Introdução aos Modelos Lineares Clássicos

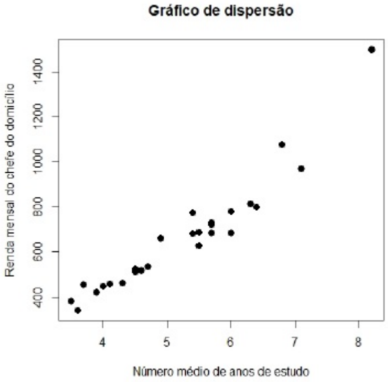
Professor: Pedro Almeida

UEPB

Roteiro

Roteiro

- Considere dados extraídos do censo de 2000 que apresenta para cada unidade da federação o **número médio de anos de estudo** e a **renda média mensal dos chefes de domicílio**.



Motivação

Queremos responder as seguintes perguntas:

- A renda mensal dos chefes de domicílio pode ser explicada através do número médio de anos de estudo destes?
- Existe um valor numérico que quantifica a relação entre as duas variáveis?
- Para um valor fixado de anos de estudo de um certo chefe de domicílio, qual será o valor previsto para a renda mensal dos chefes de domicílio?
- O quanto da variabilidade da renda mensal pode ser explicado através do número médio de anos de estudo?

- Existem situações nas quais há interesse em estudar o comportamento conjunto de uma ou mais variáveis;
- Em muitos casos, a explicação de um fenômeno de interesse pode estar associado a outros fatores (variáveis) que contribuem de algum modo para a ocorrência deste fenômeno;
- O comportamento conjunto de duas variáveis quantitativas pode ser observado por meio do gráfico de dispersão.

- O termo regressão foi utilizado pela primeira vez por Galton em 1885, em um estudo que avaliou a relação entre a altura dos pais e filhos.
- O **objetivo** do estudo foi avaliar como a altura do pai **influenciava** a altura do filho. Por esse motivo, Galton denominou de regressão, por existir uma tendência de regressão à média.
- Em muitas situações, temos interesse em verificar existências de relações entre duas ou mais variáveis. Nesse sentido, a **Análise de regressão linear** é uma técnica estatística para modelar e quantificar a relação entre duas ou mais variáveis.

Roteiro

Tipos de Modelos

Um **modelo de regressão** é um modelo estatístico em que alguma característica distribuicional da variável de interesse é afetada por outra(s) variável(is).

- É uma das técnicas de modelagem mais usadas;
- Possui ampla literatura como:
  - Modelo de Regressão Linear Múltipla
  - Modelo de regressão Não-Linear
  - Modelo Linear Generalizado

Roteiro

- No entanto, antes de propor um modelo de regressão é importante verificar o grau de correlação entre as **variáveis independentes x** e a **variável resposta y**.
- Além disso, nem sempre uma correlação elevada entre variáveis indica que faz sentido propor um modelo de regressão
  - Ex.: Produção de banana *versus* taxa de natalidade.
- A **coerência** e **intuição** do pesquisador é muito importante no momento de propor uma relação entre x e y.

Mapas de dispersão e tipos de correlação

Figura 1: Comportamento do coeficiente de correlação.

Coeficiente de Correlação Linear

Mede a intensidade e a direção da relação linear entre duas variáveis.

$$
\begin{aligned}
\rho(X,Y) &= \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \\
&= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}, \quad i = 1, \ldots, n.
\end{aligned}
$$

em que n: tamanho da amostra; y: variável dependente; x: variável independente; σ_X e σ_Y é o desvio de X e Y, respectivamente; Cov(X, Y) é a covariância entre duas variáveis.

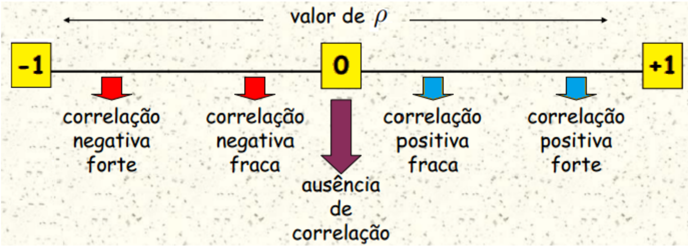


Figura 2: Variação do coeficiente de correlação.

