预测中国电影票房

小组成员：徐寅秋、叶家寅
姜昊、唐瀚林

* 中国电影票房指在国内上映的电影，制片地区无约束

背景与研究意义

在大众的精神文化消费比重增加，有关电影的营销、社交信息积累更多的时代下，我们希望能够使用影片上映时的信息预测它的最终票房

为影院在上映后期的排片提供数据支持

在电影后期的营销和发行环节：对影片的定档日期选择，营销策略有帮助。尤其是发行公司在保底发行模式中，起到重要作用。

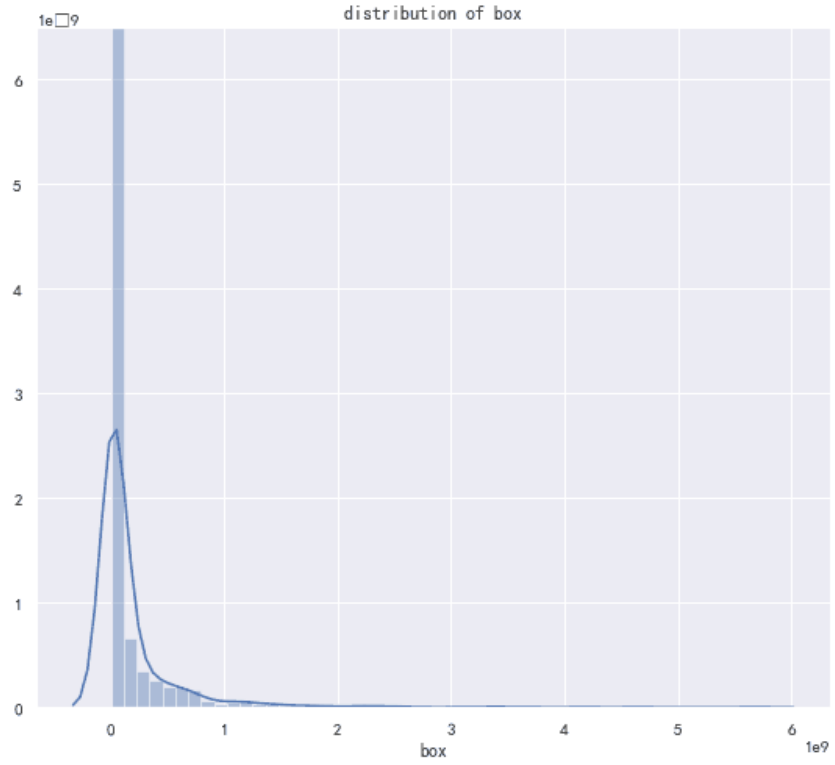
数据来源

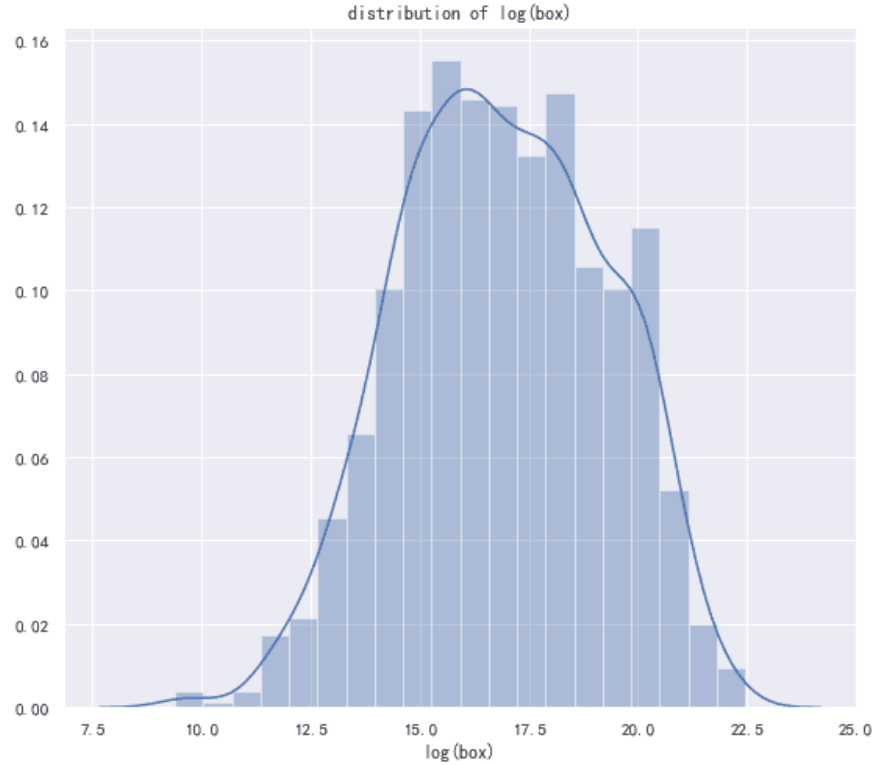
- 猫眼专业版
  - 分析页面
  - 图表数据API
- Python 爬虫
- 缺失数据人工补全

