数据分析与处理技术

时间序列预测

简单预测模型

综合过去所有数据，利用均值做平均是常用的一种预测方法，但也有它明显的局限性

```
> mean(beer2)
[1] 433.5135
> meanf(beer2,1)
        Point Forecast    Lo 80    Hi 80   Lo 95    Hi 95
2010 Q3       433.5135 377.2457 489.7813 346.801 520.2261
```

naive方法则简单用最末值做预测

```
> naive(beer2,1)
```

```
> rwf(beer2,1)
```

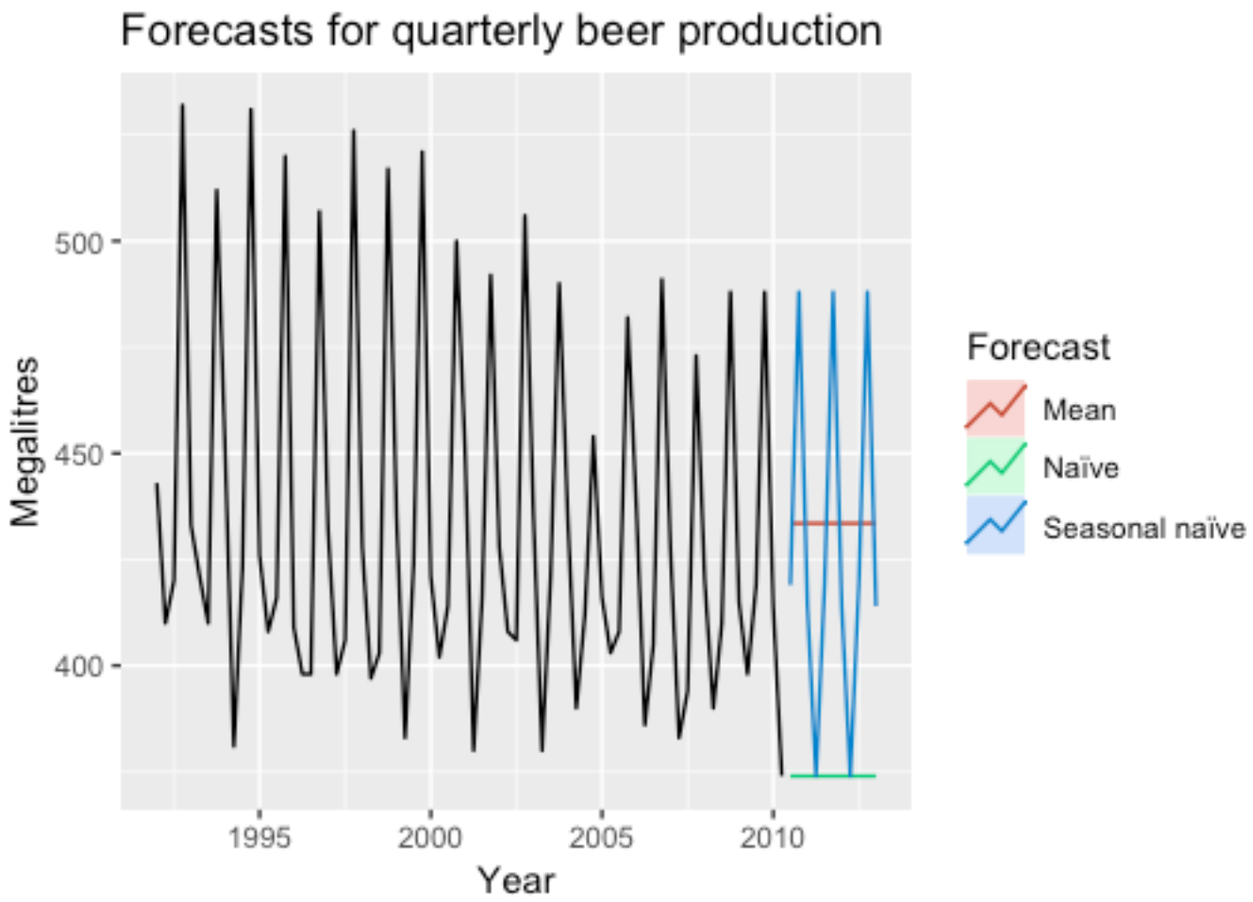
naive方法也叫做random walk forecast

以季节变动为基础的naive方法

```
> snaive(y,1)
```

加入趋势漂移的naive预测

```
> rwf(y,1,drift = T)
```



练习：尝试对fpp2中goog200序列用meanf,naive和趋势漂移的naive方法做预测并做图

