时间序列分析与预测

第六讲

黄嘉平

深圳大学｜中国经济特区研究中心
粤海校区汇文楼办公楼 1510
课程网站 https://huangjp.com/TSAF/

1. 线性回归模型

1.1. 线性回归模型

当一个时间序列变量 y 和另一个（或几个）时间序列变量 x 之间存在线性关系时，我们可以用线性回归模型描述这种关系。

在统计学中，y 被称为**因变量（dependent variable）**，x 被称为**自变量（independent variable）**。在计量经济学中，y 通常被称为**被解释（explained）变量**，x 是**解释（explanatory）变量**。当我们的目的是时间序列预测时，y 是**被预测（forecast）变量**，x 是**预测（predictor）变量**。

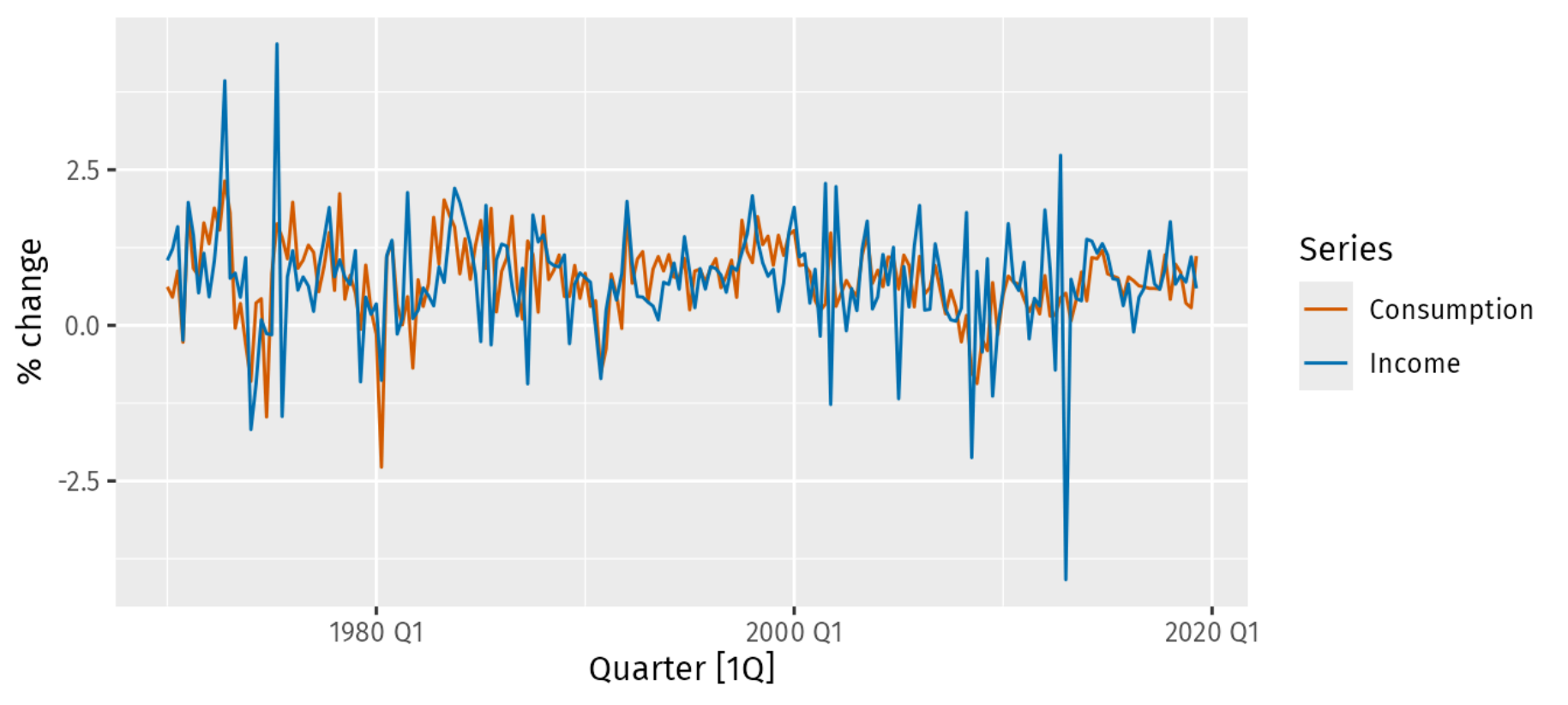
多变量线性回归模型的定义如下：

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k x_{k,t} + \varepsilon_t$$

这里 ε_t 是误差项，它包含了所有无法被 $\beta_0 + \beta_1 x_{1,t} + \ldots + \beta_k x_{k,t}$ 解释但又影响 y_t 取值的因素。系数 $\beta_1, \ldots, \beta_k$ 衡量各自对应的预测变量的边际效应，即固定其他预测变量时，该变量的变化对 y_t 变化的影响。

1.2. 单变量回归模型：收入与消费支出

下图展示了 fpp3 包提供的 `us_change` 数据集中美国人均消费支出和人均可支配收入的百分比变化（1970Q1 – 2019Q2）



1.2. 单变量回归模型：收入与消费支出

假设我们用收入（Income）作为预测变量，则回归模型可以写成

$$\text{Consumption}_t = \beta_0 + \beta_1 \text{Income}_t + \varepsilon_t$$

可以在 `model()` 函数中调用 `TSLM()` 函数进行拟合。拟合方法为最小二乘法。

```
us_change |>
  model(TSLM(Consumption ~ Income)) |> # 注意 TSLM() 中回归模型的写法
  report() # 显示拟合后的结果，和 base-R 中的 summary() 函数类似
```

