# 核心指标预测模型设计文档

# 设计目标

无线APP产品都有一些核心的KPI指标来衡量产品的发展情况,例如产品新增激活用户数、活跃用户数等。这些KPI指标的计算周期通常是天级、周级、月级，当看到这些指标时只能获得产品过去的状态。高层往往需要参考产品在未来时间的指标情况进行决策。本项目旨在对产品核心指标的时间序列数据进行建模，对指标未来走势进行预测，同时通过深入研究外部因素、指标之间的相关性提高预测模型的预测精度。

# 设计思路

根据业务需求，首先对核心KPI指标的定义进行分析。

## 指标体系

|  |  |  |
| --- | --- | --- |
| **手机百度（框，baiduboxapp）：**  定义1：包括框全平台（**除去三星**）的数据，包括**百搜pad**，不包括专版为**掌百、other**的数据  定义2：包括框android平台（**除去三星**）版本大于等于4.0的数据，不包括专版为**百搜pad**，**掌百、other**的数据  定义3：分别包括**全平台、android平台、ios平台**，渠道类型分别为**预装、后装**，专版包**括百搜、百搜pad**（二期需求） | | |
| **指标名称** | **精确定义** | **粒度** |
| 新增用户数  activate\_count | 首次安装使用产品的用户数（cuid去重加总） | 天级 |
| 活跃用户数  active\_count | 用户打开使用产品的用户数（cuid去重加总），按照不同的去重周期具体分为，日/周/月活跃用户数，本次仅考虑天级 | 天级 |
| 累计激活用户数  accumulate\_count | 全历史累计新增用户数，等于所有新增用户数的加总 | 天级 |
| 定义1数据接口：  select event\_day,sum(new\_uv),sum(active\_uv),sum(total\_uv) from insight\_searchbox\_alluser\_ed2 where (product="baiduboxapp" or product="baiduboxpad") and event\_day="${hivevar:date}" group by event\_day;  定义2数据接口：  select event\_day,sum(new\_uv),sum(active\_uv),sum(total\_uv) from insight\_searchbox\_alluser\_ed2 where product="baiduboxapp" and event\_day="${hivevar:date}" and os="android" and softv>="4.0" and softv<="9.9" group by event\_day;  定义3数据接口：  select event\_day,sum(new\_uv),sum(active\_uv),sum(total\_uv) from insight\_searchbox\_alluser\_ed2 where product="baiduboxapp" and event\_day="${hivevar:date}" and os="android" and softv>="4.0" and softv<="9.9" group by event\_day; | | |

|  |  |  |
| --- | --- | --- |
| **手机浏览器（浏览器，baidubrowser）：**  包括浏览器全平台的数据，包括专版为**浏览器、OEM预装**的数据，不包括专版为**新掌百、hao123、other**的数据 | | |
| **指标名称** | **精确定义** | **粒度** |
| 新增用户数 | 首次安装使用产品的用户数（cuid去重加总） | 天级 |
| 活跃用户数 | 用户打开使用产品的用户数（cuid去重加总），按照不同的去重周期具体分为，日/周/月活跃用户数，本次仅考虑天级 | 天级 |
| 累计激活用户数 | 全历史累计新增用户数，等于所有新增用户数的加总，由于浏览器累计活跃缺失，使用**浏览器全平台去重**的指标。 | 天级 |
| 数据获取接口：  风险：与DT平台展现有微小diff，由于底层数据切分不准  select event\_day, sum(new), sum(active), sum(acc) from (  select event\_day, sum(newuser) as new, sum(activeuser) as active, sum(0) as acc from insight\_browser\_dayuv\_bridge where (specialversion="llq" or specialversion="bdoem") and event\_day="{DATE}" group by event\_day  union all  select event\_day, sum(0) as new, sum(0) as active, sum(totaluser) as acc from insight\_browser\_dayuv\_cube where event\_day="{DATE}" and specialversion="a" group by event\_day) t\_1  group by event\_day; | | |

