Análise Multivariada
Regressão

ACH2036 – Métodos Quantitativos Aplicados à Adm. de Empresas I

Prof. Regis Rossi A. Faria

2º sem. 2020

Créditos: parte do conteúdo baseado em slides da profa. Ana Amélia Benedito Silva

Regressão

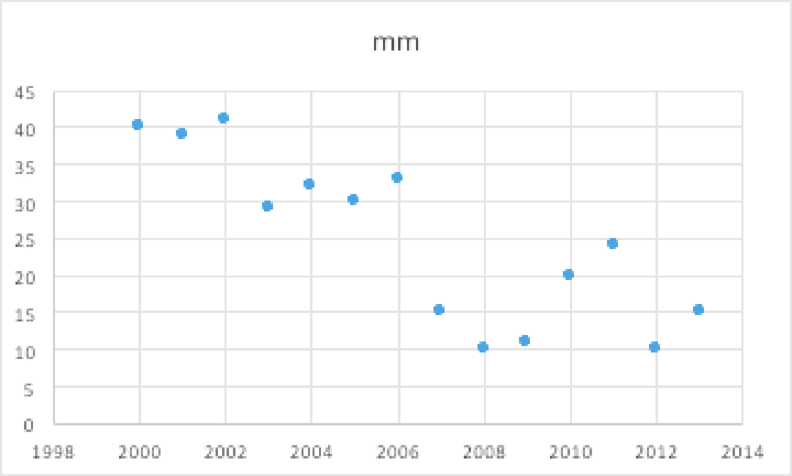
- Objetivos:
  - conjugar variáveis para predizer melhor um fenômeno de interesse;
  - encontrar relação causal entre variáveis
- Estabelece relação entre uma variável dependente (quantitativa) e duas ou mais variáveis independentes ou explicativas (quantitativas ou qualitativas)

Regressão

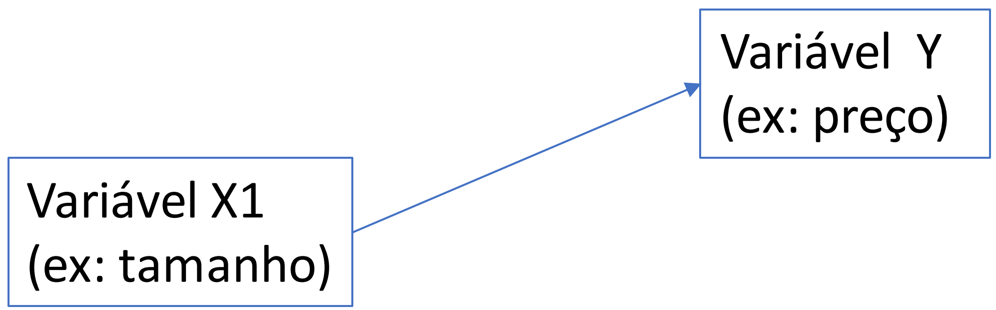
- Exemplos:
  - Prever venda de produto (variável dependente Y) a partir dos investimentos em propaganda (X1), do nível de desconto (X2)
  - Prever gasto familiar (Y) a partir da renda familiar (X1) e número de membros da família (X2)
  - Prever valor do colesterol (Y) a partir do sexo (X1), da idade (X2), do peso (X3), e da dieta (X4)
  - Prever preço de um apartamento (Y) a partir do seu tamanho (X1), idade (X2), localização (X3) e número de quartos (X4)

Regressão

- Análise de regressão: usualmente empregada para o propósito de previsão, do valor de uma variável (dependente) em função de outras (independentes)
- Pressupõe: existência de uma dependência estatística entre variáveis (ex: a precipitação de neve e o tempo em anos)

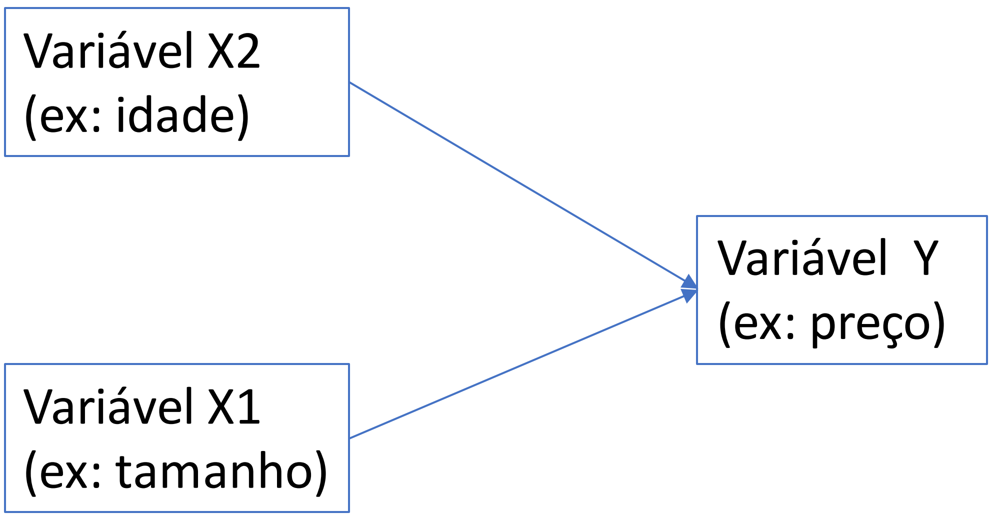


Regressão Simples



$$Y = a + b_1 X_1$$

Regressão Múltipla



$$Y = a + b_1X_1 + b_2X_2$$

Regressão Múltipla

- Uma generalização desta combinação linear é dada por

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \mathcal{E}$$

onde Y é a variável dependente, X_i são as variáveis independentes, b_i são os coeficientes ou parâmetros da regressão, Ɛ é o resíduo ou erro da regressão, e a é o intercepto (interseção da reta com o eixo Y).

