预测 2022 年社会消费品零售总额的损失
基于季节性 ARIMA 模型

杨在洲

yangzzh@shanghaitech.edu.cn

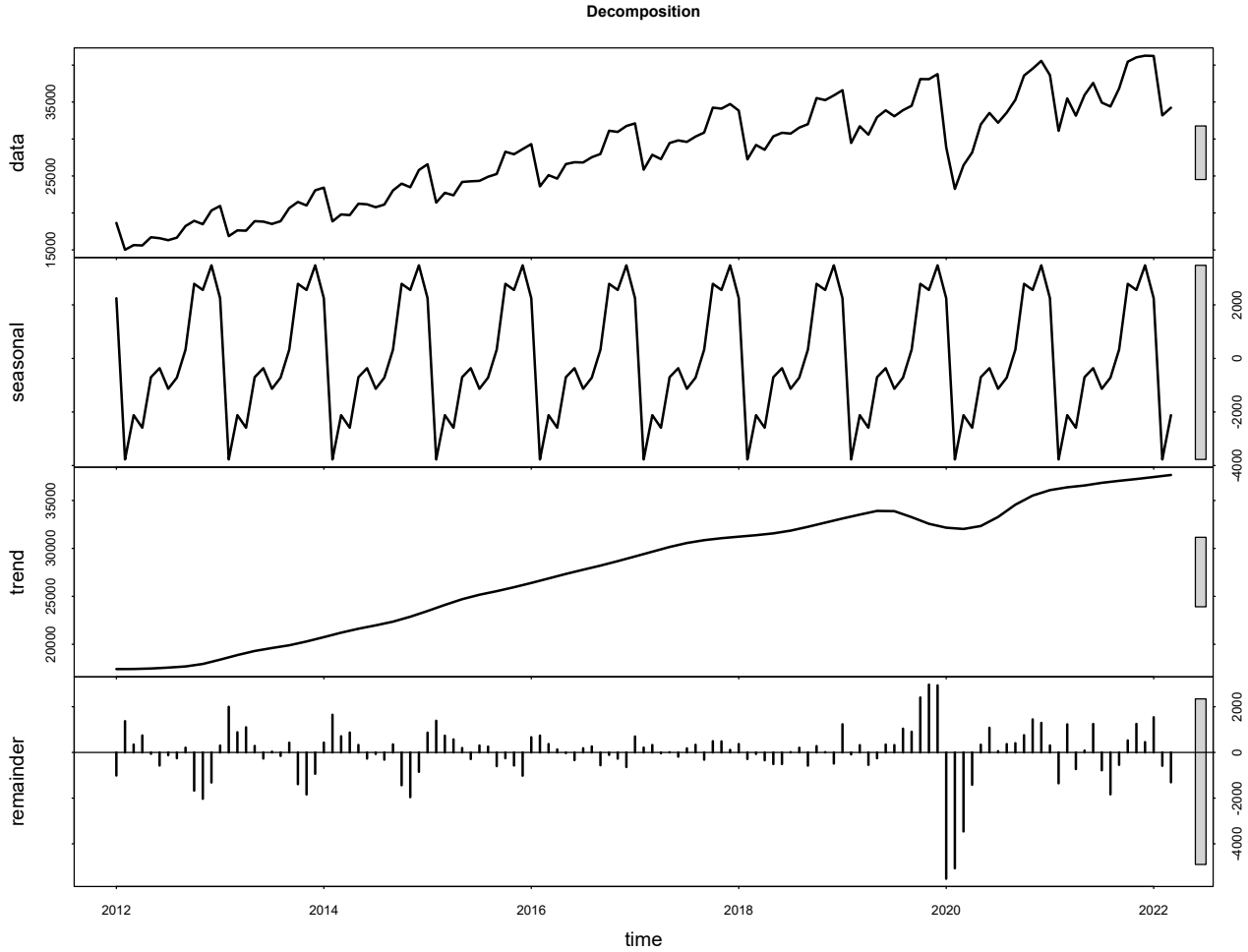
June 2, 2022

目录

数据来源

- 采用的数据为 2012 年 1 月起至 2022 年 4 月的月度数据，来自国家统计局
- 选择 2022 年 2 月份以前的数据作为训练集

季节性分解


Decomposition
data
seasonal
trend
remainder
15000
25000
35000
20000
30000
2000
0
-2000
-4000
2012
2014
2016
2018
2020
2022
time

