UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

SME0500 - Cálculo Numérico

Prof. Dr. Leandro Franco de Souza

RELATÓRIO - CÁLCULO DO MÉTODO DOS MÍNIMOS QUADRADOS USANDO REGRESSÃO LINEAR MÚLTIPLA

Vitor Antonio de Almeida Lacerda Nº USP 12544761

São Carlos - SP

2024

1. INTRODUÇÃO

A expectativa de vida é um indicador demográfico crucial que reflete as condições de saúde, bem-estar e desenvolvimento socioeconômico de uma população. Compreender os fatores que influenciam a expectativa de vida pode fornecer insights valiosos para a formulação de políticas públicas e estratégias de desenvolvimento sustentável. Neste estudo, foram analisados diversos países para determinar a relação entre a expectativa de vida e variáveis socioeconômicas, tais como Renda per Capita, Produto Interno Bruto (PIB) e População.

Para realizar essa análise, utilizou-se a técnica de regressão linear múltipla, um método estatístico que permite modelar a relação entre uma variável dependente e várias variáveis independentes. A regressão linear múltipla é fundamentada no método dos mínimos quadrados, e é capaz de fornecer estimativas precisas e interpretáveis.

O objetivo deste estudo foi desenvolver um modelo que possa prever a expectativa de vida com base nas variáveis mencionadas. Para isso, os dados de 49 países foram divididos em conjuntos de treinamento e teste. O modelo foi treinado com os dados de 44 países e testado nos dados dos 5 países restantes para avaliar sua precisão e capacidade de generalização. Os resultados obtidos foram expressos por meio de coeficientes de regressão, erro quadrático médio (MSE) e coeficiente de determinação (R^2), proporcionando uma análise detalhada da relação entre as variáveis independentes e a expectativa de vida.

2. METODOLOGIA

A regressão linear múltipla é uma técnica estatística que permite modelar a relação entre uma variável dependente e múltiplas variáveis independentes. Este método é uma extensão da regressão linear simples e é utilizado quando se deseja entender como várias variáveis preditoras afetam uma variável resposta. Na regressão linear múltipla, o objetivo é ajustar uma equação linear da forma:

$$\gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$. Onde:

- γ é a variável dependente (expectativa de vida).
- $x_1, x_2, \ldots, x_n$ são as variáveis independentes (Renda per Capita, PIB e População).
- β_0 é o intercepto do modelo.
- $\beta_1, \beta_2, \ldots, \beta_n$ são os coeficientes de regressão que representam a magnitude e a direção do efeito de cada variável independente sobre a variável dependente.
- ϵ é o termo de erro que captura a variação não explicada pelo modelo.

A estimativa dos coeficientes de regressão é realizada utilizando o método dos mínimos quadrados, que busca minimizar a soma dos quadrados dos erros (diferenças entre os valores observados e os valores previstos pelo modelo). Os métodos são descritos nos tópicos abaixo.

2.1 Coleta e Organização dos Dados.

Os dados foram coletados para 49 países, incluindo as variáveis Renda per Capita, PIB, População e Expectativa de Vida. Os dados foram fornecidos, e podem ser encontrados nesta tabela abaixo:

País	Renda per Cápita (US$)	PIB (US$ bilhões)	População (milhões)	Expectativa de Vida (anos)
Estados Unidos	65,118	21433	331.0	79.11
China	10,262	14342	1,402.0	76.91
Japão	40,847	5082	126.3	84.67
Alemanha	46,259	3861	83.2	81.33
Índia	2,338	2869	1,366.0	69.66
Brasil	8,717	2056	212.6	75.88
Reino Unido	41,059	2828	67.9	81.27
França	40,493	2778	65.3	82.57
Itália	34,32	2002	60.4	83.24
Canadá	46,233	1736	37.6	82.30
Rússia	11,585	1699	144.1	72.58
Austrália	53,825	1397	25.5	82.89

