时间序列分析基础

作者 张洋 | 发布于 2013-05-27

数据挖掘 数据分析 时间序列 ARIMA R

时间序列是现实生活中经常会碰到的数据形式。例如北京市连续一年的日平均气温、某股票的股票价格、淘宝上某件商品的日销售件数等等。时间序列分析的的目的是挖掘时间序列中隐含的信息与模式，并借此对此序列数据进行评估以及对系列的后续走势进行预测。

由于工作需要，我最近简单学习了时间序列分析相关的基础理论和应用方法，这篇文章可以看做是我的学习笔记。

文章主要内容会首先描述时间序列分析的基本概念和相关的统计学基础理论，然后着重讲述十分经典和常用的ARIMA模型，在这之后会讲述季节ARIMA模型。

由于打算以学习笔记的形式写这篇文章，所以我不会一下子写完整篇文章才发布，而是持续更新这篇文章，写的过程中也可能会对前面的内容进行修订。

文章中会穿插许多实例，分析过程中将使用R为分析工具。

基本概念

时间序列

简单来说，时间序列是一个变量在不同时间点的值所组成的有序序列。例如北京市2013年4月每日的平均气温就构成了一个30个元素的时间序列，为了方便，我们一般认为序列中相邻元素具有相同的时间间隔。

时间序列可以分为确定的和随机的。例如一个1990年出生的人，从1990年到1999年年龄可以表述为$\{0,1,2,\ldots,9\}$，这个序列并没有任何随机因素。不过现实生活中我们面对的更多的是掺杂了随机因素的时间序列，例如气温、销售量等等。

时间序列分析

时间序列分析说白了就是寻找时间序列中的模式。如果是在确定性时间序列中，这就基本等价于寻找序列的通项公式，例如上面年龄的时间序列，用差为1的等差数列公式就可以很好的描述其模式。

当然实际的时间序列分析基本都是针对随机时间序列。对于随机时间序列，情况会复杂一些，但本质上还是可以看做寻找通项公式（可以是封闭形式或递推形式），只不过我们面对的序列存在随机扰动，所以分析过程中除了确定性序列分析的技术外，还需要一些概率统计方面的知识和方法，下面一节会介绍一些相关的基础知识。

主要统计量

注意时间序列中的每一个元素都是一个普通的随机变量，如果忽略序列的时间性，那么我们面对的实际上是一个随机变量集合，所以从这个角度来说时间序列的统计分析与普通统计分析没有太大不同，相关的理论也是通用的。

对于随机变量集合来说，要完整描述其统计特性需要处理其多元联合分布，这是非常复杂的。所以实际我们往往做一些必要的简化假设，避免处理复杂的多元联合分布。

现假设我们有随机时间序列

$$\{Y_t|t=0,\pm1,\pm2,\cdots\}$$

下面先给出一些常用的统计量。后面会接着通过一些常见序列来举例说明各统计量如何计算。

均值

均值函数被定义为关于自变量t的函数：

$$\mu_t = E(Y_t)$$

t的均值函数值表示在t时刻随机变量Y_t的期望。

方差

与均值类似，方差是t时刻序列元素的方差：

$$\sigma_t^2 = E((Y_t - \mu_t)^2)$$

自协方差

自协方差是一个二元函数，其自变量为两个时间点，值是两个时间点上序列值的协方差：

$$\gamma_{t,s} = Cov(Y_t, Y_s) = E((Y_t - \mu_t)(Y_s - \mu_s))$$

当t=s时，自协方差就是t时刻的方差。

自相关系数

自相关系数是两个时刻的值的相关系数：

$$\rho_{t,s} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$$

如果忽略元素来自时间序列这一事实，各统计量的意义实际上与普通的统计学中无异。因此这些统计量的一些性质也可以无缝推广到时间序列分析。例如期望的线性性质等等。如果有需要可以自行复习一下这些统计量的相关计算性质。后面的推导会主要集中于这几个统计量的计算。

