Parte 1.1: Regressão Linear Simples e Mínimos Quadrados Ordinários

Econometria I - IBMEC

Marcus L. Nascimento

26 de agosto de 2025

1. Regressão Linear Simples

2. Mínimos Quadrados Ordinários (MQO)

3. Coeficiente de determinação (R^2)

4. Formas funcionais

Regressão Linear Simples

Introdução

- A análise de regressão estuda a relação entre o que denominamos **variável dependente** ou **variável resposta** e um conjunto de variáveis que denominamos **variáveis independentes**, **explicativas** ou **regressores**.
  - **Variável dependente**: Variável que está sendo estudada, comumente denotada por Y;
  - **Variáveis independentes**: Variáveis utilizadas para explicar a variável dependente, comumente denotadas por $X_1, X_2, \ldots, X_p$.
- A relação entre as variáveis é representada por um modelo matemático, que associa a variável dependente com as variáveis independentes.
- Este modelo é designado por **Regressão Linear Simples** quando há uma variável explicativa $(p = 1)$ e uma variável resposta.
- Analogamente, quando o modelo envolve duas ou mais variáveis explicativas passamos a denominá-lo **Regressão Linear Múltipla**.

Introdução

Através de modelos de Regressão Linear Simples, estudamos a relação linear entre duas variáveis quantitativas.

EXEMPLOS:

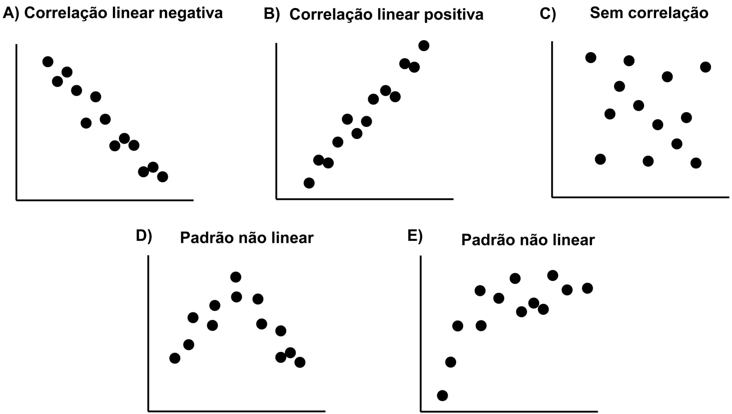
- Renda semanal e despesas de consumo;
- Variação dos salários e taxa de desemprego;
- Vendas de produtos e investimentos em publicidade.

CORRELAÇÃO VS. REGRESSÃO:

- Regressão explicita a forma como variáveis estão relacionadas;
- Correlação quantifica a força ou grau com que variáveis estão relacionadas.

Diagrama de dispersão

Diagramas de dispersão permitem decidir empiricamente se há uma relação linear entre uma variável dependente (Y) e uma variável explicativa (X).



Definição

Um modelo de Regressão Linear simples é descrito da seguinte forma:

$$Y = E(Y|X = x) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon,$$

onde:

- A função $E(Y|X = x)$ é chamada regressão de Y em X;
- Y: variável dependente;
- X: variável explicativa ou independente medida sem erro (não aleatória);
- β_0: coeficiente de regressão que representa o intercepto (parâmetro desconhecido do modelo);
- β_1: coeficiente de regressão que representa a inclinação (parâmetro desconhecido do modelo);
- ε: erro aleatório que contém a variação de Y que não pode ser explicada linearmente pelo comportamento da variável X.

Suposições do modelo

Dadas n observações do par X e Y, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, temos:

$$Y_i = E(Y_i|X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Suposições sobre os erros (ε_i):

- Independência: ε_i's são variáveis aleatórias independentes;
- $\text{Var}(\varepsilon_i|x_i) = \text{Var}(\varepsilon_i) = \sigma^2$;
- $E(\varepsilon_i|x_i) = E(\varepsilon_i) = 0$.

A partir das suposições, temos:

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i \text{ e } \text{Var}(Y_i|X = x_i) = \sigma^2.$$

Observação

Vale ressaltar que o termo regressão linear significa regressão linear nos **parâmetros**.

