# Assumptions and Justifications

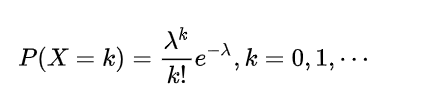
1. 上传数据大体上是真实的，符合游戏规则的
2. 数据代表的都是真实的人，他们解题的思路没有发生重大的变化

# model 1

报告的结果数量每天都有所不同。开发一个模型来解释这种变化，*并使用您的模型创建一个关于2023年3月1日报告结果数量的预测区间*。

1报告总人数和时间的关系比例

尝试1：



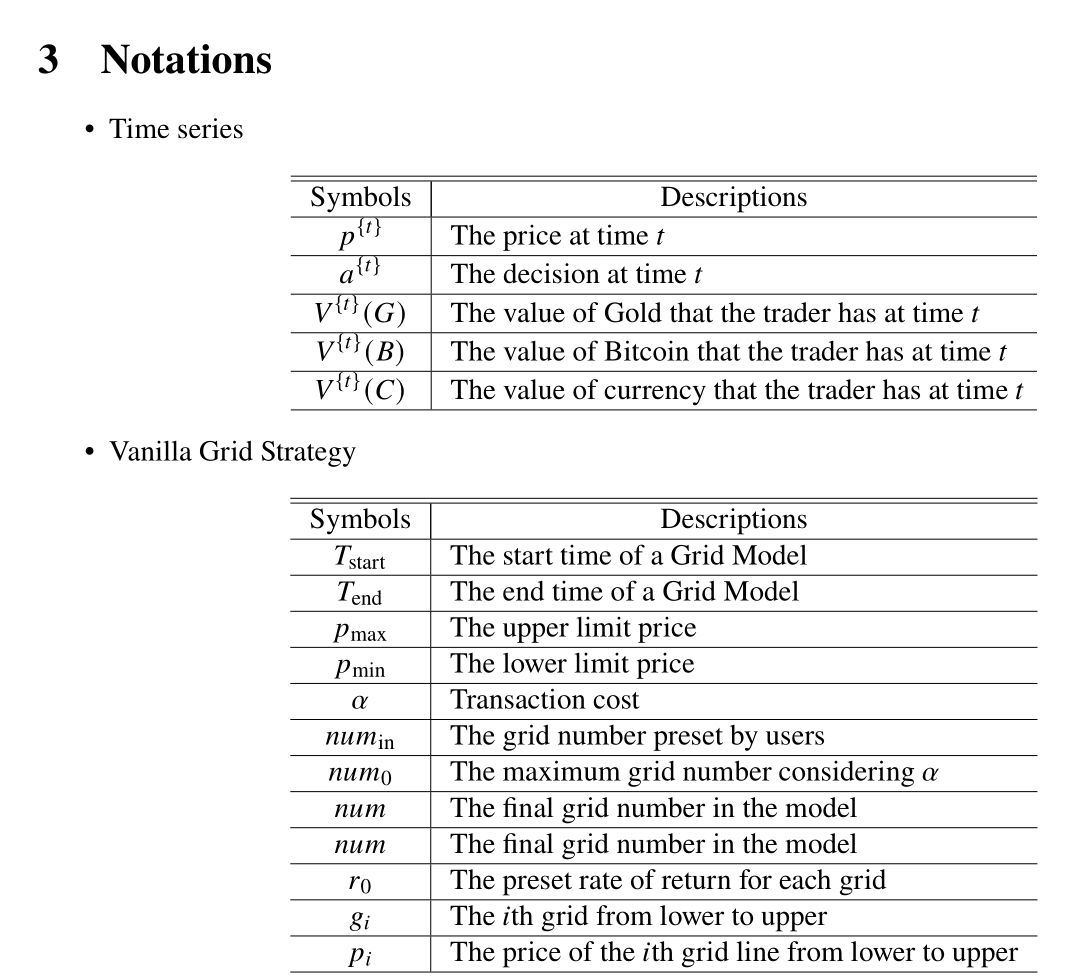
利用泊松分布模型的概率函数，创建线性回归（）

尝试1可能的问题：倍数相差太小（？2w与35w相比也有17倍）

不够拟合

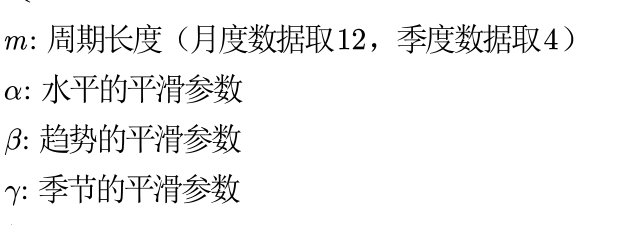
非线性回归

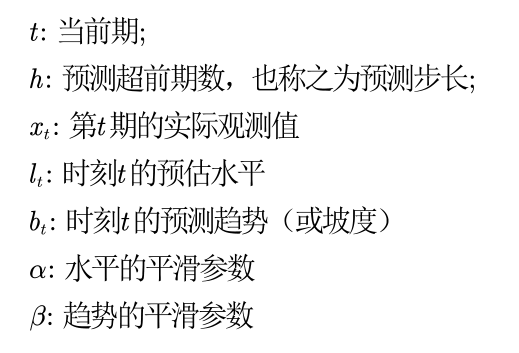
**暂时确定：**



**记号部分类似如此，分模型书写**

记号：





* 模型总览：

