Séries de Tempo

Aula 7 - Modelos multivariados

Regis A. Ely

Departamento de Economia
Universidade Federal de Pelotas

04 de setembro de 2020

Conteúdo

Regressão linear com séries de tempo

Se tivermos duas séries de tempo Y_t e X_t, podemos relacioná-las através de uma regressão linear?

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

O que muda com o subscrito t?

- Para garantirmos que o coeficiente β não é viesado, precisamos ter $Cov(X_t, \varepsilon_t) = 0$
- Agora temos uma dinâmica: $Y_{t-1} = \alpha + \beta X_{t-1} + \varepsilon_{t-1}$, que deve valer para todo t

Regressão linear com séries de tempo

E se tivermos $Cov(\varepsilon_t, \varepsilon_{t-1}) \neq 0$?

- Na regressão com dados transversais isso pode ser resolvido com erros robustos
- Em regressões com séries de tempo, a correlação dos resíduos se torna um problema maior
- Suponha que $\varepsilon_t = b\varepsilon_{t-1} + \eta_t$, sendo η_t um ruído branco não correlacionado, então:

$$Cov(X_t, \varepsilon_t) = Cov(X_t, b\varepsilon_{t-1} + \eta_t), \text{ e}$$

$$Cov(X_t, \varepsilon_{t-1}) = Cov(X_t, Y_{t-1} - \alpha - \beta X_{t-1})$$

Regressão linear com séries de tempo

- Logo, $Cov(X_t, \varepsilon_t) = 0$ só será válido se $Cov(X_t, X_{t-1}) = 0$
- Ao estudar modelos univariados, vimos que raramente as séries de tempo não apresentam autocorrelação
- **Conclusão**: em regressões lineares com séries de tempo, os problemas de endogeneidade, que geram coeficientes viesados, estão refletidos na correlação serial dos resíduos
- Assim, o principal diagnóstico para checarmos a validade de um modelo multivariado de séries de tempo continua sendo a correlação residual

Regressão linear com séries de tempo

A regressão linear $Y_t = \alpha + \beta X_t + \varepsilon_t$ sempre estará errada?

- Nem sempre, embora na maior parte dos casos sim
- Se os resíduos resultantes da regressão forem estacionários e não autocorrelacionados esta regressão é válida
- Mas e se os resíduos forem autocorrelacionados?

Modelos ADL

Suponha que você estime a regressão $Y_t = \alpha + \beta X_t + \varepsilon_t$ mas os resíduos ε_t sejam autocorrelacionados, de modo que $\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t$, sendo η_t um ruído branco. Assim:

$$Y_t = \alpha + \beta X_t + \rho \varepsilon_{t-1} + \eta_t$$

E como $\varepsilon_{t-1} = Y_{t-1} - \alpha - \beta X_{t-1}$, então:

$$Y_t = \alpha + \beta X_t + \rho Y_{t-1} - \rho\alpha - \rho\beta X_{t-1} + \eta_t$$

Modelos ADL

A partir da equação anterior pode-se especificar um modelo ADL(1,1):

$$Y_t = \alpha + \rho Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$$

Onde ADL significa *Autoregressive Distributed Lags*, sendo os números as ordens das defasagens da variável dependente e explicativa (exógena)

- Corrigimos o problema da autocorrelação dos resíduos adicionando defasagens da variável dependente e explicativa

Modelos ADL

- A seleção do número de defasagens a incluir nos modelos ADL pode ser feita através de critérios de identificação (Akaike, BIC, etc) ou diagnóstico dos resíduos
- Como vimos, há duas maneiras de estimar os modelos ADL:
    1. Especificar as defasagens das variáveis dependentes e exógenas: $Y_t = \alpha + \rho Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$
    2. Especificar a estrutura de autocorrelação dos resíduos: $Y_t = \alpha + \beta X_t + \varepsilon_t$ e $\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t$
- Essas duas formas são equivalentes mas geram interpretações distintas para os coeficientes

Modelos ADL

- No primeiro caso, os coeficientes β_0 e β_1 são os efeitos de X_t e X_{t-1} em Y_t condicionais ao valores passados $X_{t-2}, X_{t-3}, \ldots$
- No segundo caso, o coeficiente β é o efeito total de X_t em Y_t, considerando todos os efeitos que valores defasados de X_t tiveram em Y_t
- Uma terceira possibilidade na estimação dos modelos ADL é usar uma *transfer function*[1], que é uma generalização destes dois modelos e permite que estimar efeitos com decaimento ao longo do tempo ou outras formas funcionais para a relação dinâmica de X_t e Y_t

