Análise de Regressão e Correlação Linear Simples

A análise de **REGRESSÃO** e **CORRELAÇÃO** compreendem a análise de dados amostrais para saber **se** e **como** um certo conjunto de variáveis está relacionado com outra variável.

Aplicações da regressão:

- ★ Estimar valores de uma variável com base em valores conhecidos de outra;
- ★ Explicar valores de uma variável em termos de outra;
- ★ Predizer valores futuros de uma variável.

Análise de regressão: estuda o relacionamento entre uma variável chamada a variável dependente (Y) e outras variáveis chamadas variáveis independentes (X_1, X_2, . . ., X_K).

O termo variável dependente implica geralmente uma relação do tipo causa-efeito, porém a RL pode ser usada para modelar a relação funcional entre duas varáveis – ié uma relação que pode ser expressa através de uma função matemática – independentemente de existir ou não um relação do tipo causa-efeito.

- O relacionamento entre as variáveis é representado por um *modelo matemático*, isto é, por uma equação que associa a variável dependente com as variáveis independentes. Este modelo é designado por **modelo de regressão linear simples** se define uma relação linear entre a variável dependente (**Y**) e uma variável independente (**X**).

- Se em vez de uma, forem incorporadas várias variáveis independentes (X_1, X_2, . . ,X_K), o modelo passa a denominar-se **modelo de regressão linear múltipla - RLM.**

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.

As técnicas de análise de correlação e regressão estão intimamente ligadas.

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)

★ Pode suceder que dois homens adultos tenham a mesma altura e pesos diferentes e vice-versa. No entanto, *em média*, quanto maior for a altura maior será o peso.

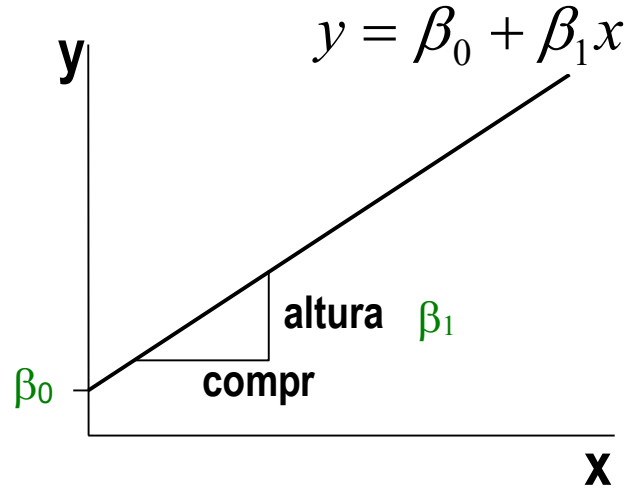
★ Relativamente ao segundo exemplo, pode também suceder que a colheitas iguais correspondam preços diferentes e vice-versa. No entanto, *em média*, quanto maior for a colheita menor será o preço do vinho.

É essa variação em média que vai ser estudada

A **Regressão Linear Simples** (RLS) constitui uma tentativa de estabelecer uma equação matemática linear (reta) que descreva o relacionamento entre duas variáveis.

O modelo de Regressão Linear Simples: $Y = \beta_0 + \beta_1 X + \varepsilon$

- Y → variável dependente
- X → variável independente
- β_0 → ordenada na origem
- β_1 → declive da reta
- ε → variável erro