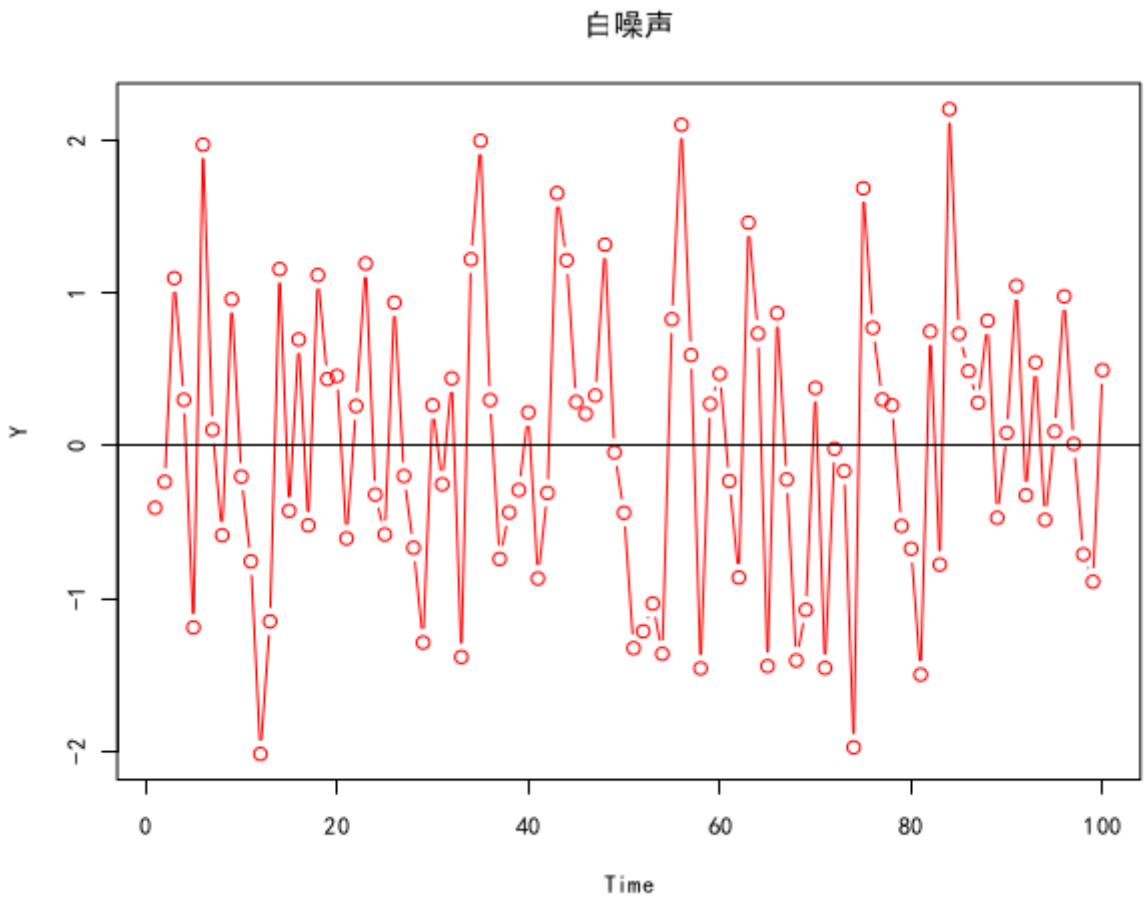
时间序列示例

下面看几个简单的随机时间序列示例。

白噪声

考虑一个时间序列，其中每一个元素为独立同分布变量，且均值为0。这种时间序列叫做白噪声。之所以叫这个名字，是因为对这种序列的频域分析表明其中平等的包含了各个频率，和物理中的白光类似。

下面是用R模拟生成的白噪声时序图。




```
1.  Y = ts(rnorm(100, mean=0, sd=1));
2.  plot(Y, family="simhei", main="白噪声", type="b", col="red");
3.  abline(h=0)
```

其中共100个元素，每个元素都独立服从标准正态分布$N(0,1)$。可以从图中看出白噪声基本是在均值附近较为平均的随机震荡。

由于每个元素服从$N(0,1)$，所以均值$\mu_t = 0$，方差$\sigma_t^2 = 1$。又因为每个元素独立，所以对于任何$t \neq s$，$\gamma_{t,s} = 0$，$\rho_{t,s} = 0$。这些统计特征与对图像的直观观察基本一致。

白噪声的重要之处在于很多其它的重要时间序列都可以通过它构造出来，这一点下文会看到。我们一般用e表示白噪声，将白噪声序列写作：

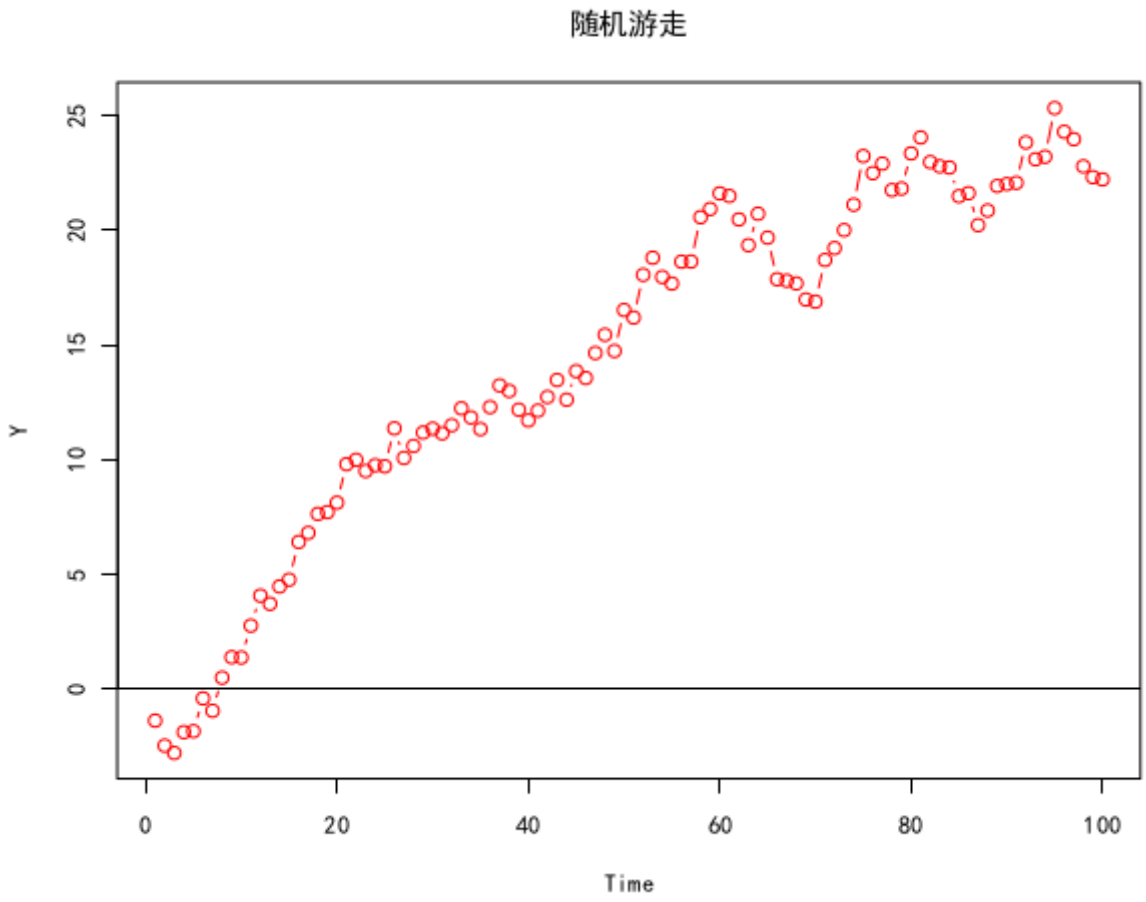
$$\{e_1, e_2, \ldots, e_t, \ldots\}$$

随机游走

下面考虑这样一个时间序列，其在t时刻的值是前面白噪声序列的前t个值之和，设$\{e_1, e_2, \ldots, e_t, \ldots\}$为标准正态分布产生的白噪声，则：

$$\begin{aligned} Y_1 &= e_1 \\ Y_2 &= e_1 + e_2 \\ &\vdots \\ Y_t &= e_1 + e_2 + \cdots + e_t \\ &\vdots \end{aligned}$$