PERGUNTAS:

- $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ é considerado um modelo de regressão linear simples?

Observação

Vale ressaltar que o termo regressão linear significa regressão linear nos **parâmetros**.

PERGUNTAS:

- $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ é considerado um modelo de regressão linear simples?
- E $\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$?

Exercícios (Wooldridge, J. M. (2003))

Suponha que Y denote o número de filhos e X seja a escolaridade medida em anos de uma mulher. Um modelo simples relacionando fertilidade com escolaridade é

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

onde ε é o erro não observado.

Exercícios (Wooldridge, J. M. (2003))

Suponha que Y denote o número de filhos e X seja a escolaridade medida em anos de uma mulher. Um modelo simples relacionando fertilidade com escolaridade é

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

onde ε é o erro não observado.

(i) Quais fatores estão contidos em ε? É provável que eles estejam correlacionados com a escolaridade?

Exercícios (Wooldridge, J. M. (2003))

Suponha que Y denote o número de filhos e X seja a escolaridade medida em anos de uma mulher. Um modelo simples relacionando fertilidade com escolaridade é

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

onde ε é o erro não observado.

(i) Quais fatores estão contidos em ε? É provável que eles estejam correlacionados com a escolaridade?

(ii) A relação entre a escolaridade e os fatores contidos em ε impacta alguma das suposições do modelo de regressão linear simples? Explique.

Exercícios (Wooldridge, J. M. (2003))

Em um modelo de regressão linear simples, $y = \beta_0 + \beta_1 x + \varepsilon$, suponha que $E(\varepsilon) \neq 0$. Sendo $E(\varepsilon) = \alpha_0$, mostre que o modelo pode ser sempre reescrito com a mesma inclinação (β_1), porém com novo intercepto e novo erro, onde o novo erro possui valor esperado igual a zero.

Mínimos Quadrados Ordinários (MQO)

Introdução

Dado um modelo de regressão linear simples, como estimamos os parâmetros β_0 e β_1?

Dado um modelo de regressão linear simples, como estimamos os parâmetros β_0 e β_1?

Mínimos Quadrados Ordinários (MQO): procedimento bastante utilizado em Econometria para obtenção de estimadores.

- Quanto menor for o erro quadrático total , $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, melhor será a estimativa;
- Desejamos **minimizar** o erro quadrático total;
- Minimizar o erro quadrático total significará encontrar valores de β_0 e β_1 que minimizem a função

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Estimadores de Mínimos Quadrados Ordinários

O mínimo de $S(\beta_0, \beta_1)$ é obtido através do cálculo de sua derivada com respeito a β_0 e β_1, igualando o resultado a zero.

$$\begin{aligned} \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} &= 0; \qquad (1) \\ \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} &= 0. \qquad (2) \end{aligned}$$

A partir de (1) e (2), obtemos:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}; \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \end{aligned}$$

onde $\bar{x} = \sum_{i=1}^{n} x_i / n$ e $\bar{y} = \sum_{i=1}^{n} y_i / n$.

Definindo

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}; \\
S_{xx} &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2; \\
S_{yy} &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2,
\end{aligned}
$$

reescrevemos $\hat{\beta}_0$ e $\hat{\beta}_1$ da seguinte forma:

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}; \\
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}.
\end{aligned}
$$

Interpretação

Seja $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ a **reta de regressão estimada**, temos:

- $x = 0$: $\hat{y} = \hat{\beta}_0$.
  $\hat{\beta}_0$ é o ponto onde a reta corta o eixo das ordenadas e pode ser interpretável ou não.

- $x \to x + 1$: $\Delta\hat{y} = [\hat{\beta}_0 + \hat{\beta}_1(x + 1)] - [\hat{\beta}_0 + \hat{\beta}_1 x] = \hat{\beta}_1$.
  $\hat{\beta}_1$ é o coeficiente angular e representa o quanto varia a média de Y para um aumento de uma unidade da variável X.

Suposições

A1. (Linearidade) No modelo populacional, a variável resposta y está relacionada a variável independente x e ao erro ε da seguinte forma:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

onde β_0 e β_1 são os parâmetros populacionais para o intercepto e a inclinação.