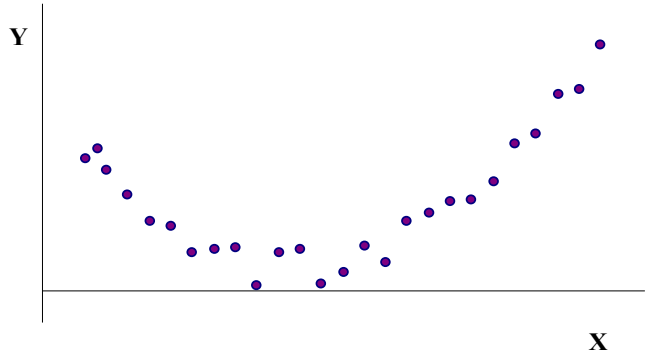


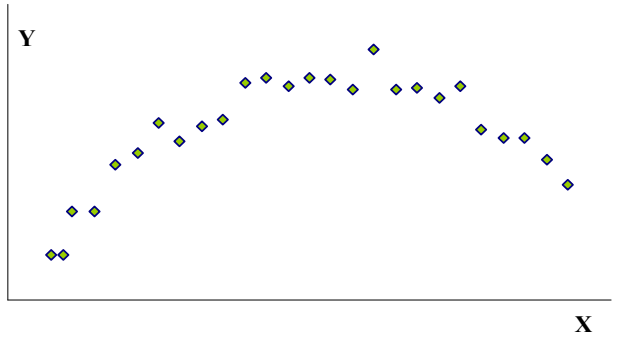
Com os dados constrói-se o **diagrama de dispersão**;

★ se este exibir uma tendência linear → regressão linear.

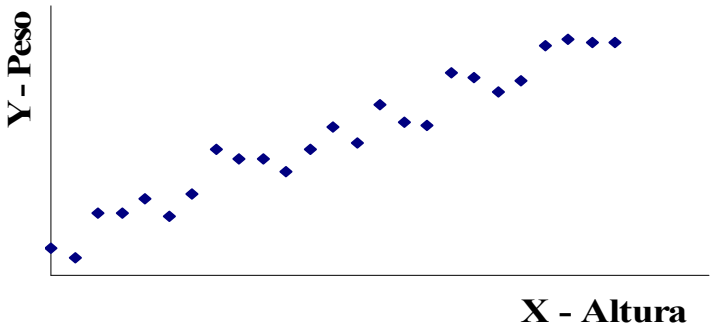
- pode-se também concluir (empiricamente) se o grau de relacionamento linear entre as variáveis **é forte** ou **fraco,** conforme o modo como se situam os pontos em redor de uma reta imaginária que passa através do enxame de pontos.
- se o declive da reta é positivo, concluímos que a **correlação entre X e Y é positiva**, i.e., os fenómenos variam no mesmo sentido. Ao contrário, se o declive é negativo, então **a correlação entre X e Y é negativa**, i.e., os fenómenos variam em sentido inverso.

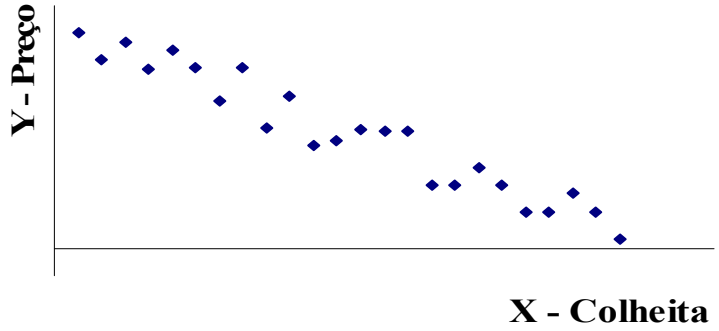
…Sugerem uma regressão não linear, i.e., a relação entre as duas variáveis poderá ser descrita por uma equação não linear





