统计学与 R 语言

第 25 讲　时间序列分析 II：SARIMA

张敬信

2022 年 5 月 23 日

哈尔滨商业大学

一. 几种典型的随机过程

(1) 白噪声过程

若 $E(y_t) = 0;\ \mathrm{Var}(y_t) = \sigma^2 < \infty;\ \mathrm{cov}(y_t, y_{t+k}) = 0, k \neq 0$，则 $\{y_t : t = 1, \cdots, T\}$ 称为白噪声过程。

(2) 随机游走过程

若 $y_t = y_{t-1} + \varepsilon_t$，其中 ε_t 为白噪声，则 y_t 称为随机游走过程。

（3）p 阶自回归过程 AR(p)

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots \phi_p y_{t-p} + \varepsilon_t \qquad \text{(AR)}$$

其中，ϕ_i 为自回归参数，ε_t 为白噪声。

记 $\Phi_p(\mathrm{L}) = \mathrm{I} - \phi_1 \mathrm{L} - \cdots - \phi_p \mathrm{L}^p$，则 AR(p) 可表示为 $\Phi_p(\mathrm{L}) y_t = \varepsilon_t$.

AR(p) 模型的自相关系数具有拖尾性；AR(p) 模型的偏自相关系数具有 p 阶截尾性。

（4）q 阶移动平均过程 MA(q)

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \qquad \text{(MA)}$$

其中，θ_j 为移动平均参数，ε_t 为白噪声。

记 $\Theta_q(\mathrm{L}) = I - \theta_1 \mathrm{L} - \cdots - \theta_q \mathrm{L}^q$，则 MA(q) 可表示为 $y_t = \Theta_q(\mathrm{L})\varepsilon_t$.

MA(q) 模型的自相关系数具有 q 阶截尾性；MA(q) 模型的偏自相关系数具有拖尾性。

二. 从 ARMA 到 SARIMA 模型

1. 自回归移动平均模型 ARMA(p,q)[1]

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

即由自回归和移动平均两部分共同构成的随机过程，也可表示为

$$\Phi_p(\mathrm{L}) y_t = \Theta_q(\mathrm{L}) \varepsilon_t \qquad \text{(ARMA)}$$

其中，$\Phi_p(\mathrm{L})$ 和 $\Theta_q(\mathrm{L})$ 分别表示 L 的 p, q 阶特征多项式。

对于 ARMA(p,q) 过程，

- 其平稳性只依赖于其自回归部分，即 $\Phi_p(\mathrm{L}) = 0$ 的全部根取值在单位圆之外（绝对值大于 1）；
- 其可逆性则只依赖于移动平均部分，即 $\Theta_q(\mathrm{L}) = 0$ 的根取值应在单位圆之外。

[1] ARMA(p,q) 模型针对的是均值为 0 的时间序列数据，若非 0，做一次平移（整体减去其样本均值）即可.

2. **差分自回归移动平均模型** ARIMA(p,d,q)

ARMA(p,q) 模型处理的平稳时间序列，若非平稳时间序列做 d 阶差分可变成平稳时间序列，接着再使用 ARMA(p,q) 模型处理，则整个模型记为 ARIMA(p,d,q) 模型：

$$\Phi_p(\mathrm{L})\Delta^d y_t = \Theta_q(\mathrm{L})\varepsilon_t \qquad \text{(ARIMA)}$$

建议使用 ARIMA 模型的自动定阶。

3. **季节性时间序列模型** SARIMA(p,d,q)×(P,D,Q)s

有些时间序列数据存在明显的周期性变化，这往往是由于季节性变化（包括季度、月度、周度等变化）或其他一些固有因素引起的，这类序列称为**季节性时间序列**。比如某地区的小时气温值序列中除了含有以天为周期的变化，还含有以年为周期的变化。在经济领域中，季节性序列更是随处可见，如季度、月度、周度时间序列等。

