Econometria | 2022/2023

Lezione 14: Time Series

Giuseppe Ragusa

https://gragusa.org

Roma, maggio 2023

Sommario

1. Serie temporali: quali peculiarità?
2. Uso di modelli di regressione per previsioni
3. Ritardi, differenze, autocorrelazione e stazionarietà
4. Autoregressioni
5. Il modello ADL (autoregressivo misto)
6. Incertezza e intervalli delle previsioni
7. Scelta della lunghezza dei ritardi: criteri di informazione
8. Non stazionarietà I: tendenze
9. Non stazionarietà II: rotture

Serie temporali: quali peculiarità?

Le serie temporali sono costituite da dati raccolti sulla stessa unità in più periodi temporali

Esempi:

1. Consumi aggregati e PIL per un paese (per esempio, 20 anni di osservazioni trimestrali = 80 osservazioni)
2. Tassi di cambio yen/$, GBP/$ ed Euro/$ (dati giornalieri per 1 anno = 365 osservazioni)
3. Consumo di sigarette pro capite in California, per anno (dati annuali)

Alcune serie temporali per dati macro e finanziari

Una serie temporale di dati finanziari giornalieri USA

Impieghi delle serie temporali

- Previsione
- Stima di effetti causali dinamici
  - Se la FED aumenta il Federal Funds rate, quale sarà l’effetto sui tassi di inflazione e disoccupazione fra 3 mesi? E fra 12 mesi?
  - Qual è l’effetto nel tempo sul consumo di sigarette di un aumento dell’imposta sulle sigarette?
  - Modellazione di rischi, usata nei mercati finanziari
- Tra le applicazioni al di là dell’economia vi sono la modellazione ambientale e climatica, ingegneristica (dinamiche di sistema), informatica (dinamica di rete)

Le serie temporali sollevano nuove problematiche tecniche

Ritardi temporali

Correlazione nel tempo (correlazione seriale, o autocorrelazione già incontrata con i dati panel)

Calcolo di errori standard quando gli errori sono serialmente correlati

Uso di modelli di regressione per la previsione

- Previsione e stima di effetti causali sono obiettivi piuttosto diversi
- Per la previsione
  - La distorsione da variabili omesse non è un problema!
  - Non ci preoccuperemo di interpretare i coefficienti nei modelli di previsione – non serve stimare effetti causali se si vogliono soltanto fare previsioni!
  - La validità esterna è fondamentale: il modello stimato usando dati storici deve valere nel (prossimo) futuro

Introduzione alle serie temporali e alla correlazione seriale

Basi per le serie temporali:

1. Notazione
2. Ritardi, differenze prime, tassi di crescita
3. Autocorrelazione (correlazione seriale)
4. Stazionarietà

Notazione

- Y_t indica il valore di Y nel periodo t
- $\{Y_1 \ldots, Y_T\}$ indica le T osservazioni sulla variabile serie temporale Y

Consideriamo soltanto osservazioni consecutive, a intervalli uniformi (per esempio mensili, dal 1960 al 1999, senza saltare mesi; dati mancanti e intervalli non uniformi introducono complicazioni tecniche)

Ritardi, differenze prime e tassi di crescita

CONCETTO CHIAVE 14.1 — **Ritardi, differenze prime, logaritmi e tassi di crescita**

- Il primo ritardo di una serie temporale Y_t è Y_{t-1}; il suo j-esimo ritardo è Y_{t-j}.
- La differenza prima di una serie, ΔY_t, è la sua variazione tra il periodo $t-1$ e il periodo t, cioè $\Delta Y_t = Y_t - Y_{t-1}$.
- La differenza prima del logaritmo di Y_t è $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- La variazione percentuale di una serie temporale Y_t tra i periodi $t-1$ e t è approssimativamente uguale a $100\Delta \ln(Y_t)$, dove l'approssimazione è più accurata quando la variazione percentuale è piccola.

Ritardi, differenze prime e tassi di crescita, ctd.

PIL = PIL Italiano (milioni di Euro)

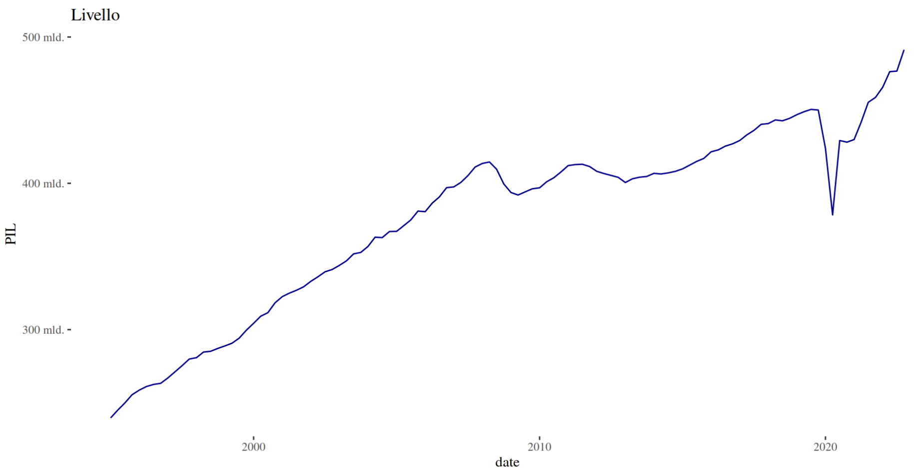
```r
cap <- "PIL Italiano. Source: Eurostat."
pil_ita <- fredr(
  series_id = "CPMNACSCAB1GQIT",
  observation_start =
    as.Date("1995-01-01"))
pil_ita <- pil_ita |>
  select(date, value) |>
  mutate(
    `PILl1` = lag(value),
    `logPIL` = log(value),
    `var. % QoQ` =
           100*(value-PILl1)/PILl1,
          `var. % QoQ (log diff)` =
          100*(logPIL-lag(logPIL)))|>
  dplyr::rename(`PIL` = value)
kable(pil_ita |>
        filter(date<as.Date("2000-01-01")),
      digits = c(0,0,0,2,2,2),
      caption = cap) |>
  kable_styling(font_size = 13)
```

PIL Italiano. Source: Eurostat.