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





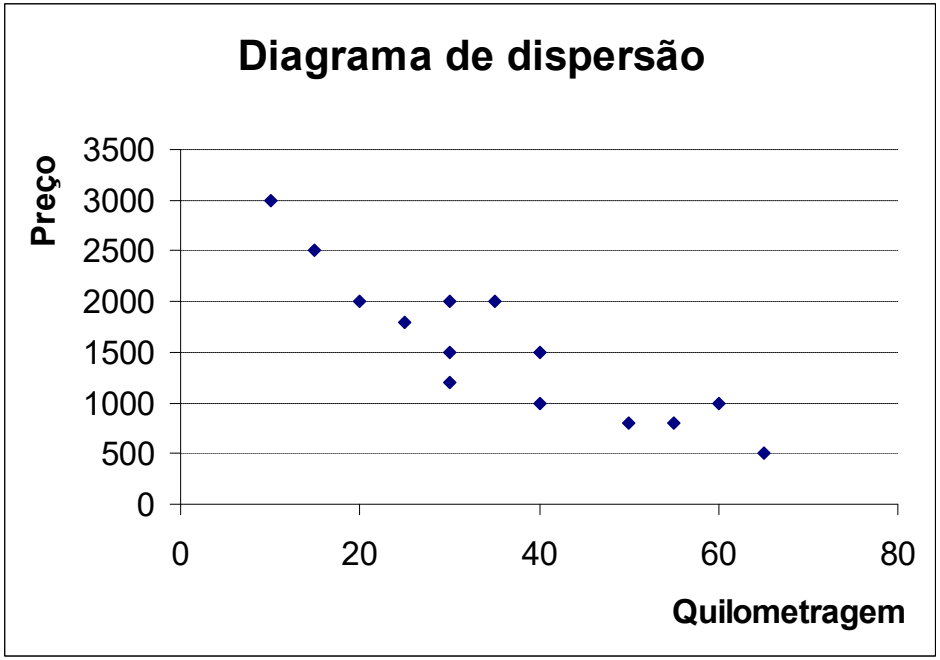
Existência de correlação positiva →em média, quanto maior for a altura maior será o peso

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

Exemplo 1:

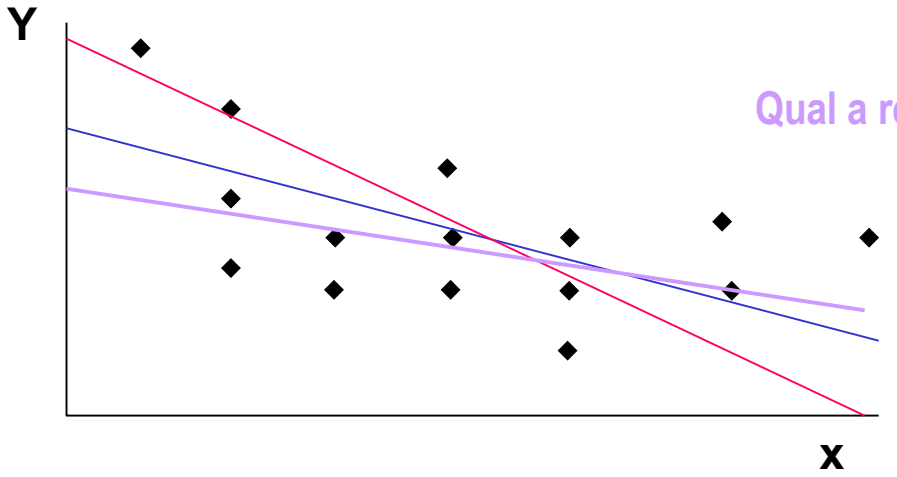
Queremos estudar a relação entre a quilometragem de um carro usado e o seu preço de venda

Carros	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**Total**
Quilometragem (1000 Km)	40	30	30	25	50	60	65	10	15	20	55	40	35	30	**505**
Preço de venda(x10 €)	1000	1500	1200	1800	800	1000	500	3000	2500	2000	800	1500	2000	2000	**21600**

