时间序列分析（TIME SERIES ANALYSIS）

主讲：吴尚
复旦大学管理学院统计与数据科学系

模型诊断与优化

- 为什么要模型诊断?
- 残差分析
- 模型选择与优化

为什么要模型诊断?

- 模型的假设和识别可能有误
- 过度差分
- 差分不够
- 模型过拟合
- 模型不充分
- 非正态
- 非平稳，需要作进一步变换
- 异方差

过度差分和差分不够

- 过度差分：如果差分后，估计所得的MA系数所构成的MA特征多项式有接近1的根，则说明过度差分了，可以减少一次差分，同时附上相应多项式趋势。

- 差分不够：如果估计所得的AR系数所构成的AR特征多项式有接近1的根，则说明差分不够，可以增加一次差分后拟合模型。

模型的残差分析

- 目的
  - *检验模型对信息的提取是否充分*
- 检验对象
  - *残差序列*
- 判定原则
  - *一个好的拟合模型应该能够提取观察值序列中几乎所有的样本相关信息，即残差序列应该为白噪声序列;*
  - *反之，如果残差序列为非白噪声序列，那就意味着残差序列中还残留着相关信息未被提取，这就说明拟合模型不够有效。*

如何计算残差

- 残差 = 实际值 − 预测值

- AR(p)模型： $\hat{e}_t = Y_t - \hat{\theta}_0 - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2} - \cdots - \hat{\phi}_p Y_{t-p}$

- ARMA(p, q)模型： $\hat{e}_t = Y_t - \sum_{j=1}^{\infty} \hat{\pi}_j Y_{t-j}$ （零均值情形）

- 残差分析的核心是分析残差是否为（近似）白噪声，注意$\hat{e}_t$约等于e_t，但并不相等。

残差的初步分析

- 散点图：作出$\hat{e}_t$对$\hat{e}_{t-k}$（或$\hat{e}_t$对Y_{t-k}）的散点图，初步考察相关性。
- 残差的ACF图：计算$\hat{e}_t$与$\hat{e}_{t-k}$（或 $\hat{e}_t$与Y_{t-k}）之间的相关系数来分析判断。若相关系数较小，则认为无相关性假设成立，即模型为适合模型；否则，认为不适合。
- Q-Q图——残差的正态性检验

Q统计量

- 将残差序列$\{\hat{e}_t\}$的自相关系数记为$\hat{r}_k$
- Q统计量（大样本情形下）
$$Q(K) = n(\hat{r}_1^2 + \hat{r}_2^2 + \cdots + \hat{r}_K^2)$$
- 如果真实模型为ARMA(p, q)，且同时用ARMA(p, q)模型拟合，则对于较大的n，$Q(K)$近似服从自由度为$K - p - q$的卡方分布。
- 最简单的例子：$p = q = 0$（思考）
- 对于白噪声，$Var(r_k) \approx \frac{1}{n}, Corr(r_k, r_j) \approx 0, k \neq j$

Ljung-Box检验

- 当样本量不够时，需要对Q统计量进行一定修正，以得到更精确的结果。

$$Q_*(K) = n(n+2)\left(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \cdots + \frac{\hat{r}_K^2}{n-K}\right)$$

- 在统计软件R当中，可以通过函数tsdiag进行Ljung-Box检验，给出的是对于若干不同的K，检验的p值：在近似分布$\chi^2(K-p-q)$下，计算比所得统计量的值$Q_*(K)$更大的概率