数据检验

- 数据平稳性检验
- 数据随机性检验

平稳性检验

平稳的时间序列

若时间序列 X_t 满足如下条件：

- 均值 $E(X_t) = \mu$, 均值 μ 是与时间 t 无关的常数
- 方差 $Var(X_t) = \sigma^2$, 方差 σ 是与时间 t 无关的常数
- 协方差 $Cov(X_t, X_{t+k}) = \gamma^2$, 协方差只与间隔 t 有关

则称时间序列 X_t 是平稳的时间序列

ADF 检验

ADF 检验

$$\Delta X_t = \delta X_{t-1} + \sum_{i=1}^{m} \beta_i X_{t-i} + \epsilon_t$$

$$\Delta X_t = \alpha + \delta X_{t-1} + \sum_{i=1}^{m} \beta_i X_{t-i} + \epsilon_t$$

$$\Delta X_t = \alpha + \beta_t + \delta X_{t-1} + \sum_{i=1}^{m} \beta_i X_{t-i} + \epsilon_t$$

三个模型原假设都是 $H_0 : \delta = 0$. 若拒绝 H_0 则为平稳序列，否则为非平稳序列。通过 *ADF* 临界值表判断是否接受 H_0

ADF 检验

对原序列做 *ADF* 检验，得到结果如下:

Table: 原时序 X_t 的 ADF 检验结果

Augmented Dickey-Fuller Test	
Lag Order:	1
Dickey-Fuller:	0.3394
P Value	0.7218

$p > 0.05$ 无法拒绝原假设，原时序非平稳

ADF 检验

为去掉了原序列线性的趋势因子，对原时序 X_t 进行一阶差分得到 $\hat{X}_t$

Table: 一阶差分时序 $\hat{X}_t$ 的 ADF 检验结果

Augmented Dickey-Fuller Test	
Lag Order:	1
Dickey-Fuller:	-7.5267
P Value	0.01

由于 $p < 0.05$ 所以拒绝原假设，差分后的序列是平稳的，

随机性检验

采用 Ljung-Box 检验 $\hat{X}_t$ 随机性，假设 H_0：为对所有的 $k > 0$，样本的自相关系数服从：

$$\hat{\rho}_k \approx N(0, \frac{1}{n})$$

得到的 Ljung-Box 检验结果为：

Table: Ljung-Box 检验

Ljung-Box test	
X-squared	494.39
df	6
p-value	$<$ 2.2e-16

由于 $p < 0.05$ 所以拒绝原假设，则 ΔX_t 为非随机序列，可进行下一步建模。

ARIMA 模型

给定一个差分 d 阶的时间序列 y_t，$ARIMA(p, d, q)$ 模型如下：

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

或者写为

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t$$

其中 ε_t 是白噪声序列，p 是自回归的阶数，q 是移动平均的阶数。

自相关系数

自相关系数 ACF

$$\rho_h = \rho(y_t, y_{t+k}) = \frac{Cov(y_t, y_{t+k})}{\sigma_t \sigma_{t+k}}$$

平稳序列的自相关函数 ACF 与时间间隔 k 有关,ACF 图显示了 y_t 与 y_{t-k} 之间相关性, 可通过 ACF 相关系数决定 q.