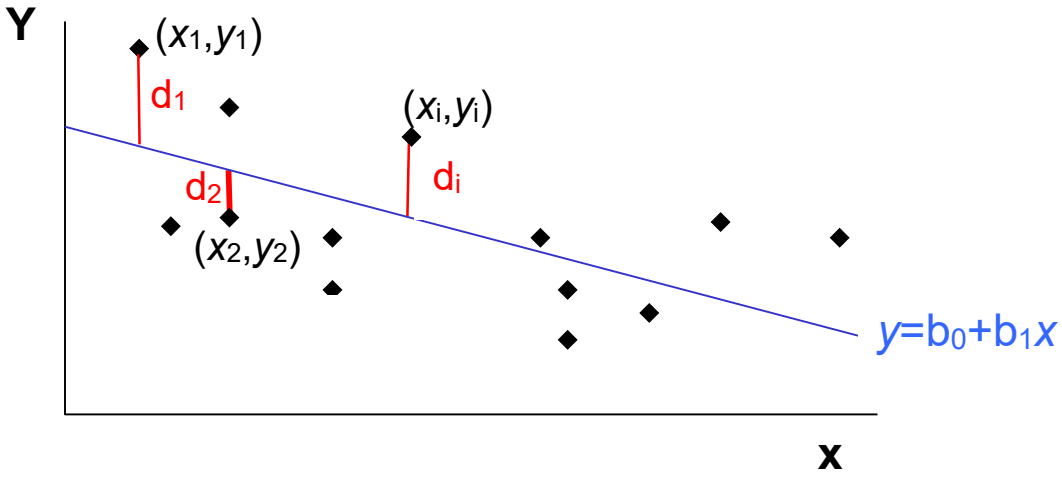


Os dados sugerem uma relação linear entre a quilometragem e o preço de venda. Existe uma **correlação negativa:** em média, quanto maior for a quilometragem menor será o preço de venda

Estimação dos coeficientes de regressão - Método dos Mínimos



Qual a reta que melhor se ajusta?



$y=b_0+b_1x$

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

é o valor dado pela reta

$d_i=y_i-(b_0+b_1x_i)$

resíduos

Objetivo: determinar b_0 e b_1 de modo a que os resíduos sejam tão pequenos quanto possível.

O método dos mínimos quadrados consiste em escolher b_0 e b_1 de modo a minimizar a soma dos quadrados dos resíduos d_i. Desta forma estamos essencialmente a escolher a reta que se aproxima o mais possível de todos os pontos dos dados simultaneamente.

$\bar{x} \rightarrow$ média dos valores observados de X $\qquad$ $\bar{y} \rightarrow$ média dos valores observados de Y

Para determinar b_0 e b_1 de modo a minimizar SSE:

$$\begin{cases} \dfrac{\partial SSE}{\partial b_0} = 0 \\ \dfrac{\partial SSE}{\partial b_1} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \\ -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0 \end{cases} \Leftrightarrow \cdots \Leftrightarrow \begin{cases} b_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\,\bar{x}^{\,2}} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}$$

Reta de regressão estimada: $\hat{y} = \hat{\mu}_{Y/x} = b_0 + b_1 x$

⇩

Estima o **valor médio de Y** para um dado valor de X

Mas é usada também para estimar o próprio valor de Y. De facto, o senso comum diz-nos que uma escolha razoável para predizer o valor de Y para um dado x de X, é o valor médio estimado $\hat{\mu}_{Y/x}$. **Por exemplo, se quiséssemos predizer o preço de venda de um carro com 57 000 km uma escolha lógica seria usar o preço médio de venda dos carros com esta quilometragem**

Voltando ao Exemplo 1:

Carros	Quilometragem X (1000 Km)	Preço de venda Y (dezena de Euros)	XY	X^2	Y^2
1	40	1000	40000	1600	1000000
2	30	1500	45000	900	2250000
3	30	1200	36000	900	1440000
4	25	1800	45000	625	3240000
5	50	800	40000	2500	640000
6	60	1000	60000	3600	1000000
7	65	500	32500	4225	250000
8	10	3000	30000	100	9000000
9	15	2500	37500	225	6250000
10	20	2000	40000	400	4000000
11	55	800	44000	3025	640000
12	40	1500	60000	1600	2250000
13	35	2000	70000	1225	4000000
14	30	2000	60000	900	4000000
Total	**505**	**21600**	**640000**	**21825**	**39960000**