下面是用R模拟的随机游走。




```
Y = ts(rnorm(100, mean=0, sd=1));
for (i in 2:length(Y)) {
    Y[i] = Y[i] + Y[i-1];
}
plot(Y, family="simhei", main="随机游走", type="b", col="red");
abline(h=0)
```

可以看到随机游走比白噪声平滑很多，并且呈现出一些“趋势性”的感觉。下面分析其相关统计特征。

均值：$\mu_t = E(e_1 + \cdots + e_t) = E(e_1) + \cdots + E(e_t) = 0$

方差：$\sigma_t^2 = Var(e_1 + \cdots + e_t) = Var(e_1) + \cdots + Var(e_t) = t\sigma^2$

对协方差的计算需要用到一个协方差性质：

$$Cov(\sum_{i=1}^{m} c_i Y_i, \sum_{j=1}^{n} d_j Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} c_i d_j Cov(Y_i, Y_j)$$

设t小于s，由于只有i=j时$Cov(Y_i, Y_j) = \sigma^2$，所以：

自协方差：$\gamma_{t,s} = t\sigma^2$

自相关系数：$\rho_{t,s} = \frac{t\sigma^2}{\sqrt{ts\sigma^4}} = \sqrt{\frac{t}{s}}$

从统计性质可以看到，随机游走的“趋势性”实际是个假象，因为其均值函数一直是白噪声的均值，不存在偏离的期望。但是方差与时间呈线性增长并且趋向于无穷大，这意味着只要时间够长，随机游走的序列值可以偏离均值任意远，但期望永远在均值处。

物理与经济学中的很多现象都被看做是随机游走，例如分子的布朗运动，股票的价格走势等等。

从协方差和相关系数看，如果起点t固定，则越接近的点相关性越大，例如$\rho_{1,2} = 0.707$，$\rho_{1,3} = 0.577$，$\rho_{1,4} = 0.500$。同时，起点不同，时滞相同自相关系数也不同，越往后同时滞自相关系数越大，例如$\rho_{2,3} = 0.816$，$\rho_{3,4} = 0.866$。

实际上从纯数学角度可以将自相关系数看成一个二元函数，自变量是时间点t和时滞s-t。认识到这点很重要，因为它与时间序列分析中一个重要的概念——平稳性有着密切的关系。

平稳性

平稳性是时间序列分析中很重要的一个概念。一般的，我们认为一个时间序列是平稳的，如果它同时满足一下两个条件：

1）均值函数是一个常数函数

2）自协方差函数只与时滞有关，与时间点无关

以上面两个时间序列为例。两个序列均满足条件1），因为标准正态分布白噪声和其形成的随机游走的均值函数都是值恒为0的常数函数。再来看条件2）。白噪声的自协方差函数可以表述为：

$$\gamma_{t,s} = \begin{cases} 1 & (t = s) \\ 0 & (t \neq s) \end{cases}$$

可以看到只有在时滞为0时值为1，其它均为0，所以白噪声是一个平稳序列。

而随机游走我们上面分析过，其自协方差为：

$$\gamma_{t,s} = t\sigma^2$$

很明显其自协方差依赖于时间点，所以是一个非平稳序列。

后面可以看到，一般的时间序列分析往往针对平稳序列，对于非平稳序列会通过某些变换将其变为平稳的，例如，对于随机游走来说，其一阶差分序列是平稳的（显然其一阶差分是白噪声）。

下一章节会介绍ARIMA模型，其中将对上面提到的平稳性和差分的概念给出更具体的说明的示例。

注意我们下面说到平稳序列时都默认其均值为0，因为具有均值μ的平稳时间序列只要将其做一个变换$Y_t - \mu$就可以得到一个均值为0的序列，假设均值为0可以使得问题分析得到简化。

ARIMA模型

基本模型

上文说过，时间序列分析实际上是寻找随机时间序列中的模式，所以首先要对时间序列做一个假设，假设其符合某个模式。具体一点，就是时间序列可以用一个函数（可以包含随机变量）来描述。函数的自变量是时刻，值是这个时刻序列的值。例如上例中的白噪声可以用$f(t) = e$描述，其中e是一个服从标准正态分布的随机变量。

时间序列分析的核心工作之一就是根据观察到的序列值来估计这个函数。

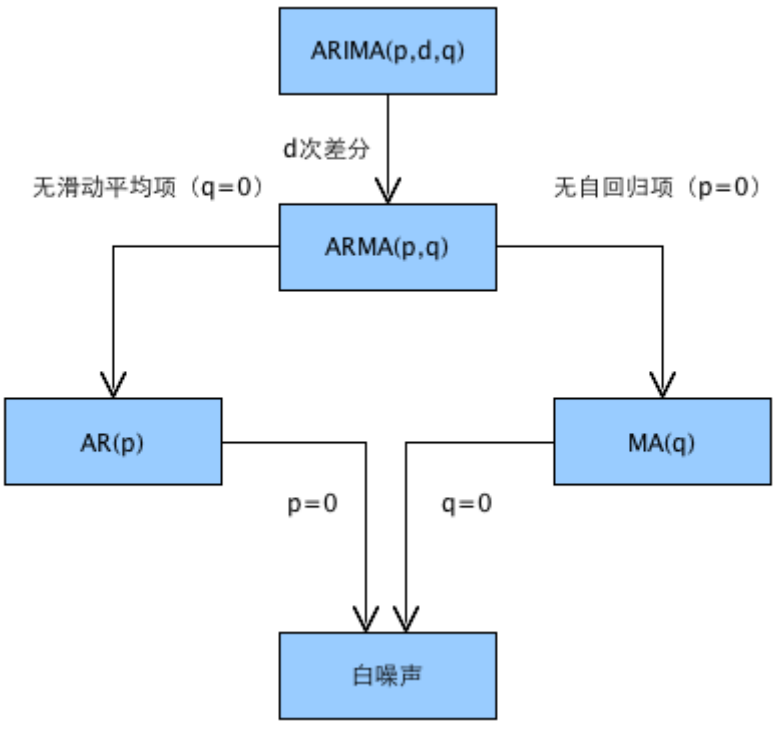
其中ARIMA模型是一个常用的函数模型。实际上ARIMA模型本身并不是一个具体描述时间序列的函数，而是一类函数的总称。ARIMA模型可以表述为$ARIMA(p, d, q)$，其中p、d和q定义域均为自然数。从这个角度看，ARIMA可以看成函数的函数，或者叫做高阶函数。这个高阶函数将一个定义在自然数上的三元空间映射到一个具体函数，具体函数可以描述一个时间序列。