拟合后的模型可以写成

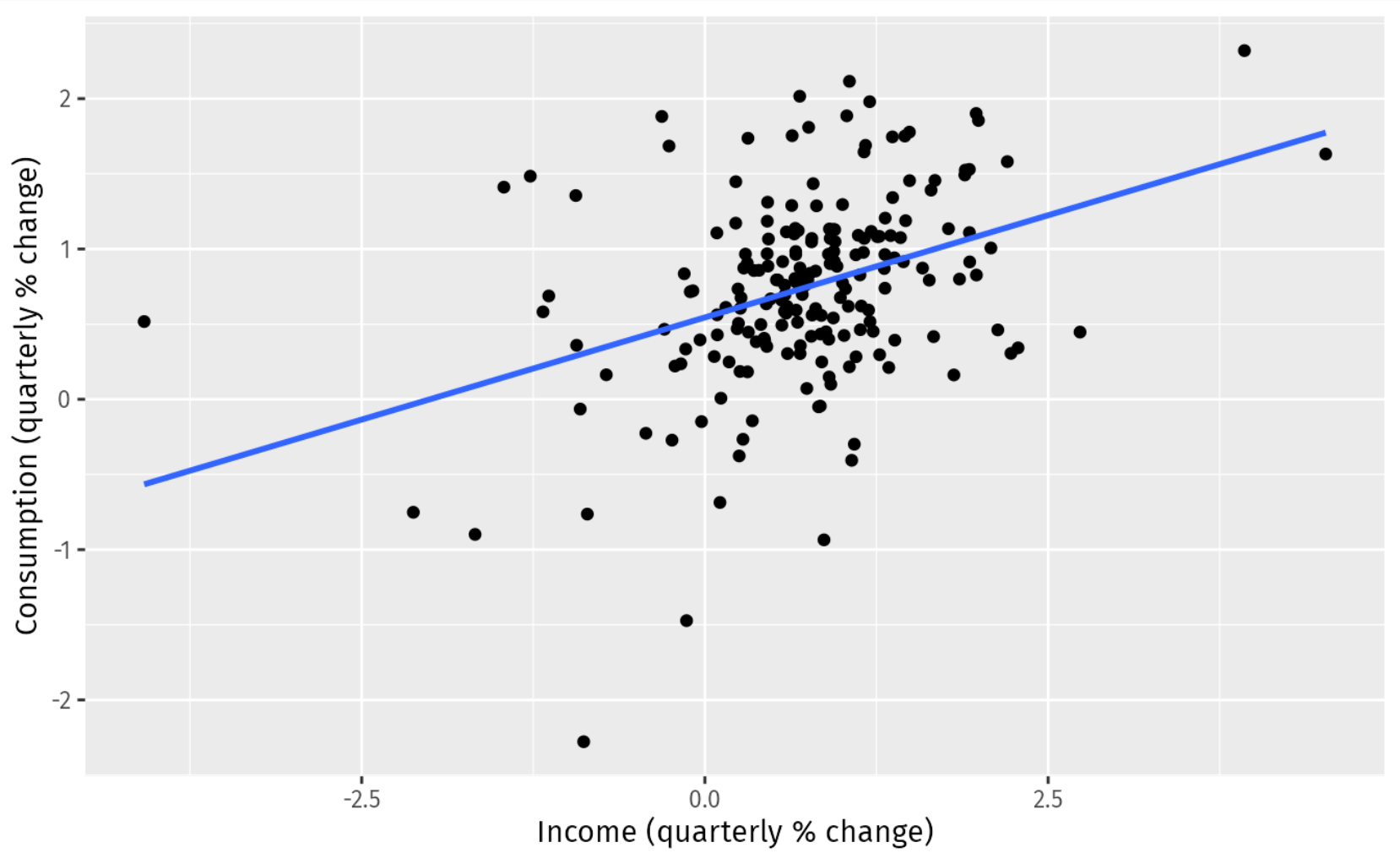
$$\widehat{\text{Consumption}}_t = \underset{(0.054)}{0.54} + \underset{(0.047)}{0.27}\ \text{Income}_t$$

注：系数下方括号内的数字是标准误。

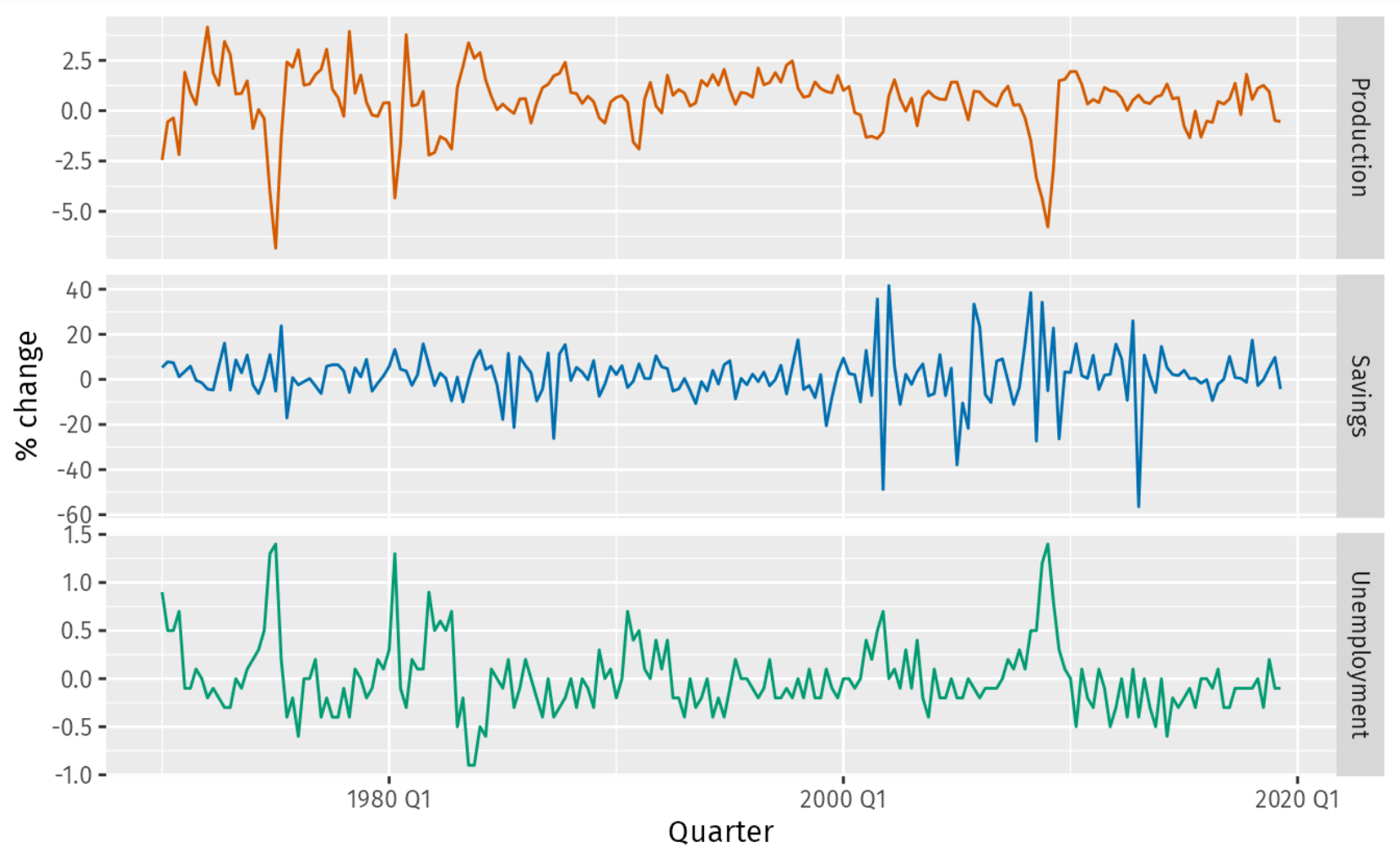
1.2. 单变量回归模型：收入与消费支出

```
Series: Consumption
Model: TSLM

Residuals:
     Min       1Q   Median       3Q      Max
-2.58236 -0.27777  0.01862  0.32330  1.42229

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54454    0.05403  10.079  < 2e-16 ***
Income       0.27183    0.04673   5.817  2.4e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5905 on 196 degrees of freedom
Multiple R-squared: 0.1472, Adjusted R-squared: 0.1429
F-statistic: 33.84 on 1 and 196 DF, p-value: 2.4022e-08
```

