Modelos de Regressão e Previsão

Análise de regressão com dados de séries temporais

Prof. Carlos Trucíos
carlos.trucios@facc.ufrj.br
ctruciosm.github.io

Faculdade de Administração e Ciências Contábeis
Universidade Federal do Rio de Janeiro

Aula 12

Introdução

Introdução

- Até agora estudamos o MRL em um contexto de dados de **corte transversal**
- Sob HRLM1–HRLM5, os $\hat{\beta}$'s são os melhores estimadores lineares não viesados e se incluirmos HRLM6, inferência estatística exata é possivel
- Dados de séries temporais se diferenciam de dados de corte transversal em que eles tem uma ordenação temporal
    - Temperatura
    - PBI
    - Indice da bolsa de valores de SP
    - Inflação
    - Taxa de desemprego
    - Demanda de produtos
- Estudaremos MRL num contexto de **séries temporais**, estabeleceremos novas hipóteses e propriedades dos estimadores

Introdução

Sejam y e z duas séries temporais

Modelos estáticos (ME)

Um modelo estático, relaciona y e z da forma:

$$y_t = \beta_0 + \beta_1 z_t + u_t \quad t = 1, \ldots, n$$

Modelos de defasagem distribuida finita (MDDF)

No MDDF, permitimos que as variáveis afetem y com defasgens,

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \cdots + \delta_k z_{t-k} + u_t$$

No ME, a interpretação no modelo estatístico não muda,já para interpretar o MDDF precisamos tomar maiores cuidados.

Hipóteses do modelo

Hipóteses do modelo

No MRL com dados de **corte transversal** tinhamos:

- **HRLM1**: O modelo populacional é linear nos parâmetros, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- **HRLM2**: $(y_1, x_{1,1}, \ldots, x_{1,k}), \cdots, (y_n, x_{n,1}, \ldots, x_{n,k})$ constituem uma a.a. de tamanho n do modelo populacional
- **HRLM3**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HRLM4**: $E(u|X) = 0$
- **HRLM5**: Os erros tem variância constante (homocedasticidade)
- **HRLM6**: O erro populacional u é independente das variaveis explicativas (X) e $u \sim N(0, \sigma^2)$

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \; t = 1, \ldots, n$

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$, $t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$, $t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0$, $t = 1, \ldots, n$ (implica que u_t é não correlacionado as v. explicativas em todos os períodos de tempo)

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$, $t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0$, $t = 1, \ldots, n$ (implica que u_t é não correlacionado as v. explicativas em todos os períodos de tempo)
- **HST4**: Homocedasticidade, $V(u_t|X) = V(u_t) = \sigma^2$, $t = 1, \ldots, n$

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \ t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0, \ t = 1, \ldots, n$ (implica que u_t é não correlacionado as v. explicativas em todos os períodos de tempo)
- **HST4**: Homocedasticidade, $V(u_t|X) = V(u_t) = \sigma^2, \ t = 1, \ldots, n$
- **HST5**: $Corr(u_t, u_s|X) = 0, \ \forall t \neq s$ (sem correlação serial)

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$, $t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0$, $t = 1, \ldots, n$ (implica que u_t é não correlacionado as v. explicativas em todos os períodos de tempo)
- **HST4**: Homocedasticidade, $V(u_t|X) = V(u_t) = \sigma^2$, $t = 1, \ldots, n$
- **HST5**: $Corr(u_t, u_s|X) = 0$, $\forall t \neq s$ (sem correlação serial)
- **HST6**: Os erros u_t são independentes de X e $u_t \sim N(0, \sigma^2)$

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \ t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0, \ t = 1, \ldots, n$ (implica que u_t é não correlacionado as v. explicativas em todos os períodos de tempo)
- **HST4**: Homocedasticidade, $V(u_t|X) = V(u_t) = \sigma^2, \ t = 1, \ldots, n$
- **HST5**: $Corr(u_t, u_s|X) = 0, \ \forall t \neq s$ (sem correlação serial)
- **HST6**: Os erros u_t são independentes de X e $u_t \sim N(0, \sigma^2)$