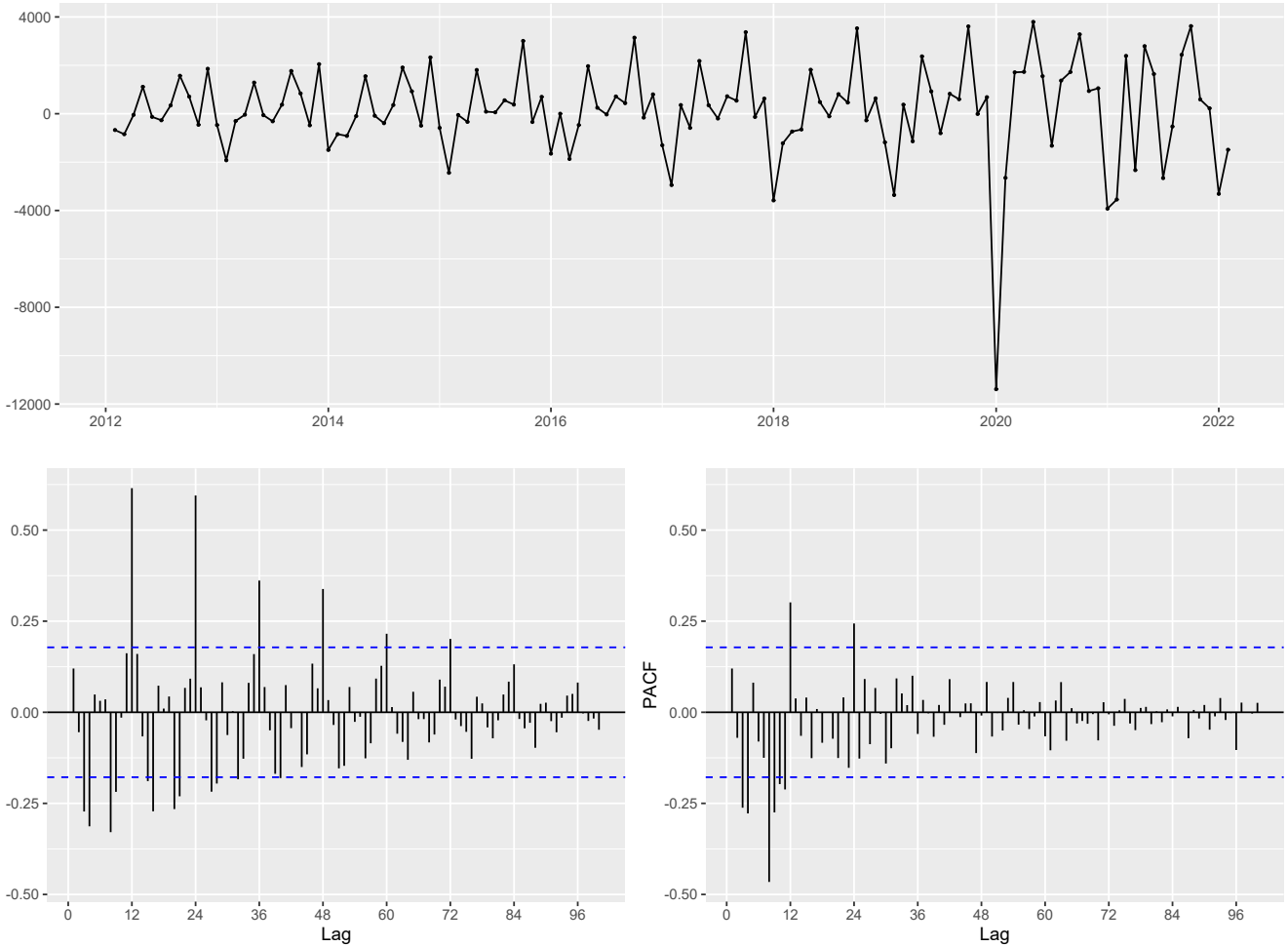
偏自相关系数

偏自相关系数 PACF

在计算相关性时移除了中间变量 $y_{t-1}, y_{t-2}, \cdots, y_{t-k+1}$ 的间接影响，直接得到 y_t 与 y_{t-k} 之间的相关性，通过 $PACF$ 估计 P 值

参数估计

根据 ACF 和 PACF 图的拖尾情况选取合适的参数 p,q:



季节性 ARIMA 模型

将一阶差分序列 ΔX_t 进行分解，写成季节，部分与非季节部分的乘积

$$\text{ARIMA} \quad (p,d,q) \qquad (P,D,Q)_m$$

例如对于 $ARIMA(1,1,1)(1,1,1)_m$ 模型:

$$(1-\phi_1 B)\,(1-\Phi_1 B^{12})\,(1-B)\,(1-B^4)y_t = (1+\theta_1 B)\,(1+\Theta_1 B^4)\varepsilon_t$$