[1] Mais informações sobre estes três tipos de modelo pode ser obtida em `https://robjhyndman.com/hyndsight/arimax/`

Cinto de segurança e mortes no trânsito

Neste exemplo vamos utilizar a base de dados `Seatbelts`, que contém o número de motoristas que foram mortos em acidentes de trânsito no Reino Unido durante o período de janeiro de 1969 a dezembro de 1984, sendo que em 31 de janeiro de 1983 foi introduzida uma lei que tornou compulsório o uso de cinto de segurança

```r
library(tidyverse)
library(tsibble)
library(feasts)
library(fable)
base <- as_tsibble(Seatbelts) %>%
  filter(key %in% c("DriversKilled", "kms", "PetrolPrice", "law"))
```

Nosso modelo terá apenas quatro variáveis, o número de motoristas que foram mortos, a distância percorrida em quilômetros, o preço da gasolina e a dummy que identifica a implementação da lei

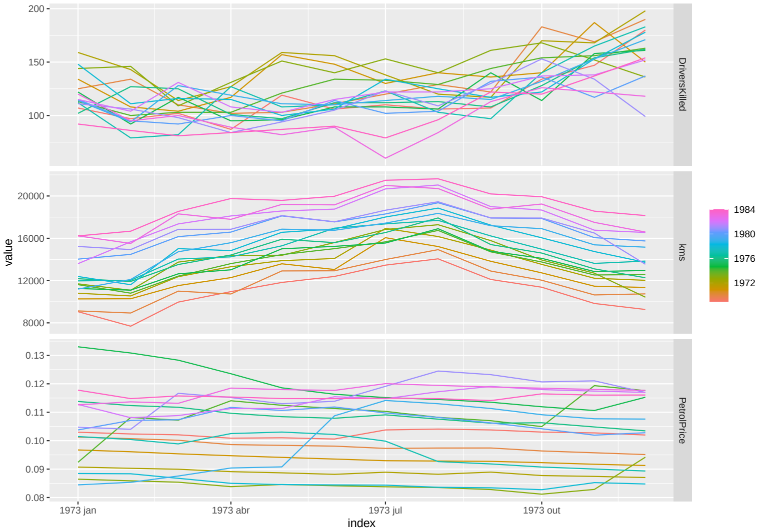
Sazonalidade

Primeiro devemos checar se as séries de tempo apresentam sazonalidade, pois isto pode afetar a relação entre as variáveis (*temporal confounder*):

```
base %>%
  filter(key != "law") %>%
  gg_season(value, period = 12)
```

Vamos inspecionar os gráficos sazonais das variáveis, com exceção da dummy de implementação da lei

Sazonalidade



Sazonalidade

Parece que as séries do número de motoristas mortos e da distância percorrida apresentam sazonalidade, enquanto que o preço da gasolina não. Podemos confirmar isso usando um dos testes de sazonalidade que vimos na Aula 2:

```
library(seastests)
base %>%
  group_by(key) %>%
  group_map(~isSeasonal(.x$value, freq = 12)) %>%
  `names<-`(unique(base$key))
```

```
##      DriversKilled kms  PetrolPrice law
## [1,] TRUE          TRUE FALSE       FALSE
```

Sazonalidade

- Para remover a sazonalidade vamos utilizar o método X11-ARIMA
- No código do próximo slide:
  1. Estimamos os componentes sazonais para as séries `DriversKilled` e `kms`
  2. Juntamos estas séries dessazonalizadas com as séries originais de `PetrolPrice` e `law`
  3. Criamos uma nova variável `value` que contém o log das séries dessazonalizadas para `DriversKilled` e `kms` e o log da série original para `PetrolPrice` (não utilizamos log da variável *dummy*)
  4. Colocamos os dados em formato `wide` para utilizarmos as séries como variáveis explicativas nos modelos

Sazonalidade

```r
base <- base %>%
  filter(key %in% c("DriversKilled", "kms")) %>%
  model(feasts:::X11(value)) %>%
  components() %>%
  select(index, key, value_adj = season_adjust) %>%
  right_join(base, by = c("index", "key")) %>%
  group_by(key) %>%
  mutate(
    value = case_when(
      key %in% c("DriversKilled", "kms") ~ log(value_adj),
      key == "PetrolPrice" ~ log(value),
      TRUE ~ value
    )
  ) %>%
  ungroup() %>%
  pivot_wider(-value_adj, names_from = "key", values_from = "value")
```