Hipóteses do modelo

Quando trabalhamos com **séries temporais**:

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$, $t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0$, $t = 1, \ldots, n$ (implica que u_t é não correlacionado as v. explicativas em todos os períodos de tempo)
- **HST4**: Homocedasticidade, $V(u_t|X) = V(u_t) = \sigma^2$, $t = 1, \ldots, n$
- **HST5**: $Corr(u_t, u_s|X) = 0$, $\forall t \neq s$ (sem correlação serial)
- **HST6**: Os erros u_t são independentes de X e $u_t \sim N(0, \sigma^2)$

- Sob HST1–HST3, $E(\hat{\beta}) = \beta$
- Sob HST1–HST5, $\hat{\beta}$ é BLUE
- HST6, nos permite fazer inferência.

Forma funcional e dummy

Forma funcional e dummy

A inclussão de formas funcionais e variáveis dummy são também válidas em um contexto de séries temporais.

[1] avgcov: proporcao dos trabalhadores cobertos pela lei do salário mínimo

Forma funcional e dummy

A inclussão de formas funcionais e variáveis dummy são também válidas em um contexto de séries temporais.

Exemplo Seja o modelo

$$\log(prepop_t) = \beta_0 + \beta_1 \log(mincov_t) + \beta_2 \log(usgnp_t) + u_t$$

em que

- $prepop_t$ é a taxa de empreg em Porto Rico no ano t
- $usgnp_t$ é o produto nacional bruto real dos Estados Unidos
- $mincov_t = (avgmin/avgwage) * avgcov$ mede a importância dos salários minimos em relação aos salários médios[1]

[1] avgcov: proporcao dos trabalhadores cobertos pela lei do salário mínimo

Forma funcional e dummy

```
library(wooldridge)
library(dplyr)
prminwge %>% select(year,prepop, mincov,usgnp) %>% glimpse()
```

```
## Rows: 38
## Columns: 4
## $ year   <int> 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957
## $ prepop <dbl> 0.470, 0.449, 0.434, 0.428, 0.415, 0.419, 0.41
## $ mincov <dbl> 0.09999498, 0.10551952, 0.12078384, 0.14966875
## $ usgnp  <dbl> 1203.7, 1328.2, 1380.0, 1435.3, 1416.2, 1494.9
```

Forma funcional e dummy

```
modelo = lm(log(prepop)~log(mincov) + log(usgnp), data = prminwge
round(summary(modelo)$coef,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0544     0.7654 -1.3776   0.1771
## log(mincov)  -0.1544     0.0649 -2.3797   0.0229
## log(usgnp)   -0.0122     0.0885 -0.1377   0.8913
```

- O aumento em 1% em mincov, implica, em média, a diminuição de prepop em -0.1544 %
- O aumento em 1% em usgnp, implica, em média, a diminuição de prepop em -0.0122 % (porém, não é estatisticamente significativa)

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)
  - antes e depois do 11-09-2001

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)
  - antes e depois do 11-09-2001
  - antes e depois de determinada politica na empresa

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)
  - antes e depois do 11-09-2001
  - antes e depois de determinada politica na empresa
  - antes e depois de determinada politica publica

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)
  - antes e depois do 11-09-2001
  - antes e depois de determinada politica na empresa
  - antes e depois de determinada politica publica
  - etc

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)
  - antes e depois do 11-09-2001
  - antes e depois de determinada politica na empresa
  - antes e depois de determinada politica publica
  - etc

Forma funcional e dummy

- Variáveis dummy são frequentemente incluidas em regressão com dados de séries temporais
- Elas são utilizadas para dividir períodos de tempo
  - posicionamento politico na presidencia: direita ou esquerda
  - período covid: antes do covid e depois do covid (março 12, 2020)
  - antes e depois do 11-09-2001
  - antes e depois de determinada politica na empresa
  - antes e depois de determinada politica publica
  - etc

Variáveis dummy nos ajudam a isolar períodos que possam ser sistemáticamente diferentes de outros períodos no conjunto de dados.