|  |  |  |
| --- | --- | --- |
| **手机输入法（输入法，baiduinput）：**  包括输入法全平台的数据，不包括**内核OEM**的数据 | | |
| **指标名称** | **精确定义** | **粒度** |
| 新增用户数 | 首次安装使用产品的用户数（cuid去重加总） | 天级 |
| 活跃用户数 | 用户打开使用产品的用户数（cuid去重加总），按照不同的去重周期具体分为，日/周/月活跃用户数，本次仅考虑天级 | 天级 |
| 活跃老用户数 | 去除活跃用户数中当天激活用户之外的活跃用户数 | 天级 |
| 累计激活用户数 | 全历史累计新增用户数，等于所有新增用户数的加总 | 天级 |
| 数据获取接口：  风险：通过平台号过滤oem内核，如果有相应变更需要修改数据接口  SELECT event\_day, sum(new\_user\_num), sum(active\_user\_withnew\_num), sum(active\_user\_withoutnew\_num), sum(cum\_user\_num\_cumulate\_by\_new) from insight\_input\_daily\_info\_new\_v2 where not array\_contains(Array("a2", "a3", "a4", "a5", "a7", "a8", "p-a1-2-11", "p-a1-2-44", "p-a1-2-46", "p-a1-2-47", "p-a1-2-52", "p-a1-2-53", "p-a1-2-58", "p-a1-2-59", "p-a1-2-60", "p-t1-2-48", "p-t1-2-49", "p-w1-2-45", "p-a1-3-77" ), platform) and event\_day="{DATE}" group by event\_day | | |

|  |  |  |
| --- | --- | --- |
| **百度手机助手（手机助手，baiduappsearch）：**  仅包括百度手机助手**移动客户端**的数据 | | |
| **指标名称** | **精确定义** | **粒度** |
| 新增用户数 | 首次安装使用产品的用户数（cuid去重加总） | 天级 |
| 活跃用户数 | 用户打开使用产品的用户数（cuid去重加总），按照不同的去重周期具体分为，日/周/月活跃用户数，本次仅考虑天级 | 天级 |
| 累计激活用户数 | 全历史累计新增用户数，目前缺乏该指标历史数据 | 天级 |
| 数据获取接口：  风险：历史数据中20130419、20130420数据缺失  select event\_day,sum(case when cor\_id="notspam" then new\_user\_count else ceil(0) end),sum(activeuser),sum(user\_acc) **from** insight\_appsearch\_uv\_daily\_v2 where event\_day="{DATE}" AND platform="appclient" group by event\_day; | | |

## 指标预测模型

**基础模型**

指标序列的数据具有类似于经济指数走势的特点，主要包含趋势项、周期项、随机项，其中不同类型指标的周期项存在差异需要分别考虑，而随机项又可以拆分为非定期因素与白噪声两部分，通过对外部影响的建模来进行非定期因素的拟合。

其中，表示在日期t的指标值，trend和season分别表示趋势项和周期项，表示外部影响项，表示白噪声项。

另外，产品的核心指标往往是多个维度的汇总值，而各维度粒度下的指标值在汇总时肯能会将典型的序列模式平滑掉，因此考虑尝试在不同维度值（例如：线下渠道、线上渠道）分别进行模型拟合，最后按加法模式进行组合。

**新增激活用户数（天级）**

新增激活用户数是指每天新增开始使用的用户数，用户的数量主要受到渠道推广力度的影响影响，推广力度是抽象聚合的概念，很难量化，因此可以将推广力度转化为几个更容易量化的指标：

周末标识h：周末用户有更多的闲暇时间，更有可能从线上、线下渠道被接入App

节日标识f：类似周末标识，但线下渠道可能会有更显著的增长

节前活动标识p：例如在春节、情人节、圣诞节前，手机的销量可能会有较大增长，导致线下渠道的增长。

同时考虑到线上、线下渠道间的差异，初步考虑渠道类型维度对新增激活用户数进行拆分，新增激活用户数的预测模型如下：（后期根据实际数据进行模型优化）

**活跃用户数（天级）**

活跃用户数是指每天主动打开App使用并产生服务端可感知行为的用户的数量，活跃用户数的影响因素较为复杂，部分是用户习惯使用，部分是由于一些热点新闻、事件偶然使用，部分是由于消息push引导使用。

月初/末效应：由于用户流量分布的特点，月初的活跃度往往高于月末的活跃度。

新版发布：新版发布后活跃用户数会有较大的变化。

消息push：每天消息推送的内容、次数、人群会引起活跃用户的较大变化。

**累计新增用户数（天级）**

累计新增用户数是指产品自发布之日起所有新增激活用户数的加总值。日期t的累计用户数等于日期t-1的累计用户数加日期t的新增用户数，因此累计新增用户数的预测通过上一期累计用户数加新增用户数预测值实现。