Exhibit 8.11 Residual Autocorrelation Values from AR(1) Model for Color

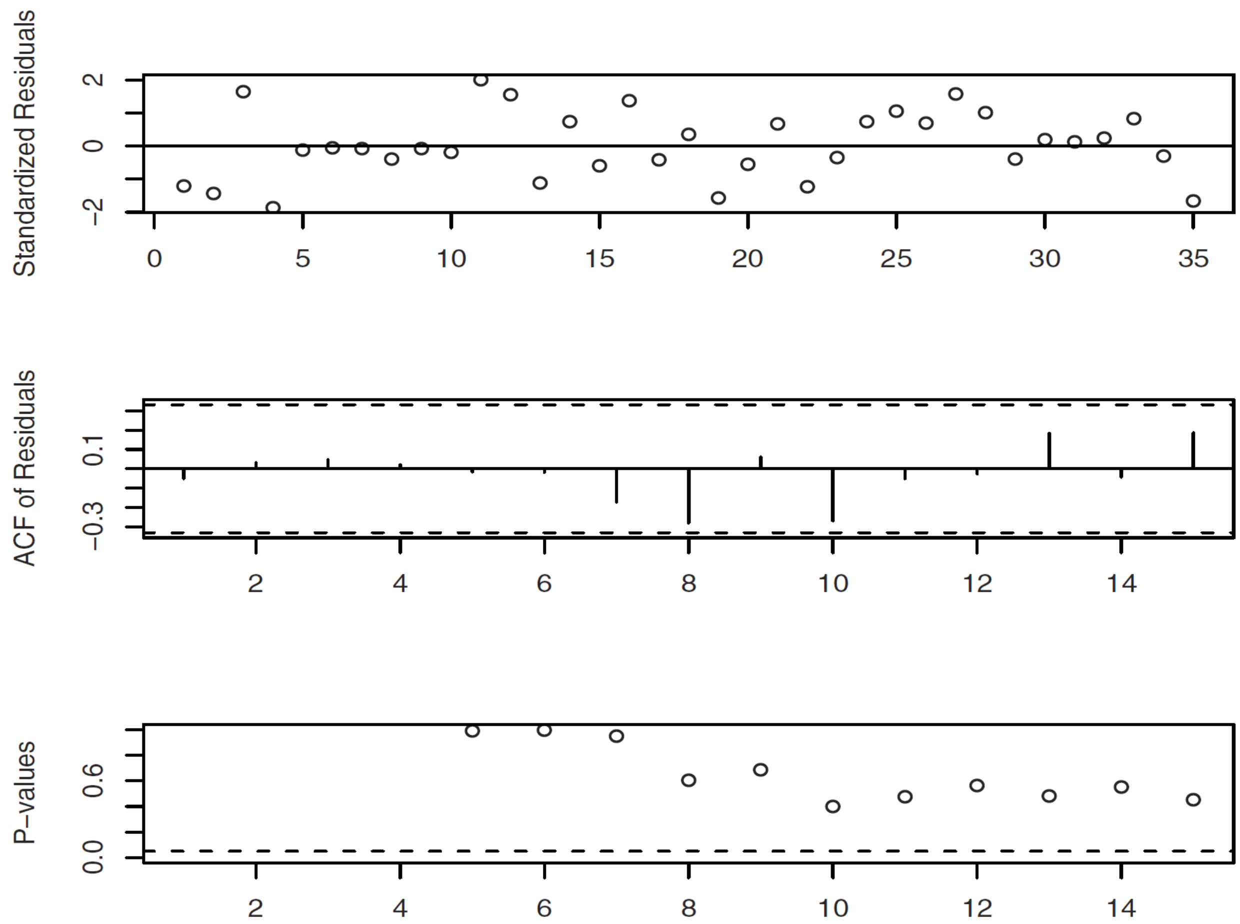
Lag k	1	2	3	4	5	6
Residual ACF	–0.051	0.032	0.047	0.021	–0.017	–0.019

```
> acf(residuals(m1.color),plot=F)$acf
> signif(acf(residuals(m1.color),plot=F)$acf[1:6],2)
> # display the first 6 acf values to 2 significant digits
```

The Ljung-Box test statistic with $K = 6$ is equal to

$$Q_* = 35(35+2)\left(\frac{(-0.051)^2}{35-1} + \frac{(0.032)^2}{35-2} + \frac{(0.047)^2}{35-3} + \frac{(0.021)^2}{35-4} + \frac{(-0.017)^2}{35-5} + \frac{(-0.019)^2}{35-6}\right) \approx 0.28$$

