1. Ao aplicar a regressão múltipla, construímos um modelo para explicar a variabilidade na variável dependente. Para isso, queremos incluir as influências simultâneas e individuais de diversas variáveis independentes. Por exemplo, suponha que quiséssemos desenvolver um modelo para predizer a margem de lucro anual para associações de poupança e empréstimo usando dados coletados ao longo de um período de anos. Uma especificação inicial do modelo indicou que a margem de lucro anual estava relacionada à receita líquida por dólar depositado e ao número de agências de poupança e empréstimo. Espera-se que a receita líquida anual aumente a margem de lucro anual, e o número de agências de poupança e empréstimo diminua a margem de lucro anual devido ao aumento da concorrência. Isso nos levaria a especificar um modelo de regressão populacional do tipo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{1}$$

onde

Y = margem de lucro anual,

X_1 = receita líquida anual por dólar depositado,

X_2 = número de agências de poupança e empréstimo naquele ano.

Na Tabela-1 temos dados sobre Poupança e Empréstimo que contêm 25 observações por ano dessas variáveis. Esses dados serão usados para desenvolver um modelo linear que prevê a margem de lucro anual em função da receita por dólar depositado e do número de agências (Spellman, 1978).

Ano	Receita por Dólar	Número de Escritórios	Margem de Lucro	Ano	Receita por Dólar	Número de Escritórios	Margem de Lucro
1	3,92	7298	0,75	14	3,78	6672	0,84
2	3,61	6855	0,71	15	3,82	6890	0,79
3	3,32	6636	0,66	16	3,97	7115	0,70
4	3,07	6506	0,61	17	4,07	7327	0,68
5	3,06	6450	0,70	18	4,25	7546	0,72
6	3,11	6402	0,72	19	4,41	7931	0,55
7	3,21	6368	0,77	20	4,49	8097	0,63
8	3,26	6340	0,74	21	4,70	8468	0,56
9	3,42	6349	0,90	22	4,58	8717	0,41
10	3,42	6352	0,82	23	4,69	8991	0,51
11	3,45	6361	0,75	24	4,71	9179	0,47
12	3,58	6369	0,77	25	4,78	9318	0,32
13	3,66	6546	0,78				

Tabela-1 Dados operacionais das associações de poupança e empréstimo.

Mas antes de podermos estimar o modelo, precisamos desenvolver e compreender o procedimento de regressão múltipla. Para começar, consideremos o modelo geral de regressão múltipla e observemos as semelhanças em relação ao modelo de regressão simples.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon. \tag{2}$$

Da mesma forma que no modelo de regressão linear simples algumas suposições devem

ser feitas sobre o termo de erro aleatório:

- O erro do modelo é uma variável aleatória com média zero;
- Os erros possuem distribuição Normal;
- A variância dos erros é constante;
- Os erros associados com quaisquer duas observações são independentes;
- O número de observações deve ser superior ao número de variáveis;
- Multicolinearidade entre as variáveis independentes (não pode existir combinação linear entre as variáveis independentes).

Nesse caso, tanto o método dos mínimos quadrados, quanto o método da máxima verossimilhança podem ser utilizados para obter os estimadores dos parâmetros do MRLM. A equação de regressão linear múltipla estimada, nesse exercício, é dada por

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \quad (3)$$

Na construção do MRLM, como mostrado em (2), consideramos um termo de erro. Este termo de erro reconhece que nenhuma relação postulada se manterá exatamente e que provavelmente haverá variáveis adicionais que também afetam o valor observado de *Y*. Assim, no cenário da aplicação, observamos o valor esperado da variável dependente, *Y* — conforme representado pelo plano na Figura 1 — mais um termo de erro aleatório, ε, que representa a parte de *Y* não incluída no valor esperado.