**新增用户数（月级）**

月级新增用户数是指对一个月内每天新增用户数的直接加总值。首先可以使用天级数据预测未来每天的指标，再将指标按月加总求和。这种方法只适合较短期预测，因此只能实现最近1~2个月的预测，更长时间预测会由于误差累计导致误差迅速增加。对于较长期的预测（未来12个月），考虑使用每个月的新增用户均值直接预测未来每个月的指标。

**活跃用户数（月级）**

同4.新增用户数（月级）

**累计新增用户数（月级）**

方法类似于4.累计新增用户数（天级），月级累计新增用户数取月末最后一天的累计用户数。

## 指标预测方法设计

**月粒度指标预测方法**

需求：利用核心指标时间序列数据预测未来12个月的指标序列，主要对指标变化的趋势性、周期性进行建模，尽可能降低平均误差率。

解决方案1：按照半月粒度对时序数据进行汇总，使用ARIMA模型对历史数据进行拟合，由于历史数据较少（不足20个月），难以拟合年度周期性，最终模型确定为order=(1,1,0)。

## 数据流设计

数据流通过串联独立的可配置job实现，不同类型的job通过读取对应的配置文件执行数据流传输、计算。数据流框架如下图所示。@ADD项目实例

主要包括3类输入数据：Row Data为产品的用户基础数据，存储在ＵＤＷ系统中，**Outer Data**为外部产品输入数据，存储方式不确定，DB Data主要存储计算用到的一些中间数据，存储在mysql系统中。

主要包括5类作业：HQL Job通过读取HQL配置从原始数据中计算出产品基础指标；SQL Job主要负责将基础指标数据同步到DB中；**Grab** Job主要负责将外部数据同步到DB中；**Preprocss** Job主要通过对DB中存储的中间数据进行数据预处理，生成**预测的输入数据集**；R Job主要利用读取训练数据以及预测脚本进行计算，输出预测结果。



## 数据流模块详细设计

**1 基础类设计图**



**2 作业执行流程**



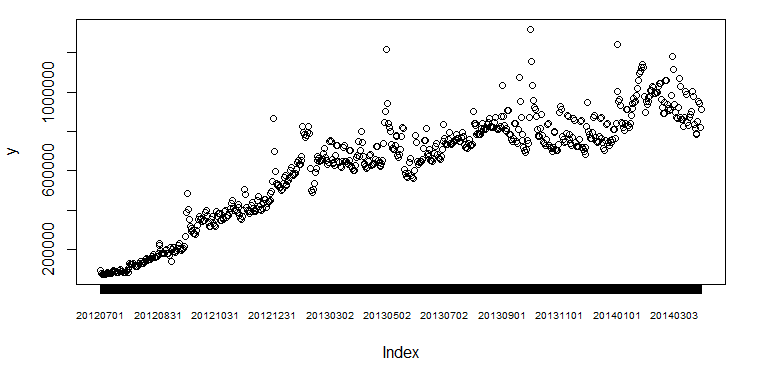
**3 作业流配置说明**



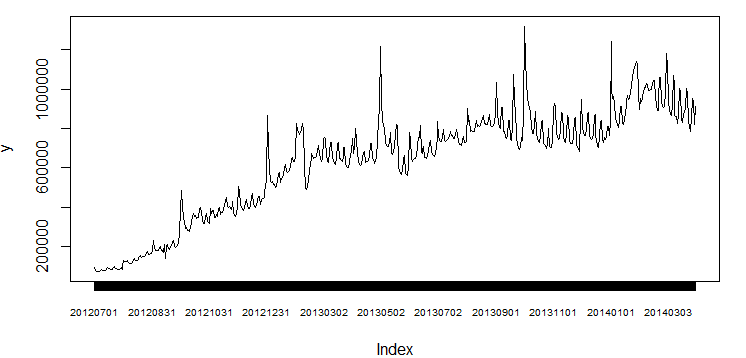
## 时间序列数据分析

### 百度框激活数据

### 1、首先加载原始时间序列数据（2012-7-1~2014-4-1），共640个数据点

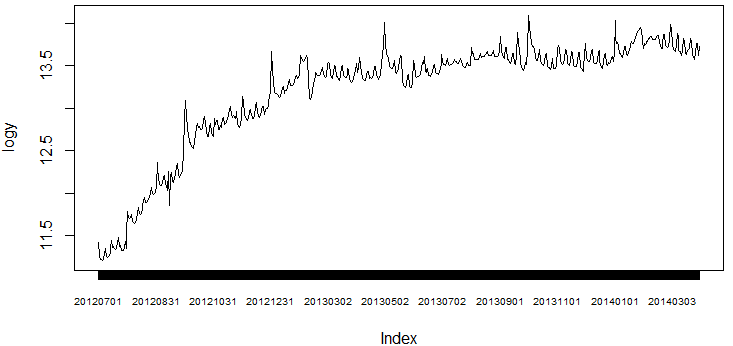


原始点序列显得非常杂乱无章，毫无规律，下面开始对这一序列进行深入分析，挖掘出其中包含的规律，建立回归模型，利用历史数据进行模型参数拟合，最后实现对未来数据的预测。