Exhibit 8.12 Diagnostic Display for the AR(1) Model of Color Property




```
> win.graph(width=4.875,height=4.5)
> tsdiag(m1.color,gof=15,omit.initial=F)
```

模型过拟合与不充分

- 如果增加模型阶数后，所得到新参数并不显著（残差平方和没有显著减小、似然函数没有显著增大），则可以认为没有必要增加阶数，模型过拟合。
- 如果拟合了AR(1)模型后，残差在1阶滞后处存在明显的相关性，则模型不充分，应该考虑ARMA(1,1)模型。
- 如果拟合了MA(1)模型后，残差在1阶滞后处存在明显的相关性，则模型不充分，应该考虑MA(2)模型。

模型选择与优化

- 问题提出：当一个拟合模型通过了检验，说明在一定的置信水平下，该模型能有效地拟合观察值序列的波动，在实际识别ARMA(p, q)模型时，有可能存在不止一组（p, q）值都能通过模型检验。
- 优化的目的：选择相对最优模型

AIC准则

- 显然，增加p与q的阶数，可增加拟合优度，但却同时增加了模型复杂性。因此，存在着模型的“简洁性”与模型的“拟合优度”的权衡选择问题。
- 指导思想
  - *似然函数值越大越好*
  - *未知参数的个数越少越好*
- Akaike’s Information Criterion：AIC信息准则
- $AIC = -2\log(L) + 2k$
- 这里$k = p + q$ (如果有常数项, 再加1)

BIC准则

- 在样本容量趋于无穷大时，由AIC准则选择的模型不收敛于真实模型，它通常比真实模型所含的未知参数个数要多。
- Bayesian Information Criterion：BIC信息准则
- $BIC = -2\log(L) + k\log(n)$
- $k = p + q$ (如果有常数项,再加1)

选取原则

- 在选择可能的模型时，AIC与BIC越小越好。
- 显然，如果添加的滞后项没有解释能力，则对似然函数的增大没有多大帮助，却增加了参数的个数，因此使得AIC或BIC的值增加。
- 需注意的是，在不同模型间进行比较时，必须选取相同的时间段。
- 另外，建模的目的是为了预测，在有多个模型都通过模型检验时，可以通过在实际预测中的表现来选择最优的模型。

Exhibit 8.13 AR(1) Model Results for the Color Property Series

Coefficients:[†]	ar1	Intercept[‡]
	0.5705	74.3293
s.e.	0.1435	1.9151

sigma^2 estimated as 24.83: log-likelihood = -106.07, AIC = 216.15

[†] `m1.color` # R code to obtain table
[‡] Recall that the intercept here is the estimate of the process mean μ—not θ_0.

Exhibit 8.14 AR(2) Model Results for the Color Property Series

Coefficients:	ar1	ar2	Intercept
	0.5173	0.1005	74.1551
s.e.	0.1717	0.1815	2.1463

sigma^2 estimated as 24.6: log-likelihood = -105.92, AIC = 217.84

```
> arima(color,order=c(2,0,0))
```