Forma funcional e dummy

Ejemplo Seja o modelo

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u_t$$

em que

- gfr_t: taxa geral de fertilidade (crianças nascidas a cada 1000 mulheres) no ano t
- pe_t: taxa de dedução de impostos no ano t
- $ww2_t$: Dummy: 1 durante o tempo que os Estados Unidos estiveram envolidos na Segunda Guerra Mundial
- $pill_t$. Dummy: 1 desde 1963 em diante (inclussão da pílula anticoncepcional)

Forma funcional e dummy

```
modelo = lm(gfr ~ pe + ww2 + pill, data = fertil3)
round(summary(modelo)$coef,5)
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  98.68176    3.20813 30.75991  0.00000
## pe            0.08254    0.02965  2.78417  0.00694
## ww2         -24.23840    7.45825 -3.24988  0.00180
## pill        -31.59403    4.08107 -7.74161  0.00000
```

- Na WW2, tivemos em média 24 nascimentos menos p/c 1000 mulheres
- Após a inclussão da pílula, o número de nascimento p/c 1000 mulheres diminuiu em 32.
- Um aumento de 12 USD em pe, implica no aumento, em média de (0.08254*12=0.99048) 1 nascimento a cada 1000 mulheres

Forma funcional e dummy

O modelo anterior

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u_t$$

é um modelo estático. Mas poderiamos também usar um modelo de defasagem distribuida finita (MDDF)

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 pe_{t-q} + \beta_3 pe_{t-1} + \beta_4 ww2_t + \beta_5 pill_t + u_t$$

Forma funcional e dummy

```
modelo = lm(gfr ~ pe + pe_1 + pe_2 + ww2 + pill, data = fertil3)
round(summary(modelo)$coef,5)
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  95.87050    3.28196 29.21138  0.00000
## pe            0.07267    0.12553  0.57891  0.56468
## pe_1         -0.00578    0.15566 -0.03713  0.97050
## pe_2          0.03383    0.12626  0.26792  0.78962
## ww2         -22.12650   10.73197 -2.06174  0.04330
## pill        -31.30499    3.98156 -7.86250  0.00000
```

Devemos confiar em essas estatísticas T?

Forma funcional e dummy

```
library(car)
vif(modelo)
```

```
##        pe      pe_1      pe_2       ww2      pill
## 22.238847 35.407682 24.092658  2.625982  1.174410
```

Suspeitamos de multicolinearidade então não podemos confiar nas estatísticas T, mas podemos fazer um teste F.

Forma funcional e dummy

$$H_0 : \beta_{pe_t} = 0, \beta_{pe_{t-1}} = 0, \beta_{pe_{t-2}} = 0 \quad vs \quad H_1 : H_0 \text{ não é verdade}$$

```
summary(modelo)$fstatistic
```

```
##      value     numdf     dendf
## 12.72844   5.00000  64.00000
```

```
qf(0.95,5,63)
```

```
## [1] 2.360684
```

Rejeitamos H_0, ou seja, pelo menos um dos β's é $\neq 0$. Agora vamos testar $H_0 : \beta_{pe_{t-1}} = 0, \beta_{pe_{t-2}} = 0 \quad vs \quad H_1 : H_0$ não é verdade

Forma funcional e dummy

```
fertil3 %>% select(pe, pe_1, pe_2) %>% glimpse()
```

```
## Rows: 72
## Columns: 3
## $ pe   <dbl> 0.00, 0.00, 0.00, 0.00, 19.27, 23.94, 20.07, 15.
## $ pe_1 <dbl> NA, 0.00, 0.00, 0.00, 0.00, 19.27, 23.94, 20.07,
## $ pe_2 <dbl> NA, NA, 0.00, 0.00, 0.00, 0.00, 19.27, 23.94, 20
```

- Existem NA's
- **Para fazer um teste F precisamos ter o mesmo número de observações**

```
modeloi = lm(gfr ~ pe + pe_1 + pe_2 + ww2 + pill, data = fertil3)
modelor = lm(gfr ~ pe + ww2 + pill, data = fertil3)
anova(modelor,modeloi) # dará erro
```