例如：

$$\begin{aligned} (ARIMA(0,0,1))(t) &= e_t - \theta e_{t-1} \\ (ARIMA(0,0,2))(t) &= e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \\ (ARIMA(1,0,0))(t) &= e_t + \phi(ARIMA(1,0,0))(t-1) \\ (ARIMA(1,0,2))(t) &= e_t + \phi(ARIMA(1,0,2))(t-1) - \theta_1 e_{t-1} - \theta_2 e_{t-2} \end{aligned}$$

当不同参数取0时ARIMA可以退化为更简单的形式，例如d=0时，模型变为ARMA，如果d和q都等于0，就变为AR模型，而如果d和p为0，则是MA模型，如果d、p和q都为0，那就是白噪声了。所以$ARIMA(0,0,0)$就是白噪声序列函数，因此白噪声也只是ARIMA模型的一个特例。

下图说明了各个模型间的关系：



图中自顶向下越来越特化，而自底向上则越来越泛化。

这一节我们从较为简单的特化ARIMA模型开始讲述，由简入难，一步一步描述各种模型。后面的几节会讲述如何对ARIMA模型进行训练和估计以及如何应用ARIMA模型进行预测。

MA模型

ARIMA中的p、d和q分别表示自相关、差分和滑动平均。当p和d均为0时，就变成了简单的滑动平均模型$MA(q)$。

一般的滑动平均模型被定义为：

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

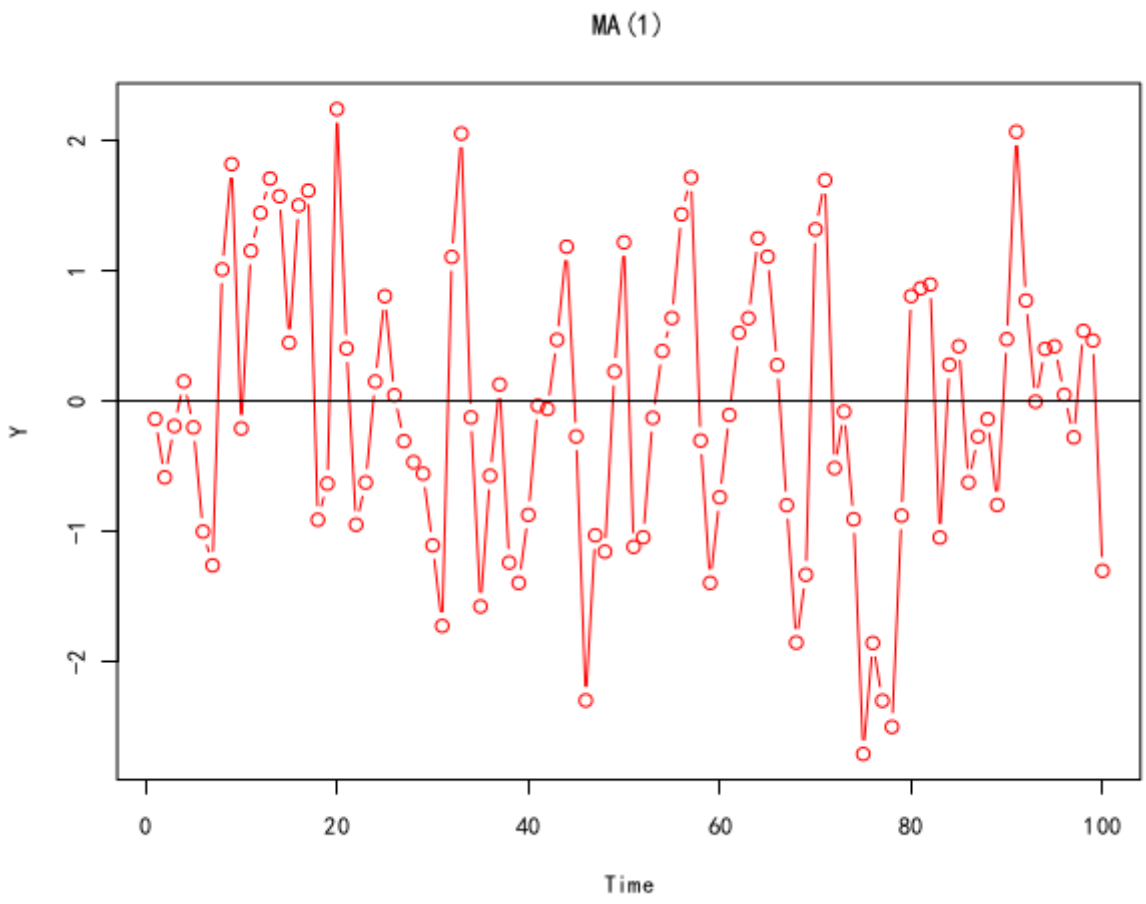
其中e是方差为σ^2的白噪声。并且要求各参数$-\theta$是定义在-1到1的闭区间上。

可以看出，MA(q)在t时刻的值就是白噪声序列t到t-q共q+1个点的线性组合，系数是$(1, -\theta_1, \ldots, -\theta_q)$。

简单起见，我们分析一下最简单的MA(1)和MA(2)模型及其统计性质。

下面是MA(1)模型的序列函数以及用R产生的模拟数据：

$$Y_t = e_t - \theta e_{t-1}$$




```r
Ye = rnorm(100, mean=0, sd=1);
Y = c();
Y[1] = Ye[1];
for (i in 2:length(Ye)) {
    Y[i] = Ye[i] - (-0.8) * Ye[i-1];
}
Y = ts(Y);
plot(Y, family="simhei", main="MA(1)", type="b", col="red");
```

```
9.   abline(h=0)
```

