第三次练习课参考答案

黄嘉平

2025-05-27

首先调用程序包（已隐去调用后显示信息）

```
library(fpp3)
```

第一题

准备数据

```
aus_cement <- aus_production |>
  select(Cement)
```

1. 起止时间与期数

```
aus_cement |>
  slice(1,n()) # 从第一行和最后一行中了解起止时间
```

```
## # A tsibble: 2 x 2 [1Q]
##   Cement Quarter
##    <dbl>   <qtr>
## 1    465 1956 Q1
## 2   2401 2010 Q2
```

可知起始时间是 1956 年第一季度，终止时间是 2010 年第二季度。

```
aus_cement |>
  count() # 获取样本量
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   218
```

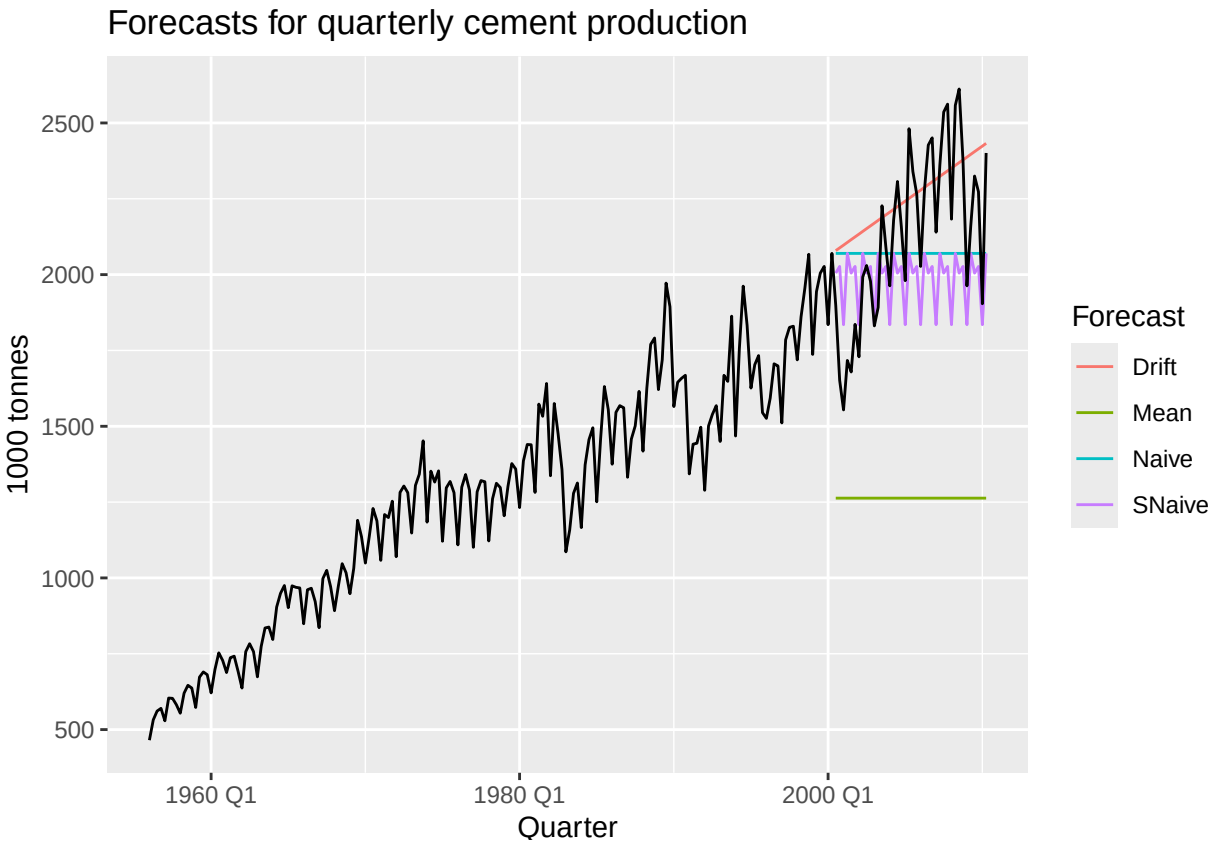
从样本量可知该样本共包含 218 期。

2. 设定训练集和测试集

```
cement_train <- aus_cement |>
  slice_head(n = 178) # 定义训练集

cement_test <- aus_cement |>
  slice_tail(n = 40) # 定义测试集
```