数据内容

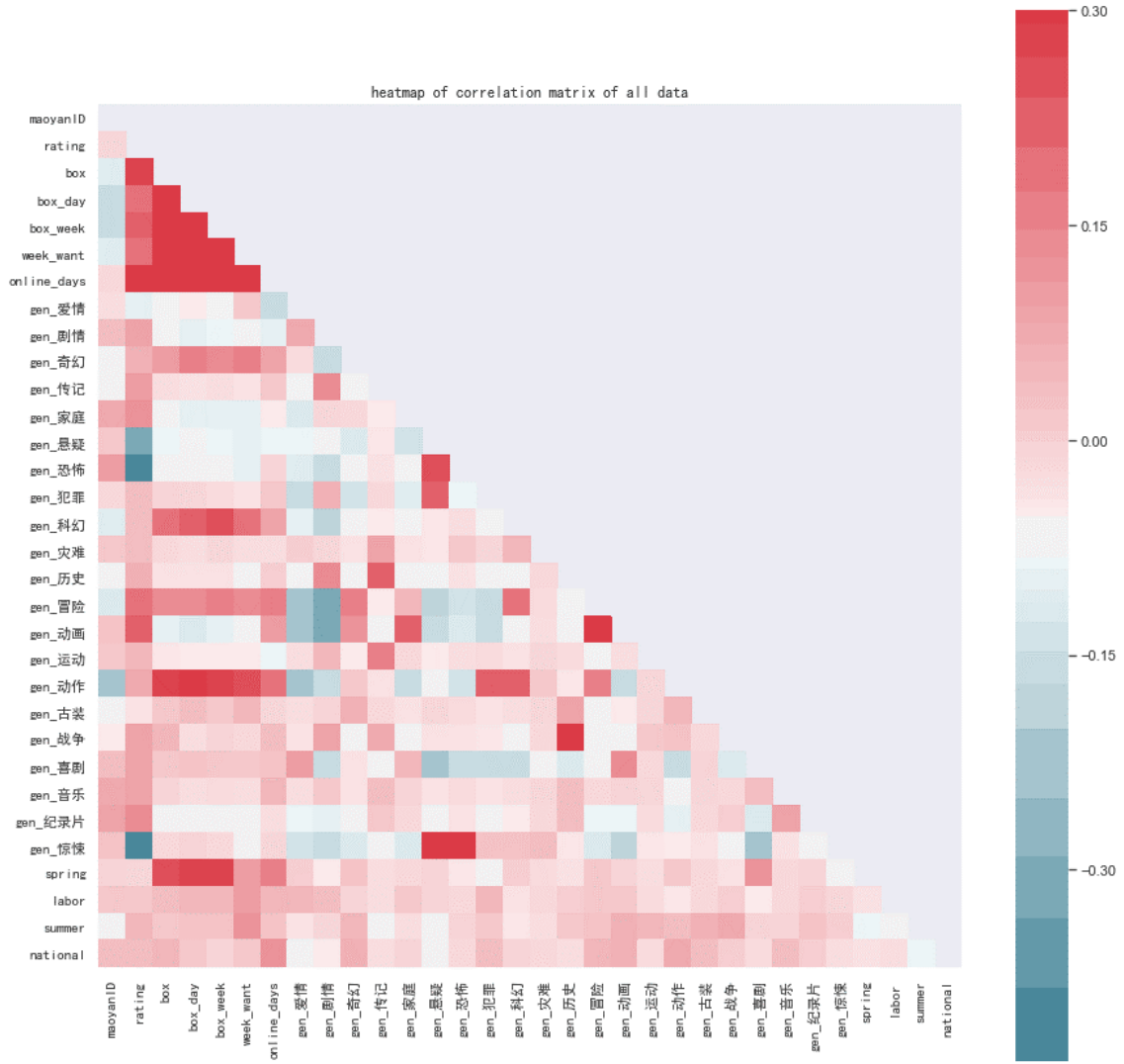
数据名程	数据解释
影片类型	科幻、喜剧、动作……
上映天数	
特殊档期	暑期、国庆、春节、五一
票房（首日，最终）	
想看人数	上映前一周猫眼新增点击想看总人数
是否为IMAX	
制片地区	
微博话题讨论量	
评分	
物料总播放量	上映前一周各大视频平台预告片/宣传片播放总量