1.2. 单变量回归模型：收入与消费支出



1.3. 多变量回归模型：消费支出的其他预测变量



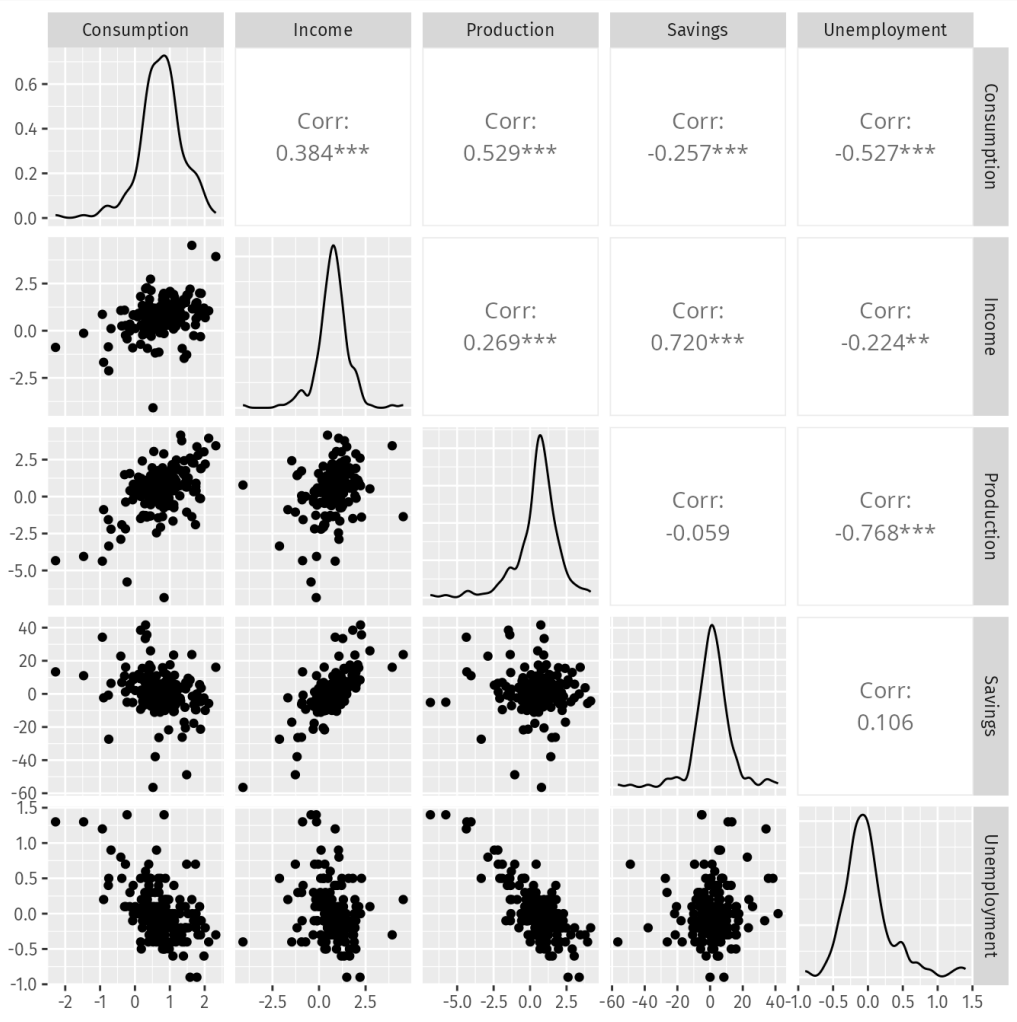
1.3. 多变量回归模型：消费支出的其他预测变量

我们可以考虑 Consumption_t 的多变量回归模型

$$\text{Cons}_t = \beta_0 + \beta_1 \text{Inc}_t + \beta_2 \text{Prod}_t + \beta_3 \text{Sav}_t + \beta_4 \text{Unemp}_t + \varepsilon_t$$

在拟合之前，可以先检查各变量见的相关系数

```
library(GGally) # 记得安装此程序包
us_change |>
  GGally::ggpairs(columns = 2:6)
```



1.3. 多变量回归模型：消费支出的其他预测变量

```
fit_consMR <- us_change |>
  model(MR = TSLM(
      Consumption ~ Income + Production + Savings + Unemployment
    ))
coef(fit_consMR)
```

```
# A tibble: 5 × 6
  .model term         estimate std.error statistic  p.value
  <chr>  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 MR     (Intercept)    0.253    0.0345       7.34 5.71e-12
2 MR     Income         0.741    0.0401      18.5  1.65e-44
3 MR     Production     0.0472   0.0231       2.04 4.29e- 2
4 MR     Savings       -0.0529   0.00292    -18.1  2.03e-43
5 MR     Unemployment  -0.175    0.0955      -1.83 6.89e- 2
```