3. 拟合与预测

```
cement_fit <- cement_train |>
  model(
    Mean = MEAN(Cement),
    Naive = NAIVE(Cement),
    SNaive = SNAIVE(Cement),
    Drift = RW(Cement ~ drift())
  ) # 分别拟合四个模型

cement_fc <- cement_fit |>
  forecast(h = 40) # 预测未来 40 期

cement_fc |>
  autoplot(aus_cement, level = NULL) +
  labs(
    y = "1000 tonnes",
    title = "Forecasts for quarterly cement production"
  ) +
  guides(colour = guide_legend(title = "Forecast"))
```



4. 预测精确度

```
accuracy(cement_fc, cement_test)
```

```
## # A tibble: 4 x 10
##   .model .type     ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Drift  Test   -123.   249.  203. -7.16  10.4   NaN   NaN 0.447
## 2 Mean   Test    870.   912.  870. 39.7   39.7   NaN   NaN 0.640
## 3 Naive  Test     62.8  282.  241.  1.23  11.4   NaN   NaN 0.640
## 4 SNaive Test    149.   293.  251.  5.55  11.5   NaN   NaN 0.791
```

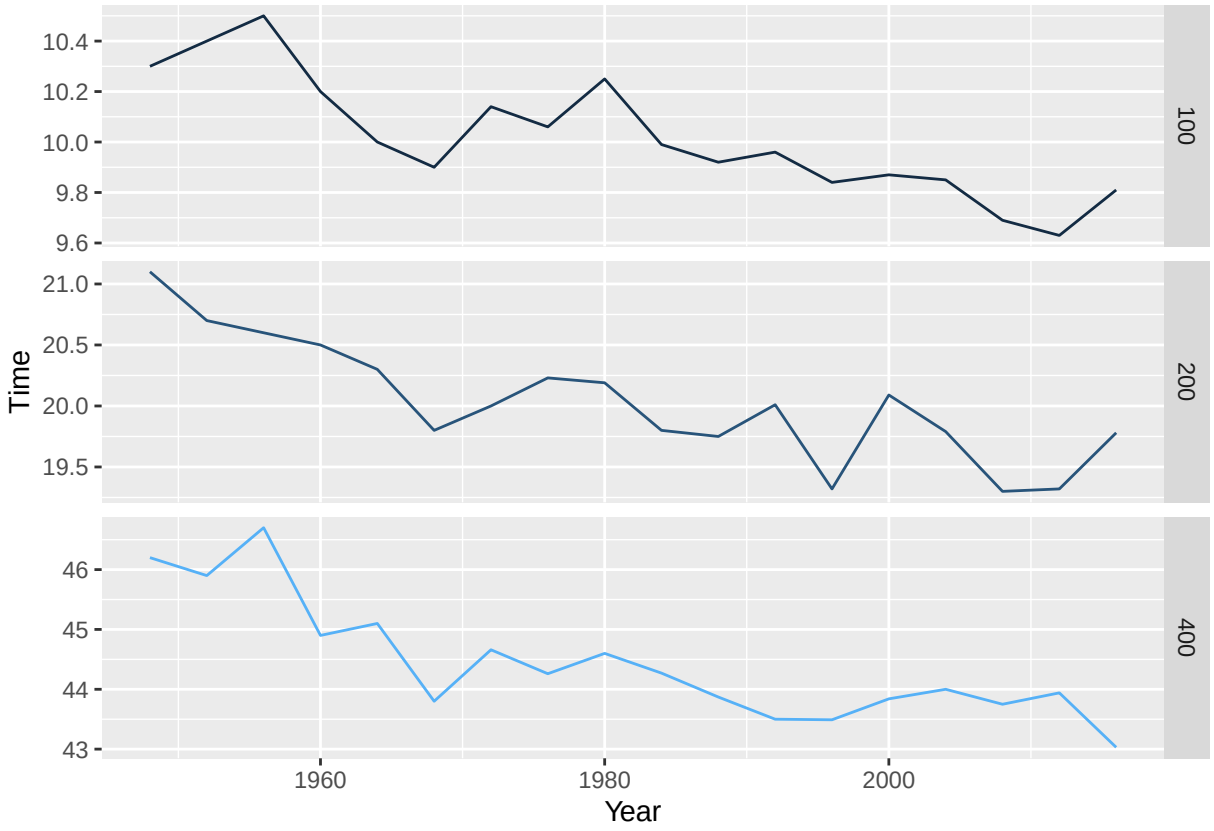
从结果中可知，RMSE 和 MAE 都显示漂移法的预测误差在四种方法中最小。

第二题

1. 选取数据并绘制时序图

```
men_short <- olympic_running |>
  filter(Sex == "men" & Year >= 1948 & Length <= 400)
```

```
men_short |>
  ggplot(aes(Year, Time, colour = Length)) +
  geom_line() +
  facet_grid(Length ~ ., scales = "free_y") +
  guides(colour = "none")
```



2. 回归分析

回归模型可以表达为

$$\text{Time}_t = \beta_0 + \beta_1 t + \varepsilon_t$$

```
short_fit <- men_short |>
  model(TSLM(Time ~ trend())) # 针对数据中每个 key 值组合进行回归

short_fit |>
  tidy() |> # report() 命令只能用于显示单一模型的结果，tidy() 命令可以显示多个模型
  filter(term == "trend()") |> # 仅显示时间项的系数
  mutate(annual_change = estimate / 4) |>  # 由于时间间隔是四年，因此年度变化应除以 4
  select(-.model)
```

```
## # A tibble: 3 x 8
##   Length Sex    term     estimate std.error statistic    p.value annual_change
##    <int> <chr> <chr>       <dbl>     <dbl>     <dbl>      <dbl>          <dbl>
```