Regressão Múltipla

- Considerações gerais
  - A variável Y é aleatória
  - A esperança matemática (média) dos resíduos é nula
  - A variância dos termos de erro é constante (e igual a σ^2) o que implica na *homoscedasticidade dos resíduos*, ou dispersão homogênea de Y em relação a cada valor de X
  - Os resíduos são independentes entre si
  - Os resíduos têm distribuição normal

Exemplos

- Gastos de uma academia
  - Planilha
  - R-Studio

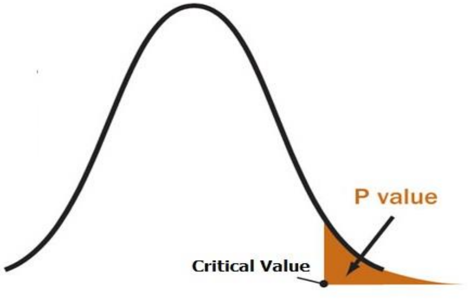
- Antes porém vamos ver alguns conceitos importantes para compreender os parâmetros extraídos dos modelos de análise

Regressão Múltipla

- Conceitos importantes
  - R: *coeficiente de correlação*: representa o grau de associação entre duas variáveis (valor entre -1 a +1)
  - R^2 e R^2 ajustado: *coeficiente de determinação* ou poder explicativo da regressão: expressa percentual de compreensão ou explicação do fenômeno; índice que mostra melhor explicação do fenômeno pelas variáveis em uso
    - R^2 ajustado é utilizado para comparar modelos com diferentes quantidades de variáveis
  - Erro padrão: medida da precisão da previsão: expressa um desvio padrão em torno da reta de regressão

Regressão Múltipla

- Conceitos importantes
  - Valor p: expressa a significância de uma evidência ou dado obtido; permite estabelecer um grau de confiabilidade do resultado
  - O valor p está associado à porção extrema da distribuição de probabiidade



Conceitos importantes

- O valor p é usado em teste de hipótese, dá evidência contra ou a favor da hipótese nula (H0)
- Quanto menor o valor p, maior a evidência de que deve-se rejeitar a hipótese nula (H0), isto é, a hipótese de que não há uma relação estatística significativa
  - Um valor p pequeno ($p<0.05$) rejeita H0, isto é, dá evidências de que esta hipótese é inválida
  - Um valor p grande ($p>0.05$) dá evidências de que a hipótese alternativa é fraca, isto é, a hipótese nula é significativa

Conceitos importantes

- Na ausência de um valor de referência (nível alfa) para significância, pode-se usar:
  - se $p > .10$ → "não significante"
  - se $p \leq .10$ → "marginalmente significante"
  - se $p \leq .05$ → "significante"
  - se $p \leq .01$ → "altamente significante"

Conceitos importantes

- Soma dos quadrados dos resíduos (SQR ou SSE na sigla em inglês): nosso objetivo é usar um modelo explicativo que sempre minimize o SQR
- Valor p do Teste F de significância global: avalia se o modelo é útil para explicar a relação conjugada da variável dependente e as independentes
  - Sig: F<0,05 → evidências de que o modelo é útil

Conceitos importantes

- Teste T ANOVA: testa o efeito conjunto de variáveis independentes sobre a variável dependente: verifica a probabilidade de que os parâmetros da regressão em conjunto sejam =0, isto é, não há relação estatística significativa
  - H0: $R^2 = 0$
  - Em oposição temos que H1: $R^2 > 0$ (hipótese alternativa tem significância)

Análise de regressão

- A análise de regressão estuda o relacionamento entre uma variável chamada variável dependente e outras variáveis chamadas variáveis independentes.
- Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes.

Análise de regressão

- Este modelo é designado por modelo de regressão linear simples se define uma relação linear entre a variável dependente e uma variável independente.
- Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se modelo de regressão linear múltipla.

Regressão simples → Regressão múltipla

$$Y = a + b_1 X_1$$

$$Y = a + b_1 X_1 + b_2 X_2$$

Análise de correlação

- A análise de correlação dedica-se a inferências estatísticas das medidas de associação linear que se seguem:
  - coeficiente de correlação simples: mede a “força” ou “grau” de relacionamento linear entre 2 variáveis;
  - coeficiente de correlação múltiplo: mede a “força” ou “grau” de relacionamento linear entre uma variável e um conjunto de outras variáveis.
- As técnicas de análise de correlação e regressão estão intimamente ligadas.

Diagrama de Dispersão

- Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

- Com os dados constrói-se o diagrama de dispersão. Este deve exibir uma tendência linear para que se possa usar a regressão linear.
- Este diagrama permite, portanto, decidir empiricamente se um relacionamento linear entre X e Y deve ser assumido.

Diagrama de Dispersão

- Exemplo: o plot de gastos (Y) *versus* horas de mão de obra direta (X) sugere uma relação linear entre as variáveis
- E por análise do diagrama de dispersão pode-se também concluir (empiricamente) se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma reta imaginária que passa através de uma nuvem de pontos.