```
autoplot(beer2) +
    autolayer(meanf(beer2, h=11),
              series="Mean", PI=FALSE) +
    autolayer(naive(beer2, h=11),
              series="Naïve", PI=FALSE) +
    autolayer(snaive(beer2, h=11),
              series="Seasonal naïve", PI=FALSE) +
    ggtitle("Forecasts for quarterly beer production") +
    xlab("Year") + ylab("Megalitres") +
    guides(colour=guide_legend(title="Forecast"))
```

autoplot只能有一个，下一个自适应图层需要变成autolayer

时间项回归

趋势与季节性是时间序列要考虑的首要特征，线性回归可以时间项作为自变量做回归预测，如 $y_t = \beta_0 + \beta_1 t + \varepsilon_t,$

在趋势项基础上加入按日期周期型出现的季节调整项如下

$$y_t = \beta_0 + \beta_1 t + \beta_2 d_{2,t} + \beta_3 d_{3,t} + \beta_4 d_{4,t} + \varepsilon_t,$$

在周期非常明显时，人为做出一系列周期变量数据进入回归建模

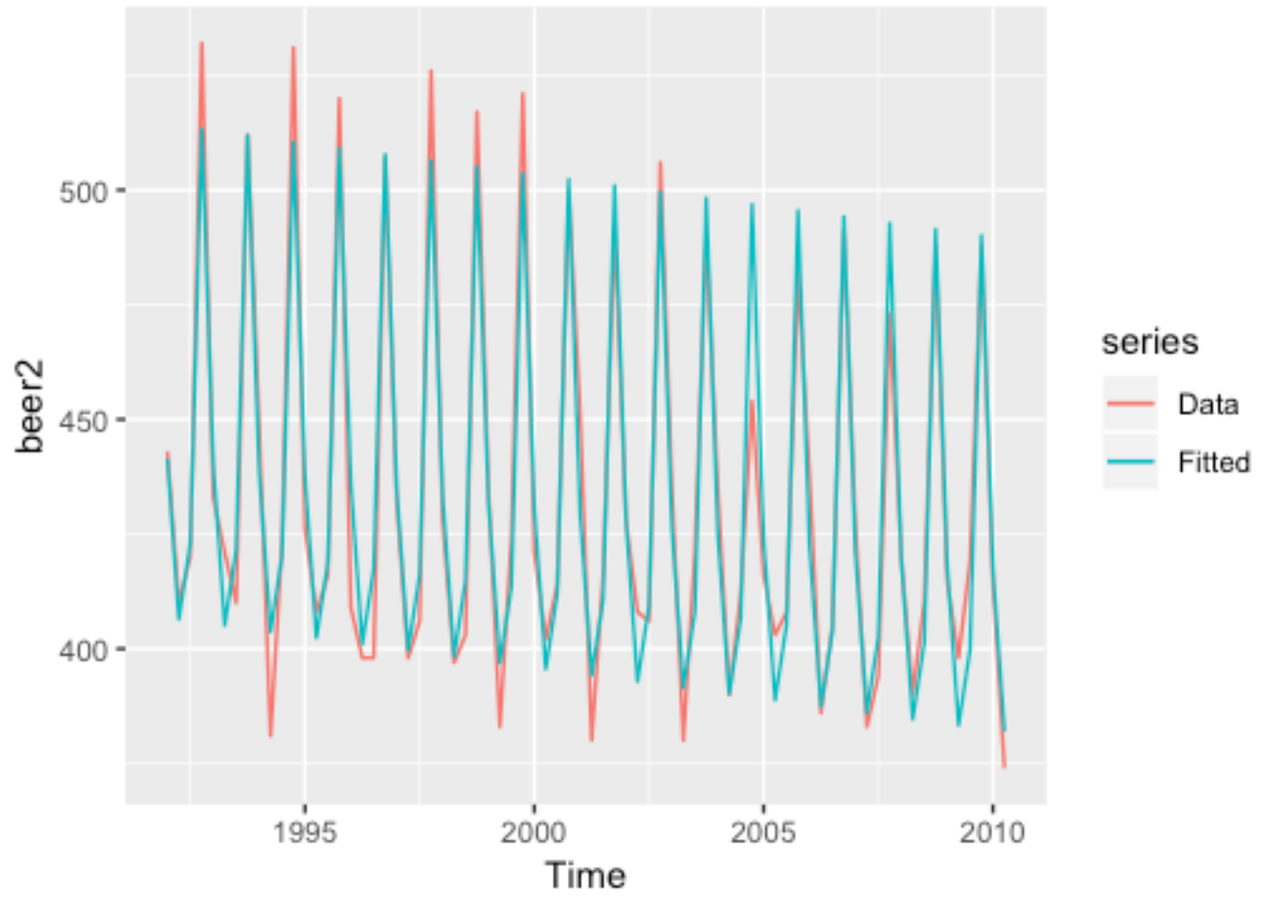
	$d_{1,t}$	$d_{2,t}$	$d_{3,t}$	$d_{4,t}$	$d_{5,t}$	$d_{6,t}$
Monday	1	0	0	0	0	0
Tuesday	0	1	0	0	0	0
Wednesday	0	0	1	0	0	0
Thursday	0	0	0	1	0	0
Friday	0	0	0	0	1	0
Saturday	0	0	0	0	0	1
Sunday	0	0	0	0	0	0
Monday	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