Estacionariedade

- Por enquanto não iremos nos preocupar com estacionariedade das séries de tempo
- Porém, será essencial checar a estacionariedade e autocorrelação dos resíduos das nossas regressões
- Se tivermos resíduos com raíz unitária ou autocorrelação a nossa regressão está mal especificada
- Na próxima aula veremos caso a caso os problemas ocasionados pela combinação de séries não estacionárias em modelos de regressão

Modelo de Regressão Linear

O primeiro modelo que podemos estimar é uma regressão linear, sem considerar as autocorrelações entre as séries:

```r
reg_linear <- base %>%
  model(LM = TSLM(DriversKilled ~ kms + PetrolPrice + law))
report(reg_linear)
```

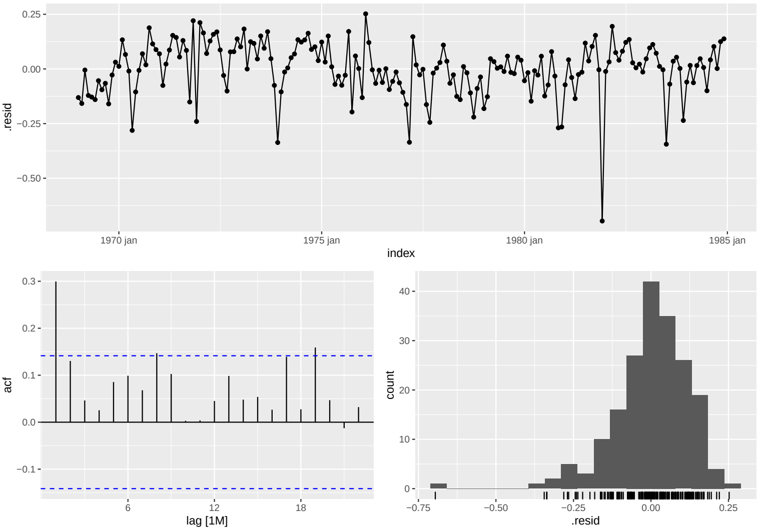
```
...
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.71629    0.75327   6.261 2.53e-09 ***
## kms         -0.09890    0.07086  -1.396 0.164453
## PetrolPrice -0.46117    0.08457  -5.453 1.54e-07 ***
## law         -0.13373    0.03394  -3.940 0.000115 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1245 on 188 degrees of freedom
## Multiple R-squared: 0.3577,  Adjusted R-squared: 0.3475
## F-statistic: 34.91 on 3 and 188 DF, p-value: < 2.22e-16
...
```

Modelo de Regressão Linear

- Inicialmente parece que a lei teve um impacto negativo no número de mortes em acidentes de trânsito da ordem de 12.5% $(1 - e^{-0.13373})$
- Para verificar se este modelo está bem especificado, podemos inspecionar os resíduos e suas funções de autocorrelação através do comando `gg_tsresiduals`

```
gg_tsresiduals(reg_linear)
```

Modelo de Regressão Linear



Modelo ADL

Os resíduos ainda apresentam alguma autocorrelação. Podemos tentar melhorar este modelo estimando um modelo ADL(1,1):

```
reg_armax <- base %>%
  model(
    LM = TSLM(
      DriversKilled ~ lag(DriversKilled) + kms + lag(kms) +
        PetrolPrice + lag(PetrolPrice) + law
    )
  )
tidy(reg_armax)
```