设季节性时间序列的变化周期为 s，即时间间隔为 s 的观测值有相似之处。首先用季节差分的方法消除周期性变化。

季节差分算子定义为

$$\Delta_s^D y_t = (\mathrm{I} - \mathrm{L}^s) y_t = y_t - y_{t-s}$$

对于非平稳季节性时间序列，有时需要进行 D 次季节差分之后才能转换为平稳的序列。接着可以建立关于周期为 s 的 P 阶自回归 Q 阶移动平均季节时间序列模型：

$$\mathrm{A}_P(\mathrm{L}^s)\Delta_s^D y_t = \mathrm{B}_Q(\mathrm{L}^s) u_t \tag{1}$$

注意： P, Q 等于 2 时，滞后算子应为 $(L^s)^2 = L^{2s}$.

对于模型 (1)，相当于假定 u_t 是平稳的、非自相关的。当 u_t 非平稳且存在 ARMA 成分时，则可以把 u_t 描述为

$$\Phi_p(\mathrm{L})\Delta^d u_t = \Theta_q(\mathrm{L})\varepsilon_t$$

其中，ε_t 为白噪声过程，p, q 分别表示非季节自回归、移动平均算子的最大阶数，d 表示 u_t 的一阶（非季节）差分次数。从而

$$u_t = \Phi_p^{-1}(\mathrm{L})\Delta^{-d}\Theta_q(\mathrm{L})\varepsilon_t \tag{2}$$

将 (2) 式代入 (1) 式，就得到季节性时间序列模型的一般表达式：

$$\Phi_p(\mathrm{L})\mathrm{A}_P(\mathrm{L}^s)\Delta^d\Delta_s^D y_t = \Theta_q(\mathrm{L})\mathrm{B}_Q(\mathrm{L}^s)\varepsilon_t \qquad \text{(SARIMA)}$$

其中，下标 P, Q, p, q 分别表示季节与非季节自回归、移动平均算子的最大滞后阶数，D, d 分别表示季节和非季节性差分次数。上式称为 $(p,d,q)\times(P,D,Q)_s$ 阶季节性时间序列模型。

- 保证 $\Delta^d\Delta_s^D y_t$ 具有平稳性的条件是：$\Phi_p(\mathrm{L})\phi_P(\mathrm{L}^s)=0$ 的根在单位圆外；
- 保证 $\Delta^d\Delta_s^D y_t$ 具有可逆性的条件是：$\Theta_q(\mathrm{L})\theta_Q(\mathrm{L}^s)=0$ 的根在单位圆外。

注：当 $P=D=Q=0$ 时，SARIMA 模型退化为 ARIMA 模型，故 ARIMA 模型是 SARIMA 模型的特例。

例如，$(1,1,1)\times(1,1,1)_{12}$ 阶月度 SARIMA 模型表示为

$$(\mathrm{I}-\phi_1\,\mathrm{L})(\mathrm{I}-\alpha_1\,\mathrm{L}^{12})(\mathrm{I}-\mathrm{L})(\mathrm{I}-\mathrm{L}^{12})y_t=(\mathrm{I}+\theta_1\,\mathrm{L})(\mathrm{I}+\beta_1\,\mathrm{L}^{12})\varepsilon_t$$

$\Delta\Delta_{12}y_t$ 具有平稳性的条件是 $|\phi_1|<1,\ |\alpha_1|<1;\ \Delta\Delta_{12}y_t$ 具有可逆性的条件是 $|\theta_1|<1,\ |\beta_1|<1.$

对季节时间序列模型的季节阶数，即周期长度 s 的识别，可以通过对实际问题的分析、时间序列图以及时间序列的相关图和偏相关图分析得到。

以相关图和偏相关图为例，如果相关图和偏相关图不是呈线性衰减趋势，而是在变化周期的整倍数时刻出现绝对值相当大的峰值并呈振荡式变化，就可以认为该时间序列可以用 SARIMA 模型描述。

4. 关于模型定阶

对于平稳非白噪声序列，根据样本自相关系数图和偏自相关系数图，利用其性质估计自相关阶数 p 和移动平均阶数 q，称为 ARMA(p,q) 模型的定阶。能否正确定阶，是 ARMA 建模成功与否的关键。