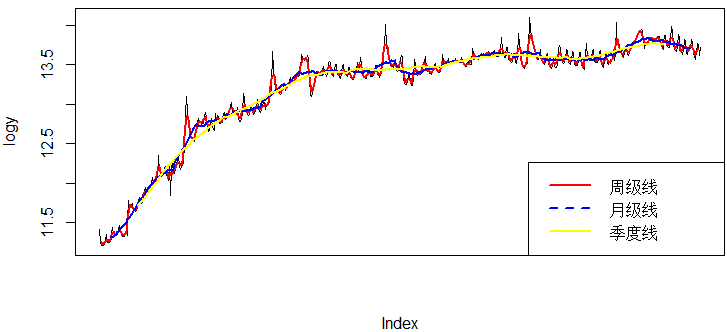
曲线图显示，模糊感觉蕴含一些变化规律

### 2、对原始数据进行log处理，消除波动项幅度的变化



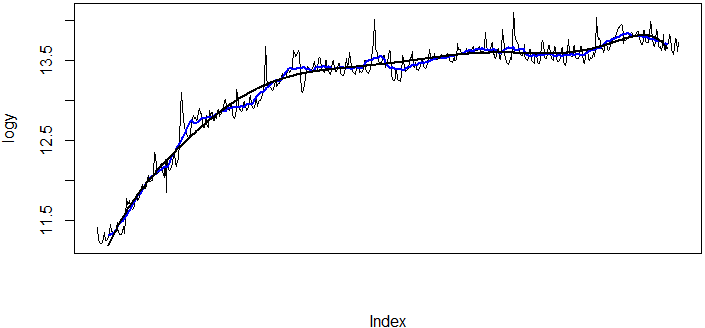
经过log处理，异常数据点的波动被显著平衡化处理

### 3、时序数据的趋势项提取（移动平均法）

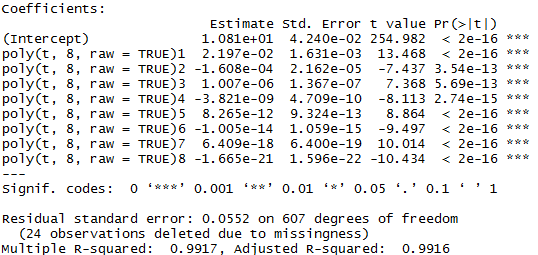


三条趋势线不同程度描述了时序数据的变化趋势，其中周级线与原始数据拟合最好，但同时带有较强的波动项，季度线较容易用低阶函数进行拟合，但忽略了过多的波动信息，因此折衷使用月级跨度表示趋势项。

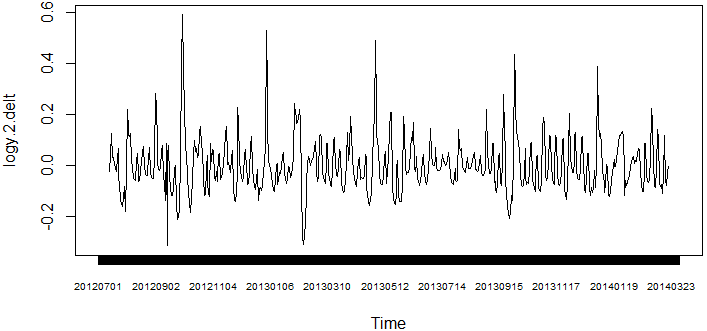
### 4、时序数据趋势项拟合，尝试多项式函数进行拟合（degree=8）



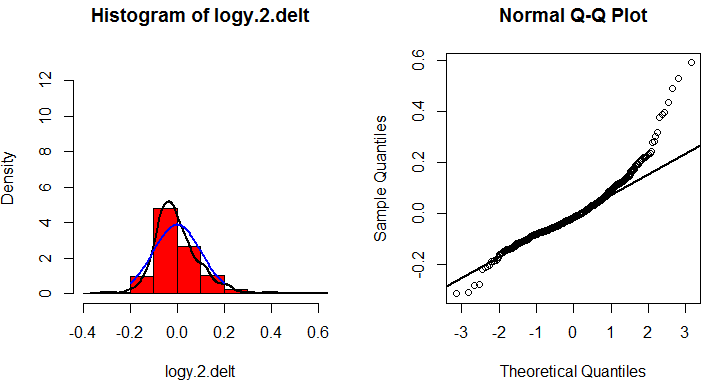
使用高阶多项式函数对月级跨度趋势项进行拟合，R-square与T-value都得到较好的效果：