- Se $\rho = 1$ implica correlação **linear positiva e perfeita**;
- Se $\rho = -1$ implica correlação **linear negativa e perfeita**;
- Se $\rho = 0$ **inexistência** de correlação linear.

Modelo de Regressão Linear Simples (MRLS)

- Seja Y uma variável resposta, e seja X uma variável denominada de regressora (FREIRE et al, 2008).
- O MRLS descreve a variável Y como uma soma de uma quantidade determinística e uma quantidade aleatória.
    - Parte determinística: uma reta em função de X.
    - Parte aleatória: denominada de erro.

Modelo de Regressão Linear Simples (MRLS)

O modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis.

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

- β_0 e β_1 são os parâmetros da regressão que são desconhecidos e devem ser estimados;
- ϵ, é o erro que é uma variável aleatória não-observável, em que é suposto que $\mathbf{E}(\epsilon) = 0$ e $\mathbf{Var}(\epsilon) = \sigma^2$.

Suposições

- S0 O modelo está correto;
- S1 $\text{E}(\epsilon_i) = 0 \ \forall_i$;
- S2 $\text{Var}(\epsilon_i) = \sigma^2 \ \forall_i \ (0 < \sigma^2 < \infty)$;
- S3 $\text{Cov}(\epsilon_i, \epsilon_s) = 0 \ \forall \ i \neq s$;
- S4 x assume pelo menos dois valores;
- S5 Normalidade.

Considere o modelo de regressão linear simples. Então a distribuição de Y, correspondente ao valor prefixado, x, de X, é dado por:

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 x; \sigma^2).$$

Prova:

$$\begin{aligned} E(Y|x) &= E\left(\beta_0 + \beta_1 x + \epsilon\right) \\ &= E\left(\beta_0 + \beta_1 x\right) + E(\epsilon) \\ &= \beta_0 + \beta_1 x + 0. \end{aligned}$$

Modelo de Regressão Linear Simples (MRLS)

$$\begin{aligned}
\mathrm{Var}(Y|x) &= \mathrm{Var}\left(\beta_0 + \beta_1 x + \epsilon\right) \\
&= \mathrm{Var}\left(\beta_0 + \beta_1 x\right) + \mathrm{Var}(\epsilon) \\
&= 0 + \sigma^2 \\
&= \sigma^2.
\end{aligned}$$

Expressando este modelo usando notação matricial. Sejam os vetores

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{e} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

E seja a matriz X:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Então,

$$\mathbf{X}\beta + \epsilon = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \epsilon_1 \\ \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \epsilon_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{y}$$

O vetor aleatório ϵ é composto de variáveis independentes, com distribuição $\mathcal{N}(0; \sigma^2)$. Desta forma, o vetor de esperanças dos elementos de ϵ é o vetor nulo de dimensão n e a matriz, cuja diagonal é formada pelas variâncias e os demais elementos são as covariâncias, logo a **matriz de variância e covariância** é dado por

$$\begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & 0 & \dots \sigma^2 \end{bmatrix} = \sigma^2 I,$$

sendo I a matriz identidade de ordem n.

Roteiro

- A estimação de β_0 e β_1 pode ser feita pelo **Método dos Mínimos Quadrados**, que não requer qualquer hipótese sobre a distribuição das componentes do vetor y, e consiste em minimizar.

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

em que y_i e x_i são os valores observados de Y_i e X_i, respectivamente, com $i = 1, 2, \cdots, n$. Ao minimizar $S(\beta_0; \beta_1)$ com respeito a β_0 e β_1, estaremos minimizando a informação perdida ao utilizar o MRLS para modelar Y.

Para encontrar esse mínimo precisamos obter as seguintes derivadas parciais:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Método dos Mínimos Quadrados

Denominado por $\hat{\beta}_0$ e $\hat{\beta}_1$ os valores que minimizam a função temos

$$-2\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right] = 0$$

$$-2\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right] x_i = 0,$$

denominado sistema de equações normais.

Para garantir que $(\hat{\beta}_0; \hat{\beta}_1)$ é de fato ponto de mínimo da função $(\beta_0; \beta_1)$, precisamos obter a matriz de segundas derivadas e mostrar que está é não-negativa definida. Prove!