有如下两种常用的模型定阶方法：

(1) **用 `acf()` 和 `pacf()` 函数绘制自相关图和偏自相关图，按如下原则定阶：**

- 若平稳序列的偏相关系数是拖尾的，且在某阶处落入置信限内，则选择 $p =$ 该阶数；
- 若平稳序列的自相关系数是截尾的，则选择 $q =$ 该截尾阶数；

季节模型定阶也是类似的，只不过看的是季节周期的倍数位置，例如季节周期为 $s = 12$, 需要看位置 $12, 24, 36, \cdots$ 若这些位置出现绝对值相当大的峰值并呈振荡式变化，说明有季节性因素，应采用 SARIMA 模型。

(2) **网格搜索 + 模型评估指标.**

用网格搜索的方式：分别拟合各阶数（不超过 3 阶）组合的模型，再根据模型评估指标来选出最优模型。

模型评估指标通常采用 AIC 和 SBC 信息准则。

- AIC 准则（最小信息量准则）或修正的 AICc 由 Akaike 提出，是一种考评综合最优配置的指标，它是拟合精度和参数未知个数的加权函数：

$$\mathrm{AIC} = -2\ln(\text{模型中极大似然函数值}) + 2 \text{ *(模型中未知参数个数)}$$

- 修正的 AIC：

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2\cdot \text{模型未知参数个数}\cdot(\text{模型未知参数个数}+1)}{T-(\text{模型未知参数个数}+1)}$$

- BIC/SBC 准则

AIC 准则未充分考虑样本量的影响会有偏差，Akaike 又提出 BIC 准则，它与 Schwartz 根据贝叶斯理论提出 SBC 准则相同，即将未知参数个数的惩罚权重由常数 2 变成了 $\ln n$:

$$\mathrm{BIC} = -2\ln(\text{模型中极大似然函数值}) + \ln n \text{ *(模型中未知参数个数)}$$

AIC 或 BIC/SBC 都是越小越好，选择其值达到最小的模型作为最优模型。

- 准确度指标

时间序列预测相当于是一种回归，也可以考虑用评估回归的准确度指标，如 RMSE、MAE 等。

手动建立 SARIMA 模型的步骤：

(1) 首先要确定 d, D. 通过差分和季节差分把原序列变换为一个平稳的序列，令 $x_t = \Delta^d \Delta_s^D y_t$

(2) 然后用 x_t 建立季节 ARMA 模型。

注意：用对数的季节时间序列数据建模时通常 D 不会大于 1，P 和 Q 不会大于 3；季节时间序列模型参数的估计、检验与前面介绍的估计、检验方法相同。利用乘积季节模型预测也与上面介绍的预测方法类似。

注：更建议直接使用 R 包提供的自动定阶。

三. 案例继续：出口额数据 SARIMA 建模

1. **读入数据，转化 tsibble**

```
library(tidyverse)
library(fpp3)
library(lubridate)
df = readxl::read_xlsx("datas/export_datas.xlsx") %>%
  mutate(Time = ymd(str_c(Time, "1 日")) |> yearmonth()) %>%
  as_tsibble(index = Time)
```

⑴ 差分定阶（对方差齐序列）

```
df1 = df %>%
  mutate(export = log(export))
unitroot_ndiffs(df1$export)          # 差分阶数
#> ndiffs
#>      1
unitroot_nsdiffs(df1$export)         # 季节差分阶数
#> nsdiffs
#>       0
```

故 $d = 1,\ D = 0.$

(2) p, q, P, Q 定阶（对平稳序列）

- **自相关系数（ACF）**

与两个连续变量的线性相关系数一样，也可以考虑时间序列 y_t 与 y_{t-1}，y_{t-2}，$\cdots$ 的线性相关系数，叫做**自相关系数**：

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}, \qquad k = 1, 2, \cdots$$

正是因为时间序列 y_t 与自己的过去序列 y_{t-1}，y_{t-2}，$\cdots$ 有线性相关性，才需要这样自回归建模：$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots$.