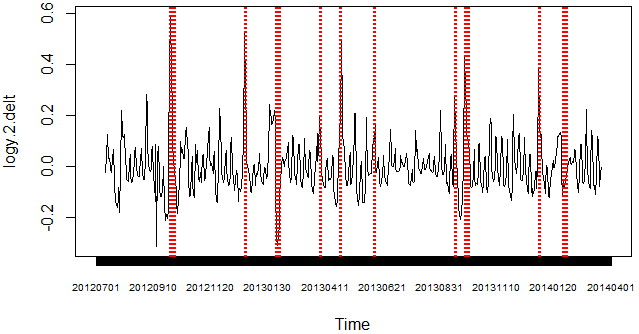
### 5、去除趋势项的时序数据分析



趋势原始时序数据中的趋势项成份，剩余部分得到波动较为随机的序列，这时可以通过检验时序数据的平稳性，直接考虑ARIMA模型对数据进行建模，这里考虑继续深入挖掘时序数据中包含的规律。

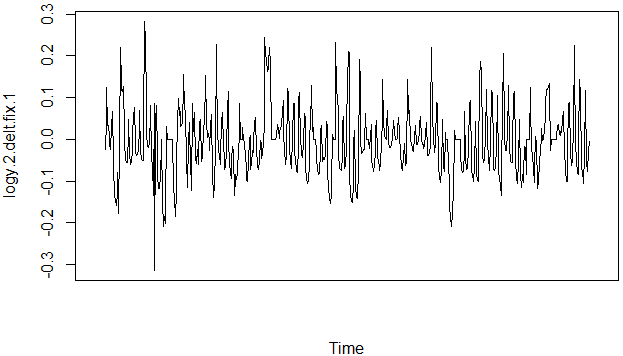


### 6、去除趋势项的时序数据分析



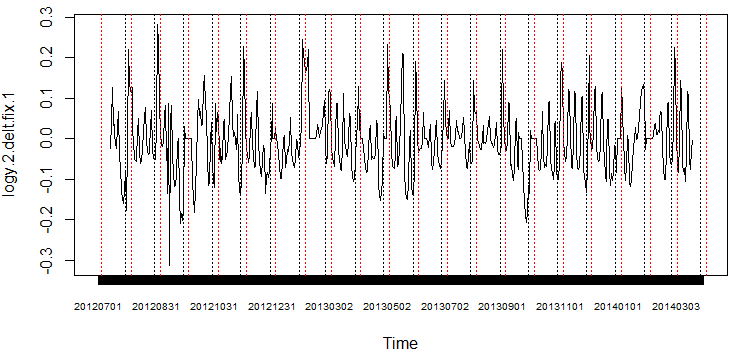
可以明显看出比较大的波动项都是在节假日时间产生的，同时十一、五一、元旦这类大长假数据产生脉冲式激增，清明、端午、中秋这类小长假上涨幅度较小，春节期间会有较大幅度的下降。

