Análise descritiva bivariada
O caso de duas variáveis numéricas

Mauro Campos

22 de novembro de 2024

Resumo

Neste material de aula apresentamos métodos para uma análise descritiva bivariada quando as duas variáveis envolvidas na análise são numéricas. Formalmente, colocaremos foco em métodos para lidar com relações lineares entre as duas variáveis numéricas envolvidas. Tais métodos estão relacionados a dois conceitos fundamentais em estatística, a saber: correlação linear e regressão linear simples. Correlação linear estuda a existência de uma relação linear entre duas variáveis numéricas com o auxílio do gráfico de dispersão e do coeficiente de correlação linear de Pearson, desenvolvido para medir a intensidade da relação linear. Regressão linear simples busca descrever matematicamente a relação linear entre as duas variáveis numérica quando tal relação existe. O método de mínimos quadrados será usado para estimar os parâmetros do modelo de regressão linear simples. Por fim, o modelo linear estimado poderá ser usado para fazer predições para uma das variáveis envolvidas na análise para diferentes valores da outra variável.

Conteúdo

1 Introdução **2**
- 1.1 Contextualização do tema através de um exemplo prático . . . 2
- 1.2 Questões de interesse numa análise conjunta de duas variáveis numéricas . . . 3

2 Correlação linear **3**
- 2.1 Gráfico de dispersão . . . 4
- 2.2 Coeficiente de correlação de Pearson . . . 5
- 2.3 Fórmula alternativa para o coeficiente de correlação de Pearson . . . 6

3 Regressão linear simples **8**
- 3.1 Estimação de mínimos quadrados . . . 8
- 3.2 Modelo estimado e predições . . . 8

Referências **10**

A Exercício **11**

B Formulário **12**

C Script R para correlação linear e regressão linear simples **13**

1 Introdução

1.1 Contextualização do tema através de um exemplo prático

▷ Considere o conjunto de dados da Tabela 1. A variável X representa a renda bruta mensal familiar e a variável Y representa a percentagem dessa renda mensal que é gasta com saúde. Tabela 1 mostra amostras emparelhadas (ou pareadas) dessas duas variáveis para 10 famílias criteriosamente selecionadas a partir de uma população de famílias que vivem numa certa localidade.

Tabela 1: Percentagem da renda bruta mensal familiar gasta com saúde para uma amostra de 10 famílias

Id	Família	Renda bruta mensal (X)	% da renda bruta mensal gasta com saúde (Y)
1	A	12	7.2%
2	B	16	7.4%
3	C	18	7.0%
4	D	20	6.5%
5	E	28	6.6%
6	F	30	6.7%
7	G	40	6.0%
8	H	48	5.6%
9	I	50	6.0%
10	J	54	5.5%

Fonte: Morettin & Bussab (2010, Cap. 4)

▷ Para todas as famílias dessa localidade, estamos interessados nas seguintes questões de estudo:

1. As variáveis X e Y estão relacionadas?
2. Se X e Y estão relacionados, essa relação é do tipo que pode ser descrita por uma relação linear?
3. Se X e Y são variáveis linearmente relacionadas, então como usar os dados observados (ou seja, os dados da Tabela 1) para quantificar a intensidade dessa relação linear?
4. Se X e Y são variáveis linearmente relacionadas, então é razoável assumir que a média de Y, denotada aqui por μ_Y, pode ser descrita em termos de X através de uma função linear da forma

$$\mu_Y(X) = a + bX. \tag{1}$$

   Sendo assim, como usar os dados observados para estimar os parâmetros a e b que definem a reta relacionando X e μ_Y? Em outras palavras, como usar os dados para encontrar um modelo linear

$$\hat{y} = \hat{\mu}_Y(x) = \hat{a} + \hat{b}x \tag{2}$$

   que estima a relação populacional $\mu_Y(X) = a + bX$?
5. Por fim, como fazer uma predição pra Y para um valor de X (geralmente não observado previamente)? Por exemplo, o valor $x = 35$ não é igual a nenhum dos valores observados de X na Tabela 1. Então, como prever a percentagem esperada $\hat{y} = \hat{\mu}_y(x)$ da renda bruta mensal familiar gasta com saúde para uma família com renda mensal de 35 unidades monetárias?