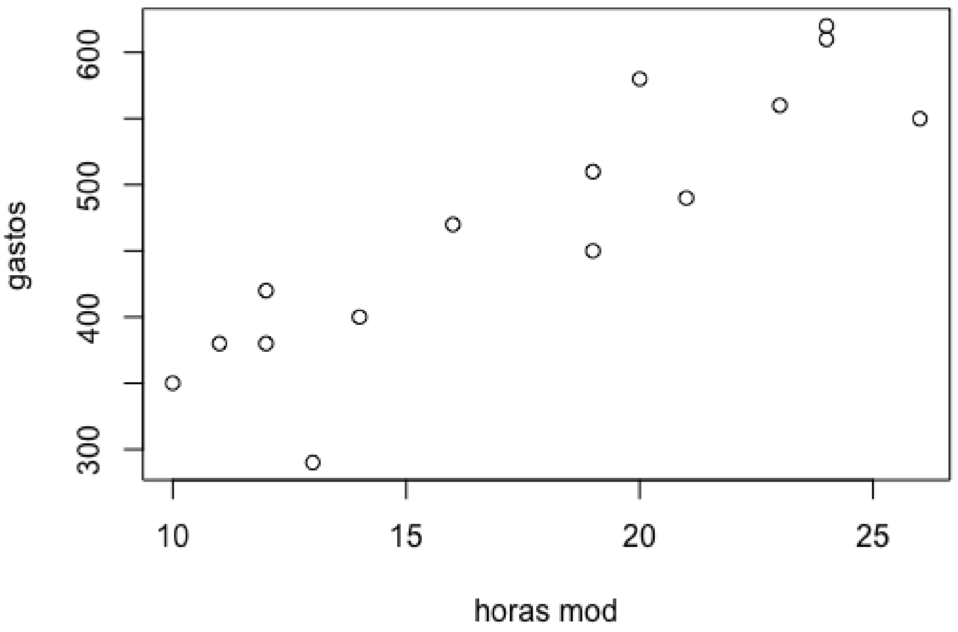
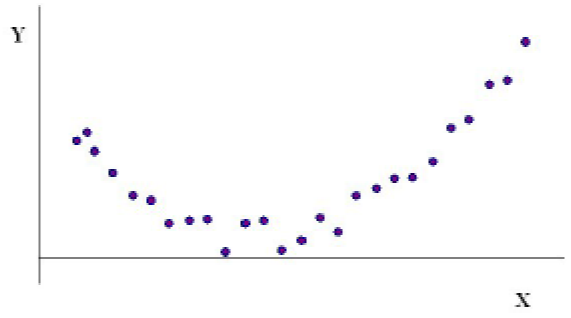


Diagrama de Dispersão

- A correlação é tanto maior quanto mais os pontos se concentram, com pequenos desvios, em relação a essa reta.
- Se o declive da reta é positivo, concluímos que a correlação entre X e Y é positiva, i.e., os fenômenos variam no mesmo sentido.
- Ao contrário, se o declive é negativo, então a correlação entre X e Y é negativa, i.e., os fenômenos variam em sentido inverso.

Diagrama de Dispersão

Sugerem uma regressão não linear
(i.e., a relação entre as duas variáveis poderá ser descrita por uma equação não linear)



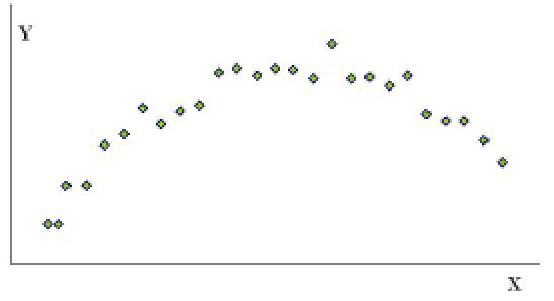
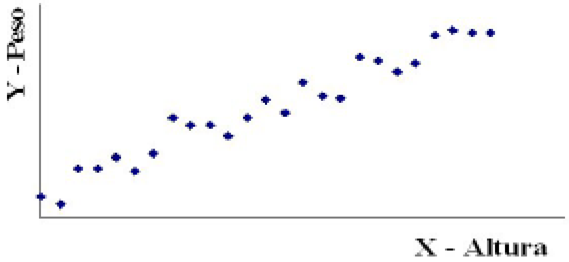
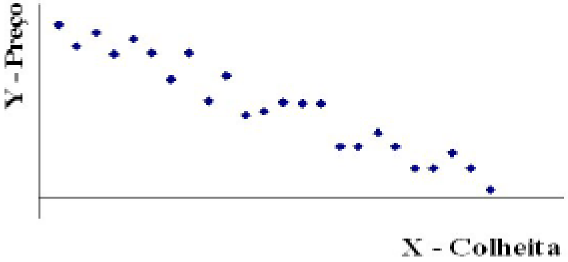


Diagrama de Dispersão

Sugerem uma regressão linear
(i.e., a relação entre as duas variáveis poderá ser descrita por uma equação linear)





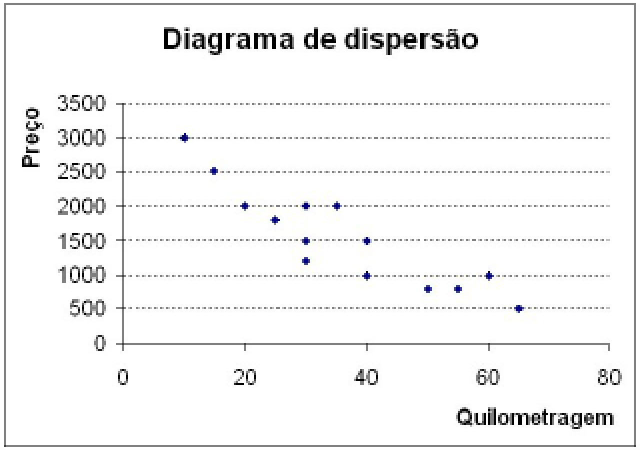
Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)

Existência de correlação negativa (em média, quanto maior for a colheita menor será o preço)

Exemplo de regressão linear simples

Queremos estudar a relação entre a quilometragem de um carro usado e o seu preço de venda

Carros	Quilometragem X (1000 Km)	Preço de venda Y (dezena de Euros)
1	40	1000
2	30	1500
3	30	1200
4	25	1800
5	50	800
6	60	1000
7	65	500
8	10	3000
9	15	2500
10	20	2000
11	55	800
12	40	1500
13	35	2000
14	30	2000
Total	**505**	**21600**



Os dados sugerem uma relação linear entre a quilometragem e o preço de venda. Existe uma **correlação negativa**: em média, quanto maior for a quilometragem menor será o preço de venda.

Modelo de regressão linear simples

$$Y = \beta_0 + \beta_1 X + E$$

X – variável explicativa ou independente medida sem erro (não aleatória);
E – variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X;
β_0 e β_1 – parâmetros desconhecidos do modelo (a estimar);
Y – variável explicada ou dependente (aleatória).