Nota: Não vai ser exigido este cálculo

$$\bar{x} = \frac{505}{14} = 36.07$$

$$\bar{y} = \frac{21600}{14} = 1542.85$$

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

$$= \frac{640000 - 14 \times 36.07 \times 1542.85}{21825 - 14 \times 36.07^2} =$$

$$= -38.5$$

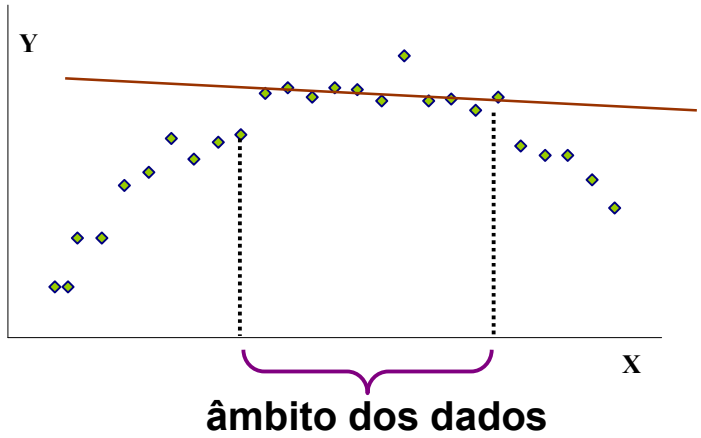
$$b_0 = \bar{y} - b_1 \bar{x} = 1542.85 +$$

$$+ 38.56 \times 36.07 = 2934$$

Reta de regressão estimada: $\hat{y} = \hat{\mu}_{Y/x} = 2934 - 38{,}56x$

- O preço esperado para um carro é de 2934 dezenas de euros menos 38.56 dezenas de Euros por cada mil quilómetros que o carro tenha andado.
- Para um carro que tenha andado 20 000 km, a equação sugere o preço:
  $\hat{y} = 2934 - 38.56 \times 20 = 2162.8$ dezenas de euros.
- Em média, por cada 1000km que o carro tenha andado, o preço de venda baixa 38.56 dezenas de euros.
- **b_0=2934** → não pode ser interpretado como sendo o preço previsto para um carro novo, *0 Km*, pois este valor de quilometragem encontra-se fora do âmbito dos dados. Apenas teria sentido fazer esta interpretação se existissem observações para x=0.
- Trata-se de uma relação média, assim um carro com determinada quilometragem não obterá necessariamente o preço de venda exato indicado pela equação.

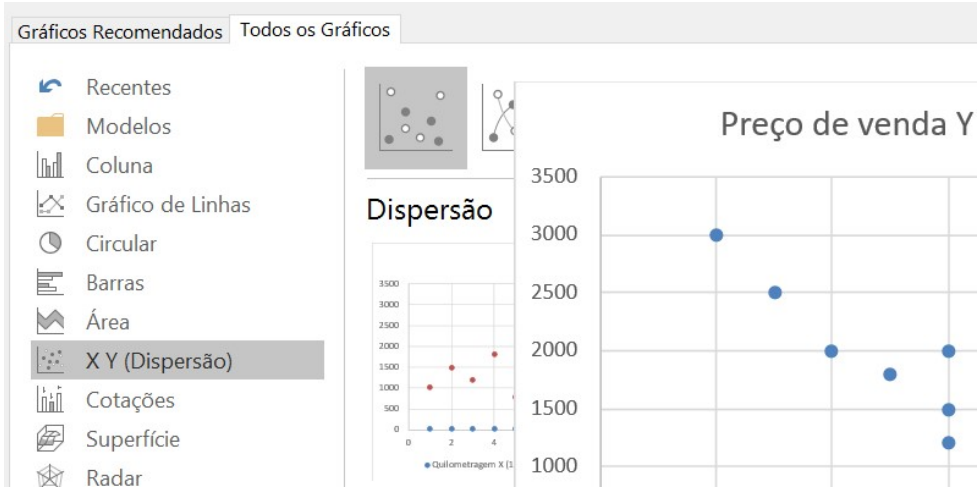
CUIDADO!!!! Um conjunto de pontos dá evidência de linearidade apenas para os valores de X cobertos pelo conjunto de dados. Para valores de X que saem fora dos que foram cobertos não há qualquer evidência de linearidade. Por isso é arriscado usar uma reta de regressão estimada para predizer valores de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