线性回归+周期项的方法在forecast包中有对应工具，省去了手动设置周期数据的麻烦

```
fit=tslm(beer2~trend+season)
```

做出图形对比

```
autoplot(beer2, series="Data") +
    autolayer(fit$fitted.values, series="Fitted")
```



时间序列模型

指数平滑法

移动平均实际将所有参与平滑的数据当作相等作用看待，而naive方法则认为最新的数据会最接近未来预测值，结合两者想法另最末的数据权重高，越远的数据权重越低，做出一种变权平均的效果。

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots$$

```
oildata=window(oil,start=1996)
```

截取oil数据集(年度石油产量数据)1996年后的部分

```
fc=ses(oildata,h=5,alpha = 0.3)
```

```
autoplot(fc)
```

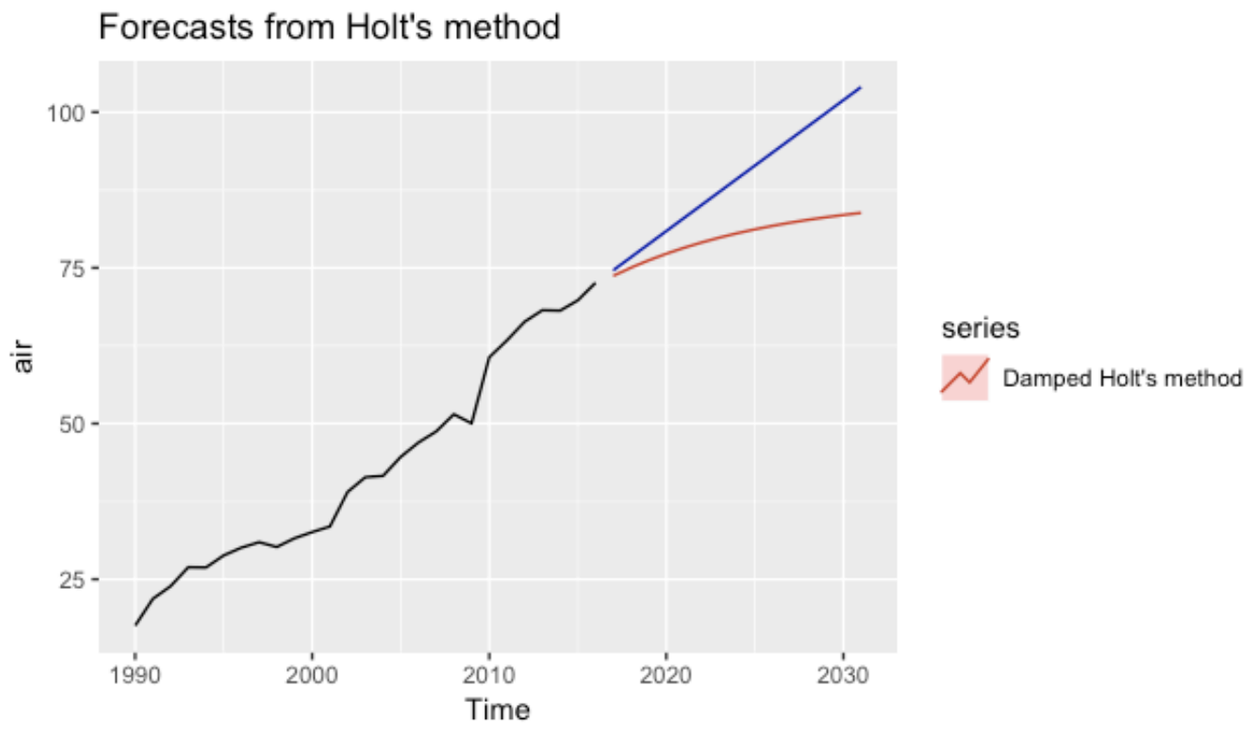
指数平滑法适用于趋势并不太明显的

带趋势的指数平滑法——Holt's 线性趋势法

指数平滑基础上改进的线性趋势法解决带有明显增长趋势的问题，公式如下

$$\hat{y}_{t+h|t} = \ell_t + hb_t$$
$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$

```
air=window(ausair,start=1990)
fc=holt(air,h=5)
```



但holt线性趋势会无限制增长，这不符合常识，任何增长都会遇到瓶颈，然后逐步放缓。阻滞线性趋势模型在holt模型基础上对预测加入了放缓增长因素

```
fc1<- holt(air, h=15)
fc2<- holt(air,damped = T,phi=0.9, h=15)
```