Coreia do Sul	31,846	1629	51.7	82.59
Espanha	29,715	1393	46.7	83.51
México	10,118	1269	128.9	75.13
Indonésia	4,135	1042	273.5	71.72
Holanda	52,331	912	17.4	82.17
Arábia Saudita	23,139	779	34.8	75.13
Turquia	8,958	761	84.3	77.93
Suíça	82,839	746	8.6	83.93
Argentina	10,549	449	45.4	76.28
Suécia	52,477	541	10.4	82.78
Polônia	15,45	596	38.3	78.47
Bélgica	46,482	533	11.5	81.69
Noruega	81,694	434	5.4	82.94
Irã	6	463	83.0	76.02
Áustria	51,641	449	8.9	81.74
Tailândia	7,808	514	69.8	77.15
Nigéria	2,149	432	206.1	54.33
Emirados Árabes	43,47	421	9.9	77.97
Egito	3,019	363	102.3	72.54
Filipinas	3,485	362	108.1	71.16
Singapura	65,233	372	5.7	83.45
Vietnã	3,537	340	97.3	75.28
Malásia	10,24	336	32.4	75.04
Chile	15,01	282	19.1	80.10
Paquistão	1,543	284	220.9	67.27
Colômbia	6,498	336	50.9	77.29
Peru	6,978	229	32.9	76.02
Ucrânia	3,727	155	41.0	71.76
Bangladesh	1,998	324	164.7	72.32
Romênia	12,896	244	19.1	75.45
Hungria	17,846	180	9.6	76.75
Grécia	19,193	209	10.4	81.04
República Checa	23,496	251	10.7	79.31
Portugal	23,333	237	10.2	81.32
Iraque	4,96	224	40.2	70.27
Nova Zelândia	42,634	212	4.9	82.42
Cuba	9,099	100	11.3	79.18

Tabela 1 - Dados disponíveis para obtenção dos resultados

2.2 Divisão dos Dados.

Os dados foram divididos em conjuntos de treinamento (44 países) e teste (5 países). O conjunto de treinamento foi utilizado para ajustar o modelo, enquanto o conjunto de teste foi utilizado para avaliar a precisão do modelo.

2.3 Preparação das Variáveis.

As variáveis independentes (Renda per Capita, PIB e População) foram organizadas em uma matriz X com uma coluna adicional de 1s para o intercepto.

A variável dependente (Expectativa de Vida) foi organizada em um vetor y

2.4. Cálculo da Matriz Transposta.

A matriz transposta X^T foi calculada.

2.5. Cálculo da Matriz Produto.

A matriz produto $X^T X$ foi calculada.

2.6. Cálculo da Inversa da Matriz Produto.

A inversa da matriz $(X^T X)^{-1}$ foi calculada.

2.7. Estimativa dos Coeficientes de Regressão (beta).

Os coeficientes de regressão foram estimados utilizando a fórmula $\beta = (X^T X)^{-1} X^T y$

2.8 Previsões.

As previsões para os dados de treinamento foram calculadas utilizando $y_{train} = X_{train}\beta$

As previsões para os dados de teste foram calculadas utilizando $y_{test} = X_{test}\beta$

2.9 Avaliação do Modelo.

O erro quadrático médio (MSE) foi calculado para os dados de treinamento e teste para avaliar a precisão do modelo.

O coeficiente de determinação (R^2) foi calculado para os dados de treinamento e teste para indicar a proporção da variância explicada pelo modelo.