这个模拟数据构造自服从标准正态分布的白噪声，其中一阶滑动参数为-0.8。从图上看，这个序列比白噪声平滑，并且比随机游走平稳一些。下面定量分析其各统计量。

均值：

$$\mu_t = E(Y_t) = E(e_t) - \theta E(e_{t-1}) = 0$$

方差：

$$\sigma_t^2 = Var(Y_t) = Var(e_t) + \theta^2 Var(e_{t-1}) = (1+\theta^2)\sigma^2$$

自协方差：

$$\gamma_{t,s} = Cov(e_t - \theta e_{t-1}, e_s - \theta e_{s-1}) = Cov(e_t, e_s) - \theta Cov(e_t, e_{s-1}) - \theta Cov(e_{t-1}, e_s) + \theta^2 Cov(e_{t-1}, e_{s-1})$$

显然，在t小于s时，只有s-t=1时，有$Cov(e_t, e_{s-1}) = Var(e_t) = \sigma^2$。所以自协方差函数只与时滞s-t有关，我们将自协方差表示为时滞k的函数γ_k，我们有：

$$\gamma_k = \begin{cases} (1+\theta^2)\sigma^2 & (k=0) \\ -\theta\sigma^2 & (k=1) \\ 0 & (k>1) \end{cases}$$

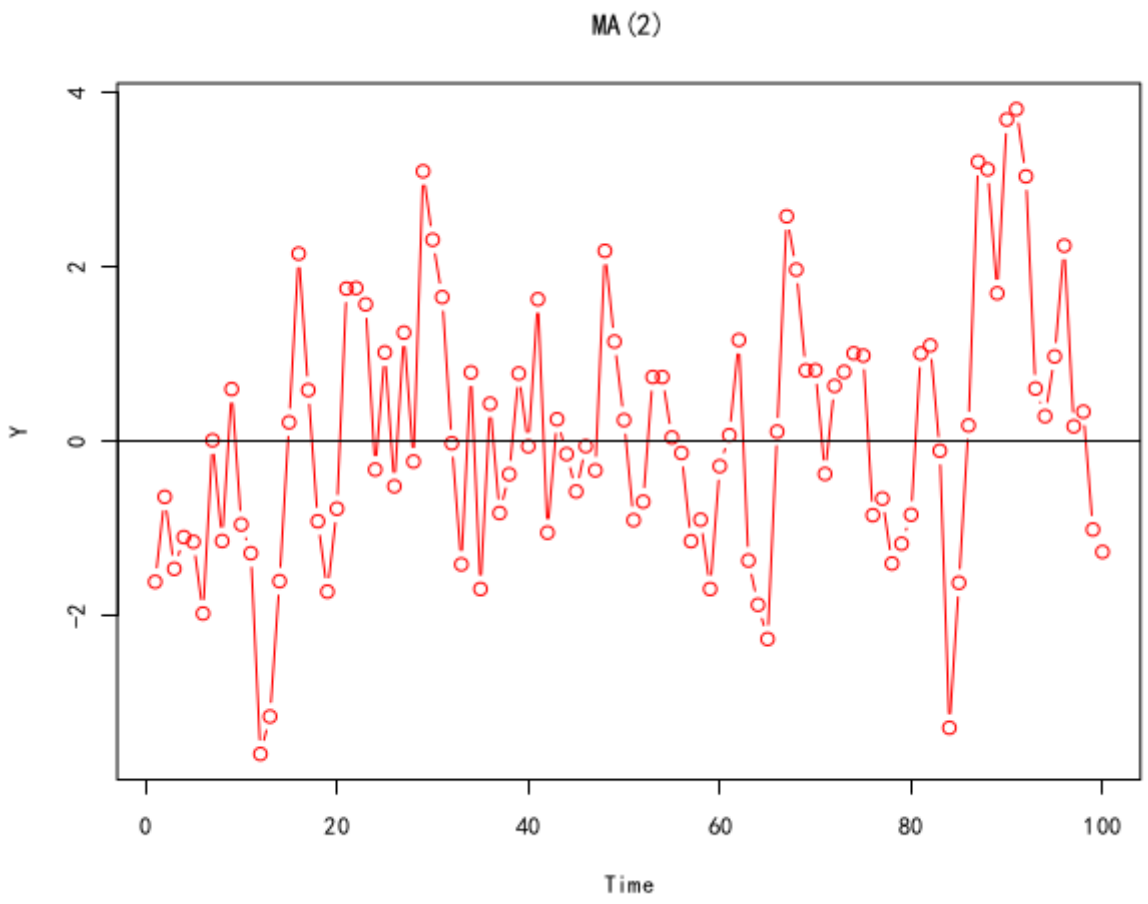
自相关系数：

$$\rho_k = \begin{cases} 1 & (k=0) \\ (-\theta)/(1+\theta^2) & (k=1) \\ 0 & (k>1) \end{cases}$$

从上面分析得出，MA(1)模型是一个平稳序列，因为其均值为常数，自协方差只与时滞相关。以后任何平稳模型，我们都将自协方差和自相关系数表示为时滞k的函数，而不再表示为t和s的函数。另外还可以发现，MA(1)模型每一个序列值只与其前一个值有相关性，而时滞超过1则无相关性，后面可以看到这个特性是识别MA模型的重要特征。另外不难证明，一阶自相关系数在θ为正负1时分别达到最强负相关-0.5和最强正相关0.5，在θ为0时，MA(1)退化为白噪声，因此自相关系数为0。

类似的，MA(2)模型可以表述为：

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$




```
1.   Ye = rnorm(100, mean=0, sd=1);
2.   Y = c();
3.   Y[1] = Ye[1];
4.   Y[2] = Ye[2] - (-0.8) * Ye[1];
5.   for (i in 3:length(Ye)) {
6.       Y[i] = Ye[i] - (-0.8) * Ye[i-1] - (-0.9) * Ye[i-2];
7.   }
8.   Y = ts(Y);
9.   plot(Y, family="simhei", main="MA(2)", type="b", col="red");
```

```
10.   abline(h=0)
```