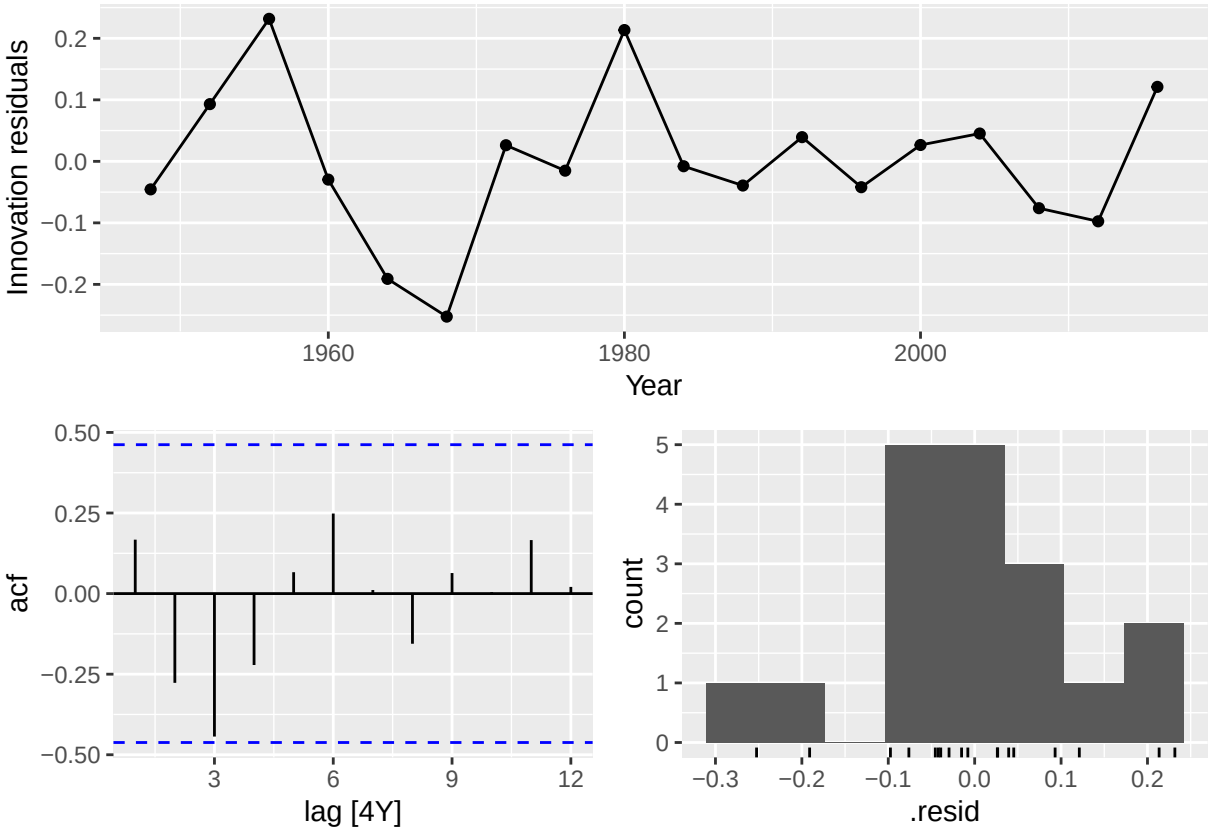
```
## 1     100 men    trend()  -0.0386    0.00567      -6.81 0.00000422       -0.00966
## 2     200 men    trend()  -0.0773    0.0126       -6.12 0.0000147        -0.0193
## 3     400 men    trend()  -0.157     0.0252       -6.23 0.0000119        -0.0393
```

由 `estimate` 列和 `annual_change` 列可知，男子 100 米的回归系数估计值为 -0.0386，平均每年缩短 0.00966 秒；男子 200 米的回归系数估计值为 -0.0773，平均每年缩短 0.0193 秒；男子 400 米的回归系数估计值为 -0.157，平均每年缩短 0.0393 秒。

