Aula 4 – Modelos de Regressão

Luiz Filipe de Almeida Xavier

Departamento de Economia da UFPE
(Decon/UFPE)

6 de junho de 2025

Definição de Regressão.

O que é uma regressão?

Definição

A **regressão** é uma análise quantitativa que busca entender como uma variável de interesse (**Y**, ou **dependente**) se comporta em função de outra(s) variável(is) (**X**, ou **dependente**). Ela é usada quando se quer **quantificar a relação** entre variáveis, prever valores futuros ou identificar padrões nos dados.

Ex. Estrutura

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Tipos de Regressão.

Regressão Linear Simples

Definição

Modela a relação entre uma variável dependente e uma única variável independente, assumindo que essa relação é linear (reta).

Estimação

As regressões simples podem ser estimadas pelo método dos mínimos quadrados ordinários (MQO ou OLS - Ordinary Least Squares).

MQO:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i \tag{2}$$

$$\varepsilon_i^2 = (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{3}$$

$$Min_{\{\beta_0, \beta_1\}}\, \varepsilon_i^2 = (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{4}$$

O Método consiste em encontrar os valores ótimos dos parâmetros (β_0 e β_1) que minimizam o erro ao quadrado.

Representação gráfica

Após estimar os parâmetros, pode-se representar graficamente a regressão em um plano cartesiano onde o eixo horizontal representa a variável independente (X) e o eixo vertical representa a variável dependente (Y).

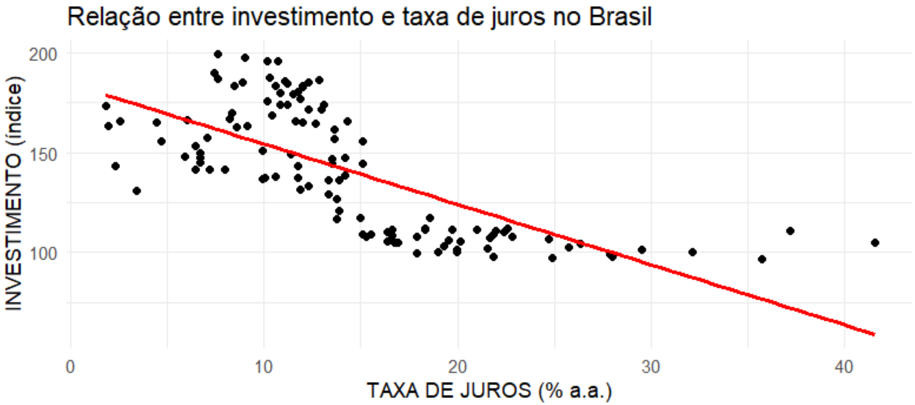
Definição

Esse tipo de representação gráfica é conhecida como **cross-section**, ou corte transversal.

Parâmetros

β_0 : Intercepto da reta (onde corta o eixo vertical).
β_1 : Coeficiente angular da reta.

Exemplo da representação gráfica de uma regressão linear simples



Fonte: Autoria própria – dados do IPEA.

Regressão

$$\widehat{Investimento} = 184.37 - 3.01 \cdot JUROS$$

Interpretação dos parâmetros

A interpretação dos parâmetros estimados vai depender muito dos tipos de dados tratados, porém, em termos gerais, podemos dizer o seguinte:

β_0 :

É a parte da variável dependente que não é explicada pela variável independente.

β_1 :

É a sensibilidade da variável dependente para cada variação de X em uma unidade.

Em termos de cálculo, $\frac{dY}{dX} = \beta_1$

Regressão no R

lm

Estimar os parâmetros de uma regressão no R é muito simples, com uma função nativa da linguagem que é a lm:

```
lm(y ~ x, banco_de_dados)
```

Parâmetros

x: vetor de dados da variável independente
y: vetor de dados da variável dependente
banco_de_dados: data frame onde esses vetores de dados estão.