1.2 Questões de interesse numa análise conjunta de duas variáveis numéricas

▷ Podemos generalizar a discussão anterior, pensando em X e Y como duas variáveis numéricas genéricas. Nesse caso, a Tabela 2 apresenta amostras emparelhadas de X e Y para n unidades experimentais criteriosamente selecionadas a partir de uma população alvo de interesse para a qual um certo estudo por amostragem está sendo desenvolvido.

Tabela 2: Conjunto de dados bivariados genérico

Id	X	Y
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

▷ Para todas as unidades que constituem nossa população alvo, estamos interessados nas seguintes questões:

Questão 1 As variáveis X e Y estão relacionadas?

Questão 2 Se X e Y estão relacionadas, essa é uma relação linear?

Questão 3 Se X e Y são variáveis linearmente relacionadas, como usar os dados para quantificar a intensidade dessa relação linear?

Questão 4 Se X e Y são variáveis linearmente relacionadas, como usar os dados para encontrar um modelo

$$\hat{y} = \hat{\mu}_Y(x) = \hat{a} + \hat{b}x \tag{3}$$

que estima relação linear populacional entre X e $\mu_Y(X)$?

Questão 5 Como fazer uma predição para Y quando $x = x_0$?

▷ A variável Y é geralmente chamada por um dos seguintes nomes: resposta, variável dependente ou variável de saída. Já a variável X é geralmente chamada por um dos seguintes nomes: covariável, variável independente ou variável de entrada. Todas as questões acima listadas estão interligadas. Um modelo linear entre X e Y só será adequado e útil, se X e Y estão linearmente relacionadas. Se esse é o caso e conseguimos encontrar boas estimativas $\hat{a}$ e $\hat{b}$ para os parâmetros a e b (na relação populacional $\mu_Y(X) = a + bX$), então teremos um modelo linear para a média de Y, denotada por μ_Y. Em outras palavras, será possível estimar o valor esperado de Y para qualquer valor fixo de x ($\hat{y} = \hat{\mu}_Y(x)$).

2 Correlação linear

▷ A análise de correlação linear busca responder as **Questões 1, 2 e 3** que foram listadas na Introdução (Seção 1) com o auxílio de duas ferramentas: o gráfico de dispersão e o coeficiente de correlação linear de Pearson.