离得越近越可能有显著的自相关，直到不再自相关显著为止，这就需要定阶。

```r
df2 = df1 %>%
  mutate(export = difference(export))
df2 %>%
  ACF(export, lag_max = 9)
#> # A tsibble: 9 x 2 [1M]
#>     lag     acf
#>   <lag>   <dbl>
#> 1    1M -0.213
#> 2    2M -0.269
#> 3    3M  0.0404
#> 4    4M  0.134
#> # ... with 5 more rows
```

- **偏自相关系数（PACF)**

在考虑 y_t 与 y_{t-k} 的自相关系数时，若剔除掉中间 $y_{t-1}, y_{t-2}, \cdots, y_{t-k+1}$ 的影响，就是**偏自相关系数**:

$$\phi_{kk} = \frac{E(y_t - \tilde{\mu})E(y_{t-k} - \tilde{\mu}_{t-k})}{E(y_{t-k} - \tilde{\mu}_{t-k})^2}$$

其中,

$$\tilde{\mu}_t = E(y_t | y_{t-1}, y_{t-2}, \cdots, y_{t-k+1})$$

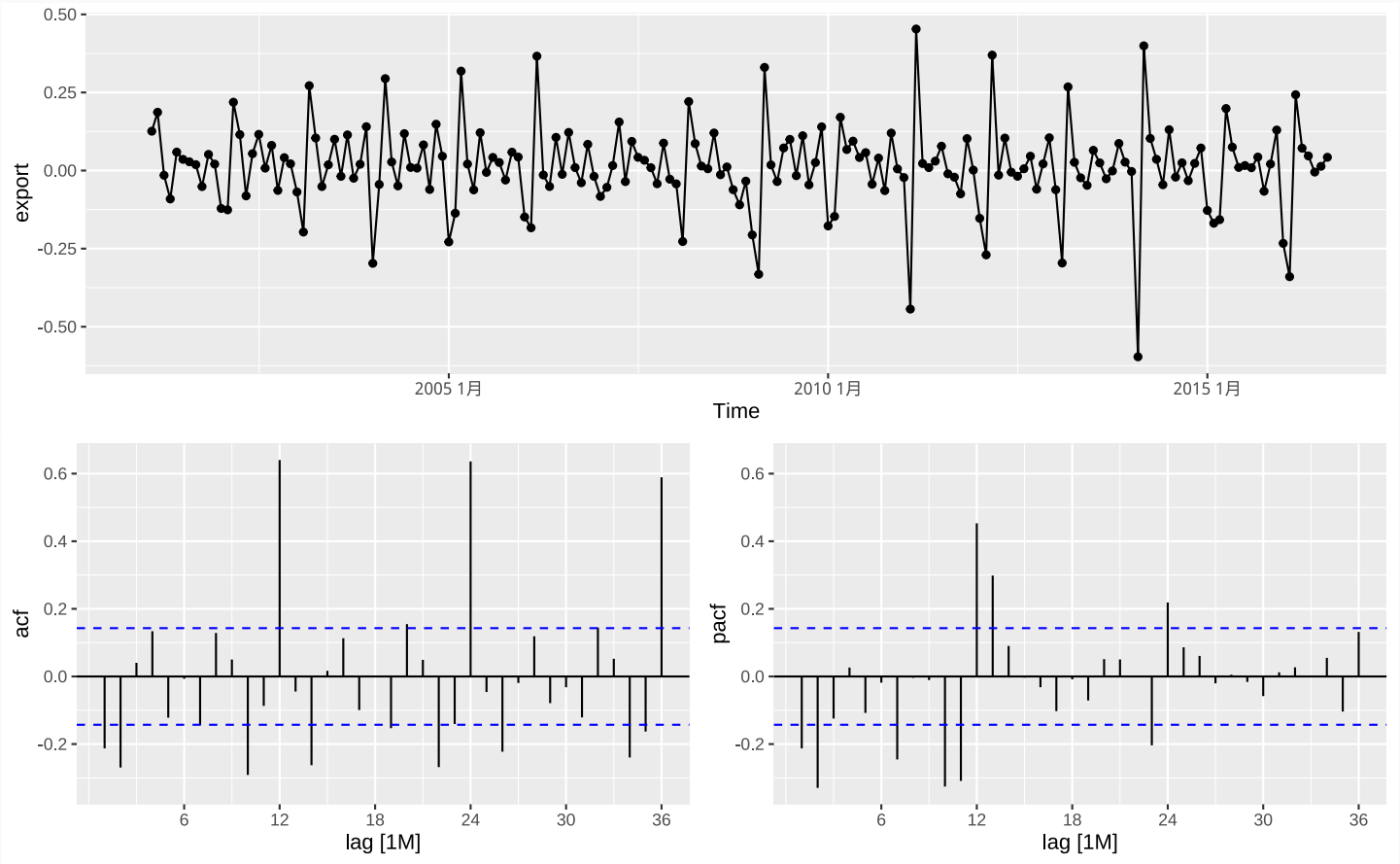
$$\tilde{\mu}_{t-k} = E(y_{t-k} | y_{t-1}, y_{t-2}, \cdots, y_{t-k+1})$$