Exemplos

1. Relação entre o peso e a altura de um homem adulto (X: altura; Y: peso)
2. Relação entre o preço do vinho e o montante da colheita em cada ano (X: montante da colheita; Y: preço do vinho)

Modelo de regressão linear simples

Num estudo de regressão temos n observações da variável X : $x_1, x_2, \ldots, x_n$ (assume-se que estas observações são medidas sem erro).

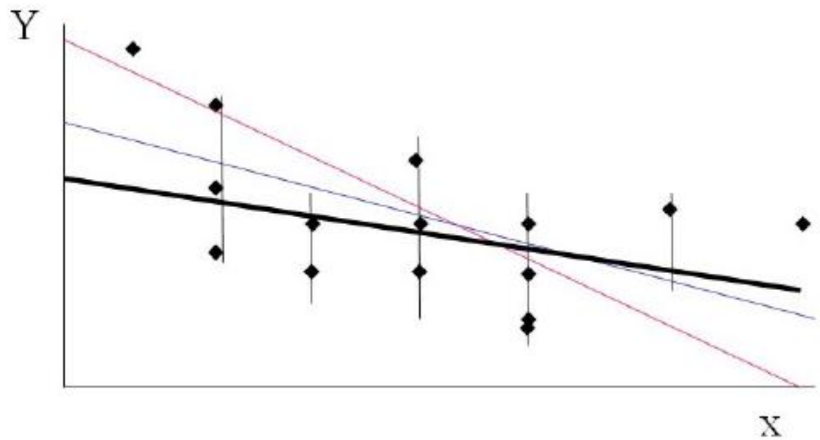
Temos então n variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$ tais que:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \qquad i = 1, ..., n$$

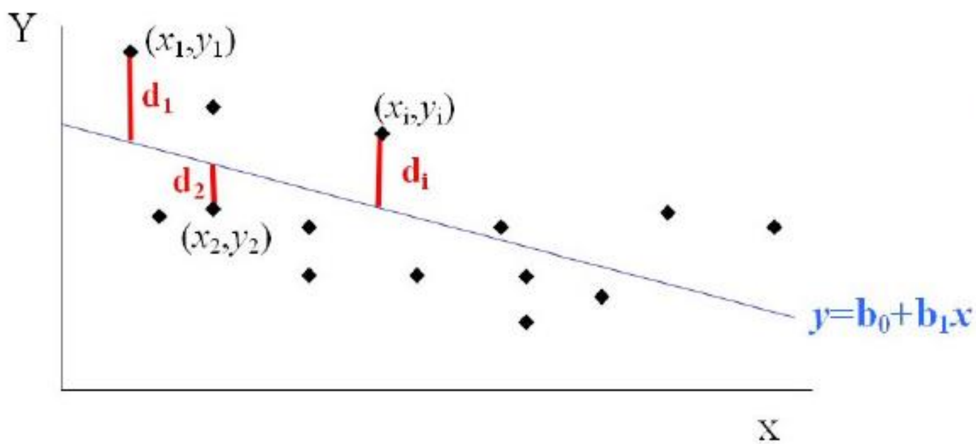
Admite-se que $E_1, E_2, ..., E_n$ são variáveis aleatórias independentes de média zero e variância σ^2.

Para qualquer valor x_i de X, Y_i é uma variável aleatória de média $\mu_{Y_i} = \beta_0 + \beta_1 x_i$ e variância σ^2

Estimação pelo método dos mínimos quadrados



Qual a recta que melhor se ajusta?



$$\hat{y}_i = b_0 + b_1 x_i$$

é o valor dado pela recta

$$d_i = y_i - (\mathbf{b_0} + \mathbf{b_1} x_i)$$

resíduos

Estimação pelo método dos mínimos quadrados

- Iremos estimar os parâmetros usando o método dos mínimos quadrados.

Seja $d_i = y_i - \hat{y}_i \rightarrow$ *i-ésimo resíduo*

- O objetivo é escolher b_0 e b_1 de modo a minimizar a soma dos quadrados destes resíduos.

$$SSE = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

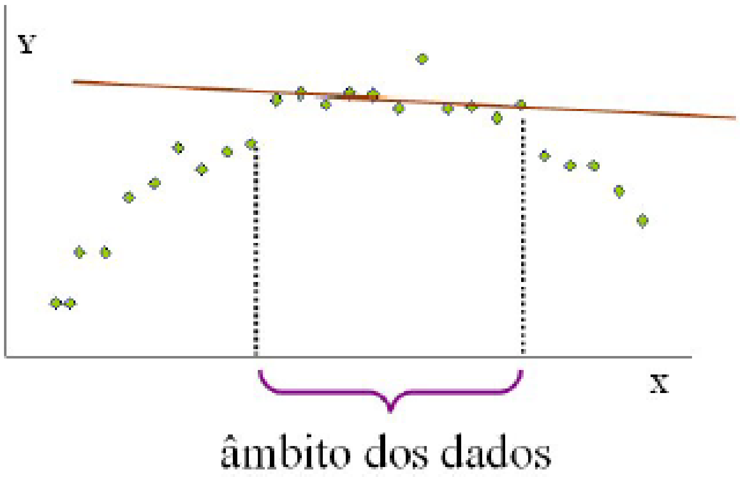
Estimação pelo método dos mínimos quadrados

Para determinar b_0 e b_1, de modo a minimizar SSE resolve-se o seguinte sistema de equações:

$$\begin{cases} \frac{\partial SSE}{\partial b_0} = 0 \\ \\ \frac{\partial SSE}{\partial b_1} = 0 \end{cases} \Leftrightarrow \ldots \Leftrightarrow \begin{cases} b_0 = \overline{y} - b_1 \overline{x} \\ \\ b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\, \overline{x}\, \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\, \overline{x}^2} \end{cases}$$

ATENÇÃO:

Um conjunto de pontos dá evidência de linearidade apenas para os valores de X cobertos pelo conjunto de dados. Para valores de X que saem fora dos que foram cobertos não há qualquer evidência de linearidade. Por isso é arriscado usar uma recta de regressão estimada para predizer valores de Y correspondentes a valores de X que saem fora do âmbito dos dados.