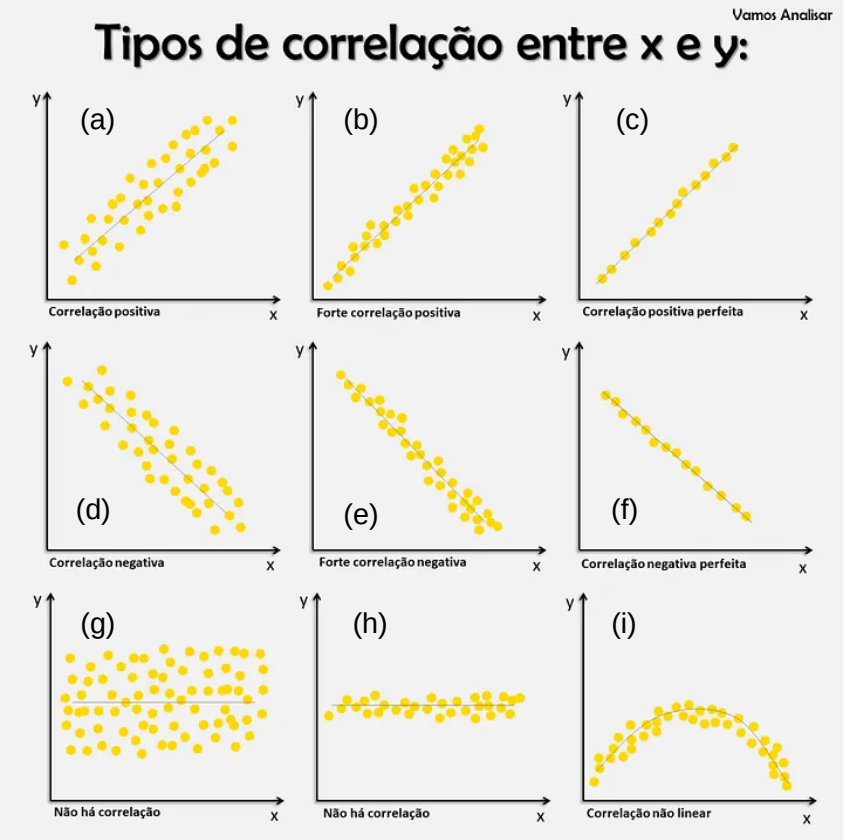

Figura 1: Exemplos de gráficos de dispersão

2.1 Gráfico de dispersão

▷ Como já descrito, a Tabela 2 apresenta amostras emparelhadas de duas variáveis X e Y para n unidades experimentais criteriosamente selecionadas a partir de uma população alvo. Uma representação gráfica importante que pode ser utilizada para visualizar conjuntos de dados estruturados conforme a Tabela 2 é o gráfico de dispersão. Esse gráfico mostra no plano cartesiano xOy todos os pares ordenados (x_i, y_i) correspondentes a todas as amostras emparelhadas de X e Y que foram listadas na Tabela 2.

▷ Figura 1 mostra diferentes exemplos de gráficos de dispersão:

1. Os gráficos (a), (b) e (c) exibem padrões onde os valores de Y crescem quando os valores de X também crescem. A única diferença é a proximidade dos pontos em relação a linha reta em torno da qual os pontos flutuam. Quanto maior é a proximidade dos pontos em relação a essa linha reta, mais intensa (ou mais forte) é a relação linear em X e Y. Nesses casos, dizemos que X e Y são variáveis que apresentam correlação positiva.

2. Os gráficos (d), (e) e (f) exibem padrões onde os valores de Y decrescem quando os valores de X crescem. Novamente, a única diferença é a proximidade dos pontos em relação a linha reta em torno da qual os pontos flutuam. Quanto maior é a proximidade dos pontos em relação a essa linha reta, mais intensa (ou mais forte) é a relação linear em X e Y. Porém, nesses casos, dizemos que X e Y são variáveis que apresentam correlação negativa.

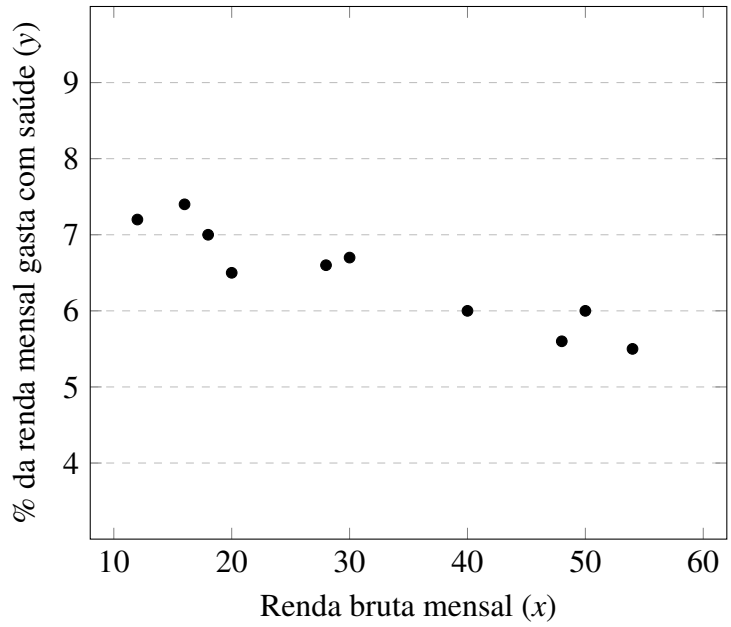

Figura 2: Diagrama de dispersão para os dados da Tabela 1

3. Os gráficos (g) e (h) não exibem nenhum padrão definido para os valores de X e Y, o que sugere nenhuma relação entre as variáveis. Nesses casos, dizemos que X e Y são variáveis não-correlacionadas.

4. O gráfico (i) exibe um padrão para os valores de X e Y, porém esse padrão não é linear visto que os pontos não flutuam em torno de uma reta. Nesse caso, dizemos que X e Y apresentam correlação não-linear.