A2. (Amostragem aleatória) Pode-se utilizar uma amostra aleatória de tamanho n, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, do modelo populacional.

A3. (Média Condicional Zero) $E(\varepsilon | x) = 0$.

Suposições

A4. (Variação amostral no regressor) Na amostra, as variáveis independentes x_i, $i = 1, 2, \ldots, n$, não são todas iguais a mesma constante.

A5. (Homocedasticidade) $\text{Var}(\varepsilon|x) = \sigma^2$.

OBSERVAÇÃO:

- $E(\varepsilon|x) = 0$ implica que todos os fatores contidos no erro devem ser não correlacionados com o regressor.

Propriedades

- Sob as suposições A1-A4, os estimadores de Mínimos Quadrados Ordinários $\hat{\beta}_0$ e $\hat{\beta}_1$ são não viesados:

$$E(\hat{\beta}_0) = \beta_0 \text{ e } E(\hat{\beta}_1) = \beta_1.$$

- Sob as suposições A1-A5, as variâncias dos estimadores de Mínimos Quadrados Ordinários dependem da dispersão da variável independente e da variância do erro:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \text{ e } \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

  - $\text{Var}(\hat{\beta}_1) \underset{n \to \infty}{\longrightarrow} 0$ para valores de x_i distribuídos ao redor da média $\bar{x}$;
  - $\text{Var}(\hat{\beta}_0) \underset{n \to \infty}{\longrightarrow} 0$ assumindo que os valores de x_i são apropriadamente selecionados (não clusterizados próximos a média).

Propriedades

Melhores estimadores lineares não-viesados (BLUE):

- Os estimadores de MQO para os parâmetros β_0 e β_1 são os melhores dentre todos os estimadores da classe dos lineares não-viesados;
- Além de serem não-viesados, apresentam a menor variância dentre os demais estimadores não-viesados, gerando estimadores com menor erro quadrático médio dentre os lineares.

Estimador para a variância do erro (σ^2)

Seja $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ o **resíduo** da regressão linear. Para obtermos um estimador não enviesado de σ^2, analisamos a dispersão em torno da reta de regressão estimada:

$$\text{SSR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ (Soma de quadrados dos resíduos)} .$$

Sob as suposições A1-A5, $E(\sum_{i=1}^{n} e_i^2) = (n-2)\sigma^2$, logo um estimador não viesado de σ^2 é

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}.$$

Cálculo das variâncias dos estimadores de MQO

Com base em uma amostra, é possível encontrar estimativas para as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$ substituindo σ^2 por $\hat{\sigma}^2$ (estimador não viesado para σ^2) nas expressões para $\text{Var}(\hat{\beta}_0)$ e $\text{Var}(\hat{\beta}_1)$:

$$\widehat{\text{Var}}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \text{ e } \widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

para os quais

$$E(\widehat{\text{Var}}(\hat{\beta}_0)) = \text{Var}(\hat{\beta}_0) \text{ e } E(\widehat{\text{Var}}(\hat{\beta}_1)) = \text{Var}(\hat{\beta}_1).$$

Coeficiente de determinação (R^2)

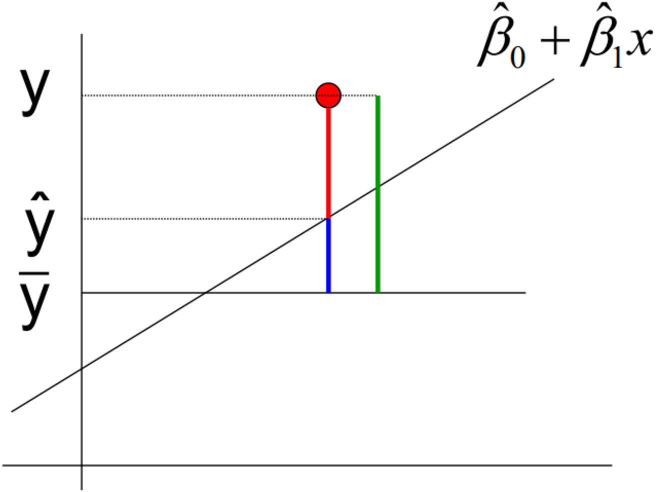
Definição

Uma vez obtida a **reta de regressão estimada**, faz-se necessário construir uma medida que indique, mesmo que de modo imperfeito, a qualidade do ajuste do modelo de regressão.