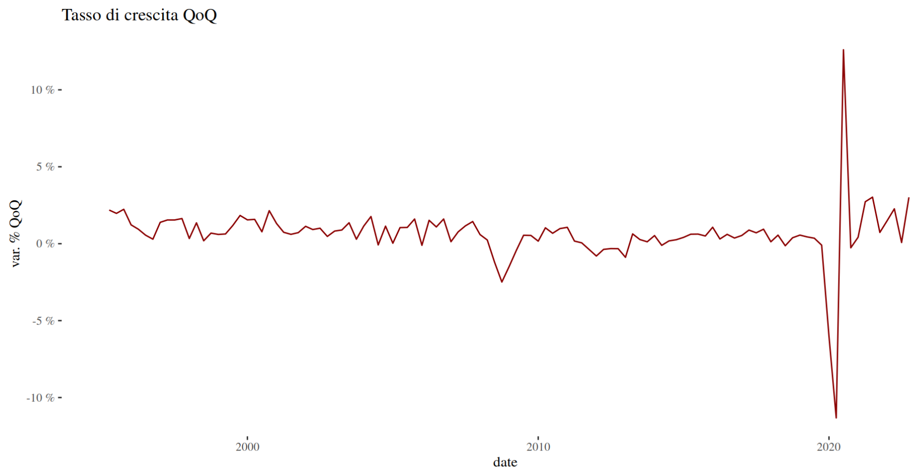
date	PIL	PILl1	logPIL	var. % QoQ	var. % QoQ (log diff)
1995-01-01	239678	NA	12.4	NA	NA
1995-04-01	244974	239678	12.4	2.21	2.19
1995-07-01	249849	244974	12.4	1.99	1.97
1995-10-01	255500	249849	12.4	2.26	2.24
1996-01-01	258647	255500	12.5	1.23	1.22
1996-04-01	261094	258647	12.5	0.95	0.94
1996-07-01	262545	261094	12.5	0.56	0.55
1996-10-01	263316	262545	12.5	0.29	0.29
1997-01-01	267000	263316	12.5	1.40	1.39
1997-04-01	271149	267000	12.5	1.55	1.54
1997-07-01	275362	271149	12.5	1.55	1.54
1997-10-01	279900	275362	12.5	1.65	1.63
1998-01-01	280846	279900	12.6	0.34	0.34
1998-04-01	284680	280846	12.6	1.37	1.36
1998-07-01	285219	284680	12.6	0.19	0.19
1998-10-01	287188	285219	12.6	0.69	0.69
1999-01-01	288921	287188	12.6	0.60	0.60
1999-04-01	290755	288921	12.6	0.63	0.63
1999-07-01	294207	290755	12.6	1.19	1.18
1999-10-01	299655	294207	12.6	1.85	1.83

Esempio: PIL e sua variazione

```r
ggplot(pil_ita, aes(y=PIL, x = date)) + geom_line(col = "darkblue") +
  ggtitle("Livello") + theme_tufte() +
  scale_y_continuous(labels = scales::unit_format(unit = "mld.", scale = 1e-03))
ggplot(pil_ita, aes(y=`var. % QoQ`, x = date)) +
  geom_line(aes(y=`var. % QoQ (log diff)`), col = "darkred") +
  ggtitle("Tasso di crescita QoQ") + theme_tufte() +
  scale_y_continuous(labels = scales::unit_format(unit = "%", scale = 1))
```



PIL Italiano (trimestrale, in milioni di Euro)



PIL Italiano tasso di crescita % (QoQ)

Autocorrelazione (correlazione seriale)

La correlazione di una serie con i suoi valori ritardati è detta autocorrelazione o correlazione seriale

La j-esima autocovarianza di Y_t è

$$cov(Y_t, Y_{t-1}) = E[(Y_t - E(Y_t))(Y_{t-j} - E(Y_{t-j}))]$$

La j-esima autocorrelazione di Y_t è

$$corr(Y_t, Y_{t-j}) = \frac{cov(Y_t, Y_{t-j})}{\sqrt{var(Y_t)var(Y_{t-j})}}$$

Queste sono correlazioni di popolazione, che descrivono la distribuzione congiunta di (Y_t, Y_{t-j})

Autocorrelazioni campionarie

- La j-esima autocorrelazione campionaria è una stima della j-esima autocorrelazione di popolazione:

$$\widehat{cov}(Y_t, Y_{t-j}) = \frac{1}{T} \sum_{t=j+1}^{T} \left(Y_t - \bar{Y}_t\right) \left(Y_{t-j} - \bar{Y}_{t-j}\right)$$

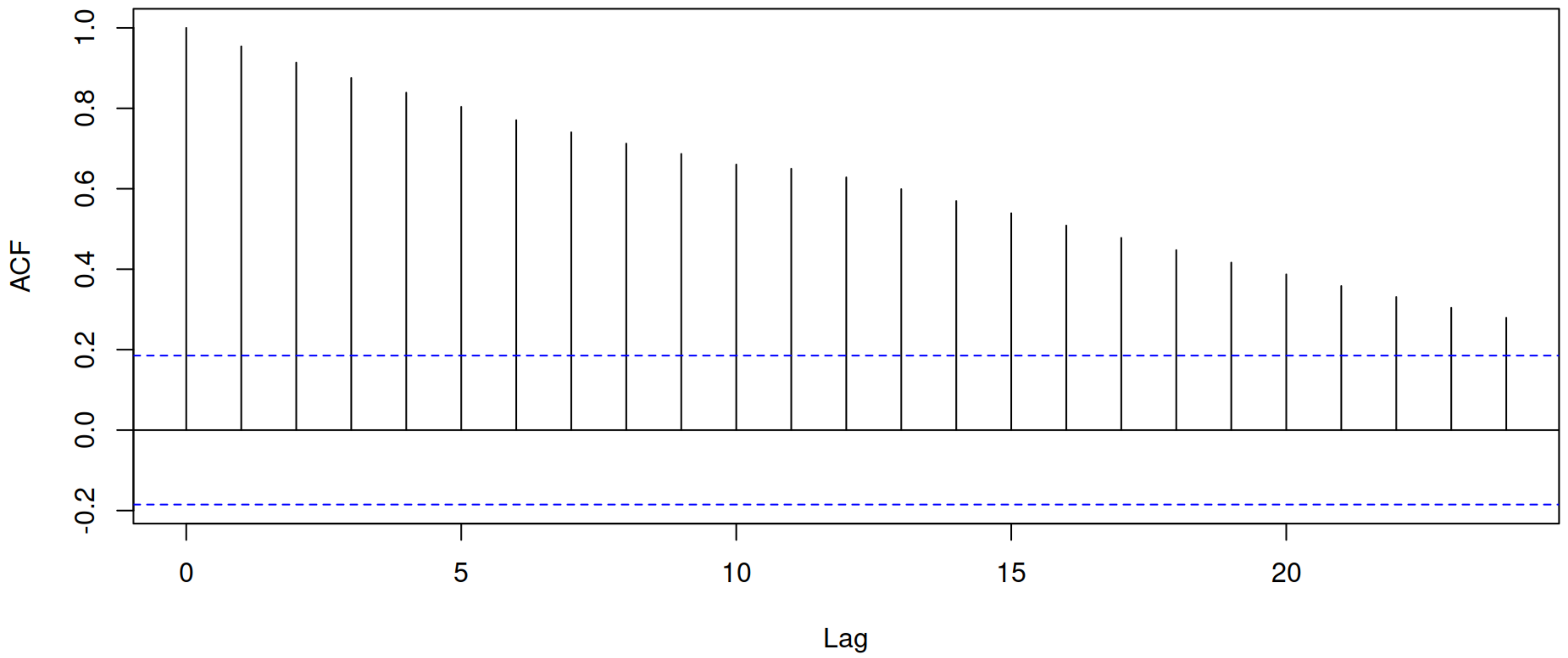
dove $\bar{Y}$ è la media campionaria di Y_t calcolata su osservazioni $t = j + 1, \ldots, T$.