Do exemplo gráfico anterior – Regressão no R

Código:

```
lm(FBCF ~ selic, df)
```

Console:

```
Call:
lm(formula = FBCF ~ selic, data = df)

Coefficients:
(Intercept)        selic
     184.37        -3.01

>
```

Regressão Linear Multipla

Definição

É uma extensão da regressão linear simples para múltiplas variáveis independentes, assumindo que a relação de cada variável dependente com a independente é linear (reta).

Estimação

Também pode ser estimada por MQO

Representação gráfica

A representação gráfica de regressões múltiplas com mais de duas variáveis independentes torna-se inviável, pois a visualização máxima da dinâmica entre as variáveis é limitada a gráficos tridimensionais...

Representação simplificada

Porém, É possível analisar graficamente a dinâmica entre as variáveis por meio de uma abordagem **estática comparativa**, que consiste em manter algumas variáveis constantes enquanto se observa a variação de outras.

Modelo

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + \ \ldots + \beta_n \cdot X_{ni} + \varepsilon_i$$

Interpretação dos parâmetros

Assim como na regressão simples, a interpretação dos parâmetros estimados vai depender muito dos tipos de dados tratados, porém, em termos gerais, podemos dizer o seguinte:

β_n , $n > 0$:

É a sensibilidade da variável dependente para cada variação de X_n em uma unidade, considerando todas as variáveis independentes constantes.

Em termos de cálculo, $\frac{\delta Y}{\delta X} = \beta_n$

Por que utilizar uma regressão múltipla em vez de uma simples?

Pode ocorrer o problema de **viés de variável omitida** em uma regressão simples, ou com poucas variáveis.

Viés de variável omitida:

O viés de variável omitida ocorre quando uma variável relevante que afeta tanto a variável dependente (Y), quanto pelo menos uma das variáveis independentes (X_n) é excluída do modelo de regressão. Isso pode distorcer as estimativas dos coeficientes das variáveis incluídas.

Condição para ter viés ao omitir a variável:

$$\widehat{\beta_n} \neq 0 \qquad em \;\; X_n = \widehat{\delta_0} + \widehat{\delta_1} \cdot X_k, \;\; \delta_1 \neq 0$$

Regressão no R

lm

Para fazer uma regressão múltipla no R basta somar os vetores de variáveis independentes:

$\texttt{lm(y \~{} x1 + x2 + ... + x}_{\texttt{n}}\texttt{, banco\_de\_dados)}$

Exemplo de uma regressão linear Múltipla no R

Regressão Simples

$$\widehat{Renda} = \widehat{\beta_0} + \widehat{\beta_1} \cdot Educação$$

```
lm(renda_mensal ~ anos_estudo, df_limpo)
```

```
Call:
lm(formula = renda_mensal ~ anos_estudo, data = df_limpo)

Coefficients:
(Intercept)  anos_estudo
     -636.6        342.6
```

Exemplo de uma regressão linear Múltipla no R

Regressão Multipla

$$\widehat{Renda} = \widehat{\beta_0} + \widehat{\beta_1} \cdot Educação + \widehat{\beta_2} \cdot Sexo + \widehat{\beta_3} \cdot Cor$$

```
lm(renda_mensal ~ anos_estudo + sexo + cor, df_limpo)
```

```
Call:
lm(formula = renda_mensal ~ anos_estudo + sexo + cor, data = df_limpo)

Coefficients:
(Intercept)  anos_estudo         sexo          cor
     -364.9        369.4      -1181.0       -571.5
```

Verificando Viés de Variável Omitida no R

1ª condição

$$\widehat{\beta_2} \neq 0 \quad e \quad \widehat{\beta_3} \neq 0$$

2ª condição

$$Educação = \widehat{\delta_0} + \widehat{\delta_1} \cdot Sexo, \;\; \delta_1 \neq 0$$

$$Educação = \widehat{\tau_0} + \widehat{\tau_1} \cdot Cor, \;\; \tau_1 \neq 0$$

```
Call:
lm(formula = anos_estudo ~ sexo,
data = df_limpo)

Coefficients:
(Intercept)         sexo
     10.205        1.726
```

```
Call:
lm(formula = anos_estudo ~ cor,
data = df_limpo)

Coefficients:
(Intercept)          cor
    11.0213      -0.7437
```

Regressão não linear

Definição

É um tipo de regressão em que a relação entre a variável dependente (Y) e as variáveis independentes (X) é representada por uma combinação não linear dos parâmetros (não é uma reta).