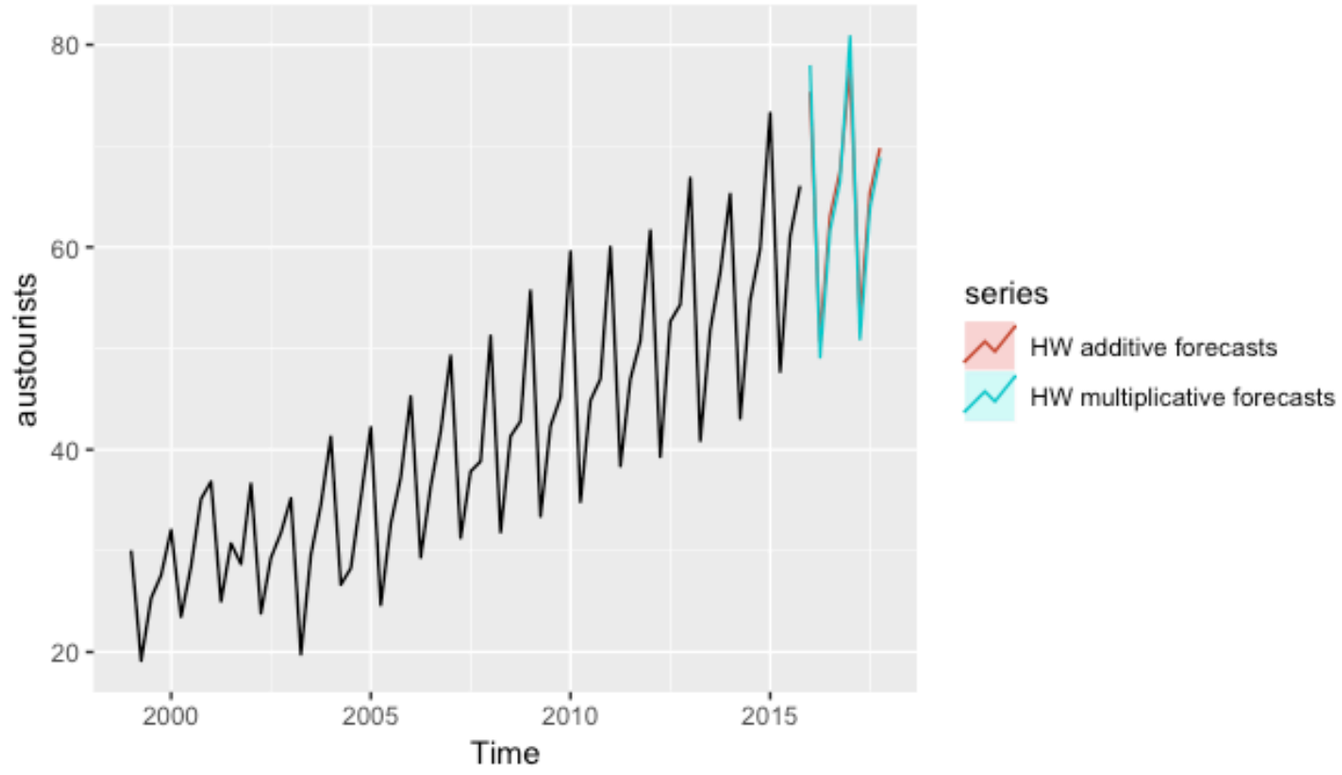
```
autoplot(fc1,series="Holt's method", PI=FALSE) +
  autolayer(fc2, series="Damped Holt's method", PI=FALSE)
```

Holt-Winter季节模型

Holt-Winter模型在Holt模型基础上解决了数据既带有明显季节性又复合了强烈趋势的问题

Holt-Winter模型需要通过将数据T-S特征分解后建模，从而出现加法型’additive’和乘法型’multiplicative’两种模型

```
fit1=hw(austourists,seasonal = 'additive')
fit2=hw(austourists,seasonal = 'multiplicative')
```




```
autoplot(austourists)+
    autolayer(fit1, series="HW additive forecasts", PI=FALSE) +
    autolayer(fit2, series="HW multiplicative forecasts",PI=FALSE)
```