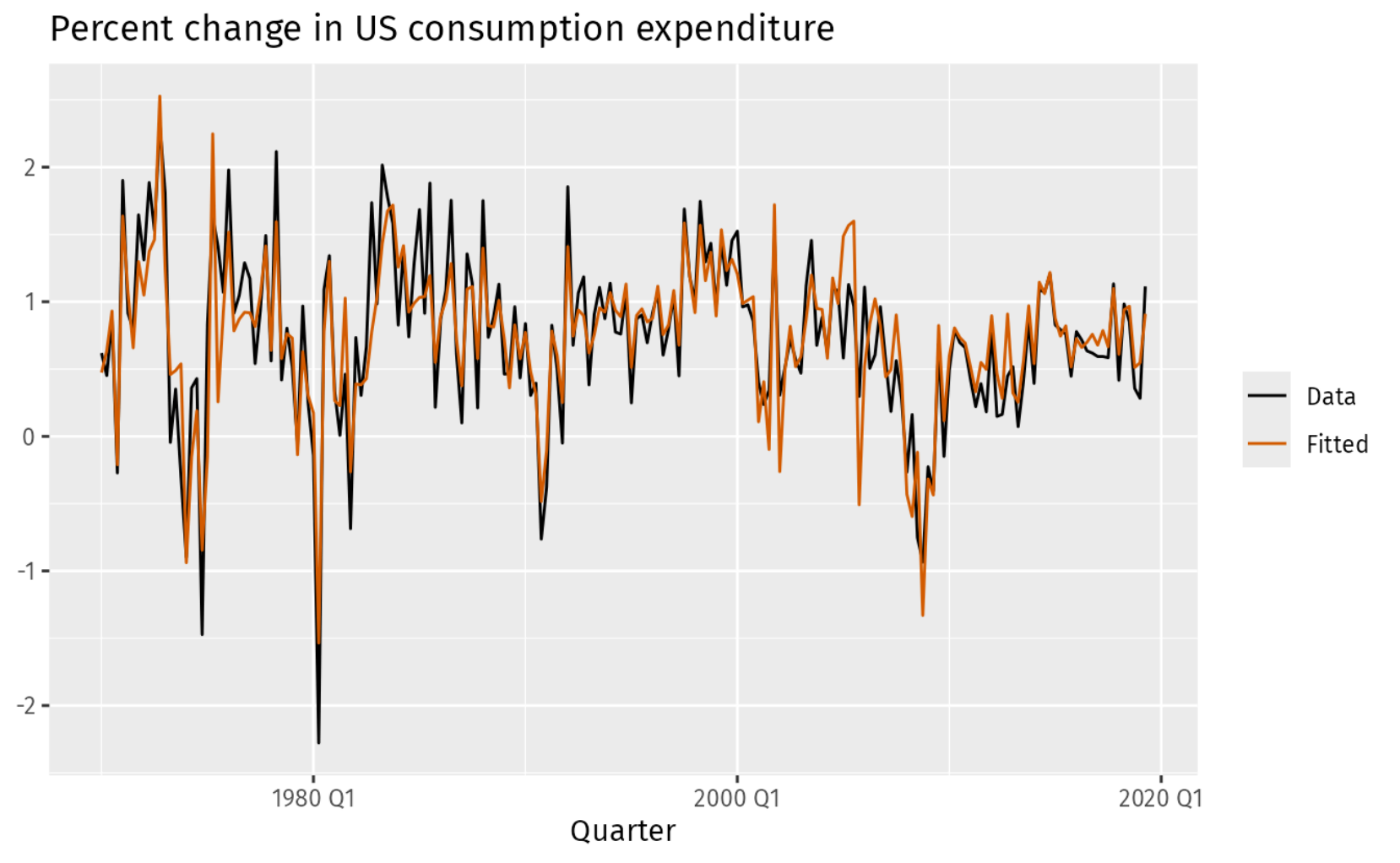
1.3. 多变量回归模型：消费支出的其他预测变量

拟合好参数后，我们就可以计算出被预测变量的拟合值 $\widehat{\text{Cons}}_t$。`augment()` 函数可以用拟合好的模型自动计算拟合值、残差和创新残差。

下面的代码将消费支出的观测值和拟合值绘制在同一个时序图中进行对比

```
augment(fit_consMR) |>
  ggplot(aes(x = Quarter)) +
  geom_line(aes(y = Consumption, colour = "Data")) +
  geom_line(aes(y = .fitted, colour = "Fitted")) +
  labs(y = NULL,
    title = "Percent change in US consumption expenditure"
  ) +
  scale_colour_manual(values=c(Data="black",Fitted="#D55E00")) +
  guides(colour = guide_legend(title = NULL))
```

1.3. 多变量回归模型：消费支出的其他预测变量



1.4. 残差诊断和伪回归

最小二乘法假设回归误差项序列 $(\varepsilon_1, \ldots, \varepsilon_T)$ 满足下列条件：

- **均值为零**。如不满足则代表预测值存在系统性偏误。
- **无自相关**。如不满足则代表数据中的信息没有被充分利用。
- **与预测变量不相关**。如不满足则意味着还有其他预测变量应该加入模型中。

此外，我们也经常假设误差项服从正态分布。这是为了方便进行系数的假设检验。

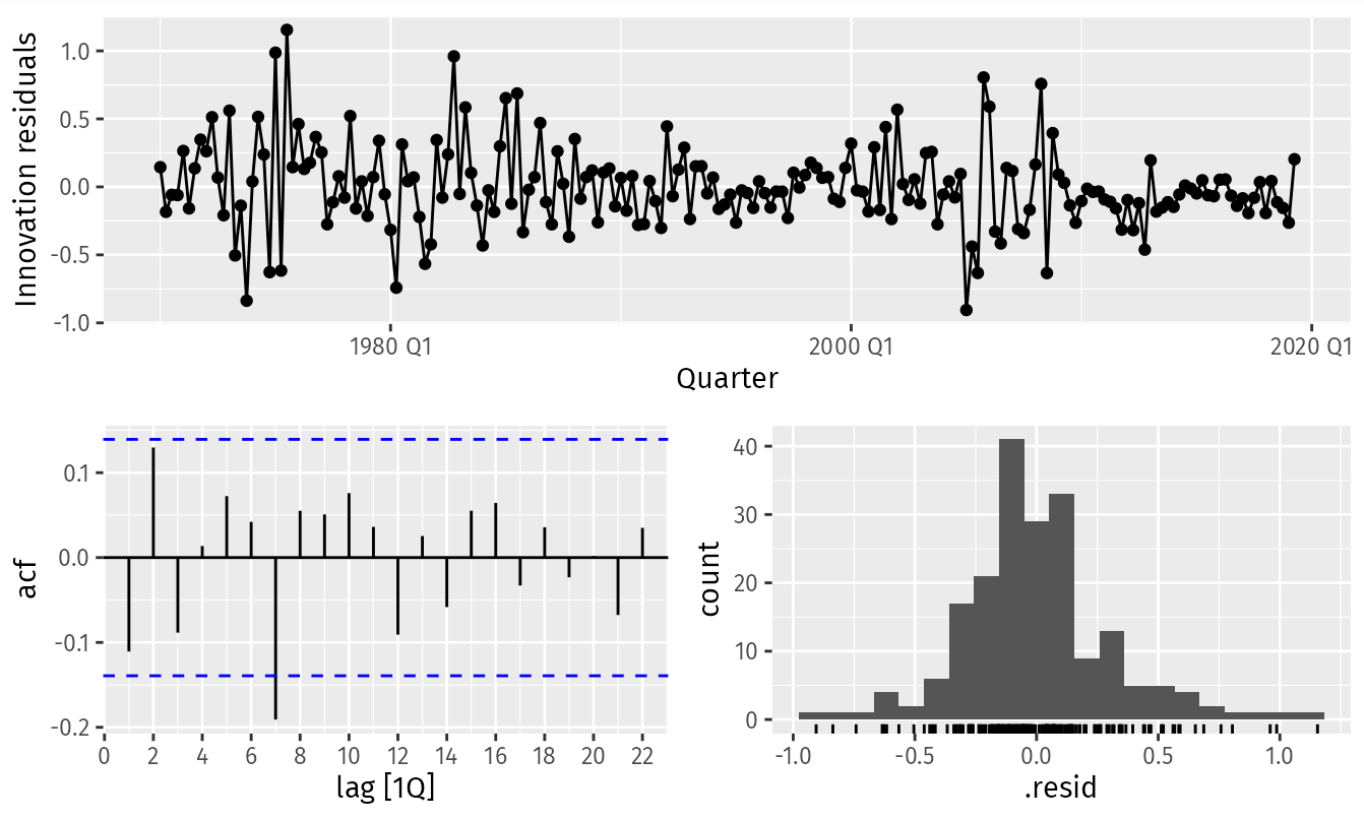
我们无法直接观察误差项，但可以利用拟合后的残差 $\hat{e}_t = y_t - \hat{y}_t$ 验证这些假设是否成立。从残差得定义可得