Note:

- La sommatoria è su $t = j + 1, \ldots, T$ (perché?)
- Il divisore è T non $T - j$ (questa è la definizione usata per le serie temporali)

Le autocorrelazioni della serie del PIL italiano:

```r
acf(pil_ita$PIL, lag.max = 24, plot = TRUE, main = "")
```



PIL Italiano: Autocorrelazioni

Stazionarietà

CONCETTO CHIAVE 14.5 **Stazionarietà**

Una serie temporale Y_t è **stazionaria** se la sua distribuzione di probabilità non cambia nel corso del tempo, cioè se la distribuzione congiunta di $(Y_{s+1}, Y_{s+2},...,Y_{s+T})$ non dipende da s indipendentemente dal valore di T; altrimenti, la serie Y_t viene detta **non stazionaria**. Due serie temporali X_t e Y_t, sono dette **congiuntamente stazionarie** se la distribuzione congiunta di $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2},...,X_{s+T}, Y_{s+T})$ non dipende da s indipendentemente dal valore di T. La stazionarietà impone che il futuro sia come il passato, almeno in senso probabilistico.

La stazionarietà indica che la storia è rilevante. Si tratta di un requisito chiave per la validità esterna della regressione di serie temporali.

Per ora assumiamo che Y_t sia stazionaria (ci torneremo più avanti).

Autoregressioni

- Un punto di partenza naturale per un modello di previsione è quello di usare valori passati di Y_t (cioè $Y_{t-1}, Y_{t-2}, \ldots$) per la previsione di Y_t.
- Un' autoregressione è un modello di regressione in cui si esegue la regressione di Y_t rispetto ai suoi valori passati.
- Il numero di ritardi usati come regressori è detto ordine dell'autoregressione.
  - In una autoregressione del primo ordine, si esegue la regressione di Y_t rispetto a Y_{t-1}
  - In una autoregressione del p-esimo ordine, si esegue la regressione di Y_t rispetto a $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$.

Il modello autoregressivo del primo ordine

Il modello di popolazione AR(1) è

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

- β_0 e β_1 non hanno interpretazioni causali

- se $\beta_1 = 0, Y_{t-1}$ non è utile per prevedere Y_t
- i parametri del modello AR(1) possono essere stimati da una regressione di Y_t rispetto a Y_{t-1}
- la verifica di $\beta_1 = 0$ v. $\beta_1 \neq 0$ fornisce un test dell'ipotesi che Y_{t-1} non sia utile per prevedere Y_t

Modello AR(1) per il tasso di crescita del PIL

```r
pil_ita <- pil_ita |>
              mutate(DY   = 100*(log(PIL)-log(lag(PIL))),
                     DYl1 = lag(DY),   ## Primo ritardo
                     DYl2 = lag(DY,2)) ## Secondo ritardo
## Escludiamo i trimestri immediatamente precedente al trimestre
## Covid (I trimestre 2020) e tutti i trimestri post-Covid
pre_covid_pil_ita <- pil_ita |> filter(date < as.Date("2019-10-01"))
range(pre_covid_pil_ita$date)
```

```
[1] "1995-01-01" "2019-07-01"
```

Usiamo $DY_t = \log(PIL_t) - \log(PIL_{t-1})$ per due motivi:

- Spesso siamo interessati a predire il tasso di variazione piuttosto che i valori in livelli e quindi usare DY sembra appropriato
- Utilizzare PIL presenta dei problemi di tipo statistico in quanto questa variabile ha un'autocorrelazione particolarmente persistente e cioè genera problemi statistici nella stima dei parametri.
- Qualora fossimo comunque interessati a predire il livello del PIL in ogni trimestre potremmo sempre utilizzare la relazione

$$\log PIL_{T+1} = DY_{T+1} + \log PIL_{T-1}$$

AR(1)

```r
## Stima AR(1)
summary(lm(DY~DYl1, data=pre_covid_pil_ita))
```

```

Call:
lm(formula = DY ~ DYl1, data = pre_covid_pil_ita)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0687 -0.3524  0.0136  0.4391  1.4458 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.2598     0.0824    3.15   0.0022 ** 
DYl1          0.5693     0.0818    6.96  4.4e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.621 on 95 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.338,	Adjusted R-squared:  0.331 
F-statistic: 48.4 on 1 and 95 DF,  p-value: 4.36e-10
```

La variazione ritardata nel tasso di crescita del PIL è un predittore utile del tasso di crescita del PIL attuale (t-test)

Previsioni: terminologia e notazione

- I valori predetti sono “dentro il campione” (definizione consueta)
- Le previsioni sono “fuori dal campione” – nel futuro
- Notazione :
  - $Y_{T+1|T}$ = previsione di Y_{T+1} basata su $Y_T, Y_{T-1}, Y_{T-1}, \ldots$ usando i coefficienti di popolazione (ignoti)
  - $\hat{Y}_{T+1|T}$ = previsione di Y_{T+1} basata su $Y_T, YT-1, \ldots$ usando i coefficienti stimati su dati al periodo T
  - Per un AR(1):
    - $Y_{T+1|T} = \beta_0 + \beta_1 Y_T$
    - $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$ dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono stimati con dati al periodo T

Errori di previsione

L’errore di previsione futura a un periodo è

$$\text{errore previsione } = \hat{Y}_{T+1|T} - Y_{T+1}$$

La distinzione tra errore di previsione e residuo è la stessa che esiste tra previsione e valore predetto:

- un residuo è “dentro il campione”

$$\text{residuo}_t = Y_t - \hat{\beta}_0 + \hat{\beta}_1 Y_{t-1}, \quad t = 1, \ldots, T$$

- un errore di previsione è “fuori campione” – il valore di Y_{T+1} non è usato nella stima dei coefficienti di regressione

$$\text{errore previsione}_t = Y_{T+1} - \hat{\beta}_0 + \hat{\beta}_1 Y_T.$$

Esempio: previsione crescita del PIL con AR(1)

```r
AR1 <- lm(DY~DYl1, data=pre_covid_pil_ita)
## Residui
AR1
```

```

Call:
lm(formula = DY ~ DYl1, data = pre_covid_pil_ita)

Coefficients:
(Intercept)         DYl1
      0.260        0.569
```

```r
## Predizione
YT = tail(pre_covid_pil_ita$DY, 1) ## Ultima osservazione PIL
Yhat <- predict(AR1, newdata = list(DYl1 = YT)) ## Predizione con ultima osservazine e' Y_{T+1|T}
Yhat
```

```
    1
0.463
```

```r
## Valore realizzato
Y0 <- pil_ita |> filter(date==as.Date("2019-10-1")) |> pull(DY)
Y0
```

```
[1] -0.0951
```

```r
## Errore prevision
Y0 - Yhat
```

```
     1
-0.558
```