Forma funcional e dummy

```
# Removendo NA's
fertil = fertil3 %>% select(gfr, pe, pe_1, pe_2, ww2, pill) %>%
         na.omit() %>% glimpse()
```

```
## Rows: 70
## Columns: 6
## $ gfr   <dbl> 125.0, 123.4, 121.0, 119.8, 111.2, 117.9, 119.8,
## $ pe    <dbl> 0.00, 0.00, 19.27, 23.94, 20.07, 15.33, 34.32, 36
## $ pe_1  <dbl> 0.00, 0.00, 0.00, 19.27, 23.94, 20.07, 15.33, 34
## $ pe_2  <dbl> 0.00, 0.00, 0.00, 0.00, 19.27, 23.94, 20.07, 15.3
## $ ww2   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ pill  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
```

Forma funcional e dummy

```
modeloi = lm(gfr ~ pe + pe_1 + pe_2 + ww2 + pill, data = fertil)
modelor = lm(gfr ~ pe + ww2 + pill, data = fertil)
anova(modelor,modeloi)
```

```
## Analysis of Variance Table
##
## Model 1: gfr ~ pe + ww2 + pill
## Model 2: gfr ~ pe + pe_1 + pe_2 + ww2 + pill
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     66 13054
## 2     64 13033  2    21.761 0.0534  0.948
```

```
# Não rejeitamos H0
```

Tendêndia e Sazonalidade

Tendêndia e Sazonalidade

Duas caracteristicas presentes nas séries temporais

Tendêndia e Sazonalidade

Duas caracteristicas presentes nas séries temporais

Tendência

Componente a longo prazo e pode ser modelada por uma função polinomial ou logaritmica.

Tendêndia e Sazonalidade

Duas caracteristicas presentes nas séries temporais

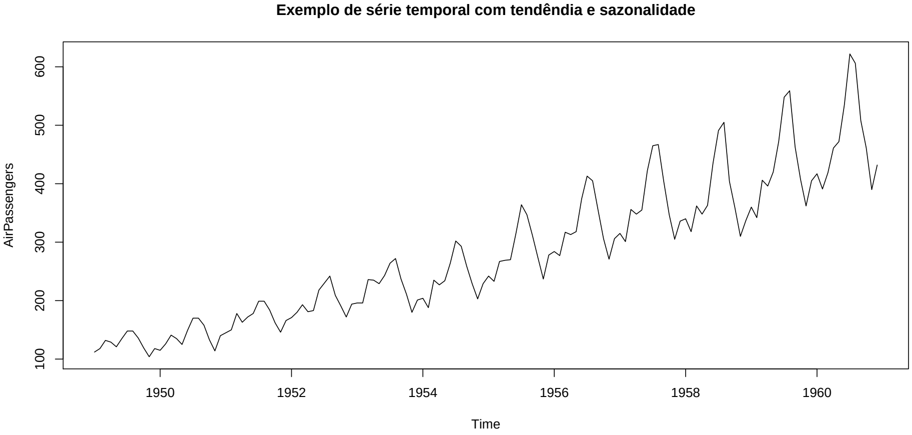
Tendência

Componente a longo prazo e pode ser modelada por uma função polinomial ou logaritmica.

Sazonalidade

Oscilações que se produzem e repetem em curtos periodos de tempo.

Tendêndia e Sazonalidade



Tendêndia

- Muitas séries temporais possuem uma tendência ao longo do tempo

Tendência

- Muitas séries temporais possuem uma tendência ao longo do tempo
- Ignorar o fato que duas (ou mais) séries podem aprensentar tendência na mesma direção (ou em direções opostas) pode nos levar a conclussões erradas no modelo de regressão

Tendência

- Muitas séries temporais possuem uma tendência ao longo do tempo
- Ignorar o fato que duas (ou mais) séries podem aprensentar tendência na mesma direção (ou em direções opostas) pode nos levar a conclussões erradas no modelo de regressão
- Muitas vezes podemos encontrar relações significativas e R^2 proximos de 1 indicando um *aparente* bom modelo, mas na verdade esses resultados são apenas obtidos por causa que as series apresentam a mesma tendência (regressão espuria)