为条件期望。

注：偏自相关系数不能通过上式直接计算，具体计算需要借助 OLS 和 Yule-Walker 方程。

```
df2 %>%
  PACF(export, lag_max = 9)
#> # A tsibble: 9 x 2 [1M]
#>      lag    pacf
#>    <lag>   <dbl>
#> 1     1M -0.213
#> 2     2M -0.330
#> 3     3M -0.125
#> 4     4M  0.0264
#> # ... with 5 more rows
```

```
df2 %>%
  gg_tsdisplay(export, plot_type = "partial", lag = 36)
```



- 两条虚线是置信限，落入置信限内就是近似为 0;
- 偏 ACF 值是拖尾的，并在 2 阶后落入置信限内，故选择 $p = 2$;
- ACF 值是截尾的，并在 2 阶后迅速跌入置信限内，故选择 $q = 2$;

故确定 ARIMA 模型阶数为 $(2, 1, 2)$.

对于季节性部分，

- 偏 ACF 值在 $12, 24$ 处落在置信限外，36 处落入置信限内，故选择 $P = 2$;
- ACF 值在 $12, 24, 36$ 处基本没有变化，故选择 $Q = 0$;

从而确定季节模型阶数为 $(2, 0, 0)_{12}$.

注：自动定阶不需要做什么，直接进入建模就行。

3. SARIMA 建模

`fpp3` 生态提供了统一的建模框架，`ARIMA()` 用于 SARIMA 建模，其中设置模型公式：

- 左端是因变量，可以对其施加任何变换；
- 右端通过 `pdq()` 和 `PDQ()` 设置定阶参数，季节周期默认为 12，若不是，在 `PQD()` 中用参数 `period` 设置；
- 右端可以包含 1，表示模型带均值项。

下面直接对原始的时间序列，同时做手动、自动定阶建模，以比较模型性能。

注意在模型公式中对因变量做了取对数变换，以保证模型预测结果直接就与原数据同量级。

```
fit = df %>%
  model(
    manual = ARIMA(log(export) ~ 1+pdq(2,1,2)+PDQ(2,0,0)),
    auto = ARIMA(log(export)))
fit
#> # A mable: 1 x 2
#>                                     manual
#>                                   <model>                      <mo
#> 1 <ARIMA(2,1,2)(2,0,0)[12] w/ drift> <ARIMA(2,1,3)(0,1,1)[
```