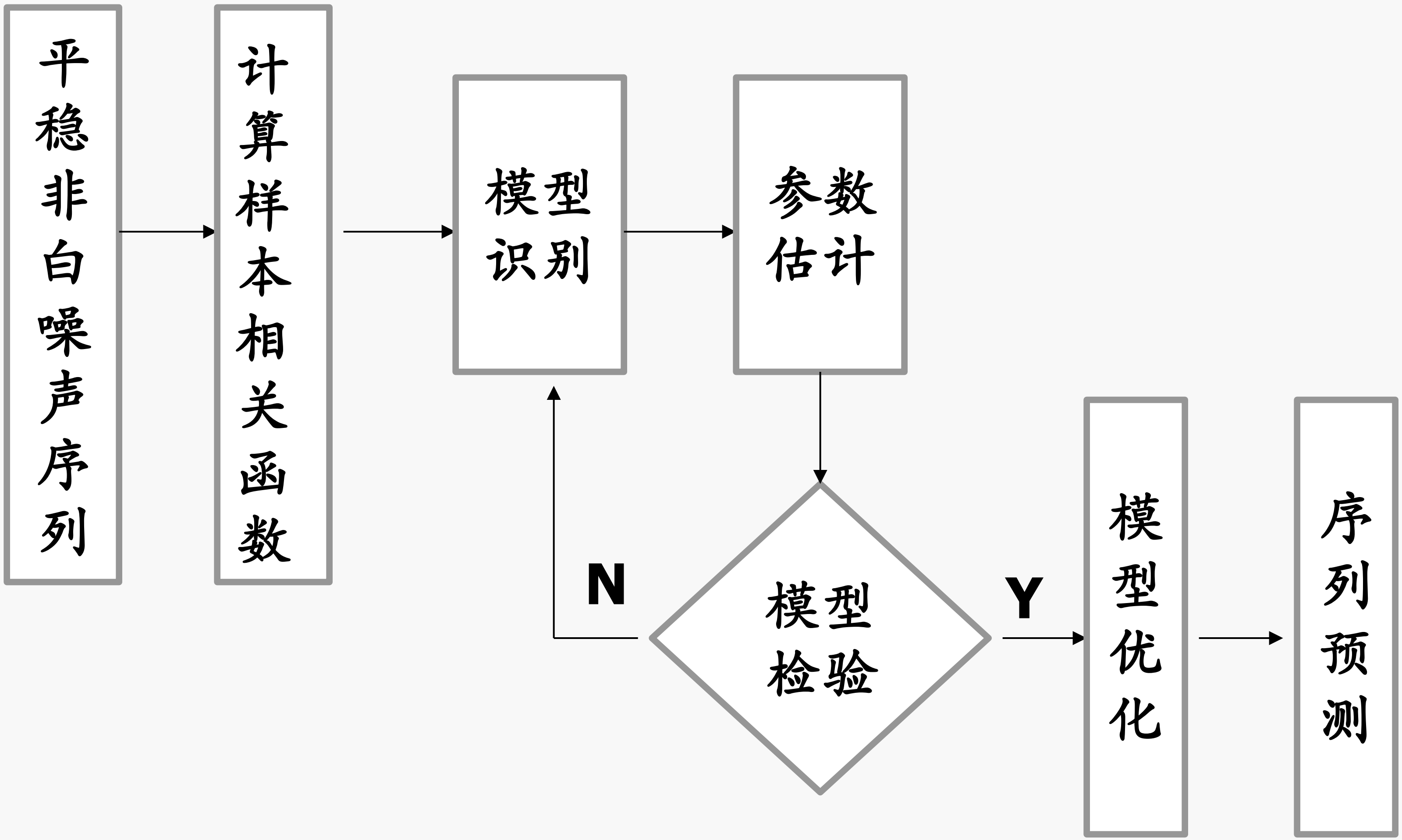
Exhibit 8.15 Overfit of an ARMA(1,1) Model for the Color Series

Coefficients:	ar1	ma1	Intercept
	0.6721	–0.1467	74.1730
s.e.	0.2147	0.2742	2.1357

sigma^2 estimated as 24.63: log-likelihood = -105.94, AIC = 217.88

```
> arima(color,order=c(1,0,1))
```

建模流程



ARMA模型预测

- 条件数学期望
- 最小均方误差预测
- ARIMA模型预测

条件数学期望

- X, Y皆为离散型随机变量, 则Y 对于给定$X = x$的条件数学期望定义为：$E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x)$
- 对于一般的函数$h(x)$有：