O perigo de extrapolar para fora do âmbito dos dados amostrais é que a mesma relação possa não mais se verificar.

Modelo de regressão linear múltipla

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + E$$

$X_1, \ldots, X_k$ – variáveis explicativas ou independentes medidas sem erro (não aleatórias);

E – variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento das variáveis $X_1, \ldots, X_k$ e os possíveis erros de medição;

$\beta_0, \ldots, \beta_k$ – parâmetros desconhecidos do modelo (a estimar);

Y – variável explicada ou dependente (aleatória).

Modelo de regressão linear múltipla

Exemplos

- Relação entre o volume de vendas (Y) efetuadas durante um dado período de tempo por um vendedor, seus anos de experiência (X_1) e seu escore num teste de inteligência (X_2).
- Vendedores com 4 anos de experiência ($x_1 = 4$) e *escore* 3 no teste de inteligência ($x_2 = 3$), podem apresentar volumes de vendas diferentes (Y´s diferentes).
- Isto é, fixando a variável anos de experiência - X_1 - num valor, por exemplo 4 anos, e X_2 noutro valor, por exemplo 3, o volume de vendas vai variar devido a outras influências aleatórias.

Para x_1 e x_2 fixos, Y é uma variável aleatória.

Modelo de regressão linear múltipla

Num estudo de regressão temos n observações de cada variável independente:

$$
\begin{array}{ccccc}
 & i=1 & i=2 & \ldots & i=n \\
X_1 & x_{11} & x_{12} & \ldots & x_{1n} \\
X_2 & x_{21} & x_{22} & \ldots & x_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
X_k & x_{k1} & x_{k2} & \ldots & x_{kn}
\end{array}
$$

Para cada i, i.e., para $x_{1i}, \ldots, x_{ki}$ fixos, Y_i é uma variável aleatória.

Temos então n variáveis aleatórias: $Y_1,\ Y_2,\ \ldots,\ Y_n$:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + E_i \qquad i = 1, ..., n$$

Modelo de regressão linear múltipla

$$Y_1 = \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{k1} + E_1$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 x_{1n} + \ldots + \beta_k x_{kn} + E_n$$

Admite-se que $E_1, \ldots, E_n$ são variáveis aleatórias independentes de média zero e variância σ^2

Então, para quaisquer valores $x_{1i}, \ldots, x_{ki}$ fixos, Y_i é uma variável aleatória de média

$$\mu_{Y_i} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

e variância σ^2.

Modelo de regressão linear múltipla

- Em outras palavras:

Os dados para a análise de regressão e de correlação múltipla são da forma:

$$(y_1, x_{11}, x_{21}, \ldots, x_{k1}),\ (y_2, x_{12}, x_{22}, \ldots, x_{k2}), \ldots, (y_n, x_{1n}, x_{2n}, \ldots, x_{kn})$$

Cada observação obedece à seguinte relação:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}_{\mu_{Y_i}} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

$\hookrightarrow$ O valor observado de uma variável aleatória (y_i), usualmente difere da sua média (μ_{Y_i}) por uma quantidade aleatória ε_i.

Modelo de regressão linear múltipla

Temos então o seguinte sistema escrito em notação matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow y = X\beta + \varepsilon$$

y - Vector das observações da variável dependente;
X - Matriz significativa do modelo;
β - Vector dos parâmetros do modelo;
ε - Vector das realizações da variável aleatória residual.

Estimação pelo método dos mínimos quadrados

A cada observação $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$ está associado um resíduo

$$d_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki})$$

O objectivo é escolher $b_0, b_1, \ldots, b_k$ de modo a minimizar a soma dos quadrados dos resíduos.

$$SSE = \sum_{i=1}^{n} d_i{}^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2$$

Estimação pelo método dos mínimos quadrados

Para determinar $b_0, b_1, \ldots, b_k$, de modo a minimizar SSE resolve-se o seguinte sistema de equações:

$$\frac{\partial SSE}{\partial b_0} = 0 \quad \wedge \quad \frac{\partial SSE}{\partial b_1} = 0 \quad \wedge \quad \ldots \quad \wedge \quad \frac{\partial SSE}{\partial b_k} = 0$$

Obtém-se $b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X^T X)^{-1} X^T y$ estimativa para $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

O estimador é $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \end{bmatrix} = (X^T X)^{-1} X^T Y.$

Estimação pelo método dos mínimos quadrados

Cada coeficiente de regressão estimado b_i, $i = 1, \ldots, k$ (estimativa de β_i), estima o efeito sobre o valor médio da variável dependente Y de uma alteração unitária da variável independente X_i, mantendo-se constantes todas as restantes variáveis independentes.

No caso $k = 1$ (regressão simples) temos:

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \left(X^T X\right)^{-1} X^T y,$$

onde X tem apenas duas colunas.

Como já vimos, b_0 e b_1 podem também ser determinados pelas relações:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

Exemplo

Os dados apresentados no quadro seguinte representam as vendas, Y, em milhares de Euros, efectuadas por 10 empregados de uma dada empresa, o nº de anos de experiência de cada vendedor, X_1 e o respectivo score no teste de inteligência, X_2.

Vendedor	Vendas (Y)	Anos de experiência(X_1)	Score no teste de inteligência (X_2)
1	9	6	3
2	6	5	2
3	4	3	2
4	3	1	1
5	3	4	1
6	5	3	3
7	8	6	3
8	2	2	1
9	7	4	2
10	4	2	2

Exemplo

Pretende-se determinar se o sucesso das vendas pode ser medido em função das duas variáveis explicativas X_1 e X_2 através de um modelo linear .