Il modello AR(p)

Il modello autoregressivo del p-esimo ordine AR(p) è

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t$$

Il modello AR(p) usa p ritardi di Y come regressori

Il modello AR(1) è un caso particolare

I coefficienti non hanno un'interpretazione causale

Per verificare l'ipotesi che $Y_{t-2}, \ldots, Y_{t-p}$ non siano utili a prevedere Y_t, oltre a Y_{t-1}, si usa un test Wald

Per determinare p

1. "criterio di informazione" (ne parleremo più avanti)

Modello AR(2) per il tasso di crescita del PIL

```r
AR2 <- lm(DY~DYl1+DYl2, data=pre_covid_pil_ita)
sAR2 <- summary(AR2)
sAR2
```

```
Call:
lm(formula = DY ~ DYl1 + DYl2, data = pre_covid_pil_ita)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2468 -0.3637  0.0676  0.4188  1.2535 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.1985     0.0851    2.33    0.022 *  
DYl1          0.4168     0.1004    4.15  7.3e-05 ***
DYl2          0.2352     0.0983    2.39    0.019 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.608 on 93 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.359, Adjusted R-squared:  0.345 
F-statistic:   26 on 2 and 93 DF,  p-value: 1.05e-09
```

- La statistica t del secondo ritardo è 2.392 (p-value: 0.019)
- Quindi, il secondo ritardo è utile a prevedere la crescita del PIL

Previsione

```r
AR2
```

```

Call:
lm(formula = DY ~ DYl1 + DYl2, data = pre_covid_pil_ita)

Coefficients:
(Intercept)         DYl1         DYl2  
      0.199        0.417        0.235  
```

$$Y_{T+1|T} = 0.199 + 0.417 \times Y_T + 0.235 \times Y_{T-1}$$

```r
YT = tail(pre_covid_pil_ita$DY,2) ## Ultima osservazione PIL YT = (Y_{T-1}, Y_T)
Yhat <- predict(AR2, newdata = list(DYl2 = YT[1], DYl1 = YT[2])) ## Ultime 2 osservazioni
Yhat
```

```
    1 
0.451 
```

Regressioni temporali con predittori aggiuntivi e modello autoregressivo misto

Finora abbiamo considerato modelli di previsione che usano solo valori passati di Y

Ha senso aggiungere altre variabili (X) che potrebbero essere predittori utili di Y, oltre ai valori predittivi dei valori ritardati di Y:

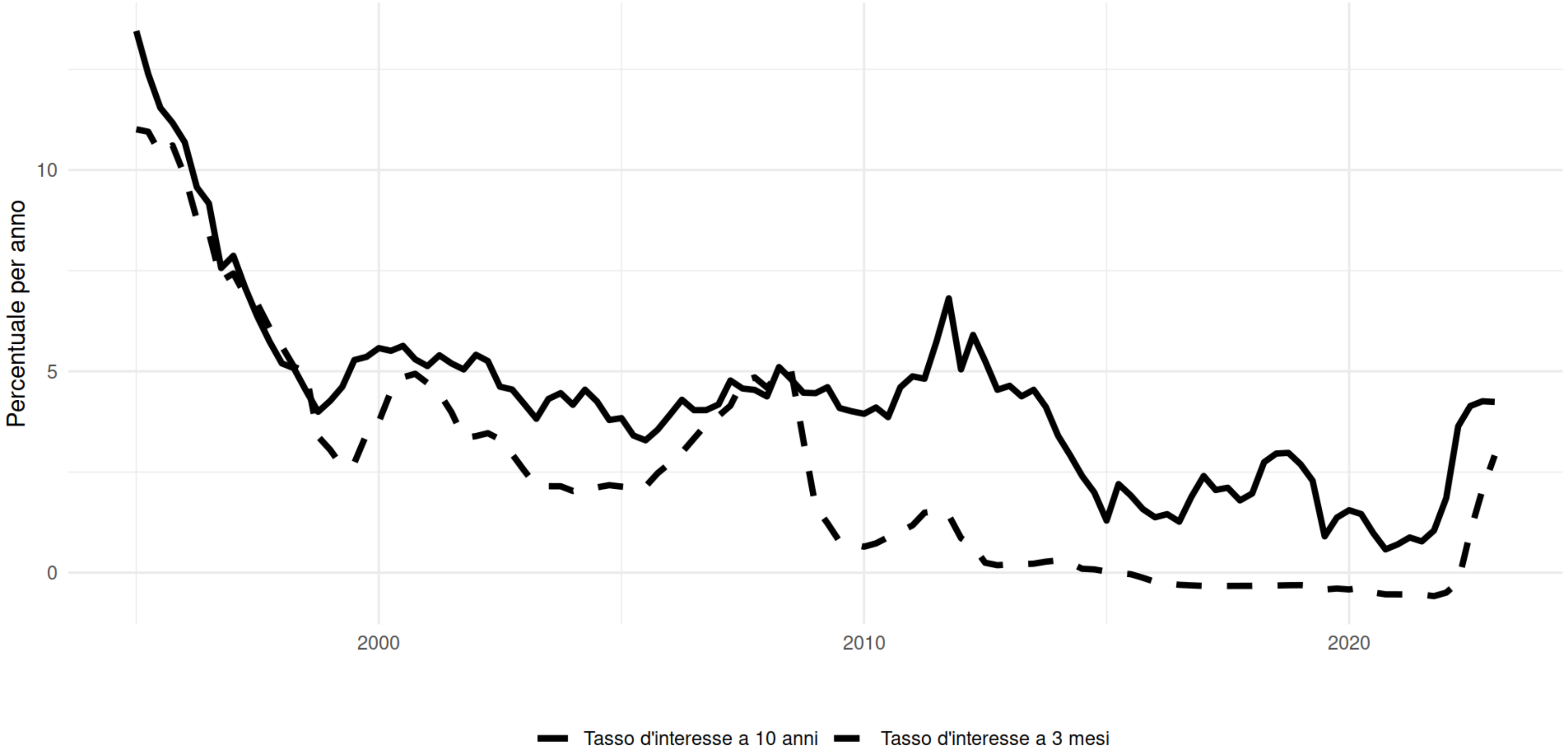
$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \cdots + \delta_r X_{t-r} + u_t$$

Questo è un modello autoregressivo misto con p ritardi di Y e r ritardi di X: ADL(p,r)