$$\sum_{t=1}^{T} \hat{e}_t = 0, \quad \sum_{t=1}^{T} x_{i,t}\hat{e}_t = 0, (i = 1, \ldots, k)$$

即残差的均值为零，残差与预测变量不相关。

1.4. 残差诊断和伪回归

```
fit_consMR |> gg_tsresiduals() # 绘制残差诊断图
```



1.4. 残差诊断和伪回归

除了利用 ACF 图判断是否存在自相关外，还可以利用 Ljung-Box 检验进行统计推断。
Ljung-Box 检验的假设是

H_0: 序列不存在自相关（是白噪声序列）。
H_1: 序列存在自相关（不是白噪声序列）。

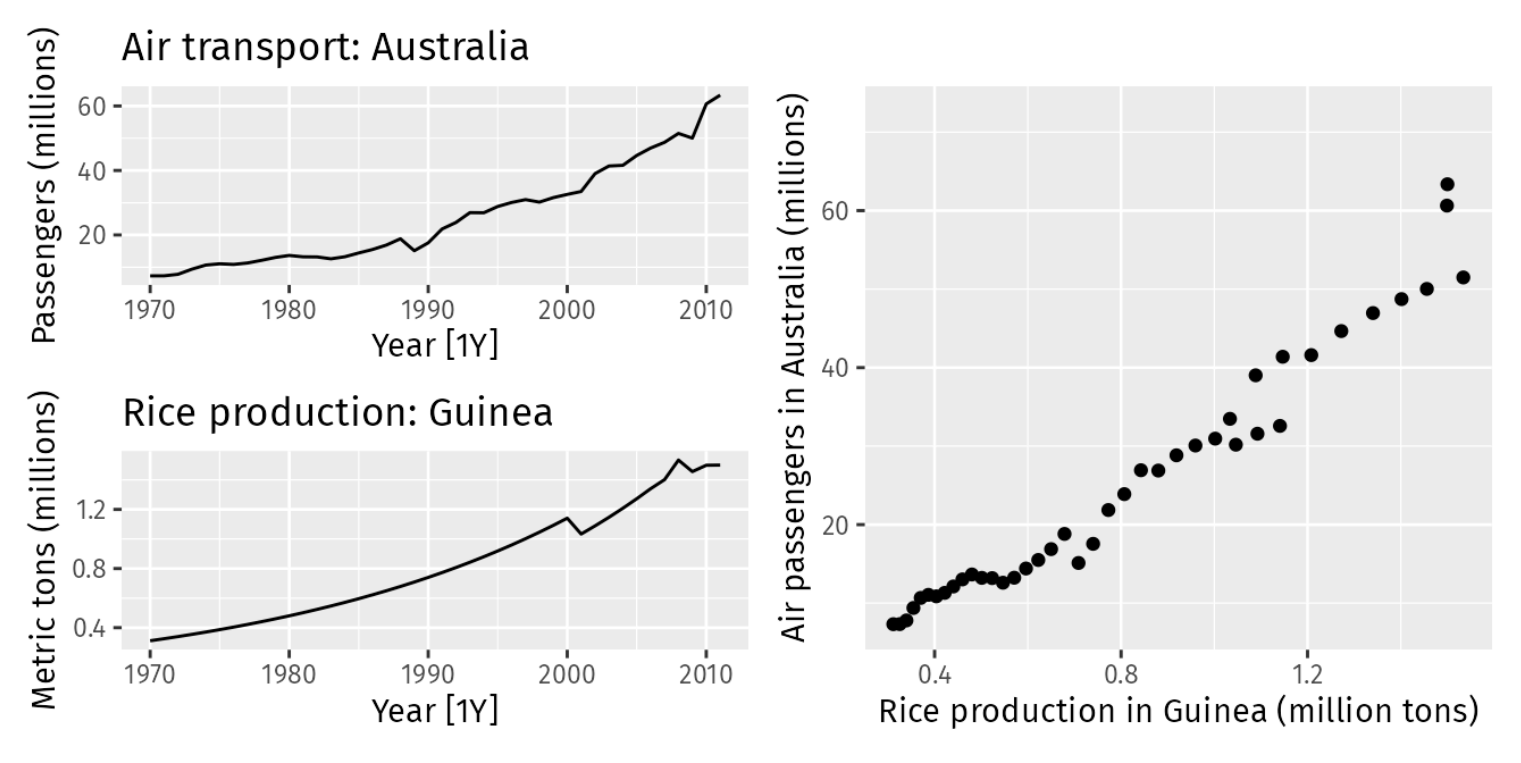
```
augment(fit_consMR) |>
  features(.innov, ljung_box, lag = 10)
```

```
# A tibble: 1 × 3
  .model lb_stat lb_pvalue
  <chr>    <dbl>     <dbl>
1 MR        18.9    0.0420
```

Ljung-Box 检验结果显示 p 值小于 0.05，即在 5% 水平上显著，提示拒绝零假设 H_0。
但 ACF 图提示自相关阶数为 7 且程度较小，可能不会造成太大影响。

1.4. 残差诊断和伪回归

多数情况下，时间序列数据呈现出明显的趋势特征。如果预测变量和被预测变量都有类似的趋势，则回归模型会检测出两者间明显的线性相关。但是这种相关并不是真正的相关。例如下图中的澳大利亚航空旅客数和几内亚大米产量。