Tendência

- Muitas séries temporais possuem uma tendência ao longo do tempo
- Ignorar o fato que duas (ou mais) séries podem aprensentar tendência na mesma direção (ou em direções opostas) pode nos levar a conclussões erradas no modelo de regressão
- Muitas vezes podemos encontrar relações significativas e R^2 proximos de 1 indicando um *aparente* bom modelo, mas na verdade esses resultados são apenas obtidos por causa que as series apresentam a mesma tendência (regressão espuria)
- Podemos *remover* a tendência incluindo na regressão uma componente de *tendência temporal*

Tendência

Que modelos capturam adequadamente o comportamento da tendência?

- Tendência Linear: $y_t = \alpha_0 + \alpha_1 t + e_t$
- Tendência Quadrática: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$
- Tendência Cúbica: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + e_t$
- Tendência Exponencial: $\log(y_t) = \alpha_0 + \alpha_1 t + e_t$ $(y_t > 0)$

Tendêndia

Que modelos capturam adequadamente o comportamento da tendência?

- Tendência Linear: $y_t = \alpha_0 + \alpha_1 t + e_t$
- Tendência Quadrática: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$
- Tendência Cúbica: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + e_t$
- Tendência Exponencial: $\log(y_t) = \alpha_0 + \alpha_1 t + e_t \; (y_t > 0)$

Como construir o modelo?

- Incluir variaveis com tendêndia temporal não violam as HST1–HST6

Tendêndia

Que modelos capturam adequadamente o comportamento da tendência?

- Tendência Linear: $y_t = \alpha_0 + \alpha_1 t + e_t$
- Tendência Quadrática: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$
- Tendência Cúbica: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + e_t$
- Tendência Exponencial: $\log(y_t) = \alpha_0 + \alpha_1 t + e_t \ (y_t > 0)$

Como construir o modelo?

- Incluir variaveis com tendêndia temporal não violam as HST1–HST6
- Contudo, devemos ter cuidado para não cair no caso de regressão espúria

Tendêndia

Que modelos capturam adequadamente o comportamento da tendência?

- Tendência Linear: $y_t = \alpha_0 + \alpha_1 t + e_t$
- Tendência Quadrática: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$
- Tendência Cúbica: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + e_t$
- Tendência Exponencial: $\log(y_t) = \alpha_0 + \alpha_1 t + e_t \ (y_t > 0)$

Como construir o modelo?

- Incluir variaveis com tendêndia temporal não violam as HST1–HST6
- Contudo, devemos ter cuidado para não cair no caso de regressão espúria
- Incluir uma tendência temporal (linear, quadrática, exponencial) elimina o problema.

Tendêndia: Exemplo

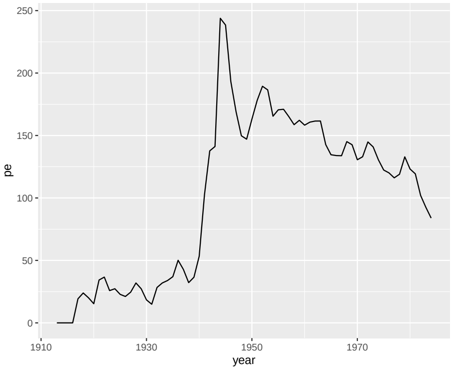
```
modelo = lm(gfr ~ pe + ww2 + pill, data = fertil3)
round(summary(modelo)$coef,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.6818     3.2081 30.7599   0.0000
## pe            0.0825     0.0296  2.7842   0.0069
## ww2         -24.2384     7.4583 -3.2499   0.0018
## pill        -31.5940     4.0811 -7.7416   0.0000
```

```
summary(modelo)$adj.r.squared
```

```
## [1] 0.4501837
```

Tendêndia: Exemplo



Tendência: Exemplo

```
modelo_t = lm(gfr ~ pe + ww2 + pill + t, data = fertil3)
round(summary(modelo_t)$coef,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 111.7694     3.3578 33.2868   0.0000
## pe            0.2789     0.0400  6.9685   0.0000
## ww2         -35.5923     6.2974 -5.6519   0.0000
## pill          0.9974     6.2616  0.1593   0.8739
## t            -1.1499     0.1879 -6.1195   0.0000
```

```
summary(modelo_t)$adj.r.squared
```

```
## [1] 0.6420465
```