Para tal, definiremos as seguintes quantidades:

- $y - \bar{y}$: erro ao se prever y pela média geral;
- $y - \hat{y}$: erro ao se prever y pelo valor estimado para $E(Y|X)$ (**resíduo**);
- $\hat{y} - \bar{y}$: "ganho" ao se prever y pelo valor estimado para $E(Y|X)$ em comparação ao se prever y pela média geral y.

Definição



Definição

A partir das quantidades definidas anteriormente, escrevemos suas respectivas somas de quadrados:

- Soma de quadrados total (SST): $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$;
- Soma de quadrados dos resíduos (SSR): $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$;
- Soma de quadrados devido à explicação (SSE): $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

Definição

$\text{RESULTADO: SST} = \text{SSE} + \text{SSR},$

onde:

- SSE: parcela da variabilidade de y que é explicada pelos regressores do modelo;
- SSR: parcela da variabilidade de y que **não** é explicada pelos regressores do modelo.

Por fim, o Coeficiente de Determinação (R^2) é expresso da seguinte forma:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

Interpretação

- Como R^2 é uma proporção, sempre será um número entre 0 e 1;
- $R^2 = 0$: indica que o modelo não explica nada da variação da variável resposta ao redor da média. A média da variável dependente prediz a mesma tão bem quanto o modelo de regressão;
- $R^2 = 1$: indica que o modelo explica toda a variação da variável resposta no entorno da média;
- $100 \times R^2$ pode ser interpretado como a proporção da variabilidade total de y que é explicada pelo regressor do modelo adotado

Exercícios (Wooldridge, J. M. (2003))

Em um conjunto de dados com registros de nascimentos nos Estados Unidos, duas variáveis de interesse são o peso do bebê ao nascer (bwght) e o número médio de cigarros que a mãe fumou por dia durante a gestação (cigs).
A regressão linear simples a seguir foi estimada utilizando $n = 1388$ nascimentos:

$$\widehat{\text{bwght}} = \hat{\beta}_0 + \hat{\beta}_1 \text{cigs},$$

onde $\hat{\beta}_0 = 110,77$ e $\hat{\beta}_1 = -0.514$

(i) Qual o valor predito para o peso ao nascer quando cigs $= 0$? Qual valor predito quando cigs $= 20$? Comente sobre a diferença.

(ii) O modelo de regressão linear simples necessariamente captura uma relação causal entre o peso do bebê ao nascer e o hábito de fumar da mãe? Explique.

Exercícios (Wooldridge, J. M. (2003))

Considerando uma função de consumo linear

$$\widehat{\text{cons}} = \hat{\beta}_0 + \hat{\beta}_1 \text{inc},$$

a propensão marginal a consumir (estimada) é dada pela inclinação $\hat{\beta}_1$, enquanto a propensão média a consumir é dada por $\widehat{\text{cons}}/\text{inc} = \hat{\beta}_0/\text{inc} + \hat{\beta}_1$.

Utilizando dados sobre a renda anual e o consumo de 100 famílias, a seguinte equação é obtida:

$$\widehat{\text{cons}} = \hat{\beta}_0 + \hat{\beta}_1 \text{inc},$$

onde $\hat{\beta}_0 = -124,84$ e $\hat{\beta}_1 = 0.853$

(i) Interprete o intercepto na equação anterior e comente sobre seu sinal e magnitude.

(ii) Qual é o consumo predito para uma família com renda de \$30.000?

Formas funcionais

Introdução

- Em muitas aplicações econômicas, a relação entre as variáveis dependente e independente não é adequadamente descrita através de um modelo linear.

- Há formas de considerar não-linearidades em modelos de regressão linear simples através da apropriada definição das variáveis dependente e independente.

- Em análise de regressão linear, estudamos modelos lineares nos parâmetros e vimos casos em que os modelos podem ou não ser lineares nas variáveis.

- As variáveis podem se tornar lineares através de transformações apropriadas: a relação não-linear pode ser "linearizável" por transformações.