Logo após alguma álgebra

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Assim, para prever valores de Y para valores fixados de X, utiliza-se

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0;$$

em que x_0 é um valor fixado para a covariável e $\hat{y}_0$ é o valor previsto para a variável resposta. O valor fixado para x_0 deve estar próximo aos limites do valores observados de X da amostra utilizada para estimar os coeficientes da regressão.

Interpretação dos Coeficientes Estimados

Para interpretar o coeficiente estimado $\hat{\beta}_0$, tome $x_i = 0$. Então, o MRLS ajustado é

$$\hat{\mu}_i = \hat{y}_i = \beta_0$$

- Note que $\hat{\beta}_0$ é o ponto onde a reta de regressão ajustada **intercepta o eixo x**.

- $\hat{\beta}_0$ é uma estimativa para a média da variável resposta quando a covariável assume valor zero.

Exemplo 2

Para interpretar $\hat{\beta}_1$, considere o aumento de uma unidade no valor da covariável, isto é, $x_0 = x + 1$. Então,

$$\begin{aligned} \hat{y} &= \hat{y}(x') - \hat{y}(x) \\ &= \hat{\beta}_0 + \hat{\beta}_1 x' - \hat{\beta}_0 - \hat{\beta}_1 x \\ &= \hat{\beta}_0 + \hat{\beta}_1 (x+1) - \hat{\beta}_0 - \hat{\beta}_1 x \\ &= \hat{\beta}_1 \end{aligned}$$

Logo, além de $\hat{\beta}_1$ ser o coeficiente angular da reta de regressão, este, também, é o quanto o valor da média estimada de Y varia quando aumentamos uma unidade da variável X.

Exemplo 2

Encontre a reta de mínimos quadrados e os resíduos para os seguintes pares de valores (FREIRE et al, 2008):

x	1	1	2	2	3	3
y	1	3	1	3	1	3

Solução:

	x	y	x^2	xy	$\hat{y}$	$\epsilon = y - \hat{y}$
	1	1	1	1	2	-1
	1	3	1	3	2	1
	2	1	4	2	2	-1
	2	3	4	6	2	1
	3	1	9	3	2	-1
	3	3	9	9	2	1
soma	12	12	28	24	12	0

Exemplo 2

Usando a soma das 4 primeiras colunas da tabela, obtemos

$$\bar{x} = \frac{12}{6} \quad \bar{y} = \frac{12}{6} = 2,$$

$$\hat{\beta}_1 = \frac{24 - \frac{1}{6}(12 \times 12)}{28 - \frac{1}{6}(12)^2} = 0 \quad \hat{\beta}_0 = 2 - 0(2) = 2$$

portanto, a reta de mínimos quadrados é $\hat{y} = 2$.

Propriedades do Ajuste de Mínimos Quadrados

- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$

$$\text{Prova:} \qquad \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i} = 0.$$

- $\sum_{i=1}^{n} e_i = 0$

$$\text{Prova:} \qquad \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0$$

$$\rightarrow \sum_{i=1}^{n} \underbrace{(y_i - \hat{y}_i)}_{e_i} = 0.$$

Propriedades do Ajuste de Mínimos Quadrados

- $\sum_{i=1}^{n} x_i e_i = 0$

Prova: $$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$
$$\rightarrow \sum_{i=1}^{n} x_i \underbrace{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}_{e_i} = 0.$$

- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Prova: $$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 - \hat{\beta}_1 x_i) e_i$$
$$= \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} e_i}_{0} - \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i e_i}_{0}.$$

Estimadores de Mínimos Quadrados para o MRLS

- $\mathrm{E}(\hat{\beta}_1) = \beta_1$ e $\mathrm{E}(\hat{\beta}_0) = \beta_0$
- $\mathrm{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ e $\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$

Precisamos estimar mais um parâmetro "a variância do erro, σ^2", o qual representa a distorção da reta. Um estimador não-viesado de σ^2 para o modelo de regressão simples é dado por

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

sob o MRLS, $\frac{(n-2)\hat{\sigma^2}}{\sigma^2} \sim \mathcal{X}^2_{(n-2)}$.