Tendêndia: Exemplo

```
modelo_t2 = lm(gfr ~ pe + ww2 + pill + t + I(t^2), data = fertil3
round(summary(modelo_t2)$coef,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 124.0919     4.3607 28.4566   0.0000
## pe            0.3478     0.0403  8.6392   0.0000
## ww2         -35.8803     5.7079 -6.2860   0.0000
## pill        -10.1197     6.3361 -1.5972   0.1150
## t            -2.5314     0.3894 -6.5011   0.0000
## I(t^2)        0.0196     0.0050  3.9454   0.0002
```

```
summary(modelo_t2)$adj.r.squared
```

```
## [1] 0.7059697
```

Tendêndia : Exemplo

```
modelo_t3 = lm(gfr ~ pe + ww2 + pill + t + I(t^2) + I(t^3), data
round(summary(modelo_t3)$coef,4)
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 142.7955     4.3377  32.9193    0e+00
## pe            0.1619     0.0413   3.9198    2e-04
## ww2         -19.0467     5.0420  -3.7776    3e-04
## pill        -25.0097     5.3456  -4.6785    0e+00
## t            -5.6122     0.5428 -10.3401    0e+00
## I(t^2)        0.1554     0.0203   7.6528    0e+00
## I(t^3)       -0.0013     0.0002  -6.8088    0e+00
```

```
summary(modelo_t3)$adj.r.squared
```

```
## [1] 0.8257366
```

Tendêndia : Exemplo

```
modelo_t4 = lm(gfr ~ pe + ww2 + pill + t + I(t^2) + I(t^3) + I(t
round(summary(modelo_t4)$coef,4)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143.6371     5.6929 25.2311   0.0000
## pe            0.1602     0.0422  3.7940   0.0003
## ww2         -19.2373     5.1459 -3.7384   0.0004
## pill        -24.2481     6.3167 -3.8387   0.0003
## t            -5.8456     1.1502 -5.0822   0.0000
## I(t^2)        0.1709     0.0701  2.4359   0.0176
## I(t^3)       -0.0016     0.0015 -1.0944   0.2779
## I(t^4)        0.0000     0.0000  0.2307   0.8183
```

```
# summary(modelo_t4)$adj.r.squared:    0.8231607
```

Tendêndia

O que estamos fazendo quando incluimos tendêncial temporal?

No fundo, o que estamos fazendo é remover a tendência das séries originais.

```r
modelo_org = lm(gfr~pe + pill +  t, data = fertil3)
# regressao de c/variavel sobre a tendência
modelogrf = lm(gfr~t, data = fertil3)
modelope = lm(pe~t, data = fertil3)
modelopill = lm(pill~t, data = fertil3)
# Calculamos os residuais
grf2 = residuals(modelogrf)
pe2  = residuals(modelope)
pill2  = residuals(modelopill)
```

Tendêndia

podemos pensar em *grf2*, *pe2* e *pill2* como variaveis que tiveram a tendência removida. Entao se fizermos

$$grf2 = \beta_0 + \beta_2 pe2 + \beta_2 pill2 + u \tag{1}$$

Os $\hat{\beta}_1$ e $\hat{\beta}_2$ da Eq (1), seram os mesmos $\hat{\beta}_1$ e $\hat{\beta}_2$ do modelo

$$grf = \beta_0 + \beta_2 pe2 + \beta_2 pill2 + \beta_3 t + u$$

Tendência

```
modeloaux = lm(grf2~pe2 + pill2,data=fertil3)
round(coef(modelo_org),4)
```

```
## (Intercept)          pe        pill           t
##    109.2111      0.1774     -2.7431     -0.8370
```

```
round(coef(modeloaux),4)
```

```
## (Intercept)         pe2       pill2
##      0.0000      0.1774     -2.7431
```

Tendêndia

Como saber se incluir uma tendência temporal no modelo?