- 评估两个模型的性能

```
glance(fit)
#> # A tibble: 2 x 8
#>   .model  sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_r
#>   <chr>    <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <lis
#> 1 manual 0.00649    202. -388. -387. -362. <cpl [26]> <cpl
#> 2 auto   0.00578    202. -389. -389. -367. <cpl [2]>  <cpl
accuracy(fit)          # 评估模型的准确度
#> # A tibble: 2 x 10
#>   .model .type       ME  RMSE   MAE    MPE  MAPE  MASE RMS
#>   <chr>  <chr>    <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <db
#> 1 manual Training -3.62  99.5  62.5 -0.434  5.63 0.369 0.5
#> 2 auto   Training -8.99  94.9  56.8 -0.875  4.94 0.336 0.4
```

可见，两个模型的 AICc 值和准确度指标基本相当，自动定阶模型稍好一点点，但模型要更复杂，我们选择更简单的手动定阶模型：

```
mdl = select(fit, manual)
tidy(mdl)
#> # A tibble: 7 x 6
#>   .model term     estimate std.error statistic       p.val
#>   <chr>  <chr>       <dbl>     <dbl>     <dbl>         <db
#> 1 manual ar1        0.142     0.534      0.266 0.790
#> 2 manual ar2       -0.157     0.168     -0.937 0.350
#> 3 manual ma1       -0.695     0.529     -1.31  0.191
#> 4 manual ma2        0.324     0.245      1.32  0.189
#> 5 manual sar1       0.425     0.0676     6.29  0.000000002
#> 6 manual sar2       0.424     0.0694     6.11  0.000000005
#> 7 manual constant   0.00138   0.00252    0.549 0.584
```

$SARIMA(p,d,q)\times(P,D,Q)[m]$ 模型与结果系数的对应关系为：

$$(\mathrm{I}-\underline{\phi_1}\mathrm{L}-\cdots-\underline{\phi_p}\mathrm{L}^p)(\mathrm{I}-\underline{\alpha_1}\mathrm{L}^m)(\underline{\mathrm{I}-\mathrm{L}})(\underline{\mathrm{I}-\mathrm{L}^m})(y_t-\underline{\mu})=(\mathrm{I}+\underline{\theta_1}\mathrm{L}+\cdots+\underline{\theta_q}\mathrm{L}^q)(1+\underline{\beta_1}\mathrm{L}^m)\varepsilon_t$$