Matriz significativa do modelo: $X = \begin{bmatrix} 1 & 6 & 3 \\ 1 & 5 & 2 \\ 1 & 3 & 2 \\ 1 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 3 & 3 \\ 1 & 6 & 3 \\ 1 & 2 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix}$

Vector das observações da var. dependente: $y = [9\ 6\ 4\ 3\ 3\ 5\ 8\ 2\ 7\ 4]^T$

Exemplo

Vector das estimativas dos coeficientes de regressão:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} -0.262712 \\ 0.745763 \\ 1.338983 \end{bmatrix}$$

Equação de regressão estimada:

$$\hat{y} = \hat{\mu}_{Y|x_1,x_2} = -0.262712 + 0.745763x_1 + 1.338983x_2$$

$\hookrightarrow$ Estima-se que o volume médio de vendas de um vendedor (em milhares de Euros) é igual a 0.745763 vezes os seus anos de experiência mais 1.338983 vezes o seu score no teste de inteligência menos 0.262712.

Exemplo

Por exemplo, o volume médio de vendas para vendedores com 4 anos de experiência e com score 3 no teste de inteligência é estimado por:

$$\hat{y} = -0.262712 + 0.745763 \times 4 + 1.338983 \times 3 = 6.737289$$

$b_1 = 0.745763 \mapsto$ Em média, um ano extra de experiência entre vendedores com o mesmo score no teste de inteligência, conduz a um aumento no volume de vendas de uma quantidade que pode ser estimada em 745.763 Euros.

$b_2 = 1.338983 \mapsto$ Em média, um vendedor com score no teste de inteligência igual a 2 vende mais 1338.983 Euros (valor estimado) do que um vendedor com a mesma experiência e score 1, e menos 1338.983 Euros do que um vendedor com a mesma experiência e com score 3.

Exemplo

Atenção:

- $b_0 = -0.262712$ não pode ser interpretado como sendo o volume médio de vendas de um vendedor hipotético sem experiência prévia e com score zero no teste de inteligência. Com efeito, vendas negativas são impossíveis. Note que valores nulos de X_1 e X_2 encontram-se fora do âmbito dos dados.
- Trata-se de uma relação média, assim um vendedor com determinados anos de experiência e determinado score no teste de inteligência não obterá necessariamente o volume de vendas exacto indicado pela equação.

Qualidade do ajustamento

A equação de regressão estimada pode ser vista como uma tentativa para explicar as variações na variável dependente Y que resultam das alterações nas variáveis independentes $X_1, X_2, \ldots, X_k$.

Seja $\overline{y}$ a média dos valores observados para a variável dependente.

Uma medida útil associada ao modelo de regressão é o grau em que as predições baseadas na equação, $\hat{y}_i$, superam as predições baseadas em $\overline{y}$.

Se a dispersão (erro) associada à equação é muito menor que a dispersão (erro) associada a $\overline{y}$, as predições baseadas no modelo serão melhores que as baseadas em $\overline{y}$.

Qualidade do ajustamento

Pode-se mostrar que:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \;=\; \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \;+\; \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

$$\begin{array}{ccccc} \downarrow & & \downarrow & & \downarrow \\ SST & = & SSE & + & SSR \end{array}$$

$SST \longmapsto$ Soma dos quadrados totais - Variação total

$SSE \longmapsto$ Soma dos quadrados dos resíduos - Variação não explicada

$SSR \longmapsto$ Soma dos quadrados da regressão - Variação explicada

Isto é:

Variação Total de Y à volta da sua média	=	Variação que o ajustamento não consegue explicar	+	Variação explicada pelo ajustamento

Coeficiente de determinação

O quociente entre *SSR* e *SST* dá-nos uma medida da proporção da variação total que é explicada pelo modelo de regressão. A esta medida dá-se o nome de coeficiente de determinação (r^2),

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = 1 - \frac{SSE}{SST}$$

Note que:

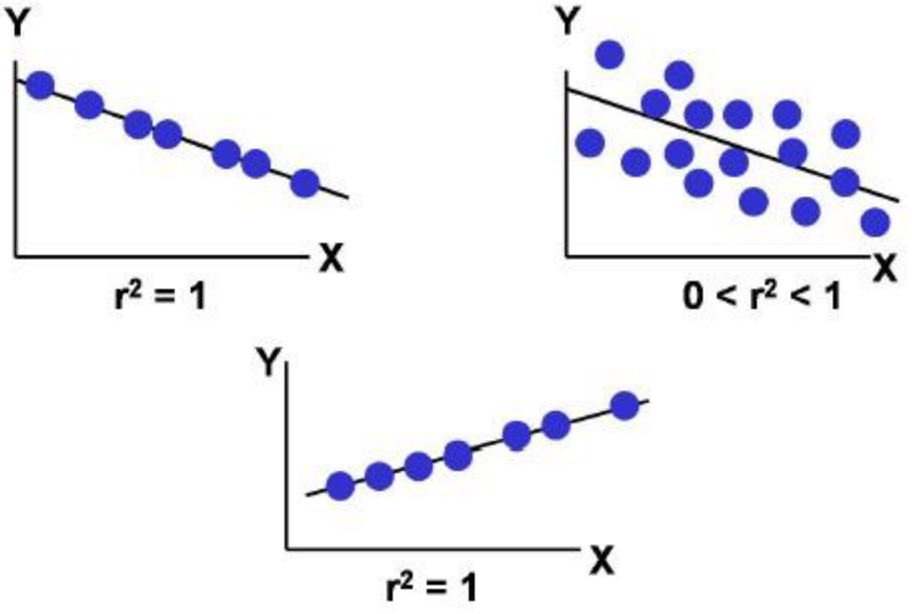
- $0 \leq r^2 \leq 1$;
- $r^2 \cong 1$ (próximo de 1) significa que grande parte da variação de Y é explicada linearmente pelas variáveis independentes;
- $r^2 \cong 0$ (próximo de 0) significa que grande parte da variação de Y não é explicada linearmente pelas variáveis independentes.

Coeficiente de determinação

Este coeficiente pode ser utilizado como uma medida da qualidade do ajustamento, ou como medida da confiança depositada na equação de regressão como instrumento de previsão:

- $r^2 \cong 0 \longmapsto$ modelo linear muito pouco adequado;
- $r^2 \cong 1 \longmapsto$ modelo linear bastante adequado.