Tipos

Existem vários tipos de regressão não linear, com os parâmetros tendo interpretações diferentes.

Log-linear:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (5)$$

Log-log:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i \quad (6)$$

linear-log:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i \quad (7)$$

Polinomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_k (X_i)^k + \varepsilon_i \quad (8)$$

Regressão no R

lm

Para fazer uma regressão que envolva ln no R, basta utilizar a função log nas variáveis:

```
lnx <- log(x)
lny <- log(y)
lm(lny ~ lnx, banco_de_dados)
```

Modelo log-linear

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = e^{(\beta_0+\beta_1 X_i+\varepsilon_i)}$$

β_n , $n > 0$:

Representa a variação % de Y quando X varia em uma unidade.

Modelo linear-log

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i$$

β_n , $n > 0$:

Representa a variação de Y quando X varia em 1 %.

Interpretação de alguns dos modelos

Modelo log-log

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i \qquad Y_i = e^{(\beta_0 + \beta_1 \ln(X_i) + \varepsilon_i)}$$

$$\frac{\delta Y_i}{\delta X_i} = e^{(\beta_0 + \beta_1 \ln(X_i) + \varepsilon_i)} \frac{\beta_1}{X_i}$$

$$\frac{\delta Y_i}{\delta X_i} = Y_i \frac{\beta_1}{X_i}$$

$$\beta_1 = \frac{\delta Y_i}{\delta X_i} \frac{X_i}{Y_i}$$

β_n , n > 0:

Representa a elasticidade de Y em relação a X, isto é, a mudança % de Y para cada variação % de X

Por que utilizar um tipo de regressão a outro (qual é melhor)?

O tipo de regressão utilizado vai depender (1) da interpretação que o pesquisador quer fazer, e (2) da melhor capacidade explicativa do modelo.

R^2

O R^2 é uma estimativa que mede justamente a capacidade explicativa de um modelo, em que, quanto mais próximo de 1, melhor é a predição do modelo.

$$0 \leq R^2 \leq 1$$

R^2 no R

R^2

Para encontrar o valor do R^2 no R, você precisa salvar os resultados da regressão em uma variável:

```
modelo <- lm(y ~ x, banco_de_dados)
```

E depois utilizar a função summary para mostrar um resumo com as estatísticas da regressão:

```
Summary(modelo)
```

Exemplo R^2

```
reg_simples <- lm(renda_mensal ~ anos_estudo, df_simples)
summary(reg_simples)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2695.7 -1063.4  -463.4  1126.9  6304.3

Coefficients:
            Estimate Std. Error t value
(Intercept)  -433.36     663.06  -0.654
anos_estudo   258.07      52.85   4.883
            Pr(>|t|)
(Intercept)    0.515
anos_estudo 4.06e-06 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
  0.1 ' ' 1

Residual standard error: 1546 on 98 degrees of freedom
Multiple R-squared:  0.1957,    Adjusted R-squared:  0.1875
F-statistic: 23.84 on 1 and 98 DF,  p-value: 4.064e-06
```