**7、去除节假日项的时序数据**



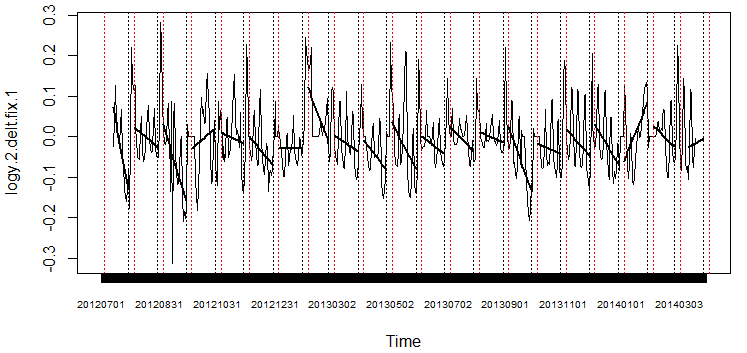
消除节假日的时序数据后，可以看到数据波动范围显著缩小为[-0.3,+0.3]。

**8、月初、月末效应检查**

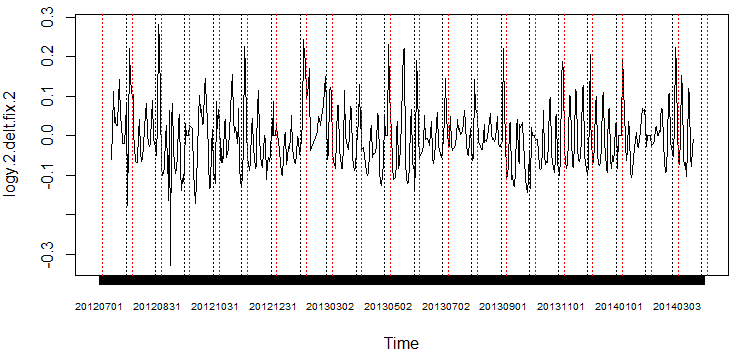


上图中每条黑线与红线部分之间的部分为月初月末交接处，范围选取为[-2,3]，月末会有较大幅度下降，月初会有较大幅度上涨。

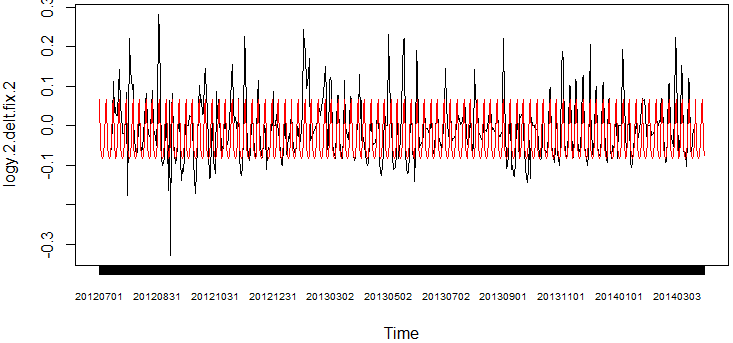
**8、月度数据趋势分析**

****

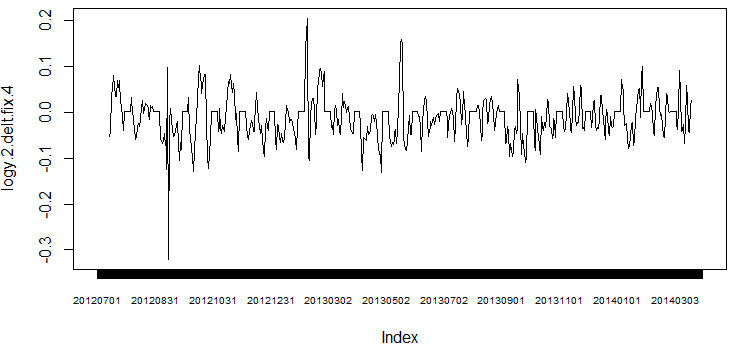
对月度数据进行线性函数拟合，除个别月份（春节），大部分月份的数据具有向下的趋势，说明每个月的时序具有下降的线性趋势，同时在月初有较大的增幅，在月末有较大的降幅。按照这一思路，在每个月内，以月中为中点对数据进行线性修正。