数据可视化 票房的分布


distribution of box
1e□9
6
5
4
3
2
1
0
0
1
2
3
4
5
6
1e9
box


distribution of log(box)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
log(box)

数据可视化 各变量间的相关性



- 第一天票房、第一周票房、上映天数、想看人数间有明显相关性
- 科幻、动作、春节档与票房明显相关

K-means

根据$screeplot$,分为三类

Cluster3:

$$box \in [2.0 \times 10^9, \infty)$$

Cluster2:

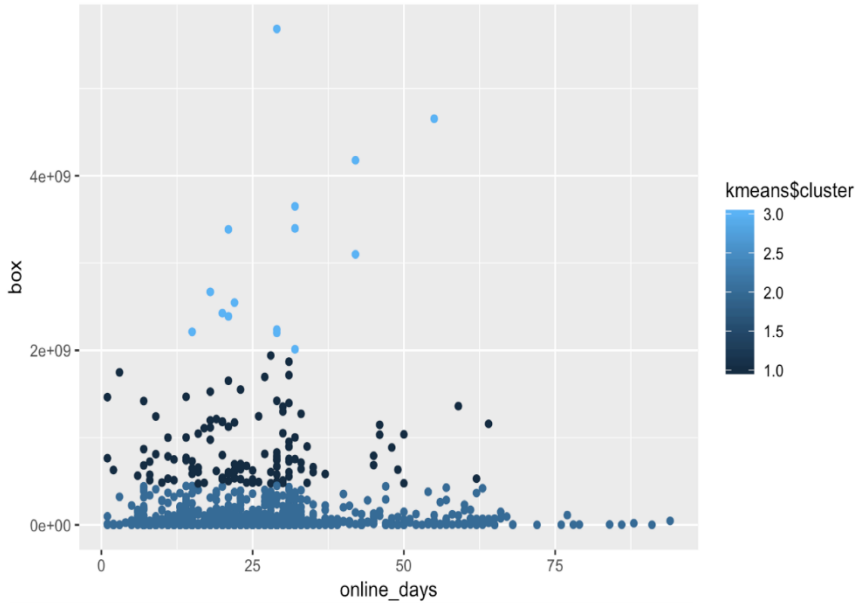
$$box \in [0, 5.0 \times 10^8\ )$$

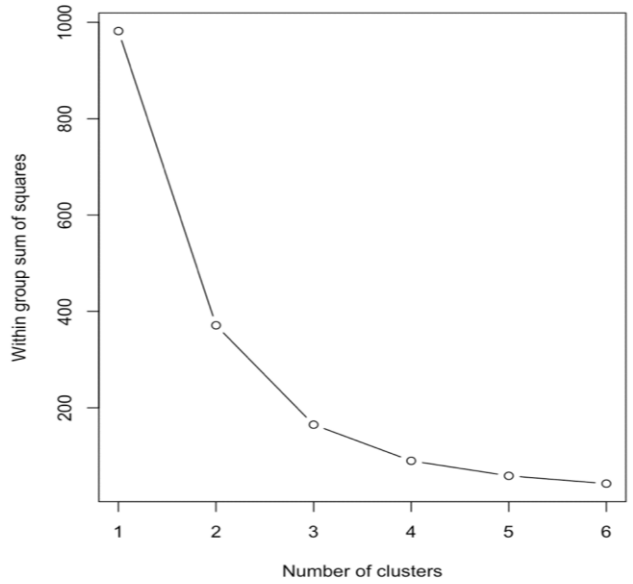
Cluster1:

$$box \in [5.0 \times 10^8, 2.0 \times 10^9)$$

由于样本量小，第一、三类仅用Fisher's LDA做粗略的分类预测。

第二类用多维线性回归模型预测。





K-means

组	评分	首日票房	IMAX比例	想看人数	占比最高地区	占比最高类型	占比最高档期
1	8.509	853336957	0.598	59280	中国大陆	动作	暑期档
2	7.687	57554062	0.085	11034	中国大陆	喜剧	暑期档
3	9.127	3018454545	0.818	104764	中国大陆	动作	春节档