Exemplo 1. Figura 2 mostra o gráfico de dispersão para os dados da Tabela 1. Lembre-se que essa tabela apresenta amostras emparelhadas para as variáveis renda bruta mensal familiar (variável X) e percentagem dessa renda mensal gasta com saúde (variável Y) para 10 famílias criteriosamente selecionadas a partir de uma população alvo numa certa localidade. Duas observações podem ser feitas a partir do gráfico de dispersão:

1. É razoável notar que o gráfico apresenta uma nuvem de pontos "flutuando" em torno de uma reta.

2. É razoável notar também que quanto maior é a renda bruta mensal familiar, menor é o percentual dessa renda mensal gasta com saúde.

Essas duas observações sugerem que para todas as famílias dessa localidade a renda bruta mensal familiar e a percentagem dessa renda gasta com saúde são negativamente correlacionada. Em outras palavras, as variáveis X e Y apresentam uma correlação linear negativa, ou seja, se uma variável aumenta a outra diminui e vice-versa.

2.2 Coeficiente de correlação de Pearson

▷ Considere novamente o conjunto de dados da Tabela 2 que mostra uma coleção de amostras emparelhadas de duas variáveis X e Y. A observação x_i de X está pareada com a observação y_i de Y, formando assim um par ordenado (x_i, y_i) de valores observados de X e Y para a i-ésima unidade experimental na amostra. Após construir o gráfico de dispersão desses dados e verificar que as variáveis em questão estão linearmente correlacionadas, o passo seguinte é quantificar a intensidade dessa correlação linear. Isso pode ser feito através do coeficiente de correlação amostral de Pearson, que foi desenvolvido exatamente para medir a intensidade da correlação linear entre duas variáveis numéricas. Formalmente o coeficiente de correlação amostral é definido como por:

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt[2]{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt[2]{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{4}$$

▷ Três observações são importantes nesse momento:

1. O coeficiente de correlação amostral r é um estimador (com boas propriedades) para a correlação populacional entre X e Y definida por:

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\text{DP}(X)\cdot\text{DP}(Y)} = \frac{\text{Covariância entre } X \text{ e } Y}{\text{Desvio padrão de } X \cdot \text{Desvio padrão de } Y}. \tag{5}$$

   Ou seja, $r = \hat{\rho}(X,Y)$.

2. $-1 \leq r \leq 1$, sendo que: $r = 0$ (ou $r \approx 0$) significa que as variáveis são não-correlacionadas (ou fracamente correlacionadas); $0 < r \leq 1$ significa que as variáveis são positivamente correlacionadas; $-1 \leq r < 0$ significa que as variáveis são negativamente correlacionadas; $r = -1$ ou $r = 1$ representam correlações perfeitas.

3. r não serve para medir a intensidade de correlações não-lineares.

Tabela 3: Conjunto de dados bivariados genérico junto com colunas auxiliares

Id	X	Y	-	-	-	-	-
1	x_1	y_1	$(x_1-\bar{x})$	$(y_1-\bar{y})$	$(x_1-\bar{x})(y_1-\bar{y})$	$(x_1-\bar{x})^2$	$(y_1-\bar{y})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$(x_n-\bar{x})$	$(y_n-\bar{y})$	$(x_n-\bar{x})(y_n-\bar{y})$	$(x_n-\bar{x})^2$	$(y_n-\bar{y})^2$
-	$\sum x_i$	$\sum x_i$	-	-	$\sum(x_i-\bar{x})(y_i-\bar{y})$	$\sum(x_i-\bar{x})^2$	$\sum(y_n-\bar{y})^2$

▷ Para entender detalhadamente como o coeficiente de correlação é calculado, basta considerar a Tabela 3 que mostra um conjunto de dados bivariados genérico junto com colunas auxiliares que são úteis nesse cálculo.

2.3 Fórmula alternativa para o coeficiente de correlação de Pearson

▷ Podemos deduzir uma fórmula alternativa para o coeficiente de correlação amostral. Basta observar as seguintes igualdades:

1. $S_{XY} = \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$, onde $\bar{x} = (1/n)\sum_{i=1}^n x_i$ e $\bar{y} = (1/n)\sum_{i=1}^n y_i$.
2. $S_X^2 = \sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.
3. $S_Y^2 = \sum_{i=1}^n (y_i-\bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$.