如果做统计分析，会发现MA(2)模型也是一个平稳模型，并且在时滞大于2后没有相关性。

一般的，MA(q)模型是一个平稳模型，并且在时滞大于q后没有相关性。此处不再给出一般MA模型的统计分析，有兴趣的朋友可以自行推导。

AR模型

现在考虑另一种模型：t时刻的序列值与其滞后p的p个时间序列呈多元线性相关：

$$Y_t = e_t + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}$$

从公式上看，AR应该比MA具有更强的自相关性，因为MA仅与滞后的白噪声因素相关，而AR是时间序列前后直接相关。由于AR模型的统计特性推导比MA复杂很多，所以我们先分析最简单的AR(1)，借此了解AR模型的特性。

AR(1)的序列公式如下：

$$Y_t = e_t + \phi Y_{t-1}$$

与MA(1)不同，这是一个递推公式，并且是一个线性递推公式，我们可以把它展开：

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots + \phi^k e_{t-k} + \cdots$$

我们会得到一个无穷级数表达式（假设原始白噪声序列有无穷多滞后项）。显然其均值为0。方差的计算如下：

$$\sigma_t^2 = Var(Y_t) = (1 + \phi^2 + \phi^4 + \cdots + \phi^{2k} + \cdots)\sigma^2 = (\lim_{n\to\infty} \frac{1-\phi^{2n}}{1-\phi^2})\sigma^2$$

显然只有当$|\phi| < 1$时级数收敛到$\frac{\sigma^2}{1-\phi^2}$

这里不加证明给出一个重要的结论：AR(1)是平稳的当且仅当$|\phi| < 1$。下面我们假设序列满足平稳条件，推导其自协方差和自相关系数。

$$\begin{aligned}
\gamma_k &= E((Y_{t-k} - \mu_{t-k})(Y_t - \mu_t)) \\
&= E(Y_{t-k}Y_t) \\
&= E(Y_{t-k}e_t + \phi Y_{t-k}Y_{t-1}) \\
&= E(Y_{t-k})E(e_t) + \phi E(Y_{t-k}Y_{t-1}) \\
&= \phi\gamma_{k-1}
\end{aligned}$$

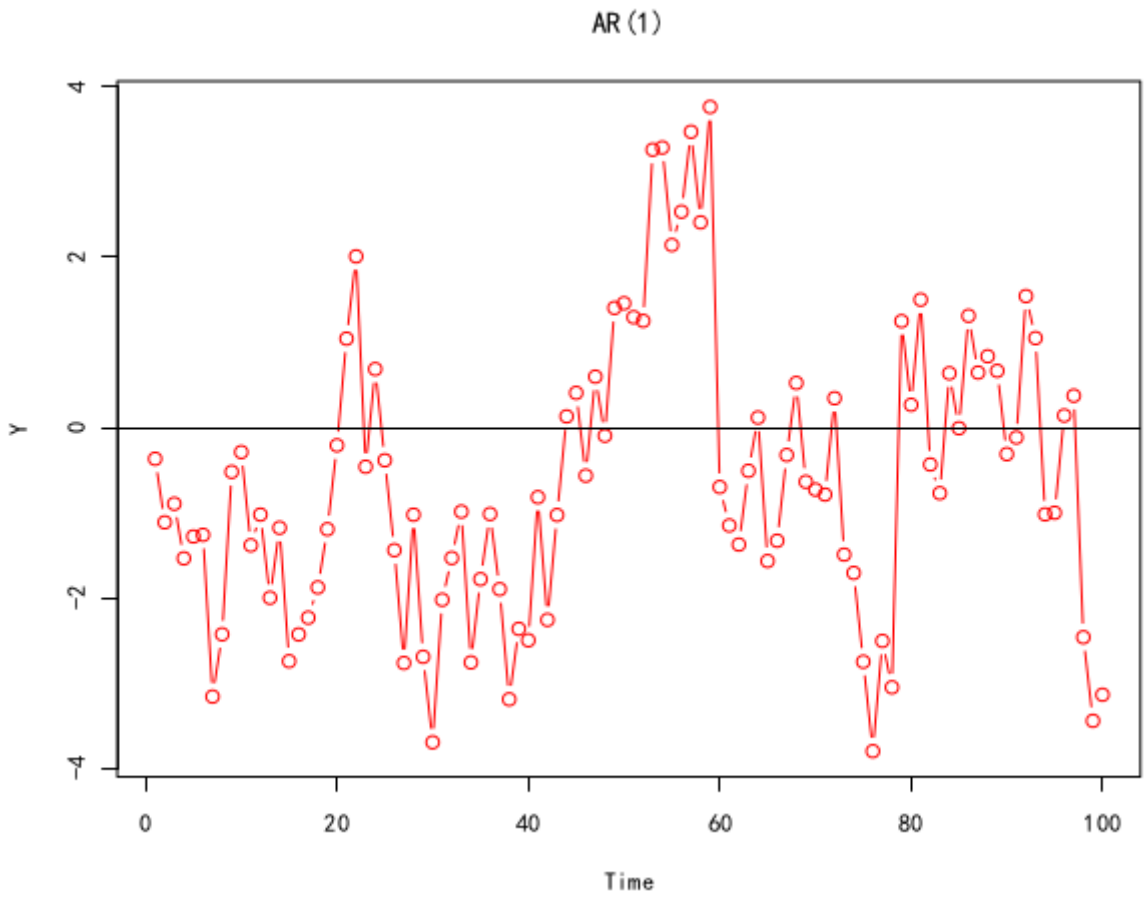
由上面的递推式得：

$$\gamma_k = \phi^k \gamma_0 = \phi^k \frac{\sigma^2}{1-\phi^2}$$

$$\rho_k = \gamma_k / \gamma_0 = \phi^k$$

从统计特性可以知道，AR(1)模型相近的时序点倾向于一起“运动”，因此可能呈现假趋势。前置节点的影响随着时滞的增大而呈指数衰减。这种特性对于模型识别非常重要。

下面是一个R模拟的AR(1)时间序列。




```r
Ye = rnorm(100, mean=0, sd=1);
Y = c();
Y[1] = Ye[1];
for (i in 2:length(Ye)) {
    Y[i] = Ye[i] + 0.8 * Y[i-1];
}
Y = ts(Y);
plot(Y, family="simhei", main="AR(1)", type="b", col="red");
abline(h=0)
```