TPM: 0.02807018

正态性检验 Kolmogorov Smirnov test

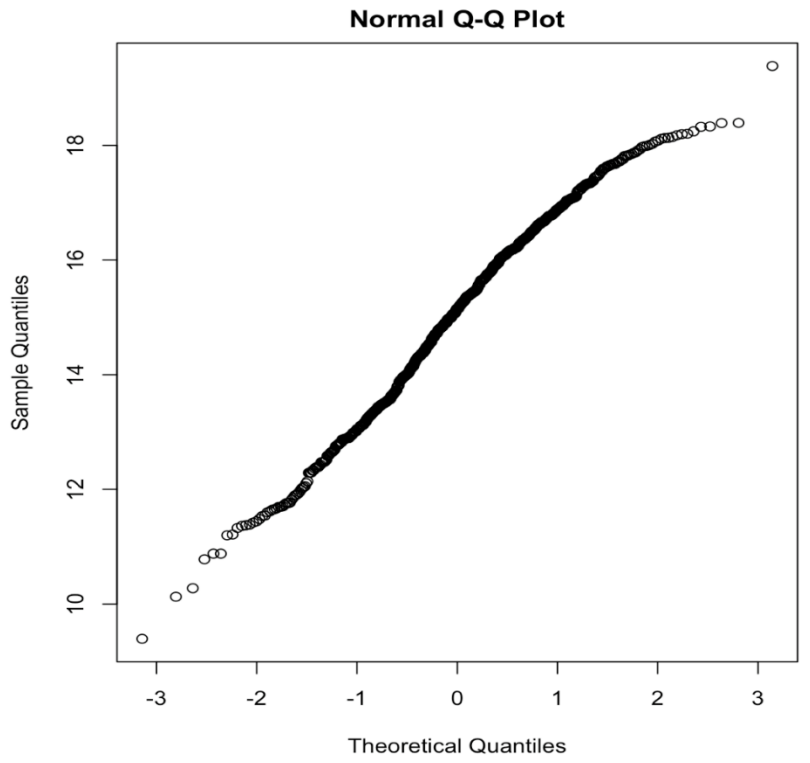
- $box \rightarrow \log(box)$

```
        Two-sample Kolmogorov-Smirnov test

data:  scale(mytablenew$box) and y
D = 0.26937, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
        Two-sample Kolmogorov-Smirnov test

data:  scale(log(mytablenew$box)) and y
D = 0.04674, p-value = 0.3369
alternative hypothesis: two-sided
```