Introdução

- Em alguns casos, será possível fazer interpretações dos parâmetros em termos do problema de interesse ao se considerar a não-linearidade.
- **Efeito Marginal**: Mede o efeito da variável X na variável Y.

$$\frac{dY}{dX}.$$

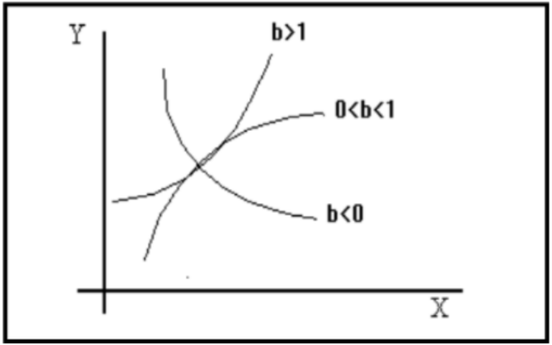
- **Elasticidade**: Mede a variação percentual de Y correspondente a uma dada variação percentual em X.

$$\frac{\%\Delta E(Y|X)}{\%\Delta X} = \frac{dY}{dX} \times \frac{X}{Y},$$

onde $\%\Delta X = 100.\frac{x - x_0}{x_0} = 100.\frac{\Delta x}{x_0}$.

Forma funcional: $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$.

Figura: Para análise do gráfico, considere que $\beta_1 = b$.



Modelo Log-Log

Interpretação associada a β_1:

$$\frac{\%\Delta E(Y|X)}{\%\Delta X} = \beta_1 \text{ (elasticidade).}$$

OBSERVAÇÕES:

- Modelo de elasticidade constante.
- O Modelo Log-Log decorre do fato de que o logaritmo aparece em ambos os membros da equação.
- Para utilizarmos esse modelo, todos os valores de Y e X devem ser positivos.

EXEMPLO:

Podemos estimar um modelo de elasticidade constante relacionando os salários de diretores de empresas e as vendas relacionadas às companhias:

$$\ln(\text{salário}) = \beta_0 + \beta_1 \ln(\text{vendas}) + \varepsilon,$$

onde as vendas representam o total anual medido em milhões de dólares.
Suponha que o seguinte resultado é obtido ao estimarmos a equação anterior via MQO:

$$\ln(\text{salário}) = 4.822 + 0.257 \ln(\text{vendas}) + \varepsilon.$$

Qual interpretação é dada para $\hat{\beta}_1 = 0.257$?

Modelo Log-Nível

Forma funcional: $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$.

Interpretação associada a β_1:

$$\frac{\%\Delta E(Y|X)}{\Delta X} = 100 \times \beta_1 \text{ (semi-elasticidade).}$$

EXEMPLO:

Considere o caso em que deseja-se estimar a relação entre salário e educação. Utilizando o salário na escala logarítmica, temos a seguinte equação:

$$\ln(\text{salário}) = \beta_0 + \beta_1 \text{educ} + \varepsilon,$$

onde a educação é medida em anos.
Suponha que ao estimarmos a equação anterior via MQO, obtemos $\hat{\beta}_0 = 0.584$ e $\hat{\beta}_1 = 0.083$. Qual interpretação é dada para $\hat{\beta}_1 = 0.083$?

Modelo Log-Nível

A partir da equação $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$, é possível observar que

$$y = e^{\beta_0 + \beta_1 x + \varepsilon}.$$

Logo,

$$\frac{dy}{dx} = \beta_1 e^{\beta_0 + \beta_1 x + \varepsilon} = \beta_1 y. \tag{3}$$

Reescrevendo (3), temos que

$$\beta_1 = \frac{dy/y}{dx},$$

ou seja, β_1 representa uma taxa de crescimento. Dessa forma, $\beta_1 = 0,04$ indica um crescimento de 4% em Y diante de uma variação de uma unidade em X.

Modelo Nível-Log

Forma funcional: $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$.

Interpretação associada a β_1:

$$\frac{\Delta E(Y|X)}{\%\Delta X} = \frac{\beta_1}{100}.$$

Neste caso, um aumento de 1% na variável independente X resulta em um aumento de $\beta_1/100$ na variável dependente Y.

Resumo

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$