Sendo assim, é simples verificar que o coeficiente r também pode ser calculado a partir da seguinte fórmula alternativa:

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt[2]{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)\cdot\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}. \tag{6}$$

Tabela 4: Conjunto de dados bivariados genérico junto com colunas auxiliares

Id	X	Y	-	-	-
1	x_1	y_1	x_1y_1	x_1^2	y_1^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	x_iy_i	x_i^2	y_i^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	x_ny_n	x_n^2	y_n^2
-	$\sum x_i$	$\sum x_i$	$\sum x_iy_i$	$\sum x_i^2$	$\sum y_n^2$

▷ Para entender detalhadamente como o coeficiente de correlação é calculado usando a Eq. (6), basta considerar a Tabela 4 que mostra um conjunto de dados bivariados genérico junto com colunas auxiliares que são úteis nesse cálculo.

Exemplo 2. Exemplo 1 mostrou a Figura 2 que apresenta o gráfico de dispersão para os dados da Tabela 1. O gráfico de dispersão desse conjunto de dados indica que a renda bruta mensal familiar (variável X) e a percentagem dessa renda mensal gasta com saúde (variável Y) são variáveis negativamente correlacionadas. Em outras palavras, X e Y apresentam uma correlação linear negativa. Podemos calcular formalmente a intensidade dessa correlação através do coeficiente de correlação r. Para isso, basta considerar a Tabela 5.

Tabela 5: Conjunto de dados da Tabela 1 junto com colunas auxiliares para o cálculo de r

Id	X	Y	XY	X^2	Y^2
1	12	7.2	86.4	144	51.84
2	16	7.4	118.4	256	54.76
3	18	7.0	126.0	324	49.00
4	20	6.5	130.0	400	42.25
5	28	6.6	184.8	784	43.56
6	30	6.7	201.0	900	44.89
7	40	6.0	240.0	1600	36.00
8	48	5.6	268.8	2304	31.36
9	50	6.0	300.0	2500	36.00
$n = 10$	54	5.5	297.0	2916	30.25
-	$\sum x_i = 316$	$\sum y_i = 64.5$	$\sum x_iy_i = 1952.4$	$\sum x_i^2 = 12128.0$	$\sum y_i^2 = 419.91$

Para os dados em questão, temos que:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt[2]{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} = \frac{1952.4 - 10(31.6)(6.45)}{\sqrt[2]{[12128.0 - 10(31.6)^2] \cdot [419.91 - 10(6.45)^2]}} = -0.94.$$

Concluímos que a variável renda bruta mensal familiar é fortemente correlacionada como a variável percentagem dessa renda mensal gasta com saúde. Confirmamos que a correlação em questão é negativa, que significa que quanto maior for a renda bruta mensal, menor será a percentagem dessa renda mensal gasta com saúde. Uma análise confirmatória pode ser realizada através de um teste de hipóteses apropriado. Não vamos detalhar essa análise confirmatória, mas o resultado do teste mostra que a correlação entre as variáveis X e Y é estatisticamente significativa (apresentando um p-value muito pequeno). Portanto, podemos concluir que não existem evidências nos dados observados que suportam a hipótese de correlação nula entre as variáveis em questão.

3 Regressão linear simples

▷ Quando X e Y são variáveis linearmente correlacionadas, a análise de regressão linear simples busca responder as **Questões 4 e 5** que foram listadas na Introdução (Seção 1). Usaremos o método de mínimos quadrados para estimar os parâmetros a e b da relação populacional

$$\mu_Y(X) = a + bX \tag{7}$$

entre X e a média de Y, denotada aqui por μ_Y. Em outras palavras, o método de mínimos quadrados usa os dados para encontrar estimadores $\hat{a}$ e $\hat{b}$ para a e b, tal que o seguinte modelo linear estimado é obtido:

$$\hat{y} = \hat{\mu}_Y(x) = \hat{a} + \hat{b}x. \tag{8}$$

Como resultado, $\hat{y}$ pode ser usado de duas maneiras: (1) como uma estimativa pontual para a média de Y dado um valor de x ($\hat{\mu}_Y(x)$); ou (2) como uma predição pontual para Y correspondente a um novo valor de x.