Esempio: tassi di interesse a lungo e a breve termine e relativo differenziale

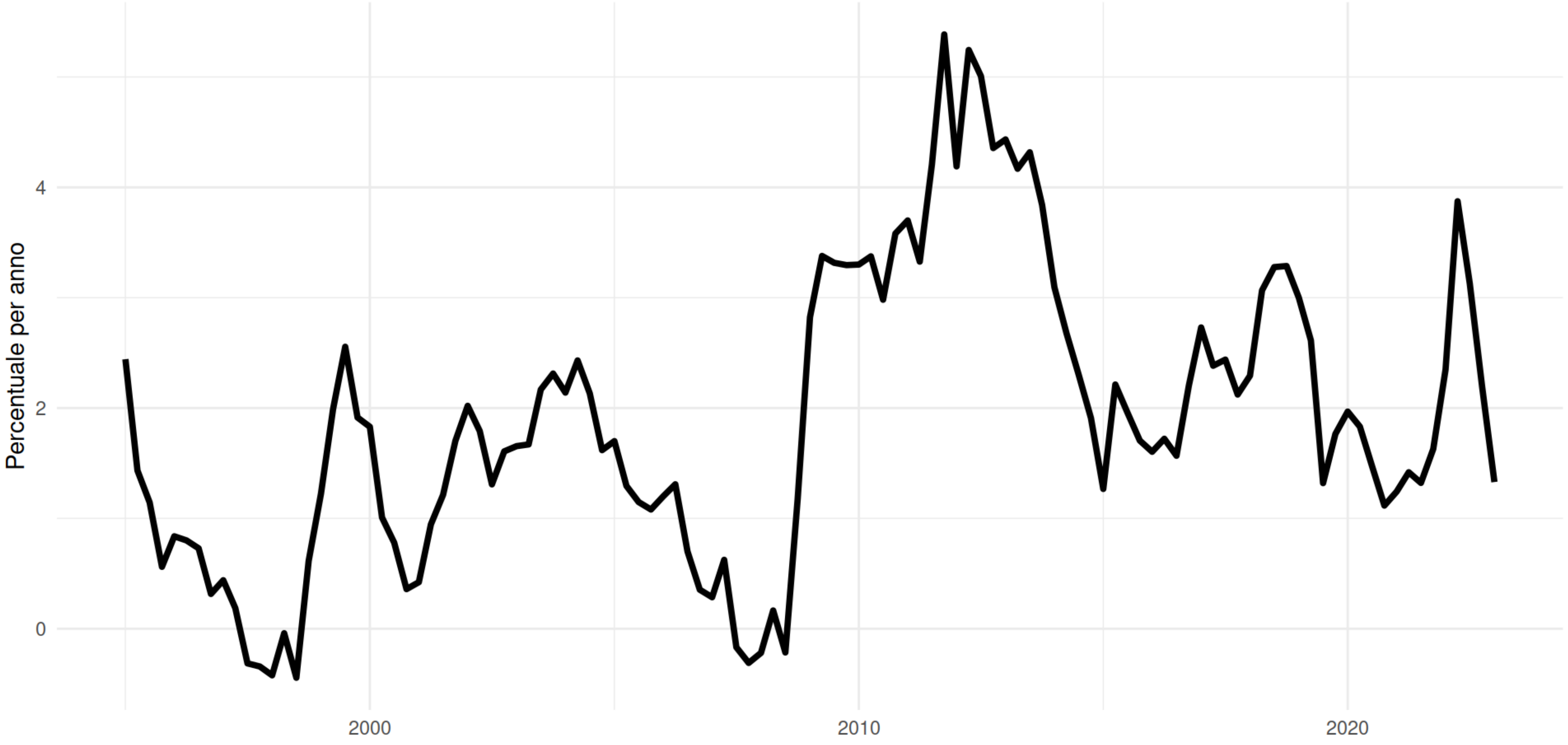
```r
# https://fred.stlouisfed.org/series/IRLTLT01ITM156N
long <- fredr("IRLTLT01ITM156N",
              frequency = "q",
              aggregation_method = "eop",
              observation_start = as.Date("1995-01-01")) |>
  rename(ilong = value)

# https://fred.stlouisfed.org/series/IR3TIB01ITM156N
short <- fredr("IR3TIB01ITM156N",
              frequency = "q",
              aggregation_method = "eop",
              observation_start = as.Date("1995-01-01")) |>
  rename(ishort = value)

intr <- inner_join(long, short, by  = "date") |>
  select(date, ishort, ilong) |>
  mutate(TSpread = ilong-ishort,
         TSpreadl1 = lag(TSpread),
         TSpreadl2 = lag(TSpread,2))

pre_covid_pil_ita <- left_join(pre_covid_pil_ita, intr)
```

```r
ggplot(intr, aes(y=ilong, x=date)) +
  geom_line(aes(lty="Tasso d'interesse a 10 anni"), lwd = 1.35) +
  geom_line(aes(y=ishort, lty = "Tasso d'interesse a 3 mesi"), lwd = 1.35) +
  ylab("Percentuale per anno") + xlab("") +
  scale_linetype_manual(name="", breaks = c("Tasso d'interesse a 10 anni", "Tasso d'interesse a 3 mesi"),
                        values = c(1,2)) +
  theme_minimal() + theme(legend.position = "bottom")
```



Tasso d’interesse a 3 mesi e asso d’interesse a 10 anni

```r
ggplot(intr, aes(y=TSpread, x=date)) +
  geom_line(lwd = 1.35) + ylab("Percentuale per anno") + xlab("") + theme_minimal()
```



Modello ADL(2,2)

```r
ADL22 <- lm(DY ~ DYl1 + DYl2 + TSpreadl1 + TSpreadl2, data = pre_covid_pil_ita)
summary(ADL22)
```

```

Call:
lm(formula = DY ~ DYl1 + DYl2 + TSpreadl1 + TSpreadl2, data = pre_covid_pil_ita)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2281 -0.3599  0.0669  0.4269  1.2558 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.1854     0.1647    1.13  0.26326    
DYl1          0.4190     0.1038    4.04  0.00011 ***
DYl2          0.2398     0.1095    2.19  0.03100 *  
TSpreadl1     0.0127     0.1450    0.09  0.93044    
TSpreadl2    -0.0082     0.1343   -0.06  0.95143    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.614 on 91 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.359, Adjusted R-squared:  0.331 
F-statistic: 12.7 on 4 and 91 DF,  p-value: 2.83e-08
```

Statistica F per coefficienti sui ritardi di TSpread:

```r
library(car)
linearHypothesis(ADL22, c("TSpreadl1=0", "TSpreadl2=0"))
```

```
Linear hypothesis test

Hypothesis:
TSpreadl1 = 0
TSpreadl2 = 0

Model 1: restricted model
Model 2: DY ~ DYl1 + DYl2 + TSpreadl1 + TSpreadl2

  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     93 34.3
2     91 34.3  2   0.00403 0.01   0.99
```

Il test dell'ipotesi congiunta che nessuna delle X sia un predittore utile, oltre ai valori pasasti di Y, si chiama test di causalità di Granger

CONCETTO CHIAVE 14.7 **Test di causalità di Granger**

La statistica per il test di causalità di Granger è la statistica F per la verifica dell'ipotesi nulla che i coefficienti su tutti i valori di una delle variabili dell'equazione (14.20) (per esempio, i coefficienti di X_{1t-1}, X_{1t-2},...,X_{1t-q_1}) siano pari a zero. Questa ipotesi nulla implica che i regressori non abbiano ulteriore potere predittivo per Y_t rispetto a quello già posseduto dagli altri regressori; il test di questa ipotesi nulla viene detto test di causalità di Granger.