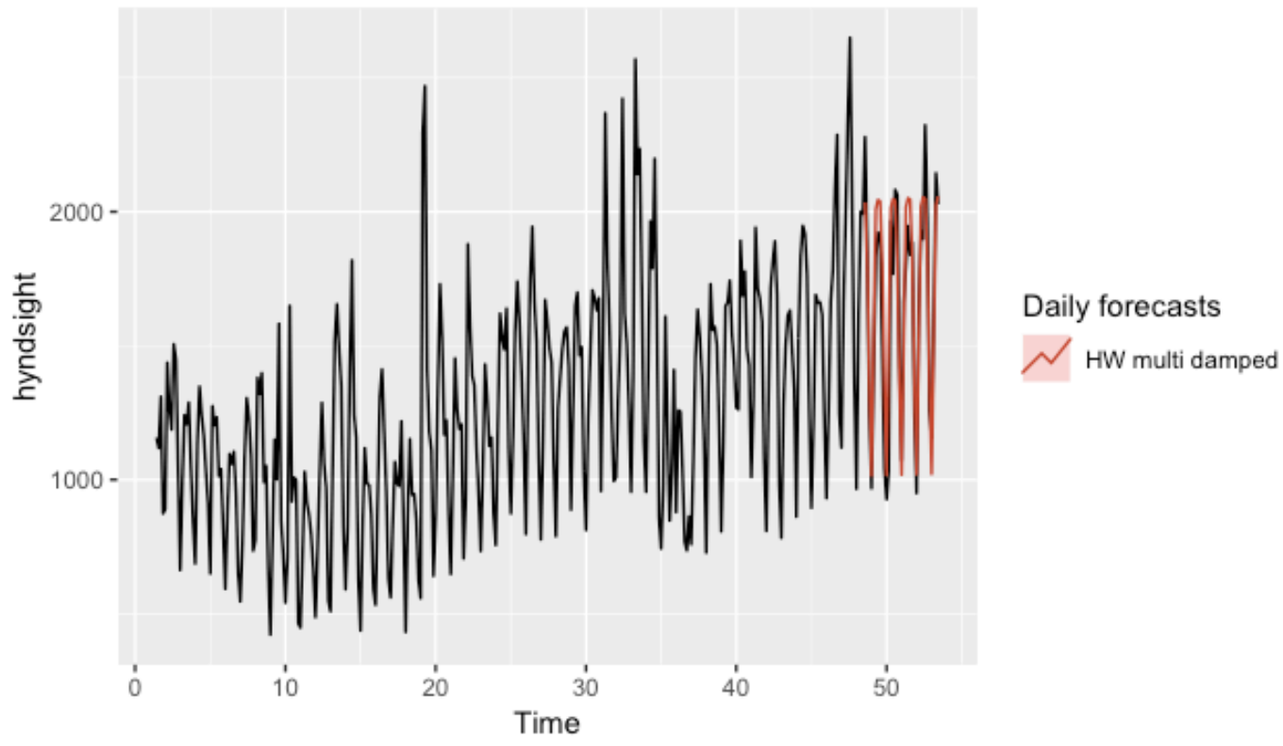
带阻滞的Holt-Winter季节模型

Holt-Winter模型同样集成了Holt模型的阻滞增长特点，在hw函数中存在与holt模型同样的阻滞逻辑参数

```
fc <- hw(subset(hyndsight,end=length(hyndsight)-35),
         damped = TRUE, seasonal="multiplicative", h=35)
```

为了检验预测效果，我们空出35个数据，其余数据用于训练模型

```
autoplot(hyndsight) +
    autolayer(fc, series="HW multi damped", PI=FALSE)+
guides(colour=guide_legend(title="Daily forecasts"))
```



差分移动平均自回归模型-Arima

前述模型都是建立在趋势较为明显的基础上，当趋势越来越复杂，直接在原始序列上做任何模型都失去了意义，并且T-C特征也无法再混合在一起

为了能预测复杂趋势特征，我们需要更多的观察角度去找到可描述的趋势特征。为此，转向在数据的差分上做分析，即前后数据之差，也叫做随机游走(random walk)