3.1 Estimação de mínimos quadrados

▷ Sejam X e Y variáveis linearmente correlacionadas tal que a seguinte relação populacional $\mu_Y(X) = a + bX$ é válida, onde a e b são parâmetros populacionais geralmente desconhecidos. O método de mínimos quadrados é um método (não-paramétrico) que pode ser usado para obter estimadores para a e b. O critério de estimação utilizado pelo método de mínimos quadrados é aquele que minimiza a soma de quadrados dos erros. O i-ésimo erro ε_i correspondente a i-ésima observação pareada (x_i, y_i) nos dados é definida por: $\varepsilon_i = y_i - (a + bx_i)$. Portanto, formalmente, o método de mínimos quadrados encontra estimadores para a e b que minimizam a seguinte soma de quadrados:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2. \tag{9}$$

Como resultado desse método de estimação, encontramos que estimadores de mínimos quadrados para a e b são dados por:

$$(\hat{a}, \hat{b}) = \arg\min \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \tag{10}$$

onde

$$\hat{b} = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{e} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}. \tag{11}$$

A dedução detalhada desses resultados será omitida, pois está além dos objetivos para o desenvolvimento do nosso material de aula. Nosso foco recai na aplicação prática desses resultados para os dados da Tabela 1. Essas aplicações serão desenvolvidas nos Exemplos 3 e 4.

3.2 Modelo estimado e predições

▷ Considere um conjunto de dados com observações emparelhadas de duas variáveis X e Y linearmente correlacionadas. Em resumo, o modelo de regressão linear simples estimado é dado por

$$\hat{y} = \hat{\mu}_Y(x) = \hat{a} + \hat{b}x \tag{12}$$

onde $\hat{a}$ e $\hat{b}$ são os estimadores de mínimos quadrados. O gráfico do modelo linear estimado é a reta de mínimos quadrados para os dados.

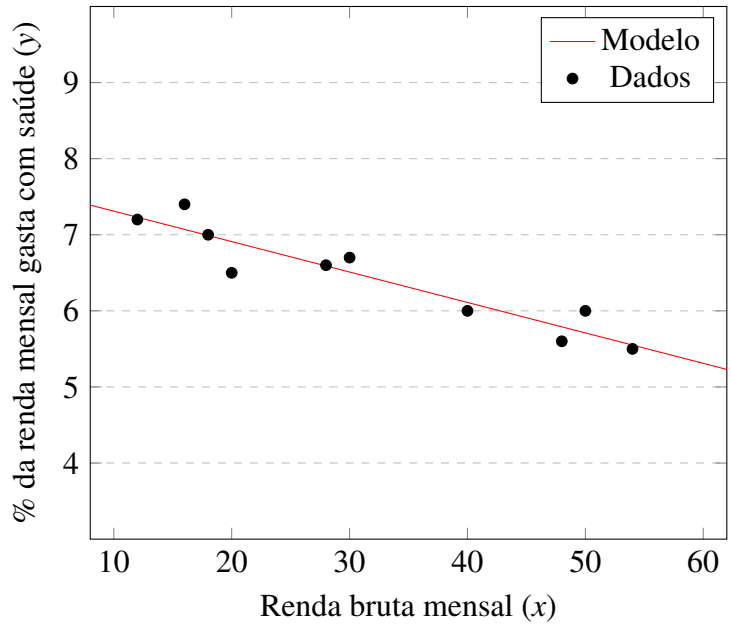

Figura 3: Diagrama de dispersão e a reta de mínimos quadrados

▷ O coeficiente de determinação, denotado por R^2, mede a variação total de Y que é explicada pelo modelo linear estimado. O coeficiente R^2 é definido por

$$R^2 = r^2 \tag{13}$$

onde r é o coeficiente de correlação linear de Pearson.

▷ O modelo linear estimado pode ser usado como uma estimativa pontual para a média de Y dado um valor fixo de x ou como uma predição pontual para Y correspondente a um novo valor de x.