1.4. 残差诊断和伪回归

存在共同趋势的时间序列变量间的回归被称为伪回归（spurious regression）。

伪回归并不能揭示变量间真正的相关关系，因此即使在短期内的预测效果很好，长期的预测效果往往较差。

如果拟合结果的 R^2 和自相关系数都很高，则说明可能存在伪回归。在用大米产量作为预测变量回归航空旅客数的拟合结果中，$R^2 = 0.958$，同时存在明显的 1 阶自相关，符合伪回归的特征。（详见教科书第 7.3 节）

2. 预测变量的选择

2.1. 预测变量的可预测性

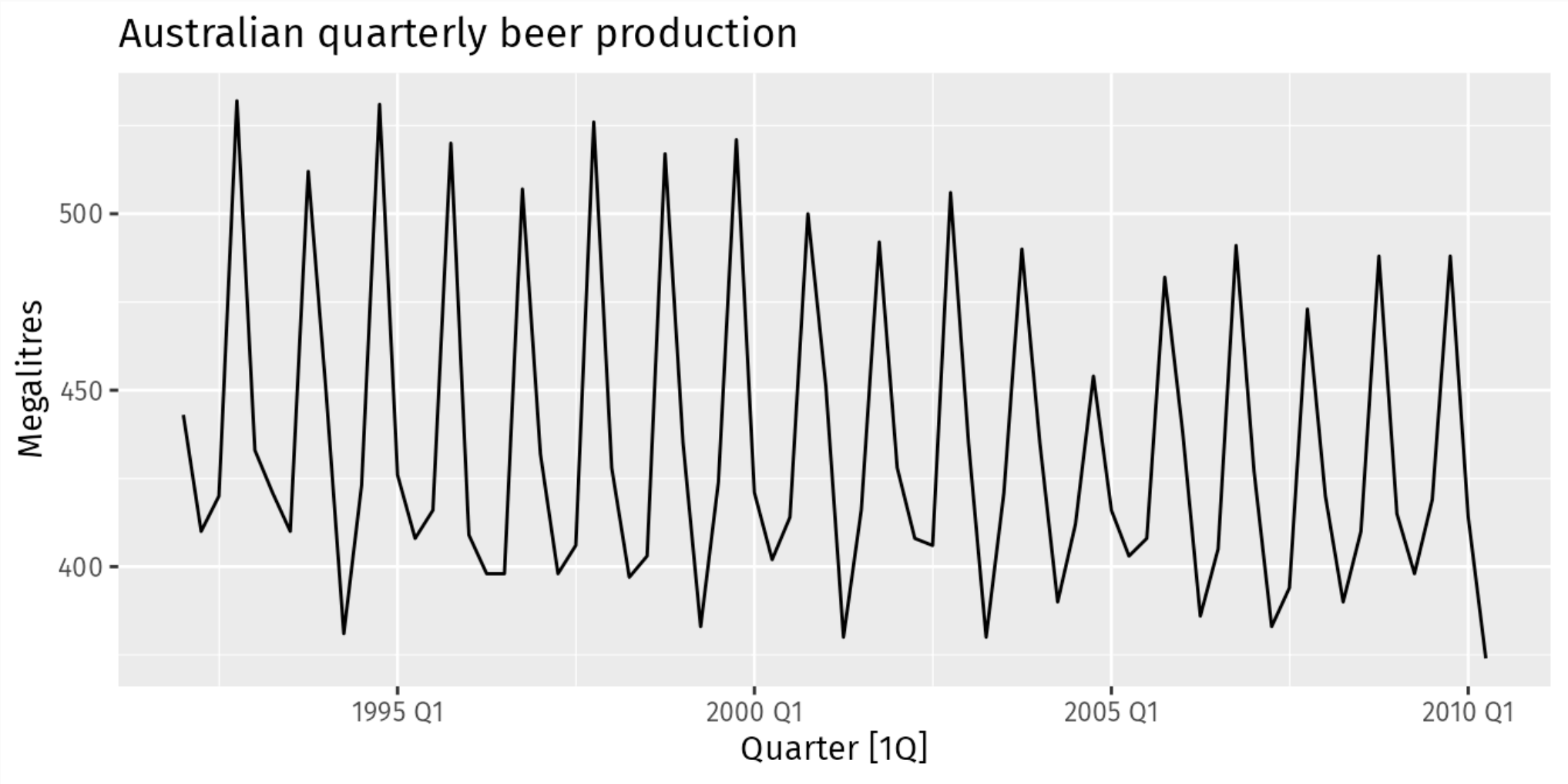
如果用回归模型进行预测，我们就必须事先知道预测变量在未来的取值。这包含两种情况：

1. 预测变量本身是很容易预测的，例如短期内的天气
2. 预测变量是时间的非随机函数
   - 用时间 t 作为预测变量：$y_t = \beta_0 + \beta_1 t + \varepsilon_t$，即趋势模型
   - 用季节或日期虚拟变量（dummy variable）作为预测变量：

$$y_t = \beta_1 + \beta_2 D_{2,t} + \beta_3 D_{3,t} + \beta_4 D_{4,t} + \varepsilon_t$$

（如果加入常数项，则虚拟变量的个数应为总类别数减一）
类似的还有法定假日虚拟变量等。

2.2. 澳大利亚季度啤酒产量的例子



2.2. 澳大利亚季度啤酒产量的例子

考虑包含趋势和季节虚拟变量的回归模型

$$y_t = \beta_0 + \beta_1 t + \beta_2 D_{2,t} + \beta_3 D_{3,t} + \beta_4 D_{4,t} + \varepsilon_t$$

拟合此模型的代码是

```
model(TSLM(Beer ~ trend() + season()))
```

这里省略拟合结果。可以在教科书第 7.4 节中找到完整代码和拟合值的时序图。

2.3. 分布滞后模型

有时预测变量作用在被预测变量上的效果是滞后的，例如广告的效果可能要在几周或几个月后才能在销售量中体现。此时，我们可以用原预测变量的滞后项（或应称为前置项，因为它的发生时间早于被预测变量）作为预测变量，这类模型称为分布滞后（distributed lag）模型。这样做的另一个好处是可以进行对未来的短期预测。

一阶滞后模型可以写为

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \varepsilon_t$$

也可以用时加入多个滞后变量

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{1,t-2} + \beta_3 x_{2,t-1} + \beta_4 x_{2,t-2} + \varepsilon_t$$