2.10 Implementação

A implementação foi realizada em Python, utilizando bibliotecas como NumPy para manipulação de matrizes e cálculos matemáticos. Para isso, como possui 49 países, utilizou-se os 44 primeiros países para a obtenção da fórmula e o erro foi verificado aplicando a fórmula nos 5 países que ficaram de fora do cálculo da fórmula. Veja o código utilizado abaixo.

```
import numpy as np



# Dados de entrada

data = {

     'País': ['Estados Unidos', 'China', 'Japão', 'Alemanha', 'Índia',
'Brasil', 'Reino Unido', 'França', 'Itália', 'Canadá',

                  'Rússia', 'Austrália', 'Coreia do Sul', 'Espanha',
'México', 'Indonésia', 'Holanda', 'Arábia Saudita', 'Turquia',

                 'Suíça', 'Argentina', 'Suécia', 'Polônia', 'Bélgica',
'Noruega', 'Irã', 'Áustria', 'Tailândia', 'Nigéria',

                 'Emirados Árabes', 'Egito', 'Filipinas', 'Singapura',
'Vietnã', 'Malásia', 'Chile', 'Paquistão', 'Colômbia',

                'Peru', 'Ucrânia', 'Bangladesh', 'Romênia', 'Hungria',
'Grécia', 'República Checa', 'Portugal', 'Iraque',

             'Nova Zelândia', 'Cuba'],

    'Renda per Capita (US$)': [65118, 10262, 40847, 46259, 2338, 8717,
41059, 40493, 34320, 46233, 11585, 53825, 31846, 29715,

                                10118, 4135, 52331, 23139, 8958, 82839,
10549, 52477, 15450, 46482, 81694, 6463, 51641, 7808,

                                  2149, 43470, 3019, 3485, 65233, 3537,
10240, 15010, 1543, 6498, 6978, 3727, 1998, 12896, 17846,

                               19418, 37926, 23566, 4774, 41243, 8707],

     'PIB (US$ bilhões)': [21137, 14140, 5156, 3846, 2875, 2056, 2829,
2716, 2001, 1736, 1641, 1381, 1658, 1403, 1195, 1042,
```

```
                           907, 778, 760, 705, 449, 552, 586, 529, 403,
454, 458, 456, 448, 421, 332, 304, 364, 262, 315, 268,

                           231, 323, 232, 130, 249, 248, 153, 142, 245,
265, 224, 122, 103, 97],

    'População (milhões)': [331, 1441, 126, 83, 1380, 213, 68, 67, 60,
38, 146, 26, 51, 47, 128, 273, 17, 34, 84, 8, 45, 10,

                            38, 11, 5, 84, 9, 70, 206, 10, 100, 109, 6,
96, 32, 19, 229, 50, 32, 43, 165, 19, 10, 11, 10, 11,

                           10, 41, 5, 11],

    'Expectativa de Vida': [78.9, 76.7, 84.5, 81.2, 69.7, 75.9, 81.3,
82.5, 83.2, 82.3, 72.7, 82.9, 83.3, 83.0, 75.0, 71.8,

                             82.1, 75.3, 78.6, 83.4, 76.5, 82.3, 77.5,
81.6, 82.4, 76.2, 81.4, 78.6, 54.5, 77.8, 70.5, 71.2,

                             83.6, 75.3, 76.9, 80.2, 67.0, 76.7, 76.4,
72.3, 72.6, 75.4, 75.8, 81.4, 79.7, 81.1, 70.8, 82.3, 79.7]

}



# Usando os 44 primeiros países para treinar o modelo e os 5 últimos
para testar

X_train = np.array([

    [1, 65118, 21433, 331.0],

    [1, 10262, 14342, 1402.0],

    [1, 40847, 5082, 126.3],

    [1, 46259, 3861, 83.2],

    [1, 2338, 2869, 1366.0],

    [1, 8717, 2056, 212.6],

    [1, 41059, 2828, 67.9],

    [1, 40493, 2778, 65.3],

    [1, 3432, 2002, 60.4],

    [1, 46233, 1736, 37.6],
```

```
    [1, 11585, 1699, 144.1],
    [1, 53825, 1397, 25.5],
    [1, 31846, 1629, 51.7],
    [1, 29715, 1393, 46.7],
    [1, 10118, 1269, 128.9],
    [1, 4135, 1042, 273.5],
    [1, 52331, 912, 17.4],
    [1, 23139, 779, 34.8],