- Técnica mais usada para verificar a adequação do ajuste do modelo de regressão a um conjunto de dados, baseada na seguinte identidade

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2\\ \text{SQT} &= \text{SQE} + \text{SQR}\end{aligned}$$

Coeficiente de determinação - R^2

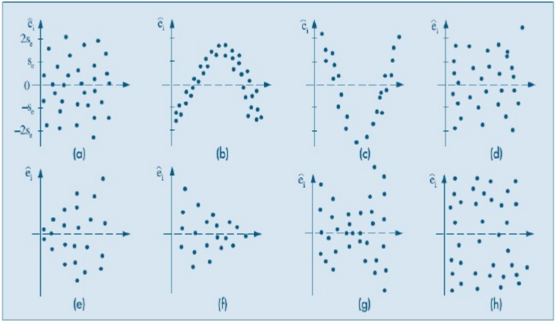
O coeficiente de correlação múltipla de Pearson (ou coeficiente de determinação) R^2 expressa o quanto o modelo explica a variabilidade total da variável y.

$$R^2 = \frac{SQE}{\mathrm{SQT}}$$

- O coeficiente R^2 é interpretado como a proporção da variação de Y que é explicada pela covariável X. ($\in (0, 1)$)
- Finalidade: Medir o poder de explicação de um modelo.

- resíduos ordinários;
- resíduos padronizados.

Figura 16.7: Gráficos de resíduos. (a) situação ideal; (b), (c) modelo não-linear; (d) elemento atípico; (e), (f), (g) heterocedasticidade; (h) não-normalidade.



Fonte: MORETTIN e BUSSAB (2010).

Tabela ANOVA

A tabela da **ANOVA** é usada para testar a adequação global do modelo de regressão

Efeito	Soma de Quadrados	G.L.	Média de Quadrados	Estatística
Regressão	SQE	1	MQE=SQE/(1)	F=MQE/MQR
Residual	SQR	$n-2$	MQR= SQR/($n-2$)	
Total	SQT	$n-1$		

Teste F- Adequação Global

- Hipóteses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- Estatística de Teste

$$F = MQE/MQR$$

- **Se** $F > F_{1,n-2}(\alpha)$ **rejeita** H_0, logo o efeito global de pelo menos algumas variáveis presentes na matriz X explica a variabilidade de y.

Teste F- Adequação Global

- A estatística do teste F representa o quociente entre **SQE** e **SQR** e têm distribuição $\mathcal{X}^2$, pelos respectivos G.L., logo temos que

$$F \sim F_{1,\, n-2},$$

que representa o valor de uma distribuição F-Snedecor com 1 e $n-2$ graus de liberdade, ao nível de significância α.

Seleção das Variáveis Explicativas

- O **Teste F** permite apenas inferir que algumas variáveis explicativas são realmente importantes (mas não sabemos quais!).
- O **Teste t** permite selecionar as variáveis independentes (explicativas) que são significativas para o modelo.

Seleção das variáveis explicativas- Teste t

- Obter um modelo parcimonioso;
- Eliminar variáveis que tem pouca ou nenhuma contribuição na variabilidade da variável dependente y.
- Hipóteses:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0.$$

- Estatística do Teste

$$T = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

- **Se** $T < t_{n-2}(\alpha/2)$ **não rejeita** H_0, logo a variável independente X não é significativa para explicar a variabilidade da resposta.

Exemplo 3

Comandos utilizados para o ajuste do modelo de regressão liner simples no *software* R.

```
#dados
x=c(5,8,7,10,6,7,9,3,8,2)
y=c(6,9,8,10,5,7,8,4,6,2)

#Ajuste do modelo ajuste=lm(y ~ x)
summary(ajuste)
```

Exemplo 3

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7949 -0.6474  0.2735  0.7265  1.2051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8889     0.9567   0.929 0.380023
x             0.8632     0.1379   6.258 0.000244 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 8 degrees of freedom
Multiple R-squared:  0.8304,    Adjusted R-squared:  0.8092
F-statistic: 39.16 on 1 and 8 DF,  p-value: 0.0002437
```

Figura 4: Resultados do ajuste obtidos no R.

Referências

Paula, Gilberto Alvarenga. (2004) Modelos de regressão: com apoio computacional. São Paulo: IME-USP.