原序列

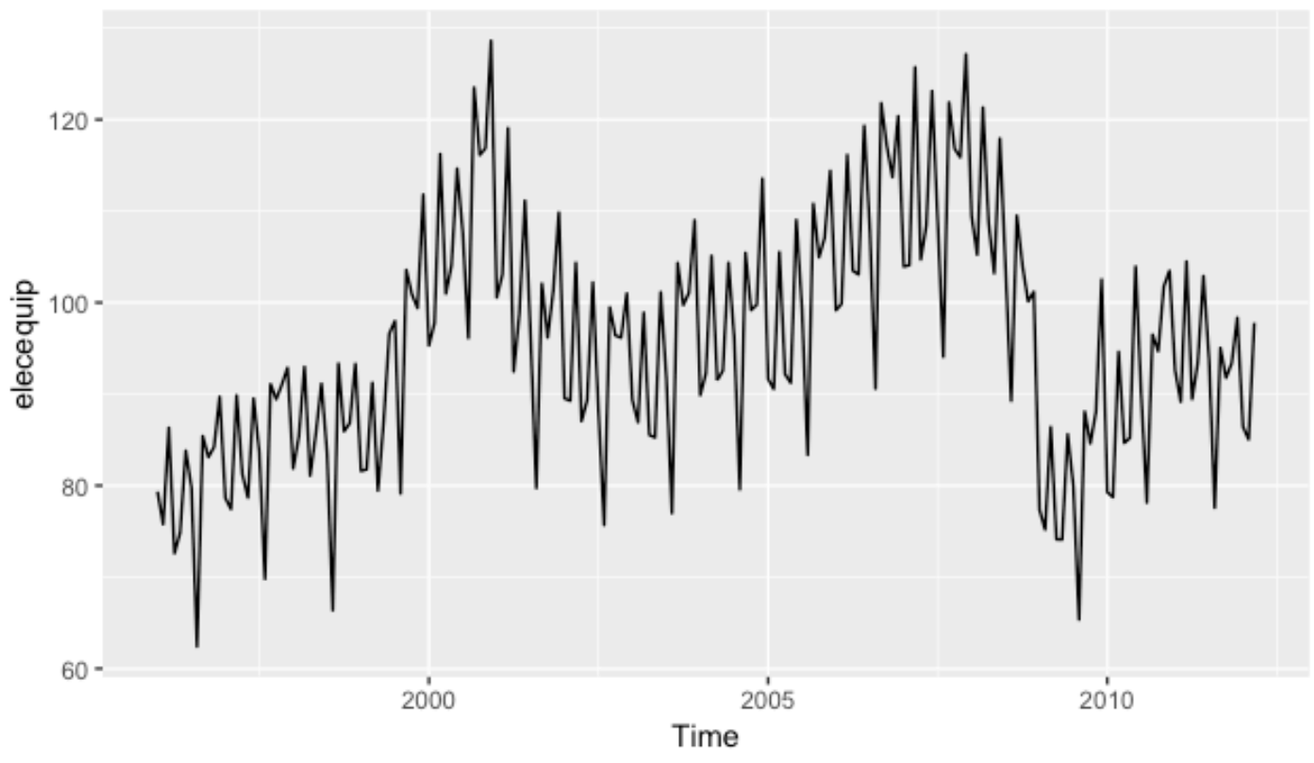
$$y_t$$

一阶差分量

$$y'_t = y_t - y_{t-1}$$

二阶差分量

$$y''_t = y'_t - y'_{t-1}$$



思考差分和导数的关系

随机漫步random walk

一阶平稳　一阶平稳对应了前变用到的所有模型

一阶差分量 $y'_t = y_t - y_{t-1}$ 为了方便记做 $y_t - y_{t-1} = \varepsilon_t$

如果一阶差分序列是平稳的，即 $y_t = y_{t-1} + \varepsilon_t$ 呈现出高度相关或者加入常数后 $y_t - y_{t-1} = c + \varepsilon_t \quad \text{or} \quad y_t = c + y_{t-1} + \varepsilon_t$ 残差也是高度相关，则意味着序列趋势是增长或者下降。如果在一阶差分找不到规律，则需要高阶或者季节性差分中寻找。

二阶平稳

当我们在一阶序列中找不到平稳状态，则进入二阶差分寻找平稳性。

$$\begin{aligned} y_t'' &= y_t' - y_{t-1}' \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2}. \end{aligned}$$

季节平稳

除连续的差分外，季节性差分平稳性也是考虑的角度之一。如滞后m期做差分。$y_t' = y_t - y_{t-m}$ 或形式变为 $y_t = y_{t-m} + \varepsilon_t$

移动平均模型

Moving average model(简称MA)不同于AR用滞后变量做回归，MA用白噪声作为自变量做回归，阶数q指模型中的滞后变量个数。

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

自回归模型

Autoregression model(简称AR)利用序列自身的滞后期作为自变量做回归，它的阶数p指模型中的自回归变量个数，记做AR(p)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

差分移动平均自回归模型

ARIMA(p,d,q)模型则是综合了AR和MA模型，其中p为自回归项数、q是移动平均项数，d则是差分阶数

确定各参数最合适的取值是一个不太容易的事情，forecast包中给出一个自动定参数的auto.arima函数，按照数据特征进行优化个参数。

```
fc=auto.arima(elecequip)

autoplot(forecast(fc,15))
```

arima模型相对于前变的简单模型要复杂的多，auto.arima或arima做出的结果仅是对模型的训练，而非直接给出预测结果，需要用forecast函数再做一次预测，类似于线性回归里的prediction函数