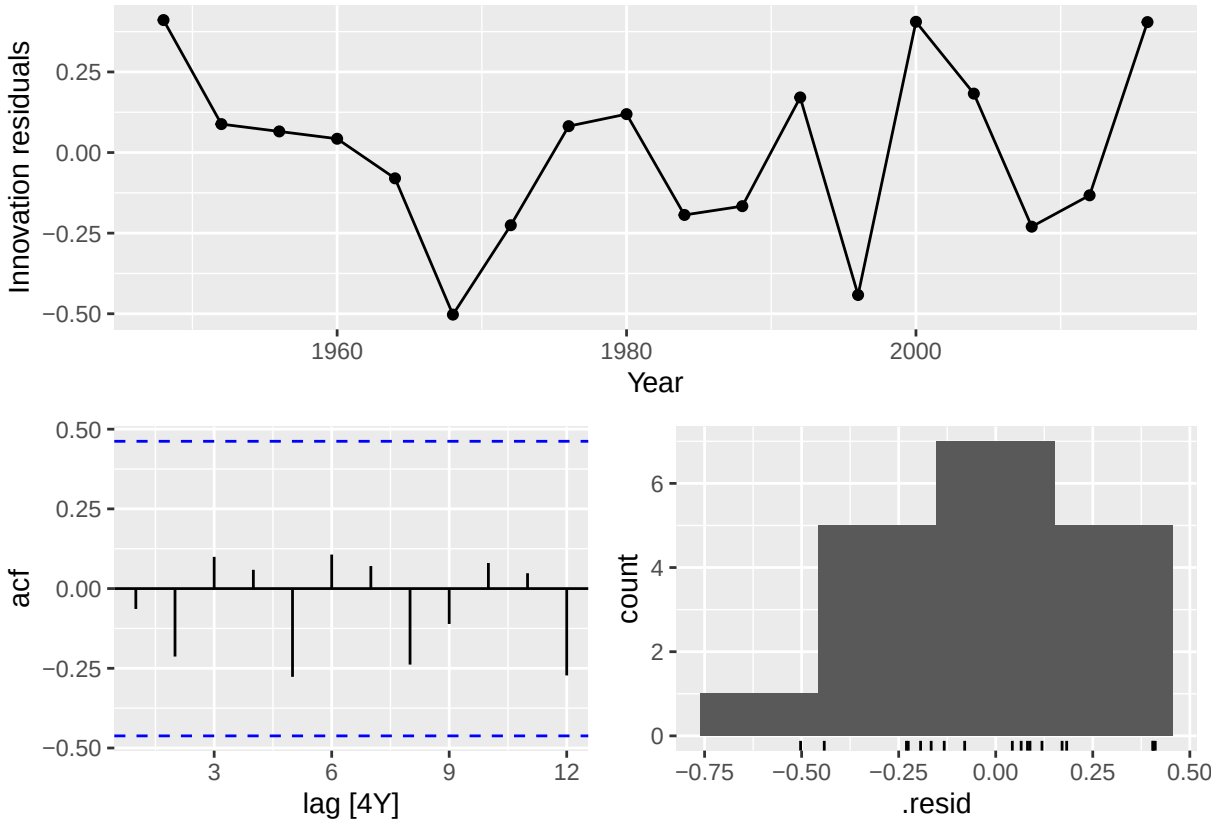
3. 残差诊断

残差诊断图

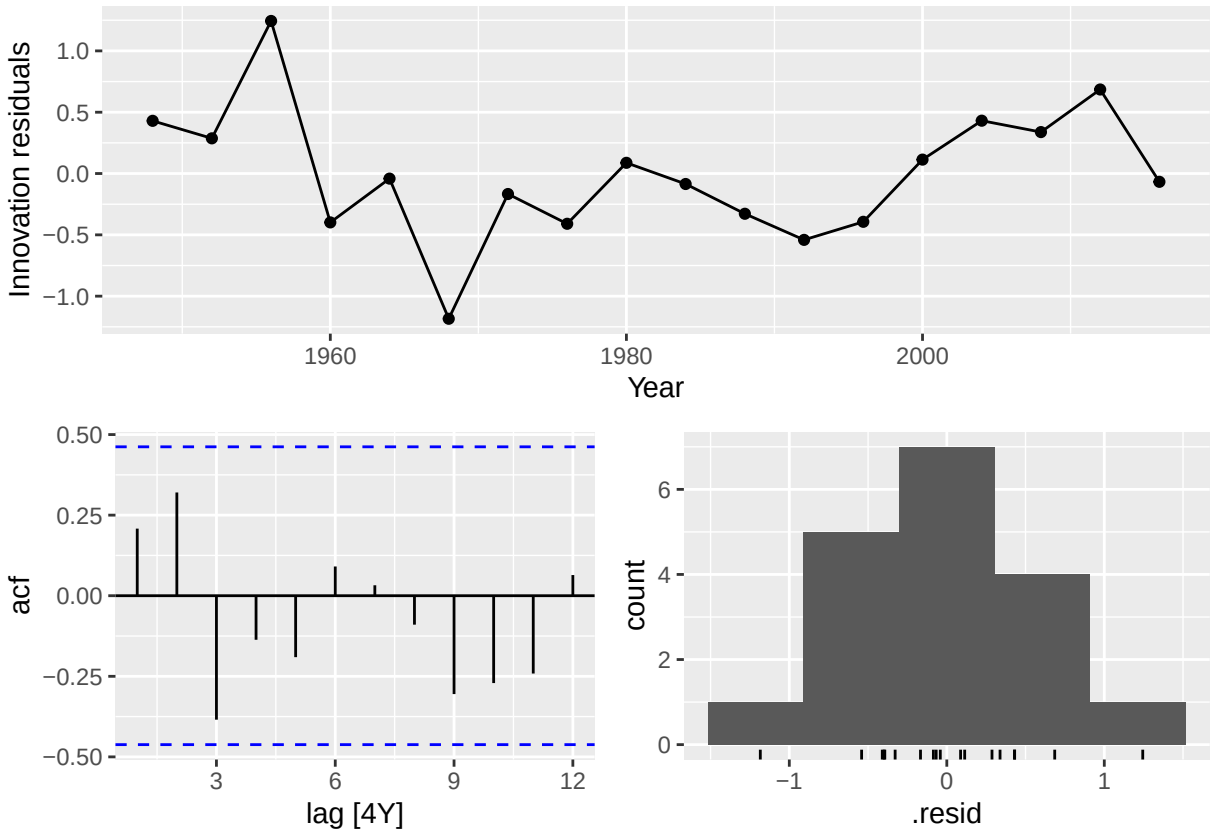
```
short_fit |>
  slice(1) |> # 男子 100 米
  gg_tsresiduals()
```




```
short_fit |>
  slice(2) |> # 男子 200 米
  gg_tsresiduals()
```




```
short_fit |>
  slice(3) |> # 男子 400 米
  gg_tsresiduals()
```