变量选择&结果 Step & Cross validation

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.826e-01  2.150e-01   3.640 0.000296 ***
log(box_day)  9.323e-01  1.297e-02  71.861  < 2e-16 ***
rating        2.115e-01  2.030e-02  10.420  < 2e-16 ***
reg_中国大陆  1.241e-01  4.607e-02   2.694 0.007261 **
reg_美国      1.797e-01  5.275e-02   3.407 0.000701 ***
reg_印度      3.325e-01  1.580e-01   2.105 0.035710 *
week_want     6.299e-06  1.542e-06   4.084 5.03e-05 ***
gen_爱情     -2.116e-01  4.604e-02  -4.597 5.24e-06 ***
gen_传记      3.368e-01  1.194e-01   2.822 0.004930 **
gen_悬疑     -8.555e-02  5.616e-02  -1.523 0.128190
gen_恐怖      4.321e-01  9.481e-02   4.557 6.28e-06 ***
gen_运动     -3.331e-01  1.261e-01  -2.642 0.008456 **
gen_纪录片    1.666e-01  1.086e-01   1.534 0.125660
gen_惊悚      1.335e-01  7.517e-02   1.775 0.076342 .
spring        3.634e-01  1.231e-01   2.952 0.003282 **
labor         3.620e-01  1.025e-01   3.532 0.000445 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4356 on 600 degrees of freedom
Multiple R-squared:  0.9549,    Adjusted R-squared:  0.9537
F-statistic: 846.1 on 15 and 600 DF,  p-value: < 2.2e-16
```

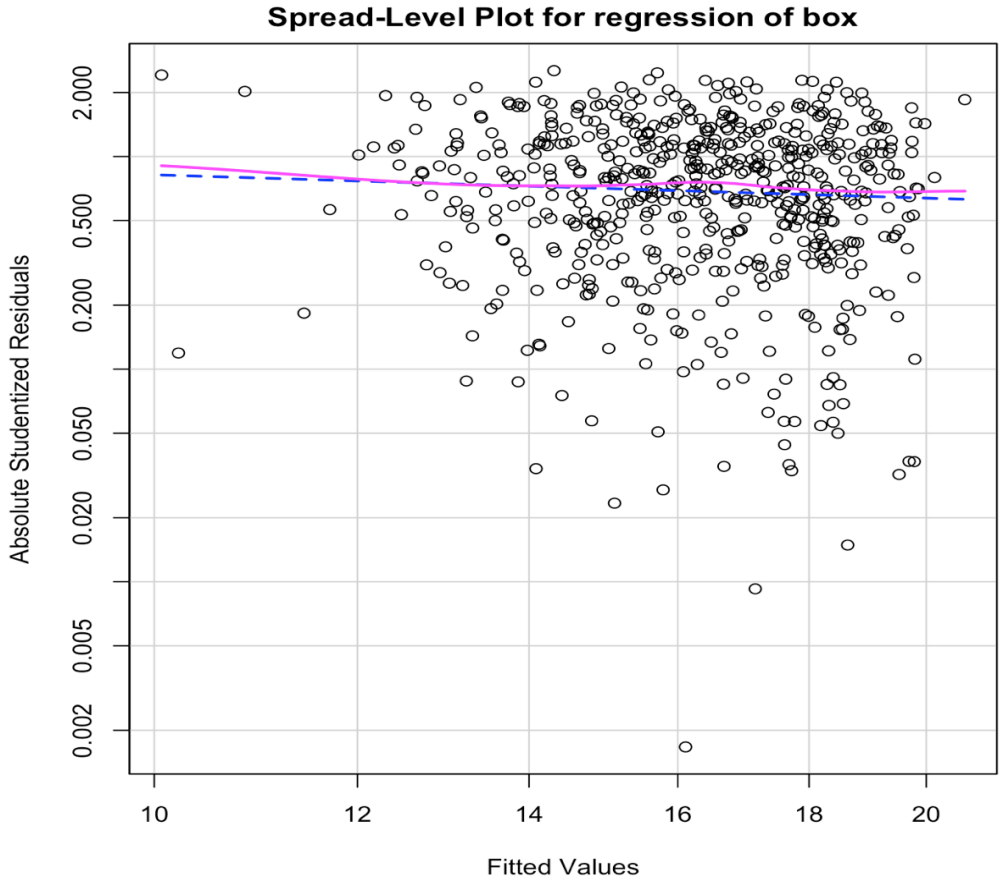
- Step

  采用AIC原则自动选择变量

- K折交叉验证，选取K=10

  根据模型的稳健性选择最优模型

模型检验

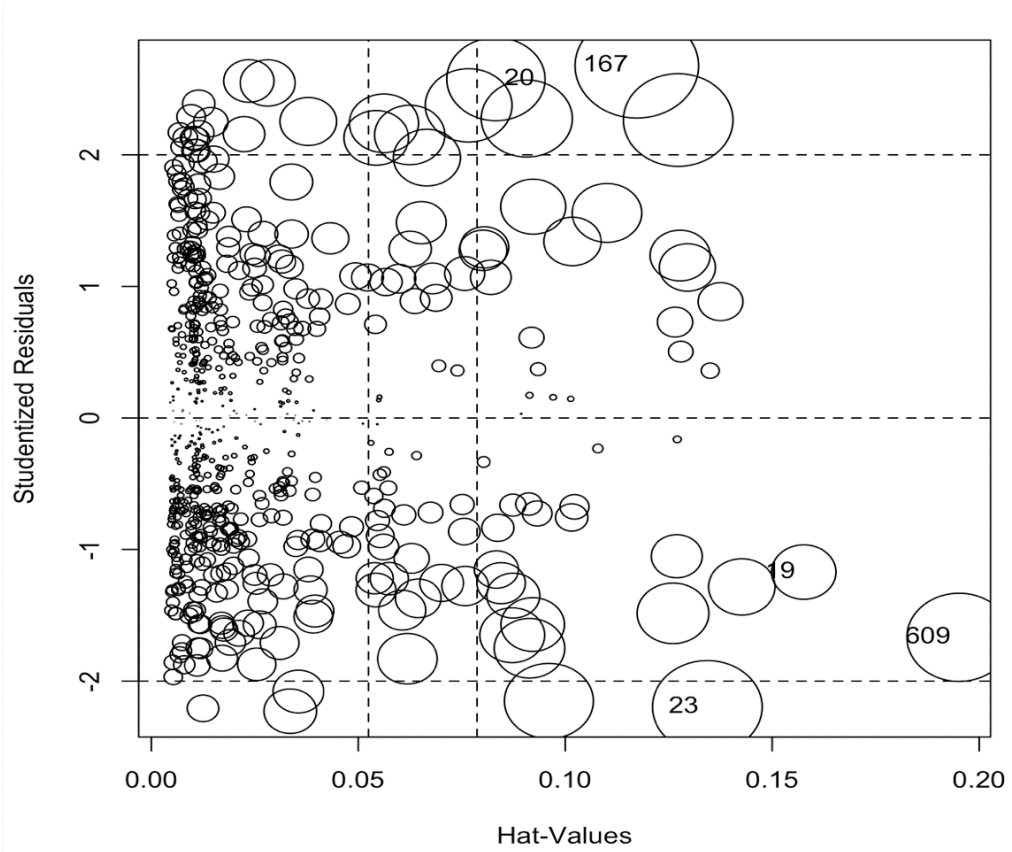


- 多重共线性　VIF
- 异方差检验

  $ncvtest\ \&\ Spread\ level\ plot$
- 强影响点检验

  $absolute\ studentized\ residual\ > 3$
  $\&\ hat\ value\ >\ 0.2$

模型检验



- 多重共线性　VIF
- 异方差检验

  $ncvtest\ \&\ Spread\ level\ plot$
- 强影响点检验

  $absolute\ studentized\ residual\ > 3$
  $\&\ hat\ value\ >\ 0.2$

以$Log(Y)$为因变量预测Y的修正

	预测模型的MAE（未修正）	预测模型的MAE（修正）