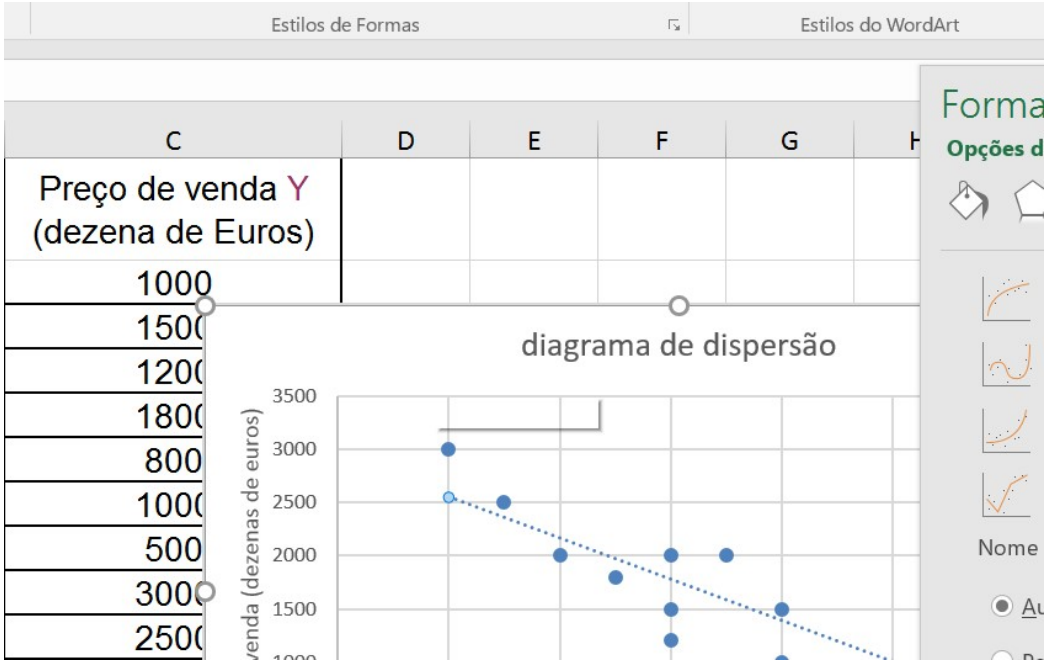
O perigo de extrapolar para fora do âmbito dos dados amostrais é que a mesma relação possa não mais se verificar.

Como fazer o diagrama de dispersão no `Excel:`

`Menu inserir>gráficos>dispersão`



Selecionando os marcadores da série, é possível adicionar a linha de regressão e obter a equação da reta estimada:



Coeficiente de determinação: para RLS pode ser calculado através da fórmula:

$$r^2 = \frac{b_0 \sum_{i=1}^{n} y_i + b_1 \sum_{i=1}^{n} y_i x_i - n\bar{y}^2}{\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2}$$

r^2 mede a proporção de variação da variável dependente Y que é explicada pelo modelo, isto é, pela equação de regressão ajustada, ou equivalentemente, é a proporção da variação de Y explicada em termos lineares pelas variáveis independentes. Neste sentido 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

- $0 \leq r^2 \leq 1$;
- $r^2 \cong 1$ (próximo de 1) significa que grande parte da variação de Y é explicada linearmente pela variável independente; quanto mais próximo de 1 estiver o coeficiente melhor é o "grau de ajustamento", ou seja maior é a "proximidade" entre os Y_i e os $\hat{Y}_i$.

- $r^2 \cong 0$ (próximo de 0) significa que grande parte da variação de Y não é explicada linearmente pela variável independente.