"Causalità" è un termine sfortunato in questo caso: la causalità di Granger si riferisce semplicemente al contenuto predittivo (marginale)

Incertezza e intervalli di previsione

Per costruire intervalli di previsione serve una misura dell'incertezza di previsione.

Consideriamo la previsione

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \hat{\delta}_1 X_T$$

L'errore di previsione è:

$$\hat{Y}_{T+1|T} - Y_{T+1} = u_{T+1} - (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta_1) X_T$$

L'errore di previsione quadratico medio (MSFE) è

$$MSFE = E\left[\left(\hat{Y}_{T+1|T} - Y_{T+1}\right)^2\right] = \underbrace{E\left[(u_{T+1})^2\right]}_{\text{errore modello}} - \underbrace{E\left[\left((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta\right.\right.}_{\text{errore dovuto alla stima dei parametri}}$$

La radice quadrata dell'errore di previsione quadratico medio (RMSFE) :

$$RMSFE = \sqrt{MSFE}$$

Tre modi per stimare l'RMSFE

1. Usare l'approssimazione $RMSFE \approx var(u_{T+1}) \approx \frac{1}{T}\sum_{i=1}^{T}\hat{u}_t^2$, così da stimare l'RMSFE mediante il SER

2. Usare un'effettiva cronologia di previsione per $\hat{t} = t+1, \ldots T$, cioé $\hat{Y}_{\hat{t}|\hat{t}}$ e quindi stimare RMSE mediante

$$\widehat{RMSE} = \sqrt{\frac{1}{T-t-1}\sum_{\hat{t}=t+1}^{T}(Y_{\hat{t}} - \hat{Y}_{\hat{t}|\hat{t}-1})^2}$$

   Solitamente non è pratico – richiede la disponibilità di una registrazione storica di previsioni effettive dal modello

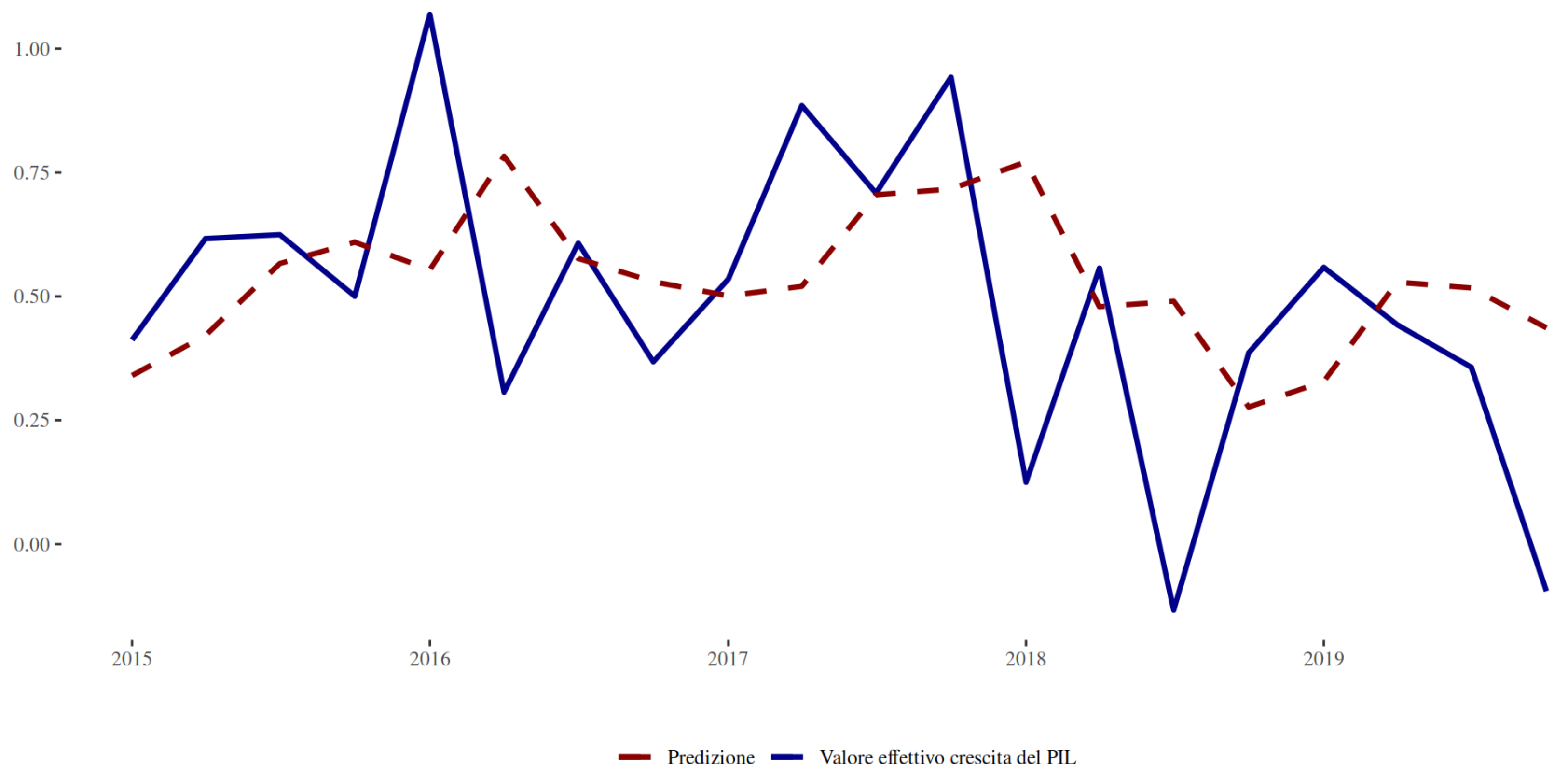
3. Usare una cronologia di previsione simulata, cioè che simuli le previsioni che si sarebbero fatte usando il modello in tempo reale quindi usare il metodo 2, con queste pseudo previsioni fuori campione