残差诊断图中的 ACF 图提示，三个回归模型的残差序列都符合白噪声的特征。

```
augment(short_fit) |>
  features(.innov, ljung_box, lag = 10)
```

```
## # A tibble: 3 x 5
##   Length Sex   .model               lb_stat lb_pvalue
##    <int> <chr> <chr>                  <dbl>     <dbl>
## 1    100 men   TSLM(Time ~ trend())   11.3      0.334
## 2    200 men   TSLM(Time ~ trend())    6.89     0.736
## 3    400 men   TSLM(Time ~ trend())   15.9      0.104
```

Ljung-Box 检验的结果显示，所有模型都无法拒绝零假设，进一步印证了残差序列符合白噪声的特征。因此，以上回归模型是合适的。

4. 预测

```
short_fc <- short_fit |>
  forecast(h = 1) # 预测未来 1 期，即 2020 年

short_fc |>
  hilo(level = c(80, 95)) |> # hilo() 函数可以计算预测区间（和均值的点预测）
  select(6:8) # 有限显示预测结果
```

```
## # A tsibble: 3 x 7 [4Y]
## # Key:       Length, Sex, .model [3]
##   .mean                  `80%`                    `95%`  Year Length Sex
##   <dbl>                 <hilo>                   <hilo> <dbl>  <int> <chr>
## 1  9.65 [ 9.471871,  9.828652]80 [ 9.377437,  9.923086]95  2020    100 men
## 2 19.3  [18.901314, 19.694895]80 [18.691266, 19.904943]95  2020    200 men
## 3 42.9  [42.147482, 43.733041]80 [41.727810, 44.152713]95  2020    400 men
## # i 1 more variable: .model <chr>
```

点预测（均值），80% 和 90% 区间预测的结果分别显示在前三列中。从 olympics.com 网站可知，2020 东京奥运会中三项冠军成绩分别为

- 男子 100 米：9.80 秒。大于点预测值，但在 80% 预测区间内。
- 男子 200 米：19.62 秒。大于点预测值，但在 80% 预测区间内。
- 男子 400 米：43.85 秒。大于点预测值，且在 80% 预测区间外，但在 95% 预测区间内。

由此可见，基于线性回归模型的预测结果一定程度上高估了成绩（实际时间没有预测时间快）。