为预测报告结果数量的变化，我们使用回归分析和时间序列分析两大方向，并比较两者利弊。观察报告结果数量数据的可视化视图，我们使用对数线性回归模型，得到了拟合系数较高的结果，但会有过拟合的结果。分析影响报告结果数量的实际因素后，我们认为一周七天中报告数量的变化有一定的周期性规律，即存在季节性波动因素。于是我们使用Holt改良后的指数平滑模型，即温特乘法模型。计算ACF和PACF得到拟合情况，计算置信区间（写的不好）

* 回归分析（先没写）
* 温特乘法模型

para1：判断应该用时间序列预测（按实际意义）

时间序列的因素：



para2：判断模型存在季节性（季节性分解+实际意义-周）

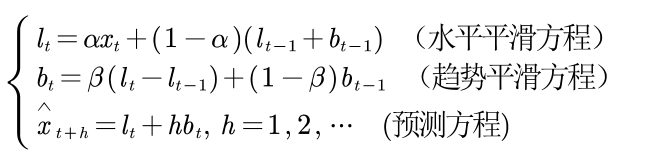


m取7（待补充）

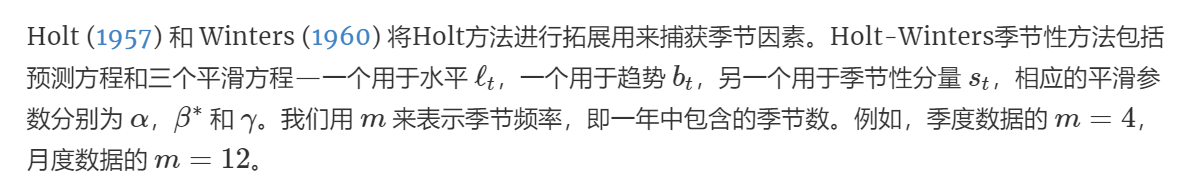
~~para2.9 holt简单季节性模型（winter模型基础）~~

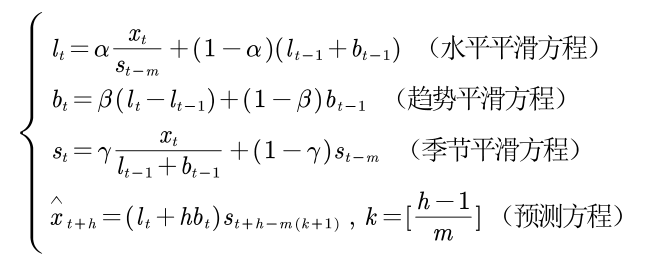
~~Holt 在1957年把简单的指数平滑模型进行了延伸，能够预测包含趋势的数据，~~