Il metodo delle pseudo previsioni fuori campione

```r
Y0 <- Yhat <- error <- YT <- NULL
for (j in 0:19) {
  if (j < 19)
    Y0[j+1] <- pre_covid_pil_ita[80+j+1,] |> pull(DY)
  else
    Y0[j+1] <- pil_ita |> filter(date == as.Date("2019-10-01")) |> pull(DY)

  df <- pre_covid_pil_ita[1:(80+j),]
  ## Stimare il modello per ogni periodo
  m <- lm(DY~DYl1+DYl2+TSpreadl1+TSpreadl2, data=df)
  ## Calcolare la “previsione” per la data  t+1 usando la stima fino a t
  YT[[j+1]] <- as.list(tail(df,1) |> select(DY, DYl1, TSpread, TSpreadl1))
  names(YT[[j+1]]) <- c("DYl1", "DYl2", "TSpreadl1", "TSpreadl2")
  Yhat[j+1] <- predict(m, newdata = YT[[j+1]])
  ## Calcolare l’errore della pseudo previsione fuori campione
  error[j+1] = Y0[j+1] - Yhat[j+1]
}
RMSE <- sqrt(mean(error^2))
RMSE
```

```
[1] 0.313
```

```r
date <- c(pre_covid_pil_ita$date[81:99], as.Date("2019-10-01"))
df <- tibble(date = date, Y0 = Y0, Yhat = Yhat, error = error)
ggplot(df, aes(y=Y0, x=date)) + geom_line(aes(col="Valore effettivo crescita del PIL"),lwd=1.1) +
  geom_line(aes(y=Yhat, col = "Predizione"), lty = 2, lwd = 1.1) +
  #geom_line(aes(y=error, col = "error")) +
  scale_color_manual(name = "",
                     values = c("Valore effettivo crescita del PIL"="darkblue",
                                "Predizione"="darkred")) +
  theme_tufte() + theme(legend.position = "bottom") + ylab("") + xlab("")
```



Usare l'RMSFE per costruire intervalli di previsione

Se u_{t+1} ha distribuzione normale, allora un intervallo di previsione al 95% può essere costruito come

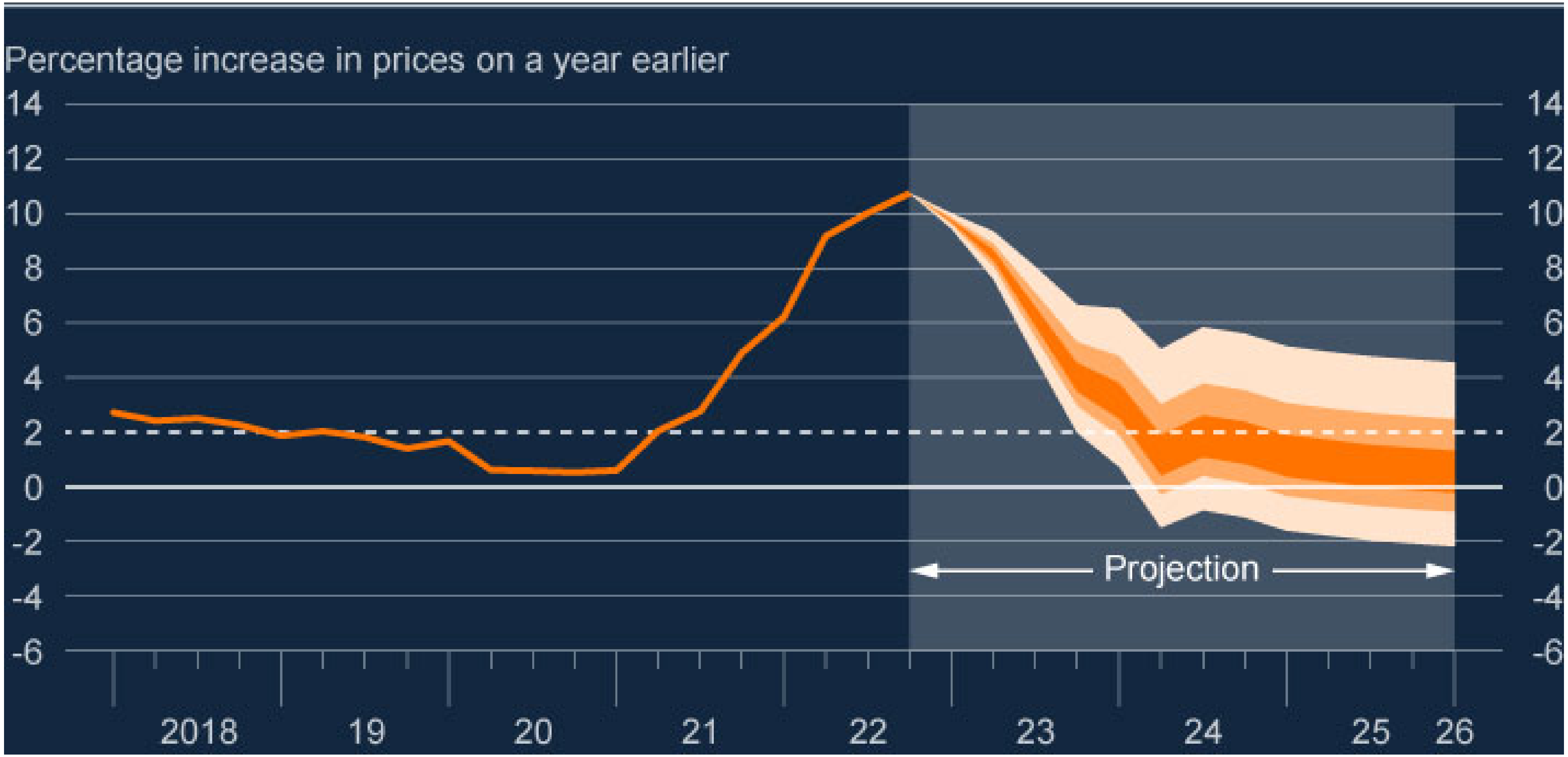
$$\hat{Y}_{t+1} \pm 1.96 \times \sqrt{RMSE}$$

Note :

- Un intervallo di previsione al 95% non è un intervallo di confidenza (Y_{T+1} non è un coefficiente non casuale, è casuale)
- Questo intervallo è valido solo se u_{T_+1} è normale – ma potrebbe comunque fornire un'approssimazione ragionevole ed è una misura comunemente usata di incertezza della previsione
- Spesso si usano intervalli di previsione "67%" :