季节性差分

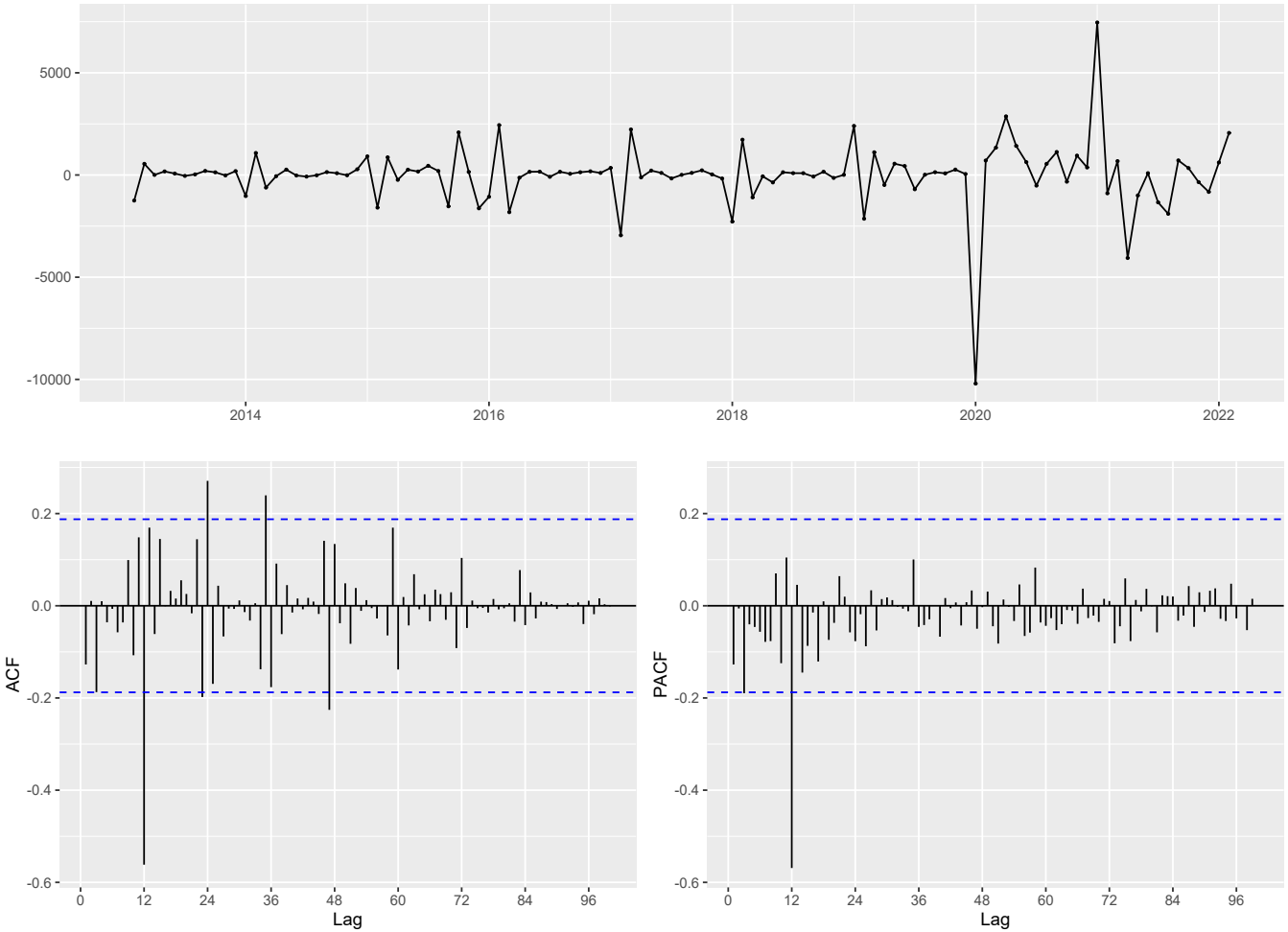
季节性差分

观测周期为 12 个月，为先消除季节型波动，对 ΔX_t 再进行差分

$$X'_t = \Delta X_t - \Delta X_{t-12}$$

季节性差分

得到序列 X_t' 的自相关图：



参数估计

参数确定

- 可确定季节部分相应系数 $P = 1, Q = 1$
- 非季节部分的 ACF 和 PACF 图较难判断产生拖尾的临界点

模型选择

利用 AIC,AICc,BIC 准则定量的确定在何种系数下的模型最优

AIC（赤池信息准则）

$$\mathrm{AIC} = -2log(L) + 2(p + q + k + 1)$$

其中 L 数据的似然函数，最后一项为参数个数 (包含了余项的方差)$k = 0$ 若 $c = 0$,$k = 1$ 若 $c \neq 0$ 对于 ARIMA 模型而言，修正过的 AIC 值可以被表示为：