Exemplo 3. Nos Exemplos 1 e 2 concluímos que para os dados da Tabela 1 a renda bruta mensal familiar é fortemente correlacionada como a percentagem dessa renda mensal gasta com saúde. A correlação em questão é negativa, que significa que quanto maior for a renda bruta mensal, menor será a percentagem dessa renda mensal gasta com saúde. Esse resultado está em conformidade com o gráfico de dispersão dos dados, apresentado na Figura 2. Podemos descrever matematicamente essa relação linear encontrando o modelo estimado $\hat{y} = \hat{\mu}_Y(x) = \hat{a} + \hat{b}x$. Para os dados em questão, as estimativas de mínimos quadrados são dadas por:

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{1952.4 - 10(31.6)(6.45)}{12128.0 - 10(31.6)^2} = \frac{-85.8}{2142.4} = -0.04$$

e

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 6.45 - (-0.04)(31.6) = 7.71.$$

Concluímos que o modelo linear estimado (ou de forma equivalente, a reta de mínimos quadrados) é dado por:

$$\hat{y} = 7.71 - 0.04x.$$

Usando os nome originais das variáveis, o modelo linear estimado também pode ser escrito como:

$$\%\text{ esperado da renda bruta mensal gasta com saúde} = 7.71 - 0.04 \cdot \text{Renda bruta mensal}.$$

Figura 3 apresenta o modelo linear estimado (ou a reta de mínimos quadrados) sobreposta aos dados no gráfico de dispersão, mostrando visualmente o ajuste do modelo aos dados em questão. O coeficiente de determinação é dado por $R^2 = (-0.94)^2 = 0.8836$, que significa que 88.36% da variação total de Y é explicada pelo modelo estimado, restando 11.64% dessa variabilidade que não é explicada pelo modelo estimado, mas que pode ser atribuída a fatores (variáveis) e incertezas não considerados no modelo.

Exemplo 4. No Exemplo 3 encontramos que o modelo de regressão linear simples estimado para os dados da Tabela 1 é dado por

$$\hat{y} = 7.71 - 0.04x.$$

Deseja-se fazer uma predição para Y quando $x = 35$. Em outras palavras, deseja-se encontrar a percentagem "esperada" da renda bruta mensal gasta com saúde para uma família com uma renda bruta mensal igual a 35 unidades monetárias (u.m.). Para responder essa questão, basta utilizar o modelo estimado fazendo $x = 35$:

$$\hat{y} = 7.71 - 0.04(35) = 6.31\%.$$

Feito isso, encontramos que para uma família com uma renda bruta mensal igual a 35 u.m., espera-se que 6.31% dessa renda mensal seja gasta com saúde.

Referências

P. A. Barbetta. Estatística Aplicada às Ciências Sociais, 5a edição. Florianópolis: Editora da UFSC, 2005.

S. Vieira. Introdução à Bioestatística, 4a edição. Rio de Janeiro: Editora Elsevier, 2011.

M. Triola. Introdução à Estatística, 7a edição. Rio de Janeiro: Editora LTC, 1999.

P. Morettin & W. Bussab. Estatística Básica, 9a edição. São Paulo: Editora Saraiva, 2010.

A Exercício

▷ Considere a Tabela 6 que apresenta observações emparelhadas para as variáveis X e Y.

1. Esboce o gráfico de dispersão dos dados.
2. Calcule o coeficiente de correlação linear.
3. Os resultados dos dois primeiros itens sugerem que X e Y são variáveis linearmente correlacionadas? Justifique sua resposta.
4. Se X e Y são variáveis linearmente correlacionadas, encontre o modelo de regressão linear simples estimado $\hat{y} = \hat{\mu}(x) = \hat{a} + \hat{b}x$ com base nos dados em questão. Calcule a variação de Y que é explicada pelo modelo linear estimado.
5. Utilize o modelo linear estimado (ou seja, a reta de mínimos quadrados) do item anterior para predizer um valor para Y quando $x = 6.3$.

Informações úteis:

- $\sum_i x_i = 25.00$.
- $\sum_i y_i = 53.70$.
- $\sum_i x_i y_i = 284.00$.
- $\sum_i x_i^2 = 135.00$.
- $\sum_i y_i^2 = 601.69$.