ar1　arp　sar1　d=1　D=1　intercept　ma1　maq　sma1

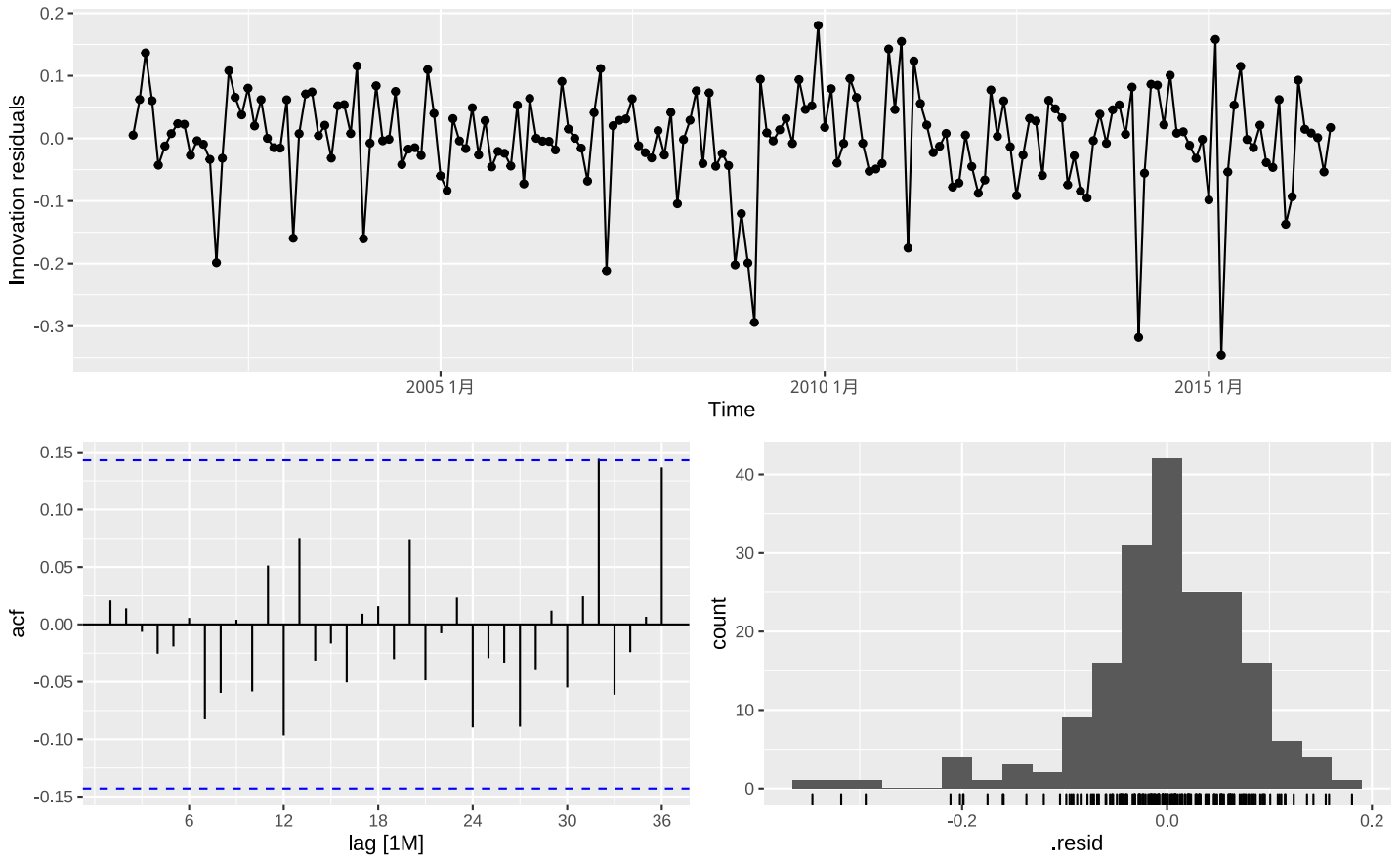
从而，可写出模型公式：

$$(I+0.142\,\mathrm{L}-0.157\,\mathrm{L}^2)(I+0.425\,\mathrm{L}^m+0.424\,\mathrm{L}^{2m})(I-\mathrm{L})\cdot(y_t-0.00138)=(I-0.695\,\mathrm{L}+0.324\,\mathrm{L}^2)\varepsilon_t$$

4. 模型检验

- 绘制残差图

```
gg_tsresiduals(mdl, lag = 36)
```



残差图和残差直方图表明残差大致服从 0 均值小标准差的正态分布（是白噪声），残差自相关图也表明，各阶自相关系数都落在置信限以内，即可认为是 0.

进一步，做残差白噪声检验（Ljung-Box 检验）：

```
augment(mdl) %>%
  features(.innov, ljung_box, lag = 24, dof = 6)
#> # A tibble: 1 x 3
#>   .model lb_stat lb_pvalue
#>   <chr>    <dbl>     <dbl>
#> 1 manual    11.3     0.882
```