$box\ office$	12.906×10^6	13.59×10^6

- $\hat{y} = e^{\widehat{\log(y)}}$会系统地低估 y
- 伍德里奇《计量经济学导论》 当u独立于x时

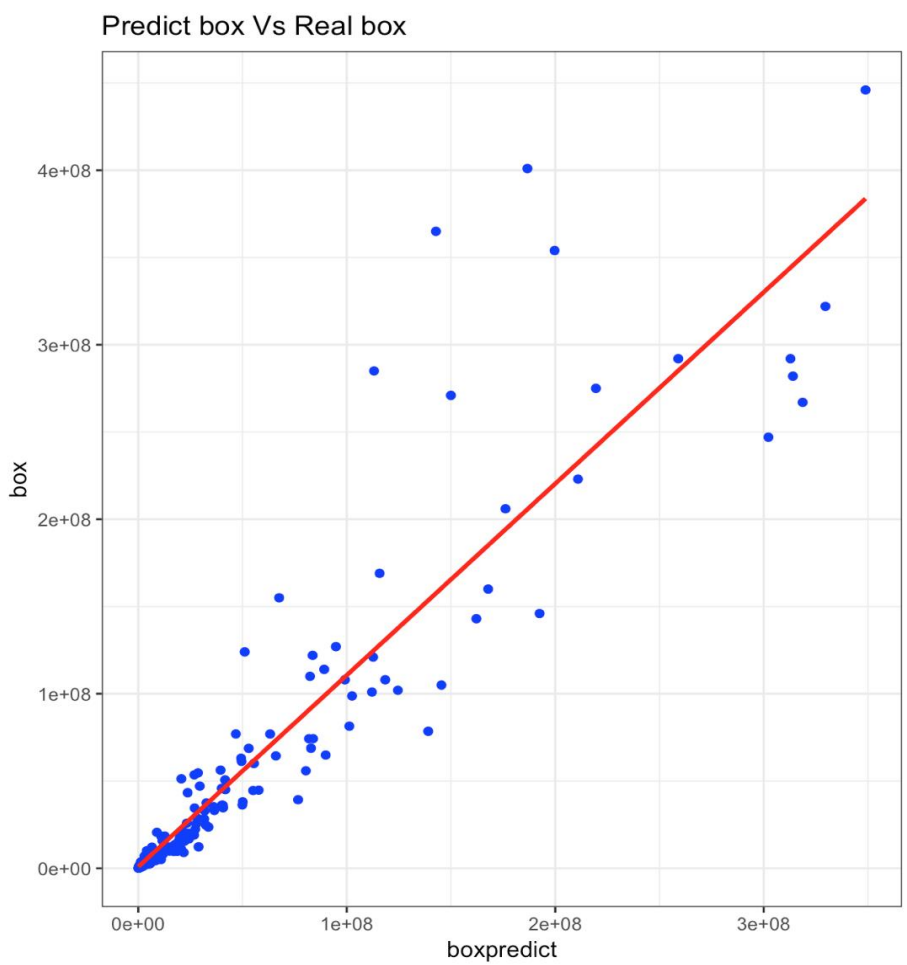
$$\log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

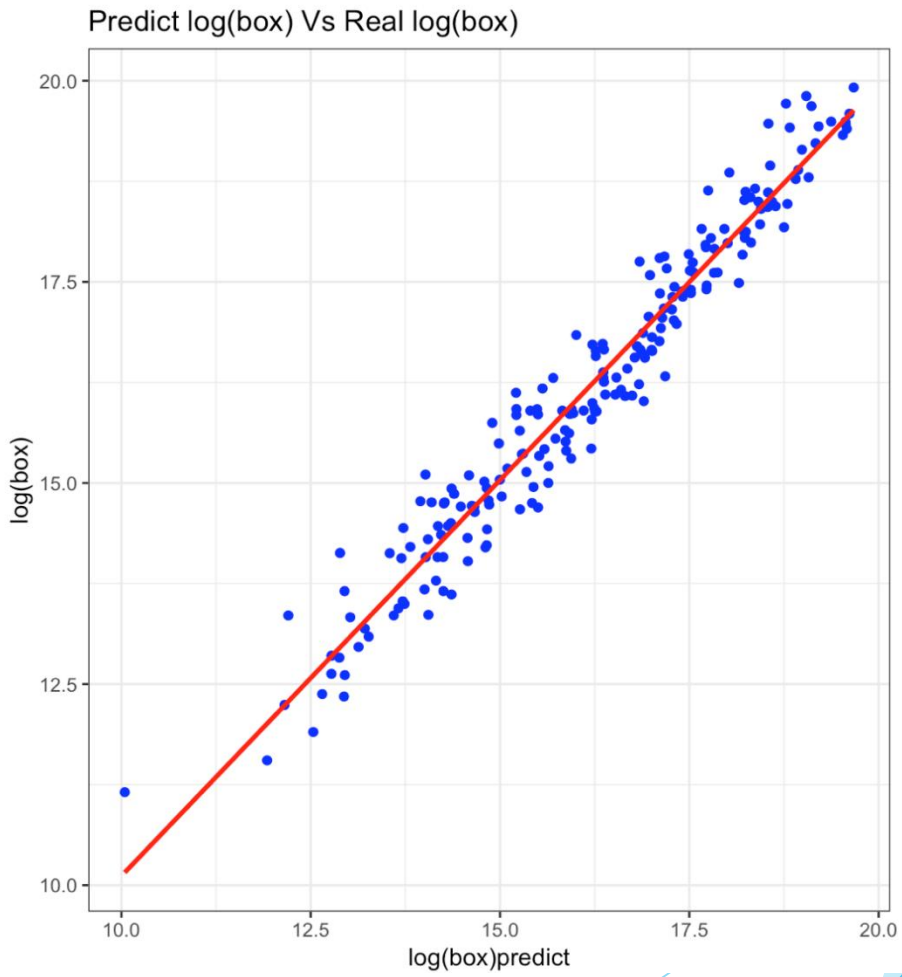
$$E(y|x) = a \times e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_n x_n}, a = e^{\epsilon}$$

- 可利用$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} exp\hat{\epsilon}_i$得到

$$\hat{y} = \hat{a} \times e^{\widehat{\log(y)}}$$

模型预测


Predict box Vs Real box
box
boxpredict
0e+00
1e+08
2e+08
3e+08
4e+08


Predict log(box) Vs Real log(box)
log(box)
log(box)predict
10.0
12.5
15.0
17.5
20.0

不足和改进

- **部分数据无法获取**
  - 排片、黄金场占比、上座率
  - 微博指数、微信指数、百度指数
  - 主观数据、营销宣传事件信息
- 时间序列
  - 使用时间序列建模，使用上映后的信息实时更新
- 相对误差
  - 回归最小化相对误差

谢谢！