Tabela 6: Conjunto de dados para o Exercício

Id	X	Y
1	3.0	8.0
2	4.0	8.5
3	5.0	11.2
4	6.0	12.0
5	7.0	14.0

B Formulário

▷ A seguinte lista resume todas as fórmulas do nosso material. Todas consideram os dados da Tabela 2:

1. $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ e $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$.
2. $S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$.
3. $S_X^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.
4. $S_Y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$.
5. $S_X = \sqrt[2]{S_X^2}$ e $S_Y = \sqrt[2]{S_Y^2}$.
6. Coeficiente de correlação linear de Pearson

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt[2]{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \cdot \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}.$$

7. Estimadores de mínimos quadrados

$$\hat{b} = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{e} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

.

8. Modelo linear estimado ou reta de mínimos quadrados

$$\hat{y} = \hat{\mu}_Y(x) = \hat{a} + \hat{b}x.$$

9. Coeficiente de determinação $R^2 = r^2$.

C Script R para correlação linear e regressão linear simples

```
> #---------------------------------------------------------------------
> # SCRIPT R PARA OS EXEMPLOS 1, 2, 3 e 4.
> n <- 10
> x <- c(12,16,18,20,28,30,40,48,50,54)
> y <- c(7.2,7.4,7.0,6.5,6.6,6.7,6.0,5.6,6.0,5.5)
> f <- LETTERS[1:10]
> df <- data.frame(Familia=f,Renda=x,PercentualGasto=y)
> df
   Familia Renda PercentualGasto
1        A    12             7.2
2        B    16             7.4
3        C    18             7.0
4        D    20             6.5
5        E    28             6.6
6        F    30             6.7
7        G    40             6.0
8        H    48             5.6
9        I    50             6.0
10       J    54             5.5
> summary(df)
    Familia      Renda      PercentualGasto
 A      :1   Min.   :12.0   Min.   :5.500
 B      :1   1st Qu.:18.5   1st Qu.:6.000
 C      :1   Median :29.0   Median :6.550
 D      :1   Mean   :31.6   Mean   :6.450
 E      :1   3rd Qu.:46.0   3rd Qu.:6.925
 F      :1   Max.   :54.0   Max.   :7.400
 (Other):4

> sd(df[,2])
[1] 15.42869
> sd(df[,3])
[1] 0.6570134
> cor(df[,2:3],method="pearson")
                    Renda PercentualGasto
Renda            1.0000000      -0.9404625
PercentualGasto -0.9404625       1.0000000
> cor.test(df[,2],df[,3],method="pearson")

 Pearson's product-moment correlation

data:  df[, 2] and df[, 3]
t = -7.826, df = 8, p-value = 5.114e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9861500 -0.7621149
sample estimates:
       cor
-0.9404625
> #---------------------------------------------------------------------
```

```
> #---------------------------------------------------------------------
> modelo <- lm(PercentualGasto~Renda,data=df)
> summary(modelo)

Call:
lm(formula = PercentualGasto ~ Renda, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41456 -0.09842 -0.01481  0.14090  0.32524

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.715534   0.178214  43.294 8.94e-11 ***
Renda       -0.040049   0.005117  -7.826 5.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2369 on 8 degrees of freedom
Multiple R-squared:  0.8845, Adjusted R-squared:   0.87
F-statistic: 61.25 on 1 and 8 DF,  p-value: 5.114e-05

> predicao <- predict(modelo,data.frame(Renda=35))
> predicao
       1
6.313835
> #---------------------------------------------------------------------
> plot(df[,2:3],type="p",xlim=c(10,60),ylim=c(4,9),
                xlab="Renda bruta mensal",
                ylab="% da renda mensal gasta com saúde",
                pch=16)
> text(38.0,3.9,"x = 35")
> text(13.5,6.40,"y.hat = 6.31%")
> abline(modelo,col="red")
> abline(v=35,col="blue")
> abline(h=predicao,col="blue")
> points(x=35,y=predicao,pch=16,col="blue")
> dev.off()
> #---------------------------------------------------------------------
```