2.4. 模型选择

选择最好的回归模型也是选择最好的预测变量组合。

错误的选择方法：

1. 根据被预测变量和预测变量间的相关系数判断是否加入该预测变量。

   ➡ 因为两者间的相关系数无法体现该变量在多变量模型中的表现。

2. 加入多个预测变量，然后将系数不显著的变量从模型中剔除。

   ➡ 因为我们的目的是预测，系数不显著并不等同于预测效果差。

2.4. 模型选择

选择模型时可以参考的测度包括

- **调整后的 R^2（adjusted R^2，记作 $\overline{R}^2$）**：越高代表模型拟合度越好
- **CV（cross-validation，交叉检验测度）**：越低代表模型拟合度越好
- **AIC（Akaike information criterion 赤池信息准则）**：越低代表模型的拟合度越好
- **AICc（修正赤池信息准则）**：AIC 倾向于选择预测变量更多的模型，AICc 修正了这个缺陷
- **BIC（Schwarz 的贝叶斯信息准则）**：越低代表模型的拟合度越好

很多人喜欢用 $\overline{R}^2$，但是它倾向于选择变量更多的模型，而变量越多意味着预测难度越大。统计学家更偏好 BIC，因为它最能体现变量间的真实关系，但是在预测中真实模型不一定最有用。本书作者推荐使用 AICc。

以上测度值都保存在拟合后的模型中，可以用 `glance()` 函数调取。

2.4. 模型选择

Table 7.1: All 16 possible models for forecasting US consumption with 4 predictors.

Income	Production	Savings	Unemployment	AdjR2	CV	AIC	AICc	BIC
●	●	●	●	0.763	0.104	-456.6	-456.1	-436.9
●	●	●		0.761	0.105	-455.2	-454.9	-438.7
●		●	●	0.760	0.104	-454.4	-454.1	-437.9
●		●		0.735	0.114	-435.7	-435.5	-422.6
●	●		●	0.366	0.271	-262.3	-262.0	-245.8
	●	●	●	0.349	0.279	-257.1	-256.8	-240.7
●			●	0.345	0.276	-256.9	-256.6	-243.7
●	●			0.336	0.282	-254.2	-254.0	-241.0
	●	●		0.324	0.287	-250.7	-250.5	-237.5
		●	●	0.311	0.291	-246.9	-246.7	-233.7
	●		●	0.308	0.293	-246.1	-245.9	-232.9
	●			0.276	0.304	-238.1	-238.0	-228.2
			●	0.274	0.303	-237.4	-237.3	-227.5
●				0.143	0.356	-204.6	-204.5	-194.7
		●		0.061	0.388	-186.5	-186.4	-176.7
				0.000	0.409	-175.1	-175.0	-168.5

3. 利用回归模型进行预测

3.1. 事前预测与事后预测

根据预测时所用信息的时间节点，可以将预测分为事前（ex-ante）预测和事后（ex-post）预测。

- **事前预测**：仅利用预测时已知的信息。

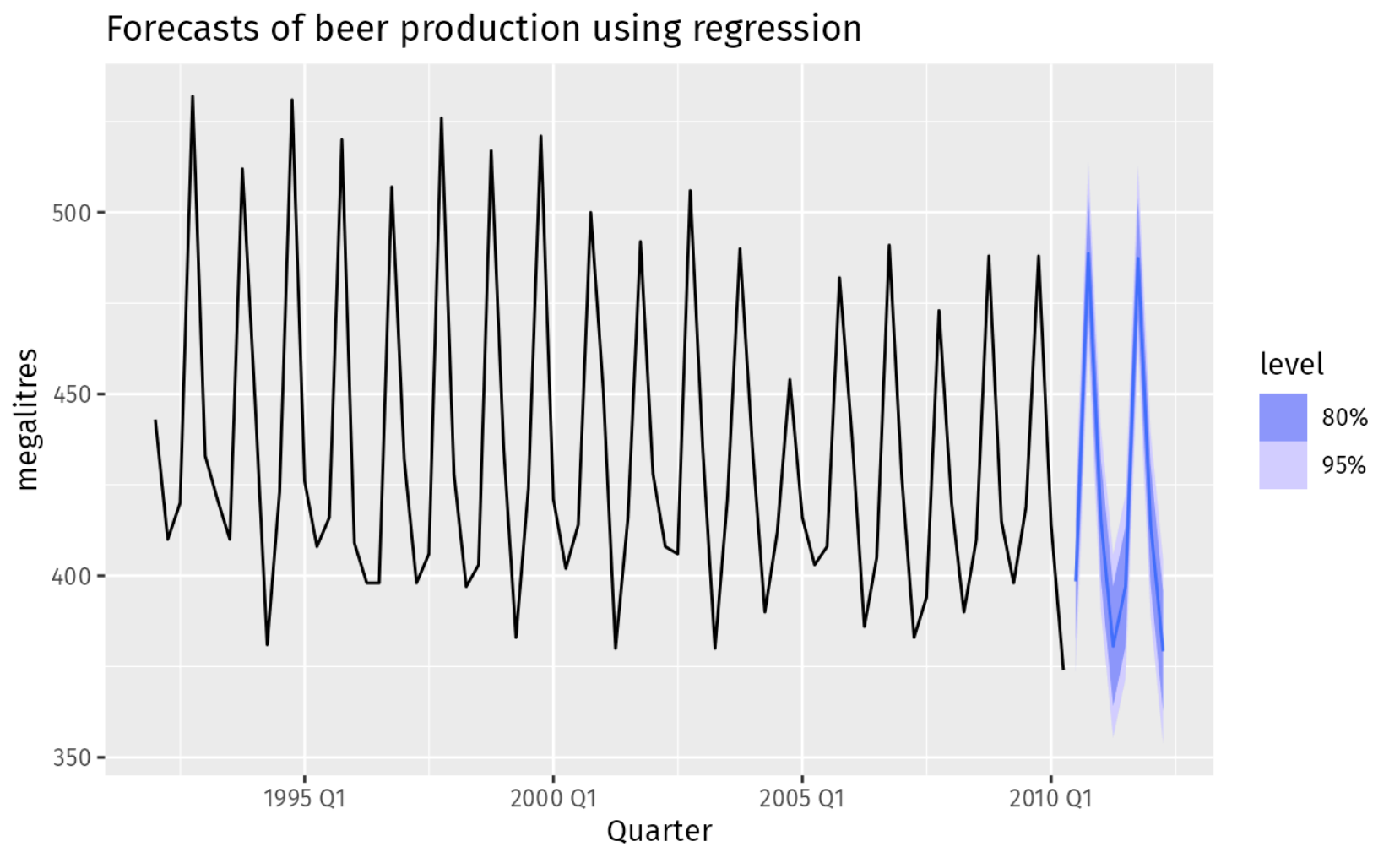
  这就是通常意义上的“预测”，包含以下几种情况：
  - 首先对 x_{T+h} 进行预测，然后利用该预测值再预测 y_{T+h}。
  - 预测变量为时间趋势、季节虚拟变量等确定函数时，可直接计算其未来值。
  - 利用政府部门或其他机构发布的 x_t 的预测数据。
  - 根据预设的情景（scenario）进行预测。
  - 利用分布滞后模型。
- **事后预测**：需要利用预测时未知的信息。