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - p - q - k - 2}$$

模型选择

AICc（赤池信息量准则）

$$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}$$

BIC（贝叶斯信息准则）

$$\text{BIC} = \text{AIC} + [\log(T) - 2](p+q+k+1)$$

模型选择

通过枚举 p,q 的值得到相应模型 AIC,AICc,BIC 如下：

相应的 ARIMA 模型	AIC	AICc	BIC
(0,1,0)(1,1,1)[12]	1876.32	1876.55	1884.4
(0,1,1)(1,1,1)[12]	1878.32	1878.7	1889.08
(0,1,2)(1,1,1)[12]	1877.96	1878.54	1891.41
(0,1,3)(1,1,1)[12]	1876.22	1877.04	1892.36
(1,1,1)(1,1,1)[12]	1880.27	1880.86	1893.73
(1,1,2)(1,1,1)[12]	1873.59	1874.41	1889.74
(1,1,3)(1,1,1)[12]	1875.51	1876.62	1894.35
(2,1,1)(1,1,1)[12]	1873.53	1874.36	1889.68
(2,1,2)(1,1,1)[12]	1875.02	1876.13	1893.86
(3,1,0)(1,1,1)[12]	1879.02	1879.84	1895.16
(3,1,1)(1,1,1)[12]	1875.53	1876.64	1894.37
(3,1,2)(1,1,1)[12]	1876.98	1878.79	1901.2

残差检验

从表??中看出，ARIMA$(2,1,1)(1,1,1)_{12}$ 是最优的 ARIMA 模型。对残差 ϵ_t 做 Ljung-Box test 检验:

Table: 残差 *Ljung − Box* 检验结果

Ljung-Box test	
df	19
p-value	0.90

$p > 0.05$ 无法拒绝原假设，所得残差为白噪声序列，残差之间不存在自相关性。

残差检验

并且得到的残差图??，残差基本符合正态分布要求：

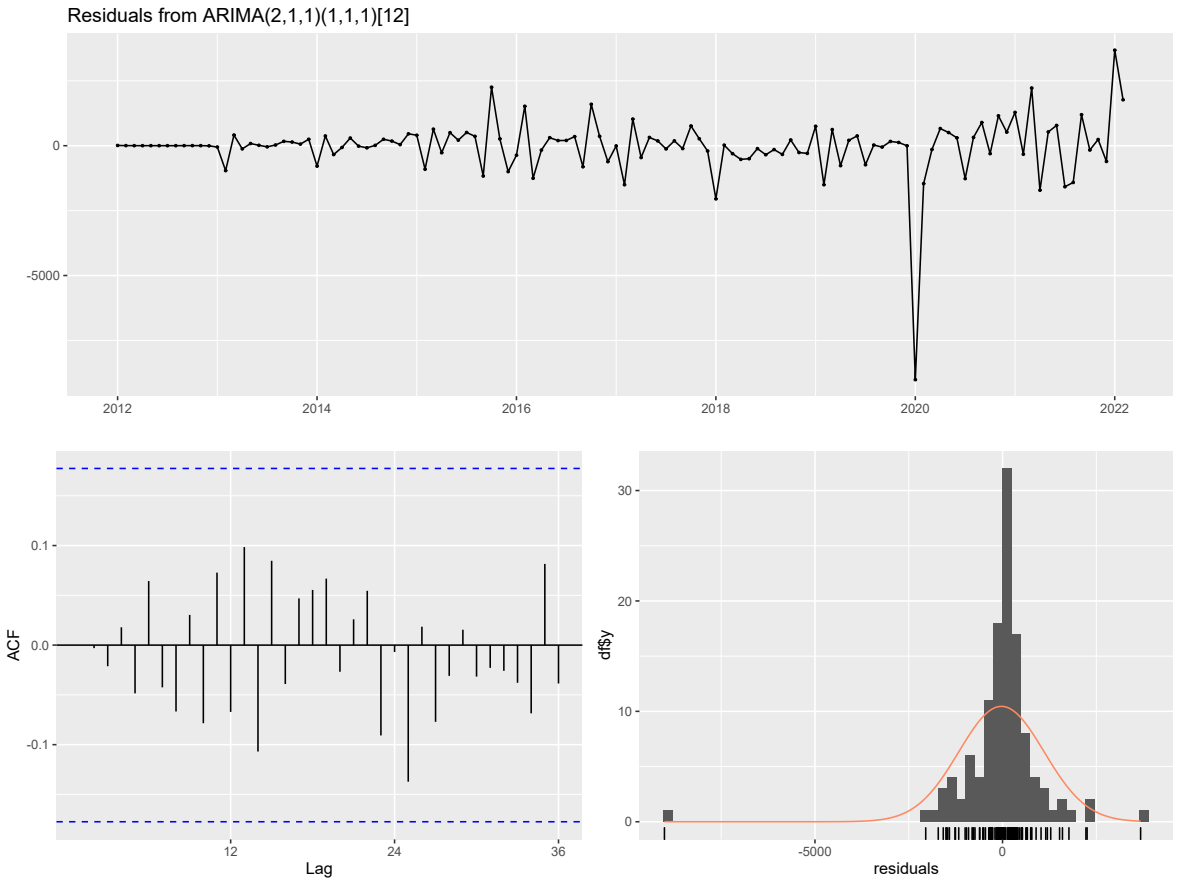


Figure: $ARIMA(2,1,1)(1,1,1)_{12}$ 的残差图

残差检验

为进一步说明，绘出正态 Q-Q 图

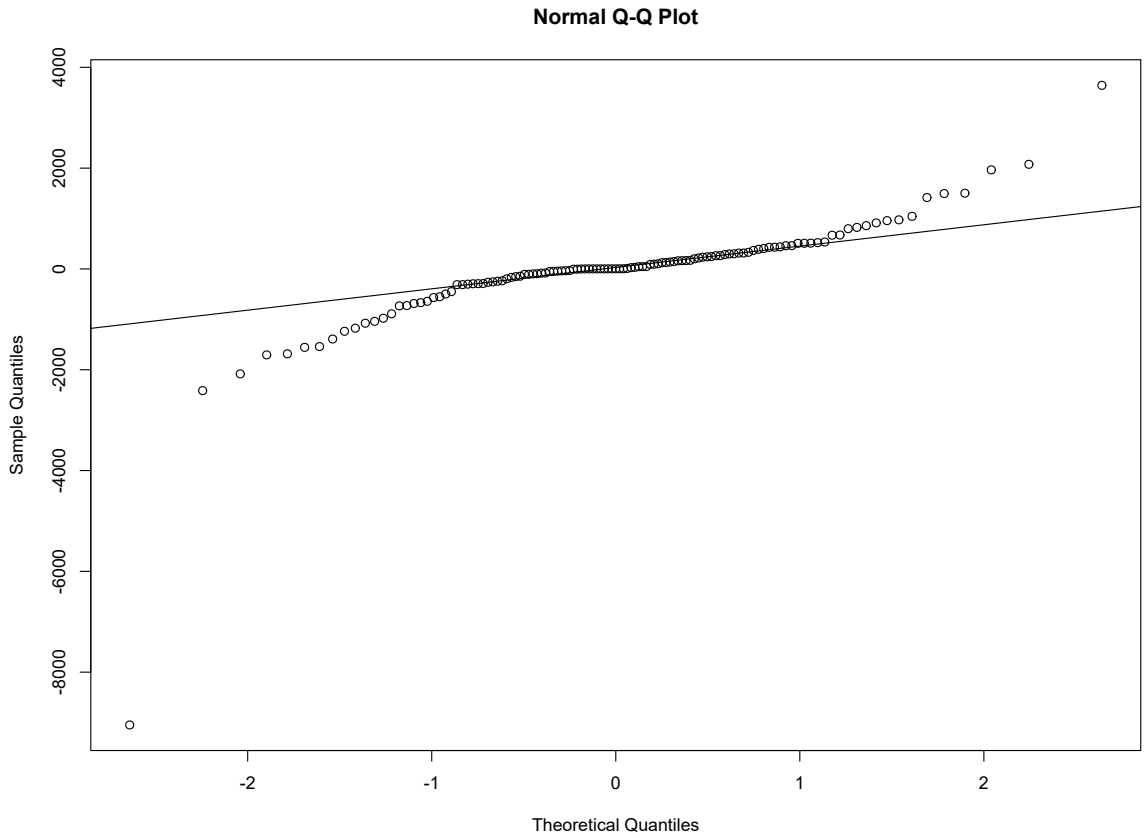


Figure: ARIMA$(2,1,1)(1,1,1)_{12}$ 的残差 Q-Q 图

预测

Figure: ARIMA 模型得到的 12 步拟合值

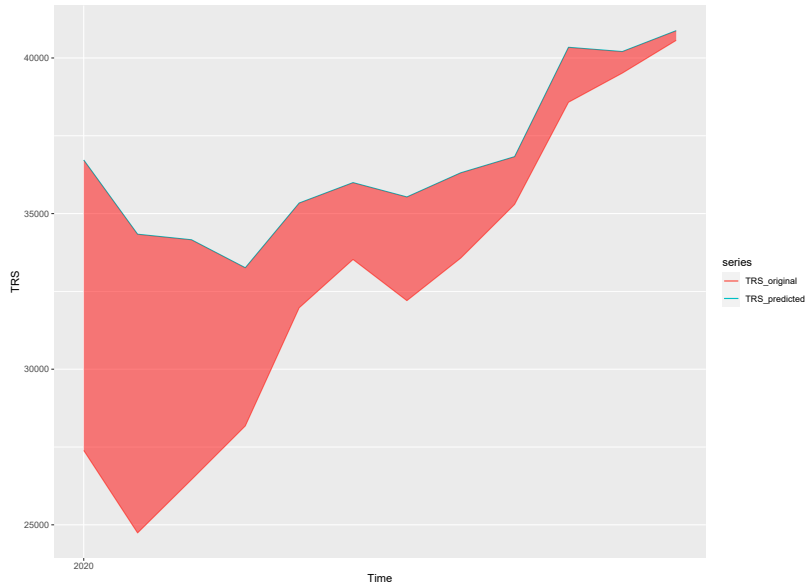