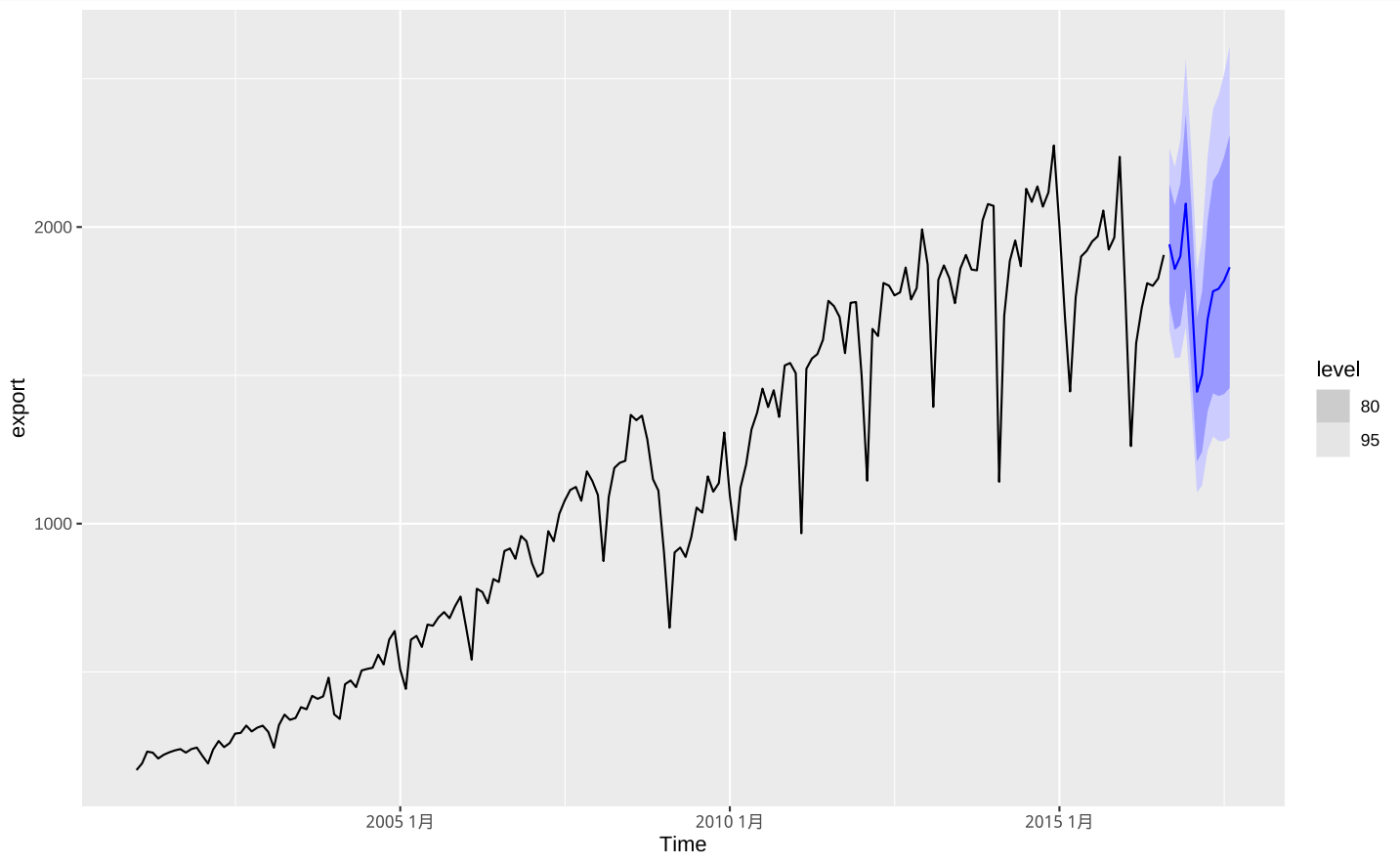
P 值 $= 0.882 > 0.05$, 接受原假设，即残差是白噪声。这表明建模成功。

注： 参数 dof 是自由度，建议设置为 $p + q + P + Q$.

5. 模型预测

```
pred = forecast(mdl, h = 12)
pred
#> # A fable: 12 x 4 [1M]
#> # Key:     .model [1]
#>   .model       Time                 export .mean
#>   <chr>       <mth>                 <dist> <dbl>
#> 1 manual  2016 9 月 t(N(7.6, 0.0065)) 1942.
#> 2 manual 2016 10 月 t(N(7.5, 0.0078)) 1859.
#> 3 manual 2016 11 月 t(N(7.5, 0.0096)) 1901.
#> 4 manual 2016 12 月  t(N(7.6, 0.012)) 2079.
#> # ... with 8 more rows
```

```
autoplot(pred, df)
```



本篇主要参阅 (张敬信, 2022), (Hyndman and Athanasopoulos, 2021), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. O Texts, 3 edition.

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

黄湘云 (2021). *Github: R-Markdown-Template*.