★ À raiz quadrada de r^2 dá-se o nome de **coeficiente de correlação** (linear de Pearson) simples ou múltiplo consoante tenhamos uma ou mais variáveis independentes

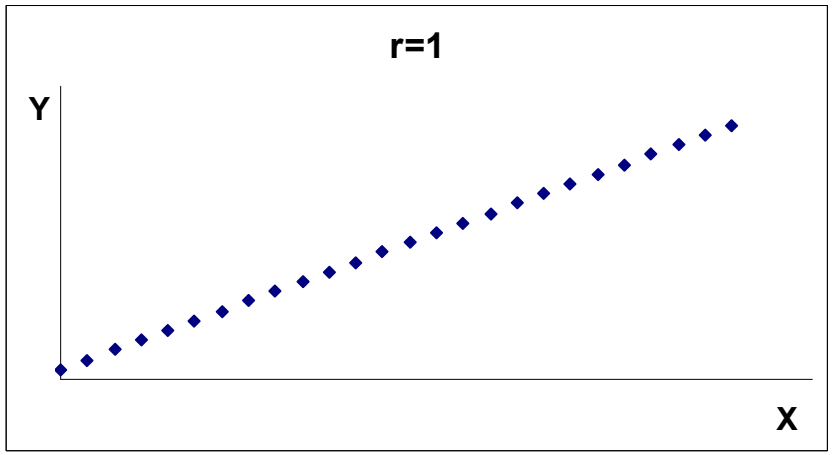
$$r = \pm\sqrt{r^2} \text{ (com o sinal do declive } b_1)$$

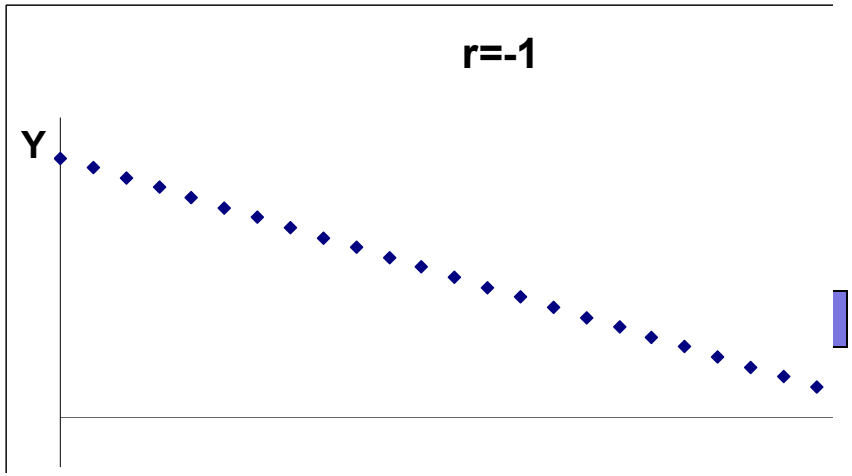
Coeficiente de correlação Simples:

É uma medida do grau de associação **linear** entre as variáveis X e Y.

- $-1 \leq r \leq 1$;
- $r > 0$ (positivo) indica que as duas variáveis tendem a variar no mesmo sentido, isto é, em média um aumento na variável X provocará um aumento na variável Y;

- r<0 (negativo) indica que as duas variáveis tendem a variar em sentido inverso, isto é, em média um aumento na variável X provocará uma diminuição na variável Y;
- r=1 e r=-1 indicam a existência de uma relação linear perfeita entre X e Y, positiva e negativa respetivamente;
- r=0 indica a inexistência de qualquer relação ou tendência linear entre X e Y podendo no entanto existir uma relação não linear entre elas. Isto é, é possível que as duas variáveis estejam fortemente associadas (movimentos numa variável estão associados a movimentos na outra) sem que o relacionamento seja linear.





Voltando ao exemplo 1 (carros preço/ km):

$$r^2 = \frac{b_0 \sum_{i=1}^{n} y_i + b_1 \sum_{i=1}^{n} y_i x_i - n\bar{y}^2}{\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2} = \frac{2934 \times 21600 - 38.56 \times 640000 - 14 \times 1542.85^2}{39960000 - 14 \times 1542.85^2} = 0.809$$

r^2=0.809→ aproximadamente 81% da variação no preço de venda dos carros está relacionada linearmente com a variação na quilometragem rodada, i.e., 81% dessa variação é explicada por variações na quilometragem. 19% não é explicada por variações na quilometragem e é resultante de outros fatores não considerados (que podem influir no preço de venda), como por exemplo:

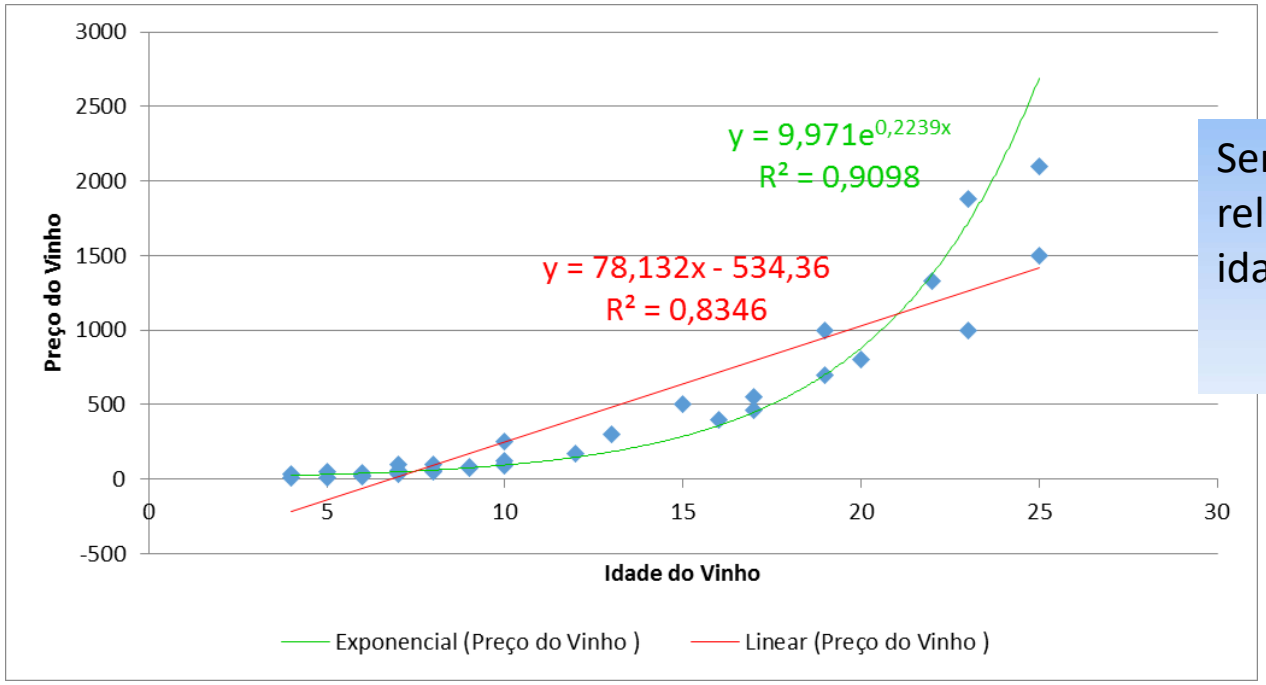
- as condições gerais do carro; a localização/reputação do vendedor;
- a necessidade que o comprador tem do carro; o nº de registos de propriedade do carro, etc.

- $r = -\sqrt{0.809} = -0.899 \rightarrow$ indica que o grau de relacionamento **linear** entre as variáveis é forte. A correlação é negativa, pois um acréscimo na quilometragem é, tendencialmente, acompanhado por um decréscimo do preço de venda.

Exemplo 2:

Na tentativa de explicar as variações do preço do vinho tinto de uma determinada zona demarcada, recolheram-se registos de preços (em unidades monetárias – u.m.) e da idade do vinho (em anos).

Idade do Vinho	25	22	20	19	23	17	17	16	19	13	12	10	10	10	9	9	10
Preço do Vinho	2100	1325	800	700	1000	550	460	400	1000	300	170	100	125	89	79	70	250
Idade do Vinho	8	8	8	7	7	7	7	6	6	6	5	5	4	4	15	25	23
Preço do Vinho	99	55	51	60	39	34	100	39	25	18	50	8	35	11	500	1500	1880



Será que existe alguma relação entre o preço e a idade do vinho?

SERÁ LINEAR?