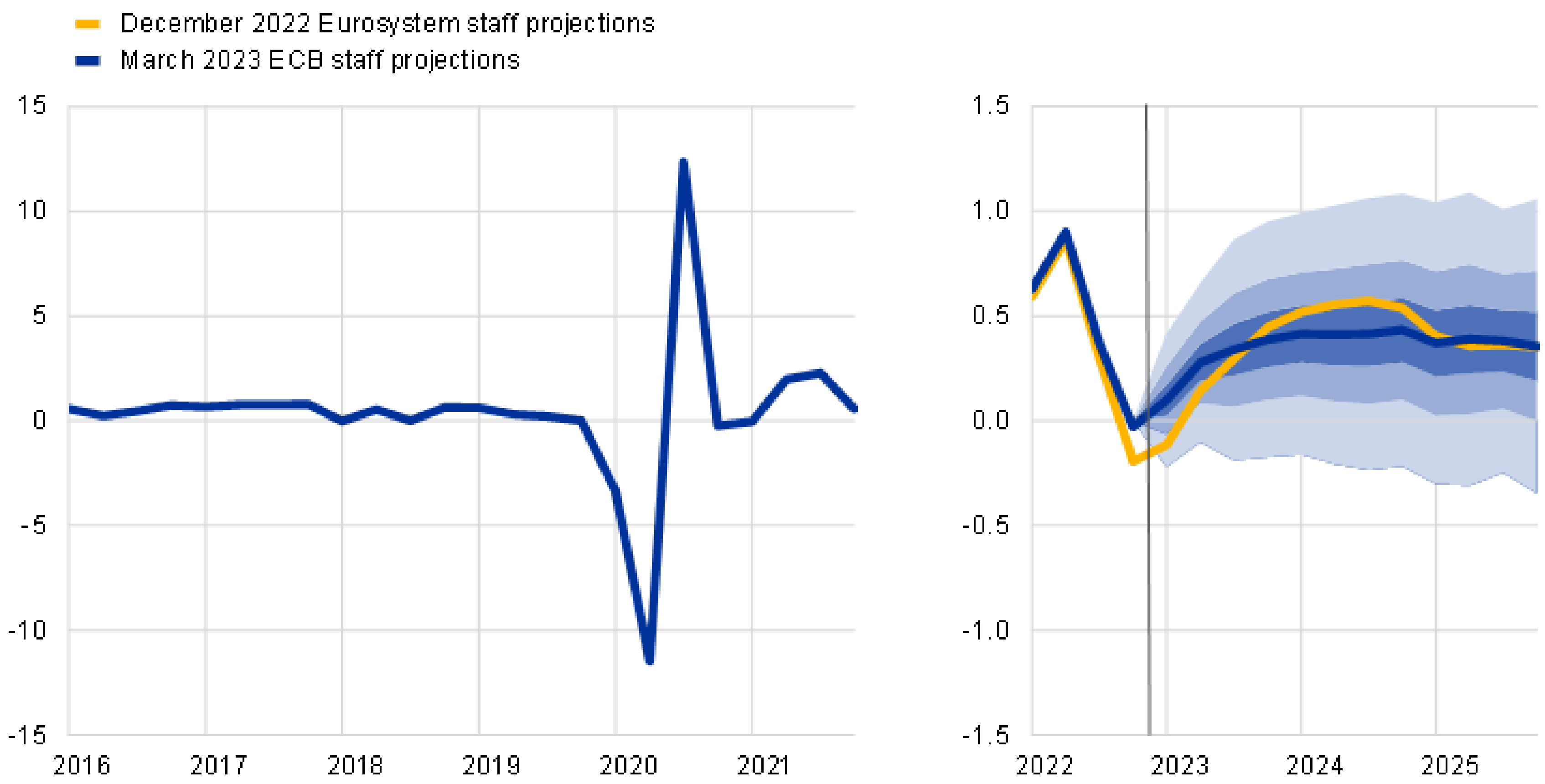
$$\hat{Y}_{t+1} \pm 0.97 \times \sqrt{RMSE}$$

Esempio 1: “grafico a ventaglio” di Bank of England, marzo 2023



https://www.bankofengland.co.uk/monetary-policy-report/2023/february-2023

Esempio 2: bollettino mensile della banca centrale europea, marzo 2023



https://www.ecb.europa.eu/pub/projections/html/ecb.projections202303_ecbstaff~77c0227058.en.htm

Scelta della lunghezza dei ritardi usando criteri d'informazione (Paragrafo 14.5)

- Come scegliere il numero di intervalli p in un AR(p)?
- La distorsione da variabili omesse è irrilevante per la previsione!
- Si possono usare sequenze di test t o F ; ma i modelli scelti tendono a essere "troppo grandi" (perché?)
- Un altro modo – migliore – per determinare la lunghezza dei ritardi è quello di usare un criterio di informazione
- I criteri di informazione bilanciano distorsione (troppo pochi ritardi) e varianza (troppi ritardi)
- Due criteri informativi sono quello di Bayes (BIC) e quello di Akaike (AIC)

Il Bayes Information Criterion (BIC)

$$BIC(p) = \log\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{\log(T)}{T}$$

dove $SSR(p)$ ě la somma dei quadrati dei residui quando l'AR ha p ritardi

Lo stimatore di p secondo il criterio BIC minimizza $BIC(p)$ fra le possibili scelte $p = 1, \ldots, p_{max}$.

- Primo termine : sempre decrescente in p (più grande è p, migliore è l'adattamento)
- Secondo terme : sempre crescente in p
  - La varianza della previsione dovuta all'errore di stima aumenta con p – perciò non si vuole un modello di previsione con troppi coefficienti
  - Questo termine è una "penalità" per l'uso di più parametri – che aumenta la varianza della previsione.
- _Minimizzando il BIC(p) si bilanciano distorsione e varianza per determinare un valore "migliore" di p per la previsione

Un altro criterio di informazione:♪Akaike Information Criterion (AIC)

$$BIC(p) = \log\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T}$$

- Il termine di penalità è più piccolo per l' AIC rispetto al BIC $2 < logT$
  - AIC stima più ritardi (p più grande) del BIC
  - e lo stimatore AIC di p non è consistente – può sovrastimare p – la penalità non è abbastanza grande

Modello AR della crescita del PIL

```r
p_max <- 6
bic <- aic <- NULL
frm <- DY ~ 1
BIC <- function(m) {
  u <- resid(m)
  T <- length(u)
  p <- length(coef(m))-1
  log(mean(u^2)) + (p+1)*log(T)/T
}

AIC <- function(m) {
  u <- resid(m)
  T <- length(u)
  p <- length(coef(m))-1
  log(mean(u^2)) + (p+1)*2/T
}

for (j in 1:p_max) {
  lags <- paste0("I(lag(DY,", 1:j, "))",
                 collapse = "+")
  frm <- as.formula(paste0("DY ~", lags))
  m <- lm(frm, data = pre_covid_pil_ita)
  bic[j] <- BIC(m)
  aic[j] <- AIC(m)
}
tibble(p=1:6, `BIC(p)`=bic,`AIC(p)`=aic) |>
  kable(digits = 13) |>
  kable_styling()
```

p	BIC(p)	AIC(p)
1	-0.878	-0.931
2	-0.885	-0.965
3	-0.872	-0.980
4	-0.815	-0.950
5	-0.760	-0.923
6	-0.699	-0.891

Generalizzazione del BIC a modelli multivariati (ADL)

Sia K = numero totale di coefficienti nel modello (intercetta, ritardi di Y, ritardi di X).

Il BIC è

$$BIC(K) = \log\left[\frac{SSR(K)}{T}\right] + K\frac{2}{T}$$

- Lo si può calcolare su tutte le possibili combinazioni di ritardi di Y e di X (ma sono tante)!

- In pratica si potrebbero scegliere ritardi di Y con il BIC, e decidere se includere o meno X usando un test di causalità di Granger con un numero fisso di ritardi (dipendente da dati e applicazione)