Figura 2: Linear Simples

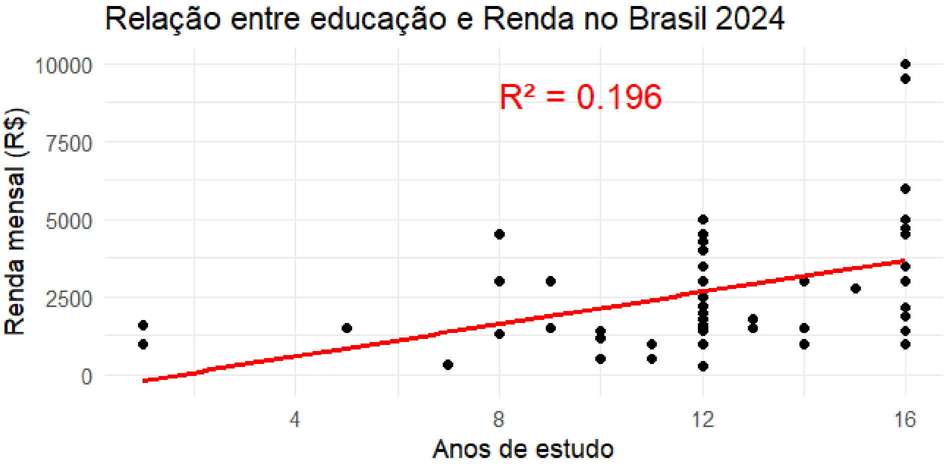


Figura 3: Linear log

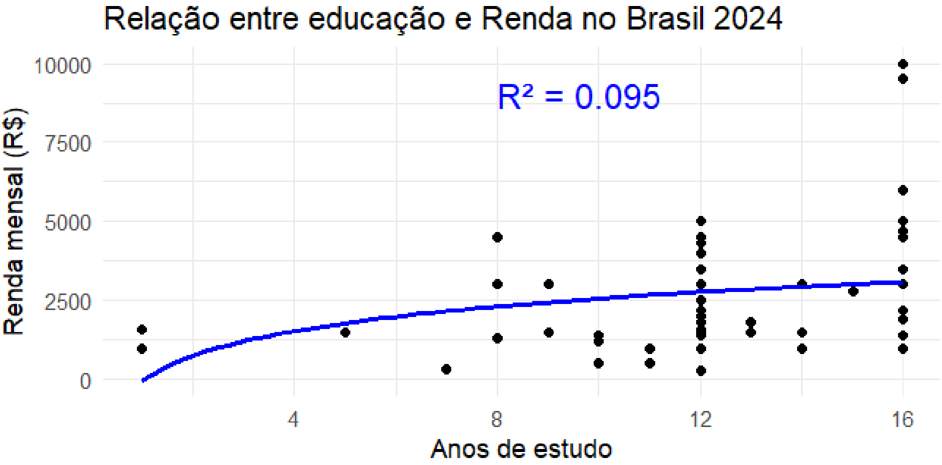
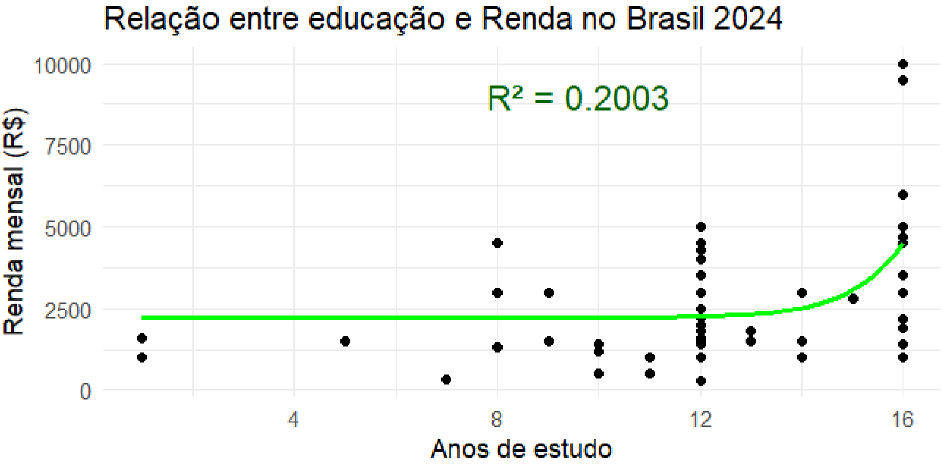


Figura 4: log linear



Outras estatísticas importantes (1)

