
1	-0.1567419	1.94346534	-1.0154547	0.05750353	-0.649709	-0.4152666	-0.0516343	-1.2069211	-1.0853392	40.8	0
7	0.05273567	-0.0734251	-0.2680916	-0.2042327	1.0115918	0.37320468	-0.3841573	0.01174736	0.14240433	93.2	0

- 建模分析信用卡欺诈
- 分析模型有效性
- 使用全部数据
- 使用更复杂模型和工具



https://www.kaggle.com/mlg-ulb/creditcardfraud

课程项目

- 选择一个**金融应用**方面数据集，进行**数据分析或者建模分析**
- 金融应用方面数据集合可以到**Kaggle**等网站去寻找
- 不要寻找过于常见的金融应用和数据集合

分组和上次包介绍一样是否可以?

- 组团自愿，人数不要太多或太少
- 数据分析和建模分析的工具和包不限制
- 建议不要过于简单，要有一定**新颖性和实用性**

- **6月22日**课堂报告
- **6月10日**初步确定分组和选题，发给助教，11-13日商量最后确定

谢谢！

孙惠平

sunhp@ss.pku.edu.cn