$$E(h(Y)|X = x) = \sum_y h(y) \cdot p_{Y|X}(y|x)$$

- 特别的：

$$Var(Y|X = x) = \sum_y \big(y - E(Y|X = x)\big)^2 \cdot p_{Y|X}(y|x)$$

$$= E(Y^2|X = x) - \big(E(Y|X = x)\big)^2$$

条件数学期望

- X,Y皆为连续型随机变量, 则Y 对于给定$X=x$的条件数学期望定义为: $E(Y|X=x)=\int_{-\infty}^{\infty} y\cdot f_{Y|X}(y|x)dy$
- 对于一般的函数$h(x)$有:

$$E(h(Y)|X=x)=\int_{-\infty}^{\infty} h(y)\cdot f_{Y|X}(y|x)\,dy$$

- 特别的:

$$Var(Y|X=x)=\int_{-\infty}^{\infty}[y-E(Y|X=x)]^2\cdot f_{Y|X}(y|x)\,dy$$
$$=E(Y^2|X=x)-\big(E(Y|X=x)\big)^2$$

例

设随机向量(X,Y)的联合密度为:

$$f_{X,Y}(x,y)=\begin{cases}6xy(2-x-y), & 0\le x\le 1, 0\le y\le 1\\ 0, & 其它\end{cases}$$

求：$E[Y|X=x], Var(Y|X=x)$

条件数学期望的两重性

(1) $E[Y|X] = g(X)$

其中，$g(x) = E[Y|X = x]$

(2)$E[Y|X_1, \cdots, X_n] = g(X_1, \cdots, X_n)$,

其中，$g(x_1, \cdots, x_n) = E[Y|X_1 = x_1, \cdots, X_n = x_n]$

条件数学期望的性质

- 全期望公式: $E[E[Y|X]]=E[Y]$

$$E[E[Y|X_1,\cdots,X_n]]=E[Y]$$

- 线性：$E[1|Y_1,\cdots,Y_n]=1$，对任意常数a,b，有

$$E[aY_1+bY_2|X_1,\cdots,X_n]$$
$$=aE[Y_1|X_1,\cdots,X_n]+bE[Y_2|X_1,\cdots,X_n]$$

- 独立公式：如果Y与$X_1,\cdots,X_n$相互独立，则

$$E[Y|X_1,\cdots,X_n]=E[Y]$$

- 分解公式：对任意n元连续函数f，有

$$E[f(X_1,\cdots,X_n)Y|X_1,\cdots,X_n]$$
$$=f(X_1,\cdots,X_n)E[Y|X_1,\cdots,X_n]$$

最小均方误差预测

- 我们的目标是用X来预测Y，标准为最小化均方误差，即需要选择一个函数$h(X)$，使得下式达到最小：
$$E[Y-h(X)]^2$$
- 不难证明，最小均方误差预测为
$$h(X)=E[Y|X]$$
- 同理，如果用$X_1,\cdots,X_n$来预测Y，最小均方误差预测为
$$h(X_1,\cdots,X_n)=E[Y|X_1,\cdots,X_n]$$

时间序列的预测

- 假设我们已知序列$Y_1, \cdots, Y_t$，预测未来l期的值Y_{t+l}，则最小均方误差预测记为
$$\hat{Y}_t(l) = E(Y_{t+l} | Y_1, \cdots, Y_t)$$
- 预测误差记为
$$e_t(l) = Y_{t+l} - \hat{Y}_t(l)$$
- 若$E[e_t(l)] = 0$，则称预测是无偏的。
- 预测误差的方差为$Var(e_t(l))$

ARIMA模型预测

- 对于可逆模型，当$j \le 0$时，$E\left[e_{t+j} \middle| Y_1, \cdots, Y_t\right] \approx e_{t+j}$
- ARIMA模型表达式两边同时对$Y_1, \cdots, Y_t$求条件期望得预测递推式:

$$\hat{Y}_t(l) = \phi_1 \hat{Y}_t(l-1) + \cdots + \phi_p \hat{Y}_t(l-p) + \theta_0 - \theta_1 e_{t+l-1}^{*} - \cdots - \theta_q e_{t+l-q}^{*}$$