- Se alguma variavel (dependênte ou independênte) apresentar tendência, então incluir t (ou t^2ou t^3) é uma boa ideia.

Tendência

Como saber se incluir uma tendência temporal no modelo?

- Se alguma variavel (dependênte ou independênte) apresentar tendência, então incluir t (ou t^2ou t^3) é uma boa ideia.
- Se ao incluir a tendência temporal, o termo de tendência for estatísticamente significativo e os resultados mudarem de forma susbtancial, os resultados sem uma tendência devem ser tratados com desconfiança.

Tendêndia

Como saber se incluir uma tendência temporal no modelo?

- Se alguma variavel (dependênte ou independênte) apresentar tendência, então incluir t (ou t^2ou t^3) é uma boa ideia.
- Se ao incluir a tendência temporal, o termo de tendência for estatísticamente significativo e os resultados mudarem de forma susbtancial, os resultados sem uma tendência devem ser tratados com desconfiança.
- Se o tendência temporal não for incluida, então nao ocorrera remoção da tendència e poderemos encontrar relaciones espúrias.

Tendência: R^2's

- Os R^2's podem ser enganosos quando trabalhamos com séries temporais.
- Precisamos ajustar os R^2's para poder ser utilizados
- Uma forma simples consite em fazer a regresão de y com a tendencia temporal e utilizar o residual $u = y*$ na regressão sobre a tendência temporal e as variaveis explicativas

Tendência: R^2's

```
summary(modelo_t)$adj.r.squared
```

```
## [1] 0.6420465
```

```
summary(modelo_t2)$adj.r.squared
```

```
## [1] 0.7059697
```

```
summary(modelo_t3)$adj.r.squared
```

```
## [1] 0.8257366
```

```
summary(modelo_t4)$adj.r.squared
```

```
## [1] 0.8231607
```

Tendência: R^2's

```
gfr_t1 = residuals(lm(gfr~t, data = fertil3))
gfr_t2 = residuals(lm(gfr~t + I(t^2), data = fertil3))
gfr_t3 = residuals(lm(gfr~t + I(t^2) + I(t^3), data = fertil3))
gfr_t4 = residuals(lm(gfr~t + I(t^2) + I(t^3) + I(t^4), data = fe

modelo_t1r = lm(gfr_t1 ~ pe + ww2 + pill + t, data = fertil3)
modelo_t2r = lm(gfr_t2 ~ pe + ww2 + pill + t + I(t^2), data = fer
modelo_t3r = lm(gfr_t3 ~ pe + ww2 + pill + t + I(t^2) + I(t^3), d
modelo_t4r = lm(gfr_t4 ~ pe + ww2 + pill + t + I(t^2) + I(t^3) +
```

Tendência: R^2's

```
summary(modelo_t1r)$adj.r.squared
```

```
## [1] 0.4960654
```

```
summary(modelo_t2r)$adj.r.squared
```

```
## [1] 0.5713144
```

```
summary(modelo_t3r)$adj.r.squared
```

```
## [1] 0.4789117
```

```
summary(modelo_t4r)$adj.r.squared
```

```
## [1] 0.3786115
```

Sazonalidade

- Número de passageiros de uma aerolinha
- Vendas de um determinado produto
- Temperatura

Algumas séries temporais exhibem um padrão que se repete ao longo do tempo. Esse padrão e chamado de sazonalidade

Como incluir essa informação no nosso modelo de regressão?