对修正后的数据进行frequency=7进行周期项拟合

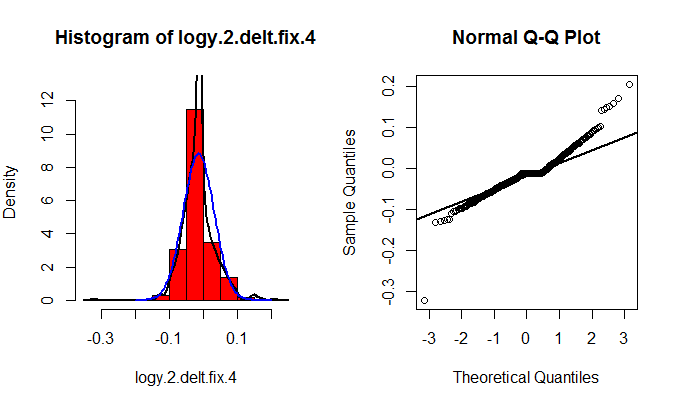


**9、消除月初月末影响数据**



消除月初月末的较大波动数据，数据整体波动基本降为[-0.1,0.1]之间

**10、随机项检验**



至此，时序数据的波动范围已经缩减到可接受的范围。

**10、时间序列数据建模**

对时序不同构成进行建模，使用加法模型构造时序数据的完整模型。

趋势项：

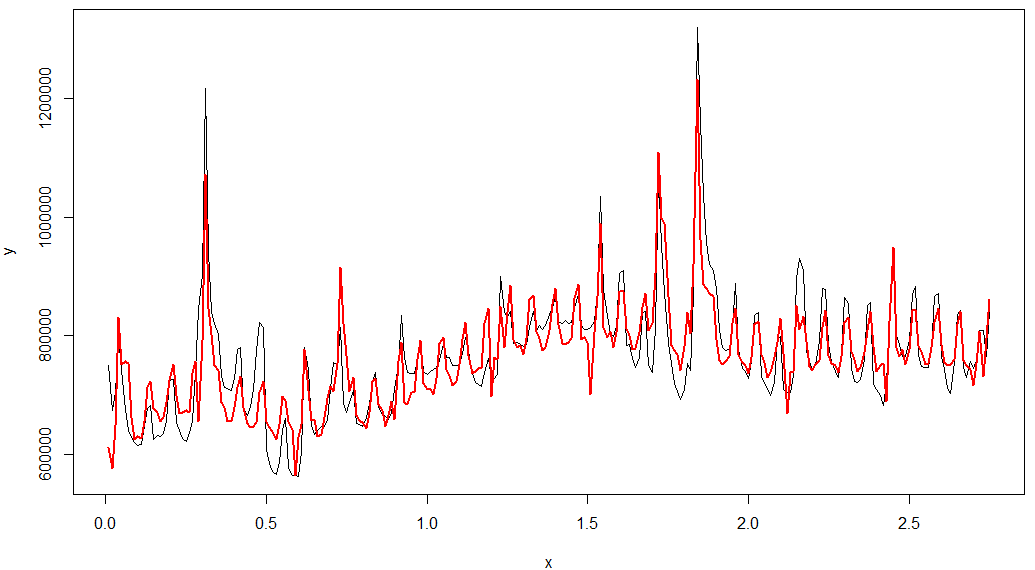
月级波动项：

周级波动项：

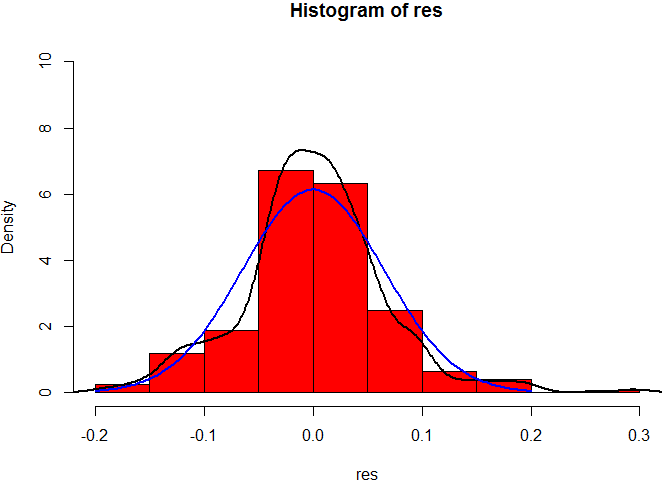
节日波动项：

整体模型：

模型拟合：

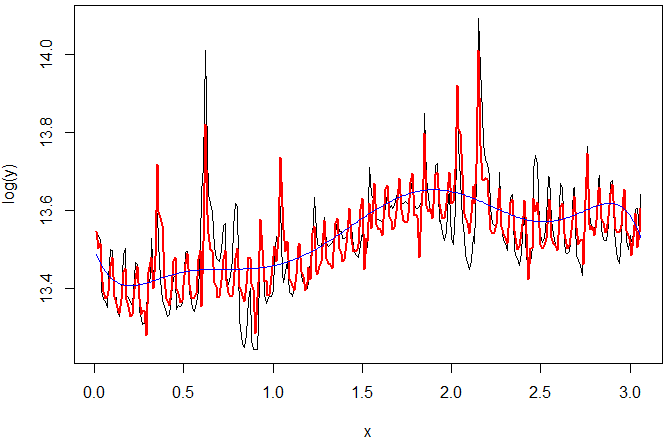


直接进行参数估计的结果与原始曲线非常接近,残差标准差为50000，占原始指标比例5.5%。平均绝对误差为36000，占原始指标比例为4%，



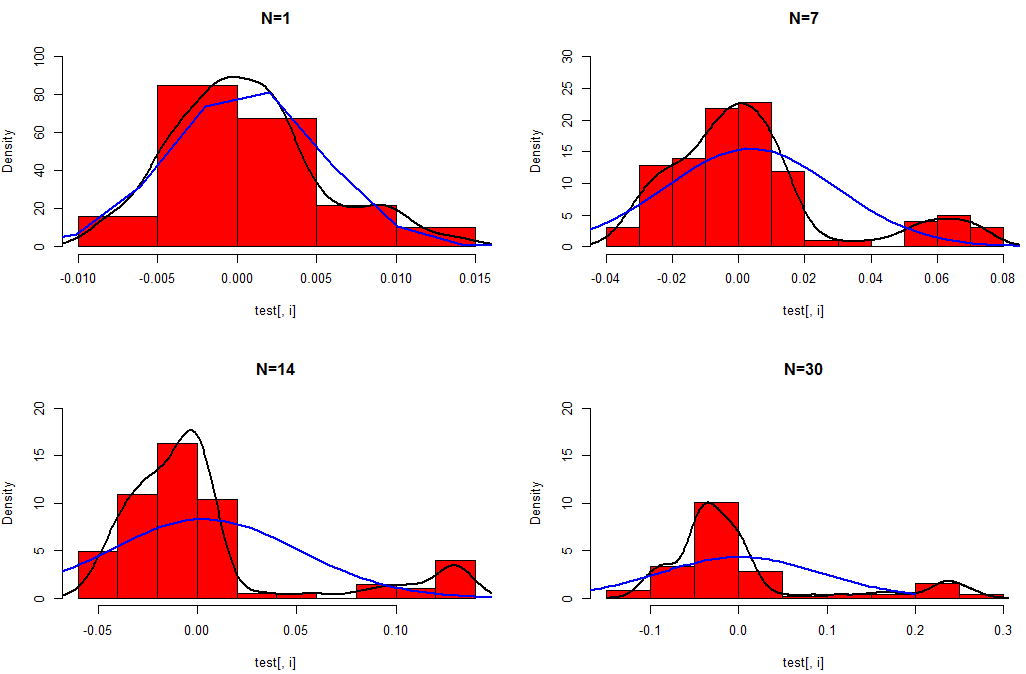
**10、从回归到预测**

如果用我们的模型来预测未来的数据，效果会如何呢？用最小二乘法拟合的模型参数仅仅是在训练集上使残差的均方差最小化的最优解，而在未知的数据集上未必会有这样的效果。



由于我们用polynominal拟合趋势项，很明显可以看到在趋势项的最后端，是大幅度下降的趋势，很快会到达零（按照经验这是不太可能的），因此需要牺牲一部分训练拟合度，提升预测拟合度。

对趋势项的长期预测（预测T+N）有两种主要思路：迭代预测与直接预测。其中迭代预测思路比较容易理解，但是难以解决误差累积的问题，包括自回归，指数平滑，移动平均等；直接预测方法包括KNN，神经网络等。由于本项目的时序数据较少，且项目需求重点关注较为近期的结果，所以采用差分项的移动平均法进行T+1~T+N项的预测。