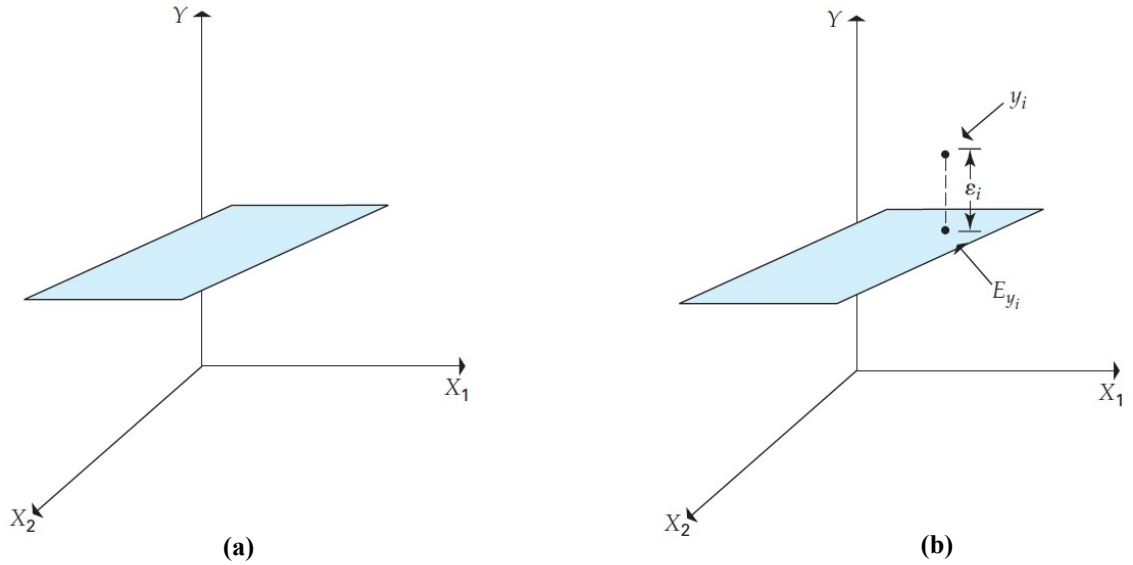

Figura-1 (a) O Plano é o Valor Esperado de *Y* como Função de X_1 e X_2; (b) Comparação dos Valores Observados e Esperados de *Y* como Função de Duas Variáveis Independentes.

As interpretações geométricas da regressão múltipla tornam-se cada vez mais complexas à medida que o número de variáveis independentes aumenta. No entanto, a analogia com a regressão linear simples é extremamente útil. Estimamos os coeficientes minimizando a soma dos desvios quadrados na dimensão *Y* em torno de uma função linear das variáveis independentes. Na regressão simples, a função é uma reta em um gráfico bidimensional. Com duas variáveis independentes, a função é um plano no espaço tridimensional. Além de duas variáveis independentes, temos vários hiperplanos complexos impossíveis de visualizar.

Primeira Parte

1.1 Construa um modelo de regressão para explicar Y, em termos das variáveis X_1 e X_2;
1.2 Analise a Tabela ANOVA da regressão gerada;
1.3 Plotar o plano ajustado ao conjunto de dados, atribuindo dinâmica 3D;
1.4 Verifique a normalidade dos resíduos;
1.5 Verifique a homocedasticidade dos resíduos;
1.6 Verifique a linearidade dos coeficientes;
1.7 Verifique a ausência de aucorrelação serial dos resíduos;
1.8 Verifique a multicolinearidade entre as variáveis independentes;
1.9 Interprete todos os testes realizados.

Segunda Parte

2.0 Construa uma RNA para explicar Y, em termos das variáveis X_1 e X_2;
2.1 Compare o modelo estimado pelo MRLM com o resultado da RNA.

Observação:

Os dados desse exercício foram retirados do Livro: *Statistics for Business and Economics*, Global Edition by Paul Newbold, William Carlson and Betty Thorne, 2013. A seguir são fornecidas **algumas** saídas da análise de regressão que estão nesse Livro. Apenas para você comparar com os resultados que vai encontrar quando usar o R ou o Python.

Referência citada no texto:

Spellman, L. J. 1978. "*Entry and Profitability in a Rate-free Savings and Loan Market*," Quarterly Review of Economics and Business 18 (2): 87–95.

Resultados da Regressão Múltipla

Regression Equation for Savings and Loan Association Profit (Minitab and Excel Output)

Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices

Predictor          Coef        SE Coef       T       P
Constant        1.56450       0.07940   19.70   0.000
X1 revenue      0.23720       0.05556    4.27   0.000
X2 offices   -0.00024908    0.00003205   -7.77   0.000

S = 0.0533022   R-Sq = 86.5%    R-Sq(adj) = 85.3%

Analysis of Variance
Source            DF        SS        MS       F       P
Regression         2   0.40151   0.20076   70.66   0.000
Residual Error    22   0.06250   0.00284
Total             24   0.46402
```

Regression coefficients b_0, b_1, b_2

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.930212915					
R Square	0.865296068					
Adjusted R Square	0.853050256					
Standard Error	0.053302217					
Observations	25					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	0.40151122	0.20075561	70.66057082	2.64962E-10	
Residual	22	0.06250478	0.002841126			
Total	24	0.464016				
	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.564496771	0.079395981	19.70498685	1.81733E-15	1.399839407	1.72915414
X1 revenue	0.237197475	0.055559366	4.269261695	0.000312567	0.121974278	0.35242067
X2 offices	–0.000249079	3.20485E-05	–7.771949195	9.50879E-08	–0.000315544	–0.00018261

Regression coefficients b_0, b_1, b_2

Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices

Predictor          Coef       SE Coef       T       P
Constant        1.56450       0.07940   19.70   0.000
X1 revenue      0.23720       0.05556    4.27   0.000
X2 offices   -0.00024908   0.00003205   -7.77   0.000

S = 0.0533022   R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source            DF       SS        MS       F       P
Regression         2  0.40151   0.20076   70.66   0.000
Residual Error    22  0.06250   0.00284
Total             24  0.46402

Source        DF      Seq SS
X1 revenus     1     0.22990
X2 offices     1     0.17161
```

Coefficients b_0, b_1, b_2

Standard error of the estimate s_e

Coefficient of determination R^2

$MSR = SSR/K$

Error variance s^2_e

$SSR = 0.40151$
$SSE = 0.06250$
$SST = 0.46402$

Number of independent X Variables, K