    [1, 8958, 761, 84.3],
    [1, 82839, 746, 8.6],
    [1, 10549, 449, 45.4],
    [1, 52477, 541, 10.4],
    [1, 1545, 596, 38.3],
    [1, 46482, 533, 11.5],
    [1, 81694, 434, 5.4],
    [1, 60, 463, 83.0],
    [1, 51641, 449, 8.9],
    [1, 7808, 514, 69.8],
    [1, 2149, 432, 206.1],
    [1, 4347, 421, 9.9],
    [1, 3019, 363, 102.3],
    [1, 3485, 362, 108.1],
    [1, 65233, 372, 5.7],
    [1, 3537, 340, 97.3],
    [1, 1024, 336, 32.4],
    [1, 1501, 282, 19.1],
    [1, 1543, 284, 220.9],
```

```
    [1, 6498, 336, 50.9],

    [1, 6978, 229, 32.9],

    [1, 3727, 155, 41.0],

    [1, 1998, 324, 164.7],

    [1, 12896, 244, 19.1],

    [1, 17846, 180, 9.6],

    [1, 19193, 209, 10.4]

])

y_train = np.array([79.11, 76.91, 84.67, 81.33, 69.66, 75.88, 81.27,
82.57, 83.24, 82.30, 72.58, 82.89, 82.59, 83.51,

                    75.13, 71.72, 82.17, 75.13, 77.93, 83.93, 76.28,
82.78, 78.47, 81.69, 82.94, 76.02, 81.74, 77.15,

                    54.33, 77.97, 72.54, 71.16, 83.45, 75.28, 75.04,
80.10, 67.27, 77.29, 76.02, 71.76, 72.32, 75.45,

                  76.75, 81.04])



X_test = np.array([

    [1, 23496, 251, 10.7],

    [1, 23333, 237, 10.2],

    [1, 496, 224, 40.2],

    [1, 42634, 212, 4.9],

    [1, 9099, 100, 11.3]

])

y_test = np.array([79.31, 81.32, 70.27, 82.42, 79.18])



# Cálculo da matriz transposta de X

X_train_T = X_train.T
```

```
# Cálculo da matriz produto X^TX

XTX = X_train_T @ X_train



# Cálculo da inversa de X^TX

XTX_inv = np.linalg.inv(XTX)



# Cálculo dos coeficientes de regressão (beta)

beta = XTX_inv @ X_train_T @ y_train



print(f"\nCoeficientes de regressão (beta):\nIntercepto:
{beta[0]:.8f}\nRenda per Capita: {

      beta[1]:.8f}\nPIB: {beta[2]:.8f}\nPopulação: {beta[3]:.8f}")



# Previsões para os dados de treinamento

y_train_pred = X_train @ beta



# Previsões para os dados de teste

y_test_pred = X_test @ beta



print("\nPrevisões para os dados de treinamento:\n", y_train_pred)

print("\nPrevisões para os dados de teste:\n", y_test_pred)



# Print Previsões vs Valores Reais para o conjunto de testes

print("\nPrevisões vs Valores Reais para o conjunto de testes:")

for i in range(len(y_test)):

              print(f"País:    {data['País'][44    +    i]},    Previsão:
{y_test_pred[i]:.2f}, Valor Real: {y_test[i]}")
```

```
# Cálculo do Erro Quadrático Médio (MSE) para os dados de treinamento
MSE_train = np.mean((y_train - y_train_pred)**2)

# Cálculo do Erro Quadrático Médio (MSE) para os dados de teste
MSE_test = np.mean((y_test - y_test_pred)**2)

# Cálculo do Coeficiente de Determinação (R²) para os dados de treinamento
R2_train = 1 - (np.sum((y_train - y_train_pred)**2) /
                np.sum((y_train - np.mean(y_train))**2))

# Cálculo do Coeficiente de Determinação (R²) para os dados de teste
R2_test = 1 - (np.sum((y_test - y_test_pred)**2) /
               np.sum((y_test - np.mean(y_test))**2))

print("\nErro Quadrático Médio (MSE) para os dados de treinamento:", MSE_train)
print("\nErro Quadrático Médio (MSE) para os dados de teste:", MSE_test)
print("\nCoeficiente de Determinação (R²) para os dados de treinamento:", R2_train)
print("\nCoeficiente de Determinação (R²) para os dados de teste:", R2_test)
```