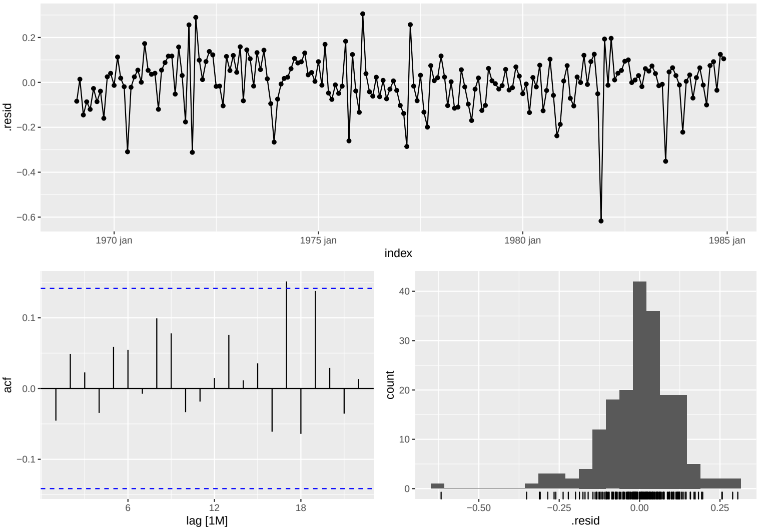
```
## # A tibble: 7 x 6
##   .model term               estimate std.error statistic   p.value
##   <chr>  <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 LM     (Intercept)          3.60      0.803       4.48 0.0000131
## 2 LM     lag(DriversKilled)   0.303     0.0690      4.39 0.0000190
## 3 LM     kms                  0.328     0.221       1.48 0.140
## 4 LM     lag(kms)            -0.428     0.222      -1.93 0.0552
## 5 LM     PetrolPrice         -0.233     0.287      -0.812 0.418
## 6 LM     lag(PetrolPrice)    -0.0816    0.289      -0.283 0.778
## 7 LM     law                 -0.0873    0.0335     -2.60 0.00997
```

Modelo ADL

- Neste modelo, o efeito da lei cai para 8.36% $(1 - e^{-0.0.0873})$
- Entretanto, este é um efeito de curto prazo agora, dada a inclusão da defasagem da variável dependente
- Para verificar se este modelo está bem especificado, podemos inspecionar os resíduos e suas funções de autocorrelação através do comando `gg_tsresiduals`

```
gg_tsresiduals(reg_armax)
```

Modelo ADL



.resid
0.2
0.0
−0.2
−0.4
−0.6
1970 jan
1975 jan
1980 jan
1985 jan
index
acf
0.1
0.0
−0.1
6
12
18
lag [1M]
count
40
30
20
10
0
−0.50
−0.25
0.00
0.25
.resid

Modelo ADL

- Os resíduos parecem menos correlacionados do que o modelo de regressão linear
- Entretanto, a interpretação do coeficiente de interesse mudou
- Podemos resolver este problema estimando um modelo com erros correlacionados através da função `ARIMA`, especificando as ordens `(p,d,q)` do modelo a ser estimado

Modelo com erros autocorrelacionados

```r
reg_arma_errors <- base %>%
  model(
    LM = ARIMA(
      DriversKilled ~ pdq(1,0,0) + kms + PetrolPrice + law
    )
  )
tidy(reg_arma_errors)
```

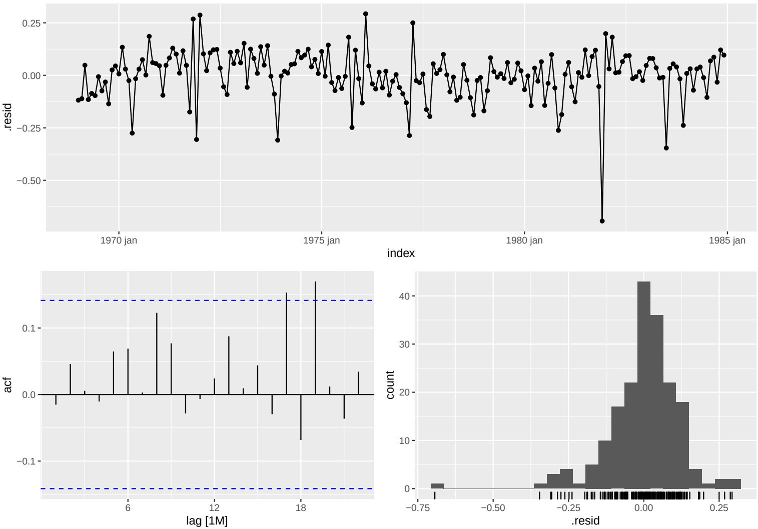
```
## # A tibble: 5 x 6
##   .model term        estimate std.error statistic    p.value
##   <chr>  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 LM     ar1            0.304    0.0694     4.38  0.0000196
## 2 LM     kms           -0.0737   0.0924    -0.798 0.426
## 3 LM     PetrolPrice   -0.465    0.111     -4.18  0.0000437
## 4 LM     law           -0.136    0.0444    -3.05  0.00260
## 5 LM     intercept      4.47     0.980      4.56  0.00000920
```