对预测长度为（1,7,14,30）的预测误差序列进行分析，发现N=30时，只有68%的置信度绝对误差率小于7.3%。

## 时间序列预测结果

|  |  |  |
| --- | --- | --- |
| 预测指标名称 | 30日测试平均误差率 | 30日预测平均误差率 |
| 全平台激活数 | 0.042649141 | 0.081824088 |
| Android激活数 | 0.044284971 | 0.084698149 |
| Iphone激活数 | 0.050759146 | 0.120847081 |
| 全平台活跃数 | 0.020376265 | 0.055076126 |
| Android活跃数 | 0.023044165 | 0.04193166 |
| Iphone活跃数 | 0.024507963 | 0.074744549 |

|  |  |  |  |
| --- | --- | --- | --- |
| 预测指标名称 | 7日平均误差率 | 14日平均误差率 | 30日平均误差率 |
| 全平台激活数 | 0.045687 | 0.049257 | 0.081824088 |
| Android激活数 | 0.050223 | 0.053431 | 0.084698149 |
| Iphone激活数 | 0.055233 | 0.080431 | 0.120847081 |
| 全平台活跃数 | 0.030505 | 0.036197 | 0.055076126 |
| Android活跃数 | 0.028012 | 0.029166 | 0.04193166 |
| Iphone活跃数 | 0.035383 | 0.048754 | 0.074744549 |