下面说一下AR模型的平稳条件。为了讨论这点，我们先引进一个定义：AR模型的特征方程。一般的，AR(p)模型的特征方程被定义为：

$$1-\phi_1 x-\phi_2 x^2-\cdots-\phi_p x^p=0$$

显然一个AR(p)的特征方程是一个一元p次方程。

已经证明：一个AR模型是平稳的当且仅当其特征方程的所有根的绝对值大于1。

利用这个结论可以将AR的平稳性问题转化为一个代数问题。例如上面的AR(1)模型，其特征方程为$1-\phi x=0$，唯一的根为$x=1/\phi$，因此AR(1)的平稳条件是$|1/\phi|>1$，等价于$|\phi|<1$，这是上面给出过的AR(1)平稳条件。

当p大于1时，因为特征方程可能存在复数根，所以平稳条件的计算涉及比较复杂的线性代数和复变函数知识，这里就不再详述了。

对于一般AR模型

$$Y_t=e_t+\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\cdots+\phi_p Y_{t-p}$$

假设AR已经满足平稳性条件，有如下统计特性：

对上式两边求期望，可得$\mu=E(Y_t)=(\phi_1+\cdots+\phi_p)\mu$，要使此等式恒成立显然：

$$\mu=0$$

对上式两边乘以Y_{t-k}，然后求期望，可得自协方差递推式：

$$\gamma_k=E(Y_tY_{t-k})=\phi_1\gamma_{k-1}+\cdots+\phi_p\gamma_{k-p}$$

再除以γ_0自相关系数递推式：

$$\rho_k=E(Y_tY_{t-k})=\phi_1\rho_{k-1}+\cdots+\phi_p\rho_{k-p}$$

而对于初始值$\rho_1,\ldots,\rho_p$的求解，根据平稳序列自相关系数的稳定性，有$\rho_{-k}=\rho_k$，再加上$\rho_0=1$，带入上面递推式可得一个含有p个未知量的线性方程组，解方程组就可以得到$\rho_1,\ldots,\rho_p$。

通过上面的分析可以发现AR模型统计特性的分析最终会归结为线性代数问题，不过在现实的时间序列分析中能否从数学意义上理解上述过程并不是重点，重点是直观理解AR(p)模型在不同的参数ϕ下其自相关系数随时滞k的变化情况。

而且现实建模时一般很少使用高于AR(2)的模型，因为过高的阶会导致复杂的模型和提高过拟合风险。因此在实际使用中了解AR(1)和AR(2)的特性一般就足够了，后面在模型识别中会结合图形描述AR(2)的自相关函数特性。

ARMA模型

如果一个时间序列兼有AR和MA部分，并且是平稳的，则构成ARMA模型。一般$ARMA(p,q)$的表达式为：

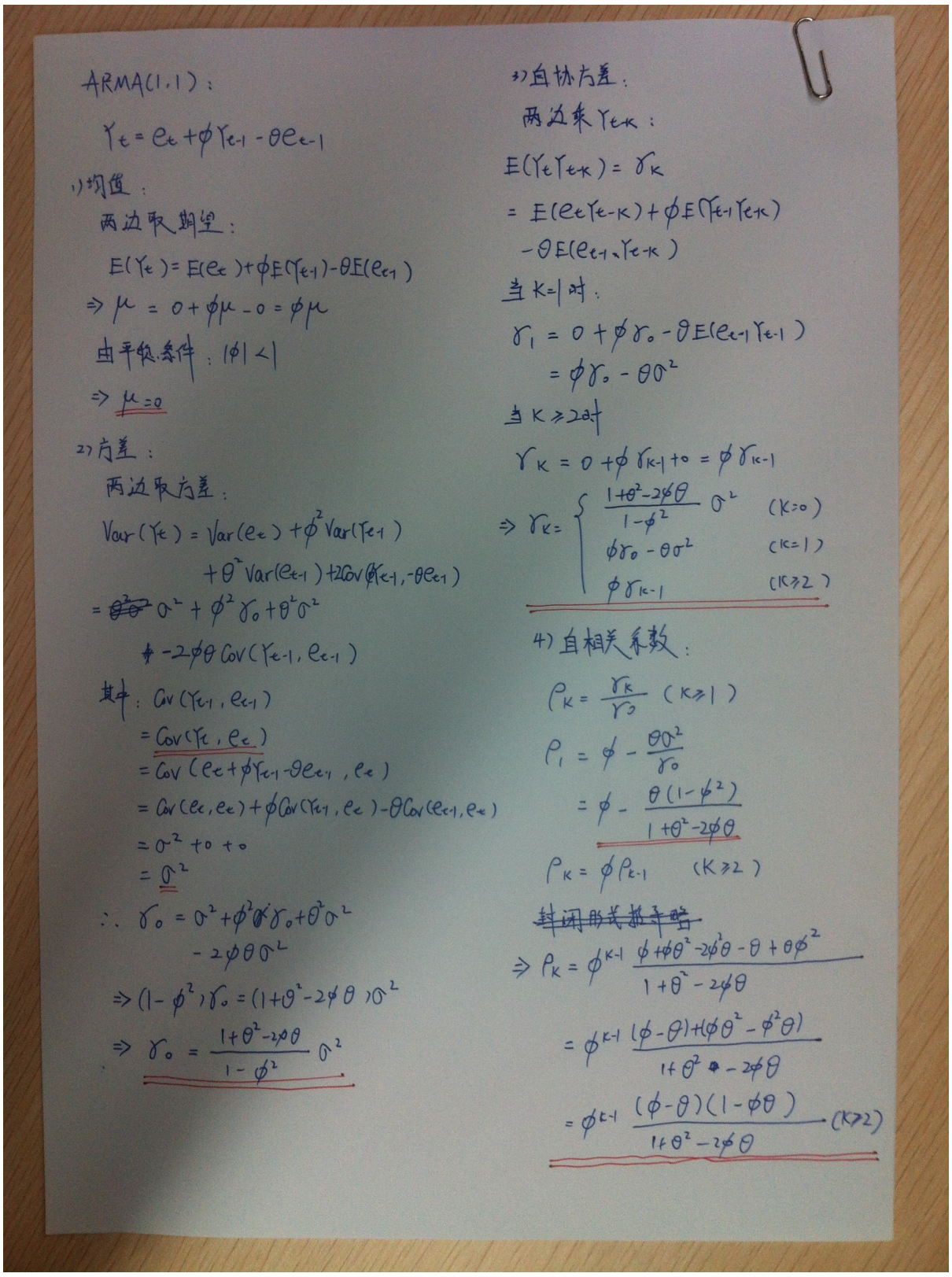
$$Y_t = e_t + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