Exemplos de diagrama de dispersão



Coeficiente de correlação

À raiz quadrada de r^2 dá-se o nome de:

- coeficiente de correlação simples se está envolvida apenas uma variável independente;
- coeficiente de correlação múltiplo se estão envolvidas pelo menos duas variáveis independentes.

Coeficiente de correlação simples

$$r = \pm\sqrt{r^2} \quad \text{(com o sinal do declive } b_1\text{)}$$

Este coeficiente é uma medida do grau de relacionamento linear entre as variáveis X e Y.

- r varia entre -1 e 1;
- $r = -1$ e $r = 1$ indicam a existência de uma relação linear perfeita (negativa e positiva respectivamente) entre X e Y;
- $r = 0$ indica a inexistência de qualquer relação ou tendência linear entre X e Y;
- $r > 0$ indica uma relação linear positiva entre as variáveis X e Y, ou seja, as variáveis tendem a variar no mesmo sentido;
- $r < 0$ indica uma relação linear negativa entre as variáveis X e Y, ou seja, as variáveis tendem a variar em sentido inverso.

Coeficiente de correlação simples

O coeficiente de correlação simples r também pode ser calculado a partir da seguinte fórmula:

$$r = \pm \sqrt{\frac{b_0 \sum_{i=1}^{n} y_i + b_1 \sum_{i=1}^{n} y_i x_i - n\bar{y}^2}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \quad \text{(com o sinal do declive } b_1\text{)}$$

Coeficiente de correlação múltiplo

É uma medida do grau de associação linear entre Y e o conjunto de variáveis $X_1, X_2, \ldots, X_k$.

- r varia entre 0 e 1;
- $r = 1$ indica a existência de uma associação linear perfeita, ou seja, Y pode ser expresso como uma combinação linear de $X_1, X_2, \ldots, X_k$;
- $r = 0$ indica a inexistência de qualquer relação linear entre a variável dependente Y e o conjunto de variáveis independentes $X_1, X_2, \ldots, X_k$.

Exemplo

Para o exemplo em estudo temos a seguinte tabela

i	y_i	x_{1i}	x_{2i}	$\hat{y}_i$	d_i $= y_i - \hat{y}_i$	d_i^2	$(y_i - \overline{y})^2$	$(\hat{y}_i - \overline{y})^2$
1	9	6	3	8,22881	0,77119	0,59473	...	...
2	6	5	2	6,14407	-0,14407	0,02076	...	...
3	4	3	2	4,65254	-0,65254	0,42581	...	...
4	3	1	1	1,82203	1,17797	1,38760	...	...
5	3	4	1	4,05932	-1,05932	1,12216	...	...
6	5	3	3	5,99153	-0,99153	0,98312	...	...
7	8	6	3	8,22881	-0,22881	0,05236	...	...
8	2	2	1	2,56780	-0,56780	0,32239	...	...
9	7	4	2	5,39831	1,60169	2,56543	...	...
10	4	2	2	3,90678	0,09322	0,00869	...	...
Total	51					SSE =7.48305	SST =48.9	SSR =41.41695

Coeficiente de determinação

$$r^2 = \frac{SSR}{SST} = \frac{41.41695}{48.9} = 0.84697$$

- 84.7% da variação nas vendas está relacionada linearmente com variações nos anos de experiência e no QI.
- As duas variáveis independentes utilizadas no modelo linear ajudam a explicar cerca de 84.7% da variação nas vendas.
- Ficam por explicar 15.3% das variações no volume de vendas, que se devem a outros fatores não considerados, como por exemplo:
  - a simpatia do vendedor
  - a reputação do vendedor
  - a condição do ambiente da loja
  - etc.

Coeficiente de correlação múltiplo

$$r = \sqrt{0.84697} = 0.92031$$

- Indica a existência de uma associação linear forte entre o volume de vendas e as variáveis independentes X_1 e X_2, anos de experiência e *escore* no teste de inteligência.

- Podemos então concluir que o modelo linear se afigura bastante adequado para descrever o relacionamento entre a variável Y, volume de vendas, e as variáveis X_1 e X_2.

Exemplos

- Planilha
- R-Studio

Operações no R

valores da distribuição dos resíduos

```
> modelo1var <- lm(gastos$gastos ~ gastos$horas)
> summary (modelo1var)

Call:
lm(formula = gastos$gastos ~ gastos$horas)

Residuals:
    Min       1Q   Median       3Q      Max
-103.802  -23.996    6.326   29.230   69.230

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   176.577     42.991   4.107  0.00124 **
gastos$horas   16.710      2.343   7.132 7.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.07 on 13 degrees of freedom
Multiple R-squared:  0.7964,     Adjusted R-squared:  0.7808
F-statistic: 50.86 on 1 and 13 DF,  p-value: 7.683e-06
```

Coeficientes:
Estimate *são os valores do intercepto (b0) e b1 (coef. da variável "horas")*
Std. Error *é o erro padrão no cálculo do coeficiente*
t value *é o valor t do teste*
Pr(>|t|) *é o p-value do teste t, = probabilidade de obter um t-value maior que o observado sob H0 válida; representa a probabilidade que a variável não seja relevante para o modelo, isto é, se é alto os coeficientes não são significantes*

Validando um modelo lm

- O que buscar para validar o modelo, para verificar a sua eficácia:
  - ✓ ***p-value****: devem ser significantes, isto é, menores que o nível de significância estatística determinado (≤ **α**, ex: ≤0.05).*
  - *No R:* ***Pr(>|t|)*** *é o p-value do teste t*
  - *A significância pode ser checada usando-se o modelo de variância ANOVA, que mostrará as somas quadráticas (SSR, SSE e SST), o valor F e o p-value*

Validando um modelo lm

- O que buscar para validar o modelo, para verificar a sua eficácia:
  - ✓ ***R^2 e R^2 ajustado****: nos permitem checar a linearidade (grau de associação linear entre as variáveis) e dizem o quanto a variação na variável dependente é explicada pelo modelo, e quanto maior (ex: > 0.7) melhor*