3. RESULTADOS

Os resultados encontrados foram expressados por diversos coeficientes e métricas, que são essenciais para entender a eficácia do modelo de regressão

linear múltipla aplicado aos dados. A figura abaixo retrata o resultado obtido através do código.

```
Coeficientes de regressão (beta):
Intercepto: 74.66473685
Renda per Capita: 0.00013675
PIB: 0.00011849
População: -0.00416310

Previsões para os dados de treinamento:
 [84.73146508 71.93084657 80.32705773 81.10191461 69.63762801 75.21535667
 80.33209287 80.25959021 75.1198428  81.03640042 75.85043426 82.08483254
 78.99755911 78.69899016 75.6621453  74.21507152 81.8567762  77.77649021
 75.62899214 86.04579712 75.97154083 81.8619232  74.78719468 81.03656264
 85.86556748 74.38226681 81.7429412  75.50282373 74.15179294 75.26787195
 74.69472144 74.73418381 83.60588369 74.78364958 74.70970076 74.82390248
 73.98977004 75.38126846 75.50916691 75.02209395 74.29069833 76.37769829
 77.08659072 77.27090258]

Previsões para os dados de teste:
 [77.86307775 77.8412097  74.59175199 80.4997782  75.87385708]

Previsões vs Valores Reais para o conjunto de testes:
País: República Checa, Previsão: 77.86, Valor Real: 79.31
País: Portugal, Previsão: 77.84, Valor Real: 81.32
País: Iraque, Previsão: 74.59, Valor Real: 70.27
País: Nova Zelândia, Previsão: 80.50, Valor Real: 82.42
País: Cuba, Previsão: 75.87, Valor Real: 79.18

Erro Quadrático Médio (MSE) para os dados de treinamento: 17.58552743647747

Erro Quadrático Médio (MSE) para os dados de teste: 9.498187792834104

Coeficiente de Determinação (R²) para os dados de treinamento: 0.442932285070919

Coeficiente de Determinação (R²) para os dados de teste: 0.48474735907950195
```

Imagem 1 - Resultados do código

Agora, explicando cada um dos resultados obtidos:

- **Intercepto**: 74.6647. Este valor representa o ponto onde a linha de regressão intercepta o eixo Y
- **Renda per Capita**: 0.000137. Relação com a qualidade de vida diretamente proporcional.
- **PIB**: 0.000118. Relação com a qualidade de vida diretamente proporcional.

- **População**: -0.00416. Relação com a qualidade de vida inversamente proporcional.

A partir destes dados foi possível obter a equação solicitada:

$$Expectativa\ de\ Vida\ = 74.66473685\ +\ 0.00013675 \times Renda\ per\ Capita$$

$$+\ 0.00011849 \times PIB\ -\ 0.00416310 \times População$$

Além disso, foram obtidos os dados relacionados ao erro, conforme supracitado na metodologia:

- As previsões para os **Dados de Treinamento** feitas pelo modelo para os dados de treinamento são valores estimados de expectativa de vida com base nos valores das variáveis independentes desses países. As previsões variam de 69.67 a 85.87, demonstrando como o modelo ajusta os dados de entrada para estimar a variável de saída.
- As previsões para os **Dados de Teste** são utilizadas para avaliar a precisão do modelo em novos dados que não foram usados durante o treinamento. As previsões, também representadas pelo Gráfico 1, para os cinco países no conjunto de teste são:
    - República Checa: Previsão: 77.86, Valor Real: 79.31
    - Portugal: Previsão: 77.84, Valor Real: 81.32
    - Iraque: Previsão: 74.59, Valor Real: 70.27
    - Nova Zelândia: Previsão: 80.50, Valor Real: 82.42
    - Cuba: Previsão: 75.87, Valor Real: 79.18