Modelo com erros autocorrelacionados

- Neste modelo, o efeito da lei volta para patamares próximos ao modelo de regressão linear, sendo 12.7% $(1 - e^{-0.136})$
- Agora estamos calculando o efeito total que a lei teve sobre a variável dependente
- Para verificar se este modelo está bem especificado, podemos inspecionar os resíduos e suas funções de autocorrelação através do comando `gg_tsresiduals`

```
gg_tsresiduals(reg_arma_errors)
```

Modelo com erros autocorrelacionados



Modelo com erros autocorrelacionados

- Os resíduos estão um pouco melhores do que o modelo de regressão linear
- Outras especificações podem ser testadas, como por exemplo um modelo ADL(0,1), em que não são incluídas defasagens da variável dependente, apenas das variáveis explicativas
- Uma outra alternativa é estimar modelos de *transfer function* através da função `arimax` do pacote `TSA`
    - Este tipo de modelo é uma generalização dos modelos ADL em que pode-se especificar efeitos graduais ao longo do tempo por exemplo

Previsão com variáveis exógenas

- Quando se faz previsões que dependem de outras variáveis, você precisará conhecer os valores futuros delas para fazermos previsões da variável dependente
- Normalmente não conhecemos os valores futuros dessas variáveis, logo, podemos contornar isso de duas maneiras:
    1. Construir cenários para os valores futuros dos previsores e então utilizar estes valores prever a variável dependente
    2. Fazer previsões univariadas para cada previsor e então utilizá-las para prever a variável dependente

Modelos ECM

O modelo de correção de erros (ECM) pode ser descrito como:

$$\Delta Y_t = \beta_0 \Delta X_t + \gamma[Y_{t-1} - \delta X_{t-1}] + \varepsilon_t$$

onde:

- β_0 captura a relação de curto prazo entre Y_t e X_t
- γ captura a taxa pela qual o modelo volta ao equilíbrio de longo prazo (proporção de desequilíbrio corrigida a cada instante t)
- Se $\gamma = 0$ não há equilíbrio de longo prazo, e se $\gamma = -1$, o desequilíbrio sempre se corrige em apenas um período
- Devemos ter $\gamma \leq 0$ e $|\gamma| < 1$

Equivalência entre modelos ADL e ECM

Podemos mostrar que os modelos ADL e ECM são equivalentes começando com um modelo ADL(1,1)[2]:

$$Y_t = \beta_0 X_t + \beta_1 X_{t-1} + \rho Y_{t-1} + \eta_t$$

Então substraímos Y_{t-1} de ambos os lados:

$$Y_t - Y_{t-1} = \beta_0 X_t + \beta_1 X_{t-1} + (\rho - 1) Y_{t-1} + \eta_t$$

E definindo $(\rho - 1) = \gamma$:

$$\Delta Y_t = \beta_0 X_t + \beta_1 X_{t-1} + \gamma Y_{t-1} + \eta_t$$

[2]Vamos omitir a constante por conveniência.

Equivalência entre modelos ADL e ECM

Fazendo $X_t = \Delta X_t + X_{t-1}$ teremos:

$$\Delta Y_t = \beta_0 \Delta X_t + (\beta_0 + \beta_1) X_{t-1} + \gamma Y_{t-1} + \eta_t$$

E rearranjando os termos:

$$\Delta Y_t = \beta_0 \Delta X_t + \gamma [Y_{t-1} + \frac{(\beta_0 + \beta_1)}{\gamma} X_{t-1}] + \eta_t$$

Logo, definindo $\delta = \frac{-(\beta_0 + \beta_1)}{\gamma}$ temos o modelo ECM mostrado anteriormente, lembrando que $\gamma = (\rho - 1)$

Resumo dos modelos ADL

Vários modelos lineares multivariados são casos específicos de um modelo ADL com erros AR(1):

$$Y_t = \beta X_t + \beta_1 X_{t-1} + \rho Y_{t-1} + \varepsilon_t, \text{ com}$$
$$\varepsilon_t = \phi \varepsilon_{t-1} + \eta_t$$

Exemplos:

- *Regressão linear simples*: $\beta_1 = \rho = \phi = 0$
- *Regressão linear com erro AR(1)*: $\beta_1 = \rho = 0$
- *Finite distributed lags*: $\rho = \phi = 0$
- *Lagged dependent variable*: $\beta_1 = \phi = 0$
- *Modelos ADL e ECM*: $\phi = 0$