Variáveis dummy sazonais:

- Se tivermos dados mensais, incluir 11 variáveis dummy
- Se tivermos dados trimestrais, incluir 3 variáves dummy
- Se tivermos dados quatrimestrais, incluir 2 variáves dummy

Sazonalidade: Exemplo

Seja o modelo

$$\log(uclms) = \beta_0 + \beta_1 ez + \beta_2 t + u$$

em que:

- uclms: número de desempregados
- ez: dummy (1 no período que a cidade de Anderson tinha uma zona empresarial)

Sazonalidade: Exemplo

```r
t = seq(1, nrow(ezanders))
modelo = lm(log(uclms) ~ ez+ t, data = ezanders)
summary(modelo)$coef
```

```
##                 Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)  9.304914440 0.090212699 103.144175 1.452922e-106
## ez          -0.371839450 0.165306482  -2.249394  2.659455e-02
## t           -0.009336544 0.002661819  -3.507580  6.692054e-04
```

vamos incluir variaveis dummy para capturar o possivel efeito sazonal

Sazonalidade

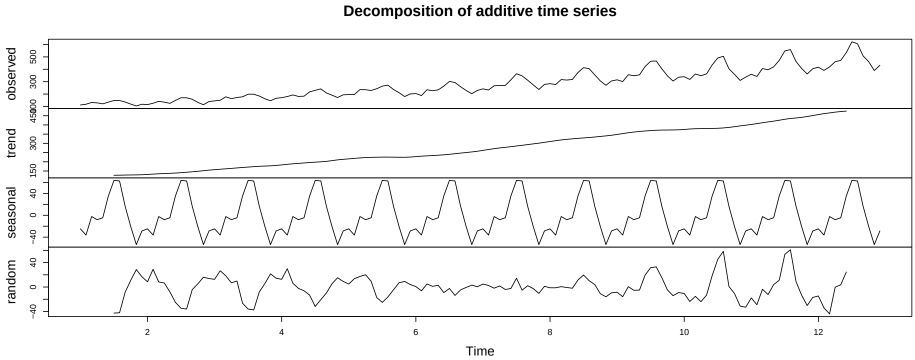
```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.2726     0.1413 65.6186   0.0000
## factor(month)AUG      0.0541     0.1768  0.3062   0.7602
## factor(month)DEC      0.0595     0.1824  0.3259   0.7452
## factor(month)FEB      0.3382     0.1766  1.9149   0.0586
## factor(month)JAN      0.3446     0.1767  1.9506   0.0541
## factor(month)JULY    -0.1339     0.1767 -0.7581   0.4503
## factor(month)JUNE    -0.1757     0.1766 -0.9947   0.3225
## factor(month)MAR      0.2920     0.1766  1.6539   0.1015
## factor(month)MAY     -0.1431     0.1766 -0.8104   0.4198
## factor(month)NOV     -0.2694     0.1773 -1.5194   0.1321
## factor(month)OCT     -0.3813     0.1771 -2.1529   0.0339
## factor(month)SEPT    -0.3541     0.1769 -2.0014   0.0483
## ez                   -0.5080     0.1457 -3.4876   0.0007
```

Conclusões

- Séries temporis podem ser incluidas num contexto de modelos de regresão
- Incluir tendència temporal (Linear, Quadrática, Cúbica, Exponencial, etc) nos ajudará na modelagem e evitara que encontremos relações espúrias
- Para incorporar o efeito sazonal, podemos incluir variáveis dummy
- Muitas vezes, detectar tendência e sazonalidade não é tarefa facil.

Conclusões

```
dados = ts(AirPassengers, frequency = 12) #dados mensais
plot(decompose(dados))
```



Conclusões

Temos trabalhado sob as hipótestes

- **HST1**: $y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$, $t = 1, \ldots, n$
- **HST2**: Não existe colinearidade perfeita entre as variáveis independêntes e nenhuma das variáveis independentes é constante.
- **HST3**: $E(u_t|X) = 0$, $t = 1, \ldots, n$
- **HST4**: Homocedasticidade, $V(u_t|X) = V(u_t) = \sigma^2$, $t = 1, \ldots, n$
- **HST5**: $Corr(u_t, u_s|X) = 0$, $\forall t \neq s$
- **HST6**: Os erros u_t são independentes de X e $u_t \sim N(0, \sigma^2)$

Na prática, algumas dessas hipóteses podem ser irreais (em especial HST3 e HST5) e os resultados obtidos devem ser visto com cuidado.

Leituras recomendadas

Leituras recomendadas

- Wooldridge, Jeffrey M. *Introdução à Econometria: Uma abordagem moderna*. (2016). Cengage Learning. – **Cap 10**