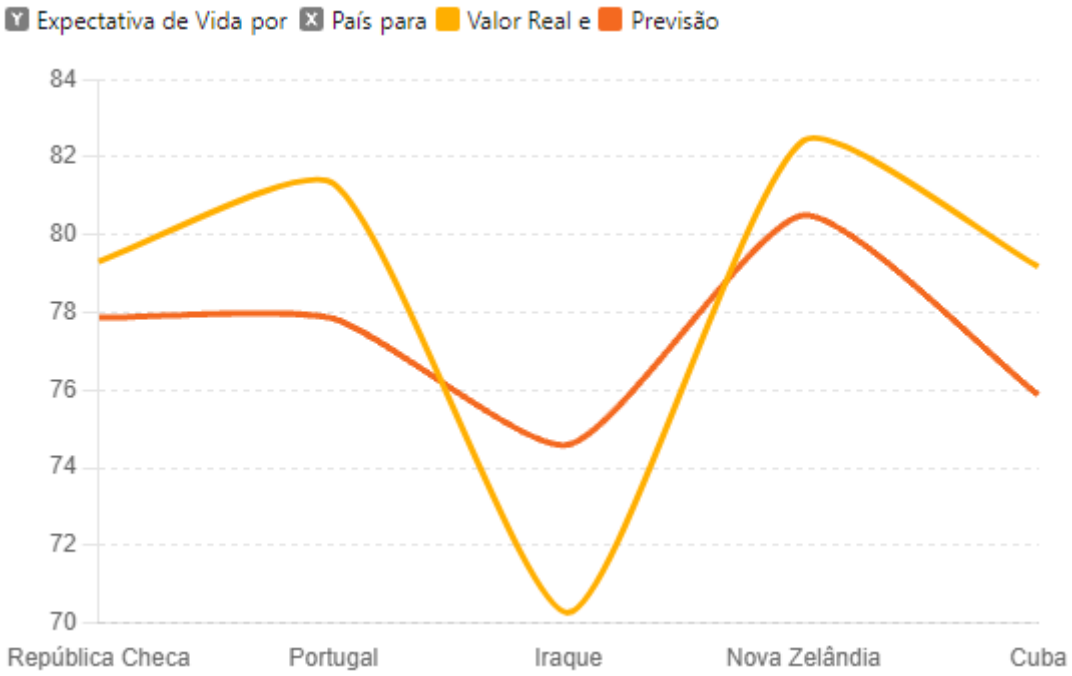

Gráfico 1 - Previsões Vs Valores Reais para o conjunto de testes

Além disso, há outros cálculos que foram possíveis por meio do código:

- **Erro Quadrático Médio MSE para os dados de treinamento**: 17.5855
    - Este valor indica o erro médio ao quadrado das previsões do modelo em relação aos dados de treinamento. Um MSE mais baixo indica um ajuste melhor do modelo aos dados de treinamento.
- **MSE para os dados de teste**: 9.4982
    - Este valor mede o erro médio ao quadrado das previsões do modelo em relação aos dados de teste. Um MSE mais baixo para os dados de teste sugere que o modelo generaliza bem para novos dados.
- **Coeficiente de Determinação (R^2)**: 0.4429
    - Este valor sugere que aproximadamente 44.29% da variância na expectativa de vida dos dados de treinamento é explicada pelas variáveis independentes no modelo.

- **R^2 para os dados de teste**: 0.4847
    - Este valor indica que cerca de 48.47% da variância na expectativa de vida dos dados de teste é explicada pelas variáveis independentes no modelo. Um R^2 mais próximo de 1 indica um melhor ajuste.

4. CONCLUSÃO

Em conclusão, os resultados mostraram que a fórmula de previsão desenvolvida, que inclui um intercepto e coeficientes para cada variável independente, conseguiu captar a relação entre as variáveis preditoras e a expectativa de vida. Embora o modelo tenha mostrado um erro quadrático médio (MSE) e coeficientes de determinação (R^2) que indicam variação na precisão, ele forneceu uma base sólida para entender como diferentes fatores econômicos e demográficos podem influenciar a expectativa de vida.

A análise gráfica das previsões versus os valores reais para os dados de teste revelou a capacidade do modelo de aproximar-se dos valores reais, embora com algumas discrepâncias. Estes resultados sugerem que, apesar de o modelo, implementado corretamente, ser um bom ponto de partida, há espaço para melhorias, como a inclusão de mais variáveis explicativas ou o uso de técnicas de modelagem mais avançadas para capturar melhor a complexidade dos dados.