已经证明，ARMA序列是平稳的当且仅当其AR特征方程的根的模大于1。因此求解ARMA平稳条件与AR平稳条件无异，只需忽略MA部分直接套用AR平稳条件求解即可。换句话说，ARMA(p,q)平稳当且仅当AR(p)平稳。

下面研究最简单的ARMA(1,1)模型。这是一个带有一阶自回归和一阶滑动平均的序列：

$$Y_t = e_t + \phi Y_{t-1} - \theta e_{t-1}$$

其平稳的条件等于AR(1)的平稳条件，也就是$|\phi| < 1$。在这个前提下，我们分析其统计特性。由于推导过程比较复杂，这里直接把我之前在纸上推导的草稿贴出来，详见下图（点击图片可放大）。



其中结论性部分我用红笔标出了。根据上面的推导结果，ARMA(1,1)的统计特性如下：

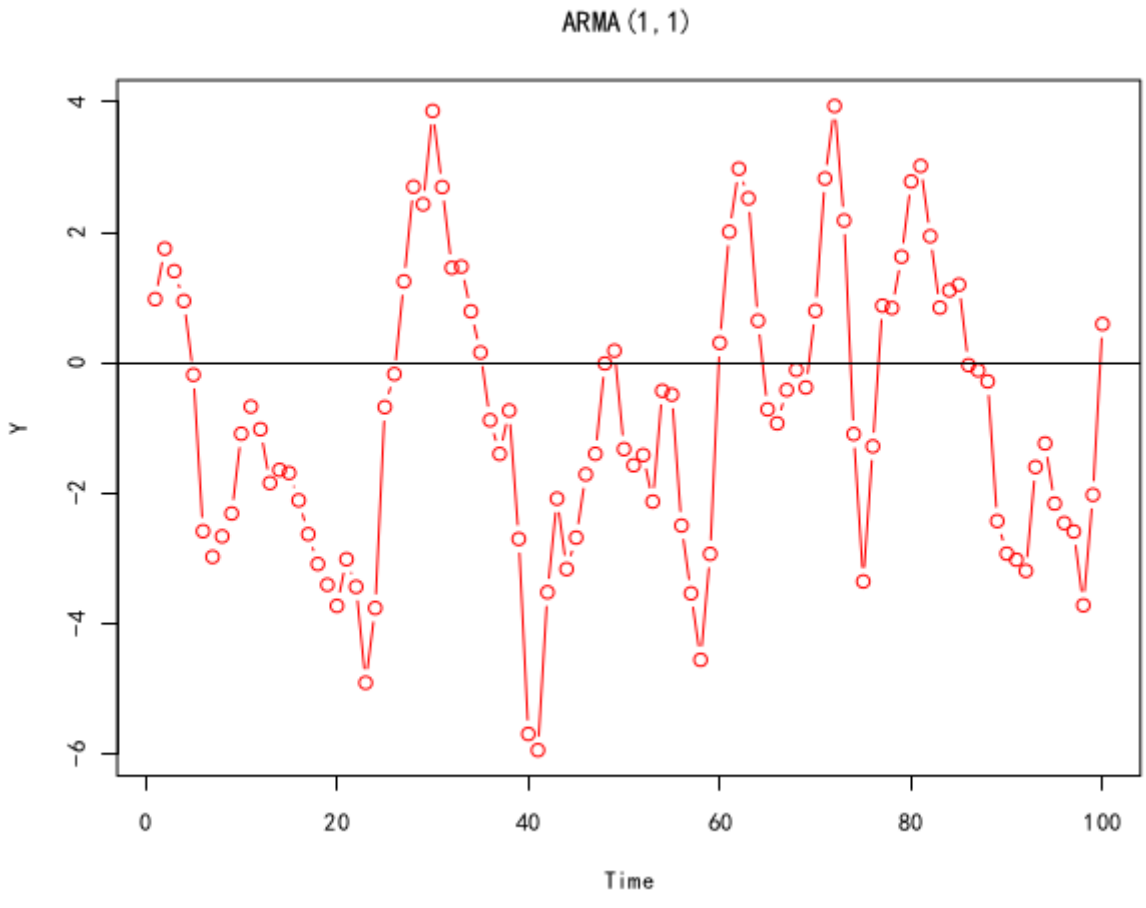
$$\mu = 0$$

$$\gamma_k = \begin{cases} \frac{1+\theta^2-2\phi\theta}{1-\phi^2}\sigma^2 & (k=0) \\ \phi\gamma_0 - \theta\sigma^2 & (k=1) \\ \phi\gamma_{k-1} & (k \geq 2) \end{cases}$$

$$\rho_k = \phi^{k-1}\frac{(\phi-\theta)(1-\phi\theta)}{1+\theta^2-2\phi\theta}$$

大约可以看到相关系数也是随时滞呈指数递减，当然不同的参数会有不同的情况，具体我们留待模型识别一节讨论。

下面给出一个模拟的ARMA(1,1)时间序列：




```
Ye = rnorm(100, mean=0, sd=1);
Y = c();
Y[1] = Ye[1];
for (i in 2:length(Ye)) {
    Y[i] = Ye[i] + 0.8 * Y[i-1] - (-0.9) * Ye[i-1];
}
Y = ts(Y);
plot(Y, family="simhei", main="ARMA(1,1)", type="b", col="red");
abline(h=0)
```

ARIMA模型

平稳性、可逆性及一般线性过程

模型识别

参数估计

模型预测

实例：使用R进行ARIMA时间序列分析

模型诊断

季节ARIMA模型