- 其中$e_{t+j}^{*} = \begin{cases} e_{t+j} & j \le 0 \\ 0 & j > 0 \end{cases}$
- 用模型表达式减去预测递推式可得误差递推式

$$e_t(l) = \phi_1 e_t(l-1) + \cdots + \phi_p e_t(l-p) + e_{t+l} - \theta_1 e'_{t+l-1} - \cdots - \theta_q e'_{t+l-q}$$

- 其中$e'_{t+j} = \begin{cases} 0 & j \le 0 \\ e_{t+j} & j > 0 \end{cases}$

ARIMA模型预测

- 当$l > q$时，预测递推式为非齐次线性差分方程:

$$\hat{Y}_t(l) = \phi_1 \hat{Y}_t(l-1) + \cdots + \phi_p \hat{Y}_t(l-p) + \theta_0$$

- 回顾讲义《线性差分方程》
- 误差递推式

$$e_t(l) = \phi_1 e_t(l-1) + \cdots + \phi_p e_t(l-p) + e_{t+l} - \theta_1 e'_{t+l-1} - \cdots - \theta_q e'_{t+l-q}$$

可以简写为

$$\Phi(B) e_t(l) = \Theta(B) e'_{t+l}$$

于是

$$e_t(l) = \Psi(B) e'_{t+l} = e_{t+l} + \psi_1 e_{t+l-1} + \cdots + \psi_{l-1} e_{t+1}$$

- 误差方差为$Var(e_t(l)) = \sigma_e^2 (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)$

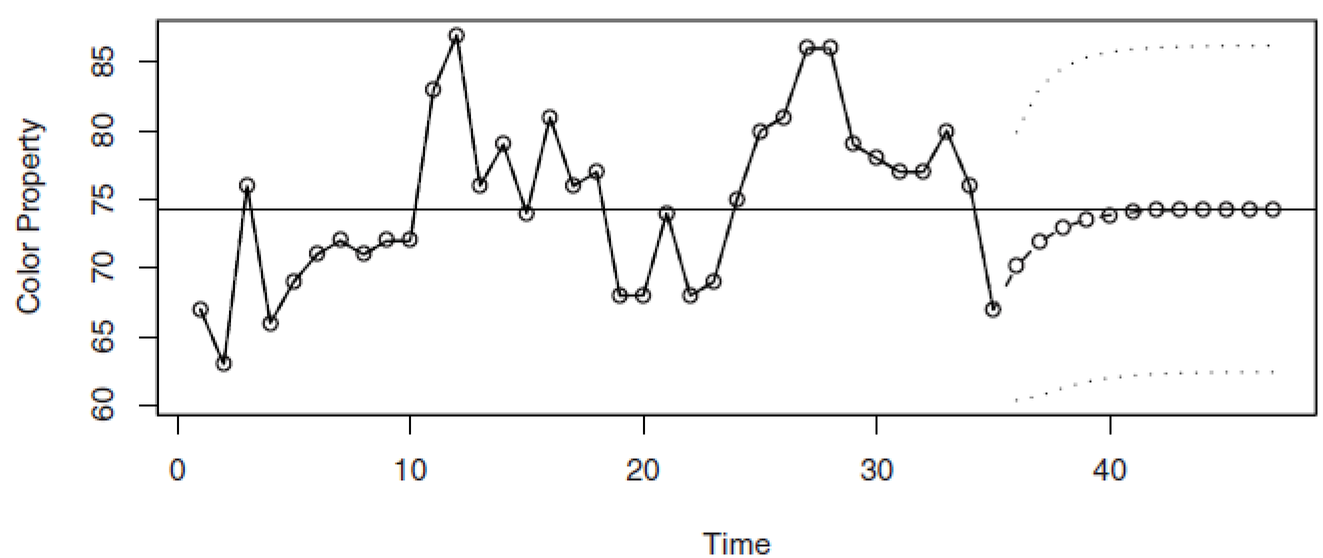
区间估计

- 已知序列$Y_1,\cdots,Y_t$，未来真实值Y_{t+l}的均值为$\hat{Y}_t(l)$，方差为$Var(e_t(l))=\sigma_e^2(1+\psi_1^2+\cdots+\psi_{l-1}^2)$
- 可以认为$\frac{Y_{t+l}-\hat{Y}_t(l)}{\sqrt{Var(e_t(l))}}$大致服从标准正态分布
- $P\left(-z_{1-\frac{\alpha}{2}}<\frac{Y_{t+l}-\hat{Y}_t(l)}{\sqrt{Var(e_t(l))}}<z_{1-\frac{\alpha}{2}}\right)\approx 1-\alpha$
- 区间估计为