Figure: 2020 年社会消费品零售总额的损失

干预分析

持续性干预变量

$$S_t^T = \begin{cases} 0 & \text{疫情发生前 t<T} \\ 1 & \text{疫情发生后 t>=T} \end{cases}$$

干预分析模型

设 ω 为干预未知的干预系数，Z_t 为疫情发生后所产生的损失的时间序列，通过一阶差分获得平稳序列，则干预后的模型可写为

$$Z_t = \delta Z_{t-1} + \omega$$

干预模型

用最小二乘法的到参数的估计值，$\delta = 0.7328, \omega = 72.1654$ 绘出回归拟合图像和残差图如下，残差符合正态分布要求，且通过 Ljung-Box 检验。：

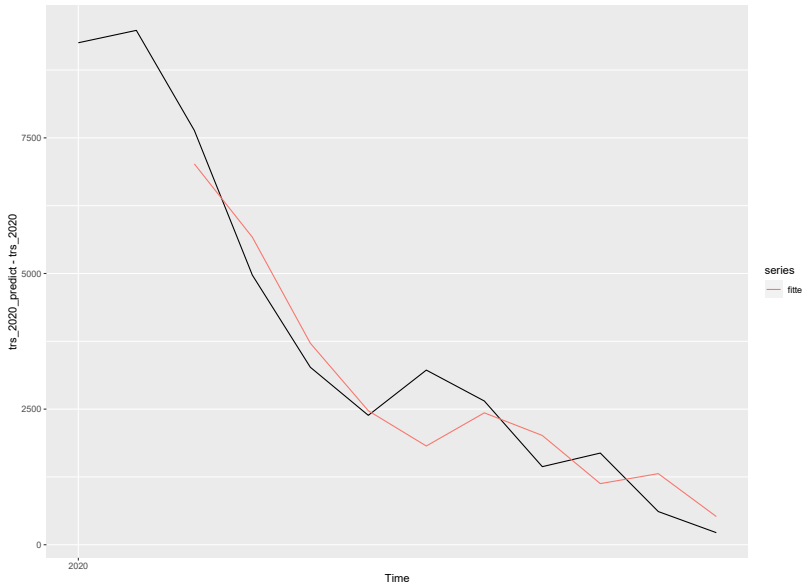


Figure: 2020 年社会消费品零售总额的损失图 (红色为回归结果)