Erro padrão

É uma medida da precisão da estimativa de um coeficiente em um modelo de regressão. Ele indica quanto a estimativa do coeficiente pode variar de uma amostra para outra.

$$Erro\ Padrão = \sqrt{Var(\widehat{\beta}_i)}$$

t–valor

Mede quantas vezes o coeficiente estimado é maior que seu erro padrão.

$$t = \frac{\widehat{\beta}_i}{\sqrt{Var(\widehat{\beta}_i)}}$$

Outras estatísticas importantes (2)

P-valor

É uma probabilidade utilizada para verificar se há ou não evidência suficiente nos dados para afirmar que ele é diferente de zero, geralmente com 95% de confiança.

Se o P-valor for < 0.05, há evidências de significância.
Se o P-valor for > 0.05, não há evidência de significância.

Estatísticas no R

Summary

Todas essas estatísticas são geradas após utilizar a função summary na regressão salva:

```
modelo <- lm(y ~ x, banco_de_dados)
summary(modelo)
```

Exemplo R²

```
reg_simples <- lm(renda_mensal ~ anos_estudo, df_simples)
summary(reg_simples)
```

Erro padrão

T-valor

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2695.7 -1063.4  -463.4  1126.9  6304.3 

Coefficients:
            Estimate Std. Error t value
(Intercept)  -433.36     663.06  -0.654
anos_estudo   258.07      52.85   4.883
            Pr(>|t|)    
(Intercept)    0.515    
anos_estudo 4.06e-06 ***
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 1546 on 98 degrees of freedom
Multiple R-squared:  0.1957,	Adjusted R-squared:  0.1875 
F-statistic: 23.84 on 1 and 98 DF,  p-value: 4.064e-06
```

P-valor

Outras estatísticas importantes (3)

Intervalo de confiança

É um intervalo de valores plausíveis para um parâmetro estimado a partir de uma amostra, em função do erro padrão, com um determinado nível de confiança.

$$Intervalo\text{: } \widehat{\beta_i} \pm f(Erro\ Padrão)$$

Função:

Se nesse intervalo houver o valor zero, o parâmetro estimado não é significante. Geralmente se utilizam os seguintes níveis de confiança: 90%, 95% e 99%

Estatísticas no R

confint

Para encontrar o intervalo de confiança no R, basta utilizar a função confint, inserindo a regressão e o nível de confiança:

```
modelo <- lm(y ~ x, banco_de_dados)
confint(modelo, lvel = 0.95)
```

*Homoscedasticidade

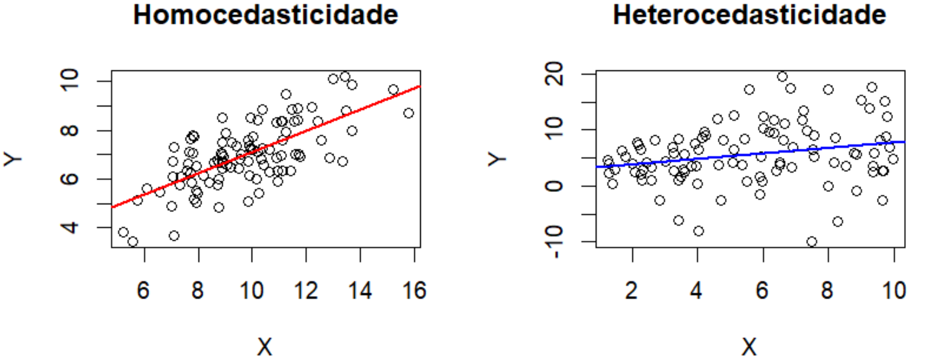
Definição

É uma suposição fundamental dos modelos de regressão linear que significa que os erros têm variância constante em todos os níveis das variáveis independentes.

Heteroscedasticidade

Ocorre quando o erro tem variância dependente de X.

Exemplo gráfico homoscedasticidade e heteroscedasticidade



Por que se preocupar com Heteroscedasticidade?

A presença de heterocedasticidade nas amostras viola uma das premissas fundamentais do método MQO: os resíduos devem ter variância constante.

Demais estatísticas:

Quando há Heteroscedasticidade na amostra, as estatísticas de p-valor, t-valor e intervalo de confiança são viesadas (podem ser maiores ou menores dependendo do nível de X, por exemplo)

Verificando heteroscedasticidade no R

bptest

Para verificar se há heteroscedasticidade na amostra, pode-se utilizar o teste Breusch-Pagan, chamado no R de bptest. É uma função disponível em um pacote de testes de regressão:

```
install.packages("lmtest")
library(lmtest)

modelo <- lm(y ~ x, banco_de_dados)
confint(modelo, lvel = 0.95)
bptest(modelo)
```

Se p-valor < 0.05, há heterocedasticidade

Teste de Heteroscedasticidade – Gráficos anteriores

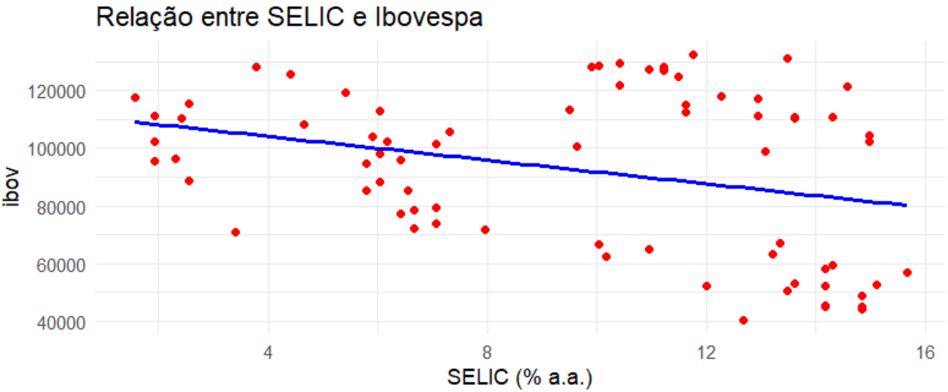
```
bptest(modelo1)
```

```
data:  modelo1
BP = 6.0034, df = 1,
p-value = 0.01428
```

```
bptest(modelo2)
```

```
data:  modelo2
BP = 0.35242, df =
1, p-value = 0.5527
```

Exemplo com dados reais




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 112271.4     7664.2  14.649   <2e-16
selic        -2048.5      734.7  -2.788   0.0068
```

```
> confint(modelo, lvel = 0.95)
                2.5 %      97.5 %
(Intercept) 96989.412 127553.4205
selic       -3513.493   -583.4877
```

```
        studentized
Breusch-Pagan test

data:  modelo
BP = 26.347, df = 1,
p-value = 2.852e-07
```

Regressão Robusta

Definição

É um tipo de regressão projetada para lidar com situações em que os dados contêm outliers (valores extremos) ou violam suposições clássicas da regressão linear, como a da homoscedasticidade.

Regressão Robusta no R

rlm

A regressão robusta no R também é muito simples. Basta utilizar a função rlm do pacote mass:

```
library(MASS)
Modelo_robusto <- rlm(y ~ x, banco_de_dados)
```

Comparando Exemplo de regressão linear Múltipla da renda R

```
lm(renda_mensal ~ anos_estudo + sexo + cor, df_limpo)

Call:
lm(formula = renda_mensal ~ anos_estudo + sexo + cor, data = df_limpo)

Coefficients:
(Intercept)  anos_estudo         sexo          cor
     -364.9        369.4      -1181.0       -571.5
```

```
rlm(renda_mensal ~ anos_estudo + sexo + cor, df_limpo)

Call:
rlm(formula = renda_mensal ~ anos_estudo + sexo + cor, data = df_limpo)
Converged in 7 iterations

Coefficients:
(Intercept) anos_estudo        sexo         cor
   434.5055    197.6729   -591.0421   -224.1540
```

Obrigado!