  事后预测主要用来理解模型的特点。

3.2. 澳大利亚季度啤酒产量的例子

```r
recent_production <- aus_production |>
  filter(year(Quarter) >= 1992) # 选取 1992 年以后的数据

fit_beer <- recent_production |>
  model(TSLM(Beer ~ trend() + season())) # 趋势 + 季节虚拟变量

fc_beer <- forecast(fit_beer) # 预测区间的默认值是两年
fc_beer |>
  autoplot(recent_production) +
  labs(
    title = "Forecasts of beer production using regression",
    y = "megalitres"
  )
```

3.2. 澳大利亚季度啤酒产量的例子



3.3. 基于情景分析预测美国的消费增长

在预测 x_t 的未来值比较困难时，预设不同的情景（scenario）是很常见的做法。

例如我们想预测美国未来四个季度的消费增长，就可以考虑以下两种情景：

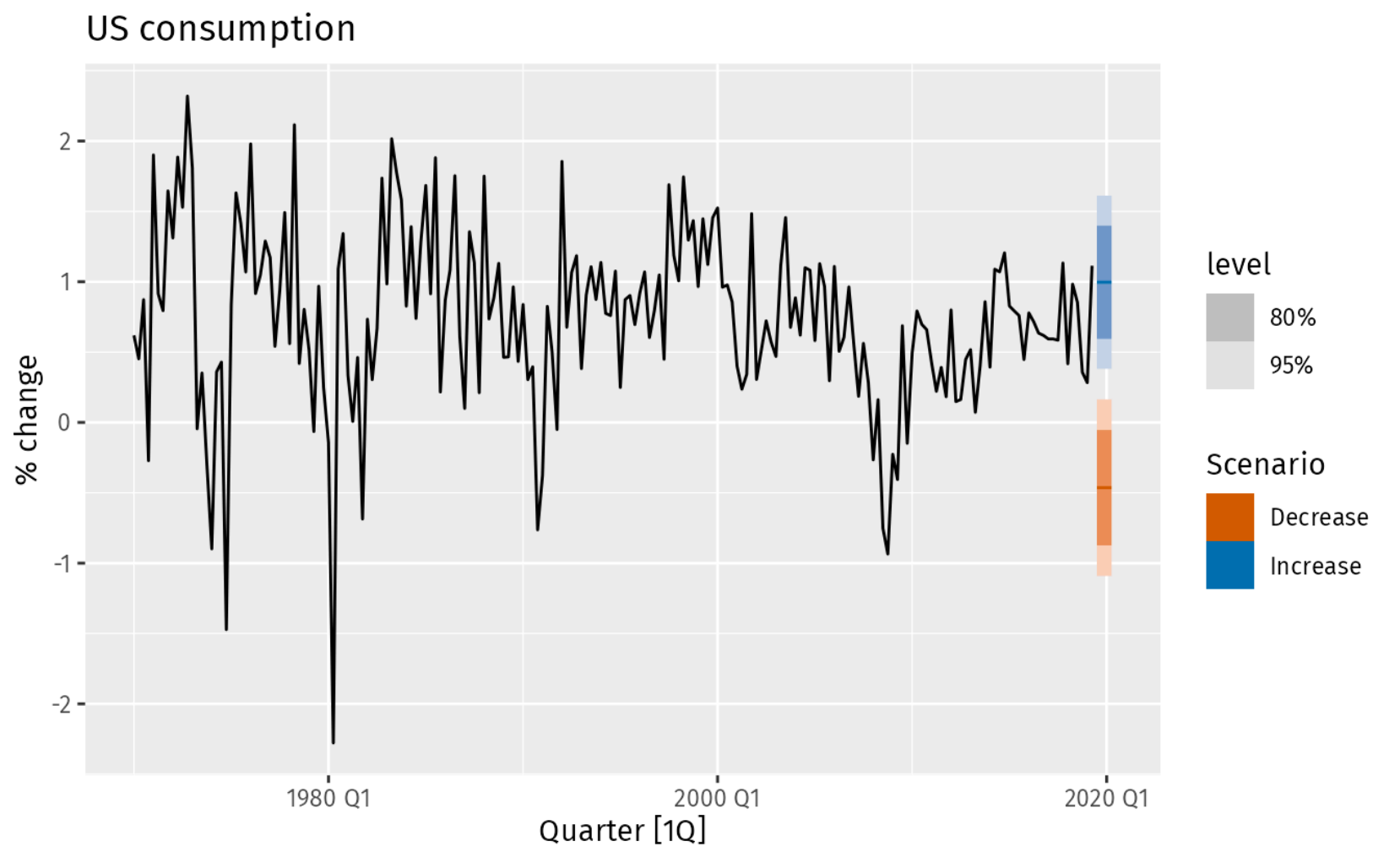
1. **增长**：收入持续增加 1%，储蓄持续增加 0.5%，失业率保持不变
2. **衰退**：收入持续减少 1%，储蓄持续减少 0.5%，失业率保持不变

可以利用 `scenario()` 函数设定情景。

3.3. 基于情景分析预测美国的消费增长

```
fit_consBest <- us_change |>
  model(lm = TSLM(Consumption ~ Income + Savings + Unemployment))
future_scenarios <- scenarios( # 生成两个情景（保存后为 list 形式）
  Increase = new_data(us_change, 4) |>
    mutate(Income=1, Savings=0.5, Unemployment=0),
  Decrease = new_data(us_change, 4) |>
    mutate(Income=-1, Savings=-0.5, Unemployment=0),
  names_to = "Scenario") # 储存情景的列的名称

fc <- forecast(fit_consBest, new_data = future_scenarios)
us_change |>
  autoplot(Consumption) + autolayer(fc) +
  labs(title = "US consumption", y = "% change")
```

3.3. 基于情景分析预测美国的消费增长



4. 课后练习

4. 课后练习

- 学习教科书第 7 章（Time Series Regression Model）中的内容，并尝试在自己的电脑上复现书中的结果。

- 利用 `us_change` 数据集回答下列问题：

  1. 用时间趋势和季节虚拟变量作为预测变量对消费增长率进行建模，拟合参数后进行预测。
  2. 先用时间趋势和季节虚拟变量作为预测变量对收入增长率进行建模，拟合参数后进行预测。然后将收入增长率作为预测变量对消费增长率进行建模，拟合参数后利用前一步获得的预测值对消费增长率进行预测。
  3. 对 (1) 和 (2) 的结果进行比较。哪个方法的预测效果更好?