~~该方法包含一个预测方程和两个平滑方程（一个用于水平，另一个用于趋势）~~

~~~~

para3：介绍winter模型





解释：

注意到lt xt是接近预测数据的

bt相当于是lt的差分

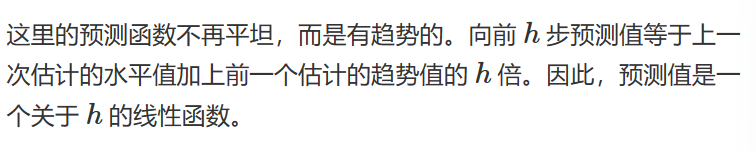
st是指数，自己与自己波动

当前时刻的预估水平由此期的实际观测值除以前一周期季节性**指数**和前一数据的【预估水平趋势和】决定。具体决定使用加权平均数，权值为水平平滑参数α，α的确定由最大化拟合系数决定（最优拟合）

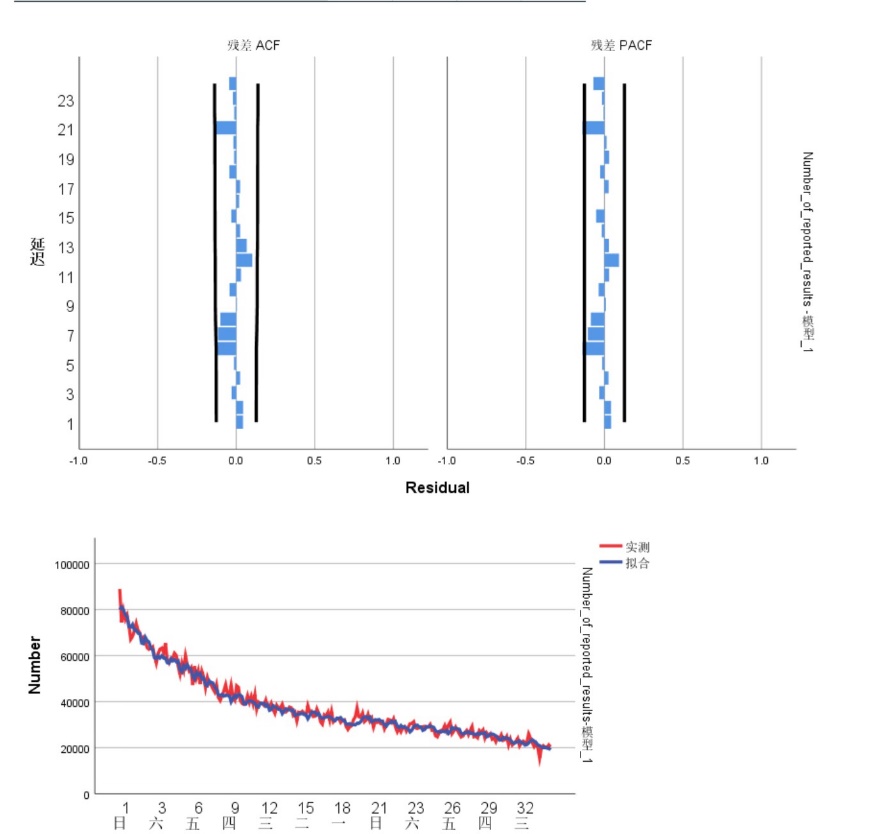
当前时刻的预估趋势由lt的差分与上一时刻的预估趋势决定。具体决定使用加权平均数，权值为趋势平滑参数β，β的确定由最大化拟合系数决定（最优拟合）

当前时刻的预估季节指数由【此期的实际观测值与【预估水平趋势和】的比】与上一时刻的预估季节指数决定。具体决定使用加权平均数，权值为季节平滑参数γ，γ的确定由最大化拟合系数决定

当前时刻的预测值为【当前时刻的预估水平趋势和】与前一期的预估季节指数的乘积，这也是为什么成为“乘法模型”



后面的具体要交给编程手



区间确定：

~~直接拿e^（kx+b）[+x]拟合（k为负数）~~

~~会过拟合~~

~~在确定预测模型后，我们根据每日的波动，确定预测区间~~

# model 2

是否有单词的属性会影响报告结果的比例？

对于未来日期的**给定解决方案单词**，开发一个模型，使您可以**预测报告结果的分布**。换句话说，预测未来日期的相关百分比（1、2、3、4、5、6、X）的分布。

*您的模型和预测有哪些不确定性？请举一个关于2023年3月1日单词EERIE的预测的具体例子。您对您模型的预测有多自信？*

*——需要我们自己寻找确定所谓的词的属性，所以一定存在某种属性影响这个比例*

注意到这里与时间无关，通过单词与报告结果分布，训练出单词预测报告分布的模型。

那我们需要对单词做出一个分析，大概是单词->特征向量->报告分布

# model 3

开发并总结一个模型，通过难度**分类解决方案单词**。**确定**与每个分类**相关**联的给定单词的**属性**。*使用您的模型，单词EERIE有多难？讨论您的分类模型的准确性。列出并描述该数据集的其他有趣特征。*

1 难是如何定义的

2 在模型2的基础上聚类分析即可