$$\left(\hat{Y}_t(l)-z_{1-\frac{\alpha}{2}}\sqrt{Var(e_t(l))},\ \hat{Y}_t(l)+z_{1-\frac{\alpha}{2}}\sqrt{Var(e_t(l))}\right)$$

例

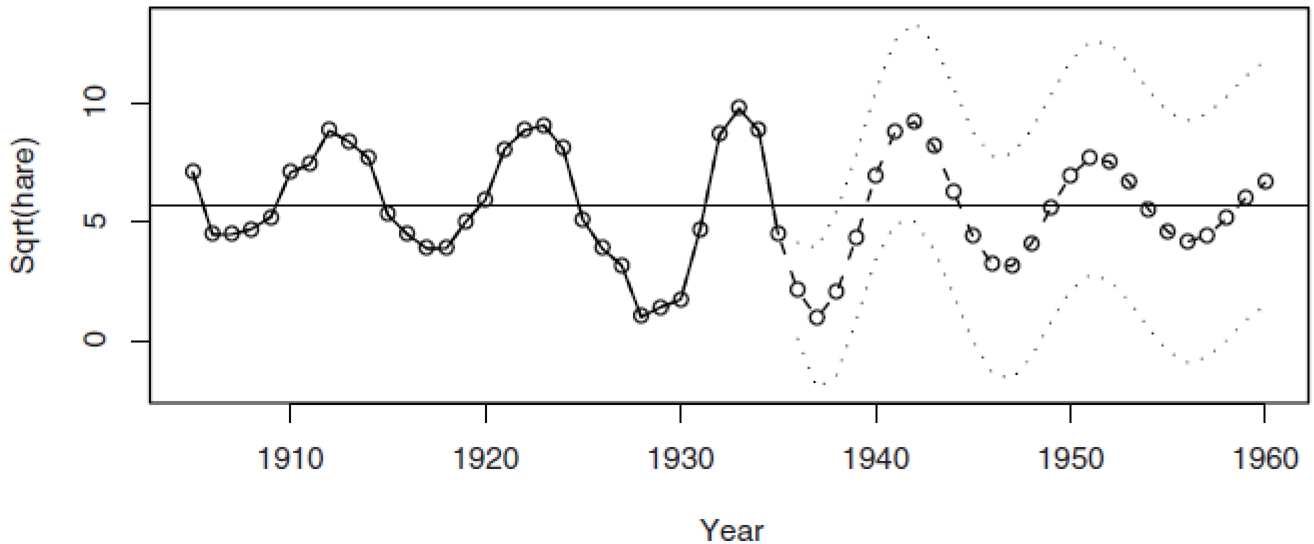
Exhibit 9.3 Forecasts and Forecast Limits for the AR(1) Model for Color




```
> data(color)
> m1.color=arima(color,order=c(1,0,0))
> plot(m1.color,n.ahead=12,type='b',xlab='Time',
   ylab='Color Property')
> abline(h=coef(m1.color)[names(coef(m1.color))=='intercept'])
```

例

Exhibit 9.4 Forecasts from an AR(3) Model for Sqrt(Hare)




```
> data(hare)
> m1.hare=arima(sqrt(hare),order=c(3,0,0))
> plot(m1.hare, n.ahead=25,type='b',
   xlab='Year',ylab='Sqrt(hare)')
> abline(h=coef(m1.hare)[names(coef(m1.hare))=='intercept'])
```