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

  - *R^2 ajustado penaliza ou ajusta o valor total de R^2 para o número de preditores usados no modelo (q)*
  - $R^2_{aj} = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-q)}{SST(n-1)}$*, onde MSE (mean squared error) e MST (mean squared total) e* $R^2_{aj} = 1 - (\frac{(1-R^2)(n-1)}{n-q})$

Validando um modelo lm

- O que buscar para validar o modelo, para verificar a sua eficácia:
  - ✓ ***Erro padrão**: indica a magnitude do desvio, e quanto mais próximo de zero melhor*

$$\text{erro padrão} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-q}}$$

  - ✓ ***Estatística F**: é uma medida do ajuste do modelo para explicar o fenômeno, e quanto mais alta, melhor*

$$\text{estatistica-f} = \frac{MSR}{MSE}$$

onde MSR (mean squared regression) é dado por $MSR = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{q-1} = \frac{SST-SSE}{q-1}$

E ainda

- Checagem de
  - **Linearidade →** pelo diagrama de dispersão e o coeficiente de determinação (R^2)
    - A linearidade pressupõe uma relação matemática representada por uma função de 1o grau
  - **Autocorrelação serial →** correlação dos resíduos ao longo do espectro de variáveis deve ser zero
    - O efeito das observações de X deve ser nulo sobre as observações seguintes de X, e os resíduos devem ser independentes entre si

E ainda

- Checagem de
  - **Homocedasticidade** (dos resíduos) → avaliação da hipótese da variância constante dos resíduos (que não devem apresentar nenhum padrão ou tendência) gerando
    - um gráfico dos *valores previstos* versus os *resíduos* de cada previsão:

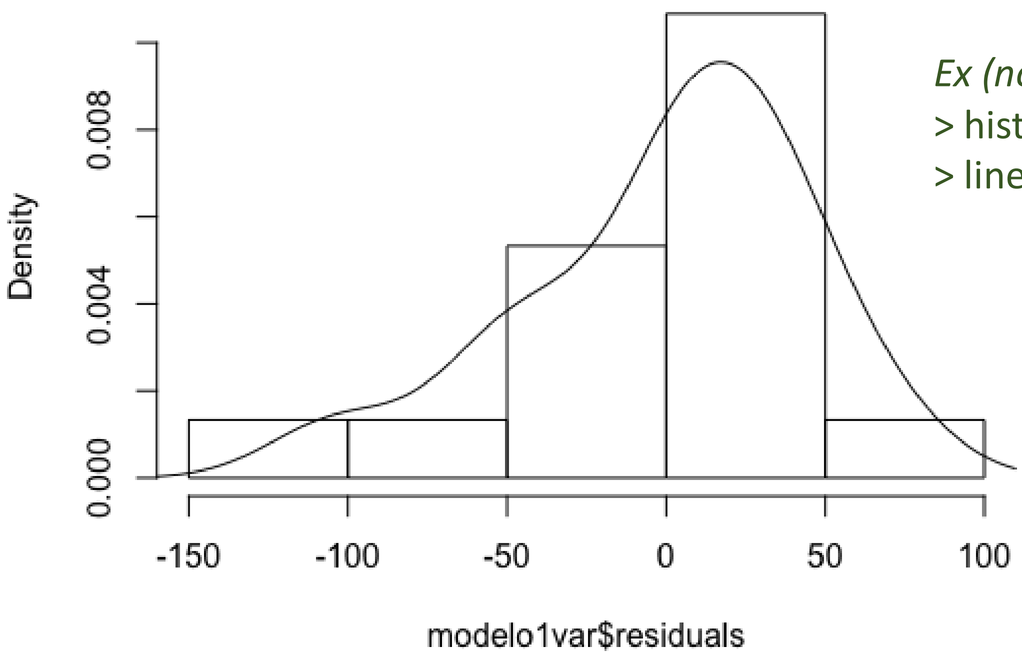
    Ex (no R): > plot(rstudent(modelo) ~ fitted(modelo))

    - um gráfico dos resíduos versus cada uma das variáveis independentes:

    Ex (no R): > plot(x=modelo$var_dep, y=modelo$var_indep)

E ainda

- Checagem de
  - Normalidade (dos resíduos) → avaliação da distribuição dos resíduos e assimetrias na curva, por meio do histograma e curva suavizada do histograma



Ex (no R):

> hist(x=modelo$residuals, probability = TRUE)
> lines (density(modelo$residuals))

Testes aplicáveis para checar normalidade:

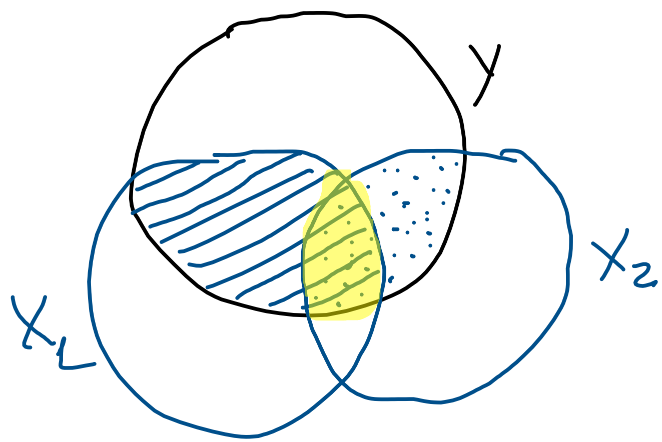
- *Kolmogorov-Smirnov*
- *Shapiro-Wilk*
- *Jarque-Bera*

Correlações parciais

- Ao acrescentar uma nova variável a um modelo de regressão, deve-se considerar também as inter-relações existentes (desconhecidas) entre as variáveis independentes

  → isto pode ser responsável por incrementos menores no poder explicativo do modelo, pois uma parte do poder preditivo é compartilhada pelas variáveis (e já estava presente na 1ª variável)

var. depend.: Y
var. independ.: X_1, X_2



Fim