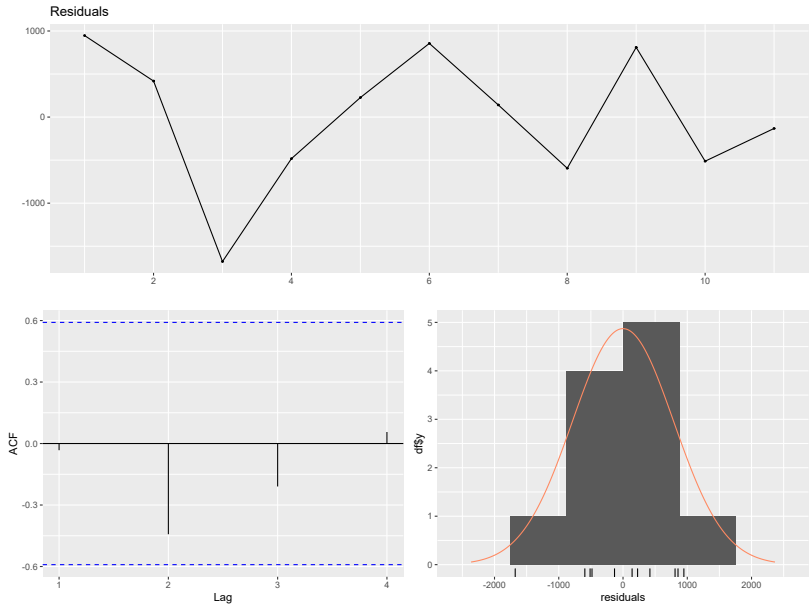


Figure: 回归结果的残差图

预测

Table: 2022 年 3 月起社会消费品零售总额的损失（单位：亿元）

月份	3	4	5	6	7	8	9	10	11	12
损失	3924	8587	6364	4735	3542	2667	2026	1557	1213	961

预测

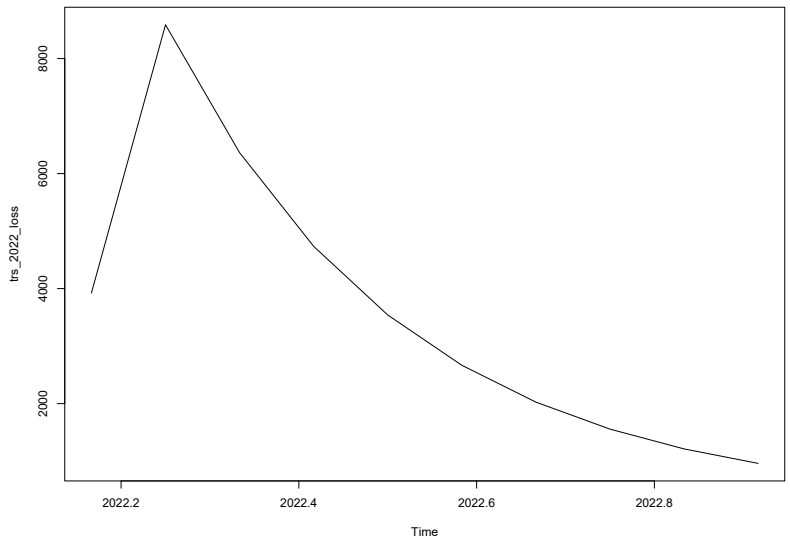


Figure: 2022 年 3 月起社会消费品零售总额的损失

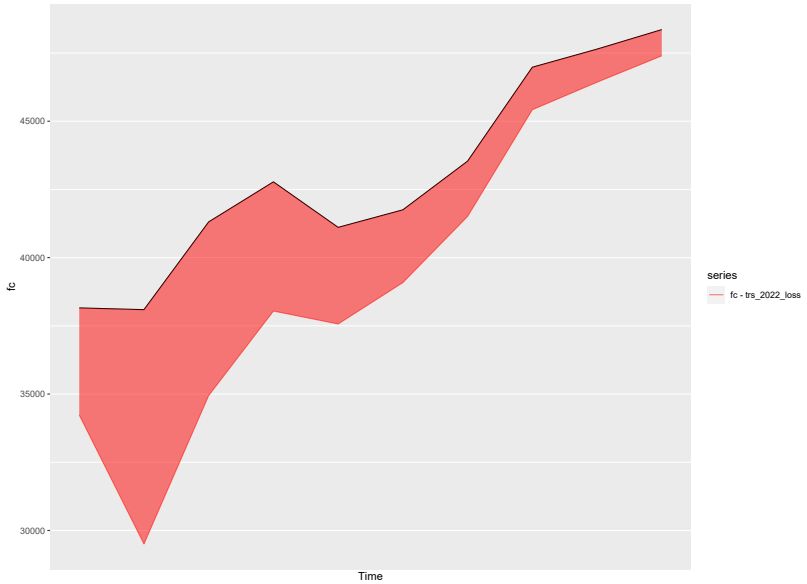


Figure: 红色面积为预测损失

损失的社会消费品零售总额共计 35576.72 亿元

参考文献